bioRxiv preprint doi: https://doi.org/10.1101/007005; this version posted July 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RESEARCH

Heteroplasmy is ubiquitous and stable at the single cell level

Anitha D Jayaprakash[1], Erica Benson[1], Raymond Liang[1], Jaehee Shim[1], Luca Lambertini[1], Mike Wigler[2], Stuart A Aaronson[1] and Ravi Sachidanandam[1*]

*Correspondence:
ravi.mssm@gmail.com
[1]Department of Oncological Sciences, Icahn School of Medicine at Mount Sinai , One Gustave L. Levy place, New York, 10029, USA
Full list of author information is available at the end of the article

Abstract

Eukaryotic cells carry two genomes, nuclear (nDNA) and mitochondrial (mtDNA), which are ostensibly decoupled in their replication, segregation and inheritance. It is increasingly appreciated that heteroplasmy, the occurrence of multiple mtDNA haplotypes in a cell, plays an important biological role, but its features are not well understood. Until now, accurately determining the diversity of mtDNA has been difficult due to the relatively small amount of mtDNA in each cell ($< 1\%$ of the total DNA), the intercellular variability of mtDNA content and copies of mtDNA pseudogenes in nDNA. To understand the nature of heteroplasmy, we developed Mseek, a novel technique that purifies and sequences mtDNA. Mseek yields high purity ($> 98\%$) mtDNA and its ability to detect rare variants is limited only by sequencing depth, providing unprecedented sensitivity and specificity. Using Mseek, we confirmed the ubiquity of heteroplasmy by analyzing mtDNA from a diverse set of cell lines and human samples. Applying Mseek to colonies derived from single cells, we find heteroplasmy is stably maintained in individual daughter cells over multiple cell divisions. Our simulations suggest the stability of heteroplasmy is facilitated by the exchange of mtDNA between cells. We also explicitly demonstrate this exchange by co-culturing cell lines with distinct mtDNA haplotypes. Our results shed new light on the maintenance of heteroplasmy and provide a novel platform to investigate various features of heteroplasmy in normal and diseased tissues.

Keywords: mitochondria; mtDNA; heteroplasmy; genetics

Introduction

Mitochondria are organelles present in almost every eukaryotic cell [1]. They enable aerobic respiration[2] to efficiently generate ATP, and play an important role in oxygen sensing, inflammation, autophagy, and apoptosis[3, 4]. Mitochondrial activity relies on over a thousand proteins, mostly coded by the nuclear DNA in humans[5], but genes from the mitochondrial genome, a small circular DNA (**mtDNA**), play a critical role in their function. In humans, the mtDNA is ≈ 17 kbp and codes thirteen proteins critical for the electron transport chain, along with twenty-two tRNAs, two rRNAs and a control region, called the displacement loop (D-loop) (Fig. S1)[6]. Their genetic code differs from the nuclear code. In mammalian mitochondria, ATA codes for Methionine instead of Isoleucine, TGA codes for Tryptophan instead of the stop codon, and AGA, AGG code for stop codons instead of Arginine hinting at a bacterial origin[7]. Mitochondria are inherited solely from the mother and reproduce without recombination. Each mitochondrion carries multiple mitochondrial genomes$(5 - 10)$[8] and each cell contains hundreds to thousands of mitochondria,

bioRxiv preprint doi: https://doi.org/10.1101/007005; this version posted July 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

depending on the tissue[9]. Inherited mutations in mtDNA have been linked to several genetic disorders including *diabetes mellitus and deafness* (DAD) and *Leber's hereditary optic neuropathy* (LHON)[10]. De novo mutations in mtDNA have also been linked to diseases[11, 12, 13, 14].

Heteroplasmy, which is the occurence of multiples mtDNA haplotypes, has been documented in a variety of human tissues[15] and in samples from the 1000 genomes project[16]. Accurate determination of heteroplasmy, especially the low-frequency haplotypes, is needed for disease-association studies with mtDNA, as well as studies of metabolic activity of cancer cells[17]. Deep sequencing is the only means to identify novel mtDNA haplotypes as well as somatic mutations in tissues and perform association studies to link the haplotypes to disease states. However, measurements of heteroplasmy are compromised by copies of large segments of mtDNA, called Nuclear-mtDNA pseudogene sequences (**Numts**), present in the mammalian nDNA[18] (Fig. S2). Thus, accurate determination of heteroplasmy requires purification of mtDNA. Without purification, Numts contaminate the measurements of mtDNA variants, and introduce inaccuracies in the estimates of heteroplasmy, especially because Numts exhibit variability and occur in variable copy numbers similar to any other part of the nDNA. Isolating mtDNA has long been a challenge. In forensics and genealogy, allele-specific primer extensions (SNaPshot) are used for genotyping mtDNA[19]. Hyper variable regions(HVR) in the D-loop have been amplified using PCR[20]. Entire mtDNA has been accessed using primers specific to mtDNA to either perform long-range PCR[21], or amplify overlapping fragments[15]. Isolation of organelles by ultra-high-speed centrifugation has also been used, though the yields are low along with contamination from fragmented nuclear DNA[22]. Computational methods have also been used to infer heteroplasmy from whole-exome[23, 16] and whole-genome data[24],

Heteroplasmy derived from PCR-based methods are error-prone, due to variability in amplification. Errors also arise from clonal amplification of variants arising from mistakes of polymerases, a common problem in PCR-amplicon sequencing. Additionally, sequence and copy number variations of Numts confound results from computational and PCR-based methods in unpredictable ways. Thus, none of the methods outlined above are able to accurately identify low-frequency variants in mtDNA.

We present here Mseek, a novel method to enzymatically purify mtDNA by depleting linear nDNA and inexpensively sequencing it. By applying Mseek to several cell-lines and human peripheral blood mononuclear cells (PBMC), we identified multiple mtDNA haplotypes in the samples. A major benefit of this method is the ability to call extremely rare variants, with sensitivity of calls only limited by the sequencing depth. Sequencing errors can also be overcome with more sequencing, which is not always possible, especially in PCR amplicon sequencing. Additionally, through clonal expansion of single cells from a variety of cell lines, we establish that heteroplasmy is stably maintained at a single cell level through multiple divisions. This suggests active intercellular exchange of mtDNA. This exchange is explicitly demonstrated by co-culturing two different cell-lines with distinct mtDNA haplotypes, labeling one cell line with GFP, and sorting the cells after many generations (> 25) to show mtDNA haplotypes unique to one cell-line selectively appear in

bioRxiv preprint doi: https://doi.org/10.1101/007005; this version posted July 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the other. These results, in conjunction with simulations, suggest that exchange of mtDNA between cells is a source of renewal and stability.

Results

Mseek: An efficient method to isolate and sequence mtDNA

The potential relevance of mtDNA to many diseases requires a method to accurately determine the diversity of mtDNA in populations of cells. However, as noted, one of the major problem of existing approaches is the presence of nuclear DNA, which contains sequences of high homology to mtDNA (Numts), making it difficult to discern mtDNA from nDNA (Fig. S2). To address this issue, we sought to take advantage of the difference in topology between nDNA and mtDNA using an exonuclease to digest the linear nDNA, while leaving intact the circular mtDNA. Total DNA was extracted from HEK 293T cells, and digested with exonuclease V or left undigested. To determine the outcome, we PCR amplified sequences specific to nDNA or mtDNA using appropriate primers. As expected, in the undigested samples of total DNA we could detect both nDNA and mtDNA (Fig.1A). In sharp contrast, in the samples treated with exonuclease V we could only detect mtDNA (Fig. 1B). The lengths of the expected pcr products are shown in Fig. 1C. Using this approach, mtDNA was prepared and sequenced on the Illumina MiSeq platform. Out of a total of 3.05 million 100nt reads, 1.233 million mapped to the mitochondrial genome and 50,000 ($< 2\%$) mapped to the nDNA. The remainder were adapter dimers, which are sequencing artifacts currently filtered out experimentally using Ampure beads. Over 98% of the mappable reads were derived from mtDNA with an average coverage $>$ 3000X (Fig. 1D). More than 50 distinct samples were processed similarly to consistently obtain high purity mtDNA sequence.

The error rate per base of the reads is approximaely 1 in 1000 (Q score $>$ 30). Using at least 10 non-clonal reads to make a variant call reduces errors from sequencing to much less than 1 in a million. This coverage also allows removal of variants with a significant bias towards one strand, a known source of errors on the Illumina platform[25]. Contamination from the small amount of nDNA left in the samples does not contribute appreciably to the noise as Numts are a small fraction of total nDNA. Thus, calling rare variants to any level of sensitivity only depends on the depth of sequencing. This approach, designated Mseek (Fig. 2), provides a means of unmatched efficiency in accurately sequencing the mtDNA contained within a population of cells

Ubiquity of heteroplasmy

Since cell lines are clonally derived, the expectation is that the nDNA (and mtDNA) are identical across cells. We decided to explore the diversity of mtDNA in a variety of cell lines to test the expectation that mtDNA would be homoplasmic in cell lines, since, either a slight fitness advantage of one haplotype or drift[26] would lead to a clonal selection and homoplasmy. We applied Mseek to thirty samples including four human PBMCs and human cell lines derived from human diploid fibroblasts (501T), glioma (A382) and breast carcinoma (HCC1806 and MDA-MB-157).

The mtDNA sequences were analyzed for variations in order to infer universal features in mtDNA variability and differences between human cell lines and blood-derived mtDNA. Repeat content of the sequences was computationally identified to

bioRxiv preprint doi: https://doi.org/10.1101/007005; this version posted July 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

estimate nDNA contamination, which ranged from $0.5 - 1.5\%$; further confirming the specificity of Mseek. Importantly, because of this high degree of mtDNA purity ($> 98\%$) we were able to multiplex all 30 samples in a single MiSeq run, with average coverage of > 100X.

Variants with a frequency of either 0 or 1 in the population arise from homoplasmic mtDNA. Intermediate frequencies between 0 and 1 imply the co-existence of multiple haplotypes in the population. Strikingly, in both cell-lines and human blood-derived mtDNA, we observed variants occurring in the $0.1 - 0.9$ frequency range (Fig. 3), indicating that multiple haplotypes were present in the samples. The tool *Mutation Assessor*[27] was used to label the variants as *high*, *medium*, *low*, or *neutral* signifying their predicted impact on protein function. Cell-lines and human PBMCs did not exhibit putative deleterious mutations at high frequency, consistent with the expectation that functioning cells should have functional mitochondria.

The mtDNA has a few non-coding regions outside of the D-loop which occur as gaps between genes. None of the samples exhibited mutations in these regions, suggesting an evolutionarily conserved role, such as in transcriptional control, for these regions. Each sample had unique, distinguishing mutations, ranging in frequency from 0.36 to 1.0. There were a number of unique variants in the four human PBMC samples (ranging in number from 5 to 15) and in the cell lines (ranging in number from 5 to 21).

Since the cell lines were derived from a variety of tissues, our findings have some level of universality. There were no key distinguishing features between cell-line and human blood-derived mtDNA, in terms of deleterious mutations or degree of heteroplasmy, contrary to findings from a study based on whole-genome sequencing of TCGA samples[24]. Our findings are consistent with another study based on colorectal cancer[15].

Stability of heteroplasmy in cell-lines

The results above indicate heteroplasmy exists within a cell population but do not establish heteroplasmy in individual cells, since a mixture of homoplasmic cells with different haplotypes would give the same result. In order to establish heteroplasmy in individual cells, we placed the severest possible bottleneck on the population by deriving colonies from single cells, utilizing MDA-MB-157 and U20S breast carcinoma and osteosarcoma lines respectively (Fig. 4). In each of the derived colonies (8 colonies), the variants from the original lines remained in the derived colonies and at approximately the same frequencies as in the original tumor lines. The sharing of mutations between the original and derived colonies suggests that the diversity in mtDNA exists in individual cells. The preservation of the frequencies between the original and derived colonies indicates further that this heteroplasmy is uniform across cells in the original line (Fig. 4). Since the new clonal lines underwent at least 25 divisions from the single-cell stage, these results also suggest that heteroplasmy is stably maintained over multiple generations with no signs of selection or drift. Over many divisions, errors in replication should have increased diversity in heteroplasmy, while small differences in fitness and drift should lead to homoplasmy. In fact, drift has been proposed as a mechanism for the selection of homoplasmic mtDNA mutations in tumors[26], which has been corroborated in other studies[15]. In light of these reports, our findings are quite unexpected.

bioRxiv preprint doi: https://doi.org/10.1101/007005; this version posted July 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

A simple model of mtDNA genetics assumes random assortment of mtDNA haplotypes between daughter cells upon cell division, along with multiplication of mitochondria. This model would predict drift towards homoplasmy, as seen in our simulation of this process (Fig. 5) and by others[26]. The rate of drift in haplotype frequencies is a function of the number of mtDNA molecules per cell and the original frequency of the haplotypes (Fig. 5). After many passages, irrespective of the original mtDNA distribution, the likelihood of two randomly selected cells having the same heteroplasmic mix would be extremely low, which is at odds with the stable and uniform heteroplasmy that we observed in the clonally-derived cell-lines. This suggests the existence of an active mechanism to counteract this drift.

Exchange of mtDNA between cells within a population is the simplest explanation for the uniformity of heteroplasmy and its stability. Exchange can counteract the effects of drift by bringing the haplotype distribution closer to the average of the distribution across cells within the population. Other explanations, such as a balancing selection[28] could also be invoked to explain the lack of drift. This can be discounted because most variants are neutral and specific to each cell line, suggesting the selection needs to be different for each cell line without an obvious selective pressure.

Experimental demonstration of mtDNA exchange between cells

In order to explicitly demonstrate the exchange of mtDNA between cells, we co-cultured cell-lines with distinct private haplotypes. Two sets of pairs including MDA-MB-157 and HCC as well as A382 and U20S were used. For each pair, one of the cell-lines was labeled with GFP (by transfection with a vector expressing GFP). After approximately 20 passages, the cells were sorted for the GFP marker by FACS, and mtDNA from the sorted cells were sequenced. The sorted cells were greater than 99% pure based on FACS.

Tables 1 and 2 shows the results of sequencing mtDNA from these co-culture experiments. We detected variants private to one cell-line in the co-cultured partner cell-line, suggesting the transfer of mtDNA between the cell-lines. Not every private variant was transferred, arguing against the results arising from errors in sorting or cytoplasmic/nuclear exchange between cells. The purity of the sorted cells, based on FACS, further suggests that nuclear exchange does not account for the findings.

Discussion

Accurate sequencing of mtDNA is important for sensitive measurements of heteroplasmy, whose variability can have clinical significance, as a biomarker and in disease progression[29]. Mseek provides a means to purify and deeply sequence mtDNA and determine heteroplasmy accurately by eliminating Numts and PCR-related biases. The sensitivity of Mseek is a function of sequencing depth alone. This is one of the most detailed and extensive survey of mtDNA from cell lines yet obtained.

Accurate identification of variant frequencies is not possible through deep sequencing methods currently in use. So far, deep sequencing approaches to mtDNA have used either long-range PCR[21, 30], or a multitude of mtDNA-specific PCR primers to amplify short overlapping mtDNA fragments ($\approx$ 650 nt) which are ligated to each other, fragmented and prepared for sequencing [15]. Mining of whole-genome[24, 31] and whole-exome[23, 16] data has also been used to identify mtDNA

bioRxiv preprint doi: https://doi.org/10.1101/007005; this version posted July 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

fragments. A new approach uses methyl-specific endonucleases MspJI and AbaSI to deplete nDNA that is likely to be methylated[32]. A failing here is the Numts are not always methylated. PCR-amplicon biases, the inability to identify polymerase errors and contamination from Numts call into question the sensitivity of these methods to low frequency variants. The ability of Numts to confound analyses is highlighted by a study that used whole-genome data from the TCGA and inferred that deleterious mtDNA mutations are more common in cancer cells compared to normal tissue[24]. In contrast, findings of low mutations rates in tumor mtDNA from a colorectal cancer study[15] are more in line with our findings that cell lines don't exhibit higher rates of deleterious mutations compared to normal cells from human tissues.

We have shown here that cells from a wide-range of cell lines and human samples exhibit heteroplasmy, in accord with results from several studies[15, 16]. This suggests that heteroplasmy might be an essential feature of mtDNA. In fact, heteroplasmy seems to provide a *fingerprint* that can identify cells. A larger survey is needed to understand the resolution of this fingerprint and its ability to distinguish cellular origins. We found that mtDNA from transformed human cell-lines and primary human lymphocytes are similar with respect to the distributions of densities and frequencies of mutations (benign and deleterious ones). Non-coding gaps between mtDNA genes are highly conserved, indicating they might be control elements.

Clonal amplification of cells does not lead to a selection of particular mtDNA haplotypes, in fact, heteroplasmy is very stable, at least over the 25 or so divisions of cell lines that we have studied. This stability of heteroplasmy in cell lines is surprising in light of 1) the higher rates of mutation in mtDNA[33] which should increase the diversity of mtDNA, and 2) drift, which should lead to homoplasmy in about 70 generations[26, 15]. The stability of heteroplasmy against drift could arise from exchanges of mtDNA between cells which can be inferred from our cell-line data (Fig. 4) in conjunction with simulations (Fig. 5) and co-culturing experiments (Tables 1 and 2). The transfer of mtDNA seems to occur in a selective manner, suggesting either there are incompatibilities between the mtDNA haplotypes or between certain haplotypes and the nuclear genome. There is some indication that the amount of transfer increases over the number of passages of co-culture, based on our limited set of experiments, establishing this definitively requires a more long-standing experiment with sampling at different time points. A co-evolution of mtDNA and nDNA has in fact been suggested earlier[33]. This is also consistent with a study in mice that suggests that mitochondria from different species cannot co-exist[34]. The selective advantage of certain mtDNA haplotypes can additionally contribute to the stability of the mtDNA.

The exact mechanisms of mtDNA transfer are not known. Horizontal transfer of genetic material between species of yeasts has been shown[35] and there is increasing interest in organelle transfer between cells through microtubule formation[36]. Within a cell, networks of mitochondria are created through fusion, mediated by *fusin*, which leads to the exchange of mtDNA[37]. This is necessary for functional mitochondria; knocking out *fusin* causes muscles to atrophy through the accumulation of deleterious mutations[37]. *In vivo*, exchanges of mitochondria between cells

bioRxiv preprint doi: https://doi.org/10.1101/007005; this version posted July 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

has also been demonstrated in the rejuvenation of cells with damaged mitochondria by transfer of functional mitochondria from mesenchymal stem cells[38]. Rejuvenation of cells containing damaged mtDNA by transfer of functional mtDNA from neighboring cells in culture has also been observed[39]. Ours is the first demonstration of mtDNA transfer between cells with functional mtDNA.

This is the first explicit demonstration of mtDNA transfer between cells with functional mtDNA whereas previous studies have shown transfer from cells with functional mtDNA into ones with non-functioning mtDNA[39, 38]. The proposed exchange of mtDNA between cells can explain its stability over the lifetime of an organism, and over generations, inferred from the relative lack of major age-related disorders originating in the mtDNA and the ability to infer geographic origins of a person from the mtDNA sequence. The stability of mtDNA against deleterious mutations could also be enhanced by a coupling between replication and transcription[40], ensuring the depletion of non-functional mtDNA by inefficiencies in their replication.

By making mtDNA sequencing economical, Mseek enables large-scale studies of heteroplasmy for GWAS applications and clinical monitoring of mtDNA in tissues. The sequencing of mtDNA in cell-lines allows us, for the first time, to understand the nature of mtDNA variability and its maintenance in cell populations. There is great value in surveying large populations in order to establish the normal range of heteroplasmy for use in GWAS studies. The transfer of functional mtDNA into diseased cells could be used as therapy to treat disorders arising from mtDNA defects. Somatic mutations in mtDNA could play a role in various human disorders and in aging, especially when the transfer between cells is impeded and mechanisms involved in mtDNA transfer might be fruitful targets for therapeutic intervention.

Methods

Mseek

We have developed a new method of isolating and sequencing mtDNA (Fig. 2). The results section contains details of its performance. Briefly, the method consists of the following steps, total DNA is isolated from the sample. The nDNA is digested using Exonuclease V. The products are purified using Ampure beads to remove short fragments. Using PCR primers specific to mtDNA and nDNA, the purity of the treated samples is tested (Fig. 1B). Following this, the sample is fragmented using Covaris and end-repaired. Barcoded adapters compatible with the sequencing platform are ligated to the fragments. The universal adapters are used to amplify the library and prepare it for deep sequencing.

Cell Culture

mtDNA was isolated and sequenced from several cell lines including, 293T (a kidney-cancer derived cell line), U2OS and Saos-2 (human osteosarcoma cell lines) and MDA-MB-157 (metastatic human breast cancer cell line).

All cells were grown in Dulbecco's modified Eagle's medium (DMEM; Invitrogen), 10% heat-inactivated fetal bovine serum (FBS; Invitrogen) and 50 U/ml penicillin and streptomycin (Pen/Strep; Invitrogen). Cultures were maintained at 37° C in 5% CO2.

bioRxiv preprint doi: https://doi.org/10.1101/007005; this version posted July 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Clonal isolation of tumor cells was performed by serial dilution into 96-well plates and visual examination of wells for single cells, which were then expanded for an additional 28-30 population doublings.

Analyses

Sequences that map to repeat elements (which occur only in the nDNA) allow reliable estimation of the level of nDNA contamination, which ranged from $0.5 - 1.5\%$.

MiST[25], a variant detection tool for whole-exome data, was used to call mtDNA variants. The reference mitochondrial genome has the accession NC_012920 from Genbank. The mtDNA annotations are from MITOMAP13, and SNP annotations are from dbSNP14. The error rate in Miseq and Hiseq reads are approximately 1 in a 1000, so requiring at least 3 non-clonal reads to have the variant to make the call, reduces the error rate to well under 1 in a million. Variants with reads predominantly in one strand are excluded to firther reduce errors, based on our previous experience[25].

We developed a pipeline to assemble the mitochondrial genome from the deep-sequencing data, to demonstrate that the reads assemble into a circle and no large deletions, duplications or other large-scale structures were detected.

Mutation Assessor[27] was used to assess the impact of mtDNA mutations on protein function. This tool uses conservation of structure across orthologues to identify mutations in the DNA (and consequent changes in amino-acids) with potentially deleterious effects. The mutations are rated *high*, *medium*, *low*, or *neutral* based on their impact on protein function. We highlight the *high* and *medium* impact mutations in our graphs, as they may affect mitochondrial function.

Competing interests

Author's contributions

AJ and RS designed Mseek and designed several experiments. SA and MW designed several experiments with cell-lines. LL provided human samples and suggested applications. AJ, JS, RL performed Mseek, EB performed cell-line work. AJ and RS wrote the paper.

Acknowledgements

RS, AJ, LL were partially supported by the grant 1R21HG007394-01 from the NIH. Some experiments were also funded by a pilot grant from the Venture Capital Research Funding Program of Children's Environmental Health Center (CEHC) at Mount Sinai. Brian Brown helped craft the message. Comments from Viviana Simon, and Benjamin tenOever added clarity. Avinash Waghray, Sunniva Bjorklund and Vessela Kristensen caught numerous errors and gave many suggestions to improve the writing.

Author details

[1]Department of Oncological Sciences, Icahn School of Medicine at Mount Sinai , One Gustave L. Levy place, New York, 10029, USA. [2]Department of Public Health, Icahn School of Medicine at Mount Sinai, One Gustave L. Levy place, New York, USA. [3]Cold Spring Harbor Laboratory, One Bungtown Road, Cold Spring Harbor, USA.

bioRxiv preprint doi: https://doi.org/10.1101/007005; this version posted July 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

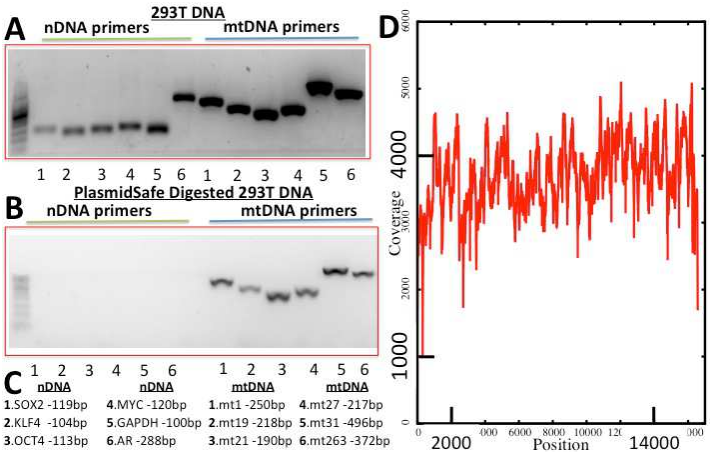
1. Wallace, D.C.: Structure and evolution of organelle genomes. Microbiological Reviews **46**(2), 208–240 (1982). PMID: 6750346
2. Wallace, D.C.: Mitochondrial diseases in man and mouse. Science (New York, N.Y.) **283**(5407), 1482–1488 (1999). PMID: 10066162
3. Liu, X., Kim, C.N., Yang, J., Jemmerson, R., Wang, X.: Induction of apoptotic program in cell-free extracts: requirement for dATP and cytochrome c. Cell **86**(1), 147–157 (1996). PMID: 8689682
4. Newmeyer, D.D., Ferguson-Miller, S.: Mitochondria: releasing power for life and unleashing the machineries of death. Cell **112**(4), 481–490 (2003). PMID: 12600312
5. Westermann, B., Neupert, W.: 'Omics' of the mitochondrion. Nature biotechnology **21**(3), 239–240 (2003). PMID: 12610566
6. Brandon, M.C., Lott, M.T., Nguyen, K.C., Spolim, S., Navathe, S.B., Baldi, P., Wallace, D.C.: MITOMAP: a human mitochondrial genome database–2004 update. Nucleic Acids Research **33**(Database issue), 611–613 (2005). PMID: 15608272
7. Jukes, T.H., Osawa, S.: The genetic code in mitochondria and chloroplasts. Experientia **46**(11-12), 1117–1126 (1990). Accessed 2014-07-08
8. Legros, F., Malka, F., Frachon, P., Lombes, A., Rojo, M.: Organization and dynamics of human mitochondrial DNA. Journal of cell science **117**(Pt 13), 2653–2662 (2004)
9. Fernandez-Vizarra, E., Enriquez, J.A., Perez-Martos, A., Montoya, J., Fernandez-Silva, P.: Tissue-specific differences in mitochondrial activity and biogenesis. Mitochondrion **11**(1), 207–213 (2011). Accessed 2013-04-25
10. Wallace, D.C.: Mitochondrial DNA mutations in diseases of energy metabolism. Journal of bioenergetics and biomembranes **26**(3), 241–250 (1994). PMID: 8077179
11. DiMauro, S., Schon, E.A.: Mitochondrial DNA mutations in human disease. American journal of medical genetics **106**(1), 18–26 (2001). PMID: 11579421
12. Lebon, S., Chol, M., Benit, P., Mugnier, C., Chretien, D., Giurgea, I., Kern, I., Girardin, E., Hertz-Pannier, L., de Lonlay, P., Rötig, A., Rustin, P., Munnich, A.: Recurrent de novo mitochondrial DNA mutations in respiratory chain deficiency. Journal of Medical Genetics **40**(12), 896–899 (2003)
13. Taylor, R.W., Turnbull, D.M.: Mitochondrial DNA mutations in human disease. Nature reviews. Genetics **6**(5), 389–402 (2005). PMID: 15861210
14. Greaves, L.C., Taylor, R.W.: Mitochondrial DNA mutations in human disease. IUBMB life **58**(3), 143–151 (2006). PMID: 16766381
15. He, Y., Wu, J., Dressman, D.C., Iacobuzio-Donahue, C., Markowitz, S.D., Velculescu, V.E., Diaz, J. Luis A, Kinzler, K.W., Vogelstein, B., Papadopoulos, N.: Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature **464**(7288), 610–614 (2010). PMID: 20200521
16. Diroma, M.A., Calabrese, C., Simone, D., Santorsola, M., Calabrese, F.M., Gasparre, G., Attimonelli, M.: Extraction and annotation of human mitochondrial genomes from 1000 genomes whole exome sequencing data. BMC Genomics **15**(Suppl 3), 2 (2014). Accessed 2014-07-07
17. Birsoy, K., Possemato, R., Lorbeer, F.K., Bayraktar, E.C., Thiru, P., Yucel, B., Wang, T., Chen, W.W., Clish, C.B., Sabatini, D.M.: Metabolic determinants of cancer cell sensitivity to glucose limitation and biguanides. Nature **advance online publication** (2014). Accessed 2014-07-22
18. Woischnik, M., Moraes, C.T.: Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. Genome Research **12**(6), 885–893 (2002). PMID: 12045142 PMCID: PMC1383742. Accessed 2013-04-25
19. Vallone, P.M.: Capillary electrophoresis of an 11-plex mtDNA coding region SNP single base extension assay for discrimination of the most common caucasian HV1/HV2 mitotype. Methods in molecular biology (Clifton, N.J.) **830**, 159–167 (2012). PMID: 22139659
20. Paull, D., Emmanuele, V., Weiss, K.A., Treff, N., Stewart, L., Hua, H., Zimmer, M., Kahler, D.J., Goland, R.S., Noggle, S.A., Prosser, R., Hirano, M., Sauer, M.V., Egli, D.: Nuclear genome transfer in human oocytes eliminates mitochondrial DNA variants. Nature **advance online publication** (2012). Accessed 2013-04-25
21. Maricic, T., Whitten, M., Paabo, S.: Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. PLoS ONE **5**(11), 14004 (2010). Accessed 2013-04-25
22. Ameur, A., Stewart, J.B., Freyer, C., Hagstrom, E., Ingman, M., Larsson, N.-G., Gyllensten, U.: Ultra-deep sequencing of mouse mitochondrial DNA: mutational patterns and their origins. PLoS Genet **7**(3), 1002028 (2011). Accessed 2013-02-19
23. Picardi, E., Pesole, G.: Mitochondrial genomes gleaned from human whole-exome sequencing. Nature methods **9**(6), 523–524 (2012). PMID: 22669646
24. Larman, T.C., DePalma, S.R., Hadjipanayis, A.G., Protopopov, A., Zhang, J., Gabriel, S.B., Chin, L., Seidman, C.E., Kucherlapati, R., Seidman, J.G.: Spectrum of somatic mitochondrial mutations in five cancers. Proceedings of the National Academy of Sciences **109**(35), 14087–14091 (2012). Accessed 2013-02-19
25. Subramanian, S., Pierro, V.D., Shah, H., Jayaprakash, A.D., Weisberger, I., Shim, J., George, A., Gelb, B.D., Sachidanandam, R.: MiST: a new approach to variant detection in deep sequencing datasets. Nucleic Acids Research (2013). PMID: 23828039. Accessed 2013-07-05
26. Coller, H.A., Khrapko, K., Bodyak, N.D., Nekhaeva, E., Herrero-Jimenez, P., Thilly, W.G.: High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection. Nature Genetics **28**(2), 147–150 (2001)
27. Reva, B., Antipin, Y., Sander, C.: Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Research (2011). Accessed 2013-02-19
28. Connallon, T., Clark, A.G.: Balancing selection in species with separate sexes: Insights from fisher's geometric model. Genetics (2014)
29. Wallace, D.C., Chalkia, D.: Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and

bioRxiv preprint doi: https://doi.org/10.1101/007005; this version posted July 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

disease. Cold Spring Harbor Perspectives in Biology **5**(11), 021220 (2013)
30. Sosa, M.X., Sivakumar, I.K.A., Maragh, S., Veeramachaneni, V., Hariharan, R., Parulekar, M., Fredrikson, K.M., Harkins, T.T., Lin, J., Feldman, A.B., Tata, P., Ehret, G.B., Chakravarti, A.: Next-generation sequencing of human mitochondrial reference genomes uncovers high heteroplasmy frequency. PLoS computational biology **8**(10), 1002737 (2012). PMID: 23133345
31. Chu, H.-T., Hsiao, W.W.L., Tsao, T.T.H., Chang, C.-M., Liu, Y.-W., Fan, C.-C., Lin, H., Chang, H.-H., Yeh, T.-J., Chen, J.-C., Huang, D.-M., Chen, C.-C., Kao, C.-Y.: Quantitative assessment of mitochondrial DNA copies from whole genome sequencing. BMC genomics **13 Suppl 7**, 5 (2012). PMID: 23282223
32. McKernan, K.J., Spangler, J., Zhang, L., Tadigotla, V., McLaughlin, S., Warner, J., Zare, A., Boles, R.G.: Expanded genetic codes in next generation sequencing enable decontamination and mitochondrial enrichment. PLoS ONE **9**(5), 96492 (2014). Accessed 2014-06-19
33. Dowling, D.K., Friberg, U., Lindell, J.: Evolutionary implications of non-neutral mitochondrial genetic variation. Trends in Ecology & Evolution **23**(10), 546–554 (2008). Accessed 2014-07-07
34. Sharpley, M.S., Marciniak, C., Eckel-Mahan, K., McManus, M., Crimi, M., Waymire, K., Lin, C.S., Masubuchi, S., Friend, N., Koike, M., Chalkia, D., MacGregor, G., Sassone-Corsi, P., Wallace, D.C.: Heteroplasmy of mouse mtDNA is genetically unstable and results in altered behavior and cognition. Cell **151**(2), 333–343 (2012). Accessed 2013-02-19
35. Marinoni, G., Manuel, M., Petersen, R.F., Hvidtfeldt, J., Sulo, P., Piškur, J.: Horizontal transfer of genetic material amongSaccharomyces yeasts. Journal of Bacteriology **181**(20), 6488–6496 (1999). PMID: 10515941. Accessed 2014-01-04
36. Rogers, R.S., Bhattacharya, J.: When cells become organelle donors. Physiology (Bethesda, Md.) **28**(6), 414–422 (2013). PMID: 24186936
37. Chen, H., Vermulst, M., Wang, Y.E., Chomyn, A., Prolla, T.A., McCaffery, J.M., Chan, D.C.: Mitochondrial fusion is required for mtDNA stability in skeletal muscle and tolerance of mtDNA mutations. Cell **141**(2), 280–289 (2010). PMID: 20403324
38. Islam, M.N., Das, S.R., Emin, M.T., Wei, M., Sun, L., Westphalen, K., Rowlands, D.J., Quadri, S.K., Bhattacharya, S., Bhattacharya, J.: Mitochondrial transfer from bone-marrow-derived stromal cells to pulmonary alveoli protects against acute lung injury. Nature medicine **18**(5), 759–765 (2012). PMID: 22504485
39. Spees, J.L., Olson, S.D., Whitney, M.J., Prockop, D.J.: Mitochondrial transfer between cells can rescue aerobic respiration. Proceedings of the National Academy of Sciences of the United States of America **103**(5), 1283–1288 (2006). Accessed 2014-07-08
40. Clayton, D.A.: Transcription and replication of mitochondrial DNA. Human reproduction (Oxford, England) **15 Suppl 2**, 11–17 (2000). PMID: 11041509

Figures

Figure 1 Performance of Mseek. (**A**) PCR products run on a 2% Agarose gel using primers for 6 nDNA genes (OCT4, MUC, KLF4, SOX2, GAPDH and AR) and 5 regions of mtDNA. before exonuclease digestion. (**B**) After digestion, the nDNA bands disappear. (**C**) Sizes of expected PCR products. (**D**) Deep sequencing, read depth (y-axis) versus position on mtDNA (x-axis). 1.23 million mtDNA reads and 50,000 nDNA reads implying $> 98\%$ pure mtDNA.

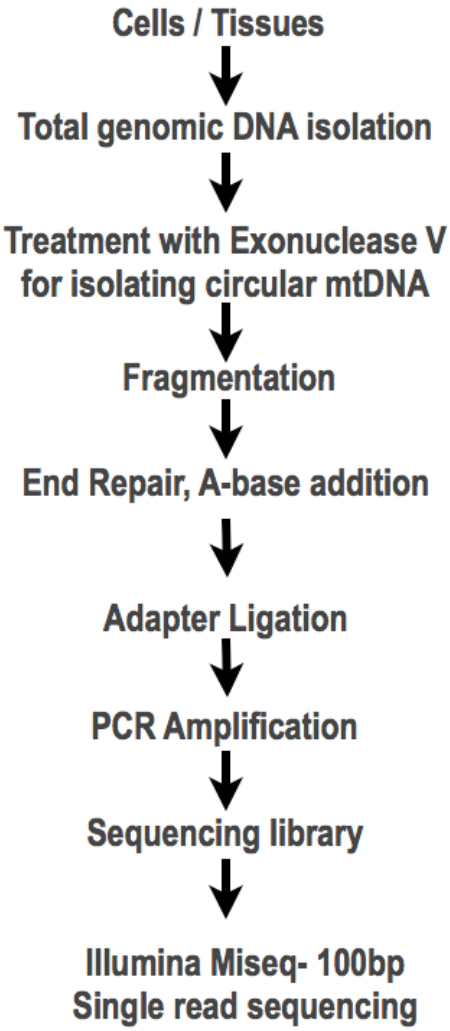


Tables

Supplementary Material

bioRxiv preprint doi: https://doi.org/10.1101/007005; this version posted July 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2 The Mseek protocol.



bioRxiv preprint doi: https://doi.org/10.1101/007005; this version posted July 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3 Novel variants in a cell-line (MDA-MB-157, circles) and a human (PBMC, crosses). Mutation frequency (y-axis) versus position on mtDNA (x-axis). Genes are colored bands at bottom of graph ($+, -$ represent the strand, cds is coding sequence). *Neutral*, *Low*, *Medium*, and *High* are the effect of the mutation on protein function (*Mutation Assessor*[27]). Except for the D-loop, most of the mtDNA codes for a transcript, with a few gaps. The longer gaps (11, 24, and 30 nt long) are marked by vertical red lines. The 45 nt long overlapping region between ATP8 and ATP6 is marked by black vertical lines. Mutation frequencies between 0 and 1 arise from the co-existence of multiple mtDNA haplotypes. Heteroplasmy at a cellular level is demonstrated in Fig. 4. There are no stark differences between the human and cell-line derived mtDNA.

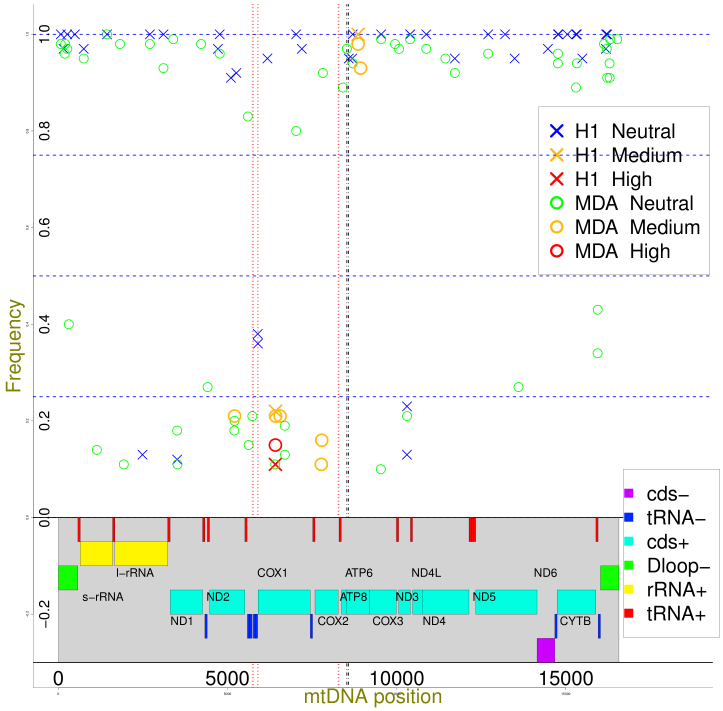
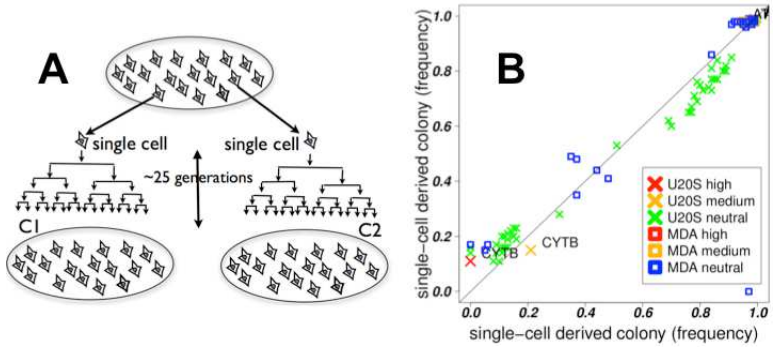
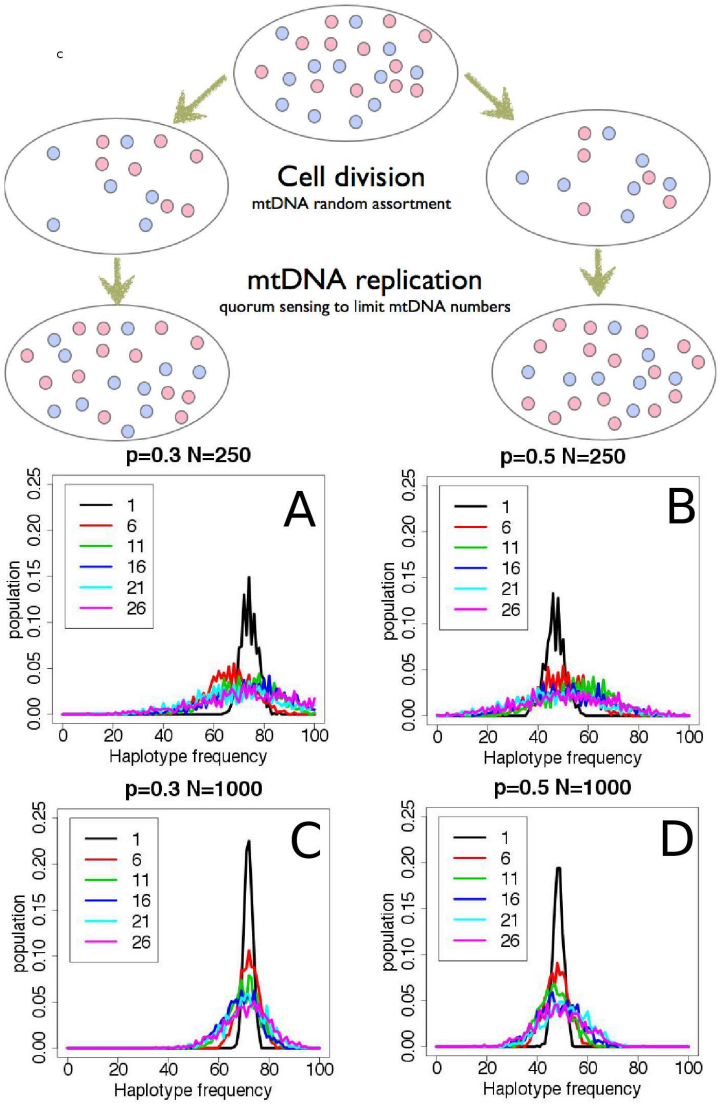


Figure 4 Stability of heteroplasmy. **A** C1 and C2 are colonies derived from single cells in the original colony, placing a severe botteneck on the mtDNA, and then passaged ≈ 25 times. **B** Mutation frequencies in C1 (x-axis) versus C2 (y-axis) for two cell-lines, U20S and MDA-MB-157. The mutations mostly lie along the diagonal; the heteroplasmic mix of mutations in the derived colonies are similar to each other. This implies the heteroplasmic mix exists at the single-cell level, and is stable over many divisions. A drift in frequencies is expected with random assortment of mtDNA haplotypes (simulations, Fig. 5). The stability of the frequencies implies active mechanisms to counteract the drift, such as the exchange of mtDNA between cells.



bioRxiv preprint doi: https://doi.org/10.1101/007005; this version posted July 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 5 Simulations of mtDNA replication. Each cell contains a mixture of mtDNA haplotypes, here we consider only two species, red and blue, for the sake of simplicity. At cell division, the mitochondria assort randomly between the daughter cells and divide (with mtDNA replication) till a quorum of mitochondria, specific to each tissue-type, is reached. Quorum sensing is not well-understood, but the number of mitochondria per cell is tissue-specific[9]. The graphs show simulations of the evolution of heteroplasmy over time, based on this model. The distribution of haplotype frequencies spreads over time (drift), implying two randomly selected cells are unlikely to have the same mtDNA haplotype distribution. The plots show the distribution of frequencies after 1, 6, 11, 16, 21 and 26 divisions. The starting number of mitochondria per cell (N) is 250 in the upper panels (A,B) and 1000 in the lower panels (C,D). One of the alleles occurs with a frequency (p) of 0.3 in the left panels (A,C) and p=0.5 in the right panels (B,D). The drift is slower and the distributions narrower with larger N and smaller deviations of p from 0.5.



bioRxiv preprint doi: https://doi.org/10.1101/007005; this version posted July 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1 Data from Mixing experiments of HCC cells co-cultured with MDA-MB-157 for 6 weeks. The column private_to identifies the cell-lines that exhibit the variant. Rows highlighted in gray are cases where a variant unique to HCC has been identified in MDA-MB-157 cells co-cultured with HCC. The light green rows are variants private to HCC that did not transfer into MDA-MB-157. Rows highlighted in blue show variants common to both cell lines. For example, at position 3796 (row 2), the A from the reference mtDNA genome is mutated to a T only in HCC, the MDA-MB-157 cultured with HCC exhibits an A, with a frequency of 0.14 (or 14%). f is the frequency and c is the coverage.

mut	gene	AA	private_to	HCC in MDA-MB-157(f|c)	MDA-MB-157 in HCC(f|c)	flank
G3666A	ND1	G120G	HCC	0.93|33	0|25	TCTGATCAGG[G|a]TGAGCATCAA
A3796T	ND1	T163S	HCC	0.88|9	0.14|21	TAACCTCTCC[A|t]CCCTTATCAC
A4104G	ND1	L266L	HCC	0.9|21	0.06|31	TTCTAACCTC[A|g]CTGTTCTTAT
C7819A	COX2	L78L	MDA-MB-157	0|35	0.76|43	CGCATCCTTT[C|a]CATAACAGAC
T7861C	COX2	D92D	HCC	0.96|29	0.05|39	CGGACTAATC[T|c]TCAACTCCTA
C8932T	ATP6	P135S	MDA-MB-157	0|25	0.9|21	CCATACTAGT[C|t]ATTATCGAAA
G9064A	ATP6	A179T	HCC	1|10	0.16|24	CCATTAACCT[G|a]CCCTCTACAC
A9072G	ATP6	S182S	HCC	0.88|9	0.18|22	CTTCCCTCTA[A|g]ACTTATCATC
T9540C	COX3	L111L	A382|MDA-MB-157|HCC	0.86|15	1|11	TACCCCCCAA[T|c]TAGGAGGGCA
A9545G	COX3	G113G	HCC	0.93|15	0|13	CCCAATTAGG[A|g]GGGCACTGGC
G10688A	ND4L	V73V	HCC	0.96|26	0.1|19	ACCTGACTCC[G|a]ACCCCTCACA
A12810G	ND5	W158W	HCC	0.92|25	0|20	AATCCTATAC[A|g]ACCGTATCGG
T13602C	ND5	Y422Y	MDA-MB-157	0 |22	0.45|24	CAAGCGCCTA[T|c]AGCACTCGAA
C14911T	CYTB	Y55Y	HCC	0.96|29	0|39	CTATATTACG[C|t]ATCATTTCTC
T15514C	CYTB	Y256Y	MDA-MB-157	0 |12	0.89| 28	CAGACAATTA[T|c]ACCCTAGCCA
C16187T	D-loop		HCC	1|9	0|31	ATCAAAACCC[C|t]CTCCCCATGC
T16209C	D-loop		MDA-MB-157	0|17	0.92| 26	TACAAGCAAG[T|c]ACAGCAATCA
C16223T	D-loop		A382|MDA-MB-157|HCC	0.95| 20	0.96| 31	GCAATCAACC[C|t]TCAACTATCA
C16292T	D-loop		MDA-MB-157	0|15	0.76| 21	CAAACCTACC[C|t]ACCCTTAACA
C16295T	D-loop		MDA-MB-157	0|14	0.72| 22	ACCTACCCAC[C|t]CTTAACAGTA
T16311C	D-loop		HCC|MDA-MB-157	0.75|12	0.8|20	CAGTACATAG[T|c]ACATAAAGCC

Table 2 Data from Mixing experiments of U20S co-cultured with A382 for 6 weeks. The column private_to identifies the cell-lines that exhibit the variant. The dark gray highlights rows where a variant unique ot A382 has been identified in U20S cells co-cultured with A382, only once case was identified, the variant at position 315 (row 2). The A382 culture from the co-culture did not yield sufficient mtDNA for sequencing. The light blue rows are variants private to A382 that did not transfer into U20S.

mut	gene	AA	private_to	U20S in A382(f|c)	flank
T146C	D-loop		U20S	0.97|37	ATTCCTGCCT[T|c]ATCCTATTAT
C315CC	D-loop		A382	0.23|30	CCCCTCCCCC[C|Cc]CTTCTGGCCA
G3010A	l-rRNA		U20S	1|87	AGGACATCCC[G|a]ATGGTGCAGC
C3699G	ND1	G131G	U20S	0.98|85	CCCTGATCGG[C|g]GCACTGCGAG
T4216C	ND1	Y303H	U20S	1|64	ACTTATATGA[T|c]ATGTCTCCAT
A8701G	ATP6	T58A	A382|MDA-MB-157|HCC	0|121	ACAAATGATA[A|g]CCATACACAA
T9540C	COX3	L111L	A382|MDA-MB-157|HCC	0|65	TACCCCCCAA[T|c]TAGGAGGGCA
A10398G	ND3	T113A	HCC|MDA-MB-157|U20S	1|35	ATTAGACTGA[A|g]CCGAATTGGT
G10685A	ND4L	A72A	U20S	1|98	GCGAAGCAGC[G|a]GTGGGCCTAG
A11251G	ND4	L164L	U20S	1|127	TCATCGCACT[A|g]ATTTACACTC
A12612G	ND5	V92V	U20S	1|62	TCATCCCTGT[A|g]GCATTGTTCG
T13281C	ND5	V315V	U20S	0.99|127	TCATAATAGT[T|c]ACAATCGGCA
G13708A	ND5	A457T	U20S	0.98|93	TAAACGCCTG[G|a]CAGCCGGAAG
A13933G	ND5	T532A	U20S	0.99|115	CCCTAGCATC[A|g]CACACCGCAC
T14798C	CYTB	F17L	U20S	1|61	TAACCACTCA[T|c]TCATCGACCT
C15263A	CYTB	P172T	U20S	0.17|129	AGTAGACAGT[C|a]CCACCCTCAC
G15301A	CYTB	L185L	A382|MDA-MB-157	0|133	ACTTCATCTT[G|a]CCCTTCATTA
C15452A	CYTB	L235I	U20S	0.97| 135	ACTTCTCTTC[C|a]TTCTCTCCTT
C16069T	D-loop		U20S	1|78	AGTATTGACT[C|t]ACCCATCAAC
C16108T	D-loop		U20S	0.98|71	ACATTACTGC[C|t]AGCCACCATG
T16126C	D-loop		U20S	1|89	ATGAATATTG[T|c]ACGGTACCAT
C16223T	D-loop		A382|MDA-MB-157|HCC	0| 168	GCAATCAACC[C|t]TCAACTATCA

bioRxiv preprint doi: https://doi.org/10.1101/007005; this version posted July 23, 2014. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S1 Organization of mtDNA. 13 protein-coding genes, 22 tRNA genes and 2 rRNA genes are encoded by a single circular nucleic acid and transcribed from three promoters: LSP (inner circle of genes - strand), HSP1 (outer circle of gene, + strand) and HSP2 (16S rRNA) on the D-loop, which is non-coding but critical for replication and transcription. The three polycistronic transcripts are processed by enzymatic excision of the tRNAs. There are a few small gaps (< 30 nt) in annotation, and a 45nt overlap between ATP6 and ATP8 which might have roles in replication.

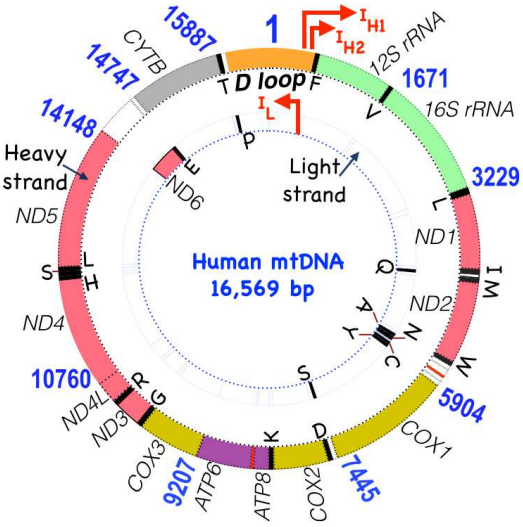
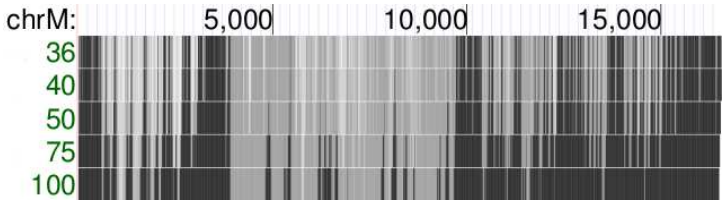



Figure S2 Mappings of mtDNA tiles. The x-axis is the position along the mtDNA. Each band shows mappings of n-mer tiles ($n = 36, 40, 50, 75, 100$) from mtDNA on the human genome. Darker regions are more unique. Even 100 nt tiles from mtDNA often map to nDNA; thus, unambiguous identification of mtDNA variants requires efficient isolation of mtDNA. The pseudo-gene copies of mtDNA in the nuclear genome are called nuclear-mtDNA pseudogene sequences or **Numts**.