SMP: Genetic Epidemiological Methods (GEM)

Project: Genome-wide Linkage Analysis and Association Studies Using SNP Genotypes from the Affymetrix 10K and 100K Chips

Klaus Rohde - **Max Delbrück Centre for Molecular Medicine (MDC), Berlin-Buch - rohde@mdc-berlin.de**

Introduction

Genome-wide linkage and association studies allow to find causal genes without having to know anything on their position on the genome, as necessary in the case of candidate gene studies. They will find such a causal gene, even if there is no assumption for its connection or contribution to the trait. Positional cloning of Mendelian traits via genome-wide linkage studies has been a powerful tool for the past two decades. Application of genome-wide linkage in the analysis of complex genetic traits, however, were only partially successful, with the lack of success caused mainly by the low heritability of those complex traits (common variants of modest effect). Higher density of markers and better powered study design may help to a certain degree, however even then a consecutive (candidate gene) association study will be necessary to locate the causal gene in the generally broader region of linkage.

Genome-wide association studies combine the generally higher power of association studies with the advantage of not having to know anything of the causal gene beforehand. However, this unbiased approach has its prize: the large number of SNP to be genotyped along the genome and the high number of individuals necessary to find also modest signals made such approaches for a long time not feasible. With the advent of the gene chip technology by Affymetrix and Perlegen such studies have become possible, first by using the 10K and 100K chips, than even a 500K chip along the genome. Great expectations raised by the potential of these excellent technologies are damped by a series of emerging problems:

- population stratification

Population admixture is the presence of subgroups within the population with different frequencies in disease and marker alleles. This may lead in an association study of unrelated individuals to an enrichment of one subgroup with higher disease allele frequency in the cases and to a (false positive) association if also the marker allele frequency is different from the controls.

One solution for this problem is the co-genotyping of a series of unlinked markers (genomic controls) which serve to assess the inflation of the chi-square statistics by admixture, or a re-matching of cases and controls using programs structure and Strat [1,2] to detect and deal with the underlying stratification.

Family-based studies are immune to the stratification problem, however need a much higher number of genotyped individuals, and may be more sensitive to effects of genotyping errors.

- technical artefacts

Genotyping via chip is an excellent technology, however, by its automated analysis vulnerable to false genotypes and more importantly to missing (not called) ones. It is a problem if at a locus only one special genotype is not called and causes deviation from Hardy-Weinberg-Equilibrium and false positive associations.

- multiple testing

In a genome-wide association study 10K up to 500K SNP will be genotyped and tested for association. This multitude of test brings up the problem that each p-value found in a single test has to be multiplied by the number of independent tests carried out (Bonferroni correction). Since not all SNP in the chip are in linkage equilibrium and independent, the real number of independent SNP may well be less than their total number on the chip, which nevertheless may serve as a conservative correction (very conservative for the 500K chip!).

A solution for this problem is the use of two samples, a screening sample and a confirmatory sample. The screening sample will be used to carry out the genome-wide study, subject to multiple testing. Instead of carrying out this correction one defines a level for the statistic which secures a necessary power in trade-off for accepting a number of false positives. All those SNP accepted in the screening set have then to be tested in a second confirmatory sample, without the burden of too high a correction factor.

Results/Project status

Linkage analyses using Gene Chip data

Linkage analysis using multiple loci of dense SNP markers from the gene chips have no major problems with population stratification and multiple testing and are the method of choice for the 10K chips with a mean SNP distance of 300 kBP, which is much greater than mean regions of linkage disequilibrium along the genome. Therefore we started at the Gene Mapping Center at the MDC with a compilation and adaptation of routines for linkage analysis of the gene chip data in an own program ALOMOHORA [3], written in Perl/Tk and running under Windows and Linux.

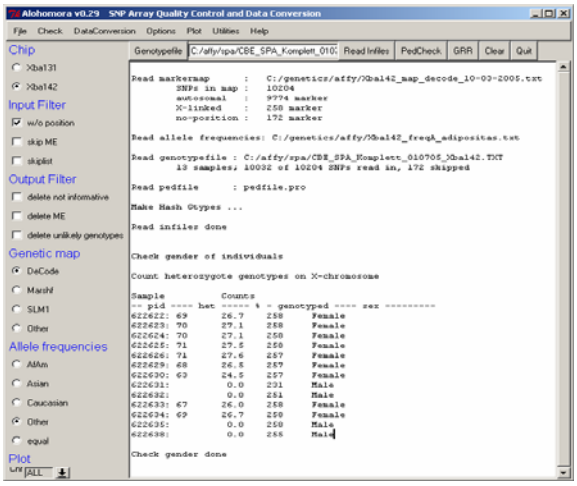

***Fig 1**: Screen shot of the ALOHOMORA main menu*

ALOHOMORA accepts genotype data as generated by the GeneChip DNA Analysis Software (GDAS v3.0) from Affymetrix, carries out conversion and control of these data for the attached analysis routines. For comprehensive quality control ALOHOMORA uses a number of freely available programs. The gender of samples may be checked by counting the heterozygous SNP genotypes on the X-chromosome and comparing it to the pedigree file information. The correct relationships within families are tested by the program GRR [4], using also genome-wide SNP information. The detection of Mendelian errors is carried out by the program PedCheck [5]. SNP with Mendelian errors and SNP that are not informative for linkage can be selectively removed from the sample. Non-mendelian errors will be detected by the option 'error' of the routine Merlin [6] and unlikely genotypes may the deleted in individuals in which they occur. Other options regard the chip

version used, since SNP contents may differ between different version of the chips, the preferred genetic map, and allele frequencies for a appropriate ethnicity.
The checked data will then be converted for the linkage analysis routines Allegro [7], Genehunter [8], Merlin [6] and Simwalk2 [9]. The high number of marker loci on the chip raised problems especially for the older programs Genehunter and Simwalk2, which were designed for the requirements of conventional low-number marker sets with a restriction for the number of markers used. This had partially be overcome by recompiled versions of the programs for a higher maximum number of markers, or as alternative, by using a subset of consecutive markers for the analysis in a moving window technique.
For the chosen program, the user can define a genetic model in case of parametrical linkage, the size of the moving window and furthermore, select linkage program-specific options.
Non-parametric LOD score calculations are preferably performed with Merlin or Allegro, chromosome by chromosome using all SNPs on a chromosome simultaneously for a multipoint analysis.
Parametric linkage analysis was carried out with Allegro v1.2, Genehunter 2.1v5 and with Simwalk2 v2.89. Due to the limitations of Genehunter and Simwalk2 with respect to the number of markers, the analysis was done with subsets of markers in a way of non-overlapping moving windows. Simwalk2 was recompiled for the use of up to 255 markers in one run. In case of large pedigrees, when Genehunter drops individuals and both Allegro and Merlin would skip the pedigree, we split the pedigree to appropriate sizes or use Simwalk2 for the pedigree as a whole.
Non-parametric LOD score (NPL) and parametric LOD score may be plotted for all chromosomes. All four programs Allegro, Merlin, Genehunter and Simwalk2 generate haplotypes, however as linkage routines on a family basis under the wrong assumption of linkage equilibrium between loci in the founder individuals. Nevertheless the higher pedigree information may overcome this disadvantage in comparison to population-based studies.

Linkage disequilibrium studies using GeneChip data

Linkage disequilibrium studies using GeneChip data comprise association studies on single marker and marker haplotype basis. Our first study in this direction was the analysis of a data set of 100 cases and 100 controls for multiple sclerosis, genotyped on a 10K basis [10]. The low sample size of this study should reveal only high risk alleles and in this way give expertise in those kind of studies and demonstrate their feasibility. We used for the analysis our haplotype estimation programs on EM (expectation maximization) basis embedded into scripts for the open-source statistics package R (http://www.R.org), which were also used for the single marker analysis.
As expected from the large mean distance between the 10K SNP the haplotype analysis proved to be not very helpful. Only in some regions with a higher SNP density reliable haplotypes could be estimated, as for example, for exceptional dense lying three SNP on chromosome 6, which revealed also an (already known) high association to the disease. The single marker analysis based on a 2 by 3 table chi-square test signalled the strongest significance also for those three SNP on chromosome 6, which remained significant even after the (conservative!) Bonferroni correction for multiple testing, and raises the hope, that strong signals may be detected even for moderate sample sizes and correction for multiple testing. All other most significant SNP selected for a confirmatory analysis in a second confirmatory data set failed a significant association.
Some of these SNP, re-genotyped by a different approach, revealed that their high significance in the chip analysis was caused by a bias of the not called SNP to a special genotype, underlining the warning to skip SNP marker with too low a calling rate from the analysis.
At present we are in preparation of a linkage and association study for the new 500K Affymetrix GeneChip.

Outlook

The 500K GeneChip with SNP genotypes at a mean distance of about 6 kBP will allow a genome-wide characterization of regions with high linkage disequilibrium and give the possibility for estimating haplotypes and their association to the trait. The choice of the sample (250 nuclear families with 2 children each) allows linkage analysis and as association studies by TDT test, so that results of the different approaches, their strength and weaknesses may easily be compared. However, an expected calling rate of around 90 percent may cause some problems. One could skip those SNP with too low a calling rate or find ways to overcome it, possibly by using haplotype analysis also over those no called genotypes, in the hope, that the population-based haplotypes will repair those not called genotypes.
To circumvent inflation of the association by multiple testing, we use a two stage study design, with the first stage using the gene chip data as explorative sample for finding a certain limited set of significant SNP, which will then be tested in a second confirmatory set of similar sample size and a much lower Bonferroni correction.
The huge amount of genome-wide genotyping data via the Affymetrix chips over all genotyping projects at the GMC present a valuable source for the analysis of the underlying population. In the framework of NGFN2 most of the studies will come from regions of Germany and even if cases (and quite often also controls) are sampled with a bias in respect to the underlying traits, a large number of genotyped chromosomes, which are not at all linked to the trait, could serve as unbiased sample. With this in mind, one could compare the genotypes on these chromosomes for population stratification, using the programs structure and Strat and look if either the population stratification, if there is any, differs between different chromosomes of the same study, or on the other side, there is a difference between the same chromosome for different studies from the same region.
This more and more increasing data base could in this way serve to elucidate the population characteristics of the underlying German population, which is of great importance for all association studies carried out on this population background.

*Lit.: **1.** Pritchard JK et al. Influence of Population Structure Using Multilocus Genotype Data., Genetics. 2000;155:945-59. **2.** Pritchard JK et al. Association mapping in structured populations. Am J. Hum Genet. 2000;67:170-81. **3.** Rueschendorf F et al. ALOHOMORA: a tool for linkage analysis using 10K SNP data. Bioinformatics. 2005;21:2123-5. **4.** Abecasis GR et al. GRR: graphical representation of relationship errors. Bioinformatics. 2001;17(8):742-3. **5.** O'Connell JR et al. PedCheck: A program for identifying genotype incompatibilities in linkage analysis. Am J Hum Genet. 1997;63:259-66. **6.** Abecasis GR et al. Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet. 2002;30(1):97-101. **7.** Gudbjartsson DF et al. Allegro, a new computer program for multipoint linkage analysis. Nat Genet. 2000 May.25(1):12-3. **8.** Kruglyak L et al. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet. 1996;58(6):1347-63. **9.** Sobel E et al. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet. 1996;58(6):1323-37. **10.** Goedde R et al. Association of the HLA region with multiple sclerosis as confirmed by a genome screen using >10,000 SNPs on DNA chips. J Mol Med. 2005;83(6):486-94.*