RESEARCH

Stable heteroplasmy at the single cell level is facilitated by inter-cellular exchange of mtDNA

Anitha D Jayaprakash[1], Erica Benson[1], Swapna Gone[4], Raymond Liang[1], Jaehee Shim[1], Luca Lambertini[1], Masoud M Toloue[4], 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. 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 mtDNA pseudogenes (Numts) in nDNA. To understand the nature of heteroplasmy, we developed Mseek, a novel technique to purify and sequence mtDNA. Mseek yields high purity ($> 90\%$) 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. We hypothesized that the stability of heteroplasmy could be facilitated by inter-cellular exchange of mtDNA. We 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 features of heteroplasmy in normal and diseased states.

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 proteins from the mitochondrial genome, a small circular DNA (**mtDNA**), play a critical role in their function. In humans, the reference mtDNA is 16,569 bp long 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]. Each mitochondrion carries multiple mitochondrial genomes $(5-10)$[7] and each cell contains hundreds to thousands of mitochondria, depending on the tissue[8]. The mtDNA replicate without recombination. mtDNA is inherited solely from the mother; inherited mutations in mtDNA have been linked to several genetic disorders including *diabetes mellitus and deafness*(DAD) and *Leber's hered-*

itary optic neuropathy(LHON)[9]. De novo mutations in mtDNA have also been linked to diseases[10, 11, 12, 13].

Heteroplasmy, which is the occurrence of multiple mtDNA haplotypes, has been documented in various studies, cancer cells[14, 15], blood samples from families[16] blood and muscle biopsies from identical twins[17], and cells from the 1000 genomes project [18, 19]. Though extensive, these studies have not established the nature of heteroplasmy at the *level of the single cell*. 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[14, 20]. 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 mitochondrial DNA (**Numts**), present in the mammalian nDNA[21] (**Fig. S2, S3**).

Without purification of mtDNA, Numts introduce unpredictable inaccuracies in the estimates of heteroplasmy, especially because they exhibit variations in sequence and copy numbers. Numts have been annotated in the reference human genome[22, 23], and there are tools to analyze high throughput sequencing data in light of these annotations [24], but a comparison of two recent versions of the reference human genome (hg19 and hg38 on the UCSC genome browser, **Fig. S2, S3**) suggests that this annotation is not complete and changes significantly with the reference genome. Additionally, the distribution of Numts might be specific to each individual's nuclear genome. A recent study of mtDNA from twins has highlighted the need for further investigation of Numts and their potential to confound analyses of heteroplasmy[25].

Isolating mtDNA has long been a challenge. In forensics and genealogy, allele-specific primer extensions (SNaPshot) are used for genotyping mtDNA[26]. Hyper variable regions (HVR) in the D-loop have been amplified using PCR[27]. Entire mtDNA has been accessed using primers specific to mtDNA to either perform long-range PCR[28], or amplify overlapping fragments[14]. Isolation of organelles by ultra-high-speed centrifugation[29] has also been used, but the yields are low and contaminated with fragmented nuclear DNA[30]. Computational methods have also been used to infer heteroplasmy from whole-exome[19, 31] and whole-genome data[15], but such data contains Numts and furthermore, generating such data for new samples is expensive. A new approach uses methyl-specific endonucleases MspJI and AbaSI to deplete nDNA that is likely to be methylated[32], but Numts need not be methylated. Heteroplasmy derived from PCR-based methods are error-prone, due to variability in amplification. Clonal amplification of errors introduced by polymerases is also a 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, it is difficult to ascertain if the methods outlined above are accurate in their measurement of frequencies of common variants and their identification of variants that occur at frequencies below 5%.

We present here Mseek, a novel method to enzymatically purify mtDNA by depleting linear nDNA and inexpensively sequencing it (**Fig. S4**). Mseek uses exonuclease V to digest away linear nDNA, leaving behind circular mtDNA. A major benefit

of this method is the ability to call extremely rare variants, with the sensitivity of calls only limited by the sequencing depth. By applying Mseek to several cell-lines and human peripheral blood mononuclear cells (PBMC), we identified mixtures of different mtDNA haplotypes in the samples. Additionally, through clonal expansion of single cells from a variety of cell lines, we establish, for the first time, that heteroplasmy is stably maintained at a single cell level over multiple divisions. We infer the stability is due to intercellular exchange of mtDNA and experimentally demonstrate this exchange by co-culturing two different cell-lines with distinct mtDNA haplotypes to show that there is transfer of mtDNA from one cell type to another.

Results

Mseek: An efficient method to isolate and sequence mtDNA

To efficiently purify mtDNA, 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 efficiency of digestion, sequences specific to nDNA and mtDNA were PCR amplified using appropriate primers (**Tables ST1 and ST2**). As expected, in the undigested samples of total DNA we detected both nDNA and mtDNA (**Fig. 1A**). In sharp contrast, in the samples treated with exonuclease V we only detected 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 rest of the reads were adapter dimers, formed during library preparation by the ligation of adapters to each other. 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. This approach, designated **Mseek** (**Fig. S4**), provides a means of unmatched efficiency in accurately sequencing the mtDNA contained within a population of cells.

Our comparisons of Mseek to other kits on the market show that Mseek substantially outperforms all of them in terms of the purity of mtDNA and yield (Supplementary material and **Fig. S5**). As currently implemented, a limitation of Mseek is its requirement of at least 4 μgrams of intact total DNA, which is a big improvement over the 50μg of total DNA that were needed for the initial experiments. Further improvements of sample preparation may reduce the input amount required, but are beyond the scope of this paper. For less than 4μg of total DNA we recommend using long-range PCR amplification with mtDNA specific primers after exonuclease V treatment, which depletes nDNA to minimize distortions arising from Numts. The primers for long range PCR are specified in table ST2 and enzyme used for amplification is from Clontech (Advantage Genomic LA Polymerase Mix catalog # S4775).

Multiple rounds of treatment with exonuclease V maybe needed to achieve high purity, a single round usually achieves around 80% purity. Assuming around 20 copies of the mtDNA per nuclear genome (which is probably an overestimate), the contamination from Numts is less than 0.01%, which can be safely ignored for most

applications. Thus, Mseek can be used to call rare variants to any level of sensitivity, only limited by the depth of sequencing. Most methods will not allow this, for example, in PCR-amplicon sequencing, sensitivity does not necessarily increase with depth of sequencing as errors introduced during amplification cannot be corrected by greater sequencing depth. This is a valuable feature, that distinguishes it from other techniques, enabling the tracking of rare variants. Duplex sequencing, which uses adaptors with random tags to track clones in order to reduce the errors introduced during sample preparation for deep-sequencing, should be used in conjunction with Mseek to identify and study rare variants (with frequencies $< 0.1\%$) [33].

Ubiquity of heteroplasmy

Since cell lines are clonally derived, the nDNA (and mtDNA) are expected to be identical across cells, since, either i) a slight fitness advantage of one haplotype or ii) drift[34], would lead to clonal selection and homoplasmy. To study clonality in mtDNA, 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). Repeat content of the sequences was computationally identified to estimate nDNA contamination, which ranged from $0.5 - 5\%$; confirming the specificity of Mseek. Importantly, because of this high degree of mtDNA purity, we were able to multiplex all 30 samples in a single MiSeq run, with average coverage of > 50X. We show the coverage at various positions in the Tables 1 and 2.

The sequences were analyzed for variants using MiST[35]. Variants with a frequency of 1 indicate homoplasmic mtDNA. Frequencies less than 1 imply the coexistence of multiple haplotypes. Strikingly, in both cell-lines and human blood-derived mtDNA, we observed variants occurring in the $0.1 - 0.9$ frequency range (**Fig. 2**), indicating the presence of multiple haplotypes. Most mutations were transitions (Table ST3), suggesting that the mutations don't arise from oxidative stress, and are most likely driven by the polymerase-γ activity[33]. The program *Mutation Assessor*[36] 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 mutations of putative high effect at a high frequency (> 5 %), consistent with the expectation that functioning cells should have functional mitochondria.

Each sample had unique, distinguishing mutations, ranging in frequency from 0.36 to 1.0. There were a number of variants unique to each of the human PBMC samples (ranging in number from 5 to 15) and each of the cell lines (ranging in number from 5 to 21). Our findings hold for cell lines derived from a variety of tissues, suggesting they are of general applicability. There were no key distinguishing features between normal or cancer cell lines and human blood-derived mtDNA, in terms of deleterious mutations or degree of heteroplasmy.

Stability of heteroplasmy in cell-lines

The results above indicate that heteroplasmy exists within a cell population but questions remain about the nature of heteroplasmy in individual cells. A mixture of homoplasmic cells with different haplotypes can appear to be heteroplasmic. 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. 3**). In each of the derived colonies (4 colonies per cell line), variants from the original lines existed in the derived colonies at approximately the same frequencies as in the original colonies. The sharing of mutations between the original and derived colonies implies the diversity in mtDNA is present 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. 3**). Low frequency mutations (frequency $< 2\%$) in the original colony do disappear in the derived colonies, suggesting the rarer variants might be present in a subset of cells, unlike the more common variants, and might be transient.

A simple model of mtDNA genetics assumes random assortment of mtDNA haplotypes between daughter cells upon cell division, followed by multiplication of mitochondria. This model would predict drift towards homoplasmy, as seen in our simulation of this process (**Fig. 4**) and by others[34]. 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. 4**). 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 across the population. Other explanations, such as a balancing selection[37] or segregated replication of groups with fixed composition[38] could also be invoked to explain the lack of drift, but involve complicated mechanisms. These can also 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 test the ability of mtDNA to transfer between cells, we co-cultured cell lines with distinct mtDNA heteroplasmy signatures. Since we were not sure which kinds of cells would allow transfer and if contact between cells were important, we tested mixtures of a pair of untransformed cell lines from embryo lung fibroblasts (IMR90 and WI-38) as well as four cancer cell lines (MDA-MB-157, U2OS, A382 and HCC1806).

One cell line in each of the co-cultured pair was labelled with constitutively expressed GFP, with 10-fold more non-GFP cells than GFP. After co-culturing for 6 weeks, the GFP labelled cells were isolated (either sorting by FACS or by using Neomycin resistance) and placed in culture again for upto 2 weeks, to obtain 10 million cells, which were then prepared for mtDNA sequencing. The details are given in the methods section. There were four experiments,

1. HCC1806-GFP mixed with MDA-MB-157 (both breast cancer) (FACS)
2. U20S-GFP mixed with A382 (both cancer cells) (FACS)
3. MDA-MB-157-GFP (breast cancer) mixed with IMR90 (normal fibroblast) (Neomycin resistance)

4 WI-38-GFP mixed with IMR90 (normal fibroblast cells) (Neomycin resistance)

In two cases, (WI-38-GFP, IMR90 **Fig. 5**, Table 1) and (HCC1806-GFP,MDA-MB-157, Tables 2), we detected variants private to the non-GFP-labelled cell-line in the co-cultured partner cell-line, suggesting the transfer of mtDNA between the cell-lines. The private variants were transferred to varying degrees, ranging from thorough mixing to no transfer, arguing against the results arising from errors in sorting or cytoplasmic/nuclear exchange between cells. The purity of the sorted cells, based on FACS, excludes nuclear exchange as an explanation for the findings. The lack of transfer in the two pairs, 1) MDA-MB-157-GFP,IMR90 (Table ST4) and 2) U20S-GFP, A382 (Table ST5) further suggests that the transfer results are not artifacts. In all cases, the cultures were visually inspected to ensure that both cell-lines were thriving.

Discussion

Sensitive detection of heteroplasmy is important as its variability may have clinical significance, as a biomarker and in disease progression[39]. Analyses of heteroplasmy is confounded by Numts, highlighted by a study that used whole-genome data from the TCGA to infer that deleterious mtDNA mutations are more common in cancer cells compared to normal tissue[15]. In contrast, findings of low mutations rates in tumor mtDNA from a colorectal cancer study[14] are more in line with our findings that cancer cell lines don't exhibit higher rates of deleterious mutations compared to normal cells from human tissues. A study of mtDNA from twins further highlights the need for further investigation of Numts[25].

Mseek provides the means to purify and deeply sequence mtDNA and determine heteroplasmy accurately by eliminating Numts and PCR-related biases. The sensitivity of Mseek increases with sequencing depth and can be made more accurate by using Duplex sequencing[33]. As currently implemented, a limitation of Mseek is its requirement of at least 4 μgrams of intact total DNA. Using Mseek we have demonstrated that cells from a wide-range of cell lines and human samples exhibit heteroplasmy, in accord with results from several studies[14, 19].

Heteroplasmy might be an essential feature of mtDNA, seemingly providing 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 mtDNAs 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).

By studying mtDNA from colonies derived from single cells we showed that heteroplasmy is stable at the single-cell level, which is surprising in light of, 1) the higher rates of mutation in mtDNA[40] which should increase the diversity of mtDNA, and 2) drift, which should lead to homoplasmy in about 70 generations[14, 34]. Another study has determined the stability of heteroplasmy using PCR to track a particular mutation (A3243G)[38], they proposed multiplication of segregating units with fixed mixtures of haplotypes to explain this stability. Intercellular exchange of mtDNA is the simplest explanation for this stability, on the basis of our cell-line data (**Fig. 3**) in conjunction with simulations (**Fig. 4**) and co-culturing experiments (**Tables 1,2** and **Fig. 5**).

The transfer of mtDNA must occur through whole mitochondrial organelle transfer, as naked mtDNA in the cytoplasm would lead to an innate immune response[41], leading to cell death. To ensure that the results are not artifacts, we used two different methods to select the GFP positive cells, visually inspected over a 100 cells to make sure there was no contamination after the sorting and inspected the cells daily to make sure the colonies were growing stably and the cultures were not taken over by one cell line. The sequencing data also provides evidence that the results are not artifacts, as the transfer of variants is not uniform, some variants show a lot of transfer while others show very little or none. Additionally, certain pairs of cell lines do not show mtDNA transfer despite being co-cultured for over 6 weeks, further suggesting that the data is not an artifact and contact between cells might be needed to facilitate organelle transfer. It is possible that co-culturing cell lines that show transfer for a sufficiently long time will lead to a thorough mixing the mtDNA between the cells.

Intercellular transfer of mtDNA has been seen in other instances. Horizontal transfer of genetic material between species of yeasts has been shown[42]. Organelle transfer between cells through microtubule formation is increasingly a focus of many studies[43]. Exchange of mtDNA between mitochondria within a cell is facilitated through networks created by fusion, mediated by the mitofusins *Mfn1* and *Mfn2*[44]. This intracellular exchange is essential for functional mitochondria; knocking out the fusins causes muscles to atrophy through the accumulation of deleterious mtDNA mutations[44]. *In vivo*, exchanges of mitochondria between cells have also been demonstrated in the rejuvenation of cells with damaged mitochondria by transfer of functional mitochondria from mesenchymal stem cells[45]. Rejuventation of cells containing damaged mtDNA by transfer of functional mtDNA from neighboring cells in culture has also been observed[46].

The present study is the first explicit demonstration of mtDNA transfer between healthy human cells in culture. Prior studies have shown transfer from cells with functional mtDNA into ones with non-functioning mtDNA using protein markers[45, 46]. The exchange of mtDNA between cells can explain its stability over the lifetime of an organism, inferred from the relative paucity of age-related disorders originating in somatic mtDNA mutations and over generations, inferred from the ability to identify the geographic origins of a person from the mtDNA sequence. The stability of mtDNA against deleterious mutations could be enhanced by a coupling between replication and transcription[47], ensuring the depletion of non-functional mtDNA by inefficiencies in replication.

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

Methods

Mseek

Our method of isolating and sequencing mtDNA, dubbed Mseek (**Fig. S4**), consists of the following steps, i) Total DNA is isolated from the sample, ii) The nDNA is digested using Exonuclease V, iii)The products are purified using Ampure beads to remove short fragments, iv) To test the results of the digestion, PCR primers specific to mtDNA and nDNA are used on 1μl of the digested sample (**Fig. 1B**), v) the rest of the sample is fragmented using Covaris and end-repaired, vi) Barcoded adapters compatible with the sequencing platform are ligated to the fragments and vii) The universal adapters are used to amplify the library for loading on deep-sequencing instruments.

Sample Processing

Total DNA was isolated from 500μl of whole blood or cell lines using the Epicentre protocol (MC85200). DNA was eluted in 100μl of TE buffer and checked for quality and quantity using a 1% agarose and nanodrop respectively. The eluted DNA was further heated at 70C for 30 minutes to inactivate any left over proteinase K that was introduced in the DNA isolation protocol. A first digestion was carried out by adding to the total DNA Sample (4-8μg in 35 μl), the following, NEB4 10X Buffer (6 μl), 10 mM ATP ATP (12 μl), ExoV from NEB-M0345S (4 μl) and H2O (3 μl). The digests were left at 37C for 48 hours, heat inactivated at 70C for 30 minutes and purified using AMPure beads (Beckman Coulter). A second digestion was carried out by adding to the ExoV treated DNA (in 35 μl) the following, NEB4 10X Buffer (6 μl), 10 mM ATP (12 μl), ExoV from NEB-M0345S (4μl) and H2O (3 μl). The digests were left at 37C for 16 hours, heat inactivated at 70C for 30 minutes and purified using AMPure beads (Beckam Coulter).

The primers listed in Tables ST1 and ST2 were used to detect presence of nuclear and mtDNA. Only if the digested DNA did not show significant PCR product for nuclear DNA, were the samples were processed for deep sequencing. In cases with nuclear DNA contamination, the digestions were repeated to improve mtDNA purification. We used Covaris for shearing and the Rapid DNA kit from Bioo Scientific for DNA library prep (5144-01).

Cell culture

293T (embryonic kidney cells containing the SV40 T-antigen), 501T (normal adult fibroblasts), U2OS (osteosarcoma), Saos-2 (osteosarcoma), HCC1806 (breast cancer), MDA-MB-157 (breast cancer) and A382 (glioblastoma) cells were grown in Dulbecco's Modified Eagle's Medium (Sigma-Aldrich, St Louis, MO, USA) supplemented with 10% fetal bovine serum (Sigma-Aldrich), 50 units/ml of penicillin/streptomycin (Gibco, Grand Island, NY, USA). All cells were sub-cultured or collected using 0.05% trypsin-EDTA (Gibco) and maintained at 37C and 90% humidity in a 5% CO2 incubator. For selection of cells carrying the neomycin selection marker, 750μg/ml G418 (Gibco) was added to the culture medium for 2 weeks. Control cells not carrying the resistance marker were used to verify cell death by G418. Clonal isolation of tumor cells was performed by serial dilution into 96-well plates and visual examination of wells for single cells. Single cells were then expanded for an additional 28-30 population doublings, expanding into larger tissue culture plates as necessary.

FACS sorting

Cultured cells were collected in PBS (Gibco) at a density 5 x 10^6 cells/ml and passed through a 40μm filter. Cells were sorted using the BD FACS Aria II cell sorter (Becton-Dickinson, Mountain View, CA, USA), using the 488nm laser. Sorting was performed in a sterile BSL2+ biosafety cabinet. FACSDiva Version 6.1.2 software (Becton-Dickinson) was used for analysis.

Lentiviral GFP expression

To ectopically express GFP, we used a NSPI-derived lentiviral vector that drives GFP expression with a constitutive PGK promoter and contains a neomycin selection marker[48]. High titer lentiviral production and infection were carried out as previously described[48].

Mixing experiments

Two strategies were used for selecting GFP-positive cells in the mixing experiments.

In the first strategy, HCC1806 and U2OS cells were transduced with a high titer of lentivirus to constitutively express GFP. Four days after infection, GFP-positive cells were isolated by utilizing the BD FACSAria II cell and FACSDiva Version software. GFP-positive cells were placed back in culture. An examination of several hundred cells showed that all were GFP positive, suggesting that FACS sorting was highly efficient and there were no GFP-negative cells.

Pairs of cell lines (one GFP-positive and the other GFP-negative) were mixed and co-cultured in 150mm tissue culture dishes. In one group of experiments, MDA-MB-157 was co-cultured with HCC1806-GFP (ratio of 10:1 ensured by counting cells) In a second group of experiments, A382 cells were co-cultured with U2OS-GFP cells. The cells were allowed to grow for 4 weeks and were sub-cultured 1:4 or 1:5 when they reached near confluency, which was about every four days. The culture was monitored visually to ensure both GFP-positive and GFP-negative cells existed in culture.

Live GFP positive and GFP negative cells were separately gated and sterile sorted using the BD FACSAria II cell and FACSDiva software according to GFP status. Sorted GFP-positive cells were placed back into culture and allowed to expand for a week or two, to obtain 5 x 10^6 cells in a 150 mm plate, giving $\approx$ 4.5 x 10^6 of unlabelled cells and 5 x 10^5 of GFP-labeled cells, which was then processed for mtDNA sequencing. The purity of the cells after sorting was further verified by visual examination under the microscope.

The second strategy was similar to the above with the following changes. 1) After the infection with the high titer GFP lentivirus with neomycin resistance, GFP-positive cells were selected for neomycin resistance. 2) After completion of the mixing time period, GFP negative cells were specifically selected against using the neomycin selection marker as an alternative to cell sorting. Visual examination was used to further verify the purity of the GFP positive population. In this way, only GFP positive/neomycin resistant cells were collected for mtDNA isolation. MDA-MB-157-GFP mixed with IMR90, and WI-38-GFP mixed with IMR90 were processed by this strategy.

Analyses

The sequencing data is generated as fastq format files. These were processed using a pipeline developed by us for whole-exome analysis called MiST which is described elsewhere[35]. In brief, the sequences were filtered for quality (sequences with more than 10 consecutive nucleotides with $Q < 20$ were eliminated) and mapped to the reference mitochondrial genome (accession NC_012920 from Genbank). Identical reads were identified as being clonal and were considered only once, irrespective of the number of copies, towards variant calling. A variant call was made only if there were at least three non-clonal reads carrying the variant, at least 10nt away from the ends, and a minimum coverage of ten was required at the variant.

Variants occurring on reads predominantly on one strand ($> 80\%$) of the mtDNA were excluded to further reduce errors, based on our prior experience[35]. The error rate in Miseq and Hiseq reads are usually less than 1 in a 1000 (phred score $Q > 30$), requiring at least 3 non-clonal reads reduces the error rate to well under 1 in a million. Nuclear contamination was estimated using sequences that map to repeat elements such as LINEs and SINEs, which only occur in nDNA. This enables reliable estimation of the level of nDNA contamination, which ranged from $0.5 - 1.5\%$.

The annotation of variants was determined using mtDNA annotations from MITOMAP13. Common SNPs (and haplotype indicators) were identified from dbSNP14. Various programs that annotate the effect of variants on protein function were tested. We eliminated programs that indicated common SNPs in mtDNA proteins were deleterious, *Mutation Assessor*[36] performed the best under this test, we used it to assess the impact of mtDNA mutations on protein function.

Mutation Assessor 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 might affect mitochondrial function.

Custom code was developed for simulations and the plots were created using Gnuplot and R.

Author's contributions

AJ and RS designed Mseek and designed several experiments. SA and MW designed key experiments with cell-lines. LL provided human samples and suggested applications. SG and MT compared Mseek to other kits. AJ, JS, RL used Mseek on samples, 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 and Eirini Papapetrou 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.

Supplementary Material

Comparison of Mseek to other techniques

We tested several kits on the market using PCR with a combination of nuclear DNA and mtDNA specific primers and followed it up with sequencing only if the PCR data looked promising. PCR-based approaches (e.g. Life technologies) can be excluded for reasons discussed in the introduction. Approaches using SNP arrays (23andMe) don't have the resolution offered by Mseek. Most other kits on the market for mtDNA purification involve lysing cells to release the organelles either chemically or by use of a dounce followed by the use of one of the following technologies,

- Purification of mitochondrial DNA from organelles using Enzyme B mix from Abcam to clean up the DNases and other proteins (kits from Abcam, PromoKine and BioVision),
- Magentic capture of mitochondria using anti-TOM22 microBeads to target the translocase of the outer mitochondrial membrane 22 (TOM 22) (MACS from Miltenyi Biotec)
- Amplification from total DNA with mtDNA specific primers (REPLI-g from Qiagen),
- Extraction of an mtDNA-enriched fraction using Qiagen's plasmid miniprep kit followed by bead purification using Agencourt AMPure XP system[49].

We first tested the mini-prep method of isolating pure mtDNA[49]. Approximately 20 million cells (HEK 293T) were used for each isolation. The cells were loaded onto a spin mini prep column and isolation was done according to the manufacturer's protocol (QIAGEN spin mini prep catalog # 27115). The mtDNA which is similar to plasmid DNA (size and circular structure) was eluted in 100 μl of elution buffer. The mtDNA enriched fraction was later purified using the AMPure XP system. A 0.4 X proportion of beads by volume were added, collected on the magnetic stand and washed twice with 80% ethanol. Post washing, the mtDNA was resuspended in 25 ul 0.1 X TE buffer. On testing with PCR, nDNA bands were seen(**Fig. S5A**), consistent with the published data[49] showing that the product had about 78% nuclear DNA[49]. In that study, a subsequent mtDNA-specific amplification yielded high purity mtDNA.

We also tested the Miltenyi biotech Mitochondria MidiMACS starting kit, human (catalog # 130-094-872), since it is used by many commercial kits. Mammalian cells ($\approx 10^7$) were lysed and mitochondria were magnetically labeled with Anti-Tom22 microbeads which bind to the translocase of the outer mitochondrial membrane 22 protein (TOM 22). The sample was then loaded onto the column placed in the MACS separator. After washing only magnetically labeled mitochondria are retained on the column. The column is detached from the separator and mitochondrial organelles are eluted. Mitochondrial DNA is further isolated from the organelles by isopropanol precipitation. Since nuclear DNA remained in the MACS-purified samples according to PCR(**Fig. S5B**), consistent with the published data[49], sequencing was not performed.

We compared Mseek to REPLI-g (Qiagen), using sequencing on mtDNA from a blood sample (in triplicate) to establish that Mseek exhibits the highest purity of mtDNA compared to kits currently available on the market(**Fig. S5C**). Mseek data from cell lines shows substantial reduction of nuclear DNA and uniform coverage across the length of mtDNA(**Fig. 1**).

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. [4]Bioo Scientific Corporation, 7050 Burleson Road, Austin TX, USA.

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. 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)
8. 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
9. Wallace, D.C.: Mitochondrial DNA mutations in diseases of energy metabolism. Journal of bioenergetics and biomembranes **26**(3), 241–250 (1994). PMID: 8077179
10. DiMauro, S., Schon, E.A.: Mitochondrial DNA mutations in human disease. American journal of medical genetics **106**(1), 18–26 (2001). PMID: 11579421
11. 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)
12. Taylor, R.W., Turnbull, D.M.: Mitochondrial DNA mutations in human disease. Nature reviews. Genetics **6**(5), 389–402 (2005). PMID: 15861210
13. Greaves, L.C., Taylor, R.W.: Mitochondrial DNA mutations in human disease. IUBMB life **58**(3), 143–151 (2006). PMID: 16766381
14. 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
15. 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
16. Goto, H., Dickins, B., Afgan, E., Paul, I.M., Taylor, J., Makova, K.D., Nekrutenko, A.: Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. Genome Biology **12**(6), 59 (2011). Accessed 2014-09-23
17. Avital, G., Buchshtav, M., Zhidkov, I., (Feder), J.T., Dadon, S., Rubin, E., Glass, D., Spector, T.D., Mishmar, D.: Mitochondrial DNA heteroplasmy in diabetes and normal adults: role of acquired and inherited mutational patterns in twins. Human Molecular Genetics **21**(19), 4214–4224 (2012). Accessed 2014-09-23
18. Ye, K., Lu, J., Ma, F., Keinan, A., Gu, Z.: Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. Proceedings of the National Academy of Sciences **111**(29), 10654–10659 (2014). Accessed 2014-10-03
19. 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
20. 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
21. 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
22. Lascaro, D., Castellana, S., Gasparre, G., Romeo, G., Saccone, C., Attimonelli, M.: The RHNumtS compilation: Features and bioinformatics approaches to locate and quantify human NumtS. BMC Genomics **9**(1), 267 (2008). Accessed 2014-09-22
23. Calabrese, F.M., Simone, D., Attimonelli, M.: Primates and mouse NumtS in the UCSC genome browser. BMC Bioinformatics **13**(Suppl 4), 15 (2012). Accessed 2014-09-22
24. Calabrese, C., Simone, D., Diroma, M.A., Santorsola, M., Guttà, C., Gasparre, G., Picardi, E., Pesole, G., Attimonelli, M.: MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. Bioinformatics, 483 (2014). Accessed 2014-09-22
25. Bouhlal, Y., Martinez, S., Gong, H., Dumas, K., Shieh, J.T.C.: Twin mitochondrial sequence analysis. Molecular Genetics & Genomic Medicine **1**(3), 174–186 (2013). Accessed 2014-09-23
26. 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
27. 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
28. 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
29. Lang, B.F., Burger, G.: Purification of mitochondrial and plastid DNA. Nature Protocols **2**(3), 652–660 (2007). PMID: 17406627
30. 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
31. Picardi, E., Pesole, G.: Mitochondrial genomes gleaned from human whole-exome sequencing. Nature methods **9**(6), 523–524 (2012). PMID: 22669646
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. Schmitt, M.W., Kennedy, S.R., Salk, J.J., Fox, E.J., Hiatt, J.B., Loeb, L.A.: Detection of ultra-rare mutations by next-generation sequencing. Proceedings of the National Academy of Sciences **109**(36), 14508–14513 (2012). Accessed 2014-09-24
34. 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)
35. 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
36. 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
37. Connallon, T., Clark, A.G.: Balancing selection in species with separate sexes: Insights from fisher's geometric model. Genetics (2014)
38. Raap, A.K., Jahangir Tafrechi, R.S., van de Rijke, F.M., Pyle, A., Wählby, C., Szuhai, K., Ravelli, R.B.G., de Coo, R.F.M., Rajasimha, H.K., Nilsson, M., Chinnery, P.F., Samuels, D.C., Janssen, G.M.C.: Non-random mtDNA segregation patterns indicate a metastable heteroplasmic segregation unit in m.3243a>g cybrid cells. PLoS ONE **7**(12), 52080 (2012). Accessed 2014-10-05
39. Wallace, D.C., Chalkia, D.: Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. Cold Spring Harbor Perspectives in Biology **5**(11), 021220 (2013)
40. 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
41. Zhang, Q., Raoof, M., Chen, Y., Sumi, Y., Sursal, T., Junger, W., Brohi, K., Itagaki, K., Hauser, C.J.: Circulating mitochondrial DAMPs cause inflammatory responses to injury. Nature **464**(7285), 104–107 (2010). Accessed 2014-10-05
42. 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
43. Rogers, R.S., Bhattacharya, J.: When cells become organelle donors. Physiology (Bethesda, Md.) **28**(6), 414–422 (2013). PMID: 24186936
44. 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
45. 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
46. 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
47. Clayton, D.A.: Transcription and replication of mitochondrial DNA. Human reproduction (Oxford, England) **15 Suppl 2**, 11–17 (2000). PMID: 11041509
48. Benson, E.K., Mungamuri, S.K., Attie, O., Kracikova, M., Sachidanandam, R., Manfredi, J.J., Aaronson, S.A.: p53-dependent gene repression through p21 is mediated by recruitment of e2f4 repression complexes. Oncogene **33**(30), 3959–3969 (2014)
49. Quispe-Tintaya, W., White, R.R., Popov, V.N., Vijg, J., Maslov, A.Y.: Fast mitochondrial DNA isolation from mammalian cells for next-generation sequencing. BioTechniques **55**(3), 133–136 (2013)
50. Jukes, T.H., Osawa, S.: The genetic code in mitochondria and chloroplasts. Experientia **46**(11-12), 1117–1126 (1990). Accessed 2014-07-08

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, **Table ST1**) and 5 regions of mtDNA (**Table ST2**) 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) demonstrating uniformity of coverage. 1.23 million mtDNA reads and 50,000 nDNA reads implying $> 98\%$ pure mtDNA.

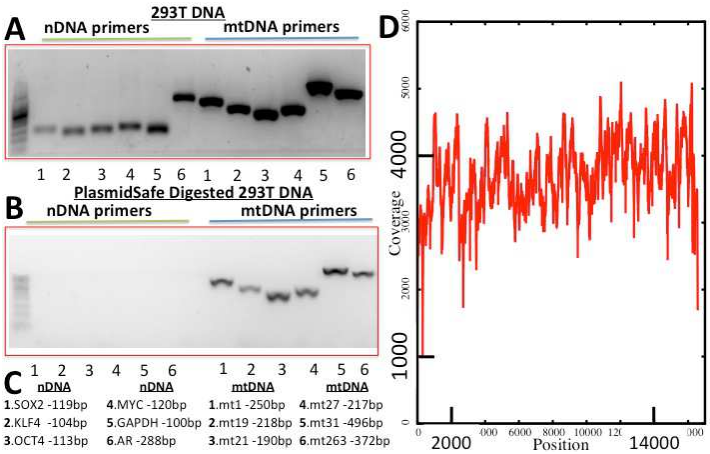

Figure 2 Variants in a cell-line (MDA-MB-157, circles) and a human (PBMC, crosses) compared to the reference mtDNA sequence. Mutation frequency (y-axis) versus position on mtDNA (x-axis). Genes are colored bands at bottom of graph ($+, -$ represent the strand, cds is coding sequence). *Neutral*, *Low*, *Medium*, and *High* are the effect of the mutation on protein function (*Mutation Assessor*[36]). 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. 3**. There are no stark differences between the human and cell-line derived mtDNA.

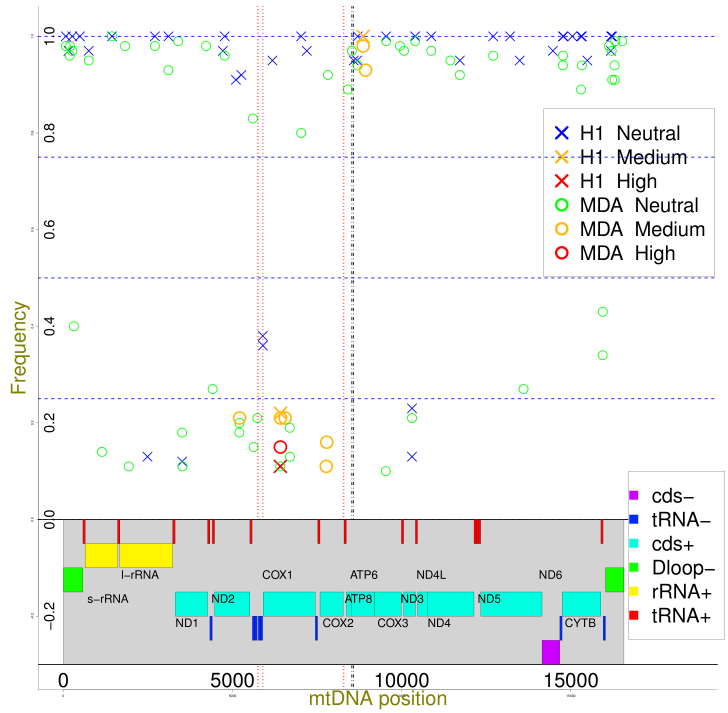

Figure 3 Stability of heteroplasmy. **A** C1 and C2 are colonies derived from single cells in the original colony, placing a severe botteneck on the mtDNA, and then passaged ≈ 25 times. **B** Mutation frequencies in C1 (x-axis) versus C2 (y-axis) for two cell-lines, U20S and MDA-MB-157. The mutations 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. 4**). The stability of the frequencies implies active mechanisms to counteract the drift, such as the exchange of mtDNA between cells.

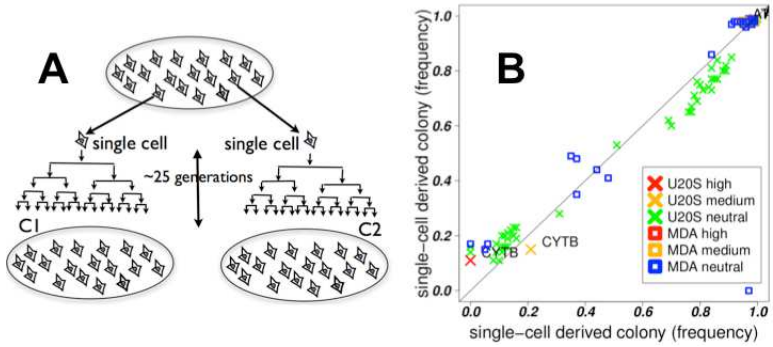

Figure 4 Simulations of mtDNA replication. Each cell contains a mixture of mtDNA haplotypes, here we consider only two species, red and blue, for the sake of simplicity. At cell division, the mitochondria assort randomly between the daughter cells and divide (with mtDNA replication) till a quorum of mitochondria, specific to each tissue-type, is reached. Quorum sensing is not well-understood, but the number of mitochondria per cell is tissue-specific[8]. 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.

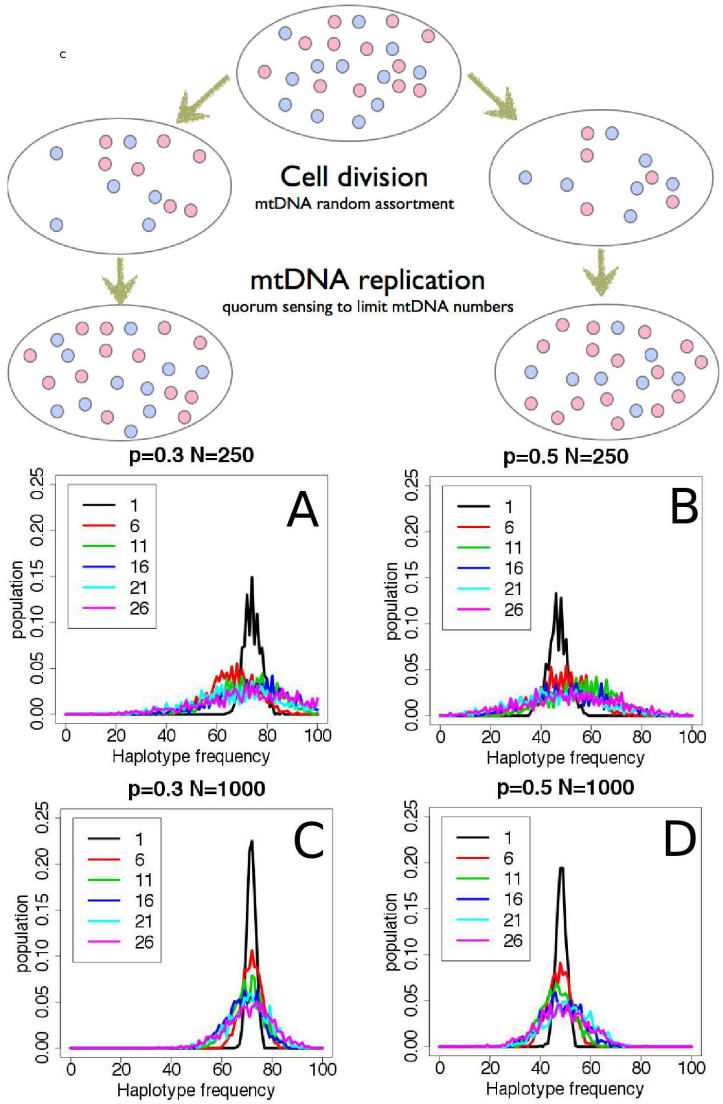

Figure 5 mtDNA transfer in co-cultured cells. Two normal fibroblast cell lines, WI-38 labeled with GFP and un-labeled IMR90 were co-cultured. The variants specific to each cell line before transfer and their frequency (Y-axis) is shown as red squares and blue triangles respectively. The variants are listed along the X-axis identified by the position on the mtDNA and the gene. After the co-culture, GFP negative cells were killed using the neomycin selection marker, which was incorporated with the GFP. The remaining GFP positive cells were regrown in culture and mtDNA variants sequenced (green circles). A clear transfer of variants from IMR90 to WI-38 cells is observed, as the green circles do not follow the pattern of the red squares. The variants on the X-axis are organized by the frequency of the green circles in the descending order.

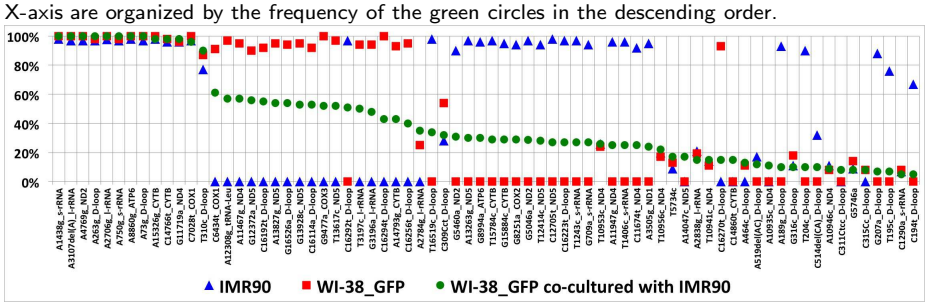

Figure S1 Organization of mtDNA. 13 protein-coding genes, 22 tRNA genes and 2 rRNA genes are encoded by a single circular nucleic acid and transcribed from three promoters: LSP (inner circle of genes - strand), HSP1 (outer circle of gene, + strand) and HSP2 (16S rRNA) on the D-loop, which is non-coding 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. The mitochondrial 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[50].

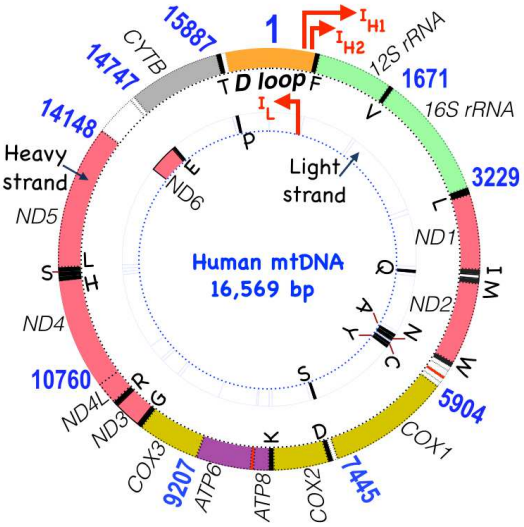

Table 1 Co-culturing experiments for WI-38-GFP(fibroblast) and IMR90 (fibroblast). The white rows are variants private to WI-38-GFP, gray rows show variants transferred from IMR90 to WI-38 and blue rows are variants common to IMR90 and WI-38. The green rows are variants private to IMR90 with minimal or no transfer. The frequency (*freq*) ranges from 0-1 and the coverage (*cov*) is the number of reads at the variant.

Variant	gene	AA	IMR90 control		WI-38-GFP control		WI-38-GFP mixed with IMR90		flank
			freq	cov	freq	cov	freq	cov	
A73g	D-loop		0.97	405	1	78	1	73	TGGGGGGT[A|g]TGCACGCG
A189g	D-loop		0.93	63	0	145	0.10	108	ACAGGCGA[A|g]CATACTTA
C194t	D-loop		0.67	31	0	140	0.05	107	CGAACATA[C|t]TTACTAAA
T195c	D-loop		0.76	30	0	143	0.07	104	GAACATAC[T|c]TACTAAAG
T204c	D-loop		0.90	62	0	132	0.10	104	TACTAAAG[T|c]GTGTTAAT
G207a	D-loop		0.88	63	0	133	0.07	104	TAAAGTGT[G|a]TTAATTAA
A263g	D-loop		0.97	335	0.98	63	1	58	GCACAGCC[A|g]CTTTCCAC
C309Cct	D-loop		0.28	280	0.54	81	0.32	64	AACCCCCC[C|Cct]TCCCCCGC
T310c	D-loop		0.77	272	0.87	79	0.90	62	ACCCCCCC[T|c]CCCCCGCT
C315Cc	D-loop		0	322	0.08	94	0.08	74	CCCTCCCC[C|Cc]GCTTCTGG
G316c	D-loop		0.11	303	0.18	95	0.10	69	CCTCCCCC[G|c]CTTCTGGC
A464c	D-loop		0	715	0.11	191	0.13	127	CCCCTCCC[A|c]CTCCCATA
C514del(CA)	D-loop		0.32	306	0	118	0.10	109	CTACCCAG[C|del(CA)]ACACACAC
A519del(AC)	D-loop		0.17	317	0	123	0.12	111	CAGCACAC[A|del(AC)]CACACCGC
G709a	s-rRNA		0.94	226	0	111	0.27	113	GCATCCCC[G|a]TTCCAGTG
A750g	s-rRNA		0.97	202	0.98	68	1	63	TCAAAAGG[A|g]ACAAGCAT
T1243c	s-rRNA		0.97	298	0	126	0.27	118	TCACCACC[T|c]CTTGCTCA
C1290a	s-rRNA		0.07	571	0.08	126	0.05	102	GATGAAGG[C|a]TACAAAGT
T1406c	s-rRNA		0.96	179	0	110	0.25	93	TATGAAAC[T|c]TAAGGGTC
A1438g	s-rRNA		0.98	170	1	58	1	50	AGTAAACT[A|g]AGAGTAGA
A2706g	l-rRNA		0.976	250	1	69	1	55	GCGGGCAT[A|g]ACACAGCA
A2784g	l-rRNA		0	421	0.25	127	0.35	108	GGTCCTAA[A|g]CTACCAAA
A2838g	l-rRNA		0.21	507	0.192	125	0.15	97	AGAACCCA[A|g]CCTCCGAG
A3107del(A)	l-rRNA		0.97	201	1	46	1	42	CTATCTAC[A|del(A)]TTCAAATT
G3196a	l-rRNA		0	507	0.94	56	0.48	94	TCAACTTA[G|a]TATTATAC
T3197c	l-rRNA		0	510	0.94	56	0.5	94	CAACTTAG[T|c]ATTATACC
A3505g	ND1	T->A	0.95	318	0	153	0.24	132	CCACATCT[A|g]CCATCACC
A4769g	ND2	M->M	0.97	233	1	60	1	50	ATCATAAT[A|g]GCTATAGC
G5046a	ND2	V->I	0.97	326	0	120	0.288	125	TAATAGCA[G|a]TTCTACCG
G5460a	ND2	A->T	0.90	407	0	113	0.31	123	CACTCATC[G|a]CCCTTACC
T5734c	tRNA-Asn-tRNA-Cys		0.09	731	0.13	174	0.17	116	TCTACTTC[T|c]CCCGCCGC
G5746a	tRNA-Asn-tRNA-Cys		0.09	650	0.14	171	0.08	118	GCCGCCGG[G|a]AAAAAAGG
C6434t	COX1	T->T	0	494	0.91	83	0.61	107	GCCATAAC[C|t]CAATACCA
C7028t	COX1	A->A	0.97	284	1	73	0.96	53	GTTGTAGC[C|t]CACTTCCA
G8251a	COX2	G->G	0.94	247	0	107	0.29	112	GAAATAGG[G|a]CCCGTATT
A8860g	ATP6	T->A	0.98	296	1	81	1	57	GAGCGGGC[A|g]CAGTGATT
G8994a	ATP6	L->L	0.96	269	0	100	0.30	115	ATAGCCCT[G|a]GCCGTACG
G9477a	COX3	V->I	0	469	1	58	0.52	97	CCTCAGAA[G|a]TTTTTTTC
T10941c	ND4	L->P	0.14	1685	0.11	384	0.15	324	CGACCCCC[T|c]AACAACCC
A10946c	ND4	T->P	0.11	1638	0.08	357	0.09	296	CCCTAACA[A|c]CCCCCCTC
T10953c	ND4	L->P	0.26	1395	0.24	297	0.26	225	AACCCCCC[T|c]CCTAATAC
T10956c	ND4	L->P	0.20	1302	0.17	281	0.22	208	CCCCCTCC[T|c]AATACTAA
A11467g	ND4	L->L	0	516	0.95	60	0.57	97	GTACTCTT[A|g]AAACTAGG
C11674t	ND4	T->T	0.92	267	0	98	0.25	108	ATCCAAAC[C|t]CCCTGAAG
G11719a	ND4	G->G	0.96	235	0.96	62	0.98	65	GCCCACGG[G|a]CTTACATC
A11947g	ND4	T->T	0.96	319	0	135	0.25	127	CTACTTAC[A|g]GGACTCAA
A12308g	tRNA-Leu		0	591	0.97	79	0.57	102	GGCCCCAA[A|g]AATTTTGG
G12372a	ND5	L->L	0	683	0.90	95	0.56	112	CTAACCCT[G|a]ACTTCCCT
T12414c	ND5	P->P	0.94	353	0	140	0.28	133	GTTAACCC[T|c]AACAAAAA
C12705t	ND5	I->I	0.98	304	0	114	0.27	121	CTACTCAT[C|t]TTCCTAAT
A13263g	ND5	Q->Q	0.97	253	0	110	0.30	112	TCAAGTCA[A|g]CTAGGACT
T13617c	ND5	I->I	0	561	0.97	72	0.52	100	CGAATAAT[T|c]CTTCTCAC
A13827g	ND5	G->G	0	533	0.95	83	0.54	111	TTCCTAGG[A|g]CTTCTAAC
G13928c	ND5	S->T	0	593	0.95	70	0.53	103	CTACCCTA[G|c]CATCACAC
C14766t	CYTB	T->I	0.96	464	0.98	56	0.98	52	ACGCAAAA[C|t]TAACCCCC
A14793g	CYTB	H->R	0	474	0.93	45	0.43	81	AATTAACC[A|g]CTCATTCA
A15326g	CYTB	T->A	0.98	309	1	83	0.98	66	CCCTAGCA[A|g]CACTCCAC
T15784c	CYTB	P->P	0.97	339	0	95	0.29	110	AGCTACCC[T|c]TTTACCAT
G15884c	CYTB	A->P	0.95	309	0	122	0.29	117	TCAAATGG[G|c]CCTGTCCT
C16114a	D-loop		0	662	0.92	79	0.53	107	GCCAGCCA[C|a]CATGAATA
C16192t	D-loop		0	601	0.92	77	0.55	100	CCCCCTCC[C|t]CATGCTTA
C16223t	D-loop		0.97	283	0	110	0.27	100	AATCAACC[C|t]TCAACTAT
C16256t	D-loop		0	520	0.95	47	0.40	72	CTCCAAAG[C|t]CACCCCTC
C16270t	D-loop		0	480	0.93	16	0.15	52	CTCACCCA[C|t]TAGGATAC
C16292t	D-loop		0.97	338	0	40	0.51	64	AACCTACC[C|t]ACCCTTAA
C16294t	D-loop		0	382	1	40	0.43	69	CCTACCCA[C|t]CCTTAACA
T16519c	D-loop		0.98	258	0	88	0.34	108	CTTCAGGG[T|c]CATAAAGC
G16526a	D-loop		0	316	0.94	75	0.54	102	GTCATAAA[G|a]CCTAAATA

Table 2 **Data from mixing experiments of MDA-MB-157 cells co-cultured with HCC1806 cells for 6 weeks. The MDA-MB-157 cells were labeled with GFP and made up approximately 10% of the mixture, which was passaged approximately 25 times before the component cells were separated using FACS. The mtDNA was sequenced to identify variants in the GFP and non-GFP cells. The private_to column identifies the cell-lines that exhibit the variant. Rows highlighted in gray are cases where a variant unique to HCC1806 has been identified in MDA-MB-157 cells. The light green rows are variants private to HCC1806 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 6), the A from the reference mtDNA genome is mutated to a T only in HCC1806, the MDA-MB-157 cultured with HCC1806 exhibits an A, with a frequency of 0.14 (or 14%). The frequency is $freq$ (which ranges from 0-1) and the coverage, number of reads covering the variant, is cov.**

variant	gene	AA	private_to	HCC1806 control		MDA-MB-157 from HCC1806		flank
				freq	cov	freq	cov	
G3666A	ND1	G120G	HCC1806	0.93	33	0	25	TCTGATCAGG[G|a]TGAGCATCAA
A9545G	COX3	G113G	HCC1806	0.93	15	0	13	CCCAATTAGG[A|g]GGGCACTGGC
A12810G	ND5	W158W	HCC1806	0.92	25	0	20	AATCCTATAC[A|g]ACCGTATCGG
C14911T	CYTB	Y55Y	HCC1806	0.96	29	0	39	CTATATTACG[C|t]ATCATTTCTC
C16187T	D-loop		HCC1806	1	9	0	31	ATCAAAACCC[C|t]CTCCCCATGC
A3796T	ND1	T163S	HCC1806	0.88	9	0.14	21	TAACCTCTCC[A|t]CCCTTATCAC
A4104G	ND1	L266L	HCC1806	0.9	21	0.06	31	TTCTAACCTC[A|g]CTGTTCTTAT
T7861C	COX2	D92D	HCC1806	0.96	29	0.05	39	CGGACTAATC[T|c]TCAACTCCTA
G9064A	ATP6	A179T	HCC1806	1	10	0.16	24	CCATTAACCT[G|a]CCCTCTACAC
A9072G	ATP6	S182S	HCC1806	0.88	9	0.18	22	CTTCCCTCTA[A|g]ACTTATCATC
G10688A	ND4L	V73V	HCC1806	0.96	26	0.1	19	ACCTGACTCC[G|a]ACCCCTCACA
C7819A	COX2	L78L	MDA-MB-157	0	35	0.76	43	CGCATCCTTT[C|a]CATAACAGAC
C8932T	ATP6	P135S	MDA-MB-157	0	25	0.9	21	CCATACTAGT[C|t]ATTATCGAAA
T13602C	ND5	Y422Y	MDA-MB-157	0	22	0.45	24	CAAGCGCCTA[T|c]AGCACTCGAA
T15514C	CYTB	Y256Y	MDA-MB-157	0	12	0.89	28	CAGACAATTA[T|c]ACCCTAGCCA
T16209C	D-loop		MDA-MB-157	0	17	0.92	26	TACAAGCAAG[T|c]ACAGCAATCA
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
T9540C	COX3	L111L	Both	0.86	15	1	11	TACCCCCCAA[T|c]TAGGAGGGCA
C16223T	D-loop		Both	0.95	20	0.96	31	GCAATCAACC[C|t]TCAACTATCA
T16311C	D-loop		Both	0.75	12	0.8	20	CAGTACATAG[T|c]ACATAAAGCC

Figure S2 Numts (Nuclear-mtDNA pseudogenes). The x-axis is the position along the mtDNA. The graph on the top shows the number of matches on the nuclear genome (hg38) of 36 nt tiles from the mtDNA. The graph in the bottom shows the changes in the mapping numbers for the tiles between hg38 and hg19, positive numbers are an increase in matches, while negative numbers are a decrease.

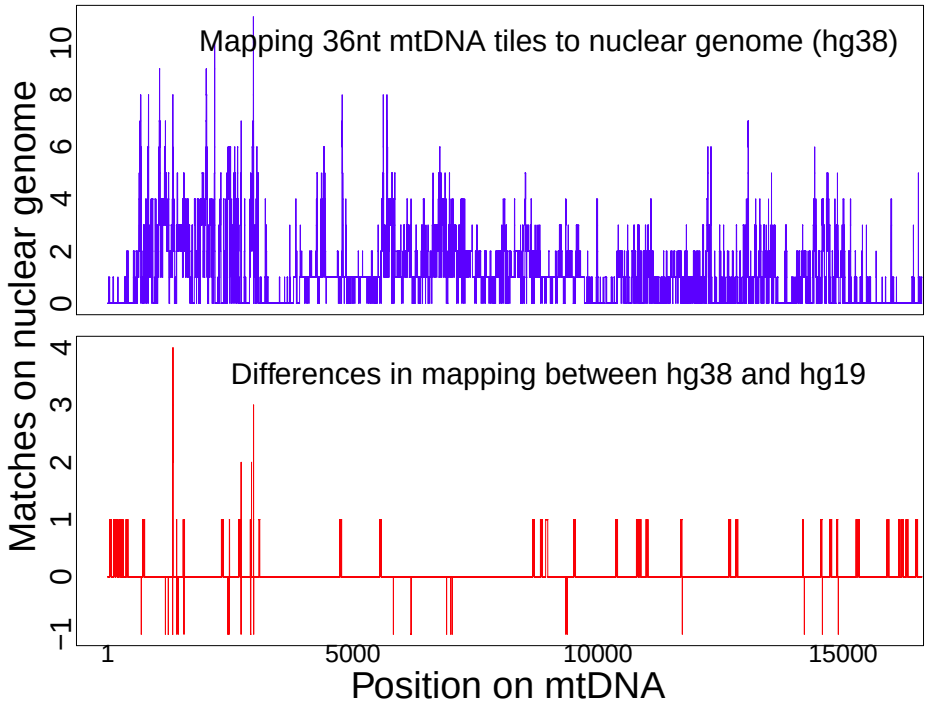

Figure S3 Numts (Nuclear-mtDNA pseudogenes). The x-axis is the position along the mtDNA. The graph on the top shows the number of matches on the nuclear genome (hg38) of 100nt tiles from the mtDNA. The graph in the bottom shows the changes in the mapping numbers for the tiles between hg38 and hg19, positive numbers are an increase in matches, while negative numbers are a decrease. Even at 100nt, there are a number of Numts in the reference genome.

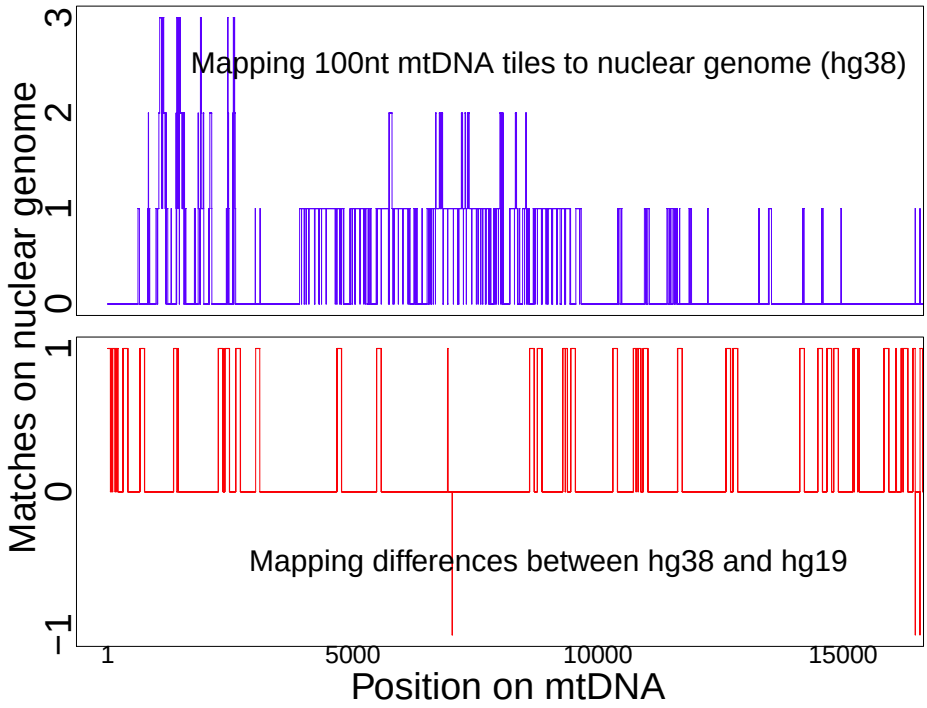

Figure S4 The Mseek protocol.

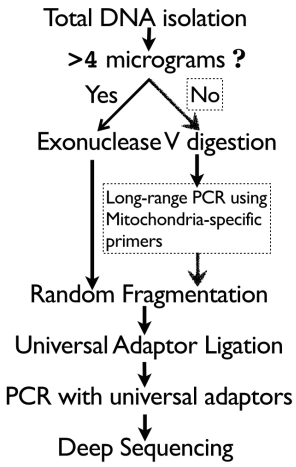

Figure S5 Mseek comparisons to other techniques. Plasmid mini-prep (panel A) and MACS (panel B) show a lack of purification of mtDNA based on PCR products run on a gel, thus they were not prepared for sequencing. Panel C shows sequencing data from REPLI-g (Qiagen) compared to Mseek, both of which show purification of mtDNA on the gel. **A** Plasmid mini prep followed by bead purification. PCR using primers specific to 18S(1), 28S(2), B-actin(3), mtDNA(4), mtDNA(5) show nDNA is present after treatment. **B** MACS. PCR using primers specific to 18S(1), 28S(2), B-actin(3), GAPDH(4), mtDNA(5), mtDNA(6) show that nDNA is present after prep.**C** Deep-sequencing data. Mseek (red bars) exhibits consistently better enrichment of mtDNA compared to REPLI-g (green bars) and untreated total DNA (blue bars). This comparison was done in triplicates, each using 8 μg of total DNA from the same blood sample. The purity of mtDNA from Mseek (80%) is lower compared to results presented in the text from cell lines, due to various factors in blood which requires further fine-tuning.

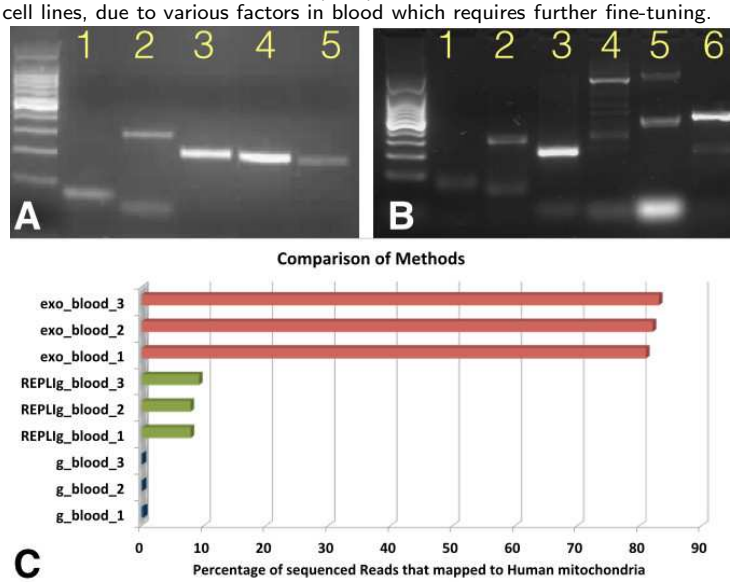

Table ST1 Primers specific to human nuclear DNA.

gene	forward primer	reverse primer	Product size
hSox2	TTTTGTCGGAGACGGAGAAG	CATGAGCGTCTTGGTTTTCC	119 bp
hKlf4	ACCCTGGGTCTTGAGGAAGT	AGGAAGGATGGGTAATTGGG	104 bp
hOct4	GTGGAGAGCAACTCCGATG	TTGATGTCCTGGGACTCCTC	113 bp
hMyc	AAGGACTATCCTGCTGCCAA	CCTCTTGACATTCTCCTCGG	120 bp
hGapdh	CTCTGCTCCTCCTGTTCGAC	AATCCGTTGACTCCGACCTT	345 bp
FMR1	GCTCAGCTCCGTTTCGGTTTCACTTCCGGT	AGCCCCGCACTTCCACCACCAGCTCCTCCA	281 bp
AR	TCCAGAATCTGTTCCAGAGCGTGC	GCTGTGAAGGTTGCTGTTCCTCAT	288 bp
18S	GCAATTATTCCCCATGAACG	GGGACTTAATCAACGCAAGC	68 bp

Table ST2 Primers specific to human mtDNA. The last row is the primer pair for amplifying the whole mtDNA using long-range PCR.

names	forward primer	reverse primer	Product size (bp)
mt pair.1	CTATCACCCTATTAACCACTCA	TTCGCCTGTAATATTGAACGTA	179
mt pair.21	AATCCAAGCCTACGTTTTCACA	AGTATGAGGAGCGTTATGGAGT	190
mt pair.31	TAACCCAAGTCAATAGAAGCCG	TAAGTTTCATAAGGGCTATCGT	497
mt pair.263	ATCTATTACTCTCATCGCTACC	TGTGATGCTAGGGTAGAATCCG	373
mt pair.19	AATCGAGTAGTACTCCCGATTG	TTCTAGGACGATGGGCATGAAA	218
mt pair.27	TCATCCCTGTAGCATTGTTCGT	TCGGGCGTATCATCAACTGATG	217
long-range	TCGGGCGTATCATCAACTGATG	CAGATGCCAACACAGCAGCCAT	16,568

Table ST3 Transitions versus Transversions for data from IMR90, WI-38 and MDA-MB-157. Of a total of 84 single base mutations there are only 6 transversions, the rest are transitions. The row labels are the original reference base, the column labels are the variant bases, so G to A changes happened 15 times. Most mutations were transitions (Table ST3), suggesting that the mutations don't arise from oxidative stress, and are most likely driven by the errors introduced by polymerase-γ activity[33].

	a	**c**	**g**	**t**
A	0	1	23	0
C	2	0	0	17
G	15	3	0	0
T	0	23	0	0

Table ST4 Co-culturing experiments for MDA-MB-157-GFP(cancer) with IMR90(fibroblast) The white rows are variants private to MDA-MB-157-GFP, gray rows are variants that transferred from IMR90 to MDA-MB-157-GFP and blue rows are variants common to IMR90 and MDA-MB-157-GFP. There were no transfers in this case. The frequency is *freq* (which ranges from 0-1) and the coverage, number of reads covering the variant, is *cov*.

Variant	gene	AA	IMR90 control		MDA-MB-157-GFP control		MDA-MB-157-GFP mixed with IMR90		flank
			freq	cov	freq	cov	freq	cov	
A73g	D-loop		0.97	405	0.97	181	0.95	286	TGGGGGGT[A|g]TGCACGCG
A189g	D-loop		0.93	63	0.93	62	0.94	169	ACAGGCGA[A|g]CATACTTA
C194t	D-loop		0.67	31	0	79	0	212	CGAACATA[C|t]TTACTAAA
T195c	D-loop		0.76	30	0	78	0	212	GAACATAC[T|c]TACTAAAG
A200g	D-loop		0	75	0.96	58	0.95	173	TACTTACT[A|g]AAGTGTGT
T204c	D-loop		0.90	62	0	81	0	222	TACTAAAG[T|c]GTGTTAAT
G207a	D-loop		0.88	63	0	85	0	234	TAAAGTGT[G|a]TTAATTAA
A263g	D-loop		0.97	335	0.99	146	0.97	246	GCACAGCC[A|g]CTTTCCAC
C309Cct	D-loop		0.28	280	0	98	0	122	AACCCCCC[C|Cct]TCCCCCGC
T310c	D-loop		0.77	272	0.82	95	0.83	116	ACCCCCCC[T|c]CCCCCGCT
C311Ctcc	D-loop		0	294	0.19	105	0.12	138	CCCCCCCT[C|Ctcc]CCCCGCTT
G316c	D-loop		0.11	303	0.05	115	0.06	147	CCTCCCCC[G|c]CTTCTGGC
C514del(CA)	D-loop		0.32	306	0	174	0	318	CTACCCAG[C|del(CA)]ACACACAC
A519del(AC)	D-loop		0.17	317	0	177	0	325	CAGCACAC[A|del(AC)]CACACCGC
G709a	s-rRNA		0.94	226	0	136	0	260	GCATCCCC[G|a]TTCCAGTG
A750g	s-rRNA		0.97	202	0.97	90	0.96	130	TCAAAAGG[A|g]ACAAGCAT
T1243c	s-rRNA		0.97	298	0	147	0	313	TCACCACC[T|c]CTTGCTCA
C1290a	s-rRNA		0.07	571	0.11	157	0.08	320	GATGAAGG[C|a]TACAAAGT
T1406c	s-rRNA		0.96	179	0	126	0	240	TATGAAAC[T|c]TAAGGGTC
A1438g	s-rRNA		0.98	170	0.98	74	0.98	122	AGTAAACT[A|g]AGAGTAGA
T1822c	l-rRNA		0	673	0.98	111	0.93	170	AAGCATAA[T|c]ATAGCAAG
A2706g	l-rRNA		0.976	250	0.96	92	0.96	149	GCGGGCAT[A|g]ACACAGCA
A2838g	l-rRNA		0.21	507	0.16	156	0.26	294	AGAACCCA[A|g]CCTCCGAG
A3107del(A)	l-rRNA		0.97	201	1	59	0.98	101	CTATCTAC[A|del(A)]TTTCAAATT
T3396c	ND1	Y->Y	0	477	0.975	80	0.93	148	CTAGGCTA[T|c]ATACAACT
A3505g	ND1	T->A	0.95	318	0	189	0	395	CCACATCT[A|g]CCATCACC
T4218c	ND1	Y->Y	0	747	0.952	125	0.94	198	ATATGATA[T|c]GTCTCCAT
G4412a	tRNA-Met		0	489	0.35	175	0.32	272	TAAGGTCA[G|a]CTAAATAA
A4769g	ND2	M->M	0.97	233	0.99	101	0.97	139	ATCATAAT[A|g]GCTATAGC
G5046a	ND2	V->I	0.97	326	0	187	0	331	TAATAGCA[G|a]TTCTACCG
G5460a	ND2	A->T	0.90	407	0	171	0	340	CACTCATC[G|a]CCCTTACC
C5601t	tRNA-Ala		0	503	0.94	92	0.94	153	CTGCAAAA[C|t]CCCACTCT
T5734c	tRNA-Asn-tRNA-Cys		0.09	731	0.11	230	0.13	436	TCTACTTC[T|c]CCCGCCGC
G5746a	tRNA-Asn-tRNA-Cys		0.09	650	0.09	232	0.11	381	GCCGCCGG[G|a]AAAAAAGG
A6692del(A)	COX1	G->	0	594	0.17	202	0.16	280	TACTCCGG[A|del(A)]AAAAAAGA
A6696del(A)	COX1	K->	0	602	0.17	218	0.16	302	CCGGAAAA[A|del(A)]AAGAACCA
C7028t	COX1	A->A	0.97	284	0.95	112	0.97	149	GTTGTAGC[C|t]CACTTCCA
C7819a	COX2	L->L	0	535	0.92	99	0.95	167	ATCGCCCT[C|a]CCATCCCT
G8251a	COX2	G->G	0.94	247	0	166	0	279	GAAATAGG[G|a]CCCGTATT
C8410t	ATP8	P->P	0	720	0.97	122	0.91	222	ATTACCCC[C|t]ATACTCCT
A8527g	ATP8	K->K	0	610	0.94	84	0.94	180	GAACCAAA[A|g]TGAACGAA
A8527g	ATP6	M->V	0	610	0.94	84	0.94	180	GAACCAAA[A|g]TGAACGAA
A8701g	ATP6	T->A	0	832	0.92	128	0.94	223	AAATGATA[A|g]CCATACAC
A8860g	ATP6	T->A	0.98	296	1	93	0.95	176	GAGCGGGC[A|g]CAGTGATT
C8932t	ATP6	P->S	0	613	0.96	85	0.91	160	CACCTACA[C|t]CCCTTATC
T9540c	COX3	L->L	0	461	0.99	111	0.99	161	CCCCCCAA[T|c]TAGGAGGG
T9950c	COX3	V->V	0	482	0.90	93	0.86	154	GATGTGGT[T|c]TGACTATT
C10070t	ND3	A->A	0	492	0.95	94	0.92	121	AACTTCGC[C|t]TTAATTTT
A10398g	ND3	T->A	0	553	0.94	108	0.97	139	TAGACTGA[A|g]CCGAATTG
T10873c	ND4	P->P	0	674	0.98	113	0.96	212	ATCATCCC[T|c]CTACTATT
T10941c	ND4	L->P	0.14	1685	0.15	614	0.13	1069	CGACCCCC[T|c]AACAACCC
A10946c	ND4	T->P	0.11	1638	0.07	579	0.09	1027	CCCTAACA[A|c]CCCCCCTC
T10953c	ND4	L->P	0.26	1395	0.33	479	0.27	873	AACCCCCC[T|c]CCTAATAC
T10956c	ND4	L->P	0.20	1302	0.29	454	0.22	816	CCCCCTCC[T|c]AATACTAA
G11440a	ND4	G->G	0	518	0.92	122	0.88	189	ATCGCTGG[G|a]TCAATAGT
C11674t	ND4	T->T	0.92	267	0	139	0	302	ATCCAAAC[C|t]CCCTGAAG
G11719a	ND4	G->G	0.96	235	0.98	94	0.95	141	GCCCACGG[G|a]CTTACATC
A11947g	ND4	T->T	0.96	319	0	200	0	397	CTACTTAC[A|g]GGACTCAA
T12414c	ND5	P->P	0.94	353	0	213	0	448	GTTAACCC[T|c]AACAAAAA
C12705t	ND5	I->I	0.98	304	0.99	116	0.96	160	CTACTCAT[C|t]TTCCTAAT
A13263g	ND5	Q->Q	0.97	253	0	180	0	336	TCAAGTCA[A|g]CTAGGACT
T13602c	ND5	Y->Y	0	565	0.38	180	0.31	316	AGCGCCTA[T|c]AGCACTCG
C14766t	CYTB	T->I	0.96	464	0.98	91	0.98	205	ACGCAAAA[C|t]TAACCCCC
A14769g	CYTB	N->S	0	494	0.94	91	0.91	202	CAAAACTA[A|g]CCCCCTAA
G15301a	CYTB	L->L	0	512	0.95	48	0.87	94	TTCATCTT[G|a]CCCTTCAT
A15326g	CYTB	T->A	0.98	309	0.98	58	0.98	92	CCCTAGCA[A|g]CACTCCAC
T15784c	CYTB	P->P	0.97	339	0	182	0	333	AGCTACCC[T|c]TTTACCAT
G15884c	CYTB	A->P	0.95	309	0	183	0	350	TCAAATGG[G|c]CCTGTCCT
T15940del(T)	tRNA-Thr		0	668	0.32	107	0.37	174	TGAAAACC[T|del(T)]TTTTCCAA
T15944del(T)	tRNA-Thr		0	654	0.45	100	0.34	173	AACCTTTT[T|del(T)]CCAAGGAC
G16129a	D-loop		0	681	0.95	93	0.92	202	TATTGTAC[G|a]GTACCATA
T16209c	D-loop		0	456	0.95	60	0.91	111	CAAGCAAG[T|c]ACAGCAAT
C16223t	D-loop		0.97	283	1	53	0.95	123	AATCAACC[C|t]TCAACTAT
C16292t	D-loop		0.97	338	1	37	0.93	64	AACCTACC[C|t]ACCCTTAA
C16295t	D-loop		0	393	0.94	38	0.93	65	CTACCCAC[C|t]CTTAACAG
T16311c	D-loop		0	457	0.98	51	0.91	80	GTACATAG[T|c]ACATAAAG
T16519c	D-loop		0.98	258	1	87	0.96	160	CTTCAGGG[T|c]CATAAAGC

Table ST5 Data from mixing experiments of U20S-GFP co-cultured with A382 for 6 weeks. The column private_to identifies the cell-lines that exhibit the variant. The light green rows are variants private to A382 that did not transfer into U20S. The frequency is *freq* (which ranges from 0-1) and the coverage, number of reads covering the variant, is *cov*.

mut	gene	AA	private_to	U20S-GFP from A382 freq	cov	flank
A8701G	ATP6	T58A	A382	0	121	ACAAATGATA[A|g]CCATACACAA
T9540C	COX3	L111L	A382	0	65	TACCCCCCAA[T|c]TAGGAGGGCA
G15301A	CYTB	L185L	A382	0	133	ACTTCATCTT[G|a]CCCTTCATTA
C16223T	D-loop		A382	0	168	GCAATCAACC[C|t]TCAACTATCA
T146C	D-loop		U20S	0.97	37	ATTCCTGCCT[T|c]ATCCTATTAT
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
A10398G	ND3	T113A	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
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