Genomics Technologies

Project Report

Group 2 - Pink

Elisabetta Callegaro elisabetta.callegaro@studenti.unitn.it
Alice Casata alice.casata@studenti.unitn.it
Alessio Gambella alessio.gambella@studenti.unitn.it
Annalisa Xamin annalisa.xamin@studenti.unitn.it

University of Trento - Via Sommarive 9, 38123 Povo (TN), Italy

Our proposal: custom primer-based PacBio sequencing

Our laboratory experience aimed to sequence and computationally analyze the mitochondrial genomes of healthy human Retinal Pigment Epithelium (RPE) cells and compare them with distinct lines of neoplastic cells to identify potential variants. After researching the current state of the art, we decided to challenge the conventional exploitation of Illumina technology for mtDNA sequencing. Our proposal is grounded on the PacBio Sequel IIe System to directly analyze pre-isolated mtDNA, in order to eliminate the need for a library preparation step:

1. The sequencing process begins with the recovery of entire mitochondria from sample cells through Fractionated Mitochondria Magnetic Separation (FMMS). Although more time-intensive when compared to alternative approaches, this sensitive technique results in optimal yields of highly purified and integral mitochondria, and can also be applied to samples derived from various tissues.
2. After the successful recovery of mitochondria, mtDNA isolation is carried out using a mitochondrial lysis buffer to release genetic material from the organelles. Subsequent steps involve phenol-chloroform extraction and treatment of the mtDNA-containing aqueous phase with RNases and proteases to remove contaminants.
3. Afterwards, a Nanodrop assessment is conducted to ensure the quality of the extracted genetic material and a Qubit fluorometer is employed to quantify the recovered mtDNA. Additionally, a low-field agarose gel electrophoresis is performed on a small fraction of the isolated material to further evaluate sample purity and verify the length and integrity of circular mtDNA molecules, exploiting a digestion by restriction enzymes.
4. The final step leverages the PacBio Sequel IIe platform for the direct sequencing of circular mtDNA: this system was chosen owing to its capability to produce long-reads, with a high-accuracy, through the Circular Consensus Sequence (CCS) mode. To perform the sequencing, an initial denaturation of the double-stranded circular molecules is required. This is followed by the annealing of a custom-made primer designed to target the highly conserved COX1 mitochondrial gene, which bears a subsequence that partially matches the canonical PacBio primer target. Therefore, due to the substantial similarity between the two sequences, it is possible to maintain the machine parameters. Finally, to avoid the potential lack of sequencing due to mutations in the COX1 subsequence, a solution is to employ a combination of primers.

This approach allows to sequence by directly analyzing the entire cyclic mtDNA, eliminating the need for time-consuming library preparation and reducing the necessary starting material. However, a key challenge of the approach is the inherent circular structure of mtDNA, causing tangled nucleic acid rings that may hinder polymerase processivity and sequencing. Still, the use of PacBio technology for the direct sequencing of small genomes without the need for library preparation has already been documented in the literature, even if not for the analysis of mtDNA.

Experimental Laboratory Activity

During the didactic laboratories, an alternative procedure, divergent from the aforementioned approach, was implemented. Initially, the QuickExtract™ DNA Extraction Solution protocol was used for the extraction of both nuclear and mtDNA from healthy RPE cells. Subsequently, a PCR amplification was performed to enhance the mitochondrial genome. More precisely, two distinct sets of primers were utilized:

- MITO primers: three primer pairs that led to the amplification of three distinct yet overlapping amplicons of the mitochondrial genome, collectively spanning its entirety.

- Region of interest (ROI) primers: comprising four primer pairs, this set directed the amplification of three ROIs that posed a greater challenge for amplification. In a preceding experiment, these regions exhibited a coverage lower than 2000x.

The obtained amplicons underwent size verification to ensure that the PCR products were as expected. MITO amplicons were analyzed by agarose gel electrophoresis (**Supplementary Data - Figure S1**), while ROI amplicons were subjected to capillary electrophoresis using the high-resolution LabChip GX Instrument. Notably, in the gel electrophoresis for MITOs, the absence of the band corresponding to the third amplicon suggests a potential issue, likely inadequate primer mixing before addition to the master mix. Subsequent Qubit quantification of MITO amplicons confirmed the unsuccessful amplification of MITO3. To proceed, a backup amplicon was utilized in the subsequent experimental steps.

In contrast to the agarose gel electrophoresis, purification of PCR products was required before the capillary electrophoresis of ROI amplicons. The purification step was performed by employing the Agencourt AMPure XP purification system, based on paramagnetic bead technology and selectively retains DNA sequences exceeding 100bp (ROI amplicons had a length of ~ 500bp). The selection of the DNA fragment length was ensured by selecting a precise DNA/PEG ratio, thus effectively discriminating against potential primer doublets and isolating PCR products. Following purification, ROI amplicons underwent capillary electrophoresis (**Supplementary Data - Figure S2**) and were subsequently quantified using the Qubit. The results confirmed that all samples exhibited a satisfactory concentration and an appropriate size.

Subsequently, a library from MITO amplicons was acquired by using the Nextera XT kit. Therefore, the same amounts of the three amplicons were combined to ensure an equal representation.

The subsequent steps involved Tag Genomic DNA, which entailed fragmentation and tagging of fragments with adapter sequences through transposase action, and library amplification. The last one involved adding i7 and i5 adapters, along with sequences for cluster generation, through a limited-cycle PCR amplification. In the case of ROI amplicons, barcodes and sequencing adapters were incorporated using a PCR process with only 6 cycles.

Finally, the libraries underwent purification using the Agencourt AMPure XP bead system to eliminate any residual reagents. Then, a quality control analysis was conducted using the Qubit dsDNA HS Assay, followed by capillary electrophoresis on the Revvity LabChip GX. The results of the MITO library are illustrated in **Figure 1**, where a notable observation is the predominant presence of fragments ranging from 200-250 bp. This suggests that during the tagmentation step the transposase overly fragmented the amplicons due to prolonged activity, and the subsequent PCR process favored the selection of shorter fragments. Ideally, library fragments are expected to be around 300bp for optimal information content. However, performing an additional size selection step would lead to further complexity depletion.

In addition, there is the need for repeated purification as bands below 150 bp are discernible: these lower-sized bands indicate the presence of primer dimers or free primers, posing potential interference with the sequencing process. For this reason, NGS analysis was performed using the Illumina MiSeq System in paired-end mode by using a backup library.

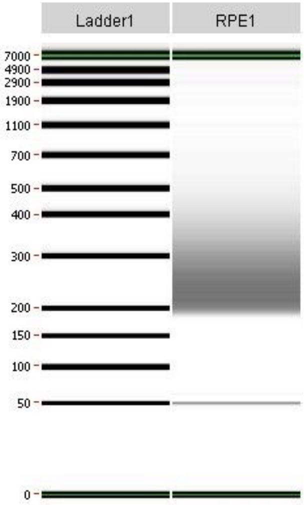


Figure 1 | Results of capillary electrophoresis on Revvity LabChip GX of the MITO library.

Computational Laboratory Activity

After sequencing the biological samples, the project proceeded into the computational analysis phase. Firstly, we evaluated the sequencing data's quality, both pre- and post-adapter removal, by utilizing FastQC. Adapters were removed by using Trimmomatic, which eliminates low-quality regions at the read ends and excludes excessively short reads. Following QC, the next step involved aligning the reads with a reference genome. The genome sequence primary assembly of GRCh38 from GENCODE was downloaded and used as a reference genome. Then, it was preprocessed to build an index, a measure that significantly reduced alignment time and expedited the overall process. BWA was used for indexing and alignment, resulting in the generation of SAM files. To enhance file readability and conserve RAM, SAM files were converted to BAM using samtools. The alignment results were finally examined using IGV.

After obtaining the BAM file from the alignment, variant calling was performed to discern discrepancies in DNA sequences between the samples and the reference genome. Employing the GATK suite, specifically Mutect2, facilitated the computation of the allelic fractions which served as a filtering criterion (as SNVs with a very low allelic fraction may arise due to sequencing errors or very rare subclonal mutations). The output of Mutect2 underwent additional refinement through FilterMutectCalls to eliminate technical artefacts and sequencing errors. Following variant calling, variant functional annotation was conducted by filtering the VCF files previously acquired. Bcftools was used to keep chromosome M only, as the focus was on the variants on the mitochondrial chromosome. Subsequently, the filtered VCF file was uploaded onto the VEP (Variant Effect Predictor) web interface. To gather additional information about the variants identified in VEP, the dbSNP database was consulted.

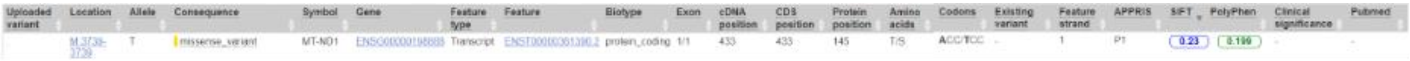
Up to this point, each of the aforementioned steps was independently executed for every sample. For comparative analysis, all samples underwent merging into a singular VCF file. Finally, the SNPRelate R package was used to compute and plot a PCA.

Results and Discussion

Following the analysis, the VCF files were loaded on the Variant Effect Predictor (VEP) by Ensembl. **Table 1** represents the VEP results of the RPE-ROI, considering only the variations for which SIFT and PolyPhen scores are available: the APPRIS parameter indicates the importance of the transcript, while SIFT and PolyPhen are predictors of the impact of the amino acid substitution on the function of the protein, based on homology and structural information, respectively.

In our case, the gene MT-ND1 (mitochondrially encoded NADH) contains a missense mutation (G>A), but SIFT and PolyPhen parameters have conflicting values, since PolyPhen predicts a benign variant, while the low SIFT score suggests the variant to be somewhat deleterious. In addition, this variant seems to be unknown, therefore it is not possible to determine which of the two parameters is more reliable.

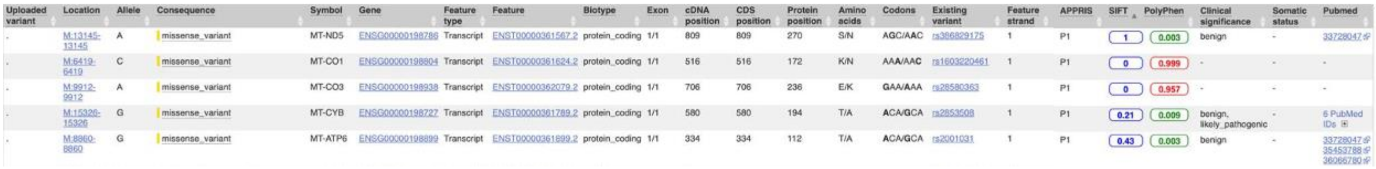
Table 1 | VPE results showing a SIFT and PolyPhen score for RPE-ROI.



Uploaded variant	Location	Allele	Consequence	Symbol	Gene	Feature type	Feature	Biotype	Exon	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Feature strand	APPRIS	SIFT	PolyPhen	Clinical significance	Pubmed
	M:3739-3739	T	missense_variant	MT-ND1	ENSG00000198888	Transcript	ENST00000361390.2	protein_coding	1/1	433	433	145	T/S	ACC/TCC	-	1	P1	0.23	0.199	-	-

Table 2 shows the VPE result of the RPE-MITO: five genes have missense mutations and two of them show conflicting SIFT and PolyPhen values (MT-NDS, MT-CO1 and MT-CO3), probably because they are taking into account different information. All the variants have already been known and documented. Notably, one of them is located on the COX1 gene which was intended for exploitation in our initial proposal; however, the variant is not included within the primer binding region we had considered for our analysis.

Table 2 | VPE results showing a SIFT and PolyPhen score for RPE-MITO.



Uploaded variant	Location	Allele	Consequence	Symbol	Gene	Feature type	Feature	Biotype	Exon	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Feature strand	APPRIS	SIFT	PolyPhen	Clinical significance	Somatic status	Pubmed
-	M:13145-13145	A	missense_variant	MT-ND5	ENSG00000198786	Transcript	ENST00000361567.2	protein_coding	1/1	809	809	270	S/N	AGC/AAC	rs386829175	1	P1	1	0.003	benign	-	33728047
-	M:6419-6419	C	missense_variant	MT-CO1	ENSG00000198804	Transcript	ENST00000361624.2	protein_coding	1/1	516	516	172	K/N	AAA/AAC	rs1603220461	1	P1	0	0.999	-	-	-
-	M:9912-9912	A	missense_variant	MT-CO3	ENSG00000198938	Transcript	ENST00000362079.2	protein_coding	1/1	706	706	236	E/K	GAA/AAA	rs28580363	1	P1	0	0.957	-	-	-
-	M:15326-15326	G	missense_variant	MT-CYB	ENSG00000198727	Transcript	ENST00000361789.2	protein_coding	1/1	580	580	194	T/A	ACA/GCA	rs2853508	1	P1	0.21	0.009	benign, likely_pathogenic	-	6 PubMed IDs
-	M:8860-8860	G	missense_variant	MT-ATP6	ENSG00000198899	Transcript	ENST00000361899.2	protein_coding	1/1	334	334	112	T/A	ACA/GCA	rs2001031	1	P1	0.43	0.003	benign	-	33728047 35453788 36066780

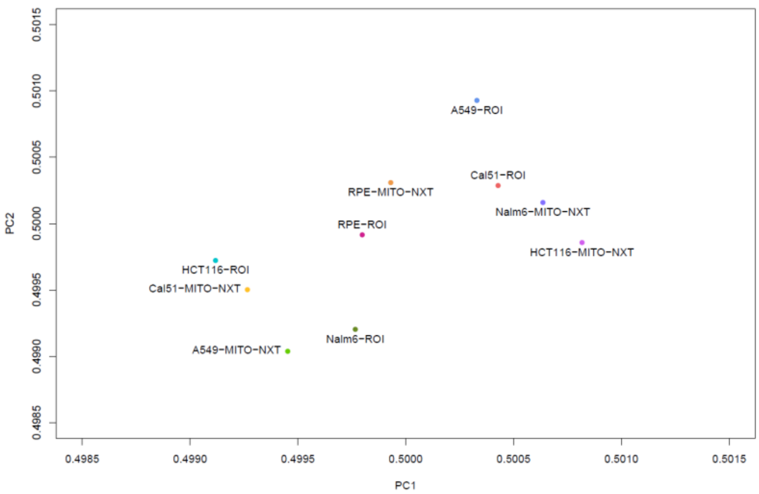


Figure 2 | PCA plot generated by RStudio, comparing RPE cells and tumor cell lines.

PCA results (**Figure 2**) diverge from our expectations. Initially, as RPE cells are a healthy cell line, we expected that the RPE-ROI and RPE-MITO points would cluster together, while appearing distinct from the other tumoral cell lines which, given their neoplastic nature, might be associated with higher mutational scores.

Contrary to these expectations, all cell lines are tightly concentrated within a very restricted region in the 2D PCA space, without a clear distinction between them. The analysis of MITO and ROI distribution within the same cell line does not reveal a discernible pattern as well. The overall lack of differentiation not only prevents us from gaining a hint of the presence of ROI mutations that were not captured by the MITO reads, but also prevents us from asserting that tumoral cell lines exhibit a higher mutational burden compared to normal RPE cells.

We hypothesize that this situation arises from a potential error during the *in-silico* analysis of the sequencing data, perhaps from conflicts between the tools utilized on the laboratory computers. Further strengthening our hypothesis there is a significant alteration in cluster distribution upon introducing a random value into the 2D PCA space, as could be expected when dealing with remarkably similar data.

Due to the unreliable outcomes of the PCA, the VCF file was consulted. As expected, tumor cell lines exhibited generally more variations than healthy RPE cells (**Supplementary Data - Table S1**). Additionally, ROIs, despite showing fewer overall variations than MITOs, proved capable of capturing variations undetected by MITO amplicons, thus confirming the usefulness of generating two distinct types of amplicons. More in depth, it was revealed that the only variant with an alternative allele shared by all neoplastic cell lines, yet absent from healthy RPE cells, is located at position ChrM:2465 (**Supplementary Data - Figure S3**), corresponding to the gene MT-RNR2. However, the clinical significance of this SNP, as well as its role in tumorigenesis, remains unknown. Furthermore, we identified eleven SNPs presenting the alternative allele in RPE cells but not in cells of neoplastic lineage (**Supplementary Data - Table S2**). Once again, the clinical significance of these alterations is currently unknown. Nonetheless, most of these variations are synonymous mutations or are localized in non-coding regions. This suggests that their impact on functionality may be irrelevant, which is coherent with the overall healthiness of the RPE cell line.

Conclusions

From this learning experience, it is clear that validating our initial proposal is crucial before exploiting it for a specific application, such as comparing mutational burdens among healthy and tumor samples. However, even the miSeq-mediated sequencing of mitochondrial genomes, though a robust and reliable approach, was not without challenges. This is exemplified by the unsuccessful amplification of MITO3, likely originating from an error related to our inexperience, and the poor outcome of the obtained library in terms of both purity and fragment size. In addition, the downstream computational analysis yielded completely unsatisfactory results, forcing us to exploit alternative manual analyses to gain some meaningful insights from the data. As a consequence, the drawn conclusions are mainly qualitative, poorly reliable and lacking a proper statistical evaluation.

Supplementary Data

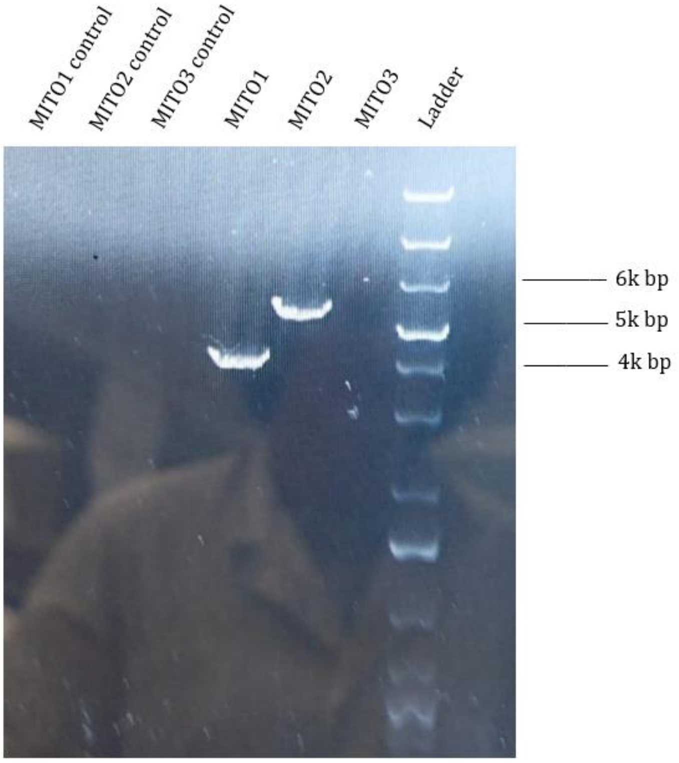


Figure S1 | Gel electrophoresis of MITO amplicons. The MITO3 band is not visible.

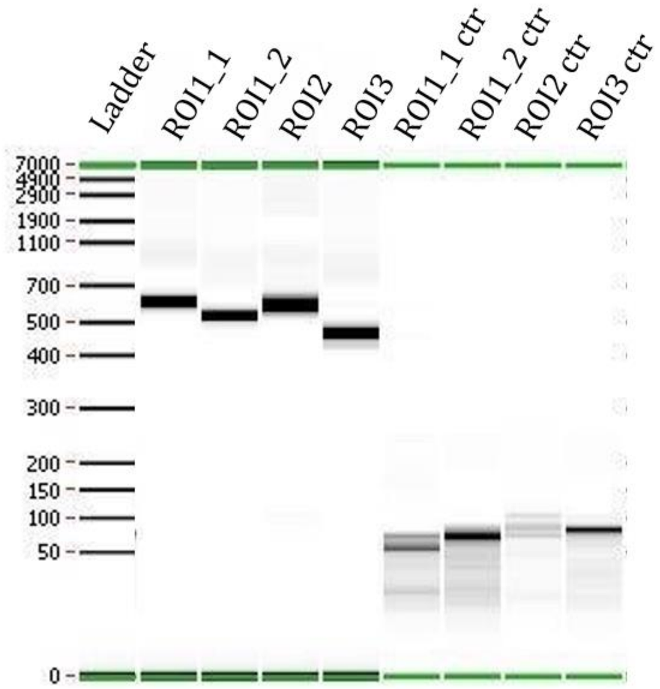
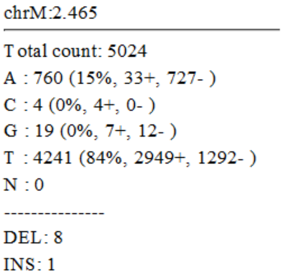


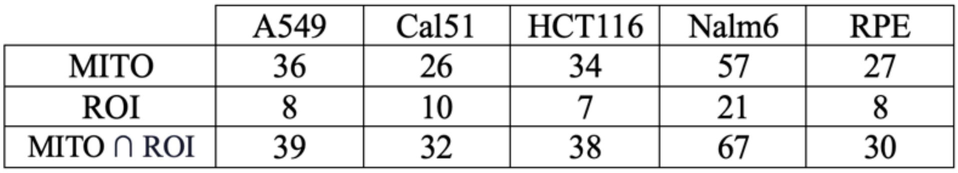
Figure S2 | Results of capillary electrophoresis of the ROI amplicons. The control bands (without DNA template) are shorter in length and not purified.



```
chrM:2.465
Total count: 5024
A : 760 (15%, 33+, 727- )
C : 4 (0%, 4+, 0- )
G : 19 (0%, 7+, 12- )
T : 4241 (84%, 2949+, 1292- )
N : 0
---------------
DEL: 8
INS: 1
```

Figure S3 | IGV base calling count in position 2456 of the mitochondrial chromosome.

Table S1 | Mutational burden of cancer vs healthy cell lines. The variations have been counted.



	A549	Cal51	HCT116	Nalm6	RPE
MITO	36	26	34	57	27
ROI	8	10	7	21	8
MITO ∩ ROI	39	32	38	67	30

Table S2 | Position and type of the SNPs found in RPE samples but not in tumor cells.

Position in ChrM	Gene	Type	RPE MITO	RPE ROI
152	-	Upstream/downstream gene variant	yes	yes
1959	MT-RNR2	Non-coding transcript exon variant	yes	no
3363	MT-ND1	-	no	yes
5318	MT-ND2	Synonymous variant	yes	no
5691	MT-TN	-	no	yes
9912	MT-CO3	Missense variant	yes	no
9950	MT-CO3	Synonymous variant	yes	no
13145	MT-ND5	Missense variant	yes	no
15466	MT-CYB	Synonymous variant	yes	no
15721	MT-CYB	Synonymous variant	yes	no
16192	-	Upstream/downstream gene variant	yes	no