Ultra-deep next generation mitochondrial genome sequencing reveals widespread heteroplasmy in Chinese hamster ovary cells.

Paul S Kelly[1†], Colin Clarke[2†], Alan Costello[1], Craig Monger[1,2], Justine Meiller[1], Heena Dhiman[2], Nicole Borth[3], Michael J Betenbaugh[4], Martin Clynes[1] and Niall Barron[1*]

[1] National Institute for Cellular Biotechnology, Dublin City University, Glasnevin, Dublin9, Ireland.

[2] National Institute for Bioprocessing Research and Training, Fosters Avenue, Blackrock, Co. Dublin, Ireland.

[3] University of Natural Resources and Life Sciences Vienna, Muthgasse 18, Vienna 1190,Austria.

[4] Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, Maryland, USA.

[†] Equal contribution

[*]Corresponding author

Email: niall.barron@dcu.ie

Phone: +353 1 700 5700

Fax: +353 1 700 5484

Keyword: Chinese hamster ovary; Biopharmaceutical; Mitochondrial; Heteroplasmy; Next generation sequencing; Genomics;

Abbreviations: Chinese hamster ovary (CHO); Metabolic flux analysis (MFA); Oxidative phosphorylation (OXPHOS); nuclear DNA (nDNA); mitochondrial DNA (mtDNA); displacement loop (D-loop); reactive oxygen species (ROS); minor allele frequency (MAF); nuclear encoded mitochondrial sequences (NumtS); single nucleotide polymorphism (SNP); insertion/deletion (INDEL);

Abstract

Recent sequencing of the Chinese hamster ovary (CHO) cell and Chinese hamster genomes has dramatically advanced our ability to understand the biology of these mammalian cell factories. In this study, we focus on the powerhouse of the CHO cell, the mitochondrion. Utilizing a high-resolution next generation sequencing approach we sequenced the Chinese hamster mitochondrial genome for the first time and surveyed the mutational landscape of CHO cell mitochondrial DNA (mtDNA). Depths of coverage ranging from ~3,319X to 8,056X enabled accurate identification of low frequency mutations (>1%) revealing that mtDNA heteroplasmy is widespread in CHO cells. A total of 197 variants at 130 individual nucleotide positions were identified across a panel of 22 cell lines with 81% of variants occurring at an allele frequency of between 1 and 99%. 89% of the heteroplasmic mutations identified were cell line specific with the majority of shared heteroplasmic SNPs and INDELs detected in clones from 2 cell line development projects originating from the same host cell line. The frequency of common predicted loss of function mutations varied significantly amongst the clones indicating that heteroplasmic mtDNA variation could lead to a continuous range of phenotypes and play a role in cell to cell, production run to production run and indeed clone to clone variation in CHO cell metabolism. Experiments that integrate mtDNA sequencing with metabolic flux analysis and metabolomics have the potential to improve cell line selection and enhance CHO cell metabolic phenotypes for biopharmaceutical manufacturing through rational mitochondrial genome engineering.

1 Introduction:

The continual improvement of bioprocesses over the last 20 years has enabled the production of g/L quantities of complex therapeutic proteins (e.g. monoclonal antibodies) from industrial scale Chinese hamster ovary (CHO) cell culture [1,2]. These dramatic improvements in performance have been achieved, in part, through understanding the nutrient requirements of CHO cells to optimise media formulations. Industrial scale cell culture processes, where possible, now utilise chemically defined media that maintain growth rate and increase titre as well as eliminate the batch variation associated with biological components such as serum [3]. The development of fed-batch cell culture strategies have also been central to achieving high product titres, counteracting the production of cellular waste products during cell culture and extending production runtimes. During exponential growth, CHO cells channel glucose and glutamine through the glycolytic pathway even in cases of high oxygen availability (aerobic glycolysis), a metabolic phenotype similar to the Warburg effect observed in cancer cells [4]. The resulting secretion of lactate and ammonium inhibit cell growth and productivity as well as initiate apoptosis and decrease product quality [3]. In a fed-batch process, cells are initially grown to a high cell density before the bioreactor environment is altered through e.g. reducing the temperature or altering the pH of cell culture. A stationary phase of cell growth is induced to shift CHO cells from a lactate production to consumption phenotype extending, viability and maximising protein production [3]. Metabolic flux analysis (MFA) and metabolomics approaches have proven to be powerful tools for understanding the molecular basis of CHO cell metabolic phenotypes to enable predictable, rapid and inexpensive optimisation of industrial bioprocesses [5]. These techniques have enabled the development of complex closed loop feeding strategies to limit glucose and glutamine concentrations as well as identify genetic engineering targets to drive CHO cells toward metabolically desirable phenotypes [6].

Recent studies of CHO cell metabolism have indicated that mitochondrial function is central to lactate production/consumption [7] and indeed the variability observed in CHO cell metabolic phenotypes [8]. Mitochondria play a central role in eukaryotic cellular energy metabolism via oxidative phosphorylation (OXPHOS) and have important functions in biological processes such as intracellular calcium signalling [9] and apoptosis [10]. While the overwhelming majority of proteins required to carry out these functions are transcribed from nuclear DNA (nDNA) [11], mitochondria also contain a distinct, double stranded circular genome. Eukaryotic cells can contain more than 1,000 copies of mitochondrial DNA (mtDNA) packaged within DNA-protein structures known as mitochondrial nucleoids (each nucleoid contains 2-10 mtDNA molecules). mtDNA copy number varies according to cell type, for instance with myocardial muscle cells containing on average 6,000 copies while leukocytes may have as few as 350 per cell [12,13]. In humans, a significant degree of variation in mtDNA copy number has been observed between the same tissues of different individuals as well as across multiple tissues from the same individual [14]. The mitochondrial genome is between 15 to 17 kb in length and contains 37 genes (28 on the guanine rich “heavy” or H-strand and 9 on the cytosine rich “light” or L-strand). mtDNA encodes 13 polypeptide subunits of OXPHOS complexes I, III, IV and V along with 2 ribosomal RNA subunits and 22 tRNAs required for intra-mitochondrial protein synthesis. The mtDNA genome is extremely compact, genes

lack introns, intergenic regions are limited to 1 or 2 nucleotides and in some cases genes can overlap (e.g. *ATP6* and *ATP8*). The only significant non-coding regulatory region is called the displacement loop (D-loop) and contains the origin of replication for the H-strand. Transcription is initiated from one of two H-strand promoters or a single promoter on the L-strand resulting in polycistronic RNA and subsequently processed to produce mRNAs, tRNAs and rRNAs. Mitochondria utilise a distinct genetic code for mRNA translation allowing translation of all codons using only 22 tRNAs [15,16]. mtDNA encodes for only two of the four nDNA stop codons ("AGA" and "AGG"), with the "UAA" stop codon added to transcribed mRNAs via polyadenylation. The nDNA stop codon "UGA" encodes tryptophan in mtDNA while the "AUA" codon that encodes for isoleucine in nDNA encodes methionine in mtDNA.

Relatively inefficient DNA repair mechanisms and close proximity to reactive oxygen species (ROS) contribute to a mitochondrial genome mutation rate at least 10 fold higher than that of the nuclear genome. This high mutation rate has seen widespread application of mtDNA sequencing for studies in evolutionary biology, population genetics and forensic science. The first pathogenic human mitochondrial mutations were identified nearly 30 years ago and since then more than 250 polymorphisms, insertions and deletions have been implicated in metabolic disorders as well as cancer and diabetes. Mitochondrial genome polyploidy can give rise to two cellular states: 1) all mtDNA copies are identical known as homoplasmy or 2) a mixture of wild-type and mutated mtDNA copies are present, known as heteroplasmy. In healthy cells, wild-type and mutated mtDNA copies can co-exist; mitochondrial dysfunction occurs when the ratio reaches a particular level known as the mitochondrial threshold effect, and in some cases, the frequency of heteroplasmy correlates with the severity of a clinical phenotype [17]. In recent years, next generation sequencing (NGS) technologies have seen widespread application for the study of heteroplasmy due to increased specificity, sensitivity and throughput in comparison to traditional Sanger sequencing. Although initially thought to be a rare phenomenon, NGS has revealed the prevalence of mitochondrial heteroplasmy in the human population as well as the age related increase of heteroplasmic variants [18]. Studies utilizing ultra-deep sequencing to identify very low frequency variants have indicated that heteroplasmy is universal with each cell containing a complex mixture of mitochondrial genotypes [17].

Efforts to understand the biology of CHO cell factories and improve industrial scale biopharmaceutical manufacturing have been dramatically enhanced since the landmark publication of the CHO-K1 genome in 2011 [19]. A wealth of sequence data is now freely available for several CHO cell lines as well as the Chinese hamster [19–22]. Direct analyses of these data have permitted the first studies of CHO cell genome instability [22], chromosomal rearrangement [23], and copy number variation [20]. Methods for expression profiling have also seen marked improvement in the CHO cell post-genomic era and overcome the reliance on homology with model species that limited early studies in the field. CHO-cell-specific sequence databases have increased the number of identifications from mass spectrometry based proteomic analysis [24]. The combination of genome sequence and NGS technology to study RNA (termed RNA-Seq) has been employed to study mRNA and small RNA [25–27] expression patterns as well as to annotate transcripts [28] and identify promoter regions [29], providing novel insights in to the CHO cell transcriptome.

While the availability of nuclear genome sequences has undoubtedly advanced CHO cell biology, we know little about the mtDNA and the impact of mutations on cell metabolism and bioprocess performance. Here, we present the first comprehensive survey of the CHO cell mitochondrial genome, spanning a panel of cell lines originating from industry, the ATCC and our laboratory. The utilisation of next generation sequencing technology enabled high-resolution detection of mtDNA mutations including those occurring at low frequency. Our results indicate that heteroplasmy is widespread in CHO cell lines, tends to be cell line specific and that these mutations could play a role in metabolic phenotype variability.

2 Materials and Methods

2.1 Extraction of DNA from Chinese hamster and mouse liver tissue

Genomic DNA was extracted from 30 mg of liver tissue from an outbred Chinese hamster and a CB17/lcr-Prkdcscid/Crl mouse liver samples (Supplementary Table 1) using a DNeasy Blood and Tissue kit (QIAGEN, 69581). Tissue samples were sheared using a Dounce homogenizer in 180 µl of ALT buffer. The purity and integrity of extracted tissue-derived genomic DNA was determined on a nano-drop and via a DNA-agarose gel stained with ethidium bromide (Supplementary Figure 1A).

2.2 CHO cell culture and mtDNA extraction

22 CHO cell lines (Supplementary Table 1) were grown in suspension unless indicated otherwise and harvested at 72 hours. All suspension cultured cell lines were seeded initially at 2×10^5 cells/mL in 5 mL of culture media. The 4 Biogen cell lines were cultured in suspension in proprietary chemically defined media supplied by the industry partner and cultured in-house. The 8 clones from Pfizer originating from 2 cell line development projects (CLD1 and CLD2), were grown in attached culture in DMEM supplemented with 5% serum. All remaining CHO cell lines were cultured in suspension in 5 mL of serum-free media at an initial density of 2×10^5 cells/mL. 17×10^6 cells were acquired at 72 hours for mitochondrial DNA extraction [30]. To reduce the contaminating nuclear DNA an additional step was carried out for CHO cell lines to enrich for double stranded mtDNA using a bacterial mini-prep kit (QIAGEN, 27104), as previously described [31]. Purity and integrity of isolated CHO mitochondrial DNA was determined using a NanoDrop and an agarose DNA gel (Supplementary Figure 1B).

2.3 Amplification of mitochondrial DNA

To further eliminate nuclear DNA from the tissue and cell line samples, we amplified mtDNA fragments using a high fidelity PCR kit (Life Technologies, 11304-011) (Supplementary Figure 1C). CHO mtDNA primers were designed using the CHO cell mtDNA sequence available on GenBank (NC_007936.1). Two overlapping ~8.5 kb mtDNA fragments were designed to span the ~16.5 kb mitochondrial genome sequence (Supplementary Table 2). Another set of overlapping primers was designed based on the *Mus musculus* mitochondrial genome sequence (NC_005089.1). PCR amplification of each paired mitochondrial genome fragment was performed using the following thermo-cycler conditions: 94°C for 2 min, 12 cycles at 94°C for 30 sec, 55°C for 30 sec and 68°C for 8.5 min. The resulting PCR products for each respective sample was cleaned using Ampure® XP DNA-binding magnetic beads (Agencourt, A63880). Two PCR primers were designed to amplify a short 454 bp amplicon of the gene CYTB

flanking two identified mutated sites (m14136 and m14378) which was used for Sanger sequencing (For: 5'-TTCAAAGATGTAGCCATACAACC-3' and Rev: 5- AACCGTAATAAACTCCTCGTCC-3').

2.4 Nextera XT Mitochondrial DNA library preparation and sequencing

For each mtDNA sample, a serial dilution was performed in nuclease-free water and quantified using the Qubit dsDNA HS assay kit (Invitrogen, Q32851) to obtain a 0.2 ng/µl stock. 1 ng of mitochondrial DNA library was prepared using the Nextera XT DNA Sample Preparation Kit (FC-131-1024) in accordance with the manufacturer's specifications. After each library was fragmented, adapters were added followed by the incorporation of sample indexes by PCR. Each of the 24 uniquely indexed samples was passed through a PCR clean-up using Ampure® XP DNA-binding magnetic beads (Agencourt, A63880). Libraries were quantified using the Qubit dsDNA HS assay kit and fragment size distribution was determined using the High Sensitivity DNA Bioanalyzer kit (Agilent, 5067-4626) to confirm the recommended 500-600bp range. Libraries were normalised to 4 nM in resuspension buffer prior to sequencing. Each 4 nM library was then pooled into a single sample and sequenced on an Illumina MiSeq (San Diego, CA) configured to produce 151bp paired end reads. Following sequencing, base calls were converted to 24 individual FASTQ format files for bioinformatics analysis (all raw data will be uploaded to NCBI's SRA database upon acceptance of publication).

2.5 Reconstruction and annotation of the Cricetulus griseus mitochondrial genome sequence.

To reconstruct the Chinese hamster mtDNA sequence the MITOBIM algorithm [32] was utilised in combination with the CHO cell mtDNA sequence available on GenBank (NC_007936). Paired-end reads were merged using FLASH [33] prior to assembly. MITOBIM assembles a mitochondrial genome by mapping sequencing reads (from Chinese hamster liver tissue) to a closely related sequence, in this case CHO cell mtDNA. The newly assembled Chinese hamster mtDNA was initially annotated using the MITOS [34] and ARWEN [35] webservers. Annotations were verified and if necessary, refined via BLAST analysis and comparison to human, mouse and rat mtDNA.

2.6 CHO cell line mapping and data pre-processing

Reads corresponding to each of the 22 CHO cell lines sequenced were subjected to quality control assessment followed by the removal of adapter sequences and reads < 50bp using *trimmomatic* [36]. The remaining sequence data was mapped to the Chinese hamster mitochondrial genome reference using the BWA-MEM algorithm [37]. Representation of a circular mitochondrial genome as a linear sequence (i.e. beginning at position 1 and ending at position 16,283) can give rise to incomplete read mapping due to the introduction of an artificial sequence break. Paired end reads that span the sequence break will be designated as unmapped and eliminated by algorithms such as BWA decreasing depth at the start and end of the mtDNA reference and affecting the ability to accurately detect variants in these regions. In this study the "double" alignment mapping strategy described by Ding *et. al.* [38] was utilised to ensure optimum alignment of reads to the Chinese hamster mitochondrial genome. Using this approach, CHO cell line reads were mapped to the original or "unshifted" reference sequence beginning at position 1 and ending at position 16,283. For the second mapping run a new reference mtDNA sequence was created by joining the start and ends of the original sequence and

introducing a new break point so that the sequence began at position 8,000 and ended at position 7999 on the original reference (Supplementary Figure 2).

2.7 Variant pre-processing

Following alignment against the “unshifted” and “shifted” reference sequence reads, with a MAPQ < 20 were designated as “unmapped” and discarded. To pre-process the “unshifted” and “shifted” mapped data for variant calling, we followed the Genome Analysis ToolKit (GATK) [39] best practice guidelines. PCR duplicates can arise during the library preparation following the amplification of the multiple copies of identical DNA fragments. Duplicates propagate errors in sample and library preparation across the dataset, violate the assumption of independence during variant calling and potentially result in the identification of false positive variants. Reads corresponding to PCR duplicates were identified using *Picard* (http://broadinstitute.github.io/picard/) and eliminated from further analyses. The remaining reads were realigned around INDELs accounting for mapping artefacts that arise from the independent read by read alignment process. Base quality score recalibration reduces the effect of systematic sequencing biases by first determining covariation between factors including nucleotide context (e.g. AC dinucleotides are often lower quality than TG) and base position within the read (bases at the ends of the reads generally have more mismatches). The Phred scaled Q values are then adjusted accordingly to reduce false positives during variant calling.

2.8 Variant discovery and annotation

To identify CHO cell line mtDNA mutations using the dual mapping strategy, variant calls were made within specific regions of the “unshifted” and “shifted” Chinese hamster reference sequences (Supplementary Figure 2). For the original “unshifted” reference, variants are called between positions 4,000 and 12,000, while sequence variants on the “shifted” reference sequence are called between positions 4,000 to 12,283, translating to the regions spanning 1 to 3,999 and 12,001 to 16,283 on the original reference sequence (encompassing the joined start and ends). The bioinformatics pipeline incorporated both VarScan [40] and LoFreq [41] for variant detection. Only those SNPs and INDELs identified by both algorithms with a minor allele frequency (MAF) ≥ 1%, minimum sequencing depth > 1,500X at the variant position and an average Phred-scaled base quality (≥ Q25) for the alternate allele were reported. Variants were eliminated if overrepresentation of reads supporting the variant was observed in either forward or reverse direction (i.e. strand bias) [42]. An additional threshold was employed for VarScan calls and only variants with $p < 0.01$ were retained. Each identified INDEL was inspected manually to confirm potential false positives. Upon completion of the mutation detection pipeline, the coordinates of “shifted” sequence variants were transformed back to the original reference coordinates and combined with those variants identified following “unshifted” sequence analysis. To determine the putative effect of each mutation we first utilised snpEff [43] to annotate each mutation (i.e. frameshift, stop codon gained, start codon mutation, missense or non-synonymous). The PROVEAN algorithm [44] was utilised to predict the functional impact of missense variants on protein function. Those missense variants with a Provean score ≤ -2.5 were classified as deleterious.

2.9 Estimation of contamination from nuclear mitochondrial sequences

Heteroplasmy detection can be confounded, particularly from whole genome sequencing data, by the presence of nuclear encoded copies of mitochondrial sequences (NumtS) [45]. In this study, potential

contamination was reduced through the long range PCR to enrich for mitochondrial DNA and in the case of cell lines, the utilisation of a bacterial mini-prep kit to eliminate non-circular DNA prior to amplification. To confirm the effectiveness of the enrichment strategy for mitochondrial DNA, we utilised the mouse mitochondrial sequencing data to assess potential contamination from NumtS. The variant identification pipeline was first used to identify SNPs and INDELs against the mouse reference sequence (NC_005089.1). To estimate the influence of NumtS, processed reads from mouse were separately aligned to the *Mus musculus* nuclear genome (mm9 assembly) using the BWA-MEM algorithm (the "–L" and "-T" parameters were set to "9,9" and "145" respectively). Those reads which mapped to known mm9 NumtS regions [46] with a MAPQ > 20 were extracted, reconverted to FASTQ files and remapped against the mouse mtDNA reference sequence using the variant discovery pipeline described above. The thresholds of the LoFreq and VarScan algorithms were modified to account for the lower depth of coverage in order to determine if NumtS were influencing variant class and heteroplasmy levels.

3 Results

3.1 Reconstruction of the Cricetulus griseus mitochondrial DNA sequence

Examination of the NC_007936 mtDNA sequence and the publication describing its acquisition [47] revealed that the sequence does not originate from *Cricetulus griseus* but the mitochondrial genome of either a CHO-K1 or CHO A_L cell line. There are also a number of discrepancies between the annotation described in the publication and the GenBank entry. For example, the GenBank annotation states that $tRNA^{Asn}$ is encoded on mtDNA H-strand while the original publication places $tRNA^{Asn}$ on the L-strand. In this study, we sequenced the Chinese hamster mitochondrial genome to produce an accurate reference sequence for comparability of CHO cell line mtDNA as well as resolving any ambiguities in annotation (Figure 1). Chinese hamster mtDNA was isolated from liver tissue and sequenced on the Illumina MiSeq platform yielding 640,142 paired-end reads. We reconstructed the mitochondrial genome from these data using the MITOBIM [32] algorithm with the NC_007936 mtDNA sequence used as the "backbone" sequence. The assembled *C.griseus* mitochondrial genome is 16,283bp in length (A=33.7%, C=22.8%, G=13.0%, T=30.5%) with an overall GC content of 35.7%.

Annotation was initially performed using MITOS [34] and ARWEN [35] and further refined through comparison of the Chinese hamster mtDNA with human, mouse and rat reference mtDNA sequences (Supplementary Table 3). The Chinese hamster mtDNA has conserved synteny with mammalian mitochondrial genomes with 13 protein-coding genes, 22 tRNAs and 2 ribosomal RNAs as well as a non-coding control region (D-loop). Nine genes are encoded on the mtDNA light strand (*ND6*, $tRNA^{Gln}$, $tRNA^{Ala}$, $tRNA^{Asn}$, $tRNA^{Cys}$, $tRNA^{Tyr}$, $tRNA^{Ser}$, $tRNA^{Glu}$ and $tRNA^{Pro}$) with the remaining 28 genes encoded by the H-strand. 9 protein-coding genes start with ATG initiation codon (*COX1*, *COX2*, *ATP8*, *ATP6*, *COX3*, *ND4L*, *ND4, ND6, CYTB*), 2 with an ATT codon (*ND2* and *ND5*), 1 with a GTG (*ND1*) and 1 with an ATA codon (*ND3*). 8 genes terminated with the TAA codon with the *ND1*, *ND2*, *COX3, ND6,* and *ND4* stop codons predicted to be completed via transcript polyadenylation. Comparison of the Chinese hamster mtDNA sequence to the currently available CHO cell line mitochondrial genome identified 7 variants, 5 in protein coding sequences with 2 mutations identified in the mtDNA D-loop

comprising 4 SNPs, 2 deletions and an insertion (Table 1). The *C.griseus* mitochondrial genome sequence and corresponding annotation have been submitted to GenBank (accession no: KX576660 [release date 28.09.2016]).

3.2 Mapping of CHO cell line next generation sequencing data to the C.gresius mitochondrial genome

A total of 22 CHO cell lines sourced from industry partners, the ATCC and from our laboratory were sequenced (Supplementary Table 1). Adapter sequences were trimmed from each of the CHO cell line datasets followed by removal of low quality reads and those less than 50bp. Upon completion of this initial pre-processing stage the number of reads remaining in each sample ranged from 1,547,006 (DCU CHO-K1 SEAP) to 465,194 (Biogen DG44 #1) cell line (Supplementary Table 4A). The BWA-MEM algorithm was used to align reads against the *C.gresius* mtDNA reference sequence as well as a modified version of the reference sequence where the original start and ends were joined and a new breakpoint introduced at 8000bp (Supplementary Figure 2). This dual mapping strategy was utilized to account for the circularity of the mitochondrial genome and remove bias arising from the use of a linear reference sequence (e.g. discarding reads that spanned the artificial breakpoint) and therefore improve our ability to detect mutations around the start and ends of the mtDNA sequence.

Following alignment to both the shifted and unshifted reference sequence, reads with a MAPQ < 20 were eliminated from further analysis. The DCU CHO-K1 SEAP sample was found to have the largest number of reads mapping with MAPQ < 20, yet this represented only ~1% of the total reads in that dataset. On average > 99% of reads mapped to the unshifted and shifted reference sequence with a MAPQ ≥ 20 (Figure 2A; Supplementary Table 4B & 4C) demonstrating the effectiveness of the mtDNA isolation and amplification method utilized in this study. To ensure PCR duplicates were ignored in downstream stages of the bioinformatics analysis duplicates were “marked” using the *Picard* tool (http://broadinstitute.github.io/picard/). The DCU CHO-K1 SEAP sample had the highest proportion of duplicates identified (~44%) in the sample set while the Biogen DG44 #1 cell sample had the lowest proportion of duplicates (~21%) (Figure 2A). The duplicate marked data was further pre-processed for variant calling by INDEL realignment and base recalibration in line with the *GATK* best practice guidelines [48,49].

3.3 Identification of CHO cell mitochondrial genome variants

The average depth of coverage and perbase coverage for each cell line sample was calculated using *samtools* [50]. The lowest average depth of coverage across the unshifted reference sequence was observed for the Biogen DG44 #1 cell line (3,319X) while the deepest coverage was observed for the miRNA-NC CHO-K1 #1 cell line (8,056X) (Figure 2B, Supplementary Table 4D & 4E). While a negligible difference in average coverage (~4X) across the entire shifted and unshifted reference sequences was observed, the effectiveness of the dual mapping strategy is illustrated by an average increase in coverage across the first and last 100bp of the mtDNA reference sequence of 1,128X and 1,079X respectively. The coverage at each individual nucleotide position of the reference sequence was found to be extremely deep (Figure 2C) permitting high resolution analysis of the CHO cell mitochondrial genome and confident identification of low frequency heteroplasmic variants across the 22 cell lines.

The Lofreq and VarScan algorithms were utilized in parallel to identify mutations in CHO cell lines when compared with Chinese hamster reference sequence. Only those SNPs and INDELs that (1) were identified by both algorithms, (2) had a minor allele frequency > 1%, (3) the sequencing depth at the mutant position was > 1,500X, (4) no strand bias was observed, and (5) the average Phred scaled base quality score for alternative allele was ≥ Q25 were retained for further analysis. For INDELs, the Q scores of the 10 flanking bases surrounding the variant position were examined and inspected manually. This procedure was carried out for both the unshifted and shifted reference sequences, and mutations within the defined calling regions, combined upon completion to produce the final variant set (Supplementary Figure 2).

In total, 197 mutations (175 SNPs and 21 deletions and an insertion) were identified across the 22 CHO cell lines (Figure 3 and Supplementary Tables 5-7). The SNPs identified corresponded to 99 nucleotide transitions (A↔G or C↔T) and 22 nucleotide transversions (A↔C, A↔T, G↔C or G↔T), yielding a 4.5:1 ts/tv ratio, similar to previous estimates for mutations in mammalian mtDNA [51]. 130 (121 SNPs and 9 deletions and an insertion) individual variant nucleotide positions were detected in one or more of the CHO cell lines. The largest number of variants was identified in the ATCC DG44 and Biogen DG44 #1 samples with 30 and 21 mutations identified respectively (Figure 4A & Supplementary Table 7) while the Biogen DG44 #2 cell line contained 9 mutated mtDNA positions. The least mutated in the cell line panel was the ATCC CHO-S cell line with only a single SNP identified while 6 SNPs were identified in the mitochondrial genome of the Biogen CHO-S cell line. Cell lines from the CHO-K1 lineages varied from 3 mutations to as many as 12 variants. The number of variant positions in the CHO-K1 cell lines originating from Pfizer cell lines derived from the same parental host ranged from 7 to 12. The remaining CHO-K1 cell lines all had less than 10 variant positions.

Of the 37 genes in the mitochondrial genome, 23 were found to have at least one variant position in one of the CHO cell lines sequenced (Figure 4B). A SNP and an insertion were identified in the mtDNA D-loop in the ATCC DG44 and Pfizer CHO-K1 #F5 cell lines. The 16S rRNA gene had the largest number of mutated positions with variants identified at 17 separate nucleotides. All of the 13 protein coding genes had at least one mutation while 8 of 22 tRNAs harboured a mutation. The *CYTB* and *COX1* were found to have the largest number of mutations amongst protein coding genes while the *ND3* gene had only one variant. A homoplasmic (MAF > 99%) SNP in $tRNA^{Val}$ (m.1074C>T) and the 16S rRNA gene (m.2235C>T) was identified in each of the 17 cell lines from the CHO-K1 lineage (Supplementary Table 7). The CHO cell lines from the S and DG44 lineages are identical to the Chinese hamster mtDNA reference sequence at these positions. The two Biogen DG44 cell lines had a shared homoplasmic mutation in $tRNA^{Val}$ (m1092A>G) yet this mutation was not detected in the ATCC DG44 cell line. The Biogen CHO S cell line contains the only homoplasmic mutation in a protein-coding gene, a SNP identified in *CYTB* (m.14311C>T).

3.4 Identification of CHO cell line heteroplasmy

81% of all CHO cell line mutations identified in this study were heteroplasmic (i.e. the MAF>1% & <99%) with a minor allele frequency spanning from 1% to 96.2% (Figure 4C). While the majority of heteroplasmic variants were identified in a single cell line, 11 of these mutations were shared in two or

more cell lines (Table 2). The effectiveness of the dual mapping strategy is further demonstrated by the identification of two variants identified in the first 100bp of the mitochondrial genome with $tRNA^{Phe}$ (m.62C>T) and the 12S rRNA gene (m.74G>A). These low frequency heteroplasmies were not identified by the variant calling pipeline using the unshifted linear reference sequence. Next generation sequencing data from the Chinese hamster liver sample was also analysed using the variant detection pipeline. No heteroplasmic variants were identified in the Chinese hamster mitochondrial genome sequence. We did, however, identify a previously reported heteroplasmic insertion [52] and deletion [53] from the mouse liver mtDNA sequencing data using the mouse mtDNA reference sequence (GenBank accession: NC_005089) and the variant identification pipeline (Supplementary Table 8).

The mouse mtDNA sequencing data was utilised to demonstrate the effectiveness of the bioinformatics pipeline and estimate the potential influence of NumtS contamination on variant calling and heteroplasmy levels. Reads originating from mouse mtDNA sequencing were first analysed using an identical bioinformatics pipeline to the Chinese hamster and CHO cell line analyses, incorporating the shifted and unshifted reference sequence mapping. From this analysis, 3 known [52,53] mouse mtDNA variants were identified - a homoplasmic SNP (m.9461T>C), a heteroplasmic deletion (m.5171delA) and an insertion (m.9820insAA) (Supplementary Table 8). To determine if these variants and their corresponding allele frequencies were influenced by NumtS, we stringently mapped reads against the mouse nuclear genome and extracted reads that aligned to known NumtS [46]. The NumtS aligned reads were remapped to the mouse mtDNA reference sequence using the same alignment and variant pre-processing approach to that of Chinese hamster and CHO cell for cell line data with the exception of variant calling thresholds that were modified to account for the lower depth of coverage. Reads were found to align predominantly to 5 regions in the mouse mtDNA reference sequence (Supplementary Figure 3), which did not overlap with the 3 variants called on the full dataset. Furthermore no SNPs or INDELs were identified in the 5 regions where NumtS reads aligned indicating that NumtS contamination was not a contributing factor in either variant detection or heteroplasmy measurement.

3.5 Prediction of the effect of mitochondrial genome variations

Of the variants identified in this study, 62% (81/130) lie within protein coding regions of the mitochondrial genome. To determine the putative effects of these variants we utilized the snpEff [43] tool to annotate each mutation. Missense variants were the most common, accounting for 60.5% (49/81) of mutations in protein coding regions in comparison to 25% (18/81) of mutations predicted to be synonymous. PROVEAN predicted that 55% (27/49) of the amino acid substitutions arising of missense mutations would affect protein function (Supplementary Table 7). The remaining variants were predicted to result in a frameshift mutation (7.5%) or premature stop codon (5%), with 1 start codon mutation identified. Of the 22 cell lines sequenced, 20 contained mutations that resulted in alteration of the protein coding sequence of at least 1 gene in the mitochondrial genome (Figure 5A). Each protein coding gene harboured at least 1 mutation that altered the amino acid sequence in at least one sample (Figure 5B). Frameshift mutations with a high probability of functional consequences were identified in *ND1*, *COX1*, *ND4*, *ND5* and *CYTB* while *COX1*, *ND4L*, *ND6* while mutations resulting in a premature stop codon were observed in *COX1*, *ND4L*, *ND6* and *CYTB*.

3.6 Is there a relationship between heteroplasmic variants and cell phenotype?

In order to investigate whether the existence of these heteroplasmic variants could infer anything about the phenotypic behaviour of the cell lines, we picked 2 closely related lines for a more detailed analysis of their growth behaviour and oxidative phosphorylation potential. While several of the SNPs and indels were predicted to impact protein function as mentioned above we chose two of the CLD1 with heteroplasmic variants in the CytB gene, CLD1 #3 and CLD1 #5. Fig X shows that Cell line #3 only reached 1.4 x 10^6 cells/ml in culture over 5 days whereas cell line #5 reached 1.8 x 10^6 cells/ml. When stained using Mitotracker, which gives an indication of mitochondrial number, it was apparent that there was not a significant difference between the 2 lines. Interestingly the staining showed a strong anti-correlation with cell growth – with functional mitochondrial content increasing considerably once the cells entered stationary phase. This is in keeping with the theory that rapidly growing cells generate ATP by aerobic glycolysis but switch to oxidative phosphorylation during stationary phase – also referred to as the metabolic shift. This shift in energy generation is typically associated with greater productivity (Ref?). To further investigate whether there were other metabolic differences between the two lines we measured oxygen utilisation by respirometry (Oroboros). This was performed during the mid-exponential phase of growth and showed that the cells that went on to reach higher cell density had reduced oxygen consumption and lower oxphos potential than the cells that peaked at 1.4 x 10^6 cells/ml. Fig Y shows that CLD1 #3 had greater leak, routine and maximum O2 consumption, indicating a higher oxphos potential. This cell line (#5) had a 36% heteroplasmic frameshift mutation in CytB at nt14136 but did not have the stop codon-forming G>A SNP at nt14378. On the other hand cell line #3 had only 7% heteroplasmic variant at nt14236 but 35% heteroplasmy at nt14378. Sanger sequencing suggested that the 2 variants are unlikely to co-exist on the same individual mitochondrial genome (Suppl Fig – Pauls data) which by extension means that there would be a higher overall percentage of mutant CytB in the #3 cell line (43%). It is also interesting to note that this cell line showed an 18% heteroplasmic variant in Cox2 which is predicted to cause an initiator codon change from ATG to ATA and would be expected to reduce translation of that protein. While far from definitive, these observations could contribute to the reduced Oxphos capacity of these cells. Finally, measuring CytB and Cox2 in both lines showed that both were expressed at a higher level in the cells with enhanced oxphos.

4 Discussion

Analytical methods including metabolic flux analysis and metabolomics have shed light on CHO cell metabolic phenotypes, informing media design, feeding strategies and cell line engineering to increase the efficiency of industrial cell culture for biopharmaceutical production [5]. Despite the clear utility of these methods, our understanding of the origins of variation in cellular metabolism in different CHO cell lines, clones and in some cases, between and over the course of production runs have not yet been completely unravelled [6]. Comparison of nuclear genome sequences has shown that CHO cell lineages

harbour distinct mutations, and that millions of SNPs and INDELs arise during the development of a new cell line [20]. This plasticity of the CHO cell nuclear genome sequence undoubtedly plays an underlying role in the range of bioprocess phenotype variation observed amongst CHO cell clones during cell line development. Eukaryotic cells also contain a separate non-nuclear polyploid genome within each mitochondrion. mtDNA sequence variants in the mitochondrial genome have been widely studied in human biology, and the association of specific mutations with a number of metabolic disorders is well established. In recent years, the development of massively parallel sequencing technology has dramatically expanded our ability to study mitochondrial genomics and permitted analysis of low frequency mtDNA heteroplasmy. Here, we present the first survey of mtDNA heterogeneity across CHO cell lineages and cell lines as well as amongst clones produced during cell line development.

Considering the genomic instability of CHO cells, Chinese hamster mtDNA is an ideal common mitochondrial reference genome to compare sequence variants across cell lines. Assessment of the suitability of an existing Chinese hamster sequence available on GenBank (NC_007936) for this purpose revealed that the sequence is from either a $CHOA_L$ or CHO-K1 CHO [47] cell line, not the Chinese hamster, and that a number of sequence features are incorrectly annotated. In order to ensure the accuracy of a reference sequence for comparison of cell lines, we sequenced mtDNA from the Chinese hamster. Comparative analysis of the Chinese hamster mtDNA sequence with the GenBank NC_007936 identified 7 mutations comprising 4 SNPs and 3 INDELs within 4 protein-coding genes and the mitochondrial D-loop control region. The 16,283bp sequence acquired in this study represents the first accurate reference for the analysis of the mitochondrial genome, providing an essential resource for future studies of CHO cell mtDNA.

To determine the prevalence of mutations in the CHO cell mitochondrial genome, 22 CHO cell lines derived from the CHO-K1, CHO-S and DG44 lineages, including industrial cell lines, engineered cell lines and clones generated from 2 cell line development projects were selected for mtDNA sequencing. The high depth of coverage achieved through massively parallel sequencing of the relatively small mitochondrial genome permits accurate identification of homoplasmic and heteroplasmic variants with minor allele frequencies as low as 1%. Established best practices were utilised to pre-process alignments before a conservative variant calling pipeline that required agreement between two algorithms for SNP and INDEL identification.

The heterogeneity of the CHO cell mitochondrial genome across the CHO cell lines analysed here is remarkable; cell lines were found to contain at least one to as many as 30 mutations in their mtDNA. We discovered both homoplasmic and heteroplasmic mutations at 130 nucleotide positions distributed across the entire mtDNA sequence with a total of 197 variants detected across the 22 CHO cell lines. A SNP or INDEL was identified in all 13 protein-coding genes, 8 tRNA genes, and both rRNA genes as well as with the D-loop region. Each of the protein coding genes were found to contain at least one heteroplasmic mutation in at least one cell line with *CYTB*, *COX1* and *ND5* each harbouring at least 10 distinct heteroplasmic variants. Of the 130 variant positions identified, only 4 (37 mutations across the 22 cell lines) were homoplasmic (MAF>99%). All 17 cell lines from the CHO-K1 lineage harboured a

homoplasmic SNP within both $tRNA^{Val}$ (m.1074C>T) and 16S rRNA (m.2235C>T) genes. We also identified a homoplasmic variant within $tRNA^{Val}$ (m.1092A>G) in the two CHO-DG44 cell lines provided by Biogen, yet this mutation was not identified in the CHO-DG44 cell line sourced from the ATCC. The fourth homoplasmic variant was identified in the *CYTB* gene (m.14311C>T) in the Biogen CHO-S cell line. The CHO-K1 and CHO-S were both derived from the original Chinese hamster ovary tissue isolate before being sent to two different laboratories [54]. We did not detect the $tRNA^{Val}$ (m.1074C>T) or 16S rRNA (m.2235C>T) common to CHO-K1 cell lines in the CHO-S cell lines or the cell line sequenced in the Partridge *et al.* study (CHO-K1 cell or derived from CHO_{AL}). In addition, the homoplasmic *CYTB* mutation (m.14311C>T) observed in the Biogen CHO-S cell line was not identified in ATCC CHO-K1 or the ATCC CHO-S cell line. These findings indicate that CHO cell mtDNA mutations have arisen independently between cell lines following isolation from the original Chinese hamster tissue and over time, have become fixed in the mitochondrial genome. The $tRNA^{Val}$ (m.1902A>G) was identified in both CHO-DG44 cell lines provided by Biogen, and once again, this mutation was not present in the ATCC CHO-DG44. These mutations in CHO-DG44, which originated from a non-CHO-K1 or CHO-S lineage [54], imply that homoplasmic mutations in CHO cell mtDNA can occur spontaneously and differ between CHO cell lines of the same lineage.

The overwhelming majority of mutations identified in this study were heteroplasmic and were detected at 126 individual nucleotide positions. Heteroplasmic mutations were identified in tRNA, rRNA and protein coding regions as well as the D-loop control region. Considering the widespread heteroplasmy identified in the mitochondrial genome of CHO cells, it was somewhat surprising that no heteroplasmy was detected in the Chinese hamster mitochondrial genome. A recent study by Li *et al.* [18] reported tissue specific patterns of heteroplasmy in more than 150 humans across 12 tissue types and demonstrated that heteroplasmy is tissue-specific and, in some cases, tissues can be free of heteroplasmy while other tissues in the same individual can contain heteroplasmic mtDNA variants. It is also possible that very low frequency (MAF<1%) variations are present in the Chinese hamster liver tissue sample. While Li *et al.* utilised a > 0.5% MAF threshold, we felt MAF>1% threshold was an appropriate choice for this study, given the depth of mtDNA coverage obtained.

In comparison to homoplasmic mutations which tended to be lineage specific, 89% of heteroplasmic mutations were identified in a single cell line and ranged from 1-96%. For example, the 2 Biogen DG44 CHO cell lines shared a common homoplasmic SNP in the $tRNA^{Val}$ gene (m.1092A>G) yet no shared heteroplasmic variants were common to both cell lines. There were also no shared heteroplasmic mutations identified between several of the CHO-K1 cell lines (i.e. Biogen CHO-K1, Pfizer CHO-K1 2B6, Pfizer CHO-K1 114 cell lines, DCU CHO-K1 SEAP or the 4 DCU sponge transfected SEAP cell lines). Of those heteroplasmic mtDNA mutations that were identified in more than one cell line, the majority were shared amongst clones originating from 2 distinct cell line development projects (CLD1 and CLD2).

PROVEAN predictions found that of the 49 amino acid substitutions arising from a missense mutation, 55% of these are highly likely to result in a functional effect. When comparing the three missense mutations in the ND5 coding sequence (m.11898G>A-18%, m.12078T>A-70%

and m.13065C>G20%), the wild-type Lysine (K) coded at position m.12078 is conserved between the Chinese hamster and Human whereas the other two amino acids are not. The 70% heteroplasmic shift from T>A thereby changing the amino acid coding sequence is likely to elicit some form of dysfunction given the level of conservation of the amino acid as well as the high frequency of heteroplasmy. This particular amino acid location has not been reported previously as a heteroplasmic variant in other cellular models, however a variety of mutations within ND5 have been reported such as m.13565C>T in human mitochondrial cybrids induced the decrease in Ca^{2+} uptake to the mitochondria as well as a dependence on glycolysis for ATP production []. One conserved amino acid location between both human and CHO is Alanine (A11) which was identified to possess a 4.1% (m.9887G>A) missense heteroplasmic mutation in the publically available CHO-DG44 line. This very same mutation has been detected previously in the mitochondrial DNA of esophageal cancer [4]. Although a considerable number of amino acids found to be mutated in this study are conserved from CHO to Human suggesting a functional role, the specific amino acid in question has not yet been identified in other studies.

Seven heteroplasmic variants were identified in 2 or more of the CLD1 clones while 4 heteroplasmies were found in 2 or more of the 3 CLD2 clones. Four shared heteroplasmic variants were identified in at least one clone from CLD1 and CLD 2. A SNP in $tRNA^{Lys}$ (m.7721A>G) was identified in 1 clone from CLD1 (MAF=47%) and 2 clones from CLD2 (MAF=13% and MAF =9.5%). A SNP in *CYTB* (m.14849G>A) was found in a single CLD1 clone and in 3 clones from CLD2. Three of the clones had an average mutation frequency of ~45% yet the MAF of the fourth clone was 1.5%, a marked difference from its counterpart clones within CLD2. The only mutation present in all 8 clones from CLD1 and CLD2 is a frameshift mutation in the *CYTB* gene (m.14136delA) with a MAF ranging from 7.7% to 52%. The m.14136delA mutation also appears in the ATCC CHO-K1 cell line (derived from an isolate of the original Chinese hamster ovary tissue) and it would seem that this mutation has been retained in the CLD project clones yet has been lost in the other CHO-K1 cell lines while the homoplasmic m.1074C>T and m.2235C>T have become fixed in all CHO-K1. The variation of mutation frequencies of heteroplasmic variants both within and across the two cell line development projects, following post single cell cloning, is in line with the model of random assortment of mtDNA upon cell division [55].

It is not possible at this point to determine the impact of rRNA or tRNA mutations without further experimentation. It would be expected, however, that perturbations within the mitochondria's translational machinery would have a considerable impact on the inner mechanics of an energy-producing mitochondrion. Lie *et al.* [*] reported that a homoplasmic T10003C mutation in the $tRNA^{Gly}$ gene caused a 70% reduction in the steady state level of $tRNA^{Gly}$ with an associated 33% reduction in mitochondrial translation. In this instance, it was predicted that this mutation interfered with the formation of the tRNA secondary structure by forming a base pair at 13C-22G in the conserved stem. In our study, 8 tRNA genes were discovered to have heteroplasmic variations present to as far as 74% for $tRNA^{Leu}$ in the case of the Pfizer CHO-K1 2B6 line (m11699G>A). Given the short sequence length of tRNAs (76-90 nt) and their functional dependency on secondary structure, it would be expected the small changes observed in this study would have an effect. Our results do, however, indicate that the mtDNA mutations are not only widespread but also likely to influence mitochondrial function with each

of the protein coding genes in the mitochondrial genome found to contain at least one SNP or INDEL in one cell line and *CYTB*, *COX1* and *ND5* each harbouring at least 10 distinct mutations. A variety of effects were predicted for mutations in protein coding genes including synonymous, non-synonymous and mutations in the initiation codon as well as variants that are likely to have a more pronounced effect on the protein sequence, including premature stop and frameshift mutations. For instance, *CYTB* was found to harbour a SNP (m.14849G>A) predicted to result in a premature stop codon as well as a frameshift mutation (m.14136delA) shared between the CLD1 and CLD2 cell lines. The presence of mtDNA mutations detected here could also play a role in CHO cell metabolic phenotype variation. This gain of stop and frameshift mutation in the *CYTB* gene are predicted to result in a loss of function and could lead to a diminished efficiency in mitochondrial aerobic respiration in the electron transport system centred on complex III. Weakened oxidative phosphorylation (OXPHOS) could signal reprogramming of cellular metabolism to rely more heavily on glycolysis [56] and maintain cellular energy balance. Exclusive reliance on glycolytic metabolism, despite being energetically inefficient, has been shown to be associated with elevated cell growth in both cancer and CHO cells due to the intermediate metabolites of glycolysis that contribute to biomass accumulation [57]. All but one protein coding variant identified were heteroplasmic and in some cases, loss of function mutations spanned a wide range. Mutation frequencies were found to vary by as much as 1.5% to 50% e.g. a *ND1* frameshift (m.3205delCT), indicating that the potential effects of these mutations could lead to a continuous distribution of phenotypes. Extensive metabolic profiling of the panel of 22 CHO cell lines utilised in this study overlapped with this comprehensive mutational data would begin to answer these questions of variant translating to phenotype.

As mentioned previously, three protein-coding genes across the 22 cell lines harboured at least 10 distinct mutations in the case of *CYTB, COX1* and *ND5*. However, in some instances, numerous mutations within the same cell line of the same gene was detected such as in the case of the *ND5* gene in the Biogen DG44 #1 cell line. In this case, *ND5* was observed to contain 4 mutations, 3 missense (m11,898G>A, m12,078T>C and m13,065C>G) occurring at a frequency of 18%, 70% and 20%, respectively, as well as a synonymous mutation at m13,157A>G (8.9%). These three missense mutations potentially changing the amino acid sequence could all impact negatively on the functional role of *ND5*. Taking into account the frequency of each mutation, if each mutant exists in isolation then a 100% dysfunctional *ND5* protein could prevail. If however, these mutations co-inhabit the same mtDNA genome, then wild-type protein could remain thereby not breaking the biochemical threshold. It has previously been shown that heteroplasmic mutations in *ND5* results in the disruption of NADH dehydrogenase assembly which is associated with a weakened OXPHOS and an increase in lactate production due to glycolytic dependency []. Cellular mitochondrial compartmentalisation would have to be determined in order to elucidate the partitioning of this mutational pool and predict whether low-frequency heteroplasmic variants remain segregated and functional []. To get a better understanding of this, we amplified the region of the *CYTB* gene that was identified to harbour two heteroplasmic mutations at m14136 and m.14378 in three clones from the cell line development panel (CLD1 #3, CLD1 #5 and CLD2 #2). TOPO cloning was performed and 10 clones from each cell line was Sanger sequenced as a means to determine the co-habitation of these mutations (Supplementary Figure S4).

When compared to the CH mitochondrial reference sequence, the early frameshift mutation, shared in all clones, was detected and reflected the heteroplasmic frequency identified through deep-sequencing. In the case of CLD1 #3 which contains both mutations, the second STOP mutation was not detected within the small panel of clones suggesting that these two mutations do not co-inhabit the same mtDNA copy further suggesting that the dysfunctional *CYTB* protein that results from these two mutants will have a synergistic effect. With such a small panel of clones, however, the later mutation was not detected. This highlights the potential impact that several low MAF variations can have on mitochondrial function if present on individual genomic copies and working in unison.

Phenotypic analysis of oxphos potential, growth characteristics and mitochondrial content of two cell lines in particular provided some tantalizing evidence of how heteroplasmic variants in important mitochondrial encoded proteins could impact on the suitability of certain cell lines for recombinant protein production. However, more targeted mitochondrial genome engineering studies in future should ascertain whether these variants are directly responsible for the phenotypes observed and indeed whether they can be manipulated to improve these characteristics. Some the molecular tools required to achieve this have only recently been developed (cell paper – mito talens).

While the results of this study demonstrate the heterogeneity of CHO cell mtDNA, the polyploid nature of the mitochondrial genome presents considerable challenges to understanding the impact of mitochondrial mutations. A mtDNA variant might be spread across the entire population or confined to a subpopulation of cells. At the subcellular level, the mutation might be distributed across multiple mitochondria or indeed confined to a limited number of mitochondria. The emergence of new methods to sequence mtDNA at the single cell level [55] will play a valuable role in future studies and increase our understanding of the implications of particular mutation in a CHO cell population. It will also be important to integrate MFA and metabolomics analysis with mtDNA sequencing to understand the biochemical threshold at which individual mutations affect CHO cell behaviour. The knowledge gained in doing so has the potential to enable precise cell line selection and ultimately rational genetic engineering to improve CHO cell phenotypes. The recent development of mtDNA CRISPR-Cas9 based methods [58] and mitoTALENS [59] for site specific mitochondrial genome editing provide routes to rational engineering of CHO cell mitochondria to enhance the metabolic performance of CHO cells for biopharmaceutical manufacture.

5 Conclusions

Widespread heteroplasmy in the CHO cell mitochondrial genome raises intriguing questions about the genetics and selection of mitochondrial mutations in CHO cells during cell culture and cell line development for biopharmaceutical production. Closely related clones derived from the same parental host and even originating from the same cell line development project can harbour distinct heteroplasmic variations. These variations in the mitochondrial genome are likely to affect mitochondrial function and could play a role in cell to cell, production run to production run and indeed clone to clone variation observed in CHO cell culture and cell line development. The combination of mtDNA sequencing with established techniques in metabolic flux analysis and metabolomics will be necessary to associate these mutations with desirable or undesirable CHO cell metabolism. The understanding of

mtDNA variation could lead to new approaches to cell line screening and ultimately engineering of CHO cell mtDNA for more productive metabolic phenotypes.

Acknowledgements

The authors gratefully acknowledge funding from Science Foundation Ireland (grant refs: 13/SIRG/2084, 13/IA/1841and 13/IA/1963) and the eCHO systems Marie Curie ITN programme (grant ref: 642663). The authors would also to acknowledge Lin Zhang (Pfizer Inc.), Scott Estes, Chapman Wright and Brian St. Germaine (Biogen Inc.) for access to cell lines and comments on the manuscript.

References

[1] Wurm, F.M., Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat. Biotechnol.* 2004, 22, 1393–1398.

[2] Li, F., Vijayasankaran, N., Shen, A. (Yijuan), Kiss, R., Amanullah, A., Cell culture processes for monoclonal antibody production. *mAbs* 2010, 2, 466–477.

[3] Butler, M., Animal cell cultures: recent achievements and perspectives in the production of biopharmaceuticals. *Appl. Microbiol. Biotechnol.* 2005, 68, 283–291.

[4] Vander Heiden, M.G., Cantley, L.C., Thompson, C.B., Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science* 2009, 324, 1029–1033.

[5] Quek, L.-E., Dietmair, S., Krömer, J.O., Nielsen, L.K., Metabolic flux analysis in mammalian cell culture. *Metab. Eng.* 2010, 12, 161–171.

[6] Young, J.D., Metabolic flux rewiring in mammalian cell cultures. *Curr. Opin. Biotechnol.* 2013, 24.

[7] Zagari, F., Jordan, M., Stettler, M., Broly, H., Wurm, F.M., Lactate metabolism shift in CHO cell culture: the role of mitochondrial oxidative activity. *New Biotechnol.* 2013, 30, 238–245.

[8] Gilbert, A., McElearney, K., Kshirsagar, R., Sinacore, M.S., Ryll, T., Investigation of metabolic variability observed in extended fed batch cell culture. *Biotechnol. Prog.* 2013, 29, 1519–1527.

[9] Rizzuto, R., De Stefani, D., Raffaello, A., Mammucari, C., Mitochondria as sensors and regulators of calcium signalling. *Nat. Rev. Mol. Cell Biol.* 2012, 13, 566–578.

[10] Wang, C., Youle, R.J., The Role of Mitochondria in Apoptosis. *Annu. Rev. Genet.* 2009, 43, 95–118.

[11] Calvo, S., Jain, M., Xie, X., Sheth, S.A., et al., Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat. Genet.* 2006, 38, 576–582.

[12] Miller, F.J., Rosenfeldt, F.L., Zhang, C., Linnane, A.W., Nagley, P., Precise determination of mitochondrial DNA copy number in human skeletal and cardiac muscle by a PCR-based assay: lack of change of copy number with age. *Nucleic Acids Res.* 2003, 31, e61.

[13] Liou, C.-W., Lin, T.-K., Chen, J.-B., Tiao, M.-M., et al., Association between a common mitochondrial DNA D-loop polycytosine variant and alteration of mitochondrial copy number in human peripheral blood cells. *J. Med. Genet.* 2010, 47, 723–728.

[14] Wachsmuth, M., Hübner, A., Li, M., Madea, B., Stoneking, M., Age-Related and Heteroplasmy-Related Variation in Human mtDNA Copy Number. *PLOS Genet* 2016, 12, e1005939.

[15] Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H.L., et al., Sequence and organization of the human mitochondrial genome. *Nature* 1981, 290, 457–465.

[16] Barrell, B.G., Anderson, S., Bankier, A.T., de Bruijn, M.H., et al., Different pattern of codon recognition by mammalian mitochondrial tRNAs. *Proc. Natl. Acad. Sci. U. S. A.* 1980, 77, 3164–3166.

[17] Payne, B.A.I., Wilson, I.J., Yu-Wai-Man, P., Coxhead, J., et al., Universal heteroplasmy of human mitochondrial DNA. *Hum. Mol. Genet.* 2013, 22, 384–390.

[18] Li, M., Schröder, R., Ni, S., Madea, B., Stoneking, M., Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations. *Proc. Natl. Acad. Sci.* 2015, 112, 2491–2496.

[19] Xu, X., Nagarajan, H., Lewis, N.E., Pan, S., et al., The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nat Biotech* 2011, 29, 735–741.

[20] Lewis, N.E., Liu, X., Li, Y., Nagarajan, H., et al., Genomic landscapes of Chinese hamster ovary cell lines as revealed by the Cricetulus griseus draft genome. *Nat Biotech* 2013, 31, 759–765.

[21] Brinkrolf, K., Rupp, O., Laux, H., Kollin, F., et al., Chinese hamster genome sequenced from sorted chromosomes. *Nat Biotech* 2013, 31, 694–695.

[22] Kaas, C.S., Kristensen, C., Betenbaugh, M.J., Andersen, M.R., Sequencing the CHO DCB11 genome reveals regional variations in genomic stability and haploidy. *BMC Genomics* 2015, 16, 160.

[23] Cao, Y., Kimura, S., Itoi, T., Honda, K., et al., Construction of BAC-based physical map and analysis of chromosome rearrangement in Chinese hamster ovary cell lines. *Biotechnol. Bioeng.* 2012, 109, 1357–1367.
[24] Meleady, P., Hoffrogge, R., Henry, M., Rupp, O., et al., Utilization and evaluation of CHO-specific sequence databases for mass spectrometry based proteomics. *Biotechnol. Bioeng.* 2012, 109, 1386–1394.
[25] Gerstl, M.P., Hackl, M., Graf, A.B., Borth, N., Grillari, J., Prediction of transcribed PIWI-interacting RNAs from CHO RNAseq data. *J. Biotechnol.* 2013, 166, 51–57.
[26] Diendorfer, A.B., Hackl, M., Klanert, G., Jadhav, V., et al., Annotation of additional evolutionary conserved microRNAs in CHO cells from updated genomic data. *Biotechnol. Bioeng.* 2015, n/a-n/a.
[27] Birzele, F., Schaub, J., Rust, W., Clemens, C., et al., Into the unknown: expression profiling without genome sequence information in CHO by next generation sequencing. *Nucleic Acids Res.* 2010, 38, 3999–4010.
[28] Rupp, O., Becker, J., Brinkrolf, K., Timmermann, C., et al., Construction of a Public CHO Cell Line Transcript Database Using Versatile Bioinformatics Analysis Pipelines. *PLoS ONE* 2014, 9, e85568.
[29] Wippermann, A., Rupp, O., Brinkrolf, K., Hoffrogge, R., Noll, T., The DNA methylation landscape of Chinese hamster ovary (CHO) DP-12 cells. *J. Biotechnol.* 2015.
[30] Clarke, C., Henry, M., Doolan, P., Kelly, S., et al., Integrated miRNA, mRNA and protein expression analysis reveals the role of post-transcriptional regulation in controlling CHO cell growth rate. *BMC Genomics* 2012, 13, 656.
[31] 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* 2013, 55, 133–136.
[32] Hahn, C., Bachmann, L., Chevreux, B., Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Res.* 2013, 41, e129.
[33] Magoč, T., Salzberg, S.L., FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinforma. Oxf. Engl.* 2011, 27, 2957–2963.
[34] Bernt, M., Donath, A., Jühling, F., Externbrink, F., et al., MITOS: improved de novo metazoan mitochondrial genome annotation. *Mol. Phylogenet. Evol.* 2013, 69, 313–319.
[35] Laslett, D., Canbäck, B., ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. *Bioinformatics* 2008, 24, 172–175.
[36] Bolger, A.M., Lohse, M., Usadel, B., Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014, 30, 2114–2120.
[37] Li, H., Durbin, R., Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* 2009, 25, 1754–1760.
[38] Ding, J., Sidore, C., Butler, T.J., Wing, M.K., et al., Assessing Mitochondrial DNA Variation and Copy Number in Lymphocytes of ~2,000 Sardinians Using Tailored Sequencing Analysis Tools. *PLOS Genet* 2015, 11, e1005306.
[39] Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., et al., in:, *Curr. Protoc. Bioinforma.*, John Wiley & Sons, Inc., 2002.
[40] Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., et al., VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012, 22, 568–576.
[41] Wilm, A., Aw, P.P.K., Bertrand, D., Yeo, G.H.T., et al., LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* 2012, 40, 11189–11201.
[42] Guo, Y., Li, J., Li, C.-I., Long, J., et al., The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* 2012, 13, 666.

[43] Cingolani, P., Platts, A., Wang, L.L., Coon, M., et al., A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* 2012, 6, 80–92.
[44] Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., Chan, A.P., Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLOS ONE* 2012, 7, e46688.
[45] Just, R.S., Irwin, J.A., Parson, W., Mitochondrial DNA heteroplasmy in the emerging field of massively parallel sequencing. *Forensic Sci. Int. Genet.* 2015, 18, 131–139.
[46] Calabrese, F.M., Simone, D., Attimonelli, M., Primates and mouse NumtS in the UCSC Genome Browser. *BMC Bioinformatics* 2012, 13 Suppl 4, S15.
[47] Partridge, M.A., Davidson, M.M., Hei, T.K., The complete nucleotide sequence of Chinese hamster (Cricetulus griseus) mitochondrial DNA. *DNA Seq. J. DNA Seq. Mapp.* 2007, 18, 341–346.
[48] Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., et al., From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al* 2013, 43, 11.10.1-33.
[49] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., et al., The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010, 20, 1297–1303.
[50] Li, H., Handsaker, B., Wysoker, A., Fennell, T., et al., The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009, 25, 2078–2079.
[51] Belle, E.M.S., Piganeau, G., Gardner, M., Eyre-Walker, A., An investigation of the variation in the transition bias among various animal mitochondrial DNA. *Gene* 2005, 355, 58–66.
[52] Fan, W., Lin, C.S., Potluri, P., Procaccio, V., Wallace, D.C., mtDNA lineage analysis of mouse L-cell lines reveals the accumulation of multiple mtDNA mutants and intermolecular recombination. *Genes Dev.* 2012, 26, 384–394.
[53] Bayona-Bafaluy, M.P., Acín-Pérez, R., Mullikin, J.C., Park, J.S., et al., Revisiting the mouse mitochondrial DNA sequence. *Nucleic Acids Res.* 2003, 31, 5349–5355.
[54] Wurm, F.M., CHO Quasispecies—Implications for Manufacturing Processes. *Processes* 2013, 1, 296–311.
[55] Jayaprakash, A.D., Benson, E.K., Gone, S., Liang, R., et al., Stable heteroplasmy at the single-cell level is facilitated by intercellular exchange of mtDNA. *Nucleic Acids Res.* 2015, gkv052.
[56] Zheng, J., Energy metabolism of cancer: Glycolysis versus oxidative phosphorylation (Review). *Oncol. Lett.* 2012, 4, 1151–1157.
[57] Templeton, N., Dean, J., Reddy, P., Young, J.D., Peak antibody production is associated with increased oxidative metabolism in an industrially relevant fed-batch CHO cell culture. *Biotechnol. Bioeng.* 2013, 110, 2013–2024.
[58] Jo, A., Ham, S., Lee, G.H., Lee, Y.-I., et al., Efficient Mitochondrial Genome Editing by CRISPR/Cas9, Efficient Mitochondrial Genome Editing by CRISPR/Cas9. *BioMed Res. Int. BioMed Res. Int.* 2015, 2015, 2015, e305716.
[59] Bacman, S.R., Williams, S.L., Pinto, M., Peralta, S., Moraes, C.T., Specific elimination of mutant mitochondrial genomes in patient-derived cells by mitoTALENs. *Nat. Med.* 2013, 19, 1111–1113.
[60] Krzywinski, M.I., Schein, J.E., Birol, I., Connors, J., et al., Circos: An information aesthetic for comparative genomics. *Genome Res.* 2009.

Figure legends

Figure 1: The *Cricetulus griseus* mitochondrial genome. mtDNA was extracted from Chinese hamster liver tissue, deep sequenced and reconstructed using the MITOBIM algorithm. The resulting 16,283bp mitochondrial DNA sequence had an average depth of coverage of ~6,417X (the depth of coverage at each base is shown as a grey histogram within the inner circle of the plot). 13 proteins, 22 tRNAs, 2 ribosomal RNAs and the non-coding D-loop control region are encoded by the Chinese hamster mitochondrial genome. 28 genes are present on the H-strand while 9 genes are on the L-strand of the mtDNA. Comparison of the Chinese hamster mtDNA sequence to the previously sequenced CHO cell line sequence on GenBank (accession no. NC_007936) revealed 7 mutations (4 SNPs and 3 INDELs) in protein coding genes (*COX3, ND4, ND5, ND6*) and the D-loop. The plot was generated using circos v.0.67 [60].

Figure 2: High resolution sequencing of the CHO cell mitochondrial genome. (A) Number of mapped reads for the 22 CHO cell lines illustrating the number of unmapped reads (MAPQ < 20), number of PCR duplicates and number of uniquely mapped reads. Note: a negligible difference in the total numbers of reads mapped against the shifted reference sequence were observed (Supplementary Table 4B & 4C). **(B)** Average sequencing depth for each CHO cell line sequenced. The average depth of coverage combines both variant calling regions within the unshifted and shifted reference sequences used for variant detection. **(C)** The mean (blue), minimum and maximum depth of coverage at each base position across the 22 CHO cell mitochondrial genomes sequenced. The utilisation of the dual mapping approach resulted in a 1,128X and 1,079X increase in coverage for the first and last 100bp of the mtDNA reference sequence respectively.

Figure 3: The mutational landscape of CHO cell line mitochondrial genome. A total of 197 mutations (175 SNPs and 22 INDELs) were identified for 22 CHO cell line samples in comparison the Chinese hamster mitochondrial genome at 130 positions in the mtDNA sequence. The nucleotide alternations detected for each cell line are shown along with the mtDNA feature harbouring the mutation. All protein coding genes along with the D-loop and 8 tRNA genes had a least 1 variant in 1 CHO cell line.

Figure 4: Mitochondrial heteroplasmy is widespread in CHO cell lines. (A) Number of SNPs and INDELs identified for 22 CHO cell line mitochondrial genomes sequenced. The mtDNA of 2 of the 3 DG44 cell lines sequenced were found to have the most variant positions while the ATCC CHO-S cell line was had only a single mutated position. **(B)** Number of variant positions for the 24 genes found to harbour at least one mutation across the 22 CHO cell lines sequenced. **(C)** The majority of variants identified were heteroplasmic - 160 mtDNA mutations out of 197 mutation variants had an allele frequency < 99%.

Figure 5: Prediction of the effects of CHO cell mitochondrial genome mutations. The majority of mutations identified in this study occurred in protein coding genes. **(A)** 22 of the CHO cell lines sequenced contained at least one mutation that altered the protein sequence. **(B)** Each protein coding gene harboured a mutation that altered the protein sequence in at least one of the cell lines.

Table legends

Table 1: Comparison of the Chinese hamster mitochondrial genome to the previously sequenced CHO cell line. 7 variants were identified within 4 protein coding genes as well as the D-loop control region. The position, mtDNA gene, type and nucleotide change observed are shown for each variant. Chinese hamster reads were mapped against the CHO cell line mtDNA sequence (GenBank accession no. NC_007936) to determine the number of reads supporting each variant. All variants were homoplasmic (allele frequency > 99%).

Table 2: Heteroplasmic mitochondrial variants identified in 2 or more CHO cell lines. In total 11 individual heteroplasmic mutations were identified in two or more cell lines. A number of heteroplasmic sites were shared from the cohort of cells acquired from an identical cell line development project. 4 heteroplasmic sites were shared amongst the 4 fast or 4 slow growing CHO cell lines clones. A heteroplasmic frameshift mutation in *Cytb* was identified in all clones analysed.

Supplementary Tables

Supplementary Table 1: Sample Information. The origin of each sample sequenced in this study along with the respective file names corresponding to mtDNA sequencing data.

Supplementary Table 2: Primer sequences for fragment amplification of the CHO cell mitochondrial genome as well as mtDNA isolated from Chinese hamster and mouse mtDNA. Primers were designed for PCR amplification of genomic CHO DNA using the CHOgenome.org database. A beta-actin control was designed against the whole CHO genome sequence with the reverse primers spanning an intron/exon junction. A small positive control (PC) for the CHO mitochondrial genome was designed using the available CHO-K1 mitochondrial genome sequence on CHOgenome.org. Primer pairs (#1 and #2) were also designed using the available CHO-K1 mitochondrial genome sequence for high-fidelity PCR of the mitochondrial genomic DNA fragments. Finally, to account for possible sequence variation between the CHO mt-DNA sequence and the Chinese hamster mt-DNA sequence, primer sets (CH1-6/8 and CH2-86/7) were designed used the available CHO-K1 mt-DNA sequence available and matched for 100% sequence similarity with both the mouse (mmu) and rat (rno) mt-DNA sequences.

Supplementary Table 3: Chinese hamster mtDNA annotation. Coordinates for each of the 38 annotated features in the *C.griseus* mitochondrial genome.

Supplementary Table 4: Pre-processed read counts, mapping rates and average depth of coverage for CHO cell line sequencing. The number of reads remaining following pre-processing along mapping and average coverage rate for each sample following shifted and unshifted sequence alignment.

Supplementary Table 5: SNP calling outputs. Variant calling outs for 22 individual CHO cell lines including the depth at coverage at each SNP, number of forward and reverse reads supporting the reference and alternative allele, average base quality (Q) for each nucleotide at the SNP position and snpEff annotation.

Supplementary Table 6: INDEL detection data. Variant calling outputs for 22 individual CHO cell lines including the depth at coverage at each INDEL, number of forward and reverse reads supporting the reference and alternative allele, average base quality (Q) at the INDEL position as well 5 upstream and downstream flanking regions nucleotide and the snpEff annotation.

Supplementary Table 7: mtDNA SNP and INDEL summary. Base changes along with allele frequency at each variant position for the 22 CHO cell lines. Mutation annotations outputted by snpEff tool are provided for each SNP and INDEL. For missense variants the Provean classification and score are included.

Supplementary Table 8: Detection of mutations in mtDNA isolated from the mouse liver sample. Using an identical dual mapping approach aligning reads to the unshifted and shifted mouse mitochondrial genome reference sequence we identified 3 mutations, two of which were heteroplasmic.

Supplementary Figures

Supplementary Figure 1: mtDNA extraction, isolation and amplification. Ethidium bromide stained agarose gels were ran to assess the quality of **(A)** Genomic DNA isolated from liver samples of Chinese Hamster using a DNeasy Kit with high integrity DNA running at ~100kb. **(B)** mitochondrial plasmid DNA isolated from CHO cell lines using a modified plasmid mini-prep kit with plasmid DNA running out as two bands (coiled and super-coiled) and **(C)** Mitochondrial genomic DNA fragments amplified by high-fidelity PCR and visualised individually and pooled

Supplementary Figure 2: Dual mapping strategy and variant calling. In this study we utilized a dual mapping strategy to account for the mitochondrial genome circularity to maximize depth of coverage at the beginning and end of the mtDNA sequence. We also required agreement between two different variant calling algorithms for a SNP or INDEL to be reported.

Supplementary Figure 3: Alignment of potential NumtS to mtDNA. The mouse mtDNA genome was utilised as a control to determine if NumtS were influencing heteroplasmy measurements. Reads were first stringently aligned to know NumtS region the mm9 genome. Those reads that aligned to NumtS were extracted and remapped against the mouse mitochondrial region. The IGV diagram illustrates that potential NumtS reads aligned to 5 regions. The grey bars illustrate the location of reads that align to both the mouse nuclear and mitochondrial genomes demonstrating that Numt contamination did not affect the 3 variants identified. In addition, no SNPs or INDELs were identified within the 5 regions.

Supplementary Figure 4: Alignment of amplified *CYTB* fragment sequences. Three clones from the two cell line development projects were selected for the amplification and Sanger sequencing of a short fragment of the CYTB gene which included an identified frameshift and STOP mutation. Sequences were aligned using the online software MultAlin. Sequence variations are highlighted in blue with full conservation in red. Each sequences starts with the initiating start ATG codon.

Tables

Table 1

Position	Gene	Variant type	CHO Cell allele	*C.griseus* allele	# reads supporting variant
9306	*COX3*	SNP	A	G	5,201
11399	*ND4*	SNP	C	A	8,765
12683	*ND5*	SNP	G	A	11,464
13436	*ND6*	Deletion	AT	A	8,815
13456	*ND6*	Insertion	C	CA	8,461
15717	D-loop	Deletion	GC	C	7,522
15888	D-loop	SNP	G	T	8,056

Table 2.

Gene	Position	Alelle		Type	Minor Alelle Freq. (%)		Effect	CHO Cell lines
		Ref	Alt		Min	Max		
12S rRNA	105	G	A	SNP	1.7	3	Non-coding	Pfizer CHO-K1 CLD1 #1 Pfizer CHO-K1 CLD1 #2 Pfizer CHO-K1 CLD1 #4
16S rRNA	1575	G	A	SNP	1.2	50	Non-coding	Pfizer CHO-K1 CLD2 #3 ATCC DGG44
16S rRNA	2151	C	T	SNP	1.2	9.9	Non-coding	Pfizer CHO-K1 CLD1 #1 Pfizer CHO-K1 CLD1 #2 Pfizer CHO-K1 CLD1 #3 Pfizer CHO-K1 CLD1 #4
ND1	3205	GCT	G	Deletion	1.5	50	Frameshift variant	Pfizer CHO-K1 CLD1 #5 Pfizer CHO-K1 CLD2 #2 Pfizer CHO-K1 CLD2 #3
COX2	6996	G	A	SNP	14	26	Initiation codon variation	Pfizer CHO-K1 CLD1 #1 Pfizer CHO-K1 CLD1 #2 Pfizer CHO-K1 CLD1 #3 Pfizer CHO-K1 CLD1 #4
tRNA^{Lys}	7721	A	G	SNP	9.5	47	Non-coding	Pfizer CHO-K1 CLD1 #5 Pfizer CHO-K1 CLD2 #1 Pfizer CHO-K1 CLD2 #3
ATP6	8067	C	T	SNP	1.8	6.8	Missense variation	Pfizer CHO-K1 CLD1 #1 Pfizer CHO-K1 CLD1 #2 Pfizer CHO-K1 CLD1 #3 Pfizer CHO-K1 CLD1 #4
ND4	11431	TCA	T	Deletion	1.7	2.0	Frameshift variant	Biogen CHO-K1 ATCC DGG44 Biogen DG44 #2
tRNA^{Leu}	11659	AG	A	Deletion	1.2	2.8	Non-coding	Pfizer CHO-K1 CLD1 #1 Pfizer CHO-K1 CLD1 #2
CYTB	14136	GA	G	Deletion	7.7	52	Frameshift variant	ATCC CHO-K1 Pfizer CHO-K1 CLD1 #1 Pfizer CHO-K1 CLD1 #2 Pfizer CHO-K1 CLD1 #3 Pfizer CHO-K1 CLD1 #4 Pfizer CHO-K1 CLD1 #5 Pfizer CHO-K1 CLD2 #1 Pfizer CHO-K1 CLD2 #2 Pfizer CHO-K1 CLD2 #3
	14378	G	A	SNP	11	36.0	Premature stop codon	Pfizer CHO-K1 CLD1 #1 Pfizer CHO-K1 CLD1 #2 Pfizer CHO-K1 CLD1 #3 Pfizer CHO-K1 CLD1 #4
	14849	G	A	SNP	1.5	53	Missense variation	Pfizer CHO-K1 CLD1 #5 Pfizer CHO-K1 CLD2 #1 Pfizer CHO-K1 CLD2 #2 Pfizer CHO-K1 CLD2 #3
	15136	C	A	SNP	1.7	19	Missense variation	Pfizer CHO-K1 CLD1 #3 Biogen DG44 #1

Figure 1

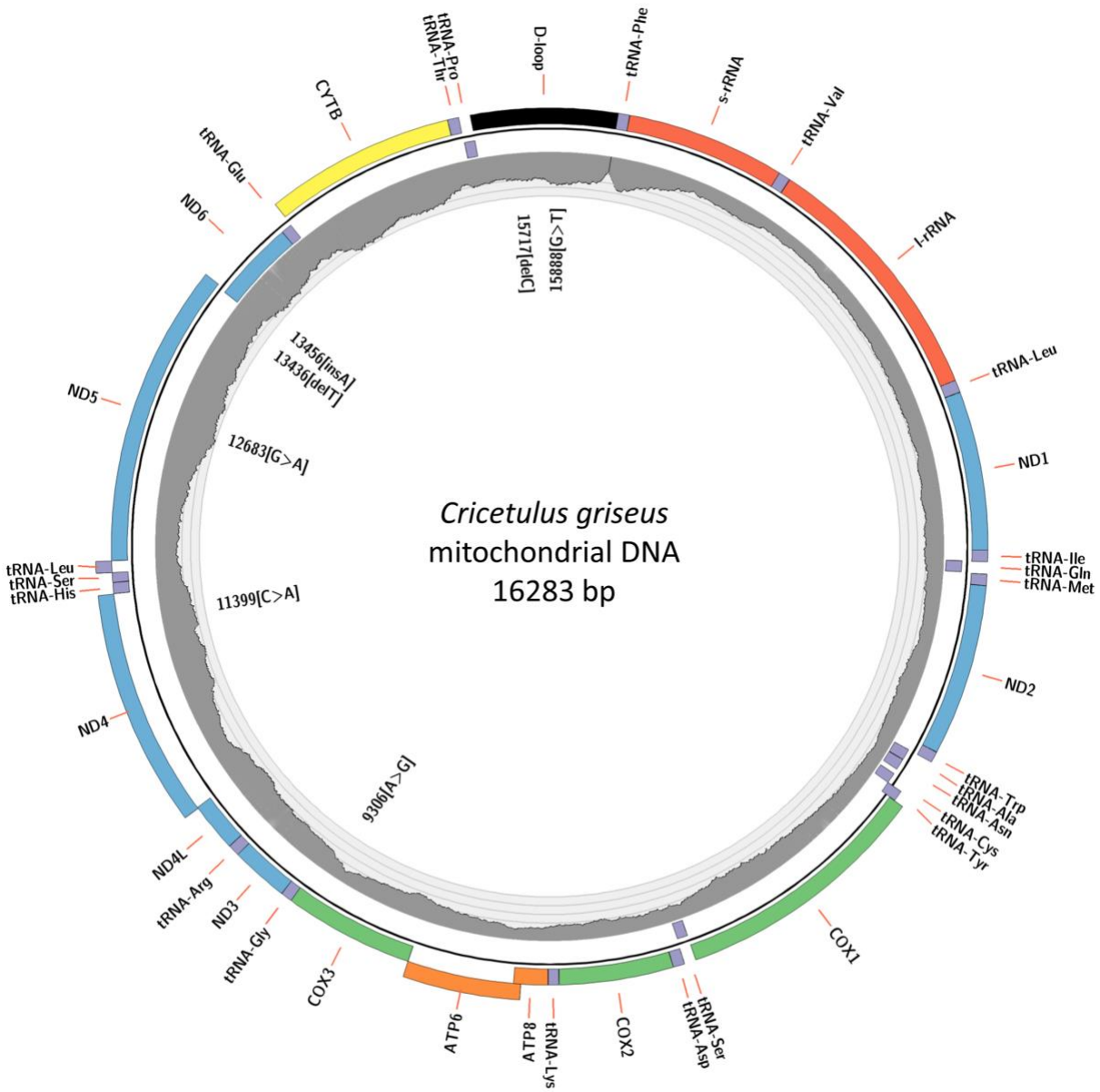

Cricetulus griseus
mitochondrial DNA
16283 bp
D-loop
tRNA-Phe
s-rRNA
tRNA-Val
l-rRNA
tRNA-Leu
ND1
tRNA-Ile
tRNA-Gln
tRNA-Met
ND2
tRNA-Trp
tRNA-Ala
tRNA-Asn
tRNA-Cys
tRNA-Tyr
COX1
tRNA-Ser
tRNA-Asp
COX2
tRNA-Lys
ATP8
ATP6
COX3
tRNA-Gly
ND3
tRNA-Arg
ND4L
ND4
tRNA-His
tRNA-Ser
tRNA-Leu
ND5
ND6
tRNA-Glu
CYTB
tRNA-Thr
tRNA-Pro
15888[G>T]
15717[delC]
13456[insA]
13436[delT]
12683[G>A]
11399[C>A]
9306[A>G]

Figure 2

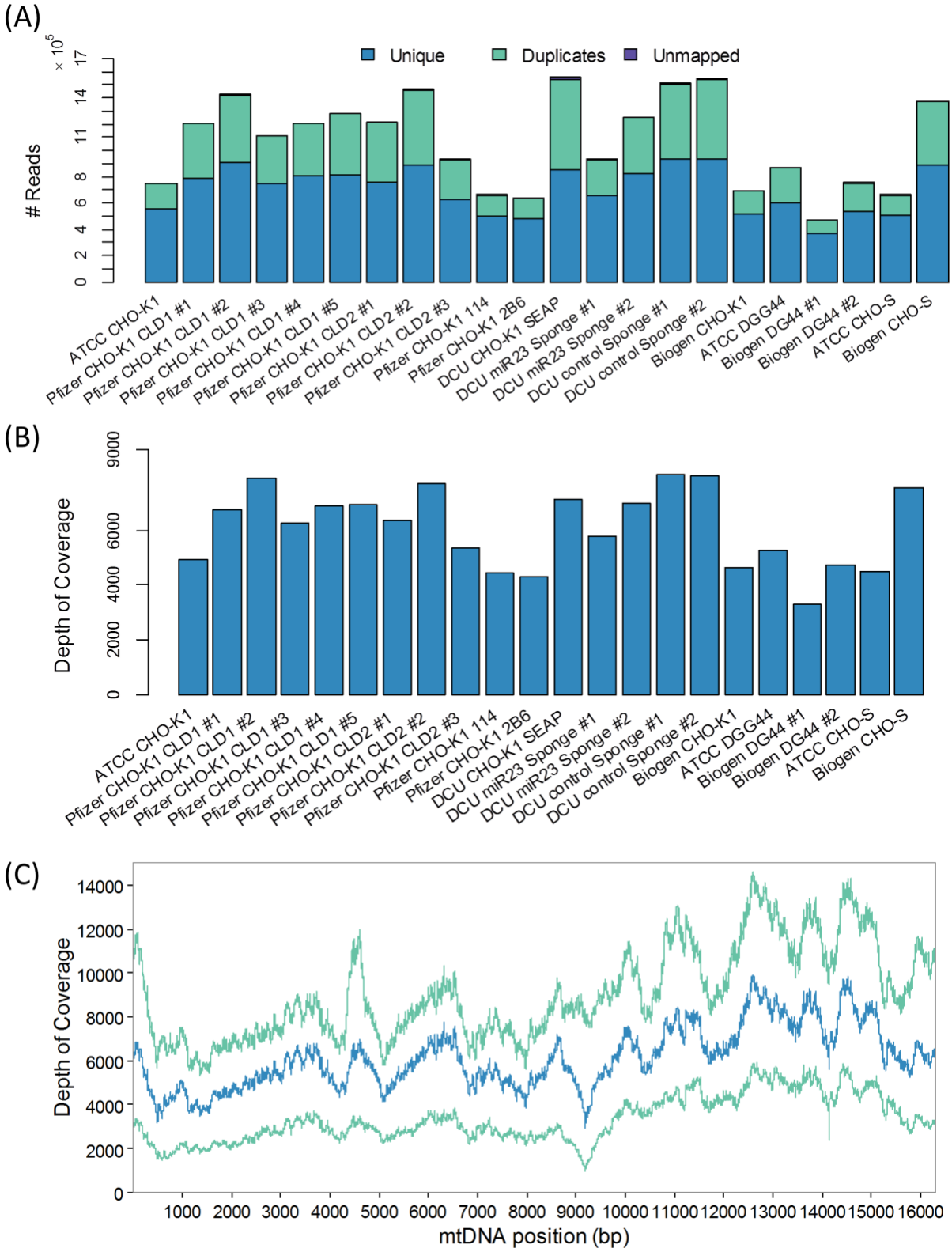

(A)
Unique
Duplicates
Unmapped
Reads
× 10^5
0
2
4
6
8
11
14
17
ATCC CHO-K1
Pfizer CHO-K1 CLD1 #1
Pfizer CHO-K1 CLD1 #2
Pfizer CHO-K1 CLD1 #3
Pfizer CHO-K1 CLD1 #4
Pfizer CHO-K1 CLD1 #5
Pfizer CHO-K1 CLD2 #1
Pfizer CHO-K1 CLD2 #2
Pfizer CHO-K1 CLD2 #3
Pfizer CHO-K1 114
Pfizer CHO-K1 2B6
DCU CHO-K1 SEAP
DCU miR23 Sponge #1
DCU miR23 Sponge #2
DCU control Sponge #1
DCU control Sponge #2
Biogen CHO-K1
ATCC DGG44
Biogen DG44 #1
Biogen DG44 #2
ATCC CHO-S
Biogen CHO-S
(B)
Depth of Coverage
0
2000
4000
6000
9000
(C)
Depth of Coverage
0
2000
4000
6000
8000
10000
12000
14000
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
mtDNA position (bp)

Figure 3

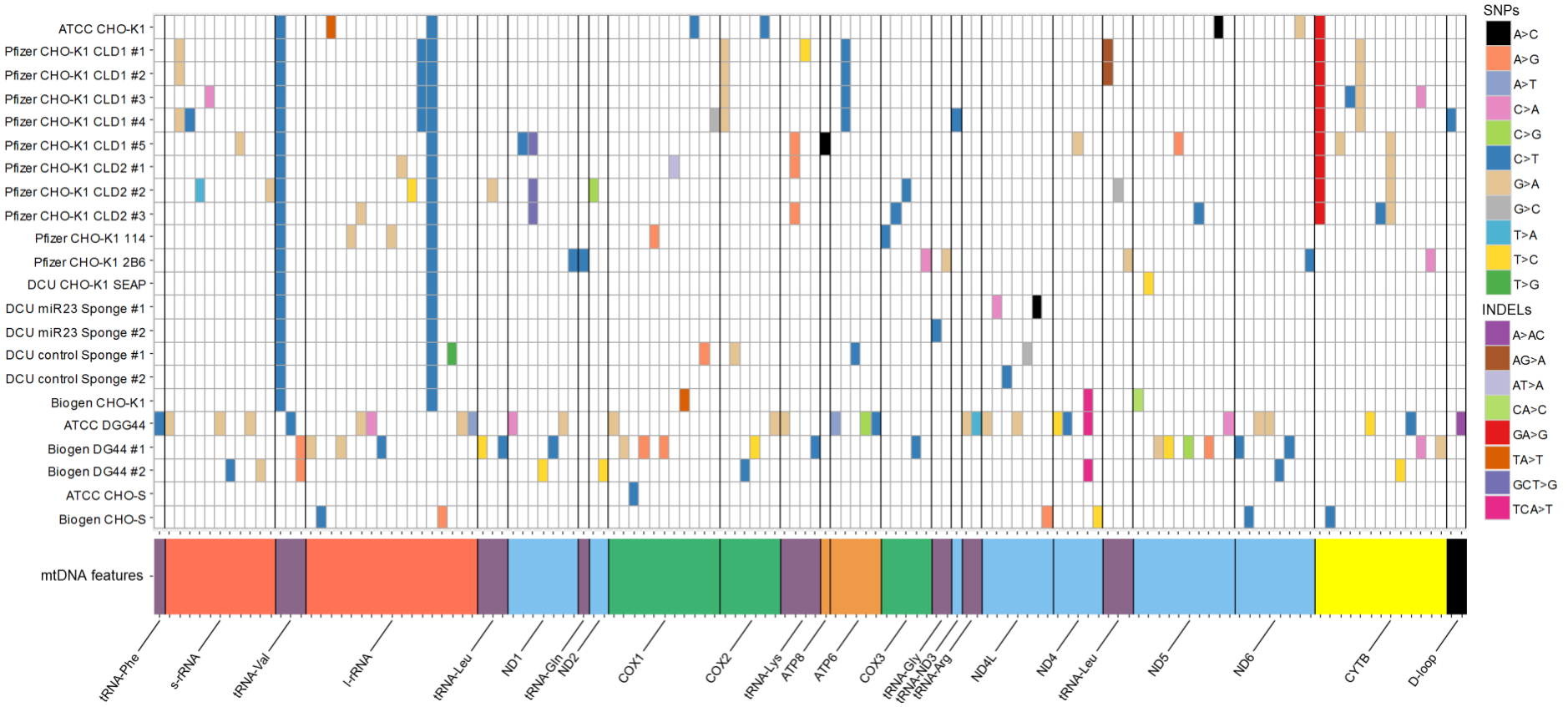

ATCC CHO-K1
Pfizer CHO-K1 CLD1 #1
Pfizer CHO-K1 CLD1 #2
Pfizer CHO-K1 CLD1 #3
Pfizer CHO-K1 CLD1 #4
Pfizer CHO-K1 CLD1 #5
Pfizer CHO-K1 CLD2 #1
Pfizer CHO-K1 CLD2 #2
Pfizer CHO-K1 CLD2 #3
Pfizer CHO-K1 114
Pfizer CHO-K1 2B6
DCU CHO-K1 SEAP
DCU miR23 Sponge #1
DCU miR23 Sponge #2
DCU control Sponge #1
DCU control Sponge #2
Biogen CHO-K1
ATCC DGG44
Biogen DG44 #1
Biogen DG44 #2
ATCC CHO-S
Biogen CHO-S
mtDNA features
tRNA-Phe
s-rRNA
tRNA-Val
l-rRNA
tRNA-Leu
ND1
tRNA-Gln
ND2
COX1
COX2
tRNA-Lys
ATP8
ATP6
COX3
tRNA-Gly
tRNA-ND3
tRNA-Arg
ND4L
ND4
tRNA-Leu
ND5
ND6
CYTB
D-loop
SNPs
A>C
A>G
A>T
C>A
C>G
C>T
G>A
G>C
T>A
T>C
T>G
INDELs
A>AC
AG>A
AT>A
CA>C
GA>G
TA>T
GCT>G
TCA>T

Figure 4

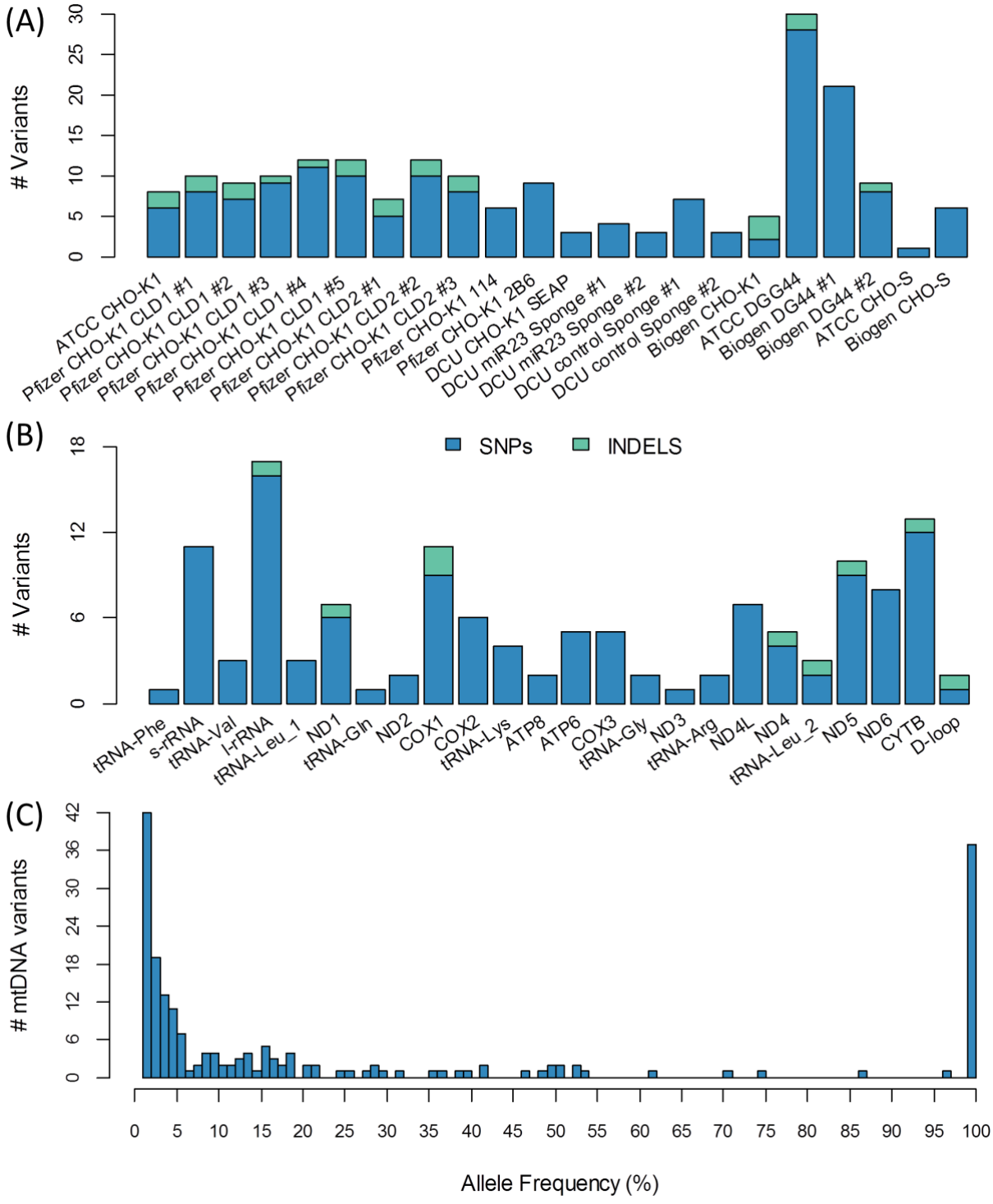
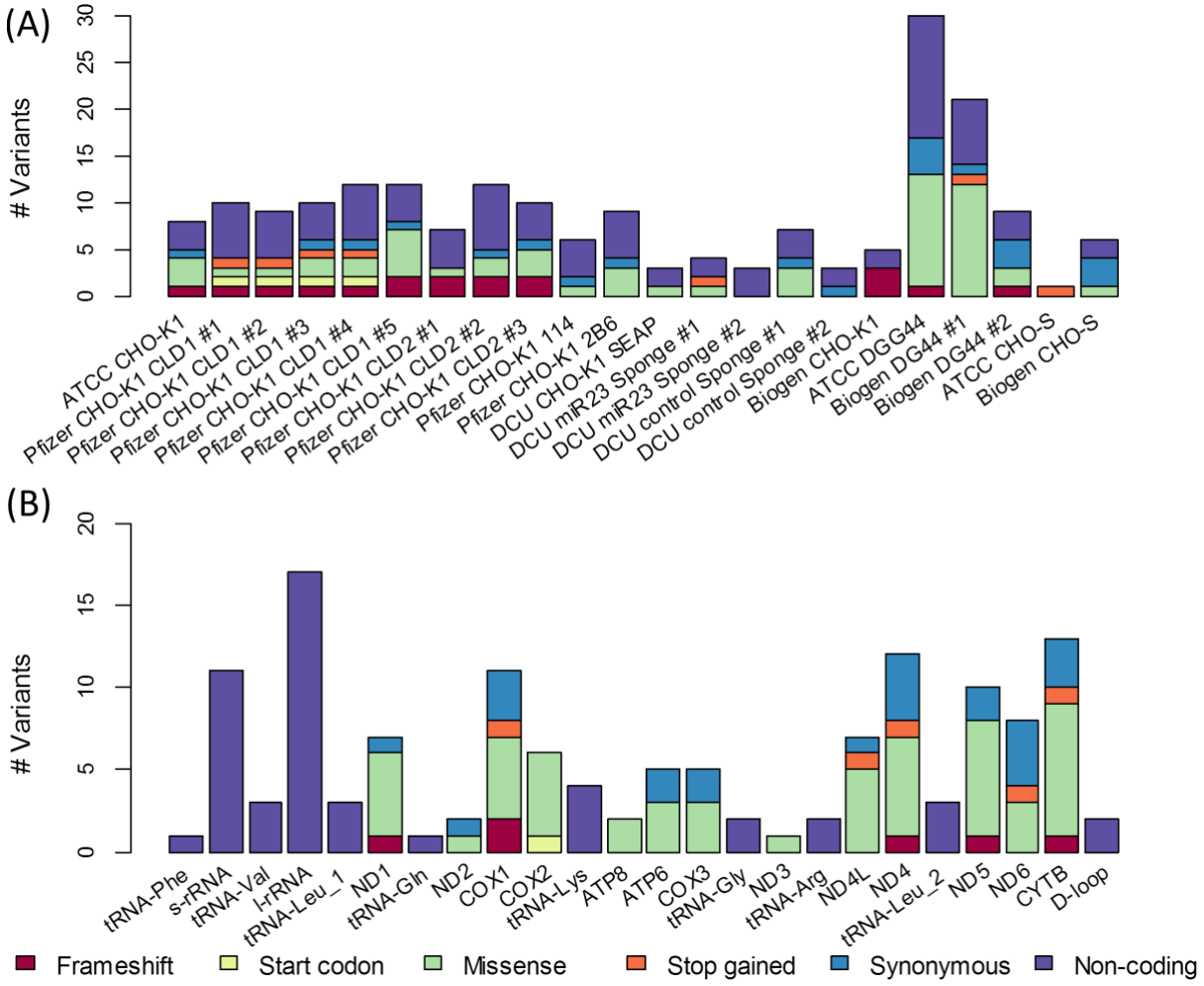
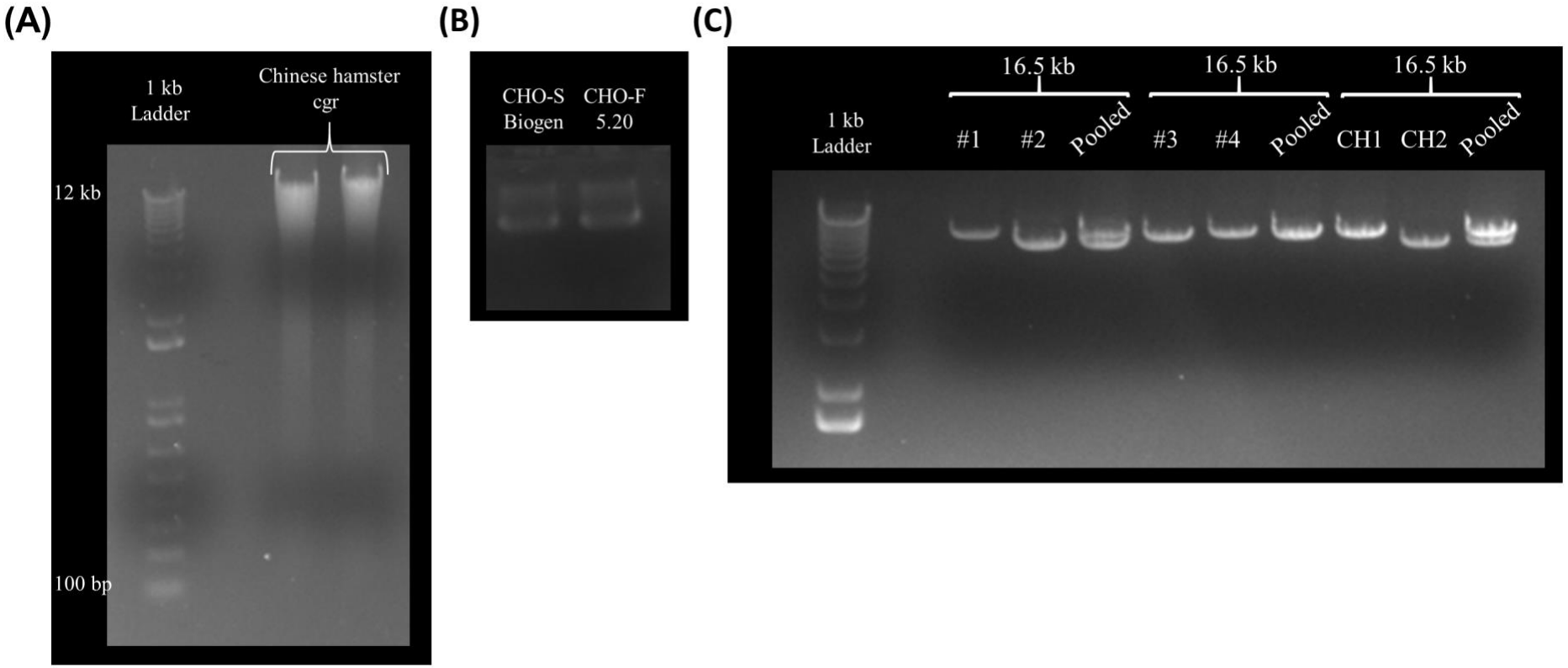

(A)
Variants
0
5
10
15
20
25
30
ATCC CHO-K1
Pfizer CHO-K1 CLD1 #1
Pfizer CHO-K1 CLD1 #2
Pfizer CHO-K1 CLD1 #3
Pfizer CHO-K1 CLD1 #4
Pfizer CHO-K1 CLD1 #5
Pfizer CHO-K1 CLD2 #1
Pfizer CHO-K1 CLD2 #2
Pfizer CHO-K1 CLD2 #3
Pfizer CHO-K1 114
DCU CHO-K1 2B6
DCU CHO-K1 SEAP
DCU miR23 Sponge #1
DCU miR23 Sponge #2
DCU control Sponge #1
DCU control Sponge #2
Biogen CHO-K1
ATCC DGG44
Biogen DG44 #1
Biogen DG44 #2
ATCC CHO-S
Biogen CHO-S
(B)
SNPs
INDELS
Variants
0
6
12
18
tRNA-Phe
s-rRNA
tRNA-Val
l-rRNA
tRNA-Leu_1
ND1
tRNA-Gln
ND2
COX1
COX2
tRNA-Lys
ATP8
ATP6
COX3
tRNA-Gly
ND3
tRNA-Arg
ND4L
ND4
tRNA-Leu_2
ND5
ND6
CYTB
D-loop
(C)
mtDNA variants
0
6
12
18
24
30
36
42
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
Allele Frequency (%)

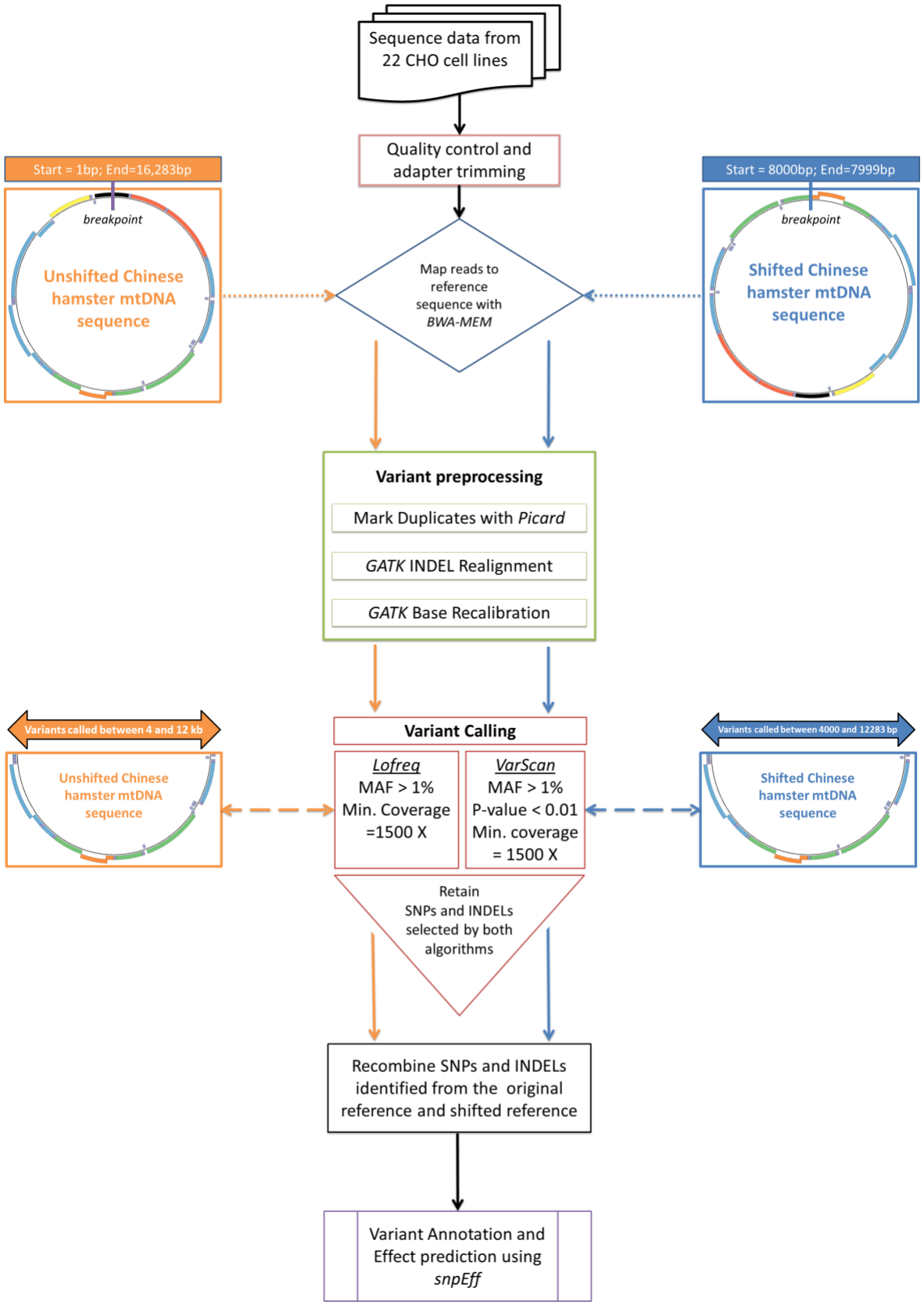
Figure 5


(A)
Variants
0
5
10
15
20
25
30
ATCC CHO-K1
Pfizer CHO-K1 CLD1 #1
Pfizer CHO-K1 CLD1 #2
Pfizer CHO-K1 CLD1 #3
Pfizer CHO-K1 CLD1 #4
Pfizer CHO-K1 CLD1 #5
Pfizer CHO-K1 CLD2 #1
Pfizer CHO-K1 CLD2 #2
Pfizer CHO-K1 CLD2 #3
Pfizer CHO-K1 114
DCU CHO-K1 2B6
DCU CHO-K1 SEAP
DCU miR23 Sponge #1
DCU miR23 Sponge #2
DCU control Sponge #1
DCU control Sponge #2
Biogen CHO-K1
ATCC DGG44
Biogen DG44 #1
Biogen DG44 #2
ATCC CHO-S
Biogen CHO-S
(B)
Variants
0
5
10
15
20
tRNA-Phe
s-rRNA
tRNA-Val
l-rRNA
tRNA-Leu_1
ND1
tRNA-Gln
ND2
COX1
COX2
tRNA-Lys
ATP8
ATP6
COX3
tRNA-Gly
ND3
tRNA-Arg
ND4L
ND4
tRNA-Leu_2
ND5
ND6
CYTB
D-loop
Frameshift
Start codon
Missense
Stop gained
Synonymous
Non-coding

Supplementary Figure 1



Supplementary Figure 2



Supplementary Figure 3