Short Report

Assembly of 809 whole mitochondrial genomes with clinical, imaging, and fluid biomarker phenotyping

Perry G. Ridge\textsuperscript{a}, Mark E. Wadsworth\textsuperscript{a}, Justin B. Miller\textsuperscript{a}, Andrew J. Saykin\textsuperscript{b}, Robert C. Green\textsuperscript{c}, the Alzheimer’s Disease Neuroimaging Initiative\textsuperscript{1}, John S. K. Kauwe\textsuperscript{a,d,*}

\textsuperscript{a}Department of Biology, Brigham Young University, Provo, UT, USA
\textsuperscript{b}Radiology and Imaging Sciences, Medical and Molecular Genetics and the Indiana Alzheimer’s Disease Center, Indiana University School of Medicine, Indianapolis, IN, USA
\textsuperscript{c}Division of Genetics, Department of Medicine, Brigham and Women’s Hospital, Partners HealthCare Personalized Medicine, The Broad Institute and Harvard Medical School, Boston, MA, USA
\textsuperscript{d}Department of Neuroscience, Brigham Young University, Provo, UT, USA

Abstract

Introduction: Mitochondrial genetics are an important but largely neglected area of research in Alzheimer’s disease. A major impediment is the lack of data sets.

Methods: We used an innovative, rigorous approach, combining several existing tools with our own, to accurately assemble and call variants in 809 whole mitochondrial genomes.

Results: To help address this impediment, we prepared a data set that consists of 809 complete and annotated mitochondrial genomes with samples from the Alzheimer’s Disease Neuroimaging Initiative. These whole mitochondrial genomes include rich phenotyping, such as clinical, fluid biomarker, and imaging data, all of which is available through the Alzheimer’s Disease Neuroimaging Initiative website. Genomes are cleaned, annotated, and prepared for analysis.

Discussion: These data provide an important resource for investigating the impact of mitochondrial genetic variation on risk for Alzheimer’s disease and other phenotypes that have been measured in the Alzheimer’s Disease Neuroimaging Initiative samples.

© 2017 The Authors. Published by Elsevier Inc. on behalf of the Alzheimer’s Association. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Alzheimer’s disease; ADNI; Mitochondrial genetics; Whole mitochondrial genomes; Next-generation sequencing

1. Introduction

Alzheimer’s disease (AD), the most common form of dementia, affects >20 million people worldwide and is the only one of the top 10 causes of death that has no effective treatments [1–3]. Full-time care is required as AD progresses, further impacting patients and their loved ones and stressing the health-care system. With incidence expected to increase to 1 in 85 people by 2050 [2], it is essential to achieve early diagnosis, effective treatments, and a better understanding of the underlying etiology.

Understanding the underlying mechanisms of risk for AD is a key for both diagnosis and treatment. Swerdlow et al. [4] proposed the Mitochondrial Cascade Hypothesis of AD. Briefly, an individual’s genetics determine the baseline mitochondrial function and how mitochondria change as a person ages and declining mitochondrial function causes AD-specific pathologies.

In addition to the evidence provided by Swerdlow et al. [4], several lines of evidence support a role for mitochondria
in AD. First, mitochondria fundamentally change in a number of ways in AD and contribute to its progression and onset [5]: metabolism decreases [6], mitochondrial fusion/fission are disrupted [7], mitochondrial concentration (i.e. the ratio of mitochondrial genomes to nuclear genomes) decreases in cerebrospinal fluid [8,9], mitochondrial morphology changes [4,10], mitochondrial-encoded enzymes in the electron transport chain are altered [5,11], amyloid plaques aggregate in mitochondria [12,13], and many of these changes take place near plaques [14].

Second, individuals with a maternal family history of AD have as high as 9 times the risk of AD compared with individuals with a paternal family history of AD [15,16], or no family history. Furthermore, individuals with a maternal family history of AD also score lower on cognitive tests [17], have a lower age of onset of AD [15,18], and have more pronounced brain abnormalities consistent with AD (e.g. cerebral metabolic changes [19], higher amyloid β burden [20], reduction in gray matter volume [21,22], and increased global Pittsburgh Compound-B uptake Pittsburgh Compound-B-positron emission tomography [23]). Moreover, we found that some of these brain abnormalities are associated with mitochondrial haplotypes [24].

This mitochondrial impact on AD risk could be influenced by several factors, including differential responses to the oxidative stress, variation in nuclear-encoded mitochondrial genes, and variation in the mitochondrial genome. In this article, we focus on an important resource for investigating mitochondrial genomic variation and others [25]. Several groups have reported a relationship between mitochondrial genetics and risk for AD (summarized in Ridge et al. [3], Table 2). Twelve different haplotypes have been implicated in mitochondrial genetic studies, but the majority of these were reported only once and not replicated [26–33], and six different studies reported no association between mitochondrial genetic variants and AD [34–39]. Among reported associations, there is no consensus, and sometimes, results appear to be contradictory. For example, Haplogroup U has been reported as both a risk and protective haplogroup [28,31,32]. However, potentially explaining the confounding nature of discoveries to date, the majority of studies used incomplete sequence data and/or had very small sample sizes [26–39], thus most were underpowered and lacked the resolution to identify correlations for all but the most common haplogroups. Only a single study used whole mitochondrial data [30], whereas most genotyped only a handful of single nucleotide polymorphisms (SNPs). Furthermore, only one study used a large data set, but in this particular data set, the authors only genotyped 138 SNPs [39]. In summary, there is strong evidence to suggest a relationship between the mitochondrial genome and AD, yet the relationship remains undefined.

The Alzheimer’s Disease Neuroimaging Initiative (ADNI) recently sequenced the whole genomes, including mitochondrial genomes, of 809 individuals. Each of the genomes was analyzed using tools and pipelines developed for diploid genomes. However, these analysis pipelines, particularly variant identification that relies on a likelihood model expecting diploid sequences, are inaccurate for use on the mitochondrial genome, which is haploid. Here, we report not only an AD data set of 809 annotated whole mitochondrial genomes with extensive phenotypes (Table 1) but also an appropriate pipeline to analyze mitochondrial genomes. We hope to facilitate research in this important area by providing a data set and analysis pipeline for future researchers to augment this initial data set.

2. Methods

2.1. Alzheimer’s Disease Neuroimaging Initiative

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership and is an ongoing, longitudinal, highly collaborative study. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and used for the early diagnosis of AD. ADNI has undergone several phases (ADNI1, ADNI GO, and ADNI 2), with each phase adding additional samples. In 2012, 818 ADNI samples were selected for whole genome sequencing to further the goals of ADNI. DNA sequence data were collected from DNA derived from whole blood. All subjects in our analyses had self-reported ancestry of non-Hispanic European American. All the data (whole genome sequence, phenotype, and newly assembled and annotated whole mitochondrial genomes) are publically available through ADNI (http://adni.loni.usc.edu/data-samples/).

2.2. Genome sequencing, assembly, and variant detection

ADNI genomes were sequenced on an Illumina HiSeq. Reads were paired-end, 100 base-pair reads. Before read mapping, adapters were removed. ADNI mapped the whole genome sequences and called variants using default settings.
in the Burrows-Wheeler Aligner [40] for mapping and standard best practices from the Genome Analysis Toolkit [41,42] for variant detection (for details see http://adni.loni.usc.edu/data-samples/genetic-data/wgs/). However, these steps needed to be redone for two reasons: First, original mappings were to Hg19. Historically, determining a reference mitochondrial genome has been a bit confusing. The standard mitochondrial genome, NC_012920, was released first in 1981 (the Cambridge Reference Sequence, or CRS) [43] and corrected in 1999 (the new Cambridge Reference Sequence, or rCRS) [44]. NC_012920 is a European haplogroup and thus is a leaf on the mitochondrial haplotype phylogenetic tree. A number of other mitochondrial genomes have been suggested as the “correct” mitochondrial genome reference, including a reconstructed hypothetical mitochondrial genome corresponding to Mitochondrial Eve [45]. Nevertheless, NC_012920 remains the standard version of the mitochondrial genome used in mitochondrial genetics studies. Hg19, which was used for read mapping by ADNI, has a version of the mitochondrial genome, represented as chrM, but sometimes corresponds to NC_001807 and sometimes NC_012920. Therefore, to be consistent with mitochondrial genetic standards, reads needed to be remapped. Second, the standard Genome Analysis Toolkit pipeline includes many steps. One of these steps is the HaplotypeCaller. The HaplotypeCaller builds a likelihood model based on possible reconstructed haplotypes in a genomic region, but it assumes sequences are diploid. Consequently, mitochondrial genomic variants identified using the HaplotypeCaller are likely to include many inaccuracies.

Since chrM and NC_012920 only differ by a few bases, we were able to extract only those reads that mapped to chrM (with SAMTools [46]), rather than all reads corresponding to the whole nuclear and mitochondrial genomes. Extracted reads were remapped to NC_012920 using Burrows-Wheeler Aligner. Next, we performed local realignments around indels and base recalibration, which are not affected by ploidy, with Genome Analysis Toolkit to refine the new mappings. Finally, we used FreeBayes (-p 1 –F 0.6, and by ploidy, with Genome Analysis Toolkit to refine the new mappings around indels and base recalibration, which are not affected by ploidy, with Genome Analysis Toolkit to refine the new mappings) to joint-call variants and converted the resulting variant call format (VCF) file to fasta with vcf2fasta (vcflib, https://github.com/vcflib/vcflib). An overview of the whole process is outlined in Supplementary Figure 1.

2.3. Genome annotation

We annotated mitochondrial variants and haplotypes for each sample. We downloaded 9228 mitochondrial DNA coding and RNA sequence variants and 2792 control region variants from MITOMAP [48]. For each variant present in the data sets downloaded from MITOMAP, we added complete information (i.e. frequency, source, locus names, etc.) to the “Info” column in the VCF file and added the corresponding annotation information to the header lines. For each variant that was reported by multiple studies in MITOMAP, we included all studies in the annotation.

Next, we annotated mitochondrial haplotypes with Phy-Mer [49]. Phy-Mer reports the five most likely mitochondrial haplotypes and a score, where 1 is a perfect score. For each of the samples, we selected the top hit. All samples had scores >0.99 except for one that had a score of 0.988.

2.4. Variant validation

SNP data from the Illumina 2.5 M Array were collected from the same DNA extraction that was used for whole genome sequence (WGS) data collected. We compared 256 mitochondrial variants genotyped on that array to the variant calls from our WGS analysis pipeline. Validation was evaluated by simply looking at the concordance of calls on the individual level for the two sets of genotype data.

3. Results

Mitochondrial genomes were sequenced to an average depth of 2986 reads and ranged from an average depth of 1515 reads to 7831 reads. We identified 1649 total mitochondrial genetic variants from these genomes, of which 1336 have been previously reported. Samples had an average of about 27 variants, with a range of 1 to 96 variants. We validated our variant calls using 256 mitochondrial genetic genotypes from an SNP array performed on the same samples. Overall, 98.18% of WGS variant calls matched SNP genotypes acquired from the array.

We identified 506 different mitochondrial haplotypes in the data set, all of which have been previously observed. The majority, 350, only appear in a single sample in the data set. The most common haplotype, K1A1B1A, is shared by 15 individuals. When considering only the major mitochondrial haplogroup (e.g. H, V, U, etc.), the majority of individuals had haplotypes in the H and U groups, 338 and 105, respectively. We report number of individuals in each major mitochondrial haplogroup in Table 2 and counts of all haplotypes in Supplementary Table 1. These frequencies are similar to those observed in other cohorts of non-Hispanic European American ancestry [50]. Several individuals have haplotypes reported to be associated with risk for AD: five have haplotype H5 (a risk haplogroup [29]).

4. Discussion

We have presented the application of a novel approach to accurately assemble and genotype a data set of whole mitochondrial genomes from the ADNI study. Several accurate algorithms exist for calling genotypes in diploid NGS data, but use models that are inappropriate for haploid samples. We used existing software, but tested and refined parameter settings to achieve high levels of genotype accuracy in our mitochondrial genome samples. We validated the identified genotypes by repurposing existing publically available data.
Table 2
Counts and frequencies of major mitochondrial haplogroups

<table>
<thead>
<tr>
<th>Major mitochondrial haplogroup</th>
<th>Count</th>
<th>Frequency (%)</th>
<th># Cases/controls/MCI/unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>0.62</td>
<td>3/1/1/0</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
<td>0.87</td>
<td>3/3/3/0</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>0.37</td>
<td>0/0/3/0</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>0.12</td>
<td>1/0/0/0</td>
</tr>
<tr>
<td>H</td>
<td>338</td>
<td>41.78</td>
<td>73/128/133/4</td>
</tr>
<tr>
<td>I</td>
<td>27</td>
<td>3.34</td>
<td>4/12/11/0</td>
</tr>
<tr>
<td>J</td>
<td>69</td>
<td>8.53</td>
<td>24/18/26/1</td>
</tr>
<tr>
<td>K</td>
<td>68</td>
<td>8.41</td>
<td>19/24/25/0</td>
</tr>
<tr>
<td>L</td>
<td>33</td>
<td>4.08</td>
<td>5/16/12/0</td>
</tr>
<tr>
<td>M</td>
<td>11</td>
<td>1.36</td>
<td>2/4/5/0</td>
</tr>
<tr>
<td>N</td>
<td>7</td>
<td>0.87</td>
<td>2/2/3/0</td>
</tr>
<tr>
<td>R</td>
<td>5</td>
<td>0.62</td>
<td>1/3/1/0</td>
</tr>
<tr>
<td>T</td>
<td>78</td>
<td>9.64</td>
<td>16/25/37/0</td>
</tr>
<tr>
<td>U</td>
<td>105</td>
<td>12.98</td>
<td>27/31/46/1</td>
</tr>
<tr>
<td>V</td>
<td>28</td>
<td>3.46</td>
<td>4/10/14/0</td>
</tr>
<tr>
<td>W</td>
<td>15</td>
<td>1.85</td>
<td>3/2/10/0</td>
</tr>
<tr>
<td>X</td>
<td>9</td>
<td>1.11</td>
<td>4/2/3/0</td>
</tr>
</tbody>
</table>

MCI, mild cognitive impairment.

Numbers refer to individuals who have the listed major mitochondrial haplogroup or a subgroup (e.g. individuals with H5 are counted as part of the H group in this table).

Our genotypes are >98% concordant with genotypes from Illumina SNP chips in the same samples. This high level of accuracy is approaching the expected error rate of SNP chips and makes it difficult to definitively determine whether our genotypes, or the SNP chip genotypes, are correct.

In addition, we have made these data publically available through the ADNI website. A multisample VCF file with mitochondrial genomic variants for each subject is available for download from the Download—Genetic Data section of the ADNI repository (http://adni.loni.usc.edu/data-samples/access-data/). Also included is an .XLSX file with mitochondrial haplotype information for each subject. The new data sets are named “ADNI WGS Whole mitochondrial genome variants” and “ADNI WGS Whole mitochondrial genome—Haplotypes”. This data set is now ready to be applied in AD studies and to help elucidate the relationship between mitochondria and AD, which has thus far alluded researchers.

Complete mitochondrial genomic data result in high resolution of haplotype definition, including large numbers of singleton haplotypes. Methods that group haplotypes in evolutionarily meaningful ways are necessary to fully leverage these data. In our previous analyses of complete mitochondrial genomics data, we incorporated TreeScanning to concentrate statistical power on evolutionarily meaningful groups of haplotypes [24,30,51,52]. This approach has been applied successfully in the study of both mitochondrial and nuclear genomic contributions to AD risk and related phenotypes [24,30,51,53,54].

The ADNI subjects include AD cases, subjects with mild cognitive impairment, and cognitively normal controls. Nearly all of these subjects are also associated with extensive clinical, imaging, and fluid biomarker data, including some longitudinal data. As such, they provide great value in evaluating the multitude of factors that are associated with AD, dementia, dementia progression, and conversion from mild cognitive impairment to dementia. While this study is underpowered to identify haplotypes associated with AD risk, we anticipate it will serve as the foundation for additional data collection and an expanded study in the future. These data will also prove valuable for validation and discovery analyses related to the multitude of other phenotypes that are available for these subjects. Association analyses using imaging and fluid biomarker data are outside the scope of this study, but we anticipate that these data will be leveraged by several groups, including ours, for these kinds of analyses.

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; Aracon 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 Genetech, 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. This work was supported by the following NIH grants: R01AG042611 (JSSK), RF1AG054052 (JSSK), U01AG024904 (AJS), P30AG010133 (AJS), R01AG019771 (AJS), R01LM011360 (AJS), U01HG006500 (RCG), U19HD077671 (RCG), R01HG02213 (RCG), R01AG047866 (RCG), and U01AG24904 (ADNI). Further, we acknowledge support from the Fulton Supercomputing Lab at Brigham Young University. We are also grateful to ADNI and its participants for their open approach to data sharing.
Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jalz.2017.11.013.

RESEARCH IN CONTEXT

1. Systematic review: We reviewed existing literature using PubMed and Google to identify data sets with whole mitochondrial genome sequences. Although several whole genome data sets have been generated for Alzheimer’s disease, none of these specifically addressed the whole mitochondrial genomes, which require specialized analytic approaches because they are haploid.

2. Interpretation: Mitochondria have a role in Alzheimer’s disease, but the precise role has thus far eluded researchers. A primary impediment is a lack of sufficient data. Our goals were to (1) develop an approach for analyzing whole mitochondrial genomes; and (2) develop a data set of whole mitochondrial genomes and make them publicly available to encourage research in this important area.

3. Future directions: We anticipate that making this data set and analytical approach will facilitate additional studies of the effect the mitochondrial genome has on risk for Alzheimer’s disease.

References


