European Mathematical Genetics Meeting held at Diepenbeek, Belgium, 13-15 April 2000

Organised by Geert Molenberghs at the Center for Statistics Limburgs Universitair Centrum

Abstracts

Theoretical Analysis of Lethal Factors in Plant Population. D. ALMORZA GOMAR, R. BOGGIO RONCEROS, M. ESTEVEZ, E. FAVRET, C. GONZALEZ, O. LOSARDO, J.C. SALERNO and O. SORRARAIN. *Universidad de Cadiz, Spain.*

In the last years, numerous experimental studies have shown that quantitative traits are controlled by some factors with major effects, and given credibility to the conclusion that major loci exist and that one might be able to study them. These experiences are closely related to previous suggestions in papers by various authors about the association of lethal infertile genes with positive yield factors. The maize has been used as a model organism in the development and evolution of molecular markers for the mapping, identification and manipulation of major genes affecting the expression of quantitative traits in plants.

Changes in population due to selection processes depend upon the genetic variability of the population; loci where mutations have very often occurred may have a deleterious effect in the homozygous state, as has been reported by several authors. These mutant genes for some reason seem to increase the adaptability of their heterozygous carrier. The fitness of a population not only increases due to the adaptation response to changes of the environment but also for a high effectivity in artificial selection on a variety of alleles.

In this paper, we are interested in the problem of the existence of some linkage groups in which lethal infertile genes associated with yield factors are present. These groups should be considered as special groups of the major genes affecting quantitative traits in plants or animals. This hypothesis is supported by the finding of natural balanced lethal systems, which maintain a short segment heterotic. These balanced lethal systems should be useful to maintain the linkage in many segments, because of the high frequency of lethals in the population, and they are also useful to show the maximum genetic load that can be tolerated by plants or animals. These considerations moved us to study the maize balanced lethal systems as a mechanism in order to identify, to isolate and to keep up through the balanced lethal systems.

In addition, it is very important to predict the number of generations that a lethal system maintains the linkage, and also the distance between the genes with two or more number of lethal genes at the repulsion phase; in this way, a mathematical method of analysis was carried out, and it will be reported in this paper and following ones. For this purpose the theory of Markov chains discrete in space and time is used for the analysis.

REFERENCES

Barucha Reid, A. T. (1960). Elements of the Theory of Markov processes and their Aplications. New York, McGraw-Hill.

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

Boggio Ronceros, R., Sorarrain, O., Salerno, J. C. & Favret, E. (1997). Theoretical Analysis of Lethal Factors in Plant Populations. *Mathematical Biosciences*.

Bosso, J. A., Sorarrain, O. M. & Favret, E. A. (1969). Application of Finite Markov Chains to Sib Maiting Population with Selection Biometrics.

Burnam, Ch. R. (1993). Maize Genetics Cooperation Newsletter **67**.

Carson, H. L. (1967). Permanent heterozygosity, Evolutionary Biology.

Crumpacker, D. W. (1967). Genetics loads in maize (*Zea mays L.*) and other cross fertilized plants and animals. *Evolutionary Biology*.

Salerno, J. C., Boggio, R. & Sorarrain, O. (1999). Análisis teórico de rendimiento en plantas reguladas por factores letales. Revista de Agricultura.

A Tool for Checking, Exploring, and Debugging Pedigrees.

R. ALVAREZ, J. A. BARO, C. CARLEOS, D. GARCIA and H. LAMELAS. *SERIDA-SIOM, Gijon, Spain*.

Introduction. A tool is presented for the preparation of genotypes of the kind used in genetic data analysis. It permits checks on consistency and complexity, and may perform error corrections with several degrees of conservativeness. Taking a file with identity-sire-dam-(genotype) alphanumeric records as input, this program builds a pedigree tree that may be used for the calculation of several coefficients of genetic diversity.

Methods. This is achieved through use of a structured type consisting of an integer (numerical identity), two pointers to string (alphanumerical identity), two pointers to structured type (sire and dam) and a real value (inbreeding coefficient). Animals are allocated within the structured vector by first piling up the three identity columns and simultaneously fetching each individual's sire and dam. A second step performs a pedigree depuration that detects errors such as change of sex, repeated non consistent records (self-sire, self-dam or different sets of parents), and circular pedigrees (animals as self-ancestors). Each error condition is documented and collected in a report file. There is a choice of implemented correction policies and the pedigree file may be overwitten with a corrected one. Additionally, the program may create subpedigrees for a given data set and a master pedigree file by recursively including all ancestors of the animals in the data set.

Several genetic diversity coefficients are calculated on the pedigree. The inbreeding coefficient (F) is calculated by exploration of every possible path between the parents of each individual that have a common ancestor as a node. The effective number of founders (fe), the effective number of ancestors (fa), and the average relatedness (AR) are calculated by similar means.

The program was coded in fortran 90, makes use of discrete mathematic tools and compiles and runs at reasonable speed in several operating systems.

REFERENCES

Boichard, D., Maignel, L. & Verrier, E. (1997). The value of using probabilities of gene origin to measure genetic variability in a population. *Genet. Sel. Evol.* **29**, 5–23.

Models for Familial Breast Cancer, Incorporating *BRCA1*, *BRCA2* and Other Genes.

A. C. ANTONIOU, D. F. EASTON, G. McMULLEN, P. D. PHAROAH and B. J. PONDER. *University of Cambridge, United Kingdom*.

Introduction. Recent studies have shown that a significant proportion of families with multiple cases of breast cancer are not due to either *BRCA1* or *BRCA2*. In computing the probability that a woman is a *BRCA1/2* gene carrier for genetic counselling purposes, it is important to allow for the

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

fact that other breast cancer susceptibility genes may exist. Objectives: We used data from both a population based series of cancer cases and high risk families in the UK, with information on *BRCA1* and *BRCA2* mutation status, to investigate the genetic models that can best explain familial breast cancer outside *BRCA1* and *BRCA2* families. We also attempted to evaluate the evidence for risk modifiers in *BRCA1* and *BRCA2* carriers.

Methods. We estimated the simultaneous effects of *BRCA1*, *BRCA2*, a third hypothetical gene "*BRCA3*", and a polygenic effect using segregation analysis. We also allowed for the modifying effect of other genes on the risks of *BRCA1/2* mutation carriers. The Hypergeometric Polygenic Model was used to approximate polygenic inheritance and the effect of risk modifiers. The models were assessed by likelihood comparisons and by comparison of the observed numbers of mutations and affected relatives with the predicted numbers. All the models were implemented in the computer program MENDEL.

Results. *BRCA1* and *BRCA2* could not explain all the observed familial clustering. The best fitting model for the residual familial breast cancer was the polygenic, and there was significant evidence for a modifying effect of other genes on the risks of *BRCA1* and *BRCA2* mutation carriers. Under this model the frequency of *BRCA1* was estimated to be 0.072% (95%CI, 0.03-0.16%) and of *BRCA2* 0.082% (95%CI, 0.04-0.16%). The breast cancer risk by age 70 for *BRCA1* carriers "free" of modifiers was estimated to be 40% and the corresponding risk for *BRCA2* was 66%.

Conclusions. These findings suggest that a number of common, low penetrance genes with additive effects may account for the residual non-*BRCA1/2* familial aggregation of breast cancer. The modifying effect could explain the previously reported differences between population based estimates for *BRCA1/2* penetrance and estimates based on high-risk families.

Beyond Infinitesimal Model BVs; On the Use of Identified QTL for the Estimation of Polygenic Breeding Values. R. ALVAREZ, J. A. BARO, C. CARLEOS, D. GARCIA and R. PONG WONG. *SERIDA-SEMIO,Gijon, Spain.*

Introduction. Polygenic breeding values can be derived by application of mixed model equations on phenotypic records of the individual and its relatives. They have been used for the identification of individuals with maximum genetic merit in most selection programs, and for monitoring actual selection response. Superation of the infinitesimal model assumption by use of an identified QTL in selection requires the joint estimation of BVs for the known QTL with those for polygenic BVs.

Methods. The effects of different genotypes at a single QTL are often estimated by Ordinary Least Squares. When some individuals are relatives, the sharing of alleles at other loci influencing the trait will induce correlations between residuals not accounted for by OLS. Similarly, mixed model BV estimation ignoring QTL genotypes may lead to incorrect values due to wrong assumptions on the similarity among individuals through sharing of QTL alleles. Results are presented for a program that takes account of Mendelian segregation of known, discrete genes and polygenic genes. This procedure jointly estimates QTL and polygenic BVs by Gibbs sampling using probabilistic estimates for the QTL incidence matrix and is thus applicable to data sets with missing genotypes, almost a certain condition for most breeding programs. It was applied to a beef cattle test station data set with animals genotyped for the muscular hypertrophy (*mh*) gene and measured for a large

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

number of zoometry and growth traits. Results were compared when the QTL genotype was fitted as a fixed effect, or completely ignored.

Conclusions. It was found that the *mh* allele is partially dominant, decreasing body length and height, testicle size, and average daily gain, but increasing thorax diameter and, obviously, muscular score. Due to the small number of connections available in the pedigree, no definite conclusions can be made on the partition of genetic variance, though allele segregation seems to account for all of the genetic variance of testicle size.

REFERENCES

Hofer A. & Kennedy B. W. (1993). Genetic evaluation for a quantitative trait controlled by polygenes and a major locus with genotypes not or only partially known. *Genet. Sel. Evol.* **25**, 537–555.

Unusual Group II Introns in Plant Mitochondria.

L. BONEN and C. CARRILLO. *Biology Department, University of Ottawa, Canada.*

Introduction. Virtually all cis-splicing and trans-splicing introns present in the mitochondrial genes of flowering plants have been categorized as group II based on secondary structural features [Michel & Ferat, 1995]. Ribozymic group II introns are characterized by six helical domains, with the highly-conserved domain 5 comprising part of the catalytic core and domain 6 containing a bulging adenosine for lariat formation in splicing. Plant mitochondrial introns, however, in many cases display weak helical structure within core sequences and show unexpected sequence divergence among homologues from different plants.

Methods. We have assessed the impact of sequence variation on the ability of the domain 5/6 regions of intron RNAs to be folded into conventional structures, and examined whether RNA editing improves helicity by the conversion of A-C mispairs to A-U pairs. In this study, we have characterized intron RNA and DNA sequences from plants such as wheat, rice, tobacco, soybean, pea and Arabidopsis, using previously described methodologies [Carrillo & Bonen, 1997].

Results. In our survey of the domain 5/6 regions from eight different introns, we observed editing at only approximately one-third of the A-C mispairs. For example, within domain 5, three out of nine candidate sites showed editing in excised intron RNAs; moreover, one of these was edited in only some plant species. Other atypical features include non-AC mispairs within domain 5/6 helices, absence of bulging adenosine within domain 6, and plant-specific variation in domain 5/6 linker length. As a consequence, such introns cannot be folded into either the classical group II structures or ones shared among plants.

Conclusions. Our comparative RNA analysis suggests that the core structures of these apparently degenerate plant mitochondrial introns are under reduced evolutionary constraint compared to conventional group II introns. This raises the possibility that certain functions have been taken over by proteins and/or small RNAs.

REFERENCES

Carrillo, C. & Bonen, L. (1997). RNA editing status of nad7 introns in wheat mitochondria. *Nucleic Acids Res.* **25**, 403–409.

Michel, F. & Ferat, J. L. (1995). Structure and activities of group II introns. *Ann. Rev. Biochem.* **64**, 435–461.

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

Estimating Effect of a QTL by Trait-selected Pooled Samples. R. ALVAREZ, J. A. BARO, C. CARLEOS, N. CORRAL, D. GARCIA and T. LOPEZ. *Escuela Superior de Ingeniera Informatica, Gijon, Spain.*

Introduction. Estimators for the effect of genotype on a quantitative phenotype are explored in the situation that allele frequencies, rather than individual typings, are available.

Genotyping is usually the most expensive investment of a genetic mapping experiment. Numerous attempts have been made to reduce the burden of genotyping, without decreasing the power of detection of genotype-phenotype association.

Stuber *et al.* (1980) suggested studying marker allele frequencies in selected lines (selective genotyping). Arnheim *et al.* (1985) introduced the idea of using pooled DNA samples in the context of case-control studies; this technique implies no knowledge about individual genotypes, only about allele frequencies. Combination of both strategies may lead to an important saving in the number of genotypings.

Methods. The phenotypic distribution of a continuous trait was modeled as a mixture of normal distributions, with mean and variance depending on genotype (heteroskedastic model). Likelihood functions are presented for different levels of availability of genotypic and phenotypic information. Various ways of performing the selection (number of selected pools in a phenotypic distribution, defined within fixed thresholds or for fixed proportions) are discussed. An alternative set of heuristic estimators is introduced that proves to be as accurate, both asymptotically and for moderate-sized samples, as the very computationally demanding Maximum-Likelihood estimators.

Conclusions. We pay special attention to the half-sib family design, and examine the factors potentially affecting the accuracy of estimators: allelic frequencies, additive and dominance gene effects, recombination rate, technical error for frequencies in pools, and family and sample sizes.

REFERENCES

Arnheim, N., Strange, C. & Erlich, H. (1985). Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of HLA class II loci. *Proc. Natl. Acad. Sci.* **82**, 6970–4.

Stuber, C. W., Moll, R. H., Goodman, M. M., Schaffer, H. E. & Weir, B. S. (1980). Allozyme frequency changes associated with selection for increased grain yield in maize (Zea mays L.). *Genetics* **95**, 225–36.

Sewall Wright's equation $\Delta q = (q(1-q)\varphi w/\varphi q)/2w$. A. W. F. EDWARDS. *Gonville and Caius College, Cambridge, U.K.*

An equation of Sewall Wright's $\Delta q = (q(1-q)\varphi w/\varphi q)/2w$ expresses the change in the frequency q of an allele under selection at a multiallelic locus as a function of the gradient of the mean fitness 'surface' w in the direction in which the relative proportions of the other alleles do not change. An attempt to derive this equation using conventional vector calculus shows that this description leads to a different equation, and that the purported gradient in Wright's equation is not a gradient of the mean fitness surface except in the diallelic case, when the two equations are the same.

Some Issues Regarding Breed Allocation Using Microsatellite Markers. R. ALVAREZ, J. A. BARO, J. CANON, C. CARLEOS and D. GARCIA. *Universidad de Madrid, Spain.*

Introduction. Allocation of individual animals to their corresponding breeds is one of the many purposes for which molecular markers can reveal themselves as a powerful tool.

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

Methods. In this study we take microsatellite markers to apply classical statistical methods of decission theory to solve this problem. Data for the estimation of allele frequencies were obtained from five Spanish cattle and seven Spanish horse breeds with 16 and 13 microsatellite markers genotyped, respectively.

Results. Animals were classified attending to a maximum probability criterion, and misclassification rates were assessed, varying from 1% error in cattle when allocating to one of the five breeds, to 10% (alpha = 3D0.1) of wrong assignments when distinguishing between two particular horse breeds.

Conclusions. However, three issues of discussion arose while performing these analyses which allow several approaches for solution. In the first one the treatement of outlier alleles is examined. These are alleles which are significantly "unusual" in the contingency tables as to characterise one breed versus any other. We deal with the problem of accurately determining the above referred significance. The second issue is somewhat a particular case of the first. It has to do with missing alleles and the fact that alleles whose estimated frequency is zero for a particular breed force us to reject individuals bearing such alleles as members of that breed despite further positive evidence of their belonging to it. Finally, allocation significance is studied, as the proposed method provides misclassification rates only in a generic sense, not for the individual animal, but for the whole of the breed. Several ways of solving each of these three topics are proposed and discussed through stochastic simulations.

REFERENCES

Paektau, D., Calvert, W., Stirling, I. & Stobeck, C. (1995). Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology* **4**, 347–354.

Comparing Huge Sets of SAGE (Serial Analysis of Gene Expression)-data. H. N. CARON and S. H. HEISTERKAMP. *University of Amsterdam, The Netherlands.*

Introduction. Advances in molecular biology allow one to gather huge amounts of gene expression data in a routine fashion. Furthermore, datasets are available on the web, which can be downloaded to compare with one's own dataset. The raw data consist for each set (bank) of "tags"and the number of times that tag appears, with no zero counts. The number of unique tags in the most common databases is less than 50 000, the total count may be as high as 900 000. Earlier publications [Audic and Claverie, 1997; Chen *et al.* 1998)] on the analysis of these datasets discussed the way 'p-values' should be interpreted when using the comparison of tag by tag count of the raw data (assuming either a binomial or a beta-binomial distribution).

Objectives. We are interested in the comparison of gene expression data from cell lines of different tumour tissues from children, in order to detect differences which may direct further research into the genes involved in the particular tumour. That is we want to 'sieve' datasets for possible differences or looking for patterns of agreement/ disagreement among several datasets.

Methods. We used the (empirical) Bayesean framework to arrive at stable estimates of the abundance of each unique tag (species), thereby allowing for the non-zero counts and possible bias towards some tags. Estimated abundancies are then used for comparison, based on a F-distribution with non-integer degrees of freedom, or used in a cluster analysis when using more datasets at one time.

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

Results. We arrived at a zero-truncated negative-binomial distribution for the numbers of each species seen. An efficient algorithm was developed to calculate the stable estimated abundancy. When comparing two datasets of 50 000 tags, of which the gene was manipulated in on or off, we still sieved 1300 tags to be of interest.

REFERENCES

Audic, S., Claverie, J.-M. (1997). The significance of digital gene expression profiles. *Genome Research* **7**, 986–995.

Chen, H., Centola, M., Altschul, S. F. & Metzger, H. (1998). Characterization of Gene expression in resting and activated mast cells. *Journal of experimental medicine*, 188, **9**, 1657–1668.

Admixture Association in Genetic Epidemiology. J. J. HOTTENGA[1,2], J. J. HOUWING[1], L. A. SANDKUIJL[1,2] and C. M. VAN DUIJN[1]. [1] *Department of Epidemiology & Biostatistics, Erasmus University Rotterdam, The Netherlands,* [2] *Department of Human Genetics, Leiden University, The Netherlands.*

The genetic association study is a powerful tool to study candidate genes via a case-control design. Like all case control studies however, it is susceptible to confounding, leading to false positives. One form of confounding occurs when the cases and controls are sampled from two or more sub-populations that differ with respect to allele frequencies. Association can then be found even if there is no causal relationship between the DNA variant and the disease. The aim of the present study is to quantify the problem of admixture by means of simulation. In this study two populations were simulated – one of which was a large population with stable allele frequencies, while the other population was small and therefore susceptible to genetic drift. With the allele frequencies from the simulated populations odds ratios were calculated under various conditions of admixture. Given an odds ratio threshold, the probability of finding a positive association was calculated. Our results show that only major differences in allele frequencies will lead to spurious admixture association. Such pronounced differences will only be observed if a sub-population has been exposed to a high level of drift due to prolonged isolation.

On the Differences between Maximum Likelihood and Regression Interval Mapping in the Analysis of Quantitative Trait Loci. C.-H. KAO. *Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, Republic of China.*

The likelihood of interval mapping model is generally a finite normal mixture (Lander & Botstein, 1989; Jansen, 1993; Zeng, 1994; Kao *et al.* 1999). In the computation of the maximum likelihood estimates (MLE) of the finite normal mixture model, the iterative expectation-maximization (EM) algorithm (Dempster *et al.* 1977) is a broadly applicable approach as algorithms such as Newton-Raphson and Fisher's score methods may turn out to be complicated. When the number of QTLs considered in the model increases, the numbers of mixture components and parameters in the likelihood increase dramatically. As a result, the maximization of the likelihood and derivation of the MLE through the EM algorithm could become difficult to evaluate and obtain; moreover, when mapping the entire genome for QTL, the estimation needs to be performed at every position of the genome. Therefore, the maximum likelihood (ML) estimation by the EM algorithm is often regarded to be complex in derivation of MLE and computationally expensive in QTL mapping (Haley & Knott, 1992; Xu, 1998). In view of these difficulties in estimation, regression (REG) interval mapping,

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

which regresses the quantitative trait value on the conditional expectation of QTL genotype, was proposed to approximate ML interval mapping as well as save time in computation (Haley & Knott, 1992; Martinez & Curnow, 1992). Although REG interval mapping lacks some attractive properties, such as consistency and asymptotic efficiency, as compared to ML interval mapping in statistical inference, and may suffer from the lack of interpretability in terms of genetic model (Haley & Knott, 1992; Jansen, 1993), it is often claimed that the two approaches provide virtually similar or identical estimates and test statistics in QTL mapping (Haley & Knott, 1992; Xu, 1998). As a consequence, the REG method has been widely accepted and applied to QTL mapping studies by many researchers (Haley, Knott & Elsen, 1994; Whittaker *et al*. 1996; Xu, 1996 1998; Lebreton *et al*. 1998; Goffinet & Mangin, 1998; Dupuis & Siegmund, 1999). Although REG may approximate ML interval mapping well in some cases as shown by Haley & Knott (1992) and Xu (1998), their differences in the estimation of QTL parameters do exist and could be significant in practical QTL mapping as shown in this article. Unfortunately, there is not much attempt to investigate these differences in the literature. Xu (1995) pointed out that the estimation of residual variance by REG interval mapping is biased. In this paper, the differences between the two approaches in the estimation of and testing for QTL parameters due to several factors, such as heritability, size of interval, relative QTL position in an interval, the difference between QTL effects, epistasis and linkage between QTL, are investigated both analytically and numerically by simulation. With the understanding of the factors affecting the differences between the two methods, a more efficient, precise and powerful strategy using both methods can be explored in QTL mapping. The QTL mapping properties under these factors are also investigated and and discussed.

The differences between ML and REG interval mapping in the analysis of QTL are investigated analytically and numerically by simulation. The analytical investigation is based on the comparison of the solution sets of the ML and REG methods in the estimation of QTL parameters. Their differences are found to relate to the similarity between the conditional posterior and conditional probabilities of QTL genotypes, and depend on several factors, such as heritability of a quantitative trait, relative QTL position in an interval, interval size, difference between the sizes of QTL, epistasis and linkage between QTL. The differences in mean squared error (MSE) of the estimates, likelihood ratio test (LRT) statistics in testing parameters and power of QTL detection between the two methods become larger as (1) the heritability becomes higher, (2) the QTL locations are positioned toward the middle of intervals, (3) the QTL are located in wider marker intervals, (4) epistasis between QTL is stronger, (5) the difference between QTL effects becomes larger and (6) the positions of QTL get closer in QTL mapping. Especially, the REG method is biased in the estimation of heritability and residual variance of a quantitative trait, and it may have serious problem in detecting closely linked QTL. Consequently, the differences between the ML and REG methods can be significant, and the approximation of the REG to ML method may be poor in practical QTL mapping. In general, the ML method tends to be more powerful and to give estimates with smaller MSE and larger LRT statistic. This implies that ML interval mapping can be more accurate, precise and powerful than REG interval mapping. The REG method is faster in computation, especially when the number of QTL considered in the model is large. Recognizing the factors affecting the approximation of the REG to ML interval mapping can help an efficient strategy using both methods in QTL mapping to be outlined. Also, QTL mapping is found to be more precise and powerful when

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

QTL are located on the boundary of intervals, in narrow intervals, loosely linked and with weaker epistatic interaction.

REFERENCES

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **39**, 1–38.

Doerege, R. W. & Churchill, G. A. (1996). Permutation test for multiple loci affecting a quantitative character. *Genetics* **142**, 284–294.

Dupuis, J. & Siegmund, D. (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics* **151**, 373–386.

Goffinet, B & Mangin, B. (1998). Comparing methods to detect more than one QTL on a chromosome. *Theor. Appl. Genet.* **96**, 628–633.

Haley, C. S. & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

Haley, C. S., Knott, S. A. & Elsen, J.-M. (1994). Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* **136**, 1195–1207.

Jansen, R. C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205–211.

Kao, C.-H., Zeng, Z.-B & Teasdale, R. D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203–1216.

Lander, E. S. & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Lebreton, C. M., Visscher, P. M., Haley, C. S., Sewmikhodskii, A. & Quarrie, S. A. (1998). A nonparametric bootstrape method for tesing close linkage vs. pleitropy of coincident quantitative trait loci. *Genetics* **150**, 931–943.

Whittaker, J. C., Thompson, R. & Visscher, P. M. (1996). On the mapping of QTL by regression phenotype on marker type. *Heredity* **77**, 23–32.

Xu, S. (1995). A comment on the simple regression method for interval mapping. *Genetics* **141**, 1657–1659.

Xu, S. (1998). Further investigation on the regression method of mapping quantitative trait loci. *Heredity* **80**, 364–373.

Zeng, Z.-B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468.

C++ Library for Simulating Genetic Transference. J. A. BARO[1], C. CARLEOS[2], H. LAMELAS[3] and F. MENENDEZ[4]. [1] *Genetic Unit, Serida, Spain,* [2] *Dpt. of Statistics, Universidad de Oviedo, Spain,* [3] *CoALA, Universidad de Oviedo, Campus de Viesques, Gijon, Spain,* [4] *EUITIG, Universidad de Oviedo, Spain.*

A C++ library for simulating the process of genetic information transference is presented. Its main application is in the study of genetic linkage, but it was designed to be as general as possible. The use of object oriented programming (OOP) allows straightforward use of the library by any programmer with total disregard of its internal structure. A user interface for graphical environments, capable of concurrent simulation management, is being developed.

Introduction. A C++ class library has been built that simulates the inheritance process, specifically the gametogenesis (meiosis) and the generation of a new individual (mating). The need for this tool arose from our genetic linkage studies with increasingly complex models that led us to unify our programming under a general "genetic objects" approach, in lieu of applying *ad hoc* Perl algorithms. The main biological entities were represented by a C++ class hierarchy.

The library is focused on linkage analysis. Notwithstanding, it is easy to reliably extend the implemented classes or to add new ones. Most usual experimental designs in animal breeding have been reproduced (F_2, back-cross, half-sib families).

Though the library is oriented towards the study of segregation of single (discrete) genes, it supports the mixed inheritance model, allowing an infinitesimal polygenic background. Multi-trait analyses are taken into account and, as well as this, single genes that may affect various traits (pleiotropy).

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

Design. At the heart of the design is a series of classes that represent biological entities: allele, locus, chromosome, genome, haplotype, trait, phenotype, individual, breed (population), species. Class "species" holds the genetic map (position of loci in chromosomes and magnitude of effects on traits) and class "breed" sets allele frequencies and environmental variances.

An individual's phenotype for a certain trait is modeled as the sum of a polygenic value, plus single gene contributions, plus an environmental Gaussian deviate. Gene interaction (epistasis) is assumed null, but allele interaction (dominance) is fully accounted for. Each trait affected by a pleiotropic locus has a symmetric matrix $m \times m$ associated with it, with m being the number of alleles at the locus; element (i, j) is the gene effect when alleles i and j are present. Gene effects are modeled as real numbers because we are mainly interested in quantitative traits, but an adaptation is under development to allow use of fuzzy numbers as implemented by the first author in the library FAIR.

Simulation of any pedigree structure may be performed from the above-defined classes. Further, three special C++ classes were created to automate tasks in simulating the most usual experimental designs: F_2, back-crosses and half-sib families. When defining a simulation, the user is offered a choice of random number generators (system call, "Mersenne Twister"), and map functions (complete interference, Haldane, Kosambi); these lists are easily extendable. A simulation result consists of a list of individuals, and several objects may summarize information, i.e. allelic counts by user-defined categories. The library uses a tag-based, HTML-like format for input and output. For input, first lines define general parameters (family type and size, random number generator, map function). Next, genome map is defined. Allele frequencies and environmental variances are set in successive sections, corresponding to populations or breeds involved in the simulation. For output, some other formats, such as Cri-Map, Genepop and Pedigree Viewer are available as well.

Discussion. There are several reasons behind the choice of C++ as the programming language. It is a widely used language, almost all platforms have C++ compilers, and an ANSI standard is available. It has a powerful support for OOP and the generated code is highly efficient. Besides, it is linkable with other programming languages as C, Java, Ada, Perl etc. Most OOP and C++ programming tools have been used in the library, in order to achieve a high degree of flexibility: inheritance, composition, polymorphism, and dynamic linking.

The OOP paradigm has been chosen for the sake of re-usability, that is, inclusion of single library objects in other projects. Resulting designs are clear and elegant, with source code easy to maintain and modify.

A window-graphic environment interface is being developed, with a multi-threaded structure. It is expected to run under several platforms (MS Windows, Unix). We hope to extend the library to manage threshold models and to automate simulations for other experimental designs.

REFERENCES

Lamelas, H. & Riesgo, E. (1998). *Computer-Aided Fuzzy Number Arithmetics and Graphical Representation*, Proyecto Fin de Carrera, EUITIG, Gijon, Asturias, Spain. Web page: http://www.coala.uniovi.es/fair

Matsumoto, M. & Nishimura, T. (1998). Mersenne Twister: a 623-dimensionally equidistributed uniform pseudo-random number generator, *ACM Trans. Model Comp. Simul.* **8**, 3–30.

A Generalized Estimating Equation Approach to Mapping of Quantitative Trait Loci (QTL).

C. LANGE and J. C. WHITTAKER. *Department of Applied Statistics, University of Reading, U.K.*

During the last years several statistical methods have been developed which allow one to map

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

quantitative trait loci (QTL) relatively to markers. The so far established methodology assumes normality or is restricted in the sense that it is only possible to map QTLs corresponding to one single trait. However it can be of great economical interest to map QTLs of traits with non-normal distributions simultaneously, e.g. the QTLs which are responsible for the height of a maize type, the numbers of grains in the maize cob, the time until harvest and the probability of being resistant against a certain disease. I am going to discuss a QTL-mapping approach which is based on generalized estimating equations and provides the above mentioned mapping facilities. With this approach the relative position of a QTL to its flanking markers and all genetic parameters, as additive or dominance effect components, can be estimated. The variance-covariance matrix of all estimates can be easily obtained by a variance sandwich-estimator which is robust against miss-specification of the correlation structure of the environmental error. Furthermore the approach is tested in several simulation experiments. These experiments show, compared with transformation of the data to near-normality and using standard mapping procedures, that considering the special trait types already during the construction of the mapping procedure yields a substantial increase of efficiency in all estimates.

REFERENCES

Hackett, C. A. & Weller, J. I. (1995). Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics* **51**, 1252–1263.

Jiang, C. & Zeng, Z. B. (1995). Multiple trait analysis of genetic mapping for quantitative trait loci. **140**, 1111–1127.

Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Prentice, R. L. & Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47**, 825–839.

Haplotype Analysis in Association Studies when Phase is Unknown.

D. G. CLAYTON and A. P. MANDER. *MRC Biostatistics Unit, Cambridge, U.K.*

Introduction. It has been a longstanding complaint of association analysis that spurious associations can be explained by admixture or ethnic stratification. In epidemiological terms when a variable is independently related to both disease status and genotype but not on the causal pathway of the genotype and disease it is said to be a confounder. The method of analysis is then by stratifying the analysis by the confounder. The algorithm discussed here was first published in [1] as a solution for clustering markers when investigating association but can also be used to perform stratified analysis. Log-linear models are fitted using iterative proportional fitting (IPF) [2]. IPF can handle huge dimensional contingency tables, even when the likelihood is badly behaved. Additionally the IPF algorithm can handle constrained estimation and hence profile likelihoods can be constructed for the odds ratio of interest.

Objectives. The main purpose of this article is to introduce a function written for the statistical package Stata that can test various hypotheses of association and estimate odds ratios and confidence intervals. Application of the function investigates power in case-control SNP association studies. In particular how much is power improved when the controls have phase known compared to phase unknown.

Conclusions. For SNP association studies it appears that the additional effort for obtaining phase for controls does not lead to huge gains in power.

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

REFERENCES

Agresti, A. (1992). Modelling patterns of agreement and disagreement. *Statistical methods in medical research*. **1**, 201–218.

Chiano, M. N. & Clayton, D. G. (1998). Fine genetic mapping using haplotype analysis and the missing data problem. *Ann. Hum. Genet.* **62**, 55–60.

QTL Mapping Using Dynamic Programming. J. FLINT and R. MOTT. *University of Oxford, U.K.*

We describe a new multipoint method using dynamic programming for determining haplotype distributions from genotype data for outbred populations. We apply the technique to map QTLs for emotionality in mice [Talbot *et al.* 1999] and show that the new method detects QTLs which are not detectable using standard linkage methods.

REFERENCES

Talbot, C. J., Nicod, A., Cherny, S. S., Fulker, D. W., Collins, A. C. & Flint, J. (1999). High-resolution mapping of quantitative trait loci in outbred mice. *Nat Genet. Mar.* **21**(3), 305–8.

Effect of the Population Structure on Linkage Disequilibrium Mapping. P. V. BARET and J. NSENGIMANA. *Unité de Génétique, Université catholique de Louvain, Louvain-la-Neuve, Belgium.*

In order to take advantage of the forthcoming availability of dense marker maps, new methods of genetical analysis based on linkage disequilibrium (LD) have been recently proposed (Terwilliger & Weiss, 1998; Kruglyack, 1999). However, results have been so far restricted to specific human populations (e.g. Häsbacka *et al.* 1992). A main issue is to determine the power of these methods in populations affected by factors such as selection or inbreeding. The effect of selection and inbreeding on LD (Lewontin, 1964; Hill & Robertson, 1968) is considered as the main potential limitation in the use of LD mapping in farm animals (Baret & Hill, 1997). A software is developed to compare the statistical power of LD mapping methods. Different family structures are simulated on a 20 generation period. Fifteen markers are evenly spaced on a 42cM chromosome. Individuals are randomly mated in each pedigree. For each individual, genotypic data (phases and alleles at each locus) and pedigree information are recorded. A new measure of LD, D1, i.e. the frequency of haplotypes with the common phase from a given sire at a pair of marker loci, is introduced. D1 is estimated with respect to the genetic distance [D1 (d)] and a graphical representation of this parameter along the chromosome and across generations is proposed. The highest values of LD are observed on the shortest chromosome segments and in the very first generations. The mean inbreeding coefficient is also estimated in generation 20 of each population and LD [D1 (d)] is plotted against this inbreeding level (IL). The higher the IL, the higher the LD. Plots of D1 (d) show a minimum value (=sill) reached at a given genetic distance (=range). A comparison of different pedigrees reveals that for every mating structure there is a unique pairing of the sill and the range. The sill and the range may then characterise the overall LD level in any population at a given generation. Deeper insight is to be gained in this study. A geostatistical approach will be used in modelling plots of D1 (d) and analytical values of the sill and the range characterising different pedigrees will be derived. The possibility of using these parameters in LD mapping models will be investigated.

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

We are grateful for the financial support of the Belgian Fund for Research in Agriculture and Industry (FRIA). Corresponding author: baret@gena.ucl.ac.be.

REFERENCES

Baret, P. V. & Hill, W. G. (1997). Gametic disequilibrium mapping: potential applications in livestock. *Anim. Breed. Abstr.* **65**, 309–319.

Häsbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A. & Lander, E. (1992). Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nat. Genet.* **2**, 204–211.

Hill, W. G. & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38**, 226–231.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22**, 139–144.

Lewontin, R. C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics.* **49**, 49–67.

Terwilliger, J. D. & Weiss, K. M. (1998). Linkage disequilibrium mapping of complex disease : fantasy or reality? *Curr. Opin. Biotech.* **9**, 578–594.

Role of Selection, Drift, Mutational Bias and Linkage in the Evolution of DNA Base Content.

C. GAUTIER, G. PIGANEAU, B. TOURANCHEAU and R. WESTRELIN. *Université Claude Bernard, Lyon, France.*

We investigated the role selection could have played in the appearance of long DNA sequences with biased $G+C/A+T$ base composition. Since the neutral mutational bias model (Sueoka, 1988), which assumes local variation in mutation pressure, no alternative selective model has been proposed. This lack of selective models is due to the mathematical complexity of the models considering selection and linkage (Hospital & Chevalet, 1996). We used computer simulations to obtain the mutation-drift-selection equilibrium value. We studied the evolution of a haploid population of N sequences of L bases W (A and T) and S (G and C). The mutation process between W and S is biased: W mutates in S at rate u, S mutates in W at rate v, and follows a Poisson distribution of mean $(u+v)*L*N$. Selection occurs as a multinomial draw of the N next sequences each affected by a fitness value. The fitness is additive beween sites and the selection coefficient s per base is positive.

We found that the currently used diffusion approximations, that predict that the equilibrium value depends only on the products Nu, Nv and Ns, do not hold for $Ns > 1$. The discrepancy between the equilibrium values for different population sizes increases with increasing s, the efficiency of selection increasing with population size. We also observed the decrease of the equilibrium with the sequence length as previously observed by Li (1987) and Comeron *et al.* (1999), raising the question of a selection limit due to linkage under classical fitness functions shemes.

This work has strong implications in molecular evolution where the dynamics of base composition is unknown under selection pressure. Obtaining a mathematical theory of the evolution of L linked sites under selection is an important goal for future work.

REFERENCES

Comeron, J. M., Kreitman, M. & Aguadé M. (1999). Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. *Genetics* **151**, 239–249.

Hospital, F. & Chevalet, C. (1996). Interactions of selection, linkage and drift in the dynamics of polygenic characters. *Genet. Res.* **67**, 77–87.

Li, W. H. (1987). Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J. Mol. Evol.* **24**, 337–345.

Sueoka, K. (1988). Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. USA* **85**, 2653–2657.

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

Experimental Design for Genomic Mismatch Scanning: Exploring Different Models for the Recombination Process. J. H. BARRETT, S. JOHN, T. H. PINEL and J. WORTHINGTON. *ARC Epidemiology unit, University of Manchester, U.K.*

Introduction. Genomic mismatch scanning (GMS) is a recently developed technique, in which identical-by-descent (IBD) regions in the DNA of two related individuals are isolated and identified. When applied to affected pairs of relatives the method provides a potentially useful tool in the mapping of disease genes. The number and spacing of markers required to identify the IBD regions have yet to be investigated.

Objectives. To estimate the number and the length of regions shared by different relatives under different models of the recombination process.

Methods. To estimate the number and length of the regions shared by two related individuals, a program was written to simulate recombination events within a single chromosome. The recombination events were at first assumed to follow a Poisson process and the model was then expanded to incorporate interference and a variable recombination rate. For the basic Poisson model, the recombination rate was set to be two. For the second more complex model, the overall rate remained the same, but the rate varied within the chromosome, with the telomeres having a rate twice as high as the central region. All the simulations were conducted under the assumption that no disease gene was present, and were run for relatives between two and ten meioses apart.

Results. There was a small but consistent decrease in the mean number of shared regions in the second model compared with the basic Poisson model; the decrease in the variance of the number and length of shared regions is more marked.

Discussion. Refinements to the Poisson model make a modest change to the mean and variance of the number and length of regions shared by a relative pair. This information will inform experimental design when applying GMS.

An Algorithm to Characterise the Noncommunicating Classes on Pedigree Data. S. A. SISSON. *Department of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, U.K.*

Because of its simplicity, the Gibbs sampler is a popular tool for sampling from the genotypic state space on pedigrees. However, unlike many other applications of the Gibbs sampler, Markov chains thus constructed in the analysis of pedigree data risk sampling from only part of the state space. Reducibility with any such sampling scheme is caused when the presence of genotypic data on typed individuals constrain other nodes in the pedigree graph to be of particular genotypic configurations.

While this problem can be overcome with the use of more general samplers, their very generality results in slow mixing and prolonged simulation time. The preferred solution is to design a fast and efficient pedigree specific sampler, based on the identification of all irreducible noncommunicating sets. An algorithm by Lin *et al.* (1994) is able to identify all such sets on a pedigree, under the specific conditions in which observed genotypes on children in a given marriage cause reducibility problems for their parents. This method, however, can be shown to fail for a number of basic examples (Jensen & Sheehan, 1997).

A new algorithm is presented, designed to overcome the limitations observed in the existing method, by considering the effect of a specifically ordered series of local state space restrictions

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

on the discrete pedigree state space as a whole. Implementation is illustrated by means of a simple worked example.

REFERENCES

Jensen, C. S., Sheehan, N. (1997). Problems with the Determination of the Noncommunicating classes for MCMC Applications in Pedigree Analysis, *Technical Report R-97-5004, Aalborg University*.

Lin, S., Thompron, E., Wijsman, E. (1994). Finding non-communicating sets for Markov chain Monte Carlo estimates on Pedigrees, *American Journal of Human Genetics* **54**, 695–704.

Statistical Power of QTL Mapping Methods Applied to Bacteria Counts. P. V. BARET (Corresponding author: baret@gena.ucl.ac.be) and P. TILQUIN. *Unité de Génétique, Université catholique de Louvain, Louvain-la-Neuve, Belgium.*

Most of the QTL mapping methods assume that phenotypes follow a normal distribution, but many phenotypes of interest are not normally distributed, e.g. bacteria counts. Our objective is to assess the efficiency of QTL mapping methods applied to bacteria counts. A tool to mimic distributions of non-normal phenotypes is developed and integrated in a simulation algorithm of QTL mapping in half-sib pedigrees (Baret *et al*. 1998). The statistical power of four QTL mapping methods is compared : 1) nested ANOVA (AN), 2) least-squares (LS) and 3) maximum-likelihood (ML) (Knott *et al*. 1996), 4) nonparametric (NP) (Kruglyak & Lander, 1995; Coppieters *et al*. 1998). A bacteria-like phenotype is simulated. Each sire (n=30) is mated to 40 dams and the trait is measured on a single offspring per mating. A single marker (16 alleles) and a single QTL (2 alleles) share the same location. Overall trait heritability is 0.25. Three levels of QTL effect are considered : 2%, 8% and 18% of phenotypic variance. NP method has significantly higher power than other methods (17% [NS], 72%, 98% for the 3 QTL levels respectively). LS and AN methods have similar powers (10%, 40% and 79%) and power of ML is significantly lower than the power of other methods (7%, 14%, 37%). When a mathematical transformation (Bosseray & Plommet, 1976) is applied on raw data prior to analysis, power of LS, AN and ML are close to the power of NP. According to these first results concerning bacteria counts, it is suggested either to apply mathematical transformations before analysis, or to use NP method. These results will be confirmed with other disease traits.

REFERENCES

Baret, P. V., Knott, S. A. & Visscher, P. M. (1998). On the use of linear regression and maximum likelihood for QTL mapping in half-sib designs. *Genet. Res.* **72**, 149–158.

Bosseray, N. & Plommet, M. (1976). Transformation normalisant la distribution du nombre de Brucella dans la rate de souris inoculées par voie intrapéritoneale. *J. Biol. Stand.* **4**, 341–351.

Coppieters, W., Kvasz, A., Farnir, F., Arranz, J. J., Grisart, B., Mackinnon, M. & Georges, M. (1998). A rank-based nonparametric method for mapping quantitative trait loci in outbred half-sibs pedigrees: application to milk production in a grand-daughter design. *Genetics* **149**, 1547–1555.

Knott, S. A., Elsen, J.-M. & Haley, C. S. (1996). Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. *Theor. Appl. Genet.* **93**, 71–80.

Kruglyak, L. & Lander, E. S. (1995). A nonparametric approach for mapping quantitative trait loci. *Genetics* **139**, 1421–1428.

The Dynamic Relationship between Production and Level of Disease Resistance from an Animal Breeding Perspective. P. BIJMA, S. C. BISCHOP, J. A. M. van ARENDONK and E. H. van der WAAIJ. *Wageningen Institute of Animals Sciences, The Netherlands.*

In animal breeding, an important aim is to optimize production level, given the animal's genetic

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

capacity in combination with a certain environment. Animals under infection pressure, as for example is the case with certain nematode infections in sheep, will show a reduced production. The relation between resistance to infection and production is modelled for a changing level of resistance. The model assumes that, given a certain infection pressure, there is a threshold for resistance below which the animal does not produce (e.g. grow). There also is a threshold above which the animal will produce at production potential. In between both thresholds animals will show a decrease in production, the size of the decrease depending on the severity of infection and the level of resistance. We have studied the changes in the relation between level of resistance and observed production level for a population under selection, where we started with a population with a mean level of disease resistance a little below the lower threshold. Selection was on observed production level, resulting in non-linear selection response for all three traits considered. With poor resistance, selection for observed production induced an increased level of resistance, resulting in increased observed production level. With increasing resistance, selection response shifts from resistance to production potential, and eventually selection for observed production is equal to selection for production potential. The rate at which observed production improved depended on the heritability, distance between the thresholds and the selection intensity. The model also showed that the phenotypic correlation between observed production and resistance increased at first and subsequently asymptoted to zero, whereas the phenotypic correlation between observed production and production potential increased and asymptoted to one.

The 2-Locus Genotype Relative Risk (GRR) Procedure. E. LESAFFRE, W. VINCK and R. VLIETINCK. *Catholic University of Leuven.*

We developed a new statistical technique to detect genetic heterogeneity and epistasis between specific loci. Our method extends the 1-locus GRR method (Schaid, 1993) to the case of 2 unlinked biallelic loci. Let A and B be two biallelic loci (alleles A: 1 & 2, B: 1 & 2). Nine 2-locus genotypes GAABB are possible, each associated with a relative risk $\Psi_{AABB} = p(D|G_{AABB})/p(D|G_{1111})$ (D stands for disease). The effect of 0, 1 or 2 loci, and 1 or 2 disease alleles at each locus on disease status, corresponds to different patterns of Ψ's. Given Hardy Weinberg proportions and random mating, the y's (and the allele frequencies, which are assumed known) determine the expected frequency of case 2-locus genotype given parental 2-locus genotype. The application of the method of maximum likelihood on genotypic data of cases and their parents allows one to estimate the Ψ's and to infer which of the genetic models best supports the data. As inferential statistics, we used the likelihood ratio statistic and the Akaike Information Criterion (AIC) for nonnested models. The model that was significantly better than the null model (all Ψ's equal) ($p < 0.01$) and had the lowest AIC, was preferred.

Power calculation was performed by random simulation of case genotypes, conditional on parental genotype for each of the proposed models, and subsequent evaluation of the data using the inferential procedure that was outlined. When we simulate a twofold increase in risk for the 2-locus genotypes at risk and when we assume the allele frequencies to be known and equal to 0.5 for both alleles at each of the two loci, it appears that sample sizes of at least 300 cases are sufficient to discriminate between the different genetic models, when the chance of a type-2 error is controlled at 0.2.

In conclusion, we showed that with our method moderate sample sizes are adequate to discriminate

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

between the proposed genetic models. Further power analyses and the application to real data will put our method into perspective with existing methods.

REFERENCES

Schaid, D. J. & Sommer, S. S. (1993). Genotype Relative Risks: Methods for Design and Analysis of Candidate-Gene Association Studies. *Am. J. Hum. Genet.* **53**, 1114–1126.

Analysis of Association at Polymorphic Marker Loci. S. MAHDI and N. WILLIAMS. E-mail: smahdi@uwichill.edu.bb *Department of Computer Science, Mathematics and Physics, University of The West Indies, Cave Hill Campus, Barbados.*

Introduction. Although multipoint methods are generally more informative than two-point methods for testing genetic association with disease, tests for allelic association involving one marker and one trait locus are still vital for preliminary screening at lower cost. We examine the properties of a likelihood ratio test proposed by Terwilliger (1995) to detect association between a disease locus and a polymorphic marker locus. This test has been successfully used in many cases (for example Nikali *et al.* 1995).

Objectives. We aim to compare the power of Terwilliger's test using different weights to that of Pearson's chi-square test. Another goal is to improve power by reformulating the conditional probability model to allow direct estimation of the allele frequencies in the affected population. An extension to the case of multiple markers, by using haplotypes consisting of consecutive pairs of markers across a map, is also considered.

Methods. Monte Carlo studies are used to obtain empirical distributions of the likelihood ratio statistic for comparison with the proposed distribution. The bootstrap methodology is also employed in estimating p-values.

Results. The results showed show that Terwilliger's test is overly conservative. Careful inspection of the association model reveals that the null value of λ, the parameter of interest, lies on the boundary of the parameter space, so that the assumed distribution is invalid. We discuss a mixture distribution suggested in Self & Liang (1987) for use in such circumstances. It was found to closely fit the asymptotic distribution of the likelihood ratio statistic, and its use leads to type-one error rates which resemble their targeted values. The reformulated model resulted in statistical tests with power greater than 65% at 0.001, 0.01 and 0.5 significance level, for m = 2,3,5,8 and 10 alleles.

AMS 1991 Subject Classification: 62A10 62P10 62H15 62H20

keywords: association, likelihood ratio test, parameters, boundary

This work was funded by the University of the West Indies (UWI), Cave Hill Campus. We are especially grateful to the UWI School for Graduate Studies and Research for its support.

REFERENCES

Nikali, K., Suomalainen, A., Terwilliger, J., Koskinen, T., Weissenbach, J. & Peltonen, L. (1995). Random search for shared chromosomal regions in four affected individuals: The assignment of a new hereditary ataxia locus. *Am. J. Hum. Genet.* **56**, 1088–1095.

Self, S. & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators in non-standard conditons. *Journal of the American Statistical Association* **82**(398), 605–610.

Terwilliger, J. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Am. J. Hum. Genet.* **56**, 777–787.

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

An Extension of the GSMA Method to Include Candidate Region Studies C. LEWIS and L. WISE. *Guy's, King's and St. Thomas' Schools of Medicine and Dentistry, London, U.K.*

Introduction. The GSMA has been proposed as a meta-analysis method for genome wide studies (Wise *et al*. 1999). An extension to the GSMA is developed which allows for inclusion of candidate studies.

Objectives. To include candidate region studies into meta-analyses using the GSMA.

Methods. The GSMA method divides the genome into N common bins which are ranked in each study using the statistic in each bin showing maximum evidence for linkage. The GSMA test statistic for a bin is the summed rank for the bin across the studies. For each candidate study we use the family structures, marker densities and allele frequencies to carry out simulations under the null hypothesis of no linkage. Each simulation is analysed using the method used in the study and the maximum value of the linkage statistic from each simulation is recorded. These maximal statistics are used to generate an empirical density for the test statistic for each study, which is divided into N quantiles. The candidate studies are assigned ranks according to which quantile the observed test statistics are in. These ranks are added to the summed ranks from the GSMA for the appropriate bins. We consider the effect of missing information and illustrate the method with reference to genome wide and candidate studies for inflammatory bowel disease (IBD).

Results. Simulation studies indicate this is a feasible meta-analysis method for candidate regions, but that missing information leads to conservative ranks being assigned to the studies. Results of the meta-analysis of IBD studies will be presented.

Conclusion. This extension of the GSMA provides a useful meta-analysis tool allowing candidate studies to be included with meta-analyses of genome wide studies. Work is in progress to develop a method of assessing the effect that publication bias may have on the results.

REFERENCES

Wise, L. H., Lanchbury, J. S. & Lewis, C. M., (1999). Meta-analysis of genome searches. *Ann. Hum. Genet.* **63**, 263–272.

Ribosomal Proteins: Homology and Motifs. W. H. WONG[1] and J. ZHANG[2]. [1] *Department of Statistics, University of California, Los Angeles, CA 90095,* [2] *EURANDOM, LG 1.06, Den Dolech 2, 5612 AZ, Eindhoven, The Netherlands.*

Protein synthesis is mediated by ribosomes in any cell. Each ribosome is composed of RNAs and ribosomal proteins. It is commonly assumed that the function of ribosomal proteins is to stablize specific RNA structures and to promote a compact folding of the large ribosomal RNAs. To investigate how the ribosomal protein and ribosomal RNA interact, we need to obtain and analyze the RNA-binding sites in these ribosomal proteins. The importance of these patterns is reflected in the recent discovery that several human and other vertebrate genetic disorders are caused by aberrant expression of RNA-binding proteins (see Burd & Dreyfuss, 1994). Ribosomal proteins are extremely ancient molecules. Some studies have already showed that there are some strong similarities between the binding-patterns (or structures) of RNA-binding proteins and DNA-binding proteins (see Wool, Chan & Gluck, 1996; Ramakrishnan & White, 1998; Draper & Reynaldo, 1999). Several binding strategies used by DNA-binding proteins are used by ribosomal proteins. Thus ribosomal proteins may provide a window into the protein evolution.

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

In this note we focus our attentions on identifying the possible motifs for various different ribosomal protein families and predicting the potential functions of these motifs.

The key step in identifying the motifs is to form ribosomal protein families with a certain degree of diversity which can be measured by percent identities among the proteins in each family. It is known that the organisms can be divided into three phylogenetic domains: eubacteria, archaebacteria and eukaryotes. Eukaryotes include mammalian organisms and viridiplantae. We choose 13, 4 and 4 prototype ribosomal proteins from each domain, respectively, and require the percent identities in each family to be not larger than 0.90 in order to avoid redundancy.

From pairwise comparison using the BLAST and FAST (Baxevanis & Ouellette, 1998), we found almost all archaebacteria have clearly evident homologies in eukaryotic ribosomal proteins (the percentage of identities based on FAST are usually higher than 30 percent). However, only some of them are homologous to the ribosomal proteins in eubacteria. This implies that archaebacteria are closer to eukaryotes than eubacteria in terms of ribosomal proteins. There are at least 31 ribosomal protein families which have homologies accross the three phylogenetic domains. For the types of ribosomal proteins which do not exist in eubacteria, we select the ribosome proteins (with percent identities larger or equal to 0.30) from viridiplantae and add them to the corresponding families.

Current knowledge shows that RNA-binding sites are highly conserved across some phylogenetic domains and that such conserved patterns (motifs) can be identified by multiple sequence alignments. However, it is possible that some motifs may be the interaction sites between different ribosomal proteins instead of the RNA-binding sites. So it is natural to ask how to verify whether a motif is the binding domain for the ribosomal RNAs. The following are available clues for identifying a RNA-binding domain: (1) including some biochemical and structual features (e.g., basic (K, H, R), aromatic (W, Y, F), hydrophobic (A, V, I, L, C, M), and hydrophilic (D, E, S, T, H, N, Q) amino acids, structures derived from NMR spectroscopy and crystal (x-ray) methods); (2) homologous to some nucleic-acid-binding-protein; (3) the significant change of the binding affinity for the target RNA shown by mutational analysis.

Among various multiple alignment approaches, MACAW and CLUSTalW (see Baxevanis & Ouellette, 1998, Ch. 8) are most useful for carrying out our task.

We use MACAW iteratively to find more than 115 motifs for these ribosomal protein families. All these motifs are available from the authors. Some of these motifs have already been in the data base for the signatures of proteins—PROSITE. To apply CLUSTalW, we present an information content method to find the most conserved motifs from a set of aligned protein sequences. The information content method has some connection with what is called the average specific binding energy (see also Stormo & Fields, 1998). For ribosomal proteins, the resulting motifs have gaps and are often different from the corresponding motifs based on MACAW only in the boundaries and in gaps.

Are the most conserved motifs located in the RNA-binding domains?

To address this issue we check a total of 19 ribosomal protein structures, which are available in the literature. The small subunits include S1, S4, S5,S6, S7, S8, S15, S17, S19. The large subunits include L1, L4, L6, L7/12, L9, L11, L14, L22, L30, L36. It is found that the most conserved motifs based on MACAW are located in the putative RNA-binding domains. As a result, we can predict the remaining most conserved motifs (whose structures are unknown currently) are the RNA-binding

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

sites with a probability near 1. For L2, we only obtained the structure of the C-terminal. We find that the sub-conserved motif is located in that region and that the most conserved motif is located in the N-terminal. However, from the above simple argument we predict that the most conserved motif of L2 may be also located in its RNA-binding domains.

The authors are grateful to Mr. Cheng Li for his help and discussions during carrying out this project.

REFERENCES

Baxevanis, A. D. & Ouelette, B. F. F. (1998). *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins.* John Wiley, New York.

Burd, C. G. & Dreyfuss, G. (1994). Conserved structures and diversity of functions of RNA-binding proteins. *Science* **265**, 615–621.

Draper, D. E. & Reynaldo, L. P. (1999). RNA binding strategies of ribosomal proteins. *Nucleic Acids Research* **27**, 381–388.

Ramakrishnan, V. & White, S. W. (1998). Ribosomal protein structures: insights into the architecture, machinery and evolution of the ribosome. *TIBS* **23**, 208–212.

Stormo, G. D. & Fields, D. S. (1998). Specificity, free energy and information content in protein-DNA interactions. *TIBS* **23**, 109–113.

Wool, I. G., Chan, Y. L. & Gluck, A. (1996). Mammalian ribosomes: the structure and the evolution of the proteins. In *Translational Control*, pp. 685–732, Cold Spring Harbor Laboratory Press.

Poster presentations

Grade-of-Membership Sibpair Linkage Analysis Maps the *IDDM11* Locus to 14q24.3-q31: a Pattern-recognition Approach. E. HEDLUND CORDER, L. L. FIELD and M. A. WOODBURY. *Duke University, Durham, U.S.A.*

Introduction. Existing approaches to genetic linkage analysis cannot simultaneously investigate marker information for numerous loci, linked and unlinked, plus numerous clinical variables. Objectives. Use a pattern recognition approach called grade-of-membership analysis (or GoM) to implement sib-pair linkage analysis for linked markers near a previously defined locus.

Methods. The 246 families and 8 markers were those which identified the *IDDM11* locus on chromosome 14q24.3-q31 (Field *et al.* 1996). IBS information on allele sharing was used for the 578 sibpairs scored 0, 1, or 2 shared alleles for each marker locus. This information was investigated in GoM models. The model that specified 3 (vs 4, or 5) typologies best fit the data based on Akaike's information criterion. They differed in the extent of allele sharing (little, some, extensive). The frequency of doubly affected pairs increased from 42% to 54% and to 62% based on typology (i.e. extent of allele sharing) – a 48% increase in relative terms.

Conclusions. GoM may be a useful approach to sibpair linkage analysis. More complex situations need to be investigated.

REFERENCES

Manton, K. G., Woodbury, M. A., Tolley, H. D. (1994). Statistical Applications Using Fuzzy Sets, New York, John Wiley & Sons. (User Documentation for DSIGoM Version 1.0. (1999) Durham, NC, Decision Systems, Inc.)

Field, L. L., Tobias, R., Thomson, G. & Plon, S. (1996). Susceptibility to insulin-dependent diabetes mellitus maps to a locus (IDDM11) on human chromosome 14 q24.3-q31. *Genomics* **33**, 1–8.

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

Genome-wide Screen for Essential Hypertension Genes in a Deep-rooted Sardinian Pedigree. A. ANGIUS[1], G. CASU[1], P. FORABOSCO[1], G. MAESTRALE[1], P. MELIS[1], A. PALA[2], M. PALERMO[2], D. PIRAS[1] and M. PIRASTU[1]. [1] *Istituto di Genetica Molecolare CNR, Alghero, Italy,* [2] *Cattedra di Endocrinologia, Univesità degli Studi di Sassari, Sassari, Italy.*

Introduction. Genetic isolates represent ideal tools in identifying complex disease genes since affected subjects are expected to share chromosomal regions identical-by-descent (IBD) from a few founders, reducing the presence of genetic heterogeneity. In ancient, geographically isolated populations, originated from a small number of founders and with stable expansion rate, genetic diversity among individuals is expected to be reduced and the extent of linkage disequilibrium (LD) to be large, due to genetic drift. In a demographically stable area of Sardinia, whose population structure consists of several small villages, we identified an ancient isolated village, presenting high endogamy and inbreeding, where we are able to trace through archival records its genealogies for 350 years. This extended genealogy information will be used as a study strategy in the search for shared IBD genomic segments originating from common ancestors.

Background

To examine Talana's genetic structure we analysed maternal and paternal lines through the study of mtDNA and Y-chromosome markers. Sequence analysis of segment 1 of mtDNA Control Region identified 17 mitochondrial lines: 8 of them, shared by 75% of the present-day population, were already present in the village in the 17th century. Analysis with several Y-chromosome microsatellites in 56 males representing paternal lineage identified only 35 different haplotypes. We found that 8 lines, already present in the village in the 17th century, account for 75% of all Y chromosomes in the present-day male population. We also tested LD, analysing 50 maternally unrelated males with 6 markers spanning 11.5 Mb on the Xq13 chromosome. The study revealed that several haplotypes comprising at least three markers show LD that extends for an average interval of 5 Mb and that 5 ancestral X chromosomes are present in 80% of the sample. This ideal population will allow fine mapping through LD in low-density scans and smaller samples.

Methods. In order to disclose specific susceptibility genes and to test the involvement of known candidate genes for hypertension we identified 13 hypertensive subjects from a large 10-generation family encompassing more than 40 severe hypertensive patients (BP¿100 mmHg). We performed genome-wide linkage analysis with 400 microsatellites using robust two-point affected-only parametric linkage analysis with FASTLINK (Cottingham *et al.* 1993) and multipoint non-parametric IBD-sharing analysis with SIMWALK2 (Sobel & Lange, 1996), that allow computation in large and complex pedigrees. SIMWALK2 is a statistical program that uses Markov chain Monte Carlo (MCMC) and simulated annealing algorithm to analyse pedigree data of any size. The program slides an imaginary trait locus across the map and estimates several sharing statistics at each position of the map. Model based linkage analysis was performed under a dominant and a recessive mode of inheritance with disease allele frequencies set to 0.01 for the dominant and to 0.1 for the recessive model and assumed 50% penetrance.

Results. First-stage genome-wide scan:

Preliminary results reveal an excess IBD sharing yielding nominal p-values < 0.05 and LOD score > 2 on chromosomes 1, 3, 4, 12, 19 and p-values < 0.01 on chromosomes 13 and 15. These

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

regions are currently being scrutinised using a denser map searching for ancestral IBD haplotypes in the whole sample of 40 patients belonging to a large genealogy. Regions where markers flanking relevant candidate genes yielded positive LOD score, or where excess IBD-sharing was observed, are also being further investigated with densely spaced markers and additional subjects.

REFERENCES

Cottingham, R. W., Idury, R. M. & Schaffer, A. A. (1993). Faster sequential genetic linkage computations. *Am. J. Hum. Genet.* **53**, 252–263.

Sobel, E. & Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker sharing statistics. *Am. J. Hum. Genet.* **58**, 1323–1337.

Bayesian Fine Scale Mapping of Disease Loci Using the Coalescent. A. MORRIS. *University of Reading, U.K.*

Abstract. Particularly with the expected availability of high density maps of single nucleotide polymorphisms (SNPs), population association studies are likely to provide a considerable improvement in resolution over classical linkage studies for the fine scale mapping of loci contributing to complex human diseases.

Multipoint models of disease-marker association rely on the assumption that the majority of affected individuals share a common ancestor bearing a disease predisposing mutation.

The development of such models poses two particular challenges.

First, correlation within case haplotypes, since chromosomes bearing the mutation will be expected to inherit, identical by descent, the ancestral marker haplotype in the vicinity of the disease locus.

Second, correlation between case haplotypes, since chromosomes bearing the mutation may share part of their recombinational ancestry in common.

In this paper, a new multipoint model for disease-marker association is presented that accounts for correlation within and between case SNP haplotypes by using the coalescent.

The model incorporates a low rate of mutation at SNPs and allows for the possibility of multiple disease mutations. We employ Markov chain Monte Carlo (MCMC) methods in a Bayesian framework to sample over the distribution of possible coalescent trees for the ancestry of the case chromosomes.

In this way, we obtain posterior distributions for the model parameters including the location of the mutation, the time to the most recent common ancestor of the case chromosomes and the ancestral marker haplotype.

The method is illustrated using real data for which we obtain reliable estimates of the ΔF508 mutation for cystic fibrosis in a candidate region of chromosome 7q31.

Variance-components approach to linkage analysis of livestock pedigree data. T. I. AXENOVICH, Y. S. and AULCHENKO, G. R. SVISCHEVA. *Institute of Cytology and Genetics, Novosibirsk, Russia.*

The extension of variance-components method (Amos, 1994; Almasy & Blangero, 1998) for mapping quantitative trait loci (QTLs) is proposed. The new method permits analysis of the pedigree data coming from crosses between genetically heterogeneous parental breeds (populations). The variance-components approach assumes that the likelihood of pedigree data is proportional to the

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press

density of multinomial normal distribution determined by the vector of the means and the variance-covariance matrix. Modern realisations of the method are aimed to analyse human pedigree data. These realisations can not be applied directly to the analysis of livestock pedigree data because the founders of livestock pedigrees may come from different breeds. Therefore modifications of the vector of the means and variance-covariance matrix are required. Recently we described the modification of mixed model for segregation analysis of pedigrees coming from interbreed crosses (Axenovich, 1999). The key assumptions, which adapt traditional segregation models to analysis of such pedigrees, are as follows: rule of genes' inheritance and the penetrance function is the same for all pedigree members. These assumptions together with the ordinary assumptions of variance-components method allowed us to construct the vector of means and the variance-covariance matrix for a pedigree coming from interbreed crosses. The elements of this matrix are described via genetic parameters introduced for parental populations and several additional parameters. The variance-components method is expected to be a powerful approach to mapping the quantitative trait loci of livestock because the method utilises the genetic information from pedigrees of arbitrary structure and not only from the fixed crosses.

REFERENCES

Almasy, L. & Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *Am. J. Hum. Genet.* **62**, 1198–1211.

Amos, C. I. (1994). Robust variance-components approach for assessing genetic linkage in pedigrees. *Am. J. Hum. Genet.* **54**, 535–543.

Axenovich, T. I. (1999). Inheritance of Quantitative Traits in Hybrid Pedigrees: Mixed Models. *Rus. J. Genet.* **35**, 444–452.

https://doi.org/10.1017/S000348000000823X Published online by Cambridge University Press