Visualisation of gene and pathway determinants of disease

Olivier Delrieu and Clive Bowman*

GlaxoSmithKline

1 Introduction

We present the basic concepts of visualising gene and pathway determinants of disease in individuals using empirically derived log Bayes Factors ('LBFs') from partial log-linear modeling defined by a simple classificatory contrast to a baseline group using single nucleotide polymorphism (SNP) data. Individualised aggregation to ontologies or 'buckets' is explained.

2 Case-control LBFs

Log diagnostic likelihood ratios (Pepe, 2003) have a long history in the the development and use of regulated medical diagnostic devices (see McNeil and Hanle, 1984; Radack *et al.*, 1986; and Weissler and Bailey, 2004). Their relevance to genetic classification can be seen by posing a metric for the individualisation of genetic differences as follows:-

The prior odds of a hypothesis H_p relative to a *baseline* hypothesis H_d is

$$\frac{\Pr(H_p|I)}{\Pr(H_d|I)}$$

where I is all the contextual information to hand.

Let H_d be : that the ith disease case individual arriving (denoted s_i) has the genome of a non-diseased control (denoted c) and let H_p be: that they instead have the genome of a different alternative group. Then, given a vector of single nucleotide polymorphism (SNP) genotypes over many loci $\mathbf{G_c}$ for a pre-specified set of control individuals, the posterior odds of H_p relative to H_d once the SNP polymorphism genotypes $\mathbf{G_{s_i}}$ for the ith case arriving are known is

$$\frac{\Pr(H_p|\mathbf{G_{s_i}}, \mathbf{G_c}, I)}{\Pr(H_d|\mathbf{G_{s_i}}, \mathbf{G_c}, I)}$$

which is equal to the prior odds multiplied by the likelihood ratio

$$\frac{\Pr(\mathbf{G_{s_i}}, \mathbf{G_c}|H_p, I)}{\Pr(\mathbf{G_{s_i}}, \mathbf{G_c}|H_d, I)}$$

This Bayes Factor or 'diagnostic' likelihood ratio (DLR) is the amount of evidence that the ith case individual is classified as not having the same genome as that of a set of controls.

Following a 'scene-anchoring' argument (Evett and Weir, 1998) or by 'extending the conversation' (Congdon, 2003), this likelihood ratio can be simplified to

$$\frac{\Pr(\mathbf{G_{s_i}}|H_p, I)}{\Pr(\mathbf{G_{s_i}}|\mathbf{G_c}, H_d, I)}$$

The denominator is known as the 'false alarm rate', the numerator the 'alarm rate' (Pepe, 2003).

The false alarm rate is the simple *draw* probability. This is posed herein as a product Bernoulli by assuming independence between loci (and for simplicity of exposition also assuming independent genotypes within a locus), given by $\prod_{j=1}^{N_{lg}} ({\theta_j}^{s_{ij}}.(1-\theta_j)^{1-s_{ij}})$ where N_{lg} is the number of genotype x locus combinations and s_{ij} is an indicator variable (0,1) that the ith case individual s_i has that jth genotype x locus combination. Here, θ_j is unknown and is estimated by the genotype x locus frequency in the controls. It is convenient to form a beta distributed Bayesian estimate $\hat{\theta_j}$ using a non-informative Uniform prior on θ_j thus avoiding numerical difficulties with zero observed counts.

Making the same assumptions, the hit rate (Pepe, 2003) is posed as $\prod_{j=1}^{N_{lg}} ({\nu_j}^{s_{ij}}.(1-\nu_j)^{1-s_{ij}})$ and a beta distributed Bayesian estimate of the genotype x locus frequency in the *arrived* cases used for $\hat{\nu_j}$. The frequency estimates from the arrived cases can include or may exclude the ith case (i.e. a *jack-knife* or *cross-validated* estimate).

Ignoring the marginal constraint within each locus, treating only the presence of each genotype as being informative and taking the log of the likelihood ratio (c.f. the LBF) gives an additive, relative-risk measure of the estimated amount of evidence for the ith case individual having the genome of a non-control

$$\sum_{j=1}^{N_{lg}} \left[(\log(\hat{\nu_j}) - \log(\hat{\theta_j}) \right] .s_{ij}$$

This extremely simple empirical Bayesian predictive measure is effectively the summation of case-control contrasts (i.e. group by SNP genotype differences or interaction) over genotypes x loci from a log-linear model estimated for each locus on its own (*partial log-linear modeling*) of all subjects in 2 groups, instantiated by the presence of the SNP genotype markers in *that* individual. For log-linear modelling see McCullagh and Nelder (1989). Each LBF is thus an estimate of the SNP genotype x locus by classification group interaction. This measure can be calculated for both cases and controls and effectively transforms the characterisation of people from a binary domain of SNPs to a rapidly calculable continuous measure with simple additive properties. Whilst statistical (and genetic) niceties have been eschewed in its derivation, this metric or distance is biologically insightful.

3 LBF 'Measure M' and 'Measure V'

For any disease case and for any non-diseased control a vector of LBF measures is a type of *profile*. This profile along the genome can be considered as a stochastic sequence and thus be characterised by its first and second moments (mean and variance respectively). These have a meaning in terms of the biological classificatory signal involved for *that* person (Figure 1). Posing other moments is possible viz. skewness, kurtosis etc or fitting appropriate distributions within and between individuals. Note that the likelihood ratio of a likelihood ratio is the likelihood ratio itself (Pepe 2003).

4 Aggregation of LBFs

The aggregation of such SNP-based discriminatory or classificatory evidence can be made over any domain to help answer biologically and clinically relevant questions. For instance, LBF values can be summarised or collapsed over:-

- genotypes within a locus to give a SNP level measure ('SNP LBF'),
- all SNP loci within a gene to produce a gene level measure ('gene LBF')

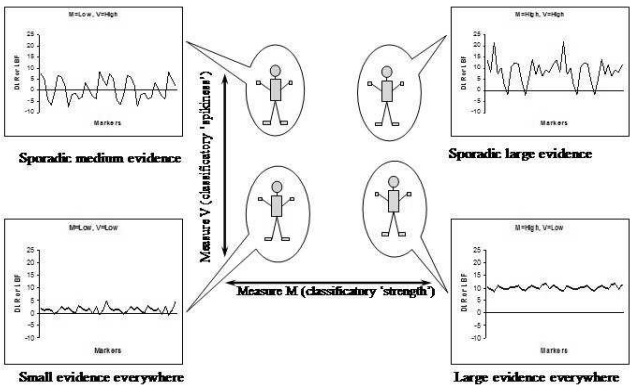

Figure 1: Typical individuals - Measure M and V

- any contrast or 'bucket' (ontology) of genes - for instance any group of genes coding for proteins involved in one biological pathway ('pathway LBF')
- the whole genome ('Measure M' - Figure 1)

using either simple summation or averages (with or without weighting). This allows the 'supervised' biological exploration of SNP data. Non-linear buckets are possible (if hard to interpret). More sophisticated aggregation of SNPs to genes by the use of inverse linkage disequilibrium 'smoothers' within a gene is possible.

5 Eigen analysis of LBFs

As an LBF profile along the genome can be considered as a stochastic sequence it can be characterised by further second order properties such as cross-marker covariances within a subject. Such covariances measure the evolutionary linkage disequilibrium (LD) used by geneticists to investigate concerted non-independence of SNPs along the genome (Sham, 1997). As only one genotype within a locus can be present, covariation within a SNP for a subject is zero.

Use of gene level aggregated mean(LBF) values for each individual to form a mean-centred overall sums of squares and cross-products matrix (cf. SSCP) across all subjects allows an 'unsupervised' eigen analysis of any common second order characteristics of the SNP profiles in cases and controls. For SSCP and eigen analysis see Mardia *et al.* (1979). This yields mutually orthogonal (un-correlated) eigen vectors of loadings (scaled coefficients) that best represent the total variation in that sample of cases and controls. Using the correlation matrix instead, *as we recommend*, allows variate re-scaling to a common metric more suitable for gene and pathway exploration where one seeks markers *relatively* implicated in case-control distinction even if they have small absolute biological effects. Gene LBF SSCPs are then rescaled to equivariance.

Since the diagnostic likelihood ratio is defined by reference to the false alarm rate, all first eigen vector loadings (scaled coefficients) must be positive by definition (bar any mathematical artifacts) and the first eigen vector *is* the case-control difference (i.e the case-control *signal*). Second and subsequent eigen vectors represent axes of heterogeneity *common* or *shared* in both cases and controls (i.e. *population stratification* or the case-control *'noise'*) which decline in their explanation of the variation in the sample (c.f. decreasing eigen values).

The larger the first eigen vector loading (scaled coefficient) for a marker, the more that marker is implicated in the case-control distinction. The larger the first eigen value the greater the detection of case-control genetic signal compared to total case-control noise. Raw coefficients for the ith eigen vector need to be rescaled by multiplication with $\sqrt{\lambda_i}$ (where λ_i is the ith eigen value) to yield comparable 'loading' vectors of equal variance.

Our procedure can be seen as simply correlation filtering or principal component regressions of partial log-linear model estimated contrasts of the gene determinants of the two groups of interest as visual partitions. Each gene is treated as a mutual 'nuisance' variable for each other gene and conditioned on each other (c.f. 'genomic control' - Devlin *et al.* (2001)). The 'signal' discriminating groups of interest is amplified by marginalising their data and re-applying to the group as an individualised contrast of the expected effect. 'Signal' and 'noise' are simultaneously estimated allowing for each other without explicit conditioning or holding one constant. *'Each person (or gene) tells you about both the signal and noise for each gene (or person)'*.

Each eigenvector is just a particular weighted set of correlated genes common in the people under study. In a genetic sense they are a measure of linkage disequilibrium on a certain metric and represent 'population diplotypes'. Mutually orthogonal eigen vectors condition out certain 'buckets' or pathways as measured by the coefficients in the equations. The second and subsequent eigenvectors are different common patterns of linkage disequilibrium amongst the people in the whole set, or put another way the pattern of loadings (scaled coefficients) for the second and subsequent eigenvectors define that type of linkage disequilibrium comprising those genes. In this way, one can see that the first eigenvector is the genes correlating (in linkage disequilibrium) with each other aliased (by definition of the LBF measure) with the case-control distinction. Put another way, the first eigenvector is that set of genes who have *correlated* changes in scaled frequency across the case-control distinction. The higher the loading (scaled coefficient) for a gene, the bigger the relative weighting for that gene in determining the aliased linkage disequilibrium diplotype.

References

Congdon, P. (2003) *Applied Bayesian Modelling.* John Wiley and Sons, Ltd.

Devlin, B., Roeder, K. and Wasserman, L. (2001). Genomic control, a new approach to genetic-based association studies. *Theor. Popul. Biol.*, **60(3)**, 155-166.

Evett, I. W. and Weir, B. S. (1998). *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists.* Sinauer Associates Inc.

Mardia, K. V., Kent, J. T and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* Chapman and Hall/CRC.

McNeil, B. J. and Hanle, J. A. (1984). Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Medical Decision Making*, **4(2)**, 137-150.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

Radack, K., Rouan, G. and Hedges, J. (1986). The Likelihood Ratio: an improved measure for reporting and evaluating diagnostic test results. *Arch. Pathol. Lab. Med.*, **110**, 689-693.

Sham, P. (1997). *Statistics in Human Genetics*. Hodder Arnold.

Weissler, A. M. and Bailey, K. R. (2004). A critique on contemporary reporting of likelihood ratios in test power analysis. *Mayo Clin. Proc.*, **79**, 1317-1318.