Sequencing and bioinformatics analysis of the human mitochondrial genome

Alexandrina Bodrug
CHU Angers, Service Génétique, équipe bioinfo

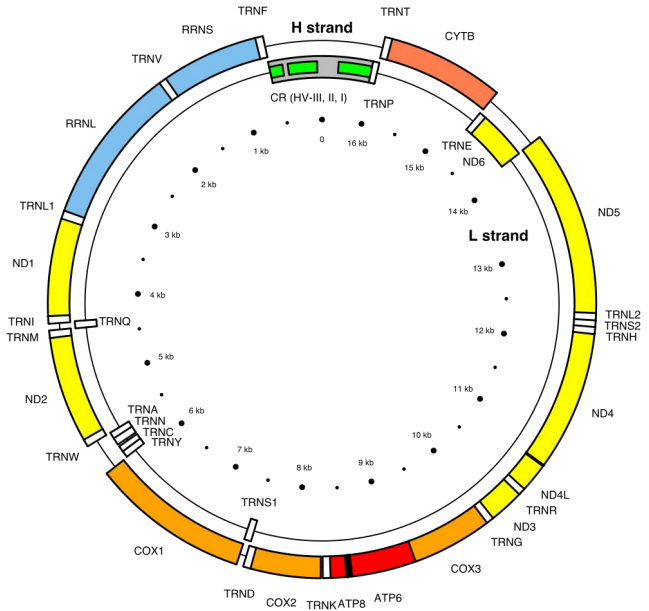
Séminaire BioInfoDiag
17.04.2023

Key points: mtDNA genome variation, proliferation and clinico-genetic heterogeneity

mtDNA key points

- 16,569 bp
- 37 genes
- Regulatory region
- multiple copies per organelle (2-10)
- multiple organelles per cell (up to 1k)

→ ntDNA, discreet allele frequency 0 / 0,5 / 1
→ mtDNA, continuous distribution



mtDNA key points

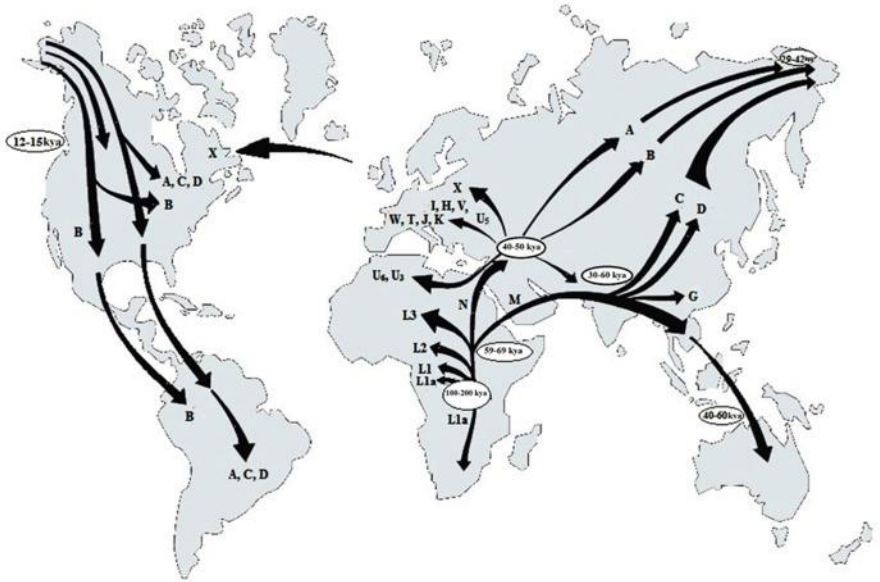
wild type ：

- rCRS: NC_012920.1 (hg38)
- NC_001807.4 (hg19)

Reference mtDNA is **haplogroup** H → Very present in databases.

Haplogroup markers of J haplogroup, in respect to the reference mtDNA:

73A>G, **263**A>G, **750**A>G, **1438**A>G, **2706**A>G, **4769**A>G, **7028**C>T, **8860**A>G, **10398**A<G, **11719**G>A, **14766**C>T, **15326**A>G



©Guha P et al. (2013)

mtDNA reference

hg19: Cambridge Reference Sequence
NC_001807.4

hg38: Revised Cambridge Reference Sequence
NC_012920.1

- Corrected 11 errors
- Noted 7 private/rare polymorphisms: 263A, 311C-315C, 750A, 1438A, 4769A, 8860A, 15326A

→ real sequence, not a consensus

©Anderson S et al. (1981)

©Andrews RM et al. (1999)

mtDNA variations and polymorphisms

Haplogroup markers → homoplasmic, fixed variations

Private and rare polymorphisms → *de novo* or inherited

Variants of unknown significance

Pathogenic variants

- ~ 100 known pathogenic variants
- Few are heteroplasmic

mtDNA variations and polymorphisms

Haplogroup markers → homoplasmic, fixed variations

Private and rare polymorphisms

Variants of unknown significance

Pathogenic variants

- ~ 100 known pathogenic variants
- Few are heteroplasmic

How and why do pathogenic mutations arise in mtDNA?

mtDNA proliferation

Vegetative segregation (dividing cells)

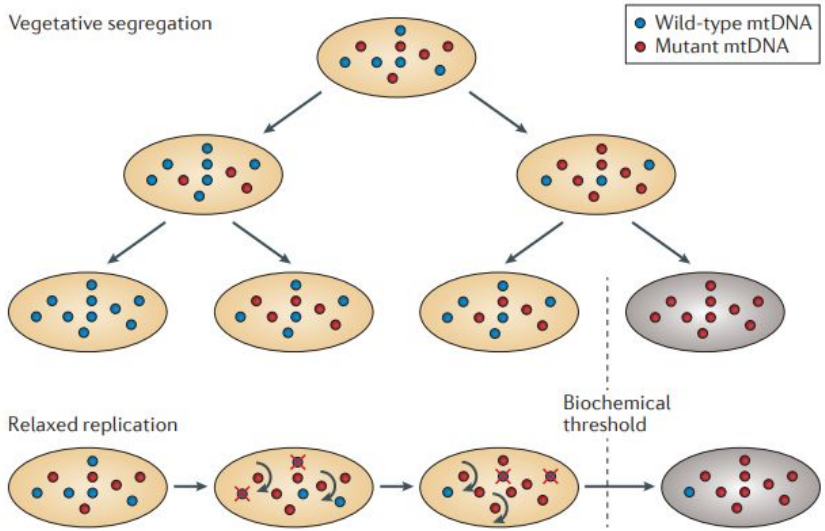
- mtDNA can segregate without replication, altering heteroplasmy in child cells

Relaxed replication (non dividing cells)

- certain mtDNA can replicate preferentially

Primordial germ cell bottleneck

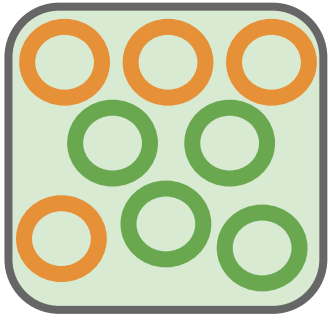
- particular selective forces in PGC
- some mutation will never be fixed because of biochemical threshold → always heteroplasmic



©Steward JB et al. (2015)

mitochondrial mutation load

50% heteroplasmy



Enough wild type genomes to maintain function

50% heteroplasmy

Not enough wild type genomes to maintain function

In mitochondrial diagnostics, heteroplasmy is often routinely reported with to NGS.

Wild type mitochondrial load is a separate sequencing experiment that is often lacking in diagnostics.

heteroplasmic mtDNA variation

Locus	Associated Diseases	Allele
MT-TL1	MELAS / Leigh Syndrome / DMDF / MIDD / SNHL / CPEO / MM / FSGS / ASD / Cardiac+multi-organ dysfunction	m.3243A>G

©MITOMAP

3243A>G is always heteroplasmic

80% of patients with a pathogenic heteroplasmic variation have the 3243A>G.

homoplasmic pathogenic mtDNA variations

Example: Leber's hereditary optic neuropathy

- Most common mitochondrial disease
- Caused mainly by 3 homoplasmic variations:
  - 11778G>A
  - 3460G>A
  - 14484T>C

→ variations present in all maternally related individuals to the proband, but not all display the disease: **incomplete penetrance**

Locus	Associated Diseases	Allele
MT-ND1	LHON MELAS overlap	m.3376G>A
MT-ND1	LHON	m.3460G>A
MT-ND1	LHON	m.3635G>A
MT-ND1	LHON	m.3700G>A
MT-ND1	LHON	m.3733G>A
MT-ND1	LHON / Leigh-like phenotype	m.4171C>A
MT-ND4L	LHON	m.10663T>C
MT-ND4	LHON / Progressive Dystonia	m.11778G>A
MT-ND5	LHON	m.13051G>A
MT-ND5	Ataxia+PEO / MELAS, LD, LHON, myoclonus, fatigue	m.13094T>C
MT-ND5	LHON	m.13379A>G
MT-ND5	Leigh Disease / MELAS / LHON-MELAS Overlap Syndrome / negative association w Carotid Atherosclerosis	m.13513G>A
MT-ND6	LHON	m.14482C>G
MT-ND6	LHON	m.14482C>A
MT-ND6	LHON	m.14484T>C
MT-ND6	LHON	m.14495A>G
MT-ND6	LHON	m.14568C>T

homoplasmic pathogenic mtDNA variations

Example: Leber's hereditary optic neuropathy

- mutant mtDNA load?
- nuclear mutations?
- transcriptomic modifiers?

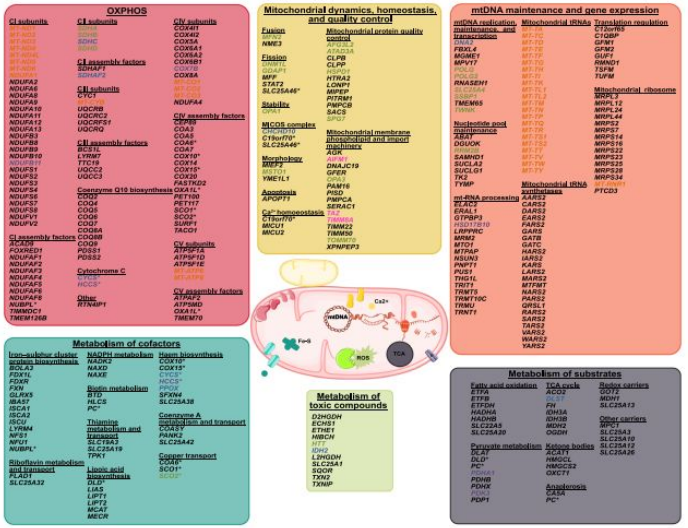
"A compensatory strategy to mitochondrial dysfunction commonly observed in mitochondrial diseases is the increase of ***mitochondrial biogenesis****, as exemplified by the massive proliferation of mitochondria in skeletal muscle fibres (DiMauro and Schon, 2003)."*

©Giordano et al. (2014)

Analysing mtDNA is not straightforward!

Analysis challenges:

- Heteroplasmy threshold
- tissue specific heteroplasmy
- tissue specific targeting
- Incomplete penetrance
- Mutation load
- *de novo* variations vs inherited variations
- variants of unknown significance
- Background variation (haplorgoup)
- Nuclear regulatory genes



©Gusic et al. (2021)

Analysing mtDNA is not straightforward!

Routine diagnostics 

- sequenced tissue != affected tissue
- missing family history
- missing mutation load info
- missing nuclear encoded mitochondrial genes

Routine diagnostics 

- multiple trusted pipelines to retrieve variants
- heteroplasmy info
- extraction of variants of interest
- nuMTs are not an issue (in case of exome data)
- increasing number of probands with multiple sequenced tissues

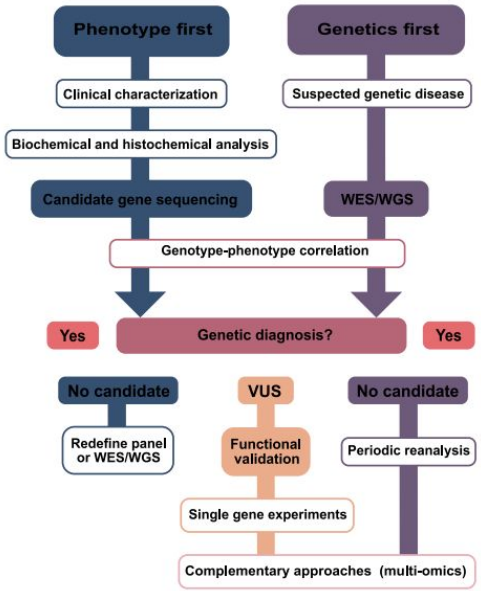
mtDNA sequencing and analysis for diagnostics

Motivation behind sequencing mtDNA

Routine and mainstream practice for:

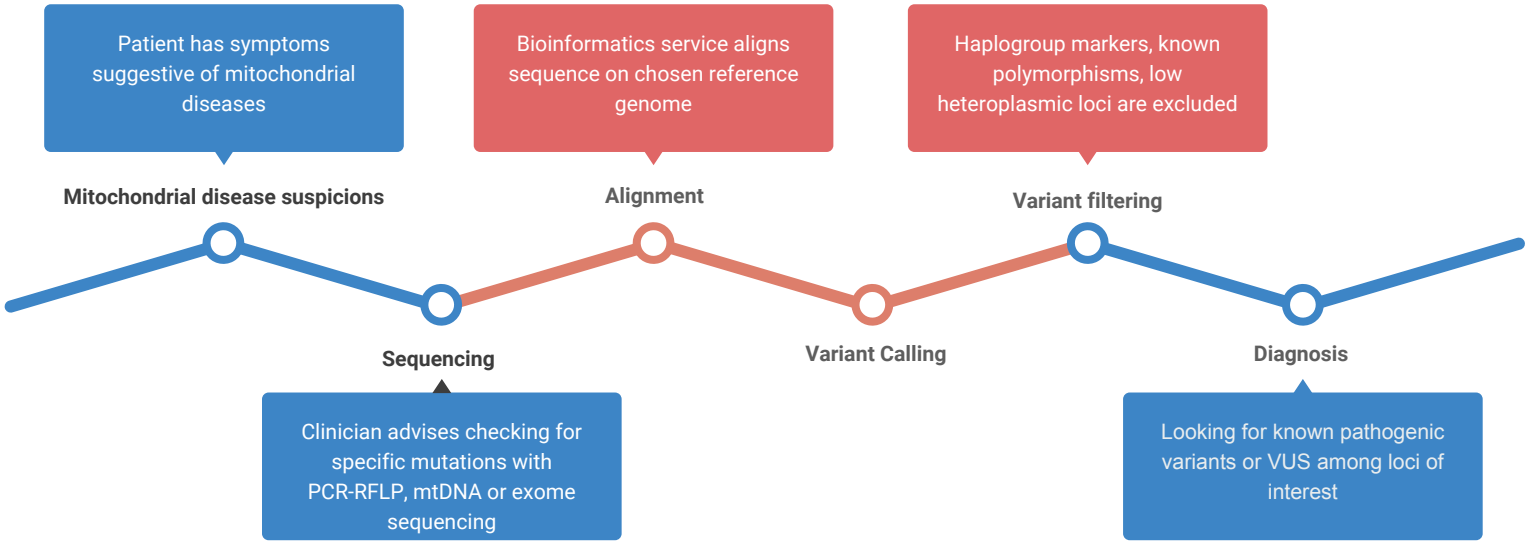
- Genetic counseling
- Treatment options
- Reproductive options

Transition from **Phenotype first** to **Genotype first**



©Gusic M et al. (2021)

General steps of mtDNA analysis for diagnostics



mtDNA sequence source

Sequence source:

- Exome off target
- Exome on target
- mtDNA sequencing
- PCR-RFLP used for confirming low heteroplasmy variants
- Dels and dups: Southern blot, real time PCR, long range PCR

Sequencing platforms:

- Illumina (CHU Angers)
- IonTorrent (CHU Angers)
- Long molecule technologies

Extracting mtDNA from exome data

Goals:

- Getting rid of nuMTs
- Obtaining enough mtDNA coverage

Motivation:

Vast phenotypic overlap with other disease and absence of reliable biomarkers → integration of unbiased methodologies early in the diagnostics

Solutions:

- MToolBox: alignment to 1) mt chromosome 2) hg19
- CHU Anger + pipelinemito (CHU Dijon): aligning to hg38 + rCRS simultaneously

GOLD projet at CHU Angers

Goal: building a methodology to extract mtDNA from WGS and evaluate the tissue influence

Methodology:

- 10 individuals from Gazel cohorte
- Blood and saliva samples
  - mtDNA IonTorrent (positive control)
  - Illumina WGS

Results:

- (re)Identification of pathogenic variants
- (re)Identification of haplogroups
- Successful exclusion of nuMTS reads and nuMTS variations
- Validated heteroplasmies, with low variation between blood and saliva

Analysis pipeline at CHU Angers

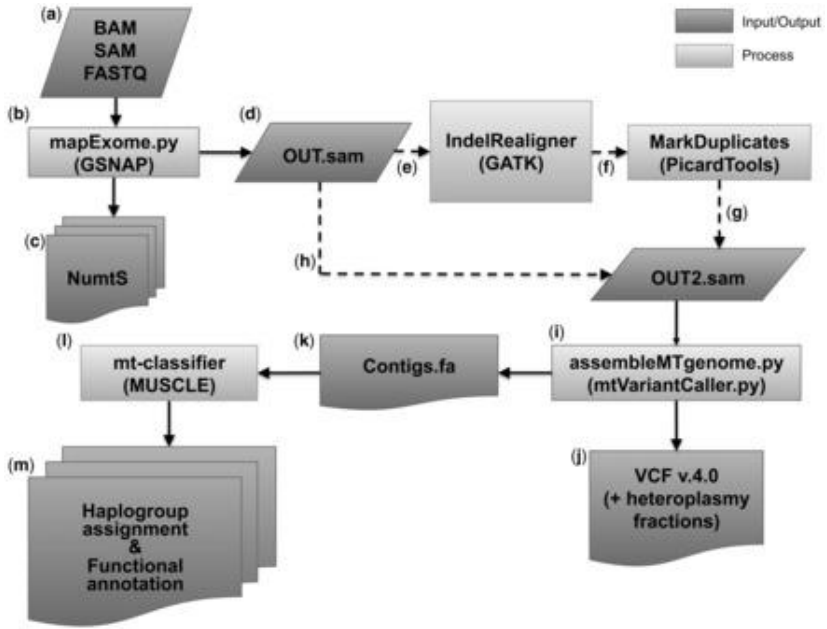
- Sequencing with IonTorrent or Illumina of **mitochondrial DNA**
- Quality check fastq
- Alignment with bwa-mem to **rCRS**
- Quality check bams
- Variant calling with: Strelka2, Mutect, TSVC, DeepVariant, GATK
- Determining the haplogroup (Haplogrep2)
- Excluding haplogroup markers, polymorphisms highly present in local and publicly available databases

Candidate variants:

- Rare variants in population (polymorphisms?)
- Rare variants in our local DB (artefacts?)
- High enough heteroplasmy
- *de novo*?
- other known similar variants (same region?)

MTOOLBox (Bologna)

©Calabrese C et al. (2014)

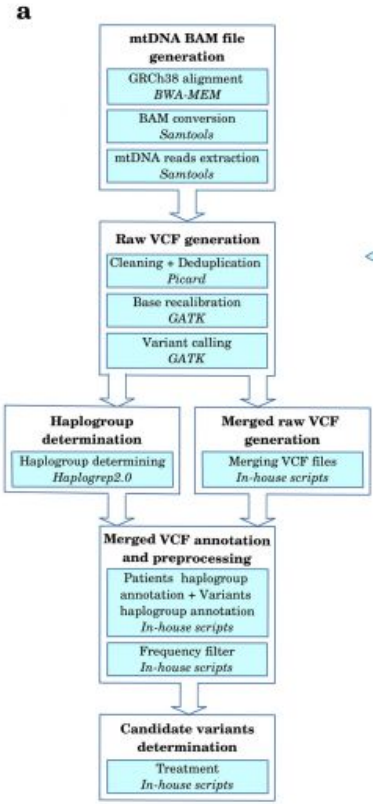


- Alignment to mtDNA (GSNAP) then to hg19
- Mitochondrial assembly for haplogroup assignment via alignment to macro-haplogroup consensus sequences
- Realignment around known indels (HmtDB & MITOMAP)

pipelinemito (Dijon)

©Garret P et al. (2020)

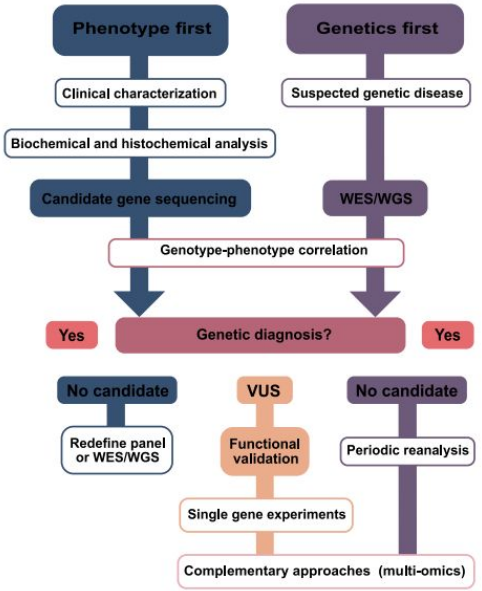
- Alignment to hg38 + mtDNA → better to avoid realignment to nuMTS and avoid bias in heteroplasmy rate
- Haplogrep2 to determine haplogroup



Going back to diagnostics

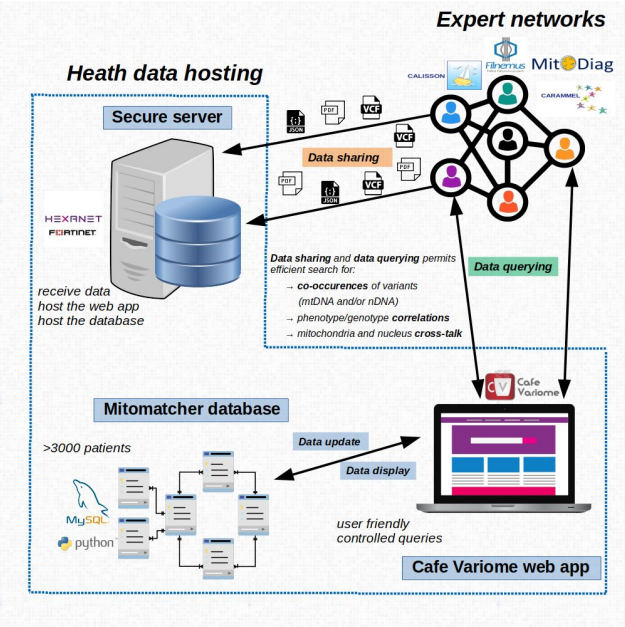
Genetics first approach:

1) Sequence mtDNA→ exome→WGS (+maternal/sibling sequencing?)
2) Diagnostics if known pathogenic variant at high heteroplasmy (>5%) OR
3) VUS with clinical symptoms corresponding to clinico-genetic litterature description



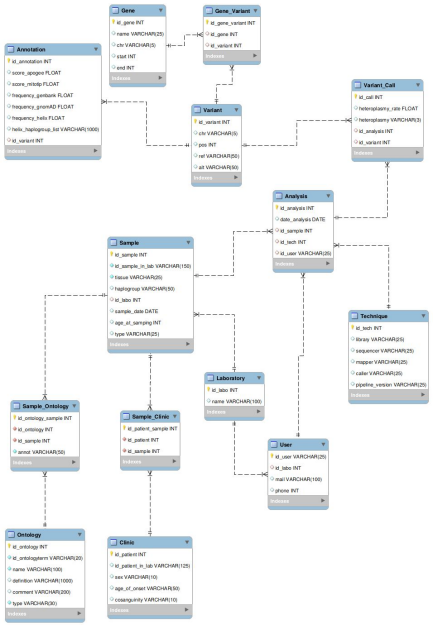
©Gusic M et al. (2021)

What to do with uncertain finds: mitomatcherDB



- Clinical-genetic database
- Probands/patients with HPO terms
- Facilitate the finding of similar patients or nomad patients

What to do with uncertain finds: mitomatcherDB



- Several samples per individual (blood, urine)
- Several variant callings per sample (sequencing, pipelines)
- One ontology per individual!

Conclusion

- Moving to genetics first diagnostics approach of mitochondrial diseases
- Several methodologies to exploit existing exome data
- Candidate variants are often rare at population level, with high enough heteroplasmy
- Inconclusive candidate variants should be stored for further investigation in DBs

Thanks to the CHU Angers bioinformatics team

Main bibliography

Diroma, M.A., Calabrese, C., Simone, D. *et al.* **Extraction and annotation of human mitochondrial genomes from 1000 Genomes Whole Exome Sequencing data**. *BMC Genomics* 15 (Suppl 3), S2 (2014). https://doi.org/10.1186/1471-2164-15-S3-S2

Garret P, Bris C, Procaccio V, Amati-Bonneau P, Vabres P, Houcinat N, Tisserant E, Feillet F, Bruel AL, Quéré V, Philippe C, Sorlin A, Tran Mau-Them F, Vitobello A, Costa JM, Boughalem A, Trost D, Faivre L, Thauvin-Robinet C, Duffourd Y. **Deciphering exome sequencing data: Bringing mitochondrial DNA variants to light**. Hum Mutat. 2019 Dec;40(12):2430-2443. doi: 10.1002/humu.23885. Epub 2019 Aug 26. PMID: 31379041.

Zhang H, Esposito M, Pezet MG, Aryaman J, Wei W, Klimm F, Calabrese C, Burr SP, Macabelli CH, Viscomi C, Saitou M, Chiaratti MR, Stewart JB, Jones N, Chinnery PF. **Mitochondrial DNA heteroplasmy is modulated during oocyte development propagating mutation transmission**. Sci Adv. 2021 Dec 10;7(50):eabi5657. doi: 10.1126/sciadv.abi5657. Epub 2021 Dec 8. PMID: 34878831; PMCID: PMC8654302.

Gusic M, Prokisch H. **Genetic basis of mitochondrial diseases**. FEBS Lett. 2021 Apr;595(8):1132-1158. doi: 10.1002/1873-3468.14068. Epub 2021 Mar 21. PMID: 33655490.