On using the correlations of divergences

Olivier Delrieu & Clive E. Bowman*

GlaxoSmithKline R&D,

Abstract

The choice, aggregation, weighting, and interactions of divergences, a natural metric of contrasts in the exponential family, is outlined from our experiences in applied genetics.

Key words: aggregation, biomarker, Bayes factor, copy number, epistasis, filtering, gene, haplotyping, interaction, LBF, LLR, log likelihood ratio, map, multivariate method, over-representation, SNP, visualisation.

Divergences (Jardine & Sibson, 1971) are the natural metric for the comparison of individuals or groups in the co-analysis or visualisation (i.e. the filtering) of multiple biomarker data types (see Fig.1 upper left). Correlations of divergences (see http://taxonomy.delrieu.org) can offer phenomenological insight and understanding. Covariances of them are scale dependent and focus on prediction. We could not have developed these ideas without the enduring personal support of Allen Roses.

Choice: Practitioners face choices within applied genetics (see Roses, 2004) and care in analyses matters. All filters sacrifice (i.e. chose to ignore) some features of the original data to highlight others. Different divergences can be used for different purposes (Table 1) and can be filtered differently. Correlations of divergences can be estimated over all individuals (Delrieu & Bowman, 2006a) or over a single group, before eigen decomposition overall, or of a single group (see Fig.1 lower left), and the biplot display of individuals. Eigen vectors of divergences can be further projected (Stem et al., 2007). Divergences can be derived for all distributions of the exponential family (see Table 2 for some useful examples). Frequentist or Bayesian parameter estimation (Congdon, 2001) is possible before ‘plug-in’ into divergences. Estimation can be marginal or joint. While -2*LLR is asymptotically distributed as $\chi^2_{\text{dof}}$ for nested hypotheses, inferences on divergences (Pardo, 2006) is best done using permutation distributions (e.g. Figs.1 upper right,4).

Aggregation: SNP log Bayes Factors ($lbfs$) are the data’s empirical genetic model and non-linearly stretch and squash the axes of multi-dimensional SNP data. $lbfs$ generate variance (i.e. scale) between groups (by definition (Delrieu & Bowman, 2005) as well as within groups (i.e. departures from a single genotype). Aggregations can strengthen correlations by reducing ‘noise’. Linkage disequilibrium (LD) generates covariance (i.e. pattern). Delrieu & Bowman (2006a) linearly aggregate these specific SNP divergences and then perform eigen analysis of correlations of these sample aggregates over individuals - much as characters are coalesced in agglomerative clustering. If SNPs are not in LD within an aggregate (e.g. ‘singleton signals’) this will correctly down-play the $lbfs$ for that aggregate. Presence of LD will reinforce aggregate signals. The correlation calculation for aggregates in Delrieu & Bowman (2006a) may discount some within aggregate information as, $\text{var}_k \Sigma_m(lbf) \neq \text{just a } \Sigma_m$ function of $\text{var}_k(lbf_m)$ (where $k$ indicates over individuals and $m$ over markers) for an aggregate necessarily. For instance, let $\overrightarrow{x}$ be a row vector of the $p$ SNP loci genotyped, and, let $\overrightarrow{lbfs}$ be the row vector of $p$ $lbfs$ variables derived from $\overrightarrow{x}$ for each person, then, the correlation matrix of $lbfs$ variables over k people is the mean-centred, rescaled average SSCP over people:

$$\Theta = \text{CORR}_k(lbf_1, \ldots, lbf_p) \equiv \text{elements } [\text{corr } (lbf_i, lbf_j)] = r_{ij} \text{ (for all } i \text{ and } j = 1 \text{ to } p),$$

where $r_{ij} = \frac{\text{cov}(lbf_i, lbf_j)}{\text{var}(lbf_i) \text{var}(lbf_j)}$ over people. So, for 3 SNPs then $\Rightarrow \Theta = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}$.
Table 1: Use of divergences (Kullback, 1997) in applied genetics. Delrieu & Bowman (2005) introduced directed log Bayes factors (lbf) or log likelihood ratios (LLR) as additive contrasts for decomposing the genetic determinants of disease. Delrieu & Bowman (2006a) exemplified the visualisation of their correlations in marker triage as well as in group and individual outlier detection. Bowman et al. (2006) used directed expected LLR ratios in the eigen analysis of outlier gene by treatment correlations. $2^{*}i^{*}(KL)$ is asymptotically $\chi^{2}_{g-1}$ for a SNP (see Kupperman, 1958). Note: Undirected Symmetric J could be used, for instance, to compare the genetics of severe, moderate, mild and no disease cohorts; or to compare placebo, ‘gold standard’ and test medicine groups; or to compare various ethnic groups genetically within a large population study etc.

Table 2: Some miscellaneous useful divergences for applied genetics. $\psi(k)$ is digamma $\Gamma(k)\Gamma'(k)$. For binary data in Delrieu & Bowman (2005, 2006a), KL = $\nu_{i}\log\frac{\nu_{i}}{\theta_{i}}$ for cases and $\log\frac{\theta_{i}}{\nu_{i}}$ for controls; Symmetric J - their average. For normal distribution formulae see Bowman et al. (2006). For multivariate normal see Dragalin & Fedorov (1999). * ⇒ Bayesian a posteriori estimate Congdon(2001) using vague Gamma(0.001,0.001) prior for Poisson parameter $\lambda$ is [sum(of number data) plus 0.001] / [count(people) plus 0.001] rather than raw estimate of (sum/count). ** ⇒ Bayesian a posteriori estimate Congdon (2001) using Gamma(a, b) prior for rate parameter $\lambda$ is [number of data points plus a] / [total of data plus b] rather than raw estimate of (1/mean). Analogous divergences of different data types can be put in simultaneous analyses e.g. ‘location model’ for binary and normal.
Thus, for a gene made of \( m \) unweighted SNPs: 

\[
lbf_G = \frac{1}{2} (lbf_1 + lbf_2)
\]

and SNP covariance matrix \( \Psi = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \). Then, let a gene \( G \) be a simple linear unweighted mean aggregate of 2 SNPs: 

\[
lbf_G = \frac{1}{2} (lbf_1 + lbf_2)
\]

Now a different choice to that of Delrieu & Bowman (2006a) is to transform \( \Theta \) into a new correlation matrix \( \Omega \) under SNP aggregation to gene \( G \) using \( \Psi \). Here \( \Omega = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} \), with 

\[
q_{12} = q_{21} = \text{corr} (lbf_G, lbf_3) \quad \text{and} \quad q_{11} = q_{22} = 1.
\]

We have:

\[
q_{12} = \text{corr} \left( \frac{1}{2} (lbf_1 + lbf_2), lbf_3 \right) = \frac{\text{cov} \left( \frac{1}{2} (lbf_1 + lbf_2), lbf_3 \right)}{\sqrt{\text{var} \left( \frac{1}{2} (lbf_1 + lbf_2) \right) \text{var} \left( lbf_3 \right)}}
\]

\[
cov \left( \frac{1}{2} (lbf_1 + lbf_2), lbf_3 \right) = \frac{1}{2} \text{cov} (lbf_1, lbf_3) + \frac{1}{2} \text{cov} (lbf_2, lbf_3)
\]

\[
\text{var} \left( \frac{1}{2} (lbf_1 + lbf_2) \right) = \frac{1}{4} \left( \text{var} (lbf_1) + \text{var} (lbf_2) + 2 \text{cov} (lbf_1, lbf_2) \right)
\]

Hence:

\[
q_{12} = \frac{c_{13} + c_{23}}{\sqrt{(\text{var} (lbf_1) + \text{var} (lbf_2) + 2c_{12}) \text{var} (lbf_3)}}
\]

Thus, for a gene made of \( m \) unweighted SNPs: 

\[
lbf_G = \frac{1}{m} \sum_{i=1}^{m} lbf_i
\]

Thus, for two linear unweighted aggregations of genes: 

\[
lbf_{G1} = \frac{1}{m} \sum_{i=1}^{m} lbf_i \quad \text{and} \quad lbf_{G2} = \frac{1}{n} \sum_{i=m+1}^{m+n} lbf_i
\]

\[
q_{G1,G2} = \frac{\text{cov} (lbf_{G1}, lbf_{G2})}{\sqrt{\text{var} (lbf_{G1}) \text{var} (lbf_{G2})}}
\]

\[
q_{G1,G2} = \frac{1}{m} \sqrt{\sum_{i=1}^{m} \text{var} (lbf_i) + 2 \sum_{1 \leq i < j} c_{ij}} \frac{1}{n} \sqrt{\sum_{i=m+1}^{m+n} \text{var} (lbf_i) + 2 \sum_{m+1 \leq i < j} c_{ij}}
\]

\[
q_{G1,G2} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} c_{ij} \sum_{i=m+1}^{m+n} \sum_{j=m+1}^{m+n} c_{ij}}{\sqrt{\sum_{i=1}^{m} \sum_{j=1}^{m} c_{ij} + \sum_{i=m+1}^{m+n} \sum_{j=m+1}^{m+n} c_{ij}}}
\]

The numerator in equation (3) represents the between gene covariation, the denominator the within gene covariation. In practice (Fig. 2 left), the
sample ‘uncorrected’ estimates of the correlations of SNP aggregates using Delrieu & Bowman (2006a) are close, on average, to those from equation (3) using the per marker lbfs (denoted ‘proper’ herein). However, this is not always true (see Fig. 2 centre) due to rounding errors and mean centring effects ⇒ Choice of filter implementation matters.

Figure 1: Upper left: Simultaneous or co-analysis of multiple data types in applied genetics. Psychiatric disease A example from Delrieu & Bowman (2006a). Biplot showing case control distinction (cases on right, controls on left - CaseCont vector inclined to horizontal axis), based upon SNP (_nnnnnn), pathway (PATHn) and clinical anxiety data (anxietySubject) together with within group variation explained by various overweight measures. Note neurological pathway 3 and anxiety correlate well with disease status supporting a postulated mechanism involved in the phenotype. Upper right: Genetic determinants of drug-induced SJS/TEN in Caucasians Delrieu & Bowman (2006b). Permutation distribution under the null of 2D case-control projected lbfs loading1s - 3587 permutations of case-control status. Observed 2D projected loading1 indicates significant genetic effect of marker HLA-B*57 (which was the highest 2D case-control projected loading1 in original dataset out of 137 SNPs and HLA markers) on the predisposition to these adverse events. Lower Left: Cases-only eigen decomposition showing distinct genetic determinants within the two drug-induced disjoint AE phenotypes SJS and TEN (data from Stern et al. (2007)). Cases as triangles, controls as crosses. Note importance of constituents of Sanger haplotypes (DBB, MANN and PGF, see http://www.sanger.ac.uk/HGP/Chr6/MHC/). Lower Right: Decomposition of two overlapping drug-induced gut phenotypes (AE ‘I’ and AE ‘S’). Green crosses - simultaneous display of 133 ADME genes together with their 1043 SNPs constituents and suspect drug metabolic pathway aggregate (Drug_MeT). Note how drug metabolism is more correlated with one AE than the other.

Weighting: A larger number of SNPs in a gene potentially increases the magnitude and granularity of $\text{mean}_k(lb_{fG1})$ evidence (where k indicates across individuals). Whilst in practice large genes have larger $\text{var}_m(lb_{fM})$ (where m indicates over markers within an aggregate), a small gene may unluckily by chance alone have a high lb (or larger $\text{var}_k(lb_{fG1})$). Choice of unweighted means of marker lbfs in the eigen analysis visualisations can favour those aggregations with larger number of constituents (Fig.2 right, 3 left). This is driven by the structure of the SNP to gene map (Fig.3) used in filtering the dataset altering the precisions of aggregate averages and the distribution of SNP lbfs results. It can be approximately adjusted for (Fig.2 left and right) using a diagonal re-weighting (rationalised below), or estimated by regression and subtracted (for modest sizes of genes - Figs.2 right and 4 right), and or obviated in permutation.
Table 3: Example of dummy factors $X \diamond Y$ to indicate categorical SNP interaction terms of $X$ with $Y$ and the choice of parameterisation for interactions. All main effects and interactions are entered into the lbf correlation matrix and eigen analysis simultaneously, i.e. to use interaction term: SNP1 $\diamond$ SNP2; also include SNP1 and SNP2; to use $HLA-B^*57 \diamond -C^*06$ also requires inclusion of dummy variables (as follows):- for genotypic analysis, $HLA-B^*57$ (scored YY, YN or NN) and HLA-C*06 (YY, YN, NN) see Delrieu & Bowman (2006b); for allelic analysis, $HLA-B^*57$ (scored +1=yes/-1=no) and HLA-C*06 (+1=yes/-1=no), see Fig.3 right. Note the different way in which heterozygotes can be dealt with.

Figure 2: Results for typical large loading1 ‘RedBox’ gene RYR2 and dataset disease A from Delrieu & Bowman (2006a). Observed data held fixed, $> 1330$ permutations of SNP to gene mapping retaining map structure (as given in Fig.3 centre). Left: Distribution of loading1 in uncorrected (Delrieu & Bowman, 2006a), weighted and ‘proper’ analysis. Note:- effect of rescaling by weighting; agreement on average of two unweighted methods; location of observed ‘proper’ loading1 under null mapping i.e. likely over-representation of large SNP effects due to the concerted LD and reinforcement of signal amongst the RYR2 SNPs. Centre: Distribution of the difference in loading1 from original aggregation method of Delrieu & Bowman (2006a) versus that of the exact method herein (denoted ‘proper’) showing that they do differ under the null mapping from time to time and significantly for this dataset ⇒ Covariance terms within an aggregation matters when using aggregations. Right: Mean over SNP to gene mapping permutations for each gene for the three approaches. Note: for unweighted methods (over this range of gene sizes), a square root relationship with higher loading1 values for aggregates containing larger number of data SNPs (given this map structure - Fig.3 centre and this distribution of SNP lbf). However, weighting effectively ameliorates reduced diagonal variance terms (see text). Note residual LD re-inforcement of unweighted signals shown by small disparities to fitted line (c.f. Fig.4 right). Dotted line indicates ‘RedBox’ boundary Delrieu & Bowman (2006a).
Figure 3: *Left:* Observed data of Disease A (Delrieu & Bowman, 2006a), observed data for [proper minus uncorrected] loading1 for each gene aggregate together with line through mean per number of data SNPs. Note broad square root relationship. Correction is especially important for large genes. *Centre:* Frequency histogram of genes with different numbers of data SNPs for map used in Fig. 2 right. Fitted distribution: Geom(0.026) or NegBin(1.0,0.026), $\chi^2$ goodness of fit $= 10^{-4}$ or; inverse quadratic curve fit $y = 1133.09x^{-2.06}$. Different map structures will yield same relationships of unweighted gene loading1s with the number of data SNPs under the null. *Right:* Interaction analysis (see Delrieu & Bowman (2006b)). Skin cancer cell lines courtesy of J Kulski, Murdoch University, Western Australia; HLA and Alu loci - 8091 main effects and interactions. Note discrete clusters of two loci marker combinations indicating specific haplotypes of interest to the right. Vertical displacement indicates different genotypes within the haplotypic loci.

Figure 4: Observed data of Disease A from Delrieu & Bowman (2006a) - 1512 genes. Map structure as in Fig. 3 centre preserved. Probability [of loading1 for a gene aggregate under the null permutations ($> 1330$) of SNP to gene map] being $< $ actual observed value (i.e. high values = significant over-representation of effects in gene) - see Fig. 2 left. *Left:* Same; by number of data SNPs in aggregate. Big genes can show high granularity and here tend to show over-representation. But some large genes appear unimportant (e.g. GRINB2), as despite their large signal, their signal is very likely to be by chance alone. Note: this display does not yet correct for number of data SNPs per gene. *Centre:* Same; by loading1 showing RedBox gene RYR2 observed could be due to chance given this map structure i.e observed RYR2 may be dominated by a, or only a few, disjoint highly associated singleton SNP(s). Horizontal line $\alpha=0.2$. Vertical dotted line indicates 'RedBox' boundary (odds of signal versus 'noise' = 4). Note: this display does not yet correct for number of data SNPs per gene. *Right:* Same, by loading1 adjusted for gene size via regression, showing EDIL3 remains an important aggregate amongst those in the top right hand quadrant of interest. Horizontal line $\alpha=0.2$. Large genes, *per force*, can easily show over-representation. Over-representation in a RedBox gene will also be indicated by small $\text{var}_m(lbf_m)$ for subjects (where $m$ indicates over markers within an aggregate).
testing (Fig.2 left and Fig.4).

Aggregations apply a design matrix $X$ to the data. A set of linear aggregations, if mutually orthogonal, certainly affect variance:- Consider $lbf$ data for 3 SNPs on K people (i.e. $lbf_{kj}$ for $j=1$ to 3 and $k=1$ to $K$) with no missing values (note: $j$ here subsumes extra index $i$ for up to 3 genotypes within a SNP; Delrieu & Bowman, 2006a). Mean-centre and divide these by the square root of the variance across people for each SNP $j$ (i.e. $d_{kj} = \frac{lbf_{kj} - \text{mean}_k(lbf_{kj})}{\sqrt{\text{var}_k(lbf_{kj})}}$),

where $k$ denotes across people. Call this $K \times 3$ data matrix, $\Delta = \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ \vdots & \vdots & \vdots \\ d_{K1} & d_{K2} & d_{K3} \end{bmatrix}$. Then

$$SSCP(\Delta) = \Delta^T \Delta = CORR\Delta(\Delta) = \begin{bmatrix} 1 & \frac{1}{K} \sum_{k=1}^{K} (d_{k2}.d_{k1}) & \frac{1}{K} \sum_{k=1}^{K} (d_{k3}.d_{k1}) \\ \frac{1}{K} \sum_{k=1}^{K} (d_{k2}.d_{k1}) & 1 & \frac{1}{K} \sum_{k=1}^{K} (d_{k3}.d_{k2}) \\ \frac{1}{K} \sum_{k=1}^{K} (d_{k3}.d_{k1}) & \frac{1}{K} \sum_{k=1}^{K} (d_{k3}.d_{k2}) & 1 \end{bmatrix}$$

$\rho$ which is $3 \times 3$ and $\frac{1}{K} \sum_{k=1}^{K} d_{kj}^2 = 1$ for all $j=1$ to 3. Now, let $X$ be the design matrix for a set of orthogonal linear aggregations i.e. $X^T.X$ is diagonal:- Say, we define a gene ($G1$) as the unweighted mean of two SNPs and leave the remaining SNP alone, i.e. $d_{k(G1)} = \frac{d_{k1} + d_{k2}}{2}$ for all $k=1$ to $K$. Let $X = \begin{bmatrix} 1 & 0 \\ \frac{1}{2} & 0 \\ 0 & 1 \end{bmatrix}$ which is $3 \times 2$. Note $X^T.X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. Then $\Delta.X = \begin{bmatrix} \frac{d_{11} + d_{12}}{2} & d_{13} \\ \frac{d_{21} + d_{22}}{2} & d_{23} \\ \frac{d_{k1} + d_{k2}}{2} & d_{K3} \end{bmatrix}$ which is $K \times 2$ and $SSCP(\Delta.X) = X^T.\Delta.X = \Upsilon$ which is $2 \times 2$.

But this is not $CORR\Delta(\Delta.X)$ as the leading diagonal of $\Upsilon$ will no longer be all 1s $\Rightarrow$ Choice of the parameterisation of $X$ matters. This can be seen by noting that although the column means of $\Delta.X$ will still be zero, the first column now has a variance $\neq 1$. By definition, $\text{mean}_k(d_{kj}) = 0$, $\text{var}_k(d_{kj}) = 1$ for all $j$, then, as a first approximation, under a null of SNP independence i.e. covariance=0 within an aggregate $\Rightarrow$ var$_k \sum_{k=1}^{K} (\frac{1}{2}(d_{k1} + d_{k2})) = \frac{1}{2}(K.\text{var}_k(d_{k1}) + K.\text{var}_k(d_{k2})) = S$ say, $\Rightarrow S = K/2$, and this could be used for $\Upsilon$ adjustment to achieve $CORR\Delta(\Delta.X)$ before eigen decomposition. Now:- $K$ is fixed by the data for all aggregations and can be ignored (i.e. subsumed into eigenvalues via $\sum(eigenvalues) = \text{total variation in sample}$); leaving $S = 1/2$ which is $< 1$ i.e. the corresponding diagonal element of $\Upsilon$ is too small compared to that of a non-aggregated SNP. To produce an equivalent matrix with a diagonal set to 1s, then, choose to simply inflate the variances in the denominator of all off-diagonal correlation calculations featuring $G1$ by $x 1/2$ i.e. $x 2$, or more simply divide elements $\upsilon_{hi}$ in $\Upsilon$ by $\sqrt{2}$ for $h \neq i=1$ to 2. This generalises for $lbf_{k,G1} = \frac{1}{m} \sum_{j=1}^{m} lbf_{kj}$ and $lbf_{k,G2} = \frac{1}{m} \sum_{i=1}^{m} lbf_{kj}$ to an adjustment proportional to $\frac{1}{\sqrt{mn}}$, or the following effective correction to the off-diagonal gene level correlation coefficients (see ‘weighted’ in Fig.2):- $q_{G1,G2}^* = \frac{G1.G2}{\sqrt{mn}}$ where G1 and G2 have (1..m) and (1..n) SNPs within them. The matrix $X$ is used with resultant eigenvectors to calculate subject gene level $lbf$ scores. This $\Upsilon$ weighting is only approximate as LD within a gene is actually likely for any mutations of biological significance. Moreover, LD would be the
origin of any over-representation of large signals within a gene. Clearly, this re-weighting is not necessary for a set of genes with broadly similar number of SNPs in each, nor should be made for an experiment where there has been a purposeful rationale for greater SNP sampling within an a priori region, or for a dense region with strong LD, or where one is content to favour more examined or more evolutionarily variable genes.

A different choice would be to re-standardise the columns of $\Delta X$ and dispense with any weighting for $\Upsilon$ (i.e. rather than say expand the diagonal variance term further with within aggregate covariation using the bilinear rule - equation (2), or even considering adjusting the cross-product terms themselves with equation (1)). Another choice would be to ensure that the sum of the squares of the terms of a column in $X$ add up to 1 i.e. $X.X^T = I$, but then this matches a gene effect of $\frac{d_k^1 + d_k^2}{\sqrt{2}}$ not the desired $\frac{d_k^1 + d_k^2}{2}$. The choice is to keep the variability (error) the same, which effectively amplifies the aggregate signal, or leave the data unstandardised keeping the signal the same in the new aggregate direction but reducing its variability ⇒ Care is needed. Correlation structure ignored in aggregation can be partly retrieved by the concomitant eigen decomposition of aggregates together with all their constituents (see Fig.1 lower right) i.e. if solely the relative directions are of interest. Mappings of one SNP to one gene, or many SNPs to one gene do not induce extra correlations in themselves, but mappings of one SNP to many genes in aggregations will ⇒ Choice of map matters. Dummy uniform data can be used in eigen analyses to diagnose such non-orthogonality (see: graphical method for missing data in EigenAnalysis.sas within Taxonomy3 software on http://taxonomy.delrieu.org). Eigen visualisations using any $X^T.X$ with off-diagonal structure would need careful interpretation as such will induce extra pattern. Even then, inferences could be done using permutations of case-control status together with permutations of SNP to gene mapping as appropriate to the question of interest.

**Interactions:** For binary divergences, categorical biomarker interactions can be examined by constructing dummy variables (see Table 3 for SNP and HLA examples) to be included in with the analysis of the main effects. The choice of how to deal with heterozygotes determines the coding. Two-marker combinations (epistasis) appearing in similar locations on biplots (i.e. correlated) can indicate the presence, span and constituents of the likely relevant haplotypes containing those markers (Delrieu & Bowman (2006b) and Fig.3).

**References**


