Genes, Genomes and Genomics ©2010 Global Science Books

A Multivariate Approach to the Proteomics of Tomato Fruit Ripening

Nahuel N. Pratta[1] • Marta Quaglino[2] • Gustavo R. Rodríguez[3] • Roxana Zorzoli[3,4] • Guillermo R. Pratta[3,5*]

[1] Graduate Dissertant from Licenciatura en Estadística, Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario, Argentina
[2] Cátedra de Análisis Multivariado, Escuela de Estadística, Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario, Argentina
[3] Cátedra de Genética, Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Argentina
[4] Consejo de Investigaciones de la Universidad Nacional de Rosario, Argentina
[5] Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina

Corresponding author: * **gpratta@unr.edu.ar**

ABSTRACT

Structural and functional genomics are useful approaches to better understand the biological process regarding the genetic composition of individuals and its expression at the phenotypic level. Numerous data are produced in studies of this area of research, multivariate statistics providing powerful tools for reducing their dimensionality. The general objective of this research was to apply principal component analysis (PCA) and multivariate analysis of variance (MANOVA) to analyse the proteomics of the tomato ripening and its association with agronomic quantitative traits. A PCA was applied as a non classic exploratory way to the binary data presence / absence of a given polypeptide in mature green, breaker and red ripe tomato fruits from F_2 plants. To verify results from PCA, a hierarquical cluster analysis was performed by UPGMA to the same data. Advantage of PCA is that the relative contribution of a given polypeptide for grouping can be measured. Then, polypeptides mostly contributing for grouping were introduced as the source of variation in a one-way MANOVA to detect association with quantitative traits, which were the dependent variables. Polypeptides mostly contributing to general variability were efficiently identified by PCA, while MANOVA verified that one of them is a putative molecular marker of the quantitative traits.

Keywords: biometrics, clusters, multivariate analysis of variance, plant breeding, principal components
Abbreviations: BR, breaker stage of tomato ripening; **DNA**, deoxi-ribonucleic acid; **HCA**, hierarquical cluster analysis; **MANOVA**, multivariate analysis of variance; **MG**, mature green stage of tomato ripening; **PC**, principal component; **PCA**, principal component analysis; **QTL**, quantitative trait loci, **RR**, red ripe stage of tomato ripening; **SDS-PAGE**, sodium dodecyl sulphate polyacrylamide gel electroforesis; **UPGMA**, unweighted pair group method with arithmetic mean

INTRODUCTION

The outstanding advances in the knowledge of gene structure and function in last decades allowed plant breeders to work with a higher level of accuracy. Genomics, transcriptomics, proteomics and metabolomics, help to better understand the biological process taking place between the genes and the phenotypes (Skalnikova *et al.* 2007). In this sense, molecular markers are useful tools to classify and manage germplasm in breeding programs (Rocco *et al.* 2006). Molecular markers based on polypeptides fractioned by SDS-PAGE are simple and effective to distinguish among genotypes. Though they are less polymorphic than DNA markers, they could provide quick information about molecular polymorphism related to the trait under study given than they vary according to the source of sample for extraction. Numerous data are produced in this area of research. Hence, a reduction in their dimensionality conserving a maximum of information is required. Multivariate statistics provides powerful tools for the analysis. Though principal component analysis (PCA) is a common technique used for quantitative data, some authors from many fields of biology (Roldán-Ruiz *et al.* 2000; Figueiredo Goulart *et al.* 2005; Du *et al.* 2006; Helánová *et al.* 2006) applied PCA for frequency or binary data ranging between 0 – 1, related to presence / absence of a given object in a given situation, and were able to identify the objects mostly contributing to the general variability. Therefore, association between molecular markers and quantitative traits has been mainly achieved by univariate analysis, such as single point analysis, in which one-way ANOVA is applied to detect a significant effect of the presence / absence of a given marker on the quantitative trait under analysis (Collard *et al.* 2005). Though this approach proved to be robust, it would be interesting to apply MANOVA for detecting quantitative trait loci (QTL), taking the markers mostly contributing to general variability identified by PCA as source of variation, and the quantitative traits as dependent variables. In this way, a global understanding of the molecular basis underlying quantitative traits would be gained (Brewer *et al.* 2007).

The cultivated tomato (*Solanum lycopersicum*, n = 12) is a dicotyledonous species, member of the small section *Lycopersicon* of the *Solanaceae* family. This section comprises 12 species having high degree of homosequenciality. Tomato is a selfing crop in which fruit quality is very important because of the great levels of human consumption. Fruit quality is mainly determined by the ripening stage at harvesting. Like in other climacteric fruits, carotenoids accumulation and softening allow distinguishing different ripening stages of the tomato fruit: mature green, breaker, and red ripe (Giovannonni 2004). Mature green (MG) is that stage in which fruit reached its maximum size but no variation in colour has still produced, at the breaker stage (BR) the 10–30% of fruit area has turned to pink, and at the red ripe stage (RR), more than 90% of fruit area has acquired red colour. A good indicator of changes occurring along fruit ripening is the protein expression (Faurobert *et*

Received: 26 June, 2010. Accepted: 30 July, 2010.

Research Note

al. 2007). A broad range of genetic variability exists for protein expression since in addition to differences among stages in standard genotypes, genetic variants (including mutants for ripening and exotic germplasms) are available in the tomato genome (van Berloo *et al.* 2008). Among mutants, the homozygous *nor* gene produces non climacteric fruits that ripe slowly or did not ripe at all, acquiring just a pale pink colour. Transition of chloroplast to chromoplast is delayed, polygalacturonase is absent and appropriate colours, flavour, savour, texture and taste are not developed in these fruits. Given that those unfavourable pleiotropic effects are maintained in the heterozygous genotypes, fruits carrying the *nor* mutant are scarcely commercialized. Among exotic relatives, *S. lycopersicum* var. *cerasiforme* genetic pool is an admixture between the cultivated tomato and wild species, and some genotypes have many agronomical important traits including diseases and pest resistances, high nourishing quality, and adverse environments tolerance.

Genetic variability is crucial in breeding program, and crosses to exotic germplasms were made by our research group to broad the tomato gene pool and to obtain new tomato genotypes (Rodriguez *et al.* 2006). At the molecular level, Pratta *et al*. (2001) and Pereira da Costa *et al.* (2009) detected differences in polypeptide profiles of the pericarp tissue at two ripening stages in different tomato genotypes including standard and mutant for ripening genotypes as well as exotic germplasm. Rodriguez *et al.* (2008) found significant cosegregations between agronomical traits related to fruit quality and some polypeptides of the pericarp tissue at MG, BR and RR. The general objectives of the present report were to identify the polypeptides mostly contributing to general variability of the proteomics of the tomato fruit ripening and to detect associations among these polypeptides and some quantitative traits by means of multivariate analyses.

MATERIALS AND METHODS

Data analyzed in this research were extracted from Rodriguez *et al.* (2008) but a univariate approach was applied there. In contrast a multivariate analysis is presented here. Briefly, field assays were carried out at the horticultural section of the experimental station "José F. Villarino" and the "Plant molecular biology and *in vitro* culture laboratory" of the Agronomy Faculty - UNR, located in Zavalla (Santa Fe, Argentina, 31 SL). The F_1 generation among a homozygote *nor* genotype of *S. lycopersicum* (accession 80462) and accession LA1385 of *S. lycopersicum* var. *cerasiforme* was obtained by hand emmasculation and pollination. The F_2 generation was obtained by selfing F_1 plants. Sixty F_2 plants were randomly chosen for quantitative and molecular assessments. Total proteins of the pericarp tissue were extracted from fruits at MG, BR and RR. Polypeptides were fractioned by SDS-PAGE, and polymorphisms among genotypes were recorded as the presence/absence of polypeptides having different molecular mass (in kDa) by stage (Rodriguez *et al.* 2008). Multivariate analyses of PCA and Clustering were applied to characterize the polypeptide polymorphism in the segregating F_2 population. Also, six quantitative traits were evaluated: days from transplant to anthesis (DF), days from anthesis to MG (DMG), days from MG to BR (DBR), days from BR to RR (DRR), fruit shelf life (SL, days from harvest at breaker to the beginning of fruit softening), and fruit mass at harvest (M, in g). Means and standard deviations were calculated, and association among polymorphic polypeptides and quantitative data were estimated by MANOVA. The Hotelling-T^2 test was used to detect differences among mean values of groups of F_2 plants defined by the presence / absence of polypeptides. Statistical software used was Infogen, developed at Universidad Nacional de Córdoba (Argentina).

RESULTS AND DISCUSSION

A PCA was applied to the binary data presence/absence of polypeptides among F_2 plants by assigning 1 to the presence and 0 to the absence. The goal of this analysis was to identify the polypeptides mostly contributing to the general variability of each data set, being those ones that had a correlation coefficient with PC1 and PC2 higher than 0.50 in absolute value. Sign was not considered because it is related to the presence/absence of a given polypeptide, being both alternatives equally important in determining the polymorphism. For polypeptides at BR, the eigenvalues (proportion of total variability explained by each principal component) were 0.14 for PC1 and 0.14 for PC2, a total of 20 PC being obtained. In this report just the first two PC were retained for further assessment in order to visualize the structure of the data set in just two dimensions. Polypeptides were identified by their molecular mass, being those of 30, 34, 45, 47, 97, and 105 kDa (which had the highest correlation with PC1) and those of 32, 80, 88, and 99 kDa (which had the highest correlation with PC2) the ones that mostly contributed to the general variability. For polypeptides at MG and RR, the eigenvalues were 0.19 for PC1 and 0.15 for PC2, and 0.16 for PC1 and 0.13 for PC2, respectively. At MG, a total of 18 PC were obtained while at RR, this number was of 20 different PC. Correlations of polypeptides at MG pointed that those of 59, 65, 69, 73, 82, and 88 kDa highly correlated to PC1 and those of 88, 97, and 99 kDa highly correlated to PC2. Correlations of polypeptides at RR pointed that those of 69, 72, and 88 kDa highly correlated to PC1 and those of 30, 45, 47, and 97 kDa, highly correlated to PC2.

PCA was also used to group the F_2 plants according to the polypeptide characterization. To verify the ability of PCA for grouping, a hierarquical cluster analysis (HCA) was performed by UPGMA method with Jaccard distances among F_2 plants. Since this method is the most commonly used for grouping, associations of plants obtained by PCA were compared to those obtained by HCA. **Fig. 1** shows, as example, the grouping of F_2 plants by the presence / absence of different polypeptides in the pericarp tissue at MG, in the space defined by PC1 and PC2. Three groups of genotypes were determined by PCA at this stage, though it is important to note here that circles dividing groups where hand-drawn once the consistency of each group was verified by HCA. In fact, the dendrogram with Jaccard distances and UPGMA method verified the consistency of groups obtained by PCA (**Fig. 2**). Both multivariate methods resulted in groupings essentially similar. When considering a Jaccard distance of 0.50, HCA also showed 3 groups, their composition of F_2 plants being in agreement to those obtained by PCA. Just a few plants were exceptions, like 3, 53, 59, and 98. Grouping of F_2 plants at MG and RR are not shown, but 5 groups were obtained at BR and 4 at RR by PCA. HCA at these stages also resulted in groupings very similar to those obtained by PCA. Plants 11, 85, and 87 at MG, and 94 at RR were the exceptions.

Means and standard deviations of quantitative traits evaluated in the F_2 tomato generation are in **Table 1**. Then, the F_2 plants were divided in two groups regarding the presence / absence of the polymorphic polypeptides mostly contributing to general variability identified by PCA, and MANOVA was applied to visualize the effect of each polypeptide on the multidimensional space defined by the quantitative traits as a whole. Hence we attempted to detect QTLs by means of association among molecular data and phenotypic traits. Just a polypeptide of 97 kDa (in italics in **Fig. 1**) at the breaker stage had a significant effect when applying this model (**Table 1**), thus indicating a pleiotropic effect of BR97 on all quantitative traits. Mean values of quantitative traits among the group of F_2 plants defined by the presence / absence of BR97 were significant at 5% according to the Hotelling-T^2 test.

We demonstrated here that the complex structure of binary data such as the polypeptides expressing during tomato fruit ripening was simplified by PCA without following any *a priori* model of its distribution. Though PCA was developed to analyse quantitative data, there are antecedents regarding its use for binary date, especially in biological studies. For instance, Roldán-Ruiz *et al.* (2000)

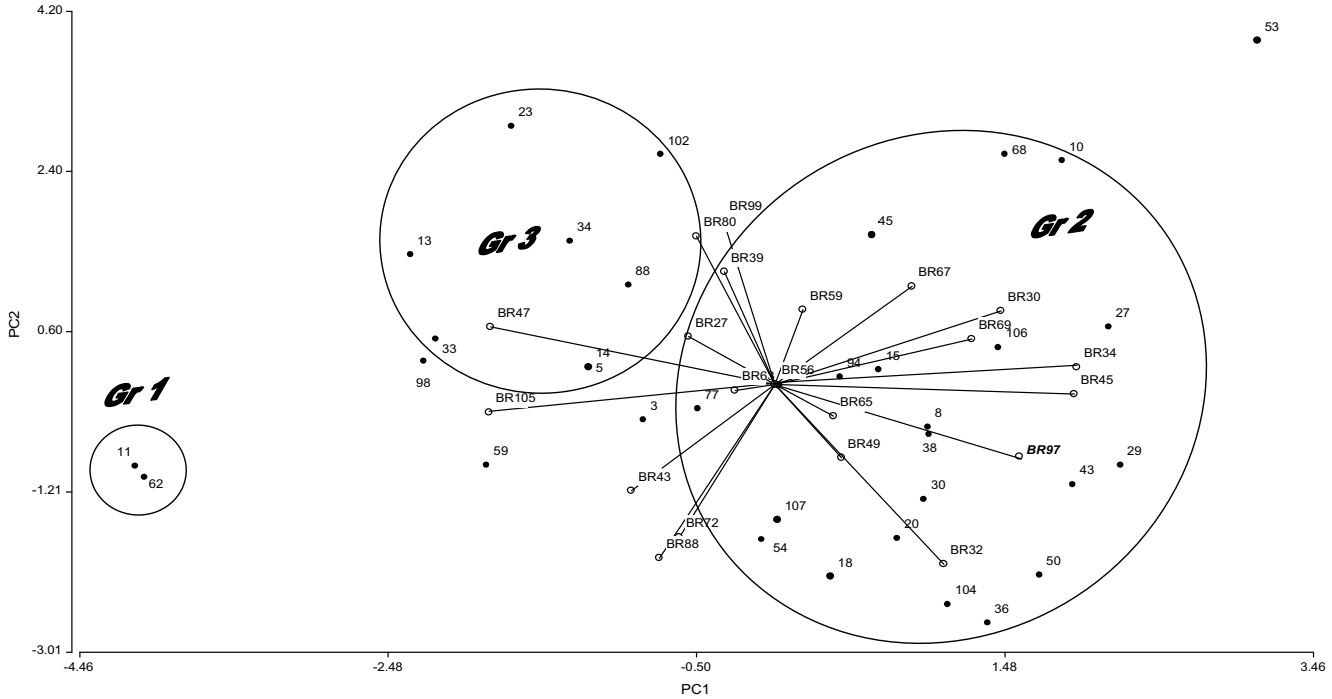

Fig. 1 Grouping of F_2 plants (dotted circles) in the plane defined by Principal Component 1 (PC1) and Principal Component 2 (PC2) from presence / absence of polypeptides (empty circles) at the breaker stage.

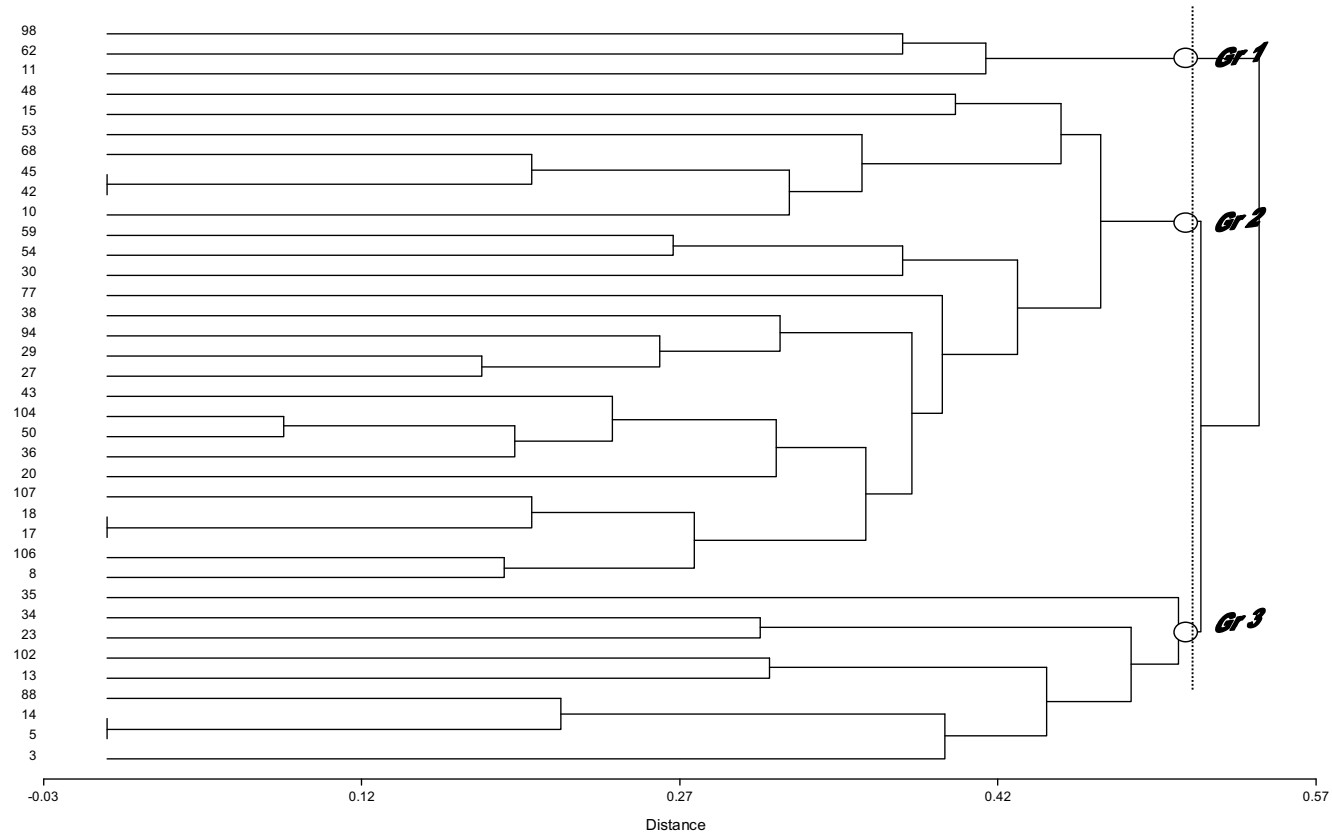

Fig. 2 Grouping of F_2 plants by Cluster Analysis with Jaccard distance calculated from presence / absence of polypeptides at the breaker stage.

applied PCA for characterizing rye-grass genotypes by the relative frequency of DNA fragments among populations, Figueiredo Goulart *et al.* (2005) segregated different ecosystems according to the to presence / absence of plant species, Du *et al.* (2006) identified properties of a given protein by its amino acid composition, and Helánová *et al.* (2006) assessed the terpenes present in fresh needles of *Picea abies*. In all cases, a complex data set was evaluated weighting its components, and it was possible to identify such components mostly contributing to the general variability. Also, a given set of objects (genotypes, polypeptides, ecosystems) was classified in those studies.

In the present study, it was possible to identify the polypeptides mostly contributing to the general variability in different stages of tomato fruit ripening. Also, F_2 plants were classified according to the presence / absence of the polypeptides. Both achievements are relevant to plant genetics and breeding, given that the most variable polypep-

Table 1 Means and standard deviations of the quantitative traits evaluated in the F_2 generation and means by groups of F_2 plants defined by the presence (1) and absence (0) of the polypeptide of 97 kDa detected at the breaker stage (BR97).

Means	DF	DMG	DBR	DRR	SL	M
F_2 generation	73.16 ± 21.29	30.63 ± 4.10	11.66 ± 4.62	7.27 ± 3.84	34.97 ± 14.04	5.61 ± 2.62
Group of F_2 plants (1)	67.45 ± 11.08	31.70 ± 1.69	10.63 ± 2.54	6.95 ± 0.57	32.26 ± 3.14	5.20 ± 0.32
Group of F_2 plants (0)	89.84 ± 8.34	25.84 ± 3.02	15.20 ± 1.43	9.60 ± 1.04	37.52 ± 1.92	7.92 ± 1.56

DF: Days from transplant to anthesis.
DMG: days from anthesis to mature green stage.
DT: days from mature green to breaker stages.
DRR: days from breaker to red ripe stages.
SL: fruit shelf life, in days from harvest at breaker stage to the beginning of fruit softening.
M: fruit mass at harvest, in grams.

tide could be used as molecular markers of agronomic traits, and classifying individuals within segregant generations would improve their management in field assays. From a biological viewpoint, the detection of an early expressed polypeptide (that of 97 kDa at BR) having pleiotropic effects on several fruit traits measured at a later stage, is important to better understanding the whole process of tomato ripening. From an agronomic viewpoint, polypeptides previously reported by Rodriguez *et al.* (2008) as molecular markers of productive quantitative traits by using univariate methods were the same that appeared as most variable in this paper. Hence, an additional advantage of the multivariate approach would be to minimize the number of univariate tests needed to identify molecular markers by the traditional statistical methods.

REFERENCES

Brewer MT, Moyseenko JB, Monforte AJ, van der Knaap E (2007) Morphological variation in tomato: a comprehensive study of quantitative trait loci controlling fruit shape and development. *Journal of Experimental Botany* **58**, 1339-1349

Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica* **142**, 169-196

Du QS, Jiang ZQ, He WZ, Li DP, Chou KC (2006) Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction. *Journal of Biomolecular Structure and Dynamics* **23**, 635-640

Faurobert M, Mihr C, Bertin N, Pawlowski T, Negroni L, Sommerer N, Causse M (2007) Major proteome variations associated with cherry tomato pericarp development and ripening. *Plant Physiology* **143**, 1327-1346

Goulart MF, Pontes Ribeiro S, Lovato MB (2005) Genetic, morphological and spatial characterization of two populations of *Mabea fistulifera* Mart. (Euphorbiaceae), in different successional stages. *Brazilian Archives of Biology and Technology* **48**, 275-284

Giovannoni JJ (2004) Genetic regulation of fruit development and ripening. *The Plant Cell* **16**, S160-S170

Helánová V, Chvílíèková I, Martinková M, Meloun M, Kubáò V (2006) Application of multivariate statistical analysis to biological data. Variations of monoterpene content in fresh needles of *Picea abies* (L.) Karst. *Chemical Analysis (Warsaw)* **51**, 551-565

Pereira da Costa JH, Rodríguez GR, Pratta GR, Zorzoli R, Picardi LA (2009) Characterization of tomato germoplasm by pericarp protein profiles and morphologic and biochemical fruit traits. *Fruit, Vegetable and Cereal Science and Biotechnology* **3**, 48-53

Pratta G, Valle EM, Zorzoli R, Carrillo N, Picardi LA (2001) Characterization of tomato genotypes that differ in their fruit shelf life by analysis of total pericarp protein patterns at two ripening stages. *Acta Horticulturae* **546**, 483-487

Rocco M, D'Ambrosio C, Arena S, Faurobert M, Scaloni A, Marra M (2006) Proteomic analysis of tomato fruits from two ecotypes during ripening. *Proteomics* **6**, 3781-3791

Rodriguez G, Pratta G, Zorzoli R, Picardi LA (2006) Recombinant lines obtained from an interspecific cross between *Lycopersicon* species selected by fruit weight and fruit shelf life. *Journal of the American Society for Horticultural Science* **131**, 651-656

Rodriguez GR, Sequin L, Pratta GR, Zorzoli R, Picardi LA (2008) Protein profiling in F_1 and F_2 generations of two tomato genotypes differing in ripening time. *Biologia Plantarum* **52**, 548-552

Roldán-Ruiz I, Calsyn E, Gilliland TJ, Coll R, van Eijk MJT, De Loose M (2000) Estimating genetic conformity between related ryegrass (*Lolium*) varieties. 2. AFLP characterization. *Molecular Breeding* **6**, 593-602

Skalnikova H, Halada P, Vodicka P, Motlik J, Rehulka P, Hørning O, Chmelik J, Nørregaard JO, Kovarova H (2007) A proteomic approach to studying the differentiation of neural stem cells. *Proteomics* **7**, 1825-1838

van Berloo R, Zhu A, Ursem R, Verbakel H, Gort G, van Eeuwijk FA (2008) Diversity and linkage disequilibrium analysis within a selected set of cultivated tomatoes. *Theoretical and Applied Genetics* **117**, 89-101