Published in final edited form as:

Genet Epidemiol. 2011 ; 35(Suppl 1): S22–S28. doi:10.1002/gepi.20645.

Quality Control Issues and the Identification of Rare Functional Variants with Next-Generation Sequencing Data

Claudia Hemmelmann¹, E. Warwick Daw², and Alexander F. Wilson³

¹Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Germany

²Division of Statistical Genomics, Washington University School of Medicine, St. Louis, MO

³Genometrics Section, Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD

Abstract

Next-generation sequencing of large numbers of individuals presents challenges in data preparation, quality control, and statistical analysis because of the rarity of the variants. The Genetic Analysis Workshop 17 (GAW17) data provide an opportunity to survey existing methods and compare these methods with novel ones. Specifically, the GAW17 Group 2 contributors investigate existing and newly proposed methods and study design strategies to identify rare variants, predict functional variants, and/or examine quality control. We introduce the eight Group 2 papers, summarize their approaches, and discuss their strengths and weaknesses. For these investigations, some groups used only the genotype data, whereas others also used the simulated phenotype data. Although the eight Group 2 contributions covered a wide variety of topics under the general idea of identifying rare variants, they can be grouped into three broad categories according to their common research interests: functionality of variants and quality control issues, family-based analyses, and association analyses of unrelated individuals. The aims of the first subgroup were quite different. These were population structure analyses that used rare variants to predict functionality and examine the accuracy of genotype calls. The aims of the family-based analyses were to select which families should be sequenced and to identify high-risk pedigrees; the aim of the association analyses was to identify variants or genes with regression-based methods. However, power to detect associations was low in all three association studies. Thus this work shows opportunities for incorporating rare variants into the genetic and statistical analyses of common diseases.

Keywords

1000 Genomes Project; association; collection of rare variants; family data; next-generation sequencing; regression; quality control

Introduction

The introduction of next-generation sequencing technologies has made the sequencing of genomes of many individuals possible at a reasonable cost and has enabled the investigation of the common disease/rare variant hypothesis [Fitze et al., 2002; Bodmer and Bonilla,

Corresponding author: Dr. Claudia Hemmelmann, Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Maria-Goeppert-Str. 1, 23562 Lübeck, Germany, Phone: +49 (451) 500-2789, Fax: +49 (451) 500-2999, hemmelmann@imbs.uni-luebeck.de.

2008]. One underlying idea of the Genetic Analysis Workshop 17 (GAW17) Group 2 contributions is that rare variants have the greatest potential to be functional in human disease and therefore are more likely to have a larger effect than common variants. Because of the rarity of the variants, the appropriateness of standard methods for analysis in genome-wide association studies, such as analysis of population structure and statistical methods, have to be revised.

GAW17 provided exome sequencing data from the 1000 Genomes Project with 24,487 variants within 3,205 genes from 697 individuals in 7 populations. Furthermore, 200 replicates of phenotypic traits for two study designs (unrelated individuals and large families) were simulated. In most of the Group 2 contributions the quantitative risk factors Q1, Q2, and Q4 were considered; however, Q4 was not associated with any genotype in the data set. Only one work group used case-control status. Further details of the simulated data can be found elsewhere [Almasy et al., 2011]. Three of the eight work groups used all 200 replicates, one group used 10 replicates, and one group used only 1 replicate; three groups did not use any of the simulated phenotypes. Furthermore, some work groups analyzed all genes, whereas other groups selected a subset of genes or variants.

Although the eight contributions covered a wide variety of topics under the general topic of identifying rare variants, they can be grouped into three categories according to the common research themes: (1) functionality of variants and quality control issues, (2) family-based analyses, and (3) association analyses of unrelated individuals. An overview of the eight contributions is given in Table I.

We should note that there was no consistent definition of the term *rare variants* in these contributions. Thus different definitions of *rare* were applied by the contributors. Four work groups defined rare variants as those with minor allele frequency (MAF) < 5% [Baye et al., 2011; Ding et al., 2011; Guo et al., 2011; Jaffe et al., 2011]; one group used MAF < 1% [Sung et al., 2011], one used MAF < 0.5% [Cai et al., 2011], and one work group did not specify any threshold [Stram, 2011]. Gagnon et al. [2011] did not use MAF in their study design. Furthermore, most of the contributors analyzed rare and common variants together. Only Baye et al. [2011] compared the results of analyzing rare versus common variants.

Functionality of Variants and Quality Control Issues

Baye et al. [2011] examined an extensive population structure analysis using common and rare functional variants. They used principal components analysis to reduce variable dimension, Structure to assess ancestry, and discriminant analysis to predict population membership [Pritchard et al., 2000; Krzanowsky, 2003]. Seven of the 11 populations were included: Caucasians from the United States with northern and western European ancestry; Yoruba from Ibadan, Nigeria; Japanese from Tokyo; Han Chinese from Beijing; Chinese in metropolitan Denver, Colorado; Luhya in Webuye, Kenya; and Tuscans in Italy. Furthermore, Baye and colleagues investigated two subsets of the data. The first subset included common functional variants (MAF $\geq$ 5%), and the second subset included only rare functional variants (MAF < 5%).

Baye et al. [2011] detected a clear distinction between the three geographic origins (Europe, Asia, and Africa) with principal components analysis using common functional variants but not between the seven different populations. The first principal component distinguished between Africans and non-Africans, and the second principal component distinguished between Europeans and non-Europeans. Altogether, this analysis required 388 principal components to account for 90% of the variation or population structure. However, the analysis based on only rare variants required 532 principal components to account for 90% of the variation, and Africans and non-Africans were distinguished only on the second

principal component. Furthermore, Baye and colleagues observed substantial variability in the ancestral genetic background based on rare variants compared with common variants. An individual with primary European ancestry in a population sample of Yoruba was identified using only rare variants. However, this individual had less inferred European ancestry when only common variants were considered. Ninety-eight percent of the individuals were assigned to their correct population using 400 common SNPs. In contrast, 1,000 rare functional variants were needed to reach the same level of individual assignment to their correct ancestry.

Baye et al. [2011] concluded that the number of principal components required to account for population structure varied with MAF. They showed that as MAF decreased, the number of SNPs required for population assignment increased. However, variants with lower MAF were less heterozygous and less informative and thus had less discriminatory power. Furthermore, including rare variants to detect outliers was effective, even among geographically close populations.

Jaffe et al. [2011] compared the findings of four nucleotide- and/or amino acid-based algorithms aimed at predicting the effect of observed nonsynonymous sequence variation. They predicted functionality for nonsynonymous SNPs using the nucleotide version of Sorting Intolerant from Tolerant (SIFT) [Ng and Henikoff, 2003], PolyPhen-2 [Sunyaev et al., 2001; Adzhubei et al., 2010], VarioWatch [Chen et al., 2008], and Multivariate Analysis of Protein Polymorphism (MAPP) [Stone and Sidow, 2005]. These four functional analysis programs fell into two categories: nucleotide- and amino acid-based. Some of the programs classified their predictions into levels of functional threats, and others provided probabilities. The predictions were dichotomized as tolerated or deleterious. Thus very high level and high-level threats were classified as deleterious, and low- and medium-level threats were classified as tolerated for VarioWatch. For MAPP, the severe and moderate levels were classified as deleterious, and the minor level was classified as tolerated. The output levels of possibly damaging and probably damaging were classified as deleterious, and the benign level was classified as tolerated for PolyPhen-2. In contrast, SIFT provided binary classification of variants as either tolerated or deleterious. Further details of the four programs can be found elsewhere [Jaffe et al., 2011]. The analyses were restricted to 101 autosomal genes with 3,781 nonsynonymous SNPs that caused a missense mutation in the subsequent amino acid.

Jaffe et al. [2011] obtained only moderate levels of pairwise agreement for the four programs (59–71%). MAPP and PolyPhen-2 showed the highest agreement with 71%; the lowest agreement (59%) was observed between VarioWatch and MAPP. To compare the results to a gold standard, Jaffe and colleagues examined the entire GAW17 data set to identify variants that were previously analyzed in experimental models providing true functional results. For this, the Online Mendelian Inheritance in Man (OMIM) database [<http://omim.org>] was searched for functional variants. Jaffe and co-workers found 13 functional variants using this method. In addition, they identified two variants (in *BRCA2* and *MFTRR*) by searching the literature. This approach was necessary because the simulated GAW17 data do not present true biological functionality.

Jaffe et al. [2011] demonstrated that 13 of these 15 variants result in a loss of function, but only 6 of these were predicted to have a deleterious effect on protein function by all four programs. PolyPhen-2 had the greatest accuracy and predicted the function of 11 (out of the 15 variants) correctly, whereas VarioWatch and SIFT predicted 10 variants correctly, and MAPP predicted 9 variants correctly.

The four programs used by Jaffe and colleagues have different methodologies and discordant results. Thus the investigators do not recommend using any program as a true substitute for functional assays. For example, a less conservative approach such as PolyPhen-2 should be used if the aim is the selection of variants for further statistical analyses.

Stram's [2011] aim was to evaluate the quality of the GAW17 data by examining the accuracy of its genotype calls. Stram compared the GAW17 genotype calls to the calls in HapMap III, release 2, for the same individuals. Stram assumed that the HapMap genotype calls were reliable enough to determine the gold standard. In his study 616 of the 697 individuals and 3,403 of the 24,487 SNPs in the GAW17 data set were also found in HapMap. Thus there were 2,096,248 GAW17 SNPs for which HapMap genotype calls had already been examined on the same individuals. Stram established the concordance rate between the GAW17 data and the HapMap genotype calls in order to compare these SNP genotype calls for each individual. Only 65.4% of the genotype calls were concordant with HapMap. The concordance rate per individual ranged from 51.8% to 73.6%. However, the comparison between the GAW17 genotypes and the HapMap calls was problematic because missing or low-quality genotype calls were filled in so that all GAW17 participants had complete data, and thus it was impossible to differentiate between actual genotype calls made from the sequence data and filled-in, imputed genotypes in the GAW17 data [Almasy et al., 2011].

Normally, genotype quality scores are considered in any meaningful HapMap concordance analysis. The GAW17 data set did not provide quality scores. Therefore Stram [2011] used the 1000 Genomes Project pilot3 release sequence alignment data of 90 European-descent samples of chromosome 1 to call genotypes using the Broad Institute's Genotype Analysis Toolkit (GATK) software [McKenna et al., 2010]. A quality score, which was a Bayesian function of relevant sequence reads, sequence-read quality scores, and sequence-read mapping quality scores, was obtained for each genotype call made by GATK. Under the assumption that these quality scores were accurate, $n(1 - p)$ genotypes were expected to be HapMap concordant for n genotypes with $1 - p$ confidence. If only the calls with 99% confidence or greater were considered, just 88.5% overall HapMap concordance would be found (and not the expected 99%). Hence Stram [2011] concluded that the resulting HapMap discordance seemed inevitable if the GAW17 data set was filtered on only these inaccurate quality scores. Thus the inaccurate prior probabilities were the prime suspect in the discordance issues. A possible explanation for this is that Bayesian genotype callers such as GATK are conditioned on a multisample prior probability. This means that allele frequencies are estimated using all available samples, and a low sample size in this prior probability estimate would yield high variance in the estimate. The GATK inaccurate quality scores appear to be at least partly explained by such variance. For those loci for which 10 or more samples were included in the prior probability calculation, 95.1% concordance could be observed vs. 54.6% for regions with less than 10 samples.

Finally, it seems that the genotype calls and the quality scores of a significant number of samples in the GAW17 data set were incorrect. However, the 1000 Genomes Project is a work in progress. Further work is required to investigate the correctness of the quality scores, especially for rare variants.

Family-Based Analyses

The original shared genomic segment (SGS) method is an approach for localizing predisposition genes for a trait segregation using extended pedigrees [Thomas et al., 2008]. In the SGS method long runs of loci in case subjects that share a common allele identically

by state (IBS) are used to localize hypothesized predisposition genes. The distribution under the null hypothesis of no genetic effect is estimated by simulation. For many common diseases the problem is that not all case subjects have to be attributable to the same underlying causal variants. In this case an SGS method based on sharing of all case subjects may have low power. Therefore Cai et al. [2011] introduced a new pairwise shared genomic segment (pSGS) method that examines the sharing among all pairs of case subjects. The test statistic for the pSGS method is defined as

$$SH^{\text{case}}(k) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N R_{ij}, \quad (1)$$

where R_{ij} is the run length (i.e., the number of SNPs in a row sharing IBS), obtained from a pairwise comparison of case i with case j at locus k . This test statistic is similar to that for a case-control haplotype-sharing method [Nolte et al., 2007]. Cai and colleagues used a significance level of $\alpha = 0.001$ and assessed the test statistic against an empirical distribution of at least 10,000 null test statistics.

Cai et al. [2011] identified high-risk pedigrees (defined as those with at least 15 total meioses between case subjects and a statistical excess of disease [$p < 0.01$]) over all 200 replicates. Note that the genome-wide probability of sharing identically by descent by chance across 15 meioses was approximately 0.05. Cai and colleagues then computed the pSGS test statistic for the identified high-risk pedigrees.

In Cai's study, 18 high-risk pedigrees were identified across all 200 replicates, and the number of related case subjects ranged from 21 to 24. At least one region containing a true and polymorphic causal variant was identified in 13 of the 18 pedigree replicates. Eighteen rare causal variants were polymorphic within each of these 18 extended high-risk pedigrees. Four of these 18 rare causal variants were identified by the pSGS method. All four true positives were identified in one pedigree, two in three pedigrees, and one in nine pedigrees.

Cai et al.'s [2011] investigation has two limitations: the small sample size and the fact that all replicates were generated from the same genotype data. This means that at most one family could be identified per replicate, and it always had the same structure. This new pSGS method has potential; however, it needs to be validated with another data set. In addition, further analyses concerning the balance of true- and false-positive findings are necessary.

Gagnon et al. [2011] also analyzed the family data by identifying rare variants with large effects on trait variance using oligogenic segregation analysis. The aim of this study was to develop a cost-efficient study design strategy to select which family should be sequenced.

Gagnon et al. [2011] estimated the mean number of quantitative trait loci (QTLs) that explained a proportion of the variance of the trait by applying an oligogenic linear model for each quantitative trait. They modeled the trait Y as

$$Y = \mu + X\beta + \sum_{i=1}^k Q_i \alpha_i + e, \quad (2)$$

where μ is the overall mean, β is the vector of covariate effects, α is the vector of the i th QTL effect, and e are normally distributed residuals. Sex, Age, and Smoking status were included as covariates. The number k of QTLs is an appropriate estimable parameter using the reversible jump Markov chain Monte Carlo algorithm implemented in Loki [Heath, 1997]. The basic idea of the additional analysis steps is that families with the additional estimated mean number of QTLs compared to that in other families are more likely to segregate functional rare variants that are not segregating in the other families. Gagnon and colleagues restricted the association tests of rare variants with the traits to regions with LOD ≥ 0.6 , and they considered only replicate 1 of the simulated GAW17 phenotypes for these analyses.

As a result, Family 7, compared to the other families, was estimated to harbor two more QTLs for trait Q1 and one additional QTL for trait Q2. They then investigated the influence of rare variants on the additional loci in Q1 and Q2. In this step, they detected 505 genes with LOD ≥ 0.6 for trait Q1 and 315 genes for Q2. Q4 was correctly identified as not being influenced by QTLs. They further analyzed the regions surrounding these genes, including 216 variants for Q1 and 85 variants for Q2. Seven variants were identified for Q1 and six for Q2. However, only two variants for Q1 and none for Q2 were true positives. Note that the genes with true variants or genes in high linkage disequilibrium with one true causal variant were genes with LOD scores greater than 4 in all families, whereas the false-positive genes had LOD scores less than 2. Thus this strategy is able to detect linkage to rare variants and helps to reduce sequencing costs. Type I error was much better than that observed for approaches used in unrelated subjects.

Interestingly, both groups that investigated family data [Cai et al., 2011; Gagnon et al., 2011] identified rare variants within Family 7, although their approaches were different. Given that the two groups considered different phenotypes, it was not surprising that different variants were identified. The only variant identified by both groups was a private mutation, C4S4935. Furthermore, the aims of both groups were quite different. Gagnon et al. [2011] was aiming to develop a strategy to select a family that should be sequenced, whereas Cai et al. [2011] intended to identify high-risk families.

Association Analyses of Unrelated Individuals

Ding et al. [2011] introduced a novel nonparametric Bayes-based clustering approach for quantitative traits. In this approach, the quantitative trait is constructed per individual with the help of a linear regression model in which the regression coefficients are modeled using a Dirichlet process. Variants with the same regression coefficient are clustered, and the SNPs in the largest cluster are considered not associated, whereas SNPs outside that cluster are rated as potential risk variants. In their study, Ding and colleagues incorporated both common and rare variants and investigated quantitative trait Q1. Thus the corresponding regression model was

$$y_i = z_i' \gamma + \sum_{j=1}^J x_{ij} \beta_j + \varepsilon_i, \quad (3)$$

where y_i is the quantitative trait of the i th individual, z_i is the vector of the corresponding covariates with the regression coefficient γ , x_{ij} denotes the genotype of the j th SNP in individual i with regard to regression coefficients β_j , which are modeled using a Dirichlet process, and ε_i denotes an error for individual i that is normally distributed with mean 0 and variance σ^2 .

Ding et al. [2011] used a truncated Dirichlet process to approximate Dirichlet process priors and developed a block Gibbs sampling method for Dirichlet process models. In each iteration of the Gibbs sampler, a clustering structure of the SNP-specific regression coefficients was generated such that the number of clusters and cluster membership of the coefficients varied across iterations. By calculating posterior pairwise probabilities of coefficients to be equal, Ding and colleagues obtained a distance measure that could be used in complete linkage hierarchical clustering to evaluate a final clustering structure of the variants. The number of clusters considered ranged from small (2 to 5 clusters) to very large (up to 100 clusters).

Ding et al. [2011] modeled 244 nonsynonymous SNPs from the vascular endothelial growth factor (VEGF) pathway. The results from their study were based on 10 of the 200 replicates of the GAW17 data. Clusters of size 2–5 resulted in detection of an increasing number of true positives with an increasing number of clusters. However, as the number of clusters increased, the proportion of false discoveries increased as well. To obtain balance between sensitivity and specificity, Ding and colleagues evaluated receiver operating characteristic (ROC) curves for each replicate. The optimal number of clusters ranged from 59 to 96, such that for the 10 replicates the average sensitivity and specificity were 0.71 and 0.72, respectively. Considering only associations that were detected in all 10 replicates, Ding and co-workers identified 10 true-positive SNPs out of 39 associated variants. However, in this case, on average, two false positives were detected. Setting a threshold of detecting associations in at least 9 of 10 replicates led to the identification of 15 out of 39 associated SNPs with no additional false positives.

Guo et al. [2011] compared a modification of linear regression based on the least absolute shrinkage and selection operator (LASSO) approach [Tibshirani, 1996] with the F test and a version of the combined multivariate and collapsing (CMC) method [Li and Leal, 2008] that is applicable to quantitative traits, called the QCMC method. In brief, a LASSO regression is similar to an ordinary regression in which an additional penalty term is considered. This model is indexed by M_1 ; M_0 is the model under the null hypothesis that β is a vector of zeros. The corresponding residual sums of squares are denoted as RSS_{M_1} and RSS_{M_0} , respectively. Then, the F statistic is constructed:

$$F_{\text{LASSO}} = \frac{(RSS_{M_1} - RSS_{M_0})/(p - 1)}{RSS_{M_1}/(n - p)}, \quad (4)$$

which asymptotically follows the F distribution with $(p - 1, n - p)$ degrees of freedom, where n is the number of individuals and p is the generalized degree of freedom (GDF), which are applied in the F tests for model M_1 . In classical linear models the number of covariates is fixed and corresponds to the number of degrees of freedom. In contrast, in LASSO regression the number of nonzero coefficients cannot accurately measure the model complexity anymore. For LASSO regression the GDF is also introduced to correct for selection bias and to accurately measure the degrees of freedom of the obtained model. Further details of LASSO regression can be found elsewhere [Dasgupta et al., 2011].

Guo et al. [2011] compared two other methods with LASSO regression: the F test based on general linear regression and the QCMC method, in which markers are divided into rare and common subgroups on the basis of predefined allele frequencies. The CMC method and further collapsing methods have been described elsewhere [Dering et al., 2011]. Guo and colleagues investigated two scenarios for the QCMC method; in the first case subgroups consisted of SNPs with $MAF \leq 0.05$, and in the second case SNPs with $MAF \leq 0.01$ were collapsed.

Guo et al. [2011] performed association tests on the gene level with respect to quantitative trait Q2. The power and type I error frequencies were based on all 200 replicates. Because of low power and comparability of the applied methods, both the Bonferroni-corrected significance level ($\alpha = 1.5 \times 10^{-5}$) and the weakened significance level of 0.01 were considered.

In general, Guo et al. [2011] found that LASSO regression outperformed linear regression. With respect to the two genes *VNN1* and *SREBF1*, where most of the causal SNPs were less frequent than 0.005, both QCMC approaches were more powerful than LASSO regression. In the remaining cases LASSO regression outperformed the QCMC approaches. In general, power increased in genes when their contribution to the phenotype variation increased. However, all four approaches had inflated type I error frequencies, and the inflation for the LASSO regression was slightly larger than that of the other three approaches. Thus this could be the reason for the larger power of the LASSO regression, and further investigations are needed to evaluate these results.

Sung et al. [2011] used tiled regression to investigate the association between genotypes and a quantitative trait. In this approach the genome was divided into independent segments based on recombination hot spots. Each variant was assigned to a tile based on its physical position, and a tile was selected for further analysis if either the multiple regression on all variants within the tile or a simple linear regression on any single variant in the tile was significant. The independent individual variants within each tile were then selected with stepwise regression. Thereafter all significant variants were combined across tiles in higher-order stepwise regressions within chromosomes and then across the entire genome. The result was a set of variants that independently contributed to trait variation in a multiple regression model. Sung and colleagues compared power and type I error from the tiled regression method to those obtained from simple linear regressions using TRAP (Tiled Regression Analysis Package) [Sorant et al., 2010] and PLINK [Purcell et al., 2007]. They investigated three different sets of variants: (1) all variants individually, (2) all rare variants with $MAF < 0.01$ collapsed, and (3) all nonsynonymous rare variants with $MAF < 0.01$ collapsed. The rare variants were collapsed to a single variant coded as the presence or absence of any rare variant for each tile. All 200 replicates were analyzed.

In general, Sung et al. [2011] found that the power to detect associations for each causal variant was low, although the power increased when rare variants were collapsed. Only variants in two associated genes had power greater than 0.3 for Q1, and one gene had power greater than 0.2 for Q2.

The most intriguing finding in Sung and colleagues' study was that the average type I error varied by about three orders of magnitude across traits for traditional simple linear regression, ranging from 10^{-7} to 10^{-3} for uncollapsed variants and from 0 to 10^{-3} for collapsed variants. The correlation structure for traits Q1, Q2, and Q4 were quite different, and this suggests that correlations within and between linkage disequilibrium blocks may be dramatically inflating type I error. In contrast, the average type I error was stable across traits for the tiled regression method with both uncollapsed and collapsed variants, most likely because only uncorrelated variants were incorporated into the final model. Although the type I errors were stable, too many noncausal variants (about 10^{-6}) were identified when the tiled regression critical level was 10^{-7} (about 10^{-6}) and too few noncausal variants (about 10^{-3}) were identified when the critical level was 10^{-2} .

All three work groups performing association analyses on unrelated individuals used regression-based methods, and the power to detect associations was low in all three studies. However, Ding et al. [2011] and Sung et al. [2011] could identify at least some of the same

variants. For example, the three variants C13S431, C13S522, and C13S523 within *FLT1* (with MAF from 0.017217 to 0.066741) were identified by both groups. Guo et al. [2011] performed their analysis on the gene level, and they estimated the power and type I error based on all 200 replicates.

Type I errors were inflated for the analyses by Ding et al. [2011] and Sung et al. [2011] and for the simple linear regressions of Guo et al. [2011]. Sung et al. [2011] noted that the type I errors for simple linear regression varied dramatically with respect to the genetic structure underlying the trait. The issues of low power and inflated type I error were also reported from other GAW17 participants. Further investigations are necessary to compare these methods using further data sets.

Final Remarks

There were three different topics throughout the papers in Group 2: (1) functionality of variants and quality control issues, (2) family-based analyses, and (3) association analyses of unrelated individuals. The aims of the first subgroup were quite different. These were population structure analyses that used rare variants to predict functionality and examine the accuracy of genotype calls. The aim of the family-based analyses was the selection of the families that should be sequenced and the identification of high-risk pedigrees. In contrast, the association analyses of unrelated individuals were aimed at the identification of variants or genes with regression-based methods, although the power to detect associations was low.

The same limitations were shared by all three topics in Group 2. The main limitation was the small sample size (except Gagnon et al.'s [2011] study): As the allelic effect decreased, larger and larger sample sizes were required for the detection of rare variants. For example, 1,250 case subjects and 1,250 control subjects were required to attain a power of 80% to detect an allelic odds ratio of 2 at $\alpha = 5 \times 10^{-8}$ with MAF = 0.05. The detection of a similar association with a MAF of 0.01 required 6,000 case subjects and 6,000 control subjects [Asimit and Zeggini, 2010]. Not unexpectedly, the power of all considered methods was low. A further problem for the unrelated individuals and family-based analyses was that all 200 replicates were based on the same genotypes. In addition, for quality control issues there was a lack of reference sequence and missing quality scores.

The contributions of GAW17 Group 2 show that the identification of rare functional variants depends on more than genotypes and phenotypes alone. Important factors to identify rare variants can be prior functional knowledge, correction for population substructure, high-quality data, and the strategic use of large families.

Thus the work of Group 2 shows opportunities for incorporating rare variants into the genetic and statistical analyses of common diseases. However, further effort is required to establish standards for the preparation and statistical analysis of next-generation sequencing data in this new, interesting, and still developing research area.

Acknowledgments

The Genetic Analysis Workshops are supported by National Institutes of Health (NIH) grant R01 GM031575 from the National Institute of General Medical Sciences. We thank the Group 2 participants for their contributions and the reviewers for their helpful comments. This summary was supported in part by the Division of Intramural Research, National Human Genome Research Institute, NIH.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. *Nat Meth.* 2010; 7:248–9.

- Almasy LA, Dyer TD, Peralta JM, Kent JW Jr, Charlesworth JC, Curran JE, Blangero J. Genetic Analysis Workshop 17 mini-exome simulation. *BMC Proc.* 2011; 5(suppl 9):S2.
- Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. *Annu Rev Genet.* 2010; 44:293–308. [PubMed: 21047260]
- Baye TM, He H, Ding L, Kurowski BG, Zhang X, Martin LJ. Population structure analysis using rare and common functional variants. *BMC Proc.* 2011; 5(suppl 9):S8.
- Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet.* 2008; 40:695–701. [PubMed: 18509313]
- Cai Z, Knight S, Thomas A, Camp NJ. Pairwise shared genomic segment analysis in high-risk pedigrees: application to Genetic Analysis Workshop 17 exome-sequencing SNP data. *BMC Proc.* 2011; 5(suppl 9):S9.
- Chen YH, Liu CK, Chang SC, Lin YJ, Tsai MF, Chen YT, Yao A. GenoWatch: a disease gene mining browser for association study. *Nucleic Acids Res.* 2008; 36:W336–W340. [PubMed: 18440974]
- Dasgupta A, Sun YV, König IR, Bailey-Wilson JE, Malley JD. Brief review of regression-based and machine learning methods in genetic epidemiology: the GAW17 experience. *Genet Epidemiol.* 2011; X(suppl X):X–X.
- Dering C, Hemmelmann C, Pugh E, Ziegler A. Statistical analysis of rare sequence variants: an overview of collapsing methods. *Genet Epidemiol.* 2011; X(suppl X):X–X.
- Ding L, Baye TM, He H, Zhang X, Kurowski BG, Martin LJ. Detection of associations with rare and common SNPs for quantitative traits: a nonparametric Bayes-based approach. *BMC Proc.* 2011; 5(suppl 9):S10.
- Fitzte G, Cramer J, Ziegler A, Schierz M, Schreiber M, Kuhlisch E, Roesner D, Schackert HK. Association between c135G/A genotype and RET proto-oncogene germline mutations and phenotype of Hirschsprung's disease. *Lancet.* 2002; 359:1200–5. [PubMed: 11955539]
- Gagnon F, Roslin NM, Lemire M. Successful identification of rare variants using oligogenic segregation analysis as a prioritizing tool for whole-genome sequencing studies. *BMC Proc.* 2011; 5(suppl 9):S11.
- Guo W, Elston RC, Zhu X. Evaluation of a LASSO regression approach on the unrelated samples of Genetic Analysis Workshop 17. *BMC Proc.* 2011; 5(suppl 9):S12.
- Heath SC. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet.* 1997; 61:748–60. [PubMed: 9326339]
- Jaffe A, Wojcik G, Chu A, Golozar A, Maroo A, Duggal P, Klein AP. Identification of functional genetic variation in exome sequence analysis. *BMC Proc.* 2011; 5(suppl 9):S13.
- Krzanowsky, W. *Principles of Multivariate Analysis.* New York: Oxford University Press; 2003.
- 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; 83:311–21. [PubMed: 18691683]
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010; 20:1297–1303. [PubMed: 20644199]
- Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 2003; 31:3812–14. [PubMed: 12824425]
- Nolte IM, de Vries AR, Spijker GT, Jansen RC, Brinza D, Zelikovsky A, Te Meerman GJ. Association testing by haplotype-sharing methods applicable to whole-genome analysis. *BMC Proc.* 2007; 1(suppl 1):S129. [PubMed: 18466471]
- Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics.* 2000; 155:945–59. [PubMed: 10835412]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007; 81:559–75. [PubMed: 17701901]
- Sorant AJM, Cai J, Sung H, Kim Y, Wilson A. Tiled Regression Analysis Package (TRAP): software implementation of tiled regression methodology. *Genet Epidemiol.* 2010; 34:984–85.

- Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* 2005; 15:978–86. [PubMed: 15965030]
- Stram AH. Comparing nominal and real quality scores on next-generation sequencing genotype calls. *BMC Proc.* 2011; 5(suppl 9):S14.
- Sung H, Kim Y, Cai J, Cropp CD, Simpson CL, Li Q, Perry BC, Sorant AJ, Bailey-Wilson JE, Wilson AF. Comparison of results from tests of association in unrelated individuals with uncollapsed and collapsed sequence variants using tiled regression. *BMC Proc.* 2011; 5(suppl 9):S15.
- Sunyaev S, Ramensky V, Koch I, Lathe W III, Kondrashov AS, Bork P. Prediction of deleterious human alleles. *Hum Mol Genet.* 2001; 10:591–97. [PubMed: 11230178]
- Thomas A, Camp NJ, Farnham JM, Allen-Brady K, Cannon-Albright LA. Shared genomic segment analysis: mapping disease predisposition genes in extended pedigrees using SNP genotype assays. *Ann Hum Genet.* 2008; 72(2):279–87. [PubMed: 18093282]
- Tibshirani R. Regression shrinkage and selection via the LASSO. *J R Stat Soc.* 1996; 58:267–88.

Summary of studies in Group 2

Table 1

Subgroup	Contribution	Aim	Phenotype	Methods/software	Power (P)/type I error (I)	Significance level (local)
Functionality of variants and quality control issues	Baye et al. [2011]	Population structure analysis	–	Principal components analysis, Structure, discriminant analysis	–	–
	Jaffe et al. [2011]	Comparison of four programs for prediction of functional relevance of variants	–	VarioWatch, MAPP, SIFT, PolyPhen-2	–	–
	Stram [2011]	Quality control in comparison to HapMap III data	–	–	–	–
Family-based analyses	Cai et al. [2011]	Identification of high-risk pedigrees by means of the pairwise shared genomic segment (pSGS) method	Case-control status	Pairwise shared genomic sequence (pSGS)	P, I	0.001
	Gagnon et al. [2011]	Oligogenic segregation analysis for identifying rare variants with large effects in families	Q1, Q2, Q4	QTLs	P, I	–
Association analyses of unrelated individuals	Ding et al. [2011]	Association test using a nonparametric Bayes-based clustering algorithm	Q1	Bayes/cluster (based on regression)	P, I	0.05/244
	Guo et al. [2011]	Comparison of LASSO regression and <i>F</i> tests at the gene level	Q2	LASSO, <i>F</i> test, QCMC	P, I	1.5×10^{-5} or 0.01
	Sung et al. [2011]	Identification of variants contributing to the variation of quantitative traits by means of tiled regression	Q1, Q2, Q4	Tiled regression	P, I	0.1, 0.01, or 0.001