Published in final edited form as:
Genet Epidemiol. 2010 September ; 34(6): 582–590. doi:10.1002/gepi.20515.

On the genome-wide analysis of copy number variants in family-based designs: Methods for combining family-based and population based information for testing dichotomous or quantitative traits, or completely ascertained samples

Amy Murphy[1,2,3], **Sungho Won**[4], **Angela Rogers**[1,2], **Jen-Hwa Chu**[1,2], **Benjamin A Raby**[1,2,3], and **Christoph Lange**[2,3,4]

[1]Channing Laboratory, Brigham and Women's Hospital, Boston, MA

[2]Harvard Medical School, Boston, MA

[3]Center for Genomic Medicine, Brigham and Women's Hospital, Boston, MA

[4]Department of Biostatistics, Harvard School of Public Health, Boston, MA

Abstract

We propose a new approach for the analysis of copy number variants (CNVs)for genome-wide association studies in family-based designs. Our new overall association test combines the between-family component and the within-family component of the data so that the new test statistic is fully efficient and, at the same time, achieves the complete robustness against population-admixture and stratification, as classical family-based association tests that are based only on the between-family component. Although all data are incorporated into the test statistic, an adjustment for genetic confounding is not needed, not even for the between-family component. The new test statistic is valid for testing either quantitative or dichotomous phenotypes. If external CNV data are available, the approach can also be used in completely ascertained samples. Similar to the approach by Ionita-Laza et al.(1), the proposed test statistic does not required a CNV-calling algorithm and is based directly on the CNV probe intensity data. We show, via simulation studies, that our methodology increases the power of the FBAT statistic to levels comparable to those of population-based designs. The advantages of the approach in practice are demonstrated by an application to a genome-wide association study for body mass index (BMI).

1. Introduction

Recent studies (2;3;4;5;6) have highlighted the importance of copy number variation in modifying the risk of certain diseases. With the increasing interest in the study of copy number variants (CNVs), a number of methodologies for analyzing these data have been developed for both population-based and family-based analyses (1;7;8;9;10;11). Recently, Ionita-Laza et al. (2008) (1) introduced a statistic for association testing of CNVs in family-based settings. Noting that, for SNP-chips, the calling of CNVs can be a difficult and error-prone process when standard calling algorithms are applied(12;13;14;15), the test statistic is constructed using genetic information based on the probe intensity data rather than by the

URL

The testing strategy has been fully implemented in the software package *PBAT*, which is freely available at http://www.biostat.harvard.edu/~clange/default.htm (31;34).

called CNVs. Since then, new, improved algorithms to call CNVs' genotypes based on intensity data from SNP-chips(16;17;18) have been developed. While the algorithms are able to provide important genomic insights into the underlying genetic architecture, the advantages of using called CNVs over probe intensities in the association test statistic are not clear. In the best-case scenario for the calling-algorithms, (i.e. genotypes called based on intensities for SNPs), Ionita-Laza et al.(1) showed that there is virtually no difference in terms of statistical power between association tests that are based on genotype data and association tests that are based on intensity data. Furthermore, the advantage of using intensity data is that it avoids bias due to incorrectly called CNV genotypes. In family-based association tests, mis-called genotypes for SNPs or CNVs can introduce a systematic bias in terms of increased α-levels (19;20;21;22;23).

Currently, the methodology by Ionita-Laza et al. (2008) (1) for family-based association test in CNV analysis solely utilizes the within-family component. Similar to the standard FBAT-approach, the between-family component can only be incorporated into test-strategies and model building steps in which the between-family component is used to inform the "final" FBAT-testing step or to construct an FBAT-statistic with maximal statistical power(24;25). In order to maintain complete robustness against confounding due to population admixture and stratification, the between-family component can not be included directly in the test statistic.

In this manuscript, we develop an overall family-based association test that simultaneously utilizes all information about the genetic association that is included in both components, i.e., the between-family component and the within-family component. However, by virtue of its design, the new overall association test for CNVs maintains the complete robustness of classical family-based association tests, e.g., TDT (26), FBAT(27;28). The new FBAT-CNV test combines an association test for the population-based component, the between-family component, with a standard FBAT test that is based only on the within-family component, using Liptak's method for combining p-values. Similar to the philosophy of genomic control(29), the significance of the association test for the between-family component is not assessed based on its asymptotic distribution, but relative to the other CNVs, using rank-based p-values. As will be shown, the use of rank-based p-values in the overall test statistic is sufficient to ensure that the approach is robust against population admixture and stratification.

The proposed overall CNV-FBAT statistic utilizes the standard FBAT-CNV statistic(1) and an association test for the between-family component that is constructed based on the conditional mean model or its non-parametric extensions (24;25). It is therefore straight-forward to apply the new approach to any trait types which have been integrated into the FBAT-framework, e.g., binary traits, continuous traits, multivariate traits, time-to-onset, etc. (30;31;32;33;34;35;36). However, for simplicity, we will illustrate the new overall FBAT-CNV statistic for scenarios in which dichotomous or quantitative traits are analyzed and for samples that are either unascertained, i.e., total population samples, or completely ascertained, i.e., every subject is affected. Using simulation studies that are based on ~900,000 CNV probes, the performance of the overall FBAT-CNV statistic is assessed under the null-hypothesis and its power levels are estimated. The features of the new approach are demonstrated by an application to a genome-wide association study for BMI.

2. Methods

Prior to describing the joint test, we briefly review the FBAT-CNV statistic introduced by Ionita-Laza and colleagues(1). We assume that genotype intensity data have been collected for nuclear families (i.e., offspring and parents). Let x_{ij} denote the probe intensity for the j^{th}

offspring in the i^{th} family. The recoded offspring phenotype is denoted by T_{ij}, which represents the offspring phenotype minus an offset value. For example, the phenotype is quantitative, T_{ij} might be residual from the regression of the offspring phenotype on important clinical predictors of the phenotype (e.g. $BMI \sim SEX + AGE$). The FBAT statistic for copy number variation, introduced by Ionita-Laza et al. (2008)(1) is given by:

$$Z_{FBAT-CNV} = \frac{\Sigma_{ij}[T_{ij}(x_{ij} - E(X_i))]}{\left[\Sigma_{ij}\Sigma_{ij'}T_{ij}T_{ij'}(x_{ij} - E(X_i))(x_{ij'} - E(X_i))\right]^{0.5}} \quad (2.1)$$

where $E(X_i)$ is the 'expected' intensity score given the parental intensities. When parental intensities are available, $E(X_i)$ is given by the mean of parental probe intensities. If parental data are missing, $E(X_i)$ may be obtained by averaging the intensity data across all offspring within a family $E(X_i) = \Sigma_j x_{ij}/n_j$. This concept is similar to the sufficient statistic given by Rabinowitz and Laird (2000) (28). It should also be noted that the empirical variance as given by Lake et al. (2000) (37), rather than the theoretical variance, is used to standardize the test statistic, since Mendelian transmissions cannot be identified in the intensity data. For further details on the derivation of the test statistic, please see Ionita-Laza et al. (2008)(1).

2.1. Quantitative Traits

Next, we turn our attention to the joint test statistic, as originally developed by Won et al. (2009)(38). Won proposes an overall family-based association test statistic, Z_i, which is a weighted sum:

$$Z_i = \omega_{FBAT} Z_{FBAT_i} + \omega_P Z_{P_i} \quad (2.2)$$

The parameters ω_{FBAT} and ω_P are specified a priori to the analysis, and are standardized so that $\omega^2_{FBAT} + \omega^2_P = 1$. Z_{FBAT_i} is the Z-score corresponding to the p-value of the i^{th} FBAT statistic, and Z_{P_i} is the Z-score corresponding to a p-value that reflects the relative ranking of the i^{th} population-based test statistic. To preserve the type I error of the test in the presence of population stratification, the p-values obtained from Z_{P_i} are based on the relative ranking of the i^{th} test statistic, rather than its asymptotic value. For details, see Won et al. (38). This method of combining independent test statistics was originally conceived by Liptak(39), and henceforth, the joint test of family-based and population based information will be referred to as the FBAT-Liptak test.

2.1.1. Implementation of FBAT-CNV Liptak test for Quantitative or dichotomous traits—

The extension of the method by Won et al. (2009)(38) to copy number variants is fairly straightforward, when using the statistic developed by Ionita-Laza et al. (2008)(1). In the joint test statistic, Z_i, the family component Z_{FBAT_i} is replaced by $Z_{FBAT-CNV_i}$, as given in equation 2.1. The between-family component of the test statistic, Z_{P_i} is obtained from a Wald test of the genetic effect size from a regression of the offspring phenotype on the expected offspring probe intensity, $E(X_i)$, using a generalization of the conditional mean model(24;30). In the case of CNVs, this quantity is given by the mean of parental probe intensities (or in the event of missing parental data, $E(X_i) = \Sigma_j x_{ij}/n_j$), rather than by the expected transmissions based on parental genotypes, assuming Mendelian transmission, as is the case with SNP data. This concept is applicable to both quantitative or dichotomous (i.e., case-control) traits, using an identity or logit link, respectively, in a generalized linear model. As noted above, to protect against potential confounding such as by population stratification, the p-values from Z_{P_i} reflect the relative rank of the CNV rather than an asymptotic value obtained from standard normal or student's t distribution. As with the standard method using SNP data, the joint test statistic reflects the summed weighted

family-based and population-based components, and may be adjusted for multiple comparisons using standard approaches. It should be noted, however, that while the FBAT-CNV statistic may detect either de novo or germline mutations, the power to detect de novo mutations will be decreased using this method, as the between-family component solely is based on parental intensities.

2.1.2. Implementation of Liptak FBAT-CNV for fully ascertained family-based samples—Recently, Lasky-Su and colleagues (2009)(36) extended the idea proposed by Won et al.(38) to family-based studies in which all offspring in the study are affected with the disease or trait of interest. From equation 2.2, it is easy to see that a standard Transmission Disequilibrium Test (TDT)(26) could be used to calculate Z_{FBAT_i}. However, family-based designs where all offspring are affected pose a problem for calculating the effect size for the between-family component, as the conditional mean model(24; 30) requires variation in the offspring phenotype. Murphy et al.(40) propose using relative risk ratios based on the parental genotypes (which assume Hardy-Weinberg Equilibrium in the general population) for estimating the between-family effect size for SNP data, but this method is not applicable in the context of CNVs. Lasky-Su et al.(36) address the issue of generating between-family information by using control genotype data from either publicly available (e.g., dbGap) or commercial (e.g., Illumnina, Affymetrix, and Perlegen) sources. However, the method by Lasky-Su et al. requires called genotype data and thus is not applicable to CNV intensities. For CNV data, we will extend the concept of using freely available genetic data to devise a test for generating the statistic Z_{P_i} in 2.2.

We assume that CNV probe intensity values will be available for a population of control subjects from a publicly available source. Let Z_{P_i} be the Z-score corresponding to the rank of the p-value from a Wald test from a generalized linear model with a logit link:

$$Y_{ij} \sim I(G(X_{ij})), \text{ where: } \begin{cases} I(G(X_{ij}))=E(X_i) \text{ if the subject is a case} \\ I(G(X_i j))=X_{ij} \text{ if the subject is a control} \end{cases} \quad (2.3)$$

where Y_{ij} is the offspring or control phenotype (i.e., Y_{ij}=1 if an affected offspring, 0 if a control), $E(X_i)$ is the expected offspring CNV probe intensity, and X_{ij} is observed CNV probe intensity in the control. As noted previously, $E(X_i)$ is given by the mean of parental probe intensities, or in the event of missing parental data, $E(X_i) = \Sigma_j x_{ij}/n_j$. After obtaining the Wald test statistic rank-based p-value of the CNV genetic effect size for the population based component (Z_{P_i}), and the $FBAT-CNV$ statistic asymptotic p-value for the family-based component (Z_{FBAT_i}), the method by Won et al. follows straightforwardly(38).

As shown in the next section, while this method has slightly less power than a standard case-control test (where the asymptotic p-value from the Wald test is used), its power loss is minimal while preserving the attractive feature of FBAT robustness against population structure.

3. Simulation Studies

3.1. Quantitative trait

Using simulation studies, we contrast the new testing strategy to a both a standard FBAT statistic, as well as a standard population-based analysis (i.e., a generalized linear model regressing the phenotype on the observed CNV intensity) of only the offspring. For the Liptak-CNV test, we varied the weights for the FBAT and Wald components of the of the statistic from 1:3 and 3:1 (i.e., the FBAT weight ranged from 1 to 3 times the weight of the Wald weight, and vice-versa), and set δ=0.5. For details on the weights and appropriate specifications for δ, see Won et al.(38). All three analytic approaches are compared under

scenarios with varying sample and effect sizes. We simulated the trio data under the assumption that all of the offspring data are available and that the genotype probe intensity of the parents are known. Additionally, we assumed independence between the CNV loci. The parental CNV data were generated from a N(0,1) distribution. The offspring intensities were generated using a N(EX,1) distribution, where EX is the mean of the parental intensities.

In each simulation, one locus is assumed to be the DSL, while the other CNVs are considered null loci. For the null loci, the offspring phenotype was simulated using a N(0,1) distribution. For the DSL, the distribution was N($a * X$), where X is the probe intensity of the CNV locus. In the simulations, the genetic effect size ranges between 0.1 and 0.2. The trio sample size varies between 500 – 2000 trios.

For each test, the p-value was adjusted using a standard Bonferroni correction. Thus, the power was defined as the proportion of replicates with an FBAT statistic p-value < $5.56\times10^{-}8$ (i.e., 0.05/900,000), to reflect the current capacity of CNV typing technology. The power was based on 10,000 replicates. In table 1 below, the standard FBAT test is denoted as "FBAT," the population-based test is denoted as "Wald", and our new integrated approach using both the population and family based information is denoted as "FBAT-Liptak."

Across all sample and effect size, the FBAT-Liptak test demonstrates considerable power gains over the standard FBAT test, and generally has less than a <10% power loss in comparison to the population-based Wald test. In the scenario with equal weighting, 1000 trios and an effect size of 0.175, the FBAT-Liptak test has a 200% power gain over the FBAT test, while the Wald test only has a 6% power gain over the FBAT-Liptak test. From the simulations, the optimal weighting scheme tends to range between weighting the FBAT component 2 to 3 times more heavily than the Wald component. In Won et al. (38), the optimal weights for the FBAT and Wald components are described as a ratio of their standardized effect sizes. Overall, the power estimates in the FBAT-Liptak test demonstrate modest power reductions in comparison to a population based test, while preserving one of the most desirable properties of the FBAT, robustness against population stratification.

3.2. Completely ascertained sample

In this section, we also compare the new testing strategy to a both the standard FBAT statistic and population-based test, except that a logit link was used in the generalized linear model for the population based analysis. The CNV intensities for the unrelated control population (i.e., to be obtained from a publicly available resource) were simulated using a N(0,1) distribution. The trio CNV intensities, were simulated using the methods described above, except that large trio (i.e., parents and offspring) population, from which "affected" offspring were sampled, using the following approach.

As above, we assumed that there was one locus contributing to disease risk and that there was no LD between the loci. We assumed that the baseline risk of the disease, K, was 5%. We used a logistic regression model to estimate the risk of disease for each offspring, where $Pr(\text{Affected}) = e^{\alpha+\beta X}/1 + e^{\alpha+\beta X}$, where $\alpha = log(K/1 - K)$, β is the log of the odds ratio, and X is the observed CNV intensity in the offspring. From the pool of trio CNV intensity data, we randomly selected a subset of affected offspring and their parents, to simulate a completely ascertained sample. The controls were all assigned an affection status of unaffected, regardless of their CNV intensity value, to reflect the fact that a proportion of subjects in the "control" dataset are likely to have the disease or trait of interest, at a rate roughly equivalent to the population prevalence of the disease. We denote these subjects as "unselected controls."

The simulations included sample sizes of 250–500 trios, where the corresponding population-based test (i.e., denoted as "Wald") comprises 250–500 pairs of unrelated individuals, the affected offspring from the trios and the unrelated control population. The effect size ranged from an odds ratio of 1.25 – 2. As above, the power was defined as the proportion of replicates with an FBAT statistic p-value < 5.56×10^{-8} (i.e., 0.05/900,000), applying standard Bonferonni correction for multiple comparisons. The power was based on 10,000 replicates. The results are displayed in table 2 below.

For completely ascertained samples, the power of both the FBAT and Liptak-CNV tests generally exceed the population-based Wald test. It should be noted that the simulations are designed such that the number of trios and number of case-control pairs (e.g. 250 trios/250 case-control pairs) are equivalent. With an equal weighting scheme, the power for the FBAT and Liptak tests are fairly comparable, although the Liptak-CNV test does give a power boost in lower power ranges. When the weights for the FBAT and Wald compon-ponents of the Liptak-CNV are varied, an up-weighting of the FBAT component boosts the power, with the 3:1 weighting of the FBAT to Wald component, respectively, consistently demonstrating the highest power.

4. Data Analysis

Asthma is a complex respiratory disorder, with both environmental and genetic components, which has been shown to have substantial heritability (41;42;43). We applied our methodology to a genome-wide association study of child asthmatics and their families. The families were identified through the Childhood Asthma Management Program (CAMP) (44) Genetics Ancillary Study. Increasing body mass index (BMI) and obesity has been previously linked to increasing asthma risk, and a recent study(45) found a gene associated with both BMI and asthma. In our analysis, we screened for potential copy number variants for associated with BMI.

In the CAMP study, SNP genotyping was performed using Illumina HumanHap 550v3 array. Of the 561,466 SNPs, 16,419 (2.9%) were removed during data cleaning due to the following reasons: 1) probe sequences did not map uniquely to hg18 genome build, 2) poor cluster separation as manually reviewed in Illumina BeadStudio software, 3) −log10(pval) for Hardy-Weinberg equilibrium >= 8, and/or 4) completion rate < 95%, 5) Mendelian error count >=5, or 6) minor allele frequency = 0. Overall, adequate DNA for genome-wide SNP genotyping was available for 1172 subjects in 403 families (43 subjects in 19 families were removed due to inadequate DNA samples)

To reduce the number of multiple comparisons (and false positive results), we applied a circular binary segmentation algorithm to identify the most likely CNV regions in our cohort. Circular binary segmentation (CBS) uses SNP intensity at sequential markers to call CNV regions(46). The goal of CBS is to parse the genome into segments of equal copy number, while accounting for the inherent noise in measuring array intensities. This is achieved by identifying change-points, locations at which the distribution preceding the change-point differs from the distribution proceeding the change-point. The change- or break- points may then be used to limit testing to regions where copy number gains or losses have occurred. Practically speaking, this algorithm allows us to restrict the testing of potential trait-associated CNVs to markers that fall within CNV regions, thus substantially reducing the number of association tests conducted. The CBS algorithm was applied to all autosomal SNP in the CAMP dataset using the Bioconductor package DNAcopy(47). In our analysis, a CNV was called in an individual when the (absolute) intensity value of a segment was greater than four standard deviations of the middle 50% quantile of all data(48). For inclusion in the analysis, we required CNV prevalence of at least 1% (i.e., present in 12

individuals). If called markers were within 50kb of one another, they were considered part of the same CNV region. Using this approach, 976 common CNVs (from 6,965 SNPs) were identified in the CAMP data set with frequency of at least 1%. For the analysis, we applied our joint test statistic to the probe intensities of all 6,965 (called CNV region) SNP, testing for potential association with BMI at baseline, adjusting for age and sex. Table 3 displays the results for the CAMP data analysis.

Interestingly, two of the top SNPs (rs1758827 and rs1886314) are in intronic regions of protocadherin genes (*PCDH15* and *PCDH9* on chromosomes 10 and 13, respectively), which are members of the cadherin superfamily. The top two results were significant (at α=0.05) after Bonferroni correction for multiple comparisons. As noted in Figure 1, there is significant enrichment in the regions surrounding the top SNP on chromosomes 9 and 10. Overall, among the top findings, there was consistency in the family-based and population-based association results.

5. Discussion

In this manuscript, we present a new method for maximizing the available information in family-based genome-wide studies of copy number variants. Our new methodology combines the population level and family level data in family-based designs into a more powerful association test. This test can be applied to both quantitative and dichotomous traits. Additionally, although public repositories of CNV data are generally not currently available, it is not unreasonable to anticipate their availability in the near future, given the current initiatives to develop a public repository of both phenotype and genome-wide SNP data (e.g. dbGap, http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gap). When control CNV data become available from a public repository, our methodology also can be applied to completely ascertained family-based samples. For continuous traits, our methodology approaches that of a strictly population-based test. For completely ascertained samples, we have shown that our method can be more powerful than a population based approach combining affected offspring with unrelated controls. And unlike a population-based approach, our method is robust against population admixture(38), like the standard FBAT statistic. The power of our methodology may be further increased by using alternative approaches for multiple comparisons adjustment. Although we have applied a very conservative Bonferroni correction to our combined test statistic, other methods, such as FDR(49) could be applied.

Acknowledgments

We thank all subjects for their ongoing participation in this study. We acknowledge the CAMP investigators and research team, supported by NHLBI, for collection of CAMP Genetic Ancillary Study data. All work on data collected from the CAMP Genetic Ancillary Study was conducted at the Channing Laboratory of the Brigham and Women's Hospital under appropriate CAMP policies and human subject's protections. The CAMP Genetics Ancillary Study is supported by U01 HL075419, U01 HL065899, U01 HL089897, U01 HL089856, P01 HL083069, R01 HL086601, and T32 HL07427 from the National Heart, Lung and Blood Institute, National Institutes of Health. C.L. is supported by the National Institutes of Health grant R01 59532.

REFERENCES

1. Ionita-Laza I, Perry G, Raby B, Klanderman B, Laird NM, et al. On the analysis of copy-number variations in genome-wide association studies: a translation of the family-based association test. Genet Epidemiol. 2008; 32:273–284. [PubMed: 18228561]
2. Yang TL, Chen XD, Guo Y, Lei SF, Wang JT, et al. Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis. Am J Hum Genet. 2008; 83:663–674. [PubMed: 18992858]

3. Vrijenhoek T, Buizer-Voskamp JE, van der Stelt I, Strengman E, Sabatti C, et al. Recurrent CNVs disrupt three candidate genes in schizophrenia patients. Am J Hum Genet. 2008; 83:504–510. [PubMed: 18940311]
4. Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, et al. Structural variation of chromosomes in autism spectrum disorder. Am J Hum Genet. 2008; 82:477–488. [PubMed: 18252227]
5. Northcott PA, Nakahara Y, Wu X, Feuk L, Ellison DW, et al. Multiple recurrent genetic events converge on control of histone lysine methylation in medulloblastoma. Nat Genet. 2009; 41:465–472. [PubMed: 19270706]
6. Kathiresan S, Voight B, Purcell S, Musunuru K, Ardissino D, et al. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat Genet. 2009; 41:334–341. [PubMed: 19198609]
7. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet. 2008; 40:1253–1260. [PubMed: 18776909]
8. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet. 2008; 40:1166–1174. [PubMed: 18776908]
9. Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, et al. A robust statistical method for case-control association testing with copy number variation. Nat Genet. 2008; 40:1245–1252. [PubMed: 18776912]
10. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, et al. Strong association of de novo copy number mutations with autism. Science. 2007; 316:445–449. [PubMed: 17363630]
11. Shrestha S, Aissani B, Tiwari H, Wiener H. Association test of multiallelic gene copy numbers in family trios. Genet Epidemiol. 2009 epub ahead of print.
12. Tuzun E, Sharp A, Bailey J, Kaul R, Morrison V, et al. Fine-scale structural variation of the human genome. Nat Genet. 2005; 37:727–732. [PubMed: 15895083]
13. Komura D, Shen F, S I, Fitch K, Chen W, et al. Genome-wide detection of human copy number variations using high-density dna oligonucleotide arrays. Genome Res. 2006; 16:1575–1584. [PubMed: 17122084]
14. Redon R, Ishikawa S, Fitch K, Feuk L, Perry G, et al. Global variation in copy number in the human genome. Nature. 2006; 7118:444–454. [PubMed: 17122850]
15. Korbel J, Urban A, Grubert F, Du J, Royce T, et al. Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. Proc Natl Acad Sci. 2007; 104:10110–10115. [PubMed: 17551006]
16. Wang K, Li M, Hadley D, Liu R, Glessner J, et al. Penncnv: an integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data. Genome Res. 2007; 11:1665–1674. [PubMed: 17921354]
17. Franke L, de Kovel C, Aulchenko Y, Trynka G, Zhernakova A, et al. Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. Am J Hum Genet. 2008; 82:1316–1333. [PubMed: 18519066]
18. Zollner S, Su G, Stewart WC, Chen Y, McInnis MG, et al. Bayesian EM algorithm for scoring polymorphic deletions from SNP data and application to a common CNV on 8q24. Genet Epidemiol. 2009; 33:357–368. [PubMed: 19085946]
19. Mitchell A, Cutlter D, Chakravarti A. Undetected genotyping errors cause apparent over-transmission of common alleles in the transmission/disequilibrium test. Am J Hum Genet. 2003; 72:598–610. [PubMed: 12587097]
20. Hao K, Li C, Rosenow C, Wong W. Estimation of genotype error rate using samples with pedigree information–an application on the genechip mapping 10k array. Genom. 2004; 84:623–630.
21. Hirschhorn J, Daly M. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet. 2005; 6:95–108. [PubMed: 15716906]
22. Cheng K, Chen J. A simple and robust tdt-type test against genotyping error with error rates varying across families. Hum Hered. 2007; 64:114–122. [PubMed: 17476111]
23. Hao K, Cawley S. Differential dropout among snp genotypes and impacts on association tests. Hum Hered. 2007; 63:219–228. [PubMed: 17347569]

24. Lange C, DeMeo D, Silverman EK, Weiss S, Laird NM. Using the noninformative families in family-based association tests: A powerful new testing strategy. Am J Hum Genet. 2003b; 79:801–811.
25. Lange C, Lyon H, DeMeo D, Raby B, Silverman E, et al. A new powerful non-parametric two-stage approach for testing multiple phenotypes in family-based association studies. Human Heredity. 2003c; 56:10–17. [PubMed: 14614234]
26. Spielman RS, McGinnis RE, Ewens WJ. Transmisson test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDm). Am J Hum Genet. 1993; 52:506–516. [PubMed: 8447318]
27. Laird NM, Horvath S, Xu X. Implementing a unified approach to family based tests of association. Genet Epi. 2000; 19:S36–S42.
28. Rabinowitz D, Laird NM. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered. 2000; 50:211–223. [PubMed: 10782012]
29. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. Theor Popul Biol. 2001; 60:155–166. [PubMed: 11855950]
30. Lange C, Silverman EK, Xu X, Weiss S, Laird NM. A multivariate family-based association test using generalized estimating equations: FBAT-GEE. Biostat. 2003a; 4:195–206.
31. Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM. PBAT: Tools for family-based association studies. Am J Hum Genet. 2004; 74:367–369. [PubMed: 14740322]
32. Lange C, Blacker D, Laird N. Family-based association tests for survival and times-to-onset analysis. Statistics in Medicine. 2004; 23:179–189. [PubMed: 14716720]
33. Lange C, Van Steen K, Andrew T, Lyon H, DeMeo D, et al. A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects. Statistical Applications in Genetics and Molecular Biology. 2004; Vol 3(No 1) Article 17 http://wwwbepresscom/sagmb/vol3/iss1/art17.
34. Van Steen K, Lange C. PBAT: a comprehensive software package for genome-wide association analysis of complex family based studies. Hum Genomics. 2005a; 2 67–6.
35. Jiang H, Harrington D, Raby BA, Bertram L, Blacker D, et al. Family-based association test for time-to-onset data with time-dependent differences between the hazard functions. Genet Epi. 2006; 30:124–132.
36. Lasky-Su J, Won S, Weiss S, Lange C. On genome-wide association studies in family-based design: An integrative analysis approach combing ascertained family-samples with unselected controls. In Preparation. 2009
37. Lake S, Blacker D, Laird N. Family-based tests of association in the presence of linkage. Amer J Hum Genet. 2000; 67:1515–1525. [PubMed: 11058432]
38. Won S, Wilk J, Mathias R, O'Donnell J, Silverman E, et al. On the analysis of genome-wide association studies in family-based designs: A universal, robust analysis approach and an application to four genome-wide association studies. PLoS Genet. 2009 under review.
39. Liptak T. On the combination of independent tests. agyar Tud Akad Mat Kutato' IntKo"zl. 1958; 3:171.
40. Murphy A, Weiss S, Lange C. Screening and replication using the same data set: Testing strategies for family-based studies in which all probands are affected. PLoS Genet. 2008; 4:e1000197. [PubMed: 18802462]
41. Tan H, Walker M, Gagnon F, Wen SW. The estimation of heritability for twin data based on concordances of sex and disease. Chronic Dis Can. 2005; 26:9–12. [PubMed: 16117840]
42. Koeppen-Schomerus G, Stevenson JRP. Genes and environment in asthma: a study of 4 year old twins. Arch Dis Child. 2001; 85:398–400. [PubMed: 11668102]
43. Mathias RA, Freidhoff LR, Blumenthal MN, Meyers DA, Lester L, et al. Genome-wide linkage analyses of total serum IgE using variance components analysis in asthmatic families. Genet Epi. 2001; 20:340–355.
44. CAMP. The childhood asthma management program (CAMP): design, rationale, and methods. childhood asthma management program research group. Control Clin Trials. 1999; 20:91–120. [PubMed: 10027502]

45. Murphy A, Tantisira KG, Soto-Quirs ME, Avila L, Klanderman BJ, et al. PRKCA: a positional candidate gene for body mass index and asthma. Am J Hum Genet. 2009; 85:87–96. [PubMed: 19576566]

46. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004; 5:557–572. [PubMed: 15475419]

47. Venkatraman ES, Olshen AB. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics. 2007; 23:657–663. [PubMed: 17234643]

48. Aguirre AJ, Brennan C, Bailey G, Sinha R, Feng B, et al. High-resolution characterization of the pancreatic adenocarcinoma genome. Proc Natl Acad Sci USA. 2004; 101:9067–9072. [PubMed: 15199222]

49. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Royal Stat Soc B. 1995; 57:289–300.

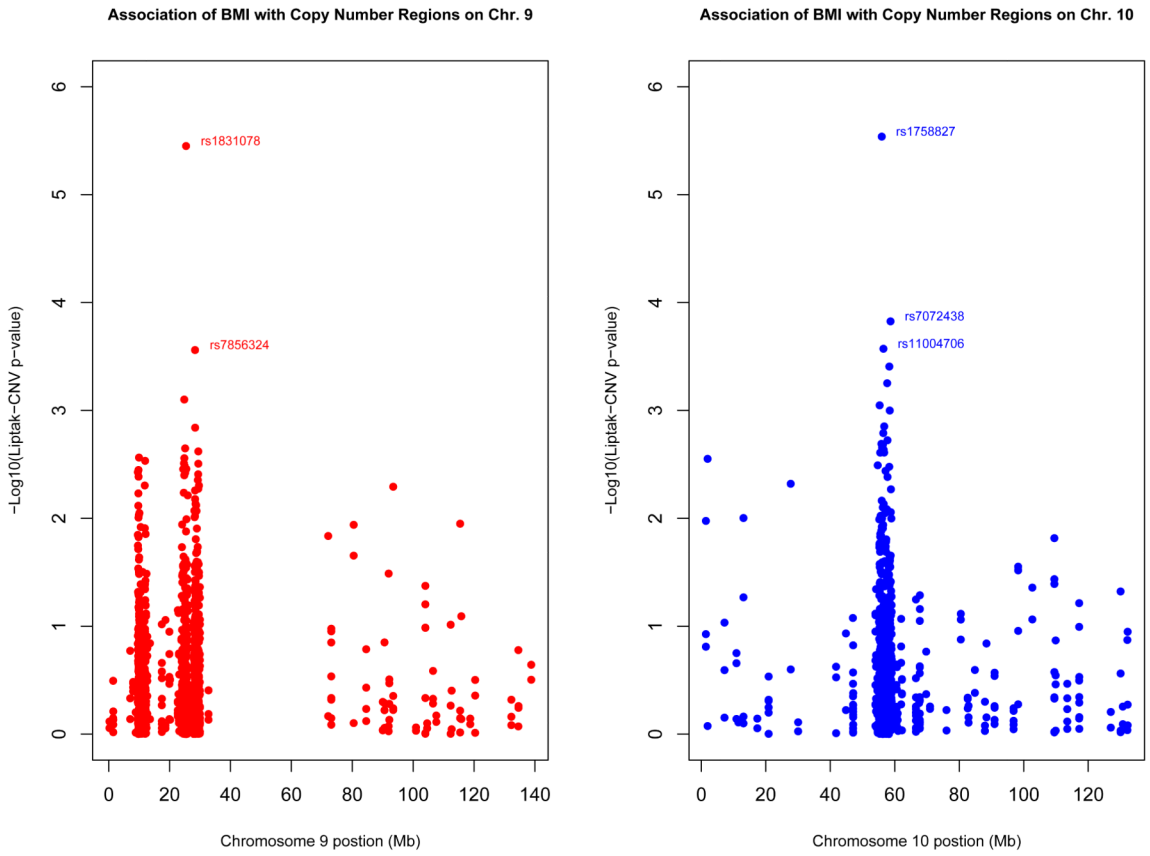

Fig. 1. Association of BMI with Copy Number Regions on Chromosomes 9 and 10

These figures plot the −Log10 Liptak-CNV p-values against the chromosomal position of SNP in identified copy number regions on chromosomes 9 and 10. There are areas of significant enrichment on both chromosomes, centered around two markers, rs1831078 and rs1758827, which met genome-wide significance criteria.

Table 1

Power for 500–2000 trios and 900,000 markers for a quantitative trait.

N	Effect			Equal weights	FBAT up-weighted		Wald up-weighted	
trios	Size	FBAT	Wald	Liptak(1:1)	Liptak(2:1)	Liptak(3:1)	Liptak(1:2)	Liptak(1:3)
500	0.1	0.000	0.004	0.003	0.001	0.002	0.002	0.001
	0.125	0.000	0.026	0.016	0.013	0.012	0.011	0.009
	0.15	0.004	0.095	0.061	0.060	0.062	0.042	0.036
	0.175	0.024	0.259	0.176	0.182	0.183	0.133	0.108
	0.2	0.075	0.504	0.376	0.396	0.392	0.309	0.251
1000	0.1	0.008	0.058	0.041	0.054	0.049	0.033	0.024
	0.125	0.043	0.265	0.210	0.242	0.231	0.172	0.136
	0.15	0.172	0.637	0.547	0.587	0.577	0.467	0.402
	0.175	0.417	0.906	0.853	0.879	0.869	0.786	0.721
	0.2	0.712	0.987	0.972	0.983	0.980	0.948	0.917
2000	0.1	0.144	0.519	0.458	0.488	0.478	0.378	0.297
	0.125	0.511	0.922	0.891	0.907	0.895	0.823	0.753
	0.15	0.867	0.997	0.994	0.996	0.995	0.985	0.972
	0.175	0.987	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999
	0.2	0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999

The standard FBAT test is denoted as “FBAT,” the population-based test is denoted as “Wald”, and our new integrated approach using both the population and family based information is denoted as “FBAT-Liptak. The numbers in parentheses, reflect the weighting of the FBAT component of the Liptak-CNV test relative to the Wald component of the Liptak-CNV, (e.g., (2:1) indicates that the FBAT component has twice the weight of the Wald component). The power reflects the proportion of replicates were the association p-value$< 5.56\times10^{-8}$.

Table 2

Power for 250–1000 trios/subjects and 900,000 markers in a completely ascertained sample

N trios/	Effect			Equal weights	FBAT up-weighted		Wald up-weighted	
pairs	Size	FBAT	Wald	Liptak(1:1)	Liptak(2:1)	Liptak(3:1)	Liptak(1:2)	Liptak(1:3)
250	1.25	0.004	0.008	0.009	0.013	0.014	0.007	0.005
	1.375	0.104	0.116	0.121	0.177	0.196	0.075	0.050
	1.5	0.446	0.462	0.470	0.586	0.623	0.304	0.217
	1.625	0.804	0.796	0.803	0.889	0.910	0.623	0.486
	1.75	0.957	0.949	0.949	0.983	0.991	0.850	0.734
500	1.25	0.128	0.128	0.147	0.208	0.227	0.086	0.054
	1.375	0.749	0.705	0.756	0.851	0.875	0.570	0.431
	1.5	0.990	0.974	0.984	0.997	0.998	0.935	0.859
	1.625	>0.999	>0.999	>0.999	>0.999	>0.999	0.997	0.984
	1.75	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	0.999
1000	1.25	0.776	0.696	0.773	0.869	0.891	0.594	0.450
	1.375	>0.999	0.997	0.999	1.000	>0.999	0.992	0.972
	1.5	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999
	1.625	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999
	1.75	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999	>0.999

The number of trios reflects the number of trios included in either the FBAT or in the family-based component of the FBAT-Liptak tests. The number of subjects reflects the total number of subjects used (including both the offspring from the trios and unselected controls) used in the Wald or in the population-based component of the Wald-Liptak test. The standard FBAT test is denoted as “FBAT,” the population-based test is denoted as “Wald”, and our new integrated approach using both the population and family based information is denoted as “FBAT-Liptak. The numbers in parentheses, reflect the weighting of the FBAT component of the Liptak-CNV test relative to the Wald component of the Liptak-CNV, (e.g., (2:1) indicates that the FBAT component has twice the weight of the Wald component). The power reflects the proportion of replicates were the association p-value< 5.56×10^{-8}.

Table 3

Top 10 Results for identified CNV regions, sorted by Liptak p-value, for the CAMP BMI analysis

SNP	Chr.	Position* (bp)	FBAT p-value	Rank-based Wald p-value	Liptak test	Liptak p-value
rs1758827	10	56295820	4.931E-03	7.128E-05	4.68	2.902E-06
rs1831078	9	25416556	2.411E-03	2.138E-04	4.64	3.547E-06
rs1886314	13	67392298	3.583E-04	1.019E-02	4.16	3.133E-05
rs9351261	6	65171419	1.999E-02	4.989E-04	3.97	7.118E-05
rs11758713	6	67036364	2.593E-04	3.899E-02	3.83	1.285E-04
rs7072438	10	59013036	2.775E-02	7.840E-04	3.79	1.495E-04
rs696279	7	16700798	3.811E-03	1.133E-02	3.66	2.547E-04
rs11004706	10	56813669	7.685E-02	3.564E-04	3.64	2.681E-04
rs7856324	9	28366643	1.814E-03	2.145E-02	3.64	2.756E-04

*SNP positions are from NCBI build 37.1. The FBAT p-value is the asymptotic p-value obtained from the Ionita-Laza FBAT – CNV test statistic. The Observed Wald p-value is the asymptotic p-value from a Wald test regressing the BMI phenotype on the expected parental CNV intensities. The Rank-based Wald p-value reflects the p-value based on the rank of the Wald test. The Liptak test is the value of the test statistic given in equation 2.2, which combines the FBAT p-value and the Rank-based Wald p-value. The Liptak p-value is the asymptotic p-value from this combined test statistic. Results with a Liptak p-value$< 7.18\times10^{-6}$ are significant at the α=0.05 level after Bonferroni correction for multiple comparisons.