



Practice of Epidemiology

Hypothesis-Driven Candidate Gene Association Studies: Practical Design and Analytical Considerations

Timothy J. Jorgensen, Ingo Ruczinski, Bailey Kessing, Michael W. Smith, Yin Yao Shugart, and Anthony J. Alberg

Initially submitted February 13, 2009; accepted for publication July 14, 2009.

Candidate gene association studies (CGAS) are a useful epidemiologic approach to drawing inferences about relations between genes and disease, especially when experimental data support the involvement of specific biochemical pathways. The value of CGAS is apparent when allele frequencies are low, effect sizes are small, or the study population is limited or unique. CGAS is also valuable for validating previous reports of genetic associations with disease in different populations. Despite the many advantages, the information generated from CGAS is sometimes compromised because of either inefficient study design or suboptimal analytical approaches. Here the authors discuss issues related to the study design and statistical analyses of CGAS that can help to optimize their usefulness and information content. These issues include judicious hypothesis-driven selection of biochemical pathways, genes, and single nucleotide polymorphisms, as well as appropriate quality control and analytical procedures for measuring main effects and for evaluating environmental exposure modifications and interactions. A study design algorithm using the example of DNA repair genes and cancer is presented for purposes of illustration.

cancer; data analysis; DNA repair; genetic epidemiology; genome-wide association study; haplotypes; polymorphism, single nucleotide; research design

Abbreviations: CGAS, candidate gene association study(ies); LD, linkage disequilibrium; MAF, minor allele frequency; SNP, single nucleotide polymorphism.

Currently, there are 2 primary research paradigms for population-based genetic association studies, both of which are based on genotyping of single nucleotide polymorphisms (SNPs)—the genome-wide association study and the candidate gene association study (CGAS). Both paradigms have been successfully employed to identify genotype associations with disease, but replication and validation have been pervasive problems. For example, an evaluation of 1,348 genetic association studies of Alzheimer's disease showed a highly significant excess ($P < 10^{-6}$) of studies reporting significant associations compared with the number "expected" (1). Some of this excess may be attributable to publication bias, but some is probably due to interstudy heterogeneity. For example, in a random sampling of Human Genome Epidemiology articles from 2001–2003, fewer

than half of the researchers reported testing for Hardy-Weinberg equilibrium, and 88% reported no adjustments for multiple comparisons (2). The topic of best practices for the design of CGAS has been inadequately addressed. Thus, our purpose in this paper is to address practical considerations in optimizing the design and implementation of CGAS.

CGAS is a deductive approach based on testing an a priori hypothesis that specific genes are associated with disease risk. CGAS thus provides a focused view of genomic regions of interest, hypothesized to be associated with a particular disease. Because CGAS is a hypothesis-driven approach, it allows for targeted evaluation of selected alleles in study populations relevant to the hypothesis. Markers can thus be concentrated in the candidate genes and can

Correspondence to Dr. Timothy J. Jorgensen, Department of Radiation Medicine, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, 3970 Reservoir Road NW, Washington, DC 20057 (e-mail: tjorge01@georgetown.edu).

purposefully target putative functional SNPs. Within the targeted candidate genes, this approach may confer inferential advantages in comparison with untargeted screening strategies such as genome-wide association studies, where coverage is spread across the whole genome and typically does not specifically target functional SNPs. This enhanced power is important when studying lower-frequency SNPs or when the study population is small (e.g., hospital-based) or unique (e.g., a rare disease with limited cases) (3). Furthermore, CGAS represents a cost-efficient approach for well-focused disease questions.

We raise here several issues worth considering in implementing CGAS. These include strategies to maximize the information gleaned using SNP selection and analysis approaches that capitalize on the hypothesis-driven nature of the SNP selection. Carefully addressing these issues may help to boost statistical power to detect weak associations and enhance the validity of the inferences drawn from CGAS. In this article, we describe an algorithm for SNP selection that incorporates into the selection process biochemical pathway knowledge, prior epidemiologic findings, platform constraints, analytical issues, and study aims. We also discuss analytic approaches that incorporate the study hypothesis and prior knowledge to stratify thresholds for statistical significance. As an illustration, we use the genes of DNA repair pathways as the candidates for investigation and cancer as the disease outcome.

CANDIDATE PATHWAYS VERSUS CANDIDATE GENES

Prior identification of specific candidate genes for investigation is the hallmark of CGAS. Currently, our knowledge of the functions of biochemical pathways is stronger than our understanding of the functions of individual genes, and we have new and better tools for assigning genes to functional pathways (4–7). Although complete knowledge of the universe of pathway genes can never be assumed, for many pathways *in vitro* reconstitution of functional activity, protein-protein interaction studies, and gene knockout experiments have helped identify the central players. Consequently, a productive strategy in CGAS is to hypothesize at the level of pathways and include all of the known genes in the pathway as candidate genes. Compared with studying individual genes, the inferences derived from a candidate pathway study are enhanced by allowing global conclusions about the association between an entire biochemical pathway and disease.

A wealth of rapidly evolving bioinformatic resources is now available to assist in pathway selection and prioritization (8–10). Additionally, a number of computer-based algorithms specifically designed for prioritization of putative disease-related genes have been developed (11–17). Many of these strategies are based on sequence comparisons with other genes with a known or suspected association with the outcome (i.e., “reference” or “training” genes). Identification of putative “new” candidate disease genes may, in turn, implicate “new” pathways associated with the disease. Thus, pathway and gene prioritization can be an iterative process, where pathway identification and prioritization im-

PLICATE new genes and gene prioritization based on sequence or other similarities can implicate new pathways. Many of these computer algorithms, however, have limited ability to integrate prioritizations emanating from different databases or to incorporate different prioritization strategies. A new bioinformatic program called Endeavor (18) has recently been developed that has several appealing features: 1) it uses multiple heterogeneous data sources, integrating them into a global ranking by means of order statistics; 2) it can be used to rank genes involved in both diseases and biologic processes; and 3) it provides user flexibility in database selection. This is an area of rapid development, and further advances in the area of computer-based pathway/gene prioritization can be anticipated.

Given the resources discussed above, known biochemical pathways can often be credibly prioritized as to their likely roles in the etiology of many diseases. Prioritizing pathways is a key to successful CGAS, since it identifies the specific candidate genes to be evaluated, drives the level of SNP coverage required for the genes, and dictates the thresholds for hypothesis testing. Each of these aspects is discussed below.

SNP SELECTION

Primarily because they logically suggest a mechanism for functional change, nonsynonymous and splice junction SNPs (i.e., “functional” SNPs) have been favored genetic markers for CGAS. Functional SNPs cause amino acid changes within a protein, which would be expected to alter the protein’s activity (19–21). Synonymous and noncoding SNPs, on the other hand, might affect function, but this effect would presumably be indirect, such as through altered transcription rates or message stability (22). However, before inferring that a functional SNP is biologically involved in disease causation, direct corroboration about changes in protein function is needed, because SNPs are often in high linkage disequilibrium (LD) with one another, and the SNP associated with the disease might only be a marker for an unmeasured functional SNP in LD. Thus, without corroborative evidence, disease associations with “functional” SNPs are in principle no more informative than any other SNPs.

While LD constrains inferences for any particular SNP, LD structure can also be exploited to allow inferences about nonmeasured SNP variants. For most genes, coverage of all common allelic variants can now be achieved by genotyping relatively few haplotype tagging SNPs. The International HapMap Project (23; <http://www.hapmap.org/>) has facilitated the haplotype tagging approach by measuring the LD between a selected subset of the known SNPs in different racial and ethnic reference populations. As seen in our example below, the 5 major DNA repair pathways—nucleotide excision repair, base excision repair, nonhomologous end joining, mismatch repair, and homologous recombination—are covered by 1,004 SNPs (see Table 1, priority ranks 1.1 and 2–5), representing an average of approximately 200 SNPs per DNA repair pathway. Examination of the KEGG pathway database (Kyoto Encyclopedia of Genes

Table 1. DNA Repair and Related Biochemical Pathways Prioritized for SNP Selection and a Summary of Their Gene, SNP, and Allele Counts^a

Priority Rank	DNA Repair and Related Pathways	No. of Genes	Gene Running Total	Total No. of Known SNPs	No. of Selected SNPs	Running SNP Total	No. of Tagged Alleles	Running Allele Total
1.1	Nucleotide excision repair	34	34	8,803	398	398	1,010	1,010
1.2	Confounder/modifier genes	30	64	8,115	278	676	1,223	2,233
2	Base excision repair	16	80	2,261	122	798	170	2,403
3	Nonhomologous end joining	7	87	3,913	119	917	333	2,736
4	Homologous recombination	20	107	8,020	184	1,101	1,197	3,933
5	Mismatch repair	10	117	4,632	181	1,282	482	4,415
6	DNA-DNA crosslink repair	8	125	3,050	84	1,366	320	4,735
7	DNA-protein crosslink repair	1	126	671	6	1,372	17	4,752
8	Direct reversal repair	3	129	1,618	31	1,403	397	5,149
9	Poly(ADP-ribose) polymerases	4	133	1,436	24	1,427	152	5,301
10	DNA damage signal transduction	8	141	1,757	56	1,483	151	5,452
11	Possible DNA repair syndrome	5	146	2,511	37	1,520	258	5,710
12	Chromatin metabolism	2	148	206	12	1,532 ^b	18	5,728
13	DNA synthesis	13	161	4,829	0			
14	Putative DNA repair genes	9	170	1,484	0			
15	Nucleases and ubiquitin genes	11	181	2,283	0			
16	Nucleotide pool genes	3	184	540	0			
	Total	184		56,129				

Abbreviations: ADP, adenosine diphosphate; CEPH, Utah residents with northern/western European ancestry (abbreviated CEU); SNP, single nucleotide polymorphism.

^a All genes in ranked pathways 1–12 (including pathways 1.1 and 1.2) were accommodated on the Illumina GoldenGate chip (Illumina, Inc., San Diego, California) (1,536-SNP capacity), representing 148 genes, 1,532 SNPs, and 5,728 “tagged alleles” (i.e., haplotype alleles defined by tagging SNPs). “Known SNPs” refers to those which appear in the database of the International HapMap Project (23; <http://www.hapmap.org/>). The Tagger SNP selection algorithm (35; Broad Institute, Cambridge, Massachusetts) with allele frequencies from the CEPH population data (i.e., HapMap CEU) was used for SNP selection. “Known SNPs” are those with a frequency greater than 0 in HapMap CEU.

^b Selection of SNPs was halted at pathway priority rank 12, so as not to exceed the capacity of the Illumina GoldenGate chip (i.e., 1,536 SNP genotyping slots).

and Genomes; <http://www.genome.ad.jp/kegg/pathway.html#cellular>) suggests that these 5 DNA repair pathways are similar in size and complexity to other human biochemical pathways. Assuming similar SNP densities among the various genes of the various pathways, it appears that most human pathways may be covered with a few hundred SNPs (24).

Easy-to-use algorithms for selecting tagging SNPs are available (25). One potential pitfall of these algorithms is the rigid use of a minimum minor allele frequency (MAF) as a selection parameter. The rationale for using a minimum MAF (typically 0.05) is that more common alleles confer greater statistical power, so that enriching for more common alleles increases the power to detect associations. However, rarer SNPs may have larger effect sizes (26), making omitting them potentially counterproductive. Furthermore, allelic frequencies are often not well established or may be based on a reference population that differs from the study population. Unless the MAFs are accurately known for the actual study population, it may be hazardous to pare SNPs purely on the basis of MAF thresholds. Alternatives include using tiered MAF cutoffs based on the priority of specific

candidate pathways or, if resources allow, assessing all validated SNPs, regardless of MAF.

The tagging SNP approach is not without limitations. For some genes in HapMap, haplotypes are defined by tagging SNPs selected on the basis of limited or fragmented LD data, making their ability to identify a true haplotype unreliable. This can result in multiple fragmented haplotype blocks that disproportionately consume tagging SNPs. In addition, SNP density among the genes in the HapMap database is not uniform, and the SNP information content among genes is biased toward genes that have attracted the most scientific interest. For example, *TP53*, which codes for the intensively studied p53 tumor suppressor gene, has greater than 6-fold more SNPs recorded in dbSNP (National Center for Biotechnology Information; <http://www.ncbi.nlm.nih.gov/projects/SNP/>) than the similar-sized but less studied *EXTL2* gene, which is required for the biosynthesis of heparin sulfate. Thus, the reliability of tagging SNPs in defining a genuine haplotype is also not uniform among the genes. To guard against genes with poor LD data disproportionately consuming SNP genotyping slots, inspecting the LD data for genes with the most numerous tagging SNPs is

recommended. Often missing or incomplete LD data make a gene appear fragmented into multiple LD blocks, giving the impression that many tagging SNPs are required for adequate coverage. If so, it may be prudent to drop the SNPs that are tagging unreliable blocks or drop tagging SNP coverage for the gene altogether in favor of incorporating more validated functional SNPs.

Another SNP selection criterion is the specification of sufficiently high genotyping call rates. High SNP call rates are essential for haplotyping success (27, 28). Call rates can be predicted a priori on the basis of prior genotyping experience with either the same SNP or SNPs with similar surrounding nucleotide sequences. The platform manufacturers use such prior information to produce “design scores.” A design score is an index of the probability of genotyping success for a specific SNP based on the DNA sequence and the required assay design parameters for the particular platform being used. Advances in genotyping technology, coupled with the use of the manufacturer’s latest design scores, have made SNP call rates as high as 99% the current norm.

ENVIRONMENTAL EXPOSURES

Testing for gene-environmental interactions is an important component of CGAS. The environmental risk factors for a disease may act at least partly via interactions with genetic risk determinants, so the validity of the genetic findings will be enhanced by accounting for these environmental exposures as either potential confounders or effect modifiers. In fact, some genetic associations (e.g., DNA repair genes and cancer) may only be relevant in the presence of certain environmental exposures (e.g., DNA-damaging agents). If not accounted for, environmental risk factors will add imprecision and potentially bias the measures of association for genetic risk. Even a true association with a large effect in an environmentally exposed subgroup can be severely diluted toward the null if that environmental exposure is not well measured and appropriately incorporated into the analysis (29, 30). Depending on whether the environmental exposure is a potential confounder and/or effect modifier, it can be incorporated into the analyses by adjusting for it and/or assessing for a potential interaction between it and the genotypes of interest.

APPLICATION

For illustration, we apply some of the principles outlined above to a practical study design situation. Basal-cell carcinoma and squamous-cell carcinoma, referred to in combination as nonmelanoma skin cancer, are the most common human malignancies. A personal history of nonmelanoma skin cancer is predictive of recurrent nonmelanoma skin cancer and malignant melanoma. We have undertaken a CGAS to test the specific hypothesis that DNA repair gene variants underlie the genetic risk of the nonmelanoma skin cancer high-risk phenotype, because we believe that the risk may be related to suboptimal repair of DNA damage. However, it would also be possible to address alternative hypotheses grounded in other plausible biologic mechanisms (e.g.,

immunodeficiency (31, 32)). This study is being conducted within the Clue II cohort, a well-characterized community-based cohort with long-term follow-up, comprised predominantly of adult Caucasians residing in Washington County, Maryland (33).

Our approach has been to genotype SNPs in all of the genes in all of the known DNA repair pathways, as well as some other biochemical pathways that may contribute to DNA repair (e.g., DNA synthesis). To maximize the efficiency and informational content of SNP genotyping, we developed a study design algorithm that incorporates both functional and haplotype tagging strategies to measure genetic diversity; it also accounts for the platform constraints of the SNP genotyping technology used.

The first step was to ascertain from the literature 184 human DNA-repair and DNA-repair-related genes (34). These genes were assigned to biochemical pathways based on current knowledge of their putative activities, functions, sequence homology, or physical associations with other DNA repair proteins. The biochemical pathways were prioritized on the basis of the strength of the scientific evidence that they were linked to nonmelanoma skin cancer (Figure 1). This evidence included biochemical plausibility, as well as *in vitro* studies and prior epidemiologic reports. We further identified 30 non-DNA repair genes with putative or potential roles in nonmelanoma skin cancer. These “analytical control” SNPs were assigned high genotyping priority because of their value in interpreting the study findings. These SNPs represent epidemiologic controls, added to maintain the best epidemiologic practices. Any inferred SNP-disease association could be confounded in that both the SNP and the disease might each be independently associated with another unknown SNP. Any SNPs previously reported to be associated with the disease outcome, therefore, represent potential confounders of the newer findings, since they could be linked to the test SNP (through LD) and to the disease outcome. Likewise, a previously reported SNP association may influence the strength of the association between the test SNP and the disease, possibly through a genetic or biochemical interaction, and could therefore represent an effect modifier. Thus, SNPs previously reported to be associated with the disease outcome should also be considered candidates for effect modifiers. Omission of such analytical control SNPs from genotyping would preclude evaluating them as confounders and effect modifiers, illustrating that sound epidemiologic study design principles apply even with the application of the newest genotyping technologies. Analysis of the control SNPs should include tests for interactions between the genotype associations, as well as assessment of r^2 and D' to evaluate associations that might be merely a consequence of LD.

Among the 184 genes identified, all known nonsynonymous SNPs were selected for genotyping, regardless of MAF, because of their strong potential for functional protein modifications. Beyond these, haplotype tagging SNPs were also selected on the basis of HapMap data using the Tagger SNP selection program (35; Broad Institute, Cambridge, Massachusetts; <http://www.broadinstitute.org/mpg/tagger/>) with aggressive multimarker tests, to minimize the number of SNPs required to identify haplotypes. We used data on the

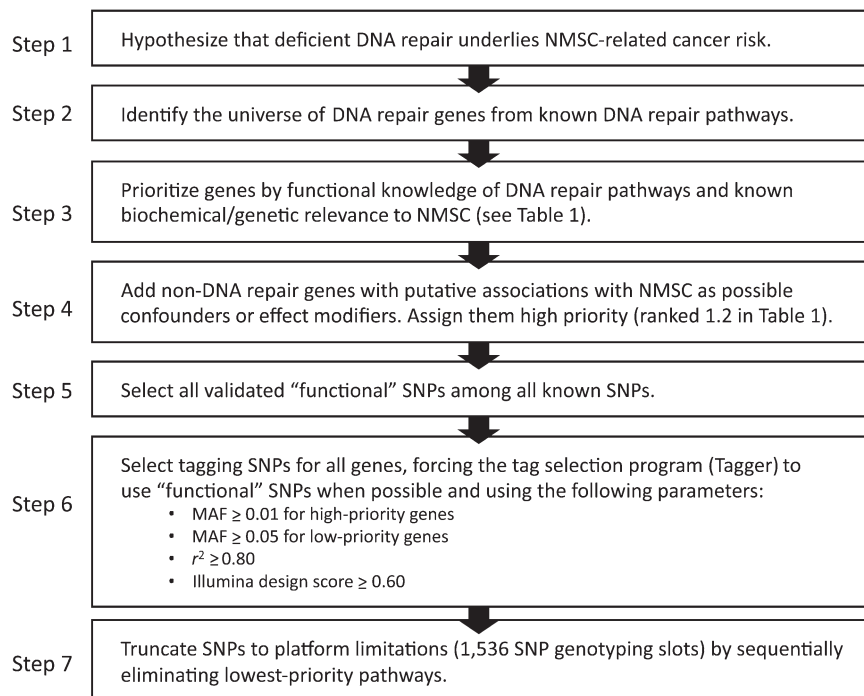


Figure 1. Flow diagram of the gene/single nucleotide polymorphism (SNP) selection process used in a candidate gene association study. Seven sequential steps identify candidate genes, prioritize them on the basis of hypothesis, select the most parsimonious combination of functional and tagging SNPs, and pare the final SNP count to the constraints of the platform. In this case, the platform is the Illumina GoldenGate chip (Illumina, Inc., San Diego, California), which has 1,536 SNP genotyping slots. See the “Application” section of the text for a full explanation of the gene/pathway prioritization algorithm. MAF, minor allele frequency; NMSC, nonmelanoma skin cancer. (The Tagger SNP selection program was produced by the Broad Institute, Cambridge, Massachusetts.)

CEPH population (Utah residents with northern/western European ancestry (abbreviated as CEU); <http://snp.cshl.org/citinghapmap.html>.en) from HapMap, because the Clue II population is more than 97% Caucasian. Selection of the correct reference population is not trivial and is an important design consideration, since haplotype blocks vary by ethnicity (36, 37). For example, on average, LD blocks are shorter in African Americans than in Caucasians, requiring more SNPs for African-American haplotype coverage. Using more markers in turn limits power, because of the necessity of correcting for more comparisons. Using an incorrect ethnic reference population for tagging SNP selection can thus compromise the validity of a CGAS.

To be parsimonious in selecting SNPs, the Tagger software was forced to select from the previously identified functional SNPs whenever feasible; thus, some SNPs served as both functional and tagging SNPs. For tagging SNPs, we used a tiered MAF cutoff. In general, a MAF cutoff greater than or equal to 0.05 was used, except among genes in the highest-priority pathways (e.g., nucleotide excision repair), where the cutoff was relaxed to MAF greater than or equal to 0.01 to maximize allelic coverage.

We used aggressive multimarker tests, with $r^2 \geq 0.8$. The 0.8 r^2 value was selected as a cutoff because it limited the total number of tagging SNPs required (i.e., a practical consideration), while incorporating some of the rarer alleles which might have relatively high effect sizes (i.e., a theoret-

ical consideration). Since power decreases as r^2 goes down, the selection of an r^2 cutoff is a tradeoff between minimizing the number of SNPs and maximizing power, given practical considerations of chip capacity and sample size. Choosing an exact r^2 is subjective, but $r^2 \geq 0.8$ is typically considered a reasonable value.

After we generated a list of SNPs for potential genotyping, it was truncated on the basis of pathway priority. Final results showed that all genes in priority pathways 1–12 (148 genes) could be accommodated with the chosen platform (Illumina GoldenGate panel; Illumina, Inc., San Diego, California) (Table 1). Coverage of bona fide DNA repair genes was virtually complete. Only priority pathways 13–16, representing genes marginally involved in DNA repair, were dropped (36 genes).

These results illustrate that virtually all known human DNA repair genes, as well as 30 non-DNA-repair genes previously implicated in nonmelanoma skin cancer, can be genotyped with 1,532 SNPs. These 1,532 SNPs cover 8.4×10^6 base pairs of total gene sequence at an average density of 1 SNP per 5,483 base pairs. The average number of SNPs per gene is 11, with a range from 1 to 36.

Coverage includes all validated functional SNPs with adequate design scores (i.e., ≥ 0.6) regardless of estimated allelic frequency, in addition to haplotype tagging for 115 of the 148 genes. Tagging allows for the potential identification of 5,483 haplotype alleles among these 115 genes,

which potentially permits inference on 30,107 known SNPs in LD within the haplotype blocks. The total frequency estimates of phased haplotypes for most of these genes, based on HapMap tagging SNPs, are in the range of 95%–100% coverage of the allelic variation of the genes within the population. Thus, by accounting for virtually all of the common diversity among DNA repair genes in the general population, this finite group of SNPs should provide sufficiently strong coverage of DNA repair pathway genes to permit a global test of the hypothesis that variant DNA repair genotypes may be the underlying explanation for the excess cancer risk seen among nonmelanoma skin cancer patients.

DATA QUALITY CONTROL AND STATISTICAL ANALYSIS

To assure the validity of the data and statistical inferences, issues related to the quality of the study samples and genomic data deserve consideration. Poor DNA quality is generally reflected in an unacceptably high fraction of failed genotyping reactions for the DNA sample SNPs, or elevated genotype uncertainty scores. The same applies to contaminated samples, often identified by an excess of heterozygote genotype calls. Relatedness between individuals can be assessed using the identity-by-state (IBS) method (38), and the violation of the independence assumption in the statistical analysis is typically addressed by either removing some of those persons or using statistical tests (such as generalized estimating equations (39)) that account for this dependence. Exploratory data analysis tools such as multidimensional scaling and principal-components analysis are easily employed and are powerful tools for detecting differences in genetic background and admixture (40). While persons with a very different genetic background than those in the study population are typically few and are usually removed from the analysis, the above tools also permit assessment of slight departures from the assumption of homogeneous genetic backgrounds in the study population, which could, if not accounted for, lead to spurious associations. (See reference 41 for a detailed discussion of sample quality control issues.)

Although SNPs with higher MAFs are typically favored in CGAS, as described above, some SNPs with low MAFs might also be deliberately included. This can yield some SNPs with little or no variation in the observed samples, rendering these SNPs noninformative. Furthermore, departures from Hardy-Weinberg equilibrium can be due to genotyping errors or population stratification and can lead to spurious associations (42). Thus, SNPs with a substantial departure from Hardy-Weinberg equilibrium are typically flagged to indicate that caution is needed when drawing inferences, taking into consideration the fact that chance departures from Hardy-Weinberg equilibrium do occur. Some SNPs are more difficult to genotype than others, because of substantial differences in the resolution of the raw fluorescent signals used in genotype calling algorithms (43). SNPs without sufficient overall genotype quality (e.g., many missing data) are often excluded from the analysis. The genotype uncertainty and missing genotype data can also

be addressed through imputation methods (44) or by directly including the genotype uncertainty in appropriate statistical tests (45, 46), which can properly quantify all available information and yield correct inferences without removing SNPs.

Similar to genome-wide association studies, the statistics assessing the associations between the individual SNPs and the outcome are of primary interest, and the type I error inflation due to multiple comparisons is typically addressed by procedures that control the family-wise error rate (such as the Bonferroni correction) or by determining the false-discovery rates and q values (47). The overall significance can be assessed using permutation tests, which leave the LD structure between SNPs intact and thus are less conservative than a Bonferroni correction (for example, see the article by Sull et al. (48)). That is, the threshold for overall significance will be lower, thereby increasing power. Further, when choosing SNPs to include in CGAS, researchers traditionally favor functional SNPs and SNPs with higher MAFs. This is based on the a priori scientific belief that those SNPs are more likely to be associated with the disease or that the statistical power to detect an existing association is higher with those SNPs. This prioritization can be objectively incorporated into the statistical analysis (49–51).

Some SNPs may be included in CGAS simply for validation; that is, they have been reported to be associated with the disease outcome in previous studies. In fact, identifying SNPs warranting consideration for validation studies has been simplified by a Web-based program called Gene Prospector (52), which selects genes with possible disease associations based on a highly curated and updated literature database of genetic association studies. This and other literature-based strategies provide a rich source of validation SNPs for potential inclusion in CGAS. Such validation SNPs differ from the SNPs included for discovery, allowing inferences to be drawn separately for these groups. Nevertheless, control of the family-wise error rate should be used for the validation SNPs. For these SNPs, disease associations have been previously observed, and thus a bound on the overall error rate (i.e., the probability of falsely rejecting at least 1 hypothesis) is highly desirable. As in a hypothesis test, failure to reach significance after, for example, a Bonferroni correction does not mean that the SNP is not associated with disease; it simply means that the SNP failed to validate. However, confidence of a true association would be greater for SNPs that reach significance after controlling for the family-wise error rate. The number of SNPs included for validation is typically small, and thus the overall significance level required would not be extreme. Further, validation SNPs are often from different genes and/or genomic regions and therefore not in high LD, giving rise to virtually independent test statistics. In these instances, the Bonferroni correction is not considered overly conservative.

CGAS allows for a targeted and focused view of particular regions of interest that are known to be or suspected of being associated with a particular phenotype. For example, SNP-SNP and SNP-environment interactions can readily be explored in CGAS. This is not only computationally feasible but also biologically meaningful, because the SNPs are often selected from specific genes in pathways of interest. In

addition, techniques for dimension reduction in association studies employing all SNPs within a gene or LD block are meaningful and scalable. For practical reasons, tagging SNPs are often favored in CGAS, because tagging SNPs typically capture the most variability in LD blocks; for a limited number of SNPs, tagging SNPs usually represent the best option (53). Other multilocus approaches that avoid phasing also exist. These methods employ the marginal test statistics or P values (such as the modified Fisher's inverse chi-square test allowing for correlation (54)) or truly multivariate approaches, using the genotype information to derive a block-specific test statistic directly (e.g., using the main principal components to assess the major axis of variation in an SNP block (55) or using score-based tests in the context of empirical Bayes models (56)).

CONCLUSION

CGAS is a powerful, hypothesis-driven epidemiologic approach that can contribute significantly to our understanding of the heritability of common diseases, particularly when preexisting biochemical data bolster the hypothesis. Further, CGAS remains the only feasible approach for studying small or unique populations. Nevertheless, CGAS sometimes wastes information because of an inefficient study design and suboptimal SNP selection strategies. In light of state-of-the-art technologic, bioinformatic, and statistical resources, we have emphasized genotyping and analysis strategies to strengthen the inferences that can be drawn from CGAS. With very few additional resources, the ability to infer associations can be enhanced. To maximize the potential impact of CGAS on improving public health, consideration ought to be given to ensuring well-conceived SNP selection strategies that take into account a priori knowledge of the relevant biochemical pathways and analytic strategies that logically flow from the prioritization used in the SNP selection strategy.

ACKNOWLEDGMENTS

Author affiliations: Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC (Timothy J. Jorgensen); Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland (Ingo Ruczinski); Laboratory of Genomic Diversity, National Cancer Institute, Frederick, Maryland (Bailey Kessing, Michael W. Smith); Genomic Research Branch, Division of Neuroscience, National Institute of Mental Health, Rockville, Maryland (Yin Yao Shugart); and Hollings Cancer Center and Department of Biostatistics, Bioinformatics, and Epidemiology, Medical University of South Carolina, Charleston, South Carolina (Anthony J. Alberg).

This work was supported in part by the National Cancer Institute (grant CA105069) and the National Heart, Lung, and Blood Institute (grant HL090577). Additionally, this project was supported in part with federal funds from the National Cancer Institute, under contract N01-CO-12400.

This research was also supported in part by the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Center for Cancer Research.

The content of this publication does not necessarily reflect the views or policies of the US Department of Health and Human Services, nor does the mention of trade names, commercial products, or organizations imply endorsement by the US government.

Conflict of interest: none declared.

REFERENCES

1. Kavvoura FK, McQueen MB, Khoury MJ, et al. Evaluation of the potential excess of statistically significant findings in published genetic association studies: application to Alzheimer's disease. *Am J Epidemiol.* 2008;168(8):855–865.
2. Yesupriya A, Evangelou E, Kavvoura FK, et al. Reporting of Human Genome Epidemiology (HuGE) association studies: an empirical assessment [electronic article]. *BMC Med Res Methodol.* 2008;8:31.
3. Wilkening S, Chen B, Bermejo JL, et al. Is there still a need for candidate gene approaches in the era of genome-wide association studies? *Genomics.* 2009;93(5):415–419.
4. Lesnick TG, Papapetropoulos S, Mash DC, et al. A genomic pathway approach to a complex disease: axon guidance and Parkinson disease [electronic article]. *PLoS Genet.* 2007;3(6):e98.
5. Bebek G, Yang J. PathFinder: mining signal transduction pathway segments from protein-protein interaction networks [electronic article]. *BMC Bioinformatics.* 2007;8:335.
6. Shriner D, Baye TM, Padilla MA, et al. Commonality of functional annotation: a method for prioritization of candidate genes from genome-wide linkage studies [electronic article]. *Nucleic Acids Res.* 2008;36(4):e26.
7. George RA, Liu JY, Feng LL, et al. Analysis of protein sequence and interaction data for candidate disease gene prediction [electronic article]. *Nucleic Acids Res.* 2006;34(19):e130.
8. Baxevas AD. The importance of biological databases in biological discovery. *Curr Protoc Bioinformatics.* 2006; Chapter 1:Unit 1.1.
9. Teufel A, Krupp M, Weinmann A, et al. Current bioinformatics tools in genomic biomedical research (review). *Int J Mol Med.* 2006;17(6):967–973.
10. Adriaens ME, Jaillard M, Waagmeester A, et al. The public road to high-quality curated biological pathways. *Drug Discov Today.* 2008;13(19-20):856–862.
11. Freudenberg J, Propping P. A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics.* 2002;18(suppl 2):S110–S115.
12. Perez-Iratxeta C, Bork P, Andrade MA. Association of genes to genetically inherited diseases using data mining. *Nat Genet.* 2002;31(3):316–319.
13. Turner FS, Clutterbuck DR, Semple CA. POCUS: mining genomic sequence annotation to predict disease genes [electronic article]. *Genome Biol.* 2003;4(11):R75.
14. Tiffin N, Kelso JF, Powell AR, et al. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res.* 2005;33(5):1544–1552.
15. Adie EA, Adams RR, Evans KL, et al. Speeding disease gene discovery by sequence based candidate prioritization [electronic article]. *BMC Bioinformatics.* 2005;6:55.

16. López-Bigas N, Ouzounis CA. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.* 2004;32(10):3108–3114.
17. Kent WJ, Hsu F, Karolchik D, et al. Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res.* 2005;15(5):737–741.
18. Aerts S, Lambrechts D, Maity S, et al. Gene prioritization through genomic data fusion. *Nat Biotechnol.* 2006;24(5):537–544.
19. Bao L, Cui Y. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics.* 2005;21(10):2185–2190.
20. Tian J, Wu N, Guo X, et al. Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines [electronic article]. *BMC Bioinformatics.* 2007;8:450.
21. Burke DF, Worth CL, Priego EM, et al. Genome bioinformatic analysis of nonsynonymous SNPs [electronic article]. *BMC Bioinformatics.* 2007;8:301.
22. Amcheslavsky A, Zou W, Bar-Shavit Z. Toll-like receptor 9 regulates tumor necrosis factor- α expression by different mechanisms. Implications for osteoclastogenesis. *J Biol Chem.* 2004;279(52):54039–54045.
23. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest.* 2008;118(5):1590–1605.
24. De Bakker PI, Graham RR, Altshuler D, et al. Transferability of tag SNPs to capture common genetic variation in DNA repair genes across multiple populations. *Pac Symp Biocomput.* 2006;11:478–486.
25. Stram DO. Tag SNP selection for association studies. *Genet Epidemiol.* 2004;27(4):365–374.
26. Gorlov IP, Gorlova OY, Sunyaev SR, et al. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet.* 2008;82(1):100–112.
27. Liu W, Yang T, Zhao W, et al. Accounting for genotyping errors in tagging SNP selection. *Ann Hum Genet.* 2007;71(pt 4):467–479.
28. Liu W, Zhao W, Chase GA. The impact of missing and erroneous genotypes on tagging SNP selection and power of subsequent association tests. *Hum Hered.* 2006;61(1):31–44.
29. Khoury MJ, Adams MJ Jr, Flanders WD. An epidemiologic approach to ecogenetics. *Am J Hum Genet.* 1988;42(1):89–95.
30. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet.* 2005;37(4):413–417.
31. Gerlini G, Romagnoli P, Pimpinelli N. Skin cancer and immunosuppression. *Crit Rev Oncol Hematol.* 2005;56(1):127–136.
32. Lu H, Ouyang W, Huang C. Inflammation, a key event in cancer development. *Mol Cancer Res.* 2006;4(4):221–233.
33. Helzlsouer KJ, Alberg AJ, Huang HY, et al. Serum concentrations of organochlorine compounds and the subsequent development of breast cancer. *Cancer Epidemiol Biomarkers Prev.* 1999;8(6):525–532.
34. Wood RD, Mitchell M, Lindahl T. Human DNA repair genes, 2005. *Mutat Res.* 2005;577(1-2):275–283.
35. de Bakker PI, Yelensky R, Pe'er I, et al. Efficiency and power in genetic association studies. *Nat Genet.* 2005;37(11):1217–1223.
36. Ouyang C, Krontiris TG. Identification and functional significance of SNPs underlying conserved haplotype frameworks across ethnic populations. *Pharmacogenet Genomics.* 2006;16(9):667–682.
37. Takeuchi F, Yanai K, Morii T, et al. Linkage disequilibrium grouping of single nucleotide polymorphisms (SNPs) reflecting haplotype phylogeny for efficient selection of tag SNPs. *Genetics.* 2005;170(1):291–304.
38. Weir BS, Anderson AD, Hepler AB. Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet.* 2006;7(10):771–780.
39. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics.* 1986;42(1):121–130.
40. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006;38(8):904–909.
41. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature.* 2007;447(7145):661–678.
42. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol.* 2001;60(3):155–166.
43. Carvalho B, Bengtsson H, Speed TP, et al. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics.* 2007;8(2):485–499.
44. Dai JY, Ruczinski I, LeBlanc M, et al. Imputation methods to improve inference in SNP association studies. *Genet Epidemiol.* 2006;30(8):690–702.
45. Marchini J, Howie B, Myers S, et al. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet.* 2007;39(7):906–913.
46. Plagnol V, Cooper JD, Todd JA, et al. A method to address differential bias in genotyping in large-scale association studies [electronic article]. *PLoS Genet.* 2007;3(5):e74.
47. Storey JD, Tibshirani R. Statistical significance for genome-wide studies. *Proc Natl Acad Sci U S A.* 2003;100(16):9440–9445.
48. Sull JW, Liang KY, Hetmanski JB, et al. Differential parental transmission of markers in *RUNX2* among cleft case-parent trios from four populations. *Genet Epidemiol.* 2008;32(6):505–512.
49. Genovese CR, Roeder K, Wasserman L. False discovery control with p -value weighting. *Biometrika.* 2006;93(3):509–524.
50. Roeder K, Devlin B, Wasserman L. Improving power in genome-wide association studies: weights tip the scale. *Genet Epidemiol.* 2007;31(7):741–747.
51. Roeder K, Bacanu SA, Wasserman L, et al. Using linkage genome scans to improve power of association in genome scans. *Am J Hum Genet.* 2006;78(2):243–252.
52. Yu W, Wulf A, Liu T, et al. Gene Prospector: an evidence gateway for evaluating potential susceptibility genes and interacting risk factors for human diseases [electronic article]. *BMC Bioinformatics.* 2008;9:528.
53. Chapman JM, Cooper JD, Todd JA, et al. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum Hered.* 2003;56(1-3):18–31.
54. Chapman J, Whittaker J. Analysis of multiple SNPs in a candidate gene or region. *Genet Epidemiol.* 2008;32(6):560–566.
55. Gauderman WJ, Murcray C, Gilliland F, et al. Testing association between disease and multiple SNPs in a candidate gene. *Genet Epidemiol.* 2007;31(5):383–395.
56. Goeman JJ, van de Geer SA, van Houwelingen HC. Testing against a high dimensional alternative. *J R Stat Soc Series B Stat Methodol.* 2006;68(3):477–493.