Published in final edited form as:
Genet Epidemiol. 2011 ; 35(Suppl 1): S107–S114. doi:10.1002/gepi.20659.

Lessons Learned from Genetic Analysis Workshop 17: Transitioning from Genome-Wide Association Studies to Whole-Genome Statistical Genetic Analysis

Alexander F. Wilson[1] and **Andreas Ziegler**[2]

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

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

Abstract

Genetic Analysis Workshop 17 (GAW17) focused on the transition from genome-wide association study designs and methods to the study designs and statistical genetic methods that will be required for the analysis of next-generation sequence data including both common and rare sequence variants. In the 166 contributions to GAW17, a wide variety of statistical methods were applied to simulated traits in population- and family-based samples, and results from these analyses were compared to the known generating model. In general, many of the statistical genetic methods used in the population-based sample identified causal sequence variants (SVs) when the estimated locus-specific heritability, as measured in the population-based sample, was greater than about 0.08. However, SVs with locus-specific heritabilities less than 0.03 were rarely identified consistently. In the family-based samples, many of the methods detected SVs that were rarer than those detected in the population-based sample, but the estimated locus-specific heritabilities for these rare SVs, as measured in the family-based samples, were substantially higher (>0.2) than their corresponding heritabilities in the population-based samples. Substantial inflation of the type I error rate was observed across a wide variety of statistical methods. Although many of the contributions found little inflation in type I error for Q4, a trait with no causal SVs, type I error rates for Q1 and Q2 were well above their nominal levels with the inflation for Q1 being higher than that for Q2. It seems likely that this inflation in type I error is due to correlations among SVs.

Keywords

linkage; association; next-generation sequencing; computer simulation

Introduction

Genetic Analysis Workshop 17 (GAW17) focused on the transition from study designs and methods for a genome-wide association study (GWAS) to the study designs and statistical genetic analysis methods that will be required for the analysis of next-generation sequence data. Several elements make up a statistical genetic analysis study: the density of the markers (ranging from single-nucleotide polymorphism [SNP] panels to whole-genome sequences), the study design (population- or family-based), and the type of trait (qualitative or quantitative). The recent focus on genome-wide association studies has often precluded

Corresponding author: Alexander F. Wilson, Ph.D., NIH/NHGRI, Genometrics Section, 333 Cassell Drive, Suite 1200, Baltimore, MD 21224, Tel: 443-740-2918, afw@mail.nih.gov.

other methods and study designs, and the term *GWAS* has come to commonly refer to a population-based study of unrelated individuals with high-density SNP panels in case-control or cohort study designs. This focus has been driven largely by the available genotyping technology, the ease of obtaining population-based samples of unrelated individuals, the focus on categorical disease, and the computational speed and simplicity of the analysis. One of the major strengths of next-generation sequence technology is that it can identify both common SNPs with minor allele frequencies greater than about 5% and rare sequence variants with minor allele frequencies of less than about 1–5%. These SNPs and rare sequence variants are collectively referred to here as sequence variants (SVs).

In a GWAS, the SNPs on the high-density genotyping platform are selected to be relatively common; the inclusion of rare variants in next-generation sequence data is problematic because variants with low minor allele frequencies are often removed during data cleaning and present difficulties in terms of traditional statistical analyses. It is clear that rare SVs require different study designs and methods of statistical analysis, including linkage analysis and intrafamilial tests of association, to identify the variants responsible for the phenotypic variation of quantitative traits and the susceptibility to qualitative disease. However, there is considerable discussion over the most appropriate study designs and methods. To provide a public forum for this discussion, the participants in GAW17 focused specifically on the type of data that would typically be seen from a "mini" exome next-generation sequencing project that included both common and rare SVs. The 166 contributions to GAW17 are organized into the 15 thematic summaries presented in this volume. Here, we briefly review the model underlying the simulations and reinterpret the causal SVs in terms of their locus-specific heritabilities; we also summarize the lessons learned from the group presentations at the GAW17 meetings in Boston, Massachusetts (October 2010), the group summaries [this issue], and the individual papers published in *BMC Proceedings* [v 5 (suppl 9), 2011]. We focus on the themes that span many of the group reports and the take-home lessons gleaned from GAW17 in terms of what matters and why.

Methods

The underlying simulation model is described in detail by Almasy et al. [2011]. In that report, the frequency of the minor allele and its effect size were given for all the causal variants for quantitative traits Q1 and Q2 and for the liability underlying the discrete trait (D); Q4 had no causal variants. The allele frequency and effect size can be combined into a more interpretable measure, the locus-specific heritability (h_L^2), at least for the quantitative traits and the underlying liability of the qualitative trait. For a specific locus L, it was assumed that there was no dominance such that:

$$h_L^2=2p(1-p)a^2, \tag{1}$$

where p is the frequency and a is the additive effect of the minor allele. For the population-based sample, h_L^2 was estimated for quantitative traits Q1 and Q2 [Falconer and Mackay, 1996] for each causal SV in each replicate, and the estimates were averaged over all 200 replicates. Results are presented in Table I. Similarly, the estimated locus-specific heritabilities for Q1 and Q2 for the family-based sample are presented in Table II. Two methods were used to estimate h_L^2 in the family-based sample—the regression of offspring on mid-parent [ROMP] method [Pugh et al., 2001; Roy-Gagnon et al., 2005] and the traditional method—but in the traditional method, all individuals were included and familial dependencies were ignored.

Given that the number of causal SVs and their effect sizes for Q1 were fewer and larger than those for Q2 and that the underlying liability for the disease trait D was based on SVs involved in Q1 and Q2, we focus primarily on results from Q1, with commentary on Q2, Q4, and D where appropriate. A comparison of the results in Tables I and II illustrates the difference between the estimated locus-specific heritability in population- and family-based study designs. In the population-based sample, the causal SVs with the highest locus-specific heritabilities were in *FLT1*: C13S523, C13S522, and C13S524, with locus-specific heritabilities of 0.15, 0.08, and 0.04, respectively. The next highest were in *KDR*: C4S1877 and C4S1889, both with heritabilities of 0.03. However, the highest locus-specific heritabilities in the family-based sample determined with the ROMP method were in *VEGFC* and *VEGFA*, namely, C4S4935 and C6S2981 with heritabilities of 0.25 and 0.23, respectively, followed by SVs in *FLT1* and *KDR* (C13S523, C4S1878, and C4S1884, with heritabilities of 0.046, 0.044 and 0.018, respectively). Estimates of -based sample, the causal SVs with the h_L^2 for SVs and their ranks were similar to those obtained with the traditional method.

For Q2, the estimated locus-specific heritabilities were considerably lower than those for Q1. Only C6S5380 (*VNN1*) had a locus-specific heritability greater than 0.01 in the population-based sample, and only C10S3109 (*SIRT1*) and C9S444 (*VLDLR*) had locus-specific heritabilities greater than 0.05 in the family-based sample. For most of the other causal SVs in Q1 and Q2 the confidence intervals included zero (the null hypothesis).

Summary of Results

The Simulation Model Matters

Comparison of empirically obtained results with that of a known simulation model that reflects the true state of nature is a powerful tool for evaluating methods and approaches in statistical genetics. Simulation studies can provide insight into the statistical validity, the type I error rate and power, and the large- and small-scale sample properties of new statistical methods. However, it is important to realize that building a simulation model that mimics large portions of the genome is problematic. Because the characteristics of the genome are neither fully measured nor understood, particularly in terms of all the correlations and interactions between millions of SVs, the most prudent approach is to use actual sequence data as the basis for the underlying model. With this approach, the known (and unknown) structure in the genome can be incorporated without having to create a model with assumptions that may not accurately reflect the true state of nature. As described by Almasy et al. [2011], complex trait phenotypes based on actual SVs can then be modeled as desired. However, in the simulated data, replications based on the same underlying sequence data were not independent, and the genotypes were identical over all the replicated samples. Thus the replicates were completely correlated with respect to genotype, and because at least some of the variation inherent in the phenotypes was generated on the basis of causal SVs, the phenotypes were correlated across replicates as well.

Cost and Sample Size Matter

The cost for next-generation sequencing is decreasing on a continuing basis. At present, a high-density SNP panel costs about $500, a whole-exome sequence costs about $2,500, and a whole-genome sequence with 30× coverage costs about $5,000. The cost to genotype one replicate of 697 individuals with a high-density SNP panel would be about $348,500; the cost to perform whole-exome sequencing for one replicate of 697 individuals would be about $1,742,500, and the cost for all 200 replicates would be about $350 million. Although the sample size of the data provided for GAW17 was fairly typical of what can be obtained or afforded at this point in time in terms of sequencing data, it was quite small compared to sample sizes for genome-wide association studies. Until the cost of sequencing approaches the cost of a GWAS, the detectable effect size for rare SVs that contribute to trait variation will remain high. Samples large enough to detect small effect sizes may be prohibitively expensive in the near future, although this will be less of a problem as sequencing costs continue to decline. Several contributors to GAW17 pooled replicates to increase the sample size, but at least in the near term, because of the cost to replicate even one sample, caution must be used in making generalizations about methods that pool replicates to expand the sample size. However, it is important to note that for most of the causal rare SVs, the estimated locus-specific heritabilities were so small that even pooling replicates did not substantially increase the power of these tests.

Minor Allele Frequency and Effect Size Matter

Rare or common, SVs have two main attributes: the frequency of the allele(s) of the variant and the size of the effect of the variant allele(s). Although much debate has focused on common versus rare alleles, what is most important is the size of the allelic effect of the variant and at what level the effect is measured (the individual, the family, or the population). The variant for familial hypercholesterolemia, for example, has a large effect in individuals with two copies of the variant, less of an effect in relatives who carry only one copy of the variant, and virtually no effect at the population level because the allele is so rare. Unless population studies are very large, rare causal variants with large effect size will be difficult to detect in population-based studies. However, in a family-based sample, because of the increased presence of a specific causal SV in relatives, the effect size will no longer be overwhelmed by its low population frequency and the ability to detect a rare variant will be amplified.

Data Quality Matters

The quality of the SV data is critical to the interpretation of the results. A number of data quality issues were raised at GAW17, some of which were identical to those for a GWAS and some of which were unique to next-generation sequence data. In terms of concordance with known SNP genotyping, Stram [2011] and Hemmelmann et al. [2011] noted that the concordance rate was only 88.5% for SVs shared by both the mini-exome data and HapMap genotyping calls for those individuals genotyped on both platforms. Furthermore, some SVs in the simulated mini-exome sequence data exhibited strong correlations with SVs on different chromosomes [Thomas et al., 2011; Tintle et al., 2011]. Again, these interchromosomal correlations may be due to sequence misalignment or chance. When the data contain a large number of rare SVs and a relatively small number of individuals, correlation between SVs can be substantial, and sometimes even complete correlation occurs simply by chance alone.

The amount of missing data and the call rates are critical measures of data quality in a GWAS. With well-characterized high-density SNP genotyping platforms, markers are often dropped if the proportion of missing data exceeds a specified threshold. Individuals can be dropped if the proportion of genotyping calls fails to reach some minimal threshold as well. This is considerably more difficult with next-generation sequence data, because, by definition, the rare SVs can be quite rare. In this situation, it is difficult to distinguish between a rare SV and a sequencing error. Family-based samples provide the opportunity for Mendelian consistency checking, which can help to distinguish between a real segregating variant and a sequencing error. Finally, cryptic relatedness (unknown relatives in the sample) remains an issue for both high-density SNP genotyping and sequence variants.

Nonindependence of Sequence Variants Matters

It is important to note that the lack of independence between SVs is a larger issue in next-generation sequencing studies than in a GWAS because of the increased density of the SVs. As in a GWAS, nonindependence between SVs in linkage disequilibrium blocks is common, although in next-generation sequence data the density of SVs is increased and thus the number of highly correlated SVs is increased as well. In these data, however, nonindependence in SVs across linkage disequilibrium blocks also occurred. This was a key finding in both the Group 9 and Group 7 summaries [Thomas et al., 2011; Tintle et al. 2011, respectively]. These summaries noted that failure to address correlations between SVs increased the type I error rate over a wide variety of methods. Unlike typical analyses in a GWAS, Thomas et al. [2011] noted that single-SNP tests are not reliable for association.

Enrichment for Phenotype and Genotype Matters

At GAW17, investigators presented a number of strategies to enrich the samples to be analyzed for the presence of rare SVs. On the phenotyping side these strategies included selection for discordant relative pairs, extreme phenotypes, and distributional extremes. For the population-based samples, Bailey-Wilson [2011], in the Group 14 summary, reported that the power to detect association was increased when subsamples were selected for extreme phenotypes [Lamina, 2011]. On the genotyping side the strategies included the use of families and rare variant counts in a case-control framework. In the Group 10 summary, Melton and Pankratz [2011] described methods that used the joint analysis of multiple correlated phenotypes to reduce the phenotypic dimensionality and methods that reduced the genotypic dimensions.

Families Matter

It is important to realize that the use of families is an enrichment strategy for both phenotype and genotype. The selection of highly loaded families focuses on those families with a high proportion of affected individuals or large trait variation. The selection of extended families will overrepresent some rare causal variants if they are present in the founders or in individuals who marry into the family, and then segregate throughout the extended family. By the same token, some rare causal variants will be lost if they are not present in the set of founders or individuals who marry into the family or if they segregate out of the family in subsequent meioses. The amplification of a rare SV was particularly relevant in Family 7, in which a rare causal SV in the *VEGFC* pathway was present in the founders and then segregated throughout the entire extended family. A number of contributors were able to identify causal SVs in the *VEGFC* and *VEGFA* genes because the estimated locus-specific heritabilities were higher in the family samples than in the population-based samples.

Family data can also provide other relevant information, such as estimates of locus-specific heritability and candidate regions based on linkage analysis. In the Group 11 summary, Hinrichs and Suarez [2011] noted that information from family-based samples could be used to identify candidate regions, identify individuals for sequencing, and provide weights for intrafamilial tests of association.

How You Treat Rare Variants Matters

When the SVs are rare, it becomes necessary to collapse rare variants into new derived variables in order to analyze them. Many different methods were used to collapse or aggregate SVs at GAW17, and these methods are summarized by König et al. [2011], Melton and Pankratz [2011], Sun et al. [2011], Sung et al. [2011], Tintle et al. [2011], and Ye and Engelman [2011]. In general, the analysis of collapsed variants had somewhat better power than that of uncollapsed variants, although the performance of collapsing and aggregating methods depended to a large extent on the underlying genetic structure and the use of unrelated individuals or family data. Summarizing a somewhat different approach, Kent [2011] noted that using common variants to tag rare variants can work, but not very reliably.

The performance of the collapsing methods is tempered somewhat by the fact that all the causal variants in the underlying simulation model had allelic effects in the same direction, which is most likely not the case in the real world [Bickeböller et al., 2011].

The methods described by Ye and Engelman [2011] and by Tintle et al. [2011], for example, illustrate a number of collapsing schemes, some previously proposed and some novel. For the most part, these methods focused on the sample of unrelated individuals. Not unexpectedly, results from these methods identified genes containing SVs with the highest locus-specific heritability, for example, SVs in *FLT1* and *KDR* in trait Q1. Ye and Engelman [2011] noted that the number of SVs in the derived collapsed variants was problematic because some genes had many SVs, whereas others only had one. Taking this one step further, Cantor and Wilcox [2011] reported a variety of methods for aggregating SVs into haplotypes and multiallelic genotypes in the Group 13 summary.

Results Matter

Based on the estimated locus-specific heritabilities for Q1 (Tables I and II), it was not surprising that many methods found causal SVs in the *FLT1* and *KDR* genes in most replicates in both the population-based and the family-based study designs. The estimated locus-specific heritabilities for SVs in these genes were greater than 0.03 and 0.02 in the population- and family-based samples, respectively. Melton and Pankratz [2011] and Bailey-Wilson et al. [2011], for example, reported that most of the methods considered were able to detect the SVs in *FLT1* and *KDR* in the population-based sample, but, as in other methods, a substantial increase in type I error over the nominal rate was observed. In the family-based sample, however, SVs in *VEGFC* and *VEGFA* had the highest locus-specific heritabilities (>0.2) followed by SVs in *FLT1* and *KDR* (Table II). This was primarily due to founder effects and the subsequent segregation of the SVs in the families, both largely the result of chance. The locus-specific heritabilities for most of the other causal variants were quite low, generally less than about 0.003, because either the allele frequency was low or the allelic effect size was low, or both.

In Q2, the estimated locus-specific heritabilities were nearly all less than 0.01 in the population-based sample and, except for SVs in *SIRT1* and *VLDLR*, all less than 0.001 in the family-based sample. In general, SVs with locus-specific heritabilities less than 0.01 were not detectable with any substantial degree of power. There were no causal SVs in Q4.

Power and Type I Error Matter

Many of the proposed new methods were evaluated by counting the number of times a true causal variant was identified as a causal variant over all replicates. It should be emphasized that, because of the nonindependence of the samples, this estimate was a surrogate for power (i.e., the proportion of samples that identified the variant as causal) rather than a true estimate of the power of the test. Furthermore, the confidence intervals on these estimates for 200 replicates were quite large, making comparisons across methods problematic, and the confidence intervals would be even larger for smaller critical values.

Similarly, just as the estimate of the proportion of samples that identified a true causal variant as a causal variant was taken to be a surrogate for power, the proportion of samples that identified a noncausal variant as a causal variant was taken to be a surrogate for type I error. There are, however, two null hypotheses that can be tested. Under the null hypothesis of no genetic component, no variants influence the phenotype; that is, the phenotype is essentially a normally distributed random variable (i.e., trait Q4). Under the alternative null hypothesis, no causal variants are among the set of variants considered; however, other unknown causal variants do, in fact, influence the phenotypic distribution (e.g., Q1, Q2, and the underlying liability of the discrete trait). These variants affect the correlation structure of the phenotypes and can be correlated with other known or unknown variants.

In general, the estimate of the type I error rate was elevated over a broad range of methods and in both the family and unrelated individuals data for traits Q1 and Q2. Even stringent critical values resulted in more observed type I errors than expected at the corresponding nominal critical value. This is relevant for the estimates of power as well, because the power of a test is related to its type I error rate (i.e., the power of the test at the null hypothesis is, in fact, the type I error rate). This may be due to the alternative null hypothesis problem, with SV correlations inflating both the type I error and power. König et al. [2011], Sun et al. [2011], and Sung et al. [2011] noted that the expected type I error rate for Q4, a trait with no causal variants, was close to its expected nominal rate. When there was no genetic component in the simulated model, most of the methods appeared to be statistically valid such that the estimated type I error rate was close to the nominal rate. This was clearly not the case for traits Q1 and Q2. This suggests that the inflated type I errors are due to correlations in the data when there is any underlying genetic component in the model. These correlations may be due to chance, the quality of the sequence alignment, linkage disequilibrium, gametic disequilibrium, or other unknown structural genetic relationships. Bailey-Wilson et al. [2011], Cantor and Wilcox [2011], Sun et al. [2011], Tintle et al. [2011], and Ye and Engelman [2011] all noted that in the population-based samples, the power to detect associations was not high, even for those SVs with the highest locus-specific heritabilities (e.g., *FLT1* and *KDR* in trait Q1), and that the type I error rate was inflated, sometimes markedly so. Tintle et al. [2011] postulated that the elevation in type I error rates could be due to either population stratification or correlations between SVs on different chromosomes. They noted that principal components analysis helps to ameliorate population stratification, although it also resulted in the loss of rare SVs. In general, methods that analyzed the population-based samples found some of the causal SVs, often those with the higher locus-specific heritabilities, but the power to detect associations with rare SVs was quite low, and the proportion of type I errors was substantially elevated over the nominal levels.

Study Design Matters

Because the simulation study provided both population- and family-based samples, the GAW17 participants used a wide range of sampling strategies. Study designs ranged from population-based case-control studies for the disease to loaded family-based designs with extended families for the quantitative traits. Methods of analysis ranged from simple linear regression, to intrafamilial tests of association in parent-offspring trios and in nuclear and extended families, to linkage analysis in nuclear and extended families. Kazma and Bailey [2011] presented the difference in results obtained from different sampling strategies and methods of analysis. Although both population- and family-based studies found SVs in *FLT1* and *KDR* for trait Q1, only the family-based intrafamilial tests of association and the linkage method found SVs in *VEGFC* and *VEGFA*. The estimated locus-specific heritabilities in the family-based sample were about 0.2 for both *VEGFC* and *VEGFA* and were considerably larger than for the SVs in *FLT1* and *KDR*; these are, in fact, the top seven SVs in Table II. It is clear that family-based designs are more effective for detecting rare variants with large effects in *families*.

Incorporating Prior Information Matters

Not unexpectedly, including prior information when it is, in fact, involved in the etiology of a trait helps to identify causal SVs. This information includes relevant covariates, correlations among SVs, and external knowledge of genes and pathways involved in the expression of the trait. As in a GWAS, incorporating information about population substructure and linkage disequilibrium can also help to identify causal SVs. For example, information about coding regions for proteins and nonsynonymous versus synonymous base substitutions can be used to help formulate collapsing strategies that could increase the power to detect rare variants. Similarly, information from noncoding elements, alternative splice sites, and gene pathways can help to identify functional domains or gene sets that can be used to increase the power to detect causal variants. In the Group 3 summary, Chen et al. [2011] reported that considering only nonsynonymous SVs, for example, increased the power of the test. Both Chen et al. [2011] and Namkung et al. [2011] reported that using different types of information improved the detection of some of the causal SVs and that in at least some situations, the power to detect associations was somewhat better for genes and pathways than for individual SVs.

Multiple Testing Still Matters

As noted by König et al. [2011], problems associated with multiple testing are magnified in next-generation sequence data because the number of SVs greatly exceeds the number of SNPs in a GWAS. The GAW17 contributors considered several approaches to adjust for multiple tests, including the traditional Bonferroni test, resampling methods, and reduction of dimensionality. Other contributors focused on model fitting (e.g., machine learning approaches) rather than on adjusting for multiple tests. Although some of the methods were quite innovative, no single method performed substantially better than the others or better than the methods used in genome-wide association studies.

Computational Issues Matter

It should be acknowledged that there is substantial additional computational burden for methods that use families, maximum-likelihood estimation, any form of penalized regression, and machine learning. Any method that uses an iterative process, a process with high dimensionality, or permutation testing will increase the computational time, perhaps only marginally for each SV, but the cumulative computational burden over millions of SVs may make the analysis impractical with present-day hardware and software [Bickeböller et al., 2011]. Extremely computationally intensive methods may not be feasible with today's current infrastructure, and analysis of whole-exome sequence data will have to be able to be completed in hours, days, or weeks, not years. Other bottlenecks include data storage capacity and data transfer rates. Care must be taken to ensure that proposed new methods are computationally tractable, and there may have to be a trade-off in terms of the appropriateness of the method and the computational time required to analyze all the sequence variants in the entire genome.

Discussion

The contributions from the GAW17 participants encompass a wide variety of study designs, approaches, and statistical methods with the primary focus on detecting effects in SVs generated from next-generation sequencing data. In general, many of the statistical genetic methods used in the population-based sample consistently found causal SVs when the estimated locus-specific heritability, as measured in the population-based sample, was greater than about 0.08; and sometimes, SVs with heritabilities as low as 0.03 were identified. However, virtually no SVs with locus-specific heritabilities less than 0.03 were consistently identified. It is important to note that, as the locus-specific heritability approached 0, the power to detect an allelic effect approached the type I error rate. Thus it should come as no surprise that the power to detect SVs with locus-specific heritabilities less than about 0.03 was not substantially greater than the type I error rate and that these SVs were, for the most part, undetectable using population-based samples.

In the family-based samples, many of the methods consistently detected SVs that were much rarer than those detected in the population-based samples, but the locus-specific heritabilities for these rare SVs, as measured in the family-based samples, were substantially higher than their corresponding locus-specific heritabilities in the population-based samples. For example, the locus-specific heritability for C4S4935 in *VEGFC* was 0.25 in the family-based samples but was only about 0.01 in the population-based samples. Although this SV was consistently detected using the family-based samples, it was rarely detected in the population-based samples. It is clear that in order to detect rare SVs, enrichment for the rare SVs is required, as demonstrated in GAW17 with a study design based on large, highly loaded families.

The persistent inflation of the type I error rate over a wide variety of statistical methods is more troublesome. Although many of the contributors found little inflation in type I error for Q4, a trait with no causal SVs, type I error rates for Q1 and Q2 were well above their nominal levels with the inflation for Q1 being higher than that for Q2. It seems likely that this inflation in type I error is due to correlations among SVs, both causal and noncausal, either within linkage disequilibrium blocks or across chromosomes.

Finally it should be remembered that we are in the early days of the development of statistical genetic analysis methods that are appropriate for SVs from next-generation sequencing data and that the challenges, problems, and pitfalls raised by these methods are eerily similar to those from the early days of the analysis of protein polymorphisms, restriction fragment length polymorphisms, short tandem repeat polymorphisms, and SNPs. What is clear is that further development is necessary for sampling methods that enrich rare SVs (e.g., family-based methods or the ClinSeq approach, which uses a population-based approach with the ability to recruit family members of individuals with rare SVs [Biesecker et al., 2009]), for strategies that collapse or aggregate rare variants into biologically meaningful derived variables, for statistical methods that are robust with respect to correlations between SVs [e.g., Sung et al., 2011], for methods that are robust with respect to causal SVs that act as outliers, and for methods that incorporate information from both families and unrelated individuals.

Acknowledgments

The Genetic Analysis Workshops are supported by National Institutes of Health grant R01 GM031575. This summary was supported, in part, by the Division of Intramural Research, National Human Genome Research Institute, National Institutes of Health; and an intramural grant from the University of Lübeck, Lübeck, Germany. We would like to thank Dr. Cristina Justice, Mera Krishnan, and Alexa J. M. Sorant for their efforts in determining the locus-specific heritabilities for the family- and population-based samples, and Drs. Joan Bailey-Wilson, Jean MacCluer, and Laura Almasy for their helpful suggestions and comments.

References

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

Bailey-Wilson JE, Brennan JS, Bull SB, Culverhouse R, Kim Y, Jiang Y, Jung J, Li Q, Lamina C, Liu Y, et al. Regression and data mining methods for analyses of multiple rare variants in the Genetic Analysis Workshop 17 mini-exome data. Genet Epidemiol. 2011 X(suppl X):X–X.

Bickeböller H, Houwing-Duistermaat JJ, Wang X, Yan X. Dealing with high dimensionality for the identification of common and rare variants as main effects and for gene-environment interaction. Genet Epidemiol. 2011 X(suppl X):X–X.

Biesecker LG, Mullikin JC, Facio FM, Turner C, Cherukuri PF, Blakesley RW, Bouffard GG, Chines PS, Cruz P, Hansen NF, et al. The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine. Genome Res. 2009; 19:1665–1674. [PubMed: 19602640]

Cantor RM, Wilcox M. Detecting rare variant associations: methods for testing haplotypes and multiallelic genotypes. Genet Epidemiol. 2011 X(suppl X):X–X.

Chen GK, Wei P, DeStefano AL. Incorporating biological information into association studies of sequencing data. Genet Epidemiol. 2011 X(suppl X):X–X.

Falconer, DS.; Mackay, TFC. Introduction to Quantitative Genetics. Essex, UK: Pearson Education; 1996.

Hemmelmann C, Daw EW, Wilson AF. Quality control issues and the identification of rare functional variants with next-generation sequencing data. Genet Epidemiol. 2011 X(suppl X):X–X.

Hinrichs AL, Suarez BK. Incorporating linkage information into a common disease/rare variant framework. Genet Epidemiol. 2011 X(suppl X):X–X.

Kazma R, Bailey JN. Population-based and family-based designs to analyze rare variants in complex diseases. Genet Epidemiol. 2011 X(suppl X):X–X.

Kent JW Jr. Rare variants, common markers: synthetic association and beyond. Genet Epidemiol. 2011 X(suppl X):X–X.

König IR, Nsengimana J, Papachristou C, Simonson MA, Wang K, Weisburd JA. Multiple testing in high-throughput sequence data: experiences from Group 8 of Genetic Analysis Workshop 17. Genet Epidemiol. 2011 X(suppl X):X–X.

Lamina C. Digging into the extremes: a useful approach for the analysis of rare variants with continuous traits? BMC Proc. 2011; 5 suppl 9:S105.

Melton PE, Pankratz N. Joint analyses of disease and correlated quantitative phenotypes using next-generation sequencing data. Genet Epidemiol. 2011 X(suppl X):X–X.

Namkung J, Raska P, Kang J, Liu Y, Lu Q, Zhu X. Analysis of exome sequences with and without incorporating prior biological knowledge. Genet Epidemiol. 2011 X(suppl X):X–X.

Pugh EW, Papanicolaou GJ, Justice CM, Roy-Gagnon M-H, Sorant AJM, Kingman A, Wilson AF. Comparison of variance components, ANOVA, and regression of offspring on mid-parent (ROMP) methods for SNP markers. Genet Epidemiol. 2001; 21 suppl 1:S794–S799. [PubMed: 11793780]

Roy-Gagnon M-H, Mathias RA, Wilson AF. Application of the regression of offspring on mid-parent (ROMP) method to detect associations between SNPs and the beta 2 EEG phenotype in the COGA data. BMC Genet. 2005; 6 suppl 1:S56. [PubMed: 16451668]

Stram AH. Comparing nominal and real quality scores on next-generation sequencing genotype calls. BMC Proc. 2011; 5 suppl 9:S14.

Sun YV, Sung YJ, Tintle N, Ziegler A. Identification of genetic association of multiple rare variants using collapsing methods. Genet Epidemiol. 2011 X(suppl X):X–X.

Sung H, Kim Y, Cai J, Cropp CD, Simpson CL, Li Q, Perry BC, Sorant AJM, 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.

Thomas A, Abel JH, Di Y, Faye LL, Jin J, Liu J, Wu Z, Paterson AD. Effect of linkage disequilibrium on the identification of functional variants. Genet Epidemiol. 2011 X(suppl X):X–X.

Tintle N, Aschard H, Hu I, Nock N, Wang H, Pugh E. Inflated type I error rates when using aggregation methods to analyze rare variants in the 1000 Genomes Project exon sequencing data in unrelated individuals: summary results from Group 7 at Genetic Analysis Workshop 17. Genet Epidemiol. 2011 X(suppl X):X–X.

Ye KQ, Engelman CD. Detecting multiple causal rare variants in exome sequence data. Genet Epidemiol. 2011 X(suppl X):X–X.

Table I

Estimates of mean locus-specific heritability ± standard deviation for the population-based samples averaged over all 200 replicates for Q1 and Q2 in descending order

Trait	Causal sequence variant	Gene	h_L^2
Q1	C13S523	*FLT1*	0.152 ± 0.025
	C13S522	*FLT1*	0.083 ± 0.018
	C13S524	*FLT1*	0.037 ± 0.013
	C4S1877	*KDR*	0.031 ± 0.013
	C4S1889	*KDR*	0.031 ± 0.013
	C13S431	*FLT1*	0.028 ± 0.011
	C4S1884	*KDR*	0.025 ± 0.011
	C1S6533	*ARNT*	0.021 ± 0.009
	C4S1878	*KDR*	0.020 ± 0.008
	C14S1734	*HIF1A*	0.017 ± 0.008
	C4S4935	*VEGC*	0.011 ± 0.007
	C5S5133	*FLT1*	0.007 ± 0.005
	C6S2981	*VEGFA*	0.005 ± 0.005
	C4S1861	*KDR*	0.005 ± 0.004
	C4S1887	*KDR*	0.004 ± 0.004
	C4S1874	*KDR*	0.004 ± 0.004
	C4S1873	*KDR*	0.004 ± 0.003
	C1S6542, C4S1890, C14S1729, C19S4831, C1S3181, C13S479, C13S514, C1S6537, C5S5156, C19S4815, C13S505, C4S1879, C13S320, C14S1718, C1S3182, C19S4799, C1S6561, C1S6540, C13S547		≤0.003 ± ≤0.003
Q2	C6S5380	*VNN1*	0.016 ± 0.008
	C6S5449	*VNN3*	0.010 ± 0.007
	C10S3050	*SIRT1*	0.010 ± 0.007
	C8S442	*LPL*	0.010 ± 0.007
	C6S5441	*VNN3*	0.009 ± 0.006
	C11S5292	*PDGFD*	0.009 ± 0.006
	C3S4875	*BCHE*	0.007 ± 0.006
	C2S354	*GCKR*	0.006 ± 0.008
	C3S4859	*BCHE*	0.006 ± 0.005
	C12S211	*VWF*	0.006 ± 0.007
	C3S679	*RARB*	0.005 ± 0.005
	C10S3048	*SIRT1*	0.005 ± 0.005
	C17S1024	*SREBF1*	0.005 ± 0.005
	C3S4873	*BCHE*	0.004 ± 0.005
	C6S5378	*VNN1*	0.004 ± 0.004
	C8S1741	*PLAT*	0.004 ± 0.004
	C9S377	*VLDLR*	0.004 ± 0.004
	C9S376	*VLDLR*	0.004 ± 0.005

Trait	Causal sequence variant	Gene	h_L^2
	C17S1046	*SREBF1*	0.004 ± 0.004
	C8S530	*LPL*	0.004 ± 0.004
	C10S3092, C3S4836, C11S5301, C8S476, C6S5446, C3S4869, C8S1758, C3S635, C3S4876, C9S444, C17S1007, C17S1055, C6S5448, C17S1043, C11S5302, C12S181, C3S4874, C8S1742, C11S5299, C3S4862, C8S1799, C9S497, C9S430, C10S3110, C3S4867, C10S3107, C8S1772, C17S1045, C17S1056, C9S443, C17S1009, C9S367, C17S1030, C3S4880, C9S391, C6S5412, C3S4856, C3S4860, C8S1811, C17S1048, C8S1770, C10S3108, C6S5426, C10S3093, C8S1773, C10S3058, C10S3109, C7S5133, C7S5144, C7S5132, C3S4834		$\leq 0.003 \pm \leq 0.004$

Table II

Estimates of mean locus-specific heritability ± standard deviation for the family-based samples averaged over all 200 replicates for Q1 and Q2 in descending order based on ROMP result

	Causal SV	**Gene**	h_L^2 **(ROMP)**	$h_L^2=2p(1-p)a^2$
Q1	C4S4935	*VEGFC*	0.250 ± 0.066	0.182 ± 0.040
	C6S2981	*VEGFA*	0.230 ± 0.057	0.173 ± 0.036
	C13S523	*FLT1*	0.046 ± 0.018	0.042 ± 0.014
	C4S1878	*KDR*	0.044 ± 0.026	0.030 ± 0.014
	C4S1884	*KDR*	0.018 ± 0.016	0.020 ± 0.011
	C13S431	*FLT1*	0.017 ± 0.018	0.019 ± 0.012
	C13S522	*FLT1*	0.008 ± 0.008	0.014 ± 0.008
	C19S4831	*HIF3A*	0.001 ± 0.007	0.004 ± 0.004
	C13S514	*FLT1*	0.000 ± 0.003	0.002 ± 0.002
	C4S1861	*KDR*	0.000 ± 0.003	0.002 ± 0.002
	C4S1873	*KDR*	0.000 ± 0.004	0.002 ± 0.002
	C1S6533	*ARNT*	−0.001 ± 0.007	0.003 ± 0.004
	C1S6540	*ARNT*	−0.001 ± 0.003	0.001 ± 0.002
	C13S320	*FLT1*	−0.001 ± 0.005	0.002 ± 0.002
	C1S3181	*ELAVL4*	*	0.002 ± 0.002
	C4S1890	*KDR*	*	0.001 ± 0.002
	C14S1734	*HIF1A*	*	0.001 ± 0.002
Q2	C10S3109	*SIRT1*	0.104 ± 0.037	0.017 ± 0.012
	C9S444	*VLDLR*	0.058 ± 0.022	0.012 ± 0.009
	C8S442	*LPL*	0.008 ± 0.010	0.026 ± 0.014
	C17S1045	*SREBF1*	0.008 ± 0.012	0.004 ± 0.006
	C6S5380	*VNN1*	0.005 ± 0.012	0.025 ± 0.014
	C17S1043	*SREBF1*	0.005 ± 0.008	0.013 ± 0.010
	C6S5441	*VNN3*	0.003 ± 0.008	0.014 ± 0.009
	C17S1024	*SREBF1*	0.003 ± 0.007	0.003 ± 0.004
	C11S5292	*PDGFD*	0.002 ± 0.005	0.003 ± 0.004
	C6S5449	*VNN3*	0.001 ± 0.006	0.006 ± 0.006
	C6S5378	*VNN1*	0.001 ± 0.005	0.005 ± 0.005
	C9S376	*VLDLR*	0.000 ± 0.007	0.009 ± 0.009
	C3S4880	*BCHE*	0.000 ± 0.003	0.004 ± 0.004
	C10S3093	*SIRT1*	0.000 ± 0.004	0.002 ± 0.002
	C6S5426	*VNN3*	0.000 ± 0.005	0.002 ± 0.003
	C2S354	*GCKR*	0.000 ± 0.003	0.002 ± 0.002
	C6S5439	*VNN3*	0.000 ± 0.003	0.001 ± 0.002
	C3S4874	*BCHE*	−0.001 ± 0.003	0.006 ± 0.006
	C8S476	*LPL*	−0.001 ± 0.004	0.003 ± 0.003
	C3S4834	*BCHE*	−0.001 ± 0.003	0.001 ± 0.002
	C8S1811	*PLAT*	−0.001 ± 0.005	0.002 ± 0.003

Causal SV	Gene	h_L^2 **(ROMP)**	$h_L^2=2p(1-p)a^2$
C10S3050	*SIRT1*	-0.002 ± 0.006	0.004 ± 0.004
C3S4856	*BCHE*	*	0.001 ± 0.002
C8S1773	*PLAT*	*	0.002 ± 0.003
C10S3108	*SIRT1*	*	0.002 ± 0.002
C12S181	*VWF*	*	0.002 ± 0.003
C12S211	*VWF*	*	0.002 ± 0.003
C17S1009	*SREBF1*	*	0.002 ± 0.002
C17S1048	*SREBF1*	*	0.001 ± 0.002

*Insufficient variation to calculate the locus-specific heritability.