MitoImpute: A Snakemake pipeline for imputation of mitochondrial genetic variants

Tim W McInerney[1], Brian Fulton-Howard[2], Christopher Patterson[3,4], Devashi Paliwal[1], Lars S Jermiin[5,6,7,8], Hardip R Patel[1], Judy Pa[3,4], Russell H Swerdlow[9], Alison Goate[2], Simon Easteal[1], Shea J Andrews[2*], for the Alzheimer's Disease Neuroimaging Initiative[†]

[1]The John Curtin School of Medical Research, The Australian National University, Canberra, Australian Capital Territory, Australia
[2]Ronald M. Loeb Center for Alzheimer's disease, Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, New York, USA
[3]Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
[4]Department of Neurology, Alzheimer's Disease Research Center, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
[5]CSIRO Land & Water, Commonwealth Scientific Industrial & Research Organization, Acton, ACT 2601, Australia
[6]Research School of Biology, Australian National University, Canberra, ACT 2601, Australia
[7]School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland
[8]Earth Institute, University College Dublin, Belfield, Dublin 4, Ireland
[9]Department 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

[†]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) to accurately assign sequences to their correct haplogroup. To address this, we created an easy to use mitochondrial DNA imputation pipeline, MitoImpute.

Results: We validated imputation accuracy by measuring haplogroup and genotype concordance in two datasets, 1000 Genomes Project and the Alzheimer's Disease Neuroimaging Initiative. In both datasets, we observed a significant improvement in haplogroup concordance and excellent genotype concordance measures. We 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.

Availability

https://github.com/sjfandrews/MitoImpute

Introduction

Mitochondrial DNA (mtDNA) variation is informative about human evolution and can be associated with disease (Gorman *et al.*, 2016). Variation in these data 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.

We present MitoImpute, a pipeline to infer mtDNA haplotypes 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) whole genome sequence (WGS) data, and real-world data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Saykin *et al.*, 2010).

Methods

Reference Alignment

We used publicly available PhyloTree (van Oven and Kayser, 2009) sequences to create a large (*n*=7,747) reference alignment with the revised Cambridge Reference Sequence (rCRS) (Andrews *et al.*, 1999) site numbering convention. We aligned sequences in batches of 50 using the L-INS-i version of MAFFT (Katoh and Standley, 2013), 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) 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) in Geneious v10.2.6 (Kearse *et al.*, 2012). 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).

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).

Imputation

We used the IMPUTE2 chromosome X protocol for imputation (Howie *et al.*, 2009; Gonçalves *et al.*, 2018). 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) 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 (k_{hap}) was estimated by setting k_{hap} 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 as haplogroup concordance and genotype concordance using Matthew's Correlation Coefficient (MCC) (Matthews, 1975), 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 k_{hap} and MAF. Pipelines for implementing our imputation protocol and reproducing our results were created in SnakeMake (Köster and Rahmann, 2012).

Results

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 k_{hap} parameter settings (k_{hap} = 100, 250, 500, 1000). Larger k_{hap} parameter settings performed comparatively poorly, displaying a reduced ability to correctly assign haplogroups for some ISMs.

Overall Microarray Performance

Using our recommended settings (k_{hap} = 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) and genotyping data (Saykin *et al.*, 2010) (Supplementary Table 9). The ADNI genotype data were mapped to the rCRS. Hi-MC (Smieszek *et al.*, 2018) 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

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. Additionally, by incorporating a globally-diverse mitochondrial sequence Reference Panel, we demonstrate MitoImpute's utility in non-European populations. 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.

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.

Andrews,R.M. *et al.* (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.*, **23**, 147.

Gonçalves,V.F. *et al.* (2018) Examining the role of common and rare mitochondrial variants in schizophrenia. *PLoS One*, **13**, e0191153.

Gorman,G.S. *et al.* (2016) Mitochondrial diseases. *Nat. Rev. Dis. Primers*, **2**, 16080.

Howie,B.N. *et al.* (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies.

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.

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.

Köster,J. and Rahmann,S. (2012) Snakemake--a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.

Lott,M.T. *et al.* (2013) mtDNA Variation and Analysis Using Mitomap and Mitomaster. *Curr. Protoc. Bioinformatics*, **44**, 1.23.1–26.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

van Oven,M. and Kayser,M. (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.*, **30**, E386–E394.

Ridge,P.G. *et al.* (2018) Assembly of 809 whole mitochondrial genomes with clinical, imaging, and fluid biomarker phenotyping. *Alzheimers. Dement.*, **14**, 514–519.

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.

Smieszek,S. *et al.* (2018) Hi-MC: a novel method for high-throughput mitochondrial haplogroup classification.