Here we propose a simple statistical algorithm for rapidly scoring loci associated with disease or traits due to recessive mutations or deletions using genome-wide single nucleotide polymorphism genotyping case–control data in unrelated individuals. This algorithm identifies loci by defining homozygous segments of the genome present at significantly different frequencies between cases and controls. We found that false positive loci could be effectively removed from the output of this procedure by applying different physical size thresholds for the homozygous segments. This procedure is then conducted iteratively using random sub-datasets until the number of selected loci converges. We demonstrate this method in a publicly available data set for Alzheimer’s disease and identify 26 candidate risk loci in the 22 autosomes. In this data set, these loci can explain 75% of the genetic risk variability of the disease.

**Introduction**

Advances in whole-genome single nucleotide polymorphism (SNP) assay technology have provided a powerful array of tools for simultaneously scoring common genetic variation. However, it is often difficult to identify loci associated with disease because of the large number of tests carried out and the associated conservative multiplicity adjustment, such as Bonferroni method. We are interested in identifying such loci associated with a disease likely due to recessive mutation or gene deletions.

High density SNP analysis readily reveals the presence of large homozygous segments in unrelated subjects (Hinds et al., 2005; Simon-Sanches et al., 2007; Wang et al., 2007). The probability of a randomly selected SNP locus being homozygous (‘AA’ or ‘BB’) based on data from HapMap is about 0.65 (Hinds et al., 2005; Rabbee and Speed, 2006) and this may lend itself to autozygosity mapping in ostensibly outbred populations; however, traditional autozygosity mapping methods (Lander and Botstein, 1987; Mueller and Bishop, 1993; Gschwend et al., 1996) based on consanguineous relationships are not appropriate for unrelated individuals. To identify loci with possible recessive effects of relatively high penetrance in outbred populations, large sample sizes are needed for genotyping. Some recent studies on homozygosity analysis of SNP assays have been attempted using different approaches (Woods et al., 2004; Lencz et al., 2007; Miyazawa et al., 2007). However, they either have some familial relationship requirements (Woods et al., 2004; Miyazawa et al., 2007) or a high false positive rate (Lencz et al., 2007).

In the context of SNP genotyping, it is often not easy to distinguish heterozygous genomic deletion from homozygosity; thus a segment with all loci genotyped being ‘AA’ or ‘BB’ in a pedigree genotype file could be either a region of genuine homozygosity or effective hemizygosity caused by genomic deletion. We call such a region ‘apparently homozygous region’ (AH). By carrying out an appropriate association analysis on AHs, one can detect not only the possible recessively mutated loci from some common ancestor but also deletions (Hunter, 2005; Klein et al., 2005; Van Eyken et al., 2007).
In this paper, we propose a simple statistical algorithm for genome-wide AH analysis (GAHA) of case–control data in unrelated subjects. It can robustly identify loci that are associated with disease by efficiently removing false positive loci. We demonstrate this method in a publicly available data set for Alzheimer’s disease (AD) (Coon et al, 2007), consisting of 502,627 SNP loci genotyped in unrelated 859 cases and 552 neurologically normal controls. A total of 26 loci from the 22 autosomes are identified and they explain 75% of the genetic risk variability of the disease.

Results and discussion

AH size threshold

In the context of the current data, it is not appropriate to use the number of loci as a measure of AH size as previously reported (Lencz et al, 2007) because of its dependence on SNP density. Here we use the number of nucleotide basepairs between the first and last loci of an AH as a measure of AH size.

Let \( C \) be a size threshold of AHs. We are interested in identifying loci proportions of which are significantly different between controls and cases in AHs with sizes \( \geq C \). As seen in Figure 1, for example, there are \( n_1 \) cases and with a given \( C \) we compute the proportion of the locus SNP-1 on AHs \( p_1 \)(number of AHs containing SNP-1)/\( n_1 \). Similarly, for \( n_0 \) controls, we find the proportion \( p_0 \) of the same locus. Using \( p_1 \) and \( p_0 \), we compute \( z \)-statistic for proportional test as described in Materials and methods. The locus is selected for further screening if \( |z| \geq z_{1-\alpha/2} \). It should be noted that the distance between \( C_1 \) and \( C_2 \) must be larger than the minimum distance between loci of the platform and may be chosen by referring to some public genotyping parameters (for example, the average distance between loci is \( \sim 9 \) kb in Affymetrix 500K GeneChip, and a median distance is \( \sim 3 \) kb in Illumina HumanHap550 BeadChip according to Gunderson et al, 2005; Steemers and Gunderson, 2007). Let \( S_1 \) be the set containing the loci selected with \( C_1 \) and \( S_2 \) with \( C_2 \), respectively. As the true AHs with size \( \geq C_2 > C_1 \) will remain using either \( C_1 \) or \( C_2 \), the loci, not in \( S_1 \cap S_2 \), should be more likely false positives and thus be removed. For example, in the AD data using a significance level \( \alpha=0.001 \), among the 25,086 loci on chromosome 1, there were 18 loci selected using \( C=10 \) kb and 12 loci using \( C=30 \) kb, respectively, with only three being common in both sets. In general, we set \( C=\{C_i, i=1, 2, \ldots, L\} \) with \( C_1 < C_2 < \cdots < C_L \) to cover a wide range of AHs and let \( S \) be the set containing all loci common in adjacent sets \( S=\{S_1 \cap S_2, S_2 \cap S_3, \ldots, S_{L-1} \cap S_L\} \). This loci-selecting procedure is called ‘procedure of adjacent-C-selection’ (PACS).

The PACS can efficiently remove false positive loci, however, for a real data set in unrelated individuals with large genetic variation, the selected loci usually still contain some false positives, many of which could be removed through further ‘purification’. To achieve this, ideally we should repeat the above steps using an independent data set from the same population to get another candidate set. Then identify the common loci from both sets. This new candidate set contains fewer false positive loci, which could be further removed by repeating above steps iteratively until the number of candidate loci converges. Although it is generally not realistic to do so, we could do the ‘purification’ using random subsets from the full data set as described below.

Let \( n_k^*=[f \times n_k] > 30 \) be the size of a random subset from the full data set of size \( n_k \), where \( k=1 \) for cases and \( k=0 \) for controls, and \( f \) be a constant with \( 0<f<f_{\text{max}} \). \( f_{\text{max}} = \frac{\text{min}_k(n_k) - 1}{\text{min}_k(n_k)} \). The randomly and independently chosen \( n_k^* \) cases and \( n_k^* \) controls form a random case–control sub-data set for further removing the false positive loci.

Algorithm for screening risk loci

We propose to use multiple \( C \) values for screening risk loci. Suppose we choose \( C_1 \) and \( C_2 \), with \( C_1 < C_2 \), for selecting candidate loci with \( |z| \geq z_{1-\alpha/2} \). It should be noted that the distance between \( C_1 \) and \( C_2 \) must be larger than the minimum distance between loci of the platform and may be chosen by referring to some public genotyping parameters (for example, the average distance between loci is \( \sim 9 \) kb in Affymetrix 500K GeneChip, and a median distance is \( \sim 3 \) kb in Illumina HumanHap550 BeadChip according to Gunderson et al, 2005; Steemers and Gunderson, 2007). Let \( S_1 \) be the set containing the loci selected with \( C_1 \) and \( S_2 \) with \( C_2 \), respectively. As the true AHs with size \( \geq C_2 > C_1 \) will remain using either \( C_1 \) or \( C_2 \), the loci, not in \( S_1 \cap S_2 \), should be more likely false positives and thus be removed. For example, in the AD data using a significance level \( \alpha=0.001 \), among the 25,086 loci on chromosome 1, there were 18 loci selected using \( C=10 \) kb and 12 loci using \( C=30 \) kb, respectively, with only three being common in both sets. In general, we set \( C=\{C_i, i=1, 2, \ldots, L\} \) with \( C_1 < C_2 < \cdots < C_L \) to cover a wide range of AHs and let \( S \) be the set containing all loci common in adjacent sets \( S=\{S_1 \cap S_2, S_2 \cap S_3, \ldots, S_{L-1} \cap S_L\} \). This loci-selecting procedure is called ‘procedure of adjacent-C-selection’ (PACS).

The PACS can efficiently remove false positive loci, however, for a real data set in unrelated individuals with large genetic variation, the selected loci usually still contain some false positives, many of which could be removed through further ‘purification’. To achieve this, ideally we should repeat the above steps using an independent data set from the same population to get another candidate set. Then identify the common loci from both sets. This new candidate set contains fewer false positive loci, which could be further removed by repeating above steps iteratively until the number of candidate loci converges. Although it is generally not realistic to do so, we could do the ‘purification’ using random subsets from the full data set as described below.

Let \( n_k^*=[f \times n_k] > 30 \) be the size of a random subset from the full data set of size \( n_k \), where \( k=1 \) for cases and \( k=0 \) for controls, and \( f \) be a constant with \( 0<f<f_{\text{max}} \). \( f_{\text{max}} = \frac{\text{min}_k(n_k) - 1}{\text{min}_k(n_k)} \). The randomly and independently chosen \( n_k^* \) cases and \( n_k^* \) controls form a random case–control sub-data set for further removing the false positive loci.

Figure 1  Scheme for computing the proportion of a locus on AHs. For a given chromosome of a subject, the symbols (●, ○) represent SNP loci. The shaded segments denote AHs with size greater than or equal to a pre-selected threshold \( C \). The proportion of a locus on AHs is computed as \( p= (\text{the number of AHs containing this locus})/(\text{the total number of individuals}) \), for example \( p_1=4/6 \) for SNP-1.
from the candidate set using the same set of $C$ values as applied to the full data set.

Let $S$ be the set containing the selected loci from the full data set and $S^*$ be that from the random sub-data set. Let $S_1^* = S^* \cap S$ containing the common loci in both sets and $N_1 = |S_1^*|$ be the number of loci in $S_1^*$. Next we generate a new $S^*$ from the second random sub-data set and let $S_2^* = S^* \cap S^*$ with $N_2 = |S_2^*|$. Repeating these steps to update the candidate loci set until the number of $N_t, t=1,2,\ldots$, converges to a constant integer $N_c$ with $N_c = 0$ if the null hypothesis of no difference between $p_1$ and $p_0$ is true and $N_c \geq 4$ if the alternative hypothesis $p_1 \neq p_0$ is true. For a given $f$, there are possible ways for selecting case–control subset, which should be much larger than the number required for reaching convergence at an appropriate level of significance. The above GAHA algorithm is summarized in Box 1.

**Figure 2**  The plot of $z$ versus nucleotide basepair of chromosome 19 in the AD data set: (A) before and (B) after the procedure of adjacent-C-selection, (C) the most significant region—the peak locus is rs4420638, (D) the most significant region with two loci on APOE ( ).
The false positive rate of a locus in the final set should be $\leq \alpha$. The false negative rates of loci selection in a random subset were estimated under the same settings for the full data set (Supplementary Table 2).

Application to AD data set

Set $C=\{1, 10 \text{ kb}, 30 \text{ kb}, 50 \text{ kb}, 100 \text{ kb}, 140 \text{ kb}, 250 \text{ kb}, 500 \text{ kb}, 1 \text{ Mb}\}$ and $\alpha = 0.001$. We identified 607 loci from 4054 loci whose $z \geq z_{1-\alpha/2}$ (Figure 2A) from the 22 autosomes in the AD data set (Coon et al., 2007).

The most significant AH region was on 19q13.2 (see Figure 2B) with positive $z$ values suggesting significantly more AHs in controls than in cases. This region, covering the whole apolipoprotein E (APOE) gene, contains four loci including rs4420638 (Figure 2C), which is in linkage disequilibrium with APOE (Coon et al., 2007). However, there were no genotypes within APOE in the AD data. We added available genotyping information (Coon et al., 2007) of two loci on APOE, rs429358 and rs7412, to the AD data. The two APOE loci define the $e2/e3/e4$ genotypes. Figure 2D shows the APOE loci indeed on the AH region where the majority controls have the $e3$ genotype, supporting the observation that APOE $e3$ is protective against the disease when compared with $e4$ (Farrer et al., 1997).

To further reduce the false positive rate within this list, we chose $f=0.9$ for generating random subsets, each with 773 cases and 497 controls. The use of $f=0.9$ may not be the statistically optimal choice; it is, however, the best we tried. The convergence of the loci number is shown in Figure 3. There were 26 loci in the final list (Figure 3B) (Table I). Based on a logistic regression model fit, the percent variation of the genetic risk explained by these 26 loci was 75.3%. Model selection removed 10 confounder loci and retained 16 loci (each with $P$-value $<0.05$), including rs4420638, in the reduced model with 74.8% of the genetic risk variation explained (Supplementary Table 3, 4).

The APOE $e4$ was carried by $\sim 40\%$ of the later-onset AD cases (Poirier et al., 1993; Laws et al., 2003). Recall that rs4420638 is in linkage disequilibrium with APOE, we found that the percent genetic risk variation explained by this locus alone was 34.2%. However, when rs4420638 was excluded from the reduced model, the percentage genetic risk variation explained by the remaining 15 loci was decreased only by 2.9% (from 74.8% to 71.9%). This suggests these loci explain the genetic risk variation of AD as a group. Several of the 26 loci identified in this screening were also found in homozygous regions identified in an early onset AD study of a consanguineous family (Clarinó et al., 2008), suggesting that one of these regions harbors a recessive genetic lesion causing AD.

The 26 loci are on 20 genes of which 13 are in known functional pathways or networks as revealed from an Ingenuity Pathway Analysis (Ingenuity Systems, www.ingenuity.com) (Supplementary Pathway/Network analysis). On the basis of the correlations among the 20 genes and AD status of subjects, we construct an AD genetic network (Supplementary Figure 2).

Summary

We propose a statistical method for GAHA of SNP case–control data in unrelated subjects to identify risk loci that are most likely associated with a disease or abnormality due to recessive mutation or deletion. The main novelty of this method over other approaches is to minimize the false positive rate of the risk candidates. We remove the false positive loci by selecting the common loci with different size thresholds of homozygous segments and repeating these steps iteratively using random sub-data sets until the number of selected loci converges. Furthermore, this method allows selects risk loci from a wider AH size range. By demonstrating of the method using a publicly available AD SNP assay data set, we identified 26 candidate risk loci from the 22 autosomes.

Materials and methods

Notes

Suppose there are $n$ SNP loci genotyped on a given chromosome (an autosome). We view the sequences of SNP loci on a chromosome as
Table I List of candidate loci associated with AD from the 22 autosome of the AD SNP genotype data (Coon et al., 2007)

| CHR | SNP ID | Location | Function | Gene | Gene ID | Effect |
|-----|--------|----------|----------|------|---------|--------|
| 1   | rs17325867^bc | 699984761 | Intron   | LRRCD^d | 57554 | Risk   |
| 1   | rs1038891 b,c | 40895642 | 0        | C1orf175 | 57554 | Risk   |
| 1   | rs7582851 b   | 192032391 | Intron   | LOC440700 | 440700 | Protect |
| 1   | rs7684615 b   | 52481466 | Intron   | LOC729817 | 729817 | Risk   |
| 1   | rs13193950   | 81593433 | –71296   | LOC647195 | 647195 | Protect |
| 1   | rs1785928 b   | 31979929 | Coding   | VTCN1^e | 79679 | Protect |
| 2   | rs9994615 b   | 40786592 | Intron   | COL2A1 | 255631 | Risk   |
| 2   | rs7582851 b   | 192032391 | Intron   | LOC440700 | 440700 | Protect |
| 3   | rs1213247 b   | 81572755 | mRNA–UTR | PVRL2 | 52899 | Risk   |
| 4   | rs10105784 d  | 40793978 | 0        | PRKCI | 5592 | Risk   |
| 5   | rs1602843 b   | 86324342 | 0        | APOB | 374977 | Risk   |
| 6   | rs13213247 b  | 163947773 | –32939  | KRT8P21 | 51514 | Risk   |
| 7   | rs10827687 b  | 59088188 | –039887  | KRT8P21 | 51514 | Risk   |
| 8   | rs12681 b     | 90220223 | –1188    | LOC642337 | 642337 | Risk   |
| 9   | rs10740548    | 40895642 | 0        | CLOTT15 | 39783 | Risk   |
| 10  | rs1358891 b,c | 104979509 | 481535  | LOC440700 | 440700 | Protect |
| 10  | rs10827687 b  | 90220223 | –1188    | LOC642337 | 642337 | Risk   |
| 10  | rs13213247 b  | 81572755 | –72008   | LOC729817 | 729817 | Risk   |
| 10  | rs13193950    | 81593433 | –71296   | LOC729817 | 729817 | Risk   |
| 11  | rs10827687 b  | 90220223 | –1188    | LOC642337 | 642337 | Risk   |
| 12  | rs7967572     | 73396086 | 51514    | KRT8P21 | 12681 | Risk   |
| 12  | rs1789292 b   | 31979929 | Coding   | ELPI | 52520 | Risk   |
| 13  | rs11879589    | 50065116 | Intron   | PVR| 5819 | Protect |
| 14  | rs4420638 b   | 50114786 | Locus region | APOC1 | 341 | Protect |
| 15  | rs10827687 b  | 90220223 | –1188    | LOC642337 | 642337 | Risk   |
| 16  | rs204907      | 50153836 | Intron   | CLPTM1 | 1209 | Protect |

^aIn nucleotide basepair.
^bLoci remained in the model on logistic regression selection with a P-value<0.05.
^cLoci in homozygous regions containing candidate loci of recessive genetic lesion causing AD (Clarimón et al., 2008).
^dA SNP in Affymetrix 500K GeneChip, but without NCBI ID.
^eLoci are on known functional pathways and networks as revealed by the use of Ingenuity Pathway Analysis (Ingenuity Systems, www.ingenuity.com).

Logistic regression

In logistic regression using the selected loci as predictor variables, let $x_{ij}=1$ if the $j$th locus of the $i$th subject is on an AH with size being equal to or larger than $C=10$ kb and $x_{ij}=0$ otherwise. Logistic regression is carried out using SAS 9.0.

Supplementary information

Supplementary information is available at the Molecular Systems Biology website (www.nature.com/msb).

Declaration

The views expressed in this article do not represent those of the US Food and Drug Administration.

Acknowledgements

This study was supported by the Intramural Program of the National Institute on Aging, National Institutes of Health and Department of Health and Human Services, project number AG1000950-07. This study used high-performance computational capabilities of the Biowulf Systems at the National Institutes of Health, Bethesda, MD (http://helix.nih.gov).

Conflict of interest

The authors declare that they have no conflict of interest.

References

Clarimón J, Djaldetti R, Lleó A, Guarrojo RJ, Molinuevo JL, Paisán-Ruiz C, Gómez-Isla T, Blesa R, Singleton A, Hardy J (2008)
Whole genome analysis in a consanguineous family with early onset Alzheimer’s disease. *Neurobiol Aging*, doi:10.1016/j.neurobiolaging.2008.02.008

Coon KD, Myers AJ, Craig DW, Webster JA, Pearson JV, Hu Lince D, Zismann VL, Beach TG, Leung D, Bryden L, Halperin RF, Marlowe L, Kaleem M, Walker DG, Ravid R, Heward CB, Rogers J, Papassotiropoulos A, Reiman EM, Hardy J et al. (2007) A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer’s disease. *J Clin Psychiatry* 68: 613–618

Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, Myers RH, Pericak-Vance MA, Risch N, van Duijn CM (1997) Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *J Am Med Assoc* 278: 1349–1356

Lencz T, Lamberta C, DeRosse P, Burdick K, Morgan TV, Kane JM, Kucherlapati R, Malhotra AK (2007) Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc Natl Acad Sci USA* 100: 9440–9445

Miyazawa H, Kato H, Awata T, Kohda M, Iwasa H, Koyama N, Tanaka T, Huqun, Kyo S, Okazaki Y, Hagiwara K (2007) Homozygosity haplotype allows a genome-wide search for the autosomal segments shared among patients. *Am J Hum Genet* 80: 1090–1102

Mueller RF, Bishop DT (1993) Autozygosity mapping, complex consanguinity, and autosomal recessive disorders. *J Med Genet* 30: 798–799

Poirier J, Davignon J, Bouthillier D, Kogan S, Bertrand P, Gauthier S (1993) Apolipoprotein E polymorphism and Alzheimer’s disease. *Lancet* 342: 697–699

Rabbee N, Speed TP (2006) A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* 22: 7–12

Simon-Sanches J, Scholz S, Fung HC, Matarin M, Hernandez D, Gibbs JR, Britton A, Warrant de Brieze P, Peckham E, Gwinn-Hardy K, Crawley A, Keen JC, Nash J, Borgaonkar D, Hardy J, Singleton A (2007) Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. *Hum Mol Genet* 16: 1–14

Steemers FJ, Gunderson KL (2007) Whole genome genotyping technologies on the BeadArray platform. *Biotechnol J* 2: 41–49

Van Eyken E, Van Camp G, Van Laer L (2007) The complexity of age-related hearing impairment: contributing environmental and genetic factors. *Audiol Neurootol* 12: 345–358