Published in final edited form as:

Genet Epidemiol. 2011 ; 35(Suppl 1): S80–S84. doi:10.1002/gepi.20655.

Rare Variants, Common Markers: Synthetic Association and Beyond

Jack W. Kent Jr.¹

¹Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX

Abstract

The phenomenon of synthetic association raises the possibility that common variant genetic markers may be coupled with functional rare variants sufficiently often to allow the rare variants to be tagged by the common ones. Using human exome sequence data from the 1000 Genomes Project, two investigative teams in Group 12 of Genetic Analysis Workshop 17 found that stochastic coupling between rare and common variants does occur, although perhaps not sufficiently often that we can expect common variant signals to reflect synthetic association; other teams considered methods for detecting association using both rare and common variants. Common themes were that synthetic association is more apparent in population strata (ancestral or familial) and that careful selection of the unit of analysis (gene, gene network, or other genomic subset) is likely to be crucial to the discovery of rare variants that contribute to risk of disease.

Keywords

synthetic association; rare variants; association; identity by state

Introduction

The common disease/common variant (CDCV) hypothesis [Chakravarti, 1999] holds that genetic variants that contribute to disease risk but that are individually of small effect (and therefore subject to weak natural selection) can increase in allele frequency in human populations, eventually establishing a basis for population-wide genetic risk. The CDCV hypothesis leads to predictions that should facilitate positional gene discovery: Major risks should be ancestral rather than familial, allowing gene discovery in unrelated subjects; and ancestral risk haplotypes should be identifiable through analysis of a limited set of haplotype-tagging marker variants. The major predicted limitation on analytical power is sample size, which could be increased readily by recruiting unrelated case and control subjects.

Numerous haplotype-tagging markers have been associated with several common diseases, although (as CDCV predicts!) the effect size of such markers is typically small. More disturbing than small effect size is the relative lack of success in identifying *functional* variants in linkage disequilibrium with the markers [Dickson et al., 2010]. In response to these concerns, Dickson et al. [2010] proposed an alternative to the CDCV hypothesis: Much of the genetic variation contributing to disease may be due to rare variants, and multiplexes of these rare variants can be *synthetically associated* with common markers as a result of stochastic patterns of identity by descent. Dickson and colleagues showed, by

simulation and analysis of well-characterized monogenic diseases, that a marker may be synthetically associated with a disease if, by chance, enough rare functional variants in a contributing gene or gene pathway are coupled with the same marker allele in different lineages. Importantly, the physical distances involved in the coupling can be much larger than those covered by population-level linkage disequilibrium [Dickson et al., 2010].

The data provided to participants of Genetic Analysis Workshop 17 (GAW17) included preliminary exome sequence data from the 1000 Genomes Project — a snapshot of the real extent of rare and common variation present in the human genome [Almasy et al., 2011]. GAW17 Group 12 considered strategies for combining information from rare and common variants on several levels (Table I). Sun et al. [2011] and Kent et al. [2011] examined the extent of correlation between rare and common variant genotypes in the source data to address the question of whether common variants can tag rare ones. W. Chen et al. [2011] evaluated four collapsing methods; Cheng et al. [2011] and Uh et al. [2011] considered pathway approaches, seeking to maximize the genetic signal by considering variation in multiple biologically connected genes. Wang et al. [2011] outlined a complex, gene-based (although conceptually extensible to other groupings of single-nucleotide polymorphisms [SNPs]), joint rare variant/common variant approach that sought to control type I error rates through a screening step and principal-components-based reduction of the common variant data. Finally, Thalamuthu et al. [2011] presented a technique for using locus-specific genetic identity by state as a predictor of phenotype.

Results and Discussion

Correlation Between Rare and Common Variant Genotypes

Although the cost of comprehensive sequencing is dropping rapidly, genome-wide inventories of the full extent of genetic variation are not yet available for most genetic studies. The preliminary exome sequence data provided for GAW17 offered an early glimpse of the actual extent of rare and common variation, at least in coding regions. Both Sun et al. [2011] and Kent et al. [2011] sought to characterize the correlation structure of these variants independent of any phenotype as a basis for understanding the underlying potential for synthetic association. Notably, both work groups made use of the ethnic assignments provided with the 1000 Genomes Project data; although the genetic models for the GAW17 simulations did not explicitly incorporate population-specific effects [Almasy et al., 2011], ethnic differences in allele frequencies evidently were highly relevant to the potential for synthetic association, as predicted by Dickson et al. [2010].

Sun et al. [2011] focused on the 697 unrelated individuals with exome data. Rare variants were defined as SNPs with minor allele frequency (MAF) < 0.01 , and common variants were defined as SNPs with MAF > 0.05 . SNPs in complete linkage disequilibrium were reduced to a single representative SNP. Briefly, the exome data were divided into nonoverlapping 1-Mb bins, and in each bin the following procedure was iterated: Five rare variants were selected at random and “collapsed” such that an individual received a score of 1 if he or she had at least one minor allele for any of the five rare variants, or 0 otherwise. Each individual was given a score equal to the number of copies of the minor allele for each common variant in the bin, and a correlation was computed for each pair of rare variant and common variant scores. The distribution of correlation coefficients was examined across the exome. The mean correlation was low (< 0.1) overall and within ethnic strata, although the range of correlation coefficients was broader within strata — a point that I discuss shortly. In a separate analysis, Sun and colleagues found that multiple correlations between rare variant scores and scores for common variants in high linkage disequilibrium were stronger than those for individual common variants, after correcting for the additional degrees of freedom. Interestingly, common variants appeared to be somewhat better at tagging putatively

nonfunctional (synonymous) rare variants than at tagging nonsynonymous ones. Based on their observations, Sun and colleagues suggested that common variants were likely to have low capacity for tagging rare variants unless population substructure was taken into account.

Kent et al. [2011] approached the same question using a slightly different strategy: Annotated gene exomes were selected if they possessed at least 10 but not more than 50 variants, with both rare variants and common variants present. (Like Sun et al. [2011], Kent and colleagues defined rare variants by $MAF < 0.01$, but they defined common variants by $MAF > 0.1$.) Kent and co-workers' collapsing method calculated a sum of the number of copies of rare variant minor alleles within each selected gene, which was then correlated with the genotype score for a randomly chosen common variant within the same gene. The outcome measure was the probability that the correlation between rare variant genotypes and common variant genotypes was 0, as determined from likelihood ratio tests of the alternative hypothesis against this null hypothesis. Population structure was addressed not by stratification, as Sun et al. [2011] did, but by principal components correction [Price et al., 2006]. Kent and colleagues also examined correlations between rare variants and common variants in the simulated family genotype data, which was based on gene dropping from 202 founders chosen from the 697 unrelated individuals [Almasy et al., 2011].

There was no evidence in Kent et al.'s [2011] study for a significant difference in common variant/rare variant correlation in synonymous vs. nonsynonymous variants. Like Sun et al. [2011], Kent and colleagues found only modest evidence of correlation between common variant genotypes and rare variant genotypes in the unrelated individuals, and correlation was reduced by principal components correction for population stratification. The extent of correlation was substantially greater in the family data, with nominally significant ($P < 0.05$) correlation in about 40% of comparisons vs. 20–30% in the unrelated individuals. Kent et al. [2011] argued that these last two observations were consistent with the finding of Sun et al. [2011] that rare variant/common variant correlation is more evident within population strata, and they attributed the increased potential for synthetic association in either ethnic or family strata to founder effects.

Both of these studies agree that common variants *can* tag rare variants —that the conditions for synthetic association do exist in natural sequence data — but that the extent of such correlation is modest at best. This suggests that, although association with a common variant may represent the effects of synthetically associated functional rare variants, common variant markers are probably not an optimal tool for rare variant discovery. Furthermore, both studies indicate that taking relatedness (ancestral or familial) into account is likely to be important for detecting rare variants that contribute to common traits.

Beyond Synthetic Association

The synthetic association hypothesis — and rare variant analysis in general — assumes that different individuals may carry different rare mutations in a gene (or gene pathway) that have similar effects on a trait of interest. Because any one of these rare variants alone poses severe distributional problems for statistical analysis, most participants in GAW17 used some method to combine rare variant genotypes according to some rule. Synthetic association reflects a natural combinatorial accident that, as we have seen, may not occur frequently enough to make rare variant tagging by common variants a reliable expectation [Sun et al., 2011]. Other positional rules include gene- or pathway-based analysis (discussed later) and sliding-window methods.

Five contributions to Group 12 looked forward to a time when partial or complete sequence data will allow routine analysis of both rare variant and common variant data. W. Chen et al. [2011] compared published combinatorial methods: the combined multivariate and

collapsing (CMC) method [Li and Leal, 2008], two weighted-sum methods proposed by Madsen and Browning [2009], and a variable threshold method of Price et al. [2010] (for a review, see Dering et al. [2011]). Analysis was confined to nonsynonymous SNPs, which are assumed more likely to be functional in an exome sequence; individual gene exomes were analyzed as functional units. Power, estimated from the number of true findings compared to the known genetic models for 200 replicates, was low for all methods examined. Interestingly, the joint rare variant/common variant methods yielded inflated family-wise error rates (FWERs), whereas analyses using rare variants only controlled FWER [Chen et al., 2011]. Although this was not addressed by Chen and colleagues, the error inflation when common variants were included could reflect underlying synthetic association, among other possibilities.

Wang et al. [2011] compared various weighting methods for collapsed scores as well as logistic regression versus logistic kernel machine methods, extending previous methods [Cook and Forzani, 2008; Wu et al., 2010]. Like Chen et al. [2011], Wang and colleagues gave priority to nonsynonymous rare variants (although by differential weighting of synonymous and nonsynonymous variant scores rather than by excluding synonymous SNPs). Rare variant data were combined with common variant scores derived using a principal-components-based method; this method was described as accounting for SNP-SNP interaction and as reducing the multiple testing burden by reducing the dimension of the overall test, but presumably it also dealt with linkage disequilibrium among the common variant SNPs. A prescreening step was used to reduce the number of genes entering final analysis. Wang and colleagues characterized this also as a dimension-reducing step, but it evidently acted as a safeguard against false positives by eliminating SNP sets (genes) with no a priori evidence of association. The combination of prescreening and common variant data reduction helped to control the inflated type I error rate noted by Chen et al. [2011]. The unit of analysis was the gene, although Wang and co-workers suggested that their methods were generalizable to other SNP sets.

What Is the Unit of Analysis?

As noted, joint analysis of multiple rare variants is likely to be required to provide an adequate signal for association, which compels the investigator to choose a unit of analysis (gene, pathway, etc.) as a basis for grouping SNPs. Reasoning that the combined effect of multiple variants in gene pathways may be more apparent than their individual effects, Cheng et al. [2011] calculated pathway risk scores using biological pathways from a public database (<http://www.broad.mit.edu/gsea/msigdb/index.jsp>). The scores were developed stepwise as follows: A collapsed rare variant score was calculated based on the number of rare variant minor alleles within a pathway, and the effect of this score on disease status was estimated by logistic regression. Common variants within the pathway were selected using LASSO (least absolute shrinkage and selection operator) analysis in a model that adjusted disease status by age, sex, smoking behavior, and ethnicity. The rare variant and common variant scores were then weighted by their respective odds ratios for association with the trait and summed (the method allowed for the possibility that no common variant was contributory). Logistic regression was performed on the summed score; because the distribution of the test statistic could not be assumed, Cheng and colleagues determined significance levels using permutation. They did not find any rare variants that showed significant association with disease in univariate tests, as expected. Out of 1,126 pathways, 772 (69%) had at least one common variant chosen by the LASSO. The number of pathways showing at least suggestive empirical association was increased by the use of the combined risk score.

In a detailed comparison of methods, Uh et al. [2011] considered two strategies for grouping variants: pathway and gene. Their investigation was limited to the vascular endothelial

growth factor (VEGF) pathway, which was, by design, a primary functional pathway in the GAW17 simulations. Unlike other methods considered so far, some of the approaches considered by Uh et al. [2011] — the gene set ridge regression in association studies (GRASS) model [L. S. Chen, 2010] and Uh and co-workers' own empirical Bayes method [Houwing-Duistermaat et al., 2011] — addressed linkage disequilibrium between common variants within genes, which likely helped to control type I error. As part of this analysis, Uh and colleagues estimated the extent of correlation between rare variant genotypes and common variant genotypes, concluding that there was limited evidence that common variants could tag rare variants [Uh et al., 2011]. They found that the power of the various methods was modest overall and recommended weighting rare variants more heavily than common variants in combined tests.

The studies by Uh et al. [2011] and Cheng et al. [2011] highlight the importance of the choice of analytical unit in combinatorial methods. Uh and colleagues offer a particularly appealing strategy: identification of pathway followed by identification of gene(s), with attention to patterns of linkage disequilibrium in the genes [Uh et al., 2011]. Both the stepwise approach and correction for nonindependence of variants reduce the overall dimension of the hypotheses and, consequently, the burden of correction for multiple tests. A limitation to this approach, however, lies in the choice of pathway, which must be based on prior knowledge.

Relatedness

The methods described so far are based on the main effects of allele counts. Thalamuthu et al. [2011] used dense genotype information to estimate the genetic similarity of paired individuals in a gene or genomic region and compared this to phenotypic similarity (presumably these should be correlated if a genomic region carries variant(s) that cause the trait of interest). Thalamuthu and colleagues had previously presented a kernel-based association test (KBAT) for common variant genotypes [Mukhopadhyay et al., 2010]; for GAW17 they extended the method to include pooled counts of rare variants and differential effects in population strata. Based on the simulated phenotype data, they found that joint rare variant/common variant KBAT analysis, compared to analysis of common variants alone, provided a modest increase in power to detect causal genes with control of type I error rates [Thalamuthu et al., 2011]. A potential criticism of their approach is that collapsing of rare variant genotypes, although essential for most main-effects methods, may actually reduce the evidence for genetic similarity and may unnecessarily limit power in this case.

Although not explicitly described as such, Thalamuthu et al.'s [2011] method is a type of linkage analysis for unrelated individuals, based on identity by state (IBS) rather than identity by descent. As such, it points to a general class of methods that have the potential to address several of the themes raised by GAW17 Group 12 participants. A genetic similarity-based approach could make efficient use of the information provided by deep sequencing, drawing on the effects of multiple variants in a genomic region while remaining agnostic about the functionality (or even the direction of effect) of any single variant. This advantage would be offset, as in family-based linkage analysis, by loss of spatial resolution and the necessity of using other analytical means to identify putative functional variants (the ultimate test of functionality, of course, being experimental). As with other methods discussed by Group 12 participants, IBS-based methods would raise the question of defining the unit of analysis: Thalamuthu et al. [2011] used gene annotations to define genomic regions, but in principle their method could be applied to sliding windows, pathway gene sets, and so on. Finally, Thalamuthu and colleagues explicitly recognized the potential for differential effects of genetic similarity in population strata, as anticipated by Sun et al. [2011] and Kent et al. [2011].

Conclusions

To answer one of the starting questions for GAW17 Group 12: Yes, common variants can tag rare variants, but not very reliably. The diverse studies contributed by this group, like many in GAW17, tended to highlight the limitations of existing methods for analyzing rare variants more than inspire confidence that we fully understand how to use the dense genotype data that will be available in the near future. Two hopeful themes did emerge, however: that careful consideration of the unit of analysis can improve our detection of associated variants when prior biological information is available; and that enhanced estimates of genetic similarity, based on dense genotyping, may help to generate those a priori hypotheses.

Acknowledgments

I am grateful to all of the Group 12 investigators for their work and participation. Hae-Won Uh and Anbupalam Thalamuthu provided especially helpful comments and guidance in the discussions. The Genetic Analysis Workshops are supported by National Institutes of Health (NIH) grant GM031575, and my participation was supported by NIH grant MH059490.

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):S2.
- Chakravarti A. Population genetics: making sense out of sequence. *Nat Genet.* 1999; 456:56–60. [PubMed: 9915503]
- Chen LS, Hutter CM, Potter JD, Liu Y, Prentice RL, Peters U, Hsu L. Insights into colon cancer etiology using a regularized approach to gene set analysis of GWAS data. *Am J Hum Genet.* 2010; 86:860–71. [PubMed: 20560206]
- Chen W, Gao X, Wang J, Sun C, Wan W, Zhi D, Liu N, Chen X, Gao G. Evaluation of association tests for rare variants using simulated data sets in the Genetic Analysis Workshop 17 data. *BMC Proc.* 2011; 5(suppl 9):S86.
- Cheng L, Hu PZ, Sykes J, Pintilie M, Liu G, Xu W. A pathway-based association analysis model using common and rare variants. *BMC Proc.* 2011; 5(suppl 9):S85.
- Cook R, Forzani L. Principal fitted components for dimension reduction in regression. *Stat Sci.* 2008; 23:485–501.
- 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.
- Dickson SP, Wang K, Krantz I, Hokonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. *PLoS Biol.* 2010; 8:e1000294. [PubMed: 20126254]
- Houwing-Duistermaat JJ, Uh H-W, Tsonaka R. Pathway analysis for family data using nested random-effects models. *BMC Proc.* 2011; 5(suppl 9):S22. [PubMed: 21645302]
- Kent JW Jr, Farook V, Göring HHH, Dyer TD, Almasy L, Duggirala R, Blangero J. Do rare variant genotypes predict common variant genotypes? *BMC Proc.* 2011; 5(suppl 9):S87.
- Li B, Leal S. 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]
- Madsen B, Browning S. A groupwise association test for rare mutation using weighted sum statistics. *PLoS Genet.* 2009; 5:e1000384. [PubMed: 19214210]
- Mukhopadhyay I, Feingold E, Weeks DE, Thalamuthu A. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genet Epidemiol.* 2010; 34:213–21. [PubMed: 19697357]
- Price AL, Kryukov GV, Bakker PIW, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet.* 2010; 86:832–8. [PubMed: 20471002]

- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006; 38:904–9. [PubMed: 16862161]
- Sun X, Namkung J, Zhu X, Elston RC. Capability of common SNPs to tag rare variants. *BMC Proc.* 2011; 5(suppl 9):S88.
- Thalamuthu A, Zhao J, Keong GTH, Kondragunta V, Mukhopadhyay I. Association tests for rare and common variants based on genotypic and phenotypic measures of similarity between individuals. *BMC Proc.* 2011; 5(suppl 9):S89.
- Uh H-W, Tsonaka R, Houwing-Duistermaat JJ. Does pathway analysis make it easier for common variants to tag rare ones? *BMC Proc.* 2011; 5(suppl 9):S90.
- Wang R, Jie P, Pei W. SNP set analysis for detecting disease association using exon sequence data. *BMC Proc.* 2011; 5(suppl 9):S91.
- Wu M, Kraft P, Epstein M, Taylor D, Chanock S, Hunter D, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet.* 2010; 86:929–42. [PubMed: 20560208]

Table I

Comparison of published contributions to GAW17 Group 12

Contribution	Unit of analysis	Methods ^a	Phenotype	Data structure
W. Chen et al. [2011]	Gene	CMC, weighted sum, variable threshold	Disease status	Unrelated individuals
Cheng et al. [2011]	Functional pathway	Stepwise collapsing (rare variants), LASSO (common variants), combined logistic regression	Disease status	Unrelated individuals
Kent et al. [2011]	Gene	Correlation	N/A	Unrelated individuals and family-based
Sun et al. [2011]	1-Mb window	Correlation	N/A	Unrelated individuals
Thalamuthu et al. [2011]	Gene	KBAT	All	Unrelated individuals
Uh et al. [2011]	Functional pathway, gene	Global test, GRASS, empirical Bayes test, CMC	Q1 (quantitative trait), disease status	Unrelated individuals
Wang et al. [2011]	Gene	Logistic regression, kernel machines	Disease status	Unrelated individuals

^aSee text for explanation of methods. CMC, combined multivariate and collapsing analysis; LASSO, least absolute shrinkage and selection operator; KBAT, kernel-based association test; GRASS, gene set ridge regression in association studies.