Published in final edited form as:
Genet Epidemiol. 2011 ; 35(Suppl 1): S48–S55. doi:10.1002/gepi.20649.

Analysis of Exome Sequences With and Without Incorporating Prior Biological Knowledge

Junghyun Namkung[1], **Paola Raska**[1], **Jia Kang**[2], **Yunlong Liu**[3], **Qing Lu**[4], and **Xiaofeng Zhu**[1]

[1]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH

[2]Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT

[3]Department of Medical and Molecular Genetics; Center for Computational Biology and Bioinformatics; and Center for Medical Genomics, Indiana University School of Medicine, Indianapolis, IN

[4]Department of Epidemiology, Michigan State University, East Lansing, MI

Abstract

Next-generation sequencing technology provides new opportunities and challenges in the search for genetic variants that underlie complex traits. It will also presumably uncover many new rare variants, but exactly how these variants should be incorporated into the data analysis remains a question. Several papers in our group from Genetic Analysis Workshop 17 evaluated different methods of rare variant analysis, including single-variant, gene-based, and pathway-based analyses and analyses that incorporated biological information. Although the performance of some of these methods strongly depends on the underlying disease model, integration of known biological information is helpful in detecting causal genes. Two work groups demonstrated that use of a Bayesian network and a collapsing receiver operating characteristic curve approach improves risk prediction when a disease is caused by many rare variants. Another work group suggested that modeling local rather than global ancestry may be beneficial when controlling the effect of population structure in rare variant association analysis.

Keywords

rare variant; association analysis; risk prediction model; population structure; biological information; receiver operating characteristic; Bayesian network

Introduction

Data for Genetic Analysis Workshop 17 (GAW17) were simulated on the basis of the exome sequencing data from the 1000 Genomes Project [Almasy et al., 2011]. Of the 3 billion bases of the human genome, the exomes are the most intensely annotated regions. The annotated information includes the starting and ending positions of translated bases, functions of translated products, and reported genetic variations and phenotypic changes that can be caused by genetic variations. This information can be easily retrieved using public database tools, such as the National Center for Biotechnology Information (NCBI), Ensemble, Vega, and the University of California Santa Cruz (UCSC), genome browsers.

Corresponding Author: Xiaofeng Zhu, Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, 2103 Cornell Road, Cleveland, OH 44106, Phone: 216-368-0201 Fax: 216-368-4880, xiaofeng.zhu@case.edu.

Many public database sources provide various types of information on genes and gene products. For example, NCBI's Clusters of Orthologous Groups of proteins (COGs) provide clusters of conserved sequences (orthologs) across species. The functional group that a gene belongs to can be found in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.ad.jp/kegg), BioCarta (http://www.biocarta.com/genes/allPathways.asp), Reactome (http://www.reactome.org/ReactomeGWT/entrypoint.html), and Gene Ontology (GO) (http://www.geneontology.org/). Protein motif information can be found in web databases such as PROSITE (http://ca.expasy.org/prosite/), Pfam (http://pfam.sanger.ac.uk/), and InterPro (http://www.ebi.ac.uk/interpro/). Protein interaction information can be located in the Biomolecular Object Network Databank (BOND) (http://bond.unleashedinformatics.com/) and the Database of Interacting Proteins (DIP) (http://dip.doe-mbi.ucla.edu/dip/Main.cgi). For nonsynonymous variants, the functional effect of the nucleotide changes can be predicted using web database programs such as PolyPhen (http://genetics.bwh.harvard.edu/pph2/) and SIFT (http://sift.bii.a-star.edu.sg/).

In population-based association studies of complex traits, various types of biological information can be used in many ways [Rebbeck et al., 2004; Ioannidis et al., 2009; Chen and Thomas, 2010]. However, such information is frequently limited to a description of the causality of the identified genetic variations located in a gene. Common features of multiple genes can be used to group genes and to test for the association with a target phenotype, such as in a pathway analysis. Prior knowledge that predicts the severity of changes resulting from a genetic variation can be incorporated into a statistical model by using different weights or prior probabilities. Biological information can also be used to prioritize a subset of genetic variants for subsequent functional analyses.

The contributors to Group 6 of GAW17 discussed the incorporation of known biological knowledge into genetic association studies of exome sequence data. Here, we summarize the nine individual contributions and the group discussion from the meeting (two of the papers are presented here as personal communications). The primary scientific questions addressed and the corresponding methods used by the individual contributors are presented in Table I. Although the study goals and approaches vary, the nine contributions can be grouped into four categories according to research similarity: (1) association tests between genetic variants and traits, (2) prioritization of genetic variants based on functional annotation, (3) construction of causal regulatory network and disease prediction using multiple risk factors, and (4) distribution inference of rare variants across populations.

The plausible biological scenario of having a biological pathway underlying the phenotype of interest was simulated for GAW17. Specifically, from the vascular endothelial growth factor (VEGF) pathway, 9 genes for Q1, 13 genes for Q2, and 15 genes for a binary trait were selected, and 51 variants of these genes were used as the causal variants for simulating the traits. In addition, the effect sizes of the selected causal variants were determined by the predicted impact of a mutation to protein function.

Data

Three quantitative traits (Q1, Q2, and Q4) and one binary trait were simulated for GAW17, and 200 replicates were included. All four traits were analyzed by two work groups of Group 6; Q1, Q2, and the binary trait were analyzed by two work groups; only the binary trait was analyzed by two work groups; Q1 and Q4 trait were analyzed by one work group; and no information on traits was used by two work groups (Table II). All but two work groups [Y. Hu et al., personal communication, 2011; Raska and Zhu, 2011] used all 200 replicates. One work group included all individuals including family members [Kang et al., 2011], seven work groups included 697 unrelated individuals, and one work group did not use individual information [Hu et al., personal communication, 2011]. Five work groups used a subset of single-nucleotide polymorphisms (SNPs), whereas the others used all of the available genotype data for the 24,487 SNPs. Covariates such as Age, Smoking status, Sex, and Ethnicity were used by four work groups. Additional biological knowledge from external sources was used by two work groups (Table II). Seven work groups used the answer information, and two work groups [Hu et al., personal communication, 2011; Raska and Zhu, 2011] did not use it.

Results

Testing Association Between Genetic Variants and Traits

Various statistical methods for conducting association analyses between genetic variants and traits were investigated by four work groups [Kumari and Chen personal communication, 2011; Lorenzo Bermejo et al., 2011]; Tong et al., 2011; Yang and Chen, 2011]. Each applied different test statistics to identify associated SNPs and compared the performances of different approaches in terms of empirical power and type I error. Each work group discussed how to deal with rare variants or the effect of the inclusion of rare variants in the analyses. Table III lists the methods used by each work group and whether biological knowledge was incorporated.

Test Statistics

Yang and Chen [2011] conducted a single-SNP association analysis using linear regression models with Q1. They used an additive effect model for each SNP and included Age and Smoking status as covariates in their models. They also conducted association tests by combining *p*-values of individual SNPs in a gene, genomic region, or pathway. SNPs having *p*-values below a predefined threshold were included in the combined *p*-value analysis using the truncated product method of Zaykin et al. [2002]. Yang and Chen [2011] investigated three different approaches: (1) analysis of common variants only, (2) analysis of rare variants only, and (3) an analysis that combined two *p*-values from common and rare variants.

Tong et al. [2011] compared single-SNP analysis with gene-based analysis. For single-SNP analysis, they applied linear regression to the three quantitative traits (Q1, Q2, and Q4) and used a logistic regression model for the binary trait. For the gene-based analysis, they used three approaches. First, they applied a regression model using multilocus genotypes that were defined as a combination of genotypes from all SNPs in a gene. Second, they classified the multilocus genotypes into three categories by comparing each multilocus genotype with a null multilocus genotype (i.e., a set of multilocus genotypes with no rare alleles). The significance for the three genotype categories was tested using permutation. Third, they used a defined similarity score-based test statistic for the binary trait. The similarity score between two distinct multilocus genotypes was computed as

$$1-\sum_{k=1}^{L}\frac{w_k\left|n_{ik}-n_{jk}\right|}{L}, \tag{1}$$

where n_{ik} and n_{jk} are the number of minor alleles of the *i*th and *j*th distinct genotypes at the *k*th SNP, L is the number of SNPs in a compared multilocus genotype, and w_k is a weight function. The weight is given as the inverse of the rare allele frequency so that a pair of genotypes sharing a rare allele has a higher similarity score than pairs that share only a common allele. Both Tong et al. [2011] and Yang and Chen [2011] applied a false discovery rate (FDR) controlling procedure to account for multiple tests [Benjamini and Hochberg, 1995].

Population stratification is one of the common causes of inflated type I error in genetic association analyses. Kumari and Chen [personal communication, 2011] proposed the following two-stage approach: filtering to select a subset of SNPs followed by the association tests. To eliminate the effect of population structure, they computed the principal components (PCs) from 1,000 SNPs in the genes that were not selected in the filtering stage. They chose the five PCs that explained the greatest genomic variance among individuals to serve as ancestry covariates. The trait values and combined genotype values were then regressed on those ancestry covariates, and the residuals were used for the testing.

Lorenzo Bermejo et al. [2011] evaluated an association analysis method proposed by Aitkin [2010] using GAW17 data. Aitkin [2010] proposed an integrated Bayes/likelihood approach using an uninformative uniform prior probability. Lorenzo Bermejo et al. [2011] compared the performance of Aitkin's method with standard logistic regression using unrelated samples with 16 SNPs of the *KDR* gene and the binary trait. They coded the SNP genotype as the number of minor alleles present and used the collapsed rare variants as independent variables. For the Bayes/likelihood approach, they obtained a distribution of deviance and deviance differences between a model with a genetic factor and the baseline model without genetic factors using a random walk Metropolis algorithm [Metropolis et al., 1953; Hastings, 1970]. The relative significance of the genetic effect was defined by the proportion of positive deviance differences from 10,000 sample draws. The larger the proportion, the less significant the model.

Rare Variant Combining Methods

Analysis of rare variants requires large sample sizes to obtain reasonable power. For individual SNPs, conventional statistical tests may not be able to obtain significant results even with thousands of samples. To tackle the difficulty of detecting rare variants, investigators have introduced many approaches that combine rare variants to define new variables, and these approaches have been reviewed by Dering et al. [2011]. Three of the Group 6 work groups adopted different rare variant collapsing approaches in their association analyses. Although the definition of rare variants can be subjective, a minor allele frequency (MAF) cutoff of 1% was used by Lorenzo Bermejo et al. [2011] and Yang and Chen [2011]. When variants are combined, the grouping criteria greatly affect the results of analysis in terms of power and interpretation. For exome sequencing data, individual genes can be a practical grouping rule for combining rare variants; this method was applied by three work groups. Kumari and Chen [personal communication, 2011] used the weighted-sum method of Madsen and Browning [2009] without applying a MAF cutoff. Lorenzo Bermejo et al. [2011] created a variable that indicates presence or absence of any rare alleles in a gene. The question of whether or not rare variants add information to common variants in an association analysis was also investigated by Yang and Chen [2011].

Power and Type I Error

Three work groups compared the performance of different methods by examining the empirical power, defined as the proportion of times that an association test resulted in significant *p*-values for a true associated SNP among 200 replicates. Because Q4 was not associated with any SNP, it was used to evaluate the type I error rate. Yang and Chen [2011] performed 1 million permutations for Q4 to obtain empirical *p*-values for individual SNPs, because Q4 did not follow a standard normal distribution. Tong et al. [2011] compared both the power of four approaches at a fixed false-positive rate (FPR) and the average FDR over the 200 replicates; the observed FPR was calculated as the number of reported false positives divided by the total number of noncausal loci, and the observed FDR was calculated as the number of reported false positives divided by the total number of significant outcomes. Lorenzo Bermejo et al. [2011] compared the distribution of *p*-values and proportions of positive deviances based on analysis of the 200 replicates.

Association Analysis Results

Yang and Chen [2011] identified four true causal genes, namely, *KDR*, *VEGFC*, *FLT1*, and *HIF1A*, associated with Q1 by using single-locus analysis. By combining their region-based analysis with statistics of rare and common SNPs, they also identified *ELAVL4* and *VEGFA* in addition to the four mentioned genes. Tong et al. [2011] detected a few genes at the genome-wide significance level ($\alpha = 10^{-5}$) with a similar number of false positives for the single-SNP analysis and the gene-based analyses. Kumari and Chen [personal communication, 2011] detected *FLT1* and *KDR* with *p*-values less than 10^{-5}.

The results of Yang and Chen [2011] suggest that gene-based analysis outperforms single-SNP analysis. The number of SNPs in a gene varies from one to hundreds. Accordingly, the number of tests required by single-SNP analysis varies, and this affects the multiple testing correction procedure. Single-SNP analysis was more powerful than gene-based analysis for only a few genes that had a highly significantly associated SNP, such as *ARNT* and *HIF3A* in the association test for trait Q1. However, Tong et al. [2011] concluded that single-SNP analysis had greater power more frequently than gene-based analysis, which uses multilocus genotypes as test covariates. For the binary trait, the single-SNP analysis had greater power to detect genes in which a causal SNP allele frequency was not too low, such as *FLT1*, which has a causal SNP with a MAF of 0.067. On the other hand, the gene-based method had greater power for genes with multiple rare causal SNPs, such as *FLT4* and *HIF1A*. Results from both work groups showed that the performances of the two approaches varied across all the tested genes. Some genes could be detected only by single-SNP analysis. Thus both groups suggested using single-SNP analysis and gene-based analysis as complementary methods. In addition, Yang and Chen [2011] conducted pathway-based analysis for the *VEGF* pathway using rare SNPs only and common SNPs only and a combined analysis; all three methods identified the pathway with significant *p*-values. However, the type I error was inflated by 5.8–8.2% at the 5% significance level.

Next-generation sequencing technologies allow us to study rare variants in addition to common variants that have been investigated in genome-wide association studies. It remains unclear whether the data obtained from the new technology will help us find the variants that contribute to missing heritability. Yang and Chen [2011] investigated the effect of inclusion of rare variants in association analysis using Q1. Their analysis showed that the inclusion of rare variants improved power by about 36% for *FLT4* and by about 47% for *VEGFA*. On the other hand, when common variants were close to the causal rare variants, most information seemed to be captured by the common variants. *ELAVL4*, *ARNT*, and *H1F1A* had substantially more power in the single-SNP analysis of common SNPs than in the combined analysis. Only three genes (*FLT4*, *VEGFA*, and *VEGFC*, in which most of the variants are rare) did not show such increase in power. In addition, Yang and Chen [2011] suggested that rare-variant-only analysis had the largest type I error. These results show that rare variant analysis using conventional regression analysis should be performed cautiously.

Lorenzo Bermejo et al. [2011] reported that both standard logistic regression and the integrated Bayes/likelihood approach performed poorly, because they produced many false positives and negatives. The power for 10 causal SNPs varied from 0% to 65% for logistic regression and from 2% to 69% for the Bayes/likelihood approach at the 5% significance level. On the basis of their comparisons, they concluded that Bayes factors might discriminate between causal variants and markers better than *p*-values from logistic regression. When collapsing rare variants, logistic regression produced more significant results than the Bayes/likelihood approach.

Integration of Known Biological Information

In the definition of new variables using multiple genetic variants, various types of biological information can be used to combine or weight SNPs. The four work groups that included association tests used a new variable defined by combining multiple SNPs, using either a gene or a biological pathway as a group. However, many other types of biological annotation information, such as gene ontology terms, can be adopted to combine SNPs. Analysis of variables that combine multiple genes will be useful when identifying groups of genes with small cumulative effects [Wu et al., 2010].

Lorenzo Bermejo et al. [2011] adopted Bayesian approaches with a noninformative prior distribution; however, the biological information can be used to determine a much more appropriate prior distribution. Bayesian approaches have been widely used in genetic studies because of their ability to incorporate knowledge from external data sources into the model specification [Beaumont and Rannala, 2004]. For example, the functional annotation of genetic variations, such as exonic, intronic, or intergenic variations, has been used as a prior weight in genetic association studies [Rannala and Reeve, 2001]. More recently, Bayesian approaches for genome-wide association studies have been shown to be useful for incorporating heterogeneous biological knowledge [Lewinger et al., 2007]. Chen and Witte [2007] and Heron et al. [2007] also adopted an empirical Bayes model method to detect associated SNPs. Although Bayesian approaches have several advantages, such as the capability of handling complex likelihood functions by using Markov chain Monte Carlo (MCMC) techniques, the intensive computation involved may limit the application of the approaches to large-scale genetic data.

Disease Modeling by Incorporating Multiple Traits and Prediction

Common complex diseases are believed to be caused by the interplay of multiple genetic and environmental risk factors. Statistical analyses that take this complexity into account will likely yield novel insights into the underlying pathophysiological and etiological processes, which will eventually promote the development of improved disease prediction and prevention strategies.

Kang et al. [2011] used a Bayesian network (BN) approach to infer the conditional independent relationships between the genetic and environmental risk predictors and disease outcomes. They constructed a BN for the genes that were selected on the basis of the association analysis. For the association analysis, they derived gene-level scores by using a weighted-sum SNP combining method that gave more weight to rare variants. In addition, they used another weight function that gave more importance to nonsynonymous SNPs than to synonymous ones. Using the gene scores, they conducted a linear regression analysis for each of the four traits, adjusting for Age, Sex, and Smoking status. Genes significantly associated with any of the four traits were used for the BN construction. Under certain constraints, such as a trait not having edges to genes, causal networks between genetic and environmental factors and traits were constructed with the optimization of the BN through MCMC analysis.

For the analysis of the GAW17 simulation data, Kang et al. [2011] showed that the BN approach was able to successfully uncover most of the true underlying relationships between the genetic and environmental risk predictors and the quantitative traits. To quantify the advantage of using a joint approach (e.g., the BN approach), where multiple traits are considered simultaneously, over a marginal approach (e.g., least absolute shrinkage and selection operator [LASSO]), where only one trait is considered, they calculated the area under the curve (AUC) of both methods to avoid cutoff selection. The calculated AUC measures how closely the detected genes agree with the true causal genes in the simulation model. Kang and colleagues found that the AUC for the BN was 0.61, whereas the AUC for LASSO was only 0.57. The BN approach also outperformed the conventional LASSO approach with greater AUC in terms of selecting true disease causal genes.

Kang et al. [2011] also demonstrated that incorporating biological knowledge, such as functional annotation of SNPs, into the analysis is beneficial. Using the BN approach, they evaluated the combined role of genetic and environmental risk predictors in disease risk prediction and found that genetic variants played a limited role compared to their environmental counterparts. This finding is consistent with the findings of Wei and Lu [2011], who also suggested a limited role of genetic variants in predicting disease outcomes from the GAW17 simulated data. In their paper, Kang and colleagues introduced a collapsing receiver operating characteristic (ROC) curve approach for risk prediction of the sequencing data. The collapsing ROC curve approach is essentially an extension of the previously developed forward ROC curve approach [Ye et al., 2011] with an additional multistage collapsing procedure. The multistage collapsing procedure was developed on the basis of the ideas of Li and Leal [2008]. At each stage, the procedure selects rare variants in a stepwise manner and then collapses them into a pseudo-common variant, which is created when no additional rare variants can further increase its accuracy. The collapsing procedure is repeated on the remaining rare variants, and it generates a set of pseudo-common variants. The forward ROC curve approach is then applied to all pseudo-common and common variants to form an optimal risk prediction model. By applying the collapsing ROC curve approach to the GAW17 simulated data, Kang and colleagues showed that additional accuracy could be gained by considering rare variants.

Prioritization of Genetic Variants Based on Functional Annotation

One challenge of genetic studies is identifying functional genetic variants that cause phenotypic differences in humans from tens of thousands of variants. Many of these variants might be indirectly associated with a phenotype through linkage disequilibrium between genotyped variants and causal variants that have a direct effect on the phenotype. Subsequent functional analysis is required to identify the causal variants from all the significant candidates. Biological knowledge accumulated from previous experimental studies enables the prioritization of a subset of candidates for further analysis. Two contributions in Group 6 [Hu et al., personal communication, 2011; Teng et al., 2011] investigated the features of two types of functional variant groups and introduced approaches for prioritizing candidate causal variants.

Teng et al. [2011] focused on the potential role of synonymous polymorphisms in affecting the binding affinities of RNA-binding proteins that may change the RNA splicing form, that is, cause alternative splicing. They proposed a statistical framework to evaluate how likely a SNP is to affect the binding of any of the nine RNA-binding proteins whose binding affinities had already been characterized. Among 10,113 synonymous polymorphisms identified in the 697 individuals in the GAW17 data, they found 1,851 candidates that potentially affected the binding affinity of at least one of the nine proteins, 182 of which were located in alternatively spliced exons. They also showed that the average MAF of SNPs located in alternatively spliced exons was similar to the average MAF of nonsynonymous SNPs, both of which were significantly lower than the average MAF of the remaining synonymous SNPs. The low MAF may indicate that the 182 identified polymorphisms were under negative selection because of the deleterious alterations in the gene product [Goddard et al., 2000]. Teng et al. [2011] also suggested a workflow to identify functional SNPs that might affect a phenotype by altering the splicing pattern of a gene.

Hu et al. [personal communication, 2011] investigated nonsynonymous SNPs in the coding regions that do not fold spontaneously into a unique three-dimensional shape. These regions are also called intrinsically disordered regions [Iakoucheva et al., 2002]. More than 40% of all human proteins contain disordered regions that are known to play critical roles in many key biological processes [Uversky and Dunker, 2010]. Hu and colleagues systematically examined the effect of SNPs on the structural changes in terms of intrinsic disorder status.

Teng et al. [2011] and Hu et al. [personal communication, 2011] demonstrated the usefulness of prioritization of SNPs by using previously reported biological knowledge to identify functional genetic variations among SNPs with association signals from population genetic studies. Teng et al. [2011] in particular showed an example of an important function of synonymous SNPs that might be missed.

Distribution of Rare Variants Across Populations

Raska and Zhu [2011] used the GAW17 exome data to demonstrate that there is a significant difference in genome-wide rare variant density across the seven studied populations. They compared regression coefficients between counts of rare variants and total variant counts per gene for each population. They also computed Tajima’s *D* values [Tajima, 1989] on each gene for each population for all 3,205 genes. They found that the populations clustered by continent for both the regression slopes and Tajima’s *D* values, with the African populations showing the highest rare variant densities and the European populations showing the lowest variant densities, findings that are consistent with the literature.

The variation in rare variant density has the potential to confound rare variant association analyses in mixed and admixed populations. Raska and Zhu [2011] showed that this confounding association existed when using both group-level statistics, which were used to compare case subjects and control subjects, and individual-based statistics. They also showed that with the group-level statistics, a mixed population had an increased rare variant density beyond the level of the individual population. This suggests that for a rare variant association analysis that uses a group-level statistic, diversity within case subjects or control subjects in addition to global ancestry would have to be taken into account.

In addition to this genome-wide variation across populations, there is variance in selection pressure throughout the genome that causes rare variant density to differ from gene to gene within populations. Raska and Zhu [2011] found that of those genes with the lowest Tajima’s *D* values, some were common across all populations, but others were population specific. The expected rare variant density for a gene will therefore depend on both the gene and the population of origin. As a consequence, genome-wide ancestral estimates may not provide adequate control in association studies for admixed individuals; instead, local gene-specific estimates may be necessary.

Discussion

The contributors in Group 6 examined various approaches for investigating the genetics underlying the traits simulated in the GAW17 data. Because the traits were simulated using the contributions of both common and rare variants, several work groups compared various approaches for searching genetic variants associated with the traits, including single-variant, gene-based, and pathway-based analyses. The performance of the different approaches was dependent on which genes contributed to trait variation through either common variants or multiple rare variants [Tong et al., 2011; Yang and Chen, 2011]. Novel statistical methods, such as principal-components-based approaches, may hold some promise, but further investigation is warranted [Kumari and Chen, personal communication, 2011]. Biological annotation on rare variants, such as information on splicing regulation and protein structure, can prove useful. Integration of such information in statistical analysis will improve statistical power to detect causal variants [Hu et al., personal communication, 2011; Teng et al. 2011]. However, the method to most efficiently incorporate this biological knowledge requires further investigation.

An interesting analytical approach is the application of the graphical model (e.g., a BN) [Kang et al., 2011]. By simultaneously modeling the conditional independence of genetic and environmental factors and multiple phenotypes, the BN approach was able to reconstruct the underlying complex topology. More important, the reconstructed topology was almost identical to the true topology of the genetic and environmental factors on the phenotypes. Furthermore, integrating additional information, such as functional annotation, into the prediction models led to improved risk prediction [Kang et al., 2011; Wei and Lu, 2011]. However, simply adding many SNPs to a prediction model does not necessarily improve prediction power [Kang et al., 2011; Wei and Lu, 2011].

The GAW17 data suggest that the rare variant distribution varies across the genome, indicating that genes may undergo balancing or positive selection [Raska and Zhu, 2011]. The difference in rare variant distribution across populations suggests that caution should be exercised in association analyses of rare variants, even when incorporating local population structure into the analysis, as suggested by Qin et al. [2011] and Wang et al. [2011].

Acknowledgments

This work was supported by National Institutes of Health (NIH) grant HL086718 from the National Heart, Lung, and Blood Institute and NIH grants HG003054 and HG005854 from the National Human Genome Research Institute. The Genetic Analysis Workshops are supported by NIH grant R01 GM031575.

References

Aitkin, M. Statistical Inference: An Integrated Bayesian/Likelihood Approach. London: CRC Press; 2010.

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; 5 suppl 9:S2.

Beaumont MA, Rannala B. The Bayesian revolution in genetics. Nat Rev Genet. 2004; 5:251–261. [PubMed: 15131649]

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

Chen GK, Thomas DC. Using biological knowledge to discover higher order interactions in genetic association studies. Genet Epidemiol. 2010; 34:863–878. [PubMed: 21104889]

Chen GK, Witte JS. Enriching the analysis of genome-wide association studies with hierarchical modeling. Am J Hum Genet. 2007; 81:397–404. [PubMed: 17668389]

Dering C, Pugh E, Ziegler A. Statistical analysis of rare sequence variants: an overview of collapsing methods. Genet Epidemiol. 2011 X(suppl X):XX–XX.

Goddard KA, Hopkins PJ, Hall JM, Witte JS. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am J Hum Genet. 2000; 66:216–234. [PubMed: 10631153]

Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970; 57:97–109.

Heron EA, O'Dushlaine C, Segurado R, Gallagher L, Gill M. Exploration of empirical Bayes hierarchical modeling for the analysis of genome-wide association study data. Biostatistics. 2007; 12:445–461. [PubMed: 21252078]

Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol. 2002; 323:573–584. [PubMed: 12381310]

Ioannidis JP, Thomas G, Daly MJ. Validating, augmenting, and refining genome-wide association signals. Nat Rev Genet. 2009; 10:318–329. [PubMed: 19373277]

Kang J, Zheng W, Li L, Lee JS, Yan X, Zhao HY. Use of Bayesian networks to dissect the complexity of genetic disease: application to the Genetic Analysis Workshop 17 simulated data. BMC Proc. 2011; 5 suppl 9:S37. [PubMed: 21645318]

Lewinger JP, Conti DV, Baurley JW, Triche TJ, Thomas DC. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. Genet Epidemiol. 2007; 31:871–882. [PubMed: 17654612]

Li BS, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008; 83:311–321. [PubMed: 18691683]

Lorenzo Bermejo J, Beckmann L, Chang-Claude J, Fischer C. Using the posterior distribution of deviance to measure evidence of association for rare susceptibility variants. BMC Proc. 2011; 5 suppl 9:S38.

Madsen B, Browning S. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009; 5:e1000384. [PubMed: 19214210]

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equations of state calculations by fast computing machines. J Chem Phys. 1953; 21:1087–1092.

Qin H, Morris N, Kang SJ, Li M, Tayo B, Lyon H, Hirschhorn J, Cooper RS, Zhu X. Interrogating local population structure for fine mapping in genome-wide association studies. Bioinformatics. 2011; 26:2961–2968. [PubMed: 20889494]

Rannala B, Reeve JP. High-resolution multipoint linkage-disequilibrium mapping in the context of a human genome sequence. Am J Hum Genet. 2001; 69:159–178. [PubMed: 11410841]

Raska P, Zhu X. Rare variant density across the genome and across populations. BMC Proc. 2011; 5 suppl 9:S39.

Rebbeck TR, Spitz M, Wu X. Assessing the function of genetic variants in candidate gene association studies. Nat Rev Genet. 2004; 5:589–594. [PubMed: 15266341]

Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989; 123:585–595. [PubMed: 2513255]

Teng M, Wang Y, Wang G, Jung J, Edenberg HJ, Sanford JR, Liu Y. Prioritizing single-nucleotide variations that potentially regulate alternative splicing. BMC Proc. 2011; 5 suppl 9:S40.

Tong L, Tayo B, Yang J, Cooper RS. Comparison of SNP-based and gene-based association studies in detecting rare variants using unrelated individuals. BMC Proc. 2011; 5 suppl 9:S41.

Uversky VN, Dunker AK. Understanding protein non-folding. Biochim Biophys Acta. 2010; 1804:1231–1264. [PubMed: 20117254]

Wang X, Zhu X, Qin H, Cooper RS, Ewens WJ, Li C, Li M. Adjustment for local ancestry in genetic association analysis of admixed populations. Bioinformatics. 2011; 27:670–677. [PubMed: 21169375]

Wei C, Lu Q. Collapsing ROC approach for risk prediction research on both common and rare variants. BMC Proc. 2011; 5 suppl 9:S42.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Hum Genet. 2010; 86:929–942. [PubMed: 20560208]

Yang HC, Chen CW. Region-based and pathway-based QTL mapping using a p-value combination method. BMC Proc. 2011; 5 suppl 9:S43.

Ye C, Cui Y, Wei C, Elston RC, Zhu J, Lu Q. A nonparametric method for building predictive genetic tests on high-dimensional data. Hum Hered. 2011; 71:161–170. [PubMed: 21778735]

Zaykin DV, Zhivotovsky LA, Westfall PH, Weir BS. Truncated product method for combining *p*-values. Genet Epidemiol. 2002; 22:170–185. [PubMed: 11788962]

Table I

Research goals and methods used by each work group in Group 6

Work group	Primary scientific question	Analysis method(s)
Hu et al. [personal communication, 2011]	To determine whether SNPs in intrinsic disordered regions can represent disease risk	Bioinformatics
Kang et al. [2011]	To determine how multiple environmental factors and genetic factors interact to determine disease phenotype and conditioning of all other variables	Graphical models (Bayesian networks)
Kumari and Chen [personal communication, 2011]	To identify group-wise association of rare variants with a disease	Weighted-sum method after adjusting for population structure using principal components
Lorenzo Bermejo et al. [2011]	To evaluate an integrated Bayes/likelihood approach and compare its performance with that of standard logistic regression	Aitkin's integrated Bayes/ likelihood approach
Raska and Zhu [2011]	To determine how rare variant distribution varies across the genome and across populations	Rare variant to total variant ratios and Tajima's *D*
Teng et al. [2011]	To determine whether synonymous SNPs can be attributed to disease risk	Bioinformatics
Tong et al. [2011]	To compare SNP- and gene-based association analyses	False discovery rate and similarity-based statistics
Wei and Lu [2011]	To develop and evaluate statistical approaches for genetic risk prediction based on both common and rare SNPs	Novel grouping ROC approach
Yang and Chen [2011]	To identify quantitative trait loci (QTLs), genes, or pathways using region- and pathway-based QTL mappings and to compare the power of several QTL mappings (i.e., rare SNPs only, common SNPs only, and combined analysis).	*P*-value combination and Monte Carlo procedure

Table II

Data used in the studies by Group 6

Work group	**Traits**	**SNPs**	**Covariates**	**Biological knowledge**
Hu et al. [personal communication, 2011]	None	Nonsynonymous SNPs	None	Bioinformatics tool (VSL2) to predict protein intrinsic disorder status
Kang et al. [2011]	Q1, Q2, Q4, and binary trait	24,487 SNPs, synonymous and nonsynonymous treated separately	Smoking, Age, Sex	SNP functional annotation
Kumari and Chen [personal communication, 2011]	Q1, Q2, and binary trait	24,487 SNPs	None	Gene
Lorenzo Bermejo et al. [2011]	Binary trait	16 SNPs in *KDR* gene	Smoking, Age, Ethnicity	None
Raska and Zhu [2011]	None	24,487 SNPs	None	Gene
Teng et al. [2011]	Q1, Q2, and binary trait	Synonymous SNPs	None	Consensus sequences of RNA-binding protein target sites (nine RNA-binding proteins)
Tong et al. [2011]	Q1, Q2, Q4, and binary trait	24,487 SNPs	Smoking, Age, Sex	Gene
Wei and Lu [2011]	Binary trait	161 disease-susceptibility SNPs from GAW17 answer sheet	None	None
Yang and Chen [2011]	Q1 and Q4	24,487 SNPs	Smoking, Age, Sex	Gene and pathway

Table III

Summary of studies on association testing

	Lorenzo Bermejo et al. [2011]	Kumari and Chen [personal communication, 2011]	Tong et al. [2011]	Yang and Chen [2011]
Statistics	Integrated Bayes/likelihood approach	Multiple correlation coefficient, Armitage trend test	Linear/logistic regression, similarity score-based statistics	Linear regression
Genome-wide analysis	–	Q1, Q2, binary trait	Q1	Q1
Adjusting covariates	Age, Smoking, Ethnicity	Principal components computed from 1,000 SNPs	Age and Smoking for Q1 and binary trait; Age, Smoking, and Sex for Q4	Age and Smoking
Group to combine SNPs	Gene	Gene	Gene	Gene/pathway
Combining SNPs	Presence/absence (alleles with MAF < 1%)	Weighted-sum statistics (Madsen and Browning [2009])	Multivariate, combined	Combining *p*-values obtained from individual SNP tests
Multiple testing	–	Bonferroni	False discovery rate	False discovery rate of 5%
Power	Binary trait	Q1, Q2, binary trait	Q1	Q1
Type I error	–	–	Q4	Q4