Title: Family Based Studies and Genetic Epidemiology: Theory and Practice.

Authors: JH Barrett¹, NA Sheehan², A Cox³, J Worthington⁴, C Cannings³, MD Teare³

Affiliations

1. Genetic Epidemiology Division, Leeds Institute of Molecular Medicine, University of Leeds, Leeds, UK

2. Department of Health Sciences and Genetics, University of Leicester, Leicester, UK

3. School of Medicine and Biomedical Sciences, University of Sheffield, Sheffield, UK

4. The Medical School, University of Manchester, Manchester, UK

Corresponding author: M. Dawn Teare, PhD

Mathematical Modelling and Genetic Epidemiology
School of Medicine, The University of Sheffield
Beech Hill Road,
Sheffield, S10 2RX, UK
Tel 44 114 271 3195
Fax 44 114 271 1711
Email: m.d.teare@sheffield.ac.uk

Keywords: population based association studies, family study design, case-control-design

Running title: Family Based Studies

Abstract

Family based studies have underpinned many successes in uncovering the causes of monogenic and oligogenic diseases. Now research is focussing on the identification and characterisation of genes underlying common diseases and it is widely accepted that these studies will require large population based samples. Population based family study designs have the potential to facilitate the analysis of the effects of both genes and environment.

These types of studies integrate the population based approaches of classic epidemiology and the methods enabling the analysis of correlations between relatives sharing both genes and environment. The extent to which such studies are feasible will depend upon population- and disease-specific factors. To review this topic, a symposium was held to present and discuss the costs, requirements and advantages of population based family study designs. This article summarises the features of the meeting held at The University of Sheffield, August 2006.

Family studies have long played a central role in genetic epidemiology. Many genes with a strong influence on disease risk have been identified through linkage studies. However the familial linkage approach has had a limited success in mapping disease risk loci in multifactorial disease. To localise and identify genes associated with moderate to low risk large sample sizes are required, and it has been argued that in this context collecting family based material is too slow and inefficient. Recent advances in the availability of dense markers coupled with the realisation that the effect size for the majority of genetic risk loci will be low, are driving genetic epidemiology towards the population-based approaches of environmental epidemiology. Genome wide association studies are demanding very large sample sizes, which are necessary to detect these low effects with sufficiently small p-values to address concerns over multiple testing.

The primary focus of the family based studies symposium was to review the advantages, requirements and efficiency of family study designs in the investigation of multifactorial disease. Keynote speakers were invited to illustrate the new opportunities and challenges brought to genetic epidemiology in the genomic era. The topics addressed by the meeting included, population based family studies, models of interaction, twin and twin-family studies, family based tests in inbred populations, computational challenges in the genetic analysis of pedigrees, and the value and utility of population-specific based resources.

While linkage studies have a demonstrated utility in mapping rare, highly penetrant disease genes, results in mapping more moderate loci have been disappointing. The opportunities for mapping low penetrance disease predisposing genes with families alone are diminishing. Risch and Merikangas [1] argued that case-control association studies present much greater power to detect the locus specific relative risks underlying polygenic disease. Ten years later genome wide association studies are a reality [2,3,4] and their ability to uncover genetic risks in complex disease will soon be assessed. To be adequately powered, case-control association studies need to be very large, though they can be constructed to be considerably

more powerful (or equivalently, less expensive) if the cases are selected on the basis of their family history [5,6].

Family based resources can be used in gene discovery [7] and much work has been done on the performance of association tests in the family setting [8], family based tests are able to distinguish between sub-models of interaction [9,10], though special consideration is needed when the sampled population is inbred [11]. However, the real utility of the family based collection is expected to be the key future role when studying gene characterisation and models of gene/gene and gene/environment interaction [12], particularly when resources are population based. The merits and practicalities of variations on this design have been reviewed by Hopper et al. [13]. In its most simple form cases and controls are ascertained and subsequently biological samples and detailed information are collected on their relatives. The extent to which this is feasible will very much depend on the disease under investigation and population/country specific regulations regarding data protection and ethics. For example, when studying childhood disease it may be relatively easy to contact and include all first degree relatives of a case and age matched control. In adult onset disease including first degree relatives may be dependent upon the proband's consent for relatives to be contacted, and subsequently the consent of the relative to be approached. The resulting opportunities for selected dropout result in loss of power and concerns about response bias. Some countries have excellent record tracking facilities that enable reconstruction of pedigrees (through birth certificates) and linking with medical records. Under these conditions, familial risks and simple models of gene/environment interaction can be investigated [14] without the need for the individual to participate directly, though detailed exposures and genetic analysis would still require consent of and contact with subjects.

Twin and twin-family studies are good examples of the family cohort design. The long term follow-up of twin pairs permits the estimation of the relative contribution of environmental vs genetic effects in many phenotypes [15]. The twin-family design enables more refined study such as cultural transmission, imprinting, genetics and environmental stability [16, 17]. Other

recently formed family cohort studies include the GRAPHIC study [18], the Norwegian Mother and Child Birth Cohort Study [19] and Generation Scotland [20].

The Utah genealogical database is an exceptional founder population resource. This archive contains health and pedigree records on over 9 million subjects. The extended pedigrees have enabled the study of the familial component of many diseases [21, 22] and have been exploited to investigate pseudo-isolate populations [23], and genetic linkage analysis [24, 25].

Power to detect linkage in oligogenic disease can be improved by consideration of co-variables [26], power can also be enhanced through linkage studies on founder or inbred populations but the pedigrees collected through these studies often place additional demands to the analysis. Genetic analysis methods need to be developed to fully integrate new forms of common genetic variation such as copy number polymorphisms [Teare, unpublished data] and to incorporate prior information including error models [27]. Recent developments in computational algorithms have made exact multipoint linkage analysis feasible in large, looped pedigrees [28]. There are still many instances when extensive multipoint analysis is only possible using Markov Chain Monte Carlo approaches. The computation time necessary to estimate the type 1 error rate for these implementations can be substantially reduced through use of fuzzy p-values [29, 30].

Although case-control studies are the most powerful strategy for gene-discovery, family based studies clearly have more potential in characterisation of both main effects and interactions. Questions relating to interaction will clearly require datasets where environmental exposures can be measured or estimated [31].

We would like to thank all participants attending the symposium and our sponsor GE Healthcare. We particularly thank the invited keynote speakers, Tim Bishop, Heather Cordell, Lisa Cannon-Albright, Doug Easton, Emmanuelle Genin, Jaakko Kaprio, and Elizabeth Thompson. The meeting was co-hosted by the North of England Genetic Epidemiology Group and the Academic Unit of Mathematical Modelling and Genetic Epidemiology, University of Sheffield.

References:

1. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* 273: 1516-1517.
2. Maraganore, DM, de Andrade, M, Lesnick, TG, Strain, KJ, Farrer MJ, Rocca, WA, Pant, PVK, Frazer, KA, Cox, DR, Ballinger, DG. (2005) High-resolution whole-genome association study of Parkinson disease. *Am J Hum Genet* 77, 685–693.
3. Duerr RH, Taylor KD, Brant SR, Rioux, JD, Silverberg MS, Daly MJ, Steinhardt AH, Abraham C, Regueiro M, Griffiths A, Dassopoulos T, Bitton A, Yang H, Targan S, Datta LW, Kistner EO, L. Schumm LP, Lee AT, Gregersen PK, Barmada MM, Rotter JI, Nicolae DL, Cho JH. (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene *Science* 1 December 2006: 1461-1463.
4. Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, Smink LJ, Guja C, Ionescu-Tirgoviste C, Widmer B, Dunger DB, Savage DA, Walker NM, Clayton DG, Todd JA. (2006) A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nature Genetics* 38 (6): 617-619.
5. Antoniou AC, Easton DF. (2003) Polygenic inheritance of breast cancer: Implications for design of association studies. *Genetic Epidemiology* 25: 190-202.
6. Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, North B, Jayatilake H, Barfoot R, Spanova K, McGuffog L, Evans DG, Eccles D, Easton DF, Stratton MR, The Breast Cancer Susceptibility Collaboration (UK), Rahman N. (2006) ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility *Nature Genetics* 38: 873-875.
7. Van Steen K, McQueen MB, Herbert A, Raby B, Lyon H, DeMeo DL, Murphy A, Su J, Datta S, Rosenow C, Christman M, Silverman EK, Laird NM, Weiss ST, Lange C. (2005) Genomic screening and replication using the same data set in family-based association testing. *Nature Genetics* 37: 683-691.
8. Biernacka JM and Cordell HJ (2006) Perils and pitfalls of permutation tests for distinguishing the effects of neighbouring polymorphisms. *Genetic Epidemiology* 30:582-9.
9. Cordell HJ, Barratt BJ, Clayton DG (2004) Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genetic Epidemiology* 26: 167-185.
10. Jonasdottir G, Palmgren J, Humphreys K. (2005) Analysis of binary traits: testing association in the presence of linkage *BMC Genetics* 6: S92.
11. Bourgain C, Genin E. (2005) Complex trait mapping in isolated populations: Are specific statistical methods required? *European Journal Of Human Genetics*, 13:698-706.
12. Howson JMM, Barratt BJ, Todd JA, Cordell HJ. (2005) Comparison of population- and family-based methods for genetic association analysis in the presence of interacting loci *Genetic Epidemiology* 29: 51-67.
13. Hopper JL, Bishop DT, Easton DF. (2005) Genetic Epidemiology 6 - Population-based family studies in genetic epidemiology. *Lancet* 366: 1397-1406.

14. Bermejo JL, Hemminki K. (2005) Familial risk of cancer shortly after diagnosis of the first familial tumor. *Journal Of The National Cancer Institute* 97: 1575-1579.
15. Boomsma D, Busjahn A, Peltonen L. (2002) Classical twin studies and beyond *Nature Reviews Genetics* 3: 872-882.
16. Kaprio J, Pulkkinen L, Rose RJ. (2002) Genetic and Environmental Factors in Health-related Behaviors: Studies on Finnish Twins and Twin Families. *Twin Research* 5: 366-371.
17. Keski-Rahkonen A, Viken RJ, Kaprio J, Rissanen A, Rose RJ. (2004) Genetic and environmental factors in breakfast eating patterns. *Behavior Genetics* 34: 503-514.
18. Tobin MD, Raleigh SM, Newhouse S, Braund P, Bodycote C, Ogleby J, Cross D, Gracey J, Hayes S, Smith T, Ridge C, Caulfield M, Sheehan NA, Munroe PB, Burton PR, Samani NJ. (2005). Association of WNK1 gene polymorphisms and haplotypes with ambulatory blood pressure in the general population. *Circulation* 112: 3423-3429.
19. Ronningen KS, Paltiel L, Meltzer HM, Nordhagen R, Lie KK, Hovengen R, Haugen M, Nystad W, Magnus P, Hoppin JA. (2006). The biobank of the Norwegian mother and child cohort Study: A resource for the next 100 years. *European Journal Of Epidemiology* 21: 619-625.
20. Smith BH, Campbell H, Blackwood D, Connell J, Connor M, Deary IJ, Anna F Dominiczak AF, Fitzpatrick B, Ford I, Jackson C, Haddow G, Kerr S, Lindsay R, McGilchrist M, Morton R, Murray G, Palmer CNA, Pell JP, Ralston SH, St Clair D, Sullivan F, Watt G, Wolf R, Wright A, Porteous D, Morris AD. (2006) Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Medical Genetics* 7: Art. No. 74.
21. Goldfarb-Rumyantzev AS, Cheung AK, Habib AN, Wang BJ, Lin S, Baird BC, Naiman N, Cannon-Albright L. (2006) A Population-Based Assessment of the Familial Component of Chronic Kidney Disease Mortality *American Journal Of Nephrology* 26: 142-148.
22. Thomas A, Cannon-Albright L, Bansal A, Skolnick MH. (1999) Familial associations between cancer sites. *Computers and Biomedical Research* 32: 517-529.
23. Cannon-Albright LA, Farnham JM, Thomas A, Camp NJ. (2005) Identification and study of Utah pseudo-isolate populations - Prospects for gene identification. *American Journal Of Medical Genetics* 137A: 269-275.
24. Camp NJ, Farnham JM, Cannon-Albright LA. (2006) Localization of a prostate cancer predisposition gene to an 880-kb region on chromosome 22q12.3 in Utah high-risk pedigrees. *Cancer Research* 66: 10205-10212.
25. Camp NJ, Cannon-Albright LA. (2005) Dissecting the genetic etiology of major depressive disorder using linkage analysis *Trends In Molecular Medicine* 11: 138-144.
26. Nsengimana J, Samani NJ, Hall AS, Balmforth AJ, Mangino M, Yuldasheva N, Maqbool A, Braund P, Burton P, Bishop DT, Ball SG, Barrett JH for the British Heart Foundation Family Heart Study Research Group. (2006) Enhanced linkage of a locus on chromosome 2 to premature coronary artery disease in the absence of hypercholesterolemia. *European Journal of Human Genetics* advance online publication, December 6, 2006.

27. Egeland T, Sheehan NA (2007) Structured incorporation of prior information in identification problems. *Annals of Human Genetics in press*.
28. Silberstein M, Tzemach A, Dovgolevsky N, Fishelson M, Schuster A, Geiger D. (2006) Online system for faster multipoint linkage analysis via parallel execution on thousands of personal computers. *American Journal Of Human Genetics* 78: 922-935.
29. Thompson, E. A. and Geyer, C. J. (2007) Fuzzy p-values in latent variable problems. *Biometrika in press*.
30. Thompson, E. A. (2006) Uncertainty in Inheritance: Assessing evidence for linkage. Paper presented at the Third University of Washington Biostatistics Symposium: Nov 2005. Technical report No. 498, Department of Statistics, University of Washington. (<http://www.stat.washington.edu/www/research/reports/2006/tr498.pdf>)
31. Hemminki K, Bermejo JL, Forsti A. (2006) The balance between heritable and environmental aetiology of human disease. *Nature Reviews Genetics* 7: 958-965.