Candidate Gene Identification Approach: Progress and Challenges

Ms. N. Varsha[1] Ms. M. Varshinee[2] Mrs. V. Deepa[3]
[1,2]B.Sc Student [3]Assistant Professor
[1,2,3]Department of Computer Science
[1,2,3]PSGR Krishnammal College for Women, Coimbatore, India

Abstract— Gene expression profile analysis is the study of the way in which genes are transcribed to produce functional gene products (functional RNA species or protein products). There has been tremendous innovation in gene expression technologies, including high-throughput assays such as microarrays, and sequence-based techniques such as RNA-Seq. The Gene expressions are collected and analyzed for normal or disease genes. Using this project the system can diagnose types of diseases and abnormalities which may affect the user.

Keywords: Candidate gene approach, information bottleneck, digital candidate gene approach

I. INTRODUCTION

The candidate gene approach to conducting genetic association studies focuses on associations between genetic variation within pre-specified genes of interest and phenotypes or disease states. This is in contrast to genome-wide association studies, which scan the entire genome for common genetic variation. The candidate gene (CG) approach has been applied in plant genetics in the past decade for the characterisation and cloning of Mendelian and quantitative trait loci (QTLs). It constitutes a complementary strategy to map-based cloning and insertional mutagenesis.

II. 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 implicate 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.

III. 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.

IV. 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.

V. 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 r2 and D′ to evaluate associations that might be merely a consequence of LD.

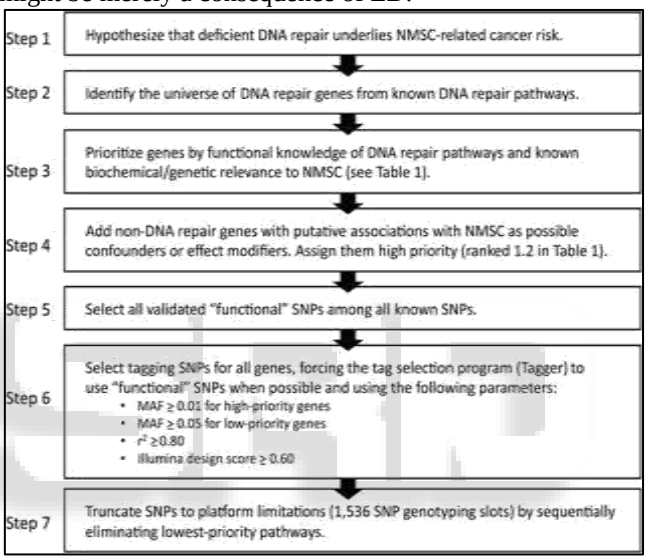

Fig. 1:

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 Golden Gate 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.)

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 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 $r2 \geq 0.8$. The 0.8 r2 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 theoretical consideration). Since power decreases as r2 goes down, the selection of an r2 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 r2 is subjective, but $r2 \geq 0.8$ 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 Golden Gate 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 × 106 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.

VI. 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.

References

[1] Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat Rev Genet. 2002; 3: 391-397.
[2] Fujii J, Otsu K, Zorzato F, et al. Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. Science. 1991; 253: 448-451.
[3] Johnson PL, McEwan JC, Dodds KG, et al. A directed search in the region of GDF8 for quantitative trait loci affecting carcass traits in Texel sheep. J Anim Sci. 2005;83:1988-2000.
[4] Bellamy R. Identifying genetic susceptibility factors for tuberculosis in Africans: a combined approach using a candidate gene study and a genome-wide screen. ClinSci (Lond). 2000; 98: 245-250.
[5] Grisart B, Coppieters W, Farnir F, et al. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res. 2002; 12: 222-231.
[6] Wayne ML, McIntyre LM. Combining mapping and arraying: An approach to candidate gene identification. Proc NatlAcadSci USA. 2002; 99: 14903-14906.
[7] Thaller G, Kuhn C, Winter A, et al. DGAT1, a new positional and functional candidate gene for intramuscular fat deposition in cattle. Anim Genet. 2003; 34: 354-357.
[8] Clop A, Marcq F, Takeda H, et al. A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nature Genetics. 2006; 38: 813-818.
[9] Stratil A, Geldermann H. Analysis of porcine candidate genes from selected QTL regions affecting production traits. AnimSci Pap Rep. 2004; 22: 123-125.

