桉树杂交种与其亲本的近红外光谱判别

卢万鸿，李 鹏*，王楚彪，林 彦，罗建中

国家林业和草原局桉树研究开发中心，广东 湛江 524022

摘 要 研究桉树控制授粉后目标性状的基因作用方式是探索其基因重组规律的重要内容。常规的数量统计分析精度往往不高，而DNA分析的专业要求高，且费时费力。该研究利用近红外光谱(NIRs)研究不同基因型桉树杂交种、亲本及杂交种与亲本间近红外光谱信息的关系，探索NIRs用于桉树杂交种与其亲本判别的可行性和准确性。以控制授粉的桉树亲本及其杂交F1代材料为对象，每种基因型从各自田间试验分别选取10个单株，采集树冠中上部新鲜健康叶片。用手持式近红外仪Phazir Rx(1624)采集桉树杂交种与其亲本叶片的NIRs信息。每单株选10片完全生理成熟的健康叶片，避开叶脉扫描其正面光谱5次，以50条NIRs信息的均值代表单个叶片的NIRs信息，最终每个基因型获得10条NIRs信息。对原始NIRs采用二阶多项式S. G一阶导数预处理。预处理后的NIRs用于多元统计分析，首先对桉树杂交亲本和子代样本进行主成分分析(PCA)，直观展示不同基因型的分类情况。然后运用簇类独立软模式(SIMCA)和偏最小二乘判别分析(PLS-DA)两种有监督的判别模式验证NIRs用于桉树杂交种与其亲本树种的分类判别效果。PCA结果显示，不同的亲本间、杂交种间及杂交种与亲本间样本的主因子得分可以清晰地将各基因型分开。SIMCA模式判别分析中，桉树杂交种样本到亲本PCA模型的样本距离显示，待判别样本能够形成单独的聚类，且能直观反映两者的遗传相似。PLS-DA判别结果显示，桉树杂交亲本的PLS模型能通过预测其杂交子代的响应变量将其与亲本准确分开。结果表明，桉树叶片的NIRs信息可以准确地反映桉树杂交子代遗传信息的传递规律，NIRs判别模型可以准确地将各种基因型予以区分。因此，NIRs信息不仅可用于桉树杂交种和纯种的定性判别，还可以分析桉树基因重组过程中加性遗传效应的大小，从而为桉树遗传基础分析及其育种改良研究提供理论支撑。

关键词 有监督的模型；主成分分析(PCA)；簇类独立软模式(SIMCA)；偏最小二乘判别分析(PLS-DA)

中图分类号：S722.34 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)03-0873-05

引 言

研究桉树(桉属*Eucalyptus*、伞房属*Corymbia*和杯果木属*Angophora*)控制授粉后目标性状的基因作用方式，有助于分析亲本组合时的基因重组规律，为开发优良杂交组合的亲本选配提供理论支撑。常规的数量标记分析精度往往不高[1-2]，而分子生物学方法[3-4]又需要强的专业知识，且程序繁复，很难满足林木育种改良研究中对大量群体材料的快速分析。

化学计量学和光谱学的发展促进了近红外光谱(near infrared spectroscopy，NIRs)用于植物来源分类及其物质成分快速预测研究的繁荣。研究人员采集了伞房属(*Corymbia*)桉树纯种及其杂交种叶片的NIRs信息，建立的伞房属桉树NIRs模型的判别准确率为76%－90%[5]。采用人为混合松树(*Pinus*)松针模拟杂交种，建立了纯种松树的NIRs判别模型，其判别准确率也达到了90%以上[6]。伞房属桉树杂交子代与其亲本萜烯含量的差异非常小，但NIRs对这种极微小的差异反映很敏感，能检测到常规方法检测不到的微小差异[7]。关于玉米自交系遗传距离与其NIRs光谱距离间关系的研究显示，其光谱距离与其遗传距离间的相关性超过0.9，表明NIRs可以反映玉米自交系间的遗传关系。Diepeveen等的研究更具体定位了影响NIRs特征的含有遗传信息的小麦染色体片段[8]。

不同来源的物种在特定条件下内在的遗传物质差异从根本上决定了其组织成分的差异，这是NIRs用于物种成分预

收稿日期：2019-01-23，**修订日期**：2019-04-07

基金项目：国家自然科学基金青年项目(31700599)，国家自然科学基金面上项目(31670680)资助

作者简介：卢万鸿，1982年生，国家林业和草原局桉树研究开发中心副研究员 e-mail：luwanhong@outlook.com

*通讯联系人 e-mail：462227809@qq.com

测和判别分析的主要依据[9-11]。本研究以桉树控制授粉材料为对象，分析桉树叶片 NIRs 与其遗传基础间的关系，并用簇类独立软模式(soft independent modeling of class analog，SIMCA)和偏最小二乘判别分析(partial least squares discrimination analysis，PLS-DA)两种判别分析方法进行桉树杂交种及其亲本的分类判别，探索 NIRs 用于桉树杂交种与其亲本判别的可行性及准确性。

1 实验部分

1.1 材料

试验材料包括粗皮桉(*E. pellita*)、尾叶桉(*E. urophylla*)、细叶桉(*E. tereticonis*)3 个纯种及其 5 个杂交种，杂交种分别为 3 个亲本树种间的杂交 F1 代，外加一个目前商用的桉树无性系 DH3229(见表 1)。从田间试验林中采集各基因型的叶片用于 NIRs 的扫描。

表 1 测试桉树杂交组合及其亲本信息

Table 1 The details of the hybrids and their parents of eucalypt

分类	序号	编号	母本	亲本
杂交种	1	K50	*E. pellita* 粗皮桉	*E. urophylla* 尾叶桉
	2	EC126	*E. urophylla* 尾叶桉	*E. pellita* 粗皮桉
	3	EC148	*E. pellita* 粗皮桉	*E. urophylla*×*tereticonis* 尾细桉
	4	EC149	*E. tereticonis* 细叶桉	*E. urophylla* 尾叶桉
	5	U6	*E. urophylla* 尾叶桉	*E. tereticonis* 细叶桉
	6	DH3229	*E. urophylla* 尾叶桉	*E. grandis* 巨桉
亲本	1	P47	*E. pellita* 粗皮桉	
	2	U110	*E. urophylla* 尾叶桉	
	3	T0105	*E. tereticonis* 细叶桉	

1.2 仪器

手持式近红外仪 Phazir Rx (1624) (Polychromix，Thermo Scientific，USA)用于 NIRs 数据的采集。Phazir Rx (1624)波长范围为 1 600～2 400 nm，光学分辨率 12 nm，内置基于 MEMS 技术的可编程微衍射光栅，自带背景校正片。

1.3 方法

1.3.1 光谱采集

每种基因型分别选取 10 个单株，采集树冠中上部的新鲜健康叶片。每单株选 10 片完全生理成熟的健康叶片，用 Phazir Rx(1624)扫描其正面光谱共 5 次，以均值代表单个叶片的 NIRs 信息[10]。每种基因型获得 10 条 NIRs。

1.3.2 NIRs 数据的预处理和分析

对原始 NIRs 进行二阶多项式 S.G 一阶导数预处理[10-11]。通过主成分分析直观判断 NIRs 对桉树不同基因型的分类效果。建立簇类独立软模式(SIMCA)和偏最小二乘判别分析(PLS-DA)两种有监督方式的判别模型检验 NIRs 的树种判别效果。建立 PLS-DA 模型时，分别对 3 个亲本树种人为赋值，即 1，2 和 3[12]。数据预处理和分析过程均在 The Unscrambler x10.4(CAMO，Oslo，Norway)中实现，主成分分析(principal component analysis，PCA)和 PLS 过程均采用全交互式内部交叉验证算法。

2 结果与讨论

2.1 桉树杂交种叶片的 NIRs 信息的差异

NIRs 主要是物质有机分子的倍频与合频吸收光谱，不仅能得到物质的分子结构、组成和状态信息，也能反映密度、粒度、高分子物的聚合度及纤维形态等物质的物理状态信息[10]。图 1 是 6 个桉树杂交组合原始 NIRs 的平均值曲线，通过 NIRs 原始光谱的直观变化很难发现其特征峰，6 种组合的 NIRs 信息在全波段变化趋势基本一致，且存在明显的重叠。在波长 1 860 nm 之前和波长 1 940 nm 之后，6 个杂交组合的 NIRs 反射率在一定程度上存在差异，但不足以据此进行树种判别。

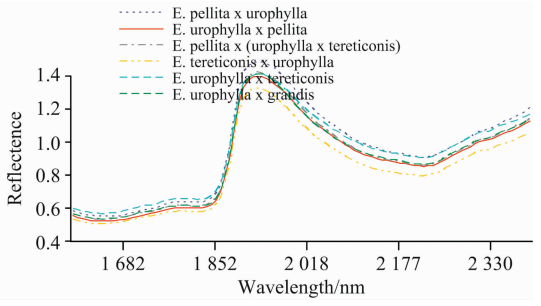

图 1 桉树杂交种的原始 NIRs 反射率

Fig. 1 The raw NIRs reflectance of eucalypt hybrids

2.2 桉树杂交种及其亲本的 PCA 聚类

PCA 可以简化多维数据中大量重叠的信息，因子得分可以反映受试样本间的距离关系。图 2 是 6 个桉树杂交种叶片 NIRs 数据的 PCA 因子得分图。杂交种 EC126，EC148 和 EC149 能清晰地聚类[图 2 上]，U6，K50 和 DH3229 也能清

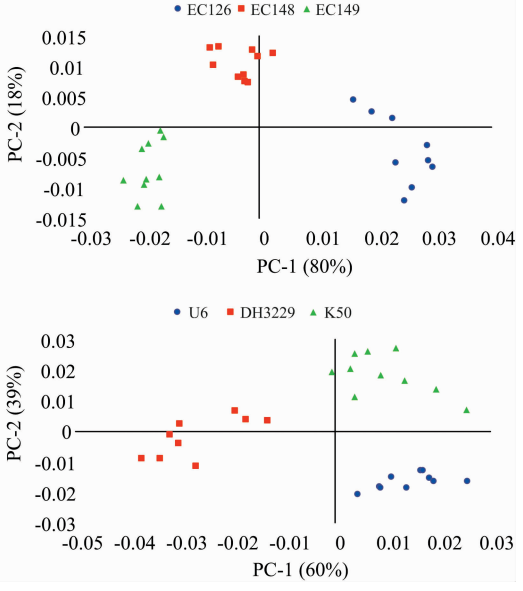

图 2 桉树杂交子代的 PCA 因子得分图

Fig. 2 The PCA scores plot of NIRs from eucalypt hybrid progenies

晰地分开[图 2 下]，表明 NIRs 能够区分不同的基因型，可对遗传差异做出响应。

将 6 个杂交种同时进行 PCA 分析时，各杂交种存在一定程度的重叠(未展示)，这主要与其亲本的遗传亲缘关系有很大的关系。另外，因子得分的聚集度也反映了杂交种遗传变异的大小，聚集度高的变异小，反之亦然，如图 2 中 EC126，K50 和 EC149 的变异可能较大，EC148 和 U6 的变异则相对较小。

图 3 是 3 个亲本 NIRs 数据 PCA 的因子得分图。从图中可以看出，粗皮桉相对较为分散，细叶桉聚集度最高，尾叶桉聚集度居中。从分类的聚集度来看，粗皮桉的变异最大，细叶桉的变异则最小。总体来看，因子得分图能够将 3 个亲本树种清晰地分开，真实地反映了不同树种内在的遗传差异。

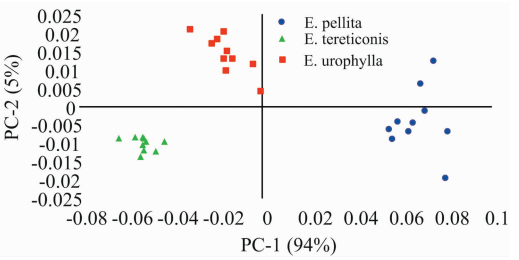

图 3　桉树杂交亲本的 PCA 因子得分图

Fig. 3　The PCA scores plot of NIRs from eucalypt cross parents

2.3　桉树亲本模型对其杂交种的 SIMCA 判别

用 3 个亲本样本建立 PCA 模型，设定临界概率水平为 0.05。图 4 为每组亲本 PCA 模型对杂交子代进行 SIMCA 判别的结果，结果以杂交种样本与亲本 PCA 模型中心间的距离表示。图 4 显示，6 个杂交种均可以与其亲本清晰地分开，其中杂交种 K50 与其母本粗皮桉的距离相对较近，EC126 与其父本粗皮桉间的距离较近，EC148 与其母本粗皮桉间的距离非常接近，EC149 距其父本尾叶桉较近，U6 基本居于其父本尾叶桉和母本细叶桉之间，商用杂交种 DH3229 基本居于尾叶桉和粗皮桉中间(未采集到巨桉样本)。SIMCA 模式判别中的样本距离更直观地反映了杂交种与其亲本样本间的遗传相似性(遗传距离)。

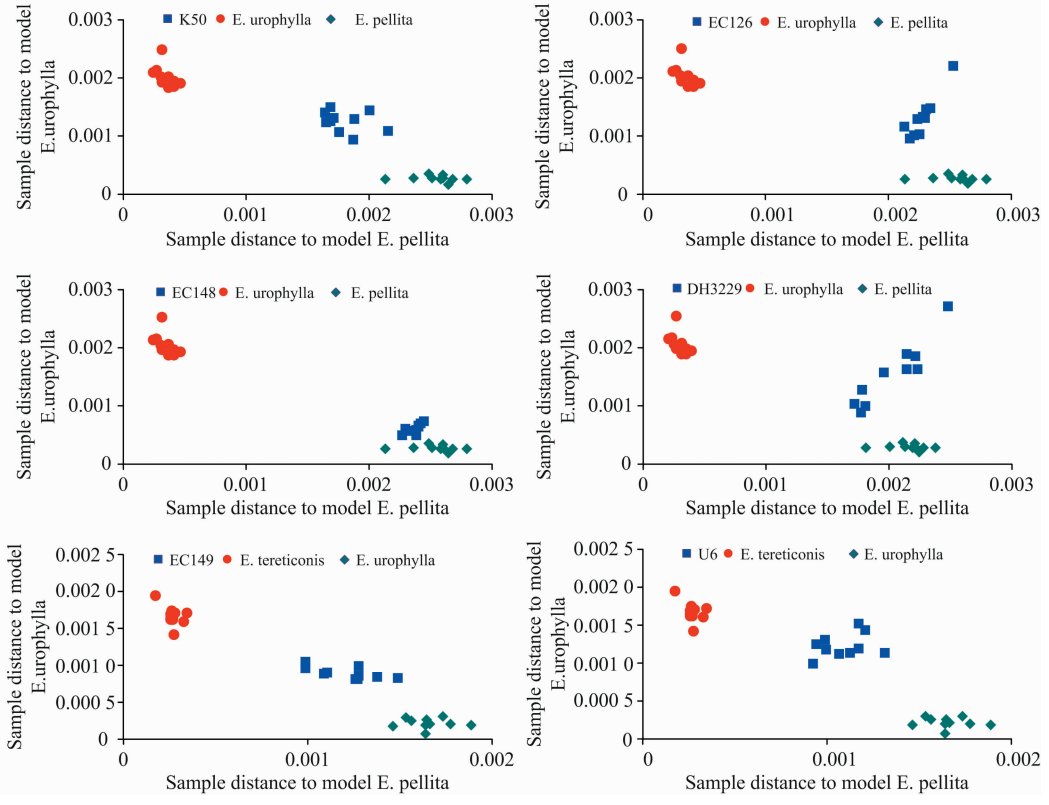

图 4　SIMCA 分析中 6 个桉树杂交种样本到亲本模型中心距离的 Cooman 图

Fig. 4　Cooman plot from the SIMCA analysis showing the distance to the cross parents PCA model centers for six eucalypt hybrids

2.4　桉树亲本模型对其杂交种的 PLS-DA 判别

图 5 展示了 5 个组合的桉树杂交亲本 PLS 模型对亲本和杂交子代样本的预测结果。结果显示，每个亲本的预测值都集中在各自响应变量周围(1，2 和 3)，且集中度很高。杂交种 K50 的预测值为 1.5～1.7，EC126 的预测值为 1.1～1.2，EC148 的预测值为 2.1～2.3，EC149 的预测值为 1.2～1.5，U6 的预测值为 1.5～1.8。预测值显示，EC126 和 EC148 的预测值高度重叠，EC149 预测值的变异幅度最大，

K50 和 U6 预测值的变幅居中。PLS-DA 判别可以清晰地将不同基因型区分开来，不过，任何判别方法都需要专业知识来配合解读分析结果。

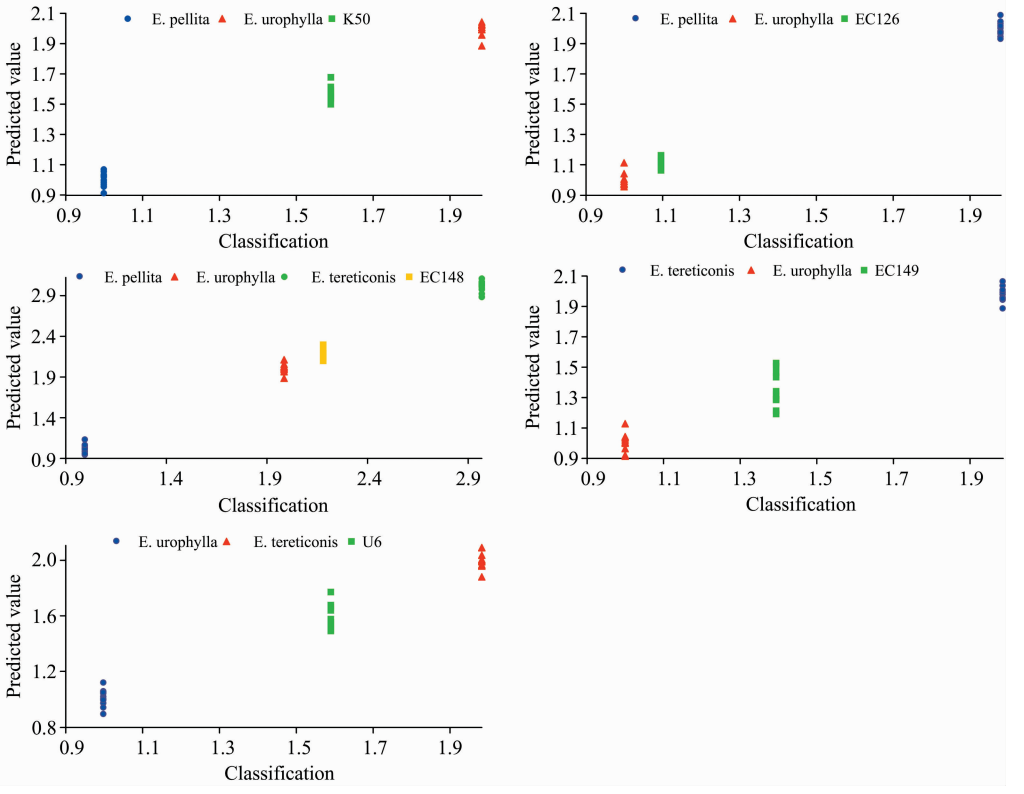

图 5 PLS-DA 模型对桉树亲本与其杂交种样本响应变量的预测

Fig. 5 PLS-DA prediction of eucalypt parents and the hybrids samples

图 6 是 PLS-DA 判别分析时所建模型第一个主因子的载荷，第二和第三个主因子的载荷分布与第一个相似，强度略有差异，本文没有展示。图中小方块指此处波段所对应的有机化合物的 NIRs 特征峰。1 890 nm 处为化学键 O—H 和 C—O的伸缩振动(stretching)吸收峰，对应的化合物主要为纤维素。1 980～2 000 nm 处为O—H键伸缩振动、水分子中 O—H 键变形(deformation)吸收峰[13-14]。

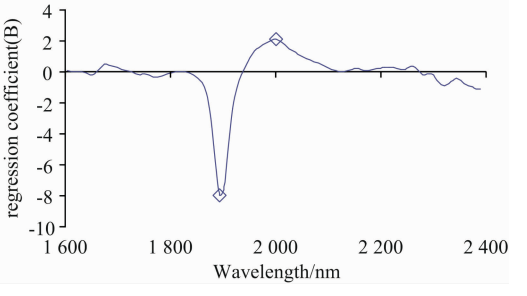

图 6 PLS-DA 判别模型第一主因子不同波段的载荷

◇：化学键的 NIRs 特征峰

Fig. 6 The loading weights for the first latent variable of the PLS-DA regression for assigned values

The small square ◇ indicates NIRs characteristic peaks of the chemical bands

3 结 论

采用 SIMCA 和 PLS-DA 两种有监督的判别模型，有效地解决了桉树叶片 NIRs 信息复杂、重叠的问题。两种模式的判别效果均显示，NIRs 可以将桉树杂交种、亲本、杂交种与其亲本清晰地区分。

结果表明 NIRs 真实地反映了不同基因型的遗传信息。其次，NIRs 可以反映桉树的遗传变异程度，即桉树杂交 F1 代来自亲本的加性遗传效应的大小。所以，NIRs 能够用于桉树树种的区分，也可根据 PCA 聚类离散度或 PLS 的预测响应变量分析基因重组过程中的加性遗传效应。

References

[1] Akbari M, Salehi H, Niazi A. Molecular Biotechnology, 2018, 60(4): 259.
[2] Miyata K, Morita S, Dejima H, et al. Diagnostic Cytopathology, 2017, 45(11): 963.
[3] Gungordu A, Uckun M, Yologlu E. Chemosphere, 2016, 144(125): 2024.

[4] Corral M A, Paula F M, Meisel D M C L, et al. Parasitology, 2016, 144(2): 1.
[5] Abasolo M, Lee D J, Raymond C, et al. Forest Ecology & Management, 2013, 304: 121.
[6] Espinoza J, Hodge G, Dvorak W. Journal of Near Infrared Spectroscopy, 2012, 20(4): 437.
[7] Hayes R A, Nahrung H F, Lee D J. Australian Journal of Botany, 2013, 61: 52.
[8] Diepeveen D, Clarke G P Y, Ryan K, et al. Journal of Cereal Science, 2012, 55(1): 6.
[9] Sandak A, Sandak J, Negri M. Wood Science & Technology, 2011, 45(1): 35.
[10] LU Wan-hong, YANG Gui-li, LIN Yan, et al(卢万鸿, 杨桂丽, 林 彦, 等). Scientia Silvae Sinicae(林业科学), 2017, 53(5): 16.
[11] Yang G L, Lu W H, Lin Y, et al. Journal of Tropical Forest Science, 2017, 29(1): 121.
[12] Meder R, Kain D, Ebdon N, et al. Journal of Near Infrared Spectroscopy, 2014, 22(5): 337.
[13] Schwanninger M, Rodrigues J C, Fackler K. Journal of Near Infrared Spectroscopy, 2011, 19: 287.

Identifying Eucalypt Hybrids and Cross Parents by Near Infrared Spectroscopy

LU Wan-hong, LI Peng*, WANG Chu-biao, LIN Yan, LUO Jian-zhong
China Eucalypt Research Centre, Zhanjiang 524022, China

Abstract Studying the gene control pattern of interested traits after control pollination is one of the key fields in exploring the law of gene recombination in eucalypt improvement. The accuracy of conventional quantitative analysis for that is often low, and the DNA analysis for that has high professional requirements, and is time consuming and laborious commonly. The aim of the current study is to study the relationship among different genotypes of hybrids, parents, hybrids and their parents in eucalypt based on the near infrared spectroscopy (NIRs) of foliage, and to discuss the practicability and the accuracy of the NIRs discriminant model for the classifying of eucalypt hybrids and their parents. The genetical materials in the study contained three eucalypt parents and their F1 progenies by control pollination. Fresh and healthy leaves from middle to upper crowns in a tree from their field testing trials were collected, and 10 individuals were chosen per genotype. The handheld portable near infrared spectrometer Phazir Px (1624) was used to scan the NIRs of foliage collected. 10 healthy current-year leaves were chosen per individual tree, five scans for NIRs from each side of the middle part of the frontal vein of the leaves were taken, calculated the average of 50 scans as the NIRs data of a leaf, thus 10 NIRs were got for every genotype in totally. The transform of S. G first derivative with second order polynomial fit was performed for the raw NIRs in present study. The successive multivariate analysis was conducted after NIRs pretreatment. To demonstrate the classification of different genotypes by the principal component analysis (PCA) with the NIRs data of hybrids and the patents in eucalypt. Then, two supervised discriminant models, soft independent modeling of class analogy (SIMCA) and partial least squares-discriminant analysis (PLS-DA) pattern recognition, were used to test the accuracy of NIRs model in the classifying for eucalypt hybrids and their cross parents. The scores plot of PC1 and PC2 in PCA demonstrated strong groups among different genotypes, such as cross parents, hybrids, and between hybrids and their parents. The sample distance to parents PCA model in the SIMCA analysis showed that the hybrids to be distinguished can form a clear group differentiated with their parents, and demonstrated the genetic similarity between parents and their progenies directly. The PLS-DA pattern recognition analysis indicated that the hybrids can be discriminated with cross parents by the response values of hybrids predicted by the parents PLS model. All the findings in present study showed that NIRs information of eucalypt leave truly reflects the transmission of genetic information occurring in the process of control pollination, and different genotypes including hybrids and their parents can be discriminated accurately by NIRs models, suggesting that NIRs can be used not only for qualitative identification between eucalypt hybrids and the cross parents, but also for assessing the extent of additive genetic effects in gene recombination, which can provide theoretical reference for the genetic basis analysis and breeding improvement in eucalypt.

Keywords Supervised model; Principal component analysis (PCA); Soft independent modeling of class analogy (SIMCA); Partial least squares-discriminant analysis (PLS-DA)

* Corresponding author

(Received Jan. 23, 2019; accepted Apr. 7, 2019)