**MitoImpute: A Snakemake pipeline for imputation of mitochondrial genetic variants**

Tim W McInerney1, Brian Fulton-Howard2, Christopher Patterson3,4, Devashi Paliwal1, Lars S Jermiin5,6,7,8, Hardip R Patel1, Judy Pa3,4, Russell H Swerdlow9, Alison Goate2, Simon Easteal1, Shea J Andrews2\*, for the Alzheimer’s Disease Neuroimaging Initiative†

1The John Curtin School of Medical Research, The Australian National University, Canberra, Australian Capital Territory, Australia

2Ronald M. Loeb Center for Alzheimer’s disease, Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, New York, USA

3Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

4Department of Neurology, Alzheimer’s Disease Research Center, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

5CSIRO Land & Water, Commonwealth Scientific Industrial & Research Organization, Acton, ACT 2601, Australia

6Research School of Biology, Australian National University, Canberra, ACT 2601, Australia

7School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland

8Earth Institute, University College Dublin, Belfield, Dublin 4, Ireland

9Department of Neurology, Alzheimer’s Disease Center, University of Kansas, Fairway, KS, USA

\*Correspondence to: Shea Andrews, The Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Place, New York, NY 10029, USA.

Tel: +1-212-659-8632; E-mail: [shea.andrews@mssm.edu](mailto:shea.andrews@mssm.edu)

†Data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at:

<http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf>

# **Abstract**

Motivation: Many available genotyping microarrays do not include sufficient mitochondrial single nucleotide variants (mtSNVs), limiting the utility of microarray chips to infer missing genotypes and subsequently to accurately assign sequences to their correct haplogroup. To address this, we created an easy to use mitochondrial DNA imputation pipeline, MitoImpute, which infers missing genotypes using a reference panel of 36,960 publicly available human mitochondrial genome sequences.

Results: We validated imputation accuracy by measuring haplogroup and genotype concordance in two datasets; (a) the 1,000 Genomes Project and (b) the Alzheimer’s Disease Neuroimaging Initiative. In the 1,000 Genome Datasets, we observed a mean Matthew’s correlation coefficient of MCC=0.64, leading to a mean 42.7% improvement in haplogroup assignment. In ADNI, we infer missing genotypes and demonstrated that MitoImpute can be utilised by long-term studies whose older datasets have limited mtSNV genotypes, thus making them useful for combining with resequenced datasets.

Availability

<https://github.com/sjfandrews/MitoImpute>

Supplementary Information.

Supplementary data are available at Bioinformatics online

## 

## **Introduction**

Mitochondrial DNA (mtDNA) variation is informative about human evolution and can be associated with disease [(Gorman *et al.*, 2016)](https://paperpile.com/c/6nYaGH/uCdp). Variation in these datasets is often described in the context of established haplotype groups (haplogroups), which represent branch points in the mtDNA phylogeny, with higher-order branch points representing major macro-haplogroups. However, microarrays used for typing mtDNA single nucleotide variants (mtSNVs) may not include sufficient mtSNVs to accurately define mtDNA haplogroups. A more complete approach is required.

The ability to accurately impute missing mtSNVs is dependent on the quality of the reference panel being used. To ensure that allelic frequencies, and therefore probabilities of observing a given haplotype, are accurately estimated, it is pertinent that nucleotide sites are correctly aligned to their homologous reference positions. Variation in sequence length between samples due to insertions and deletions causes gap character states to be placed in sequences so that nucleotides sites will be aligned to their correct homologous position. Where these gaps are placed, however, is dependent on the sampled sequences; differing random subsamples can lead to inconsistent gap placement. Aligning newly sampled sequences to a reference alignment incorporates prior decisions of gap placement into the alignment process, whereas aligning to just a reference sequence incorporates no prior knowledge. Therefore, it is pertinent that an imputation reference panel be constructed by aligning sequences to a high-quality reference alignment to ensure imputation accuracy.

Imputation of missing variants depends upon the haplotypes of the sequences in the Reference Panel. Samples from well-represented populations, such as Europeans and East Asians [(Popejoy and Fullerton, 2016; Sirugo *et al.*, 2019)](https://paperpile.com/c/6nYaGH/uXSH+sOjx), are likely to have their haplotypes represented in a large reference panel even if some of the samples have rare haplotypes. Samples from under-represented populations, however, are less likely to have their haplotypes represented, even if they are common in their population, due to an ascertainment bias towards European and East Asian populations [(Popejoy and Fullerton, 2016)](https://paperpile.com/c/6nYaGH/uXSH). For instance, indigenous populations from the Pacific Islands and Australia are not represented in the 1,000 Genomes Project phase 3 dataset. This presents a significant problem for both medical and evolutionary research, as historical datasets from these populations would need to have missing variants imputed from other populations that do not include haplotypes present in the sample panel populations. Inability to capture haplotypes with a disease-causing variant will lead to a failure to understand the genetics behind that disease [(Popejoy and Fullerton, 2016; Sirugo *et al.*, 2019)](https://paperpile.com/c/6nYaGH/uXSH+sOjx).

We present MitoImpute, a pipeline to infer missing mtDNA genotypes and assign haplogroups from globally-representative reference panels of mtDNA sequences. The performance of MitoImpute is validated using *in silico* microarrays (ISMs) derived from 1,000 Genomes Project [(1000 Genomes Project Consortium *et al.*, 2015)](https://paperpile.com/c/6nYaGH/B4Lf) whole genome sequence (WGS) data, and empirical data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [(Saykin *et al.*, 2010)](https://paperpile.com/c/6nYaGH/EPF5).

## **Methods**

### **Reference Alignment**

We used publicly available PhyloTree [(van Oven and Kayser, 2009)](https://paperpile.com/c/6nYaGH/bEt5) sequences to create a large (*n*=7,747) reference alignment with the revised Cambridge Reference Sequence (rCRS) [(Andrews *et al.*, 1999)](https://paperpile.com/c/6nYaGH/kw3a) site numbering convention. The rCRS remains within the reference alignment so that site numbering conventions can be maintained and verified as new sequences are aligned. We aligned sequences in batches of 50 using the L-INS-i version of MAFFT [(Katoh and Standley, 2013)](https://paperpile.com/c/6nYaGH/xYfv), then combined the batches, resolving inconsistent gap placements manually. rCRS site numbers were preserved by removing sites at which gaps were introduced in the rCRS during the alignment process.

### **Reference Panel**

Whole human mtDNA sequences were downloaded from GenBank on 2018-07-18 by adapting the MitoMap [(Lott *et al.*, 2013)](https://paperpile.com/c/6nYaGH/rRzH) search term (Supplementary Methods). This returned 44,299 complete human mtDNA sequences and excluded archaic and ancient sequences. These sequences were aligned to the reference alignment in batches of 2,500 using the MAFFT algorithm [(Katoh and Standley, 2013)](https://paperpile.com/c/6nYaGH/xYfv) in Geneious v10.2.6 [(Kearse *et al.*, 2012)](https://paperpile.com/c/6nYaGH/gPUL). Sites introducing gaps in the reference alignment were removed to maintain consistent nucleotide position numbering with rCRS. To improve the quality of the Reference Panel, sequences containing ≥5 ambiguous characters or ≥8 gaps were removed from the alignment. This threshold was set to enable inclusion of haplogroup B sequences which averaged 7 gaps relative to other sequences. Following this quality control, the Reference Panel contained 36,960 sequences (Supplementary Table 1).

The Reference Panel differs from the Reference Alignment as it is a well-curated exhaustive dataset of pre-aligned sequences that researchers can use to impute missing mtSNVs and subsample sequences from, without the need to perform an alignment process.

### **Validation Panel**

ISMs were created by subsetting mtSNVs present in 1000 Genomes Project Phase 3 WGS data (*n*=2,535) to those included on existing commercially available microarrays. Microarray information was obtained from strand orientation files available from the Wellcome Centre (<http://www.well.ox.ac.uk/~wrayner/strand/>), with 101 strand files containing mtSNVs (Supplementary Table 2). Haplogroup assignment for the WGS data and the ISMs was performed using Hi-MC [(Smieszek *et al.*, 2018)](https://paperpile.com/c/6nYaGH/s4qF).

### **Imputation**

We used the IMPUTE2 chromosome X protocol for imputation [(Howie *et al.*, 2009; Gonçalves *et al.*, 2018)](https://paperpile.com/c/6nYaGH/OAat+tAaX). No recombination was assumed; therefore, we applied a uniform recombination rate of *r*=0 across all sites. The Markov chain Monte Carlo step in IMPUTE2 is used to account for phase uncertainty in recombining diploid data [(Howie *et al.*, 2009)](https://paperpile.com/c/6nYaGH/OAat) but we did not perform this step as our data is non-recombining and haploid.

The effect of varying the number of sequences in the reference alignment (khap) was estimated by setting khap to 100, 250, 500, 1,000, 2,500, 5,000, 10,000, 20,000, and 30,000. We tested the ability of our pipeline to impute rare variants by filtering the Reference Panel to minor allele frequencies (MAF) of 1%, 0.5% and 0.1%, resulting 409, 682 and 1874 mtSNVs, respectively (Supplementary Tables 3). Imputation accuracy was assessed using Matthew’s Correlation Coefficient (MCC) [(Matthews, 1975)](https://paperpile.com/c/6nYaGH/QoS2) for genotype concordance and Hi-MC [(Smieszek et al. 2018)](https://paperpile.com/c/6nYaGH/s4qF) for haplogroup assignment, with the WGS data used as the truth set. Linear mixed-model ANOVA was used to assess the meaningful difference in haplogroup assignment and MCC (mean of mtSNVs per ISM) for different parameters tested for khap and MAF. Pipelines for implementing our imputation protocol and reproducing our results were created in SnakeMake [(Köster and Rahmann, 2012)](https://paperpile.com/c/6nYaGH/jAgj).

## **Results**

### Reference Alignment and Reference Panel

The Reference Panel contained 7,128 (19.3%) sequences for which GenBank metadata on geographic provenance were available (Table 1) representing 49 countries. Within country provenance for was avaliable for xxx samples, taking the total number of identified regions of provenance to 103. This included smaller ethnic groups such as Yami Taiwanese, Morrocan Berbers, Pacific Islanders, Indigenous Australians, and people from Central Asia and Siberia. There is, however, still a distinct bias towards European (3,855; 54.1% of sequences with provenance) and East Asian (2,065; 29.0%) samples. All major haplogroups are represented in the Reference Panel (Table 1), including rare haplogroups such as haplogroup S which is endemic to Indigenous Australians, haplogroup L5 which is found in Mbuti Pygmies, haplogroup L6 which is found in low frequencies in Yemen and Ethiopia, and haplogroups O and Q which are found exclusively in the Pacific Islands. Haplogroup B was most filtered haplogroup after quality control on the Reference Panel, removing 3,395 sequences (or, 46% of removed sequences), leaving only 273 haplogroup B sequences. Haplogroup H was also heavily filtered following QC (1,376; 19%), however haplogroup H was still highly represented in the final reference panel (n=7,644). All other haplogroups had only a small fraction of their sequences removed during QC.

### ***In silico* Microarrays**

#### ***Parameter Tuning***

When compared to un-imputed data, haplogroup concordance improved by 42.7%, 44.6%, and 43.3% for MAF = 1%, 0.5%, and 0.1%, respectively (Supplementary Table 4; Supplementary Figure 1). Variation in this success rate was within the expected range (AVOVA, *p*=0.6). For genotype concordance, the best results were obtained for MAF = 1%; here the variation in success rate was significant (ANOVA, *p*<0.0001, Supplementary Table 5; Supplementary Figure 2). The number of reference haplotypes used had a noticeable effect on haplogroup and genotype concordance (ANOVA, *p*<0.0001, Supplementary Table 6; Supplementary Figure 3, 4). There was no significant difference between the top four khap parameter settings (khap = 100, 250, 500, 1000). Larger khap parameter settings performed comparatively poorly, displaying a reduced ability to correctly assign haplogroups for some ISMs.

#### ***Overall Microarray Performance***

Using our recommended settings (khap = 500, MAF = 1%), the average haplogroup assignment accuracy was 89.3% (95% Confidence Interval [CI] = 87.4, 91.2) following imputation, an increase of 42.7% (95% CI = 40.1%, 45.23%) (Supplementary Table 7). The best-performing ISM (Illumina HumanHap 240S) correctly assigned 99.8% of haplogroups after haplotype imputation, a small improvement of 0.8%. The worst performing group of ISMs (HumanOmni1-Quads) correctly assigned 52.3% of haplogroups after imputation compared to 12.9% before imputation. Correct assignment for the worst performing individual ISM (HumanOmni 2.5) increased from 4.9% to 64.0% after imputation. The greatest improvement was 64.8% for the HumanCore ISMs. In terms of genotype concordance, the mean MCC = 0.64 (95% CI = 0.60, 0.68, Supplementary Table 7), to MCC = 0.97 for the best performing ISM (Infinium Global Screening Array-24v2) and to MCC = 0.10 for the worst performing ISM (HumanOmni 2.5).

#### ***Overall Haplogroup Concordance***

Concordance of individual haplogroups was estimated at the macro-haplogroup level. Prior to imputation, less than 49% of sequences from haplogroups M, HV, D, L, A, H, J, W, I, V were assigned to their correct haplogroup (Supplementary Table 8). Imputation improved haplogroup assignment by between 30% and 83%. Microarray assignment was relatively good (>74%) for haplogroups R, B, U, N, C, T, K, so improvement from imputation was, correspondingly, minor to moderate (0.1%-18%). Haplogroups JT and X showed no improvement.

### **Alzheimer’s Disease Neuroimaging Initiative**

To illustrate the utility of MitoImpute, we tested our pipeline on 258 participants from the ADNI dataset who had both WGS [(Ridge *et al.*, 2018)](https://paperpile.com/c/6nYaGH/WWBy) and genotyping data [(Saykin *et al.*, 2010)](https://paperpile.com/c/6nYaGH/EPF5) (Supplementary Table 9). The ADNI genotype data were mapped to the rCRS. Hi-MC [(Smieszek *et al.*, 2018)](https://paperpile.com/c/6nYaGH/s4qF) was used to assign haplogroups to the WGS, genotyped, and imputed data. Genotype data assigned the correct haplogroup to 31.4% of samples, which improved to 91.9% (Supplementary Table 10) after imputation. The corresponding improvement for macro-haplogroups was 37.2% to 95%. Eight of nineteen macro-haplogroups showed no improvement as the genotype data provided perfect or near-perfect haplogroup assignment. Haplogroups J, L2, M, V, W, X all improved from 0% to 100% correct assignment. Haplogroup H was the most frequently observed and showed an improvement of 5.8% to 100%. Haplogroups N & R were the worst performing post-imputation at 25% and 36.4%, respectively (Supplementary Table 11). Following imputation, the mean genotype concordance per mtSNV was MCC = 0.71 (95% CI = 0.66, 0.75).

## **Discussion**

Utilising complete mitochondrial genomes is key to understanding the genetic basis of mitochondrial disease and explaining patterns of human diversity. However, due to cost and technological limitations, older datasets and those generated from genotyping microarrays often only contain a subset of mtDNA variants, thereby precluding any missing variants from downstream analysis. We address this issue by developing MitoImpute, a pipeline for easy imputation of missing mtDNA variants. We determined imputation accuracy by using concordance of assigned haplogroups and Matthew’s correlation coefficient of genotypes. Our results show using the MitoImpute pipeline improves haplogroup assignment compared to haplogroups assigned from just microarray variants.

The MitoImpute pipeline improves haplogroup assignment in many commonly used microarrays, as demonstrated in the IMS analysis. By applying MitoImpute to the ADNI dataset, we further demonstrated that MitoImpute can be utilised by long-term studies whose older datasets have limited mtSNV genotypes, thus making them comparable with newer resequenced datasets. MitoImpute provides an opportunity for datasets with limited mitochondrial genetic variation to be analyzed with a more complete set of genetic variants and a more accurate assignment of haplogroups. The global disparity in medical research is evident in the high proportion of European individuals (~78%) association study catalogues [(Sirugo et al. 2019)](https://paperpile.com/c/6nYaGH/sOjx). The 1,000 Genomes Project phase 3 includes 2,504 individuals from 26 populations, however these individuals were often sampled from 1-3 cities within geographically diverse countries, such as China. Our MitoImpute Reference Panel contains sequences from at least 103 regions in at least 49 countries, capturing a more globally-representative sample of mitochondrial genetic diversity. The diversity included in the MitoImpute Reference Panel will allow researchers to perform imputation in under-represented human populations, contributing to solving the disparity in medical genomic research.

Performance testing of the MitoImpute pipeline revealed some counterintuitive results. One would expect that including rarer haplotypes into the Reference Panel by decreasing the MAF threshold would lead to increases in imputation accuracy. However, the best performing of the Reference Panel parameter settings was the highest MAF threshold (MAF ≥ 1%; Supplemental Figure 2). We suspect the decrease in imputation accuracy is due ISMs with few mtSNVs being unable to ‘decide’ which reference haplotype to impute from, in some cases making an erroneous decision. However, we have not investigated this further, and therefore recommend the parameter setting MAF ≥ 1%. Another seemingly counterintuitive result is the decrease in imputation accuracy as the khap parameter increases. Increasing the khap parameter increases the number of haplotypes in the Reference Panel from which IMPUTE2 will impute. We suspect that increasing the number of reference haplotypes beyond 1,000 leads to a greater chance of mismatch between the incomplete sample haplotypes and the Reference Panel haplotypes, particularly in ISMs with few mtSNVs. The limitations of the MAF and khap parameters, we suspect, is due to a dearth of mtSNVs in some ISMs. Therefore, we recommend that users take care with imputing variants in sample panels with few mtSNVs.

Manual curation of alignments requires time and expertise, which may not be available to all researchers. As a publicly available resource, the Reference Alignment can be used by future mitochondrial studies to ensure consistent alignment of homologous positions and placement of gap character states. The Reference Panel can also lend itself to utilities outside of it original purpose of MitoImpute. As a well-curated alignment, the Reference Panel can be subsampled, allowing for non-experts to use these sequences for medical studies, phylogenetics and population genetics, and other studies.

To facilitate the reanalysis of older mitochondrial datasets with missing variants and mtDNA research in general, we have: 1) created a reference alignment to aid with consistent placement of gap character states during the alignment process; 2) created a Reference Panel containing globally diverse human mtDNA sequences that can be used for imputation or subsampled as a ready-to-use dataset; 3) created a pipeline that allows researchers without proficient bioinformatic skills to perform imputation of mtSNVs. As the MitoImpute Reference Panel is updated to include more sequences as they become available on GenBank, the power to impute variants in diverse datasets will only increase.

# **Acknowledgments**

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

# **Funding**

Dr. Judy Pa and Christopher Patterson were supported by the National Institute on Aging (R01AG054617 PI: Judy Pa). SJA, BFH and AMG are supported by the JPB Foundation (<http://www.jpbfoundation.org>). RHS is supported by P30 AG035982.

# **Conflicts of interest**

AMG served on the scientific advisory board for Denali Therapeutics from 2015-2018. She has also served as a consultant for Biogen, AbbVie, Pfizer, GSK, Eisai and Illumina.

## **References**

[1000 Genomes Project Consortium *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.](http://paperpile.com/b/6nYaGH/B4Lf)

[Andrews,R.M. *et al.* (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.*, **23**, 147.](http://paperpile.com/b/6nYaGH/kw3a)

[Gonçalves,V.F. *et al.* (2018) Examining the role of common and rare mitochondrial variants in schizophrenia. *PLoS One*, **13**, e0191153.](http://paperpile.com/b/6nYaGH/tAaX)

[Gorman,G.S. *et al.* (2016) Mitochondrial diseases. *Nat. Rev. Dis. Primers*, **2**, 16080.](http://paperpile.com/b/6nYaGH/uCdp)

[Howie,B.N. *et al.* (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies.](http://paperpile.com/b/6nYaGH/OAat)

[Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.](http://paperpile.com/b/6nYaGH/xYfv)

[Kearse,M. *et al.* (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.](http://paperpile.com/b/6nYaGH/gPUL)

[Köster,J. and Rahmann,S. (2012) Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.](http://paperpile.com/b/6nYaGH/jAgj)

[Lott,M.T. *et al.* (2013) mtDNA Variation and Analysis Using Mitomap and Mitomaster. *Curr. Protoc. Bioinformatics*, **44**, 1.23.1–26.](http://paperpile.com/b/6nYaGH/rRzH)

[Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.](http://paperpile.com/b/6nYaGH/QoS2)

[van Oven,M. and Kayser,M. (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.*, **30**, E386–E394.](http://paperpile.com/b/6nYaGH/bEt5)

[Popejoy,A.B. and Fullerton,S.M. (2016) Genomics is failing on diversity. *Nature*, **538**, 161–164.](http://paperpile.com/b/6nYaGH/uXSH)

[Ridge,P.G. *et al.* (2018) Assembly of 809 whole mitochondrial genomes with clinical, imaging, and fluid biomarker phenotyping. *Alzheimers. Dement.*, **14**, 514–519.](http://paperpile.com/b/6nYaGH/WWBy)

[Saykin,A.J. *et al.* (2010) Alzheimer’s Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans. *Alzheimers. Dement.*, **6**, 265–273.](http://paperpile.com/b/6nYaGH/EPF5)

[Sirugo,G. *et al.* (2019) The Missing Diversity in Human Genetic Studies. *Cell*, **177**, 26–31.](http://paperpile.com/b/6nYaGH/sOjx)

[Smieszek,S. *et al.* (2018) Hi-MC: a novel method for high-throughput mitochondrial haplogroup classification.](http://paperpile.com/b/6nYaGH/s4qF)

## **Supplementary Information**

**Supplementary Methods**

The following search term was used to identify whole human mtDNA sequences from GenBank on 2018-07-18:

(016500[SLEN]:016600[SLEN]) AND Homo[Organism] AND mitochondrion[FILT] AND complete genome NOT (Homo sp. Altai OR Denisova hominin OR neanderthalensis OR heidelbergensis OR consensus OR ancient human remains OR shotgun)

## **Supplementary Tables**

Please find all tables at the following [link](https://docs.google.com/spreadsheets/d/1Km7TI6vZeAgH10Gye6SwFALLsd_WQB2UGWGbMFDA4Bk/edit?usp=sharing):