Detection of mtDNA heteroplasmy from second generetion sequencing data.

Jclub in bioinformatics
Tarmo Puurand
06.02.2012

Heteroplasmy detection methods

- Sanger sequencing
- High-performance liquid chromatography (HPLC)
- Pyrosequencing
- SnaPshot
- High-resolution melt (HRM) profiling
- Temporal temperature gradient gel electroforesis (TTGE)
- Invader assay
- Amplification refractory mutation system
- Surveyor nuclease

Data

147 individuals from Georgia, Armenia, Azerbaijan, Iran and Turkey.
Two overlapping PCR products 9.7 and 7.3 kb in length.
Illumina GAII platform, multiplex sequencing protocol.

Individuals	Read length	Coverage
97	36	65x
17	36	78x (replicated runs)
17	76	211x
4	76	(additional runs)

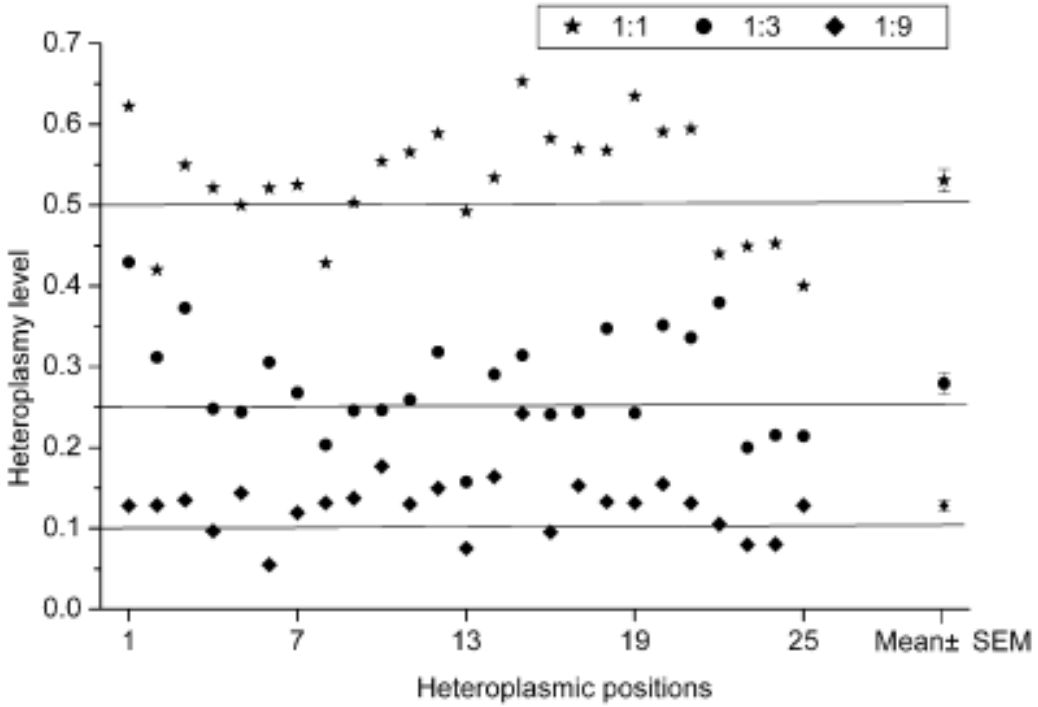
PhiX174 analysis

Table 1. Number of Heterozygous Positions Detected in the PhiX174 Genome under Different Criteria

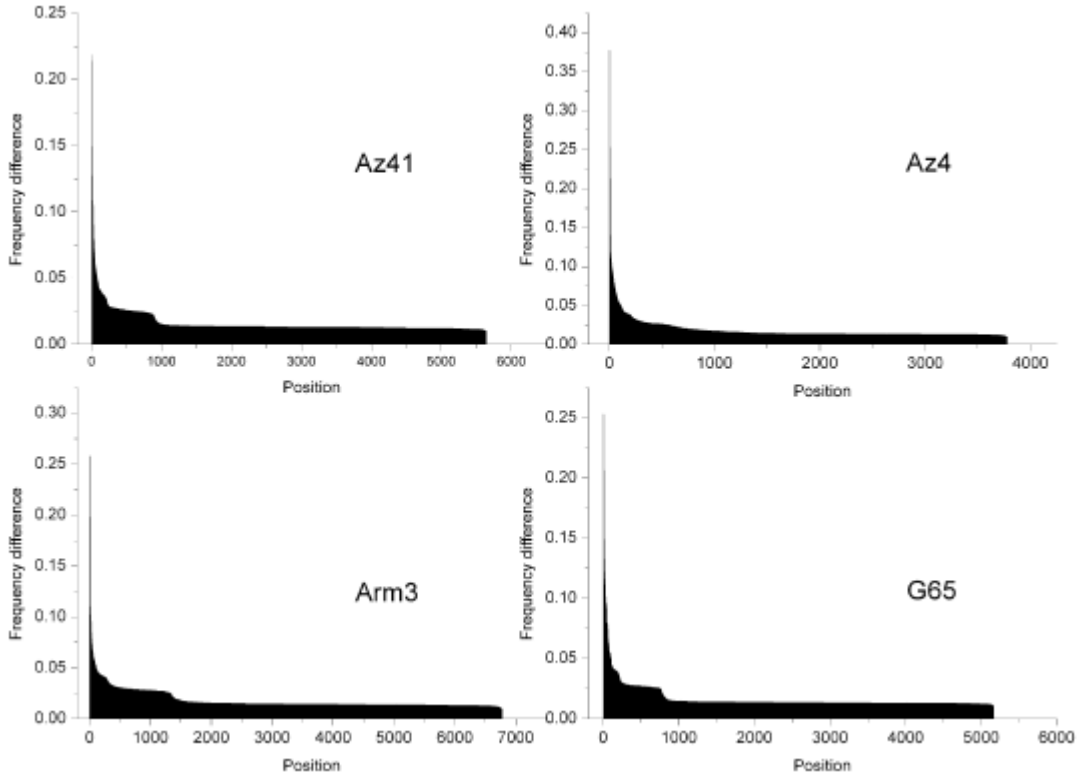
	No Quality Filter	QS ≥ 20	QS ≥ 23
Validated by one strand	582.9 ± 40.9[a]	192.8 ± 14.0	126.3 ± 24.6
Validated by two strands	17.9 ± 4.7	3.0 ± 1.4	1.9 ± 1.0

[a] Standard deviation based on 100 resamplings of the data.

Artificial mixes



Validation of replicates



Detected polymorfic positions

Position	Individual	Coverage	Major Allele	Frequency	Minor Allele	Frequency	Gene Annotation[a]
64	Ir28	82	T	0.87	C	0.13	CR
146	Ir17	62	T	0.90	C	0.10	CR
146	G65	186	T	0.80	C	0.20	CR
150[b]	Arm17	69	C	0.88	T	0.12	CR
152	Ir11	71	T	0.89	C	0.11	CR
195[b]	G67	67	T	0.90	C	0.10	CR
203	Az5	67	A	0.66	G	0.34	CR
204	Arm17	75	T	0.89	C	0.11	CR
1552	Ir54	69	G	0.87	A	0.13	12S
3014	Az10	44	G	0.86	T	0.14	16S
3492	Arm25	60	A	0.65	C	0.35	NS(*ND1*; Lys>Asn)

Position	Individual	Coverage	Allele 1	Frequency	Allele 2	Frequency	Gene Annotation
57	Ir28	82	C	0.84	-	0.16	CR
15940	Arm2	208	-	0.85	T	0.15	*tRNA-THR*
248	Ir10	68	A	0.84	-	0.16	CR

Samples

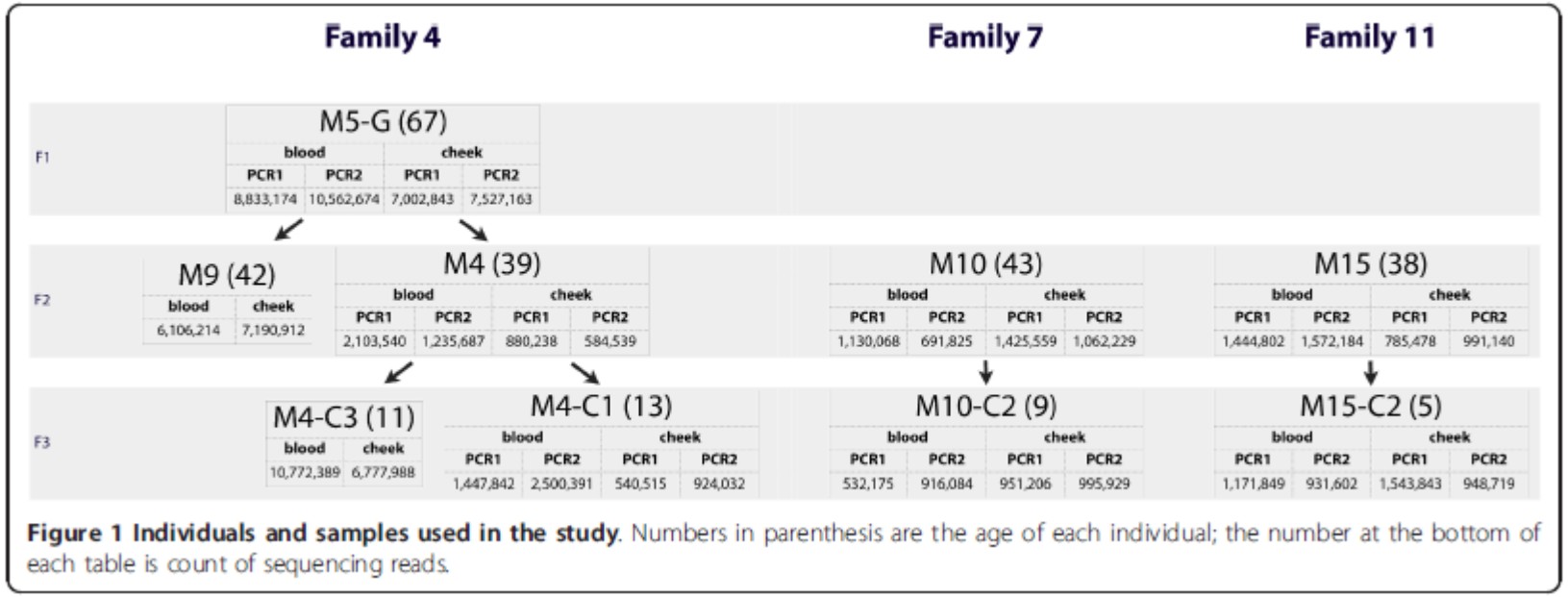


Figure 1 Individuals and samples used in the study. Numbers in parenthesis are the age of each individual; the number at the bottom of each table is count of sequencing reads.

Polymorfic positions types

- Sites without allele frequency shifts
- Sites with allele frequency shift
- Sites with *de novo* mutations

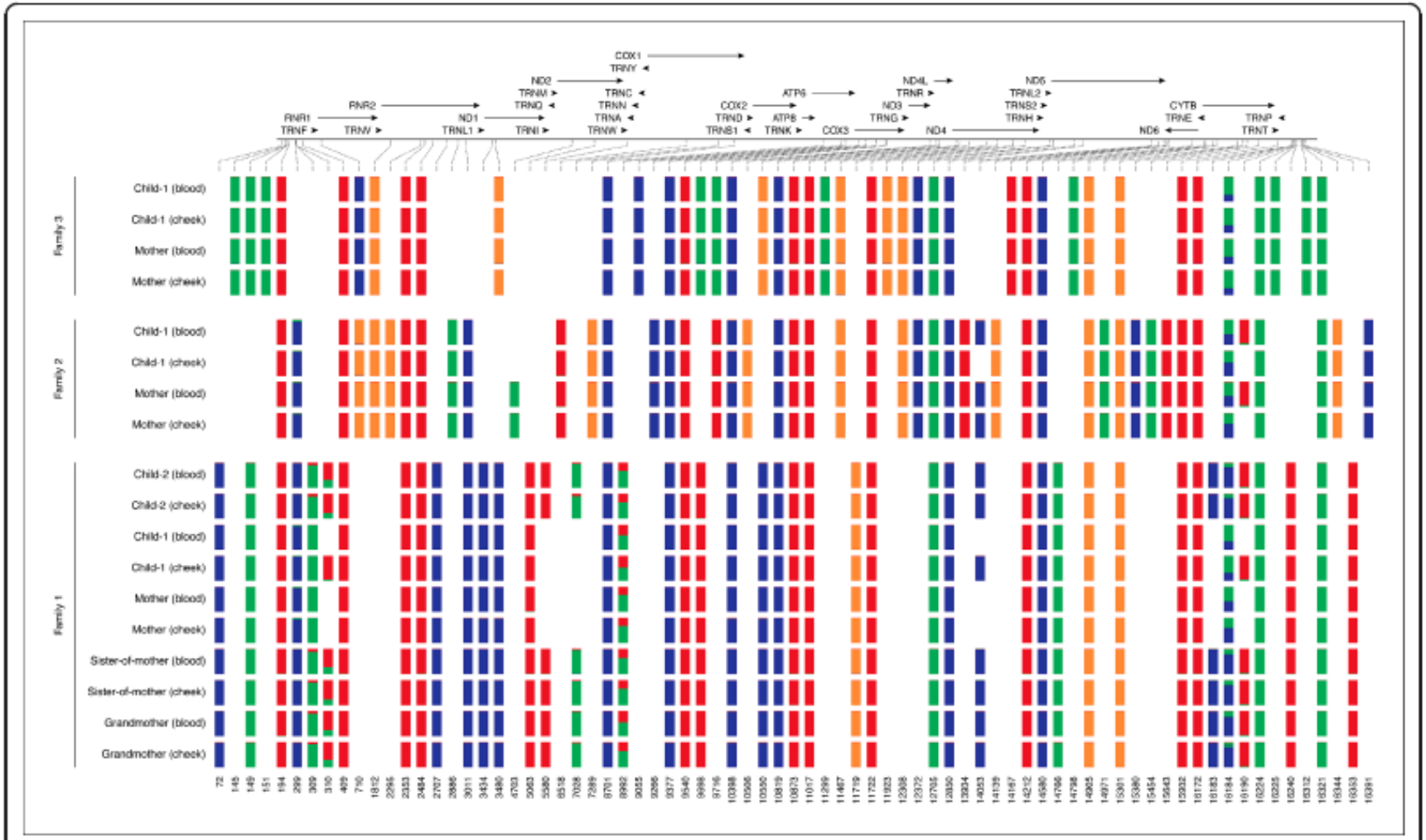


Figure 3 A representation of all differences found between each sequenced individual and the reference human mtDNA from genome build hg19. The colored bars (blue = A, green = C, orange = G, red = T) represent the frequency of a given allele in each sample. For example, at position 8,992 one can clearly see a heteroplasmy with two high frequency alleles C and T. Lines on top of the image represent location and orientation of mitochondrial genes. F1 = Family F4, F2 = Family F7, F3 = Family F11.

Detected polymorfic positions in family 4

Table 1 Allele frequencies at heteroplasmic sites in Family F4.

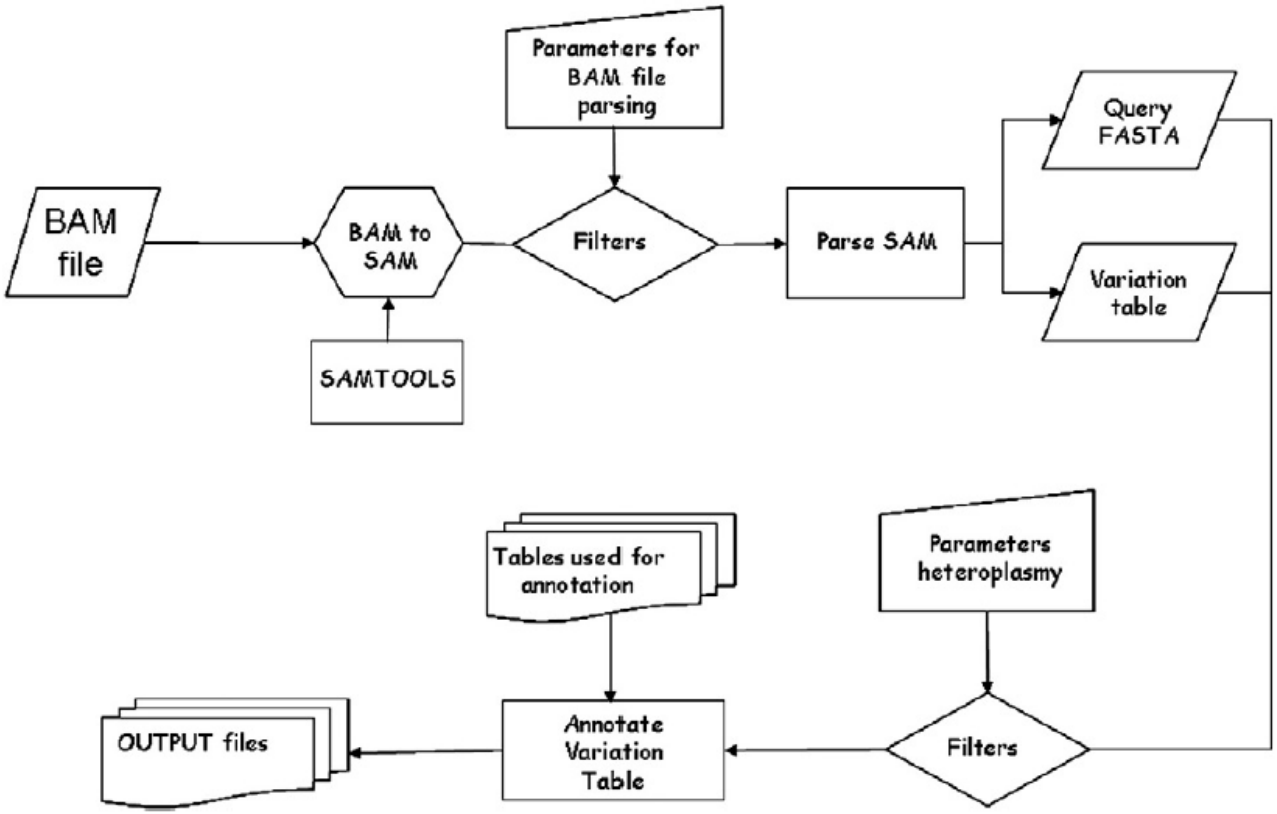
			Family F4																								
Tissue	Site	Ref	M5G (grandmother)					M9 (daughter of M5G)					M4 (daughter of M5G)					M4-C1 (child of M4)					M4-C3 (child of M4)				
			A	C	G	T	cvrg	A	C	G	T	cvrg	A	C	G	T	cvrg	A	C	G	T	cvrg	A	C	G	T	cvrg
blood	5063	T	0.000	0.001	0.000	*0.998*	81,207	0.000	0.001	0.000	*0.999*	21,069	0.000	0.016	0.000	*0.984*	12,376	0.000	0.001	0.000	*0.999*	5,228	0.000	0.001	0.000	*0.999*	50,019
	7028	T	0.002	*0.975*	0.001	*0.021*	5,739	0.001	*0.966*	0.001	*0.032*	1,671	0.000	*0.975*	0.000	*0.025*	5,102	no data	0.002	*0.910*	0.000	*0.088*	4,036				
	8992	C	0.000	*0.652*	0.000	*0.347*	52,519	0.000	*0.659*	0.000	*0.341*	15,597	0.000	*0.672*	0.000	*0.327*	14,174	0.000	*0.526*	0.000	*0.474*	4,585	0.000	*0.670*	0.000	*0.330*	35,005
			A	*C*	*T*	*G*	cvrg	*A*	*C*	*T*	*G*	cvrg	*A*	*C*	*T*	*G*	cvrg	*A*	*C*	*T*	*G*	cvrg	*A*	*C*	*T*	*G*	cvrg
cheek	5063	T	0.000	0.001	0.000	*0.999*	59,896	0.000	0.001	0.000	*0.999*	20,635	0.000	*0.020*	0.000	*0.980*	2,294	0.000	0.002	0.000	*0.998*	2,073	0.000	0.001	0.000	*0.998*	29,013
	7028	T	0.001	*0.982*	0.001	0.015	3,905	0.001	*0.965*	0.001	*0.033*	1,526	no data					no data					0.001	*0.965*	0.000	*0.034*	2,071
	8992	C	0.000	*0.545*	0.000	*0.454*	38,968	0.000	*0.639*	0.000	*0.360*	14,624	0.000	*0.686*	0.000	*0.314*	1,931	0.001	*0.578*	0.000	*0.421*	1,433	0.000	*0.669*	0.000	*0.330*	19,214

The frequencies were calculated by dividing the number of reads supporting a given allele by the quality adjusted coverage listed in "coverage" column. Quality adjusted coverage = number of reads where the base aligning over a given position has a phred score equal or higher than 30.

Galaxy workflow

- The workflow starts with the sequencing reads, maps them using BWA mapper, splits the results into two strandspecific branches (one for the plus strand and one for the minus strand), transforms datasets from read-centric (Sequence Alignment/ Map (SAM)) to genome-centric form (pileup) and performs a number of filtering and thresholding steps before merging the branches and generating a list of sites that contain allelic variants with the frequency above 0.01.

Pipeline


Parameters for BAM file parsing
Query FASTA
BAM file
BAM to SAM
Filters
Parse SAM
Variation table
SAMTOOLS
Tables used for annotation
Parameters heteroplasmy
OUTPUT files
Annotate Variation Table
Filters

Options

The parameters used by MitoBamAnnotator.

Parameter name	Brief description	Default value
The parameters used to reject reads and bases		
Min base quality	A base is ignored if its quality (as provided in the BAM file) is below this threshold	23
Min worst base quality	A read is rejected if the lowest quality of any of its bases is below this threshold	5[a] (23)
Max mismatch rate	A read is discarded if it has a fraction of mismatches above this threshold. Insertions and deletions are considered as mismatches	3/36[b]
Min coverage for consensus	Minimal number of high-quality bases for calling a consensus base	3
Min mapping score	A read is discarded if its mapping score is below this threshold	23
Min read length	Minimal length of reads to be considered	25[b] bp (36)
Max ambiguous bases	Maximal number of ambiguous ('N') bases that are allowed in a read	0[a] (3/36)
Remove PCR duplicates	Use the samtools rmdup command to eliminate PCR duplicates. Require that the mode of sequencing (single- or paired-end) is specified.	Yes[c]
The parameters used to imply heteroplasmy		
Min position coverage	Minimal coverage for a position to be considered for heteroplasmy analysis	10
Min secondary fraction	Minimal fraction of base calls (in% of high-quality bases) that support a secondary base	1.6[a] (1.6)
Min secondary strand coverage	Minimal coverage of the secondary base on either strand (forward or reverse). Note that the minimal combined coverage considering both strands is *x*2 larger than this value	2
Min secondary fraction per strand	Minimal fraction of base calls (in% of high-quality bases) on either strand (forward or reverse).	0.8[c]
Min reads in each strand supporting second best base	Minimal support for a secondary base, expressed as the number of base calls that support it in both in the reverse and the forward orientation.	5[b] (1)
Filter out low complexity regions	Ignore nucleotides flanked by low complexity regions	Yes[c]

The parameters used by MitoBamAnnotator for filtering reads, bases and suspected heteroplasmy. Reads, bases and variants are only reported if they meet all the criteria described in this table. All of these parameters can be changed.

Output files

- MitoBamAnnotator provides the user with five files by email:

(i) the consensus sequence in FASTA format;

(ii) (ii) a list of single nucleotide differences from the Fig. 1. Revised Cambridge Reference Sequence ("rCRS", NC_012920) in the format required by HaploGrep (Kloss-Brandstatter et al., 2010). This file is required to examine the possibility that the secondary reads reflect contamination with an external source of DNA by considering the possibility that the secondary reads originate from an individual carrying mtDNA from a different haplogroup (Salas et al., 2005). Note that even if the underlying alignment was generated using the Yoruba mtDNA sequence (NC_001807.4, which differs from the rCRS by few SNPs and indel positions) as a reference, this report will include a comparison to the rCRS;

(iii) an annotated variation table (a detailed table description and an example of this table can be found in Supplementary Table 1);

(iv) a subset of the annotated variation table, including only putative heteroplasmic sites; and

(v) a log file containing descriptive statistics summarizing the filtration process.

References

- Li M, Schönberg A, Schaefer M, Schroeder R, Nasidze I, Stoneking M. Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. Am J Hum Genet. 2010 Aug 13;87(2):237-49.
- Goto H, Dickins B, Afgan E, Paul IM, Taylor J, Makova KD, Nekrutenko A. Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. Genome Biol. 2011 Jun 23;12(6):R59. [Epub ahead of print]
- Zhidkov I, Nagar T, Mishmar D, Rubin E. MitoBamAnnotator: A web-based tool for detecting and annotating heteroplasmy in human mitochondrial DNA sequences. Mitochondrion. 2011 Nov;11(6):924-8. Epub 2011 Aug 22.