CIDR
Center for Inherited Disease Research
Johns Hopkins University

Calling mitochondrial DNA (mtDNA) variants from whole exome sequencing (WES) data

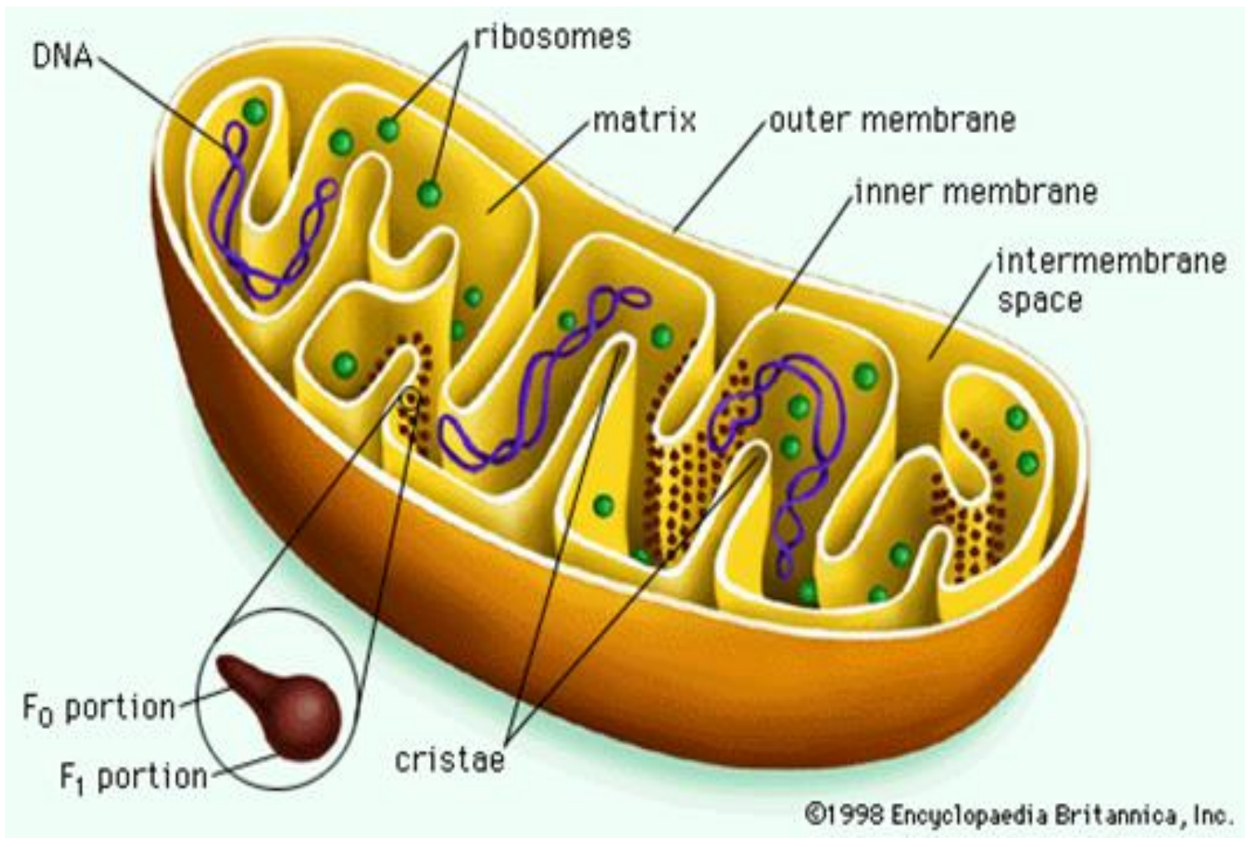
JOHNS HOPKINS
MEDICINE

Peng Zhang[1], Hua Ling[1], Kurt Hetrick[1], Elizabeth Pugh[1], Dane Witmer[1], Jun Ding[2], Nara Sobreira[3], David Valle[3], Kimberly Doheny[1]

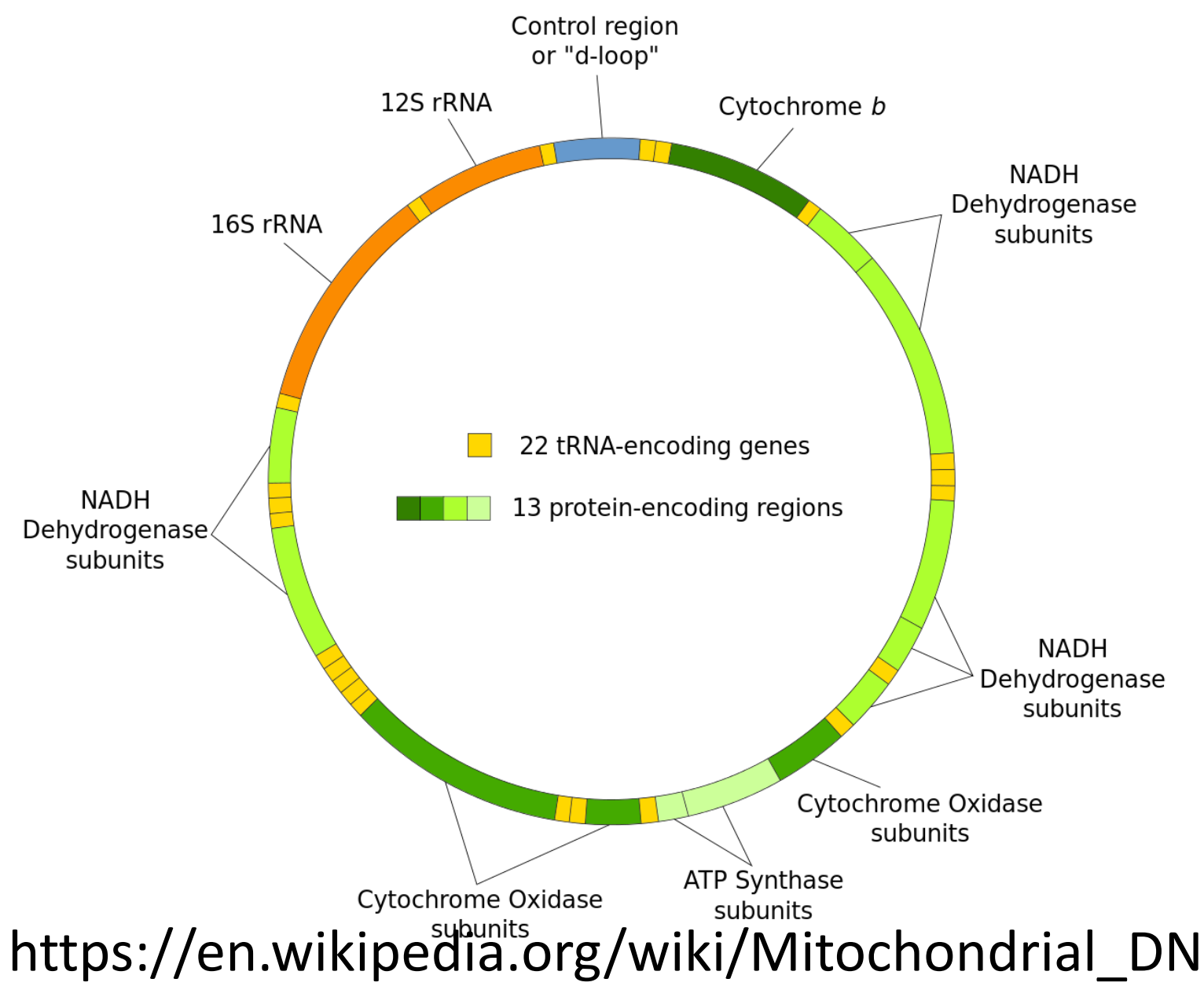
[1]Center for Inherited Disease Research, Institute of Genetic Medicine, The Johns Hopkins School of Medicine
[2]Laboratory of Genetics, National Institute on Aging, NIH, Baltimore, MD
[3]Institute of Genetic Medicine, Johns Hopkins University School of Medicine

Introduction

Mitochondrion functions as the “cellular power plant” and generates most of the cell’s supply of ATP. Human mitochondrial DNA (mtDNA) is a circular molecule of 16,569 bases, 100 to 10,000 copies per cell, and encodes 37 genes (2 rRNAs, 22 tRNAs and 13 polypeptides). Each mitochondrion encodes some of its constituent proteins in resident DNA. Mitochondrion is also involved in a range of other processes such as signaling, cellular differentiation, and cell death.



Mitochondrion diagram. Credit: ffis.es


Control region or "d-loop"
12S rRNA
Cytochrome *b*
16S rRNA
NADH Dehydrogenase subunits
22 tRNA-encoding genes
13 protein-encoding regions
NADH Dehydrogenase subunits
NADH Dehydrogenase subunits
Cytochrome Oxidase subunits
ATP Synthase subunits
Cytochrome Oxidase subunits

https://en.wikipedia.org/wiki/Mitochondrial_DNA

As part of the Baylor Hopkins Center for Mendelian Genomics (BHCMG) (http://www.mendelian.org/), CIDR has performed whole exome sequencing (WES) for over 900 samples to discover the genetic basis for Mendelian conditions. In this study, we examined whether the off-target reads of mitochondrial DNA (mtDNA) generated by WES can be used to study mtDNA variation. As many large-scale genetic studies are collecting WES data, the method and protocol developed by us could have wide application.

We extracted off-target mtDNA reads from our WES data and called **homoplasmies** (positions with all the same non-reference alleles) and **heteroplasmies** (positions with a mixture of two or more alleles) using MitoCaller, a software application designed to call both homoplasmies and heteroplasmies from next-generation whole genome sequencing data (Ding et al. PLoS Genetics 2015*, ASHG Poster# 1287W*). We re-evaluated the data quality control steps for our WES data. We then used ANNOVAR to get the functional annotations of the mtDNA variant calls.

Materials and methods

Sample selection:

- 787 BHCMG samples with WES data.

Sequencer and reagents:

- Exome Capture: Agilent SureSelect HumanAllExonV4.
- Illumina HiSeq2500 platform (Majority).
- TruSeq Rapid SBS-HS 100 bp Paired Ends (Majority).

Sequencing data processing:

- BWA mem 0.7.8 alignment, local alignment and base call quality score recalibration with GATK 3.1-1.
- Samtools 0.1.18 extract mtDNA reads, exclude secondary alignment.
- MitoCaller to call mtDNA variants.
- ANNOVAR annotate homoplasmy and heteroplasmy calls.

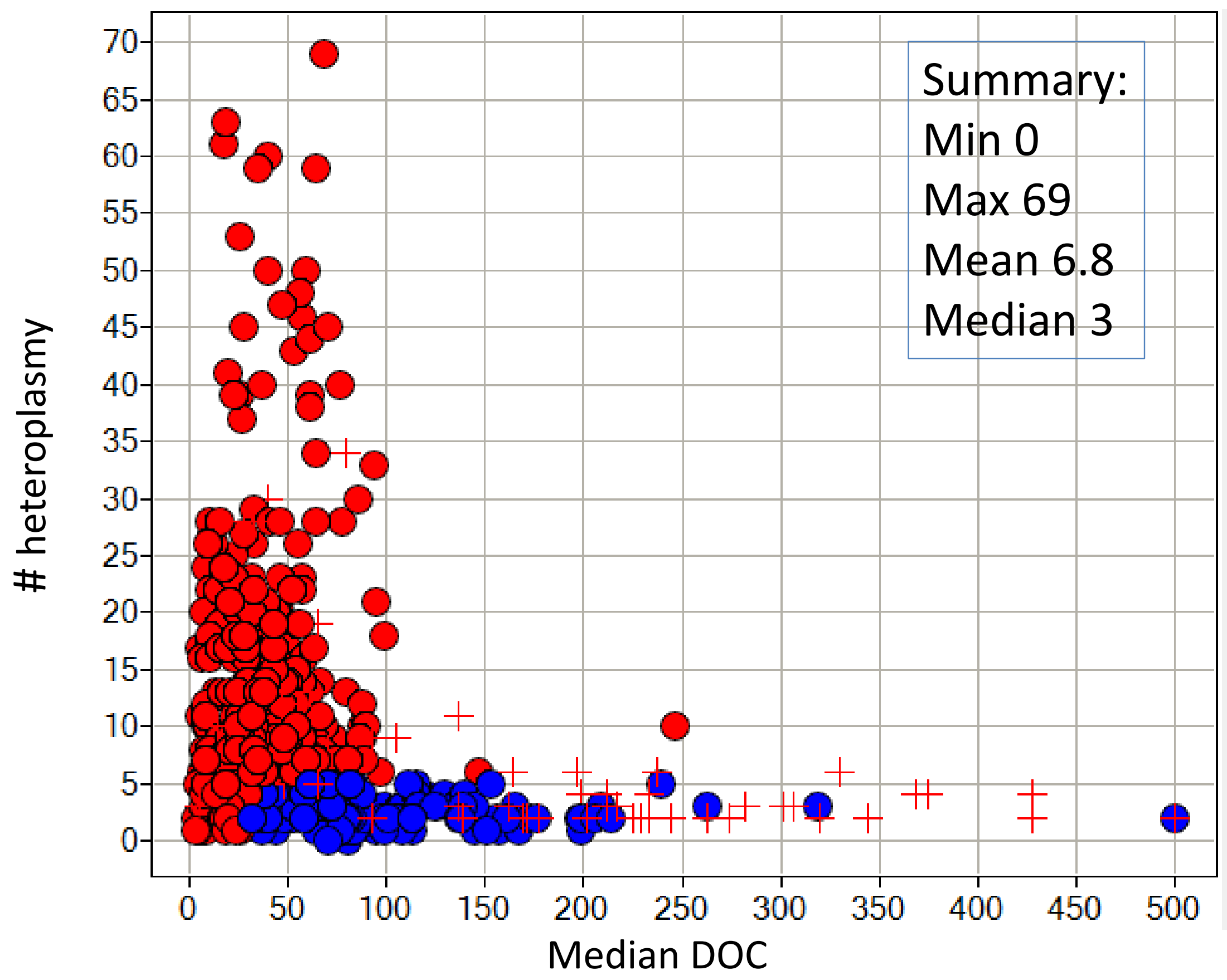
Results

The DOC in the targeted regions of our samples ranged from 50x to 322x with a median of 93x (percentage DOC on target at least 10x ranges from 85.9% to 99.6%).

For the off-target reads, we observed a mtDNA median depth of coverage (DOC) ranging from 3x to 500x in the 787 samples, and DOC varied significantly by DNA source.

Ding et al. (2015) suggest a DOC threshold of 100x for mtDNA, but since we only have a 21.5% of samples achieved a median DOC of 100x or above, we evaluated lower threshold of DOC. We looked at number of heteroplasymies called and DNA source. For a subset of data we manually evaluated 1) maternal and paternal inheritance 2) consistency of calls between replicates. 3) used downsampling to see how the number of calls was affected by various DOCs.

Using a 4% minor allele fraction threshold for calling heteroplasmies, we present results using samples with DOC of 30x and above, number of heteroplasmies call of 5 or less, and DNA source from blood or saliva for downstream analysis (364 calls from 174 samples).



Scatter plot of heteroplasmy call (vertical axis) versus median DOC (horizontal axis) for 787 samples. The summary statistics of calls over 787 samples. The blue ● are the samples selected for downstream analysis, the red ● represents the samples filtered from down stream analysis, red + are other types or unknown DNA sources.

Distribution of homoplasmies and heteroplasmies in functional categories.

	Intergenic	Non-coding RNA	Protein-coding		
			Synonymous	Non-synonymous	Stop gain/loss
Homoplasmy (n = 4490)	1123 (25.0%)	819 (18.2%)	1588 (35.4%)	957 (21.3%)	3 (loss) (0.07%)
Heteroplasmy (n = 364)	97 (26.6%)	66 (18.1%)	91 (25.0%)	105 (28.8%)	5 (gain) (1.37%)

Distribution of protein-coding heteroplasmies versus disease status.

	Affected (n =93)	Not affected (n = 31)
Synonymous	65 (74.7%)	22 (25.3%)
Non-synonymous	71 (72.4%)	27 (27.6%)
Stopgain	4 (80.0%)	1 (20.0%)

Summary

- We have showed the feasibility of calling mtDNA variants using off-target reads from WES data.
- We have annotated the called mtDNA variants into five functional categories with ANNOVAR.
- Because of the target region pooling step in WES, the off-target mtDNA reads from WES are usually lower than those from WGS (Whole Genome Sequencing) data.
- Adding a threshold for the number of heteroplasmies and restricting sample DNA source allowed us to use a lower threshold for DOC.
- Future work includes 1) refining the threshold of calling heteroplasmies with other data quality control statistics and 2) identifying potential causal mtDNA variants for phenotype of interests.