ARTICLE

A Rare Variant Nonparametric Linkage Method for Nuclear and Extended Pedigrees with Application to Late-Onset Alzheimer Disease via WGS Data

Linhai Zhao,[1] Zongxiao He,[1] Di Zhang,[1] Gao T. Wang,[2] Alan E. Renton,[3] Badri N. Vardarajan,[4] Michael Nothnagel,[5,6] Alison M. Goate,[3,7] Richard Mayeux,[4] and Suzanne M. Leal[1,4,8,*]

To analyze family-based whole-genome sequence (WGS) data for complex traits, we developed a rare variant (RV) non-parametric linkage (NPL) analysis method, which has advantages over association methods. The RV-NPL differs from the NPL in that RVs are analyzed, and allele sharing among affected relative-pairs is estimated only for minor alleles. Analyzing families can increase power because causal variants with familial aggregation usually have larger effect sizes than those underlying sporadic diseases. Differing from association analysis, for NPL only affected individuals are analyzed, which can increase power, since unaffected family members can be susceptibility variant carriers. RV-NPL is robust to population substructure and admixture, inclusion of nonpathogenic variants, as well as allelic and locus heterogeneity and can readily be applied outside of coding regions. In contrast to analyzing common variants using NPL, where loci localize to large genomic regions (e.g., >50 Mb), mapped regions are well defined for RV-NPL. Using simulation studies, we demonstrate that RV-NPL is substantially more powerful than applying traditional NPL methods to analyze RVs. The RV-NPL was applied to analyze 107 late-onset Alzheimer disease (LOAD) pedigrees of Caribbean Hispanic and European ancestry with WGS data, and statistically significant linkage (LOD $\geq$ 3.8) was found with RVs in *PSMF1* and *PTPN21* which have been shown to be involved in LOAD etiology. Additionally, nominally significant linkage was observed with RVs in *ABCA7*, *ACE*, *EPHA1*, and *SORL1*, genes that were previously reported to be associated with LOAD. RV-NPL is an ideal method to elucidate the genetic etiology of complex familial diseases.

Introduction

In recent years, there has been a great effort to understand the genetic contribution of rare variants (RVs) to the etiology of complex traits and diseases. The ability to study RVs has been greatly influenced by the availability of massively parallel sequencing, which led to the generation of whole-genome and -exome sequence data for hundreds of thousands of individuals. Most whole-genome and -exome sequence-based complex trait studies are performed using either case-control or population-based data,[1,2] but several studies have generated sequence data on families.[3,4] Although there are many RV aggregate association methods to analyze case-control and population-based data,[5–11] only a few have been developed to analyze families.[4,12–16] In addition to using family-based RV association methods to identify disease loci, linkage analysis can also be performed. Although parametric linkage analysis is inappropriate for complex trait analysis, non-parametric linkage (NPL)[17] (also known as model free or allele sharing methods) is a powerful approach to identify disease loci in families. Although RV parametric linkage methods have been developed,[18] this is not the case for NPL analysis.

Analyzing families segregating complex diseases can increase power to detect association signals compared to analyzing simplex cases, because pathogenic susceptibility variants segregating in pedigrees with multiple affected family members tend to have larger effect sizes.[19,20] Additional power can be obtained by analyzing multiple affected family members since frequencies of RVs tend to be increased and their heterogeneity reduced, compared to analyzing samples of unrelated case and control subjects. Therefore, when available, it is highly beneficial to analyze family data for complex traits.

For family-based association analysis, unaffected family members must be included in the analysis, but these unaffected individuals may be asymptomatic carriers of susceptibility variants because for complex traits penetrance can be incomplete.[21] This is even a greater problem for diseases with late onset age, since many unaffected family members will be below or within the age of onset. Additionally, RV aggregate association methods are generally sensitive to inclusion of non-causal variants.[22]

For NPL, the underlying assumption is that affected relatives will share identical by descent (IBD) susceptibility alleles or alleles that are in linkage disequilibrium (LD) with pathogenic variants.[23] Several NPL methods for common variants (minor allele frequency [MAF] > 0.05) have been developed for nuclear and extended families. For nuclear families, allele sharing is compared for affected

[1]Center for Statistical Genetics, Baylor College of Medicine, Houston, TX 77030, USA; [2]Department of Human Genetics, The University of Chicago, Chicago, IL 60637, USA; [3]Department of Neuroscience and Ronald M. Loeb Center for Alzheimer Disease, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA; [4]Department of Neurology, Taub Institute on Alzheimer Disease and the Aging Brain, and Gertrude H. Sergievsky Center, Columbia University, New York, NY 10027, USA; [5]Cologne Center for Genomics, Department of Statistical Genetics and Bioinformatics, University of Cologne, 50931 Cologne, Germany; [6]University Hospital Cologne, 50937 Cologne, Germany; [7]Departments of Genetics and Genomic Sciences and Neurology, Mount Sinai School of Medicine, New York, NY 10029, USA; [8]Center for Statistical Genetics, Columbia University, New York, NY 10027, USA
*Correspondence: sml3@cumc.columbia.edu
https://doi.org/10.1016/j.ajhg.2019.09.006.

sibpairs and it is determined if sharing deviates from the expectation under the null using the chi-square *goodness-of-fit* test,[24,25] maximum LOD score test (MLS),[26–28] mean test,[29] proportion test,[30,31]or minmax test.[32] Methods for extended pedigrees include the Affected Pedigree Member (APM) method,[33] which is obsolete since it analyzes identical-by-state (IBS) sharing, rather than IBD, and has increased type I and II errors. Kruglyak and colleagues developed an NPL approach,[17] which is based upon IBD sharing between affected family members. For this NPL method, there are several sharing measures, which include the commonly used S_{Pairs}, which measures IBD sharing within and between relative pairs, and S_{All}, which estimates the number of alleles from distinct affected pedigree members that are IBD. NPL methods have also been extended to perform multipoint linkage analysis.[17,26,34–36] These methods, although widely used, had limited success because causal susceptibility variants could not be identified due to disease loci mapping to large genetic intervals, e.g., > 50 Mb. Large intervals occur within families due to long-range LD between common variants. For common variants, locus heterogeneity is an additional problem, because it can dilute the linkage signal and increase the size of the mapped region.[37–39] Despite these limitations for the analysis of common variants, applying NPL to analyze RVs can overcome these problems. To further increase the power of analyzing RV using the NPL, an RV-specific method was developed, the RV-NPL, which examines sharing of only RV minor alleles to calculate the test statistic.

For the RV-NPL, an extension of the Kruglyak et al. NPL approach,[17] a regional locus is generated to analyze RVs in aggregate using the collapsed haplotype pattern (CHP) method.[18] IBD sharing is determined using the pedigree-specific regional locus and two IBD methods were developed: CHP-NPL and RV-NPL. As was performed for the NPL method, the CHP-NPL estimates IBD sharing for both major and minor alleles, i.e., haplotypes with at least one RV and haplotypes without any RVs. The RV-NPL estimates IBD sharing only for minor alleles, i.e., haplotypes that carry at least one RV. Using simulation studies and analyzing RVs (MAF < 0.01), the performance of CHP-NPL, RV-NPL, and multipoint NPL was compared. Although the power for the multipoint NPL and CHP-NPL are similar, the RV-NPL is substantially more powerful. When parental genotype data are missing, multipoint NPL analysis can have considerable inflated type I error rates[40] since the assumption that markers are in linkage equilibrium may be violated. Although markers can be pruned to remove LD, this can lead to a loss in power.[41] Therefore, it is not advisable to use multipoint NPL analysis when there is missing parental genotypes, which often occurs in family-based studies. For CHP-NPL and RV-NPL, type I error rates are well controlled. CHP-NPL and RV-NPL are both robust to population substructure and admixture between and within families, inclusion of nonpathogenic variants, and allelic and locus heterogeneity. In contrast to performing NPL analysis with common variants, the CHP-NPL and RV-NPL usually detect linkage to a small region, e.g., a gene, due to the low levels of LD between RVs. Both methods can also be used to analyze either gene regions or complete genomes using recombination events as boundaries for the regional locus. However, due to the superior power of the RV-NPL, it is recommended to use this method instead of the CHP-NPL.

The RV-NPL was applied to analyze whole-genome sequence (WGS) data generated on 107 nuclear and extended pedigrees with late-onset Alzheimer disease (LOAD) from the Alzheimer disease Sequencing Project (ADSP, dbGaP accession phs000572.v7.p4). Alzheimer disease (AD) is a neurodegenerative disease characterized by dementia that typically begins with subtle or poorly recognized failure of memory and slowly becomes more severe and incapacitating (see GeneReviews in Web Resources). AD is genetically heterogeneous with an estimated heritability of $h^2 = 60\%–80\%$.[43] Although genome-wide association studies (GWASs) of common variants have successfully identified LOAD loci, with the exception of *APOE*, each locus only accounts for a small fraction of disease susceptibility, and a large proportion of LOAD heritability remains unexplained.[44] Therefore, there is great interest in investigating the role RVs play in the etiology of AD. Application of the RV-NPL to ADSP WGS Caribbean Hispanic and European-ancestry pedigree data found significant evidence of linkage (LOD score $\geq$ 3.8)[45] between LOAD and nonsynonymous RVs in *PSMF1* (20p13 [MIM: 617858], GenBank: NM_178578.3, LOD = 3.87) and *PTPN21* (14q31.3 [MIM: 603271], GenBank: NM_007039, LOD = 3.81). *PSMF1* was previously shown to be associated with AD.[46–48] *PTPN21* was identified as a risk gene for AD in a Bayesian machine learning mediation analysis.[49] Functional studies suggest that both of these genes are potentially involved in AD etiology neurons.[50–53] Additionally, nominally suggestive linkage ($p < 0.05$) was observed with RVs in *ABCA7* (19p13.3 [MIM: 605414], GenBank: NM_019112.3), *ACE* (17q23.3 [MIM: 106180], GenBank: NM_000789.4), *EPHA1* (7q35 [MIM: 179610], GenBank: NM_005232.5), and *SORL1* (11q24.1 [MIM: 602005], GenBank: NM_003105.6). These genes were previously reported to be associated at a genome-wide significance level with AD.[47,48,54–57]

Material and Methods

Rare Variant Extension of NPL

For each pedigree, all variants are phased using an extension of the Lander-Green Algorithm.[58,59] After phasing, CHP-based regional loci[18] are constructed using RVs with MAFs below a given threshold criterion, e.g., < 1%. Regional loci can include either all RVs or only those that meet specific annotation specifications, e.g., missense, CADD c-score > 20. When there are missing founder or parental genotypes, regional loci genotypes are reconstructed or inferred based upon CHP genotypes from offspring and their family-specific CHP allele frequencies which are estimated

based on MAFs obtained from databases, e.g., gnomAD.[60] If the analyzed families are from different populations, then ancestry-specific MAFs should be used for each pedigree. Variants not observed in the relevant database population are assigned a MAF of (1 − k)/2N, where N is the number of individuals for the specific population in the database and k is the fraction of singletons observed.[61] If a sufficiently large sample is analyzed, MAFs can be estimated from pedigree founders and reconstructed founders. Only haplotypes that are observed in a pedigree are considered possible haplotypes to impute missing data for that pedigree. Frequencies for CHP alleles are calculated from the M observed RVs in the sample families, i.e., $\prod_{i=1}^{M}(1-f_i)$ for the wild-type CHP allele where f_i is the MAF for i^{th} observed RV. For alternative CHP alleles, the occurrence of minor allele in a haplotype with given haplotype pattern $[x_1,x_2,\ldots,x_M]$, $x_k \in [0,1]$ can be approximated by a multivariate Poisson distribution (details on the calculation can be found in Wang et al.[18]) and the individual frequency for each of the H observed alternative CHP alleles in a pedigree (i.e., RV-carrying haplotype h_k) is calculated by $\frac{P(h_k)}{\sum_{k=1}^{H}P(h_k)} \times [1-\prod_{i=1}^{M}(1-f_i)]$ where $P(h_k)$ is the probability from Poisson distribution for haplotype h_k such that the cumulative MAF for the alternative CHP alleles is $1-\prod_{i=1}^{M}(1-f_i)$. Based on the CHP allele frequencies, the missing parental genotypes can be reconstructed.

The alleles of the regional locus are scored to ensure that each haplotype within a pedigree is unique, so there is no loss of linkage information. Additional information on generating regional loci can be found in Wang et al.[18]

Each CHP regional locus is used to examine IBD (0, 1, or 2) allele sharing among affected pedigree members. For CHP-NPL, IBD sharing is calculated using both haplotypes with at least one RV and those haplotypes without any RVs. On the other hand, the test statistics for the RV-NPL are calculated only based on sharing of haplotypes that carry at least one RV, i.e., sharing of haplotypes that do not contain a RV does not contribute to linkage signals. Statistics are calculated using two different scoring functions, NPL_{All}, and NPL_{Pairs}, for both CHP-NPL and RV-NPL.

For NPL_{Pairs}, the sum of pairwise IBD sharing for affected pedigree members for the regional CHP locus is obtained for the j^{th} family with n_j affected relative-pairs by calculating the score,

$$S_{pairs,j} = \sum_{p=1}^{n_j} \tau_p,$$

where τ_p is the IBD sharing value for the p^{th} affected relative-pair, for RV-NPL, τ_p is the sharing of RV carrying haplotypes, with the score measuring the overall pairwise allele-sharing within the j^{th} family.

When there is no allelic heterogeneity, an increase in power can be obtained for families with more than two affected members using the *all* score which is implemented in the NPL_{All} test statistic. The *all* score was proposed by Whittemore and Halpern.[62] It can be calculated as follows,

$$S_{all,j}(v) = 2^{-a}\sum_{h}\left[\prod_{i=1}^{2f} b_i(h)!\right],$$

where a is the number of affected individuals in the j^{th} family, h is a collection of alleles from the region loci obtained by choosing one allele from each of the affected pedigree members (e.g., $h = [A_{11},A_{21},\ldots,A_{a1}]$ with A_{11} representing the 1st allele selected from 1st affected member and h has a total of 2^a possible combinations), and $b_i(h)$ denotes the number of times that the i^{th} founder (f) allele appears in h (for i = 1,...,2f). The sum is taken over all 2^a possible ways to choose h. For RV-NPL, $b_i(h)$ will be set to 0 if the i^{th} founder has a wild-type CHP allele, i.e., haplotype that does not contain any RVs. The score S_{all} is generated using an inheritance vector v for the j^{th} family, and the computation details of inheritance vector can be found in Whittemore and Halpern.[62]

To extend the analysis to the situation where precise IBD sharing values are unknown, the expected values over all possible inheritance patterns can be obtained by

$$S_j = \sum_{v} P_j(v_k)\cdot S_j(v_k),$$

where $P_j(v_k)$ is the posterior probability of inheritance vector v_k for the j^{th} family. For both approaches, a standardized score

$$Z_j = \frac{[S_j - \mu_j]}{\sigma_j}$$

is calculated for the j^{th} family, where S_j is the score for family j and μ_j and σ_j represent the mean and standard deviation of S_j under the null hypothesis, respectively. The null distribution of S_j is determined by enumerating every possible inheritance vector under the null for family j. Additionally, for RV-NPL, the null distribution is determined while maintaining the CHP genotypes of founders, since S_j for RV-NPL is dependent on the founder genotypes. The overall Z score is obtained as a linear combination of Z_j scores for a total of all families m,

$$Z = \sum_{j=1}^{m} \gamma_j Z_j,$$

where the weight is $\gamma_j = 1/\sqrt{m}$. Using either the S_{All} or S_{Pairs} score, the NPL_{All} or NPL_{Pairs} test statistic can be obtained.

When analyzing linkage across multiple RVs, unlike multipoint NPL, which can provide an NPL score at any map positions, CHP-NPL and RV-NPL give a single NPL score for a region. Moreover, in contrast to traditional NPL methods, CHP-NPL and RV-NPL use family-specific CHP allele frequencies enabling the inclusion of correct allele frequencies when joint analysis of families from different populations is performed. For the original version of the NPL,[17] it was demonstrated that the analytical p values were overly conservative when descent information is incomplete, i.e., missing genotype data.[63] Therefore, for CHP-NPL the Kong and Cox[63] extension was implemented to correct for the conservative nature of the NPL. For RV-NPL, empirical p values are obtained through permutations, by retaining founder haplotypes and then based on Mendelian segregation randomly assigning haplotypes to non-founders. For founders with missing genotypes, their haplotypes are reconstructed using genotype information from offspring and CHP marker allele frequencies obtained as described above. These haplotypes are then assigned to offspring based on Mendelian segregation. For family members with missing sequence data, their simulated genotypes are removed before analysis. Adaptive permutation is used to reduce computational time; p values are evaluated at pre-defined checkpoints, and permutation is terminated for tests that are not significant.

Simulation Framework

Type I error of RV-NPL and CHP-NPL were evaluated, through simulation studies. Additionally, the power of RV-NPL, CHP-NPL, and multipoint NPL were assessed and compared. Genotypes were simulated for 17,987 autosomal genes across the genome based on the observed variant sites and their corresponding

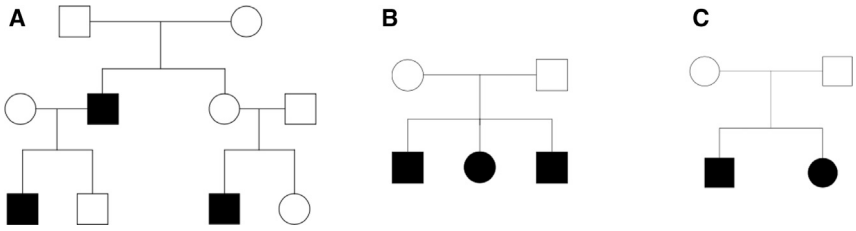


Figure 1. Pedigree Structures Used in the Simulation Studies
Three different pedigree structures were used for the simulation studies to evaluate type I error and power: Multi-generational pedigree with three affected family members (A), nuclear pedigree with three affected siblings (B), and nuclear pedigree with two affected siblings (C).

MAFs obtained from 33,370 Non-Finnish Europeans (NFE) recorded in the Exome Aggregation Consortium (ExAC)[60] database. Genetic maps distances and recombination rates were estimated from the Rutgers Combined Linkage-Physical map[64] using interpolation. Using RarePedSim,[65] sequence variant data were generated for three pedigree structures: nuclear families with either two or three affected siblings and an extended pedigree with two branches and three affected family members (Figure 1). Data were generated unconditional on affection status to evaluate type I error and conditional on disease status and phenotype model to evaluate power. Genotype phase information is removed from the simulated data to mimic real-world sample sequences and the data is phased using the Lander-Green Algorithm.[58,59] For both the evaluation of type I and II errors, genes with at least one variant with a ExAC NFE MAF ≤ 0.01 were analyzed.

Type I Error Evaluation

To evaluate type I error, genotype data were simulated for all autosomal genes across the genome unconditional on the affection status of the family members, i.e., odds ratio (OR) = 1.0. Type I error was evaluated for 100 extended families (Figure 1A); 300 nuclear families with three affected siblings (Figure 1B); and 2,000 nuclear families with two affected siblings (Figure 1C). Data were generated for pedigrees with no missing genotypes, and also for pedigrees with a percentage of founders missing all variant data. One thousand replicates of complete exomes were generated and every gene with one or more RVs was analyzed in each exome. p values were obtained both analytically (CHP-NPL) and empirically (RV-NPL) using one million permutations. Nominal p values were evaluated at 5.0×10^{-2} (LOD score 0.8), 5.0×10^{-3} (LOD score 1.45), and 1.5×10^{-5} (LOD score 3.8) levels and quantile-quantile (QQ) plots with results from all exome replicates were also generated. Type I error evaluation was performed for RV-NPL$_{Pairs}$, RV-NPL$_{All}$, CHP-NPL$_{Pairs}$, and CHP-NPL$_{All}$.

Power Evaluation

Power was evaluated for 100 extended families; 2,000 nuclear families with two affected siblings; and 300 nuclear families with three affected siblings. RV genotypes in each gene were generated conditional on affection status of the pedigree members assuming a multiplicative model for which each causal RV within a gene region has an OR of 5.0 and the disease has a prevalence of 0.01. Although a complex trait is being studied, an OR = 5.0 was selected, since variants for which familial aggregation is observed usually have larger effect sizes than susceptibility variants underlying sporadic disease. For every power evaluation, an exome was generated with 17,987 autosomal genes each linked to the disease, i.e., genotypes generated conditional on the affection status. Every gene with at least one RV was analyzed. It should be noted that since each gene was analyzed individually, genotypes at the other loci do not affect the results. For all analyses the power was determined by the ratio of number of genes with LOD > 3.8 (p value $\leq 1.5 \times 10^{-5}$), the genome-wide significance level proposed by Lander and Kruglyak,[45] to the total number of genes analyzed. Power was evaluated for RV-NPL$_{Pairs}$, RV-NPL$_{All}$, CHP-NPL$_{Pairs}$, and CHP-NPL$_{All}$ and for comparison purposes RVs in each gene region were also analyzed using multipoint NPL$_{Pairs}$ and NPL$_{All}$ as implemented in MERLIN.[58] p values were obtained analytically for CHP-NPL and multipoint NPL, and empirically using one million permutations for RV-NPL.

Simulations were performed under different scenarios to evaluate and compare the power. To estimate the effect of non-causal variants on power, two different scenarios were used. First, all nonsense, missense, and splice site variants were analyzed where 100%, 75%, and 50% are susceptibility variants (OR = 5.0) and the remaining variants (0%, 25%, and 50%) are neutral (OR = 1). Here the number of variants analyzed was kept constant and as the number of non-causal variants increased the number of susceptibility variants declined. Second, missense, nonsense, and splice site variants were assigned an OR = 5.0 and synonymous variants an OR = 1.0 (non-causal). The data were analyzed including and excluding synonymous variant to evaluate robustness of the methods to including non-causal variants while keeping the number of causal variants consistent. To evaluate the effect of missing data on power, analyses were performed with 10%, 30%, and 50% of the pedigrees having all founders missing their sequence data.

To appraise the effect of locus heterogeneity, data were generated under linkage homogeneity ($\alpha = 1.0$) and heterogeneity ($\alpha = 0.67$) and the power was compared. First pedigrees were generated with linkage (all nonsense, missense, and splice site variants have an OR = 5.0) with RVs in every autosomal gene generated conditional on the affection status and then an additional dataset 50% of the sample size of the linked families was generated under the null (all RVs unlinked with an OR = 1.0, i.e., generated unconditional on the pedigree affection status). For the extended pedigrees, 100 were generated with linkage and 50 unlinked. The linked pedigrees were first analyzed separately and then together with the unlinked ones.

Families with intra-familial heterogeneity (inclusion of simplex case subjects) were also simulated to evaluate the performance of RV-NPL$_{Pairs}$ and RV-NPL$_{All}$. Exome data with RVs in every autosomal gene were generated conditional on disease status for extended families with one branch containing two affected siblings and the other branch with an unaffected and affected sibling (Figures S1A). For the analysis, all children had an affected disease status (Figure S1B), to generate data with intra-familial heterogeneity.

Application to Alzheimer Disease Data

The RV-NPL was used to analyze families segregating LOAD. WGS data from 107 LOAD families with 486 members of which 446 have a LOAD diagnosis were available for analysis. The ADSP data were obtained from dbGaP (accession phs000572.v7.p4). WGS data for ADSP were generated at Baylor College of Medicine Human Genome Sequencing Center, Broad Institute Genome Center, and Genome Institute at Washington University. This dataset consists

Table 1. Power of CHP-NPL and RV-NPL

	Pairs						All					
	Sibpair[a]		Triplet[b]		Extended[c]		Sibpair		Triplet		Extended	
	CHP[d]	RV[e]	CHP	RV	CHP	RV	CHP	RV	CHP	RV	CHP	RV
100% causal[f]	0.801	0.954	0.817	0.932	0.789	0.859	0.801	0.954	0.817	0.930	0.780	0.858
75% causal	0.693	0.918	0.729	0.893	0.711	0.798	0.693	0.918	0.729	0.890	0.703	0.798
50% causal	0.521	0.825	0.588	0.797	0.580	0.672	0.521	0.825	0.588	0.793	0.569	0.671
NS & S[g]	0.686	0.890	0.771	0.901	0.761	0.856	0.686	0.890	0.771	0.899	0.756	0.856
Locus Het	0.778	0.947	0.808	0.925	0.787	0.858	0.778	0.947	0.808	0.924	0.778	0.858
10% MF[h]	0.798	0.953	0.815	0.931	0.788	0.859	0.798	0.953	0.815	0.930	0.779	0.858
30% MF	0.791	0.952	0.813	0.930	0.786	0.858	0.791	0.952	0.813	0.929	0.776	0.858
50% MF	0.784	0.951	0.808	0.929	0.780	0.857	0.784	0.951	0.808	0.928	0.769	0.858

Abbreviations: Het, heterogeneity; MF, missing founders; NS, nonsynonymous; and S, synonymous.
[a]2,000 nuclear families with 2 affected siblings
[b]300 nuclear families with 3 affected siblings
[c]100 extended families
[d]CHP-NPL method
[e]RV-NPL method
[f]Percentage of causal functional variants
[g]Analysis of causal nonsynonymous (NS) and non-causal synonymous (S) RVs
[h]Percentage of founders with missing genotypes

of 112 LOAD pedigrees from different populations: African American (1), Dominican (64), European ancestry (44), and Puerto Rican (3). For the analysis, family members were considered affected if their phenotype was defined as "definite AD," "probable AD," "possible AD," and "family-reported AD." The mean age of onset for AD was 72.63 years with a standard deviation of 8.46. *APOE* (MIM: 107741) status was also obtained for all family members and families were selected for WGS sequencing if no more than 75% of affected family members were heterozygous for *APOE* ε4 allele and none were homozygous. The African American family was excluded from the analysis, due to only a single family being available from this ancestry group. Three additional pedigrees were also excluded due to only one affected family member with available WGS data, making these families incompatible for linkage analysis. An additional pedigree was also excluded due to a high level of missing genotype data. A total of 42 families of European ancestry and 65 Caribbean Hispanic families (62 Dominican and 3 Puerto Rican) were analyzed. The pedigree structures and their ancestries are displayed in Figure S2 and Table S1, respectively.

In addition to the initial quality control (QC) performed by the ADSP QC working group,[4] genotypes with a genotype quality score (GQ) < 20 were removed. Only variant sites that were flagged as "PASS" for both the Broad and BCM pipelines, had a missing rate $\leq 10\%$, and had no Mendelian inconsistencies were included in analysis. Gene regions were assigned using RefSeq definitions. MAFs were annotated using the gnomAD database from the NFE and Latino (AMR) populations. For missing genotypes, CHP regional markers were constructed using gnomAD allele frequencies that corresponded to the family's ancestry, i.e., NFE or AMR. ANNOVAR was used to perform functional annotations.[66] RV-NPL$_{All}$ and RV-NPL$_{Pairs}$ were used to analyze every gene that had at least one RV site. Analysis was performed constructing regional markers within gene regions using frameshift, missense, nonsense, and splice sites variants with a MAF < 0.01 in gnomAD. European and Caribbean Hispanic families were analyzed jointly and separately to elucidate whether there were any association specific to one ancestry.

Results

Type I Error Evaluation

For nuclear (with two or three affected siblings) and extended pedigrees simulated under the null hypothesis of no association, nominal p values were evaluated at 5.0×10^{-2}, 5.0×10^{-3}, and 1.5×10^{-5} (Table S2) and quantile-quantile (QQ) plots were also generated (Figures S3–S6). These results suggest that type I error is well controlled for RV-NPL$_{All}$, RV-NPL$_{Pairs}$, CHP-NPL$_{All}$, and CHP-NPL$_{Pairs}$. It was also demonstrated that the type I error for CHP-NPL and RV-NPL (*all* and *pairs*) is well controlled when founder data were missing (Table S2 and Figures S3–S6).

Power Evaluation

Power was evaluated for nuclear (two and three affected siblings) and extended families analyzing RVs with MAF < 0.01 for RV-NPL$_{Pairs}$, RV-NPL$_{All}$, CHP-NPL$_{Pairs}$, CHP-NPL$_{All}$, multipoint NPL$_{Pairs}$, and multipoint NPL$_{All}$. Performance of the NPL methods was investigated when sequence data were missing for founders, when non-causal variants were included in the analysis, and in the presences of intra- and inter-familial heterogeneity. Since it has been established that multipoint NPL has increased type I error when there are missing parental data and LD is ignored,[40] analyses were not performed using multipoint-NPL when founders had their genotype data missing. For multipoint NPL and CHP-NPL for both S_{all} and S_{pairs} statistics, the power was identical for various scenarios when no data were missing (Table 1 and Figures 2 and S7–S10). For example, when simulated missense, nonsense, frameshift, and splice variants (MAF $<$ 0.01 and all causal) were analyzed for 300 nuclear families

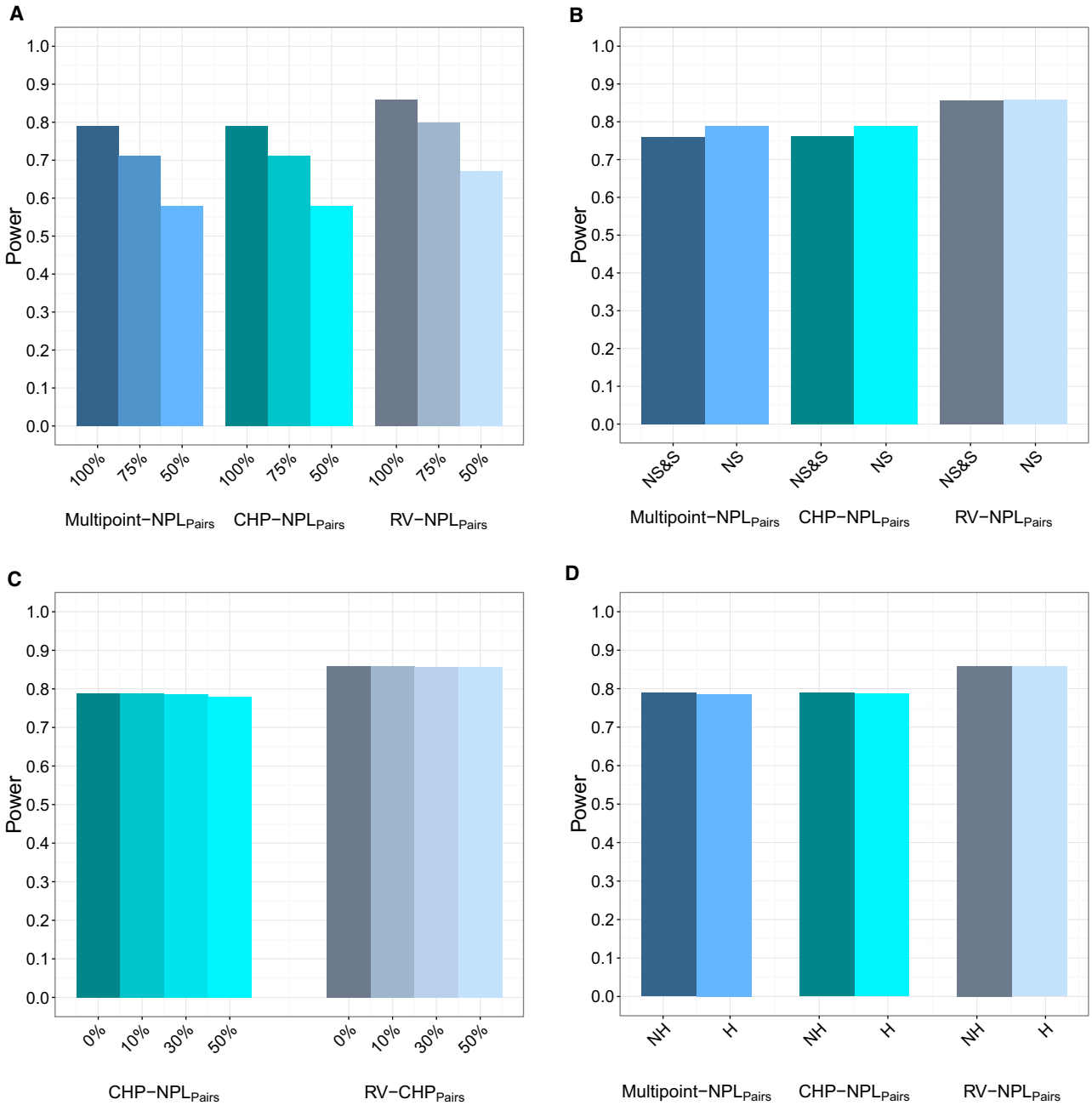


Figure 2. Exome-wide Power Comparison for RV-NPL$_{Pairs}$
Genotypes were generated for 100 extended families, conditional on affection status assuming a multiplicative model in which each causal variant within a gene region has an OR of 5.0. Analysis was performed using RV-NPL$_{Pairs,}$ CHP-NPL$_{Pairs}$, and Multipoint-NPL: with 100%, 75%, and 50% of the variant being causal and the remaining non-causal (OR = 1) (A); with only causal nonsynonymous (NS) variants as well as with causal nonsynonymous (NS) and non-causal synonymous (S) variants (B); with 0%, 10%, 30%, and 50% of the founders missing all genotype data (C); and with no heterogeneity (NH), i.e., 100 linked families as well as with locus heterogeneity (H), i.e., 100 linked and 50 unlinked families (D).

with three affected siblings, the power for multipoint NPL$_{Pairs}$ and CHP-NPL$_{Pairs}$ are both 0.817. Similarly, for extended families, the power for both multipoint NPL$_{Pairs}$ and CHP-NPL$_{Pairs}$ are 0.789. Since NPL$_{Pairs}$ and NPL$_{All}$ give identical results for affected sibpairs, the power is only displayed for NPL$_{Pairs}$.

For each scenario, the power for RV-NPL is consistently higher than for CHP-NPL and multipoint NPL for both S_{all} and S_{pairs}. The power is displayed for affected sibpairs, nuclear families with three affected siblings, and extended families in Table 1 and Figures 2 and S7–S10. When rare causal missense, nonsense, and splice site variants were analyzed for all autosomal genes for 2,000 affected sibpairs, the power for RV-NPL$_{Pairs}$ is 19.1% higher than for CHP-NPL$_{Pairs}$ and multipoint NPL$_{Pairs}$. For the same scenario, the power increases by 14.2% (300 nuclear families with

three affected siblings) and 8.8% (100 extended pedigrees) when RV-NPL$_{Pairs}$ instead of CHP-NPL$_{Pairs}$ or multipoint NPL$_{Pairs}$ was used to analyze the data (Table 1 and Figures 2, S7, and S8). Similar results are observed for NPL$_{All}$ (Table 1 and Figures S9 and S10).

For 100 extended pedigrees when 50% of the founders are missing their genotype data and all variants are causal, there is a 10.1% increase in power for RV-NPL$_{Pairs}$ compared to CHP-NPL$_{Pairs}$. Similarly, for 2,000 affected sibpairs when 50% of founders are missing all genotype data and all variants are causal, the power for RV-NPL$_{Pairs}$ is 21.3% higher than for CHP-NPL$_{Pairs}$ (Table 1, Figures 2C and S8C).

The impact of non-causal variants on power was also examined. In the first scenario, there is a set number of nonsynonymous variants, but the proportion that are causal was reduced from 100% to 50%. In the second scenario, all nonsynonymous variants are causal and analyzed and then both the causal nonsynonymous and non-causal synonymous variants were analyzed together so that the ratio of nonsynonymous (causal) to synonymous (non-causal) variants is 2:1, to mimic observed ratios of nonsynonymous to synonymous variants.[60] The first scenario was designed to evaluate a lower-powered yet possibly more realistic etiology for complex traits; the second scenario was designed to assess robustness of the methods to non-causal variants.

In the first scenario, when there is a set number of nonsynonymous variants, for RV-NPL, CHP-NPL, and multipoint NPL, the power decreases with decreasing proportion of causal variants and increasing non-causal variants. For example, compared to 100% causal variants, when only 50% of variants are causal and the rest non-causal, the power of RV-NPL$_{Pairs}$ decreases by 13.5%, 14.5%, and 21.7% for 2,000 affected sibpairs, 300 nuclear families with three affected siblings, and 100 extended families, respectively. For CHP-NPL$_{Pairs}$ and multipoint NPL$_{Pairs}$, the power for 300 nuclear families with three affected siblings both dropped from 0.817 to 0.588 (by 28.0%) when the proportion of causal variants decreased from 100% to 50% and non-causal variants increased from 0% to 50%. Similarly, for extended families, the power for both CHP-NPL$_{Pairs}$ and multipoint NPL$_{Pairs}$ decreased from 0.789 to 0.580 (by 26.5%). Regardless of the proportion of causal variants, RV-NPL consistently displayed higher power than CHP-NPL and multipoint NPL, e.g., when the proportions of causal and non-causal variants are each 50%, the power for RV-NPL$_{Pairs}$ is 58.2%, 35.5%, and 15.9% higher than for CHP-NPL$_{Pairs}$ and multipoint NPL$_{Pairs}$ for affected sibpairs, nuclear families with three affected siblings, and extended families, respectively (Table 1 and Figures 2A, S7A, and S8A). Similar results were observed for NPL$_{All}$ (Table 1 and Figures S9A and S10A). There is a greater loss of power for CHP-NPL and multipoint NPL compared to RV-NPL as the proportion of causal variants decreases and the non-causal variants increases, e.g., for affected sibpairs when the proportion of causal variants were reduced from 100% to 75%, RV-NPL$_{Pairs}$ displays a modest 3.7% reduction in power while the power for CHP-NPL$_{Pairs}$ and multipoint NPL$_{Pairs}$ is reduced by 13.4%; when the proportion of causal variants was further decreased from 100% to 50%, RV-NPL$_{Pairs}$ has a 13.5% reduction in power, while the power reduction, 34.9%, for the CHP-NPL$_{Pairs}$ and multipoint NPL$_{Pairs}$ is again more severe, which results in an increased power discrepancy between RV-NPL and CHP-NPL from 19.1% to 58.2%. Similarly, for extended pedigrees, when the proportion of causal variants was reduced from 100% to 75%, the power loss for RV-NPL$_{Pairs}$ is 7.1% compared to a 9.9% for CHP-NPL$_{Pairs}$ and multipoint NPL$_{Pairs}$, and similar results were observed when the proportion of causal variants was reduced from 100% to 50%. This same trend is also observed for affected nuclear families with three affected siblings as well as for NPL$_{All}$, suggesting that RV-NPL is more robust to a reduction in causal variants and an inclusion of non-causal variants than CHP-NPL and multipoint NPL (Table 1, Figures 2A, S7A, S8A, S9A, and S10A).

In the second scenario, inclusion of non-causal synonymous variants in the analysis causes substantial reductions in power for CHP-NPL$_{Pairs}$ and multipoint NPL$_{Pairs}$, yet RV-NPL power remains robust to the inclusion of non-causal variants. For example, the initial power for RV-NPL$_{Pairs}$ was 29.8%, 16.8%, and 12.4% higher than for CHP-NPL$_{Pairs}$ and multipoint NPL$_{Pairs}$ for 2,000 affected sibpairs, 300 nuclear families with three affected siblings, and 100 extended families, respectively. The reduction in power for RV-NPL$_{Pairs}$ when non-causal variants were included in the analysis is 6.6% (2,000 affected sibpairs), 3.4% (300 nuclear families with three affected siblings), and 0.4% (100 extended families) while for both CHP-NPL$_{Pairs}$ and multipoint NPL$_{Pairs}$ the drop in power when non-causal variants were included in the analysis is 14.4% (2,000 affected sibpairs), 5.5% (300 nuclear families with three affected siblings), and 3.5% (100 extended families), respectively (Table 1 and Figures 2B, S7B, and S8B) and similar results for NPL$_{All}$ can be found in Table 1 and Figures S9B and S10B. These results again support that RV-NPL is more robust to non-causal variants than CHP-NPL and multipoint NPL.

Furthermore, the power of RV-NPL is largely maintained when there is missing genotype data. When simulating 10%, 30%, and 50% of families with all founders missing their sequence data, the power of both RV-NPL and CHP-NPL decreases as the proportion of pedigrees missing founder data increases. For all pedigree structures, while the power loss for RV-NPL and CHP-NPL are both very minor, RV-NPL is still more robust to missing genotype data. For example, for each of the three pedigree structures, RV-NPL$_{Pairs}$ loses 0.1% power on average when 30% of the pedigrees are missing sequence data for all founders compared to when no data is missing, while CHP-NPL$_{Pairs}$ loses 0.7% power on average. When 50% of the pedigrees with all founders missing all sequence data, RV-NPL$_{Pairs}$ loses 0.2% power on average compared to when there is

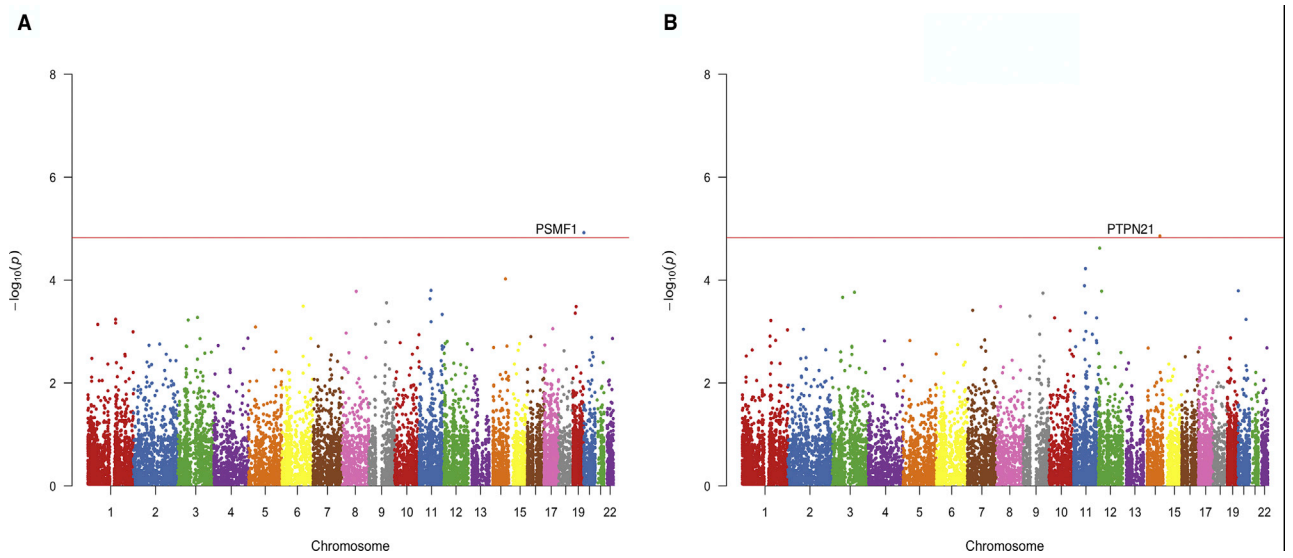


Figure 3. Manhattan Plots Displaying the RV-NPL Results for the Analysis of the ADSP Pedigrees
The results from analyzing functional RVs in 107 ADSP pedigrees with European and Hispanic origin for RV-NPL$_{Pairs}$ (A) and RV-NPL$_{All}$ (B) are displayed with the red line indicating the significance threshold of LOD = 3.8.

no missing data, and on average the power loss for CHP-NPL$_{Pairs}$ was 1.4% for each of the pedigree structures (Table 1, Figures 2C, S7C, and S8C).

Simulation results for all pedigree structures also demonstrate that RV-NPL is robust to locus heterogeneity, e.g., with only a 0.7% loss in power when 3,000 affected sib-pairs were analyzed (1/3 [1,000 pedigrees] unlinked to the disease locus and 2/3 [2,000 pedigrees] linked to all simulated disease loci [$\alpha = 0.67$]) compared to when only 2,000 affected sibpairs with linkage ($\alpha = 1.0$) were analyzed. For this same scenario, CHP-NPL and multipoint NPL have < 3% loss of power. Additionally, only very small decreases in power were observed for the analysis of RVs when there was locus heterogeneity for nuclear families with three affected siblings and the extended pedigrees regardless of whether the analysis was performed using RV-NPL, CHP-NPL, or multipoint-NPL (Table 1, Figures 2D, S7D, S8D, S9D, and S10D).

For RV-NPL$_{All}$ and RV-NPL$_{Pairs}$, there is no difference in power for affected sibpairs, since for this family structure the methods are equivalent. For nuclear families with three affected siblings and extended families, the power for S_{pairs} was slightly higher than for S_{all}, e.g., when analyzing 100% causal variants for nuclear families with three affected siblings, the power of RV-NPL$_{Pairs}$ is 0.932 compared to 0.930 for RV-NPL$_{All}$. This slight difference is due to intra-familial heterogeneity since there can be simplex case subjects in the families due to the OR and disease prevalence used to generate the RV data conditional on the affection status. The power discrepancy between RV-NPL$_{Pairs}$ and RV-NPL$_{All}$ increases when the proportion of causal variants was decreased from 100% to 50% causal, e.g., for 300 nuclear families with three affected siblings, the difference in power changed from 0.2% (0.932 for RV-NPL$_{Pairs}$ and 0.930 for RV-NPL$_{All}$) to 0.5% (0.797 for RV-NPL$_{Pairs}$ and 0.793 for RV-NPL$_{All}$). A similar trend was observed in extended families (Table 1, Figures S7A and S10A). Due to the design of the test, S_{all} is less robust to intra-familial heterogeneity than S_{pairs}. However, in the absence of intra-familial heterogeneity, S_{all} can provide higher test statistics than S_{pairs}. It was observed that 69.1% of the generated genes have a higher test statistic for S_{all} than S_{pairs} (Figure S1 and Table S3). Additionally, we used the proportion of families that have only one type of RV haplotype (i.e., all RV haplotypes observed in a family are the same) as a proxy of allelic homogeneity within a family to further demonstrate its impact on the power of S_{all} and S_{pairs}. It was observed that for genes with a higher test statistic for S_{all} than S_{pairs}, 79% of the family have only one RV haplotype, while for genes with a higher test statistic for S_{pairs} than S_{all}, 66% of families have only one RV haplotype. We also evaluated power for RV-NPL$_{Pairs}$ and RV-NPL$_{All}$ when 100 extended pedigrees were generated with intra-familial heterogeneity (changing the affection status to increase the number of simplex cases) were analyzed, and observed that *pairs* is more powerful than S_{all}, with the power of RV-NPL$_{Pairs}$ being 4.4% higher than that of RV-NPL$_{All}$ when all simulated nonsynonymous RVs are causal.

Analysis of Alzheimer Disease Sequencing Project Data

Joint analysis of the Caribbean Hispanics and Europeans identified linkage with two genes, *PSMF1* (LOD: 3.87) and *PTPN21* (LOD: 3.81), observed reaching the significance threshold of a LOD score $\geq$ 3.8[45] (Figure 3). No genes reached a significant LOD score of $\geq$ 3.8 when analyses were performed separately for Caribbean Hispanics and Europeans.

Additionally, nominal significance was observed for several genes that were demonstrated to be associated

with LOAD: *ABCA7* (RV-NPL$_{Pairs}$ $p = 3.0 \times 10^{-2}$, RV-NPL$_{All}$ $p = 6.0 \times 10^{-3}$) and *EPHA1* (RV-NPL$_{Pairs}$ $p = 7.0 \times 10^{-3}$, RV-NPL$_{All}$ $p = 6.0 \times 10^{-3}$) display nominal significance in Caribbean Hispanic families while *ACE* (RV-NPL$_{Pairs}$ $p = 2.8 \times 10^{-2}$, RV-NPL$_{All}$ $p = 2.8 \times 10^{-2}$) and *SORL1* (RV-NPL$_{Pairs}$ $p = 1.6 \times 10^{-2}$, RV-NPL$_{All}$ $p = 1.5 \times 10^{-2}$) are nominally significant in European families.

For *PSMF1* (RV-NPL$_{Pairs}$ $p = 1.2 \times 10^{-5}$, RV-NPL$_{All}$ $p = 1.6 \times 10^{-4}$), seven out of eight missense RVs observed segregate in 14 families with increased RV minor allele sharing, and five RVs are located in conserved nucleotide sites (Table S4). For *PTPN21* (RV-NPL$_{Pairs}$ $p = 9.5 \times 10^{-5}$, RV-NPL$_{All}$ $p = 1.4 \times 10^{-5}$), six missense RVs were observed segregating in eight families with enhanced minor RV allele sharing (Table S5). For *ABCA7*, 13 missense RVs segregate in 20 pedigrees with RV allele sharing greater than expected under the null hypothesis of no linkage (Table S6). For *ACE*, three missense RVs were observed in three linked pedigrees, and two of these RVs are conserved and deemed deleterious by at least six bioinformatics tools (Table S7). Two missense RVs in *EPHA1* were observed in three pedigrees with linkage, and both RVs are located in conserved sites and deemed deleterious by at least six bioinformatics tools (both have CADD scaled C-scores = 35, Table S8). For *SORL1*, seven missense RVs were segregating in seven pedigrees with increased sharing. Five of the segregating RVs are conserved and deemed deleterious by at least four of seven bioinformatics tools (Table S9). Pedigrees segregating variants in *ABCA7*, *ACE*, *EPHA1*, *PSMF1*, *PTPN21*, and *SORL1* are shown in Figure S2 and Table S1.

Discussion

We developed the RV-NPL, to perform aggregated rare-variant NPL analysis, using the CHP method.[18] Based on simulation studies, we demonstrated that RV-NPL has well-controlled type I error. It is a powerful approach to map complex trait loci with familial aggregation and is robust to locus and allelic heterogeneity as well as inclusion of non-causal variants.

Parametric linkage analysis should be used for Mendelian traits, since NPL methods will be less powerful. However, when the genetic model is unknown (which is usually the case for complex traits), NPL is more powerful than parametric linkage analysis,[67] since for parametric linkage analysis incorrect specification of the disease and penetrance model will lead to a severe loss in power. The power of the NPL is not affected by an unknown underlying genetic model, since it is not specificed.[17] Therefore for the analysis of complex trait family data, NPL and not parametric linkage analysis should be used.

RV-NPL has several major advances over traditional NPL methods. First, it is more powerful than traditional multipoint NPL methods under a variety of simulation scenarios. Additionally, analyzing RVs instead of common ones provides better resolution of the linkage region, usually to within a gene or a small genomic region, which is demonstrated in the analysis of the ADSP pedigrees. Applying NPL methods to analyze common variants led to large genetic intervals, due to their LD structure in families.[68,69] In contrast, two factors that aid in fine mapping of loci for RVs are their low levels of LD and the fact that linked variants often differ between families. Moreover, resolution can be further refined when recombination events occur within a gene region, allowing linkage signals to be mapped to sub-units of a gene divided by recombination.

Unlike for parametric linkage analysis where locus heterogeneity can be modeled in the linkage framework, NPL methods do not allow for the incorporation of linkage admixture into the analysis, which for common variant analysis can greatly attenuate power. For RV linkage analysis, there is very little loss in power with the presence of locus heterogeneity because unlinked regions usually do not contain informative variants and thus do not contribute to confounding RV-NPL analysis. This is advantageous for the analysis of complex traits due to extensive locus heterogeneity.

It has previously been demonstrated that NPL methods are robust to population substructure and admixture between and within families.[20] For linkage analysis, in the presence of missing data, type I error can be increased when incorrect allele frequencies are used in the analysis.[40] For each family, population-specific allele frequencies should be used. For example, when the ADSP admixed Caribbean Hispanic families were analyzed, allele frequencies were obtained from the gnomAD AMR population while for the ADSP European-Americans allele frequencies from the NFE were used, to avoid an increase in type I error due to the use of incorrect allele frequencies. No inflation of type I error was observed when mega-analysis was used to analyze families of European and Caribbean Hispanic ancestry. Although for family-based RV aggregate association analysis type I error can be well controlled when there is population admixture and substructure,[20,70] an additional problem is that inclusion of non-causal variant can attenuate the signal when the underlying genetic etiology varies by ancestry. Using RV-NPL, families of different ancestries can be analyzed jointly, since linkage is robust to inclusion of families that are not linked to the same loci and non-causal variants.

Though family data can provide several benefits in mapping causal RVs, family-based studies do have drawbacks. The recruitment of probands and their relatives is more time consuming and expensive compared to the ascertainment of unrelated individuals. Pedigrees often have diverse structures and so it is necessary to be able to analyze multiplex pedigrees. The NPL method lends itself well to this situation. Additionally, parental data are often unavailable for families, in particular for late-onset diseases. RV-NPL has only a minimal power loss when founders and parents were missing their genotype data, and for the analysis of

the ADSP pedigrees that include complex multi-generational pedigrees that have a large proportion of founders missing all variant data, type I error was well controlled.

As suggested by Lander and Kruglyak,[45] a LOD score of 3.8 was used as the significance threshold to control the family-wise error rate and provide a genome-wide significance level of 0.05 regardless of the marker loci density. Although Lander and Kruglyak proposed different thresholds depending on the observed pedigree relationships, e.g., sibpairs or uncle-nephew, and then weighting significance levels based on the proportion of each family type, here we apply the most stringent level suggested, a LOD score of 3.8 ($p = 1.5 \times 10^{-5}$) regardless of pedigree structure.

Our study observed excess RV sharing for *PSMF1* among affected members of fourteen pedigrees and for *PTPN21* among affected members of eight pedigrees (Table S1). None of the affected RV carriers in these pedigrees are positive for the *APOE* ε4 allele. For *PSMF1*, there are three European families and eleven Caribbean Hispanic families with RV allele with increased minor allele sharing. Among the eight pedigrees with increased minor allele sharing in *PTPN21*, seven are Caribbean Hispanic and one is European. Different RVs were found in Caribbean Hispanic and European pedigrees: in *PSMF1,* seven of the eleven Caribbean Hispanic pedigrees and none of the European pedigrees displayed linkage to rs79465651 which is a conserved nucleotide site. For *PTPN21*, rs150736820 had increased minor allele sharing in the European pedigree, but was not observed in the seven Caribbean Hispanic pedigrees; while rs3825676 had increased minor allele sharing in three out of seven Caribbean Hispanic pedigrees. For both genes, the linkage signals from ancestry specific analyses are weaker than those from the combined pedigrees, suggesting the potential benefit of performing mega-analysis.

LOAD associations with common and rare variants in *PSMF1* were previously reported. A small LOAD GWAS study (124 cases) of Israeli Arabs with a low frequency of APOE ε4 carriers reported several common variants associations in the *PSMF1* gene region with the most significant SNV having a $p = 3.6 \times 10^{-5}$.[46] Associations were also observed in the Alzheimer disease Genetics Consortium (ADGC) and the International Genomics of Alzheimer's Project (IGAP) datasets. For the ADGC study with 1,968 African- American LOAD cases an association was observed with variant rs35517343 (MAF 0.014 $p = 1.9 \times 10^{-6}$) which is in the splice region of *PSMF1*.[47] Rs35517343 is extremely rare in non-African populations. The discovery stage of the IGAP study in individuals of European ancestry observed a nominal significance of $p = 1.6 \times 10^{-3}$ with RV rs202107404 (MAF = 0.00002) which lies in the intronic region of *PSMF1*.[48] Functional studies also provide support to *PSMF1* potential role in AD etiology. *PSMF1* encodes a protein that, through the 11S and 19S regulators, inhibits the activation of the 26S proteasome, which regulates Aβ metabolism and tau degradation. The functional impairment of 26S proteasome, especially in neurons, decreases the activity of α-secretase and leads to the production and accumulation of Aβ,[52] which is an important feature of AD, therefore suggesting the potential involvement of *PSMF1* in AD etiology via inhibiting the function of 26S proteasome. For *PTPN21*, a previous causal mediation analysis combining large-scale GWAS and brain gene expression data for Europeans, identified this gene as a strong causal mediator for AD.[49] It has also been found to promote neuron survival through ErbB4/NRG3 pathway and increase neuritic length,[53] which is vital for maintaining normal neuronal function, suggesting a potential important role in neural development. A previous GWAS study found *PTPN21* significantly associated with schizophrenia,[51] suggesting its involvement in the pathogenesis of neural diseases.

Four known AD-associated genes (*ABCA7*, *ACE, EPHA1*, and *SORL1*) displayed linkage signals with nominal significance in RV-NPL analysis of the ADSP data. A variety of common variants in *ABCA7* have been identified as susceptibility loci for LOAD through several GWAS analyses in European and African American populations.[47,48,54,56] For Europeans, RVs were also reported in several association analyses of LOAD that were performed using targeted sequencing of AD- associated genes.[71–73] In a French case-control sample, gene-level RV association analysis identified a significance association between *ABCA7* and early-onset AD.[74] In our study, *ABCA7* displayed nominal significance in Caribbean Hispanic, but not in European families. Only weak association with RVs in *ABCA7* were previously reported for LOAD in Caribbean Hispanics[72] and our finding lend support to its involvement in this population. *ACE* was identified as a risk gene for LOAD with significant association in European population[48] and an Israeli Arab community,[75] it could also impact the risk of LOAD by regulating the level of Aβ.[76] *ACE* reached nominal significance only in the analysis of the European pedigrees. Previous associations for *ACE* were only for common variants and this study suggests that functional rare variants may also be involved. *EPHA1* was first implicated in LOAD etiology through the association of rs11767557, a common variant in the promoter region, which was reported in two LOAD GWAS meta-analyses of European populations.[54,55] Another associated common variant was later found by a GWAS in European population.[56] Additionally, a targeted sequencing study identified RV rs202178565 to be significantly enriched in Caribbean Hispanics LOAD patients.[72] In our study, although rs202178565 was not present, *EPHA1* still displayed nominal significance in Caribbean Hispanic pedigrees, supporting its involvement in Hispanics. Evidence of linkage in Europeans was not observed for RVs in *EPHA1*. *SORL1* is associated with increased risk of both early- and late-onset AD[77,78] and it is involved in the AD etiology through aberrant trafficking and metabolism of the amyloid precursor protein (APP)[79] that could increase Aβ. Both common variants and RVs have been reported as LOAD risk loci in

SORL1. In a family-based joint linkage and association study on targeted sequence data, *SORL1* RV rs143571823 showed significant segregation with disease in 87 Caribbean Hispanic LOAD families.[57] For Europeans, several GWAS meta-analyses identified significant associations between common variants in *SORL1* and LOAD.[56,77] Although none of the known risk variants were present in ADSP analysis, *SORL1* displayed nominal significance in European but not in Caribbean Hispanic pedigrees. For pedigrees that display linkage to *ABCA7*, *ACE, EPHA1*, and *SORL1* none of the affected pedigree members are positive for the *APOE* ε4 allele. Considering that the application of RV-NPL in ADSP pedigrees focused on RVs, these findings suggest that genes implicated with AD may harbor both common and rare susceptibility variants.

Although RV-NPL was used to analyze gene regions in genomes, it can also be implemented to analyze complete genomes, using recombination events as boundaries for the regional locus. The ability to use recombination events to aggregate variants is an advantage to RV association methods where prior knowledge or a sliding window are necessary to aggregate RVs outside of gene regions.

RV-NPL is a robust and powerful tool to map RVs for complex disease segregating in families. Results from extensive simulation studies and the analysis of the ADSP data demonstrate the power and robustness of RV-NPL, as well as its ability to fine map loci and to detect linkage to individual genes. These characteristics make RV-NPL an ideal method to elucidate the genetic etiology of complex familial diseases. RV-NPL is implemented primarily in Python with C++ extensions, and the software package is publicly available online.

Accession Numbers

The dbGaP accession number for the genome sequences reported in this paper is phs000572.v7.p4.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2019.09.006.

Acknowledgments

We wish to thank the family members who participated in the Alzheimer Disease Sequencing Project and made this research possible. The datasets used for the analyses in this manuscript were obtained from the database of Genotypes and Phenotypes (dbGaP) at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000572.v7.p4 through dbGaP accession study number phs000572.v7.p4. We would like to thank dbGaP for distributing the data used in this study. The National Institute on Aging (NIA)-LOAD study supported the collection of samples used in this study through NIA grants U24AG026395 and R01AG041797. We thank contributors, including the Alzheimer Disease Centers who collected samples used in the NIA-LOAD study, as well as patients and their families, whose help and participation made this work possible. Data collection for this project was also supported by the Genetic Studies of Alzheimer Disease in Caribbean Hispanics (EFIGA) funded by the NIA grants 5R37AG015473, RF1AG015473, and R56AG051876. We acknowledge the EFIGA study participants and the EFIGA research and support staff for their contributions to this study. This work was also supported by grants from the National Human Genome Research Institute R01 HG008972 and NIA RF1 AG058131. Complete acknowledgments can be found in the Supplemental Acknowledgments.

Declaration of Interests

The authors declare no competing interests.

Received: May 11, 2019
Accepted: September 5, 2019
Published: October 3, 2019

Web Resources

ADSP, https://www.niagads.org/adsp/content/home
ANNOVAR, http://annovar.openbioinformatics.org/en/latest/
CADD, https://cadd.gs.washington.edu/
dbGAP, https://www.ncbi.nlm.nih.gov/gap/
dbSNP, https://www.ncbi.nlm.nih.gov/projects/SNP/
ExAC Browser, http://exac.broadinstitute.org/
fathmm, http://fathmm.biocompute.org.uk/
GenBank, https://www.ncbi.nlm.nih.gov/genbank/
GeneReviews, Bird, T.D. (1993). Alzheimer disease overview. https://www.ncbi.nlm.nih.gov/books/NBK1161/
gnomAD, https://gnomad.broadinstitute.org/
LRT, http://www.genetics.wustl.edu/jflab/lrt_query.html
Merlin, http://csg.sph.umich.edu/abecasis/merlin/
Mutation Taster, http://www.mutationtaster.org/
OMIM, https://www.omim.org/
PLINK 1.9, https://www.cog-genomics.org/plink2/
PolyPhen-2, http://genetics.bwh.harvard.edu/pph2/
PROVEAN, http://provean.jcvi.org/
RV-NPL, https://github.com/statgenetics/rvnpl
UCSC Genome Browser, http://genome.ucsc.edu/

References

1. Walter, K., Min, J.L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J.R., Xu, C., Futema, M., Lawson, D., et al.; UK10K Consortium (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82–90.
2. Schick, U.M., Auer, P.L., Bis, J.C., Lin, H., Wei, P., Pankratz, N., Lange, L.A., Brody, J., Stitziel, N.O., Kim, D.S., et al.; Cohorts for Heart and Aging Research in Genomic Epidemiology; and National Heart, Lung, and Blood Institute GO Exome Sequencing Project (2015). Association of exome sequences with plasma C-reactive protein levels in >9000 participants. Hum. Mol. Genet. *24*, 559–571.
3. Engelman, C.D., Greenwood, C.M.T., Bailey, J.N., Cantor, R.M., Kent, J.W., Jr., König, I.R., Bermejo, J.L., Melton, P.E., Santorico, S.A., Schillert, A., et al. (2016). Genetic Analysis Workshop 19: methods and strategies for analyzing human sequence and gene expression data in extended families and unrelated individuals. BMC Proc. *10* (*Suppl 7*), 67–70.

4. He, Z., Zhang, D., Renton, A.E., Li, B., Zhao, L., Wang, G.T., Goate, A.M., Mayeux, R., and Leal, S.M. (2017). The Rare-Variant Generalized Disequilibrium Test for Association Analysis of Nuclear and Extended Pedigrees with Application to Alzheimer Disease WGS Data. Am. J. Hum. Genet. *100*, 193–204.
5. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. *83*, 311–321.
6. Morris, A.P., and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet. Epidemiol. *34*, 188–193.
7. Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. *5*, e1000384.
8. 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.
9. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.
10. Auer, P.L., Reiner, A.P., Wang, G., Kang, H.M., Abecasis, G.R., Altshuler, D., Bamshad, M.J., Nickerson, D.A., Tracy, R.P., Rich, S.S., Leal, S.M.; and NHLBI GO Exome Sequencing Project (2016). Guidelines for large-scale sequence-based complex trait association studies: lessons learned from the NHLBI Exome Sequencing Project. Am. J. Hum. Genet. *99*, 791–801.
11. Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., and Lander, E.S. (2014). Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. USA *111*, E455–E464.
12. 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.
13. De, G., Yip, W.-K., Ionita-Laza, I., and Laird, N. (2013). Rare variant analysis for family-based design. PLoS ONE *8*, e48495.
14. 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.
15. 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., and Satten, G.A. (2015). A statistical approach for rare-variant association testing in affected sibships. Am. J. Hum. Genet. *96*, 543–554.
16. 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.
17. Kruglyak, L., Daly, M.J., Reeve-Daly, M.P., and Lander, E.S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. Am. J. Hum. Genet. *58*, 1347–1363.
18. Wang, G.T., Zhang, D., Li, B., Dai, H., and Leal, S.M. (2015). Collapsed haplotype pattern method for linkage analysis of next-generation sequence data. Eur. J. Hum. Genet. *23*, 1739–1743.
19. 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.
20. Ott, J., Kamatani, Y., and Lathrop, M. (2011). Family-based designs for genome-wide association studies. Nat. Rev. Genet. *12*, 465–474.
21. Hopper, J.L., Bishop, D.T., and Easton, D.F. (2005). Population-based family studies in genetic epidemiology. Lancet *366*, 1397–1406.
22. Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet. *95*, 5–23.
23. Ziegler, A., König, I.R., and Pahlke, F. (2010). A Statistical Approach to Genetic Epidemiology: Concepts and Applications, with an E-learning platform (John Wiley & Sons).
24. Motro, U., and Thomson, G. (1991). Affected kin-pair IBD methods: genetic models. Genet. Epidemiol. *8*, 317–327.
25. Holmans, P. (1998). Affected sib-pair methods for detecting linkage to dichotomous traits: review of the methodology. Hum. Biol. *70*, 1025–1040.
26. Risch, N. (1990). Linkage strategies for genetically complex traits. I. Multilocus models. Am. J. Hum. Genet. *46*, 222–228.
27. Risch, N. (1990). Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am. J. Hum. Genet. *46*, 229–241.
28. Hauser, E.R., Boehnke, M., Guo, S.W., and Risch, N. (1996). Affected-sib-pair interval mapping and exclusion for complex genetic traits: sampling considerations. Genet. Epidemiol. *13*, 117–137.
29. Blackwelder, W.C., and Elston, R.C. (1985). A comparison of sib-pair linkage tests for disease susceptibility loci. Genet. Epidemiol. *2*, 85–97.
30. Day, N.E., and Simons, M.J. (1976). Disease susceptibility genes–their identification by multiple case family studies. Tissue Antigens *8*, 109–119.
31. Suarez, B.K., Rice, J., and Reich, T. (1978). The generalized sib pair IBD distribution: its use in the detection of linkage. Ann. Hum. Genet. *42*, 87–94.
32. Whittemore, A.S., and Tu, I.P. (1998). Simple, robust linkage tests for affected sibs. Am. J. Hum. Genet. *62*, 1228–1242.
33. Weeks, D.E., and Lange, K. (1988). The affected-pedigree-member method of linkage analysis. Am. J. Hum. Genet. *42*, 315–326.
34. Kruglyak, L., and Lander, E.S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am. J. Hum. Genet. *57*, 439–454.
35. Marlow, A.J., John, S., and Worthington, J. (1997). Multipoint analysis of quantitative traits. Genet. Epidemiol. *14*, 845–850.
36. O'Connell, J.R. (2001). Rapid multipoint linkage analysis via inheritance vectors in the Elston-Stewart algorithm. Hum. Hered. *51*, 226–240.
37. Kruglyak, L., and Lander, E.S. (1996). Limits on fine mapping of complex traits. Am. J. Hum. Genet. *58*, 1092–1093.
38. Finch, S.J., Chen, C.-H., Gordon, D., and Mendell, N.R. (2001). A study comparing precision of the maximum multipoint heterogeneity LOD statistic to three model-free multipoint linkage methods. Genet. Epidemiol. *21*, 315–325.
39. Greenberg, D.A., and Abreu, P.C. (2001). Determining trait locus position from multipoint analysis: accuracy and power of three different statistics. Genet. Epidemiol. *21*, 299–314.
40. Huang, Q., Shete, S., and Amos, C.I. (2004). Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. Am. J. Hum. Genet. *75*, 1106–1112.

41. Moskvina, V., Schmidt, K.M., Vedernikov, A., Owen, M.J., Craddock, N., Holmans, P., and O'Donovan, M.C. (2012). Permutation-based approaches do not adequately allow for linkage disequilibrium in gene-wide multi-locus association analysis. Eur. J. Hum. Genet. *20*, 890–896.
43. 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.
44. 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.
45. Lander, E., and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat. Genet. *11*, 241–247.
46. Sherva, R., Baldwin, C.T., Inzelberg, R., Vardarajan, B., Cupples, L.A., Lunetta, K., Bowirrat, A., Naj, A., Pericak-Vance, M., Friedland, R.P., et al. (2011). Identification of Novel Candidate Genes for Alzheimer Disease by Autozygosity Mapping Using Genome Wide SNP Data From an Israeli-Arab Community. J. Alzheimers Dis. *23*, 349–359.
47. Reitz, C., Jun, G., Naj, A., Rajbhandary, R., Vardarajan, B.N., Wang, L.-S., Valladares, O., Lin, C.-F., Larson, E.B., Graff-Radford, N.R., et al.; Alzheimer Disease Genetics Consortium (2013). Variants in the ATP-binding cassette transporter (ABCA7), apolipoprotein E ε4,and the risk of late-onset Alzheimer disease in African Americans. JAMA *309*, 1483–1492.
48. Kunkle, B.W., Grenier-Boley, B., Sims, R., Bis, J.C., Damotte, V., Naj, A.C., Boland, A., Vronskaya, M., van der Lee, S.J., Amlie-Wolf, A., et al.; Alzheimer Disease Genetics Consortium (ADGC); European Alzheimer's Disease Initiative (EADI); Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium (CHARGE); and Genetic and Environmental Risk in AD/Defining Genetic, Polygenic and Environmental Risk for Alzheimer's Disease Consortium (GERAD/PERADES) (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat. Genet. *51*, 414–430.
49. Park, Y., Sarkar, A., He, L., Davila-Velderrain, J., De, P.J., and Kellis, M. (2017). A Bayesian approach to mediation analysis predicts 206 causal target genes in Alzheimer's disease. bioRxiv. https://doi.org/10.1101/219428.
50. Keller, J.N., Hanni, K.B., and Markesbery, W.R. (2000). Impaired proteasome function in Alzheimer's disease. J. Neurochem. *75*, 436–439.
51. Chen, J., Lee, G., Fanous, A.H., Zhao, Z., Jia, P., O'Neill, A., Walsh, D., Kendler, K.S., Chen, X.; and International Schizophrenia Consortium (2011). Two non-synonymous markers in PTPN21, identified by genome-wide association study data-mining and replication, are associated with schizophrenia. Schizophr. Res. *131*, 43–51.
52. Morawe, T., Hiebel, C., Kern, A., and Behl, C. (2012). Protein homeostasis, aging and Alzheimer's disease. Mol. Neurobiol. *46*, 41–54.
53. Plani-Lam, J.H.-C., Chow, T.-C., Siu, K.-L., Chau, W.H., Ng, M.-H.J., Bao, S., Ng, C.T., Sham, P., Shum, D.K.-Y., Ingley, E., et al. (2015). PTPN21 exerts pro-neuronal survival and neuritic elongation via ErbB4/NRG3 signaling. Int. J. Biochem. Cell Biol. *61*, 53–62.
54. Hollingworth, P., Harold, D., Sims, R., Gerrish, A., Lambert, J.-C., Carrasquillo, M.M., Abraham, R., Hamshere, M.L., Pahwa, J.S., Moskvina, V., et al.; Alzheimer's Disease Neuroimaging Initiative; CHARGE consortium; and EADI1 consortium (2011). Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. Nat. Genet. *43*, 429–435.
55. Naj, A.C., Jun, G., Beecham, G.W., Wang, L.-S., Vardarajan, B.N., Buros, J., Gallins, P.J., Buxbaum, J.D., Jarvik, G.P., Crane, P.K., et al. (2011). Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. Nat. Genet. *43*, 436–441.
56. Lambert, J.C., Ibrahim-Verbaas, C.A., Harold, D., Naj, A.C., Sims, R., Bellenguez, C., DeStafano, A.L., Bis, J.C., Beecham, G.W., Grenier-Boley, B., et al.; European Alzheimer's Disease Initiative (EADI); Genetic and Environmental Risk in Alzheimer's Disease; Alzheimer's Disease Genetic Consortium; and Cohorts for Heart and Aging Research in Genomic Epidemiology (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat. Genet. *45*, 1452–1458.
57. Vardarajan, B.N., Zhang, Y., Lee, J.H., Cheng, R., Bohm, C., Ghani, M., Reitz, C., Reyes-Dumeyer, D., Shen, Y., Rogaeva, E., et al. (2015). Coding mutations in SORL1 and Alzheimer disease. Ann. Neurol. *77*, 215–227.
58. Abecasis, G.R., Cherny, S.S., Cookson, W.O., and Cardon, L.R. (2002). Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat. Genet. *30*, 97–101.
59. Abecasis, G.R., and Wigginton, J.E. (2005). Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am. J. Hum. Genet. *77*, 754–767.
60. 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.
61. Brenner, C.H. (2010). Fundamental problem of forensic mathematics–the evidential value of a rare haplotype. Forensic Sci. Int. Genet. *4*, 281–291.
62. Whittemore, A.S., and Halpern, J. (1994). A class of tests for linkage using affected pedigree members. Biometrics *50*, 118–127.
63. Kong, A., and Cox, N.J. (1997). Allele-sharing models: LOD scores and accurate linkage tests. Am. J. Hum. Genet. *61*, 1179–1188.
64. Matise, T.C., Chen, F., Chen, W., De La Vega, F.M., Hansen, M., He, C., Hyland, F.C., Kennedy, G.C., Kong, X., Murray, S.S., et al. (2007). A second-generation combined linkage physical map of the human genome. Genome Res. *17*, 1783–1786.
65. 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.
66. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164–e164.
67. Goldgar, D.E. (2001). Major strengths and weaknesses of model-free methods. Adv. Genet. *42*, 241–251.
68. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J., and Lander, E.S. (2001). High-resolution haplotype structure in the human genome. Nat. Genet. *29*, 229–232.
69. Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F.,

Ward, R., and Lander, E.S. (2001). Linkage disequilibrium in the human genome. Nature *411*, 199–204.

70. McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. *9*, 356–369.
71. Steinberg, S., Stefansson, H., Jonsson, T., Johannsdottir, H., Ingason, A., Helgason, H., Sulem, P., Magnusson, O.T., Gudjonsson, S.A., Unnsteinsdottir, U., et al.; DemGene (2015). Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease. Nat. Genet. *47*, 445–447.
72. Vardarajan, B.N., Ghani, M., Kahn, A., Sheikh, S., Sato, C., Barral, S., Lee, J.H., Cheng, R., Reitz, C., Lantigua, R., et al. (2015). Rare coding mutations identified by sequencing of Alzheimer disease genome-wide association studies loci. Ann. Neurol. *78*, 487–498.
73. Cuyvers, E., De Roeck, A., Van den Bossche, T., Van Cauwenberghe, C., Bettens, K., Vermeulen, S., Mattheijssens, M., Peeters, K., Engelborghs, S., Vandenbulcke, M., et al. (2015). Mutations in ABCA7 in a Belgian cohort of Alzheimer's disease patients: a targeted resequencing study. Lancet Neurol. *14*, 814–822.
74. Le Guennec, K., Nicolas, G., Quenez, O., Charbonnier, C., Wallon, D., Bellenguez, C., Grenier-Boley, B., Rousseau, S., Richard, A.-C., Rovelet-Lecrux, A., et al.; CNR-MAJ collaborators (2016). ABCA7 rare variants and Alzheimer disease risk. Neurology *86*, 2134–2137.
75. Meng, Y., Baldwin, C.T., Bowirrat, A., Waraska, K., Inzelberg, R., Friedland, R.P., and Farrer, L.A. (2006). Association of polymorphisms in the Angiotensin-converting enzyme gene with Alzheimer disease in an Israeli Arab community. Am. J. Hum. Genet. *78*, 871–877.
76. Jochemsen, H.M., Teunissen, C.E., Ashby, E.L., van der Flier, W.M., Jones, R.E., Geerlings, M.I., Scheltens, P., Kehoe, P.G., and Muller, M. (2014). The association of angiotensin-converting enzyme with biomarkers for Alzheimer's disease. Alzheimers Res. Ther. *6*, 27.
77. Miyashita, A., Koike, A., Jun, G., Wang, L.-S., Takahashi, S., Matsubara, E., Kawarabayashi, T., Shoji, M., Tomita, N., Arai, H., et al.; Alzheimer Disease Genetics Consortium (2013). SORL1 is genetically associated with late-onset Alzheimer's disease in Japanese, Koreans and Caucasians. PLoS ONE *8*, e58618.
78. Pottier, C., Hannequin, D., Coutant, S., Rovelet-Lecrux, A., Wallon, D., Rousseau, S., Legallic, S., Paquet, C., Bombois, S., Pariente, J., et al.; PHRC GMAJ Collaborators (2012). High frequency of potentially pathogenic *SORL1* mutations in autosomal dominant early-onset Alzheimer disease. Mol. Psychiatry *17*, 875–879.
79. Rogaeva, E., Meng, Y., Lee, J.H., Gu, Y., Kawarai, T., Zou, F., Katayama, T., Baldwin, C.T., Cheng, R., Hasegawa, H., et al. (2007). The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. Nat. Genet. *39*, 168–177.

The American Journal of Human Genetics, Volume *105*

Supplemental Data

A Rare Variant Nonparametric Linkage Method for Nuclear and Extended Pedigrees with Application to Late-Onset Alzheimer Disease via WGS Data

Linhai Zhao, Zongxiao He, Di Zhang, Gao T. Wang, Alan E. Renton, Badri N. Vardarajan, Michael Nothnagel, Alison M. Goate, Richard Mayeux, and Suzanne M. Leal

Supplemental Material

Supplemental Figures and Tables

Figure S1. Pedigree structures used for evaluation of intra-familial locus heterogeneity

Pedigrees with intra-familial locus heterogeneity were simulated by generating genotypes on extended families with two branches with three of the four children in the last generation being affected (panel A) and analyzing pedigrees with all children affected (panel B), to mimic intra intra-familial locus heterogeneity

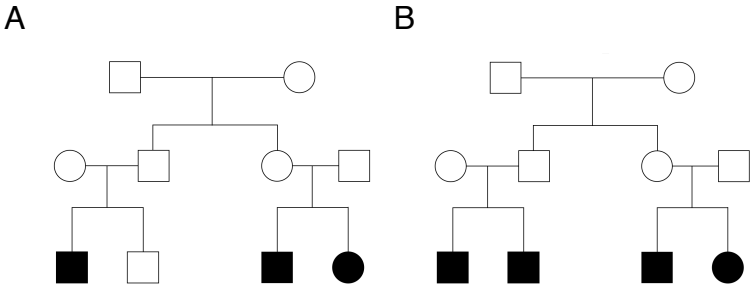
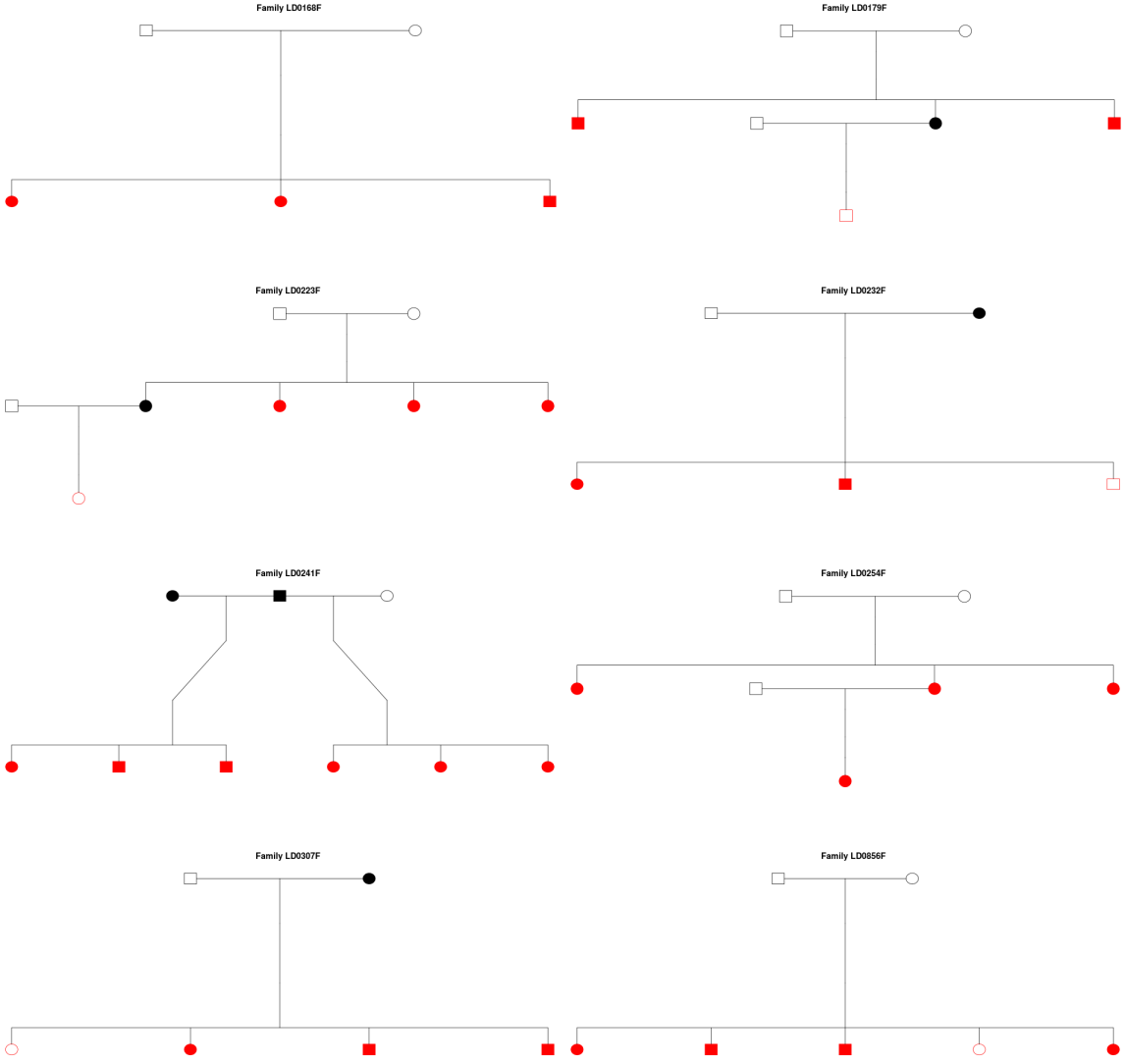


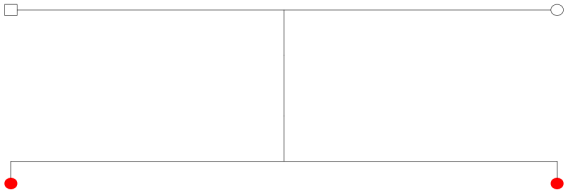
Figure S2. Pedigrees included in the analysis from the Alzheimer's Disease Sequencing Project

The 107 pedigrees (42 European pedigrees and 65 Hispanic pedigrees) which were analyzed. Squares represent males and circles females. Filled symbols are individuals affected Alzheimer's disease and open symbols represent unaffected family members. Those individuals shown in red have whole genome sequence data available, while those in black do not have available genotype data.

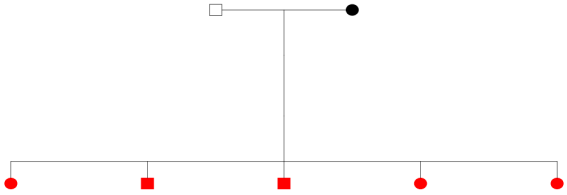
1. European pedigrees



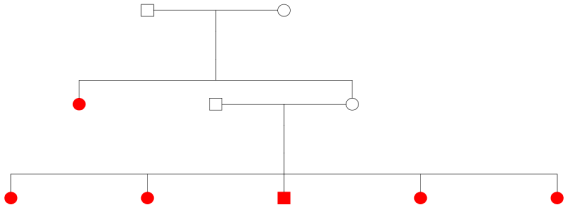
Family LD0949F



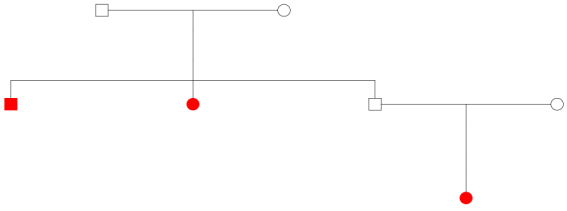
Family LD1012F



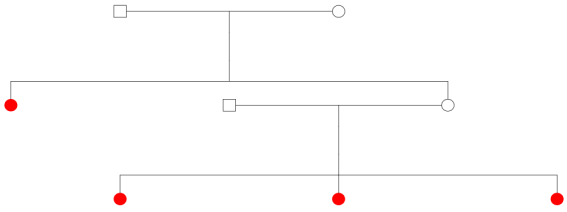
Family LD1223F



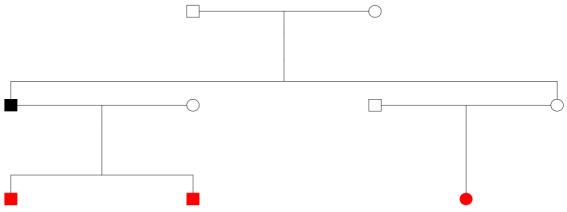
Family LD1260F



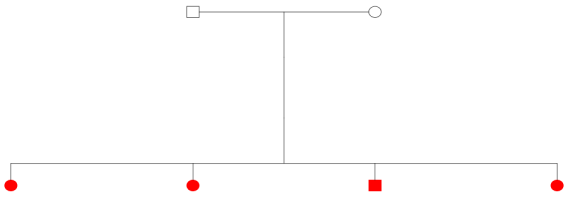
Family LD1265F



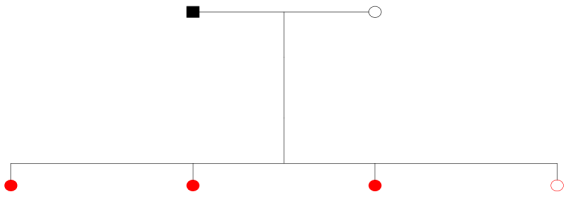
Family LD1315F



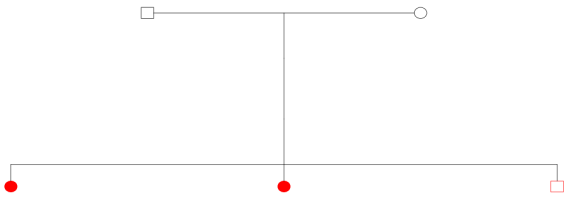
Family LD1329F



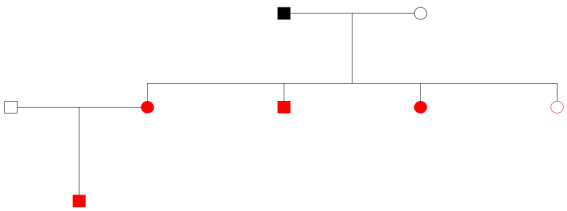
Family LD1579F



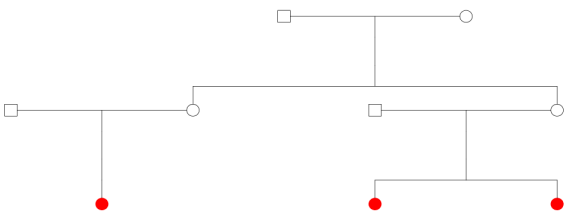
Family NC0049F



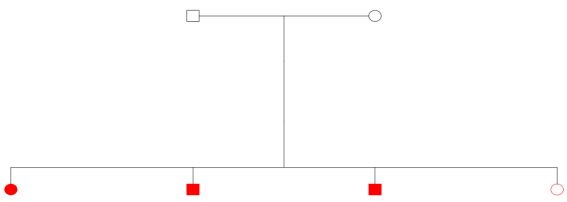
Family NC0131F

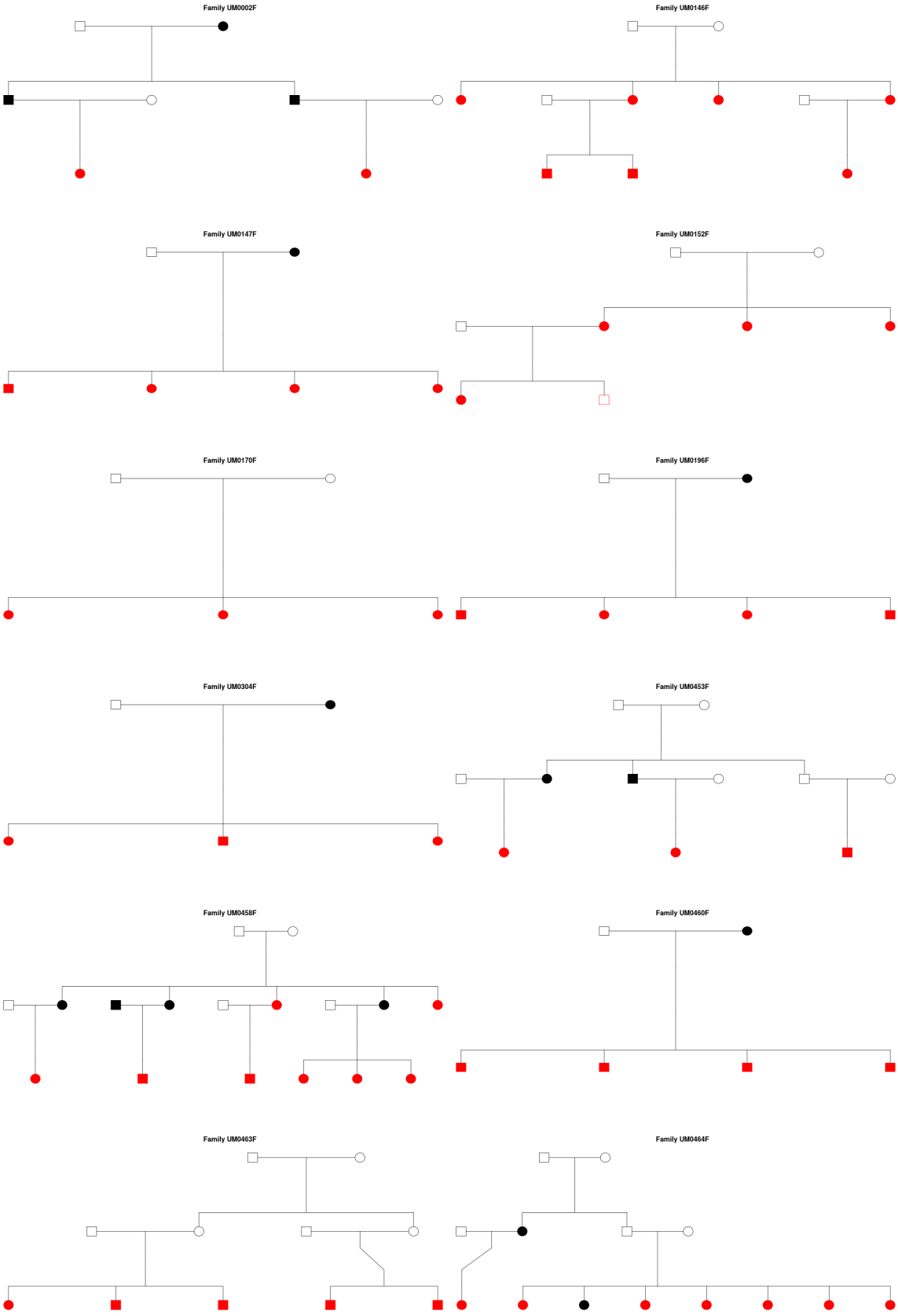


Family NC0205F

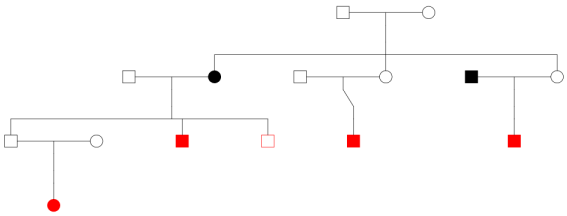


Family NC0302F

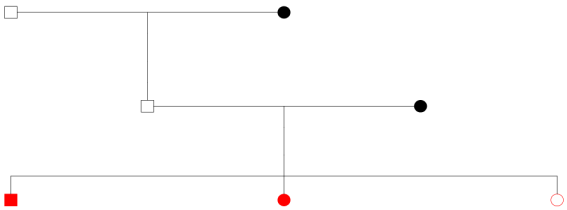



Family UM0002F
Family UM0146F
Family UM0147F
Family UM0152F
Family UM0170F
Family UM0196F
Family UM0304F
Family UM0453F
Family UM0458F
Family UM0460F
Family UM0463F
Family UM0464F

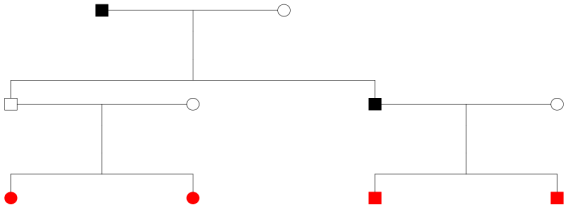
Family UP0001F



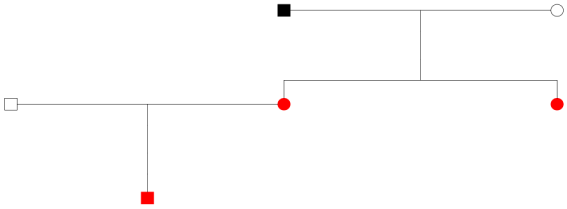
Family UP0002F



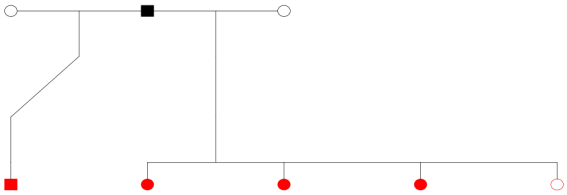
Family UP0002F-2



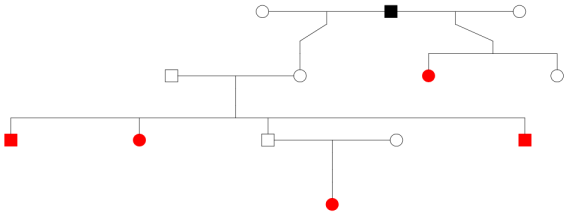
Family UP0003F



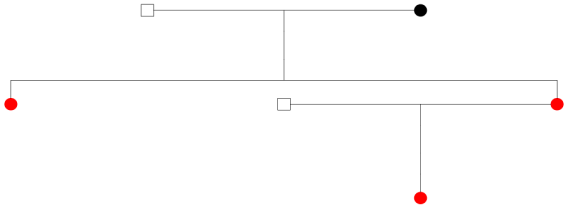
Family UP0004F



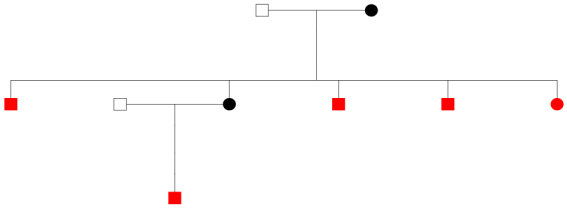
Family UP0005F



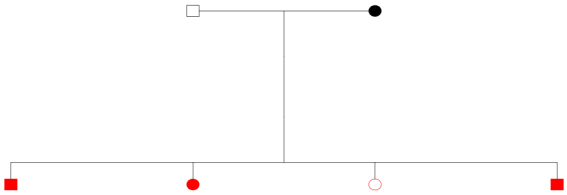
Family UP0006F



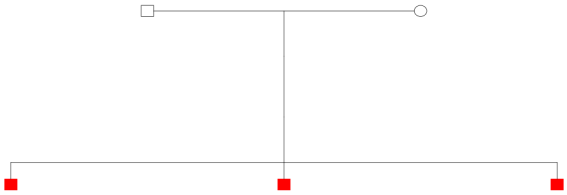
Family UP0007F

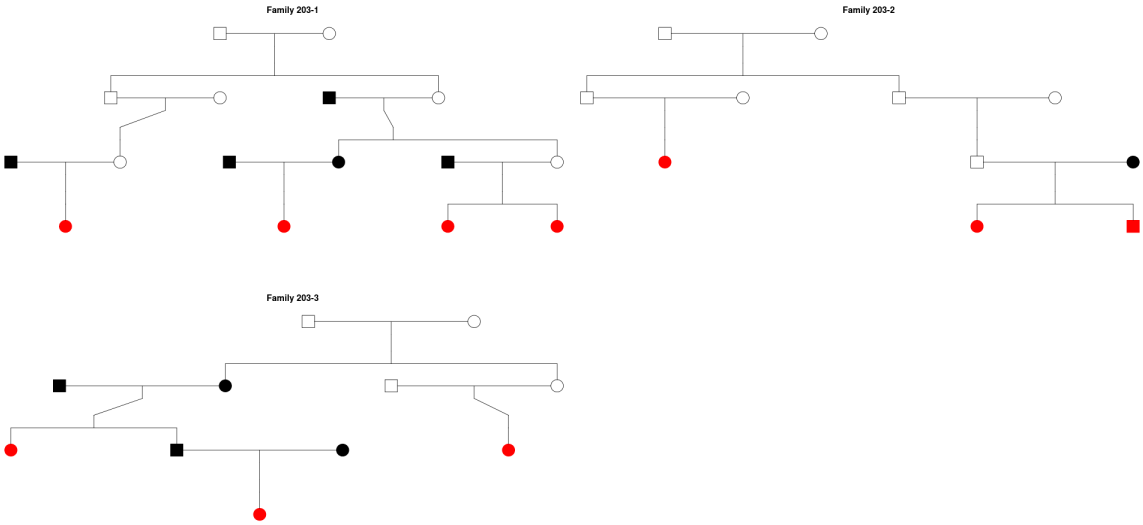


Family UP0008F

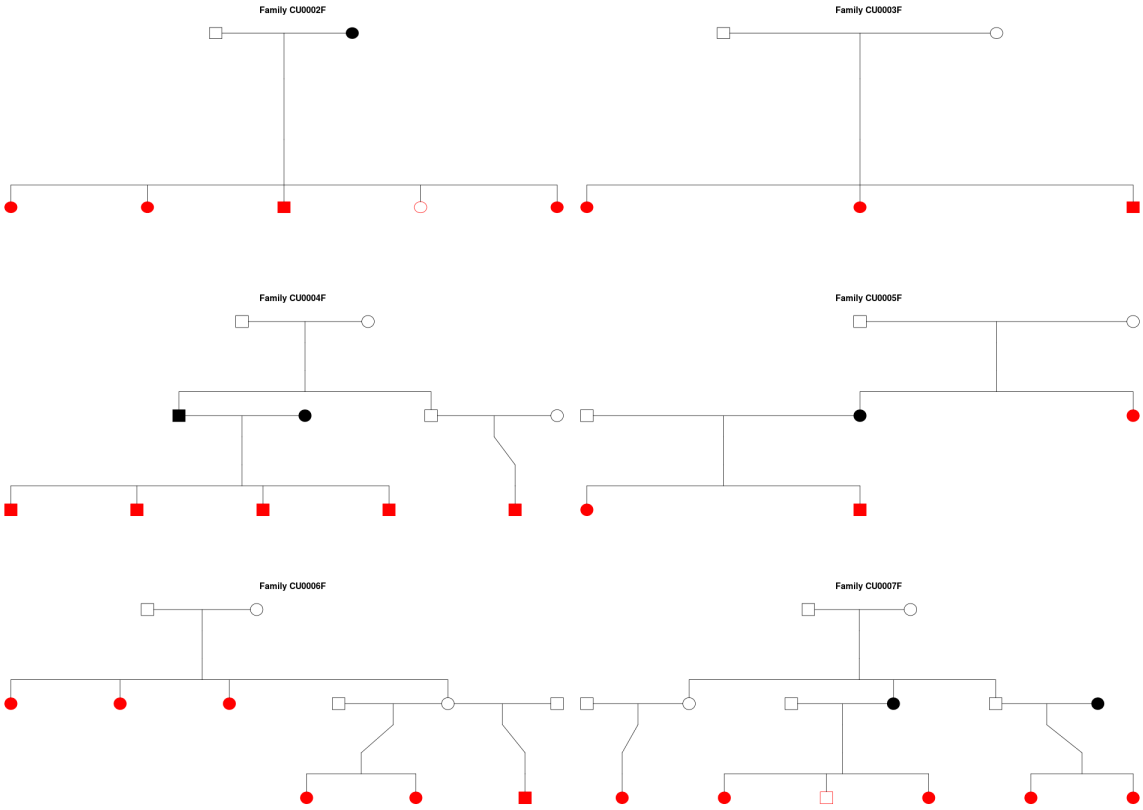


Family VU0072F

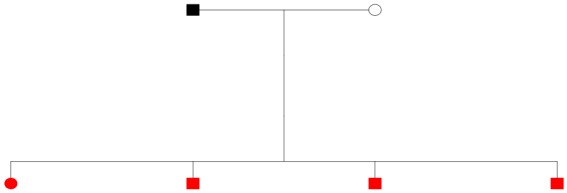




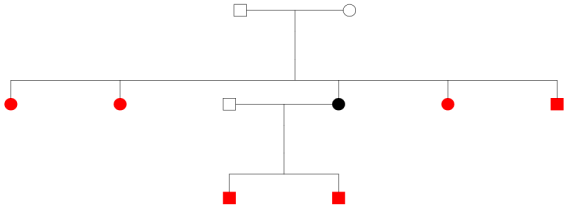
2. Hispanic Pedigrees



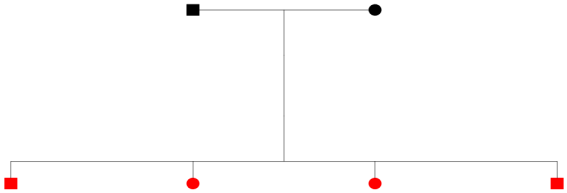
Family CU0008F



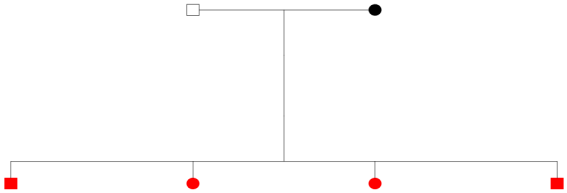
Family CU0009F



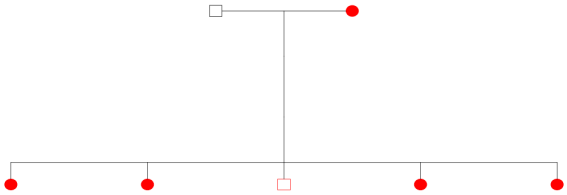
Family CU0010F



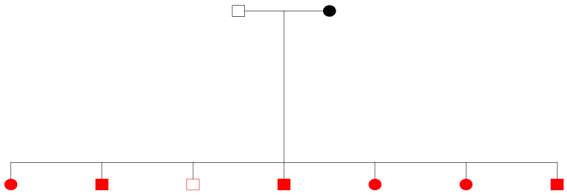
Family CU0012F



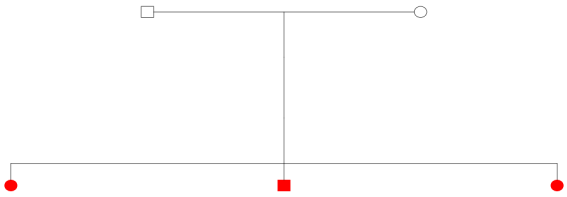
Family CU0013F



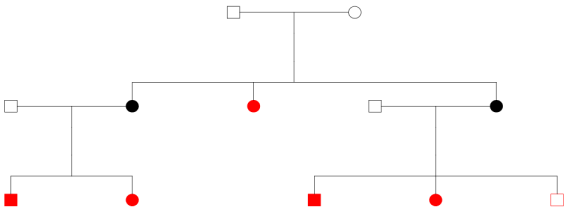
Family CU0014F



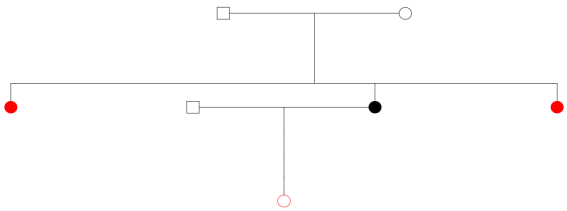
Family CU0015F



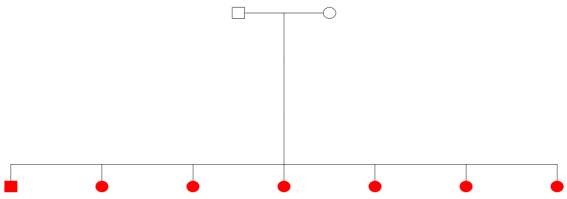
Family CU0016F



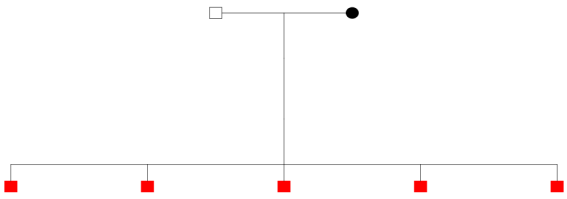
Family CU0017F



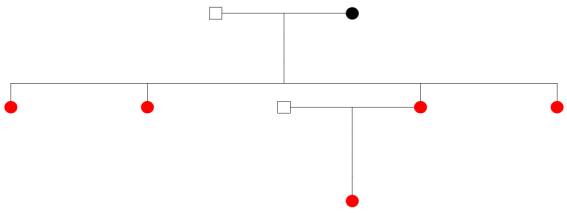
Family CU0018F



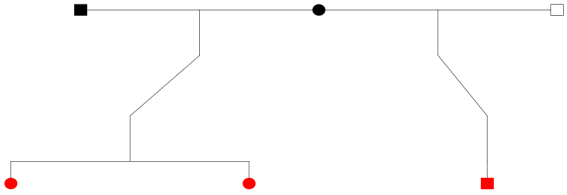
Family CU0019F



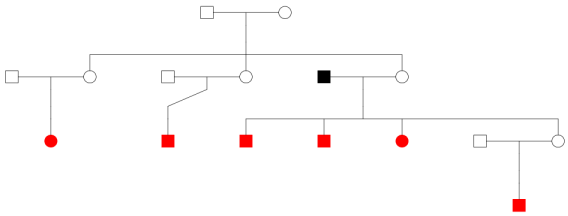
Family CU0020F



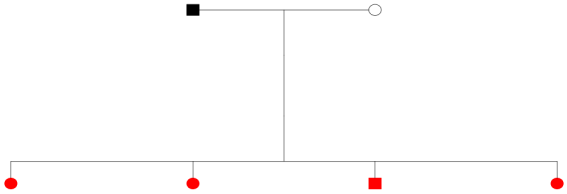
Family CU0022F



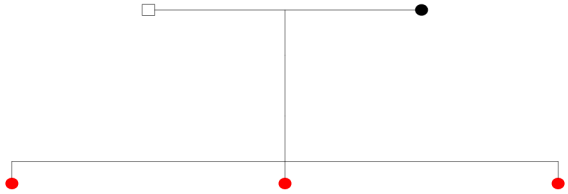
Family CU0023F



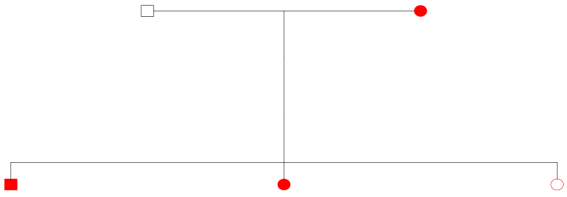
Family CU0024F



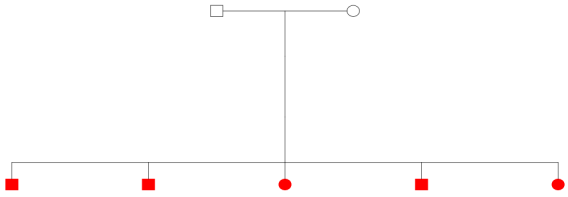
Family CU0025F



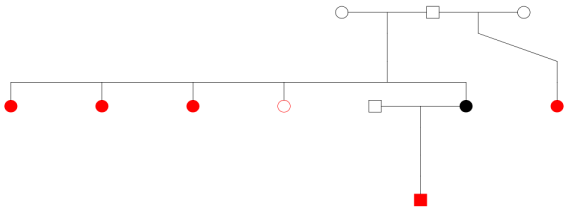
Family CU0026F



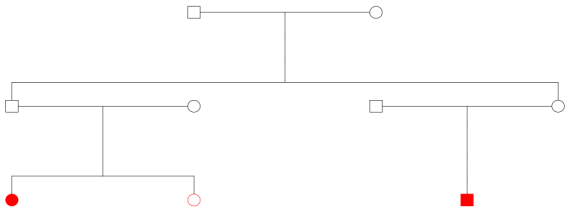
Family CU0029F



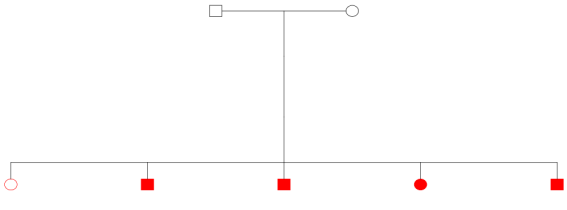
Family CU0030F



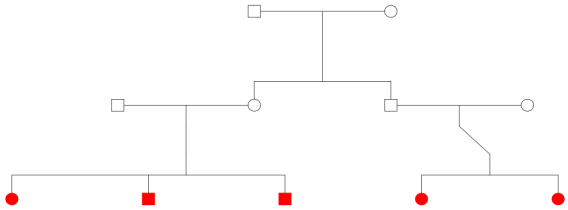
Family CU0033F



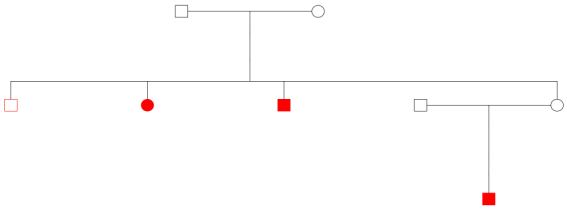
Family CU0035F



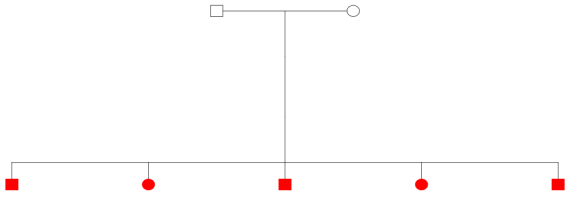
Family CU0036F



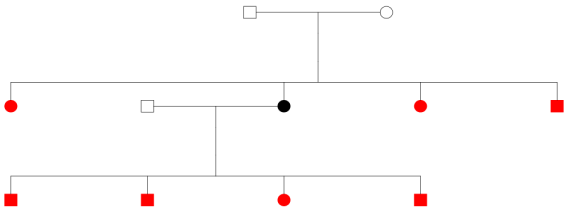
Family CU0037F



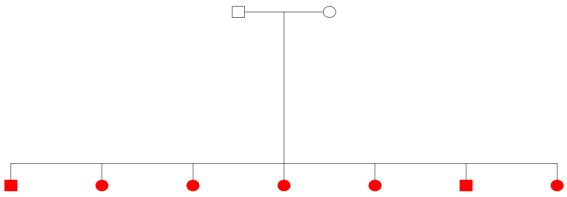
Family CU0038F



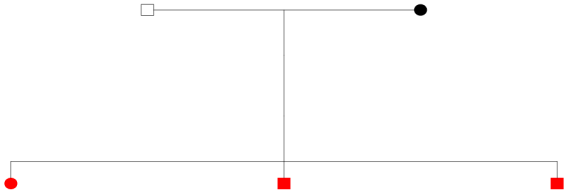
Family CU0039F



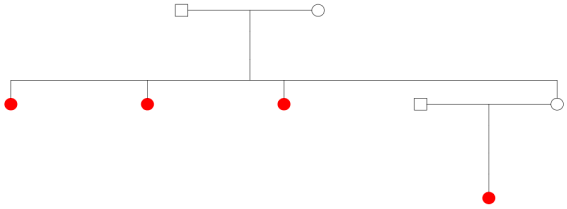
Family CU0040F



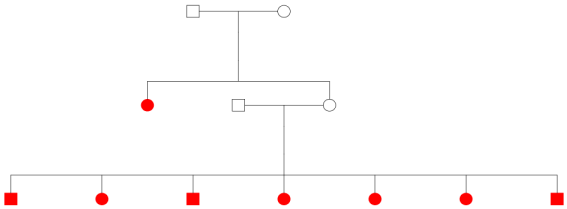
Family CU0041F



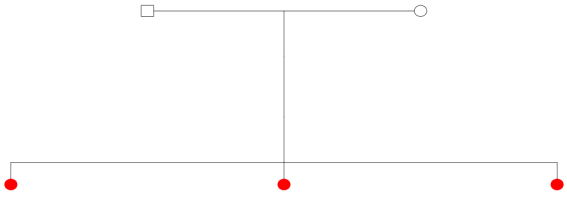
Family CU0043F



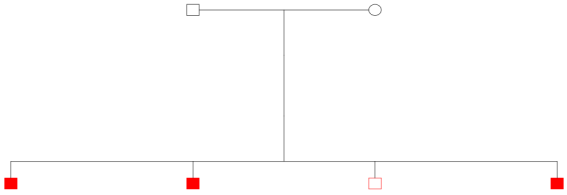
Family CU0044F



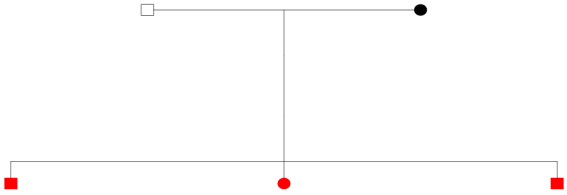
Family CU0044F-2



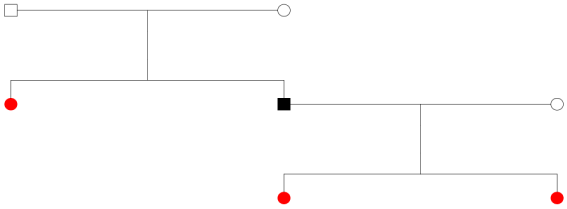
Family CU0045F



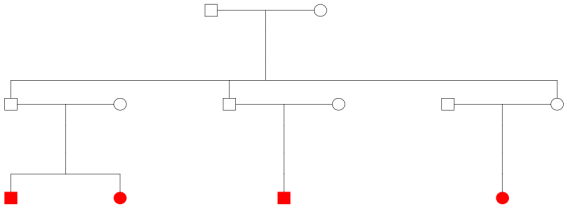
Family CU0046F



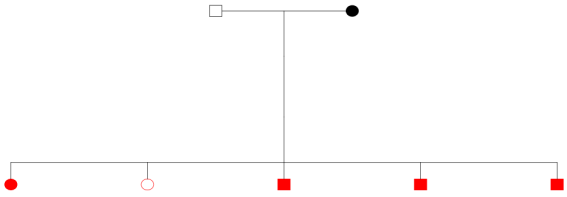
Family CU0047F



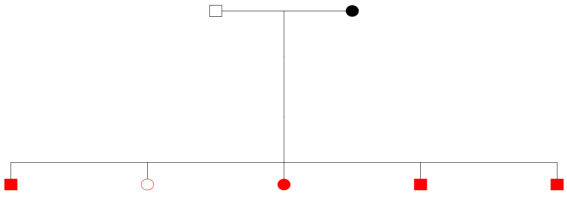
Family CU0048F



Family CU0049F



Family CU0050F



Family CU0052F

Family CU0053F

Family CU0055F

Family CU0057F

Family CU0058F

Family CU0059F

Family CU0060F

Family CU0064F

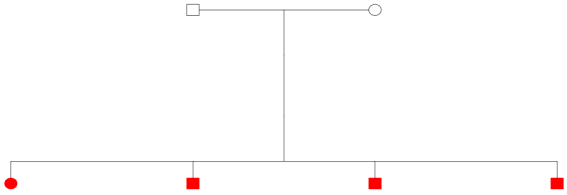
Family CU0065F

Family CU0067F

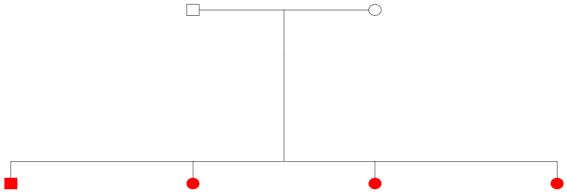
Family CU0068F

Family CU0070F

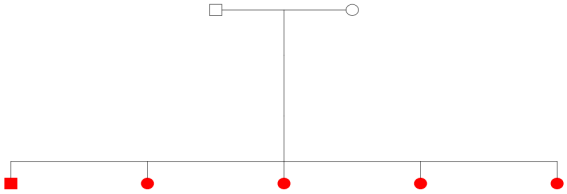
Family CU0071F



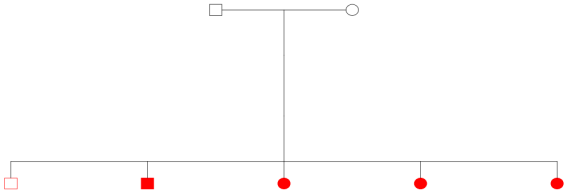
Family CU0073F



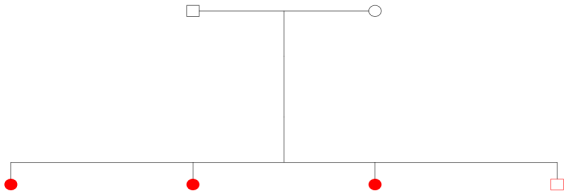
Family CU0075F



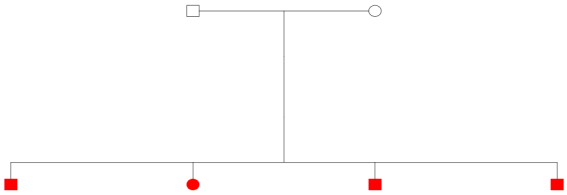
Family CU0076F



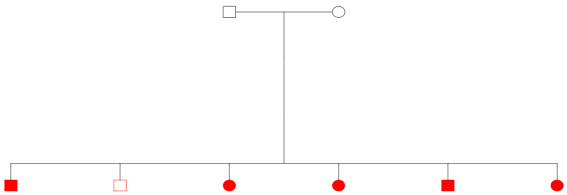
Family CU0078F



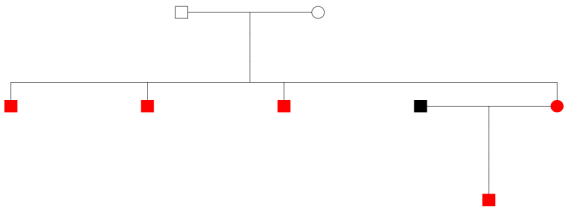
Family CU0079F



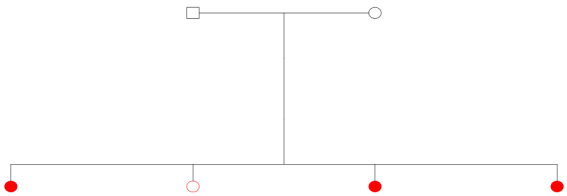
Family CU0081F



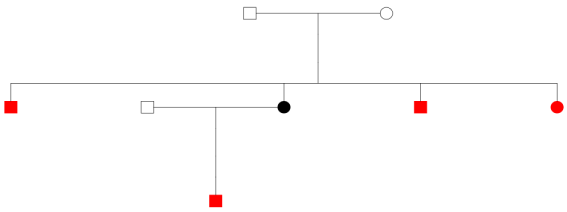
Family CU0082F



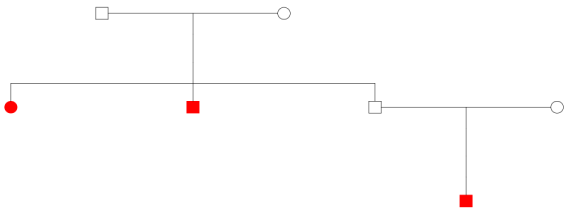
Family CU0083F



Family CU0032F



Family CU0042F



Family CU0051F

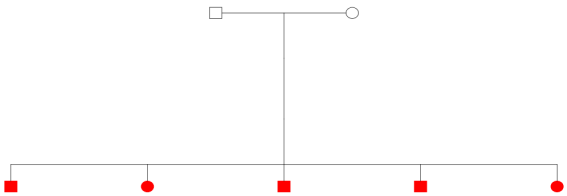


Figure S3. QQ plots for CHP-NPL$_{Pairs}$ under the null hypothesis of no linkage

One-thousand replicates of exomes were generated under the null hypothesis of no linkage for 2,000 affected sib-pairs without missing genotype date (panel A); 2,000 affected sib-pairs with founders missing all exome data (panel B); 300 nuclear families with three affected siblings without missing genotype data (panel C); 300 nuclear families with three affected siblings with founders missing all exome data (panel D); 100 extended families without missing genotype data (panel E); and 100 extended families with founders missing all exome data (panel F); and analyzed using CHP-NPL$_{Pairs}$ obtaining analytical p-values.

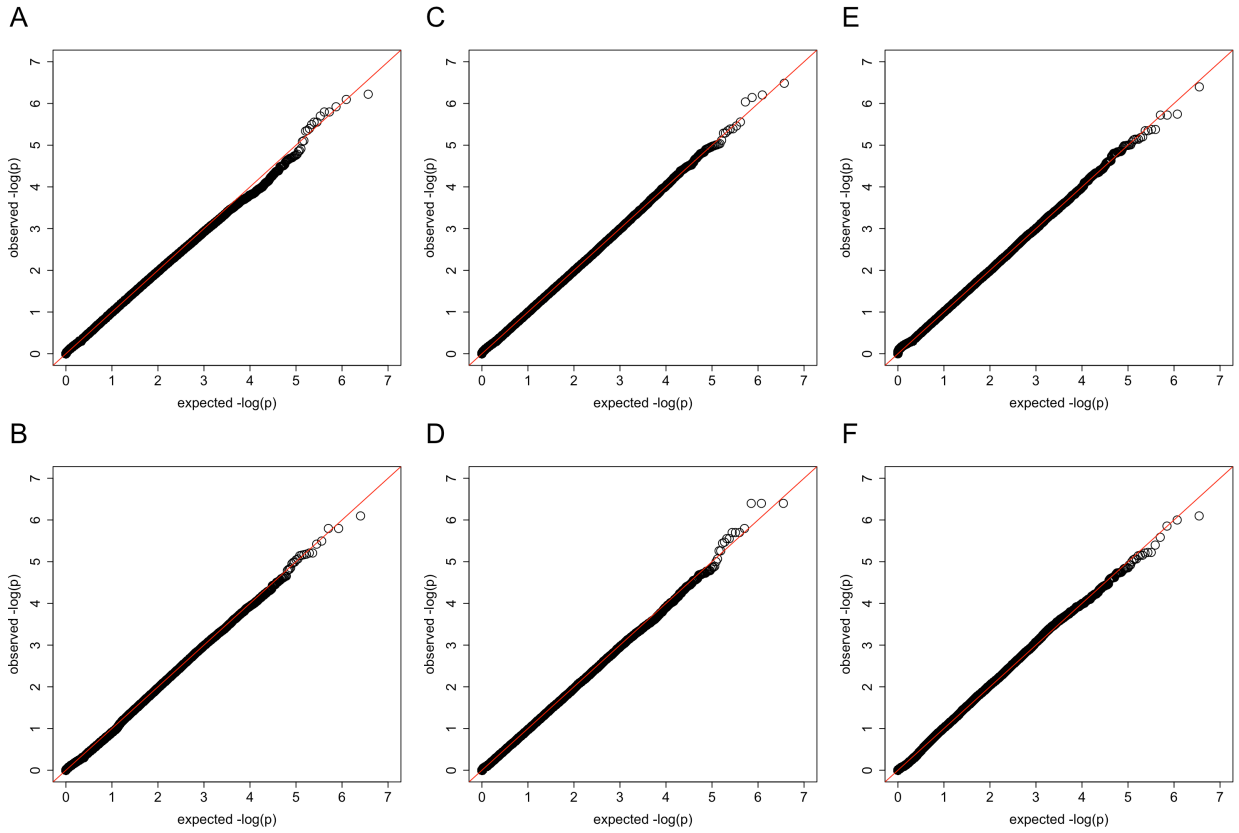


Figure S4. QQ plots for CHP-NPL$_{All}$ under the null hypothesis of no linkage

One-thousand replicates of exomes were generated under the null hypothesis of no linkage for 2,000 affected sib-pairs without missing genotype date (panel A); 2,000 affected sib-pairs with founders missing all exome data (panel B); 300 nuclear families with three affected siblings without missing genotype data (panel C); 300 nuclear families with three affected siblings with founders missing all exome data (panel D); 100 extended families without missing genotype data (panel E); and 100 extended families with founders missing all exome data (panel F); and analyzed using CHP-NPL$_{All}$ obtaining analytical p-values.

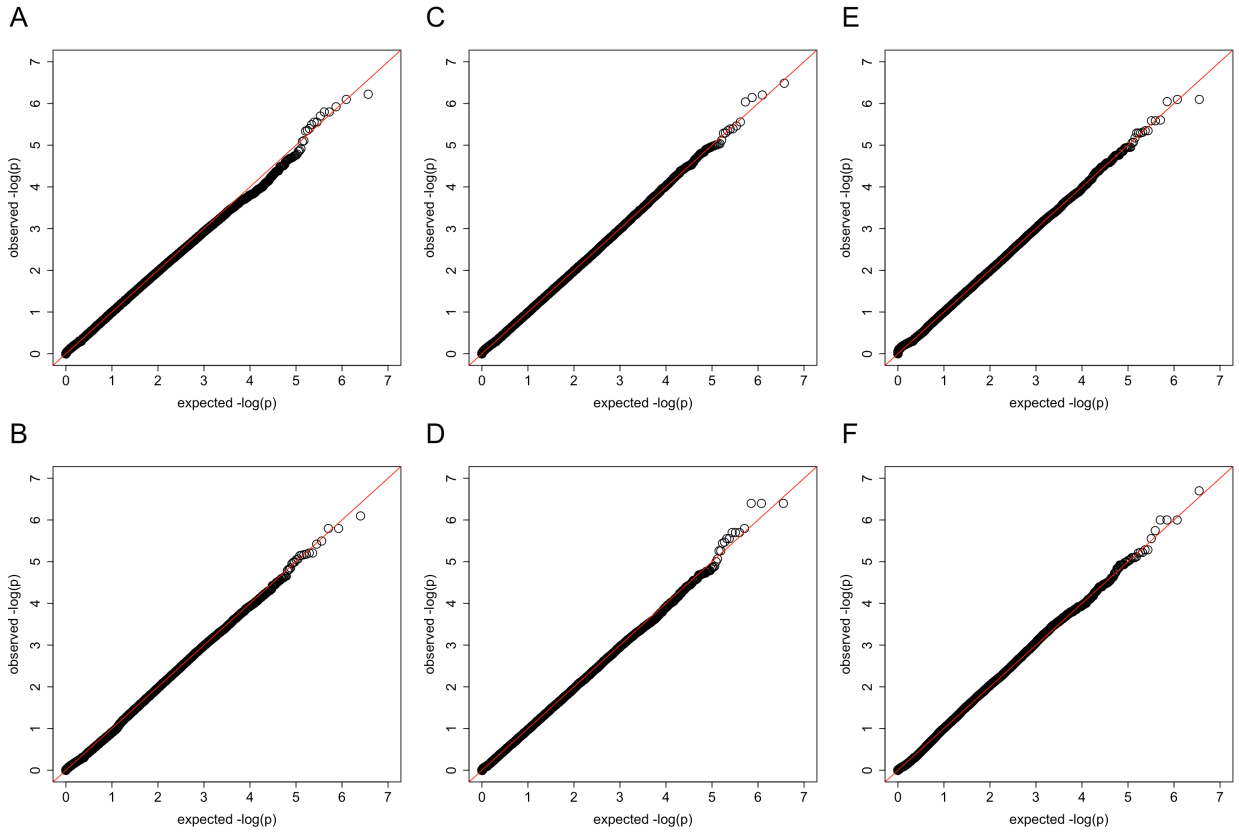


Figure S5. QQ plots for RV-NPL$_{Pairs}$ under the null hypothesis of no linkage

One-thousand replicates of exomes were generated under the null hypothesis of no linkage for 2,000 affected sib-pairs without missing genotype date (panel A); 2,000 affected sib-pairs with founders missing all exome data (panel B); 300 nuclear families with three affected siblings without missing genotype data (panel C); 300 nuclear families with three affected siblings with founders missing all exome data (panel D); 100 extended families without missing genotype data (panel E); and 100 extended families with founders missing all exome data (panel F); and analyzed using RV-NPL$_{Pairs}$ obtaining empirical p-values using 1,000,000 permutations. The observed plateau is due to the number of permutations performed.

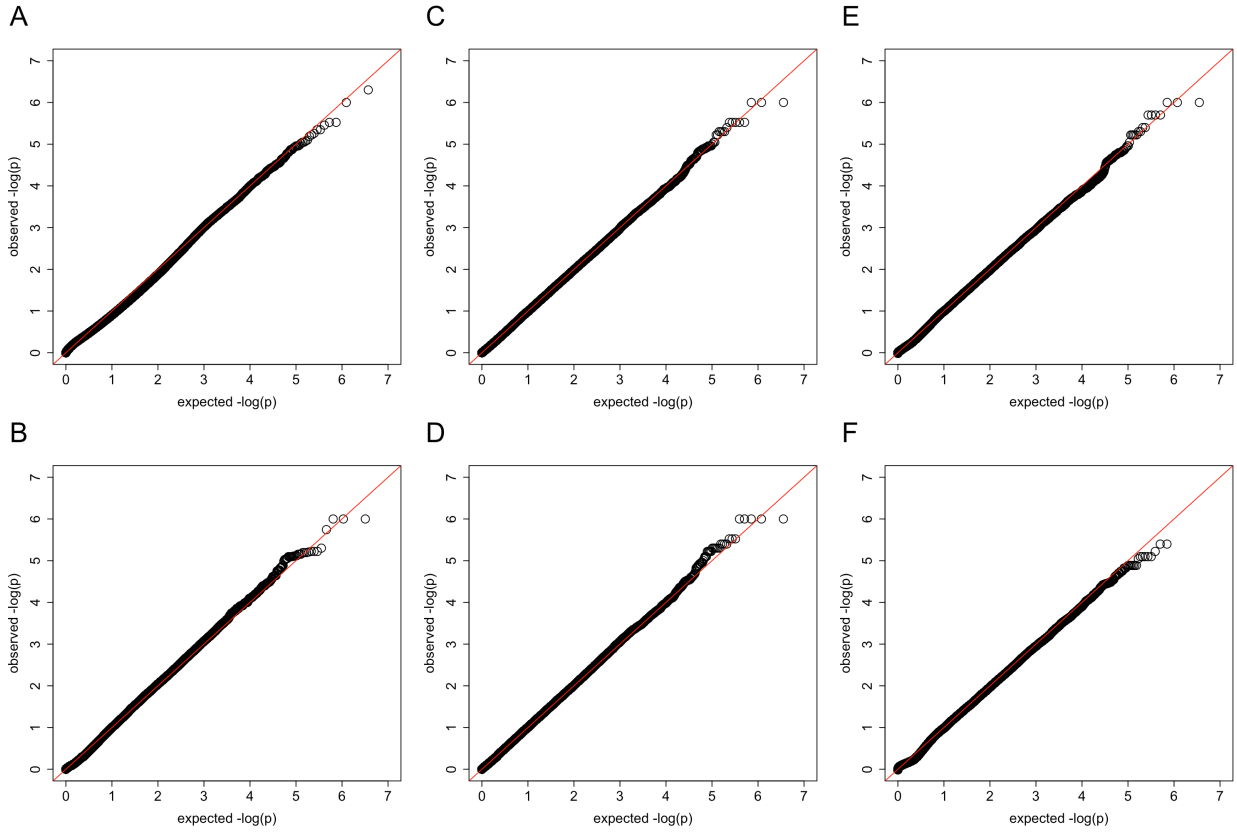


Figure S6. QQ plots for RV-NPL$_{All}$ under the null hypothesis of no linkage

One-thousand replicates of exomes were generated under the null hypothesis of no linkage for 2,000 affected sib-pairs without missing genotype date (panel A); 2,000 affected sib-pairs with founders missing all exome data (panel B); 300 nuclear families with three affected siblings without missing genotype data (panel C); 300 nuclear families with three affected siblings with founders missing all exome data (panel D); 100 extended families without missing genotype data (panel E); and 100 extended families with founders missing all exome data (panel F); and analyzed using RV-NPL$_{All}$ obtaining empirical p-values using 1,000,000 permutations. The observed plateau is due to the number of permutations performed.

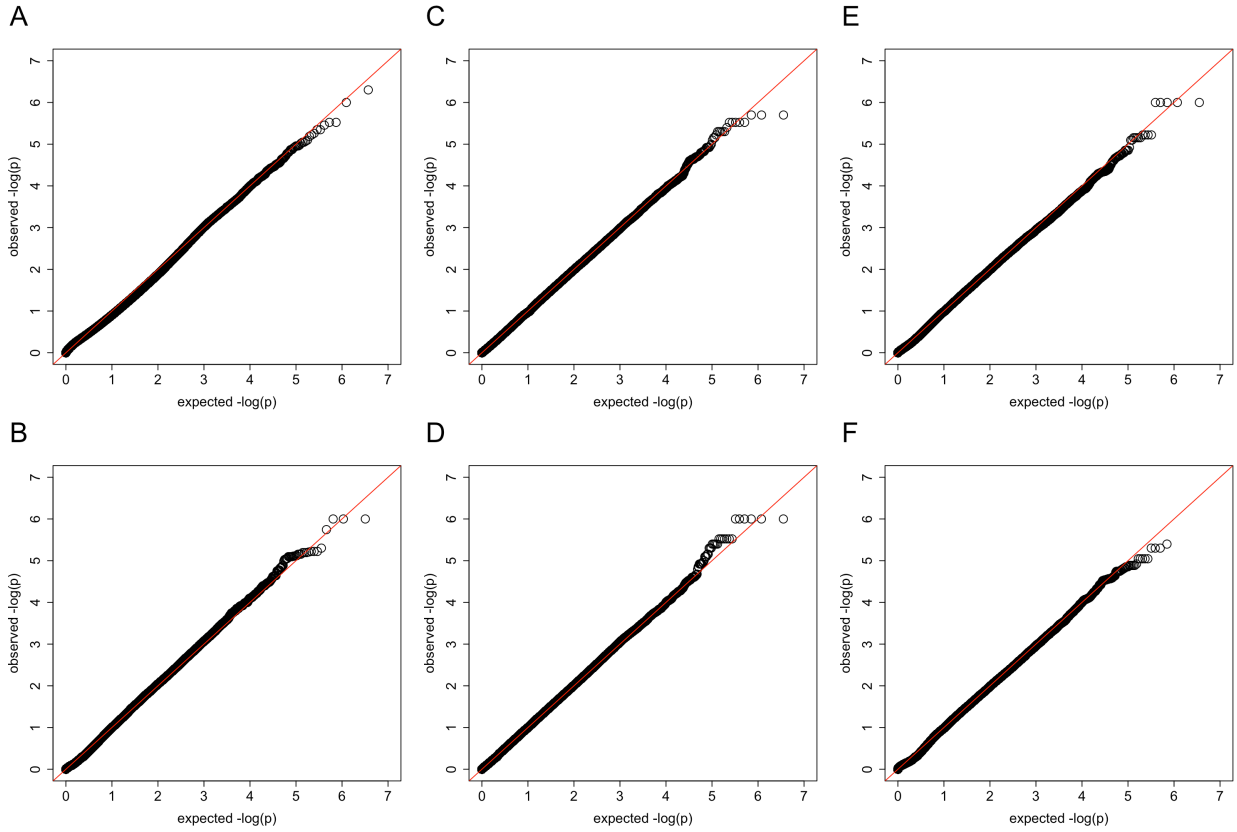


Figure S7. Power comparison for NPL_{Pairs} for nuclear families with three affected siblings

Genotypes were generated for 300 nuclear families with three affected siblings conditional on affection status assuming a multiplicative model in which each causal variant within a gene region has an OR of 5.0. Analysis was performed using RV-NPL_{Pairs}, CHP-NPL_{Pairs}, and Multipoint-NPL_{Pairs}: with 100%, 75% and 50% of the variant being causal and the remaining non-causal (OR=1.0) (panel A); with only causal nonsynonymous (NS) variants as well as with causal nonsynonymous (NS) and non-causal synonymous (S) variants (panel B); with 0%, 10%, 30%, and 50% of the founders missing all genotype data (panel C); and with no heterogeneity (NH), i.e. 300 linked families as well as with locus heterogeneity (H), i.e., 300 linked and 150 unlinked families (panel D).

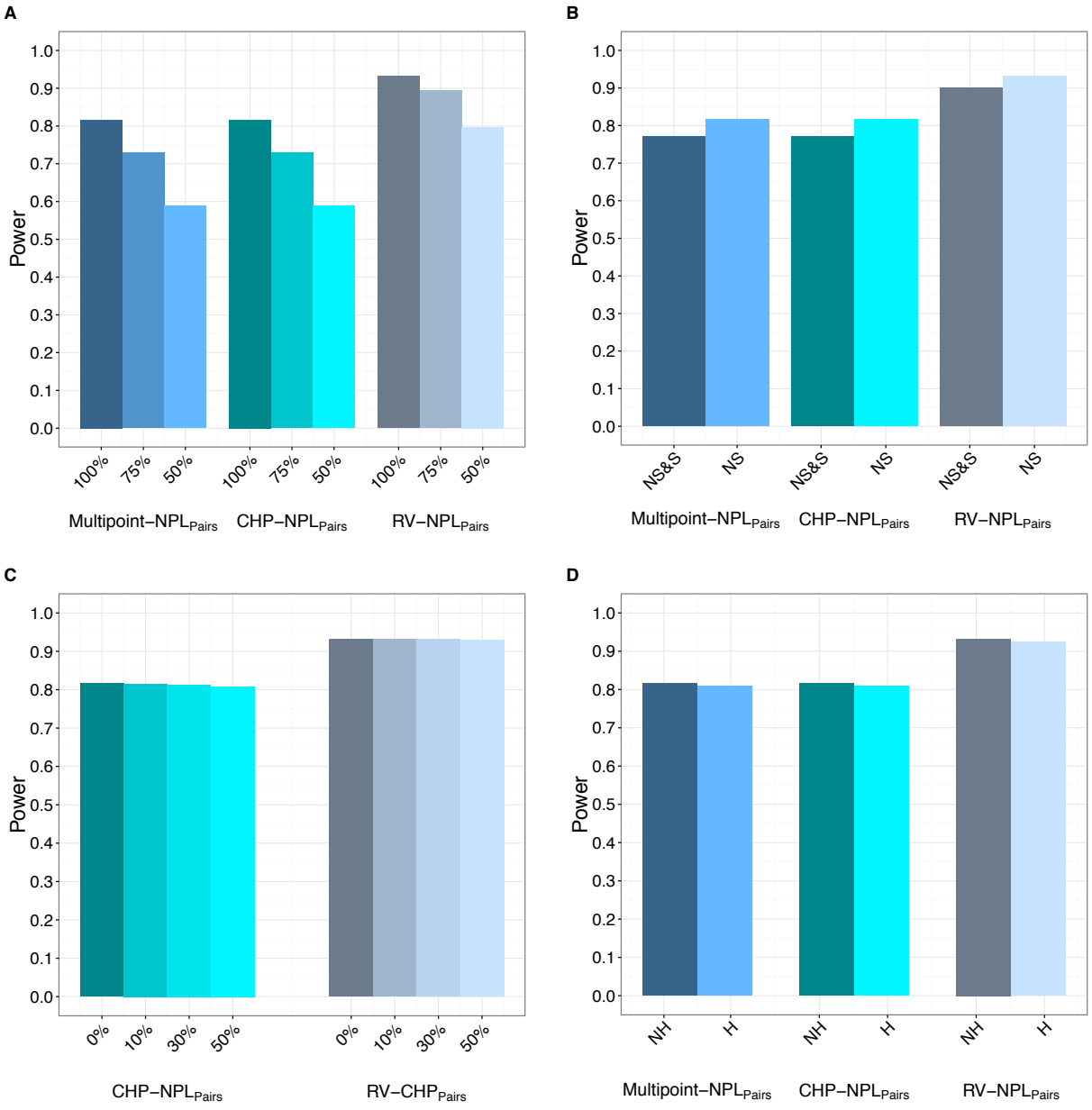

A
Power
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
100%
75%
50%
100%
75%
50%
100%
75%
50%
Multipoint–NPL$_{Pairs}$
CHP–NPL$_{Pairs}$
RV–NPL$_{Pairs}$
B
Power
NS&S
NS
NS&S
NS
NS&S
NS
Multipoint–NPL$_{Pairs}$
CHP–NPL$_{Pairs}$
RV–NPL$_{Pairs}$
C
Power
0%
10%
30%
50%
0%
10%
30%
50%
CHP–NPL$_{Pairs}$
RV–CHP$_{Pairs}$
D
Power
NH
H
NH
H
NH
H
Multipoint–NPL$_{Pairs}$
CHP–NPL$_{Pairs}$
RV–NPL$_{Pairs}$

Figure S8. Power comparison for NPL_{Pairs} on affected sibpairs

Genotypes were generated for 2,000 nuclear families with affected sibpairs conditional on affection status assuming a multiplicative model in which each causal variant within a gene region has an OR of 5.0. Analysis was performed using RV-NPL_{Pairs}, CHP-NPL_{Pairs}, and Multipoint-NPL_{Pairs}: with 100%, 75% and 50% of the variant being causal and the remaining non-causal (OR=1.0) (panel A); with only causal nonsynonymous (NS) variants as well as with causal nonsynonymous (NS) and non-causal synonymous (S) variants (panel B); with 0%, 10%, 30%, and 50% of the founders missing all genotype data (panel C); and with no heterogeneity (NH), i.e. 2,000 linked families as well as with locus heterogeneity (H), i.e., 2,000 linked and 1,000 unlinked families (panel D).

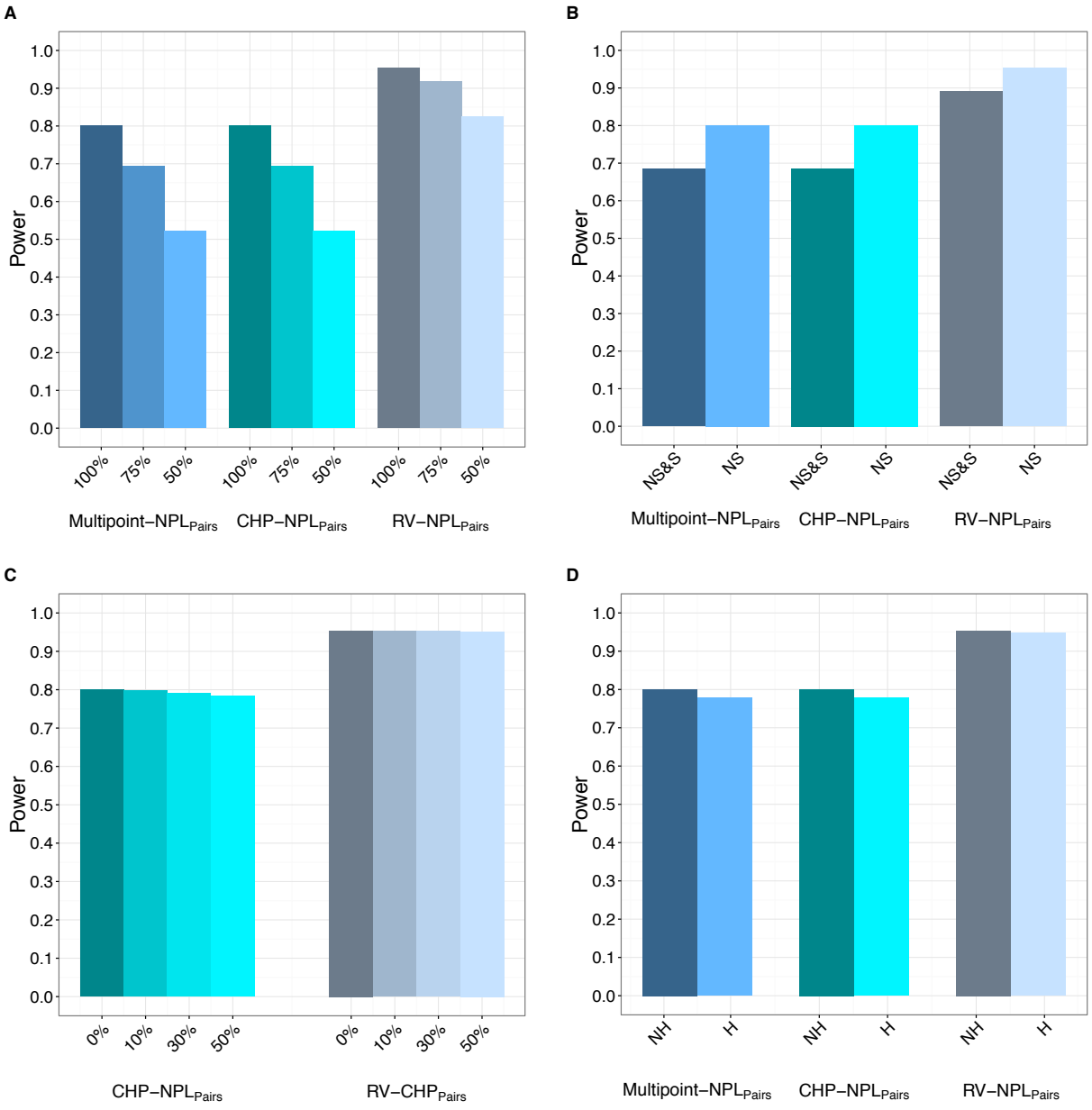

A
Power
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
100%
75%
50%
100%
75%
50%
100%
75%
50%
Multipoint–NPL$_{Pairs}$
CHP–NPL$_{Pairs}$
RV–NPL$_{Pairs}$
B
Power
NS&S
NS
NS&S
NS
NS&S
NS
Multipoint–NPL$_{Pairs}$
CHP–NPL$_{Pairs}$
RV–NPL$_{Pairs}$
C
Power
0%
10%
30%
50%
0%
10%
30%
50%
CHP–NPL$_{Pairs}$
RV–CHP$_{Pairs}$
D
Power
NH
H
NH
H
NH
H
Multipoint–NPL$_{Pairs}$
CHP–NPL$_{Pairs}$
RV–NPL$_{Pairs}$

Figure S9. Power comparison for NPL_{All} on extended families

Genotypes were generated for 100 extended families conditional on affection status assuming a multiplicative model in which each causal variant within a gene region has an OR of 5.0. Analysis was performed using RV-NPL_{All}, CHP-NPL_{All}, and Multipoint-NPL_{All}: with 100%, 75% and 50% of the variant being causal and the remaining non-causal (OR=1.0) (panel A); with only causal nonsynonymous (NS) variants as well as with causal nonsynonymous (NS) and non-causal synonymous (S) variants (panel B); with 0%, 10%, 30%, and 50% of the founders missing all genotype data (panel C); and with no heterogeneity (NH), i.e. 100 linked families as well as with locus heterogeneity (H), i.e., 100 linked and 50 unlinked families (panel D).

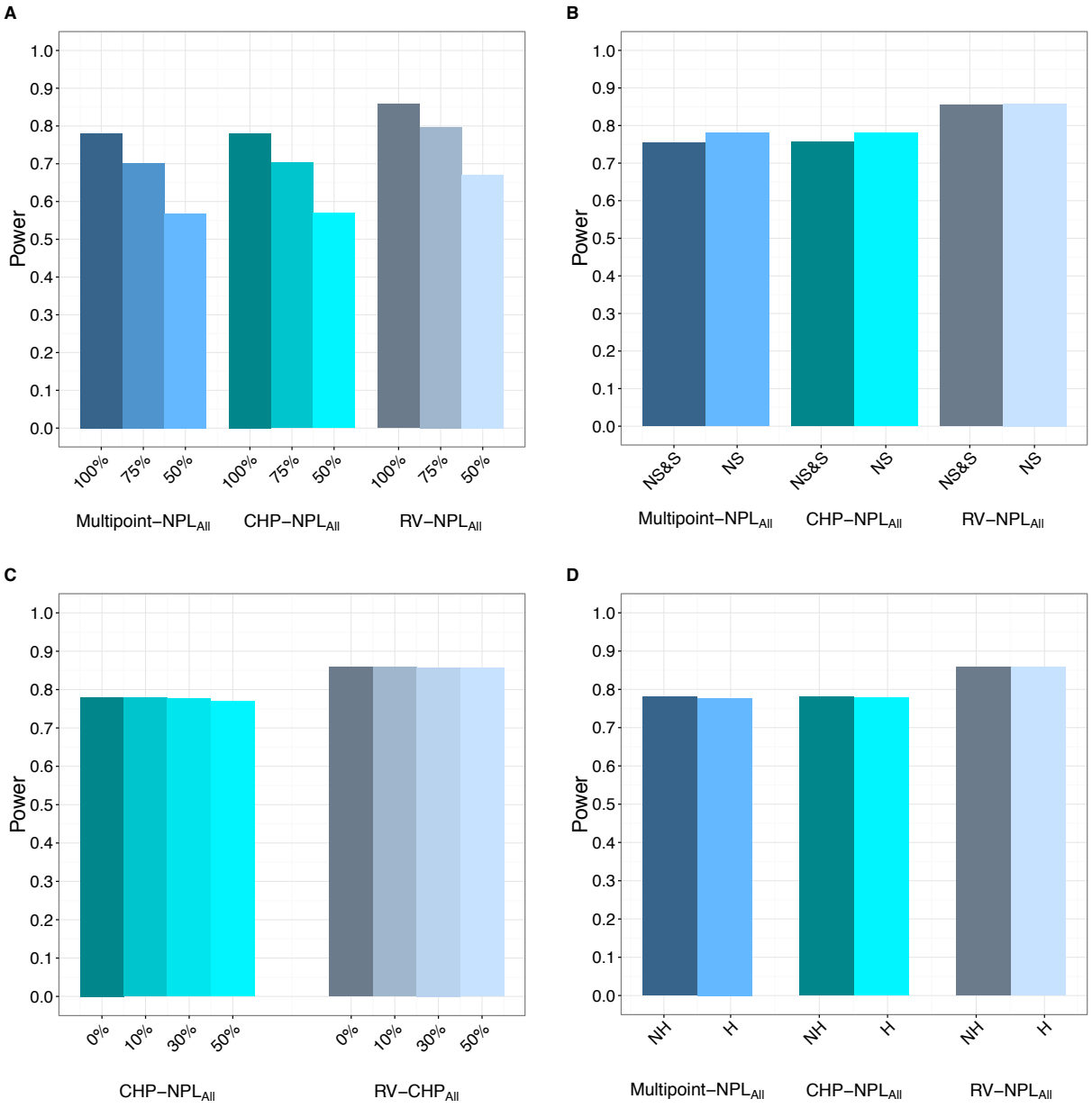

A
Power
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
100%
75%
50%
100%
75%
50%
100%
75%
50%
Multipoint–NPL$_{All}$
CHP–NPL$_{All}$
RV–NPL$_{All}$
B
Power
NS&S
NS
NS&S
NS
NS&S
NS
Multipoint–NPL$_{All}$
CHP–NPL$_{All}$
RV–NPL$_{All}$
C
Power
0%
10%
30%
50%
0%
10%
30%
50%
CHP–NPL$_{All}$
RV–CHP$_{All}$
D
Power
NH
H
NH
H
NH
H
Multipoint–NPL$_{All}$
CHP–NPL$_{All}$
RV–NPL$_{All}$

Figure S10. Power comparison for NPL_{All} on nuclear families with three affected siblings

Genotypes were generated for 300 nuclear families with three affected siblings conditional on affection status assuming a multiplicative model in which each causal variant within a gene region has an OR of 5.0. Analysis was performed using RV-NPL_{All}, CHP-NPL_{All}, and Multipoint-NPL_{All}: with 100%, 75% and 50% of the variant being causal and the remaining non-causal (OR=1.0) (panel A); with only causal nonsynonymous (NS) variants as well as with causal nonsynonymous (NS) and non-causal synonymous (S) variants (panel B); with 0%, 10%, 30%, and 50% of the founders missing all genotype data (panel C); and with no heterogeneity (NH), i.e. 100 linked families as well as with locus heterogeneity (H), i.e., 300 linked and 150 unlinked families (panel D).

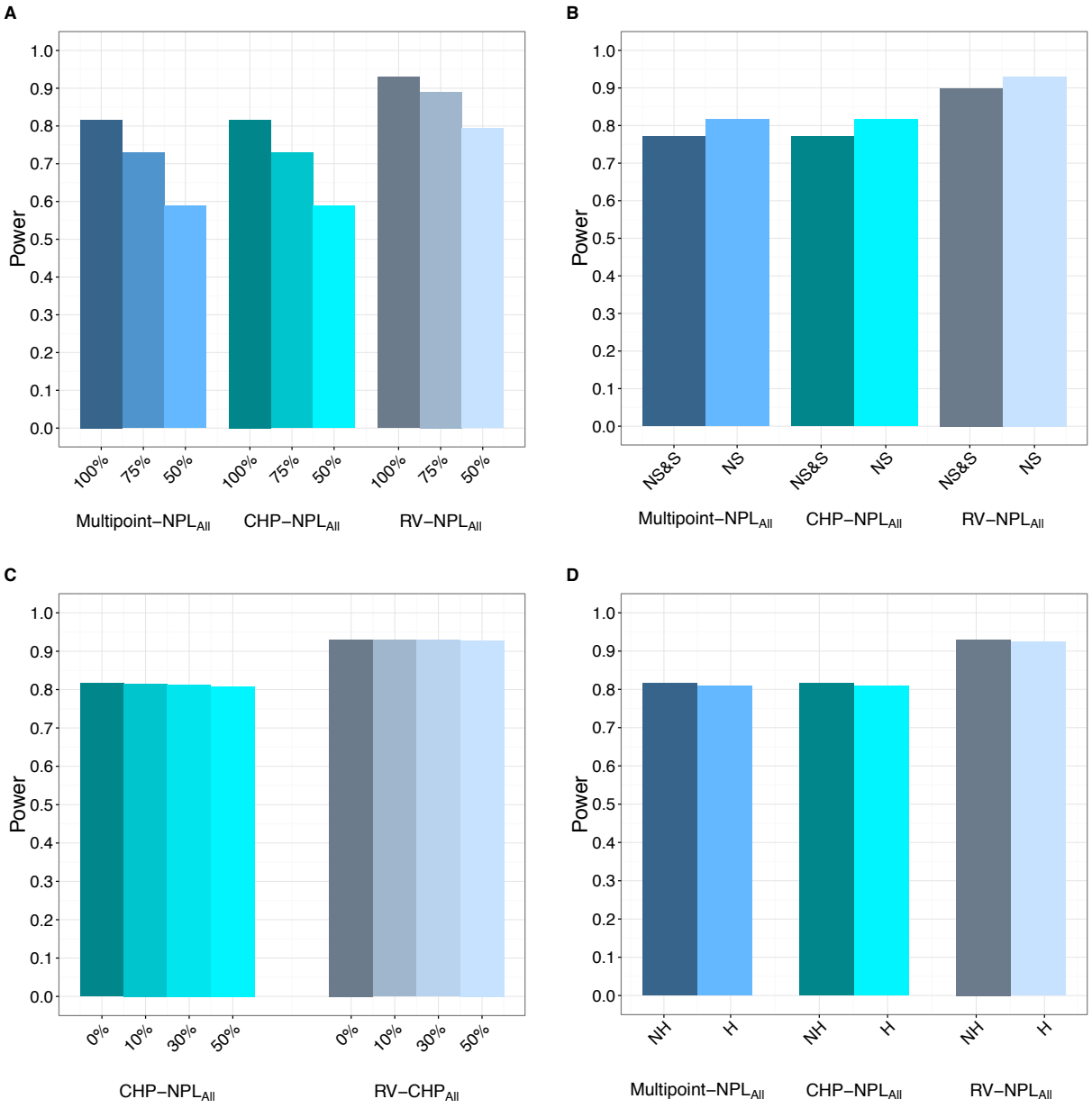

A
Power
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
100%
75%
50%
100%
75%
50%
100%
75%
50%
Multipoint–NPL$_{All}$
CHP–NPL$_{All}$
RV–NPL$_{All}$
B
Power
NS&S
NS
NS&S
NS
NS&S
NS
Multipoint–NPL$_{All}$
CHP–NPL$_{All}$
RV–NPL$_{All}$
C
Power
0%
10%
30%
50%
0%
10%
30%
50%
CHP–NPL$_{All}$
RV–CHP$_{All}$
D
Power
NH
H
NH
H
NH
H
Multipoint–NPL$_{All}$
CHP–NPL$_{All}$
RV–NPL$_{All}$

Table S1: The Ethnicities of Alzheimer's disease families included in the analysis

Ethnicity	Number of Families	Family IDs
Dominican	62	CU0002F, CU0003F[a], CU0004F[^], CU0005F[a], CU0006F[*a], CU0007F, CU0008F, CU0009F[c], CU0010F, CU0012F, CU0013F[a], CU0014F, CU0015F, CU0016F[a], CU0017F[^], CU0018F, CU0019F[^a], CU0020F, CU0022F, CU0023F[*], CU0024F, CU0025F, CU0026F, CU0029F, CU0030F[*a], CU0033F, CU0035F[^], CU0036F, CU0037F, CU0038F[*], CU0039F[a], CU0040F[a], CU0041F[a], CU0043F[c], CU0044F[a], CU0045F[c], CU0046F, CU0047F, CU0048F[^], CU0049F[*], CU0050F, CU0052F, CU0053F, CU0055F[^], CU0057F, CU0058F, CU0059F, CU0060F[a], CU0064F[a], CU0065F, CU0067F[*a], CU0068F[^], CU0070F[^a], CU0071F[a], CU0073F[^a], CU0075F[^], CU0076F[*], CU0078F[a], CU0079F[^], CU0081F[a], CU0082F, CU0083F[a]
European Descent	41	LD0168F, LD0179F, LD0223F, LD0232F, LD0241F[d], LD0254F, LD0307F, LD0856F, LD0949F, LD1012F[d], LD1223F, LD1260F, LD1265F, LD1315F[b], LD1329F[*], LD1579F[d], NC0049F, NC0131F, NC0205F[^], NC0302F, UM0002F, UM0146F[^b], UM0147F[d], UM0152F, UM0170F, UM0196F[d], UM0304F, UM0453F, UM0458F, UM0460F, UM0463F[^b], UM0464F, UP0001F, UP0002F, UP0003F, UP0004F[d], UP0005F, UP0006F, UP0007F, UP0008F, VU0072F
Puerto Rican	3	CU0032F, CU0042F, CU0051F
Dutch Isolate	1	203[d]

[^]Pedigrees with excess RV sharing for gene *PSMF1*.
*Pedigrees with excess RV sharing for gene *PTPN21*.
[a]Pedigrees with excess RV sharing for gene *ABCA7*; [b]Pedigrees with excess RV sharing for gene *ACE*;
[c]Pedigrees with excess RV sharing for gene *EPHA1*; [d]Pedigrees with excess RV sharing for gene *SORL1*.

Table S2. Type I error rate of CHP-NPL and RV-NPL at α-level of 0.05 and 0.005

		Nuclear pedigree with two affected siblings			Nuclear pedigree with three affected siblings			Extended pedigree		
	α-level	5.0×10^{-2}	5.0×10^{-3}	1.5×10^{-5}	5×10^{-2}	5.0×10^{-3}	1.5×10^{-5}	5.0×10^{-2}	5.0×10^{-3}	1.5×10^{-5}
No missing genotype	CHP-NPL_{Pairs}	4.8×10^{-2}	4.5×10^{-3}	1.0×10^{-5}	4.9×10^{-2}	4.8×10^{-3}	1.5×10^{-5}	5.0×10^{-2}	5.0×10^{-3}	1.5×10^{-5}
	CHP-NPL_{All}	4.8×10^{-2}	4.5×10^{-3}	1.0×10^{-5}	4.9×10^{-2}	4.8×10^{-3}	1.5×10^{-5}	5.0×10^{-2}	4.9×10^{-3}	1.3×10^{-5}
	RV-NPL_{Pairs}	4.6×10^{-2}	4.3×10^{-3}	1.4×10^{-5}	4.9×10^{-2}	4.8×10^{-3}	1.7×10^{-5}	4.9×10^{-2}	4.8×10^{-3}	1.7×10^{-5}
	RV-NPL_{All}	4.6×10^{-2}	4.3×10^{-3}	1.4×10^{-5}	4.9×10^{-2}	4.9×10^{-3}	1.6×10^{-5}	4.9×10^{-2}	4.9×10^{-3}	1.0×10^{-5}
All founders missing genotype	CHP-NPL_{Pairs}	4.6×10^{-2}	4.8×10^{-3}	1.4×10^{-5}	4.6×10^{-2}	4.5×10^{-3}	1.0×10^{-5}	5.1×10^{-2}	5.3×10^{-3}	1.4×10^{-5}
	CHP-NPL_{All}	4.6×10^{-2}	4.8×10^{-3}	1.4×10^{-5}	4.6×10^{-2}	4.5×10^{-3}	1.0×10^{-5}	5.1×10^{-2}	5.2×10^{-3}	1.7×10^{-5}
	RV-NPL_{Pairs}	5.1×10^{-2}	5.2×10^{-3}	1.7×10^{-5}	5.1×10^{-2}	5.2×10^{-3}	1.7×10^{-5}	4.9×10^{-2}	4.6×10^{-3}	1.5×10^{-5}
	RV-NPL_{All}	5.1×10^{-2}	5.2×10^{-3}	1.7×10^{-5}	5.1×10^{-2}	5.2×10^{-3}	1.6×10^{-5}	5.0×10^{-2}	4.8×10^{-3}	1.5×10^{-5}

Exome-wide type I error was evaluated using data generated for 1000 exomes and analyzing each gene. Three different values for α-level are shown here: $5.0\text{x}10^{-2}$, $5.0\text{x}10^{-3}$ and 1.5×10^{-5}. Type I error rate was calculated by dividing the total number of genes with a p-value equal or smaller than the α-level value by the number of genes analyzed across all 1000 generated exomes.

Table S3. Power comparison of NPL_{Pairs} and NPL_{All} in intra-familial locus heterogeneity

	RV-NPL_{Pairs}	RV-NPL_{All}	$Z_{All} > Z_{Pairs}$[a]
Without intra-familial locus heterogeneity	0.6410	0.6411	69.08%
With intra-familial locus heterogeneity	0.2997	0.2870	38.52%

Power was compared between RV-NPL_{Pairs} and RV-NPL_{All} in extended families with and without intra-familial locus heterogeneity.

[a]Proportion of total genes that have Z-scores of RV-NPL_{All} higher than that of RV-NPL_{Pairs}

Table S4: Bioinformatic evaluation and frequencies of analyzed rare variants within *PSMF1*

dbSNP rsID	rs751905514*^	rs35236223*^	rs148476395*	rs146300768^	rs146612629	rs79465651*	rs148156083*^	rs758812434*
hg19 position	20:1106192	20:1106214	20:1115798	20:1115864	20:1115870	20:1143797	20:1145081	20:1145111
Reference Allele	A	G	A	C	T	T	G	G
Alternate Allele	G	A	G	T	A	C	A	A
cDNA change	c.181A>G	c.203G>A	c.400A>G	c.466C>T	c.472T>A	c.575T>C	c.725G>A	c.755G>A
ACC	p.Asn61Asp	p.Arg68Gln	p.Ile134Val	p.Arg156Trp	p.Phe158Ile	p.Val192Ala	p.Arg242His	p.Ser252Asn
MAF[a]	$7.22x10^{-6}$	$3.61x10^{-5}$	$6.90x10^{-5}$	$4.08x10^{-4}$	$2.78x10^{-4}$	$5.61x10^{-3}$	$1.49x10^{-3}$	$3.66x10^{-5}$
MAF (NFE)[b]	$1.58x10^{-5}$	$3.16x10^{-5}$	$1.07x10^{-4}$	$5.53x10^{-5}$	$3.95x10^{-4}$	$7.26x10^{-4}$	$2.50x10^{-3}$	0
MAF (AMR)[c]	0	$2.91x10^{-5}$	$5.96x10^{-5}$	$3.20x10^{-4}$	$5.23x10^{-4}$	$3.31x10^{-3}$	$9.30x10^{-4}$	0
GERP score	4.93	4.93	4.12	2.07	5.03	5.26	5.11	4.12
PhyloP score	3.37	6.37	2.02	0.27	0.81	4.56	8.44	3.37
CADD score[d]	18.7	34.0	6.0	23.9	22.6	12.7	28.4	12.1
FATHMM	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated
MutationTaster	disease causing	disease causing	disease causing	polymorphism	disease causing	polymorphism	disease causing	polymorphism
Polyphen-2 HVAR	possibly damaging	probably damaging	benign	probably damaging	benign	benign	benign	benign
PROVEAN	neutral	deleterious	neutral	deleterious	neutral	neutral	deleterious	neutral
SIFT	tolerated	damaging	tolerated	damaging	tolerated	tolerated	tolerated	tolerated
LRT	deleterious	deleterious	deleterious	neutral	neutral	neutral	deleterious	deleterious

Abbreviations: ACC, amino acid change; MAF, minor allele frequency; NFE, Non-Finnish European; AMR, Latino; CADD, Combined Annotation Dependent Depletion; FATHMM, Functional Analysis through Hidden Markov Models; PROVEAN, Protein Variation Effect Analyzer; SIFT, Sorting Intolerant From Tolerant.

[a]MAFs are from gnomAD (Genome Aggregation Database) combining all populations; [b]MAFs are from gnomAD NFE population; [c]MAFs are from gnomAD AMR population; [d]Scaled CADD score.

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

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

Table S5: Bioinformatic evaluation and frequencies of analyzed rare variants within *PTPN21*

dbSNP rsID	rs141951135*^	rs150736820*^	rs143571855	rs3825676*^	rs149927113	rs138752198*	rs146847601*^
hg19 position	14:88935348	14:88935351	14:88945312	14:88945407	14:88945485	14:88974290	14:89016641
Reference Allele	G	G	G	C	C	T	C
Alternate Allele	A	A	C	G	G	C	A
cDNA change	c.3308C>T	c.3305C>T	c.2463C>G	c.2368G>C	c.2290G>C	c.425A>G	c.121G>T
ACC	p.Pro1103Leu	p.Pro1102Leu	p.Asp821Glu	p.Gly790Arg	p.Val764Leu	p.Gln142Arg	p.Val41Leu
MAF[a]	$5.41x10^{-5}$	$1.61x10^{-3}$	$1.95x10^{-3}$	$1.84x10^{-2}$	$2.17x10^{-4}$	$3.58x10^{-4}$	$4.94x10^{-4}$
MAF (NFE)[b]	$3.58x10^{-5}$	$2.73x10^{-3}$	$1.85x10^{-5}$	$1.95x10^{-2}$	0	$6.00x10^{-4}$	$7.90x10^{-6}$
MAF (AMR)[c]	$2.08x10^{-4}$	$1.16x10^{-4}$	$7.87x10^{-4}$	$3.86x10^{-3}$	$3.74x10^{-5}$	$2.06x10^{-4}$	$1.16x10^{-4}$
GERP score	5.90	5.90	-6.17	4.66	-2.01	5.36	5.50
PhyloP score	9.48	3.71	-1.29	5.10	0.88	2.42	7.60
CADD score[d]	34.0	22.9	0.04	19.8	0.1	7.9	23.6
FATHMM	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated	tolerated
MutationTaster	disease_causing	disease causing	polymorphism	disease causing	polymorphism	disease causing	disease_causing
Polyphen-2 HVAR	probably damaging	benign	benign	probably damaging	benign	benign	probably damaging
PROVEAN	deleterious	deleterious	neutral	neutral	neutral	neutral	neutral
SIFT	damaging	damaging	tolerated	damaging	tolerated	tolerated	tolerated
LRT	deleterious	deleterious	deleterious	deleterious	neutral	neutral	deleterious

Abbreviations: ACC, amino acid change; MAF, minor allele frequency; NFE, Non-Finnish European; AMR, Latino; CADD, Combined Annotation Dependent Depletion; FATHMM, Functional Analysis through Hidden Markov Models; PROVEAN, Protein Variation Effect Analyzer; SIFT, Sorting Intolerant From Tolerant.

[a]MAFs are from gnomAD (Genome Aggregation Database) combining all populations; [b]MAFs are from gnomAD NFE population; [c]MAFs are from gnomAD AMR population; [d]Scaled CADD score.

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

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

Table S6: Bioinformatic evaluation and frequencies of analyzed rare variants within *ABCA7*

dbSNP rsID	rs146597357	rs151054304	NA*^	rs138055574	rs76282929^	rs149949633*^	rs111940546*
hg19 position	19:1041922	19:1042353	19:1043175	19:1044672	19:1048898	19:1048950	19:1051209
Reference Allele	C	C	A	G	G	G	G
Alternate Allele	A	T	G	A	C	A	T
cDNA change	c.253C>A	c.455C>T	c.715A>G	c.1144G>A	c.2274G>C	c.2326G>A	c.2740G>T
ACC	p.Leu85Met	p.Pro152Leu	p.Asn239Asp	p.Gly382Ser	p.Gln758His	p.Gly776Arg	p.Ala914Ser
MAF[a]	$3.08x10^{-4}$	$1.42x10^{-3}$	.	$5.34x10^{-5}$	$4.28x10^{-3}$	$1.13x10^{-4}$	$1.86x10^{-4}$
MAF (NFE)[b]	$4.95x10^{-5}$	$2.44x10^{-5}$	.	$2.38x10^{-5}$	$6.58x10^{-5}$	$1.95x10^{-4}$	0
MAF (AMR)[c]	$2.41x10^{-4}$	$7.63x10^{-4}$	.	0	$1.59x10^{-3}$	$8.91x10^{-5}$	$1.75x10^{-4}$
GERP score	1.68	2.06	3.04	2.83	3.99	3.99	3.4
PhyloP score	0.83	-0.85	2.17	0.73	-5.34	6.44	1.39
CADD score[d]	13.4	7.4	20.4	8.8	25.3	28.6	8.16
FATHMM	damaging	damaging	damaging	damaging	damaging	tolerated	tolerated
MutationTaster	polymorphism	polymorphism	polymorphism	polymorphism	polymorphism	disease_causing	polymorphism
Polyphen-2 HVAR	benign	benign	probably damaging	benign	probably damaging	probably damaging	benign
PROVEAN	neutral	neutral	deleterious	neutral	deleterious	deleterious	neutral
SIFT	tolerated	tolerated	damaging	tolerated	damaging	damaging	tolerated
LRT	.	.	.	.	.	.	.

Abbreviations: ACC, amino acid change; MAF, minor allele frequency; NFE, Non-Finnish European; AMR, Latino; CADD, Combined Annotation Dependent Depletion; FATHMM, Functional Analysis through Hidden Markov Models; PROVEAN, Protein Variation Effect Analyzer; SIFT, Sorting Intolerant From Tolerant.

[a]MAFs are from gnomAD (Genome Aggregation Database) combining all populations; [b]MAFs are from gnomAD NFE population; [c]MAFs are from gnomAD AMR population; [d]Scaled CADD score.

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

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

Table S6: Bioinformatic evaluation and frequencies of analyzed rare variants within ABCA7 (continued)

dbSNP rsID	rs947668738*	rs114614802^	rs369849959	rs184590335*^	rs73505232*^	rs114782266^
hg19 position	19:1053401	19:1054324	19:1056127	19:1057919	19:1058635	19:1059056
Reference Allele	G	G	G	C	C	G
Alternate Allele	C	A	A	T	T	A
cDNA change	c.3294G>C	c.3710G>A	c.4301G>A	c.4886C>T	c.5168C>T	c.5435G>A
ACC	p.Glu1098Asp	p.Arg1237His	p.Arg1434His	p.Ser1629Leu	p.Ser1723Leu	p.Arg1812His
MAF[a]	$6.37x10^{-5}$	$2.38x10^{-3}$	$2.52x10^{-5}$	$1.29x10^{-3}$	$1.21x10^{-2}$	$1.06x10^{-2}$
MAF (NFE)[b]	0	$2.49x10^{-5}$	$1.60x10^{-5}$	0	$1.74x10^{-4}$	$6.41x10^{-3}$
MAF (AMR)[c]	$1.19x10^{-3}$	$1.03x10^{-3}$	$8.76x10^{-5}$	$9.47x10^{-3}$	$5.01x10^{-3}$	$5.35x10^{-3}$
GERP score	1.25	3.64	-2.23	4.22	4.23	0.81
PhyloP score	2.42	0.36	-0.98	7.64	2.03	4.26
CADD score[d]	22.9	32.0	2.8	35.0	33.0	21.8
FATHMM	tolerated	damaging	damaging	damaging	damaging	damaging
MutationTaster	polymorphism	polymorphism	polymorphism	disease_causing	polymorphism	polymorphism
Polyphen-2 HVAR	benign	probably damaging	benign	benign	benign	benign
PROVEAN	neutral	deleterious	neutral	deleterious	deleterious	deleterious
SIFT	tolerated	damaging	tolerated	damaging	damaging	damaging
LRT	.	.	.	.	.	.

Abbreviations: ACC, amino acid change; MAF, minor allele frequency; NFE, Non-Finnish European; AMR, Latino; CADD, Combined Annotation Dependent Depletion; FATHMM, Functional Analysis through Hidden Markov Models; PROVEAN, Protein Variation Effect Analyzer; SIFT, Sorting Intolerant From Tolerant.

[a]MAFs are from gnomAD (Genome Aggregation Database) combining all populations; [b]MAFs are from gnomAD NFE population; [c]MAFs are from gnomAD AMR population; [d]Scaled CADD score.

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

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

Table S7: Bioinformatic evaluation and frequencies of analyzed rare variants within *ACE*

dbSNP rsID	rs148943954*^	rs3730043*^	rs765069550
hg19 position	17:61560846	17:61568577	17:61574683
Reference Allele	C	C	C
Alternate Allele	G	T	T
cDNA change	c.1513C>G	c.2747C>T	c.3877C>T
ACC	p.Pro505Ala	p.Thr916Met	p.His1293Tyr
MAF[a]	$5.37x10^{-4}$	$4.02x10^{-3}$	$2.26x10^{-5}$
MAF (NFE)[b]	$1.24x10^{-4}$	$6.50x10^{-3}$	$4.38x10^{-5}$
MAF (AMR)[c]	$7.34x10^{-4}$	$1.55x10^{-3}$	0
GERP score	4.90	4.25	-0.28
PhyloP score	3.27	2.39	0.93
CADD score[d]	25.4	28.8	15.1
FATHMM	damaging	tolerated	tolerated
MutationTaster	disease_causing	disease_causing	polymorphism
Polyphen-2 HVAR	possibly damaging	probably damaging	benign
PROVEAN	deleterious	deleterious	neutral
SIFT	damaging	damaging	damaging
LRT	deleterious	deleterious	neutral

Abbreviations: ACC, amino acid change; MAF, minor allele frequency; NFE, Non-Finnish European; AMR, Latino; CADD, Combined Annotation Dependent Depletion; FATHMM, Functional Analysis through Hidden Markov Models; PROVEAN, Protein Variation Effect Analyzer; SIFT, Sorting Intolerant From Tolerant.

[a]MAFs are from gnomAD (Genome Aggregation Database) combining all populations; [b]MAFs are from gnomAD NFE population; [c]MAFs are from gnomAD AMR population; [d]Scaled CADD score.

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

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

Table S8: Bioinformatic evaluation and frequencies of analyzed rare variants within *EPHA1*

dbSNP rsID	rs139482378*^	rs139711610*^
hg19 position	7:143088584	7:143091417
Reference Allele	C	C
Alternate Allele	T	T
cDNA change	c.2897G>A	c.2372G>A
ACC	Arg966His	p.Arg791His
MAF[a]	$6.01x10^{-4}$	$3.26x10^{-4}$
MAF (NFE)[b]	$1.12x10^{-3}$	$1.55x10^{-5}$
MAF (AMR)[c]	$3.67x10^{-4}$	$3.11x10^{-4}$
GERP score	5.24	4.67
PhyloP score	2.51	7.56
CADD score[d]	35.0	35.9
FATHMM	tolerated	damaging
MutationTaster	disease_causing	disease_causing
Polyphen-2 HVAR	probably damaging	probably damaging
PROVEAN	deleterious	deleterious
SIFT	damaging	damaging
LRT	deleterious	deleterious

Abbreviations: ACC, amino acid change; MAF, minor allele frequency; NFE, Non-Finnish European; AMR, Latino; CADD, Combined Annotation Dependent Depletion; FATHMM, Functional Analysis through Hidden Markov Models; PROVEAN, Protein Variation Effect Analyzer; SIFT, Sorting Intolerant From Tolerant.

[a]MAFs are from gnomAD (Genome Aggregation Database) combining all populations; [b]MAFs are from gnomAD NFE population; [c]MAFs are from gnomAD AMR population; [d]Scaled CADD score.

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

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

Table S9: Bioinformatic evaluation and frequencies of analyzed rare variants within *SORL1*

dbSNP rsID	rs1051430452*^	rs150609294*^	rs139710266*^	rs62617129	rs62622819	rs140327834*^	rs142884576*^
hg19 position	11:121360768	11:121384931	11:121384991	11:121444958	11:121485599	11:121495816	11:121498300
Reference Allele	A	A	A	A	T	A	C
Alternate Allele	G	C	G	G	A	T	T
cDNA change	c.707A>G	c.1112A>C	c.1172A>G	c.3346A>G	c.5439T>A	c.6194A>T	c.6401C>T
ACC	p.Asp236Gly	p.Asn371Thr	p.Tyr391Cys	p.Ile1116Val	p.His1813Gln	p.Asp2065Val	p.Thr2134Met
MAF[a]	$3.98x10^{-6}$	$1.37x10^{-3}$	$3.18x10^{-5}$	$5.31x10^{-3}$	$4.99x10^{-3}$	$2.54x10^{-3}$	$3.29x10^{-4}$
MAF (NFE)[b]	$8.79x10^{-6}$	$2.17x10^{-3}$	$4.40x10^{-5}$	$8.25x10^{-3}$	$8.97x10^{-3}$	$4.10x10^{-3}$	$5.89x10^{-4}$
MAF (AMR)[c]	0	$1.41x10^{-4}$	0	$2.65x10^{-3}$	$1.89x10^{-3}$	$1.53x10^{-3}$	$5.64x10^{-5}$
GERP score	5.68	5.66	5.56	-5.57	-8.35	5.32	5.74
PhyloP score	8.73	9.24	9.24	-0.74	-1.34	8.64	2.63
CADD score[d]	33.0	24.1	25.0	0.05	9.4	25.5	23.9
FATHMM	tolerated	tolerated	tolerated	damaging	tolerated	tolerated	damaging
MutationTaster	disease_causing	disease_causing	disease_causing	polymorphism	disease_causing	disease_causing	disease_causing
Polyphen-2 HVAR	probably damaging	possibly damaging	probably damaging	benign	benign	probably damaging	benign
PROVEAN	deleterious	deleterious	deleterious	neutral	neutral	deleterious	neutral
SIFT	damaging	damaging	tolerated	tolerated	tolerated	tolerated	damaging
LRT	deleterious	deleterious	deleterious	neutral	neutral	deleterious	neutral

Abbreviations: ACC, amino acid change; MAF, minor allele frequency; NFE, Non-Finnish European; AMR, Latino; CADD, Combined Annotation Dependent Depletion; FATHMM, Functional Analysis through Hidden Markov Models; PROVEAN, Protein Variation Effect Analyzer; SIFT, Sorting Intolerant From Tolerant.

[a]MAFs are from gnomAD (Genome Aggregation Database) combining all populations; [b]MAFs are from gnomAD NFE population; [c]MAFs are from gnomAD AMR population; [d]Scaled CADD score.

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

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

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.