Published in final edited form as:

Genet Epidemiol. 2018 September ; 42(6): 500–515. doi:10.1002/gepi.22133.

Analysis of Pedigree Data in Populations with Multiple Ancestries: Strategies for Dealing with Admixture in Caribbean Hispanic Families from the ADSP

Rafael A. Nafikov¹, Alejandro Q. Nato Jr¹, Harkirat Sohi¹, Bowen Wang², Lisa Brown³, Andrea R. Horimoto¹, Badri N. Vardarajan⁴, Sandra M. Barral⁴, Giuseppe Tosto⁴, Richard P. Mayeux⁴, Timothy A. Thornton³, Elizabeth Blue¹, and Ellen M. Wijsman^{1,3,*}

¹Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, USA

²Department of Statistics, University of Washington, Seattle, USA

³Department of Biostatistics, University of Washington, Seattle, USA

⁴Department of Neurology, Columbia University, New York, USA

Abstract

Multipoint linkage analysis is an important approach for localizing disease-associated loci in pedigrees. Linkage analysis, however, is sensitive to misspecification of marker allele frequencies. Pedigrees from recently-admixed populations are particularly susceptible to this problem because of the challenge of accurately accounting for population structure. Therefore, increasing emphasis on use of multi-ethnic samples in genetic studies requires re-evaluation of best-practices, given data currently available. Typical strategies have been to compute allele frequencies from the sample, or to use marker allele frequencies determined by admixture proportions averaged over the entire sample. However, admixture proportions vary among pedigrees and throughout the genome in a family specific manner. Here we evaluate several approaches to model admixture in linkage analysis, providing different levels of detail about ancestral origin. To perform our evaluations, for specification of marker allele frequencies we used data on 67 Caribbean Hispanic admixed families from the Alzheimer's Disease Sequencing Project. Our results show that choice of admixture model has an effect on the linkage analysis results. Variant-specific admixture proportions, computed for individual families, provide the most detailed regional admixture estimates, and, as such, are the most appropriate allele frequencies for linkage analysis. This likely decreases the number of false positive results, and is straightforward to implement.

Keywords

complex trait; missing data; Markov Chain Monte Carlo; large pedigrees; late-onset disease

*Correspondence to: Ellen M. Wijsman, Division of Medical Genetics, Department of Medicine, BOX 357720, University of Washington, Seattle, WA 98195-7720, USA, wijsman@u.washington.edu.

INTRODUCTION

Multipoint linkage analysis remains an important approach for localizing trait loci (Ott, Wang, & Leal, 2015) prior to gene/region evaluation from sequence data (Franke et al., 2006; Krauthammer, Kaufmann, Gilliam, & Rzhetsky, 2004). Linkage analysis works best for localizing trait loci that segregate rare alleles, which may be the cause of much unexplained variance in complex diseases (Manolio et al., 2009; McClellan & King, 2010; Mitchell, 2012). Analysis can be carried out with exact computation in small pedigrees in the presence of many markers (Lander & Green, 1987) or in large pedigrees in the presence of only a few markers (Elston & Stewart, 1971). However, Markov Chain Monte Carlo (MCMC) computation (Heath, 1997; Sobel & Lange, 1996; Thompson & Heath, 1999; Thompson, 2011) is required for multipoint linkage analysis with many markers in large and possibly complex pedigrees, particularly in the presence of extensive missing data. A complete analysis typically involves determination of probabilities defining chromosomal inheritance patterns followed by inclusion of the probability of the trait data given these inheritance patterns (Kruglyak, Daly, Reeve-Daly, & Lander, 1996; Lange & Sobel, 1991; Sobel & Lange, 1996).

Past uses of linkage analysis in studies of the genetic basis of Alzheimer's disease (AD) successfully led to identification of several genes for rare early-onset autosomal-dominant forms of AD. Genes well-established in early-onset AD are the amyloid precursor protein (APP; Goate et al., 1991) and the presenilins (PSEN1, PSEN2; Levy-Lahad et al., 1995; Sherrington, 1995). Pedigree-based linkage studies also identified a few genomic regions associated with the more common late-onset forms of AD (LOAD; Butler et al., 2009; Lee et al., 2008; Rademakers et al., 2005) but genetic heterogeneity of the disease (St George-Hyslop et al., 1990) and sparse available sequence data limited progress towards gene identification. The apolipoprotein E gene (APOE; Corder et al., 1993; Corder et al., 1994) was the only one identified for LOAD as part of these earlier studies. Genome-wide association studies (GWAS) have more recently discovered a number of variants significantly associated with LOAD risk (Lambert et al., 2009; Seshadri et al., 2010; Sims et al., 2017) but GWAS do not identify the causal genes, and GWAS-identified variants explain only a small fraction of the LOAD-based heritability, estimated to be ~79% (Gatz et al., 2006).

A large set of pedigrees was recently selected for investigation as part of the AD Sequencing Project (ADSP). This project was launched in response to the 2012 United States National Alzheimer's Project Act to identify AD risk and protective gene variants with a long-term goal of discovering AD-associated pathways to facilitate the development of therapeutic approaches and preventatives for the disease. The Caribbean Hispanic (CH) component of this sample had a genome scan with SNP markers (Barral et al., 2015), using only single-marker linkage analysis. Chromosome-wide multipoint linkage analysis, however, is preferable because of better resolution of regions with positive linkage signals as well smoothing of the noisy effects observed with single marker linkage analysis (Kruglyak et al., 1996).

Linkage analysis can be sensitive to allele frequency misspecification (Ott, 1992). This sensitivity is particularly problematic when founder genotypes in a pedigree are missing (Schaid, McDonnell, Wang, Cunningham, & Thibodeau, 2002; Sieh, Yu, Bird, Schellenberg, & Wijsman, 2007), which is the case in AD families. Use of sample-specific allele frequencies was one early proposed solution to the problem (Knapp, Seuchter, & Baur, 1993); in the context of analysis of an admixed population, this is equivalent to assuming allele frequencies that represent the population average admixture derived from the set of contributing ancestral populations. The issue of allele frequency misspecification, however, is more challenging when recently admixed populations such as the ADSP CH families are used. Here admixture fractions vary substantially across individuals (Bryc et al., 2010), and use of the population average frequency may have an impact on the linkage analysis results. Recent availability of dense genomic data and multiple reference populations now makes it possible to perform a more detailed ancestry evaluation in admixed populations than was previously possible (Bryc et al., 2010). These scientific advances created a need to re-evaluate existing approaches for allele frequency specification in linkage analysis in admixed populations.

This paper has two goals. First, we evaluate several different approaches for specifying SNP allele frequencies in recently-admixed populations in the context of pedigree-based multipoint linkage analysis. Second, we provide results from full multipoint linkage analysis in the individual ADSP CH pedigrees. Our results demonstrate the potential importance of accounting for admixture at the sub chromosomal level in linkage analysis, and highlight regions with evidence for linkage in individual families in this sample.

METHODS

Sample and genomic data

All 67 CH families from the Dominican Republic that were part of the ADSP discovery project were used in the current study (Beecham et al., 2017) with the approval from institutional review boards at participating institutions and receipt of a written informed consent from all participants. For the purpose of the current analysis, the AD phenotype of each individual was coded as: affected for subjects with a clinical diagnosis of definite, probable, or possible AD; unaffected for subjects with a clinical confirmation of no AD or other dementia at most recent examination; and unknown for all other subjects (McKhann et al., 2011). Some individuals within each family were genotyped with one of several SNP genotyping array platforms, providing a total of 545 genotyped individuals. The SNP genotype data, based on NCBI Build 37, were subjected to quality control, filtered to remove all non-diallelic variants, and merged together into a single joint panel using forward strand orientation with PLINK version 1.07 (Purcell et al., 2007) and retaining only those SNPs with an overall genotype completion rate of 80 %. As part of QC, subjects with sex chromosome information that did not match their reported sex or with estimated kinship coefficients that did not match that of their position in pedigree were removed, and two virtual parents were added. To avoid unnecessary computation that would have no impact on results, pedigree trimming was performed to remove individuals with no information relevant to the analysis (i.e. individuals with both no genotype and no phenotype data at the

bottom of a pedigree). As a result, the total number of individuals in CH families decreased from 2,806 to 1,379. In this paper the following ADSP CH families were used to provide examples in the results section: CU0005F, CU0018, CU0030F, CU0040F, CU0042F, and CU0048F. The selected families had a number of individuals ranging from 9 to 36 and a ratio of individuals with GWAS data to the total number of individuals in a pedigree ranging from 0.28 to 0.78.

Statistical analyses

Overview—In brief, our analysis of genomic data involved separate computations of inheritance patterns and probability of trait data, given these inheritance patterns. We started with multipoint computation of the inheritance patterns of a panel of markers, selected from the merged GWAS arrays to meet specific criteria for multipoint analysis, described further below. We refer to this panel of markers as a “framework panel”, because it consists of only a small fraction of the GWAS array and frames downstream analysis. The inheritance patterns are represented as inheritance vectors (IVs) at each framework marker position (Kruglyak et al., 1996). We then used these IVs to compute LOD scores for linkage analysis with `gl_lods` from the MORGAN package (Tong & Thompson, 2007). Because of the admixed background of the families, we performed principal component analysis (PCA) for validating ancestral composition of the families, and we estimated their regional ancestry genome-wide, in both cases using the same denser set of markers that differed from the framework marker panels. We then used the ancestry estimates to compute appropriate framework marker allele frequencies for the multipoint computation of inheritance vectors. We propose several admixture models to compute such marker allele frequencies. After computing IVs for multiple sets of framework markers that were generated using each admixture model, we computed LOD scores. The effect of various choices of admixture model on linkage analysis was then evaluated.

Local ancestry estimation—We used SNPs from the merged reference panel to estimate 3-way admixture between European, African, and Native American populations, locally for each subject, in all subjects with GWAS SNP data (Figure 1). Three samples provided reference population data to estimate ancestry locally along chromosomes in our sample: European and African ancestries were represented by 165 (CEU) and 203 (YRI) samples, respectively, both from the International HapMap Project, phase 3, release 3 (International HapMap 3 Consortium, 2010). Because the HapMap project did not include a Native American sample, we used a surrogate for Native American ancestry represented by 63 samples from the Americas of the Human Genome Diversity Project (HGDP) database (Cann et al., 2002; Rosenberg et al., 2002; Rosenberg et al., 2005). We merged HapMap and HGDP reference databases by keeping only variants common to the three reference populations mentioned above. This resulted in 603,611 SNPs with an overall genotyping rate of 0.998. The combined reference panel was then merged with SNP markers from the ADSP CH samples (Figure 1), retaining only SNPs with missing rate below 7 %, a threshold selected based upon the missing data distribution in the merged sample. The resulting data set consisted of 273,523 SNPs with an overall genotyping rate of 0.996, was also used in PCA, and was different from framework marker panels used in multipoint linkage analysis.

Alleles in this final data set were then phased to haplotypes (Figure 1) using Beagle version 3.3.2 (S. Browning & B. Browning, 2007).

We carried out local ancestry estimation in the ADSP CH samples with RFMix, version 1.5.4 (Maples, Gravel, Kenny, & Bustamante, 2013) using the PopPhased option to allow for genotype phase uncertainty (Figure 1). RFMix was used because, relative to other options, it has higher computational efficiency and accuracy, it has the ability to differentiate between closely related populations, and it handles more than two reference populations of relatively small sizes (Maples et al., 2013). The method implemented in RFMix is one of the first methods for local ancestry estimation that uses information about linkage disequilibrium to connect ancestry along an admixed chromosome to observed haplotypes in reference populations. Local ancestry estimation was performed by ignoring the relatedness among ADSP CH subjects because including such information confounds the admixture estimates and results in biased estimates (Conomos, Miller, & Thornton, 2015). The combined SNP panel used in the analysis with RFMix included only samples from unrelated reference-population individuals (112 CEU, 147 YRI, and 63 Native Americans) excluding subjects who were less than second-degree relatives. These groups of unrelated individuals were identified as described below in the PCA subsection of the Methods section. The subjects retained for use differ slightly from earlier published lists (Pemberton et al., 2010), likely because of differences in approaches used to estimate relatedness in a structured sample. We computed global ancestry estimates over the whole sample by averaging across local ancestry values for all 273,523 SNPs that had local ancestry estimates and used these estimates in PCA.

PCA—We obtained PCs on the ADSP CH samples both to validate the ancestral composition and as a part of the process of identifying unrelated individuals in the reference populations for local ancestry estimation. The above step involved an iterative PCA and pairwise kinship coefficient estimation and the use of the same merged sample as in the local ancestry estimation. We first computed the pairwise ancestry divergence measures for the ADSP samples and each of the three populations in the combined reference panel to identify subsets of unrelated and related individuals (Figure 1). The kinship option of the Kinship-based Inference for Gwas robust (KING-robust) method (Manichaikul et al., 2010) is insensitive to population structure, such as exists in the CH sample, and was used to do the computation. Here ancestry divergence measures for related individuals were used as pairwise kinship coefficients. We then used the pairwise ancestry divergence measures for unrelated and related individuals from the ADSP CH samples to perform an initial PCA using the PC-AiR function (Conomos et al., 2015) in the GENetic ESTimation and Inference in Structured samples (GENESIS) package (Conomos, Thornton, Gogarten, & Brown, 2018; Figure 1). The PC-AiR calculates PCs for a mutually unrelated set of subjects and then projects PC values for the related subjects accounting for known and cryptic relatedness in the sample. The PC-AiR function was used with the MAF, kin.thresh, and div.thresh options to threshold MAF, kinship coefficients, and divergence measures at 0.05, 0.025, and -0.025, respectively. Next we computed initial pairwise kinship coefficients with the PC-Relate function (Conomos et al., 2016) in the GENESIS package (Conomos et al., 2018), which allows for population and pedigree structure (Figure 1). The PC-Relate function was used

with a threshold of 0.05 for the MAF and the pcMAT option to adjust for global proportional ancestry estimated, as described in the Local ancestry estimation subsection of the Methods section. After the initial estimation of pairwise kinship coefficients, we repeated the PCA which resulted in final PCs for the ADSP CH samples used to create a PCA plot (Figure 1). We also computed final pairwise kinship coefficients and used them to exclude related individuals from reference populations for local ancestry estimation, as described earlier (Figure 1).

Framework marker selection—The software package Pedigree-Based Analysis Pipeline (PBAP) was used to select three non-overlapping marker panels. These panels, which were different from the marker panel used for local ancestry estimation and PCA, comprised an average of 5,507 genome-wide SNP framework markers selected to be ideal for linkage analysis (Nato et al., 2015). We used these markers to compute IVs for multipoint linkage analysis. Only one panel of framework markers was used for the genome-scan analysis. The remaining two panels were used in regions of interest to make sure that results were robust to the choice of panel. The framework markers were selected based on the following criteria: data completion of >95% (resulting in $97.63 \pm 0.15\%$ data completion), in linkage equilibrium ($r^2 < 0.04$), sparsely and uniformly spaced at an interval of > 0.5 cM (resulting in an average of 0.66 ± 0.013 cM) to accommodate requirements of MCMC computations, and having relatively high minor allele frequency > 0.25 (resulting in mean MAF = 0.36 ± 0.005) to be informative. In general, a range of spacing of framework markers should suffice. However, our intention was to use chromosomal inheritance patterns at framework markers for a variety of follow-up analyses, some of which benefit from relatively close spacing (Cheung, Thompson, & Wijsman, 2013). Allele frequencies derived from individuals of only European (EUR) descent of the 1000 Genomes Project data, Phase I, version 3 (Genomes Project, 2012) were used to define MAF-based marker selection thresholds. Framework marker positions were obtained from the Rutgers map (Kong et al., 2004; Matise et al., 2007), which was based on dbSNP Build 134. Map positions were then converted to those based on the Haldane map function. Direct matching using marker rs number and/or physical position or a linear interpolation, if a marker was not in the map database, was used to determine positions of framework markers on the same map as was used for estimating the IVs.

Allele frequency modifications of framework markers in the admixed CH

population—Framework marker allele frequencies used in multipoint analysis to compute IVs were derived from the 1000 Genomes Project phase-1, version 3 data (Genomes Project, 2012) using weighted averages of allele frequencies for the EUR, African (AFR), and Native American (AMR) reference populations. The weights were specified by admixture proportions calculated from local ancestry estimates for each of the four admixture models described below. Admixture proportions were calculated by averaging local admixture estimates computed earlier (1) across all the families and chromosomes (Global), (2) across all the chromosomes, genome-wide, but within each family (Family-Based Genome-Wide, FBGW), (3) within each chromosome and family (Family-Based Chromosome-Wide, FBCW), and (4) by determining regional admixture estimates for each SNP within each chromosome and family (Local). We developed a Perl script ADMIXFRQ available on

<https://github.com/RafPrograms/ADMIXFRQ>, which calculates admixture proportions and specifies marker allele frequencies under the four different admixture models mentioned above.

The reference populations used had adequate sample sizes to provide reasonably accurate allele frequency estimates for the framework markers. The number of individuals in the EUR, AFR, and AMR reference populations of the 1000 Genomes Project were 379, 246, and 181, respectively, yielding a standard error (SE) of 0.025 for an allele frequency of 0.36 (the mean MAF of our framework markers) in the smallest AMR reference population and 0.017 in the largest EUR reference population. As is typical in very late-onset diseases, extremely limited data availability on founders in the ADSP CH families precluded use of within-pedigree observations for allele frequency specification.

Multipoint computation of IVs—We generated samples of IVs from the posterior distribution of possible IVs, given the pedigree, meiotic map, and observed multipoint marker data. For this purpose we used the `gl_auto` program of the Morgan package version 3.3 (Tong & Thompson, 2007) which samples from the exact distribution (Lander & Green, 1987) for smaller pedigrees, and uses MCMC with a hybrid sampler (Heath, 1997; Thompson & Heath, 1999; Tong & Thompson, 2007) for larger pedigrees. Here “small” represents pedigrees with 20 meioses. To generate IVs, we used a panel of genome scan framework markers as described above. Sequential imputation (Kong, Cox, Frigge, & Irwin, 1993) with 20 Monte Carlo (MC) iterations was used to obtain a starting configuration prior to the MCMC iterations. The number of initial burn-in iterations was set to be either 10% or 15% of the total number of MC main iterations, which were set to either 50,000 or 75,000 (Table S1). The number of saved iterations ranged from 1,000 to 3,000, depending on pedigree size. For MCMC computation both the locus and meiosis samplers (L- and M-sampler) were used with 20% and 80% of updates made by each, respectively (Tong & Thompson, 2007). A ‘sample by scan’ sampling method was used to update successfully all marker loci and meioses in an order determined by random permutation. The outputs of the analysis were sampled IVs that describe possible IBD patterns among individuals in a pedigree. The analysis was run four times in all CH families using framework marker allele frequencies computed from admixture proportions for each of the four admixture models described above.

LOD Score computations—We computed LOD scores with the `gl_lods` program of the Morgan package (Tong & Thompson, 2007). This computation used the IVs generated by `gl_auto` at the positions of the markers used to compute IVs, a file with information on AD status, coded as a discrete trait, and specification of trait model parameters. We used a dominant model with penetrance probabilities of 0.001, 0.9, and 0.9 for low-risk homozygotes, heterozygotes, and high-risk homozygotes, respectively, with the frequency for the high risk allele 0.05. Reduced penetrance of 0.9 for high-risk-allele carriers and low penetrance of 0.001 for non-carriers allowed for the fact that AD is a complex trait defined by combined action of multiple factors, with genetic heterogeneity both within and between individual families. An AD allele frequency of 0.05 provided further robustness to the model for the effects of multiple genes entering pedigrees through different founders.

RESULTS

Ancestral origin of the ADSP CH sample and admixture modeling

European ancestry was predominant in the CH families followed by African and, finally, Native American ancestries. The PCA for the ADSP CH sample showed the 3-way admixture between the European, African, and Native American populations (Figure 2), with estimates of Global admixture proportions of 0.65, 0.27, and 0.09, respectively. The FBGW admixture proportions, however, did not always follow the same trend in admixture proportions established by the Global admixture model. As with the Global admixture model, the European ancestry was generally the predominant one among FBGW admixture proportions, but ranged from 0.31 to 0.83 (Figure 3A). A few CH families, however, had the African ancestry as the major contributor, although, the African ancestry was generally the second most predominant, ranging from 0.09 to 0.65. With the exception of a few families, the Native American ancestry had the smallest contribution, with its proportion ranging from 0.03 to 0.14 (Figure 3A). Only one pedigree had a Native American ancestral proportion as high as 0.25 making it the second-largest ancestral background in this family.

Admixture proportions differ among families and different admixture models. FBGW admixture proportions for the European, African, and Native American ancestries in CU0048F, for example, were 0.77, 0.11, and 0.11, respectively, showing more European and Native American and less African ancestry than the Global ancestry estimates. A distribution of the FBCW admixture proportions across each of the 22 chromosomes in CU0048F ranged from 0.62 to 0.91 for the European ancestry, from 0.06 to 0.22 for the Native American ancestry, and from 0.02 to 0.18 for the African ancestry (Figures S1). As for the local admixture proportions, the chromosome-wide average vary continuously genome-wide and their specificity depends on an ancestral background of a family at a particular genomic location. For the majority of genomic regions, the Local admixture proportions were different from those for the Global, FBGW, and FBCW admixture models (Figures 4A, 5, and 6A).

The choice of admixture model has a strong effect on marker allele frequencies. Scatter plots of alternative allele frequencies for 5481 framework markers and all possible combinations of the four admixture models shows large differences between the models (Figures 3B–D and S2). The biggest difference in allele frequencies was between the Local and Global admixture models (Figure 3B) followed by the differences of the Global and Local admixture models versus both FBGW and FBCW models. The smallest difference in allele frequencies was observed between FBGW and FBCW models (Figure 3D).

Linkage signals and admixture models

The use of the Local admixture model can result in lower LOD scores in linkage analysis than the use of one of the other models, consistent with preventing LOD score inflation. For example, for CU0048F positive LOD scores for the Global, FBGW, and FBCW admixture models were detected on chr 18 between 67.4 and 87.1 cM, with maximum LOD scores of 1.18, 1.30, and 1.16, respectively. In the same genomic region, the LOD score for the Local admixture model was negative or only slightly above zero (Figure 4A). These differences in

LOD scores between admixture models coincides with switching in ancestry along the chromosome that is only captured using the Local admixture model (Figure 4A). This switching results in an increased proportion of European ancestry and decreased proportions of African and Native American ancestries in the Local admixture model. Admixture proportions for the Global, FBGW, and FBCW admixture models are, of course, constant throughout the chromosome. As a result, ancestry in this region, when compared with the proportions from the Local admixture model, is lower than the local estimates for European ancestry and higher than the local estimates for the African and Native American ancestries. Genome-wide results of linkage analysis in CU0048F evaluated by comparing LOD scores for the Local versus Global, FBGW, and FBCW admixture models show that differences were observed for both positive and negative ranges of LOD scores (Figures 4B–D).

Another instance of possible LOD score inflation on chr 7 is observed in CU0030F when the Global rather than the Local admixture model is used. The use of the Global admixture model resulted in a maximum LOD score of 1.88 detected on chr 7 between 59.20 and 74.17 cM, while LOD scores computed with the Local, FBCW, and FBGW admixture models for the same genomic region did not exceed 0.66 (Figure 5A). Within the linkage region there was an increase in the proportion of European ancestry for the Local admixture model that surpassed the proportions of European ancestry computed using the Global, FBGW, and FBCW admixture models (Figure 5A). At the same time, the proportion of African ancestry for the Local admixture model decreased and reached a value below that observed for other admixture models.

Probable underestimation of the evidence for linkage can also occur if the Local admixture model is not used in the analysis. For CU0005F, for example, the maximum LOD score for the Local admixture model detected on chr 22 was 1.48 between 35.54 and 72.16 cM, while the use of Global, FBGW, and FBCW admixture models gave LOD scores of only 0.30, 0.47, and 0.31, respectively (Figure 5B). At the left boundary of the linkage area the proportion of European ancestry is the smallest for the Local admixture model but gradually increases over the linkage region, reaching the largest value in the region equivalent to the one computed with the Global admixture model. The opposite trend occurs for the proportion of African ancestry calculated by the Local admixture model. The proportion of Native American ancestry specified by the Local admixture model only slightly decreases over this linkage region.

In some cases, the Local admixture model can produce the largest LOD score. Linkage analysis on chr 17 in CU0042F, for example, revealed a maximum LOD score of 2.06 for the Local admixture model between 43.57 and 56.95 cM (Figure 5C). Smaller LOD scores of 1.34, 1.12, and 1.49 were determined in the same genomic region for the Global, FBGW, and FBCW admixture models, respectively. Here a LOD score increase for the Local admixture model was associated with a rise in the proportion of its Native American ancestry (Figure 5C). Meanwhile, the proportion of European ancestry for the Local admixture model decreased and the proportion of African ancestry for the same model fluctuated. As compared with the Local admixture model, ancestry proportions for the Global, FBGW, and FBCW admixture models were lower for Native Americans and higher for Europeans.

The choice of admixture model to compute marker allele frequencies for linkage analysis can influence not only a magnitude of linkage signal but also the boundaries of the linkage region. Thus, multipoint linkage analysis in CU0018F on chr 13, for example, identified linkage signals with similar LOD scores of 1.32, 1.25, 1.34, and 1.34 for Local, Global, FBGW, and FBCW admixture models, respectively (Figure 5D). The boundaries of the linkage signals, however, differ between the admixture models. For example, the right boundary for FBGW and FBCW admixture models was located at 69.33 cM while the same boundary for Local admixture model was further along the chromosome at 85.35 cM (Figure 5D). The telomeric boundary for the same linkage signals differed between the models as well. For Global and FBGW admixture models, the telomeric boundary was at 19.25 and 24.37 cM, respectively that was some distance away from the boundary point for Local and FBCW admixture models and was at 17.26 cM. Changes in a location of the boundaries of linkage region were also associated with ancestral switching observed for the Local admixture model. The proportions of African and European ancestries for the Local admixture model increased and decreased, respectively, above and below the corresponding values for Global, FBGW, and FBCW admixture models.

Overall, the use of the Global admixture model in linkage analysis leads to overall increases in LOD scores, consistent with LOD score inflation. Distributions of cumulative LOD score differences between the Local and the other three admixture models show a skew to the left (Figure S3, Table S2), with the greatest difference between the Local and Global admixture models. This indicates that on average the Local admixture model produced the lowest LOD scores among the three models. Total LOD scores summed over all families at each marker position were negative, genome-wide, with a maximum of -35.35 at 263.43 cM on chr 2 for the Local admixture model (Tables S5–8).

There are also situations where the choice of admixture model has little or no effect on the results. Linkage analysis on chr 3 in CU0040F, for example, produced maximum LOD scores of 0.84 for all admixture models within the same genomic region between 103.51 and 133.81 cM (Figure 6A). Admixture proportions for European, African, and Native American ancestries calculated with the Local admixture model varied within the linkage region. However, these changes had no effect on LOD scores in this particular example, which consisted of seven affected siblings with both genotype and phenotype data, and two parents with neither type of data. Genome-wide LOD scores computed on CU0040F for the Local admixture model versus LOD scores for the Global, FBCW, and FBGW (Figures 6B–D) admixture models show insensitivity of the analysis to ancestry specification in this family.

Summary of multipoint linkage analysis in the ADSP CH sample

Multipoint linkage analysis in the ADSP CH sample run using the Local admixture model to compute marker allele frequencies identified 20 suggestive or significant genome-wide linkage regions (Table 1). The largest LOD score of 3.02 was identified on chr 9 between 165.36 and 166.83 cM in CU0023F. The second-largest LOD score of 2.42 was identified on chr 9, as well, but between 83.42 and 94.83 cM and in a different family, CU0016F (Table 1). In general, linkage signals with LOD score above 1.9 were identified on a wide range of chromosomes and in different families. There was not a single case of linkage region overlap

among these strongest linkage signals identified in different families. The average size of linkage regions was 11.47 cM with the widest region of 38 cM and the maximum LOD score of 2.32 identified on chr 15 in CU0076F. For three linkage signals with LOD scores of 3.02, 2.13, and 1.96 - identified on chr 9, chr 10, and chr 1 in CU0023F, CU0007F, and CU0081F, respectively - the size of the linkage regions was small, ranging from 1.47 to 2.22 cM. These narrow linkage regions might be indicative of possible false positive signals.

New regions with evidence of linkage were identified when the Local admixture model was used to specify framework marker allele frequencies in multipoint linkage analysis. Results of linkage analysis for marker positions reported in Table 1 were expanded by adding LOD scores computed using the Global, FBGW, and FBCW admixture models (Table S3). Similarly, linkage regions with notable maximum LOD score > 1.9 , obtained using the Global admixture model, are reported in Table S4, along with the corresponding LOD scores from the three other admixture models. Comparison between Tables S3 and S4 shows that 5 of the 20 linkage regions identified as notable by one admixture model (Local or Global) are not notable under the other admixture model. For example, use of the Local model identifies a signal with LOD=3.02 on chr 9 in CU0023F while the Global model only attains LOD=1.03 at that position, and the Global model identifies a signal with LOD=2.35 on chr 3 in CU0005F while the Local model obtains LOD=-1.64 at the same position. The results in Tables S3 and S4 also reinforce the concept that the locations of the maximum LOD scores and boundaries of linkage regions vary when different admixture models are used to specify framework marker allele frequencies in linkage analysis.

DISCUSSION

Here we showed that different approaches for specifying admixture-informed marker allele frequencies can affect the results of multipoint linkage analysis in pedigrees from admixed populations, including the ADSP CH pedigrees. We evaluated four models for specifying marker allele frequencies for estimation of IVs, for models ranging from treating all pedigrees as drawn from a single set of allele frequencies defined by the average ancestry proportions in the sample, to models with increasing resolution of ancestry proportions within individual families and along the genome. We showed that both the magnitude and location of linkage signals can vary substantially. We also show that the genome-wide distribution of LOD scores is consistent with an increasing positive bias for allele frequency models that fail to take into account different ancestry proportions among families, and along the genome. We provide a Perl script ADMIXFRQ, available on <https://github.com/RafPrograms/ADMIXFRQ>. This script calculates genome-wide marker allele frequencies for use with MORGAN (Tong & Thompson, 2007) for easing marker allele frequency specification for different admixture models.

Accounting for admixture in linkage analysis is, in a sense, an old research problem which is reevaluated in this paper in light of new types of data. The problem of misspecification of allele frequencies in linkage analysis has been previously addressed (Knapp et al., 1993) but in a context of unstructured populations. In the past, sparse genotype data and absence of good reference samples made it impossible to perform detailed admixture estimation in admixed populations. Current availability of the dense genomic data, coupled with

feasibility of ancestry-specific allele frequency estimation (Browning et al., 2016), however, allowed reevaluation here of the problem of allele frequency misspecification in linkage analysis, but in the context of admixed populations.

The question of how best to account for admixture in computation of IBD in pedigrees is a special-case of the general problem of marker allele-frequency specification. For computation of marker IBD in pedigrees, the importance of allele frequency specification has been known for many years, both with respect to choice of individual marker allele frequencies (Knapp et al., 1993) and haplotype frequencies to the population under investigation (Sieh et al., 2007). However, specific issues that are important in the context of admixed samples has not previously been discussed, and in the absence of dense SNP or sequence data needed to estimate local ancestry, only an approach such as the Global model is possible. With current genomic data, however, relatively accurate estimates of ancestry along the genome are now available, and should be taken into account.

Admixture proportions in the ADSP CH families are typical of those reported for CH from the Dominican Republic. Average admixture proportions for European, African, and Native American ancestries in the general USA Latino population are 65.1%, 6.2%, and 18.0%; in Latinos who self-identify as Dominicans, these proportions are 56%, 28%, and 7% (Bryc, Durand, Macpherson, Reich, & Mountain, 2015). These proportions were calculated using relatively small sample sizes from the 23andMe data set. Average admixture proportions across the ADSP CH sample computed as the Global admixture proportions were similar, with the European ancestry being the largest and the Native American ancestry the smallest of the three ancestries.

Allele frequencies in admixed populations are influenced by the choice of admixture model used for their calculation. In our study the Global, FBGW, FBCW, and Local admixture models generally provided different admixture proportions for any particular variant considered. As a result, allele frequencies computed with these admixture models are also different. This implies that for analytical procedures such as multipoint linkage analysis where a choice of allele frequencies matters, the use of appropriate admixture modeling is important for accurate analysis.

The Local admixture model appears to be the most appropriate for specifying marker allele frequencies from an appropriate set of reference allele frequencies. The Local model uses the most refined information regarding admixture proportions for any particular variant in a family. Use of the Global, FBGW, and FBCW admixture models does not account for frequent switches in ancestry of chromosomal segments inferred along chromosomes in individuals from an admixed population. As a result, admixture proportions provided by these three admixture models are not as precise as those provided by the Local admixture model.

False positive or negative linkage signals may occur if the Local admixture model is not used to calculate allele frequencies for linkage analysis. A genomic location of linkage signal is also affected by the choice of admixture model used to calculate framework marker allele frequencies for linkage analysis. Earlier work showed that misspecification of linkage

marker allele frequencies can result in false-positive linkage signals (Ott, 1992). In studies of familial AD where genotype data on individuals from earlier generations of a pedigree are unavailable, the effect of linkage marker allele frequency misspecification is particularly problematic (Knapp et al., 1993). The proposed solution to the problem was to calculate allele frequencies from actual family data. To the best of our knowledge, the problem of allele frequency misspecification in linkage analysis in admixed families has not been addressed at all. Our proposed use of the Local admixture model to calculate linkage marker allele frequencies, however, is a better solution than the use of allele frequencies from actual pedigree data because population data provide better estimates of allele frequencies than small data within a family.

The use of the Local admixture model in linkage analysis could more accurately define haplotypes co-segregating with a disease phenotype. This could justify a preference for the Local admixture model in linkage analysis over more general ones that do not take into account regional admixture. Most of the time ancestral switchers occur at left and right boundaries of a linkage region and no substantial switching occurs in the middle. Absence of ancestral switching within linkage regions insures the predominance of a particular ancestry which haplotype might be responsible for driving linkage signal. Use of proper admixture proportions helps to correctly calculate allele frequencies necessary for defining segregating haplotype. Misspecification of admixture proportions at variant level, however, leads to false framework marker allele frequencies which could define a nonexistent haplotype outside a real linkage region. This situation might lead to false-positive linkage signals if the nonexistent haplotype happened by chance to segregate with a disease phenotype. The opposite could be true as well when misspecified allele frequencies fail to define segregating haplotype in true linkage regions leading to false-negative results.

In small pedigrees, linkage analysis is not sensitive to the choice of framework marker allele frequencies. This phenomenon was mostly observed in pedigrees with less than 15 individuals where the choice of admixture model to calculate framework marker allele frequencies did not affect the magnitude and location of linkage signals. Neither number of individuals nor generations in a pedigree by itself seems to have an effect on LOD score differences computed with different admixture models. A plausible explanation for the lack of differences in LOD scores between different admixture models here is that in small pedigrees with SNP genotype data available on informative individuals it is possible to accurately infer IVs and in that case information about framework marker allele frequencies does not have an effect on IVs computation.

In summary, our proposed use of the Local admixture model to compute framework marker allele frequencies for multipoint linkage analysis in admixed families could decrease the occurrence of false-positive and –negative linkage signals and could improve the accuracy of linkage signal localization. This approach for calculating allele frequencies in admixed families is particularly important when there are missing genotype data in upper generations of pedigrees that is commonly the case for the late-onset diseases such as AD. Our refined approach for allele frequency calculation in admixed families could also be used in other analyses where the choice of allele frequencies matters.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially supported by the National Institute of Health (NIH) grants P50AG005136, U01AG049505, U01AG049506, and U01AG049507. The authors declare no conflict of interest.

References

- Barral S, Cheng R, Reitz C, Vardarajan B, Lee J, Kunkle B, ... Mayeux R. Linkage analyses in Caribbean Hispanic families identify novel loci associated with familial late-onset Alzheimer's disease. *Alzheimer's & Dementia*. 2015; 11(12):1397–1406. DOI: 10.1016/j.jalz.2015.07.487
- Beecham GW, Bis J, Martin E, Choi SH, DeStefano A, van Duijn C, ... Schellenberg G. The Alzheimer's Disease Sequencing Project: Study design and sample selection. *Neurology Genetics*. 2017; 3(5):e194.doi: 10.1212/NXG.000000000000194 [PubMed: 29184913]
- Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*. 2007; 81(5):1084–1097. DOI: 10.1086/521987 [PubMed: 17924348]
- Browning SR, Grinde K, Plantinga A, Gogarten SM, Stilp AM, Kaplan RC, ... Laurie CC. Local Ancestry Inference in a Large US-Based Hispanic/Latino Study: Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *G3 (Bethesda)*. 2016; 6(6):1525–1534. DOI: 10.1534/g3.116.028779 [PubMed: 27172203]
- Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *The American Journal of Human Genetics*. 2015; 96(1):37–53. DOI: 10.1016/j.ajhg.2014.11.010 [PubMed: 25529636]
- Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, Auton A, ... Ostrer H. Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences*. 2010; 107(Suppl 2):8954–8961. DOI: 10.1073/pnas.0914618107
- Butler AW, Ng MY, Hamshere ML, Forabosco P, Wroe R, Al-Chalabi A, ... Powell JF. Meta-analysis of linkage studies for Alzheimer's disease - a web resource. *Neurobiology of Aging*. 2009; 30(7):1037–1047. DOI: 10.1016/j.neurobiolaging.2009.03.013 [PubMed: 19362756]
- Cann HM, De Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, ... Cambon-Thomsen A. A human genome diversity cell line panel. *Science*. 2002; 296(5566):261–262. DOI: 10.1126/science.296.5566.261b [PubMed: 11954565]
- Cheung CY, Thompson EA, Wijsman EM. GIGI: an approach to effective imputation of dense genotypes on large pedigrees. *American Journal of Human Genetics*. 2013; 92(4):504–516. DOI: 10.1016/j.ajhg.2013.02.011 [PubMed: 23561844]
- Conomos MP, Laurie CA, Stilp AM, Gogarten SM, McHugh CP, Nelson SC, ... Laurie CC. Genetic diversity and association studies in US Hispanic/Latino populations: applications in the Hispanic Community Health Study/Study of Latinos. *The American Journal of Human Genetics*. 2016; 98(1):165–184. DOI: 10.1016/j.ajhg.2015.12.001 [PubMed: 26748518]
- Conomos MP, Miller MB, Thornton TA. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genetic Epidemiology*. 2015; 39(4):276–293. DOI: 10.1002/gepi.21896 [PubMed: 25810074]
- Conomos MP, Thornton T, Gogarten SM, Brown L. GENESIS: GENetic ESTimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness. R package version 2.8.1. 2018; doi: 10.18129/B9.bioc.GENESIS
- Corder E, Saunders A, Strittmatter W, Schmechel D, Gaskell P, Small G, ... Pericak-Vance MA. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*. 1993; 261(5123):921–923. DOI: 10.1126/science.8346443 [PubMed: 8346443]

- Corder E, Saunders AM, Risch N, Strittmatter W, Schmechel D, Gaskell P, ... Schmechel K. Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nature Genetics*. 1994; 7(2):180–184. DOI: 10.1038/ng0694-180 [PubMed: 7920638]
- Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. *Human Heredity*. 1971; 21(6):523–542. DOI: 10.1159/000152448 [PubMed: 5149961]
- Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *The American Journal of Human Genetics*. 2006; 78(6):1011–1025. DOI: 10.1086/504300 [PubMed: 16685651]
- Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, ... Pedersen NL. Role of genes and environments for explaining Alzheimer disease. *Archives of General Psychiatry*. 2006; 63(2):168–174. DOI: 10.1001/archpsyc.63.2.168 [PubMed: 16461860]
- Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, ... McVean GA. Genomes Project, C. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012; 491(7422):56–65. DOI: 10.1038/nature11632 [PubMed: 23128226]
- Goate A, Chartier-Harlin MC, Mullan M, Brown J, Crawford F, Fidani L, ... Hardy J. Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature*. 1991; 349(6311):704–706. DOI: 10.1038/349704a0 [PubMed: 1671712]
- Heath SC. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *The American Journal of Human Genetics*. 1997; 61(3):748–760. DOI: 10.1086/515506 [PubMed: 9326339]
- International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010; 467(7311):52–58. DOI: 10.1038/nature09298 [PubMed: 20811451]
- Knapp M, Seuchter SA, Baur MP. The effect of misspecifying allele frequencies in incompletely typed families. *Genetic Epidemiology*. 1993; 10(6):413–418. DOI: 10.1002/gepi.1370100614 [PubMed: 8314036]
- Kong A, Cox N, Frigge M, Irwin M. Sequential imputation and multipoint linkage analysis. *Genetic Epidemiology*. 1993; 10(6):483–488. DOI: 10.1002/gepi.1370100626 [PubMed: 8314048]
- Kong X, Murphy K, Raj T, He C, White P, Matise T. A combined linkage-physical map of the human genome. *The American Journal of Human Genetics*. 2004; 75(6):1143–1148. DOI: 10.1086/426405 [PubMed: 15486828]
- Krauthammer M, Kaufmann CA, Gilliam TC, Rzhetsky A. Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proceedings of the National Academy of Sciences*. 2004; 101(42):15148–15153. DOI: 10.1073/pnas.0404315101
- Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. Parametric and nonparametric linkage analysis: a unified multipoint approach. *The American Journal of Human Genetics*. 1996; 58(6):1347. [PubMed: 8651312]
- Lambert JC, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, ... Amouyel P. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nature Genetics*. 2009; 41(10):1094–1099. DOI: 10.1038/ng.439 [PubMed: 19734903]
- Lander ES, Green P. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences*. 1987; 84(8):2363–2367.
- Lange K, Sobel E. A random walk method for computing genetic location scores. *The American Journal of Human Genetics*. 1991; 49(6):1320. [PubMed: 1746559]
- Lee JH, Barral S, Cheng R, Chacon I, Santana V, Williamson J, ... Stern Y. Age-at-onset linkage analysis in Caribbean Hispanics with familial late-onset Alzheimer's disease. *Neurogenetics*. 2008; 9(1):51–60. DOI: 10.1007/s10048-007-0103-3 [PubMed: 17940814]
- Levy-Lahad E, Wasco W, Poorkaj P, Romano DM, Oshima J, Pettingell WH, ... Wang K. Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science*. 1995; 269(5226):973–977. DOI: 10.1126/science.7638622 [PubMed: 7638622]
- Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics*. 2010; 26(22):2867–2873. DOI: 10.1093/bioinformatics/btq559 [PubMed: 20926424]

- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, ... Visscher PM. Finding the missing heritability of complex diseases. *Nature*. 2009; 461(7265):747–753. DOI: 10.1038/nature08494 [PubMed: 19812666]
- Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*. 2013; 93(2): 278–288. DOI: 10.1016/j.ajhg.2013.06.020 [PubMed: 23910464]
- Matisse TC, Chen F, Chen W, De La Vega FM, Hansen M, He C, ... Buyske S. A second-generation combined linkage physical map of the human genome. *Genome Research*. 2007; 17(12):1783–1786. DOI: 10.1101/gr.7156307 [PubMed: 17989245]
- McClellan J, King MC. Genetic heterogeneity in human disease. *Cell*. 2010; 141(2):210–217. DOI: 10.1016/j.cell.2010.03.032 [PubMed: 20403315]
- McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, ... Phelps CH. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*. 2011; 7(3):263–269. DOI: 10.1016/j.jalz.2011.03.005
- Mitchell KJ. What is complex about complex disorders? *Genome Biology*. 2012; 13(1):237.doi: 10.1186/gb-2012-13-1-237 [PubMed: 22269335]
- Nato AQ Jr, Chapman NH, Sohi HK, Nguyen HD, Brkanac Z, Wijsman EM. PBAP: a pipeline for file processing and quality control of pedigree data with dense genetic markers. *Bioinformatics*. 2015; 31(23):3790–3798. DOI: 10.1093/bioinformatics/btv444 [PubMed: 26231429]
- Ott J. Strategies for characterizing highly polymorphic markers in human gene mapping. *The American Journal of Human Genetics*. 1992; 51(2):283. [PubMed: 1642229]
- Ott J, Wang J, Leal SM. Genetic linkage analysis in the age of whole-genome sequencing. *Nature Reviews Genetics*. 2015; 16(5):275–284. DOI: 10.1038/nrg3908
- Pemberton TJ, Wang C, Li JZ, Rosenberg NA. Inference of unexpected genetic relatedness among individuals in HapMap Phase III. *The American Journal of Human Genetics*. 2010; 87(4):457–464. DOI: 10.1016/j.ajhg.2010.08.014 [PubMed: 20869033]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, ... Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*. 2007; 81(3):559–575. DOI: 10.1086/519795 [PubMed: 17701901]
- Rademakers R, Cruts M, Sleegers K, Dermaut B, Theuns J, Aulchenko Y, ... Van Broeckhoven C. Linkage and association studies identify a novel locus for Alzheimer disease at 7q36 in a Dutch population-based sample. *The American Journal of Human Genetics*. 2005; 77(4):643–652. DOI: 10.1086/491749 [PubMed: 16175510]
- Rosenberg NA. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Annals of Human Genetics*. 2006; 70(6):841–847. DOI: 10.1111/j.1469-1809.2006.00285.x [PubMed: 17044859]
- Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics*. 2005; 1(6):e70.doi: 10.1371/journal.pgen.0010070 [PubMed: 16355252]
- Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. Genetic structure of human populations. *Science*. 2002; 298(5602):2381–2385. DOI: 10.1126/science.1078311 [PubMed: 12493913]
- Schaid DJ, McDonnell SK, Wang L, Cunningham JM, Thibodeau SN. Caution on pedigree haplotype inference with software that assumes linkage equilibrium. *The American Journal of Human Genetics*. 2002; 71(4):992.doi: 10.1086/342666 [PubMed: 12387273]
- Seshadri S, Fitzpatrick AL, Ikram MA, DeStefano AL, Gudnason V, Boada M, ... Lambert JC. Genome-wide analysis of genetic loci associated with Alzheimer disease. *JAMA*. 2010; 303(18): 1832–1840. DOI: 10.1001/jama.2010.574 [PubMed: 20460622]
- Sherrington R, Rogaev EI, Liang YA, Rogaeva EA, Levesque G, Ikeda M, ... Tsuda T. Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature*. 1995; 375(6534):754–760. DOI: 10.1038/375754a0 [PubMed: 7596406]
- Sieh W, Yu CE, Bird TD, Schellenberg GD, Wijsman EM. Accounting for linkage disequilibrium among markers in linkage analysis: impact of haplotype frequency estimation and molecular

- haplotypes for a gene in a candidate region for Alzheimer's disease. *Human Heredity*. 2007; 63(1): 26–34. DOI: 10.1159/000098459 [PubMed: 17215579]
- Sims R, Van Der Lee SJ, Naj AC, Bellenguez C, Badarinarayan N, Jakobsdottir J, ... Martin ER. Rare coding variants in *PLCG2*, *ABI3*, and *TREM2* implicate microglial-mediated innate immunity in Alzheimer's disease. *Nature Genetics*. 2017; 49(9):1373–1384. DOI: 10.1038/ng.3916 [PubMed: 28714976]
- Sobel E, Lange K. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *The American Journal of Human Genetics*. 1996; 58(6):1323. [PubMed: 8651310]
- St George-Hyslop P, Haines J, Farrer L, Polinsky R, Van Broeckhoven C, Goate A, ... Sorbi S. Genetic linkage studies suggest that Alzheimer's disease is not a single homogeneous disorder. *Nature*. 1990; 347(6289):194–197. DOI: 10.1038/347194a0 [PubMed: 2395471]
- Thompson E. The structure of genetic linkage data: from LIPED to 1M SNPs. *Human Heredity*. 2011; 71(2):86–96. DOI: 10.1159/000313555 [PubMed: 21734399]
- Thompson EA, Heath SC. Estimation of conditional multilocus gene identity among relatives. *Lecture Notes-Monograph Series*. 1999:95–113.
- Tong L, Thompson E. Multilocus lod scores in large pedigrees: combination of exact and approximate calculations. *Human Heredity*. 2007; 65(3):142–153. DOI: 10.1159/000109731 [PubMed: 17934317]

Figure 1.

Local ancestry estimation and PCA. Solid and dashed line boxes represent data sets and analytical procedures used on these data sets, respectively, with heavy lines indicating data sets used in additional analyses reported here. ADSP, Alzheimer's Disease Sequencing Project; CH, Caribbean Hispanics; GENESIS, GENetic ESTimation and Inference in Structured samples package; GWAS, Genome-Wide Association Study; KING-robust, Kinship-based INference for Gwas robust method; PCA, principal component (PC) analysis.

Figure 2.

PCA of the ADSP admixed CH population. Dots and pluses represent “unrelated” and “related” subsets of individuals, respectively. PCA, principal component (PC) analysis.

Figure 3.

Ancestry modeling and allele frequency specification in the ADSP CH families. (A) Family-based genome-wide (FBGW) admixture proportions of European, African, and Native American ancestries for each of 67 families. Two horizontal black lines drawn across the bar chart serve as reference points for the values of admixture proportions of the above average ancestries in the ADSP CH sample shown in the first bar as the Global admixture proportions for the ancestries in the entire ADSP CH sample. (B–D) are scatter plots of genome-wide framework markers' alternative allele frequencies (AF) which were computed in CU0048F using one of our four admixture models and the 1000 Genomes Project populations' data. Dashed lines define borders of a region where alternative AF differences are within ± 0.1 limit. FBCW, family-based chromosome-wide; ADSP and CH are defined in Figure 1 legend.

Figure 4.

Possible inflation of logarithm of odds (LOD) scores in multipoint linkage analysis when generalized admixture models are used to compute framework marker allele frequencies. (A) Upper part of the plot shows LOD scores for multipoint linkage analysis with incomplete penetrance model on chr 18 in CU0048F. Global, Family-Based Genome-Wide (FBGW), Family-Based Chromosome-Wide (FBCW), and Local models of ancestry specification were used to compute framework marker allele frequencies used in the linkage analysis. Lower parts of the plot show admixture proportions for European (EUR), African (AFR), and Native American (AMR) ancestries computed using four different admixture models mentioned above. (B), (C), and (D) are scatter plots of LOD scores for which allele frequencies were computed with one of the four different admixture models mentioned above. ADSP and CH are defined in Figure 1 legend.

Figure 5.

Possible inflation and deflation of logarithm of odds (LOD) scores and linkage region misspecification in multipoint linkage analysis when generalized admixture models used to compute framework marker allele frequencies. Upper part of the plots shows LOD scores for multipoint linkage analysis with incomplete penetrance model on chr 7 (A), chr 22 (B), chr 17 (C), and chr 13 (D) in CU0030F, CU0005F, CU0042F, and CU0018F, respectively. Global, Family-Based Genome-Wide (FBGW), Family-Based Chromosome-Wide (FBCW), and Local models of ancestry specification were used to compute framework marker allele frequencies used in the linkage analysis. Lower parts of the plots show admixture proportions for European (EUR), African (AFR), and Native American (AMR) ancestries computed in the four different families using four different admixture models mentioned above. ADSP and CH are defined in Figure 1 legend.

Figure 6.

Insensitivity of multipoint linkage analysis to allele frequency misspecification in pedigree CU0040F. (A) Upper part of the plot shows logarithm of odds (LOD) scores for multipoint linkage analysis with incomplete penetrance model on chr 3. Global, Family-Based Genome-Wide (FBGW), Family-Based Chromosome-Wide (FBCW), and Local models of ancestry specification were used to compute framework marker allele frequencies used in the linkage analysis. Lower parts of the plot show admixture proportions for European (EUR), African (AFR), and Native American (AMR) ancestries computed using four different admixture models mentioned above. (B), (C), and (D) are scatter plots of LOD scores for which allele frequencies were computed with one of the four different admixture models mentioned above. ADSP and CH are defined in Figure 1 legend.

Table 1

Results of multipoint linkage analysis with incomplete penetrance model and Local admixture model used to specify linkage marker frequencies in the ADSP CH sample, for individual pedigrees where the maximum LOD score > 1.9.

Family ID	N of Indiv	LOD Score	Chr	cM	Linkage Region Boundaries (cM)		bp	rs#
					left	right		
CU0023F	68	3.02	9	166.83	165.36	166.83	140581700	rs10732688
CU0016F	27	2.42	9	87.18	83.42	94.83	88371919	rs10868345
CU0042F	36	2.39	16	54.55	52.97	73.04	26729319	rs4787866
CU0076F	16	2.32	15	21.12	1.36	39.36	31197564	rs4779794
CU0039F	33	2.29	22	14.03	12.43	19.60	20793914	rs1035239
CU0007F	60	2.27	17	74.98	49.23	79.67	47793042	rs2017835
CU0076F	16	2.27	1	27.99	22.72	38.42	12654035	rs12045736
CU0014F	23	2.21	2	43.74	41.50	53.61	22098922	rs10200675
CU0023F	68	2.19	16	125.34	123.28	133.72	86094005	rs6540248
CU0005F	33	2.17	8	136.21	129.70	140.48	127767127	rs7822787
CU0042F	36	2.16	13	2.12	0.01	11.99	21512850	rs11619673
CU0007F	60	2.13	10	162.32	161.76	163.98	130706844	rs1886380
CU0042F	36	2.06	17	54.18	43.57	56.95	29038032	rs1808255
CU0081F	17	2.00	10	20.11	13.23	25.77	7280550	rs10905129
CU0039F	33	1.98	1	199.94	186.44	208.20	193901506	rs574986
CU0037F	51	1.97	16	30.13	27.04	35.73	12292138	rs4780416
CU0081F	17	1.97	6	123.20	120.34	125.01	115886188	rs12207371
CU0081F	17	1.96	1	159.06	157.80	159.68	157319351	rs7535100
CU0042F	36	1.92	4	8.68	0.16	10.60	5550882	rs10937654
CU0044F	48	1.92	3	162.52	161.15	170.27	151361249	rs4680584

N of Indiv, number of individuals in a family used for analysis; LOD Score, logarithm of odds score; Chr, chromosome number; cM, centimorgan position of a marker; bp, base pair position of a marker; rs#, rs number. ADSP and CH are defined in Figure 1 legend.