V4.1 on 2019.10.31

Supplemental materials

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

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

Then the composite conditional log-likelihood can be constructed as

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

in which we set

$$\omega_{ik} = 1$$

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

$$U_c(\theta) = \frac{\partial \ell_c(\theta)}{\partial \theta}$$

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

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

The corresponding estimators of *H* and *V* are denoted by $\widehat{H}$ and $\widehat{V}$ evaluated at $\hat{\theta}_c$. The modified composite likelihood under boundary constraints was given by Chen et al(1) as

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

where

$$T(\theta) = N^{-1/2}\widehat{H}^{-1}U_c\left(\hat{\theta}_c\right) - N^{1/2}\left(\theta - \hat{\theta}_c\right)$$

$$\widehat{H}_A = \widehat{H}\widehat{V}^{-1}\widehat{H}$$

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

Thus, we derive the adjusted likelihood ratio test

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

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

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

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

Then, the expected number of bases g with e is

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

Thus,

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

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

where C is a finite constant. Then we derive

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

As a result, $\widehat{V}^{-1}$ tends to $\boldsymbol{0}$ in the model, which means the adjustment is not necessary. To be clear, the special form of matrix V with all equal elements is due to the infinite N which ensures all possible e and g occur in the function $\ell_c(\theta)$. The V does not have a special form when N is a finite number. The simulation results are concordant with the theoretical results (Figure S1). The power of the model is evaluated in Figure S2.

Figure S1. The comparison between theoretical and empirical P-values from Monte Carlo procedures under truly distributed sequencing error rates. With the null hypothesis, one million simulations were conducted.

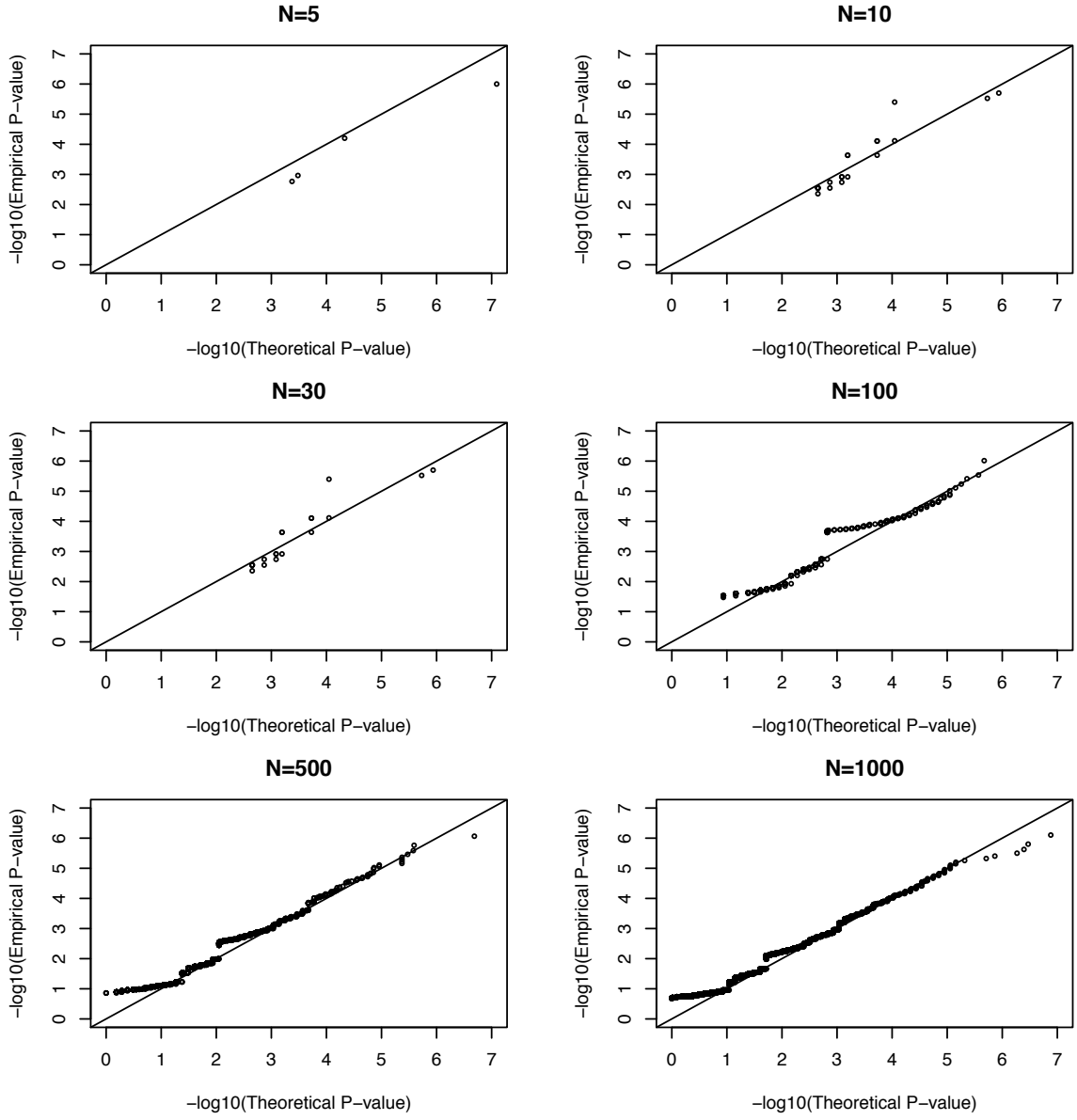

Figure S2. Receiver Operating Characteristic (ROC) curves of the likelihood-based model under uniformly distributed sequencing errors (Q20). We simulated 10,000 SNVs with a frequency equal to 1% or 0.1% and 1000 SNVs for 0.01%.

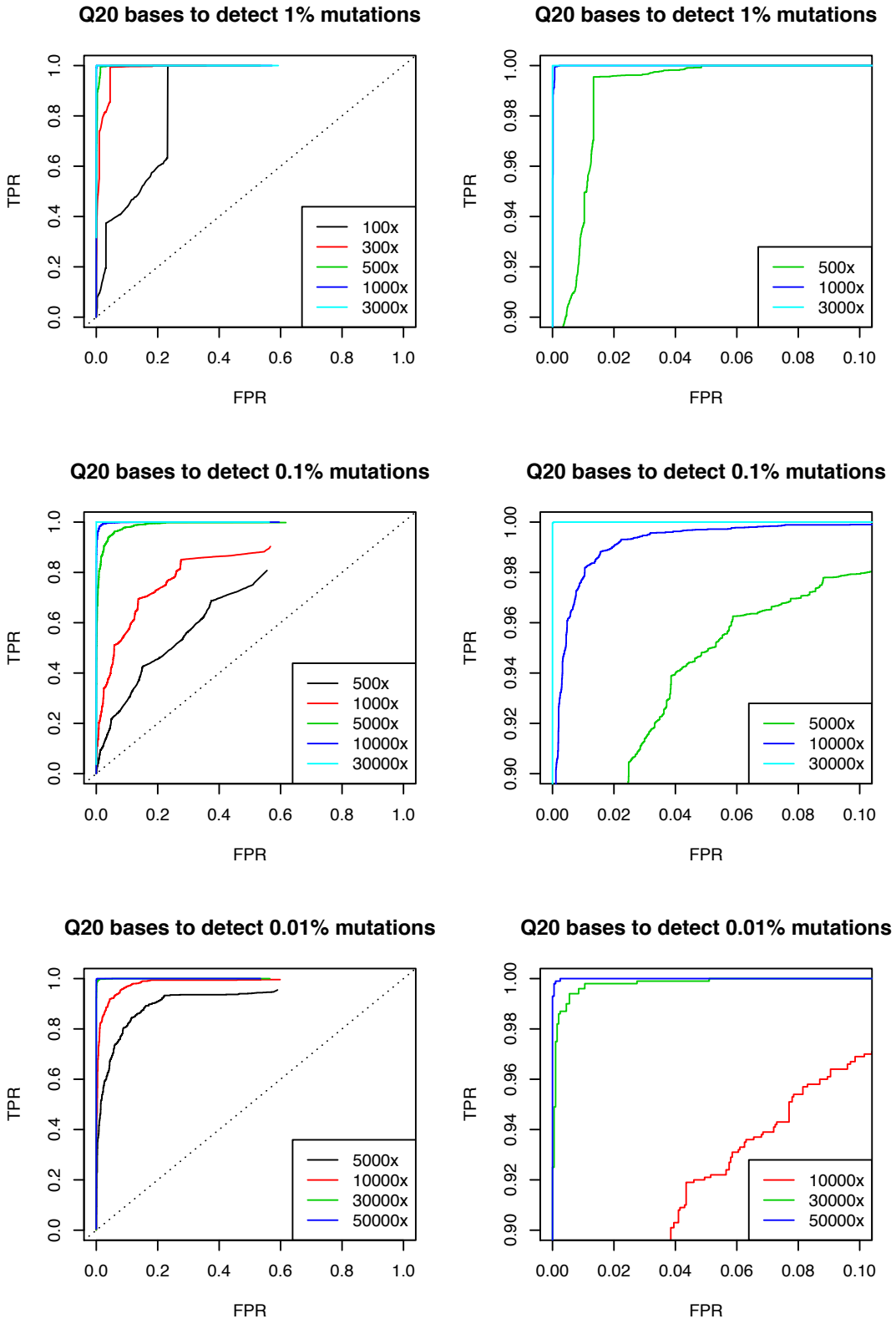

Figure S3. Estimation of allele frequency with the likelihood-based model under uniformly distributed sequencing errors.

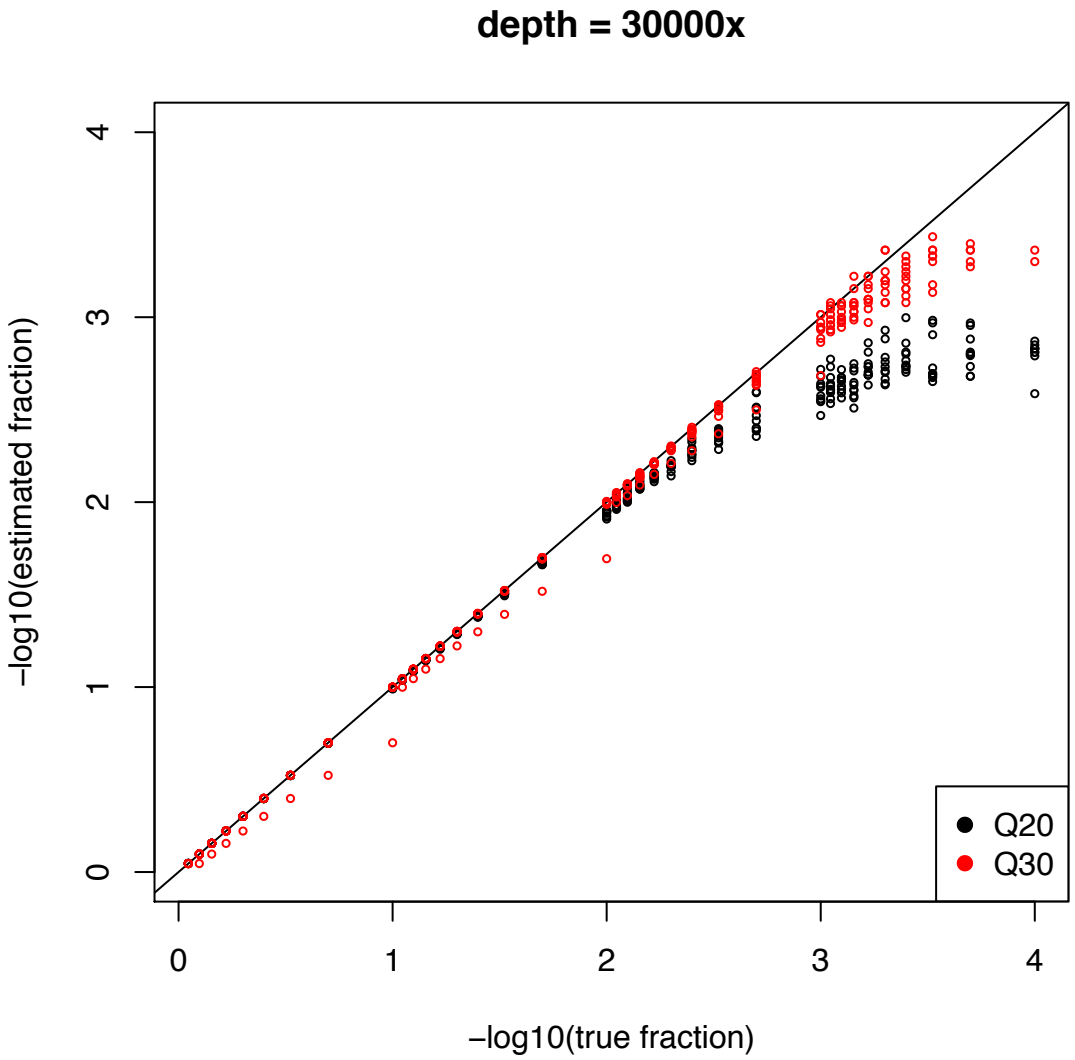

Figure S4. Pipelines of DS and LFMD.

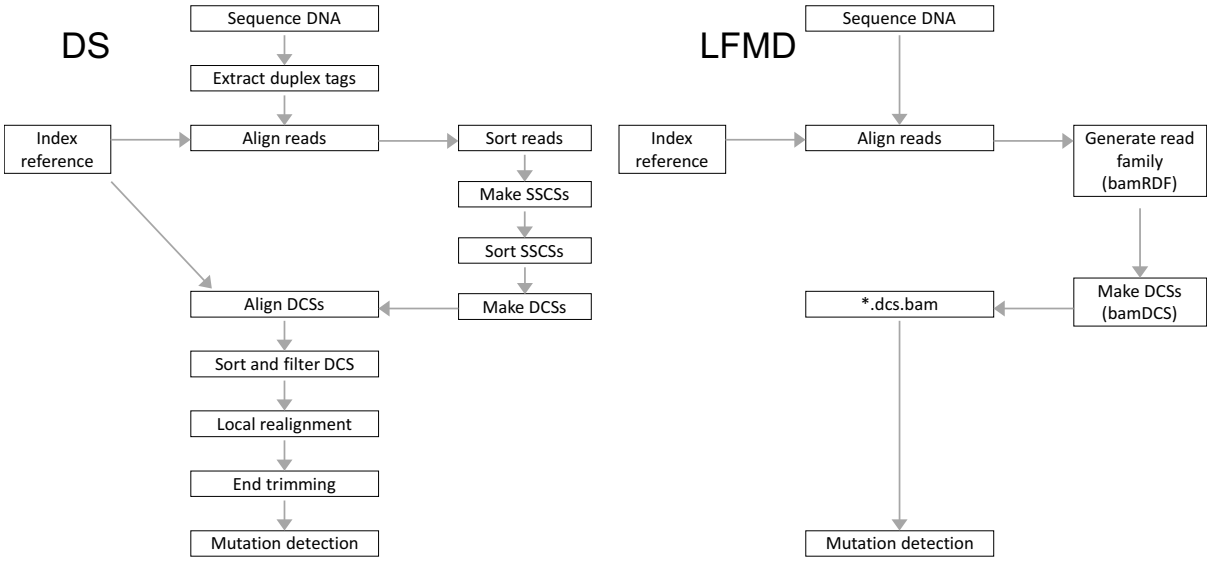

Table S1. The number of true positives detected by DS and LFMD. There are 67 single nucleotide variants (SNVs), 13 insertions (INSs), and 3 deletions (DELs) in the simulated data at every level of alternative allele frequency (AAF).

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

Table S2. The number of false positives detected by DS and LFMD.

AAF	SNV		INS		DEL	
	DS	LFMD	DS	LFMD	DS	LFMD
1.0E-04	0	0	0	0	0	0
2.0E-04	0	0	0	0	0	0
3.0E-04	0	0	0	0	0	0
4.0E-04	0	0	0	0	0	0
5.0E-04	1	0	0	0	0	0
6.0E-04	0	0	0	0	0	0
7.0E-04	0	0	0	0	0	0
8.0E-04	0	0	0	0	0	0
9.0E-04	0	0	0	0	0	0
1.0E-03	0	1	0	0	0	0
2.0E-03	0	0	0	0	0	0
3.0E-03	2	0	0	0	0	0
4.0E-03	0	0	0	0	0	0
5.0E-03	0	0	0	0	0	0
1.0E-02	0	0	0	0	0	0

Table S3. DS vs LFMD on 26 human mtDNA samples from Prof. Kennedy's laboratory.

Sample	DS_only	Overlap	LFMD_only	DS_only /Overlap	LFMD_only /Overlap
1440B	27	928	110	2.91%	11.85%
1440E	10	491	66	2.04%	13.44%
2384H	13	500	171	2.60%	34.20%
2384P	4	200	60	2.00%	30.00%
3080H	5	231	68	2.16%	29.44%
3080P	23	504	104	4.56%	20.63%
334B	14	592	100	2.36%	16.89%
334E	13	1332	142	0.98%	10.66%
409B	20	649	76	3.08%	11.71%
409E	10	994	134	1.01%	13.48%
511H	15	669	104	2.24%	15.55%
523B	2	494	57	0.40%	11.54%
523E	6	675	73	0.89%	10.81%
533B	1	216	52	0.46%	24.07%
533E	1	111	35	0.90%	31.53%
547H	4	411	104	0.97%	25.30%
547P	10	799	94	1.25%	11.76%
552B	14	467	87	3.00%	18.63%
552E	12	576	76	2.08%	13.19%
558P	7	82	40	8.54%	48.78%
626H	6	189	101	3.17%	53.44%
626P	5	165	76	3.03%	46.06%
652B	10	684	78	1.46%	11.40%
652E	3	595	54	0.50%	9.08%
670B	8	753	73	1.06%	9.69%
670E	1	116	41	0.86%	35.34%
Median	/	/	/	2.02%	16.22%

Table S4. The number of mutations found in mtDNA of 8 YH cell lines. Under the hypothesis that true mutations should be identified from at least two samples, we detected 68 "true" mutations and then calculated TP, FP, TPR, and FDR.

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

Table S5. Five low-frequency SNVs found only by LFMD. AA is short for amino acid.

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

Table S6. 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	

Table S7. Mutations simulated in the mixed double-strand sequencing data. The number of original background DNA fragments in the simulation is 10^6.

Chromosome	Position	Reference	Alternative	Type
MT	70	G	T	snv
MT	270	A	C	snv
MT	470	A	C	snv
MT	670	C	G	snv
MT	870	C	G	snv
MT	1070	C	CGA	ins
MT	1270	T	A	snv
MT	1470	A	C	snv
MT	1670	A	C	snv
MT	1870	A	C	snv
MT	2070	C	CGA	ins
MT	2270	A	C	snv
MT	2470	G	T	snv
MT	2670	C	G	snv
MT	2870	G	T	snv
MT	3070	G	GTCT	ins
MT	3270	C	G	snv
MT	3470	T	A	snv
MT	3670	G	T	snv
MT	3870	C	G	snv
MT	4070	A	AC	ins
MT	4270	T	A	snv
MT	4470	A	C	snv
MT	4670	C	G	snv
MT	4870	A	C	snv
MT	5070	A	AC	ins
MT	5270	C	G	snv
MT	5470	C	G	snv
MT	5670	A	C	snv

MT	5870	T	A	snv
MT	6070	T		del
MT	6270	G	T	snv
MT	6470	A	C	snv
MT	6670	A	C	snv
MT	6870	T	A	snv
MT	7070	C	CGA	ins
MT	7270	T	A	snv
MT	7470	C	G	snv
MT	7670	A	C	snv
MT	7870	T	A	snv
MT	8070	C	CGA	ins
MT	8270	C	G	snv
MT	8470	A	C	snv
MT	8670	A	C	snv
MT	8870	T	A	snv
MT	9070	T		del
MT	9270	C	G	snv
MT	9470	C	G	snv
MT	9670	A	C	snv
MT	9870	C	G	snv
MT	10070	C	CGA	ins
MT	10270	T	A	snv
MT	10470	A	C	snv
MT	10670	C	G	snv
MT	10870	C	G	snv
MT	11070	T		del
MT	11270	C	G	snv
MT	11470	A	C	snv
MT	11670	A	C	snv
MT	11870	C	G	snv
MT	12070	G	GTCT	ins
MT	12270	T	A	snv

MT	12470	T	A	snv
MT	12670	C	G	snv
MT	12870	C	G	snv
MT	13070	C	CGA	ins
MT	13270	C	G	snv
MT	13470	A	C	snv
MT	13670	A	C	snv
MT	13870	A	C	snv
MT	14070	A	AC	ins
MT	14270	A	C	snv
MT	14470	T	A	snv
MT	14670	T	A	snv
MT	14870	A	C	snv
MT	15070	C	CGA	ins
MT	15270	T	A	snv
MT	15470	T	A	snv
MT	15670	T	A	snv
MT	15870	A	C	snv
MT	16070	A	AC	ins
MT	16270	C	G	snv
MT	16470	G	T	snv

References

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