Mathematisches Forschungsinstitut Oberwolfach

Report No. 7/2003

Medical Statistics - Current Developments in Statistical Methodology for Genetic Architecture of Complex Diseases

February 2nd – February 8th, 2003

The meeting was organised by H. Schäfer (Marburg), C.Amos (Houston), and M.P. Baur (Bonn). During the 5 days of the conference, 29 talks and a tutorial on Variance-component Methods were given, 43 scientists from Germany (# 20), USA (# 16), Great Britain (# 3) France (# 3), and Canada (# 1) participated. The conference was dedicated to current statistical developments in the field of statistical genetics and genetic epidemiology. After the impressing advances in uncovering the etiology of monogenic diseases in the last decades, the grand challenge to genetic epidemiology is now the identification of genes for complex diseases. The rapid developments in molecular genetics, such as the detection of highly polymorphic microsatellite markers and of up to 1.500.000 SNPs in the human genome, has been paralleled by great advances in the statistical methodology. The conference gave an impressive overview of current statistical developments in this field. The fruitful scientif! ic interaction during the conference will certainly stimulate further cooperation and research.

The organisers and participants thank the "Mathematisches Forschungsinstitut Oberwolfach" to make the conference possible in the usual comfortable and inspiring setting.

Abstracts

Population Genomics: A Paradigm for Complex Disease Studies in the Post-Genome Era

Ranajit Chakraborty and Kosuke Teshima

Even though complex diseases constitute the major public health burden in all societies around the world, success in determining etiology of such diseases has been rather limited for several reasons. This presentation starts with a brief outline of possible reasons of the difficulties involved in elucidating the genetic basis of complex diseases. From these discussions, we argue that population-based association studies are more likely to provide insights of genetic basis of complex diseases, rather than traditional family-based study designs. However, since disease-gene association at population level stems from inter-locus association of alleles, a thorough understanding of population genetic properties of linkage disequilibrium (LD) is needed for appropriate genetic interpretation of disease-gene association data. To this effect, some properties of genome-wide background LD are examined through a coalescence-based simulation study. We show that when microsatellite loci ar! e used as genomic markers for disease-gene association studies, the expectation of the weighted normalized LD between two loci decreases with recombination distance between loci. However, the extent and trend of such decay is dependent on the rate and pattern of mutations as well as on the demographic history of populations. For example, for any specified recombination distance, the simulation results show that the power of detection of LD is larger in populations of constant smaller size. In a growing population, the power of detecting LD is substantially reduced, making it comparable to that expected in a constant population of the largest size reached by the population. In presence of population growth, the enhancement of LD detection power with increasing sample size is less conspicuous than in populations of constant size. Power of detection of LD is also larger for loci with higher mutation rate in populations of constant size, although under population growth, the eff! ect of mutation rate is reversed, particularly for markers of l! arger recombination distances. Multistep forward-backward mutations at microsatellite loci actually increase the power. Finally, presence of multiple alleles at microsatellite loci makes such markers more powerful to detect LD, than the common single nucleotide polymorphism sites (SNPs) residing at the same recombination distance. The relevance of such population genomic paradigm for complex traits is also illustrated with examples, where studying them individually may not readily recognize the underlying mutations affecting complex traits, but when studied in the context of haplotypes, their effects become statistically significant. (Research supported by US Public Health Research Grants from the National Institutes of Health).

Mapping Genes With the Use of a Local Approcimation to the Ancestral Recombination Graph

Sebastian Zoellner and Jonathan Pritchard

Fine mapping with Linkage Disequilibrium (LD) is the method of choice to locate disease-causing mutations by using LD to detect haplotypes that share a common ancestry among cases. But this method relies on a sufficiently high level of association between a marker and a disease phenotype. However, for complex disease genes this association does not always exist. Therefore, a more powerful approach is required. Here we propose a novel method to infer the position of a disease mutation by inferring the ancestry of a genomic region from marker data consisting of a sample of diploid cases and controls with known phase. For a given location on the marker-map, we use the marker information of both cases and controls to reconstruct local approximations of the full recombination graph using Markov Chain Monte Carlo. We sample from the space of trees that are most likely, given the marker data. On each tree, the likelihood that the given location harbours the disease mutation is estimated, based on the distribution of cases and controls among the tips. By integrating over the space of trees, we estimate the likelihood of a disease mutation at the given location. Repeating this process for each possible position of the disease mutation allows the estimation of a likelihood curve in the genomic region under consideration. From this, we can determine the most likely location and construct a corresponding confidence interval. This method also allows us to estimate the penetrances of the disease mutation(s) for any given location of the disease mutation. Furthermore, this approach can be extended to analyze quantitative traits.

Bayesian modelling of complex metabolic pathways

Duncan C. Thomas

Much of molecular epidemiology is concerned with studying pathways that may involve multiple genes and/or multiple environmental exposures. For example, well done red meat (E_1) and tobacco smoking (E_2) are known sources of heterocyclic amines (HCAs) and polycyclic aromatic hydrocarbons (HCAs), which are converted to active carcinogens by a number of activating enzymes and detoxified by other enzymes, each regulated by specific genes (for example, *CYP1A2, NAT1, NAT2* for HCAs, *CYP1A1, mEH, GSTM3* for PAHs ($G_1 - G_6$), amongst others). These various factors have been implicated in epidemiological studies of colorectal polyps (Y) or colorectal cancer, for which polyps are a known precursor lesion.

Epidemiological data are commonly presented in terms of contingency tables, considering one or two exposures or genes at a time, singly or in pairwise combinations, thereby neglecting the potentially confounding effects of other factors or more complex interactions. While multiple logistic regression offers an approach to developing more sophisticated models, uncertainty remains about the choice of model when interpreting the effects included in any particular model. For example, in fitting these factors to data on 466 polyps cases and 509 controls drawn from a sigmoidoscopy clinic (Haile et al., 1997), we found significant main effects only for smoking, but a number of gene-environment ($G \times E$), $E \times E$, and $G \times G$ interactions, as well as some 3- and 4-way interactions, but the strength of these interactions varied depending upon what other terms were in the model. We suggest two alternative approaches to this problem: Bayes model averaging (George and Fost! er, 2000; Raftery et al., 1997); and pharmacokinetic modelling (Gelman et al., 1996).

In the former approach, we postulate a logistic model of the form

$$\operatorname{logit} \Pr(Y = 1|\mathbf{E}, \mathbf{G}) = \beta_0 + \boldsymbol{\beta}_1'\mathbf{E} + \boldsymbol{\beta}_2'\mathbf{G} + \boldsymbol{\beta}_3'\mathbf{E} \otimes \mathbf{E} + \boldsymbol{\beta}_4'\mathbf{G} \otimes \mathbf{E} + \boldsymbol{\beta}_5'\mathbf{G} \otimes \mathbf{G} + ...$$

together with a prior distribution for $\boldsymbol{\beta}$ as a mixture

$$\Pr(\boldsymbol{\beta}) = \sum_{m=1}^{M} \pi_m \prod_{p=1}^{P} \left[(1 - W_{mp})\delta(0) + W_{mp}N(0, \tau_p^2)\right]$$

where π_m denotes the prior probability for model m and W_{mp} an indicator variable for whether effect p is in model m. Following Chipman (1996), we restrict attention to the class $\mathcal{H}$ of hierarchical models, such that an interaction effect can have $W = 1$ only if all of its main effects and lower-order interactions are also included. Hence we take

$$\Pr(\mathbf{W}_m) \propto \prod_p I(\mathbf{W}_m \in \mathcal{H}) \sum_p \text{Dirichlet}(\mathbf{1})$$

Fitting this model to the polyps data, only the smoking effect remained significant in the marginal model, while most of the interaction effects disappeared.

In the pharmacokinetic modelling approach, we include the intermediate metabolites as unobserved (latent) variables Z_{im} for metabolite m, which are related to each other by a series of person-specific activation rates λ_{im} and detoxification rates μ_{im}, which in turn are determined by the individual's genotypes. Specifically we used first-order linear kinetic models, which yield the steady state solution

$$Z_{i,m+1} = Z_{im} \frac{\lambda_{im}}{\lambda_{i,m+1} + \mu_{im}}$$

and assumed that $\log \lambda_{im} \sim N(\bar{\lambda}_{mG_{im}}, \sigma_m^2)$. This model can be fitted using Markov chain Monte Carlo methods, as described by Cortessis et al. (2002). Simulation studies of the statistical properties of the approach are underway.

References

1. Chipman H. Bayesian variable selection with related predictors. *Can J Statist* 1996; **24**:17-36.
2. Cortessis V, Thomas DC. Toxicokinetic genetics: An approach to gene-environment and gene-gene interactions in complex metabolic pathways. In: Bird P, Boffetta P, Buffler Pet al (Ed.),Mechanistic considerations in the molecular epidemiology of cancer. Lyon, France, IARC Scientific Publications. 2002; : in press.
3. Gelman A, Bois F, Jiang J. Physiological pharmacokinetic analysis using population modelling and informative prior distributions. *J Am Statist Assoc* 1996; **91**:1400-1412.
4. George EI, Foster DP. Calibration and empirical Bayes variable selection. *Biometric* 2000; **87**:731-747.
5. Haile RW, Witte JS, Longnecker MP, et al. A sigmoidoscopy-based case-control study of polyps: macronutrients, fibre, and meat consumption. *Int J Cancer* 1997; **73**:497-502.
6. Raftery AE, Madigan D, Hoeting JA. Bayesian model averaging for linear regression models. *J Am Statist Assoc* 1997; **92**:179-191.

Model-based methods for the genetic analysis of complex diseases: what can they bring?

Florence Demenais

Identification of genes underlying complex diseases is a challenging task which can be conducted by two main analytical approaches: 1) model-free methods that do not assume a genetic model for the disease (or for the phenotype of interest); 2) model-based methods that specify a genetic model. Models-free methods are the methods of choice to characterize the chromosomal regions that may harbour disease-predisposing genes by genome-wide linkage analyses and, then, to identify these genes by association studies. Model-based methods can be used advantageously to confirm linkage results obtained by model-free methods but also permit, in a given region, not only to identify the genetic variants associated with the disease (or any intermediate quantitative phenotype) but also to model the effects of these variants and to search for gene-gene and gene-environment interactions.

The regressive models that include the effect of the gene under investigation (possibly linked and eventua! lly in linkage disequilibrium (LD) with a marker or a set of markers), other sources of familial correlations (due to other genes and/or shared environmental factors), and covariates, are particularly suitable to achieve these goals [1, 2, 3]. The performances of combined segregation-linkage analysis based on these models were presented in the two following situations: 1) detection of potential causal variants influencing the variability of a quantitative phenotype; 2) assessment of the power of this approach to detect gene-environment (GxE) interactions underlying the liability to a disease.

We showed that combined segregation-linkage analysis, based on the class D regressive model, was able to detect significant effects of four of five polymorphisms of the AGT (Angiotensinogen) gene influencing plasma AGT levels. The most significant result was found at C-532T polymorphism (P = 0.000001), which accounted for 4.3

Power of the class D regressive-threshold model [5,6] to detect GxE was investigated by simulations in nuclear families [7]. The generated model assumed a liability to disease depending on a common gene with a small effect, a polygenic component and an environmental factor interacting with the gene; parameter values of these different components were varied. The disease gene was assumed tightly linked to a diallelic marker (SNP) with varying strength of LD between the 2 loci (D' being equal to 0, 0.5 or 1). Power of the regressive threshold model to detect GxE was estimated by the proportion of 100 replicates of 165 nuclear families (with varying sibship size) to reject the null hypothesis of no interaction versus GxE while estimating all parameters of the model. This power was between 65-82when D' = 1, 25-28factor and the nature of this factor (binary or quantitative). Moreover, when comparing two alternative formulations of the regressive models, the power to detect GXE appeared higher when analyzing the data with the regressive threshold model [5] than using the regressive logistic model [2]. Alternatively, ignoring the presence of GxE in the analysis may affect the detection of the true LD model and, therefore, the identification of the putative functional variant. Evidence for complete LD was reduced by about 60-70% when ignoring GxE as compared to taking it into account and this impact was slightly greater for a dominant gene than for a recessive one and for a quantitative environmental factor than for a binary factor. Thus, use of models that can take into account both LD and GxE may be of major importance to disentangle the mechanisms underlying complex diseases. This was illustrated by combined segregation analysis of cutaneous malignant melanoma and CDKN2A gene (a known melanoma predisposing gene) in 53 French melanoma-prone families. Cumulative risk of melanoma depended on presence of mutations of this gene and this risk was modified by pigmentary phenotypes (high number of nevi, presence of atypical nevi) and reactions to sun exposure (propensity to sun! burn) [8]. These latter results can have important consequences in risk assessment.

References

1. Bonney GE. On the statistical determination of major gene mechanisms in continuous human traits: regressive models. Am J Med Genet, 1984, 18: 731-749
2. Bonney GE. Regressive logistic model for familial disease and other binary traits. Biometrics, 1986, 42:611-625.
3. Bonney GE, Lathrop GM, Lalouel JM. Combined linkage and segregation analysis using regressive models. Am J Hum Genet, 1988, 43, 29-37
4. Brand E, Chatelain N, Paillard F, Tiret L, Visvikis S, Lathrop M, Soubrier F, Demenais F. Detection of putative functional angiotensinogen (AGT) gene variants controlling plasma AGT levels by combined segregation-linkage analysis. Eur J Hum Genet, 2002,

10:715-723.
5. Demenais FM. Regressive logistic models for familial diseases: a formulation assuming an underlying liability model . Am J Hum Genet, 1991, 49: 773-785.
6. Briollais L, Demenais F. Regressive threshold model for the familial analysis of complex diseases with variable age of onset. Genet Epidemiol, 2002, 23:375-397
7. Chaudru V, Rosenberg M, Demenais F. Gene-environment interactions and linkage disequilibrium in familial analysis of complex diseases. Am J Hum Genet, 2001, 69(4):407 (abstract)
8. Chaudru V, Chompret A, Avril MF, Bressac-De Paillerets B, The French Hereditary Melanoma Study Group and Demenais F. Effects of CDKN2A gene, naevus phenotypes and sun-related covariates in French melanoma-prone families. Melanoma Research, 2001, 11 (suppl 1): S143(abstract).

Detection of microdeletions using SNPs and other genetic markers

Chris Amos

Microdeletions have been found to cause a variety of human diseases and conditions such as Prader-Willi Syndrome, diGeorge syndrome and autism. Syndromes that sometimes present with microdeletions such as Wilms Tumor due to deletion of the WT1 gene and nearby loci (WAGR association syndrome) are virtually all de novo and show little familiality although they are genetic. In 2002, microdeletions were commonly found in regions of the genome supporting linkage to familial autism and these are all stably inherited, with a normal parent transmitting a microdeletion to an affected child. The increasing precision provided by characterizing families with much denser maps of genetic markers suggests increased ability to identify microdeletions. In this talk, I developed statistical approaches for specifying the likelihood of family data to model normal transmission of alleles, de novo microdeletions and stably inherited microdeletions. Complications arise in trying to model transitions from normal to microdeleted regions and I gave one approach assuming a first-order Markov process. I also discussed the complexity of modelling family data when there are associations among the alleles at closely linked loci. I briefly described an approach for error identification in families. I also briefly discussed potential development of methods to scan for microdeletions using cases with a disease or cases and controls.

Scoring method for linkage from Bayesian MCMC oligogenic combined segregation analysis

E.Warwick Daw

In the dissection of complex genetic traits, where exact parametric models become computationally intractable, Bayesian Monte Carlo Markov chain (MCMC) techniques have shown promise. The methods implemented by Heath (1997, American Journal of Human Genetics 61:748760), in the program Loki, have been able to localize genes for complex traits in both real and simulated data sets. Loki uses an iterative MCMC process to estimate the posterior probability over a model space that includes quantitative trait loci (QTL) locations on a chromosome. We consider marginal posterior probability not only over location, but also over variance attributed to a QTL, producing a 2-dimensional probability surface. This 2-dimensional surface allows us to estimate the relative contribution, which is important in an oligogenic analysis. Unfortunately, interpretation of the results and assessment of their significance has been difficult. We have introduced score, the Log

Of the Posterior placement ! probability ratio (LOP), for assessing oligogenic QTL detection and localization (Daw et al., 2003 Genetic Epidemiology 24:181-190). The LOP is the log of the posterior probability of linkage to the real chromosome divided by the posterior probability of linkage to an unlinked pseudo chromosome with marker informativeness similar to the marker data on the real chromosomes. The pseudo chromosome is created by random gene drop using the observed pedigree structures, allele frequencies, and missing data patterns. Since the LOP cannot be calculated exactly, we estimate it in simultaneous analysis of both real and pseudo chromosomes. While lod scores are calculated under a single linkage model, the LOP is calculated with Monte Carlo integration over a large number of model parameters, including the number of trait loci and the additive and dominance effects at each locus. We have examined several ways to estimate LOP: by counting QTL placements on the real and pseudo chromosomes,! by calculating the relative probabilities on the real and pseu! do chromosomes at the location a QTL is proposed, and by calculating the relative probabilities on the real and pseudo chromosomes at fixed locations along the chromosome. We have begun to investigate the distributional properties of each of these estimates of LOP in the presence and absence of trait genes. We have demonstrated how to obtain an empirical p-value for chromosome-wide linkage with LOP.

Use of Decision and Regression Trees in Genetic Epidemiology

Michael Krawczak

One of the major aims of genetic epidemiology is to partition, on the basis of environmental or genotypic covariates, a data set comprising qualitative or quantitative trait values. Decision and regression trees represent a powerful tool to achieve this goal. In genetic epidemiology, however, this type of exploratory data mining methodology has not hitherto attained the attention it deserves. Possible applications are manifold, including the use of decision trees for characterizing gene-disease associations. Here, the major advantages of the technique are that it

(1) is genotype-based and therefore avoids prior hypotheses about the genotype-phenotype relationships involved,
(2) can include a virtually unlimited number of genetic markers,
(3) allows for complex gene-gene and gene-environment interactions,
(4) requires no haplotype reconstruction,
(5) entails only a moderate loss of information in comparison to haplotype-genotype-based inference tools.

The practical utility of decision and regression trees in a genetics context is highlighted by an application to 23 SNP from the human CARD15 gene region in relation to Crohn Disease (CD). The three SNPs known to be associated with CD together with an additional SNP from the 5UTR constitute the final, pruned tree.

In another study of the in vitro expression profile of 15 SNPs from the human growth hormone (GH1) region, 6 SNP defining 11 haplotypes are shown to capture the bulk of phenotypic variation observed. These SNP exerted there increasing/decreasing influence upon expression

Sum statistics for capturing joint effects of multiple disease loci

Jurg Ott

When multiple susceptibility loci are jointly responsible for a disease, it must be inefficient to carry out association (or linkage) studies for one marker at a time. We previously addressed the problem of multiple disease loci by ordering SNP markers according to their individual statistic for association and then building sums containing varying numbers of SNPs with the highest statistics. The smallest p-value associated with any of the largest n sums was then taken as our overall statistic with associated study-wise significance level ("Set Association" analysis, Hoh et al., Genome Res 11:2115, 2001).

A drawback of this approach is that it relies on marker-specific statistics (main effects). We propose an extension of our method that includes pair-wise interactions between SNPs. Theoretical calculations on the basis of a purely epistatic 3-locus disease model (Culverhouse et al., Am J Hum Genet 70:461, 2002) demonstrate that the penalty for multiple testing of all possible pairs of up to 50,000 SNPs still leaves expected results of association tests highly significant.

Computational approaches to detecting gene-gene interactions

Jason H. Moore

One goal of genetic epidemiology is to identify polymorphisms associated with common, complex multifactorial diseases. Success in achieving this goal will depend on a research strategy that recognizes and addresses the importance of interactions among multiple genetic and environmental factors in the etiology of common diseases. One traditional approach to modelling the relationship between discrete predictors such as genotypes and discrete clinical outcomes is logistic regression. However, logistic regression is limited in its ability to deal with interaction data involving many simultaneous factors because of the curse of dimensionality. In response to this limitation, we developed the multifactor dimensionality reduction (MDR) approach that seeks to reduce the dimensionality of multilocus genotype space to facilitate the identification of gene-gene interactions. This approach is nonparametric and genetic-model free and is directly applicable to the analysis of case-c! ontrol and discordant sib-pair study designs. Further, an MDR software package is available. Empirical studies with both simulated and real data have indicated MDR has good power for identifying high-order gene-gene interactions. We anticipate that MDR will be a useful addition to the repertoire of new approaches for the detection and characterization of gene-gene and gene-environment interactions.

Detection and modelling of heterogeneity in allele sharing

Shelley B.Bull

Linkage studies of complex disease may be designed to find chromosomal regions that harbour susceptibility genes by genome-wide search or to assess a particular gene with known location by a candidate gene approach. Common designs involve recruitment of families with affected sibling pairs and/or other affected relatives in whom a set of genetic markers are typed across the genome or in a region. In complex disease, the ability to detect linkage is compromised by heterogeneity, due for example to having a mixture of families that are linked or unlinked to a single disease gene locus. Modelling of variation in identical-by-descent allele sharing can be useful to increase power to detect linkage, identify covariate-defined subgroups that are linked to particular marker regions, identify characteristics for phenotype refinement, and improve the design of subsequent studies to localize genes and characterize their effects in combination with other genes and/or environmental fact! ors.

To this end, we extended the linear and exponential linkage likelihoods of Kong and Cox (Am J Hum Genet 1997) for affected relatives to incorporate a binary covariate that allows differences in NPL (non-parametric linkage) scores between two groups of families. We compared the performance of a likelihood ratio (LR) test statistic for group differences to that of a simple two-sample t-statistic for mean NPL differences. In simulation studies of families with affected siblings or affected cousins exhibiting locus heterogeneity, we found that, under the null hypothesis of linkage without heterogeneity, the distribution of the LR test statistic depends on the extent of linkage, particularly so in the linear model due to constraints on the parameters. On the other hand, the distribution of the t-statistic may be biased by differences between groups in information content. Thus, although these approaches are useful in several respects, the interpretation of covariate effects in al! lele-sharing models requires caution.

Methods and algorithms for linkage analysis of genetically complex traits: extensions for imprinting and two-locus models

Konstantin Strauch Johannes Dietter

Genetically complex traits are often determined by more than one gene. In addition, there may be non-canonical inheritance even in the context of a single gene, such as overdominance or genomic imprinting. In order to adequately model imprinting in the context of parametric single-trait-locus linkage analysis, individuals heterozygous at the disease locus need to be distinguished by the parent who transmitted the mutation. Therefore, an adequate imprinting model contains two heterozygote penetrances instead of only one, that is, four penetrance parameters altogether. Parametric linkage analysis with four-penetrance imprinting models has been implemented into the program GENEHUNTER-IMPRINTING. Since, for parametric linkage analysis, the power to detect linkage decreases if the trait model is misspecified, it can be useful to maximize the LOD score with respect to the disease model parameters, i.e., penetrances and disease allele frequency. Therefore, GENEHUNTER-IMPRINTING has recently been extended to perform such a maximization procedure, which is called 'MOD-score analysis' or 'maximizing the maximum LOD score' (MMLS).

Another program extension, GENEHUNTER-TWOLOCUS, allows to explicitly model two trait loci with parametric and nonparametric linkage analysis in the multi-marker context. For traits which are in fact determined by two loci, the power of two-trait-locus analysis has proven to be higher than with single-trait-locus methods, provided that at least one marker linked to each disease locus is included into the analysis. Due to the nature of the Lander-Green algorithm, for which the computation time and memory requirements increase exponentially with the number of individuals in a pedigree, the pedigrees to be analyzed with the original version of GENEHUNTER-TWOLOCUS were restricted to 12-13 bits. Recently, the two-trait-locus algorithm of the program has been thoroughly optimized; this increased the speed by a factor of ten. In addition, the time-intensive part has been parallelized. The obtained speed-up is perfect, and so the computation time further decreases by a factor which is equal to the number of processors used. Altogether, with the optimized and parallelized version of GENEHUNTER-TWOLOCUS, the size of pedigrees which can be analyzed increases from 12-13 to 17 bits and more. In addition, the new version of GENEHUNTER-TWOLOCUS calculates LOD and NPL scores as a function of both trait-locus positions, not just of one trait-locus position as before, with the position of the other disease locus being fixed. The LOD or NPL results are conveniently displayed in a three-dimensional plot, and can be viewed e.g. with GNUPLOT, or any other suitable graphics package.

Bayesian spatial modelling of haplotype associations

Duncan C. Thomas

Multilocus genotypes of tightly linked loci can be a powerful tool for mapping a disease gene by linkage disequilibrium or for characterizing the effect of a candidate gene. To fully exploit the information from multilocus genotypes $\mathbf{G}_i tr = (G_{i\ell})_{\ell=1,\dots L}$, where ℓ indexes the loci and i the subjects, one must first arrange the genotypes into two haplotypes $\mathbf{H}_i = (h_{i1}, h_{i2})$, the sequence of alleles on each chromosome. These haplotypes can usually not be determined with certainty from the observed genotypes. Hence, we use a likelihood of the form

$$\mathcal{L}(\boldsymbol{\beta}, \mathbf{q}) = \sum_{\mathbf{h}\sim\mathbf{G}_i} P_{\boldsymbol{\beta}}(Y_i|\mathbf{H}_i = \mathbf{h})\ P_{\mathbf{q}}(\mathbf{H}_i = \mathbf{h})$$

where Y_i is a subject's phenotype, $\mathbf{h} \sim \mathbf{G}$ indicates the set of haplotype pairs that are compatible with the observed genotypes, and $(\boldsymbol{\beta}, \mathbf{q})$ are parameters for haplotype relative risks and haplotype population frequencies, respectively. For example, we might adopt a logistic model of the form

$$\operatorname{logit}\Pr(Y_i = 1|\mathbf{H}_i) = \beta_0 + \beta_{h_{i1}} + \beta_{h_{i2}}$$

The model can be fitted using E-M methods (Chiano and Clayton, 1998; Excoffier and Slatkin, 1995; Schaid et al., 2002; Stram et al., 2002), provided the number of loci (or haplotypes) is not too large; otherwise Markov chain Monte Carlo (MCMC) methods can be used (Liu et al., 2001; Niu et al., 2002).

However, in the case of many haplotypes, we are quickly confronted with the interrelated problems of multiple comparisons and sparse data, for which Bayesian shrinkage estimators appear to be a natural solution. In doing so, we wish to exploit the notion that structurally similar haplotypes in the neighbourhood of a disease predisposing locus are more likely to harbor the same susceptibility allele and hence to have similar βs. Thomas et al. (2001) and Molitor et al. (Molitor et al., 2002) considered a conditional autoregressive (CAR) model of the form

$$\boldsymbol{\beta} \sim MVN(\mathbf{0}, \sigma^2\mathbf{I} + \tau^2\mathbf{W})$$

where $\mathbf{W}$ is a matrix of "similarities" of each pair (h, k) of haplotypes, such as the length $L_{hk}(x)$ of the segment shared *ibs* surrounding a candidate mutation location x.

More recently, we have been considering the Potts (Green and Richardson, 2001) and Voronoi (1908) spatial clustering models, of the form

$$\operatorname{logit}\Pr(Y_i = 1|\mathbf{H}_i) = \beta_0 + \beta_{c_{h_{i1}}} + \beta_{c_{h_{i2}}}$$

where c_h denotes the "cluster" to which haplotype h belongs. For the Potts model,

$$\Pr(c_h = c|\mathbf{c}_{-h}) = \frac{\exp\left[\psi \sum_{h\sim k} I(c_k = c)\right]}{\sum_{c'} \exp\left[\psi \sum_{h\sim k} I(c_k = c')\right]}$$

where $h \sim k$ means h and k are neighbours (e.g., they differ at only a single SNP). In the Voronoi model, haplotypes are assigned deterministically the the cluster containing the "nearest" ancestral haplotype A_c. MCMC methods are used to update $\boldsymbol{\beta}$, x, σ^2, τ^2 (in the CAR models), $\mathbf{c}$, ψ (for the Potts model), and A_c (for the Voronoi model). Reversible jump MCMC could be used to update the number of clusters.

These spatial clustering models implicitly assume a "star-shaped" genealogy, i.e., the present day haplotypes in each cluster are assumed to be independently derived from the corresponding ancestral haplotype. We are currently trying to relate this approach to coalescent models, by allowing each haplotype to have its own β_h, but to allow the covariance of pairs of haplotypes in the same cluster to depend upon their estimated time T_{hk} to a common ancestor, based on their shared length L_{hk}. Letting $\mathrm{cov}(\beta_h, \beta_k|T_{hk}) = \sigma_c^2 \exp(-\varphi T_{hk})$, and using the fact that $L_{hk} \sim \Gamma(2, 2T_{hk})$ and $T_{hk}\ \Gamma(1, 1/(2N))$, where N is the effective population size, the marginal covariance can be shown to be

$$\mathrm{cov}(\beta_n, \beta_k|c_h{=}c_k) = \left(\frac{4\varphi\sigma_c^2}{N}\right)\left(\frac{L_{hk}}{\phi + 1/(2N) + 2\text{Ł}_{hk}}\right)$$

We are also exploring alternative approaches based on hierarchical models for inferences on the genotype level, without having to infer haplotypes (Conti and Witte, 2003). In this approach, the first level model is expressed in terms of relative risks coefficients for the genotypes at each locus separately, and the second level model defines the prior means and covariances of these coefficients.

References

1. Chiano M, Clayton D. Fine genetic mapping using haplotype analysis and the missing data problem. *Ann Hum Genet* 1998; **62**:55-60.

2. Conti DV, Witte JS. Hierarchical modelling of linkage disequilibrum: genetic structure and spatial relations. *Am J Hum Genet* 2003; **72**:351-63.

3. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 1995; **12**:921-7. Green P, Richardson S (2001). Hidden Markov models and disease mapping.

4. Liu JS, Sabatti C, Teng J, et al. Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res* 2001; **11**:1716-24.

5. Molitor J, Marjoram P, Thomas D. Application of Bayesian spatial statistical methods to the analysis of haplotype effects and gene mapping. *Genetic Epidemiology* 2002; :under review.

6. Niu T, Qin ZS, Xu X, et al. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet* 2002; **70**:157-169.

7. Schaid DJ, Rowland CM, Tines DE, et al. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics* 2002; **70**:425-34.

8. Stram DO, Pearce L, Bretsky P, et al. (2002). Modelling and E-M estimation of haplotype-specific relative risks from genotype data for a case-control study of unrelated individuals.

9. Thomas DC, Morrison J, Clayton DG. Bayes estimates of haplotype effects. *Genet Epidemiol* 2001; **21** (Suppl 1):S712-S717.

10. Voronoi. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. *J. Reine Angew. Math.* 1908; **34**:198-287.

Statistical Issues in Haplotype Analysis

Hongyu Zhao

Haplotype analyses may provide more information on the complex relation between DNA variation and phenotypes than single marker analyses. However, there are many statistical issues that need to be resolved and understood to fully take advantage of haplotype data. In this talk, we cover a few topics that are of great interest to human geneticists.

The first issue is that the identification and analysis of haplotypes require knowledge of phase information on the sampled individuals, which is not generally available from unrelated individuals. We describe a likelihood-based approach to assessing associations between traits and haplotypes based on haplotype-based logistic regression models. In this model, the outcome is case or control, and the predictor variables include the number of copies (0,1,or 2) of each haplotype, as well as other explanatory variables. We consider the underlying unobserved haplotype pairs and phase information as missing data and use the method of weights to estimate model parameters. Our methods provide both global test and haplotype-specific tests. In addition, one major advantage of our method is that we are able to estimate specific haplotype effects with respect to the baseline haplotype. We show the results of our method applied to study associations between p53 and breast cancer and betwe! en the human opioid receptor gene and substance dependence using real data sets. Our methods can be directly applied to cohort studies and extended to analyze quantitative traits and other types of data by constructing corresponding generalized linear models.

The second issue is SNP marker selection. Although millions of genetic polymorphisms have been identified in the human genome, a small proportion of these markers may be needed to capture the majority of the diversity due to linkage disequilibrium. Several methods have been proposed to select representative markers. One use of these representative markers is to select a set of markers to detect an association of disease with haplotype (disease gene mapping). We investigate the usefulness of three methods designed to select markers that preserve information or diversity for identifying disease associations with haplotypes in case-control studies. We examine five genes with known disease associations. The three procedures designed to preserve information or diversity often lead to the selection of markers with poor power to detect the disease association. An alternative strategy, such as a two-stage design that uses data on cases and controls in the initial stage, may se! lect more powerful markers for disease gene mapping in the second stage.

The third issue addressed in this talk is measurement errors in genotypings. In general, two different genotyping strategies may be employed to establish associations between genotypes and phenotypes: (1) collecting individual genotypes or (2) quantifying allele frequencies in DNA pools. These two technologies have their respective advantages. Individual genotyping gathers more information, whereas DNA pooling may be more cost effective. Recent technological advances in DNA pooling have generated great interest in using DNA pooling in association studies. We investigate the impacts of measurement errors in genotyping on the identification of genetic associations with these two genotyping strategies. We find that, with current technologies, a larger sample is generally required to achieve the same power using DNA pooling compared to individual genotyping. We further consider the use of DNA pooling as a screening tool to identify candidate regions for follow up studies. We fi! nd that a majority of the positive regions identified from DNA pooling results may represent false positives if measurement errors are not appropriately considered in the design of the study.

Haplotype estimation for multilocus SNP genotype data and their association to genetic traits

Klaus Rhode

With exception of direct haplotyping the majority of genetic data are large samples of multilocus biallelic (SNP) genotypes with the phase of the alleles at each locus not known. The problem is to estimate the underlying haplotype pairs behind the phase-unknown genotypes in order to find some association between a certain haplotype and a genetic trait. We use a three-stage approach:
- EM-algorithm (extended to nuclear families) for estimating haplotype frequencies and all compatible haplotype pair configurations of the sample.
- Conventional statistics ANOVA (parents or individuals) or TDT (nuclear families) at the most likely haplotype pair configuration of the sample.
- Sampling the statistics over all compatible haplotype pair configurations with their weight in order to take also the less likely haplotype pair configurations into account.

Applications for simulated and real data (ACE polymorphisms in a Caucasian and Nigerian sample) were shown.

Detecting association between disease and a group of “haplotype tagging” single nucleotide polymorphisms

David Clayton

Recently it has been suggested that association between disease and a small genetic region may be most efficiently detected by typing a few “haplotype tagging” SNP’s (htSNP’s). Typically these will have been selected from a larger set of markers in order to most closely identify the most frequent haplotypes in the population. Less attention has been given to the problem of how such studies should be analysed when completed and how the initial data which was used to select the htSNPs should be incorporated into the analysis. This paper discussed this problem for both population– and family–based association studies. The role of the R^2 measure of association between a causal locus and various methods of scoring of marker haplotypes is highlighted. Lack of knowledge of gametic phase was shown to result in little loss of power, and a new “multi-locus TDT” was proposed in passing.

Design Issues in Using htSNPs for Association Studies

John S. Witte and Deborah Thompson

The increasing availability of regularly spaced SNPs throughout the human genome has prompted many groups to use SNPs in association studies to search for genetic variants responsible for common diseases. More information can be gained from basing such studies on haplotypes, rather than considering each SNP separately. Several groups have demonstrated the existence of discrete blocks of low haplotype diversity within the human genome. High linkage disequilibrium within blocks means that information from a proportion of the SNPs is redundant, and the majority of the haplotypes can be distinguished using a much smaller number of SNPs, known as ’tagging SNPs’. Several approaches have been suggested for identifying the optimal tagging SNPs. We focus on the program tagSNPs (Stram et al., 2003), which selects tagging SNPs to minimise the uncertainty in predicting common haplotypes for individuals with unphased genotype data. Optimal tagging SNPs can be estimated using a small subgroup of the study population that have been genotyped for a dense SNP map, and it is just these tagging SNPs that are genotyped in the remainder of the samples. We used simulations to investigate how the size of the subsample affects the power of such an association study, and whether the subsample should consist of controls, or a mixture of cases and controls.

Inference of Specific Haplotype Effects in Case-Control Studies Using Unphased Genotype Data

Michael P. Epstein and Glen A. Satten

Haplotype-based association methods are powerful and popular procedures for identifying genes that influence complex disease. For a case-control study design, a variety of statistical methods exist that detect haplotype-disease association by comparing haplotype frequencies among sampled cases and controls. Given unrelated samples often consist of unphased genotype data (resulting in haplotype ambiguity), many such statistical methods account for missing haplotype information by using the Expectation-Maximization (EM) algorithm for inference. While existing haplotype-based association methods are important, the majority fail to determine how specific haplotypes influence disease. Inference of specific haplotype effects is valuable-particularly for identifying functional variants of a candidate gene. Therefore, we develop a retrospective likelihood for estimating and testing the effects of specific SNP-based haplotypes on disease in a case-control study assuming unphased geno! type data. Our proposed method has a flexible structure that allows modelling of main and interaction effects of specific haplotypes on disease. For statistical inference, we apply an Expectation-Conditional-Maximization (ECM) algorithm to account for the ambiguous haplotype information in the genotype data. Using simulation studies, our results suggest that our method returns unbiased estimates of specific haplotype effect size and has excellent power to detect such effects.

Association studies with tightly linked markers

Michael Knapp

Recently, Zhao et al. (2000) proposed a modification of the TDT which allows to analyze multiple tightly linked marker loci simultaneously, even if haplotypes are not directly observed. This talk presents a general approach for constructing size α tests in case of random variables for which the set of possible realizations is finite and in case that some of these realizations are equally probable under the null hypothesis. It is shown that the method by Zhao et al. (2000) can be viewed as a special case of this principle. By this, it is obvious that several intuitive modifications of Zhao's method which may increase the power of his approach will not affect its validity. In addition, the general approach shows how to extend Zhao's method to allow the analysis of general nuclear families with an arbitrary number of affected and unaffected children.

Using the non-informative families in family-based association tests: a powerful new testing strategy

Christoph Lange

For genetic association studies with multiple phenotypes, we propose a new testing strategy for family-based association tests (FBATs) that increases the power by both using all available family data and reducing the number of hypotheses tested. Using conditional power calculations, the approach identifies the subset of phenotypes which has optimal power when tested for association either by univariate or by multivariate FBATs. Further, using all available families, the approach provides estimates of the effects sizes without biasing the nominal significance level. In simulation studies, we compare our testing strategy with standard methodology. An application of our strategy to an asthma study shows the practical relevance and the robustness of the proposed methodology. The popular TDT is included as a special case.

Adaptive designs for candidate gene association studies with trios for data-driven design modifications

Andre Scherag

The use of conventional TDT test statistics in the analysis of candidate gene association studies requires the precise and complete pre-specification of the total number of trios to be sampled in order to control the error risk of a false positive result (type I error risk). Looking at the collected data before the pre-specified end and performing a conventional statistical level alpha test for early stopping in general will increase the risk of false positive results. In the case of most of these studies only little information about the genetic effects will be available beforehand and thus it will be difficult to specify the alternative hypothesis and to fix the sample size. In this situation we propose to allow for an interim analysis and explicitly plan to make appropriate sample size adjustments or changes based on estimates of the genetic effects obtained from the interim analysis (Scherag, Dempfle, Hinney, Hebebrand and Schäfer, in revision). We apply the method for data-driven design modifications developed by Mller and Schäfer (2001) in the context of a clinical trial which allows to change statistical design elements such as the sample size or to include an interim analysis for early stopping when no formal rule for early stopping was foreseen or to increase or reduce the number of planned interim analyses, without affecting the type I error risk. The method is based on the conditional rejection probability of a decision function.

IBD (Identity by Descent) Processes

Chris Cannings

The concept of Identity by Descent (IBD) is immensely valuable in capturing the information regarding genealogical structure, either for a specific sample or for a random process within a population. Using it we can investigate complex models of population including diploids, polygamy, proscription of certain types of marriages, clan structure etc.

IBD is usually used with respect to a single locus but in a more general setting IBD can be treated as a process across the genome, the IBD state changing at points corresponding to recombination events in the history of the genealogy.

Contributions to the understanding of this process were initiated by Donnelly(1983) who considered unilineal relationships between two individuals (ones who can only share one gene IBD), and showed that in this case the IBD process was equivalent to a random walk on a hypercube (the dimension matching the depth of the genealogy). Bickeboller and Thompson(1996(considered the case of half sibs using the Poisson clumping heuristic, and Stefanov(2000) cases similar to those of Donnelly exploiting methods developed for exponential families. Explicit expressions for the distribution of the total IBDlength in unilineal relationships are given by Walters and Cannings (20031), and for the number of segments by Walters and Cannings(2003b).

For more general genealogies and larger (than 2) sets of individuals the exact derivation of the distribution is intractable. One method is to simulate the flow through the genealogy as per Dimitropoulou and Cannings (2003). Another is to derive the matrix of transition rates between the possible IBD states, and to deduce from this, possibly by simulation, statistics of interest. The latter method has the advantage that the simulation is then far easier to carry out because the genealogy has been encapsulated into a much simpler structure. It is this latter method which will be described here, and the particular example of repeated sib-mating used as an illustration.

References

1. Bickeboller H and Thompson EA (1996) Distribution of genome shared IBD by half-sibs: approximation by the Poisson clumping heuristic. Theoret Pop Biol 50 66-90
2. Dimitropoulou P and Cannings C (2003) RECSIM and INDSTATS: Probabilities if Identity in General Genealogies. Bioinformatics In Press
3. Donnelly K (1983) The probability that related individuals share some section of the genome identical by descent. Theoret Pop Biol 23 34-63
4. Stefanov VT (2000) Distribution of genome shared identical by descent by two individuals in a grandparent-type relationship Genetics 156 1403-1410
5. Walters K and Cannings C(2003a) The probability density of the total IBD length over a single autosome in grandparent-type relationships Submitted for publication
6. Walters K and Cannings C(2003The distribution of the number of IBD segments in a single autosome in grandparent-type relationships Submitted for publication

Genome sharing in large pedigrees

Elizabeth Thompson

Gene identity by descent underlies patterns of phenotypic similarity among related individuals. Individuals concordant for a trait phenotype have increased probability of ibd at causal loci, and hence also at linked markers. Using MCMC methods, patterns of ibd can be realized jointly over individuals and jointly over loci, conditional on all marker data $\mathbf{Y_M}$ observed on the pedigree. Such MCMC methods have the advantage that ibd can be scored jointly over loci. Also the samplers can be adapted to gender-specific maps, genetic interference, and linkage disequilibrium among marker loci in pedigree founders. Often measures W of ibd are used to test for linkage, using statistics of the form $T \equiv \text{Exp}(W \mid \mathbf{Y_M})$. This is undesirable, since $\text{var}(T) \leq \text{var}(W)$ and suitable pedigree-based measures W often have highly skewed distributions, given $\mathbf{Y_M}$. Instead we propose p-values for linkage of the form

$$\text{Exp}(P(W^* \geq W | \mathbf{Y}^*_\mathbf{M}, \mathbf{Y_M}))$$

where the expectation is taken over data-set realizations $\mathbf{Y}^*_\mathbf{M}$ having the distribution of marker data having the same genetic map, marker-locus characteristics, and data-availability as the observed data $\mathbf{Y_M}$, but in the absence of linkage. The cumulative distribution functions of $W^*(\lambda)$ and $W(\lambda)$ are estimated for all of a set of chromosomal locations $\lambda \in \Lambda$ by MCMC given $\mathbf{Y}^*_\mathbf{M}$ and $\mathbf{Y_M}$, respectively. Combined p-values for linkage are then estimated by taking expectations over multiple data sets $\mathbf{Y}^*_\mathbf{M}$.

Meta-Analysis of Linkage Genome Scans

Astrid Dempfle and Sabine Loesgen

Linkage genome scans for complex diseases have low power with the usual sample sizes. Meta-analysis of several scans for the same disease might therefore be a promising approach. Appropriate data are getting accessible. Here, we give an overview of the available statistical methods and recent applications. In a simulation study, we compare the power of different methods to combine multipoint linkage scores, namely Fishers p-value combination, the truncated product method (Zaykin et al., 2002), the Genome Search Meta-Analysis (GSMA, Wise et al., 1999) method and a weighting method for nonparametric linkage scores (Loesgen et al., 2001). In particular, we investigate the effects of heterogeneity introduced by different genetic marker sets and sample sizes between genome scans. The weighting methods explicitly take those differences into account and have more power in the simulated scenarios than the other methods.

Group sequential study designs in genetic-epidemiological case control studies

I.R. Koenig, A. Ziegler

In the past years, the focus of genetic-epidemiological studies has shifted to the analysis of complex diseases. Here, a single gene often contributes only little to the manifestation of a trait; hence, many patients have to be included in a study to reliably detect small effects. To reduce the number of required phenotypings and genotypings in a study and thus facilitate the analysis of complex traits, sequential study designs can be applied. For the sequential analysis of candidate genes in association studies, we describe at first the procedure by Sobell et al. [1]. This includes the successive testing of many candidate genes with an adjustment of the significance level. Thus detected associations are validated in independent samples. Based on results from Monte-Carlo simulations, we discuss the efficiency of this procedure. Secondly, we present the adaptation of group sequential study designs by Pampallona and Tisatis [2] to the analysis of candidate genes. In this procedure, the sample of cases and controls is enlarged sequentially; after the genotyping of each subsample, association is analyzed in the cumulative data. Error rates and the efficiency of this proceeding are similarly investigated by Monte-Carlo simulations. Finally, we compare both procedures regarding error rates, efficiency, and practical applicability.

References

1. Sobell J. L., Heston L. L., Sommer S. S. (1993) Am J Med Genet 48:28-35.
2. Pampallona S., Tsiatis, A. A. (1999) J Stat Plan Inf 42:19-35.

Pairwise IBD estimation

Simon Heath

Estimates of pairwise measures of Identity by Descent (IBD) describe the identity relationship between the four alleles carried by any two individuals at a given locus. These measures are required for (amongst other uses) variance component methods of linkage analysis which use them to give the covariance structure of the major gene variance components. The problem of estimating the IBD measures is technically similar to classical linkage analysis in that it involves summation over all configurations of gene flow in pedigrees which are consistent with the observed data, and algorithms for linkage analysis can be simply adapted to the estimation of IBD measures. It is normally desirable to use all available marker and pedigree information, but this leads to computational difficulties unless pedigree are very small or few marker loci are considered. MCMC methods can be applied to this problem to perform the summation over gene flow patterns. Use of MCMC approaches has the advantages that large pedigrees with many marker loci can be used, and more complicated estimates can be easily obtained such as joint IBD sharing at multiple linked loci. However it can be difficult to obtain a well mixing sampler, particularly when markers are tightly linked. It can be useful to construct a MCMC sampler which has multiple types of sample updates to help the sampler efficiently move around the sample space. The sampler presented here has two main components: a locus sampler which updates jointly the state of all pedigree members at an individual marker conditional on neighbouring loci, and a meiosis sampler which jointly updates the grand-parental origin of the genes carried by an individual at all loci. Further improvement in the meiosis sampler by jointly updating blocks of meioses will also be presented.

Modelling of Periodic Screening for Lung Cancer of Smokers with Increased Mutagen Sensitivity and Reduced DNA Repair Capacity: Impact on Population–Based Mortality

Olga Yu Gorlova and Marek Kimmel

It has been shown that lung cancer cases consistently show higher mutagen induced chromosomal break scores and poorer DNA repair capacity (DRC) than age- and ethnicity-matched controls. These associations can help to identify individuals (smokers) at higher risk to develop lung cancer. In this study, we applied the data from both these assays in 149 cases of non-small cell lung cancer and 149 matched controls, to a previously developed and validated stochastic model of lung cancer natural history and detection (Flehinger and Kimmel, Biometrics 1987, 43: 127-144). This model allows estimating the mortality reduction associated with early detection of lung cancer followed by appropriate treatment (Flehinger et al., Cancer 1993, 72: 1573-1580; Gorlova et al, Cancer 2001, 92: 1531-1540). The lifetime susceptibility to get lung cancer distinguishes mutagen sensitive and insensitive individuals. We estimated the lifetime susceptibility using our previous estimate (17.4%) of lung ! cancer susceptibility based on the Mayo Lung Project and the case-control lung cancer study at the Epidemiology Department at the MD Anderson Cancer Center. As an example, the estimate of lifetime susceptibility to lung cancer of smokers who exhibited the mutagen sensitive phenotype equals 25%, as opposed to 14% for non-sensitive smokers. Modelling shows that annual CT screening of all smokers for 20 years can reduce mortality by 36%. If, within the same group, only individuals with elevated BPDE sensitivity and reduced DRC are screened, the mortality reduction is 19%. Further modelling to consider costs associated with screening the highest-risk segments of the population is warranted.

Edited by Tim Becker

Participants

Dr. Goncalo Abecasis
goncalo@umich.edu
Center for Statistical Genetics
Department of Biostatistics
SPH II
1420 Washington Heights
Ann Arbor MI 48109-2029 – USA

Prof. Dr. Chris Amos
camos@request.mdacc.tmc.edu
MD Anderson Cancer Center
Dept. of Epidemiology
Box 0189
1515 Holcombe Blvd.
Houston TX 77030 – USA

Prof. Dr. Max P. Baur
Max.Baur@ukb.uni-bonn.de
Institut für Med. Biometrie,
Informatik und Epidemiologie
Universität Bonn
Sigmund-Freud-Straße 25
D–53127 Bonn

Dr. Tim Becker
becker@imsdd.meb.uni-bonn.de
Institut für Med. Biometrie,
Informatik und Epidemiologie
Universität Bonn
Sigmund-Freud-Straße 25
D–53127 Bonn

Lars Beckmann
l.beckmann@dkfz-heidelberg.de
Deutsches Krebsforschungszentrum
Abteilung Klinische Epidemiologie
Im Neuenheimer Feld 280
D–69120 Heidelberg

Prof. Dr. Shelley Bull
University of Toronto
Samuel Lunenfeld Research Institute
Suite 850
600 University Avenue,
Toronto ONT M5G 1X5 – Canada

Prof. Dr. Chris Cannings
c.cannings@shef.ac.uk
University of Sheffield
Royal Hallamshire Hospital
GB-Sheffield S10 2JF

Prof. Dr. Ranajit Chakraborty
ranajit.chakraborty@uc.edu
Center for Genome Information
Dept. of Environmental Health
Univ. of Cincinnati, Kettering Lab.
3223 Eden Ave., Room 110
Cincinnati, OH 45267-0056 - USA

Dr. Jenny Chang-Claude
j.chang-claude@dkfz-heidelberg.de
j.chang-claude@dkfz.de
Deutsches Krebsforschungszentrum
Abteilung Klinische Epidemiologie
Im Neuenheimer Feld 280
D– 69120 Heidelberg

Dr. Nilanjan Chatterjee
chattern@mail.nih.gov
NCI, DCEG,
Biostatistics Branch
Epidemiologic Methods Section
EPS 8046
Bethesda MD 20892-7244 - USA

Dr. David G. Clayton
David.Clayton@mrc-bsu.cam.ac.uk
David.Clayton@cimr.cam.ac.uk
Diabetes & Inflammation Laboratory
Cambridge Institute for Medical
Research, Welcome Trust, MRC Build.
Addenbrook's Hospital
GB-Cambridge CB2 2XY

Prof. Dr. E. Warwick Daw
ewdaw@mdanderson.ort
MD Anderson Cancer Center
Dept. of Epidemiology
Box 0189
1515 Holcombe Blvd.
Houston TX 77030 - USA

Florence Demenais
demanais@evry.inserm.fr
INSERM EMI 00-06
Tour Evry 2
523 Place des Terrasses de
l'Agora
F-91034 Evry Cedex

Astrid Dempfle
dempfle@med.uni-marburg.de
Institut für Medizinische Biometrie
und Epidemiologie
der Philipps-Universität Marburg
Bunsenstr. 3
D–35037 Marburg

Prof. Dr. Kim-Anh Do
kim@mdanderson.org
Department of Biostatistics
Box 0447
U.T.M.D. Anderson Cancer Center
1515 Holcombe Boulevard
Houston TX 77030 – USA

Prof. Dr. Michael P. Epstein
mepstein@genetics.emory.edu
Emory University
School of Medicine
Department of Human Genetics
615 Michael Street, Suite 301
Atlanta GA 30322 – USA

Dipl. Math. Christine Fischer
christine_fischer@ukl.uni-heidelberg.de
Christine-Fischer@med.uni-heidelberg.
Institut für Humangenetik
Universität Heidelberg
Im Neuenheimer Feld 328
D–69120 Heidelberg

Frank Geller
geller@mailer.uni-marburg.de
Institut für Medizinische Biometrie
und Epidemiologie
der Philipps-Universität Marburg
Bunsenstr. 3
D–35037 Marburg

Dr. Simon Heath
heath@cng.fr
Centre National de Genotypage
CP 5721
2 rue Gaston Cremieux
F-91057 Evry Cedex

Prof. Dr. Marek Kimmel
kimmel@stat.rice.edu
Dept. of Statistics
Rice University
Houston, TX 77251 – USA

Dr. Michael Knapp
umt70e@uni-bonn.de
knapp@imsdd.meb.uni-bonn.de
Institut für Med. Biometrie,
Informatik und Epidemiologie
Universität Bonn
Sigmund-Freud-Straße 25
D–53127 Bonn

Dr. Inke R. König
koenigir@imbs.mu-luebeck.de
Institut für Medizinische Biometrie
und Statistik
Universität zu Lübeck
Ratzeburger Allee 160
D–23538 Lübeck

Prof. Dr. Michael Krawczak
krawczak@medinfo.uni-kiel.de
Institut für Medizinische
Informatik und Statistik des
Universitätsklinikums Kiel
Brunswiker Str. 10
D–24105 Kiel

Christoph Lange
hsimbe@mailer.uni-marburg.de
Dept. of Biostatistics
Harvard School of Public Health
677 Huntington Ave.
Boston, MA 02115 – USA

Dr. Berthold Lausen
berthold.lausen@rzmail.uni-erlangen.de
Institut für Medizininformatik,
Biometrie and Epidemiologie
Universität Erlangen
Waldstraße 6
D–91054 Erlangen

Dr. Nikolas Maniatis
N.Maniatis@soton.ac.uk
Human Genetics Division
School of Medicine
Southampton General Hospital
Duthie Building (MP 808)
GB-Southampton SO16 6YD

Dr. Ulrich Mansmann
mansmann@imbi.uni-heidelberg.de
Institut für Medizinische Biometrie
und Informatik
Im Neuenheimer Feld 305
D–69120 Heidelberg

Jason Moore
moore@phg.mc.vanderbilt.edu
Program in Human Genetics
Department of Molecular Physiology
and Biophysics, 519 Light Hall
Vanderbilt U., Medical School
Nashville TN 37232-0700 - USA

Dr. Hans-Helge Müller
muellerh@mailer.uni-marburg.de
Institut für Medizinische Biometrie
und Epidemiologie
der Philipps-Universität Marburg
Bunsenstr. 3
D–35037 Marburg

Prof. Dr. Jürg Ott
ott@rockefeller.edu
Rockefeller University
1230 York Avenue
New York, NY 10021 – USA

Dr. Klaus Rohde
rohde@mdc-berlin.de
Max-Delbrück-Centrum für
Molekulare Medizin
Berlin-Buch
Robert-Rössle-Str. 10
D–13125 Berlin

Prof. Dr. Helmut Schäfer
hsimbe@mailer.uni-marburg.de
Institut für Medizinische Biometrie
und Epidemiologie
der Philipps-Universität Marburg
Bunsenstr. 3
D–35037 Marburg

Andre Scherag
scherag@med.uni-marburg.de
Institut für Medizinische Biometrie
und Epidemiologie
der Philipps-Universität Marburg
Bunsenstr. 3
D–35037 Marburg

Dr. Konstantin Strauch
ktrauch@imsdd.meb.uni-bonn.de
Institut für Med. Biometrie,
Informatik und Epidemiologie
Universität Bonn
Sigmund-Freud-Straße 25
D–53127 Bonn

Prof. Dr. Duncan C. Thomas
dthomas@usc.edu
Department of Preventive Medicine
Division of Biostatistics
School of Medicine
1540 Alcazar Street, CHP-220
Los Angeles CA 90033 - USA

Elizabeth A. Thompson
thompson@stat.washington.edu
Department of Statistics
University of Washington
Box 35 43 22
Seattle, WA 98195-4322 - USA

Prof. Dr. H.-Erich Wichmann
wichmann@gsf.de
GSF-Inst. für Epidemiologie
Ingolstädter Landstr. 1
D–85764 Neuherberg

Prof. Dr. Thomas F. Wienker
wienker@uni-bonn.de
Institut für Med. Biometrie,
Informatik und Epidemiologie
Universität Bonn
Sigmund-Freud-Straße 25
D–53127 Bonn

Dr. Scott Williams
smwilliams@phg.mc.vanderbilt.edu
Program in Human Genetics
Department of Molecular Physiology
and Biophysics, 519 Light Hall
Vanderbilt U., Medical School
Nashville TN 37232-0700 – USA

Prof. Dr. John S. Witte
witte@darwin.cwru.edu
Genetic Epidemiolopy Unit
International Agency for Research
on Cancer
150, cours Albert Thomas
F-69372 Lyon Cedex 08

Dr. Hongyu Zhao
hongyu.zhao@yale.edu
Yale University School of Medicine
Epidemiology & Public Health
60 College Street
New Haven CT 06520 - USA

Prof. Dr. Andreas Ziegler
ziegler@imbs.mu-luebeck.de
Institut für Medizinische Biometrie
und Statistik
Universität zu Lübeck
Ratzeburger Allee 160
D–23538 Lübeck

Dr. Sebastian Zöllner
szoellne@genetics.uchicago.edu
The University of Chicago
The Division of Biological Sciences
Dept. of Human Genetics
920 E.58th ST.-CLSC 507
Chicago IL 60637 - USA