ARTICLE

Detection, Imputation, and Association Analysis of Small Deletions and Null Alleles on Oligonucleotide Arrays

Lude Franke,[1,3] Carolien G.F. de Kovel,[1] Yurii S. Aulchenko,[2] Gosia Trynka,[3] Alexandra Zhernakova,[1] Karen A. Hunt,[4] Hylke M. Blauw,[5] Leonard H. van den Berg,[5] Roel Ophoff,[1,6] Panagiotis Deloukas,[7] David A. van Heel,[4] and Cisca Wijmenga[1,3,*]

Copy-number variation (CNV) is a major contributor to human genetic variation. Recently, CNV associations with human disease have been reported. Many genome-wide association (GWA) studies in complex diseases have been performed with sets of biallelic single-nucleotide polymorphisms (SNPs), but the available CNV methods are still limited. We present a new method (TriTyper) that can infer genotypes in case-control data sets for deletion CNVs, or SNPs with an extra, untyped allele at a high-resolution single SNP level. By accounting for linkage disequilibrium (LD), as well as intensity data, calling accuracy is improved. Analysis of 3102 unrelated individuals with European descent, genotyped with Illumina Infinium BeadChips, resulted in the identification of 1880 SNPs with a common untyped allele, and these SNPs are in strong LD with neighboring biallelic SNPs. Simulations indicate our method has superior power to detect associations compared to biallelic SNPs that are in LD with these SNPs, yet without increasing type I errors, as shown in a GWA analysis in celiac disease. Genotypes for 1204 triallelic SNPs could be fully imputed, with only biallelic-genotype calls, permitting association analysis of these SNPs in many published data sets. We estimate that 682 of the 1655 unique loci reflect deletions; this is on average 99 deletions per individual, four times greater than those detected by other methods. Whereas the identified loci are strongly enriched for known deletions, 61% have not been reported before. Genes overlapping with these loci more often have paralogs ($p = 0.006$) and biologically interact with fewer genes than expected ($p = 0.004$).

Introduction

It has become apparent that copy-number variation (CNV) accounts for a considerable amount of genetic variation[1–5] and has been implicated as a causal mechanism for several disorders.[6–8] Specialized comparative genomic hybridization (CGH) arrays that contain large-insert clones that hybridize to complementary DNA[1,5,9,10] have provided much insight into the properties of CNVs. These studies have shown that individuals usually carry many small-deletion and duplication CNVs that can be found with high population frequencies.

Recently, much effort has been devoted to detecting CNVs with single-nucleotide polymorphism (SNP) genotype data in both familial and unrelated samples.[2,4,11–19] An important resource so far has been the HapMap project,[20] in which over three million SNPs have been typed for 270 samples. In addition, growing resources of genotype data from oligonucleotide arrays that usually assay at least 300,000 SNPs have been generated for genome-wide association (GWA) studies. Although there are technical challenges to detecting CNVs with these arrays,[21] various methods have been developed. Some have been designed to work on single samples,[13,14,17–19,22] using similar principles as used for array CGH, whereas others take multiple samples jointly into consideration.[2,4,15,22] The single-sample methods typically require that multiple, consecutive (usually at least three) SNPs show deviations in the allele intensity signals. When multiple samples are analyzed together, genotype calls, based on biallelic SNP assumptions, can provide circumstantial evidence that CNVs span these SNPs. SNPs that map within common CNVs are expected to show deviations from Hardy-Weinberg equilibrium (HWE) and an increased number of missing genotype calls. If family data are present, a control for Mendelian segregation is routinely performed. Usually this is done to determine genotyping accuracy, but if for a given SNP segregation inconsistencies are observed, these can also be caused by violations of the assumption that the SNP is biallelic: Duplications, deletions, or the presence of a third allele at the locus that is not labeled by the assay can all lead to observations of Mendelian inconsistency.

One limitation of the available CNV detection methods is the resolution because nearly all require that multiple consecutive SNPs show aberrant intensity characteristics.[4,13,14,16–19,22] One method has a resolution as high as a single SNP,[15] but it can only be applied to families.

Here, we describe a new genotype-calling method ("*TriTyper*") that can reliably detect deletions in unrelated samples that span only one SNP. Our algorithm detects SNPs

[1]Complex Genetics Section, DBG-Department of Medical Genetics, University Medical Centre Utrecht, 3584 CG Utrecht, The Netherlands; [2]Department of Epidemiology & Biostatistics, Erasmus MC Rotterdam, 3000 CA Rotterdam, The Netherlands; [3]Genetics Department, University Medical Centre Groningen and University of Groningen, 9700 RB Groningen, The Netherlands; [4]Institute of Cell and Molecular Science, Barts and The London School of Medicine and Dentistry, London, E1 2AT, UK; [5]Department of Neurology, Rudolf Magnus Institute of Neuroscience, University Medical Center Utrecht, 3584 CX Utrecht, The Netherlands; [6]Center for Neurobehavioral Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095, USA; [7]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK
*Correspondence: c.wijmenga@umcutrecht.nl
DOI 10.1016/j.ajhg.2008.05.008.

with an extra, untyped allele (including deletion CNVs encompassing these SNPs) with raw intensity data from Illumina Infinium HumapHap300 and HumanHap550 BeadChip arrays.[23] Using *TriTyper*, we identified 1880 SNPs with a common extra allele (frequency >0.5%) in a collection of 3102 DNA samples from individuals of Northwest European origin. Our method can accurately assign genotypes by utilizing local linkage disequilibrium (LD) with nearby SNPs.[1,24,25] We show that our procedure results in correct genotype assignments through a Mendelian segregation analysis in white European HapMap trios, in which many segregation inconsistencies, observed under biallelic-calling assumptions, are resolved when triallelic genotypes have been assigned. Of the 1880 triallelic SNPs, 1204 can be fully imputed from surrounding SNPs without the need to use raw intensity data. This is helpful when analyzing triallelic SNPs in publicly available and other data sets for which only genotype calls have been made available. We show how these triallelic genotypes can be used for association studies and that our test statistic shows no inflation in significant signals as exemplified in an analysis of celiac disease (MIM 212750). Yet, like other imputation methods,[26,27] our method has superior power to detect true positive associations, when contrasted to an association analysis of nearby biallelic SNPs, used for imputing the triallelic SNPs. The identified triallelic loci are strongly enriched for known deletions, but the majority of identified deletions have not yet been described. We support previous findings that genes, mapping within these deletions, more often have paralogs, but we also found that the genes usually tend to interact biologically with fewer genes than expected. With *TriTyper*, more genetic information can be captured, triallelic SNP genotypes can be imputed, and interesting phenomena, including small-deletion CNVs, can be detected in numerous case-control cohorts that have already been typed on oligonucleotide platforms.

Material and Methods

Triallelic-Genotype-Calling Algorithm

Oligonucleotide assays, available for high-throughput SNP genotyping, usually measure the intensities of two fluorescent labels that are attached to two known alleles, A and B. Throughout this paper, these are plotted on the x axis ($intensity_a$) and y axis ($intensity_b$), respectively. When an extra, untyped allele (a "null" or 0 allele) is present, up to six clusters (representing AA, AB, BB, A0, B0, and 00 genotypes) in the raw intensity plot will become visible (Figure 1A). Usually, these A0 and B0 clusters partly overlap with the AA and BB clusters, respectively, whereas the 00 cluster has a very low Euclidian intensity. We refer to this as a "triallelic" pattern or "triallelic" SNP. If the presence of this null allele is not recognized, standard calling algorithms will typically call A0 and B0 genotypes as AA and BB, respectively, and 00 genotypes as "failed." Under biallelic assumptions, deviations from HWE are then likely to become apparent.

We used these deviations under biallelic assumptions as the basis for our triallelic genotype-calling algorithm (*TriTyper*). *TriTyper* extends a biallelic genotype-calling algorithm we have recently developed[28] and models triallelic genotypes by using a maximum-likelihood estimation (MLE) procedure that optimizes HWE under triallelic assumptions[29] (Figures 1A–1D; for details, see Appendix A). Another key aspect of our method is that it uses the presence of local LD between this null allele and nearby biallelic SNPs[1,24,25] to gain evidence that the extra allele has been correctly identified. Once this has been established, it takes advantage of these biallelic SNPs to improve the triallelic-genotype assignments by using a fairly straightforward imputation method (Figures 1E–1G; for details see Appendix A) that borrows some ideas from methods that impute genotypes for biallelic SNPs.[26,27] This imputation methodology often allows for accurately discriminating between A0 and AA and between B0 and BB samples; such discrimination is particularly helpful because these clusters usually overlap somewhat (Figure 1G, green arrow).

Data Sets for Triallelic-SNP Discovery

Initial analyses were performed on a cohort that comprised 1422 unrelated control individuals[28] from the 1958 British birth cohort that passed quality control (QC) and had been typed on the Illumina Infinium II Human Hap550 BeadChip platform for 571,738 SNPs. To also detect triallelic SNPs with lower null-allele frequencies, we added three more cohorts. These included 778 unrelated UK celiac disease cases,[28] 450 unrelated Dutch controls,[30] and 472 unrelated Dutch amyotrophic lateral sclerosis (MIM 105400) cases[30] that all passed QC and had been typed on the Illumina Infinium II Human Hap300 BeadChip platform for 317,503 SNPs. In this combined analysis, 313,505 SNPs could be analyzed because they were present on both the Hap300 and Hap550 platforms. A total of 20 samples (0.6%) showed aberrant intensity signals for many of the triallelic SNPs and were removed from the analyses.

Association Analysis

An analysis for marginal association effects on the biallelic SNPs used for imputation of the triallelic SNPs was performed as follows: Analyses were confined to SNPs for which the null allele was not in complete LD with a biallelic SNP because for these SNPs, Fisher's exact test for association would be identical to the association analysis of the triallelic null allele. Only triallelic SNPs, in which one biallelic SNP could help to discriminate between A0 and AA genotypes and another biallelic SNP could help to discriminate between B0 and BB genotypes, were included in the analysis.

To assess the marginal effect on the SNPs used for imputing the triallelic SNPs, we simulated three different scenarios of triallelic SNP association (Fisher's exact test for the triallelic null allele of 10^{-4}, 10^{-6}, and 10^{-8}). For each triallelic SNP, an equal number of controls and cases were chosen, but case and control labels were assigned in such a way that association for the triallelic SNP yielded a Fisher's exact p value for the null-allele that approximated the p value of the scenario under investigation. This allowed for determining the marginal association effect on the two biallelic SNPs used for imputing each triallelic SNP. We repeated this 100 times to gain accurate estimates. Subsequently, for each triallelic SNP, the average marginal effect on the biallelic SNP that was associated most significantly was recorded. Once this was performed for all the triallelic SNPs, the median marginal effect could be determined for each scenario.

The triallelic SNP null-allele association analysis was performed on a celiac disease GWA data set[28] and was confined to those triallelic SNPs for which imputation could help to discriminate

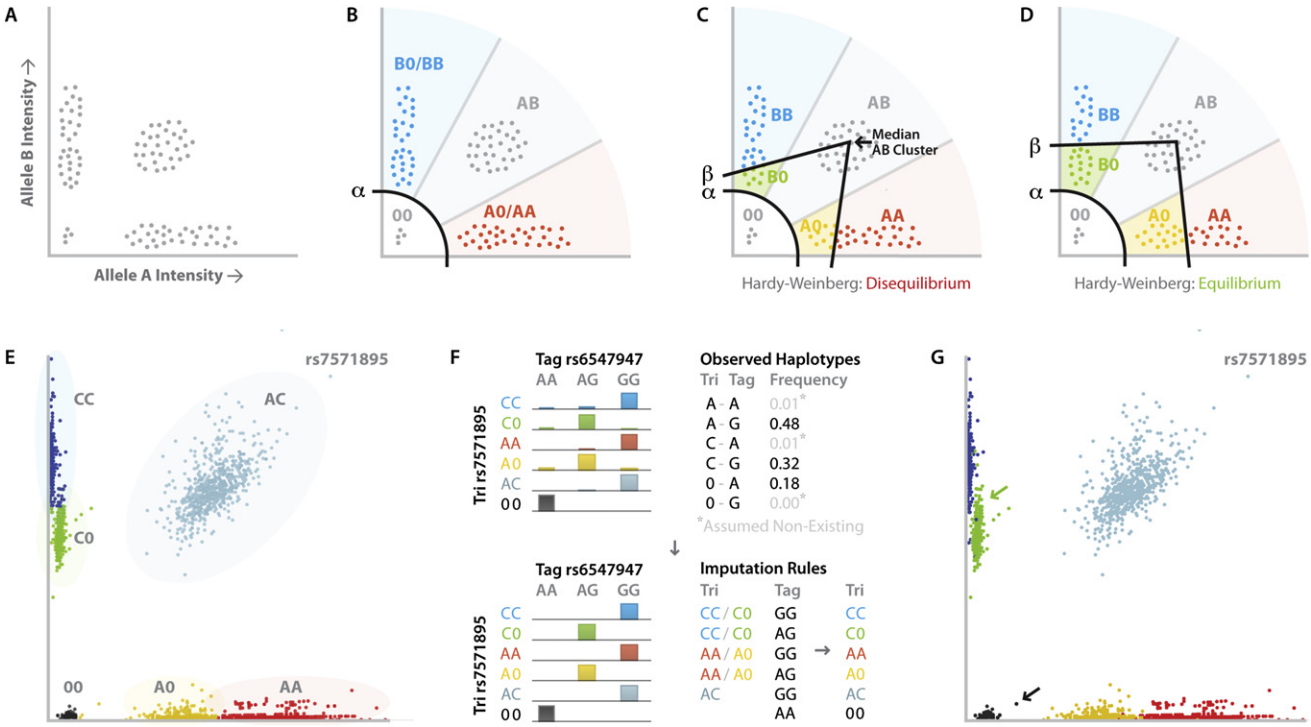

Figure 1. Genotyping Methodology for SNPs with a Third, Untyped Allele

The graphs show the intensities of the A labeled probe (x axis) and B labeled probe (y axis) of both a theoretical SNP with an third, untyped allele (top figures) and a real SNP (rs7571895, bottom figures). (A) shows that six genotypes for a triallelic SNP exist. The A0 and AA, and B0 and BB, genotype clusters usually overlap somewhat. (B) shows that initially 00 genotypes are assigned to samples that have an intensity lower than threshold α. The remaining samples are designated an initial A0/AA, AB, or B0/BB genotype with an existing calling algorithm. As shown in (C), parameter β is then used to discriminate between A0 and AA and between B0 and BB genotypes (see text). This allows for determining whether Hardy-Weinberg equilibrium is observed. As shown in (D), parameters α and β are then optimized (with a maximum-likelihood-estimation procedure) until the SNP does adhere to Hardy-Weinberg equilibrium conditions. Triallelic-genotype assignments, based on the MLE procedure for SNP rs7571895, are shown in (E). (F) shows that subsequent analysis of neighboring SNPs results in the identification of biallelic SNP rs654797, which is in strong LD with the null allele of rs7571895. Although LD does not seem to be perfect ($r^2 < 1$), we assume that this is probably because of imperfections in the initial genotype assignments and that some of the haplotypes (indicated with an asterisk) are not actually present. This allows for identifying a set of triallelic genotype imputation rules that are applied to the data and result in (G) improved genotype assignments for rs7571895, as is clearly visible when distinguishing C0 from CC samples (green arrow) and 00 from A0 samples (black arrow).

between both the A0 and AA samples and between the B0 and BB samples. We did this because different arrays had been used to genotype cases and controls. Although these arrays for most SNPs show highly comparable intensity characteristics, for some SNPs, subtle differences are present. When nearby biallelic SNPs can only help to discriminate between A0 and AA or between B0 and BB, spurious associations are to be expected because of the way our calling algorithm initially discriminates between A0 and AA and between B0 and BB genotypes. Because of the normally low frequency of the null allele, a Fisher's exact test was performed for testing the association significance. Type I errors were ascertained by a quantile-quantile (Q-Q) plot, generated by plotting the observed ordered null-allele associations against the ordered expected associations. Then we fitted a line to the lower 90% of the distribution, of which the slope ($\lambda_{inflation}$) denotes either the inflation or deflation of the test statistic.

Segregation Analysis

A segregation analysis was performed on 16 CEU trios for which biallelic-genotype data had been generated on the Illumina Infinium II Human Hap650 platform (containing 660,918 SNPs). We chose this data set because no genotypes for many of the identified triallelic SNPs were available in the Phase II release from HapMap; this was because of the fact that SNPs showing segregation inconsistencies in multiple trios were not included in this release.

Triallelic SNPs were included for analysis if genotypes could be imputed on the basis of the biallelic calls; thus without directly relying upon the raw intensity data, this method required that genotype calls for these SNPs and the biallelic SNPs used for imputation were available. Imputation allowed us to inspect visually whether the raw-intensity-data patterns corresponded well to the imputed genotype assignments. Subsequently, we used these imputed triallelic genotypes to assess how many of the Mendelian segregation inconsistencies observed under biallelic assumptions could be resolved. We took a conservative approach, because we did not score segregation inconsistencies in the analysis of the biallelic-genotype calls in trios in which a genotype had not been called for either the mother or the father.

Identity of Untyped Alleles

Various sources can result in the detected null alleles within the identified triallelic SNPs. Deletion CNVs that span these SNPs

will result in these triallelic intensity characteristics, whereas a previously unknown, third nucleotide at the physical position of the SNP gives the same results. Alternatively, it is possible that within the immediately adjacent locus that is complementary to the 50 bp primer of the SNP (used in the Illumina Infinium chemistry), there is a secondary polymorphism that affects the hybridization efficacy of the primer and that will consequently result in the same triallelic pattern.[31]

To discriminate between these three possible explanations, we investigated whether there was any evidence that these SNPs reside within deletion CNVs. If a deletion CNV is large enough to span multiple assayed SNPs, these SNPs should all show a triallelic intensity characteristic. It is likely they will all be identified by our calling method, but some might be missed (type II error). To overcome this, for each triallelic SNP we assessed whether its neighboring SNPs showed characteristics suggesting the presence of a triallelic pattern. It is expected that if this is the case, a neighboring SNP (such as the triallelic SNP) will show Euclidian intensities for the triallelic A0 and B0 samples that are significantly lower than the intensities of the samples with a triallelic AA, AB, or BB genotype.

We first corrected for differences in probe intensity characteristics within these neighboring SNPs through ranking the Euclidian intensities of the samples that had an AA genotype for the neighboring SNP and through ranking the Euclidian intensities of the samples that had a BB genotype for the neighboring SNP. We linearly scaled these two rankings to [0, 1] and assigned a value of 0.5 to samples that were heterozygous for the neighboring SNP. We then compared the ranked intensities of the samples that had been assigned triallelic 00, A0, or B0 genotypes with the ranked intensities of samples with triallelic AA or BB genotypes and required that ranked intensities of the 00, A0, and B0 samples were significantly lower (one-sided Wilcoxon-Mann-Whitney test p value $< 10^{-5}$). We then called genotypes under biallelic assumptions for the neighboring SNP. We also required that loss of heterozygosity (LOH) was observed (Fisher's exact test p value < 0.01) in the samples that had been assigned 00, A0, or B0 genotypes for the triallelic SNP. However, we only tested for this if the minor allele frequency of the neighboring SNP was high enough, such that in a theoretical situation in which no AB samples were present, the LOH Fisher's exact test p value would be below 0.001.

We first performed this analysis for the immediately adjacent SNPs and then moved farther to the left and right, continuing as long as the above conditions applied. Because the A0 and AA clusters and B0 and BB clusters usually overlap somewhat, we reasoned that if a deletion spans several SNPs, a better separation between A0 and AA samples and between B0 and BB samples would be obtained if we averaged the ranked intensities of these SNPs per sample. We applied this as an extra criterion for determining how far a deletion is likely to extend. Apart from the above criteria, we also required that, when we included more neighboring SNPs to the left and right of the triallelic SNP, the averaged ranked intensity differences between the samples with an A0 or B0 genotype and the samples with an AA and BB genotype should consistently become more significant.

These criteria meant we could determine the locus size for each fitted triallelic SNP. Immediately overlapping and adjacent loci were concatenated, resulting in loci that ranged in size between one SNP and loci that contained multiple fitted SNPs and/or neighboring SNPs that showed aberrant intensity characteristics and LOH.

To identify SNPs for which the observed triallelic intensity characteristic was due to a polymorphism in the primer region, we derived the physical genomic positions in which the 50 bp primers annealed and determined whether more polymorphisms had been described within these loci in dbSNP (build 127). All analyses were performed on the NCBI build 36 genome assembly.

All the triallelic loci identified were categorized into loci that contained multiple consecutive triallelic SNPs, loci that contained one SNP for which no polymorphisms within the primer were known, and loci that contained one single triallelic SNP and for which a primer polymorphism was known.

Resequencing

We selected 23 triallelic SNPs for resequencing. Two were selected to corroborate our prediction that the null allele for these was caused by primer polymorphisms. We selected an additional 21 triallelic SNPs to get an estimate of what proportion of the identified null alleles reflects primer polymorphisms and what proportion reflects deletions. To assess the quality of the genotype predictions, we selected triallelic SNPs with different inferred genotype qualities. We selected samples for all six genotypes when possible. Primers were designed such that we PCR amplified ~500 base pairs around the triallelic SNPs. On average, nine samples were sequenced per SNP. Sequencing was performed according to standard protocols on an ABI 3730 (Applied Biosystems) sequencer.

Genomic Properties of Triallelic Loci

Ensembl[32] version 41.36c was used for annotation purposes and mapping of gene identifiers to Ensembl gene names. The size of each identified locus was defined by taking the physical distance between the two immediate biallelic SNPs that enclosed it. The significance of underrepresentations or overrepresentations for each of the various genomic properties was empirically determined by permuting all loci across the genome 1000 times, through defining the loci randomly around SNPs that were present on the Illumina Hap550 chip, and ensuring that the size of these permuted loci was equal to the real distribution. Known deletion CNVs were derived from the Database of Genomic Variants[3] (March 2007 release, NCBI build 36 mapping). We assessed enrichment of the loci for these deletions by determining how many loci overlapped with known deletion CNVs and by fitting an extreme value distribution (EVD) on the permuted loci with the EVD add-on package[33] to R (R Development Core Team 2003, version 2.4.1). The Online Mendelian Inheritance in Man[34] morbid map (downloaded on 6 December 2006) was used for the enrichment analysis of disease genes that overlapped with our loci. Enrichment analysis of genes with known paralogs was determined empirically by dreviation of all known paralogs from Ensembl and assessment of whether the number of genes that overlapped with the identified loci with known paralogs was higher than within the permutations. Known biological interactions were derived from KEGG,[35] BioGrid,[36] Reactome,[37] BIND,[38] HPRD,[39] and IntAct[40] (all downloaded on 17 April 2007). Interaction-depletion analysis for the genes, overlapping with the identified loci, was determined by contrasting the distribution of the number of interactions ("degree") for each of these genes against the distribution of the degree of the genes that were present within the 1000 permutations, with a Wilcoxon-Mann-Whitney test.

Results

Identification of 1880 Triallelic SNPs

TriTyper initially determines which SNPs show deviation from HWE under biallelic assumptions, which provides

evidence that an extra, untyped allele might be present for these SNPs (see Figure 1A and details in Appendix A). For these SNPs, we tried to fit "triallelic" genotypes (Figure 1A, see details in Appendix A). Initially, we used parameter α to identify a putative set of samples with 00 genotypes and assigned preliminary A0/AA, AB, and B0/BB genotypes to the remaining samples (Figure 1B). We used parameter β to distinguish both between A0 and AA samples and between B0 and BB samples (Figure 1C). By adjusting α and β, and using a maximum-likelihood estimation procedure, we could then find a triallelic-genotype assignment in which HWE was observed (Figure 1D). We then looked for circumstantial evidence that this untyped allele had been correctly identified (Figure 1E) by searching nearby biallelic SNPs that are in near perfect LD with this null allele (Figure 1F). Because some of the initially assigned genotypes might be incorrect, we can use this LD to improve upon the triallelic genotyping through imputation (Figure 1G, green and black arrows) (see details in Appendix A).

By applying this algorithm to 1,417 unrelated UK controls, genotyped for 571,738 SNPs (Illumina Human Hap550 array), we identified 1,535 triallelic SNPs (median null-allele frequency = 8.6%). To be able to detect triallelic SNPs with a lower null-allele frequency, we increased the sample size to 3102, by adding 768 unrelated UK celiac patients, 445 unrelated Dutch controls, and 472 unrelated Dutch amyotrophic lateral sclerosis patients. Because these samples had been typed on the Illumina Human Hap300 array, this analysis was restricted to the 313,505 SNPs that were present on both array types. We identified 958 triallelic SNPs, of which 345 (median null-allele frequency = 4.7%) had not been identified in the smaller cohort. Cluster plots of all 1880 triallelic SNPs are available on the *TriTyper* website.

The presence of LD between these null alleles and nearby biallelic SNPs provides strong evidence that an untyped allele has been correctly identified for these triallelic SNPs. In addition, once the presence of this LD had been established, we utilized it to partly impute the triallelic genotypes. For 1204 (64%) of the 1880 triallelic SNPs, imputation is capable of discriminating both between A0 and AA and between B0 and BB samples. In these cases, biallelic-genotype calls suffice to infer these "fully imputable" triallelic genotypes. This allows for performing association analysis of triallelic SNPs in GWA studies for which only biallelic-genotype calls have been made publicly available[41,42] or when different genotyping assays have been used.

To assess how well imputation functions when only biallelic-genotype calls and no raw intensity data were available, we performed a Mendelian segregation analysis on genotype data from 16 CEU trios. For these samples, biallelic-genotype calls were available for 1153 (96%) of the 1204 fully imputable triallelic SNPs (see Material and Methods). A total of 431 (37%) SNPs showed segregation inconsistencies under biallelic assumptions. When imputing triallelic genotypes, this decreased to 319 (28%). This indicates that some segregation inconsistencies can indeed be resolved. We reasoned that if the LD was high between the null allele and the biallelic SNPs used for imputation, the genotypes should mostly be correct and would resolve most of the observed segregation inconsistencies. To assess this, we confined the analysis to those triallelic SNPs in our cohort for which the observed concordance between the preliminary triallelic genotypes determined and the subsequently imputed triallelic genotypes was at least 90%. Of these 596 triallelic SNPs, 257 (43%) showed Mendelian segregation inconsistencies when they were called under biallelic assumptions, compared to 60 (10%) when the imputed triallelic genotypes (individual segregation plots are available at the *TriTyper* website) were used. This implies that for the great majority of the identified SNPs, an extra allele has indeed been typed but that most of these triallelic genotypes can be correctly imputed when the LD is sufficiently high. Additionally, the concordance between the preliminary assigned triallelic genotype and eventually imputed genotypes serves as a quality statistic measure of the triallelic-genotype calling.

Association Analysis

Because most GWA studies aim to identify new susceptibility loci for diseases, it is essential that accurate association analysis can also be performed on the triallelic SNPs identified. We first investigated whether such an analysis has higher statistical power than an analysis of biallelic SNPs that are in LD with these triallelic SNPs, because we expected some marginal effect on these nearby biallelic SNPs to be observed as well. To assess the strength of this marginal effect, we simulated null-allele associations for 600 triallelic SNPs under three association scenarios (association $p = 10^{-4}$, $p = 10^{-6}$, and $p = 10^{-8}$, see Material and Methods). For each scenario, case and control labels for each triallelic SNP were assigned in such a way that the association p value for the null allele of this SNP approximated the p value of the scenario under investigation. Then the association strength of the SNPs used for imputation purposes could be determined (Figure 2A). The median marginal effect was 3×10^{-3}, 3×10^{-4}, and 2×10^{-5} for the three scenarios, respectively, indicating that marginal effects on the SNPs used for imputation are usually present but much weaker than for the imputed triallelic SNP. It can thus be concluded that the statistical power to detect associations for the null alleles of these triallelic SNPs is considerably higher than an analysis of the biallelic SNPs that are in LD with them.

We performed a celiac disease association analysis on the triallelic SNPs identified in the data set[28] that comprised 1417 UK controls and 768 celiac disease cases. Celiac disease is a common (1% prevalence), inflammatory condition of the small intestine induced by intake of gluten in wheat, rye, and barley. Most of the heritability is explained by the human leukocyte antigen (HLA) component,[43] because the majority of individuals with celiac disease possess HLA-DQ2 (and the remainder mostly have HLA-DQ8).[44]

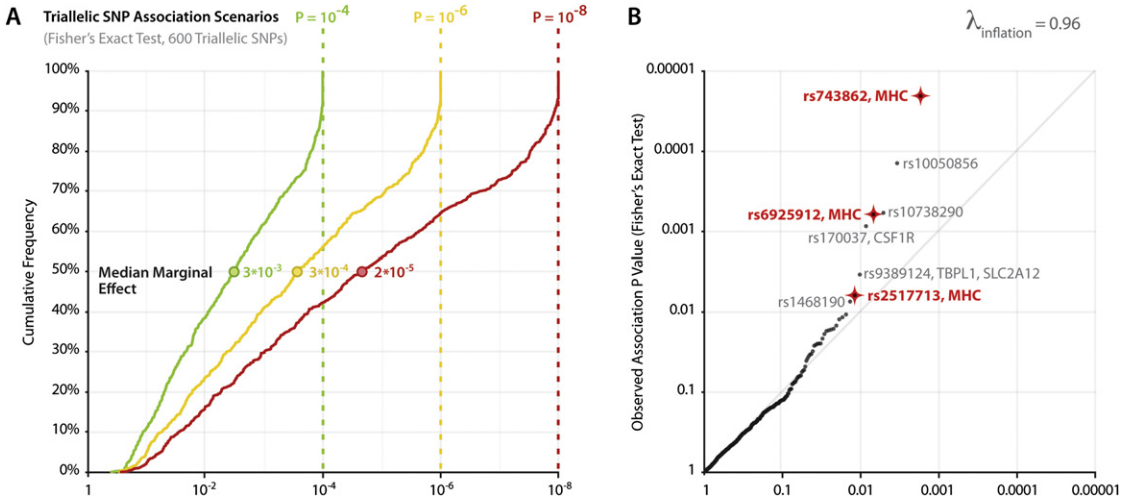

Figure 2. Association Analysis with Triallelic SNPs and Marginal Effect on SNPs, Used for Imputation

(A) Marginal association signals of SNPs, used for imputing triallelic SNPs, with disease. Fixed associations for the null allele of 600 triallelic SNPs were defined in such a way that each of the triallelic SNPs approximated a Fisher's exact test p value of 10^{-4}, 10^{-6}, or 10^{-8}. We then assessed whether a marginal association signal was present within the SNPs that had been used to impute the triallelic genotypes. The median marginal effect and the cumulative distribution of the marginal association p value for each of these SNPs are shown, ranked on significance (see text for details).

(B) Quantile-quantile plot of observed versus expected p values in a triallelic SNP null-allele association analysis in celiac disease, for which cases and controls had been typed on different platforms. Eight triallelic SNPs with a Fisher's exact test p value < 0.01 are indicated. The $\lambda_{inflation}$ factor is 0.96, suggesting no inflation of the test statistic. Three SNPs map within the major histocompatibility complex region (indicated in red).

Recently, we identified additional susceptibility loci in a GWA study,[28,45,46] in which we performed an association analysis on 585 fully imputable triallelic SNPs (see Material and Methods). The results (Figure 2B) indicate that an association analysis on these triallelic SNPs does not lead to inflated test statistics, because $\lambda_{inflation} = 0.96$ when calculated on the lower 90% of the distribution ($\lambda_{inflation} = 1.08$ when calculated with all test statistics). This suggests that our imputation methodology prevents spurious associations; such a finding is quite encouraging because the cases and controls had been typed on different arrays (Illumina Human Hap300 versus Illumina Human Hap550). Eight triallelic SNPs showed a Fisher's exact test p value below 0.01 (Table 1). When we expanded the control cohort by adding 445 Dutch controls, all eight SNPs retained a p value < 0.01. Three of these (rs743862, rs6925912, and rs2517713, marked red in Figure 2B) map within or very close to the major histocompatibility complex (MHC) that is highly polymorphic, has extended LD, and contains the strongly associated HLA-DQA1 (MIM 146880) and HLA-DQB1 (MIM 604305) genes. As such, these null alleles probably reflect nearby polymorphisms (located on a celiac-disease-associated haplotype) that affect the annealing of the triallelic SNP primers. On the basis of dbSNP (build 127), this is known to be the case for rs743862 (rs28366194 at +1bp) and rs2517713 (rs9260378 at +3 bp). Although such a secondary "primer polymorphism" is not known for rs6925912, this cannot be excluded as the MHC is highly polymorphic. For the remaining five triallelic SNPs, there is little evidence for their potential involvement in celiac disease, with the notable exception of rs170037. This SNP maps within a known susceptibility locus (CELIAC2 [MIM 609754] on 5q31-33) that has been identified in independent linkage studies[47–49] and was significantly linked in a meta-analysis of four populations.[50] It maps in an intron of the colony stimulating factor 1 receptor (CSF1R [MIM 164770]) that is involved in monocyte to macrophage differentiation and innate immunity.[51] For CSF1R, some weak association has also been reported with Crohn's disease,[52] another inflammatory gastrointestinal disorder for which molecular mechanisms, comparable to celiac disease, have been implicated.[46]

It is relevant to note that if the null allele itself is not associated with disease, but the A or B alleles are, biallelic assumptions will result in either an overestimation or underestimation of the effect, depending on whether the effect is dominant or recessive, respectively (see details and Figure 3). Although these triallelic SNPs are usually excluded from biallelic association analyses, because of observed HWE deviations, it is possible these deviations remain under the threshold used (usually in GWA studies an exact HWE p value < 0.0001 is used to exclude SNPs from subsequent association analysis[28]). This is likely to be the case if the sample size is small, indicating that when associations are observed for any identified triallelic SNP under biallelic assumptions, one should proceed with caution.

Identity of Null Alleles

The detected null alleles within the 1880 triallelic SNPs can originate from different sources. These SNPs might map

Table 1. Triallelic SNPs with Null Allele, Associated with Celiac Disease

Triallelic SNP	Chr.	Position (bp)	Overlapping Genes (nearby genes)	Null-Allele Frequency, UK Cases	Null-Allele Frequency, UK Controls	Null-Allele Frequency, Dutch Controls	Association p Value of UK Samples (Fisher's exact test)	Allele Frequency p Value of SNPs Used for Imputation on UK Samples (1 df χ^2 test)	
rs743862[a]	*6*	*32,489,917*	*(BTNL2, HLA-DRA)*	*12.5%*	*8.4%*	*6.3%*	*2.02×10^{-5}*	rs9501626, rs3817963	*$0.0132.63 \times 10^{-10}$*
rs10050856[a]	5	23,407,397	(PRDM9)	13.5%	9.6%	10.8%	1.40×10^{-4}	rs10038792, rs3924616	0.2740.0978
rs10738290[b]	9	12,730,906	(TYRP1, C9orf150)	3.5%	5.8%	5.8%	5.86×10^{-4}	rs970946, rs391858	0.8630.002
rs6925912	*6*	*26,084,906*	*TRIM38*	*12.0%*	*15.8%*	*15.8%*	*6.10×10^{-4}*	*rs199750, rs199741*	*9.09×10^{-6} 1.70×10^{-4}*
rs170037	5	149,420,837	CSF1R	4.9%	7.5%	6.1%	8.58×10^{-4}	rs216148	0.028
rs9389124	6	134,355,478	TBPL1,SLC2A12	4.4%	6.5%	6.0%	0.0034	rs6902440	0.017
rs2517713[a,b]	*6*	*30,026,078*	*HLA-A*	*2.7%*	*4.4%*	*5.8%*	*0.0061*	*rs2860580, rs2256902*	*1.11×10^{-16} 0.005*
rs1468190	16	13,265,389	(ERCC4)	17.3%	20.7%	21.0%	0.0074	rs10492781	0.004

SNPs mapping within major histocompatibility complex (MHC) is indicated in italics; $p < 0.01$.
[a] Known polymorphism present within the primer of SNP (dbSNP, build 127).
[b] Known deletion locus (Database of Genomic Variants, March 2007 release).

within deletion CNVs, and such a mapping will result in the observed triallelic intensity characteristics, but the null allele might also reflect an unknown, third nucleotide at the physical position of the SNP (e.g., an A/C SNP in fact is an A/C/G SNP). Another explanation could be that, within the immediately adjacent locus that is complementary to the 50 bp primer of the SNP, a secondary polymorphism is present that affects the hybridization efficacy of the primer and consequently results in the same triallelic pattern.[31] To gain insight into these classes, we defined nonoverlapping loci (see Figure 4 and Table 2) by concatenating immediately adjacent triallelic SNPs. A total of 208 of the SNPs that were immediately adjacent to the triallelic SNPs, but which had not been deemed triallelic, were also added because they showed aberrant intensity characteristics and loss of heterozygosity (see Material and Methods). This resulted in the identification of 1655 different loci in total.

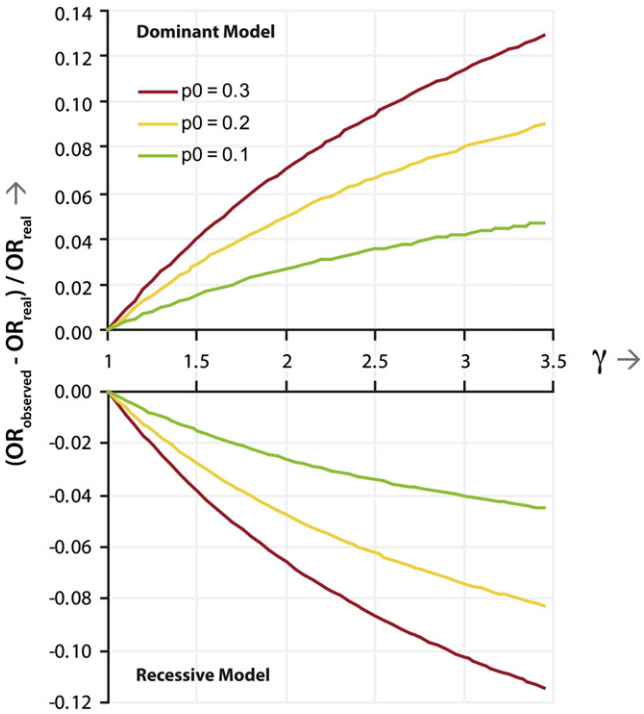

Figure 3. Consequences of Mistyping a Null Allele for Case-Control Association Studies
It is assumed allele A is the true risk allele for various values of γ (relative risk of AA homozygote) and frequencies of the null allele (p0). The overestimation of the effect under a dominant model (top figure) and the underestimation of the effect under a recessive model (bottom figure) are shown.

A total of 145 loci spanned multiple adjacent SNPs, which suggests these loci reflect deletions and this is supported by an analysis of the Database of Genomic Variants. Seventy-seven (53%) were already known to be deletions in this database, and this is much more than expected (Extreme Value Distribution p value $< 10^{-50}$).

For the remaining 1510 loci that contained only one SNP, the origin of the extra allele was less obvious: One explanation could be that polymorphisms map within the locus that is complementary to the 50 bp primer of the SNP, affecting the hybridization efficacy of the primer and resulting in this triallelic pattern. These primer polymorphisms were observed in 437 (29%) of these loci (Table 2), a finding that is considerably higher than expected because secondary polymorphisms are known within the primer region for 85,045 (16%) of the 550,123 Human Hap550 SNPs with known mapping (Fisher's exact test p value $< 10^{-18}$). Interestingly, when assessing how far these primer polymorphisms map away from the triallelic SNP, the two distributions showed a markedly different distribution (see Figure 5). Primer polymorphisms were

Figure 4. Overview of 1655 Triallelic Loci Identified on Autosomes and Chromosome X

Immediately to the left of each chromosome are depicted all the SNPs present on the Illumina Human Hap550 platform. CNVs known in the Database of Genomic Variants are shown to the right of each chromosome. Next to this, the triallelic loci are shown for which the length of each bar denotes the null-allele frequency. Blue indicates a single triallelic SNP locus, orange indicates a locus in which multiple adjacent triallelic SNPs have been identified, and gray indicates a single triallelic SNP locus for which polymorphisms are known within the region complementary to the primer of the triallelic SNP (dbSNP build 127).

usually much closer to the investigated triallelic SNP compared to the distribution of the other SNPs with known primer polymorphisms (Wilcoxon Mann-Whitney p value $< 10^{-76}$). This implies that primers on the Illumina platform usually tolerate polymorphisms well, as long as these do not map too close (>10 bp) to the SNP to be typed.

For the 1073 loci without known primer polymorphisms, we observed a strong enrichment of deletions,

Table 2. Overview of the Genomic Properties of Identified Triallelic SNPs

Initial data set	3102 samples: 2,185 UK (1417 on Hap550; 768 on Hap300), 917 Dutch samples (Hap300)		
Identified triallelic SNPs	1,880		
Identified loci	1,655 (immediate adjacent triallelic SNPs have been concatenated in a single locus)		
Locus size	**1 triallelic SNP (1510 Loci)**		**Loci with ≥ 2 adjacent triallelic SNPs (145 loci)**
	No known primer polymorphism	Known primer polymorphism	Multiple triallelic SNPs within locus
	Possible origin of null-allele: - Primer polymorphism - Deletion - Extra allele	Origin of null-allele: Primer polymorphism Typed SNP / Untyped SNP in primer (Wildtype allele) Genome TGGACTAGGTACGTGACTACACGTCAT Primer TGATCCATGCACTGATGTGCAG → Signal Alternative allele TGGACTCGGTACGTGACTACACGTCAT TGATCCATGCACTGATGTGCAG → No signal	Probable origin of null-allele: Deletion Typed SNP 1 / SNP 2 TAGTGGACTAGGTACTGACGTACACGTCATGA Deletion TAGTGGTCATGA
Number of Loci	**1073**	**437**	**145**
Overlap with known CNV deletions	136 (enrichment $p < 10^{-50}$)	50	77 (enrichment $p < 10^{-50}$)
No. of unique Ensembl genes	490	216	105 (enrichment p = 0.035)
No. of loci that contain Ensembl genes	485	207	59 (depletion p = 0.013)
Median no. of interactions of Ensembl genes	1	1	0 (depletion p = 0.004)
No. of genes with paralogs	359	140	84 (enrichment p = 0.006)
No. of OMIM disease genes	63	30	10
Enriched cytogenetic bands (p < 0.05)	2q, 3p, 6p	6p, 8p, 22q	5p, 8p

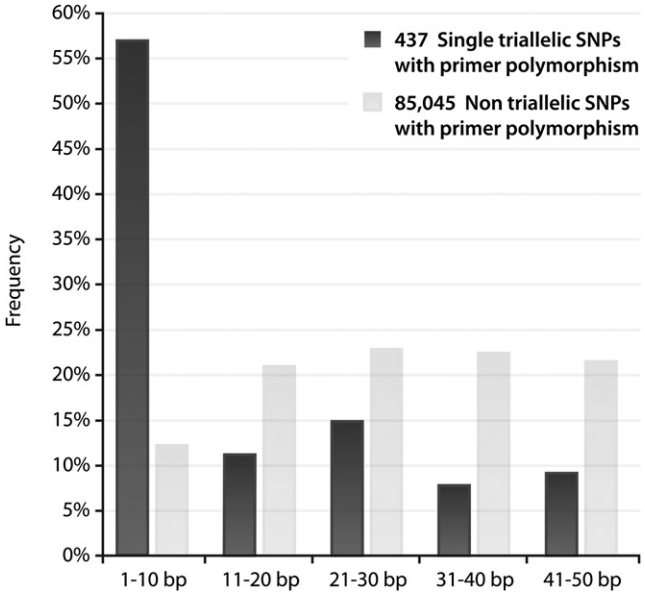

Figure 5. Distribution of Distance of Secondary Polymorphisms Present within Primers of Human Hap550 SNPs

Distribution plot of the distance of secondary polymorphisms present within primers of Human Hap550 SNPs to the actual SNP. Polymorphisms are known in dbSNP (build 127) for 85,045 of the SNPs present on the Illumina Hap550 platform within the 50 bp long primers. For the 1880 fitted triallelic SNPs, this is the case for 437 of the SNPs (expected 235, Fisher's exact p value $< 10^{-18}$). When investigating how far away these secondary polymorphisms are from the actual SNPs, it turns out that within the triallelic SNPs, these secondary polymorphisms usually map much closer to the actual SNP than for the nontriallelic SNPs (Wilcoxon-Mann-Whitney p value $< 10^{-76}$).

known in the Database of Genomic Variants, in light of the fact that 136 (13%) had been reported in this database (Extreme Value Distribution p value $< 10^{-50}$). Earlier estimates show that 50%[31]–60%[5] of these loci reflect deletions. This suggests we have detected at least 682 small-deletion CNV regions (assuming 50% of the 1073 loci reflect deletions and adding the 145 multiple SNP loci). With an observed median null-allele frequency of 7.6% for these loci, this suggests we have identified 99 deletions per individual on average. An exponential distribution fits the observed triallelic-locus-size distribution (Figure 6A, median size = 7290 bp), supporting previous observations that small CNVs strongly outnumber larger ones.[4,53] A negative binomial distribution fits the observed allele frequency distribution (Figure 6B) well.

Resequencing

We resequenced 23 triallelic SNPs to assess the predicted proportion of deletions among the identified triallelic SNPs (Table 3). For two triallelic SNPs (rs13213842 and rs7678151), we confirmed that the observed null allele was indeed due to a primer polymorphism. For the other 21 triallelic SNPs, we observed that the null allele reflects a primer polymorphism in ten SNPs. Small deletions were identified in two SNPs (rs7822381 and rs2486674). For the other nine triallelic SNPs, no primer polymorphism was identified. Additionally, for the samples for which we had predicted a homozygote deletion, no product was observed, suggesting these reflect deletions that are bigger than the loci we had amplified. These results support our estimate that ~50% of the triallelic SNPs represent deletions. We also assessed how well the predicted genotypes correspond to the resequenced genotypes. Seventeen SNPs showed perfect concordance, whereas for six SNPs, this was not the case. However, for each of these SNPs, the predicted quality of genotype inference (based on the concordance between the preliminary triallelic genotypes and imputed genotypes) was lower than 0.90, suggesting that genotypes are usually correctly inferred for 1052 (56%) of the 1,880 triallelic SNPs, because these have a concordance value over 0.90 (Table 3, indicated by the black horizontal bar).

Genomic Properties

To gain insight into the enrichment or depletion of certain genomic features within these loci, we analyzed the three triallelic-locus categories separately (Table 2, if enrichments and depletions p value was below 0.05, these are indicated). Fewer multiple-SNP loci than expected contained genes (empiric p value = 0.013), but when the loci contained genes, the number of genes was higher than expected (empiric p value = 0.035). No depletion or enrichment for these measures was observed in the two other classes of loci. It has been demonstrated that genes within CNVs have more paralogs than expected.[54] We also observed this for the multiple SNP loci (empiric p = 0.006), but not for the other two loci classes. Because genes within known deletions tend to be buffered by paralogs that usually have quite similar functions, it is likely that genes within these CNVs are biologically less important. To assess this in a different way, we investigated the number of known interactions these genes have because various studies have shown[36,55,56] that essential genes tend to have more interactions than nonessential genes. We assessed this by analyzing a collection of 80,350 known biological interactions (see Material and Methods) and indeed observed for the genes within the multiple-SNP loci that the number of interactions they have is usually significantly less than expected (Wilcoxon-Mann-Whitney p value = 0.004). In addition, various cytogenetic arms (2q, 3p, 5p, 6p, 8p, and 22q) were enriched for triallelic loci (empiric p value < 0.05).

Summary statistics for the 1880 triallelic SNPs are provided as Supplemental Data available online. *TriTyper* is freely available for downloading from the author's website, along with Java source code. It provides functionality for discovering triallelic SNPs in data sets in which raw intensity data is available. When only biallelic-genotype calls are available, *TriTyper* allows for imputing triallelic genotypes for 1204 triallelic SNPs of the 1880 SNPs we have identified in this study. After assigning triallelic genotypes, *TriTyper* can perform association analysis.

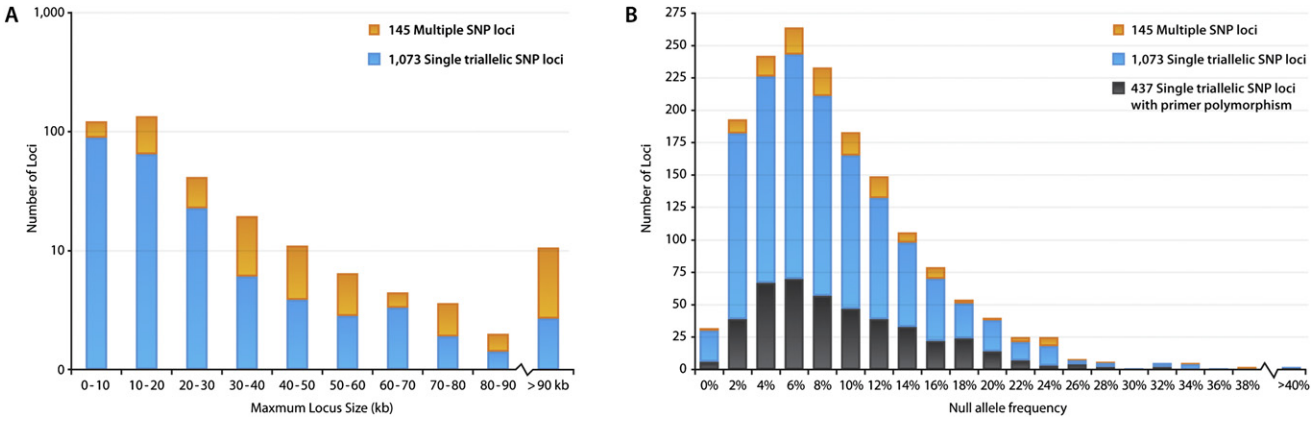

Figure 6. Distribution of Triallelic-Locus Size and Null-Allele Frequency

(A) The triallelic loci for which no polymorphism within the primer is known in dbSNP (build 127) are plotted against the maximum potential size of each locus, assuming these can reflect deletions (by taking the physical position of the immediately adjacent SNPs that look normal on the Illumina Human Hap550 platform).

(B) The number of triallelic loci is plotted against the null-allele frequency.

Discussion

In this paper, we have described a method (*TriTyper*) that uses raw intensity data from the Illumina genotyping platform to identify SNPs with an extra untyped, but common allele. Our method is the first to our knowledge to do this in case-control data sets by utilizing the presence of local LD to improve genotype assignments. Through this approach we identified 1880 triallelic SNPs, and for 1204 of these, the LD patterns permitted inferring the triallelic genotypes without needing access to raw intensity data. This enables association analyses on these SNPs in white European data sets that have similar LD patterns, but for which only genotype calls have been made available, or those that have been generated with completely different platforms.

With the triallelic-genotype calls from *TriTyper*, highly robust association analyses can be performed. We have shown this in a triallelic null-allele association analysis in celiac disease, for which cases had been run on a different type of array than that used for the controls, and we saw no inflation of the test statistic. Simulations indicate that our method has superior power to detect these associations, compared to an association analysis on the biallelic SNPs that are in LD and have been used to infer the triallelic genotypes. The triallelic SNPs identified also have ramifications for association analyses that are based on biallelic assumptions. If, for any of the triallelic SNPs, the null allele is not associated but the A and B alleles are, the real effect of the association will be overestimated or underestimated, depending on a dominant or recessive model, respectively.

The reported associations in celiac disease did not survive multiple testing when we assumed hundreds of thousands of biallelic association tests have already been performed in a GWA analysis. These findings, however, do provide new hypotheses for further replication in independent cohorts.

The identity of each of the triallelic SNPs identified remains to be established. We observed that 437 triallelic SNPs showed a triallelic pattern because of a polymorphism in the region of the primer, usually within 10 bp from the target SNP (see Figure 5). This artifact should serve as a warning for all oligonucleotide-based assays, and we urge researchers to validate putative CNVs with different techniques. For the remaining 1218 unique loci (in which immediately adjacent triallelic SNPs had been concatenated), we observed a strong enrichment for deletions, known in the Database of Genomic Variants. We estimate that, of these loci, 682 reflect deletions, suggesting that on average 99 deletion CNVs per individual were identified. This is approximately four times more than what has been found by other methods using identical oligonucleotide arrays (between 10 and 27 CNVs on average per individual[1,14,22]). The high resolution of our method and the fact that we take LD into account probably explain this difference.

Loci that contained multiple SNPs overlapped with fewer genes than expected, although the total number of genes for these loci was higher than expected. Comparable analyses[1,54] conflict with each other and as such warrants further clarification. As shown before,[54] genes within these loci have paralogs more often than expected (p value = 0.006). We are the first to our knowledge to show that the genes within these loci also biologically interact with significantly fewer genes than expected (p value = 0.004).

Various avenues for extending *TriTyper* can be envisaged. A drawback of our current imputation methodology is that we assume certain haplotypes have a zero frequency, which might not reflect the reality because of lower LD than assumed. Therefore, for some of the triallelic SNPs, it is likely that some of the imputed genotypes will be incorrect. Consequently, an association analysis using imputed triallelic genotypes will have lower statistical power compared to an ideal situation, in which accurate triallelic genotypes would be available. We argue this sacrifice in calling accuracy and power because of imputation is

Table 3. Resequencing Results of Triallelic SNPs

SNP	Known Primer Polymorphism	Genotyped Samples (predicted inferred genotypes)	Predicted Genotype Quality	Observed Origin of Null Allele	Resequenced Discordant Genotypes (Explanation)
rs10504729	-	6 (1 A0, 1 AA, 1 AG, 2 G0, 1 GG)	0.66	Primer polymorphism (C/T, −1 bp)	0
rs2675899[a]	-	10 (2 A0, 2 AA, 2 CA, 2 C0, 2 CC)	0.71	Probably deletion	0
rs13213842	rs35678510, A/G, +1 bp	8 (1 A0, 1 AA, 2 AG, 2 G0, 2 GG)	0.71	Primer polymorphism (A/G, +1 bp)	1 (G0→GG)
rs3131755	-	5 (1 A0, 1 AA, 1 B, 2 BB)	0.75	Primer polymorphism (T/G, +4 bp)	0
rs195738	-	12 (1 00, 4 A0, 1 AA, 4 G0, 1 GG)	0.79	Probably deletion	2 (A0→AA)
rs8053391[a]	-	4 (2 A0, 1 AA, 1 GG)	0.81	Primer polymorphism (C/G, +4 bp)	0
rs7678151	rs28542567, A/G, −3 bp	11 (1 00, 2 A0, 2 AA, 2 AG, 2 G0, 2 GG)	0.83	Primer polymorphism (A/G, −3 bp)	1 (G0→GG)
rs2871198[a]	-	10 (4 A0, 1 AA, 1 AG, 3 G0, 1 GG)	0.86	Probably deletion	0
rs9355606	-	6 (1 A0, 1 AG, 2 G0, 2 GG)	0.86	Probably deletion	0
rs495991	-	4 (2 G0, 2 AG)	0.86	Primer polymorphism (C/G, −1 bp)	1 (G0→GG)
rs10510312	-	6 (3 G0, 1 GG, 2 AG)	0.87	Primer polymorphism (A/C, −1 bp)	1 (G0→GG)
rs2486674	-	18 (9 A0, 1 AA, 1 AG, 6 G0, 1 GG)	0.88	Deletion (TGAGTATAGTAdel→AGTTTins/+)	5 (3 A0→AA, 2 G0→GG)
rs11834116	-	4 (1 AA, 1 A0, 1 AG, 1 G0)	0.91	Primer polymorphisms (C/T, −8 bp, A/G, + 1bp)	0
rs7083969	-	9 (1 A0, 1 AA, 1 AG, 5 G0, 1 GG)	0.92	Probably deletion	0
rs7083969	-	11 (6 A0, 1 AA, 1 AG, 3 G0, 1 GG)	0.92	Probably deletion	0
rs1109374	-	4 (1 A0, 1 AA, 1 AG, 1 GG)	0.92	Primer polymorphism (C/T, +3 bp)	0
rs9361448	-	14 (1 00, 5 A0, 1 AA, 1 AC, 5 C0, 1 CC)	0.94	Probably deletion	0
rs11533655[a]	-	15 (2 00, 5 A0, 1 AA, 1 AG, 5 G0, 1 GG)	0.95	Probably deletion	0
rs2254039	-	6 (3 G0, 2 GG, 1 AG)	0.95	Primer polymorphism (C/T, −1 bp)	0
rs2894386[a]	-	17 (2 A0, 8 AA)	0.95	Primer polymorphism (C/T, −4 bp)	0
rs7133541	-	10 (5 A0, 1 AA, 1 AG, 2 G0, 1 GG)	0.96	Probably deletion	0
rs7822381	-	18 (4 A0, 7 AA, 5 AG, 2 G0)	0.97	Deletion (1 bp deletion in primer)	0
rs1551821	-	6 (3 A0, 2 AC, 1 AA)	0.98	Primer polymorphism (A/C, +1 bp)	0

In total, two known primer polymorphisms, ten previously unknown primer polymorphisms, and 11 probable deletions were found in the observed origin of null alleles.

[a] Known deletion locus (Database of Genomic Variants, March 2007 release).

acceptable, because it considerably reduces type I errors in association testing. If different platforms or batches have been used for genotyping and cases and controls are not evenly spread[28] over these, spurious associations are to be expected because of the way our calling algorithm initially discriminates between A0 and AA and between B0 and BB genotypes. If these genotypes can be imputed with nearby biallelic SNPs, false-positive associations will be prevented. Although highly sophisticated imputation algorithms have been described for biallelic SNPs,[26,57] it is not straightforward to use these to resolve this issue. This is mostly due to the fact that we currently cannot rely upon phased haplotypes from HapMap, because all the SNPs within HapMap have been called under biallelic assumptions. Another complication is the difficulty to estimate r^2 and to interpret D′ if the number of alleles between two markers differ.[58,59] However, we expect that by incorporating some of the concepts underlying these biallelic imputation methodologies, the accuracy of the imputed triallelic genotypes can be improved.

Currently, *TriTyper* can only detect SNPs with a common extra but untyped allele. We envisage that adaptations to both our calling algorithm and LD-based genotype imputation methodology will probably allow identification of very small but common duplications. In addition, studies that aim to identify rare de novo deletions and duplications can immediately benefit from our work. Because the number of samples we have studied is reasonably high (3102), we were able to identify common triallelic SNPs that had a null-allele frequency as low as 0.5%. If researchers are not aware of these common triallelic SNPs and use smaller cohorts, they might deem these SNPs rare and potentially biologically interesting when aberrant characteristics are observed in only a few samples. Methodologically, the resolution of de novo CNV detection methods[14,22] can also be improved by incorporating LD-based frameworks: Conceptually, if two SNPs are in very strong LD, but in one sample a recombination seems to be present, a de novo duplication or deletion that spans one of these SNPs could be an alternative explanation.

The Illumina BeadChip arrays we have used here are strongly biased against CNVs, because SNPs that showed low call rates, HWE deviations, or many Mendelian segregation inconsistencies in a subset of the HapMap samples had been removed during the design of these chips. This also explains why the observed median null-allele frequency of the identified triallelic SNPs was only 7.6%. Because we did not use the most current Illumina

chips, we expect the newer ones that are better tailored to target CNVs (e.g., Illumina HumanHap370 and HumanHap1M), to lead to greater insight into CNVs.

The Human Gene Mutation Database[60] reports 73,411 variants that mostly have a phenotypic effect, of which ~16% are microdeletions and 7% are microinsertions (smaller than 20 bp), whereas larger deletions and insertions constitute 6% and 1% of the variants, respectively. This clearly indicates the importance of structural variants and deletions in both rare and common diseases.[6–8] New statistical CNV detection methods (such as *TriTyper*) and more extensive oligonucleotide arrays will undoubtedly result in the identification of many more variants, of which quite a few will turn out to be associated with disease.

Appendix A. Genotype Calling

Conventional Biallelic-Genotype Calling

When the minor allele frequency (MAF) is sufficiently high, assigning genotypes to biallelic SNPs is usually fairly straightforward: Three separate clusters will appear (reflecting the AA, AB, and BB genotypes) that can usually be well separated with a clustering algorithm we recently described.[28] This algorithm uses per-sample polar angle θ [$\theta = 2/\pi$ * arctan (intensity$_b$/intensity$_a$)] to identify three clusters of sample for which the standard deviations of the θ values for each cluster are low. This is achieved by exploring a 2D search space (in which one parameter discriminates between AA and AB samples and the other discriminates between AB and BB samples). The method then settles upon a certain clustering for which the three calculated standard deviations have a sum that has been minimized.

Preliminary Triallelic-Genotype Calling

When a SNP is triallelic, but the SNP has been called under biallelic assumptions for sufficient samples, it is likely that HWE deviations will be observed. Assuming HWE for the true alleles A, B, and 0, we can compute the expected frequencies of observed genotypes AA, AB, and BB. From these we can compute the observed allele frequencies for A and B. Now the deviation from the Hardy Weinberg equilibrium in those observed genotypes AA, AB, and BB relative to the genotype frequencies expected from the observed allele frequencies A and B can be computed. It turns out that the resulting χ^2 depends on the true frequency of the 0 allele, and of course on the sample size, but not on the frequencies of the A and B alleles:

$$\chi^2 = n \cdot {p_0}^2 \cdot (4 - 8p_0 + 5{p_0}^2),$$

where n is sample size and p_0 the frequency of the 0 allele.

Calculations show that if 3000 samples are typed, a null allele with a frequency of 2% or higher will on average cause a HWE deviation that can be demonstrated at the level of $p = 0.05$. Figure 7 illustrates how the HWE test statistic depends on the sample size and the frequency of the 0 allele.

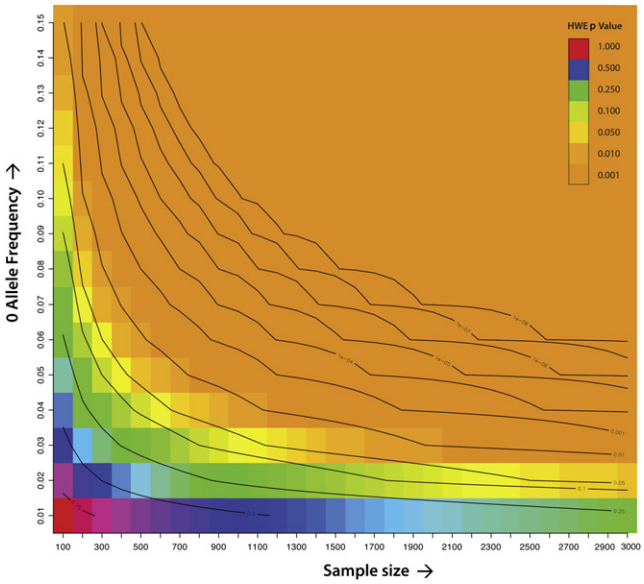

Figure 7. HWE Test Statistics, when Analyzing Triallelic SNPs, Called under Biallelic Assumptions

These calculations show the HWE test statistic for various sample sizes and different frequencies of the 0 allele. If we incorrectly assume triallelic SNPs are biallelic, analysis of sample sizes that are representative of current genome-wide association studies will result in significant HWE deviations, even when the 0 allele has a fairly low frequency, e.g., when testing 3000 samples, and assuming a call rate of 100% for samples having one or two copies of the A or B allele, triallelic SNPs with a null-allele frequency above 2% have an expected deviation from HWE with $p < 0.05$.

Although these HWE deviations can also arise because of failed assays, they are explained by an unlabelled allele in a substantial number of cases.[31] We followed up SNPs when, under biallelic assumptions, the exact HWE p value was below 0.05 or when the call rate was below 98%. For these SNPs, we determined whether triallelic genotypes could be called by introducing two additional parameters (α and β) to our calling algorithm.

In the initial triallelic genotype-calling procedure, genotypes 00 are assigned to samples that have a Euclidian intensity below α. For the remaining samples, we use the aforementioned calling algorithm to identify three clusters of samples that are either A0 or AA (A0/AA), are AB, or either are B0 or BB (B0/BB) (Figure 1B).

Subsequently we partition both the A0 and AA samples and the B0 and BB samples using parameter β. Nonpseudoautosomal chromosome X SNPs provide detailed insight into the intensity characteristics of these A0, B0, AA, and BB samples. For these SNPs, females will usually have two copies, whereas males will only have one copy (Figure 8A). We investigated 11,652 nonpseudoautosomal chromosome X SNPs, present on the Illumina Human Hap550 platform, for which 1417 unrelated UK samples from the 1958 British birth cohort had been typed.[28] For each of these SNPs, we linearly scaled the probe intensities, such that the center

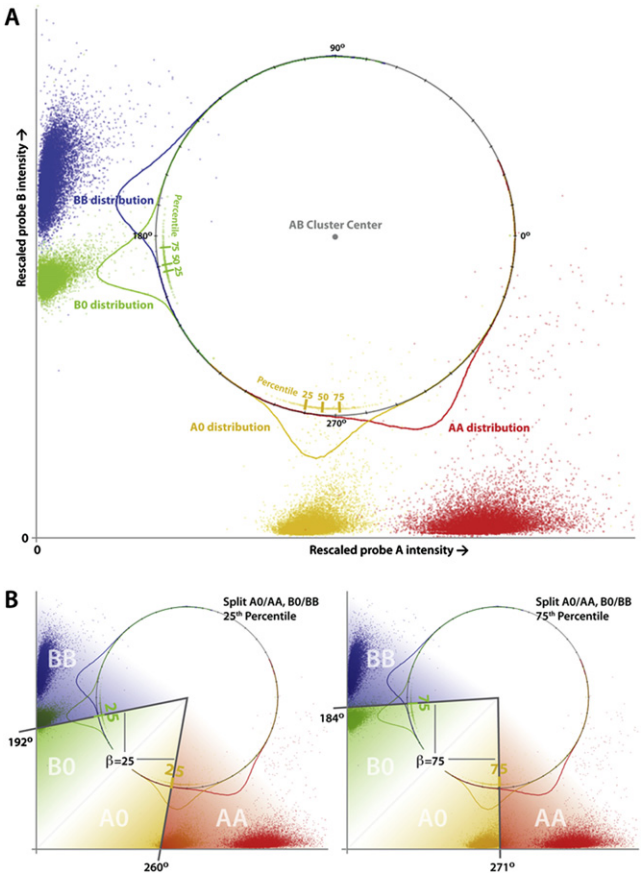

Figure 8. Distribution of *A* and *B* Allele Intensities of 11,652 Chromosome X SNPs, Present on the Illumina Human HapMap550 Platform

Each dot represents the median coordinate of the A0 (males, green), AA (females, blue), B0 (males, yellow), BB (females, red), or AB (females, gray) cluster for a single SNP. The A and B intensities have been scaled in such a way that for each SNP, the median AB cluster center is identical for all chromosome X SNPs. As shown in (A), it is evident that single-copy genotypes (A0 and B0) clearly show different intensity characteristics than AA and BB genotypes. Another observation is that the A and B probes have slightly different characteristics, because the A0 and AA distributions overlap slightly less than the B0 and BB distributions, indicating that on average, A0 and AA samples can be better distinguished from each other. To correct for these differences in intensity characteristics, parameter β is calibrated on these chromosome X SNP distributions. As shown in (B), the genotype-calling algorithm uses parameter β to distinguish between A0 and AA and between B0 and BB. For a given β, an angle for the A0 distribution is determined where the A0 distribution percentile equals β. The same holds for the angle of the B0 distribution. In the present example, increasing β increases the angle of the *A* line slightly more than it increases the angle of the *B* line. Examples are shown where β is 25 ([B], left) and where β is 75 ([B], right), resulting in different genotype assignments (A0, AA, B0, and BB genotypes assignments are indicated in yellow, red, green, and blue, respectively).

of the AB cluster was at coordinate (1, 1). We then moved the origin of the Cartesian coordinate system to this coordinate and converted to a polar coordinate system, allowing us to determine a 1D angle distribution for the A0, the AA, and the B0 and BB samples. These distributions allow us to introduce parameter β (range [0, 100]), which denotes both the percentile of the A0 and the percentile of the B0 distributions. We use this parameter to distinguish between one and two copies (Figure 1C) because the corresponding percentile corresponds to two different Cartesian rays that both start from the AB cluster center but have different angles, for which one ray (reflecting the percentile within the chromosome X A0 distribution) allows us to divide the A0/AA samples in A0 and AA samples and another ray (reflecting the percentile within the chromosome X B0 distribution) allows us to divide the B0/BB samples in B0 and BB samples (Figure 8B). For example, when $\beta = 25$ (Figure 8B, left), for the samples which are either AA or A0, the samples having an angle to the AB cluster location below 260° will be designated A0 and having an angle above 260° will be designated AA. For samples that are either BB or B0, those having an angle to the AB cluster location below 192° will be designated BB and those having an angle above 192° will be designated B0. When $\beta = 75$ (Figure 8B, right), the thresholds for these angles are 271° and 184°, respectively.

It is evident that different α and β values will result in different triallelic-genotype assignments. To optimize these, we use an MLE procedure that assumes HWE under a triallelic model, through the following log likelihood formula:[29]

$$\begin{aligned}\log(likelihood) = {} & \log[(n_{aa}+n_{bb}+n_{ab}+n_{a0}+n_{b0}+n_{00})!] \\ & -[\log(n_{aa}!)+\log(n_{bb}!)+\log(n_{ab}!)+\log(n_{a0}!) \\ & +\log(n_{b0}!)+\log(n_{00}!)]+n_{aa}*\log(p_a*p_a) \\ & +n_{ab}*\log(2*p_a*p_b)+n_{bb}*\log(p_b*p_b) \\ & +n_{a0}*\log(2*p_a*p_0)+n_{b0}*\log(2*p_b*p_0) \\ & +n_{00}*\log(p_0*p_0)\end{aligned}$$

where n_{aa}, n_{bb}, n_{ab}, n_{a0}, n_{b0}, and n_{00} are the number of individuals with assigned genotype AA, BB, AB, A0, B0, and 00, respectively, and p_a, p_b, and p_0 are the allele frequencies of allele A, B, and 0, respectively.

Through analysis of the entire search space, the values for α and β for which this likelihood is maximal can be determined (Figure 1D), indicating that the assigned genotype distribution most closely resembles the distribution expected under triallelic HWE. Identified triallelic SNPs are included for follow-up analysis, if the null-allele frequency is over 0.5% and the fitted β parameter value is between 6 and 97.

Eventual Triallelic-Genotype Calling through Imputation

To improve upon the initially assigned triallelic genotypes, we take advantage of local linkage disequilibrium, because the presence of LD between biallelic SNPs can often be utilized to improve genotype assignments.[26,27,57] Because LD has been described for deletion CNVs as well,[1,2,24,25,61] we

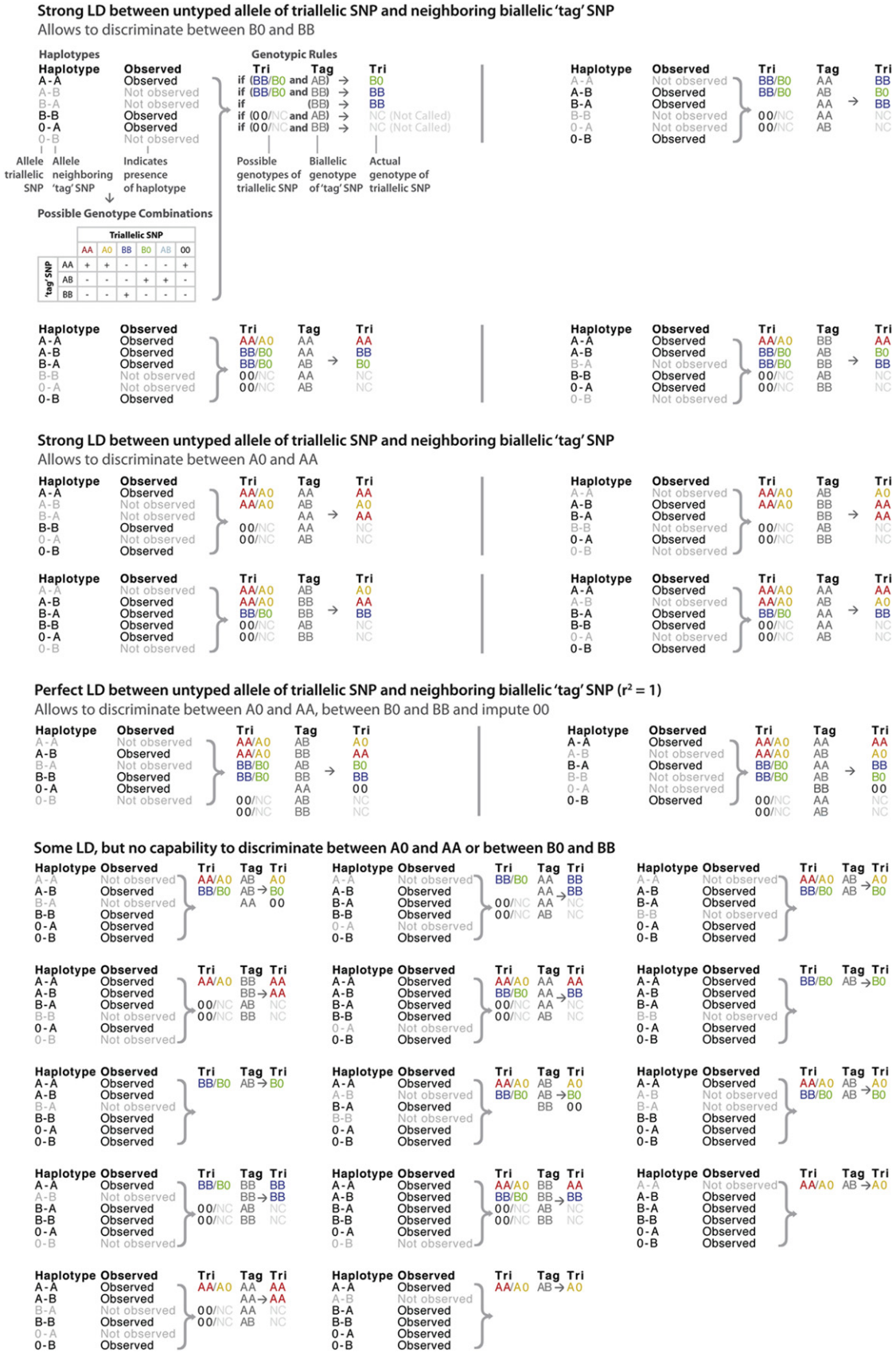

Figure 9. Imputation Scenarios

When assuming all alleles have a nonzero frequency for both the triallelic SNP and the neighboring biallelic SNP and that some LD is present (i.e., at least one haplotype has not been observed), there are 24 different imputation scenarios possible. For ten of these

assumed these triallelic genotypes can potentially also be inferred through LD.

To assess this, we require that at least one of the six haplotypes should have a zero frequency and that all alleles are present for the biallelic SNP and triallelic SNP, resulting in the identification of 24 "haplotype scenarios" that each have a different set of haplotypes that have not been observed (Figure 9). For each of these scenarios, a set of triallelic-genotype imputation rules can be easily deduced. It turns out that ten scenarios are capable of discriminating between A0 and AA and/or between B0 and BB triallelic genotypes. This is very helpful because in the initial genotype-assignment procedure, a somewhat rough division is made between the A0 and AA genotypes and between the B0 and BB genotypes (through optimization of parameter β). As such, it is likely that some incorrect genotypes (Figure 1E) have initially been assigned to samples that cluster in the vicinity of the two dividing rays determined by parameter β (e.g., the initially assigned A0 genotype should actually be AA and vice versa). This is resolved if nearby biallelic SNPs allow for discrimination between A0 and AA and between B0 and BB samples. We concentrate on any of these ten scenarios throughout this paper and will assess these for each triallelic SNP.

We first assess the LD for each triallelic SNP identified with the immediately adjacent biallelic SNPs (10 to the left and 10 to the right): For each pair, haplotype frequencies (h_{aa}, h_{ab}, h_{ba}, h_{bb}, h_{0a}, and h_{0b}) are estimated with an expectation-maximization algorithm.[62] If the frequencies of some of these haplotypes are zero (e.g., haplotypes h_{aa}, h_{ba}, and h_{0b} have a zero frequency, as in Figure 1F), it is determined whether this configuration of observed and nonobserved haplotypes matches one of the ten haplotype scenarios for which the biallelic SNP helps to discriminate between some of the triallelic genotypes, and we use the neighboring SNP for imputation. Because of the uncertainties mentioned for the initially assigned triallelic genotypes, certain estimated haplotypes frequencies will be incorrect, resulting in haplotypes with nonzero frequencies that in reality should have a zero frequency (Figure 1F). In order to overcome this, we relaxed our method for assessing the imputation potential of each neighboring biallelic SNP: We assumed that haplotypes with low, but nonzero frequencies in reality might have a zero frequency. For each haplotype, it was determined whether the frequency was lower than the frequency of the haplotype with the same triallelic allele, but with a different biallelic allele. If this was the case, we assumed that this haplotype in reality might have a zero frequency. To ascertain this, we tested all possible haplotype scenarios (through systematic inclusion and exclusion of these potentially zero-frequency haplotypes) and assessed whether any of these scenarios could help to discriminate between A0 and AA or between B0 and BB. If this was observed, we searched for evidence that our zero-frequency assumption for these haplotypes was indeed correct, by imputing the A0 and AA or B0 and BB genotypes and testing whether the Euclidian intensities of the imputed A0 or B0 samples were significantly lower (Wilcoxon-Mann-Whitney test $p < 10^{-3}$) than the Euclidian intensities of the AA or BB samples. In addition, we tested whether the concordance between the imputed and observed genotypes was higher than 60%. If this was observed, we assumed this haplotype scenario could be used for imputation purposes and stored it in a vector. Once all haplotype scenarios had been assessed for each of the 20 biallelic neighboring SNPs, we selected the imputation scenario that had the highest genotypic concordance and that could help to discriminate between A0 and AA and the imputation scenario with the highest genotypic concordance that could help to discriminate between B0 and BB. This sometimes resulted in the identification of one single biallelic SNP, in perfect LD with the untyped allele of the triallelic SNP that could be used to discriminate both between A0 and AA and between B0 and BB genotypes.

Appendix B. Consequences of Miscalling Null Alleles in Case-Control Studies

If the presence of a null allele is not recognized, this will have consequences for case-control association studies. The easiest case is when the null allele is itself the risk allele. If it is not recognized as such, the SNP will give no signal at all when assuming the A0 and B0 genotypes confer the same risk. However, it is likely that these SNPs will be removed from the analysis because HWE deviations are expected to appear and lower call rates will become apparent.

It is more complicated for cases in which allele A is the risk allele. Taking the above scenario, we can calculate the odds ratio (OR) of allele A versus nonallele A for the situations in which the null allele is recognized and not recognized. For simplicity, we will limit ourselves to a dominant and a recessive model. In the dominant model, for the observed OR (allele A versus nonallele A) in which the null allele is not recognized, we get:

$$OR_A(obs) = \frac{\gamma\left[(\alpha-1)(p_B+2p_0)+(\alpha\gamma-1)p_A\right]}{(\alpha\gamma-1)(2p_0+\gamma p_A+p_B)}.$$

Also, if the null allele is typed correctly:

$$OR_A(real) = \frac{\gamma\left[(\alpha-1)(p_B+p_0)+(\alpha\gamma-1)p_A\right]}{(\alpha\gamma-1)(p_0+\gamma p_A+p_B)},$$

scenarios, the biallelic SNP can help to discriminate between B0 and BB and/or between A0 and AA for the triallelic SNP. For the first imputation scenario, a detailed description of this procedure is provided: With this set of observed and unobserved haplotypes, a limited number of genotype combinations exist. This allows for deducing a set of genotypic rules that can help to discriminate between B0 and BB genotypes for the triallelic SNP, on the basis of the genotype of the neighboring biallelic SNP.

where p_A, p_B, p_0 are the allele frequencies of the respective alleles, α is the disease risk for genotypes not containing A, and $\alpha\gamma$ is the disease risk for individuals carrying one or two A alleles. Note the difference of $2p_0$ and p_0 in both denominator and numerator between the two equations.

For the recessive model, in which penetrance for AA homozygotes is still $\alpha\gamma$ and penetrance for all other genotypes is α:

$$OR_A(obs) = \frac{(\alpha-1)(p_B + 2p_0 + \gamma p_A)}{(\alpha-1)(2p_0 + p_B) + (\alpha\gamma - 1)p_A}.$$

Also, if the null-allele is typed correctly:

$$OR_A(real) = \frac{(\alpha-1)(p_B + p_0 + \gamma p_A)}{(\alpha-1)(p_0 + p_B) + (\alpha\gamma - 1)p_A}.$$

Figure 3 depicts the consequences of mistyping on the observed OR: OR is overestimated for the dominant model and underestimated for the recessive model. The amount of overestimation or underestimation depends on the relative penetrance (γ) of the risk allele and the null-allele frequency.

Supplemental Data

One spreadsheet is available at http://www.ajhg.org/.

Acknowledgments

We thank Jackie Senior, Madelien van de Beek, Ritsert Jansen, and members of the Complex Genetics Section, UMC Utrecht for critically reading the manuscript. We thank D. Simpkin, T. Dibling and C. Hand for genotyping (Sanger Institute) and D. Strachan and W.L. McArdle for 1958 birth cohort samples. We thank Illumina for providing HapMap genotype data. We thank Dutch and UK clinicians who collected samples[28,30] and sample donors. We thank the Genomics Center Utrecht for computational resources. Statistical analyses were carried out on the Genetic Cluster Computer in Amsterdam, which is financially supported by the Netherlands Organization for Scientific Organization (NWO, grant 480-05-003). We acknowledge funding from Coeliac UK; the Netherlands Organization for Scientific Research (NWO, grant 918-66-620); Netherlands Organization for Health Research and Development (ZonMW grant 917-66-315); the Coeliac Disease Consortium (an innovative cluster approved by the Netherlands Genomics Initiative and partly funded by the Dutch government [grant BSIK03009]); the Netherlands Genomics Initiative (grant 050-72-425 and fellowship grant to L.F.); Prinses Beatrix Fonds (L.H.v.d.B.); and the Wellcome Trust (GR068094MA Clinician Scientist Fellowship to D.A.v.H. and support for the work of P.D.). The authors acknowledge use of genotypes from the British 1958 birth cohort collection, funded by the UK Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02.

Received: January 27, 2008
Revised: March 21, 2008
Accepted: May 13, 2008
Published online: June 5, 2008

Web Resources

The URL for data presented here are as follows:

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim
TriTyper, http://www.ludesign.nl/trityper

References

1. Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. (2006). Global variation in copy number in the human genome. Nature *444*, 444–454.
2. McCarroll, S.A., Hadnott, T.N., Perry, G.H., Sabeti, P.C., Zody, M.C., Barrett, J.C., Dallaire, S., Gabriel, S.B., Lee, C., Daly, M.J., et al. (2006). Common deletion polymorphisms in the human genome. Nat. Genet. *38*, 86–92.
3. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W., and Lee, C. (2004). Detection of large-scale variation in the human genome. Nat. Genet. *36*, 949–951.
4. Conrad, D.F., Andrews, T.D., Carter, N.P., Hurles, M.E., and Pritchard, J.K. (2006). A high-resolution survey of deletion polymorphism in the human genome. Nat. Genet. *38*, 75–81.
5. de Smith, A.J., Tsalenko, A., Sampas, N., Scheffer, A., Yamada, N.A., Tsang, P., Ben-Dor, A., Yakhini, Z., Ellis, R.J., Bruhn, L., et al. (2007). Array CGH analysis of copy number variation identifies 1284 new genes variant in healthy white males: Implications for association studies of complex diseases. Hum. Mol. Genet. *16*, 2783–2794.
6. Aitman, T.J., Dong, R., Vyse, T.J., Norsworthy, P.J., Johnson, M.D., Smith, J., Mangion, J., Roberton-Lowe, C., Marshall, A.J., Petretto, E., et al. (2006). Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature *439*, 851–855.
7. Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R.J., Freedman, B.I., Quinones, M.P., Bamshad, M.J., et al. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science *307*, 1434–1440.
8. Fellermann, K., Stange, D.E., Schaeffeler, E., Schmalzl, H., Wehkamp, J., Bevins, C.L., Reinisch, W., Teml, A., Schwab, M., Lichter, P., et al. (2006). A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. Am. J. Hum. Genet. *79*, 439–448.
9. Wong, K.K., deLeeuw, R.J., Dosanjh, N.S., Kimm, L.R., Cheng, Z., Horsman, D.E., MacAulay, C., Ng, R.T., Brown, C.J., Eichler, E.E., et al. (2007). A comprehensive analysis of common copy-number variations in the human genome. Am. J. Hum. Genet. *80*, 91–104.
10. Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U., Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R., et al. (2005). Segmental duplications and copy-number variation in the human genome. Am. J. Hum. Genet. *77*, 78–88.
11. Simon-Sanchez, J., Scholz, S., Fung, H.C., Matarin, M., Hernandez, D., Gibbs, J.R., Britton, A., de Vrieze, F.W., Peckham, E., Gwinn-Hardy, K., et al. (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.

12. Pinto, D., Marshall, C., Feuk, L., and Scherer, S.W. (2007). Copy-number variation in control population cohorts. Hum. Mol. Genet. *16 Spec No. 2*, R168–R173.
13. Komura, D., Shen, F., Ishikawa, S., Fitch, K.R., Chen, W., Zhang, J., Liu, G., Ihara, S., Nakamura, H., Hurles, M.E., et al. (2006). Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. Genome Res. *16*, 1575–1584.
14. Colella, S., Yau, C., Taylor, J.M., Mirza, G., Butler, H., Clouston, P., Bassett, A.S., Seller, A., Holmes, C.C., and Ragoussis, J. (2007). QuantiSNP: An objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res. *35*, 2013–2025.
15. Kohler, J.R., and Cutler, D.J. (2007). Simultaneous discovery and testing of deletions for disease association in SNP genotyping studies. Am. J. Hum. Genet. *81*, 684–699.
16. Kosta, K., Sabroe, I., Goke, J., Nibbs, R.J., Tsanakas, J., Whyte, M.K., and Teare, M.D. (2007). A Bayesian approach to copy-number-polymorphism analysis in nuclear pedigrees. Am. J. Hum. Genet. *81*, 808–812.
17. Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wang, L., Hangaishi, A., Kurokawa, M., Chiba, S., Bailey, D.K., Kennedy, G.C., et al. (2005). A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res. *65*, 6071–6079.
18. Zhang, J., Feuk, L., Duggan, G.E., Khaja, R., and Scherer, S.W. (2006). Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet. Genome Res. *115*, 205–214.
19. Leykin, I., Hao, K., Cheng, J., Meyer, N., Pollak, M.R., Smith, R.J., Wong, W.H., Rosenow, C., and Li, C. (2005). Comparative linkage analysis and visualization of high-density oligonucleotide SNP array data. BMC Genet. *6*, 7.
20. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.
21. Carter, N.P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. Nat. Genet. *39*, S16–S21.
22. Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S.F., Hakonarson, H., and Bucan, M. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. *17*, 1665–1674.
23. Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G., and Chee, M.S. (2005). A genome-wide scalable SNP genotyping assay using microarray technology. Nat. Genet. *37*, 549–554.
24. Hinds, D.A., Kloek, A.P., Jen, M., Chen, X., and Frazer, K.A. (2006). Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat. Genet. *38*, 82–85.
25. Locke, D.P., Sharp, A.J., McCarroll, S.A., McGrath, S.D., Newman, T.L., Cheng, Z., Schwartz, S., Albertson, D.G., Pinkel, D., Altshuler, D.M., et al. (2006). Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am. J. Hum. Genet. *79*, 275–290.
26. Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. *39*, 906–913.
27. Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U., et al. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science *316*, 1341–1345.
28. van Heel, D.A., Franke, L., Hunt, K.A., Gwilliam, R., Zhernakova, A., Inouye, M., Wapenaar, M.C., Barnardo, M.C., Bethel, G., Holmes, G.K., et al. (2007). A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. Nat. Genet. *39*, 827–829.
29. Ceppellini, R., Siniscalco, M., and Smith, C.A. (1955). The estimation of gene frequencies in a random-mating population. Ann. Hum. Genet. *20*, 97–115.
30. van Es, M.A., van Vught, P.W., Blauw, H.M., Franke, L., Saris, C.G., Van den Bosch, L., de Jong, S.W., de Jong, V., Baas, F., van't Slot, R., et al. (2008). Genetic variation in DPP6 is associated with susceptibility to amyotrophic lateral sclerosis. Nat. Genet. *40*, 29–31.
31. Carlson, C.S., Smith, J.D., Stanaway, I.B., Rieder, M.J., and Nickerson, D.A. (2006). Direct detection of null alleles in SNP genotyping data. Hum. Mol. Genet. *15*, 1931–1937.
32. Hubbard, T.J., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., et al. (2007). Ensembl 2007. Nucleic Acids Res. *35*, D610–D617.
33. Stephensen, A.G. (2002). EVD: Extreme value distributions. R-News *2*, 31–32.
34. McKusick, V.A. (2007). Mendelian inheritance in man and its online version, OMIM. Am. J. Hum. Genet. *80*, 588–604.
35. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. Nucleic Acids Res. *32*, D277–D280.
36. Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J., Cusick, M.E., Roth, F.P., et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature *430*, 88–93.
37. Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., et al. (2007). Reactome: A knowledge base of biologic pathways and processes. Genome Biol. *8*, R39.
38. Alfarano, C., Andrade, C.E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., et al. (2005). The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res. *33*, D418–D424.
39. Mishra, G.R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T.M., et al. (2006). Human protein reference database–2006 update. Nucleic Acids Res. *34*, D411–D414.
40. Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., et al. (2007). IntAct–open source resource for molecular interaction data. Nucleic Acids Res. *35*, D561–D565.
41. Schymick, J.C., Scholz, S.W., Fung, H.C., Britton, A., Arepalli, S., Gibbs, J.R., Lombardo, F., Matarin, M., Kasperaviciute, D., Hernandez, D.G., et al. (2007). Genome-wide genotyping in amyotrophic lateral sclerosis and neurologically normal controls: First stage analysis and public release of data. Lancet Neurol. *6*, 322–328.
42. Fung, H.C., Scholz, S., Matarin, M., Simon-Sanchez, J., Hernandez, D., Britton, A., Gibbs, J.R., Langefeld, C., Stiegert,

M.L., Schymick, J., et al. (2006). Genome-wide genotyping in Parkinson's disease and neurologically normal controls: First stage analysis and public release of data. Lancet Neurol. *5*, 911–916.
43. Sollid, L.M. (2000). Molecular basis of celiac disease. Annu. Rev. Immunol. *18*, 53–81.
44. Karell, K., Louka, A.S., Moodie, S.J., Ascher, H., Clot, F., Greco, L., Ciclitira, P.J., Sollid, L.M., and Partanen, J. (2003). HLA types in celiac disease patients not carrying the DQA1*05–DQB1*02 (DQ2) heterodimer: Results from the European Genetics Cluster on Celiac Disease. Hum. Immunol. *64*, 469–477.
45. Monsuur, A.J., de Bakker, P.I., Alizadeh, B.Z., Zhernakova, A., Bevova, M.R., Strengman, E., Franke, L., van't Slot, R., van Belzen, M.J., Lavrijsen, I.C., et al. (2005). Myosin IXB variant increases the risk of celiac disease and points toward a primary intestinal barrier defect. Nat. Genet. *37*, 1341–1344.
46. Hunt, K.A., Zhernakova, A., Turner, G., Heap, G., Franke, L., Bruinenberg, M., Romanos, J., Dinesen, L.C., Ryan, A.W., Panesar, D., et al. (2008). Novel coeliac disease genetic risk loci with links to adaptive immunity. Nat. Genet., in press.
47. Liu, J., Juo, S.H., Holopainen, P., Terwilliger, J., Tong, X., Grunn, A., Brito, M., Green, P., Mustalahti, K., Maki, M., et al. (2002). Genomewide linkage analysis of celiac disease in Finnish families. Am. J. Hum. Genet. *70*, 51–59.
48. Greco, L., Babron, M.C., Corazza, G.R., Percopo, S., Sica, R., Clot, F., Fulchignoni-Lataud, M.C., Zavattari, P., Momigliano-Richiardi, P., Casari, G., et al. (2001). Existence of a genetic risk factor on chromosome 5q in Italian coeliac disease families. Ann. Hum. Genet. *65*, 35–41.
49. Greco, L., Corazza, G., Babron, M.C., Clot, F., Fulchignoni-Lataud, M.C., Percopo, S., Zavattari, P., Bouguerra, F., Dib, C., Tosi, R., et al. (1998). Genome search in celiac disease. Am. J. Hum. Genet. *62*, 669–675.
50. Babron, M.C., Nilsson, S., Adamovic, S., Naluai, A.T., Wahlstrom, J., Ascher, H., Ciclitira, P.J., Sollid, L.M., Partanen, J., Greco, L., et al. (2003). Meta and pooled analysis of European coeliac disease data. Eur. J. Hum. Genet. *11*, 828–834.
51. Riccioni, R., Saulle, E., Militi, S., Sposi, N.M., Gualtiero, M., Mauro, N., Mancini, M., Diverio, D., Lo Coco, F., Peschle, C., et al. (2003). C-fms expression correlates with monocytic differentiation in PML-RAR alpha+ acute promyelocytic leukemia. Leukemia *17*, 98–113.
52. Zapata-Velandia, A., Ng, S.S., Brennan, R.F., Simonsen, N.R., Gastanaduy, M., Zabaleta, J., Lentz, J.J., Craver, R.D., Correa, H., Delgado, A., et al. (2004). Association of the T allele of an intronic single nucleotide polymorphism in the colony stimulating factor 1 receptor with Crohn's disease: A case-control study. J. Immune Based Ther. Vaccines *2*, 6.
53. Estivill, X., and Armengol, L. (2007). Copy number variants and common disorders: Filling the gaps and exploring complexity in genome-wide association studies. PLoS Genet. *3*, 1787–1799.
54. Nguyen, D.Q., Webber, C., and Ponting, C.P. (2006). Bias of selection on human copy-number variants. PLoS Genet. *2*, e20.
55. Jeong, H., Mason, S.P., Barabasi, A.L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. Nature *411*, 41–42.
56. Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., and Barabasi, A.L. (2007). The human disease network. Proc. Natl. Acad. Sci. USA *104*, 8685–8690.
57. Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. *78*, 629–644.
58. Hedrick, P.W. (1987). Gametic disequilibrium measures: Proceed with caution. Genetics *117*, 331–341.
59. Zapata, C. (2000). The D' measure of overall gametic disequilibrium between pairs of multiallelic loci. Evolution Int. J. Org. Evolution *54*, 1809–1812.
60. Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S., Abeysinghe, S., Krawczak, M., and Cooper, D.N. (2003). Human Gene Mutation Database (HGMD): 2003 update. Hum. Mutat. *21*, 577–581.
61. Yu, Z., and Schaid, D.J. (2007). Methods to impute missing genotypes for population data. Hum. Genet. *122*, 495–504.
62. Slatkin, M., and Excoffier, L. (1996). Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. Heredity *76*, 377–383.