GENOMICS AGE: READING AND UNDERSTANDING OUR GENOMES

HANA GARAJ[a], ORIOL NAVARRO TRAVESSET[a], MAGDALENA ŽIVKOVIĆ[a], TUGCE BILGIN SONAY[a,b]

[a] S3++ Summer School of Science, Pozega, Croatia

[b] University of Zurich, Switzerland

ABSTRACT

Cancer is a complex, lethal disease very frequent in humans. One of its causes is mutations in our DNA, which can be identified through recent sequencing techniques and genomics research. The aim of this project is to process, analyse and compare raw short reads from the mitochondrial DNA of healthy individuals and cancer patients to find out mutations that can be related to cancer. Our results revealed three cancer specific mutations in genes that code for NADH dehydrogenase complex and cytochrome b. Mutations on these genes have been found to facilitate tumour growth and we suggest that this might be because malfunctioning of their proteins causes higher oxidative stress in the cell due to electron leakage from transportation paths.

INTRODUCTION

Cancer is a disease caused by an uncontrolled division of abnormal cells in a part of the body. About 14.6 % of deaths in 2012 were caused by cancer (8.2 million deaths). All tumour cells show the six hallmarks of cancer. These include: resisting programmed cell death, insensitivity to anti-growth signals and overgrowing (tumour suppressor genes don't work properly), sustaining proliferative signalling (oncogenes occur by mutation of DNA), including angiogenesis (their own blood supply) and activating an invasion and metastasis.

There are various alternations that can cause cancer. Genes can acquire mutations in their sequence, leading to different variants, known as alleles, in the population. These alleles encode slightly different versions of a protein, which cause different phenotype traits. Mutations including single nucleotide substitutions, insertions, deletions, inversions and copy number changes can cause cancer. Also, gene expression mechanisms can be modified by processes like methylation, non-coding RNA, or DNA-binding proteins. But what is actually a gene?

A gene is a locus region of DNA which is made up of nucleotides and is the molecular unit of heredity. The Human Genome Project has estimated 20,000 human protein-coding genes. But now it is increasingly clear that the non-coding DNA has a very important role to play. That role is still largely unknown but is likely to include regulating which genes are ‘switched on’ or ‘switched off’ in each cell.

The human genome is the complete set of nucleic acid sequence for humans, encoded as DNA within the 23 chromosome pairs in cell nuclei and in a small DNA molecule found within individual mitochondria. Genomics is the discipline that sequences, assembles, and analyses the function and structure of genomes using recombinant DNA, DNA sequencing methods, and bioinformatics.

In our project, we use data acquired by the method of shotgun sequencing. Shotgun sequencing is a sequencing method designed for analysis of DNA sequences longer than 1000 base pairs, up to and including entire chromosomes. Here, DNA polymerase (the enzyme in cells that synthesizes DNA) is used to generate a new strand of DNA from a strand of interest. In the sequencing reaction, the enzyme incorporates into the new DNA strand individual nucleotides that have been chemically tagged with a fluorescent label. As this happens, the nucleotide is excited by a light source, and a fluorescent signal is emitted and detected. The signal is different depending on which of the

four nucleotides was incorporated. This method can generate 'reads' of 125 nucleotides in a row and billions of reads at a time.

After sequencing individual fragments, the sequences can be reassembled on the basis of their overlapping regions. This allows the longer sequence to be assembled from shorter pieces. In this process, each base has to be read not just once, but at least several times in the overlapping segments to ensure accuracy.

These recent technologies allow us to sequence DNA for less than 1000 dollars in a day, and as such have revolutionized the study of genomics and molecular biology. Researchers can use DNA sequencing to search for genetic variations and/or mutations that may play a role in the development or progression of a disease. Single nucleotide polymorphisms (SNPs) are variations in a single nucleotide at a specific position in the genome and represent the most common kinds of mutations (incidence higher than 1% of population). Most SNPs are harmless but some can cause proteins to be manufactured or fold incorrectly and thus cause diseases.

For this project, we are looking at and analysing mutations of the mitochondrial DNA (mtDNA) because of its important role in the control of a cell's energy metabolism. In a healthy cell, most energy is produced through the oxidative phosphorylation system (OXPHOS), and almost 4/5ths of protein complexes it consists of are encoded by the mtDNA [1]. Mutations on the mtDNA can significantly change metabolic pathways in the cell. These mutations allow organisms to adapt to new environments, and this phenomenon can even be observed in humans where specific mtDNAs lineages can be found only on some continents. But, like all mutations, mutations on the mtDNA are mostly deleterious and have been correlated with a broad range of metabolic and degenerative diseases, from diabetes and autism to early childhood death [2]. The mutation rate of mtDNA is on average 10 times higher than that of nuclear DNA (nDNA) [2] because of its closeness to reactive oxygen species (ROS) by-products of respiration and thus higher oxidative stress [3]. Mutation rate also increases with age. To safeguard against deleterious mutations in a single cell, there are, unlike in the nucleus, multiple copies of mtDNA present in the mitochondria at any time, and a typical mammalian cell typically contains 10^3-10^4 copies in total. For any mutation to thus be significant, most copies have to possess it [3].

Tumour cells survive in a different environment than other cells in the sense of different goals (growth and proliferation at the fastest rate achievable versus collaboration). When normal cells detect that a mtDNA is damaged beyond repair, apoptosis is triggered, but not in cancer cells who have apoptosis inhibited [1]. Homoplasmic (meaning present in almost all copies in a

cell) mutations in the mtDNA have been found in the most common cancers in the regions that code for the Krebs cycle [3]. Back in 1927 Warburg noticed that tumour cells produce most of their energy through the process of glycolysis rather than OXPHOS [1]. Glycolysis produces only 2 ATPs and a lot of lactate, so it is hypothesized that this is because it is a much faster process and so preferable to cancer cells that do not worry about efficiency but only personal growth and proliferation. Changes in the metabolic pathways trigger a kind of a vicious circle causing higher oxidative stress in the entire cell which then in turn causes even more mutations in both the nDNA and mtDNA. For this reason cancer cells can evolve and adapt relatively fast to new environments. Research has found that cancer cells have less anti-oxidants present which usually work to keep the oxidative homeostasis, most notably cytochrome c and b [3], and also that the aggressive and rapidly growing ones have the highest incidence of mitochondrial dysfunction.

In this project we will check the significance of SNPs identified in the genomes of different individuals (both healthy and cancer patients), gathered from online databases and processed using online platforms. Our final aim is to learn how to use bioinformatics tools to analyse raw next generation sequencing data and obtain whole genome sequences that we can then compare and analyse to search for possible cancer related mutations.

METHODS

In order to find variants between some individuals' genome and the reference genome, we first had to gather the short reads of the sequenced mitochondrial DNA from the specific online platforms. Then, these data was processed using an online platform called Galaxy [4], which made possible to obtain the final consensus file that contained the whole sequence from each individual's mitochondrial DNA. With this information, we were able to elaborate a phylogenetic tree with all the individuals (a mother and a child, both healthy, and a cancer patient). Finally, we analysed the variants between the different sequences and we gathered information about the genes that contained these highly possible mutations.

In this project we analysed the short reads from three mother and child pairs from a genomic study [5] on blood and buccal mitochondrial DNA, with each sample sequenced ~20,000× per site. These short reads from the different healthy individuals have been gathered from the NCBI (National Center for Biotechnology Information) [6] database, an online platform that contains data of the sequenced genome from some healthy individuals in a Fastq format. The UCSC (University of California, Santa Cruz) [7] Table Browser has been used for gathering the reference sequenced mitochondrial DNA.

We used the program FastQC [8] to analyse these short reads and its quality, and to decide which parameters we should use when filtering these data. The short reads were trimmed and filtered using the online tool Galaxy in order to increase its overall quality (Fig. 1). The trimming was made because we observed that the quality of the reads at their ends was considerably worse than the rest, and the filtering parameters were set so that we wouldn't lose more than 25% of our data.

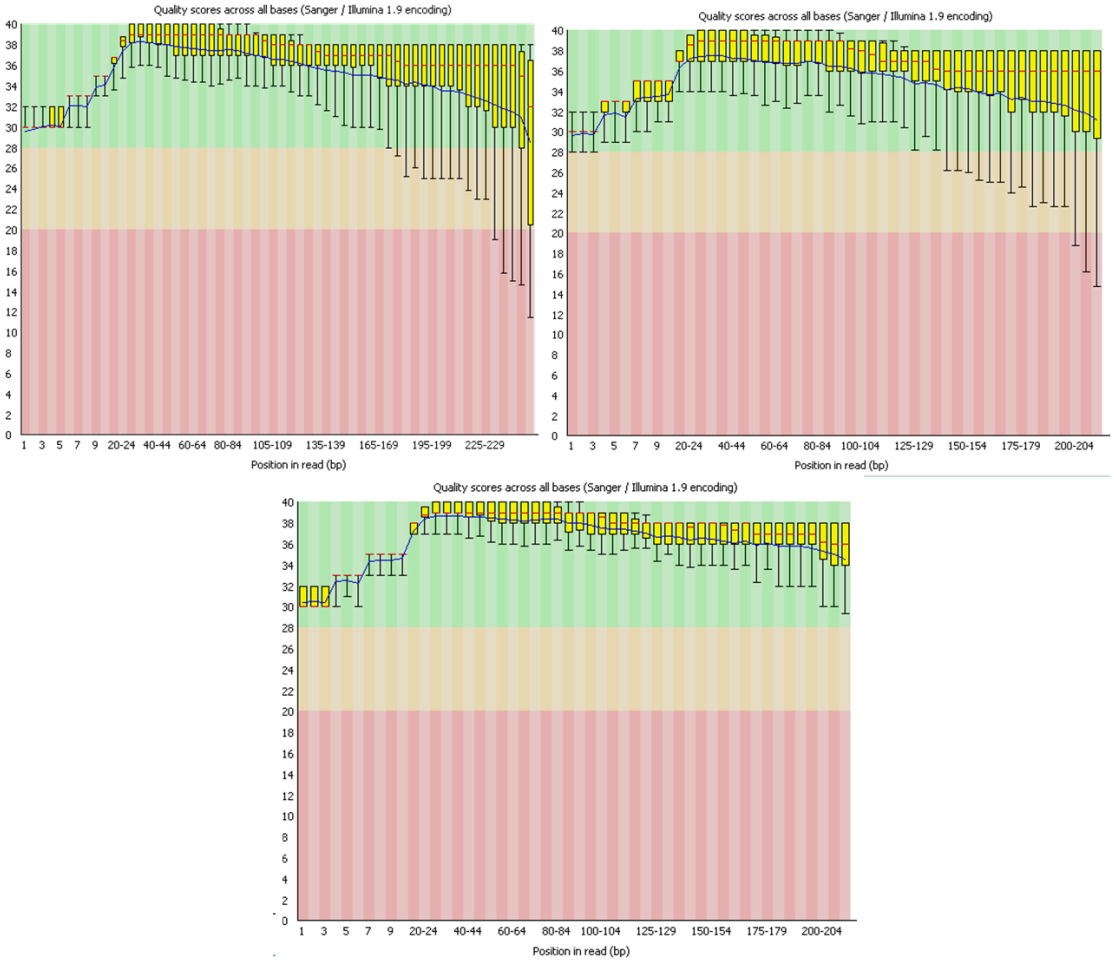

Figure 1: Example of a quality score distribution for a healthy individual sample: first figure corresponds to the raw short reads, second figure corresponds to the trimmed short reads, and third figure corresponds to the filtered short reads.

The next step was aligning the short filtered reads against the reference genome using the Bowtie2 [9] tool in the Galaxy platform, obtaining as a result a BAM (Binary sequence Alignment/Map) file. Then, different SAMtools [10] were used to filter and generate a pileup from the BAM file. At the end, we generated the final consensus in a fasta format. The same procedures were applied on a BAM file which contained the sequenced mitochondrial DNA of an acute myeloid leukemia patient.

Once we obtained the desired fasta files, we concatenated them (each pair of mother and child's sequenced genomes and the cancer patient's

sequenced genome) and obtained the multiple aligned file (Fig. 2). Phylogenetic and molecular evolutionary analyses were conducted using MEGA version 4 (Tamura, Dudley, Nei, and Kumar 2007) [11], and a phylogenetic tree of all the treated individuals was generated.

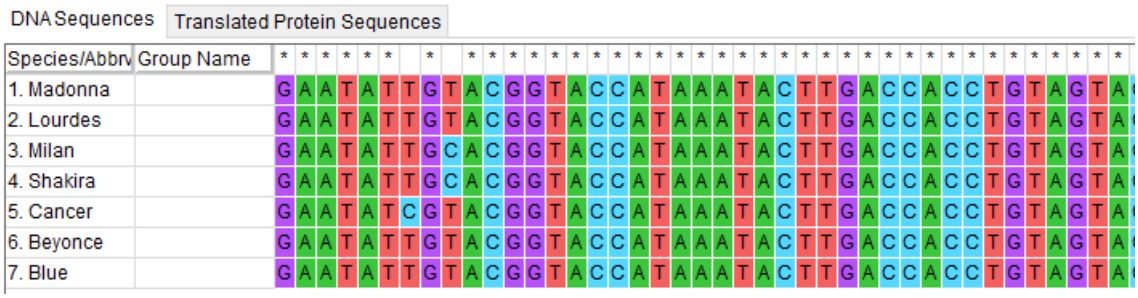

Figure 2: Multiple genomes from the individuals aligned together in MEGA.

The pileup we generated was in a VCF (Variant Call File) format, and it was used to identify the variants between the individuals and the reference genome (chromosome M). The common SNPs (Single Nucleotide Polymorphisms) in humans were downloaded from the UCSC Table Browser, and used to identify its intersection with the variants we identified. That allowed us to search for the genes that contained these SNPs and to see if these were related with cancer.

RESULTS

The aim of this project was to analyse raw sequence data from six healthy individuals and one acute myeloid leukemia cancer patient in order to find variants between them and deduce whether or not these variants could be mutations which potentially lead to cancer. We gathered all the data from online databases and processed it to make a phylogenetic tree and to find the variants between them, studying the genes that contained these variants and their relation to cancer.

For each individual we got at least 736000 short reads after filtering them, which confirms that the data had a very good quality. We used these reads to generate a consensus file and then we concatenated the data from every individual in order to generate a phylogenetic tree of the individuals (Fig. 3):

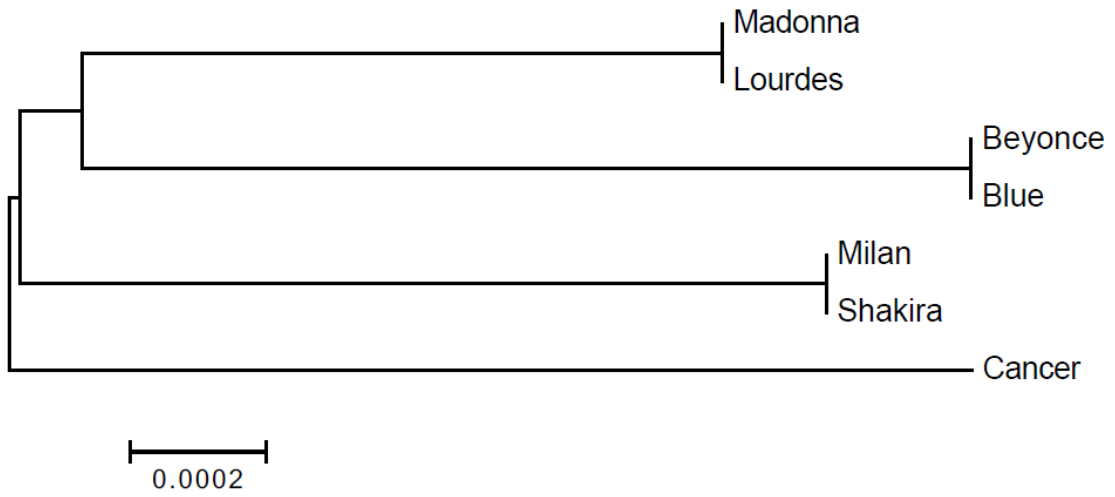

Figure 3: Phylogenetic tree of the 3 mother and child pairs and the cancer patient. The scale represents the amount of genetic changes.

In Fig. 3 we can see that all the mother and child pairs are clustered together and that the cancer patient is the farthest one.

After that, we identified the variants between the individuals and the reference chromosome genome and we determined which of them overlap with known SNPs in human populations. On average, we identified 69 variants in our individuals' sequenced genome, but on the end an average of 13 of them overlapped with the known SNPs. 6 of those were found only in the cancer genome, and 3 of them corresponded to a missense variant.

The genes where these variants were found are the following: the CYTB gene (which encodes for cytochrome b), the MT-ND5 gene (which encodes for NADH dehydrogenase) and the MT-ND3 gene (which also encodes for NADH dehydrogenase).

DISCUSSION

After processing the initial data from all individuals, we have been able to create a phylogenetic tree to see genetic differences between them, and we have also found which genes contained the SNPs that were most likely to cause cancer.

The phylogenetic tree shows all mother and child pairs clustered together. That's most probably because the child has inherited the mitochondrial DNA only from the mother, and so their evolutionary distance is almost 0. The first mother and child pair (Madonna & Lourdes) is clustered together with Beyonce and Blue's branch, not with Shakira and Milan's branch; a possible explanation for this phenomenon could be that both Shakira and Milan have accumulated more mutations than the other pairs, and so their evolutionary distance is greater. However, the cancer patient is the one who's

further from the rest, and that makes sense with our hypothesis because the cancer patient is likely to have more mutations than the rest.

We have found non-synonymous mutations (mutations which change the peptide in a given protein, potentially influencing its function) in the regions coding for complex I and complex III. The MT-ND3 and MT-ND5 code for core proteins in the complex I, the largest and most complicated of all protein complexes involved in the respiration, and are responsible for catalysing binding of and electron transport from NADH molecules. Since the electrons that escape from this transfer (0.15-2% in a healthy cell) [1] are responsible for most ROS creation, we suggest that this might be one way that mutations of these genes benefit tumor cell proliferation. The CYBT gene codes for cytochrome b in complex III. This complex contributes to the proton gradient and cytochrome b is responsible for the further transport of electrons. When the electron transfer through cytochrome b is, for whatever reason, reduced, complex III might leak electrons that then form superoxides with molecular oxygen. Since that causes additional oxidative stress, this mutation's benefits to tumor growth might be explained by the reduction the effectiveness of cytochrome b in electron transport.

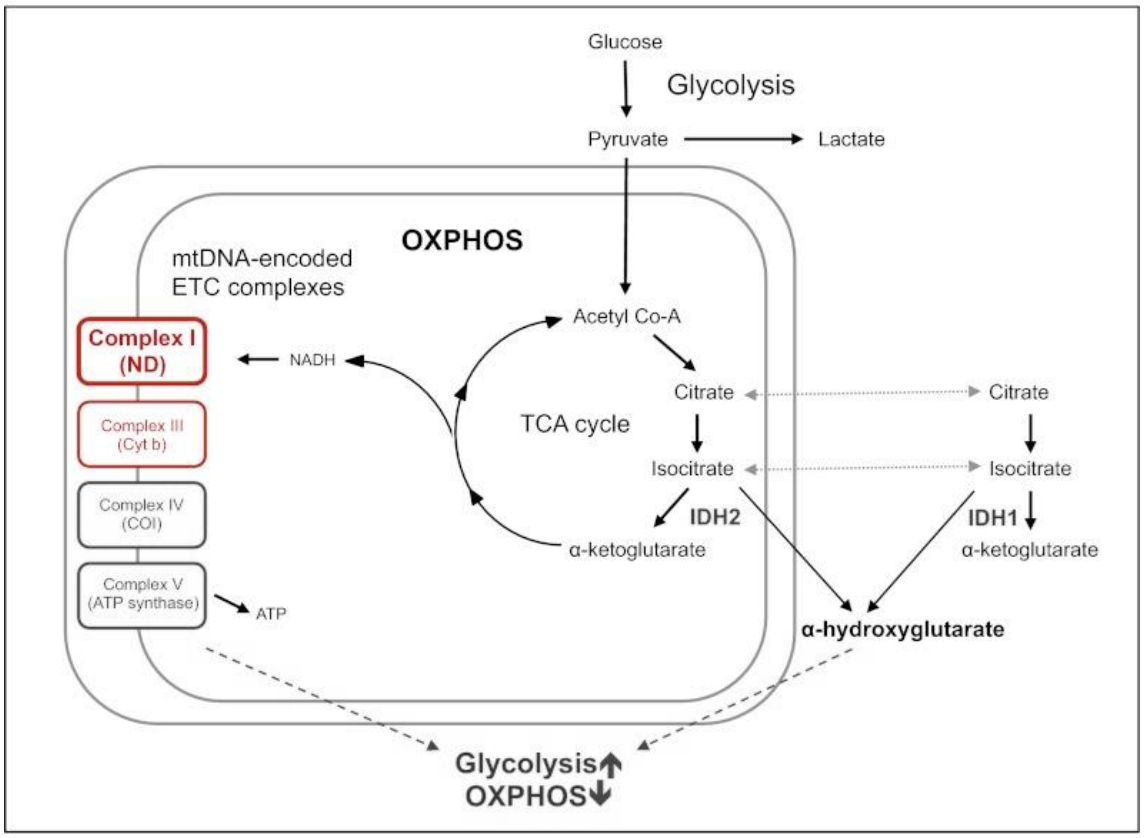

Figure 4: Proposed pathway by which the mutations we identified lead to metabolic deregulation altering the relative amounts of oxidative phosphorylation and glycolysis. Figure altered [12].

Through comparison with literature, we have found that the mutation of MT-ND3 gene is associated with higher cancer incidence [13]. Dasgupta et al. have conducted experiments with overexpression of the mutated CYBT gene in

in vitro tumor cells and have noticed accelerated growth, apoptosis resistance and higher ROS concentrations [14], while in a similar experiment with overexpressed mutated MT-ND5 they have noticed higher proliferation, invasiveness and superoxide concentration [15]. Although the exact mechanism by which these mutations benefit tumors is still unclear, we propose that these mtDNA mutations also encourage tumor growth and survival in our cancer patient too, and that we have found a likely cause of cancer in our patient's genome.

REFERENCES

[1] VAN GISBERGEN MW, VOETS, AM, STARMANS MH, DE COO IF ET AL. How do changes in the mtDNA and mitochondrial dysfunction influence cancer and cancer therapy? Challenges, opportunities and models. Mutat. Res. Rev. Mutat. Res. 2015, 764, 16–30

[2] DOUGLAS W. Genetics: Mitochondrial DNA in evolution and disease. Nature PY 2016 print 535 7613 498 500 Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. 0028-0836 http://dx.doi.org/10.1038/nature18902 L3 - 10.1038/nature18902 News & Views

[3] CHATTERJEE A, MAMBO E, SIDRANSKY D. Mitochondrial DNA mutations in human cancer Oncogene 2006 0000///print 25 34 4663 4674 0950-9232 http://dx.doi.org/10.1038/sj.onc.1209604

[4] AFGAN E, BAKER D, VAN DEN BEEK M, BLANKENBERG D ET AL. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Research 2016 doi: 10.1093/nar/gkw343

[5] REBOLLEDO-JARAMILLO B, SHU-WEI SU M, STOLER N, MCELHOE J ET AL. Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA PNAS 2014 111 (43) 15474-15479

[6] http://www.ncbi.nlm.nih.gov/sra

[7] KAROLCHIK D, HINRICHS AS, FUREY TS, ROSKIN KM ET AL. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D493-6

[8] ANDREWS S. FastQC: a quality control tool for high throughput sequence data. 2010 Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

[9] LANGMEAD B, SALZBERG SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012, 9: 357-359. 10.1038/nmeth.1923

[10] LI H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011 Nov 1;27(21):2987-93. Epub 2011 Sep 8. [PMID: 21903627]

[11] TAMURA K, DUDLEY J, NEI M & KUMAR S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution 2007 24:1596-1599. (Publication PDF at http://www.kumarlab.net/publications)

[12] LARMAN T, DEPALMA S, HADJIPANAYIS A, PROTOPOPOV A ET AL. Spectrum of somatic mitochondrial mutations in five cancers PNAS 2012 109 (35) 14087-14091; published ahead of print August 13, 2012, doi:10.1073/pnas.1211502109

[13] CZARNECKA AM, KRAWCZYK T, ZDROŻNY M ET AL. Breast Cancer Res Treat 2010 121: 511. doi:10.1007/s10549-009-0358-5

[14] DASGUPTA S, HOQUE MO, UPADHYAY S, SIDRANSKY D. Forced cytochrome B gene mutation expression induces mitochondrial proliferation and prevents apoptosis in human uroepithelial SV-HUC-1 cells. 2009 International Journal of Cancer, vol. 125, no. 12, pp. 2829–2835

[15] DASGUPTA S, SOUDRY E, MUKHOPADHYAY N, SHAO C ET AL. Mitochondrial DNA mutations in respiratory complex-I in never-smoker lung cancer patients contribute to lung cancer progression and associated with EGFR gene mutation. 2012 J Cell Physiol 227:2451–2460