Published in final edited form as:
Genet Epidemiol. 2011 ; 35(Suppl 1): S41–S47. doi:10.1002/gepi.20648.

Population-based and Family-based Designs to Analyze Rare Variants in Complex Diseases

Rémi Kazma[1,*] and **Julia N Bailey**[2,3]

[1]Department of Epidemiology and Biostatistics and Institute for Human Genetics, University of California, San Francisco, California

[2]Department of Epidemiology, David Gethen School of Medicine, University of California, Los Angeles, California

[3]Research Service, VA GLAHS Epilepsy Center of Excellence, Epilepsy Genetics/Genomics Laboratories, Los Angeles, California, USA

Abstract

Genotyping of rare variants on a large scale is now possible using next-generation sequencing. The sample selection is a crucial step in designing the genetic study of a complex disease and knowledge of the efficiency and limitations of the population-based and family-based designs can help making the appropriate choice.

The 9 contributions to the Group 5 of the Genetic Analysis Workshop 17 evaluated the population-based and family-based designs by comparing the results obtained with various methods applied on the mini-exome simulations. These simulations consisted of 200 replicates comprising unrelated individuals and 8 extended pedigrees with genotypes and various phenotypes. The methods tested for association with a population-based and/or a family-based design, tested for linkage with a family-based design or estimated heritability.

In this paper, we summarize the strength and weaknesses of both designs. While a population-based design seems more suitable to detect the effect of multiple rare variants, a family-based design can potentially enrich the sample in very rare variants, for which the effect would be concealed at the population level. However, as of today, the main limitation is still the expensive cost of next-generation sequencing.

Keywords

Genetic Analysis Workshop; 1000 genome project; next-generation sequencing; linkage; association; aggregation; familial relatedness; population stratification; heritability

INTRODUCTION

Complex diseases result from the interplay of multiple genetic and environmental factors. To locate and identify common genetic variants involved in complex diseases, many statistical methods based on genetic linkage and/or association have been proposed. In the last five years, genome-wide association studies have successfully identified hundreds of associations of common variants with complex diseases and traits [Manolio, 2010]. Yet, this success is partially shadowed by the small portion of heritability explained by these

[*]Correspondence to: Rémi Kazma, University of California San Francisco, Box 3110, 1450 3rd Street, CA 94143-3110. KazmaR@humgen.ucsf.edu Tel: +1 (415) 514 9793 Fax: +1 (415) 476 1356.

associations. One possible explanation for this missing heritability is that rare variants with moderate to strong effects might play a role in the inheritance of complex diseases [Bansal et al., 2010; Eichler et al., 2010; Maher, 2008]. With recent improvements in next-generation sequencing, testing this hypothesis is now at reach, but choosing the appropriate sampling design needs to be considered early in the study. Furthermore, sequencing a whole genome or exome is still expensive. So, comparing the power achieved while using population-based or family-based designs is essential to select the most cost-effective design.

To study the genetic basis of complex diseases, two broad types of sampling designs are often used: the population-based and the family-based designs. The population-based design consists in sampling affected and unaffected individuals who are unrelated, such as population cohorts or case-control samples. The family-based design consists in either ascertaining affected individuals with relatives, such as parents (trios), parents and siblings (nuclear families) or larger families (extended pedigrees) or in collecting unascertained related individuals (usually nuclear or extended pedigrees).

The Group 5 of the Genetic Analysis Workshop 17 (GAW 17) evaluated the population-based and family-based designs by comparing the results obtained with various methods applied on the mini-exome simulations. In this paper, we summarize the main topics discussed by the 9 papers of Group 5 and the strengths and weaknesses of both sampling designs.

GAW 17 DATA

The GAW 17 mini-exome simulations consisted of 200 replicates comprising 697 unrelated individuals and 697 members of 8 extended pedigrees. Genotypes of 24,487 variants located in 2305 genes were based on the pilot 3 study of the 1000 Genomes Project [Durbin et al., 2010]. Fully informative markers provided the Identity By Descent (IBD) for pairs of relatives at each gene, assuming no recombination within genes. Age, sex, population and genotypes of individuals were the same for the 200 replicates. The dichotomous disease phenotype, three quantitative traits (Q1, Q2, and Q4) and the smoking status were simulated in each replicate according to the simulation model described in [Almasy et al., 2011].

The 9 contributions to the group 5 used different outcome variables. The outcome studied was the dichotomous disease for 4 contributions [Fardo et al., 2011; Kazma et al., 2011; Lin et al., 2011; Liu and Thalamuthu, 2011] Q1 for 5 contributions [Kazma et al., 2011; Mahachie John et al., 2011; Saad et al., 2011; Shetty et al., 2011; Zhang et al., 2011b], Q2 for 1 contribution [Zhang et al., 2011a], and Q4 for 1 contribution [Shetty et al., 2011]. Two contributions [Fardo et al., 2011; Kazma et al., 2011] split the 8 pedigrees provided in the GAW 17 simulations into 194 trios without loss of individuals, whereas the 7 other contributions used the 8 extended pedigrees [Lin et al., 2011; Liu and Thalamuthu, 2011; Mahachie John et al., 2011; Saad et al., 2011; Shetty et al., 2011; Zhang et al., 2011a; Zhang et al., 2011b] (Table I).

METHODS

The 9 contributions to the Group 5 applied various methods with different purposes to analyze the 2 sampling designs available. Using the population-based design, 6 contributions tested for genetic association through a generalized linear model or a non-parametric test [Kazma et al., 2011; Lin et al., 2011; Liu and Thalamuthu, 2011; Mahachie John et al., 2011; Saad et al., 2011; Zhang et al., 2011b]. Using the family-based design, 7 contributions tested for genetic association and 2 contributions tested for genetic linkage [Kazma et al., 2011; Lin et al., 2011; Liu and Thalamuthu, 2011; Mahachie John et al., 2011; Saad et al., 2011; Zhang et al., 2011a; Zhang et al., 2011b]. In addition to single marker tests, all participants aggregated rare variants when testing for association with both sampling designs. Finally, Fardo et al., [2011] explored 3 methods combining the population-based and family-based designs to test for genetic association and Shetty et al. [2011] applied 3 methods to estimate heritability using the population-based or the family-based designs (Table I).

Association tests

Aggregation of rare variants—To improve the power of association tests with rare variants, several methods proposed to aggregate rare variants within a specific region, *e.g.* a gene [Han and Pan, 2010; Hoffmann et al., 2010; Li and Leal, 2008; Madsen and Browning, 2009; Morgenthaler and Thilly, 2007; Morris and Zeggini, 2010; Price et al., 2010]. For a comprehensive review of these methods see in this volume of the *Journal* [Dering et al., 2011]. In the Cohort Allelic Sum Test (CAST), all variants below a frequency threshold are aggregated into an indicator variable set 0 for individuals with no variant and 1 for those with at least 1 variant in the gene. A proportion score corresponding to the proportion of rare variant sites with a rare allele can also be used to aggregate rare variants in the CAST [Li and Leal, 2008]. The Combined Multivariate and Collapsing (CMC) method consists in aggregating rare variants below a frequency threshold and including this score with all common variants in a single model. Several contributions assessed whether aggregating rare variants improved the power to detect “causal genes” (*i.e.*, genes harboring causal variants), in particular when using a family-based design [Lin et al., 2011; Liu and Thalamuthu, 2011; Mahachie John et al., 2011; Saad et al., 2011; Zhang et al., 2011b].

Saad et al. [2011] and Zhang X. et al. [2011] compared the power and false positive rates of the single marker test and 5 aggregating methods applied to population-based and family-based association tests : the indicator and proportion scores used in CAST, the CMC, the Variable Threshold and the Weighted Sum methods [Li and Leal, 2008; Madsen and Browning, 2009; Price et al., 2010]. Setting the frequency threshold defining rare variants at 1 % or 5 % was also assessed and a modified version of CMC was proposed, where variants with a minor allele frequency (MAF) below 1 % and variants with a MAF between 1 % and 5 % are aggregated separately [Saad et al., 2011].

However, when multiple covariates need to be adjusted for, the CMC method often fails to converge and to provide parameter estimates [Kazma et al., 2011]. To deal with this issue, the Principal Component and Collapsing (PCC) method uses a Principal Component Analysis (PCA) to summarize the genetic information of aggregated rare variants and individual common variants in a gene. The first Principal Component (PC) is then used in a regression model to test for the effect of the gene [Kazma et al., 2011]. The PCC was compared to the Weighted Sum [Price et al., 2010] and to the Step-Up [Hoffmann et al., 2010] methods. An extension of PCC to analyze trios is also implemented.

A non-parametric association test, the Model-based Multifactorial Dimensionality Reduction (MB-MDR) [Calle et al., 2008], was recently extended to family-based designs: FAM-MDR [Cattaert et al., 2011]. Mahachie et al. [2011] compared the power and false positive rates of MB-MDR and FAM-MDR to their parametric counterparts, the Lasso penalized regression [Tibshirani, 1996] and the PBAT screening approach followed by the FBAT statistic [Van Steen et al., 2005]. In particular, they assessed whether aggregating rare variants with an indicator score could increase the power to detect association using a non-parametric test.

Familial relatedness—When testing for association using large pedigrees, affected and unaffected individuals are related and the correlation between their genotypes must be accounted for. Three methods to adjust for familial relatedness have been assessed by contributions to the Group 5.

The first method uses a Unified Mixed Model (UMM) correcting for population stratification and familial relatedness [Zhang et al., 2011a]. Population stratification is accounted for by incorporating in the model the first 10 PCs using all variants with a MAF above 10 % as a fixed effect variable, whereas familial relatedness is accounted for by incorporating the kinship matrix as a random effect variable. Zhang Q et al. [2011] evaluated two methods to derive the kinship matrix using either the pedigree information or the Identity By State (IBS) allele sharing matrix. The second method uses the kinship matrix derived from the pedigrees to adjust the logistic model with generalized estimating equations (GEE) [Liu and Thalamuthu, 2011].

Finally the third method uses the Modified Quasi-Likelihood Score test (MQLS) adapted from the Case-Control Quasi-Likelihood Score test (CC-QLS) [Bourgain et al., 2003]. Compared to the CC-QLS test, the MQLS test can improve power because it relies on the property that variants are enriched in affected individuals with affected relatives [Thornton and McPeek, 2007]. Lin et al. [2011] extended this method to analyze aggregated rare variants.

Combination of family-based and population-based designs—A valuable approach involves combining the population-based and family-based designs. Such a situation often arises when a family sample used to test for linkage is enriched with a sample of unrelated cases and controls to test for association. Fardo et al. [2011] examined the properties of 3 methods to combine family-based and population-based designs. The first approach uses simultaneously two samplings: the trios and the unrelated controls with the cases from the trios. After verifying that the estimators of the genetic effect obtained by each sampling designs are consistent, a weighted least square estimator is constructed and used to test for the genetic effect [Chen and Lin, 2008]. The second and third approaches use trios and unrelated cases and controls. In the second approach, unrelated individuals (cases, controls and parents of trios) are used to calculate PCs and correct for population stratification. Then a covariance between genotypic and phenotypic residuals is calculated and used as test statistic adjusting for within family correlation [Zhu et al., 2008]. In the third approach, a multivariate GEE-based score test [Lange et al., 2003] is calculated for the unrelated sample (cases-controls and parents) and for the related sample (offspring) separately and then both score tests are added. Adjustment for population stratification is done using a standard PCA for the unrelated sample and a TDT-like PCA for the related sample [Zhang et al., 2009].

Linkage tests

Two genetic linkage approaches were applied to the family-based design (Table I–B). In the first approach, a 2 degrees of freedom chi-square statistic compared the IBD distributions between pairs of cases and pairs of controls sampled from the 8 pedigrees [Liu and Thalamuthu, 2011]. In the second approach, two-points linkage LOD scores at the 9 genes containing causal variants were calculated using the Sequential Oligogenic Linkage Analysis Routines (SOLAR) [Almasy and Blangero, 1998] and the provided IBD scores [Zhang et al., 2011b].

Heritability estimation

Heritability was estimated for Q1 using the population-based (200 replicates) and the family-based (4 random replicates) designs, adjusting for age, sex, and the smoking status [Shetty et al., 2011] (Table I–C). In the family-based design, heritability was estimated using a polygenic mixed effect model applying the George-Elston transformation to normalize the distribution of residuals obtained after adjustment for age, sex, and smoking status [George and Elston, 1988]. In the population-based design, heritability was estimated as the proportion of variance described by all variants under an additive model using the Ordinary Least Square (OLS) [Yang et al., 2010] and the Restricted Maximum Likelihood (REML) methods.

RESULTS

Type I error rate and power across association methods

The type I error rates of the methods was assessed by 5 of the contributions [Fardo et al., 2011; Liu and Thalamuthu, 2011; Mahachie John et al., 2011; Saad et al., 2011; Zhang et al., 2011a]. Three of them assessed the type I error rate using the average over the 200 replicates of the false positive detections among all non-causative variants [Fardo et al., 2011], among all non-causative genes [Liu and Thalamuthu, 2011], or for each of 7 non-causative genes selected with characteristics (number of variants and MAF distributions) similar to the causative ones [Saad et al., 2011]. Mahachie John et al. [2011] used the family-wise error rate (FWER) and Zhang et al. [2011a] interpreted visually quantile-quantile plots.

When using the population-based design to test for association with the quantitative trait Q1, most of the methods aggregating rare variants had inflated false positive rates. But after adjusting on the first 5 PCs of the PCA of all common variants, most false positive rates had the expected nominal α value (5%) [Saad et al., 2011].

MB-MDR and FAM-MDR also had inflated false positive rates (FWER of 0.13 and 0.065, respectively), in particular for rare variants (MAF < 1 %) while the power to detect the association of Q1 with the 11 causal variants located on chromosome 4 was quite low. Moreover, when aggregating rare variants, the FWER increased drastically, rendering the power comparisons difficult. The elevated FWER observed can be attributed to a limited number of “problematic” rare variants [Mahachie John et al., 2011].

Whether using the pedigree-based kinship matrix or the IBS allele sharing matrix, adjusting for familial relatedness using the UMM decreased false positive rates and increased the power to detect the association of Q2 using a family-based design. However correcting for population stratification using the first 10 PCs had little impact on false positive and power rates of the association test with Q2 [Zhang et al., 2011a].

When adjusting for familial relatedness using GEE, the test of association with the disease phenotype using a family-based design had an inflated false positive rate (on average 9.6% at α = 5 %) [Liu and Thalamuthu, 2011].

Population-based versus family-based methods

Most methods to test for association, whether using a population-based or a family-based design, had a high power to detect the effects of the genes *FLT1* and *KDR* on Q1. With a population based-design, the power of the single marker was 100 % and 74 % to detect association of Q1 with any variant of *FLT1* and *KDR*, respectively, at α = 5 %. With a family-based design the power to detect association of Q1 with *FLT1* and *KDR* was 95 % and 61 %, respectively. When aggregating rare variants in the population-based design, the power to detect association of Q1 with *FLT1* was maintained except when rare variants are defined with a lower threshold (1 % versus 5 %). However, when using the family-based design, aggregating rare variants in *FLT1* diminished substantially the power [Saad et al., 2011]. The *FLT1* and *KDR* genes harbor 13 and 4 causal variants among 35 and 16 variants, respectively. Each of those 2 genes has 1 common causal variant with a population MAF of

6.67 % and 16.50 % and an Odds-Ratio (OR) of 1.92 and 1.15 (association with Q1), respectively [Almasy et al., 2011].

Interestingly, association tests with the family-based design had a high power to detect the effects of *VEGFA* and *VEGFC* on Q1 (100 % at α = 0.01 %) [Saad et al., 2011; Zhang et al., 2011b]. The *VEGFA* gene harbors 1 causal variant (C6S2981) with a population MAF of 0.22 % and an OR of 3.34 among 6 variants, whereas *VEGFC* harbors only 1 causal variant (C4S4935) with a population MAF of 0.017 % and an OR of 3.89. However, in the family dataset the MAFs of the causal variants in *VEGFA* (C6S2981) and *VEGFC* (C4S4935) are 3.3 % and 2.2 %, respectively.

In some cases, aggregating rare variants proved to be detrimental in particular when common variants are not included in the model, such as in the CAST. Conversely, aggregating rare variants improved the power to detect some other genes with causal rare variants, but no single method did systematically better than the others. Analyzing Q1 using the PCC approach to summarize common variants and the aggregated rare variants gave results similar to a Weighted Sum or to the Step-Up method with the population-based design. Using the PCC approach with the family-based design, *VEGFA* and *FLT1* were the 2 genes with the strongest association, but did not reach the significance threshold after correcting for multiple testing [Kazma et al., 2011].

When adjusting for familial relatedness using the MQLS method, the power to detect the association of Q1 with *VEGFC* and *VEGFA* were respectively 99 % and 94.5 % at $\alpha = 1.56 \times 10^{-5}$. When collapsing rare variants with an indicator variable at a threshold of 1 %, the power decreased drastically for the genes *SIRT1* and *VLDLR*, but it increased slightly for the genes *SREBF1*, *PIK3R3*, *PLAT* and *FLT4* [Lin et al., 2011].

Combination of population-based and family-based designs

While the false positive rate was well controlled, the power of the 3 association tests combining population-based and family-based designs were very low for rare variants (MAF < 5 %). Averaging over all causal SNPs, the 3 tests had a power to detect association of the disease phenotype with rare variants lower than the power to detect it with common variants. However, high powers were observed for a few rare variants in particular when using the multivariate GEE-based score test [Zhang et al., 2009]. For this test, the power to detect the association of the disease phenotype with causal rare variants in *VEGFA* and *VEGFC* were 93.5 % and 80.5 % respectively, at α = 5%, but it also detected other causal rare variants in *SIRT1*, *VLDLR*, *KDR* and *PIK3C2B* with power rates between 73 % and 40 %.

Type I error rate and power across linkage methods

The linkage test based on the differences of IBD distributions between pairs of cases and pairs of controls suffered from an inflated false positive rate (on average 11.2 % at α = 5 %) [Liu and Thalamuthu, 2011]. The two-point linkage analysis detected a strong linkage (LOD score ≥ 3) for *VEGFA* and *VEGFC* in, respectively, 55 % and 63 % of the replicates. However, the false positive rate of this method was not evaluated [Zhang et al., 2011b].

Heritability estimation

The mean heritability estimate for Q1 was 0.65 using the family-based design and 0.53 using the population-based design (calibrated REML). Heritability estimates using the family-based design are reasonable but slightly overestimating the true heritability used to simulate the data (0.58 for Q1). Conversely, the population-based methods are not adapted to estimate heritability with data containing rare variants.

DISCUSSION

The population-based and family-based designs detect different causal genes with rare variants. Association tests using the population-based design detected the association of *FLT1* and *KDR* with Q1, even when using the single marker test. Indeed, both genes have a causal common variant and multiple causal rare variants. Aggregating rare variants improved the power to detect association with some genes, such as *KDR*, when the threshold defining rare variants was not too low and when common variants were included in the final model [Kazma et al., 2011; Saad et al., 2011; Zhang et al., 2011b]. However when looking at each causal gene, there are no particular genetic models (number and MAFs of causal variants) which seem to have a better power with any of the methods [Saad et al., 2011].

The PCC method provides a quick and reliable alternative to consider common variants and aggregated rare variants when multiple adjustments are done [Kazma et al., 2011]. Nonetheless, further assessments of the PCC method are needed. In particular the number of PCs required in the regression model should be evaluated.

Several contributions identified inflated type I error [Liu and Thalamuthu, 2011; Mahachie John et al., 2011; Saad et al., 2011; Zhang et al., 2011a]. Adjusting for population stratification corrected for the inflated type I error when testing for the association with Q1 using a population-based design [Saad et al., 2011], but not when testing for the association with Q2 using a family-based design [Zhang et al., 2011a].

The MB-MDR and FAM-MDR methods are flexible non-parametric tests that can handle different types of genetic models including epistasis and gene-environment interactions with population-based and family-based designs. With the GAW 17 simulation, those methods had severely inflated false positive rates with low power to detect association with rare variants [Mahachie John et al., 2011]. However, comparison of the false positive rates of MB-MDR and FAM-MDR to the false positive rates of the other methods in Group 5 is not possible. For MB-MDR and FAM-MDR, false positive rates were considered family-wise (FWER), i.e. the proportion of datasets for which at least one non-causal variant (or non-causal gene) has been declared significant. In contrast, for other methods, false positive rates were considered for each non-causal variant (or non-causal gene) separately, not accounting for multiple testing. However, Mahachie John et al. [2011] observed that removing a few spurious rare variants decreased the FWER. This issue has also been reported in other GAW 17 groups [Luedtke et al., 2011].

Association tests using the family-based design detected *FLT1* and *KDR* with a lower power than those using the population-based design, as well as the causal rare variants of *VEGFA* and *VEGFC* [Lin et al., 2011; Liu and Thalamuthu, 2011; Saad et al., 2011; Zhang et al., 2011b]. However, caution should be taken to avoid inflation of false positive rates when cases and controls are related (*e.g.*, using extended pedigrees for a standard association test). Using the IBS allele sharing matrix to adjust for familial relatedness requires no information about the pedigree structure but has computational burdens. Conversely, if the pedigree structure is available, the kinship matrix has been shown to be a good alternative [Zhang et al., 2011a]. Assessing the empirical p-value using the classical permutation procedure, where case-control statuses (or continuous trait values) are randomly reassigned to genotypes, might be invalid in association tests with extended pedigree [Bourgain and Génin, 2005]. Consequently, methods that account for the correlation between individuals should be preferred when the pedigree structure is available.

Linkage tests using the family-based design detected *VEGFA* and *VEGFC* too, but did not detect *FLT1*, nor *KDR* [Liu and Thalamuthu, 2011; Zhang et al., 2011b]. The linkage signal of *VEGFA* is only observed in family 7, whereas the linkage signal of *VEGFC* is only observed in families 2 and 7 [Reference of the linkage group summary paper?].

Although, methods combining both designs had an overall power to detect causal rare variants quite low, such approaches could be a good compromise to take advantage of both designs [Fardo et al., 2011]. Incorporating aggregation of rare variants might increase further the power to detect association with rare variants.

In a population sample “large enough”, all the rare variants present in the population will be sampled. Of course, the larger the sample, the lower the frequency of rare variants sampled. Conversely, in an extended pedigree, rare variants are sampled through a limited number of founders and are then transmitted (or not) to their offspring. Therefore, family samples can be enriched with one or multiple rare variants. In the GAW 17 simulations, we observed this “founder effect” for the causal rare variants of the genes *VEGFA* (C6S2981) and *VEGFC* (C4S4935) which had different frequencies between the population and the family datasets. The fact that the genotypes were fixed among the replicates limited the possibility for other rare variants to be selected by founders and transmitted to the offspring. This simulation procedure explains why the power to detect the 2 causal rare variants of *VEGFA* and *VEGFC* is very high (often 100%) and why the power of other variants, which never occurred in any founder, is very low. Several rare variants in *KDR* and *FLT1* genes had lower MAFs in the family datasets than in the population datasets. In consequence, the power for detecting association with these 2 genes using family datasets was reduced.

To take advantage of this “founder effect” in the family-based design, it will be important to focus on developing clever ascertainment tools to select families, such as the sampling of extreme phenotypes [Guey et al., 2011]. Comparing different family-based designs (trios, nuclear families, various sizes of pedigrees) should be further investigated with more realistic settings, such as generating missing data, or estimating IBD.

In conclusion, while a population-based design seems more suitable to detect the effect of multiple rare variants, a family-based design can potentially enrich the sample in very rare variants, for which the effect would be concealed at the population level. However, as of today, the main limitation is still the expensive cost of next-generation sequencing.

Acknowledgments

The authors thank all the research teams of the GAW 17 Group 5 for the dynamic and fruitful discussions. The primary contributing authors include: T. Cattaert, D.W. Fardo, T.J. Hoffmann, B. Li, P. Lin, T. Liu, J.M. Mahachie John, H. Qin, M. Saad, P.B. Shetty, Q. Zhang, X. Zhang. The Genetic Analysis Workshop 17 was supported by NIH grant R01 GM031575. Rémi Kazma is supported by NIH grant R25CA112355. Julia N. Bailey is supported by NIH grant R01NS055057

REFERENCES

Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet. 1998; 62(5):1198–211. [PubMed: 9545414]

Almasy L, Dyer TD, Peralta JM, Kent JW Jr, Charlesworth JC, Curran JE, Blangero J. Genetic Analysis Workshop 17 Mini-Exome Simulation. BMC Proc. 2011

Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet. 2010; 11(11):773–85. [PubMed: 20940738]

Bourgain C, Génin E. Complex trait mapping in isolated populations: Are specific statistical methods required? Eur J Hum Genet. 2005; 13(6):698–706. [PubMed: 15785775]

Bourgain C, Hoffjan S, Nicolae R, Newman D, Steiner L, Walker K, Reynolds R, Ober C, McPeek MS. Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. Am J Hum Genet. 2003; 73(3):612–26. [PubMed: 12929084]

Calle ML, Urrea V, Vellalta G, Malats N, Steen KV. Improving strategies for detecting genetic patterns of disease susceptibility in association studies. Stat Med. 2008; 27(30):6532–46. [PubMed: 18837071]

Cattaert T, Calle ML, Dudek SM, Mahachie John JM, Van Lishout F, Urrea V, Ritchie MD, Van Steen K. Model-based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise. Ann Hum Genet. 2011; 75(1):78–89. [PubMed: 21158747]

Chen YH, Lin HW. Simple association analysis combining data from trios/sibships and unrelated controls. Genet Epidemiol. 2008; 32(6):520–7. [PubMed: 18348203]

Dering C, Hemmelmann C, Pugh E, A. Z. Statistical analysis of rare sequence variants: An overview of collapsing methods. Genetic Epidemiology. 2011

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010; 467(7319): 1061–73. [PubMed: 20981092]

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010; 11(6):446–50. [PubMed: 20479774]

Fardo DW, Druen AR, Liu J, Mirea L, Infante-Rivard C, Breheny P. An exploration and comparison of methods for combining population- and family-based genetic association using the Genetic Analysis Workshop 17 mini-exome. BMC Proc. 2011

George VT, Elston RC. Generalized modulus power transformations. Commun Statist - Theory Meth. 1988; 17:2933–2952.

Guey LT, Kravic J, Melander O, Burtt NP, Laramie JM, Lyssenko V, Jonsson A, Lindholm E, Tuomi T, Isomaa B, et al. Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. Genet Epidemiol. 2011

Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered. 2010; 70(1):42–54. [PubMed: 20413981]

Hoffmann TJ, Marini NJ, Witte JS. Comprehensive approach to analyzing rare genetic variants. PLoS One. 2010; 5(11):e13584. [PubMed: 21072163]

Kazma R, Hoffmann TJ, Witte JS. On the use of principal components to aggregate rare variants in case-control and family-based association studies in the presence of multiple covariates. BMC Proc. 2011

Lange C, Silverman EK, Xu X, Weiss ST, Laird NM. A multivariate family-based association test using generalized estimating equations: FBAT-GEE. Biostatistics. 2003; 4(2):195–206. [PubMed: 12925516]

Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83(3):311–21. [PubMed: 18691683]

Lin P, Hamm M, Hartz S, Zhang Z, Rice JP. Challenges and directions: An analysis of GAW17 data by collapsing rare variants within family data. BMC Proc. 2011

Liu T, Thalamuthu A. Identity by descent and association analysis of dichotomous traits based on large pedigrees. BMC Proc. 2011

Luedtke A, Powers S, Petersen A, Sitarik A, Bekmetjev A, Tintle N. Evaluating methods for the analysis of rare variants in sequence data. BMC Proc. 2011

Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009; 5(2):e1000384. [PubMed: 19214210]

Mahachie John JM, Cattaert T, De Lobel L, Van Lishout F, Empain A, Van Steen K. Comparison of genetic association strategies in the presence of rare alleles. BMC Proc. 2011

Maher B. Personal genomes: The case of the missing heritability. Nature. 2008; 456(7218):18–21. [PubMed: 18987709]

Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010; 363(2):166–76. [PubMed: 20647212]

Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res. 2007; 615(1–2):28–56. [PubMed: 17101154]

Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol. 2010; 34(2):188–93. [PubMed: 19810025]

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010; 86(6):832–8. [PubMed: 20471002]

Saad M, Saint Pierre A, Bohossian N, Macé M, Martinez M. A comparative study of statistical methods for detecting association with rare variants in exome-resequencing data. BMC Proc. 2011

Shetty PB, Qin H, Namkung J, Elston RC, Zhu X. Estimating heritability using family and unrelated data. BMC Proc. 2011

Thornton T, McPeek MS. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. Am J Hum Genet. 2007; 81(2):321–37. [PubMed: 17668381]

Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Stat Soc B. 1996; 58(1):267–288.

Van Steen K, McQueen MB, Herbert A, Raby B, Lyon H, Demeo DL, Murphy A, Su J, Datta S, Rosenow C, et al. Genomic screening and replication using the same data set in family-based association testing. Nat Genet. 2005; 37(7):683–91. [PubMed: 15937480]

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010; 42(7):565–9. [PubMed: 20562875]

Zhang L, Pei YF, Li J, Papasian CJ, Deng HW. Univariate/multivariate genome-wide association scans using data from families and unrelated samples. PLoS One. 2009; 4(8):e6502. [PubMed: 19652719]

Zhang Q, Doyoung C, Aldi K, Borecki II, Province MA. Methods for adjusting population structure and familial relatedness in association test for collective effect of multiple rare variants on quantitative traits. BMC Proc. 2011a

Zhang X, He H, Ding L, Baye TM, Kurowski BG, Martin LJ. Family and population based designs identify different rare causal variants. BMC Proc. 2011b

Zhu X, Li S, Cooper RS, Elston RC. A unified association analysis approach for family and unrelated samples correcting for stratification. Am J Hum Genet. 2008; 82(2):352–65. [PubMed: 18252216]

TABLE I

Summary of the data and methods used by Group 5 contributions

A. ASSOCIATION TESTS

Sampling design	Method	Aggregation methods[a]	Adjustment variables	Contribution (outcome)
Population Based	Logistic/Linear regression	SM, CMC, CASTi, CASTp, VT, WS	Population	Saad M et al. (Q1)
		SM, CASTi, CASTp, WS	Age, sex, smoking Population	Zhang X et al. (Q1)
		PCC, WS, Step-Up	Age, sex, smoking Population	Kazma R et al. (Disease, Q1)
		SM, CASTi	Population	Lin P et al. (Disease)
	MB-MDR	SM, CASTi		Mahachie J et al. (Q1)
	Lasso penalized regression	SM, CASTi		Mahachie J et al. (Q1)
	Logistic regression	SM, CASTp	Age, smoking status	Liu T et al. (Disease)
Family Based[b]	Measured genotype test (QTDT)	SM, CMC, CASTi, CASTp, WS		Saad M et al. (Q1)
		SM, CASTi	Age, sex, smoking Population	Zhang X et al. (Q1)
	Score test (FBAT)	PCC	Age, sex, smoking	Kazma R et al. (Disease, Q1)
	FAM-MDR	SM, CASTi		Mahachie J et al. (Q1)
	PBAT screening + FBAT	SM		Mahachie J et al. (Q1)
	Logistic regression| + GEE	SM, CASTp		Liu T et al. (Disease)
	Modified quasi-likelihood score test	SM, CASTi	Population	Lin P et al. (Disease)
	Unified Mixed Model	CASTi	Population	Zhang Q et al. (Q2)
Population and Family Based[b]	Weighted least square estimator of combined effects	SM		Fardo DW et al. (Disease)
	Genotype-Phenotype covariance	SM	Population	
	FBAT-GEE	SM	Population	

B. LINKAGE TESTS

Sampling design[a]	Model	Adjustment variables	Contribution (outcome)
Family Based	IBD analysis		Liu T et al. (Disease)
	Two point linkage (SOLAR)	Age, sex, smoking Population	Zhang X et al. (Q1)

C. HERITABILITY ESTIMATION

Sampling design[a]	Model	Adjustment variables	Contribution (outcome)
Population Based	Linear mixed model (OLS / REML)	Age, sex, smoking	Shetty PB et al. (Q1)
Family Based	Polygenic mixed effect (S.A.G.E. ASSOC)	Age, sex, smoking	Shetty PB et al. (Q1, Q4)

[a] SM: Single marker test; CASTi: Cohort Allelic Sum Test (indicator variable); CASTp: Cohort Allelic Sum Test (proportion variable); CMC: Combined Multivariate and Collapsing; VT: Variable Threshold; WS: Weighted Sum

[b] Kazma et al. and Fardo et al. split the 8 extended pedigrees into 194 trios IBD: Identity By Descent; IBS: Identity By State; MB-MDR: Model-Based Multifactor Dimensionality Reduction; FAM-MDR: Familial MB-MDR; MQLS: Maximum Quasi-Likelihood Score; OLS: Ordinary Least

Square; PC: Principal Component; PCC: Principal Component and Collapsing; QTDT: Quantitative Transmission Disequilibrium Test; REML Restricted Maximum Likelihood; SOLAR: Sequential Oligogenic Linkage Analysis Routine