Supplementary Material - mtDNA-Server

Supplementary Figure 1: Sample Mix-ups of HM625679.1 and KC286589.1

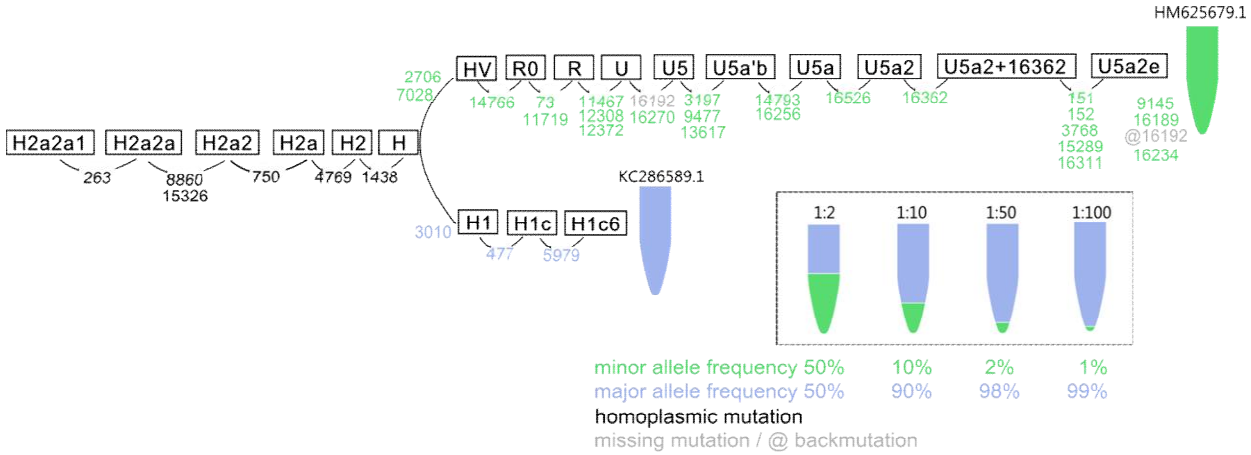

Figure Description:
We validated our approach based on 4 different sample-mix ups on Illumina HiSeq and on the IonTorrent PGM. Therefore, we mixed 2 samples in the laboratory as follows: M1 – 1:2 (50%), M2 – 1:10 (10%), M3 – 1:50 (2%), M4 – 1:100 (1%).
Sample 1: HM625679.1 http://www.ncbi.nlm.nih.gov/nuccore/301505851 (Haplogroup U5a2e)
Sample 2: KC286589.1 http://www.ncbi.nlm.nih.gov/nuccore/445067603) (Haplogroup H1c6)

Excluded from Sample 1: 15372
Excluded from Sample 2: 7076, 9462, 11150, 15236, 16129

27 expected sites: 73, 151, 152, 477, 2706, 3010, 3197, 3768, 5979, 7028, 9145, 9477, 11467, 11719, 12308, 12372, 13617, 14766, 14793, 15289, 16189, 16234, 16256, 16270, 16311, 16362, 16526

PIPELINE COMPARISION

We executed mtDNA-Server against different web services and command line tools. Since the mix-ups were too large for upload (> 50 MB), we used two publicly available data sets. For mtDNA-Server no parameters needed to be adjusted.

Evaluated data sets:

- 1000G Project Phase1 sample HG00096
  - http://mtdna-server.uibk.ac.at/assets/bam/HG00096.mapped.ILLUMINA.bwa.GBR.low_coverage.20101123.bam
- Tumor/Benign sample TCGA-BH-A0BM-01A-11W-A071-09 / TCGA-BH-A0BM-01A-01W-A071-09
  - https://github.com/riverlee/MitoSeek/blob/v1.3/Examples/brca_tumor.bam
  - https://github.com/riverlee/MitoSeek/blob/v1.3/Examples/brca_normal.bam

Compared Pipelines including used parameter sets:

- **MitoSeek (version 1.3)**
  perl mitoSeek.pl -i <input.bam> -t 4 -sb 0 -hp 1 -d 5 -str 4 -sp 1 -sa 0

- **MToolBox on MSeqDR**
  Input format: BAM, reference sequence hg19+rCRS, Filtering and extra option: none, Minimum distance of ins/dels from read end: 5 bps, and Heteroplasmy threshold for FASTA consensus sequence: 0.8.

- **Galaxy Naive Variant Caller**
  Minimum base quality, Minimum mapping quality and Minimum number of reads needed to consider a REF/ALT needed = 20, ploidy = 1

- **Mit-o-matic**
  Read length 101, files converted to FASTQ with BamToFastq, alignment Tool BWA, Data Single-End, Heteroplasmy cut-off 10%.
  For HG00096 which was too big for upload, we used the command-line version:
  perl mitomatic.pl -c -t bwa -o se -f 10 -d hg00096_10 -i HG00096.fastq

- **LoFreq**
  lofreq call HG00096.bam -o HG00096.vcf –f rCRSreference.fasta

- **MitoBamAnnotator**
  Unfortunately we were not able to run either of the samples on MitobamAnnotator.

Supplementary Table 1: HG00096 Sample Evaluation

Mutation	LoFreq high coverage (Gold standard)	LoFreq low coverage	Galaxy Naïve Variant Caller	MitoSeek	MSeqDR	Ye et al.	**mtDNA-Server**
1456 T/C	**1.09%**				1.20%		**1.01%**
2746 T/C	**1.84%**	2.34%	2.47%	2.40%	2.40%	2.29%	**2.47%**
3200 T/C	**0.93%**				1.00%		**1.02%**
12410 A/G	**1.07%**	1.27%	1.25%	1.20%	1.00%		**1.14%**
14071 A/G	**0.99%**	1.02%	1.16%		1.20%		**1.08%**
14569 G/A	**50.15%**	57.62%	57.71%	58.00%	59.30%	56.17%	**57.56%**
15463 A/G	**0.89%**				1.30%	1.08%	**1.25%**
16093 T/C	**56.83%**	60.19%	60.91%		60.20%	59.46%	**59.63%**
16360 C/T	**39.43%**	39.43%	38.80%		39.50%	37.78%	**38.56%**
*3488 T/A				1.10%	1.10%		
*6419 A/C			4.52%	1.50%	1.70%		
**10306 A/C			6.28%	2.50%	1.80%		

Supplementary Table 1: HG00096 high coverage has been analysed with LoFreq (~15,000 x) as a defined gold standard (bold). HG00096 low coverage data (~1,300 x) has then be executed on all web services and pipelines. Mutations highlighted in green are expected, red unexpected mutations. For unexpected, transversions only found on one strand are considered as artefacts and marked with *. Error hot spot mutations reported by Li et al ((3)) are marked with **. Mit-o-matic resulted in over 528 heteroplasmic sites when using 1% heteroplasmic threshold and 20 heteroplasmic sites with a 10% threshold, with a resulting haplogroup U8b1b1 instead of the expected H16a1 and was therefore excluded.

Supplemental Table 2: Tumor Sample Evaluation

Tumor mtDNA Positions found as heteroplasmies (mean cov. 197 x)					
mit-o-matic	Galaxy Naive Variant Caller	LoFreq	MToolbox on MSeqDR	MitoSeek	mtDNA-Server
				83	
				153	
195	195	195	195		195 (81%)
				217	
				290	
1149		1149	1149		1149 (14%)
		2960	2960		2960 (7%)
			4878 (G/GC)		
			5181 (A/G)		
6419					

	8165	8165	8165		8165 (36%)
			8940 (C/T)		
10306*					
12414		12414	12414		12414 (98%)
12661	12661	12661	12661		12661 (24%)
	15612	15612	15612		15612 (37%)
16271	**16271**	**16271**	**16271**		**16271 (15%)**

Supplementary Table 2: The original mutations reported by MitoSeek couldn't be confirmed with either of the web-servers or LoFreq. Entries in the table represent heteroplasmic mutations annotated by the positions on the rCRS. Mutations highlighted in red are possible false positives. Mutations on 6419 and 10306 are transversions. Mutations marked with * are reported error hot spot by Li et al, 2010 (see Paper for reference). Additional mutations found with MToolBox on 4878, 5181, 8940 can be explained either by length heteroplasmies or sequencing issues and can't be interpreted as correct, nor false positives without further investigation.

Supplemental Table 3: Benign Sample Evaluation

Benign mtDNA Positions found as heteroplasmies (mean cov. 55 x)					
mit-o-matic	Galaxy Naive Variant Caller	LoFreq	MToolbox on MSeqDR	MitoSeek	mtDNA-Server
				45	
				48	
				98	
				99	
				195	
213T/A					
				239	
4657A/G					
4658A/G					
6419A/C					
10197G/C					
10306A/C*		10306*			
16271	**16271**	**16271**	**16271**		**16271 (16%)**

Supplementary Table 3: The original mutations reported by MitoSeek (http://htmlpreview.github.io/?https://github.com/riverlee/MitoSeek/blob/release/brca_tumor/mitoSeek.html) couldn't be confirmed with either of the web-servers or LoFreq. Entries in the table represent heteroplasmic mutations annotated by the positions on the rCRS. Mutations highlighted in red are possible

false positives. *Mutation on 10306 is reported as error hot spot by Li et al. Mutation 16271 found in tumor and benign hints to a germline mutation.