Correlation laplacians, haplotype networks and residual pharmacogenetics

Clive Bowman[1*] and Olivier Delrieu[2]

[1]School of Biological Sciences, The University of Reading, Whiteknights, Reading, RG6 6AH, UK

[2]PGX-IS, Kingsmead Business Park, Frederick Place, High Wycombe, HP11 1LA, UK

[*](correspondence to: c.e.bowman@reading.ac.uk tel: +44-7810-506362 fax:+44-8700-557753)

Abstract

The walk laplacian transform of genetic correlation matrices is described and the network display and interpretation of ensuing haplotype relations exemplified. Pharmacogenetic epistatic interactions can be decomposed using residual projections.

Key words:- divergences, filtering, SNP, gene, log likelihood ratio, multivariate projection, covariance visualisation, Bayes Factor, eigen analysis, bio-marker interaction, PAJEK, nodes, edges, linkage disequilibrium, epistasis.

Introduction: Delrieu and Bowman (2005), Bowman *et al.* (2006), Delrieu and Bowman (2007) and Charalambous *et al.* (2008) present a triage framework for effects in supervised genetic analyses (Bowman and Delrieu (2009a)) - building upon the work of Jardine and Sibson (1971). This R-technique uses correlation matrices of information contrasts (individualised log likelihood ratios or log Bayes Factors - *lbf*s) derived on variables from the exponential family of distributions. Eigen analysis of such offers simultaneous multidimensional insight in drug discovery (Delrieu and Bowman (2006a)) producing cascading orthogonal latent sets of (partial) redundancies amongst markers of diminishing relevance to the phenotypic contrast - see http://taxonomy.delrieu.org for software. Pirmohamed *et al.* (2007) and Alifrevic *et al.* (2009) have recently used such to decompose the aetiology of complex drug-induced disorders. Orthogonal ordination is just one visualisation of correlation structures, directed graphs (weighted networks) another topology.

Correlation laplacian: For an undirected graph, take the estimated *lbf* correlation matrix $\hat{\Sigma}$ ($\equiv MSSSCP$ matrix) i.e. $\begin{bmatrix} \hat{\rho}_{1,1} & \hat{\rho}_{1,2} & \hat{\rho}_{1,3} \\ \hat{\rho}_{2,1} & \hat{\rho}_{2,2} & \hat{\rho}_{2,3} \\ \hat{\rho}_{3,1} & \hat{\rho}_{3,2} & \hat{\rho}_{3,3} \end{bmatrix}$ as, for example, $\hat{\Sigma} = \begin{bmatrix} 1 & 0.9 & 0.1 \\ 0.9 & 1 & -0.2 \\ 0.1 & -0.2 & 1 \end{bmatrix}$, then threshold each element using a specified value, say ± 0.2 (as the 'size' of a 'link of importance') to form an 'adjacency matrix', resetting the diagonal elements to zero, to yield $A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$. Following Skillicorn (2007), form the row sum of A as a diagonal matrix $D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Calculate the 'laplacian' L as L = D-A $= \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$. Then calculate the 'representation matrix' R - which is the 'walk laplacian' or transition matrix R=D^{-1}.L

(equivalently R=I-D^{-1}.A) $= \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix}$. Give R to a network plotting routine like PAJEK (http://pajek.imfm.si/doku.php). In the above, A has all positive (or zero) entries for adjacency; all the negative correlations in $\hat{\Sigma}$ are showing is the *type* of edge connecting 2 nodes (markers). For directed graphs, do as above, but reset the sign of the off-diagonal elements of the final R ($r_{i,j}, i \neq j$) to match the sign of the original $\hat{\rho}_{i,j}$. Note, that you could also do no thresholding but discard the sign in $\hat{\Sigma}$ then weights flow through as the row sum which is then the sum of abs(correlation) etc. If $\hat{\rho}_{k,m} = corr(lbf_{k,m})$ for SNP (or gene) k and m in a whole genome scan, then the edge between node k and node m in the resultant network represents the haplotype relationships between k and m in the space (context) of the contrast (trait) of interest - see Bowman and Delrieu (2009b). Figure 1 shows an example of such a directed haplotype graph 'appearing from the mist' as the covariance sensitivity threshold decreases. This graphical result is a minimum spanning 'tree' of marker relationships without the topological strict branching constraint.

Haplotype networks: The empirical distribution of the number of edges for each node is an important indicator for a network's function (see Barabási (2002)). Figure 2 illustrates that for drug-induced Stevens-Johnson syndrome/Toxic epidermal necrolysis (see Pirmohamed *et al*. (2007)), the haplotype relationships (shown in detail in Bowman and Delrieu (2009b)) neither match an Erdős-Rényi random network, nor a scale-free topology. High values for the degree of a node can be because it is a part of an extended haplotype (like HLA-C*06, HLA-B*57 and HLA-DRB1*07 in MHC 57.1 with co-occurring TNF mutations) - with the network edge density being locally high around it; or, it can be because it is a 'lynch pin' hub where sets of different haplotypes 'cross' in common (like perhaps for HLA-DRB3*02) - see Bowman and Delrieu (2009b). The non-random mixture of haplotype sizes of direct relevance to the trait suggested by such displays as Figure 2 can inform further follow-up, recalling that in this approach 'haplotypes' are not physically constrained to colocation nearby or even on the same chromosome.

Residual pharmacogenetics: In a case-control association analysis, given *lbf* values for two markers of interest (Delrieu and Bowman (2006a)), form a dummy interaction marker (as in Delrieu and Bowman (2007)) and estimate its *lbf* for each individual. Note, at this point, do not necessarily simultaneously display ordinations of marker by marker interactions and marker effects as in Delrieu and Bowman (2006b). Rather, regress the dummy interaction *lbf* values (without an intercept) on the per-person *lbf* values of both the original constituent markers within that interaction. Take the residual of this regression (aka 'projection') as a measure of the more 'bang for your buck' given by epistasis between the two original markers with respect to the contrast (trait) of interest. Do for all pairs of markers. Triage by performing eigen decomposition of the estimated correlation matrix of *these* measures over all the two marker interactions (including a dummy case-control variable). Use the magnitude of the (case-control projected) loading 1 from *this* eigen analysis as the off-diagonal elements of a *new* matrix S such that $s_{i,j}$ is the (case-control projected) loading 1 for the (residual) dummy interaction marker between markers i and j. This matrix can be censored by an interaction-analysis 'RedBox' cutoff over the interaction markers (see Delrieu and Bowman (2006a)) to yield the first (major) latent set of two-marker interactions (=extended haplotypes). Set diag(*S*)=0, threshold and make positive as appropriate (see correlation laplacian above) to form an adjacency matrix A and proceed as before to display a network - see Figure 3 for an example. The resultant network represents the major (first) set of how each marker - potentially influencing the contrast of interest - *interacts*

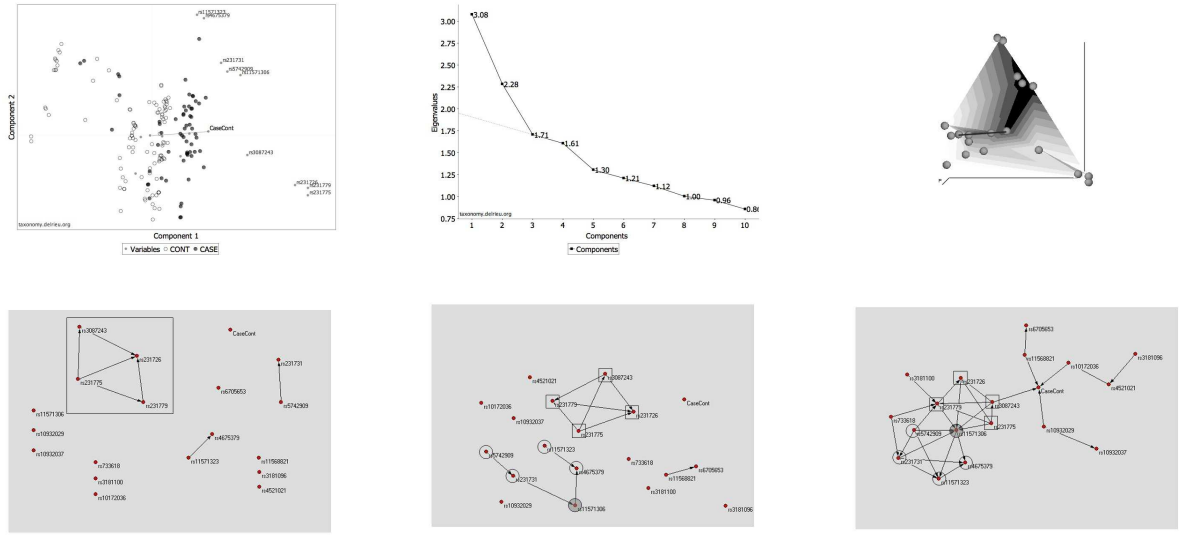

Figure 1: Example of carbamazepine induced HSR (105 controls; 61 cases mild and severe; 4 genes - CD28, CTLA4, ICOS, PD1 - comprising 18 SNPs; ex Eunice Zhang). *Top*: Genetic ordination (see Delrieu and Bowman (2006a)). *Left*:- Biplot. Note substantial heterogeneity on both components. *Middle*:- Scree plot showing only 2 latent components for trait over and above genetic 'noise'. *Right*:- Heat map of the 'this-sample' case-control permutation p-value (see Bowman (2009)) over the ordination space for each SNP. Darker colour means more significant. Spheres are the markers as in ordination on the top left. SNP rs111571306 (ICOS) was the most significant p=0.074 (10,000 permutations). *Bottom*: Walk laplacian: normalised graph laplacian of thresholded *lbf* correlations of SNP carriage displayed as weights using PAJEK. Edges between nodes in such plots define sets of correlated SNPs (cf. 'haplotypes') in the space of the contrast of phenotype of interest. Solid lines = positive correlation. Dotted lines = negative correlation. Orientation of each simplex is for convenience of visual clarity only. *Left*:- High threshold (=0.30) showing haplotypes forming within CTLA4 and within ICOS genes. SNP group in large square are all in CTLA4. *Middle*:- Diminishing threshold (=0.28). Markers in circles load heavily on the 2nd ordination component (comparison of severe HSR to others), those in squares on the 1st component (comparison within controls and with mild HSR) - see ordination top left. *Right*:- Fully linked network (low threshold=0.16) showing complete haplotype joinings of relevance to carbamazepine HSR. Squares represent group heterogeneity markers, circles between group markers. Covariance threshold set to achieve no orphan nodes. Note SNP rs111571306 (ICOS) highlighted is a haplotype 'hub' marker (in fact, a 'lynch pin' - see text) between the, encircled ICOS-dominated and the ensquared CTLA4-dominated, haplotypes leading towards pre-disposition to severe HSR and mild HSR respectively). *Note*: Relative loading positions, permutation p values and PAJEK network topology are invariant under addition of linear dummy severity variables - mild and severe (which are included for ease of interpretation). Such simply linearly translate the scores and alter the absolute magnitude of eigenvalues.

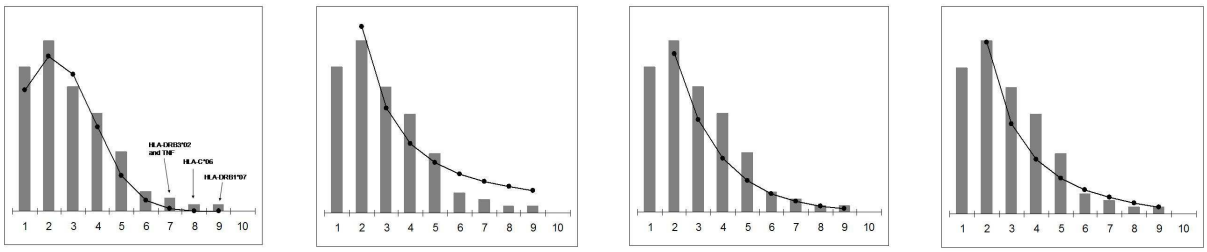

Figure 2: Nodal degree histograms for drug-induced SJS/TEN haplotype network in Bowman and Delrieu (2009b) (given no orphan nodes). Top 14 markers with (X) edges are:- HLA-DRB1*07 (9), HLA-C*06 (8), HLA-DRB3*02 (7), TNF (7), BAT1 (6), HLA-B*57 (6), HLA-B*47 (5), HLA-C*03 (5), HLA-DQB1*03 (5), HLA-DRB3*03 (5), HLA-DRB5*01 (5), HSPA1L (5), HSPA1S (5), and MICA (5). *Far Left*:- Normal distribution fit. Mean location around 2 = popular 'chaining' of markers randomly to one neighbour either side. Note consistent excess of 'hub' markers caused by presence of extended haplotypes giving dense 'local' relations. *Middle Left*:- Power law fit (for >2 edges) showing how any apparent 'hub' excess here is mild compared to scale-free networks. *Middle Right*:- Exponential fit (for >2 edges) showing that memory-less stochastic behaviour of links is seen within the dense network areas 'tail'. This could be taken to infer an *actual* excess of 3-5 edge nodes in this trait. *Far Right*:- Power law fit with vertical offset (for >2 edges) showing that the observed is still lower than the theory of scale-free topology. Again note possible inference of an *actual* excess of 3-5 edge nodes for this trait - a measure of the pre-dominance of particular types of LD patterns pertaining to the trait (contrast).

with each other to influence the phenotype. It is an initial epistasis relationship network i.e. the first (major) latent pattern of epistasis with respect to the trait. The edges represent the 'more bang for your buck' in influencing the phenotype. Making the colour of the edges or the size or 2-D position of the interaction nodes = the magnitude of the (case-control projected) loading 1 from a standard 'non-interaction' single marker analysis (see Delrieu and Bowman (2006a)) simultaneously visualises the absolute importance of the marker interaction *constituent* nodes in the first (major) latent set of single markers influencing the phenotypic trait (contrast of interest). Further eigenvectors yield cascading diminishing sets of minor epistatic relations. One can trim the networks back to the original single marker level only 'RedBox' (see Delrieu and Bowman (2006a)) if epistatic marker interactions only amongst the 'important' (first major latent set of) single markers are required (i.e. if there is a desire for 'model' marginality). Aggregation of markers into genes offers the opportunity to visualise extended gene 'haplotypes' and epistatic networks in a similar fashion (not illustrated). The distinction and role of interaction and residual analysis is summarised in Figure 4.

Acknowledgment: Allen Roses, whose vision forged the way for us to develop these ideas - that have been kindly fostered for many years by Kanti Mardia. Eunice Zhang (Univ. of Liverpool) for use of her data. Sir David Wallace for the gift of Skillicorn. Peter Grindrod for various discussions.

References

Alfirevic A, Vilar F J, Alsbou M, Jawaid A, Thomson W, Ollier W E R, Bowman C E, Delrieu O, Park B K and Pirmohamed M (2009) TNF, LTA, HSPA1L and HLA-DR gene polymorphisms in HIV positive patients with hypersensitivity to co-trimoxazole *Pharmacogenomics* (in press)

Barabási A-L (2002) *Linked. The new science of networks* Perseus Publishing

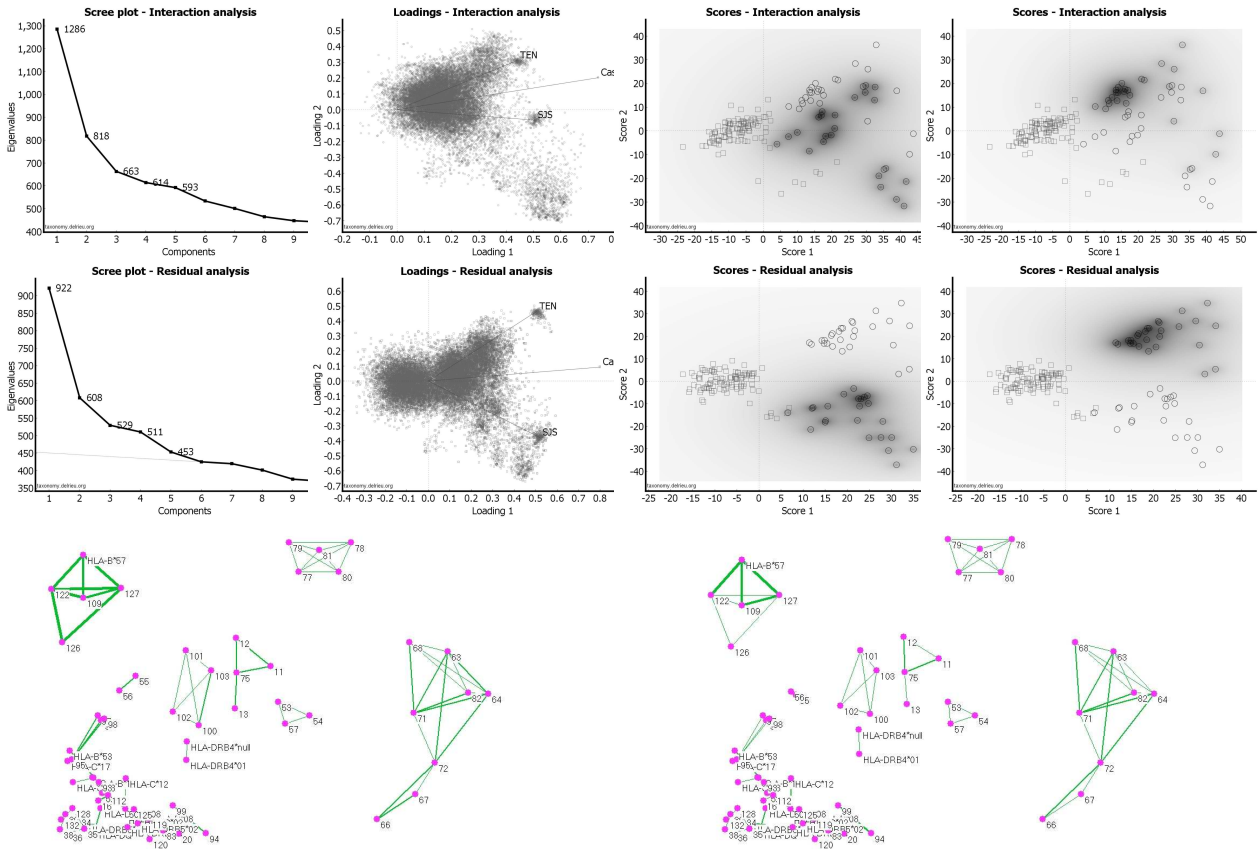

Figure 3: Drug-induced SJS/TEN (see Pirmohamed *et al.* (2007)) ordination (50 cases, 112 controls, 132 SNPs + 82 HLA alleles). *UPPER*: SNP level interaction and *MIDDLE*: SNP level residual analysis, including dummy disease indicator variables. Showing (left to right):- eigenvalue scree-plot; loadings plot annotated by direction of TEN (upper diagonal) and SJS (lower diagonal); and, heat maps for SJS (left plot) and TEN (right plot) overlain on score plots. *Note*: Small group of controls that look like cryptic future SJS cases (squares intercalated in SJS group); Two possible bi-marker driven extended haplotype groups amongst SJS subjects ; yet, probably only one epistatic SJS group. *LOWER*: Walk laplacian network (at correlation threshold $\geq$0.7 to control number of nodes displayed). On left - SNP level interaction analysis; on right - SNP level residual analysis. In both loading1 threshold $\geq$0.5 so as to highlight only important relations. Nodes arranged in absolute 2-D space according to a single marker SNP/HLA ordination (not shown). Diagonally up the page indicates greater relevance to the trait, *now* vertically = SJS predisposition, horizontally = TEN predisposition. Boldness of line (edge or arc) indicates strength of extended haplotype or epistasis between nodes of relevance to the phenotype contrast. SNPs 122, 126, 127 are in TNF, 109 in MICA. *Note*: Strong epistasis between a subset of elements in the extended haplotype around HLA-B*57. The extended haplotype and epistatic marker group to the right hand side contains SNPs in IL1F5 and IL1F8 but is dominated by those in IL1F10 (SNPs 62-68). Whilst carriage of IL1F7 markers (SNPs 77-81) might be individually of relevance to the skin phenotype overall, there is only a weak extended haplotype and no major epistasis between its constituents of relevance to the trait.

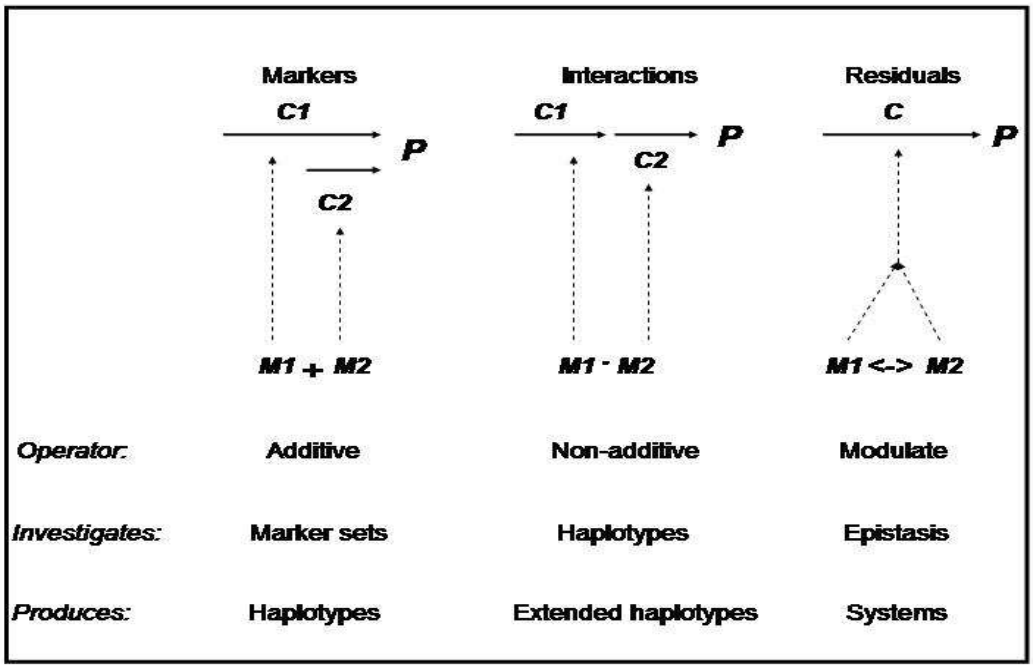

Figure 4: Overall scheme of SNP analysis. P = phenotype. C, C1, C2 = clinical pathways. *Left*: Over domain of individual markers - after Delrieu and Bowman (2005), Delrieu and Bowman (2006a) etc. *Middle*: Over domain of bi-marker haplotypes - after Delrieu and Bowman (2006b). *Right*: Over pairs of markers - after *this paper*.

Bowman C E (2009) Megavariate genetics: What you find is what you go looking for. *19th Altenberg Workshop of Theoretical Biology. Measuring Biology - Quantitative methods: Past and Future.* (eds Bookstein F and K Schaeffer) Konrad Lorenz Institute, Austria (in press)

Bowman C and Delrieu O (2009a) Immunogenetics of drug-induced 'blistering' disorders. I - Perspective *Pharmacogeneomics* (in press)

Bowman C and Delrieu O (2009b) Immunogenetics of drug-induced 'blistering' disorders. II - Synthesis *Pharmacogeneomics* (in press)

Bowman C, Delrieu O and Roger J (2006) Filtering pharmacogenetic signals In: S Barber, P D Baxter, K V Mardia and R E Walls *(Eds) Interdisciplinary Statistics and Bioinformatics.* University of Leeds, 184pp. 41-47

Charalambous C, Delrieu O and Bowman C (2008) Whole genome scan algebra and smoothing. In: Barber, S, Baxter P D, Gusnanto, A and Mardia, K V (eds) *The Art and Science of Statistical Bioinformatics.* Univ. of Leeds 21-27

Delrieu O and Bowman C (2005) Visualisation of gene and pathway determinants of disease In: S Barber, P D Baxter, K V, Mardia and R E Walls *(Eds.), Quantitative Biology, Shape Analysis, and Wavelets*. University of Leeds, 180pp. 21-24

Delrieu O and Bowman C (2006a) Visualising gene determinants of disease in drug discovery *Pharmacogenomics* 7, (3), 311-329

Delrieu O and Bowman C (2006b) Visualisation of gene by gene interactions in pharmacogenetics *International Congress Of Human Genetics, Brisbane Australia, 6-11th August 2006* (poster)

Delrieu O and Bowman C (2007) On using the correlations of divergences In: S Barber, P D Baxter, and Mardia, K V *(Eds), Systems Biology and Statistical Bioinformatics*. University of Leeds, 144pp. 27-35

Jardine N and Sibson R (1971) *Mathematical Taxonomy* John Wiley

Pirmohamed M, Arbuckle J, Bowman C, Brunner M, Burns D, Delrieu O, Dix L, Twomey J and Stern R (2007) Investigation into the multi-dimensional genetic basis of drug-induced Stevens-Johnson syndrome and toxic epidermal necrolysis *Pharmacogenomics* 8 (12), 1661-1691

Skillicorn D (2007) *Understanding complex datasets* Chapman and Hall/CRC