Statistical analysis of metabolomic data

Jim McNicol, Susan R. Verrall, Tom Shepherd, Gary Dobson, D. Wynne Griffiths, Gavin Ramsay, Howard V. Davies & Derek Stewart

Metabolic compositional profiles, consisting of hundreds of compound intensities, are increasingly used at SCRI to characterise samples of different genotypic and environmental backgrounds. We illustrate two exploratory statistical analyses using profiles consisting of 78 GC-MS polar and 52 non-polar compounds from 156 potato tuber samples. The mother plants, grown from true seed in a glasshouse, represent the four main cultivated groups of potato within the broad definition of *Solanum tuberosum*, Andigena, Phureja, Stenotomum and Chilean Tuberosum, with 78, 43, 24 and 11 accessions, respectively.

The first approach is to identify the main sources of variation using all 130 compounds. Principal components (PCs) achieve this by partitioning all the variation into uncorrelated variables, the first few of which often summarise broad effects. Interpretation of the principal components is achieved through 'loadings', the relative contributions of each metabolite, and 'scores', the sample values on the components.

Loadings show that the first PC describes total metabolite content and Fig. 1 (a) shows that this is negatively correlated with % dry matter content for each group

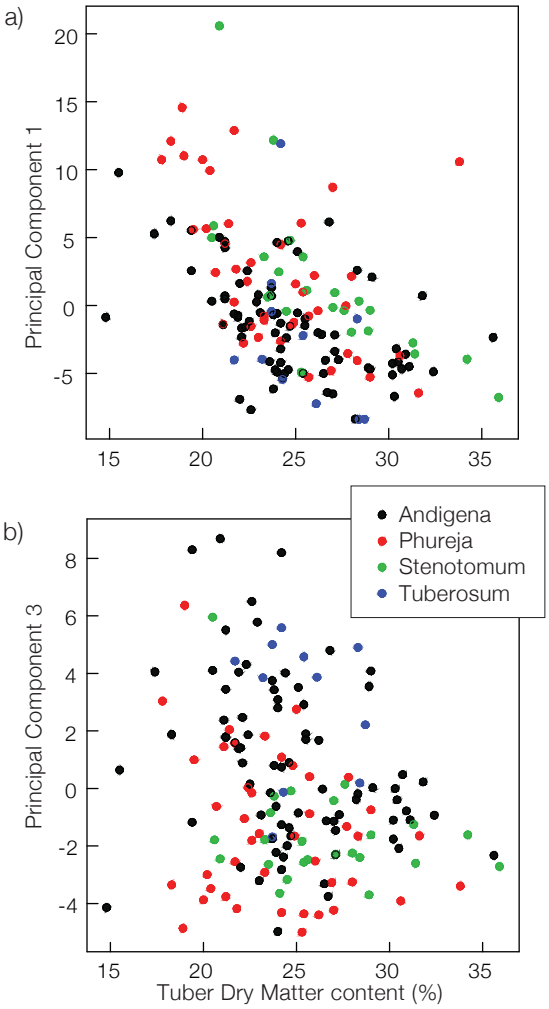

Figure 1 Relationship between principal components and % dry matter (a) PC1 – Total metabolite content (b) PC3 – sugar content.

except Tuberosum. The third PC is dominated by fructose, glucose and sucrose and Fig. 1 (b) suggests that

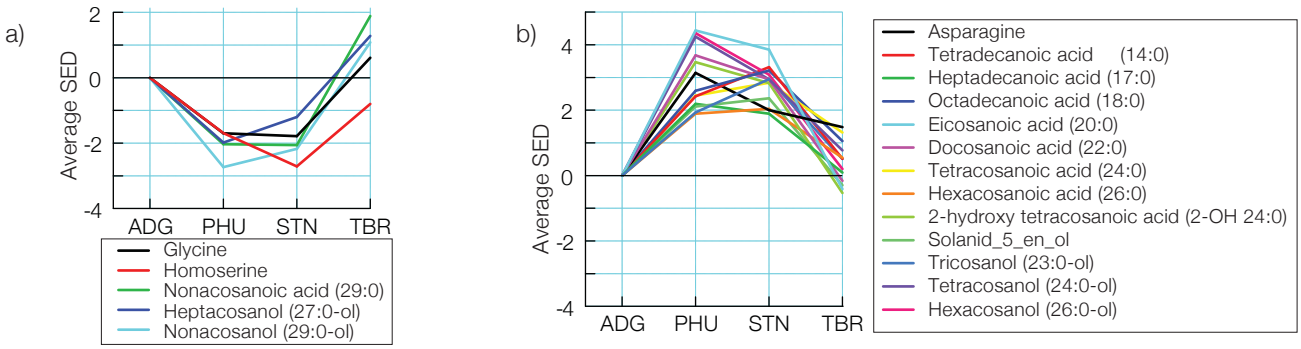

Figure 2 Accumulated expression clusters
(a) predominantly long chain odd-carbon fatty acids and alcohols (b) predominantly even-carbon fatty acid and alcohol homologues.

levels of these sugars are relatively high for Tuberosum and low for Stenotomum.

The second approach identifies individual metabolites which are differentially accumulated among the groups. This is achieved by analysis of variance of each metabolite separately. A significance cut-off corresponding to a false discovery rate of 2% identified 59 metabolites which were accumulated differentially. Hierarchical clustering, based on pairwise standard errors of difference between groups, partitioned these metabolites into groups of similar significance patterns across the four groups.

Three main groups of biosynthetically-related, non-polar metabolites (composed of saturated fatty acids and fatty alcohols) were identified, differing predominantly in the length of their carbon chains and in the presence of chains with odd and even numbers of carbon atoms. It is also of note that unsaturated fatty acids, the major constituents within the non-polar metabolites, do not appear in any of the clusters and therefore do not show any significant inter-group variation. This could reflect differences in the specificity of the enzyme systems responsible for synthesis of long carbon chains, similar to those observed for various leaf lipids (Shepherd, 2003; Shepherd & Griffiths, 2006). Fig. 2 (a) shows the group which includes long chain odd-carbon fatty acids and alcohols, whereas in Fig. 2 (b) the cluster consists mainly of even-carbon fatty acid and alcohol homologues. The different patterns probably reflect the existence of parallel pathways for synthesis of odd and even carbon components, and a shift in the partitioning of precursors between the pathways in Phureja and Stenotomum in comparison with Tuberosum and Andigena.

References

Shepherd, T. 2003. Wax pathways. In: Thomas, B., Murphy, D. & Murray, B. (eds). *Encyclopaedia of Applied Plant Sciences*, Academic Press, London, 1204-1225.

Shepherd, T. & Griffiths D. W. 2006. The effects of stress on plant cuticular waxes. *New Phytologist* **171**, 469-499.