mity: A highly sensitive mitochondrial variant analysis pipeline for whole genome sequencing data

Clare Puttick[1*], Kishore R Kumar[1,2,3,4], Ryan L Davis[1,2], Mark Pinese[5,6,9], David M Thomas[5,7], Marcel E Dinger[1,8], Carolyn M Sue[1,2,4,*] and Mark J Cowley[1,6,7,9*]

[1]Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia, [2]Department of Neurogenetics, Kolling Institute, Royal North Shore Hospital and University of Sydney, St Leonards, NSW, Australia, [3]Molecular Medicine Laboratory, Concord Hospital, Sydney, Australia, [4]Department of Neurology, Royal North Shore Hospital, Northern Sydney Local Health District, St Leonards, NSW, Australia, [5]Cancer Division, Garvan Institute of Medical Research, Darlinghurst, NSW, Australia, [6]Children's Cancer Institute, University of New South Wales, Randwick, NSW, Australia, [7]St Vincent's Clinical School, University of New South Wales, Sydney, Australia, [8]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, Australia, [9]School of Women's and Children's Health, University of New South Wales, Sydney, Australia

*To whom correspondence should be addressed.

Abstract

Motivation: Mitochondrial diseases (MDs) are the most common group of inherited metabolic disorders and are often challenging to diagnose due to extensive genotype-phenotype heterogeneity. MDs are caused by mutations in the nuclear or mitochondrial genome, where pathogenic mitochondrial variants are usually heteroplasmic and typically at much lower allelic fraction in the blood than affected tissues. Both genomes can now be readily analysed using unbiased whole genome sequencing (WGS), but most nuclear variant detection methods fail to detect low heteroplasmy variants in the mitochondrial genome.

Results: We present *mity*, a bioinformatics pipeline for detecting and interpreting heteroplasmic SNVs and INDELs in the mitochondrial genome using WGS data. In 2,980 healthy controls, we observed on average 3,166× coverage in the mitochondrial genome using WGS from blood. *mity* utilises this high depth to detect pathogenic mitochondrial variants, even at low heteroplasmy. *mity* enables easy interpretation of mitochondrial variants and can be incorporated into existing diagnostic WGS pipelines. This could simplify the diagnostic pathway, avoid invasive tissue biopsies and increase the diagnostic rate for MDs and other conditions caused by impaired mitochondrial function.

Availability: *mity* is available from https://github.com/KCCG/mity under an MIT license.

Contact: clare.puttick@crick.ac.uk, carolyn.sue@sydney.edu.au, MCowley@ccia.org.au

1 Supplementary information: Attached here

2 Introduction

The human mitochondrial (MT) genome is a 16,569bp circular chromosome, with potentially thousands of copies in a given cell (Davis et al., 2018), and a mutation rate 19x higher than the nuclear genome (Tuppen et al., 2010). Mitochondrial diseases (MDs) are highly heterogeneous genetic disorders, characterised by mitochondrial respiratory chain impairment (Gorman et al., 2016) and are caused by pathogenic variants in either the MT or nuclear genome. Over 300 pathogenic MT variants have been reported (Wallace, 2018), and variants in over 300 nuclear genes associated with disease (Davis et al., 2018). Variants can be present in variable proportions of mitochondrial genomes, known as heteroplasmy. Furthermore, heteroplasmy changes with age and between tissues, being higher in the affected tissue and lower in more accessible tissues such as blood (Sue et al., 1998).

Current clinical-grade whole genome sequencing (WGS) at 30–40× nuclear coverage provides high coverage of the MT genome (1,000–100,000×), suggesting very low levels of heteroplasmy could be reliably detected. With high-coverage however, systematic sequencing errors accumulate, particularly in certain sequence contexts, making it challenging to discern true pathogenic variants from noise (Griffith et al., 2015). Most popular variant callers are optimised for diploid analysis and are thus incapable of identifying low heteroplasmy MT variants. Existing approaches for MT-DNA analysis are web-based (Lee et al., 2008; Weissensteiner et al., 2016) or GUI-based (Ishiya and Ueda, 2017). They are thus less

amenable to high-throughput, reproducible analysis, or only validated only on high heteroplasmy variants (Santorsola et al., 2016).

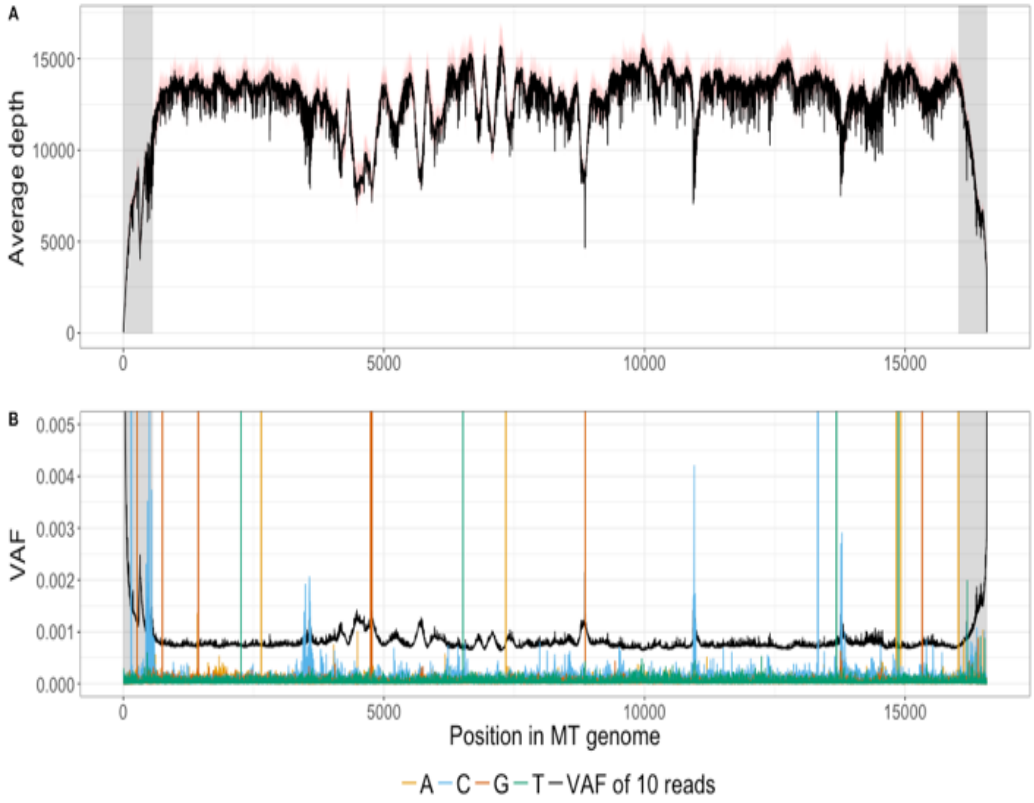
Here, we present *mity*, a bioinformatics pipeline to detect MT SNVs and INDELs from WGS, to assist clinicians and researchers with the diagnosis of MDs. *mity* was optimised to identify low heteroplasmy variants, to be easily integrated into existing high-throughput analysis pipelines, and to generate a highly interpretable report to aid molecular diagnosis.

3 Approach

mity consists of three modules that easily integrate MT sequence analysis into existing nuclear WGS analysis pipelines (Supp. Fig 1). The first module, *mity-call* analyses a BAM file to call, filter and normalise MT SNVs and INDELs, producing a *mity* VCF. The second module, *mity-report* creates easily interpretable spreadsheet reports with extensive annotations. Finally, the *mity-merge* module merges nuclear and *mity* VCFs to produce a single high-quality VCF. This allows for seamless integration of *mity* into existing production or clinical-grade analysis pipelines for subsequent variant interpretation. *mity* has been optimised for 30–40× Illumina HiSeq X data aligned to the GRCh37 + decoy (hs37d5) reference genome using BWA-MEM.

2.1 mity-call

We first assessed the performance of GATK HaplotypeCaller (Poplin et al., 2018), FreeBayes (Garrison and Marth, 2012), and LoFreq (Wilm et al., 2012) for MT variant detection (Supp. Results). We selected FreeBayes as *mity-call*'s underlying



variant caller, as it was more sensitive than HaplotypeCaller, particularly for low heteroplasmy variants and more specific than LoFreq (Supp. Fig 2). Using default FreeBayes settings, the reproducibility from 13 WGS replicates of the NA12878 cell line was poor (Supp. Fig 3a). We thus optimised the mapping quality (MQ≥30), and base quality (BQ≥24) filters, resulting in an average of 17.5±0.5 variants per replicate, with variant allele frequency (VAF) >0.01 (Supp. Fig 3a). These filters caused a minimal reduction of sequencing depth in 2,980 healthy controls (Lacaze et al., 2019; Pinese et al., n.d.) (Supp. Fig 3b). Excluding the D-loop, sequencing depth was generally >10,000× (Figure 1A) in the 13 replicates of NA12878, with low per-base noise, supporting a conservative lower-heteroplasmy limit of >0.002 (Figure 1B). Multi-allelic variants are normalised using a custom algorithm (Supp. Results; Supp. Fig 4).

Fig. 1. The high average depth across the MT genome in WGS data means that *mity* can detect very low heteroplasmy variants. A The average depth (black; red = interquartile range) across 13 replicates of NA12878, with BQ≥24 and MQ≥30. **B** The VAF of all three possible non-reference bases, at each position in the MT genome is typically far lower than the VAF corresponding to 10 high-quality reads (black). Spikes of alternate reads with VAF>0.002 outside the D-loop (grey) correspond to true genetic variants.

The default variant quality score in FreeBayes penalises low VAF variants, due to the overwhelmingly high numbers of reference reads. However, we reasoned that the evidence to support low heteroplasmy MT variants should a) only consider the alternate read count, b) scale with the alternate read count, and c) be reported using a similar scale to other variant quality methods. To achieve this, we used a binomial model to implement a Phred-scaled variant quality score, q. Assuming a noise level p, q is the Phred-scaled probability of observing at least n alternate reads by chance, given the total number of reads covering the variant position (Supp. Results). The calculation of q is fast, VAF independent, and has a default threshold of q≥30, which can be tuned to favour sensitivity or specificity.

mity includes three additional filters: a strand bias filter to exclude variants with >90% or <10% alternative reads from one strand, a read depth filter set to <15× depth, and a region filter to exclude variants in the homopolymeric regions at m.302-319, or m.3105-3109, where there is an 'N' at m.3107, in the GRCh37 version of the mitochondrial sequence.

2.2 mity-report

End-users of *mity* include genome researchers and clinicians, so *mity-report* was developed to produce easily interpretable spreadsheet reports containing comprehensively annotated MT variant lists. Variants are tiered by VAF to aid prioritisation: tier 1, VAF ≥0.01; tier 2, VAF <0.01 with >10 supporting reads; and tier 3 are the remaining variants. Variant impact is estimated using Variant Effect Predictor (VEP) (McLaren et al., 2016), with additional annotation from MITOMAP (Brandon et al., 2005), allele frequency from 2,980 WGS healthy individuals from MGRB and additional VCF fields. The mitochondrial haplogroup is determined using PhyloTree (van Oven and Kayser, 2009).

2.3 mity-merge

In order to integrate *mity* into existing WGS analysis pipelines, *mity-merge* replaces the MT variants from a genome-wide VCF (e.g., GATK), with those from the *mity* VCF. The merged VCF can then be used with existing downstream tools for annotation and filtering.

2.4 NUMT homology

Nuclear mitochondrial DNA (NUMT) are fragments of the mitochondrial genome integrated into the nuclear genome (Lopez et al., 1994) that can potentially confound heteroplasmy (Parr et al., 2006; Santibanez-Koref et al., 2019). Informed by our previous work in pseudogene homology (Mallawaarachchi et al., 2016), and sequence homology analysis (Supp. Methods & Supp. Results), we conclude that NUMT:MT homology to known NUMT is highly unlikely to cause false-positive tier 1 *mity* variants (Supp. Figure 7). It is challenging to rule out the presence of very rare, or patient-specific NUMT, so we recommend low heteroplasmy variants, particularly tiers 2 and 3 be validated using orthogonal approaches.

2.5 Performance

mity can operate on either a WGS or MT-only BAM file (<2.1Gb), with a run-time of <10 minutes using a single-core and <8Gb RAM.

4 Conclusion

mity overcomes many of the challenges of accurate low heteroplasmy variant identification in the MT genome. Additional work is needed to identify MT structural variation, variant phasing, and prioritisation of novel MT variants. *mity* can be easily incorporated into existing high-throughput analysis pipelines, while simultaneously producing user-friendly reports. *mity* was developed on 242 MD patients (Davis et al, in prep) and 2,980 healthy individuals. By extending the scope of variant analysis in patient data, *mity* helps support further adoption of clinical WGS.

Acknowledgements

We thank members of the Kinghorn Centre for assistance with data generation and patients for contributing their samples to this research.

Funding

MJC and RLD were supported by NSW Health Early-Mid Career Fellowships. KRK was supported by an NHMRC Early Career Fellowship (APP1091551). CMS was supported by an NHMRC Practitioner Fellowship (APP1008433). This work was supported by a NSW Genomics Collaborative grant (CMS, MED, RLD, KRK) and the NSW Health-funded Medical Genomics Reference Bank (DT, MED). We acknowledge financial support from the Kinghorn Foundation, without which this research would not have been possible.

Conflict of Interest: none declared.

References

Brandon, M.C., Lott, M.T., Nguyen, K.C., Spolim, S., Navathe, S.B., Baldi, P., Wallace, D.C., 2005. MITOMAP: a human mitochondrial genome database--2004 update. Nucleic Acids Res. 33, D611-613. https://doi.org/10.1093/nar/gki079

Davis, R.L., Liang, C., Sue, C.M., 2018. Mitochondrial diseases. Handb. Clin. Neurol. 147, 125–141. https://doi.org/10.1016/B978-0-444-63233-3.00010-5

Garrison, E., Marth, G., 2012. Haplotype-based variant detection from short-read sequencing. ArXiv12073907 Q-Bio.

Gorman, G.S., Chinnery, P.F., Dimauro, S., Hirano, M., Koga, Y., McFarland, R., Suomalainen, A., Thorburn, D.R., Zeviani, M., Turnbull, D.M., 2016. Mitochondrial diseases. Nat. Publ. Group 2, 1–23. https://doi.org/10.1038/nrdp.2016.80

Griffith, M., Miller, C.A., Griffith, O.L., Krysiak, K., Skidmore, Z.L., Ramu, A., Walker, J.R., Dang, H.X., Trani, L., Larson, D.E., Demeter, R.T., Wendl, M.C., McMichael, J.F., Austin, R.E., Magrini, V., McGrath, S.D., Ly, A., Kulkarni, S., Cordes, M.G., Fronick, C.C., Fulton, R.S., Maher, C.A., Ding, L., Klco, J.M., Mardis, E.R., Ley, T.J., Wilson, R.K., 2015. Optimizing Cancer Genome Sequencing and Analysis. Cell Syst. 1, 210–223. https://doi.org/10.1016/j.cels.2015.08.015

Ishiya, K., Ueda, S., 2017. MitoSuite: a graphical tool for human mitochondrial genome profiling in massive parallel sequencing. PeerJ 5, e3406. https://doi.org/10.7717/peerj.3406

Lacaze, P., Pinese, M., Kaplan, W., Stone, A., Brion, M.-J., Woods, R.L., McNamara, M., McNeil, J.J., Dinger, M.E., Thomas, D.M., 2019. The Medical Genome Reference Bank: a whole-genome data resource of 4000 healthy elderly individuals. Rationale and cohort design. Eur. J. Hum. Genet. 27, 308–316. https://doi.org/10.1038/s41431-018-0279-z

Lee, H.Y., Song, I., Ha, E., Cho, S.-B., Yang, W.I., Shin, K.-J., 2008. mtDNAmanager: a Web-based tool for the management and quality analysis of mitochondrial DNA control-region sequences. BMC Bioinformatics 9, 483. https://doi.org/10.1186/1471-2105-9-483

Lopez, J.V., Yuhki, N., Masuda, R., Modi, W., O'Brien, S.J., 1994. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J. Mol. Evol. 39, 174–190. https://doi.org/10.1007/bf00163806

Mallawaarachchi, A.C., Hort, Y., Cowley, M., McCabe, M.J., Minoche, A., Dinger, M.E., Shine, J., Furlong, T.J., 2016. Whole-genome sequencing overcomes pseudogene homology to diagnose autosomal dominant polycystic kidney disease. Eur. J. Hum. Genet. EJHG 24, 1584–1590. https://doi.org/10.1038/ejhg.2016.48

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., Cunningham, F., 2016. The Ensembl Variant Effect Predictor. Genome Biol. 17, S8. https://doi.org/10.1093/nar/gku1206

Parr, R.L., Maki, J., Reguly, B., Dakubo, G.D., Aguirre, A., Wittock, R., Robinson, K., Jakupciak, J.P., Thayer, R.E., 2006. The pseudo-mitochondrial genome influences mistakes in heteroplasmy interpretation. BMC Genomics 7, 185. https://doi.org/10.1186/1471-2164-7-185

Pinese, M., Lacaze, P., Rath, E.M., Stone, A., Brion, M.-J., Ameur, A., Nagpal, S., Puttick, C., Husson, S., Degrave, D., Navin Cristina, T., Silva Kahl, V.F., Statham, A.L., Woods, R.L., McNeil, J.J., Riaz, M., Barr, M., Nelson, M.R., Reid, C.M., Murray, A.M., Shah, R.C., Wolfe, R., Atkins, J.R., Fitzsimmons, C., Cairns, H.M., Green, M.J., Carr, V.J., Cowley, M., Pickett, H.A., James, P.A., Powell, J.E., Kaplan, W., Gibson, G., Gyllensten, U., Cairns, M.J., McNamara, M., Dinger, M.E., Thomas, D.M., n.d. The Medical Genome Reference Bank: Whole genomes and phenotype of 2,570 healthy elderly. https://doi.org/10.1101/473348

Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Auwera, G.A.V. der, Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M.J., Neale, B., MacArthur, D.G., Banks, E., 2018. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv 201178. https://doi.org/10.1101/201178

Santibanez-Koref, M., Griffin, H., Turnbull, D.M., Chinnery, P.F., Herbert, M., Hudson, G., 2019. Assessing mitochondrial heteroplasmy using next generation sequencing: A note of caution. Mitochondrion 46, 302–306. https://doi.org/10.1016/j.mito.2018.08.003

Santorsola, M., Calabrese, C., Girolimetti, G., Diroma, M.A., Gasparre, G., Attimonelli, M., 2016. A multi-parametric workflow for the prioritization of mitochondrial DNA variants of clinical interest. Hum. Genet. 135, 121–136. https://doi.org/10.1007/s00439-015-1615-9

Sue, C.M., Quigley, A., Katsabanis, S., Kapsa, R., Crimmins, D.S., Byrne, E., Morris, J.G.L., 1998. Detection of MELAS A3243G point mutation in muscle, blood and hair follicles. J. Neurol. Sci. 161, 36–39. https://doi.org/10.1016/S0022-510X(98)00179-8

Tuppen, H.A.L., Blakely, E.L., Turnbull, D.M., Taylor, R.W., 2010. Mitochondrial DNA mutations and human disease. Biochim. Biophys. Acta BBA - Bioenerg. 1797, 113–128. https://doi.org/10.1016/j.bbabio.2009.09.005

van Oven, M., Kayser, M., 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum. Mutat. 30, E386-394. https://doi.org/10.1002/humu.20921

Wallace, D.C., 2018. Mitochondrial genetic medicine. Nat. Genet. 50, 1642–1649. https://doi.org/10.1038/s41588-018-0264-z

Weissensteiner, H., Forer, L., Fuchsberger, C., Schöpf, B., Kloss-Brandstätter, A., Specht, G., Kronenberg, F., Schönherr, S., 2016. mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. Nucleic Acids Res. 44, W64-69. https://doi.org/10.1093/nar/gkw247

Wilm, A., Aw, P.P.K., Bertrand, D., Yeo, G.H.T., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L., Nagarajan, N., 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 40, 11189–11201. https://doi.org/10.1093/nar/gks918

5 Supplementary Methods

Patient recruitment

We recruited 10 adult patients reviewed at the Mitochondrial Disease Clinic at Royal North Shore Hospital, Sydney, Australia, between 2013-2015. The research was approved by the Northern Sydney Local Health District Human Research Ethics Committee (HREC/10/HAWKE/132). Total genomic DNA was isolated from peripheral blood using standard methods. All participants provided written informed consent. NA12878 reference material was sourced from Genome in a Bottle.

Sequencing and read alignment

Sequencing libraries were created from nine patients in singlicate, one patient in duplicate, and 13 replicates from NA12878, using Illumina TruSeq Nano HT v2.5 library preparation kits, using Hamilton Star instruments. Sequencing was performed on Illumina HiSeq X instruments following the manufacturers specifications at the Kinghorn Centre for Clinical Genomics, Sydney. Sequence reads were aligned to the human genome reference assembly GRCh37 decoy genome (hs37d5) using BWA-MEM (v0.7.12-r1039, settings -M; (Li, 2013)). Reads were further processed using GATK Indel Realignment, and GATK Base Recalibration (version 3.3; (Van der Auwera et al., 2013)). Depth of coverage was performed using bedtools genomecov (Quinlan, 2014), and depth of alternate alleles in Figure 1B with samtools mpileup (Li et al., 2009).

Nuclear insertions of mitochondrial origin (NUMTs)

From the comprehensive RHNumtS.2 catalogue of NUMT (Ramos et al., 2011), we identified HSA_NumtS_001 as the only candidate NUMT with sufficient length and homology to be investigated (Supp. Fig 7). We aligned the mitochondrial sequence chrM[GRCh37]:m.5842-9755, corresponding to HSA_NumtS_001 using BLASTN (Altschul et al., 1990), and characterised the distribution of genetic variation within the corresponding NUMT region at chr1[GRCh37]:g.566,391-570,303.

Supplementary Results

Evaluation of variant callers

GATK HaplotypeCaller (Poplin et al., 2018) is a popular genome-wide variant caller, but it has three major limitations for MT variant calling: 1) it down-samples the reads to a maximum of 500× depth, 2) it uses a diploid model by default, which is insensitive to low heteroplasmy variants and 3) does not provide a minimum VAF setting. LoFreq (Wilm et al., 2012) was developed specifically for detecting low-frequency variants in somatic cancer data. FreeBayes (Garrison and Marth, 2012) is a popular haplotype-aware, genome-wide variant caller, which allows for control over the minimum VAF and the minimum number of alternative-reads. Whilst FreeBayes and HaplotypeCaller both have ploidy parameters that can theoretically be tuned to prioritise low VAF variants, from a high-ploidy sample, the execution runtime becomes exponentially slower and computationally infeasible in practice for MT analysis.

We assessed the performance of GATK HaplotypeCaller, FreeBayes, and LoFreq for MT variant detection. We first sequenced ten patient genomes using a single lane of Illumina HiSeq X, resulting in 30–40x average nuclear coverage. We ran each of the three callers, on each of the ten samples, and identified a median of 28, 41 and 56 MT variants, respectively (Supp. Fig 1A). We manually inspected every variant and found LoFreq to be overly susceptible to systematic sequencing artefacts, and thus too sensitive. As expected, we found HaplotypeCaller to be insensitive to low VAF variants (Supp. Fig 1B). FreeBayes produced very few false positives and was selected as the *mity-call* variant caller.

mity-call: variant normalisation

As FreeBayes is a haplotype-aware variant caller, it can merge two nearby variants in *cis* into a longer multi-nucleotide variant (MNV). In practice, sequencing noise from high MT coverage can artificially inflate the number of MNVs and reduce the number of reads supporting a variant of interest (Supp. Figure 4). Existing variant decomposition methods including vt normalize (Tan et al., 2015) and vcflib (Garrison, 2016) do not decompose all of the INFO and FORMAT annotations of MNVs, which are critical components for the *mity-report*, and downstream analysis tools. We thus implemented *mity-normalise* to decompose and normalise all variants, as well as the variant metadata within the INFO and FORMAT fields in the VCF (Supp. Figure 4c).

***mity-call*: assessing variant quality**

We implemented a variant quality score, *q*, that is fast and VAF independent, similar to the challenge of identifying rare variants within tumours (Rheinbay et al., 2017). *q* is defined as the Phred-scaled probability of seeing at least the observed number of alternate reads by chance, given a noise threshold and assuming a binomial distribution. That is, given a noise threshold *p* and position *i*, and n_i alternate bases, from a total depth of N_i, the variant quality q_i is:

$$q_i = -10log_{10}(1-F(n_i|p,N_i)) ,$$

where *F* is the binomial cumulative distribution function. The noise level parameter *p* may vary for each dataset and so should be separately estimated for each dataset.

Estimating *p* in the 13 NA12878 replicates.

To estimate *p*, we used samtools mpileup (Li et al., 2009) to calculate the VAF of all three alternate bases at every position, in each of the 13 replicates of NA12878 with MQ>30 and BQ>24 (Supp. Figure 5a). There remained a persistent set of alternate reads, with VAF up to 0.001. Thus, taking a conservative approach, we set p=0.002 to reduce the impact of sequencing noise. At $q \geq 30$, we identified 18 variants in all 13 replicates of NA12878 (Supp. Figure 5b). However, there were 5 additional variants seen in a subset of replicates, all of which failed manual review.

Estimating *p* in the two replicates of the MD patient

In the same way as above, in two replicates of an MD patient, we estimated a higher noise floor of p=0.003 (Supp. Figure 6a) and found 59 variants in both replicates (Supp. Fig 6b), including a known pathogenic m.3243A>G variant (VAF=0.15% in both samples). There were one and three variants private to each replicate, of which two passed manual review, but which just failed the *q* threshold in the other sample.

As with any classification problem, when using *mity* there is a need to balance sensitivity with specificity. We have set the default threshold as $q \geq 30$, which favours sensitivity over specificity. In larger cohorts, the estimation of *p* could be fine-tuned by estimating a base or region-specific *p*, similar to the ideas presented by Gerstung and colleagues (Gerstung et al., 2014).

Nuclear insertions of mitochondrial origin (NUMTs)

There have been a number of reports of nuclear mitochondrial DNA (NUMT), which are fragments of the mitochondrial genome integrated into the nuclear genome (Lopez et al., 1994). The presence of NUMT can potentially confound heteroplasmy (Parr et al., 2006; Santibanez-Koref et al., 2019) by causing reads to align to the wrong genome, which would change the variant allele frequency, and thereby the estimation of variant heteroplasmy. Many known NUMT were integrated into the human genome over evolutionary timescales, allowing the accumulation of genetic changes to make it possible to resolve these sequences. To our knowledge, the rates of rare (patient- or family-specific) NUMT have not been systematically investigated and will remain challenging to resolve using short-read sequencing.

Here, we investigated whether any of the known NUMT could create false-positive MT SNV and INDEL variants caused by misalignment of known NUMT reads.

We have previously shown that when using the same sequencing technology (Illumina HiSeq X), and read aligner (BWA-MEM), that 150bp paired-end WGS can correctly align sequencing reads to the *PKD1* gene, despite *PKD1* having six pseudogenes with up-to 97.7% sequence homology (Mallawaarachchi et al., 2016). In a follow-up study of 145 clinical patients, no false positives due to sequencing read misalignment were detected using the same approach (Mallawaarachchi *et al, in review*). These results suggest that any NUMT with homology <97.7% should be readily resolved using 150bp paired-end sequencing reads. We further reasoned that short NUMT less than 300bp, i.e. less than the size of a typical read-pair of 300-450bp, would be easily resolved using existing read alignment. Furthermore, *mity* relies on even more stringent read mapping thresholds than reported above, of MQ>=30, meaning a 1:1000 chance of read misalignment.

Based on the RHNumtS.2 catalogue of NUMT (Ramos et al., 2011), just a single NUMT is longer than 300bp with sequence homology higher than 97.7%: HSA_NumtS_001. This NUMT is 5,839bp long and aligns to chr1:564,464-

570,303 (hg19) and chrM:3914-9755 with 98.55% homology. 98.55% homology represents 85 genetic differences over the whole length, or on average, 4.4 mismatches per 300bp read pair.

We performed pairwise sequence alignment of HSA_NumtS_001 and the corresponding MT sequence using BLASTN (Altschul et al., 1990). We reasoned that the genetic differences between the two would allow reads to be properly aligned (Supp Fig 7a), but when the distance between genetic changes approached the read length, that read alignment may become difficult to resolve (Supp Fig 7b). For every base in the alignment, the distance to the nearest mismatch was calculated (Supp Fig 7c) and only one stretch of 24 bases are more than 150 bp from a mismatch, (i.e., the length of a single-read), and none approached 300bp, the paired-end read-length. In this region of high homology, there is still another 150bp read in the read-pair that can be used to align the read pair appropriately

Furthermore, given an average nuclear sequencing depth of 30× and our stringent MQ threshold of 30 (1:1000 read misalignment) even in the unlikely scenario that all of the reads in a given genomic location misalign from the NUMT to the MT, this is highly unlikely to cause a false positive tier 1 variant. Taken together, we find it highly unlikely that given existing 150bp, paired-end reads from WGS, and high-quality read aligners, that NUMT:MT homology to the previously reported NUMT will be a significant source of false positive MT variants.

Supplementary References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Garrison, E., 2016. Vcflib, a simple C++ library for parsing and manipulating VCF files.

Garrison, E., Marth, G., 2012. Haplotype-based variant detection from short-read sequencing. ArXiv12073907 Q-Bio.

Gerstung, M., Papaemmanuil, E., Campbell, P.J., 2014. Subclonal variant calling with multiple samples and prior knowledge. Bioinformatics 30, 1198–1204. https://doi.org/10.1093/bioinformatics/btt750

Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv.org q-bio.GN. https://doi.org/10.6084/m9.figshare.963153

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Subgroup, 1000 Genome Project Data Processing, 2009. The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. 25, 2078–2079.

Lopez, J.V., Yuhki, N., Masuda, R., Modi, W., O'Brien, S.J., 1994. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J. Mol. Evol. 39, 174–190. https://doi.org/10.1007/bf00163806

Mallawaarachchi, A.C., Hort, Y., Cowley, M., McCabe, M.J., Minoche, A., Dinger, M.E., Shine, J., Furlong, T.J., 2016. Whole-genome sequencing overcomes pseudogene homology to diagnose autosomal dominant polycystic kidney disease. Eur. J. Hum. Genet. EJHG 24, 1584–1590. https://doi.org/10.1038/ejhg.2016.48

Parr, R.L., Maki, J., Reguly, B., Dakubo, G.D., Aguirre, A., Wittock, R., Robinson, K., Jakupciak, J.P., Thayer, R.E., 2006. The pseudo-mitochondrial genome influences mistakes in heteroplasmy interpretation. BMC Genomics 7, 185. https://doi.org/10.1186/1471-2164-7-185

Poplin, R., Ruano-Rubio, V., DePristo, M.A., Fennell, T.J., Carneiro, M.O., Auwera, G.A.V. der, Kling, D.E., Gauthier, L.D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M.J., Neale, B., MacArthur, D.G., Banks, E., 2018. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv 201178. https://doi.org/10.1101/201178

Quinlan, A.R., 2014. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinforma. 47, 11.12.1-11.12.34. https://doi.org/10.1002/0471250953.bi1112s47

Ramos, A., Barbena, E., Mateiu, L., del Mar González, M., Mairal, Q., Lima, M., Montiel, R., Aluja, M.P., Santos, C., 2011. Nuclear insertions of mitochondrial origin: Database updating and usefulness in cancer studies. Mitochondrion 11, 946–953. https://doi.org/10.1016/j.mito.2011.08.009

Rheinbay, E., Parasuraman, P., Grimsby, J., Tiao, G., Engreitz, J.M., Kim, J., Lawrence, M.S., Taylor-Weiner, A., Rodriguez-Cuevas, S., Rosenberg, M., Hess, J., Stewart, C., Maruvka, Y.E., Stojanov, P., Cortés, M.L., Seepo, S., Cibulskis, C., Tracy, A., Pugh, T.J., Lee, J., Zheng, Z., Ellisen, L.W., Iafrate, A.J., Boehm, J.S., Gabriel, S.B., Meyerson, M., Golub, T.R., Baselga, J., Hidalgo-Miranda, A., Shioda, T., Bernards, A., Lander, E.S., Getz, G., 2017. Recurrent and functional regulatory mutations in breast cancer. Nature 547, 55–60. https://doi.org/10.1038/nature22992

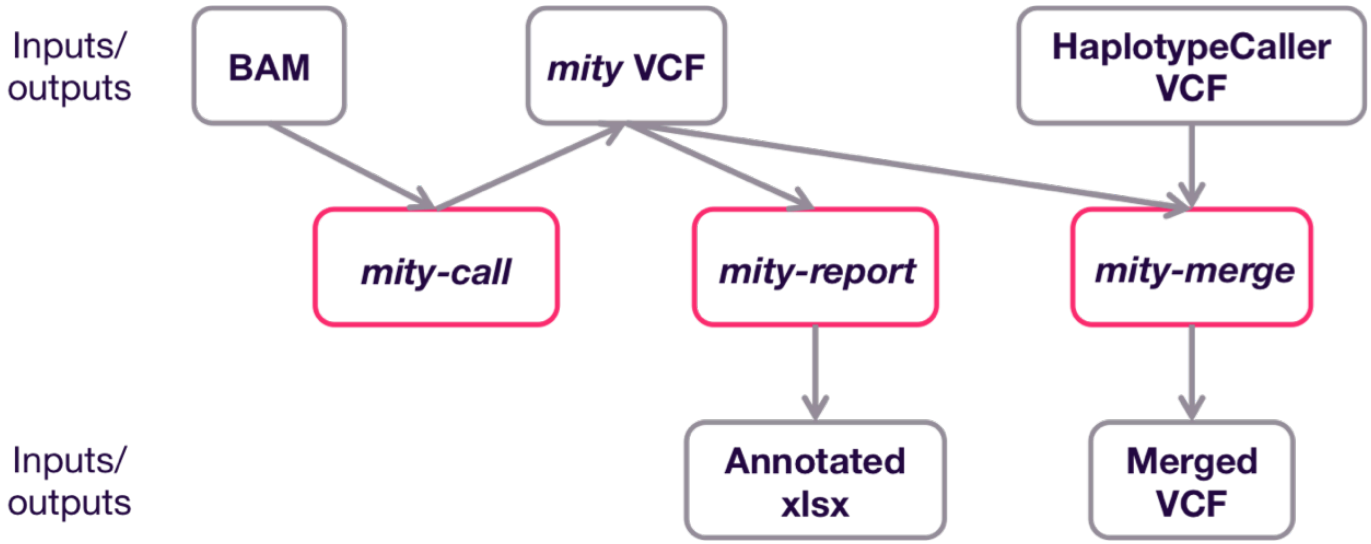
Santibanez-Koref, M., Griffin, H., Turnbull, D.M., Chinnery, P.F., Herbert, M., Hudson, G., 2019. Assessing mitochondrial heteroplasmy using next generation sequencing: A note of caution. Mitochondrion 46, 302–306. https://doi.org/10.1016/j.mito.2018.08.003

Tan, A., Abecasis, G.R., Kang, H.M., 2015. Unified representation of genetic variants. Bioinforma. Oxf. Engl. 31, 2202–2204. https://doi.org/10.1093/bioinformatics/btv112

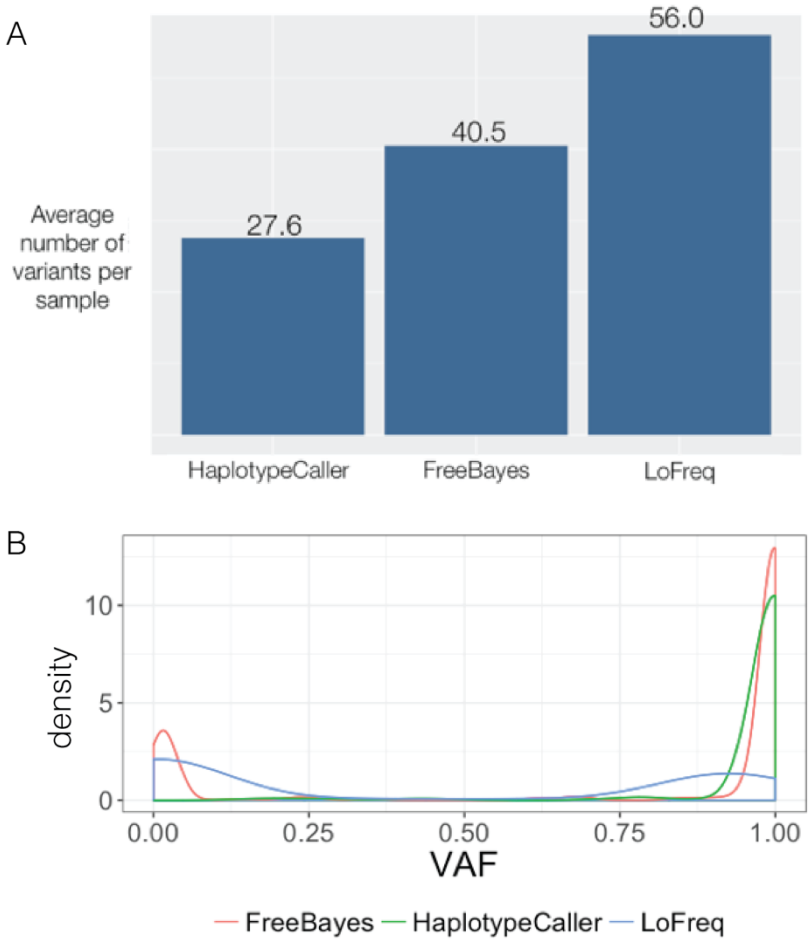
Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S., DePristo, M.A., 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline: The Genome Analysis Toolkit Best Practices Pipeline, in: Bateman, A., Pearson, W.R., Stein, L.D., Stormo, G.D., Yates, J.R. (Eds.), Current Protocols in Bioinformatics. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 11.10.1-11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

Wilm, A., Aw, P.P.K., Bertrand, D., Yeo, G.H.T., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L., Nagarajan, N., 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 40, 11189–11201. https://doi.org/10.1093/nar/gks918

6 Supplementary Figures

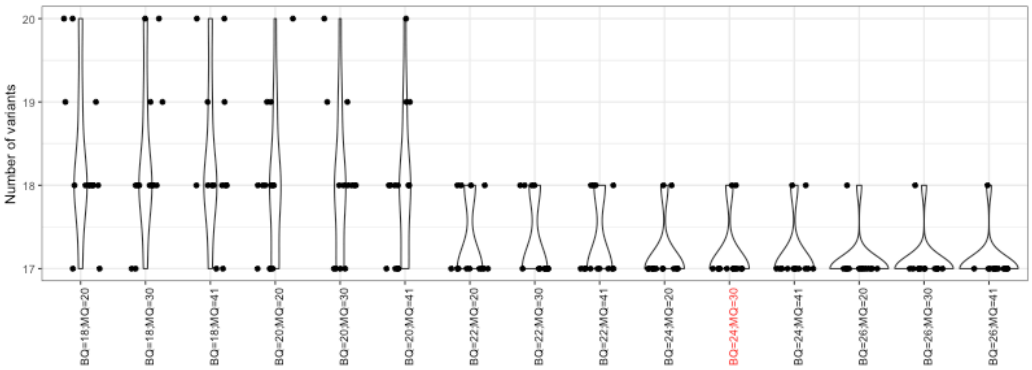


Supp Figure 1: The *mity* analysis pipeline. *mity* consists of three modules: *mity-call*, to call, filter and normalise variants in the MT genome; *mity-report*, to produce a clinician and researcher-friendly annotated MT variant report; *mity-merge*, to integrate *mity* analysis into nuclear analyses.

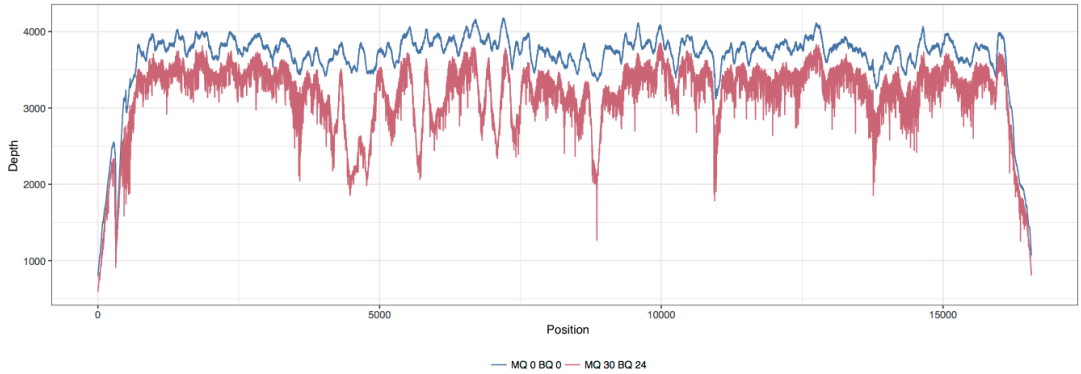


Supp Figure 2: The impact of variant callers on MT variant detection. A: Ten patients with MD were sequenced using WGS, and analysed by three different variant callers using default settings: GATK HaplotypeCaller, FreeBayes and LoFreq. The average number of variants found per sample is reported for each caller. **B**: A density plot of the VAF of the variants detected by HaplotypeCaller, FreeBayes and LoFreq. VAF: variant allele frequency.

A

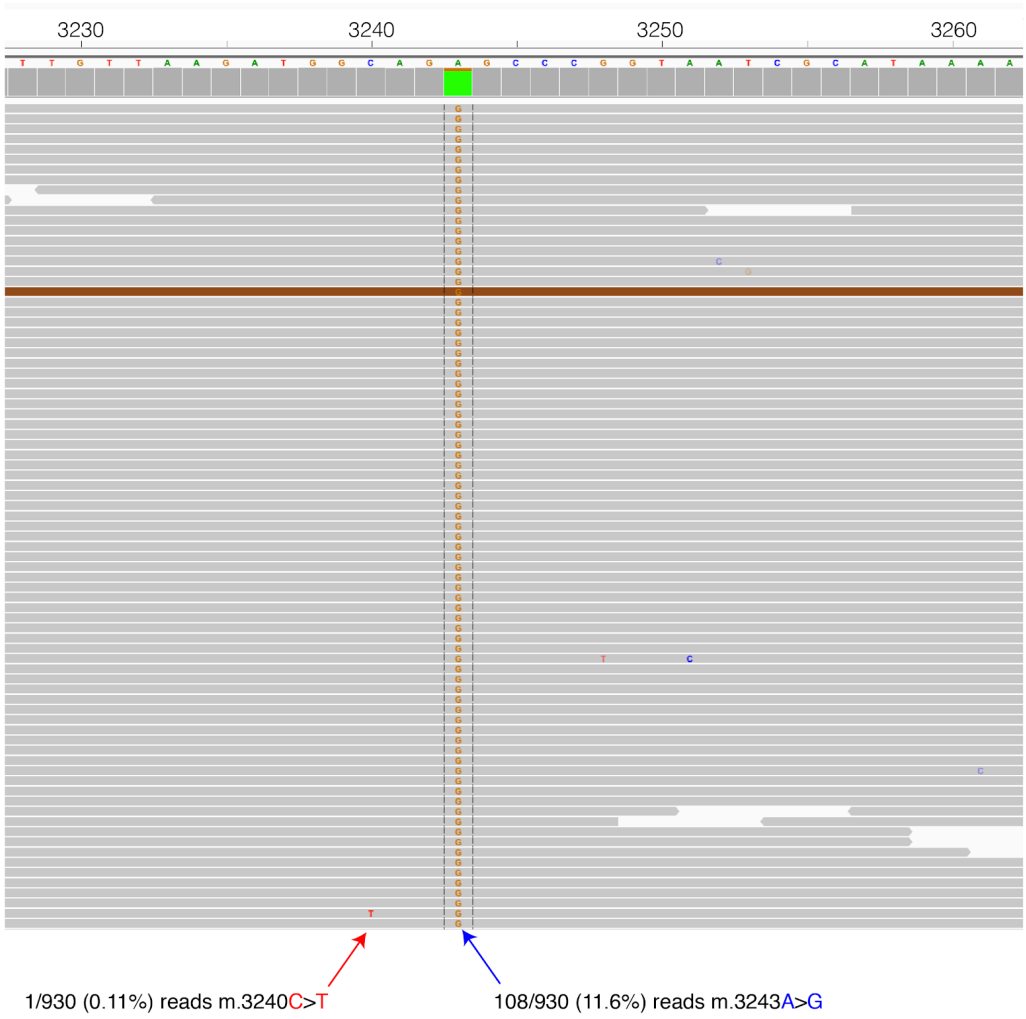


B



Supp Figure 3: Optimising the mapping (MQ) and base (BQ) quality filters. A: The number of variants with VAF>0.01 identified in 13 replicates of NA12878 across varying BQ and MQ combinations. The optimal combination of BQ≤24 and MQ≤30 is highlighted red. **B**: Filtering reads with MQ≤30 and BQ≤24 caused a modest reduction in coverage from 3,666× (MQ=0, BQ=0) to 3,166× (MQ≤30, BQ≤24) in a cohort of 2,980 healthy unrelated controls MGRB. VAF: variant allele frequency; MGRB: medical genomics reference bank.

A



B

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation count">
##FORMAT=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observation count">
#CHROM POS  ID REF  ALT       QUAL FILTER  INFO             FORMAT      B49
MT     3240 .  CAGA TAGG,CAGG .    .       TYPE=complex,snp GT:DP:RO:AO 0/1:930:822:1,107
```

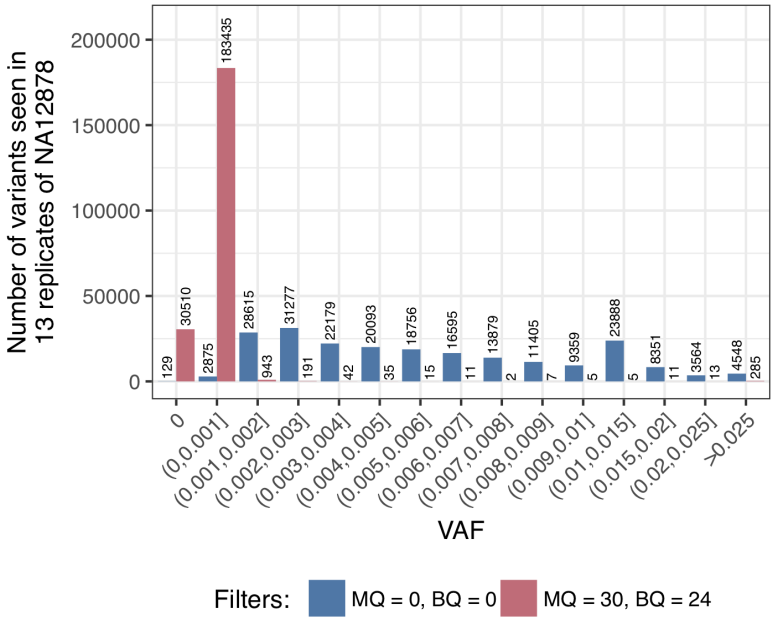
C

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=RO,Number=1,Type=Integer,Description="Reference allele observation count">
##FORMAT=<ID=AO,Number=A,Type=Integer,Description="Alternate allele observation count">
#CHROM POS  ID REF ALT QUAL FILTER  INFO     FORMAT      B49
MT     3240 .  C   T   .    FAIL    TYPE=snp GT:DP:RO:AO 0/1:930:929:1
MT     3243 .  A   G   .    PASS    TYPE=snp GT:DP:RO:AO 0/1:930:823:108
```

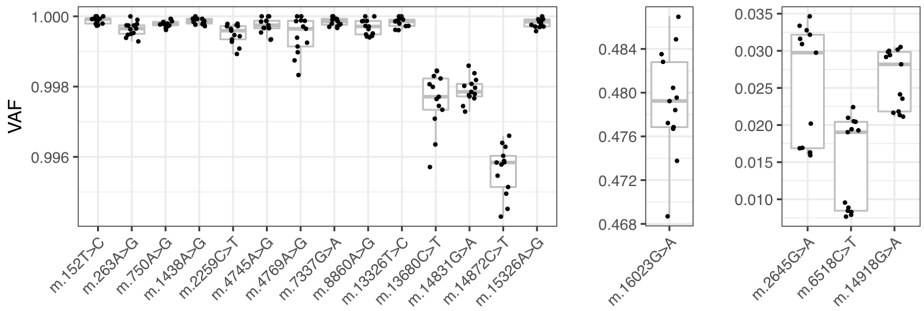
Supp Figure 4: Variant normalisation. A Raw sequencing reads from a patient with the pathogenic m.3243A>G variant at 11.6% VAF (blue arrow). A single read had an m.3240C>T variant, with BQ=30 (red arrow). **B** By default, FreeBayes will merge variants on the same haplotype, thus creating a multi-nucleotide polymorphism. Of the 930 total reads, 822 match the CAGA reference sequence, one matches the TAGG sequence, and 107 match the CAGG sequence. Most variant annotation tools, including VEP, which is used by *mity* would fail to annotate this as the well-known pathogenic m.3243A>G variant. **C** After *mity-normalise*, this multi-nucleotide variant is decomposed into the m.3240C>T variant with just one supporting read (red), and the pathogenic m.3243A>G variant with 108 supporting

reads (blue). Most variant annotation tools would now easily annotate the m.3243A>G variant as pathogenic. VAF: variant allele frequency, VEP: variant effect predictor, BQ: base quality.

A

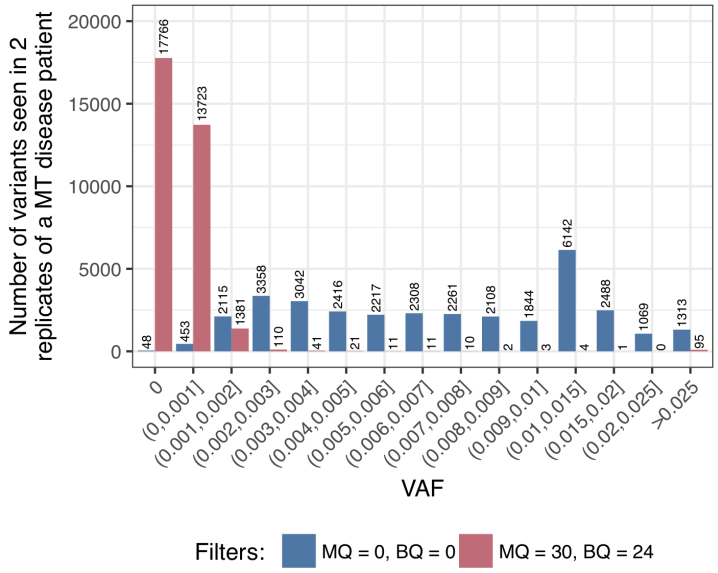


B

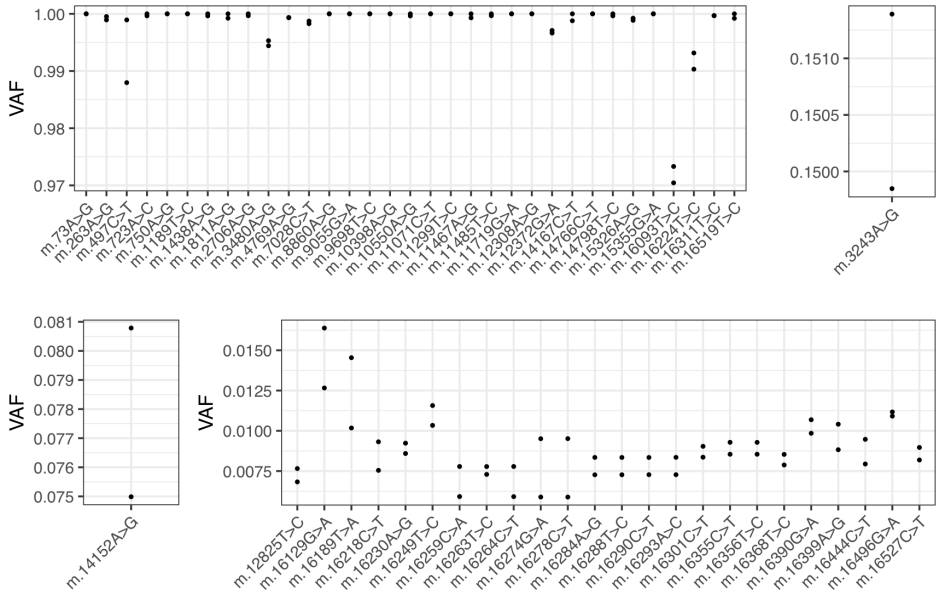


Supp Figure 5: Variant heteroplasmy is highly reproducible using control cell lines. *mity* was run on 13 replicates of NA12878, and we established a noise threshold *p* of 0.002. **A**: The VAF of all alternate alleles found in 13 replicates of NA12878. **B**: The VAF of all *mity* variants identified in all 13 replicates of NA12878, with p=0.002, $q \geq 30$, MQ≥30 and BQ≥24. VAF: variant allele frequency, MQ: mapping quality, BQ: base quality, *p*: noise threshold, *q*: Phred-scaled variant quality.

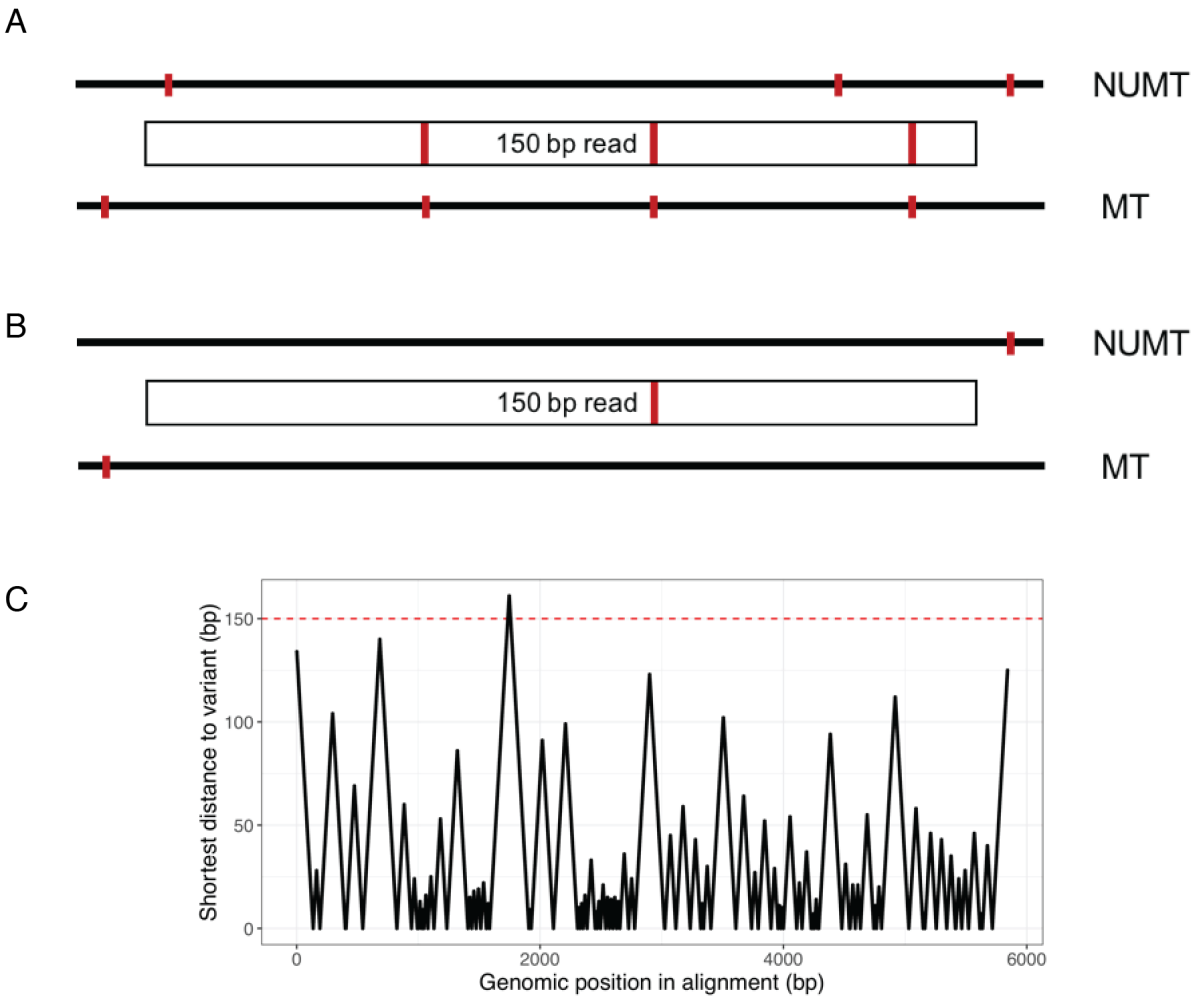
A



B



Supp Figure 6: **Variant heteroplasmy is highly reproducible using patient material.** *mity* was run on two replicates of the same patient with mitochondrial disease. **A**: The VAF of all alternate alleles found in the two replicates. From this, the noise threshold *p* was set to 0.003. **B**: The VAF of all *mity* variants identified in the two replicates, including the m.3243A>G pathogenic variant, with p=0.003, q≥30, MQ≥30 and BQ≥24. VAF: variant allele frequency, MQ: mapping quality, BQ: base quality, *p*: noise threshold, *q*: Phred-scaled variant quality.



Supp Figure 7: NUMT:MT homology is not expected to be a significant source of false-positive MT variants. Informed by our previous work, we reasoned that reads from any short NUMT (<300bp), or those from a NUMT:MT pair with <97.7% sequence homology, would be correctly aligned and that most NUMT would fall below this homology threshold. In the example shown in panel **A**, the sequence mismatches allow the read to be correctly aligned to the MT. In panel **B**, if the homology is too high, and the read cannot be unambiguously aligned to the MT. **C**: From the RHNumtS.2 catalogue of NUMT(Ramos et al., 2011), there is only one NUMT with length >300bp, and NUMT:MT homology >97%: HSA_NumtS_001. We measured the distance to the nearest mismatch at each position in the alignment of HSA_NumtS_001 and the MT genome using BLASTN. This reveals that there is only a single region of 24 bases where there is more than 150bp between variants. NUMT: nuclear mitochondrial DNA; MT: mitochondrial DNA.