ESF EMRC Exploratory Workshop:

Finding quantitative trait loci (QTL) for complex traits in humans. Statistical analysis of data from selected groups of related individuals: An application to anxiety/depression.

Oxford, 11-13 January 2001

Organisers: Dorret Boomsma (Vrije Universiteit, Amsterdam)
Lon Cardon (Welcome Trust, Oxford, UK)

SCIENTIFIC REPORT

Summary

A very active and still growing field of research is concerned with the discovery of genes that contribute to the development of human diseases. The search for such genes has been successful for some relative simple diseases, that is, diseases in which only one or at most a few genes are involved. Studies that compare the similarities of monozygotic and dizygotic twins have repeatedly demonstrated that more complex diseases or, more generally, quantitative human traits also often have a genetic basis. However, for such traits each contributing gene can be expected to have a much smaller effect, which reduces the power of statistical methods and research designs commonly used for simple diseases to reliably identify the individual genes involved in quantitative traits. Consequently, the identification of the genes involved becomes much harder, even to the extent that some scientists have expressed their doubts about the prospects of success for this endeavour.

However, it has recently become clear that a combination of measures may increase the statistical power sufficiently to identify at least some of the more influential genes. These measures can be broadly categorised as methods to improve the design of the data collection and statistical methods and models to extract more information from the data once they have been obtained. Moreover, some statistical methods can be more informative for some designs than for others and some designs require modification of existing statistical methods. The participants in the workshop presented and discussed known and new results pertaining to these issues.

Content

This was a highly successful meeting, which was held in Oxford (UK) January 11-13, 2001. There were 23 participants (14 senior faculty, 7 students and postdocs and 2 statistical geneticists from Gemini, UK).

The first day started with two lectures on general aspects of design and data collection. Next came a presentation which compared the informativeness of several designs for variance components models. In the afternoon a number of current projects were discussed: their sampling features, problems encountered in the study and ways to tackle those problems. The presentations on the second day concentrated on more technical statistical aspects of analysing data from selected samples. Summaries of the presentations are given below.

Introduction, study designs, data
Dorret. Boomsma

The meeting started with an introduction that outlined the rationale for selection of families in QTL studies of complex traits. The power to detect genes that influence only a relatively small part of the genetic variance of complex / quantitative traits is low. Therefore, selection of informative families is a necessary step in gene-hunting projects, such as the studies of anxiety and depression, currently carried out in the UK, The Netherlands, Australia and Virginia (USA).

Selection of siblings for genotyping can be based on:
- extreme phenotypic scores
- sibship size
- parental phenotypes
- availability of parents; willingness to participate

After selection, the selected families participate in:
- genotyping (microsatellites, SNPs)
- additional phenotyping (e.g. psychiatric interviews and intermediate phenotypes)

After genotyping, a second selection can be carried out for:
- fine-mapping / association analysis
- additional phenotyping: 'expensive' phenotypes in part of the sample (e.g. in IBD = 0 or 2 sibling pairs)

Problems in the analysis of genotypes – phenotypes involve:
- parameter bias in linkage and association analysis: dependent on statistical approach
- mode of selection (concordant / discordant)
- background genetic correlation
- QTL allele frequency
- distributional properties of the data
- the analysis of genotyped & untyped Ss within the same family
- multicenter studies: replication (possibly using different phenotypes / markers)

Genetic models for complex phenotypes
Lodewijk Sandkuijl

Sandkuijl presented an overview of considerations that may or should determine the design and analysis of a study aimed at discovering genes or a region linked to a gene for complex diseases or phenotypes. A major considerations is whether the disease or phenotype is rare or relatively common. Whatever the type, stratified or biased sampling is in general preferable to random sampling.
- For common phenotypes and common alleles, families or subjects can be selected at random, or extremes subjects and controls may be sampled.
- For a genome search, families with extreme probands or extreme / affected sib pairs can be selected.

- Families with at least two extreme members or extreme subjects in isolated populations are useful for discovering QTL's for a rare phenotype.
- For screening, the ASP (affected sib pair) design, extended families and patients in isolated populations are useful (these designs require a progressively smaller sample).
- For fine mapping, extended families or patients in isolated populations may be used, but conclusions for isolated populations may not generalise to other populations.
- For estimation of risk contribution, cases versus controls, the ASP design (less suitable but usable), or isolated populations can be used.

Another concern is the number of a priori assumptions that should be put into the model. For ASP designs comparatively little assumptions are needed as compared to extended family designs (which for example require much better estimates of gene frequencies).
The relative information of affected versus unaffected sib pairs depends on the penetrance. They are just as informative for100% penetrance, but if the penetrance is only 50 %, the affected sibs are much more informative. Thus, specification of the penetrance can be highly influential for the outcome of the analysis.

Locus heterogeneity may also be a problem within families; if it exists, one should use nuclear families rather than extended families.

Study design: optimal selected sampling for complex traits
Shaun Purcell

In variance components analysis, several subject selection methods are being used: randomly selected subjects, probands, extreme discordant sib pairs, extreme concordant and discordant sib pairs (EDAC design), the ASP design, maximally dissimilar subjects. For multidimensional problems, maximal dissimilar subjects may be selected using the Mahalanobis distance, which also generalises to larger sibships. No selection scheme is optimal under all conditions (e.g., concordants are informative for rare recessive diseases).

To judge the informativeness of a selection, the noncentrality parameter of the loglikelihood-ratio is useful. Results of a simulation study were presented in which the various selection schemes were evaluated. The ASP and proband selections did not work well. Extreme discordant pairs and especially the EDAC design were much more informative. For linkage analyses, selection of sibships that are likely to contain pairs that are IBD zero or two is powerful. For association relatively large numbers of individuals who are homozygous for the QTL should be selected.

For DNA pooling, simulations and calculations suggest that the top and bottom 27 % of the distribution should be selected. It can also be useful to select multiple pools from the initial sample.

Overview of current projects employing selected sampling from large phenotyped populations with a focus on anxiety/depression

In this session a number of participants described the design and data collection of current projects.

Flint discussed the progress in collecting a large sample of families, screened with the Eysenck Personality Questionnaire, to be used for searching for the genetic determinants of neuroticism. He described the development of an automated procedure for mailing siblings between the ages of 30 and 50 and how they have collected 88,000 individuals, among which are 24,000 sibling pairs suitable for selecting extremes for genotyping.

Flint stressed the importance of minimizing error rates because even small errors in genotyping result in a very large loss of power to detect genetic effects. To do so, they have designed a database and an automated allele caller so that errors in data handling can be reduced to the minimum. The database is relational, running on a PC with web interfaces so that it is available to all users, across all platforms. The allele-caller is rule-based, rather than relying on learning algorithms, because of the greater robustness this affords. The critical problem here is to identify those alleles that may be misinterpreted and allow those to be manually checked. The program is written in C and has yet to be fully tested in conjunction with the database. It will be made available to the scientific community.

So far, they have collected 3,000 cheek swabs for DNA extraction, and have automated the process of plate-layout and data collection from two genotyping machines, an ABI 3700 and the Megabace automated sequencer. Genotyping will be completed in about 9 months.

Sham presented an overview of the GENESIS project and the strategies used in that project. The main variables in the study are the EPQ (neuroticism) and MASQ (anxious arousal / high positive affect). Data were obtained of 10000 sibships for a total of 25000 respondents out of 80000 that were asked to answer the questionnaires. The 800 most informative sibships are selected for linkage and association analysis and a psychiatric interview.

Kaprio gave an overview of the resources of the Finnish Twin Register. Around 20.000 like-sex twin pairs born before 1958 have participated in survey studies in 1975, '81 and '90 (in 1996-97 opposite-sex twins were also asked to participate). These surveys included questionnaires that assessed Neuroticism and Depression. For 2500 of these twin pairs (1576 DZ) DNA samples are available. In samples of younger twins (3065 pairs born between 1975-79 and 3112 pairs born in 1983-87) assessment of depression by parents, teachers and the twins themselves is available. Some samples also include siblings.

Testing for linkage replication
Mike Neale and Hermien Maes

N&M discussed the difficulty of finding linkage and testing for linkage replication of complex phenotypes. Traditionally a high significance level, corresponding to a lod score of 3.6 or higher, has been used as a criterion for linkage. But for complex traits this stringent criterion is usually not adhered to. An alternative is the use of loglikelihood-ratio support intervals, or the place where the loglikelihood-ratio drops one point.

A problem in evaluating a replication is that the lod score peak may occur on a different position. Peaks will be spread more if not all families with relevant alleles are affected. What is measured in different studies may be conceptually similar but yet not equivalent. Moreover, genes with similar effects could cluster in the same region. Simulation studies suggest that the estimated QTL position can on average differ 10 cm for 100 sib pairs and 2.6 cm for 1000 sib pairs. The number of linked families has a large effect on the estimated QTL position. The GASP programme can be used for simulation of genotype and phenotype data.

Combining data of several studies in one analysis can provide a formal test of the equivalence of the QTL position by comparing loglikelihood-ratios for models in which the position is and is not constrained to be the same across the studies.

They also presented an overview of Mx models for QTL analysis. Selection of subjects can in principle be modelled in Mx, but this may become quite complicated. Instead, a weight formula can be used (different weights for different pi-hats) for estimating a mixture distribution. In Mx sib pair means, covariances and weights may differ. Mx can handle missing values and estimate threshold models.

Finally, it was shown how the computer programme Mx can be used for the estimation of haplotype frequencies, a type of analyses for which it is not normally used.

Likelihood in selected and unselected sample
Pak Sham

Analysing a selected sample by standard methods results in higher type one errors. Also, selection usually causes non-normality. Several solutions of this problem have been proposed:

- Impute pi values of nonselected subjects
- Use Haseman-Elston regression
- Adjust the test statistic
- Model the non-normal distribution
- Use robust methods, for example GEE (generalised estimation equations) or nonparametric methods

Sham proposed a new method based on adjustment of the test statistic. The reasoning behind it is in analogy to the proband method, where selection does not cause problems. Simulation results were presented that confirmed the validity of the method if the data are normally distributed. For non-normal data, the type one error rate is larger than the nominal level.

Sham also presented a modified Haseman-Elston regression which should theoretically be almost as powerful as a variance components analysis (which is known to be more powerful than traditional Haseman-Elston regression). The theoretical calculations were demonstrated and confirmed in a simulation study.

Full IBD distribution and pi-hat approach in selected samples: sib pairs and larger families
Nick Martin and Andrew Birley

Martin discussed the design of a study on Australian twins and their sibs. Variables of interest are neuroticism and anxiety. Families with extreme scoring subjects were selected for genotyping. The linkage analyses of the study was discussed by Birley. In particular the use of larger sibships and their full IBD distribution instead of using only the pairwise IBD probability estimates.

Combined association and linkage analysis
Lon Cardon and George Abacasis

Besides giving some results on design issues, Cardon and Abacasis demonstrated the effects of genotyping errors on the results of QTL / association analyses. A one percent error in genotyping is tolerable, but 10 percent error can reduce the lod score by as much as 30-40 percent. For selected samples or rare alleles the effects are even worse. Error detection can be based on Mendelian transmission errors, unlikely recombinations and unlikely haplotypes. Results of doing the required calculations were presented for three programmes: a new programme called Merlin (Cardon & Abacasis), Genehunter and Allegro. Merlin is much faster than the other programmes and also requires much less memory. Some computationally intensive and memory hungry analyses (e.g. checking errors in large families) can only be performed within reasonable time and memory limits by Merlin.

Selection after genotyping
Bas Heymans

In this presentation empirical results were presented for a genome scan in Dutch sib pairs. The analyses suggested a significant QTL effect, but closer inspection of the data showed that only a few families / sib pairs contributed to the QTL effect. These sib pairs, all of which were IBD zero, seemed to have rather extremely dissimilar scores as compared to the rest of the sample. A major problem for the interpretation of the results thus becomes whether these are real or spurious effects. The effects might be spurious because the sib pairs are, for one reason or another, outliers. As most of the information for a QTL can be expected to become from the tails of a phenotype distribution, such problems of interpretation are bound to also occur in other studies.

Assessment of the results and contributions to the future direction of the field

The results presented in the workshop warrant a guarded optimism for discovering genes involved in complex traits. There was considerable agreement in the results on which designs are most powerful for linkage analysis as judged from a variety of perspectives. Hence these designs, in which subjects are selected for genotyping using one of a number of criteria, will be used more often in ongoing and future studies. However, using selected subjects poses the problem of how to analyse the data so that genetic effects or linkage are estimated consistently. One participant presented a solution to this problem which appears to work well. However, this problem will remain an issue for future research.

Results were presented that demonstrate that accurate genotyping is extremely important, even with fairly large samples. Because obtaining statistically reliable linkage results for quantitative traits requires fairly large samples even if subjects are first selected for genotyping, genotyping will be done by automated procedures. An automated procedure for genotyping was presented, but it was also made clear that validating the accuracy of the procedure is a problem that is not easily solved. Thus, how to increase and assess the accuracy of such procedures is an important topic for future research.

Although the presented results demonstrated that linkage can be obtained with appropriate designs and statistical methods, candidate regions for genes involved may still be quite large. This was verified with an empirical example. The results and discussion also made clear, that finding ways to narrow this region size is still to a large extent an open and important problem.

Address List of Participants:

1. **Beem**, Leo, Dept. of Biological Psychology, Vrije Universiteit, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands
2. **Birley**, Andrew, , Queensland Institute of Medical Research, 300 Herstin Road, Brisbane QLD 4029, Australia (current address)
3. **Boomsma**, Dorret, Dept. of Biological Psychology Vrije Universiteit, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands
4. **Cardon**, Lon, The Wellcome Trust Centre for Human Genetics, University of Oxford,Roosevelt Drive, Oxford OX3 7BN, UK
5. **Cherny**, Stacy, The Wellcome Trust Centre for Human Genetics, University of Oxford,Roosevelt Drive, Oxford OX3 7BN, UK
6. **Dudbridge**, Frank, Wellcome Trust Centre for Molecular Mechanisms in Disease, University of Cambridge, Hills Road, Cambridge CB2 2XY, UK
7. **Flint**, Jonathan, Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK
8. **Heijmans**, Bas, Dept. of Molecular Epidemiology, Leids Universitair Medisch Centrum, Wassenaarseweg 72, 2333 AL Leiden, The Netherlands
9. **Kaprio**, Jaakko, University of Helsinki, Dept. of Public Health, Mannerheimintie 172, 00014 Helsinki, Finland
10. **Maes**, Hermien, Dept. of Human Genetics, Virginia Commonwealth University, P.O. Box 980003 Richmond, VA 23298-003, USA
11. **Martin**, Nick, Queensland Institute of Medical Research, 300 Herstin Road, Brisbane QLD 4029, Australia
12. **McGuffin**, Peter, Social, Genetics and Developmental Psychiatry Research Centre Institute of Psychiatry, De 'Crespigny Park, Denmark Hill, London SE5 8AF, UK
13. **Mittalbag**, Rita, Gemini Genomics, 162 Science Park. Milton road, Cambridge, CB4 0GH, UK
14. **Neale**, Mike, Dept. of Human Genetics, Virginia Commonwealth University, P.O. Box 980003 Richmond, VA 23298-003, USA
15. **Nobelzinen**, Pebbe, University of Helsinki, Dept. of Public Health, Mannerheimintie 172, 00014 Helsinki, Finland
16. **Posthuma**, Daniëlle, Dept. of Biological Psychology, Vrije Universiteit, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands
17. **Purcell**, Shaun, Social, Genetic and Developmental Research Centre, Institute of Psychiatry, De 'Crespigny Park, Denmark Hill, London SE5 8AF, UK
18. **Putter**, Hein, , Dept. of Medical Statistics, Leids Universitair Medisch Centrum,, 2300 RC Leiden, The Netherlands
19. **Sandkuijl**, Lodewijk, Dept. of Medical Statistics, Leids Universitair Medisch Centrum,, 2300 RC Leiden, The Netherlands
20. **Sham**, Pak, Social, Genetic and Developmental Research Centre, Institute of Psychiatry, London, UK
21. **Slagboom**, Eline, Dept. of Molecular Epidemiology, Leids Universitair Medisch Centrum, Wassenaarseweg 72, 2333 AL Leiden, The Netherlands
22. **Veriensalo**, Pia, University of Helsinki, Dept. of Public Health, Mannerheimintie 172, 00014 Helsinki, Finland
23. **Williamson**, Richard, address unknown.