The Rare-Variant Generalized Disequilibrium Test for Association Analysis of Nuclear and Extended Pedigrees with Application to Alzheimer Disease WGS Data

Zongxiao He,[1] Di Zhang,[1] Alan E. Renton,[2] Biao Li,[1] Linhai Zhao,[1] Gao T. Wang,[1,4] Alison M. Goate,[2] Richard Mayeux,[3] and Suzanne M. Leal[1,*]

Whole-genome and exome sequence data can be cost-effectively generated for the detection of rare-variant (RV) associations in families. Causal variants that aggregate in families usually have larger effect sizes than those found in sporadic cases, so family-based designs can be a more powerful approach than population-based designs. Moreover, some family-based designs are robust to confounding due to population admixture or substructure. We developed a RV extension of the generalized disequilibrium test (GDT) to analyze sequence data obtained from nuclear and extended families. The GDT utilizes genotype differences of all discordant relative pairs to assess associations within a family, and the RV extension combines the single-variant GDT statistic over a genomic region of interest. The RV-GDT has increased power by efficiently incorporating information beyond first-degree relatives and allows for the inclusion of covariates. Using simulated genetic data, we demonstrated that the RV-GDT method has well-controlled type I error rates, even when applied to admixed populations and populations with substructure. It is more powerful than existing family-based RV association methods, particularly for the analysis of extended pedigrees and pedigrees with missing data. We analyzed whole-genome sequence data from families affected by Alzheimer disease to illustrate the application of the RV-GDT. Given the capability of the RV-GDT to adequately control for population admixture or substructure and analyze pedigrees with missing genotype data and its superior power over other family-based methods, it is an effective tool for elucidating the involvement of RVs in the etiology of complex traits.

Introduction

The inability of common variants identified by genome-wide association studies (GWASs) to explain much of the heritability of most complex diseases and the advances of next-generation sequencing (NGS) technologies have led to an increased interest in investigating the etiology of complex disease due to rare variants.[1,2] Most NGS association studies use a population-based design, for which a large number of rare-variant association methods have been developed. There is also great interest in performing NGS association studies with the use of family data given that causal variants that aggregate in families usually have larger effect sizes than those found in sporadic cases. Family-based studies can therefore be more powerful than population-based studies given an equivalent number of cases.[3] Study designs with familial cases might be preferred over case-control design when families with multiple affected individuals are available for study, especially for complex diseases for which loci with large effects have not been detected.[4] Another advantage of a family-based design is its ability to avoid confounding due to population admixture or substructure. Population-based association studies can suffer from inflated false-positive rates as a result of population admixture or substructure, which is an even greater problem for rare variants.[5] Rare variants are more likely to have more recent origins and are therefore more likely to be population specific than common variants, and there can be considerable differences in the rare-variant allelic spectrum, even between European ethnic groups. These differences can be more extreme in the study of admixed populations, such as African Americans and Hispanics. Family-based designs are robust against population admixture or substructure, and significant findings always imply association with the causal variant or association with a variant that is in linkage disequilibrium (LD) with the pathogenic variant.

A few tests have been proposed for family-based designs for the analysis of rare variants in sequence data. For example, the transmission disequilibrium test (TDT)[6] has been extended to test rare-variant association by grouping information across multiple variants within a genomic region.[7,8] The extensions combine the benefits of rare-variant association analysis and family-based design, providing a robust and powerful approach to identifying and characterizing rare disease-susceptibility variants. However, these methods are not valid tests of LD for nuclear families with more than one affected child. Moreover, when extended pedigrees with multiple nuclear families and/or discordant sib-pairs are available, it is advantageous to include them in analysis because they also provide association information. The family-based association test (FBAT)[9] has been extended to analyze sequence data in the rare-variant burden association test.[10] A variance-component extension of FBAT has also been proposed, but the implemented software is applicable only

[1]Center for Statistical Genetics, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA; [2]Department of Neuroscience and Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY 10029, USA; [3]Department of Neurology, Taub Institute on Alzheimer's Disease and the Aging Brain and Gertrude H. Sergievsky Center, Columbia University, New York, NY 10027, USA
[4]Present address: Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA
*Correspondence: sleal@bcm.edu
http://dx.doi.org/10.1016/j.ajhg.2016.12.001.

to case-parent trio data.[11] However, both FBAT extensions suffer from potential loss of power, which can be substantial for extended pedigrees, because these methods ignore parental phenotypes. Epstein et al. proposed a statistical approach for rare-variant association testing in affected sibships,[12] which is less than optimal because it can be used only for analyzing affected sib-pairs in nuclear families (Michael Epstein, personal communication). Recently, Sul et al. proposed RareIBD for the analysis of large extended families of arbitrary structure.[13] A main assumption of RareIBD is that only one founder in a family carries a rare variant in a given gene, which is often violated especially for extended pedigrees, and violation of this assumption will result in inflated type I error rates.

Here, we propose the rare-variant extension of the generalized disequilibrium test (GDT). The GDT utilizes genotype differences in all discordant relative pairs to assess associations within a family.[14] The GDT has increased power by efficiently incorporating information beyond first-degree relatives. Moreover, quantitative or qualitative covariates, e.g., age, body mass index, and smoking status, can be incorporated in the analysis to control for confounding. The rare-variant extension of GDT (RV-GDT) aggregates a single-variant GDT statistic over a genomic region of interest, which is usually a gene. Additionally, the RV-GDT can incorporate weights that are based on allele frequencies or bioinformatics tools. Using simulated genetic data, we demonstrated that the RV-GDT method has well-controlled type I error rates, even when applied to admixed populations, populations with substructure, and pedigrees with family members missing genotype data. As a comparison, we also extended the pedigree disequilibrium test (PDT)[15,16] to analyze rare variants in general pedigrees. The PDT breaks a general pedigree into case-parent trios and discordant sib-pairs and then combines their contributions into a statistic that takes into account their non-independence. The rare-variant extension of the PDT (RV-PDT) is a weighted or unweighted combination of the single-variant PDT statistic over a genomic region of interest. The type I error was also evaluated in our simulated data for the RV-PDT and RareIBD. Although the RV-PDT had well-controlled type I error, the type I error for RareIBD was inflated, especially for extended pedigrees. The power of the RV-GDT was always substantially more powerful than that of Epstein's affected-sib-pair (ASP) method and had similar or slightly higher power for nuclear families than the FBAT and RV-PDT under a variety of disease models. However, when applied to extended pedigrees and/or pedigrees in which family members were missing genotypes data, the RV-GDT was more powerful than these methods.

To further illustrate application of the proposed methods, we analyzed whole-genome sequence (WGS) data for 81 families affected by Alzheimer disease (AD [MIM: 104300]) from the Alzheimer Disease Sequencing Project (ADSP; dbGaP: phs000572.v6.p4). AD is a neurodegenerative disease characterized by dementia and typically begins with subtle and poorly recognized memory failure and slowly becomes more severe and incapacitating (see GeneReviews in Web Resources). AD is genetically heterogeneous and has an estimated heritability of 60%–80%.[17] Although GWASs have successfully identified disease-associated loci, each locus accounts for only a small fraction of AD susceptibility, and a large proportion of AD heritability still remains unexplained.[18] There is great interest in investigating the role of rare variants in the etiology of AD. Application of the RV-GDT identified suggestive associations between AD and rare variants in *AXIN1* (MIM: 603816; GenBank: NM_003502.3) and *TNK1* (MIM: 608076; GenBank: NM_001251902.1). An association between AD and a common variant in *TNK1* was previously identified,[19] and evidence of TNK1 involvement in AD has been further supported by experimental studies.[20,21] Although *AXIN1* has not been previously shown to be associated with AD, experimental studies suggest that there might be a link between *AXIN1* and AD.[22–25] These findings could provide new insights in the understanding of AD etiology.

Material and Methods

RV-GDT

The GDT utilizes the genotype differences in all discordant relative pairs within a family to assess the association.[14] For the i^{th} pedigree, n_i is the total number of genotyped individuals, n_i^{A} is the number of genotyped individuals who are affected, and $n_i^{\text{U}} = n_i - n_i^{\text{A}}$ is the number of genotyped individuals who are unaffected. The single-variant GDT statistic for the i^{th} pedigree is defined as

$$G_i = \sum_{j=1}^{n_i^{\text{A}}} \sum_{k=1}^{n_i^{\text{U}}} (X_{ij} - X_{ik}) C_{ijk},$$

where X_{ij} and X_{ik} are the numbers of minor alleles in the j^{th} and k^{th} unaffected individuals, respectively. C_{ijk} is $1/n_i$ if no covariates are included in the model; otherwise, it is given as

$$C_{ijk} = \frac{8}{n_i} \frac{\exp\left\{ (Z_{ij} - Z_{ik})^T \alpha \right\}}{\left(1 + \exp\left\{ (Z_{ij} - Z_{ik})^T \alpha \right\}\right)^3},$$

where Z_{ij} and Z_{ik} are the covariate vectors for the j^{th} and k^{th} unaffected individuals, respectively. Values in vector α are log odds ratios (ORs) for the association between the covariates and the trait, which are estimated from a logistic regression model that includes the phenotypes and covariates. The single-variant GDT statistic for a dataset with N independent families is given as

$$Z^{\text{GDT}} = \frac{\sum_{i=1}^{N} G_i}{\sqrt{\sum_{i=1}^{N} G_i^2}},$$

which asymptotically follows a standard normal distribution under the null hypothesis of no association.[14]

It has been shown that an association test performed with individual rare variants (minor allele frequency [MAF] $\leq$ 1%) is underpowered[26] given the small number of observed alternative alleles and the stringent multiple-testing correction. In order to increase power, it is advantageous to aggregate rare-variant information

across a region. Similar to the burden of rare variants (BRV) method,[27] here we aggregate the contributions of M variants across a region of interest, which is given as

$$G_{i\cdot} = \sum_{m=1}^{M} G_{im},$$

where G_{im} is the single-locus GDT statistic on the m^{th} variant for the i^{th} pedigree. In addition to aggregating the information across multiple variants, an alternative approach is to take a weighted sum of the contributions of each single variant, i.e.,

$$G_{i\cdot} = \sum_{m=1}^{M} w_m G_{im},$$

where w_m is the weight assigned to the m^{th} variant. The weights can be inferred from MAF in control[28] or complete[29] samples or from the predicted functionality of the variant,[30] such as the C-score from the Combined Annotation Dependent Depletion (CADD) tool.[31] The RV-GDT statistic is defined as

$$Z^{\text{GDT}} = \frac{\sum_{i=1}^{N} G_{i\cdot}}{\sqrt{\sum_{i=1}^{N} G_{i\cdot}^2}}.$$

To infer its statistical significance, we apply a permutation procedure to derive empirical p values. We fix the genotypes and covariates for each pedigree and randomly shuffle the phenotypes among subjects within each pedigree. The vector α, a covariate adjustment, is also re-calculated after the phenotypes are shuffled. To reduce computational time, we use an adaptive permutation that evaluates the estimated p value at pre-defined checkpoints and stops further permutations for non-significant tests.

For rare-variant association tests, a common approach is to select a fixed MAF threshold and analyze only variants that meet the criterion. To determine whether the MAF of a variant is below the cutoff, one can obtain information on allele frequencies either from the sample or from public databases, e.g., the Exome Aggregation Consortium (ExAC) Browser. To avoid the implicit assumption about the relationship between allele frequency and variant functionality, we can use the variable threshold as an alternative approach to determine which variants should be analyzed.[30] The intuition is that there exists an unknown threshold T for which variants with MAF $< T$ are more likely to be functional than variants with MAF $> T$. In this approach, the RV-GDT score is calculated for each allele-frequency threshold, and the final RV-GDT statistic is defined as the maximum score. The p values must be inferred empirically for multiple-testing correction.

RV-PDT

The PDT takes into account the difference in the number of transmitted and non-transmitted minor alleles from parents to affected siblings and the difference in the number of minor alleles between affected and unaffected siblings.[15,16] For the i^{th} pedigree, n_T is the number of case-parent trios from informative nuclear families (at least one affected child, both parents genotyped at the marker, and at least one heterozygous parent), and n_S is the number of informative discordant sib-pairs (at least one affected and one unaffected sibling with different marker genotypes). The single-variant PDT statistic for the i^{th} pedigree is defined as

$$P_i = \sum_{k=1}^{n_T} T_{ik} + \sum_{j=1}^{n_S} S_{ij},$$

where T_{ik} is the difference between the number of minor alleles transmitted and the number of minor alleles not transmitted from a heterozygous parent in the k^{th} trio, and S_{ij} is the difference between the number of minor alleles in affected siblings and those in unaffected siblings in the j^{th} discordant sib-pair. Let N be the number of independent informative pedigrees (at least one informative nuclear family and/or discordant sibship); then, the single-variant PDT statistic is defined as

$$Z^{\text{PDT}} = \frac{\sum_{i=1}^{N} P_i}{\sqrt{\sum_{i=1}^{N} P_i^2}}.$$

We can sum the contributions of each single variant across a region by

$$P_{i\cdot} = \sum_{m=1}^{M} P_{im},$$

where P_{im} is the single-variant PDT statistic on the m^{th} variant in the i^{th} pedigree, and M is the total number of variants in the region of interest. We can also consider the weighted sum of multiple variants, which is similar to the RV-GDT. The RV-PDT statistic is defined as

$$Z^{\text{PDT}} = \frac{\sum_{i=1}^{N} P_{i\cdot}}{\sqrt{\sum_{i=1}^{N} P_{i\cdot}^2}}.$$

The p values can be inferred empirically via haplotype permutation, which is able to control the type I error in the TDT-based tests.[7] For each pedigree, we fix the founders' genotypes and obtain the genotypes of the non-founders by pairing a randomly selected paternal and maternal haplotype. Adaptive permutation can also be used to reduce computational time.

Simulation Framework

Generation of Family Data

To evaluate the performance of the RV-GDT, we compared it to other family-based association methods, including FBAT, RV-PDT, and Epstein's ASP method, through simulating and analyzing family-based exome sequence data. Genotypes were simulated for autosomal genes across the genome on the basis of the observed variant sites and their corresponding MAFs obtained from the non-Finnish European and African and African American populations recorded in the ExAC Browser[32] (17,987 autosomal genes for 33,370 subjects in the non-Finnish European population and 17,892 autosomal genes for 5,203 subjects in African and African American populations). Family data were generated with RarePedSim,[33] which is able to effectively simulate sequence-based genotypes for any arbitrarily complex pedigree structure by conditioning on observed phenotypic data and incorporating a user-specified phenotype model and variant information. Using ExAC MAFs, we generated genotypes under linkage equilibrium (after assigning haplotypes to founders), which then segregated within the generated pedigrees.

Disease Model

The disease prevalence is assumed to be 1%, and the disease status for each subject is assigned on the basis of the multisite genotypes consisting of rare nonsense, missense, and splice-site variants (MAF $\leq$ 1% in its corresponding ExAC population). An OR of 2.5 is assigned to each variant that is deemed causal, and the disease probabilities of all variants within a gene are computed on the basis of a multiplicative mode of inheritance.[33]

Evaluation of Type I Error

To evaluate the type I error rate of the RV-GDT and RV-PDT, we set the OR of the causal variant to 1 (no association between genetic

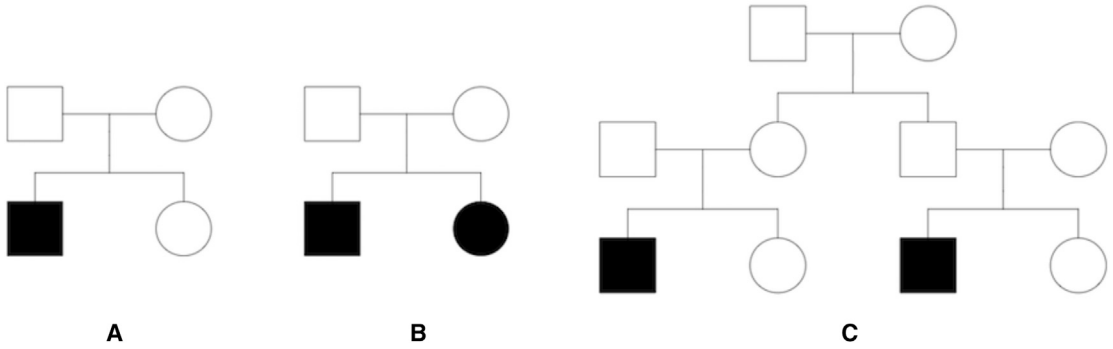


Figure 1. Pedigree Structures Used in the Simulation Studies
(A) Discordant nuclear sib-pair: the family contains parents, an affected child, and an unaffected child.
(B) Affected sib-pair: the nuclear family contains parents and two affected children.
(C) Extended three-generation pedigree.

variant and phenotype) and used the variant information from the ExAC non-Finnish European population to generate the family data. We considered four different types of family data: 1,000 nuclear families with one affected child and one unaffected child (discordant sib-pair; Figure 1A), 1,000 nuclear families with two affected children (affected sib-pair; Figure 1B), 1,000 extended pedigrees with two affected individuals in the third generation (extended pedigree; Figure 1C), and a mixture of these three pedigree structures (500 discordant sib-pairs, 250 affected sib-pairs, and 250 extended pedigrees). Genotype data were simulated for all autosomal genes across the genome, and genes with at least three informative variant sites were analyzed. Type I error rates were evaluated as the proportion of genes with a p value less than 0.05 and 0.005. The p values were obtained empirically via 100,000 permutations. Moreover, to demonstrate that the RV-GDT can appropriately handle pedigrees with missing data, we analyzed the data after removing genotype data from 50% of the founders.

We also evaluated the type I error rates of RV-GDT when there is population admixture or population substructure. Genotype data were simulated for 17,873 autosomal genes present in both the ExAC non-Finnish European and African and African American populations. To generate family data from an admixed population, we randomly generated the haplotypes of the founders in each pedigree from the European or African population with a probability of 20% or 80%, respectively. To generate family data with population substructure, we simulated 50% of the families with ExAC non-Finnish European variant information and simulated the other 50% of the families with ExAC African and African American variant information.

We also evaluated the type I error rates of RareIBD (version 1.1)[13] in our simulation framework. We applied a maximum of 100,000 inheritance vector samplings to pre-compute the mean and standard deviation of the statistic and used 10,000 gene-dropping permutations to estimate the p values.

Power Evaluation

To evaluate the power of RV-GDT, we simulated 1,000 families of each pedigree structure shown in Figure 1 and a mixture of the three pedigree structures (500 discordant sib-pairs, 250 affected sib-pairs, and 250 extended pedigrees) by using ExAC non-Finnish European variant information. Genotype data were generated for autosomal genes across the genome when 75% of the rare nonsense, missense, and splice-site variants were randomly selected to be causal with an OR of 2.5. Genes with at least three informative variant sites were analyzed, and power was evaluated as the proportion of genes with a p value less than 0.05. To assess the influence of missing founder genotype data on power, we used a probability to determine which founders were missing all of their genotype data and considered three different probabilities (0%, 25%, and 50%). To further evaluate the power of RV-GDT for extended pedigrees in which family members are missing genotype data, we determined whether each parent, regardless of being a founder or non-founder (subjects in the first two generations), had a 0%, 25%, 50%, or 75% probability of missing all of their genotype data. Power was evaluated when 50%, 75%, and 100% of the randomly selected rare nonsense, missense, and splice-site variants were causal with an OR of 2.5.

We compared the power of the RV-GDT method to that of other family-based association tests, including Epstein's ASP method (one-sided test),[12] RV-PDT, and FBAT[10] (version 2.0.4, with the "-v0" option to calculate unweighted rare-variant statistics). Epstein's ASP method requires estimation of identity by descent (IBD) sharing between affected siblings. Because the IBD-sharing information is known in the simulated data, we used the exact IBD sharing in the power evaluation. The phase information generated during simulation of family data was used for haplotype permutation in the RV-PDT. Both Epstein's ASP method and FBAT software report analytical p values. For RV-GDT and RV-PDT, we performed one-sided tests and obtained p values empirically by performing 2,000 permutations.

Application to AD Data

Description of the ADSP Data

The WGS data from 112 families were downloaded from dbGaP: phs000572.v6.p4. Study subjects with phenotypes coded as "definite AD," "probable AD," or "possible AD" were labeled as affected, and all other subjects were labeled as unaffected. The mean age of onset for AD was 72.63 years with a standard deviation of 8.46. In all 112 families selected for generation of WGS data, no more than 75% of affected members were positive for *APOE4* (MIM: 107741), and no family members were homozygous for *APOE4*. Families in whom all sequenced subjects were affected were excluded from our analysis, which resulted in 81 families (21 nuclear and 60 extended), including 414 subjects with WGS data (316 affected and 98 unaffected; 167 male and 247 female) and 418 subjects without WGS data (22 affected and 396 unaffected; 221 male and 197 female). Their pedigrees and ethnicities (46 Dominican,

Table 1. Type I Error Rate for RV-GDT at α Levels of 0.05 and 0.005

	Discordant Sib-Pair		Affected Sib-Pair		Extended Pedigree		Mixed Family Types	
	$\alpha = 0.05$	$\alpha = 0.005$	$\alpha = 0.05$	$\alpha = 0.005$	$\alpha = 0.05$	$\alpha = 0.005$	$\alpha = 0.05$	$\alpha = 0.005$
RV-GDT	0.047	0.0048	0.050	0.0051	0.051	0.0049	0.051	0.0048
Each Founder Has a 50% Probability of Missing All Genotype Data								
RV-GDT	0.051	0.0050	0.051	0.0049	0.049	0.0047	0.051	0.0047
80% African and 20% European Population Admixture								
RV-GDT	0.051	0.0049	0.051	0.0048	0.048	0.0047	0.050	0.0051
50% African and 50% European Families								
RV-GDT	0.050	0.0051	0.049	0.0053	0.048	0.0052	0.051	0.0048

We simulated 1,000 families for each pedigree structure shown in Figure 1 and mixed pedigree structures. Genotype data were generated for all autosomal genes across the genome with an OR of 1.0, and type I error rate was defined as the proportion of genes with a p value less than 0.05 or 0.005. We used variant information for 17,987 autosomal genes from the ExAC non-Finnish European population to generate family data when each founder had a 0% or 50% probability of missing genotype data. We used variant information for 17,873 autosomal genes present in both the ExAC non-Finnish European population and African and African American populations to generate family data with population admixture and substructure.

31 of European descent, 2 Puerto Rican, 1 Dutch isolate, and 1 African American) are shown in Figure S2 and Table S2, respectively.

Generation, Quality Control, Annotation, and Analysis of WGS Data

Genomic DNA were sequenced at the Broad Institute, Human Genome Sequencing Center at the Baylor College of Medicine, and McDonnell Genome Institute at Washington University. Reads were mapped to the GRCh37 reference genome assembly with the Burrows-Wheeler Aligner.[34] BAM files from all three sequencing centers were collected, and genotype calling and primary quality control (QC) were performed by both the Broad Institute (Broad pipeline) and the Baylor College of Medicine Human Genome Sequencing Center (Baylor pipeline). The Broad and Baylor pipelines used the Genome Analysis Toolkit HaplotypeCaller[35] and Atlas2 software,[36] respectively, for genotype calling.

For the Broad pipeline, those variants that did not "pass" Variant Quality Score Recalibration were removed. For the Baylor pipeline, variants with a mapping score < 0.80 were deleted, and genotypes that did not "pass" the Sample Genotype Filter or that had a read depth < 10 or an out-of-range ratio of variant reads to total read depth (≤ 0.75 or ≥ 0.25) were deleted. For both pipelines, the following types of variants were excluded: monomorphic variants, those with a call rate $\leq 80\%$, those with excessive heterozygosity, and those with an average mean read depth > 500.

Once primary QC was completed for the Broad and Baylor pipelines, consensus genotypes were determined by keeping concordant variants and excluding variants in which a different alternative allele was called between the two pipelines. After consensus calling, a second round of variant-level QC was applied to remove any variants that were monomorphic, had >20% missing genotypes, or had an excessive number of heterozygous genotypes. QC was performed by the QC working group of the ADSP.

We used the RefGene database to select variants located in exon regions and included only single-nucleotide variants (SNVs) within the autosomal exome coding region in our analysis. Mendelian inconsistencies were identified and removed with PLINK software.[37] Gene regions were assigned according to RefSeq definitions, and ANNOVAR was used to annotate variant sites.[38] Variants within regions containing copy-number variants or pseudogenes were excluded, and variants that were either nonsynonymous or putative splice sites were included in the analysis. Only variants that were absent or had a MAF $\leq 2\%$ in the ExAC Browser were analyzed. We used Variant Association Tools (VAT) to perform the variant-selection procedures described above.[39] Only genes with at least three variant sites were analyzed, leaving 8,891 genes for analysis.

Results

Type I Error Rate

When the family data were generated under the null hypothesis of no association with ExAC non-Finish European variant information, type I error of the RV-GDT was well controlled for the family structures shown in Figure 1 and also for the mixed family structures. Type I error rates were evaluated at $\alpha = 0.05$ and $\alpha = 0.005$, the results of which are shown in Table 1. Additionally, the quantile-quantile plot of the $-\log_{10}$p values demonstrates that type I error was well controlled (Figure S1). The type I error was also well controlled for RV-PDT for all family structures (data not shown). When founders had a 50% probability of missing all of their genotype data, the RV-GDT method still had proper control of type I error rates for all pedigree and mixed family structures (Table 1 and Figure S1).

To demonstrate that the RV-GDT can adequately control for population admixture, we generated data for pedigrees with 80% African and 20% European admixture. We also evaluated whether type I error was well controlled in the presence of European-African substructure by simulating and analyzing data with 50% European pedigrees and 50% African pedigrees. For both scenarios, type I error was well controlled (Table 1 and Figure S1), suggesting that the RV-GDT is robust to population admixture and substructure.

We also evaluated the type I error rates of RareIBD by using simulated family data, and the results are shown in Table S1. We observed inflated type I error rates for discordant sib-pairs and extended pedigrees in each scenario evaluated (i.e., non-Finnish European population, missing founder data, African-European substructure, and

Table 2. Power Comparison of Epstein's ASP Method, RV-PDT, FBAT, and RV-GDT when Founders Are Missing Different Proportions of Genotype Data

	Discordant Sib-Pair			Affected Sib-Pair			Extended Pedigree			Mixed Family Types		
Method	**0%[a]**	**25%[a]**	**50%[a]**	**0%**	**25%**	**50%**	**0%**	**25%**	**50%**	**0%**	**25%**	**50%**
Epstein's ASP[b]	–	–	–	0.23	0.23	0.23	–	–	–	0.10	0.10	0.10
FBAT	0.46	0.40	0.35	0.80	0.71	0.54	0.42	0.38	0.34	0.61	0.52	0.38
RV-PDT	0.51	0.45	0.42	0.79	0.70	0.52	0.48	0.45	0.41	0.63	0.55	0.43
RV-GDT	0.53	0.48	0.45	0.81	0.75	0.62	0.64	0.62	0.60	0.66	0.62	0.56

Genetic variant data were generated for 1,000 families of each pedigree structure shown in Figure 1 and mixed pedigree structures with the use of ExAC non-Finnish European variant information. Genotype data were generated for 17,987 autosomal genes across the genome when 75% of the rare nonsense, missense, and splice-site variants were randomly selected to be causal with an OR of 2.5, and power was evaluated as the proportion of genes with a p value < 0.05.

[a]Probability that each founder is missing all genotype data.

[b]Power was evaluated under the assumption that the exact IBD sharing between affected sib-pairs is known. Unknown IBD sharing and non-simulated data would reduce the power.

African-European admixture) and deflated type I error rates for affected sib-pairs. For example, when data were generated with substructure (50% African and 50% European) and analyzed, the discordant sib-pairs and extended pedigrees had type I error rates of 0.063 and 0.477, respectively, at $\alpha = 0.05$, whereas the affected sib-pairs had a type I error rate of 0.030.

Power Evaluation

We compared the power of RV-GDT with that of Epstein's ASP method, RV-PDT, and FBAT for a variety of pedigree structures (Figure 1) when 75% of the rare variants were causal with an OR of 2.5. Additionally, we evaluated the effect of missing founder data on power by using different probabilities (0%, 25%, and 50%) to determine whether a founder was missing genotype data. The power of RV-GDT and other methods when founders were missing genotype data is shown in Table 2. When none of the founders were missing data, the differences in power between RV-GDT, RV-PDT, and FBAT were small for nuclear families (discordant and affected sib-pairs), but all methods were considerably more powerful than Epstein's ASP method. When 1,000 discordant sib-pairs were analyzed, the power of FBAT, RV-PDT, and RV-GDT was 0.46, 0.51, and 0.53, respectively, whereas Epstein's ASP method was unable to analyze these data because there were no affected sibships. When 1,000 affected sib-pairs were analyzed, the power of FBAT, RV-PDT, and RV-GDT was 0.80, 0.79, and 0.81, respectively, whereas the power of Epstein's ASP method was 0.24. However, RV-GDT had considerably higher power than the other methods, i.e., FBAT and RV-PDT, which can analyze extended pedigrees. When 1,000 extended pedigrees were analyzed, the power of FBAT and RV-PDT was 0.42 and 0.48, respectively, whereas the power of RV-GDT was 0.64. Moreover, the power of RV-GDT was higher than that of the other methods when founders were missing genotype data. When ~25% of the founders were missing their genotype data, the power of FBAT, RV-PDT, and RV-GDT for 1,000 extended pedigrees was 0.38, 0.45, and 0.62, respectively; when the missing probability was increased to 50%, the power of FBAT and RV-PDT was 0.34 and 0.41, respectively, whereas RV-GDT still had a power of 0.60.

To further evaluate the power of RV-GDT when family members are missing genotype data, we simulated 1,000 extended pedigrees under different proportions of causal variants (50%, 75%, and 100%) with an OR of 2.5. When none of the pedigree members were missing genotype data, the power of the RV-GDT was higher than that of RV-PDT and FBAT (Figures 2A–2C). For example, when 100% of the rare nonsense, missense, and splice-site variants were causal and none of the pedigree members were missing genotype data, the power of FBAT, RV-PDT, and RV-GDT was 0.63, 0.69, and 0.81, respectively (Figure 2C). When founders had a 25%, 50%, or 75% probability of missing genotype data, the power decreased for each method as the percentage of founders missing genotype data increased (Figures 2A–2C). However, the RV-GDT still had considerably higher power than the other methods, not only because its initial power was higher than that of the other methods but also because it lost less power as the percentage of founders missing genotype data increased. For example, when 100% of the rare variants were causal and the probability that founders were missing genotype data was increased from 0 to 50% (Figure 2C), the FBAT and RV-PDT power was reduced by 12.70% (from 0.63 to 0.55) and 10.14% (from 0.69 to 0.62), respectively, whereas RV-GDT had a 3.84% (from 0.81 to 0.78) loss of power. Similar patterns of decreasing power were observed when family members in the first two generations had a 25%, 50%, or 75% probability of missing all of their genotype data, regardless of whether they were founders or non-founders (Figures 2D–2F). For the model in which 100% of the rare variants were causal (Figure 2F), the power of the FBAT and RV-PDT was reduced by 20.63% (from 0.63 to 0.50) and 15.94% (from 0.69 to 0.58), respectively, when the probability of individuals missing their genotype data was increased from 0% to 50%, whereas the power for RV-GDT was reduced by 8.64% (from 0.81 to 0.74).

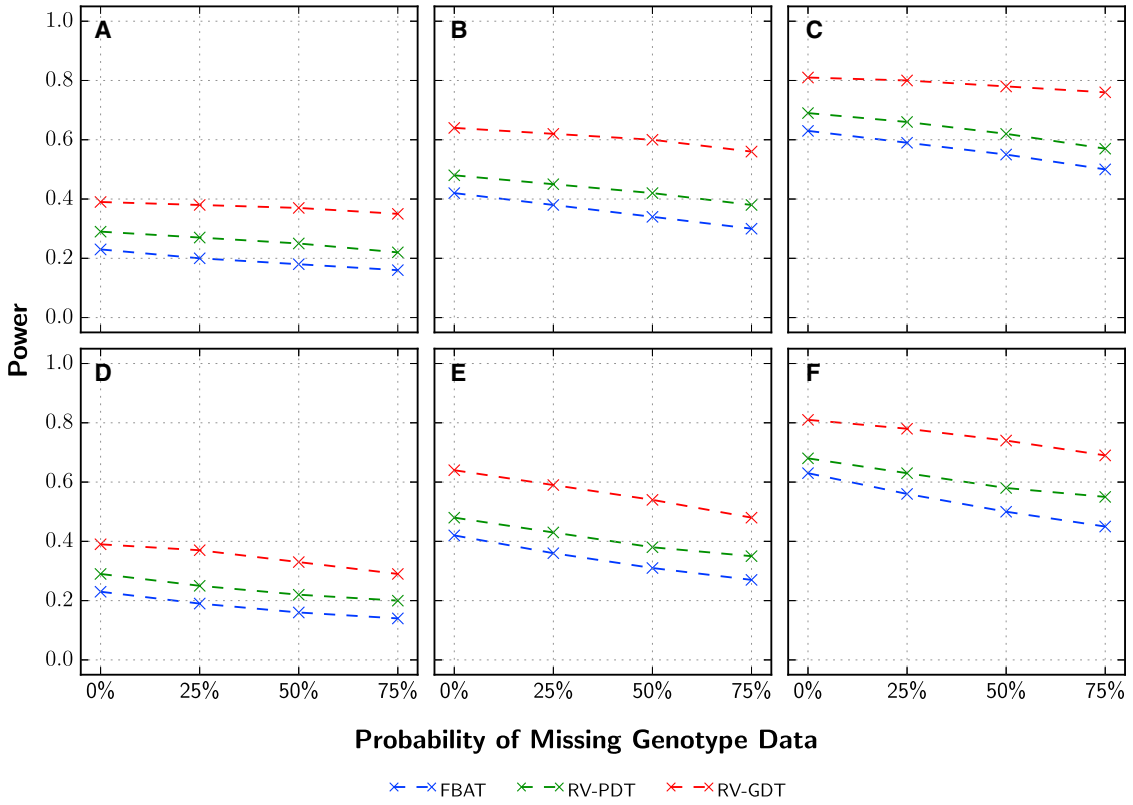


Figure 2. Power Comparisons of FBAT, RV-PDT, and RV-GDT for Extended Pedigrees with Family Members Missing Genotype Data
Genetic variant data were generated for 1,000 extended pedigrees with ExAC non-Finnish European variant information. Different proportions of the rare nonsense, missense, and splice-site variants were deemed to be causal: 50% (A and D), 75% (B and E), and 100% (C and F) with an OR of 2.5.
(A–C) Power comparisons when the probability that each founder was missing all genotype data ranged from 0% to 75%.
(D–F) Power comparisons when the probability that each parent (founder or non-founder) was missing all genotype data ranged from 0% to 75%.

Application to AD Data

We applied the RV-GDT method to analyze WGS data from the ADSP dataset. All pedigrees have at least one parental family member who is missing WGS data. Given the small sample size, application of the RV-GDT did not detect associations with exome-wide significance of 2.50×10^{-6} (Bonferroni correction for 20,000 genes). The most significant associations with AD were observed for *MARCH10* (MIM: 613337; GenBank: NM_001100875.1; p value = 5.0×10^{-5}), *AMBN* (MIM: 601259; GenBank: NM_016519.5; p value = 9.0×10^{-5}), *TCOF1* (MIM: 606847; GenBank: NM_000356.3; p value = 2.0×10^{-4}), *AXIN1* (p value = 2.5×10^{-4}), and *TNK1* (p value = 6.0×10^{-4}). The ExAC MAFs of the variants within these genes, along with annotations from dbNSFP (version 2.9)[40] (which include GERP,[41] PhyloP,[42] and CADD[31] scores and Functional Analysis through Hidden Markov Models [fathmm],[43] MutationTaster,[44] PolyPhen-2,[45] PROVEAN,[46] and SIFT[47] prediction), are shown in Table S3–S7.

Eight missense variants were observed in *MARCH10* (Table S3). 53 alternative alleles were observed in family members with AD, whereas six were observed in unaffected individuals. Except for SNV rs13801568, which had the same number of alternative alleles in affected and unaffected family members, all other variant sites had higher alternative-allele counts in affected subjects than in unaffected family members. Five SNVs occurred at conserved nucleotides, and two SNVs were deemed to be deleterious by at least three of six bioinformatics tools (CADD, fathmm, MutationTaster, PolyPhen-2, PROVEAN, and SIFT). Eight nonsynonymous variants were observed in *AMBN*, and 41 and 4 alternative alleles of these variants were observed in affected and unaffected family members, respectively (Table S4). All SNVs except rs150017698 had a higher number of alternative alleles in affected family members than in unaffected family members. Four variants in *AMBN* were deemed to be conserved by both GERP and PhyloP and were predicted to be deleterious by at least three of six bioinformatics tools. 19 missense variants were observed in *TCOF1*—78 and 12 alternative alleles were observed in affected and unaffected family members, respectively (Table S5). Although only four variants were deemed to be conserved by both GERP and PhyloP, 12 variants were judged to be deleterious by at least three of six bioinformatics tools. *AXIN1* included eight nonsynonymous variants; 34 alternative alleles were observed in affected family members, and one alternative allele was

observed in an unaffected family member, and all variants had a higher number of alternative alleles in affected family members than in unaffected family members (Table S6). Six variants occurred at conserved nucleotides; however, none of the eight variants were deemed to be deleterious by at least three of six bioinformatics tools. Of the six variants observed in *TNK1*, 21 alternative alleles were observed in affected family members, and no alternative alleles were observed in unaffected family members (Table S7). Four variants were deemed conserved by both GERP and PhyloP, and four variants were deemed to be deleterious by at least three of six bioinformatics tools. rs201180891 occurs at a highly conserved residue and is predicted to be deleterious in all available bioinformatics results, and four affected family members were observed to be carriers of the alternative allele of this variant. The variant c.923T>A (p.Met308Lys) (GenBank: NM_003985.3) was not found in ExAC samples, and two individuals affected by AD are carriers of an alternative allele. This variant also occurs at a highly conserved residue and is predicted to be deleterious by five of six bioinformatics tools.

Discussion

We extended the family-based GDT to allow for the analysis of rare variants so that the method could be applied to association analysis of WES or exome sequence data. Our simulation studies demonstrated that the RV-GDT has well-controlled type I error rates, even when applied to admixed populations or populations with substructure. The RV-GDT has greater power than other family-based rare-variant association methods and is substantially more powerful when applied to extended pedigrees and/or pedigrees in which family members are missing genotype data. There are advantages to performing family-based association studies over employing population-based designs. Family-based studies can have higher power given an equivalent number of cases because they can involve more pathogenic susceptibility variants with larger effect sizes than those observed for sporadic disease.[3] Additionally, many family-based association methods can control for population admixture and substructure on a local level, whereas for population-based designs, the inclusion of principal or multi-dimensional scaling components can control for population admixture and substructure only on a global level, which might not be sufficient for rare-variant association studies.[48] However, family-based designs do have their drawbacks; compared with population-based studies, they require more resources for the recruitment of probands and their relatives. For family-based designs, genotype data are often missing because of unascertainable family members, e.g., non-paternity and deceased parents from late-onset disease. Usually, family data are composed of many different types of pedigree structures, and there are family members without genotype data, as observed in the 81 AD-affected families analyzed here (see Figure S2). The ability to analyze family data, including extended pedigrees and/or pedigrees in which family members are missing genotypes data, with minimal loss of power makes the RV-GDT an extremely valuable method for detecting associations and elucidating the genetic etiology of complex traits.

RareIBD has a main assumption that only one founder in each family carries a mutation for a specific rare variant. Violation of this assumption will result in an inflated test statistic.[13] Complex traits, such as AD and coronary heart disease, have relatively high prevalences; therefore, a pedigree might have multiple affected individuals who do not have the same causal variants. Unlike for Mendelian diseases, the assumption that only one founder in a family carries the pathogenic susceptibility variant might not be valid for complex traits. Despite the fact that RareIBD (version 1.1) excludes variants violating this assumption, our simulations showed extremely inflated type I error rates for extended pedigrees. Our simulation framework is based on ExAC variant information for all genes, which represents exome sequence data more realistically than generating data for a single genomic region by using a population demographic model (the latter was previously used for evaluating type I error of RareIBD[13]).We also evaluated the power of RareIBD when 75% of rare nonsense, missense, and splice-site variants were randomly selected to be causal with an OR of 2.5. Although type I error was slightly inflated, the power of RareIBD was 0.32 when 1,000 discordant sib-pairs were analyzed, whereas the power of FBAT, RV-PDT, and RV-GDT was 0.46, 0.51, and 0.53, respectively. When 1,000 affected sib-pairs were analyzed, the power of RareIBD was 0.79, comparable to that of FBAT (0.80), RV-PDT (0.79), and RV-GDT (0.81). We did not evaluate power for extended pedigrees because it would not be valid given the extremely inflated type I error rates for RareIBD.

RareIBD (version 1.1) can analyze only families from single populations because of differences in the allelic spectrum between populations. Families of different ancestries need to be analyzed separately, and meta-analysis needs to be performed, which can lead to a loss of power. Genetic studies of complex traits are often composed of families ascertained from multiple populations. For example, ADSP includes both families of European descent and African American and Dominican families. Even the analysis of families of European descent can still be problematic, given that for rare variants, the allelic spectrum can differ greatly even between adjacent populations, e.g., Ashkenazi and other Eastern European populations.[49]

Neither RV-GDT nor FBAT requires haplotyping, IBD-sharing estimation, or imputation of missing genotypes, which avoids the potential decrease in power due to loss of information and/or inclusion of noise. For Epstein's ASP method, IBD sharing between siblings must be estimated before statistical analysis. RV-PDT requires haplotype information in order to perform haplotype permutation, which is necessary to control type I error in the

presence of LD between variants.[7] Even though some algorithms can perform haplotyping and/or IBD-sharing estimation with acceptable accuracy, the potential loss of information and inclusion of noise can greatly jeopardize power. In our power evaluation of Epstein's ASP method, we used exact IBD sharing in the simulated data. There would be a loss of power if the IBD-sharing information were inferred, for example, via MERLIN. Especially when founders are missing, it would introduce more uncertainties and lead to a further decrease in power. RareIBD requires family data without missing genotypes. Missing genotypes are imputed, and the most likely genotypes are analyzed. Although family-based imputation can reach relatively high accuracy, association testing with the most likely imputed genotype can lead to type I error inflation.[50] More experiments are needed to investigate how imputed genotypes can be correctly incorporated in family-based association studies.

Adjustment for non-confounding covariates that are known to influence the trait can reduce spurious associations due to sampling artifacts or biases in study design.[51] However, caution should be exercised for decisions about whether to incorporate covariates in the association analysis of binary traits. It has been shown for GWASs that including known covariates can reduce the power to identify associated variants when the disease prevalence is low. On the other hand, including non-confounding predictive covariates when disease prevalence is sufficiently high (>20%) will often lead to an increase in power.[52] The RV-GDT can incorporate covariates in the analysis, but it does not provide an evaluation of covariate significance and also cannot be used for covariate selection. The FBAT can also incorporate covariates, whereas Epstein's ASP method and RV-PDT have not been extended to adjust for covariates.

It was previously shown that integrating information on variant allele frequencies from population-based data into family-based studies can be useful for association studies of rare variants.[53] Jiang et al. suggested incorporating population-control-based weights into the TDT framework to potentially up-weight pathogenic susceptibility variants, down-weight neutral variants, and also assign the direction of the effect for pathogenic variants.[8] In our simulation, no improvement in power was observed for either RV-GDT or RV-PDT when weights were incorporated with data from population controls. In fact, the incorporation of weights from population controls led to slightly less power than not using any weights. Genetic data for 1,000 extended pedigrees were simulated with ExAC non-Finnish European variant information, and 75% of the rare nonsense, missense, and splice-site variants were randomly selected to be causal with an OR of 2.5. We also generated 20,000 population controls, and the weights inferred the control data, as suggested by Jiang et al. The power of the RV-GDT decreased from 0.81 to 0.78, and the power of the RV-PDT also decreased from 0.79 to 0.77. This is not surprising given that it has previously been shown that decreases in power can occur when weights are not optimal.[54] In our simulations, the controls were generated from the same population as the family data, but because of random variability, they were not always optimal. The reduction in power could be even greater if controls are drawn from a different population. Moreover, how to handle variants that are not present in population controls is a practical problem that needs to be addressed. For example, 10.41% of variants analyzed in the AD pedigrees are not present in the ExAC Browser, which is one of the largest publicly available databases.

The application of the RV-GDT in the analysis of AD pedigrees highlights its applicability to family-based studies. The pedigree structures of this dataset are highly heterozygous, and each pedigree has at least one family member who has not been sequenced (see Figure S2). RV-GDT has fewer constraints on pedigree structure than Epstein's ASP method and RV-PDT, and it can analyze any pedigree as long as it includes both affected and unaffected subjects. FBAT can analyze most pedigree structures, but those with missing parental data often cannot be analyzed, especially when there are no unaffected offspring. Epstein's ASP method is only applicable to analyzing affected sibships in nuclear pedigrees, and it needs the estimated IBD sharing between sibships, which is problematic for pedigrees missing parental genotypes. Without parental genotypes, IBD sharing must be estimated from identity by state (IBS) sharing and variant allele frequencies. It is unclear how well Epstein's ASP method performs in this situation. The majority of affected sibships in the AD pedigrees are missing both parental genotypes, and no single affected sibship has genotype data for both parents. The RV-PDT requires nuclear families with informative case-parent trios and/or discordant sib-pairs to detect association, but the haplotype permutation is necessary for the RV-PDT to control type I error,[7] and the haplotype permutation needs the complete nuclear family or reconstructed haplotypes for the missing founders. No single AD pedigree has complete parental WGS data, and the RV-PDT hasn't been extended to analyze family data with reconstructed haplotypes; therefore, it is not possible to analyze the AD pedigrees with the RV-PDT. The AD data could not be analyzed with RareIBD given the extreme inflation of type I error for extended pedigrees. Of the 81 AD-affected families, 25 cannot be analyzed by the FBAT (see Figure S2). Moreover, the FBAT returned "NaN" (not a number) when the genotype data for *TCOF1* in the AD data were analyzed. The p values of FBAT for *MARCH10*, *AMBN*, *AXIN1*, and *TNK1* were 0.06, 0.30, 0.04, and 0.04, respectively, and all p values were considerably less significant than those obtained from the RV-GDT.

The application of the RV-GDT on WGS data from 81 AD-affected families identified potential involvement of *AXIN1* (16p13.3) and *TNK1* (17p13.1) in this neurodegenerative disease. Previously, a SNP-based association study detected an association between rs1554948, which

is within *TNK1*, and late-onset AD.[19] Our study implicates multiple rare variants in *TNK1* as a potential underlying cause of AD. We observed 21 alternative alleles in 21 affected family members and no alternative alleles in unaffected family members. Out of 21 family members carrying *TNK1* alternative alleles, only two of them were *APOE4* positive. The association between rare variants in *TNK1* and AD is consistent with functional studies of its protein. *TNK1* encodes a non-receptor tyrosine kinase and mediates intracellular signaling. Activated TNK1 has been reported to facilitate tumor necrosis factor alpha (TNFα)-induced apoptosis, which suggests its involvement in TNFα signaling and neuronal cell death.[20] The involvement of TNK1 in AD pathogenesis could also be through its interaction with phospholipase C (PLC). TNK1 has been reported to be associated with PLC gamma 1,[21] and multiple studies have observed aberrant PLC activity in AD brains.[19] Identification of the involvement of *TNK1* in AD etiology provides new insights that could be used for prevention and treatment. Although the association between AD and variants in *AXIN1* has not been reported previously, the identification of *AXIN1* is also consistent with experimental findings. *AXIN1* encodes a scaffolding protein that plays a critical role in regulating GSK3-mediated phosphorylation of the protein tau.[22] Axin negatively affects tau phosphorylation by GSK3,[23] and phosphorylated tau has a decreased capacity to bind and stabilize microtubules.[24] The abnormally phosphorylated tau has been observed in populations of AD-affected individuals at clinicopathological levels.[25] These findings suggest that abnormal expression of AXIN1 might contribute to tau pathology in AD. No previous association studies have implicated *MARCH10*, *AMBN*, or *TCOF1* in the etiology of AD. Additionally, no functional studies support the involvement of these genes in AD etiology. The associations between AD and *MARCH10*, *AMBN*, and *TCOF1* might be false positives, and future replication studies will allow for the elucidation of whether these genes are involved in the etiology of AD.

The WGS data on the analyzed 81 families is part of the ADSP Discovery Phase, which will be followed by the Discovery Extension Phase and the Follow-Up Phase. The Discovery Extension Phase will include WGS data on 107 additional family members of the pedigrees studied in the Discovery Phase and additional WGS data on 213 family members from new pedigrees. The Follow-Up Phase will include more families undergoing whole-genome or exome sequencing, and individual investigators will share their sequence data with the ADSP. These additional sequence data will permit a replication study.

The RV-GDT method provides a robust and powerful way to use family-based sequence data to identify associations between complex disease and rare variants. Given its capability of adequately controlling for population admixture or substructure and its superior power over other methods for extended pedigrees and pedigrees with missing data, the RV-GDT is extremely beneficial in elucidating the involvement of rare variants in the etiology of complex traits. The RV-GDT is applicable to exome and genome sequence data and rare variants obtained from genotyping arrays. VAT is an all-in-one software pipeline package available for pre-processing data from various input formats, and the RV-GDT package is implemented to analyze rare-variant data exported by VAT. The RV-GDT is implemented in Python, and the software package and documentation are publicly available online.

Supplemental Data

Supplemental Data include Supplemental Acknowledgments, two figures, and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.12.001.

Acknowledgments

We wish to thank the family members who participated in the Alzheimer Disease Sequencing Project and made this research possible. This work was supported by National Human Genome Research Institute grant R01 HG008972. Complete acknowledgments can be found in the Supplemental Acknowledgments.

Received: October 4, 2016
Accepted: December 6, 2016
Published: January 5, 2017

Web Resources

ADSP, https://www.niagads.org/adsp/content/home
ANNOVAR, http://annovar.openbioinformatics.org/en/latest/
CADD, http://cadd.gs.washington.edu/
dbNSFP, https://sites.google.com/site/jpopgen/dbNSFP
dbSNP, http://www.ncbi.nlm.nih.gov/projects/SNP/
Epstein software, http://genetics.emory.edu/labs/epstein/software
ExAC Browser, http://exac.broadinstitute.org/
FBAT, http://www.biostat.harvard.edu/fbat/fbat.htm
FreeBayes, https://github.com/ekg/freebayes
GeneReviews, Bird, T.D. (1993). Alzheimer Disease Overview, https://www.ncbi.nlm.nih.gov/books/NBK1161/
Genome Analysis Toolkit (GATK), https://www.broadinstitute.org/gatk/
OMIM, http://www.omim.org/
PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/
RareIBD, http://jaehoonsullab.semel.ucla.edu/rareibd/
RV-GDT, https://statgen.research.bcm.edu/index.php/Main_Page#Statistical_Genetics_Software
Variant Association Tools (VAT), http://varianttools.sourceforge.net/Association/

References

1. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. *11*, 446–450.
2. Cirulli, E.T., and Goldstein, D.B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat. Rev. Genet. *11*, 415–425.

3. Ott, J., Kamatani, Y., and Lathrop, M. (2011). Family-based designs for genome-wide association studies. Nat. Rev. Genet. *12*, 465–474.
4. Li, M., Boehnke, M., and Abecasis, G.R. (2006). Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. Am. J. Hum. Genet. *78*, 778–792.
5. Mathieson, I., and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nat. Genet. *44*, 243–246.
6. Spielman, R.S., McGinnis, R.E., and Ewens, W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet. *52*, 506–516.
7. He, Z., O'Roak, B.J., Smith, J.D., Wang, G., Hooker, S., Santos-Cortez, R.L.P., Li, B., Kan, M., Krumm, N., Nickerson, D.A., et al. (2014). Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data. Am. J. Hum. Genet. *94*, 33–46.
8. Jiang, Y., Satten, G.A., Han, Y., Epstein, M.P., Heinzen, E.L., Goldstein, D.B., and Allen, A.S. (2014). Utilizing population controls in rare-variant case-parent association tests. Am. J. Hum. Genet. *94*, 845–853.
9. Laird, N.M., Horvath, S., and Xu, X. (2000). Implementing a unified approach to family-based tests of association. Genet. Epidemiol. *19* (*Suppl 1*), S36–S42.
10. De, G., Yip, W.-K., Ionita-Laza, I., and Laird, N. (2013). Rare variant analysis for family-based design. PLoS ONE *8*, e48495.
11. Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J.D., and Lin, X. (2013). Family-based association tests for sequence data, and comparisons with population-based association tests. Eur. J. Hum. Genet. *21*, 1158–1162.
12. Epstein, M.P., Duncan, R., Ware, E.B., Jhun, M.A., Bielak, L.F., Zhao, W., Smith, J.A., Peyser, P.A., Kardia, S.L.R., and Satten, G.A. (2015). A statistical approach for rare-variant association testing in affected sibships. Am. J. Hum. Genet. *96*, 543–554.
13. Sul, J.H., Cade, B.E., Cho, M.H., Qiao, D., Silverman, E.K., Redline, S., and Sunyaev, S. (2016). Increasing Generality and Power of Rare-Variant Tests by Utilizing Extended Pedigrees. Am. J. Hum. Genet. *99*, 846–859.
14. Chen, W.-M., Manichaikul, A., and Rich, S.S. (2009). A generalized family-based association test for dichotomous traits. Am. J. Hum. Genet. *85*, 364–376.
15. Martin, E.R., Monks, S.A., Warren, L.L., and Kaplan, N.L. (2000). A test for linkage and association in general pedigrees: the pedigree disequilibrium test. Am. J. Hum. Genet. *67*, 146–154.
16. Martin, E.R., Bass, M.P., Gilbert, J.R., Pericak-Vance, M.A., and Hauser, E.R. (2003). Genotype-based association test for general pedigrees: the genotype-PDT. Genet. Epidemiol. *25*, 203–213.
17. Van Cauwenberghe, C., Van Broeckhoven, C., and Sleegers, K. (2016). The genetic landscape of Alzheimer disease: clinical implications and perspectives. Genet. Med. *18*, 421–430.
18. Del-Aguila, J.L., Koboldt, D.C., Black, K., Chasse, R., Norton, J., Wilson, R.K., and Cruchaga, C. (2015). Alzheimer's disease: rare variants with large effect sizes. Curr. Opin. Genet. Dev. *33*, 49–55.
19. Grupe, A., Abraham, R., Li, Y., Rowland, C., Hollingworth, P., Morgan, A., Jehu, L., Segurado, R., Stone, D., Schadt, E., et al. (2007). Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. Hum. Mol. Genet. *16*, 865–873.
20. Azoitei, N., Brey, A., Busch, T., Fulda, S., Adler, G., and Seufferlein, T. (2007). Thirty-eight-negative kinase 1 (TNK1) facilitates TNFalpha-induced apoptosis by blocking NF-kappaB activation. Oncogene *26*, 6536–6545.
21. Felschow, D.M., Civin, C.I., and Hoehn, G.T. (2000). Characterization of the tyrosine kinase Tnk1 and its binding with phospholipase C-γ1. Biochem. Biophys. Res. Commun. *273*, 294–301.
22. Salahshor, S., and Woodgett, J.R. (2005). The links between axin and carcinogenesis. J. Clin. Pathol. *58*, 225–236.
23. Stoothoff, W.H., Bailey, C.D.C., Mi, K., Lin, S.-C., and Johnson, G.V.W. (2002). Axin negatively affects tau phosphorylation by glycogen synthase kinase 3beta. J. Neurochem. *83*, 904–913.
24. Spittaels, K., Van den Haute, C., Van Dorpe, J., Geerts, H., Mercken, M., Bruynseels, K., Lasrado, R., Vandezande, K., Laenen, I., Boon, T., et al. (2000). Glycogen synthase kinase-3beta phosphorylates protein tau and rescues the axonopathy in the central nervous system of human four-repeat tau transgenic mice. J. Biol. Chem. *275*, 41340–41349.
25. Kolarova, M., García-Sierra, F., Bartos, A., Ricny, J., and Ripova, D. (2012). Structure and pathology of tau protein in Alzheimer disease. Int. J. Alzheimers Dis. *2012*, 731526.
26. Gorlov, I.P., Gorlova, O.Y., Sunyaev, S.R., Spitz, M.R., and Amos, C.I. (2008). Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am. J. Hum. Genet. *82*, 100–112.
27. Auer, P.L., Wang, G., and Leal, S.M. (2013). Testing for rare variant associations in the presence of missing data. Genet. Epidemiol. *37*, 529–538.
28. Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. *5*, e1000384.
29. Lin, D.Y., and Tang, Z.Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. Am. J. Hum. Genet. *89*, 354–367.
30. Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. *86*, 832–838.
31. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet. *46*, 310–315.
32. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
33. Li, B., Wang, G.T., and Leal, S.M. (2015). Generation of sequence-based data for pedigree-segregating Mendelian or Complex traits. Bioinformatics *31*, 3706–3708.
34. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.
35. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.
36. Challis, D., Yu, J., Evani, U.S., Jackson, A.R., Paithankar, S., Coarfa, C., Milosavljevic, A., Gibbs, R.A., and Yu, F. (2012).

An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics *13*, 8.
37. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
38. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.
39. Wang, G.T., Peng, B., and Leal, S.M. (2014). Variant association tools for quality control and analysis of large-scale sequence and genotyping array data. Am. J. Hum. Genet. *94*, 770–783.
40. Liu, X., Jian, X., and Boerwinkle, E. (2013). dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum. Mutat. *34*, E2393–E2402.
41. Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., Sidow, A.; and NISC Comparative Sequencing Program (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. *15*, 901–913.
42. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110–121.
43. Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L.A., Edwards, K.J., Day, I.N.M., and Gaunt, T.R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum. Mutat. *34*, 57–65.
44. Schwarz, J.M., Cooper, D.N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. Nat. Methods *11*, 361–362.
45. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.
46. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., and Chan, A.P. (2012). Predicting the functional effect of amino acid substitutions and indels. PLoS ONE *7*, e46688.
47. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. *4*, 1073–1081.
48. Liu, J., Lewinger, J.P., Gilliland, F.D., Gauderman, W.J., and Conti, D.V. (2013). Confounding and heterogeneity in genetic association studies with admixed populations. Am. J. Epidemiol. *177*, 351–360.
49. Atzmon, G., Hao, L., Pe'er, I., Velez, C., Pearlman, A., Palamara, P.F., Morrow, B., Friedman, E., Oddoux, C., Burns, E., and Ostrer, H. (2010). Abraham's children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern Ancestry. Am. J. Hum. Genet. *86*, 850–859.
50. Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. Annu. Rev. Genomics Hum. Genet. *10*, 387–406.
51. Bush, W.S., and Moore, J.H. (2012). Chapter 11: Genome-wide association studies. PLoS Comput. Biol. *8*, e1002822.
52. Pirinen, M., Donnelly, P., and Spencer, C.C.A. (2012). Including known covariates can reduce power to detect genetic effects in case-control studies. Nat. Genet. *44*, 848–851.
53. He, X., Sanders, S.J., Liu, L., De Rubeis, S., Lim, E.T., Sutcliffe, J.S., Schellenberg, G.D., Gibbs, R.A., Daly, M.J., Buxbaum, J.D., et al. (2013). Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. PLoS Genet. *9*, e1003671.
54. Liu, D.J., and Leal, S.M. (2012). Estimating genetic effects and quantifying missing heritability explained by identified rare-variant associations. Am. J. Hum. Genet. *91*, 585–596.

The American Journal of Human Genetics, Volume *100*

Supplemental Data

The Rare-Variant Generalized Disequilibrium Test for Association Analysis of Nuclear and Extended Pedigrees with Application to Alzheimer Disease WGS Data

Zongxiao He, Di Zhang, Alan E. Renton, Biao Li, Linhai Zhao, Gao T. Wang, Alison M. Goate, Richard Mayeux, and Suzanne M. Leal

Supplemental Acknowledgements

The Alzheimer's Disease Sequencing Project (ADSP) is comprised of two Alzheimer's Disease (AD) genetics consortia and three National Human Genome Research Institute (NHGRI) funded Large Scale Sequencing and Analysis Centers (LSAC). The two AD genetics consortia are the Alzheimer's Disease Genetics Consortium (ADGC) funded by NIA (U01 AG032984), and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) funded by NIA (R01 AG033193), the National Heart, Lung, and Blood Institute (NHLBI), other National Institute of Health (NIH) institutes and other foreign governmental and non-governmental organizations. The Discovery Phase analysis of sequence data is supported through UF1AG047133 (to Drs. Schellenberg, Farrer, Pericak-Vance, Mayeux, and Haines); RF1AG015473 to Dr. Mayeux; U01AG049505 to Dr. Seshadri; U01AG049506 to Dr. Boerwinkle; U01AG049507 to Dr. Wijsman; and U01AG049508 to Dr. Goate.

The ADGC cohorts include: Adult Changes in Thought (ACT), the Alzheimer's Disease Centers (ADC), the Chicago Health and Aging Project (CHAP), the Memory and Aging Project (MAP), Mayo Clinic (MAYO), Mayo Parkinson's Disease controls, University of Miami, the Multi-Institutional Research in Alzheimer's Genetic Epidemiology Study (MIRAGE), the National Cell Repository for Alzheimer's Disease (NCRAD), the National Institute on Aging Late Onset Alzheimer's Disease Family Study (NIA-LOAD), the Religious Orders Study (ROS), the Texas Alzheimer's Research and Care Consortium (TARC), Vanderbilt University/Case Western Reserve University (VAN/CWRU), the Washington Heights-Inwood Columbia Aging Project (WHICAP) and the Washington University Sequencing Project (WUSP), the Columbia University Hispanic- Estudio Familiar de Influencia Genetica de Alzheimer (EFIGA), the University of Toronto (UT), and Genetic Differences (GD**)**.

The CHARGE cohorts, with funding provided by 5RC2HL102419 and HL105756, include the following: Atherosclerosis Risk in Communities (ARIC) Study which is carried out as a collaborative study supported by NHLBI contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C), Austrian Stroke Prevention Study (ASPS), Cardiovascular Health Study (CHS), Erasmus Rucphen Family Study (ERF), Framingham Heart Study (FHS), and Rotterdam Study (RS).

The three LSACs are: the Human Genome Sequencing Center at the Baylor College of Medicine (U54 HG003273), the Broad Institute Genome Center (U54HG003067), and the Washington University Genome Institute (U54HG003079).

Biological samples and associated phenotypic data used in primary data analyses were stored at Study Investigators institutions, and at the National Cell Repository for Alzheimer's Disease (NCRAD, U24AG021886) at Indiana University funded by NIA. Associated Phenotypic Data used

in primary and secondary data analyses were provided by Study Investigators, the NIA funded Alzheimer's Disease Centers (ADCs), and the National Alzheimer's Coordinating Center (NACC, U01AG016976) and the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS, U24AG041689) at the University of Pennsylvania, funded by NIA, and at the Database for Genotypes and Phenotypes (dbGaP) funded by NIH. Contributors to the Genetic Analysis Data included Study Investigators on projects that were individually funded by NIA, and other NIH institutes, and by private U.S. organizations, or foreign governmental or nongovernmental organizations.

Supplemental Figures and Tables

Figure S1. QQ plot of the $-\log_{10}$ p-values for data simulated under the null hypothesis of no association.

P-values for the RV-GDT were obtained empirically using 100,000 permutations. Confidence intervals are highlighted in gray. Four different types of family data were investigated: 1,000 discordant sib-pairs (Figure 1 pedigree structure A), 1,000 affected sib-pairs (Figure 1 pedigree structure B), 1,000 extended pedigrees (Figure 1 pedigree structure C), and mixed family types including 500 discordant sib-pairs, 250 affected sib-pairs and 250 extended pedigrees. Panel A: genotypes simulated using ExAC Non-Finnish European variant information; Panel B: genotypes simulated using ExAC Non-Finnish European variant information with ~50% of the founders missing their genotype data. Panel C: genotypes were simulated for an admixed population (20% Non-Finnish Europeans and 80% African/African American); Panel D: genotypes were simulated for a population with substructure (50% Non-Finnish European families and 50% African/African American families).

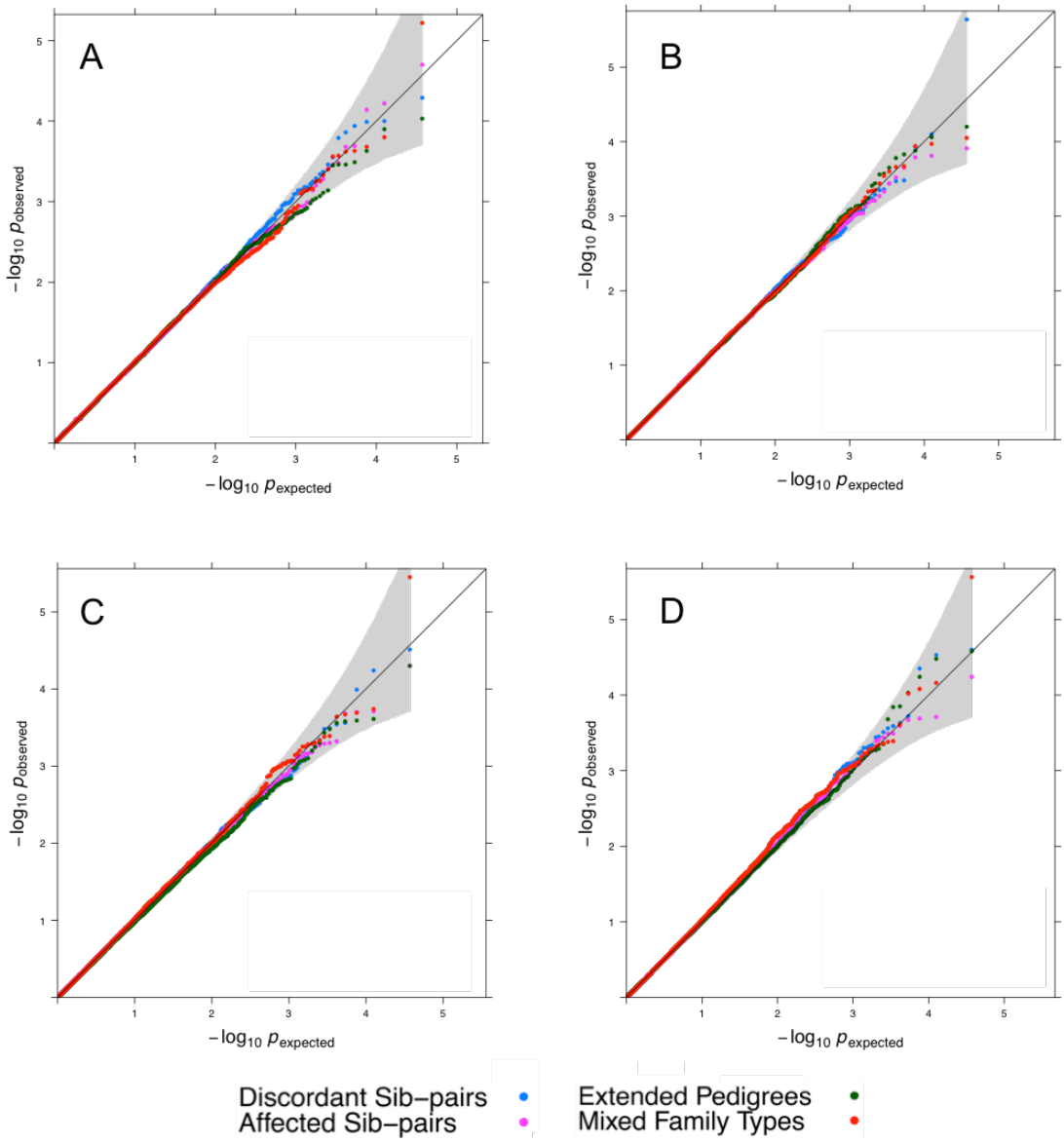
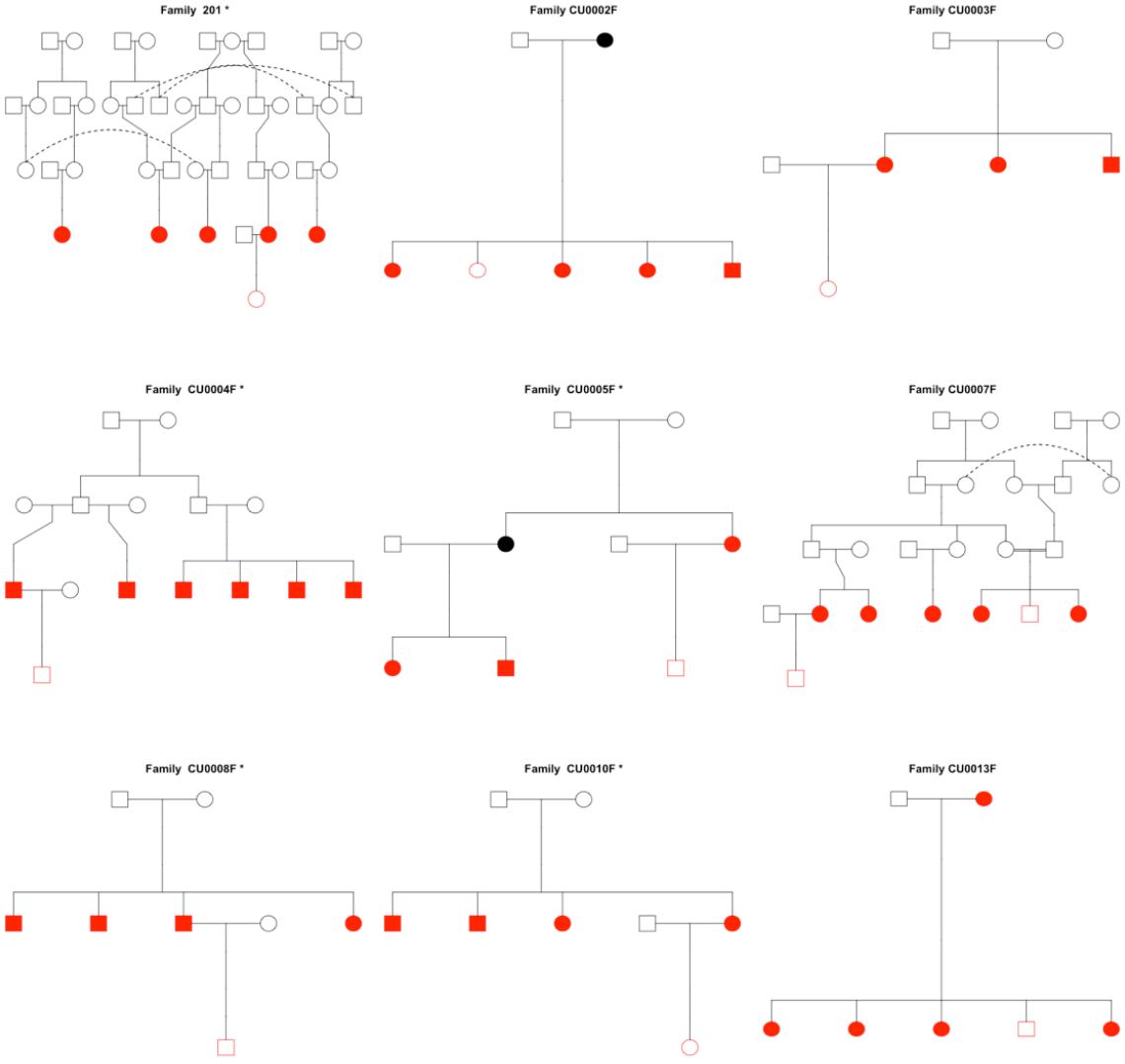
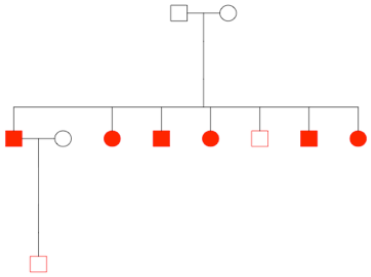

A
B
C
D
$-\log_{10} p_{observed}$
$-\log_{10} p_{expected}$
Discordant Sib−pairs
Affected Sib−pairs
Extended Pedigrees
Mixed Family Types

Figure S2. Eighty-one Alzheimer's disease pedigrees included in the analysis from Alzheimer's Disease Sequencing Project.

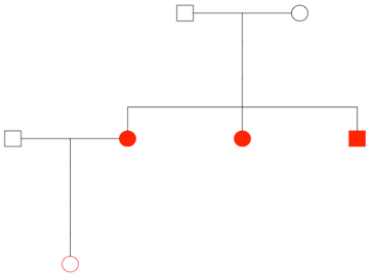
The dataset includes 338 individuals with Alzheimer's disease of which 316 have whole genome sequence data and 494 unaffected individuals of which 98 have whole genome sequence data. Filled squares and circles represent males and females with Alzheimer's disease, respectively, while family members with open squares and circles represent unaffected individuals. Individuals represented in red have whole genome sequence data available, while those in black do not have genotype data. There are 25 families that cannot be analyzed by the FBAT software, and these families are marked with asterisks (*).



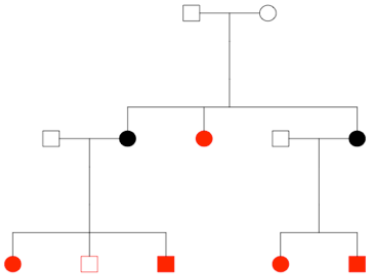
Family CU0014F



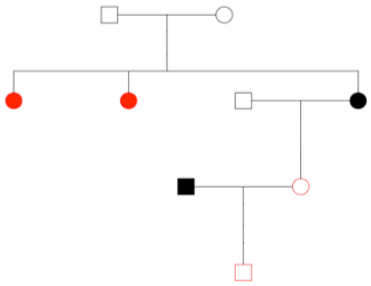
Family CU0015F



Family CU0016F

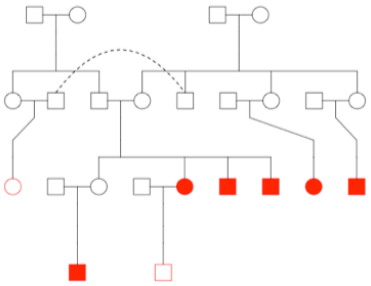


Family CU0017F *

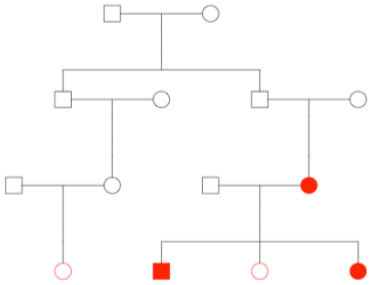


Family CU0021F

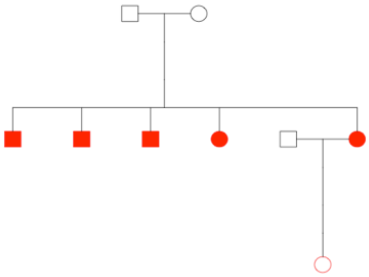
Family CU0023F



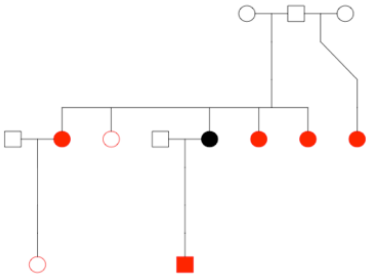
Family CU0026F



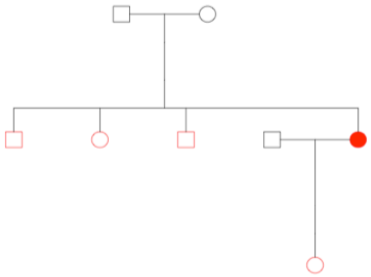
Family CU0029F *



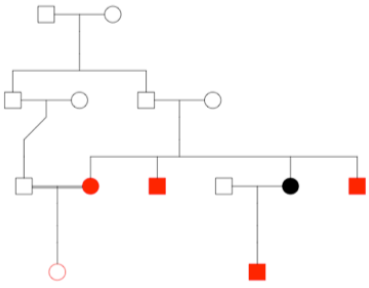
Family CU0030F



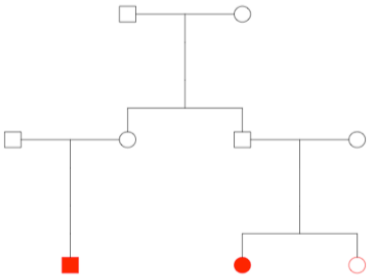
Family CU0031F



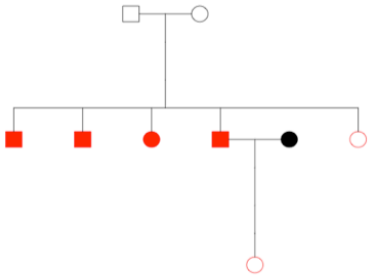
Family CU0032F



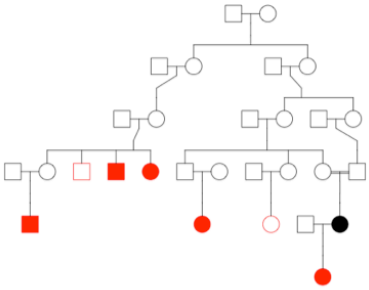
Family CU0033F



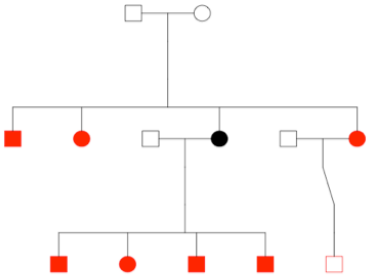
Family CU0035F



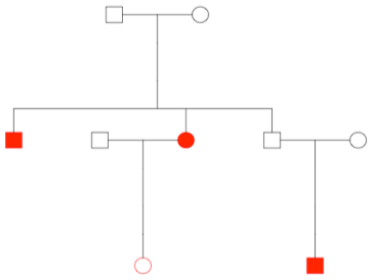
Family CU0037F



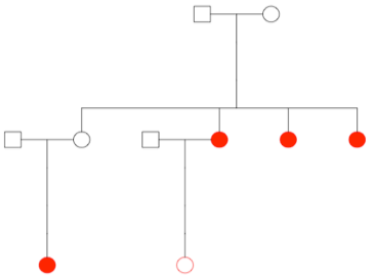
Family CU0039F *



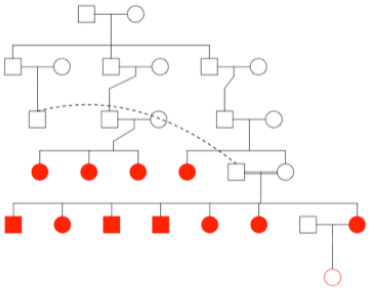
Family CU0042F *



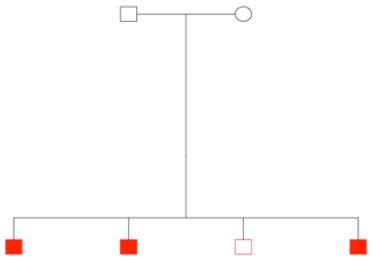
Family CU0043F



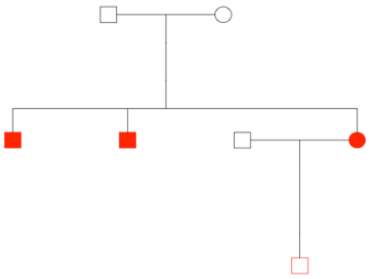
Family CU0044F



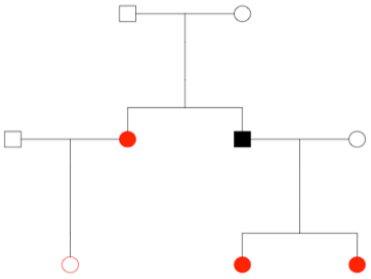
Family CU0045F



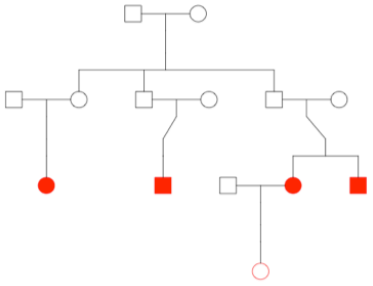
Family CU0046F



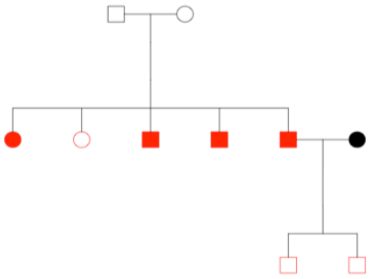
Family CU0047F *



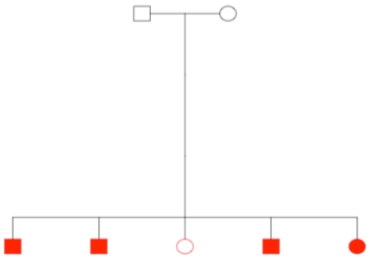
Family CU0048F *



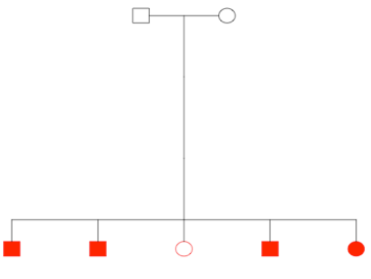
Family CU0049F



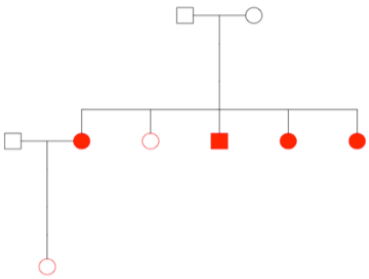
Family CU0050F



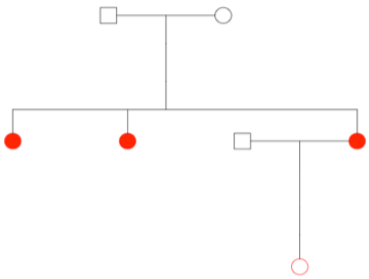
Family CU0052F



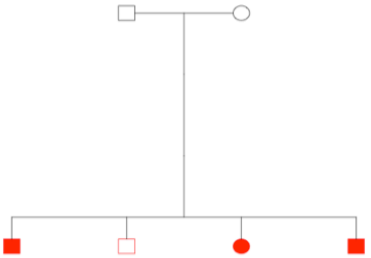
Family CU0053F



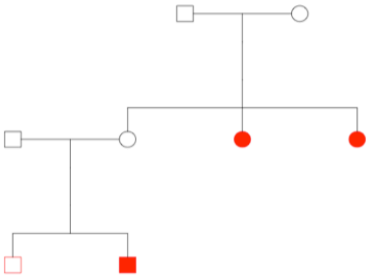
Family CU0057F



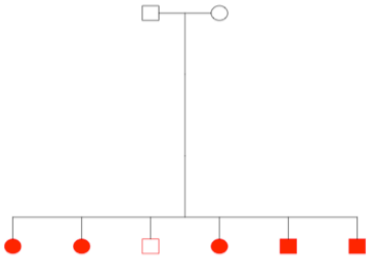
Family CU0058F



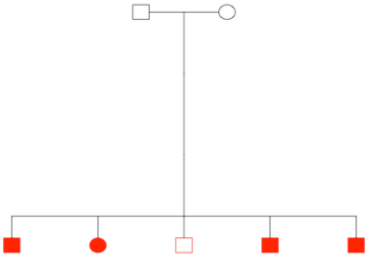
Family CU0060F



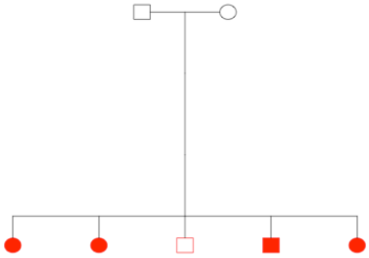
Family CU0064F



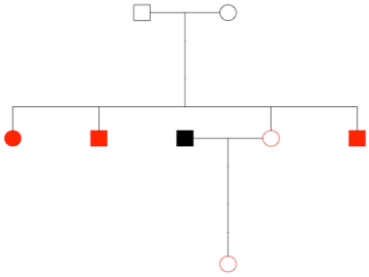
Family CU0065F



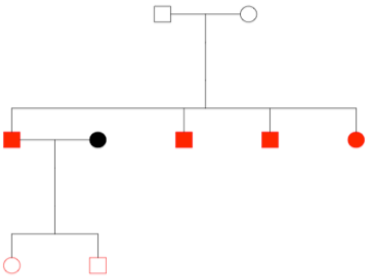
Family CU0067F



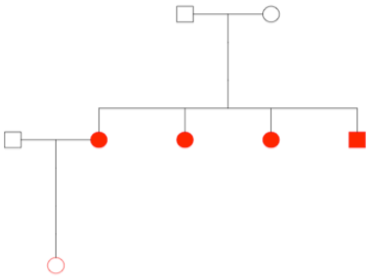
Family CU0068F



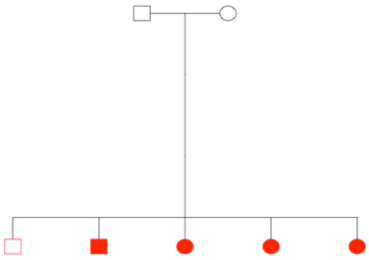
Family CU0071F *

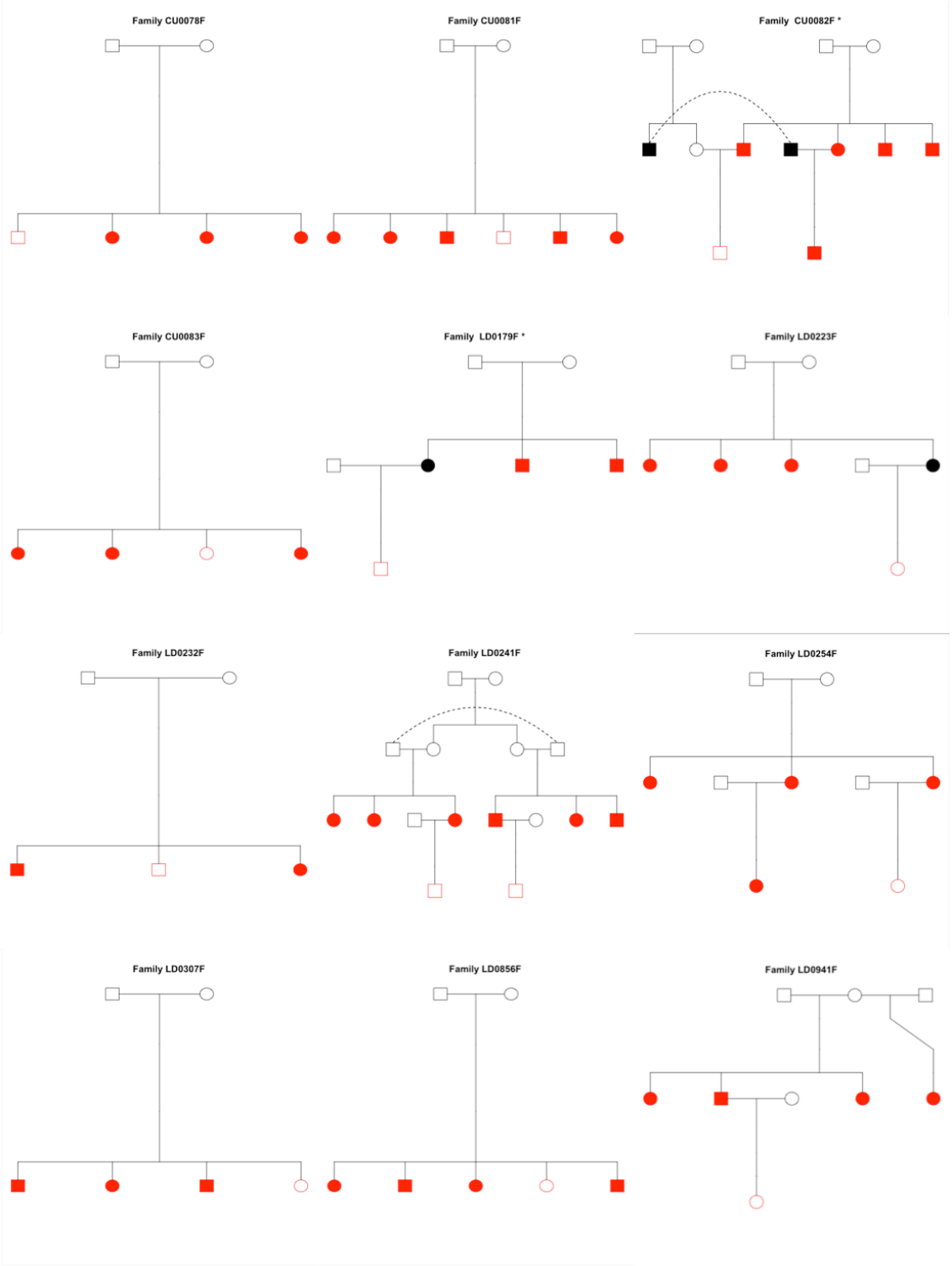


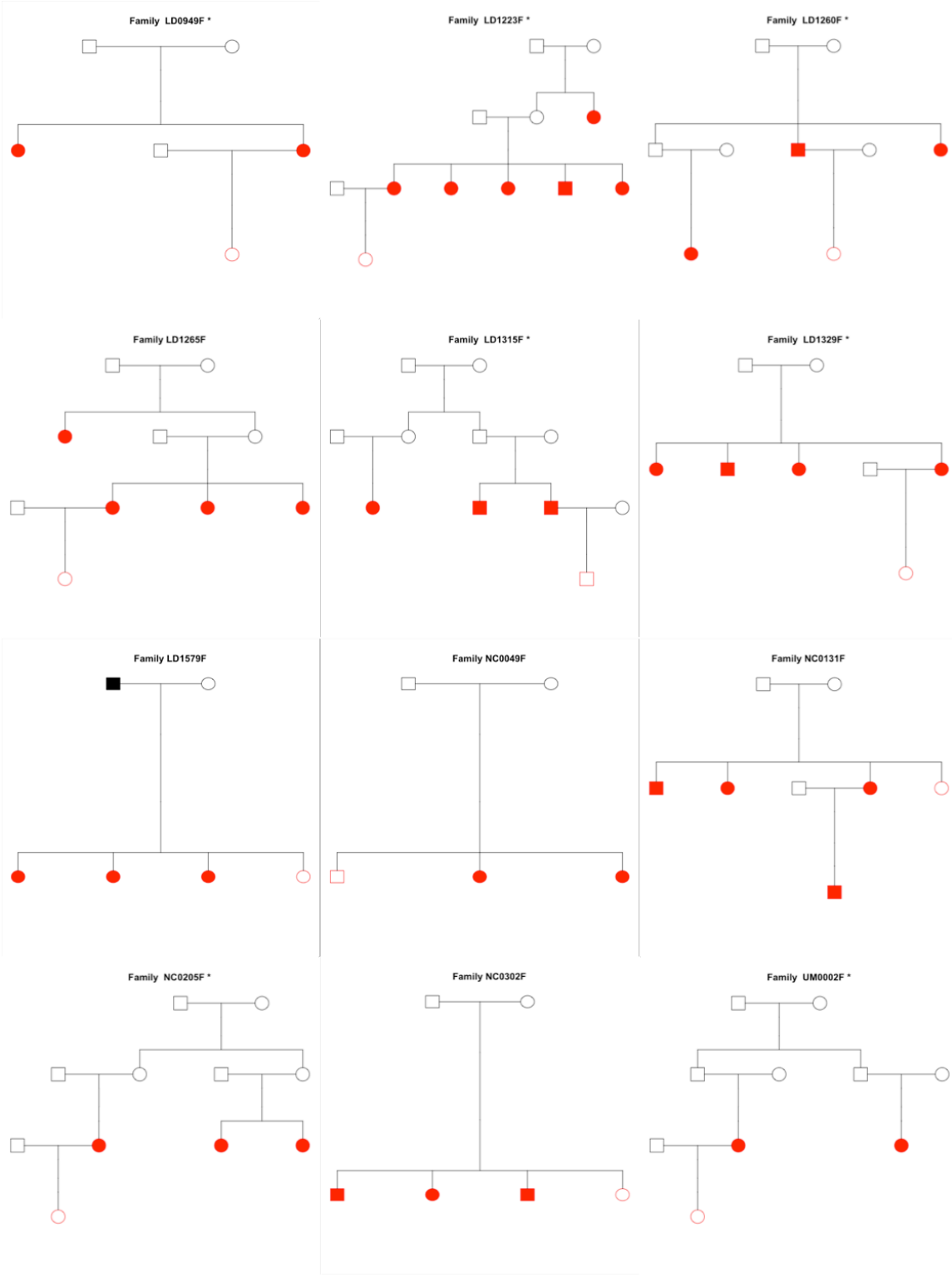
Family CU0073F *



Family CU0076F




Family CU0078F
Family CU0081F
Family CU0082F *
Family CU0083F
Family LD0179F *
Family LD0223F
Family LD0232F
Family LD0241F
Family LD0254F
Family LD0307F
Family LD0856F
Family LD0941F


Family LD0949F *
Family LD1223F *
Family LD1260F *
Family LD1265F
Family LD1315F *
Family LD1329F *
Family LD1579F
Family NC0049F
Family NC0131F
Family NC0205F *
Family NC0302F
Family UM0002F *

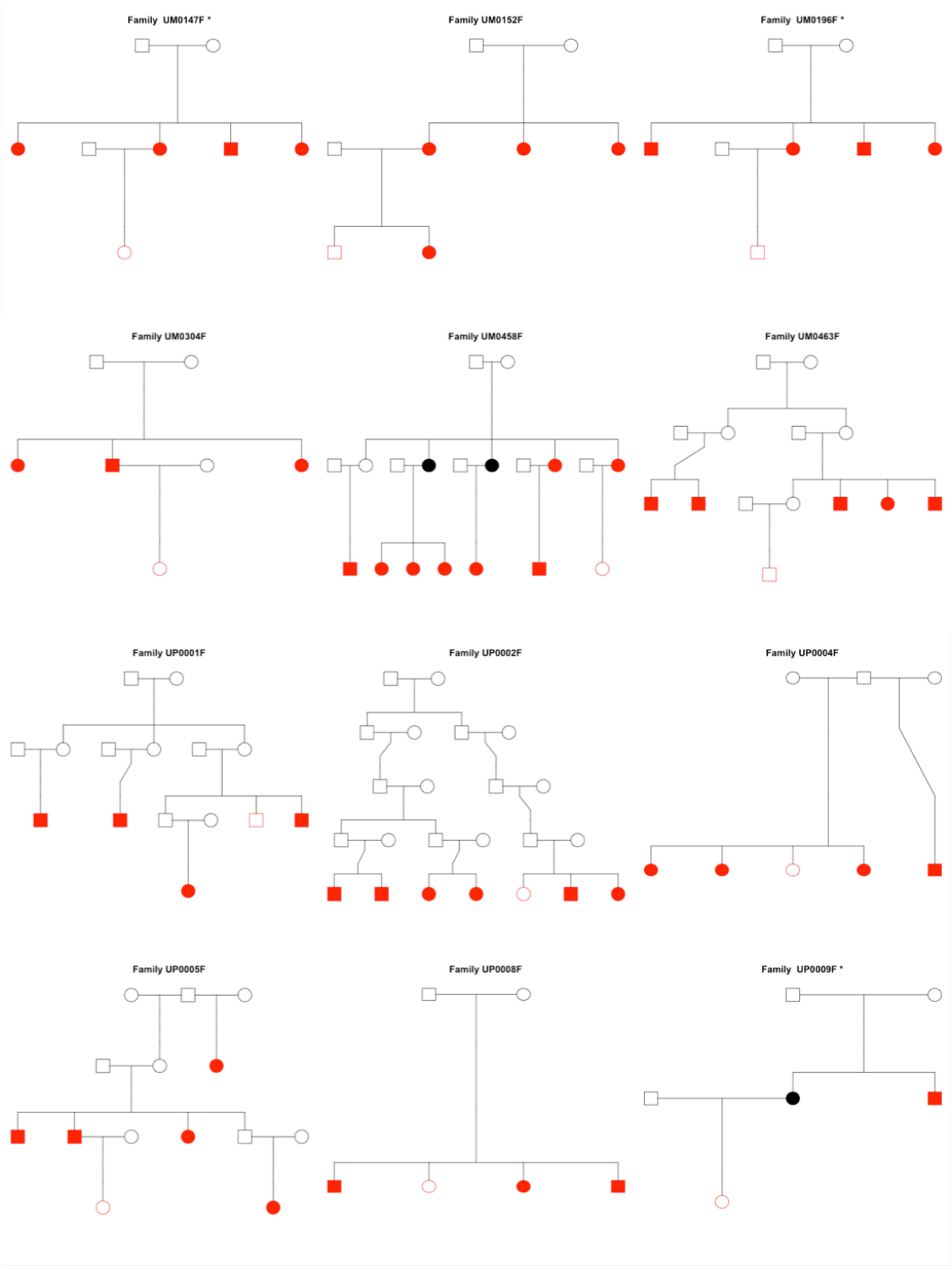

Family UM0147F *
Family UM0152F
Family UM0196F *
Family UM0304F
Family UM0458F
Family UM0463F
Family UP0001F
Family UP0002F
Family UP0004F
Family UP0005F
Family UP0008F
Family UP0009F *

Table S1. Type I error rate for RareIBD at α levels of 0.05

	Discordant Sib-pair	Affected Sib-pair	Extended Pedigree	Mixed Family Types
RareIBD	0.064	0.033	0.465	0.094
	Each Founder has a probability of 50% to be missing all of their genotype data			
RareIBD	0.064	0.046	0.296	0.091
	80% African and 20% European population admixture			
RareIBD	0.059	0.028	0.395	0.086
	50% African and 50% European families			
RareIBD	0.063	0.030	0.476	0.089

One-thousand families for each pedigree structure shown in Figure 1 and mixed pedigree structures were simulated. Genotype data were generated for all autosomal genes across the genome with an OR = 1.0, and type I error rate was defined as the proportion of genes with a p-value less than 0.05. Variant information for 17,987 autosomal genes from ExAC Non-Finnish European were used to generate family data when each founder has a probability (0% or 50%) to be missing their genotype data. Variant information for 17,873 autosomal genes that are present in both ExAC Non-Finnish European and African/African American populations was used to generate family data with population admixture and substructure.

Table S2. The Ethnicities of Alzheimer's disease families included in the analysis

Ethnicity	Number of Families	Family IDs
Dominican	46	CU0002F, CU0003F, CU0004F, CU0005F, CU0007F, CU0008F, CU0010F, CU0013F, CU0014F, CU0015F, CU0016F, CU0017F, CU0021F, CU0023F, CU0026F, CU0029F, CU0030F, CU0031F, CU0033F, CU0035F, CU0037F, CU0039F, CU0043F, CU0044F, CU0045F, CU0046F, CU0047F, CU0048F, CU0049F, CU0050F, CU0052F, CU0053F, CU0057F, CU0058F, CU0060F, CU0064F, CU0065F, CU0067F, CU0068F, CU0071F, CU0073F, CU0076F, CU0078F, CU0081F, CU0082F, CU0083F
European Descent	31	LD0179F, LD0223F, LD0232F, LD0241F, LD0254F, LD0307F, LD0856F, LD0949F, LD1223F, LD1260F, LD1265F, LD1315F, LD1329F, LD1579F, NC0049F, NC0131F, NC0205F, NC0302F, UM0002F, UM0147F, UM0152F, UM0196F, UM0304F, UM0458F, UM0463F, UP0001F, UP0002F, UP0004F, UP0005F, UP0008F, UP0009F
Puerto Rican	2	CU0032F, CU0042F
African American	1	LD0941F
Dutch Isolate	1	201

Table S3. Bioinformatic evaluation and frequencies of analyzed rare variants within *MARCH10*

dbSNP rsID	rs146326363*	rs116835087*^	rs147046907*	rs60472825	rs138015683	rs374880698*^	rs141415486*	rs78457484
hg19 Position	17:60782924	17:60813470	17:60813550	17:60813944	17:60813982	17:60814417	17:60837208	17:60837337
Reference allele	G	C	G	G	T	A	C	G
Alternate allele	C	T	A	A	C	C	T	A
cDNA change	c.2347C>G	c.1759G>A	c.1679C>T	c.1285C>T	c.1247A>G	c.812T>G	c.370G>A	c.241C>T
Amino acid change	p.Gln783Glu	p.Gly587Ser	p.Thr560Ile	p.His429Tyr	p.Asn416Ser	p.Phe271Cys	p.Glu124Lys	p.Pro81Ser
ExAC all MAF	0.0005	0.0037	0.0068	0.0056	0.0010	1.63E-05	0.0011	0.0072
Number of alternative alleles - AD pedigree members (n=316)	2	4	10	19	1	3	1	13
Number of alternative alleles - unaffected pedigree members (n=98)	1	0	0	4	1	0	0	0
GERP score	4.24	4.4	3.32	3	-0.978	2.61	4.79	0.166
PhyloP score	2.254	1.601	1.565	0.095	-0.474	1.492	2.695	0.711
CADD score, scaled	19.6	22.4	6.022	12.46	4.402	22.4	15.74	9.87
FATHMM	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated
MutationTaster	disease-causing	disease-causing	polymorphism	polymorphism	polymorphism	disease-causing	disease-causing	polymorphism automatic
Polyphen-2 HVAR	benign	possibly damaging	benign	benign	Benign	probably damaging	benign	possibly damaging
PROVEAN	neutral	neutral	neutral	neutral	neutral	deleterious	neutral	neutral
SIFT	tolerated	damaging	tolerated	damaging	tolerated	damaging	tolerated	damaging

Abbreviations are as follows: ExAC, Exome Aggregation Consortium; MAF, minor allele frequency; CADD, Combined Annotation Dependent Depletion; FATHMM, Functional Analysis through Hidden Markov Models; PROVEAN, Protein Variation Effect Analyzer; SIFT, Sorting Intolerant From Tolerant. Conservation scores and bioinformatics results as compiled by dbNSFP v.2.9.

*Variant is deemed as conserved nucleotide (both GERP and PhyloP scores > 1).

^Variant is deemed damaging by at least three of six bioinformatics tools (variant with CADD scaled score >15 is deemed to be deleterious).

Table S4. Bioinformatic Evaluation and Frequencies of Rare Missense Variants within *AMBN*

dbSNP rsID	rs143940501*^	rs146167261*^	rs115723025	rs113506649	rs139319140*^	rs150017698*^	NA	rs76503327
hg19 Position	4:71469167	4:71471905	4:71471958	4:71471985	4:71472001	4:71472053	4:71472146	4:71472235
Reference allele	C	G	G	C	G	A	C	G
Alternate allele	T	A	A	A	A	C	T	A
cDNA change	c.743C>T	c.802G>A	c.855G>A	c.882C>A	c.898G>A	c.950A>C	c.1043C>T	c.1132G>A
Amino acid change	p.Ala248Val	p.Gly268Arg	p.Met285Ile	p.His294Gln	p.Gly300Ser	p.Glu317Ala	p.Ala348Val	p.Val378Ile
ExAC all MAF	0.0007	0.0026	0.0103	0.0053	0.0004	0.0004	NA	0.0068
Number of alternative alleles - AD pedigree members (n=316)	2	4	14	4	1	0	3	13
Number of alternative alleles - unaffected pedigree members (n=98)	0	0	1	0	0	1	0	2
GERP score	5.07	5.79	2.73	-1.87	4.95	4.66	1.98	4.07
PhyloP score	2.394	3.499	0.453	-1.137	2.596	3.297	0.204	0.908
CADD score, scaled	24.6	26.3	10.97	0.023	26.1	19.31	0.002	10.97
FATHMM	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated
MutationTaster	disease-causing	disease-causing	polymorphism	polymorphism	disease-causing	disease-causing	polymorphism	polymorphism
Polyphen-2 HVAR	possibly damaging	probably damaging	benign	benign	probably damaging	probably damaging	benign	benign
PROVEAN	neutral	deleterious	neutral	neutral	deleterious	deleterious	neutral	neutral
SIFT	damaging	damaging	damaging	tolerated	damaging	tolerated	tolerated	tolerated

Abbreviations are as follows: NA, Not Available; ExAC, Exome Aggregation Consortium; MAF, minor allele frequency; CADD, Combined Annotation Dependent Depletion; FATHMM, Functional Analysis through Hidden Markov Models; PROVEAN, Protein Variation Effect Analyzer; SIFT, Sorting Intolerant From Tolerant. Conservation scores and bioinformatics results as compiled by dbNSFP v.2.9.

*Variant is deemed as conserved nucleotide (both GERP and PhyloP scores > 1).

^Variant is deemed damaging by at least three of six bioinformatics tools (variant with CADD scaled score >15 is deemed to be deleterious).

Table S5. Bioinformatic Evaluation and Frequencies of Rare Missense Variants within *TCOF1*

dbSNP rsID	rs56180593^	rs142477153*^	rs181203524	rs144327167*^	rs189476787^	rs75181211^	rs114326915	rs201458471	rs75583421^
hg19 Position	5:149740732	5:149747428	5:149748403	5:149753894	5:149754226	5:149754229	5:149754325	5:149754948	5:149755362
Reference allele	C	A	C	G	C	C	C	T	G
Alternate allele	T	G	T	A	T	T	T	C	A
cDNA change	c.122C>T	c.326A>G	c.503C>T	c.797G>A	c.899C>T	c.902C>T	c.998C>T	c.1304T>C	c.1552G>A
Amino acid change	p.Ala41Val	p.Asn109Ser	p.Thr168Met	p.Ser266Asn	p.Pro300Leu	p.Ala301Val	p.Ser333Leu	p.Met435Thr	p.Val518Ile
ExAC all MAF	0.0023	0.0001	0.0004	0.0034	0.0002	0.0035	0.0038	0.0001	0.0077
Number of alternative alleles - AD pedigree members (n=316)	4	1	7	2	3	13	2	3	6
Number of alternative alleles - unaffected pedigree members (n=98)	0	0	0	0	0	2	1	1	1
GERP score	2.98	4.33	-1.6	2.98	1.5	2.63	-0.0656	1.02	1.02
PhyloP score	0.602	2.068	-0.037	4.082	0.661	0.386	1.055	-0.049	0.48
CADD score, scaled	22.3	23.2	7.967	22.9	12.91	16.34	6.996	0.009	18.7
FATHMM	damaging	damaging	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated
MutationTaster	polymorphism	polymorphism	polymorphism	polymorphism	polymorphism	polymorphism	polymorphism	polymorphism	polymorphism
Polyphen-2 HVAR	benign	probably damaging	probably damaging	probably damaging	probably damaging	benign	benign	benign	possibly damaging
PROVEAN	neutral	deleterious	neutral	neutral	deleterious	deleterious	deleterious	neutral	neutral
SIFT	damaging	damaging	damaging	damaging	damaging	damaging	damaging	tolerated	damaging

Abbreviations are as follows: NA, Not Available; ExAC, Exome Aggregation Consortium; MAF, minor allele frequency; CADD, Combined Annotation Dependent Depletion; FATHMM, Functional Analysis through Hidden Markov Models; PROVEAN, Protein Variation Effect Analyzer; SIFT, Sorting Intolerant From Tolerant. Conservation scores and bioinformatics results as compiled by dbNSFP v.2.9.

*Variant is deemed as conserved nucleotide (both GERP and PhyloP scores > 1).

^Variant is deemed damaging by at least three of six bioinformatics tools (variant with CADD scaled score >15 is deemed to be deleterious).

Table S5. Bioinformatic Evaluation and Frequencies of Rare Missense Variants within *TCOF1* (continued)

rs201843484*^	rs73270846^	rs149384321	rs376431510	rs150515843^	rs114689020	NA^	rs200424706^	rs144793475*	rs55980697^
5:149755712	5:149756088	5:149756091	5:149759108	5:149759198	5:149759201	5:149759213	5:149771660	5:149772288	5:149776067
C	C	A	G	C	C	C	T	A	A
T	T	G	A	T	T	G	G	T	G
c.1730C>T	c.2014C>T	c.2017A>G	c.2441G>A	c.2531C>T	c.2534C>T	c.2546C>G	c.3207T>G	c.3304A>T	c.3773A>G
p.Ala577Val	p.Pro672Ser	p.Thr673Ala	p.Gly814Glu	p.Pro844Leu	p.Ser845Leu	p.Ala849Gly	p.Ser1069Arg	p.Thr1102Ser	p.Lys1258Arg
0.0003	0.0040	0.0001	4.08E-05	0.0015	0.0088	NA	0.0018	0.0001	0.0104
1	2	0	18	2	1	3	2	1	2
0	0	1	3	0	0	1	0	0	0
3.92	4.38	-10.7	1.61	2.82	0.389	0.811	-8.13	3.19	4.98
2.275	0.147	-4.381	0.338	0.68	-0.568	-0.368	-1.384	1.026	0.612
22.5	22.8	0.001	7.651	23.2	9.89	22	20.7	13.47	22.2
tolerated	tolerated	tolerated	tolerated	damaging	tolerated	tolerated	tolerated	tolerated	tolerated
polymorphism	polymorphism	polymorphism	polymorphism	polymorphism	polymorphism	polymorphism	polymorphism	polymorphism	disease-causing
possibly damaging	possibly damaging	benign	benign	benign	benign	probably damaging	probably damaging	benign	probably damaging
deleterious	deleterious	neutral	deleterious	deleterious	neutral	deleterious	deleterious	neutral	neutral
damaging	tolerated	tolerated	tolerated	damaging	damaging	damaging	damaging	damaging	tolerated

Table S6. Bioinformatic Evaluation and Frequencies of Rare Missense Variants within *AXIN1*

dbSNP rsID	rs34015754*	rs117208012*	rs367788267*	rs141148118	rs149849071	rs146947903*	rs116350678*	rs140151215*
hg19 Position	16:338189	16:347063	16:347143	16:347930	16:347957	16:348021	16:348233	16:396221
Reference allele	C	C	G	C	C	G	C	G
Alternate allele	T	T	A	T	T	C	T	C
cDNA change	c.2522G>A	c.1948G>A	c.1868C>T	c.1576G>A	c.1549G>A	c.1485C>G	c.1273G>A	c.805C>G
Amino acid change	p.Arg841Gln	p.Gly650Ser	p.Ser623Leu	p.Ala526Thr	p.Val517Ile	p.Asp495Glu	p.Gly425Ser	p.Gln269Glu
ExAC all MAF	0.0085	0.0168	0.0004	0.0007	0.0016	0.0098	0.0075	0.0009
Number of alternative alleles - AD pedigree members (n=316)	2	3	1	3	1	2	19	3
Number of alternative alleles - unaffected pedigree members (n=98)	0	0	0	0	0	0	1	0
GERP score	4.43	2.63	4.98	-5.51	-10.1	3.79	3.31	5.1
PhyloP score	4.824	1.19	2.46	-1.489	-2.359	1.208	2.732	7.611
CADD score, scaled	13.68	8.312	16.57	0.018	0.001	10.82	13.84	23.6
FATHMM	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated
MutationTaster	disease-causing	polymorphism	polymorphism	polymorphism	polymorphism	polymorphism	polymorphism	disease-causing
Polyphen-2 HVAR	possibly damaging	benign	benign	benign	benign	benign	benign	benign
PROVEAN	neutral	neutral	neutral	neutral	neutral	neutral	neutral	neutral
SIFT	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated

Abbreviations are as follows: ExAC, Exome Aggregation Consortium; MAF, minor allele frequency; CADD, Combined Annotation Dependent Depletion; FATHMM, Functional Analysis through Hidden Markov Models; PROVEAN, Protein Variation Effect Analyzer; SIFT, Sorting Intolerant From Tolerant. Conservation scores and bioinformatics results as compiled by dbNSFP v.2.9.

*Variant is deemed as conserved nucleotide (both GERP and PhyloP scores > 1).

Table S7. Bioinformatic Evaluation and Frequencies of Rare Missense Variants within *TNK1*

dbSNP rsID	rs201180891*^	rs61730812*^	NA*^	rs56093628^	rs80015268*	rs141588799
hg19 Position	17:7286889	17:7287832	17:7287859	17:7291869	17:7291943	17:7292117
Reference allele	T	C	T	C	G	G
Alternate allele	G	A	A	G	C	A
cDNA change	c.380T>G	c.896C>A	c.923T>A	c.1637C>G	c.1711G>C	c.1802G>A
Amino acid change	p.Phe127Cys	p.Ala299Asp	p.Met308Lys	p.Ser546Cys	p.Gly571Arg	p.Ser601Asn
ExAC all MAF	3.42E-05	0.0114	NA	0.0022	0.0027	0.0005
Number of alternative alleles - AD pedigree members (n=316)	4	8	2	4	1	2
Number of alternative alleles - unaffected pedigree members (n=98)	0	0	0	0	0	0
GERP score	1.99	4.03	5.3	4.18	3.27	1.89
PhyloP score	5.177	2.355	6.029	0.375	3.52	0.736
CADD score, scaled	26.2	22.1	22.1	20.4	9.985	5.874
FATHMM	damaging	damaging	tolerated	tolerated	tolerated	tolerated
MutationTaster	disease-causing	disease-causing	disease-causing	polymorphism	polymorphism	polymorphism
Polyphen-2 HVAR	probably damaging	possibly damaging	probably damaging	possibly damaging	benign	benign
PROVEAN	deleterious	neutral	deleterious	neutral	neutral	neutral
SIFT	damaging	tolerated	damaging	damaging	damaging	damaging

Abbreviations are as follows: NA, Not Available; ExAC, Exome Aggregation Consortium; MAF, minor allele frequency; CADD, Combined Annotation Dependent Depletion; FATHMM, Functional Analysis through Hidden Markov Models; PROVEAN, Protein Variation Effect Analyzer; SIFT, Sorting Intolerant From Tolerant. Conservation scores and bioinformatics results as compiled by dbNSFP v.2.9.

*Variant is deemed as conserved nucleotide (both GERP and PhyloP scores > 1).

^Variant is deemed damaging by at least three of six bioinformatics tools (variant with CADD scaled score >15 is deemed to be deleterious).