LFMD: detecting low-frequency mutations in high-depth genome sequencing data without molecular tags

Rui Ye[1,2,7*], Xuehan Zhuang[3*], Jie Ruan[2,7*], 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).

Rui Ye, yerui@connect.hku.hk

Jie Ruan, rj_0625405006@126.com

Xuehan Zhuang, sunnyzxh@connect.hku.hk

Yanwei Qi, ywqi.cu@gmail.com

Yitai An, ayt_1988@126.com

Jiaming Xu, xujm@haplox.com

Timothy Mak, timmak@yahoo.com

Xiao Liu, airson_liu@163.com

Xiuqing Zhang, zhangxq@genomics.cn

Huanming Yang, yanghm@genomics.cn

Xun Xu, xuxun@genomics.cn

Larry Baum, lwbaum@hku.hk

Chao Nie, niechao@genomics.cn

Pak Chung Sham, pcsham@hku.hk, (852) 3917 9557

Abstract

As next-generation sequencing (NGS) and liquid biopsy become more prevalent in research and in the clinic, there is an increasing need for better methods to reduce cost and improve sensitivity and specificity of low-frequency mutation detection (where the Alternative Allele Frequency, or AAF, is less than 1%). Here we propose a likelihood-based approach, called Low-Frequency Mutation Detector (LFMD), which combines the advantages of duplex sequencing (DS) and the bottleneck sequencing system (BotSeqS) to maximize the utilization of duplicate reads. Compared with the existing state-of-the-art methods, DS, Du Novo, UMI-tools, and Unified Consensus Maker, our method achieves higher sensitivity, higher specificity ($< 4 \times 10^{-10}$ errors per base sequenced) and lower cost (reduced by ~70% at best) without involving additional experimental steps, customized adapters or molecular tags. LFMD is useful in areas where high precision is required, such as drug resistance prediction and cancer screening. As an example of LFMD's applications, mitochondrial heterogeneity analysis of 28 human brain samples across different stages of Alzheimer's Disease (AD) showed that the canonical oxidative damage related mutations, C:G>A:T, are significantly increased in the mid-stage group. This is consistent with the Mitochondrial Free Radical Theory of Aging, suggesting that AD may be linked to the aging of brain cells induced by oxidative damage.

Keywords

Low-frequency mutation, liquid biopsy, drug resistance, cancer diagnosis, mitochondria heterogeneity, somatic mutation detection

Background

Low-frequency mutations (LFMs) are defined as somatic mutations with allele frequency lower than 1% in an individual's DNA or mixed DNA. LFMs can indicate early stages of diseases (such as cancer[1-3], Alzheimer's Disease (AD)[4] and Parkinson's Disease (PD)[5]), study aging[5, 6], identify disease-causing variants[7], predict potential drug resistance[8], diagnose mitochondrial disease before tri-parental *in vitro* fertilization[9], and track the mutational spectrum of viral genomes, malignant lesions, and somatic tissues[8, 10]. To effectively improve signal-to-noise ratio (SNR) and detect LFMs, researchers have developed methods with stringent thresholds, complex experimental procedures[4, 11], single cell sequencing[12-15], circle sequencing[16], o2n-seq[17], Sef-SeqS[18, 19], and more precise analytic models[6, 20-22]. The bottleneck sequencing system[23] (BotSeqS) and duplex sequencing[24, 25] (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

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[15]. 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.

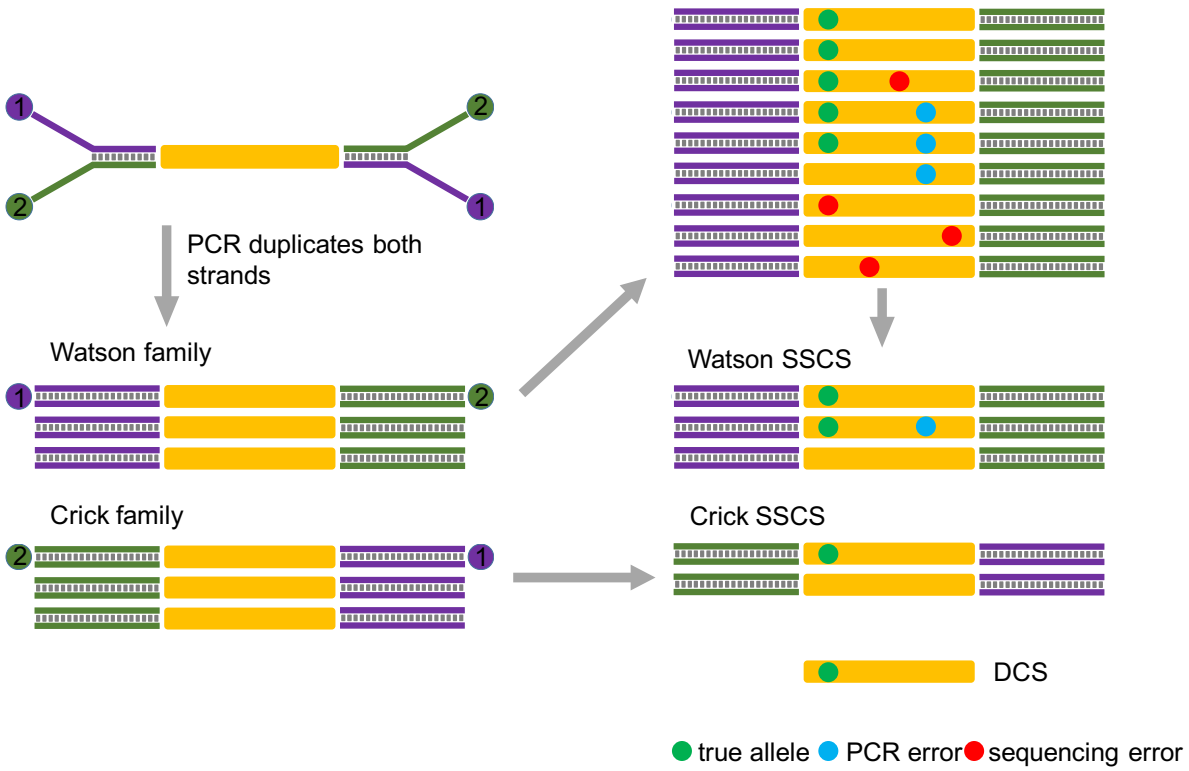
Disadvantages of circle sequencing and o2n-seq

87 Circle sequencing and o2n-seq only utilize single-strand DNA templates, so their specificities
88 are limited by the error rate of PCR. Circle sequencing controls errors to a rate as low as $7.6 \times$
89 10^{-6} per base sequenced[16] and o2n-seq's error rate ranges from 10^{-5} to 10^{-8} per base
90 sequenced[26], while DS can achieve 4×10^{-10} errors per base sequenced[24].
91
92 **Disadvantages of BotSeqS**
93 In contrast, BotSeqS uses endogenous molecular tags to group reads from the same DNA
94 template and construct double-strand consensus reads. As a result, it can detect very rare
95 mutations ($<10^{-6}$) while being cheap enough to sequence the whole human genome[23].
96 However, it requires highly diluted DNA templates before PCR amplification to reduce
97 endogenous tag conflicts and ensure sufficient sequencing of each DNA template. Thus, it has
98 a high specificity with poor sensitivity. Also, it discards clonal variants and small
99 insertions/deletions (indels) to limit false positives.
100
101 **Disadvantages of DS**
102 DS ligates exogenous random molecular tags (also known as barcode, unique molecular
103 identifier, UID or UMI) to both ends of each DNA template prior to PCR amplification.
104 Although sensitive and accurate, much sequencing data is wasted on sequencing tags, fixed
105 sequences and a large proportion of read families that contain less than three read pairs. The
106 errors on molecular tags lead to false construction of read families, smaller read families, and
107 lower utilization rate of sequencing data. In addition, since tags are synthesized with
108 customized adapters, batch effects may occur during DNA library construction.
109
110 **Disadvantages of tag clustering**

111 To solve the problem induced by the errors on tags, multiple methods (e.g. UMI-tools) have
112 been developed to cluster similar tags[27-29], where one or two mismatches are allowed in
113 merging two tags. Although tag clustering does improve the sensitivity, it is still not a
114 straightforward way to solve the problem. Inappropriate tag clustering might occur because
115 unstable synthesis of random tags can result in distinct but similar tags.
116
117 **A new approach**
118 To avoid the aforementioned problems, we present here a new, efficient approach that
119 combines the advantages of BotSeqS and DS. The method uses a likelihood-based model[6,
120 20, 30] to dramatically reduce endogenous tag conflicts. Then it groups reads into read families
121 and constructs double-strand consensus reads to detect ultra-rare mutations accurately while
122 maximizing the utilization of duplicate read pairs. This simplifies the DNA sequencing
123 procedure, saves data and cost, achieves higher sensitivity and specificity, and can be used in
124 whole genome sequencing if coupled with a dilution step prior to PCR amplification. In
125 addition, our new method offers a statistical solution to the problem of PCR duplication in the
126 basic analysis pipeline of NGS data and can improve sensitivity and specificity of other variant
127 calling algorithms without specific experimental designs. As the price of sequencing is falling,
128 the depth and the rate of PCR duplication are rising. The method we present here might help
129 deal with such high-depth data more accurately and efficiently.
130
131 Since the LFMs of multiple samples need to be compared collectively to unveil the relationship
132 between LFMs and phenotypes, the detection sensitivities in all samples should be estimated
133 carefully, based on the differences in sequencing depth, initial DNA concentration, mutation
134 frequency, PCR amplification efficiency, double-strand consensus sequence (DCS) depth, etc.
135 Meanwhile, it is known that these factors significantly affect the sensitivity of LFM detection.

136 If the results are compared directly without adjustment, systematic bias may conceal the real
137 differences between samples. Therefore, it is essential to estimate the sensitivity of the method
138 for each position at which any mutation was detected.
139
140 **Results**
141
142 **Overview of LFMD**
143 The strategy of LFMD (pronounced "leaf mid") is quite intuitive, like "looking for a needle in
144 a haystack": split the haystack into smaller ones, and then look for the needle in the smaller
145 haystacks. In LFMD, the needle is a DNA fragment that has a mutation and several PCR
146 duplicates; the clustering of read pairs with the same genome coordinates is splitting the
147 haystack into smaller ones. Because the allele frequency in the smaller haystack is much higher
148 than that in the original haystack, it is much easier to find the needle; i.e., sensitivity is increased.
149
150 To group reads from the same DNA template, the simplest idea is to group properly mapped
151 reads with the same coordinates (i.e., chromosome, start position, and end position) because
152 random shearing of DNA molecules can provide natural differences, called endogenous tags,
153 between templates. A group of reads is called a read family. However, as the length of each
154 DNA template is only approximately determined, random shearing cannot provide enough
155 differences to distinguish each DNA template. Thus, it is common that two original DNA
156 templates share the same coordinates. If two or more DNA templates shared the same
157 coordinates, and their reads are grouped into a single read family, it is difficult to determine,
158 using only their frequencies as a guide, whether an allele is a potential error or a mutation.
159 Thus, BotSeqS introduced a strategy of dilution before PCR amplification to dramatically
160 reduce the number of DNA templates in order to reduce the probability of endogenous tag

161 conflicts. And DS introduced exogenous molecular tags before PCR amplification to
162 dramatically increase the differences between templates. Thus, BotSeqS sacrifices sensitivity
163 and DS sequences extra data: the tags.
164
165 Here we introduce a third strategy to eliminate tag conflicts. It is a likelihood-based approach
166 based on an intuitive hypothesis: that if reads of two or more DNA templates group together
167 due to endogenous tags, a true allele's frequency in this read family is high enough to
168 distinguish the allele from background sequencing errors. The overview of the LFMD pipeline
169 is shown in Figure 1.



170
171 **Figure 1 Overview of LFMD.** The Y-shaped adapters determine read 1 (purple bar) and 2
172 (green bar). The directions of reads determine +/- strands. So, after the first cycle of the PCR
173 amplification, the Watson and Crick families are well defined. Then within a read family, true
174 alleles (green dots) and accumulated PCR errors (blue dots) are detected via the likelihood-
175 base model and given a combined error rate. Sequencing errors (red dots) are eliminated.

176 Combining single-strand consensus sequences (SSCSs) of paired read families, high-quality
177 double-strand consensus sequences (DCSs) with estimated error rates are generated and used
178 in the downstream analysis.
179
180 **Performance of the method on simulated data**
181 We modified Python scripts developed by the Du Novo[31] team to simulate mixed double-
182 strand sequencing data. A total of 800 single nucleotide variants (SNVs), 120 insertions (INSs),
183 and 120 deletions (DELs) were simulated in the human mitochondrial genome. Although the
184 simulation may not be entirely realistic, the results are still useful to evaluate the performance
185 and the potential drawbacks of LFMD, Du Novo[31], DS, UMI-tools, and Unified Consensus
186 Maker (UniC, an advanced version of DS). Because the underlying true mutations are known,
187 the true positives and false positives can be calculated (Figures 2 and 3). Among these tools,
188 DS and UniC are from the same lab. UniC is an updated version of DS. DS uses the genome
189 positions to groups read pairs based on their genome positions and considers molecular tags
190 later, while UniC groups read pairs into read families based on molecular tags prior to read
191 alignment. UMI-tools is used to cluster molecular tags to increase the size of read families. Du
192 Novo uses the same strategy as UniC but requires an extremely large amount of memory, and
193 its output consensus reads have embedded false indels which cause many false positives,
194 several magnitudes higher than that of other tools. So, in the following comparison, we do not
195 compare Du Novo with other tools. With different sequencing depths—33k, 65k, and 130k—
196 and distinct mutation frequencies, LFMD consistently has the best performance in terms of
197 sensitivity and specificity (Figures 2 and 3).

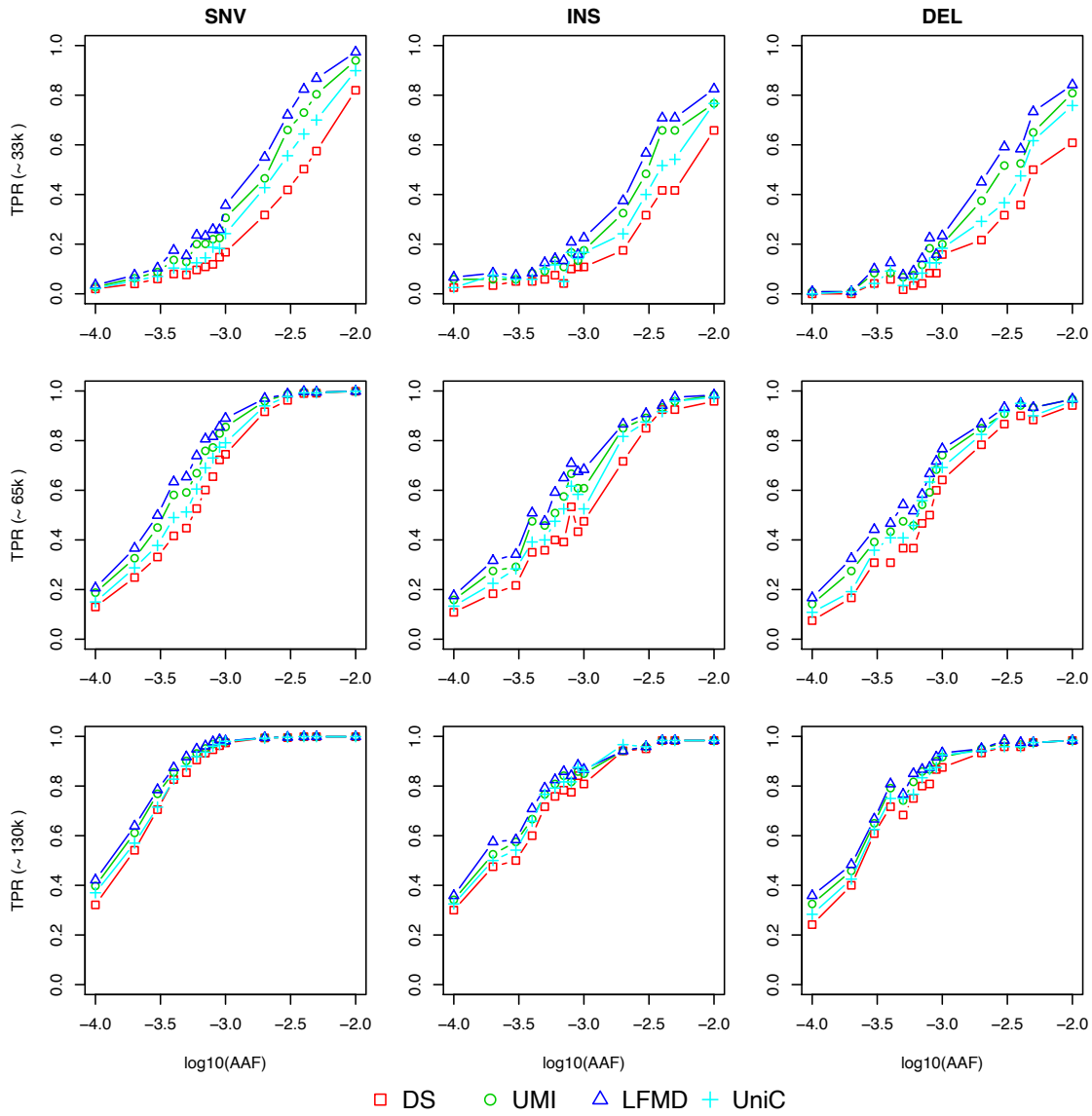


Figure 2 True positive rates of four tools on simulated data with depth around 33k, 65k, and 130k. DS, duplex sequencing; UMI, UMI-tools; UniC, Unified Consensus Maker; SNV, single nucleotide variant; INS, small insertion; DEL, small deletion; TPR, true positive rate; AAF, alternative allele frequency;

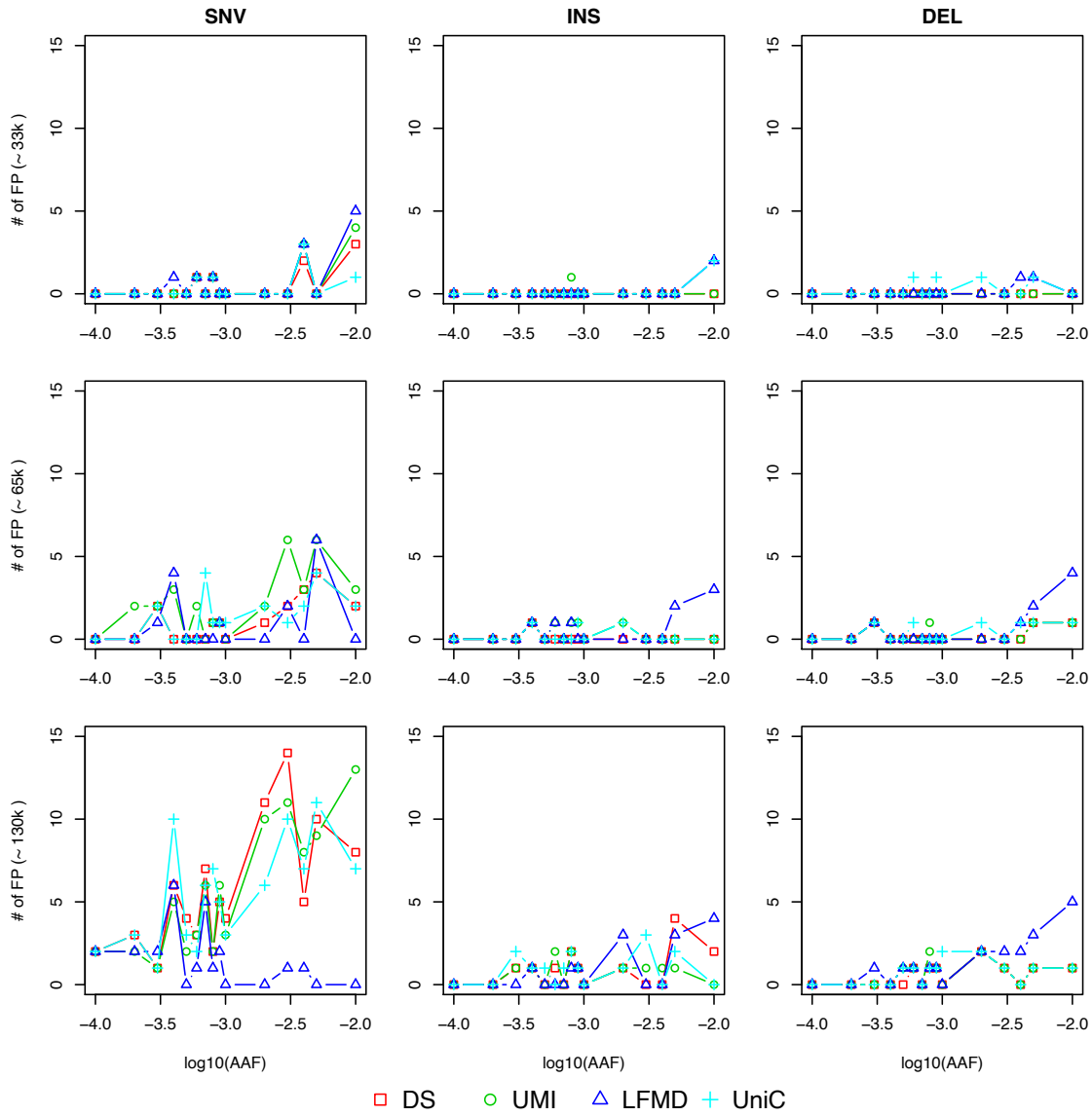


Figure 3 The number of false positives of four tools on simulated data with depth around 33k, 65k, and 130k. DS, duplex sequencing; UMI, UMI-tools; UniC, Unified Consensus Maker; SNV, single nucleotide variant; INS, small insertion; DEL, small deletion; FP, false positive; AAF, alternative allele frequency.

Performance of the method on real data

To evaluate the performance of LFMD on real data, we used consistent parameters to rerun the whole pipelines of DS, UMI-tools (UMI), Unified Consensus Maker (UniC), and LFMD in a

212 mouse mitochondrial DNA (mtDNA) dataset: SRR1613972 (Figure 4a). The pipelines are
213 designed to only allow fundamental differences between the four strategies and fix other
214 parameters in the analysis pipeline . Thus, the comparison between the results indicates the
215 underlying differences of the tools. To reduce the influence of mapping quality, we only use
216 unique, properly mapped, read pairs to call mutations.
217
218 We investigated the discordant mutations one by one. Three mutations only detected by DS
219 and UMI-tools may be false positives due to poor mapping qualities (MT:7G>A and
220 MT:3418T>del) and low sequencing qualities (MT:5462T>G). The mapping errors are from
221 the following: 1) DS and UMI-tools assigned the highest base quality for all bases of DCSs[24]
222 even if the base was N and 2) DS and UMI-tools introduced some Ns in DCS reads because
223 the minimum proportion of "true" bases in a read family is set to 0.7 by default (it can be set
224 to 0.5 to improve sensitivity). Forty-three mutations only detected by LFMD were due to four
225 technical reasons: 1) sequencing and PCR errors on tags led to smaller read families in DS,
226 decreasing its sensitivity; 2) some low complexity tags were discarded; 3) the tools except
227 LFMD assumed "true" mutations should occupy most reads (proportion >= 0.7) in the read
228 family, although this assumption is unreasonable considering that PCR errors occur in the first
229 PCR cycle; 4) potential false positives of LFMD occurred near indels due to mapping errors.
230
231 LFMD displayed a type of error which led to 6 possible false negatives compared to other tools.
232 This error may occur when a short indel error in a read family splits the big read family into
233 smaller ones, reducing the ability to detect mutations in the read family. UniC detected some
234 mutations that were not detected by other tools. This may be related to the fact that UniC tended
235 to report more false positives in the simulated data due to barcode conflict.

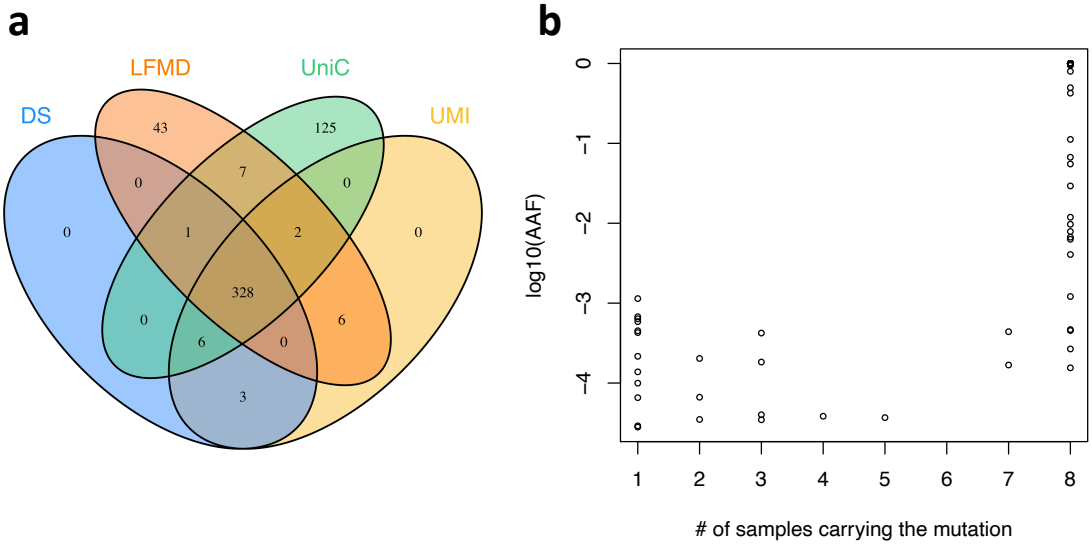


Figure 4 Venn diagram and robustness of LFMD on real data. (a) Venn diagram of mutations detected by four tools in mouse mtDNA. (b) Distribution of mutations found in mtDNA of the Yan Huang cell line. DS, duplex sequencing; UMI, UMI-tools; UniC, Unified Consensus Maker; AAF, alternative allele frequency.

Consistency evaluation using 8 technical replicates of Yan Huang (YH) cell line

We sequenced the YH cell line (passage 19) in 8 independent experiments to evaluate the stability of LFMD (see Materials for details). The results are highly consistent in terms of numbers of mutations detected (range 61~68), shown in Figure 4b and Table 1. Under the hypothesis that true mutations should be identified from at least two samples, we detected 68 "true" mutations, with the mean true positive rate (TPR) and false discovery rate (FDR) around 91.4% and 2.4%, respectively. Due to the high specificity and low sensitivity of LFMD in simulated data, it is very unlikely that any of the mutations are false positives, and the above TPR and FDR only indicate the consistency of parallel experiments.

252 **Table 1 The number of mutations found in mtDNA of 8 YH cell lines.** Under the
253 hypothesis that true mutations should be identified from at least two samples, we detected 68
254 "true" mutations and then calculated TP, FP, TPR, and FDR.

Samples	# of mutations	TP	FP	TPR	FDR
L01_501	64	63	1	92.7%	1.56%
L01_502	68	67	1	98.5%	1.47%
L01_503	62	62	0	91.2%	0.00%
L01_504	65	63	2	92.7%	3.08%
L01_505	62	60	2	88.2%	3.23%
L01_506	61	59	2	86.8%	3.28%
L01_507	65	62	3	91.2%	4.62%
L01_508	62	61	1	89.7%	1.61%
Mean	63.6	62.1	1.50	91.4%	2.4%
SD	2.3	2.4	0.93	3.6%	1.5%

255 TP, true positive; FP, false positive; TPR, true positive rate; FDR, false discovery rate; SD,
256 standard deviation.
257

258 **Mitochondrial DNA Data from 28 Samples with Alzheimer's Disease**

259 The dataset used in this section was provided by Prof. Scott R. Kennedy[4]. They studied point
260 mutations and indels in mtDNA of 28 samples with AD and reported the possibility that the
261 accumulation of low frequency mutations in mtDNA plays a key role in the occurrence and
262 development of AD. They also claimed that mtDNA mutations are inconsistent with oxidative
263 damage. The samples are flash frozen tissues obtained from the hippocampus and parietal lobes
264 of human brain. According to the criteria from DSM-IV and Braak Staging[32], cases were
265 stratified into three groups: ND-Low, ND-High, and D-High, representing early-, mid-, and
266 late-stages, respectively. This dataset was reanalysed with four tools: DS, UMI-tools (UMI),
267 Unified Consensus Maker (UniC), and LFMD. Due to the higher sensitivity, LFMD detected
268 more mutations than other tools, especially for C:G>A:T mutations in the ND-High group
269 (Supplementary Figure S7).

After adjustment for sensitivity, the canonical signature of oxidative damage on DNA[33-35], C:G>A:T, was significantly more common ($p<0.01$) in the ND-High group than in the ND-Low group (Figure 5). This is inconsistent with the main conclusion of Hoekstra et al.'s paper[4] (the origin of the dataset), but consistent with other papers which focus on high frequency mutations in mtDNA[36-38]. In conclusion, a more sensitive tool, coupled with appropriate adjustments, can help uncover the true relationship between LFMs and phenotypes. The low frequency C:G>A:T mutations in mtDNA are markers of oxidative DNA damage and play a role in AD pathogenesis.

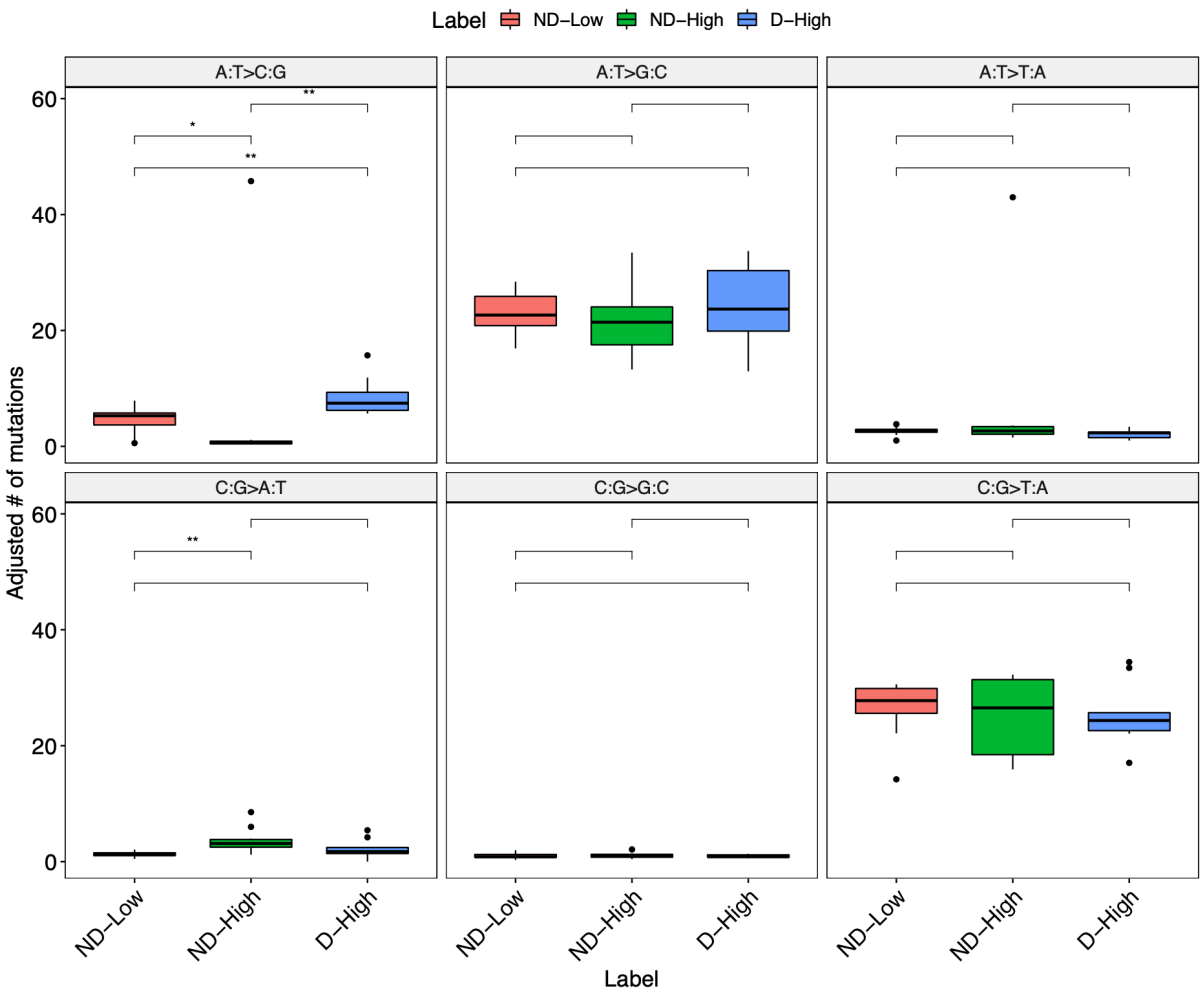


Figure 5 LFMD-revealed mutation spectrum of mtDNA samples with Alzheimer's Disease (AD) after sensitivity adjustment.

282 **Ultra-low frequency mutations on *ABL1* associated with potential drug resistance**

283 Using the duplex sequencing method in 2015, Schmitt et al. analysed an individual with chronic
284 myeloid leukaemia who relapsed after targeted therapy with the drug, Imatinib (the Short Read
285 Archive under accession SRR1799908). We analysed this individual and found 5 extra LFMs.
286 Two of them are in the coding region of the *ABL1* gene and change amino acids: E255G and
287 V256G (E, glutamic acid; G, glycine; V, valine). In the drug resistance database of
288 COSMIC[39], we found that 1) E255VDK, change of the 255th amino acid to V, D, or K (D,
289 aspartic acid; K, lysine), is associated with resistance to the drugs Dasatinib, Imatinib, and
290 Nilotinib; and 2) V256L (L, leucine) is related to resistance to the drug Imatinib. Although the
291 amino acid changes, in this case, are not the same as those in the database, these two additional
292 LFMs still infer potential resistance to the drug Imatinib and provide an additional explanation
293 for the clinical relapse of leukemia. Another useful finding is the prediction of potential
294 resistance to Dasatinib and Nilotinib, a result which may guide clinical practice. The annotation
295 results of 5 LFMs are shown in Table 2.

296

297 **Table 2 Five low-frequency SNVs on *ABL1* detected by LFMD.**

Position (chr9)	Variant	Transcript (NM_)	Function	cDNA Position	CDS Position	AA Position	AA Change
133738364	A>G	005157	exon	767	764	255	E>G
133738364	A>G	007313	exon	1260	821	274	E>G
133738367	T>G	005157	exon	770	767	256	V>G
133738367	T>G	007313	exon	1263	824	275	V>G
133748236	C>T	005157	intronic				
133748236	C>T	007313	intronic				
133748343	T>G	005157	exon	1007	1004	335	V>G
133748343	T>G	007313	exon	1500	1061	354	V>G
133756073	A>C	005157	intronic				
133756073	A>C	007313	intronic				

298 cDNA, complementary DNA; CDS, coding sequence; AA, amino acid. One-letter codes of amino acids:
299 E, glutamic acid; G, glycine; and V, valine.

300

301 **Discussion**

302

303 LFMD is still expensive for target regions >2 Mbp in size because of the need for high
304 sequencing depth. However, as the cost of sequencing continues to fall, it will become more
305 and more practical. In order to sequence a larger region with lower cost, the dilution step of
306 BotSeqS can be introduced into the LFMD pipeline. Because LFMD can deal with tag conflicts,
307 the dilution level can be decreased 1 or 2 magnitudes to increase the sensitivity. Additional
308 experiments will be done soon.

309

310 Only accepting randomly sheared DNA fragments, not working on short amplicon sequencing
311 data, and only working on pair-end sequencing data are known limitations of LFMD. Moreover,
312 LFMD's precision is limited by the accuracy of the alignment software. Although tags are
313 excluded in the model of LFMD described in this paper, LFMD still has the potential to utilize
314 tags and deal with amplicon sequencing data. The basic assumption in our model, that error
315 rates are the same for mutation from the original base to any of the other three bases, is not
316 close enough to reality according to experimental data at present. These remaining issues may
317 be solved in the next version of LFMD.

318

319 To estimate the theoretical limit of LFMD, let read length be 100 bp and the standard deviation
320 (SD) of insert size be 20 bp. Furthermore, let N represent the number of read families across
321 one point. Then, N = (2 reads * 100 bases) * (2 directions * 3 SDs * 20 bp) = 24000 if only
322 considering $\pm$3 SD. As the sheering of DNA is not random in the real world, it is safe to set N

323 as 20,000 according to our experience. Ideally, the likelihood ratio test can detect mutations
324 whose frequency is greater than 0.2% in a read family with Q30 bases (refer to Supplementary
325 Figure S2.). Thus, the theoretical limit of minor allele frequency is around 1e-7 (= 0.002 /
326 20000).
327
328 LFMD reduces the cost partly because it discards tags. First, for a typical 100 bp read, the
329 lengths of the tags and the fixed sequences between the tag and the true sequence are 12 bp and
330 5 bp respectively. So (12+5) / 100 = 17% of data are saved if we do not use tags. Second, the
331 efficiency of targeted capture decreases by about 10 to 20% because of the tags and the fixed
332 sequences, according to in-house experiments. Third, LFMD can work on short read data
333 (PE50) of BGISEQ, and then 30% to 40% of the cost can be saved because of the cheaper
334 sequencing platform. Overall, costs can be reduced by about 70% at best.
335
336 Based on the concept of molecular tags, Du Novo[31], UMI-tools[29] and others have made
337 significant improvements since 2012. In contrast, LFMD uses a completely different approach
338 that discards tags in the first place and provides potential advantages. As DS data and high-
339 depth conventional whole genome sequencing data accumulate in public databases, LFMD
340 provides opportunities for further mining of existing datasets, especially mitochondrial DNA.
341
342 **Conclusion**
343
344 To eliminate endogenous tag conflicts, we use a likelihood-based model to separate the read
345 family of the minor allele from that of the major allele. Without additional experimental steps
346 or customized adapters, LFMD achieves higher sensitivity, higher specificity, and lower cost
347 than other methods. LFMD is useful in areas that require ultra-high precision, such as

348 prediction of drug resistance and early screening of cancer. It is a general method that can be
349 used in several cutting-edge areas, and its mathematical methodology can be generalized to
350 increase the power of other NGS tools.
351
352 The oxidative damage related mutation, C:G>A:T, is significantly enriched in the mid-stage
353 AD group, which is inconsistent with previous low-frequency mutation studies but consistent
354 with the Mitochondrial Free Radical Theory of Aging. The results suggest that AD may be
355 linked to the aging of brain cells induced by oxidative damage. This success also shows that a
356 tool like LFMD, with better sensitivity and consistency across samples, is essential for further
357 study of disease aetiology.
358
359 **Methods and Materials**
360
361 **Subject recruitment and sampling**
362 A lymphoblastoid cell line (Yan Huang, or YH cell line) established from the first Asian
363 genome donor[40] was used. Total DNA was extracted with the MagPure Buffy Coat DNA
364 Midi KF Kit (MAGEN). The DNA concentration was quantified by Qubit (Invitrogen). DNA
365 integrity was examined by agarose gel electrophoresis. The extracted DNA was kept frozen at
366 -80°C until further processing.
367
368 **Mitochondrial genome DNA isolation**
369 Mitochondrial DNA (mtDNA) was isolated and enriched by one or two pairs of primers
370 amplifying the complete mitochondrial genome. The samples were isolated using a single
371 primer set (LR-PCR4) by ultra-high-fidelity Q5 DNA polymerase, following the protocol of
372 the manufacturer (NEB) (Table 3).

373

374 **Table 3 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	

375

376

377 **Library construction and mitochondrial whole genome DNA sequencing**

378 For the BGISeq-500 sequencing platform, mtDNA PCR products were fragmented directly by
379 Covaris E220 (Covaris, Brighton, UK) without purification. Sheared DNA ranging from 150
380 bp to 500 bp without size selection was purified with an Axygen™ AxyPrep™ Mag PCR
381 Clean-Up Kit. 100 ng of sheared mtDNA was used for library construction. End-repairing and
382 A-tailing was carried out in a reaction containing 0.5 U Klenow Fragment (ENZYMATICS™
383 P706-500), 6 U T4 DNA polymerase (ENZYMATICS™ P708-1500), 10 U T4 polynucleotide
384 kinase (ENZYMATICS™ Y904-1500), 1 U rTaq DNA polymerase (TAKARA™ R500Z), 5
385 pmol dNTPs (ENZYMATICS™ N205L), 40 pmol dATPs (ENZYMATICS™ N2010-A-L), 1
386 X PNK buffer (ENZYMATICS™ B904) and water with a total reaction volume of 50 µl. The
387 reaction mixture was placed in a thermocycler running at 37°C for 30 minutes and heat-
388 denatured at 65°C for 15 minutes with the heated lid at 5°C above the running temperature.
389 Adaptors with 10 bp tags (Ad153-2B) were ligated to the DNA fragments by T4 DNA ligase

390 (ENZYMATICS™ L603-HC-1500) at 25°C. The ligation products were PCR amplified.
391 Twenty to twenty-four purified PCR products were pooled together in equal amounts and then
392 denatured at 95°C and ligated by T4 DNA ligase (ENZYMATICS™ L603-HC-1500) at 37°C
393 to generate a single-strand circular DNA library. Pooled libraries were made into DNA
394 Nanoballs (DNB). Each DNB was loaded into one lane for sequencing.
395
396 Sequencing was performed according to the BGISeq-500 protocol (SOP AO) employing the
397 PE100 mode. For reproducibility analyses, YH cell line mtDNA was processed four times
398 following the same protocol as described above to serve as library replicates, and one of the
399 DNBs from the same cell line was sequenced twice as sequencing replicates. A total of 8
400 datasets were generated using the BGISeq-500 platform. MtDNA sequencing was performed
401 on the BGISeq-500 with 100 bp paired-end reads. The libraries were processed for high-
402 throughput sequencing with a mean depth of ~60000x.
403
404
405
406 **Method to Generate a Simulated Dataset**
407 We modified Python scripts developed by the Du Novo[31] team to simulate mixed double-
408 strand sequencing data. Although the simulation may not be entirely realistic, the results are
409 still useful to evaluate the performance and the potential drawbacks of LFMD. Of note, the
410 Python scripts use Dr. Heng Li's well-known program, wgsim, to generate NGS read pairs
411 with controlled sequencing errors, including point errors and indels.
412
413 The original script is "sim2.py", which was developed by the Du Novo team. It can simulate
414 pair-end sequencing with molecular tags. We added a feature that makes the sequencing and

415 PCR error rates slightly higher at both ends of reads than in the middle of reads. The new script
416 is named "sim2_endHighErr.py". Basically, the error rate at the 5 bp ends of each read is 5
417 times larger than that in the middle, but not larger than 0.1. This feature makes the simulation
418 data more realistic, although in the real data each base has a different error rate, and bases at
419 read ends do not always have higher error rates.
420
421 Apart from calculating the error rate at read ends, we also wrote a script, getFrag_randShear.pl,
422 to randomly shear genome DNA. Because the basic assumption of LFMD is that read pairs
423 from random sheared DNA fragments can be grouped into read families which conduce to
424 reduction of sequencing and PCR error. We set the mean length of sheared fragments as 600
425 bp and the standard deviation as 20 bp. The lengths of sheared fragments are set to be normally
426 distributed.
427
428 With the randomly sheared fragments, we used sim2_endHighErr.py to generate NGS read
429 pairs with original and mutation-inserted mitochondrial reference genomes (NC_012920.1).
430 The base quality in the middle of each read is set as Q20, which means a sequencing error rate
431 of 1%. The PCR error rate is set as 10^{-5}, the read length as 100 bp, the length of a molecular
432 tag 12 bp, and the length of the fixed sequence between tag and sequence 5 bp. To be consistent
433 with the data of the Duplex Sequencing system, the fixed sequence is also "TGACT". In this
434 case, the simulated data can be used as input data for not only LFMD, but also for Duplex
435 Sequencing, UMI-tools, Du Novo, Unified Consensus Maker, etc. Thus, the comparison
436 between different methods can be done on the same dataset. Because LFMD does not need tag
437 information to group reads, we added an extra step to remove tags in the LFMD pipeline to
438 analyze simulated data and Duplex Sequencing data. So, the true read length of data used in

439 LFMD is 83 bp (= 100 - 12 - 5). The results showed that LFMD used 83% of data but
440 outperformed the rival methods which used tags.
441
442 The mutation-inserted genome of the mitochondrial chromosome is generated by an in-house
443 script named "fakeMut_rand.pl". The script accepts parameters like reference genome,
444 numbers of SNVs, number of insertions (INSs), number of deletions (DELs), and maximum
445 lengths of INSs and DELs. Seeds that were used for generating random numbers were included
446 in the output, which is useful because the script can generate exactly the same mutation-
447 inserted genome when the same seed is used. Here is the seed list of the 9 batches: 105901112,
448 1294553942, 1803264703, 2924365934, 2587769895, 859786530, 2672906100, 3210264384,
449 4040679464. As the simulation mainly focused on short indels, we set the maximum lengths
450 of indels to 3 bp. We simulated 9 batches of data. Batch 1 has 50 SNVs, 30 INSs, and 30 DELs
451 on a single haplotype of mitochondrial genome. The mutations are distributed randomly, but
452 the minimum distance between two mutations is 7 bp. Batch 2 has 50 SNVs, 20 INSs, and 20
453 DELs. Batches 3-9 have 100 SNVs, 10 INSs, and 10 DELs. Thus, in 9 batches of data, we
454 simulated a total of 800 SNVs, 120 INSs, and 120 DELs.
455
456 Due to the fact that indels may fall next to similar sequences in the genome and result in shifted
457 genome positions in detection results, an additional script, named "blast2mutList.pl", was used
458 to adjust their genome positions and update their mutation lists. Although most of the shifts
459 can be detected and adjusted, there are several mutations that cannot be dealt with correctly.
460 For example, in the third batch, we found position shifts at 2388 and 11399. These turn out to
461 be issues related to "blastn 2.6.0". Specifically, the DEL at position 11399, which deletes "TA",
462 is described falsely by blast as a DEL which deletes "AT" at position 11400. The correct

463 location should be 11399 according to the general standard. Thus, a false positive in the LFMD
464 result may be due to an error in the very beginning induced by the extra tool, “blastn”.
465
466 The mutation-inserted mitochondrial genomes were used as templates to generate 15 different
467 numbers of DNA fragments: 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 2000, 3000,
468 4000, 5000, and 10000. Then, based on these fragments, NGS read pairs were generated by
469 simu.randomShear_frag.sh. Similarly, the mitochondrial reference genome without any
470 mutations was also used as template to generate 15 different numbers of DNA fragments: 1e6
471 - 100, 1e6 - 200, 1e6 - 300, 1e6 - 400, 1e6 - 500, 1e6 - 600, 1e6 - 700, 1e6 - 800, 1e6 - 900,
472 1e6 - 1000, 1e6 - 2000, 1e6 - 3000, 1e6 - 4000, 1e6 - 5000, and 1e6 - 10000. Thus, after
473 generating NGS read pairs based on these DNA fragments and merging them with read pairs
474 of the corresponding fragments of the mutation-inserted genome, we simulated 15 datasets for
475 each batch. The designed allele frequencies of the 15 datasets are 1e-4, 2e-4, 3e-4, 4e-4, 5e-4,
476 6e-4, 7e-4, 8e-4, 9e-4, 1e-3, 2e-3, 3e-3, 4e-3, 5e-3, and 1e-2. The overall read depth is around
477 130,000x for all datasets. Thus, we got simulated sequencing data of 135 (9 batches of data x
478 15 fragment sets) distinguishable samples.
479
480 To evaluate the performance of methods under different sequencing depths, we randomly
481 selected read pairs from these datasets. The fractions of selected read pairs were 0.1, 0.2, 0.25,
482 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0. In this way, sequencing data with a total of 11 levels of
483 read depth were generated. The fraction 0.25 corresponded to a read depth of 32,500x, while
484 the fraction 0.5 represented 65,000x.
485
486 **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 families, 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

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

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[41] 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[42] presented by Chen et al in 2017 (Details in Supplementary Materials).

533 Interestingly, we finally arrive at a general conclusion that the further adjustment of χ_1^2 is not
534 helpful in similar cases although the asymptotic distribution we use is not perfect when N is
535 small (e.g., N<5). Alternative approaches might be derived in the future. We also compared
536 theoretical P-values with empirical P-values from Monte Carlo procedures (Supplementary
537 Figure S1), explored the power of our model under truly and uniformly distributed sequencing
538 errors (Supplementary Figure S2), and evaluated the accuracy of allele frequency
539 (Supplementary Figure S3 and S4). The simulation results support the theoretical conclusion
540 sufficiently when P-value is equal or less than 10^-3.
541
542 Because the null and alternative hypotheses have two and three free variables respectively, the
543 Chi-square distribution has 1 degree of freedom. The P-value of the allele g can then be given
544 $$P_g = 1 - \mathrm{cdf}(t_g)$$
545
546 where $\mathrm{cdf}(x)$ is the cumulative density function of the χ_1^2 distribution. If P_g is less than a given
547 threshold α, the null hypothesis is rejected and the allele g is treated as a candidate allele of
548 the read family.
549
550 Although P_g cannot be interpreted as the probability that $H_{0,g}$ is true and allele g is an error, it
551 is a proper approximation of the error rate of allele g when $P_g \leq 10^{-3}$. We only reserve alleles
552 with $P_g \leq \alpha$ in both Watson and Crick families and substitute others with "N". Then Watson
553 and Crick families are compressed into several single-strand consensus sequences (SSCSs).
554 The SSCSs might contain haplotype information if more than one heterozygous site were
555 detected. Finally, SSCSs which are consistent in both Watson and Crick families are claimed
556 as double-strand consensus sequences (DCSs).
557

For each allele on a DCS, let $P_w(g)$ and $P_c(g)$ represent the relative error rates of the given allele g in the Watson and Crick family respectively, and let $P_{wc}(g)$ denote the error rate of the allele g on the DCS. 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[43]. 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 frequency 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

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

582 Then $Q(g)$ is transferred to an ASCII character, and a series of characters make a base quality
583 sequence for the DCS. Finally, we generate a BAM file with DCSs and their quality sequences.
584
585 With the BAM file which contains all the high-quality DCS reads, the same approach is used
586 to give each allele a P-value at each genomic position which is covered by DCS reads. Adjusted
587 P-values (q-values) are given via the Benjamin-Hochberg procedure. The threshold of q-values
588 is selected according to the total number of tests conducted and false discovery rate (FDR)
589 which can be accepted.
590
591 A similar mathematical model was described in detail in previous papers by Ding et al[6] and
592 Guo et al[20]. Ding et al. used this model to reliably call mutations with frequency > 4%. In
593 contrast, we use this model to deal with read families rather than non-duplicate reads. In a
594 mixed read family, most of the minor allele frequencies are larger than 4%, so the power of the
595 model meets our expectation.
596
597 For those reads containing indels, the CIGAR strings in BAM files contain I or D. It is obvious
598 that reads with different CIGAR strings cannot fit into one read family. Thus, CIGAR strings
599 can also be used as part of endogenous tags. In contrast, the soft-clipped part of CIGAR strings
600 cannot be ignored when considering start and end positions because low-quality parts of reads
601 tend to be clipped, and the coordinates after clipping are not a proper endogenous tag for the
602 original DNA template.
603
604 **Method to estimate sensitivity**
605

606 One of the major concerns of low frequency mutation detection is the relatively-low sensitivity
607 and high-depth requirement due to the fact that the sensitivity of LFMD might be affected by
608 total sequencing depth, depth distribution, randomness of DNA shearing, distribution of length
609 of sheared DNA fragment, DCS depth, PCR bias, number of PCR cycles during library
610 construction, initial amount of DNA, GC content, etc. Thus, comparison among multiple
611 samples with various library construction and sequencing status is challenging (Supplementary
612 Figure S5). To overcome this obstacle, it is essential to carry out a robust sensitivity estimation
613 for all samples before comparing their mutation calling results. With the estimated sensitivity,
614 the number of called mutations can be normalized into a pseudo-number of mutations, then the
615 comparison between the samples will be fair (Supplementary Figure S6). In addition, accurate
616 estimation of the number of fragments in a read family can help to estimate the PCR error rate
617 accurately and thus improve the specificity of LFMD.

618 As shown in Figure 6, insert size is defined as the length of DNA fragment after random
619 shearing in the library construction step; bin is defined as a possible genome position that can
620 be occupied by a DNA fragment with given insert size. As the read length is given and fixed,
621 the total number of bins are determined and fixed. The positions that are not occupied are empty
622 bins.

623

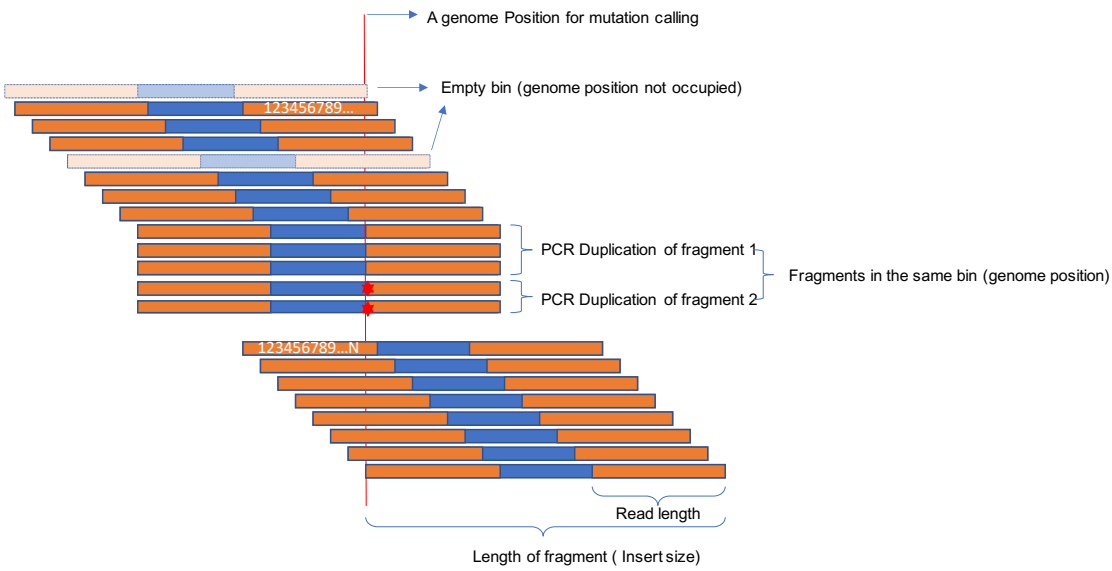


Figure 6 The definition of insert size and bin in the LFMD context.

For each DNA fragment, let Y denote the number of sequenced read pairs of the fragment. Assume Y is a Poisson distributed random variable with mean λ_2 and variance λ_2, denoted as $Y \sim Pois(\lambda_2)$. So, its probability mass function can be written as

$$P(Y = y) = \frac{{\lambda_2}^y e^{-\lambda_2}}{y!}$$

For k fragments in a bin,

$$E\left(\sum_{j=1}^{k} Y_j\right) = k\lambda_2$$

$$Var\left(\sum_{j=1}^{k} Y_j\right) = \sum_{j=1}^{k} Var(Y_j) = k \cdot Var(Y_j) = k\lambda_2$$

Let $Y' = \sum_{j=1}^{k} Y_i$ and $\lambda' = k\lambda_2$, then $Y' \sim Pois(\lambda')$, where Y' is the number of sequenced read pairs in a bin which has k fragments. Let X denote the number of fragments in a bin. Assume X follows a Poisson distribution, $X \sim Pois(\lambda)$, with mean λ and variance λ.

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

So, the probability that y read pairs found in a bin is

$$P(Y' = y) = \sum_{k=0}^{\infty} P(Y' = y | X = k) P(X = k) = \sum_{k=0}^{\infty} \left(\frac{(k\lambda_2)^y e^{-k\lambda_2}}{y!} \cdot \frac{\lambda^k e^{-\lambda}}{k!} \right)$$

Let N_b be the number of bins, and Z be the total number of all read pairs in all bins, then

$$\lambda_2 = \frac{Z}{N_b \lambda} = \frac{\sum_{i=1}^{N_b} y_i}{N_b \lambda} > 0, \phi > 0, \lambda > 0$$

Let N_{empty} denote the number of empty bins, then $P(Y' = 0) \approx N_{empty}/N_b$. Therefore, λ can be calculated via solving the following equation:

$$P(Y' = 0) = \sum_{k=0}^{\infty} \left(\frac{(k\lambda_2)^0 e^{-k\lambda_2}}{0!} \cdot \frac{\lambda^k e^{-\lambda}}{k!} \right)$$

$$= \sum_{k=0}^{\infty} \left(e^{-k\lambda_2} \cdot \frac{\lambda^k e^{-\lambda}}{k!} \right) = \sum_{k=0}^{\infty} \left(e^{-k\lambda_2 - \lambda} \cdot \frac{\lambda^k}{k!} \right) = \sum_{k=0}^{\infty} \left(e^{-\frac{kZ}{N_b \lambda} - \lambda} \cdot \frac{\lambda^k}{k!} \right) \approx \frac{N_{empty}}{N_b}$$

However, the equation, $\lambda_2 = \frac{Z}{N_b \lambda}$ and $P(Y' = 0) \approx N_{empty}/N_b$, may not hold when the amount of data is small. So, the simple estimation of λ is efficient but biased.

To estimate λ and λ_2 generally, a better method is to select them to maximize the log-likelihood function

$$\ell(\lambda, \lambda_2) = \sum_{i=1}^{N_b} log\big(P(Y' = y_i)\big) = \sum_{i=1}^{N_b} log\left[\sum_{k=0}^{\infty}\left(\frac{(k\lambda_2)^{y_i} e^{-k\lambda_2}}{y_i!} \cdot \frac{\lambda^k e^{-\lambda}}{k!}\right)\right]$$

$$= \sum_{i=1, y_i>0}^{N_b} log\left[\sum_{k=1}^{\infty}\left(\frac{(k\lambda_2)^{y_i} e^{-k\lambda_2}}{y_i!} \cdot \frac{\lambda^k e^{-\lambda}}{k!}\right)\right] + \sum_{i=1, y_i=0}^{N_b} log\left[e^{-\lambda} + \sum_{k=1}^{\infty}\left(e^{-k\lambda_2-\lambda} \cdot \frac{\lambda^k}{k!}\right)\right]$$

Once the Maximum Likelihood Estimators (MLEs) are obtained for the log-likelihood function, it is possible to calculate the probability of k fragments in a bin given observed y read pairs according to Bayes' theorem:

$$P(X = k|Y' = y) = \frac{P(Y' = y|X = k)P(X = k)}{P(Y' = y)}$$

We assume only one fragment in the bin has the mutation, considering the low frequency of the mutation. Thus, in the bin, the mutation frequency is $1/k$. Let s denote the number of reads supporting the mutation. Given y reads and the threshold that at least 3 reads support the mutation, the probability that the mutation cannot be detected is

$$P(s < 3;\ y) = \sum_{s=0}^{2} \binom{y}{s} (1/k)^s (1 - 1/k)^{y-s}$$

If considering the sequencing error as e, it is possible to add the probability that 3 PCR products of the fragment support the mutation but at least one of them has a sequencing error, wherein the mutation cannot be detected:

$$P(s = 3;\ y) * (1 - (1 - e)^3) = \binom{y}{3} (1/k)^3 (1 - 1/k)^{y-3} (1 - (1 - e)^3)$$

Let y_w and y_c denote the number of read pairs in Watson and Crick families respectively. We get the probability that the mutation cannot be detected as

$$P(s < 3; y_w) + P(s < 3; y_c) - P(s < 3;\ y_w) * P(s < 3; y_c)$$

So, we can estimate the sensitivity at the position as

$$1 - \prod_{i=1}^{N_b} P(not\ detected\ in\ the\ i^{th}\ bin)$$

To fit the empirical sensitivities better, more sophisticated models, including overdispersion of the Poisson Model and stand bias, were considered in the supplementary materials. In fact, the simple two-Poisson model is good enough to fit the sensitivity curve, especially when sequencing depth is low (<60,000x).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors have approved the manuscript for submission.

Availability of data and materials

The datasets supporting the conclusions of this article are available in the CNGBdb repository (https://db.cngb.org/cnsa/) with accession code CNP0000297. Analysis codes used in this paper can be accessed at https://github.com/RainyEricYe/LFMD.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Shenzhen Municipal Government of China [JCYJ20170412153100794]. The funders had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

P.C.S. and C.N. supervised the entire work. R.Y. wrote the manuscript. R.Y. and X.H.Z. wrote the code and evaluated the performance of LFMD. Y.W.Q. assisted the analysis of mitochondrial data. R.J. led all wet-lab experiments and tested the LR-PCR on mitochondria. R.J., Y.T.A., and J.M.X. prepared the samples and did the experiments. T.M. and P.C.S. helped the mathematical part. P.C.S, C.N., L.B., X.X., H.M.Y, X.Q.Z, X.L., and T.M. contributed to critical revision of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Prof. Scott R. Kennedy for kindly sharing their valuable data. We thank Jeremiah Wala and Rameen Beroukhim for developing a wonderful C++ API and SeqLib[44], and we thank S. Bochkanov, V. Bystritsky, and other contributors for the development and maintenance of ALGLIB[45], which accelerated the coding progress of LFMD significantly. We also thank Heng Li, Richard Durbin and other developers for creating bwa[46] and samtools[47]. We are so grateful to McKenna et al. and Broad Institute for the introduction and maintenance of the GATK[48] framework. And we do appreciate the open source scripts of Nicholas Stoler et al. to simulate tagged sequences[31]. Without the open source spirit and knowledge sharing, our work could not be done.

710 **References**

711 1. Newman AM, Lovejoy AF, Klass DM, Kurtz DM, Chabon JJ, Scherer F, Stehr H, Liu CL,
712 Bratman SV, Say C: **Integrated digital error suppression for improved detection of**
713 **circulating tumor DNA.** *Nature biotechnology* 2016, **34:**547.

714 2. Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, Anagnostou V, Fiksel J,
715 Cristiano S, Papp E: **Direct detection of early-stage cancers using circulating tumor**
716 **DNA.** *Science translational medicine* 2017, **9:**eaan2415.

717 3. Newman AM, Bratman SV, To J, Wynne JF, Eclov NC, Modlin LA, Liu CL, Neal JW,
718 Wakelee HA, Merritt RE: **An ultrasensitive method for quantitating circulating**
719 **tumor DNA with broad patient coverage.** *Nature medicine* 2014, **20:**548.

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

723 5. Simon DK, Lin MT, Zheng L, Liu G-J, Ahn CH, Kim LM, Mauck WM, Twu F, Beal MF,
724 Johns DR: **Somatic mitochondrial DNA mutations in cortex and substantia nigra in**
725 **aging and Parkinson's disease.** *Neurobiology of aging* 2004, **25:**71-81.

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

730 7. Wallace DC, Chalkia D: **Mitochondrial DNA genetics and the heteroplasmy**
731 **conundrum in evolution and disease.** *Cold Spring Harbor perspectives in biology*
732 2013, **5:**a021220.

733 8. Schmitt MW, Fox EJ, Prindle MJ, Reid-Bayliss KS, True LD, Radich JP, Loeb LA:
734 **Sequencing small genomic targets with high efficiency and extreme accuracy.**
735 *Nature methods* 2015, **12:**423.
736 9. Dimond R: **Social and ethical issues in mitochondrial donation.** *British medical*
737 *bulletin* 2015, **115:**173.
738 10. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R: **Accurate sampling and**
739 **deep sequencing of the HIV-1 protease gene using a Primer ID.** *Proceedings of the*
740 *National Academy of Sciences* 2011, **108:**20166-20171.
741 11. Marquis J, Lefebvre G, Kourmpetis YA, Kassam M, Ronga F, De Marchi U, Wiederkehr
742 A, Descombes P: **MitoRS, a method for high throughput, sensitive, and accurate**
743 **detection of mitochondrial DNA heteroplasmy.** *BMC genomics* 2017, **18:**326.
744 12. Kang E, Wang X, Tippner-Hedges R, Ma H, Folmes CD, Gutierrez NM, Lee Y, Van
745 Dyken C, Ahmed R, Li Y: **Age-related accumulation of somatic mitochondrial DNA**
746 **mutations in adult-derived human iPSCs.** *Cell Stem Cell* 2016, **18:**625-636.
747 13. Blandini F, Greenamyre JT, Nappi G: **The role of glutamate in the pathophysiology**
748 **of Parkinson's disease.** *Functional neurology* 1996, **11:**3-15.
749 14. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, Cook K, Stepansky A,
750 Levy D, Esposito D: **Tumour evolution inferred by single-cell sequencing.** *Nature*
751 2011, **472:**90.
752 15. Baslan T, Hicks J: **Single cell sequencing approaches for complex biological systems.**
753 *Current opinion in genetics & development* 2014, **26:**59-65.
754 16. Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, Sawyer SL: **High-**
755 **throughput DNA sequencing errors are reduced by orders of magnitude using circle**

756 **sequencing.** *Proceedings of the National Academy of Sciences* 2013, **110:**19872-
757 19877.
758 17. Wang K, Lai S, Yang X, Zhu T, Lu X, Wu C-I, Ruan J: **Ultrasensitive and high-efficiency**
759 **screen of de novo low-frequency mutations by o2n-seq.** *Nature communications*
760 2017, **8:**15335.
761 18. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B: **Detection and**
762 **quantification of rare mutations with massively parallel sequencing.** *Proceedings of*
763 *the National Academy of Sciences* 2011, **108:**9530-9535.
764 19. Vogelstein B, Kinzler KW, Papadopoulos N, Kinde I: **Safe sequencing system.** Google
765 Patents; 2016.
766 20. Guo Y, Li J, Li C-I, Shyr Y, Samuels DC: **MitoSeek: extracting mitochondria**
767 **information and performing high-throughput mitochondria sequencing analysis.**
768 *Bioinformatics* 2013, **29:**1210-1211.
769 21. Kim J, Kim D, Lim JS, Maeng JH, Son H, Kang H-C, Nam H, Lee JH, Kim S: **The use of**
770 **technical replication for detection of low-level somatic mutations in next-**
771 **generation sequencing.** *Nature communications* 2019, **10:**1047.
772 22. Wang TT, Abelson S, Zou J, Li T, Zhao Z, Dick JE, Shlush LI, Pugh TJ, Bratman SV: **High**
773 **efficiency error suppression for accurate detection of low-frequency variants.**
774 *Nucleic acids research* 2019.
775 23. Hoang ML, Kinde I, Tomasetti C, McMahon KW, Rosenquist TA, Grollman AP, Kinzler
776 KW, Vogelstein B, Papadopoulos N: **Genome-wide quantification of rare somatic**
777 **mutations in normal human tissues using massively parallel sequencing.**
778 *Proceedings of the National Academy of Sciences* 2016, **113:**9846-9851.

779 24. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA: **Detection of ultra-rare**
780 **mutations by next-generation sequencing.** *Proceedings of the National Academy of*
781 *Sciences* 2012, **109:**14508-14513.
782 25. Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, Prindle MJ, Kuong KJ,
783 Shen JC, Risques RA, Loeb LA: **Detecting ultralow-frequency mutations by Duplex**
784 **Sequencing.** *Nat Protoc* 2014, **9:**2586-2606.
785 26. Wang K, Lai S, Yang X, Zhu T, Lu X, Wu C-I, Ruan J: **Ultrasensitive and high-efficiency**
786 **screen of de novo low-frequency mutations by o2n-seq.** *Nature communications*
787 2017, **8:**1-11.
788 27. Peng Q, Satya RV, Lewis M, Randad P, Wang Y: **Reducing amplification artifacts in**
789 **high multiplex amplicon sequencing by using molecular barcodes.** *BMC genomics*
790 2015, **16:**589.
791 28. Kou R, Lam H, Duan H, Ye L, Jongkam N, Chen W, Zhang S, Li S: **Benefits and**
792 **challenges with applying unique molecular identifiers in next generation**
793 **sequencing to detect low frequency mutations.** *PloS one* 2016, **11:**e0146638.
794 29. Smith T, Heger A, Sudbery I: **UMI-tools: modeling sequencing errors in Unique**
795 **Molecular Identifiers to improve quantification accuracy.** *Genome Res* 2017,
796 **27:**491-499.
797 30. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, Khor CC, Petric R, Hibberd
798 ML, Nagarajan N: **LoFreq: a sequence-quality aware, ultra-sensitive variant caller**
799 **for uncovering cell-population heterogeneity from high-throughput sequencing**
800 **datasets.** *Nucleic Acids Research* 2012, **40:**11189-11201.
801 31. Stoler N, Arbeithuber B, Guiblet W, Makova KD, Nekrutenko A: **Streamlined analysis**
802 **of duplex sequencing data with Du Novo.** *Genome biology* 2016, **17:**180.

803 32. Braak H, Braak E: **Neuropathological stageing of Alzheimer-related changes.** *Acta*
804 *neuropathologica* 1991, **82:**239-259.
805 33. Cheng KC, Cahill DS, Kasai H, Nishimura S, Loeb LA: **8-Hydroxyguanine, an abundant**
806 **form of oxidative DNA damage, causes G----T and A----C substitutions.** *Journal of*
807 *Biological Chemistry* 1992, **267:**166-172.
808 34. Cooke MS, Evans MD, Dizdaroglu M, Lunec J: **Oxidative DNA damage: mechanisms,**
809 **mutation, and disease.** *The FASEB Journal* 2003, **17:**1195-1214.
810 35. Marnett LJ: **Oxyradicals and DNA damage.** *carcinogenesis* 2000, **21:**361-370.
811 36. Lovell MA, Soman S, Bradley MA: **Oxidatively modified nucleic acids in preclinical**
812 **Alzheimer's disease (PCAD) brain.** *Mechanisms of ageing and development* 2011,
813 **132:**443-448.
814 37. Wang J, Xiong S, Xie C, Markesbery W, Lovell M: **Increased oxidative damage in**
815 **nuclear and mitochondrial DNA in Alzheimer's disease.** *Journal of neurochemistry*
816 2005, **93:**953-962.
817 38. Wang J, Markesbery WR, Lovell MA: **Increased oxidative damage in nuclear and**
818 **mitochondrial DNA in mild cognitive impairment.** *Journal of neurochemistry* 2006,
819 **96:**825-832.
820 39. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG,
821 Creatore C, Dawson E: **COSMIC: the catalogue of somatic mutations in cancer.**
822 *Nucleic acids research* 2018, **47:**D941-D947.
823 40. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, et al:
824 **The diploid genome sequence of an Asian individual.** *Nature* 2008, **456:**60-65.
825 41. Drton M: **Likelihood ratio tests and singularities.** *The Annals of Statistics* 2009,
826 **37:**979-1012.

827 42. Chen Y, Huang J, Ning Y, Liang K-Y, Lindsay BG: **A conditional composite likelihood**
828 **ratio test with boundary constraints.** *Biometrika* 2017, **105:**225-232.
829 43. Weiss G, Von Haeseler A: **A coalescent approach to the polymerase chain reaction.**
830 *Nucleic acids research* 1997, **25:**3082-3087.
831 44. Wala J, Beroukhim R: **SeqLib: a C++ API for rapid BAM manipulation, sequence**
832 **alignment and sequence assembly.** *Bioinformatics* 2016, **33:**751-753.
833 45. Bochkanov S, Bystritsky V: **Alglib.** *Available from: www alglib net* 2013, **59**.
834 46. Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler**
835 **transform.** *Bioinformatics* 2009, **25:**1754-1760.
836 47. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
837 Durbin R: **The sequence alignment/map format and SAMtools.** *Bioinformatics* 2009,
838 **25:**2078-2079.
839 48. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K,
840 Altshuler D, Gabriel S, Daly M: **The Genome Analysis Toolkit: a MapReduce**
841 **framework for analyzing next-generation DNA sequencing data.** *Genome research*
842 2010, **20:**1297-1303.
843