FAMILIAL STUDIES IN

WHOLE EXOME AND GENOME SEQUENCING

Janice L. Farlow

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Medical and Molecular Genetics,
Indiana University

May 2015

Accepted by the Graduate Faculty, of Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____
Tatiana Foroud, Ph.D.
Chair

Doctoral Committee

_____
David Flockhart, M.D. Ph.D.

March 16, 2015

_____
Yunlong Liu, Ph.D.

_____
Quyen Hoang, Ph.D.

DEDICATION

To my family

for their boundless

love, understanding, and sacrifice

ACKNOWLEDGEMENTS

The first ‘mentor’ was Athena, the Greek goddess of wisdom, who assumed the appearance of Mentor, a friend of Odysseus, in order to guide Odysseus’ son through uncertain times. I have been fortunate to have many mentors, each of whom have guided me in one aspect of life or another, but I have truly been blessed to have Tatiana Foroud as a wise mentor not just for my research, but for navigating life as a whole. I have been forever inspired by her absolute devotion to her family, selfless contributions to others, commitment to scientific truth, and endless energy and perseverance. Although my sentences will always be too long for her liking, I hope that she realizes how pervasive her impact on my career and other aspirations in life has been.

Along with Tatiana, I would like to thank my committee members David Flockhart, Yunlong Liu, and Quyen Hoang. They have all always been willing and available for guidance on this work, from important minute details to critical review of the purpose and direction of this research. I owe many thanks to the many members of the Foroud group, especially Dongbing Lai, Leah Wetherill, and Dan Koller. These individuals and many more have each spent much time with me coaching me through programming, statistics, patient information databases, and so much more.

Many have proclaimed collaborative science as the way of the future, and I have been fortunate to be involved in collaborations from the start of this journey. I would like to thank in particular the team at the Center for Inherited Disease Research at Johns Hopkins (Hua Ling, Kurt Hetrick, Elizabeth Pugh, and Kim Doheny among others), who have patiently answered my questions, suggested novel directions, and been steadfast partners in all of this work. My thanks also go to Joe Broderick from the University of Cincinnati and the many collaborators involved in the Familial Intracranial Aneurysm study, who ventured with me into this first foray with whole exome sequencing data. I would also like to thank those at the HudsonAlpha Institute for Biotechnology and Baylor College of Medicine for their contributions to the Parkinson disease work, as well as those who have and will continue to devote their efforts to the research on X-linked ataxia dementia.

Of course none of this work would be possible without the many patients and families involved in each of these studies. I count myself most fortunate to have their trust and their stories as a lifetime of inspiration.

I also need to thank the Indiana University Medical Scientist Training Program (MSTP) family – directors Maureen Harrington and Raghu Mirmira, Jan Receveur, and all of the other students who have served as companions through this adventure. My thanks also go to Wade Clapp, veteran of the MSTP, who brought me to the Indiana University School of Medicine and connected me with

Tatiana. The support of the MSTP program, as well as funding from the Clinical Translational Sciences Institute, the National Institutes of Health and their sequencing contracts, and the National Center for Genome Analysis Support, was critical for this work.

Lastly, I would like to express my unending gratitude to my family. To my parents William and Jane, thank you for a lifetime of support and inspiration, for which you can never be fully repaid. To Grace and Grant, how in the world did we end up on these divergent yet parallel tracks? Thank you for all of our random conversations and last minute review of my writing. Aunt Fen, your support and love of Layla has been and continues to be an incredible blessing. I am indebted to my in-laws Marty and Diane, who have always been incredibly engaged and supportive. Erin, thank you for taking me in under your roof (dogs, baby, friends and all). You have truly been an amazing ‘big sister’ and friend. And to the other sister wife Gaury, thank you for doing more than any good friend should ever have to do over the years!

And to Nate, Layla, and our little one to be… words cannot express how much you have taught me over these few years. Thank you for always being there, for all the memories, and for all the adventures to come!

Janice L. Farlow

FAMILY STUDIES IN WHOLE EXOME AND GENOME SEQUENCING

Population genetics has been revolutionized by the advent of high-throughput sequencing (HTS) methods in the 21st century. Modern day sequencers are now capable of sequencing entire exomes and genomes at unprecedented speed and accuracy. An explosion of bioinformatics software and data analysis tools now makes sequencing accessible for gene discovery in both rare Mendelian and complex disease. Family-based sequencing studies in particular have great potential for elucidating the genetic basis for many more diseases.

We apply both whole exome and genome sequencing to three different cases of familial disease: intracranial aneurysm (IA), Parkinson disease (PD), and X-linked ataxia dementia (XLAD). IA and PD are both common, complex traits that inflict a devastating disease burden worldwide, mostly due to few effective therapeutic interventions. Little of the heritability of both IA and PD has been explained to date, especially as it relates to the impact of rare variation on disease. XLAD is an extremely rare neurological disease described thus far in one kindred. Although promising results have been achieved through previous genetic study designs, the causative gene has not yet been identified. For all three diseases, HTS offers an opportunity to explore the role of rare variation in disease pathogenesis. In each study, we explore the opportunities and challenges of family-based HTS for different disease models. The work

presented herein contributes effective practices for study design, analysis, and interpretation in a rapidly growing field still replete with questions about how best to implement HTS in studying familial disease.

Tatiana Foroud, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AAA	Abdominal aortic aneurysm
BWA	Burrows Wheeler Aligner
CADD	Combined Annotation Dependent Depletion
CIDR	Center for Inherited Disease Research
CLIA	Clinical Laboratory Improvement Amendments
CNV	Copy number variant
dbGaP	Database of Genotypes and Phenotypes
dbSNP	Database of Single Nucleotide Variants
ESP	Exome Sequencing Project
ExAC	Exome Aggregation Consortium
FDR	False discovery rate
GATK	Genome Analysis Toolkit
GO	Gene Ontology
GWA	Genome wide association
HGSC	Human Genome Sequencing Center
HTS	High-throughput sequencing
IA	Intracranial aneurysm
IGV	Integrated Genomics Viewer
Indel	Insertion-deletion
IRB	Institutional Review Board
MAF	Minor allele frequency
MRA	Magnetic resonance angiography
NCBI	National Center for Biotechnology Information
NVPD	Non-verified Parkinson disease
PCR	Polymerase chain reaction
PD	Parkinson disease
PSG	Parkinson Study Group
RVIS	Residual variation intolerance score
SAH	Subarachnoid hemorrhage
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
UKPDBB	United Kingdom Parkinson Disease Brain Bank
UTR	Untranslated region
VPD	Verified Parkinson disease
VQSR	Variant Quality Score Recalibration
VUS	Variant of unknown significance
WES	Whole exome sequencing
WGS	Whole genome sequencing
XLAD	X-linked ataxia dementia

Some of the text in this dissertation has published in *PLoS One*. Farlow et al. Lessons learned from whole exome sequencing in multiplex families affected by a complex genetic disorder, intracranial aneurysm. *PLoS One.* 2015.

INTRODUCTION

A brief history of population genetics

Since the days of Hippocrates, collecting and analyzing a family's history of disease has been an integral part of the practice of medicine. Physicians knew that a family medical history can indicate a higher risk for certain diseases and conditions, even if the causative mechanism for the risk was unknown. It was not until the 1800's, when Gregor Mendel completed his meticulous breeding experiments with sweet peas, that the laws of hereditary genetics began to be defined. When his work was rediscovered a century later, modern genetics was born, and scientists quickly began to investigate Mendelism as it applied to human families with disease.

In the middle of the 20th century, principles of genetic linkage and recombination were explored, leading to the ability to map chromosomes. This enabled scientists to conduct linkage studies, in which genetic regions harboring causative mutations could be mapped by observing the segregation patterns of disease with the inheritance of genetic markers. Linkage analysis was very successful in mapping a number of monogenic diseases, including cystic fibrosis[1-4] and Huntington's disease.[5] This methodology, however, is not successful at exploring all disease traits. For instance, large multiplex families required for robust linkage signals sometimes do not exist for extremely rare diseases that are fatal well before child-bearing years. Additionally, linkage

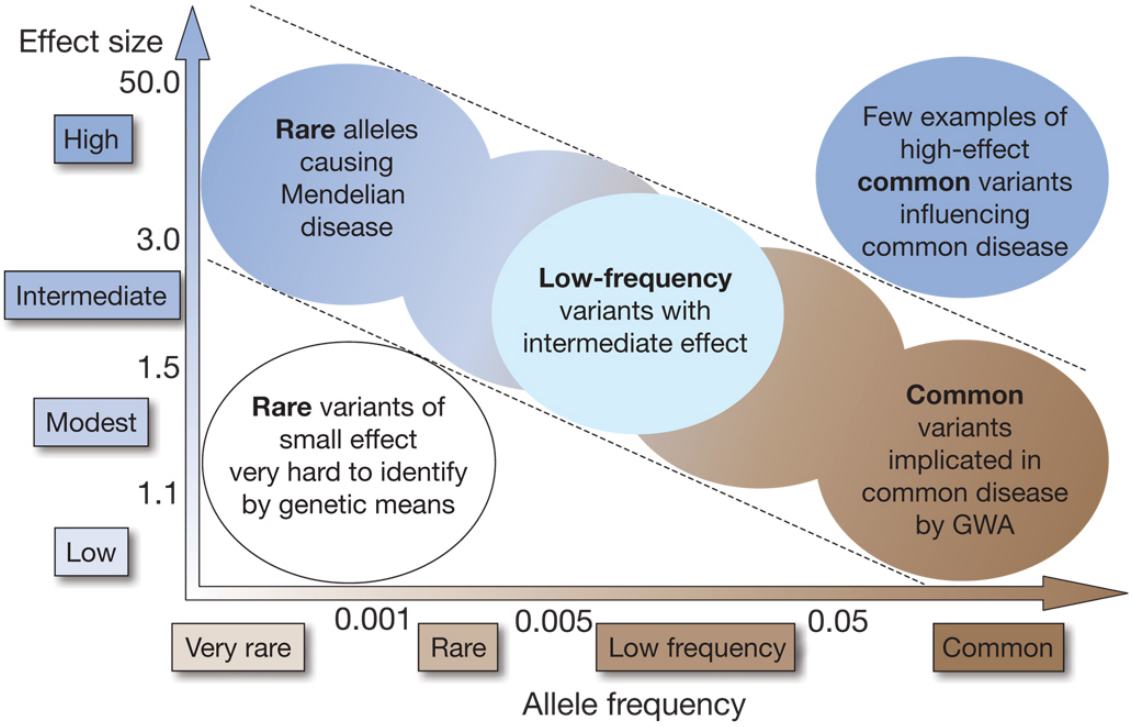
studies conducted for complex traits, such as diabetes and schizophrenia, sometimes seemed to conflict with previous findings for the same trait.[6-8] The heterogeneity of complex traits, which are caused by a combination of multiple genetic and/or environmental factors, limits the statistical power of linkage analysis in families. Furthermore, for those linkage studies that do reach statistical significance, the implicated interval may contain hundreds of genes, requiring intensive molecular work to identify the causative gene.

Candidate gene studies constitute another approach for studying human disease. After forming a hypothesis that a particular candidate gene is involved in the underlying pathophysiology of a disease, researchers can statistically test whether a particular allele of the gene is more frequently observed in cases than controls. Thus, population-based samples can be used instead of the unique families required for robust linkage mapping. Additionally, greater statistical power to detect genes of small effect sizes can be obtained through candidate gene association tests than through linkage analysis.[9] While most candidate gene studies have focused on common variation, the most notable consistent finding of which has been the *APOE* association with Alzheimer's disease,[10] some studies have also been utilized to explore rare variation.[11,12] Despite some promising findings, candidate gene approaches have faced criticisms for poor replication of findings[13-15] and for the significant limitation of requiring *a priori* knowledge about disease mechanisms.

With the development of genome-wide single nucleotide polymorphism (SNP) arrays, unbiased genome-wide association (GWA) studies became possible. GWA studies, which became popular starting in 2005, began to elucidate many common variants of smaller effect size important in complex diseases.[16] While many GWA studies have focused on large cohorts of unrelated individuals, some have also explored familial diseases.[17] Many GWA studies, however, suffer from the inability to narrow down an associated genetic region to a causative mutation, and the necessity to gather large samples numbering in the tens of thousands presents challenges for many disease models.

Findings from candidate gene, linkage, and GWA studies together have identified the genetic basis for a small percentage of diseases, and for some traits, they have accounted for only a portion of the estimated heritability of the disease. In particular, linkage studies have been able to explore rare variants with large effect sizes (Figure 1). GWA studies, on the other hand, are better powered to find common variation, or variants found in at least 5% of the general population, with lower effect sizes. Thus, researchers have suggested that some of the remaining heritability lies with rare variation of moderate to high effect sizes.[18-22] Such variation cannot be explored effectively using historical population genetics approaches, but this all changed with the genesis of high-throughput sequencing (HTS).

Figure 1. Genetic variant frequencies and effect sizes. Manolio et al, 2009.[23]



Advent of high-throughput sequencing

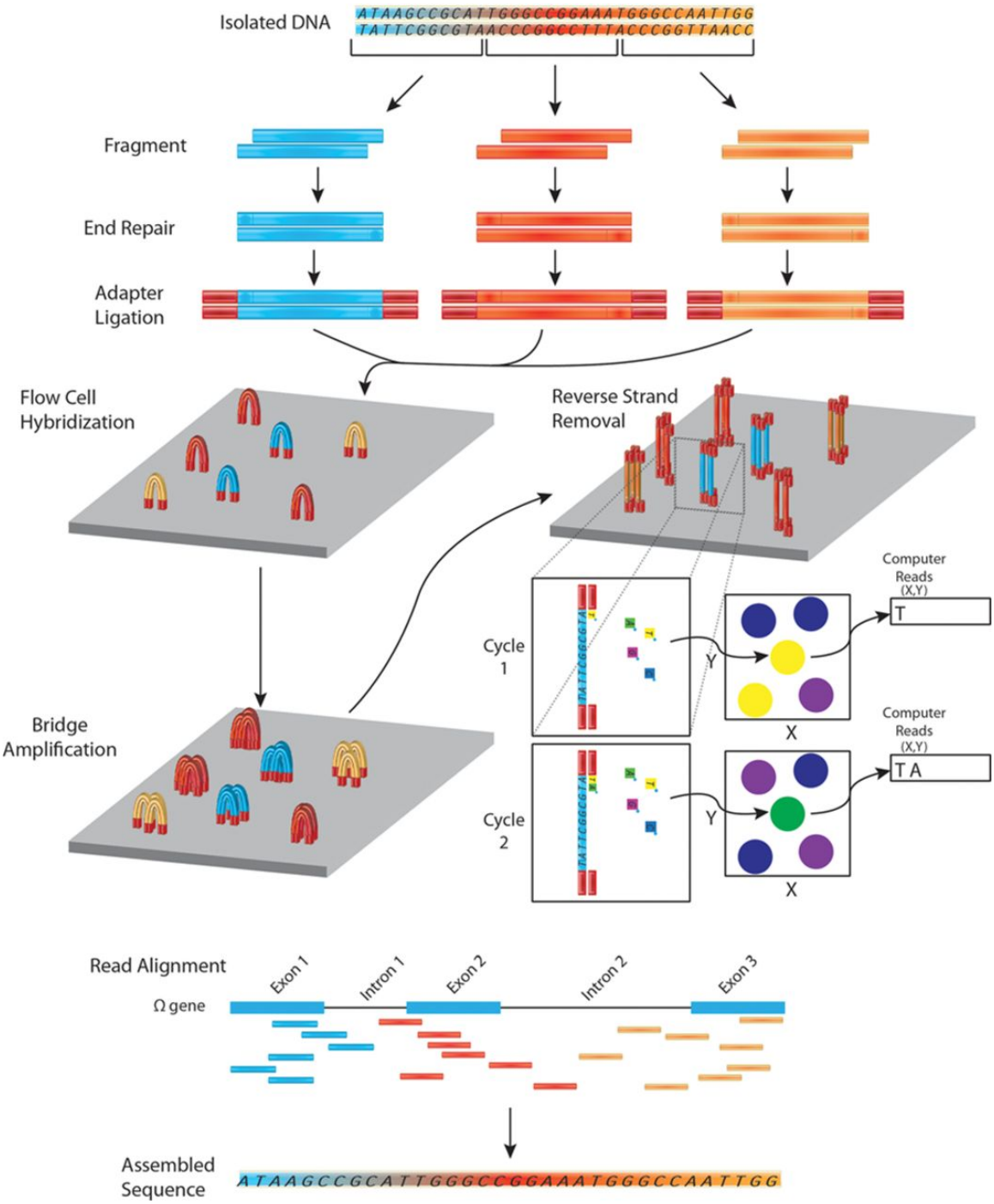
In the 1970's, researchers ascertained the first DNA sequences, paving the way for the sequencing of the first genome in 1977.[24] The Sanger sequencing method[24,25] that made this advance possible remained the primary way to sequence DNA until the 21st century. In this method, a labeled primer is annealed to a known segment of DNA juxtaposed to the unknown sequence of interest. Catalytic polymerization reactions then occur in four different tubes, each containing a different nucleotide, until the random addition of a specially-labeled chain-terminator nucleotide. The separation of the resultant fragments of DNA on a polyacrylamide gel by fragment size then allows the researcher to determine the nucleotide sequence. This basic technique with modifications, coupled with

the development of the polymerase chain reaction (PCR) in the 1980's, eventually led to the sequencing of the first human genome in 2001.[26]

In 2004, HTS was introduced as 'next-generation' sequencing. HTS, which relies on parallel sequencing of millions of small stretches of DNA, exponentially increased the number of bases that could be sequenced given a finite cost and timeframe. Competition and technological advances drove the rapid evolution of commercially available sequencers. At the time of the work presented herein, Illumina products dominated much of the HTS market. Like traditional Sanger sequencing, this technology (Figure 2) first sonicates genomic DNA into small fragments, which are ligated to adapters. After hybridization to a flow cell, the fragments are amplified creating clusters of fragments with the same nucleotide sequence. The complementary strand of DNA is removed, and the sequencer then adds fluorescently-labeled nucleotides sequentially and visually records the resulting fluorescence. The produced image is then reverted to a short string of nucleotide sequence, referred to as a sequencing read. After the sequencing run is complete, these reads and accompanying quality metrics can then be processed using a number of bioinformatics methods, which are described later.

Figure 2. High-throughput DNA sequencing on the Illumina platform.

Churko et al, 2013.[27]



Combined with the development of bioinformatics methods to leverage the reference genome sequence, HTS became feasible for researchers around the world. Techniques were developed not only for sequencing and detecting variation for whole genomes, but also for targeted parts of the genome and for more complex genetic features. Whole exome sequencing (WES), for instance, provides sequence data for just the exome, or the coding portion of the genome. Although the exome only comprises 1% of the genome, most mutations associated with Mendelian disease to date have been found in the coding portion of the genome.[28-31] Thus, many researchers interested in studying sequence variation in disease have opted for the sequencing of more individuals using the more cost-effective WES approach instead of WGS.[32] The transcriptome, or the RNA produced from DNA, can also be sequenced through HTS. Variation in the transcriptome can provide clues about alternative splicing, differential expression, gene fusion events, and functional non-coding RNAs. A wide variety of epigenomic data can also be obtained through HTS, allowing researchers an unprecedented genome-wide perspective on effects and mechanisms of transcriptional control. A sample of possible applications of HTS is listed in Table 1.

Table 1. **Applications of high-throughput sequencing**. Adapted from Shendure et al, 2012.[33] FAIRE = formaldehyde-assisted isolation of regulatory elements; MAINE = MNase-assisted isolation of nucleosomes; ChIP = chromatin immunoprecipitation; RIP = RNA-binding protein immunoprecipitation; CLIP = cross-linking immunoprecipitation; HITS = high-throughput sequencing of RNA isolated by CLIP; ChIA-PET = chromatin interaction analysis paired-end tag

Method	**Feature Examined**
DNA-Seq	Genome
Targeted DNA-Seq	Subset of a genome, e.g. whole exome sequencing
Methyl-Seq	Sites of DNA methylation
DNase-Seq, Sono-Seq, FAIRE-Seq	Active regulatory chromatin
MAINE-Seq	Histone-bound DNA
ChIP-Seq	Protein-DNA interactions
RIP-Seq, CLIP-Seq, HITS-CLIP	Protein-RNA interactions
RNA-Seq	Transcriptome
Hi-C	Three-dimensional genomic structure
ChIA-PET	Long-range interactions mediated by a protein
Ribo-Seq	Ribosome-protected mRNA fragments (mRNA under active translation)

In 2009, two landmark reports of the use of WES for gene discovery were published.[34,35] In the first study, WES was applied to 10 unrelated individuals affected by Kabuki syndrome, a rare multi-system disorder that has only been reported in about 400 cases worldwide.[34] A careful review of the variants identified and comparison of them to public databases and control exomes eventually led to the identification of *MLL2* as the causative gene. In the second study, the exomes of two siblings and two unrelated individuals affected by Miller syndrome were sequenced.[35] A similar method of retaining only rare variants, coupled with application of different inheritance models eventually singled out *DHODH* as the culpable gene. In both of these cases, Ng and colleagues were able to identify not only the genetic region linked with a disease, but the exact variation causing the disease as well. This was accomplished using only a few individuals, unlike the large multi-generational pedigrees required for linkage analysis or the thousands of individuals necessary for statistically robust GWA studies.

Within the next few years, the costs for sequencing and the storage of large datasets fell rapidly, and large collaborative efforts like the 1000 Genomes Project[36] produced resources that enabled small and large research efforts alike to conduct their own HTS experiments. To date, the genetic basis of over 100 rare Mendelian[30] and complex diseases[37] has been discovered using WES and whole genome sequencing (WGS).[38]

A general framework for analysis of sequencing studies

The bioinformatics community has developed thousands of techniques to make these discoveries possible. After sample preparation and sequencing (Figure 2) the numerous steps of bioinformatics processing can largely be boiled down to the general categories of alignment and variant detection, with measures taken for assessing quality throughout the process.

Alignment consists of matching each read to the reference genome. A series of steps are taken to ensure that excess mismatches at particular bases, within reads, and within regions are reviewed; duplicate reads are removed; and recalibration is performed for insertion/deletions (indels) and other types of structural variation. Although Burrows-Wheeler Aligner (BWA) is often the aligner of choice and is used throughout the present work, dozens of alignment programs exist and can be employed for different kinds of scenarios. Additionally, some software also permits *de novo* assembly, where a reference genome is not used in aligning the sequence. Such programs thus can be used to identify novel sequence or in species where a reference genome does not exist.

A number of programs exist for the next stage in data processing, or variant calling. The most frequently used programs currently can robustly detect single nucleotide variants (SNVs) and small indels. Some of the newer programs designed to identify larger indels and other types of structural variation are still being vetted by the bioinformatics community. Variant detection software typically

looks at the sequencing reads that span each position (or a stretch of bases for structural variation) to identify whether nucleotide calls for some or all of the reads differ from the reference genome. Some programs also examine haplotypes, or a collection of linked alleles, to increase the quality of variant calls; such programs are generally helpful for most sequencing studies but may not be the variant caller of choice when looking at extremely rare variants. Many programs recalibrate their variant calls based on a number of factors, sometimes dictated by the algorithm employed and other times chosen through machine learning approaches.

Throughout both the alignment and variant detection steps, various methods of quality control exist. General quality metrics such as sequencing depth and percentage of bases covered at particular depths can help determine if samples need to be re-sequenced. The number of different types of variants per sample, as well as statistics such as the transition to transversion ratio and the percentage of variants previously identified, can be compared to normally expected numbers and ratios. Even with a number of different quality metrics, typically the alignment and quality sequencing reads of variants of interest should be reviewed using inspection software, and many groups validate the variants using genotyping arrays or the traditional Sanger sequencing method. Some basic quality control measures are reviewed by Do and colleagues.[39]

After variant detection and quality control measures, a typical WES experiment will generate over 20,000 exonic SNVs per individual.[30] Of these variants, 10,000-12,500 will typically be synonymous variants, 9,500-12,000 are expected to be missense, and 100-200 will be splice altering or stop variants.[39] WGS, on the other hand, will typically yield over 3 million SNVs per individual.[40] A typical healthy person's genome harbors about 100 genuine loss of function variants, most of which represent heterozygous variants in nonessential genes, with only about 20 variants inactivating a gene's function entirely.[41]

Given the large number of variants identified in WES and WGS, several methods can winnow down the number of variants to key candidate variants. Many studies, such as the first WES studies[34,35] and the work presented herein, use filtering strategies based on hypotheses about the characteristics of causative variants expected. Larger studies, especially those without familial samples, have opted for statistical association tests, either designed for single variants or clusters of variants.

As a first step in both filtering strategies and association analysis, sequencing studies typically apply annotation and *in silico* prediction programs. Through annotation software, researchers are able to assign putative function (e.g. location within or outside of an exon, name of the nearest gene, effect on the mRNA or protein sequence if any, variant frequencies, etc.) to variants. *In silico* prediction programs help researchers determine the impact of a particular

variant. Some *in silico* programs measure the level of evolutionary conservation at a locus (where a more conserved location might imply that the locus has an important biological function, as deleterious mutations undergo purifying selection). Others focus on the variant's effect on the structure, enzymatic function, or other important characteristics of the protein. A third class combines both conservation and protein effect predictions, and some even compute and aggregate predictions from multiple other *in silico* programs.

Statistical association tests that can be applied to sequencing data are being developed at a rapid pace. Most of the available programs allow for testing the association of a single variant with the trait of interest, but these tests are generally not well powered for rare variants due to their infrequent observation in the dataset. In order to increase their power, most rare variant association tests combine rare variants in some fashion. The most popular grouping of variants to date has been by gene, although one might feasibly also look at variants across a group of genes or a pathway. Two major types of these collapsing or aggregative association tests exist. One version tabulates the number of rare alleles in a gene for each sample and compares the general 'burden' of rare alleles in cases versus controls. The other type, termed as a variance-component test, allows for variants to have either deleterious or protective effects by comparing the number of variants with non-zero effect sizes to expected scenarios. Although these tests have not been employed in the work presented herein, future studies will undoubtedly use later generations of these programs. A

recent review by Lee and colleagues lists some of the specific programs and considerations for each category of test.[42]

The bioinformatics programs used in the present work are listed in Table 2.

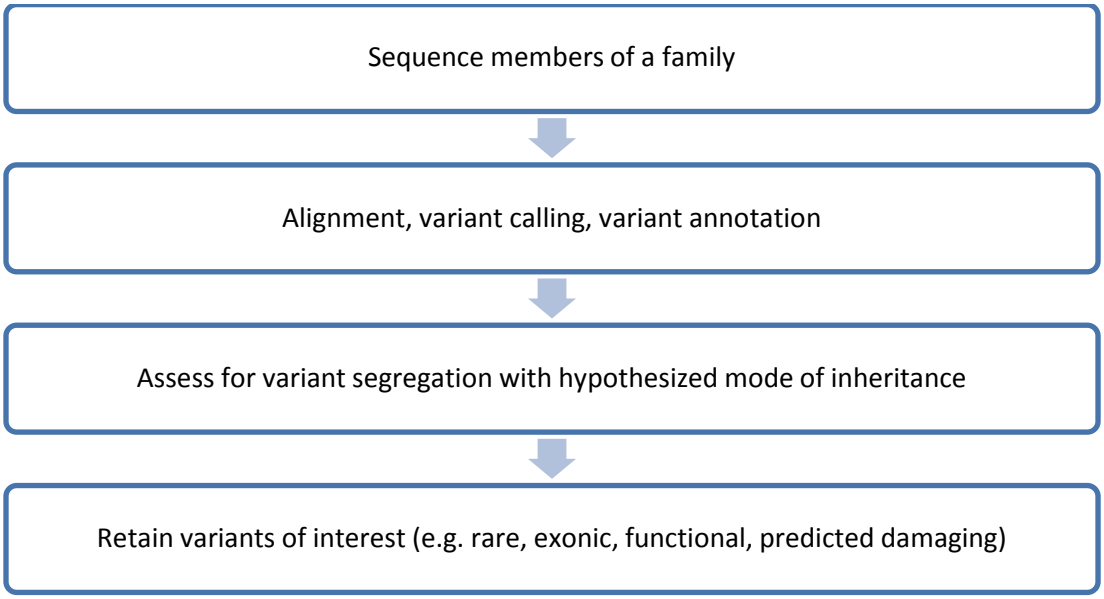
Table 2. **Bioinformatics programs utilized.** SNV = single nucleotide variant; indel = insertion/deletion

Program category	Program Name	Description
Alignment	Burrows Wheeler Aligner (BWA)[43]	Common aligner used for its general speed and accuracy
Variant detection	Genome Analysis Toolkit (GATK)[44] 1. Unified Genotyper 2. Haplotype Caller	Common variant caller used mostly for SNVs and indels; also recalibrates alignments
	SAMTools[45]	Common variant caller used mostly for SNVs and indels
In silico prediction	Combined Annotation Dependent Depletion (CADD)[46]	Scores the relative deleteriousness of a variant based on a number of other *in silico* prediction programs
	DDIG-in[47]	Predicts locus conservation for non-frameshifting indels
	GERP[48]	Predicts locus conservation
	MutPred[49]	Predicts locus conservation and variant impact on resultant protein
	PolyPhen[50]	Predicts variant impact on resultant protein
	Residual Variation Intolerance Score (RVIS)[51]	Computes a relative score for how well a gene tolerates mutation
	SIFT[52]	Predicts locus conservation for SNVs
	SIFT-Indel[53]	Predicts locus conservation for frameshifting indels
Other	Picard (http://picard.sourceforge.net/)	Used in this work for removal of duplicate reads
	ANNOVAR[54]	Provides comprehensive annotation features
	Integrated Genomics Viewer (IGV)[55]	Manual inspection of read alignments and variant calls
	Merlin[56]	Performs rapid linkage analysis

Sequencing applied to families

Family-based studies have formed the foundation for WES and WGS approaches. Many initial studies focused on rare Mendelian diseases in families where initial linkage analysis had yielded a promising genetic interval, but a specific gene had not been identified. Instead of using the laborious Sanger sequencing method to examine all the large number of genes in these intervals, researchers could now utilize WES or WGS to quickly identify the causative mutation in the linkage families.[57-61] Other studies applied WES or WGS to pedigrees that were uninformative for linkage because there were too few affected members or meioses, such as the Miller Syndrome case[35] and the more recent studies of *de novo* germline mutations using trios.[62-64] The typical workflow for a family-based sequencing study is depicted in Figure 3.

Figure 3. Schematic workflow of family-based sequencing studies.



More recently, WES and WGS have been applied to more common and complex diseases. Some multi-institutional collaborations have leveraged large cohorts of unrelated individuals, leading to discoveries of rare genetic risk factors involved in diabetes,[65] cardiovascular phenotypes,[66] and other traits. Other groups have continued using families to explore complex traits, either by including only unrelated familial samples or by sequencing multiple members of families per study. Some of the techniques developed in family-based sequencing studies, such as incorporating identity-by-descent (IBD) into quality control processes,[67] have also aided in the application of HTS to sporadic disease as well.

Challenges in sequencing studies

Despite the early successes of sequencing applied to familial disease and the resulting number of research groups focused in this area, a number of challenges remain in the field. Some of the 'rate limiters,' as described by Shendure et al, include the cost and effort of acquiring and storing samples, constructing libraries, sustaining technological infrastructure, and maintaining labor and expertise.[33] General computational and bioinformatics challenges and potential solutions are also discussed by Berger and colleagues.[68]

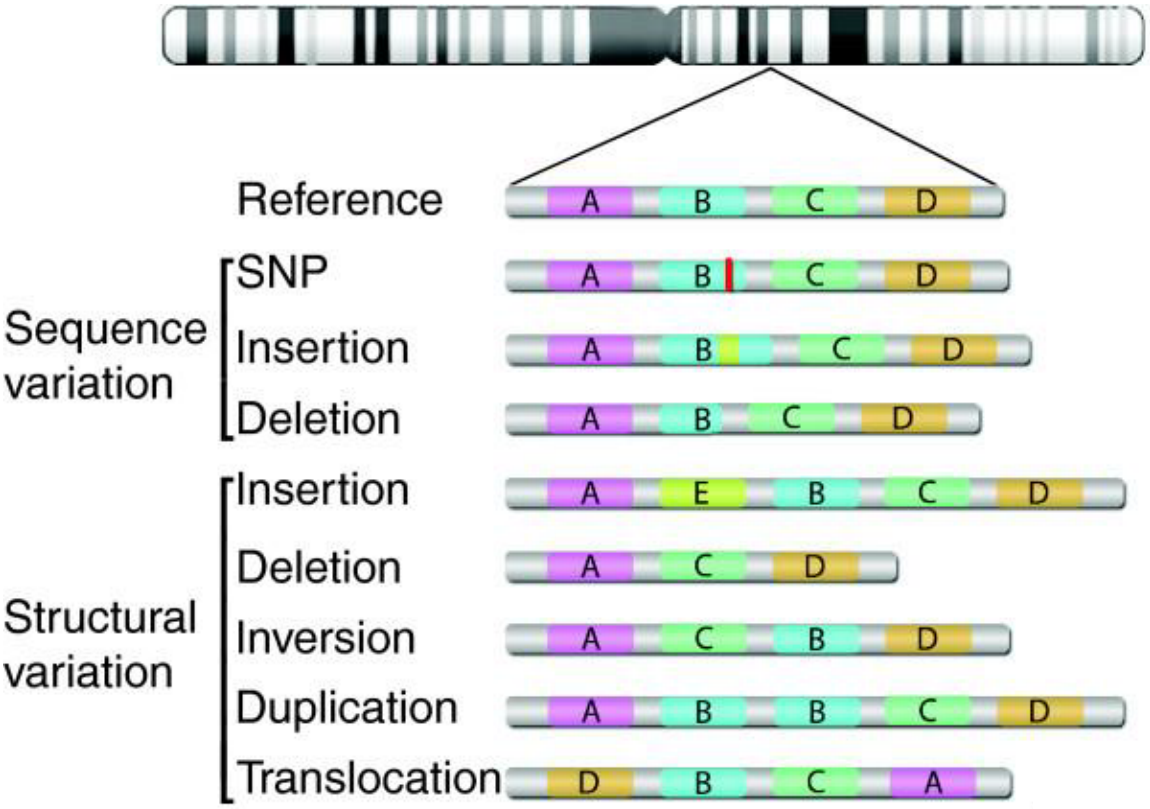
The interpretation of results from sequencing studies offers another multifaceted challenge. Variants of unknown significance (VUS), or variants whose biochemical and/or clinical significance has not been identified or confirmed, often are the sole products of sequencing experiments. *In silico* prediction

programs can highlight particular variants of interest or provide ideas on how to molecularly characterize the variants identified, but they serve only as the first step in understanding the functional impact of a variant on disease. Bell and colleagues, for instance, showed that 27% of 406 published sequencing variants indicated as severe disease mutations actually were common SNPs or did not have enough evidence to confirm pathogenicity.[69] Others have urged caution in assigning pathogenicity to variants before rigorous follow-up in order to avoid adverse consequences for patients and research.[70] In addition to the confusion and discrepancies in nomenclature for the relatively new field of sequencing, the field lacks a gold standard for following-up VUS. Depending on the nature of each VUS and the disease of interest, different follow-up studies may include additional population genetic study designs, expression studies, molecular characterization of a gene or pathway, or more. Such studies can take anywhere from a few years to decades. Thus, there is a delicate balance between accumulating and publishing results from sequencing and biochemical analyses in order to advance research findings and ultimately serve pressing clinical needs.

Furthermore, despite the incredible potential of WES and WGS, these technologies are still limited. WES only captures about 1% of the genome, and captures vary in their target intervals and weaknesses.[71] WGS, although it covers the entire genome by definition, may not have enough sequencing depth in some regions to accurately call variants. Additionally, the relative costs of WES versus

WGS may make it more cost effective to choose WES and sequence a larger number of individuals. For either technology, certain regions of the genome like highly repetitive sequence and CG-rich intervals are still difficult to sequence accurately.[72] For the variants that can be detected by WES or WGS, different analysis methods must be used to look for different types of variants present. Different bioinformatics programs, spanning the alignment step to variant calling, are more sensitive or specific to different types of variation. While some variant calling programs that call SNVs have been extensively tested, programs targeting indels, copy number variations (CNV), and other types of variants are currently less sophisticated (types of variants that can be identified through sequencing are depicted in Figure 4). For variants identified by these approaches, more extensive confirmation steps are necessary. Additionally, even for SNVs, important discrepancies still exist between variant callers. For instance, O'Rawe et al found only a 57.4% concordance across 5 SNV variant calling pipelines used on 15 exomes.[73]

Figure 4. **Types of variants that can be identified through high-throughput sequencing.** Each letter (A, B, C, D) corresponds to a distinct gene. Single nucleotide variants (SNVs) are <5% minor allele frequency (MAF), whereas single nucleotide polymorphisms (SNPs) are >5% MAF. Sequence variation refers to SNV/SNPs or small (<1kb) indels. Rahim et al, 2008.[74]



Researchers rely on accurate annotation of variants to narrow down groups of variants to study further. While there are some common sources for these annotations, there are still discrepancies between even commonly used databases.[75] Bioinformatics programs that facilitate annotation of sequencing data may also resolve discrepancies in slightly different ways,[39] adding more difficulty in attempts to combine or replicate results. Additionally, the number and types of *in silico* prediction programs have exploded in the past few years. Sensitivities and specificities of each program differ from one another,[76-79] and there is still not a gold standard approach to their application.

As sequencing is applied to more complex diseases, questions about study design, including the number and type of samples to include as well as the appropriate statistical tests to employ, become more intricate.[80] In order to expand sample sizes, some research groups have combined sequencing data from other groups and/or data repositories. Batch effects, or variation due to varying experimental conditions (e.g. temperature, time, personnel), can be widespread in such situations.[81,82] Furthermore, the possible introduction of false signals from minor differences in ancestry is likely to be more misleading in analysis of rare sequencing variants as opposed to GWA studies.[39] Researchers are actively working on identifying and correcting for such quality control errors to increase reproducibility.

To add another complicating factor, allelic architecture, or the numbers and types of risk variants, is known to differ substantially between diseases.[23,83] Manolio and colleagues, for instance, cite how the majority of heritability for age-related macular degeneration can be explained by 5 loci, whereas several dozen loci of smaller effect size have been identified for Crohn's disease.[23] Unlike previous linkage and GWA studies, sequencing has the potential of identifying variants across the spectrum of frequencies and effect sizes, but different types of analyses may need to be performed to determine robust associations.

Statement of Purpose

In the present study, WES and WGS are applied to familial cases of 3 different disease models: intracranial aneurysm (IA), Parkinson disease (PD), and X-linked ataxia dementia (XLAD). Through these examples of familial disease, we explore the opportunities and challenges of family-based HTS. The work presented herein contributes effective practices for study design, analysis, and interpretation in a rapidly growing field still replete with questions about how best to implement HTS in studying familial disease.

CHAPTER I: FAMILIAL INTRACRANIAL ANEURYSM

Introduction

Subarachnoid hemorrhage (SAH) is the most devastating subtype of stroke. Fatality from SAH between 21 days to one month of the hemorrhage ranges from 25-35% in high-income countries to almost 50% in low- to middle-income countries.[84] Up to 80-90% of SAH cases are caused by rupture of IA, which are present in approximately 3% of the population.[85] Smoking and hypertension are important risk factors, increasing the risk of IA rupture by 3.1 and 2.6 times respectively.[86] The risk of an IA and for IA rupture is also increased among individuals having a first-degree relative with a history of an IA [85,87,88]. The location and number of IAs in a given individual also appears to be influenced by a family history.[89] Thus, several lines of evidence suggest that IA is due to both genetic and environmental risk factors. Unfortunately, until more is understood about these risk factors, the severe morbidity and mortality associated with this disease will continue to be a large public health burden.

Several approaches have been employed to identify genes contributing to IA. Initial studies utilized pedigrees having multiple affected members. Analyses in these initial studies detected linkage to several chromosomal regions (1p34.3-36.13[90,91], 4q32.2[92], 6p23[93],7q11[90], 7q36.3[92], 8q12.1[92], 11q24-25[93-95], 12q21.33[92], and 14q23-31[95]); however, the causative gene was not identified in any of these regions. More recently, GWA studies have focused on the role of

common variants that might individually have a small effect on disease risk. Analyses have consistently detected association to SNPs in *CDKN2BAS*, also known as *ANRIL*, on chromosome 9p21.3[96-98], as well as *SOX17* on chromosome 8q12.1[96-98]. Association has also been reported to *EDNRA* on chromosome 4q31[99], *CNNM2* on chromosome 10q24[97], *KL/STARD13* on chromosome 13q13[97], and *RBBP8* on chromosome 18q11[97]. Together, these genes only explain a fraction of the population attributable risk for IA.

Advances in technology, especially in the development of HTS, now make it possible to efficiently search for rare variants having a large effect on disease risk. These rare variants may point to novel genes and pathways that are critical to improve the molecular understanding of IA and methods of predicting those at greatest risk. In the present work, WES was applied to a unique set of families densely affected with IA to investigate the role of rare genetic variation in disease susceptibility and to demonstrate important study design considerations for WES studies in complex disease.

Materials and methods

Families selected for whole exome sequencing

Individuals were recruited as part of the Familial Intracranial Aneurysm (FIA) Study.[100] Study approval was granted by the institutional review boards at Indiana University, University of Cincinnati, and all participating study sites. Written consent was obtained from all study participants.

Families were recruited to ensure that DNA could be obtained from at least two living affected relatives and that the family would be informative for linkage analysis. Exclusion criteria included (i) a fusiform-shaped unruptured IA of a major intracranial trunk artery; (ii) an IA that is part of an arteriovenous malformation; (iii) a family or personal history of polycystic kidney disease, Ehlers Danlos syndrome, Marfan's syndrome, fibromuscular dysplasia, or Moya-Moya disease; or (iv) failure to obtain informed consent from the patient or family members. To identify unruptured IA, magnetic resonance angiography (MRA) was offered to first degree relatives of affected family members who had a higher risk of IA as defined by: 1) 30 years of age or older and 2) either a 10 pack year history of smoking or an average blood pressure of ≥140 mmHg systolic or ≥90 mmHg diastolic.

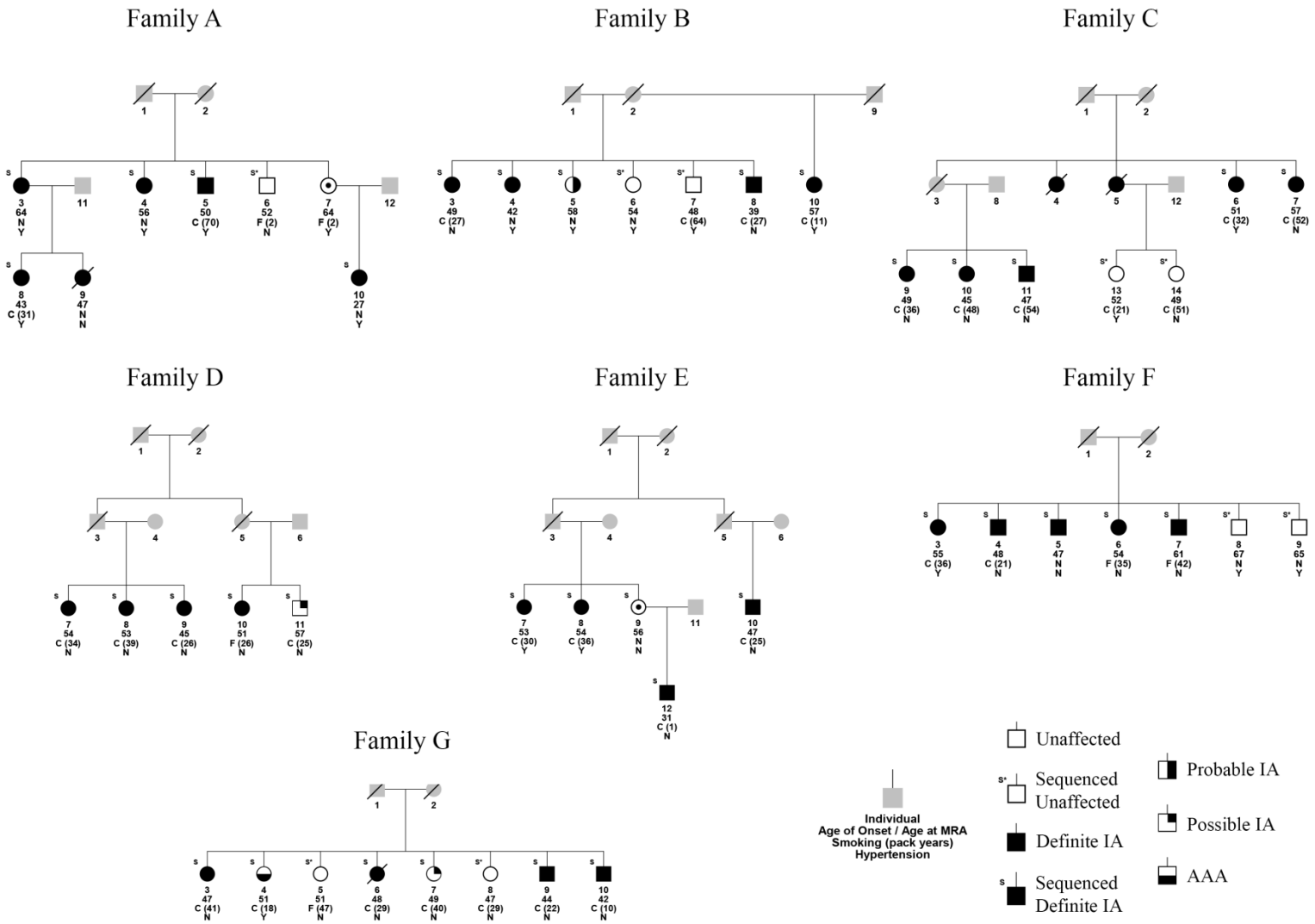
Only individuals having an IA based on an intra-arterial angiogram, operative report, autopsy, or size ≥7 mm on non-invasive imaging (MRA) were considered "definite" cases (Table 3). Two neurologists independently reviewed each record to determine if a subject met all inclusion and exclusion criteria. In case of disagreement, a third neurologist reviewed the data.

Table 3. **Intracranial aneurysm disease phenotypes.** Classification of affected status of samples in the Familial Intracranial Aneurysm Study.

Classification	Definition
Definite	Medical records document an intracranial aneurysm (IA) on angiogram, operative report, autopsy, or a non-invasive imaging report (MRA, CTA) demonstrates an IA measuring 7mm or greater.
Probable	Death certificate mentions probable IA without supporting documentation or autopsy. Death certificate mentions subarachnoid hemorrhage (SAH) without mention of IA and a phone screen is consistent with ruptured IA (severe headache or altered level of consciousness) rapidly leading to death. An MRA documents an IA that is less than 7 mm but greater than 3 mm.
Possible	Non-invasive imaging report documents an aneurysm measuring between 2 and 3 mm or SAH was noted on death certificate, without any supporting documentation, autopsy or recording of headache or altered level of consciousness on phone screen. Death certificate lists ‘aneurysm’ without specifying cerebral location or accompanying SAH.
Not a case	There is no supporting information for a possible IA.

Seven families of European American descent with the highest density of affected individuals who also had DNA available were selected for WES[101] (Figure 5). All affected individuals for which sufficient DNA was available were selected for sequencing. Unaffected individuals were selected only if there was an MRA conducted that confirmed the absence of an IA at 45 years or older and if there was sufficient DNA available. One clinically unaffected individual in family E was assumed to be an obligate carrier and was sequenced with her offspring to allow confirmation of allele transmission. Within the seven families, 45 individuals were chosen for WES.

Figure 5. Simplified pedigrees for the intracranial aneurysm whole exome sequencing families. Only sequenced individuals and those needed to preserve generational structure are shown to protect the anonymity of the pedigree. IA=intracranial aneurysm. All affected individuals are definite IA unless noted as a probable IA, possible IA, or aortic abdominal aneurysm (AAA). Criteria for defining definite, probable, and possible IA statuses are outlined in Table 1. All unaffected individuals, with the exception of individual E-9, had an MRA performed that did not show evidence of an IA. Grey indicates an unknown phenotype. An ‘S’ above an individual denotes that the individual was selected for sequencing.


Family A
Family B
Family C
Family D
Family E
Family F
Family G
Individual
Age of Onset / Age at MRA
Smoking (pack years)
Hypertension
Unaffected
Sequenced Unaffected
Definite IA
Sequenced Definite IA
Probable IA
Possible IA
AAA

Whole exome sequencing

WES was performed at the Center for Inherited Disease Research (CIDR, Johns Hopkins University). Exonic sequences were captured using the Agilent SureSelect Human All Exon 50Mb kit, and paired-end sequencing was performed on the Illumina HiSeq 2000 system, using Flowcell version 3 and TruSeq Cluster Kit version 3. All samples were genotyped using the Illumina HumanOmniExpress-12v1_C platform for quality assurance. Two HapMap samples and two study duplicates were used to ensure library preparation batch quality.

Whole exome sequencing bioinformatics

Primary analysis was done using HiSeq Controls Software and Runtime Analysis Software. The CIDRSeqSuite pipeline was used for secondary bioinformatics analysis, which consists mainly of alignment using BWA (version 0.5.9)[43] to the human genome reference sequence (build hg19) and applying GATK (version 1.0.4705)[44] to perform local realignment and base quality score recalibration. Duplicate molecules were flagged and mate-pair information synchronized using Picard (version 1.52, http://picard.sourceforge.net/). The GATK Unified Genotyper (GATK version 1.2-29) was used for multi-sample variant calling. The dataset, consisting of called variants, subject phenotypes, and pedigree information for the multiplex IA families can be requested directly from the National Center for Biotechnology Information (NCBI) Database of Genotypes and Phenotypes (dbGaP) (accession phs000636). Mapped reads are available

on the Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) (accession SRX329208-SRX329252).

GATK VQSR (GATK version 1.2-38)[102] created a high-quality call set for SNVs by using an adaptive error model to estimate the likelihood of true genotype calls based on aggregating information across multiple quality metrics. As recommended by GATK, HapMap 3.3 and the Illumina Omni 2.5M chip sites, available from the GATK bundle 1.2, were used as training sets and the annotations of Quality by Depth, Haplotype Score, Mapping Quality Rank Sum, Read Position Rank Sum, Fisher Strand Bias Test, and Mapping Quality were used as quality metrics for the recalibration. SNVs were filtered until 99% of the overlapping HapMap 3.3 sites were retained after application of VQSR. Insertion/deletions were removed if they had a quality by depth < 2.0, ReadPosRankSum < -20.0 (Z-score from Wilcoxon rank sum test of alternative versus reference read position bias), Fisher's Strand Bias > 200.0 (phred-scaled p-value using Fisher's exact test to detect strand bias), and/or a homopolymer run > 5.

ANNOVAR[54] was used to annotate variants for location, predicted effect on the protein across three gene databases (RefSeq, UCSC, and Ensembl), and corresponding gene and transcript length. Allele frequencies within European American populations in 1000 Genomes (February 2012 release, http://www.1000genomes.org)[36] and the Exome Sequencing Project (ESP)

(ESP6500 release with insertion/deletions and chromosome X and Y calls, http://evs.gs.washington.edu/EVS/)[103] were recorded using custom scripts. The scripts mapped variants to 1000 Genomes and ESP based on chromosomal position and reference and alternate alleles to determine allele frequencies. If a variant was not found in 1000 Genomes or ESP, the alternate allele frequency was set to 0. If a variant was found in both 1000 Genomes and ESP, the smaller alternate allele frequency was taken as the consensus frequency.

Variants were annotated for binned minor allele frequencies from 290 samples without a known cardiovascular phenotype that were exome sequenced at CIDR using identical capturing and sequencing technology, although SAMtools[45] was used for variant calling instead of GATK Unified Genotyper. Variants that were monomorphic across all samples were also flagged.

Variants were also annotated using custom scripts for Gene Ontology (GO) (http://www.geneontology.org)[104] terms that were hypothesized to play a role in IA pathophysiology. GO terms used included GO:0001944 (vasculature development), GO:0001570 (vasculogenesis), GO:0003018 (vascular process in circulatory system), GO:0005581 (collagen), GO:0005604 (basement membrane), and GO:0051541 (elastin metabolic process).

Two programs were used to predict the pathogenicity of SNVs: SIFT[52] and PolyPhen-2[105]. Scores of damaging for SIFT, or possibly or probably damaging

for PolyPhen-2, were accepted as evidence for pathogenicity. Two additional programs were used to analyze the effect of insertions and deletions: SIFT-INDEL[53] for those that cause a frameshift, and DDIG-in for those that do not cause a frameshift[47]. Variants were also annotated for C-scores from the Combined Annotation Dependent Depletion (CADD) webserver (http://cadd.gs.washington.edu)[46]. C-scores of 10 or greater, corresponding to the 10% most deleterious substitutions in the human genome according to CADD, were considered damaging predictions.

Biological filtering retained loci if they: 1) were autosomal variants; 2) were predicted to be nonsynonymous SNVs or insertion/deletions in an exonic and/or splicing region (within 2 bp of a splicing junction, as annotated by ANNOVAR) based on RefSeq, UCSC, and Ensembl annotations; 3) had an allele frequency in European American populations <1% (1000 Genomes, ESP); 4) had an allele frequency less than 1% in CIDR binned minor allele frequencies and were not monomorphic across all samples; 5) were predicted most likely to be damaging by CADD and by at least one other protein prediction program; and 6) segregated with all individuals with a definite IA and obligate carriers in at least one family. All alignments for variants passing these biological filters were visually inspected using the Integrated Genomics Viewer (IGV)[55] to confirm presence of a variant. Visual inspection for each variant included reviewing read pair orientations, mappability, and soft-clipping; variants that were called nearby; overall depth of sequencing and genomic features that might have inhibited

coverage at that locus; and repetitive sequence that might have influenced the position or variant allele called for the locus. In addition to the filters described above, insertion/deletions were also compared against a different dataset consisting of approximately 500 samples without a known cardiovascular phenotype. This comparison dataset used GATK Unified Genotyper (version 2.3-9) for variant calling and the Agilent SureSelect Human All Exon 51Mb capture kit. If the allele designations and/or positions did not match between the two datasets but were within 10 bp, manual review of both the IA and comparison BAM files with IGV was done to reconcile differences in allele designations and position assignments between the two datasets.

Loci were also annotated if they: A) segregated with all aneurysms (including probable and possible IA and the one abdominal aortic aneurysm case in family G) and B) were not found in any sequenced unaffected individuals, excluding assumed obligate carriers.

Linkage

The 7 families were included as part of a larger linkage study of 2,317 individuals from 394 families using the 6K Illumina array.[92] Multipoint parametric linkage analysis (autosomal dominant inheritance, 1% disease allele frequency was performed using Merlin.[56] Only genotypic data from family members with definite IA and obligate carriers were included in the linkage analysis. WES variants were annotated for the highest LOD score obtained by linkage markers within a 10Mb

window centered on the sequencing variant. A maximum possible LOD score for each family was calculated by simulating a hypothetical fully informative marker using the aforementioned model parameters and the pedigrees for each family.

Tissue collection for RNA expression

Aneurysm biopsies from the aneurysm fundus distal to the clip were collected from patients undergoing neurosurgical clipping of an IA at the Department of Neurology and Neurosurgery in the University Medical Center Utrecht in the Netherlands. These patients were completely independent of the families included for WES. Patients undergoing surgery because of intractable epilepsy were included as controls, and part of a superficial cortical artery in the resected part of the brain was excised as control vessel tissue. Samples were collected from 44 aneurysm biopsies (22 ruptured, 21 unruptured, 1 with unknown rupture status) and 16 control biopsies. All samples were immediately snap frozen in liquid nitrogen less than 1 minute after excision and stored at -80 °C until further use.

RNA isolation, sample preparation, and sequencing

RNA isolation, sample preparation, and sequencing was conducted at the University Medical Center Groningen in Groningen, the Netherlands. Each sample was homogenized with zirconia/silica beads in the BeadBeater machine (BioSpec products, Inc.). After homogenization, total RNA was extracted and purified using an RNeasy microkit (Qiagen, Valencia, CA, USA) according to the

manufacturer's instructions. An initial quality check of the samples by capillary electrophoresis and RNA quantification for each sample was performed using the LabChip GX (PerkinElmer, Waltham, Massachusetts, USA). Samples with a minimum amount of 7 ng non-degraded RNA were selected for subsequent sequencing analysis. Sequence libraries were generated using the TruSeq RNA sample preparation kit from Illumina (San Diego, USA) using the Sciclone NGS Liquid Handler (Perkin Elmer). To remove contamination of adapter-duplexes, an extra purification of the libraries was performed with the automated agarose gel separation system Labchip XT (Perkin Elmer). The obtained cDNA fragment libraries were sequenced on an Illumina HiSeq2000 using default parameters (single read 1x100bp) in pools of 10 or 11 samples. Processing of the raw data, including a demultiplexing step, was performed using Casava software (Illumina) with standard settings.

Differential expression analysis

Sequencing reads with quality score under Phred Score <30 were discarded. The quality filtered trimmed fastQ files were then aligned to the human reference genome (hg19) using the STAR aligner,[106] allowing for 2 mismatches. SAMtools version 0.1.18[45] was used to sort the aligned reads. Gene level quantification was performed by HTSeq-0.5.4[107] using parameters --mode=union --stranded=no and Ensembl version 71 as the gene annotation database.

R version 3.1.0 was used for differential expression analysis. The counts per gene for each sample obtained after alignment were used as input for the analysis. Low count genes (genes with less than 1 read per million in *n* of the samples, where *n* is the size of the smallest group of replicates, i.e. *n*=16) were filtered out since there is little power to detect significant evidence of differential expression in these genes.[108]

The Bioconductor (version 2.14) packages edgeR (version 3.6.2) and limma (version 3.20.2) were used for subsequent steps. To correct for technical influences, edgeR adjusts for varying sequencing depths between samples and normalizes for the RNA composition of the sample. A generalized linear model was used to test for differential expression between aneurysmal and control tissue. Other factors included in the model were age and sex of patients, as well as rupture status. Common and tagwise dispersion estimates were calculated with the Cox-Reid profile adjusted likelihood method to be able to correct for the technical and biological variation when fitting the multivariate negative binomial model. In estimating the tagwise dispersion, the program default for degrees of freedom (df=10) was used. A negative binomial generalized log-linear model, using the tagwise dispersion estimates, was fitted to the read counts for each gene, and a gene-wise statistical test was performed. Then, a likelihood ratio test was performed. Benjamini Hochberg false discovery rates (FDR) for a transcriptome-wide experiment were calculated to correct for multiple testing. All

genes with an FDR adjusted p-value <0.05 were considered individual genes of interest.

Results

Sequencing data quality

The average study duplicate reproducibility of SNV and insertion/deletion calls were 99.13% and 94.42%, respectively, and genotypes for non-reference calls per sample from the WES data achieved an average 99.57% concordance with genotype calls from the Illumina® HumanOmniExpress-12v1_C array. The average sensitivity to heterozygote calls on the array was 98.13%. After application of GATK quality filters, 98,351 SNVs and 5,851 insertion/deletions were retained. The transition-transversion ratio for exonic variants and percent of SNVs in dbSNP 137, both measures of the quality of the data, were 3.3 and 94.79% respectively.

Biological Filtering

The number of variants retained after each biological filter employed in the Methods is shown in Tables 4-5 for SNVs and insertion/deletions, respectively. The list of SNVs and insertion/deletions satisfying biological filters 1-6 is shown in Table 6. The final candidate variants passing biological filters 1-6 and manual inspection include 67 SNVs and 1 deletion. The sets of variants that A) segregate with all aneurysmal phenotypes and B) are not carried in unaffected individuals

are included in Table 6 as subsets of the 68 final variants. The limitations of only considering these sets of variants are described in the Discussion.

Table 4. Intracranial aneurysm whole exome sequencing single nucleotide variant filtering pipeline.

Numbers in parentheses refer to filtering steps described in the Methods. IA = intracranial aneurysm

Family	A	B	C	D	E	F	G	All
All variants found in at least one definite IA	46168	41978	44689	44515	49142	39495	37809	98351
(1) Autosomal variants	45390	41280	43994	43701	48376	38925	37251	96552
(2) Variants predicted to be functional	12261	11158	11849	11841	13203	10578	10025	29194
(3) Rare variants	1020	889	953	1298	1356	843	823	7845
(4) Variants not found or of low frequency in the internal allele frequency database	793	725	740	1028	1049	676	658	6428
(5) Variants predicted damaging	393	345	369	442	470	297	306	3008
(6) Variants segregating with all definite IA in at least one family	13	11	2	10	4	8	24	67
Variants passing visual inspection	13	11	2	10	4	8	24	67
A. Variants segregating with all IA (definite, probable, possible) or AAA in at least one family	13	9	2	8	3	8	7	46
B. Variants not found in unaffected individuals	5	2	1	7	3	1	0	19

Table 5. **Intracranial aneurysm whole exome sequencing insertion deletion filtering pipeline**. Numbers in parentheses refer to filtering steps described in the Methods. IA = intracranial aneurysm

Family	A	B	C	D	E	F	G	All
All variants found in at least one definite IA	3316	2736	3226	3166	3396	2987	2966	5851
(1) Autosomal variants	3264	2705	3178	3102	3345	2940	2921	5737
(2) Variants predicted to be functional	538	457	560	541	581	511	465	1126
(3) Rare variants	284	221	299	277	299	266	260	589
(4) Variants not found or of low frequency in the internal allele frequency database	178	159	188	171	192	165	157	453
(5) Variants predicted damaging	60	59	65	50	59	55	42	194
(6) Variants segregating with all definite IA in at least one family	24	22	23	19	23	24	19	26
Variants passing visual inspection and manual review with internal database calls	0	0	0	0	0	0	1	1
A. Variants segregating with all IA (definite, probable, possible) or AAA in at least one family	0	0	0	0	0	0	0	0
B. Variants not found in unaffected individuals	0	0	0	0	0	0	0	0

Table 6. Candidate variants identified through whole exome sequencing in the intracranial aneurysm whole exome sequencing families. Chr = chromosome, Pos = position, Ref = reference allele, Alt = alternate allele. Alt Freq = alternate allele frequency (consensus frequency for the alternate allele from 1000 Genomes and/or Exome Sequencing Project, as described in the Methods), LOD = maximum LOD score for linkage markers found within a 10Mb window of the sequencing variant, Fam = family, Unaff = number of sequenced unaffected individuals who carry the variant, logFC = log fold change of expression differential (N/A indicates no expression data is available for the gene), FDR = false discovery rate-adjusted p-value. All variants are predicted to be non-synonymous exonic variants except the deletion at the end of the Table. A plus sign (+) denotes a damaging prediction. For variants segregating in families B, D, or G, a (§) indicates that variant was also shared by an individual in the same family with a probable or possible IA or an abdominal aortic aneurysm.

							Protein Prediction Programs								
Chr	Pos	Ref	Alt	Gene	Full_Name	Alt Freq	Poly Phen	SIFT	CADD	Amino Acid Change	LOD	Fam	Unaff	logFC	FDR
1	6631121	C	T	TAS1R1	taste receptor, type 1, member 1	0.0001		+	16.77	NM_177540:exon2:c.C344T:p.T115M	1.08	D§	0	N/A	N/A
1	15905363	G	T	AGMAT	agmatine ureohydrolase	0.0026		+	15.62	NM_024758:exon4:c.C711A:p.N	0.83	F	1	-0.127	0.952

					(agmatinase)					237K					
1	28206319	G	A	C1orf38	chromosome 1 open reading frame 38	0.0001	+	+	17.71	NM_001105556:exon3:c.G400A:p.A134T	0.57	G§	0	N/A	N/A
1	28477192	T	C	PTAFR	platelet-activating factor receptor	0.0052	+	+	20.80	NM_001164721:exon3:c.A341G:p.N114S	0.57	G§	0	-0.506	0.867
1	33760820	G	A	ZNF362	zinc finger protein 362	0.0000	+		21.80	NM_152493:exon8:c.G1060A:p.A354	0.85	B§	1	0.336	0.784
1	36638206	G	A	MAP7D1	MAP7 domain containing 1	0.0011	+	+	34.00	NM_018067:exon4:c.G602A:p.R201Q	0.47	D§	0	0.157	0.792
1	111968011	G	A	OVGP1	oviductal glycoprotein 1, 120kDa	0.0000	+	+	12.85	NM_002557:exon4:c.C311T:p.T104I	0.57	G§	1	-0.023	0.988
1	177899689	C	A	SEC16B	SEC16 homolog B (S. cerevisiae)	0.0010	+	+	21.60	NM_033127:exon25:c.G3102T:p.Q1034H	0.87	C	0	N/A	N/A
1	197072434	T	A	ASPM	asp (abnormal spindle) homolog, microcephaly	0.0013	+		14.55	NM_018136:exon18:c.A5947T:p.M1983L	0.57	G§	1	1.195	0.642

					associated (Drosophila)										
1	204418411	C	T	PIK3C2B	phosphoinositide-3-kinase, class 2, beta polypeptide	0.0007	+	+	35.00	NM_002646:exon15:c.G2248A:p.G750S	0.57	G§	1	-0.505	0.672
1	212799290	C	A	FAM71A	family with sequence similarity 71, member A	0.0000	+		13.78	NM_153606:exon1:c.C1071A:p.S357R	0.57	G§	1	N/A	N/A
1	228290051	T	G	C1orf35	chromosome 1 open reading frame 35	0.0093	+		21.10	NM_024319:exon5:c.A407C:p.E136A	-0.29	A	0	-0.079	0.934
2	101866509	C	T	KLF11	Kruppel-like factor 11	0.0003	+	+	14.69	NM_001177718:exon2:c.C224T:p.P75L	1.41	A	0	-0.129	0.892
2	55825844	A	G	SMEK2	SMEK homolog 2, suppressor of mek1 (Dictyostelium)	0.0026	+	+	23.90	NM_001122964:exon4:c.T629C:p.F210S	1.43	E	0	-0.222	0.631
2	73718061	A	G	ALMS1	Alstrom syndrome 1	0.0000	+	+	12.02	NM_015120:exon10:c.A8972G:p.D2991G	1.13	D§	0	-0.264	0.749

2	74757348	T	C	HTRA2	HtrA serine peptidase 2	0.0030	+	+	11.98	NM_013247:exon1:c.T215C:p.L72P	1.43	E	0	0.267	0.595
2	161029157	G	C	ITGB6	integrin, beta 6	0.0001	+	+	17.45	NM_000888:exon6:c.C844G:p.L282V	-0.84	G	1	N/A	N/A
3	126137556	G	A	CCDC37	coiled-coil domain containing 37	0.0052	+		12.36	NM_182628:exon7:c.G589A:p.A197T	-0.84	G	2	N/A	N/A
3	180334458	C	T	CCDC39	coiled-coil domain containing 39	0.0026	+		20.70	NM_181426:exon18:c.G2432A:p.R811H	0.22	A	1	0.167	0.882
3	186508024	A	C	RFC4	replication factor C (activator 1) 4, 37kDa	0.0000		+	12.98	NM_002916:exon10:c.T903G:p.H301Q	0.83	F	1	0.125	0.906
4	106158134	C	T	TET2	tet oncogene family member 2	0.0000	+	+	12.41	NM_017628:exon3:c.C3035T:p.P1012L	0.57	G§	1	-0.231	0.878
*4	106639176	T	A	GSTCD	glutathione S-transferase, C-terminal domain containing	0.0047	+		22.90	NM_001031720:exon2:c.T406A:p.C136S	0.57	G§	1	-0.199	0.781
5	11018087	T	C	CTNND2	catenin (cadherin-	0.0000	+		25.80	NM_001332:exon18:c.	-0.29	A	0	-1.940	0.401

					associated protein), delta 2 (neural plakophilin-related arm-repeat protein)					A3083G:p.K1028R					
5	140801897	C	T	PCDHGA11	protocadherin gamma subfamily A, 11	0.0007		+	18.54	NM_018914:exon1:c.C1103T:p.A368V	0.57	G	1	-0.587	0.624
5	140955835	C	T	DIAPH1	diaphanous homolog 1 (Drosophila)	0.0007	+		36.00	NM_005219:exon14:c.G1423A:p.E475K	0.57	G	1	0.344	0.612
5	149901055	G	A	NDST1	N-deacetylase/N-sulfotransferase (heparan glucosaminyl) 1	0.0036	+		18.54	NM_001543:exon2:c.G239A:p.R80H	1.43	E	0	-0.157	0.806
5	157053610	T	C	SOX30	SRY (sex determining region Y)-box 30	0.0013	+		15.84	NM_178424:exon5:c.A2000G:p.N667S	0.83	F	0	N/A	N/A
6	133169909	G	T	TBC1D7	TBC1 domain family, member 7	0.0042	+	+	23.60	NM_001143965:exon5:c.C413A:p.A138D	0.86	G§	1	-0.372	0.758

6	149856802	C	T	PPIL4	peptidylprolyl isomerase (cyclophilin)-like 4	0.0000	+	+	34.00	NM_139126:exon5:c.G394A:p.G132S	-0.29	A	0	0.100	0.900
6	159420630	A	T	RSPH3	radial spoke 3 homolog (Chlamydomonas)	0.0002	+	+	15.37	NM_031924:exon1:c.T379A:p.C127S	0.57	G	1	-0.140	0.858
6	167709705	G	A	UNC93A	unc-93 homolog A (C. elegans)	0.0052	+		24.10	NM_001143947:exon3:c.G455A:p.G152D	0.85	B	1	N/A	N/A
6	168317794	A	C	MLLT4	myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 4	0.0000	+	+	26.90	NM_001207008:exon18:c.A2522C:p.K841T	0.57	G§	1	-0.150	0.884
8	72958750	T	A	TRPA1	transient receptor potential cation channel, subfamily A, member	0.0000	+		14.64	NM_007332:exon17:c.A2059T:p.N687Y	-0.96	G	0	N/A	N/A

					1										
9	21166077	T	C	IFNA21	interferon, alpha 21	0.0002		+	10.42	NM_002175:exon1:c.A535G:p.K179E	1.12	D§	0	N/A	N/A
9	35404008	G	A	UNC13B	unc-13 homolog B (C. elegans)	0.0006	+	+	34.00	NM_006377:exon39:c.G4754A:p.R1585H	0.83	F	1	-0.377	0.658
10	13240791	C	A	MCM10	minichromosome maintenance complex component 10	0.0049	+		17.17	NM_018518:exon16:c.C2222A:p.T741K	0.85	B	1	1.318	0.563
10	47087309	G	C	PPYR1	pancreatic polypeptide receptor 1	0.0000	+	+	15.45	NM_005972:exon3:c.G526C:p.A176P	0.57	G§	1	N/A	N/A
10	82187167	G	A	C10orf58	chromosome 10 open reading frame 58	0.0013	+		36.00	NM_032333:exon5:c.G491A:p.R164Q	0.56	G	1	N/A	N/A
10	105218301	C	G	CALHM1	calcium homeostasis modulator 1	0.0001	+		16.88	NM_001001412:exon1:c.G208C:p.V70L	-0.29	A	1	N/A	N/A
10	105727572	C	G	SLK	FYN oncogene related to	0.0000	+	+	20.60	NM_014720:exon1:c.C69G:p.H2	-0.29	A	1	0.182	0.774

					SRC, FGR, YES					3Q					
ǂ10	105797397	G	A	COL17A1	collagen, type XVII, alpha 1	0.0005		+	14.75	NM_000494:exon46:c.C3205T:p.R1069W	0.57	G	1	N/A	N/A
10	105893436	T	G	WDR96	WD repeat domain 96	0.0005	+		23.90	NM_025145:exon35:c.A4538C:p.D1513A	-0.29	A	1	0.048	0.988
11	400124	C	G	PKP3	plakophilin 3	0.0013	+	+	12.37	NM_007183:exon6:c.C1431G:p.N477K	-0.71	G§	0	N/A	N/A
11	73074872	G	A	ARHGEF17	Rho guanine nucleotide exchange factor (GEF) 17	0.0003	+	+	18.47	NM_014786:exon16:c.G5327A:p.C1776Y	1.13	D§	0	0.162	0.931
11	108277861	C	T	C11orf65	chromosome 11 open reading frame 65	0.0064	+	+	21.30	NM_152587:exon4:c.G190A:p.A64T	1.13	C	1	N/A	N/A
11	124742851	G	A	ROBO3	roundabout, axon guidance receptor, homolog 3 (Drosophila)	0.0004	+	+	20.20	NM_022370:exon9:c.G1402A:p.V468M	1.31	A	0	-0.019	0.993

11	126147035	T	G	FOXRED1	FAD-dependent oxidoreductase domain containing 1	0.0013	+	+	18.40	NM_017547:exon10:c.T1171G:p.L391V	-0.58	F	1	-0.152	0.815
‡12	29680094	G	T	FOXM1	forkhead box M1	0.0000	+	+	13.37	NM_202003:exon8:c.C1957A:p.P653T	0.29	D§	0	0.885	0.615
*12	12630140	T	G	DUSP16	dual specificity phosphatase 16	0.0026	+		16.34	NM_030640:exon7:c.A1625C:p.D542A	-0.69	B§	2	-0.324	0.686
*12	49498284	T	G	LMBR1L	limb region 1 homolog (mouse)-like	0.0040	+		16.10	NM_018113:exon5:c.A382C:p.M128L	0.83	F	2	0.156	0.824
12	56335802	T	C	DGKA	diacylglycerol kinase, alpha 80kDa	0.0000		+	17.40	NM_001345:exon16:c.T1271C:p.V424A	1.11	D	0	0.551	0.544
*12	96374381	C	A	HAL	histidine ammonia-lyase	0.0006	+	+	25.70	NM_002108:exon17:c.G1472T:p.G491V	1.14	D§	1	-0.479	0.922
12	126139069	C	T	TMEM132B	transmembrane protein 132B	0.0002	+	+	10.88	NM_052907:exon9:c.C3050T:p.S1017L	1.14	D	0	-2.626	0.023
15	75014793	T	A	CYP1A1	cytochrome P450,	0.0003	+	+	14.09	NM_000499:exon2:c.	0.83	F	1	N/A	N/A

					family 1, subfamily A, polypeptide 1					A646T:p.S216C					
16	44949	G	A	NME4	non-metastatic cells 4, protein expressed in	0.0000		+	11.74	NM_005009:exon3:c.G296A:p.R99H	0.85	B§	1	-0.076	0.943
*16	21337701	G	A	TSC2	tuberous sclerosis 2	0.0040	+		12.84	NM_001114382:exon32:c.G3820A:p.A1274T	0.85	B§	1	-0.229	0.658
16	11785220	G	A	TXNDC11	thioredoxin domain containing 11	0.0014	+		18.28	NM_015914:exon8:c.C1826T:p.A609V	0.85	B§	1	0.134	0.896
16	20796338	G	A	ACSM3	acyl-CoA synthetase medium-chain family member 3	0.0013	+	+	22.00	NM_005622:exon8:c.G1052A:p.S351N	0.57	G	0	0.751	0.496
16	53321892	A	G	CHD9	chromodomain helicase DNA binding protein 9	0.0076		+	18.22	NM_025134:exon27:c.A5213G:p.K1738R	0.65	G	0	0.095	0.910

17	5425076	A	G	NLRP1	NLR family, pyrin domain containing 1	0.0042		+	10.35	NM_033007:exon12:c.T3461C:p.M1154T	-0.56	D§	0	0.293	0.727
17	48762223	G	A	ABCC3	ATP-binding cassette, sub-family C (CFTR/MRP), member 3	0.0013	+	+	22.70	NM_003786:exon29:c.G4267A:p.G1423R	0.85	B§	0	-0.043	0.994
17	61432613	T	A	TANC2	tetratricopeptide repeat, ankyrin repeat and coiled-coil containing 2	0.0000	+	+	25.00	NM_025185:exon12:c.T2222A:p.F741Y	0.85	B§	0	-0.213	0.859
19	11598418	G	A	ZNF653	zinc finger protein 653	0.0000		+	16.16	NM_138783:exon4:c.C860T:p.A287V	1.41	A	1	0.168	0.829
19	13226094	G	A	TRMT1	TRM1 tRNA methyltransferase 1 homolog (S. cerevisiae)	0.0002	+	+	20.70	NM_017722:exon4:c.C640T:p.R214W	1.41	A	1	0.222	0.737

19	57175814	C	G	ZNF835	zinc finger protein 835	0.0009	+		18.91	NM_001005850:exon2:c.G753C:p.E251D	0.86	G	1	-0.797	0.556
19	57723459	C	T	ZNF264	zinc finger protein 264	0.0000	+	+	11.70	NM_003417:exon4:c.C994T:p.R332W	0.86	G	1	-0.132	0.882
20	44463002	A	G	SNX21	sorting nexin family member 21	0.0000	+		22.20	NM_152897:exon2:c.A184G:p.S62G	0.85	B§	1	-0.220	0.797
6	153312343	TTTTA	T	MTRF1L	mitochondrial translational release factor 1-like	0.0000	NA	+(SIFT-INDEL)	14.77	NM_019041:exon6:c.915_918del:p.305_306del	0.57	G	1	-0.095	0.924

Of the 68 retained variants, five variants (found in the genes *GSTCD, DUSP16, LMBR1L, HAL*, and *TSC2*) were found in definite IAs in two families; in all of these cases, the variant segregated fully with definite IA in only one family. Two other variants (found in the genes *COL17A1* and *FOXM1*) were the only variants of the 68 retained variants that were labeled with vascular-related GO annotations (i.e. GO:0005604 basement membrane and GO:0005581 collagen; and GO:0001570 vasculature development and GO:0001570 vasculogenesis; respectively).

Linkage

The distribution of genome-wide LOD scores for each family is depicted in Figures 6-12, with the WES variants satisfying biological filters 1-6 superimposed. The maximum possible LOD score for each family given the model parameters and the specific pedigree structure is also reported in Figures 6-12. The highest LOD score obtained by linkage markers within a 10Mb window centered on each sequencing variant is recorded in Table 6. Of the 68 WES variants satisfying biological filters 1-6 and manual inspection, 23 variants had a LOD score for a linkage marker within 10Mb of the sequencing variant that fell within 0.01 of the highest possible LOD score for that family.

Figure 6. Summary of genome-wide multipoint linkage analysis for intracranial aneurysm whole exome sequencing Family A. Positions of candidate single nucleotide variants and insertion/deletions identified in the sequencing data are denoted by diamonds and crosses, respectively.

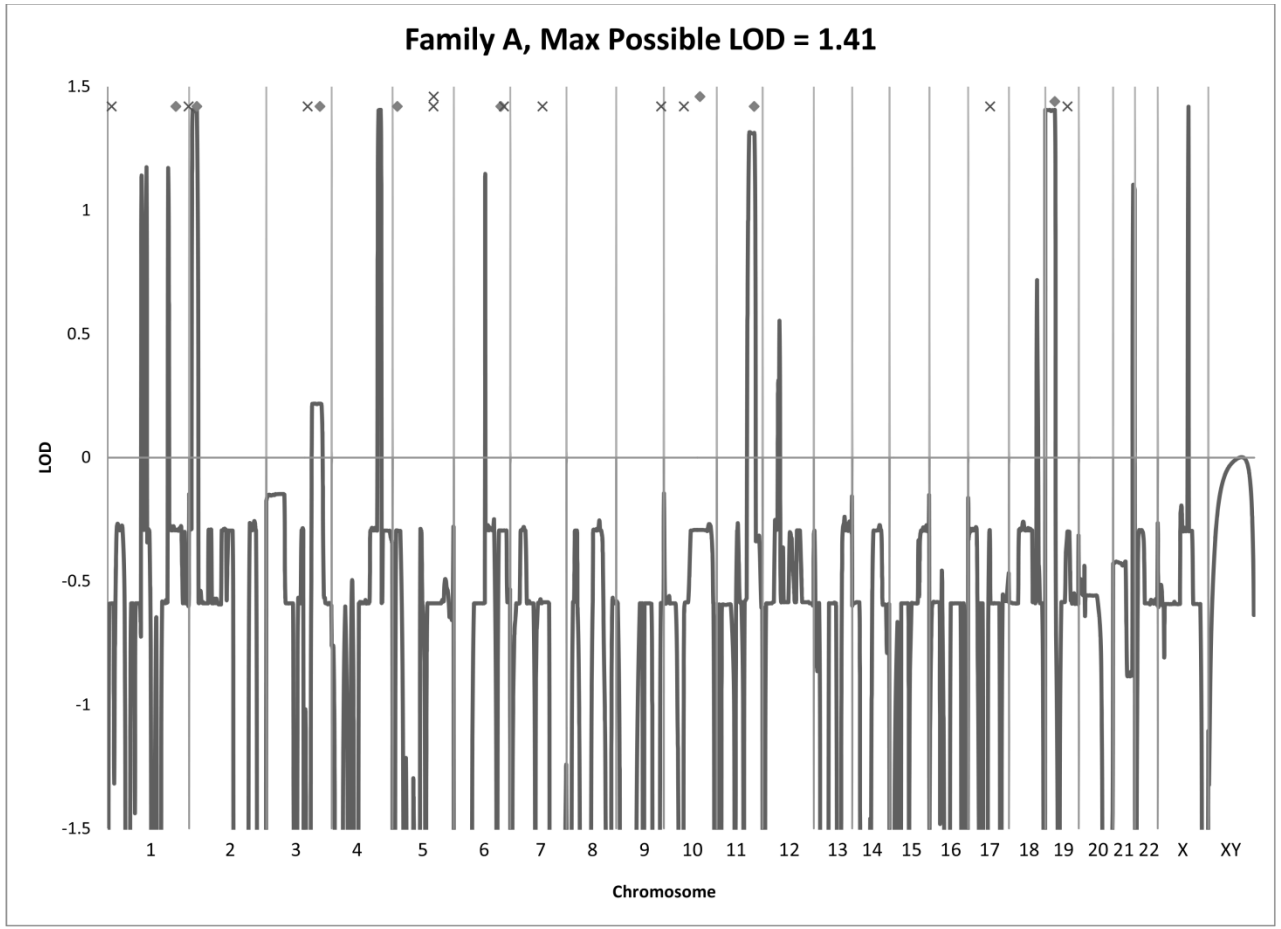


Figure 7. Summary of genome-wide multipoint linkage analysis for intracranial aneurysm whole exome sequencing Family B. Positions of candidate single nucleotide variants and insertion/deletions identified in the sequencing data are denoted by diamonds and crosses, respectively.

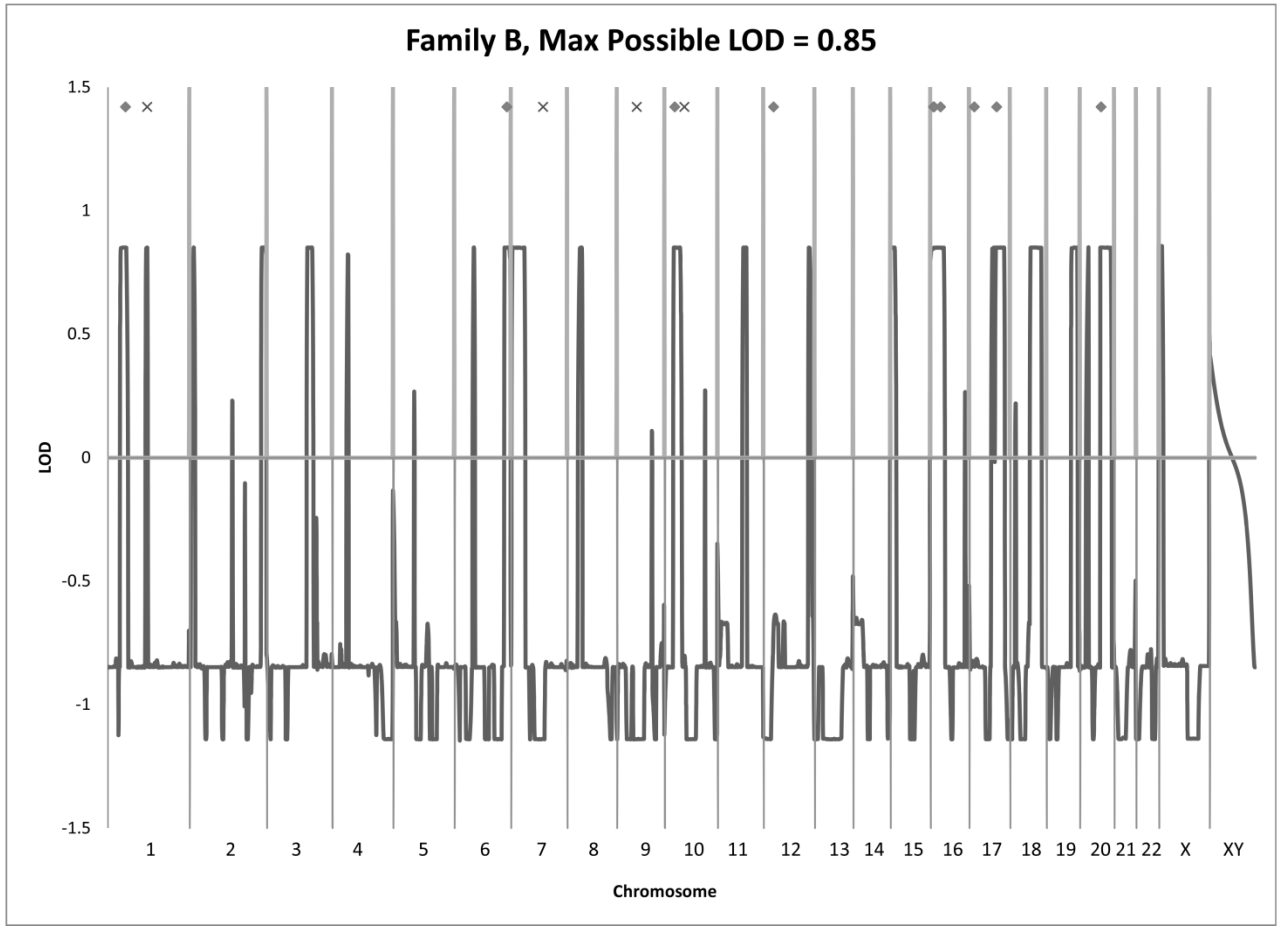


Figure 8. Summary of genome-wide multipoint linkage analysis for intracranial aneurysm whole exome sequencing Family C. Positions of candidate single nucleotide variants and insertion/deletions identified in the sequencing data are denoted by diamonds and crosses, respectively.

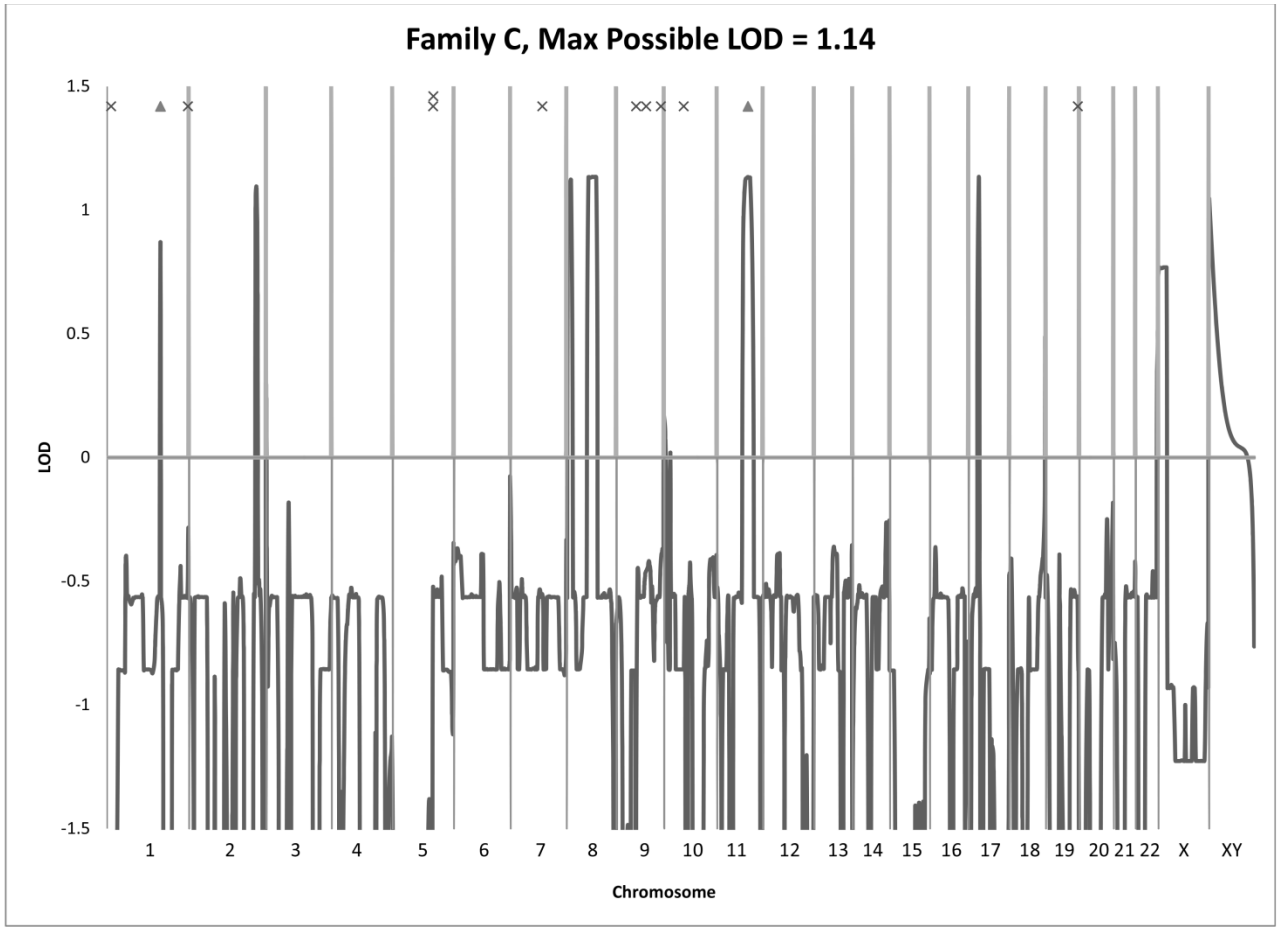


Figure 9. Summary of genome-wide multipoint linkage analysis for intracranial aneurysm whole exome sequencing Family D. Positions of candidate single nucleotide variants and insertion/deletions identified in the sequencing data are denoted by diamonds and crosses, respectively.

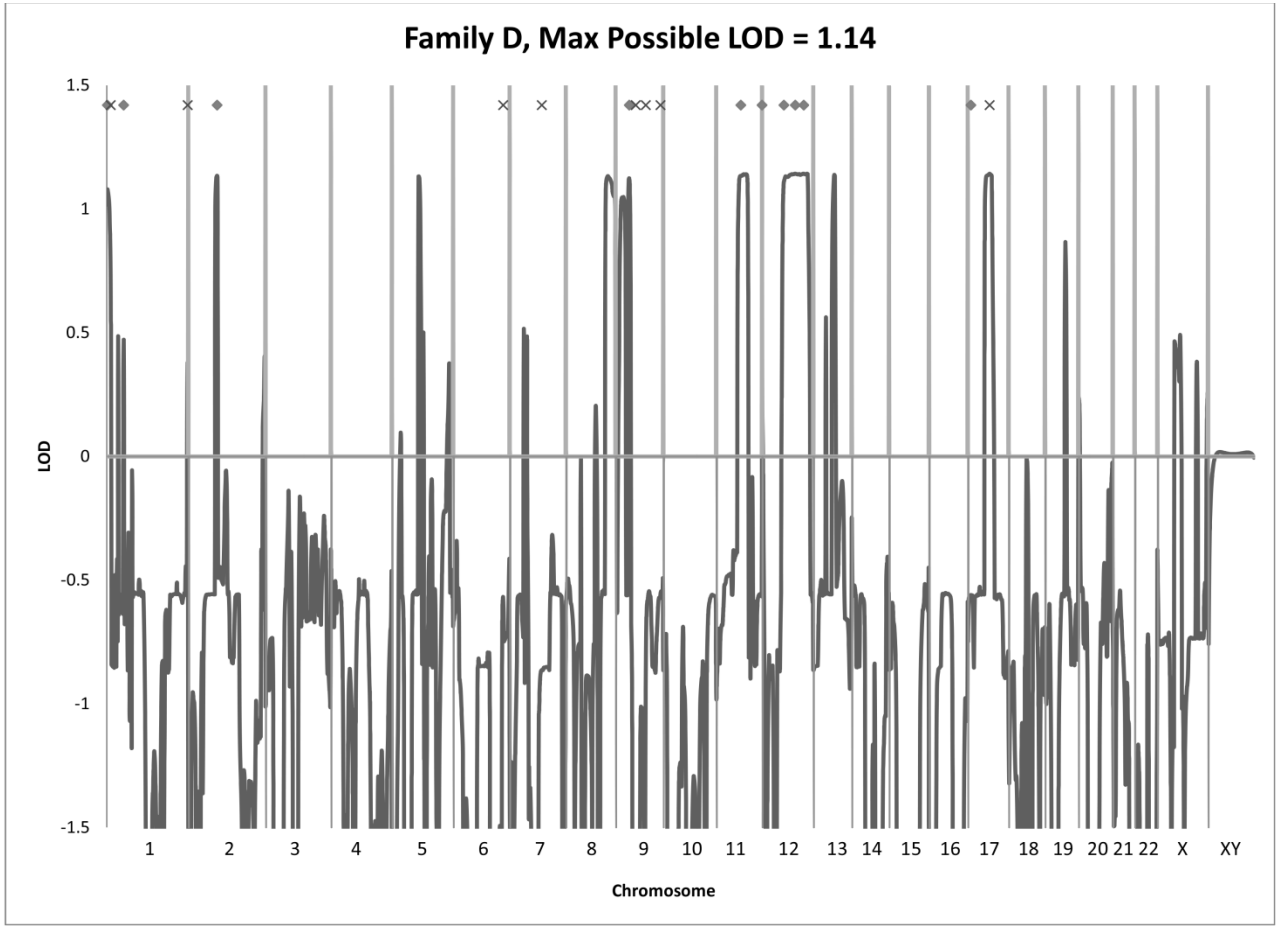


Figure 10. Summary of genome-wide multipoint linkage analysis for intracranial aneurysm whole exome sequencing Family E. Positions of candidate single nucleotide variants and insertion/deletions identified in the sequencing data are denoted by diamonds and crosses, respectively.

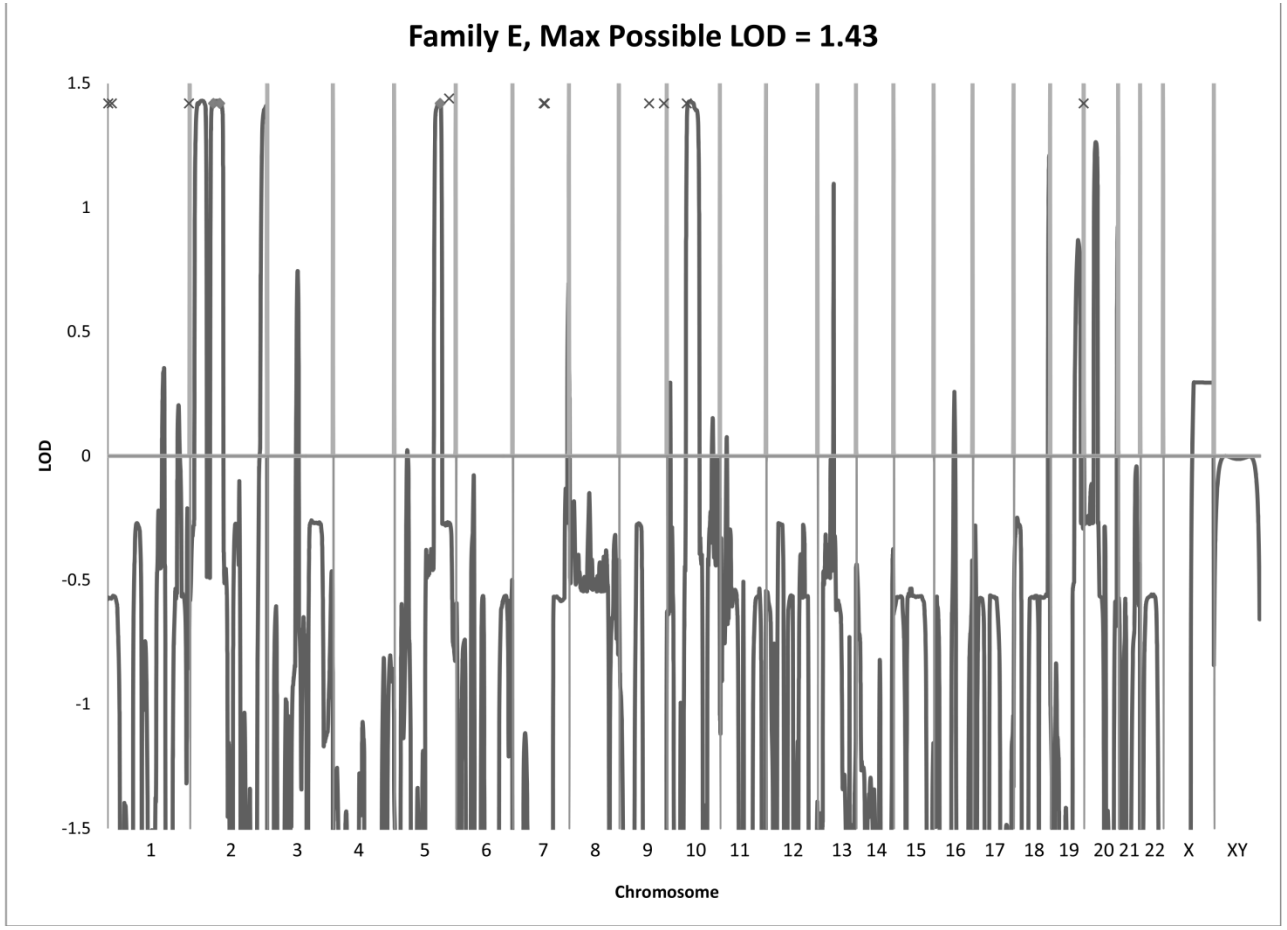


Figure 11. Summary of genome-wide multipoint linkage analysis for intracranial aneurysm whole exome sequencing Family F. Positions of candidate single nucleotide variants and insertion/deletions identified in the sequencing data are denoted by diamonds and crosses, respectively.

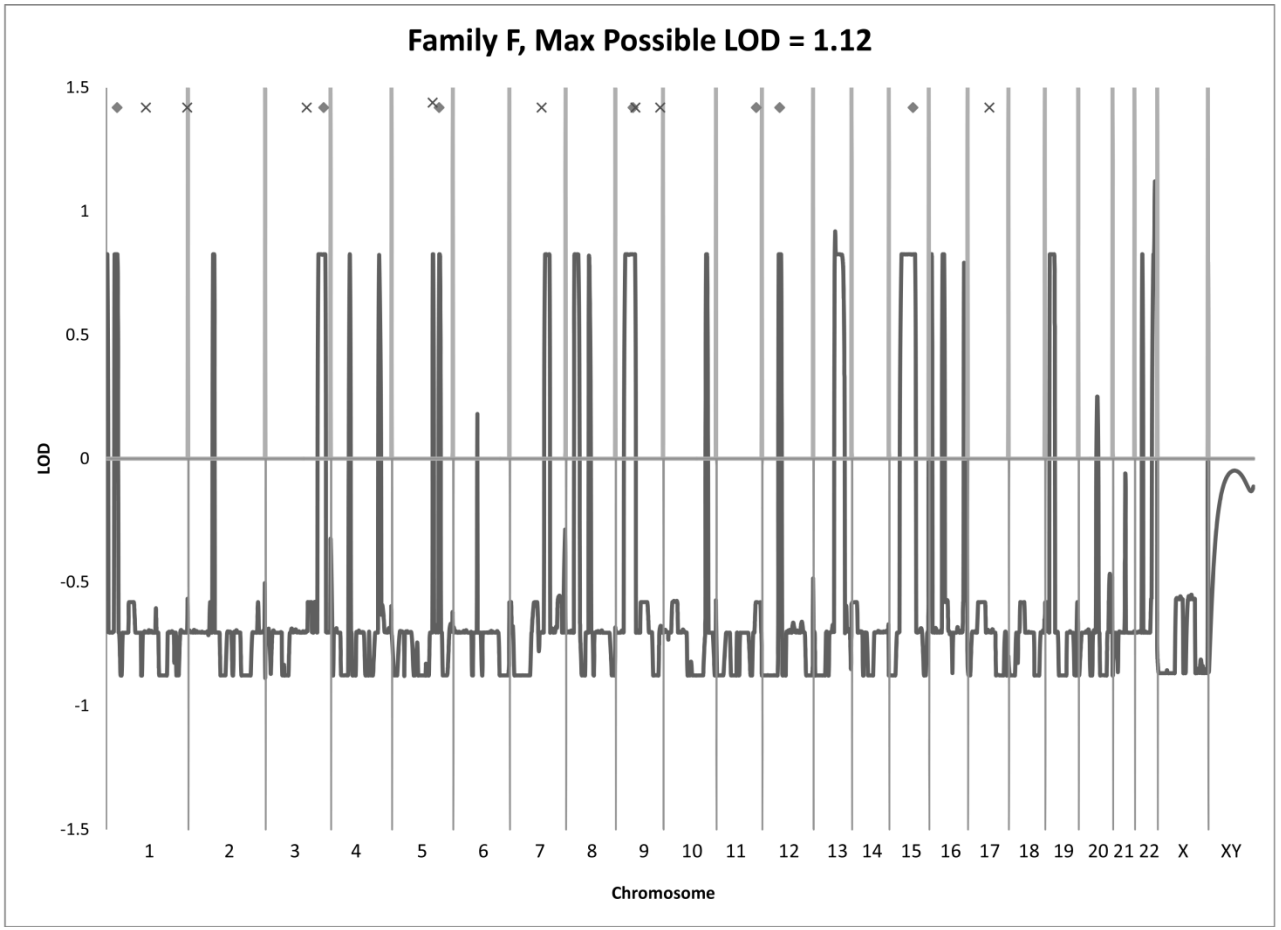
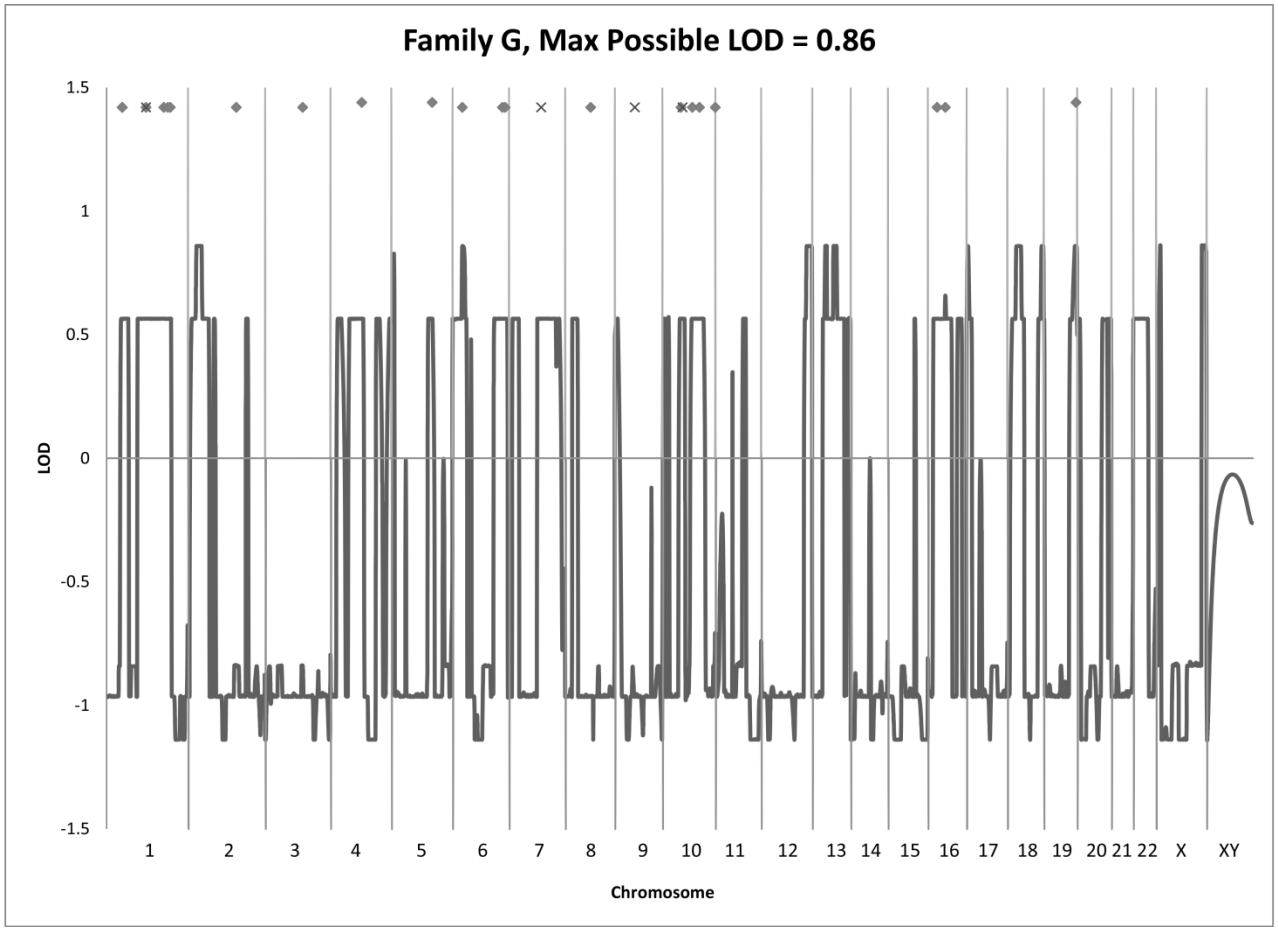


Figure 12. Summary of genome-wide multipoint linkage analysis for intracranial aneurysm whole exome sequencing Family G. Positions of candidate single nucleotide variants and insertion/deletions identified in the sequencing data are denoted by diamonds and crosses, respectively.



The 23 variants within a possible linkage peak were distributed among all families except family F, where the highest LOD score for a linkage marker within 10Mb of a filtered sequencing variant was 0.83 but the highest possible LOD score for the family was 1.12. Family B had the most retained variants within possible linkage peaks (n=9); followed by family D (n=4); families A, E, and G (n=3); and family C (n=1). Of the 23 variants, only 8 also met the optional prioritization criteria of segregating with all aneurysmal phenotypes and not being carried by an unaffected individual (*KLF11* variant in family A, variants in *ABCC3* and *TANC2* in family B, variants in *ALMS1* and *ARHGEF17* in family D, and variants in *SMEK2, HTRA2, and NDST1* in family E).

While none of the 68 variants coincided with well-established GWAS association signals, 6 of the variants were found within IA linkage peaks identified in previously published family studies, independent of the families in this report. Four variants (found in the genes *C1orf38*, *PTAFR*, *ZNF362*, and *MAP7D1*) fell within the linkage peak 1p34.3-36.13,[90,91] while 2 variants (found in the genes *ROBO3* and *FOXRED1*) were located in the linkage peak 11q24-25 [93-95]. None of these 6 genes were suggested as candidate genes by the authors of the published linkage studies. The linkage regions each cover hundreds of genes, as they span approximately 24 and 14 Mb respectively.

RNA expression

Expression data was obtained in 51 of the 68 candidate genes in an independent set of IA cases and controls. Log fold changes and FDR-adjusted p-values for each gene is displayed in Table 6. Only 1 gene (*TMEM132B*) of the 51 genes showed differential expression (overexpressed with log fold change=2.63, FDR-adjusted p-value=0.023).

Discussion

TMEM132B

Exome sequencing presents an opportunity to explore the contribution of rare variation to complex disorders like IA. We have used this approach to identify 68 rare variants in 68 genes that segregate within 7 densely affected families. Of the 51 genes that were expressed in IA tissue, one gene (*TMEM132B*) was found to be significantly overexpressed in IA tissue in comparison to control vascular tissue.

TMEM132B, or transmembrane protein 132B, is a relatively uncharacterized protein of unknown function. The variant segregating in the family is rare (0.024% frequency in European American samples in the Exome Sequencing Project and not found in 1000 Genomes) and predicted damaging by SIFT, PolyPhen-2, and CADD due to a change from the polar amino acid serine to the nonpolar amino acid leucine at a highly conserved position. Each of the individuals with a definite IA in family D was heterozygous for the variant. Mutations inherited in a dominant

manner often lead to a disease phenotype through a gain of function mechanism, which would be supported by the overexpression of *TMEM132B* in IA tissue as compared to control vessels. It is also possible, however, that dominantly-inherited mutations exert their effect via haploinsufficiency or dominant negative mechanisms. Further studies are required to confirm the role of *TMEM132B* in IA and through what mechanism the variant identified in this study may act.

The *TMEM132B* variant was not inherited by individual 11 in family D. Individual 11 was diagnosed as a possible IA due to the presence of a small aneurysm identified through non-invasive imaging (i.e. 1-2mm, verified by 3 independent neurologists). This is in contrast to the definite IAs clearly identified in this individual's sibling and cousins; thus, individual 11 is most likely actually unaffected.

Prioritization of variants within families

Expression information was only available for 51 of the 68 candidate genes; thus, RNA expression cannot confirm or rule out the role of the remaining 17 genes in IA pathophysiology. Additionally, a subset of the other 50 variants with expression data may also contribute to IA in ways not captured by the RNA expression experiment and should be explored. In order to further study the cause of IA in each of the remaining families, candidate variants in each family must be prioritized.

In families C and E, segregation analysis reduced the number of prioritized variants to only 2 and 4 variants, respectively. For family C, the two variants have a CADD score >20. The variant in *SEC16B* is not found within a potential linkage peak; however, in support of its potential significance in disease susceptibility, it is not carried by any tested unaffected family member. The variant in *C11orf65*, on the other hand, is found within a potential linkage peak but is also inherited by an unaffected family member. For family E, three variants segregate in the family (and a fourth variant in *GSTCD* is found in only one individual in family E but segregates fully in family G). The three variants that segregate in the family (in genes *SMEK2*, *HTRA2*, and *NDST1*) all are found within potential linkage peaks. Data are not available from any unaffected family members. Therefore, further prioritization among these three variants could incorporate CADD scores, which range from 11.98 for the *HTRA2* variant to 23.9 for the *SMEK2* variant.

Considerations for pedigree and phenotypic data

It is possible that genetic heterogeneity, phenocopies, or gene-environment interactions could explain one or more IAs in the families chosen for this study. In this case, the criterion requiring all affected individuals to share a variant would miss important disease-contributing variants. Similar family-based sequencing studies in the future could relax this segregation criterion with the recognition that a much larger number of variants will be retained. Family-based aggregative association tests that incorporate different penetrance models could also be employed with a larger number of samples.

The availability and quality of clinical data is also critical to consider in complex disease WES studies. In this study, several families also had individuals with probable and possible IAs (see Table 3 for phenotype definitions), and one family also had an occurrence of an abdominal aortic aneurysm (Figure 5). Given the high density of definite IAs in these families, it is likely that some or all of the probable and possible IAs have disease due to the same disease-contributing variant. Additionally, given the possible genetic link between different forms of aneurysms,[109] the abdominal aortic aneurysm may also share the same genetic etiology within that family. We thus flagged variants that segregated fully among all individuals with an aneurysm (definite, probable, or possible IA, or an abdominal aortic aneurysm) (Tables 4-5). This represents a possible method for prioritizing variants for further study, with the caveat that including non-definite IAs increases the likelihood of genetic heterogeneity, phenocopies, and gene-environment interactions.

Another approach to prioritize variants for further study is to utilize genotypic data from unaffected individuals. The ability of this approach to rapidly narrow down the number of variants under consideration is readily apparent from this study (Tables 4-5), but there are major concerns about inflating false negative rates by using unaffected individuals. Given the traditionally late age of onset for intracranial aneurysms, only individuals who had an MRA confirming the absence of IA at age 45 or older were sequenced as unaffected samples. Despite these precautions, the unaffected individuals in this study were still relatively close in

age to the age at diagnosis of their relatives who had an IA, and it is possible that the unaffected individuals will actually develop an IA later in life due to a genetic predisposition.

The difficulty in defining an unaffected also surfaces when considering the putative obligate carriers in these families. In Family A, individual A-7 also had an MRA done at age 64 that excluded the presence of an IA, yet we would posit that this individual likely passed a causative genetic variant to her daughter (A-10), whose IA is more likely to have a genetic basis due to her young age of onset. Without the daughter's data, individual A-7 would have likely been chosen as an unaffected individual for sequencing, especially given that she had major environmental risk factors (a history of smoking and hypertension). In Family E, the sequenced individual E-9 is also an obligate carrier under our model. Unlike individual A-7, an MRA could not be obtained on individual E-9, and she did not have a history of smoking or hypertension. Since all affected individuals in family E had at least one environmental risk factor and individual E-9 did not, it is possible that the causative genetic variant in family E requires an additional environmental insult to lead to IA development. The importance of strong environmental risk factors such as smoking to the development of aneurysms, even in the context of rare causal genetic variants, cannot be underestimated. Alternative methods of prioritization of variants that incorporate this possibility should be explored. Thus, unaffected status in this study was used as a mechanism for possible prioritization but not for automatic exclusion of variants.

The ability to use unaffected individuals will vary in studies of different diseases and will likely be more fruitful in those diseases that appear to have a smaller environmental/lifestyle contribution.

For future family-based sequencing studies in complex disease, it may not be feasible to sequence as many individuals per pedigree as was done for this study. Thus, it is critical to carefully select samples based on the quality of phenotyping and the pedigree structure. Recently developed tools offer statistical methods to select related subjects for sequencing based on genetic distance,[110] samples that span multiple generations (Exome Picks, http://genome.sph.umich.edu/wiki/ExomePicks), and a combination of both of these methods.[111] As evident from Tables 4-5, selecting families with more closely related individuals, such as families with full siblings as in Families F and G, will yield a smaller number of initially called variants across the family. Yet, the power to narrow down the number of variants segregating with disease is diminished in such families due to the naturally larger percentage of alleles shared, as compared to families with individuals in multiple generations such as in Family C. Thus, where possible, selection of more distantly related family members for sequencing studies will have greater power to generate a narrowed list of prioritized variants.

For some families, it may be possible to combine linkage and sequencing data to find causative variants. The families sequenced in this study were included as a

part of a larger linkage study reported previously.[92] The same model parameters used for WES variant filtering was applied for multipoint linkage analysis. Since any given marker may have been uninformative for a family, a maximum LOD score was reported within a 10Mb window of the sequence variant's chromosomal position. Although only modest evidence of linkage was obtained, several sequencing variants lay within the linkage regions in these families (Figures 6-12). Many variants, however, did not overlap with any evidence of linkage, suggesting that these families were either not fully informative for robust linkage analysis near these loci, or the sequencing variants identified are not causative genetic variants in these families.

Considerations for exonic variation

In recent years, WES has emerged as a practical method for systemically exploring rare coding variation. Since the majority of known genetic causes of Mendelian disorders affect protein coding regions,[31] the exome is a logical starting place to identify potentially causative variants in diseases that exhibit Mendelian inheritance. The densely-affected families sequenced in this study appear to display autosomal dominant inheritance; therefore, we hypothesized that coding variants may explain most or even all of these cases.

Due to imperfect capture and alignment, WES generates some off-target, non-exonic variant calls. While it is possible that important variation exists in these off-target regions, a higher percentage of calls in these regions are of poorer

quality. Thus, only those variants within exonic or splicing regions were retained in this experiment. Since different databases contain different numbers of and boundaries for genes and exons,[112] a consensus prediction of gene and exon boundaries was made to determine those variants that fell within exonic or splicing regions. In order to minimize the type I error rate by using functional predictions of the highest confidence, the intersection of functional predictions from three different databases (RefSeq, UCSC, and Ensembl) was used for this study. Thus, variants were only retained if they were predicted by all three databases to be within exonic or splicing regions. Other WES studies may choose to generate a larger set of variants by prioritizing all variants in the union rather than the intersection of functional predictions from multiple databases; however, appropriate methods for validating variants with functional predictions that differ by database should be employed.

It is possible that non-coding variants and/or epistatic interactions are important in IA development in these families and in other complex diseases, in which case alternate study designs should be utilized. At the time of this study, whole genome sequencing could have only been employed at the expense of sequencing fewer individuals, and annotations and bioinformatics tools available for non-coding sequence were less robust. Given that whole genome sequencing generates about 3 million SNVs per genome,[113] annotations and bioinformatics tools are even more critical for practical prioritization of candidate variants. In the future, techniques like whole genome sequencing, as well as targeted

resequencing, transcriptome sequencing, and other high throughput study designs, can be applied to fully catalogue the role of genetic variation in IA development.

Considerations for allele frequency

The average individual has around 15,000 exonic SNVs differing from the reference human genome sequence.[103] In order to narrow down the number of variants identified by a WES study, initial studies[34,35] focused on rare diseases and limited analysis to novel variants. This strategy is too restrictive for more common diseases such as IA. In the particular subset of families used for this study, there is a uniquely high incidence of IA, which enriches for the possibility of identifying rare, highly penetrant variants of larger effect sizes. Rare variants and less common variants are typically defined as less than 1% and 1-5% minor allele frequency, respectively.[114,115] Given the rarity of families that are as densely affected as the ones in this study, a 1% minor allele frequency threshold was set. It is possible, however, that a variant of higher minor allele frequency causes IA in one or more of these families. Future studies with a much larger sample size could employ aggregative association tests[42] with relaxation of the allele frequency threshold.

In this study, allele frequencies specifically from European American populations were available from public databases. Given that rare variants can be population-specific,[116] the selection of appropriate allele frequency databases is critical. In

lieu of publicly available allele frequencies, future studies may consider sequencing a large number of internal controls and possibly requesting commonly available controls to sequence as well. While not feasible for the current study, such a design would help control for platform- and pipeline-specific artifacts in sequencing while ensuring phenotyping quality for controls.

While it is standard for WES studies to utilize public databases to filter variants, it is also valuable to use internal frequency databases that are specific to the sequencing and variant calling pipeline. Because variant calling can be lab-specific due to the technology used, in this study variants were annotated for binned minor allele frequencies from 290 unrelated samples without a known cardiovascular phenotype that were exome sequenced at CIDR. Thus variants that would have otherwise been considered rare or novel when compared against public databases, but that were actually a recurring artifact of the sequencing, were captured as having a high CIDR binned minor allele frequency. Given that the bioinformatics pipeline used in this study differed slightly from that of the internal database, the internal database filter may have missed some artifacts specific to the variant calling method. Variants that were monomorphic (i.e. all heterozygous or homozygous for the alternate allele) across all samples were also removed since it is highly unlikely that the identical rare disease-causing allele would be shared by both affected and unaffected individuals in multiple families.

Indel allele frequencies in both internal and external databases are inherently less accurate than frequencies for SNVs, due to the increased difficulty and variation in calling structural variants. Also, differences in how position coordinates are assigned as well as reference and alternate allele designations further makes comparison challenging. The 26 indels that passed biological filters 1-6 (described in the Methods) in all cases except for one were shared in almost all or all of the 7 families sequenced in this study. Just as variants that were monomorphic across all datasets were removed as probable sequencing or pipeline artifacts, it is very unlikely that any given rare disease-causing insertion/deletions would also be shared across all or almost all families in a complex disease. It is possible that multiple families may carry different disease-causing insertion/deletions in the same gene, but this pattern was not seen. Thus, a second internal frequency comparison set of 500 samples that had a more similar bioinformatics pipeline to the IA samples sequenced in this study (i.e. use of GATK Unified Genotyper for variant calling) was used for manual review in combination with IGV visual inspection for the 26 indels remaining after application of biological filters. Manual review as described in the Methods excluded all but one of the 26 indels, demonstrating that manual inspection and use of an internal dataset generated by a similar bioinformatics pipeline are critical for reviewing insertion/deletions in sequencing experiments. Future studies may also consider utilizing newer local re-assembly-based methods for variant calling, such as FreeBayes[117] or GATK's HaplotypeCaller, which may improve the accuracy of insertion/deletion calls.

Considerations for functional predictions of exonic variation

More severe amino acid substitutions are more likely to present clinically,[31] so most WES studies to date have focused on non-synonymous SNVs and insertion/deletions. In this study, we also opted to focus on these variants, as predicted by the intersection of the three gene databases (RefSeq, UCSC, and Ensembl). Future studies focused on exonic variation could also study the effect of synonymous variation, which has been shown to also play an important role in human disease.[118] At the time of this study, fewer validated tools existed to examine the role of synonymous variation in sequencing data.

In this study, several programs were used to measure the level of conservation of a locus and the predicted pathogenicity of a variant. The programs have varying degrees of sensitivity and specificity for different kinds of variants, particularly due to the use of different but not completely independent data sources when generating predictions.[79] The bioinformatics community is working to develop tools that will be able to better integrate information to provide a more informed pathogenicity prediction. One such tool, the CADD program,[46] was recently introduced but has not been applied to a large number of datasets. Since there are few published studies implementing CADD, we have conservatively removed only variants with a C-score <10, thus retaining variants that are predicted by CADD to be among the 10% most deleterious substitutions in the human genome.

Considerations for biological processes and pathways

The filtering schema did not employ any assumptions about biological processes or pathways. Variants were annotated for GO terms chosen for possible relation to IA formation; however, only two variants in the final candidate variant list (variants found in the genes *COL17A1* and *FOXM1*) had one or more of these GO annotations. While using GO annotations as a filter is a powerful method for narrowing a list of variants, such an approach would depend on the comprehensiveness of GO annotations, as well as the reliability of investigator-chosen GO terms. To avoid subjectivity in selecting biological processes or pathways, future studies with larger sample sizes should consider employing formal gene set enrichment analysis, which eliminates the need to choose pathways *a priori*. Even for smaller datasets, use of GO annotations may help determine which gene variants to pursue first in additional experiments to explore possibly causal associations between the variant and disease of interest.

Summary

This is one of the few studies published to date that apply WES in a cohort of well-characterized families densely affected with a common complex disease without an *a priori* focus on a particular pathway or genomic region. We have laid out many considerations for future WES studies in complex disease, including the use of pedigree and phenotypic data, defining gene and exon boundaries, sources for allele frequency estimates, proper interpretation of *in silico* functional predictions, the role of environmental factors in the determination of potentially

causal rare variants, and the possible utility of combining pathway information with sequencing data.

In this study, 68 rare exonic variants in 68 genes were identified. Of these genes, one gene (*TMEM132B*) was significantly differentially expressed in IA versus control tissue. Further studies are needed to confirm and explore the *TMEM132B* variant, as well as the possible contribution of the other 67 variants. Replication and/or meta-analysis with similar sequencing studies using larger sample sizes could be used to gather further evidence for specific genes on this list. Additionally, a subset of these variants, which can be prioritized through any of the methods discussed in this study, could be explored through functional studies in models where vascular phenotypes can be easily observed, such as zebrafish. Targeted gene editing, such as through the CRISPR-Cas system, could help test whether a given variant disrupts the normal functioning of the relevant gene and whether such a disruption leads to a phenotype of interest. Ultimately, such a model should also enable investigation of whether the disrupted phenotype can be rescued by reintroduction of the wild type allele or interference with the variant allele. For comprehensive exploration of the variants identified in this study, multiple methods of experimental validation may be necessary. This study represents a necessary first step in the evaluation of role of rare variants in a common complex disease. Further evaluation in other familial and sporadic samples, as well as multi-ethnic samples, will be essential

CHAPTER II: PARKINSON DISEASE

Introduction

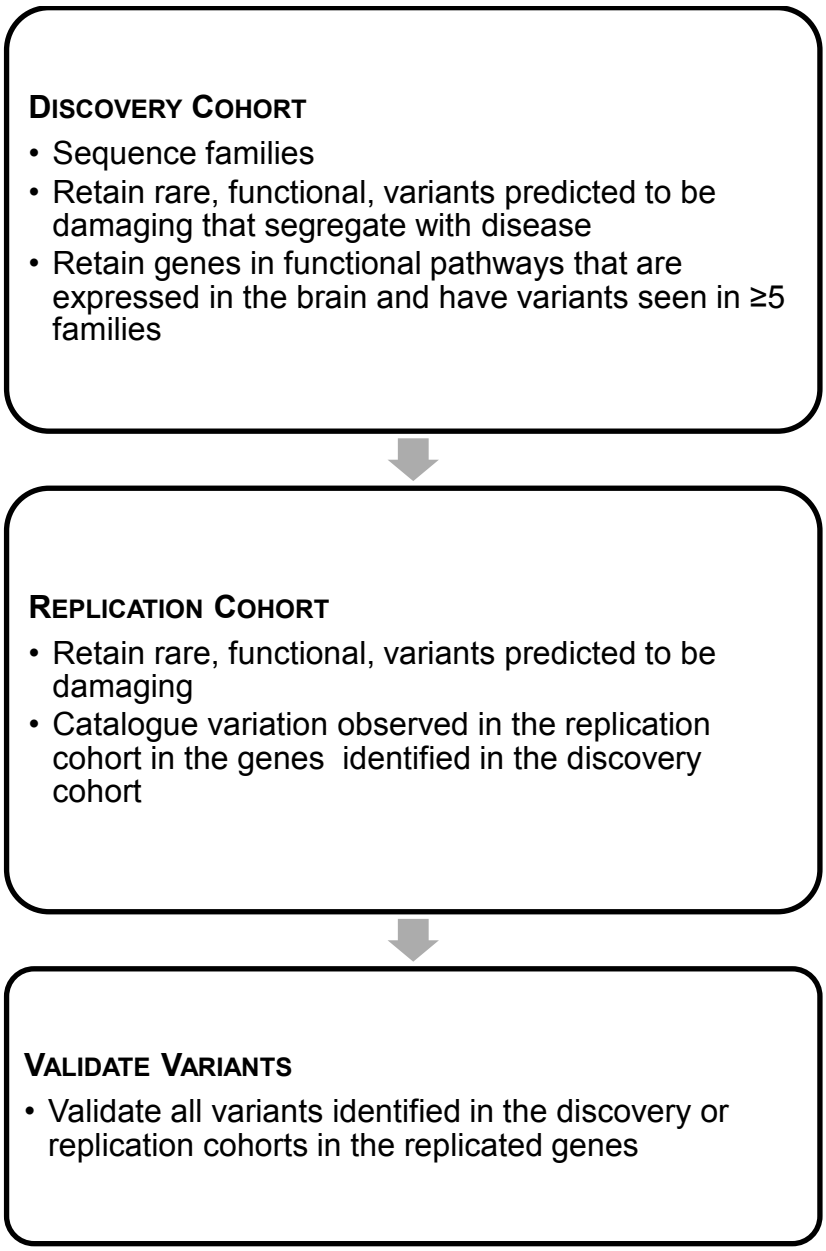
Parkinson's disease (PD) is a progressive neurodegenerative disease for which susceptibility is linked to genetic and environmental risk factors. Linkage studies have previously identified very rare variants in multigenerational families[119-123] that have a large effect on disease risk. Genome-wide association studies have recently revealed common loci that have relatively small individual effects on PD susceptibility.[113,124] Despite these advances, currently only about 6-7% of the heritability of PD has been explained.[125]

One approach to identify other potential genes and variants contributing to disease risk is through the analysis of personal genomes using high-throughput sequencing to highlight variants that exert a significant effect on disease susceptibility. The likelihood of detecting such variants can be enriched through the sequencing of PD patients with a family history of PD, who may be more likely to have a genetic contribution to disease susceptibility. Whole exome sequencing (WES) typically yields over 20,000 exonic single nucleotide variants (SNVs) per individual sequenced,[30] requiring a strategy to narrow the number of variants of interest. Successful approaches have included aggregative association tests in large samples of unrelated individuals[65,126,127] and filtering strategies within large and densely affected pedigrees[128,129] or consanguineous families.[130] Cohort studies of unrelated individuals are potentially limited by the

cost of WES in large numbers of subjects. While family-based strategies facilitate variant prioritization based on allele sharing and segregation, they are potentially insensitive to incompletely penetrant variants, intra-familial heterogeneity, and oligogenic inheritance, all of which are considerations in complex genetic disorders such as PD.[131]

WES in PD has been reported in studies involving one or a few families[128,129,132-135] or in candidate gene investigations.[136] In this report, we sequenced exomes from a discovery cohort of 93 individuals in 32 multiplex PD families. We then analyzed the genes with variants of interest in a replication cohort of familial PD probands to identify a subset of candidate genes containing rare, potentially functional variants that may contribute to disease risk (Figure 13).

Figure 13. Parkinson disease whole exome sequencing study design.



Materials and methods

Discover cohort subjects

The study protocol was approved by the Indiana University Institutional Review Board (IRB) as well as the ethics boards of all study sites. Families with at least one pair of living siblings diagnosed with PD were recruited and evaluated throughout North America by Parkinson Study Group (PSG) movement disorder neurologists. Written informed consent was obtained from all participants. Validated diagnostic checklists[137,138] implementing UK PD Brain Bank (UKPDBB) criteria, modified to allow for familial PD, were completed for all study participants. PD patients were classified as having either verified PD (VPD) or non-verified PD (NVPD). NVPD cases displayed clinical symptoms similar to PD but either failed to meet all UKPDBB inclusion criteria or met at least one of the exclusion criteria. All study subjects were offered the opportunity to participate in a brain-only autopsy program. Peripheral blood for DNA extraction was obtained from all consented individuals.

Exome sequencing of discovery samples

WES was performed on 32 families with the largest number of VPD cases, without another segregating neurological disorder, and without a known causative PD mutation in *LRRK2* or *parkin*. Among the 32 families were 90 subjects meeting criteria for VPD who were included for sequencing. An additional 3 individuals who were initially classified as NVPD were also included. Two of these individuals had subsequent neuropathological findings consistent

with PD. The third subject met all PD clinical inclusion criteria (including onset after 20 years of age, bradykinesia, persistent asymmetry, diagnosis by a movement disorder neurologist) and had significant supporting criteria (including rigidity, postural instability, a resting tremor, disease progression, and a positive response to levodopa) but met the solitary exclusion criterion of having concomitant Alzheimer disease and sensory deficits. For the purposes of subsequent analyses, this individual was considered to be affected. Of the 32 families, 6 families had 2 affected members sequenced, 23 families had 3 affected members sequenced, and 3 families had 4 affected members sequenced.

All samples were sequenced at one of two centers (44 samples representing 15 families at the Center for Inherited Disease Research [CIDR], and 53 samples representing 18 families at the HudsonAlpha Institute for Biotechnology). One family with 4 members was sequenced at both centers for quality assurance. The Agilent SureSelect 50Mb Human All Exon Kit (CIDR) and Nimblegen 44.1Mb SeqCap EZ Exome Capture version 2.0 (HudsonAlpha) were used for capture, and the Illumina HiSeq 2000 system was used for 100bp paired-end sequencing.

For the one family sequenced at both centers, summary sequencing statistics were compared to assess quality. Because two different captures were used, statistics were calculated only for those loci targeted by both capture kits. The intersection of variants found at both sequencing centers (32,280) and those sets

of SNVs found only at one center (3,061 for CIDR and 640 for HudsonAlpha) were examined. For SNVs found at the intersection, the genotype concordance rate was 99.0%, transition/transversion ratio was 3.0, and 99.0% were found in dbsnp137. Of those SNVs only identified at CIDR, the transition/transversion ratio was 2.4, and 98.3% were found in dbSNP137. For those SNVs found only at HudsonAlpha, the transition/transversion ratio was 1.8, and 86.1% were identified in dbSNP137.

Samples were aligned to the human genome reference sequence (build hg19) using Burrows Wheeler Aligner[43]. The Genome Analysis Toolkit (GATK)[44] was used for local realignment, base quality score recalibration, and multi-sample variant calling (Unified Genotyper) for the samples sequenced at CIDR and HudsonAlpha separately. GATK Variant Quality Score Recalibration[102] and recommended GATK training sets (i.e. HapMap 3.3 and Illumina Omni 2.5M chip sites, available from GATK bundle 1.2) were used to create a high-quality set of variant calls.

Annotation

ANNOVAR[54] was used to annotate high quality variants for predictions of variant location and function (using the RefSeq and UCSC databases). Custom scripts annotated variants for their allelic frequency in 1000 Genomes European American populations (2012 release, http://www.1000genomes.org)[36], Exome Sequencing Project (ESP) European American populations (5400 exomes

release, http://evs.gs.washington.edu/EVS/)[103], and dbSNP 137 (http://www.ncbi.nlm.nih.gov/SNP/)[139]. Allele frequencies were also obtained from an internal frequency database of 283 unrelated samples without a known neurological phenotype sequenced at CIDR. SIFT[52], Polyphen2[50], MutPred[49], and Gerp[48] were used to predict mutation deleteriousness and degree of locus conservation. Custom scripts annotated genes that fell within Gene Ontology (GO) (http://www.geneontology.org)[104] categories of interest (GO:0042417 dopamine metabolic process, GO:0050780 dopamine receptor binding, GO:0007270 neuron-neuron synaptic transmission, GO:0050804 regulation of synaptic transmission, GO:0007212 dopamine receptor signaling pathway, GO:0004952 dopamine receptor activity, GO:0006511 ubiquitin-dependent protein catabolic process, GO:0006979 response to oxidative stress, GO:0016567 protein ubiquitination, GO:0031396 regulation of protein ubiquitination). Genes were determined to be expressed in the brain based on significant expression above the background, as computed and normalized across Allen Brain Institute samples, following the Allen Human Brain Atlas protocols (http://www.brain-map.org, downloaded 05/17/2012).

Filtering

Variants were retained if they were: 1) predicted to be SNVs or insertion/deletions (indels) in an exonic and/or splicing region based on one or more gene databases; 2) had an allele frequency <3% in European American reference populations in 1000 Genomes and ESP, as well as in the internal

frequency database; 3) were predicted damaging by at least one *in silico* protein functional and structural effect prediction program or were located in a highly conserved region (Gerp>0.5); and 4) segregated with at least two PD cases in the same family. Genes were then retained if they: A) were in a GO category of interest; B) were expressed in the brain; and C) had retained variants that were observed in at least 5 of the 32 families sequenced.

Replication and variant confirmation

The prioritized genes were examined in a replication cohort of 49 unrelated PD patients with a family history of PD that had WES performed at the Human Genome Sequencing Center (HGSC) at Baylor College of Medicine (BCM) through the Baylor-Hopkins Center for Mendelian Genomics initiative. All individuals were clinically diagnosed with PD based on examination by experienced movement disorders neurologists and reported at least one first-degree relative diagnosed with PD. Written informed consent was obtained from all participants, and the study was approved by the BCM IRB. Preparation and sequencing of genomic DNA was performed as previously described in detail.[140] The BCM HGSC Core-developed library (VCRome 2.1)[141] was used for capture (covered all genes nominated from the discovery analysis at a depth of 50X or greater), and the Illumina HiSeq 2000 system was used for sequencing. With sequencing yields averaging 9.9 Gb per sample, samples achieved an average of 94% of the targeted exome bases covered to a depth of 20X or greater. Sequencing data were processed through the HGSC-developed Mercury pipeline

using the Atlas2 variant calling method[142,143] and annotated using the Cassandra annotation pipeline[144] based on ANNOVAR. Compared to the discovery pipeline, a more stringent allele frequency filter was employed—all variants considered in replication were <1% in European American reference populations (1000 Genomes and ESP). Potentially deleterious and highly conserved variants were identified using SIFT, Polyphen2, MutPred, and Gerp. Variants present in the 8 genes prioritized from the discovery analysis were extracted.

All variants in replicated genes were reviewed in the Exome Aggregation Consortium beta version 0.2 (ExAC, Cambridge, MA [http://exac.broadinstitute.org)] [November 17, 2014]) to ensure that allele frequencies obtained through the >60,000 exomes in ExAC corresponded to those obtained in 1000 Genomes and ESP 5400. Targeted PCR and Sanger sequencing were used to confirm all variants for genes with consistent evidence supporting links to familial PD in both the discovery and replication cohorts. Variants were annotated for C-scores from the Combined Annotation Dependent Depletion (CADD) webserver (http://cadd.gs.washington.edu),[46] where C-scores ≥10 correspond to the 10% most deleterious substitutions in the human genome, as predicted by CADD. Genes with evidence of replication were also annotated for residual variation intolerance score (RVIS) percentiles, in which lower percentiles correspond to genes that are most intolerant of functional mutations.[51]

Results

Discovery Cohort

Clinical characteristics of individuals from the 32 multiplex PD families in the discovery cohort are summarized in Table 7.

Table 7. Clinical characteristics of the Parkinson disease patients in the discovery and replication cohorts. 1 - Data not available for 2 of 93 cases. 2 - Data not available for 6 of 49 cases.

Clinical Characteristic	Discovery Cohort	Replication Cohort
Number of individuals (number of families)	93 (32)	49 (49)
% Female, % Male	47.9%, 52.1%	32.6%, 67.3%
Average age of onset (mean ± SD)	61.8 ± 9.97[1]	50.1 ± 15.7[2]
Ethnicity	90 self-reported, non-Hispanic, European Americans 3 self-reported, non-Hispanic Asians	37 self-reported, non-Hispanic European 8 individuals of Hispanic descent 3 self-reported, non-Hispanic Asians 1 individual of Middle Eastern descent

Each sample sequenced at CIDR achieved a mean coverage of 98X for targeted bases, with an average of 93% of targeted bases covered at least 8X. The transition/transversion ratio was 3.3, and 94.4% of variants were found in dbsnp137. The sequencing data achieved 99.6% concordance with OmniExpress GWAS array genotype calls performed on the same individuals. Each sample sequenced at HudsonAlpha achieved a mean coverage of 57X for targeted bases, with an average of 93% of targeted bases covered at least 8X. The average transition/transversion ratio per sample was 3.2, and 91.5% of variants were found in dbSNP137.

Application of GATK quality filters resulted in 149,055 SNVs and 9,378 indels across all samples (range of 22,188-28,230 total variants per sample) (Figure 14). Nonsynonymous SNVs or indels within an exon (as annotated by at least one of two gene databases, i.e. RefSeq, UCSC) having an allele frequency of <3% (1000 Genomes, ESP) were retained. After removing variants that were predicted to be benign by all three protein prediction programs and that were not in a highly conserved region, approximately 10% of the original variants remained. Further filtering was performed based on segregation within families, leaving 7,729 SNVs and 305 indels. Prioritization based on brain expression, GO annotation, and presence of variants in at least 5 families yielded 21 variants (21 SNVs, 0 indels) across 8 genes for evaluation in the replication cohort (Table 8).

Figure 14. Parkinson disease discovery cohort variant filtering. SNV = single nucleotide variant; MAF = minor allele frequency; GO = Gene Ontology

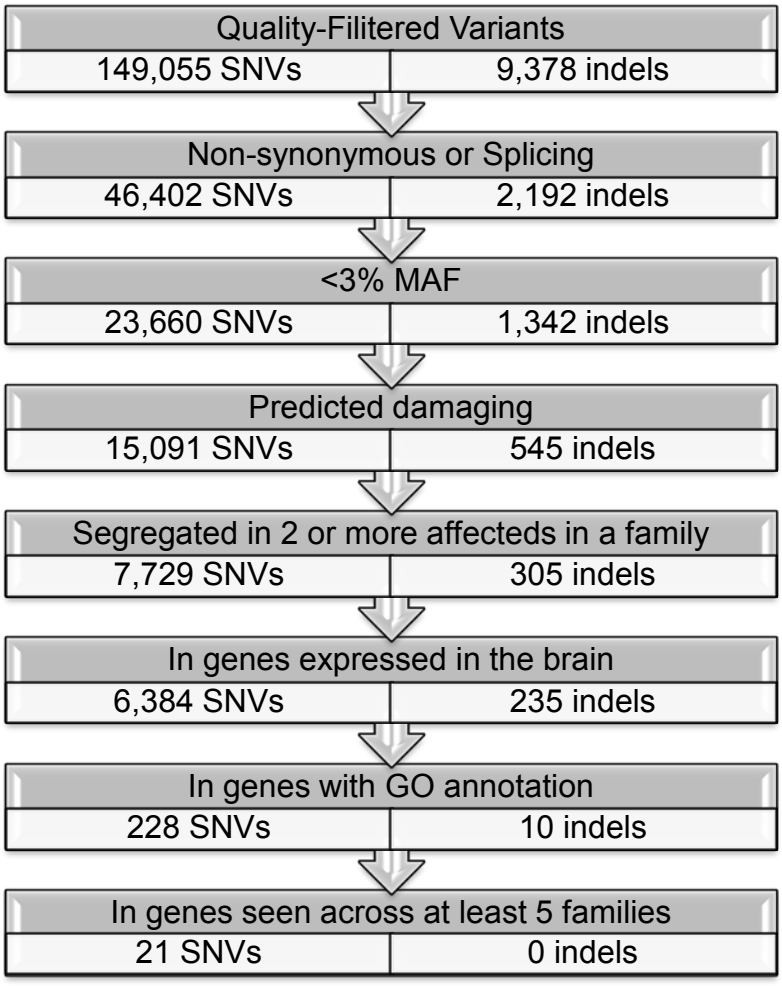


Table 8. Variants identified in the Parkinson disease discovery cohort. All variants are predicted to be nonsynonymous single nucleotide variants (SNV). Chr = chromosome, Ref = reference allele, Alt = alternate allele, ExAc Freq = total frequency in Exome Aggregation Consortium. *A single family shared two variants of interest (denoted by asterisks) in the same gene.

Gene Symbol	Chr	Position	Ref	Alt	ExAc Freq	Amino Acid Change	CADD score	Discovery Samples/ Families	Families with 2 or more members sharing variant
CBLC	19	45295664	A	G	0.0081	NM_001130852:c.A892G:p.M298V	17.0	8/4	2
	19	45296767	G	A	0.0045	NM_001130852:c.G1036A:p.E346K	13.0	3/2	1
CHAT	10	50824106	C	T	0.0095	NM_001142933:c.C8T:p.P3L	4.5	4/2	2
	10	50863188	G	A	0.0075	NM_001142929:c.G1328A:p.R443Q	29.1	4/2	1
KIF1B	1	10363664	G	T	0.018	NM_183416:c.G2421T:p.M807I	4.2	3/2*	1
	1	10363944	G	A	8.17e-06	NM_183416:c.G2701A:p.E901K	14.0	3/1	1
	1	10364260	A	G	0.0061	NM_183416:c.A3017G:p.E1006G	11.3	4/3*	1
MYLK2	20	30408306	C	G	0.013	NM_033118:c.C430G:p.P144A	23.8	7/5	1
	20	30407387	G	A	0.00070	NM_033118:exon2:c.G4A:p.A2T	23.8	2/1	1
TNK2	3	195594494	C	T	0.015	NM_005781:c.G2630A:p.R877H	21.7	11/4	4
	3	195595212	C	T	0.0026	NM_005781:exon12:c.G1912A:p.V638M	29.3	2/1	1
TNR	1	175355171	T	C	N/A	NM_003285:exon8:c.A1774G:p.T592A	24.2	2/1	1

	1	175372714	T	G	0.0043	NM_003285:c.A538C:p.N180H	24.5	2/1	1
	1	175375355	T	C	0.0044	NM_003285:c.A496G:p.T166A	13.6	5/3	1
	1	175375388	A	T	7.33e-05	NM_003285:c.T463A:p.C155S	22.0	2/1	2
TRIM56	7	100731638	C	T	0.028	NM_030961:c.C1045T:p.L349F	7.74	6/4	2
	7	100732376	G	A	0.0040	NM_030961:exon3:c.G1783A:p.A595T	10.78	3/1	1
TOPORS	9	32541880	G	C	0.0028	NM_001195622:exon2:c.C2448G:p.H816Q	12.1	3/1	1
	9	32542278	T	C	0.017	NM_001195622:c.A2050G:p.N684D	15.0	3/2	2
	9	32550896	G	C	0.0085	NM_005802:exon2:c.C74G:p.S25W	19.6	2/1	1
	9	32550953	G	A	4.50e-05	NM_005802:exon2:c.C17T:p.P6L	22.9	2/1	1

Replication cohort

Clinical characteristics of the 49 familial PD probands in the replication cohort are summarized in Table 7.

Rare variants predicted to be damaging in the 8 genes prioritized in the discovery analysis were extracted from WES data for the replication cohort. Three genes (*KIF1B, TNK2,* and *TNR*) that harbored variants of interest (as defined in the Methods) in the discovery cohort were also found to have variants of interest in the replication cohort (Table 9). One variant in *KIF1B* (p.E1006G) was observed in both discovery and replication samples (3 and 1 samples, respectively). For the other 2 replicated genes, distinct variants were identified in the discovery and replication cohorts.

Table 9. Parkinson disease candidate genes identified through whole exome sequencing. RVIS=Residual Variation Intolerance Score (lower percentile corresponds to more mutation intolerant genes). *Asterisks indicate variants that were seen in both discovery and replication cohorts.

Gene Symbol	Gene Name	Map Location	Gene Ontology	Transcript Size (base pairs)	Genic Intolerance (RVIS Score Percentile)	No. of Variants/ Families in Discovery Cohort	No. of Variants/ Families in Replication Cohort
KIF1B	kinesin family member 1B, transcript variant 2	1p36.2	neuron-neuron synaptic transmission	7,680	3.93%	3/5	1/1*
TNK2	tyrosine kinase, non-receptor, 2	3q29	protein ubiquitination	4,476	14.28%	2/5	2/2
TNR	tenascin R (restrictin, janusin)	1q24	neuron-neuron synaptic transmission; regulation of synaptic transmission	5,190	20.04%	4/6	1/1

In total, the 3 replicated genes were observed to harbor 12 distinct potentially functionally relevant variants (Table 10). All 12 variants were confirmed by targeted PCR and Sanger sequencing in all relevant samples, and allele frequencies obtained from ExAC corresponded to those from 1000 Genomes and ESP. Genic intolerance RVIS score percentiles[51] and CADD scores[46] were computed for the retained genes and variants in order to further characterize the potential impact of functional mutations at the gene- and variant-level respectively. The 3 replicated genes had a mean RVIS score percentile of 12.8% (SD=8.2%), and the mean CADD score for the 12 retained variants was 20.4 (SD=8.7). The calculated RVIS score percentiles fall within ranges that reflect purifying selection, or probable greater intolerance for mutations within the gene than most genes. Similarly, the computed CADD scores, with the exception of one variant (p.M807I in *KIF1B*), place all of the 12 retained variants at or above the 10% most deleterious variants in the human genome, as predicted by a comprehensive range of predictions used in the CADD algorithm.

Table 10. Variants identified in the Parkinson disease candidate genes. Chr = chromosome, Ref = reference allele, Alt = alternate allele, ExAc Freq = total frequency in Exome Aggregation Consortium. *Asterisk indicates variant that was seen in both discovery and replication cohorts.

Gene Symbol	Chr	Position	Ref	Alt	ExAc Freq	Exonic Prediction	Amino Acid Change	CADD score	Discovery Families / Replication Probands
KIF1B	1	10363664	G	T	0.018	nonsynonymous SNV	NM_183416:c.G2421T:p.M807I	4.2	2/0
	1	10363944	G	A	8.17e-06	nonsynonymous SNV	NM_183416:c.G2701A:p.E901K	14.0	1/0
	1	10364260	A	G	0.0061	nonsynonymous SNV	NM_183416:c.A3017G:p.E1006G	11.3	3/1*
TNK2	3	195594092	G	A	5.83e-05	nonsynonymous SNV	NM_005781:exon13:c.C2930T:p.A977V	17.8	0/1
	3	195594494	C	T	0.015	nonsynonymous SNV	NM_005781:c.G2630A:p.R877H	21.7	4/0
	3	195595212	C	T	0.0026	nonsynonymous SNV	NM_005781:exon12:c.G1912A:p.V638M	29.3	1/0
	3	195605390	A	G	2.45e-05	nonsynonymous SNV	NM_005781:exon8:c.T1088C:p.V363A	25.9	0/1
TNR	1	175355171	T	C	N/A	nonsynonymous SNV	NM_003285:exon8:c.A1774G:p.T592A	24.2	1/0
	1	175355213	G	A	N/A	stopgain SNV	NM_003285:exon8:c.C1732T:p.R578X	36.0	0/1
	1	175372714	T	G	0.0043	nonsynonymous SNV	NM_003285:c.A538C:p.N180H	24.5	1/0
	1	175375355	T	C	0.0044	nonsynonymous SNV	NM_003285:c.A496G:p.T166A	13.6	3/0
	1	175375388	A	T	7.33e-05	nonsynonymous SNV	NM_003285:c.T463A:p.C155S	22	1/0

Discussion

Using WES in a discovery and replication cohort of familial PD patients, we detected 12 likely deleterious, rare, exonic variants in 3 genes (*KIF1B*, *TNK2*, and *TNR*) that may play a role in susceptibility to PD. All variants were found in the heterozygous form, suggesting that they are inherited in a dominant manner, as expected from the pedigree structures of the families sequenced, and may lead to a disease phenotype either through a gain-of-function, haploinsufficiency, or a dominant negative mechanism.

KIF1B, or kinesin family member 1B, is a gene on 1p36.2 that encodes a motor protein that transports synaptic vesicle precursors and mitochondria.[145-148] Mutations in *KIF1B* were linked with Charcot-Marie-Tooth disease type 2A (CMT2A);[145] however, a recent study more conclusively implicates the nearby *MFN2* gene in CMT2A.[149] Three rare, nonsynonymous variants in *KIF1B* were found in this study, including one variant (p.E1006G) that was present in both the discovery and replication cohorts. The variants do not overlap known *KIF1B* protein domains in Ensembl, although they do all cluster on the most 3 prime coding exon of the gene. Further work is needed to confirm the effects of these variants on protein structure and/or function.

TNK2 encodes for a non-receptor tyrosine kinase (activated CDC42 kinase 1) that is important for cell growth, survival, and migration. Studies suggest that *TNK2* is involved in synaptic function and plasticity,[150-152] and a recent report

suggests that mutations in the gene may cause autosomal recessive infantile onset epilepsy.[153] Other studies exploring the role of *TNK2* in cancer have established links between the *TNK2* protein and the epidermal growth factor receptor (EGFR).[154,155] In the discovery and replication cohorts, 4 unique rare nonsynonymous *TNK2* variants were identified. One variant (p.V363A) is found in the EGFR inhibitor Mig-6 domain (IPR021619, PF11555). Binding of Mig-6 to the kinase domain of EGFR inactivates the receptor, which suggests that this domain in the *TNK2* protein may also be important for appropriate regulation of its function.

TNR, or tenascin R, encodes an extracellular matrix glycoprotein only found in the central nervous system.[156] Tenascin R is thought to be involved in neurite growth, neural cell adhesion, and sodium channel functioning.[157,158] Of the 6 unique variants prioritized in *TNR*, 5 were found only in the discovery cohort as rare nonsynonymous variants. One variant (p.R578X) was found only in the replication cohort and results in the addition of a stop site at position 578 of a 1358 amino acid protein. This variant, along with one other variant (p.T592A), are found in the fibronectin-3 domain (IPR003961) of the protein, which is important for cell surface binding.

Variants from the discovery analysis present in the 5 genes lacking evidence of replication (*CHAT, CBLC, MYLK2, TRIM56,* and *TOPORS*) are listed in Table 8. While some or all of these genes may represent false positives, differences

between the discovery and replication analysis (captures, sequencing chemistries, bioinformatics pipelines, allele frequency threshold, etc.) may have prevented replication. Additionally, some genes may not have been prioritized to look for evidence of replication due to differences between the two captures and other possible batch effects in the discovery phase limiting the effectiveness of the across families filter.

Unlike previous studies focused on a single large pedigree or extensive datasets of unrelated individuals, our blended approach leveraged a well-characterized set of moderately-sized families and an additional set of unrelated familial probands. A major advantage of this study is that both the discovery and replication cohorts only included familial PD cases, unlike many other studies where discovery samples are in families and replication cohorts include sporadic cases. Families with multiple affected members are more likely to be enriched for causative, moderately rare variants having a modest or large effect size. By requiring variant segregation within a family, we limited the number of false positives in the discovery phase. Furthermore, our two-phase study design decreases the chance of false positives and thus increases the likelihood that the 3 candidate genes identified in this study are truly involved in PD etiology, though further replication in other datasets is warranted.

Another strength of the study is that locus heterogeneity could be explored both within and between families. In the discovery analysis, we required variants to

segregate with at least two PD cases in a family, thereby allowing any remaining cases in the family to potentially have a distinct genetic or environmental cause. Our experimental design contrasts with recent efforts that employ sequencing approaches in large family pedigrees to identify variants with fully penetrant effects and therefore responsible for strictly Mendelian PD; this category of variants appears to account for rare causes of PD.[128,129,133] While our study design allows for detection of such mutations, the employed strategy also permits the discovery of rare variants with intermediate penetrance, such as *LRRK2* G2019S[159] and mutations in *GBA*.[160] Since 10-20% of PD patients report having at least one first-degree relative affected by PD,[161-163] it is possible that variants of this class remain a major contributor to PD heritability. Complex genetic etiology has been previously observed in PD; for example, reports have shown that in some families segregating Mendelian forms of PD (*SNCA* or *LRRK2* positive families), not all affected family members carry a mutation.[121,164,165] Our study is also robust to detect interfamilial allelic heterogeneity, or unique variants in the same gene segregating in different families.

One limitation of our approach was that larger genes might be prioritized by chance because of their size rather than due to the enrichment of rare functional variants associated with PD. Exome sequencing by design also misses possibly important variation in intronic and regulatory regions, as well as forms of structural variation. Use of the GO filter to focus on pathways of interest might have excluded important genes that were either poorly annotated or in pathways

thus far not associated with PD. The GO filter used, as seen in Figure 14, narrowed the number of variants under consideration from 6,635 to 228 SNVs, ultimately prioritizing 21 variants across 8 genes for further study. Had the GO filter not been applied, the 6,635 SNVs would have only been narrowed to 300 SNVs (87 genes) using the across families filter. Future studies with larger sample sizes could employ formal gene set enrichment analysis to bypass the potential limitation of relying on pre-specified pathways for variant filtering.

Summary

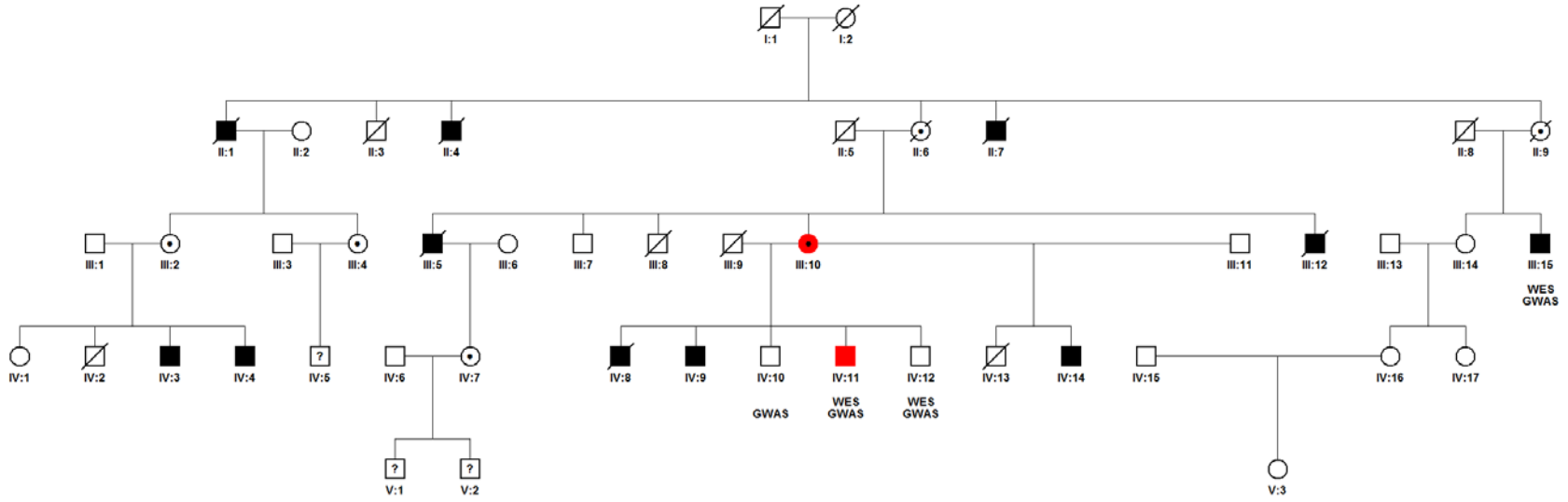
In summary, we employed a two-stage strategy to identify and replicate genes that may harbor rare variants contributing to PD susceptibility. Both the discovery and replication samples were comprised of familial PD patients, who may be more likely to segregate relatively rare variants of larger effect on disease risk. The 3 genes nominated in this study warrant further evaluation for their potential role in PD pathophysiology.

CHAPTER III: X-LINKED ATAXIA DEMENTIA

Introduction

X-linked ataxia dementia (XLAD), also known as X-linked spinocerebellar ataxia type 4, is an extremely rare neurodegenerative disorder. During their childhood, affected individuals develop ataxia, or uncoordinated movement. Dementia occurs later in life, along with variable onset of upper motor neuron disease. Increasing motor, emotional, and mental instability occurs throughout the second through fifth decades of life, with death typically in the sixth decade. Moderate phenotypic variability is observed in affected males, and carrier females sometimes show a milder phenotype including cognitive and motor abnormalities. The disease appears to segregate in an X-linked pattern in the one kindred ever described with this syndrome (Figure 15). Clinical and laboratory investigations in this family indicate cerebellar and pyramidal system involvement with severe cerebral cortex deficiencies.[166]

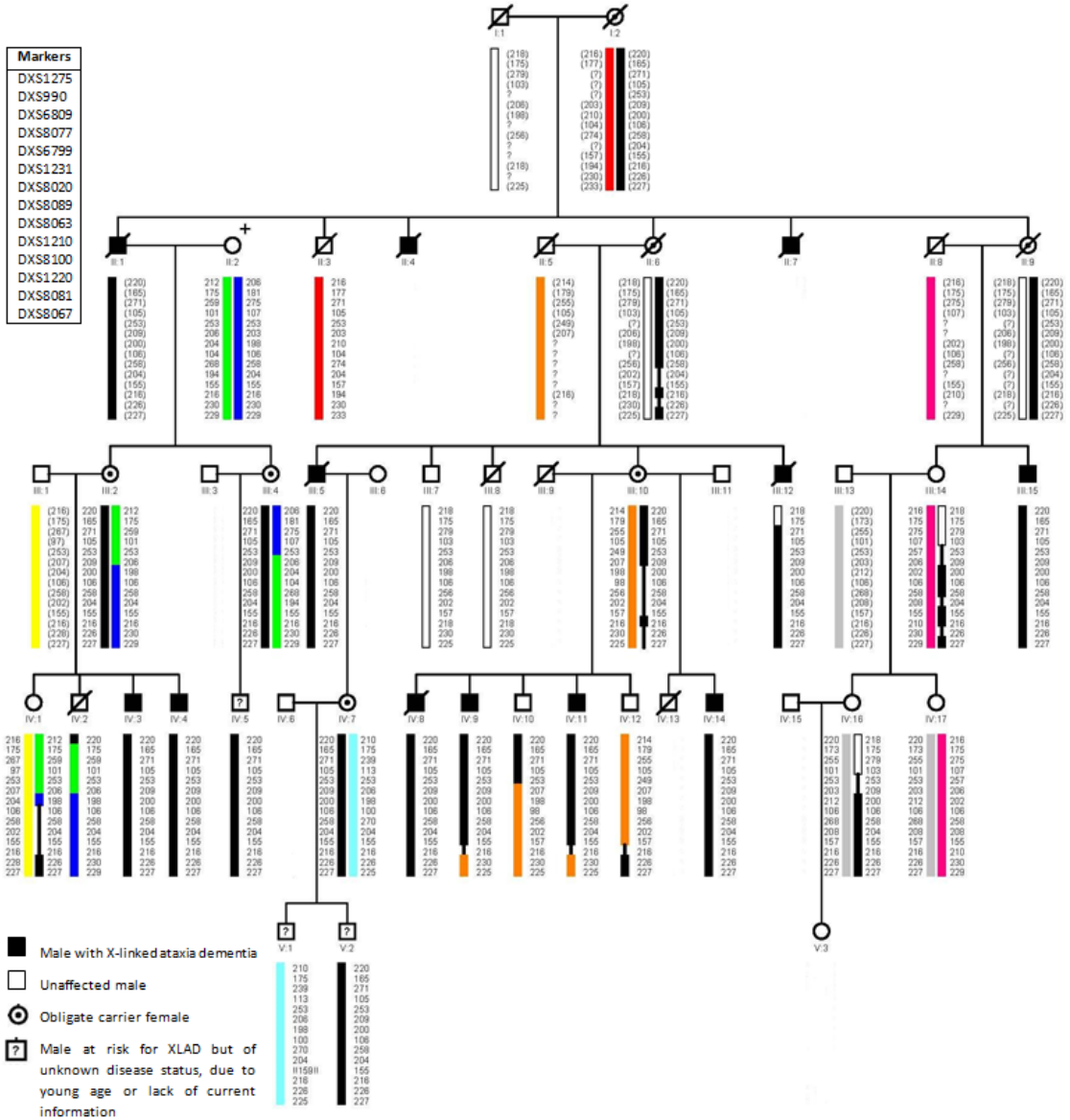
Figure 15. Simplified pedigree for the X-linked ataxia dementia family. Whole genome sequencing was performed on individuals in red. WES = whole exome sequencing; GWAS = Omni1-Quad Genome Wide Association Study Array; ? = male at risk for XLAD but of unknown disease status due to young age at time of assessment



Other ataxia-dementia syndromes such as olivopontocerebellar atrophy, Gerstmann-Straussler-Sheinker disease, and adrenoleukodystrophy were all considered but ruled out in this family. Two other reports exist of ataxia and dementia both segregating in a X-linked manner, but affected members of those families have compounding extrapyramidal symptoms. There are also X-linked syndromes displaying either ataxia or dementia but not both reported in the literature.[167,168]

Previous linkage studies using microsatellite markers on the X chromosome were conducted (unpublished data). In the most recent study, microsatellite markers at approximately 5 cM intervals were used, all females were classified as unaffected, and penetrance was set at 95%. After genotyping of a second set of microsatellite markers to narrow the interval, a LOD score of 5.29 was obtained in the region Xq21.33-q23. To rule out fragile X-associated ataxia, individuals III-12, IV-8, and IV-11 were also tested for fragile X using PCR; all 3 samples were normal. (Figure 16)

Figure 16. X-linked ataxia dementia family structure and haplotype analysis. Analysis and figure generation by Jill Rosenfeld (unpublished data). Parentheses indicate inferred genotypes. The regions of a narrowed bar for individuals II-6, III-10, and III-14 indicate unknown phase, while all other narrowed bars indicate regions of recombination.



Further genotyping of two distantly related affected males (Figure 15, III:15, IV:11) and two unaffected males (IV:10, IV:12) on an Illumina Omni1-Quad array narrowed the region of interest to 19Mb and ruled out a large CNV. Over 100 genes are contained in this interval.

WES was conducted for 3 individuals (III:15, IV:11, IV:12) in a previous study (Agilent SureSelect Human X Chromosome Demo Kit, 75bp paired-end sequencing, Illumina GAIIx, BWA, SAMtools). Five SNVs were identified that were present on the disease haplotype but not present in dbSNP. Three of these variants were identified either in the pilot 1 dataset for the 1000 Genomes Project or in the genotyping results from 2000 female controls; due to the rarity of the disease, it was hypothesized that any causative variant would be completely novel, and thus the 3 variants were excluded from further study. The remaining two SNVs did not have an obvious mechanism of disease causation. Four short indels also were identified on the disease haplotype, but all were present in the 1000 Genome Project. (Table 11)

Table 11. X-linked ataxia dementia whole exome sequencing variants on the disease haplotype and not present in dbSNP.

Chr	Position	Ref	Fam	Gene	FunctionGVS	1000 Genomes	# control chrs
X	99992932	G	A	NOX1	coding-synonymous	0/182	2/4000
X	102833316	G	A	none	intergenic	2/182	-
X	105083643	A	T	NRK	intron (39 bp from exon)	0/182	0/4000
X	102932576	G	A	PLP1	utr-3	1/182	17/4000
X	107221681	C	G	ATG4A	utr-5 (31 bp from start codon)	0/182	0/4000

In the present study, 2 individuals (III:10, IV:11) were chosen for WGS in order to expand the search space for rare variants that may be involved in causing the disease.

Materials and methods

Subjects

The simplified pedigree for the XLAD family is depicted in Figure 15. All subjects submitted written consent, and the study was approved by the Indiana University IRB.

The Agilent SureSelectXT2 Library Prep Kit and Illumina HiSeq2000 (Flowcell v3, TruSeq Cluster Kit v3, TruSeq SBS v3) were used to generate 100 bp paired end sequencing data. Paired end alignment was performed to the GRCh37 reference genome with BWA v.0.5.10, and duplicates were marked using Picard v.1.74.

GATK v.2.3-4 was used for indel realignment, base call quality score recalibration, multi-sample variant calling (Unified Genotyper), and VQSR. Bedtools v.2.19.1 was used for coverage analysis of coding exons (defined by UCSC coordinates) within the region of interest. Variants were annotated by ANNOVAR and the recently-developed CADD program.

High-quality variants located in the region of interest that were novel (not present in 1000 Genomes, the Exome Sequencing Project, and dbSNP137) and located in the region of interest were retained. All variants identified in an exonic, splicing, UTR, or regulatory region were retained. Variants of interest were visually inspected using IGV v.2.3.34. The entire region was also visually inspected using IGV for alignment issues (regions of soft-clipping, unmapped pairs, abnormal pair orientations) that may point to a small to medium-sized structural variant.

Results

The average transition/transversion ratio for exonic variants and all variants was 3.10 and 2.12 respectively, and the percentage of variants found in dbSNP137 was 98.98%. Mean autosomal coverage was 27X, and 97.5% of autosomal regions were covered >8X.

Coverage analysis revealed that very few small intervals of coding exons within the region of interest were not covered at all or had low coverage (Table 12).

Table 12. X-linked ataxia dementia whole genome sequencing coverage analysis. DP = sequencing depth; bp = base pairs

Sample	No Coverage Total [Average Interval]	Low Coverage (DP<5) Total [Average Interval]
III:10	15bp [5bp]	12bp [6bp]
IV:11	48bp [24bp]	261bp [15bp]

There were 7,901 variants identified in the region of interest. Of these, 5,798 were heterozygous in the mother and hemizygous in the son, and 505 of these were not observed in 1000 Genomes, the Exome Sequencing Project, and dbSNP137. Variants were retained if they were found in an exonic, splicing, UTR, or regulatory region. Using these criteria, 22 variants (3 SNVs and 19 indels) were retained, all of which were located in UTR (Table 13) or regulatory regions.

Table 13. X-linked ataxia dementia whole genome sequencing variants identified in untranslated regions (UTR). Variants listed by decreasing CADD c-score value. *All variants are in the 3'-UTR except MORF4L2 (in 5'-UTR). Chr = chromosome, Pos = position (build hg19), Ref = reference allele, Alt = alternate allele, SNV = single nucleotide variant, DEL = deletion, INS = insertion

Chr	Pos	Ref	Alt	Type	Gene	C-Score
X	103045920	G	A	SNV	PLP1	10.63
X	102930671	TA	T	DEL	MORF4L2	9.853
X	100350299	TGC	T	DEL	TMEM35	4.265
X	106313096	TA	T	DEL	RBM41	3.451
X	101913597	TA	T	DEL	GPRASP1	0.534
X	105881497	C	CT	INS	CXorf57	0.325
X	102941052	TC	T	DEL	MORF4L2*	0

The PLP1 (associated with Pelizaeus-Merzbacher disease, which includes childhood ataxia and cognitive impairment) variant was previously identified in the WES experiment but was ruled out when genotyped in control chromosomes. Further examination at CIDR identified the same variant in two males affected with a common disease sequenced with the same capture as the XLAD subjects. All other variants have not to date been identified in other datasets.

Upon visual inspection with IGV, no regions were found with an obvious structural variant shared in the mother and son.

Discussion

The magnitude of the linkage signal at Xq21.33-q23 strongly suggests that the variant that causes XLAD exists in this interval. Both WES and WGS, however, have not identified an exonic, and putatively functional variant within this region that segregates in a X-linked recessive manner. Together, the WES and WGS experiments suggest that adequate coverage has been achieved over the entire region of interest (Table 12). There may be variation within the interval however that is difficult to assess with sequencing methods (e.g. areas of repetitive sequence). Study designs not reliant on the current methods of sequencing would be required to detect these variants.

A number of novel variants in predicted non-coding regions of the X chromosome agree with the hypothesized segregation pattern (Table 13). Although the severity of phenotype suggests that the causative variant is within a coding region, a non-coding variant may also lead to the disease. For instance, changes in promoter regions or enhancers can affect gene transcription, while UTR sequence alterations can influence the regulation of translation. Possible links between both 5' and 3'-UTRs and diseases including X-linked Charcot-Marie-tooth disease, Fragile X syndrome, epidermolysis bullosa simplex, and a number of other diseases have been suggested.[169] Targeted mutagenesis and subsequent examination of translation efficiency could be utilized to study the variants in Table 13. As noted by Ward and colleague however,[170] landmark studies linking non-coding variants to some diseases required extensive

experimental follow-up, and rigorous study will be required to confirm the association of any variants identified in this study with XLAD pathogenesis.

Another possible reason for a lack of a positive exonic finding thus far may be due to limitations in current methods and data sources for annotation. Thus, a variant nominated in the WGS data actually may be a coding variant that has not been assigned to a gene yet. As annotation sources improve over time, periodic re-examination of the WGS data using the existing pipeline is warranted. A review of annotations for non-coding variants could be relevant as well, especially as systematic efforts such as the ENCODE Project[171] and the Roadmap Epigenomics Mapping Consortium[172] continue to release data.

This work also has not conclusively ruled out a structural variant as a cause for the disease. At the time of the study, algorithms to effectively detect and conclusively call medium-sized structural variants were still in development. Because automated methods for detecting structural variation are not yet optimized, we manually reviewed the entire 19 Mb region of interest for evidence that a structural variant might be present (as described in the Methods), but found no signs at this time that there is a structural variant shared in the mother and son. Future studies in XLAD could focus on structural variants, since variants such as simple repeat expansions have been clearly linked with several neurological diseases, including Fragile X[173-175] and Friedreich's ataxia.[176] Emerging bioinformatics tools for application to HTS could be used, although

methods other than sequencing might be warranted due to the difficulty of designing sequencing baits and accurately calling sequencing variants in highly repetitive regions. Another option for future study is to investigate whether the insertion of novel sequence could lead to the disease. This type of variation would require a very different analysis pipeline, most likely including computationally-intensive *de novo* assembly or even different sequencing options (e.g. selecting a technology that will produce longer read lengths).[177] Additional members of the family could be screened for identified candidate structural variants, and molecular studies to characterize the potential role of segregating variants should be conducted.

Summary

A series of genetic study designs have been applied in a single kindred segregating a rare neurodegenerative disease. Initial linkage studies pointed to a strongly significant interval on the X chromosome, but genome-wide genotyping and WES failed to identify a promising candidate gene. Further WGS has not identified a clearly causative variant for this region. A few variants in UTR and regulatory regions are possible candidates for further validation and exploration of involvement in disease causation. Although preliminary visual inspection was conducted of the region of interest, further work remains to identify structural variants. Additional alignment and calling algorithms can be used to circumvent potential biases in the current bioinformatics pipeline used.

CONCLUSION

Within the field of family-based sequencing, there are many study designs that can help elucidate the genetic basis of both rare and complex diseases. In the present study, WES was applied to both familial IA and PD, and both WES and WGS were employed to study a family with XLAD. For IA, we used WES of a small set of densely affected families to describe considerations for other WES studies in complex disease, including use of family and clinical data, sources and definitions for gene and variant annotations, interpretation of *in silico* predictions, and more. Our PD WES study was a two-stage design, blending the use of moderately-sized families and an independent set of familial probands that allowed for the exploration of locus heterogeneity within and between families. The XLAD study presented an opportunity to compare and combine WES and WGS results, and although a definitely causative gene was not identified, important groundwork has been laid for future studies. As HTS technology and analysis methods improve and decrease in cost and labor intensiveness, WGS will likely supplant WES due to lower bias and broader coverage.[30] Thus, experience with applying this technology to families will become increasingly important.

Advantages of family-based sequencing studies

There are several advantages for using familial data for sequencing studies. First, such studies are enriched for samples that are actually linked by a genetic

cause of disease[178,179] and can control for type 1 error rates due to population stratification.[180] Given the difficulty of narrowing down the enormous number of variants identified in WES and WGS, sequencing multiple individuals per family can dramatically aid the filtering process as demonstrated in the current work. Furthermore, family-based sequencing can somewhat offset the expense of following-up a large number of candidate variants, since putative causative or protective alleles can first be genotyped in other family members to confirm or refute segregation. Such segregation studies will be an obvious next step in the follow-up of the variants identified in our IA, PD, and XLAD families. Unexpected patterns of segregation can be checked against possible locus or allelic heterogeneity, or even environmental causes for the disease. Furthermore, although not explored in the present work, homozygosity mapping of familial sequencing data can also be an effective method for determining the genetic basis of an autosomal recessive disease.[30]

Stringent quality control measures for HTS are critical, and family-based sequencing studies have the benefit of additional sources for quality metrics.[67,178] Careful examination of expected and computed pedigree structure can verify that samples are labeled correctly and can identify cryptic relatedness. It is not uncommon to find individuals related to one another in the same study, even for the larger studies being conducted with ‘unrelated’ cases; such relatedness can easily confound the results of even a well-designed study. Additionally, data from multiple family members can improve variant calling, especially for structural

variation. In fact, some variant detection algorithms like PennCNV,[181] FamSeq,[182] PolyMutt[183] already make use of familial information. Since many types of structural variants have not been ruled out for all our IA, PD, and XLAD families, future studies should utilize these tools and emerging ones. As was mentioned for XLAD, certain types of structural variation have not been completely ruled out as the causative mutation in the family.

Caveats for family-based sequencing studies

While there are many advantages to using familial sequencing data, there are some important caveats as well. If using statistical association tests, particular care must be taken to account for relatedness, or special algorithms designed to incorporate pedigree information should be used.[184,185] Since increased computational resources are required for incorporating pedigree information, these programs have been slower to develop. For these tests or for manual filtering, as was used in the studies presented in this work, researchers must be aware that assumed inheritance models may not actually reflect genuine allelic inheritance. Studies may be broadly designed to examine multiple inheritance models, as was done for the FIA and PD studies presented.

Great attention must be paid toward careful phenotyping before assigning strict inheritance hypotheses, especially for complex diseases. As demonstrated in the FIA study, these considerations are important for designating affected and unaffected status, as well as assessing for reduced penetrance and

heterogeneity. For instance, careful examination of smoking and hypertension data for individuals is critical when studying IA genetics, given the important contribution of both environmental risk factors to the disease.[86] Similarly, information about carriers of known mutations like *LRRK2* G2019S or individuals whose clinical history points to exposure to certain chemical agents or a history of head injury should be factored into variant segregation analysis in PD.

Additionally, not all disease models are best studied through use of family data. For diseases with a low sibling recurrence risk ratio like autism, it may be more advantageous to study unrelated affected individuals rather than familial samples,[179] unless large pedigrees with high familial aggregation are used.[178] Finally, for studies assessing *de novo* mutations that lead to drastically reduced fitness, the benefit of additional segregation analyses may not be present.

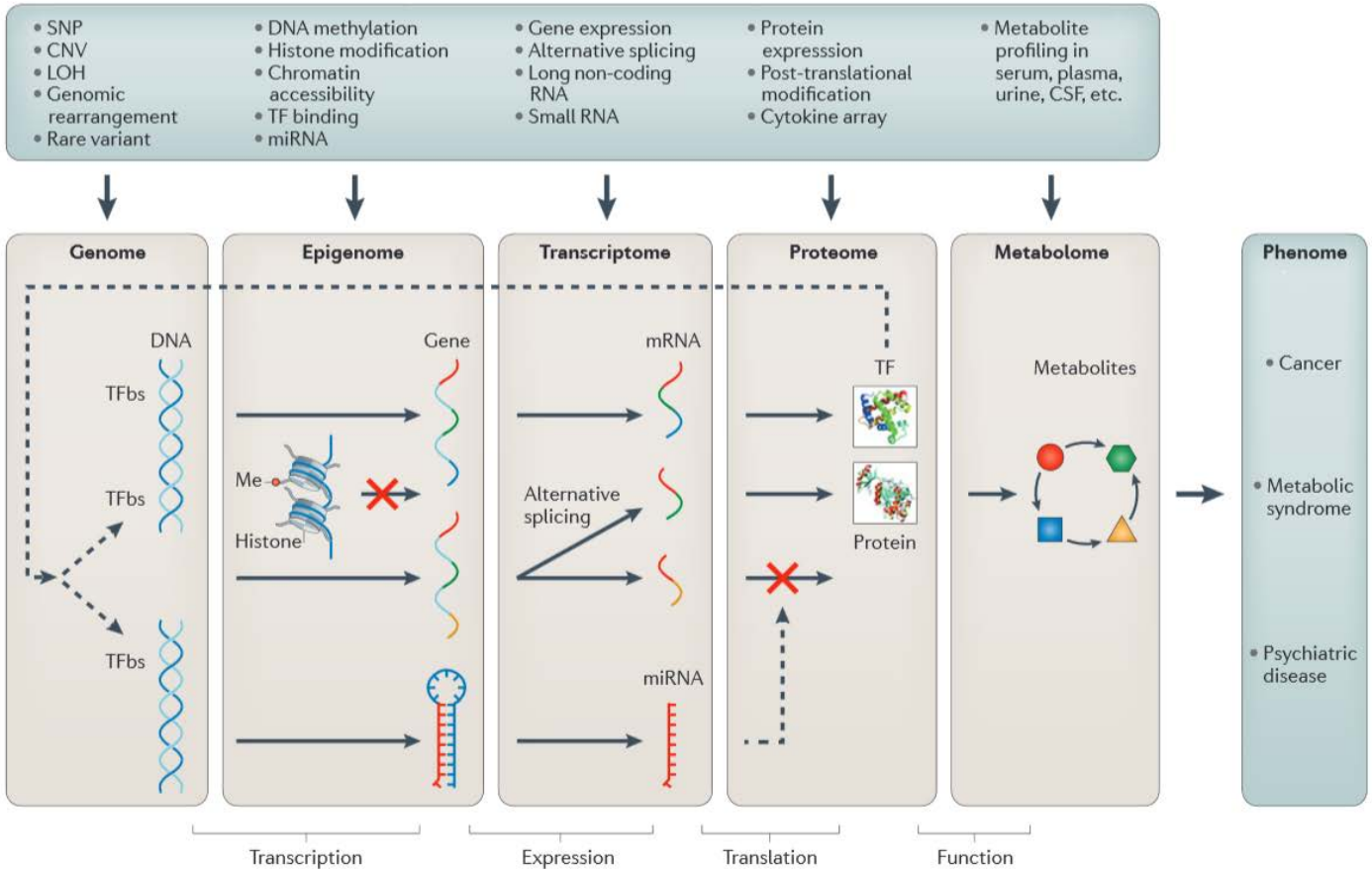
Future genetic studies in IA, PD, and XLAD

Our studies in IA, PD, and XLAD illustrate both the opportunities and challenges of family-based HTS. Candidate variants and genes have been identified in all three studies, although much work remains to fully characterize these variants and confirm their role in disease pathogenesis. Evidence from aneurysmal expressions studies in IA and a replication WES cohort in PD serve as preliminary steps in this effort, and further population genetic study designs for IA and PD are underway. Collaborative efforts to combine sequencing data with

other groups may yield further evidence toward the genes nominated in this work.

Although current efforts in our group are limited to genomic sequencing, we expect to utilize other high-throughput designs (Table 1 and Figure 17) in the future. A future challenge will be to integrate various ‘omic’ approaches with more targeted molecular studies to get a more complete picture of disease pathogenesis in individuals and populations. Even more questions exist about appropriate study designs for data integration, but the need to draw conclusions across many types of high-throughput data is well recognized.[186]

Figure 17. Data integration of high-throughput 'omics.' SNP = single nucleotide polymorphism, CNV = copy number variant, LOH = loss of heterozygosity, TF = transcription factor, bs = binding site, Me = methylation, CSF = cerebrospinal fluid. Ritchie et al, 2015.[186]



Molecular characterization of any candidate gene or variant is crucial. While the specific experiments utilized will largely depend on the gene of interest, there are general approaches and tools that could be applied broadly. Specifically, genome engineering experiments[187] can make targeted alterations to the genome that reflect the variants identified, allowing the researcher to then observe transcriptional and translational efficiency; stability, localization, binding, and functions of resultant proteins; and other potential effects of the sequence perturbations in cellular and animal models.

When designing such studies, characteristics of the disease being studied are important to consider. For instance, the phenotype of IA development and rupture may only be replicated in tissue models with careful hemodynamic control. Examination of the effects of a sequence alteration on a protein in endothelial cells may not be enough to model the complementary effects of vascular smooth muscle cells and fibroblasts, as well as how the overall vascular structure responds to hemodynamic stress or toxins introduced systemically from smoking. Additionally, while a possible genetic link has been established between IA and extracranial aneurysms,[109] the particular properties of intracranial arteries as opposed to their extracranial counterparts should be considered when constructing a model. Established differences include the distribution of elastic components, the thickness of layers of the arterial wall, and the perivascular support of cerebrospinal fluid for cerebral arteries.[188]

In PD, studying both neuronal and glial cells may be important to recapitulate the phenotype. Multiple model systems may need to be employed, as current models do not individually reproduce all aspects of PD. In fact, a current obstacle in PD animal model research is that even the observation of an aggregation of alpha-synuclein, a hallmark histopathological marker of PD, does not always correspond to a quantifiable motor phenotype in animal models. Additionally, motor symptoms in animal models do not completely translate to motor manifestations in humans affected with PD. Models also frequently do not recapitulate the common non-motor symptoms of PD, including sleep disturbances, dysfunction of the gastrointestinal system, and depression.[189]

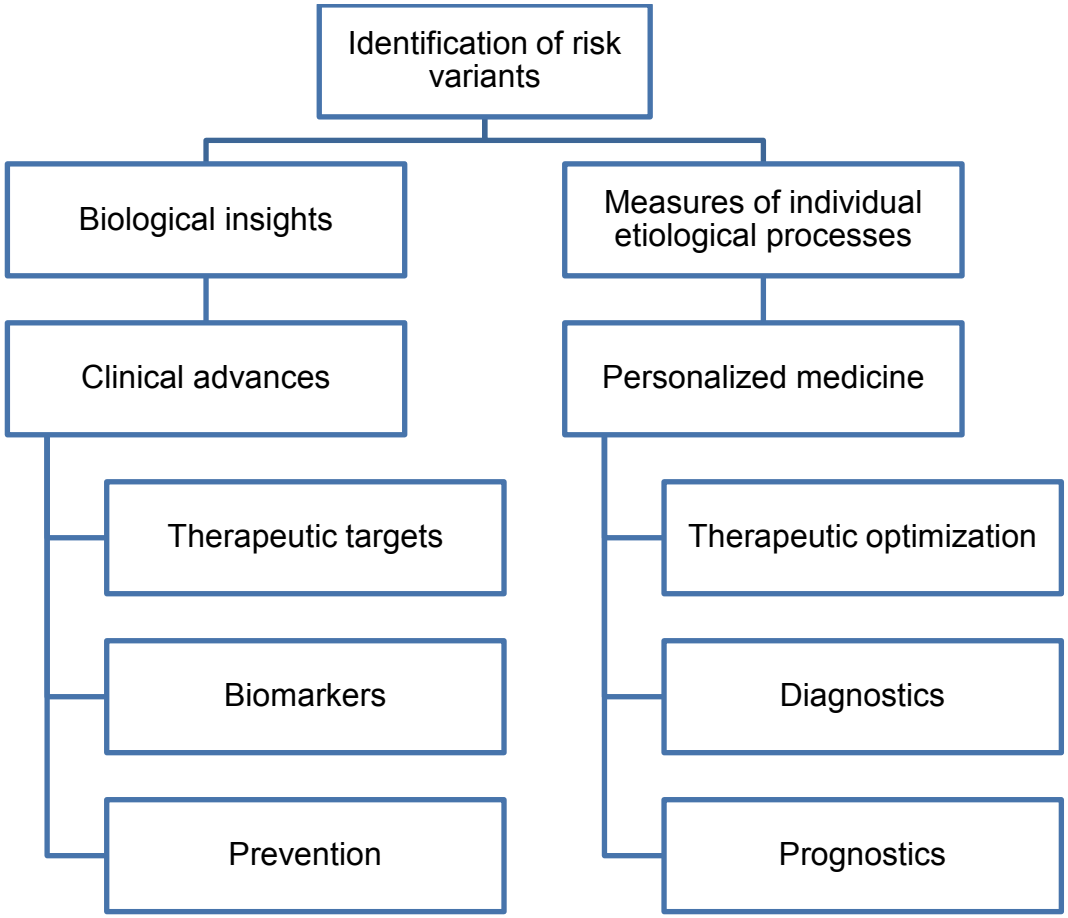
Since specific brain regions have not been implicated in XLAD other than through typical clinical symptom manifestations and their localizations, future studies in XLAD may need global studies of transcription and translation in the brain before diving into targeted study of a particular cell type or mechanism. Finally, IA, PD, and XLAD are all complicated by the fact that they are intracranial, which limits the accessibility to both affected and unaffected human tissue.

Potential clinical applications of sequencing findings in IA, PD, and XLAD

Family-based sequencing studies hold great promise for gene discovery in these three diseases, but the ultimate goal for this research is to advance the biological understanding of the disease that is necessary for benefiting patient care and

outcomes. As demonstrated in Figure 18, translation of sequencing findings to the clinic can take many forms.

Figure 18. Translation of sequencing findings to clinic. Adapted from McCarthy et al, 2008.[190]



When the role of a particular gene in IA or PD has been confirmed, functional assays may be designed to review other variants identified in the gene by other research groups or in the clinic more efficiently. As noted by those studying *BRCA1* and *BRCA2*, two genes known to be involved in multiple cancers, functional assays have been valuable in systematically determining whether a rare VUS is functional.[191,192] They do note that the multifaceted function of each gene and still imperfect knowledge of the pathophysiology behind each cancer means that multiple functional assays may be necessary to characterize a variant, and that positive or negative tests do not always translate directly to developing or not developing the cancer. Nevertheless, a growing database of genetic variation and potential mechanisms for disease in IA and/or PD would be valuable for translating sequencing findings to clinical practice.

For IA, we believe our studies will ultimately help us to better ascertain the risk of developing or rupturing an IA in a family. Given that fatality from IA rupture and resultant SAH is estimated between 25-35% in high-income countries and almost 50% in other countries,[84] better risk prediction models based on identifying high-risk patients clinically (e.g. family history of IA or SAH, smoking and/or hypertension, etc.) and combining the clinical information with a genetic profile are warranted. Understanding the biological basis behind aneurysm formation and rupture may also lead to therapeutic interventions that could help predict IA formation or rupture, and possibly halt or reverse the progression of the disease process. Such therapeutics could ultimately replace deficient protein or

chemicals, inactivate mutant substrates, increase or decrease gene expression, or possibly even introduce corrected sequence. Targeted drug delivery may serve as the major obstacle for a therapeutic agent taken systemically, although localized application may be possible during clinically-advisable neurosurgical clipping of IAs. If intracranial vasculature can still be accessed by a systemic therapy that has little harm on extracranial tissues, and the effects on extracranial vasculature in particular can be characterized, then systemic therapeutics may be feasible. Given the significant morbidity and mortality of this disease, the risk-to-benefit ratios of increased screening, monitoring, and/or intervention may be more palatable to high-risk patients and their clinicians.

In PD, no therapeutic intervention thus far is effective at neuroprotection at an early stage in the disease, and many current treatments also have severe side effects.[193] For instance, the gold standard of levodopa therapy is only efficacious without major side effects for 4-6 years.[193] Basic gene discovery projects like ours are necessary in order to provide novel therapeutic targets, but also to be able to offer early detection and monitoring of the disease progression. They may also provide new insights into other neurodegenerative diseases. Still, if findings from sequencing studies like ours only contribute to better risk prediction, diagnosis, and prognosis, they will serve less to improve clinical management of this incurable disease and more to fuel better designed studies to discover effective therapeutic interventions. Such PD therapies may assist in restoring neurochemical balance, supporting fragile dopaminergic neurons, removing

buildup of toxic substances, and addressing other changes that have been noted in PD pathophysiology.

Finally, confirming the genetic cause of XLAD could potentially offer more options for management for this currently incurable disease. With the appropriate cultural and ethical caveats, such as those suggested for genetic screening in individuals of Ashkenazi Jewish descent,[194] reproductive counseling and general carrier screening could be employed in this family. Future studies centered on the implicated gene, pathway, or other biological mechanism may offer targets for halting or even reversing the neurodegenerative process for future descendants in the family. Although this disease has only been reported in one family thus far, identification of a causative genetic mechanism in this family could also enhance our knowledge of the intersection of ataxia and dementia, both devastating symptoms of many other neurological disease processes.

Challenges of moving toward everyday genomic medicine

While not directly addressed in our research findings, our work and others[195-197] have raised some important considerations for the adoption of HTS into clinical settings. Much enthusiasm toward clinical HTS applications has been generated from some successful applications of WGS and WES in the clinic, especially in relation to rare diseases.[198,199] At the time of this work, the President of the United States unveiled a precision medicine initiative designed to funnel $215

million into researching and applying genomics in clinical care. Despite this recognition of the potential of clinical genomics, concerns include how clinically actionable variants are, the infrastructure and logistics required, and provider and patient expectations and readiness for genomic medicine. Such questions are relevant to downstream findings from our studies, as well as the numerous other HTS studies being conducted on a host of rare and complex diseases.

Clinical utility of variants associated with disease

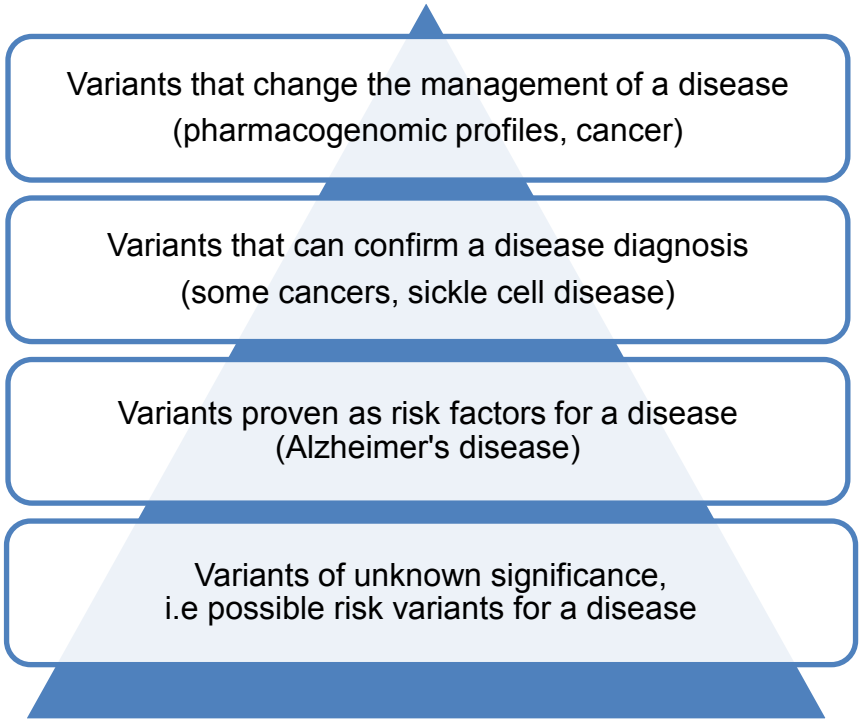
Our gene discovery projects are currently focused on identifying any variant associated with disease. In clinical applications, however, genomic variants must be stratified to facilitate their incorporation into practice (Figure 19). For instance, it is recognized that certain genetic variants have well-described effects on a patient's response to certain medications; knowledge of a patient's pharmacogenomic profile is thus clinically actionable.[200] Variants that may alter management of disease are also prevalent in cancer, where precision genomics is being employed to subtype cancers, determine prognoses, select drug regimens, and more.[201]

In some cases, knowledge that a patient carries a variant does little more than confirm the diagnosis of a disease and does not alter management at all. One such notable case is with sickle cell disease, for which the molecular basis of the disease has been known since the 1950's. Vernon Ingram himself, who demonstrated the amino acid substitution critical for sickle cell disease, declared

that “the discovery of the molecular basis of the disease…was of limited benefit to the patient population.”[202] Similarly, the genetic etiology of Huntington’s disease was established in the early 1990’s,[203] but no curative treatment currently exists. Still, some research suggests that predictive testing for Huntington’s disease, regardless of the risk profile obtained for an individual, may improve the person’s psychological well-being.[204]

In other fields, there are proven risk variants, but the small or unknown risk to an individual patient may mean that no clinical action is warranted. For instance, the *APOE* ε4 allele is the most prevalent risk factor for sporadic Alzheimer’s disease, yet not all individuals carrying the allele develop the disease while some without the allele do, and no therapeutic interventions to date have been successfully created based on knowledge of this risk variant.[205] Finally, the largest category of genomic variants includes all those that have been identified through human or animal studies, but the effect of the variants have not been confirmed or characterized. Such variants, such as those identified in our studies in IA and PD, are deposited in public databases and the literature in order to advance scientific research, but these information sources are also queried by geneticists, commercial developers of genetic products, and increasingly engaged patients. As previously mentioned, VUS are abundant in genomic research currently, and it is unclear how patient expectations and clinical practice may change in response to returned reports of these types of variants.

Figure 19. Stratification of genomic disease variants.



Infrastructure for genomic medicine

Although many preach that genomic testing is just another type of clinical test, there are few arguments that the logistics of widespread adoption of genomic medicine are enormously complicated.[195] While there is currently great commercial interest in developing user-friendly interfaces to display clinical genomic data, research to establish guidelines and standardized technological practices for incorporating genomic medicine into the medical record is still in its infancy.[206] Furthermore, infrastructure must be developed to create and maintain databases of clinically actionable genomic variants across medical specialties, as well as the technology to distribute these data appropriately to the laboratories,

clinics, and other access points for patients and healthcare providers. Given the highly identifying nature of genomic data, great care must be taken to ensure the security of any data storage and transfer. Infrastructure needs also include modifications to the Clinical Laboratory Improvement Amendments (CLIA) environment, establishing practical reimbursement schemes, and other legal, political, and regulatory necessities.[207]

Considerations for providers and patients

Beyond technological infrastructure, much debate exists about whether healthcare providers and patients are ready for genomic medicine. Many believe that there is currently inadequate genomics education in the health professions,[195,208-210] leading to a prohibitive level of physician discomfort in interpreting and applying genomic information in everyday practice.[211,212] The genetics community is working to suggest what type of results to return and when,[213-215] but there is currently no gold standard across medical specialties. The number of potential incidental findings, findings whose implications may change quickly over time as research advances, is unprecedented in genomic data.

Many advocate for patient choice in the return of genomic data, and recently developed direct-to-consumer options encourage active patient engagement but raise concerns in the clinical and research communities.[216-219] Studies have shown that most patients would prefer to have all or most information returned,

even if such data is not deemed actionable by the clinical genetics community.[220,221] This poses the question of whether the return of some information, such as VUS, may actually cause more harm than good. Such harm could be psychological or, in the case of unnecessary testing and treatment, physical. Additionally, there will likely be increased strain on healthcare resources, as time, labor, and money are redirected toward pre- and post-test counseling and following up potential findings. The released genetic information may also have implications for family members of the patient, which raises issues about informed consent and counseling of entire families, especially in regard to pediatric patients. While efforts to provide genomics education to the public are beneficial,[222] it is unlikely that these initiatives will be enough to ensure that genomic data are appropriately received, internalized, and utilized. As a result of questions about value, potential harm, cost, and feasibility, some advocate for the limited return of select incidental findings to particular patient populations based on disease state[223] or the stage of lifespan and purpose of the test.[224]

Genomics holds incredible potential to revolutionize our knowledge of disease, as well as the practice of medicine in general. Our understanding of the genetic basis behind diseases like IA, PD, and XLAD will undoubtedly advance through different high-throughput technologies. The many advantages of family-based sequencing studies in both rare and complex disease position them to become strategies of choice for gene discovery projects, with important caveats to ensure appropriate study design and molecular characterization of implicated genes.

Ultimately, such research should be directed toward improved and novel clinical applications. Much work remains, however, to ensure that unintended implications of the widespread adoption of genomic medicine are premeditated and thoughtfully handled.

REFERENCES

1 Tsui, L. C. *et al.* Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science* **230**, 1054-1057 (1985).

2 Knowlton, R. G. *et al.* A polymorphic DNA marker linked to cystic fibrosis is located on chromosome 7. *Nature* **318**, 380-382 (1985).

3 Wainwright, B. J. *et al.* Localization of cystic fibrosis locus to human chromosome 7cen-q22. *Nature* **318**, 384-385 (1985).

4 White, R. *et al.* A closely linked genetic marker for cystic fibrosis. *Nature* **318**, 382-384 (1985).

5 Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**, 234-238 (1983).

6 Roberts, S. B., MacLean, C. J., Neale, M. C., Eaves, L. J. & Kendler, K. S. Replication of linkage studies of complex traits: an examination of variation in location estimates. *Am J Hum Genet* **65**, 876-884, doi:S0002-9297(07)62338-6 [pii] 10.1086/302528 (1999).

7 Risch, N. Genetic-Linkage and Complex Diseases, with Special Reference to Psychiatric-Disorders. *Genetic Epidemiology* **7**, 3-16, doi:DOI 10.1002/gepi.1370070103 (1990).

8 Altmuller, J., Palmer, L. J., Fischer, G., Scherb, H. & Wjst, M. Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet* **69**, 936-950, doi:S0002-9297(07)61310-X [pii] 10.1086/324069 (2001).

9 Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516-1517 (1996).

10 Strittmatter, W. J. *et al.* Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc Natl Acad Sci U S A* **90**, 1977-1981 (1993).

11 Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat Genet* **37**, 161-165, doi:ng1509 [pii] 10.1038/ng1509 (2005).

12 Farooqi, S. & O'Rahilly, S. Genetics of obesity in humans. *Endocr Rev* **27**, 710-718, doi:er.2006-0040 [pii] 10.1210/er.2006-0040 (2006).

13 Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A. & Contopoulos-Ioannidis, D. G. Replication validity of genetic association studies. *Nat Genet* **29**, 306-309, doi:10.1038/ng749 [pii] (2001).

14 Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet Med* **4**, 45-61, doi:10.109700125817-200203000-00002 (2002).

15 Hardy, J. The real problem in association studies. *Am J Med Genet* **114**, 253, doi:10.1002/ajmg.10294 [pii] (2002).

16 Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362-9367, doi:10.1073/pnas.0903103106 [pii] (2009).

17 Pankratz, N. *et al.* Genomewide association study for susceptibility genes contributing to familial Parkinson disease. *Hum Genet* **124**, 593-605, doi:10.1007/s00439-008-0582-9 (2009).

18 Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* **80**, 727-739, doi:S0002-9297(07)61104-5 [pii] 10.1086/513473 (2007).

19 Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* **69**, 124-137, doi:S0002-9297(07)61452-9 [pii] 10.1086/321272 (2001).

20 Goldstein, D. B. *et al.* Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet* **14**, 460-470, doi:10.1038/nrg3455 [pii] (2013).

21 Gorlov, I. P., Gorlova, O. Y., Frazier, M. L., Spitz, M. R. & Amos, C. I. Evolutionary evidence of the effect of rare variants on disease etiology. *Clin Genet* **79**, 199-206, doi:10.1111/j.1399-0004.2010.01535.x (2011).

22 Gibson, G. Rare and common variants: twenty arguments. *Nat Rev Genet* **13**, 135-145, doi:10.1038/nrg3118 [pii] (2011).

23 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753, doi:10.1038/nature08494 [pii] (2009).

24 Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687-695 (1977).

25 Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94**, 441-448 (1975).

26 Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351, doi:10.1126/science.1058040 291/5507/1304 [pii] (2001).

27 Churko, J. M., Mantalas, G. L., Snyder, M. P. & Wu, J. C. Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. *Circ Res* **112**, 1613-1623, doi:10.1161/CIRCRESAHA.113.300939 112/12/1613 [pii] (2013).

28 Majewski, J., Schwartzentruber, J., Lalonde, E., Montpetit, A. & Jabado, N. What can exome sequencing do for you? *J Med Genet* **48**, 580-589, doi:10.1136/jmedgenet-2011-100223 [pii] (2011).

29 Stenson, P. D. *et al.* The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Hum Genomics* **4**, 69-72, doi:U8K3X868GR637691 [pii] (2009).

30 Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* **12**, 745-755, doi:10.1038/nrg3031 [pii] (2011).

31 Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* **33 Suppl**, 228-237, doi:10.1038/ng1090 [pii] (2003).

32 Teer, J. K. & Mullikin, J. C. Exome sequencing: the sweet spot before whole genomes. *Hum Mol Genet* **19**, R145-151, doi:10.1093/hmg/ddq333 [pii] (2010).

33 Shendure, J. & Lieberman Aiden, E. The expanding scope of DNA sequencing. *Nat Biotechnol* **30**, 1084-1094, doi:10.1038/nbt.2421 [pii] (2012).

34 Ng, S. B. *et al.* Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* **42**, 790-793, doi:10.1038/ng.646 [pii] (2010).

35 Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet* **42**, 30-35, doi:10.1038/ng.499 [pii] (2010).

36 Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 [pii] (2010).

37 Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nat Genet* **44**, 623-630, doi:10.1038/ng.2303 [pii] (2012).

38 Rabbani, B., Mahdieh, N., Hosomichi, K., Nakaoka, H. & Inoue, I. Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. *J Hum Genet* **57**, 621-632, doi:10.1038/jhg.2012.91 [pii] (2012).

39 Do, R., Kathiresan, S. & Abecasis, G. R. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum Mol Genet* **21**, R1-9, doi:dds387 [pii] 10.1093/hmg/dds387 (2012).

40 Pelak, K. *et al.* The characterization of twenty sequenced human genomes. *PLoS Genet* **6**, e1001111, doi:10.1371/journal.pgen.1001111 [pii] (2010).

41 MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-828, doi:10.1126/science.121504 0335/6070/823 [pii] (2012).

42 Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* **95**, 5-23, doi:10.1016/j.ajhg.2014.06.009 S0002-9297(14)00271-7 [pii] (2014).

43 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 [pii] (2010).

44 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:10.1101/gr.107524.110 [pii] (2010).

45 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 [pii] (2009).

46 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315, doi:10.1038/ng.2892 [pii] (2014).

47 Zhao, H. *et al.* DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. *Genome Biol* **14**, R23, doi:gb-2013-14-3-r23 [pii] 10.1186/gb-2013-14-3-r23 (2013).

48 Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901-913, doi:gr.3577405 [pii] 10.1101/gr.3577405 (2005).

49 Li, B. *et al.* Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* **25**, 2744-2750, doi:10.1093/bioinformatics/btp528 [pii] (2009).

50 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-249, doi:10.1038/nmeth0410-248 [pii] (2010).

51 Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* **9**, e1003709, doi:10.1371/journal.pgen.1003709 PGENETICS-D-13-00588 [pii] (2013).

52 Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-3814 (2003).

53 Hu, J. & Ng, P. C. Predicting the effects of frameshifting indels. *Genome Biol* **13**, R9, doi:10.1186/gb-2012-13-2-r9 [pii] (2012).

54 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164 (2010).

55 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192, doi:10.1093/bib/bbs017 [pii] (2013).

56 Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**, 97-101, doi:10.1038/ng786 [pii] (2002).

57 Corbett, M. A. *et al.* A focal epilepsy and intellectual disability syndrome is due to a mutation in TBC1D24. *Am J Hum Genet* **87**, 371-375, doi:10.1016/j.ajhg.2010.08.001 S0002-9297(10)00413-1 [pii] (2010).

58 Corbett, M. A. *et al.* A mutation in the Golgi Qb-SNARE gene GOSR2 causes progressive myoclonus epilepsy with early ataxia. *Am J Hum*

Genet **88**, 657-663, doi:10.1016/j.ajhg.2011.04.011 S0002-9297(11)00152-2 [pii] (2011).

59 Koenekoop, R. K. *et al.* Mutations in NMNAT1 cause Leber congenital amaurosis and identify a new disease pathway for retinal degeneration. *Nat Genet* **44**, 1035-1039, doi:10.1038/ng.2356 [pii] (2012).

60 Reversade, B. *et al.* Mutations in PYCR1 cause cutis laxa with progeroid features. *Nat Genet* **41**, 1016-1021, doi:10.1038/ng.413 [pii] (2009).

61 Nikopoulos, K. *et al.* Next-generation sequencing of a 40 Mb linkage interval reveals TSPAN12 mutations in patients with familial exudative vitreoretinopathy. *Am J Hum Genet* **86**, 240-247, doi:10.1016/j.ajhg.2009.12.016 S0002-9297(09)00610-7 [pii] (2010).

62 Allen, A. S. *et al.* De novo mutations in epileptic encephalopathies. *Nature* **501**, 217-221, doi:10.1038/nature12439 [pii] (2013).

63 Jiang, Y. H. *et al.* Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *Am J Hum Genet* **93**, 249-263, doi:10.1016/j.ajhg.2013.06.012 S0002-9297(13)00281-4 [pii] (2013).

64 Lim, E. T. *et al.* Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. *Neuron* **77**, 235-242, doi:10.1016/j.neuron.2012.12.029 S0896-6273(13)00033-0 [pii] (2013).

65 Estrada, K. *et al.* Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population. *JAMA* **311**, 2305-2314, doi:10.1001/jama.2014.65111878720 [pii] (2014).

66 Regalado, E. S. *et al.* Exome sequencing identifies SMAD3 mutations as a cause of familial thoracic aortic aneurysm and dissection with intracranial and other arterial aneurysms. *Circ Res* **109**, 680-686, doi:10.1161/CIRCRESAHA.111.248161 [pii] (2011).

67 Bahlo, M., Tankard, R., Lukic, V., Oliver, K. L. & Smith, K. R. Using familial information for variant filtering in high-throughput sequencing studies. *Hum Genet* **133**, 1331-1341, doi:10.1007/s00439-014-1479-4 (2014).

68 Berger, B., Peng, J. & Singh, M. Computational solutions for omics data. *Nat Rev Genet* **14**, 333-346, doi:10.1038/nrg3433 [pii] (2013).

69 Bell, C. J. *et al.* Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med* **3**, 65ra64, doi:10.1126/scitranslmed.3001756 3/65/65ra4 [pii] (2011).

70 MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469-476, doi:10.1038/nature13127 [pii] (2014).

71 Hedges, D. J. *et al.* Comparison of three targeted enrichment strategies on the SOLiD sequencing platform. *PLoS ONE* **6**, e18595, doi:10.1371/journal.pone.0018595 PONE-D-11-01617 [pii] (2011).

72 Sulonen, A. M. *et al.* Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol* **12**, R94, doi:10.1186/gb-2011-12-9-r94 [pii] (2011).

73 O'Rawe, J. *et al.* Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* **5**, 28, doi:10.1186/gm432 [pii] (2013).

74 Rahim, N. G., Harismendy, O., Topol, E. J. & Frazer, K. A. Genetic determinants of phenotypic diversity in humans. *Genome Biol* **9**, 215, doi:10.1186/gb-2008-9-4-215 [pii] (2008).

75 McCarthy, D. J. *et al.* Choice of transcripts and software has a large effect on variant annotation. *Genome Med* **6**, 26, doi:10.1186/gm543 [pii] (2014).

76 Gnad, F., Baucom, A., Mukhyala, K., Manning, G. & Zhang, Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics* **14 Suppl 3**, S7, doi:10.1186/1471-2164-14-S3-S7 [pii] (2013).

77 Chan, P. A. *et al.* Interpreting missense variants: comparing computational methods in human disease genes CDKN2A, MLH1, MSH2, MECP2, and tyrosinase (TYR). *Hum Mutat* **28**, 683-693, doi:10.1002/humu.20492 (2007).

78 Wei, Q., Wang, L., Wang, Q., Kruger, W. D. & Dunbrack, R. L., Jr. Testing computational prediction of missense mutation phenotypes: functional characterization of 204 mutations of human cystathionine beta synthase. *Proteins* **78**, 2058-2074, doi:10.1002/prot.22722 (2010).

79 Thusberg, J., Olatubosun, A. & Vihinen, M. Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat* **32**, 358-368, doi:10.1002/humu.21445 (2011).

80 Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* **111**, E455-464, doi:10.1073/pnas.1322563111 [pii] (2014).

81 Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* **11**, 733-739, doi:10.1038/nrg2825 [pii] (2010).

82 Huang, Y. & Gottardo, R. Comparability and reproducibility of biomedical data. *Brief Bioinform* **14**, 391-401, doi:10.1093/bib/bbs078 [pii] (2013).

83 Pritchard, J. K. & Cox, N. J. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet* **11**, 2417-2423 (2002).

84 Feigin, V. L., Lawes, C. M., Bennett, D. A., Barker-Collo, S. L. & Parag, V. Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review. *Lancet Neurol* **8**, 355-369, doi:10.1016/S1474-4422(09)70025-0 [pii] (2009).

85 Vlak, M. H., Algra, A., Brandenburg, R. & Rinkel, G. J. Prevalence of unruptured intracranial aneurysms, with emphasis on sex, age, comorbidity, country, and time period: a systematic review and meta-analysis. *Lancet Neurol* **10**, 626-636, doi:10.1016/S1474-4422(11)70109-0 [pii] (2011).

86 Feigin, V. L. *et al.* Risk factors for subarachnoid hemorrhage: an updated systematic review of epidemiological studies. *Stroke* **36**, 2773-2780, doi:01.STR.0000190838.02954.e8 [pii] 10.1161/01.STR.0000190838.02954.e8 (2005).

87 Broderick, J. P. *et al.* Greater rupture risk for familial as compared to sporadic unruptured intracranial aneurysms. *Stroke* **40**, 1952-1957, doi:10.1161/STROKEAHA.108.542571 [pii] (2009).

88 Bromberg, J. E. *et al.* Subarachnoid haemorrhage in first and second degree relatives of patients with subarachnoid haemorrhage. *BMJ* **311**, 288-289 (1995).

89 Mackey, J. *et al.* Unruptured intracranial aneurysms in the Familial Intracranial Aneurysm and International Study of Unruptured Intracranial Aneurysms cohorts: differences in multiplicity and location. *J Neurosurg* **117**, 60-64, doi:10.3171/2012.4.JNS111822 (2012).

90 Ruigrok, Y. M. & Rinkel, G. J. Genetics of intracranial aneurysms. *Stroke* **39**, 1049-1055, doi:10.1161/STROKEAHA.107.497305 [pii] (2008).

91 Nahed, B. V. *et al.* Mapping a Mendelian form of intracranial aneurysm to 1p34.3-p36.13. *Am J Hum Genet* **76**, 172-179, doi:S0002-9297(07)62555-5 [pii] 10.1086/426953 (2005).

92 Foroud, T. *et al.* Genome screen in familial intracranial aneurysm. *BMC Med Genet* **10**, 3, doi:10.1186/1471-2350-10-3 [pii] (2009).

93 Worrall, B. B. *et al.* Genome screen to detect linkage to common susceptibility genes for intracranial and aortic aneurysms. *Stroke* **40**, 71-76, doi:10.1161/STROKEAHA.108.522631 [pii] (2009).

94 Vaughan, C. J. *et al.* Identification of a chromosome 11q23.2-q24 locus for familial aortic aneurysm disease, a genetically heterogeneous disorder. *Circulation* **103**, 2469-2475 (2001).

95 Ozturk, A. K. *et al.* Molecular genetic analysis of two large kindreds with intracranial aneurysms demonstrates linkage to 11q24-25 and 14q23-31. *Stroke* **37**, 1021-1027, doi:01.STR.0000206153.92675.b9 [pii] 10.1161/01.STR.0000206153.92675.b9 (2006).

96 Bilguvar, K. *et al.* Susceptibility loci for intracranial aneurysm in European and Japanese populations. *Nat Genet* **40**, 1472-1477, doi:10.1038/ng.240 [pii] (2008).

97 Yasuno, K. *et al.* Genome-wide association study of intracranial aneurysm identifies three new risk loci. *Nat Genet* **42**, 420-425, doi:10.1038/ng.563 [pii] (2010).

98 Foroud, T. *et al.* Genome-wide association study of intracranial aneurysms confirms role of Anril and SOX17 in disease risk. *Stroke* **43**, 2846-2852, doi:10.1161/STROKEAHA.112.656397 [pii] (2012).

99 Yasuno, K. *et al.* Common variant near the endothelin receptor type A (EDNRA) gene is associated with intracranial aneurysm risk. *Proc Natl Acad Sci U S A* **108**, 19707-19712, doi:10.1073/pnas.1117137108 [pii] (2011).

100 Broderick, J. P. *et al.* The Familial Intracranial Aneurysm (FIA) study protocol. *BMC Med Genet* **6**, 17, doi:1471-2350-6-17 [pii] 10.1186/1471-2350-6-17 (2005).

101 Foroud, T. Whole exome sequencing of intracranial aneurysm. *Stroke* **44**, S26-28, doi:10.1161/STROKEAHA.113.001174 44/6_suppl_1/S26 [pii] (2013).

102 DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43**, 491-+, doi:Doi 10.1038/Ng.806 (2011).

103 Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-69, doi:10.1126/science.1219240 [pii] (2012).

104 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).

105 Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20, doi:10.1002/0471142905.hg0720s76 (2013).

106 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 [pii] (2013).

107 Anders, S., Pyl, P. T. & Huber, W. HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics*, doi:btu638 [pii] 10.1093/bioinformatics/btu638 (2014).

108 Anders, S. *et al.* Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc* **8**, 1765-1786, doi:10.1038/nprot.2013.099 [pii] (2013).

109 Helgadottir, A. *et al.* The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. *Nat Genet* **40**, 217-224, doi:10.1038/ng.72 [pii] (2008).

110 Staples, J., Nickerson, D. A. & Below, J. E. Utilizing graph theory to select the largest set of unrelated individuals for genetic analysis. *Genet Epidemiol* **37**, 136-141, doi:10.1002/gepi.21684 (2013).

111 Cheung, C. Y., Marchani Blue, E. & Wijsman, E. M. A statistical framework to guide sequencing choices in pedigrees. *Am J Hum Genet* **94**, 257-267, doi:10.1016/j.ajhg.2014.01.005 S0002-9297(14)00006-8 [pii] (2014).

112 Clark, M. J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol* **29**, 908-914, doi:10.1038/nbt.1975 [pii] (2011).

113 Nalls, M. A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet* **46**, 989-993, doi:10.1038/ng.3043 [pii] (2014).

114 Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* **11**, 415-425, doi:10.1038/nrg2779 [pii] (2010).

115 Panoutsopoulou, K., Tachmazidou, I. & Zeggini, E. In search of low-frequency and rare variants affecting complex traits. *Hum Mol Genet* **22**, R16-21, doi:10.1093/hmg/ddt376 [pii] (2013).

116 Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* **44**, 243-246, doi:10.1038/ng.1074 [pii] (2012).

117 Garrison E, M. G. Haplotype-based variant detection from short-read sequencing. *ArXiv e-prints*, 3907 (2012).

118 Sauna, Z. E. & Kimchi-Sarfaty, C. Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet* **12**, 683-691, doi:10.1038/nrg3051 [pii] (2011).

119 Bonifati, V. *et al.* Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism. *Science* **299**, 256-259, doi:10.1126/science.1077209 [pii] (2003).

120 Zimprich, A. *et al.* Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron* **44**, 601-607, doi:S0896627304007202 [pii] 10.1016/j.neuron.2004.11.005 (2004).

121 Polymeropoulos, M. H. *et al.* Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* **276**, 2045-2047 (1997).

122 Kitada, T. *et al.* Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. *Nature* **392**, 605-608, doi:10.1038/33416 (1998).

123 Valente, E. M. *et al.* Hereditary early-onset Parkinson's disease caused by mutations in PINK1. *Science* **304**, 1158-1160, doi:10.1126/science.1096284 [pii] (2004).

124 Nalls, M. A. *et al.* Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet* **377**, 641-649, doi:10.1016/S0140-6736(10)62345-8 [pii] (2011).

125 Do, C. B. *et al.* Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genet* **7**, e1002141, doi:10.1371/journal.pgen.1002141 PGENETICS-D-11-00444 [pii] (2011).

126 Lange, L. A. *et al.* Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am J Hum*

Genet **94**, 233-245, doi:10.1016/j.ajhg.2014.01.010 S0002-9297(14)00011-1 [pii] (2014).

127 Yang, Y. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312**, 1870-1879, doi:10.1001/jama.2014.14601 1918774 [pii] (2014).

128 Zimprich, A. *et al.* A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease. *Am J Hum Genet* **89**, 168-175, doi:10.1016/j.ajhg.2011.06.008 S0002-9297(11)00261-8 [pii] (2011).

129 Vilarino-Guell, C. *et al.* VPS35 mutations in Parkinson disease. *Am J Hum Genet* **89**, 162-167, doi:10.1016/j.ajhg.2011.06.001 S0002-9297(11)00242-4 [pii] (2011).

130 Yu, T. W. *et al.* Using whole-exome sequencing to identify inherited causes of autism. *Neuron* **77**, 259-273, doi:10.1016/j.neuron.2012.11.002 S0896-6273(12)00993-2 [pii] (2013).

131 Singleton, A. B., Farrer, M. J. & Bonifati, V. The genetics of Parkinson's disease: progress and therapeutic implications. *Mov Disord* **28**, 14-23, doi:10.1002/mds.25249 (2013).

132 Neri, M. *et al.* Whole exome sequencing filtered by novel candidate genes as tool for gene discovery in a recessive family with Parkinson and ataxia. *Neuromuscular Disord* **22**, 810-810, doi:DOI 10.1016/j.nmd.2012.06.029 (2012).

133 Vilarino-Guell, C. *et al.* DNAJC13 mutations in Parkinson disease. *Hum Mol Genet* **23**, 1794-1801, doi:10.1093/hmg/ddt570 [pii] (2014).

134 Edvardson, S. *et al.* A deleterious mutation in DNAJC6 encoding the neuronal-specific clathrin-uncoating co-chaperone auxilin, is associated with juvenile parkinsonism. *PLoS One* **7**, e36458, doi:10.1371/journal.pone.0036458 PONE-D-12-04451 [pii] (2012).

135 Koroglu, C., Baysal, L., Cetinkaya, M., Karasoy, H. & Tolun, A. DNAJC6 is responsible for juvenile parkinsonism with phenotypic variability. *Parkinsonism Relat Disord* **19**, 320-324, doi:10.1016/j.parkreldis.2012.11.006 S1353-8020(12)00439-7 [pii] (2013).

136 Nuytemans, K. *et al.* Whole exome sequencing of rare variants in EIF4G1 and VPS35 in Parkinson disease. *Neurology* **80**, 982-989, doi:10.1212/WNL.0b013e31828727d4 [pii] (2013).

137 Pankratz, N. *et al.* Genome screen to identify susceptibility genes for Parkinson disease in a sample without parkin mutations. *Am J Hum Genet* **71**, 124-135, doi:S0002-9297(07)60040-8 [pii] 10.1086/341282 (2002).

138 Nichols, W. C. *et al.* Linkage stratification and mutation analysis at the Parkin locus identifies mutation positive Parkinson's disease families. *J Med Genet* **39**, 489-492 (2002).

139 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311 (2001).

140 Yuan, B. *et al.* Global transcriptional disturbances underlie Cornelia de Lange syndrome and related phenotypes. *J Clin Invest*, doi:77435 [pii] 10.1172/JCI77435 (2015).

141 Bainbridge, M. N. *et al.* Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome Biol* **12**, R68, doi:10.1186/gb-2011-12-7-r68 [pii] (2011).

142 Shen, Y. *et al.* A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res* **20**, 273-280, doi:10.1101/gr.096388.109 [pii] (2010).

143 Reid, J. G. *et al.* Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. *BMC Bioinformatics* **15**, 30, doi:10.1186/1471-2105-15-30 [pii] (2014).

144 Bainbridge, M. N. *et al.* Whole-genome sequencing for optimized patient management. *Sci Transl Med* **3**, 87re83, doi:10.1126/scitranslmed.3002243 3/87/87re3 [pii] (2011).

145 Zhao, C. *et al.* Charcot-Marie-Tooth disease type 2A caused by mutation in a microtubule motor KIF1Bbeta. *Cell* **105**, 587-597, doi:S0092-8674(01)00363-4 [pii] (2001).

146 Nangaku, M. *et al.* KIF1B, a novel microtubule plus end-directed monomeric motor protein for transport of mitochondria. *Cell* **79**, 1209-1220, doi:0092-8674(94)90012-4 [pii] (1994).

147 Niwa, S., Tanaka, Y. & Hirokawa, N. KIF1Bbeta- and KIF1A-mediated axonal transport of presynaptic regulator Rab3 occurs in a GTP-dependent manner through DENN/MADD. *Nat Cell Biol* **10**, 1269-1279, doi:10.1038/ncb1785 [pii] (2008).

148 Yonekawa, Y. *et al.* Defect in synaptic vesicle precursor transport and neuronal cell death in KIF1A motor protein-deficient mice. *J Cell Biol* **141**, 431-441 (1998).

149 Kijima, K. *et al.* Mitochondrial GTPase mitofusin 2 mutation in Charcot-Marie-Tooth neuropathy type 2A. *Hum Genet* **116**, 23-27, doi:10.1007/s00439-004-1199-2 (2005).

150 La Torre, A., del Rio, J. A., Soriano, E. & Urena, J. M. Expression pattern of ACK1 tyrosine kinase during brain development in the mouse. *Gene Expr Patterns* **6**, 886-892, doi:S1567-133X(06)00049-4 [pii] 10.1016/j.modgep.2006.02.009 (2006).

151 Urena, J. M. *et al.* Expression, synaptic localization, and developmental regulation of Ack1/Pyk1, a cytoplasmic tyrosine kinase highly expressed in the developing and adult brain. *J Comp Neurol* **490**, 119-132, doi:10.1002/cne.20656 (2005).

152 La Torre, A. *et al.* A role for the tyrosine kinase ACK1 in neurotrophin signaling and neuronal extension and branching. *Cell Death Dis* **4**, e602, doi:10.1038/cddis.2013.99 [pii] (2013).

153 Hitomi, Y. *et al.* Mutations in TNK2 in severe autosomal recessive infantile onset epilepsy. *Ann Neurol* **74**, 496-501, doi:10.1002/ana.23934 (2013).

154 Howlin, J., Rosenkvist, J. & Andersson, T. TNK2 preserves epidermal growth factor receptor expression on the cell surface and enhances migration and invasion of human breast cancer cells. *Breast Cancer Res* **10**, R36, doi:10.1186/bcr2087 [pii] (2008).

155 Galisteo, M. L., Yang, Y., Urena, J. & Schlessinger, J. Activation of the nonreceptor protein tyrosine kinase Ack by multiple extracellular stimuli. *Proc Natl Acad Sci U S A* **103**, 9796-9801, doi:0603714103 [pii] 10.1073/pnas.0603714103 (2006).

156 Pesheva, P., Spiess, E. & Schachner, M. J1-160 and J1-180 are oligodendrocyte-secreted nonpermissive substrates for cell adhesion. *J Cell Biol* **109**, 1765-1778 (1989).

157 Jakovcevski, I., Miljkovic, D., Schachner, M. & Andjus, P. R. Tenascins and inflammation in disorders of the nervous system. *Amino Acids* **44**, 1115-1127, doi:10.1007/s00726-012-1446-0 (2013).

158 Anlar, B. & Gunel-Ozcan, A. Tenascin-R: role in the central nervous system. *Int J Biochem Cell Biol* **44**, 1385-1389, doi:10.1016/j.biocel.2012.05.009 S1357-2725(12)00175-6 [pii] (2012).

159 Healy, D. G. *et al.* Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: a case-control study. *Lancet Neurol* **7**, 583-590, doi:10.1016/S1474-4422(08)70117-0 [pii] (2008).

160 Alcalay, R. N. *et al.* Comparison of Parkinson risk in Ashkenazi Jewish patients with Gaucher disease and GBA heterozygotes. *JAMA Neurol* **71**, 752-757, doi:10.1001/jamaneurol.2014.313 1861751 [pii] (2014).

161 McDonnell, S. K. *et al.* Complex segregation analysis of Parkinson's disease: The Mayo Clinic Family Study. *Ann Neurol* **59**, 788-795, doi:10.1002/ana.20844 (2006).

162 Kurz, M., Alves, G., Aarsland, D. & Larsen, J. P. Familial Parkinson's disease: a community-based study. *Eur J Neurol* **10**, 159-163, doi:532 [pii] (2003).

163 Elbaz, A. *et al.* Familial aggregation of Parkinson's disease: a population-based case-control study in Europe. EUROPARKINSON Study Group. *Neurology* **52**, 1876-1882 (1999).

164 Nichols, W. C. *et al.* Genetic screening for a single common LRRK2 mutation in familial Parkinson's disease. *Lancet* **365**, 410-412, doi:S0140-6736(05)17828-3 [pii] 10.1016/S0140-6736(05)17828-3 (2005).

165 Klein, C., Chuang, R., Marras, C. & Lang, A. E. The curious case of phenocopies in families with genetic Parkinson's disease. *Mov Disord* **26**, 1793-1802, doi:10.1002/mds.23853 (2011).

166 Farlow, M. R., DeMyer, W., Dlouhy, S. R. & Hodes, M. E. X-linked recessive inheritance of ataxia and adult-onset dementia: clinical features and preliminary linkage analysis. *Neurology* **37**, 602-607 (1987).

167 Evidente, V. G., Gwinn-Hardy, K. A., Caviness, J. N. & Gilman, S. Hereditary ataxias. *Mayo Clin Proc* **75**, 475-490, doi:S0025-6196(11)64217-1 [pii] 10.4065/75.5.475 (2000).

168 Cohn-Hokke, P. E., Elting, M. W., Pijnenburg, Y. A. & van Swieten, J. C. Genetics of dementia: update and guidelines for the clinician. *Am J Med Genet B Neuropsychiatr Genet* **159B**, 628-643, doi:10.1002/ajmg.b.32080 (2012).

169 Chatterjee, S. & Pal, J. K. Role of 5'- and 3'-untranslated regions of mRNAs in human diseases. *Biol Cell* **101**, 251-262, doi:10.1042/BC20080104 [pii] (2009).

170 Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* **30**, 1095-1106, doi:10.1038/nbt.2422 [pii] (2012).

171 An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 [pii] (2012).

172 Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* **28**, 1045-1048, doi:10.1038/nbt1010-1045 [pii] (2010).

173 Oberle, I. *et al.* Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. *Science* **252**, 1097-1102 (1991).

174 Verkerk, A. J. *et al.* Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905-914, doi:0092-8674(91)90397-H [pii] (1991).

175 Kremer, E. J. *et al.* Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. *Science* **252**, 1711-1714 (1991).

176 Campuzano, V. *et al.* Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* **271**, 1423-1427 (1996).

177 Raphael, B. J. Chapter 6: Structural variation and medical genomics. *PLoS Comput Biol* **8**, e1002821, doi:10.1371/journal.pcbi.1002821 PCOMPBIOL-D-12-01585 [pii] (2012).

178 Wijsman, E. M. The role of large pedigrees in an era of high-throughput sequencing. *Hum Genet* **131**, 1555-1563, doi:10.1007/s00439-012-1190-2 (2012).

179 Ionita-Laza, I. & Ottman, R. Study designs for identification of rare disease variants in complex diseases: the utility of family-based designs. *Genetics* **189**, 1061-1068, doi:10.1534/genetics.111.131813 [pii] (2011).

180 Evangelou, E., Trikalinos, T. A., Salanti, G. & Ioannidis, J. P. Family-based versus unrelated case-control designs for genetic associations. *PLoS Genet* **2**, e123, doi:10.1371/journal.pgen.0020123 (2006).

181 Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**, 1665-1674, doi:gr.6861907 [pii] 10.1101/gr.6861907 (2007).

182 Peng, G. *et al.* Rare variant detection using family-based sequencing analysis. *Proc Natl Acad Sci U S A* **110**, 3985-3990, doi:10.1073/pnas.1222158110 [pii] (2013).

183 Li, B. *et al.* A likelihood-based framework for variant calling and de novo mutation detection in families. *PLoS Genet* **8**, e1002944, doi:10.1371/journal.pgen.1002944 PGENETICS-D-12-00106 [pii] (2012).

184 Mathew, G., George, V. & Xu, H. Comparison of several sequence-based association methods in pedigrees. *BMC Proc* **8**, S48, doi:10.1186/1753-6561-8-S1-S48 [pii] (2014).

185 Thomas, D. C., Yang, Z. & Yang, F. Two-phase and family-based designs for next-generation sequencing studies. *Front Genet* **4**, 276, doi:10.3389/fgene.2013.00276 (2013).

186 Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet*, doi:10.1038/nrg3868 [pii] (2015).

187 Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262-1278, doi:10.1016/j.cell.2014.05.010 S0092-8674(14)00604-7 [pii] (2014).

188 Hayashi, K., Handa, H., Nagasawa, S., Okumura, A. & Moritake, K. Stiffness and elastic behavior of human intracranial and extracranial arteries. *J Biomech* **13**, 175-184 (1980).

189 Le, W., Sayana, P. & Jankovic, J. Animal models of Parkinson's disease: a gateway to therapeutics? *Neurotherapeutics* **11**, 92-110, doi:10.1007/s13311-013-0234-1 (2014).

190 McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**, 356-369, doi:10.1038/nrg2344 [pii] (2008).

191 Guidugli, L. *et al.* Functional assays for analysis of variants of uncertain significance in BRCA2. *Hum Mutat* **35**, 151-164, doi:10.1002/humu.22478 (2014).

192 Millot, G. A. *et al.* A guide for functional analysis of BRCA1 variants of uncertain significance. *Hum Mutat* **33**, 1526-1537, doi:10.1002/humu.22150 (2012).

193 Kalinderi, K., Fidani, L., Katsarou, Z. & Bostantjopoulou, S. Pharmacological treatment and the prospect of pharmacogenetics in Parkinson's disease. *Int J Clin Pract* **65**, 1289-1294, doi:10.1111/j.1742-1241.2011.02793.x (2011).

194 Gross, S. J., Pletcher, B. A. & Monaghan, K. G. Carrier screening in individuals of Ashkenazi Jewish descent. *Genet Med* **10**, 54-56,

doi:10.1097/GIM.0b013e31815f247c 00125817-200801000-00008 [pii] (2008).

195 Biesecker, L. G., Burke, W., Kohane, I., Plon, S. E. & Zimmern, R. Next-generation sequencing in the clinic: are we ready? *Nat Rev Genet* **13**, 818-824, doi:10.1038/nrg3357 [pii] (2012).

196 Green, E. D. & Guyer, M. S. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**, 204-213, doi:10.1038/nature09764 [pii] (2011).

197 Brownstein, C. A. *et al.* An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. *Genome Biol* **15**, R53, doi:10.1186/gb-2014-15-3-r53 [pii] (2014).

198 Gahl, W. A. *et al.* The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. *Genet Med* **14**, 51-59, doi:10.1038/gim.0b013e318232a005 [pii] (2012).

199 Mayer, A. N. *et al.* A timely arrival for genomic medicine. *Genet Med* **13**, 195-196, doi:10.1097/GIM.0b013e3182095089 (2011).

200 Harper, A. R. & Topol, E. J. Pharmacogenomics in clinical practice and drug development. *Nat Biotechnol* **30**, 1117-1124, doi:10.1038/nbt.2424 [pii] (2012).

201 Roychowdhury, S. & Chinnaiyan, A. M. Translating genomics for precision cancer medicine. *Annu Rev Genomics Hum Genet* **15**, 395-415, doi:10.1146/annurev-genom-090413-025552 (2014).

202 Ingram, V. M. Sickle-cell anemia hemoglobin: the molecular biology of the first "molecular disease"--the crucial importance of serendipity. *Genetics* **167**, 1-7, doi:167/1/1 [pii] (2004).

203 A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. *Cell* **72**, 971-983, doi:0092-8674(93)90585-E [pii] (1993).

204 Wiggins, S. *et al.* The psychological consequences of predictive testing for Huntington's disease. Canadian Collaborative Study of Predictive Testing. *N Engl J Med* **327**, 1401-1405, doi:10.1056/NEJM199211123272001 (1992).

205 Michaelson, D. M. APOE epsilon4: The most prevalent yet understudied risk factor for Alzheimer's disease. *Alzheimers Dement* **10**, 861-868, doi:S1552-5260(14)02499-6 [pii] 10.1016/j.jalz.2014.06.015 (2014).

206 Kullo, I. J. *et al.* Return of results in the genomic medicine projects of the eMERGE network. *Front Genet* **5**, 50, doi:10.3389/fgene.2014.00050 (2014).

207 Meric-Bernstam, F., Farhangfar, C., Mendelsohn, J. & Mills, G. B. Building a personalized medicine infrastructure at a major cancer center. *J Clin Oncol* **31**, 1849-1857, doi:10.1200/JCO.2012.45.3043 [pii] (2013).

208 Demmer, L. A. & Waggoner, D. J. Professional medical education and genomics. *Annu Rev Genomics Hum Genet* **15**, 507-516, doi:10.1146/annurev-genom-090413-025522 (2014).

209 Korf, B. R. *et al.* Framework for development of physician competencies in genomic medicine: report of the Competencies Working Group of the Inter-Society Coordinating Committee for Physician Education in Genomics. *Genet Med* **16**, 804-809, doi:10.1038/gim.2014.35 [pii] (2014).

210 Patay, B. A. & Topol, E. J. The unmet need of education in genomic medicine. *Am J Med* **125**, 5-6, doi:10.1016/j.amjmed.2011.05.005 S0002-9343(11)00388-3 [pii] (2012).

211 Stanek, E. J. *et al.* Adoption of pharmacogenomic testing by US physicians: results of a nationwide survey. *Clin Pharmacol Ther* **91**, 450-458, doi:10.1038/clpt.2011.306 [pii] (2012).

212 Klitzman, R. *et al.* Attitudes and practices among internists concerning genetic testing. *J Genet Couns* **22**, 90-100, doi:10.1007/s10897-012-9504-z (2013).

213 Scheuner, M. T. *et al.* Reporting genomic secondary findings: ACMG members weigh in. *Genet Med* **17**, 27-35, doi:10.1038/gim.2014.165 [pii] (2015).

214 ACMG policy statement: updated recommendations regarding analysis and reporting of secondary findings in clinical genome-scale sequencing. *Genet Med* **17**, 68-69, doi:10.1038/gim.2014.151 [pii] (2015).

215 McEwen, J. E. *et al.* The Ethical, Legal, and Social Implications Program of the National Human Genome Research Institute: reflections on an ongoing experiment. *Annu Rev Genomics Hum Genet* **15**, 481-505, doi:10.1146/annurev-genom-090413-025327 (2014).

216 Murray, M. F. Why we should care about what you get for "only $99" from a personal genomic service. *Ann Intern Med* **160**, 507-508, doi:10.7326/M13-2804 1828555 [pii] (2014).

217 Annes, J. P., Giovanni, M. A. & Murray, M. F. Risks of presymptomatic direct-to-consumer genetic testing. *N Engl J Med* **363**, 1100-1101, doi:10.1056/NEJMp1006029 (2010).

218 Statement of the ESHG on direct-to-consumer genetic testing for health-related purposes. *Eur J Hum Genet* **18**, 1271-1273, doi:10.1038/ejhg.2010.129 [pii] (2010).

219 Prainsack, B. & Vayena, E. Beyond the clinic: 'direct-to-consumer' genomic profiling services and pharmacogenomics. *Pharmacogenomics* **14**, 403-412, doi:10.2217/pgs.13.10 (2013).

220 Fernandez, C. V. *et al.* Attitudes of parents toward the return of targeted and incidental genomic research findings in children. *Genet Med* **16**, 633-640, doi:10.1038/gim.2013.201 [pii] (2014).

221 O'Daniel, J. & Haga, S. B. Public perspectives on returning genetics and genomics research results. *Public Health Genomics* **14**, 346-355, doi:10.1159/000324933 [pii] (2011).

222 Brazas, M. D., Lewitter, F., Schneider, M. V., van Gelder, C. W. & Palagi, P. M. A quick guide to genomics and bioinformatics training for clinical and public audiences. *PLoS Comput Biol* **10**, e1003510, doi:10.1371/journal.pcbi.1003510 PCOMPBIOL-D-13-01935 [pii] (2014).

223 Bennette, C. S., Gallego, C. J., Burke, W., Jarvik, G. P. & Veenstra, D. L. The cost-effectiveness of returning incidental findings from next-generation genomic sequencing. *Genet Med*, doi:10.1038/gim.2014.156 [pii] (2014).

224 Topol, E. J. Individualized medicine from prewomb to tomb. *Cell* **157**, 241-253, doi:10.1016/j.cell.2014.02.012 S0092-8674(14)00204-9 [pii] (2014).

CURRICULUM VITAE

Janice L. Farlow

EDUCATION

2009-present **Indiana University**, Indianapolis, IN
M.D. (expected May 2017)

2009-2015 **Indiana University**, Indianapolis, IN
Medical Scientist Training Program

Ph.D. Medical and Molecular Genetics (2015)
Thesis: "Familial studies in whole exome and whole genome sequencing"
Advisor: Tatiana Foroud, Ph.D.
Minor: Life Sciences

2005-2009 **Indiana University**, Bloomington, IN
Wells Scholar Program
Hutton Honors College
Summa Cum Laude, General Honors Notation

B.S. Biology, Honors (2009)
Thesis: "Cleaved Amplified Polymorphic Sequences (CAPS) marker mapping of pollen sterility genes in *Solanum*"
Advisor: Leonie C. Moyle, Ph.D.

B.A. Individualized Major (Romance Languages), Honors (2009)
Thesis: "Language learning social networking sites and their implications for formal language education."
Advisor: James L. Franklin, Ph.D.

Minors: Chemistry, Economics

HONORS, AWARDS, FELLOWSHIPS

2015 Excellence in Public Health Award, United States Public Health Service

2015 Dr. Charles R. Bantz Award for Excellence, Indiana University Purdue University Indianapolis

2015 Premier 10, Best in School, and Elite 50, Indiana University Purdue University Indianapolis

2015	William M. Plater Civic Engagement Medallion, Indiana University Purdue University Indianapolis
2015	Wells Graduate Fellowship, Indiana University
2014	Outstanding Student Contributions, Curricular Reform, Indiana University School of Medicine
2013	Student Travel Scholarship, Central Group on Educational Affairs, Association of American Medical Colleges
2013	Educational Enhancement Grant, Graduate Professional Student Government, Indiana University School of Medicine
2012	AMA Foundation Leadership Award, American Medical Association
2012-2014	Clinical and Translational Sciences Institute Pre-Doctoral Training Grant, Indiana Clinical and Translational Sciences Institute
2011	Educational Enhancement Grant, Graduate Professional Student Government, Indiana University School of Medicine
2011	Chapter Outreach Grant, American Academy of Pediatrics
2010	Medical Student Service Project Award, Alpha Omega Alpha
2009-present	Medical Scientist Training Program Fellowship, Indiana University School of Medicine
2009	Hutton Honors College / Individualized Major Program Capstone Award, Indiana University
2008	Hutton Honors College Research Grant, Indiana University
2007	Phi Beta Kappa Appointment, Indiana University
2006	Norton-Mavor Classical Studies Prize, Indiana University
2005-2009	Founder’s Scholar, Indiana University
2005-2009	Dean’s List, Indiana University
2005-2009	Herman B Wells Scholar, Indiana University

2005	Presidential Scholar, United States of America
2005	National Merit Scholar, National Merit Scholarship Program
2005	AP Scholar with Honor, The College Board
2005	Hoosier Scholar, State of Indiana
2005	Academic All-Star, Indianapolis Star

GRADUATE AND UNDERGRADUATE LEADERSHIP EXPERIENCE

2014-present	Student Lead, Independent Student Analysis Committee, Liaison Committee on Medical Education Accreditation Self-Study Task Force, Indiana University School of Medicine
2014-present	Nominating Committee, Administrative Board, Organization of Student Representatives, Association of American Medical Colleges
2014-present	Chair, Central Region, Organization of Student Representatives, Association of American Medical Colleges
2014-present	Organization of Student Representatives Representative, Steering Committee, Central Group on Student Affairs, Association of American Medical Colleges
2014-present	Editor and Reviewer, Journal of Student-Run Clinics
2014-2015	Member of the Planning Committee and Abstract Reviewer, 2015 Central Group on Educational Affairs / Central Group on Student Affairs / Central Region of the Organization of Student Representatives Regional Meeting
2014-2015	Member of the Conference Committee and Abstract Reviewer, Society of Student-Run Free Clinics
2014-2015	Chair, Search and Screen Committee for the Student Trustee, Indiana University

2013-2015	Trustee, Indiana University Board of Trustees
2013-2014	Organization of Student Representatives Representative, Executive Committee, Central Group on Educational Affairs, Association of American Medical Colleges
2012-2014	Regional Delegate for Medical Education, Central Region, Organization of Student Representatives, Association of American Medical Colleges
2012-2013	Chair, Indiana University Student Outreach Clinic
2012-2013	Member, Search Committee for Dean of the Indiana University School of Medicine and Indiana University Vice President for Clinical Affairs, Indiana University
2012-2013	Senior Co-Chair, Medical Student Service-Learning Group, Indiana University School of Medicine
2012	Executive President, Medical Student Council
2011-2014	Member, Curricular Reform Committees, Indiana University School of Medicine
2011-2012	Member, Administrative Review Committee for Dean of the Indiana University School of Medicine, Indiana University
2011-2012	Research Chair, Indiana University Student Outreach Clinic
2011-2012	Junior Co-Chair, Medical Student Service-Learning Group, Indiana University School of Medicine
2011	Secretary, Medical Student Council
2010-present	Representative, Organization of Student Representatives, Association of American Medical Colleges
2010-2012	Organizing Committee, Westside Health Fair

2010-2012	Philanthropy Co-Chair, Indiana University Chapter, American Medical Women's Association
2010-2011	Vice-President, Indiana University Chapter, American Medical Women's Association
2010	Member, ENLACE Medical Spanish Trip to Honduras
2009-2011	Member, Community Leadership Mentor Program
2009-2011	Medical Student Liaison, Indiana Chapter, American Academy of Pediatrics
2008-2011	Member, Programming Services Committee, Indiana Chapter, March of Dimes
2008-2009	Member, Board of Aeons, Indiana University
2008-2009	Vice-President, Indiana University Chapter, Circle K International
2007-2009	Chapter Youth on Board, Board of Directors, Indiana Chapter, March of Dimes
2007-2009	Volunteer Co-Chair, Indiana University Chapter, Alpha Chi Sigma
2007	District Awards Chair, Indiana District, Circle K International
2006-2008	Vice-Chair, Western Region, National Youth Council, March of Dimes
2006-2008	District Secretary, Indiana District, Circle K International
2005-2009	Member, National Youth Council, March of Dimes
2005-2008	Co-Chair, Wells Activism and Volunteer Effort, Wells Scholar Program, Indiana University

2005-2006 Member, International Ad-Hoc Diversity Committee, Circle K International

2005-2006 Advocacy Chair, Indiana University Chapter, Asian American Association

2005-2006 Service-Fundraising Chair, Indiana University Chapter, Circle K International

PEER-REVIEWED PUBLICATIONS

1. **Farlow JL**, Robak LA, Hetrick K, Bowling K, Boerwinkle E, Akdemir ZC, Gambin T, Gibbs RA, Gu S, Preti J, Jankovic J, Jhangiani SN, Kaw K, Lin H, Ling H, Liu Y, Lupski JR, Muzny D, Porter P, Pugh E, White J, Doheny K, Myers RM, Shulman JM, Foroud T. Whole exome sequencing identifies candidate genes for Parkinson's disease. *Submitted*.

2. **Farlow JL,** Lin H, Sauerbeck L, Lai D, Koller DL, Pugh E, Hetrick K, Ling H, Kleinloog R, van der Vlies P, Deelen P, Swertz MA, Verweij BH, Regli L, Rinkel GJE, Ruigrok YM, Doheny K, Liu Y, Broderick J, Foroud T. Lessons learned from whole exome sequencing in multiplex families affected by a complex genetic disorder, intracranial aneurysm. PLoS One. 2015;10(3):e0121104. PMID: 25803036.

3. **Farlow JL**, Goodwin C, Sevilla JM. Interprofessional education through service learning: lessons from a student-led free clinic. Journal of Interprofessional Care 2015 Jan 7:1-2 [Epub]. PMID: 25565371.

4. **Farlow JL**, Pankratz N, Wojcieszek J, Foroud T. Parkinson Disease Overview. GeneReviews [Internet]. Seattle(WA): University of Washington, Seattle; 1993-2014. PMID: 2031402.

5. **Farlow JL**, Foroud T. Genetics of Dementia. Seminars in Neurology. 2013;33(4):417-422. PMID: 24234360.

ORAL PRESENTATIONS AND WORKSHOPS

1. White J, Kammeyer R, **Farlow JL**, Gregory E, Young V, Kinney M, Emdin J, Ebbens C, Sevilla J. Development of inter-professional partnerships for attaining patient care, promotional, and educational goals. (panel) Society for Student Run Free Clinics Annual Meeting, 2015. Atlanta, GA.

2. **Farlow JL**. Advancing student engagement across Indiana University. (workshop) Indiana University Student Leaders Colloquium, 2014. Indianapolis, IN.

3. Arwood B, **Farlow JL**. Agenda setting: we have a plan, how do we make it happen? (workshop) Indiana University Kokomo Student Government Association Student Leadership Conference, 2014. Kokomo, IN.

4. **Farlow JL**. The search for rare genetic variants in intracranial aneurysms. (presentation) Clinical Translational Sciences Institute Predoctoral Grantee Series, 2014. Lafayette, IN.

5. **Farlow JL**, Wadhwani A, Pacl K, Patel A, London D. Medical student engagement in their medical education. (workshop) Organization of Student Representatives Central Business Meeting of the Association of American Medical Colleges Organization of Student Representatives (OSR) /Group on Student Affairs (GSA) / Group on Diversity and Inclusion (GDI) Spring Meeting, 2014. San Diego, CA.

6. Althoff M, Rogers AJ, **Farlow JL**, Jelenchick L, Modi A. Developing multi-institutional student run clinic research projects. (presentation) Society for Student Run Free Clinics Annual Meeting, 2014. Nashville, TN.

7. **Farlow JL,** Rochford B, Young V, Kammeyer R, Carroll E, LeMay E, Mullen K. Models for effective solicitation, incorporation, and integration of multiple professions and disciplines within a student-run clinic. (workshop) Society for Student Run Free Clinics Annual Meeting, 2014. Nashville, TN.

8. **Farlow JL,** Chan KM, Rammaha R, Hawthorne C, George L, Rochford BT, Piron C, Heskett R**.** The IU-SOC: How can we become truly interprofessional? (workshop) Society for Student Run Free Clinics Annual Meeting, 2013. San Antonio, TX.

9. **Farlow JL**. Investigation of genetic variants that cause Parkinson disease. (presentation) Clinical Translational Sciences Institute Predoctoral Grantee Series, 2012. Lafayette, IN.

10. **Farlow JL**. Application of whole exome sequencing to Parkinson disease. (presentation) Indiana University School of Medicine Medical and Molecular Genetics Research Club, 2012. Indianapolis, IN.

11. **Farlow JL.** Investigation of genetic variants that cause Parkinson disease. (presentation) Indiana University Medical Scientist Training Program Student Seminar Series, 2012. Indianapolis, IN.

12. Broderick J, **Farlow JL**, Lin H, Hetrick K, Ling H, Lai D, Sauerbeck L, Woo D, Langefeld C, Brown R, Pugh E, Doheny K, Liu Y, Foroud T. Genetics of intracranial aneurysm: exome sequencing in FIA families. (presentation). International Stroke Genetic Consortium Annual Meeting, 2012. Krakow, Poland.

13. **Farlow JL**, Lin H, Hetrick K, Ling H, Lai D, Sauerbeck L, Woo D, Langefeld C, Brown R, Pugh E, Doheny K, Liu Y, Foroud T, Broderick J. The use of linkage data to prioritize results from whole exome sequencing in familial intracranial aneurysm. (presentation) American Academy of Neurology Annual Meeting, 2012. New Orleans, LA.

14. **Farlow JL.** An interprofessional service-learning environment: the Indiana University Student Outreach Clinic. (presentation, poster) Student Programming Showcase, Association of American Medical Colleges (AAMC) Group on Student Affairs (GSA) Central/Southern Region Meeting, March 2012. Clearwater, FL.

15. Broderick J, Brown R, Sauerbeck L, Huston J, Woo D, Deka R, Meissner I, Worrall B, Ko N, Langefled C, Rouleau G, Connelly S, Anderson C, Pugh E, Hetrick K, Doheny K, **Farlow JL**, Lin H, Foroud T. (presentation) American Stroke Association, 2012. New Orleans, LA.

16. **Farlow JL**. Whole exome sequencing in familial intracranial aneurysm. (presentation) Indiana University School of Medicine Medical and Molecular Genetics Research Club, December 2011. Indianapolis, IN.

17. **Lin J**. Reaching new heights with the March of Dimes. (workshop) Circle K International Convention, 2008. Denver, CO.

18. **Lin J.** Road trip around the world: engaging a diverse audience. (workshop) March of Dimes Team Youth National Conference, 2008. Washington DC.

19. **Lin J.** Club Secretary Training. (workshop) Circle K Club Officer Training Conference, 2008. Indianapolis, IN.

20. **Lin J**. Two colors one cause. (workshop) Family Career and Community Leaders of America National Cluster Meeting, 2007. Buffalo, NY.

21. **Lin J.** Club Secretary Training. (workshop) Circle K Club Officer Training Conference, 2007. Indianapolis, IN**.**

22. **Lin J.** Diversity and Circle K International. (workshop) Circle K International Convention, 2006. Boston, MA.

23. **Lin J.** Two colors one cause. (workshop) Circle K International Convention, 2006. Boston, MA.

24. **Lin J.** Public relations toolkit. (workshop) Circle K Club Officer Training Conference, 2006. Indianapolis, IN.

25. **Lin J.** Club Treasurer Training. (workshop) Circle K Club Officer Training Conference, 2006. Indianapolis, IN.

26. **Lin J.** Key Club International. (workshop) California-Nevada-Hawaii Key Club District Convention, 2005. Long Beach, CA.

27. **Lin J.** Time for Change. (workshop) California-Nevada-Hawaii Key Club District Convention, 2005. Long Beach, CA.

28. **Lin J.** What will you do for the pink and blue. (workshop) California-Nevada-Hawaii Key Club District Convention, 2005. Long Beach, CA.

29. **Lin J.** Trustee Address. (platform presentation) California-Nevada-Hawaii Key Club District Convention, 2005. Long Beach, CA.

30. **Lin J.** Trustee Address. (platform presentation) Michigan Key Club District Convention, 2005. Battle Creek, MI.

31. **Lin J.** Public relations. (workshop) Indiana Key Club District Convention, 2004.

32. **Lin J.** Major Emphasis Program. (workshop) Indiana Key Club District Convention, 2004.

33. **Lin J**. Key to College. (workshop) Indiana Key Club Joint Division Council Meeting, 2003. Indianapolis, IN.

POSTER PRESENTATIONS

* Entries noted by an asterisk had content presented at multiple conferences. In each instance, the presentation was made to disseminate the material to a wider audience, and the conference did not disallow previously presented work.

1. **Farlow JL**, Hetrick K, Ling H, Craig B, Farlow M, Pugh E, Doheny K, Foroud T. Genome sequencing in X-Linked Ataxia Dementia. (poster) American Society of Human Genetics Annual Meeting, 2014. San Diego, CA.

2. Bains A, Mendez J, **Farlow JL**, Kirchhoff S, Margalit R, Cauley K. Curriculum for Student Run Free Clinics at CGEA Schools. (poster) Association of American Medical Colleges (AAMC) Group on Educational Affairs (GEA) Central Region Meeting, 2014. Cleveland, OH.

3. **Farlow JL**, Carroll E, Rochford B, Young V, Albrecht L, White J, Kammeyer R, Weaver G, Wanaselja A, Baughman C, Secor L, Faller M, Wahle B, Meyers K, Geros K, Woods C, Mustaklem S, Freeman A, LeMay E, Mullen K. The Indiana University Student Outreach Clinic: community-based interprofessional care. (poster) Society for Student Run Free Clinics Annual Meeting, 2014. Nashville, TN.
   *Updated poster from Heskett et al, Society for Student Run Free Clinics Annual Meeting, 2013.

4. **Farlow JL,** Keshvani N, Agarwal D, O'Neill B, Osborn K. Indiana University School of Medicine: AMA Medical Student Section Advocacy in Action. (poster) Association of American Medical Colleges (AAMC) Annual Meeting, 2013. Philadelphia, PA.

5. **Farlow JL,** Goodwin CB, Piron C, Sevilla-Martir J, Ribera T, Loftus A, Kirchhoff S. The IU Student Outreach Clinic: A Model for Community-Based Interprofessional Education. (poster) Association of American Medical Colleges (AAMC) Annual Meeting, 2013. Philadelphia, PA.
*Poster from Farlow et al, Association of American Medical Colleges (AAMC) Group on Educational Affairs (GEA) Central Region Meeting, 2013.

6. **Farlow JL**, Lin H, Hetrick K, Ling H, Pugh E, Bowling K, Jain P, Liu Y, Doheny K, Myers RM, Foroud T. Prioritization of results from whole exome sequencing in Parkinson Disease. (poster) Indiana Clinical and Translational Sciences Institute Annual Meeting, 2013. Indianapolis, IN.
* Poster from Farlow et al, American Society of Human Genetics Annual Meeting, 2012.

7. **Farlow JL**, Lin H, Hetrick K, Ling H, Pugh E, Bowling K, Jain P, Liu Y, Doheny K, Myers RM, Foroud T. Prioritization of results from whole exome sequencing in Parkinson Disease. (poster) Department of Medical and Molecular Genetics Poster Session, 2013. Indianapolis, IN.
* Poster from Farlow et al, American Society of Human Genetics Annual Meeting, 2012.

8. **Farlow JL**, Lin H, Hetrick K, Ling H, Pugh E, Bowling K, Jain P, Liu Y, Doheny K, Myers RM, Foroud T. Prioritization of results from whole exome sequencing in Parkinson Disease. (poster) National Clinical and Translational Sciences Predoctoral Programs Meeting, 2013. Rochester, MN.
* Poster from Farlow et al, American Society of Human Genetics Annual Meeting, 2012.

9. Heskett R, Chan KM, Rammaha R, Hawthorne C, George L, Rochford BT, Piron C, **Farlow JL.** A retrospective case-study of community acquired healthcare. (poster) Robert E. Bringle Civic Engagement Showcase and Symposium, 2013. Indianapolis, IN.
*Poster from Heskett et al, Society for Student Run Free Clinics Annual Meeting, 2013.

10. **Farlow JL**, Johnson J, Agarwal D, O'Neill B, Flaherty A, Mehta R. 9 Campuses, 1 School: Coordination of Student Activities and Student Affairs Among Regional Campuses. (poster) Association of American Medical Colleges (AAMC) Group on Student Affairs (GSA) and Organization of Student Representatives (OSR) Central Region Meeting, 2013. St. Louis, MO.

11. **Farlow JL**, Hodgdon K, Abhyankar R, Johnson J, Agarwal D, O'Neill B, Mehta R, Flaherty A. Passport to Wellness. (poster) Association of American Medical Colleges (AAMC) Group on Student Affairs (GSA) and Organization of Student Representatives (OSR) Central Region Meeting, 2013. St. Louis, MO.

12. **Farlow JL,** Goodwin CB, Piron C, Sevilla-Martir J, Ribera T, Loftus A, Kirchhoff S. The IU Student Outreach Clinic: A Model for Community-Based Interprofessional Education. (poster) Association of American Medical Colleges (AAMC) Group on Educational Affairs (GEA) Central Region Meeting, 2013. Cincinnati, OH.

13. Heskett R, Chan KM, Rammaha R, Hawthorne C, George L, Rochford BT, Piron C, **Farlow JL.** A retrospective case-study of community acquired healthcare. (poster) Society for Student Run Free Clinics Annual Meeting, 2013. San Antonio, TX.

14. **Farlow JL**, Lin H, Hetrick K, Ling H, Pugh E, Bowling K, Jain P, Liu Y, Doheny K, Myers RM, Foroud T. Prioritization of results from whole exome sequencing in Parkinson Disease. (poster) American Society of Human Genetics Annual Meeting, 2012. San Francisco, CA.

15. **Farlow JL**, Lin H, Hetrick K, Ling H, Lai D, Sauerbeck L, Woo D, Langefeld C, Brown R, Pugh E, Doheny K, Liu Y, Foroud T, Broderick J. Prioritization of results from whole exome sequencing in familial intracranial aneurysm. (poster) Department of Medical and Molecular Genetics Poster Session, 2012. Indianapolis, IN.
*Poster from IUPUI Research Day, 2012.

16. **Farlow JL**, Lin H, Hetrick K, Ling H, Lai D, Sauerbeck L, Woo D, Langefeld C, Brown R, Pugh E, Doheny K, Liu Y, Foroud T, Broderick J. Prioritization of results from whole exome sequencing in familial intracranial aneurysm. (poster) IUPUI Research Day, 2012. Indianapolis, IN.

17. **Farlow JL.** An interprofessional service-learning environment: the Indiana University Student Outreach Clinic. (presentation, poster) Student Programming Showcase, Association of American Medical Colleges (AAMC) Group on Student Affairs (GSA) and Organization of Student Representatives (OSR) Central/Southern Region Meeting, March 2012. Clearwater, FL.
*Updated poster from Association of American Medical Colleges Annual Meeting, 2011.

18. **Farlow JL.** Lessons learned from an interprofessional student-run free clinic: the Indiana University Student Outreach Clinic and its transformative impact on the health of downtown Indianapolis. (poster) Joseph Taylor Symposium, February 2012. Indianapolis, IN.

19. **Farlow JL**, Chen E, Johnson J, Agarwal D. An interprofessional service-learning environment: The Indiana University Student Outreach Clinic. (poster) Association of American Medical Colleges Annual Meeting, 2011. Denver, CO.

20. Foroud T, Koller D, Lai D, **Farlow JL**, Lin H, Pankratz N, Liu Y, Deka R, Sauerbeck L, Ling H, Hetrick K, Doheny K, Pugh E, Broderick J. Using linkage data to prioritize analysis of data from whole exome sequencing. (poster) American Society of Human Genetics Annual Meeting, 2011. Montreal, Canada.

21. **Farlow JL**, Martens G. Engaged learning through a student-run clinic. (poster) National Outreach Scholarship Conference, 2011. East Lansing, MI.

22. Ansari S, Martens G, Stilger B, Sarkissian A, Donaldson J, Goodwin C, Li Y, McHenry A, Morone P, **Lin J.** The operating model of the IU Student Outreach Clinic. (poster) Society of Student Run Free Clinics National Conference, 2011. Houston, TX.

23. Ansari S, Martens G, Stilger B, Sarkissian A, Donaldson J, Goodwin C, Li Y, McHenry A, Morone P, **Lin J.** The operating model of the IU Student Outreach Clinic. (poster) IUPUI Annual Civic Engagement Showcase, 2011. Indianapolis, IN.

NON-PEER-REVIEWED EDUCATIONAL RESOURCES

Farlow JL, London DA, Doo F, Golden A, Pacl K, Patel A, Wadhwani A. Medical education – curricular fundamentals: a primer for medical students involved in curricular planning. iCollaborative; 2014. Resource ID: 2315.