European Journal of Human Genetics (2006) 14, 1147–1148

www.nature.com/ejhg

NEWS AND COMMENTARY

Gene mapping

# Balance among quality, quantity and cost of data in the era of whole-genome mapping for complex disease

Derek Gordon

*European Journal of Human Genetics* (2006) **14,** 1147–1148.
doi:10.1038/sj.ejhg.5201693; published online 12 July 2006

The authors Chang *et al*[1] are to be commended for their efforts to educate the larger scientific public about the impact of data quality in complex disease gene mapping. Specifically, the authors document four types of data errors that are general to most linkage studies: errors in phenotype, pedigree structure, marker information, and marker genotypes. These authors also pose a set of excellent questions regarding the importance of data quality and provide empirical answers through their experience with GenNet whole-genome linkage studies (part of The Family Blood Pressure Program[2]).

Among the questions posed are: (i) How much of the genome is covered by a 10 cM linkage scan (when data are removed owing to low genotyping quality or Mendelian inconsistency)? (ii) Do allele shifting markers (ie, markers in which identical alleles are sometimes called differently because different flanking primers, allele sizing software, or allele binning methods have been used when STR genotyping of a data set is performed in multiple batches over several years) affect linkage evidence? (iii) Do family structure errors significantly reduce linkage signal? (iv) Is removal of Mendelian inconsistencies an adequate substitution for comprehensive data cleaning? The answers from the GenNet example document that comprehensive data cleaning can result in both the removal of false-positive evidence of linkage as well as a potentially substantial increase in linkage evidence for true positives. Perhaps most important, these authors (as other authors have recently carried out[3]) provide a comprehensive protocol that helps guarantee good data quality and therefore maximal power to localize disease genes for complex traits.

This work raises the larger question of allocation of resources in the era of whole-genome mapping for complex diseases. With the advent of genotyping technologies that can produce genotype calls for hundreds of thousands of genotypes across the whole genome[4] and a widely publicised successful gene localisation for age-related macular degeneration using these technologies,[5] there is an understandably strong attraction to thinking that methods that involve increasing data quantity will be a panacea for the ills of unsuccessful complex gene mapping studies. However, there are a number of factors involved in designing successful gene-mapping studies.

We illustrate three major factors and their relationship schematically in Figure 1. The figure is a triangle, with each vertex representing one aspect of a study that must be balanced with respect to the other two. The vertices are: study cost, in terms of time and money; data quantity, which includes the number of subjects for whom genotype and phenotype (diagnosis) information is obtained, and also the number of genotypes per individual obtained; and data quality, which represents the accuracy of the genotype, phenotype (or diagnostic), and other information (eg, environmental covariates) for each individual in the study. To illustrate use of this figure, consider some examples. If a research team wants to phenotype large numbers of individuals and genotype them, and in addition, wants to insure high accuracy rates, then the team must

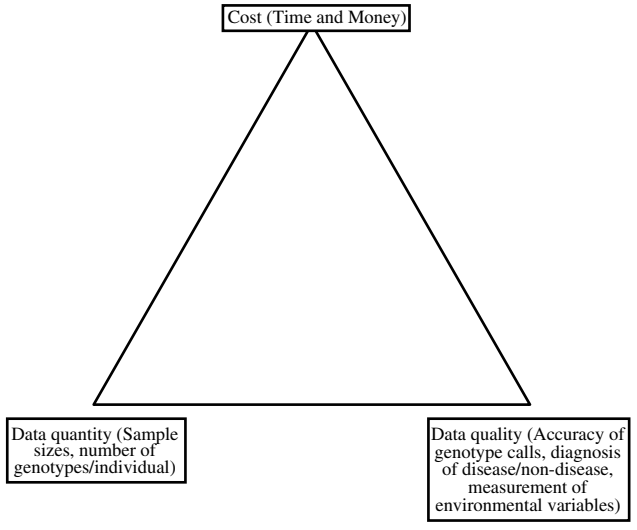


**Figure 1** Graphical representation of balance among the allocation of resources (factors) in whole-genome mapping studies.

increase either time or money invested. Similarly, if a team is working with a fixed budget (either time-wise or money-wise), and the team wants to insure high-quality phenotypes and genotypes, then it will most likely have to reduce the sample size studied.

It is critically important to bear these points in mind before embarking on large-scale linkage or association studies. Recent research (including that of Chang *et al*[1]) indicates that sacrificing data quality will reduce power to detect loci (or equivalently, will increase sample size requirements[6–10] for a fixed power level) and/or will shift the apparent location of susceptibility genes,[11,12] so that in effect, research teams may end up trying to find a 'moving target'. For example, the original sample size requirements needed if the data were highly accurate may be insufficient because accuracy is reduced to collect individuals in a given time frame.

In summary, the importance of Chang *et al*'s[1] work cannot be overstated. Data quality is a critically important factor to be considered if research teams are to be successful isolating complex disease loci in this new era of whole-genome mapping.

As a final thought, this author recommends that, in the balance among the three factors presented in Figure 1, research teams sacrifice neither data quality nor data quantity in their searches for complex trait susceptibility loci. That is, research teams might consider increasing cost, especially in terms of time, when performing their studies. The monetary costs per year can be kept smaller by increasing the time in which studies are completed. With this type of strategy, teams will have the added advantage that additional samples collected may be used as replication for initial linkage and/or association signals. This strategy has been employed to successfully localize susceptibility genes for previously thought intractable diseases as schizophrenia.[13]

## Acknowledgements

*We gratefully acknowledge the contributions of Dr Stephen J Finch (with whom the author has enjoyed a long and fruitful collaboration) on the subject of statistical methods to address misclassification error*■

*Dr D Gordon is at the Department of Genetics, Rutgers University, 145 Bevier Road, Room 128, Piscataway, NJ 08854, USA.*
*E-mail: gordon@biology.rutgers.edu*

## References

1 Chang YC, Kim JD, Schwander K *et al*: The impact of data quality on the identification of complex disease genes: experience from the Family Blood Pressure Program. *Eur J Hum Genet* 2006; **4**: 469–477.
2 Multi-center genetic study of hypertension: The Family Blood Pressure Program (FBPP). *Hypertension* 2002; **39**: 3–9.
3 Pompanon F, Bonin A, Bellemain E *et al*: Genotyping errors: causes, consequences and solutions. *Nat Rev Genet* 2005; **6**: 847–859.
4 Steemers FJ, Chang W, Lee G *et al*: Whole-genome genotyping with the single-base extension assay. *Nat Methods* 2006; **3**: 31–33.
5 Klein RJ, Zeiss C, Chew EY *et al*: Complement factor H polymorphism in age-related macular degeneration. *Science* 2005; **308**: 385–389.
6 Zheng G, Tian X: The impact of diagnostic error on testing genetic association in case–control studies. *Stat Med* 2005; **24**: 869–882.
7 Seaman SR, Holmans P: Effect of genotyping error on type-I error rate of affected sib pair studies with genotyped parents. *Hum Hered* 2005; **59**: 157–164.
8 Gordon D, Haynes C, Blumenfeld J *et al*: PAWE-3D: visualizing power for association with error in case–control genetic studies of complex traits. *Bioinformatics* 2005; **21**: 3935–3937.
9 Kang SJ, Gordon D, Finch SJ: What SNP genotyping errors are most costly for genetic association studies? *Genet Epidemiol* 2004; **26**: 132–141.
10 Gordon D, Finch SJ, Nothnagel M *et al*: Power and sample size calculations for case–control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Hum Hered* 2002; **54**: 22–33.
11 Clayton DG, Walker NM, Smyth DJ *et al*: Population structure, differential bias and genomic control in a large-scale, case–control association study. *Nat Genet* 2005; **37**: 1243–1246.
12 Barral S, Haynes C, Levenstien MA *et al*: Precision and type I error rate in the presence of genotype errors and missing parental data: a comparison between original TDT and TDTae statistics. *BMC Genet* 2005; **6**: S150.
13 Brzustowicz LM, Hodgkinson KA, Chow EW *et al*: Location of a major susceptibility locus for familial schizophrenia on chromosome 1q21–q22. *Science* 2000; **288**: 678–682.