Investigating metabolic dysfunction and arrhythmogenesis in an early-onset atrial fibrillation patient cohort

by

Brent Flodin

B.Sc. (Hons), Simon Fraser University, 2013

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Biomedical Physiology and Kinesiology
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2021

Declaration of Committee

Name: **Brent Flodin**

Degree: **Master of Science**

Thesis title: **Investigating metabolic dysfunction and arrhythmogenesis in an early-onset atrial fibrillation patient cohort**

Committee:

Chair: **Xiaowei Song**
Adjunct Professor, Biomedical Physiology and Kinesiology

Glen Tibbits
Supervisor
Professor, Biomedical Physiology and Kinesiology

Zachary Laksman
Committee Member
Clinical Assistant Professor, Medicine
University of British Columbia

Ryan Morin
Committee Member
Associate Professor, Molecular Biology and Biochemistry

Peter Ruben
Examiner
Professor, Biomedical Physiology and Kinesiology

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Updated Spring 2016

Abstract

Despite the prevalence of atrial fibrillation (AF) and the burden it places on health care systems, there remains much that is unknown regarding heritable factors influencing its development and progression. In this study, I investigated whole-exome sequencing (WES) data from a cohort of patients presenting with early-onset AF to explore the role that metabolic dysfunction might play in contributing to disease onset. I curated a metabolism-related gene panel and, following in silico prediction of variant pathogenicity, performed gene-level burden testing using reference data from the Genome Aggregation Database (gnomAD) and the human mitochondrial genome database MITOMAP. I further explored genes associating with AF in the UK Biobank data set, and discovered associations with several AF comorbidities including diabetes, hypertension, and stroke.

Keywords: atrial fibrillation; metabolism; whole-exome sequencing; diabetes; hypertension; stroke

This document is dedicated to the noble mitochondria.

Because, as we all know, the mitochondria are the powerhouse of the cell.

Acknowledgements

There are a large number of people that I would like to thank for their support and the instrumental ways in which they contributed to my pursuit and completion of this degree.

First, I would like to thank my supervisory committee:
Zachary Laksman, for his endless clinical expertise, encouragement, and for providing crucial structure to my work through frequent updates and feedback,
Ryan Morin, for his irreplaceable technical advice which helped me turn a superficial understanding of genomic methods into a functional analytical pipeline, and
Glen Tibbits, for his support, patience, and insidious style of intellectual encouragement which somehow managed to trick me into realizing I'm smarter than I thought I was.

Many people aided in my development from someone who had never even taken a genetics course at the beginning of this project to someone capable of understanding the complexities of bioinformatics studies and applying them in a productive way:
Fiona Brinkman, for allowing me to take her bioinformatics course with no experience with the subject matter and no guarantee I would even know what was going on,
Brandon Chalazan, for thoughtful conversation which inadvertently convinced me not to re-map to GRCh38, which made my life far easier than it would have otherwise been,
Alexandra Elbakyan, for removing barriers in the way of science,
Roselle Gélinas, for her invaluable metabolomics knowledge and her willingness to share that expertise when I was developing my gene panel,
Robert Hegele, Jason Roberts, and John Robinson, for their time in helping me to understand the initial sequencing and processing of the patient data that I used,
Kate Huang, Mark Trinder, and Liam Brunham, for graciously accessing the UKB data that I requested and manipulating the output into something that was far quicker and easier for me to analyze than it would have been without their extra effort,
Julieta Lazarte, for not only her willingness to answer any questions I had, using her knowledge of the same data set I used, but for allowing me to call her during her lunch break to do so, rather than just enjoying her meal and emailing me later,
Anna Lehman and Jill Mwenifumbo, for sharing their expertise and helping me to navigate these topics and data during the crucial period where I had learned a lot of things but didn't yet really understand how they all fit together as a cohesive whole,
Janet Liew and Jeremy Parker, without whom I would not only have no idea how to write

and submit an application for research ethics approval, but would also have no idea how to later locate that ethics approval to satisfy my thesis submission requirements,
Chris Rushton, for his extensive constructive feedback and advice on my initially-proposed analytical pipeline, and
Lauren Tindale, for her personal advice, direction to helpful resources, and friendly encouragement as I struggled to orient myself in a daunting field of study.

More broadly, my movement through my academic career has involved numerous supporters and advocates:
Scott Alexander, Leigh Bloomfield, Kim Lajoie, and Dan Marigold, for making my first exposure to research possible, without which I might not have even started this degree,
Craig Asmundson, for inviting me to apply for my first teaching position, which turned out to be a transformative experience and a catalyst for considerable personal growth,
Diana Bedoya, for keepin' it real in a way that few others can,
Jim Carter, Ryan Dill, Tony Leyland, and Van Truong, for their feedback and encouragement in expanding my skill, breadth, and opportunities in teaching,
Tom Claydon and Will Cupples, for their thoughtfulness and assistance in navigating the bureaucratic frustration of a degree plagued by medical and personal catastrophe,
Cierrah DiCesare, for deftly conveying compassionate understanding and hard perspective in equal measure, each precisely when I needed it,
Doc Hedges, for enthusiastic encouragement toward, and opportunities for, academic, intellectual, and personal growth beyond measure,
Christina Hull, for aggressive support and a mystifying and unending belief in what I could accomplish that so far eclipsed my own as to be almost comical, and
Marija Jovanovic, for accidentally facilitating profound and serendipitous personal discovery through the hieroglyphics of our modern age.

No thanks would be complete without mention of my endlessly supportive family: my dad, my sisters Hailey and Jenni, and my uncle John, without any of whom things would be very different for me. This is no less true for my small friends Rex and Marlie, who have carried me through more than it is possible for them to ever know.

Last, but certainly not least, I am eternally grateful to the Fellows of the Glum Birch Society, whose writing accountability parties were the primary reason that this thesis eventually saw completion at all.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1. Atrial fibrillation

Atrial fibrillation (AF) is a disease of the heart that is defined by the rapid, uncoordinated excitation of the atria and associated compromise in mechanical function. It is the most common clinically-diagnosed cardiac arrhythmia, with an estimated lifetime risk of 25%.[1] Its actual prevalence is likely even higher, as it is often transient or asymptomatic, and may therefore be under-detected.[2–5] Even subclinical AF can predispose to risk of morbidity and mortality, and it progresses to persistent AF in up to 50% of cases through self-propagating electrical and structural remodeling, an evolution often described as “AF begets AF.”[6, 7] When it does present with noticeable symptoms, it does so spanning a wide clinical range. Though it has been estimated that over 50% of patients have mild symptoms such as discomfort or impaired exercise capacity, AF is also related to a significant decline in quality of life.[5, 8, 9]

In addition to an association with myocardial infarction, heart failure, chronic kidney disease, and dementia, it has been estimated that AF is causative in one-third of strokes in patients over the age of 65.[6] In patients under age 65 with undiagnosed AF, stroke is the first related clinical symptom identified in 36% of cases.[10] In fact, sub-clinical AF confers a 2.5-fold increase in risk for ischemic stroke or systemic embolism, independent of other risk factors.[4] Over 30% of first-time strokes do not have an identified cause, and it is suspected that subclinical AF could be a factor in these cases.[11] Ultimately, AF confers almost a two-fold increased risk of mortality.[12]

Due to its high prevalence and potentially serious complications, AF represents an enormous burden on national health care systems.[13] Improvements in detection and treatment, as well as increasing incidence, are predicted to see the global AF population double by 2050.[14] Since the financial impact on health care systems of someone with AF is approximately five-fold that of someone without, AF is expected to require an even greater proportion of health expenditures as the average age of the population increases.[15]

AF has a complex etiology; while it is frequently related to advancing age and complicating cardiovascular conditions (e.g. hypertension), the current known clinical risk factors fail to explain AF in many cases.[16, 17] AF occurring in patients under 60 years of age and also in the absence of diabetes, hypertension, and structural cardiac irregularities has historically been known as “lone AF” and family studies have since shown that the strongest predictor of developing AF is having a direct relative with AF, far beyond the predictive capacity of traditional risk factors. It is now widely accepted that there is a multifaceted genetic contribution to AF development.[18] Dozens of genes appear to have involvement in AF, and the underlying physiology of its development remains opaque, particularly since known causative mutations display incomplete penetrance for reasons not yet understood.[19, 20] To complicate matters further, more recent evidence suggests that, in addition to rare variants with a strong physiological effect, there exists a large collection of common variants that may increase AF in independently small, but synergistic, ways.[21]

Owing to the complexity involved in regulating the cardiac action potential (AP), many of the genes associated with susceptibility to AF code for ion channel proteins, and pathogenic variants interfere with the carefully-orchestrated electrical propagation of the AP through the cardiac syncytium.[22] These variants tend to increase the general level of cellular excitability during a vulnerable period in the cardiac cycle through one of several mechanisms. Increased current density due to gain of function in a repolarizing ion channel, for example, can shorten the atrial AP, reducing cells’ effective refractory period (ERP) and increasing their susceptibility to delayed afterdepolarizations (DADs) and electrical re-entry loops.[23, 24] Conversely, loss of function in these repolarizing channels can result in a lengthening of the atrial AP, increasing the probability of early afterdepolarizations (EADs) and provide a vulnerable window of arrhythmogenic opportunity.[25] In agreement with these contrasting phenomena, it has been shown that patients are at increased AF risk with both lengthened and shortened corrected QT (QTc) interval.[26] Over time, persistent AF can drive changes in gene expression which contribute to further deterioration, such as decreased expression of L-type Ca^{2+} channels, ultimately resulting in shortened AP duration (APD) and thus ERP.[27]

Though the majority of identified causal variants are in ion channel genes, variants causing altered electrical dynamics have also been found in other genes, such as those coding for gap junction proteins, which may cause impaired conduction velocity

(CV).[28, 29] The consequences of slowed CV can be similar to those of shortened APD, whereby cells that would normally be refractory during the arrival of a potentially re-entrant depolarization are given extra time to recover and thus can inappropriately depolarize in response to the errant AP.

Tissue remodeling, such as fibrosis, may also act as a substrate for arrhythmia.[30] Since fibrotic tissue is electrically silent, it can not only slow conduction velocity, but also create regions of conduction block which interfere with normal waveform propagation. When a re-entrant wave encounters such a region, it can divide into numerous smaller wavelets which radiate outward in all directions, resulting in the asynchronous electrical disorganization characteristic of AF.[31]

AF is coincident with many conditions that are themselves associated with fibrosis of cardiac tissue, such as both hypertrophic and dilated cardiomyopathy, chronic atrial dilation, and aging.[32–34] Since fibrosis may result from AF, in addition to being a contributor, it facilitates a positive feedback cycle in which AF reinforces its own increasing severity over time.[35] Because of this progressive nature, early diagnosis and intervention is critical. Treatment falls within two categories: rate control and rhythm control. Rate control is aimed at mitigating the effect of the increased atrial rate on the ventricular rate so as to avoid ventricular tachyarrhythmia and associated pathological remodeling, and is accomplished using various anti-arrhythmic drugs (AADs). Rhythm control aims to restore sinus rhythm through AADs, cardioversion, or ablation therapy.[36]

Irrespective of management method, anticoagulation therapy is typically prescribed as a means of reducing risk of thromboembolism.[37] Additional treatment depends, to some degree, on disease advancement. Paroxysmal AF tends to have less-progressed fibrosis and ectopic waveforms of lower complexity, and responds more favorably to therapies aimed at preventing transmission of ectopic triggers from the pulmonary veins, which are critical sites of origin. Radiofrequency or cryoablation, which electrically isolate the ectopic regions, have been very successful in controlling paroxysmal AF.[38] Efficacy of AAD and ablation therapy both appear to be inversely related to AF progression, and once it has become persistent or permanent there is a decrease in relief following treatment.[39, 40] While rhythm control has traditionally been the favored approach, rhythm control drugs can have profound side effects, and some (e.g. sotalol) appear to actually increase risk of mortality.[41] Further, a pharmacological rhythm

control approach does not appear to offer any morbidity or mortality benefit when compared with a rate control strategy.[42]

The resistance to treatment, potentially serious complications, and growing prevalence make better understanding of the development and progression of AF a crucial area of research. While there are many genetic loci associated with AF, it is estimated that a large proportion of the heritable component of the disease remains unknown, emphasizing the need to investigate novel pathways and mechanisms of action.[43]

1.2. Metabolic arrhythmogenesis

A growing body of evidence suggests that a potential area of new insight is the link between metabolomics and arrhythmias. Though it includes numerous pathways that mainly function in parallel, many catabolic processes eventually converge to support synthesis of adenosine triphosphate (ATP), whether directly, through oxidative phosphorylation in the mitochondria or substrate level phosphorylation in the cytosol, or indirectly, through the production of substrate molecules to facilitate these processes.

It has long been understood that reperfusion following ischemia can lead to arrhythmogenic cellular decompensation and cell death, broadly implicating metabolic disturbance as a potential trigger for arrhythmia.[44] Both animal model and human studies show evidence of reduced atrial perfusion in AF, and rapid atrial pacing is accompanied by increased ATP synthase activity, suggesting an imbalance between energy production and demand.[45–47] An increase in the ratio of adenosine monophosphate (AMP) to ATP is indicative of increased energy consumption, and these changes are detected by the AMP-activated protein kinase (AMPK) which increases net cellular energy generation. AF patients display this increased activation relative to controls, further supporting the association between AF and changing energy dynamics.[48]

In a pertinent example of a metabolic gene variant causing electrical dysfunction, a variant in *PRKAG2*, which encodes the γ-subunit of AMPK, was observed in a family with Wolff-Parkinson-White (WPW) syndrome.[49] This same variant was also observed in an AF patient, and functional studies confirmed the variant was responsible for the aberrant electrical activity.[50] Intuitively, the best-understood of these variants tend to be

those which are involved earlier in their respective energy pathways, where there is a clearer relationship between individual proteins and their direct functional roles, prior to convergence at the electron transport chain. Numerous reported variants in *ACADVL*, for instance, result in very long chain acyl CoA dehydrogenase deficiency (VLCAD) which compromises β-oxidation of very long chain fatty acids and can result in fatal cardiomyopathy and arrhythmia.[51]

The literature describes many cases, however, of variants in genes further downstream that are related to similar cardiovascular complications, but for which direct mechanisms are far more elusive. Variants in mitochondrial complex I (NADH-ubiquinone oxidoreductase) are frequently associated with mitochondrial complex I deficiency, but because so many gene products are involved in the assembly and ongoing operation of the holoenzyme, and because many of these do not have a clearly understood function, genotype-phenotype relationships are difficult to infer. Despite this, associated variants have been observed in many of the subunits, and there is often considerable similarity in symptoms.[52] Leigh syndrome, and similar clinical presentations, are commonly associated with complex I variants, as are various cardiomyopathies. A similar trend exists with variants in complex II (succinate dehydrogenase) complex III (ubiquinol-cytochrome c reductase) and complex IV (cytochrome c oxidase) as well. As with deficiency in complex I, Leigh syndrome is a common clinical presentation associated with these variants, but dysfunction in these other complexes also exhibit the same puzzling range of symptoms, in terms of both severity and kind, which also include various metabolic symptoms such as lactic acidosis and disrupted blood glucose regulation, in addition to cardiomyopathies and arrhythmias.[53–55] Mechanisms for these relationships have rarely been proposed, since the specific functions of many of the mitochondrial complex subunits are still unknown. However, given the degree of similarity between the different conditions, it may be the case that the crucial process in disease onset is not specific to any of the mitochondrial complexes, and that disruption of ATP synthesis and overall energy balance as a whole is the primary contributor to these diseases.

As the site of the majority of ATP synthesis in cardiomyocytes, mitochondria present an obvious target for the investigation of such a relationship, particularly since rapid electrical pacing induces a pathological mitochondrial phenotype.[56, 57] Numerous studies have established that mitochondria are large contributors to the generation of

reactive oxygen species (ROS) within cells, predominantly complex I and complex III.[58] The repercussions of increased activity of these ROS-generating pathways can be seen following acute physiological insult. Transient ischemia and subsequent reperfusion, for example, leads to a variety of mitochondrial decompensations, such as inappropriate Ca^{2+} flux, and overproduction of ROS with simultaneous depolarization of the mitochondrial inner membrane.[59]

These same homeostatic disruptions can be triggered experimentally through photo-oxidation of a single mitochondrion, and can then lead to further increases in ROS production, a phenomenon known as ROS-induced ROS release (RIRR).[60,61] This process ultimately triggers a self-propagating cycle of mitochondrial damage which can then spread to other nearby mitochondria.[62] It has been further demonstrated that this dysfunction can progress to the point of generating a synchronized, cell-wide oscillation in mitochondrial inner membrane potential ($\Delta\Psi m$) and mitochondrial oxidation state. Cardiomyocyte APD also oscillates in phase with $\Delta\Psi m$ and this has been shown to directly cause arrhythmias.[60, 63] Coincidentally, upset Ca^{2+} homeostasis with subsequent disruption of $\Delta\Psi m$ has also been observed in VLCAD, a common consequence of variants in *ACADVL*.[64] While the mechanisms ultimately responsible are not yet conclusively determined, computational modeling has suggested that one possibility is through ROS activation of CaMKII.[65] More broadly, oxidative stress, like fibrosis, also generally increases with age, and increased ROS levels have been shown to be a factor in both structural and electrical remodeling, perhaps contributing to the well-established association between advancing age and AF.[66–68]

ROS overproduction is not exclusively the result of experimental intervention or severe physiological insult, however, and is also the consequence of pathological variants which cause mitochondrial enzyme dysfunction.[69, 70] Exploring these relationships using conventional methods is challenging, however, since the transfer of energy from primary substrate molecules to ATP involves the products of multiple hundreds of genes. As such, this is an area that stands to greatly benefit from the analytical power of recent genomics techniques. These techniques have already had success in identifying rare variants that influence other cardiovascular factors, such as blood pressure and plasma cholesterol, and it may be possible to utilize them for additional complex conditions which represent the manifestation of numerous, subtle contributors, such as AF.[71, 72]

1.3. Gene panel curation

I assembled a panel of candidate genes that code for proteins with evidence of involvement in cellular metabolic function, based on extensive literature review. I emphasized genes with more narrowly-defined functions so that a mechanism of any association with AF would be more easily intuited. For example, enzymes and subunits of multi-protein catalytic complexes are abundant since the downstream effect of a dysfunctional enzyme can be conjectured with some degree of confidence. In contrast, high-level transcription factors, which are likely to have a broad and less transparent influence through the complex interaction of multiple downstream cellular pathways, are less well-represented. The candidate gene list was assembled with the invaluable feedback of Dr. Roselle Gélinas and is viewable in Appendix A, with associated metabolic pathways and molecular functions predominantly annotated via the Human Protein Atlas.[73]

1.4. Genomics

Genetic data can be analyzed through a large and growing number of approaches which are continuously becoming more sophisticated. Still, the capacity to generate sequencing data far surpasses the capacity to actually analyze those data. Whole-exome sequencing (WES) is a popular investigative approach, and one that is becoming optimized to the point where it is now economically viable to include as a regular part of diagnosis and treatment of some diseases. This is possible because WES focuses on sequencing only the coding regions of a patient's DNA (the exome) rather than the entire genome. Because the vast majority of genetic diseases are currently believed to originate from exome variants, and because the exome represents less than 2% of the genome, WES combines high relevance of information with comparatively lower cost and higher throughput.

In typical case/control experiments, a group of patients with a phenotype of interest (e.g. a disease) is sequenced along with a control group, the members of which do not exhibit the phenotype. After identifying the sites at which each participant diverges from the human reference genome, the groups are compared to identify variants that are enriched in either the patient cohort, suggesting the potential for a causative relationship with the disease, or in the control group, suggesting the possibility

of some protective effect. These studies are challenging for many reasons, including the degree of genetic variation displayed across the human population that is unrelated to disease, that some variants appear to be pathogenic in certain human sub-populations but not others, and that some variants cause disease only under certain very specific conditions, such as the simultaneous presence of several variants, or variants which are disease-causing in a heterozygous, but not homozygous, state.

1.4.1. Population effects

If a particular variant is believed to be disease-causing, then there is an evolutionary selective pressure working against the widespread propagation of that variant. By extension, variants that are highly-observed in a population are generally regarded as benign or of otherwise having little impact on reproductive fitness. As a result, sequencing research has historically focused on rare variants that are defined as variants that are observed in a reference population below a semi-arbitrary threshold which varies from study to study, e.g. 1% or 0.1%, as these have a greater likelihood of imparting a larger functional effect.[74, 75] Variants that are thus defined as rare, however, may only be rare in particular sub-populations owing to unequal human population expansion and lower relative genetic diversity within certain groups combined with increased geographic relocation. Since rare variants do tend to be population-specific, by analyzing a diverse group of people collectively, variants that are rare in one group but more common in another may lead to spurious associations, and so compensating for these disparities is necessary to improve reliability of results.[76] One straight-forward method of accomplishing this is to subdivide an experimental group based on the genetic lineage of the subjects and consider each group in isolation of the others, using appropriately-matched reference populations for each group.

Further complicating this process is the fact that someone's genetic ancestry may be distinct from their socially-influenced perception of their identity and lead them to, whether unknowingly or deliberately, misclassify themselves. This can add more noise to the data if experimental grouping is done on the basis of the self-described identity of the participants. Principal component analysis (PCA) is a common tool for dimensional reduction in many areas of research, and in genomics can be used to mathematically cluster participants based on their overall genetic variation in a more objective fashion.[77] Each individual's data are reduced to a single value that represents

their overall genetic variation, and then these values are re-arranged in coordinate space in such a way as to explain as much variance as possible in the data set. This is typically achieved by plotting the first two principal components against each other to visually or mathematically assess participant grouping. Often, the greatest source of variation in an ethnically diverse population coincides with the genetic lineage of the participants, allowing effective stratification and subsequent reduction in unrelated genotype-phenotype association.

1.4.2. Assessment of variant pathogenicity

Even after controlling for population effects, unrelated genotype-phenotype association remains a common issue in genomics studies, and it can be difficult to establish a causative relationship. Variants may have a legitimate association with a phenotype of interest yet not necessarily be disease-causing, or even mechanistically related in any fashion, e.g. a variant in linkage disequilibrium (LD) with the true pathogenic variant. As such, before causation can be determined, other investigations must be done, such as functional evaluation of the variant in vitro, or pedigree studies that show clear evidence of genotype-phenotype segregation in family members. Validating experiments can be time- and labor-intensive, and expensive, so it is useful in exploratory research to narrow the scope of the investigation whenever and however possible.

One such method of focusing on a more manageable number of candidate variants is through the application of computational scores which predict whether or not a given variant is likely to have a deleterious effect on the resultant protein. Some of these algorithms base their assessments on the evolutionary conservation of a given site, whereas others attempt to predict how the chemical properties of the substituted amino acid would alter protein function. Since various approaches appear to offer orthogonal data, there exist algorithms that derive a weighted aggregate from multiple other scores, designed to improve upon any of the components in isolation.

Many tools that attempt to score variants according to their predicted degree of deleteriousness are based on similar underlying principles. Because of this, the American College of Medical Genetics and Genomics (ACMG) has issued guidelines indicating that predicted deleteriousness by multiple scores should be considered one

collective piece of evidence as to a variant's likelihood of pathogenicity.[78] However, the guidelines also observe that it is nonetheless desirable to use multiple tools, as each invokes those underlying principles in unique ways. Thus, the relative strengths and weaknesses of each algorithm can be mitigated to some degree by considering their output in aggregate. In observation of the ACMG guidelines, I used several predictive scores in the evaluation of the deleteriousness of candidate variants, which I selected following extensive literature review in order to assemble a diverse panel according to the data on which their algorithms are based and weighted, their degree of use and validation in the field, and their performance.

1.4.3. Statistical analyses

I expected the statistical analyses in this project to have relatively poor power, owing to the small sample size in the patient cohort, and thus took several approaches in an attempt to increase the likelihood of uncovering an association. First, I restricted my investigation to variants in a panel of specific candidate genes rather than the entire exome. Genome-wide and, to a lesser extent, exome-wide association studies undoubtedly overlook relevant variants that do not achieve statistical significance because of the multiple test corrections performed to account for the sheer number of variants that are considered. By narrowing the focus to genes that may have some direct involvement in metabolic processes, I greatly reduced the number of statistical tests performed. Second, as mentioned earlier, I only evaluated rare variants, as those are more likely to result in an overt change in protein function and thus have a higher propensity to be disease causing than more common variants when considered on an individual basis. Third, rather than assessing variants on an individual basis, I collapsed them to the gene level and performed burden testing on the summed variants.

Looking exclusively at rare variants in a small cohort makes it less likely that a given variant will be observed in multiple participants, and because of that it may not be evident if the variant truly associates with the disease. By pooling the variants in a gene and considering them collectively, it may then be more readily observed that variation in that gene as a whole may have a pathological effect, albeit without the ability to identify which of the variants are specifically responsible.[79]

2. Study goals

With this project I aimed to develop and employ an analysis pipeline to investigate the possible relationship between the development of early-onset AF and the presence of predicted-deleterious variants in metabolism-associated genes. If these variants lead to dysfunctional gene products which disrupt regular metabolic processes and destabilize mitochondrial or cellular energy dynamics, and if the result of this is overproduction of ROS and compromised ROS homeostasis, then accumulation of ROS could follow, which is an arrhythmogenic substrate. I thus hypothesize that this early-onset AF population will be enriched for deleterious variants relative to the general population in a panel of such metabolic genes.

3. Methods

3.1. Participant cohorts

Prior to this study, a cohort of patients was established with inclusion criteria of an AF diagnosis prior to 60 years of age, and without evidence of typical clinical risk factors, such as hypertension, diabetes, valvular heart disease, heart failure, obstructive sleep apnea, and coronary artery disease. An array of patient data was collected, examples of which include blood samples for whole-exome sequencing, 12-lead ECG results, anti-arrhythmic drug (AAD) prescription status, and family history of AF and other disorders. General demographic data are collected in Table 1.

Table 1. Early-onset AF participant demographic data

Population	Participants (n)	Enrolled age (years)	BMI
Total	211	50.7	26.7
Male	175	50.3	24.2
Female	36	52.8	24.4
White	174	50.9	27.0
Asian	32	52.0	25.5
Hispanic	3	45.9	24.9
First Nations	2	24.8	26.6

Number of total participants as well as divided by sex and self-identified ethnicity. Included are the mean age for the group's clinical enrolment and mean BMI following pruning of missing values.

Since this patient group underwent WES as part of their diagnostic and treatment program, a matched control group was unavailable, and I instead used data from the Genome Aggregation Database (gnomAD) 2.1 release as a comparator.[74] It is the largest publicly accessible WES data set and includes well-defined sub-populations, and it provided the allele counts for the control side of many of the case/control statistical analyses performed. For mitochondrial variants, I retrieved frequency data from Mitomap, a large and diverse database of mitochondrial variation.[80]

I utilized data from the March 2019 release of the UK Biobank (UKB) to act as a more rigorously-matched case/control set in an attempt to validate any findings from the patient group in a larger population.[81] The initial release contains WES and extensive clinical data for 50,000 participants. After data pruning, the effective size of the data set used for this study was 49,901 people, including 27,230 women and 22,671 men.

3.2. Exome sequencing

As part of their clinical assessment and treatment, blood samples were collected from participants and whole-exome sequencing was performed by Génome Québec (Montréal, QC) using the NovaSeq 6000 sequencer (Illumina, San Diego, CA). Coding regions were enriched using the SeqCap EZ Human Exome Capture v.3.0 probes (Roche NimbleGen, Pleasanton, CA).

Post-sequencing processing used the GenPipes DNA-Seq pipeline, which is an implementation based on the BROAD Institute Genome Analysis Toolkit (GATK) Best Practices.[82, 83] To summarize, primer sequences were trimmed from the raw reads which were then mapped to the human reference genome using the Burrows Wheeler Aligner bwa-mem tool.[84] Base qualities were adjusted near indels by GATK indels realignment and base recalibration in order to account for the decreased reliability of reference alignment surrounding indels relative to SNVs. Picard markduplicates was used to identify and flag reads likely to be duplicate sequences generated by the exome enrichment PCR amplification rather than legitimate sequencing output, and SNPs and indels were identified using GATK haplotype caller.

3.3. Quality control measurement

I filtered participant VCFs based on several quality control metrics before further analysis. First, I merged the variants of the entire patient cohort into one large file for aggregate analysis and quality filtering using bcftools, and then split multiallelic sites and left-aligned variants.[85] Because some quality metrics require different filtering thresholds for SNVs and indels, I analyzed these variant types separately.

I imported all variants into R and then visualized them across several quality metrics.[86] As per GATK germline variant hard filtering recommendations, I made an effort to tailor the filtering thresholds to this data set in order to strike a balance between retaining true positives and eliminating false positives. As GATK's filtering thresholds are deliberately very lenient, I filtered variants according to the more stringent visually-determined cut-offs using GATK's VariantFiltration feature, and then recombined passing SNVs and indels for further analysis.

3.4. Variant annotation and filtering

I annotated variants with the Ensembl Variant Effect Predictor (VEP) using initially broad settings (e.g. annotations for all available transcripts in which a variant appears) in order to capture as much information as possible for downstream processes.[87] To later focus these data, I then performed several filtering steps. First, I restricted variants to those falling within the candidate gene panel. I retrieved gene border coordinates from the University of California Santa Cruz (UCSC) Genome Browser, and iteratively compared the end coordinates from all available transcripts for each gene until the pair of highest and lowest coordinates were identified.[88] After determining the widest possible base range for each gene, I then used these coordinates with bcftools to subset the relevant genes from the overall data. Using VEP's filter_vep functionality, I then excluded variants less likely to have an impact on protein function (e.g. intronic variants), as well as those with a global allele frequency observed in greater than 1% of the gnomAD population. Mitochondrial variants were evaluated similarly, instead using frequencies from the Mitomap reference data.

Not all deleteriousness predictors score all available transcripts, and different transcripts can have different scores, so I made an effort to determine the most reasonable transcript to use for assessment. In many cases this was the canonical transcript, but I accessed tissue-specific expression data via the Genotype Tissue Expression Project (GTEx) and, where possible, I utilized the transcript most prevalent in atrial appendage samples.[89]

Using bcftools, I measured the transition/transversion ratio (Ti/Tv) of SNVs and the insertion/deletion ratio of indels after various filtering steps in order to compare them to expected values from the literature as an additional means of assessing variant calling accuracy.

3.5. Deleteriousness scoring

Variant deleteriousness scores were included in the initial VEP annotation using both the native VEP cache as well as the dbNSFP and dbscSNV plugins.[90–92] In several cases multiple scores may have been provided for a given variant according to the number of transcripts considered. I accessed the most recent release of each tool's

score database and, wherever possible, cross-referenced all variants with these updated tools to remove ambiguity and revise their assessment to reflect the most recent data available. To demarcate possibly damaging from benign variants I relied on the scores suggested by the respective authors of each tool.

I assessed putative loss-of-function (LOF) variants according to their phred-scaled CADD scores, since certain variant types (e.g. frameshift variants) are not scored by all tools.[93] In addition to CADD scores, in my analysis of splice site variants I also considered their Ada and RF scores from dbscSNV, which estimate whether or not a given substitution is likely to alter splicing at that location. I excluded any LOF variants that were flagged as low confidence by the Loss-Of-Function Transcript Effect Estimator (LOFTEE) plugin for VEP.[74] Whereas LOF variants with sufficient CADD score were included based on that score alone, I scored nuclear missense variants with CADD, PROVEAN, REVEL, SIFT, and VEST4, and only considered variants further if there was unanimity in the predictions.[93–97] Mitochondrial variants I instead evaluated using APOGEE, Mtoolbox, and CAROL.[98–100] Again, I included variants for downstream analysis only if all three scores agreed that a variant was likely to have a deleterious impact on the protein.

3.6. Sub-population clustering

Both the early-onset AF cohort and the UKB participant group included self-identified ethnicity among their demographic data. In addition to partitioning on the basis of this self-identified ethnicity, I explored the patient cohort sub-populations using principal component analysis to examine if that changed the grouping of any subjects, and what effect that ultimately had on candidate gene evaluation.

A cohort-wide VCF of all variants that passed initial quality filters, as described in section 3.3, was converted to BED format using Plink.[101] I used the indep-pairwise function to remove rare variants in the dataset (MAF < 0.01) and then to identify and remove those pairs of variants in linkage disequilibrium (LD). This was done using a sliding window of 50 bases, incrementing 5 bases per step. If a pair of variants within the window had a squared correlation (r^2) of greater than 0.2 at each step, variants were pruned until there were no longer any such pairs remaining. Following this, I employed Plink's pca function to calculate the eigenvalues and eigenvectors for the dataset, which

I then imported into R. I iteratively plotted combinations of the first 20 principal components against each other for visual estimation of population grouping, which I then validated using R's kmeans function. With consideration of the relatively small number of participants from the other groups, I primarily emphasized distinguishing between white and non-white subjects.

3.7. Statistical analyses

I performed statistical analyses in R. After all previous variant filtering steps were completed, I summed individual variant counts to collapse results to the gene level for burden testing. I compared the total number of observed variants in each gene in the patient cohort with the sum of those same variants observed in the gnomAD or Mitomap reference data and calculated odds ratios (ORs) with 95% confidence intervals (CIs) and p values using Fisher's exact test, which I then adjusted to constrain the false discovery rate (FDR) to 5%. In gnomAD, the total allele number is inconsistent between variants as a result of differing numbers of participants whose base calls passed or failed at a particular locus. To account for this when variants are considered at the gene level rather than the variant level, I used the maximum possible number of observations in a given gnomAD population (e.g. 133,770 for non-Finnish Europeans) in these calculations.

3.8. Positive control analysis

A prior study used similar early-onset AF exome data to investigate the relationship between AF and LOF variants and copy number variants (CNV) in cardiomyopathy genes and identified several LOF variants in the gene *TTN*.[102] As a positive control, I applied the variant filtering and evaluation methods described earlier in this thesis to *TTN* variants in this patient exome data, to assess whether this analysis pipeline would also recognize an association between *TTN* variants and AF. With the exception of the gene coordinates used to select the region of interest, all procedures were performed identically to those previously discussed.

3.9. Validation of statistical associations

Based on the results of the prior statistical analyses, I further investigated promising genes that associated with AF in the UKB data set in an attempt to validate my findings in a more rigorous case-control comparison. First, variants from selected genes that were used in the previous burden testing were converted to GRCh38 coordinates using the NCBI Genome Remapping Service.[103] Variant data for UKB participants were then retrieved and analyzed. Any participants that had later withdrawn consent from UKG were excluded from analysis, as well as those with a missing genotype measurement batch or a mismatch between reported and genetically-determined sex. I greatly appreciate the efforts of Mark Trinder and Kate Huang in accessing and merging the relevant genotype and phenotype data from the UKB data set, and without their assistance none of the UKB investigation would have been possible.

First, I excluded non-white participants in order to better match the group from which the tested variants originated. I then classified UKB participants among the AF group if any of the following applied: 1. self-reported atrial fibrillation or flutter during their intake interview at time of program enrolment, 2. automated diagnosis of either condition from resting ECG data during a UKB assessment, or 3. associated hospital in-patient admission at any point with ICD-10 code I48, the category code for atrial fibrillation and flutter. There were 2,964 such individuals. To more specifically identify participants as part of an early-onset subset of this AF population, I applied exclusionary criteria that were as similar as possible to those for the original early-onset AF patient cohort: diagnosis before 60 years of age (n=622), and without recorded evidence of other pathologies: hypertension, heart failure, valvular diseases, cardiomyopathy, diabetes, COPD, sleep apnea, and hyperthyroidism (n=320). As described previously, I counted variants in the UKB early-onset AF group and the UKB non-AF group (n=43,593) and compared using Fisher's exact test with FDR-correction, and calculated ORs and 95% CIs. I then performed a secondary analysis without the age restriction, and/or without comorbidities, to investigate if there was any broader AF association as well.

3.10. Atrial fibrillation comorbidity association

Since AF is a multifaceted disease that interacts with numerous conditions, it is possible that variants that appear to associate with AF are actually associated with one or more of its comorbidities, even if that comorbidity exists at a subclinical level and has not yet been diagnosed. To investigate this possibility, I divided the UKB data set into several other case/control groups for hypertrophic cardiomyopathy (HCM), diabetes (both types I and II, as well as those for which the type was unspecified in the UKB data), hypertension, and stroke. I summed variants as before in both cases and controls and compared using the aforementioned statistical tests.

4. Results

4.1. Quality control measurement

A representative example of the visual evaluation of quality control metrics described in section 3.3 is displayed in Figure 1.

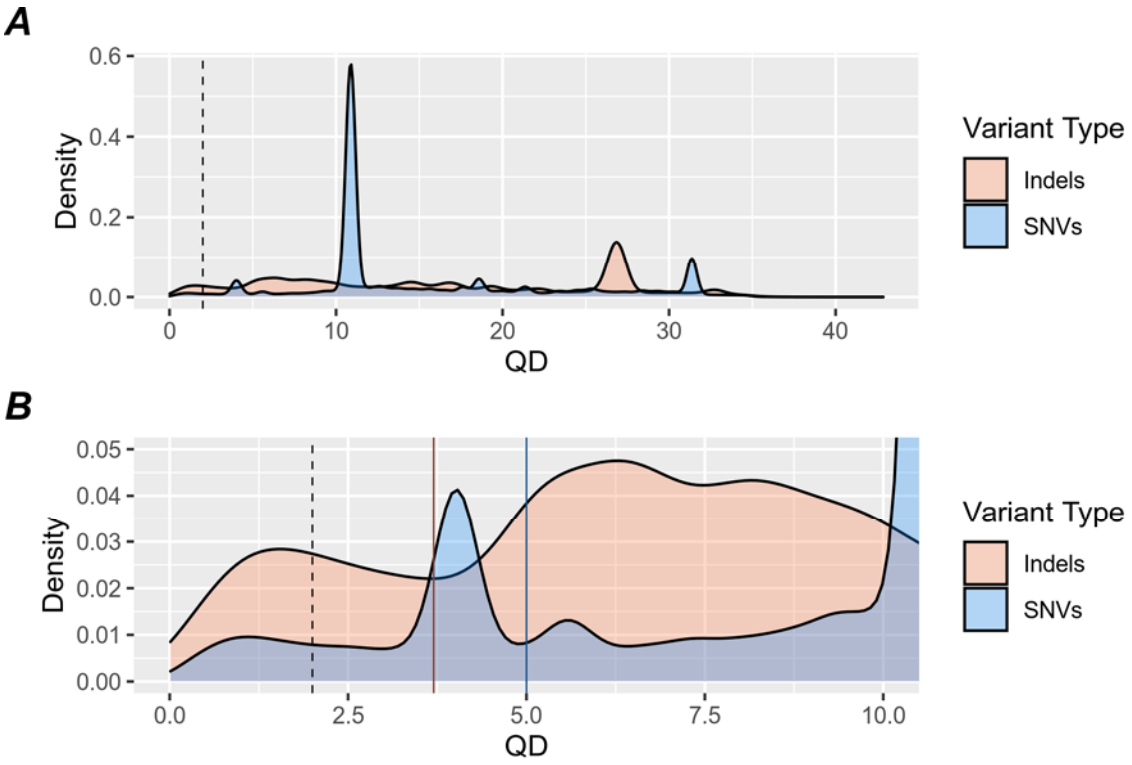


Figure 1. Density plot of variants by QD value

A. SNVs and indels are individually plotted according to variant quality by depth (QD) value. The dashed line is placed at the GATK recommended threshold for both SNVs and indels (2). B. Zoomed view of the region where the cutoffs are located. The colored lines correspond to the visually determined filter cut offs, red for indels (3.7) and blue for SNVs (5).

Tailoring the thresholds to these data yielded filtering cut offs that were more stringent than the GATK default recommendations, and resulted in a larger number of removed variants, shown in Table 2. Note that columns do not sum to the total number of variants removed under each set of filtering conditions, as a single variant can fail multiple criteria.

Table 2. Number of variants removed by each quality filter

		SNV				INDEL			
		GATK		Graph		GATK		Graph	
QD	<	66,765	2.0	262,506	5.0	39,302	2.0	71,224	3.7
FS	>	1,786	60	139,477	5.0	50	300	30,901	5.0
MQ	<	269,260	40	324,369	41				
MQRS	<	416	-12.5	53,729	-2.5				
RPRS	<	16	-8.0	65	-6.0	0	-20	180	-6.25
SOR	>	72,854	3.0	72,854	3.0	16,876	10	67,588	3.0

The number of variants removed by a given filtering threshold is shown, along with the direction of removal. E.g. using the GATK default value of filtering out variants with a QD lower than 2.0, 66,785 variants are removed. In contrast, using the 5.0 threshold derived visually from the density plots, 262,506 variants were removed. QD: QualByDepth, FS: FisherStrand, MQ: RMSMappingQuality, MQRS: MappingQualityRankSumTest, RPRS: ReadPosRankSumTest, SOR: StrandOddsRatio.

I applied both filter sets independently in order to quantify the total number of variants that were retained in each case, and these results are shown in Table 3. Observing the less forgiving filtering criteria removed an additional 8% of variants.

Table 3. Comparison between filtering thresholds

	Unfiltered	GATK	%	Graph	%
SNV	4,198,663	3,821,624	91.02	3,498,380	83.32
Indel	794,364	755,029	95.05	671,383	84.52
Total	4,993,027	4,576,653	91.66	4,169,763	83.51

The number and proportion of variants retained after applying the collection of GATK default values or the graph-derived values to the collection of cohort-wide variants.

The expected transition/transversion ratios for exon-located synonymous and non-synonymous SNVs are 4.9 and 2.1 respectively, for an overall exonic SNV Ti/Tv of 2.8. Initially, the cohort-wide Ti/Tv was 1.68, but this improved after discarding variants from outside exon boundaries. These values increased further after the application of quality filters to slightly higher than the expected values. and may be evidence of data bias due to artifacting introduced during the sequencing or variant calling processes.

Table 4. Transition/transversion ratios following data processing steps

	Ti/Tv ratio		
Data processing step	sSNV	nsSNV	Combined
Initial variants			1.68
Exon restriction	4.69	2.06	2.78
Quality filtering	5.09	2.22	2.98

Transition/transversion ratio is given for synonymous SNVs (sSNV) non-synonymous SNVs (nsSNV) and both types collectively (Combined).

4.2. Principal component analysis

Population comparison used iterative combinations of the first 20 principal components. The first principal component provided the best capacity to distinguish between populations, and the point at which to separate the white and non-white participants was visually estimated at a PC1 value of 0.

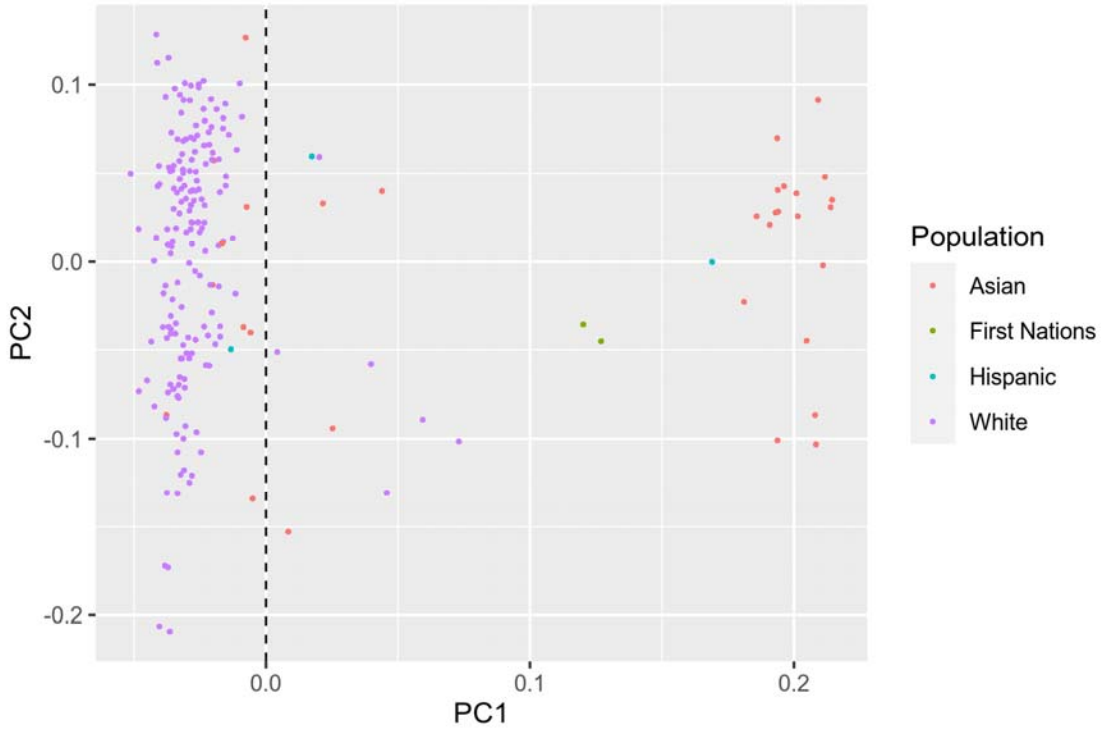


Figure 2. Principal component analysis and population clustering

Populations are colored according to self-identified ethnicity. The vertical line at 0 represents the visual estimation of the point of division between the white and non-white populations.

Based on this division, nine self-identified Asian participants were grouped within the white population, as well as one Hispanic participant. Conversely, six patients who identified as white were excluded from that group. I validated group demarcation using k-means clustering, and found that three clusters were ideal for isolating the white participants, shown in Figure 3. Examples of other explored k-means values can be viewed in Appendix D.

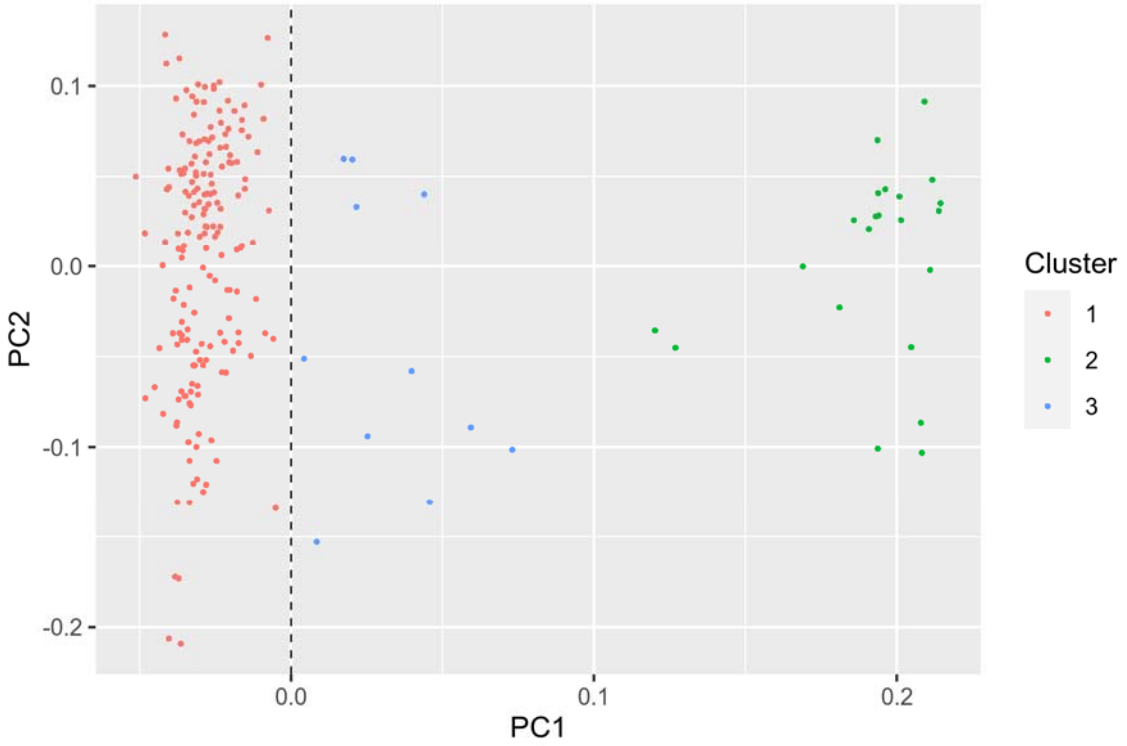


Figure 3. Computational determination of population clustering using k-means

Population distribution was computationally explored to validate the visual estimation using k-means clustering. Three clusters were ideal for separating the white and non-white subjects, and coincided with the visual estimation.

The PCA recategorization resulted in an increase in the white population by four members, and an overall change in its participant composition by approximately ten percent. Because of the magnitude of these changes, I ran a revised statistical analysis to evaluate the potential change in identified AF-associating genes.

4.3. Early-onset burden testing

4.3.1. Cohort-wide analysis

Of 87 nuclear genes that were candidates for burden testing, 53 had evidence of a positive association with AF (FDR-adjusted p<0.05, 95% CI lower bound > 1), with three of seven mitochondrial genes showing similar results. Acyl-CoA dehydrogenases were highly represented, including *ACAD9* (p=0.016), *ACADS* (p=0.0014), and *ACADVL* (p=0.0014). along with several other genes related to acyl-CoA or acetyl-CoA. A large number of genes that contribute products to mitochondrial complexes I through V also had evidence of an association. A small selection of tested genes is shown in Table 5. Due to the large number of genes with an apparent association, I investigated further to ensure these results were not influenced by population effects.

Table 5. Selected results from full-cohort burden testing

Gene	OR	95% CI	P	Gene	OR	95% CI	P
ACACB	2.2	1.1 - 4.0	0.049	*ATP5F1A*	10.6	2.86 - 27.8	0.005
ACAD9	199	3.8 - 2648	0.016	*MT-ATP8*	4.3	1.16 - 11.4	0.031
ACADS	240	22.7 - 1448	0.001	*MT-CYB*	244	17.64 - 3568	0.002
ACADVL	20.2	5.4 - 53.5	0.001	*MT-ND5*	12.5	2.49 - 38.8	0.007
ACSM4	149	3.0 - 1544	0.018	*PC*	13.5	4.32 - 32.3	0.001
ACSS2	19.0	2.2 - 71.8	0.014	*PRKAB2*			0.006
AGL	7.59	1.55 - 22.6	0.018	*SLC22A5*	7.5	2.40 - 17.7	0.005

Odds ratios (OR) are shown with 95% confidence intervals (95% CI) and FDR-corrected p values (P). The *PRKAB2* variant did not exist in gnomAD, and as such no OR could be calculated.

4.3.2. Sub-populations analyses

I repeated burden testing on the self-identified white participants (n=174) and self-identified Asian participants (n=32) separately and a comparison of all analyses is given in Table 6. Of the white cohort, 33 of 66 nuclear genes retained a positive AF association, whereas none of the five tested mitochondrial genes did. The Asian cohort had 51 nuclear genes and six mitochondrial genes tested, with 43 and five respectively statistically associating with AF, suggesting that the Asian subpopulation was a powerful contributor to the collective results. When comparing the results from the white participants with those of the entire cohort, three genes gained an association from analyzing the white population independently, whereas the genes that lost an

association numbered 16. Thirty genes were associated in both cases. It was further revealed that the ethnicity coding in the patient data allowed no way to distinguish between genetically distinct groups that might all describe themselves as Asian, a situation not unlikely given the presence of large numbers of both South Asian and East Asian residents in the geographic region from which the patient cohort was recruited, and one which would result in population-reference mismatch and inflation of results.

4.3.3. Population adjustment

Following the discovery of the ambiguous nature of the ethnicity coding, I performed principal component analysis and repeated gene burden testing with the revised participant categorization. 21 associated genes were shared between the PCA-defined white participants and the self-identified white participants, whereas association was lost for seven genes, but gained for one other. In total, 30 of 77 nuclear genes and one of two tested mitochondrial genes achieved an association. The associated genes and specific variants are collected in Appendix B.

Table 6. Comparison of population-specific gene burden testing results

Population	Genes tested	Genes associating	%
All participants	94	56	59.6
Asian (self-identified)	57	48	87.7
White (self-identified))	71	33	46.5
White (PCA-defined)	80	30	39.2

The total number of genes tested for each cohort are compared with the proportion which had FDR-adjusted $p<0.05$ and an OR 95% CI that did not include 1.

4.4. Replication of Titin association

Across the entire patient group there were 191 variants in *TTN* which passed initial quality filtering. These comprise one frameshift deletion, three in-frame deletions, one splice acceptor variant, four stop gained variants, and 182 missense variants. The splice acceptor variant was discarded after being flagged by LOFTEE due to the presence of a rescue acceptor sequence several bases away. Only 35 SNVs remained after deleteriousness scoring. While the majority were observed in only a single participant, there was one missense variant that was found in four people, two that were found in three people, and four variants with two observations, one of which was a

homozygous substitution. Of the five LOF variants, two were observed in the white population and two in the Asian population following PCA-adjustment. One was in an individual who did not definitively cluster with either group, but who identified as white.

Following burden testing of the white participants, this analytical protocol did identify *TTN* as having an association with AF with an odds ratio of 11.8 (95% CI 8.16 – 16.64), and $p=1.70*10^{-26}$.

4.5. UK Biobank burden testing

Despite the strength of the associations discovered in the smaller data set, none of the genes selected for UKB comparison yielded similar results in the larger data set. Many of the selected variants were not observed at all in the UKB case groups, and there was no observed association for any of diagnosis under 60 years of age, either with or without comorbidities, or diagnosis at any age, again with or without comorbidities. Results from UKB subset with early diagnosis and no comorbidities are shown in Table 7, and are representative of the results from the other comparisons.

Table 7. Comparison of burden testing results between data sets

Gene	Variants (AF)	Variants (UKB)	Hits (AF)	Hits (UKB)	P (AF)	P (UKB)
ACADL	2	1	3	6	0.224	1
ACSM5	2	1	2	6	0.092	1
EHHADH	2	3	2	5	0.012	1
FBP2	2	2	11	7	0.070	1
GYS1	3	1	3	1	0.016	1
H6PD	2	1	5	2	0.079	1
HK3	2	2	4	3	0.180	1
PC	4	1	4	1	0.0005	1
PYGM	5	1	5	2	0.035	1
SLC22A5	5	2	5	6	0.024	1
SUCLG2	1	1	1	6	0.162	1

The variants columns (2 and 3) show the number of unique variants within each gene that contributed to the burden testing results. The hits columns (4 and 5) illustrate the total number of variants counted in each gene, including those that were observed multiple times. The AF columns (2, 4, 6) are data from the early-onset AF patient cohort, and the UKB columns (3, 5, 7) are data from the UKB data set. P values are FDR-corrected. Genes without any observed variants in the UKB data set are not included in this table.

4.6. Comorbidity association

Several genes associated with AF comorbidities. While there were none found with either the specific type I or type II diabetes groups, *ACSS2* was positively associated with the group of participants with an unspecified diabetes type (OR 2.24, 95% CI 1.27-3.71, p = 0.043). *ACSS2* (OR 1.56, 95% CI 1.16-2.08, p=0.023), and *COX15* (OR 1.71, 95% CI 1.21-2.38, p=0.028) were both associated with hypertension, and *EHHADH* was also found to associate with stroke (OR 2.76, 95% CI 1.61-4.44, p=0.002).

5. Discussion

This project involved the development of an analytical pipeline for the evaluation of rare variants and the subsequent association between the metabolism-related genes harboring those deleterious variants and atrial fibrillation. I successfully validated this pipeline by replicating a previously published association between variants in the gene *TTN* and early-onset AF in the same data set, and then applied the protocol to uncover a number of possible associations between AF and genes contributing to cellular metabolic function. While I was not able to replicate any of these associations in the larger UKB data set, in further exploring the data I uncovered associations between several of these genes and other diseases that are comorbid with AF, including diabetes, hypertension, and stroke.

5.1. Pipeline validation

The authors of the previous study pooled the patient population used in this project with a second of similar clinical presentation, giving them an experimental group of 195 after PCA and restriction to white patients, whereas this study had only the 178 from one clinical group.[102] While this difference in population introduces some degree of uncertainty when comparing the results of the *TTN* validation, the patient pool utilized for this project does represent the majority of that larger combined group, so results should generalize reasonably well. Additionally, the control group used for allele frequency data by Lazarte et al. was drawn from the 1000 Genomes Project (1KG) data, whereas I utilized gnomAD.[104] Since gnomAD is a much larger database, and has itself absorbed the 1KG data set, it is possible that differences in number of allele observations between the two could influence significance of results. However, given its size and participant diversity, gnomAD likely offers a more accurate representation of overall allele frequency, particularly for rare variants which may be less reliably quantified in smaller samples. Though there are several discrepancies between our relative approaches, the goal of demonstrating that this variant assessment pipeline was capable of reproducing an association between AF and *TTN* variants was successful.

5.2. Early-onset cohort gene association

Among the genes that appeared to associate with an increased risk of AF, there were several functional clusters. Eight are relevant to processes involving coenzyme A (CoA), including acyl-CoA dehydrogenases which catalyze the initial and rate-limiting step of fatty acid β oxidation (*ACAD9, ACADS, ACADVL*), acyl-CoA synthetases which produce acetyl-CoA to fuel the tricarboxylic acid (TCA) cycle among other destinations (*ACSM4, ACSM5, ACSS2, ACSS3*), and a pantothenate kinase, which catalyzes the first and rate-limiting step of CoA synthesis (*PANK1*). There was also one gene which codes a TCA enzyme (*MDH2*). Four other genes are also involved with β oxidation, one in the mitochondria (*HADHB*) and three in peroxisomes (*DECR2, EHHADH, HSD17B4*). Nine genes contribute to ATP synthesis through the electron transport chain, and are distributed between complex I (*NDUFS1, NDUFV1*, *NDUFV3*), complex III (*MT-CYB*), complex IV (*COA3, COX10, COX15*), and ATP synthase (*ATP5F1A, ATP5F1C*). *ACAD9*, mentioned prior, also acts as an assembly factor for complex I. Six genes are involved in carbohydrate metabolism via glycolysis (*PKM*), gluconeogenesis (*PC*), glycogenolysis (*PYGB, PYGM*), and glycogen synthesis (*AGL, GYS1*). The remaining two have functions in ketogenesis (*HMGCL*) and carnitine transport (*SLC22A5*). Under normal physiological conditions, the majority of the mature heart's energy demands are met through oxidative phosphorylation, with β oxidation of fatty acids the primary contributor.[105] As such, the number of associated genes that are involved with lipid metabolism in some fashion is interesting, and particularly the number which code for crucial enzymes involved in the rate-limiting steps of their respective pathways.

Metabolism represents the complicated intersection of many energy pathways and even more regulatory networks and as such, dysfunction in these semi-independent systems has substantial overlap with respect to disease presentation.[106, 107] Intuitively, compromised energy pathways tend to manifest most obviously in tissues which are highly metabolically active, and arrhythmias, cardiomyopathies, and optic neuropathies are common, but there exists a remarkable range of presentations, even between patients harboring the same variant. In fact, it is not uncommon to see a given variant appear to be responsible in one case for widespread multi-organ failure, whereas in another case only a single tissue is impacted.[108, 109] Many individuals with one of these conditions tend to be diagnosed early in life due to the severity of their symptoms, and

often die very young, often within a number of years, if not weeks or even days. However, age of onset is variable, and some individuals do not become symptomatic until years or decades later in life, and tend to have a much better prognosis.[110, 111] Perhaps unsurprisingly, disease severity is equally unpredictable with some disease-associated variants appearing somewhat frequently in healthy populations, suggesting a more complex interaction with some combination of other genetic, or perhaps environmental, factors.[112]

Many of the genes, and even specific variants, found here to be associated with AF have also been identified in other research as imparting greater risk for, or even in some cases causing, some of these potentially serious diseases, yet the patients in this study cohort do not display such symptoms. There is some evidence that disease severity can depend upon the amount of retained protein function, but that too is unpredictable and even individuals with variants known to be-pathogenic have been found to be asymptomatic.[112–114] As one might expect, severe complications are often the result of homozygous variants, but heterozygous associations are not uncommon and there is not often a detectable genotype-phenotype correlation.[106] This could be due to interactions with other variants in the gene, a compounding effect with variants in genes that are responsible for downstream processes, or perhaps modulating effects of variants in promoter regions.

As the heart depends on β-oxidation for supplying much of its considerable energy requirements, it comes as no surprise that dysfunction in the acyl-CoA dehydrogenases can be devastating. *ACAD9* dysfunction, for example, is a common cause of complex I deficiency which often manifests as cardiomyopathy that proves fatal within several years, though as mentioned before the range of severity of presentation is wide.[115] *ACADVL* variants have been observed in patients with atrial and ventricular ectopics, and fibroblast-derived human induced pluripotent stem cells differentiated into cardiomyocytes (hiPSC-CMs) from these individuals displayed multiple substrates for arrhythmia compared with control cells: shorter action potentials, increased frequency of delayed afterdepolarizations, and higher cytosolic Ca^{2+} concentrations which persisted through both systole and diastole.[116] In addition to this Ca^{2+} dysregulation, *ACADVL* variants also appear to be responsible for increased ROS production and destabilization of the mitochondrial membrane potential, also triggers for arrhythmia.[64, 117] The c.1700G>A variant found in this cohort has published evidence of decreasing levels of

protein expression while leaving the patient clinically asymptomatic, which implies a potential subtle threshold effect wherein they may initially experience a subclinical reduction in oxidative phosphorylation throughput but gradually progress to something that manifests later in life, as can be the case in acyl-CoA dehydrogenase dysfunction.

Also important for the process of β-oxidation is the carnitine transporter OCTN2, the product of *SLC22A5*, which transfers long-chain fatty acids into the mitochondrial matrix so they can be processed. Primary carnitine deficiency (PCD) is an autosomal recessive disease, though in some diagnosed patients only a single heterozygous risk allele is identifiable. The functional impact of variants in this gene have a tremendous range, from complete abolition of transporter activity to an observed 50% gain of function.[118, 119] Three cohort variants have been previously identified in individuals with PCD, c.34G>A, c.136C>T, and c.1451G>T.[120] C.34G>A has been observed alongside both mild symptoms, such as intolerance to fasting or increased fatigability, and severe outcomes, such as cardiac arrest.[119] The most commonly observed variant in PCD patients is the c.136C>T substitution, which reduces but does not completely eradicate transport.

Ultimately, the majority of the reducing equivalents produced through β-oxidation and other metabolic processes contribute electrons to complex I of the ETC and, as such, reduced function in many of the related ETC proteins can have profoundly harmful effects. Variants in the complex I genes such as *NDUFS1*, like *ACAD9*, can lead to complex I deficiency, through either reduced presence or reduced activity of the holoenzyme. These variants can also see a compromised mitochondrial membrane potential and increased susceptibility to oxidative damage.[121, 122] ROS hypersensitivity was demonstrated with *NDUFV1* variants as well, reinforcing the relationship between energy pathway dysfunction and ROS intolerance as a mechanism for arrhythmia.[123]

COX10 and *COX15* are both instrumental in complex IV assembly and function through their contributions to heme A synthesis.[124] One of the *COX15* variants found in this patient cohort, c.520G>A, has been described in prior research in a heterozygous state in an individual with arrhythmogenic right ventricular cardiomyopathy, though it did not segregate with the phenotype within family members as expected.[125] Like some of the previous discrepancies discussed, perhaps this is because it is not sufficient to cause disease, or perhaps it is due to incomplete penetrance owing to a more complex

genotype-phenotype relationship. Another cohort variant, c.452C>G, forms a premature stop codon (p.S151X) in exon 4. The child of a healthy father heterozygous for this variant, and a healthy mother heterozygous for a different variant, was born with severe lactic acidosis and HCM.[126] It was also observed that the effects were far more severe in the heart than in skeletal muscle, emphasizing the continuum of presentation and, perhaps in milder cases, the potential for tissue specificity to mask symptoms that would otherwise lead to a diagnosis.

5.3. UK Biobank validation

There are a number of reasons that associated findings may not have been reproducible in the UKB cohort. The initial investigation, even after PCA adjustment of ethnicity, still treated all white subjects as genetically equivalent, which is not reliably assumed. The gnomAD data are divided into Finnish and non-Finnish groups for allele frequency reporting purposes, but the latter can be just as easily expanded to a number of well-defined sub-populations, e.g. Swedish, southern European, north-western European.[74] This semi-arbitrary grouping, particularly when done inconsistently between case and control populations, can lead to uncontrolled population effects. Similarly, while participants in gnomAD are self-certified as healthy, AF is a condition that is rarely observed earlier in life. As such, gnomAD participants who carry risk alleles for AF and will eventually develop it later in life have no way of knowing that at their time of participation, so the use of large public databases as healthy control populations is dubious in some cases, particularly for conditions like AF which are relatively common, and where symptoms can take decades to manifest.

There are also obvious weaknesses in comparing data generated using two distinct sequencing and processing pipelines. Each will have its own biases and sources of sequencing artifact, and this mismatch between the two will exaggerate these effects further. This will continue to be emphasized as downstream filtering processes are further removed from each other and may be reflected in the number of variants in the AF cohort that are not observed in gnomAD. Being processed collectively, the UKB data minimizes these effects, and so it is possible that despite the quality control efforts in which I engaged, some of the variants contributing to the AF associations in the patient cohort were erroneous.

Even in a best-case scenario from a technical standpoint, the patient cohort was quite small for a rare variant study. If, by coincidence, a given variant was observed in even two people instead of one, it could appear far more common in the AF cohort than it actually is in that population, which could inflate statistical comparisons relative to a much larger and more representative data set.

5.4. Comorbidity investigation

The product of *ACSS2* is the enzyme acyl-CoA synthetase short chain family member 2. Consuming acetate and CoA as substrates it produces acetyl-CoA, which has many physiological applications including fueling the TCA cycle and fatty acid synthesis, both of which are processes dysregulated in diabetes. Current research is incomplete, but both acetate and acetyl-CoA seem to be involved in aspects of metabolic regulation, and *ACSS2* appears to play an important role. In diabetic patients, for example, plasma acetate levels are increased.[127] Studies using both rat and mouse models have also indicated that reduced expression is correlated with increased plasma concentrations of both triglycerides and glucose and also with decreased glucose tolerance.[128, 129] Interestingly, supplemental acetate lowers plasma glucose and insulin in a mouse model of diabetes.[130] Diabetes is a complex phenotype and research into the contribution that *ACSS2* has is preliminary, as well as whether it participates in the progression to a diabetic state, or its modulated expression is exclusively part of a compensatory response.

As a crucial factor in complex IV assembly and function, variation in cytochrome c oxidase assembly homolog, the product of *COX15*, has implications in disruption of ATP generation via the ETC and subsequent overproduction of ROS. Although investigations of specific mechanisms have not yielded a great degree of enlightenment, it is clear that oxidative stress is a powerful contributor to the onset and progression of hypertension, and that hypertension also contributes to worsening mitochondrial dysfunction and ROS dysregulation.[131]

EHHADH codes for enoyl-CoA hydratase and 3-hydroxyacyl-CoA dehydrogenase, which are components of the peroxisomal trifunctional enzyme and are enzymes involved in the peroxisomal fatty acid oxidation pathway. While there is little research in how this gene might interact with stroke, there is considerable research

available on upstream genes, such as *PPARA,* which codes for the peroxisome proliferator activated receptor-alpha and is a powerful regulator of peroxisomal lipid metabolism. This receptor is activated by fatty acids and a mouse knockout model displayed a similar phenotype to that of humans who have pathogenic variants in fatty acid oxidation genes.[132, 133] Prophylactic treatment through exogenous activation of the receptor in mice decreased both incidence and severity of stroke, an effect which was lost when the receptor agonist was applied to mice not expressing the protein.[134] Although upstream of *EHHADH* and having additional physiological influences, *PPARA* does regulate peroxisomal beta oxidation, and as such it is possible that dysfunction in this gene could have some degree of symptomatic overlap with dysfunction in *EHHADH*.

6. Conclusion

In this project I developed an exome rare variant analysis pipeline that I used to successfully replicate an association with variants in *TTN* in an early-onset AF patient cohort. I then applied the analytical protocol to this same clinical population to investigate genes involved in cellular metabolic processes, yielding a number of strong associations. Although I was then equally successful at being unable to validate any of the most promising candidates in the UKB data set, I found further secondary associations with diabetes, hypertension, and stroke, all risk factors for, or complications due to, the presence of AF. While none of the specific variants which contributed to the associations between these genes and their respective afflictions appeared in any publications supporting the associations at the time of this writing, several of them did appear in the literature implicated for other diseases, both cardiovascular and otherwise, establishing at least some degree of general pathological capacity. Given the complicated and bidirectional regulatory relationships between genes, tissues, and energy systems, it appears entirely possible that a pathogenic variant which disrupts one aspect of metabolism could induce a multitude of other effects, both deleterious and ameliorating, the delicate balance of which ultimately determines the onset of physiological decompensation and disease. While much more research is needed due to the scope and complexity of the underlying pathophysiology in the numerous interconnected systems, this project does contribute several new interesting directions for further investigation.

References

1. Lloyd-Jones DM. Lifetime Risk for Development of Atrial Fibrillation: The Framingham Heart Study. *Circulation*. 2004; 110: 1042-1046.
2. Roche F, Gaspoz JM, Da Costa A et al. Frequent and prolonged asymptomatic episodes of paroxysmal atrial fibrillation revealed by automatic long-term event recorders in patients with a negative 24-hour Holter. *Pacing Clin Electrophysiol*. 2002; 25: 1587-1593.
3. Savelieva I, Camm AJ. Clinical relevance of silent atrial fibrillation: prevalence, prognosis, quality of life, and management. *J Interv Card Electrophysiol*. 2000; 4: 369-382.
4. Healey JS, Connolly SJ, Gold MR et al. Subclinical atrial fibrillation and the risk of stroke. *N Engl J Med*. 2012; 366: 120-129.
5. Boriani G, Laroche C, Diemberger I et al. Asymptomatic atrial fibrillation: clinical correlates, management, and outcomes in the EORP-AF Pilot General Registry. *Am J Med*. 2015; 128: 509-18.e2.
6. Vlachos K, Letsas KP, Korantzopoulos P et al. Prediction of atrial fibrillation development and progression: Current perspectives. *World J Cardiol*. 2016; 8: 267-276.
7. Wijffels MC, Kirchhof CJ, Dorland R, Allessie MA. Atrial Fibrillation Begets Atrial Fibrillation : A Study in Awake Chronically Instrumented Goats. *Circulation*. 1995; 92: 1954-1968.
8. Thrall G, Lane D, Carroll D, Lip GY. Quality of life in patients with atrial fibrillation: a systematic review. *Am J Med*. 2006; 119: 448.e1-19.
9. Singh SN, Tang XC, Singh BN et al. Quality of life and exercise performance in patients in sinus rhythm versus persistent atrial fibrillation: a Veterans Affairs Cooperative Studies Program Substudy. *J Am Coll Cardiol*. 2006; 48: 721-730.
10. Jaakkola J, Mustonen P, Kiviniemi T et al. Stroke as the First Manifestation of Atrial Fibrillation. *PLoS One*. 2016; 11: e0168010.
11. Saber H, Thrift AG, Kapral MK et al. Incidence, recurrence, and long-term survival of ischemic stroke subtypes: A population-based study in the Middle East. *Int J Stroke*. 2017; 12: 835-843.
12. Benjamin EJ, Wolf PA, D'Agostino RB, Silbershatz H, Kannel WB, Levy D. Impact of Atrial Fibrillation on the Risk of Death : The Framingham Heart Study. *Circulation*. 1998; 98: 946-952.
13. Chugh SS, Havmoeller R, Narayanan K et al. Worldwide epidemiology of atrial fibrillation: a Global Burden of Disease 2010 Study. *Circulation*. 2014; 129: 837-847.
14. Go AS, Hylek EM, Phillips KA et al. Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study. *JAMA*. 2001; 285: 2370-2375.
15. Wu EQ, Birnbaum HG, Mareva M et al. Economic burden and co-morbidities of atrial fibrillation in a privately insured population. *Curr Med Res Opin*. 2005; 21: 1693-1699.
16. Andrade J, Khairy P, Dobrev D, Nattel S. The clinical profile and pathophysiology of atrial fibrillation: relationships among clinical features, epidemiology, and mechanisms. *Circ Res*. 2014; 114: 1453-1468.

17. Alonso A, Krijthe BP, Aspelund T et al. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium. *J Am Heart Assoc*. 2013; 2: e000102.
18. Fox CS, Parise H, D'Agostino RB et al. Parental atrial fibrillation as a risk factor for atrial fibrillation in offspring. *JAMA*. 2004; 291: 2851-2855.
19. Campuzano O, Perez-Serra A, Iglesias A, Brugada R. Genetic basis of atrial fibrillation. *Genes Dis*. 2016; 3: 257-262.
20. Ellinor PT, Yoerger DM, Ruskin JN, MacRae CA. Familial aggregation in lone atrial fibrillation. *Hum Genet*. 2005; 118: 179-184.
21. Tucker NR, Clauss S, Ellinor PT. Common variation in atrial fibrillation: navigating the path from genetic association to mechanism. *Cardiovasc Res*. 2016; 109: 493-501.
22. Christophersen IE, Ellinor PT. Genetics of atrial fibrillation: from families to genomes. *J Hum Genet*. 2016; 61: 61-70.
23. Voigt N, Heijman J, Wang Q et al. Cellular and molecular mechanisms of atrial arrhythmogenesis in patients with paroxysmal atrial fibrillation. *Circulation*. 2014; 129: 145-156.
24. Nattel S. New ideas about atrial fibrillation 50 years on. *Nature*. 2002; 415: 219-226.
25. Wakili R, Voigt N, Kääb S, Dobrev D, Nattel S. Recent advances in the molecular pathophysiology of atrial fibrillation. *J Clin Invest*. 2011; 121: 2955-2968.
26. Nielsen JB, Graff C, Pietersen A et al. J-shaped association between QTc interval duration and the risk of atrial fibrillation: results from the Copenhagen ECG study. *J Am Coll Cardiol*. 2013; 61: 2557-2564.
27. Nattel S, Harada M. Atrial remodeling and atrial fibrillation: recent advances and translational perspectives. *J Am Coll Cardiol*. 2014; 63: 2335-2345.
28. Gollob MH, Jones DL, Krahn AD et al. Somatic mutations in the connexin 40 gene (GJA5) in atrial fibrillation. *N Engl J Med*. 2006; 354: 2677-2688.
29. Igarashi T, Finet JE, Takeuchi A et al. Connexin gene transfer preserves conduction velocity and prevents atrial fibrillation. *Circulation*. 2012; 125: 216-225.
30. Zahid S, Cochet H, Boyle PM et al. Patient-derived models link re-entrant driver localization in atrial fibrillation to fibrosis spatial pattern. *Cardiovasc Res*. 2016; 110: 443-454.
31. Allessie MA, de Groot NM, Houben RP et al. Electropathological substrate of long-standing persistent atrial fibrillation in patients with structural heart disease: longitudinal dissociation. *Circ Arrhythm Electrophysiol*. 2010; 3: 606-615.
32. Ohtani K, Yutani C, Nagata S, Koretsune Y, Hori M, Kamada T. High prevalence of atrial fibrosis in patients with dilated cardiomyopathy. *J Am Coll Cardiol*. 1995; 25: 1162-1169.
33. Verheule S, Wilson E, Everett T, Shanbhag S, Golden C, Olgin J. Alterations in atrial electrophysiology and tissue structure in a canine model of chronic atrial dilatation due to mitral regurgitation. *Circulation*. 2003; 107: 2615-2622.
34. Lie JT, Hammond PI. Pathology of the senescent heart: anatomic observations on 237 autopsy studies of patients 90 to 105 years old. *Mayo Clin Proc*. 1988; 63: 552-564.
35. Li D, Fareh S, Leung TK, Nattel S. Promotion of atrial fibrillation by heart failure in dogs: atrial remodeling of a different sort. *Circulation*. 1999; 100: 87-95.
36. Morin DP, Bernard ML, Madias C, Rogers PA, Thihalolipavan S, Estes NA. The State of the Art: Atrial Fibrillation Epidemiology, Prevention, and Treatment. *Mayo Clin Proc*. 2016; 91: 1778-1810.

37. Wyse DG, Waldo AL, DiMarco JP et al. A comparison of rate control and rhythm control in patients with atrial fibrillation. *N Engl J Med*. 2002; 347: 1825-1833.
38. Haïssaguerre M, Jaïs P, Shah DC et al. Spontaneous initiation of atrial fibrillation by ectopic beats originating in the pulmonary veins. *N Engl J Med*. 1998; 339: 659-666.
39. Chiang CE, Naditch-Brûlé L, Murin J et al. Distribution and risk profile of paroxysmal, persistent, and permanent atrial fibrillation in routine clinical practice: insight from the real-life global survey evaluating patients with atrial fibrillation international registry. *Circ Arrhythm Electrophysiol*. 2012; 5: 632-639.
40. Cheniti G, Vlachos K, Pambrun T et al. Atrial Fibrillation Mechanisms and Implications for Catheter Ablation. *Front Physiol*. 2018; 9: 1458.
41. Lafuente-Lafuente C, Valembois L, Bergmann JF, Belmin J. Antiarrhythmics for maintaining sinus rhythm after cardioversion of atrial fibrillation. *Cochrane Database Syst Rev*. 2015; CD005049.
42. Roy D, Talajic M, Nattel S et al. Rhythm control versus rate control for atrial fibrillation and heart failure. *N Engl J Med*. 2008; 358: 2667-2677.
43. Weng LC, Choi SH, Klarin D et al. Heritability of Atrial Fibrillation. *Circ Cardiovasc Genet*. 2017; 10:
44. Jennings RB, Ganote CE, Reimer KA. Ischemic tissue injury. *Am J Pathol*. 1975; 81: 179-198.
45. van Bragt KA, Nasrallah HM, Kuiper M, Luiken JJ, Schotten U, Verheule S. Atrial supply–demand balance in healthy adult pigs: coronary blood flow, oxygen extraction, and lactate production during acute atrial fibrillation. *Cardiovascular Research*. 2014; 101: 9-19.
46. Skalidis EI, Hamilos MI, Karalis IK, Chlouverakis G, Kochiadakis GE, Vardas PE. Isolated atrial microvascular dysfunction in patients with lone recurrent atrial fibrillation. *J Am Coll Cardiol*. 2008; 51: 2053-2057.
47. Barbey O, Pierre S, Duran MJ, Sennoune S, Lévy S, Maixent JM. Specific up-regulation of mitochondrial F0F1-ATPase activity after short episodes of atrial fibrillation in sheep. *J Cardiovasc Electrophysiol*. 2000; 11: 432-438.
48. Lenski M, Schleider G, Kohlhaas M et al. Arrhythmia causes lipid accumulation and reduced glucose uptake. *Basic Research in Cardiology*. 2015; 110:
49. Zhang LP, Hui B, Gao BR. High risk of sudden death associated with a PRKAG2-related familial Wolff-Parkinson-White syndrome. *J Electrocardiol*. 2011; 44: 483-486.
50. Ben Jehuda R, Eisen B, Shemer Y et al. CRISPR correction of the PRKAG2 gene mutation in the patient's induced pluripotent stem cell-derived cardiomyocytes eliminates electrophysiological and structural abnormalities. *Heart Rhythm*. 2018; 15: 267-276.
51. Leslie ND, Valencia CA, Strauss AW, Zhang K. Very Long-Chain Acyl-Coenzyme A Dehydrogenase Deficiency. In: Adam MP, Ardinger HH, Pagon RA et al., eds. *GeneReviews*. Seattle (WA): University of Washington, Seattle; 1993
52. Pagniez-Mammeri H, Loublier S, Legrand A, Bénit P, Rustin P, Slama A. Mitochondrial complex I deficiency of nuclear origin I. Structural genes. *Mol Genet Metab*. 2012; 105: 163-172.
53. Fullerton M, McFarland R, Taylor RW, Alston CL. The genetic basis of isolated mitochondrial complex II deficiency. *Mol Genet Metab*. 2020; 131: 53-65.
54. Fernández-Vizarra E, Zeviani M. Nuclear gene mutations as the cause of mitochondrial complex III deficiency. *Front Genet*. 2015; 6: 134.
55. Brischigliaro M, Zeviani M. Cytochrome c oxidase deficiency. *Biochim Biophys Acta Bioenerg*. 2021; 1862: 148335.

56. Stanley WC, Recchia FA, Lopaschuk GD. Myocardial substrate metabolism in the normal and failing heart. *Physiol Rev*. 2005; 85: 1093-1129.
57. Ausma J, Wijffels M, Thoné F, Wouters L, Allessie M, Borgers M. Structural changes of atrial myocardium due to sustained atrial fibrillation in the goat. *Circulation*. 1997; 96: 3157-3163.
58. Murphy MP. How mitochondria produce reactive oxygen species. *Biochem J*. 2009; 417: 1-13.
59. Crompton M, Virji S, Doyle V, Johnson N, Ward JM. The mitochondrial permeability transition pore. *Biochem Soc Symp*. 1999; 66: 167-179.
60. Aon MA, Cortassa S, Marbán E, O'Rourke B. Synchronized whole cell oscillations in mitochondrial metabolism triggered by a local release of reactive oxygen species in cardiac myocytes. *J Biol Chem*. 2003; 278: 44735-44744.
61. Zorov DB, Filburn CR, Klotz LO, Zweier JL, Sollott SJ. Reactive oxygen species (ROS)-induced ROS release: a new phenomenon accompanying induction of the mitochondrial permeability transition in cardiac myocytes. *J Exp Med*. 2000; 192: 1001-1014.
62. Xie W, Santulli G, Reiken SR et al. Mitochondrial oxidative stress promotes atrial fibrillation. *Sci Rep*. 2015; 5: 15.
63. Akar FG, Aon MA, Tomaselli GF, O'Rourke B. The mitochondrial origin of postischemic arrhythmias. *J Clin Invest*. 2005; 115: 3527-3535.
64. Cecatto C, Amaral AU, da Silva JC et al. Metabolite accumulation in VLCAD deficiency markedly disrupts mitochondrial bioenergetics and Ca^{2+} homeostasis in the heart. *FEBS J*. 2018; 285: 1437-1455.
65. Yang R, Ernst P, Song J et al. Mitochondrial-Mediated Oxidative Ca^{2+}/Calmodulin-Dependent Kinase II Activation Induces Early Afterdepolarizations in Guinea Pig Cardiomyocytes: An In Silico Study. *J Am Heart Assoc*. 2018; 7: e008939.
66. Dai DF, Rabinovitch PS. Cardiac aging in mice and humans: the role of mitochondrial oxidative stress. *Trends Cardiovasc Med*. 2009; 19: 213-220.
67. Youn JY, Zhang J, Zhang Y et al. Oxidative stress in atrial fibrillation: an emerging role of NADPH oxidase. *J Mol Cell Cardiol*. 2013; 62: 72-79.
68. Korantzopoulos P, Kolettis TM, Galaris D, Goudevenos JA. The role of oxidative stress in the pathogenesis and perpetuation of atrial fibrillation. *Int J Cardiol*. 2007; 115: 135-143.
69. Slane BG, Aykin-Burns N, Smith BJ et al. Mutation of succinate dehydrogenase subunit C results in increased O2.-, oxidative stress, and genomic instability. *Cancer Res*. 2006; 66: 7615-7620.
70. Sajnani K, Islam F, Smith RA, Gopalan V, Lam AK. Genetic alterations in Krebs cycle and its impact on cancer pathogenesis. *Biochimie*. 2017; 135: 164-172.
71. Ji W, Foo JN, O'Roak BJ et al. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet*. 2008; 40: 592-599.
72. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*. 2004; 305: 869-872.
73. Uhlen M, Fagerberg L, Hallstrom BM et al. Tissue-based map of the human proteome. *Science*. 2015; 347: 1260419-1260419.
74. Karczewski KJ, Francioli LC, Tiao G et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020; 581: 434-443.
75. Zhu Q, Ge D, Maia JM et al. A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *Am J Hum Genet*. 2011; 88: 458-468.

76. Tennessen JA, Bigham AW, O'Connor TD et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. 2012; 337: 64-69.
77. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006; 38: 904-909.
78. Richards S, Aziz N, Bale S et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015; 17: 405-424.
79. Guo MH, Plummer L, Chan Y-M, Hirschhorn JN, Lippincott MF. Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. *The American Journal of Human Genetics*. 2018; 103: 522-534.
80. Lott MT, Leipzig JN, Derbeneva O et al. mtDNA Variation and Analysis Using Mitomap and Mitomaster. *Curr Protoc Bioinformatics*. 2013; 44: 1.23.1-26.
81. Sudlow C, Gallacher J, Allen N et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015; 12: e1001779.
82. Bourgey M, Dali R, Eveleigh R et al. GenPipes: an open-source framework for distributed and scalable genomic analyses. *Gigascience*. 2019; 8:
83. Van der Auwera GA, Carneiro MO, Hartl C et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. 2013; 11.10.1-11.10.33.
84. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25: 1754-1760.
85. Li H, Handsaker B, Wysoker A et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009; 25: 2078-2079.
86. R Core Team. R: A language and environment for statistical computing. Accessed January 8, 2020. https://www.R-project.org/
87. McLaren W, Gil L, Hunt SE et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016; 17: 122.
88. Karolchik D, Hinrichs AS, Furey TS et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 2004; 32: D493-6.
89. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015; 348: 648-660.
90. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat*. 2011; 32: 894-899.
91. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine*. 2020; 12:
92. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res*. 2014; 42: 13534-13544.
93. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*. 2019; 47: D886-D894.
94. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS One*. 2012; 7: e46688.
95. Ioannidis NM, Rothstein JH, Pejaver V et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet*. 2016; 99: 877-885.

96. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 2003; 31: 3812-3814.
97. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics*. 2013; 14 Suppl 3: S3.
98. Castellana S, Fusilli C, Mazzoccoli G et al. High-confidence assessment of functional impact of human mitochondrial non-synonymous genome variations by APOGEE. *PLoS Comput Biol*. 2017; 13: e1005628.
99. Calabrese C, Simone D, Diroma MA et al. MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. *Bioinformatics*. 2014; 30: 3115-3117.
100. Lopes MC, Joyce C, Ritchie GR et al. A combined functional annotation score for non-synonymous variants. *Hum Hered*. 2012; 73: 47-51.
101. Purcell S, Neale B, Todd-Brown K et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007; 81: 559-575.
102. Lazarte J, Laksman ZW, Wang J et al. Enrichment of loss-of-function and copy number variants in ventricular cardiomyopathy genes in 'lone' atrial fibrillation. *EP Europace*. 2021;
103. Coordinate remapping service: NCBI. National Center for Biotechnology Information (NBCI). Accessed August 3, 2020. https://www.ncbi.nlm.nih.gov/
104. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*. 2010; 467: 1061-1073.
105. Lopaschuk GD, Belke DD, Gamble J, Itoi T, Schönekess BO. Regulation of fatty acid oxidation in the mammalian heart in health and disease. *Biochim Biophys Acta*. 1994; 1213: 263-276.
106. Repp BM, Mastantuono E, Alston CL et al. Clinical, biochemical and genetic spectrum of 70 patients with ACAD9 deficiency: is riboflavin supplementation effective. *Orphanet J Rare Dis*. 2018; 13: 120.
107. Hock DH, Robinson DRL, Stroud DA. Blackout in the powerhouse: clinical phenotypes associated with defects in the assembly of OXPHOS complexes and the mitoribosome. *Biochem J*. 2020; 477: 4085-4132.
108. Bates MG, Bourke JP, Giordano C, d'Amati G, Turnbull DM, Taylor RW. Cardiac involvement in mitochondrial DNA disease: clinical spectrum, diagnosis, and management. *Eur Heart J*. 2012; 33: 3023-3033.
109. Yamada K, Taketani T. Management and diagnosis of mitochondrial fatty acid oxidation disorders: focus on very-long-chain acyl-CoA dehydrogenase deficiency. *J Hum Genet*. 2019; 64: 73-85.
110. Spiekerkoetter U, Sun B, Khuchua Z, Bennett MJ, Strauss AW. Molecular and phenotypic heterogeneity in mitochondrial trifunctional protein deficiency due to beta-subunit mutations. *Hum Mutat*. 2003; 21: 598-607.
111. Aintablian HK, Narayanan V, Belnap N, Ramsey K, Grebe TA. An atypical presentation of ACAD9 deficiency: Diagnosis by whole exome sequencing broadens the phenotypic spectrum and alters treatment approach. *Mol Genet Metab Rep*. 2017; 10: 38-44.
112. Pedersen CB, Kølvraa S, Kølvraa A et al. The ACADS gene variation spectrum in 114 patients with short-chain acyl-CoA dehydrogenase (SCAD) deficiency is dominated by missense variations leading to protein misfolding at the cellular level. *Hum Genet*. 2008; 124: 43-56.

113. Andresen BS, Olpin S, Poorthuis BJ et al. Clear correlation of genotype with disease phenotype in very-long-chain acyl-CoA dehydrogenase deficiency. *Am J Hum Genet*. 1999; 64: 479-494.
114. Schiff M, Haberberger B, Xia C et al. Complex I assembly function and fatty acid oxidation enzyme activity of ACAD9 both contribute to disease severity in ACAD9 deficiency. *Hum Mol Genet*. 2015; 24: 3238-3247.
115. Kadoya T, Sakakibara A, Kitayama K et al. Successful treatment of infantile-onset ACAD9-related cardiomyopathy with a combination of sodium pyruvate, beta-blocker, and coenzyme Q10. *J Pediatr Endocrinol Metab*. 2019; 32: 1181-1185.
116. Knottnerus SJG, Mengarelli I, Wüst RCI et al. Electrophysiological Abnormalities in VLCAD Deficient hiPSC-Cardiomyocytes Can Be Improved by Lowering Accumulation of Fatty Acid Oxidation Intermediates. *Int J Mol Sci*. 2020; 21:
117. Seminotti B, Leipnitz G, Karunanidhi A et al. Mitochondrial energetics is impaired in very long-chain acyl-CoA dehydrogenase deficiency and can be rescued by treatment with mitochondria-targeted electron scavengers. *Human Molecular Genetics*. 2019; 28: 928-941.
118. Li F-Y, El-Hattab AW, Bawle EV et al. Molecular spectrum of SLC22A5 (OCTN2) gene mutations detected in 143 subjects evaluated for systemic carnitine deficiency. *Human Mutation*. 2010; 31: E1632-E1651.
119. Frigeni M, Balakrishnan B, Yin X et al. Functional and molecular studies in primary carnitine deficiency. *Hum Mutat*. 2017; 38: 1684-1699.
120. Longo N, Frigeni M, Pasquali M. Carnitine transport and fatty acid oxidation. *Biochim Biophys Acta*. 2016; 1863: 2422-2435.
121. Iuso A, Scacco S, Piccoli C et al. Dysfunctions of cellular oxidative metabolism in patients with mutations in the NDUFS1 and NDUFS4 genes of complex I. *J Biol Chem*. 2006; 281: 10374-10380.
122. Garnham JO, Roberts LD, Espino-Gonzalez E et al. Chronic heart failure with diabetes mellitus is characterized by a severe skeletal muscle pathology. *J Cachexia Sarcopenia Muscle*. 2020; 11: 394-404.
123. Grad LI, Lemire BD. Mitochondrial complex I mutations in Caenorhabditis elegans produce cytochrome c oxidase deficiency, oxidative stress and vitamin-responsive lactic acidosis. *Hum Mol Genet*. 2004; 13: 303-314.
124. Soto IC, Fontanesi F, Liu J, Barrientos A. Biogenesis and assembly of eukaryotic cytochrome c oxidase catalytic core. *Biochim Biophys Acta*. 2012; 1817: 883-897.
125. Fedida J, Fressart V, Charron P et al. Contribution of exome sequencing for genetic diagnostic in arrhythmogenic right ventricular cardiomyopathy/dysplasia. *PLoS One*. 2017; 12: e0181840.
126. Alfadhel M, Lillquist YP, Waters PJ et al. Infantile cardioencephalopathy due to a COX15 gene defect: report and review. *Am J Med Genet A*. 2011; 155A: 840-844.
127. Moffett JR, Puthillathu N, Vengilote R, Jaworski DM, Namboodiri AM. Acetate Revisited: A Key Biomolecule at the Nexus of Metabolism, Epigenetics, and Oncogenesis - Part 2: Acetate and ACSS2 in Health and Disease. *Front Physiol*. 2020; 11: 580171.
128. Martínez-Micaelo N, González-Abuín N, Ardévol A et al. Leptin signal transduction underlies the differential metabolic response of LEW and WKY rats to cafeteria diet. *J Mol Endocrinol*. 2016; 56: 1-10.
129. Wopereis S, Radonjic M, Rubingh C et al. Identification of prognostic and diagnostic biomarkers of glucose intolerance in ApoE3Leiden mice. *Physiol Genomics*. 2012; 44: 293-304.

130. Yamashita H, Fujisawa K, Ito E et al. Improvement of obesity and glucose tolerance by acetate in Type 2 diabetic Otsuka Long-Evans Tokushima Fatty (OLETF) rats. *Biosci Biotechnol Biochem*. 2007; 71: 1236-1243.
131. Dikalov SI, Dikalova AE. Crosstalk Between Mitochondrial Hyperacetylation and Oxidative Stress in Vascular Dysfunction and Hypertension. *Antioxidants & Redox Signaling*. 2019; 31: 710-721.
132. Fu J, Gaetani S, Oveisi F et al. Oleylethanolamide regulates feeding and body weight through activation of the nuclear receptor PPAR-α. *Nature*. 2003; 425: 90-93.
133. Leone TC, Weinheimer CJ, Kelly DP. A critical role for the peroxisome proliferator-activated receptor alpha (PPARalpha) in the cellular fasting response: the PPARalpha-null mouse as a model of fatty acid oxidation disorders. *Proc Natl Acad Sci U S A*. 1999; 96: 7473-7478.
134. Deplanque D, Gelé P, Pétrault O et al. Peroxisome proliferator-activated receptor-alpha activation as a mechanism of preventive neuroprotection induced by chronic fenofibrate treatment. *J Neurosci*. 2003; 23: 6264-6271.

Appendix A.

Gene panel

Gene symbol	Metabolic pathway	Molecular function	1	2	3	4
ABCD1	Lipid degradation	Translocase				
ACAA2	Fatty acid metabolism	Transferase	■			■
ACACB	Fatty acid biosynthesis	Ligase	■			■
ACAD10	Fatty acid metabolism	Oxidoreductase				
ACAD8	Fatty acid metabolism	Oxidoreductase	■			
ACAD9	Fatty acid metabolism	Oxidoreductase				
ACADL	Fatty acid metabolism	Oxidoreductase				■
ACADM	Fatty acid metabolism	Oxidoreductase				
ACADS	Fatty acid metabolism	Oxidoreductase	■	■	■	■
ACADSB	Fatty acid metabolism	Oxidoreductase				
ACADVL	Fatty acid metabolism	Oxidoreductase	■	■	■	■
ACAT1	Fatty acid metabolism	Transferase				
ACO2	Tricarboxylic acid cycle	Lyase				■
ACOX1	Fatty acid metabolism	Oxidoreductase				
ACSL1	Fatty acid metabolism	Ligase				
ACSL4	Fatty acid metabolism	Ligase				
ACSL5	Fatty acid metabolism	Ligase	■			
ACSL6	Fatty acid metabolism	Ligase				
ACSM1	Fatty acid metabolism	Ligase				
ACSM3	Fatty acid metabolism	Ligase				
ACSM4	Fatty acid metabolism	Ligase	■		■	■
ACSM5	Fatty acid metabolism	Ligase		■	■	
ACSS1	Acyl-CoA synthesis	Ligase				
ACSS2	Acyl-CoA synthesis	Ligase	■	■	■	
ACSS3	Acyl-CoA synthesis	Ligase	■	■	■	
AGL	Glycogen biosynthesis	Glycosidase	■	■	■	■
ALDH9A1	Carnitine synthesis/transport	Oxidoreductase				■
ALDOA	Glycolysis	Lyase				■
ALDOB	Glycolysis	Lyase	■			
ALDOC	Glycolysis	Lyase	■	■		■
ATP5F1A	ATP synthesis	Structural	■		■	■
ATP5F1B	ATP synthesis	Translocase				
ATP5F1C	ATP synthesis	Structural	■		■	
ATP5F1D	ATP synthesis	Structural				
ATP5F1E	ATP synthesis	Hydrolase				

Gene symbol	Metabolic pathway	Molecular function	1	2	3	4
ATP5IF1	ATP synthesis	Structural				
ATP5MC1	ATP synthesis	Structural				
ATP5MC2	ATP synthesis	Structural				
ATP5MC3	ATP synthesis	Structural				
ATP5MD	ATP synthesis	Structural				
ATP5ME	ATP synthesis	Structural				
ATP5MF	ATP synthesis	Structural				
ATP5MG	ATP synthesis	Structural				
ATP5MGL	ATP synthesis	Structural				
ATP5MPL	ATP synthesis	Structural				
ATP5PB	ATP synthesis	Structural				
ATP5PD	ATP synthesis	Structural				
ATP5PF	ATP synthesis	Structural				
ATP5PO	ATP synthesis	Structural				
ATP5S	ATP synthesis	Structural				
ATPAF1	ATP synthesis	Chaperone				■
ATPAF2	ATP synthesis	Chaperone	■			■
BCS1L	Electron transport chain	Chaperone				
BDH1	Lipid metabolism	Oxidoreductase				
BDH2	Lipid metabolism	Oxidoreductase				
BPGM	Glycolysis	Hydrolase				
CKB	Creatine-phosphate synthesis	Kinase				
CKM	Creatine-phosphate synthesis	Kinase				
CKMT2	Creatine-phosphate synthesis	Kinase				
COA1	Electron transport chain	Assembly factor				
COA3	Electron transport chain	Assembly factor	■	■	■	
COA6	Electron transport chain	Assembly factor				
COX10	Electron transport chain	Transferase	■		■	■
COX11	Electron transport chain	Assembly factor	■	■	■	■
COX14	Electron transport chain	Assembly factor				
COX15	Electron transport chain	Assembly factor				
COX16	Electron transport chain	Assembly factor				
COX17	Electron transport chain	Chaperone				
COX18	Electron transport chain	Assembly factor				
COX20	Electron transport chain	Assembly factor				
COX4I1	Electron transport chain	Oxidoreductase				
COX4I2	Electron transport chain	Oxidoreductase				
COX5A	Electron transport chain	Assembly factor				
COX5B	Electron transport chain	Assembly factor				
COX6A2	Electron transport chain	Oxidoreductase				
COX6B1	Electron transport chain	Assembly factor				

Gene symbol	Metabolic pathway	Molecular function	1	2	3	4
COX6C	Electron transport chain	Assembly factor				
COX7A1	Electron transport chain	Oxidoreductase				
COX7A2	Electron transport chain	Assembly factor				
COX7B	Electron transport chain	Assembly factor				
COX7C	Electron transport chain	Assembly factor				
COX8A	Electron transport chain	Assembly factor				
COX8C	Electron transport chain	Assembly factor				
CPT1B	Fatty acid metabolism	Acyltransferase				
CRLS1	Lipid synthesis	Transferase				
CS	Tricarboxylic acid cycle	Transferase				
CYBA	Electron transport chain	Oxidoreductase				
CYBB	Electron transport chain	Oxidoreductase				
CYC1	Electron transport chain	Translocase				
DAGLA	Lipid degradation	Hydrolase				
DAGLB	Lipid degradation	Hydrolase				
DECR1	Fatty acid metabolism	Oxidoreductase				
DECR2	Fatty acid metabolism	Oxidoreductase				
DLAT	Carbohydrate metabolism	Transferase				
DLD	Tricarboxylic acid cycle	Oxidoreductase				
DLST	Tricarboxylic acid cycle	Transferase				
ECH1	Fatty acid metabolism	Isomerase				
ECHDC1	Fatty acid metabolism	Lyase				
ECHDC2	Fatty acid metabolism	Lyase				
ECHDC3	Fatty acid metabolism	Lyase				
ECHS1	Fatty acid metabolism	Lyase				
ECI1	Fatty acid metabolism	Isomerase				
ECI2	Fatty acid metabolism	Isomerase				
ECSIT	Electron transport chain	Assembly factor				
EHHADH	Fatty acid metabolism	Isomerase				
ENO3	Glycolysis	Lyase				
ETFA	Lipid degradation	Transferase				
ETFB	Lipid degradation	Transferase				
ETFDH	Lipid degradation	Oxidoreductase				
FBP2	Carbohydrate metabolism	Hydrolase				
FH	Tricarboxylic acid cycle	Lyase				
FLAD1	FAD synthesis	Transferase				
G6PC3	Gluconeogenesis	Hydrolase				
GAPDH	Glycolysis	Oxidoreductase				
GBE1	Glycogen biosynthesis	Transferase				
GCK	Glycolysis	Kinase				
GK	Glycerol metabolism	Kinase				

Gene symbol	Metabolic pathway	Molecular function	1	2	3	4
GOT1	Amino-acid biosynthesis	Transferase				
GOT2	Lipid transport	Transferase				
GPI	Glycolysis	Isomerase	■			
GPT	Nitrogen metabolism	Transferase	■			■
GPX3	Antioxidant	Oxidoreductase				
GPX7	Antioxidant	Oxidoreductase				
GPX8	Antioxidant	Oxidoreductase				
GYG1	Glycogen biosynthesis	Transferase				
GYS1	Glycogen biosynthesis	Transferase	■		■	■
GYS2	Glycogen biosynthesis	Transferase				
H6PD	Carbohydrate metabolism	Hydrolase	■	■		■
HADH	Fatty acid metabolism	Oxidoreductase				■
HADHA	Fatty acid metabolism	Lyase				
HADHB	Fatty acid metabolism	Transferase	■		■	■
HK1	Glycolysis	Kinase				
HK2	Glycolysis	Kinase				
HK3	Glycolysis	Kinase	■	■		■
HMGCL	Ketogenesis	Lyase	■	■	■	
HMGCLL1	Ketogenesis	Lyase				■
HMGCS2	Cholesterol biosynthesis	Transferase				
HSD17B12	Lipid biosynthesis	Oxidoreductase				
HSD17B14	Lipid metabolism	Oxidoreductase				
HSD17B4	Fatty acid metabolism	Isomerase	■		■	■
IDH2	Tricarboxylic acid cycle	Oxidoreductase	■			■
LDHA	Glycolysis	Oxidoreductase				
LDHB	Glycolysis	Oxidoreductase				
LIAS	Lipoic acid synthesis	Transferase				
LIPE	Cholesterol metabolism	Hydrolase	■			■
LPL	Lipid degradation	Hydrolase	■			■
MB	Oxygen transport	Muscle protein				
MDH1	Tricarboxylic acid cycle	Oxidoreductase	■			■
MDH2	Tricarboxylic acid cycle	Oxidoreductase	■	■	■	■
MECR	Fatty acid biosynthesis	Oxidoreductase				
MGAM	Starch digestion	Glycosidase				■
MPI	D-mannose synthesis	Isomerase	■	■	■	
MT-ATP6	ATP synthesis	Accessory subunit				■
MT-ATP8	ATP synthesis	Accessory subunit	■			■
MT-CO1	Electron transport chain	Translocase				■
MT-CO2	Electron transport chain	Translocase				
MT-CO3	Electron transport chain	Translocase				
MT-CYB	Electron transport chain	Translocase	■		■	■

Gene symbol	Metabolic pathway	Molecular function	1	2	3	4
MT-ND1	Electron transport chain	Translocase				
MT-ND2	Electron transport chain	Translocase				
MT-ND3	Electron transport chain	Translocase				
MT-ND4	Electron transport chain	Translocase				
MT-ND4L	Electron transport chain	Translocase				
MT-ND5	Electron transport chain	Translocase				
MT-ND6	Electron transport chain	Translocase				
MUT	Lipid degradation	Isomerase				
MVD	Cholesterol biosynthesis	Lyase				
NCF1	Superoxide synthesis	Oxidase				
NCF2	Superoxide synthesis	Oxidase				
NCF4	Superoxide synthesis	Oxidase				
NDUFA1	Electron transport chain	Accessory subunit				
NDUFA10	Electron transport chain	Accessory subunit				
NDUFA11	Electron transport chain	Accessory subunit				
NDUFA12	Electron transport chain	Accessory subunit				
NDUFA13	Electron transport chain	Accessory subunit				
NDUFA2	Electron transport chain	Accessory subunit				
NDUFA3	Electron transport chain	Accessory subunit				
NDUFA4	Electron transport chain	Accessory subunit				
NDUFA5	Electron transport chain	Accessory subunit				
NDUFA6	Electron transport chain	Accessory subunit				
NDUFA7	Electron transport chain	Accessory subunit				
NDUFA8	Electron transport chain	Accessory subunit				
NDUFA9	Electron transport chain	Accessory subunit				
NDUFAB1	Electron transport chain	Accessory subunit				
NDUFAF1	Electron transport chain	Chaperone				
NDUFAF2	Electron transport chain	Chaperone				
NDUFAF3	Electron transport chain	Chaperone				
NDUFAF4	Electron transport chain	Transferase				
NDUFB1	Electron transport chain	Accessory subunit				
NDUFB10	Electron transport chain	Accessory subunit				
NDUFB11	Electron transport chain	Accessory subunit				
NDUFB2	Electron transport chain	Accessory subunit				
NDUFB3	Electron transport chain	Accessory subunit				
NDUFB4	Electron transport chain	Accessory subunit				
NDUFB5	Electron transport chain	Accessory subunit				
NDUFB6	Electron transport chain	Accessory subunit				
NDUFB7	Electron transport chain	Accessory subunit				
NDUFB8	Electron transport chain	Accessory subunit				
NDUFB9	Electron transport chain	Accessory subunit				

Gene symbol	Metabolic pathway	Molecular function	1	2	3	4
NDUFC1	Electron transport chain	Accessory subunit				
NDUFC2	Electron transport chain	Accessory subunit				
NDUFS1	Electron transport chain	Oxidoreductase				
NDUFS2	Electron transport chain	Oxidoreductase				
NDUFS3	Electron transport chain	Oxidoreductase				
NDUFS4	Electron transport chain	Accessory subunit				
NDUFS5	Electron transport chain	Accessory subunit				
NDUFS6	Electron transport chain	Accessory subunit				
NDUFS7	Electron transport chain	Oxidoreductase				
NDUFS8	Electron transport chain	Oxidoreductase				
NDUFV1	Electron transport chain	Oxidoreductase				
NDUFV2	Electron transport chain	Oxidoreductase				
NDUFV3	Electron transport chain	Accessory subunit				
NMNAT1	NAD biosynthesis	Transferase				
NMNAT2	NAD biosynthesis	Transferase				
NMNAT3	NAD biosynthesis	Transferase				
NOX4	Superoxide synthesis	Oxidoreductase				
OGDH	Glycolysis	Oxidoreductase				
OXSM	Fatty acid biosynthesis	Transferase				
PANK1	Coenzyme A biosynthesis	Kinase				
PANK4	Coenzyme A biosynthesis	Hydrolase				
PC	Gluconeogenesis	Ligase				
PCK1	Gluconeogenesis	Decarboxylase				
PDHA1	Tricarboxylic acid cycle	Oxidoreductase				
PDHB	Tricarboxylic acid cycle	Oxidoreductase				
PFKM	Glycolysis	Kinase				
PFKP	Glycolysis	Kinase				
PGAM1	Glycolysis	Hydrolase				
PGK1	Glycolysis	Kinase				
PGM1	Glycolysis	Isomerase				
PKM	Glycolysis	Kinase				
PMVK	Cholesterol biosynthesis	Kinase				
PNPLA2	Lipid degradation	Hydrolase				
PNPLA4	Lipid degradation	Hydrolase				
PNPLA8	Lipid degradation	Hydrolase				
PPARA	Transcription	Activator				
PPARD	Transcription	Activator				
PPARG	Transcription	Activator				
PPARGC1A	Transcription	Activator				
PRKAA2	Fatty acid biosynthesis	Kinase				
PRKAB2	Fatty acid biosynthesis	Regulatory subunit				

Gene symbol	Metabolic pathway	Molecular function	1	2	3	4
PRKAG1	Fatty acid biosynthesis	Regulatory subunit				
PRKAG2	Fatty acid biosynthesis	Regulatory subunit				
PRKAG3	Fatty acid biosynthesis	Regulatory subunit				
PYGB	Carbohydrate metabolism	Transferase				
PYGM	Carbohydrate metabolism	Transferase				
RFK	FAD synthesis	Kinase				
RPE	Carbohydrate metabolism	Isomerase				
RPIA	Pentose phosphate pathway	Isomerase				
SCO1	Electron transport chain	Chaperone				
SCO2	Electron transport chain	Chaperone				
SDHA	Electron transport chain	Oxidoreductase				
SDHB	Electron transport chain	Oxidoreductase				
SDHC	Electron transport chain	Structural				
SDHD	Electron transport chain	Structural				
SIRT2	Regulation	Transferase				
SIRT3	Regulation	Transferase				
SLC22A5	Carnitine synthesis/transport	Translocase				
SLC25A20	Carnitine synthesis/transport	Translocase				
SLC25A4	ADP/ATP transport	Translocase				
SLC27A1	Fatty acid metabolism	Translocase				
SLC27A6	Fatty acid metabolism	Translocase				
SLC2A4	Carbohydrate metabolism	Translocase				
SLC7A4	Amino-acid transport	Translocase				
SOD1	Antioxidant	Oxidoreductase				
SOD2	Antioxidant	Oxidoreductase				
SOD2	Antioxidant	Oxidoreductase				
SOD3	Antioxidant	Oxidoreductase				
SUCLA2	Tricarboxylic acid cycle	Ligase				
SUCLG1	Tricarboxylic acid cycle	Ligase				
SUCLG2	Tricarboxylic acid cycle	Ligase				
SURF1	Electron transport chain	Assembly factor				
TECR	Fatty acid biosynthesis	Oxidoreductase				
TECRL	Fatty acid biosynthesis	Oxidoreductase				
TKTL1	Pentose phosphate pathway	Transferase				
TMLHE	Carnitine synthesis/transport	Dioxygenase				
TPI1	Gluconeogenesis	Isomerase				
TPK1	Thiamine pyrophosphate synthesis	Kinase				
TSPO	Lipid transport	Receptor				
UQCR10	Electron transport chain	Accessory subunit				
UQCR11	Electron transport chain	Accessory subunit				
UQCRB	Electron transport chain	Accessory subunit				

Gene symbol	Metabolic pathway	Molecular function	1	2	3	4
UQCRC1	Electron transport chain	Accessory subunit				
UQCRC2	Electron transport chain	Accessory subunit				
UQCRFS1	Electron transport chain	Translocase				
UQCRH	Electron transport chain	Accessory subunit				
UQCRQ	Electron transport chain	Accessory subunit				
VPS9D1	Electron transport chain	Assembly factor				

Gene symbols, metabolic pathways, and molecular functions of genes in the candidate panel. Highlighted cells in the numbered columns show which genes were found to have an association in the early-onset AF patient group in each of the analyses, the entire cohort (1), self-identified Asian participants (2), self-identified white participants (3), and PCA-identified white participants (4).

Appendix B.

Genes associated with AF

Gene	Product	OR	95%CI	P
ACAD9	Acyl-CoA dehydrogenase family member 9	1256	2.4 - 1522	0.035
ACADS	Acyl-CoA dehydrogenase short chain	151	14 - 934	0.004
ACADVL	Acyl-CoA dehydrogenase very long chain	9.7	2.0 - 29	0.023
ACSM4	Acyl-CoA synthetase medium chain family member 4	94.2	1.9 - 981	0.039
ACSM5	Acyl-CoA synthetase medium chain family member 5	84.1	8.8 - 404	0.005
ACSS2	Acyl-CoA synthetase short chain family member 2	12.0	1.4 - 46	0.039
ACSS3	Acyl-CoA synthetase short chain family member 3	374	4.8 - 16384	0.024
AGL	Amylo-alpha-1, 6-glucosidase, 4-alpha-glucanotransferase	50.3	5.6 - 219	0.010
ATP5F1A	ATP synthase F1 subunit alpha	374	4.8 - 16384	0.024
ATP5F1C	ATP synthase F1 subunit gamma			0.016
COA3	Cytochrome c oxidase assembly factor 3	126	2.4 - 522	0.035
COX10	COX10, heme A:farnesyltransferase cytochrome c oxidase assembly factor	188	3.2 - 3850	0.031
COX15	COX15, cytochrome c oxidase assembly homolog	10.1	1.2 - 38	0.048
DECR2	2,4-dienoyl-CoA reductase 2	4.1	1.3 - 9.8	0.031
EHHADH	Enoyl-CoA hydratase and 3-hydroxyacyl CoA dehydrogenase	42.0	4.7 - 177	0.012
GYS1	Glycogen synthase 1	11.4	2.3 - 34	0.016
HADHB	Hydroxyacyl-CoA dehydrogenase trifunctional multienzyme complex subunit beta			0.016
HMGCL	3-hydroxymethyl-3-methylglutaryl-CoA lyase	75.3	1.6 - 689	0.043
HSD17B4	Hydroxysteroid 17-beta dehydrogenase 4	151	14 - 934	0.004
MDH2	Malate dehydrogenase 2	126	12 - 705	0.004
MPI	Mannose phosphate isomerase			0.016
MT-CYB	Mitochondrially encoded cytochrome b	195	10 - 10399	0.005
NDUFS1	NADH:ubiquinone oxidoreductase core subunit S1			0.016
NDUFV1	NADH:ubiquinone oxidoreductase core subunit V1	188	3.2 - 3850	0.031
NDUFV3	NADH:ubiquinone oxidoreductase subunit V3	75.3	1.6 - 689	0.043
PANK1	Pantothenate kinase 1	94.2	1.9 - 981	0.039
PC	Pyruvate carboxylase	38.0	9.8 - 106	0.001
PKM	Pyruvate kinase M1/2	188	3.2 - 3850	0.031
PYGB	Glycogen phosphorylase B	16.2	3.3 - 50	0.010
PYGM	Glycogen phosphorylase, muscle associated	4.0	1.3 - 9.4	0.035
SLC22A5	Solute carrier family 22 member 5	4.7	1.5 - 11	0.024

Genes that associated with AF in the PCA cohort. Some variants did not appear in gnomAD precluding OR calculation.

Appendix C.

Technical references

Reference data

Human reference genome: 1000 Genomes Project GRCh 37

Genome Aggregation Database (gnomAD) v2.1 (exomes only)

Genotype Tissue Expression Project (GTEx) Median gene-level TPM by tissue v1.19

Mitomap Retrieved Aug.13 2020 (unversioned)

UK Biobank Mar 2019 release of 50,000 WES data

Gene border coordinates

University of California Santa Cruz (UCSC) Table Browser

Nuclear deleteriousness scoring

Combined Annotation Dependent Depletion (CADD) v1.6

Protein Variation Effect Analyzer (PROVEAN) v1.1.3

Rare Exome Variant Ensemble Learner (REVEL) retrieved Mar 5 2020 (unversioned)

Sorting Intolerant From Tolerant (SIFT) included with PROVEAN

Variant Effect Scoring Tool (VEST) v4

Mitochondrial deleteriousness scoring

MitImpact v3.0.1 includes the following:

Pathogenicity Prediction Through Logistic Model Tree (APOGEE)

Combined Annotation Scoring Tool (CAROL)

MToolBox

Software

Ubuntu v18.04.4 LTS

Genome Analysis Toolkit (GATK) v4.1.7.0 (Docker)

Variant Effect Predictor (VEP) v100 (Docker)

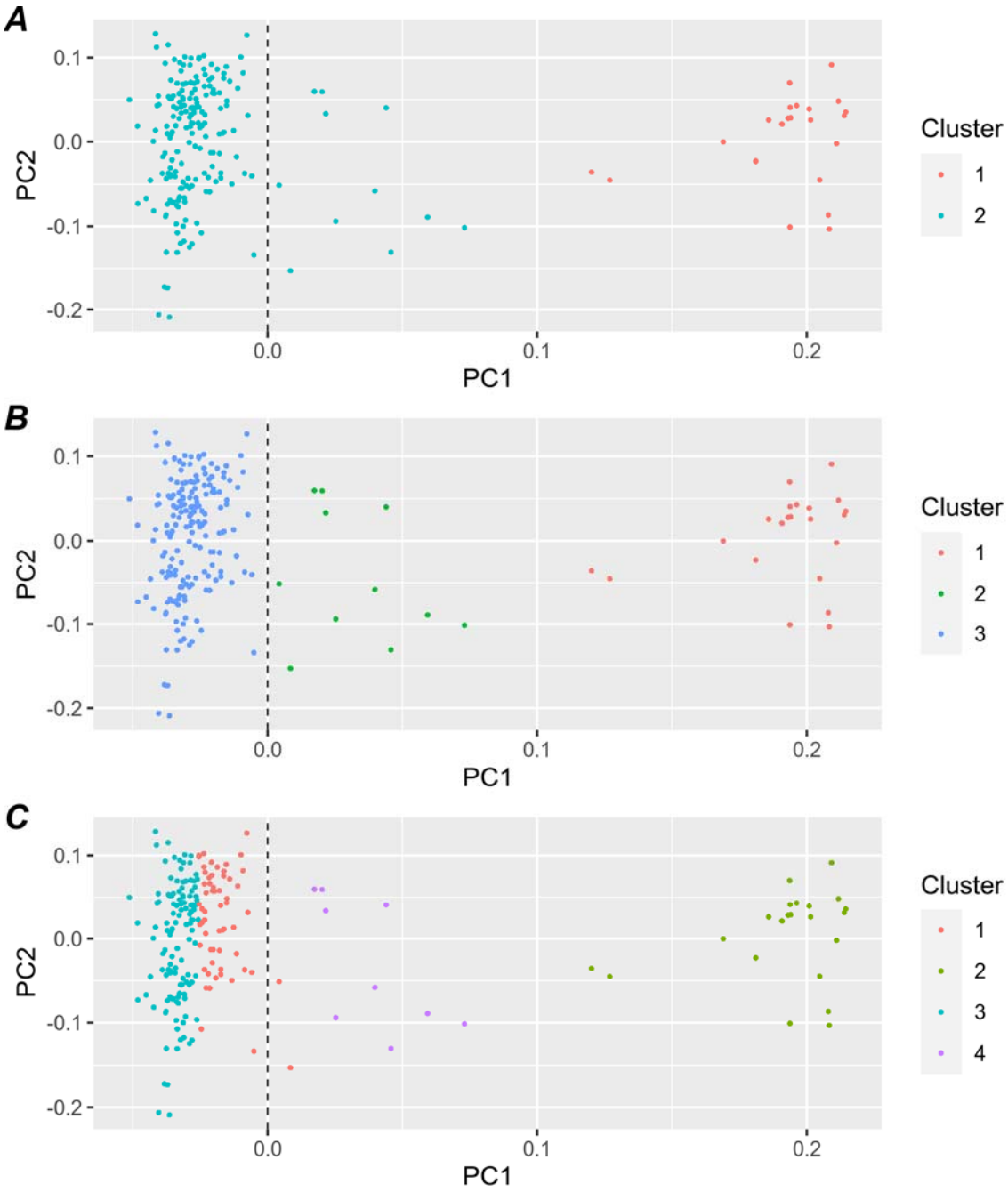
BCFtools v1.7 with htslib v1.9

VCF2bed v2.4.26

R v3.6.3 with Rstudio v1.2.5033

Appendix D.

K-means clustering



The k-means grouping of the early-onset AF cohort was compared using A. two clusters, B. three clusters, and C. four clusters.