Bioinformatics, 31(23), 2015, 3790–3798
doi: 10.1093/bioinformatics/btv444
Advance Access Publication Date: 30 July 2015
Original Paper

Genetics and population analysis

PBAP: a pipeline for file processing and quality control of pedigree data with dense genetic markers

Alejandro Q. Nato Jr[1], Nicola H. Chapman[1], Harkirat K. Sohi[1], Hiep D. Nguyen[1], Zoran Brkanac[2] and Ellen M. Wijsman[1,3,4,*]

[1]Division of Medical Genetics, Department of Medicine, [2]Department of Psychiatry and Behavioral Sciences, [3]Department of Biostatistics and [4]Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on April 2, 2015; revised on July 7, 2015; accepted on July 25, 2015

Abstract

Motivation: Huge genetic datasets with dense marker panels are now common. With the availability of sequence data and recognition of importance of rare variants, smaller studies based on pedigrees are again also common. Pedigree-based samples often start with a dense marker panel, a subset of which may be used for linkage analysis to reduce computational burden and to limit linkage disequilibrium between single-nucleotide polymorphisms (SNPs). Programs attempting to select markers for linkage panels exist but lack flexibility.

Results: We developed a pedigree-based analysis pipeline (PBAP) suite of programs geared towards SNPs and sequence data. PBAP performs quality control, marker selection and file preparation. PBAP sets up files for MORGAN, which can handle analyses for small and large pedigrees, typically human, and results can be used with other programs and for downstream analyses. We evaluate and illustrate its features with two real datasets.

Availability and implementation: PBAP scripts may be downloaded from http://faculty.washington.edu/wijsman/software.shtml.

Contact: wijsman@uw.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Technical advances have allowed the acquisition and use of tremendous amounts of marker data for both population- and pedigree-based genetic studies (Lambert *et al.*, 2013; Mardis, 2008, 2011; Metzker, 2010; Ragoussis, 2009; Schadt *et al.*, 2010; Shendure and Ji, 2008; Wang *et al.*, 2013). For human population-based studies, multiple computational tools and workflows already exist for file manipulation, quality control (QC) and basic analysis of genotype data (Fan and Song, 2012; Fuchsberger *et al.*, 2012; Gogarten *et al.*, 2012; Laurie *et al.*, 2010; Matise *et al.*, 2011; Patel and Jain, 2012; Pongpanich *et al.*, 2010; Purcell *et al.*, 2007; Ritchie *et al.*, 2010; Zhou *et al.*, 2013). For pedigree-based studies, preparation and manipulation of files is similarly crucial for downstream analyses. Some tools used to format files and perform QC in population-based studies can also be used in pedigree studies, but additional pedigree-specific preparations are also necessary in the search for rare, high risk sequence variants.

Pedigree-based designs are again being used in the search for rare disease-causing variants. Pedigree-based datasets are a useful resource (Wijsman, 2012), and diverse analysis programs exist that can handle smaller pedigrees (e.g. Merlin) (Abecasis *et al.*, 2002) and large pedigrees (e.g. Simwalk2, MORGAN) (Sobel and Lange, 1996; Thompson, 2011) with extensive marker data. For computational reasons, smaller pedigrees are typically less than ~27 bits,

while larger pedigrees are at least ~27 bits (bits = 2*n*–*f*; where *n* = number of non-founders and *f* = number of founders). Searches for rare variants often start with linkage analysis for a particular trait (Hinrichs and Suarez, 2011; Keramati *et al.*, 2014; Musunuru *et al.*, 2010; Rosenthal *et al.*, 2011; Zhao *et al.*, 2013) with follow-up on sequence data in the linkage regions from the same families. Considerable software exists for carrying out computations on pedigrees to extract maximal information on identity-by-descent (IBD) (Abecasis *et al.*, 2002; Gudbjartsson *et al.*, 2000; Heath, 1997; Sobel and Lange, 1996; Tong and Thompson, 2008). Smaller pedigrees can be analyzed with exact computation, while larger pedigrees require Markov chain Monte Carlo (MCMC) sampling (Lange and Sobel, 1991; Thompson, 1994). It is important to note that full pedigree analysis based on IBD probability estimation can be substantially more powerful than approaches that split the pedigrees followed by exact IBD computation, thus emphasizing the need for use of MCMC sampling methods in the context of large pedigrees (Saint-Pierre *et al.*, 2014).

Careful sub-selection of markers is necessary for some analyses when dense markers are available on pedigrees (Santorico and Edwards, 2014). In a multipoint analysis, it is important to minimize linkage disequilibrium (LD) between markers to avoid inflated type I error rates (Abecasis *et al.*, 2002; Huang *et al.*, 2004; Schaid *et al.*, 2002; Webb *et al.*, 2005) and to comply with intrinsic assumptions in the computational algorithm (Lander and Green, 1987) used by most programs that carry out multipoint analyses. It is also important to minimize the number of markers used to reduce computational time, while retaining enough markers to achieve high information concerning inheritance patterns. It is not necessary to use all markers (Wilcox *et al.*, 2005). The block-based approach, which was developed for smaller pedigrees to circumvent the LD problem, is no longer as effective as it originally was due to the density of current marker panels coupled with the constraints on the block size (Abecasis *et al.*, 2002). All three goals can be achieved by judicious selection of a subset of markers from a dense panel of markers.

Most modern linkage analysis programs make use of inheritance vectors (IVs), which represent the flow of chromosome positions through pedigrees (Kruglyak *et al.*, 1996; Lander and Green, 1987; Lange and Sobel, 1991; Thompson, 2011). IVs are typically used internally for linkage analysis using the Lange–Sobel estimator (Lange and Sobel, 1991). MCMC estimation is employed by several linkage analysis programs [e.g. MORGAN, Simwalk2 and Superlink-Online single-nucleotide polymorphism (SNP)] for large pedigrees (Silberstein *et al.*, 2013; Sobel and Lange, 1996; Tong and Thompson, 2008). IV probabilities can also be used to estimate IBD sharing between pairs of individuals (Koepke and Thompson, 2013; Thompson, 2011). These probabilities may be stored for later reuse, providing significant computational advantages in downstream analyses. IVs may also be used to augment sequence data by genotype imputation (Abecasis *et al.*, 2002; Cheung *et al.*, 2013) and subsequently evaluated by family-based single-SNP or SNP-set association analyses (Chen *et al.*, 2013; Fridley *et al.*, 2010; Kang *et al.*, 2010; Saad and Wijsman, 2014).

Despite the availability of various computational tools for pedigree analyses, we still lack a unified comprehensive system that eases the crucial steps needed to work with dense markers. Older file manipulation and QC tools exist that were developed for use with sparse microsatellite or SNP panels (Abecasis *et al.*, 2002; Epstein *et al.*, 2000; McPeek and Sun, 2000; O'Connell and Weeks, 1998; Sun *et al.*, 2002) but are not always suitable or accurate for use with current, dense genotyping. File manipulation tools for dense data are available (Mukhopadhyay *et al.*, 2005; Purcell *et al.*, 2007), but preparing files can be challenging especially when studies involve data for large pedigrees or when they entail combining data from multiple file sources. At least two programs exist that begin to address selection of a subset of markers for linkage analysis: MArker SElection for Linkage (MASEL) (Bellenguez *et al.*, 2009) and LINKDATAGEN (Bahlo and Bromhead, 2009). MASEL selects SNPs based on a user-specified LD threshold and weights for several parameters (Bellenguez *et al.*, 2009) but does not impose a minimum intermarker distance (MID) and requires prior steps to filter SNPs. LINKDATAGEN performs error checks, removes all Mendelian inconsistent errors and selects SNPs based on user-specified bin size, minimum distance between markers and HapMap population-specific allele frequencies (Bahlo and Bromhead, 2009; Frazer *et al.*, 2007) but does not specifically use LD to select SNPs. Neither programs have options to force inclusion nor exclusion of specific markers. It is therefore possible to perform file manipulation, marker selection and QC checks by using multiple existing tools, but a unified system that performs all desired functions would be useful.

We have developed a pedigree-based analysis pipeline (PBAP) suite of programs that carry out file manipulations and QC checks prior to downstream analyses, with a focus on human data. PBAP prepares data files in a MORGAN-compatible format, carries out marker and pedigree QC analyses and automates marker selection for a good linkage panel while allowing the user considerable flexibility for inclusion or exclusion of markers. Results obtained after running programs of the MORGAN package (Thompson, 2011) on the formatted files can be used for various downstream analyses.

2 Methods

2.1 Description of PBAP

2.1.1 Data input files

PBAP, schematically illustrated in Figure 1, uses two project data input files: a pedigree file and a genotype file. An optional phenotype file may also be present. For the genotype file, two formats may be used: (i) a normal file format, e.g. PLINK *.ped (Purcell *et al.*, 2007), has the genotype data for individuals in rows (i.e. row: subject), as is typical in linkage analysis software (Abecasis *et al.*, 2002; Cottingham *et al.*, 1993; Lathrop *et al.*, 1984), while (ii) a transposed file format has genotype data for each individual in columns (i.e. row: marker), as is used in genotype imputation and in GWAS software (Browning and Browning, 2009; Marchini *et al.*, 2007; Purcell *et al.*, 2007). PBAP carries out further steps with the transposed file format, which allows processing of dense genotype data. Input files that are in the normal file format are first converted into the transposed file format. Details of these two file formats and ancillary files are provided in the PBAP documentation. We note that although the emphasis here is on use of SNP arrays, PBAP also allows inclusion of short tandem repeats (STRs) in a marker subpanel (described further below).

PBAP uses two main types of reference data input files: map files [e.g. PLINK *.map (Purcell *et al.*, 2007)] and main population genotype files [e.g. derived from the 1000 Genomes Project [1000 G] data (Altshuler *et al.*, 2010)]. We have constructed such files from public data, although PBAP users may construct and use any reference files of their choice that contain the necessary information. These reference files should be prepared prior to using PBAP. For the reference map files, we combined the Rutgers smoothed framework map and the Rutgers map of all dbSNP Build 134 variants (Matise *et al.*, 2007). On the merged map files, we converted

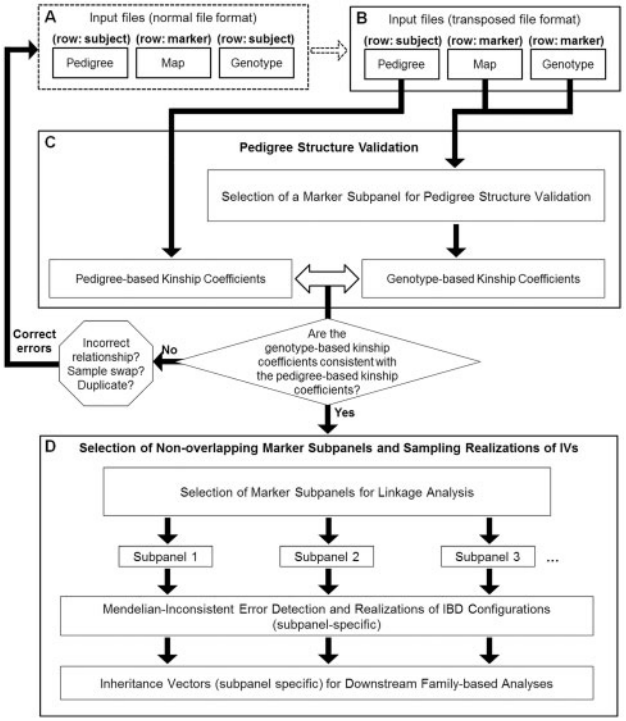

Fig. 1. Schematic diagram of PBAP. (**A** and **B**) Input files that are in the normal file format are first converted to the transposed file format. (**C**) Relationship or pedigree errors are identified. (**D**) One or more non-overlapping marker subpanels suitable for linkage analysis may be selected from the dense panels. The IVs for each marker subpanel are sampled by gl_auto of the MORGAN package (Thompson, 2011)

Kosambi distances between markers to Haldane distances, thus providing Haldane genetic locations (cM). For the reference genotype files, we downloaded 1000 G data. We used a custom Perl script, which is included in the PBAP release, to pre-process the 1000 G genotype files (i.e. to exclude indels and duplicate entries) for the different main populations [African (AFR), Admixed American (AMR), East Asian (ASN) and European EUR)] (Altshuler *et al.*, 2010). In all examples presented here, we used the genotype data for the European population. In selecting marker subpanels described below, PBAP requires pedigree files, these pre-processed reference genotype files, and uses PLINK (Purcell *et al.*, 2007) to calculate main population allele frequencies and to generate LD estimates.

2.1.2 Marker subpanels

We define a marker subpanel to be a subset of markers selected from an original dense marker panel. From a dense panel, one or more non-overlapping marker subpanels are selected in a stepwise manner. The goal is to obtain good marker subpanels, which contain highly informative markers for linkage analysis or IBD estimation (Koepke and Thompson, 2013; Thompson, 2011). These markers are chosen with three constraints: (i) to minimize LD among selected markers (Abecasis *et al.*, 2002; Huang *et al.*, 2004; Schaid *et al.*, 2002; Webb *et al.*, 2005), (ii) to be spaced at a distance that is compatible with MCMC-based approaches (Tong and Thompson, 2008; Wijsman *et al.*, 2006) and (iii) to minimize the number of markers chosen within constraints 1–2 to minimize computational time while retaining overall IBD information. For pedigree structure validation, which may be necessary as simple sample QC, only one marker subpanel across all chromosomes is needed.

For linkage analysis, two or more additional non-overlapping marker subpanels may be useful as a check that undetected genotype errors are not influential (Cheung *et al.*, 2014). PBAP can select five or more marker subpanels depending on the size of the dense panel. PBAP users may also pre-specify markers that should be excluded from the marker subpanel (i.e. exclusion markers) and/or included in the marker subpanel (i.e. inclusion markers). Inclusion/exclusion options are particularly useful when troubleshooting, comparing results from other studies or combining sparse STR marker panels with denser SNP panels. For the inclusion markers, users should specify whether a marker should always be included in the marker subpanel (i.e. core inclusion marker) or will be given priority but should pass thresholds for minimum MAF (in the reference dataset), minimum marker completion and should not be monomorphic in the dataset (i.e. auxiliary inclusion marker). PBAP users should specify the direction in which the markers will be processed [i.e. from upstream to downstream markers (forward direction) or from downstream to upstream markers (reverse direction)] and starting marker (i.e. markers upstream or downstream of this marker will be excluded when user chooses forward or reverse directions, respectively).

The selection of marker subpanels for pedigree structure validation comprises three processes. The first process excludes markers upstream or downstream of the user-specified starting marker and uses a fifth of the user-defined MID in centimorgans (cM) to create a distance-based reduced pre-subpanel to improve computational efficiency. For LD-based SNP pruning on this pre-subpanel (Purcell *et al.*, 2007), PBAP uses a maximum LD threshold (r^2) value specified by the user and a window size spanning ~1 cM that allows the user to obtain an LD-reduced pre-subpanel. The second process returns inclusion markers to this LD-reduced pre-subpanel. Core inclusion markers are permanently added into the marker subpanel, while auxiliary markers must pass thresholds specified above. The third process selects a set of markers for each of the marker subpanels using following criteria: (i) choice of main reference population to allow PBAP to use population allele frequencies appropriate for the dataset, (ii) minimum and maximum MAF in the reference population, (iii) minimum marker completion threshold of genotyped subjects, (iv) inclusion/exclusion of monomorphic markers in the dataset, (v) direction in which markers will be processed, (vi) starting marker and (vii) MID. In a real pedigree dataset, the realized MAF may be lower than it is in the reference population. PBAP selects markers based on criteria (i) through (iv) before using criteria (v) through (vii).

The same processes are also used for selecting marker subpanels for linkage analysis or marker IBD estimation with one additional step. This step identifies regions where adjacent markers in the pre-subpanel currently have a distance of ≥2 MID (i.e. gaps). These gaps are then filled in by allowing use of slightly less common variants (i.e. minimum MAF used is 0.05 lower than user-defined minimum MAF) and a slightly smaller MID (i.e. ~0.1 cM smaller than the user-defined MID) but still using the same maximum MAF and same thresholds for marker completion and r^2. Filling in these gaps minimizes loss of information that may be important for downstream analyses. PBAP cumulatively stores markers selected and excludes them in subsequent subpanels to obtain non-overlapping subpanels. These criteria or parameters are listed in Table 1.

2.1.3 Pedigree structure validation

Relationship or sample errors are identified by comparing theoretical coefficients based on pedigree structure (Karigl, 1981) with the

Table 1. Various PBAP applications and criteria/parameters

Applications	Criteria/parameters
Transposition of normal file format	Marker inclusion/exclusion
	Family ID/individual ID substitution
Selection of marker subpanels	LD (r^2)
	MAF
	Marker completion
	Monomorphic markers
	Direction of marker processing
	Starting marker
	MID (cM)
	Main population
	Source of MAF information
	Types of markers with genotype data
	Number of marker subpanels
	STRs as core inclusion markers
	Gap filling (fixed values)
Genotype-based kinship estimation	Source of MAF information
Preparation of files for MORGAN	L-sampler probability
	Number of burn-in iterations
	Number of MCMC iterations
	Number of IV realizations saved

coefficients estimated using genotype data. PBAP uses a likelihood-based estimator for computing kinship coefficients ($\hat{\varphi}$) and genotype-based identity by descent probabilities of sharing one allele ($\hat{k}_1$) (Choi *et al.*, 2009). Likelihood estimators are more efficient than moment estimators (Anderson and Weir, 2007), and unlike moment estimators that require dense arrays to achieve equivalent accuracy, a likelihood estimator is reasonably accurate for use even with SNP arrays that may sometimes consist of only sparse linkage panels (Milligan, 2003). In the process of calculating φ and k_1 based on the pedigree structure, PBAP also identifies all pairs of simple relationships as close as or closer than great-grandparent-great-grandchild. Using empirical confidence intervals (CIs; see below) for both $\hat{k}_1$ and $\hat{\varphi}$ for a particular pair of individuals, PBAP compares pedigree-based pairwise k_1 and φ and genotype-based pairwise estimated $\hat{k}_1$ and $\hat{\varphi}$ to detect possible relationship errors. Once a potential error is detected, the user must decide whether or not to correct it in the pedigree and genotype input files (Fig. 1).

To generate empirical CIs for both $\hat{k}_1$ and $\hat{\varphi}$, we performed several steps. First, we created a dataset that contains all pairs of simple relationships, i.e. as close as or closer than great-grandparent-great-grandchild. Second, we used allele frequencies of Illumina 6K array markers from a real dataset (Allen-Brady *et al.*, 2010; Cannon *et al.*, 2010; Coon *et al.*, 2010) and simulated 10 000 replicates of full marker panels genome-wide on multiple copies of the dataset using the program genedrop from the MORGAN package (Thompson, 2011). We represented 3 K, 6 K and 9 K panels, respectively, by 2819, 5624 and 8924 markers chosen from this Illumina 6 K array. For the additional 3300 markers in the 9 K panel, we randomly selected allele frequencies of markers from the same Illumina 6 K array. We used the three panels to have good representation of a genome-wide evenly spaced linkage panel (i.e. contains 5000–6000 markers). Third, we estimated the three identity coefficients ($\hat{k}_0$, $\hat{k}_1$ and $\hat{k}_2$) from these simulated genotypes, where $\hat{k}_i$ is the probability of sharing i alleles IBD. Using these three identity coefficients, values for $\hat{\varphi}$ were estimated (Choi *et al.*, 2009). Last, the empirical CIs of $\hat{k}_1$ and $\hat{\varphi}$ for different numbers of panel markers in each of the relationship categories were estimated based on the minimum and maximum values obtained for the three panels above (Fig. 2; Supplementary Figs S1 and S2) using a linear model with quadratic coefficients. The combined set of empirical CIs for each pair of simple relationships mentioned above for k_1 and φ, whose expectations are the pedigree-based coefficients, are unique and are therefore sufficient to distinguish these relationships from one another.

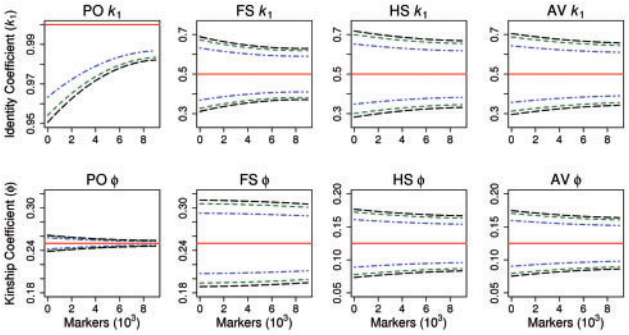

Fig. 2. Empirical CIs for genotype-based identity ($\hat{k}_1$) and kinship ($\hat{\varphi}$) coefficients. Horizontal lines represent the expected k_1 and φ for each of the different pairwise relationships. PO, parent–offspring; FS, full sibling; HS, half sibling; AV, avuncular. Dash-dots, dashes and long dashes represent the 95%, 99% and 99.5% CIs, respectively. See Supplementary Figures S1 and S2 for more comprehensive set of close relationships

2.1.4 Sampling realizations of IVs

For each marker subpanel generated, PBAP prepares MORGAN-compatible files and executes the program gl_auto of the MORGAN package (Thompson, 2011) to sample realizations of IVs at each subpanel marker location and to check for Mendelian inconsistencies. When found, PBAP zeroes out the genotypes of all individuals in that pedigree for that particular marker. There are three major advantages to use MORGAN for pedigree computations. First, use of gl_auto for preliminary analyses dramatically speeds up subsequent analyses. Second, MORGAN handles large and complex (e.g. with cross generational marriages, inbreeding or marriage loops) pedigrees with MCMC-based computations faster and more accurately than other programs (Wijsman *et al.*, 2006). Third, MORGAN combines exact and MCMC computation, allowing exact computation for smaller pedigrees without user intervention. The sampled IVs are then available for use in a multitude of possible downstream analyses. As part of the file preparation with PBAP, the user specifies conditions required for gl_auto (Table 1). PBAP calculates an appropriate number of burn-in iterations and executes gl_auto to perform the analysis. Sampled IVs are pedigree- and chromosome specific.

2.2 Evaluation of PBAP

To illustrate the performance of PBAP, we used two real autism spectrum disorder (ASD) datasets that comprise (i) 71 small to large pedigrees (ASD1) (2–9 generations) (Allen-Brady *et al.*, 2010; Cannon *et al.*, 2010; Coon *et al.*, 2010) and (ii) a five-generation large pedigree with evidence of linkage to chromosome 22 (Marchani *et al.*, 2012) (ASD2). We used genome-wide data of ASD1 and ASD2 to demonstrate pedigree structure validation and chromosome 22 data of ASD2 to demonstrate marker subpanel selection. For ASD1, we had SNP data comprising markers in the Illumina 6 K and Illumina OmniExpress arrays where we used the latter array for pedigree structure validation in the presence of sample swaps described below. For ASD2, we had dense marker panel

data comprising markers in the Illumina OmniExpress array. We carried out all steps described above with PBAP on both ASD1 and ASD2 including (i) formatting the input files, (ii) validating relationships, (iii) generating marker subpanels suitable for linkage analysis from either the OmniExpress array or the 6 K array and (iv) running gl_auto to sample realizations of IVs. All of these steps take a few seconds or minutes to run with two exceptions. First, while calculation of genotype-based kinship coefficients in ASD1 (195 individuals, 5789 markers) and ASD2 (42 individuals, 5678 markers) took only ~5 min and ~18 s using a single core on a multi-core 2.26 GHz Intel Xeon computer, respectively, execution time is proportional to the number of markers and to the square of the number of genotyped individuals and thus increases rapidly with sample size. Second, selection of one marker subpanel from the ASD2 panel of ~3500 markers on chromosome 22 used here took ~39 s but execution time increases with each additional non-overlapping subpanel generated (e.g. obtaining five non-overlapping subpanels from this panel took about ~4 min). In all of the analyses where PLINK was needed, we used PLINK v1.07. For comparison to an alternative, we also carried out marker selection using the marker subpanel selection tool, MASEL (Bellenguez *et al.*, 2009).

2.2.1 Pedigree structure validation

We performed pedigree structure validation on ASD1 and ASD2. For each dataset, we selected one marker subpanel using the following criteria: (i) MID ≥ 0.5 cM, (ii) marker completion $\geq 80\%$, (iii) $r^2 \leq 0.25$ (less stringent since kinship estimation is not strongly affected by LD), (iv) $0.3 \leq \text{MAF} \leq 0.5$ in 1000 G data since more common variants provide better kinship estimates and (v) starting marker. In all subpanel selections performed in this article, we excluded markers that were monomorphic in the dataset. Based on the specified pedigree structures of ASD1 (Illumina 6 k panel) and ASD2, we used PBAP to calculate the pedigree-based φ and to identify close relatives. We then computed genotype-based $\hat{k}_1$ and $\hat{\varphi}$ for all pairs with genotype data. Pedigree- and genotype-based coefficients were compared to detect possible relationship errors. In addition, we created four different sample swaps simultaneously in three families of ASD1 (Illumina OmniExpress array) to further illustrate pedigree QC with PBAP. The relationship errors introduced were (i) parent–offspring, (ii) first-cousin, (iii) full-sibling-unrelated and (iv) parent-avuncular sample swaps. We created a full-sibling-unrelated swap between a full-sibling (FS) and a sister-in-law, and a parent-avuncular swap between an FS pair where at least one of the siblings has one or more offspring.

2.2.2 Generating marker subpanels and performing linkage analysis

Using PBAP, we selected three non-overlapping marker subpanels to probe possible influential errors. We used the same MID and marker completion criteria used for pedigree structure validation but with $r^2 \leq 0.04$ (more stringent to minimize LD between markers), $0.2 \leq \text{MAF} \leq 0.5$ in 1000 G data (common variants with lower minimum to allow inclusion of more markers) and different starting markers (any number from 1 to 10) to facilitate variability. Since chromosome 22 spans ~80 cM, a target intermarker distance of ~0.5 cM results in a maximum of ~120–160 markers, as was achieved using PBAP. Since MASEL also uses LD as a criteria in selecting marker subpanels, we compared MASEL with PBAP. To achieve a fair comparison, we pre-filtered the SNPs prior to using MASEL through the same MAF and marker completion criteria as for PBAP. To select marker subpanels for linkage analysis using MASEL, Bellenguez *et al.* (2009) used four different weighting schemes: (1,1,1), (2,1,1), (1,2,1) and (1,1,2), representing different choices for weighting heterozygosity, intermarker distance and set size (i.e. number of SNPs removed due to LD or distance when a particular SNP is selected), respectively, in selecting markers. For example, for the first weighting scheme, MASEL uses equal weights for heterozygosity, intermarker distance and set size. Although distance is one of the parameters used by MASEL, the program does not allow the user to choose a specific MID. Using the same set of four weighting schemes and an $r^2 \leq 0.04$ (as for PBAP), we selected four different marker subpanels using MASEL. This resulted in 400–500 markers spanning chromosome 22 with many very close adjacent markers (0.002 cM). To obtain a sparser marker subpanel with MASEL, we used a lower maximum r^2 of 0.01 to select another set of three and four marker subpanels using both PBAP and MASEL, respectively. For each of these 14 marker subpanels (seven with r^2 of 0.04 and seven with r^2 of 0.01), we used PBAP and gl_auto to perform computations. For every marker subpanel, each analysis included 2500 burn-in iterations, sampling by scan, generation of 50 000 MCMC iterations, 50% L-sampler and saving 1000 realizations of IVs for inference.

To compare the impact of marker subpanels generated by MASEL versus PBAP, we carried out linkage analysis with the IVs obtained from gl_auto by running PBAP using each marker subpanel, followed by the use of gl_lods of the MORGAN package (Koepke and Thompson, 2013; Thompson, 2011). For this purpose, we used data from a genome scan that identified positive evidence of linkage ($\text{lod}_{\max} = 1.8$) in linkage analysis on chromosome 22 for the ASD2 dataset and used the same reduced-penetrance dominant model used previously (Marchani *et al.*, 2012). The program gl_lods makes use of equivalence classes of locus-specific IVs (Koepke and Thompson, 2013) to avoid redundant calculation, thus dramatically increasing speed in the lod score computations.

3. Results

3.1 Reference data files and pedigree structure validation

We used the reference map files together with the 1000 G reference genotype files to select a set of non-overlapping marker subpanels with PBAP (Table 2). We also used the allele frequencies from the pre-processed 1000 G data in analysis with gl_auto to obtain IVs. We compared the pedigree- and genotype-based $\hat{k}_1$ and $\hat{\varphi}$ of the ASD1 and ASD2 datasets using PBAP. For FS pairs, there was one detected MZ twin pair in ASD1, while no relationship errors were detected in ASD2. For the four different sample swaps created in ASD2, PBAP flagged extreme points outside the empirical CIs for these parameters in all cases as shown in the different k_0–k_1 plots (Fig. 3), illustrating its ability to detect these types of errors.

3.2 Marker subpanels generated through PBAP or MASEL

Characteristics of the marker subpanels that were generated using PBAP and MASEL are listed in Table 2 as ranges of values from three and four subpanels, respectively. Marker completion and MAF for all 14 marker subpanels from ASD2 are $\geq 95\%$ and ≥ 0.15, respectively. Minimum MAFs of marker subpanels with MASEL (0.20–0.21) are relatively higher than with PBAP (0.15–0.17), since PBAP lowers MAF in regions that require gap filling. MAF (ASD2 dataset) for all 14 marker subpanels is ≥ 0.012, which is lower than the user-specified minimum MAF.

Table 2. Characteristics of marker subpanels generated through PBAP and MASEL

Program	Maximum LD (r^2)	Parameter	Number of markers	Mean	Standard deviation	Minimum	Maximum
PBAP	0.04	ID (cM)	124–148	0.532–0.646	0.169–0.219	0.388–0.401	1.455–1.828
		MAF (1000 G)		0.311–0.337	0.088–0.098	0.150–0.172	0.489–0.500
		MAF (ASD2)		0.279–0.308	0.114–0.127	0.012–0.048	0.488–0.500
MASEL	0.04	ID (cM)	468–491	0.161–0.169	0.169–0.180	0.002–0.002	1.455–1.494
		MAF (1000 G)		0.395–0.403	0.082–0.083	0.201–0.201	0.500–0.500
		MAF (ASD2)		0.345–0.348	0.106–0.108	0.012–0.024	0.500–0.500
MASEL (thinned)	0.04	ID (cM)	156–164	0.484–0.509	0.338–0.360	0.013–0.051	1.633–1.976
		MAF (1000 G)		0.390–0.405	0.078–0.087	0.201–0.207	0.500–0.500
		MAF (ASD2)		0.335–0.361	0.101–0.111	0.036–0.048	0.500–0.500
PBAP	0.01	ID (cM)	115–137	0.572–0.691	0.212–0.239	0.400–0.401	1.455–1.961
		MAF (1000 G)		0.312–0.336	0.090–0.101	0.150–0.165	0.499–0.500
		MAF (ASD2)		0.278–0.313	0.116–0.130	0.012–0.024	0.488–0.500
MASEL	0.01	ID (cM)	79–85	0.935–0.999	0.585–0.771	0.036–0.070	3.158–4.151
		MAF (1000 G)		0.411–0.437	0.066–0.082	0.210–0.214	0.500–0.500
		MAF (ASD2)		0.345–0.376	0.094–0.114	0.024–0.107	0.488–0.500

ID, intermarker distance; MAF, minor allele frequency; 1000G, 1000 Genomes Project data (Altshuler *et al.*, 2010); ASD2, autism spectrum disorder dataset 2 (Marchani *et al.*, 2012)

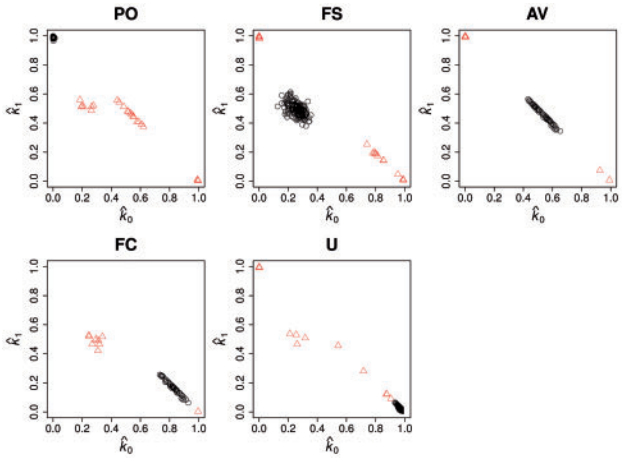

Fig. 3. Scatterplot of genotype-based identity coefficients ($\hat{k}_0$ and $\hat{k}_1$) in the presence of sample swaps. Coefficients within CI (correct relationships) are represented by circles while extreme points outside the empirical CIs that were flagged by PBAP, which are also the incorrect relationships that we created, are represented by triangles. PO, parent–offspring; FS, full sibling; AV, avuncular; FC, first cousins; U, unrelated

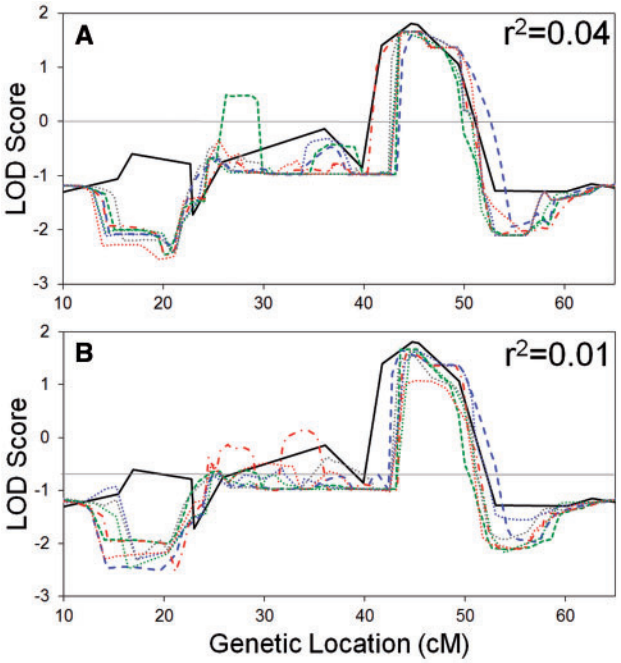

Fig. 4. Linkage analysis using PBAP versus MASEL marker subpanels. (**A**) $r^2 = 0.04$. (**B**) $r^2 = 0.01$. Solid lines represent the previous linkage analysis results using an STR panel (Marchani *et al.*, 2012). Short dashes, long dashes and dash-dots represent the linkage analysis results using PBAP marker subpanels 1, 2 and 3, respectively. Dotted lines represent the linkage analysis results using MASEL marker subpanels

Since we used the same threshold for pre-filtering SNPs, dataset marker completion and MAF were all within the ranges specified for both types of panels. However, we obtained differences in the mean intermarker distances between marker subpanels obtained by PBAP and MASEL. For an r^2 of 0.04, the mean intermarker distances of PBAP and MASEL marker subpanels are 0.53–0.65 cM (range: 0.39–1.83 cM) and 0.16–0.17 cM (range: 0.002–1.49 cM), respectively. After thinning, these MASEL marker subpanels mean intermarker distances increased to 0.48–0.51 cM (range: 0.01–1.98 cM) similar to that from PBAP, but there are markers that are still near each other even after thinning. For an r^2 of 0.01, the mean intermarker distances of PBAP and MASEL marker subpanels are 0.57–0.69 cM (range: 0.4–1.96 cM) and 0.94–1.0 cM (range: 0.04–4.15 cM), respectively. Note that the mean intermarker distances of the PBAP marker subpanels are near 0.5 cM for both r^2 of 0.04 and 0.01, while that of the MASEL marker subpanels varied from 0.16 to 1.0 cM. Moreover, maximum intermarker distances (cM) of the PBAP marker subpanels are <2 cM, with the larger gaps due to regions where there are no markers in the dense panel, whereas the larger gaps of MASEL marker subpanels had maximum intermarker distances >4 cM (Table 2) with no way to prevent this.

3.3 Linkage analysis using sampled IVs

Both PBAP and thinned MASEL subpanels generated by using a maximum r^2 of 0.04 performed well in the region with the linkage signal (40–53 cM) (Fig. 4A). Among all panels, PBAP subpanel 3 almost matched the original linkage signal obtained with STRs, which was fully informative at the STRs in the peak. Although most of the panels give similar lod scores across the chromosome, PBAP

subpanel 2 has a higher signal compared with the rest of the panels at positions ~25–30 cM. The marker(s) that provide more information against linkage may have not been in this subpanel. Moreover, all panels have more negative lod scores from ~12–22 cM and ~53–60 cM compared with the original linkage lod scores in regions between two STRs.

A similar trend was observed in marker subpanels obtained by PBAP and MASEL using a maximum r^2 threshold of 0.01 (Fig. 4B) where almost all subpanels performed relatively well at the linkage signal. However, with fewer markers, there was no longer a subpanel that almost matched the original linkage signal. From position ~26–40 cM, where there was a small set of negative lod scores with portions that are less negative in the original signal: the PBAP marker subpanel 3 has two of this type of small set, one of which has positive lod scores. In addition, all marker subpanels also have more negative lod scores from ~12–22 cM and ~53–60 cM compared with the original linkage lod scores. At this r^2 threshold, MASEL marker subpanels generated with different weighting schemes give more variable linkage results at the highest original linkage signal (~40–53 cM), while the non-overlapping PBAP subpanels still gave relatively consistent results.

4 Discussion

The use of dense markers in pedigree-based data analyses is challenging and requires nontrivial file manipulation. Selection of subsets of markers, i.e. marker subpanels, from a dense panel can be a complex process based on joint optimization across several parameters. Here, we introduce PBAP, a pipeline of programs that automatically selects ideal subpanel(s) of markers from the complete available panel and performs basic QC steps. One advantage of PBAP is its flexibility, unlike some other existing programs that allow file preparation and analysis of pedigrees but have constraints in how they select marker subpanels from dense panels. Finally, PBAP prepares files in a format suitable for pedigree-based analyses (i.e. MORGAN file format) and accesses software (i.e. gl_auto) (Thompson, 2011) to generate information suitable for pedigree-based analyses involving IBD information. This is particularly useful when samples include large pedigrees where estimation through MCMC is necessary, adding additional constraints to selection of marker subpanels. It is important to select an informative marker subpanel from a dense panel of SNPs to perform pedigree-based analyses involving IBD. PBAP automatically selects a subpanel with adequate flexibility for the user so that differences in dense panels and evolving genotyping technologies should not dramatically affect the results. In particular, PBAP includes specification of constraints needed to use methods that allow analysis of extended, as well as smaller, pedigrees. One of the main goals in selecting marker subpanels is to minimize significant LD between SNPs. The strategy used by PBAP avoids the need for considerable data and computing resources imposed by trying to model LD during the analysis and also scales easily to use of much denser data, such as is present in high throughput sequencing. A second goal was to facilitate use of computational methods that allow intact large pedigrees. As this currently requires MCMC, PBAP controls the MID and MAF chosen. Moreover, PBAP internally pre-filters SNPs based on user-specified MID, dataset marker completion and MAF range. With minute intermarker distances, MCMC-based methods suffer (Sieh *et al.*, 2005; Thompson and Heath, 1999; Wilcox *et al.*, 2005).

Similar results were observed from linkage analysis using both MASEL and PBAP marker subpanels. A difference was the amount of prior manipulation required to use MASEL for selecting subpanels, which is done internally by PBAP based on user-specified parameters. Comparable linkage results for both PBAP and MASEL subpanels may, in part, be due to the fact that ASD2 is an example of well-genotyped complex pedigree, where problematic MCMC mixing due to very small intermarker distances, and loss of information due to very large intermarker distances may not have a large impact on the linkage results. An important additional feature of PBAP is that it allows generation of non-overlapping subpanels, which may facilitate detection of influential genotyping errors based on the linkage results (Cheung *et al.*, 2014). In addition, achieving similar linkage results from non-overlapping subpanels strengthens the evidence for a linkage signal and absence of a genotype error on a particular genomic region.

A dataset may or may not have already undergone some level of QC before using PBAP. Our pipeline performs basic QC checks to ensure detection of errors that may have been missed by other QC procedures. When the dataset has relatively few genotyped individuals, there are advantages to the use of external population allele frequency estimates, and when the number of available markers common across subjects is modest, a more accurate likelihood-based estimator is a safer choice for relationship estimation than is a moment estimator. Pedigree structure validation in PBAP is useful in distinguishing close relationship errors, which is sufficient (in most cases) to detect sample swaps and misspecified relationships. Some caution is needed to interpret flagged relationships since these may or may not indicate a true relationship error. Confirmed relationship errors should be corrected by the user and may entail dropping certain individuals, editing the pedigree and/or genotype files and going back to the beginning of the pipeline.

Although optimized for human pedigrees, PBAP may also be used for other organisms as long as there is no selfing involved. The pedigree structures of other organisms may be larger and more complex than in humans but this does not provide a barrier. Except for the pedigree structure validation step, which would require use of other existing software, PBAP may be applied.

We have presented a unified comprehensive pipeline that performs critical steps to handle dense genotype data and prepare files for family-based downstream analyses. PBAP selects a good panel of sparse markers and samples IVs. This framework allows the development of additional modules that will carry out a variety of computations given the sampled IVs as illustrated though our computation of lod scores. PBAP thus enables a variety of analyses on large pedigrees with dense markers with the opportunity to extend to other programs for downstream analyses.

Acknowledgement

We acknowledge discussions with Mohamad Saad and Sulgi Kim.

Funding

This work was supported by the National Institutes of Health (R01 MH092367, R01 MH094293, R01 AG039700, R37 GM046255, P50 AG005136 and U01 AG016976).

Conflict of Interest: none declared.

References

Abecasis,G.R. *et al.* (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.*, **30**, 97–101.

Allen-Brady,K. *et al.* (2010) Genome-wide linkage in Utah autism pedigrees. *Mol. Psychiatry*, **15**, 1006–1015.

Altshuler,D. *et al*. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Anderson,A.D. and Weir,B.S. (2007) A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics*, **176**, 421–440.

Bahlo,M. and Bromhead,C.J. (2009) Generating linkage mapping files from Affymetrix SNP chip data. *Bioinformatics*, **25**, 1961–1962.

Bellenguez,C. *et al*. (2009) Linkage analysis with dense SNP maps in isolated populations. *Hum. Hered.*, **68**, 87–97.

Browning,B.L. and Browning,S.R. (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.*, **84**, 210–223.

Cannon,D.S. *et al*. (2010) Genome-wide linkage analyses of two repetitive behavior phenotypes in Utah pedigrees with autism spectrum disorders. *Mol. Autism*, **1**, 3.

Chen,H. *et al*. (2013) Sequence kernel association test for quantitative traits in family samples. *Genet. Epidemiol.*, **37**, 196–204.

Cheung,C.Y.K. *et al*. (2013) GIGI: an approach to effective imputation of dense genotypes on large pedigrees. *Am. J. Hum. Genet.*, **92**, 504–516.

Cheung,C.Y.K. *et al*. (2014) Detection of Mendelian consistent genotyping errors in pedigrees. *Genet. Epidemiol.*, **38**, 291–299.

Choi,Y. *et al*. (2009) Case-control association testing in the presence of unknown relationships. *Genet. Epidemiol.*, **35**, 668–678.

Coon,H. *et al*. (2010) Genome-wide linkage using the Social Responsiveness Scale in Utah autism pedigrees. *Mol Autism*, **1**, 8.

Cottingham,R.W. *et al*. (1993) Faster sequential genetic-linkage computations. *Am. J. Hum. Genet.*, **53**, 252–263.

Epstein,M.P. *et al*. (2000) Improved inference of relationship for pairs of individuals. *Am. J. Hum. Genet.*, **67**, 1219–1231.

Fan,Y.H. and Song,Y.Q. (2012) IPGWAS: an integrated pipeline for rational quality control and association analysis of genome-wide genetic studies. *Biochem. Biophys. Res. Commun.*, **422**, 363–368.

Frazer,K.A. *et al*. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–856.

Fridley,B.L. *et al*. (2010) Utilizing genotype imputation for the augmentation of sequence data. *PLoS One*, **5**, e11018.

Fuchsberger,C. *et al*. (2012) GWAtoolbox: an R package for fast quality control and handling of genome-wide association studies meta-analysis data. *Bioinformatics*, **28**, 444–445.

Gogarten,S.M. *et al*. (2012) GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics*, **28**, 3329–3331.

Gudbjartsson,D.F. *et al*. (2000) Allegro, a new computer program for multipoint linkage analysis. *Nat. Genet.*, **25**, 12–13.

Heath,S.C. (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am. J. Hum. Genet.*, **61**, 748–760.

Hinrichs,A.L. and Suarez,B.K. (2011) Incorporating linkage information into a common disease/rare variant framework. *Genet. Epidemiol.*, **35**, S74–S79.

Huang,Q.Q. *et al*. (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.

Kang,H.M. *et al*. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348–U110.

Karigl,G. (1981) A recursive algorithm for the calculation of identity coefficients. *Ann. Hum. Genet.*, **45**, 299–305.

Keramati,A.R. *et al*. (2014) A form of the metabolic syndrome associated with mutations in DYRK1B. *N. Engl. J. Med.*, **370**, 1909–1919.

Koepke,H. and Thompson,E. (2013) Efficient identification of equivalences in dynamic graphs and pedigree structures. *J. Comput. Biol.*, **20**, 551–570.

Kruglyak,L. *et al*. (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.*, **58**, 1347–1363.

Lambert,G. *et al*. (2013) Single nucleotide polymorphism genotyping using BeadChip microarrays. *Curr. Protoc. Hum. Genet.*, **Chapter 2**, Unit 2 9.

Lander,E.S. and Green,P.J. (1987) Construction of multilocus genetic maps in humans. *Proc. Natl. Acad. Sci. USA*, **84**, 2363–2367.

Lange,K. and Sobel,E. (1991) A random walk method for computing genetic location scores. *Am. J. Hum. Genet.*, **49**, 1320–1334.

Lathrop,G.M. *et al*. (1984) Strategies for multilocus linkage analysis in humans. *Proc. Natl. Acad. Sci. USA*, **81**, 3443–3446.

Laurie,C.C. *et al*. (2010) Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.*, **34**, 591–602.

Marchani,E.E. *et al*. (2012) Identification of rare variants from exome sequence in a large pedigree with autism. *Hum. Hered.*, **74**, 153–164.

Marchini,J. *et al*. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, **39**, 906–913.

Mardis,E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, **9**, 387–402.

Mardis,E.R. (2011) A decade's perspective on DNA sequencing technology. *Nature*, **470**, 198–203.

Matise,T.C. *et al*. (2007) A second-generation combined linkage-physical map of the human genome. *Genome Res.*, **17**, 1783–1786.

Matise,T.C. *et al*. (2011) The next PAGE in understanding complex traits: design for the analysis of population architecture using genetics and epidemiology (PAGE) study. *Am. J. Epidemiol.*, **174**, 849–859.

McPeek,M.S. and Sun,L. (2000) Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am. J. Hum. Genet.*, **66**, 1076–1094.

Metzker,M.L. (2010) Sequencing technologies—the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

Milligan,B.G. (2003) Maximum-likelihood estimation of relatedness. *Genetics*, **163**, 1153–1167.

Mukhopadhyay,N. *et al*. (2005) Mega2: data-handling for facilitating genetic linkage and association analyses. *Bioinformatics*, **21**, 2556–2557.

Musunuru,K. *et al*. (2010) Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *N. Engl. J. Med.*, **363**, 2220–2227.

O'Connell,J.R. and Weeks,D.E. (1998) PedCheck: a program for identification of genotype incompatibilities in linkage analysis. *Am. J. Hum. Genet.*, **63**, 259–266.

Patel,R.K. and Jain,M. (2012) NGS QC toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One*, **7**, e30619.

Pongpanich,M. *et al*. (2010) A quality control algorithm for filtering SNPs in genome-wide association studies. *Bioinformatics*, **26**, 1731–1737.

Purcell,S. *et al*. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Ragoussis,J. (2009) Genotyping technologies for genetic research. *Annu. Rev. Genomics Hum. Genet.*, **10**, 117–133.

Ritchie,M. *et al*. (2010) Quality control pipeline for genome-wide association studies in the eMERGE network: comparing single site QC to a merged QC approach. *Genet. Epidemiol.*, **34**, 957–957.

Rosenthal,E.A. *et al*. (2011) Linkage and association of phospholipid transfer protein activity to LASS4. *J. Lipid Res.*, **52**, 1837–1846.

Saad,M. and Wijsman,E.M. (2014) Power of family-based association designs to detect rare variants in large pedigrees using imputed genotypes. *Genet. Epidemiol.*, **38**, 1–9.

Saint-Pierre,A. *et al*. (2014) SNP-based linkage analysis in extended pedigrees: comparison between two alternative approaches. *Hum. Hered.*, **78**, 27–37.

Santorico,S.A. and Edwards,K.L. (2014) Challenges of linkage analysis in the era of whole-genome sequencing. *Genet. Epidemiol.*, **38**, S92–S96.

Schadt,E.E. *et al*. (2010) A window into third-generation sequencing. *Hum. Mol. Genet.*, **19**, R227–R240.

Schaid,D.J. *et al*. (2002) Caution on pedigree haplotype inference with software that assumes linkage equilibrium. *Am. J. Hum. Genet.*, **71**, 992–995.

Shendure,J. and Ji,H.L. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.

Sieh,W. *et al*. (2005) Comparison of marker types and map assumptions using Markov chain Monte Carlo-based analysis of COGA data. *BMC Genet.*, **6**(Suppl. 1), S11.

Silberstein,M. *et al*. (2013) A system for exact and approximate genetic linkage analysis of SNP data in large pedigrees. *Bioinformatics*, **29**, 197–205.

Sobel,E. and Lange,K. (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am. J. Hum. Genet.*, **58**, 1323–1337.

Sun,L. *et al.* (2002) Enhanced pedigree error detection. *Hum. Hered.*, **54**, 99–110.

Thompson,E.A. (1994) Monte Carlo likelihood in the genetic mapping of complex traits. *Philos. Trans. R. Soc. Lond. Ser. B*, **344**, 345–351.

Thompson,E.A. (2011) The structure of genetic linkage data: from LIPED to 1 M SNPs. *Hum. Hered.*, **71**, 86–96.

Thompson,E.A. and Heath,S.C. (1999) Estimation of conditional multilocus gene identity among relatives. In: Seillier-Moseiwitch, F., Donnelly, P. and Waterman, M. (eds.) *Statistics in Molecular Biology and Genetics: Selected Proceedings of the 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics in Molecular Biology*. Institute of Mathematical Statistics, Hayward, CA, pp. 93–113.

Tong,L.P. and Thompson,E. (2008) Multilocus lod scores in large pedigrees: combination of exact and approximate calculations. *Hum. Hered.*, **65**, 142–153.

Wang,Z. *et al.* (2013) The role and challenges of exome sequencing in studies of human diseases. *Front. Genet.*, **4**, 160.

Webb,E.L. *et al.* (2005) SNPLINK: multipoint linkage analysis of densely distributed SNP data incorporating automated linkage disequilibrium removal. *Bioinformatics*, **21**, 3060–3061.

Wijsman,E.M. (2012) The role of large pedigrees in an era of high-throughput sequencing. *Hum. Genet.*, **131**, 1555–1563.

Wijsman,E.M. *et al.* (2006) Multipoint linkage analysis with many multiallelic or dense diallelic markers: MCMC provides practical approaches 'for genome scans on general pedigrees. *Am. J. Hum. Genet.*, **79**, 846–858.

Wilcox,M.A. *et al.* (2005) Comparison of single-nucleotide polymorphisms and microsatellite markers for linkage analysis in the COGA and simulated data sets for Genetic Analysis Workshop 14: presentation groups 1, 2, and 3. *Genet. Epidemiol.*, **29**(Suppl. 1), S7–S28.

Zhao,Y. *et al.* (2013) Exome sequencing and linkage analysis identified tenascin-C (TNC) as a novel causative gene in nonsyndromic hearing loss. *PLoS One*, **8**, e69549.

Zhou,Q. *et al.* (2013) QC-Chain: fast and holistic quality control method for next-generation sequencing data. *PLoS One*, **8**, e60234.