Integrative analysis and visualization of multi-omics data of mitochondria-associated diseases

Po Yee Annie Yim

München 2022

DISSERTATION

DER FAKULTÄT FÜR BIOLOGIE

DER LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Integrative analysis and visualization of multi-omics data of mitochondria-associated diseases

Vorgelegt von

PO YEE ANNIE YIM

aus Hongkong

München, den 09.02.2022

1. Gutachter: Prof. Dr. Barbara Conradt
2. Gutachter: Prof. Dr. Wolfgang Enard

Dissertation eingereicht am: 09.02.2022
Datum der Disputation: 09.12.2022

Eidesstattliche Erklärung

Ich versichere hiermit an Eides statt, dass die vorgelegte Dissertation von mir selbständig und ohne unerlaubte Hilfe angefertigt ist.

Erklärung

Hiermit erkläre ich, dass die Dissertation nicht ganz oder in wesentlichen Teilen einer anderen Prüfungskommission vorgelegt worden ist.

München, den 07.02.2022

Po Yee Annie Yim

List of Publications

Yim A, Koti P, Bonnard A, Marchiano F, Dürrbaum M, Garcia-Perez C, Villaveces J, Gamal S, Cardone G, Perocchi F, Storchova Z, Habermann BH. **mitoXplorer, a visual data mining platform to systematically analyze and visualize mitochondrial expression dynamics and mutations.** Nucleic Acids Res. 2020 Jan 24;48(2):605-632. doi: 10.1093/nar/gkz1128. PMID: 31799603; PMCID: PMC6954439.

Krempel R, Kulkarni P, Yim A, Lang U, Habermann BH, Frommolt P. **Integrative analysis and machine learning on cancer genomics data using the Cancer Systems Biology Database (CancerSysDB).** BMC Bioinformatics. 2018 Apr 24;19(1):156. doi: 10.1186/s12859-018-2157-7. PMID: 29699486; PMCID: PMC5921751.

Arechederra M, Daian F, Yim A, Bazai KB, Richelme S, Dono R, Saurin AJ, Habermann BH, Maina F. **Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer.** Nat Commun. 2018 Aug 8;9(1):3164. doi: 10.1038/s41467-018-05550-5. Erratum in: Nat Commun. 2018 Sep 25;9(1):3976. PMID: 30089774; PMCID: PMC6082886.

Fan Y, Bazai SK, Daian F, Arechederra M, Richelme S, Temiz NA, Yim A, Habermann BH, Dono R, Largaespada DA, Maina F. **Evaluating the landscape of gene cooperativity with receptor tyrosine kinases in liver tumorigenesis using transposon-mediated mutagenesis.** J Hepatol. 2019 Mar;70(3):470-482. doi:10.1016/j.jhep.2018.11.027

Declaration of contributions as co-first authors

Yim A, Koti P, Bonnard A, Marchiano F, Dürrbaum M, Garcia-Perez C, Villaveces J, Gamal S, Cardone G, Perocchi F, Storchova Z, Habermann BH. **mitoXplorer, a visual data mining platform to systematically analyze and visualize mitochondrial expression dynamics and mutations.** Nucleic Acids Res. 2020 Jan 24;48(2):605-632. doi: 10.1093/nar/gkz1128. PMID: 31799603; PMCID: PMC6954439.

Po Yee Annie Yim contributed to this work by designing and implementing of the visualizations and data analysis presented in Figure 3, 4, 5, 7 and 8, as well as the design and development of the platform architecture and the web server. In addition, she contributed to the design and creation of figures, and interpretation of the results.

Prasanna Koti contributed to the initial idea of this work and was responsible for the annotation work documented in Table 1, as well as the design and implementation of the visualization presented in Figure 2. He also contributed to the design and development of the platform architecture and the web server.

München, den 07.02.2022

Po Yee Annie Yim Prof. Dr. Barbara Conradt Prof. Dr. Bianca Habermann

Dr. Prasanna Koti

Declaration of contributions as co-author

Krempel R, Kulkarni P, Yim A, Lang U, Habermann BH, Frommolt P. **Integrative analysis and machine learning on cancer genomics data using the Cancer Systems Biology Database (CancerSysDB).** BMC Bioinformatics. 2018 Apr 24;19(1):156. doi: 10.1186/s12859-018-2157-7. PMID: 29699486; PMCID: PMC5921751.

Po Yee Annie Yim contributed to this work by designing the interactive analysis workflow specifically for mitochondrial functions, presented in Figure 3. Further, she performed data analysis using this workflow presented in Table 4 and interpretation of the results.

Arechederra M, Daian F, Yim A, Bazai KB, Richelme S, Dono R, Saurin AJ, Habermann BH, Maina F. **Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer.** Nat Commun. 2018 Aug 8;9(1):3164. doi: 10.1038/s41467-018-05550-5. Erratum in: Nat Commun. 2018 Sep 25;9(1):3976. PMID: 30089774; PMCID: PMC6082886.

Po Yee Annie Yim contributed to this work by performing computational work using the human HCC cohort from TCGA database and data analysis presented in Figure 1e, 1f, 1g and 1h. In addition, she contributed to the interpretation of the results.

München, den 07.02.2022

Po Yee Annie Yim Prof. Dr. Barbara Conradt Prof. Dr. Bianca Habermann

Table of Contents

1 Abbreviations

ACP	Acyl Carrier Protein
AMP	Adenosine Monophosphate
ATP	Adenosine Triphosphate
Bcl-2	B-Cell Lymphoma 2
BAK	Bcl-2 Homologous Antagonist/Killer Protein
BAX	Bcl-2-Associated X Protein
BWA	Burrows-Wheeler Aligner
CSS	Cascading Styling Sheets
ChIP-seq	Chromatin Immunoprecipitation Sequencing
cDNA	Complementary DNA
Hsa21	Copy Number Variations To Chromosome 21
CGIs	CpG Islands
dNTP	Deoxynucleoside Triphosphate
DHODH	Dihydroorotate Dehydrogenase
D-loop	Displacement Loop
DNA-seq	DNA-Sequencing
DOM	Document Object Model
DS	Down Syndrome
DRP	Dynamin-Related Protein
ETC	Electron Transport Chain
ER	Endoplasmic Reticulum
Eci1	Enoyl-Coenzyme A Delta Isomerase 1
ELISA	Enzyme-Linked Immunosorbent Assays
FAS	Fatty Acid Synthesis
FAD	Flavin Adenine Dinucleotide
GEO	Gene Expression Omnibus
GDC	Genomic Data Commons
GTP	Guanosine Triphosphate
HSP	H-Strand Promoter
HCC	Hepatocellular Carcinoma
HCT116	Human Colon Cancer Cell Line
RPE1	Human Retinal Pigment Epithelial-1

HIF	Hypoxia-Inducible Factor
IMM	Inner Mitochondrial Membrane
IGV	Integrative Genomics Viewer
IGV	Integrative Genomics Viewer
JSON	Javascript Object Notation
KIRP	Kidney Renal Papillary
LSP	L-Strand Promoter
LHON	Leber Hereditary Optic Neuropathy
mtFAS	Mitochondria Fatty Acid Synthesis
MCU	Mitochondrial Ca2+ Uniproter
mtDNA	Mitochondrial DNA
MELAS	Mitochondrial Encephalopathy
MOMP	Mitochondrial Outer Membrane Permeabilization
MRPS21	Mitochondrial Ribosomal Protein S21
POLRMT	Mitochondrial RNA Polymerase
Mfn	Mitofusins
MEF	Mouse Embryonic Fibroblast
NCI	National Cancer Institute
NCBI	National Center For Biotechnology Information
NHGRI	National Human Genome Research Institute
NGS	Next-Generation Sequencing
NAD	Nicotinamide Adenine Dinucleotide
NoSQL	Non-Structured Query Language
Opa1	Optic Atrophy-1
Oat	Ornithine Aminotransferase
OMM	Outer Mitochondrial Membrane
OXPHOS	Oxidative Phosphorylation
PID	Persistent Identifier
PCA	Principle Component Analysis
PINK1	Pten-Induced Putative Kinase 1
ROS	Reactive Oxygen Species
FADH/FADH2	Reduced Form of Flavin Adenine Dinucleotide
NADH	Reduced Form of Nicotinamide Adenine Dinucleotide
RelA	Rel-Associated Protein

RT-PCR	Reverse Transcription Polymerase Chain Reaction
rRNA	Ribosomal RNA
RNA-seq	RNA-Sequencing
SVG	Scalable Vector Graphics
SAGE	Serial Analysis of Gene Expression
STAR	Spliced Transcripts Alignment To A Reference
SILAC	Stable Isotope Labeling in Cell Culture
SQL	Structured Query Language
SUCLG1	Succinate-Coa Ligase Subunit Alpha
SUCLG2	Succinate-Coa Ligase Subunit Beta
SOD-1	Superoxide Dismutase
t-SNE	T-Distributed Stochastic Neighbor Embedding
THF	Tetrahydrofolate
TCGA	The Cancer Genome Atlas
TPM	Transcript Per Million
TFAM	Transcription Factor A
tRNA	Transfer RNA
TIM	Translocase of The Inner Mitochondrial Membrane
TOM	Translocase of The Outer Mitochondrial Membrane
TCA	Tricarboxylic Acid
T21	Trisomy 21
VDAC	Voltage-Dependent Anion Channel
Yap1	Yes-Associated Protein 1

2 List of Figures

3 Summary

Mitochondria are subcellular organelles that play a crucial role in cellular bioenergetics and apoptosis, and thus are essential in maintaining normal cell function and regulating cell death pathways. In addition, they are also involved in the biosynthesis and balance of metabolite, cell signaling and anti-oxidant defense. Dysfunction of mitochondrial processes, potentially due to mutations of mitochondrial genes, could therefore lead to detrimental consequences such as severe neurodegenerative, cardiovascular and multisystemic metabolic disorders, and has also been implicated in many other diseases including cancer.

Despite being associated with various diseases, platforms or tools that are specifically designed for the exploration of expression and mutation landscapes of mitochondrial genes did not exist. Therefore, this thesis aimed to develop a visual data mining tool exclusive for mitochondrial genes, that could help enhance our understanding of the role of mitochondria in disease pathology, by leveraging the availability of high volume of -omics data.

In the first study, mitoXplorer, a web-based visual data mining platform with a set of dynamic, interactive and intuitive visualization tools was developed. We manually assembled and curated lists of genes with annotations that consist of ~1200 genes for four model species including human. These lists (mitochondrial interactomes) were integrated into the visualization tools to allow in-depth analysis of mitochondrial mutations and expression dynamics on public data sets hosted on mitoXplorer. The analysis of the transcriptome and proteome data of trisomy 21 (aneuploidy) cell lines inferred defects in mitochondria respiration, which was then verified experimentally, hence proving the predictive power of the platform, and its ability to provide testable hypotheses that could lead to the discovery of underlying molecular mechanisms of diseases.

In the second study, we have integrated a visualization module and the human mitochondrial interactome as a workflow on CancerSysDB, a platform that allows user to perform integrated analyses across multiple data types of public cancer dataset from The Cancer Genome Atlas (TCGA) research network. The visual approach analysis of expression and clinical data of KIRP (kidney renal papillary cell carcinoma) patients revealed the association between the expression of two Tricarboxylic acid (TCA) cycle genes - Succinate-CoA ligase subunits, SUCLG1 and SUCLG2, and cancer stages. This proved the practicality such visual data mining tool when analyzing large -omics dataset, and also the importance of mitochondrial functions in cancer development and progression. To continue along this line, in the third study, we explored also the methylome and transcriptome data from both a mouse model and public cancer dataset of hepatocellular

carcinoma (HCC). We discovered another possible operating mechanism in this cancer type, where the hypermethylation in CpG islands (CGIs) leads to up-regulations of a set of genes, including the gene Jun, that belongs to the process Transcription in our human mitochondrial interactome. The enrichment of this gene set observed in the mouse model was also found in 56% of HCC patients from the TCGA dataset, which were characterized to belong to an aggressive HCC subclass, hence suggesting the clinical relevance of this gene set and its potential to be used as biomarkers for patient stratifications.

Taken together, mitochondria have a multi-faceted role in cell function, as well as the pathology of various diseases. Analysis tools dedicated for the investigation of mitochondrial mutation and expression dynamics is thus necessary. The visual data mining approach adopted by mitoXplorer and CancerSysDB has been proven in this thesis to be a robust way for the exploration of data in the context of mitochondrial functions, that could help delineating molecular mechanisms in different disease conditions, through suggesting testable hypothesis for further experimental validation.

4 Introduction

4.1 Mitochondria – more than a powerhouse

Mitochondria are subcellular organelles that exist in many different forms, from round punctate to filaments generating complex networks, in cells. While they are best known for producing cellular energy in the form of ATP by a process called oxidative phosphorylation (OXPHOS) (van der Giezen & Tovar, 2005), they are also involved in many other cellular functions such as signaling, balance of metabolites, or anti-oxidant defense, just to name a few. Their many cellular roles make them one of the most important organelles with an immense impact on metabolism and homeostasis of most eukaryotic cells (Dyall et al., 2004).

4.1.1 Morphology of mitochondria

Mitochondria have a double membrane system, that consists of the outer mitochondrial membrane (OMM) and the inner mitochondrial membrane (IMM), both composed of phospholipid bilayers and proteins (Alberts et al., 1994) and separated by the intermembrane space. The outer membrane is relatively permeable to small molecules due to the presence of porins, but it also regulates the movement of larger molecules and proteins. The translocase on the outer membrane (TOM) and the inner membrane (TIM) together form a complex that recognizes signal sequences and transport proteins across the mitochondrial membrane. The inner mitochondrial membrane is mostly impermeable and encloses an area referred to as mitochondrial matrix that contains essential enzymes for mitochondrial functions and multiple copies of the highly compacted mitochondrial DNA genome (mtDNA). The inner membrane also protrudes into the matrix to form multiple foldings called cristae, which accommodate the electron transport chain (ETC) complexes for oxidative phosphorylation and structurally varies from tissues to tissues (Duchen, 2004).

The shapes of mitochondria also vary and they can form a complex interconnecting network of mitochondria called syncytium (Friedman & Nunnari, 2014). However, they can also exist as individual structure called puncta and rods (Anesti & Scorrano, 2006). The networks and copy number of mitochondria are constantly changing and are maintained by the balance between mitochondrial fusion and fission, biogenesis and mitophagy (Chan, 2012; Hoitzing et al., 2015). These processes are important in bioenergetic homeostasis and the maintenance of mitochondrial function. Mitochondrial fusion generates extended network and allows mixing of mitochondrial content and gene products, which counteracts the damages to mtDNA by reactive oxygen species (ROS). In case of insufficient functional mitochondria, the network could also be replenished from external sources (other cells or

intercellular space) through tunneling nanotubules (Rustom, 2004) or vesicles (Tkach & Théry, 2016). Together with biogenesis of mitochondria, these two processes tend to increase bioenergetics efficiency to meet high demands (Gomes et al., 2011; Y. J. Liu et al., 2020). On the other hand, mitochondrial fission facilitates turnover by creating new daughter organelles, where the dysfunctional ones are removed by mitophagy. This provides a mechanism to segregate damaged mitochondria and helps to maintain a healthy mitochondrial population and hence the bioenergetic capacity (Twig et al., 2008; Westermann, 2012).

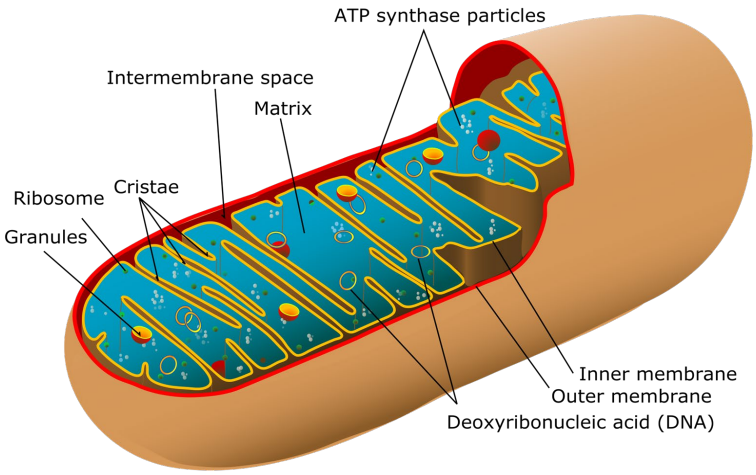


Figure 4.1 Basic structure of a mitochondrion.

Mitochondria have an inner and outer membrane, with an intermembrane space between them. The inner membrane is arranged into cristae to increase surface area for the accommodation of enzymes such as ATP synthase for oxidative phosphorylation. The space enclosed by the inner membrane is known as the matrix, which consists of essential enzymes, as well as mitochondrial DNA, ribosomes and calcium granules. Taken from https://commons.wikimedia.org/wiki/Mitochondrion, under the terms and conditions of the Creative Commons Universal (CC0) License.

4.1.2 Mitochondrial genome

The mitochondrial genome (mtDNA) is the only source of extranuclear DNA in animals (Mita et al., 1990). In human, it is a circular, double-stranded DNA of 16,569 base-pair that encodes for 37 genes, 13 of which are core proteins of the mitochondrial respiratory complexes I-IV within the OXPHOS system, two ribosomal RNAs and 22 tRNAs (S. Anderson et al., 1981). Most of the genes are encoded on the heavy strand (H-strand), where as the light strand (L-strand) encodes for only 8 tRNA and a complex I subunit. The mtDNA is a very compact genome with no introns and very few non-coding bases between genes. The only major non-coding region is the displacement loop (D-loop), with a structure that two genomic strands are separated and displaced by a third strand of DNA. The D-loop in mtDNA is around 1.1 kb and controls its transcription and replication (Shadel & Clayton, 1997).

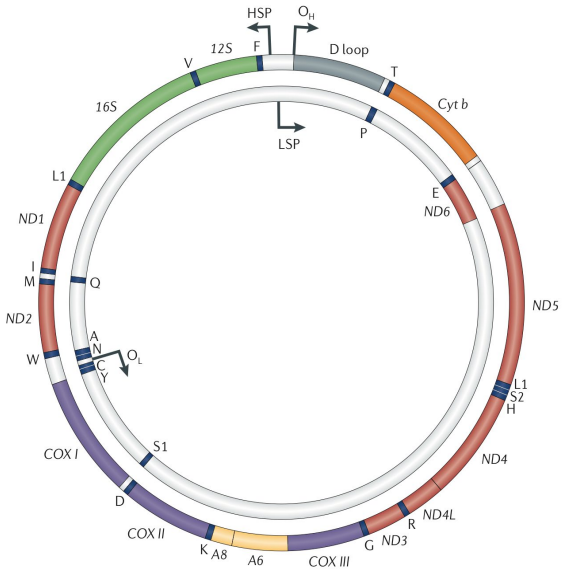


Figure 4.2 The human mitochondrial genome.

The mitochondrial genome human is a circular, double-stranded DNA and encodes for 37 genes, including 13 subunits of the mitochondrial respiratory complexes I-IV within the OXPHOS system (7 subunits of complex I (in red), 1 subunit of complex III (in orange), 3 subnits of complex IV(in purple), 2 subunits of complex V (in yellow)), 2 rRNAs (12S and 16S, in green), and 22 tRNAs. The outer circle represents the heavy strand and the inner one represents the the light strand. The diagram also shows the origins of replication of both the heavy (O_H) and the light (O_L) strands. Taken from (Schon et al., 2012), under the terms and conditions of the Creative Commons by Attribution (CC-BY) License.

The mitochondrial genome exists in multiple copies within mitochondria. Highly compacted mtDNA-protein complexes called nucleoids anchor it to the inner mitochondrial membrane (Kukat & Larsson, 2013). One single cell can contain up to thousands of nucleoids, each with a copy of mtDNA. Apart from the copy number, the mode of transmission of mtDNA is also drastically different from nuclear DNA. mtDNA is inherited maternally in human, as the paternal mtDNA is disintegrated soon after fertilization (al Rawi et al., 2011; Sato & Sato, 2011). As the segregation occurs rapidly from one generation to the next, a 'bottleneck' exists so that only a small amount of mtDNA is passed to the progeny and hence potentially eliminating defective mitochondria and mtDNA (Lieber et al., 2019; Pepling et al., 2007). The replication and segregation of mtDNA also occurs for only part of the nucleoids in a cell at a given time (Meeusen & Nunnari, 2003). Unlike the nuclear genome, the process is not strictly coupled to the cell cycle.

The multi-copy nature of mtDNA in human also gives rise to its unique feature of heteroplasmy. Heteroplasmy is referred to the presence of two or more mitochondrial genotypes in a cell. In contrast, homoplasmy is a state where the copies of all mtDNA are identical. Heteroplasmy occurs in cells of most of the outbred populations. The ratio of wild-type and mutated mitochondrial genomes usually varies in different cells and tissues. This

partly explains the high heterogeneity in phenotypes of mitochondrial diseases that are caused by mutation in mtDNA and complicates the analysis of mitochondrial genetics (N. Lane, 2011).

4.1.3 The origin of mitochondria and the cross-talk with the nuclear genome

The widely accepted theory of the origin of mitochondria is that, originally being a free living α-proteobacterium, they were engulfed and incorporated inside a precursor of modern eukaryotic cells two billion years ago (N. Lane & Martin, 2010). This theory is supported by evidence from phylogenetic analysis of ribosomal RNA (Gray, 1999; Sicheritz-Pontén et al., 1998). This endosymbiotic relationship allowed the early eukaryotes to perform aerobic respiration, hence harvesting energy from organic matter in a more efficient manner by consuming oxygen, originally a toxic substance for the archaeal cells. The new ability was a driving force for the evolution into multicellular life. During the process of evolution, the majority of the genomic material of the protomitochondrion was transferred from the mitochondria to the nucleus of the host and subsequently lost due to redundancy (Gabaldón & Huynen, 2004). Studies suggested that only 22% of mitochondrial proteins in human are descendants from the protomitochondrial ancestor (Gabaldon, 2003).

The nuclear DNA now encodes about 1200 proteins that contribute to the mitochondrial proteome. These proteins are required for mitochondrial functions, maintenance and mtDNA replication, thus for processes such as the Tricarboxylic acid (TCA) cycle, amino acid-, nucleic acid- and lipid biosynthesis, mitochondrial transcription, mtDNA replication or translation. These proteins are translated in the cytosol, and subsequently sorted and imported into mitochondria through the TIM/TOM complex to their target locations such as the outer membrane, inner membrane, intermembrane space or the mitochondrial matrix (Hensen et al., 2014; Mokranjac & Neupert, 2005).

The proper function of mitochondria, and hence the survival of the cell, are heavily dependent on the communication between mitochondria and the nuclear genome. For example, mitochondria constantly update the cell concerning its bioenergetic status by sending mitochondrial stress signals whenever bioenergetic demands cannot be met due to OXPHOS dysfunction or defective mitochondria (Cagin & Enriquez, 2015). This is referred to as retrograde signaling and include signals such as ATP/AMP, NADH/NAD+, cytosolic Ca^{2+}, or reduction in the mitochondrial membrane potential (A. W. E. Jones et al., 2012) . The nucleus responds to the signals by regulating the expression of genes that promote mtDNA repair, mitochondrial transcription and biogenesis to meet the energy

requirements of the cell. The cell might also respond by switching to a glycolytic metabolism to adapt to decreased energy production from mitochondria (Freije et al., 2012).

4.1.4 Mitochondrial functions

Mitochondria are often referred to as the powerhouses of the cell because of their ability to generate cellular energy in the form of ATP, a form that is usable by cells. In fact, they are also involved in many other processes that are crucial for the functioning and survival of cells. This is accomplished by the expression of around 1200 genes in both the mitochondrial and nuclear genomes that encode the mitochondrial proteome. Previous work attempted to compile a list of mitochondria-associated genes/proteins in different animals with approaches such as mass spectrometry (Gaucher et al., 2004; Taylor et al., 2003) and computation or machine learning (C. Guda et al., 2004), grouping them according to their functions (Calvo et al., 2016). Mitochondrial functions cover 6 main areas: Bioenergetic Pathways, ROS Defense, Apoptosis, Cell Signaling, Biosynthetic Pathways/Homeostasis and Maintenance of Mitochondria. This section aims to discuss these functions according to available literature. The list of associated genes in human are documented at Appendix I.

4.1.4.1 Bioenergetic pathways

Glycolysis and Fructose Metabolism

Glucose is a major source of energy for cells. It is broken down to pyruvate by the process of glycolysis, which is then further broken down to release energy in the form of ATP through the tricarboxylic acid (TCA) cycle and oxidative phosphorylation (OXPHOS) in mitochondria. Glycolysis also produces a net of 2 ATP per molecule of glucose during its breakdown. Although the efficiency is much lower than oxidative phosphorylation (36 ATP per glucose molecule), the speed is much faster (Lunt & vander Heiden, 2011).

Glycolysis occurs solely in the cytoplasm, yet it is part of the bioenergetic pathway and is closely linked to the TCA cycle. It is also affected in case of mitochondrial dysfunction. Therefore, genes related to glycolysis are considered to be associated to mitochondrial functions as well. Fructose has also been a part of human diet and often occurs with glucose in fruits in high concentration (Sun & Empie, 2012). Though the two sugars have different metabolic pathways (Mayes, 1993), fructose can also be utilized in oxidative phosphorylation after being converted and broken down into pyruvate.

Pyruvate Metabolism

Pyruvate, being the product of Glycolysis, is taken up by mitochondria to be broken down through the TCA cycle. Pyruvate is transported across the mitochondrial membranes by transmembrane carrier proteins (Bricker et al., 2012). It is then converted to acetyl-CoA by pyruvate dehydrogenase complexes, an irreversible process of oxidative decarboxylation, before entering the TCA cycle. These complexes and their inhibitors therefore regulate metabolite flux through the TCA cycle and hence the homeostasis of glucose metabolism.

Tricarboxylic Acid Cycle (TCA cycle)

Tricarboxylic Acid (TCA) Cycle, also known as Krebs cycle or citric acid cycle, happens within the mitochondrial matrix. It releases energy from acetyl-CoA through a series of redox reactions. The process oxidizes acetyl-CoA into CO2 and gives rise to GTP (guanosine triphosphate). It also, by the transfer of high energy electrons, converts NAD+ and FAD into NADH and FADH respectively, which are then fed to the ETC to generate ATP through OXPHOS.

Oxidative Phosphorylation (OXPHOS)

The electron transport chain (ETC) is located at the inner mitochondrial membrane and comprises of a series of complexes encoded by both the mitochondrial and nuclear genome. It is the mitochondrial enzyme cascade responsible for oxidative phosphorylation (OXPHOS) for energy production. Through the transfer of electrons from one complex to the next (complex I-IV, cytochrome C and ubiquinones), protons (H+ ions) are pumped out of the matrix to the intermembrane space, creating an electrochemical gradient across the inner membrane. Energy from this gradient is then harvested to convert ADP to ATP by ATP synthase (complex V), where energy is stored chemically at the phosphate bond. This process is termed oxidative phosphorylation (OXPHOS) and provides 90% of ATP in a cell (Davis & Williams, 2012).

The electron donors are NADH and FADH2 produced from the TCA cycle and other metabolic processes. They are oxidized by respective dehydrogenases and their electrons are passed through the ETC to oxygen and reduced to water. The ETC regulates itself by responding to negative feedbacks of ATP (Dagda et al., 2009). This helps maintaining the homeostasis of the bioenergetic status of the cell.

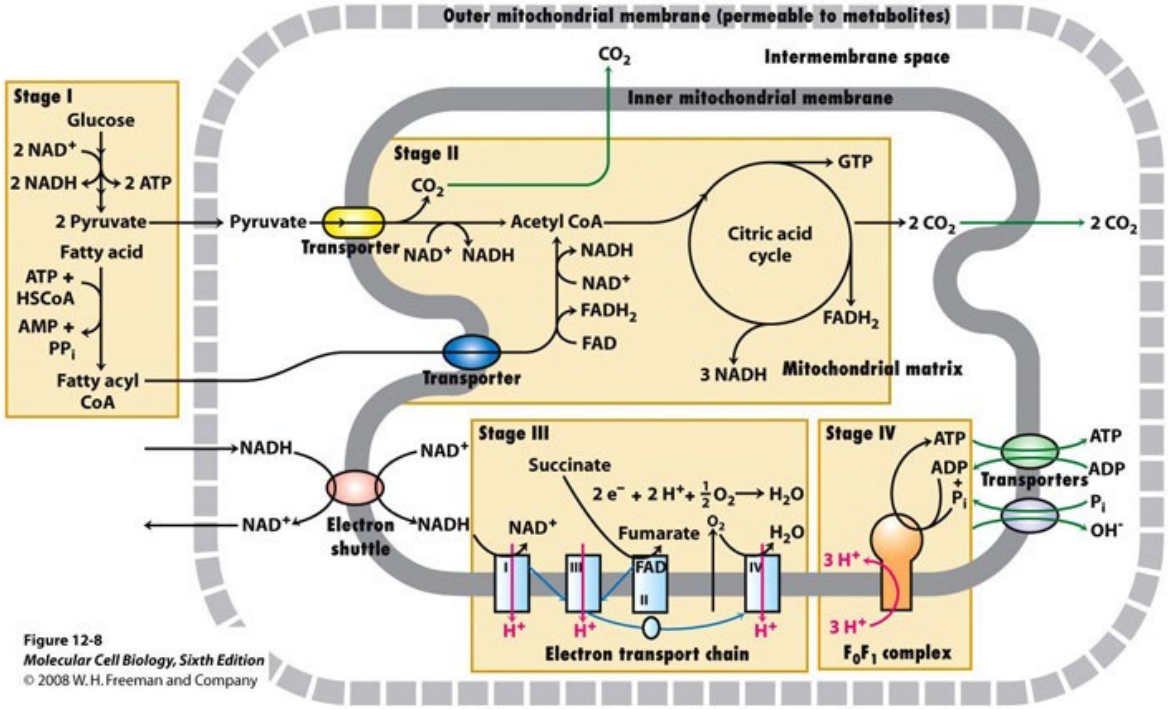


Figure 4.3 Bioenergetic pathways in mitochondria.

Mitochondria generate cellular energy in the form of ATP via a couple of stages. It starts with glycolysis (Stage I), where glucose is broken down into pyruvate. This pyruvate is then taken up by mitochondria and converted to acetyl-CoA before being further broken down through the TCA cycle (Stage II). The NADH and FADH2 produced from the TCA cycle serve as electron donors and the ETC where an electrochemical gradient is built up across the inner membrane (Stage III). Finally energy from this gradient is harvested to convert ADP to the final product. Taken from (Lodish et al., 2008), under the terms and conditions of the Creative Commons by Attribution (CC-BY) License.

Fatty Acid Degradation and Beta-oxidation

Fatty acid is another source of energy apart from glucose and has an important role in energy homeostasis. When glucose and other sources are not adequate to sustain the energy requirement of the body, fatty acids are released from adipose tissue. Beta-oxidation is the major pathway for the degradation of fatty acids and takes place within the mitochondrial matrix. Fatty acids are activated by attaching to an acyl-CoA and carried into mitochondria where they are broken down through oxidation. Each beta-oxidation cycle release two carbon units from the fatty acid chain and gives rise to acetyl-CoA that enters the TCA cycle. The process at the same time reduces NAD+ and FAD into NADH and FADH respectively, which can be fed to the ETC for OXPHOS.

4.1.4.2 ROS Defense

Reactive oxygen species (ROS) are byproducts of OXPHOS, due to the leakage of electrons at complex I and III at the ETC. These electrons leads to partial reduction of oxygen to form free-radical superoxide. Though ROS could act as signaling molecules, they cause oxidative damage to the mitochondrial genome at high concentration due to the

proximity of mtDNA to the ETC. This makes mtDNA susceptible to mutations and might contribute to disease and aging (Kujoth et al., 2005; Wallace, 2005). Lipids and proteins could also suffer from damages induced by ROS (Y. Chen et al., 2016; Holzerová & Prokisch, 2015). Mitochondria therefore have an antioxidant defense system to eliminate ROS. For example, superoxide dismutase decomposes superoxides into hydrogen peroxide, which can be further degraded into water and oxygen by peroxiredoxins.

4.1.4.3 Apoptosis

Apoptosis is a type of programmed cell death, which triggers morphological and biochemical changes in a cell, such as cell shrinkage, DNA cleavage, chromatin condensation, and cell fragmentation. It is part of the defense mechanism of multicellular organism to remove damaged or dangerous cells, and to regulate cell number. There are two main mechanisms of apoptosis: The intrinsic and the extrinsic pathway. Mitochondria are the central executioners of the intrinsic pathway (Kaufmann & Earnshaw, 2000).

In response to cytotoxic agents like radiation, nitrogen monoxide, mercury or oxidative stress due to high levels of ROS (Soga et al., 2012), apoptosis is induced in mitochondria and pores are formed in mitochondrial membranes. This results in the disruption of mitochondrial transmembrane potential and hence the release of pro-apoptotic proteins. One of the most important factors is cytochrome C (Verhagen et al., 2000), which activates the caspase pathway and eventually leads to programmed cell death.

The process is tightly regulated as it is unstoppable once it has begun (Böhm & Schild, 2003). For example, the Bcl-2 family proteins can decide the fate of the cell by controlling the formation of pores (Cory & Adams, 2002). The improper activation of apoptosis causes atrophy and diseases such as ischemic strokes and neurodegenerative diseases; whereas the lack of apoptosis can contribute to autoimmune diseases and is linked to oncogenesis (Fuchs & Steller, 2011).

4.1.4.4 Cell Signaling

Mitochondrial and Calcium Signaling

Mitochondria are heavily involved in cell signaling circuitry. They do not only respond to stimuli but also act as physical platforms for protein-protein signaling to take place, and regulate the level of signaling molecules such as calcium ions (Ca^{2+}) and ROS. Mitochondria therefore serve as the effector, transducer and initiator in multiple signaling pathways and have been implicated in processes including hypoxic stress response, differentiation, growth factor signaling, immune response and apoptosis as mentioned

above (Antico Arciuch et al., 2012; Chandel, 2010; Finkel, 2011; Gunter et al., 2004; Kawai et al., 2005).

Calcium ions (Ca^{2+}) are among the most important intracellular signaling molecules. They can stimulate ATP production by regulating the activity of components of the TCA cycle, trigger cell death by necrosis or apoptosis, and have an important role in the control of autophagy (Brini et al., 1999; Cárdenas et al., 2010; Szalai et al., 1999; Visch et al., 2004). For instance, as a response to increased bioenergetic demands, Ca^{2+} could bind to and activate dehydrogenases that increase the availability of NADH and hence the supply of electrons to the ETC. Thus, the level of Ca^{2+} in cells has to be carefully regulated and its homeostasis is maintained by mitochondria.

Mitochondria serve as buffers for Ca^{2+} by, for example, regulating their calcium channels to take in the Ca^{2+} released from the endoplasmic reticulum (ER), the reservoir of Ca^{2+}; or releasing Ca^{2+} from the mitochondrial matrix (Jouaville et al., 1995; Mitchell & Moyle, 1967). The uptake of mitochondrial Ca^{2+} is through the voltage-dependent anion channels (VDACs) on the outer membrane and the mitochondrial Ca^{2+} uniproter (MCU) located at the impermeable inner membrane. The VDACs are able to establish a close interaction with the ER to ensure rapid transfer of Ca^{2+}. The MCU then transports Ca^{2+} into the matrix across the electrochemical gradient.

4.1.4.5 Biosynthetic pathways/Homeostasis

Fatty Acid Metabolism, Biosynthesis and Elongation / Lipoic Acid Metabolism

Fatty acids can be synthesized from substrates of the TCA cycle in a reversed manner of beta-oxidation. This is to maintain the homeostasis of fatty acids in cells and store the extra energy. Fatty acid synthesis (FAS) occurs mostly in the cytosol but also in mitochondria (mtFAS) through acyl carrier protein (ACP) and a number of reductases or transferases. During each cycle of elongation, the fatty acyl chain attached on ACP is extended by two carbons through a condensation reaction, followed by a few reduction and dehydration reactions that reduce the acyl chain to a saturated fatty acid (Brody et al., 1997; Mikolajczyk & Brody, 1990). The product of mtFAS is octanoate, an eight-carbon saturated fatty acid that could be convert to lipoic acid. Lipoic acid is an important cofactor required for catalysis of a number of mitochondrial 2-ketoacid dehydrogenase complexes. It is also responsible for the regulation and stabilization of these multienzyme complexes and therefore is critical for normal mitochondrial activity (Nowinski et al., 2020; Wada et al., 1997).

Metabolism of Lipids and Lipoproteins / Cardiolipin Biosynthesis / Bile Acid Synthesis

Lipids are basic blocks of cellular and mitochondrial membranes. The control of lipid synthesis is therefore crucial for the integrity and function of both the cell and mitochondria, the latter of which are constantly reshaping by fission and fusion processes. Mitochondria are able to synthesize some of the lipids, such as phosphatidylglycerol, cardiolipin, and control their distribution (Daum & Vance, 1997; Mannella et al., 1998). Cardiolipin is an essential component of mitochondrial membranes and is involved in processes such as respiration and energy conversion. It also serves a signaling platform as it could recruit and interact with molecules engaged in process such as mitophagy (W. Huang et al., 2012) and apoptosis (Lovell et al., 2008; Sorice et al., 2004). Mitochondria are also involved in the metabolism of other important lipids, lipoproteins and their derivatives include cholesterol and bile acids. For example, cholesterol is converted to steroid precursors after being imported into mitochondria, to allow further processing at the endoplasmic reticulum (ER) (Soccio & Breslow, 2004). Bile acid is normally produced at the liver, but could also be synthesize by the “acidic” pathway inside mitochondria (Pandak et al., 2002).

Fe-S cluster biosynthesis / Heme Biosynthesis

Iron-sulfur (Fe-S) clusters and heme are both iron-containing cofactors that have crucial roles in the maintenance of mitochondria. They are components of the complexes in the ETC (Steffens et al., 1987; Tyler, 1992) and catalyze electron transfer during the formation of ATP (Beinert et al., 1997). These prosthetic groups also have diverse roles in signaling, metabolism and the defense system against oxidative damages (Beinert & Kiley, 1999; Porello et al., 1998; Ryter & Tyrrell, 2000), and are synthesized and utilized in mitochondria.

Protein Stability and Degradation / Amino Acid Metabolism

Mitochondrial proteins are under the exposure of oxidative stress due to the production of ROS during oxidative phosphorylation. Failure to control protein quality or remove misfolded protein could lead to mitochondrial dysfunction. Thus mitochondria have a system, consisting of proteases, that recognizes and degrades unwanted proteins (Livnat-Levanon & Glickman, 2011; Martinelli & Rugarli, 2010). Mitochondria are also a platform for metabolism of amino acids, the building blocks of proteins. The TCA cycle provides as well as consumes amino acids; and the metabolic pathways of 17 out of 20 amino acids utilize mitochondrial enzymes (P. Guda et al., 2007).

Nucleotide Metabolism

The balance of cellular nucleotide pool is important for genome replication and repair. Mitochondrial deoxynucleoside triphosphate (dNTP) pools are separated from the cytosolic pool because of its membrane system (L. Wang, 2016). The dNTPs are either imported from the cytosol or synthesized with substrates from metabolic processes within mitochondria (A. N. Lane & Fan, 2015). One of the biosynthetic steps for pyrimidine occurs in mitochondria, where dihydroorotate is converted into orotate via orotate dehydrogenase reaction, catalyzed by the mitochondrial protein dihydroorotate dehydrogenase (DHODH) (Evans & Guy, 2004). Whereas for purine, the entire synthesis process is cytoplasmic. However, studies showed that the glycine precursor for purine synthesis could also be synthesized inside mitochondria with its glycine cleavage system (Lewis et al., 2014).

Nitrogen Metabolism / Urea Cycle

Nitrogen is assimilated into the body in proteins of the human diet. The excess amino acids in the body are deaminated to form ammonia or ammonium ions (NH^{4+}). Since these ions are toxic, they are converted to urea in the urea cycle and then secreted. Part of the process takes place in mitochondria where three of the urea cycle enzymes are located (Adeva et al., 2012). Together these enzymes convert ammonia and ornithine to citrulline, which is then transported to the cytoplasm to complete the urea cycle. Transporters for the substrates (ornithine and citrulline) that reside at the mitochondrial membrane are also required for the normal function of the urea cycle (Brusilow & Horwich, 2001).

Metabolism of Vitamins and Co-Factors / Folate and Pterine Metabolism

Vitamins are organic molecules that are essential for various biochemical functions within the human body, ranging from growth and development (Vitamin A), regulation of cell metabolism (B vitamins) to maintaining the immune system (Vitamin C). Mitochondria have a role in the metabolism of some of these vitamins as they are also crucial for mitochondrial functions. For example, retinoids (a class of vitamin A) is a substrate to a dehydrogenase located at inner mitochondrial membrane and might have a protective effect against oxidative stress in mitochondria (Belyaeva et al., 2008). The metabolism of folate (Pterine derivates, Vitamin B9) also occurs in mitochondria, where its reduced form tetrahydrofolate (THF) serves as single carbon unit, which is important for the maintenance of redox and regulation of methylation status (Desai et al., 2016). Co-factors, like vitamins, are crucial for both cell and mitochondrial functions as they are involved in diverse biochemical reactions. One of them is nicotinamide adenine dinucleotide (NAD), which is involved in the redox reactions at the ETC. Therefore its metabolism is carefully regulated in mitochondria to maintain the balance of the ratio of NAD vs NADH (the reduced form of NAD) (Tzameli, 2012).

4.1.4.6 Mitochondrial maintenance

Mitochondrial Dynamics

The mitochondrial network is constantly reshaping through cycles of fusion and fission events in order to meet bioenergetic demands and control the quality of mitochondria. Fusion of mitochondria leads to long, networked filaments that is the predominant mitochondrial form in some cells; faulty or mutated mtDNA copies can on the other hand be isolated by fission and subsequently eliminated through mitophagy (Carelli et al., 2015). Mitophagy is initiated by mitochondrial dysfunction due to, for example, hypoxia through the PTEN-induced putative kinase 1 (PINK1)/Parkin pathway. Fusion is achieved by several GTPases: the mitofusins (Mfn) and optic atrophy-1 (Opa1) that fuse the outer and inner membrane, respectively. Fission results from the recruitment of dynamin-related protein 1 (Drp1), that constricts both the outer and inner membrane, and is regulated by kinases responding to stress conditions and distinct cell cycle phases (Busch et al., 2014; Mishra & Chan, 2016).

Replication and Transcription

The mtDNA is a closed-circular genome with two strands of DNA, the heavy strand (H-strand) and the light strand (L-strand). A proportion of the mtDNA contains a three-stranded structure called the displacement loop (D-loop), where a third DNA strand displaces the H-strand. Due to its unique structure, the replication mechanism is distinct from that of its nuclear counterpart. A widely accepted model is that the replication starts with the leading H-strand and advances around two-thirds of the mtDNA before the replication fork reaches the replication origin of the lagging L-strand and initiates its synthesis (Clayton, 1982). Transcription of mtDNA is initiated at two sites located at the D-loop. The site encompassed by the H-strand promoter (HSP) directs the transcription of H-strand, and the other one encompassed by the L-strand promoter (LSP) directs the transcription of the L-strand. The regulation of both replication and transcription of the mitochondrial genome is complex and key enzymes involved in these processes are unique to mitochondria (Taanman, 1999). During transcription initialization, mitochondrial transcription factor A (TFAM) recognizes these binding sites and induces sharp U-turn bends in mtDNA (Kukat & Larsson, 2013). Such conformational changes facilitate the access and binding of mitochondrial RNA polymerase (POLRMT) to mtDNA and allow it to orchestrate the process (Fisher & Clayton, 1988). On the other hand, TFAM is an important factor for packaging mtRNA into compact nucleoid, also via the induction of U-turn bends (Alam et al., 2003), which make it less permissive for replication and transcription. Therefore, TFAM may also operate as an

epigenetic regulator over transcription and replication by controlling the number of mtDNA molecules available for the process (Farge & Falkenberg, 2019; Gilkerson et al., 2013).

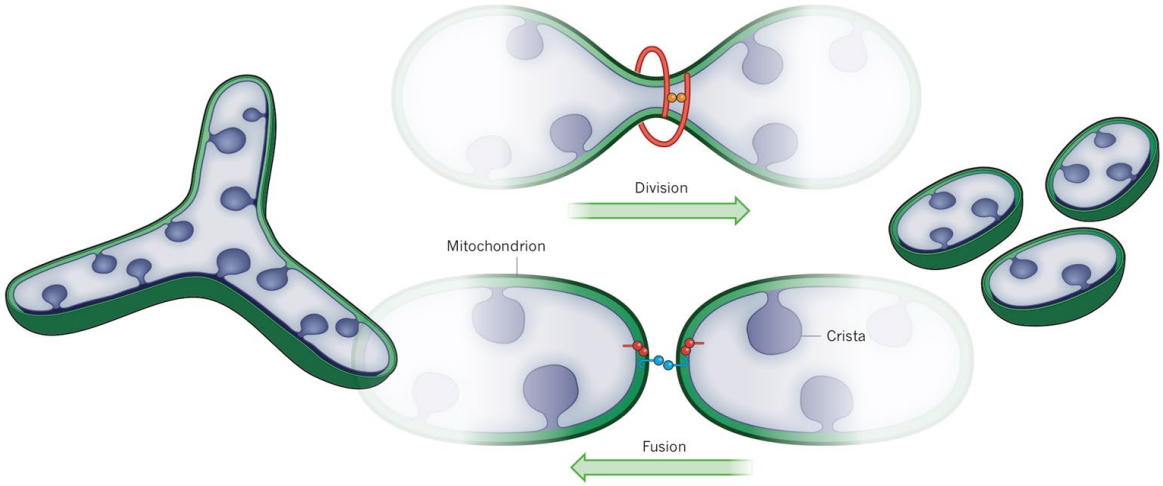


Figure 4.4 Organization of mitochondria.

Mitochondrial networks are highly dynamic and are constantly changing by the combined actions of mitochondrial division and fusion. During division, Dynamin-related protein (DRP) is recruited on the outside of the organelle to form a helical structure, which then constricts the membranes and mediates a scission. Mitochondrial fusion is achieved through interactions of several GTPases (mitofusins and optic atrophy-1) that fuse the outer and inner. Taken from (Friedman & Nunnari, 2014), under the terms and conditions of the Creative Commons by Attribution (CC-BY) License.

Translation

The translation machinery of the mitochondrial genome is also different from the nuclear genome. Mitochondria contain their own ribosomes (mitoribosomes) with different physical and chemical properties. rRNAs within mitoribosome are encoded by the mtDNA (Attardi & Ojala, 1971; Brega & Vesco, 1971). The mitochondrial mRNAs, unlike the cytosolic ones, have no upstream leader sequences that help with ribosome binding, making the translation process distinct from that of the nuclear genome (Cantatore et al., 1987). Structural studies have contributed significantly to our understanding of the structure and function of the mitoribosomes and how they recruit mitochondrial mRNAs (Sharma et al., 2003).

Import and Sorting / Mitochondrial Carrier

Mitochondria import 99% of their proteins from the cytoplasm to maintain their functions. These proteins must be correctly identified by mitochondria to be imported and delivered to target locations. Mito-proteins imported into the cellular matrix usually have an extension at the N-terminus that serves as a targeting signal, which can be recognized by the receptors of TOM (translocase of the outer membrane) complexes (Kutik et al., 2008). Once inside the intermembrane space, proteins are directed to their destinations through sorting

pathways. The TIM (translocase of the inner membrane) complexes, for example, transfer proteins into the matrix (Chacinska et al., 2009). Other complexes recognize different mitochondrial targeting signals and direct proteins to either the intermembrane space or the outer membrane (Hoppins & Nargang, 2004; Paschen et al., 2003).

4.1.5 Mitochondrial diseases

Mitochondria have diverse roles in cell metabolism and homeostasis. Dysfunction of mitochondria thus has a huge impact on the function and survival of cells, and can lead to a number of pathological disorders. Mitochondrial diseases can be caused by mutations in mtDNA or nuclear DNA that encodes for the mitochondrial proteome, or other factors that disrupt mitochondrial functions. They are complex and cover almost every field of medicine (Ylikallio & Suomalainen, 2012), such as neurodegenerative diseases, metabolic disorders, obesity and cancer.

4.1.5.1 Mitochondrial diseases associated with point mutations in mtDNA

MtDNA encodes for 13 polypeptides that are components of the OXPHOS system, as well as 22 tRNAs and two rRNAs that are essential in the synthesis of these proteins. Mutations in these genes might disturb the energy harvesting process, leading to the failure of mitochondria to meet cellular energy demand. Thus far, over 300 pathogenic mtDNA mutations have been identified (Kogelnik et al., 1998). Disorders include: Leber hereditary optic neuropathy (LHON) (mutation in Complex I gene) (Johns et al., 1992), Leigh syndrome (Complex V) (Shoffner et al., 1992) and Mitochondrial encephalopathy (MELAS) (tRNAs) (Goto et al., 1990).

As multiple copies of mtDNA exist in a single cell, heteroplasmy is often observed and the ratio of mutated/wild-type mtDNA varies among cell types and tissues. This complicates the interpretation of mitochondrial diseases and the underlying genetics (N. Lane, 2011).Phenotypic variability also arises from this condition. The wild-type mtDNA can often complement the mutated ones as mutations are usually recessive. It appears that the mutated/wild-type ratio has to reach a certain threshold of around 50 - 60%, depending on the mutations and tissue type, before the phenotypes of a disease can be observed (Hayashi et al., 1991; Mita et al., 1990; Moraes et al., 1992; Shoubridge, 1994).

4.1.5.2 Aneuploidy

Aneuploidy is the status where there is an abnormal number of chromosomes in a cell, which is fatal for human embryos in most of the cases. Trisomy 21 (three copies of chromosome 21), which results in Down syndrome (DS), is the most common type of

aneuploidy that infants can survive with; however, they suffer from many pathological conditions, including neurological deficiencies and immune disorders. Some studies suggest that the symptoms of DS are associated with oxidative stress (Anneren & Epstein, 1987; Brás et al., 1989; Brooksbank & Balazs, 1984; Busciglio & Yankner, 1995; Jovanovic et al., 1998). It was found that ROS levels in the neurons of DS patients increased 3- to 4-fold (Busciglio & Yankner, 1995), leading to elevated lipid peroxidation and possibly neuronal death.

One of the mainstream views is that the observed pathologies result from the increased dosage effect due to the presence of an extra chromosome 21 (Hsa21), which increases transcript and protein levels of the coding genes on Hsa21, and at the same time alters the expression of non-Hsa21 genes through actions of transcription factors or chromatin modifiers (Antonarakis, 2017). It has been suggested that the overexpression of mitochondrial genes located at chromosome 21, such as superoxide dismutase (SOD-1) is related to the oxidative stress in DS (Sinet, 1982), though the results from different studies remain controversial (Anneren & Epstein, 1987; de La Torre et al., 1996; Shapiro, 1999). It has also been reported that mitochondrial dysfunction is observed in DS patients, with a lack of certain mitochondrial enzymes (Prince et al., 1994); furthermore, morphological abnormalities of mitochondria have been identified in a mouse model of DS, together with an impaired energy metabolism (Bersu et al., 1998).

4.1.5.3 Cancer

The alteration of mitochondrial functions in cancer was first observed by Warburg 90 years ago (Warburg et al., 1927), namely that tumors fermented glucose to produce excess lactate even in the presence of oxygen (aerobic glycolysis). This phenomenon was termed Warburg effect in the 1970s (Racker, 1972) and was also observed in proliferating and developing cells, where the rate of glucose uptake and lactate production increase significantly. Aerobic glycolysis is a relatively inefficient, but a much faster way to generate ATP, compared to mitochondrial respiration (Heiden et al., 2009). This might give cells a selective advantage in an environment with limited resources (Pfeiffer et al., 2001), such as in tumor microenvironments where there are limited glucose and nutrients. The production and secretion of lactate also decreases the pH in the microenvironment and could enhance invasiveness of tumor cells (Estrella et al., 2013). It has also been proposed that Warburg effect is an adaptation mechanism to sustain biosynthesis under uncontrolled proliferation. The excess carbon from aerobic glycolysis can be diverted for the generation of molecules such as lipids and nucleotides (Levine & Puzio-Kuter, 2010). On the other hand, the altered metabolism can possibly activate stress response pathways that promote tumorigenesis. Changes in mitochondrial redox potential alter the generation of reactive

oxygen species (ROS) (Locasale & Cantley, 2011), which is involved in signaling processes related to cell proliferation.

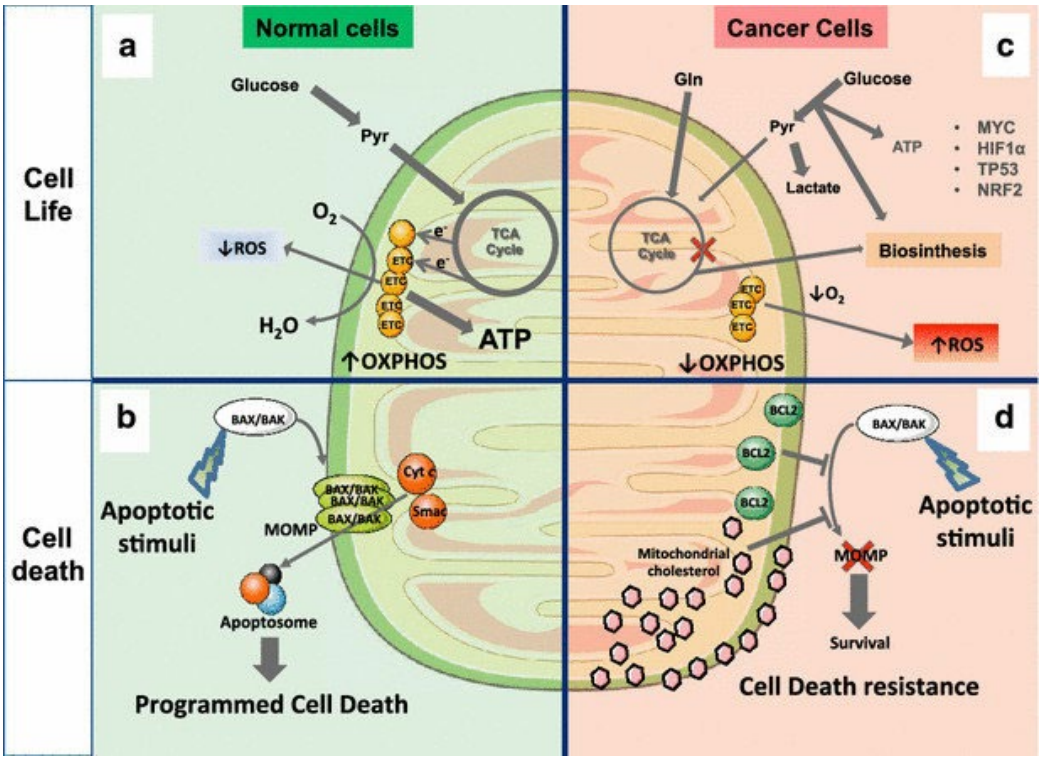


Figure 4.5 Cancer cell metabolism

(a) In normal cells, the pyruvate results from the glycolysis pathway goes into the mitochondrial TCA cycle, which produces reduced equivalents that could be fed into the ETC for the generation of ATP. (b) Upon apoptotic stimuli, BAK/ BAX and mitochondrial outer membrane permeabilization (MOMP) are activated, followed by the release of cytochrome c and the formation of the apoptosome, which results in apoptotic cell death. (c) Cancer cells exhibit an altered metabolism, where the glucose consumption rate is much higher so that the excess carbon could be diverted for biosynthesis to support the growth of the cells. ATP is generated mostly through the fermentation of pyruvate instead of the TCA cycle. (d) Changes in gene expression in cancer cells (e.g. overexpression of anti-apoptotic proteins or inactivation of pro-apoptotic proteins), could possibly counteract the action of BAX/BAK and avoid MOMP formation, hence resisting apoptosis and increasing the survival of the tumor. Taken from (Ribas et al., 2016), under the terms and conditions of the Creative Commons by Attribution (CC-BY) License.

The change in mitochondrial metabolism in cancer seems to be caused by certain mtDNA and nuclear DNA mutations, especially those in key enzymes involved in the TCA cycle. This is observed in paragangliomas, pheochromocytomas, myomas, and gliomas (Baysal et al., 2000; Tomlinson et al., 2002; Yan et al., 2009). The activities of transcription factors, such as HIF1α , FOS or JUN (Abate et al., 1990; Kurelac et al., 2011), can also be changed due to increased ROS production from altered mitochondrial metabolism, as well as mutations of mitochondrial enzymes such as succinate or fumarate dehydrogenases, as their dysfunction results in the accumulation of succinate and fumarate, which helps the activation of HIF (King et al., 2006). This change in activities alters gene expression and promotes cell proliferation. Despite the increased ROS levels, apoptosis can be limited in

cancer cells by resisting mitochondrial outer membrane permeabilization (MOMP) through impairing BAK/BAX oligomerization, thus increasing the survival of cancer cells (Green & Kroemer, 2004; Martinez-Caballero et al., 2009).

4.2 Understanding human biological systems with new technologies

The recent technological advancement in biology, especially in the field of molecular biology and genetics, allows us to have a much deeper understanding in biological systems. Next-generation sequencing (NGS), for example, is able to sequence genomes or transcriptomes in unprecedented speed with much lower costs, thus has completely revolutionized research in genomics and transcriptomics. This brings about new insights into the functions and regulation of cellular system, and the etiology or pathology of human diseases, including mitochondrial diseases. In this section, the technologies in different fields that are applied to studying and investigating biological systems will be discussed.

4.2.1 Genomics and Transcriptomics

Information of cells is stored in DNA molecules, that give cells instructions to carry out their activities. In eukaryotes, this includes both nuclear DNA and mitochondrial DNA (mtDNA), and the complete set of these DNA molecules forms the genome. The coding regions of DNA, or genes, are transcribed to RNA molecules, which are translated to protein. The sum of all RNA transcripts at a given time is referred to as the transcriptome. Studying the genome and transcriptome is therefore essential to understanding cell functions.

The sequencing of the entire human genome for the first time using the Sanger chain termination sequencing method, initiated by the Human Genome Project, was finished in 2003, took 13 years and costed 2.7 billion USD. Today with the advent of NGS methods, the sequencing of the genome (DNA-seq) has broken the 1000 USD barrier and takes less than a day. NGS platforms of different sequencing approaches, possible sequence length, error rate, and speed are available nowadays (reviewed in (Mardis, 2013) and (Goodwin et al., 2016)), with the Illumina system being one of the most popular sequencers. The Illumina system adopts a short-read sequencing approach, in which the DNAs or cDNAs are fragmented and ligated to universal adapters for amplification, and then sequenced by synthesis (Mardis, 2013). NGS sequencing can also be applied to quantifying transcriptomes (RNA-seq), in which RNAs are reverse-transcribed to cDNAs before being sequenced. Measuring transcriptome was mostly achieved by microarray previously. Though still used today, the detection of RNA with microarray has limited dynamic range

and is restricted by the availability of known RNA molecules that could be hybridized on arrays.

The sequencing of genome and transcriptome is followed by bioinformatics analysis. Quality control is usually the first step, that serves to identify and remove low quality data (reads), before the reads are aligned (or mapped) to a reference genome. A number of mappers with different algorithms, speed, features, and accuracy under different conditions (such as number of mismatches) are available (reviewed in (Fonseca et al., 2012; Hatem et al., 2013)) and users should choose according to their needs. One of the most widely used aligners for (short read) DNA-seq is BWA-aligner, which has a better performance overall compared to others even with longer reads (Hatem et al., 2013). Mutations and structural variants can further be revealed by variant callers. Again, numerous callers serving different functions with different sensitivity levels exist (reviewed in (Roberts et al., 2013; Q. Wang et al., 2013)).

Mappers for RNA-seq data usually use a different algorithm from the ones for DNA-seq, as they have to account for splicing events. Among various mappers, STAR (Spliced Transcripts Alignment to a Reference) is a very popular tool because of its relatively high accuracy and speed (Dobin et al., 2013). In addition to the identification of somatic mutations like DNA-seq data, RNA-seq data can also be used to quantify gene expression, as the number of reads that align to a transcript is proportional to its abundance in cells. After normalization to the length of genes and then the total number of mapped reads, the numbers of reads for each transcript of gene are converted to TPM (Transcript per million). Since the sums of all TPMs in each sample from the same experiment are theoretically the same, this allows the direct comparison of the proportion of reads mapped to a gene in different samples. Statistical tests and tools have also developed to account for variability between biological replicates. For example, DESeq2 (Love et al., 2014) pools information regarding variance across different genes, in order to detect true differentially expressed genes (i.e. the difference in read counts is not due to random variation), even under the condition where there are limited samples.

4.2.1.1 Mitochondrial genome and transcriptome

The complete mitochondrial genome was first sequenced and annotated as early as 1981 (S. Anderson et al., 1981). A slightly different approach has to be adopted for the study of mitochondrial genome and transcriptome due to the status of heteroplasmy and high copy numbers of mtDNA in cells. For example, variant calling has to be done in a conservative setting (i.e. high threshold for believing in calls), and separately for the plus and minus strand. The transcriptome also has to be analyzed separately from the nuclear

transcriptome with adjusted parameters, as they are sometimes discarded for having too many mapped reads.

4.2.2 Proteomics

Protein molecules, depending on their amino acid sequence, have diverse biochemical properties as well as secondary, tertiary and quaternary structures. They perform myriads of functions in our cells, acting as enzymes, structural and signaling molecules. Proteins are translated from mRNA. The study of protein, or proteomics, is usually the next step after genomics and transcriptomics to provide additional insights into the biological systems. The study of the proteome can be complicated because the proteome, unlike the genome, varies from cell to cell or even at different time points. Post-translational modifications such as phophorylation, ubiquitination and methylation add another level of complexity to the analysis of the entirety of proteins in a cell.

Major methods to study proteins are immunoassays and mass spectrometry. Western blot and enzyme-linked immunosorbent assays (ELISA) detect and quantify certain proteins (antigens) with specific antibodies. Mass spectrometry based methods, on the other hand, are used in protein profiling as they quantify almost all known proteins in cells in parallel with high resolution (Glish & Vachet, 2003). This “bottom-up” approach is widely used for both protein identification and quantification, in which proteins from a sample are digested into smaller peptides with enzymes and analyzed in a mass spectrometer. The mass-to-charge ratios of peptides are then determined by the spectrometer and appear as peaks on a mass spectrum. These peaks reflect the characteristic patterns of proteins and are then used for their identification.

For the quantification of protein abundance, stable isotope labeling in cell culture (SILAC) is one popular method (X. Chen et al., 2015). In such an approach, cells are cultivated in two different conditions: for one, in a medium with normal amino acids; in the other condition with heavy isotopes labeled amino acids, which will be eventually incorporated into the cells as they grow. The samples from both conditions are then pooled together and undergo mass spectrometry. The abundance ratios of proteins between labelled and unlabeled proteins is then determined by the peak intensities in the mass spectrum of the two samples. This method allows the quantification of many peptides in cells and thus enabling high-throughput proteomics study.

4.2.3 Available hubs for biological data

There are a couple of databases or platforms that store data from high-throughput functional genomics experiments, such as Expression Atlas (https://www.ebi.ac.uk/gxa/home) (Papatheodorou et al., 2020), ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) (Parkinson et al., 2005) and Genomic Expression Archive (https://www.ddbj.nig.ac.jp/gea/index-e.html) (Kodama et al., 2019). One of the largest databases is Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) (Edgar et al., 2002), which is a publicly available genomic database supported by the National Center for Biotechnology Information (NCBI). It accepts both raw and processed data submitted from high throughput gene expression studies, which is acquired with different technologies such as DNA microarrays, high throughput nucleic acid sequencing, RT-PCR, SAGE. Apart from gene expression data, GEO also collects data from studies on genome methylation, genome variation/copy number, etc. A submitted series of data comes with the description and methodology of the experiment, and the attributes of samples. At the time of writing (https://www.ncbi.nlm.nih.gov/geo/summary/?type=platforms), it is holding more than 4 million samples from over 146,452 series, mainly from studies on human (Homo sapiens) and mouse (Mus musculus). Researchers could either download and analyze the data files of interest, or execute queries to look for relevant data (using keywords such as gene name, diseases or nucleotide sequence), perform analysis (e.g. differential expression on curated datasets) and visualization (as a genome track) on the data directly on the GEO website.

Disease-associated data are also available on specific platforms, such as the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/) from The Cancer Genome Atlas (TCGA) project (Weinstein et al., 2013), which contains large-scale sequencing results focusing on cancer genomic datasets. TCGA is a project supervised by the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), and aims to profile and analyze human tumors and discover aberrations at different levels from genome, transcriptome, proteome to epigenome. It has generated sequencing, expression, single nucleotide variation, copy number variation and methylation data from 33 cancer types (including 10 rare cancers). The main technologies used include RNA sequencing, whole genome/exome sequencing, genotyping array and methylation array. Currently, more than 84000 cases, most of them with comprehensive clinical data like tumor staging, survival, age and gender, and could be accessed by the research community on its repository. Simple analysis on the datasets, such as cohort comparison and clinical data analysis are also available on the TCGA portal.

4.3 Big data, big challenges

The advancement in sequencing and profiling technologies in genomics, transcriptomics, and proteomics generates a huge amount of data that are still growing exponentially. The amount of sequencing data and the speed of its generation already qualify it as “big data”, which is characterized by large volume and high velocity of production (Jain, 2016) . Like the big data in any other fields, the volume and complexity of these data pose a true challenge to their analysis and interpretation. Bioinformatics tools, combined with statistical analyses enable us to identify mutations, differential expression of genes and proteins, gene regulations events, etc. On the other hand, increased computing power and recent development in data science, especially in the area of data storage, analysis and visualization, allow data to be handled more efficiently and effectively (Manyika et al., 2011).

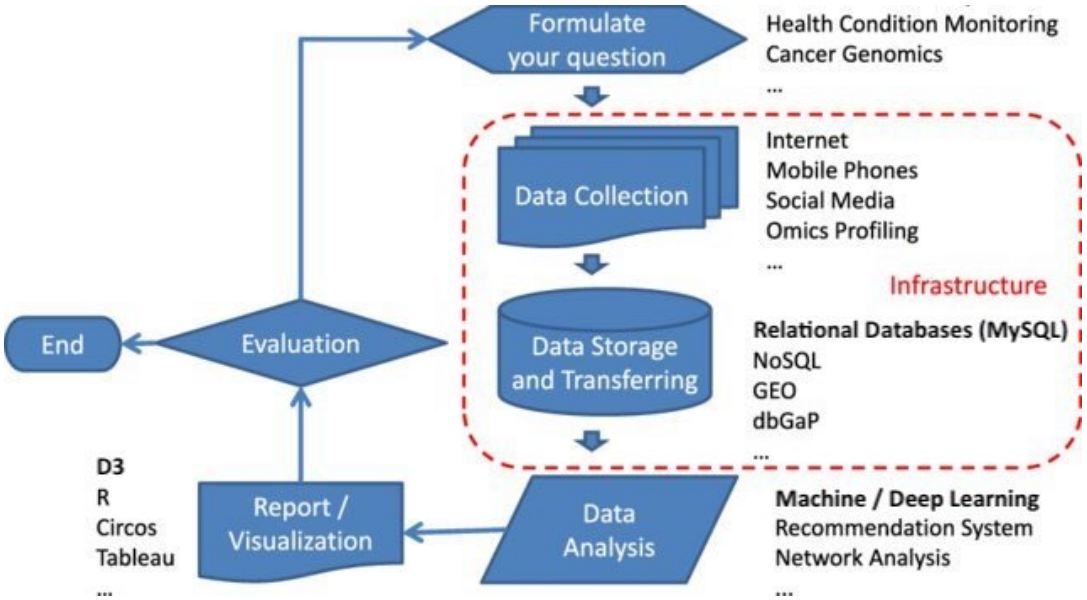


Figure 4.6 Workflow and essential stages of typical Big Data projects in healthcare.

After the question or objective has been formulated, the right kind of data has to be collected through various sources. Different types of data require different ways of storage to achieve optimal efficiency for archiving and fetching. To retrieve information from collected data, various approaches or algorithms are available to perform data analysis, depending on the nature and structure of the data. Finally, to help gaining insight from the analyzed data and perform evaluation on the results, the data could be visualized either with visualization tools or custom visualizations developed using packages from popular programming languages. The flow is usually an iterative process, with a lot of back-and-forth testing and adjustment at different stages. Taken and modified from (T. Huang et al., 2015), under the terms and conditions of the Creative Commons by Attribution (CC-BY) License.

To leverage the increased computational capacity for the extraction of useful information from data and the facilitation of knowledge discovery, good data management is particularly crucial. Recently, a consortium of scientists and organizations have published the FAIR Guiding Principles, which data should meet the principles of Findability, Accessibility, Interoperability, and Reusability (Wilkinson et al., 2016). The implementation of these

principles within the scientific community is strongly encouraged, so that computational systems could easily find, access, interoperate, and reuse data with minimal human intervention. This could be achieved by using persistent identifiers (PIDs) on the data and annotate it with sufficient metadata. The PIDs should be globally unique to be readily identifiable and the associated data and metadata should be always accessible and indexed in a searchable source (Findability and Accessibility). By using a standard language across all the data and metadata, they could also be easily reused and integrated with other data (Interoperability and Reusability).

4.3.1 Data storage and retrieval

In order to be efficiently accessed or retrieved for analysis, data are usually stored in an organized way within databases. Two main types of database systems or models in today's technology are relational database, or sometimes commonly referred as **SQL** database, that use Structured Query Language (SQL) for querying (https://www.iso.org/standard/63555.html); and **NoSQL** database (http://nosql-database.org/) that refers to non SQL or non-relational databases.

SQL databases store data in tables with definite rows and columns (structured data). Each entry (row) has a unique key, and tables can be connected by "relationships". Therefore, relational operations, such as joining two tables based on a column (e.g. retrieving gene annotation from one table and gene expression from another based on gene names), can be easily achieved. Storing different data in different but connected tables also avoids data redundancy by minimizing duplicated information. Popular relational database management systems are MySQL, PostgreSQL (both open-source) and Oracle (proprietary).

NoSQL databases, in contrast, store data in a schema-free version, in which the data need not to be in a defined structure. They could be stored as key-value pairs, free form JSON (JavaScript Object Notation) documents that consist of attribute-value pairs, or presented in graphs. This allows more flexibility as there are virtually no restrictions on the data model, and each entry can store different types and lengths of data. Common NoSQL databases include MongoDB (document-based), Redis (key-pair values) and Neo4j (graphs). Moreover, data storage can be split over several physical entities (storage units or computers), thus not restricting the size of the database.

4.3.1.1 SQL vs NoSQL

The choice of the type of database depends on the data structure, the need for reliability and scalability, performance and many other factors. SQL is a mature and stable database system with a long history that has been tested extensively. NoSQL, on the other hand, is relatively new and gaining popularity because of its flexibility to store unstructured data. Therefore, it is particularly useful when storing text documents (e.g. Twitter entries) or relationships between nodes (Facebook users' network). It is also highly scalable and fast when dealing with simply queries.

However, the relaxed data schema of NoSQL means data entries do not have to be validated for their types or checked for duplications. NoSQL also compromises consistency of data for speed, and lack the property of ACID (Atomicity, Consistency, Isolation and Durability) as in SQL. Although SQL might be not as performant as NoSQL with simple queries, it is usually faster when dealing with complex queries (e.g. join operations) as it is well-structured. Thus, SQL database are preferred if the data has a defined structure and relational queries are often called, or if robustness is an important factor.

4.3.2 Data analysis

Some biological data, such as sequencing data, are unstructured but are usually processed with bioinformatics tools to give relatively structured data as output (e.g. rows of entries of gene/protein expression value, mutations location, etc) for downstream analysis. The analysis typically involves data cleaning (e.g. dealing with missing or corrupted values in microarray data), normalization (e.g. RNA-sequencing data), transformation (e.g. Log transformation of skewed data) and statistical test (e.g. differential expression of genes). These steps usually are part of the bioinformatics analysis pipeline.

Two mainstream programming languages of nowadays for the purpose of data analysis are Python and R (Costa, 2020). Python is famous for its readability and has become one of the fastest growing language in recent years (https://www.python.org/). R has a strong focus on statistical and graphical techniques and is widely used among statisticians (https://www.r-project.org/). Both languages are popular in the community because of the availability of libraries or packages that can process sequencing data (e.g. Bioconductor packages in R; Biopython in Python), manipulate data (Pandas in Python) and perform statistical analysis (e.g. SciPy in Python).

Learning from the data to make predictions or recognize patterns is part of the data analysis as well, and this is how machine learning algorithms come into play. There are two broad

categories of machine learning, depending on the tasks: Supervised learning and unsupervised learning. Supervised learning algorithms try to learn a function that could map an input to an output. A set of labelled data, consisting of both the input and the output has to be fed to the learning algorithm for training, so that it can make predictions for new, unlabeled data. The output can be either a discrete (classification) or a continuous value (regression). For example, cancer patients can be classified into respective risk groups in terms of survival (output) based on their gene expression profile (input) (Y. C. Chen et al., 2014). There are a number of supervised leaning algorithms and among the most popular ones are Support Vector Machines, Decision Trees, Naive Bayes and the more complex Neural Network. Unsupervised learning algorithms infer a function that describes the unlabeled input data, for instance, by grouping the samples based on their similarity. An example would be clustering patients into previous unknown subgroups based on their gene expression profiles (Angermueller et al., 2016). Common clustering algorithms are k-means and hierarchical clustering.

Learning patterns in data can help give insights and reveal meaningful information. However, sequencing and profiling data often have the issue of high dimensionality (i.e. large number of input variables), which complicate the analysis. Dimensionality reduction is therefore often an essential step in data analysis. It is a process to reduce the number of variables while retaining the maximum amount of variation within the dataset. Principle Component Analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) are examples of dimensionality reduction techniques. Libraries for performing machine learning tasks and related functions are abundant in both Python (scikit-learn (https://scikit-learn.org/), tensorflow (https://www.tensorflow.org/), keras (https://keras.io/)) and R (randomForest (https://www.stat.berkeley.edu/~breiman/RandomForests), tree (https://cran.r-project.org/web/packages/tree/index.html), nnet (http://www.stats.ox.ac.uk/pub/MASS4/)).

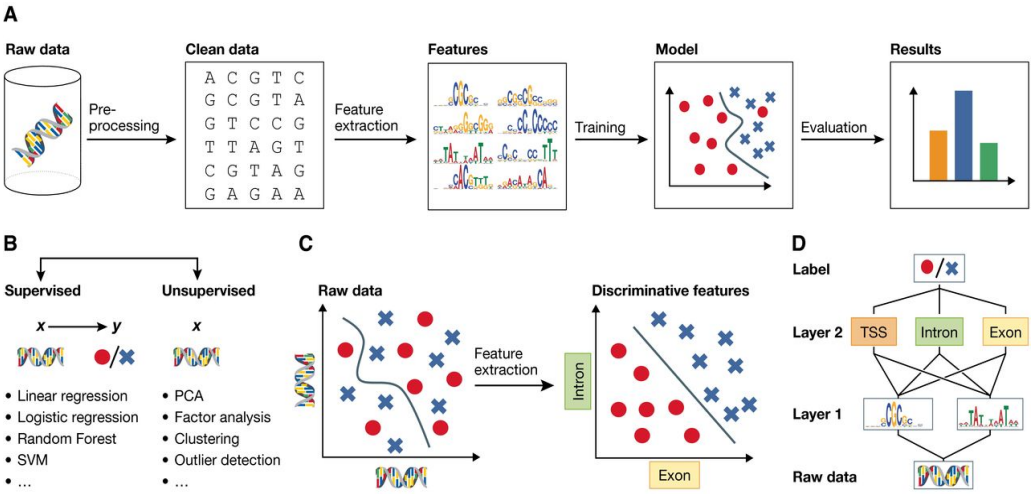


Figure 4.7 Data Analysis and Machine learning workflow.

(a) Data analysis and modelling can be performed once data has been collected. A typical workflow involves four steps: data cleaning and pre-processing, feature extraction, model fitting (training) and evaluation. (b) The two categories of machine learning algorithms are supervised machine learning, which relate input features to output label; and unsupervised machine learning, which learns the pattern in the data without labels. (c) High-dimensional omics data is often challenging for data analysis. Performing feature extractions, either manually or with the help of unsupervised models could help with the analysis and obtaining insights. Taken and modified from (Angermueller et al., 2016), under the terms and conditions of the Creative Commons by Attribution (CC-BY) License.

4.3.3 Data exploration and visualization

Representing data in a visual manner is useful for the exploration and interpretation of data. Effective visualization gives users intuition and allows them to better understand complex data, especially those in more than one dimension or with correlations. This helps with analysis and for making conclusions. Depending on the nature of the data and the aim of the analysis, the type of visualization can be as simple as a scatter plot (e.g. correlation between the expressions of two genes) or as complex as a network of nodes (e.g. interactions between proteins).

While a number of programs are able to create different types of charts and diagrams, some programming languages offer libraries that make it possible to build more customizable and diverse visualizations. Python, for example, has the libraries matplotlib (https://matplotlib.org/) and seaborn (https://seaborn.pydata.org/); whereas R comes with packages like ggplot2 (https://ggplot2.tidyverse.org/) and Lattice (https://www.rdocumentation.org/packages/lattice). Most of the libraries in these two languages build only static graphs, but some also create visualizations with a certain degree of interactivity such as highchart (https://www.highcharts.com/) and plotly (https://plotly.com/).

4.3.3.1 D3 in Javascript

D3.js (Data-Driven Documents) (https://d3js.org/) is one of the most powerful open-sourced visualization libraries in Javascript, a dominant language in web programming. Capable of supporting almost all the chart types, D3 is the golden standard for visualization in Javascript and many other libraries are based on it. It is able of generating interactive and dynamic visualization in web browsers, without the need of installing any software. This is achieved by creating an SVG (Scalable Vector Graphics) object, a scalable and resolution independent image, and dynamically styling it through CSS (Cascading Styling Sheets), a language that characterizes the styles and contents of web elements (https://developer.mozilla.org/en-US/docs/Web/CSS).

The power of D3 lies in its ability to bind data to the DOM (Document Object Model) of a web document, which describes the structure the elements of the document. Given a dataset, either in the format of JSON (JavaScript Object Notation), or CSV (Comma-Separated Values), D3 can create for each data point an SVG element within the DOM, with properties (colors, size, shape, coordinates) and behaviors associated to the value of that data point. End-users can interact with the visualization (selection by mouse click, filtering by keyboard entries) that gives dynamic response as D3 modifies the style of SVGs (each representing a data node) in parallel.

D3 is superior to other libraries in Javascript or even other languages in terms of interactivity and customizability (https://github.com/d3/d3/wiki), as its API, together with Javascript, enable creators to have control over low-level details of the visualization and be very specific about the properties and behaviors of each elements on it. However, unlike Python and R, Javascript does not provide many methods and libraries for data manipulation, or statistics and analysis, thus data have to be processed in other languages before feeding them to D3 for visualization.

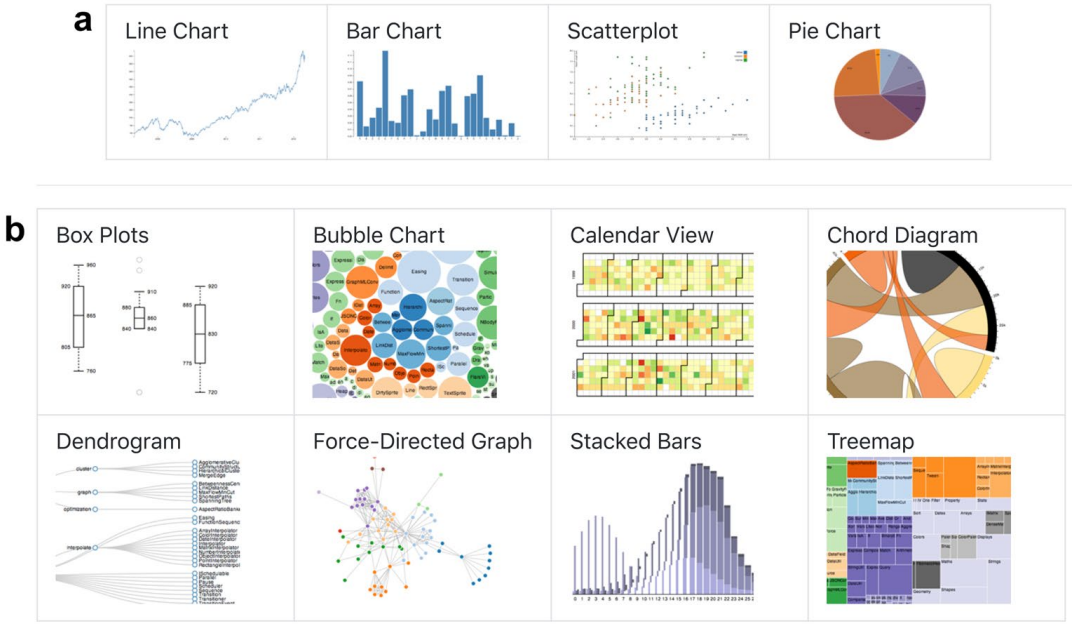


Figure 4.8 Visualizations with d3.js.

D3.js is a Javascript library for generating interactive and dynamic visualizations in typical web browsers. It creates SVG (Scalable Vector Graphics) objects that could be formatted or styled dynamically through CSS (Cascading Styling Sheets) standards. The visualizations produced using D3.js are highly customizable. Both simple graphs (a) and more complex visualizations (b) could be generated with existing methods from the library. Taken from https://d3js.org/, under the terms and conditions of the Creative Commons by Attribution (CC-BY) License.

4.3.4 Current available tools for data analysis and visualization

As mentioned in Section 4.2.3, some of the platforms or databases of omics data come with tools for the analysis and visualization of their own data. For example, on the GEO website, users could perform differential expression analysis on curated datasets managed by GEO and visualize genome tracks; On the portal where TCGA data is hosted, users could carry out cohort comparison and clinical data analysis on the data; Where as ArrayExpress provides an Bioconductor package to access the data and build data structures upon it for further analysis in R.

There are also various other tools developed for the analysis and visualizations of data from the above repositories or user-prepared data. These tools can either be a web tool or a desktop application, and they usually visualize data in one of these three types: Heatmaps, genomic coordinates and networks. To name a few examples, cBio Cancer Genomics Portal (Cerami et al., 2012) is a web platform offers the analysis of differential expression, mutation frequency and copy number alterations on pre-calculated TCGA data (and other curated cancer datasets) and visualizations in heatmaps; UCSC Cancer Genomics Browser (Zhu et al., 2009) is another web tool for the quantitative analysis of TCGA data, but it also allows users to upload their own data for analysis. Like cBioPortal, it provides

views for expression data in heatmaps, and also for mutation and copy number data; Integrative Genomics Viewer (IGV) (Robinson et al., 2011) is a desktop application developed by the Broad Institute to explore genomic datasets provided by users in genomic coordinates visualizations. The tracks and genomic regions can be annotated with integrated metadata; Finally, Cytoscape (Shannon et al., 2003), also a desktop application, serves as a network analytics software to analyze genetic interaction and gene regulatory events and visualize them in networks. All these tools are very helpful for the analysis of high volume omics data and could facilitate the discovery of underlying biological patterns out of these complex data.

4.3.5 Challenges in the analysis of mitochondrial genes addressed in this thesis

Despite the availability of visual analytics platforms and software for omics data, tools dedicated for the analysis of certain context-specific gene sets, such as mitochondrial genes, are very limited. One of the reasons is that it is often time consuming to construct such a gene list that could include most, if not all, of the related functions and pathways in that specific context. Most of the existing lists or electronic repositories of mitochondrial genes are not comprehensive since they are constructed with computational approaches and could either be susceptible to overfitting of the training data (MitoMiner (Smith et al., 2012)), or lack experimental confirmation (MitoPred (C. Guda et al., 2004)). MitoCarta (Calvo et al., 2016) is one of repositories that contains a more complete set of mitochondria-associated genes. However it currently includes only human and mouse mitochondrial genes. Another challenge is to assign proper functional annotations to the genes to facilitate further analysis, which is also a tedious task that requires a lot of literature review and hence human involvement.

Equally challenging is the development of the appropriate analysis and visualization tools. They should be intuitive enough for users to understand readily what the purpose of the visual analytics tool is, how to use the tool and what the result means. While most of above-mentioned tools could achieve these, they are usually not dynamic and interactive enough for users to explore further according to their own needs, and the types of visualizations are relatively limited. The usability of the tool is also important, as complex requirements often deter researchers from using it. For example, visualizations of some tools require high-end computer graphics to display (Caleydo (Streit et al., 2009)), and some requires users to possess certain programming knowledge (Arrayexpress (Parkinson et al., 2005)). Therefore an ideal tool should have minimum pre-requisites on the specifications of users' machines or users' programming knowledge to operate it, and be as user-friendly as possible.

5 Objectives

In light of the need for a visual analytics tool specific for a genes with mitochondrial functions, this thesis aims mainly to develop a data mining and visualization tool dedicated for the exploration and analysis of expression and mutation landscapes of mitochondrial genes. A critical component of such platform, which is currently lacking, is a comprehensive list of mitochondrial genes with annotations, that could be integrated into the analysis and visualization tools. Therefore, we first addressed this issue by assembling and manually curating lists of genes with annotated mitochondrial processes, which we referred to as mitochondrial interactome, for human and a few other model organisms. To ensure its completeness and accuracy, we started with a few published mitochondrial proteomic data and supplemented with all non-coding mitochondrial genes and potentially missing genes from different sources, including inventories that provide pathway or ortholog information. Then we performed manual cleaning to remove false-positives by, for example, searching the literature for the localization of gene products. As for the annotations, we referred to multiple established databases and decided on a set of controlled vocabulary to group the mitochondrial genes.

The next point to address is the development of the analysis and visualization tools. To tackle with the problem of a general lack of interactivity in other analytics platforms, we have developed a number of user-friendly, dynamic and interactive visualization tools that are modular and could be incorporated into different platforms. We have also built a platform, mitoXplorer, to host public dataset and visualization tools that have been integrated with the interactome, in order for users to freely explore both public or uploaded data. To allow the platform to be widely available and accessible, we decided to build it as a web application, so that it is agnostic to operating systems and does not require any installation or programming knowledge on the users' side.

The final objective is to prove robustness and the predictive power of such visual data mining tools. We analyzed the transcriptome and proteome data of trisomy 21 cell line to demonstrate the ability of the tools to explore expression dynamics of mitochondrial genes. We also validated with experiments the observations from the analysis, to show how testable hypothesis could be generated that lead to discovery about mechanisms of mitochondrial defects in trisomy 21 patients. Next, we analyzed the expression and clinical data of cancer patients from public cancer dataset (TCGA), namely kidney renal papillary cell carcinoma, to show how visual data mining tools could also help generating hypothesis on the role mitochondrial functions in terms of cancer development and progression.

6 Results

6.1 mitoXplorer, a visual data mining platform to systematically analyze and visualize mitochondrial expression dynamics and mutations

Mitochondria are important organelles that have diverse roles apart from energy production, such as ROS defense, cell homeostasis, and the control of signaling pathways. Impaired mitochondrial functions can lead to various diseases and metabolic disorders, and the field has attracted increased attention. However, tools or platforms specific for the analysis of omics data of genes with mitochondrial functions did not exist previously, despite the availability of public data from high throughput gene expression studies. Therefore, in this study we developed a web-based visual data mining platform – mitoXplorer, for the in-depth analysis and visualization of the expression and mutation landscape of mitochondrial genes.

mitoXplorer provides a set of dynamic, interactive and intuitive visualizations, that allows users to explore and interact with the analysed data. We have also integrated manually curated mitochondrial interactomes that consist of ~1200 genes grouped in 38 mitochondrial processes for four organisms (Human, Mouse, Fruit Fly and Budding Yeast), with accurate and updated annotations. Together with the visualization tools, mitoXplorer serves as user-friendly platform for the mining of mitochondrial expression dynamics and mutations across various public or user-owned datasets.

To demonstrate the predictive power of mitoXplorer, transcriptome and proteome data from cell lines with trisomy 21 was analysed and investigated in detail. We found that, in one of the trisomy 21 cell lines (RPE1 T21), there were significant differences in the regulation of transcriptome and proteome in the some of the mitochondrial processes. Notably the dysregulation of several mitochondrial ribosome proteins related to Translation, which potentially causes the down-regulation of OXPHOS genes along the entire respiratory chain that could result in defects in oxidative phosphorylation. The prediction of mitochondrial respiration failure in such cell line has then be confirmed experimentally. This showed the robustness of mitoXplorer as a visual data mining platform, which could help discover underlying molecular mechanisms and patterns for the generation of hypotheses for experimental validation.

Published online 4 December 2019

Nucleic Acids Research, 2020, Vol. 48, No. 2 ***605–632***
doi: 10.1093/nar/gkz1128

mitoXplorer, a visual data mining platform to systematically analyze and visualize mitochondrial expression dynamics and mutations

Annie Yim[1,†], Prasanna Koti[1,†], Adrien Bonnard[2], Fabio Marchiano[3], Milena Dürrbaum[1], Cecilia Garcia-Perez[4], Jose Villaveces[1], Salma Gamal[1], Giovanni Cardone[1], Fabiana Perocchi[4], Zuzana Storchova[1,5] and Bianca H. Habermann[1,3,*]

[1]Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany, [2]Aix-Marseille University, INSERM, TAGC U1090, 13009 Marseille, France, [3]Aix-Marseille University, CNRS, IBDM UMR 7288, 13009 Marseille, France, [4]Functional Genomics of Mitochondrial Signaling, Gene Center, Ludwig Maximilian University (LMU), Munich, Germany and [5]Department of Molecular Genetics, TU Kaiserslautern, Paul Ehrlich Strasse 24, 67663 Kaiserslautern, Germany

Received May 30, 2019; Revised October 30, 2019; Editorial Decision November 13, 2019; Accepted November 18, 2019

ABSTRACT

Mitochondria participate in metabolism and signaling. They adapt to the requirements of various cell types. Publicly available expression data permit to study expression dynamics of genes with mitochondrial function (mito-genes) in various cell types, conditions and organisms. Yet, we lack an easy way of extracting these data for mito-genes. Here, we introduce the visual data mining platform mitoXplorer, which integrates expression and mutation data of mito-genes with a manually curated mitochondrial interactome containing ∼1200 genes grouped in 38 mitochondrial processes. User-friendly analysis and visualization tools allow to mine mitochondrial expression dynamics and mutations across various datasets from four model species including human. To test the predictive power of mitoXplorer, we quantify mito-gene expression dynamics in trisomy 21 cells, as mitochondrial defects are frequent in trisomy 21. We uncover remarkable differences in the regulation of the mitochondrial transcriptome and proteome in one of the trisomy 21 cell lines, caused by dysregulation of the mitochondrial ribosome and resulting in severe defects in oxidative phosphorylation. With the newly developed Fiji plugin mitoMorph, we identify mild changes in mitochondrial morphology in trisomy 21. Taken together, mitoXplorer (http://mitoxplorer.ibdm.univ-mrs.fr) is a user-friendly, web-based and freely accessible software, aiding experimental scientists to quantify mitochondrial expression dynamics.

INTRODUCTION

Enormous amounts of transcriptomic data are publicly available for exploration. This richness of data gives us the unique opportunity to explore the behavior of individual genes or groups of genes within a vast variety of different cell types, developmental or disease conditions or in different species. By integrating these data in a sophisticated way, we may be capable to discover new dependencies between genes or processes.

Specific databases are available for mining and exploring disease-associated data, such as The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov/) (1), or the International Cancer Consortium Data Portal (ICGC, https://dcc.icgc.org/) (2). Especially cancer data portals allow users to perform deeper exploration of expression changes of individual genes or gene groups in different tumor types ((1–3); for a review on available cancer data portals, see (4)). Expression Atlas (https://www.ebi.ac.uk/gxa/home) on the other hand provides pre-processed data from a large variety of different studies in numerous species (5). Indeed, the majority of transcriptomic datasets are not related to cancer and are stored in public repositories such as Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) (6), DDBJ Omics Archive (https://www.ddbj.nig.ac.jp/gea/index-e.html) (7), or ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) (8). Currently, it is not straightforward to integrate data from these repositories without at least basic programming knowledge.

*To whom correspondence should be addressed. Tel: +33 4 91 26 92 36; Email: bianca.habermann@univ-amu.fr
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Next to extracting reliable information from -omics datasets, it is equally important to support interactive data visualization. This is a key element for a user-guided exploration and interpretation of complex data, facilitating the generation of biologically relevant hypotheses—a process referred to as visual data mining (VDM, reviewed e.g. in (9)). Therefore, essentially all online data portals provide graphical tools for data exploration.

What is fundamentally lacking is a user-centric, web-based and interactive platform for data integration of a set of selected genes or proteins sharing the same cellular function(s). The benefits of such a tool are evident: first, it would give us the possibility to explore the expression dynamics and the presence of mutations in this set of selected genes across many different conditions, tissues and species. Second, by integrating data using enrichment techniques, for instance with epigenetic data or by network analysis using the cellular interactome(s), it would allow us to identify the mechanisms that regulate the expression dynamics of the selected gene set.

One interesting set of genes are mitochondria-associated genes (mito-genes): in other words all genes, whose encoded proteins localize to mitochondria and fulfill their cellular function within this organelle. Mito-genes are well-suited for such a systematic analysis, because we have a relatively complete knowledge of their identity and can categorize them according to their mitochondrial functions (10). This *a priori* knowledge can help us in mining and exploring the expression dynamics of mito-genes and functions in various conditions and species.

Mitochondria are essential organelles in eukaryotic cells that are required for producing cellular energy in form of ATP and for numerous other metabolic and signaling functions (10). Attributable to their central cellular role, mitochondrial dysfunctions were found to be associated with a number of human diseases such as obesity, diabetes, neurodegenerative diseases and cancer (11–15). However, mitochondria are not uniform organelles. Their structural and metabolic diversity, both of which influence each other, has been well described in literature (16–20). This mitochondrial heterogeneity in different tissues is reflected in their molecular composition (21). The total number of proteins that contribute to mitochondrial functions and localize to mitochondria is currently not precisely known and might differ between tissues and species (22,23). Yet, based on proteomic data from several organisms, it is likely that mitochondria contain >1000 proteins (23–30). Mitochondria have their own genome, whose size in animals is between 11 and 28 kb (31). Most metazoan mitochondria encode 13 essential proteins of the respiratory chain required for oxidative phosphorylation (OXPHOS), all rRNAs of the small and large mitochondrial ribosomal subunits, as well as most mitochondrial tRNAs (32). All other proteins found in mitochondria (mito-proteins) are encoded by genes in the nucleus; the protein products of these nuclear-encoded mitochondrial genes (NEMGs) are transported to and imported into mitochondria.

Based on data from mitochondrial proteomic studies or genome-scale prediction of mito-proteins, several electronic repositories of the mitochondrial interactome have been created (24,33–36), though they often lack a proper functional assignments of mito-proteins. Moreover, proteomic studies describing the mitochondrial proteome can suffer from a high false-positive rate (23), whereas computational prediction or machine learning in most cases lack experimental confirmation (37). As a consequence, none of the published mitochondrial interactomes available to date can be taken without further manual curation. Moreover, these lists are not integrated with any available data analysis tool to explore mitochondrial expression dynamics under varying conditions or in different tissues or species.

In this study, we present mitoXplorer, a web-based, highly interactive visual data mining (VDM) platform designed to specifically mine the dynamics of a manually curated gene set with mitochondrial functions in transcriptome, proteome, as well as mutation-based data. To achieve this, mitoXplorer integrates -omics data with our hand-curated mitochondrial interactomes for currently four different model species. With mitoXplorer, we can explore the expression dynamics, as well as mutations of mito-genes and their associated mitochondrial processes (mito-processes) across a large variety of different -omics datasets without the need of programming knowledge. MitoXplorer provides users with dynamic and interactive figures, which instantly display information on mitochondrial gene functions and protein-protein interactions. Users can analyze publicly available data stored in our mitoXlorer database or upload their own data for integration with our hand-curated mitochondrial interactome. In order to demonstrate the analytical and predictive power of mitoXplorer and to experimentally verify mitoXplorer predictions, we generated transcriptome and proteome data from aneuploid cell lines, carrying trisomy 21 (T21), the most common chromosome abnormality in humans which is also known to cause substantial mitochondrial dysfunctions (38). We used mitoXplorer to analyze and integrate our data with publicly available trisomy 21 data. MitoXplorer enabled us to predict respiratory failure in one of our T21 cell lines, which we experimentally confirmed, thus demonstrating the predictive power of mitoXplorer.

MATERIALS AND METHODS

Implementation of mitoXplorer

Web interface of mitoXplorer (front-end). The web interface of mitoXplorer at the *front-end* allows users to access, interact and visualize data from its database, including the interactome and expression/mutation data. The interactive elements and visualizations on mitoXplorer are all built with Javascript, a dynamic programming language that enables interactivity on webpages by manipulating elements through DOM (Document Object Model). DOM is a representation of document, such as HTML, in a tree structure, with each element as a node or an object. Through Javascript and its libraries, visualizations in mitoXplorer can react to users' action and dynamically change the properties (size, color, coordinates) of web elements and display interactivity. All the visualization components in mitoXplorer described below are modular by design and can be deployed individually or incorporated into web platforms easily.

Mitochondrial Interactome (D3—data binding and selection). The visualization of the interactome is created with the implementation of a Javascript library, D3 (d3.js) (39). D3 (data-driven documents) is capable of binding data, usually in JSON (Javascript-oriented notation) format, to the elements of the DOM so that their properties are entirely based on given data. In the Interactome View, D3 creates an SVG (Scalable Vector Graphic) element for each gene within the DOM in the form of a bubble, with sizes and colors dependent on the associated log2 fold change (log2FC) values. The coordinates of bubbles are also calculated according to the data (e.g. the largest one being at the center) so that the layout of the whole interactome is visually appealing. Upon hovering over any bubble (gene), D3 selects the element and passes additional data bound to that element to the corresponding web element (sidebar) for display.

Comparative plot (D3—transition and sorting). The comparative plot combines three interdependent visualizations (scatterplot, bar chart and heatmap) built upon D3. Apart from data-binding and selection, these visualizations exploit the functionality of D3 for transition and sorting through its API. In the scatterplot, genes are displayed as nodes, whose colors and positions again depend on the data (log2FC). When another mito-process is selected at the bar chart, D3 updates the data bound to the node and the properties of the nodes are changed. The transition (changes in color and position) is smooth and gives users the impression that the visualization is truly dynamic and interactive. D3 can manipulate not only the elements, but also the data bound to the elements. Upon clicking the dataset or gene names on the heatmap, the data can be sorted accordingly and an index is assigned to each element (tile on the heatmap) to indicate its position.

Hierarchical clustering (mpld3—visualization in Python implemented in D3). The heatmap displaying the results of hierarchical clustering is built with mpld3, a Python library that exports graphics made with Python's Matplotlib-based libraries to JSON objects that can be displayed on web browsers. Mpld3 benefits from D3's data-binding property and allows users to create a plugin that interacts with the data on the visualization. The advantage of using mpld3 is that analyses and visualizations made in Python can be directly translated to JSON and deployed in Javascript on webpages without re-programming. In the case of hierarchical clustering, since libraries for both clustering analysis and visualization of results in a heatmap with a dendrogram are available in Python (described below), it is exported to JSON with mpld3 and a Javascript tooltip plugin that allows users to select data or display information with D3.

Principal component analysis (three.js—3D visualization). The visualization of the result of Principal Component Analysis (PCA) is 3-dimensional, with each dimension representing one of the first three principal components (PCs). This is achieved through the implementation of three.js, a Javascript library that enables animated 3D graphics to be created and displayed in a web browser. It starts with building a 'scene', or a canvas, on which 3D objects will be created. Then a 'camera' is set up that controls the view of objects on the scene from the users' perspective, such as the field of view (width, height, depth) and its ratio; and a 'renderer' that renders the scene at short time intervals so objects are displayed as animated objects (either they are animated by themselves or moved around on the scene by users). Objects of different texture, geometry and color can now be added to and rendered on the scene. Finally, the scene with objects is attached to the DOM of a webpage to become visible. In the PCA visualization, each dataset is represented and rendered as a small sphere, with coordinates (x, y, z) depending on the values of its first three PCs, and colors on the grouping of that dataset. When users drag around on the canvas or zoom in or out, all objects are re-rendered in such a way that the scene appears to be a 3-dimensional space.

*MitoXplorer database (*back-end*).* A MySQL database hosted at the *back-end* of mitoXplorer contains the interactomes of mito-genes, including the mito-process, gene ontology and the interactions between gene products; and the expression and mutation data from public databases. Each entry of the expression and mutation data has a foreign link to the interactome and file directory (dataset table). This ensures that the expression and mutation data will be updated together with the interactome, or when a dataset is updated or deleted. Users can upload their own differential expression and/or mutation data, which will be processed and integrated with the interactome by extracting mito-genes, and stored in the mitoXplorer database for up to 7 days.

Data analysis and communication between front- *and* back-end. A Python application serves as a bridge between the *front-* and *back-end* of mitoXplorer. Upon the users' request to access the database or perform analysis at the web interface, an AJAX-asynchronous call directed to the Python application is made, so the request can be performed in the background and the webpage is updated without reloading. The Python application then processes the request by connecting to the MySQL database and analyzes the data retrieved from it. The application also handles the user uploads (e.g. data cleaning) before saving it to the MySQL database. The main libraries used by the Python application for analysis include: (i) Scikit-learn: a machine learning library that provides tools for PCA, to perform dimensionality reduction on the expression of all mito-genes and of each mito-process. The first three principal components are extracted for each dataset. (ii) SciPy: a mathematical library that provides modules for hierarchical clustering, to calculate 2D distance matrices between genes and between datasets based on expression values, for each mito-process. (iii) Seaborn: a statistical visualization library built on top of SciPy to create heatmaps from the results. All the results are produced in JSON format, which are then sent via the HTTP protocol back to the *front-end* and visualized with Javascript.

The usage of mitoXplorer does not require installation or programming knowledge. Documentation and tutorials are available online and on GitLab (https://gitlab.com/habermannlab/mitox). MitoXplorer is also available for download and installation on a local server, if users wish to

build their own gene list and apply the interactive features and database of mitoXplorer, which stores the available expression and mutation data for all genes. Setup instructions are also available on the mitoXplorer GitLab repository, as is a docker version of mitoxplorer (https://gitlab.com/habermannlab/mitox, branch docker-version).

Processing of public transcriptomic and proteomic data. Proteomic and transcriptomic data from Kühl (40), as well as Liu (41), Letourneau (42), Sullivan (43) and Spletter (44) were uploaded as provided by the authors.

Public NGS datasets downloaded from GEO which did not already contain differential expression data in the form of log2FC and *P*-value were analyzed according to available pre-analyzed data: datasets with available raw read counts were analyzed using DESeq2 (version 3.9, (45)). These included data from Chowdhury (46) and Garipler (47). Datasets for which only the normalized read counts were available, the $\log_2$FC was calculated for each sample, using the corresponding wild-type samples as control (or the mean of normal samples if there were no paired samples) following best-practice guidelines. We applied this to data from TCGA (1), downloaded from the NCI GDC Data Portal (https://portal.gdc.cancer.gov/)), from Fleischer (48) and from Huang (49). Finally, microarray time-course data of yeast meisosis (GEO accession: GSE75257) were analyzed with GEO2R (50).

Metadata of the datasets (e.g. cell types, analysis pipeline and genome version used for mapping) were also downloaded and stored in the mitoXplorer database. The links to the experiments for each dataset are available at the DATABASE summary page of mitoXplorer.

Transcriptomics and proteomics of aneuploid cell lines. The proteome analysis of the trisomic cell lines was previously described (51,52).

The raw reads from RNA-sequencing were processed to remove low quality reads and adapter sequences (using TrimGalore v0.4.5 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), which uses Cutadapt (53)) and FastQC (Andrew, S. (2010) FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc)), and aligned to the human reference genome (version hg19) with TopHat2 (v2.0.11) (54). Cuffdiff from the Cufflinks package (v2.2.1) (55) was used with standard parameters to calculate the expression difference between two samples (aneuploid versus diploid) of multiple replicates and test the statistical significance. Transcriptome and proteome information are available in public repositories: NGS data have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO series accession number GSE102855 and GSE131249.

Cell culture and treatment. The human cell line RPE-1 hTERT (referred to as RPE) was a kind gift by Stephen Taylor (University of Manchester, UK). Human HCT116 cells (referred to as HCT) were obtained from ATCC (No. CCL-247). Trisomic cell lines were generated by microcell-mediated chromosome transfer as described previously (51). The A9 donor mouse cell lines were purchased from the Health Science Research Resources Bank (HSRRB), Osaka 590-0535, Japan. All cell lines were maintained at 37°C with 5% CO_2 atmosphere in Dulbecco´s modified Eagle's medium (DMEM) containing 10% fetal bovine serum (FBS), 100 U penicillin and 100 U streptomycin.

MitoTracker staining and imaging. Mitochondria were stained in 96-well plates. The cells were incubated for 30 min at 37°C with 100 nM MitoTracker deep Red FM (M22426, Invitrogen®) dye prior to fixation. Cells were fixed with 3% PFA in DMEM for 5 min at room temperature. After washing twice with 1xPBST, plates were stored with 1× PBS containing 0.01% sodium azide. Plates were stored at 4°C in the dark. Imaging was carried out on an inverted Zeiss Observer.Z1 microscope with a spinning disc and 473, 561 and 660 nm argon laser lines. Imaging devices were controlled, and images were captured, stored and processed with the SlideBook Software in Fiji (56). The images were captured automatically on multiple focal planes (step size: 700 nm) with a 40× magnification air objective.

Metabolic profiling of wild-type and T21 cell lines. RPE and HCT cells and their T21 derivatives were seeded at 25 000 or 36 000 cells/well respectively, on XF96 cell plates (Seahorse Bioscience, Agilent Technologies), 30 h before being assayed. Optimization of reagents as well as CCCP and digitonin titrations were performed as described by the manufacturer's protocols (Seahorse Bioscience). The experiments were performed using the mitochondrial and glycolytic stress test assay protocol as suggested by the manufacturer (Seahorse Bioscience, Agilent Technologies). By employing the Seahorse Bioscience XF Extracellular Flux Analyzer, the rate of cellular oxidative phosphorylation (oxygen consumption rate (OCR)) and glycolysis (cellular proton production rate (PPR)) were measured simultaneously.

For OCR measurement, DMEM media was supplemented with 25 mM glucose, 1 mM pyruvate and 2 mM glutamine. Basal rate was recorded and additions for the mito stress test were as follows: 1.5 μM oligomycin, CCCP, 2 μM rotenone + 4 μM antimycin A. For PPR measurement, DMEM media was supplemented with 2 mM glutamine. Basal rate was recorded and additions for the glycolysis stress test were as follows: 10 mM glucose, 1.5 μM oligomycin and 100 mM 2-deoxyglucose.

For intact cells, the CCCP concentrations were 7 and 1.5 μM for RPE1 and HCT116 cells, respectively. The assays of intact cells were performed in 96-well plates with at least 10 replicates per cell line. For the permeabilized RPE1 cell lines, the CCCP and digitonin concentrations were 10 and 40 μM, respectively. For OCR measurement, Mannitol–sucrose buffer (MAS) was prepared according to Seahorse Biosciences. For permeabilization, digitionin was added to MAS buffer together with the respective respiratory substrates: 10 mM pyruvate/2 mM malate, 10 mM succinate/2 μM rotenone or 0.5 mM TMPD/2 mM ascorbate/2 μM antimycin A. Basal respiration was recorded, as were additions of 4 mM ADP, 1.5 μM oligomycin, CCCP and 2 μM rotenone ± 4 μM antimycin A or 20 mM Na-azide. The as-

says in permeabilized cells were performed in poly-D-lysine-coated 96-well plates with at least five replicates per cell line.

Normalization was performed with the CyQuant cell proliferation assay kit (Life Technologies) in the same plate used for the assay of intact cells; and in a parallel plate for the permeabilized cells. Data analysis was done according to (57).

The mitoMorph plugin for morphological characterization of mitochondria by image analysis. Classification and measurement of mitochondria were performed using the software ImageJ (58), complemented with all the default plugins provided by Fiji (56) and with the additional plugin FeatureJ. A set of functions were developed to assist the user in the preparation and analysis of the data, either in interactive or batch processing mode.

Using this toolset, after all the cells of interest were manually outlined in each image, the mitochondria were segmented and characterized. For each processing step, the algorithms used are reported as described in ImageJ, and their parameters are specified in physical units.

The images were pre-processed by first suppressing the background signal (rolling ball background subtraction, kernel radius: 2.5 μm) and then enhancing the mitochondria signal (Laplacian of Gaussian, smoothing scale: 1 μm, followed by contrast limited adaptive histogram equalization, CLAHE, kernel size: 2.5 μm). Mitochondria candidates were obtained by segmentation, using Yen thresholding algorithm (59), and subjected to classification based on a set of determined features.

Objects that were too small were excluded from the analysis, and the remaining ones were assigned to one of four categories: filamentous networked (filaments), puncta, rods and swollen (60). Objects that were quasi-round, compact in intensity, and larger than the puncta were classified as swollen. All objects with an intermediate phenotype between fragmented puncta and network of filaments were classified as rods.

Classification was performed by sequentially verifying different selection criteria, one set for each class, based on the following measured features: area (A), aspect ratio (AR), circularity (C), solidity (S), minimum Feret diameter (here indicated as minimum linear extension, MLE) and longest shortest-path (here indicated as extension, E). While all the other measures are directly derived from the segmentation, the extension is measured as the longest shortest-path between any two end points in the skeleton derived from the segmentation. The selection criteria are evaluated sequentially as reported in Supplementary Table S1.

We would like to note that analysis of mitochondrial morphology on projected images is limited, as mitochondrial structures might not be resolved properly.

Image analysis using mitoMorph and data processing. Image processing and analysis was done in Fiji. Image stacks were *Z*-projected, cells were manually selected and the resulting images were saved for further batch processing using mitoMorph. Resulting network statistics of mitochondrial features for each individual cell were used for further processing (Supplementary Table S1). All statistical processing and data visualization of mitoMorph results was done using RStudio (v1.1.423, R-version: 3.6.1). Data were averaged over both clones of the two T21 cell lines.

RESULTS

The outline of the mitoXplorer web-platform is illustrated in Figure 1: at the *back-end*, manually curated mitochondrial interactomes from human, mouse, *Drosophila* and budding yeast, as well as expression and mutation data from these four species are stored in a MySQL database (details on the implementation of the *back-end* are available in Materials and Methods, as well as Supplementary Figure S1).

The user interacts with the mitoXplorer web-platform via the *front-end*, which offers different visualization and analysis methods. Users can either browse stored public data or upload their own data.

The mitochondrial interactomes

The main component of mitoXplorer is the mitochondrial interactome. Its accurate annotation and completeness are essential for performing a meaningful mitoXplorer-based analysis. To establish mitochondrial interactomes, we have assembled and manually curated lists of genes with annotated mitochondrial processes (mito-processes). Currently, the interactomes of four organisms are available on mitoXplorer: *Homo sapiens* (human), *Mus musculus* (mouse), *Drosophila melanogaster* (fruit fly) and *Saccharomyces cerevisiae* (budding yeast). We started from published mitochondrial proteomic data (27,61) for the selected species and manually cleaned the data, as well as supplemented missing mito-genes in the following way: we removed obvious false-positives from the datasets; these included mainly proteins for which there is compelling evidence in literature that they are not localized to mitochondria. Next, we supplemented likely missing genes from the proteomic data using information from Mitocarta (24), KEGG (62) information of genes associated with mitochondrial pathways, as well as orthologs across species from the four proteomic datasets. To establish whether a protein in question is primarily localized to mitochondria or not, we relied on several sources: (i) evidence from the literature on a specific gene; (ii) information from the respective gene entry at NCBI (63), Flybase (64) or the *Saccharomyces* genome database (SGD) (65); (iii) information from the GeneCards database (66); (iv) information from UniProt (67). We supplemented the lists with all non-coding genes present in the mitochondrial genomes, namely all mitochondrial rRNAs, as well as tRNAs. After manual curation, we obtained 1229 human, 1222 mouse, 1139 *Drosophila* and 988 budding yeast mitogenes. We grouped the genes in mito-processes using controlled vocabulary. In addition to purely mitochondrial processes, we added cytosolic processes coupled to mitochondrial functions, including Glycolysis, the Pentose phosphate pathway, Apoptosis or the regulation of transcription of nuclear-encoded mitochondrial genes (Transcription (nuclear)). This resulted in a total of 38 mito-processes (Table 1). We selected the correct mito-process for each gene primarily using information from the same sources mentioned before (NCBI gene entry, Flybase, SGD, GeneCards, UniProt, KEGG). According to our current annotation,

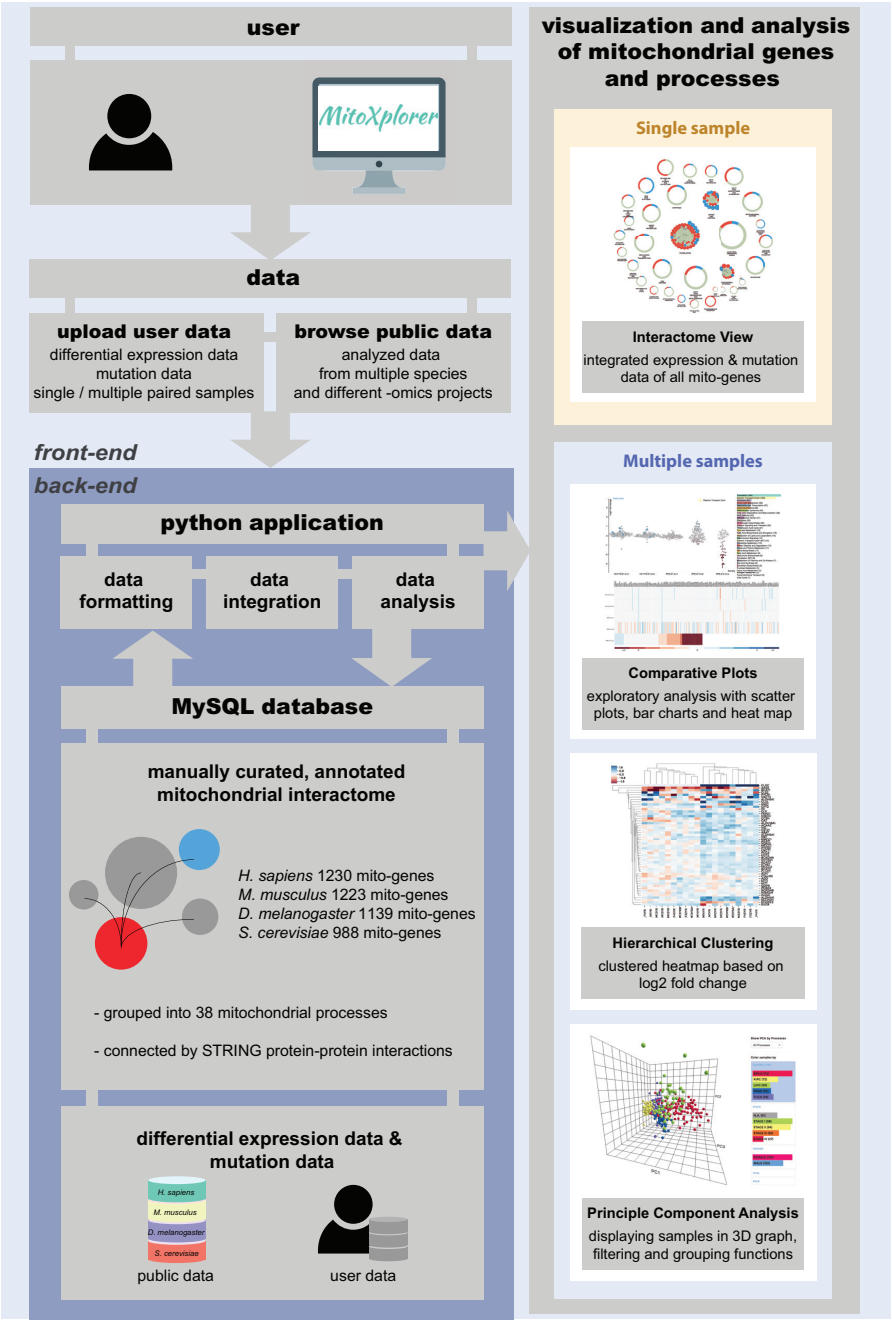


Figure 1. Setup of the mitoXplorer web-based visual data mining platform. A manually curated, annotated mitochondrial interactome represents the central part of the mitoXplorer software, for which we have assembled 1229 mito-genes in human, 1222 mito-genes in mouse, 1139 mito-genes in fruit fly and 988 mito-genes in budding yeast in 38 mitochondrial processes (mito-processes). We have connected gene products using protein-protein interactions from STRING (69). Publicly available expression and mutation data from repositories such as TCGA or GEO are provided for data integration, analysis and visualization and are stored together with species interactomes in a MySQL database. Users can provide their own data, which are temporarily stored and only accessible to the user. A set of Python-based scripts at the *back-end* of the platform handle data formatting, integration and analysis (Supplementary Figure S1). The user interacts with mitoXplorer via several visual interfaces to analyze, integrate and visualize his private, as well as public data. Four interactive visualization interfaces are offered: (i) the Interactome View allows at-a-glance visualization of the entire mitochondrial interactome of a single dataset (see Figure 2); (ii) comparative plots, consisting of a scatterplot and a sortable heatmap allows comparison of up to six datasets, whereby a single mito-process is analyzed at a time (see Figure 3); (iii) hierarchical clustering allows comparison of a large number of datasets which are clustered according to their expression values. Hierarchical clustering plots are zoom-able and interactive (see Figure 4); (iv) principal component analysis displays PCA-analyzed datasets in 3D, providing filtering and grouping functions. There is in principle no limit to the number of datasets that can be analyzed using PCA (see Figure 5).

one gene is part of only a single mito-process. We acknowledge that this annotation strategy has limitations, as a mito-gene can be part of more than one mito-process: as an example, in KEGG, the human gene GPI is part of the pathways Glycolysis, as well as Pentose phosphate pathway. In GeneCards, GPI is primarily associated with Glycolysis, thus we assigned it to the mito-process Glycolysis. For other genes, the existing annotations were less clear and we had to decide on the primary process a gene should belong to. For some mito-processes, we solved this by introducing the ontologically higher ranked term: for instance, mito-genes involved in fatty acid (FA) metabolic pathways can be associated with either the biosynthesis or degradation of FAs, or both: genes implicated in both processes were allocated to the term that would rank higher, namely Fatty acid metabolism. Genes involved in the transport of molecules across the mitochondrial membrane were divided into three groups: those involved in import & sorting of all mito-proteins that are encoded in the nuclear genome and translated in the cytosol into mitochondria; those that are part of the mitochondrial carrier family (68); and finally those transmembrane proteins that are involved in mitochondrial transmembrane transport and cannot be associated with either of the other two groups. Finally, to complete the mitochondrial interactome and reveal potential interactions between mito-processes, we added protein-protein interaction information from STRING (69) for all mito-genes.

Since we cannot guarantee that our current annotation is either complete or exempt from mis-annotations, and with the goal of nucleating a community-based effort to further complete and improve the annotation of the mitoXplorer mito-interactomes, we provide a 'FEEDBACK' page. Users can submit comments and suggestions on genes and their annotations using this page, suggest new genes and new mito-processes or provide any other feedback.

Mito-genes of human, mouse, *Drosophila* and budding yeast annotated with mito-processes are available in Supplementary Table S2A–D. The mito-interactomes can also be downloaded from the INTERACTOME page of mitoXplorer. These manually curated and annotated interactomes enable a meaningful analysis and visualization of mitochondrial expression dynamics of mito-genes and mito-processes by comparing differential expression of two or more conditions in mitoXplorer.

The mitoXplorer expression and mutation database

To foster the analysis of mitochondrial expression dynamics and mutations, mitoXplorer hosts expression and mutation data from public repositories in a MySQL database.

Expression data encompass analyzed data of differentially expressed genes from RNA-seq studies and are available in the form of $\log_2$ fold change (log2FC) and *P*-value. One differential dataset thus includes two experimental conditions with all replicates. Mutation data include analyzed data of identified SNPs of one sample against a publicly available reference genome or transcriptome.

Pre-analyzed public data are taken as provided by the authors of the respective study: information on software and genome version used for read mapping, as well as software and settings used for differential expression analysis can generally be found at the GEO-link of the respective project listed on the DATABASE page. We have ensured that only high-quality data with replicates, as well as a properly described analysis strategy are available in the mitoXplorer database. If only raw read counts were available, we analyzed the data using state-of-the-art software (DESeq2 (45), for details see Methods). Finally, whenever only normalized read counts were available, which is typical for large population-based studies, we calculated log2FC according to (70). It should be noted that due to the heterogeneity of the available formats of the provided data, the algorithms and their settings, as well as the genome version used for read mapping might differ for available projects in mitoXplorer.

The largest public resource imported into mitoXplorer covers publicly available expression data of human cancers from The Cancer Genome Atlas (TCGA) (1). We have included all paired samples. This resulted in a total of 523 differential datasets from six different cancer types: kidney cancer (KIRK), breast cancer (BRCA), liver cancer (LIHC), thyroid cancer (THCA), lung cancer (LUAD) and prostate cancer (PRAD). Changes in mitochondrial metabolism have been described in many cancer types (for a review, see (71)). As mitoXplorer is thus far the only resource that allows a focused analysis of mito-genes across different cancer types or patient groups, this resource should be especially useful to shed light on the expression dynamics of mito-genes in cancer and to classify the mitochondrial metabolic profiles of tumor types and sub-types. Users can moreover integrate proprietary data with differential expression data from different tumor types and subtypes.

We also uploaded expression data from cultivated fibroblasts of healthy human donors ranging from 1 to 94 years of age (48) (GEO accession: GSE113957). Since decline in mitochondrial quality and activity are well-known contributors to age-related conditions and diseases (72), this dataset should help uncover the contribution of altered mito-gene expression dynamics to the ageing process.

We made available several datasets from mouse knockout studies: we uploaded differential transcriptomic and proteomic data of five different mouse conditional heart knock-out strains of genes involved in mitochondrial replication, transcription and translation (40) (Lrpprc, Mterf4, Tfam, Polrmt, Twnk (Twinkle), (GEO accession: GSE96518)). These data are especially helpful in unraveling the transcriptional and post-transcriptional effects on mito-genes upon disruption of gene expression at different levels in mitochondria.

Furthermore, we added data from a mouse model of a known mitochondria-associated condition, the Barth syndrome. Barth syndrome patients develop severe cardiomyopathy (73). This syndrome is caused by mutations in or loss of the TAZ gene coding for the protein Tafazzin which is involved in cardiolipin biosynthesis (74). Failure of enzyme activity of Tafazzin leads to altered mitochondrial membrane composition, structure and metabolism (74,75). We provide differential expression data of Taz knock-out mouse embryonic fibroblasts (MEFs) compared to wild-type in normoxic and hypoxic conditions generated by

Table 1. Mito-processes and number of associated mito-genes in human, mouse, *Drosophila* and budding yeast

Mito-process	Human	Mouse	*Drosophila*	Budding yeast
Amino acid metabolism	81	79	67	44
Apoptosis	56	55	43	6
Bile acid synthesis	2	2	7	0
Calcium signaling & transport	23	23	12	4
Cardiolipin biosynthesis	6	6	5	5
Fatty acid biosynthesis & elongation	22	22	15	13
Fatty acid degradation & beta-oxidation	30	31	26	9
Fatty acid metabolism	15	13	20	8
Fe-S cluster biosynthesis	25	26	19	23
Folate & pterin metabolism	13	13	9	13
Fructose metabolism	7	7	3	14
Glycolysis	38	37	35	33
Heme biosynthesis	9	9	9	5
Import & sorting	51	51	61	55
Lipoic acid metabolism	3	3	4	3
Metabolism of lipids & lipoproteins	34	36	17	14
Metabolism of vitamins & co-factors	17	18	19	9
Mitochondrial carrier	46	45	46	23
Mitochondrial dynamics	61	59	48	39
Mitochondrial signaling	18	18	10	11
Mitophagy	21	21	13	11
Nitrogen metabolism	9	9	16	7
Nucleotide metabolism	15	15	12	23
Oxidative phosphorylation	167	164	173	115
Oxidative phosphorylation (MT)	13	13	13	9
Pentose phosphate pathway	7	7	6	14
Protein stability & degradation	27	27	20	25
Pyruvate metabolism	26	25	24	12
Replication & transcription	51	52	32	50
ROS defense	34	34	30	24
Transcription (nuclear)	24	24	25	6
Translation	185	184	192	210
Translation (MT)	24	24	24	37
Transmembrane transport	20	20	21	24
Tricarboxylic acid cycle	21	22	29	26
Ubiquinone biosynthesis	9	9	9	12
Unknown	12	12	20	46
UPRmt	7	7	4	6

Chowdhury *et al.* (46) (GSE accession: GSE119775); these data should help reveal the effect of Tafazzin loss of function during hypoxia.

To extend mitoXplorer to other model organisms, we added data from *D. melanogaster*, namely expression data from 185 wild-derived, inbred strains (males and females) from the *Drosophila* Genetics Reference Panel (DGRP2) (49): this set of lines stems from an out-crossed population in Raleigh, North Carolina. These wild-derived fly strains display a substantial quantitative genetic variation in gene expression. The availability of these data on mitoXplorer allows a focused analysis of mito-genes to elucidate whether mitochondrial expression dynamics is equally impacted in these strains.

Moreover, we have uploaded data from a recently published systematic study of flight muscle development in *D. melanogaster* (44) (GEO accession: GSE107247). This enables the analysis of mitochondrial expression dynamics during the development and differentiation of a tissue that is highly dependent on an efficient mitochondrial metabolism and especially ATP production for proper functioning.

Regarding budding yeast, we imported data from a time-course expression profiling experiment of meiosis of a synchronized cell culture (Hanlon SE, Lieb JD (unpublished), GEO accession: GSE75257), allowing users to mine the expression dynamics of mito-genes over 12h of sporulation. This project is the only microarray-based dataset we have uploaded on mitoXplorer.

Finally, we uploaded data from Protein Phosphates 2A (PP2A) yeast deletion strains; these strains show a diminished response of nuclear gene expression associated with mtDNA damage compared to wild-type (47) (GEO accession: GSE52242). This dataset should help shed light on the role of the conserved protein phosphatase PP2A in protecting cells from mtDNA damage.

To verify mitoXplorer predictions experimentally, we use data from human trisomy 21: we provide data from human trisomy 21 patients (GEO accession numbers: GSE55426; GSE79842; (42,43)), from trisomy 21 studies in mouse (GSE5542 (42), GSE79842 (43)), as well as differential datasets generated in the course of this study from human trisomic cell lines (11 datasets) which have been partially published elsewhere (51,52) (GEO accessions: GSE39768; GSE47830; GSE102855). These transcriptomic, as well as proteomic datasets should help understand the role of mitochondria and the mitochondrial metabolism in trisomy 21.

All available data can be viewed and accessed from the mitoXplorer DATABASE web-page.

User-provided expression and/or mutation data

Researchers can upload and explore their own data in mitoXplorer, given that they originate from one of the species contained in the mitoXplorer platform. Data must be pre-analyzed. Differential expression data must contain the dataset ID (describing the experimental condition), the gene name and the log2FC. Optional values include the *P*-value, as well as the averaged read counts (or intensities) of the replicates of the compared conditions. Mutation data must contain the dataset ID, gene name, the chromosome, the position, as well as reference and alternative allele. Optional values include the effect, as well as the consequence of the mutation. Users have the option to either generate their own data according to the format described on our website. We recommend to follow the best-practice-guidelines available for analyzing transcriptomic or proteomic data prior to uploading data to mitoXplorer (see for instance (76,77) for differential expression analysis or (78) for variant calling). Alternatively, users may use the RNA-seq pipeline for differential expression analysis and mutation calling that we provide at https://gitlab.com/habermannlab/mitox_rnaseq_pipeline/.

The entire list of genes from a study should be uploaded to the platform for several reasons: first, a restriction to only differentially expressed or mutated genes will suppress links between proteins in the interactome; second, an integration of user data with publicly provided data is difficult with incomplete datasets; third, mitoXplorer will automatically select the mito-genes from the user data. Uploaded data will be checked for correct formatting and integrated with the interactome of the chosen species. User data are only visible to the owner and are stored in the mitoXplorer MySQL database for 7 days. Users can integrate their own data with available public data on mitoXplorer to perform various analyses and visualizations as described below (Figure 1).

Analysis and visualization tools in mitoXplorer

The mitoXplorer web-platform provides a set of powerful, easy-to-read and highly interactive visualization tools to analyze and visualize public, as well as user-provided data by VDM (Figure 1): an **Interactome View** to analyze the overall expression and mutation dynamics of all mito-processes of a single dataset containing differentially expressed genes between two conditions and potential mutations in mito-genes; the **Comparative Plot**, consisting of an interactive scatterplot, as well as an interactive heatmap for comparing up to six datasets; the **Hierarchical Clustering**, as well as the **Principal Component Analysis** for comparative analysis of many datasets.

INTERACTOME VIEW

The Interactome View can be used to get an at-a-glance view of the overall expression dynamics of all mito-processes of a single dataset of differentially expressed mito-genes and potential mutations (Figure 2A). It allows users to identify the most prominently changed mito-processes or -genes in a dataset. The genes are grouped according to mito-processes and displayed in the process they are assigned to. The Interactome View is highly dynamic and can be adjusted by users to their needs.

When the Interactome View is launched, each mito-process is primarily shown as a grey circle with elements colored in grey, blue and/or red, indicating up- or down-regulated genes within the process, respectively (Figure 2A). Thus, mito-processes with the most up- or down-regulated genes can be quickly identified.

When clicking on a process name, its circle opens up to display all its member genes as bubbles. The size of the bubble relates to the strength of the differential regulation while the color indicates up- (blue) or down- (red) regulation of the gene (Figure 2B). If information about mutations is included in the dataset, this is indicated by a thicker, black border of the gene bubble.

Hovering over a gene will display the gene name, its function, its mito-process, the log2FC and the *P*-value of the differential expression analysis, as well as potential mutations in the *information panel* (Figure 2C). If a gene physically interacts with other mito-genes, hovering over it or over the process circle will in addition display these connections (Figure 2C). Thus, the user is immediately informed about the location and connectivity of the protein of interest within the mitochondrial interactome. Users can also search for specific genes using the 'FIND A GENE' box at the top of the page.

The Interactome View can be launched by clicking on the 'eye' symbol next to dataset names from the ANALYSIS page of mitoXplorer, after having chosen the organism, the project and the dataset. Alternatively, users can access single datasets from the DATABASE page of the platform, by clicking on the eye symbol of a listed dataset after having chosen a species, as well as a project. A new page will be opened for the Interactome View, which allows opening and comparing multiple datasets at the same time. This is especially useful for comparing the overall expression change of mito-processes of multiple datasets.

COMPARATIVE PLOT

The Comparative Plot visualization combines several interactive graphs to analyze one mito-process, allowing the comparison of up to six datasets. It includes a scatterplot with a dynamic *y*-axis, as well as an interactive heatmap at the bottom of the page. The mito-process to be visualized can be selected in the *process panel* (Figure 3A). Red and blue coloring of the dots and the heatmap indicates the directionality of differential expression (blue: upregulated; red: down-regulated); bright blue, larger gene bubbles in the scatterplot indicate mutations, if available from the dataset. This Comparative Plot offers an overview of the expression dynamics of all members of one mito-process for up to six individual datasets and thus can be helpful in identifying co-regulated genes e.g. in time-course data, patients or multiple mutant datasets.

Hovering over a gene bubble, or over a tile in the heatmap will again display the respective associated information of the gene in the *information panel* (gene name, function, mito-process, $\log_2$FC, *P*-value, potential mutations) (Figure 3B). The heatmap can be sorted according to the dataset, as well as the differential expression values within one

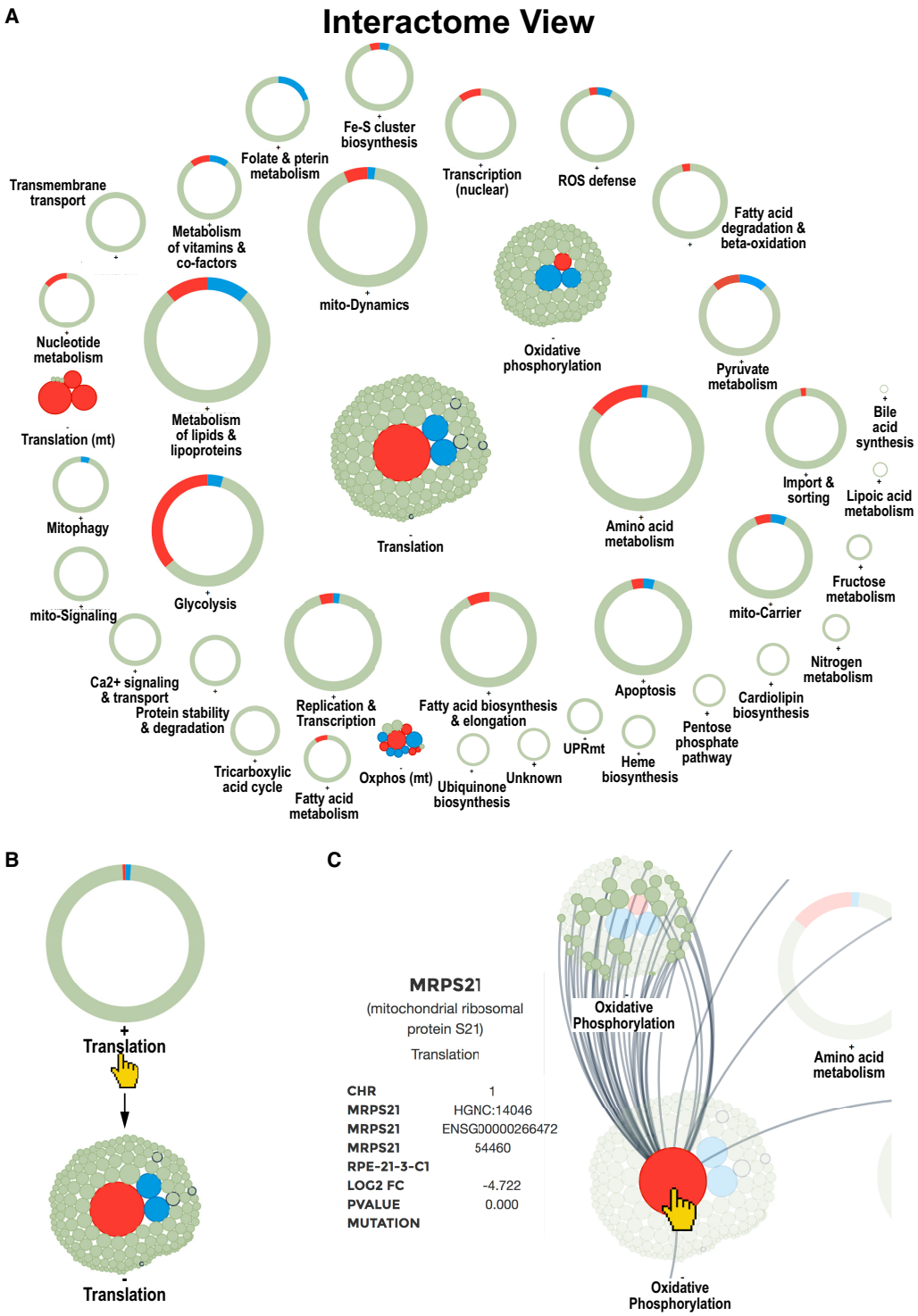


Figure 2. Interactome View of the mitoXplorer platform. (**A**) Overview of all mito-processes of one dataset. A process can either be shown as one circle with colored segments according to the number of dysregulated genes, or upon clicking on the process, by showing all individual genes being part of this process, see (**B**): by clicking on the process Translation or the adjacent '+', the circle is replaced by individual bubbles representing genes of this process. Clicking on the process again, or on the adjacent '–' will revert to the circular display. (**C**) Hovering over a gene bubble will display the name of the gene and associated information (gene name, description, chromosomal location, mitochondrial process, accession numbers, as well as $\log_2$ fold change, *P*-value and observed mutations), as well as all connections to mito-genes in other processes. Compared were the retinal epithelial cell line RPE1 (RPE) wild-type to RPE1 with Trisomy 21 (RPE_T21).

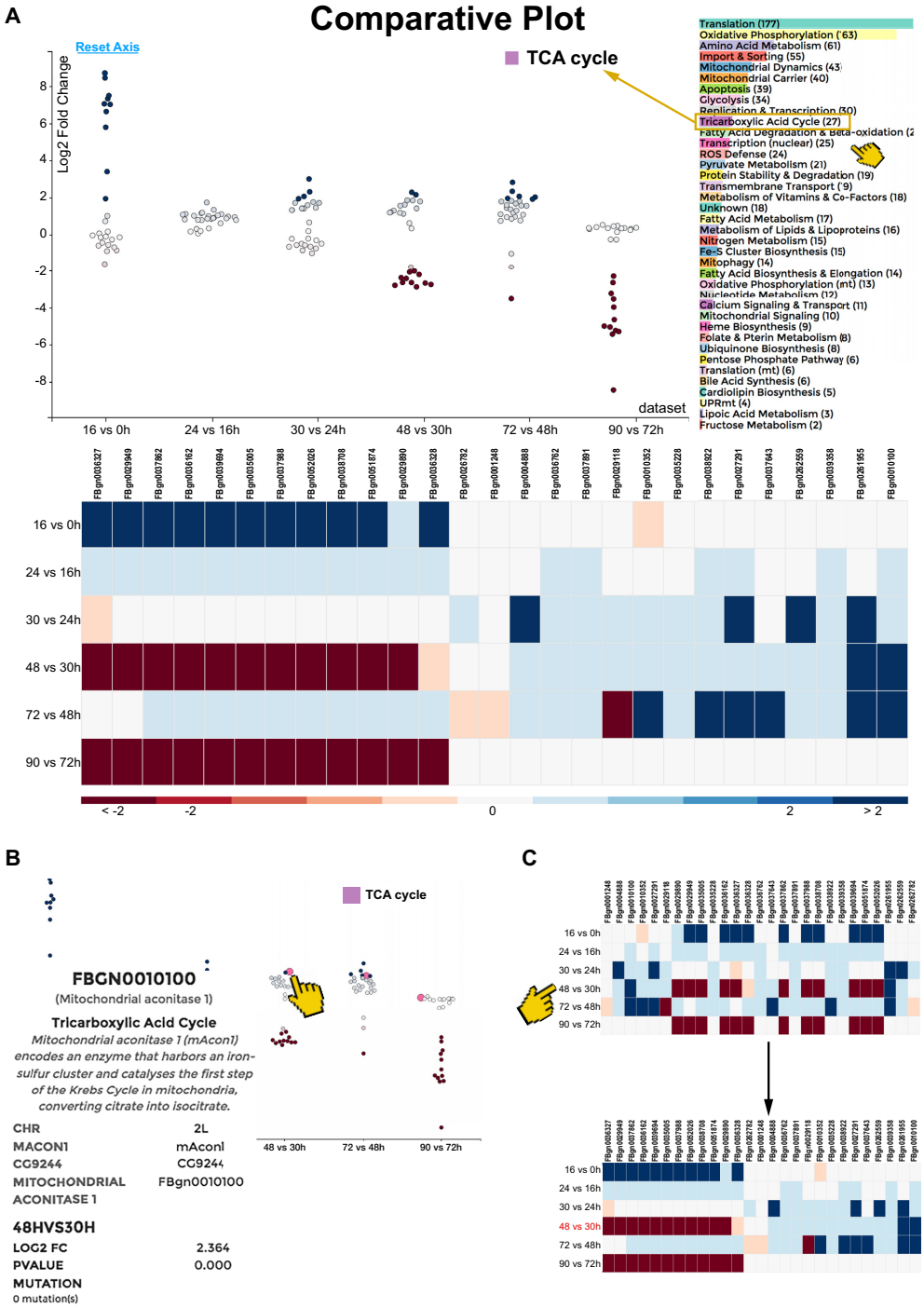


Figure 3. Comparative Plot of the mitoXplorer platform. (**A**) The Comparative Plot display is composed of a scatterplot, sortable heatmap and a bar chart for the selection of mito-processes. The scatterplot shows the $\log_2$ fold change (y-axis) and the datasets (x-axis). Each bubble represents one gene, whereby red bubbles indicate down-regulated, and blue ones upregulated genes. The process to be shown can be selected by clicking on the process name in the bar chart next to the scatterplot, the chosen process being indicated on its top. In this case, TCA cycle was chosen. The heatmap at the bottom shows the individual genes and the datasets, whereby the genes are colored according to their $\log_2$ fold change (indicated at the bottom of the plot). (**B**) Hovering over a gene bubble (or a gene tile in the heatmap) will display available information (in case of fly: gene name, mitochondrial process, gene description, chromosomal location, gene symbol, as well as $\log_2$ fold change, *P*-value and observed mutations). (**C**) The heatmap is sortable by $\log_2$ fold change (as indicated by the pointer in C), as well as by dataset. Clicking on one of the datasets will sort the heatmap according to the $\log_2$ fold change of all genes in this dataset, as is illustrated here. Clicking on one of the genes will sort the heatmap according to its $\log_2$ fold changes across different datasets. The time-series study of developing flight muscle (44) was used to demonstrate the functionality of this visualization method.

dataset (Figure 3C). The Comparative Plot is especially useful for performing a detailed, comparative, mito-process based analysis of differential expression dynamics between different datasets.

We applied this analysis method to visualize differential expression data from a time-series study of flight muscle development during pupal stages in *Drosophila* (44) (Figure 3). While enrichment analysis has revealed a general positive enrichment of processes like Tricarboxylic acid cycle (TCA Cycle) in the course of flight muscle development, mitoXplorer identifies 12 genes of TCA cycle that are co-regulated. This group of genes is strongly upregulated between 0 and 16 h after puparium formation (APF), when myoblasts divide and fuse to myotubes. The same group of genes is consecutively down-regulated in two phases at time-points 30–48 h and 72–90 h APF, when myotubes differentiate to mature muscle fibers. This is surprising as in mature muscle fibers the TCA cycle should be important for proper functioning. Their strong induction between the first two time-points could be responsible for downregulation at later stages.

THE HEATMAP: HIERARCHICAL CLUSTERING

Hierarchical Clustering visualization allows the analysis of up to 100 datasets, analyzing one process at a time. This creates a heatmap with mito-genes, as well as -datasets, which are clustered according to the $\log_2$FC using hierarchical clustering (Figure 4 a). The results are displayed as a clustered heatmap, with a dendrogram indicating the distance between datasets or between genes.

Hovering over a gene will display its associated information, as well as dataset information in the *information panel* (Figure 4B). The user can furthermore zoom into parts of the heatmap to get a more detailed view of the data. The heatmap is particularly useful for discovering groups of similarly regulated mito-genes or datasets within one mito-process.

We applied this visualization tool to display transcriptome and proteome data from a recent, systematic study of mouse conditional knock-out strains for five genes involved in mitochondrial replication (*Twinkle (Twnk)*), mtDNA maintenance (*Tfam*), mito-transcription (*Polrmt*), mito-mRNA maturation (*Lrpprc*) and mito-translation (*mTerf4*) (40). Interestingly, the expression dynamics of the mitochondrial transcriptomes and proteomes in heart tissue did not cluster together for the mutants, suggesting strong post-transcriptional effects or protein stability changes of mito-proteins upon the loss of any of these genes. In accordance with this, the expression of some mito-genes in the process Pyruvate metabolism that is shown here differs on transcriptome and proteome level. This demonstrates the usefulness of hierarchical clustering and the heatmap display in identifying the correlation or divergence between genes as well as datasets.

PRINCIPAL COMPONENT ANALYSIS

A larger number of datasets can be compared using Principal Component Analysis (PCA), either for an individual mito-process, or considering all mito-genes together (Figure 5A). In PCA, the expression value (e.g. $\log_2$FC) of each gene is considered as one dimension, and each dataset represents one data point. In the resulting 3D PCA plot, the three axes represent the first three principal components and each bubble represents one dataset. The PCA is again interactive. Mito-processes can be selected via a drop-down menu on the top of the page. The plot can be turned and moved in 3D and has a zooming function.

Hovering over a bubble will give the information associated with an individual dataset in the *information panel*, including the values of the first three principal components (Figure 5B). The information differs for each project chosen.

Individual datasets can be selected and colored via the *dataset panel* next to the plot (Figure 5C). For instance with data from TCGA the filter and coloring can be used to highlight or to limit the plot to data from different tumors, different tumor stages or according to any other additional information provided. The PCA is especially useful for analyzing a large number of datasets and displaying specific trends in sub-groups.

We used the PCA plot to visualize data from the TCGA for four cancer types stored in mitoXplorer in Figure 5A, whereby the colors of the bubbles represent the different tumor types. The plot clearly highlights the variance of the different tumor types. In particular, kidney and liver cancer are highly distinct with respect to the first three components of all mito-genes (Figure 5A).

GROUPS function

In order to allow a more detailed, gene-centered analysis of correlated datasets, we added the possibility to select and group datasets in the Heatmap and the PCA views. Groups of datasets can be compared against each other with the Comparative Plot, whereby the $\log_2$FC is averaged over the data within a group. This functionality is useful, for instance when different groups of donors with similar expression patterns should be compared to each other.

We demonstrate the usability of the GROUPS function in Supplementary Figure S2, where we analyzed the averaged expression patterns of ageing human fibroblasts from healthy donors from 1 to 94 years of age (48). A first analysis using the PCA plot revealed that individuals which are older than 80 years were separated from the rest of the donors (Supplementary Figure S2A).

We next applied the GROUPS function to analyze different age groups using the mitoXplorer Comparative Plot. We chose to group individuals based on age, whereby we generated six age groups from age 40 to 100 years. As the age group 80–90 years showed two distinct clusters, we split this group in individuals that cluster with younger donors (g1), and those that cluster with the age group over 90 (g2). Our analysis using mitoXplorer GROUPS helped reveal a strong downregulation of a substantial number of mito-genes in nearly all mito-processes starting from the age of 85 (Supplementary Figure S2B), suggesting a general mitochondrial decline in old age.

Taken together, mitoXplorer provides a versatile, interactive and integrative set of tools to visualize and analyze the expression dynamics as well as mutations of mito-genes

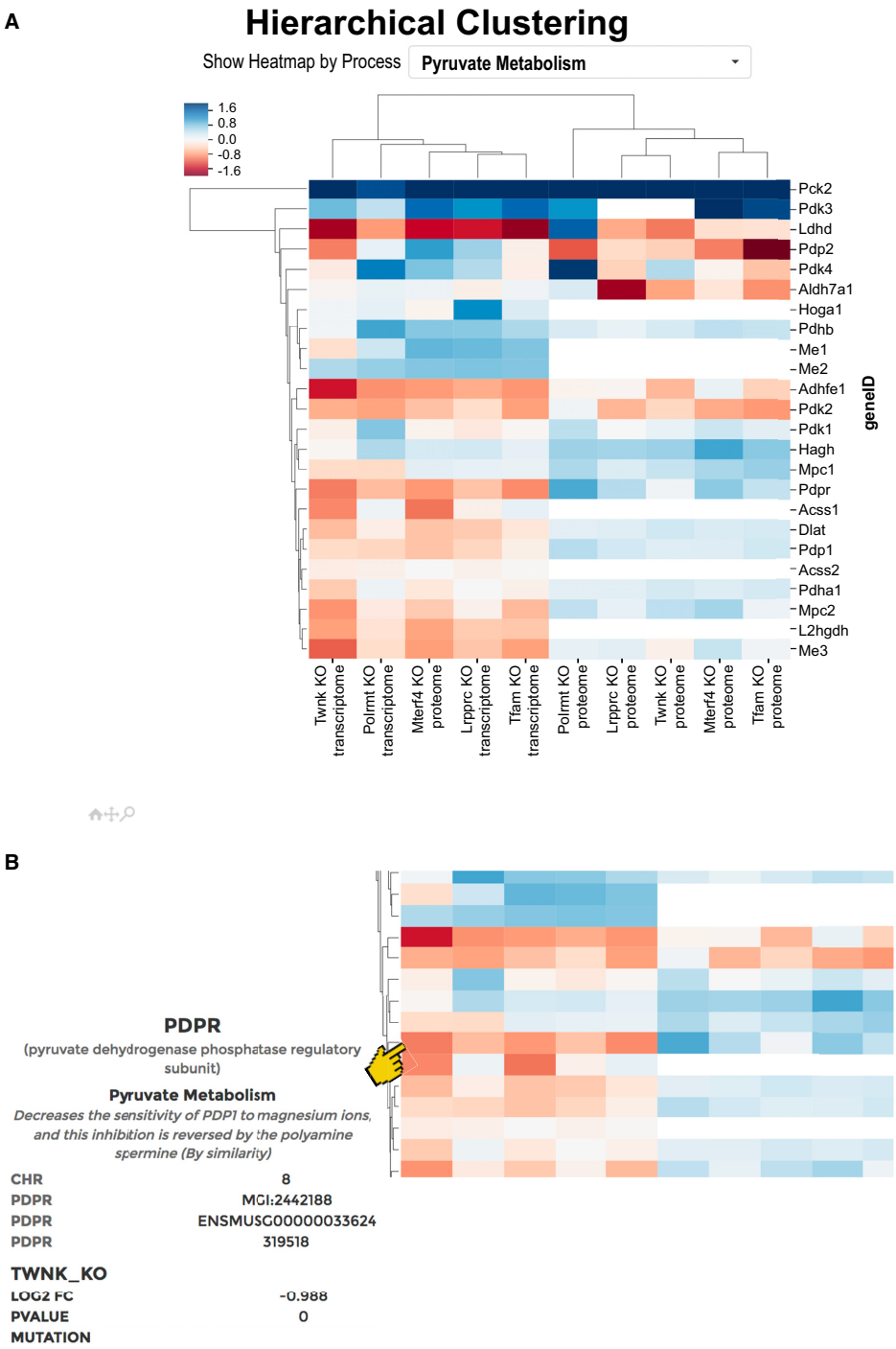


Figure 4. Hierarchical Clustering and heatmap plot of the mitoXplorer platform. Hierarchical Clustering of expression data results in a so-called heatmap. (**A**) Heatmap of transcriptome and proteome data of mouse knock-out strains of genes involved in mitochondrial replication, DNA-maintenance, transcription and RNA processing (taken from (40)). Data are clustered according to genes, as well as datasets. Gene tiles are colored according to their $\log_2$ fold change. At the top of the heatmap, the user can choose the mito-process to be displayed. (**B**) Hovering over one of the gene tiles will display information on the gene, such as the gene name, mito-process, $\log_2$ fold change, *P*-value and – if available – observed mutations. The heatmap is also zoom-able by clicking on the magnification glass at the bottom of the plot, so that large datasets can be visualized and analyzed efficiently. Datasets can be selected within the heatmap by grouping. To do this, first a group name has to be defined; second, the datasets belonging to this group have to be selected by clicking on one of the gene boxes of the dataset. This process can be repeated and the resulting groups can then be analyzed using Comparative Plots (see Supplementary Figure S2).

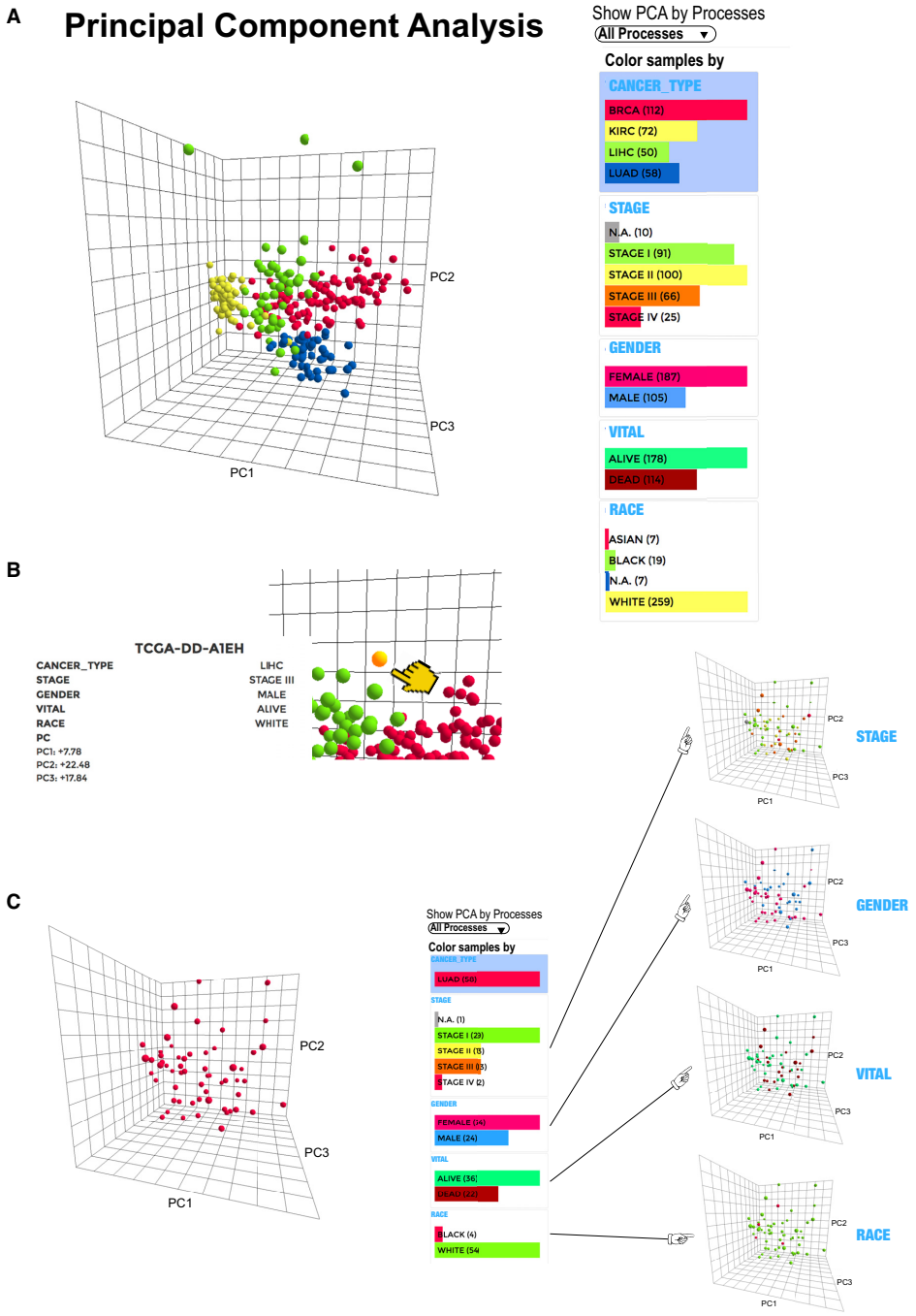


Figure 5. Principal component analysis and PCA plot of the mitoXplorer platform. (**A**) PCA analysis and plot of transcriptome data of The Cancer Genome Atlas (TCGA) database (1), showing four different cancer types: breast cancer (BRCA), kidney cancer (KIRK), liver cancer (LIHC) and lung cancer (LUAD). Each bubble represents one dataset, in this case, one cancer patient. At the right side at the top of the plot, the mito-process to be shown can be selected. In this case, 'All Processes' were selected, containing data from all mito-genes. At the right side next to the plot, different colors, as well as filters can be chosen. In this case, the Cancer Type was chosen for coloring, showing the four different cancer types in four different colors. (**B**) Hovering over a bubble will display associated information on the dataset, including the dataset name, and in case of the TCGA, information on the cancer type, the stage, the gender, the vital status, as well as skin color. In addition, the three PC components are shown. (**C**) Selecting color schemes on the right-hand side will change the coloring of the bubbles. In this case, only lung cancer is shown, and coloring is done according to Stage, Gender, Vital, and Skin color. This panel can also be used for selecting specific datasets. For instance, clicking on one of the stages will only display the chosen stage and omit datasets from other stages. As in the heatmap, datasets can be selected from the PCA for grouping. To do this, first a group name has to be defined; second, the datasets belonging to this group have to be selected by clicking on one of the dataset bubbles. This process can be repeated and the resulting groups can then be analyzed using Comparative Plots (see Supplementary Figure S2).

and mito-processes, facilitating a detailed understanding of observed changes at a molecular level.

Analyzing transcriptomes of mitochondria-associated health conditions using mitoXplorer

To demonstrate the analytical and predictive power of mitoXplorer, we explored available transcriptome data from health conditions associated with mitochondrial functions. We first performed mitoXplorer analysis of data from a mouse model of Barth syndrome (46), a mitochondrial disorder caused by mutations of the Taz gene which encodes the protein Tafazzin. Second, to show that we could verify mitoXplorer predictions experimentally, we analyzed our set of trisomy 21 data using mitoXplorer (51,52).

Analyzing the effects of Tafazzin loss of function in normoxic and hypoxic conditions using mitoXplorer

The Taz gene which encodes the cardiolipin acyl transferase Tafazzin is required for the remodeling of cardiolipin, an essential lipid component of the mitochondrial inner membrane (79). Loss of Tafazzin function leads to an abnormal fatty acid composition and a decrease in cardiolipin levels, resulting in abnormal mitochondrial morphology and dynamics, decreased stability of respiratory supercomplexes and increased oxidative stress (see (80) and references therein). Loss of Tafazzin function is also the primary cause of Barth syndrome (73), a rare, recessive, X-linked disorder that is characterized by cardiomyopathy, skeletal myopathy, growth retardation and neutropenia (81). Cardiolipin has been implicated in many mitochondrial processes, including mitochondrial protein import, mitochondrial carrier function, mitochondrial morphology and dynamics, respiratory chain function and metabolism (see (82) and references therein). Chowdhury *et al.* (46) subjected Tafazzin-mutant mouse embryonic fibroblasts (MEFs) to hypoxic stress to unravel the mechanism of impaired hypoxia-response in this cellular model of Barth syndrome. They observed that in hypoxia, a reduction of ROS levels in Tafazzin-deficient cells prohibited the induction of the NF-κB pathway, resulting in reduced Hif1α expression levels and subsequently the inability to respond to hypoxia.

Using the mitoXplorer Interactome View, we first could show that loss of Tafazzin function leads to a substantial perturbation of mito-gene expression in normoxic conditions (Supplementary Figure S3A), among which are the mito-processes Mitochondrial dynamics and Mitochondrial carrier (Figure 6A, B). In hypoxia, the expression profile of Tafazzin-deficient MEFs showed markedly different expression dynamics from wild-type cells (Supplementary Figure S3B).

In agreement with the findings from Chowdhury *et al.*, using mitoXplorer we found Hif1α down-regulated in Taz-deficient cells under normoxic and hypoxic conditions, while it was induced in hypoxia in wild-type cells (Figure 6C). Using Harmonizome (83), we also found a number of predicted and verified Hif1α target genes, some of which were de-regulated under hypoxic conditions in the Tafazzin-deficient cells (marked by an asterisk in Figure 6A–D). Next to Hif1α we found Yap1 de-regulated with the same expression dynamics. Yap1 is part of the Hippo signaling pathway and was shown to be stimulated in hypoxia (84), where it binds to and stabilizes Hif1α. As shown in Figure 6C, Yap1 was down-regulated upon loss of Tafazzin function in normoxic and hypoxic conditions, while it was upregulated in hypoxia in wild-type cells.

mitoXplorer analysis also revealed the failure of inducing RelA in hypoxia in Tafazzin-deficient cells. This transcription factor is a member of the Nf-κB family and was shown to be induced and activated early in hypoxia to modulate NF-κB target gene expression. The NF-κB dependent response to hypoxia was equally impaired upon loss of Tafazzin function.

Chowdhury and colleagues proposed that reduced ROS levels in response to hypoxia prevent proper NF-κB activation in Tafazzin-deficient cells. We thus investigated the expression dynamics of mito-genes in ROS defense using mitoXplorer and found that the Mitochondrial Methionine Sulfoxide Reductase B2 (MsrB2) was strongly induced in Tafazzin-deficient MEFs (log_2FC of 6.67 in normoxic and 5.13 in hypoxic conditions, respectively), whereas MsrB2 levels remained constant in wild-type cells in hypoxia (Figure 6D). MsrB2 reduces methionine (R)-sulfoxide to methionine and thus decreases reactive oxygen species in the cell due to its quenching properties; thus, its overexpression protects cells from oxidative stress (85). Its strong induction could be responsible for reduced ROS levels observed in Tafazzin-deficient cells.

In conclusion, using mitoXplorer we could not only identify previously described mito-processes affected by loss of Taffazin in normoxic conditions and confirm expression changes of Hif1α in Tafazzin-deficient cells in hypoxia. The expression profiles revealed by mitoXplorer analysis furthermore suggest that loss of Yap1 contributes to the observed phenotype by de-stabilizing Hif1α. Moreover, our analysis using mitoXplorer indicates that massive induction of the MsrB2 gene could be responsible for reduced ROS levels in Tafazzin-deficient cells in hypoxia, leading to a failure of induction of the NF-κB pathway and the transcription factor RelA.

Analyzing cell lines carrying trisomy 21 using the mitoXplorer platform

We next wanted to experimentally verify predictions made with mitoXplorer. To this end, we analyzed the transcriptome and proteome of a set of aneuploid cell lines carrying an extra copy of chromosome 21 (trisomy 21, T21). Mitochondrial dysfunction has been repeatedly found in T21 patients, whereby mostly oxidative stress, as well as—potentially resulting—mitochondrial respiratory deficiency have been shown to contribute to some of the observed clinical features (see for instance (86–99)). Transcriptome studies of different T21 tissues using microarrays (100–110) and more recently RNA sequencing (42,43,111) and proteomics (41,112–115) have revealed a complex picture of gene expression changes, with a marked dissimilarity in differential expression of mito-genes on mRNA and protein levels, indicating a potential post-transcriptional regulatory effect of some mito-genes in T21 (41). Yet, mito-gene and protein expression data in different tissues or under varying conditions in T21 remain sparse and a coherent

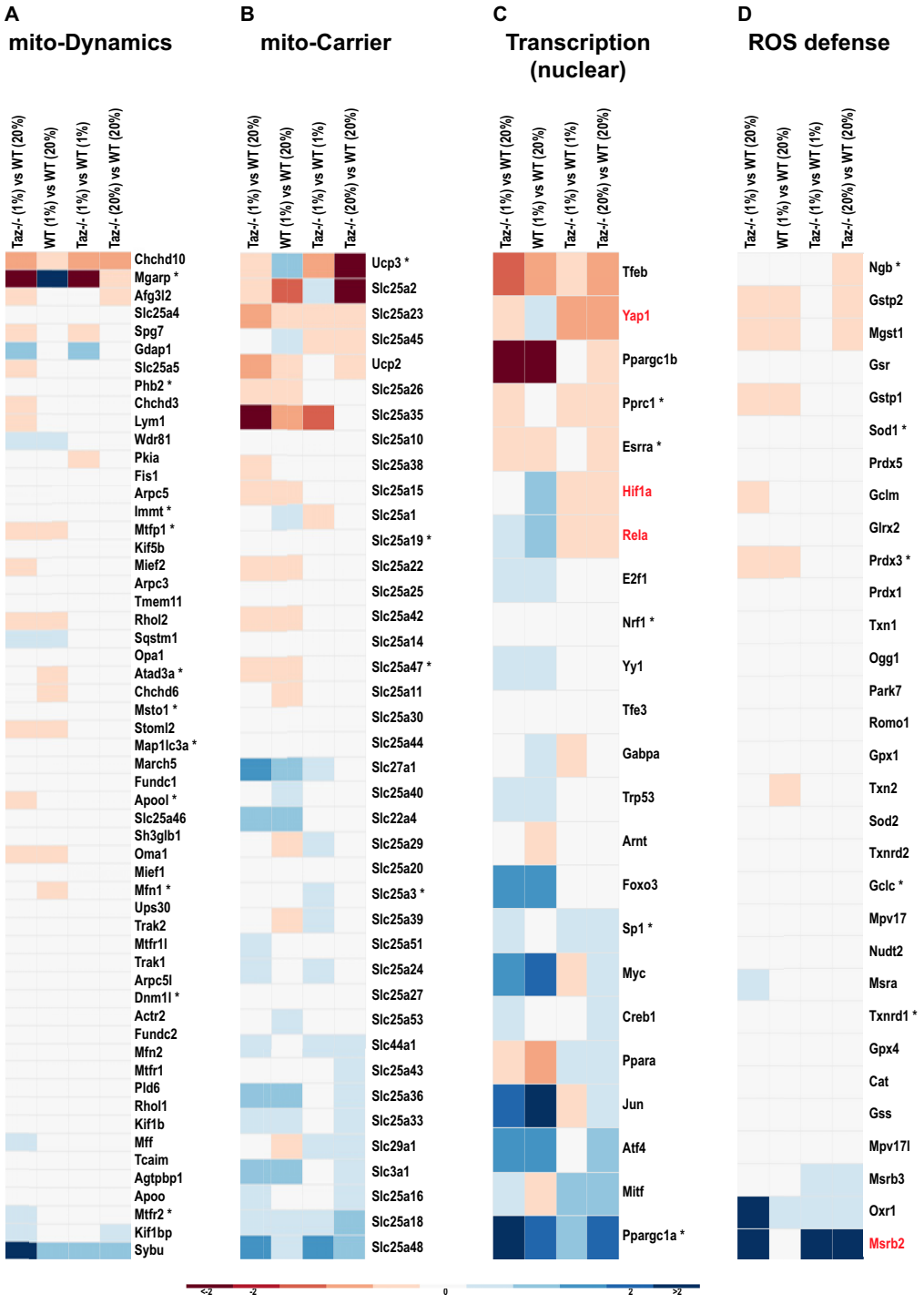


Figure 6. Mito-gene expression dynamics of Tafazzin-deficient mouse embryonic fibroblasts in normoxic and hypoxic conditions. The mito-processes Mitochondrial dynamics (mito-Dynamics), Mitochondrial carrier (mito-Carrier), Transcription (nuclear) and ROS defense are shown. (**A**) Among the genes involved in mito-Dynamics, the predicted Hif1α target Mgarp is strongly down-regulated in Tafazzin-deficient cells and fails to be induced in hypoxia. Mgarp is required for movement of mitochondria along microtubules during hypoxia (126). (**B**) The mitochondrial carriers Ucp3, a predicted Hif1α target, as well as Slc25a45 are differentially regulated in Tafazzin-deficient cells in hypoxia. (**C**) The transcription factors Hif1α, Yap1 (Hippo pathway) and Rela (NF-κB pathway)—all highlighted in red—are differentially regulated in Tafazzin knock-out cells in hypoxia compared to wild-type: they are induced in hypoxia and are down-regulated in Tafazzin-mutant cells in normoxic, as well as hypoxic conditions. Mitf and Ppara show opposite expression behavior, being induced in Tafazzin-deficient cells and down-regulated in hypoxia in wild-type cells. (**D**) The ROS defense gene MsrB2 (marked in red) is strongly induced in Tafazzin-deficient cells, while it remains constant in hypoxia in wild-type conditions. MsrB2 has ROS quenching properties.

hypothesis of the underlying mechanisms leading to the mitochondrial deficiencies in T21 patients is still missing.

We used trisomy 21 cell lines derived from either the euploid human colon cancer cell line HCT116 or from the retinal pigmented epithelial cell line RPE1, to which an extra copy of chromosome 21 was added (51). We used two RPE1-derived and two HCT116-derived clones trisomic for chromosome 21 (Supplementary Table S3 a), which were validated by fluorescent *in situ* hybridization and by whole genome sequencing. We used transcriptomic data of the original euploid RPE1 line and its two trisomic derivatives (RPE_T21 clone 1 and 2 (c1, c2) (51)), as well as for HCT116, and its trisomic derivatives (HCT_T21 (c1, c3)). We included proteomic data for RPE1 and one of its T21 derivatives (RPE_T21 c1). We performed bioinformatic analysis to determine differential expression of the above conditions (Supplementary Table S3B–E, for details on bioinformatic data analysis see Materials and Methods) and uploaded the differential expression data of the transcriptome and proteome on the mitoXplorer platform for further in-depth, mitochondrial analysis.

Differences between trisomy 21 cell lines

MitoXplorer analysis of data comparing HCT116- and RPE1-derived T21 cell lines using the Interactome View revealed that T21 induced strong effects with respect to the overall expression changes in mito-genes (Figure 7). HCT_T21 showed a subtle, but consistent up-regulation of mito-genes (Figure 7A). In contrast, RPE_T21 cells showed a strong downregulation of a few genes involved in several mito-processes, such as Fatty acid metabolism, Glycolysis or Mitochondrial dynamics (Figure 7B). Remarkably, quantitative proteome data from RPE_T21 c1 cells suggested that all mitochondria-encoded genes involved in OXPHOS, as well as the majority of nuclear-encoded OXPHOS-genes are down-regulated (Figure 7C). In conclusion, mitoXplorer analysis facilitated the finding of significant differences in mito-gene expression between the different cell lines. Importantly in RPE_T21 cells, proteome data showed a remarkable difference to transcriptome data.

mitoXplorer analysis suggests mitochondrial ribosomal assembly defects in RPE_T21 cell lines

To investigate the differences further, we next performed a more detailed analysis of expression changes in these T21 cell lines using Comparative Plots in mitoXplorer. Transcriptome and proteome data from RPE_T21, but not from HCT_T21 cell lines revealed that several subunits of the small mitochondrial ribosome (mitoribosome) were significantly down-regulated on either RNA or protein level, or both (Figure 8A). MRPS21 was strongly reduced on RNA- and protein-level. The genes MRPS33, MRPS14 and MRPS15 were largely normal on RNA level, while their protein levels decreased more than 2-fold (log2FC: MRPS33: -2.147; MRPS14: −1.827; MRPS15: −1.057). Mitoribosomal subunits are encoded in the nuclear genome and their protein products are imported into the mitochondria, where they assemble with mitochondrial ribosomal RNAs to form the large and small subunits of the mitoribosome. The mitoribosome is responsible for translating the 13 mt-mRNAs encoded in the mitochondrial genome, all of which code for key subunits of the respiratory chain required for OXPHOS (116,117). In accordance with a disrupted mitochondrial translation machinery, all quantifiable mitochondria-encoded OXPHOS proteins (Complex I: MT-ND1 and MT-ND5; Complex IV: MT-CO2) were severely diminished on protein-, but not on RNA-level in RPE1_T21 cells (Figure 8B).

Interestingly, 36 of the quantifiable OXPHOS proteins encoded in the nuclear genome were also found to be down-regulated at the proteome, but not at transcriptome level in RPE_T21 cells (Figure 8C). These include subunits of the NADH dehydrogenase (complex I), ubiquinol−cytochrome *c* reductase (complex III) and cytochrome c oxidase (complex IV). It is important to note that there is no general downregulation of mitochondrial proteins in these cells and only a few, specific proteins are strongly down-regulated (Figure 7C). Together, these data demonstrate the power of mitoXplorer to help identify the cause of important changes in mito-gene expression, here the downregulation of mitoribosomal subunits at the transcription level and the resulting consequences, in this case the downregulation of the majority of OXPHOS proteins.

RPE_T21 cells are defective in oxidative phosphorylation

The massive downregulation of OXPHOS proteins in RPE_T21 cells suggests that these cells should suffer from a severe OXPHOS deficiency. To test this hypothesis experimentally, we analyzed cellular respiration and glycolysis in T21 cell lines using a Seahorse XF96 analyzer to quantify oxygen consumption rate (OCR) as an indicator of mitochondrial respiration (Figure 9A−D, F), as well as the proton production rate (PPR) as an indicator of glycolysis (Figure 9E, G). In intact RPE_T21 cells, we indeed observed dramatically reduced levels of cellular respiration in comparison to the diploid control (Figure 9A).

As a complex I deficiency has been reported in trisomy 21 patients (95), we next asked whether RPE_T21 cells selectively suffer from a complex I deficiency, or whether the entire respiratory chain is affected, as suggested by our proteomic data. We used permeabilized cells to test each individual complex with the Seahorse analyzer, supplementing with pyruvate/malate, succinate and TMPD/ascorbate for assessing complex I, II or IV functionality, respectively. As expected from our proteomic analysis, RPE_T21 cells displayed a severe deficiency of the entire respiratory chain (Figure 9B−D). The glycolytic rate of RPE_T21 cells in the presence of glucose was similar to the diploid control cells. Inhibition of ATP-production was not able to stimulate the cells to a higher glycolytic rate (Figure 9E), which agrees with the already low OXPHOS levels observed in these cells. HCT_T21 cells, on the other hand, displayed normal respiration, as well as glycolysis (Figure 9F, G). This suggests that the respiratory chain, as well as the mitochondrial translational machinery is not generally affected in all T21 cells. Taken together, mitoXplorer helped uncover OXPHOS deficiencies in RPE_T21 cells, which we verified experimentally, demonstrating the power of an in-depth analysis of mitochondrial expression dynamics to identify the potential molecular cause of the observed phenotype.

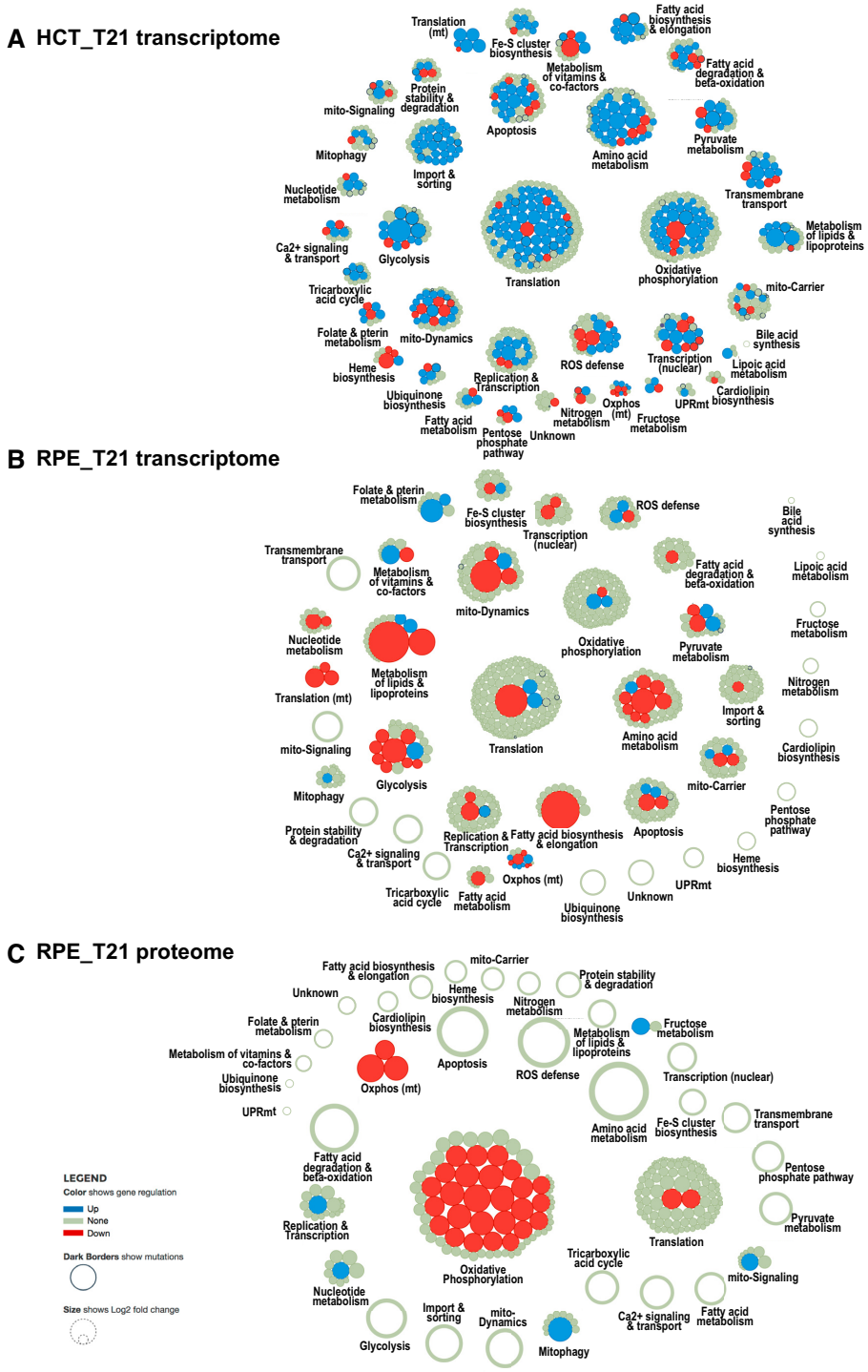


Figure 7. Interactome View of the transcriptome and proteome of cell lines carrying trisomy 21. Trisomy 21 samples were compared against their wild-type counterpart. Transcriptomic analysis of (**A**) HCT116_T21 (trisomy 21 against wild-type, c3) and (**B**) RPE21_T21 (trisomic against wild-type, c1); (**C**) proteomic analysis of RPE_T21 cells (trisomy 21 against wild-type, c1). Transcriptome changes are different between the two trisomy 21 cell lines HCT116 and RPE1. Expression changes of HCT_T21 cells are mild and genes tend to be upregulated (**A**), while some genes are strongly down-regulated in RPE_T21 cells (**B**). The transcriptome (**B**) and the proteome (**C**) of RPE_T21 cells respond quite differently, with a strong downregulation of components of the process Oxidative phosphorylation (OXPHOS) at proteome level which is not observed on transcriptome level. Most genes differentially expressed at transcript level, on the other hand, show no significant changes on proteome level. Red bubbles indicate downregulation, blue ones indicate upregulated genes. The size of the bubble corresponds to the $\log_2$ fold change.

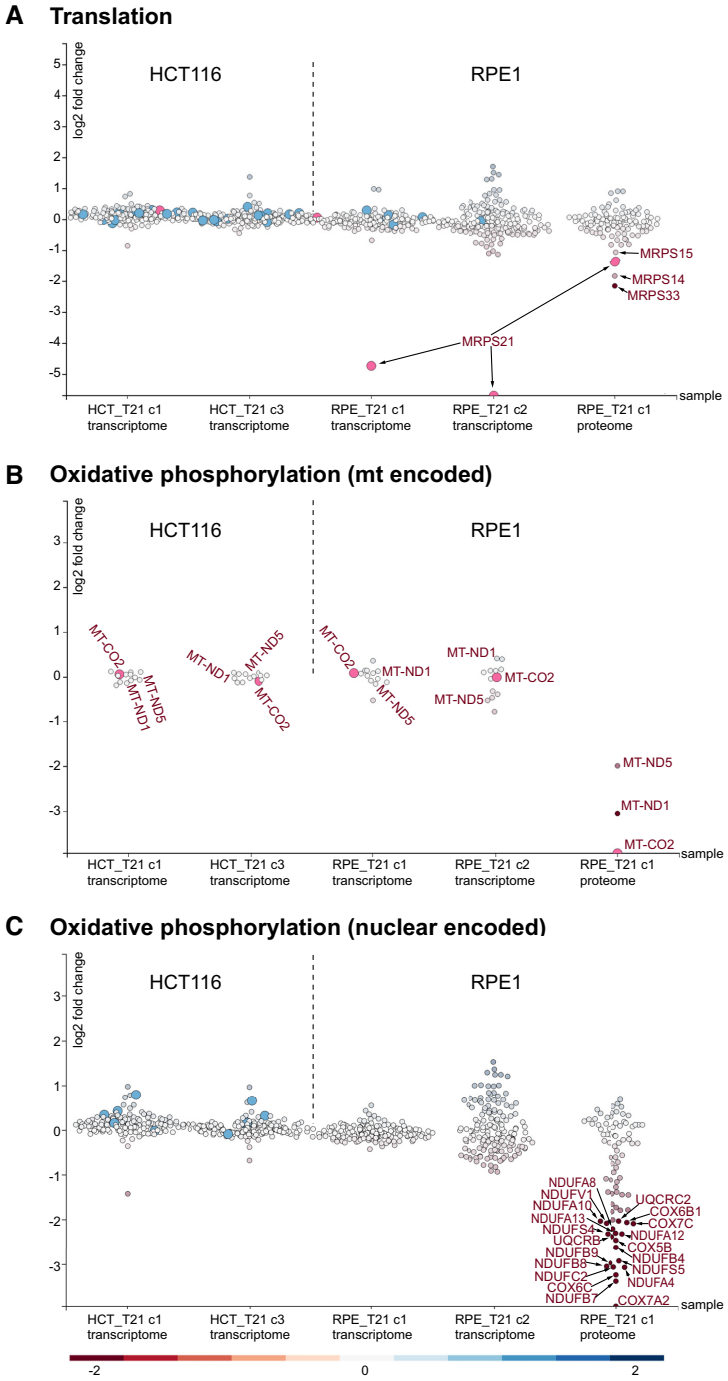


Figure 8. Scatterplots of Translation, mitochondrial- as well as nuclear components of Oxidative phosphorylation of trisomy 21 cells. (**A**) In the mitochondrial process Translation, MRPS21 is strongly down-regulated on transcriptome level in RPE_T21 cells as compared to RPE1 wild-type (wt) cells. No change is observed in HCT_T21 cells. On proteome level, several components of the mitoribosome small subunit (SSU) are down-regulated in RPE_T21 cells. (**B**) Transcript levels of mitochondria-encoded genes of Oxidative phosphorylation (OXPHOS) are not affected, while those genes are strongly diminished on protein level. (**C**) A significant number of components of OXPHOS are down-regulated on protein level in RPE_T21 cells, while no significant or only mild reduction can be observed on transcriptome level in trisomy 21 cell lines. Scatterplots are taken from the mitoXplorer Comparative Plot interface. Each bubble represents one gene, light blue dots indicate mutated genes. On the y-axis, the log2 fold change is plotted, the cell lines (transcriptome of HCT116 T21 (HCT_T21) clone 1 (c1) and clone 3 (c3) versus wild-type, as well as transcriptome of RPE1 T21 (RPE_T21) clone 1 (c1) and clone 2 (c2) versus wild-type and proteome of RPE1 T21 clone 1 (RPE_T21 c1) versus wild-type) are plotted on the x-axis. The gene highlighted in pink has been selected on the web-server: MRPS21 for the process Translation; MT-CO_2 for the process Oxidative phosphorylation (mt); and no gene has been selected in the process Oxidative phosphorylation.

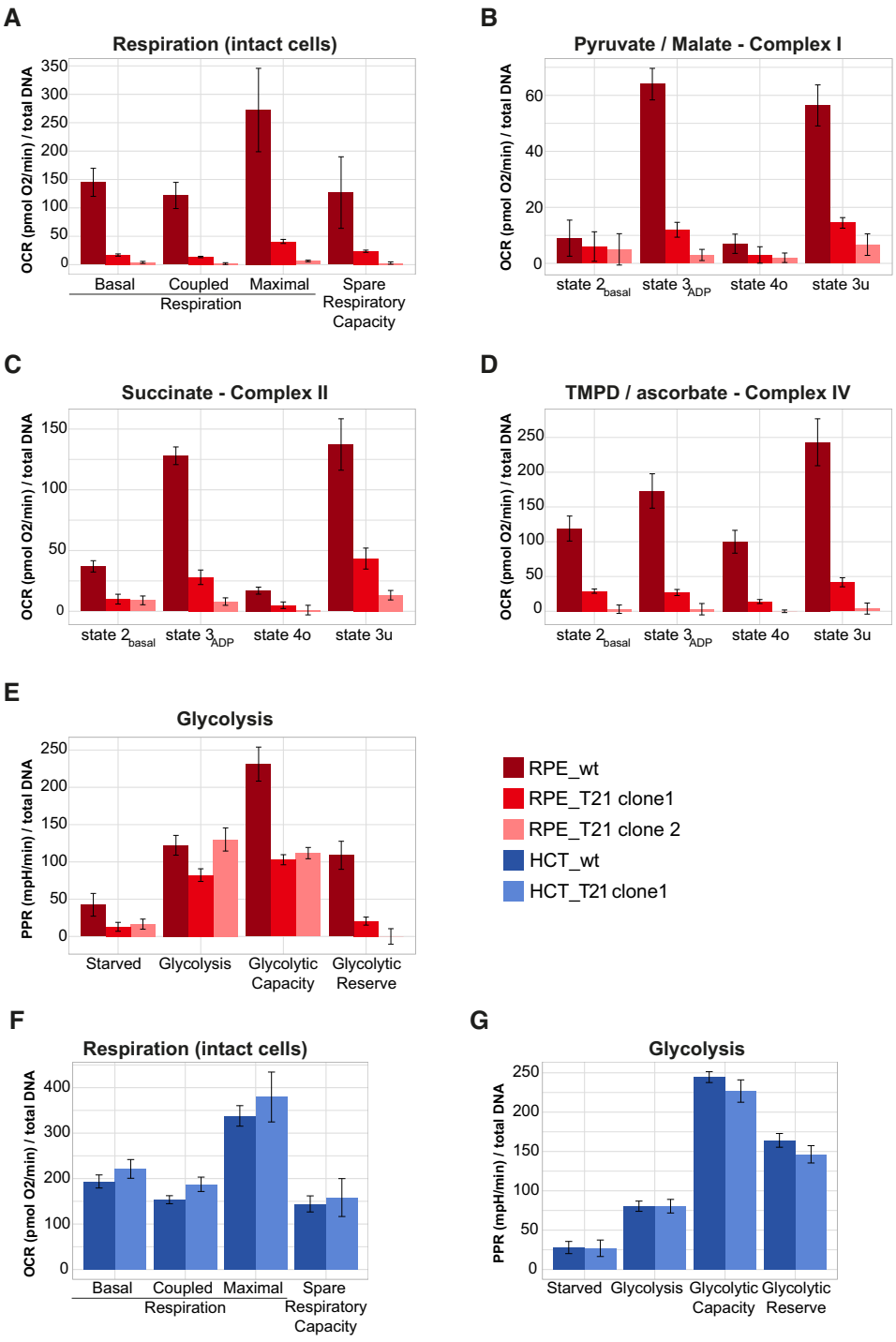


Figure 9. Mitochondrial respiration and glycolysis is strongly affected in RPE_T21 cells and not affected in HCT_T21 cells. (**A**) Respiration in intact RPE_T21 cells is greatly decreased compared to wild-type. (**B–D**) Permeabilized RPE_T21 cells supplemented for substrates of complex I, II and IV as indicated in the header of each plot, showed equally dysfunctional OXPHOS, suggesting a general break-down of the respiratory chain. (**E**) RPE_T21 cells do not have any spare glycolytic reserve. Respiration (**F**), as well as glycolysis (**G**) is virtually unchanged in HCT_T21 cells compared to their wild-type counterparts. Bright red: RPE_T21 clone 1; light red: RPE_T21 clone 2; dark red: RPE wild-type; dark blue: HCT wild-type; light blue: HCT_T21 clone 1. Measurements of cellular respiration in intact and permeabilized cells, as well as glycolytic potential were done using the Seahorse Bioscience XF Extracellular Flux Analyzer (Seahorse Biosciences). The experiments were performed using the mitochondrial and glycolytic stress test assay protocol as suggested by the manufacturer; the rate of cellular oxidative phosphorylation (oxygen consumption rate (OCR)) and glycolysis (cellular proton production rate (PPR)) were measured simultaneously.

Quantification of mitochondrial network morphology using mitoMorph

We further wanted to investigate, if T21 and the defective OXPHOS had a consequence on mitochondrial morphology and the mitochondrial network structure was changed in T21 cell lines. To quantify mitochondrial morphology in RPE_T21 cells, we stained mitochondria using the MitoTracker Deep Red FM dye. In order to quantify the characteristics of mitochondrial morphology, we developed a new Fiji plugin for quantification of mitochondrial network features, which we called mitoMorph. MitoMorph is based on the scripts provided by Leonard *et al.* (60) for quantifying mitochondrial network features such as *filaments* (corresponding to filamentous networked structures longer than 11 μm), *rods* (corresponding to filamentous networked structures shorter than 11 μm), *puncta* (corresponding to round structures below a radius of 0.6 μm) and *swollen* (corresponding to round structures above a radius of 0.6 μm) mitochondrial structures (see Methods for implementation details). MitoMorph reports the percentages of filaments, rods, puncta and swollen for each individual cell, as well as for all selected cells in a batch analysis (Figure 10A, B). Moreover, it provides the lengths and areas of filaments and rods. Figure 10C–F shows the distribution of mitochondrial network features for the two wild-type and T21 cell lines. MitoMorph analysis revealed that in both backgrounds, T21 cells had fewer mitochondrial filaments than their wild-type counterparts, but instead possessed a slightly higher number of rods, which was significant in HCT_T21 cells. Both T21 cell lines had significantly more swollen structures than their wild-type counterparts. Length and area distribution of filaments and rods were not significantly different between the wild-type and the trisomy 21 cells (Supplementary Figure S4A–D). We looked at expression dynamics of mito-genes associated with the process Mitochondrial dynamics using mitoXplorer. The only gene that is consistently, though only mildly down-regulated in both RPE_T21 clones is GDAP1 (Supplementary Figure S5). GDAP1 was shown to regulate the mitochondrial network by promoting mitochondrial fission (118). Its down-regulation could be contributing to or be a consequence of the observed phenotype. In conclusion, mitochondrial morphology based on light-microscopy is mildly affected in trisomy 21.

Data integration with publicly available trisomy 21 datasets

After discovering this differential OXPHOS defect in our RPE_T21 cell lines, we were interested in the overlap of the mito-transcriptome and -proteome of RPE_T21 cells with data from trisomy 21 patients. We used proteomic and transcriptomic data from a monozygotic twin study discordant for chromosome 21 (41,42). In agreement with our RPE_T21 data, systematic proteome and proteostasis profiling of fibroblasts from monozygotic twins discordant for T21 revealed a significant, although milder downregulation of the mitochondrial proteome, including proteins involved in OXPHOS, which is not apparent from transcriptomic analysis of the same cells (see Supplementary Figure S6A, B).

We next looked at proteomic data of fibroblasts from 11 unrelated individuals with trisomy 21 (41) (Supplementary Figure S6C–E). Virtually all T21 patients showed reduction in at least a few mitochondrial- and nuclear-encoded subunits of the respiratory chain (Supplementary Figure S6D, E). However, we could not confirm the strong reduction of the MRPS21 protein in all individuals. The only measurable mitoribosome subunit that was consistently, though in some cases only mildly, down-regulated was MRPL19 (Supplementary Figure S6C). Taken together, though the precise molecular mechanisms remain elusive, our analysis of these datasets with mitoXplorer nevertheless suggests a post-transcriptional effect leading to reduced expression levels of proteins involved in OXPHOS in trisomy 21.

DISCUSSION

The web-based mitoXplorer platform for mito-centric data exploration

MitoXplorer is a practical web tool with an intuitive interface for users who wish to gain insight from -omics data in mitochondrial functions. It is the first tool that takes advantage of the breadth of -omics data available to date to explore expression variability of mito-genes and -processes. It does so by integrating a hand-curated, annotated mitochondrial interactome with -omics data available in public databases or provided by the user.

MitoXplorer has been conceived and implemented as a visual data mining (VDM) platform: by iteratively interacting, visualizing and by allowing manipulation of the graphical display of data, the user can effectively discover complex data to extract knowledge and gain deeper understanding of the data. MitoXplorer provides a set of particularly interactive and flexible visualization tools, with a fine-grained, function- as well as gene-based resolution of the data. Clustering, as well as PCA-analysis help in addition to mine a larger number of -omics data effectively by grouping datasets with similar expression patterns.

VDM-based knowledge discovery is offered by a large number of resources and platforms. However, to the best of our knowledge, no currently available tool allows to explore expression variation of a specific subset of genes in a large number of -omics datasets. It permits users to exploit publicly available transcriptome, proteome or mutation data to study the variation and thus, the adaptability of a defined gene set in different conditions or species. While mitoXplorer offers the exploration of mito-genes, we have designed the platform in such a way that users interested in a different gene group can download a local version of mitoXplorer and upload their own interactome, which may contain any gene group of interest. Thus, mitoXplorer can be flexibly adjusted to any user-defined gene set.

Identification of putative causes of ROS-downregulation in Tafazzin-deficient cells using mitoXplorer

We have analyzed data from a mouse model of Barth syndrome to demonstrate that mitoXplorer can help identify de-regulated pathways in mitochondria-associated diseases. Barth syndrome results from a dis-balance of cardiolipin

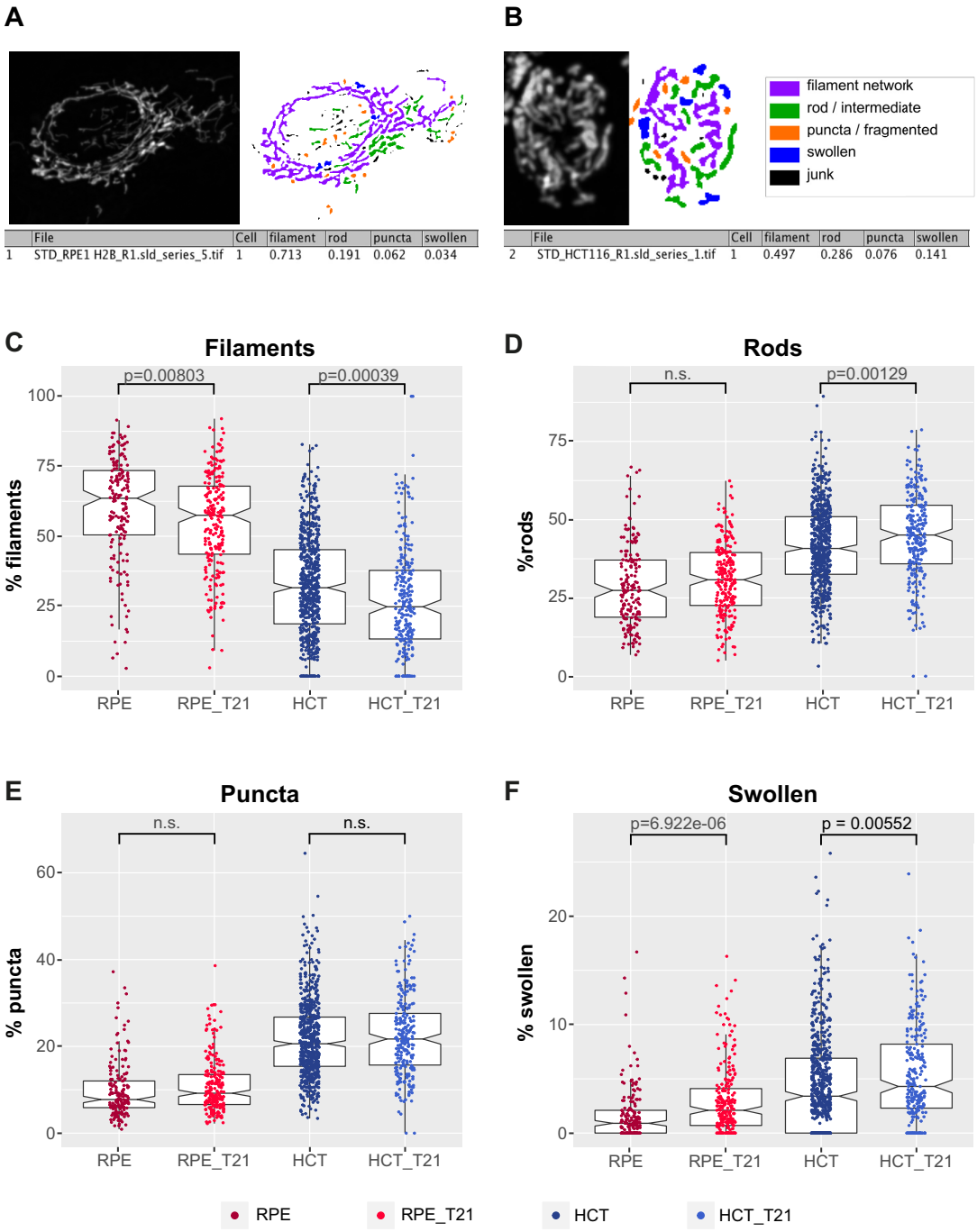


Figure 10. Mitochondrial morphology is slightly changed in trisomy 21 cells. We have stained the mitochondrial network and analyzed the network morphology, measuring the percentage of filamentous networked (filaments), rods, puncta and swollen using the Fiji plugin mitoMorph. (**A**) Sample of mitoMorph analysis of an RPE1 wild-type cell and (**B**) of an HCT116 wild-type cell. Filamentous networks are highlighted in lilac, rods in green, puncta (referring to fragmented mitochondria) are highlighted in orange and swollen ones in blue. The percentage filaments, rods, puncta and swollen mitochondria are automatically scored and reported to the user. (**C**) the percentage of filaments is slightly, but significantly reduced in RPE_T21, as well as HCT_T21 cells compared to wild-type. (**D**) There is no significant change in the percentage of rods in RPE_T21 cells and slightly higher percentage in HCT_T21 cells compared to wild-type. (**E**) The percentage of puncta is unchanged in both T21 cell lines. (**F**) There are significantly more swollen mitochondria in both, RPE_T21, as well as HCT_T21 cells compared to wild-type. Results have been averaged over both clones of each cell line, respectively. Underlying numerical values are provided in Supplementary Table S1, sample images are shown in Supplementary Figure S7.

species due to the loss-of-function of the Tafazzin protein and displays defects in many mitochondrial processes. Chowdhury *et al.* have specifically tested the response to hypoxia of Tafazzin-deficient cells. They have identified reduced ROS levels in Tafazzin-deficient MEFs that lead to a failure to induce the NF-κB transcription factor RelA and finally, the transcription factor Hif1α. Using the more comprehensive and manually curated mito-interactome of mitoXplorer, we could not only confirm the failure of induction of RelA and Hif1α. We furthermore propose that failed induction of the Hippo pathway protein Yap1 which stabilizes Hif1α could contribute to the observed phenotype. Finally, using mitoXplorer we identified MrsB2 as the putative cause of lowered ROS levels in Tafazzin-deficient cells, as this enzyme has ROS-quenching activities, normally protecting cells from oxidative damage.

Cell type-specific de-regulation of mito-genes in trisomy 21

We experimentally verified mitoXplorer predictions by expression profiling of mito-genes in T21 cell lines. Mito-genes were strongly deregulated in both trisomic cell types tested, the non-cancerous retinal pigment epithelial cell line RPE1 and the cancer cell line HCT116. Yet, the changes in expression were quite different in the two cell lines. It is not unexpected that mito-genes are differentially expressed in different cell types, reflecting the divergent cellular energy- and metabolic demands (20). Gene expression is moreover tightly regulated in a cell-type specific manner by regulating transcription, translation and the epigenetic state of the cell. Thus, also divergent and cell-type specific expression changes of mito-genes upon introduction of an extra chromosome is not surprising.

mitoXplorer assisted in revealing divergent de-regulation of mitochondrial transcriptome and proteome in trisomy 21

We found a remarkable difference between transcriptome and proteome levels of mito-genes in RPE_T21 cells. In particular the OXPHOS proteins were strongly down-regulated at protein, but not mRNA level. This can be explained by essential components of the respiratory chain being encoded in the mitochondrial genome and thus requiring a functioning mitochondrial replication system, as well as intact mitochondrial transcription and translation. Thus, there is as strong post-transcriptional regulation of the mitochondrial proteome. In case of the RPE_T21 cell line, the disintegration of the mitoribosome and thus a failure of mitochondrial translation is likely causative for the downregulation of OXPHOS components on protein-level, possibly by proteolysis, as the essential mitochondrial subunits are not produced and thus complexes cannot assemble. This conclusion is further supported by the fact that we could not observe a significant difference in mitochondrial transcript levels, with some mt-mRNAs even being upregulated; thus, mtDNA -maintenance, -replication as well as mito-transcription seem to be unaffected.

MitoXplorer analysis of previously published data of the mito-proteome of fibroblasts isolated from monozygotic twins discordant for T21, as well as 11 unrelated individuals with trisomy 21 (41) confirmed a similar post-transcriptional effect as we found in our T21 model cell lines. Taken together, our data uncovered a significant post-transcriptional regulation of the mitochondrial process OXPHOS in our model system of trisomy 21 that could bring new insight into the mechanisms of mitochondrial defects in trisomy 21 patients.

mitoXplorer helped identify mitochondrial ribosomal protein S21 (MRPS21) as potentially causative for OXPHOS failure

The most notable difference in RPE_T21 cells compared to wild-type is the >10-fold downregulation of mitochondrial ribosomal protein S21 (MRPS21) on transcript level, as well as the downregulation of Mrps21 protein and other proteins of the small and—to a lesser extend—large mitoribosome subunits. Thus, our data suggest that the integrity of the mitoribosome is compromised, leading to its disintegration and subsequently, the downregulation of mitochondrial proteins of the respiratory chain. Mrps21 is a late-assembly component and lies at the outer rim of the body (or bottom) of the small subunit (SSU) of the mitoribosome. Nevertheless, it interacts with a number of other proteins of the SSU and also directly contacts bases of the 12S rRNA (119,120). Thus, its absence could destabilize the SSU of the mitoribosome. The two most down-regulated proteins are Mrps33 and Mrps14, both of which directly interact with each other and several other proteins in the SSU and are localized to the head of the SSU. Furthermore, together with another down-regulated component, Mrps15, they are proteins that are incorporated late in the mitoribosome assembly process (120). This raises the possibility that late-assembly proteins disintegrate more readily from the mitoribosome, leading to their enhanced degradation and thus ribosome malfunction.

Based on promoter analysis using MotifMap (121), potential binding motifs of two transcription factors located on chromosome 21, GABPA and ETS2, can be found in the promoter region of the MRPS21 gene. Gabpα, which is also known as nuclear respiratory factor 2, has already been implicated in mitochondrial biogenesis by regulating Tfb1m expression (122): its depletion in mouse embryonic fibroblasts showed reduced mitochondrial mass, ATP production, oxygen consumption and mito-protein synthesis, but had no effect on mitochondrial morphology, membrane potential or apoptosis. Direct or indirect regulation of mitoribosomal proteins could be another regulatory function of this transcription factor. GABPA is not affected on transcriptome level, but is down-regulated on protein-level in RPE_T21 cells. ETS2 on the other hand has so far not been implicated in mitochondrial biogenesis or functional regulation.

We see consistent downregulation of proteins involved in OXPHOS in other trisomy 21 proteomic datasets and OXPHOS defects have been reported in trisomy 21 before. MRPS21 seems deregulated only in a few T21 individuals. Thus, the causes of OXPHOS deficiencies seem to depend on genomic background or on the cell type studied. Trisomy 21 patients develop different degrees of severity of symptoms and it is likely that the genomic variability of chromosome 21 contributes to the varying phenotypes (123,124). In conclusion, while defects in OXPHOS seem a common phenotype in trisomy 21, their severity as well as the un-

derlying mechanisms might differ depending on the cellular model or the genomic background.

Limitations and future developments of mitoXplorer

The proper assembly and annotation of the mito-interactomes turned out to be a demanding task. While proteomic studies of mitochondria are available for different model species and humans, the data cannot be taken without manual intervention due to significant numbers of false-positives and false-negatives. A manual curation of the data is therefore mandatory. Though we carefully curated the mito-interactomes used in mitoXplorer, we do not claim that they are complete or free of false-positives.

We decided to assign one mito-gene with one mito-process only, as it was the most straightforward solution to implement. Furthermore, assigning multiple processes to one gene would artificially increase the mitochondrial interactome, making the analysis also more difficult. Nevertheless, the current annotation might not represent all experimentally validated biological functions of a mito-gene. In future releases, we will therefore consider allowing two or more processes for one mito-gene at least in a limited number of cases, whenever there is strong experimental evidence that a protein or protein complex is contributing to two or more mito-processes. We hope that the scientific community working on mitochondria will help us further clean, complete and correctly annotate our mito-interactomes by using the FEEDBACK page.

The current version of mitoXplorer does not provide mito-process enrichment analysis. Our reasoning behind this decision was to allow users to mine their data in an unbiased and detailed way, considering all mito-processes rather than only focusing on enriched ones. In future releases, we will consider adding information on mito-process enrichment of a dataset to guide users in their analysis, for example by visually highlighting enriched processes in the Interactome View.

MitoXplorer has been optimized for mining expression data and currently it is not meaningful to analyze mito-gene mutations alone. We realize that this is a limitation when considering large, population-wide studies. We will therefore consider implementing a visual data mining interface that is specifically tailored for analyzing mito-gene mutations alone in future versions of the software.

MitoXplorer expects users to provide data for which differential expression analysis or mutation calling has already been performed, as mitoXplorer was conceived as a visual data mining platform. There are already many tools and pipelines available to perform differential expression analysis and mutation calling, yet too few tools that allow in-depth data mining, such as mitoXplorer. To make prior data analysis as easy as possible for users, we provide a pipeline for differential expression analysis and mutant calling, which is available in our git-repository (https://gitlab.com/habermannlab/mitox_rnaseq_pipeline/). Nevertheless, we recognize that the potential heterogeneity of analyzed data from different studies has limitations, especially concerning comparative analysis between different projects.

MitoXplorer is integrating, clustering and visualizing numerical data resulting from expression studies (transcriptome, proteome), as well as mutation data. Thus, it is currently limited to analyzing mito-genes without offering the ability to explore their embedding in a broader, cellular context and thus to learn about potential regulatory mechanisms of observed expression changes of mito-genes. Therefore, in the next release of mitoXplorer, we plan to fully embed mito-genes within the cellular gene regulatory, as well as signaling network by adding information from epigenetic studies (ChIP-seq, methylation data), as well as from the cellular interactome. We will provide the tools to perform enrichment analysis of observed transcription factors binding in the promoter regions of co-regulated mito-genes. We will embed a method to analyze promoter regions of mito-genes, as we have shown here for MRPS21, or to identify targets of transcription factors as here demonstrated for Hif1α. Furthermore, we will make available network analysis methods such as viPEr (125) to explore the cellular network regulating mito-genes. Other analysis methods we plan to provide include correlation analysis, as well as cross-species data mining. Depending on user requests, we could also add the mitochondrial interactomes of other species. As mitoXplorer stores the mitochondrial interactomes and the associated -omics data in a MySQL database, all technical requirements for extending the functionalities of mitoXplorer are already implemented.

CONCLUSIONS

mitoXplorer is a powerful, web-based visual data mining platform that allows users to in-depth analyze and visualize mutations and expression dynamics of mito-genes and mito-processes by integrating a manually curated mitochondrial interactome with -omics data in various tissues and conditions of four model species, including human. We used transcriptome and proteome data from cell lines with trisomy 21 to demonstrate the value of mitoXplorer in analyzing in detail the expression dynamics of mito-genes and -processes. We have used mitoXplorer to integrate these data with publicly available datasets of patients with trisomy 21. Using mitoXplorer for data mining, we predicted failure of mitochondrial respiration in one of the trisomy 21 cell lines, which we verified experimentally. Our results demonstrate the power of a visual data mining platform such as mitoXplorer to explore expression dynamics of a specified mito-gene set in a detailed and focused manner, leading to discovery of underlying molecular mechanisms and providing testable hypotheses for further experimental studies.

DATA AVAILABILITY

The mitoXplorer web-server is freely available at http://mitoxplorer.ibdm.univ-mrs.fr/. The source code of mitoXplorer is available at https://gitlab.com/habermannlab/mitox. The pipeline for differential expression analysis and mutation calling of RNA-seq data is available at https://gitlab.com/habermannlab/mitox_rnaseq_pipeline. MitoMorph is freely available at https://github.com/giocard/mitoMorph. RNA-seq data published with this study are available via the Gene Expression Omnibus (GEO) database (accession number: GSE131249).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We want to thank Michael Volkmer and Edlira Nano for help and advice in web-server management and development. We want to thank Stephen Taylor (University of Manchester, UK) for providing cell lines. We thank Alice Carrier, Friedhelm Pfeiffer and Frank Schnorrer for critical reading of the manuscript. We thank the Max Planck Society, the Max-Planck Institute for Biochemistry, the Aix-Marseille University, the CNRS and the Institute of Developmental Biology Marseille (IBDM) for their generous support.

Author's contributions: A.Y. and P.K. were the main developers of the mitoXplorer web-server with the help of J.V., S.G., A.B. and F.M. J.V. developed the Interactome View of the web-server. Data analysis was done by A.Y., P.K., M.D. and B.H.H. MitoTracker staining and imaging was carried out by M.D. C.M. and F.P. carried out metabolic measurements. Handling of cells and cell culture was done by M.D. and Z.S. G.C. conceived and developed the mitoMorph Fiji plugin, image analysis with mitoMorph was done by B.H.H. The project was conceived by B.H.H., the manuscript was written by A.Y. and B.H.H. with contributions from S.Z., C.G., M.D. and C.M. All authors read and approved the final version of the manuscript.

FUNDING

DFG grant [HA 6905/2-1] from the German research foundation; A*MIDEX [2HABERRE/RHRE/ID17HRU288] from Aix-Marseille University; ANR grant [ANR-18-CE45-0016-01 to B.H.H.]; ERASMUS+ Traineeship program (to A.Y.); Munich Center for Systems Neurology [SyNergy EXC 1010]; Bert L & N Kuggie Vallee Foundation (to F.P.); Bavarian Molecular Biosystems Research Network [D2-F5121.2-10c/4822 to C.M.]. Funding for open access charge: Agence Nationale de la Recherche [ANR-18-CE45-0016-01].

Conflict of interest statement. None declared.

REFERENCES

1. Cancer Genome Atlas Research Network, Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R.M., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
2. Zhang,J., Baran,J., Cros,A., Guberman,J.M., Haider,S., Hsu,J., Liang,Y., Rivkin,E., Wang,J., Whitty,B. *et al.* (2011) International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data. *Database*, **2011**, bar026.
3. Krempel,R., Kulkarni,P., Yim,A., Lang,U., Habermann,B. and Frommolt,P. (2018) Integrative analysis and machine learning on cancer genomics data using the Cancer Systems Biology Database (CancerSysDB). *BMC Bioinformatics*, **19**, 156.
4. Klonowska,K., Czubak,K., Wojciechowska,M., Handschuh,L., Zmienko,A., Figlerowicz,M., Dams-Kozlowska,H. and Kozlowski,P. (2016) Oncogenomic portals for the visualization and analysis of genome-wide cancer data. *Oncotarget*, **7**, 176–192.
5. Papatheodorou,I., Fonseca,N.A., Keays,M., Tang,Y.A., Barrera,E., Bazant,W., Burke,M., Füllgrabe,A., Fuentes,A.M.-P., George,N. *et al.* (2018) Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.*, **46**, D246–D251.
6. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
7. Kodama,Y., Mashima,J., Kaminuma,E., Gojobori,T., Ogasawara,O., Takagi,T., Okubo,K. and Nakamura,Y. (2012) The DNA Data Bank of Japan launches a new resource, the DDBJ Omics Archive of functional genomics experiments. *Nucleic Acids Res.*, **40**, D38–D42.
8. Parkinson,H., Sarkans,U., Kolesnikov,N., Abeygunawardena,N., Burdett,T., Dylag,M., Emam,I., Farne,A., Hastings,E., Holloway,E. *et al.* (2011) ArrayExpress update–an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **39**, D1002–D1004.
9. Simoff,S.J., Böhlen,M.H. and Mazeika,A. (2008) Visual Data Mining: An Introduction and Overview. In *Visual Data Mining*, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, Vol. **4404**, pp. 1–12.
10. Scheffler,I.E. (2007) *Mitochondria*. John Wiley & Sons, Inc., Hoboken, NJ.
11. Nunnari,J. and Suomalainen,A. (2012) Mitochondria: in sickness and in health. *Cell*, **148**, 1145–1159.
12. Suomalainen,A. and Battersby,B.J. (2018) Mitochondrial diseases: the contribution of organelle stress responses to pathology. *Nat. Rev. Mol. Cell Biol.*, **19**, 77–92.
13. Zong,W.-X., Rabinowitz,J.D. and White,E. (2016) Mitochondria and Cancer. *Mol. Cell*, **61**, 667–676.
14. Wallace,D.C. (2012) Mitochondria and cancer. *Nat. Rev. Cancer*, **12**, 685–698.
15. Schapira,A.H.V. (2012) Mitochondrial diseases. *Lancet.*, **379**, 1825–1834.
16. Mannella,C.A. (2008) Structural diversity of mitochondria: functional implications. *Ann. N. Y. Acad. Sci.*, **1147**, 171–179.
17. Vafai,S.B. and Mootha,V.K. (2012) Mitochondrial disorders as windows into an ancient organelle. *Nature*, **491**, 374–383.
18. Wai,T. and Langer,T. (2016) Mitochondrial dynamics and metabolic regulation. *Trends Endocrinol. Metab.*, **27**, 105–117.
19. Benard,G., Bellance,N., James,D., Parrone,P., Fernandez,H., Letellier,T. and Rossignol,R. (2007) Mitochondrial bioenergetics and structural network organization. *J. Cell Sci.*, **120**, 838–848.
20. Woods,D.C. (2017) Mitochondrial heterogeneity: evaluating mitochondrial subpopulation dynamics in stem cells. *Stem Cells Int.*, **2017**, 7068567–7068567.
21. Mootha,V.K., Bunkenborg,J., Olsen,J.V., Hjerrild,M., Wisniewski,J.R., Stahl,E., Bolouri,M.S., Ray,H.N., Sihag,S., Kamal,M. *et al.* (2003) Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell*, **115**, 629–640.
22. Jensen,R.E., Dunn,C.D., Youngman,M.J. and Sesaki,H. (2004) Mitochondrial building blocks. *Trends Cell Biol.*, **14**, 215–218.
23. Pagliarini,D.J., Calvo,S.E., Chang,B., Sheth,S.A., Vafai,S.B., Ong,S.-E., Walford,G.A., Sugiana,C., Boneh,A., Chen,W.K. *et al.* (2008) A mitochondrial protein compendium elucidates complex I disease biology. *Cell*, **134**, 112–123.
24. Calvo,S.E., Clauser,K.R. and Mootha,V.K. (2016) MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Res.*, **44**, D1251–D1257.
25. Gray,M.W. (2015) Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 10133–10138.
26. Meisinger,C., Sickmann,A. and Pfanner,N. (2008) The mitochondrial proteome: from inventory to function. *Cell*, **134**, 22–24.
27. Lotz,C., Lin,A.J., Black,C.M., Zhang,J., Lau,E., Deng,N., Wang,Y., Zong,N.C., Choi,J.H., Xu,T. *et al.* (2014) Characterization, design, and function of the mitochondrial proteome: from organs to organisms. *J. Proteome Res.*, **13**, 433–446.
28. Gaucher,S.P., Taylor,S.W., Fahy,E., Zhang,B., Warnock,D.E., Ghosh,S.S. and Gibson,B.W. (2004) Expanded coverage of the human heart mitochondrial proteome using multidimensional liquid chromatography coupled with tandem mass spectrometry. *J. Proteome Res.*, **3**, 495–505.

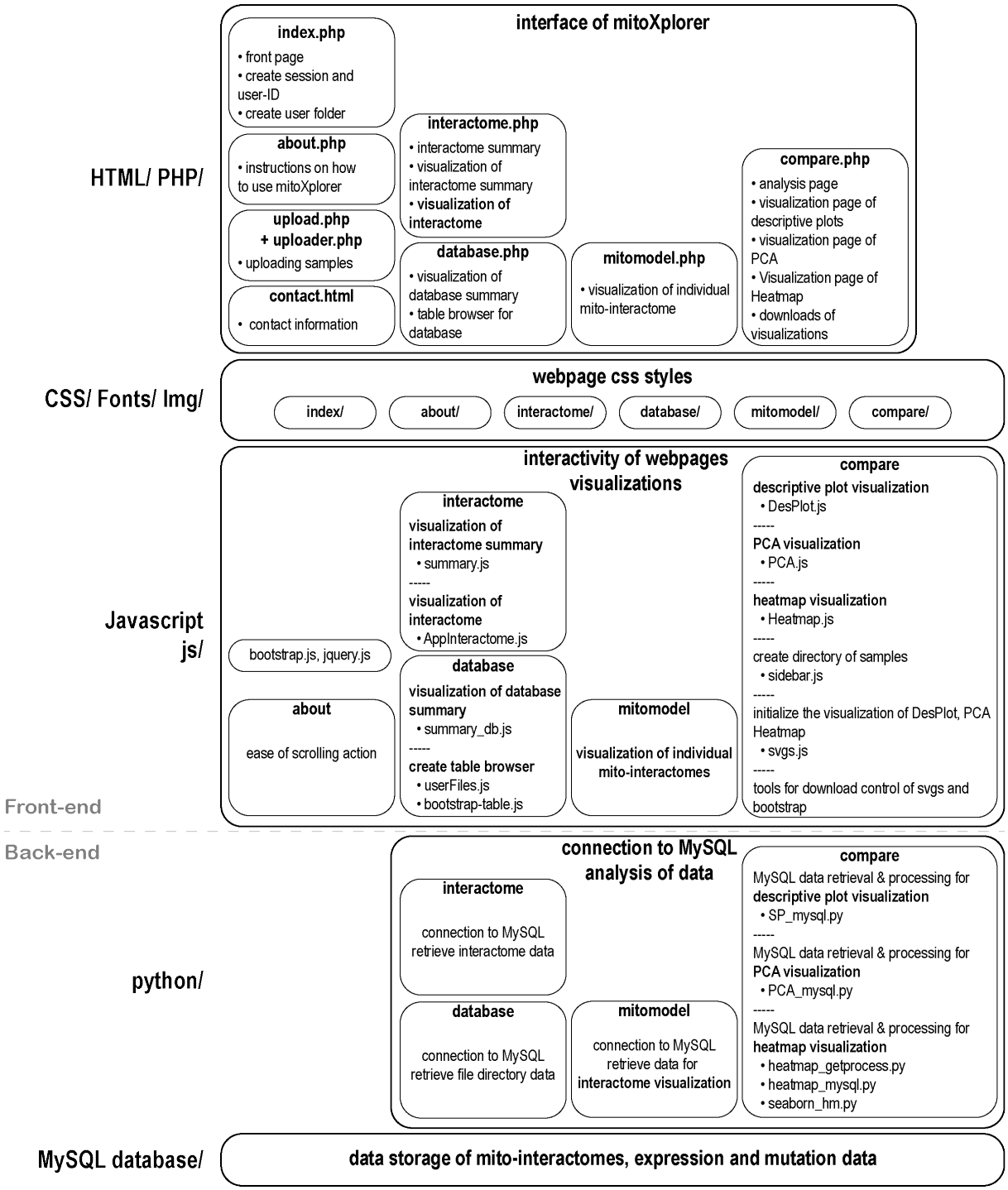
29. Taylor,S.W., Fahy,E., Zhang,B., Glenn,G.M., Warnock,D.E., Wiley,S., Murphy,A.N., Gaucher,S.P., Capaldi,R.A., Gibson,B.W. *et al.* (2003) Characterization of the human heart mitochondrial proteome. *Nat. Biotechnol.*, **21**, 281–286.
30. Gonczarowska-Jorge,H., Zahedi,R.P. and Sickmann,A. (2017) The proteome of baker's yeast mitochondria. *Mitochondrion*, **33**, 15–21.
31. Kolesnikov,A.A. and Gerasimov,E.S. (2012) Diversity of mitochondrial genome organization. *Biochemistry Mosc.*, **77**, 1424–1435.
32. Hällberg,B.M. and Larsson,N.-G. (2014) Making proteins in the powerhouse. *Cell Metab.*, **20**, 226–240.
33. Catalano,D., Licciulli,F., Turi,A., Grillo,G., Saccone,C. and D'Elia,D. (2006) MitoRes: a resource of nuclear-encoded mitochondrial genes and their products in Metazoa. *BMC Bioinformatics*, **7**, 36.
34. Smith,A.C. and Robinson,A.J. (2016) MitoMiner v3.1, an update on the mitochondrial proteomics database. *Nucleic Acids Res.*, **44**, D1258–D1261.
35. Godin,N. and Eichler,J. (2017) The Mitochondrial Protein Atlas: a database of experimentally verified information on the human mitochondrial proteome. *J. Comput. Biol.*, **24**, 906–916.
36. Cotter,D., Guda,P., Fahy,E. and Subramaniam,S. (2004) MitoProteome: mitochondrial protein sequence database and annotation system. *Nucleic Acids Res.*, **32**, D463–D467.
37. Guda,C., Fahy,E. and Subramaniam,S. (2004) MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics*, **20**, 1785–1794.
38. Izzo,A., Mollo,N., Nitti,M., Paladino,S., Calì,G., Genesio,R., Bonfiglio,F., Cicatiello,R., Barbato,M., Sarnataro,V. *et al.* (2018) Mitochondrial dysfunction in down syndrome: molecular mechanisms and therapeutic targets. *Mol. Med.*, **24**, 2–8.
39. Bostock,M., Ogievetsky,V. and Heer,J. D^3 Data-Driven Documents (2011) *IEEE Trans. Visual. Comput. Graph.*, **17**, 2301–2309.
40. Kühl,I., Miranda,M., Atanassov,I., Kuznetsova,I., Hinze,Y., Mourier,A., Filipovska,A. and Larsson,N.-G. (2017) Transcriptomic and proteomic landscape of mitochondrial dysfunction reveals secondary coenzyme Q deficiency in mammals. *Elife*, **6**, 1494.
41. Liu,Y., Borel,C., Li,L., Müller,T., Williams,E.G., Germain,P.-L., Buljan,M., Sajic,T., Boersema,P.J., Shao,W. *et al.* (2017) Systematic proteome and proteostasis profiling in human Trisomy 21 fibroblast cells. *Nat. Commun.*, **8**, 1212.
42. Letourneau,A., Santoni,F.A., Bonilla,X., Sailani,M.R., Gonzalez,D., Kind,J., Chevalier,C., Thurman,R., Sandstrom,R.S., Hibaoui,Y. *et al.* (2014) Domains of genome-wide gene expression dysregulation in Down's syndrome. *Nature*, **508**, 345–350.
43. Sullivan,K.D., Lewis,H.C., Hill,A.A., Pandey,A., Jackson,L.P., Cabral,J.M., Smith,K.P., Liggett,L.A., Gomez,E.B., Galbraith,M.D. *et al.* (2016) Trisomy 21 consistently activates the interferon response. *Elife*,**5**, 1709.
44. Spletter,M.L., Barz,C., Yeroslaviz,A., Zhang,X., Lemke,S.B., Bonnard,A., Brunner,E., Cardone,G., Basler,K., Habermann,B.H. *et al.* (2018) A transcriptomics resource reveals a transcriptional transition during ordered sarcomere morphogenesis in flight muscle. *Elife*, **7**, 1361.
45. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550–521.
46. Chowdhury,A., Aich,A., Jain,G., Wozny,K., Lüchtenborg,C., Hartmann,M., Bernhard,O., Balleiniger,M., Alfar,E.A., Zieseniss,A. *et al.* (2018) Defective Mitochondrial Cardiolipin Remodeling Dampens HIF-1α Expression in Hypoxia. *Cell Rep.*, **25**, 561–570.
47. Garipler,G., Mutlu,N., Lack,N.A. and Dunn,C.D. (2014) Deletion of conserved protein phosphatases reverses defects associated with mitochondrial DNA damage in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 1473–1478.
48. Fleischer,J.G., Schulte,R., Tsai,H.H., Tyagi,S., Ibarra,A., Shokhirev,M.N., Huang,L., Hetzer,M.W. and Navlakha,S. (2018) Predicting age from the transcriptome of human dermal fibroblasts. *Genome Biol.*, **19**, 221–228.
49. Huang,W., Carbone,M.A., Magwire,M.M., Peiffer,J.A., Lyman,R.F., Stone,E.A., Anholt,R.R.H. and Mackay,T.F.C. (2015) Genetic basis of transcriptome diversity in Drosophila melanogaster. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E6010–E6019.
50. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
51. Stingele,S., Stoehr,G., Peplowska,K., Cox,J., Mann,M. and Storchova,Z. (2012) Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Mol. Syst. Biol.*, **8**, 608.
52. Dürrbaum,M., Kuznetsova,A.Y., Passerini,V., Stingele,S., Stoehr,G. and Storchova,Z. (2014) Unique features of the transcriptional response to model aneuploidy in human cells. *BMC Genomics*, **15**, 139.
53. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, **17**, 10–12.
54. Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.
55. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.
56. Schindelin,J., Arganda-Carreras,I., Frise,E., Kaynig,V., Longair,M., Pietzsch,T., Preibisch,S., Rueden,C., Saalfeld,S., Schmid,B. *et al.* (2012) Fiji: an open-source platform for biological-image analysis. *Nat. Methods*, **9**, 676–682.
57. Divakaruni,A.S., Paradyse,A., Ferrick,D.A., Murphy,A.N. and Jastroch,M. (2014) Analysis and interpretation of microplate-based oxygen consumption and pH data. *Methods Enzymol.*, **547**, 309–354.
58. Schneider,C.A., Rasband,W.S. and Eliceiri,K.W. (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods*, **9**, 671–675.
59. Yen,J.C., Chang,F.J. and Chang,S. (1995) A new criterion for automatic multilevel thresholding. *IEEE Trans. Image Process*, **4**, 370–378.
60. Leonard,A.P., Cameron,R.B., Speiser,J.L., Wolf,B.J., Peterson,Y.K., Schnellmann,R.G., Beeson,C.C. and Rohrer,B. (2015) Quantitative analysis of mitochondrial morphology and membrane potential in living cells using high-content imaging, machine learning, and morphological binning. *Biochim. Biophys. Acta*, **1853**, 348–360.
61. Morgenstern,M., Stiller,S.B., Lübbert,P., Peikert,C.D., Dannenmaier,S., Drepper,F., Weill,U., Höß,P., Feuerstein,R., Gebert,M. *et al.* (2017) Definition of a high-confidence mitochondrial proteome at quantitative scale. *Cell Rep.*, **19**, 2836–2852.
62. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
63. NCBI Resource Coordinators (2018) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **46**, D8–D13.
64. Thurmond,J., Goodman,J.L., Strelets,V.B., Attrill,H., Gramates,L.S., Marygold,S.J., Matthews,B.B., Millburn,G., Antonazzo,G., Trovisco,V. *et al.* (2019) FlyBase 2.0: the next generation. *Nucleic Acids Res.*, **47**, D759–D765.
65. Cherry,J.M., Hong,E.L., Amundsen,C., Balakrishnan,R., Binkley,G., Chan,E.T., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R. *et al.* (2012) Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, D700–D705.
66. Stelzer,G., Rosen,N., Plaschkes,I., Zimmerman,S., Twik,M., Fishilevich,S., Stein,T.I., Nudel,R., Lieder,I., Mazor,Y. *et al.* (2016) The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinformatics*, **54**, 1.30.1–1.30.33.
67. Consortium,The UniProt (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
68. Kuan,J. and Saier,M.H. (1993) The mitochondrial carrier family of transport proteins: structural, functional, and evolutionary relationships. *Crit. Rev. Biochem. Mol. Biol.*, **28**, 209–233.
69. Szklarczyk,D., Morris,J.H., Cook,H., Kuhn,M., Wyder,S., Simonovic,M., Santos,A., Doncheva,N.T., Roth,A., Bork,P. *et al.* (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.*, **45**, D362–D368.

70. Quackenbush,J. (2002) Microarray data normalization and transformation. *Nat. Genet.*, **32**, 496–501.
71. DeBerardinis,R.J. and Chandel,N.S. (2016) Fundamentals of cancer metabolism. *Sci. Adv.*, **2**, e1600200.
72. Bratic,I. and Trifunovic,A. (2010) Mitochondrial energy metabolism and ageing. *Biochim. Biophys. Acta*, **1797**, 961–967.
73. Bione,S., D'Adamo,P., Maestrini,E., Gedeon,A.K., Bolhuis,P.A. and Toniolo,D. (1996) A novel X-linked gene, G4.5. is responsible for Barth syndrome. *Nat. Genet.*, **12**, 385–389.
74. Ikon,N. and Ryan,R.O. (2017) Barth syndrome: connecting cardiolipin to cardiomyopathy. *Lipids*, **52**, 99–108.
75. Ikon,N. and Ryan,R.O. (2017) Cardiolipin and mitochondrial cristae organization. *Biochim. Biophys. Acta Biomembr.*, **1859**, 1156–1163.
76. Conesa,A., Madrigal,P., Tarazona,S., Gomez-Cabrero,D., Cervera,A., McPherson,A., Szcześniak,M.W., Gaffney,D.J., Elo,L.L., Zhang,X. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 13–19.
77. Conesa,A., Madrigal,P., Tarazona,S., Gomez-Cabrero,D., Cervera,A., McPherson,A., Szcześniak,M.W., Gaffney,D.J., Elo,L.L., Zhang,X. *et al.* (2016) Erratum to: A survey of best practices for RNA-seq data analysis. *Genome Biol.*, **17**, 181–182.
78. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
79. Vreken,P., Valianpour,F., Nijtmans,L.G., Grivell,L.A., Plecko,B., Wanders,R.J. and Barth,P.G. (2000) Defective remodeling of cardiolipin and phosphatidylglycerol in Barth syndrome. *Biochem. Biophys. Res. Commun.*, **279**, 378–382.
80. Schlame,M. (2013) Cardiolipin remodeling and the function of tafazzin. *Biochim. Biophys. Acta*, **1831**, 582–588.
81. Jefferies,J.L. (2013) Barth syndrome. *Am. J. Med. Genet C Semin. Med. Genet.*, **163C**, 198–205.
82. Dudek,J. (2017) Role of Cardiolipin in Mitochondrial Signaling Pathways. *Front. Cell Dev. Biol.*, **5**, 90.
83. Rouillard,A.D., Gundersen,G.W., Fernandez,N.F., Wang,Z., Monteiro,C.D., McDermott,M.G. and Ma'ayan,A. (2016) The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, **2016**, baw100.
84. Ma,B., Chen,Y., Chen,L., Cheng,H., Mu,C., Li,J., Gao,R., Zhou,C., Cao,L., Liu,J. *et al.* (2015) Hypoxia regulates Hippo signalling through the SIAH2 ubiquitin E3 ligase. *Nat. Cell Biol.*, **17**, 95–103.
85. Cabreiro,F., Picot,C.R., Perichon,M., Castel,J., Friguet,B. and Petropoulos,I. (2008) Overexpression of mitochondrial methionine sulfoxide reductase B2 protects leukemia cells from oxidative stress-induced cell death and protein damage. *J. Biol. Chem.*, **283**, 16673–16681.
86. Valenti,D., de Bari,L., De Filippis,B., Henrion-Caude,A. and Vacca,R.A. (2014) Mitochondrial dysfunction as a central actor in intellectual disability-related diseases: an overview of Down syndrome, autism, Fragile X and Rett syndrome. *Neurosci. Biobehav. Rev.*, **46**, 202–217.
87. Tiano,L. and Busciglio,J. (2011) Mitochondrial dysfunction and Down's syndrome: is there a role for coenzyme Q(10)? *Biofactors*, **37**, 386–392.
88. Pagano,G. and Castello,G. (2012) Oxidative stress and mitochondrial dysfunction in Down syndrome. *Adv. Exp. Med. Biol.*, **724**, 291–299.
89. Ogawa,O., Perry,G. and Smith,M.A. (2002) The 'Down's' side of mitochondria. *Dev. Cell*, **2**, 255–256.
90. Prince,J., Jia,S., Båve,U., Annerén,G. and Oreland,L. (1994) Mitochondrial enzyme deficiencies in Down's syndrome. *J. Neural Transm. Park Dis. Dement. Sect.*, **8**, 171–181.
91. Roat,E., Prada,N., Ferraresi,R., Giovenzana,C., Nasi,M., Troiano,L., Pinti,M., Nemes,E., Lugli,E., Biagioni,O. *et al.* (2007) Mitochondrial alterations and tendency to apoptosis in peripheral blood cells from children with Down syndrome. *FEBS Lett.*, **581**, 521–525.
92. Piccoli,C., Izzo,A., Scrima,R., Bonfiglio,F., Manco,R., Negri,R., Quarato,G., Cela,O., Ripoli,M., Prisco,M. *et al.* (2013) Chronic pro-oxidative state and mitochondrial dysfunctions are more pronounced in fibroblasts from Down syndrome foeti with congenital heart defects. *Hum. Mol. Genet.*, **22**, 1218–1232.
93. Phillips,A.C., Sleigh,A., McAllister,C.J., Brage,S., Carpenter,T.A., Kemp,G.J. and Holland,A.J. (2013) Defective mitochondrial function in vivo in skeletal muscle in adults with Down's syndrome: a 31P-MRS study. *PLoS ONE*, **8**, e84031.
94. Aburawi,E.H. and Souid,A.-K. (2012) Lymphocyte respiration in children with Trisomy 21. *BMC Pediatr*, **12**, 193.
95. Valenti,D., Manente,G.A., Moro,L., Marra,E. and Vacca,R.A. (2011) Deficit of complex I activity in human skin fibroblasts with chromosome 21 trisomy and overproduction of reactive oxygen species by mitochondria: involvement of the cAMP/PKA signalling pathway. *Biochem. J.*, **435**, 679–688.
96. Valenti,D., Tullo,A., Caratozzolo,M.F., Merafina,R.S., Scartezzini,P., Marra,E. and Vacca,R.A. (2010) Impairment of F1F0-ATPase, adenine nucleotide translocator and adenylate kinase causes mitochondrial energy deficit in human skin fibroblasts with chromosome 21 trisomy. *Biochem. J.*, **431**, 299–310.
97. Abu Faddan,N., Sayed,D. and Ghaleb,F. (2011) T lymphocytes apoptosis and mitochondrial membrane potential in Down's syndrome. *Fetal Pediatr Pathol*, **30**, 45–52.
98. Izzo,A., Nitti,M., Mollo,N., Paladino,S., Procaccini,C., Faicchia,D., Calì,G., Genesio,R., Bonfiglio,F., Cicatiello,R. *et al.* (2017) Metformin restores the mitochondrial network and reverses mitochondrial dysfunction in Down syndrome cells. *Hum. Mol. Genet.*, **26**, 1056–1069.
99. Busciglio,J., Pelsman,A., Wong,C., Pigino,G., Yuan,M., Mori,H. and Yankner,B.A. (2002) Altered metabolism of the amyloid beta precursor protein is associated with mitochondrial dysfunction in Down's syndrome. *Neuron*, **33**, 677–688.
100. Lockstone,H.E., Harris,L.W., Swatton,J.E., Wayland,M.T., Holland,A.J. and Bahn,S. (2007) Gene expression profiling in the adult Down syndrome brain. *Genomics*, **90**, 647–660.
101. Halevy,T., Biancotti,J.-C., Yanuka,O., Golan-Lev,T. and Benvenisty,N. (2016) Molecular characterization of down syndrome embryonic stem cells reveals a role for RUNX1 in neural differentiation. *Stem Cell Rep.*, **7**, 777–786.
102. Olmos-Serrano,J.L., Kang,H.J., Tyler,W.A., Silbereis,J.C., Cheng,F., Zhu,Y., Pletikos,M., Jankovic-Rapan,L., Cramer,N.P., Galdzicki,Z. *et al.* (2016) Down syndrome developmental brain transcriptome reveals defective oligodendrocyte differentiation and myelination. *Neuron*, **89**, 1208–1222.
103. Jiang,J., Jing,Y., Cost,G.J., Chiang,J.-C., Kolpa,H.J., Cotton,A.M., Carone,D.M., Carone,B.R., Shivak,D.A., Guschin,D.Y. *et al.* (2013) Translating dosage compensation to trisomy 21. *Nature*, **500**, 296–300.
104. Helguera,P., Seiglie,J., Rodriguez,J., Hanna,M., Helguera,G. and Busciglio,J. (2013) Adaptive downregulation of mitochondrial function in down syndrome. *Cell Metab.*, **17**, 132–140.
105. Ripoll,C., Rivals,I., Ait Yahya-Graison,E., Dauphinot,L., Paly,E., Mircher,C., Ravel,A., Grattau,Y., Bléhaut,H., Mégarbane,A. *et al.* (2012) Molecular signatures of cardiac defects in Down syndrome lymphoblastoid cell lines suggest altered ciliome and Hedgehog pathways. *PLoS ONE*, **7**, e41616.
106. Li,C., Jin,L., Bai,Y., Chen,Q., Fu,L., Yang,M., Xiao,H., Zhao,G. and Wang,S. (2012) Genome-wide expression analysis in Down syndrome: insight into immunodeficiency. *PLoS ONE*, **7**, e49130.
107. Chou,C.Y., Liu,L.Y., Chen,C.Y., Tsai,C.H., Hwa,H.L., Chang,L.Y., Lin,Y.S. and Hsieh,F.J. (2008) Gene expression variation increase in trisomy 21 tissues. *Mamm. Genome*, **19**, 398–405.
108. Altug-Teber,O., Bonin,M., Walter,M., Mau-Holzmann,U.A., Dufke,A., Stappert,H., Tekesin,I., Heilbronner,H., Nieselt,K. and Riess,O. (2007) Specific transcriptional changes in human fetuses with autosomal trisomies. *Cytogenet. Genome Res.*, **119**, 171–184.
109. Conti,A., Fabbrini,F., D'Agostino,P., Negri,R., Greco,D., Genesio,R., D'Armiento,M., Olla,C., Paladini,D., Zannini,M. *et al.* (2007) Altered expression of mitochondrial and extracellular matrix genes in the heart of human fetuses with chromosome 21 trisomy. *BMC Genomics*, **8**, 268.
110. Mao,R., Wang,X., Spitznagel,E.L., Frelin,L.P., Ting,J.C., Ding,H., Kim,J.-W., Ruczinski,I., Downey,T.J. and Pevsner,J. (2005) Primary and secondary transcriptional effects in the developing human Down syndrome brain and heart. *Genome Biol.*, **6**, R107.

111. Hibaoui,Y., Grad,I., Letourneau,A., Sailani,M.R., Dahoun,S., Santoni,F.A., Gimelli,S., Guipponi,M., Pelte,M.F., Béna,F. *et al.* (2014) Modelling and rescuing neurodevelopmental defect of Down syndrome using induced pluripotent stem cells from monozygotic twins discordant for trisomy 21. *EMBO Mol. Med.*, **6**, 259–277.
112. Engidawork,E., Gulesserian,T., Fountoulakis,M. and Lubec,G. (2003) Aberrant protein expression in cerebral cortex of fetus with Down syndrome. *Neuroscience*, **122**, 145–154.
113. Cheon,M.S., Fountoulakis,M., Dierssen,M., Ferreres,J.C. and Lubec,G. (2001) Expression profiles of proteins in fetal brain with Down syndrome. *J. Neural Transm. Suppl.*, **61**, 311–319.
114. Cabras,T., Pisano,E., Montaldo,C., Giuca,M.R., Iavarone,F., Zampino,G., Castagnola,M. and Messana,I. (2013) Significant modifications of the salivary proteome potentially associated with complications of Down syndrome revealed by top-down proteomics. *Mol. Cell Proteomics*, **12**, 1844–1852.
115. Sullivan,K.D., Evans,D., Pandey,A., Hraha,T.H., Smith,K.P., Markham,N., Rachubinski,A.L., Wolter-Warmerdam,K., Hickey,F., Espinosa,J.M. *et al.* (2017) Trisomy 21 causes changes in the circulating proteome indicative of chronic autoinflammation. *Sci. Rep.*, **7**, 14818.
116. Chacinska,A., Koehler,C.M., Milenkovic,D., Lithgow,T. and Pfanner,N. (2009) Importing mitochondrial proteins: machineries and mechanisms. *Cell*, **138**, 628–644.
117. Sylvester,J.E., Fischel-Ghodsian,N., Mougey,E.B. and O'Brien,T.W. (2004) Mitochondrial ribosomal proteins: candidate genes for mitochondrial disease. *Genet. Med.*, **6**, 73–80.
118. Niemann,A., Ruegg,M., La Padula,V., Schenone,A. and Suter,U. (2005) Ganglioside-induced differentiation associated protein 1 is a regulator of the mitochondrial network: new implications for Charcot-Marie-Tooth disease. *J. Cell Biol.*, **170**, 1067–1078.
119. Amunts,A., Brown,A., Toots,J., Scheres,S.H.W. and Ramakrishnan,V. (2015) Ribosome. The structure of the human mitochondrial ribosome. *Science*, **348**, 95–98.
120. Bogenhagen,D.F., Ostermeyer-Fay,A.G., Haley,J.D. and Garcia-Diaz,M. (2018) Kinetics and mechanism of mammalian mitochondrial ribosome assembly. *Cell Rep.*, **22**, 1935–1944.
121. Daily,K., Patel,V.R., Rigor,P., Xie,X. and Baldi,P. (2011) MotifMap: integrative genome-wide maps of regulatory motif sites for model species. *BMC Bioinformatics*, **12**, 495.
122. Yang,Z.-F., Drumea,K., Mott,S., Wang,J. and Rosmarin,A.G. (2014) GABP transcription factor (nuclear respiratory factor 2) is required for mitochondrial biogenesis. *Mol. Cell Biol.*, **34**, 3194–3201.
123. Lana-Elola,E., Watson-Scales,S.D., Fisher,E.M.C. and Tybulewicz,V.L.J. (2011) Down syndrome: searching for the genetic culprits. *Dis. Model. Mech.*, **4**, 586–595.
124. Antonarakis,S.E., Lyle,R., Dermitzakis,E.T., Reymond,A. and Deutsch,S. (2004) Chromosome 21 and down syndrome: from genomics to pathophysiology. *Nat. Rev. Genet.*, **5**, 725–738.
125. Garmhausen,M., Hofmann,F., Senderov,V., Thomas,M., Kandel,B.A. and Habermann,B.H. (2015) Virtual pathway explorer (viPEr) and pathway enrichment analysis tool (PEANuT): creating and analyzing focus networks to identify cross-talk between molecules and pathways. *BMC Genomics*, **16**, 790.
126. Li,Y., Lim,S., Hoffman,D., Aspenstrom,P., Federoff,H.J. and Rempe,D.A. (2009) HUMMR, a hypoxia- and HIF-1alpha-inducible protein, alters mitochondrial distribution and transport. *J. Cell Biol.*, **185**, 1065–1081.

Supplementary Figures 1 – 7, Supplementary Figures Legends

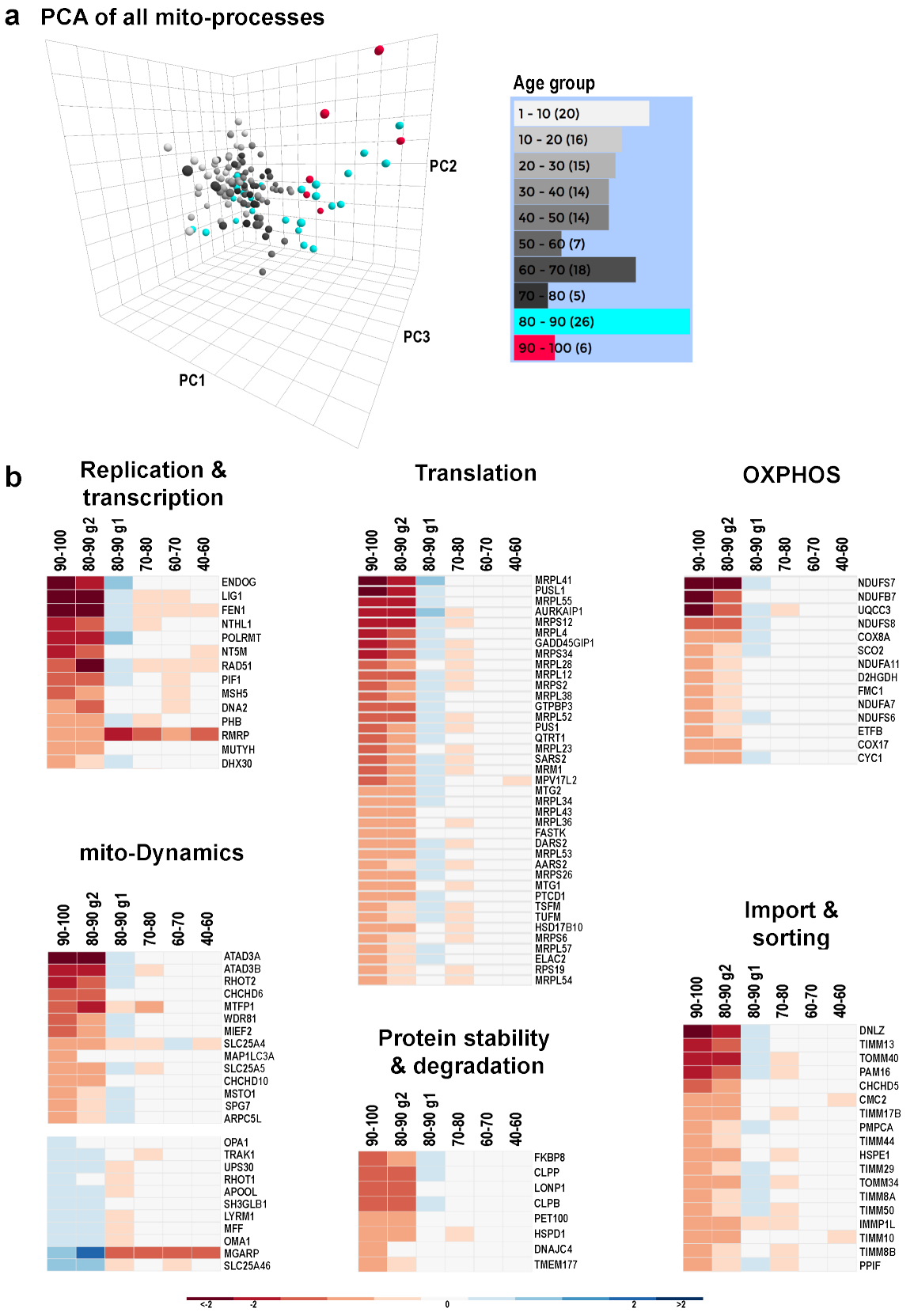
Supplementary Figure S1



Supplementary Figure S1: programmatic skeleton of the mitoXplorer web-platform. In the *back-end*, a MySQL database stores the mito-interactomes, as well as expression and mutation data that are publicly available. User-uploaded data are stored temporarily and only available to the user. A set of python-scripts connect to the MySQL database for data retrieval of both, mito-interactomes and expression and mutation data. The mitomodel script connects to the MySQL database directly for the visualization of the Interactome View. A set of scripts perform comparative analysis, for generating Comparative Plots, Heatmap and PCA visualization. In the *front-end*, a set of javascripts handle the visualizations of the plots: the 'interactome' and 'database' scripts handle the data presentation of the

mito-interactome and the available public data for the web-site; mitomodel visualizes the Interactome View and the scripts in the compare box are responsible for visualizing Comparative Plot, Heatmap and PCA. The CSS layer handles the css-styles of the page and finally, the HTML/PHP layer creates the actual interface for the user.

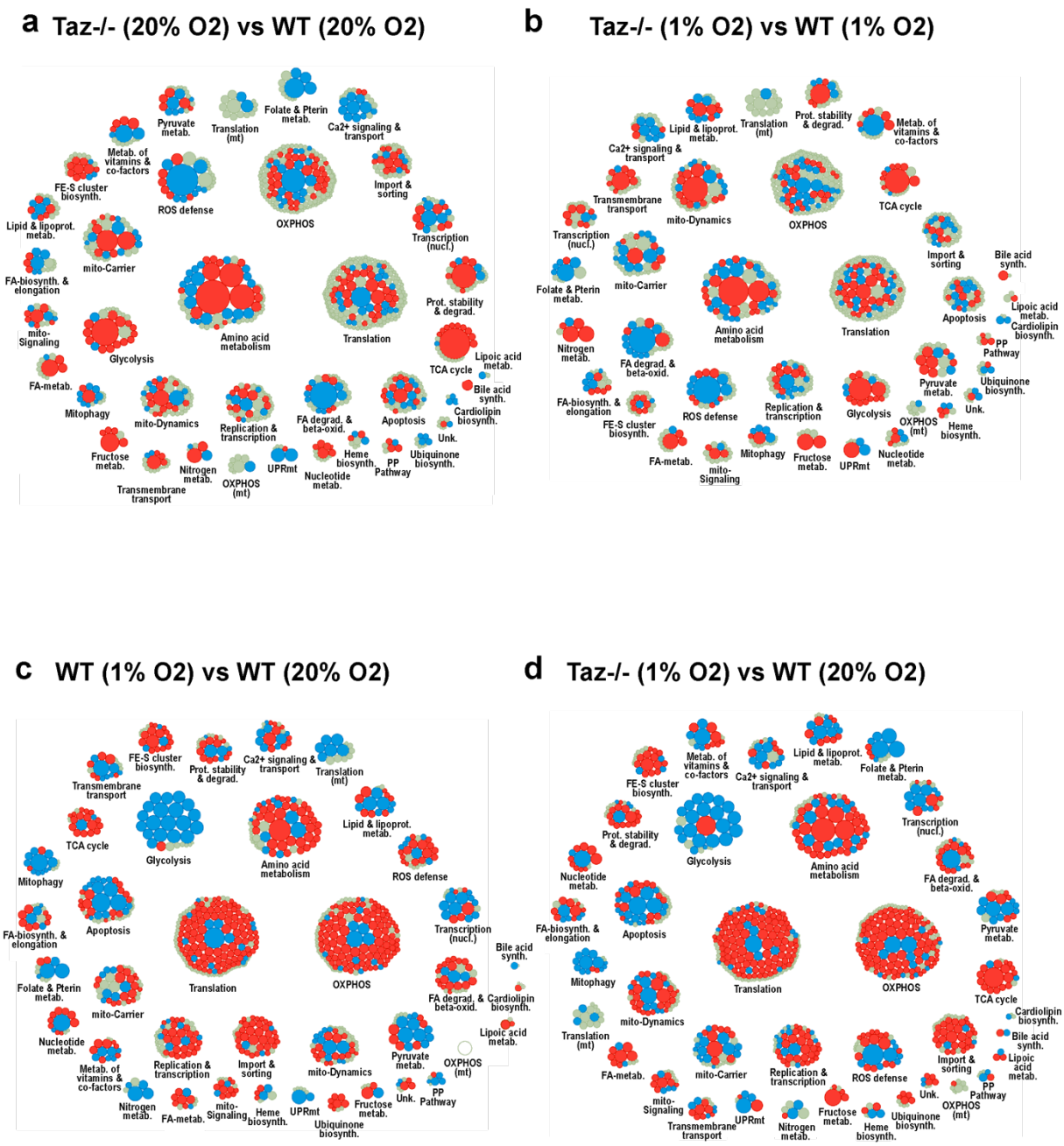
Supplementary Figure S2



Supplementary Figure S2: Human fibroblasts from healthy donors between ages 1 and 94 reveal strong down-regulation of mito-genes in different processes at ages above 85. To demonstrate the GROUP function of mitoXplorer, we use data from human fibroblasts from different age groups (1). A number of mitochondrial processes have been implicated in ageing and ageing-related diseases,

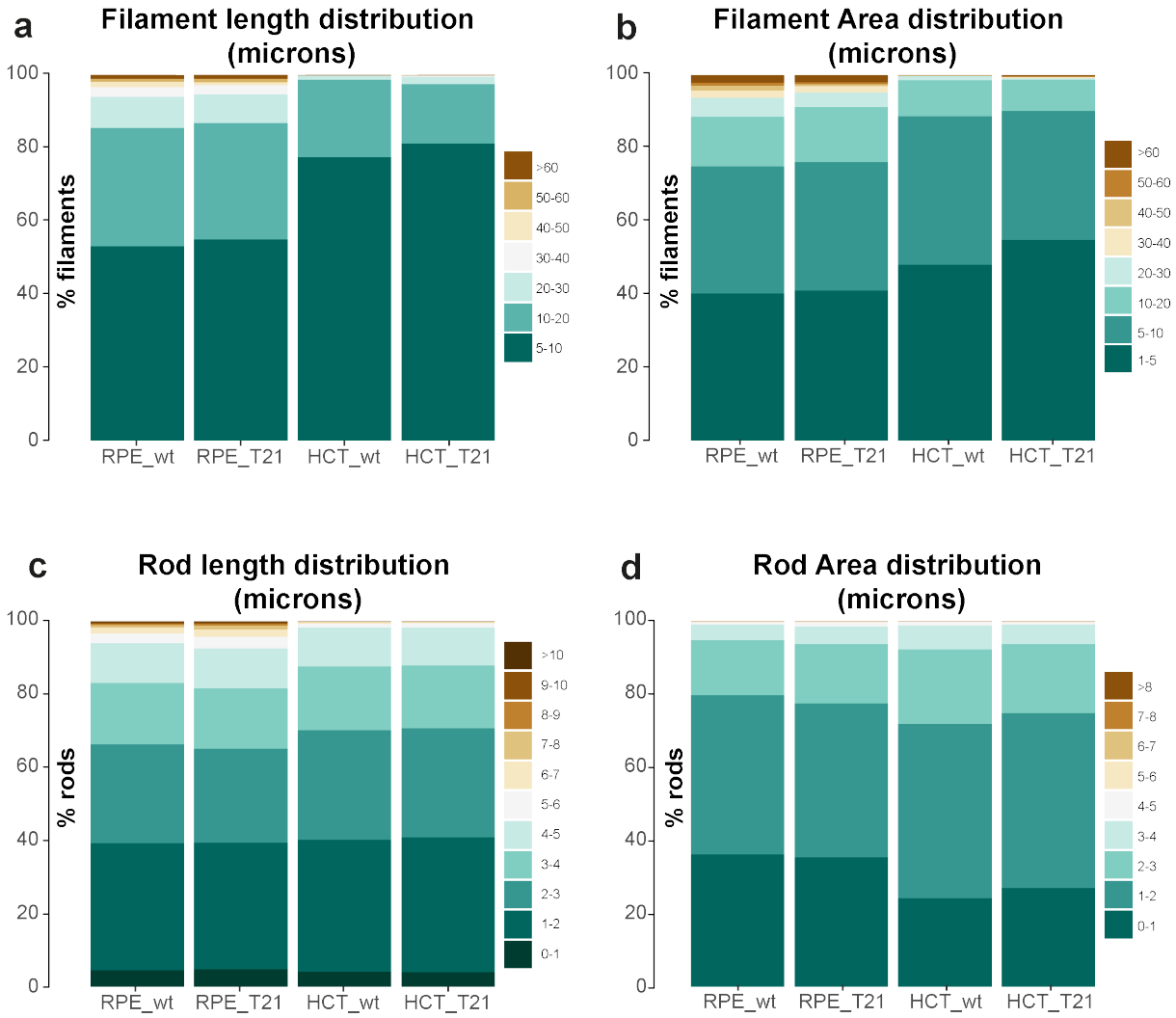
though the precise contribution of mitochondria to ageing is so far not clear (2) . However there seems to be an increase in mtDNA mutations and resulting impaired respiratory chain functions during ageing (3). We wanted to investigate whether the expression dynamics of mito-genes changes in the ageing tissue. To this end, we uploaded data from ageing human dermal fibroblasts from a study of Fleischer et al. (1). The authors chose this tissue because of its ease of availability. Moreover, the cells in the dermal layer are not prone to environmentally-induced mtDNA mutations and thus lifestyle-induced mito-gene expression changes (4). The dataset is composed of single replicates of 133 healthy donors from age 1 to 94. We calculated log2FC from RPKM values provided by the authors and uploaded the data to mitoXplorer. **(a)** We first performed a PCA analysis of mito-gene expression over all age groups. To our surprise, the mito-gene expression profiles seemed very robust up to the age of 80. Between 80 and 90 years of age, half the age group, as well as all individuals from age group 90-100 showed a very distinct expression profiles of many mito-genes and in most mito-processes. In fact, only few mito-processes were not affected. This was the case for all individuals older than 85 years except for one male of 88 years, who grouped with younger individuals. **(b)** In order to explore the nature of the de-regulation at high age, we generated groups using the mitoXplorer GROUP function. We focused on individuals between age 40 and 100. We merged age group 40-50 and 50-60 and split age group 80-90 in one group g1 with mito-gene expression dynamics consistent with younger individuals, as well as g2 with mito-gene expression profiles similar to those found in individuals above 90 years of age. Not unexpectedly we found that a number of mito-genes were generally down-regulated.

Supplementary Figure S3



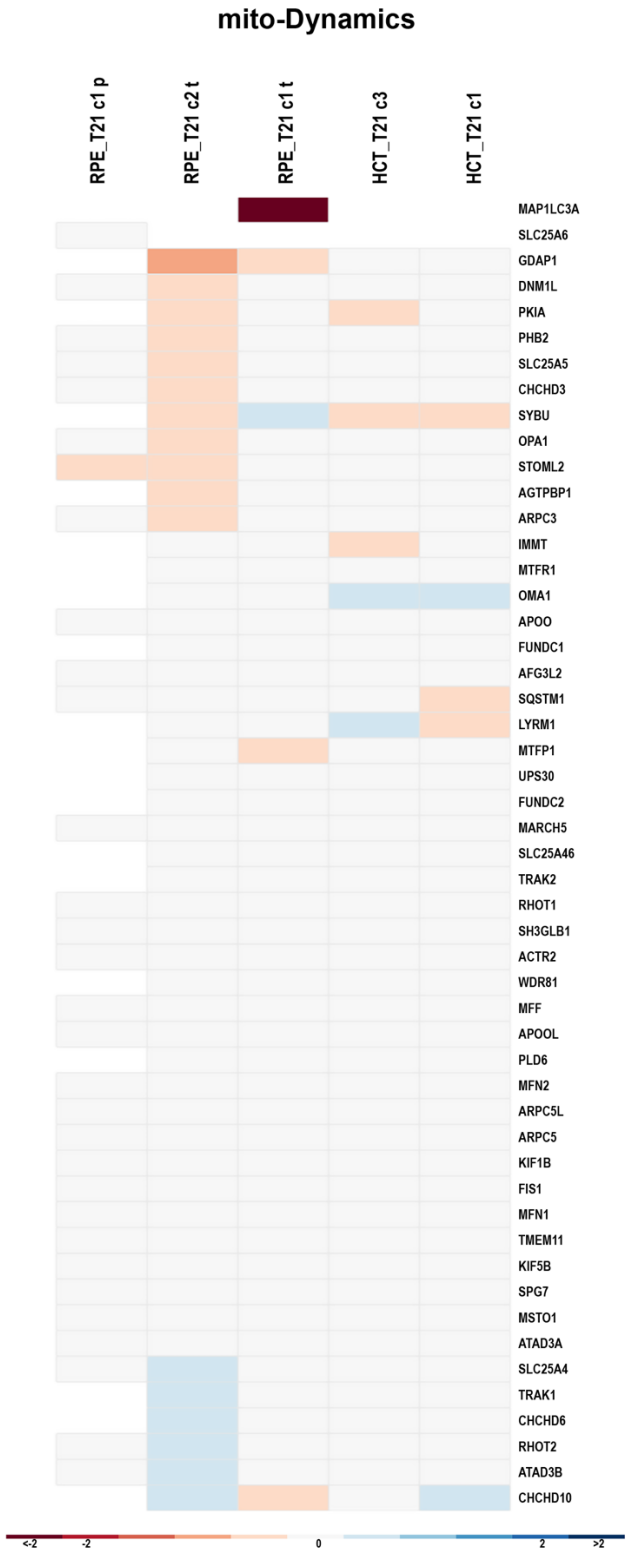
Supplementary Figure S3: Interactome Views of Tafazzin-deficient cells, as well as wild-type cells in normoxic and hypoxic conditions. (a) Tafazzin-deficient cells compared to wild-type cells. The expression dynamics from virtually all mito-processes is changed, with strongest perturbation in Amino acid metabolism, TCA cycle, Glycolysis, Fatty acid degradation & beta-oxidation, Mitochondrial carrier as well as Protein stability & degradation. **(b)** Taz-deficient cells compared to wild-type in hypoxic conditions. Changes in expression dynamics compared to normoxic conditions can for instance be observed in Mitochondrial dynamics, Transcription (nuclear), or UPRmt. **(c)** Differential expression dynamics of wild-type cells in normoxic and hypoxic conditions. In response to hypoxia, OXPHOS genes are down-regulated, while genes required for Glycolysis are induced. **(d)** Tafazzin-deficient cells in hypoxia are compared to wild-type cells in normoxic conditions. Strong changes can be observed for instance in ROS defense, Ca^{2+} Signaling & transport or UPRmt.

Supplementary Figure S4



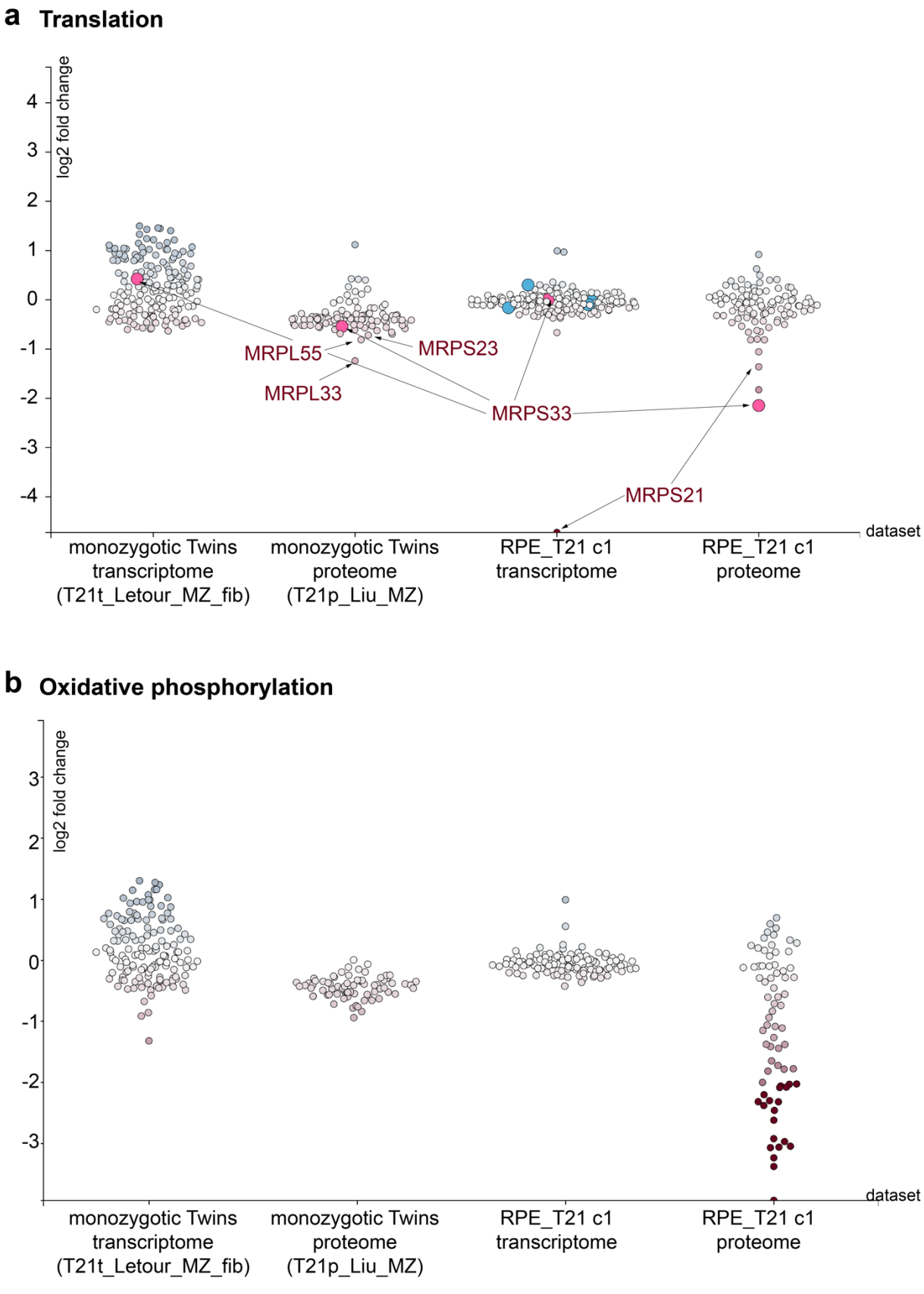
Supplementary Figure S4: Length and area distribution of filaments and rods in wild-type and T21 derived RPE1 and HCT116 cells. (a) Stacked bar-plots of filament length distribution of RPE1 wild-type (labeled RPE_wt), RPE1 21/3 (labeled RPE_T21), HCT116 wild-type (labeled HCT_wt) and HCT116 21/3 (labeled HCT_T21) cells. Overall, shorter filaments are more frequent in HCT116 than in RPE1 cells. In T21, filaments tend to be slightly shorter. **(b)** Stacked bar-plots of filament area distribution of RPE_wt, RPE_T21, HCT_wt wild-type and HCT_T21 cells. Overall, less area is occupied by filaments in HCT116 than in RPE1 cells. In HCT_T21 cells, a notably smaller area is assigned to filaments, while in RPE_T21 cells, this change is much less pronounced. **(c)** Stacked bar-plots of rod length distribution of RPE_wt, RPE_T21, HCT_wt and HCT_T21 cells. Overall, in the range between 4 and 10 microns, more rods are found in RPE1 cells. Between wild-type and T21 cells, no real length difference is observable. **(d)** Stacked bar-plots of rod area distribution of RPE_wt, RPE_T21, HCT_wt and HCT_T21 cells. Overall, there is a tendency of slightly larger rod areas in HCT116 cells. In HCT116 cells, rods seem to occupy slightly smaller areas when carrying the extra copy of chromosome 21. Data were averaged over the two clones of RPE_T21 and HCT_T21, respectively.

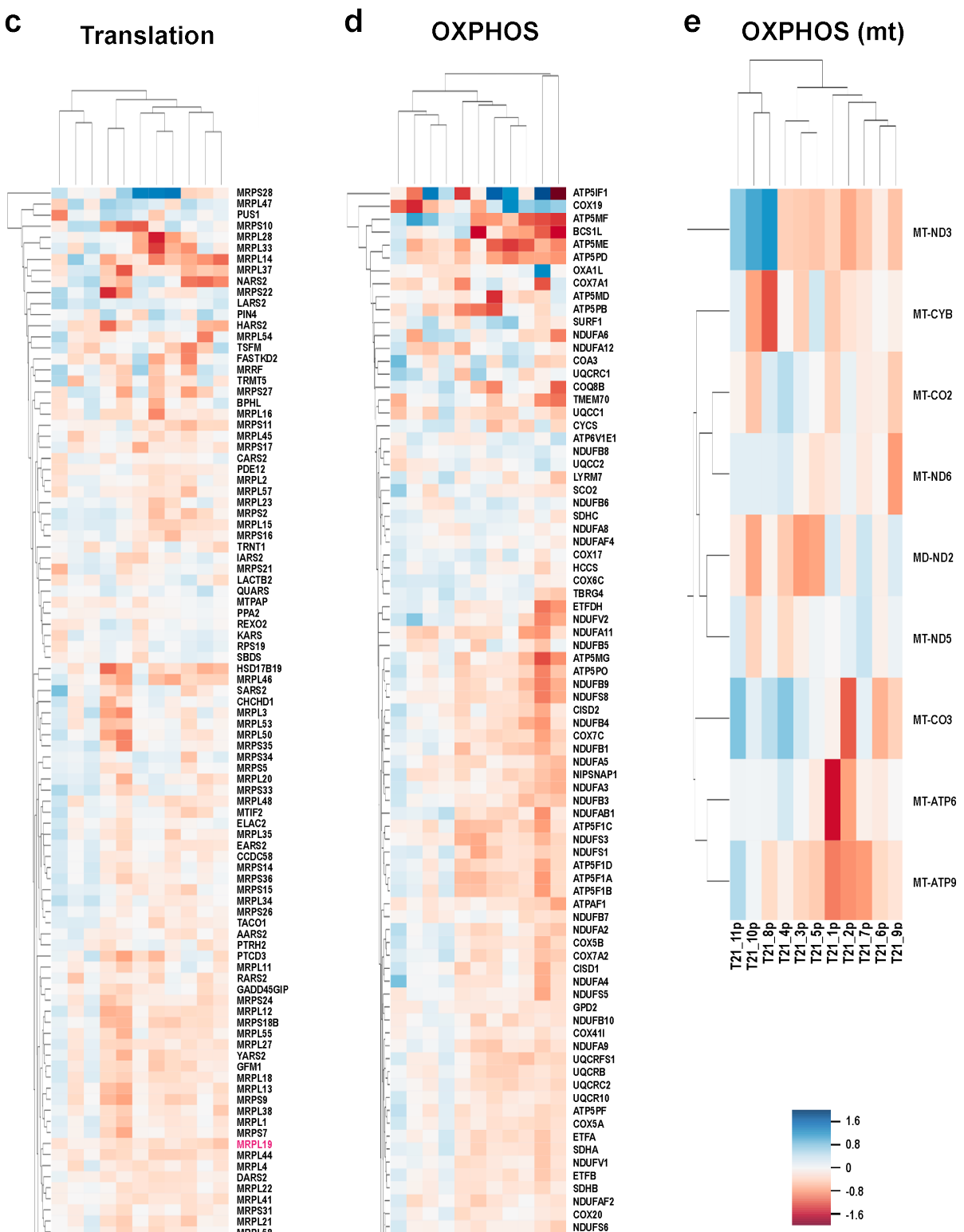
Supplementary Figure S5



Supplementary Figure S5: The mito-gene GDAP1 is consistently down-regulated in RPE_T21 clones. Shown is the heatmap of the Comparative Analysis page of mitoXplorer. Both clones of the RPE21_T21 cells show significant down-regulation of the GDAP1 gene; clones 1 and 3 of HCT_T21 cells also show slight, though statistically non-significant reduction of this gene (RPE_T21 c1: log2FC = -1.103; RPE_T21 c2: log2FC = -1.604; HCT_T21 c1: -0.238; HCT_T21 c3: -0.059).

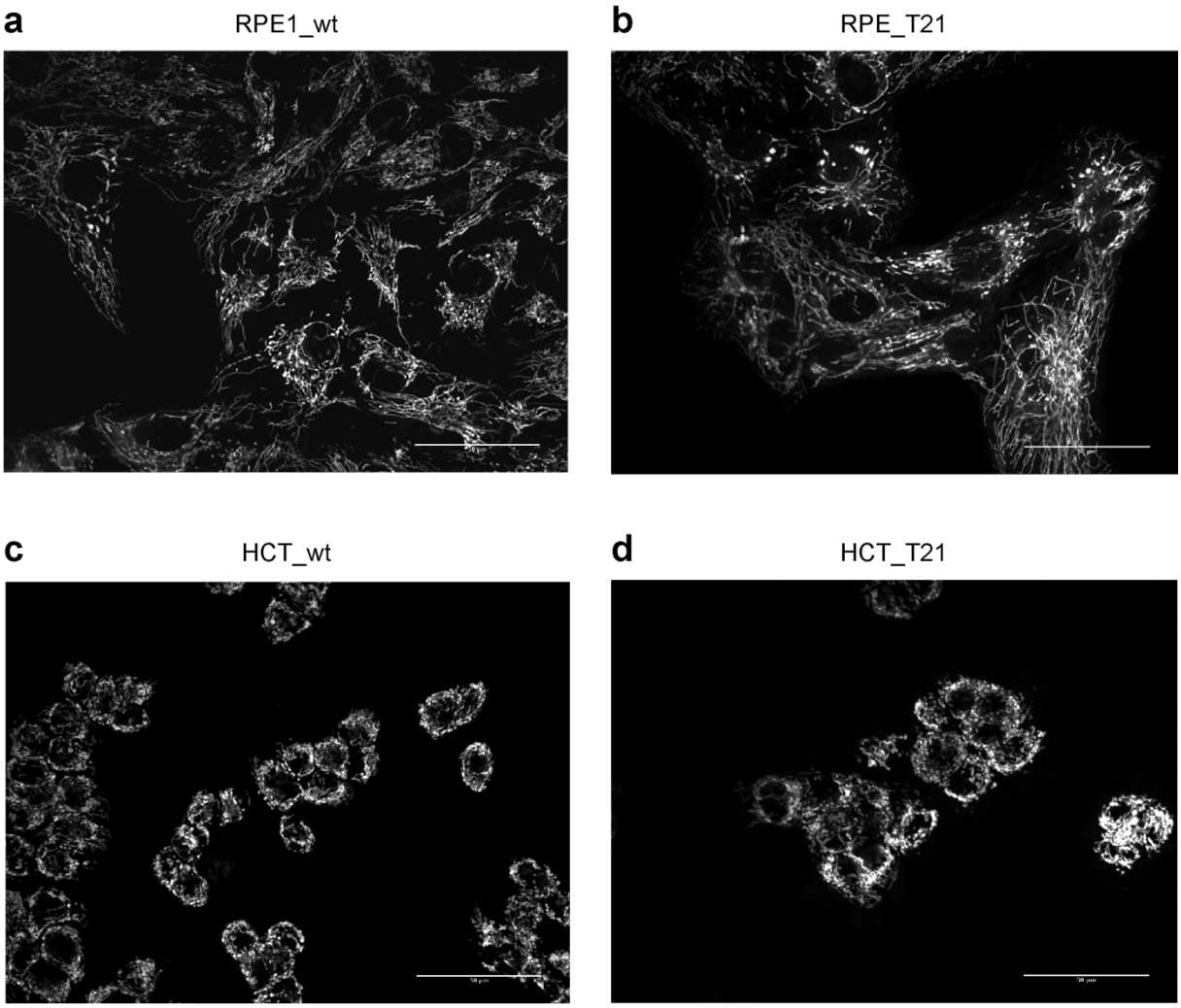
Supplementary Figure S6




c
Translation
MRPS28
MRPL47
PUS1
MRPS10
MRPL28
MRPL33
MRPL14
MRPL37
NARS2
MRPS22
LARS2
PIN4
HARS2
MRPL54
TSFM
FASTKD2
MRRF
TRMT5
MRPS27
BPHL
MRPL16
MRPS11
MRPL45
MRPS17
CARS2
PDE12
MRPL2
MRPL57
MRPL23
MRPS2
MRPL15
MRPS16
TRNT1
IARS2
MRPS21
LACTB2
QUARS
MTPAP
PPA2
REXO2
KARS
RPS19
SBDS
HSD17B19
MRPL46
SARS2
CHCHD1
MRPL3
MRPL53
MRPL50
MRPS35
MRPS34
MRPS5
MRPL20
MRPS33
MRPL48
MTIF2
ELAC2
MRPL35
EARS2
CCDC58
MRPS14
MRPS36
MRPS15
MRPL34
MRPS26
TACO1
AARS2
PTRH2
PTCD3
MRPL11
RARS2
GADD45GIP
MRPS24
MRPL12
MRPS18B
MRPL55
MRPL27
YARS2
GFM1
MRPL18
MRPL13
MRPS9
MRPL38
MRPL1
MRPS7
MRPL19
MRPL44
MRPL4
DARS2
MRPL22
MRPL41
MRPS31
MRPL21
MRPL58
MRPL39
TRMT10C
TUFM
MRPL24
MRPS6
MRPL17
MRPL43
MRPL49
DAP3
MRPS23
MRPL30
LRPPRC
MRPL9
PNPT1
PRORP
SLIRP
T21_11p
T21_3p
T21_4p
T21_1p
T21_2p
T21_8p
T21_10p
T21_6p
T21_9p
T21_5p
T21_7p
d
OXPHOS
ATP5IF1
COX19
ATP5MF
BCS1L
ATP5ME
ATP5PD
OXA1L
COX7A1
ATP5MD
ATP5PB
SURF1
NDUFA6
NDUFA12
COA3
UQCRC1
COQ8B
TMEM70
UQCC1
CYCS
ATP6V1E1
NDUFB8
UQCC2
LYRM7
SCO2
NDUFB6
SDHC
NDUFA8
NDUFAF4
COX17
HCCS
COX6C
TBRG4
ETFDH
NDUFV2
NDUFA11
NDUFB5
ATP5MG
ATP5PO
NDUFB9
NDUFS8
CISD2
NDUFB4
COX7C
NDUFB1
NDUFA5
NIPSNAP1
NDUFA3
NDUFB3
NDUFAB1
ATP5F1C
NDUFS3
NDUFS1
ATP5F1D
ATP5F1A
ATP5F1B
ATPAF1
NDUFB7
NDUFA2
COX5B
COX7A2
CISD1
NDUFA4
NDUFS5
GPD2
NDUFB10
COX41l
NDUFA9
UQCRFS1
UQCRB
UQCRC2
UQCR10
ATP5PF
COX5A
ETFA
SDHA
NDUFV1
ETFB
SDHB
NDUFAF2
COX20
NDUFS6
NDUFA10
NDUFS4
COX6B1
UQCRH
COA6
NDUFA13
NDUFA7
SCO1
NDUFB11
NDUFAF3
NDUFS7
NDUFC2
CYC1
NDUFS2
T21_11p
T21_8p
T21_3p
T21_4p
T21_5p
T21_10p
T21_9p
T21_6p
T21_7p
T21_1p
T21_2p
e
OXPHOS (mt)
MT-ND3
MT-CYB
MT-CO2
MT-ND6
MD-ND2
MT-ND5
MT-CO3
MT-ATP6
MT-ATP9
T21_11p
T21_10p
T21_8p
T21_4p
T21_3p
T21_5p
T21_1p
T21_2p
T21_7p
T21_6p
T21_9p
1.6
0.8
0
-0.8
-1.6

Supplementary Figure S6: mitoXplorer analysis of the mito-processes 'Translation', 'Oxidative Phosphorylation' and 'Oxidative Phosphorylation (mt)' of monozygotic twins discordant for T21 and 11 unrelated trisomy 21 patients. (a) The mRNA of mitoribosome small subunit component MRPS21 is strongly down-regulated in RPE_T21 cells. In monozygotic twins discordant for T21, other subunits of the small and large mitoribosome are down-regulated (T21_MZ fibroblasts: T21_Letour_MZ_fib, T21_Liu_MZ). Mitoribosome proteins are more mildly affected in T21_MZ fibroblasts. **(b)** Oxidative phosphorylation components encoded in the nucleus are downregulated on protein level in both, RPE_T21, as well as T21_MZ fibroblasts, whereby deregulation is milder in T21_MZ. In both conditions, the Oxidative phosphorylation transcriptome is mostly unaffected. **(c - e)** Mito-protein expression dynamics from fibroblasts of unrelated trisomy 21 patients in the mito-processes 'Translation' **(c)**, 'Oxidative phosphorylation' **(d)** and 'Oxidative phosphorylation (mt)' **(e)**. Protein levels of mitochondrial- and nuclear-encoded subunits of the respiratory chain is very heterogeneous between unrelated individuals with trisomy 21; yet, in most trisomy 21 patients several mitochondrial-, as well as nuclear-encoded subunits of the respiratory chain are down-regulated on protein level. A similar heterogeneity can be observed for protein expression levels involved in translation of mitochondrial proteins **(c)**. At least a few proteins involved in 'Translation' are down-regulated in each individual with trisomy 21. The only consistently, though sometimes only mildly reduced protein in this mito-process is mitoribosomal protein MRPL19 (highlighted in red), while MRPS21 shows variable expression patterns in unrelated trisomy 21 patients. Protein expression data are taken from Liu et al. (5). The log2FC of individual trisomy 21 patients versus the average of 11 unrelated healthy individuals was calculated using the log2 of the SWATH-MS intensities provided by the authors.

Supplementary Figure S7



Supplementary Figure S7: Mitochondrial network of wild-type and T21 cells. MitoTracker stainings **(a-d)** of RPE_wt **(a)** and RPE_T21 **(b)**, as well as HCT_wt **(c)** and HCT_T21 **(d)**. **(a, b)** The mitochondrial network is largely intact in RPE_T21 cells, with only slightly lower percentage filaments and an increased number of swollen mitochondria. **(c, d)** In HCT116 cells, the mitochondrial network is overall less abundant, with more rod-like and fragmented mitochondria (puncta). With trisomy 21, cells show an even more pronounced presence of rods at the cost of longer filaments, as well as more puncta and swollen mitochondria. The scale bar is 50 μm. Mitochondria were stained with MitoTracker deep Red FM from Invitrogen. Staining was done in 96-well plates. The cells were incubated for 30 min at 30°C with 100 nM MitoTracker dye prior to fixation. Cells were fixed with 3% PFA in DMEM for 5 min at room temperature. After washing with 1xPBS, 1xPBS with 0.01% sodium azide was added. Plates were stored at 4°C in the dark. Imaging was carried out on an inverted Zeiss Observer.Z1 microscope with a spinning disc and 473 nm, 561 nm and 660 nm argon laser lines. The images were captured automatically on multiple focal planes (step size 700 nm) with a 40x magnification air objective. Image stacks were Z-projected using Fiji for further analysis.

References

1. Fleischer,J.G., Schulte,R., Tsai,H.H., Tyagi,S., Ibarra,A., Shokhirev,M.N., Huang,L., Hetzer,M.W. and Navlakha,S. (2018) Predicting age from the transcriptome of human dermal fibroblasts. *Genome Biol.*, **19**, 221–8.

2. Bratic,I. and Trifunovic,A. (2010) Mitochondrial energy metabolism and ageing. *Biochim. Biophys. Acta*, **1797**, 961–967.

3. Park,C.B. and Larsson,N.-G. (2011) Mitochondrial DNA mutations in disease and aging. *J. Cell Biol.*, **193**, 809–818.

4. Stout,R. and Birch-Machin,M. (2019) Mitochondria's Role in Skin Ageing. *Biology (Basel)*, **8**, 29.

5. Liu,Y., Borel,C., Li,L., Müller,T., Williams,E.G., Germain,P.-L., Buljan,M., Sajic,T., Boersema,P.J., Shao,W., *et al.* (2017) Systematic proteome and proteostasis profiling in human Trisomy 21 fibroblast cells. *Nat Commun*, **8**, 1212.

6.2 Integrative analysis and machine learning on cancer genomics data using the Cancer Systems Biology Database (CancerSysDB)

Having diverse roles in cellular functions such as bioenergetic pathways, ROS defense and programmed cell death, mitochondria functions and metabolism has long been implicated in different steps of oncogenesis like malignant transformation and tumor progression. In this study, we have integrated a workflow specific for the analysis of mitochondrial genes into Cancer Systems Biology Database (CancerSysDB) that hosts public cancer dataset.

CancerSysDB is a platform that allows user to make customized queries and perform integrated analyses across multiple data types (transcription data, mutation data, clinical data, etc) and cancer cohorts from The Cancer Genome Atlas (TCGA) research network. This saves users from dealing with the diverse file formats and structures of data obtained through the Genomic Data Commons (GDC) Data Portal, and makes big data analytics these datasets readily accessible to them.

We have developed a workflow that integrates a manually curated human mitochondrial interactome and displayed analyzed data on an interactive dashboard as a visual data mining tool. It allows the exploration of differential expression of genes of various mitochondrial function, as well as correlation analysis with clinical features. With this workflow, we discovered a remarkable difference in the expression of Tricarboxylic acid (TCA) cycle genes, Succinate-CoA ligase subunits SUCLG1 and SUCLG2, in KIRP (kidney renal papillary cell carcinoma) patients of later stages. Succinate-CoA ligase is an essential enzyme in the production of ATP. This observation proposed that SUCLG1, together with SUCLG2 as suggested in a previous study, could be a promising indicator for later stage clear cell renal carcinomas. It also demonstrated how such a visual data mining tool could provide insights to the role mitochondrial functions in the development and progression of cancer.

Krempel et al. BMC Bioinformatics (2018) 19:156
https://doi.org/10.1186/s12859-018-2157-7

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Integrative analysis and machine learning on cancer genomics data using the Cancer Systems Biology Database (CancerSysDB)

Rasmus Krempel[1], Pranav Kulkarni[2], Annie Yim[3,4], Ulrich Lang[1], Bianca Habermann[3,4] and Peter Frommolt[2*]

Abstract

Background: Recent cancer genome studies on many human cancer types have relied on multiple molecular high-throughput technologies. Given the vast amount of data that has been generated, there are surprisingly few databases which facilitate access to these data and make them available for flexible analysis queries in the broad research community. If used in their entirety and provided at a high structural level, these data can be directed into constantly increasing databases which bear an enormous potential to serve as a basis for machine learning technologies with the goal to support research and healthcare with predictions of clinically relevant traits.

Results: We have developed the Cancer Systems Biology Database (CancerSysDB), a resource for highly flexible queries and analysis of cancer-related data across multiple data types and multiple studies. The CancerSysDB can be adopted by any center for the organization of their locally acquired data and its integration with publicly available data from multiple studies. A publicly available main instance of the CancerSysDB can be used to obtain highly flexible queries across multiple data types as shown by highly relevant use cases. In addition, we demonstrate how the CancerSysDB can be used for predictive cancer classification based on whole-exome data from 9091 patients in The Cancer Genome Atlas (TCGA) research network.

Conclusions: Our database bears the potential to be used for large-scale integrative queries and predictive analytics of clinically relevant traits.

Background

Large-scale cancer genome studies based on Next-Generation Sequencing (NGS) technology have enabled extensive research on tumorigenesis and treatment rationales [14]. The amount of data that has been generated and made available contrasts its limited accessibility to the research community. There is an increasing demand for customized queries to the data in a way that is accessible to scientists and physicians without any knowledge in bioinformatics. Genomic data from studies in The Cancer Genome Atlas (TCGA) research network obtained through the Genomic Data Commons (GDC) Data Portal (https://portal.gdc.cancer.gov) are available for multiple molecular layers and are provided in formats processed through appropriate software packages for the analysis of the raw data for every data type. The size of these processed data is orders of magnitude smaller than the raw data, in particular for whole-genome sequencing experiments, but provided in a diverse range of file formats in which the data are variably well structured. Thus, it is particularly challenging to transform these file-based data into a structure which allows a technically reasonable way to integrate data obtained by multiple technologies with manually curated data recorded in a clinical context. This underlines the need for highly flexible database structures which are suitable to model data from TCGA studies, but are generic enough to also combine TCGA data with locally acquired data obtained in a clinical context.

We present here the newly developed Cancer Systems Biology Database (CancerSysDB) portal which allows integrated analyses across multiple data types and across multiple cancer cohorts from The Cancer Genome Atlas (TCGA) research network, but also from locally acquired data in a clinical context. With its current

* Correspondence: peter.frommolt@uni-koeln.de
[2]Bioinformatics Facility, CECAD Research Center, University of Cologne, Cologne, Germany
Full list of author information is available at the end of the article

workflows, our system allows fast integrative analysis of whole-exome (WXS) and transcriptome (RNA-Seq) sequencing data. By making use of standardized JSON-based meta data formats, the CancerSysDB can be integrated into existing analysis workflows. The CancerSysDB enables highly structured organization of data from multi-OMICS technologies and makes them accessible for big data analytics on the entirety of all data ever processed on a particular site. Conceptually, this includes the prediction of clinically relevant parameters such as therapeutic response from existing pharmacogenomic data in the CancerSysDB.

Methods

Implementation

The CancerSysDB was written in Groovy on the Grails framework based on the JVM stack which bundles state-of-the-art web frameworks behind a simple interface. The CancerSysDB is a web application which needs a database instance and an application server and can run Linux shell scripts and other executables from a command line. The data source is behind a hibernate facade keeping the system independent from the database implementation used and the optimization in the background. The delivered versions are based on a docker file to automatically build an environment and run the database application for personal use. A demo instance can be used to make personalized queries to the database using publicly available TCGA data. The source code of the CancerSysDB is available on GitHub (https://github.com/RRZK/CancerSysDB).

The system can be configured to run in two different modes. The *public mode* can be used to query publicly available data without any login. The publicly available main instance of the CancerSysDB available on http://cancersys.uni-koeln.de is running in public mode and provides access to data on 11,410 patients from the Cancer Genome Atlas (TCGA) research network. This instance includes data on somatic mutations (based on WXS data), differential gene expression (based on comparative RNA-Seq analysis between tumors and tissue-derived normals), somatic copy number alterations (based on Affymetrix SNP 6.0 microarrays) as well as all clinically derived annotations of the TCGA patient data. These data types provide a powerful basis for arbitrary queries defined by the user. All TCGA data types provided through the CancerSysDB are open access data and can be obtained from the TCGA data portal without exclusive access. Users have to adhere to the TCGA data access policies that apply to these open access data (https://gdc.cancer.gov/access-data/data-access-policies). On the other hand, the *private mode* requires a login for any interaction. This mode is strongly recommended if you are working with restricted data. The University of Cologne is operating a *private mode* instance of the CancerSysDB for the organization of genomic data from in-house studies. It is used in combination with the recently published cancer genomics data processing workflow system *QuickNGS Cancer* [1] which extends our NGS bioinformatics suite *QuickNGS* [15] and allows highly scalable and standardized analysis of cancer NGS data with minimum hands-on analysis time. Various features of the CancerSysDB are compared to those of other cancer genome data integration tools in Table 1.

Data model and queries

The maintainer of a CancerSysDB instance can describe the connection between data and the main structure of the application in JSON files to bring the context structure of data into the database. The database consists of four main data types:

- *Structural data* manages the patients and samples,
- *Molecular data* is derived from cancer genome analysis,
- *Clinical data* is associated to the clinical course of a patient's disease,
- *Genomic annotation* provides information on genes and meta data about these genes.

Table 1 Comparison of various features of the CancerSysDB with those of other cancer genomics data integration tools

	CancerSysDB	TCGAbiolinks	RTCGA	cBio portal
GUI	Web framework based on Groovy/Grails	Based on Shiny	None	Web framework based on Spring Java
Query schema	Hibernate	R scripting	R scripting	SQL
Data upload	Parametrized CSV file upload	Direct access to GDC through API	Data packages available on Bioconductor	CSV files plus meta file
Query definition	JSON-based	Combination of R commands	Combination of R commands	REST-based API
Portability	Native Docker implementation	Hosted on Bioconductor	Hosted on Bioconductor	Hosted on GitHub
Programming skills required	No	Yes	Yes	No

The data model and principles how to develop database queries is further described on GitHub at https://github.com/RRZK/CancerSysDB/tree/master/web-app/data/Workflows . Data can be uploaded through the API or manually with the web front end. The API enables automated uploads from processing infrastructures like high performance computing (HPC) environments. A collection of Python scripts for upload automation is delivered with the database. We are using these scripts to link the analysis workflows on the *QuickNGS Cancer* pipeline to the CancerSysDB. The internal design of the web application empowers the maintainer to easily extend the data model, extend the import behavior and integrate custom data structures.

The maintainer of an instance of the CancerSysDB is provided with a fully controllable environment for the development of custom workflows. A custom workflow can be described in a JSON file and extended with analysis scripts and static data in a zip file which can be dynamically uploaded into the database (documentation available on the GitHub). The actual data is retrieved using queries written in the Hibernate Query Language (HQL) and the results of the queries are saved as CSV files in order to increase reproducibility on a dynamically updated database. Subsequent computations can rely on arbitrary executables in a Linux environment. The container architecture provides the encapsulation for the workflows. To control the command line based execution, packages and libraries can be installed on creation of the docker container or wrapped directly into the files to be executed by the workflow.

Data preparation

All TCGA data were obtained as level 3 data from the Legacy Archive of The Cancer Genome Atlas (TCGA) data portal. Data on somatic mutations were based on whole-exome sequencing with MAF files obtained from the Firehose pipeline of the Genome Data Analysis

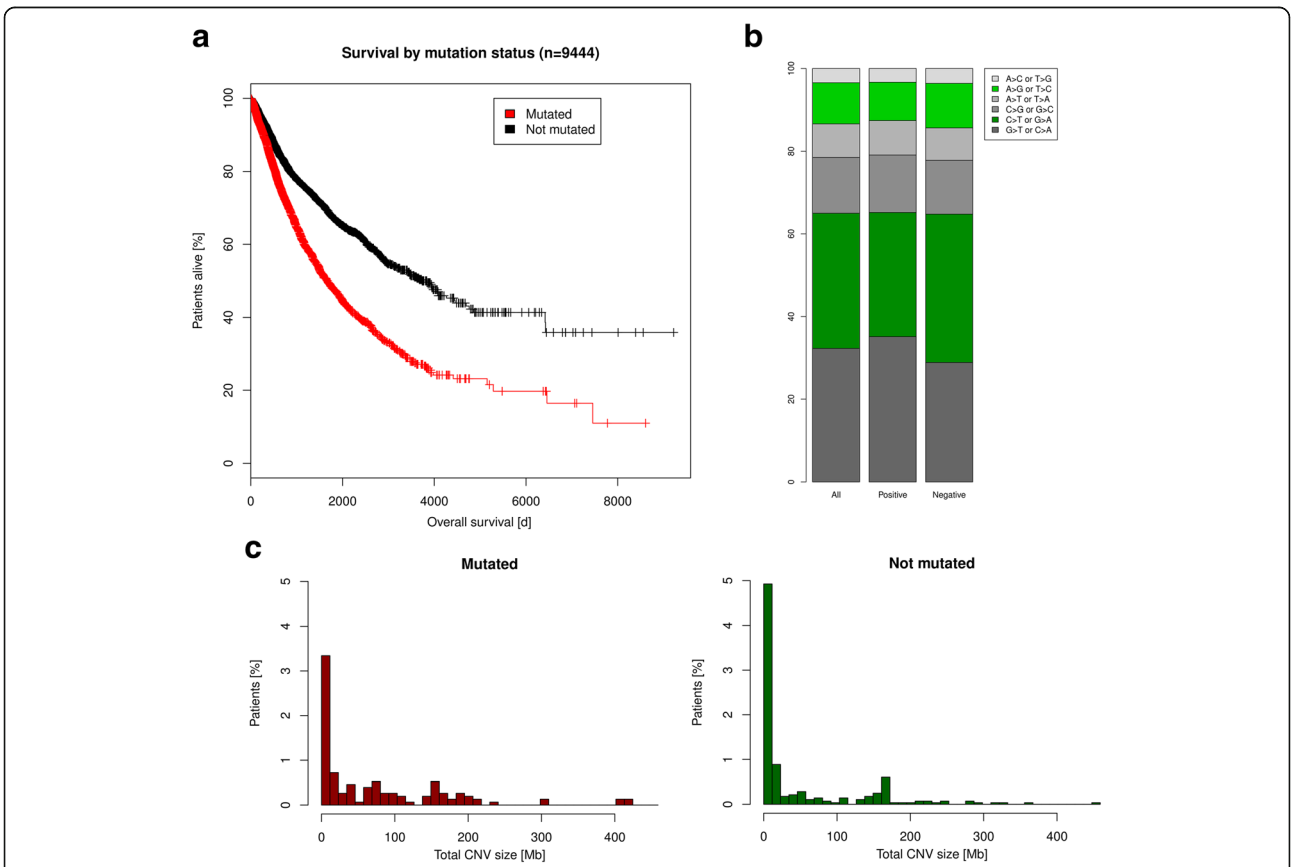


Fig. 1 Analysis results for workflows splitting multiple TCGA cohorts into TP53-mutant and non-mutant patients: **a** Overall survival is significantly different between TP53-mutant (red curve) and non-mutant patients (black curve) with a more favorable for non-mutant patients (gain in median survival: 2066 days, $p < 0.0001$, $n = 9444$). **b** The distribution of the mutations types in lung adenocarcinoma is strongly shifted towards an increase of G > T transversions in TP53 mutant compared to non-mutant patients ($p = 0.0006$, $n = 584$). **c** Genomic stability is quantified in terms of the overall size of somatic copy number alterations (sCNA) compared between tumor and normal. sCNA are considered as genomic amplifications above a level of 3 and as genomic deletions below a level of 1 for the signal ratio between tumor and paired normal sample. The difference between TP53 mutant and non-mutant patients is highly significant in glioblastoma multiforme ($p = 0.0132$, $n = 379$)

Center (GDAC) at the Broad Institute. Data on somatic copy number alterations were based on the SNP 6.0 microarray platform (Affymetrix Inc., CA, USA) given as genomic segments of equal copy number derived from the Circular Binary Segmentation (CBS) algorithm [8]. For gene expression analysis, raw RNA-Seq read counts were re-processed and compared between tumor tissues and tissue-derived normal samples using version 1.21.1 of the DESeq2 algorithm and its implementation as an R package [6]. These tissue-derived normal controls are available from only a minority of the patients in TCGA, but we consider them more suitable for a comparative tumor/normal analysis than the blood-derived normals existing for most patients. The currently existing workflows were implemented using version 3.3.3 of the functional statistics language R (http://www.r-project.org). The random forest workflow was implemented with the R package 'randomForest', version 4.6–12.

Results and discussion

In order to demonstrate how the CancerSysDB can help to obtain analysis results of immediate relevance for research projects or clinical prognosis, we showcase the analytical power by three example queries, by one machine learning workflow on the CancerSysDB and by an interactive workflow of visualizing mitochondrial pathways. The results of these showcases can be reproduced using the query and analysis source code provided in Additional file 1.

TP53-dependent analysis of overall survival, genome stability, and mutation types

The tumor suppressor gene TP53 is the most frequently deleted and mutated gene across all tumor types [3]. In the TCGA cancer cohorts, its mutation rate is highly variable and ranges up to > 75% in some cancer types [16]. The CancerSysDB enables comparative genomic analyses of patients with and without mutations in TP53 by employing three different query workflows which we operate across > 11,000 patients from 33 TCGA studies.

- *Overall survival depending on mutation status:* Across all TCGA cohorts, patients with a mutation in TP53 show an unfavorable prognosis regarding overall survival compared to TP53 wild type patients ($p < 0.0001$, $n = 9444$; Fig. 1a; Table 2a).
- *Transversions and transitions depending on mutation status:* The somatic mutational landscape of patients with lung adenocarcinoma exhibits a significant shift towards G > T transversions when compared between patients with and without mutations in TP53 ($p = 0.0006$, $n = 584$; Fig. 1b; Table 2b). G > T transversions have been shown to be induced by oxidative stress in lung cancers of tobacco smokers [12]. Their enrichment in patients with mutated TP53 is likely caused by the impaired induction of apoptosis upon these exogenic damages.

Table 2 Results of TP53-dependent analysis of genomic and clinical characteristics

(a)

	Patients	Events	5-year survival rate [%]	Median survival	95% CI
TP53 mutant	3772	1237	47.4	1670	[1526; 1818]
TP53 non-mutant	5672	1128	66.9	3736	[3262; 4267]

(b)

	Patients	CNAs [Mb]
TP53 mutant	133	74.5
TP53 non-mutant	246	50.5

(c)

		TP53			ATM		
VarType	All	Mutant [%] ($n = 320$)	Non-mutant [%] ($n = 265$)	*p*-value	Mutant [%] ($n = 49$)	Non-mutant [%] ($n = 536$)	*p*-value
A > C or T > G	3.5	3.3	3.6	< 0.0001	3.9	3.4	0.2160
A > G or T > C	9.9	9.2	10.7	< 0.0001	9.6	9.9	0.7695
A > T or T > A	8.1	8.4	7.8	0.0005	8.6	8.1	0.4584
C > G or G > C	13.6	13.9	13.1	< 0.0001	13.2	13.6	0.3790
C > T or G > A	32.7	30.0	36.0	< 0.0001	28.6	33.0	0.5121
G > T or C > A	32.3	35.2	28.8	0.0001	36.0	32.0	0.4940

- *Genomic complexity depending on mutation status:* Among the patients with glioblastoma multiforme, those with TP53 mutations are characterized by, on average, stronger genomic instability than the TP53 wild type patients ($p = 0.0132$, $n = 379$; Fig. 1c; Table 2c). This general loss of genomic stability in TP53-mutated patients can be attributed to the role of TP53 as a mediator of apoptosis in response to somatically acquired DNA damage of cancer cells and has been described in previous studies [7].

Technically, the workflows start with database queries for the TCGA barcodes of the patients with and without TP53 mutations. Subsequent queries obtain the overall survival of all patients, the overall size of genomic copy number aberrations in glioblastoma multiforme, and a list of all mutations in the cohort of patients with lung adenocarcinoma. These query results are stored as CSV files on the CancerSysDB server and are processed through workflow analysis scripts to restructure, analyze and visualize the data. The scripts for this TP53-dependent analysis of TCGA data were written in the functional statistics language R.

Prediction of cancer types with random forests

In order to demonstrate the potential of our database for predictive analytics of clinically relevant traits, we have evaluated a workflow for the classification of a yet

Table 3 Classes of carcinomas used for random forest prediction of cancer types

Class name	TCGA cohorts	Sample size		
		Total	Training set	Test set
Adrenal gland	Adrenocortical carcinoma (ACC)	271	179	92
	Pheochromocytoma and paraganglioma (PCPG)			
Bladder	Urothelial carcinoma (BLCA)	411	272	139
Brain	Lower grade glioma (LGG)	515	340	175
Breast	Breast invasive carcinoma (BRCA)	1077	711	366
Gastrointestinal	Esophageal carcinoma (ESCA)	1237	817	420
	Stomach adenocarcinoma (STAD)			
	Colon adenocarcinoma (COAD)			
	Rectum adenocarcinoma (READ)			
	Cholangiocarcinoma (CHOL)			
Head & Neck	Head and neck squamous cell carcinoma (HNSC)	590	390	200
	Uveal melanoma (UVM)			
Hematologic	Acute myeloid leukemia (LAML)	321	212	109
	Diffuse large B-cell lymphoma (DLBC)			
	Thymoma (THYM)			
Kidney	Kidney Chromophobe (KICH)	738	488	250
	Renal clear cell carcinoma (KIRC)			
	Renal papillary cell carcinoma (KIRP)			
Liver	Hepatocellular carcinoma (LIHC)	321	212	109
Ovary	Ovarian serous cystadenocatcinoma (OV)	437	289	148
Pancreas	Pancreatic adenocarcinoma (PAAD)	184	122	62
Prostate	Prostate adenocarcinoma (PRAD)	498	329	169
Skin	Cutaneous melanoma (SKCM)	104	69	35
Testis	Testicular germ cell tumors (TGCT)	150	99	51
Thoracic	Lung adenocarcinoma (LUAD)	1143	755	388
	Lung squamous cell carcinoma (LUSC)			
	Mesothelioma (MESO)			
Thyroid	Thyroid carcinoma (THCA)	496	327	169
Uterus	Uterine carcinosarcoma (UCS)	598	395	203
	Uterine corpus endometrial carcinoma (UCEC)			

uncharacterized sample into one of the cancer types available in the CancerSysDB. This workflow can be applied, for instance, to predict the primary site of a tumor from a metastatic tissue specimen of unknown origin. The workflow is basically composed of two steps:

- In the *training phase*, a random forest consisting of 1000 trees is trained on all data available in the CancerSysDB. The workflow is composed of an HQL query with subsequent submission of the query results to a high-performance compute cluster. In order to control for the relatively strong imbalance in the class sizes, the workflow was implemented using a stratified sampling approach in the random forest training procedure. The random forest is then trained in 100 parallel processes with 10 trees in each process. Subsequently, the forest is loaded back into the CancerSysDB. The entire procedure must be repeated any time new data is being uploaded into the CancerSysDB. Random forests were chosen because of their good adaption to (binary) mutation data and their convenience in parallelization.
- In the *prediction phase*, a list of mutations of a yet unclassified sample can be uploaded into the CancerSysDB and is classified according to the random forest obtained in the training phase. As usual, the classification is determined by a majority vote between the 1000 classification trees in the forest.

In the current workflow on the public instance, the training phase was carried out on data from 9091 patients in the CancerSysDB. To demonstrate that the predictions produced in this workflow are of sufficient accuracy to make them practically applicable, we split the 9091 patients in a training set of 6006 patients (66.6% in each cohort) and evaluated the predictions in a test set comprising 3085 patients (33.3% in each cohort; Table 3). Out of these 3085 patients in the test set, 1521 (49.3%) were assigned to the correct class (Fig. 2), whereas a random guess of the class would have produced a correct class assignment in only 182 cases (5.9%). Further evaluations of the workflow performance show that the success rate of the predictions does not increase with the number of trees nor the number of variables evaluated at each split, but strongly depends on the number of training samples (Additional file 2: Figure S1). In particular, Additional file 2: Figure S1c suggests that the accuracy could potentially be improved given a

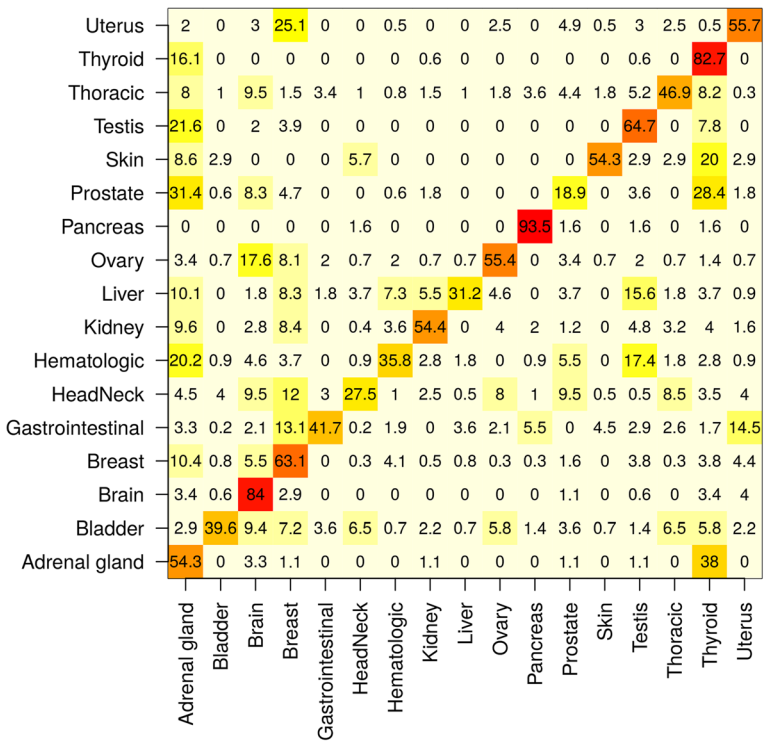


Fig. 2 Results of a cross validation of the random forest prediction of cancer types in the CancerSysDB. The predictions are based on a random forest learned on the training set comprising 6006 patients from 30 TCGA studies (Table 2). Displayed are the predictions of the classes in the 3085 patients in the training set. The accuracy strongly varies across the particular subclasses, but sums up to a total of 1521 correctly classified patients (49.3%)

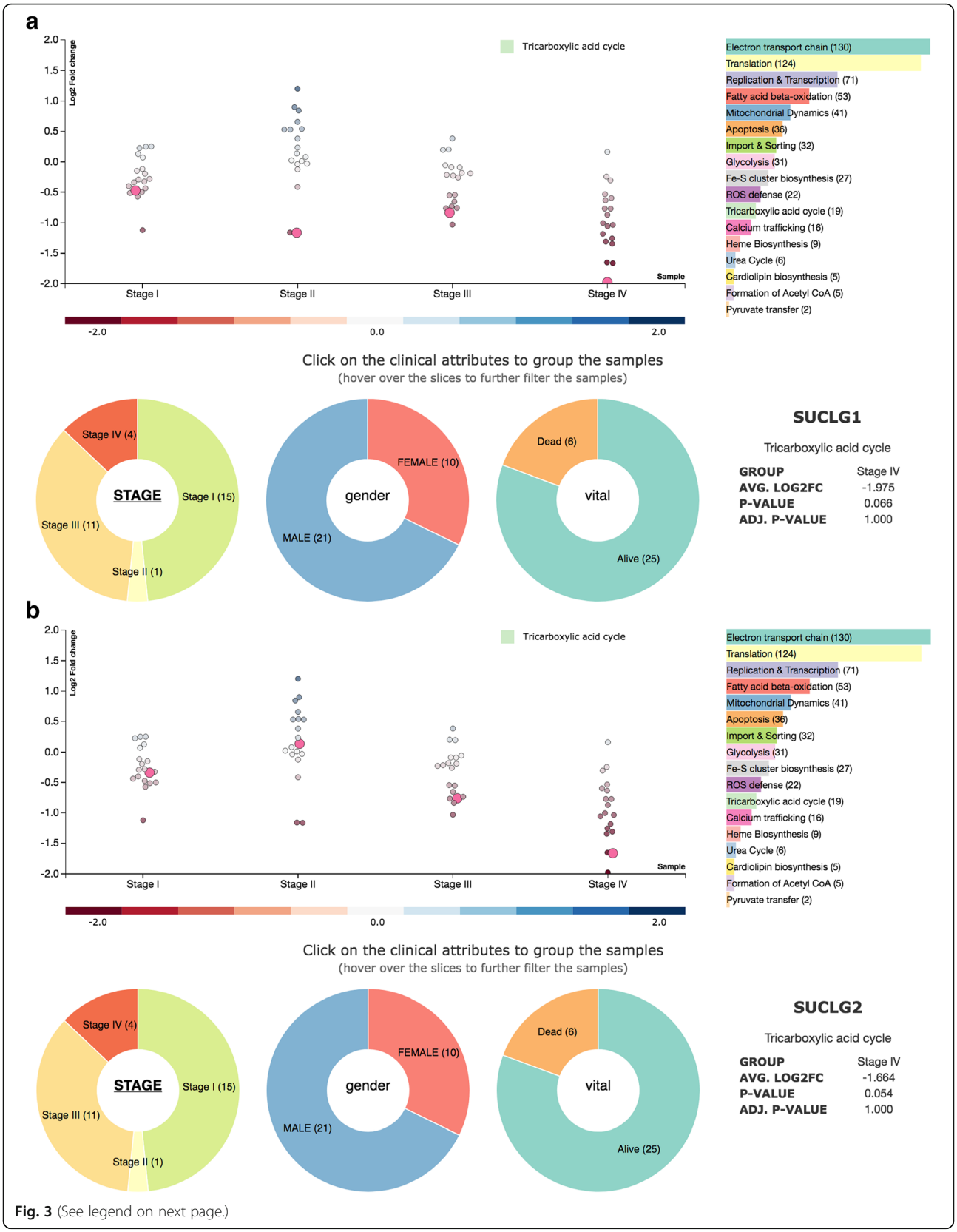


Fig. 3 (See legend on next page.)

(See figure on previous page.)
Fig. 3 In-depth analysis of the dynamics of the TCA pathway in KIRP cancer patients. Interactive view bee-swarm scatter plot on the Tricarboxylic acid cycle (TCA) pathway from KIRP cancer patients is shown. The log2-fold changes are averaged for patients according to tumor grade (Stage I-IV). The dashboard gives the number of patients per grade and allows for further filtering according to gender or vital status (see also Additional file 2: Figure S1). **a** The SUCLG1 gene is selected (pink bubble in bee-swarm scatter plot). **b** The SUCLG2 gene is selected. Both genes show a strong, averaged down-regulation in Stage IV KIRP cancer patients (see Table 4 for averaged log2-fold changes)

constantly growing amount of data in the CancerSysDB. However, we assume that the accuracy could be most stronly improved when including additional data types such as gene expression to the predictive algorithms.

Analyzing TCA-cycle genes in kidney renal papillary cell carcinoma (KIRP)

We have implemented one interactive workflow, which allows users to perform an in-depth analysis of specific groups of genes or pathways. For the public instance of the CancerSysDB, we have chosen a set of mitochondrial functions. The interactive workflow consists of a bee swarm scatter plot displaying the differential expression (log2-fold change) of all genes in a selected pathway, as well as an interactive dashboard, where users can select the desired features for data display on the bee swarm scatter plot (see Additional file 3: Figure S2). Pathways to be shown can be selected on the right-hand side of the scatter plot. Features that can be chosen include the stage of the tumor, gender of the patients, as well as vital status. Differential expression is averaged over all individuals associated with a specific feature. If one feature is selected (e.g. stage of tumor) and the user hovers over any other fields of the dashboards, the data presented in the scatter plot are filtered accordingly. Hovering over one of the stages will give information on gender and vital status of all subjects within this stage (see for instance Additional file 3: Figure S2b, where hovering over Stage IV returns the information on gender (4 males) and vital status (3 alive, 1 dead) of all subjects of this tumor stage). Hovering over one of the other dashboards will change the data for averaging accordingly. For instance, when hovering over FEMALE, data are averaged over 10 patients in two stages (Stage I and Stage III), with 2 individuals with the vital status Dead and 8 ones with vital status Alive.

We have used this workflow to observe the dynamics of the TCA pathway in KIRP (kidney renal papillary cell carcinoma) patients during tumor progression. We observed a strong down-regulation of the Succinate-CoA ligase subunits SUCLG1 and SUCLG2 in Stage IV KIRP patients (Fig. 3 and Table 4), which is independent of the vital status of the patients. We have not observed this specific down-regulation of both Succinate-CoA ligase subunits for any stage-specific cohort of any other tumor type imported from TCGA. An equally strong down-regulation of both subunits could only be observed for two sarcoma patients where no staging is done (SARC cohort in TCGA, data not shown).

Succinate-CoA ligase (SUCL) catalyses the conversion of succinyl-CoA and ADP or GDP to succinate and ATP or GTP. Substrate specificity is determined by the beta-subunit of the complex, which is either SUCLA2 (ATP) or SUGLG2 (GTP), while the alpha-subunit (SUCGL1) does not differ for either substrate [4]. SUCLG2 is predominately expressed in anabolic tissues such as liver or kidney [4, 5]; for these tissues, GTP is more important, as it is involved in processes such as gluconeogenesis or protein synthesis. Mutations of SUCLG1 lead to loss of SUCLG1 protein expression and subsequently to depletion of mtDNA; clinically, affected individuals suffer from severe acidosis and lactic aciduria [9]. Expression changes of SUCLG1 and 2 mRNA [2, 13], as well as protein [11, 17] were also identified in several studies as potential markers for kidney cancers. More notably, down-regulation of SUCLG2 protein levels are furthermore indicative for late stages in clear cell renal carcinomas [10].

Table 4 Averaged log2-fold changes of SUCLG1 and SUCLG2 mRNAs in different tumor stages of KIRP cancer patients

Stage	# Patients	Female/ Male	Alive/ Dead	SUCLG1		SUCLG2	
				log2 FC	*p*-value	log2 FC	*p*-value
I	15	5 / 10	13 / 2	−0.473	0.132	−0.338	0.307
II	1	0 / 1	1 / 0	−1.163	0.082	0.137	0.431
III	11	5 / 6	8 / 3	−0.835	0.018	−0.760	0.028
IV	4	0 / 4	3 / 1	−1.975	0.066	−1.664	0.054

Conclusions

The CancerSysDB enables highly flexible analyses of cancer data across multiple OMICS data types and clinical data. We have demonstrated that the system can be used for cross-data type queries with clinically relevant information on prognosis, genome stability and mutation types of patients with and without mutations in the tumor suppressor TP53. In addition, we have given an example how machine learning technology on only one single data type (somatic mutations) can be used to achieve confident predictions of clinically relevant traits. Finally, we have provided an example how our system

can be used as a platform for interactive analysis of different OMICS data types. The information provided by the TCGA data currently used in the public instance of the CancerSysDB is still very limited compare to the amount of data that can be expected in the near future when genomic analyses in a clinical context are becoming more and more a routine analysis. The CancerSysDB offers an appropriate framework to employ machine learning algorithms on much larger data volumes to predict, for instance, the overall survival of a patient and the response to a particular therapy given a patient's molecular background.

Additional files

Additional file 1: The source code of the database queries and workflow scripts for the three use cases reported in the paper. The results can be reproduced using the query results and analysis scripts provided. File query1.csv contains the barcodes of all samples for which mutation data do exist. File query2.csv contains the barcodes of all samples which carry a mutation in the gene of interest. Finally, query3.csv contains the survival data (according to Fig. 1a), a list of all mutations of patients in the cohort of interest (according to Fig. 1b), or a list of all genomic segments with aberrant copy number in the cohort of interest (according to Fig. 1c). There are small discrepancies between the number of patients with mutation data and the number of patients with survival data (Fig. 1a) and copy number data (Fig. 1c). (ZIP 4981 kb)

Additional file 2: Figure S1 Overall success rate of the prediction of tumor types by random forests depending on (a) the number of samples per stratum in the random forest, (b) the number of variables picked randomly for each tree in the forest and (c) the number of trees learned in the forest. Importantly, the accuracy is increasing monotonically with the number of samples, indicating that the overall strategy is suitable, in particular, for a database with continuously growing amounts of data. In contrast, the success rate does not so much depend on the parameters chosen for the training phase of the random forest. (PNG 34 kb)

Additional file 3: Figure S2 Interactive workflow of mitochondrial pathways. Shown is the Tricarboxylic acid cycle (TCA) pathway for KIRP cancer patients. The central view of this workflow is a bee-swarm scatterplot, which contains the averaged log2-fold changes of patient groups according to either tumor stage, gender or vital status. Each dot is represents the averaged log2-fold change of one gene that has been assigned to the chosen function. Functions can be selected on the right-hand side of the scatter plot. The dashboard below the scatter plot can be used to change the averaging according to a different feature ((a), which shows averaging according to stage), to display information on the composition of the selected feature ((b), which informs the user that all individuals of stage II, which was hovered over in this case, are male and that one individual is dead, while three of the patients are alive); or to further select individual patients and thus modify the averaging shown in the scatter plot ((c), where only female patients were chosen for stage-dependent averaging; as female patient data are only available for two stages (I and III), the scatter plot is changed accordingly). (PNG 679 kb)

Abbreviations

API: Application Programming Interface; CBS: Circular Binary Segmentation; CSV: Character-separated variables; GDAC: Genome Data Analysis Center; GDC: Genomic Data Commons; HPC: High-performance computing; HQL: Hibernate Query Language; JSON: JavaScript Object Notation; KIRP: Kidney renal papillary cell carcinoma; MAF: Mutation annotation format; mRNA: Messenger ribonucleic acid; NGS: Next-Generation Sequencing; SARC: Sarcoma; TCGA: The Cancer Genome Atlas

Acknowledgements

The authors thank Prasanna Koti for assistance in manual curation of mitochondrial pathways.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) with grants PF 3313/2–1 to PF, HA 6905/2–1 to BH and LA 919/6–1 to UL and by the German Ministry for Economics and Energy with grant KF2429610MS2 to PF. BH acknowledges support by the Max Planck Society and the Centre National de la Recherche Scientifique (CNRS). The funding bodies did not play any role neither in the design of the study nor in the collection, analysis, and interpretation of data or the writing of the manuscript.

Availability of data and materials

The datasets analysed in the current study are available on the in the Genomic Data Commons (GDC) Data Portal at https://gdc.cancer.gov/.
Database URL: https://cancersys.uni-koeln.de.
Source code: https://github.com/RRZK/CancerSysDB.

Authors' contributions

RK implemented the database application. PK managed and processed the data available in the database. UL operated the IT infrastructure. AY, BH and PF conceived the analysis workflows. PF and BH wrote the paper. PF, BH and UL and designed the overall concept of the project. All authors read and approved the final version of the manuscript.

Ethics approval and consent to participate

The re-analysis of TCGA samples is a retrospective case report that does not require ethics committee approval at our institution.

Consent for publication

All data used in this study was obtained from The Cancer Genome Atlas research network which originally required written informed consent from all participants.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Regional Computing Center of the University of Cologne (RRZK), Cologne, Germany. [2]Bioinformatics Facility, CECAD Research Center, University of Cologne, Cologne, Germany. [3]Institut de Biologie du Développement, Aix-Marseille University, Marseille, France. [4]Max Planck Institute for Biochemistry, Martinsried, Germany.

Received: 12 December 2017 Accepted: 16 April 2018
Published online: 24 April 2018

References

1. Crispatzu G, Kulkarni P, Toliat MR, Nürnberg P, Herling M, Herling CD, Frommolt P. Semi-automated cancer genome analysis using high-performance computing. Hum Mutat. 2017;38(10):1325–35.
2. Hakimi A, Reznik E, Lee C, Creighton C, Brannon A, Luna A, Aksoy B, Liu E, Shen R, Lee W, Chen Y, Stirdivant S, Russo P, Chen Y, Tickoo S, Reuter V, Cheng EH, Sander C, Hsieh J. An integrated metabolic atlas of clear cell renal cell carcinoma. Cancer Cell. 2016;29(1):104–16.
3. Hollstein M, Sidransky D, Vogelstein B, Harris CC. p53 mutations in human cancers. Science. 1991;253(5015):49–53.
4. Johnson J, Mehus J, Tews K, Milavetz B, Lambeth D. Genetic evidence for the expression of ATP- and GTP-specific succinyl-CoA synthetases in multicellular eucaryotes. J Biol Chem. 1998;273(42):27580–6.
5. Lambeth D, Tews K, Adkins S, Frohlich D, Milavetz B. Expression of two succinyl-CoA synthetases with different nucleotide specificities in mammalian tissues. J Biol Chem. 2004;279(35):36621–4.
6. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(1):550.

7. Negrini S, Gorgoulis VG, Halazonetis TD. Genomic instability - an evolving hallmark of cancer. Nat Rev Mol Cell Biol. 2010;11(3):220–8.
8. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5(4):557–72.
9. Ostergaard E, Christensen E, Kristensen E, Mogensen B, Duno M, Shoubridge E, Wibrand F. Deficiency of the alpha subunit of succinate-coenzyme a ligase causes fatal infantile lactic acidosis with mitochondrial DNA depletion. Am J Hum Genet. 2007;81(2):383–7.
10. Perroud B, Ishimaru T, Borowsky A, Weiss R. Grade-dependent proteomics characterization of kidney cancer. Mol Cell Proteomics. 2009;8(5):971–85.
11. Perroud B, Lee J, Valkova N, Dhirapong A, Lin P, Fiehn O, Kültz D, Weiss R. Pathway analysis of kidney cancer using proteomics and metabolic profiling. Mol Cancer. 2006;5:64.
12. Pfeifer GP, Denissenko MF, Olivier M, Tretyakova N, Hecht SS, Hainaut P. Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. Oncogene. 2002;21(48):7435–51.
13. Sanders E, Diehl S. Analysis and interpretation of transcriptomic data obtained from extended Warburg effect genes in patients with clear cell renal cell carcinoma. Oncoscience. 2015;2(2):151–86.
14. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546–58.
15. Wagle P, Nikolić M, Frommolt P. QuickNGS elevates next-generation sequencing to a new level of automation. BMC Genomics. 2015;16(1):487.
16. Wang X, Sun Q. TP53 mutations, expression and interaction networks in human cancers. Oncotarget. 2016;8(1):624–43.
17. White N, Masui O, Desouza L, Krakovska O, Metias S, Romaschin A, Honey R, Stewart R, Pace K, Lee J, Jewett M, Bjarnason G, Siu K, Yousef G. Quantitative proteomic analysis reveals potential diagnostic markers and pathways involved in pathogenesis of renal cell carcinoma. Oncotarget. 2014;5(2):506–18.

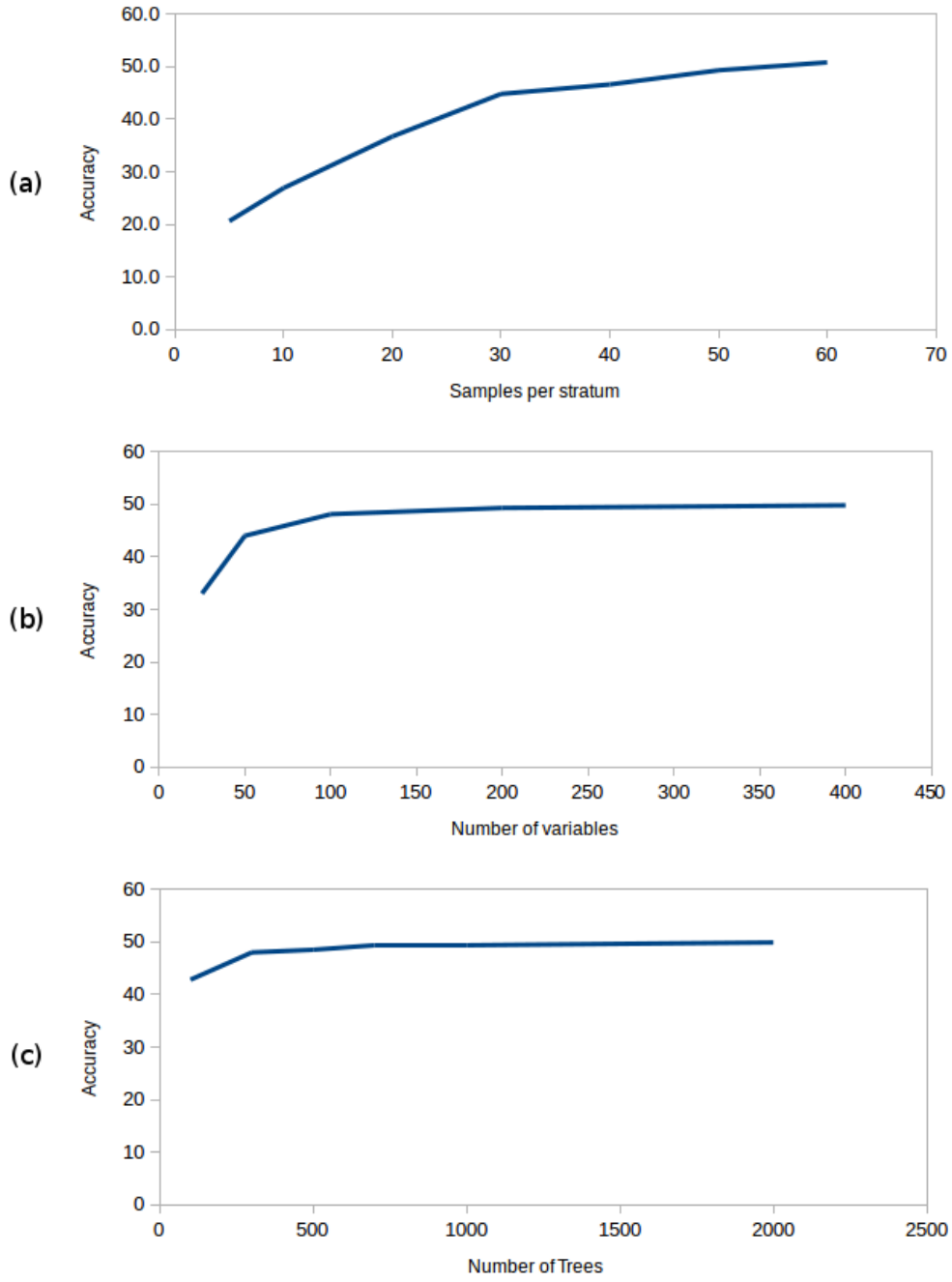


Figure S1 Overall success rate of the prediction of tumor types by random forests depending on (a) the number of samples per stratum in the random forest, (b) the number of variables picked randomly for each tree in the forest and (c) the number of trees learned in the forest. Importantly, the accuracy is increasing monotonically with the number of samples, indicating that the overall strategy is suitable, in particular, for a database with continuously growing amounts of data. In contrast, the success rate does not so much depend on the parameters chosen for the training phase of the random forest.

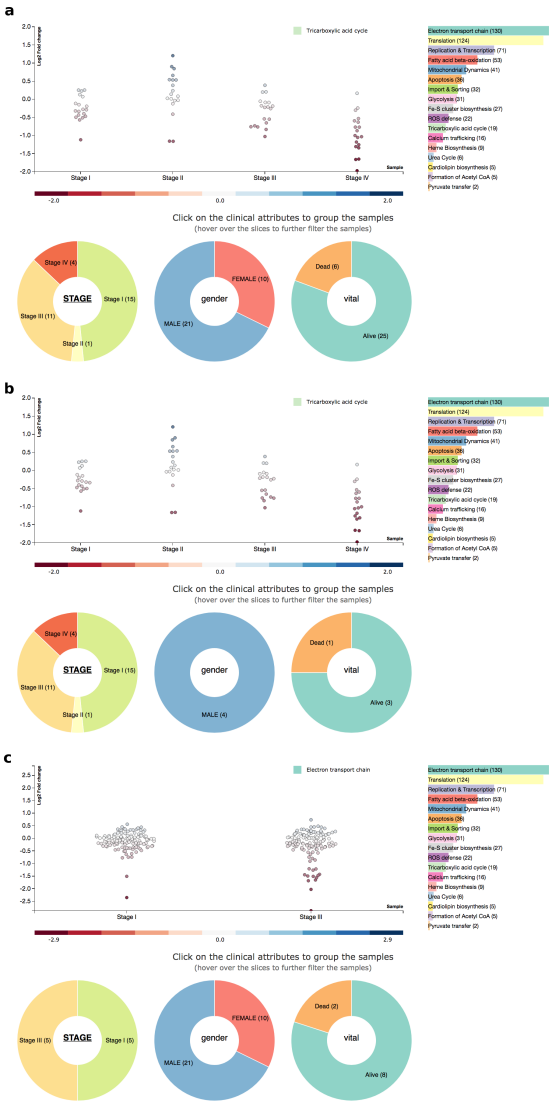


Figure S2 Interactive workflow of mitochondrial pathways. Shown is the Tricarboxylic acid cycle (TCA) pathway for KIRP cancer patients. The central view of this workflow is a bee-swarm scatterplot, which contains the averaged log2-fold changes of patient groups according to either tumor stage, gender or vital status. Each dot is represents the averaged log2-fold change of one gene that has been assigned to the chosen function. Functions can be selected on the right-hand side of the scatter plot. The dashboard below the scatter plot can be used to change the averaging according to a different feature ((a), which shows averaging according to stage), to display information on the composition of the selected feature ((b), which informs the user that all individuals of stage II, which was hovered over in this case, are male and that one individual is dead, while three of the patients are alive); or to further select individual patients and thus modify the averaging shown in the scatter plot ((c), where only female patients were chosen for stage-dependent averaging; as female patient data are only available for two stages (I and III), the scatter plot is changed accordingly).

6.3 Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer

The altered metabolism in cancer is closely related to aberrant gene expression, which depends on both genetic and epigenetic information (Esteller, 2011). DNA methylation is one of the essential epigenetic mechanisms that determines the accessibility of genetic loci to transcriptional machinery and hence levels of gene expression. Therefore, any alterations could lead to significant changes in gene expression landscape, like the down-regulation of tumor suppressor genes observed in cancer cells as a result of hypermethylation in CpG islands (CGIs) at the promoter regions (P. A. Jones & Baylin, 2007). Changes in DNA methylation could be caused by mutations in epigenetic modifiers (Baylin & Jones, 2011) and have been observed in patients of certain cancer types (Spencer et al., 2017); Although it could also occur in the absence of specific mutations, and can be controlled or modulated by signaling pathways or molecules as shown in previous researches (Forloni et al., 2016; Spangle et al., 2016). Genes identified in these studies with functional relevance in cancer could potentially be used as biomarkers for cancer prognosis.

The current study combines the analysis of methylome and transcriptome in a clinically relevant hepatocellular carcinoma (HCC) mouse model, in order to study the epigenetic mechanism that influences transcription and gene expression through DNA methylation. A group of genes were found to have focal hypermethylation in CGIs, accompanied with lower expression (H+E-), which is the conventional mechanism in cancer to downregulate tumor suppressor through promoter hypermethylation. This group includes the mitochondrial genes Ornithine aminotransferase (Oat) and Enoyl-Coenzyme A delta isomerase 1 (Eci1), that are involved in Amino Acid Metabolism and Fatty Acid Degradation & Beta-oxidation respectively.

Surprisingly, another set of genes were found to be hypermethylated in CGIs and overexpressed (H+E+) at the same time in the mouse model. This includes the mitochondrial gene Jun, that belongs to the process Transcription and is considered as an oncogene. Further analysis of the TCGA dataset showed that enrichment of this gene set characterizes 56% of the HCC patients, who belong to an aggressive HCC subclass. Collectively, the data suggested that this set of oncogenes could be of clinical relevance and used as biomarkers for patient stratifications; and that their up-regulations characterized by hypermethylation of CGIs could be a potent operating mechanism in cancer.

nature COMMUNICATIONS

Corrected: Publisher correction

ARTICLE

DOI: 10.1038/s41467-018-05550-5 OPEN

Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer

Maria Arechederra[1], Fabrice Daian[1], Annie Yim[2], Sehrish K. Bazai[1], Sylvie Richelme[1], Rosanna Dono[1], Andrew J. Saurin[1], Bianca H. Habermann[1] & Flavio Maina[1]

Epigenetic modifications such as aberrant DNA methylation reshape the gene expression repertoire in cancer. Here, we used a clinically relevant hepatocellular carcinoma (HCC) mouse model (*Alb-R26*^{Met}) to explore the impact of DNA methylation on transcriptional switches associated with tumorigenesis. We identified a striking enrichment in genes simultaneously hypermethylated in CpG islands (CGIs) and overexpressed. These hypermethylated CGIs are located either in the 5′-UTR or in the gene body region. Remarkably, such CGI hypermethylation accompanied by gene upregulation also occurs in 56% of HCC patients, which belong to the "HCC proliferative-progenitor" subclass. Most of the genes upregulated and with hypermethylated CGIs in the *Alb-R26*^{Met} HCC model undergo the same change in a large proportion of HCC patients. Among reprogrammed genes, several are well-known oncogenes. For others not previously linked to cancer, we demonstrate here their action together as an "oncogene module". Thus, hypermethylation of gene body CGIs is predictive of elevated oncogene levels in cancer, offering a novel stratification strategy and perspectives to normalise cancer gene dosages.

[1] Aix Marseille Univ, CNRS, Developmental Biology Institute of Marseille (IBDM), Parc Scientifique de Luminy, Aix Marseille Univ, 13009 Marseille, France. [2] Computational Biology Group, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany. Correspondence and requests for materials should be addressed to F.M. (email: flavio.maina@univ-amu.fr)

Appropriate timing and dosage of gene expression in healthy cells is ensured by complex processes integrating genetic and epigenetic information. Alterations of these mechanisms are frequent in cancer and underline functional changes in genes acting as oncogenes or tumour suppressors[1–3]. The use of high-throughput sequencing has contributed considerably to our understanding on how epigenetic modifications switch genomic regions from an inaccessible closed conformation to an open state–and vice-versa–contributing to changes in the transcriptome landscape[4–6]. DNA methylation is an essential epigenetic mechanism influencing gene expression levels in cells and alterations lead to dramatic changes in malignant cells. The cancer landscape is generally characterised by a diffuse DNA hypomethylation and by focal hypermethylation in CpG-rich regions known as CpG islands (CGIs)[1,7]. CGI hypermethylation at promoters represses transcription of genes acting as tumour suppressors, a well-known mechanism operating in cancer[8]. However, DNA methylation at intergenic regions and gene bodies is gaining relevance for its impact on gene expression[9,10]. Aberrant DNA methylation of large clusters of transcriptional enhancers, known as super-enhancers, leads to dramatic transcriptional changes of gene sets in cancer[11]. A large fraction of DNA methylation is also observed in gene body CGIs, with an apparent intriguing positive correlation between methylation and gene expression[12,13]. Such contradiction on DNA methylation effects in promoter versus gene body CGIs remains poorly understood.

The relevance of epigenetics in tumorigenesis has been further emphasised through recent large-scale screen analyses focused on cancer patients carrying either histone mutations or alterations in genes regulating DNA methylation–histone modifications[2]. Results from these studies highlighted how such mutations dramatically modify the epigenetic and gene expression landscapes. For example, aberrant DNA methylation has been recently reported in acute myeloid leukaemia patients with DNMT3A mutations[14]. Abnormal recruitment of PRC2 complex and DNA methylation occurs in paediatric glioblastoma with Histone H3 mutant variants[15]. Gene expression changes caused by histone H3K36 mutation is associated with sarcomagenesis[16]. Nevertheless, the epigenetic reshape occurs also in the absence of specific mutations in chromatin modulators[17]. It is the case of classical oncogenes and tumour suppressors, which can trigger profound chromatin alterations with consequences on gene expression[18,19]. For example, an oncogenic splice variant of EGFR leads to genome-wide activation of putative enhancers in glioblastoma[20]. Oncogenic EGFR leads to DNA methylation-mediated transcriptional silencing of tumour suppressors in lung cancer and glioblastoma[21]. Deregulated Ras signalling reshapes the enhancer landscape leading to aberrant oncogene expression[22]. PI3K/Akt pathway activation induces promoter-associated gene activation in breast cancer[23]. Overall, such screen approaches have also contributed to identify new genes, whose functional relevance in cancer was previously unknown and/or which deregulations can be used as cancer biomarkers for prognosis/patient stratification.

We recently reported a cancer mouse model in which slight increases in wild-type Met receptor tyrosine kinase (RTK) levels in the liver are sufficient for spontaneous tumours in mice (*Alb-R26^{Met}*). These genetic studies conceptually illustrate how the shift from physiological to pathological conditions results from perturbations in subtle signalling dosage. Through gene expression analysis, the *Alb-R26^{Met}* mice were shown to model a HCC patient subgroup corresponding to the so-called "proliferative-progenitor" subclass[24], demonstrating the clinical relevance of this genetic system. The uniqueness of this genetic system was also illustrated by its usefulness to identify new synthetic lethal interactions as potential therapies for HCC subgroups[24]. Here, we employed the *Alb-R26^{Met}* cancer model for integrative genome-wide studies combining methylome and transcriptome outcomes and compared them with those from HCC patients. Results show an enrichment in genes overexpressed and with hypermethylated CGI, with expression levels positively correlating with the CGI distance to the ATG. Whereas most of the upregulated genes are well-known oncogenes, the implication of others in cell tumorigenic properties is demonstrated here through functional studies. Enrichment of genes both overexpressed and with hypermethylated CGIs characterises the "proliferative-progenitor" HCC patient subset, which is modelled by the *Alb-R26^{Met}* genetic system. Collectively, these results show that an epigenetic reprogramming process ensuring increased dosage of an "oncogenic module" involving multiple genes operates in tumorigenesis.

Results

***Alb-R26^{Met}* tumours recapitulate DNA methylation changes of HCC patient subgroups**. We recently showed how the *Alb-R26^{Met}* genetic system is a unique tool to model: (a) the tumorigenic program, (b) the "proliferative-progenitor" HCC patient subgroup and (c) functionality of signalling alteration for drug discovery[24]. For its use to study the contribution of epigenetic modifications linked to cancer, we reasoned that it was first necessary to determine whether the *Alb-R26^{Met}* tumorigenesis occurs in a stable genomic context or is associated with chromosomal deletions/duplications. Comparative genomic hybridisation analyses on DNA inputs from 16 *Alb-R26^{Met}* tumours and 8 control livers excluded chromosomal instability (Supplementary Fig. 1). These findings therefore reinforce the appropriateness of the *Alb-R26^{Met}* cancer model as a relevant genetic system to study the epigenetic reprogramming associated with cancer, which we addressed by bioinformatically integrating data from methylome and transcriptome screens (Fig. 1a).

DNA methylation changes were scored by performing MethylMiniSeq EpiQuest sequencing on 10 *Alb-R26^{Met}* tumours (previously histologically identified as HCC[24]) and 3 control livers (Supplementary Fig. 2A). Mean methylation levels were modestly, yet significantly, different across all measured CpGs (*P*-value = 2.4E−03; Fig. 1b), being able to group tumours and controls into two distinct clusters (Fig. 1c). A remarkable predominance of global hypomethylation was observed in tumours compared with livers (Fig. 1b, Supplementary Fig. 2B). Accordingly, we observed an enrichment in hypomethylated CpGs located outside CGIs (*P*-value = 3E−04; Fig. 1b, Supplementary Fig. 2C). In contrast, a significant enrichment of hypermethylation at CpGs located within CGIs characterised *Alb-R26^{Met}* tumours compared with control livers (*P*-value = 3.9E−03; Fig. 1b, d, Supplementary Fig. 2D). These traits of CpG methylation changes, according to the CpG location with respect to CGIs, are consistent with those largely reported in the literature[1]. Focusing on differentially methylated CpGs located at annotated CGIs, we identified 513 CGIs with a β-value methylation difference of ±0.2 and a false discovery rate (FDR) <0.05 (Fig. 1d, Supplementary Fig. 2D, Supplementary Data 1). These CGIs were homogeneously distributed amongst all 19 autosomal and 1 sex chromosome mouse pairs (Supplementary Fig. 2E). Among CGIs with differentially methylated CpGs, 82% were hypermethylated in *Alb-R26^{Met}* HCC compared to controls (Fig. 1d).

To explore the relevance of these methylation changes in the context of human HCC disease, we used genome-wide DNA methylation data from a cohort of 41 HCC patients, for which data are available for both: (a) methylation and expression; (b)

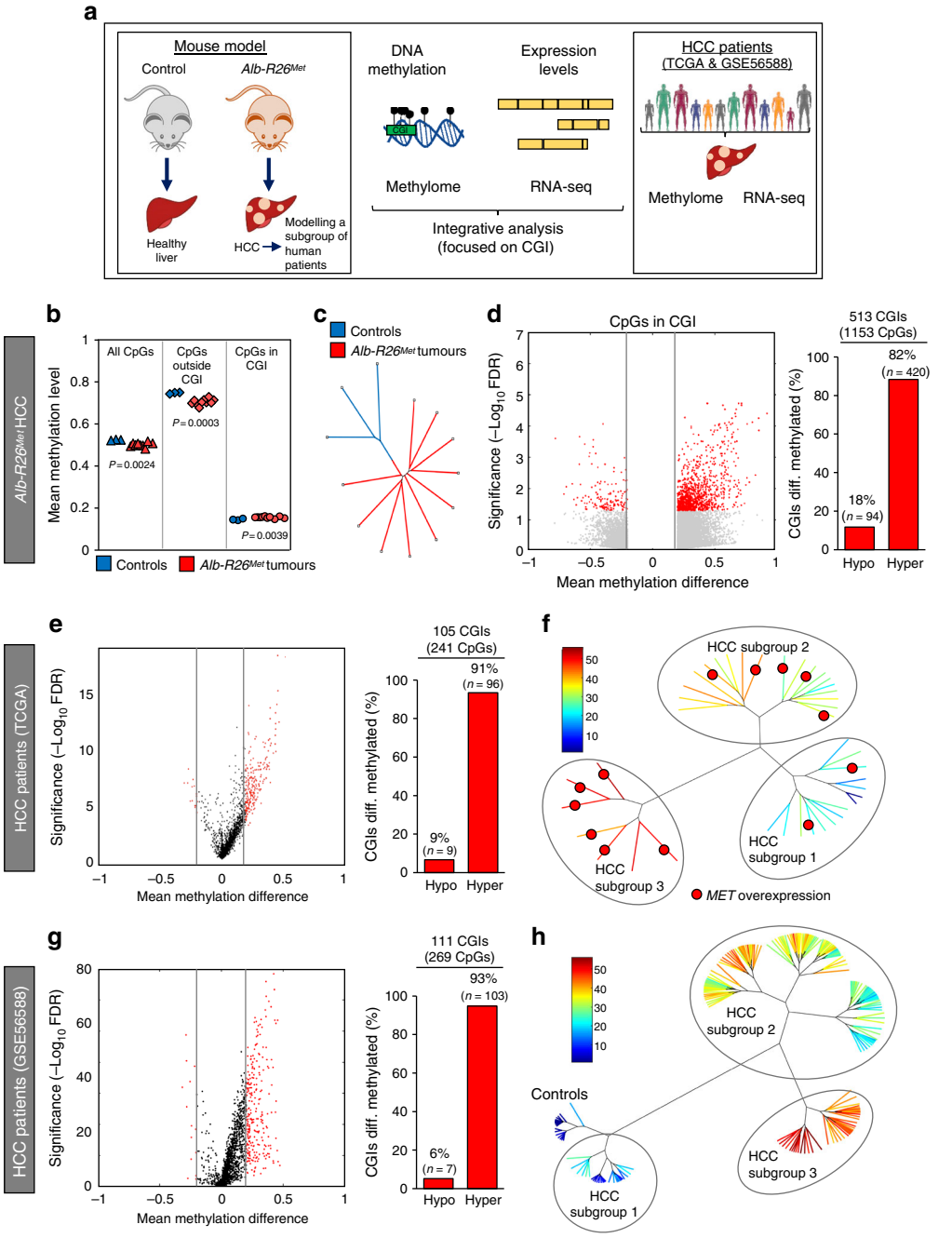


Fig. 1 Methylome studies identify an enrichment of CGI hypermethylation in *Alb-R26*^{Met} tumours, also present in a subset of HCC patients. **a** Schematic representation of the overall strategy employed. DNA methylation and gene expression levels were analysed in *Alb-R26*^{Met} tumours and control livers. Outcomes were compared with HCC human database. **b** Mean methylation levels in controls and *Alb-R26*^{Met} tumours, focusing on all CpGs, CpGs outside CGIs and CpGs in CGIs. **c** Unrooted distance tree using the overall DNA methylation content subdivides *Alb-R26*^{Met} tumours and controls in two distinct clusters. **d** Volcano plot reporting methylation differences with significance (expressed as $-\text{Log}_{10}$ FDR) for CpGs in CGIs in *Alb-R26*^{Met} tumours versus control (left). Significant differences (methylation difference > 0.2 and FDR < 0.05) are shown in red. Graph reporting the percentage (and numbers) of hypomethylated versus hypermethylated CGIs (right). **e** Volcano plot reporting the mean methylation differences with significance (expressed as $-\text{Log}_{10}$ FDR) in HCC patients from TCGA (left) for differentially methylated CGIs identified in *Alb-R26*^{Met} tumours. Similar methylation levels in HCC patients and controls are reported in black, whereas changes (>0.2) are reported in red. On the right, the graph reports the percentage (and numbers) of hypomethylated versus hypermethylated CGIs. **f** Unrooted distance tree of the 41 TCGA HCC patients showing patient segregation in three distinct subgroups, according to the 416 CGIs found differentially methylated in *Alb-R26*^{Met} tumours. Red dots highlight patients in which *MET* is overexpressed ($\log_2$FC > 1). Patients are reported in different colours according to the percentage of overlap (the scale in percentage is shown on the left). Note the striking correlation between differentially methylated CGIs and *MET* overexpression in the HCC patient subgroup 3. **g**, **h** Volcano plot (**g**) and unrooted distance tree (**h**) from studies using a second cohort of 224 HCC patients and 10 controls (GSE56588 dataset)

tumour and adjacent liver as control (from The Cancer Genomic Atlas; TCGA[25]). For comparisons between mouse and human data, we first mapped the 513 identified mouse CGIs (mm9) to the corresponding CGIs in human (hg19), using the UCSC toolbox. 501 out of 513 CGIs were successfully matched between the two genomes. For 416 CGIs, human methylome data were available in TCGA dataset (Supplementary Fig. 3A; Supplementary Data 2). We extracted the methylation β-value for the CpGs within these human CGIs and calculated the mean methylation difference for each CpG and for each of the 41 HCC patients. While the majority of these CpGs showed a methylation difference below 0.2, a proportion of CGIs (24%) were differentially methylated (FDR < 0.05) with a hypermethylation enrichment score similar to the *Alb-R26^{Met}* HCC (91%; Fig. 1e, Supplementary Fig. 3B). As the analysed cohort includes patients with widely diverse aetiologies and characteristics, we next analysed the 416 CGIs in the individual patients. Intriguingly, hierarchical clustering analysis segregated these HCC patients into three distinct subgroups, with one subgroup composed of seven patients reaching 43–56% overlap with the *Alb-R26^{Met}* list (subgroup-3; Fig. 1f, Supplementary Fig. 3C). The relevance of the *Alb-R26^{Met}* methylation changes in the context of human HCC was further assessed in a second distinct cohort of 234 human samples (224 HCC patients and 10 control individuals[26]). 27% of CGIs differentially methylated in *Alb-R26^{Met}* HCC are also altered in human HCCs, again with an enrichment in hypermethylation (93%; Fig. 1g, Supplementary Fig. 4A). Moreover, these methylation changes distinguished controls from HCC patients, which further segregate into three subgroups. HCC subgroup-3 reaches about 50% CGI overlap with the *Alb-R26^{Met}* list (Fig. 1h, Supplementary Fig. 4B).

Next, we asked whether there would be any correlation between MET alterations with the three human HCC subgroups identified by the *Alb-R26^{Met}* methylome screening. Concerning the HCC patient cohort from TCGA, we were able to perform correlative studies as RNA-seq and mutation data are available. In particular, we analysed *MET* mutations and *MET* expression levels for each patient belonging to the 3 different HCC subgroups. All HCC patients carry the wild-type form of *MET*, which is in agreement with rare mutations of *MET* in HCC. Concerning expression levels, *MET* is overexpressed in 86% (6/7) of HCC patients belonging to subgroup-3 (which best overlaps with CGI methylation changes in *Alb-R26^{Met}*), in 32% (6/19) to HCC subgroup-2, and only in 13% (2/15) to HCC subgroup-1 (Fig. 1f; patients with *MET* overexpression are highlighted with a red dot; Supplementary Fig. 3C–F). For the HCC patient cohort from GSE56588, expression data (array) are only available for some patients and without information about mutations. Therefore, correlative studies were not possible with this HCC cohort. Together, these findings show that liver cancer modelled by the *Alb-R26^{Met}* genetic system is characterised by methylation changes of specific CGIs, with a predominant hypermethylation profile. A high proportion of these alterations are also found in HCC patient subgroups. Furthermore, there is a striking correlation between differentially methylated CGIs and *MET* overexpression in the HCC patient subgroup modelled by the *Alb-R26^{Met}* genetic setting.

Enrichment in CGI hypermethylation is necessary for *Alb-R26^{Met}* tumorigenesis. The overall enrichment in CGI hypermethylation in the *Alb-R26^{Met}* genetic system prompted us to determine its relevance for cell tumorigenic properties. We designed different demethylating experimental conditions using low doses of Decitabine (0.3 µM; Fig. 2a), according to previously reported protocols[12]. We used three different *Alb-R26^{Met}* HCC cell lines, established from individual *Alb-R26^{Met}* tumours[24]. Decitabine treatment does not affect cell viability of *Alb-R26^{Met}* HCC cells, as well as of MLP-29 cells, a mouse liver progenitor cell line that is not tumorigenic as illustrated by its inability to form colonies in anchorage-independent growth assays (Fig. 2b). Instead, Decitabine treatment interferes with *Alb-R26^{Met}* cell tumorigenic properties, irrespective of the HCC cell line used, as exemplified by: (a) reduced colony numbers when cells are grown in an anchorage-independent manner (Fig. 2c); (b) reduced number and size of foci when cells are grown in an anchorage-dependent manner (Fig. 2d); (c) reduced tumour spheres when cells are grown in self-renewal conditions (Fig. 2e). The effect of global demethylation on cell tumorigenicity was further analysed in vivo by performing xenografts in nude mice. The tumour volume was significantly reduced in mice either injected with Decitabine pre-treated *Alb-R26^{Met}* HCC cells or when Decitabine pulses were administered to mice injected with untreated *Alb-R26^{Met}* HCC cells (Fig. 2f–left). Decitabine doses used in vivo were not toxic, as revealed by no significant changes on the mouse weight during the treatment (Fig. 2f–right). Together, these results indicate that the overall enrichment in CGI hypermethylation is functionally relevant for tumorigenesis modelled by the *Alb-R26^{Met}* genetic system.

CGI hypermethylation correlates with gene upregulation in *Alb-R26^{Met}* HCC. Alterations in DNA methylation are known to impact gene expression. We analysed the expression levels of the 431 genes with differentially methylated CGIs in *Alb-R26^{Met}* tumours using high-coverage RNA-seq data (4 *Alb-R26^{Met}* tumours and 4 control livers). Studies highlighted 93 genes differentially expressed ($\log_2$FC > 1, FDR < 0.05; Supplementary Data 3). According to the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database, several cancer-related pathways were significantly enriched, such as MAPK signalling, viral carcinogenesis, pathways in cancer, TGF-β signalling, cell cycle, renal cell carcinoma (Fig. 3a, Supplementary Fig. 5), strengthening the significance of genes differentially methylated and expressed in the *Alb-R26^{Met}* cancer model. Remarkably, the top-ranked MAPK pathway is coherent with its essential functionality for *Alb-R26^{Met}* tumorigenicity, as previously reported[24]. Among genes differentially methylated and expressed, 36 genes showed the expected inverse correlation between methylation and expression where 20 genes are hypomethylated and overexpressed, and 16 genes are hypermethylation and downregulated (Fig. 3b, Supplementary Data 3). Unexpectedly, 55 genes (59%) were found hypermethylated and overexpressed (Fig. 3b). Thus, tumorigenesis modelled by the *Alb-R26^{Met}* mice is characterised by a set of genes with changes in CGI methylation accompanied by a reprogramming of transcript levels.

The intriguing enrichment in hypermethylated and overexpressed genes drove us to analyse the position of the hypermethylated CGIs with respect to the ATG. Interestingly, the CGI of overexpressed genes is either close to the ATG or in the gene body region, in contrast to the CGI position of downregulated genes exclusively located around the ATG (Supplementary Fig. 6). Concerning the 55 genes hypermethylated and overexpressed, they can be subdivided into two groups. Group-I includes 31 genes, for which the CGIs are located between −50% and 30% relative to the ATG (predominantly into the 5′-UTR). Group-II includes 24 genes, whose CGIs are located much further from the ATG (from 30% of the gene body relative to the ATG to the transcription termination site), corresponding to gene body regions (Fig. 3c, Supplementary Fig. S6). Next, we analysed whether the CGI location influences gene expression. Intriguingly, the extent of overexpression is significantly higher

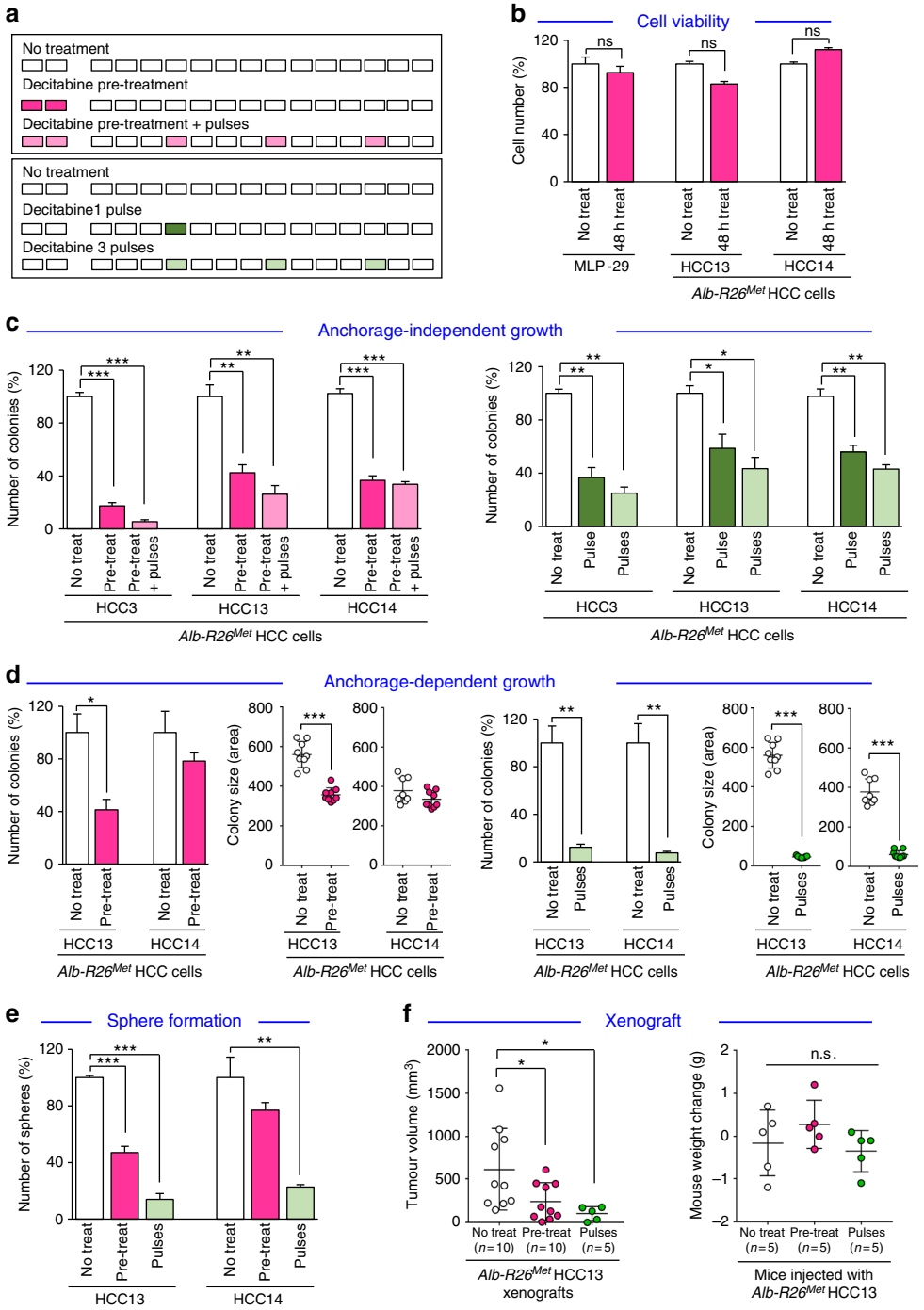


Fig. 2 Global CGI hypermethylation is functionally relevant for *Alb-R26*^{Met} tumorigenesis. **a** Scheme reporting demethylating treatments (Decitabine; 0.3 μM) used for in vitro and in vivo experiments with *Alb-R26*^{Met} HCC cells. Cells were pre-treated (48 h) before using them for experiments with or without subsequent Decitabine pulses every 3 days (top; pink). Untreated cells were exposed to pulses of demethylating treatments during the assay (bottom; green). **b** Effect of Decitabine treatment (48 h) on cell viability of mouse liver progenitor MLP-29 cells and of *Alb-R26*^{Met} HCC cells (HCC13 and HCC14). **c** Anchorage-independent growth (soft agar) assay using 3 different *Alb-R26*^{Met} HCC cell lines (HCC3, 13 and 14) showing effects of demethylating treatments described in **a**. **d** Effect of Decitabine treatment in anchorage-dependent growth (foci formation) assay of 2 different *Alb-R26*^{Met} HCC cell lines (HCC13 and 14). Graphs report number and size of colonies. **e** Decitabine (pre-treatment and treatment) reduces numbers of tumour spheres derived from *Alb-R26*^{Met} HCC cells (HCC13 and HCC14). **f** Left: Graph reporting the tumour volume of mice injected either with untreated cells (No treat), with Decitabine pre-treated *Alb-R26*^{Met} HCC cells (Pre-treat), or with untreated cells following pulses of Decitabine in vivo treatments (Pulses). Note that tumour volume was significantly reduced in mice injected either with Decitabine pre-treated *Alb-R26*^{Met} HCC cells or with untreated cells following in vivo Decitabine pulses. Right: Mouse weight of the indicated groups was measured before and after xenograft experiments. No significant changes were observed, indicating that the dose of Decitabine used in vivo was not toxic. Significant differences between groups are indicated on the top. Not significant (ns): $P > 0.05$, $^{*}P < 0.05$, $^{**}P < 0.01$; $^{***}P < 0.001$

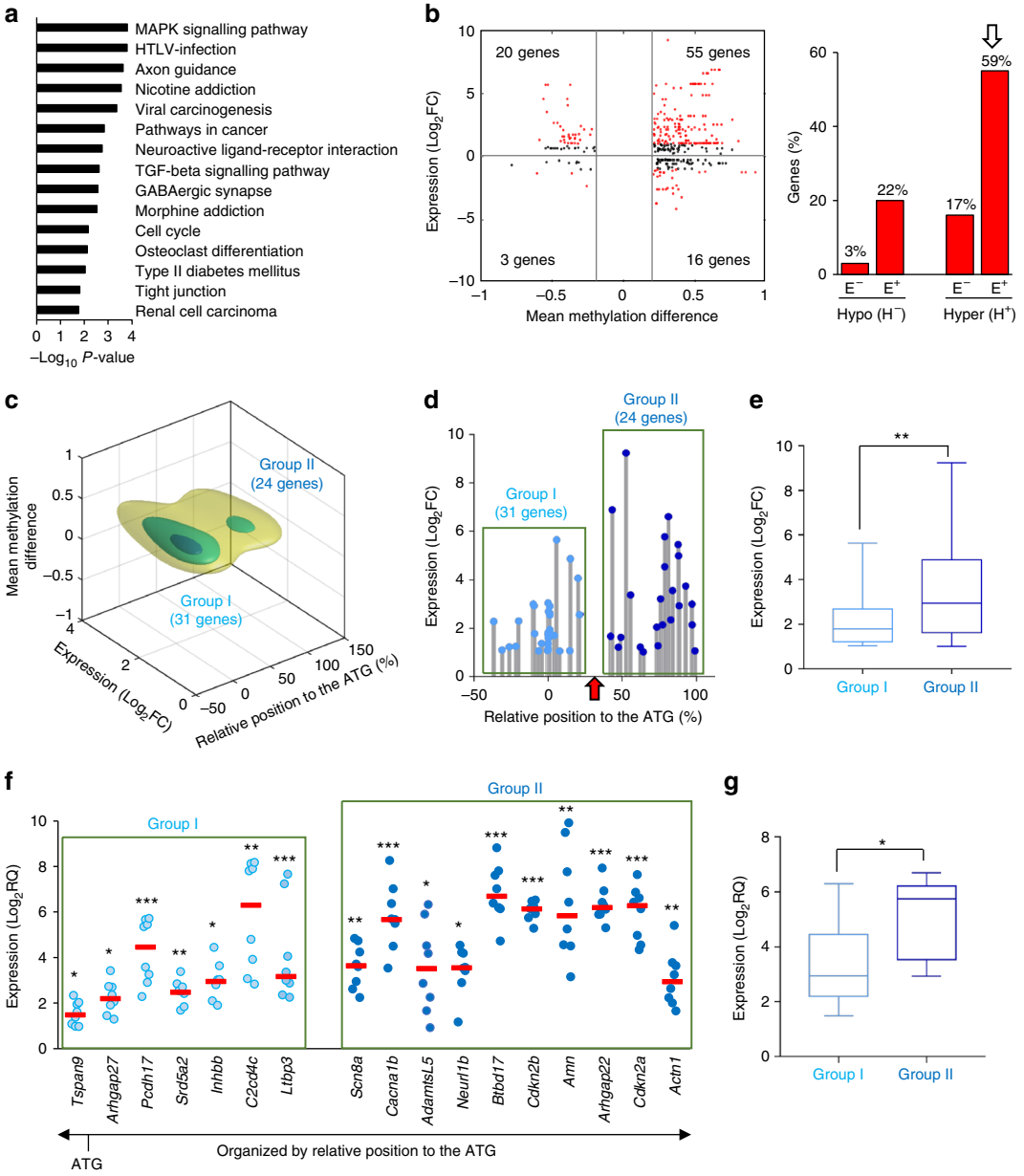


Fig. 3 *Alb-R26*^{Met} tumours are characterised by an enrichment in genes overexpressed and with hypermethylated CGIs. **a** Histogram reporting the KEGG pathway enrichment analysis for genes with changes in CGI methylation and expression in *Alb-R26*^{Met} tumours, ordered according to the $-\text{Log}_{10}$ *P*-value. **b** Left: Methylation differences versus expression for all genes with CGIs hypermethylated (H^+) or hypomethylated (H^-) in *Alb-R26*^{Met} tumours. Expression values are relative to controls. Dots correspond to single differentially methylated CpG and the corresponding gene expression (genes which expression is significantly below or above $\text{Log}_2\text{FC} \pm 1$ are indicated in red). Right: Graph reporting the percentage of downregulated (E^-) and upregulated (E^+) genes among those with a hypomethylated (H^-) or hypermethylated (H^+) CGI. Note the enrichment of genes overexpressed and with hypermethylated CGIs (indicated by an arrow), on which subsequent studies were focused. **c** For the 55 genes overexpressed and with hypermethylated CGI in *Alb-R26*^{Met} tumours, 3D density plot shows their distribution according to relative position to the ATG (as percentage), gene expression level (as Log_2FC) and CGI methylation (as β-value difference). Note that genes segregate into two groups, according to their relative position to the ATG. **d** Graph reporting the individual expression level (as Log_2FC; from RNA-seq data) of hypermethylated and overexpressed genes found in *Alb-R26*^{Met} tumours compare to controls. Note that the relative position of the hypermethylated CGI to the ATG well segregate the two groups (indicated with a red arrow). **e** Box plot illustrating the global expression levels of genes in Group-I and Group-II. **f** Graph reporting individual expression levels (as Log_2FC; by RT-qPCR) of genes belonging to the two groups in *Alb-R26*^{Met} tumours ($n = 8$) relative to control livers ($n = 6$). Red lines report the median Log_2FC in expression. Note that genes are distributed according to the location of the hypermethylated CGI relative to the ATG. **g** Box plot showing the global Log_2FC in expression (according to data in F) of genes in Group-I and Group-II. In **e** and **g**, the median is reported by a line and bars extend to the minimum/maximum values. Significance is indicated on the top. $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$

for Group-II than Group-I (median $log_2FC = 2.15 \pm$ versus median $log_2FC = 3.38 \pm 0.4$; Fig. 3d, e). Importantly, the promoter CGI methylation status of genes belonging to Group-II was similar in *Alb-R26^{Met}* tumours and control livers, thus excluding that changes in promoter methylation could influence gene expression levels (Supplementary Data 4). We corroborated these results through RT-qPCR analysis of a subset of genes belonging to both groups in *Alb-R26^{Met}* tumours ($n = 8$) relative to control livers ($n = 6$). Results showed consistent upregulation of all genes (Fig. 3f), with significant higher expression levels for those within Group-II (Fig. 3g). Together, these results highlight a set of overexpressed genes with hypermethylated CGIs in *Alb-R26^{Met}* tumours and identify a correlation between the location of hypermethylated CGIs and transcription status, where CGIs located further from the ATG showing predominantly increased transcription.

Next, we analysed whether Decitabine treatment would affect in vivo the expression levels of genes found hypermethylated and overexpressed in the *Alb-R26^{Met}* tumours. Focusing on a set of genes, we examined both their expression levels and the methylation levels of their corresponding CGIs in dissected tumours from *Alb-R26^{Met}* mice either untreated or treated with Decitabine. RT-qPCR results showed that Decitabine treatment significantly decreases the expression levels of 7/8 analysed genes (Fig. 4). Bisulfite sequencing studies revealed decreased methylation levels of most CpGs within the gene body CGIs (Fig. 4, Supplementary Data 5). Thus, CGI hypermethylation of genes belonging to Group-I and Group-II ensures their increased expression levels in *Alb-R26^{Met}* tumours as demethylating treatment leads to a reduction of both CGI methylation content and transcription.

A CGI hypermethylation and gene overexpression signature defines a HCC patient subset. Next, we explored the relationship between changes in CGI methylation and gene expression in the above cohort of 41 HCC patients. Because of expected epigenomic heterogeneity between human samples, we reasoned it relevant to perform analyses in individual patients. We integrated transcriptome and methylome data to extract the expression levels of genes with differentially methylated CGIs (Fig. 5a), then classified patients according to the highest percentage of genes: (a) overexpressed with hypermethylated CpGs (H^+E^+); (b) overexpressed with hypomethylated CpGs (H^-E^+); (c) underexpressed with hypermethylated CpGs (H^+E^-); (d) underexpressed with hypomethylated CpGs (H^-E^-). Intriguingly, 23/41 patients (56%) showed an enrichment of genes overexpressed and with hypermethylated CpGs (H^+E^+ patient-subset; Fig. 5a, Supplementary Data 6), similar to the *Alb-R26^{Met}* model (Fig. 3b). Analysis of *MET* levels in HCC patients revealed that the mean *MET* levels in the H^+E^+ subset is $0{,}77 \pm 0{,}16$ (9/23; 39% patients with *MET* levels ≥ 1), whereas in the "NO H^+E^+" subset is $0{,}2 \pm 0{,}24$ (5/18; 27% patients with *MET* levels ≥ 1; Supplementary Data 6). Interestingly, all 7 patients belonging to the HCC subgroup-3 (in Fig. 1f) are characterised by more than 37% of genes both hypermethylated and overexpressed, and 5/7 patients belong to the H^+E^+ subset (these 7 patients are highlighted with a red square and red % in Fig. 5a). Next, we asked whether the H^+E^+ patient subset could be also identified according to global gene expression or methylation features. Unsupervised cluster analysis of either global methylome or transcriptome data did not lead to the same patient clustering (Supplementary Fig. 7), thus strengthening the usefulness of combining methylation-expression features to identify specific HCC patient subsets.

The remarkable correlation between data obtained in the *Alb-R26^{Met}* HCC model and analysed patient samples prompted us to perform integrative studies using another HCC model, for which methylation and expression data are available: the hepatitis-B virus-X mice (*HBx^{tg}*; GSE48052[27]). We first identified all CpGs differentially methylated in *HBx^{tg}* HCC model, then correlated them with gene expression levels. We identified 115 genes both differentially methylated and differentially expressed (a very similar number to the 97 genes found in the *Alb-R26^{Met}* genetic setting). Nevertheless, we found a different distribution compared to that of the *Alb-R26^{Met}* HCC, with an enrichment in genes both hypomethylated and downregulated (Supplementary Fig. 8). Next, we performed correlative analyses with the 41 HCC patients (reported in Fig. 5a): amongst the 18 "NO H^+E^+" subset, 8 patients (20%) share the same enrichment of hypomethylated and downregulated genes modelled by the *HBx^{tg}* mice. Unexpectedly, only 1/8 of these patients is reported positive for HBV. Thus, an epigenetic rewiring of gene sets through hypomethylation and downregulation occurs in a fraction of HCC patients, who do not appear to be characterised by the HBV-associated risk. Collectively, these findings indicate a rather intriguing specificity in how genes are epigenetically reprogrammed in HCC patients: an enrichment in hypermethylated and upregulated genes (for those corresponding to the *Alb-R26^{Met}* model) versus an enrichment in hypomethylated and downregulated genes (for those corresponding to the *HBx^{tg}* model).

For the several genes found overexpressed and with hypermethylated CGIs in the H^+E^+ patient subset, such as *WT1*, *DLK1*, *TP73*, *EEF1A2*, *IGF1R*, *DKK1*, *SPOCK1*, *ITPKA*, *HOXA3*, *NOX4*, *FZD10*, *VASH2*, *GATA2*, *SOX8*, their upregulation in HCC samples is coherent with their reported function as oncogenes in cancer. Concerning the H^+E^+ patient subset, based on clinical data from TCGA, no association was found with a specific risk factor, such as HBV/HCV infection, high-alcohol intake or non-alcoholic fatty liver disease (NAFLD) (Supplementary Fig. 9). Instead, the H^+E^+ patient subset is distinguished by specific HCC molecular markers[28]. In particular, analysis of available RNA-seq data revealed a significant upregulation of *alpha-FETOPROTEIN* (*AFP;* a HCC marker when expressed in adult livers), *JAG1, NOTCH3, NOTCH4, SOX9, VIM* (progenitor markers) and *CD24* (a HCC prognosis marker; Fig. 5b). Importantly, these markers are also upregulated in *Alb-R26^{Met}* HCC (Fig. 5c), as we recently reported[24]. Together, these results show that an enrichment in genes characterised by "CGI hypermethylation and overexpression" occurs in HCC patients belonging to the so-called "proliferative-progenitor" subclass[28]. Moreover, these HCC patients share common features with the *Alb-R26^{Met}* liver cancer model: the epigenetic H^+E^+ signature and the "proliferative-progenitor" cell feature.

Overexpressed genes with hypermethylated CGIs in 5'-UTR or gene body regions act as oncogenes. The intriguing overlap between the *Alb-R26^{Met}* model and the H^+E^+ patient subset prompted us to explore the relevance in cancer of the 55 genes found in *Alb-R26^{Met}* tumours both overexpressed and with hypermethylated CGIs either in the 5′-UTR or in the gene body region. For this analysis, transcriptome data from HCC patients were available for 51/55 genes (Supplementary Data 7). Remarkably, most genes are overexpressed in a large proportion of HCC patients (Fig. 6a, Supplementary Data 8), with a significant higher number in the H^+E^+ patient subset compared with the other (Fig. 6b). These genes include *PRRX1* (28 patients out of 41), *CLDN7* (20), *DBN1* (25), *PCDH17* (30), *PTK7* (21), *ADAMTSL5* (21), *ARHGAP21* (30), *NFKB2* (23), *CDKN2B* (30), *RELB* (22), *DUSP8* (24), *SSBP4* (20), *IRX3* (27), *NEURL1B* (19). Differences in the expression of these 51 genes permitted segregating the H^+E^+ patient subset from the other (Fig. 6c).

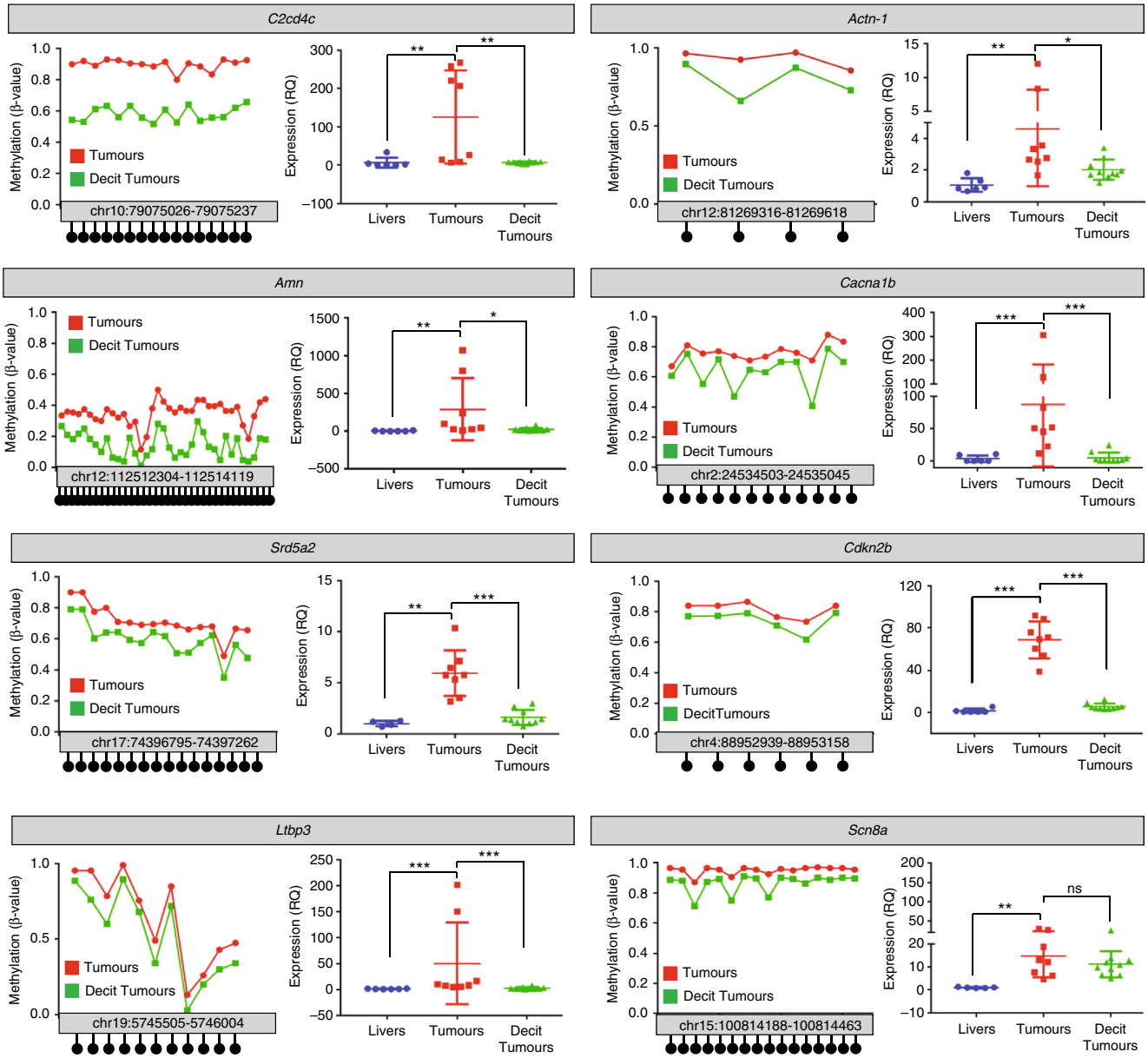


Fig. 4 Decitabine treatment decreases the expression and the CGI methylation levels of genes hypermethylated and overexpressed in *Alb-R26*^{Met} tumours. Expression and CGI methylation levels of a set of genes found hypermethylated and overexpressed in the *Alb-R26*^{Met} tumours were analysed in dissected tumours from *Alb-R26*^{Met} mice either untreated (red) or treated with Decitabine (green). For each indicated gene, graphs report the methylation levels of CpGs within the CGI of interest (left) and the expression levels of genes (right) in tumours from *Alb-R26*^{Met} mice either untreated (red) or treated with Decitabine (green), compared to control livers (blue). Note that demethylating treatment significantly decreased transcription levels. Concerning the *Scn8a* gene, the methylation levels of its gene body CGI was reduced in Decitabine treated tumours compared to untreated tumours. This was accompanied by a trend in downregulation of its expression levels, although not significant. It is possible that for *Scn8a*, the demethylation extent caused by the dose of Decitabine used is suboptimal to significantly influence its expression levels. Alternatively, a more complex mechanism could be involved in the regulation of *Scn8a* expression. Significance is indicated on the top. Not significant (ns): $P > 0.05$, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$

Furthermore, for each HCC patient we analysed the methylation levels of the CGIs corresponding to the 55 genes. We took into account that the number of CGIs for each gene varies between genes (Supplementary Data 7). 53/55 genes successfully lifted-over from mouse to human, and both methylation and expression data are available for 51 genes. These analyses revealed that 42/51 (82%) genes are both hypermethylated and overexpressed in at least 1 patient, and that 40/41 (97,5%) patients have at least 1 gene both hypermethylated and overexpressed (Fig. 6d, e, Supplementary Fig. 10, Supplementary Data 9). Additionally, there is a significant higher number of genes both hypermethylated and overexpressed in the H^+E^+ patient subset compared to the "NO H^+E^+" subset (H^+E^+ versus "NO H^+E^+": P-value < 0.001; Supplementary Fig. 10).

Curiously, in the *Alb-R26*^{Met} cancer model *Cdkn2a*, rather considered as a tumour suppressor, is overexpressed and hypermethylated in its gene body CGI, whereas no methylation changes were observed in its promoter CGI (Supplementary Data 4). We examined whether this phenomenon would also occur in HCC patients by analysing *CDKN2A* methylation and expression in HCC patients from TCGA and GSE56588 cohorts (for which methylation and expression data are available: 205/224 patients).

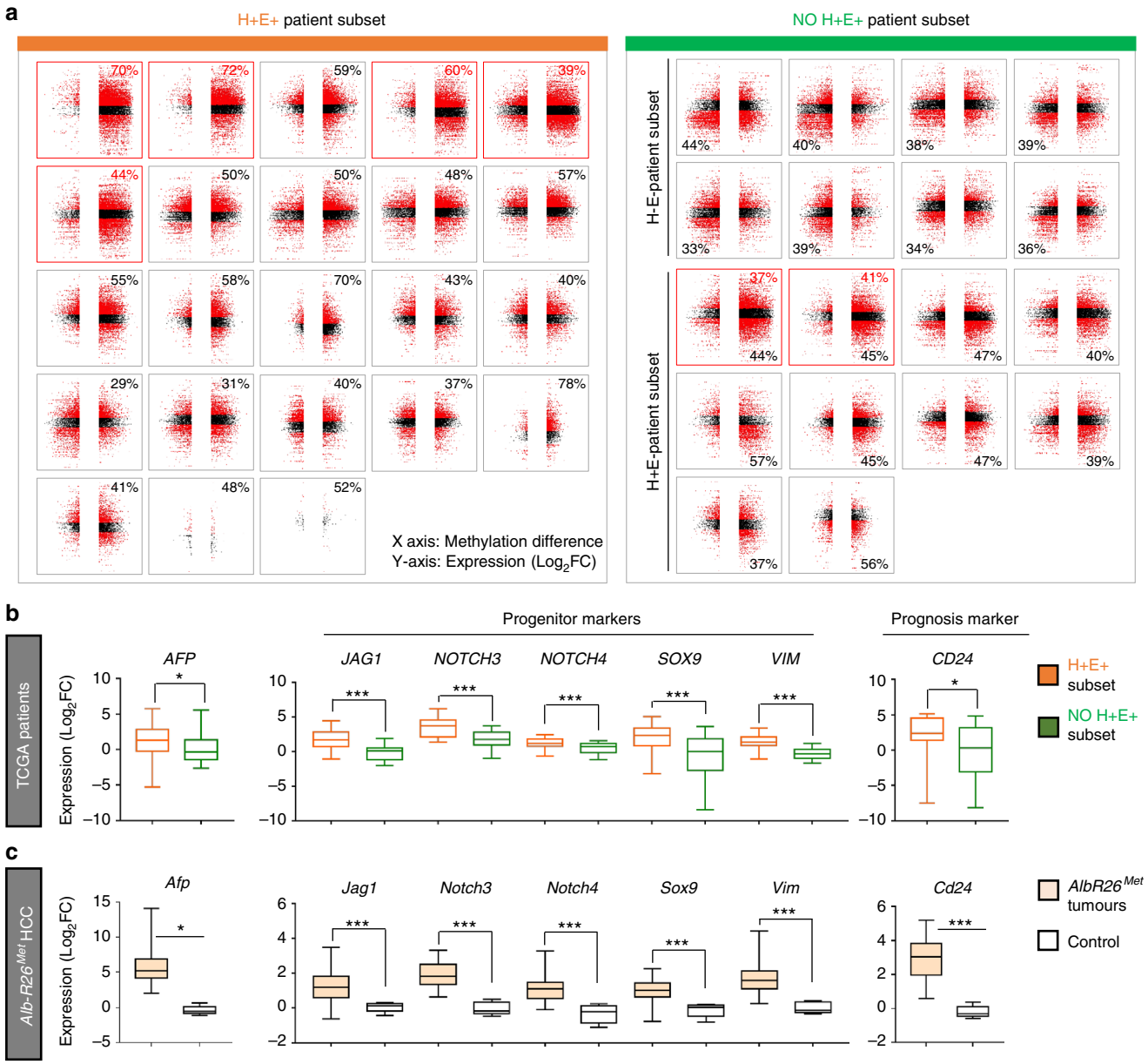


Fig. 5 A HCC patient subset, which is characterised by an enrichment of genes overexpressed and with hypermethylated CGIs, belongs to the HCC "proliferative-progenitor" subclass. **a** The 41 HCC patients are classified according to the highest percentage of genes over- versus underexpressed and with hyper- versus hypomethylated CGIs. In orange (left), patients with an enrichment of genes overexpressed and with CGI hypermethylation (H^+E^+ patient subset). Patients are organised according to the absolute number of hypermethylated CGIs. The percentage of genes overexpressed and hypermethylated is reported on the top. In green (right), all other patients are reported (NO H^+E^+ patient subset). Note that this patient subset is characterised by an enrichment in downregulated genes. Patients are organised according to an enrichment of genes with CGI hypomethylation (top) and hypermethylation (bottom). Concerning the 7 patients of the HCC subgroup 3 identified in Fig. 1f (corresponding to the best overlap patients), 5 of them belong to the H^+E^+ subset. Notably, all of these 7 patients are characterised by more than 37% of genes both hypermethylated and overexpressed (highlighted in panel with a red square and a red percentage of genes overexpressed with hypermethylated CGI). The *X*-axis reports methylation differences, whereas the *Y*-axis reports expression as Log_2FC. **b** Transcript levels (from RNA-seq data) of the indicated genes in H^+E^+ patient subset (in orange) versus the others (in green). Note significant high transcript levels of *AFP, JAG1, NOTCH3, NOTCH4, SOX9, VIM* and *CD24* in the H^+E^+ patient subset. **c** Transcript levels by RT-qPCR for the same genes shown in **b** analysed in *Alb-R26*^{Met} tumours versus control livers, displaying the same profile of gene upregulation as in the H^+E^+ patient subset. Data have been reported in ref.[24]. Significance is indicated on the top. $^*P < 0.05$, $^{***}P < 0.001$

Mouse *Cdkn2a* has two CGIs: one in the promoter and another in the gene body. Instead, human *CDKN2A* has 5 CGIs: one in the promoter and four in the gene body. Data are available only for the CGI in the promoter and for one of the four CGIs located in gene body. Notably, in both cohorts we found an enrichment of patients with an overexpression of *CDKN2A* (39/41 and 166/204, in the respective cohorts), which is associated to a hypermethylation of the gene body CGI (21/39 and 163/166, in the respective cohorts). In contrast, not methylation changes have been detected in the promoter CGI for both HCC cohorts (Supplementary Fig. 11).

Analysing pathway enrichments in KEGG pathways of genes overexpressed with hypermethylated CGIs, we identified a significant enrichment of several cancer-related pathways, such

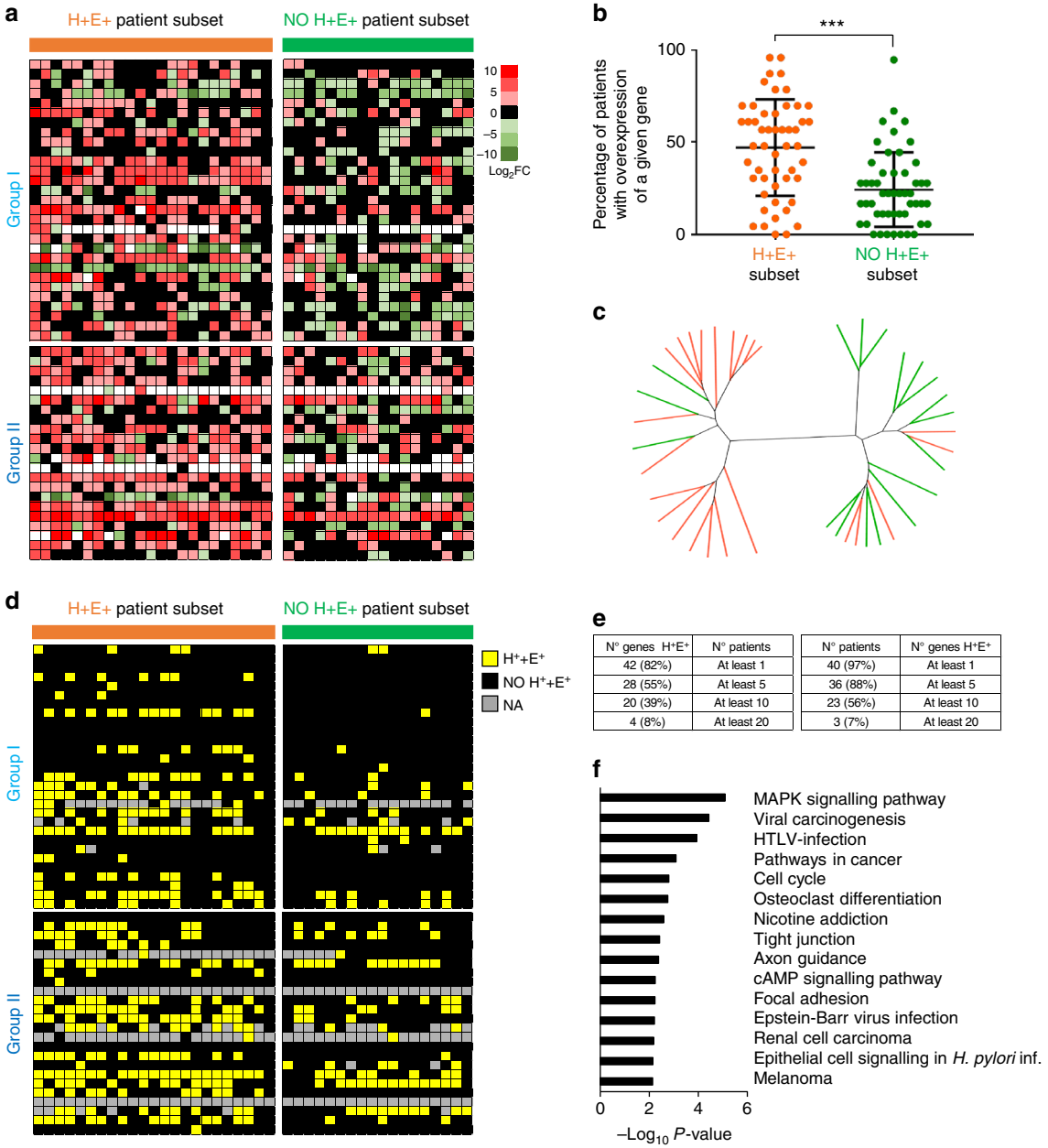


Fig. 6 The 55 genes identified in *Alb-R26^{Met}* tumours are also upregulated in the H$^+$E$^+$ patient subset, and a large proportion of them is characterised by hypermethylated CGIs. **a** Heat-map reporting expression levels of genes found overexpressed and with hypermethylated CGIs in *Alb-R26^{Met}* tumours versus control livers (rows; subdivided in Group I and II and organised according to the relative position to the ATG) in individual HCC patients (columns; organised as in Fig. 5a). Red: upregulated genes; green: downregulated genes. Black: not differentially expressed. White: data not available. The scale is shown on the right (expressed as Log_2 FC). The H$^+$E$^+$ patient subset is highlighted in orange (left), whereas all other patients in green (right). **b** In the graph, each dot corresponds to a given gene (the total 51 genes are reported). Their position corresponds to the percentage of patients in which the gene is overexpressed. In orange (left), for the H$^+$E$^+$ patient subset. In green (right), for all other patients. **c** Unrooted distance tree of HCC patients based on the 51 genes identified in the *Alb-R26^{Met}* tumours. In orange: H$^+$E$^+$ patient subset. In green: all other patients. **d** Heat-map highlighting with a yellow square genes overexpressed and with hypermethylated CGI in the corresponding HCC patient. The heat-map is organised as in panel A. Black square: genes not simultaneously overexpressed and with hypermethylated CGI in the corresponding HCC patient. Grey: data not available. **e** Tables reporting the numbers with percentages of genes overexpressed and with hypermethylated CGI in at least 1, 5, 10 or 20 HCC patients (left), as well as numbers with percentages of patients with at least 1, 5, 10 or 20 genes overexpressed and with hypermethylated CGI out of the 55 genes identified in the *Alb-R26^{Met}* model (right). **f** Histogram reporting the KEGG pathway enrichment analysis for the 55 genes identified in *Alb-R26^{Met}* tumours, ordered according to the $-Log_{10}$*P*-value. Significance is indicated on the top. ***$P < 0.001$

as MAPK signalling, viral carcinogenesis, pathways in cancer, cell cycle (Fig. 6f, Supplementary Fig. 12). Consistently, some of these genes are well-known oncogenes, such as *GRB10*, *MAP3K6*, *JUN* (which belong to MAPK pathway), *NFKB2*, *RELB* (which belong to NFkB and MAPK pathways), *MET*, *PTK7* (Supplementary Data 10). The presence of poorly characterised genes among well-established oncogenes prompted us to explore their functional relevance in cell tumorigenic properties. Focusing on *Scn8a*, *Actn1*, *Srd5a*, *NFkB2* and *Neurl1b*, we used shRNA-mediated targeting to lower their expression levels in *Alb-R26^{Met}* HCC cells

(Fig. 7a, Supplementary Fig. 13). Stable clones were used to assess cell tumorigenic properties in vitro and in vivo. These genes were selected because: (1) of their overexpression in HCC patients (*SCN8A* in 41%, *ACTN1* in 22%, *SRD5A2* in 5%, *NFkB2* in 56%, *NEURL1B* in 46%); (2) of their action as oncogenes in cancer cells (and particularly in HCC) has been less explored in previous studies (with the exception of *NFkB2* and *SRD5A*). Downregulation of either *Scn8a*, *Actn1*, *Srd5a*, *NFkB2* or *Neurl1b* interferes with the capability of cells to form: (a) colonies in anchorage-independent assays (Fig. 7b); (b) foci in anchorage-dependent assays, as revealed by a significant smaller foci size even if numbers were similar (Fig. 7c); (c) tumour spheres when cells were grown in self-renewal conditions (Fig. 7d); (d) tumours in nude mice xenografts (Fig. 7e). Collectively, these data show that most of the 55 genes identified in the *Alb-R26^{Met}* cancer model are also overexpressed and with hypermethylated CGIs in a large proportion of HCC patients, with a set of them acting together as an "oncogenic module".

Discussion

The increasing knowledge on how epigenetic modifications such as DNA methylation influence patterns of gene expression in cancer holds great promises for understanding biological processes, as well as for their use in prognosis, patient stratifications and therapeutic intervention[3,29]. This is well exemplified by reports showing correlations between changes in CGI methylation and a remarkable resetting of transcriptional networks in cancer. In the present study, we employed a clinically relevant cancer mouse model in which tumorigenesis is triggered by a slight perturbation in signalling dosages rather than drastic genetic modifications, to examine the DNA methylation landscape associated with tumorigenic acquisition. We reasoned that such a genetic tool offers a unique way to model DNA methylation changes occurring in human cancerogenesis in the absence of drastic alterations of epigenetic modulators. Our genome-wide strategy highlighted key correlations between site-specific DNA methylation changes and transcriptional dosages of the corresponding genes. The type of changes found for some genes belong to the well-known mechanism of downregulation of tumour suppressors through promoter DNA hypermethylation, which was the case of *Oat* and *Igfbp5* that can act as tumour suppressors in certain cellular contexts. Quite unexpectedly, however, there is an enrichment of genes both overexpressed and with hypermethylated CGIs. Several of them are well-known oncogenes, such as *Grb10*, *Map3k6*, *Jun*, *RelB*, *Met*, *Ptk7*, as well as *NF-KB2*, *Srd5a2*, which have been functionally validated in this study together with others poorly investigated so far: *Scn8a*, *Actn1*, *Neurl1b*. Results from our functional assays in *Alb-R26^{Met}* HCC cells demonstrate how downregulating each individual oncogene reduces, but not abolishes, cell tumorigenic properties. These results conceptually illustrate that, although each oncogene contributes to the tumorigenic properties of cancer cells, they operate in a cooperative manner as an "oncogenic module" for ensuring robustness of the tumorigenic program.

Our integrative studies using human HCC databases demonstrate that enrichment in genes both overexpressed and with hypermethylated CGIs also characterises 56% of the HCC patients, which we named as the "H^+E^+ patient subset". For several genes, upregulation in expression levels is coherent with them being bona fide oncogenes. For example, it is the case of *WT1*, *DLK1*, *TP73*, *EEF1A2*, *IGF1R*, *DKK1*, *SPOCK1*, *ITPKA*, *HOXA3*, *NOX4*, *FZD10*, *VASH2*, *GATA2*, *SOX8*. Thus, our genetic studies together with a revisited analysis of human cancer databases reveal that raising dosages of oncogene sets characterised by hypermethylated CGIs is a robust mechanism operating in cancer. The existence of such events in human pathology supports the clinical relevance of these findings. Remarkably, the H^+E^+ patient subset belongs to the "HCC proliferative-progenitor" subclass, thus attributing an additional feature to this aggressive HCC subtype. For clinical implementation, integrative methylome and transcriptome analyses on additional HCC cohorts will demonstrate the robustness of H^+E^+ patient subset classification. These findings also raise the question as to whether the H^+E^+ patient subset could be most sensitive to therapies based on demethylation agents[30].

Our expression analyses revealed that H^+E^+ genes can be segregated into two groups, according to their relative position to the ATG, with overall significant higher expression levels observed the further the CGI is located from the ATG (Group-II). Whereas for Group-I the relative position of the hypermethylated CGIs falls predominantly with the 5′-UTR, their location in Group-II is in the gene body. The positive correlation between gene body hypermethylation and expression is coherent with previous studies based on in vitro modulation of the methylation content in cancer cell lines[12]. Additionally, single-base resolution DNA methylation profiling combined with transcriptome analysis correlated changes in gene expression levels with the CpG methylation content in gene body[31–34]. The significance of gene body DNA methylation on transcriptional regulation is strengthened by studies exploring correlations with chromatin modifications. It has been reported that in the gene body: (a) H3K4me3 association to alternative promoters depends on their CpG methylation content, impacting alternative transcript products[35]; (b) H3K36me3 associates with methylated DNA in gene body and permits transcription[36]; (c) CTCF binding is lost in hypermethylated CGI, influencing splicing, in addition to the well-known action of CTCF in maintenance of chromatin architecture through generation of chromatin barriers[37,38]; (d) H3K27me3 and H3K9me3, known as repressive histone marks, are not associated with methylated DNA[39]. Future studies integrating methylome, transcriptome and ChIP-seq with several chromatin marks like those mentioned above will contribute to uncover the underlying mechanisms of action of oncogene upregulation through gene body methylation. Taking into account the variety of chromatin factors found associated in gene body, it is likely that different sets of genes are modulated by different mechanisms of action.

For translating these findings into therapies, an intriguing question is whether and to what extent the epigenetic reprogramming of a set of genes acting as an "oncogene module" still leaves space for tumour vulnerability. Our functional studies show that targeting each individual oncogene reduces, but not abolishes, tumorigenicity, indicating that each oncogene provides a net contribution to the whole tumorigenic properties of cancer cells. Such context may be particularly relevant for tumours that are not predominantly "addicted" to genetic mutation(s)[40], such as HCC. This likely explains the partial response of HCC patients even with most promising drugs targeting one or at least a restricted number of targets (e.g., Sorafenib). Such scenario contrasts with exceptional cases of effectiveness, due to the stringent addiction of cancer cells to a given oncogene, such as BCR-ABL in chronic myeloid leukaemia, ERBB2 in breast cancer, ERBB1 in non-small cell lung cancer, B-RAF in metastatic melanoma. To identify vulnerability, an approach could be to extract enriched pathways that are deregulated from the whole list of epigenetically reprogramed genes. In the case of tumorigenesis modelled by the *Alb-R26^{Met}* mice, the MAPK signalling cascade is on the top of the list of enriched pathways (with 11 genes differentially methylated in tumour versus control livers). Through a phosphokinome-based educated guess drug screen, we recently reported that tumorigenesis modelled by the *Alb-R26^{Met}* genetic

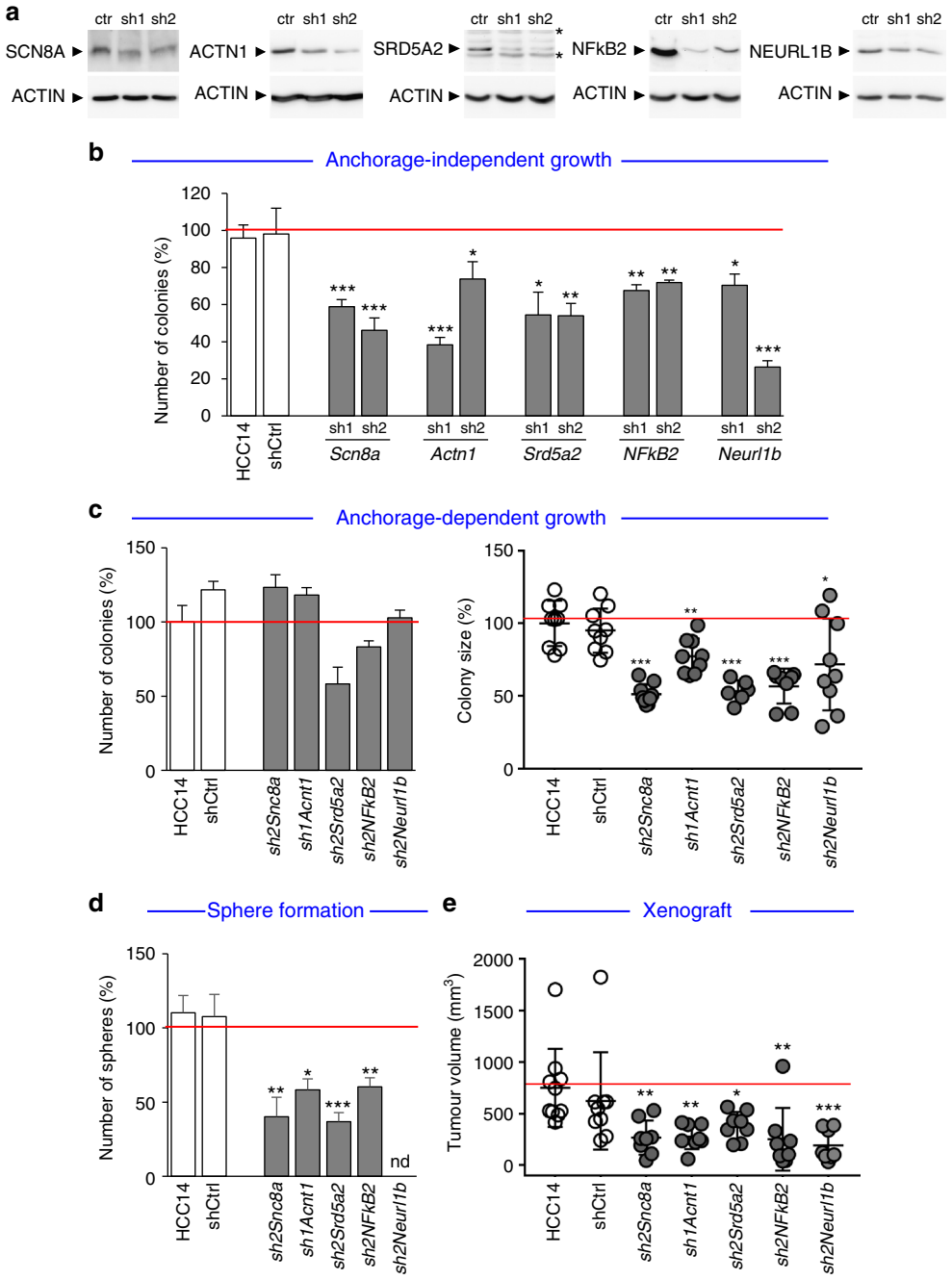


Fig. 7 Downregulation of overexpressed genes with hypermethylated gene body CGI in *Alb-R26^{Met}* HCC cells interferes with their tumorigenic properties both in vitro and in vivo. **a** Western blots showing SCN8A, ACTN1, SRD5A2, NFkB2 and NEURL1B protein levels in stable clones established after transfection of *Alb-R26^{Met}* HCC14 cells with plasmids carrying a shRNA sequence targeting the corresponding gene. Protein levels were compared to control cells (ctr). ACTIN was used as a loading control in all western blots. The asterisk indicates nonspecific bands detected using anti-SRD5A2 antibodies. **b**–**e** Biological assays to assess functional properties of *Alb-R26^{Met}* HCC14 cells carrying a shRNA sequence targeting candidate genes. Effects were compared to HCC14 cells either untransfected or transfected with a control shRNA (shCtrl). **b** Graph reporting the number of colonies formed in anchorage-independent growth assays using 2 different shRNA targeting sequences for each candidate gene. Note a decrease in colony number formation of cells with downregulated candidate genes compared with control cells. **c** Graphs reporting the number (left) and the size (right) of colonies formed in anchorage-dependent growth assays. Whereas no significant changes in colony numbers were detected, note a significant decrease in colony size when the candidate gene is downregulated. **d** Graph reporting number of spheres formed in tumour sphere assays. Note that downregulation of candidate genes significantly reduces sphere number formation. **e** Graph reporting the tumour volume of mice injected either with *Alb-R26^{Met}* HCC14 control cells or with *Alb-R26^{Met}* HCC14 cells carrying a shRNA sequence targeting candidate genes. Note that downregulation of candidate genes significantly interferes with the in vivo tumorigenic properties of *Alb-R26^{Met}* HCC14 cells. Significant differences between groups are indicated on the top. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$ (nd: no determined)

system is vulnerable to Ras pathway targeting, provided that its inhibition occurs concomitantly while destabilising the stress support mitochondrial pathway[24]. It is therefore tempting to speculate that a proportion of tumours, particularly those with epigenetic reprogramming rather than those with drastic genomic instability, still maintains vulnerability to synthetic lethal interactions. Alternatively, the use of epigenetic modulating agents to reprogram a set of genes in cancer cells, ideally used at minimal doses to limit side effects[3,29], could reinforce the action of promising targeted therapies that, when used alone, have been unsatisfactory in clinical trials, as reported for chronic myeloid leukaemia[41]. We show here that Decitabine treatment reduces the methylation levels of gene body CGIs and the expression levels of the corresponding genes. Such event correlates with reduced tumorigenic properties of *Alb-R26^{Met}* HCC cells. Nevertheless, cells for 10 days in culture after 48 h with Decitabine recover their tumorigenic properties (Supplementary Fig. 14), illustrating how reduced tumorigenicity by transient demethylation treatment is reversible, likely by resetting increased levels of oncogenes. This would be coherent with previous studies showing the capability of cancer cells to restore acquired epigenetic modifications[12]. Collectively, these findings support the possibility of achieving effective response in cancer combining epigenetic modulating agents with targeted treatments.

In conclusion, by exploring epigenetic changes associated with tumorigenesis in a clinically relevant mouse model, we discovered that for oncogene sets, characterised by hypermethylated CGIs either in their 5′-UTR or in the gene body, their expression levels are raised in cancer. The use of a mouse model in which tumorigenesis is not caused by drastic genetic manipulations strengthens the advantage of disrupting multiple oncogenes through an epigenetic reprogramming. Delineating the relationship between aberrant DNA methylation and expression of oncogenes/tumour suppressors will likely contribute to identify biomarkers for patient stratifications, functional pathways operating in cancer and strategies for an epigenetic restoration of deregulated genes in combination with molecular therapies.

Methods

Mice. Ethics Statement: All procedures involving the use of animals were performed in accordance with the European Community Council Directive of 22 September 2010 on the protection of animals used for experimental purposes (2010/63/UE). The experimental protocols were carried out in compliance with institutional Ethical Committee guidelines for animal research (comité d'éthique pour l'expérimentation animale—Comité d'éthique de Marseille; agreement number D13-055-21 by the Direction départementale des services vétérinaires—Préfecture des Bouches du Rhône).

Alb-R26^{Met} mice: *R26^{stopMet}* and *Alb-R26^{Met}* mice have been described previously[42,43]. Briefly, *R26^{stopMet}* mice (international nomenclature *Gt(ROSA)26Sor$^{tm1(Actb-Met)Fmai}$*) carrying a conditional mouse–human chimeric *Met* transgene in the *Rosa26* locus were crossed with *Albumin-Cre* mice (*B6.Cg-Tg(Alb-cre)21Mgn/J*) obtained from the Jackson Laboratory. All mice were maintained on a 50% mixed 129/SV and C57BL6 background. Mice were genotyped via PCR analysis of genomic DNA as reported in previous studies[42,43]. Mice were housed under pathogen-free conditions.

Mice drug treatment: For in vivo demethylation experiments to asses methylation levels of selected CGIs, as well as expression levels of the corresponding genes, *Alb-R26^{Met}* mice were treated with intraperitoneal injections of 2.5 mg/kg of Decitabine, twice per week (for a total of three injections).

DNA/RNA-related experiments. Genomic DNA isolation: Genomic DNA from *Alb-R26^{Met}* tumours and control livers was prepared using the ZR Genomic DNA Tissue Miniprep (Zymo Research Company), according to the manufacturer's instructions.

Total RNA extraction: Total RNA from frozen tissues and cultured cells was isolated using the RNeasy Mini Kit (Qiagen), according to the manufacturer's instructions. DNase (Qiagen) treatment was included to avoid possible genomic DNA contamination. Regarding frozen samples, 20 mg of tissue were first homogenised in the specific lysis buffer by 6300 r.p.m. 2 × 30 s using Precellys 24 (Bertin technologies), then the RNeasy Mini Kit (Qiagen) was used.

cDNA and quantitative RT-PCR analysis: cDNA was synthesised using a Reverse Transcription Kit (Bio-Rad). PCR reactions were performed using 2X SYBR Green qPCR SuperMix-UDG with Rox (ThermoFisher Scientific) and specific primers (1 µM; qPCR primer sequences are listed in Supplementary Data S11). Expression levels were quantified using the comparative Ct method ($2^{-\Delta\Delta CT}$ method) with the house-keeping gene *Hprt* as a control for internal normalisation, and results are expressed as $RQ = 2^{-\Delta\Delta CT}$.

High-throughput sequencing. Comparative Genomic Hybridisation analysis: Genomic DNA form dissected *Alb-R26^{Met}* tumours ($n = 16$) and control livers ($n = 8$) was analysed by the "Plateforme Biopuces et Sequencage IGBMC" (Illkirch, France) using an Agilent Oligonucleotide Array-Based CGH for Genomic DNA Analysis (CGH microarray 4 × 180 K).

Genome-wide DNA methylation analysis: Methyl-MiniSeq EpiQuest genome-wide sequencing was perform using genomic DNA from dissected *Alb-R26^{Met}* tumours ($n = 10$) and control livers ($n = 3$) to analyse the DNA methylation profile by the Zymo Research Corporation (Irvine, CA, USA).

Library construction. Libraries were prepared from 200–500 ng of genomic DNA digested with 60 units of TaqαI and 30 units of MspI (NEB) sequentially, then extracted with Zymo Research DNA Clean and Concentrator™-5 kit (Cat#: D4003). Fragments were ligated to pre-annealed adapters containing 5′-methyl-cytosine instead of cytosine according to Illumina's specified guidelines (www.illumina.com). Adaptor-ligated fragments of 150–250 bp and 250–350 bp in size were recovered from a 2.5% NuSieve 1:1 agarose gel (Zymoclean™ Gel DNA Recovery Kit, Zymo Research Cat#: D4001). The fragments were then bisulfite-treated using the EZ DNA Methylation-Lightning™ Kit (Zymo Research, Cat#: D5020). Preparative-scale PCR was performed and the resulting products were purified (DNA Clean and Concentrator™–Zymo Research, Cat#D4005) for sequencing on an Illumina HiSeq.

Alignments and data analysis. Sequence reads from bisulfite-treated EpiQuest libraries were identified using standard Illumina base-calling software and then analysed using a Zymo Research proprietary analysis pipeline, which is written in Python and used Bismark (http://www.bioinformatics.babraham.ac.uk/projects/bismark/) to perform the alignment. Index files were constructed using the Bismark-genome-preparation command and the entire reference genome. The non-directional parameter was applied while running Bismark. All other parameters were set to default. Filled-in nucleotides were trimmed off when doing methylation calling. The methylation level of each sampled cytosine was estimated as the number of reads reporting a C, divided by the total number of reads reporting a C or T (β-value).

Overall sequencing results (for 13 samples) are: (a) mean total read: 30 million read pairs, (b) mean mapping efficiency: 40%, (c) mean unique CpGs: 4.1 millions, (d) mean average CpG coverage: 16×, (e) mean bisulfite conversion rate: 98%. Data accessibility: Methylome datasets generated from this study are deposited with the Gene Expression Omnibus (accession GSE90093).

Identification of differentially methylated CpGs. A total of 1.085.757 unique single CpG sites, present in all samples, were analysed. β-value ranged from 0 (not methylated) to 1 (fully methylated). To identify differentially methylated CpGs, the methylation difference per CpG was calculated as the mean β-value of tumours minus the mean β-value of controls. Those with a methylation difference > 0.2 were filtered to retain the ones with a FDR < 0.05 (Student's two-sided *T*-test and Benjamini–Hochberg False Discovery Rate for *P*-value correction). A CpG is classified as "hypomethylated" when the methylation difference is <−0.2 and as "hypermethylated" when the methylation difference is >0.2. A global analysis was first carried out with all measured CpGs, then dividing the CpGs according to their location within or outside a CGI (CpG Island bedfile downloaded from UCSC). According to the Methyl-MiniSeq EpiQuest coverage, the CGI coverage by CpGs was 87.5%. Studies were focused on CGI regions. The overlap with CGI and the annotated gene was performed using the CGI track from the UCSC genome browser, and Refseq gene annotations based on the NCBI37/mm9 mouse reference. We discarded "ubiquitous CpGs" located in more than one annotated gene, and we extended the gene/CGI annotation to the gene's promoter region to −1.5 kb upstream the TSS.

Targeted Bisulfite Sequencing: Genomic DNA from *Alb-R26^{Met}* tumours dissected from mice treated with Decitabine (2.5 mg/kg; twice per week, for a total of three treatments; $n = 4$) and without treatment ($n = 2$) was used to asses CpG methylation levels in selected regions within the candidate CGIs through bisulfite sequencing by the Zymo Research Corporation (Irvine, CA, USA).

Assay Design, Sample Preparation and Multiplex Targeted Amplification. After assessment of DNA concentration and quality, DNA samples were bisulfite converted using the EZ DNA Methylation-LightningTM Kit (ref Cat#D5030) according to the manufacturer's instructions. Primers were designed with Rosefinch, Zymo Research's proprietary sodium bisulfite converted DNA-specific primer design tool (primer sequences are listed in Supplementary Data 5). Multiplex amplification of all samples using the specific primer pairs and the Fluidigm Access ArrayTM System was performed according to the manufacturer's instructions. The resulting amplicons were pooled for harvesting and subsequent barcoding according to the Fluidigm instrument's guidelines. After barcoding, samples were purified (ZR-96 DNA Clean and Concentrator™ –ZR, Cat#D4023), then prepared for parallel sequencing using a MiSeq V2 300 bp Reagent Kit and paired-end sequencing protocol, according to the manufacturer's guidelines.

Targeted Sequence Alignments and Data Analysis. Sequence reads were identified using standard Illumina base-calling software and then analysed using a Zymo Research proprietary analysis pipeline, which is written in Python. Sequence reads were aligned back to the reference genome using Bismark (http://www.bioinformatics.babraham.ac.uk/projects/bismark/), an aligner optimised for bisulfite sequence data and methylation calling[44]. Paired-end alignment was used as default thus requiring both read 1 and read 2 be aligned within a certain distance, otherwise both read 1 and read 2 were discarded. Index files were constructed using the bismark_genome_preparation command and the entire reference genome. The non-directional parameter was applied while running Bismark. All other parameters were set to default. The methylation level of each sampled cytosine was estimated as the number of reads reporting a C, divided by the total number of reads reporting a C or T.

Transcriptome analysis by RNA-seq: Total RNA from dissected *Alb-R26^{Met}* tumours ($n = 4$) and control livers ($n = 4$) was processed for transcriptome analysis. RNA quality was controlled using the Agilent RNA 6000 Pico Kit and Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, California) according to the manufacturer's recommendations. Total RNA (1 µg per sample) was used for library preparation using the TruSeq RNA Sample Preparation Kit (Illumina) by GATC Biotech (Mulhouse; NGSelect service). Sequencing was performed on a HiSeq 2500 (Illumina; 2 × 50 bp paired end) and base calling performed using RTA (Illumina). Quality control of raw reads was done using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were mapped to the reference genome mm9 with STAR aligner[45] using default parameters; differential expression was calculated using the Cufflinks package[46].

Cell culture-related experiments. Cell lines: *Alb-R26^{Met}* HCC cell lines (HCC3, HCC13 and HCC14) were established, characterised and cultured as previously described[24]; cells were regularly tested by PCR-based assay to confirm their maintenance in free *Mycoplasma* culture condition.

shRNA-mediated downregulation of candidate genes: The functional relevance of candidate oncogenes was determined using shRNA targeting sequences (Sigma; shRNA sequences are reported in Supplementary Data 12). In particular, plasmids carrying the shRNA sequence were transfected in cells using Lipofectamine 2000 reagent, according to the manufacturer' instructions (ThermoFisher Scientific). After 1 week of puromycin selection, pools of resistant clones were used to verify downregulation of gene expression levels (by RT-qPCR and western blots) and to perform biological assays.

Cell drug treatment: For the demethylation experiments, cells were exposed to 0.3 µM of Decitabine (5-Aza-2'-deoxycytidine; Selleckchem) for 48 h. After treatment, cells were used for cell viability, anchorage-dependent growth assay, anchorage-independent growth assay, tumour sphere formation assay and xenograft studies. For experiments shown in Supplementary Fig. 14, after 48 h of Decitabine treatment, cells were cultured for 10 days with complete media before performing the experiments.

Survival assay: Cells were seeded in a 150-µl volume per well in 96 well plates (10,000 cells/well) in 10% serum for 24 h, then Decitabine treatment was applied at 0.3 µM in the corresponding wells. After 48 h, cell viability was assessed in a Cell Titer Glo Luminescent Assay (Promega) and luminescent signals were measured with a luminometer microplate reader (Berthold). Data are expressed as means ± SEM of three independent experiments performed in triplicate.

Anchorage-dependent growth assay (focus formation assay): To measure anchorage-dependent growth, 300 cells were seeded in 10 ml complete media in a 10 cm dish. After 7 days, foci were stained with a 0.2% crystal violet solution (2% methanol). The total number of foci and individual foci size were quantified using ImageJ program. Data are expressed as means ± SEM of three independent experiments performed in triplicate.

Anchorage-independent growth assay (soft agar assay): Assays were performed as previously described[47–49]. Briefly, cells were cultured in 12-well plates containing two layers of agar. Cells (6×10^3) were resuspended in 0.5% agar diluted in complete medium and poured onto a 1% layer of agar (diluted in medium). Fresh medium was added to the top layer every 3 days. After 2 weeks, colonies were stained with MTT, pictures were taken using a dissecting microscope, and colonies were counted using ImageJ software. Numbers are expressed as means ± SEM of three independent experiments performed in triplicate.

Tumour sphere forming assay: Cells were cultured at a density of 2×10^4/35 mm dishes in a stem cell-permissive media. In particular, cells were cultured for one week in DMEM/F12 medium supplemented with 1% N-2 Supplement, 2% B27, 50 mg/ml of Penicillin-Streptomycin, glutamine (Gibco), 0.01% Bovine Serum Albumin (BSA), 5 mg/ml of insulin (Sigma) and growth factors including 10 ng/ml of basic fibroblast growth factor (bFGF), 20 ng/ml of epidermal growth factor (EGF) and 10 ng/ml of hepatocyte growth factor (HGF; Peprotech). After one week, pictures of the whole dish were taken using a dissecting microscope, and spheres were counted using ImageJ software. Numbers are expressed as means ± SEM of three independent experiments performed in triplicate.

In vivo tumorigenesis assays (xenografts in nude mice): For in vivo demethylation studies, xenografts were performed using *Alb-R26^{Met}* HCC cells either untreated or pre-treated for 48 h with Decitabine (0.3 µM). Cells (5×10^6) were then resuspended in a 1:1 Matrigel:PBS solution (Corning BV) and inoculated subcutaneously into the flank-leg region of nude mice (S/SOPF SWISS NU/NU; Charles River). After 5 days of cell inoculation, mice were treated with intraperitoneal injections of vehicle or Decitabine (2.5 mg/kg) twice per week for 3 weeks. Mice were then sacrificed and tumour volume was measured as length × width × height. For assessment of in vivo tumorigenic capacity of candidate genes, xenografts were performed using *Alb-R26^{Met}* HCC cells (1×10^6) either un-transfected, transfected with shControl, or with a shRNA sequence targeting the candidate gene. Tumour volume was followed every week. After 6 weeks mice were sacrificed and tumour volume after dissection was determined as length × width × height.

Western blots: Protein extracts from HCC cells were prepared and western blot analysis was performed as previously described[43,48,49]. For SCN8A detection, protein lysates were run on a 5% SDS gel and transferred overnight at 300 mA in the presence of 0.1% SDS. The acquisition of ECL signal was performed using the MyECL imager system (ThermoFisher Scientific)(Supplementary Fig. 15).

Antibodies: Antibodies used were: anti-SCN8A (Abcam, #ab65166; 1:500), anti ACTN-1 (Cell Signalling, #6487; 1:3000), anti-SRD5A2 (ThermoFisher Scientific, #PA5-25465; 1:1000), anti-NFkB2 (Cell Signalling; #4882; 1:1500), anti-NEURL1B (Abcam, #ab156988; 1:3000), anti-ACTIN (Sigma, A3853; 1:5000), anti-rabbit IgG-peroxidase or anti-mouse IgG-peroxidase (Jackson; 1:4000).

Computational analyses. Unsupervised hierarchical clustering analysis: Clustering statistics was determined by using the methylation values of all CGIs for each sample. We applied the Principal Component Analysis and the Agglomerative Distance Tree using the "linkage" function with unweighted average euclidean distance for calculating the similarity matrix of samples and the "dendrogram", as well as "phylotree" function to plot the hierarchical and distant trees (both are from Matlab Statistical Toolbox). For studies reported in Supplementary Fig. 7, clustering analysis of both methylome and expression data was performed using the function "hclust" on an Euclidean distance matrix of samples, which was computed with the function "dist". "hclust" then returned a tree-like structured object that could be plotted as dendrogram by "plot" (R, version 3.3.1).

Identification of human CGIs corresponding to the mouse CGIs of interest: To compare methylome outcomes identified in the *Alb-R26^{Met}* genetic system with those available for human studies, genomic coordinates were converted from mm9 to GRCh37/hg19 by using the "Lift-Over" tool available from UCSC (https://genome-euro.ucsc.edu/cgi-bin/hgLiftOver). This allowed us to successfully map 501 out of 513 CGIs from mouse to human regions (Supplementary Data 2). Among them, we only kept 501 unique human regions by discarding duplicate lift-overs. We also discarded 14 human regions not overlapping with any human CGI. We then check into TCGA patient datasets the presence of methylation data for those CGIs. We focused the analysis on the patient having both tumour and control samples (adjacent liver) methylation data, and we discarded the CGIs having no entry into any of the TCGA patient dataset. Finally, the total CGIs used for comparative analyses between mouse and human is 416.

Analysis of public available DNA methylome data: The human methylome data is available through firebrowse (www.firebrowse.org) by the BROAD Institute and is based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. The publicly available methylome data (Level 3 data) of HCC patients from TCGA is generated with the platform Illumina Infinium Human DNA Methylation 450 and contains beta values for 485778 CpGs. Patients with both tumour and control samples were extracted and calculation of methylation difference per CpG was applied (β-values of tumour–β-values of control). Student's *T*-test was used to compare between tumour and normal samples, and the *P*-values were corrected with Benjamini–Hochberg False Discovery Rate (FDR). As our methylome screen focused on CGIs, we revisited the human data (from TCGA and from GSE56588) to generate a list of all CpGs within CGIs with the corresponding methylation β-values. By applying the same methylation difference and FDR thresholds used for *Alb-R26^{Met}* methylome data, we extracted a list of differentially methylated CpGs from the human HCC dataset.

Methylome overlap between *Alb-R26^{Met}* outcomes and human data: A methylation overlap between *Alb-R26^{Met}* and human HCC was considered only when a given CGI was differentially methylated in both species. To define the methylation status of a given CGI, the CpG with an absolute maximum methylation difference among all patient samples was chosen as a representative probe (with *P*-value threshold and fold change cut-off defined above). This CpG was analysed in all HCC patients. An overlap score (in percentage) was determined by calculating the number of human CGIs differentially methylated versus the total number of lifted-over CGIs subset.

Analysis of public available RNA-seq data: The human RNA-seq data from TCGA was available through firebrowse. The data is generated with the platform Illumina HiSeq 2000 Sequencing System and uses MapSplice[50] to do the alignment and RSEM[51] to perform the quantitation. The scaled estimate from RSEM output was used as this value could be multiplied by 10^6 to obtain a measure in terms of transcripts per million (TPM), which is preferred over RPKM[52] or FPKM[53] as it is independent of the mean transcript length and therefore more comparable across samples[51]. The TPM is calculated for each gene and the calculation of Log_2 Fold Change (Log_2(tumour sample)−Log_2(control sample)) was applied to each patient with available data from both tumour and control samples.

Calculation of the relative position to the ATG: For calculating the position of CpGs, we used the longest transcript for each gene. The gene length was reported with values ranging from −100% and +100% (transcription end site: TES), with the ATG at position 0. The relative position for each CpG was then reported

relative to its distance to the ATG. A positive relative position corresponds to a genomic region located downstream the ATG, whereas a negative relative position stands for a genomic region located upstream the ATG.

Analysis of public available data from a mouse HCC model carrying the viral hepatitis B virus X expression: Using available methylome and expression data based on a HCC model induced by the viral hepatitis B virus X (*HBx*^{tg}; GSE48052[27]), we performed the same analysis done for the *Alb-R26*^{Met} model (Fig. 3b). For each CpG, the methylation difference between *HBx*^{tg} tumour and control sample was calculated as the difference of the RPKM. For those CpGs found differentially methylated, the expression of the corresponding gene was then calculated as the difference of the RPKM sum within the TSS and TSE.

Pathway enrichment analysis: For these analyses (shown in Supplementary Figs. 5, 9), identified genes were used as an input for KEGG pathway enrichment analysis with the REST API tool (http://rest.kegg.jp). Pathways were further ranked by $-\log_{10}$ *P*-value after applying the hypergeometric probability density function (Matlab function "hygepdf" from Statistical Toolbox).

Statistical analysis: All data were analysed using GraphPad Prism software (version 7.01) and Matlab Statistical Toolbox (version R2015b). Results are expressed as the median (indicated by a line) or as the mean ± standard error of the mean (SEM), according to sample distributions. Statistically significant differences were estimated by applying unpaired Student *t*-tests to data showing normal distributions, and Mann–Whitney tests in all other situations. Moreover, one-way-ANOVA was used to determine differences between the means of independent groups (in vivo xenograft experiments in Figs. 2h and 7d), and Fisher's exact test for categorical variables (risk factors in Supplementary Fig. 9). All statistical tests were two-sided. Statistical significance was defined as not significant (ns): $P > 0.05$; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$. Significance is indicated in figures

Data availability. Raw and processed data of bisulfite sequencing have been deposited to the Gene Expression Omnibus (GEO) [GEO: GSE90093]. The authors declare that all data supporting the findings of this study are available within the article and its Supplementary Information files, or from the authors upon reasonable request.

Received: 1 November 2017 Accepted: 10 July 2018
Published online: 08 August 2018

References

1. Esteller, M. Epigenetic changes in cancer. *F1000 Biol. Rep.* **3**, 9 (2011).
2. Baylin, S. B. & Jones, P. A. A decade of exploring the cancer epigenome-biological and translational implications. *Nat. Rev. Cancer* **11**, 726–734 (2011).
3. Dawson, M. A. & Kouzarides, T. Cancer epigenetics: from mechanism to therapy. *Cell* **150**, 12–27 (2012).
4. Laird, P. W. Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.* **11**, 191–203 (2010).
5. Ball, M. P. et al. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.* **27**, 361–368 (2009).
6. Woo, H. G. et al. Integrative analysis of genomic and epigenomic regulation of the transcriptome in liver cancer. *Nat. Commun.* **8**, 839 (2017).
7. Feinberg, A. P. & Tycko, B. The history of cancer epigenetics. *Nat. Rev. Cancer* **4**, 143–153 (2004).
8. Jones, P. A. & Baylin, S. B. The epigenomics of cancer. *Cell* **128**, 683–692 (2007).
9. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
10. Murtha, M. & Esteller, M. Extraordinary cancer epigenomics: thinking outside the classical coding and promoter box. *Trends Cancer* **2**, 572–584 (2016).
11. Heyn, H. et al. Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer. *Genome Biol.* **17**, 11 (2016).
12. Yang, X. et al. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* **26**, 577–590 (2014).
13. Jjingo, D., Conley, A. B., Yi, S. V., Lunyak, V. V. & Jordan, I. K. On the presence and role of human gene-body DNA methylation. *Oncotarget* **3**, 462–474 (2012).
14. Spencer, D. H. et al. CpG island hypermethylation mediated by DNMT3A is a consequence of AML progression. *Cell* **168**, 801–816 e813 (2017).
15. Bender, S. et al. Reduced H3K27me3 and DNA hypomethylation are major drivers of gene expression in K27M mutant pediatric high-grade gliomas. *Cancer Cell* **24**, 660–672 (2013).
16. Lu, C. et al. Histone H3K36 mutations promote sarcomagenesis through altered histone methylation landscape. *Science* **352**, 844–849 (2016).
17. Lennartsson, A. External signals shape the epigenome. *Genome Biol.* **17**, 18 (2016).
18. Liu, F., Wang, L., Perna, F. & Nimer, S. D. Beyond transcription factors: how oncogenic signalling reshapes the epigenetic landscape. *Nat. Rev. Cancer* **16**, 359–372 (2016).
19. Yao, X., Xing, M., Ooi, W. F., Tan, P. & Teh, B. T. Epigenomic consequences of coding and noncoding driver mutations. *Trends Cancer* **2**, 585–605 (2016).
20. Liu, F. et al. EGFR mutation promotes glioblastoma through epigenome and transcription factor network remodeling. *Mol. Cell* **60**, 307–318 (2015).
21. Forloni, M. et al. Oncogenic EGFR represses the TET1 DNA demethylase to induce silencing of tumor suppressors in cancer cells. *Cell Rep.* **16**, 457–471 (2016).
22. Nabet, B. et al. Deregulation of the Ras-Erk signaling axis modulates the enhancer landscape. *Cell Rep.* **12**, 1300–1313 (2015).
23. Spangle, J. M. et al. PI3K/AKT signaling regulates H3K4 methylation in breast cancer. *Cell Rep.* **15**, 2692–2704 (2016).
24. Fan, Y. et al. A Phosphokinome-based screen uncovers new drug synergies for cancer driven by liver-specific gain of non-oncogenic RTKs. *Hepatology* **66**, 1644–1661 (2017).
25. Cancer Genome Atlas Research, N. et al. The cancer genome atlas Pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
26. Villanueva, A. et al. DNA methylation-based prognosis and epidrivers in hepatocellular carcinoma. *Hepatology* **61**, 1945–1956 (2015).
27. Lee, S. M. et al. HBx induces hypomethylation of distal intragenic CpG islands required for active expression of developmental regulators. *Proc. Natl Acad. Sci. USA* **111**, 9555–9560 (2014).
28. Zucman-Rossi, J., Villanueva, A., Nault, J. C. & Llovet, J. M. Genetic landscape and biomarkers of hepatocellular carcinoma. *Gastroenterology* **149**, 1226–1239 e1224 (2015).
29. Azad, N., Zahnow, C. A., Rudin, C. M. & Baylin, S. B. The future of epigenetic therapy in solid tumours--lessons from the past. *Nat. Rev. Clin. Oncol.* **10**, 256–266 (2013).
30. Brien, G. L., Valerio, D. G. & Armstrong, S. A. Exploiting the epigenome to control cancer-promoting gene-expression programs. *Cancer Cell* **29**, 464–476 (2016).
31. Bhasin, J. M. et al. Methylome-wide sequencing detects DNA hypermethylation distinguishing indolent from aggressive prostate cancer. *Cell Rep.* **13**, 2135–2146 (2015).
32. Kulis, M. et al. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat. Genet.* **44**, 1236–1242 (2012).
33. Lister, R. et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
34. Varley, K. E. et al. Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.* **23**, 555–567 (2013).
35. Maunakea, A. K. et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**, 253–257 (2010).
36. Hahn, M. A., Wu, X., Li, A. X., Hahn, T. & Pfeifer, G. P. Relationship between gene body DNA methylation and intragenic H3K9me3 and H3K36me3 chromatin marks. *PLoS ONE* **6**, e18844 (2011).
37. Karch, F. In vivo studies of the Drosophila insulator factor CTCF reach a Catch 22. *BMC Biol.* **13**, 71 (2015).
38. Shukla, S. et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74–79 (2011).
39. Hon, G. C. et al. Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.* **22**, 246–258 (2012).
40. Maina, F. Strategies to overcome drug resistance of receptor tyrosine kinaseaddicted cancer cells *Curr. Med. Chem.* **21**, 1607–1617 (2014).
41. Scott, M. T. et al. Epigenetic reprogramming sensitizes CML stem cells to combined EZH2 and tyrosine kinase inhibition. *Cancer Discov.* **6**, 1248–1257 (2016).
42. Genestine, M. et al. Enhanced neuronal Met signalling levels in ALS mice delay disease onset. *Cell Death Dis.* **2**, e130 (2011).
43. Fan, Y. et al. Tissue-specific gain of RTK signalling uncovers selective cell vulnerability during embryogenesis. *PLoS Genet.* **11**, e1005533 (2015).
44. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
45. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
46. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
47. Furlan, A. et al. Met acts through Abl to regulate p53 transcriptional outcomes and cell survival in the developing liver. *J. Hepatol.* **57**, 1292–1298 (2012).
48. Furlan, A. et al. Combined drug action of 2-phenylimidazo[2,1-b] benzothiazole derivatives on cancer cells according to their oncogenic molecular signatures. *PLoS ONE* **7**, e46738 (2012).
49. Furlan, A. et al. Abl interconnects oncogenic Met and p53 core pathways in cancer cells. *Cell Death Differ.* **18**, 1608–1616 (2011).
50. Wang, K. et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38**, e178 (2010).
51. Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26**, 493–500 (2010).

52. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
53. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).

Acknowledgements

These results are in part based upon public data generated by TCGA Research Network: http://cancergenome.nih.gov/. We are particularly grateful to F. Helmbacher for extremely valuable feedback on the study. We thank: all members of our labs for helpful discussions and comments; F. Castets for advises on biochemical analysis; J. Zucman-Rossi and E. Letouzé for valuable discussion and feedback on the bioinformatics; V. Girod-David and L. Jullien for excellent help with mouse husbandry; C. Giaccherini, A. Dobric and E. Chrabaszcz for their contributions to studies on molecular and functional characterisations of shRNA-targeted HCC cells. This work was funded by INCa (Institut National du Cancer), FdF (Fondation de France) and GEFLUC–Les Entreprises contre le Cancer to F.M. M.A. was supported by a FdF fellowship. A.Y. was supported by DFG grant HA 6905/2-1. S.K.B. was supported by the Higher Education Commission (HEC) of Pakistan–France Campus. The contribution of the Region Provence Alpes Côtes d'Azur and of the Aix-Marseille Université to the IBDM animal facility is also acknowledged. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

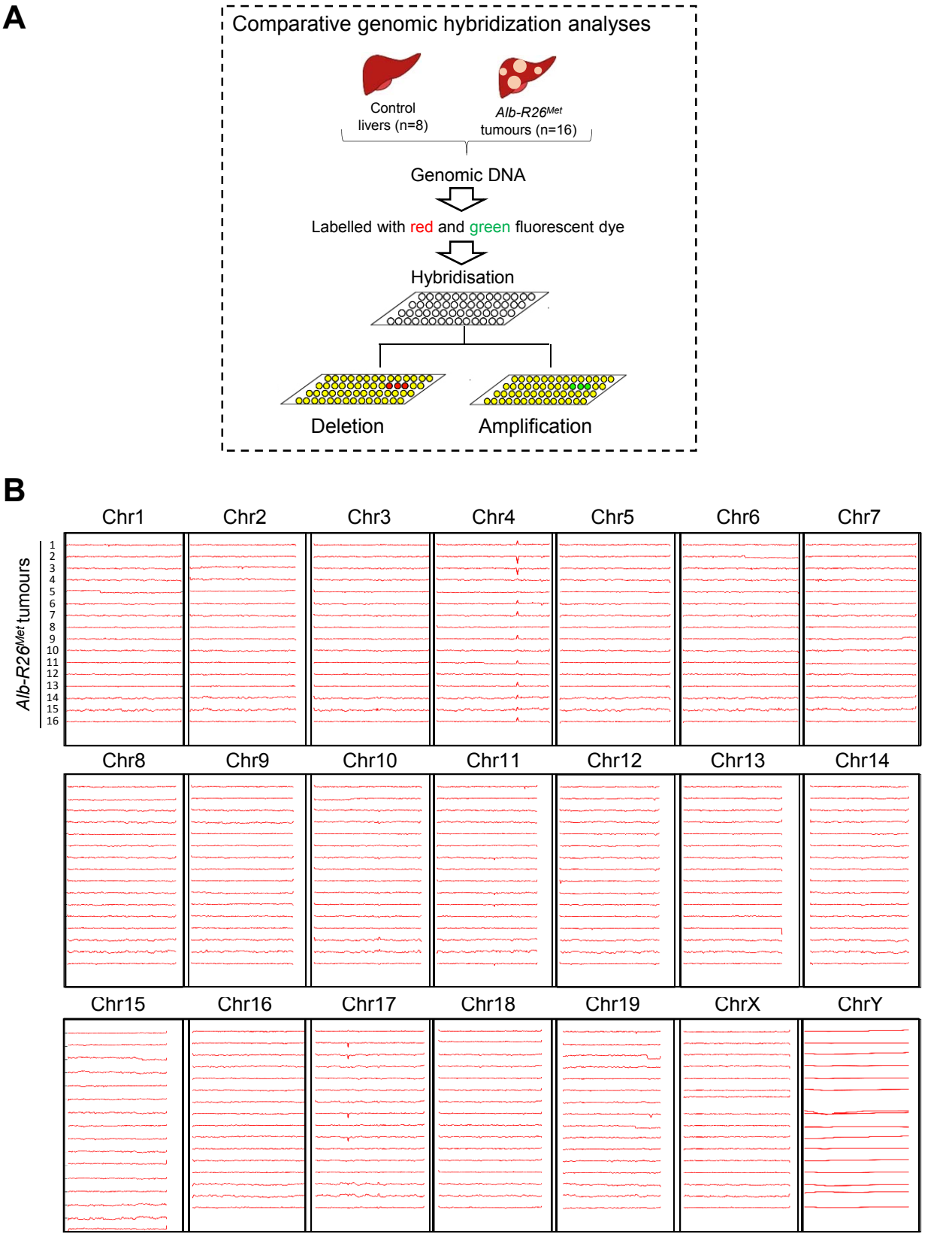
Author contributions

M.A.: performed the majority of the experiments, data analysis, interpretation, contributed to computational work and writing the paper. F.D.: performed the majority of the computational work, data analysis and interpretation. A.Y.: performed computational work using the human HCC cohort from TCGA database, data analysis and interpretation. S.K.B.: contributed to molecular and functional studies with HCC cells. S.R.: prepared tumour samples and genomic DNA for methylome; contributed to xenograft studies. R.D.: contributed to establishing the mouse model, to interpreting data and provided input on writing the paper. A.J.S.: provided support for computational work, contributed to interpreting data and provided input on writing the paper. B.H.H.: analysed and interpreted transcriptome data from mouse model, supervised studies on human HCC data and provided input on writing the paper. F.M.: designed the study, performed experiments, analysed and interpreted data, ensured financial support and wrote the paper.

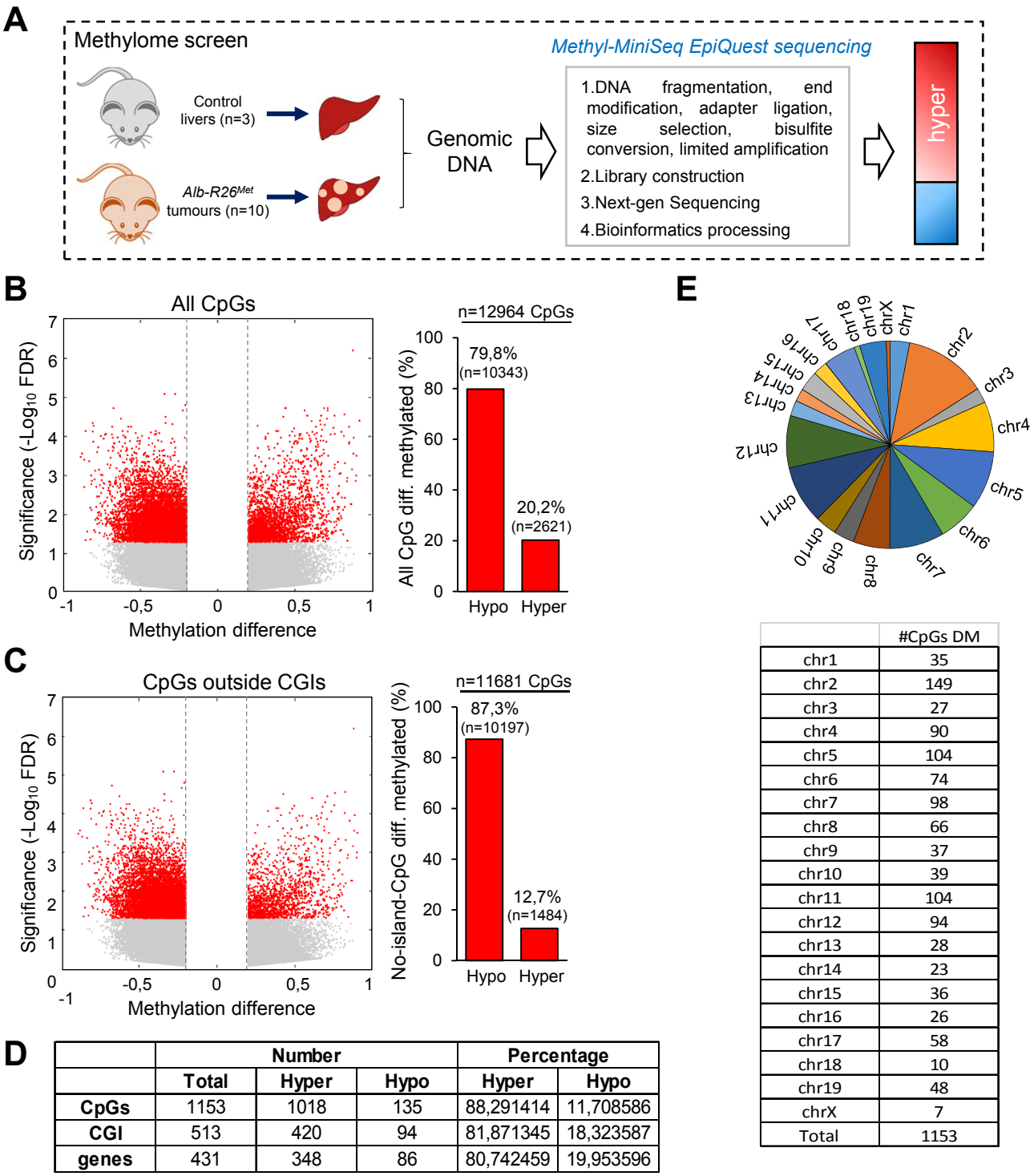
Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-05550-5.

Competing interests: The authors declare no competing interests.



Supplementary Figure 1. Tumorigenesis modelled by the *Alb-R26^{Met}* mice is characterized by a stable chromosome context. (A) Schematic representation of the experimental setting used to analyse genomic DNA from *Alb-R26^{Met}* HCC (n=16) and control livers (n=8) through Comparative Genomic Hybridization (CGH) Microarray. (B) Overall picture of the CGH results in the 19 autosomal and the 2 sex chromosome pairs of the analysed *Alb-R26^{Met}* tumours. Note that tumours develop in a stable chromosome context; only 2 regions in chromosome 4 and chromosome 17 showed some variations in a proportion of analysed tumours.

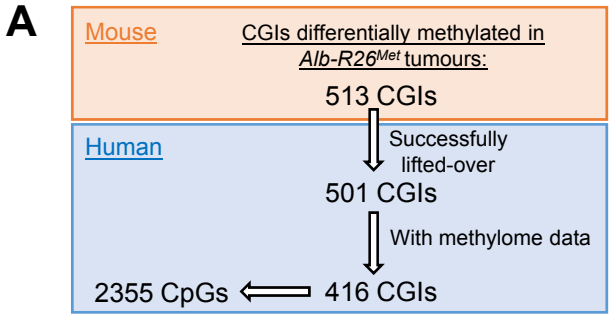


	#CpGs DM
chr1	35
chr2	149
chr3	27
chr4	90
chr5	104
chr6	74
chr7	98
chr8	66
chr9	37
chr10	39
chr11	104
chr12	94
chr13	28
chr14	23
chr15	36
chr16	26
chr17	58
chr18	10
chr19	48
chrX	7
Total	1153

D

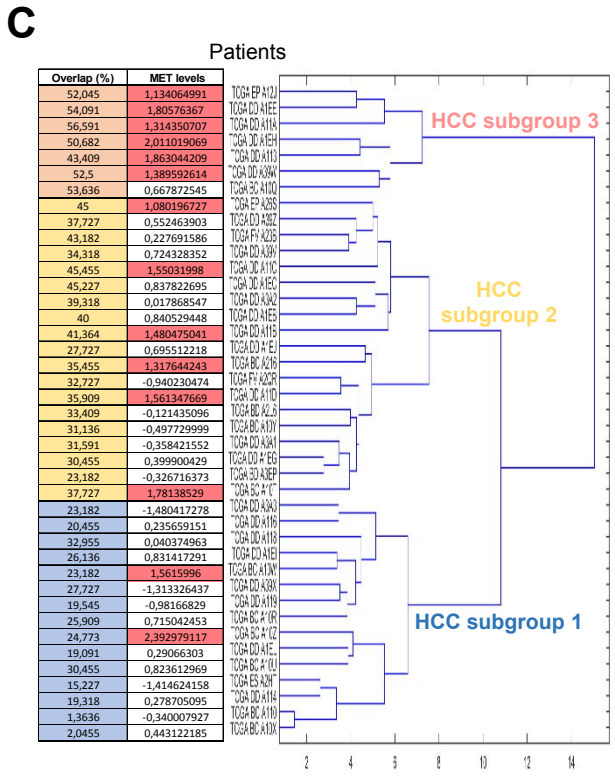
	Number			Percentage	
	Total	**Hyper**	**Hypo**	**Hyper**	**Hypo**
CpGs	1153	1018	135	88,291414	11,708586
CGI	513	420	94	81,871345	18,323587
genes	431	348	86	80,742459	19,953596

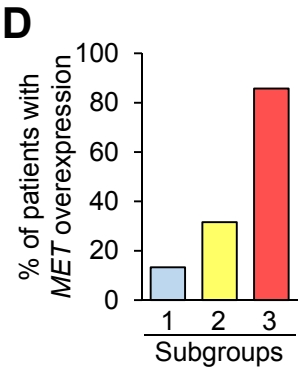
Supplementary Figure 2. Focal CGI hypermethylation and widespread CpG hypomethylation in *Alb-R26^{Met}* tumours. (A) Schematic representation of experimental settings employed for the methylome screen. Genomic DNA from *Alb-R26^{Met}* HCC (n=10) and controls (n=3) was used to examine the DNA methylation status of a CGI-enriched fraction and raw data were used for bioinformatics processing. (B and C) Volcano plot reporting methylation differences with significance (expressed as -Log$_{10}$ FDR) for all measured CpGs (B) and CpGs located outside CGIs (C) in *Alb-R26^{Met}* tumours versus control (left). Significant differences (methylation difference>0.2 and FDR<0.05) are shown in red. Graph reporting the percentage (and numbers) of hypomethylated versus hypermethylated CpGs (right). (D) Table reporting the number and percentage of CpGs, the corresponding CGIs and genes, differentially methylated, specifying the ones hypermethylated versus those hypomethylated. (E) Chromosomic distribution of the 1153 CpGs differentially methylated in *Alb-R26^{Met}* tumours compared to control livers.

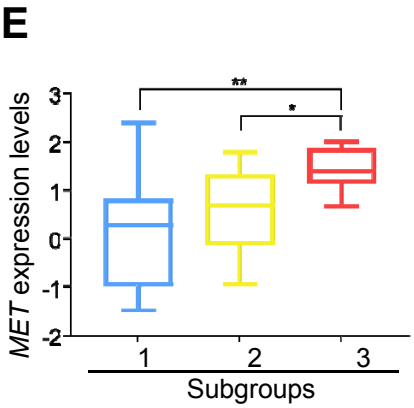


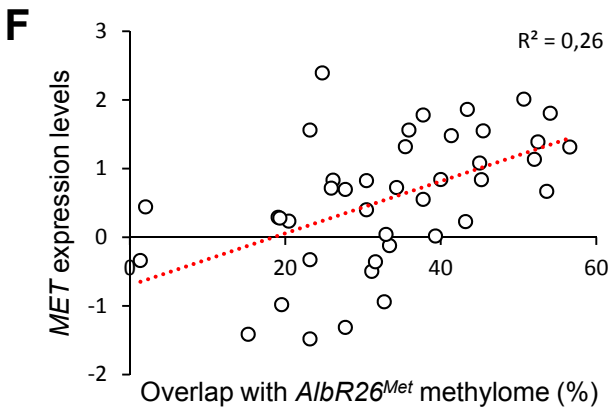
B

	CpGs	CGIs	genes
Diff. Methyl	241	105	94
Hyper-	225	96	85
Hypo-	16	9	9







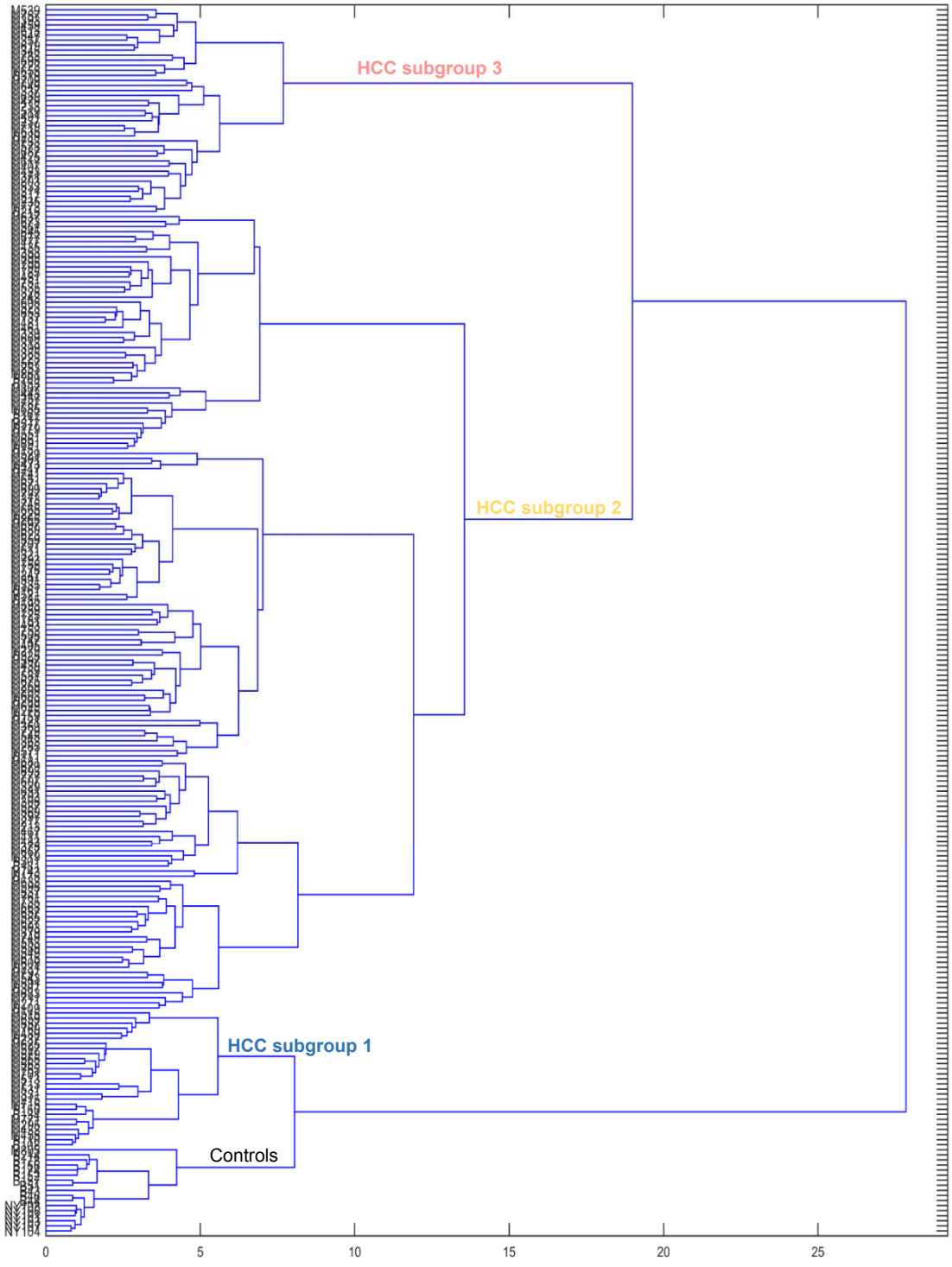


Supplementary Figure 3. Comparison of methylome outcomes of *Alb-R26^{Met}* tumours with those of the TCGA cohort (41 HCC patients). (A) Schematic representation of the mouse-human CGI lift-over performed to extract the human CGIs corresponding to the 513 mouse CGIs differentially methylated in *Alb-R26^{Met}* tumours. 501 CGIs were successfully mapped in human, and methylome data were available for 416 CGIs. (B) Table reporting the number and percentage of CpGs, the corresponding CGIs and genes, differentially methylated, specifying the ones hypermethylated versus those hypomethylated. Data refer to those reported in Figure 1E. (C) Hierarchical clustering of HCC patients based on the 416 CGIs differentially methylated in *Alb-R26^{Met}* tumours. Data refer to those reported in Figure 1F, specifying the patient ID, the percentage overlap, and the *MET* expression levels. (D) Graph showing the percentage of patients with *MET* overexpression in the three HCC subgroups reported in Figure 1F. Note that *MET* is overexpressed in 86% (6/7) of HCC patients belonging to the HCC subgroup 3 (the subgroup that best overlap with CGI methylation changes in *Alb-R26^{Met}*), 32% (6/19) to the HCC subgroup 2, and only 13% (2/15) to the HCC subgroup 1. (E) Graph reporting *MET* expression levels in the three HCC subgroups reported in Figure 1F. (F) For the 41 HCC patients, correlation between *MET* expression levels and percentage of overlap with the CGIs found differentially methylated in the *Alb-R26^{Met}*genetic setting.

A

	CpGs	CGIs	genes
Diff. Methyl	269	110	100
Hyper-	255	103	93
Hypo-	14	7	7

B

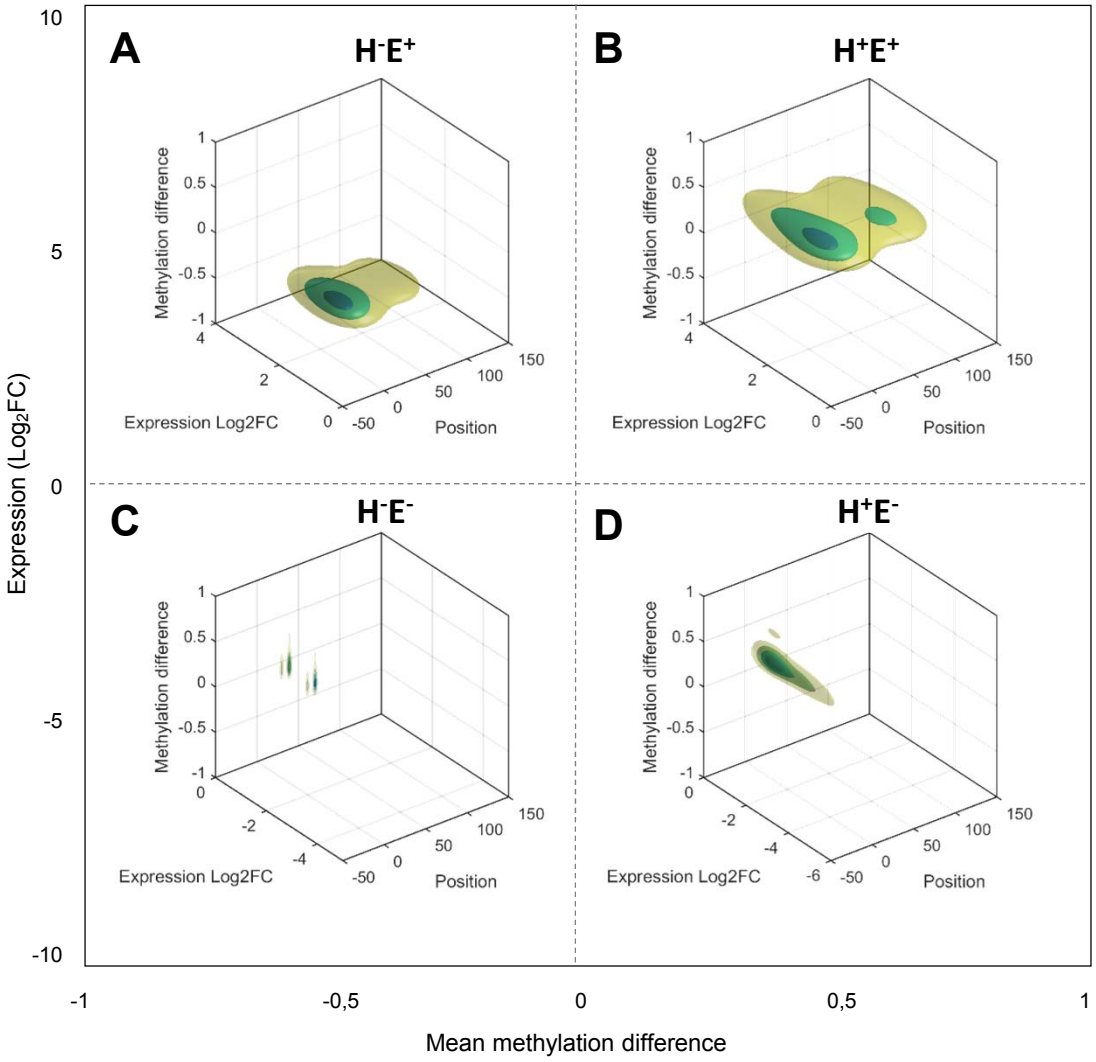


Supplementary Figure 4. Comparison of methylome outcomes of *Alb-R26^{Met}* tumours with those of a cohort of 234 human samples (from GSE56588; 224 HCC patients and 10 control individuals). (A) Table reporting the number and percentage of CpGs, the corresponding CGIs and genes, differentially methylated, specifying the ones hypermethylated versus those hypomethylated. Data refer to those reported in Figure 1G. (B) Hierarchical clustering of HCC patients based on the 416 CGIs differentially methylated in *Alb-R26^{Met}* tumours. Data refer to those reported in Figure 1H.

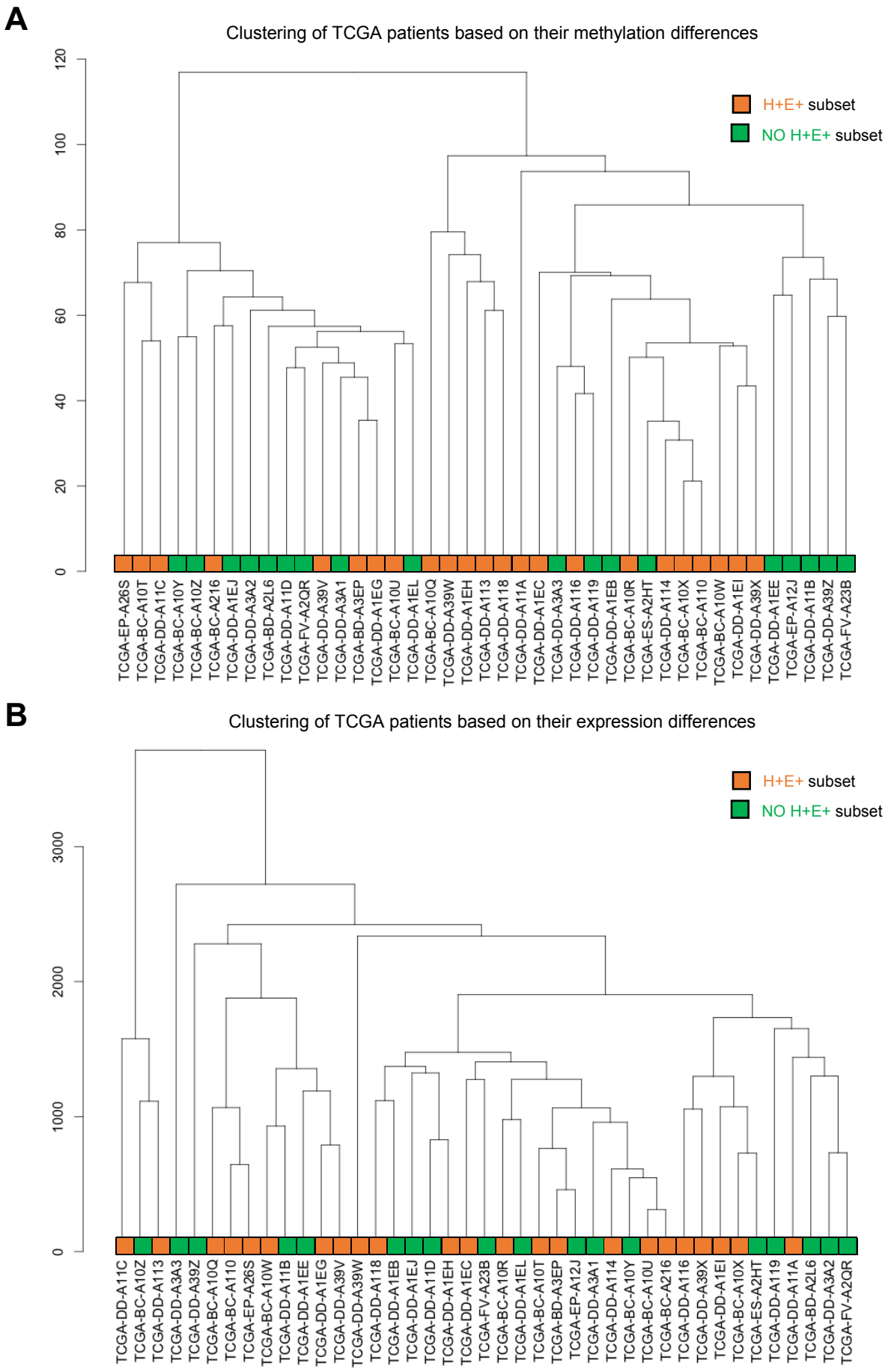
Pathway enrichment analysis: genes differentially methylated and differentially expressed in *Alb-R26^{Met}* tumours

Pathway	p value
MAPK signaling pathway - Homo sapiens (human)	0.00015777
HTLV-I infection - Homo sapiens (human)	0.00016107
Axon guidance - Homo sapiens (human)	0.0002355
Nicotine addiction - Homo sapiens (human)	0.00028279
Viral carcinogenesis - Homo sapiens (human)	0.00043804
Pathways in cancer - Homo sapiens (human)	0.0014551
Neuroactive ligand-receptor interaction - Homo sapiens (human)	0.0017757
TGF-beta signaling pathway - Homo sapiens (human)	0.0023777
GABAergic synapse - Homo sapiens (human)	0.0027043
Morphine addiction - Homo sapiens (human)	0.002966
Cell cycle - Homo sapiens (human)	0.0068301
Osteoclast differentiation - Homo sapiens (human)	0.007425
Type II diabetes mellitus - Homo sapiens (human)	0.0092091
Tight junction - Homo sapiens (human)	0.015343
Renal cell carcinoma - Homo sapiens (human)	0.017431
Calcium signaling pathway - Homo sapiens (human)	0.018154
Epithelial cell signaling in Helicobacter pylori infection - Homo sapiens (human)	0.01891
Melanoma - Homo sapiens (human)	0.019414
Adherens junction - Homo sapiens (human)	0.020952
cAMP signaling pathway - Homo sapiens (human)	0.022263
Focal adhesion - Homo sapiens (human)	0.022533
Epstein-Barr virus infection - Homo sapiens (human)	0.023628
Rap1 signaling pathway - Homo sapiens (human)	0.025603
Protein digestion and absorption - Homo sapiens (human)	0.031017
NF-kappa B signaling pathway - Homo sapiens (human)	0.034037
Endocrine resistance - Homo sapiens (human)	0.034652
Amoebiasis - Homo sapiens (human)	0.034652
Estrogen signaling pathway - Homo sapiens (human)	0.035891
Chagas disease (American trypanosomiasis) - Homo sapiens (human)	0.038408
Leukocyte transendothelial migration - Homo sapiens (human)	0.044908
Cholinergic synapse - Homo sapiens (human)	0.044908
Cytokine-cytokine receptor interaction - Homo sapiens (human)	0.045221
Serotonergic synapse - Homo sapiens (human)	0.045573

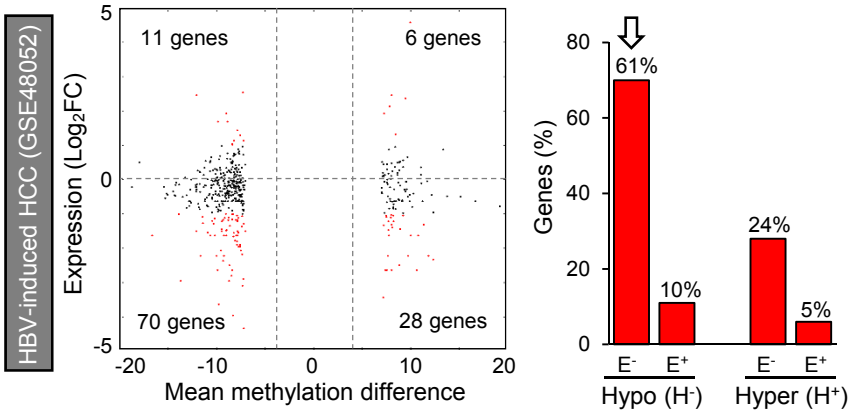
Supplementary Figure 5. KEGG pathway enrichment analysis for genes with changes in CGI methylation and expression in *Alb-R26^{Met}* tumours, ranked according to their p-value. Data reported in Figure 3A are in pink.



Supplementary Figure 6. 3D density plots showing the distribution of genes according to their relative position to the ATG (as percentage), gene expression level (as Log_2 FC), and CGI methylation difference (as β-value). Genes: overexpressed with hypomethylated CGI (A); overexpressed with hypermethylated CGI (B); downregulated with hypomethylated CGI (C); downregulated with hypermethylated CGI (D).



Supplementary Figure 7. Hierarchical clustering analysis of the TCGA cohort (41 HCC patients) based on either global DNA methylome (A) or transcriptome (B) outcomes. In orange: H^+E^+ patient subset. In green: "NO H^+E^+" patient subset.

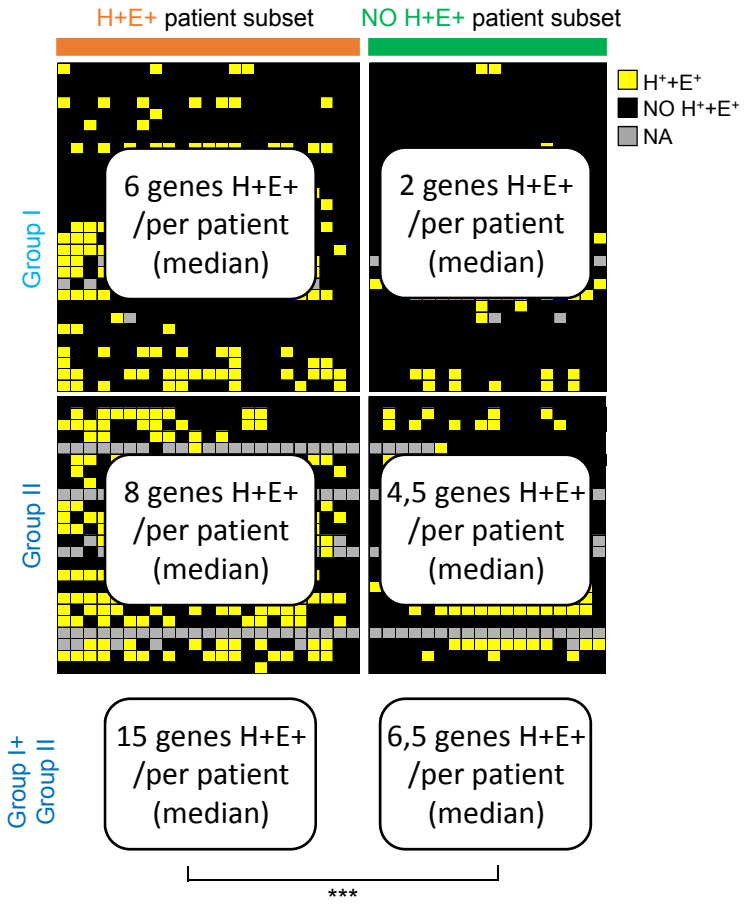


Supplementary Figure 8. Mouse *HBx^{tg}* tumours are characterized by an enrichment in genes downregulated and with hypomethylated CGIs. Left: Methylation differences versus expression for all genes with CGIs hypermethylated (H^+) or hypomethylated (H^-) in *HBx^{gt}* tumours. Expression values are relative to controls. Dots correspond to single differentially methylated CpG and the corresponding gene expression (genes which expression is significantly below or above Log_2 fold change (FC) ±1 are indicated in red). Right: Graph reporting the percentage of downregulated (E^-) and upregulated (E^+) genes among those with a hypomethylated (H^-) or hypermethylated (H^+) CGI. Note the enrichment of genes downregulated and with hypomethylated CGIs in *HBx^{gt}* tumours (indicated by an arrow), in contrast to the enrichment of gene overexpressed and with hypermethylated CGIs in *Alb-R26^{Met}* tumours.

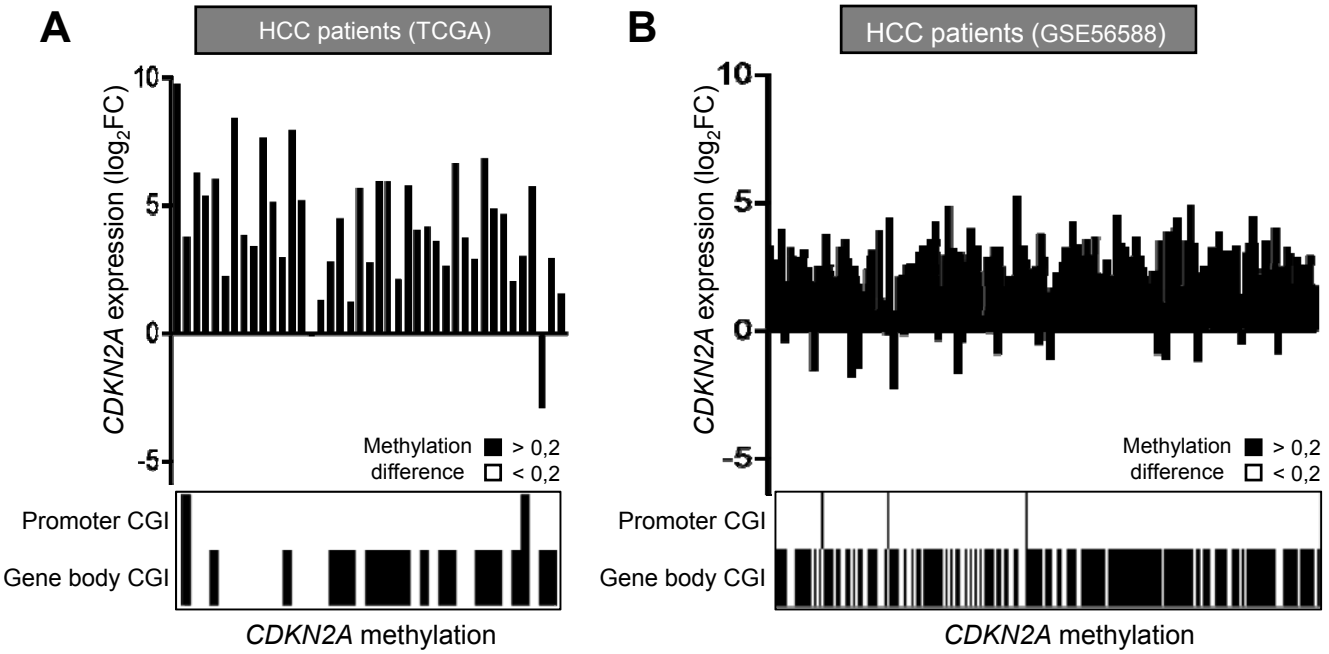
	Patients	Alcohol	NAFLD	HepatitisB	HepatitisC	Hemochromatosis	TOTAL
H^+E^+ patient subset	TCGA-BC-A10Q	0	0	0	0	0	0
	TCGA-DD-A1EH	0	0	1	0	0	1
	TCGA-DD-A1EC	0	0	0	0	0	0
	TCGA-DD-A113	0	0	0	0	0	0
	TCGA-DD-A11A	0	0	0	0	0	0
	TCGA-DD-A39W	0	0	0	0	0	0
	TCGA-BC-A216	0	0	0	0	0	0
	TCGA-EP-A26S	1	0	0	0	0	1
	TCGA-DD-A11C	0	0	0	0	0	0
	TCGA-BC-A10T	0	0	0	0	0	0
	TCGA-DD-A118	0	0	0	0	0	0
	TCGA-DD-A39X	0	0	0	0	0	0
	TCGA-DD-A1EI	0	0	1	0	0	1
	TCGA-BC-A10R	0	0	0	0	0	0
	TCGA-BC-A10W	0	0	1	0	0	1
	TCGA-BC-A10U	1	0	0	0	0	1
	TCGA-DD-A39V	0	0	0	0	0	0
	TCGA-BD-A3EP	0	0	0	0	0	0
	TCGA-DD-A1EG	1	1	0	0	0	2
	TCGA-DD-A114	0	0	0	1	0	1
	TCGA-DD-A116	0	0	1	0	0	1
	TCGA-BC-A110	1	0	0	0	0	1
	TCGA-BC-A10X	0	0	0	0	0	0
NO H+E+ patient subset	TCGA-BD-A2L6	1	0	0	0	0	1
	TCGA-BC-A10Z	0	0	0	0	0	0
	TCGA-BC-A10Y	1	0	0	0	0	1
	TCGA-FV-A2QR	0	0	0	0	0	0
	TCGA-DD-A1EJ	0	0	0	0	0	0
	TCGA-DD-A1EL	0	0	1	0	0	1
	TCGA-DD-A11D	0	1	0	0	0	1
	TCGA-DD-A3A1	0	0	0	0	0	0
	TCGA-DD-A1EE	0	0	0	1	0	1
	TCGA-EP-A12J	0	0	0	1	0	1
	TCGA-FV-A23B	0	0	0	1	0	1
	TCGA-DD-A11B	0	1	0	0	0	1
	TCGA-DD-A3A3	0	0	0	0	0	0
	TCGA-DD-A1EB	0	0	0	0	0	0
	TCGA-DD-A3A2	0	0	0	0	0	0
	TCGA-DD-A39Z	1	0	0	0	0	1
	TCGA-DD-A119	0	0	1	0	0	1
	TCGA-ES-A2HT	0	0	0	1	0	1

Fisher's exact test
p-value=0,2146

Supplementary Figure 9. H^+E^+ patients are not characterised by any specific risk factors, while show a trend of better prognosis. Table reporting risk factors associated to individual HCC patients. The presence of a risk factor is indicated as 1, whereas the absence by 0. No significant differences were found between H^+E^+ versus "NO H^+E^+" patient subsets.



Supplementary Figure 10. Among the 55 genes identified in *Alb-R26^{Met}* tumours, a significant higher number of them is both hypermethylated and overexpressed in the H^{+}E^{+} patient subset compared to the "NO H^{+}E^{+}" subset. Heat-map showed in Figure 6D in which the number of genes both hypermethylated and overexpressed per patient (median) is reported for each subgroup. Significance is indicated on the bottom. ***: P<0.001.

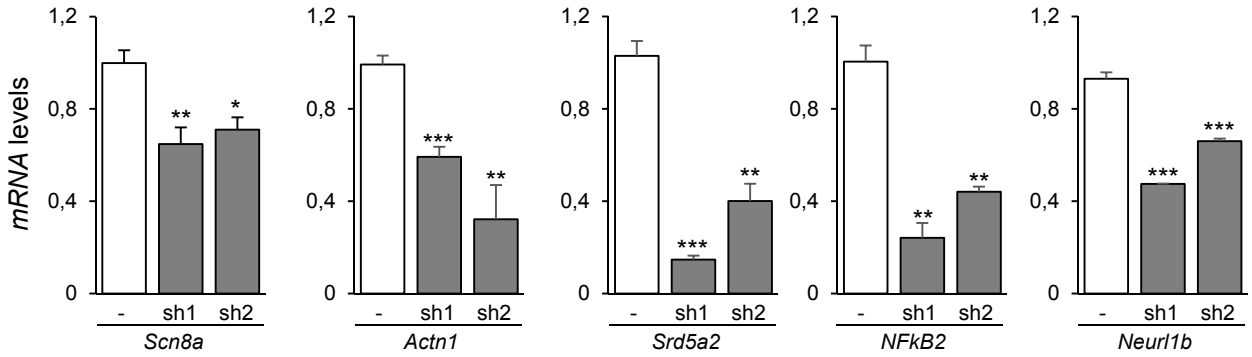


Supplementary Figure 11. The *CDKN2A* gene is overexpressed and with hypermethylated gene body CGI in the majority of HCC patients. Top: graphs reporting the expression levels of *CDKN2A* in HCC patients from the TCGA (A) and GSE56588 (B) cohorts. Bottom: schemes reporting a black line when the *CDKN2A* promoter or gene body CGI is hypermethylated (methylation difference >0.2) in HCC patients from the TCGA (A) and GSE56588 (B) cohorts. Notably, in both cohorts the majority of HCC patients carry an overexpression of *CDKN2A* (39/41 and 166/204, in the respective cohorts), which is associated with a hypermethylation of the gene body CGI (21/39 and 163/166, in the respective cohorts). In contrast, not methylation changes are detected in the promoter CGI for both HCC cohorts.

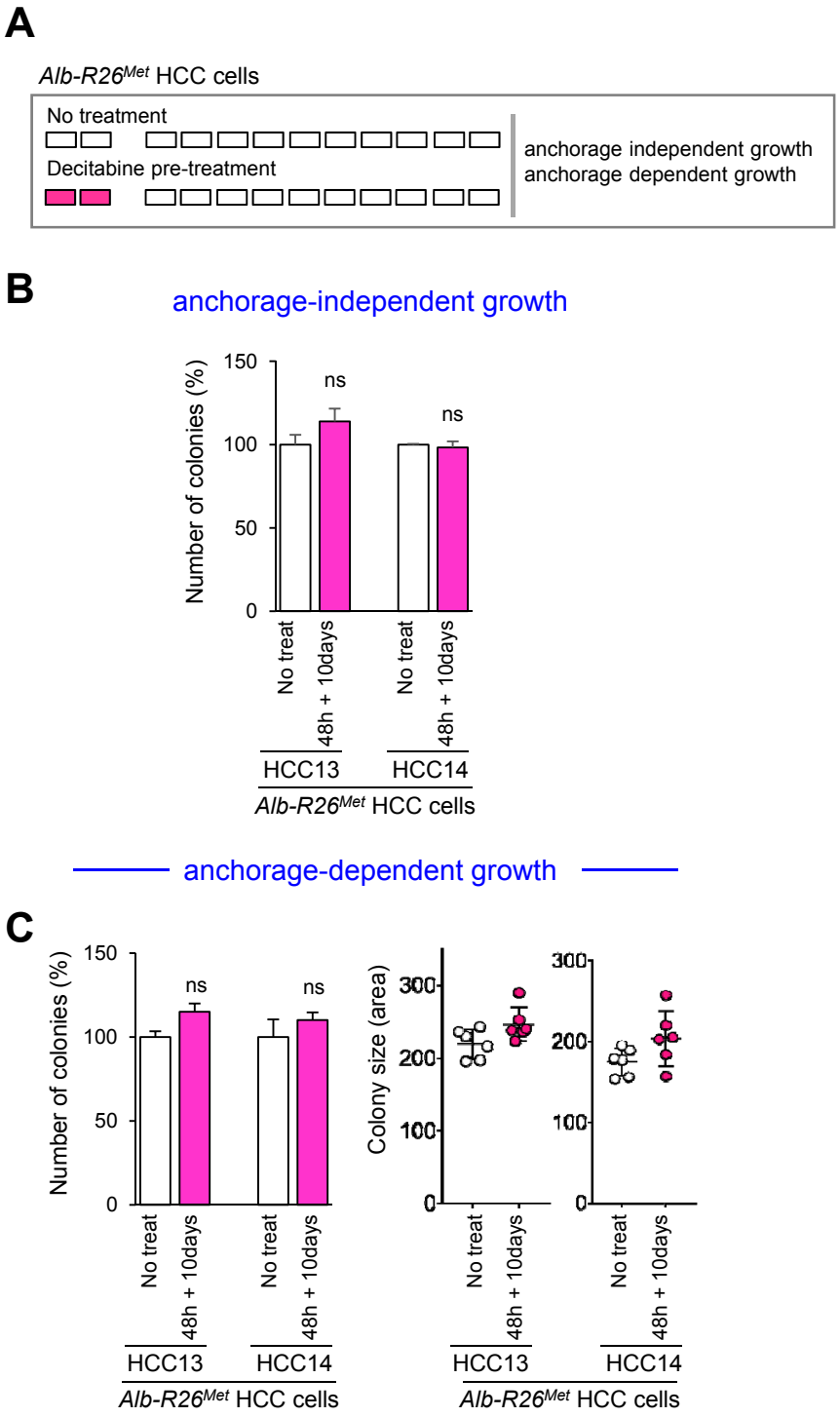
Pathway enrichment analysis: overexpressed genes with hypermethylated CGI in *Alb-R26*^{Met} tumours

Pathway	p value
MAPK signaling pathway - Homo sapiens (human)	8.2428e-06
Viral carcinogenesis - Homo sapiens (human)	3.7666e-05
HTLV-I infection - Homo sapiens (human)	0.000116
Pathways in cancer - Homo sapiens (human)	0.00080777
Cell cycle - Homo sapiens (human)	0.0016141
Osteoclast differentiation - Homo sapiens (human)	0.001764
Nicotine addiction - Homo sapiens (human)	0.0025759
Tight junction - Homo sapiens (human)	0.0038538
Axon guidance - Homo sapiens (human)	0.0041683
cAMP signaling pathway - Homo sapiens (human)	0.0058036
Focal adhesion - Homo sapiens (human)	0.0058818
Epstein-Barr virus infection - Homo sapiens (human)	0.0062006
Renal cell carcinoma - Homo sapiens (human)	0.0065617
Epithelial cell signaling in Helicobacter pylori infection - Homo sapiens (human)	0.0071469
Melanoma - Homo sapiens (human)	0.0073467
Adherens junction - Homo sapiens (human)	0.0079605
TGF-beta signaling pathway - Homo sapiens (human)	0.010621
GABAergic synapse - Homo sapiens (human)	0.011577
Morphine addiction - Homo sapiens (human)	0.012317
NF-kappa B signaling pathway - Homo sapiens (human)	0.013332
Endocrine resistance - Homo sapiens (human)	0.01359
Estrogen signaling pathway - Homo sapiens (human)	0.014114
Leukocyte transendothelial migration - Homo sapiens (human)	0.017989
Cholinergic synapse - Homo sapiens (human)	0.017989
Wnt signaling pathway - Homo sapiens (human)	0.027757
Breast cancer - Homo sapiens (human)	0.028096
Maturity onset diabetes of the young - Homo sapiens (human)	0.047146

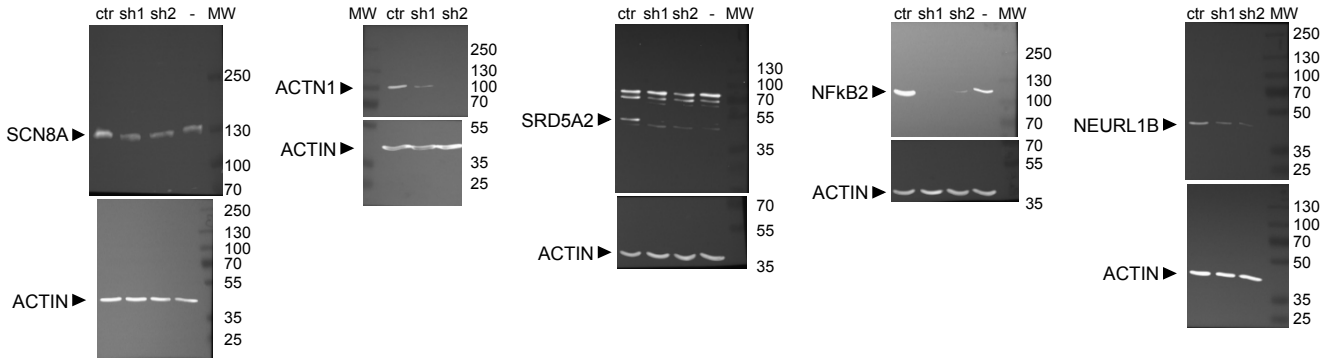
Supplementary Figure 12. KEGG pathway enrichment analysis for overexpressed genes with hypermethylated CGI in *Alb-R26*^{Met} tumours, ranked according to their p-value. Data reported in Figure 6F are in blue.



Supplementary Figure 13. Expression levels of *Scn8a*, *Actn1*, *Srd5a*, *NFkB2*, and *Neurl1b* in *Alb-R26^{Met}* HCC cells transfected with two different shRNA targeting sequences versus controls. After molecular validation, cells were used for in vitro studies reported in Figure 7.



Supplementary Figure 14. Global CGI hypermethylation is functionally relevant for *Alb-R26^{Met}* tumorigenesis. (A) Scheme reporting demethylating treatment (Decitabine; 0.3µM) used for in vitro experiments with *Alb-R26^{Met}* HCC cells. Cells were pre-treated (48 h) with Decitabine, then cultured for 10 days without any treatment before using them for experiments. (B, C) Anchorage-independent (B) and anchorage-depend (C) growth assays using 2 different *Alb-R26^{Met}* HCC cell lines (HCC13 and 14) showing effects of demethylating treatments described in A. Note that HCC cells recover their tumorigenic properties when experiments are performed with cell cultured 10 days after Decitabine pre-treatment. Significant differences between groups are indicated on the top. Not significant (ns).



Supplementary Figure 15. Full blots of gels in which the acquisition of ECL signal performed using the MyECL imager system (without negative conversion) was merged with a picture of the membranes. The corresponding molecular weights, visible on the membranes, are indicated. These results are reported in Figure 7A.

7 Discussion

Mitochondria and their functions have first been studied mostly in the context of energy production (Davis & Williams, 2012). In recent years, there is a growing interest in the extensive roles of mitochondria in other fields such as signaling pathway control, balance of metabolites, and anti-oxidant defense. The dysfunction of mitochondrial processes and pathways is also implicated in cancer and various diseases such as neurodegenerative, cardiovascular and metabolic disorders. While mitochondrial physiology and pathology have attracted increased attention, platforms or tools that are exclusive for the exploration of expression and mutation landscapes of mitochondrial genes were previously non-existent. To this end, a data mining and visualization platform specific for mitochondrial genes has been developed as the central objective of this thesis to allow interested users to study variations in mitochondrial genes and processes under different conditions with publicly available or their own -omics data.

7.1 Visual Data Mining as a tool for the exploration of -omics data

The recent advancement of sequencing technologies brings about high volume of -omics data. There are currently a lot of tools and pipelines that perform analysis on such data from quality check, mapping to differential expression analysis and mutation calling (Roumpeka et al., 2017), but most of them return results as tabular data, which is not always easy to interpret. Visualization – the transformation of raw numeric data into illustration, is a robust way to help the audience understand the data and reveal underlying patterns. In this thesis, **mitoXplorer** (**Section 6.1**) and the mitochondrial genes workflow in **CancerSysDB** (**Section 6.2**) were introduced as visual data mining platforms that allow users to visualize and analyze -omics data regarding mutations and expressions of mitochondrial genes.

7.1.1 Interactive visualization aids the discovery of underlying patterns

Both mitoXplorer and CancerSysDB provide not only static visualization of analyzed -omics data, but also a set of dynamic, interactive and intuitive visualizations, which gives users the liberty to explore the data in their own ways. For example, in the **Comparative Plot** and **Hierarchical Clustering** analysis on mitoXplorer, an interactive heatmap was used to display Log_2 fold-change (Log2FC) level of mitochondrial genes for each dataset. Unlike the static heatmap visualizations (or those with limited interactivity) that are available on many bioinformatics analytic platforms (cBioPortal (Cerami et al., 2012), UCSC (Zhu et al.,

2009)), the interactive heatmap on mitoXplorer supports sorting on differential expression level by sample or by gene, which makes it easier to do comparison within the matrix. Although being a very popular visualization to represent -omics data due to its flexibility and easiness to explore patterns (cBioPortal, UCSC, Cancer Genome Workbench (Jinghui Zhang et al., 2007), Caleydo (Streit et al., 2009), Gitools (Perez-Llamas & Lopez-Bigas, 2011)), heatmaps are not able to show the extra features of interested genes (e.g. the location of genes, mutation data) (Schroeder et al., 2013). mitoXplorer overcomes this problem by displaying functional annotation and additional information on the side panel and highlighting mutated genes on the scatter plot on another panel for the Comparative plot, which helps to discover relations between data and generate relevant hypotheses.

The categorical scatter plot (or beeswarm plot) in the **Comparative Plot** of mitoXplorer and Interactive Workflow of CancerSysDB is another visualization for the differential expression of genes. Such a scatter plot is different from the ones seen in other platforms that simply shows the relationship between two features (e.g. expression of two genes, number of mutation vs fraction of altered genome, etc) (cBioPortal, Caleydo, tranSMART (https://transmart-app.readthedocs.io/)) as it aims to facilitate the comparison of (up to six) samples (first dimension) by displaying the distribution of Log2FC values (second dimension) with non-overlapping points. Moreover, it includes mutation data as a third dimension using different color and sizes. To study a larger number of samples with additional data (e.g. clinical data), users could make use of the **Principle Component Analysis** (PCA) function in mitoXplorer. PCA, while being very useful when working with high-dimensional data and being able to potentially reveal clusters within samples, is seldom implemented in bioinformatics visualization platforms, probably due to the requirement of relatively high of computational power. In mitoXplorer, only a subset of genes related to mitochondrial functions are included. This makes such an analysis possible. Also, compared to the PCA visualization on other platforms (tranSMART), the one on mitoXplorer has much more interactivity and allows users to filter and color samples by user-defined groups, or groupings by clinical data or demographic data (if available), which makes it a very informative visualization and could potentially help to identify previously unknown subgroups.

Since the functional relationships between genes or proteins are difficult to be represented in a heatmap or scatter plot, a bubble chart is adopted as the **Interactome View** of mitoXplorer. Similar to force-directed visualizations on other platforms or tools (cBioPortal, Cytoscape), users could examine the expression and mutations of genes in a dataset, while viewing the connectivity information (usually from external sources) among them. However in the case of mitoXplorer, the genes, nodes and edges of the network that are visualized

are defined by the **mitochondrial interactomes** which are manually curated (as part of the work of this thesis) and it does not need any input from users.

7.1.2 The mitochondrial interactomes

One thing that makes mitoXplorer and CancerSysDB stand out from other bioinformatics analysis/visualization platforms is the manually curated mitochondrial interactomes with accurate and updated annotations, which are integrated into the platforms for the analysis of transcriptome, proteome and mutation data of a specific subset of genes. Although there are existing lists or electronic repositories of mitochondrial genes from different proteomic studies or genome-scale prediction of mito-proteins (MitoCarta (Calvo et al., 2016), MitoMiner (Smith et al., 2012), MitoRes (Catalano et al., 2006), MitoPred (C. Guda et al., 2004)), none of them is sufficient to be taken directly for the purpose of in-depth analysis of mitochondrial interactome due to various reasons. Only MitoCarta, which has recently released a new version of its mitochondrial protein content for humans only, has a comparable complete set of mitochondria-associated genes that are grouped into mitochondrial processes.

Studies that use computational approaches or machine learning for the construction of mitochondrial gene repositories are either susceptible to overfitting of the training data (Support Vector Machine used by MitoMiner), or requires certain assumptions (conditional independence for Naïve Bayes method used by Mitocarta); or simply lacks experimental confirmation (MitoPred). On the other hand, proteomic studies that adopted MS-based approaches can suffer from a high false-positive rate (Pagliarini et al., 2008). Therefore, the mitochondrial interactomes described in this thesis have been manually curated by referencing extensive literature and databases (UniProt, NCBI, Flybase, SGD, GeneCards) in order to be as comprehensive as possible. All interactomes have been examined meticulously and referenced to experimental evidence to avoid including false-positives. Orthologs across species are moreover included in the species-specific mito-interactomes to provide consistency.

Most importantly, the genes are grouped in mitochondrial processes and a set of controlled vocabulary was used for functional annotation, which is lacking in most of the available repositories. Such annotation of the interactome facilitates meaningful analysis and visualization of mitochondrial genes expression dynamics when comparing differential expression of various conditions. These carefully curated interactomes make up an important component of mitoXplorer and CancerSysDB as mitochondrial genes-specific visual data mining platforms, and their integration into both platforms with rich visualizations

has allowed users to explore the expression and mutation landscape of mitochondrial genes under different conditions, which, to the best of my knowledge, has not been achieved in any other available tool so far.

Finally, mitoXplorer and CancerSysDB are implemented as web platforms, which do not require installation nor depend on certain operating system like stand-alone software or software packages (Caleydo, tranSMART, Cytoscape, IGV, bigPint (Rutter & Cook, 2020), Igloo-plot (Kuntal et al., 2014)). They also come with user-friendly interfaces, and do not assume users to possess any programming knowledge. Users can choose to either use the publicly available data or upload their own data for analysis, and can easily download and share their results. Whereas some other web platforms only allow users to analyze pre-calculated data within their own database and do not accept user-provided dataset (Navigator (Brown et al., 2009)). In the first two sections of Results, it is demonstrated how such kind of visual data mining tools could help identify mitochondrial genes or proteins in deregulated pathways that might potentially lead to different pathogenic conditions.

7.2 Dynamics of mitochondrial genes expression in mitochondria-associated disease and aneuploidy conditions

In order to demonstrate the analytical and predictive power of mitoXplorer (**Section 6.1**) as a visual data mining tool, the transcriptome data from several conditions associated with mitochondrial functions were explored. The data from a mouse model of Barth syndrome, an X-linked mitochondria-associated disease characterized by cardiomyopathy, was first analyzed on mitoXplorer. Next, a set of trisomy 21 data was studied with the analytical and visualization tools on mitoXplorer. These predictions were then verified experimentally by our collaborators.

7.2.1 Intuitive visualization with functional annotations helps to reveal impaired pathways in mitochondrial diseases

Barth syndrome results from a disturbed metabolism of cardiolipin due to mutations of the Taz gene, which causes mitochondrial defects as cardiolipin is a phospholipid that composes the inner membrane of mitochondria (Bione et al., 1996). It has been shown in a previous study (Chowdhury et al., 2018) that Tafazzin-deficient mouse embryonic fibroblasts (MEFs) displayed reduced ROS level under hypoxia condition, which impaired the activation of NF-κB pathway and hence decreased the Hif-1α expression. Using the **heatmap visualization** that is integrated with the manually curated mito-interactome on mitoXplorer, it could be confirmed that the induction of RelA (transcription factor of NF-κB) and Hif-1α was indeed diminished. Moreover, it was observed that Yap1 (yes-associated

protein 1), a transcription factor that also belongs to the same mito-process “Transcription” as RelA and Hif-1α, was also down-regulated, which has lead to the speculation that the impaired induction of the Hippo pathway protein Yap1 could contribute to the phenotype observed in Tafazzin-deficient MEFs by destabilizing Hif-1α.

Due to the unique feature of the heatmap visualization on mitoXplorer that it groups and displays a set of selected mitochondrial genes of the same mito-process, the association between Yap1, RelA and Hif-1α in the role of NF-κB activation could be readily revealed from the visualization. Whereas on other platforms, either all genes are displayed in the same heatmap (cBioPortal), which does not provide any information on functional classifications; or the genes to be visualized have to be provided by users as a list (UCSC), which relies on users’ prior knowledge on the gene of interests.

The sorting function further facilitates the identification of closely associated genes in terms of expression profile, as it allows users to quickly recognize genes (of the same function) with similar expression levels under different conditions. Although such a function is also available in the heatmap of some other platform (GiTools), the sorted heatmap would not be able to give any meaningful insights when the displayed genes are not organized in proper groupings.

7.2.2 Deregulation of mitochondrial transcriptome and proteome in trisomy 21 conditions identified by mitoXplorer

As discussed in **Section 4.1.5.2 (Aneuploidy)**, mitochondrial dysfunction has been observed in patients with trisomy 21 (T21) condition, where oxidative stress could potentially cause some of the clinical features of Down Syndrome (Jovanovic et al., 1998). Transcriptome and proteomics studies of T21 tissues found that, while there is a genome-wide transcriptional change including mitochondrial genes located on Hsa21 or other chromosomes (Letourneau et al., 2014; Sinet, 1982), the protein levels of those deregulated genes do not necessarily correlate with their mRNA levels (Lockstone et al., 2007), which suggests a post-transcriptional regulatory effect of mitochondrial genes (Yansheng Liu et al., 2017). However, comprehensive studies on the transcriptome and proteome of mitochondrial genes under T21 conditions are limited and the underlying mechanism of mitochondrial dysfunction in DS patients remains elusive.

In light of that, the differential expression data of the transcriptome and proteome of two trisomy 21 cell lines (HCT116 and RPE1) was uploaded to mitoXplorer for a thorough mitochondrial analysis. The **Interactome View** provides a general overview of differential expression of mitochondrial genes in individual samples. In contrast to heatmaps that are

adopted by most analytic platforms to visualize gene expression where genes are represented using rectangles of same sizes, the interactome view renders a bubble chart, where each circle represents a gene and its size and color reflect the extend and significance of differential expression, respectively. Together with the grouping of genes (circles) according to their functional annotations of the curated mitochondrial interactome, it was observed that while there was a strong deregulation in both cell lines, the patterns were significantly different. The differences in gene expression profiles of the two cell lines was not unanticipated as the energy and metabolic demands often depend on the cell type (Woods, 2017). The Interactome View's algorithm to position the gene groups also allows users to quickly recognize the most disrupted mitochondrial function (the bigger and more center-positioned the gene group, the more disrupted that function is). This is not easily achievable in conventional heatmaps as it is much more difficult to comprehend the addictive values of colors of certain rectangles than the total size of a group of circles. In the Interactome View of the proteome data from RPE1 T21 cell lines, the majority of OXPHOS genes were down-regulated and stood out from the rest of mitochondrial functions and dominated the View; whereas in the Interactome View of the transcriptome data, the same stark contrast in size and color of OXPHOS gene group could not be seen, but a strong down-regulation of genes involved in other mito-processes was observed instead. The comparison between the Interactome View of transcriptome and proteome data from RPE1 T21 cell lines hence revealed a large discrepancy in its expression at mRNA and protein levels.

7.2.3 mitoXplorer assisted in unraveling the potential cause for OXPHOS deficiency in RPE1 T21 cells

Further analysis with the **Comparative Plot** also showed a considerable difference in the transcriptome and proteome data from RPE1 T21 regarding OXPHOS genes. The Comparative Plot in mitoXplorer allows side-by-side comparison of differential expression data of different -omics data, and highlighted the fact that the majority of OXPHOS genes were down-regulated at protein, but not mRNA level. The formation of intact respiratory chain components that are encoded in the mitochondrial genome rely strongly on mitochondrial replication, transcription and translation. Since the mitochondrial transcript levels were not significantly different in T21 tissues, mtDNA-maintenance, -replication as well as mito-transcription appeared to be unaffected. However the comparative plot of genes belonging to the mitochondrial process **Translation** showed down-regulation of several **mitochondrial ribosomal protein**, implying a defective mitochondrial translation process that potentially causes the failure in the assembly of respiratory complexes due to missing mitochondrial subunits. This explains the extensive down-regulation of OXPHOS

proteins and implies a severe OXPHOS deficiency along the entire respiratory chain, which has been confirmed experimentally.

Mitochondrial ribosomal protein S21 (MRPS21) was among other mitochondrial ribosomal protein the most drastically down-regulated (>10-fold) at transcript level as seen from the comparative plot. Mrps21 is encoded by the nuclear genome and is a late-assembly component of mitoribosome small subunit (SSU) and could interact with other proteins of the SSU. Together with the down-regulation of other mitoribosome components (Mrps33, Mrps14 and Mrps15), this suggested that the mitochondrial translation process could be impaired due to ribosome malfunction as the late-assembly proteins break down and leads to mitoribosome degradation. Altogether, the findings made with mitoXplorer on post-transcriptional regulation in the trisomy 21 model system brought about new understanding in the mechanisms of mitochondrial defects in trisomy 21 patients.

7.3 Implications of mitochondria dysfunction in cancer

Being involved in various critical cellular pathways or functions like bioenergetic pathways, ROS defense and programmed cell death, mitochondrial metabolism has gradually been recognized to be influencing different steps of oncogenesis such as malignant transformation and tumor progression (Porporato et al., 2018; Vyas et al., 2016; Wallace, 2012). There are currently plenty of tools for the analysis of public cancer dataset (e.g. TCGA) and/or the visualization of pre-analyzed datasets (**Section 4.3.4**). However, rarely do they offer solutions for the specific analysis of mitochondrial genes in these datasets.

7.3.1 Expression of TCA-cycle genes as a potential indicator for late stage kidney renal papillary cell carcinoma (KIRP)

CancerSysDB (**Section 6.2**) is a platform that enables users to make customized queries and perform analyzes across multiple data types (somatic mutation, differential gene expression, clinical data) and cancer cohort from TCGA dataset. It also includes an interactive workflow that has integrated with the manually curated human mitochondrial interactome. The result is a dashboard that display expression and clinical data with interactive visualizations, which allows not only the in-depth analysis of differential expression of genes of various mitochondrial function, but also correlation analysis with clinical features, and could potentially provide new insights to the role mitochondrial metabolism in the development of cancer.

With this interactive workflow, an interesting dynamics of Tricarboxylic acid (TCA) cycle in KIRP (kidney renal papillary cell carcinoma) patients during tumor progression was

observed. In later stages of KIRP, a significant number of TCA cycle related genes were down-regulated, especially the Succinate-CoA ligase subunits SUCLG1 and SUCLG2. There have been accumulating evidences that TCA cycle plays a key role in cancer metabolism (Sajnani et al., 2017), that certain tumor suppressors and oncogenes regulate the expression of fuel transporters and/or activity of enzymes of TCA cycle in cancer cells in order to control both the uptake and breakdown of fuel sources (N. M. Anderson et al., 2018; J. Q. Chen & Russo, 2012). Succinate-CoA ligase (SUCL) is an enzyme that catalyzes the conversion of succinyl-CoA to succinate and coupling of phosphate and nucleoside diphosphate molecule to give ATP or GTP. Expression changes of SUCLG1 and 2 at mRNA and protein level have been identified in various studies of kidney cancer (Hakimi et al., 2016; Sanders & Diehl, 2015). The observation of stage-specific down-regulation of both SUCL subunits with CancerSysDB further proposes that SUCLG1, along with SUCLG2 as suggested in a previous study, could be a potential indicator for late stages in clear cell renal carcinomas (Perroud et al., 2009).

7.3.2 Epigenetics modifications and its effect on expression of mitochondrial genes in a mouse liver cancer model

Apart from TCA cycle, many other mitochondrial functions or processes have also been shown to be involved in oncogenesis or cancer development, or plays an important role in cancer metabolism (Wallace, 2012). The altered metabolism in cancer cells could be a result of expression changes in related genes, which is influenced by both genetic and epigenetic information (Esteller, 2011). In **Section 6.3**, a clinically relevant hepatocellular carcinoma (HCC) mouse model was used to study the epigenetic mechanism that influences transcription and gene expression through DNA methylation. Focal hypermethylation in CpG islands (CGIs), accompanied with lower expression (H+E-), was found in some genes, which is a well-known mechanism in cancer to downregulate tumor suppressor through promoter hypermethylation. H+E- genes found in the model include Ornithine aminotransferase (**Oat**) and Enoyl-Coenzyme A delta isomerase 1 (**Eci1**), which belong to the mitochondrial processes Amino Acid Metabolism and Fatty Acid Degradation & Beta-oxidation respectively.

Ornithine aminotransferase (OAT) is a mitochondrial matrix enzyme that located mainly in liver, brain and kidney, which catalyzes the reaction to convert ornithine into glutamate semi-aldehyde by transferring the delta-amino group (Ginguay et al., 2017). It has an inverse relationship with ornithine levels (Ventura et al., 2009) and OAT deficiency results in elevated levels of ornithine in blood. Ornithine is a precursor of polyamine (Pegg, 2009), which is involved in cell growth, proliferation, and apoptosis, and its abnormal accumulation is associated with various diseases including cancer (He et al., 2017). However, some other

literatures suggested an inverse relationship between ornithine level and breast cancer risk (Jiayi Zhang et al., 2020), and OAT could promote proliferation and invasion of non-small cell lung cancer (Yanfeng Liu et al., 2019). The role of ornithine metabolism and OAT in tumorigenesis and tumor progression thus remains unclear and is probably dependent on the cancer type.

Remarkably, there is a group of genes, including the mitochondrial gene Jun, which were found to be both hyper-methylated and overexpressed (H+E+) in the mouse model. Jun encodes for the transcription factor c-Jun and is considered as an oncogene as c-Jun could promote cell proliferation by changing gene expression (Martinez-Caballero et al., 2009). The analysis of the TCGA dataset showed that enrichment of this set of genes characterizes 56% of the HCC patients, who belong to an aggressive HCC subclass. These data suggested that a certain set of oncogenes with up-regulated expression characterized by hypermethylation of CGIs could be a useful as biomarkers for patient stratifications; and that the altered expressions of certain sets of genes, including a few mitochondrial genes, due to epigenetic modifications could help reveal the biological processes that possibly contribute to the disease condition.

7.4 Limitations and future work

The mitochondrial interactomes with proper gene annotations is central to the analysis and visualization made on the visual data mining platforms mentioned in this thesis. I am aware of the possibility of having false-positives or missing genes despite extensive human curation. Continuous revision of the interactomes is important to keep the platforms updated so that they could stay useful to the scientific community. A Feedback section is therefore available on mitoXplorer to collect comments from users who are experienced in the field of mitochondrial research, which might help with further cleaning and completing the annotations and the interactomes. Currently, each gene is also intentionally assigned to only one mitochondrial process to simplify the analysis, which might not represent all biological functions of a gene. However, in cases where strong experimental evidence that a protein or protein complex is involved in multiple process, the assignment of the corresponding gene to more than one process should be considered in the future.

At the moment, mitoXplorer allows the analysis and integration of only expression (transcriptomic and proteomic) and mutation (genomic) data. However, it is evident that the regulatory mechanisms of mitochondrial genes also operate beyond these levels as seen from previous discussion. It is therefore important to further integrate different types of -omics data, such as epigenomics data (e.g. ChIP-seq data from epigenetics studies), for a

comprehensive analysis of cross -omics data of mitochondrial genes. On the other hand, clinical data such as disease stages and survival data could be incorporated to help users gain further insights into the role of mitochondrial genes in the progression of certain diseases.

With more -omics data at different levels in possible future releases, it is sensible to introduce more knowledge-driven analysis into the platform, such as enrichment analysis, which could be helpful for gaining insights into biological mechanisms. The current mitochondrial interactomes could provide *a priori* gene sets that are grouped by their involvement in the same mitochondrial pathways for enrichment analysis. Apart from manual curation, such gene sets could possibly be further expanded by combining prior biological knowledge with the application of statistical methods. For example, metabolic pathways have been predicted using a Bayesian probabilistic graphical model with the constraints of known gene-gene interactions by sampling co-expressed genes from gene clusters derived from gene expression data (Qi et al., 2014).

The data-driven analysis methods on mitoXplorer (e.g. Hierarchical Clustering and Principal Component Analysis (PCA)) are useful to reveal patterns and identify groups. Nevertheless, every approach has its own limitations. For example, PCA could sometimes be highly affected by outliers; t-SNE, another dimensionality reduction algorithms, could handle outliers and capture non-linear relationship between features, yet it is non-deterministic and computationally complex. Depending on the objective, one approach could be more favorable then the other. More options should therefore be available to users to suit their analysis. And to further enhance user experience, additional visualizations (e.g. violin plots, box plots, network and pathways graphs) for analysis using current or potential new approaches, and features for customization such as providing more color schemes, adjusting font size, resolution, etc. could also be considered, so that users could download publication-ready graphs after data mining and exploration.

Finally, given the practicality of mitoXplorer, it should be considered to extend the usage to more organisms or even other gene sets by compiling more “interactomes”, either manually or with the combination of statistical and computation methods mentioned above, in order to benefit a broader scientific community.

7.5 Conclusions

The work presented in this thesis has proven that visual data mining tools could be a robust instrument to explore and analyze the mutation and expression dynamics of a defined gene

set. mitoXplorer and CancerSysDB have been developed as web platforms to provide data mining and visualization service specifically for mitochondrial genes, by integrating manually curated annotations of different species with a set of dynamic, interactive and intuitive visualizations for the analysis of -omics data. This allows the exploration and mining of data in the context of mitochondrial functions, using intuitive visualizations that could often help discovering hidden patterns. The analysis of transcription and expression data of aneuploidy cells and cancer patients, as well as the experimental verification of the observation made by mitoXplorer on the phenotypes of trisomy cell lines, have demonstrated how these tools could help to discover underlying molecular mechanisms in different disease conditions through suggesting testable hypothesis for further experimental validation.

8 References

Abate, C., Patel, L., Rauscher, F. J., & Curran, T. (1990). Redox regulation of fos and jun DNA-binding activity in vitro. *Science*, *249*(4973), 1157–1161. https://doi.org/10.1126/science.2118682

Adeva, M. M., Souto, G., Blanco, N., & Donapetry, C. (2012). Ammonium metabolism in humans. *Metabolism: Clinical and Experimental*, *61*(11), 1495–1511. https://doi.org/10.1016/j.metabol.2012.07.007

al Rawi, S., Louvet-Vallée, S., Djeddi, A., Sachse, M., Culetto, E., Hajjar, C., Boyd, L., Legouis, R., & Galy, V. (2011). Postfertilization autophagy of sperm organelles prevents paternal mitochondrial DNA transmission. *Science (New York, N.Y.)*, *334*(6059), 1144–1147. https://doi.org/10.1126/science.1211878

Alam, T. I., Kanki, T., Muta, T., Ukaji, K., Abe, Y., Nakayama, H., Takio, K., Hamasaki, N., & Kang, D. (2003). Human mitochondrial DNA is packaged with TFAM. In *Nucleic Acids Research* (Vol. 31, Issue 6, pp. 1640–1645). Nucleic Acids Res. https://doi.org/10.1093/nar/gkg251

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (1994). *Molecular Biology of the Cell*. Garland Publishing Inc.

Anderson, N. M., Mucka, P., Kern, J. G., & Feng, H. (2018). The emerging role and targetability of the TCA cycle in cancer metabolism. In *Protein and Cell* (Vol. 9, Issue 2, pp. 216–237). Higher Education Press. https://doi.org/10.1007/s13238-017-0451-1

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R., & Young, I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, *290*(5806), 457–465. http://www.ncbi.nlm.nih.gov/pubmed/7219534

Anesti, V., & Scorrano, L. (2006). The relationship between mitochondrial shape and function and the cytoskeleton. *Biochimica et Biophysica Acta*, *1757*(5–6), 692–699. https://doi.org/10.1016/j.bbabio.2006.04.013

Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, *12*(7), 878. https://doi.org/10.15252/msb.20156651

Anneren, K. G., & Epstein, C. J. (1987). Lipid peroxidation and superoxide dismutase-1 and glutathione peroxidase activities in trisomy 16 fetal mice and human trisomy 21 fibroblasts. *Pediatric Research*, *21*(1), 88–92. https://doi.org/10.1203/00006450-198701000-00019

Antonarakis, S. E. (2017). Down syndrome and the complexity of genome dosage imbalance. In *Nature Reviews Genetics* (Vol. 18, Issue 3, pp. 147–163). Nature Publishing Group. https://doi.org/10.1038/nrg.2016.154

Attardi, G., & Ojala, D. (1971). Mitochondrial ribosomes in HeLa cells. *Nature New Biology*, *229*(5), 133–136. https://doi.org/10.1038/newbio229133a0

Baylin, S. B., & Jones, P. A. (2011). A decade of exploring the cancer epigenome-biological and translational implications. In *Nature Reviews Cancer* (Vol. 11, Issue 10, pp. 726–734). Nat Rev Cancer. https://doi.org/10.1038/nrc3130

Baysal, B. E., Ferrell, R. E., Willett-Brozick, J. E., Lawrence, E. C., Myssiorek, D., Bosch, A., van der Mey, A., Taschner, P. E. M., Rubinstein, W. S., Myers, E. N., Richard, C. W., Cornelisse, C. J., Devilee, P., & Devlin, B. (2000). Mutations in SDHD, a mitochondrial complex II gene, in hereditary paraganglioma. *Science*, *287*(5454), 848–851. https://doi.org/10.1126/science.287.5454.848

Beinert, H., Holm, R. H., & Münck, E. (1997). Iron-sulfur clusters: nature's modular, multipurpose structures. *Science (New York, N.Y.)*, *277*(5326), 653–659. http://www.ncbi.nlm.nih.gov/pubmed/9235882

Beinert, H., & Kiley, P. J. (1999). Fe-S proteins in sensing and regulatory functions. *Current Opinion in Chemical Biology*, *3*(2), 152–157. https://doi.org/10.1016/S1367-5931(99)80027-1

Belyaeva, O. v., Korkina, O. v., Stetsenko, A. v., & Kedishvili, N. Y. (2008). Human retinol dehydrogenase 13 (RDH13) is a mitochondrial short-chain dehydrogenase/reductase with a retinaldehyde reductase activity. *FEBS Journal*, *275*(1), 138–147. https://doi.org/10.1111/j.1742-4658.2007.06184.x

Bersu, E. T., Ahmad, F. J., Schwei, M. J., & Baas, P. W. (1998). Cytoplasmic abnormalities in cultured cerebellar neurons from the trisomy 16 mouse. *Developmental Brain Research*, *109*(1), 115–120. https://doi.org/10.1016/S0165-3806(98)00070-4

Bione, S., D'Adamo, P., Maestrini, E., Gedeon, A. K., Bolhuis, P. A., & Toniolo, D. (1996). A novel X-linked gene, G4.5. is responsible for Barth syndrome. *Nature Genetics*, *12*(4), 385–389. https://doi.org/10.1038/ng0496-385

Böhm, I., & Schild, H. (2003). Apoptosis: the complex scenario for a silent cell death. *Molecular Imaging and Biology : MIB : The Official Publication of the Academy of Molecular Imaging*, *5*(1), 2–14. http://www.ncbi.nlm.nih.gov/pubmed/14499155

Brás, A., Monteiro, C., & Rueff, J. (1989). Oxidative stress in trisomy 21: A possible role in cataractogenesis. *Ophthalmic Genetics*, *10*(4), 271–277. https://doi.org/10.3109/13816818909009882

Brega, A., & Vesco, C. (1971). Ribonucleoprotein particles involved in HeLa mitochondrial protein synthesis. *Nature New Biology*, *229*(5), 136–139. https://doi.org/10.1038/newbio229136a0

Bricker, D. K., Taylor, E. B., Schell, J. C., Orsak, T., Boutron, A., Chen, Y.-C., Cox, J. E., Cardon, C. M., van Vranken, J. G., Dephoure, N., Redin, C., Boudina, S., Gygi, S. P., Brivet, M., Thummel, C. S., & Rutter, J. (2012). A mitochondrial pyruvate carrier required for pyruvate uptake in yeast, Drosophila, and humans. *Science (New York, N.Y.)*, *337*(6090), 96–100. https://doi.org/10.1126/science.1218099

Brody, S., Oh, C., Hoja, U., & Schweizer, E. (1997). Mitochondrial acyl carrier protein is involved in lipoic acid synthesis in Saccharomyces cerevisiae. *FEBS Letters*, *408*(2), 217–220. http://www.ncbi.nlm.nih.gov/pubmed/9187370

Brooksbank, B. W. L., & Balazs, R. (1984). Superoxide dismutase, glutathione peroxidase and lipoperoxidation in Oown's syndrome fetal brain. *Developmental Brain Research*, *16*(1), 37–44. https://doi.org/10.1016/0165-3806(84)90060-9

Brown, K. R., Otasek, D., Ali, M., McGuffin, M. J., Xie, W., Devani, B., van Toch, I. L., & Jurisica, I. (2009). NAViGaTOR: Network analysis, visualization and graphing Toronto. *Bioinformatics*, *25*(24), 3327–3329. https://doi.org/10.1093/bioinformatics/btp595

Brusilow, S. W., & Horwich, A. L. (2001). Urea cycle enzymes. In C. R. Scriver, A. L. Beaudet, W. S. Sly, & D. Valle (Eds.), *The Metabolic & Molecular Bases of Inherited Disease*. New York, NY: McGraw-Hill.

Busch, K. B., Kowald, A., & Spelbrink, J. N. (2014). Quality matters: How does mitochondrial network dynamics and quality control impact on mtDNA integrity? In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 369, Issue 1646). Royal Society of London. https://doi.org/10.1098/rstb.2013.0442

Busciglio, J., & Yankner, B. A. (1995). Apoptosis and increased generation of reactive oxygen species in down's syndrome neurons in vitro. *Nature*, *378*(6559), 776–779. https://doi.org/10.1038/378776a0

Cagin, U., & Enriquez, J. A. (2015). The complex crosstalk between mitochondria and the nucleus: What goes in between? *The International Journal of Biochemistry & Cell Biology*, *63*, 10–15. https://doi.org/10.1016/j.biocel.2015.01.026

Calvo, S. E., Clauser, K. R., & Mootha, V. K. (2016). MitoCarta2.0: An updated inventory of mammalian mitochondrial proteins. *Nucleic Acids Research*, *44*(D1), D1251–D1257. https://doi.org/10.1093/nar/gkv1003

Cantatore, P., Flagella, Z., Fracasso, F., Lezza, A. M. S., Gadaleta, M. N., & de Montalvo, A. (1987). Synthesis and turnover rates of four rat liver mitochondrial RNA species. *FEBS Letters*, *213*(1), 144–148. https://doi.org/10.1016/0014-5793(87)81480-1

Carelli, V., Maresca, A., Caporali, L., Trifunov, S., Zanna, C., & Rugolo, M. (2015). Mitochondria: Biogenesis and mitophagy balance in segregation and clonal expansion of mitochondrial DNA mutations. In *International Journal of Biochemistry and Cell Biology* (Vol. 63, pp. 21–24). Elsevier Ltd. https://doi.org/10.1016/j.biocel.2015.01.023

Catalano, D., Licciulli, F., Turi, A., Grillo, G., Saccone, C., & D'Elia, D. (2006). MitoRes: A resource of nuclear-encoded mitochondrial genes and their products in Metazoa. *BMC Bioinformatics*, *7*. https://doi.org/10.1186/1471-2105-7-36

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., & Schultz, N. (2012). The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, *2*(5), 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095

Chacinska, A., Koehler, C. M., Milenkovic, D., Lithgow, T., & Pfanner, N. (2009). Importing Mitochondrial Proteins: Machineries and Mechanisms. In *Cell* (Vol. 138, Issue 4, pp. 628–644). Cell. https://doi.org/10.1016/j.cell.2009.08.005

Chan, D. C. (2012). Fusion and Fission: Interlinked Processes Critical for Mitochondrial Health. *Annual Review of Genetics*, *46*(1), 265–287. https://doi.org/10.1146/annurev-genet-110410-132529

Chen, J. Q., & Russo, J. (2012). Dysregulation of glucose transport, glycolysis, TCA cycle and glutaminolysis by oncogenes and tumor suppressors in cancer cells. In *Biochimica et Biophysica Acta - Reviews on Cancer* (Vol. 1826, Issue 2, pp. 370–384). Biochim Biophys Acta. https://doi.org/10.1016/j.bbcan.2012.06.004

Chen, X., Wei, S., Ji, Y., Guo, X., & Yang, F. (2015). Quantitative proteomics using SILAC: Principles, applications, and developments. In *Proteomics* (Vol. 15, Issue 18, pp. 3175–3192). Wiley-VCH Verlag. https://doi.org/10.1002/pmic.201500108

Chen, Y. C., Ke, W. C., & Chiu, H. W. (2014). Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. *Computers in Biology and Medicine*, *48*(1), 1–7. https://doi.org/10.1016/j.compbiomed.2014.02.006

Chen, Y., Zhang, H., Zhou, H., Ji, W., & Min, W. (2016). Mitochondrial Redox Signaling and Tumor Progression. *Cancers*, *8*(4), 40. https://doi.org/10.3390/cancers8040040

Chowdhury, A., Aich, A., Jain, G., Wozny, K., Lüchtenborg, C., Hartmann, M., Bernhard, O., Balleiniger, M., Alfar, E. A., Zieseniss, A., Toischer, K., Guan, K., Rizzoli, S. O., Brügger, B., Fischer, A., Katschinski, D. M., Rehling, P., & Dudek, J. (2018). Defective Mitochondrial Cardiolipin Remodeling Dampens HIF-1α Expression in Hypoxia. *Cell Reports*, *25*(3), 561-570.e6. https://doi.org/10.1016/j.celrep.2018.09.057

Clayton, D. A. (1982). Replication of animal mitochondrial DNA. In *Cell* (Vol. 28, Issue 4, pp. 693–705). Cell. https://doi.org/10.1016/0092-8674(82)90049-6

Cory, S., & Adams, J. M. (2002). The Bcl2 family: regulators of the cellular life-or-death switch. *Nature Reviews. Cancer*, *2*(9), 647–656. https://doi.org/10.1038/nrc883

Costa, C. D. (2020). *Top Programming Languages for Data Science in 2020*.

Dagda, R. K., Zhu, J., & Chu, C. T. (2009). Mitochondrial kinases in Parkinson's disease: converging insights from neurotoxin and genetic models. *Mitochondrion*, *9*(5), 289–298. https://doi.org/10.1016/j.mito.2009.06.001

Daum, G., & Vance, J. E. (1997). Import of lipids into mitochondria. *Progress in Lipid Research*, *36*(2–3), 103–130. http://www.ncbi.nlm.nih.gov/pubmed/9624424

Davis, R. E., & Williams, M. (2012). Mitochondrial function and dysfunction: an update. *The Journal of Pharmacology and Experimental Therapeutics*, *342*(3), 598–607. https://doi.org/10.1124/jpet.112.192104

de La Torre, R., Casado, A., López-Fernández, E., Carrascosa, D., Ramírez, V., & Sáez, J. (1996). Overexpression of copper-zinc superoxide dismutase in trisomy 21. *Experientia*, *52*(9), 871–873. https://doi.org/10.1007/BF01938872

Desai, A., Sequeira, J. M., & Quadros, E. v. (2016). The metabolic basis for developmental disorders due to defective folate transport. In *Biochimie* (Vol. 126, pp. 31–42). Elsevier B.V. https://doi.org/10.1016/j.biochi.2016.02.012

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Duchen, M. R. (2004). Mitochondria in health and disease: perspectives on a new mitochondrial biology. *Molecular Aspects of Medicine*, *25*(4), 365–451. https://doi.org/10.1016/j.mam.2004.03.001

Dyall, S. D., Brown, M. T., & Johnson, P. J. (2004). Ancient invasions: from endosymbionts to organelles. *Science (New York, N.Y.)*, *304*(5668), 253–257. https://doi.org/10.1126/science.1094884

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–210. https://doi.org/10.1093/nar/30.1.207

Esteller, M. (2011). Epigenetic changes in cancer. In *F1000 Biology Reports* (Vol. 3, Issue 1). F1000 Biol Rep. https://doi.org/10.3410/B3-9

Estrella, V., Chen, T., Lloyd, M., Wojtkowiak, J., Cornnell, H. H., Ibrahim-Hashim, A., Bailey, K., Balagurunathan, Y., Rothberg, J. M., Sloane, B. F., Johnson, J., Gatenby, R. A., & Gillies, R. J. (2013). Acidity generated by the tumor microenvironment drives local invasion. *Cancer Research*, *73*(5), 1524–1535. https://doi.org/10.1158/0008-5472.CAN-12-2796

Evans, D. R., & Guy, H. I. (2004). Mammalian pyrimidine biosynthesis: Fresh insights into an ancient pathway. In *Journal of Biological Chemistry* (Vol. 279, Issue 32, pp. 33035–33038). J Biol Chem. https://doi.org/10.1074/jbc.R400007200

Farge, G. É. R., & Falkenberg, M. (2019). Organization of DNA in mammalian mitochondria. In *International Journal of Molecular Sciences* (Vol. 20, Issue 11). MDPI AG. https://doi.org/10.3390/ijms20112770

Fisher, R. P., & Clayton, D. A. (1988). Purification and characterization of human mitochondrial transcription factor 1. *Molecular and Cellular Biology*, *8*(8), 3496–3509. https://doi.org/10.1128/mcb.8.8.3496

Fonseca, N. A., Rung, J., Brazma, A., & Marioni, J. C. (2012). Tools for mapping high-throughput sequencing data. In *Bioinformatics* (Vol. 28, Issue 24, pp. 3169–3177). Bioinformatics. https://doi.org/10.1093/bioinformatics/bts605

Forloni, M., Gupta, R., Nagarajan, A., Sun, L. S., Dong, Y., Pirazzoli, V., Toki, M., Wurtz, A., Melnick, M. A., Kobayashi, S., Homer, R. J., Rimm, D. L., Gettinger, S. J., Politi, K., Dogra, S. K., & Wajapeyee, N. (2016). Oncogenic EGFR Represses the TET1 DNA Demethylase to Induce Silencing of Tumor Suppressors in Cancer Cells. *Cell Reports*, *16*(2), 457–471. https://doi.org/10.1016/j.celrep.2016.05.087

Freije, W. A., Mandal, S., & Banerjee, U. (2012). Expression profiling of attenuated mitochondrial function identifies retrograde signals in Drosophila. *G3 (Bethesda, Md.)*, *2*(8), 843–851. https://doi.org/10.1534/g3.112.002584

Friedman, J. R., & Nunnari, J. (2014). Mitochondrial form and function. *Nature*, *505*(7483), 335–343. https://doi.org/10.1038/nature12985

Fuchs, Y., & Steller, H. (2011). Programmed cell death in animal development and disease. *Cell*, *147*(4), 742–758. https://doi.org/10.1016/j.cell.2011.10.033

Gabaldon, T. (2003). Reconstruction of the Proto-Mitochondrial Metabolism. *Science*, *301*(5633), 609–609. https://doi.org/10.1126/science.1085463

Gabaldón, T., & Huynen, M. A. (2004). Shaping the mitochondrial proteome. *Biochimica et Biophysica Acta*, *1659*(2–3), 212–220. https://doi.org/10.1016/j.bbabio.2004.07.011

Gaucher, S. P., Taylor, S. W., Fahy, E., Zhang, B., Warnock, D. E., Ghosh, S. S., & Gibson, B. W. (2004). Expanded coverage of the human heart mitochondrial proteome using multidimensional liquid chromatography coupled with tandem mass spectrometry. *Journal of Proteome Research*, *3*(3), 495–505. http://www.ncbi.nlm.nih.gov/pubmed/15253431

Gilkerson, R., Bravo, L., Garcia, I., Gaytan, N., Herrera, A., Maldonado, A., & Quintanilla, B. (2013). The mitochondrial nucleoid: Integrating mitochondrial DNA into cellular homeostasis. *Cold Spring Harbor Perspectives in Biology*, *5*(5). https://doi.org/10.1101/cshperspect.a011080

Ginguay, A., Cynober, L., Curis, E., & Nicolis, I. (2017). Ornithine aminotransferase, an important Glutamate-Metabolizing enzyme at the crossroads of multiple metabolic pathways. In *Biology* (Vol. 6, Issue 1). MDPI AG. https://doi.org/10.3390/biology6010018

Glish, G. L., & Vachet, R. W. (2003). The basics of mass spectrometry in the twenty-first century. In *Nature Reviews Drug Discovery* (Vol. 2, Issue 2, pp. 140–150). Nat Rev Drug Discov. https://doi.org/10.1038/nrd1011

Gomes, L. C., Benedetto, G. di, & Scorrano, L. (2011). During autophagy mitochondria elongate, are spared from degradation and sustain cell viability. *Nature Cell Biology*, *13*(5), 589–598. https://doi.org/10.1038/ncb2220

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. In *Nature Reviews Genetics* (Vol. 17, Issue 6, pp. 333–351). Nature Publishing Group. https://doi.org/10.1038/nrg.2016.49

Goto, Y. I., Nonaka, I., & Horai, S. (1990). A mutation in the tRNALeu(UUR) gene associated with the MELAS subgroup of mitochondrial encephalomyopathies. *Nature*, *348*(6302), 651–653. https://doi.org/10.1038/348651a0

Gray, M. W. (1999). Mitochondrial Evolution. *Science*, *283*(5407), 1476–1481. https://doi.org/10.1126/science.283.5407.1476

Green, D. R., & Kroemer, G. (2004). The pathophysiology of mitochondrial cell death. In *Science* (Vol. 305, Issue 5684, pp. 626–629). Science. https://doi.org/10.1126/science.1099320

Guda, C., Fahy, E., & Subramaniam, S. (2004). MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics (Oxford, England)*, *20*(11), 1785–1794. https://doi.org/10.1093/bioinformatics/bth171

Guda, P., Guda, C., & Subramaniam, S. (2007). Reconstruction of Pathways Associated with Amino Acid Metabolism in Human Mitochondria. *Genomics, Proteomics and Bioinformatics*, *5*(3–4), 166–176. https://doi.org/10.1016/S1672-0229(08)60004-2

Hakimi, A. A., Reznik, E., Lee, C. H., Creighton, C. J., Brannon, A. R., Luna, A., Aksoy, B. A., Liu, E. M., Shen, R., Lee, W., Chen, Y., Stirdivant, S. M., Russo, P., Chen, Y. B., Tickoo, S. K., Reuter, V. E., Cheng, E. H., Sander, C., & Hsieh, J. J. (2016). An Integrated Metabolic Atlas of Clear Cell Renal Cell Carcinoma. *Cancer Cell*, *29*(1), 104–116. https://doi.org/10.1016/j.ccell.2015.12.004

Hatem, A., Bozdağ, D., Toland, A. E., & Çatalyürek, Ü. v. (2013). Benchmarking short sequence mapping tools. *BMC Bioinformatics*, *14*(1). https://doi.org/10.1186/1471-2105-14-184

Hayashi, J. I., Ohta, S., Kikuchi, A., Takemitsu, M., Goto, Y. I., & Nonaka, I. (1991). Introduction of disease-related mitochondrial DNA deletions into HeLa cells lacking mitochondrial DNA results in mitochondrial dysfunction. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(23), 10614–10618. https://doi.org/10.1073/pnas.88.23.10614

He, W., Roh, E., Yao, K., Liu, K., Meng, X., Liu, F., Wang, P., Bode, A. M., & Dong, Z. (2017). Targeting ornithine decarboxylase (ODC) inhibits esophageal squamous cell

carcinoma progression. *Npj Precision Oncology*, *1*(1). https://doi.org/10.1038/s41698-017-0014-1

Heiden, M. G. V., Cantley, L. C., & Thompson, C. B. (2009). Understanding the warburg effect: The metabolic requirements of cell proliferation. In *Science* (Vol. 324, Issue 5930, pp. 1029–1033). Science. https://doi.org/10.1126/science.1160809

Hensen, F., Cansiz, S., Gerhold, J. M., & Spelbrink, J. N. (2014). To be or not to be a nucleoid protein: A comparison of mass-spectrometry based approaches in the identification of potential mtDNA-nucleoid associated proteins. *Biochimie*, *100*, 219–226. https://doi.org/10.1016/j.biochi.2013.09.017

Hoitzing, H., Johnston, I. G., & Jones, N. S. (2015). What is the function of mitochondrial networks? A theoretical assessment of hypotheses and proposal for future research. *BioEssays*, *37*(6), 687–700. https://doi.org/10.1002/bies.201400188

Holzerová, E., & Prokisch, H. (2015). Mitochondria: Much ado about nothing? How dangerous is reactive oxygen species production? *The International Journal of Biochemistry & Cell Biology*, *63*, 16–20. https://doi.org/10.1016/j.biocel.2015.01.021

Hoppins, S. C., & Nargang, F. E. (2004). The Tim8-Tim13 Complex of Neurospora crassa Functions in the Assembly of Proteins into Both Mitochondrial Membranes. *Journal of Biological Chemistry*, *279*(13), 12396–12405. https://doi.org/10.1074/jbc.M313037200

Huang, T., Lan, L., Fang, X., An, P., Min, J., & Wang, F. (2015). Promises and Challenges of Big Data Computing in Health Sciences. In *Big Data Research* (Vol. 2, Issue 1, pp. 2–11). Elsevier Inc. https://doi.org/10.1016/j.bdr.2015.02.002

Huang, W., Choi, W., Hu, W., Mi, N., Guo, Q., Ma, M., Liu, M., Tian, Y., Lu, P., Wang, F. L., Deng, H., Liu, L., Gao, N., Yu, L., & Shi, Y. (2012). Crystal structure and biochemical analyses reveal Beclin 1 as a novel membrane binding protein. *Cell Research*, *22*(3), 473–489. https://doi.org/10.1038/cr.2012.24

Jain, A. (2016). *The 5 V's of big data*. Watson Health Perspectives.

Johns, D. R., Neufeld, M. J., & Park, R. D. (1992). An ND-6 mitochondrial DNA mutation associated with leber hereditary optic neuropathy. *Biochemical and Biophysical Research Communications*, *187*(3), 1551–1557. https://doi.org/10.1016/0006-291X(92)90479-5

Jones, A. W. E., Yao, Z., Vicencio, J. M., Karkucinska-Wieckowska, A., & Szabadkai, G. (2012). PGC-1 family coactivators and cell fate: roles in cancer, neurodegeneration, cardiovascular disease and retrograde mitochondria-nucleus signalling. *Mitochondrion*, *12*(1), 86–99. https://doi.org/10.1016/j.mito.2011.09.009

Jones, P. A., & Baylin, S. B. (2007). The Epigenomics of Cancer. In *Cell* (Vol. 128, Issue 4, pp. 683–692). Cell. https://doi.org/10.1016/j.cell.2007.01.029

Jovanovic, S. v., Clements, D., & MacLeod, K. (1998). Biomarkers of oxidative stress are significantly elevated in Down syndrome. *Free Radical Biology and Medicine*, *25*(9), 1044–1048. https://doi.org/10.1016/S0891-5849(98)00137-3

Kaufmann, S. H., & Earnshaw, W. C. (2000). Induction of apoptosis by cancer chemotherapy. *Experimental Cell Research*, *256*(1), 42–49. https://doi.org/10.1006/excr.2000.4838

King, A., Selak, M. A., & Gottlieb, E. (2006). Succinate dehydrogenase and fumarate hydratase: Linking mitochondrial dysfunction and cancer. In *Oncogene* (Vol. 25, Issue 34, pp. 4675–4682). Oncogene. https://doi.org/10.1038/sj.onc.1209594

Kodama, Y., Mashima, J., Kosuge, T., & Ogasawara, O. (2019). DDBJ update: The Genomic Expression Archive (GEA) for functional genomics data. *Nucleic Acids Research*, *47*(D1), D69–D73. https://doi.org/10.1093/nar/gky1002

Kogelnik, A. M., Lott, M. T., Brown, M. D., Navathe, S. B., & Wallace, D. C. (1998). MITOMAP: A human mitochondrial genome database - 1998 update. *Nucleic Acids Research*, *26*(1), 112–114. https://doi.org/10.1093/nar/26.1.112

Kujoth, G. C., Hiona, A., Pugh, T. D., Someya, S., Panzer, K., Wohlgemuth, S. E., Hofer, T., Seo, A. Y., Sullivan, R., Jobling, W. A., Morrow, J. D., van Remmen, H., Sedivy, J. M., Yamasoba, T., Tanokura, M., Weindruch, R., Leeuwenburgh, C., & Prolla, T. A. (2005). Mitochondrial DNA mutations, oxidative stress, and apoptosis in mammalian aging. *Science (New York, N.Y.)*, *309*(5733), 481–484. https://doi.org/10.1126/science.1112125

Kukat, C., & Larsson, N.-G. (2013). mtDNA makes a U-turn for the mitochondrial nucleoid. *Trends in Cell Biology*, *23*(9), 457–463. https://doi.org/10.1016/j.tcb.2013.04.009

Kuntal, B. K., Ghosh, T. S., & Mande, S. S. (2014). Igloo-plot: A tool for visualization of multidimensional datasets. *Genomics*, *103*(1), 11–20. https://doi.org/10.1016/j.ygeno.2014.01.004

Kurelac, I., Romeo, G., & Gasparre, G. (2011). Mitochondrial metabolism and cancer. In *Mitochondrion* (Vol. 11, Issue 4, pp. 635–637). Mitochondrion. https://doi.org/10.1016/j.mito.2011.03.012

Kutik, S., Rissler, M., Guan, X. L., Guiard, B., Shui, G., Gebert, N., Heacock, P. N., Rehling, P., Dowhan, W., Wenk, M. R., Pfanner, N., & Wiedemann, N. (2008). The translocator maintenance protein Tam41 is required for mitochondrial cardiolipin biosynthesis. *Journal of Cell Biology*, *183*(7), 1213–1221. https://doi.org/10.1083/jcb.200806048

Lane, A. N., & Fan, T. W. M. (2015). Regulation of mammalian nucleotide metabolism and biosynthesis. In *Nucleic Acids Research* (Vol. 43, Issue 4, pp. 2466–2485). Oxford University Press. https://doi.org/10.1093/nar/gkv047

Lane, N. (2011). Mitonuclear match: Optimizing fitness and fertility over generations drives ageing within generations. *BioEssays*, *33*(11), 860–869. https://doi.org/10.1002/bies.201100051

Lane, N., & Martin, W. (2010). The energetics of genome complexity. *Nature*, *467*(7318), 929–934. https://doi.org/10.1038/nature09486

Letourneau, A., Santoni, F. A., Bonilla, X., Sailani, M. R., Gonzalez, D., Kind, J., Chevalier, C., Thurman, R., Sandstrom, R. S., Hibaoui, Y., Garieri, M., Popadin, K., Falconnet, E., Gagnebin, M., Gehrig, C., Vannier, A., Guipponi, M., Farinelli, L., Robyr, D., … Antonarakis, S. E. (2014). Domains of genome-wide gene expression dysregulation in Down's syndrome. *Nature*, *508*(7496), 345–350. https://doi.org/10.1038/nature13200

Levine, A. J., & Puzio-Kuter, A. M. (2010). The control of the metabolic switch in cancers by oncogenes and tumor suppressor genes. In *Science* (Vol. 330, Issue 6009, pp. 1340–1344). Science. https://doi.org/10.1126/science.1193494

Lewis, C. A., Parker, S. J., Fiske, B. P., McCloskey, D., Gui, D. Y., Green, C. R., Vokes, N. I., Feist, A. M., vander Heiden, M. G., & Metallo, C. M. (2014). Tracing Compartmentalized NADPH Metabolism in the Cytosol and Mitochondria of Mammalian Cells. *Molecular Cell*, *55*(2), 253–263. https://doi.org/10.1016/j.molcel.2014.05.008

Lieber, T., Jeedigunta, S. P., Palozzi, J. M., Lehmann, R., & Hurd, T. R. (2019). Mitochondrial fragmentation drives selective removal of deleterious mtDNA in the germline. *Nature*, *570*(7761), 380–384. https://doi.org/10.1038/s41586-019-1213-4

Liu, Y. J., McIntyre, R. L., Janssens, G. E., & Houtkooper, R. H. (2020). Mitochondrial fission and fusion: A dynamic role in aging and potential target for age-related disease. *Mechanisms of Ageing and Development*, *186*. https://doi.org/10.1016/j.mad.2020.111212

Liu, Yanfeng, Wu, L., Li, K., Liu, F., Wang, L., Zhang, D., Zhou, J., Ma, X., Wang, S., & Yang, S. (2019). Ornithine aminotransferase promoted the proliferation and metastasis of non-small cell lung cancer via upregulation of miR-21. *Journal of Cellular Physiology*, *234*(8), 12828–12838. https://doi.org/10.1002/jcp.27939

Liu, Yansheng, Borel, C., Li, L., Müller, T., Williams, E. G., Germain, P. L., Buljan, M., Sajic, T., Boersema, P. J., Shao, W., Faini, M., Testa, G., Beyer, A., Antonarakis, S. E., & Aebersold, R. (2017). Systematic proteome and proteostasis profiling in human Trisomy 21 fibroblast cells. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-01422-6

Livnat-Levanon, N., & Glickman, M. H. (2011). Ubiquitin-Proteasome System and mitochondria - Reciprocity. In *Biochimica et Biophysica Acta - Gene Regulatory*

Mechanisms (Vol. 1809, Issue 2, pp. 80–87). Biochim Biophys Acta. https://doi.org/10.1016/j.bbagrm.2010.07.005

Locasale, J. W., & Cantley, L. C. (2011). Metabolic flux and the regulation of mammalian cell growth. In *Cell Metabolism* (Vol. 14, Issue 4, pp. 443–451). Cell Metab. https://doi.org/10.1016/j.cmet.2011.07.014

Lockstone, H. E., Harris, L. W., Swatton, J. E., Wayland, M. T., Holland, A. J., & Bahn, S. (2007). Gene expression profiling in the adult Down syndrome brain. *Genomics*, *90*(6), 647–660. https://doi.org/10.1016/j.ygeno.2007.08.005

Lodish, H. F., Berk, A., Kaiser, C., Monty, K., Scott, M. P., Bretscher, A., Ploegh, H. L., & Matsudaira, P. T. (2008). *Molecular Cell Biology, 6th edition*. W.H. Freeman.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12). https://doi.org/10.1186/s13059-014-0550-8

Lovell, J. F., Billen, L. P., Bindner, S., Shamas-Din, A., Fradin, C., Leber, B., & Andrews, D. W. (2008). Membrane Binding by tBid Initiates an Ordered Series of Events Culminating in Membrane Permeabilization by Bax. *Cell*, *135*(6), 1074–1084. https://doi.org/10.1016/j.cell.2008.11.010

Lunt, S. Y., & vander Heiden, M. G. (2011). Aerobic glycolysis: meeting the metabolic requirements of cell proliferation. *Annual Review of Cell and Developmental Biology*, *27*, 441–464. https://doi.org/10.1146/annurev-cellbio-092910-154237

Mannella, C. A., Buttle, K., Rath, B. K., & Marko, M. (1998). Electron microscopic tomography of rat-liver mitochondria and their interaction with the endoplasmic reticulum. *BioFactors (Oxford, England)*, *8*(3–4), 225–228. http://www.ncbi.nlm.nih.gov/pubmed/9914823

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation

Mardis, E. R. (2013). Next-generation sequencing platforms. *Annual Review of Analytical Chemistry*, *6*, 287–303. https://doi.org/10.1146/annurev-anchem-062012-092628

Martinelli, P., & Rugarli, E. I. (2010). Emerging roles of mitochondrial proteases in neurodegeneration. In *Biochimica et Biophysica Acta - Bioenergetics* (Vol. 1797, Issue 1, pp. 1–10). Biochim Biophys Acta. https://doi.org/10.1016/j.bbabio.2009.07.013

Martinez-Caballero, S., Dejean, L. M., Kinnally, M. S., Oh, K. J., Mannella, C. A., & Kinnally, K. W. (2009). Assembly of the mitochondrial apoptosis-induced channel, MAC. *Journal of Biological Chemistry*, *284*(18), 12235–12245. https://doi.org/10.1074/jbc.M806610200

Mayes, P. A. (1993). Intermediary metabolism of fructose. *The American Journal of Clinical Nutrition*, *58*(5 Suppl), 754S-765S. https://doi.org/10.1093/ajcn/58.5.754S

Meeusen, S., & Nunnari, J. (2003). Evidence for a two membrane-spanning autonomous mitochondrial DNA replisome. *The Journal of Cell Biology*, *163*(3), 503–510. https://doi.org/10.1083/jcb.200304040

Mikolajczyk, S., & Brody, S. (1990). De novo fatty acid synthesis mediated by acyl-carrier protein in Neurospora crassa mitochondria. *European Journal of Biochemistry*, *187*(2), 431–437. http://www.ncbi.nlm.nih.gov/pubmed/2137086

Mishra, P., & Chan, D. C. (2016). Metabolic regulation of mitochondrial dynamics. *Journal of Cell Biology*, *212*(4), 379–387. https://doi.org/10.1083/jcb.201511036

Mita, S., Rizzuto, R., Moraes, C. T., Shanske, S., Arnaudo, E., Fabrizi, G. M., Koga, Y., DiMauro, S., & Schon, E. A. (1990). Recombination via flanking direct repeats is a major cause of large-scale deletions of human mitochondrial DNA. *Nucleic Acids Research*, *18*(3), 561–567. http://www.ncbi.nlm.nih.gov/pubmed/2308845

Mokranjac, D., & Neupert, W. (2005). Protein import into mitochondria. *Biochemical Society Transactions*, *33*(Pt 5), 1019–1023. https://doi.org/10.1042/BST20051019

Moraes, C. T., Ricci, E., Petruzzella, V., Shanske, S., DiMauro, S., Schon, E. A., & Bonilla, E. (1992). Molecular analysis of the muscle pathology associated with mitochondrial DNA deletions. *Nature Genetics*, *1*(5), 359–367. https://doi.org/10.1038/ng0892-359

Nowinski, S. M., Solmonson, A., Rusin, S. F., Maschek, J. A., Bensard, C. L., Fogarty, S., Jeong, M. Y., Lettlova, S., Berg, J. A., Morgan, J. T., Ouyang, Y., Naylor, B. C., Paulo, J. A., Funai, K., Cox, J. E., Gygi, S. P., Winge, D. R., Deberardinis, R. J., & Rutter, J. (2020). Mitochondrial fatty acid synthesis coordinates oxidative metabolism in mammalian mitochondria. *ELife*, *9*, 1–35. https://doi.org/10.7554/ELIFE.58041

Pagliarini, D. J., Calvo, S. E., Chang, B., Sheth, S. A., Vafai, S. B., Ong, S. E., Walford, G. A., Sugiana, C., Boneh, A., Chen, W. K., Hill, D. E., Vidal, M., Evans, J. G., Thorburn, D. R., Carr, S. A., & Mootha, V. K. (2008). A Mitochondrial Protein Compendium Elucidates Complex I Disease Biology. *Cell*, *134*(1), 112–123. https://doi.org/10.1016/j.cell.2008.06.016

Pandak, W. M., Ren, S., Marques, D., Hall, E., Redford, K., Mallonee, D., Bohdan, P., Heuman, D., Gill, G., & Hylemon, P. (2002). Transport of cholesterol into mitochondria is rate-limiting for bile acid synthesis via the alternative pathway in primary rat hepatocytes. *Journal of Biological Chemistry*, *277*(50), 48158–48164. https://doi.org/10.1074/jbc.M205244200

Papatheodorou, I., Moreno, P., Manning, J., Fuentes, A. M. P., George, N., Fexova, S., Fonseca, N. A., Füllgrabe, A., Green, M., Huang, N., Huerta, L., Iqbal, H., Jianu, M., Mohammed, S., Zhao, L., Jarnuczak, A. F., Jupp, S., Marioni, J., Meyer, K., … Brazma,

A. (2020). Expression Atlas update: From tissues to single cells. *Nucleic Acids Research*, *48*(D1), D77–D83. https://doi.org/10.1093/nar/gkz947

Parkinson, H., Sarkans, U., Shojatalab, M., Abeygunawardena, N., Contrino, S., Coulson, R., Farne, A., Lara, G. G., Holloway, E., Kapushesky, M., Lilja, P., Mukherjee, G., Oezcimen, A., Rayner, T., Rocca-Serra, P., Sharma, A., Sansone, S., & Brazma, A. (2005). ArrayExpress - A public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, *33*(DATABASE ISS.). https://doi.org/10.1093/nar/gki056

Paschen, S. A., Waizenegger, T., Stan, T., Preuss, M., Cyrklaff, M., Hell, K., Rapaport, D., & Neupert, W. (2003). Evolutionary conservation of biogenesis of β-barrel membrane proteins. *Nature*, *426*(6968), 862–866. https://doi.org/10.1038/nature02208

Pegg, A. E. (2009). Mammalian polyamine metabolism and function. In *IUBMB Life* (Vol. 61, Issue 9, pp. 880–894). IUBMB Life. https://doi.org/10.1002/iub.230

Pepling, M. E., Wilhelm, J. E., O'Hara, A. L., Gephardt, G. W., & Spradling, A. C. (2007). Mouse oocytes within germ cell cysts and primordial follicles contain a Balbiani body. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(1), 187–192. https://doi.org/10.1073/pnas.0609923104

Perez-Llamas, C., & Lopez-Bigas, N. (2011). Gitools: Analysis and visualisation of genomic data using interactive heat-maps. *PLoS ONE*, *6*(5). https://doi.org/10.1371/journal.pone.0019541

Perroud, B., Ishimaru, T., Borowsky, A. D., & Weiss, R. H. (2009). Grade-dependent proteomics characterization of kidney cancer. *Molecular and Cellular Proteomics*, *8*(5), 971–985. https://doi.org/10.1074/mcp.M800252-MCP200

Pfeiffer, T., Schuster, S., & Bonhoeffer, S. (2001). Cooperation and competition in the evolution of ATP-producing pathways. *Science*, *292*(5516), 504–507. https://doi.org/10.1126/science.1058079

Porello, S. L., Cannon, M. J., & David, S. S. (1998). A substrate recognition role for the [4Fe-4S]2+ cluster of the DNA repair glycosylase MutY. *Biochemistry*, *37*(18), 6465–6475. https://doi.org/10.1021/bi972433t

Porporato, P. E., Filigheddu, N., Pedro, J. M. B. S., Kroemer, G., & Galluzzi, L. (2018). Mitochondrial metabolism and cancer. In *Cell Research* (Vol. 28, Issue 3, pp. 265–280). Nature Publishing Group. https://doi.org/10.1038/cr.2017.155

Prince, J., Jia, S., Båve, U., Annerén, G., & Oreland, L. (1994). Mitochondrial enzyme deficiencies in Down's syndrome. *Journal of Neural Transmission - Parkinson's Disease and Dementia Section*, *8*(3), 171–181. https://doi.org/10.1007/BF02260938

Qi, Q., Li, J., & Cheng, J. (2014). Reconstruction of metabolic pathways by combining probabilistic graphical model-based and knowledge-based methods. *BMC Proceedings*, *8*(Suppl 6 Proceedings of the Great Lakes Bioinformatics Confer). https://doi.org/10.1186/1753-6561-8-S6-S5

Racker, E. (1972). Bioenergetics and the problem of tumor growth. *American Scientist*, *60*(1), 56–63.

Ribas, V., García-Ruiz, C., & Fernández-Checa, J. C. (2016). Mitochondria, cholesterol and cancer cell metabolism. *Clinical and Translational Medicine*, *5*(1). https://doi.org/10.1186/s40169-016-0106-5

Roberts, N. D., Kortschak, R. D., Parker, W. T., Schreiber, A. W., Branford, S., Scott, H. S., Glonek, G., & Adelson, D. L. (2013). A comparative analysis of algorithms for somatic SNV detection in cancer. In *Bioinformatics* (Vol. 29, Issue 18, pp. 2223–2230). Bioinformatics. https://doi.org/10.1093/bioinformatics/btt375

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. In *Nature Biotechnology* (Vol. 29, Issue 1, pp. 24–26). NIH Public Access. https://doi.org/10.1038/nbt.1754

Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I., & Watson, M. (2017). A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. In *Frontiers in Genetics* (Vol. 8, Issue MAR). Frontiers Research Foundation. https://doi.org/10.3389/fgene.2017.00023

Rustom, A. (2004). Nanotubular Highways for Intercellular Organelle Transport. *Science*, *303*(5660), 1007–1010. https://doi.org/10.1126/science.1093133

Rutter, L., & Cook, D. (2020). bigPint: A Bioconductor visualization package that makes big data pint-sized. *PLoS Computational Biology*, *16*(6). https://doi.org/10.1371/journal.pcbi.1007912

Ryter, S. W., & Tyrrell, R. M. (2000). The heme synthesis and degradation pathways: role in oxidant sensitivity. Heme oxygenase has both pro- and antioxidant properties. *Free Radical Biology & Medicine*, *28*(2), 289–309. http://www.ncbi.nlm.nih.gov/pubmed/11281297

Sajnani, K., Islam, F., Smith, R. A., Gopalan, V., & Lam, A. K. Y. (2017). Genetic alterations in Krebs cycle and its impact on cancer pathogenesis. In *Biochimie* (Vol. 135, pp. 164–172). Elsevier B.V. https://doi.org/10.1016/j.biochi.2017.02.008

Sanders, E., & Diehl, S. (2015). Analysis and interpretation of transcriptomic data obtained from extended Warburg effect genes in patients with clear cell renal cell carcinoma. *Oncoscience*, *2*(2), 151–186. https://doi.org/10.18632/oncoscience.128

Sato, M., & Sato, K. (2011). Degradation of paternal mitochondria by fertilization-triggered autophagy in C. elegans embryos. *Science (New York, N.Y.)*, *334*(6059), 1141–1144. https://doi.org/10.1126/science.1210333

Schon, E. A., Dimauro, S., & Hirano, M. (2012). Human mitochondrial DNA: Roles of inherited and somatic mutations. In *Nature Reviews Genetics* (Vol. 13, Issue 12, pp. 878–890). Nat Rev Genet. https://doi.org/10.1038/nrg3275

Schroeder, M. P., Gonzalez-Perez, A., & Lopez-Bigas, N. (2013). Visualizing multidimensional cancer genomics data. In *Genome Medicine* (Vol. 5, Issue 1). Genome Med. https://doi.org/10.1186/gm413

Shadel, G. S., & Clayton, D. A. (1997). Mitochondrial DNA maintenance in vertebrates. *Annual Review of Biochemistry*, *66*, 409–435. https://doi.org/10.1146/annurev.biochem.66.1.409

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Shapiro, B. L. (1999). The Down syndrome critical region. In *Journal of Neural Transmission, Supplement* (Vol. 57, Issue 57, pp. 41–60). Springer Wien. https://doi.org/10.1007/978-3-7091-6380-1_3

Sharma, M. R., Koc, E. C., Datta, P. P., Booth, T. M., Spremulli, L. L., & Agrawal, R. K. (2003). Structure of the mammalian mitochondrial ribosome reveals an expanded functional role for its component proteins. *Cell*, *115*(1), 97–108. https://doi.org/10.1016/S0092-8674(03)00762-1

Shoffner, J. M., Fernhoff, P. M., Krawiecki, N. S., Caplan, D. B., Holt, P. J., Koontz, D. A., Takei, Y., Newman, N. J., Ortiz, R. G., Polak, M., Ballinger, S. W., Lott, M. T., & Wallace, D. C. (1992). Subacute necrotizing encephalopathy: Oxidative phosphorylation defects and the ATPase 6 point mutation. *Neurology*, *42*(11), 2168–2174. https://doi.org/10.1212/wnl.42.11.2168

Shoubridge, E. A. (1994). Mitochondrial DNA diseases: Histological and cellular studies. *Journal of Bioenergetics and Biomembranes*, *26*(3), 301–310. https://doi.org/10.1007/BF00763101

Sicheritz-Pontén, T., Kurland, C. G., & Andersson, S. G. E. (1998). A phylogenetic analysis of the cytochrome b and cytochrome c oxidase I genes supports an origin of mitochondria from within the Rickettsiaceae. *Biochimica et Biophysica Acta - Bioenergetics*, *1365*(3), 545–551. https://doi.org/10.1016/S0005-2728(98)00099-1

Sinet, P. M. (1982). Metabolism of Oxygen Derivatives in Down's Syndrome. *Annals of the New York Academy of Sciences*, *396*(1), 83–94. https://doi.org/10.1111/j.1749-6632.1982.tb26845.x

Smith, A. C., Blackshaw, J. A., & Robinson, A. J. (2012). MitoMiner: A data warehouse for mitochondrial proteomics data. *Nucleic Acids Research*, *40*(D1). https://doi.org/10.1093/nar/gkr1101

Soccio, R. E., & Breslow, J. L. (2004). Intracellular cholesterol transport. In *Arteriosclerosis, Thrombosis, and Vascular Biology* (Vol. 24, Issue 7, pp. 1150–1160). Arterioscler Thromb Vasc Biol. https://doi.org/10.1161/01.ATV.0000131264.66417.d5

Soga, M., Matsuzawa, A., & Ichijo, H. (2012). Oxidative Stress-Induced Diseases via the ASK1 Signaling Pathway. *International Journal of Cell Biology*, *2012*, 439587. https://doi.org/10.1155/2012/439587

Sorice, M., Circella, A., Cristea, I. M., Garofalo, T., di Renzo, L., Alessandri, C., Valesini, G., & Degli Esposti, M. (2004). Cardiolipin and its metabolites move from mitochondria to other cellular membranes during death receptor-mediated apoptosis. *Cell Death and Differentiation*, *11*(10), 1133–1145. https://doi.org/10.1038/sj.cdd.4401457

Spangle, J. M., Dreijerink, K. M., Groner, A. C., Cheng, H., Ohlson, C. E., Reyes, J., Lin, C. Y., Bradner, J., Zhao, J. J., Roberts, T. M., & Brown, M. (2016). PI3K/AKT Signaling Regulates H3K4 Methylation in Breast Cancer. *Cell Reports*, *15*(12), 2692–2704. https://doi.org/10.1016/j.celrep.2016.05.046

Spencer, D. H., Russler-Germain, D. A., Ketkar, S., Helton, N. M., Lamprecht, T. L., Fulton, R. S., Fronick, C. C., O'Laughlin, M., Heath, S. E., Shinawi, M., Westervelt, P., Payton, J. E., Wartman, L. D., Welch, J. S., Wilson, R. K., Walter, M. J., Link, D. C., DiPersio, J. F., & Ley, T. J. (2017). CpG Island Hypermethylation Mediated by DNMT3A Is a Consequence of AML Progression. *Cell*, *168*(5), 801-816.e13. https://doi.org/10.1016/j.cell.2017.01.021

Steffens, G. C., Biewald, R., & Buse, G. (1987). Cytochrome c oxidase is a three-copper, two-heme-A protein. *European Journal of Biochemistry*, *164*(2), 295–300. http://www.ncbi.nlm.nih.gov/pubmed/3032614

Streit, M., Lex, A., Kalkusch, M., Zatloukal, K., & Schmalstieg, D. (2009). Caleydo: Connecting pathways and gene expression. *Bioinformatics*, *25*(20), 2760–2761. https://doi.org/10.1093/bioinformatics/btp432

Sun, S. Z., & Empie, M. W. (2012). Fructose metabolism in humans - what isotopic tracer studies tell us. *Nutrition & Metabolism*, *9*(1), 89. https://doi.org/10.1186/1743-7075-9-89

Taanman, J. W. (1999). The mitochondrial genome: Structure, transcription, translation and replication. In *Biochimica et Biophysica Acta - Bioenergetics* (Vol. 1410, Issue 2, pp. 103–123). Biochim Biophys Acta. https://doi.org/10.1016/S0005-2728(98)00161-3

Taylor, S. W., Fahy, E., Zhang, B., Glenn, G. M., Warnock, D. E., Wiley, S., Murphy, A. N., Gaucher, S. P., Capaldi, R. A., Gibson, B. W., & Ghosh, S. S. (2003). Characterization of the human heart mitochondrial proteome. *Nature Biotechnology*, *21*(3), 281–286. https://doi.org/10.1038/nbt793

Tkach, M., & Théry, C. (2016). Communication by Extracellular Vesicles: Where We Are and Where We Need to Go. *Cell*, *164*(6), 1226–1232. https://doi.org/10.1016/j.cell.2016.01.043

Tomlinson, I. P. M., Alam, N. A., Rowan, A. J., Barclay, E., Jaeger, E. E. M., Kelsell, D., Leigh, I., Gorman, P., Lamlum, H., Rahman, S., Roylance, R. R., Olpin, S., Bevan, S.,

Barker, K., Hearle, N., Houlston, R. S., Kiuru, M., Lehtonen, R., Karhu, A., … Aaltonen, L. A. (2002). Germline mutations in FH predispose to dominantly inherited uterine fibroids, skin leiomyomata and papillary renal cell cancer the multiple leiomyoma consortium. *Nature Genetics*, *30*(4), 406–410. https://doi.org/10.1038/ng849

Twig, G., Elorza, A., Molina, A. J. A., Mohamed, H., Wikstrom, J. D., Walzer, G., Stiles, L., Haigh, S. E., Katz, S., Las, G., Alroy, J., Wu, M., Py, B. F., Yuan, J., Deeney, J. T., Corkey, B. E., & Shirihai, O. S. (2008). Fission and selective fusion govern mitochondrial segregation and elimination by autophagy. *EMBO Journal*, *27*(2), 433–446. https://doi.org/10.1038/sj.emboj.7601963

Tyler, D. D. (1992). *The Mitochondrion in Health and Disease*. VHC Publishers.

Tzameli, I. (2012). The evolving role of mitochondria in metabolism. In *Trends in Endocrinology and Metabolism* (Vol. 23, Issue 9, pp. 417–419). Trends Endocrinol Metab. https://doi.org/10.1016/j.tem.2012.07.008

van der Giezen, M., & Tovar, J. (2005). Degenerate mitochondria. *EMBO Reports*, *6*(6), 525–530. https://doi.org/10.1038/sj.embor.7400440

Ventura, G., de Bandt, J. P., Segaud, F., Perret, C., Robic, D., Levillain, O., le Plenier, S., Godard, C., Cynober, L., & Moinard, C. (2009). Overexpression of ornithine aminotransferase: Consequences on amino acid homeostasis. *British Journal of Nutrition*, *101*(6), 843–851. https://doi.org/10.1017/S0007114508043389

Verhagen, A. M., Ekert, P. G., Pakusch, M., Silke, J., Connolly, L. M., Reid, G. E., Moritz, R. L., Simpson, R. J., & Vaux, D. L. (2000). Identification of DIABLO, a mammalian protein that promotes apoptosis by binding to and antagonizing IAP proteins. *Cell*, *102*(1), 43–53. http://www.ncbi.nlm.nih.gov/pubmed/10929712

Vyas, S., Zaganjor, E., & Haigis, M. C. (2016). Mitochondria and Cancer. In *Cell* (Vol. 166, Issue 3, pp. 555–566). Cell Press. https://doi.org/10.1016/j.cell.2016.07.002

Wada, H., Shintani, D., & Ohlrogge, J. (1997). Why do mitochondria synthesize fatty acids? Evidence for involvement in lipoic acid production. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(4), 1591–1596. https://doi.org/10.1073/pnas.94.4.1591

Wallace, D. C. (2005). A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annual Review of Genetics*, *39*, 359–407. https://doi.org/10.1146/annurev.genet.39.110304.095751

Wallace, D. C. (2012). Mitochondria and cancer. In *Nature Reviews Cancer* (Vol. 12, Issue 10, pp. 685–698). Nat Rev Cancer. https://doi.org/10.1038/nrc3365

Wang, L. (2016). Mitochondrial purine and pyrimidine metabolism and beyond. In *Nucleosides, Nucleotides and Nucleic Acids* (Vol. 35, Issues 10–12, pp. 578–594). Taylor and Francis Inc. https://doi.org/10.1080/15257770.2015.1125001

Wang, Q., Jia, P., Li, F., Chen, H., Ji, H., Hucks, D., Dahlman, K. B., Pao, W., & Zhao, Z. (2013). Detecting somatic point mutations in cancer genome sequencing data: A comparison of mutation callers. *Genome Medicine*, *5*(10). https://doi.org/10.1186/gm495

Warburg, O., Wind, F., & Negelein, E. (1927). The metabolism of tumors in the body. *Journal of General Physiology*, *8*(6), 519–530. https://doi.org/10.1085/jgp.8.6.519

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Sander, C., Stuart, J. M., Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., Ally, A., Balasundaram, M., Birol, I., Butterfield, Y. S. N., Chu, A., … Kling, T. (2013). The cancer genome atlas pan-cancer analysis project. In *Nature Genetics* (Vol. 45, Issue 10, pp. 1113–1120). Nature Publishing Group. https://doi.org/10.1038/ng.2764

Westermann, B. (2012). Bioenergetic role of mitochondrial fusion and fission. *Biochimica et Biophysica Acta - Bioenergetics*, *1817*(10), 1833–1838. https://doi.org/10.1016/j.bbabio.2012.02.033

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*. https://doi.org/10.1038/sdata.2016.18

Woods, D. C. (2017). Mitochondrial Heterogeneity: Evaluating Mitochondrial Subpopulation Dynamics in Stem Cells. In *Stem Cells International* (Vol. 2017). Hindawi Limited. https://doi.org/10.1155/2017/7068567

Yan, H., Parsons, D. W., Jin, G., McLendon, R., Rasheed, B. A., Yuan, W., Kos, I., Batinic-Haberle, I., Jones, S., Riggins, G. J., Friedman, H., Friedman, A., Reardon, D., Herndon, J., Kinzler, K. W., Velculescu, V. E., Vogelstein, B., & Bigner, D. D. (2009). IDH1 and IDH2 Mutations in Gliomas . *New England Journal of Medicine*, *360*(8), 765–773. https://doi.org/10.1056/nejmoa0808710

Ylikallio, E., & Suomalainen, A. (2012). Mechanisms of mitochondrial diseases. *Annals of Medicine*, *44*(1), 41–59. https://doi.org/10.3109/07853890.2011.598547

Zhang, Jiayi, Tao, B., Chong, Y., Ma, S., Wu, G., Zhu, H., Zhao, Y., Zhao, S., Niu, M., Zhang, S., Wang, T., Yang, S., Qiao, W., Vuong, A. M., Li, J., Zhu, D., & Tao, W. (2020). Ornithine and breast cancer: a matched case–control study. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-72699-9

Zhang, Jinghui, Finney, R. P., Rowe, W., Edmonson, M., Sei, H. Y., Dracheva, T., Jen, J., Struewing, J. P., & Buetow, K. H. (2007). Systematic analysis of genetic alterations in

tumors using Cancer Genome WorkBench (CGWB). *Genome Research*, *17*(7), 1111–1117. https://doi.org/10.1101/gr.5963407

Zhu, J., Sanborn, J. Z., Benz, S., Szeto, C., Hsu, F., Kuhn, R. M., Karolchik, D., Archie, J., Lenburg, M. E., Esserman, L. J., Kent, W. J., Haussler, D., & Wang, T. (2009). The UCSC cancer genomics browser. In *Nature Methods* (Vol. 6, Issue 4, pp. 239–240). Nat Methods. https://doi.org/10.1038/nmeth0409-239

9 Acknowledgements

First of all, I would like to thank my supervisor Prof. Dr. Bianca Habermann for giving me the opportunity to work on this exciting topic in computational biology. Bianca has been very supportive throughout my PhD years. With her guidance, I had the chance to work at different labs and with different people, and gained valuable experiences that helped me grow.

I would also like to thank Prof. Dr. Barbara Conradt for her continuous support, both as my Doktormutter and one of the Thesis Advisory Committee members. My gratitude extends to other TAC members, Prof. Dr. Zuzana Storchova and Prof. Dr. Caroline Friedel for all their suggestions and feedback during the annual TAC meetings.

My sincere thanks to all members of the Habermann group, both in Munich and in Marseille, for providing a friendly and enjoyable working environment during my time in the lab. Thank you for always being so warm and helpful. It was great spending time with you, and having all these stimulating discussion over various topics.

Finally, I am deeply grateful for the constant support from my family and my partner during my pursue of the PhD work. Thank you for being there all the time and letting me know that you always got my back, which really means a lot to me.

10 Curriculum vitae

11 Appendix

11.1 Appendix I - List of mitochondria-associated genes in Human

Gene ID	Gene Name	Mitochondrial Process	Chromosome
AADAT	aminoadipate aminotransferase	Amino Acid Metabolism	4
AARS2	alanyl-tRNA synthetase 2, mitochondrial	Translation	6
AASS	aminoadipate-semialdehyde synthase	Amino Acid Metabolism	7
ABAT	4-aminobutyrate aminotransferase	Amino Acid Metabolism	16
ABCB10	ATP binding cassette subfamily B member 10	Transmembrane Transport	1
ABCB6	ATP binding cassette subfamily B member 6 (Langereis blood group)	Transmembrane Transport	2
ABCB7	ATP binding cassette subfamily B member 7	Transmembrane Transport	X
ABCB8	ATP binding cassette subfamily B member 8	Transmembrane Transport	7
ABCD1	ATP binding cassette subfamily D member 1	Metabolism of Lipids & Lipoproteins	X
ABCD3	ATP binding cassette subfamily D member 3	Transmembrane Transport	1
ABCE1	ATP binding cassette subfamily E member 1	Transmembrane Transport	4
ABCF2	ATP binding cassette subfamily F member 2	Transmembrane Transport	7
ABHD10	abhydrolase domain containing 10	Amino Acid Metabolism	3
ABHD11	abhydrolase domain containing 11	Amino Acid Metabolism	7
ABHD16A	abhydrolase domain containing 16A	Amino Acid Metabolism	6
ACAA2	acetyl-CoA acyltransferase 2	Fatty Acid Degradation & Beta-oxidation	18
ACACA	acetyl-CoA carboxylase alpha	Metabolism of Lipids & Lipoproteins	17
ACACB	acetyl-CoA carboxylase beta	Fatty Acid Degradation & Beta-oxidation	12
ACAD10	acyl-CoA dehydrogenase family member 10	Fatty Acid Degradation & Beta-oxidation	12
ACAD11	acyl-CoA dehydrogenase family member 11	Fatty Acid Degradation & Beta-oxidation	3
ACAD8	acyl-CoA dehydrogenase family member 8	Fatty Acid Degradation & Beta-oxidation	11
ACAD9	acyl-CoA dehydrogenase family member 9	Fatty Acid Degradation & Beta-oxidation	3
ACADL	acyl-CoA dehydrogenase long chain	Fatty Acid Degradation & Beta-oxidation	2
ACADM	acyl-CoA dehydrogenase medium chain	Fatty Acid Degradation & Beta-oxidation	1
ACADS	acyl-CoA dehydrogenase short chain	Fatty Acid Degradation & Beta-oxidation	12
ACADSB	acyl-CoA dehydrogenase short/branched chain	Fatty Acid Degradation & Beta-oxidation	10
ACADVL	acyl-CoA dehydrogenase very long chain	Fatty Acid Degradation & Beta-oxidation	17
ACAT1	acetyl-CoA acetyltransferase 1	Amino Acid Metabolism	11
ACAT2	acetyl-CoA acetyltransferase 2	Fatty Acid Degradation & Beta-oxidation	6
ACO1	aconitase 1	Tricarboxylic Acid Cycle	9
ACO2	aconitase 2	Tricarboxylic Acid Cycle	22
ACOT2	acyl-CoA thioesterase 2	Fatty Acid Biosynthesis & Elongation	14
ACOT7	acyl-CoA thioesterase 7	Fatty Acid Biosynthesis & Elongation	1
ACOT9	acyl-CoA thioesterase 9	Fatty Acid Metabolism	X
ACP6	acid phosphatase 6, lysophosphatidic	Metabolism of Lipids & Lipoproteins	1
ACSBG2	acyl-CoA synthetase bubblegum family member 2	Fatty Acid Degradation & Beta-oxidation	19
ACSF2	acyl-CoA synthetase family member 2	Fatty Acid Metabolism	17
ACSF3	acyl-CoA synthetase family member 3	Fatty Acid Metabolism	16
ACSL1	acyl-CoA synthetase long chain family member 1	Fatty Acid Biosynthesis & Elongation	4
ACSL3	acyl-CoA synthetase long chain family member 3	Fatty Acid Degradation & Beta-oxidation	2
ACSL4	acyl-CoA synthetase long chain family member 4	Fatty Acid Degradation & Beta-oxidation	X
ACSL5	acyl-CoA synthetase long chain family member 5	Fatty Acid Biosynthesis & Elongation	10
ACSL6	acyl-CoA synthetase long chain family member 6	Fatty Acid Biosynthesis & Elongation	5
ACSM1	acyl-CoA synthetase medium chain family member 1	Fatty Acid Biosynthesis & Elongation	16
ACSM2A	acyl-CoA synthetase medium chain family member 2A	Fatty Acid Biosynthesis & Elongation	16

ACSM2B	acyl-CoA synthetase medium chain family member 2B	Fatty Acid Biosynthesis & Elongation	16
ACSM3	acyl-CoA synthetase medium chain family member 3	Fatty Acid Biosynthesis & Elongation	16
ACSM4	acyl-CoA synthetase medium chain family member 4	Fatty Acid Biosynthesis & Elongation	12
ACSM5	acyl-CoA synthetase medium chain family member 5	Fatty Acid Biosynthesis & Elongation	16
ACSM6	acyl-CoA synthetase medium chain family member 6	Fatty Acid Metabolism	10
ACSS1	acyl-CoA synthetase short chain family member 1	Pyruvate Metabolism	20
ACSS2	acyl-CoA synthetase short chain family member 2	Pyruvate Metabolism	20
ACSS3	acyl-CoA synthetase short chain family member 3	Metabolism of Lipids & Lipoproteins	12
ACTR2	actin related protein 2	Mitochondrial Dynamics	2
ADCK1	aarF domain containing kinase 1	Oxidative Phosphorylation	14
ADHFE1	alcohol dehydrogenase iron containing 1	Pyruvate Metabolism	8
ADO	2-aminoethanethiol dioxygenase	Amino Acid Metabolism	10
AFG1L	AFG1 like ATPase	Protein Stability & Degradation	6
AFG3L2	AFG3 like matrix AAA peptidase subunit 2	Mitochondrial Dynamics	18
AGK	acylglycerol kinase	Fatty Acid Metabolism	7
AGMAT	agmatinase	Amino Acid Metabolism	1
AGPAT5	1-acylglycerol-3-phosphate O-acyltransferase 5	Metabolism of Lipids & Lipoproteins	8
AGTPBP1	ATP/GTP binding protein 1	Mitochondrial Dynamics	9
AGXT	alanine--glyoxylate and serine--pyruvate aminotransferase	Amino Acid Metabolism	2
AGXT2	alanine--glyoxylate aminotransferase 2	Amino Acid Metabolism	5
AIFM1	apoptosis inducing factor mitochondria associated 1	Apoptosis	X
AIFM2	apoptosis inducing factor mitochondria associated 2	Apoptosis	10
AIFM3	apoptosis inducing factor mitochondria associated 3	Apoptosis	22
AK2	adenylate kinase 2	Nucleotide Metabolism	1
AK3	adenylate kinase 3	Nucleotide Metabolism	9
AK4	adenylate kinase 4	Nucleotide Metabolism	1
AKAP1	A-kinase anchoring protein 1	Mitochondrial Signaling	17
AKAP10	A-kinase anchoring protein 10	Mitochondrial Signaling	17
AKR1B1	aldo-keto reductase family 1 member B	Fructose Metabolism	7
AKR1B10	aldo-keto reductase family 1 member B10	Metabolism of Vitamins & Co-Factors	7
AKR1B15	aldo-keto reductase family 1 member B15	Metabolism of Lipids & Lipoproteins	7
AKT1	AKT serine/threonine kinase 1	Mitochondrial Signaling	14
ALAS1	5'-aminolevulinate synthase 1	Heme Biosynthesis	3
ALAS2	5'-aminolevulinate synthase 2	Heme Biosynthesis	X
ALDH18A1	aldehyde dehydrogenase 18 family member A1	Amino Acid Metabolism	10
ALDH1A1	aldehyde dehydrogenase 1 family member A1	Fructose Metabolism	9
ALDH1B1	aldehyde dehydrogenase 1 family member B1	Fatty Acid Degradation & Beta-oxidation	9
ALDH1L1	aldehyde dehydrogenase 1 family member L1	Folate & Pterin Metabolism	3
ALDH1L2	aldehyde dehydrogenase 1 family member L2	Folate & Pterin Metabolism	12
ALDH2	aldehyde dehydrogenase 2 family member	Fatty Acid Degradation & Beta-oxidation	12
ALDH3A2	aldehyde dehydrogenase 3 family member A2	Fatty Acid Degradation & Beta-oxidation	17
ALDH4A1	aldehyde dehydrogenase 4 family member A1	Amino Acid Metabolism	1
ALDH5A1	aldehyde dehydrogenase 5 family member A1	Amino Acid Metabolism	6
ALDH6A1	aldehyde dehydrogenase 6 family member A1	Amino Acid Metabolism	14
ALDH7A1	aldehyde dehydrogenase 7 family member A1	Pyruvate Metabolism	5
ALDOA	aldolase, fructose-bisphosphate A	Glycolysis	16

ALDOB	aldolase, fructose-bisphosphate B	Fructose Metabolism	9
ALDOC	aldolase, fructose-bisphosphate C	Glycolysis	17
ALKBH1	alkB homolog 1, histone H2A dioxygenase	Replication & Transcription	14
ALKBH7	alkB homolog 7	Apoptosis	19
AMACR	alpha-methylacyl-CoA racemase	Bile Acid Synthesis	5
AMBRA1	autophagy and beclin 1 regulator 1	Mitophagy	11
AMT	aminomethyltransferase	Amino Acid Metabolism	3
ANP32A	acidic nuclear phosphoprotein 32 family member A	Apoptosis	15
APAF1	apoptotic peptidase activating factor 1	Apoptosis	12
APEX1	apurinic/apyrimidinic endodeoxyribonuclease 1	Replication & Transcription	14
APEX2	apurinic/apyrimidinic endodeoxyribonuclease 2	Replication & Transcription	X
APOO	apolipoprotein O	Mitochondrial Dynamics	X
APOOL	apolipoprotein O like	Mitochondrial Dynamics	X
APOPT1	apoptogenic 1, mitochondrial	Apoptosis	14
ARG1	arginase 1	Nitrogen Metabolism	6
ARG2	arginase 2	Nitrogen Metabolism	14
ARMC10	armadillo repeat containing 10	Apoptosis	7
ARNT	aryl hydrocarbon receptor nuclear translocator	Transcription (nuclear)	1
ARPC3	actin related protein 2/3 complex subunit 3	Mitochondrial Dynamics	12
ARPC5	actin related protein 2/3 complex subunit 5	Mitochondrial Dynamics	1
ARPC5L	actin related protein 2/3 complex subunit 5 like	Mitochondrial Dynamics	9
ASAH2	N-acylsphingosine amidohydrolase 2	Metabolism of Lipids & Lipoproteins	10
ASAH2B	N-acylsphingosine amidohydrolase 2B	Metabolism of Lipids & Lipoproteins	10
ASL	argininosuccinate lyase	Nitrogen Metabolism	7
ASS1	argininosuccinate synthase 1	Amino Acid Metabolism	9
ATAD1	ATPase family AAA domain containing 1	Protein Stability & Degradation	10
ATAD3A	ATPase family AAA domain containing 3A	Mitochondrial Dynamics	1
ATAD3B	ATPase family AAA domain containing 3B	Mitochondrial Dynamics	1
ATF4	activating transcription factor 4	Transcription (nuclear)	22
ATF5	activating transcription factor 5	UPRmt	19
ATG5	autophagy related 5	Mitophagy	6
ATG9B	autophagy related 9B	Mitophagy	7
ATIC	5-aminoimidazole-4-carboxamide ribonucleotide formyltransferase/IMP cyclohydrolase	Nucleotide Metabolism	2
ATP23	ATP23 metallopeptidase and ATP synthase assembly factor homolog	Protein Stability & Degradation	12
ATP5F1A	ATP synthase F1 subunit alpha	Oxidative Phosphorylation	18
ATP5F1B	ATP synthase F1 subunit beta	Oxidative Phosphorylation	12
ATP5F1C	ATP synthase F1 subunit gamma	Oxidative Phosphorylation	10
ATP5F1D	ATP synthase F1 subunit delta	Oxidative Phosphorylation	19
ATP5F1E	ATP synthase F1 subunit epsilon	Oxidative Phosphorylation	20
ATP5IF1	ATP synthase inhibitory factor subunit 1	Oxidative Phosphorylation	1
ATP5MC1	ATP synthase membrane subunit c locus 1	Oxidative Phosphorylation	17
ATP5MC2	ATP synthase membrane subunit c locus 2	Oxidative Phosphorylation	12
ATP5MC3	ATP synthase membrane subunit c locus 3	Oxidative Phosphorylation	2
ATP5MD	ATP synthase membrane subunit DAPIT	Oxidative Phosphorylation	10
ATP5ME	ATP synthase membrane subunit e	Oxidative Phosphorylation	4
ATP5MF	ATP synthase membrane subunit f	Oxidative Phosphorylation	7
ATP5MG	ATP synthase membrane subunit g	Oxidative Phosphorylation	11
ATP5MGL	ATP synthase membrane subunit g like	Oxidative Phosphorylation	22
ATP5MPL	ATP synthase membrane subunit 6.8PL	Oxidative Phosphorylation	14
ATP5PB	ATP synthase peripheral stalk-membrane subunit b	Oxidative Phosphorylation	1
ATP5PD	ATP synthase peripheral stalk subunit d	Oxidative Phosphorylation	17

ATP5PF	ATP synthase peripheral stalk subunit F6	Oxidative Phosphorylation	21
ATP5PO	ATP synthase peripheral stalk subunit OSCP	Oxidative Phosphorylation	21
ATP6V1E1	ATPase H+ transporting V1 subunit E1	Oxidative Phosphorylation	22
ATPAF1	ATP synthase mitochondrial F1 complex assembly factor 1	Oxidative Phosphorylation	1
ATPAF2	ATP synthase mitochondrial F1 complex assembly factor 2	Oxidative Phosphorylation	17
AUH	AU RNA binding methylglutaconyl-CoA hydratase	Amino Acid Metabolism	9
AURKAIP1	aurora kinase A interacting protein 1	Translation	1
BAD	BCL2 associated agonist of cell death	Apoptosis	11
BAK1	BCL2 antagonist/killer 1	Apoptosis	6
BAX	BCL2 associated X, apoptosis regulator	Apoptosis	19
BBC3	BCL2 binding component 3	Apoptosis	19
BCAT1	branched chain amino acid transaminase 1	Amino Acid Metabolism	12
BCAT2	branched chain amino acid transaminase 2	Amino Acid Metabolism	19
BCKDHA	branched chain keto acid dehydrogenase E1, alpha polypeptide	Amino Acid Metabolism	19
BCKDHB	branched chain keto acid dehydrogenase E1 subunit beta	Amino Acid Metabolism	6
BCKDK	branched chain ketoacid dehydrogenase kinase	Amino Acid Metabolism	16
BCL2	BCL2 apoptosis regulator	Apoptosis	18
BCL2L1	BCL2 like 1	Apoptosis	20
BCL2L11	BCL2 like 11	Apoptosis	2
BCL2L13	BCL2 like 13	Apoptosis	22
BCL2L2	BCL2 like 2	Apoptosis	14
BCS1L	BCS1 homolog, ubiquinol-cytochrome c reductase complex chaperone	Oxidative Phosphorylation	2
BDH1	3-hydroxybutyrate dehydrogenase 1	Fatty Acid Metabolism	3
BECN1	beclin 1	Mitophagy	17
BECN2	beclin 2	Mitophagy	1
BID	BH3 interacting domain death agonist	Apoptosis	22
BIK	BCL2 interacting killer	Apoptosis	22
BIRC8	baculoviral IAP repeat containing 8	Apoptosis	19
BNIP1	BCL2 interacting protein 1	Apoptosis	5
BNIP3	BCL2 interacting protein 3	Apoptosis	10
BNIP3L	BCL2 interacting protein 3 like	Apoptosis	8
BOK	BCL2 family apoptosis regulator BOK	Apoptosis	2
BOLA1	bolA family member 1	Fe-S Cluster Biosynthesis	1
BOLA3	bolA family member 3	Fe-S Cluster Biosynthesis	2
BPHL	biphenyl hydrolase like	Translation	6
C12orf10	chromosome 12 open reading frame 10	Unknown	12
C12orf65	chromosome 12 open reading frame 65	Translation	12
C15orf48	chromosome 15 open reading frame 48	Oxidative Phosphorylation	15
C16orf91	chromosome 16 open reading frame 91	Unknown	16
C5orf63	chromosome 5 open reading frame 63	Unknown	5
C8orf82	chromosome 8 open reading frame 82	Unknown	8
CA5A	carbonic anhydrase 5A	Nitrogen Metabolism	16
CA5B	carbonic anhydrase 5B	Nitrogen Metabolism	X
CALCOCO2	calcium binding and coiled-coil domain 2	Mitophagy	17
CARS2	cysteinyl-tRNA synthetase 2, mitochondrial	Translation	13
CASP2	caspase 2	Apoptosis	7
CASP3	caspase 3	Apoptosis	4
CASP7	caspase 7	Apoptosis	10
CASP8	caspase 8	Apoptosis	2
CASP9	caspase 9	Apoptosis	1
CAT	catalase	ROS Defense	11
CBR4	carbonyl reductase 4	Fatty Acid Biosynthesis & Elongation	4
CCDC58	coiled-coil domain containing 58	Translation	3

CCDC90B	coiled-coil domain containing 90B	Calcium Signaling & Transport	11
CDK5RAP1	CDK5 regulatory subunit associated protein 1	Replication & Transcription	20
CDS1	CDP-diacylglycerol synthase 1	Cardiolipin Biosynthesis	4
CDS2	CDP-diacylglycerol synthase 2	Metabolism of Lipids & Lipoproteins	20
CHCHD1	coiled-coil-helix-coiled-coil-helix domain containing 1	Translation	10
CHCHD10	coiled-coil-helix-coiled-coil-helix domain containing 10	Mitochondrial Dynamics	22
CHCHD2	coiled-coil-helix-coiled-coil-helix domain containing 2	Mitochondrial Signaling	7
CHCHD3	coiled-coil-helix-coiled-coil-helix domain containing 3	Mitochondrial Dynamics	7
CHCHD4	coiled-coil-helix-coiled-coil-helix domain containing 4	Import & Sorting	3
CHCHD5	coiled-coil-helix-coiled-coil-helix domain containing 5	Import & Sorting	2
CHCHD6	coiled-coil-helix-coiled-coil-helix domain containing 6	Mitochondrial Dynamics	3
CHCHD7	coiled-coil-helix-coiled-coil-helix domain containing 7	Unknown	8
CHDH	choline dehydrogenase	Amino Acid Metabolism	3
CIAO1	cytosolic iron-sulfur assembly component 1	Fe-S Cluster Biosynthesis	2
CIAO2A	cytosolic iron-sulfur assembly component 2A	Fe-S Cluster Biosynthesis	15
CIAO2B	cytosolic iron-sulfur assembly component 2B	Fe-S Cluster Biosynthesis	16
CIAPIN1	cytokine induced apoptosis inhibitor 1	Fe-S Cluster Biosynthesis	16
CISD1	CDGSH iron sulfur domain 1	Oxidative Phosphorylation	10
CISD2	CDGSH iron sulfur domain 2	Oxidative Phosphorylation	4
CISD3	CDGSH iron sulfur domain 3	Oxidative Phosphorylation	17
CITED2	Cbp/p300 interacting transactivator with Glu/Asp rich carboxy-terminal domain 2	Mitophagy	6
CKMT1A	creatine kinase, mitochondrial 1A	Amino Acid Metabolism	15
CKMT1B	creatine kinase, mitochondrial 1B	Amino Acid Metabolism	15
CKMT2	creatine kinase, mitochondrial 2	Amino Acid Metabolism	5
CLIC4	chloride intracellular channel 4	Apoptosis	1
CLPB	ClpB homolog, mitochondrial AAA ATPase chaperonin	Protein Stability & Degradation	11
CLPP	caseinolytic mitochondrial matrix peptidase proteolytic subunit	Protein Stability & Degradation	19
CLPX	caseinolytic mitochondrial matrix peptidase chaperone subunit	Protein Stability & Degradation	15
CLYBL	citrate lyase beta like	Metabolism of Vitamins & Co-Factors	13
CMC1	C-X9-C motif containing 1	Oxidative Phosphorylation	3
CMC2	C-X9-C motif containing 2	Import & Sorting	16
CMC4	C-X9-C motif containing 4	Import & Sorting	X
CMPK2	cytidine/uridine monophosphate kinase 2	Replication & Transcription	2
COA1	cytochrome c oxidase assembly factor 1 homolog	Oxidative Phosphorylation	7
COA3	cytochrome c oxidase assembly factor 3	Oxidative Phosphorylation	17
COA4	cytochrome c oxidase assembly factor 4 homolog	Oxidative Phosphorylation	11
COA5	cytochrome c oxidase assembly factor 5	Oxidative Phosphorylation	2
COA6	cytochrome c oxidase assembly factor 6	Oxidative Phosphorylation	1
COA7	cytochrome c oxidase assembly factor 7 (putative)	Oxidative Phosphorylation	1
COASY	Coenzyme A synthase	Metabolism of Vitamins & Co-Factors	17
COQ10A	coenzyme Q10A	Oxidative Phosphorylation	12
COQ10B	coenzyme Q10B	Oxidative Phosphorylation	2
COQ2	coenzyme Q2, polyprenyltransferase	Ubiquinone Biosynthesis	4
COQ3	coenzyme Q3, methyltransferase	Ubiquinone Biosynthesis	6
COQ4	coenzyme Q4	Ubiquinone Biosynthesis	9
COQ5	coenzyme Q5, methyltransferase	Ubiquinone Biosynthesis	12
COQ6	coenzyme Q6, monooxygenase	Ubiquinone Biosynthesis	14
COQ7	coenzyme Q7, hydroxylase	Ubiquinone Biosynthesis	16
COQ8A	coenzyme Q8A	Oxidative Phosphorylation	1

COQ8B	coenzyme Q8B	Oxidative Phosphorylation	19
COQ9	coenzyme Q9	Ubiquinone Biosynthesis	16
COX10	cytochrome c oxidase assembly factor heme A:farnesyltransferase COX10	Oxidative Phosphorylation	17
COX11	cytochrome c oxidase copper chaperone COX11	Oxidative Phosphorylation	17
COX14	cytochrome c oxidase assembly factor COX14	Oxidative Phosphorylation	12
COX15	cytochrome c oxidase assembly homolog COX15	Oxidative Phosphorylation	10
COX16	cytochrome c oxidase assembly factor COX16	Oxidative Phosphorylation	14
COX17	cytochrome c oxidase copper chaperone COX17	Oxidative Phosphorylation	3
COX18	cytochrome c oxidase assembly factor COX18	Oxidative Phosphorylation	4
COX19	cytochrome c oxidase assembly factor COX19	Oxidative Phosphorylation	7
COX20	cytochrome c oxidase assembly factor COX20	Oxidative Phosphorylation	1
COX4I1	cytochrome c oxidase subunit 4I1	Oxidative Phosphorylation	16
COX4I2	cytochrome c oxidase subunit 4I2	Oxidative Phosphorylation	20
COX5A	cytochrome c oxidase subunit 5A	Oxidative Phosphorylation	15
COX5B	cytochrome c oxidase subunit 5B	Oxidative Phosphorylation	2
COX6A1	cytochrome c oxidase subunit 6A1	Oxidative Phosphorylation	12
COX6A2	cytochrome c oxidase subunit 6A2	Oxidative Phosphorylation	16
COX6B1	cytochrome c oxidase subunit 6B1	Oxidative Phosphorylation	19
COX6B2	cytochrome c oxidase subunit 6B2	Oxidative Phosphorylation	19
COX6C	cytochrome c oxidase subunit 6C	Oxidative Phosphorylation	8
COX7A1	cytochrome c oxidase subunit 7A1	Oxidative Phosphorylation	19
COX7A2	cytochrome c oxidase subunit 7A2	Oxidative Phosphorylation	6
COX7A2L	cytochrome c oxidase subunit 7A2 like	Oxidative Phosphorylation	2
COX7B	cytochrome c oxidase subunit 7B	Oxidative Phosphorylation	X
COX7B2	cytochrome c oxidase subunit 7B2	Oxidative Phosphorylation	4
COX7C	cytochrome c oxidase subunit 7C	Oxidative Phosphorylation	5
COX8A	cytochrome c oxidase subunit 8A	Oxidative Phosphorylation	11
COX8BP	cytochrome c oxidase subunit 8B, pseudogene	Oxidative Phosphorylation	11
COX8C	cytochrome c oxidase subunit 8C	Oxidative Phosphorylation	14
CPOX	coproporphyrinogen oxidase	Heme Biosynthesis	3
CPS1	carbamoyl-phosphate synthase 1	Nitrogen Metabolism	2
CPT1A	carnitine palmitoyltransferase 1A	Fatty Acid Degradation & Beta-oxidation	11
CPT1B	carnitine palmitoyltransferase 1B	Fatty Acid Degradation & Beta-oxidation	22
CPT1C	carnitine palmitoyltransferase 1C	Fatty Acid Degradation & Beta-oxidation	19
CPT2	carnitine palmitoyltransferase 2	Fatty Acid Degradation & Beta-oxidation	1
CRAT	carnitine O-acetyltransferase	Fatty Acid Degradation & Beta-oxidation	9
CREB1	cAMP responsive element binding protein 1	Transcription (nuclear)	2
CRLS1	cardiolipin synthase 1	Cardiolipin Biosynthesis	20
CS	citrate synthase	Tricarboxylic Acid Cycle	12
CSNK2A1	casein kinase 2 alpha 1	Apoptosis	20
CTSB	cathepsin B	Apoptosis	8
CYB5A	cytochrome b5 type A	Metabolism of Vitamins & Co-Factors	18
CYB5B	cytochrome b5 type B	Metabolism of Vitamins & Co-Factors	16
CYB5R1	cytochrome b5 reductase 1	Metabolism of Lipids & Lipoproteins	1
CYB5R2	cytochrome b5 reductase 2	Metabolism of Lipids & Lipoproteins	11
CYB5R3	cytochrome b5 reductase 3	Metabolism of Lipids & Lipoproteins	22
CYC1	cytochrome c1	Oxidative Phosphorylation	8
CYCS	cytochrome c, somatic	Oxidative Phosphorylation	7
CYP11A1	cytochrome P450 family 11 subfamily A member 1	Metabolism of Lipids & Lipoproteins	15
CYP11B1	cytochrome P450 family 11 subfamily B member 1	Metabolism of Lipids & Lipoproteins	8

CYP11B2	cytochrome P450 family 11 subfamily B member 2	Metabolism of Lipids & Lipoproteins	8
CYP24A1	cytochrome P450 family 24 subfamily A member 1	Metabolism of Vitamins & Co-Factors	20
CYP27A1	cytochrome P450 family 27 subfamily A member 1	Metabolism of Lipids & Lipoproteins	2
CYP27B1	cytochrome P450 family 27 subfamily B member 1	Metabolism of Vitamins & Co-Factors	12
D2HGDH	D-2-hydroxyglutarate dehydrogenase	Oxidative Phosphorylation	2
DAP3	death associated protein 3	Translation	1
DARS2	aspartyl-tRNA synthetase 2, mitochondrial	Translation	1
DBT	dihydrolipoamide branched chain transacylase E2	Amino Acid Metabolism	1
DDAH1	dimethylarginine dimethylaminohydrolase 1	Nitrogen Metabolism	1
DDIT3	DNA damage inducible transcript 3	UPRmt	12
DDX28	DEAD-box helicase 28	Translation	16
DECR1	2,4-dienoyl-CoA reductase 1	Fatty Acid Degradation & Beta-oxidation	8
DELE1	DAP3 binding cell death enhancer 1	Apoptosis	5
DGLUCY	D-glutamate cyclase	Amino Acid Metabolism	14
DGUOK	deoxyguanosine kinase	Replication & Transcription	2
DHFR	dihydrofolate reductase	Folate & Pterin Metabolism	5
DHFR2	dihydrofolate reductase 2	Folate & Pterin Metabolism	3
DHODH	dihydroorotate dehydrogenase (quinone)	Nucleotide Metabolism	16
DHRS2	dehydrogenase/reductase 2	Metabolism of Vitamins & Co-Factors	14
DHRS4	dehydrogenase/reductase 4	Metabolism of Vitamins & Co-Factors	14
DHTKD1	dehydrogenase E1 and transketolase domain containing 1	Amino Acid Metabolism	10
DHX29	DExH-box helicase 29	Replication & Transcription	5
DHX30	DExH-box helicase 30	Replication & Transcription	3
DIABLO	diablo IAP-binding mitochondrial protein	Apoptosis	12
DLAT	dihydrolipoamide S-acetyltransferase	Pyruvate Metabolism	11
DLD	dihydrolipoamide dehydrogenase	Tricarboxylic Acid Cycle	7
DLST	dihydrolipoamide S-succinyltransferase	Tricarboxylic Acid Cycle	14
DMAC1	distal membrane arm assembly complex 1	Oxidative Phosphorylation	9
DMAC2	distal membrane arm assembly complex 2	Oxidative Phosphorylation	19
DMAC2L	distal membrane arm assembly complex 2 like	Oxidative Phosphorylation	14
DMGDH	dimethylglycine dehydrogenase	Metabolism of Lipids & Lipoproteins	5
DNA2	DNA replication helicase/nuclease 2	Replication & Transcription	10
DNAJA1	DnaJ heat shock protein family (Hsp40) member A1	Protein Stability & Degradation	12
DNAJA3	DnaJ heat shock protein family (Hsp40) member A3	Protein Stability & Degradation	16
DNAJC11	DnaJ heat shock protein family (Hsp40) member C11	Protein Stability & Degradation	1
DNAJC15	DnaJ heat shock protein family (Hsp40) member C15	Import & Sorting	13
DNAJC19	DnaJ heat shock protein family (Hsp40) member C19	Import & Sorting	3
DNAJC30	DnaJ heat shock protein family (Hsp40) member C30	Protein Stability & Degradation	7
DNAJC4	DnaJ heat shock protein family (Hsp40) member C4	Protein Stability & Degradation	11
DNLZ	DNL-type zinc finger	Import & Sorting	9
DNM1L	dynamin 1 like	Mitochondrial Dynamics	12
DONSON	downstream neighbor of SON	Oxidative Phosphorylation	21
DTYMK	deoxythymidylate kinase	Nucleotide Metabolism	2
DUT	deoxyuridine triphosphatase	Nucleotide Metabolism	15
DYNLL1	dynein light chain LC8-type 1	Apoptosis	12
E2F1	E2F transcription factor 1	Transcription (nuclear)	20
EARS2	glutamyl-tRNA synthetase 2, mitochondrial	Translation	16
ECH1	enoyl-CoA hydratase 1	Fatty Acid Biosynthesis & Elongation	19
ECHDC1	ethylmalonyl-CoA decarboxylase 1	Fatty Acid Biosynthesis & Elongation	6
ECHDC2	enoyl-CoA hydratase domain containing 2	Fatty Acid Biosynthesis & Elongation	1

ECHDC3	enoyl-CoA hydratase domain containing 3	Fatty Acid Biosynthesis & Elongation	10
ECHS1	enoyl-CoA hydratase, short chain 1	Fatty Acid Biosynthesis & Elongation	10
ECI1	enoyl-CoA delta isomerase 1	Fatty Acid Degradation & Beta-oxidation	16
ECI2	enoyl-CoA delta isomerase 2	Fatty Acid Degradation & Beta-oxidation	6
ECSIT	ECSIT signalling integrator	Oxidative Phosphorylation	19
EFHD1	EF-hand domain family member D1	Calcium Signaling & Transport	2
EIF2AK3	eukaryotic translation initiation factor 2 alpha kinase 3	Translation	2
ELAC2	elaC ribonuclease Z 2	Translation	17
ENDOG	endonuclease G	Replication & Transcription	9
ENO1	enolase 1	Glycolysis	1
ENO2	enolase 2	Glycolysis	12
ENO3	enolase 3	Glycolysis	17
ERAL1	Era like 12S mitochondrial rRNA chaperone 1	Translation	17
ESRRA	estrogen related receptor alpha	Transcription (nuclear)	11
ETFA	electron transfer flavoprotein subunit alpha	Oxidative Phosphorylation	15
ETFB	electron transfer flavoprotein subunit beta	Oxidative Phosphorylation	19
ETFDH	electron transfer flavoprotein dehydrogenase	Oxidative Phosphorylation	4
ETFRF1	electron transfer flavoprotein regulatory factor 1	Oxidative Phosphorylation	12
ETHE1	ETHE1 persulfide dioxygenase	Amino Acid Metabolism	19
ETNPPL	ethanolamine-phosphate phospho-lyase	Amino Acid Metabolism	4
EXOG	exo/endonuclease G	Replication & Transcription	3
FABP6	fatty acid binding protein 6	Metabolism of Lipids & Lipoproteins	5
FAHD1	fumarylacetoacetate hydrolase domain containing 1	Amino Acid Metabolism	16
FAHD2A	fumarylacetoacetate hydrolase domain containing 2A	Amino Acid Metabolism	2
FAHD2B	fumarylacetoacetate hydrolase domain containing 2B	Amino Acid Metabolism	2
FAM162A	family with sequence similarity 162 member A	Apoptosis	3
FAM210B	family with sequence similarity 210 member B	Transmembrane Transport	20
FARS2	phenylalanyl-tRNA synthetase 2, mitochondrial	Translation	6
FASTK	Fas activated serine/threonine kinase	Translation	7
FASTKD1	FAST kinase domains 1	Translation	2
FASTKD2	FAST kinase domains 2	Translation	2
FASTKD3	FAST kinase domains 3	Translation	5
FASTKD5	FAST kinase domains 5	Translation	20
FBP1	fructose-bisphosphatase 1	Glycolysis	9
FBP2	fructose-bisphosphatase 2	Glycolysis	9
FBXL4	F-box and leucine rich repeat protein 4	Protein Stability & Degradation	6
FDPS	farnesyl diphosphate synthase	Metabolism of Lipids & Lipoproteins	1
FDX1	ferredoxin 1	Fe-S Cluster Biosynthesis	11
FDX2	ferredoxin 2	Fe-S Cluster Biosynthesis	19
FDXR	ferredoxin reductase	Fe-S Cluster Biosynthesis	17
FECH	ferrochelatase	Heme Biosynthesis	18
FEN1	flap structure-specific endonuclease 1	Replication & Transcription	11
FH	fumarate hydratase	Tricarboxylic Acid Cycle	1
FIS1	fission, mitochondrial 1	Mitochondrial Dynamics	7
FKBP8	FKBP prolyl isomerase 8	Protein Stability & Degradation	19
FMC1	formation of mitochondrial complex V assembly factor 1 homolog	Oxidative Phosphorylation	7
FOXO3	forkhead box O3	Transcription (nuclear)	6
FOXRED1	FAD dependent oxidoreductase domain containing 1	Oxidative Phosphorylation	11
FPGS	folylpolyglutamate synthase	Folate & Pterin Metabolism	9
FTMT	ferritin mitochondrial	Heme Biosynthesis	5
FUNDC1	FUN14 domain containing 1	Mitochondrial Dynamics	X

FUNDC2	FUN14 domain containing 2	Mitochondrial Dynamics	X
FXN	frataxin	Fe-S Cluster Biosynthesis	9
G6PC	glucose-6-phosphatase catalytic subunit	Glycolysis	17
G6PD	glucose-6-phosphate dehydrogenase	Pentose Phosphate Pathway	X
GABARAP	GABA type A receptor-associated protein	Mitophagy	17
GABPA	GA binding protein transcription factor subunit alpha	Transcription (nuclear)	21
GADD45GIP1	GADD45G interacting protein 1	Translation	19
GAPDH	glyceraldehyde-3-phosphate dehydrogenase	Glycolysis	12
GAPDHS	glyceraldehyde-3-phosphate dehydrogenase, spermatogenic	Glycolysis	19
GARS1	glycyl-tRNA synthetase	Translation	7
GATB	glutamyl-tRNA amidotransferase subunit B	Translation	4
GATC	glutamyl-tRNA amidotransferase subunit C	Translation	12
GATD3A	glutamine amidotransferase like class 1 domain containing 3A	Unknown	21
GATM	glycine amidinotransferase	Amino Acid Metabolism	15
GCAT	glycine C-acetyltransferase	Amino Acid Metabolism	22
GCDH	glutaryl-CoA dehydrogenase	Fatty Acid Degradation & Beta-oxidation	19
GCK	glucokinase	Glycolysis	7
GCLC	glutamate-cysteine ligase catalytic subunit	ROS Defense	6
GCLM	glutamate-cysteine ligase modifier subunit	ROS Defense	1
GCSH	glycine cleavage system protein H	Amino Acid Metabolism	16
GDAP1	ganglioside induced differentiation associated protein 1	Mitochondrial Dynamics	8
GFER	growth factor, augmenter of liver regeneration	Apoptosis	16
GFM1	G elongation factor mitochondrial 1	Translation	3
GFM2	G elongation factor mitochondrial 2	Translation	5
GHITM	growth hormone inducible transmembrane protein	Apoptosis	10
GK	glycerol kinase	Replication & Transcription	X
GK2	galactokinase 2	Replication & Transcription	15
GLDC	glycine decarboxylase	Amino Acid Metabolism	9
GLRX2	glutaredoxin 2	ROS Defense	1
GLRX5	glutaredoxin 5	Fe-S Cluster Biosynthesis	14
GLS	glutaminase	Amino Acid Metabolism	2
GLS2	glutaminase 2	Amino Acid Metabolism	12
GLUD1	glutamate dehydrogenase 1	Amino Acid Metabolism	10
GLUD2	glutamate dehydrogenase 2	Amino Acid Metabolism	X
GLUL	glutamate-ammonia ligase	Amino Acid Metabolism	1
GLYAT	glycine-N-acyltransferase	Translation	11
GLYATL2	glycine-N-acyltransferase like 2	Fatty Acid Metabolism	11
GLYATL3	glycine-N-acyltransferase like 3	Translation	6
GLYCTK	glycerate kinase	Fructose Metabolism	3
GOT2	glutamic-oxaloacetic transaminase 2	Amino Acid Metabolism	16
GPAM	glycerol-3-phosphate acyltransferase, mitochondrial	Metabolism of Lipids & Lipoproteins	10
GPAT2	glycerol-3-phosphate acyltransferase 2, mitochondrial	Metabolism of Lipids & Lipoproteins	2
GPD1	glycerol-3-phosphate dehydrogenase 1	Oxidative Phosphorylation	12
GPD2	glycerol-3-phosphate dehydrogenase 2	Oxidative Phosphorylation	2
GPI	glucose-6-phosphate isomerase	Glycolysis	19
GPT2	glutamic--pyruvic transaminase 2	Amino Acid Metabolism	16
GPX1	glutathione peroxidase 1	ROS Defense	3
GPX4	glutathione peroxidase 4	ROS Defense	19
GRPEL1	GrpE like 1, mitochondrial	Import & Sorting	4
GRPEL2	GrpE like 2, mitochondrial	Import & Sorting	5
GSK3B	glycogen synthase kinase 3 beta	Mitochondrial Signaling	3
GSR	glutathione-disulfide reductase	ROS Defense	8
GSS	glutathione synthetase	ROS Defense	20

GSTA1	glutathione S-transferase alpha 1	ROS Defense	6
GSTA2	glutathione S-transferase alpha 2	ROS Defense	6
GSTA4	glutathione S-transferase alpha 4	ROS Defense	6
GSTP1	glutathione S-transferase pi 1	ROS Defense	11
GSTZ1	glutathione S-transferase zeta 1	Amino Acid Metabolism	14
GTPBP10	GTP binding protein 10	Translation	7
GTPBP3	GTP binding protein 3, mitochondrial	Translation	19
GUF1	GUF1 homolog, GTPase	Translation	4
HADH	hydroxyacyl-CoA dehydrogenase	Fatty Acid Degradation & Beta-oxidation	4
HADHA	hydroxyacyl-CoA dehydrogenase trifunctional multienzyme complex subunit alpha	Fatty Acid Degradation & Beta-oxidation	2
HADHB	hydroxyacyl-CoA dehydrogenase trifunctional multienzyme complex subunit beta	Fatty Acid Degradation & Beta-oxidation	2
HAGH	hydroxyacylglutathione hydrolase	Pyruvate Metabolism	16
HARS2	histidyl-tRNA synthetase 2, mitochondrial	Translation	5
HCCS	holocytochrome c synthase	Oxidative Phosphorylation	X
HDHD5	haloacid dehalogenase like hydrolase domain containing 5	Fatty Acid Metabolism	22
HEMK1	HemK methyltransferase family member 1	Translation	3
HIBADH	3-hydroxyisobutyrate dehydrogenase	Amino Acid Metabolism	7
HIBCH	3-hydroxyisobutyryl-CoA hydrolase	Amino Acid Metabolism	2
HIF1A	hypoxia inducible factor 1 subunit alpha	Transcription (nuclear)	14
HIGD1A	HIG1 hypoxia inducible domain family member 1A	Oxidative Phosphorylation	3
HIGD2A	HIG1 hypoxia inducible domain family member 2A	Oxidative Phosphorylation	5
HINT2	histidine triad nucleotide binding protein 2	Calcium Signaling & Transport	9
HK1	hexokinase 1	Glycolysis	10
HK2	hexokinase 2	Glycolysis	2
HK3	hexokinase 3	Glycolysis	5
HKDC1	hexokinase domain containing 1	Glycolysis	10
HMBS	hydroxymethylbilane synthase	Heme Biosynthesis	11
HMGCL	3-hydroxy-3-methylglutaryl-CoA lyase	Amino Acid Metabolism	1
HMGCS2	3-hydroxy-3-methylglutaryl-CoA synthase 2	Amino Acid Metabolism	1
HOGA1	4-hydroxy-2-oxoglutarate aldolase 1	Pyruvate Metabolism	10
HSCB	HscB mitochondrial iron-sulfur cluster cochaperone	Fe-S Cluster Biosynthesis	22
HSD17B10	hydroxysteroid 17-beta dehydrogenase 10	Translation	X
HSD17B8	hydroxysteroid 17-beta dehydrogenase 8	Fatty Acid Biosynthesis & Elongation	6
HSD3B1	hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 1	Metabolism of Lipids & Lipoproteins	1
HSD3B2	hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 2	Metabolism of Lipids & Lipoproteins	1
HSDL1	hydroxysteroid dehydrogenase like 1	Unknown	16
HSPA1A	heat shock protein family A (Hsp70) member 1A	Protein Stability & Degradation	6
HSPA1B	heat shock protein family A (Hsp70) member 1B	Protein Stability & Degradation	6
HSPA9	heat shock protein family A (Hsp70) member 9	Import & Sorting	5
HSPD1	heat shock protein family D (Hsp60) member 1	Protein Stability & Degradation	2
HSPE1	heat shock protein family E (Hsp10) member 1	Import & Sorting	2
HTRA2	HtrA serine peptidase 2	Apoptosis	2
IARS2	isoleucyl-tRNA synthetase 2, mitochondrial	Translation	1
IBA57	iron-sulfur cluster assembly factor IBA57	Fe-S Cluster Biosynthesis	1
IDE	insulin degrading enzyme	Mitochondrial Signaling	10
IDH1	isocitrate dehydrogenase (NADP(+)) 1, cytosolic	Tricarboxylic Acid Cycle	2
IDH2	isocitrate dehydrogenase (NADP(+)) 2, mitochondrial	Tricarboxylic Acid Cycle	15
IDH3A	isocitrate dehydrogenase 3 (NAD(+)) alpha	Tricarboxylic Acid Cycle	15
IDH3B	isocitrate dehydrogenase 3 (NAD(+)) beta	Tricarboxylic Acid Cycle	20

IDH3G	isocitrate dehydrogenase 3 (NAD(+)) gamma	Tricarboxylic Acid Cycle	X
IMMP1L	inner mitochondrial membrane peptidase subunit 1	Import & Sorting	11
IMMP2L	inner mitochondrial membrane peptidase subunit 2	Import & Sorting	7
IMMT	inner membrane mitochondrial protein	Mitochondrial Dynamics	2
ISCA1	iron-sulfur cluster assembly 1	Fe-S Cluster Biosynthesis	9
ISCA2	iron-sulfur cluster assembly 2	Fe-S Cluster Biosynthesis	14
ISCU	iron-sulfur cluster assembly enzyme	Fe-S Cluster Biosynthesis	12
ITPR1	inositol 1,4,5-trisphosphate receptor type 1	Calcium Signaling & Transport	3
ITPR2	inositol 1,4,5-trisphosphate receptor type 2	Calcium Signaling & Transport	12
ITPR3	inositol 1,4,5-trisphosphate receptor type 3	Calcium Signaling & Transport	6
IVD	isovaleryl-CoA dehydrogenase	Amino Acid Metabolism	15
JUN	Jun proto-oncogene, AP-1 transcription factor subunit	Transcription (nuclear)	1
KARS	lysyl-tRNA synthetase	Translation	16
KDM6B	lysine demethylase 6B	UPRmt	17
KHK	ketohexokinase	Fructose Metabolism	2
KIF1B	kinesin family member 1B	Mitochondrial Dynamics	1
KIF1BP	KIF1 binding protein	Mitochondrial Dynamics	10
KIF5B	kinesin family member 5B	Mitochondrial Dynamics	10
KMO	kynurenine 3-monooxygenase	Amino Acid Metabolism	1
KYAT3	kynurenine aminotransferase 3	Amino Acid Metabolism	1
L2HGDH	L-2-hydroxyglutarate dehydrogenase	Pyruvate Metabolism	14
LACTB	lactamase beta	Metabolism of Lipids & Lipoproteins	15
LACTB2	lactamase beta 2	Translation	8
LAP3	leucine aminopeptidase 3	Amino Acid Metabolism	4
LARS2	leucyl-tRNA synthetase 2, mitochondrial	Translation	3
LCLAT1	lysocardiolipin acyltransferase 1	Cardiolipin Biosynthesis	2
LDHA	lactate dehydrogenase A	Glycolysis	11
LDHAL6B	lactate dehydrogenase A like 6B	Glycolysis	15
LDHB	lactate dehydrogenase B	Glycolysis	12
LDHC	lactate dehydrogenase C	Glycolysis	11
LDHD	lactate dehydrogenase D	Pyruvate Metabolism	16
LETM1	leucine zipper and EF-hand containing transmembrane protein 1	Calcium Signaling & Transport	4
LETM2	leucine zipper and EF-hand containing transmembrane protein 2	Calcium Signaling & Transport	8
LETMD1	LETM1 domain containing 1	Apoptosis	12
LIAS	lipoic acid synthetase	Lipoic Acid Metabolism	4
LIG1	DNA ligase 1	Replication & Transcription	19
LIG3	DNA ligase 3	Replication & Transcription	17
LIPT1	lipoyltransferase 1	Lipoic Acid Metabolism	2
LIPT2	lipoyl(octanoyl) transferase 2	Lipoic Acid Metabolism	11
LONP1	lon peptidase 1, mitochondrial	Protein Stability & Degradation	19
LRPPRC	leucine rich pentatricopeptide repeat containing	Translation	2
LYPLA1	lysophospholipase 1	Fatty Acid Metabolism	8
LYPLAL1	lysophospholipase like 1	Fatty Acid Metabolism	1
LYRM1	LYR motif containing 1	Mitochondrial Dynamics	16
LYRM4	LYR motif containing 4	Fe-S Cluster Biosynthesis	6
LYRM7	LYR motif containing 7	Oxidative Phosphorylation	5
MAIP1	matrix AAA peptidase interacting protein 1	Calcium Signaling & Transport	2
MALSU1	mitochondrial assembly of ribosomal large subunit 1	Translation	7
MAOA	monoamine oxidase A	Amino Acid Metabolism	X
MAOB	monoamine oxidase B	Amino Acid Metabolism	X
MAP1LC3A	microtubule associated protein 1 light chain 3 alpha	Mitochondrial Dynamics	20
MAPK1	mitogen-activated protein kinase 1	Apoptosis	22

MAPK3	mitogen-activated protein kinase 3	Apoptosis	16
MAPK8	mitogen-activated protein kinase 8	Apoptosis	10
MARCH1	membrane associated ring-CH-type finger 1	Nucleotide Metabolism	4
MARCH2	membrane associated ring-CH-type finger 2	Nucleotide Metabolism	19
MARCH5	membrane associated ring-CH-type finger 5	Mitochondrial Dynamics	10
MARS2	methionyl-tRNA synthetase 2, mitochondrial	Translation	2
MAVS	mitochondrial antiviral signaling protein	Mitochondrial Signaling	20
MCAT	malonyl-CoA-acyl carrier protein transacylase	Fatty Acid Biosynthesis & Elongation	22
MCCC1	methylcrotonoyl-CoA carboxylase 1	Amino Acid Metabolism	3
MCCC2	methylcrotonoyl-CoA carboxylase 2	Amino Acid Metabolism	5
MCEE	methylmalonyl-CoA epimerase	Amino Acid Metabolism	2
MCL1	MCL1 apoptosis regulator, BCL2 family member	Apoptosis	1
MCU	mitochondrial calcium uniporter	Calcium Signaling & Transport	10
MCUB	mitochondrial calcium uniporter dominant negative beta subunit	Calcium Signaling & Transport	4
MCUR1	mitochondrial calcium uniporter regulator 1	Calcium Signaling & Transport	6
MDH1	malate dehydrogenase 1	Tricarboxylic Acid Cycle	2
MDH2	malate dehydrogenase 2	Tricarboxylic Acid Cycle	7
ME1	malic enzyme 1	Pyruvate Metabolism	6
ME2	malic enzyme 2	Pyruvate Metabolism	18
ME3	malic enzyme 3	Pyruvate Metabolism	11
MECR	mitochondrial trans-2-enoyl-CoA reductase	Fatty Acid Biosynthesis & Elongation	1
METAP1D	methionyl aminopeptidase type 1D, mitochondrial	Import & Sorting	2
METTL17	methyltransferase like 17	Translation	14
MFF	mitochondrial fission factor	Mitochondrial Dynamics	2
MFN1	mitofusin 1	Mitochondrial Dynamics	3
MFN2	mitofusin 2	Mitochondrial Dynamics	1
MGARP	mitochondria localized glutamic acid rich protein	Mitochondrial Dynamics	4
MGME1	mitochondrial genome maintenance exonuclease 1	Replication & Transcription	20
MGST1	microsomal glutathione S-transferase 1	ROS Defense	12
MICOS10	mitochondrial contact site and cristae organizing system subunit 10	Mitochondrial Dynamics	1
MICOS13	mitochondrial contact site and cristae organizing system subunit 13	Mitochondrial Dynamics	19
MICU1	mitochondrial calcium uptake 1	Calcium Signaling & Transport	10
MICU2	mitochondrial calcium uptake 2	Calcium Signaling & Transport	13
MICU3	mitochondrial calcium uptake family member 3	Calcium Signaling & Transport	8
MIEF1	mitochondrial elongation factor 1	Mitochondrial Dynamics	22
MIEF2	mitochondrial elongation factor 2	Mitochondrial Dynamics	17
MIPEP	mitochondrial intermediate peptidase	Import & Sorting	13
MITF	melanocyte inducing transcription factor	Transcription (nuclear)	3
MLYCD	malonyl-CoA decarboxylase	Fatty Acid Metabolism	16
MMAA	metabolism of cobalamin associated A	Metabolism of Vitamins & Co-Factors	4
MMAB	metabolism of cobalamin associated B	Metabolism of Vitamins & Co-Factors	12
MMADHC	metabolism of cobalamin associated D	Metabolism of Vitamins & Co-Factors	2
MMS19	MMS19 homolog, cytosolic iron-sulfur assembly component	Fe-S Cluster Biosynthesis	10
MMUT	methylmalonyl-CoA mutase	Amino Acid Metabolism	6
MPC1	mitochondrial pyruvate carrier 1	Pyruvate Metabolism	6
MPC1L	mitochondrial pyruvate carrier 1 like	Pyruvate Metabolism	X
MPC2	mitochondrial pyruvate carrier 2	Pyruvate Metabolism	1
MPST	mercaptopyruvate sulfurtransferase	Amino Acid Metabolism	22
MPV17	mitochondrial inner membrane protein MPV17	ROS Defense	2
MPV17L	MPV17 mitochondrial inner membrane protein like	ROS Defense	16
MPV17L2	MPV17 mitochondrial inner membrane protein like 2	Translation	19

MRM1	mitochondrial rRNA methyltransferase 1	Translation	17
MRM2	mitochondrial rRNA methyltransferase 2	Translation	7
MRM3	mitochondrial rRNA methyltransferase 3	Translation	17
MRPL1	mitochondrial ribosomal protein L1	Translation	4
MRPL10	mitochondrial ribosomal protein L10	Translation	17
MRPL11	mitochondrial ribosomal protein L11	Translation	11
MRPL12	mitochondrial ribosomal protein L12	Translation	17
MRPL13	mitochondrial ribosomal protein L13	Translation	8
MRPL14	mitochondrial ribosomal protein L14	Translation	6
MRPL15	mitochondrial ribosomal protein L15	Translation	8
MRPL16	mitochondrial ribosomal protein L16	Translation	11
MRPL17	mitochondrial ribosomal protein L17	Translation	11
MRPL18	mitochondrial ribosomal protein L18	Translation	6
MRPL19	mitochondrial ribosomal protein L19	Translation	2
MRPL2	mitochondrial ribosomal protein L2	Translation	6
MRPL20	mitochondrial ribosomal protein L20	Translation	1
MRPL21	mitochondrial ribosomal protein L21	Translation	11
MRPL22	mitochondrial ribosomal protein L22	Translation	5
MRPL23	mitochondrial ribosomal protein L23	Translation	11
MRPL24	mitochondrial ribosomal protein L24	Translation	1
MRPL27	mitochondrial ribosomal protein L27	Translation	17
MRPL28	mitochondrial ribosomal protein L28	Translation	16
MRPL3	mitochondrial ribosomal protein L3	Translation	3
MRPL30	mitochondrial ribosomal protein L30	Translation	2
MRPL32	mitochondrial ribosomal protein L32	Translation	7
MRPL33	mitochondrial ribosomal protein L33	Translation	2
MRPL34	mitochondrial ribosomal protein L34	Translation	19
MRPL35	mitochondrial ribosomal protein L35	Translation	2
MRPL36	mitochondrial ribosomal protein L36	Translation	5
MRPL37	mitochondrial ribosomal protein L37	Translation	1
MRPL38	mitochondrial ribosomal protein L38	Translation	17
MRPL39	mitochondrial ribosomal protein L39	Translation	21
MRPL4	mitochondrial ribosomal protein L4	Translation	19
MRPL40	mitochondrial ribosomal protein L40	Translation	22
MRPL41	mitochondrial ribosomal protein L41	Translation	9
MRPL42	mitochondrial ribosomal protein L42	Translation	12
MRPL43	mitochondrial ribosomal protein L43	Translation	10
MRPL44	mitochondrial ribosomal protein L44	Translation	2
MRPL45	mitochondrial ribosomal protein L45	Translation	17
MRPL46	mitochondrial ribosomal protein L46	Translation	15
MRPL47	mitochondrial ribosomal protein L47	Translation	3
MRPL48	mitochondrial ribosomal protein L48	Translation	11
MRPL49	mitochondrial ribosomal protein L49	Translation	11
MRPL50	mitochondrial ribosomal protein L50	Translation	9
MRPL51	mitochondrial ribosomal protein L51	Translation	12
MRPL52	mitochondrial ribosomal protein L52	Translation	14
MRPL53	mitochondrial ribosomal protein L53	Translation	2
MRPL54	mitochondrial ribosomal protein L54	Translation	19
MRPL55	mitochondrial ribosomal protein L55	Translation	1
MRPL57	mitochondrial ribosomal protein L57	Translation	13
MRPL58	mitochondrial ribosomal protein L58	Translation	17
MRPL9	mitochondrial ribosomal protein L9	Translation	1
MRPS10	mitochondrial ribosomal protein S10	Translation	6
MRPS11	mitochondrial ribosomal protein S11	Translation	15
MRPS12	mitochondrial ribosomal protein S12	Translation	19
MRPS14	mitochondrial ribosomal protein S14	Translation	1

MRPS15	mitochondrial ribosomal protein S15	Translation	1
MRPS16	mitochondrial ribosomal protein S16	Translation	10
MRPS17	mitochondrial ribosomal protein S17	Translation	7
MRPS18A	mitochondrial ribosomal protein S18A	Translation	6
MRPS18B	mitochondrial ribosomal protein S18B	Translation	6
MRPS18C	mitochondrial ribosomal protein S18C	Translation	4
MRPS2	mitochondrial ribosomal protein S2	Translation	9
MRPS21	mitochondrial ribosomal protein S21	Translation	1
MRPS22	mitochondrial ribosomal protein S22	Translation	3
MRPS23	mitochondrial ribosomal protein S23	Translation	17
MRPS24	mitochondrial ribosomal protein S24	Translation	7
MRPS25	mitochondrial ribosomal protein S25	Translation	3
MRPS26	mitochondrial ribosomal protein S26	Translation	20
MRPS27	mitochondrial ribosomal protein S27	Translation	5
MRPS28	mitochondrial ribosomal protein S28	Translation	8
MRPS30	mitochondrial ribosomal protein S30	Translation	5
MRPS31	mitochondrial ribosomal protein S31	Translation	13
MRPS33	mitochondrial ribosomal protein S33	Translation	7
MRPS34	mitochondrial ribosomal protein S34	Translation	16
MRPS35	mitochondrial ribosomal protein S35	Translation	12
MRPS36	mitochondrial ribosomal protein S36	Translation	5
MRPS5	mitochondrial ribosomal protein S5	Translation	2
MRPS6	mitochondrial ribosomal protein S6	Translation	21
MRPS7	mitochondrial ribosomal protein S7	Translation	17
MRPS9	mitochondrial ribosomal protein S9	Translation	2
MRRF	mitochondrial ribosome recycling factor	Translation	9
MRS2	magnesium transporter MRS2	Transmembrane Transport	6
MSH5	mutS homolog 5	Replication & Transcription	6
MSRA	methionine sulfoxide reductase A	ROS Defense	8
MSRB2	methionine sulfoxide reductase B2	ROS Defense	10
MSRB3	methionine sulfoxide reductase B3	ROS Defense	12
MSTO1	misato mitochondrial distribution and morphology regulator 1	Mitochondrial Dynamics	1
MT-ATP6	mitochondrially encoded ATP synthase 6	Oxidative Phosphorylation (mt)	MT
MT-ATP8	mitochondrially encoded ATP synthase 8	Oxidative Phosphorylation (mt)	MT
MT-CO1	mitochondrially encoded cytochrome c oxidase I	Oxidative Phosphorylation (mt)	MT
MT-CO2	mitochondrially encoded cytochrome c oxidase II	Oxidative Phosphorylation (mt)	MT
MT-CO3	mitochondrially encoded cytochrome c oxidase III	Oxidative Phosphorylation (mt)	MT
MT-CYB	mitochondrially encoded cytochrome b	Oxidative Phosphorylation (mt)	MT
MT-ND1	mitochondrially encoded NADH dehydrogenase 1	Oxidative Phosphorylation (mt)	MT
MT-ND2	mitochondrially encoded NADH dehydrogenase 2	Oxidative Phosphorylation (mt)	MT
MT-ND3	mitochondrially encoded NADH dehydrogenase 3	Oxidative Phosphorylation (mt)	MT
MT-ND4	mitochondrially encoded NADH dehydrogenase 4	Oxidative Phosphorylation (mt)	MT
MT-ND4L	mitochondrially encoded NADH 4L dehydrogenase	Oxidative Phosphorylation (mt)	MT
MT-ND5	mitochondrially encoded NADH dehydrogenase 5	Oxidative Phosphorylation (mt)	MT
MT-ND6	mitochondrially encoded NADH dehydrogenase 6	Oxidative Phosphorylation (mt)	MT
MT-RNR1	mitochondrially encoded 12S RNA	Translation (MT)	MT
MT-RNR2	mitochondrially encoded 16S RNA	Translation (MT)	MT
MT-TA	mitochondrially encoded tRNA alanine	Translation (MT)	MT
MT-TC	mitochondrially encoded tRNA cysteine	Translation (MT)	MT
MT-TD	mitochondrially encoded tRNA aspartic acid	Translation (MT)	MT
MT-TE	mitochondrially encoded tRNA glutamic acid	Translation (MT)	MT

MT-TF	mitochondrially encoded tRNA phenylalanine	Translation (MT)	MT
MT-TG	mitochondrially encoded tRNA glycine	Translation (MT)	MT
MT-TH	mitochondrially encoded tRNA histidine	Translation (MT)	MT
MT-TI	mitochondrially encoded tRNA isoleucine	Translation (MT)	MT
MT-TK	mitochondrially encoded tRNA lysine	Translation (MT)	MT
MT-TL1	mitochondrially encoded tRNA leucine 1 (UUA/G)	Translation (MT)	MT
MT-TL2	mitochondrially encoded tRNA leucine 2 (CUN)	Translation (MT)	MT
MT-TM	mitochondrially encoded tRNA methionine	Translation (MT)	MT
MT-TN	mitochondrially encoded tRNA asparagine	Translation (MT)	MT
MT-TP	mitochondrially encoded tRNA proline	Translation (MT)	MT
MT-TQ	mitochondrially encoded tRNA glutamine	Translation (MT)	MT
MT-TR	mitochondrially encoded tRNA arginine	Translation (MT)	MT
MT-TS1	mitochondrially encoded tRNA serine 1 (UCN)	Translation (MT)	MT
MT-TS2	mitochondrially encoded tRNA serine 2 (AGU/C)	Translation (MT)	MT
MT-TT	mitochondrially encoded tRNA threonine	Translation (MT)	MT
MT-TV	mitochondrially encoded tRNA valine	Translation (MT)	MT
MT-TW	mitochondrially encoded tRNA tryptophan	Translation (MT)	MT
MT-TY	mitochondrially encoded tRNA tyrosine	Translation (MT)	MT
MTCH1	mitochondrial carrier 1	Apoptosis	6
MTCH2	mitochondrial carrier 2	Apoptosis	11
MTERF1	mitochondrial transcription termination factor 1	Replication & Transcription	7
MTERF2	mitochondrial transcription termination factor 2	Replication & Transcription	12
MTERF3	mitochondrial transcription termination factor 3	Translation	8
MTERF4	mitochondrial transcription termination factor 4	Translation	2
MTFMT	mitochondrial methionyl-tRNA formyltransferase	Translation	15
MTFP1	mitochondrial fission process 1	Mitochondrial Dynamics	22
MTFR1	mitochondrial fission regulator 1	Mitochondrial Dynamics	8
MTFR1L	mitochondrial fission regulator 1 like	Mitochondrial Dynamics	1
MTFR2	mitochondrial fission regulator 2	Mitochondrial Dynamics	6
MTG1	mitochondrial ribosome associated GTPase 1	Translation	10
MTG2	mitochondrial ribosome associated GTPase 2	Translation	20
MTHFD1	methylenetetrahydrofolate dehydrogenase, cyclohydrolase and formyltetrahydrofolate synthetase 1	Folate & Pterin Metabolism	14
MTHFD1L	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 1 like	Folate & Pterin Metabolism	6
MTHFD2	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2, methenyltetrahydrofolate cyclohydrolase	Folate & Pterin Metabolism	2
MTHFD2L	methylenetetrahydrofolate dehydrogenase (NADP+ dependent) 2 like	Folate & Pterin Metabolism	4
MTHFS	methenyltetrahydrofolate synthetase	Folate & Pterin Metabolism	15
MTIF2	mitochondrial translational initiation factor 2	Translation	2
MTIF3	mitochondrial translational initiation factor 3	Translation	13
MTO1	mitochondrial tRNA translation optimization 1	Translation	6
MTOR	mechanistic target of rapamycin kinase	Mitochondrial Signaling	1
MTPAP	mitochondrial poly(A) polymerase	Translation	10
MTRES1	mitochondrial transcription rescue factor 1	Translation	6
MTRF1	mitochondrial translation release factor 1	Translation	13
MTRF1L	mitochondrial translational release factor 1 like	Translation	6
MTX1	metaxin 1	Import & Sorting	1
MTX2	metaxin 2	Import & Sorting	2
MTX3	metaxin 3	Import & Sorting	5
MUL1	mitochondrial E3 ubiquitin protein ligase 1	Protein Stability & Degradation	1

MUTYH	mutY DNA glycosylase	Replication & Transcription	1
MYC	MYC proto-oncogene, bHLH transcription factor	Transcription (nuclear)	8
NADK2	NAD kinase 2, mitochondrial	Amino Acid Metabolism	5
NAGS	N-acetylglutamate synthase	Nitrogen Metabolism	17
NARF	nuclear prelamin A recognition factor	Fe-S Cluster Biosynthesis	17
NARS2	asparaginyl-tRNA synthetase 2, mitochondrial	Translation	11
NAXD	NAD(P)HX dehydratase	Metabolism of Vitamins & Co-Factors	13
NAXE	NAD(P)HX epimerase	Metabolism of Vitamins & Co-Factors	1
NBR1	NBR1 autophagy cargo receptor	Mitophagy	17
NDUFA1	NADH:ubiquinone oxidoreductase subunit A1	Oxidative Phosphorylation	X
NDUFA10	NADH:ubiquinone oxidoreductase subunit A10	Oxidative Phosphorylation	2
NDUFA11	NADH:ubiquinone oxidoreductase subunit A11	Oxidative Phosphorylation	19
NDUFA12	NADH:ubiquinone oxidoreductase subunit A12	Oxidative Phosphorylation	12
NDUFA13	NADH:ubiquinone oxidoreductase subunit A13	Oxidative Phosphorylation	19
NDUFA2	NADH:ubiquinone oxidoreductase subunit A2	Oxidative Phosphorylation	5
NDUFA3	NADH:ubiquinone oxidoreductase subunit A3	Oxidative Phosphorylation	19
NDUFA4	NDUFA4 mitochondrial complex associated	Oxidative Phosphorylation	7
NDUFA5	NADH:ubiquinone oxidoreductase subunit A5	Oxidative Phosphorylation	7
NDUFA6	NADH:ubiquinone oxidoreductase subunit A6	Oxidative Phosphorylation	22
NDUFA7	NADH:ubiquinone oxidoreductase subunit A7	Oxidative Phosphorylation	19
NDUFA8	NADH:ubiquinone oxidoreductase subunit A8	Oxidative Phosphorylation	9
NDUFA9	NADH:ubiquinone oxidoreductase subunit A9	Oxidative Phosphorylation	12
NDUFAB1	NADH:ubiquinone oxidoreductase subunit AB1	Oxidative Phosphorylation	16
NDUFAF1	NADH:ubiquinone oxidoreductase complex assembly factor 1	Oxidative Phosphorylation	15
NDUFAF2	NADH:ubiquinone oxidoreductase complex assembly factor 2	Oxidative Phosphorylation	5
NDUFAF3	NADH:ubiquinone oxidoreductase complex assembly factor 3	Oxidative Phosphorylation	3
NDUFAF4	NADH:ubiquinone oxidoreductase complex assembly factor 4	Oxidative Phosphorylation	6
NDUFAF5	NADH:ubiquinone oxidoreductase complex assembly factor 5	Oxidative Phosphorylation	20
NDUFAF6	NADH:ubiquinone oxidoreductase complex assembly factor 6	Oxidative Phosphorylation	8
NDUFAF7	NADH:ubiquinone oxidoreductase complex assembly factor 7	Oxidative Phosphorylation	2
NDUFAF8	NADH:ubiquinone oxidoreductase complex assembly factor 8	Oxidative Phosphorylation	17
NDUFB1	NADH:ubiquinone oxidoreductase subunit B1	Oxidative Phosphorylation	14
NDUFB10	NADH:ubiquinone oxidoreductase subunit B10	Oxidative Phosphorylation	16
NDUFB11	NADH:ubiquinone oxidoreductase subunit B11	Oxidative Phosphorylation	X
NDUFB2	NADH:ubiquinone oxidoreductase subunit B2	Oxidative Phosphorylation	7
NDUFB3	NADH:ubiquinone oxidoreductase subunit B3	Oxidative Phosphorylation	2
NDUFB4	NADH:ubiquinone oxidoreductase subunit B4	Oxidative Phosphorylation	3
NDUFB5	NADH:ubiquinone oxidoreductase subunit B5	Oxidative Phosphorylation	3
NDUFB6	NADH:ubiquinone oxidoreductase subunit B6	Oxidative Phosphorylation	9
NDUFB7	NADH:ubiquinone oxidoreductase subunit B7	Oxidative Phosphorylation	19
NDUFB8	NADH:ubiquinone oxidoreductase subunit B8	Oxidative Phosphorylation	10
NDUFB9	NADH:ubiquinone oxidoreductase subunit B9	Oxidative Phosphorylation	8
NDUFC1	NADH:ubiquinone oxidoreductase subunit C1	Oxidative Phosphorylation	4
NDUFC2	NADH:ubiquinone oxidoreductase subunit C2	Oxidative Phosphorylation	11
NDUFS1	NADH:ubiquinone oxidoreductase core subunit S1	Oxidative Phosphorylation	2
NDUFS2	NADH:ubiquinone oxidoreductase core subunit S2	Oxidative Phosphorylation	1
NDUFS3	NADH:ubiquinone oxidoreductase core subunit S3	Oxidative Phosphorylation	11

NDUFS4	NADH:ubiquinone oxidoreductase subunit S4	Oxidative Phosphorylation	5
NDUFS5	NADH:ubiquinone oxidoreductase subunit S5	Oxidative Phosphorylation	1
NDUFS6	NADH:ubiquinone oxidoreductase subunit S6	Oxidative Phosphorylation	5
NDUFS7	NADH:ubiquinone oxidoreductase core subunit S7	Oxidative Phosphorylation	19
NDUFS8	NADH:ubiquinone oxidoreductase core subunit S8	Oxidative Phosphorylation	11
NDUFV1	NADH:ubiquinone oxidoreductase core subunit V1	Oxidative Phosphorylation	11
NDUFV2	NADH:ubiquinone oxidoreductase core subunit V2	Oxidative Phosphorylation	18
NDUFV3	NADH:ubiquinone oxidoreductase subunit V3	Oxidative Phosphorylation	21
NEU4	neuraminidase 4	Metabolism of Lipids & Lipoproteins	2
NFE2L2	nuclear factor, erythroid 2 like 2	Replication & Transcription	2
NFS1	NFS1 cysteine desulfurase	Fe-S Cluster Biosynthesis	20
NFU1	NFU1 iron-sulfur cluster scaffold	Fe-S Cluster Biosynthesis	2
NGB	neuroglobin	ROS Defense	14
NIF3L1	NGG1 interacting factor 3 like 1	Replication & Transcription	2
NIPSNAP1	nipsnap homolog 1	Oxidative Phosphorylation	22
NIPSNAP2	nipsnap homolog 2	Oxidative Phosphorylation	7
NIPSNAP3A	nipsnap homolog 3A	Transmembrane Transport	9
NIPSNAP3B	nipsnap homolog 3B	Transmembrane Transport	9
NIT1	nitrilase 1	Amino Acid Metabolism	1
NIT2	nitrilase family member 2	Amino Acid Metabolism	3
NLN	neurolysin	Protein Stability & Degradation	5
NLRX1	NLR family member X1	Mitochondrial Signaling	11
NME1	NME/NM23 nucleoside diphosphate kinase 1	Nucleotide Metabolism	17
NME4	NME/NM23 nucleoside diphosphate kinase 4	Nucleotide Metabolism	16
NME6	NME/NM23 nucleoside diphosphate kinase 6	Nucleotide Metabolism	3
NMNAT3	nicotinamide nucleotide adenylyltransferase 3	Metabolism of Vitamins & Co-Factors	3
NNT	nicotinamide nucleotide transhydrogenase	Tricarboxylic Acid Cycle	5
NOA1	nitric oxide associated 1	Translation	4
NRDC	nardilysin convertase	Protein Stability & Degradation	1
NRF1	nuclear respiratory factor 1	Transcription (nuclear)	7
NSUN2	NOP2/Sun RNA methyltransferase 2	Replication & Transcription	5
NSUN3	NOP2/Sun RNA methyltransferase 3	Translation	3
NSUN4	NOP2/Sun RNA methyltransferase 4	Translation	1
NT5M	5',3'-nucleotidase, mitochondrial	Replication & Transcription	17
NTHL1	nth like DNA glycosylase 1	Replication & Transcription	16
NUBP1	nucleotide binding protein 1	Fe-S Cluster Biosynthesis	16
NUBP2	nucleotide binding protein 2	Fe-S Cluster Biosynthesis	16
NUBPL	nucleotide binding protein like	Oxidative Phosphorylation	14
NUDT13	nudix hydrolase 13	Nucleotide Metabolism	10
NUDT19	nudix hydrolase 19	Metabolism of Lipids & Lipoproteins	19
NUDT2	nudix hydrolase 2	ROS Defense	9
NUDT9	nudix hydrolase 9	Nucleotide Metabolism	4
OAT	ornithine aminotransferase	Amino Acid Metabolism	10
OGDH	oxoglutarate dehydrogenase	Tricarboxylic Acid Cycle	7
OGDHL	oxoglutarate dehydrogenase like	Tricarboxylic Acid Cycle	10
OGG1	8-oxoguanine DNA glycosylase	ROS Defense	3
OMA1	OMA1 zinc metallopeptidase	Mitochondrial Dynamics	1
OPA1	OPA1 mitochondrial dynamin like GTPase	Mitochondrial Dynamics	3
OPA3	OPA3 outer mitochondrial membrane lipid metabolism regulator	Metabolism of Lipids & Lipoproteins	19
OPTN	optineurin	Mitophagy	10
OSGEPL1	O-sialoglycoprotein endopeptidase like 1	Translation	2
OTC	ornithine carbamoyltransferase	Nitrogen Metabolism	X
OXA1L	OXA1L mitochondrial inner membrane protein	Oxidative Phosphorylation	14

OXCT1	3-oxoacid CoA-transferase 1	Amino Acid Metabolism	5
OXCT2	3-oxoacid CoA-transferase 2	Amino Acid Metabolism	1
OXR1	oxidation resistance 1	ROS Defense	8
OXSM	3-oxoacyl-ACP synthase, mitochondrial	Fatty Acid Biosynthesis & Elongation	3
PACRG	parkin coregulated	Apoptosis	6
PACS2	phosphofurin acidic cluster sorting protein 2	Calcium Signaling & Transport	14
PAM16	presequence translocase associated motor 16	Import & Sorting	16
PANK2	pantothenate kinase 2	Metabolism of Vitamins & Co-Factors	20
PARK7	Parkinsonism associated deglycase	ROS Defense	1
PARL	presenilin associated rhomboid like	Apoptosis	3
PARS2	prolyl-tRNA synthetase 2, mitochondrial	Translation	1
PC	pyruvate carboxylase	Glycolysis	11
PCCA	propionyl-CoA carboxylase subunit alpha	Amino Acid Metabolism	13
PCCB	propionyl-CoA carboxylase subunit beta	Fatty Acid Metabolism	3
PCK1	phosphoenolpyruvate carboxykinase 1	Glycolysis	20
PCK2	phosphoenolpyruvate carboxykinase 2, mitochondrial	Pyruvate Metabolism	14
PDE12	phosphodiesterase 12	Translation	3
PDF	peptide deformylase, mitochondrial	Amino Acid Metabolism	16
PDHA1	pyruvate dehydrogenase E1 alpha 1 subunit	Pyruvate Metabolism	X
PDHA2	pyruvate dehydrogenase E1 alpha 2 subunit	Pyruvate Metabolism	4
PDHB	pyruvate dehydrogenase E1 beta subunit	Pyruvate Metabolism	3
PDHX	pyruvate dehydrogenase complex component X	Tricarboxylic Acid Cycle	11
PDK1	pyruvate dehydrogenase kinase 1	Pyruvate Metabolism	2
PDK2	pyruvate dehydrogenase kinase 2	Pyruvate Metabolism	17
PDK3	pyruvate dehydrogenase kinase 3	Pyruvate Metabolism	X
PDK4	pyruvate dehydrogenase kinase 4	Pyruvate Metabolism	7
PDP1	pyruvate dehyrogenase phosphatase catalytic subunit 1	Pyruvate Metabolism	8
PDP2	pyruvate dehyrogenase phosphatase catalytic subunit 2	Pyruvate Metabolism	16
PDPR	pyruvate dehydrogenase phosphatase regulatory subunit	Pyruvate Metabolism	16
PDSS1	decaprenyl diphosphate synthase subunit 1	Ubiquinone Biosynthesis	10
PDSS2	decaprenyl diphosphate synthase subunit 2	Ubiquinone Biosynthesis	6
PET100	PET100 cytochrome c oxidase chaperone	Protein Stability & Degradation	19
PET117	PET117 cytochrome c oxidase chaperone	Oxidative Phosphorylation	20
PFKL	phosphofructokinase, liver type	Glycolysis	21
PFKM	phosphofructokinase, muscle	Glycolysis	12
PFKP	phosphofructokinase, platelet	Glycolysis	10
PGAM1	phosphoglycerate mutase 1	Glycolysis	10
PGAM2	phosphoglycerate mutase 2	Glycolysis	7
PGAM4	phosphoglycerate mutase family member 4	Glycolysis	X
PGAM5	PGAM family member 5, mitochondrial serine/threonine protein phosphatase	Mitophagy	12
PGD	phosphogluconate dehydrogenase	Pentose Phosphate Pathway	1
PGK1	phosphoglycerate kinase 1	Glycolysis	X
PGK2	phosphoglycerate kinase 2	Glycolysis	6
PGLS	6-phosphogluconolactonase	Pentose Phosphate Pathway	19
PGS1	phosphatidylglycerophosphate synthase 1	Cardiolipin Biosynthesis	17
PHB	prohibitin	Replication & Transcription	17
PHB2	prohibitin 2	Mitochondrial Dynamics	12
PHF8	PHD finger protein 8	UPRmt	X
PHYKPL	5-phosphohydroxy-L-lysine phospho-lyase	Amino Acid Metabolism	5
PIF1	PIF1 5'-to-3' DNA helicase	Replication & Transcription	15
PIN4	peptidylprolyl cis/trans isomerase, NIMA-interacting 4	Translation	X
PINK1	PTEN induced kinase 1	Mitophagy	1

PISD	phosphatidylserine decarboxylase	Metabolism of Lipids & Lipoproteins	22
PITRM1	pitrilysin metallopeptidase 1	Protein Stability & Degradation	10
PKIA	cAMP-dependent protein kinase inhibitor alpha	Mitochondrial Dynamics	8
PKLR	pyruvate kinase L/R	Glycolysis	1
PKM	pyruvate kinase M1/2	Glycolysis	15
PLD6	phospholipase D family member 6	Mitochondrial Dynamics	17
PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1	Apoptosis	18
PML	promyelocytic leukemia	Calcium Signaling & Transport	15
PMPCA	peptidase, mitochondrial processing alpha subunit	Import & Sorting	9
PMPCB	peptidase, mitochondrial processing beta subunit	Import & Sorting	7
PNKD	PNKD metallo-beta-lactamase domain containing	Unknown	2
PNPLA4	patatin like phospholipase domain containing 4	Fatty Acid Metabolism	X
PNPLA8	patatin like phospholipase domain containing 8	Fatty Acid Metabolism	7
PNPT1	polyribonucleotide nucleotidyltransferase 1	Translation	2
POLD4	DNA polymerase delta 4, accessory subunit	Replication & Transcription	11
POLDIP2	DNA polymerase delta interacting protein 2	Replication & Transcription	17
POLG	DNA polymerase gamma, catalytic subunit	Replication & Transcription	15
POLG2	DNA polymerase gamma 2, accessory subunit	Replication & Transcription	17
POLRMT	RNA polymerase mitochondrial	Replication & Transcription	19
PPA2	pyrophosphatase (inorganic) 2	Translation	4
PPARA	peroxisome proliferator activated receptor alpha	Transcription (nuclear)	22
PPARG	peroxisome proliferator activated receptor gamma	Transcription (nuclear)	3
PPARGC1A	PPARG coactivator 1 alpha	Transcription (nuclear)	4
PPARGC1B	PPARG coactivator 1 beta	Transcription (nuclear)	5
PPIF	peptidylprolyl isomerase F	Import & Sorting	10
PPM1K	protein phosphatase, Mg2+/Mn2+ dependent 1K	Amino Acid Metabolism	4
PPOX	protoporphyrinogen oxidase	Heme Biosynthesis	1
PPRC1	PPARG related coactivator 1	Transcription (nuclear)	10
PPTC7	PTC7 protein phosphatase homolog	Tricarboxylic Acid Cycle	12
PRDX1	peroxiredoxin 1	ROS Defense	1
PRDX3	peroxiredoxin 3	ROS Defense	10
PRDX5	peroxiredoxin 5	ROS Defense	11
PRELID1	PRELI domain containing 1	Apoptosis	5
PRELID2	PRELI domain containing 2	Transmembrane Transport	5
PRELID3A	PRELI domain containing 3A	Metabolism of Lipids & Lipoproteins	18
PRELID3B	PRELI domain containing 3B	Metabolism of Lipids & Lipoproteins	20
PRKACA	protein kinase cAMP-activated catalytic subunit alpha	Mitochondrial Signaling	19
PRKCE	protein kinase C epsilon	Mitochondrial Signaling	2
PRKN	parkin RBR E3 ubiquitin protein ligase	Mitophagy	6
PRODH	proline dehydrogenase 1	Amino Acid Metabolism	22
PRODH2	proline dehydrogenase 2	Amino Acid Metabolism	19
PRORP	protein only RNase P catalytic subunit	Translation	14
PSTK	phosphoseryl-tRNA kinase	Translation	10
PTCD1	pentatricopeptide repeat domain 1	Translation	7
PTCD2	pentatricopeptide repeat domain 2	Oxidative Phosphorylation	5
PTCD3	pentatricopeptide repeat domain 3	Translation	2
PTPMT1	protein tyrosine phosphatase mitochondrial 1	Cardiolipin Biosynthesis	11
PTRH2	peptidyl-tRNA hydrolase 2	Translation	17
PTS	6-pyruvoyltetrahydropterin synthase	Folate & Pterin Metabolism	11
PUS1	pseudouridine synthase 1	Translation	12
PUSL1	pseudouridine synthase like 1	Translation	1

PYCR1	pyrroline-5-carboxylate reductase 1	Amino Acid Metabolism	17
PYCR2	pyrroline-5-carboxylate reductase 2	Amino Acid Metabolism	1
QARS	glutaminyl-tRNA synthetase	Translation	3
QDPR	quinoid dihydropteridine reductase	Amino Acid Metabolism	4
QRSL1	glutaminyl-tRNA amidotransferase subunit QRSL1	Translation	6
QTRT1	queuine tRNA-ribosyltransferase catalytic subunit 1	Translation	19
RAB7B	RAB7B, member RAS oncogene family	Mitophagy	1
RAD51	RAD51 recombinase	Replication & Transcription	15
RARS2	arginyl-tRNA synthetase 2, mitochondrial	Translation	6
RBFA	ribosome binding factor A	Translation	18
RDH13	retinol dehydrogenase 13	Metabolism of Vitamins & Co-Factors	19
RELA	RELA proto-oncogene, NF-kB subunit	Transcription (nuclear)	11
REXO2	RNA exonuclease 2	Translation	11
RHOT1	ras homolog family member T1	Mitochondrial Dynamics	17
RHOT2	ras homolog family member T2	Mitochondrial Dynamics	16
RIDA	reactive intermediate imine deaminase A homolog	Translation	8
RMND1	required for meiotic nuclear division 1 homolog	Translation	6
RMRP	RNA component of mitochondrial RNA processing endoribonuclease	Replication & Transcription	9
RNASEH1	ribonuclease H1	Replication & Transcription	2
RNASEL	ribonuclease L	Translation	1
RNF185	ring finger protein 185	Protein Stability & Degradation	22
RNF5	ring finger protein 5	Protein Stability & Degradation	6
ROMO1	reactive oxygen species modulator 1	ROS Defense	20
RPE	ribulose-5-phosphate-3-epimerase	Pentose Phosphate Pathway	2
RPIA	ribose 5-phosphate isomerase A	Pentose Phosphate Pathway	2
RPS19	ribosomal protein S19	Translation	19
RPS6KB1	ribosomal protein S6 kinase B1	Apoptosis	17
RPUSD3	RNA pseudouridine synthase D3	Translation	3
RPUSD4	RNA pseudouridine synthase D4	Translation	11
RRAS2	RAS related 2	Mitophagy	11
RRM2B	ribonucleotide reductase regulatory TP53 inducible subunit M2B	Replication & Transcription	8
RSAD1	radical S-adenosyl methionine domain containing 1	Unknown	17
RTN4IP1	reticulon 4 interacting protein 1	Oxidative Phosphorylation	6
SAMM50	SAMM50 sorting and assembly machinery component	Import & Sorting	22
SARDH	sarcosine dehydrogenase	Metabolism of Lipids & Lipoproteins	9
SARS2	seryl-tRNA synthetase 2, mitochondrial	Translation	19
SATB1	SATB homeobox 1	UPRmt	3
SATB2	SATB homeobox 2	UPRmt	2
SBDS	SBDS ribosome maturation factor	Translation	7
SCO1	SCO cytochrome c oxidase assembly protein 1	Oxidative Phosphorylation	17
SCO2	SCO cytochrome c oxidase assembly protein 2	Oxidative Phosphorylation	22
SDHA	succinate dehydrogenase complex flavoprotein subunit A	Oxidative Phosphorylation	5
SDHAF1	succinate dehydrogenase complex assembly factor 1	Oxidative Phosphorylation	19
SDHAF2	succinate dehydrogenase complex assembly factor 2	Oxidative Phosphorylation	11
SDHAF3	succinate dehydrogenase complex assembly factor 3	Oxidative Phosphorylation	7
SDHAF4	succinate dehydrogenase complex assembly factor 4	Oxidative Phosphorylation	6
SDHB	succinate dehydrogenase complex iron sulfur subunit B	Oxidative Phosphorylation	1
SDHC	succinate dehydrogenase complex subunit C	Oxidative Phosphorylation	1
SDHD	succinate dehydrogenase complex subunit D	Oxidative Phosphorylation	11

SERAC1	serine active site containing 1	Metabolism of Lipids & Lipoproteins	6
SFXN1	sideroflexin 1	Transmembrane Transport	5
SFXN2	sideroflexin 2	Transmembrane Transport	10
SFXN3	sideroflexin 3	Transmembrane Transport	10
SFXN4	sideroflexin 4	Transmembrane Transport	10
SFXN5	sideroflexin 5	Transmembrane Transport	2
SH3GLB1	SH3 domain containing GRB2 like, endophilin B1	Mitochondrial Dynamics	1
SHC1	SHC adaptor protein 1	Mitochondrial Signaling	1
SHMT2	serine hydroxymethyltransferase 2	Folate & Pterin Metabolism	12
SIRT1	sirtuin 1	Mitochondrial Signaling	10
SIRT2	sirtuin 2	Mitochondrial Signaling	19
SIRT3	sirtuin 3	Mitochondrial Signaling	11
SIRT4	sirtuin 4	Mitochondrial Signaling	12
SIRT5	sirtuin 5	Mitochondrial Signaling	6
SLC16A7	solute carrier family 16 member 7	Mitochondrial Carrier	12
SLC22A4	solute carrier family 22 member 4	Mitochondrial Carrier	5
SLC25A1	solute carrier family 25 member 1	Mitochondrial Carrier	22
SLC25A10	solute carrier family 25 member 10	Mitochondrial Carrier	17
SLC25A11	solute carrier family 25 member 11	Mitochondrial Carrier	17
SLC25A12	solute carrier family 25 member 12	Calcium Signaling & Transport	2
SLC25A13	solute carrier family 25 member 13	Calcium Signaling & Transport	7
SLC25A14	solute carrier family 25 member 14	Mitochondrial Carrier	X
SLC25A15	solute carrier family 25 member 15	Mitochondrial Carrier	13
SLC25A16	solute carrier family 25 member 16	Mitochondrial Carrier	10
SLC25A18	solute carrier family 25 member 18	Mitochondrial Carrier	22
SLC25A19	solute carrier family 25 member 19	Mitochondrial Carrier	17
SLC25A2	solute carrier family 25 member 2	Mitochondrial Carrier	5
SLC25A20	solute carrier family 25 member 20	Mitochondrial Carrier	3
SLC25A21	solute carrier family 25 member 21	Mitochondrial Carrier	14
SLC25A22	solute carrier family 25 member 22	Mitochondrial Carrier	11
SLC25A23	solute carrier family 25 member 23	Mitochondrial Carrier	19
SLC25A24	solute carrier family 25 member 24	Mitochondrial Carrier	1
SLC25A25	solute carrier family 25 member 25	Mitochondrial Carrier	9
SLC25A26	solute carrier family 25 member 26	Mitochondrial Carrier	3
SLC25A27	solute carrier family 25 member 27	Mitochondrial Carrier	6
SLC25A28	solute carrier family 25 member 28	Fe-S Cluster Biosynthesis	10
SLC25A29	solute carrier family 25 member 29	Mitochondrial Carrier	14
SLC25A3	solute carrier family 25 member 3	Mitochondrial Carrier	12
SLC25A30	solute carrier family 25 member 30	Mitochondrial Carrier	13
SLC25A31	solute carrier family 25 member 31	Mitochondrial Dynamics	4
SLC25A32	solute carrier family 25 member 32	Folate & Pterin Metabolism	8
SLC25A33	solute carrier family 25 member 33	Mitochondrial Carrier	1
SLC25A34	solute carrier family 25 member 34	Mitochondrial Carrier	1
SLC25A35	solute carrier family 25 member 35	Mitochondrial Carrier	17
SLC25A36	solute carrier family 25 member 36	Mitochondrial Carrier	3
SLC25A37	solute carrier family 25 member 37	Fe-S Cluster Biosynthesis	8
SLC25A38	solute carrier family 25 member 38	Mitochondrial Carrier	3
SLC25A39	solute carrier family 25 member 39	Mitochondrial Carrier	17
SLC25A4	solute carrier family 25 member 4	Mitochondrial Dynamics	4
SLC25A40	solute carrier family 25 member 40	Mitochondrial Carrier	7
SLC25A41	solute carrier family 25 member 41	Mitochondrial Carrier	19
SLC25A42	solute carrier family 25 member 42	Mitochondrial Carrier	19
SLC25A43	solute carrier family 25 member 43	Mitochondrial Carrier	X
SLC25A44	solute carrier family 25 member 44	Mitochondrial Carrier	1
SLC25A45	solute carrier family 25 member 45	Mitochondrial Carrier	11

SLC25A46	solute carrier family 25 member 46	Mitochondrial Dynamics	5
SLC25A47	solute carrier family 25 member 47	Mitochondrial Carrier	14
SLC25A48	solute carrier family 25 member 48	Mitochondrial Carrier	5
SLC25A5	solute carrier family 25 member 5	Mitochondrial Dynamics	X
SLC25A51	solute carrier family 25 member 51	Mitochondrial Carrier	9
SLC25A52	solute carrier family 25 member 52	Mitochondrial Carrier	18
SLC25A53	solute carrier family 25 member 53	Mitochondrial Carrier	X
SLC25A6	solute carrier family 25 member 6	Mitochondrial Dynamics	X
SLC27A1	solute carrier family 27 member 1	Mitochondrial Carrier	19
SLC27A3	solute carrier family 27 member 3	Metabolism of Lipids & Lipoproteins	1
SLC29A1	solute carrier family 29 member 1 (Augustine blood group)	Mitochondrial Carrier	6
SLC2A1	solute carrier family 2 member 1	Glycolysis	1
SLC2A2	solute carrier family 2 member 2	Glycolysis	3
SLC2A3	solute carrier family 2 member 3	Glycolysis	12
SLC2A4	solute carrier family 2 member 4	Glycolysis	17
SLC2A5	solute carrier family 2 member 5	Glycolysis	1
SLC3A1	solute carrier family 3 member 1	Mitochondrial Carrier	2
SLC44A1	solute carrier family 44 member 1	Mitochondrial Carrier	9
SLC8B1	solute carrier family 8 member B1	Calcium Signaling & Transport	12
SLIRP	SRA stem-loop interacting RNA binding protein	Translation	14
SMDT1	single-pass membrane protein with aspartate rich tail 1	Calcium Signaling & Transport	22
SOD1	superoxide dismutase 1	ROS Defense	21
SOD2	superoxide dismutase 2	ROS Defense	6
SORD	sorbitol dehydrogenase	Fructose Metabolism	15
SP1	Sp1 transcription factor	Transcription (nuclear)	12
SPATA19	spermatogenesis associated 19	Unknown	11
SPG7	SPG7 matrix AAA peptidase subunit, paraplegin	Mitochondrial Dynamics	16
SQOR	sulfide quinone oxidoreductase	Amino Acid Metabolism	15
SQSTM1	sequestosome 1	Mitochondrial Dynamics	5
SSBP1	single stranded DNA binding protein 1	Replication & Transcription	7
STAR	steroidogenic acute regulatory protein	Bile Acid Synthesis	8
STARD7	StAR related lipid transfer domain containing 7	Metabolism of Lipids & Lipoproteins	2
STOML2	stomatin like 2	Mitochondrial Dynamics	9
SUCLA2	succinate-CoA ligase ADP-forming beta subunit	Tricarboxylic Acid Cycle	13
SUCLG1	succinate-CoA ligase alpha subunit	Tricarboxylic Acid Cycle	2
SUCLG2	succinate-CoA ligase GDP-forming beta subunit	Tricarboxylic Acid Cycle	3
SUGCT	succinyl-CoA:glutarate-CoA transferase	Amino Acid Metabolism	7
SUOX	sulfite oxidase	Amino Acid Metabolism	12
SUPV3L1	Suv3 like RNA helicase	Replication & Transcription	10
SURF1	SURF1 cytochrome c oxidase assembly factor	Oxidative Phosphorylation	9
SYBU	syntabulin	Mitochondrial Dynamics	8
SYNJ2BP	synaptojanin 2 binding protein	Mitochondrial Signaling	14
TACO1	translational activator of cytochrome c oxidase I	Translation	17
TALDO1	transaldolase 1	Pentose Phosphate Pathway	11
TAMM41	TAM41 mitochondrial translocator assembly and maintenance homolog	Import & Sorting	3
TARS2	threonyl-tRNA synthetase 2, mitochondrial	Translation	1
TAX1BP1	Tax1 binding protein 1	Mitophagy	7
TAZ	WW domain containing transcription regulator 1	Cardiolipin Biosynthesis	3
TBC1D15	TBC1 domain family member 15	Mitophagy	12
TBK1	TANK binding kinase 1	Mitophagy	12
TBRG4	transforming growth factor beta regulator 4	Oxidative Phosphorylation	7

TCAIM	T cell activation inhibitor, mitochondrial	Mitochondrial Dynamics	3
TCHP	trichoplein keratin filament binding	Calcium Signaling & Transport	12
TEFM	transcription elongation factor, mitochondrial	Replication & Transcription	17
TERT	telomerase reverse transcriptase	Replication & Transcription	5
TFAM	transcription factor A, mitochondrial	Replication & Transcription	10
TFB1M	transcription factor B1, mitochondrial	Replication & Transcription	6
TFB2M	transcription factor B2, mitochondrial	Replication & Transcription	1
TFE3	transcription factor binding to IGHM enhancer 3	Transcription (nuclear)	X
TFEB	transcription factor EB	Transcription (nuclear)	6
THEM4	thioesterase superfamily member 4	Fatty Acid Biosynthesis & Elongation	1
THG1L	tRNA-histidine guanylyltransferase 1 like	Translation	5
TIMM10	translocase of inner mitochondrial membrane 10	Import & Sorting	11
TIMM10B	translocase of inner mitochondrial membrane 10B	Import & Sorting	11
TIMM13	translocase of inner mitochondrial membrane 13	Import & Sorting	19
TIMM17A	translocase of inner mitochondrial membrane 17A	Import & Sorting	1
TIMM17B	translocase of inner mitochondrial membrane 17B	Import & Sorting	X
TIMM21	translocase of inner mitochondrial membrane 21	Import & Sorting	18
TIMM22	translocase of inner mitochondrial membrane 22	Import & Sorting	17
TIMM23	translocase of inner mitochondrial membrane 23	Import & Sorting	10
TIMM23B	translocase of inner mitochondrial membrane 23 homolog B	Import & Sorting	10
TIMM29	translocase of inner mitochondrial membrane 29	Import & Sorting	19
TIMM44	translocase of inner mitochondrial membrane 44	Import & Sorting	19
TIMM50	translocase of inner mitochondrial membrane 50	Import & Sorting	19
TIMM8A	translocase of inner mitochondrial membrane 8A	Import & Sorting	X
TIMM8B	translocase of inner mitochondrial membrane 8 homolog B	Import & Sorting	11
TIMM9	translocase of inner mitochondrial membrane 9	Import & Sorting	14
TIMMDC1	translocase of inner mitochondrial membrane domain containing 1	Import & Sorting	3
TK2	thymidine kinase 1	Replication & Transcription	17
TKFC	triokinase and FMN cyclase	Fructose Metabolism	11
TKT	transketolase	Pentose Phosphate Pathway	3
TLDC2	TBC/LysM-associated domain containing 2	ROS Defense	20
TMEM11	transmembrane protein 11	Mitochondrial Dynamics	17
TMEM126A	transmembrane protein 126A	Unknown	11
TMEM126B	transmembrane protein 126B	Oxidative Phosphorylation	11
TMEM143	transmembrane protein 143	Unknown	19
TMEM173	transmembrane protein 173	Apoptosis	5
TMEM177	transmembrane protein 177	Protein Stability & Degradation	2
TMEM186	transmembrane protein 186	Oxidative Phosphorylation	16
TMEM65	transmembrane protein 65	Calcium Signaling & Transport	8
TMEM70	transmembrane protein 70	Oxidative Phosphorylation	8
TMLHE	trimethyllysine hydroxylase, epsilon	Fatty Acid Metabolism	X
TOMM20	translocase of outer mitochondrial membrane 20	Import & Sorting	1
TOMM20L	translocase of outer mitochondrial membrane 20 like	Import & Sorting	14
TOMM22	translocase of outer mitochondrial membrane 22	Import & Sorting	22
TOMM34	translocase of outer mitochondrial membrane 34	Import & Sorting	20
TOMM40	translocase of outer mitochondrial membrane 40	Import & Sorting	19

TOMM40L	translocase of outer mitochondrial membrane 40 like	Import & Sorting	1
TOMM5	translocase of outer mitochondrial membrane 5	Import & Sorting	9
TOMM6	translocase of outer mitochondrial membrane 6	Import & Sorting	6
TOMM7	translocase of outer mitochondrial membrane 7	Import & Sorting	7
TOMM70	translocase of outer mitochondrial membrane 70	Import & Sorting	3
TOP1MT	DNA topoisomerase I mitochondrial	Replication & Transcription	8
TOP3A	DNA topoisomerase III alpha	Replication & Transcription	17
TP53	tumor protein p53	Transcription (nuclear)	17
TPI1	triosephosphate isomerase 1	Glycolysis	12
TRAK1	trafficking kinesin protein 1	Mitochondrial Dynamics	3
TRAK2	trafficking kinesin protein 2	Mitochondrial Dynamics	2
TRAP1	TNF receptor associated protein 1	Apoptosis	16
TRIAP1	TP53 regulated inhibitor of apoptosis 1	Apoptosis	12
TRIT1	tRNA isopentenyltransferase 1	Translation	1
TRMT10C	tRNA methyltransferase 10C, mitochondrial RNase P subunit	Translation	3
TRMT11	tRNA methyltransferase 11 homolog	Translation	6
TRMT2B	tRNA methyltransferase 2 homolog B	Translation	X
TRMT5	tRNA methyltransferase 5	Translation	14
TRMT61B	tRNA methyltransferase 61B	Translation	2
TRMU	tRNA 5-methylaminomethyl-2-thiouridylate methyltransferase	Translation	22
TRNT1	tRNA nucleotidyl transferase 1	Translation	3
TRUB2	TruB pseudouridine synthase family member 2	Replication & Transcription	9
TSFM	Ts translation elongation factor, mitochondrial	Translation	12
TSPO	translocator protein	Apoptosis	22
TST	thiosulfate sulfurtransferase	Amino Acid Metabolism	22
TSTD1	thiosulfate sulfurtransferase like domain containing 1	Oxidative Phosphorylation	1
TTC19	tetratricopeptide repeat domain 19	Oxidative Phosphorylation	17
TUFM	Tu translation elongation factor, mitochondrial	Translation	16
TWNK	twinkle mtDNA helicase	Replication & Transcription	10
TXN	thioredoxin	ROS Defense	9
TXN2	thioredoxin 2	ROS Defense	22
TXNRD1	thioredoxin reductase 1	ROS Defense	12
TXNRD2	thioredoxin reductase 2	ROS Defense	22
TYMS	thymidylate synthetase	Nucleotide Metabolism	18
UBA1	ubiquitin like modifier activating enzyme 1	Protein Stability & Degradation	X
UBL5	ubiquitin like 5	UPRmt	19
UCP1	uncoupling protein 1	Mitochondrial Carrier	4
UCP2	uncoupling protein 2	Mitochondrial Carrier	11
UCP3	uncoupling protein 3	Mitochondrial Carrier	11
ULK1	unc-51 like autophagy activating kinase 1	Mitophagy	12
UNG	uracil DNA glycosylase	Replication & Transcription	12
UQCC1	ubiquinol-cytochrome c reductase complex assembly factor 1	Oxidative Phosphorylation	20
UQCC2	ubiquinol-cytochrome c reductase complex assembly factor 2	Oxidative Phosphorylation	6
UQCC3	ubiquinol-cytochrome c reductase complex assembly factor 3	Oxidative Phosphorylation	11
UQCR10	ubiquinol-cytochrome c reductase, complex III subunit X	Oxidative Phosphorylation	22
UQCR11	ubiquinol-cytochrome c reductase, complex III subunit XI	Oxidative Phosphorylation	19
UQCRB	ubiquinol-cytochrome c reductase binding protein	Oxidative Phosphorylation	8
UQCRC1	ubiquinol-cytochrome c reductase core protein 1	Oxidative Phosphorylation	3

UQCRC2	ubiquinol-cytochrome c reductase core protein 2	Oxidative Phosphorylation	16
UQCRFS1	ubiquinol-cytochrome c reductase, Rieske iron-sulfur polypeptide 1	Oxidative Phosphorylation	19
UQCRH	ubiquinol-cytochrome c reductase hinge protein	Oxidative Phosphorylation	1
UQCRQ	ubiquinol-cytochrome c reductase complex III subunit VII	Oxidative Phosphorylation	5
UROD	uroporphyrinogen decarboxylase	Heme Biosynthesis	1
UROS	uroporphyrinogen III synthase	Heme Biosynthesis	10
USP15	ubiquitin specific peptidase 15	Mitophagy	12
USP30	ubiquitin specific peptidase 30	Mitochondrial Dynamics	12
USP8	ubiquitin specific peptidase 8	Mitophagy	15
VARS2	valyl-tRNA synthetase 2, mitochondrial	Translation	6
VDAC1	voltage dependent anion channel 1	Transmembrane Transport	5
VDAC2	voltage dependent anion channel 2	Transmembrane Transport	10
VDAC3	voltage dependent anion channel 3	Transmembrane Transport	8
WARS2	tryptophanyl tRNA synthetase 2, mitochondrial	Translation	1
WDR81	WD repeat domain 81	Mitochondrial Dynamics	17
XIAP	X-linked inhibitor of apoptosis	Apoptosis	X
XPNPEP3	X-prolyl aminopeptidase 3	Protein Stability & Degradation	22
YAP1	Yes associated protein 1	Transcription (nuclear)	11
YARS2	tyrosyl-tRNA synthetase 2	Translation	12
YME1L1	YME1 like 1 ATPase	Import & Sorting	10
YRDC	yrdC N6-threonylcarbamoyltransferase domain containing	Replication & Transcription	1
YY1	YY1 transcription factor	Transcription (nuclear)	14