A Complete Workflow for scmtDNAseq in CHO cells, from Cell Culture to Bioinformatic Analysis

Alan Foley[1,3], Nga Lao[1], Colin Clarke[2,3], Niall Barron[1,3]

[1]Cell Engineering Group, NIBRT, Dublin, Ireland

[2]Bioinformatics Group, NIBRT, Dublin, Ireland

[3]School of Chemical and Bioprocess Engineering, University College Dublin, Ireland.

*** Correspondence:**
niall.barron@nibrt.ie

Keywords: Chinese hamster ovary, CHO, Long range PCR, Single cell, Mitochondrial DNA, mtDNA, Single cell mitochondrial DNA sequencing, scmtDNAseq

1 Abstract

Chinese Hamster Ovary (CHO) cells have a long history in the biopharmaceutical industry and currently produce the vast majority of recombinant therapeutic proteins. The key step in controlling process and product consistency is the development of a producer cell line derived from a single cell clone. However, it is recognised that genetic and phenotypic heterogeneity between individual cells in a clonal CHO population tends to arise over time. Previous bulk analysis of CHO cell populations has revealed considerable variation within the mtDNA sequence (heteroplasmy) which could have implications for the performance of the cell line. By analysing heteroplasmy of single cells within the same population, this heterogeneity can be characterised with greater resolution. Such analysis may identify heterogeneity in the mitochondrial genome which impacts the overall phenotypic performance of a producer cell population, and potentially reveal routes for genetic engineering. A critical first step is the development of robust experimental and computational methods to enable single cell mtDNA sequencing (termed scmtDNAseq). Here, we present a protocol from cell culture to bioinformatic analysis and provide preliminary evidence of significant mtDNA heteroplasmy across a small panel of single CHO cells.

2 Introduction

Chinese Hamster Ovary (CHO) cells are the most commonly used mammalian host for the production of recombinant proteins (Walsh et al., 2022). Optimisation of biopharmaceutical production in CHO has led to titers routinely in the 3-8 g/L range (Kelley et al., 2018). Due to their importance in energy production, understanding mitochondrial function in product-producing CHO cell lines is of particular importance. While most mitochondrial proteins are encoded by nuclear DNA, a small number are encoded by mitochondrial DNA (mtDNA). The CHO mitochondrial genome contains 37 genes, all of which support oxidative phosphorylation (OXPHOS). Thirteen protein-encoding subunits are accompanied by 2 rRNAs and 22 tRNAs in a 16,283bp plasmid-like circular structure (NCBI, 2023). mtDNA is highly compact, with the only significant non-coding region in the D-loop (**Fig. 1A**).

Assuming a CHO cell has typical numbers of mitochondria per cell (100-10,000), each with 2-10 copies of mtDNA, the total genome copy per cell is large (Dhiman et al., 2019). In 'homoplasmy' all copies of mtDNA within a cell are identical; however, mitochondria can also exist in a state of 'heteroplasmy' where mutated versions of mtDNA co-exist with wild-type mtDNA within the same cell; and possibly even within the same mitochondrion. When the proportion of mutant mtDNA is above a particular threshold, mitochondrial dysfunction can occur (Dimauro and Davidzon, 2005). In human disease, this could mean development of metabolic disease including neurodegenerative disorders (Keogh and Chinnery, 2015); in CHO cell culture it could manifest as a change in bioreactor performance.

Previous bulk analysis identified several levels of heteroplasmy between CHO cell lines (Kelly. P et al., 2017); laying a theoretical explanation for the metabolic heterogeneity often observed in CHO cell cultures (Gilbert et al., 2013). Single cell sequencing of mtDNA (scmtDNAseq) has previously been employed in non-CHO cell lines using high PCR cycle numbers of 40 (Zambelli et al., 2017) and 45 (Maeda et al., 2020). Higher PCR cycle numbers are associated with greater risk of undesirable secondary products such as PCR artefacts (Lorenz, 2012). In single cell mtDNA analysis, starting mtDNA copy number is low (<100,000) therefore even small contaminations can confound true mutation nucleotide identification. In fact, to call heteroplasmic mutations at 0.015 allele frequency (a conservative level) PCR amplification should ideally not exceed 30 cycles (Zambelli et al., 2017). Also, Maeda et al. focused on specific mutations, not the whole mtDNA genome, precluding identification of as yet unknown mutations. There is real value in novel whole mtDNA single cell analysis with a low PCR cycle number.

Here, we sought to develop an optimised method to amplify mtDNA and sequence from single cells. To demonstrate the method, we analysed four single CHO cells and a bulk (multiple cells) sample for comparison. Single cells were isolated by FACS into lysis buffer with an emphasis on simple and reproducible gating (**Fig. 1B**). After optimisation of the lysis buffer, PCR kit and purification system, long-range PCR (LRPCR) cycle number (35x) was kept lower than previously reported methods. Importantly, this provides more confidence in calling low-frequency heteroplasmy. To ensure exclusion of contaminating nuclear mitochondrial DNA (Numts), primers were designed to exclusively map to CHO mtDNA and amplicons size-selected via gel electrophoresis. By confirming mtDNA amplification by agarose gel, we were able to improve the efficiency of our sequencing – since only successful reactions were brought forward for library preparation. Illumina DNA libraries were generated and iSeq100-derived sequencing output was processed and analysed using a bespoke bioinformatics pipeline. Preprocessing was performed in Linux and data analysis in R.

Fig. 1

3 Materials and Equipment

2.1 CHO cell culture

1. 125mL bioreactor flasks (Nalgene 10266432).
2. Appropriate CHO cell culture medium (e.g. Gibco CD FortiCHO 10887640).

3. CHO cell lines of interest (e.g. **Table 1**).

2.2 Immunolabelling and Staining

1. DPBS
2. Nuclease-free water
3. Trypan Blue 0.4% (Gibco 15250061)
4. Luna II (or other appropriate cell counter)
5. DAPI (Invitrogen D1306)
6. Goat F(ab')2-Fluorescein anti-Human IgG (Sigma Aldrich SAB3701254-2MG) to label IgG-producing cells if desired. Other appropriate fluorescent stains could also be employed (e.g. CellTracker Green Invitrogen 11570166).

2.3 FACS

1. 70% IPA.
2. FACS with appropriate lasers for DAPI and FITC detection. Here, a BD FACS Melody was used.
3. FACS polystyrene tubes (Falcon Corning 1018640)
4. U-bottom 96-well plates (Corning 3799)
5. Parafilm
6. TCL Buffer (QIAGEN 1070498)

2.4 AMPure purification

1. AMPure XP Beads (10136224)
2. 70% ethanol
3. Elution buffer (QIAGEN 19086)
4. Sterile PCR tubes (autoclaved)
5. 0.2mL tube magnetic stand (New England Biolabs S1515S)

6. 10uL multichannel pipette (optional)

2.5 Long Range PCR

Primers were designed using NCBI Primer-BLAST to specifically bind to mtDNA, and not to any known CHO nuclear DNA sequences to minimise Numt contamination.

1. SuperFi II Plat Taq (Invitrogen 12361010).
2. PCR thermocycler
3. 10uM forward and reverse primers (**Table 2**) (IDT)
4. 10mM dNTP Mix (Thermo Scientific R0192)

2.6 Agarose Gel

1. Agarose powder.
2. TAE buffer.
3. SafeView (NBS Biologicals). Ethidium Bromide is an alternative.
4. GeneRuler 1kb Plus Ladder (Thermo Scientific SM1333).
5. Gel Viewer/transilluminator.
6. Disposable lab scalpel.
7. Eppendorf tubes.

2.7 Gel Purification

1. QIAquick Gel Extraction Kit (QIAGEN 28706X4). Other gel extraction kits could also be utilised.

2.8 Qubit

1. Qubit 4 Fluorometer (Invitrogen)
2. Qubit 1x dsDNA HS Kit (Invitrogen Q33230)

2.9 Sequencing

1. iSeq100 (Illumina) PE150, 8 million reads
2. Illumina DNA Prep, (M) Tagmentation (24 Samples) (Illumina 20018704)
3. IDT for Illumina DNA/RNA UD Indexes Set A, Tagmentation (96 Indexes, 96 Samples) (Illumina 20027213)
4. iSeq 100 i1 Reagent v2 (300-cycle) (Illumina 20031371)
5. PhiX v3 (Illumina FC-110-3001)

4 Methods

All steps up to the completion of the LRPCR for the 4 samples (**Table 1**) were performed in sterile conditions (BSC).

Table 1

4.1 PCR Component Storage

Since the LRPCR amplifies from less than 5,000 copies of mtDNA, PCR components must have optimal efficacy. This was ensured by making small (20uL) aliquots of dNTPs (Thermo Scientific R0192) and primers (IDT) and storing at -80°C. New aliquots were used for each lot of PCR performed and subsequently discarded.

Table 2

4.2 CHO cell culture

CHO-GS cells were cultured in FORTICHO (Gibco CD FortiCHO 10887640) at 37°C, 5% CO2, 85% humidity, 125rpm with 25mm orbit in a shaking incubator in 125mL bioreactor flasks (Nalgene 10266432). Every 3-4 days, cells were passaged at 0.2*10^5 cells/mL in 30mL media in 125mL culture shaking flasks. A growth curve was established to ensure samples were taken at exponential cell phase (**Table 1**).

4.3 DAPI Stain

A working concentration of 0.1ug/mL DAPI was determined as optimal for CHO cells. DAPI (Invitrogen D1306) solutions were protected from light wherever possible. In a BSC, 10mg DAPI powder was completely dissolved in 2mL sterile deionised water to make a 5 mg/mL DAPI stock solution. This was aliquoted and stored at -20°C. Solutions were stable for at least six months. 1uL of DAPI stock solution was added to 5mL DPBS for a 1ug/mL stock 2 DAPI working solution. 1mL of 1ug/mL stock 2 solution was added to 9mL of DPBS to prepare a 0.1ug/mL DAPI working solution.

4.4 Staining Cells

Here, an AB-FITC (Sigma Aldrich SAB3701254-2MG) conjugate was used which at 4°C can bind to IgG on the cell membrane in the process of being excreted by the cell as previously demonstrated for CHO cells (Gallagher and Kelly, 2017). This allowed the sorting of cells based on the productivity of an IgG-based antibody. Cell samples were prepared as per **Table 1**. Cells were counted using trypan blue (Gibco 15250061) and a hemacytometer as per the manufacturer's instructions. 1*10^6 of viable cells were centrifuged at 200 x g for 5 minutes and supernatant discarded. Cells were washed in 1mL DPBS, centrifuged at 200 x g for 5 minutes and supernatant discarded. This was repeated for a total of 2 washes. Cells were resuspended in 1mL of DPBS using 2uL of anti-human IgG (Sigma Aldrich SAB3701254-2MG). Cells were incubated at 4°C for 30 mins at 1000rpm, protected from light. Cells were washed twice with DPBS as per steps 6 & 7 for a total of 2 washes. Cells were resuspended in 1mL of cold DPBS or cold DAPI working solution, incubated on ice for 5 mins and immediately transferred on ice to the Fluoresence-activated cell sorting (FACS) lab for immediate analysis.

4.5 Setting Single cell Gating

The FACS Melody was setup as per manufacturer's instructions. A U-bottom 96-well plate was prepared (Corning 3799) with 5uL of 1x TCL buffer (QIAGEN 1070498) in the centre of each functional well using a multichannel pipette. The plate was tapped firmly on a flat surface to encourage the central location of the TCL buffer. The size threshold was set to >12um. Using Sample 4 (**Table 1**), voltages were set to allow the representation of cells in a SSC-A against FSC-A logarithmic scale graph. Gate 1 (G1) excluded instrument noise and cell debris as per **Fig. 2A**. Using Sample 4, data was brought forward and gate 2 (G2) set using FSC-H against FSC-A as per **Fig. 2B** to exclude doublets. Using Samples 1 and 2, the G2 gate was brought forward and a range gate (G3) was set to only include live cells as per **Fig. 2C** and **2.4D**. DAPI positive was considered dead cells. Using Samples 1 and 3, the G3 gate was brought forward and a gate (G4) set for FITC-positive cells as per **Fig. 2E** and **2F**. G4 was the sorting gate for live, singlet cells. After gates had been set, it was important to record data for a large number of events (e.g. 10,000 cells) and to save FCS files.

4.6 Single Cell Sort

The flow rate was kept at a minimum to reduce the chance of doublets. Sample 1 was loaded FACS set to "single cell" and "96-well plate" modes. Desired wells were selected for sorting with a splash shield present. The lid was removed and immediately inserted into the FACS to proceed with sorting. For the positive control, sort mode was changed to "purity". After the sort was complete, the well plate was removed and immediately covered with the lid. An airtight seal was created around the edges with parafilm and the plate immediately placed in a -80°C freezer. FCS files were saved for all samples.

STOPPING POINT: Samples can be stored for up to 6 months at -80°C.

4.7 AMPure purification

AMPure beads (10136224) benefit from scalable purification – adapting to single cell samples, volumes can simply be reduced. Sequences of mtDNA sometimes migrate and integrate into the nuclear genome – known as Nuclear mitochondrial sequences (Numts). In a previous bulk analysis of mtDNA, the miniprep step purified the plasmid-like mtDNA from contaminating linear nuclear DNA (Kelly et al., 2017). Here, AMPure purification was used, leaving both mtDNA and nDNA in the sample. Blast searching primer sequences against the CHO cell line reference genome (taxid: 10029) and gel purifying 8.5kb bands provided additional protection against Numts. All steps were performed in a BSC. The subsequent LRPCR is extremely sensitive and could potentially amplify small contaminations. The 96-well plate was thawed at room temperature. Multiple samples were taken through AMPure purification in batches (to a maximum of 12 samples). The 5uL lysed sample was transferred to a labelled micro-centrifuge tube. AMPure beads were resuspended by vortexing the bottle for 1 min. 9uL of AMPure beads were added per sample (if the lysed cell sample was greater, 1.8x volume of AMPure beads was used) and pipette mixed 10 times. They were left at room temperature for 5 mins. Tubes were placed on a magnetic stand (New England Biolabs S1515S) for 2 minutes. Keeping the tubes on the magnetic stand, the cleared solution was removed and discarded leaving the beads. It was then washed with 40uL 70% ethanol. The supernatant was discarded, leaving the beads. The ethanol wash was repeated. On the second wash, remaining ethanol was removed by using a P10 pipette while avoiding removing any beads. Tubes were removed from the magnetic stand and 18uL of elution buffer (QIAGEN 19086) was added to the bead aggregate and pipette mix 10 times or until fully resuspended. Tubes were incubated for 5 minutes at room temperature. Tubes were placed on the magnetic stand for 2 minutes. Eluate was split into two 8.5uL aliquots – leaving the bead aggregate. Microcentrifuge tubes were labelled to identify which samples came from the same single cell.

4.8 SuperFi II Plat Taq LRPCR

The bottleneck of single cell sequencing is the DNA amplification. Amplification techniques that would work for bulk sequencing proved to be incompatible with single cell: mechanical purifications took too much of the sample, bacterial lysis buffers did not release enough mtDNA, and components lost effectiveness for the sensitive PCR. Once enough DNA is amplified, established protocols for bulk sequencing can be followed (Kelly et al., 2017). The SuperFi II PCR kit (Invitrogen 12361010) has 300x fidelity compared to Plat Taq. SuperFi II was better able to amplify from small samples compared to Plat Taq. Higher fidelity also means greater confidence in lower-level heteroplasmy.

In addition to the below LRPCR protocol, single cell samples post AMPure purification were diluted to 1/10, 1/100, 1/1,000 and 1/10,000; to demonstrate the limits of the high fidelity LRPCR kit. All steps were performed in a BSC while maintaining samples at all steps on ice. Fresh aliquots of primers and dNTPs were thawed at room temperature then stored on ice. SuperFi II 5x Buffer was thawed and stored on ice. DNA Polymerase was maintained at -20°C and only removed briefly when needed. Components were briefly vortexed and centrifuged before use – except for the DNA Polymerase. mtDNA LRPCR was performed in 2 separate fragments (termed X and Y). The eluate from a single cell had been split into 2 from AMPure Purification; 1 half was amplified using X primers, the other half by Y primers (**Table 2**). A mastermix was generated with 10% overage, for each X primer and Y primer, as per the example in **Table 3**. SuperFi II DNA Polymerase was added last by briefly removing it from the -20°C freezer – to minimise the time spent at room temperature. The mastermix was gently vortexed, centrifuged at 500xg and kept on ice. 16.5uL mastermix was added to 8.5uL AMPure purified DNA. The sample was gently vortexed, centrifuged at 500xg and kept on ice. Samples were placed in a PCR machine and set to a PCR cycle as per **Table 4**. Reaction volume was set to 25uL with a lid temperature of 105°C. The cycle was run overnight. On completion, samples were removed and stored at 4°C.

STOPPING POINT: Samples can be stored at -20◦C for 2 weeks.

Table 3

Table 4

4.9 Agarose Gel

Limit of detection: an 8.5kb band was still observable when taking a 1/1000 dilution of a single cell. You would expect around 100-10,000 mitochondria (Dhiman et al., 2019). Theoretically, this PCR may be on the edge of viability for single-mitochondrial sequencing.

1g of agarose was added to 100mL TAE buffer in a conical flask and microwaved for 2.5 mins, or until fully dissolved. The flask was left to cool to about 50°C. 10uL of SafeView (NBS Biologicals) was added and the mixture poured into the gel tray with a well comb. After a brief period, the gel cooled and hardened at room temperature. The gel was placed in a gel box with TAE buffer just covering the gel. Loading dye was added to all samples as per the manufacturer's instructions. Entire samples were loaded into gel wells with an appropriate DNA ladder (Thermo Scientific SM1333). Gels were run at 100V until bands were 70% down the gel. The power was turned off and the gel was carefully placed in a gel viewer. Photos of the gel were taken.

4.10 Gel excision

Under a gel visualiser, 8.5kb bands were identified indicative of single cell reactions, as illustrated by red rectangles in **Fig. 2G**. UV light exposure was minimised to limit degradation of DNA. Using a new sterile disposable scalpel, the 8.5kb band was excised and placed in a 1.5mL Eppendorf tube. The blade was thoroughly cleaned with 70% IPA and then reused. The Y single cell sample was equally isolated and placed in a separate 1.5mL Eppendorf tube.

4.11 Gel purification

The QIAquick Gel Extraction protocol for "QIAquick Gel Extraction using a Microcentrifuge". 10uL of elution buffer was used to encourage a higher final concentration.
STOPPING POINT: Samples can be stored at -20°C for 2 weeks.

4.12 Equimolar combination

The Qubit 1x dsDNA HS kit (Invitrogen Q33230) was used to quantify dsDNA. Kit components were allowed to equilibrate to room temperature for 30 mins. 10uL of Standard 1 was added to a Qubit tube, and 10uL Standard 2 to a separate Qubit tube. 190uL of 1x buffer was added to each. 1uL of each X and Y fragments was added to the separate Qubit tubes and 199uL of 1x buffer was added to each. Tubes were vortexed for 2-3 seconds and left at room temperature for 2 mins. The concentration of standards 1 and 2 were measured using the Qubit (Invitrogen). The concentration of samples was measured using the Qubit. The volume required to aliquot 1ng of the X fragment and Y fragment from the same cell was calculated and these volumes combined in a new Lo-bind tube. There should be a total of 2ng of mtDNA from each single cell. Low concentrations were expected from our single cell samples.

4.13 Library Prep

The Illumina DNA Prep protocol was followed, using IDT for Illumina DNA/RNA UD Indexes Set A, Tagmentation (96 Indexes, 96 Samples) (Illumina 20027213). Each single cell should have a unique pair of indexes. The library quality of the cleaned-up library was checked by running 1uL on a Tapestation D5000 microwell. The libraries were combined and diluted to a 2nM starting concentration as per the manufacturer's instructions.
STOPPING POINT: Samples can be stored at -20°C for 30 days.

4.14 Sequencing

Libraries generated using Illumina DNA Prep were compatible with a wide range of Illumina sequencers including HiSeq, iSeq100, MiniSeq, NextSeq and NovaSeq technologies.

The iSeq cartridge and flow cell were prepared as per the manufacturer's instructions (Illumina 20031371). 2% PhiX (Illumina FC-110-3001) spike-in was added. The sample sheet loaded onto iSeq was checked to ensure correspondence to the sample sheet from Library Prep. The cartridge was loaded, and the run performed as per manufacturer's instructions. After running, the data was downloaded and backed-up on an external hard drive.

4.15 Data preprocessing

GitHub repository: https://github.com/alanfoleynibrt/SingleCellmtDNA

The bioinformatics pipeline is available in the above GitHub repository. Initial processing of data is performed in Linux and figures are generated in R. All raw FASTQ data analysed is made available in this pipeline. A step-by-step protocol and all materials are also available.

Briefly, trim_galore (0.4.3) used to trim adapter sequences in FASTQ files. Bowtie-2 (2.3.4.1) used to map reads to the KX576660.1 CHO mtDNA reference genome. Picard (1.199) tools identified duplicates (MarkDuplicates), added read groups (AddOrReplaceReadGroups) and built a BAM index (BuildBamIndex). Gatk3.8-0 implemented to realign indels (IndelRealigner) and recalibrate bases (BaseRecalibrator). Two separate mutation calling software programs were used: lofreq_star-2.1.2 and varscan.v2.3.9. When a mutation was called by both, it was brought forward for analysis. If a mutation allele frequency was between 0.04 and 0.96, it was considered "heteroplasmic". The potential impact of identified mutations was predicted using SnpEff. In tandem, analysis was repeated using a shifted mtDNA reference genome to complete coverage over the D-loop region. Unshifted mutation calls were concatenated with those from the shifted reference sequence to provide full coverage. ggplot2 in R was used to generate figures.

5 Results

5.1 Single Cell Isolation

The overall aim of this project was to create a workflow for single cell mtDNA analysis in CHO cells. We first sought to isolate single cells in a reproducible manner. Clonal populations of CHO cells are often generated using the "serial dilution" method whereby a known number of cells is progressively diluted to approximate a single cell per unit volume. However, the nature of cell distribution in each dilution means the final dispensed sample could indeed have a single cell, but it could also have 0 cells or multiple cells. We reasoned that a FACS-based method would be more accurate and reproducible. Additionally, there is less manual work when scaling to generate large numbers of clones. We loaded the sorter (FACS Melody, BD) with CHO-GS cells which had been stained with an AB-FITC conjugate and DAPI.

Fig. 2

We followed basic guidelines for flow cytometry (Bio-Rad, 2022). We first used an SSC-A against FSC-A dot plot with calibrated voltages (**Fig. 2A**). A gate of the main cell population was selected, excluding instrument noise and cell debris. We next used an FSC-H against FSC-A graph (**Fig. 2B**). Since forward scatter determines the "size" of particles, the "height" against the "area" determines the ratio of the cell height against the cell area. A singlet will have 1x area, 1x width and 1x height. Doublets would have 2x area, 2x width but 1x height. Thus to discriminate between singlets and doublets, the ratio of area to height is considered (Bio-Rad, 2022). We used a gate to select only singlets.

We next focused on FITC staining (due to our AB-FITC conjugate) and BV510-A (due to DAPI). DAPI increases in fluorescence when binding to DNA; therefore it has applications for live/dead cell gating in flow cytometry. Live cells would have lower BV510-A fluorescence, while dead cells would have greater BV510-A fluorescence. We first loaded a population of dead/dying CHO cells with 5% viability and viewed the population in a histogram of BV510-A fluorescence (**Fig. 2C**). We calibrated the BV510-A voltage to allow space for "lower" BV510-A fluorescence to which we set a 5% gate. Thus, only 5% of the dead/dying population was within our lower BV510-A fluorescence gate. We then ran live cells with 95% viability (**Fig. 2D**). As expected, the population migrated to our "live" gate, thus allowing us to select for live cells.

To identify mAb-secreting CHO cells, we exploited the AB-FITC staining to select for cells with a mAb in "stasis" in the CHO cell membrane. We first ran a sample with a non-producing CHO cell line and ran the population on a histogram of FITC fluorescence (**Fig. 2E**). If the stain specifically binds to mAb-producing CHO cells, there should be a shift of the population to greater FITC fluorescence from non-producing to producing. We therefore set a gate for greater FITC fluorescence with 0% of cells from the non-producing cell line. When we ran the same settings for our producer CHO cell line (**Fig. 2F**), we observed an overall "shift" of the population towards greater FITC fluorescence. This finalised our FACS-based method which selects for viable, singlet and mAb-producing CHO cells. Cells that fulfilled our gating strategy were sorted into wells of a 96-well plate with lysis buffer; choosing a U-bottom plate to encourage a central location for the 5uL of lysis buffer within each well.

5.2 DNA Purification

Previous bulk analysis of CHO mtDNA used mini-prep kits (QIAGEN) to enrich the plasmid-like mtDNA and reduce capture of the linear nuclear DNA (Kelly. P et al., 2017). This method proved to be inefficient for single cell samples whose mtDNA mass was much smaller. The physical filtration system required quenching meaning that low initial volumes, as with single cell samples, were lost. AMPure bead purification emerged as a viable alternative since adaptation to lower volume samples simply required volume reduction of all reagents. Here, purification was performed by adding AMPure beads to single cell samples, applying a magnet, washing with ethanol and eluting with elution buffer. However, caution is advised since nuclear DNA is also captured.

The miniprep kit had concomitantly provided some protection against Numts since it is designed to purify circular mtDNA away from linear nuclear DNA (Kelly. P et al., 2017). Having eliminated the miniprep step, we sought to incorporate additional protection against Numts. We performed a BLAST search of our mtDNA amplification primer sequences against the nuclear CHO reference genomes and found no matches, suggesting there are no nuclear sequences to which our primers should bind. Also, Numts tend to be shorter sequences, with 78% shorter than 500bp in human mtDNA (Wei et al. 2022). Therefore, we reasoned that specific gel purification of 8.5kb amplicons would be unlikely to be contaminated with Numts.

5.3 DNA Amplification

The biggest bottleneck for single cell mtDNA sequencing was the DNA amplification step. Adaptation of the DNA amplification centred on compatibility with as much as a million-fold lower starting DNA material compared to extracting from a standard cell population sample. We first selected a high-fidelity, long-range LR-PCR kit (SuperFi II. Invitrogen 12361010)). We then designed primers to amplify the mtDNA in two separate overlapping fragments, as previously done for bulk analysis (Kelly et al., 2017). The mini-prep kit, as previously used by Kelly et al., utilised a bacterial-specific lysis method. We anticipated that a mammalian-specific lysis buffer might liberate more mtDNA than a bacterial lysis buffer, accounting for the presence of both the outer cell membrane and the double-membrane of the mitochondria (**Fig. 1A**).

We found that PCR component storage and utilisation was a critical element of the protocol being successful. Immediately on component arrival, dNTPs and primers were diluted to the desired concentrations, aliquoted into microtubes and stored at -80°C. For each LRPCR reaction, a new aliquot was thawed, used and the aliquot discarded. As a precaution, we sorted 0 cells into a 96-well plate (**Fig. 2G**) to test for sources of contamination during the protocol that could lead to non-specific amplification. We did not observe any amplicons from these wells. We then applied the method to a single cell sample, and also to a 1000 cell sample (**Fig. 2G**). We observed successful amplification of both amplicons of mtDNA from both 1000 cells and a single cell.

5.4 Sample Generation and Library Preparation

Once amplification of mtDNA from single cells was achieved, we reasoned the later steps of library preparation, sequencing and bioinformatic analysis should be largely unchanged. To verify this, four single cells and a bulk sample (4000 cells) for comparison were sorted into 5uL of TCL lysis buffer. Samples were purified, split into two equal portions and separately amplified by LRPCR; which were then visualised on agarose gel. mtDNA-specific bands were excised for both amplified fragments. Amplicons were recovered by gel purification. After quantifying this dsDNA using the Qubit 1x dsDNA HD kit, equimolar quantities of each fragment from the same cell were added into a single tube.

For library preparation, the Illumina DNA Prep protocol (20018705) with IDT for Illumina DNA/RNA UD Indexes Set A was implemented. Each single cell was separately assigned indices to be used to demultiplex single cells later. Separate libraries were combined, diluted to 2nM and loaded onto the Illumina iSeq for PE150 sequencing. A 2% PhiX library control spike-in was added. The iSeq had the option to include a "sample sheet" to which the index combinations were added. The iSeq was run to completion with an output of fastq.gz files ready for the bioinformatics pipeline.

5.5 Bioinformatic Analysis

Adapter sequences were removed from the reads which were then mapped against the CHO KX576660.1 mtDNA reference genome. During the PCR step of the Illumina DNA library prep, adapter-ligated fragments are PCR amplified. This can lead to multiple sequencing reads deriving from the same original fragment; possibly resulting in overrepresentation of certain alleles. As was performed previously in bulk analysis of CHO mtDNA, PCR duplicates were identified (**Fig. 3A**) and removed from the analysis (Kelly et al. 2017). The range of duplicate reads was 23.8% to 29.3% with the highest proportion in the mixed population. Further, indel realignment and base recalibration were used to cater for the effects of INDELS on read mapping.

Fig. 3

After excluding duplicate and unmapped reads, all samples had an average sequencing depth of >1500x – above the 1000x required for “ultra-deep sequencing” categorisation (**Fig. 3B**). Perbase coverage of all samples confirmed complete and even mapping of sequencing reads (**Fig. 3C**). The mapping indicated no strong bias for any particular region. Together, this confirmed our scmtDNAseq protocol had been successful.

The great value of single cell sequencing mtDNA at such great depth is the potential to analyse at high confidence the differences in the sequenced reads when compared to the reference genome; i.e. a mutation. The multi-copy nature of mtDNA within each cell is represented by the proportion of reads with a particular mutation. E.g. If 50/100 reads contain a mutation, the allele frequency is determined as 0.5. This infers that out of all copies of mtDNA within that cell, 50% would contain the mutation.

As was performed for the previous bulk analysis, LoFreq and VarScan were used to call mutations (Kelly et al., 2017). When both tools called a mutation above 0.04, it was brought forward for analysis. A total of 43 mutations were called among the 4 samples of which 17 were indels and 26 SNPs (**Fig. S1B**). Of mutations heteroplasmic in the bulk sample, the single cell average allele frequency varied dramatically from bulk (**Fig. 4A**). For example, the 5462T>C mutation was 0.05 in the bulk, but over 0.5 in the single cell average. This is likely a consequence of the small number of single cell samples sequenced but also suggests that some mutations may exist sporadically at a high frequency in a small number of cells within the population (scenario 2 in **Fig. 5**).

Fig. 4

To better assess the variability in allele frequency among single cells, we developed a list of “most variable” mutations which must be heteroplasmic in at least 2 out of the 4 cells. There was a wide range of allele frequencies among the single cells (**Fig. 4B**). Had only bulk analysis been performed, this range of allele frequency would not have been captured; demonstrating the enhanced resolution possible from single cell analysis.

We next considered whether the mutations among the samples were concordant. We generated a heatmap of mutations from all samples, with each mutation type represented by a colour (**Fig. 4C**). The mutations 5244TA>T and 14136T>A were present in all samples. Considering the nature of a bulk sample, one might expect that all the mutations present in the single cells should be present in the bulk sample. However, for the mutation to be called with confidence in the bulk sample, it must be above 0.04; and it must therefore be present in individual cells in the population at sufficient frequency to average above 0.04. All mutations from the bulk sample were found in at least 1 single cell; apart from the intragenic 3733G>A. The limited number of single cells analysed likely resulted in this observation. On the other hand, 73% (19/26) of mutation locations were exclusively found in the single cells but not in the bulk sample, again evidencing the high degree of resolution from single cell analysis.

The lowest number of mutations were in the bulk population with 13 (joint with Single cell 1) (**Fig. 4C**). This was not unexpected because single cells can contain additional rare mutations that average out below 0.04 in the bulk sample. On the other hand, individual single cells do not necessarily contain all the mutations found in the bulk sample. Ultimately, the greater the number of single cells analysed, the greater the certainty about the nature of individual mutations across a population of cells when compared to bulk cell sequencing.

5.6 Predicted Phenotypic Impact of Mutations

Although the presence of mutations is useful for demonstrating intercellular diversity, their phenotypic impact may be limited by many factors including the mutation's effect on, for example, amino acid sequence or tRNA structure. We therefore analysed mutations based on predicted impact on phenotype using the snpEff software tool (**Fig. 4D**) (**Fig. S1D**). Of particular interest, frameshift mutations were observed in protein-coding genes COX1, CYTB and ND4 (**Fig. S1E**). At least 1 frameshift mutation was called in all samples. Only 4 heteroplasmic mutations were above 0.5 allele frequency, with most mutations present at low levels. Mapping of mutations against allele frequency showed the vast majority to be in the 0.04-0.5 range (**Fig. S1F**).

6 Discussion

The phenotypic manifestations of disease typically occur when the responsible mtDNA mutation allele frequency reaches a certain threshold within a cell. Previous bulk analysis of CHO mtDNA identified heteroplasmy in clones derived from the same parental host, indicating at least three levels of heterogeneity: (1) production run to production run, (2) cell line to cell line and (3) clone to clone (Kelly. P et al., 2017). However, bulk analysis of heteroplasmy fails to identify the allele frequency differences between individual cells in a population (**Fig. 1A**). Single cell analysis is therefore critical to reveal the true phenotypic effect of heteroplasmy.

To illustrate this, consider if there was a hypothetical mutation with a threshold of 0.7, above which phenotypic changes would manifest in an individual cell. If bulk analysis identified this critical mutation at a frequency of 0.1 in 1 million cells (**Fig. 5**), three very different conclusions could be arrived at: (1) All cells contain the mutation at a 0.1 allele frequency, and are therefore all unaffected phenotypically; (2) 10% of cells contain the mutation at a 1.0 allele frequency, and therefore only 10% of cells are affected phenotypically or (3) cells contain the mutation at a variable rate (0-1.0), and therefore the population is affected at a variable rate.

Fig. 5

Bearing in mind the strive for homogeneity in drug production, the implications of these scenarios are significant. If a particular heteroplasmy profile affected product quality for example, perhaps only a subset of the cells produce the product at a high quality; in which case the remainder could be identified and potentially excluded to improve bioreactor performance. Equally, perhaps a cell line could be engineered with a favourable heteroplasmy profile to improve bioreactor performance. Further work is clearly needed to understand the link between mitochondrial heteroplasmy and cellular behaviour in recombinant protein production, but single cell analysis should contribute significantly in this regard.

Though the 5 samples here are demonstrative, and not enough for strong statistical conclusions, certain observations were made. The bulk population had the lowest number of reliably detectable mutations (**Fig. S1B**). All mutations in the bulk population, bar one, were found in at least one single cell (**Fig. 4C**). This demonstrated the improved resolution of mutation detection using a single cell approach. A great range of allele frequencies in "most variable" mutations was observed (**Fig. 4B**);

further indicating an uneven spread of heteroplasmy among the 4 cells, reminiscent of scenario 3 in **Fig. 5**.

High-impact mutations observed here in *CYTB* (**Fig. S1D**) would change the encoded amino acid sequence. The phenotypic effects of *CYTB* mutations are well established in human disease where patients experience highly variable severities of myopathy and muscle weakness (Blakely et al., 2005). *CYTB* mutations in yeast models can cause severe decreases in respiratory function (Fisher et al., 2004). In a bioreactor, *CYTB* mutated single cells (above a phenotypic threshold) may be one of many contributing factors to the heterogeneity observed among clonally derived CHO populations.

In conclusion, a reliable method to amplify and analyse mtDNA from single CHO cells was demonstrated (scmtDNAseq). This approach should help better understand the degree and likely impact of heteroplasmy on recombinant protein production in CHO cells.

7 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

8 Author Contributions

Alan Foley: Developed novel method under supervision of Niall Barron. Adapted bioinformatics code from Colin Clarke, and wrote manuscript.

Niall Barron: Supervisor of project. Guided method development. Manuscript review/editing.

Colin Clarke: Co-supervisor. Wrote original code for mtDNA analysis which was later adapted. Manuscript review.

9 Funding

Funded by the STACCATO project as part of the Marie Sklodowska-Curie Action of Horizon 2020 (Grant No. 813453).

10 Acknowledgments

NIBRT, Ireland: a base for method development and bioinformatics.

Niall Barron: Supervised project.

Colin Clarke: Co-supervisor. Wrote original code for mtDNA analysis which was later adapted.

Nga Lao: Significant aid in wet lab work.

11 References

Bio-Rad (2022). Doublet Discrimination - Flow Cytometry Guide. Bio-Rad. Available at: https://www.bio-rad-antibodies.com/flow-cytometry-doublet-discrimination.html [Accessed January 17, 2022].

Blakely, E. L., Mitchell, A. L., Fisher, N., Meunier, B., Nijtmans, L. G., Schaefer, A. M., et al. (2005). A mitochondrial cytochrome b mutation causing severe respiratory chain enzyme deficiency in humans and yeast. FEBS J 272, 3583–3592. doi: 10.1111/j.1742-4658.2005.04779.x.

Dhiman, H., Gerstl, M. P., Ruckerbauer, D., Hanscho, M., Himmelbauer, H., Clarke, C., et al. (2019). Genetic and Epigenetic Variation across Genes Involved in Energy Metabolism and Mitochondria of Chinese Hamster Ovary Cell Lines. Biotechnology Journal 14, 1800681. doi: 10.1002/biot.201800681.

Dimauro, S., and Davidzon, G. (2005). Mitochondrial DNA and disease. Ann Med 37, 222–232. doi: 10.1080/07853890510007368.

Fisher, N., Castleden, C. K., Bourges, I., Brasseur, G., Dujardin, G., and Meunier, B. (2004). Human disease-related mutations in cytochrome b studied in yeast. J Biol Chem 279, 12951–12958. doi: 10.1074/jbc.M313866200.

Gallagher, C., and Kelly, P. S. (2017). Selection of High-Producing Clones Using FACS for CHO Cell Line Development. Methods Mol Biol 1603, 143–152. doi: 10.1007/978-1-4939-6972-2_9.

Gilbert, A., McElearney, K., Kshirsagar, R., Sinacore, M. S., and Ryll, T. (2013). Investigation of metabolic variability observed in extended fed batch cell culture. Biotechnol Prog 29, 1519–1527. doi: 10.1002/btpr.1787.

Kelley, B., Kiss, R., and Laird, M. (2018). A Different Perspective: How Much Innovation Is Really Needed for Monoclonal Antibody Production Using Mammalian Cell Technology? Adv Biochem Eng Biotechnol 165, 443–462. doi: 10.1007/10_2018_59.

Kelly, P. S., Clarke, C., Costello, A., Monger, C., Meiller, J., Dhiman, H., et al. (2017). Ultra-deep next generation mitochondrial genome sequencing reveals widespread heteroplasmy in Chinese hamster ovary cells. Metabolic Engineering 41, 11–22. doi: 10.1016/j.ymben.2017.02.001.

Keogh, M. J., and Chinnery, P. F. (2015). Mitochondrial DNA mutations in neurodegeneration. Biochimica et Biophysica Acta (BBA) - Bioenergetics 1847, 1401–1411. doi: 10.1016/j.bbabio.2015.05.015.

Lorenz, T. C. (2012). Polymerase chain reaction: basic protocol plus troubleshooting and optimization strategies. J Vis Exp, e3998. doi: 10.3791/3998.

Ludwig, L. S., Lareau, C. A., Ulirsch, J. C., Christian, E., Muus, C., Li, L. H., et al. (2019). Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. Cell 176, 1325-1339.e22. doi: 10.1016/j.cell.2019.01.022.

Maeda, R., Kami, D., Maeda, H., Shikuma, A., and Gojo, S. (2020). High throughput single cell analysis of mitochondrial heteroplasmy in mitochondrial diseases. Sci Rep 10, 10821. doi: 10.1038/s41598-020-67686-z.

NCBI (2023). National Center for Biotechnology Information. Available at: https://www.ncbi.nlm.nih.gov/ [Accessed June 5, 2023].

Walsh, G., and Walsh, E. (2022). Biopharmaceutical benchmarks 2022. Nat Biotechnol 40, 1722–1760. doi: 10.1038/s41587-022-01582-x.

Wei, W., Schon, K. R., Elgar, G., Orioli, A., Tanguy, M., Giess, A., et al. (2022). Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes. Nature 611, 105–114. doi: 10.1038/s41586-022-05288-7.

Zambelli, F., Vancampenhout, K., Daneels, D., Brown, D., Mertens, J., Van Dooren, S., et al. (2017). Accurate and comprehensive analysis of single nucleotide variants and large deletions of the human mitochondrial genome in DNA and single cells. Eur J Hum Genet 25, 1229–1236. doi: 10.1038/ejhg.2017.129.

12 Data Availability Statement

The datasets for this study can be found in the GitHub repository: https://github.com/alanfoleynibrt/SingleCellmtDNA

13 Tables

Table 1: Samples 1-4 required for single cell sort. In this data, Late Exponential had a viability of 95%, Dead of 5%.

	Cells	Growth Phase	Stain	Function
1	Protein Producing CHO	Late Exponential	DAPI + FITC-AB	Sorting Sample
2	Protein Producing CHO	Dead	DAPI + FITC-AB	Gate Live/Dead Cells
3	Non-producing CHO	Late Exponential	DAPI + FITC-AB	Gate FITC negative
4	CHO	Late Exponential	None	Gate FSC, SSC. Gate FITC positive

Table 2: Primer sequences for LRPCR. Other cell lines may need adaptation of these sequences.

Primer	Sequence
mt-490 F (X)	5’ - GGA TTA GAT ACC CCA CTA TGC TT – 3’
mt-9304 R (X)	5’ – ATG CTG CGG CTT CAA ATC CG – 3’
mt-9180 F (Y)	5’ – ATA GCA ACA GGT TTT CAC GG – 3’
mt-598 R (Y)	5’- CGC CAA GTC CTT TGA GTT TTA – 3’

Table 3: PCR components

Reagent	**Volume per Rx (uL)**	**10x Mastermix (uL) X**	**10x Mastermix (uL) Y**
5x Buffer	5	50	50
10mM DNTP mix	0.5	5	5
10uM Primer F	1	10 (X primer)	10 (Y primer)
10uM Primer R	1	10 (X primer)	10 (Y primer)
Nuclease-free H2O	8.5	85	85
SuperFi II DNA Polymerase	0.5	5	5
TOTAL	**16.5**	**165**	**165**

Table 4: PCR settings

Step	Temperature (°C)	Time

1. Initial Denature	94	2 mins
2. (x35) Denature	94	30 s
Annealing	55	30 s
Extension	68	9 mins
3. Final extension	68	10 mins
4. Hold	4	Infinite hold

14 Figure Legends

Fig. 1: (**A**) An explanation of the multi-copy nature (heteroplasmy) of mitochondrial DNA. Numbers are true for CHO cells, though vary by eukaryotic cell type. (**B**) Method overview for scmtDNAseq. Made with BioRender.

Fig. 2: (**A-F**) Gating strategy to sort alive, singlet, antibody-producing CHO cells. (**G**) Agarose gel illustrating amplification of CHO cell mtDNA from a single cell. Also included is positive control of 1000 cells and negative control of 0 cells. "X" and "Y" refers to the 2 separate halves of the mtDNA molecule. Together, 1X and 1Y represent amplification of the whole mtDNA molecule from a single cell in two separate reactions. Red rectangles illustrate gel extraction boundaries to exclude bands other than the desired 8.5kb amplicon. Made with BioRender.

Fig. 3: (**A**) Read mapping of samples to the KX576660 CHO mtDNA reference genome. **(B)** Average per base sequencing depth of each sample. **(C)** Perbase coverage of 4 single cells and a bulk sample with correction around 0 coordinate. Made with BioRender.

Fig. 4: (A) Comparison of bulk sample allele frequencies to the average of 4 single cells' allele frequencies. The mutation must be heteroplasmic (between 0.04 and 0.96 allele frequency) in the bulk sample. **(B)** Violin plot of "most variable" mutations which must be heteroplasmic in at least 2 of 4 single cells. **(C)** Base change heatmap of heteroplasmic mutations (between 0.04 and 0.96 allele frequency) in 4 single cells and a bulk sample. **(D)** snpEff predicted impact of heteroplasmic (between 0.04 and 0.96 allele frequency) mtDNA mutations of 4 single cells and a bulk sample. Made with BioRender.

Fig. 5: How bulk analysis of heteroplasmy can obfuscate single cell orientations. Made with BioRender.

Fig. S1: **(A)** Tapestation (4200) image of mtDNA LRPCR for dilutions of a single cell. Negative control is 0 cells, positive control is 10 cells. Dilutions were made from 1/10 to 1/100,000. **(B)**

Number of heteroplasmic mutations (between 0.04 and 0.96 allele frequency) in 4 single cells and a bulk sample. **(C)** Allele Frequency heatmap of heteroplasmic (between 0.04 and 0.96 allele frequency) mutations in 4 single cells and a bulk sample. **(D)** snpEff predicted impact of heteroplasmic (between 0.04 and 0.96 allele frequency) mutations in 4 single cells and a bulk sample. **(E)** Number of mutations per gene in 4 single cells and a bulk sample. **(F)** Allele frequency distribution of heteroplasmic (between 0.04 and 0.96 allele frequency) mutations in 4 single cells and a bulk sample. Made with BioRender.