Published in final edited form as:

Hum Genet. 2009 August ; 126(2): 303–315. doi:10.1007/s00439-009-0672-3.

Detection of disease-associated deletions in case–control studies using SNP genotypes with application to rheumatoid arthritis

Chih-Chieh Wu,

Unit 1340, Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, 1155 Pressler Street, Houston, TX 77030, USA

Sanjay Shete,

Unit 1340, Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, 1155 Pressler Street, Houston, TX 77030, USA

Wei V. Chen,

Unit 1340, Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, 1155 Pressler Street, Houston, TX 77030, USA

Bo Peng,

Unit 1340, Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, 1155 Pressler Street, Houston, TX 77030, USA

Annette T. Lee,

Center for Genomics and Human Genetics, North Shore-Feinstein Medical Research Institute, Manhasset, NY, USA

Jianzhong Ma,

Unit 1340, Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, 1155 Pressler Street, Houston, TX 77030, USA

Peter K. Gregersen, and

Center for Genomics and Human Genetics, North Shore-Feinstein Medical Research Institute, Manhasset, NY, USA

Christopher I. Amos

Unit 1340, Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, 1155 Pressler Street, Houston, TX 77030, USA

Chih-Chieh Wu: ccwu@mdanderson.org

Abstract

Genomic deletions have long been known to play a causative role in microdeletion syndromes. Recent whole-genome genetic studies have shown that deletions can increase the risk for several psychiatric disorders, suggesting that genomic deletions play an important role in the genetic basis of complex traits. However, the association between genomic deletions and common, complex diseases has not yet been systematically investigated in gene mapping studies. Likelihood-based statistical methods for identifying disease-associated deletions have recently been developed for familial studies of parent-offspring trios. The purpose of this study is to develop statistical approaches for detecting genomic deletions associated with complex disease in case–control studies. Our methods are designed to be used with dense single nucleotide polymorphism (SNP)

genotypes to detect deletions in large-scale or whole-genome genetic studies. As more and more SNP genotype data for genome-wide association studies become available, development of sophisticated statistical approaches will be needed that use these data. Our proposed statistical methods are designed to be used in SNP-by-SNP analyses and in cluster analyses based on combined evidence from multiple SNPs. We found that these methods are useful for detecting disease-associated deletions and are robust in the presence of linkage disequilibrium using simulated SNP data sets. Furthermore, we applied the proposed statistical methods to SNP genotype data of chromosome 6p for 868 rheumatoid arthritis patients and 1,197 controls from the North American Rheumatoid Arthritis Consortium. We detected disease-associated deletions within the region of human leukocyte antigen in which genomic deletions were previously discovered in rheumatoid arthritis patients.

Introduction

Genomic deletions have long been known to cause a variety of genetic disorders such as DiGeorge syndrome, Prader-Willi syndrome, Williams syndrome, and Wilms tumor (Lindsay 2001). Microdeletion syndromes are generally caused by relatively large deletions and involve multiple genes. Recent studies have increasingly shown that deletions are common in patients with psychiatric disorders such as autism and schizophrenia, suggesting that genomic deletions also play an important role in the genetic basis of complex traits in the human genome. Deletions of chromosome 22q11 have been frequently observed in patients with schizophrenia or schizoaffective disorder (Liu et al. 2002; Shifman et al. 2006). Recently, three large recurrent deletions associated with schizophrenia have been discovered on chromosomes 1 and 15 ranging from 0.47 to 1.57 Mb in size (Stefansson et al. 2008). Autism patients have been found to have interstitial deletions of varying sizes ranging from 5.9 to 593 kb on chromosomes 7, 8, and 16 (Yu et al. 2002; Weiss et al. 2008).

A series of three concurrent studies investigated common deletion polymorphisms in healthy individuals from single nucleotide polymorphism (SNP) genotype data in the whole human genome; these studies provided a source of baseline information for studies of human disease and genome evolution (Conrad et al. 2006; Hinds et al. 2006; McCarroll et al. 2006). The current molecular technologies, including SNP genotyping methods, are not normally capable of identifying genomic deletions. Instead, SNPs in the regions of hemizygous deletions are generally miscalled and observed as homozygous for the allele that is present (Conrad et al. 2006; McCarroll et al. 2006). These genotyping “miscalls” result from footprints left by segregating deletions in SNP genotyping data; existing methods for detecting genomic deletions are often based on the observation of deviations from Mendelian transmission patterns in parent-offspring trios, the Hardy–Weinberg equilibrium, and null genotypes for a run of consecutive SNPs (McCarroll et al. 2006). Algorithms for detecting genomic deletions were developed using these “failed” SNP genotype patterns (Conrad et al. 2006; McCarroll et al. 2006). These concurrent studies showed that non-diseased individuals carry many various-sized genomic deletions with high population frequencies. Recently, Amos et al. (2003) and Kohler and Cutler (2007) proposed likelihood-based statistical methods for detecting disease associated deletions from familial SNP genotype data. Their methods harnessed the information on SNP transmission patterns between true genotypes and observed (miscalled) genotypes within parent-offspring trios to detect deletions.

While SNPs have long been known to be a predominant form of genomic variation and can make important contributions to phenotypic variations, copy number variation (CNV), which is one type of structural variants, has attracted attention over the last few years. CNVs have been found to be associated with rare disorders such as Alzheimer disease and autism

(Rovelet-Lecrux et al. 2006; Sebat et al. 2007) but their contribution to the risk for common, complex disorders remains largely unclear. Recent whole-genome CNV studies using either SNP genotyping arrays or clone-based comparative genomic hybridization (CGH) reported thousands of candidate CNV segments, suggesting that CNVs are ubiquitous and can be an important form of genomic variation in the human genome (Redon et al. 2006; Wong et al. 2007). SNP- and clone-based CNV detection methods have different sensitivities in various genomic regions and variations (Redon et al. 2006). Current CNV detection methods have technical challenges that result in relatively limited resolutions for whole-genome array scans. For example, SNP-based CNV detection methods typically require consecutive (at least 3) SNPs with aberrant intensity characteristics. A recent study suggests that imputation can be used to increase the sensitivity of CNV detection with extra, un-typed alleles (Franke et al. 2008). Array CGH assays do not provide absolute copy numbers but rather the change compared with the copy numbers in the reference sample (Freeman et al. 2006). Thus, using CNV methods to unambiguously determine deletions or duplications in the human genome and further assess the association between CNVs or genomic deletions and complex disease is challenging without comprehensive knowledge of normal and causative variants (Redon et al. 2006). The fine-scale approaches such as digital karyotyping or forsmid end sequencing are more robust at identifying genomic deletions and CNVs than are SNP-based detection methods (Wang et al. 2002; Tuzun et al. 2005); however, technologies based on DNA-sequencing are expensive and time consuming, making it difficult to perform whole-genome genetic studies.

Current microarray platforms are relatively affordable and are capable of genotyping hundreds of thousands of SNPs in an experiment, making them more feasible for large-scale or whole-genome genetic studies. High-density SNP genotype data on a variety of common, complex diseases have been generated for genome-wide association studies, which usually genotype 300,000 or more SNPs from oligonucleotide arrays. Thus, development of sophisticated statistical approaches is needed that use these SNP genotype data. In contrast with recent whole-genome studies that focused on detecting genomic deletions in healthy individuals or were structured to be used in familial studies of parent-offspring trios, our study was designed to detect genomic deletions associated with common, complex diseases and for case-control studies.

In this study, we developed two statistical approaches to detecting disease-associated deletions from high-density SNP genotype data in genome-wide association studies. The first method was designed for single SNP analyses and the second was designed to utilize evidence from multiple SNPs combined using cluster based approaches. We performed simulation studies and found that our proposed methods were useful for detecting disease-associated deletions and were robust in the presence of linkage disequilibrium (LD) using simulated SNP data sets. We applied the proposed methods to SNP genotype data on chromosome 6p in 868 rheumatoid arthritis (RA) patients and 1,197 controls from the North American Rheumatoid Arthritis Consortium. Our proposed methods identified the disease-associated deletions that encompassed HLA_DRB1 and C4 genes in the human leukocyte antigen (HLA) region in which genomic deletions were previously discovered in RA patients.

Statistical methods and data simulations

In this section, we describe two proposed SNP-based statistical methods in case-control studies. These statistical approaches aim to detect genomic deletions that are associated with susceptibility to complex diseases. The first method is structured for SNP-by-SNP analyses, and the second is structured for cluster analyses based on combined evidence between SNPs, which can be used for subsequent investigations into the outcomes of SNP-by-SNP analyses.

The SNP genotype data were simulated according to the International HapMap project. Three different-sized deletions were created at various locations and imposed on a portion of cases to account for the genetic heterogeneity and complexity of disease.

Statistical method for SNP-by-SNP analyses

Because current molecular techniques and SNP genotyping methods are not capable of effectively identifying genomic deletions and because the SNPs within the regions of hemizygous deletions are generally miscalled and observed as homozygous for the present allele, Amos et al. (2003) and Kohler and Cutler (2007) proposed likelihood-based methods for detecting disease-associated deletions from familial SNP genotype data. Here, we propose to compare the level of homozygosity on every locus between cases and controls by using normal approximations to test the significance of difference in homozygosity proportions. This test infers the presence of genomic deletions associated with disease by assessing the statistical significance of higher homozygosity proportions in cases than in controls. This method is structured for SNP-by-SNP analyses. Letting $\hat{p}_1$ and $\hat{p}_2$ be the respective estimates of homozygosity proportions in cases and controls at a single SNP locus and $\bar{p}$ be their weighted average, the normal deviate Z is based on the difference in proportion quantities, $\hat{p}_1 - \hat{p}_2$, divided by its standard error, $\sqrt{\bar{p}(1 - \bar{p})(1/n_1 + 1/n_2)}$, where n_1 and n_2 represent the sample sizes of cases and controls, respectively.

Statistical method for cluster analyses

In our cluster-based scheme, the evidence of genomic deletions that are associated with disease can be enhanced by the observation of successive or neighboring SNPs with excess homozygosity in cases compared with in controls. Thus, we develop a statistical test for cluster analyses to further characterize and evaluate the combined evidence between SNPs. This test is useful for subsequent investigations into the outcomes of SNP-by-SNP analyses and is designed to assess the statistical significance of multiple clusters of SNP loci with excess homozygosity in cases compared with in controls.

Suppose that T SNP loci over a chromosomal region are tested using the z-score test for SNP-by-SNP analyses in which k SNP loci have significantly higher homozygosity proportions in cases than in controls. Consider the frequency of significant SNP loci occurring within a narrow segment of interest compared with the frequency of significant SNP loci over the whole region. Suppose that the narrow segment of interest encompasses w SNP loci, among which x SNP loci have significant excess homozygosity in cases aggregating in this segment. What interests us is to evaluate whether the observation of x significant SNP loci in the segment that contains w SNP loci is statistically significant compared with the occurrence of $k - x$ significant SNP loci over $T - w$ SNP loci. Assuming that each of the T SNP loci tested is independently and equally likely to have a significant excess homozygosity proportion in cases and assuming that X represents the number of significant SNP loci within the segment that contains w SNP loci, the proposed statistical test for cluster analyses is based on the random variable X with a binomial distribution. The p -value formula for this cluster test under the null hypothesis of random allocations of significant SNP loci is expressed as follows:

$$P\left(X \geq x | w, p_0 = \frac{k}{T}\right) = \sum_{i \geq x} \binom{w}{i} p_0^i (1 - p_0)^{w-i}, \quad (1)$$

where x represents the observed number of significant SNP loci within the segment that contains w SNP loci. A small p -value of expression (1) indicates that the occurrence of x significant SNP loci aggregating in a w SNP interval cannot be explained by chance alone.

Detecting regions of excess homozygosity across multiple adjacent or neighboring SNPs forms the basis of this method for cluster analyses. Related statistical methods have been developed for detecting temporal and space-time clusters or anomalies of disease in epidemiology studies (Grimson 1993; Grimson and Mendelsohn 2000; Wu et al. 2008).

Our proposed cluster test based on combined evidence between SNPs is designed to assess the statistical significance of multiple clusters, as shown in expression (1). In contrast, the scan test is structured to detect the largest cluster. The scan test employs a moving window of predetermined length w and finds the maximum number of cases revealed through the window as it slides over the entire region (Wallenstein and Neff 1987). When only the largest cluster is being assessed or only 1 observed cluster is present, the scan test will be useful.

The z-score test is proposed for SNP-by-SNP analyses along a chromosomal region of interest. On the basis of outcomes of SNP-by-SNP analyses, we developed a statistical test for cluster analyses based on combined evidence between SNPs. These statistical approaches can be used to detect commonly shared deletions among patients with a genetic disorder, providing strong evidence that the genes in the deleted region predispose patients to disease.

Data simulations

To simulate data that show a realistic LD structure while maintaining similarity across replicate simulations, we used a forward-time simulation approach adapted from a method described by Peng et al. (2007). We selected markers with a minor allele frequency of ≥ 0.05 on chromosome 2q (ranging from 150 to 240 cM) with additional criteria on the maximum distance and LD allowed between adjacent SNP markers. The simulations started with 60 CEU parents of European ancestry sampled in Utah, USA, from the international HapMap project and followed a demographic model that roughly mimicked the evolution of an isolated population that expanded exponentially during the past 10,000 years (Wall and Przeworski 2000). A recombination model that uses an estimated fine-scale genetic map was applied during the evolution (McVean et al. 2004). The simulation was accelerated ten times using a generation scale technique described by Hoggart et al. (2007). Simulation parameters, including present population size, mutation rate, and scaling rate, were calibrated using a set of statistics so that the simulated samples resembled the allele frequencies and LD structure of the HapMap sample (Schaffner et al. 2005).

To evaluate statistical sensitivity in the presence of LD, we used two criteria to select SNP markers with the desired LD levels over the genome. One criterion was the maximum value of LD (MaxLD) allowed between adjacent SNPs, in which the magnitude of LD is defined by the coefficient

$$r^2 = \frac{(P_{AB} - P_A P_B)^2}{P_A(1 - P_A)P_B(1 - P_B)},$$

where P_A and P_B are the frequencies of alleles A and B, respectively, and P_{AB} is the frequency of haplotype AB. The other criterion was the maximum distance (MaxGap) allowed between adjacent SNPs. SNP markers were selected sequentially. A marker was chosen if its r^2 value with the last selected marker was less than the value of MaxLD or if its distance from the last selected marker reached the value of MaxGap. Several simulated SNP data sets were generated jointly using different values of Max-LD ($r^2 = 0.0, 0.10$, and 0.15) and MaxGap (10, 30, and 50 kb).

Assuming that the population prevalence of the disease of interest is 0.005, the prevalence of the disease-associated deletion is 0.01, and the penetrance of the disease-associated deletion is 0.16, the probability of carrying the disease-associated deletion is 0.32 for affected individuals and 0.0084 for unaffected individuals. Because the size of deletions plays an important role in the statistical sensitivity of the proposed statistical methods, we created disease-associated deletions of varying sizes. First, we created a disease-associated deletion of size uniformly ranging between 280 and 320 kb. The mean size of this created deletion is 300 kb. Similarly, we created a disease-associated deletion of size uniformly ranging between 80 and 120 kb and between 480 and 520 kb. The mean sizes of the deletions are 100 and 500 kb, respectively.

To accurately characterize and evaluate the statistical sensitivity of the proposed statistical methods, we simulated 100 replicates in the various simulation settings and randomly selected 300 cases and 300 controls from each replicate. In this study, we performed the statistical analyses and summarized the outcomes on the basis of 100 replicates for each of the various SNP simulation settings, which vary jointly by different values of mean disease-associated deletion size (100, 300, and 500 kb), MaxLD ($\gamma^2 = 0.0, 0.10, \text{ and } 0.15$), and MaxGap (10 kb, 30 kb, and 50 kb).

Results

Our proposed statistical approaches in case-control studies were useful for detecting disease-associated deletions and were robust in the presence of LD. First, we applied our proposed methods to the simulated SNP data set with a maximum LD of 0.15 allowed between adjacent SNPs (that is, MaxLD = Max $\gamma^2 = 0.15$) and the largest gap of 30 kb allowed between adjacent SNPs (that is, MaxGap = 30). In this setting, 6,745 SNPs were selected over the study region. The mean value of the distance between adjacent SNPs was ~13.3 kb, and the mean value of the γ^2 was 0.064. We created a random disease-associated deletions (mean size of 300 kb ranging between [280, 320]) centered at the 3,000th SNP. The disease-associated deletion of 300 kb encompassed the region from the 2,989th SNP through the 3,021st SNP, containing a total of 33 SNP loci.

Preliminary analysis at 10^{-3} nominal significance level

For SNP-by-SNP analyses, we used normal approximations to test of significance for difference in homozygosity proportions between the group of 300 cases and the group of 300 controls on each of 6,745 contiguous loci. We output the z-scores by locus, as shown in Fig. 1. We first used a nominal significance level of 1×10^{-3} , which corresponds to a z-score of 3.08. Figure 1 shows several “hits” clustered in the disease-associated deletion region, as indicated by a red arrow. Nineteen of 6,745 SNP loci had z-scores of ≥ 3.08 , 6 of which appeared in the 300 kb-deletion segment. In this case, the deletion detection rate for association with disease (that is, the probability of detecting disease-associated deletions among significant SNPs) was 0.32 (=6/19) per SNP from this single simulation. On the other hand, 13 significant SNPs lay outside the deletion, which can be classified as false-positive findings. The disease-associated deletion in mean size of 300 kb contained 33 SNP loci, indicating that the false-positive deletion detection rate for association with disease (that is, the probability of inferring disease-associated deletions when they do not exist) was 1.94×10^{-3} (=13/(6,745 – 33)) per SNP in this case.

Twenty-seven of 33 SNP loci in the disease-associated deletion had z-scores of < 3.08 and were not statistically significant at the nominal significance level of 1×10^{-3} . The major reasons might include (1) genetic heterogeneity: less than one-third of cases had disease-associated deletions, (2) excess homozygosity proportions in the disease-associated deletion in a substantial number of controls occurring by chance, (3) deletions shorter than the mean

value of 300 kb in a substantial portion of cases (because the size of disease-associated deletion ranges from 280 to 320 kb), hampering the detection sensitivity of SNP loci near the deletion boundaries, and (4) sample size in this simulation study is smaller than would typically be employed in a genome-wide association study.

Next, we applied our proposed statistical method for cluster analyses to investigate the outcomes of SNP-by-SNP analyses based on combined evidence between SNPs. As shown in Fig. 1, 4 observed clusters of successive or neighboring significant SNPs appeared with z -scores of ≥ 3.08 . We used expression (1) to calculate the p -value of the first cluster, which encompassed a short segment between the 2,284th SNP and 2,288th SNP. The null probability was $p_0 = 2.82 \times 10^{-3}$ ($=19/6,745$) because 6,745 SNP loci were tested and 19 of them had z -scores of ≥ 3.08 . This candidate deletion segment contained $w = 5$ SNP loci from the 2,284th SNP to 2,288th SNP. The observed number of significant SNP loci within this segment was 3 because the 2,284, 2,286, and 2,288th SNPs had z -scores of 3.15, 3.89, and 3.11, respectively. The p -value of this cluster using expression (1) was $Pr(X \geq 3|w = 5, p_0 = 2.82 \times 10^{-3}) = 2.23 \times 10^{-7}$. We subsequently analyzed the other three clusters; the corresponding results are shown in Table 1. Because the four observed clusters of multiple neighboring significant SNPs shown in Table 1 can occur at many other locations along the whole chromosomal region, we must take these into account when we assess statistical significance using the proposed statistical test for cluster analyses. It is note-worthy that the p -values shown in Table 1 were directly obtained using expression (1) with no correction imposed. We suggest the use of a Bonferroni-type correction to adjust the p -value thresholds.

The proposed method for cluster analyses is based on the comparison of occurrence of x significant SNP loci in a segment that contains w SNP loci with that of $k - x$ significant SNP loci over the remaining $T - w$ SNP loci. The number of ways to select a segment that contains exactly w contiguous SNP loci along T SNP loci is $(T - w + 1)$. Assuming the use of a nominal significance level of α , we propose to divide α by $(T - w + 1)$ as an adjusted p -value threshold. This is a Bonferroni-type adjustment as it provides an upper bound on the overall nominal significance level of α . Bonferroni corrections are designed to control family-wise error rates and are widely considered to be overly conservative. Our proposed correction can be relatively conservative as well. At an overall nominal significance level of 0.05, the corresponding adjusted nominal significance levels for the four observed clusters shown in Table 1 are 7.42×10^{-6} ($=0.05/(6,745 - 5 + 1)$), 7.43×10^{-6} ($=0.05/(6,745 - 15 + 1)$), 7.42×10^{-6} ($=0.05/(6,745 - 10 + 1)$), and 7.43×10^{-6} ($=0.05/(6,745 - 15 + 1)$). In this application, the numbers of w are tiny compared with $T = 6,745$. We can simply divide α by T as an approximate corrected p -value threshold in this situation. That is, 7.41×10^{-6} ($=0.05/6,745$) can serve as a corrected p -value threshold at a nominal significance level of 0.05. When a nominal significance level of 0.01 is chosen, the corrected p -value threshold becomes 1.48×10^{-6} ($=0.01/6,745$). In summary of Table 1, the first and second clusters are significant at a nominal significance level of 0.01, but the third and fourth clusters are not significant at a nominal significance level of 0.05. The second cluster lies in the range of disease-associated deletion and can be classified as a true-positive finding.

Analyses at various nominal significance levels

In addition to our statistical test for cluster analyses, we also need to account for the effect of multiple comparisons in SNP-by-SNP analyses. The Bonferroni procedure for multiple testing of significance is valuable because it provides an upper bound on the overall nominal significance level regardless of the proportions of the true null hypotheses and the correlations among hypotheses. For SNP-by-SNP analyses, the z -score test was used simultaneously for a large number of contiguous SNP loci. The Bonferroni correction was overly conservative in this application. The reasons for this are many. Most importantly, the

neighboring SNPs are not mutually independent, nor are the corresponding hypothesis tests. Thus, we chose a nominal significance level of 1×10^{-3} , instead of the Bonferroni correction 7.4×10^{-6} ($=0.05/6,745$), for the preliminary SNP-by-SNP analyses; this is presented above.

The nominal significance levels that are chosen for the proposed statistical methods may govern the outcomes of the analyses as shown in Fig. 1 and Table 1. In this application, the choice of appropriate nominal significance levels depends on many factors, including the (1) degree or extent of genetic heterogeneity, (2) values of prevalence and penetrance, (3) size of disease-associated deletions, (4) density and number of SNPs tested, (5) sample sizes of cases and controls, and (6) magnitudes of correlation between adjacent or neighboring SNPs. Because several of these factors are involved in the etiology of the disease of interest, which is usually unknown before the analysis, it is difficult to pre-determine the most appropriate nominal significance level. In response to this problem and to elucidate the corresponding effect, we suggest analyzing the data and comparing the outcomes using various nominal significance levels. Therefore, in addition to the nominal significance level of 1×10^{-3} for the preliminary analysis, we used the nominal significance levels of 5×10^{-4} and 1×10^{-4} , as described below.

When we used a nominal significance level of 5×10^{-4} , which corresponds to a z-score of 3.29, 11 SNP loci were statistically significant in the SNP-by-SNP analyses. Six of them appeared in the disease-associated deletion of 300 kb. In this case, the deletion detection rate for association with disease was 0.55 ($=6/11$) per SNP compared with 0.32 ($=6/19$) per SNP at the nominal significance level of 1×10^{-3} . The 5 SNP loci located outside the disease-associated deletion had z-scores of ≥ 3.29 , indicating that the false-positive deletion detection rate for association with disease was 7.45×10^{-4} ($=5/(6,745 - 33)$) per SNP compared with 1.94×10^{-3} ($=13/(6,745 - 33)$) per SNP at the nominal significance level of 1×10^{-3} . Only 1 cluster had 6 neighboring significant SNPs from the 2,992nd SNP to 3,006th SNP. This cluster was covered by a short segment that contained a total of 15 SNP loci. Using formula (1) to calculate the p -value of this cluster, we obtained

$Pr(X \geq 6|w=15, p_0=\frac{11}{6,745}=1.63 \times 10^{-3})=9.27 \times 10^{-14}$. We applied the Bonferroni-type correction to obtain the adjusted p -value by multiplying the above p -value with 6,745; this gave us an adjusted p -value of 6.25×10^{-10} .

We used a nominal significance level of 1×10^{-4} , which corresponds to a z-score of 3.72; 5 SNP loci were statistically significant in the SNP-by-SNP analyses (2,286, 2,994, 2,995, 2,997, and 2,999th SNPs with z-scores of 3.89, 4.05, 4.14, 3.79, and 3.97, respectively). Only 1 cluster had 4 neighboring significant SNPs from the 2,994th SNP through the 2,999th SNP; this was covered by a short segment of 6 SNP loci and appeared to be in the disease-associated deletion. In this application, the deletion detection rate was 0.80 ($=4/5$) per SNP, and the false-positive deletion detection rate was 1.49×10^{-4} ($=1/(6,745 - 33)$) per SNP. We used expression (1) to calculate the p -value and obtained

$Pr(X \geq 4|w=6, p_0=\frac{5}{6,745}=7.41 \times 10^{-4})=4.52 \times 10^{-12}$. The adjusted p -value of this cluster was 3.05×10^{-8} .

Comparison of proposed cluster test and scan test

We also applied the scan test to our data and calculated the corresponding p -values of the largest cluster using different nominal significance levels. The model of the scan test that we applied here is based on the assumption of a uniform distribution of events; that is, the model assumes that the times of events are distributed independently of each other and are equally likely to occur at any point in the time period (Wallenstein and Neff 1987). Table 2 shows the detection rates and false-positive detection rates for association with disease for

SNP-by-SNP analyses and the outcomes of the cluster analyses using our proposed cluster test and the scan test at three different nominal significance levels. Because the p -value formulae for the scan test have been adjusted accounting for multiple comparison procedures, we presented the adjusted p -values of the proposed cluster test in comparison with those of the scan test in the Table 2.

For the cluster analyses based on combined evidence between SNPs, both our proposed method and the scan test showed high statistical significance at all three nominal significance levels when used to detect the observed largest cluster of significant SNPs, as shown in Table 2. These clusters were all located in the range of disease-associated deletion centered at the 3,000th SNP. The adjusted p -values of our proposed cluster test with a Bonferroni-type correction are very similar to but slightly higher than the corresponding values of the scan test across various nominal significance levels.

In contrast with the scan test that is structured to detect the largest cluster only, our proposed statistical test for cluster analyses is designed to assess the statistical significance of multiple clusters. As shown in Table 1, we found four observed clusters, and our proposed cluster test can be used to test for each of them. The first cluster, which was located outside the range of disease-associated deletion, was significant at the nominal significance level of 0.01. Thus, it can be classified as a false-positive finding. The second cluster lay in the range of disease-associated deletion and was a true-positive finding. The third and fourth clusters were not significant at the nominal significance level of 0.05.

For SNP-by-SNP analyses, the z -score test was most efficient at the nominal significance level of 1×10^{-4} on the basis of a highest detection rate of 0.80 and a lowest false-positive detection rate of 1.49×10^{-4} , as shown in Table 2. In contrast, it was least efficient at the nominal significance level of 1×10^{-3} , as the detection rate of 0.32 was the lowest and the false-positive detection rate of 1.94×10^{-3} was the highest.

Assessment of linkage disequilibrium effects

To determine the effects of LD on our proposed methods, we evaluated the statistical sensitivity of the methods from various magnitudes of LD, adjusting for deletion length. We generated several simulation data sets with Max-LD = 0.0, 0.10, and 0.15 and MaxGap = 10, 30, and 50 kb. We created three sets of disease-associated deletions with sizes that followed uniform distributions (mean values of 100, 300, and 500 kb with the size ranges of [80, 120], [280, 320], and [480, 520], respectively). These deletions were centered at the 1,000, 3,000, and 5,000th SNPs, respectively. The population prevalence of the disease and the prevalence and penetrance of the disease-associated deletion were set to be the same as those described in the data simulations section.

In contrast to the previous analysis that was based on 1 replicate of simulation, we assessed the effects of LD on our proposed methods based on 100 replicates. We performed the analysis for each data set with 300 cases and 300 controls selected per replicate. Table 3 shows the mean number of significant SNP loci that had z -scores of ≥ 3.08 , the mean number of SNP loci with z -scores of ≥ 3.08 and in the range of disease-associated deletion, and the number of SNP loci encompassed by the mean size of disease-associated deletion, adjusting for various values of MaxLD, MaxGap, and deletion size. Using the data shown in Table 3, we calculated the corresponding deletion detection rates and false-positive rates (shown in Table 4).

The proposed z -score test for SNP-by-SNP analyses is robust in the presence of LD. We first assessed the effect of MaxLD, adjusting for MaxGap and deletion size, and found that the performance of SNP-by-SNP analyses when Max-LD = 0.15 was similar to that when

MaxLD = 0.10. In fact, the z-score test is generally better or more robust in terms of deletion detection rates, when MaxLD = 0.15 shown in Table 4. Similarly, we evaluated the effect of MaxGap, after controlling for deletion size and MaxLD, and found that the proposed method was more effective when MaxGap = 0.30 than when MaxGap = 0.50. Both MaxLD = 0.15 and MaxGap = 30 resulted in a higher SNP density than did MaxLD = 0.10 and MaxGap = 50, suggesting that the proposed method can be increasingly effective at an increased SNP density except when MaxLD = 0.15, MaxGap = 30, and mean deletion size = 300 kb. When the LD was negligible (MaxLD ~ 0.0), the z-score test provided better deletion detection rates when MaxGap = 10 kb than when MaxGap = 30 kb, as it had denser SNPs.

The size of disease-associated deletions is apparently important to the sensitivity of the proposed z-score test. We investigated this statistical property using various-sized deletions located at three different sites. Longer disease associated deletions contained substantially more significant SNPs (that is, SNP loci with z-scores of ≥ 3.08 that appeared in the range of the disease-associated deletion) than did shorter ones (shown in Table 3), after adjusting for the values of MaxLD and MaxGap. In terms of deletion detection rates, the proposed method was substantially robust at an increased deletion size. It is noteworthy that the false-positive detection rates were relatively stable and low across various settings. Most were below 1.0×10^{-3} , as shown in Table 4.

In addition to the above study of 100 replicates of simulation data based on the nominal significance level of 1.0×10^{-3} for SNP-by-SNP analyses, we investigated the association between the choice of the nominal significance level and the sensitivity of the proposed z-score test. We calculated the corresponding deletion detection rates and false-positive finding rates using the nominal significance levels of 5.0×10^{-4} and 1.0×10^{-4} for SNP-by-SNP analyses; our results are shown in Table 5. We obtained similar conclusions to those shown in Table 2. The proposed z-score test resulted in higher deletion detection rates and lower false-positive rates at the nominal significance level of 1.0×10^{-4} than at the nominal significance levels of 1.0×10^{-3} and 5.0×10^{-4} , after adjusting for the values of deletion size, MaxLD, and MaxGap. We obtained lowest deletion detection rates and highest false-positive rates at the nominal significance level of 1.0×10^{-3} . Thus, 1.0×10^{-4} was the most appropriate nominal significance level for this application. However, as we emphasized earlier, the choice of appropriate nominal significance levels depends on several key factors related to disease etiology, hypothesis tests, and the genetic data; the factors related to the knowledge of disease etiology are generally unknown prior to analyses. Therefore, the appropriate nominal significance level should be chosen for each analysis on the basis of genetic data and the disease being studied. Our investigation illustrated and characterized the effects of this choice on the outcomes of the analyses.

It is not possible to perform an ad hoc cluster investigation in the subsequent analysis for each of the 100 replicates across various settings of deletion size, MaxLD, MaxGap, and nominal significance levels, as shown in Tables 1 and 2. However, Tables 4 and 5 show that our proposed statistical method for analyzing clusters on the basis of the combined evidence between SNPs should be particularly useful at the nominal significance level of 1.0×10^{-4} , as the deletion detection rates were highest and the false-positive finding rates were lowest. However, cluster analyses would not be useful when the size of disease-associated deletion is small relative to SNP density. In this application, a cluster is formed by 2 or more significant SNPs gathering within a short segment. In the 9th row of Table 3, the mean number of significant SNP loci that appear in the range of a 100-kb deletion is less than 2, except when MaxLD ~ 0.0 and MaxGap = 10, indicating that we are not likely to even observe clusters of significant SNPs. In this situation, the SNP density is too sparse for cluster analyses to detect such short disease-associated deletions. Except when MaxLD ~ 0.0 and MaxGap = 10, the mean distance between adjacent SNPs is about 13–16 kb, indicating

that a 100-kb deletion contains only 6–8 SNPs on average. Because the size of disease-associated deletions with a mean value of 100 kb uniformly ranges between 80 and 120 kb, nearly half of cases with deletions have deletions shorter than 100 kb. That is, the number of SNPs within the commonly shared deletion region among patients is actually even smaller than 6–8 SNPs; hampering the detection of SNP loci near the deletion boundaries. This effect is particularly substantial when disease-associated deletions are small (a deletion of 100 kb in this case) relative to SNP density compared with when they are long (a deletion of 300 or 500 kb).

Application to rheumatoid arthritis

Rheumatoid arthritis is a chronic inflammatory disease that leads to the most common form of inflammatory polyarthritis in adults. The precise etiology of the disease remains unknown, but the presence of familial aggregations of RA cases and the increased concordance in monozygotic twins compared with dizygotic twins indicate the presence of genetic components to susceptibility (Seldin et al. 1999). Four separate whole-genome scans of microsatellite data from Caucasian RA families have been performed in North America, the United Kingdom, and France (Cornelis et al. 1998; Jawaheer et al. 2001, 2003; MacKay et al. 2002). The results showed that genes in the HLA region of chromosome 6, such as DRB1, make the largest genetic contribution to RA susceptibility. To evaluate the potential role of deletions in influencing case–control status, we used SNP data from a whole-genome association analysis of 868 RA cases and 1,197 controls (Plenge et al. 2007).

We applied the proposed statistical methods to the SNP data for chromosome 6p. There were 13,964 SNPs over the region of chromosome 6p with positions ranging from 0.11 Mb through 58.82 Mb. The mean distance was 4.21 kb and mean LD was $r^2 = 0.12$ between adjacent SNPs. For SNP-by-SNP analyses, we performed the z-score test to assess the statistical significance of differences in homozygosity proportions between the 868 cases and the 1,197 controls on each of 13,964 contiguous loci. We used a nominal significance level of 1×10^{-7} that corresponds to a z-score of 5.20 and output the z-scores by their corresponding physical positions, as shown in Fig. 2. We chose a relatively stringent threshold for statistical significance at a nominal significance level of 1×10^{-7} in this study, because we had many more cases and controls and higher SNP density than those data sets we analyzed in simulation studies.

Exactly 100 SNP loci with z-scores of ≥ 5.20 were statistically significant over the whole study region of chromosome 6p at the nominal significance level of 1×10^{-7} . We identified 80 significant SNP loci gathering in a small region, ranging from 31.13 to 33.96 Mb. An expanded display of the z-scores between 30.00 and 34.00 Mb is shown in Fig. 3. We used a maximum distance of ≤ 100 kb allowed between two successive significant SNP loci to define a cluster of significant SNPs. Under this criterion, each cluster will begin with a significant SNP and end with another significant SNP; we identified six distinct clusters of multiple neighboring significant SNPs that were all located between 31.13 and 33.84 Mb. Each of the six clusters is described and shown in Table 6. Most notably, there were 37 significant SNPs over a segment of 157 successive SNPs for the third cluster. Because the two ends of the cluster segment were significant SNPs, 35 significant SNPs were allocated inside the cluster. In this case, $w = 157$ and $x = 37$ were used for formula (1). The fourth cluster was also large with $w = 183$ and $x = 28$ under the criterion. We used expression (1) to calculate the p -values of these clusters; the results are shown in Table 6. The p -values of the first, second, fifth, and sixth clusters were higher than the adjusted p -value threshold 3.58×10^{-6} ($=0.05/13,964$) at the nominal significance level of 0.05; therefore, they were not statistically significant. The third and fourth clusters appeared to be the largest clusters of significant SNPs, and both were highly significant, as their adjusted p -values were $3.31 \times$

10^{-40} ($=2.37 \times 10^{-44} \times 13,965$) and 3.44×10^{-24} ($=2.46 \times 10^{-28} \times 13,965$), respectively. We applied the scan test to the third cluster, the largest one under the given criterion, and obtained the p -value of 4.09×10^{-42} . Thirty-seven significant SNP loci lay in the third cluster of 302 kb and 28 lay in the fourth cluster of 260 kb. The third and fourth clusters, separated by ~144 kb, were closer to each other than were any other adjacent clusters.

Genomic deletions in the region of HLA_DRB1 (32.36 Mb) are a common characteristic of HLA class II haplotypes, and the major DR4 and DR9 haplotypes associated with RA belong to related haplotype family with multiple DRB loci, including several pseudogenes (Spies et al. 1985; Beck and Trowsdale 1999). In contrast, the DR1 haplotypes associated with RA are members of a distinct family of haplotypes that have much fewer DRB loci. Thus, we would expect to find differences in copy number of DRB genes between RA cases and controls, which contain haplotype families with more variable numbers of DRB loci. HLA_DRB1 and HLA_DQB1 (32.74 Mb) are located in the third (between 32.23 and 32.54 Mb) and fourth clusters (between 32.68 and 32.94 Mb), respectively. Compared with HLA_DRB1, the evidence of genomic deletions and CNVs at HLA_DQB1 has been less established. The outcomes of our proposed cluster analysis should enable investigators to delineate the minimal extent of disease-associated deletions that are common among RA patients.

In addition to the outcomes of cluster analyses, our proposed SNP-by-SNP analyses detected 80 significant SNP loci between 31.13 and 33.96 Mb at the nominal significance level of 1×10^{-7} , which overlaps the HLA region of 6p21. The SNP-by-SNP analyses identified strong evidence of disease-associated deletions in the regions of HLA_C (31.35 Mb) and C4 (32.01 Mb). The deletions and CNVs in the C4 region have also been identified previously (Beck and Trowsdale 1999). Thus, our proposed SNP-by-SNP analyses and cluster analyses were useful for detecting disease associated-deletions in regions of HLA_DRB1 and C4. Further investigation into the correlations between these regions, HLA_DR, and HLA_DQ in association with RA is warranted.

Discussion

Whole-genome studies of CNVs—and genomic deletions in particular—have been performed extensively over the past few years. Many of these studies focus on investigating genetic variations in non-diseased individuals and can provide fundamental and profound resource of baseline information for the study of human disease and genomic evolution (Conrad et al. 2006; Hinds et al. 2006; McCarroll et al. 2006; Redon et al. 2006; Wong et al. 2007). These studies have shown that various-sized genomic deletions and CNVs are widespread and account for considerable genomic variations in the human genome. However, assessing these effects and associating them with susceptibility to common, complex diseases are challenging (Freeman et al. 2006; Redon et al. 2006; Stranger et al. 2007). Existing statistical or molecular approaches for detecting genomic deletions in whole-genome studies are generally based on the observation of Mendelian transmission failure patterns in parent-child trios, Hardy–Weinberg disequilibrium, or null genotypes for successive genetic markers (McCarroll et al. 2006). However, genotyping errors and missing data can also account for deviations from the Hardy–Weinberg equilibrium and Mendelian transmission inheritance. Therefore, testing for Hardy–Weinberg equilibrium and Mendelian transmission inheritance for the purpose of quality control filtering in genetic markers can be confounded with and hindered by discovering deletions using these approaches.

In contrast with recent whole-genome studies that focused on detecting the “benign” deletions in healthy individuals, our study was designed to identify genomic deletions associated with common, complex diseases. In contrast with the existing statistical

approaches developed by Amos et al. (2003) and Kohler and Cutler (2007) that were structured to detect disease-associated deletions from familial studies of parent-offspring trios, our proposed statistical methods were designed for case-control studies.

We proposed two complementary statistical methods for identifying disease-associated deletions in high-density SNP data from whole-genome association studies: the z-score test for SNP-by-SNP analyses and the statistical test for cluster analyses based on combined evidence between SNPs, which is structured to subsequent investigation into outcomes of SNP-by-SNP analyses. The identification of excess homozygosity regions in multiple cases forms the basis of our proposed methods. The evidence of genomic deletions associated with a disease can be enhanced by the observation of multiple neighboring SNPs with excess homozygosity in cases compared with in controls. Our proposed methods for case-control studies do not confound with filtering of genotyping accuracy, nor are they limited to identify inherited deletions from family data.

In this report, we used the deletion detection rate and false-positive deletion detection rate to assess the performance of our proposed statistical approaches instead of calculating empirical statistical power. These and related measures are commonly used to assess performance in genome-wide association studies (Gail et al. 2008).

We demonstrate that our proposed methods are useful for detecting disease-associated deletions and are robust in the presence of LD. Under various settings of LD magnitude and disease-associated deletion size on dense simulated SNP genotype data, the z-score test was effective at identifying disease-associated deletions in terms of the deletion detection rate and the false-positive detection rate. More importantly, its performance was increasingly efficient at an increased SNP density, after adjusting for disease-associated deletion size. We found that the proposed cluster test could evaluate multiple observed clusters of successive or neighboring significant SNPs and delineate common deleted regions among patients as long as SNP density was not too sparse relative to deletion sizes. In contrast, the existing scan test is structured to detect and assess only the largest cluster. Because our proposed statistical methods are designed to simultaneously test of significance for a large number of SNPs, we determined and articulated the effects of multiple comparison procedures on outcomes of the analyses, corresponding corrections, and choice of appropriate nominal significance levels jointly.

We applied the proposed statistical methods to 13,964 contiguous SNP loci of chromosome 6p from 868 RA patients and 1,197 controls. We detected several significant SNP loci at a nominal significance level of 1×10^{-7} . We then performed the cluster test for the outcomes of the SNP-by-SNP analyses and identified two clusters of neighboring significant SNPs. The chromosomal segments that contained disease-associated deletions detected by the SNP-by-SNP analyses and cluster analyses included the regions of HLA_DRB1 and C4 in which deletions were previously discovered. Our proposed statistical methods were useful and effective for identifying disease-association deletions in this analysis. Compared with existing approaches in the literature, our proposed statistical test for cluster analyses not only provides enhanced evidence of disease-associated deletions but can also delineate or define the extent of the minimal region of common deletions among patients, indicating the critical region of the disease.

Many recent genetic studies have found that genomic deletions can play a causative and crucial role in psychiatric disorders; therefore, efficient and powerful whole genome approaches are needed to detect interstitial deletions associated with common, complex diseases. More and more high-density SNP genotype data are becoming available for genome-wide association studies; thus, statistical methods are needed that use these

genotype data to detect disease-associated deletions. To our knowledge, ours are the first statistical approaches to case–control studies that systematically focus on whole-genome detection of genomic deletions associated with disease. In this report, we proposed, illustrated, and elucidated SNP-based statistical methods for detecting and assessing disease associated genomic deletions using dense SNP genotypes of genome-wide association studies without the need for additional SNP genotyping.

Currently, four methods have been used to validate candidate deletions: fluorescent in situ hybridization (FISH), two-color fluorescence intensity measurements, PCR amplification, and quantitative PCR (McCarroll et al. 2006). These methods can be used to distinguish true deletions from homozygous genotypes in association with disease. When molecular technologies such as SNP genotyping methods become better able to identify deletions, we may be capable of distinguishing hemizygous genotypes from homozygous genotypes on a genomewide scale. In this situation, our proposed statistical approaches will be more useful and specific because the hemizygous deletions will not be confounded with homozygous genotypes. Thus, we will be able to more accurately assess their association with disease using SNP genotyping data.

Acknowledgments

This research was supported by the US National Cancer Institute grants 2P01-CA034936 and 1R03-CA128103.

References

- Amos CI, Shete S, Chen J, Yu RK. Positional identification of microdeletions with genetic markers. *Hum Hered* 2003;56:107–118. [PubMed: 14614244]
- Beck S, Trowsdale J. Sequence organization of the class II region of human MHC. *Immunol Rev* 1999;167:201–210. [PubMed: 10319262]
- Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet* 2006;38(1):75–81. [PubMed: 16327808]
- Cornelis F, Faure S, Martinez M, Prud'homme JF, Fritz P, Dib C, Alves H, Barrera P, de Vries N, et al. New susceptibility locus for rheumatoid arthritis suggested by a genome-wide linkage study. *Proc Natl Acad Sci USA* 1998;95:10746–10750. [PubMed: 9724775]
- Franke L, de Kovel CG, Aulchenko YS, Trynka G, Zernakova A, Hunt KA, Blauw HM, van den Berg LH, Ophoff R, Deloukas P, et al. Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. *Am J Hum Genet* 2008;82(6):1316–1333. [PubMed: 18519066]
- Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, et al. Copy number variation: new insights in genome diversity. *Genome Res* 2006;16:949–961. [PubMed: 16809666]
- Gail MH, Pfeiffer RM, Wheeler W, Pee D. Probability of detecting disease-associated single nucleotide polymorphisms in case–control genome-wide association studies. *Biostatistics* 2008;9:201–215. [PubMed: 17873152]
- Grimson RC. Disease clusters, exact distributions of maxima, and *p*-values. *Stat Med* 1993;12:1773–1794. [PubMed: 8272660]
- Grimson RC, Mendelsohn S. A method for detecting current temporal clusters of toxic events through data monitoring by poison control center. *J Toxicol Clin Toxicol* 2000;38(7):761–765. [PubMed: 11192463]
- Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet* 2006;38(1):82–85. [PubMed: 16327809]
- Hoggart CJ, Chadeau-Hyam M, Clark TG, Lampariello R, Whittaker JC, De Iorio M, Balding DJ. Sequence-level population simulations over large genomic regions. *Genetics* 2007;177(3):1725–1731. [PubMed: 17947444]

- Jawaheer D, Seldin MF, Amos CI, Chen WV, Shigeta R, Monteiro J, Kern M, Criswell LA, Albani S, Nelson JL, et al. A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases. *Am J Hum Genet* 2001;68:927–936. [PubMed: 11254450]
- Jawaheer D, Seldin MF, Amos CI, Chen WV, Shigeta R, Etzel C, Damle A, Xiao X, Chen D, Lum RF, et al. Screening the genome for rheumatoid arthritis susceptibility genes: a replication study and combined analysis of 512 multicase families. *Arthritis Rheum* 2003;48:906–916. [PubMed: 12687532]
- Kohler JR, Cutler DJ. Simultaneous discovery and testing of deletions for disease association in SNP genotyping studies. *Am J Hum Genet* 2007;81(4):684–699. [PubMed: 17846995]
- Lindsay EA. Chromosomal microdeletions: dissecting del22q11 syndrome. *Nat Rev Genet* 2001;2(11):858–868. [PubMed: 11715041]
- Liu H, Abecasis GR, Heath SC, Knowles A, Demars S, Chen YJ, Roos JL, Rapoport JL, Gogos JA, Karayiorgou M. Genetic variation in the 22q11 locus and susceptibility to schizophrenia. *Proc Natl Acad Sci USA* 2002;99(26):16859–16864. [PubMed: 12477929]
- MacKay K, Eyre S, Myerscough A, Milicic A, Barton A, Laval S, Barrett J, Lee D, White S, John S, et al. Whole-genome linkage analysis of rheumatoid arthritis susceptibility loci in 252 affected sibling pairs in the United Kingdom. *Arthritis Rheum* 2002;46:632–639. [PubMed: 11920398]
- McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, et al. International HapMap Consortium. Common deletion polymorphisms in the human genome. *Nat Genet* 2006;38(1):86–92. [PubMed: 16468122]
- McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. *Science* 2004;304(5670):581–584.
- Peng B, Amos CI, Kimmel M. Forward-time simulations of human populations with complex diseases. *PLoS Genet* 2007;3(3):e47. [PubMed: 17465681]
- Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, Liew A, Khalili H, Chandrasekaran A, Davies LR, et al. TRAF1-C5 as a risk locus for rheumatoid arthritis—a genome-wide study. *N Engl J Med* 2007;357:1199–1209. [PubMed: 17804836]
- Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. Global variation in copy number in the human genome. *Nature* 2006;444:444–454. [PubMed: 17122850]
- Rovelet-Lecrux A, Hannequin D, Raux G, Le Meur N, Laquerrière A, Vital A, Dumanchin C, Feuillet S, Brice A, Vercelletto M, et al. APP locus duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat Genet* 2006;38(1):24–26. [PubMed: 16369530]
- Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 2005;15(11):1576–1583. [PubMed: 16251467]
- Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, et al. Strong association of de novo copy number mutations with autism. *Science* 2007;316:445–449. [PubMed: 17363630]
- Seldin MF, Amos CI, Ward R, Gregersen PK. The genetics revolution and the assault on rheumatoid arthritis. *Arthritis Rheum* 1999;42:1071–1079. [PubMed: 10366098]
- Shifman S, Levit A, Chen ML, Chen CH, Bronstein M, Weizman A, Yakir B, Navon R, Darvasi A. A complete genetic association scan of the 22q11 deletion region and functional evidence reveal an association between DGCR2 and schizophrenia. *Hum Genet* 2006;120(2):160–170. [PubMed: 16783572]
- Spies T, Sorrentino R, Boss JM, Okada K, Strominger JL. Structural organization of the DR subregion of the human major histocompatibility complex. *Proc Natl Acad Sci USA* 1985;82:5165–5169. [PubMed: 3860851]
- Stefansson H, Rujescu D, Cichon S, Pietiläinen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, et al. Large recurrent microdeletions associated with schizophrenia. *Nature* 2008;455:232–236. [PubMed: 18668039]

- Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 2007;315:848–853. [PubMed: 17289997]
- Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE. Fine-scale structural variation of the human genome. *Nat Genet* 2005;37:727–732. [PubMed: 15895083]
- Wall JD, Przeworski M. When did the human population size start increasing? *Genetics* 2000;155(4):1865–1874. [PubMed: 10924481]
- Wallenstein S, Neff N. An approximation for the distribution of the scan statistic. *Stat Med* 1987;6(2):197–207. [PubMed: 3589249]
- Wang TL, Maierhofer C, Speicher MR, Lengauer C, Vogelstein B, Kinzler KW, Velculescu VE. Digital karyotyping. *Proc Natl Acad Sci USA* 2002;99:16156–16161. [PubMed: 12461184]
- Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MA, Green T, et al. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* 2008;358(7):667–675. [PubMed: 18184952]
- Wong KK, de Leeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE, et al. A comprehensive analysis of common copy-number variations in the human genome. *Am J Hum Genet* 2007;80(1):91–104. [PubMed: 17160897]
- Wu CC, Grimson RC, Amos CI, Shete S. Statistical methods for anomalous discrete time series based on minimum cell count. *Biom J* 2008;50(1):86–96. [PubMed: 17853406]
- Yu CE, Dawson G, Munson J, D'Souza I, Osterling J, Estes A, Leutenegger AL, Flodman P, Smith M, Raskind WH, et al. Presence of large deletions in kindreds with autism. *Am J Hum Genet* 2002;71(1):100–115. [PubMed: 12058345]

Fig. 1.

z-Scores of SNP-by-SNP analyses with MaxLD = 0.15 and MaxGap = 30 kb for 300 cases and 300 controls. A z-score of 3.08, which corresponds to a nominal significance level of 1×10^{-3} , was used as a threshold value. Several “hits” with z-scores of ≥ 3.08 were clustered in the range of the disease-associated deletion, which is indicated by a red arrow. The location of disease-associated deletions is centered at the 3,000th SNP

Fig. 2.

z-Scores of SNP-by-SNP analyses for rheumatoid arthritis on chromosome 6p. A z-score of 5.20, which corresponds to a nominal significance level of 1×10^{-7} , was used as a threshold value for SNP-by-SNP analyses. Exactly 100 SNP loci had z-scores of ≥ 5.20 and were statistically significant at the nominal significance level of 1×10^{-7} .

Fig. 3.

Amplification of z-scores of SNP-by-SNP analyses for rheumatoid arthritis on chromosome 6p between 30 and 34 Mb. Eighty significant SNP loci with z-scores of ≥ 5.20 gathered in a small chromosomal region, ranging from 31.13 to 33.96 Mb

Table 1

Cluster analysis of simulated SNP data with $\text{MaxLD} = 0.15$ and $\text{MaxGap} = 30$ kb for 300 cases and 300 controls at a nominal significance level of 10^{-3} for SNP-by-SNP analyses

Cluster number	1	2	3	4
No. of significant SNP loci	3	6	2	2
Observed cluster location	2,284–2288th	2,992–3006th	3,858–3867th	5,134–5148th
No. of SNP loci encompassed the cluster	5	15	10	15
<i>P</i> -value of Formula (1)	2.23×10^{-7}	2.46×10^{-12}	3.53×10^{-4}	8.15×10^{-4}

Nineteen SNP loci were statistically significant at a nominal significance level of 1×10^{-3} , 13 of which were included in the above four clusters. The disease-associated deletion of 300 kb was centered at the 3,000th SNP and encompassed the segment from the 2989th SNP to the 3021st SNP. The *p*-values shown in the fifth row represent the figures using expression (1), with no Bonferroni correction

Table 2

Sensitivity of detecting deletions in association with disease at various nominal significance levels with MaxLD = 0.15 and MaxGap = 30 kb for 300 cases and 300 controls

Nominal significance level	1×10^{-3}	5×10^{-4}	1×10^{-4}
Deletion detection rate	0.32	0.55	0.80
False-positive deletion detection rate	1.94×10^{-3}	7.45×10^{-4}	1.49×10^{-4}
Location of observed largest cluster	2992nd–3006th	2992nd–3006th	2994th–2999th
Adjusted <i>P</i> -value of Formula (1)	1.66×10^{-8}	6.25×10^{-10}	3.05×10^{-8}
<i>P</i> -value of scan test	8.55×10^{-9}	1.49×10^{-10}	1.41×10^{-8}

The table shows the *p*-values of the largest clusters of significant SNPs for our proposed cluster test with Bonferroni corrections and a scan test performed at three different nominal significance levels for SNP-by-SNP analyses

Effects of linkage disequilibrium on SNP-by-SNP analyses at a nominal significance level of 10^{-3} for 300 cases and 300 controls based on 100 replicates

Table 3

MaxLD	0.15	0.15	0.10	0.10	0.10	~0.0	~0.0
MaxGap (kb)	50	30	50	30	30	30	10
Mean LD	0.041	0.064	0.034	0.034	0.060		
Mean Gap (kb)	14.0	13.3	15.4	14.5	14.5	14.5	7.6
Total number of SNP Loci tested	6,436	6,745	5,845	6,213	6,213	6,213	11,904
Mean deletion size of 100 kb							
No. of SNP loci Encompassed by deletion	11	9	7	6	6	6	14
Mean no. of SNP loci had z-scores of ≥ 3.08	7.31	7.34	6.33	7.14	6.69	6.69	17.30
Mean no. of SNP loci had z-scores of ≥ 3.08 and lay within deletion	1.54	1.66	0.97	1.56	1.52	1.52	5.16
Mean deletion size of 300 kb							
No. of SNP loci encompassed by deletion	26	33	33	26	26	26	51
Mean no. of SNP loci had z-scores of ≥ 3.08	13.42	8.28	9.28	9.70	9.51	9.51	25.63
Mean no. of SNP loci had z-scores of ≥ 3.08 and lay within deletion	7.68	3.21	4.15	4.47	4.55	4.55	15.00
Mean deletion size of 500 kb							
No. of SNP loci encompassed by deletion	39	48	32	37	37	37	66
Mean no. of SNP loci had z-scores of ≥ 3.08	17.03	22.89	11.17	17.63	17.42	17.42	35.88
Mean no. of SNP loci had z-scores of ≥ 3.08 and lay within deletion	10.48	17.10	5.89	12.11	12.27	12.27	24.62

Deletion detection and false-positive rates for SNP-by-SNP analyses at a nominal significance level of 10^{-3} for 300 cases and 300 controls on the basis of 100 replicates

Table 4

MaxLD	0.15	0.15	0.10	0.10	0.10	~0.0	~0.0
MaxGap (kb)	50	30	50	30	50	30	10
Mean deletion size of 100 kb							
Deletion detection rate	0.21	0.23	0.15	0.22	0.22	0.23	0.30
False-positive rate (10^{-4})	8.98	8.43	9.18	8.99	8.99	8.33	10.2
Mean deletion size of 300 kb							
Deletion detection rate	0.57	0.39	0.45	0.46	0.46	0.48	0.59
False-positive rate (10^{-4})	8.95	7.55	8.83	8.45	8.45	8.02	8.97
Mean deletion size of 500 kb							
Deletion detection rate	0.62	0.75	0.53	0.69	0.69	0.70	0.69
False-Positive Rate (10^{-4})	10.2	8.65	9.08	8.94	8.94	8.34	9.51

Deletion detection and false-positive rates for SNP-by-SNP analyses at nominal significance levels of 5×10^{-4} and 10^{-4} for 300 cases and 300 controls on the basis of 100 replicates

Table 5

MaxLD	0.15	0.15	0.10	0.10	0.10	-0.0	-0.0
MaxGap (kb)	50	30	50	30	30	30	10
$\alpha = 5.0 \times 10^{-4}$							
Mean deletion size of 100 kb							
Deletion detection rate	0.36	0.31	0.21	0.34	0.36	0.42	
False-positive rate (10^{-4})	3.98	4.32	4.49	4.33	3.87	5.39	
Mean deletion size of 300 kb							
Deletion detection rate	0.73	0.50	0.59	0.61	0.64	0.72	
False-positive rate (10^{-4})	3.84	3.78	4.11	4.11	3.59	4.29	
Mean deletion size of 500 kb							
Deletion detection rate	0.75	0.83	0.67	0.80	0.81	0.79	
False-positive rate (10^{-4})	4.88	4.64	4.27	4.13	3.84	4.68	
$\alpha = 1.0 \times 10^{-4}$							
Mean deletion size of 100 kb							
Deletion detection rate	0.58	0.31	0.25	0.51	0.61	0.66	
False-positive rate (10^{-4})	1.21	2.08	1.94	1.59	1.03	1.61	
Mean deletion size of 300 kb							
Deletion detection rate	0.90	0.45	0.72	0.73	0.87	0.90	
False-positive rate (10^{-4})	0.83	2.23	1.50	1.75	0.71	0.83	
Mean deletion size of 500 kb							
Deletion detection rate	0.86	0.91	0.88	0.92	0.93	0.92	
False-positive rate (10^{-4})	1.55	1.49	0.86	0.89	0.78	1.12	

Table 6
Cluster analysis of rheumatoid arthritis at a nominal significance level of 10^{-7} for SNP-by-SNP analyses

Cluster number	1	2	3	4	5	6
No. of significant SNP loci	2	4	37	28	2	2
Location of observed cluster (Mb)	31.13–31.20	31.62–31.78	32.23–32.54	32.68–32.94	33.19–33.29	33.78–33.84
No. of SNP loci encompassed cluster	37	51	157	183	46	24
P-value of Formula (1)	2.90×10^{-1}	5.05×10^{-4}	2.37×10^{-44}	2.46×10^{-28}	4.32×10^{-2}	1.28×10^{-2}

The *p*-values shown in the fifth row represent the figures using expression (1) with no Bonferroni correction