x`*Bioinformatics*, 2021, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: 10 September 2021
Manuscript Category

Sequence analysis

A framework for tracing mtDNA population dynamics through cell lineages

William Fulton[1*]

[1]School of Computing, Newcastle University

*To whom correspondence should be addressed

Availability and implementation: Framework and pipeline is freely available at https://github.com/Thomas-Fulton/Ludwig_2019.

1 Introduction

1.1 Mitochondrial polyploidy and clonal expansion

Mitochondria are vital for normal cellular function; they provide energy for cellular reactions through oxidative phosphorylation in the respiratory chain (Friedman and Nunnari, 2014). The mitochondrial genome (mtDNA) is highly polyploid: a typical cell has between 10^3 and 10^4 copies (Legros *et al.*, 2004; Satoh and Kuroiwa, 1991), since each mitochondrion may have two to ten copies of its genome (Robin and Wong, 1988). The highly polyploid mitochondrial genome contains many different mutations at once; new mutations are estimated to arise between 10 and 100 times more frequently in mitochondria than in nuclear mutations (Li *et al.*, 2014; Kaufman *et al.*, 2019; Floros *et al.*, 2018). This is because reactive oxygen free radicals from the respiratory chain can induce mutations primarily through damage to the Polymerase γ enzyme, which is then more likely to introduce transitions (Ziada *et al.*, 2020; Trifunovic *et al.*, 2004), but also through direct damage to the DNA, and the comparatively inferior mitochondrial DNA repair mechanisms (Michikawa *et al.*, 1999; Matkarimov and Saparbaev, 2020; Evans *et al.*, 2004; Zinovkina, 2018).

Mitochondrial mutations may be present at different proportions within a cell, clone or tissue, depending on the number of copies which contain the mutation. Homoplasmic alleles are fixed at an allele frequency (AF) of 1, or lost (AF = 0), where heteroplasmic alleles are not fixed and exist in a state of fluctuating allele frequency, driven by mtDNA replication and degradation (Lawless *et al.*, 2020; Bernardino Gomes *et al.*, 2021; Stewart and Chinnery, 2015). Biochemical defects and disease can result when the mutation load of a pathogenic mutation reaches a certain threshold within a cell; possible defects include reduced respiratory chain oxidative phosphorylation and ATP production (Yu-Wai-Man *et al.*, 2010; Durham *et al.*, 2007). Since ATP production is so essential to cellular activity, dysfunctional mitochondria are the source of a myriad of severe neuromuscular and neurodegenerative mitochondrial diseases (Bernardino Gomes *et al.*, 2021; Tuppen *et al.*, 2010), including Leigh Syndrome (Goldstein and Falk, 2003), MELAS (Hirano *et al.*, 1992), and Parkinson's disease (Dölle *et al.*, 2016). Dysfunctional energy metabolism can even promote tumor growth (Smith *et al.*, 2020; Ju *et al.*, 2014). MtDNA mutations are normally present and fluctuating at low levels in cells throughout a person's life (Morris *et al.*, 2017), but pathogenic mtDNA mutations may only show a disease-causing phenotype with time, once a certain mutation load threshold is reached (Li *et al.*, 2014; Greaves *et al.*, 2014; Elson *et al.*, 2001; Rossignol *et al.*, 2003). Consequently age-related diseases in particular have been associated with the accumulation of mutations in stem cell populations over time (Yu-Wai-Man *et al.*, 2010; Durham *et al.*, 2007; Michikawa *et al.*, 1999), leading to the reduced cellular function (Nooteboom *et al.*, 2010) and tissue regeneration (Sharpless and DePinho, 2007) that occurs in old age. Low level pathogenic mutations can have no effect on a cell, but new or inherited pathogenic mutations can increase to detrimental levels, and take over the cell through clonal expansion.

Clonal expansion is a dynamic process through which the proportion of a mtDNA mutation increases within mtDNA populations, both between successive generations of cells and within individual cells (Payne *et al.*, 2013; Lawless *et al.*, 2020). The mutation load, or allele frequency, of a heteroplasmic mutation may increase or decrease with time, and potentially be either lost or fixed in the mtDNA. Under neutral conditions this is driven by random processes which affect the rate of genetic drift, principly relaxed replication and vegetive segregation (Zakirova *et al.*, 2021). Different rates of relaxed replication and degradation may increase the relative proportion of certain mtDNA copies (Elson *et al.*, 2001; Zakirova *et al.*, 2021; Bogenhagen and Clayton, 1977; Chinnery and Samuels, 1999). Vegetive segregation, the random and uneven distribution of mtDNA between daughter cells during cell division, acts as a bottleneck, both increasing the impact of genetic drift by decreasing the effective mtDNA population size in each daughter, and directly changing the initial mutation load of the daughter cells because a random subset of mtDNA copies are inherited (Birky and William Birky, 2001; Stamp *et al.*, 2018). Mathematical modelling work has demonstrated that genetic drift alone can be sufficient for clonal expansion of point mutations to pathogenic levels, given enough time (Coller *et al.*, 2001; Elson *et al.*, 2001).

Mechanisms which confer a selective advantage or disadvantage on mtDNA populations have also been proposed to promote or inhibit clonal expansion of point mutations. The “survival of the sickest" theory suggests that mutant mitochondria accumulate if they produce fewer reactive oxygen species, because they are less likely to undergo mitophagy (Yoneda *et al.*, 1992), and the “negative feedback loop” theory states a mtDNA mutation which leads to a reduced levels of a transcription or replication inhibitor, will undergo increased mtDNA replication and accumulate (Kowald and Kirkwood, 2018, 2014). Deletions are not called in this variant calling pipeline, but they are also a major cause of mitochondrial diseases (Tuppen *et al.*, 2010), and could be subject to alternate selection pressures like the survival of the smallest theory, which postulates that a smaller copy of mtDNA replicates faster and therefore accumulates (Wallace, 1989).

1.2 Modelling mtDNA population dynamics

The rate at which mtDNA mutates and replicates, along with neutral and selective mechanisms affecting clonal expansion, contribute to the progression of mitochondrial diseases. Improved understanding and study of these factors and their relative impacts on clonal expansion is needed if treatments for mitochondrial diseases are to one day be developed (Lawless *et al.*, 2020). Unfortunately, direct longitudinal observations of mtDNA population dynamics through sequencing is not possible; because cells are destroyed during library preparation, mtDNA mutation levels cannot be observed across points in time, either within a single cell or between directly related generations of cells. Simulating clonal expansion from mathematical models of mtDNA population dynamics can allow the effect of different processes on mtDNA population dynamics (mutation rate, selection pressures etc.) to be hypothesised and their consequences predicted (Lawless *et al.*, 2020). The mathematical model, and any subsequent hypotheses, can then be tested through comparison with experimental data (Lawless *et al.*, 2020; Henderson *et al.*, 2009). Since influential factors like mtDNA copy number can vary wildly, eg. between cell types, mathematical models must incorporate specific assumptions and parameters to be a comparable and accurate model of mtDNA population dynamics. Model specific parameter inference can allow observation and understanding of underlying mechanisms affecting clonal expansion and mitochondrial diseases that are otherwise difficult to observe, like mutation rates, replication rates and the impact of selection pressures.

In order to create models of clonal expansion, inherited mutations undergoing clonal expansion must be identified. Allele frequency autocorrelation between generations of sub-clones is expected to be indicative of clonal expansion; because a random proportion of the parents' mutation loads are inherited, the allele frequency of mutations in daughter cells depend on the initial mutational load in the mother cell. Vegetive segregation of the mitochondria, and relaxed replication among the populations of mitochondrial genomes however give rise to allele frequency variation between generations. Therefore, a stochastic, autocorrelated, increase or decrease in allele frequency is seen from one generation to the next.

Ludwig *et al.* (2019), aimed to show that lineage tracing was possible using mtDNA reads from bulk Assay for Transposase-Accessible Chromatin sequencing data (ATAC-seq), and RNA sequencing data. As proof of principle they generated an experimental lineage tree of 66 clones (Fig. 1) by isolating and expanding a single cell from the bulk parent clone into a sub-clone, followed by single cell isolation and expansion from each sub-clone, every three weeks, for up to eight times in a single lineage. The experimental setup and TF1 clone lineage structure allows indirect observation of inherited mutations.

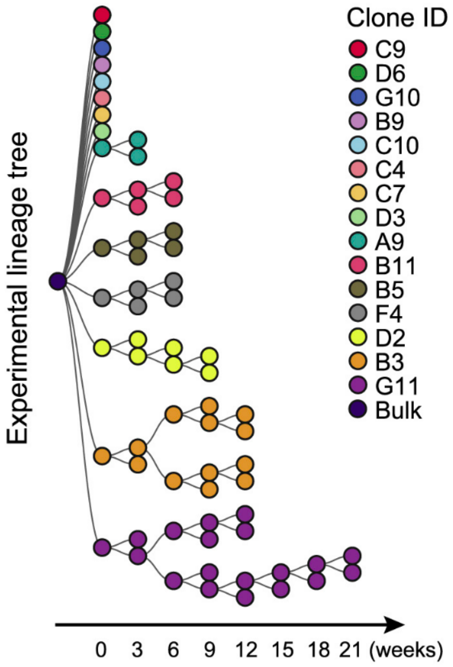

Fig. 1: Experimental setup of TF1 clones generation by Ludwig et al. (2019). Multiple observations of inherited mutations among related clones are needed to model clonal expansion, and the experimental setup of the TF1 clones lineages provided a suitable dataset for this analysis. (Adapted from: Figure 1 Ludwig *et al.*, 2019)

This research project aimed to develop a framework to conduct a reproducible, updated, start-to-finish variant calling pipeline, specialised for identifying heteroplasmic mutations undergoing clonal expansion, and built for use on the Newcastle University Rocket high performance computing service. The pipeline is applied here to reproduce the analysis previously carried out by Ludwig *et al.* (2019), and provide the allele frequencies of inherited mutations in TF1 clones that may display clonal expansion, for the improvement of mathematical models of mtDNA population dynamics. Moreover this project aims to provide the allele frequency data of mutations which putatively show clonal expansion, to explore the idea that clonal expansion can be identified through allele frequency autocorrelation between generations of sub-clones. The pipeline has been implemented in an adaptable framework, designed to be easily modified and applied to other suitable datasets. This specialised pipeline has been initially developed on, and compared to, bulk ATAC-sequencing data and variant calling of TF1 clones previously carried out by Ludwig *et al.* (2019), and leverages the almost direct nature of clones' relationships to improve detection of low-level mutations.

2 Methods

This pipeline was used to reanalyse TF1 clone bulk ATAC-sequencing data, previously carried out by Ludwig *et al.* (2019). The results between the variant calling pipelines were then compared.

2.1 Analysis of TF1 clones

2.1.1 Data retrieval

Raw reads generated from bulk ATAC-seq sequencing runs were downloaded from the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/, GEO accession: GSE115218), checked, and converted to .fastq format using the SRA-toolkit v2.11.0 (SRA Toolkit Development Team, http://ncbi.github.io/sra-tools/) and converted to .fastq format. There were 69 runs of bulk ATAC-sequencing, which included one for each of the 66 expanded TF1 clones, a technical replicate run of the bulk parent population, and two runs containing a mixture of clones mtDNA (Ludwig *et al.*, 2019).

2.1.2 Pre-alignment quality control

The raw number of reads, average base quality, GC content, and estimated number of duplicate sequences were calculated for the forward and reverse reads using the software FastQC v0.11.8 (Andrews, 2010), and visualised with MultiQC v1.7 (Ewels *et al.*, 2016), and R v3.10 (R Core Team, 2020). Data manipulation and plotting throughout was conducted in Rstudio (Rstudio team, 2020 http://www.rstudio.com/), with the packages ggplot2 v3.3.5 (Wickham, 2009) and tidyr v1.1.3 (Wickham, 2021).

2.1.3 Read alignment

The hg38 genome (Schneider *et al.*, 2017) was indexed, and the paired reads were aligned to the whole genome using Bowtie2 v2.3.4.2 (Langmead and Salzberg, 2012). Soft-clipping of reads accounted for adaptor sequences, (--local --sensitive). Reads were aligned to the entire genome so that nuclear mitochondrial sequences (NUMTs) did not misalign to the mitochondrial genome, as suggested by Santibanez-Koref *et al.* (2019) and Marquis *et al.* (2017). Samtools v1.12 (Li *et al.*, 2009) was used to sort and index the .bam files, before filtering the reads for quality. The mapping qualities, base qualities, read depths per base, and genome coverage was calculated with and without filtering for reads with a base quality >20, and mapping quality >18. Post alignment adjustments like base quality calibration and realignment around indels were excluded, as both can introduce strand biases (Guo *et al.*, 2012), and are not necessary for point mutation calling (Weissensteiner *et al.*, 2016).

2.1.4 Variant calling

The variant caller Mutserve (Weissensteiner *et al.*, 2016) was used to initially call and filter mutations. The following filters were applied when variant calling to maximise the number of true positive calls (false positives were removed post-variant calling): read alignment quality >30, base quality >20, mapping quality >18, and minimum heteroplasmy 0.001. Mutserve applies further filters: it also excludes sites covered by less than 10 reads per strand, or 3 reads per allele per strand. Finally, it applies a maximum likelihood model (Ye *et al.*, 2014; Weissensteiner *et al.*, 2016), to account for sequencing errors. The allele frequencies on the forward reads and reverse reads are calculated, and the weighted mean of each strand is taken as the mutation's allele frequency.

The revised Cambridge Reference Sequence (rCRS) mtDNA genome (Andrews *et al.*, 1999) consists of 16,569 base positions, including a base of "N" at position 3,107 representing a historical sequencing error (Ju *et al.*, 2014). Mutations at that position were removed. Mutations with allele frequency >0.990 represented homoplasmic mutations fixed in the parent bulk population, in comparison to the rCRS reference genome, and so were also excluded. Mutations were further filtered through lineage validation: if a mutation was present in more than one clone in a lineage, and had an allele frequency >0.01 in at least one clone in a lineage, the mutations were retained.

2.1.5 Identifying clonal expansion of mutations

The allele frequencies of eight paths through six lineages (each with >3 generations) were visually explored to find autocorrelation between generations. Firstly, all mutations, in all generations of clones in a defined lineage path, were plotted (Fig. S1) and visually inspected to identify inherited mutations potentially showing an autocorrelated series of increases and decreases in allele frequency across the generations, which is indicative of clonal expansion. Finally, selected mutations were plotted individually and visually examined for patterns of autocorrelation. The position was noted if a mutation was present in more than three clones.

2.2 Comparison with previous analysis of TF1 clones

Ludwig *et al.* (2019) previously called 44 mutations in the TF1 clone dataset, and used them to reconstruct the clone lineages. In order to compare our variants, the positions of the 44 mutations were converted from the index in hg19 mitochondrial genome to the index of rCRS mitochondrial genome, used in the hg38 genome assembly (positions >= 315 and < 3107 were decreased by 2; positions >= 3107 and <16193 were decreased by 1; positions >= 16193 were decreased by 2). The genomic positions of Ludwig *et al.'s* (2019) mutations are referred to here by their rCRS indexed position.

A heatmap clustering clones using our calculated allele frequencies for the 44 mutations called in Ludwig *et al.* (2019) was created and plotted with the ComplexHeatmap R package (Gu *et al.*, 2016). Pearson's correlation coefficients were calculated between ours and Ludwig's allele frequencies, for each of the 40 mutations. Allele frequencies were plotted against those of Ludwig *et al.* (2019) to observe the relationship between our variant calls. Bulk replicate allele frequencies were plotted against each other.

3 Results

3.1 Variant calling pipeline for ATAC-seq data

3.1.1 Quality of raw sequencing reads

There was a median of 10.1M reads in the forward and reverse raw .fastq files (min=4.2M, max=26.5M, Fig. 2A), and the median percentage of duplicated reads was 63.0% (min=33.3%, max=84.2% Fig. 2B). Read lengths were either 2x35bp or 2x75bp for each sequencing run. The average base quality was >30.0 for all positions in all reads (Fig. S2), and the average percentage GC content was 46.2% +/-0.62.

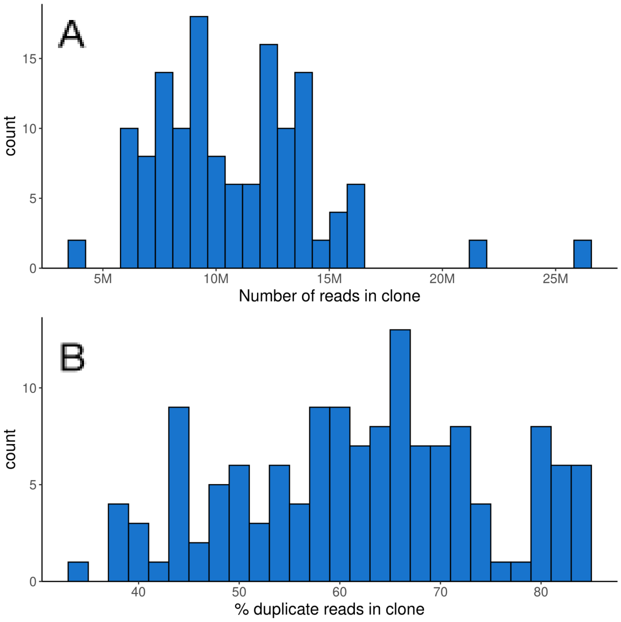

Fig. 2. Large range in the raw number of reads (A) and percentage of duplicated reads (B) across sequencing runs of 69 clones. The high duplication levels may have affected variant calling by introducing high levels of strand bias later in the pipeline.

3.1.2 Read Alignment

On average, 98.5% of each clone's reads were aligned to the mitochondrial genome (min=97.4%). Before filtering for mapping and base quality, the average base quality and mapping quality (Fig. S3)of the aligned reads were 33.4 and 26.1 respectively. The clone's mitochondrial genomes were covered at 34700X on average, with a mean read depth of 38900. Post-quality filtering (base quality >20, and mapping quality >18), the average base quality and mapping quality of the aligned reads were 34.9 and 26.8 respectively. The clone's mitochondrial genomes were covered at 28700X on average, post-quality filtering, with a mean read depth of 22600 (Fig. 3A).

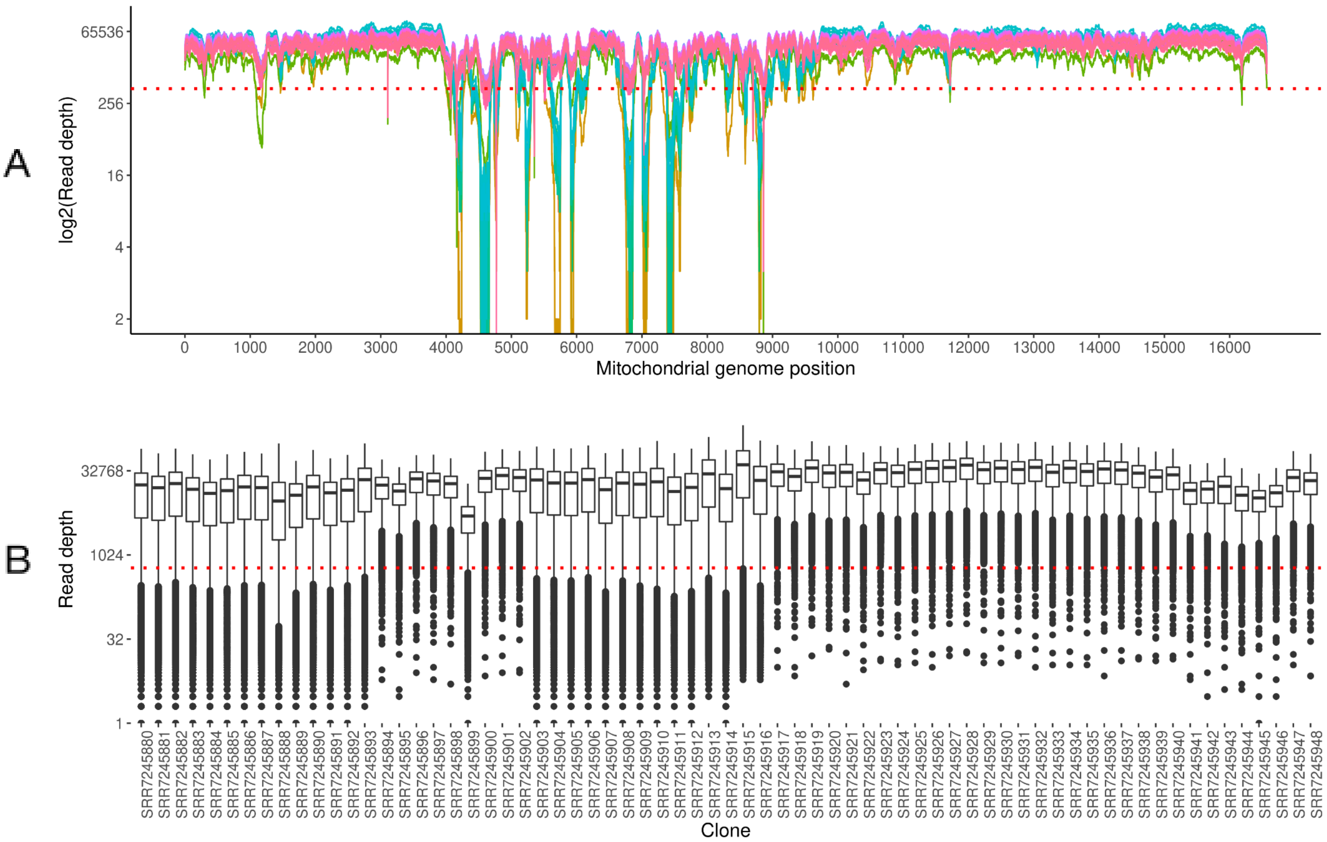

Fig. 3. Post-quality filtering mtDNA coverage and read depth for the 69 sequencing runs. A: Very high average coverage across the mitochondrial genomes despite short regions of very low coverage. Coloured by sequencing run. B: Consistent coverage among the 69 sequencing runs. Red dotted line indicates a read depth of 600: mutations with an allele frequency of <0.01 will not be called if covered by fewer than 600 reads, as the variant caller excludes alleles covered by fewer than 3 reads on each strand (Weissensteiner *et al.*, 2016).

3.1.3 Variant calling and visual identification of autocorrelation

Without filtering for strand bias or lineage validation, 2,339 heteroplasmic point mutations were called at 544 positions across the 69 clonal mitochondrial genomes. A total of 1,463 mutations at 216 positions were retained after lineage validation, and 1042 of these were flagged with strand bias. The average Ti/Tv ratio was 0.63 (564 transitions and 899 transversions).

Of the validated mutations in the 8 defined lineages: 104 different genomic positions showed autocorrelated changes in allele frequency between generations of sub-clones in a lineage, (Table 1, Fig. 6A and B). Autocorrelation was seen both at very low and very high levels of mutation load (Fig. 4A and 3C), with some large jumps between generations (Fig. 4E) including mutations present in only one generation in the lineage path (Fig. 4D).

Some mutations were seen to have been lost in a generation, before reappearing in the next generation consistent with the otherwise autocorrelated pattern of allele frequency changes. Overall increase and decrease of mutation loads were seen at different sites.

Table 1. Lineage validated mutations which have been visually identified as showing clonal expansion-like patterns of allele change across generations in a lineage (Fig. 4). Previously identified mutations (Ludwig *et al.*, 2019) are highlighted in bold. Lineage path name describes the route through a lineage in Fig. 1.

Lineage path	No. genomic positions (Previously identified, new)	Mutations demonstrating autocorrelation of allele frequencies
B11 longest lower	**2,** 8	**8002,12789,** 301,310,2954,3078,7789,12475,14424,16183
B5 longest upper	**2,** 2	**8002,13708,** 627,3577
F4 longest upper	**5,** 3	**1410,8206,11403,13288,15640,** 297,310,16183
D2 longest lower	**2,** 13	**4037,8002,** 297,301,567,3078,3565,3885,3927,8701,12475,15291,15914,16066,16161
B3 longest lower	**3,** 9	**8002,12789,** 301,302,310,3078,7789,3078,3565,3577,7789
B3 longest upper	**3,** 14	**8002,12789,11711,** 3078,3565,3577,3584,3885,3927,7789,12475,12750,15291,15914,16066, 16175,16183
G11 longest upper	**7,** 23	**822,1493,3173,5007,5862,6962,1579,** 297,302,310,822,1493,3078,3173,3577,3885,3886,3927, 5007,5351,5862,6962,7853,8138, 8701, 12475,12651,15797,15794,16172

G11 shorter upper	**7,** 2	**822,1493,2816,3173,4769,5862,6962,** 4769,5862

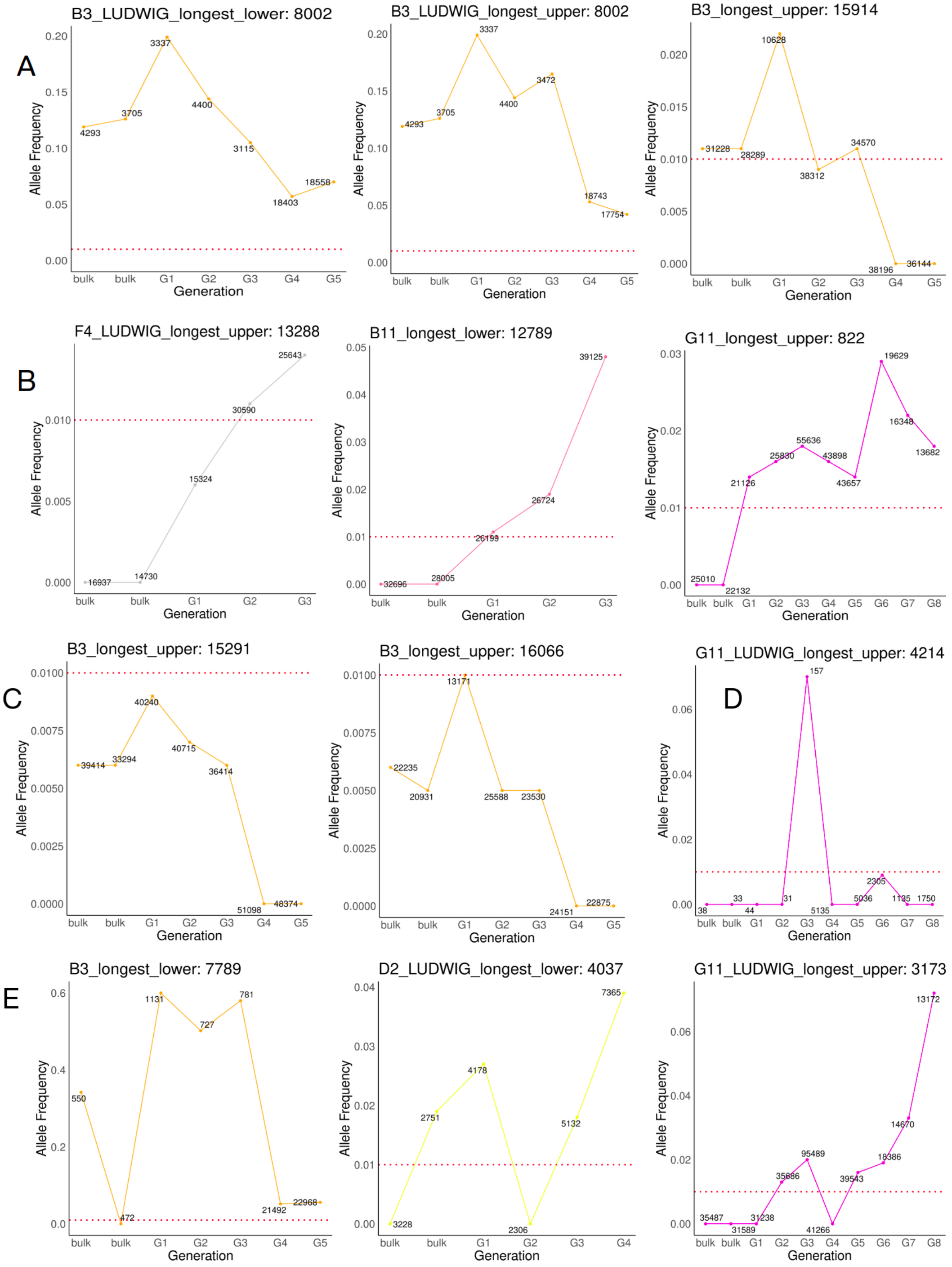

Fig. 4. Changing levels of allele frequencies for validated mutations in a lineage path. Low level mutations were considered valid if an allele frequency >0.01 was seen at the same site in at least one clone of the lineage (dotted red line). Various different patterns are seen, including visual autocorrelation indicative of clonal expansion. A: Example of mutations inherited from the bulk population, and the autocorrelation-like pattern indicative of clonal expansion. B: examples of a new mutation's allele frequency increasing across generations within a lineage. C: Low level mutations also demonstrated patterns of clonal expansion in their allele frequencies. D) and E): Large changes in allele frequencies between generations were occasionally observed. Some otherwise autocorrelated mutations were not present in all generations. Coloured according to lineages in Fig. 5. Coverage is given for each generation.

3.1.4 Comparison to previous analysis of TF1 clones

The quality and number of the reads before mapping was almost identical to that of Ludwig *et al.*'s (2019, data not shown). Our data showed a higher fold coverage than Ludwig *et al.*'s (2019), pre-quality filtering (34,900X and 33,800X respectively).

This framework's pipeline called 40 out of the 44 mutations which were previously identified in the TF1 clones dataset (Ludwig *et al.*, 2019). Using the allele frequencies of those 40 mutations from our framework, most clones clustered with others in their lineage (Fig. 5), although the clones in some lineages were fragmented: lineage D2 and B3 in particular, (yellow and orange respectively, Fig. 5). The 40 mutations were all present in the same clones as detected by Ludwig *et al.* (2019), but were called with generally lower allele frequencies (Fig. 6B). There is strong linear positive correlation between the allele frequencies called in this pipeline and the frequencies seen in Ludwig *et al.*'s data (2019) at each genomic position, (ie. the same mutation present in different clones); the mean Pearson's coefficients = 0.89. (Table 2, Fig. 6B). There were low correlations among mutations within a clone (data not shown).

The bulk technical sequencing replicates were seen to have some mutations which were only called in one of either bulk technical replicates (SRR7245880 and SRR7245881, Fig. 6B). Of the mutations which were called in both replicates, the allele frequencies were very similar. All but one of Ludwig's variants were only called in one of the bulk technical replicates (red points, Fig. 6B).

Table 2. Strong allele frequency correlation between allele frequencies of 44 previously called mutations with the allele frequencies called in this framework's pipeline. Four mutations were not called in our pipeline (bracketed). Significant p-values (<0.05) are highlighted in bold.

Position	Pearson's Correlation	
	Correlation Coefficient	p-value
182	0.973	**0.000**
309	0.189	0.119
822	0.986	**0.000**
847	0.764	**0.000**
1410	0.911	**0.000**
1493	0.992	**0.000**
(1795)	-	-
1970	0.449	**0.000**
2108	0.936	**0.000**
2816	0.985	**0.000**
3173	0.955	**0.000**
3910	0.923	**0.000**
4037	0.859	**0.000**
(4413)	-	-
4214	0.515	**0.000**
4446	0.317	**0.008**
4512	-	-
5007	0.892	**0.000**
5563	0.789	**0.000**
5862	0.817	**0.000**
(6075)	-	-
6962	0.982	**0.000**
(7074)	-	-
7789	0.812	**0.000**
8002	0.957	**0.000**
8206	0.985	**0.000**
8921	0.626	**0.000**
10371	0.992	**0.000**
11184	0.996	**0.000**
11403	0.982	**0.000**
11711	0.964	**0.000**
12061	0.956	**0.000**
12253	0.898	**0.000**
12789	0.911	**0.000**
12838	0.937	**0.000**
13288	0.917	**0.000**
13412	0.994	**0.000**
13708	0.984	**0.000**
14436	0.975	**0.000**
15088	0.992	**0.000**
15488	0.934	**0.000**
15640	0.952	**0.000**
15797	0.980	**0.000**
16250	0.910	**0.000**

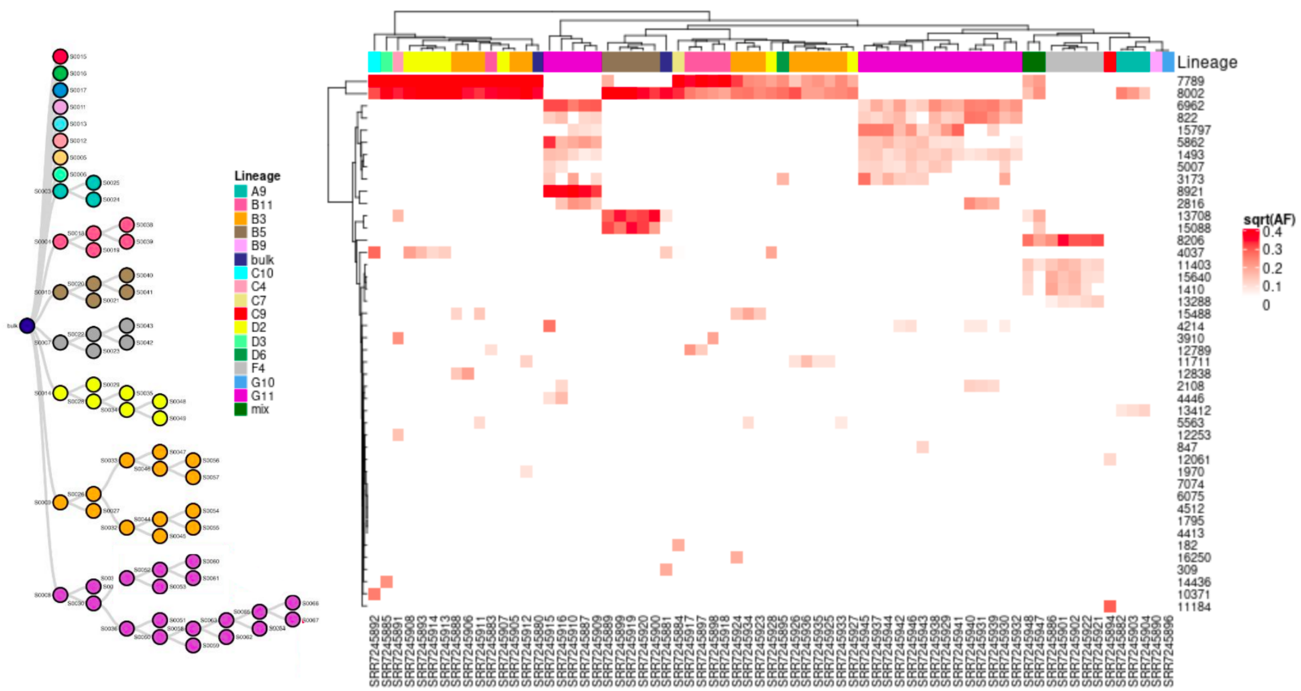

Fig. 5. This framework's variant calling pipeline called 40 out of 44 previously identified mutations (Ludwig *et al*. 2019). A: Lineage tree of clones (Adapted from: Fig. S1, Ludwig *et al.* 2019). B: Clustering of clones using the square root of a mutation's allele frequency. Genomic position of a mutation is on the y axis, colour indicates the lineage of a clone.

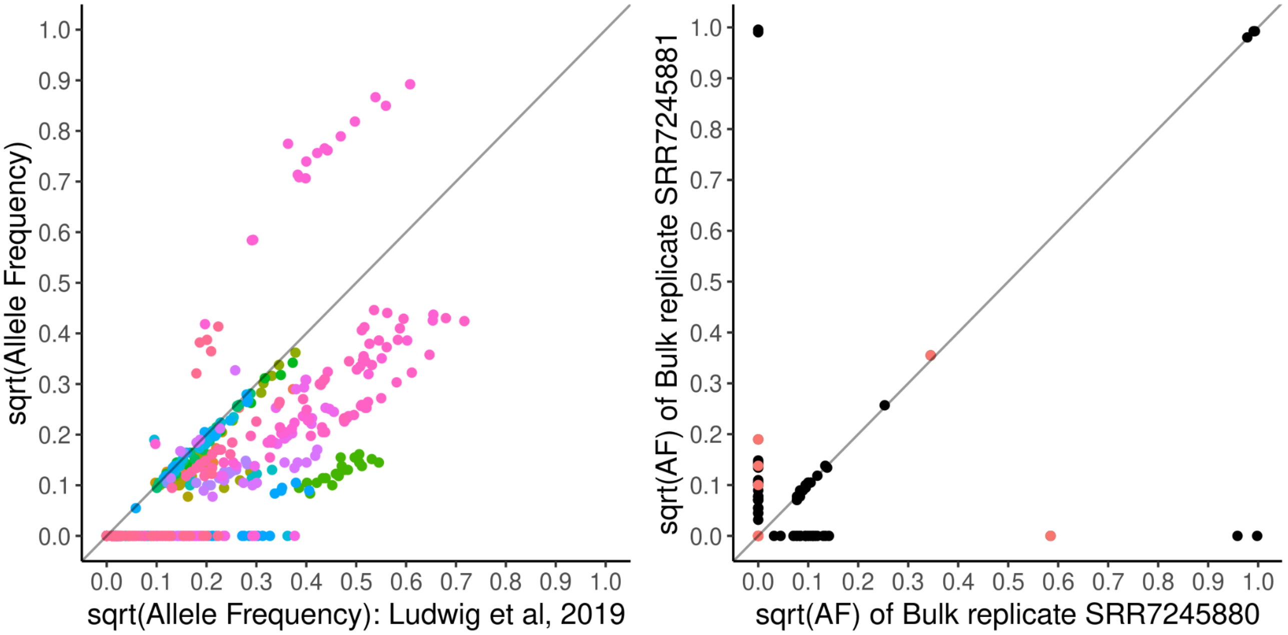

Fig. 6: Variation in allele frequencies (AF) detected: both between mutations detected by the pipeline of Ludwig *et al.* (2019)**, and those detected by this framework's pipeline (A), and between all heteroplasmic (AF<0.990) mutations in sequencing replicates of the same clone in our pipeline (B)**: All heteroplasmic mutations detected in bulk clones. Each point represents the square root of the allele frequency (to give weight to low-level mutations) of a point mutation. Points in red indicate one of the 44 high confidence mutations called by Ludwig *et al.* (2019). (**B**): root allele frequency of 44 high confidence mutations detected by Ludwig *et al.* (2019), as detected in their pipeline against our pipeline. Colour indicates mutation position. The cluster of points along the x axis indicate point mutations which were called by Ludwig *et al.* (2019), but not by this framework.

4 Discussion

This framework successfully expanded upon the analysis of Ludwig *et al.* (2019), by implementing a modified variant calling pipeline tailored towards detecting allele frequency signature of clonal expansion, on the ATAC-sequencing data of TF1 clone lineages, and taking advantage of the inherited nature of clones' mutations to validate low-level calls while compensating for potential errors.

4.1 Variant calling and mitigating biases

Specific properties are expected in mutations which show clonal expansion: they should be heteroplasmic, inherited, and demonstrate allele frequency autocorrelation between generations. These properties also provide a unique opportunity to validate variant calls: calls ambiguous in an individual clone can be called with more confidence if the mutations are inherited, a property which is already necessary for observing clonal expansion. 'Lineage validation' improved variant calling because we are only interested in inherited mutations, so some of the limitations when calling mutations in single clones with high confidence can be ignored. Consequently this pipeline initially calls all low-level mutations, before simultaneously removing potential false positive calls in individual clones and identifying mutations showing clonal expansion through lineage validation. Baslan and Hicks (2014) recommend a similar approach in single cell sequencing; only mutations present in at least two cells should be called. Lineage validation is able to increase confidence in low-level mutations because false positive mutations caused by the majority of biases and errors are random, and are very unlikely to be seen at the same genomic position in different clones, or are otherwise distinguishable from true, inherited mutations. Different bias and errors are compensated for throughout the variant calling pipeline, as described below.

Reads with low mapping quality can affect variant calls. Bowtie2 (Langmead and Salzberg, 2012) calculates mapping quality as a representation of the likelihood that a read is aligned to the correct position in the genome, which also incorporates the mapping quality of a second most likely alignment. This pipeline used a mapping quality filter threshold of 18, which is slightly lower than the Mutserve default (20, Weissensteiner *et al.*, 2016). This was chosen to increase read coverage in 28 clones which had mean read mapping qualities of ~19.5 (Fig. S3), at the cost of a relatively small increase in potentially misaligned reads which could introduce false mutation calls. This lower threshold increased the number of reads kept for variant calling and increased read coverage across the genome. In theory, more retained reads for low-level alleles, means more accurate allele frequencies can be calculated and a higher number of true positive calls are made, and subsequent lineage validation of mutations can remove the corresponding increase in false positive calls. The increased sensitivity from retaining reads is preferable in this pipeline, despite including potentially false mutations from misaligned reads.

In addition to false positives from misaligned reads, sequencing errors may resemble very low level mutations, especially at low coverage (Ekblom *et al.*, 2014; Meacham *et al.*, 2011). However, both random (Ebbert *et al.*, 2016) and context-specific sequencing errors (Allhoff *et al.*, 2013; Nakamura *et al.*, 2011) only occur on one strand. It is highly unlikely to occur at the same position on both strands (Allhoff *et al.*, 2013; Meacham *et al.*, 2011; Chen *et al.*, 2017). This pipeline accounts for potential false positive calls from misaligned reads and sequencing errors by only calling mutations if they were present on both strands: specifically if >3 reads supported an allele for both strands. Mutserve's default base quality filter was used (20, Weissensteiner *et al.*, 2016), in order to include more reads, where the more conservative threshold of 23.8 was used in Ludwig *et al.*'s (2019) variant calling pipeline; since they did not incorporate per-strand information, a higher base quality threshold of 23.8 was more applicable. Typically, noise from sequencing or PCR error is adjusted for by using a minimum cutoff of 0.01 (1%) allele frequency in high coverage datasets when calling mitochondrial mutations (Koboldt *et al.*, 2012; Rebolledo-Jaramillo *et al.*, 2014; Zhidkov *et al.*, 2011; Ding *et al.*, 2015; González *et al.*, 2020). Sequence specific sequencing errors could be present at the same position in multiple samples, but will be completely absent in one strand of a read (Allhoff *et al.*, 2013). Both types are accounted for by Mutserve's filters. However, mistakes introduced by polymerase infidelity during PCR amplification during library preparation on one strand, result in a mutation on both strands of a read, due to the semiconservative replication (Potapov and Ong, 2017; Shaw, 2002). Moreover these can sometimes be sequence specific (Potapov and Ong, 2017), so could appear as inherited mutations where they occur in the same position in different clones. While this is possible, PCR errors are unlikely to have occurred in multiple, related clones in a particular lineage. Therefore, calls were initially made using the reduced minimum heteroplasmy threshold of 0.001, but were only validated if 1) the mutation was present in more than one clone in its lineage, and 2) had an allele frequency greater than the standard sequencing error threshold, 0.01, in at least one clone in its lineage. In this way, confidence in a mutation with an allele frequency above the standard minimum threshold could be extended to low-level mutations (which would otherwise be indistinguishable from PCR error) at the same genomic position of related clones, in the same lineage.

PCR amplification bias skews the copy number of a strand during the amplification stage of library preparation (Aird *et al.*, 2011; Sims *et al.*, 2014), and sequencing bias preferentially affects one strand over the other (Ross *et al.*, 2013). Both can exaggerate strand bias levels: when the genotype or allele frequency from sequencing differs between the forward and reverse reads (Guo *et al.*, 2012). Guo *et al.* (2012) concluded that strand bias was not present in the same locations in different samples.

Because PCR errors, sequencing errors, PCR bias and sequencing bias all introduce strand bias, there are many different strand bias filters used in modern mitochondrial variant calling pipelines. EMBLEM (Xu *et al.*, 2019), a mitochondrial lineage tracing pipeline designed specifically for bulk and single cell ATAC sequencing data, calls alleles with >2 reads on each strand, and no less than 30% of the total reads supporting an allele can be on either strand. Other measures of strand bias include: the Symmetric Odds Ratio test, used by GATK (Van der Auwera and O'Connor, 2020); Varscan2 excludes mutation if >90% of an allele's reads are from one strand (Koboldt *et al.*, 2012); and samtools implements fishers' exact test to determine the probability the strand bias is random (Li *et al.*, 2009). Mutserve was chosen because is an easy to use, specialised variant caller, which was specifically built for mitochondrial genomes, unlike other commonly used software, eg. Mutect2 (Benjamin *et al.*, 2019). Additionally it can be installed locally, so large bam files do not have to be uploaded to a web server-based mitochondrial variant callers (Santorsola *et al.*, 2016; Zhidkov *et al.*, 2011; Vellarikkal *et al.*, 2015). Mutserve also extensively annotates mutations: strand bias is flagged rather than excluded, forward and reverse read coverages are calculated, gene and region of the mutation is given, amongst more relevant information (Weissensteiner *et al.*, 2016).

While mutations showing extreme strand bias (top 10% of strand bias scores) show a high potential high false-positive rate and should be excluded from variant calls in diploid genomes (Guo *et al.*, 2012; Koboldt *et al.*, 2012), calling low-level mutations of highly polyploid data, calls are unlikely to be missed, but allele frequency estimations will have been skewed. Although lineage validation of a mutation enables the presence of a mutation to be called with high confidence, which might otherwise be discarded, many of our calculated allele frequencies will have been skewed by the high proportion of strand bias in our calls; 1,042 mutations out of 1,463 were flagged with some degree of strand bias. While duplicate removal has been shown to have almost insignificant impact on variant calling (Ebbert *et al.*, 2016), at low sequencing depths duplicates can result in higher incidence of false negative and false positive calls (Li *et al.*, 2010). Since Guo *et al.* (2012) concluded that alignment artifacts were unlikely to be a cause of strand bias, and strand bias mostly occurred randomly, despite sequence-specific bias like sequencing bias, it is likely that the high number of duplicated reads (Fig. 2B) may have been responsible for not only the high variation in the number of raw reads (Fig. 2A), but also the strand bias prevalent throughout our variant calls, and may have skewed the proportion of reads containing an allele on the duplicated strand, in clones where the mutation has low coverage. Consequently, higher coverage mutations were preferred for clonal expansion plots and mutations.

This pipeline does not account for the circularity of the mitochondrial genome, unlike other mitochondrial pipelines (Ding *et al.*, 2015). Consequently, reads which should align to the ends of our reference in the control region, are not retained. Given that none of our mutations were called within 150 bps of the linear reference sequence ends, no false mutations were introduced, but some may have been missed.

Overall, this pipeline allows mutations to be visually compared between generations of clones in a lineage, in order to keep mutations otherwise removed for strand bias. Our relatively relaxed quality thresholds were implemented to minimise removal of false negative calls; ambiguous low-level mutations which are otherwise indistinguishable from sequencing noise and PCR errors, were kept and true positives subsequently validated because they were also observed in related clones.

4.2 Variants in the TF1 clones

ATAC-seq enables a 17-fold or greater enrichment of mtDNA compared to exome sequencing or whole genome sequencing (Xu *et al.*, 2019), which provides high read depths necessary for confident low level variant calls. The TF1 mitochondrial genomes had high sequencing coverage (28700X post filtering, Fig. 3A and 2B) but certain short regions showed lower coverage. This could have been due to inconsistent coverage of ATAC-seq data from library preparation and sequencing bias in Illumina platforms (Ekblom *et al.*, 2014). Alternatively, homopolymers and short repeated motifs in certain regions of the mitochondrial genome can cause ambiguous read mapping (Xu *et al.*, 2019), and false mutation calling. This is compensated for by the removal of reads with low mapping quality, but since our reduced minimum mapping quality may have retained some misaligned reads, a few of our mutations may be false positive calls. The small dip in coverage at around position 3,107 (Fig. 3A), could have been caused by the removal of reads which misaligned around the deletion of the position of a historical sequencing error in the rCRS reference genome (Ju *et al.*, 2014). This dip in coverage, when compared with the much longer regions (relative to the length of homopolymers and short repeats), suggests that the inconsistent coverage is more likely caused by innate properties of ATAC-seq library preparation and Tn5 transposase cutting bias (Ekblom *et al.*, 2014; Buenrostro *et al.*, 2013).

Some clonal expansion plots showed autocorrelation but were generally below 0.01 allele frequency, and so could have represented sequencing errors (eg. Fig. 4C).

In comparison to the number of mutations initially called before lineage validation (2,339) there were few mutations visually identified as showing clonal expansion-like patterns of allele change across generations (104), at least in the eight the eight lineage paths assessed (Table 1). Between the lineage validation threshold of 0.01 allele frequency and the proportionally high number of random mutations only seen in one clone, many mutations showing clonal expansion may have been missed. A suitable change to our variant calling pipeline could be as follows: implementing a statistical test for allele frequency autocorrelation across generations, for all positions in the mitochondrial genome, could allow quick identification of positions which show generational clonal expansion of a mutation. Statistically confirming autocorrelation could allow mutations for which all clones have allele frequencies <0.01 could improve confidence in low-level variant calls, and clonal expansion to be observed and modelled at very low allele frequencies, even if none of the allele frequencies in the lineage exceeded 0.01. However, without adjustments, this approach would not identify positions in which a call was missed for one clone in the lineage (eg. Fig. 4C); visual inspection can interpret allele frequency dip as missing call. It would be interesting to test if a statistical test for autocorrelation could distinguish between false calls introduced through misaligned reads for example, and true inherited mutation showing clonal expansion, where visual examination may not.

Some mutations showed allele frequency increasing steadily (Fig. 4B), where others decreased until completely lost. This could be due to cumulative effects of random mechanisms on mutation load across the lineage, like random rates of relaxed replication among copies of the genome (Elson *et al.*, 2001; Zakirova *et al.*, 2021), or through the cumulative genetic drift, vegetive segregation. While this may have minimal effect on the overall mutation load of a clone, there would be increased variation among individual cells within the clone. Single cell sorting creates a bottleneck in the same way as vegetive segregation, but only takes a tiny random fraction of the clones' mtDNA from once clone to another across the three week generations. Genetic drift may have been especially influential because each clone underwent a bottleneck before expansion into a colony: the small effective population size of mitochondrial genomes in the bottleneck of single cell sorting that each generation goes through increases the effect of genetic drift (Li *et al.*, 2014; Schaack *et al.*, 2020; Kimura, 1968).

Alternatively selection pressures can select for or against mutations eg. (Palozzi *et al.*, 2018). Interestingly the same position in two different lineage paths showed an almost identical decrease in mutation load across the generations (8002, Fig. 4A). One possible explanation would be that the mutation was under a negative selection, although this is more likely caused by chance, as 8002 C to T encodes a silent mutation.

Some mutations were seen to disappear in a generation (Fig. 4D), before reappearing in the next generation, and continuing to follow an otherwise autocorrelated pattern of changes in allele frequency. Read coverage at the position did not seem to be the cause (Fig. 4E). Moreover some mutations were missing in one or the other of technical replicates (Fig. 6B). This could be explained by the methodology of our pipeline: true mutations are only visually validated through the lineage *after* individual variant calling with Mutserve (Weissensteiner *et al.*, 2016). If a mutation was very low-level and indistinguishable from sequencing error it could have been discarded in that clone by Mutserve's filtering algorithms, despite our less stringent thresholds of 0.001 minimum allele frequency and minimum mapping quality of 18. Additionally, as coverage decreases, more low-level mutations may be missed because Mutserve does not call mutations supported by fewer than three reads on each strand, an effective absolute minimum read depth of six. Any mutation at 0.01 allele frequency with absolutely no strand bias, covered by less than 600 reads, would be excluded by Muterve's filters. Fig. 3 shows some regions with a read depth lower than 600 (red dotted line, Fig. 3A). This strict filtering may also explain missing mutations in some generations.

The exploration of clonal expansion in the TF1 clones' could be furthered by investigating differences between clonal expansion in silent vs missense or nonsense mutations.

4.3 Comparison of analysis pipelines for variant calling in the TF1 clones

Comparison of this framework's mutations with those called by Ludwig *et al.* (2019) revealed differences between our variant calls: both in the number and position of mutations, and in the allele frequencies (Fig. 6A). Our pipeline mapped an average of 98% of the reads which aligned in their pipeline, and almost identical high coverage was achieved. Differences in our variant calls were the result of our different aims, and consequent changes in our post-alignment quality filtering and variant calling pipelines.

Firstly, their pipeline excluded all but uniquely mapped reads before variant calling; our pipeline filtered reads by excluding any with a mapping quality lower than 18, as discussed above. Using these less stringent quality filters we expected to call a higher number of mutations, and this is reflected in our variant calls: 1463 mutations were called in a total of 216 positions in our pipeline, many more than the 44 positions of Ludwig *et al.* (2019).

Moreover, allele frequencies were calculated differently. Their pipeline calculated the allele frequency of variant calls as the total proportion of all reads at a position which supported the allele. Since sequencing errors can resemble low-level variants when strand bias is not accounted for, Ludwig *et al.* (2019) calculated their own base quality cutoff by plotting base quality scores, per-base, per-allele, and fitting three gaussian distributions to the base quality scores. The lower base quality scores may have represented sequencing bias. They determined a conservative threshold of 23.8 base quality that gave a 99% chance that the base quality fell into the highest distribution, which represented correctly called bases, unlikely to have been a sequencing error. This conservative threshold was less appropriate for our pipeline because sequencing errors generally only occur on one strand, and our variant calling incorporates per strand, per allele frequency information.

It is likely that the overall effect of the different methods and parameters used in our pipelines (mapping quality filters, variant calling methods, allele frequency calculation, etc.) caused different sets of reads, and numbers of reads, to cover each genomic position. Therefore, it is unsurprising that (while there were generally calculated allele frequencies that were different between our pipelines), consistent, strong correlations between our's and Ludwig's allele frequencies were observed within the same position of each mutation, (Fig. 6A, Table 2), but varied between each position. This suggests that variant calling and allele frequency calculations were consistent within our pipelines, or a less linear relationship and wider spread of points would have been seen for each position.

Using the allele frequencies of the 40 (out of 44) overlapping calls, clones were clustered into their lineages less accurately (Figure S1D, Ludwig *et al.* 2019). The D2 and B3 lineage clones (yellow and orange respectively, Fig. 5B) in particular were not separated. Since the missing variant at position 7,074 was present only in B3 clones, we can infer that 7,074 was an influential mutation when clustering the B3 clones into their sub-lineages. The other three mutations were only present in single clones in Ludwig *et al.*'s (2019) results, and were consequently not called in this pipeline, but would not have improved the clustering of our clones into lineages. Remaining differences between our clustering can be attributed to the different allele frequencies, or the clustering methods used for the heatmap.

4.4 Evaluation of use for exploring mtDNA population dynamics

Our aim was to observe clonal expansion of heteroplasmic mutations among generations, through visually identifying allele frequency autocorrelation between generations of clones. Multiple observations of inherited mutations among related clones are needed to model clonal expansion, and the experimental setup of the TF1 clones lineages provided a suitable dataset for this analysis. Direct observation of mutations between clones is still not possible but the cumulative mutation loads of daughter cells in a clone expanded from a single mother cell could be used to model clonal expansion with a high effective sample size: the ~16kbp long mtDNA represents 16,000 observations for copy of mtDNA in each cell in each clone.

Bulk ATAC-seq data of the TF1 clones was chosen for this pipeline over the RNA-sequencing data because we wanted to look at the clonal expansion of DNA mutations, despite the more inconsistent coverage of ATAC-seq data (Fig. 3A) from library preparation and sequencing bias in Illumina platforms (Ekblom *et al.*, 2014), ATAC-seq enables a 17-fold or greater enrichment of mtDNA compared to exome sequencing or whole genome sequencing (Xu *et al.*, 2019), which allows high read depths necessary for low level variant calling. RNA-seq showed more consistent coverage of the TF1 clones' genomes (Ludwig *et al.*, 2019), but RNA reads introduce mutations that are not present in the DNA through RNA-editing. Mutations attributed to RNA editing were previously identified in this TF1 clone dataset (Ludwig *et al.*, 2019). Our coverage was high enough for low level variant calling (González *et al.*, 2020), and of 0.01 allele frequency. However, there were some regions of the genome that were covered with fewer than 300 reads (Fig. 3A) which caused difficulty when calling low-level mutations in these regions (eg. Fig. 4D). Additionally, the length of time between generational single cell isolation and expansion into a sub clone, is recorded as roughly 3 weeks (Ludwig *et al.*, 2019); a more precise value would decrease the uncertainty in models of mtDNA population dynamics.

Mutations are extensively annotated by Mutserve (Weissensteiner *et al.*, 2016), and calls can be subsetted to account for confounding factors in a model. For example, when estimating mtDNA mutation and replication rates, different regions of the genomes have been shown to mutate at different rates; the d-loop control region has a higher mutation rate than the rest of the mitochondrial genome (Parsons *et al.*, 1997). Different mutations may have selective advantages or disadvantages: missense mutations are predominantly selectively neutral (Ju *et al.*, 2014), where protein-truncating nonsense mutations are selected against in germline cells (Rebolledo-Jaramillo *et al.*, 2014), but could also undergo positive selection by the negative feedback loop theory of positive selection (Kowald and Kirkwood, 2018, 2014), or the "survival of the smallest" replicative advantage (Wallace, 1989). Mutserve annotations are therefore valuable with respect to the aims of this pipeline.

A major limitation of this pipeline is the potentially high deviation of our calculated allele frequencies from the true allele frequency, due to the pervasive strand bias. This is especially important with respect to the eventual aim of understanding mtDNA population dynamics. Therefore a significant improvement to this pipeline would be to obtain per strand allele counts to get a better idea of the level of strand bias for each mutation. Additionally, read duplicate removal may significantly decrease strand bias so more accurate allele frequencies could be calculated. Further developments for this pipeline could include detection of mtDNA deletions since they show potentially different selection pressures, and are a major course of dysfunctional mitochondria and resulting diseases (Goldstein and Falk, 2003); deletions also typically have a lower critical mutation load threshold for the detrimental phenotype than point mutations (Tuppen *et al.*, 2010).

4.5 Evaluation as a framework

This specialised pipeline can be easily adapted and customised to suit the sequencing data and research goals, assisted by its clear workflow; a different script is used for each stage in the analysis. The pipeline is implemented within a very flexible framework. This dataset-specific adaptability was demonstrated in the TF1 clones: an informed decision to adjust the minimum mapping quality was made in response to the post-alignment quality control step.

The adaptable framework enables easy analysis of other datasets from Ludwig et al.'s (2019) research. The framework is structured to allow control over which groups and individuals are included in different stages of the analysis: different subset of sequencing runs can be easily defined in `categories.txt` for use in the whole pipeline; new paths through lineages can be created by listing which clones, and in what order they should be plotted, and lineage-specific individual positions in the genome can be chosen to plot by editing `lineages.txt`. The parameters for alignment, variant calling and data exploration can be easily adjusted in response to quality of the data, or goal of an analysis, and optimised according to the data/requirements of the analysis of future research. Additionally, source control via github allows changes to the framework to be tracked.

In addition to the flexibility within Ludwig's different datasets, clearly documentation allows scripts to be modified, improved, and/or optimised so that this analysis can be developed. For example, the number of cells in each TF1 clone is unknown, but the mitochondrial genome copy number could be estimated by dividing the average sequencing depth of mtDNA by the average sequencing depth of a nuclear reference gene, and multiplying by two (Zhou *et al.*, 2020; Johnston and Jones, 2016). This is important when modelling clonal expansion because the effect of mechanisms like genetic drift, vegetative segregation and selection, on clonal expansion are influenced by the effective mtDNA population size (Elson *et al.*, 2001; Stamp *et al.*, 2018; Lawless *et al.*, 2020; Kimura, 1968). mtDNA copy numbers could potentially be estimated from the read coverages of TF1 clone NUMTs. However this estimate would likely be inaccurate as the coverage of the reference nuclear gene is dependent on transposase accessibility in ATAC-sequencing reads. In this regard, both the RNA and ATAC sequencing data of TF1 clones are not ideal (NUMTs' RNA reads are dependent on transcription level of the gene). However, the flexibility of this framework allows this variant calling pipeline to be applied to other, more suitable datasets in the sequence read archive, which may provide a better basis for mtDNA copy number estimation in addition to clonal expansion observations.

Future improvements could be made to this framework: while the R script that produces the plots and allele frequency data is highly automated (for each defined lineage: a large exploratory plot like Fig. S1, and each specified point mutation's allele frequency plot, like Fig. 4, along with tables of all allele frequencies and annotations of: all mutations in the lineage, heteroplasmic mutations, heteroplasmic and filtered for strand bias, and lineage validated mutations etc.). The R script to produce this exploratory data is well annotated, and so can be easily modified, but it is very tailored towards exploration of the bulk TF1 clones' mutations; significant adjustment with suitable programming experience would be necessary to give the same functionality to other datasets. Additionally, error catching and tracebacks are not well implemented.

Overall, this research project achieved the aims it set out: reproduction of Ludwig *et al.*'s (2019) analysis demonstrated that the updated variant calling pipeline has comparable results, in addition to calling and validating many more mutations in the TF1 clones through lineage validation, despite the different aims of our pipelines. It is well suited to call mutations and prepare data in order to model population dynamics in mtDNA: the allele frequencies of mutations that were visually identified as demonstrating clonal expansion have been provided for modelling of mtDNA population dynamics, and the hypothesis that autocorrelation indicates clonal expansion can be modeled. Subsets of allele frequency data can be provided, along with extensive useful annotations. The pipeline is mostly automated, well described, and has potential to be improved and developed to incorporate more datasets, types of analysis, and provide more data to inform modelling of mtDNA population dynamics.

The framework and pipeline is available at: https://github.com/Thomas-Fulton/Ludwig_2019.

Acknowledgements

With thanks to my excellent supervisors.

Conflict of Interest: none declared.

References

Aird,D. *et al.* (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.*, **12**, R18.

Allhoff,M. *et al.* (2013) Discovering motifs that induce sequencing errors. *BMC Bioinformatics*, **14 Suppl 5**, S1.

Andrews,R.M. *et al.* (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.*, **23**, 147.

Baslan,T. and Hicks,J. (2014) Single cell sequencing approaches for complex biological systems. *Current Opinion in Genetics & Development*, **26**, 59–65.

Benjamin,D. *et al.* (2019) Calling Somatic SNVs and Indels with Mutect2.

Bernardino Gomes,T.M. *et al.* (2021) Mitochondrial DNA disorders: From pathogenic variants to preventing transmission. *Hum. Mol. Genet.*

Birky,C.W. and William Birky,C. (2001) The Inheritance of Genes in Mitochondria and Chloroplasts: Laws, Mechanisms, and Models. *Annual Review of Genetics*, **35**, 125–148.

Bogenhagen,D. and Clayton,D.A. (1977) Mouse L cell mitochondrial DNA molecules are selected randomly for replication throughout the cell cycle. *Cell*, **11**, 719–727.

Buenrostro,J.D. *et al.* (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.

Chen,L. *et al.* (2017) DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*, **355**, 752–756.

Chinnery,P.F. and Samuels,D.C. (1999) Relaxed Replication of mtDNA: A Model with Implications for the Expression of Disease. *The American Journal of Human Genetics*, **64**, 1158–1165.

Coller,H.A. *et al.* (2001) High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection. *Nat. Genet.*, **28**, 147–150.

Ding,J. *et al.* (2015) Correction: Assessing Mitochondrial DNA Variation and Copy Number in Lymphocytes of ~2,000 Sardinians Using Tailored Sequencing Analysis Tools. *PLoS Genet.*, **11**, e1005549.

Dölle,C. *et al.* (2016) Defective mitochondrial DNA homeostasis in the substantia nigra in Parkinson disease. *Nat. Commun.*, **7**, 13548.

Durham,S.E. *et al.* (2007) Normal Levels of Wild-Type Mitochondrial DNA Maintain Cytochrome c Oxidase Activity for Two Pathogenic Mitochondrial DNA Mutations but Not for m.3243A→G. *The American Journal of Human Genetics*, **81**, 189–195.

Ebbert,M.T.W. *et al.* (2016) Evaluating the necessity of PCR duplicate removal from next-generation sequencing data and a comparison of approaches. *BMC Bioinformatics*, **17**.

Ekblom,R. *et al.* (2014) Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. *BMC Genomics*, **15**, 467.

Elson,J.L. *et al.* (2001) Random intracellular drift explains the clonal expansion of mitochondrial DNA mutations with age. *Am. J. Hum. Genet.*, **68**, 802–806.

Evans,M.D. *et al.* (2004) Oxidative DNA damage and disease: induction, repair and significance. *Mutat. Res.*, **567**, 1–61.

Ewels,P. *et al.* (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.

Floros,V.I. *et al.* (2018) Author Correction: Segregation of mitochondrial DNA heteroplasmy through a developmental genetic bottleneck in human embryos. *Nat. Cell Biol.*, **20**, 991.

Friedman,J.R. and Nunnari,J. (2014) Mitochondrial form and function. *Nature*, **505**, 335–343.

Goldstein,A. and Falk,M.J. (2003) Mitochondrial DNA Deletion Syndromes. In, Adam,M.P. *et al.* (eds), *GeneReviews*. University of Washington, Seattle, Seattle (WA).

González,M.D.M. *et al.* (2020) Sensitivity of mitochondrial DNA heteroplasmy detection using Next Generation Sequencing. *Mitochondrion*, **50**, 88–93.

Greaves,L.C. *et al.* (2014) Clonal expansion of early to mid-life mitochondrial DNA point mutations drives mitochondrial dysfunction during human ageing. *PLoS Genet.*, **10**, e1004620.

Guo,Y. *et al.* (2012) The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics*, **13**, 666.

Gu,Z. *et al.* (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.

Henderson,D.A. *et al.* (2009) Bayesian Emulation and Calibration of a Stochastic Computer Model of Mitochondrial DNA Deletions in Substantia Nigra Neurons. *Journal of the American Statistical Association*, **104**, 76–87.

Hirano,M. *et al.* (1992) Melas: an original case and clinical criteria for diagnosis. *Neuromuscul. Disord.*, **2**, 125–135.

Johnston,I.G. and Jones,N.S. (2016) Evolution of Cell-to-Cell Variability in Stochastic, Controlled, Heteroplasmic mtDNA Populations. *Am. J. Hum. Genet.*, **99**, 1150–1162.

Ju,Y.S. *et al.* (2014) Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *Elife*, **3**.

Kaufman,B.A. *et al.* (2019) Mitochondrial DNA, nuclear context, and the risk for carcinogenesis. *Environ. Mol. Mutagen.*, **60**, 455–462.

Kimura,M. (1968) Evolutionary Rate at the Molecular Level. *Nature*, **217**, 624–626.

Koboldt,D.C. *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.

Kowald,A. and Kirkwood,T.B.L. (2018) Resolving the Enigma of the Clonal Expansion of mtDNA Deletions. *Genes* , **9**.

Kowald,A. and Kirkwood,T.B.L. (2014) Transcription could be the key to the selection advantage of mitochondrial deletion mutants in aging. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 2972–2977.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Lawless,C. *et al.* (2020) The rise and rise of mitochondrial DNA mutations. *Open Biol.*, **10**, 200061.

Legros,F. *et al.* (2004) Organization and dynamics of human mitochondrial DNA. *Journal of Cell Science*, **117**, 2653–2662.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,M. *et al.* (2010) Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. *Am. J. Hum. Genet.*, **87**, 237–249.

Li,R. *et al.* (2014) Somatic point mutations occurring early in development: a monozygotic twin study. *J. Med. Genet.*, **51**, 28–34.

Ludwig,L.S. *et al.* (2019) Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell*, **176**, 1325–1339.e22.

Marquis,J. *et al.* (2017) MitoRS, a method for high throughput, sensitive, and accurate detection of mitochondrial DNA heteroplasmy. *BMC Genomics*, **18**, 326.

Matkarimov,B.T. and Saparbaev,M.K. (2020) DNA Repair and Mutagenesis in Vertebrate Mitochondria: Evidence for Asymmetric DNA Strand Inheritance. *Adv. Exp. Med. Biol.*, **1241**, 77–100.

Meacham,F. *et al.* (2011) Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, **12**, 451.

Michikawa,Y. *et al.* (1999) Aging-dependent large accumulation of point mutations in the human mtDNA control region for replication. *Science*, **286**, 774–779.

Morris,J. *et al.* (2017) Pervasive within-Mitochondrion Single-Nucleotide Variant Heteroplasmy as Revealed by Single-Mitochondrion Sequencing. *Cell Rep.*, **21**, 2706–2713.

Nakamura,K. *et al.* (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.*, **39**, e90.

Nooteboom,M. *et al.* (2010) Age-associated mitochondrial DNA mutations lead to small but significant changes in cell proliferation and apoptosis in human colonic crypts. *Aging Cell*, **9**, 96–99.

Palozzi,J.M. *et al.* (2018) Mitochondrial DNA Purifying Selection in Mammals and Invertebrates. *J. Mol. Biol.*, **430**, 4834–4848.

Parsons,T.J. *et al.* (1997) A high observed substitution rate in the human mitochondrial DNA control region. *Nat. Genet.*, **15**, 363–368.

Payne,B.A.I. *et al.* (2013) Universal heteroplasmy of human mitochondrial DNA. *Hum. Mol. Genet.*, **22**, 384–390.

Potapov,V. and Ong,J.L. (2017) Examining Sources of Error in PCR by Single-Molecule Sequencing. *PLoS One*, **12**, e0169774.

Rebolledo-Jaramillo,B. *et al.* (2014) Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 15474–15479.

Robin,E.D. and Wong,R. (1988) Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. *J. Cell. Physiol.*, **136**, 507–513.

Rossignol,R. *et al.* (2003) Mitochondrial threshold effects. *Biochem. J*, **370**, 751–762.

Ross,M.G. *et al.* (2013) Characterizing and measuring bias in sequence data. *Genome Biol.*, **14**, R51.

Santibanez-Koref,M. *et al.* (2019) Assessing mitochondrial heteroplasmy using next generation sequencing: A note of caution. *Mitochondrion*, **46**, 302–306.

Santorsola,M. *et al.* (2016) A multi-parametric workflow for the prioritization of mitochondrial DNA variants of clinical interest. *Hum. Genet.*, **135**, 121–136.

Satoh,M. and Kuroiwa,T. (1991) Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell. *Exp. Cell Res.*, **196**, 137–140.

Schaack,S. *et al.* (2020) Disentangling the intertwined roles of mutation, selection and drift in the mitochondrial genome. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **375**, 20190173.

Schneider,V.A. *et al.* (2017) Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.*, **27**, 849–864.

Sharpless,N.E. and DePinho,R.A. (2007) How stem cells age and why this makes us grow old. *Nat. Rev. Mol. Cell Biol.*, **8**, 703–713.

Shaw,C.A. (2002) Theoretical consideration of amplification strategies. *Neurochem. Res.*, **27**, 1123–1131.

Sims,D. *et al.* (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.*, **15**, 121–132.

Smith,A.L.M. *et al.* (2020) Age-associated mitochondrial DNA mutations cause metabolic remodeling that contributes to accelerated intestinal tumorigenesis. *Nature Cancer*, **1**, 976–989.

Stamp,C. *et al.* (2018) Predominant Asymmetrical Stem Cell Fate Outcome Limits the Rate of Niche Succession in Human Colonic Crypts. *EBioMedicine*, **31**, 166–173.

Stewart,J.B. and Chinnery,P.F. (2015) The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nat. Rev. Genet.*, **16**, 530–542.

Trifunovic,A. *et al.* (2004) Premature ageing in mice expressing defective mitochondrial DNA polymerase. *Nature*, **429**, 417–423.

Tuppen,H.A.L. *et al.* (2010) Mitochondrial DNA mutations and human disease. *Biochim. Biophys. Acta*, **1797**, 113–128.

Van der Auwera,G.A. and O'Connor,B.D. (2020) Genomics in the Cloud: Using Docker, GATK, and WDL in Terra O'Reilly Media.

Vellarikkal,S.K. *et al.* (2015) mit-o-matic: a comprehensive computational pipeline for clinical evaluation of mitochondrial variations from next-generation sequencing datasets. *Hum. Mutat.*, **36**, 419–424.

Wallace,D.C. (1989) Mitochondrial DNA mutations and neuromuscular disease. *Trends Genet.*, **5**, 9–13.

Weissensteiner,H. *et al.* (2016) mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. *Nucleic Acids Res.*, **44**, W64–9.

Wickham,H. (2009) ggplot2: Elegant Graphics for Data Analysis Springer Science & Business Media.

Xu,J. *et al.* (2019) Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. *Elife*, **8**.

Ye,K. *et al.* (2014) Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 10654–10659.

Yoneda,M. *et al.* (1992) Marked replicative advantage of human mtDNA carrying a point mutation that causes the MELAS encephalomyopathy. *Proc. Natl. Acad. Sci. U. S. A.*, **89**, 11164–11168.

Yu-Wai-Man,P. *et al.* (2010) Somatic mitochondrial DNA deletions accumulate to high levels in aging human extraocular muscles. *Invest. Ophthalmol. Vis. Sci.*, **51**, 3347–3353.

Zakirova,E.G. *et al.* (2021) Natural and Artificial Mechanisms of Mitochondrial Genome Elimination. *Life*, **11**.

Zhidkov,I. *et al.* (2011) MitoBamAnnotator: A web-based tool for detecting and annotating heteroplasmy in human mitochondrial DNA sequences. *Mitochondrion*, **11**, 924–928.

Zhou,K. *et al.* (2020) A Novel Next-Generation Sequencing-Based Approach for Concurrent Detection of Mitochondrial DNA Copy Number and Mutation. *J. Mol. Diagn.*, **22**, 1408–1418.

Ziada,A.S. *et al.* (2020) Updating the Free Radical Theory of Aging. *Front Cell Dev Biol*, **8**, 575645.

Zinovkina,L.A. (2018) Mechanisms of Mitochondrial DNA Repair in Mammals. *Biochemistry (Moscow)*, **83**, 233–249.

Supplementary Material

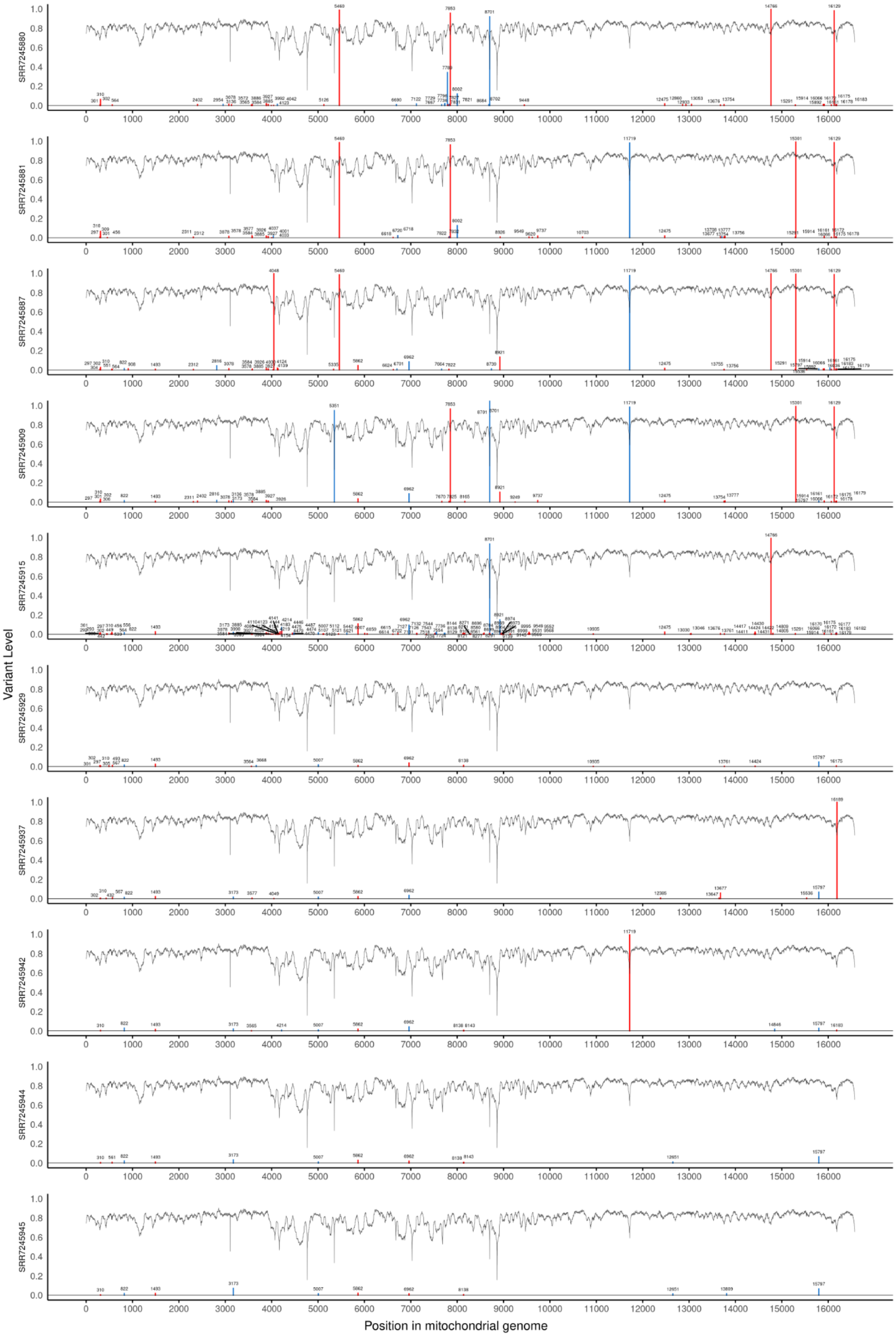

Fig. S1: Exploratory plot of all mutations for each clone, ordered by generation. Red

indicates strand bias.

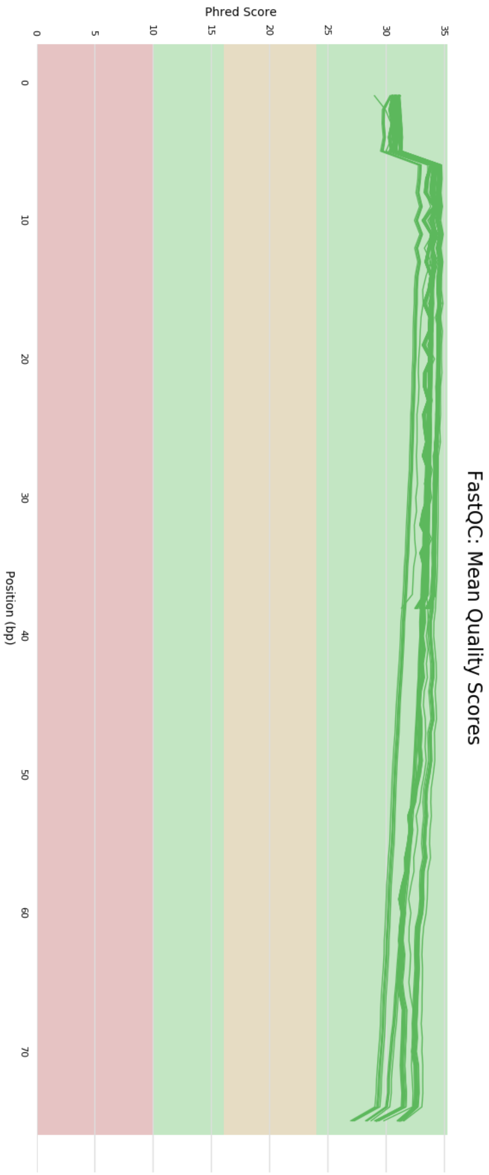

Fig. S2. Mean base quality scores across reads for each clone.

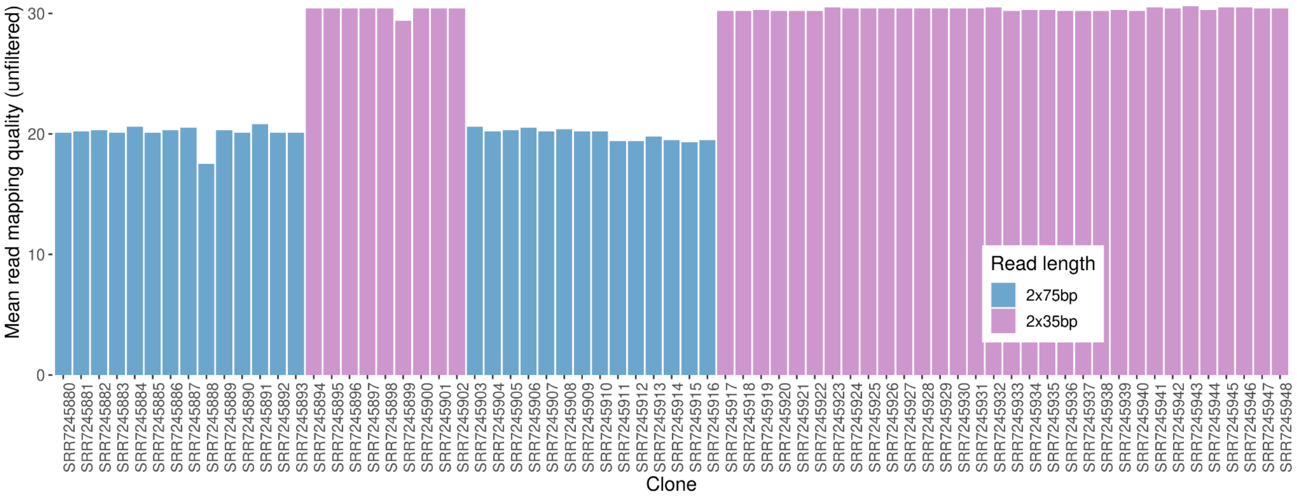

Fig. S3: Mean read mapping qualities per clone, before quality filtering. Mean mapping quality groups according to sequencing read length.