Two Problems with Multiple Genetic Studies of the Same Disease

Alejandro Schäffer
CBB/NCBI
Dept. of Health and Human Services
National Institutes of Health
Bethesda, MD 20892
schaffer@helix.nih.gov

November 2004

Each year, hundreds of studies are published providing statistical evidence that variation in a specific chromosomal region is associated with a genetic disease. These studies are often precursors to successful identification of a disease-causing gene and mutations in that gene. A variety of test statistics are used, but the most commonly used statistics such as LOD (Logarithm of ODds) scores and NPL (Non-Parametric Linkage) scores are defined in such a way that scores from multiple families can be added. This makes it appealing to collect multiple families with the same phenotype, analyze them in the same way, and sum the results. Moreover, if multiple studies are conducted on different family sets in different laboratories, they it would be nice to combine the results in a ``meta-analysis'' to see what the combined evidence shows.

An obvious problem in achieving this goal is that different laboratories use different marker sets and different methods of analysis, publishing only a summary of the results. One moderately successful approach to meta-analysis of different studies has been proposed by Lewis and colleagues (Wise *et al.* 1999) and it is called GSMA (genome scan meta-analysis). This method divides the genome into small intervals and then assigns for each interval and study a rank for that interval, higher rank meaning more favorable evidence. The set of ranks is then combined and a statistical test is used to see if the combined ranks for the best intervals are better than would be expected by chance. The initial application of GSMA was to multiple sclerosis; some recent usages are for cleft lip and palate (Marazita *et al.* 2004), heart disease (Chiodini and Lewis 2003), psoriasis (Sagoo *et al.* 2004). Although one simulation study (Levinson *et al.* 2003) suggests that the statistical power of GSMA is good, it is hard to believe that GSMA is better than reanalysis of the combined genotype data, if that were available. Basic barriers to combining the original data are: 1) lack of permission to do so in human-subjects protocols 2) rivalries between investigators and 3) different experiments done in different labs. The first two might be overcome by some cryptographic approaches.

The combination of genetic data sets raises a second problem, which is rarely considered and that is that the same family may be participating in two studies of the same disease. One example I know of is for a rare disease called Bardet-Biedl syndrome, in which the same family was used by two competing groups (Katsanis *et al.* 2000; Slavotinek *et al.* 2000) that identified mutations in the MKKS gene on chromosome 20. The fact that two groups identified the same mutation in the same patient is not a serious problem, just a waste of resources. However, when the same families are used by multiple groups and then the evidence is combined by a method such as GSMA, then this can lead to inflated evidence for a genetic locus to due to implicit double

counting of the same data point. To alleviate this problem some method to identify whether two studies share any families without breaching confidentiality or HIPAA rules could be useful.

References

Chiodini BD, Lewis CM, ``Meta-analysis of 4 coronary heart disease genome-wide linkage studies confirms a susceptibility locus on chromosome 3q,'' *Arteriosclerosis Thrombosis and Vascular Biology*, 23 (2003), 1863-1868.

Katsanis N, Beales PL, Woods MO, Lewis RA, Green JS, Parfrey PS, Ansley SJ, Davidson WS, Lupski JR, ``Mutations in MKKS cause obesity, retinal dystrophy, and renal malformations associated with Bardet-Biedl syndrome,'' *Nature Genetics*, 26 (2000), 67-70.

Levinson DF, Levinson MD, Segurado R, Lewis CM, ``Genome scan meta-analysis of schizophrenia and bipolar disorder, part I: Methods and power analysis,'' *American Journal of Human Genetics*, 73 (2003), 17-33.

Marazita ML, Murray JC, Lidral AC, Arcos-Burgos M, Cooper ME, Goldstein T, Maher BS, Daack-Hirsch S, Schultz R, Mansilla MA, Field LL, Liu Y, Prescott N, Malcolm S, Winter R, Ray A, Moreno L, Valencia C, Neiswanger K, Wyszynski DF, Bailey-Wilson JE, Albacha-Hejazi H, Beaty TH, McIntosh I, Hetmanski JB, Tuncbilek G, Edwards M, Harkin L, Scott R, Roddick LG, ``Meta-analysis of 13 genome scans reveals multiple cleft lip/palate genes with novel loci on 9q21 and 2q32-35,'' *American Journal of Human Genetics*, 75 (2004), 161-173.

Sagoo GS, Tazi-Ahnini R, Barker JWN, Elder JT, Nair RP, Samuelsson L, Traupe H, Trembath RC, Robinson DA, Iles MM, ``Meta-analysis of genome-wide studies of psoriasis susceptibility reveals linkage to chromosomes 6p21 and 4q28-q31 in Caucasian and Chinese Hans populations,'' *Journal of Investigative Dermatology*, 122 (2004), 1401-1405.

Slavotinek AM, Stone EM, Mykytyn K, Heckenlively JR, Green JS, Heon E, Musarella MA, Parfrey PS, Sheffield VC, Biesecker LG, ``Mutations in MKKS cause Bardet-Biedl syndrome,'' *Nature Genetics*, 26 (2000), 15016.

Wise LH, Lanchbury JS, Lewis CM, ``Meta-analysis of genome searches,'' *Annals of Human Genetics*, 63 (1999), 263-272.