CGMCP2019120 种质资源分析报告

项目编号：CGMCP2019120

报告单位：中玉金标记（北京）生物技术股份有限公司

联系电话：010-53275858

报告日期：2020年04月22日

目录

育种芯片介绍

为提高我国农作物的育种效率和精准性，把分子育种技术应用到育种过程中，中玉金标记推出了“中芯系列育种芯片”，运用Affymetrix® Axiom平台开发了一款高质量，高稳定，高性价比的玉米育种专用芯片 –中玉芯1号。

中玉芯1号的特点如下：

1. 有效标记多，质量高，稳定性好
2. 适用于多种种质材料
3. 基因组信息完善
4. 标记分布均匀

本图展示的是中玉芯1号设计的标记在染色体上的分布情况，由图可见，标记基本均匀覆盖玉米全基因组。

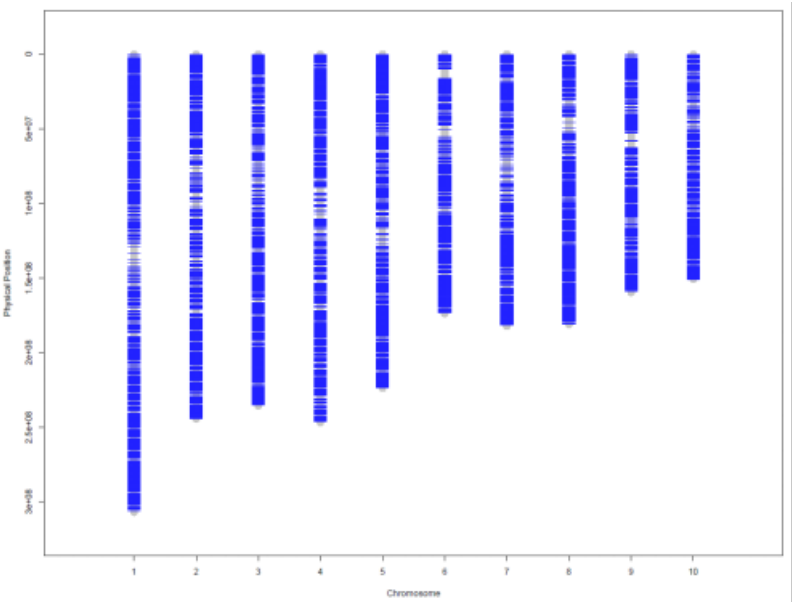

图1 标记在玉米染色体中的位置

横坐标：染色体编号
纵坐标：染色体上物理位置

表中展示的是中玉芯1号设计的标记在染色体上统计情况，由表可见：每条染色体上两标记间的物理距离平均都在200k左右。

表1 中玉芯1号标记染色体分布信息

染色体	标记数	标记间平均距离(bp)
1	1444	212632
2	1067	229091
3	1104	213469
4	1035	238647
5	1019	219725
6	770	226013
7	869	252604
8	869	208423
9	720	221903
10	676	223343

第一列：染色体编号
第二列：染色体上标记数
第三列：染色体上标记间平均物理距离

数据质控

使用中玉芯1号对客户材料做分析，通过对DQC>0.82和CR（标记检出率）>97 的样品进行SNP位点质控，剩余8970 个标记，统计结果见表2，将客户数据与公司种质资源库数据合并，进行miss<0.1和maf>0.05过滤，最终获得4052 个可用标记。

标记质控结果汇总详细内容如下：

表2 标记质控结果汇总

ConversionType	Count(Percentage)	Count_filter(Percentage)
ALL	9433	
NoMinorHom	150	1.59%
MonoHighResolution	21	0.22%
Other	1004	10.64%
PolyHighResolution	7218	76.52%
CallRateBelowThreshold	29	0.31%
OTV	1011	10.72%

ConversionType：标记类型；Count：标记类型个数；Count_filter：过滤完后剩余的标记类型个数

样品杂合率统计

样品杂合率的计算方式是样本的杂合标记数除以总标记数。样品杂合率可以体现材料的纯合程度，自交代数越高的材料样品杂合率越低。高代自交系的样品杂合率一般不会超过10%。

样品杂合率结果的取值范围是从0到1，代表从0%到100%标记杂合。

结果见附件CGMCP2019120.Sample_basic_statsistic_nuc.csv。部分结果见表3。

表3 杂合率统计

sample	calling_rate	het_ratio
H100	0.998	0.009
H101	1.000	0.007
H102	1.000	0.003
H103	0.999	0.005

sample：样品名称
calling_rate：标记检出率
het_ratio：样品杂合率

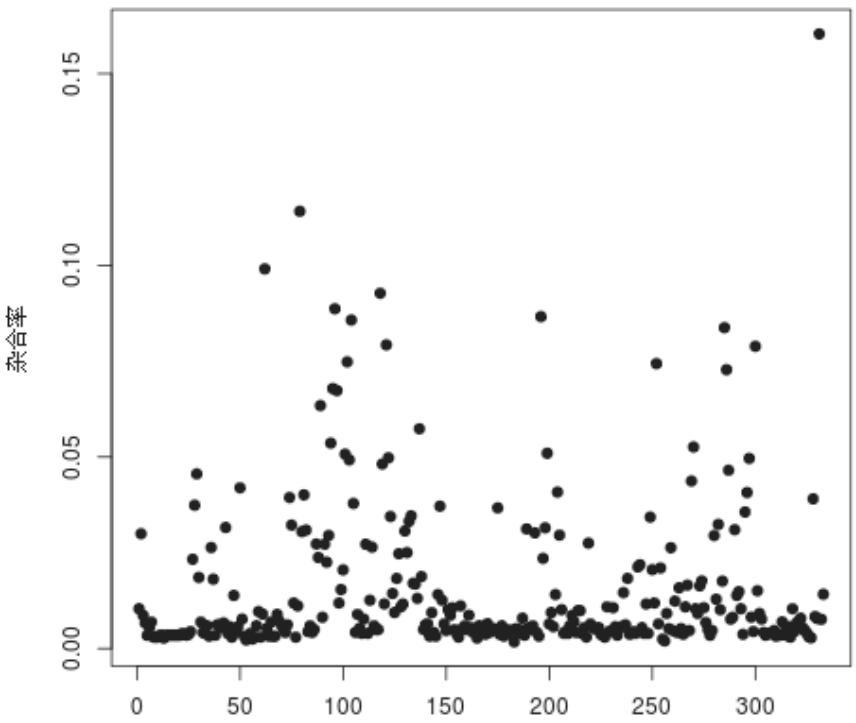

图2 样品杂合率分布图

群体结构分析

1）进化树分析

系统进化树（phylogenetic tree，又称evolutionary tree进化树）就是描述群体间进化顺序的分支图或树，用来表示群体间的进化关系。根据群体的物理或遗传学特征等方面的共同点或差异可以推断出它们的亲缘关系远近。

该图展示的是使用Treebest软件的nj-tree模型构建的进化树[1]，图中每一个分支代表一个样品，分支分离的越早代表遗传关系越远。

该项目分析结果见图3，矢量图见附件tree.pdf，图中不同颜色代表不同的群体类型。

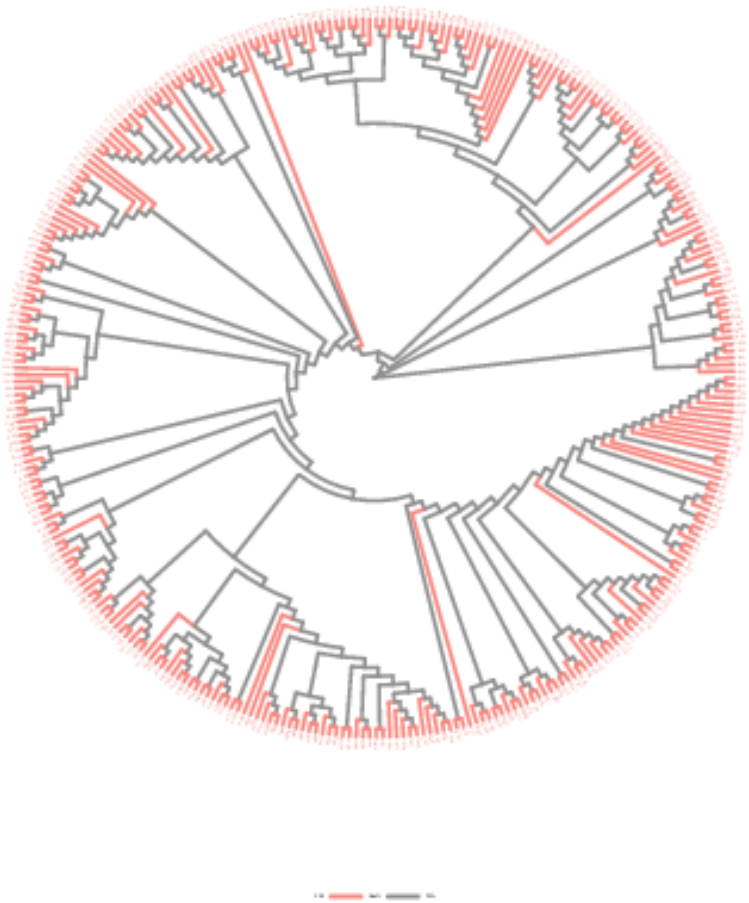

图3 tree图

2）PCA分析

主成分分析（PCA）是一种纯数学的运算方法，可以将多个相关变量经过线性转换选出较少个数的重要变量。PCA应用到很多学科，在遗传学当中，主要用于聚类分析，它是基于个体基因组SNP差异程度，按照不同性状特征将个体按主成分聚类成不同的亚群，同时用于和其它方法做相互验证。项目采用GCTA对样品进行PCA的分析[2]。

该项目分析结果见下图4，矢量图见附件PCA1v2.pdf, PCA1v3.pdf, PCA2v2.pdf, PCA1v2v3.pdf。

三张PCA图，分别代表三种维度搭配，PCA1v2v3.pdf为三维图。PCA结果实际上是一个很多维的数据，而前三个维度是能够解释所有材料是遗传变异最多的三个维度。我们分别将维度1vs2 、1vs3 、2vs3提取出来并做成二维图，其中PCA1v2.pdf最能代表所有材料的遗传变异，PCA1v3.pdf和PCA2v3.pdf次之。PCA图中不同亚群的骨干自交系会用不同颜色标注，见下图3：

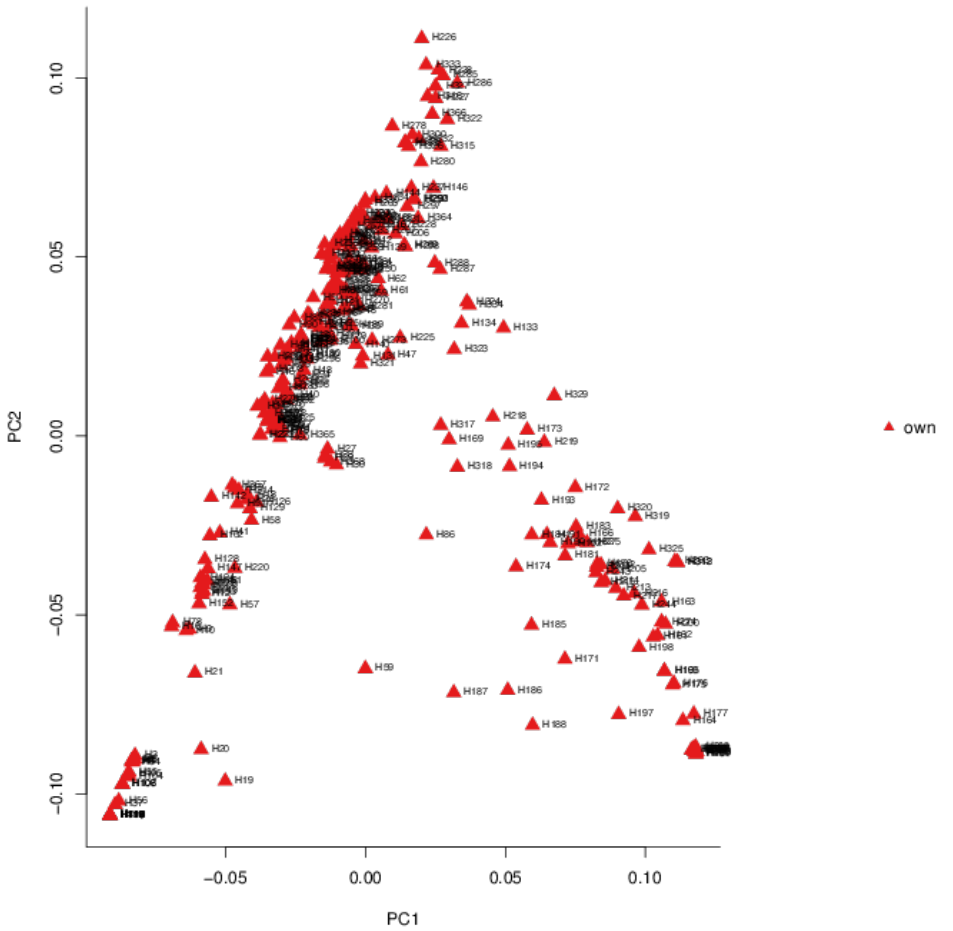

图4 PCA1vs2 图

图中所示（图中右侧展示不同颜色对应的群体类型）：

335F：335父本血缘；335M：335母本血缘；EUROF1：欧洲父本血缘1；EUROF2：欧洲父本血缘2；EUROM：欧洲母本血缘；LAN：兰卡斯特血缘；PB：PB血缘；REID：改良瑞德血缘；SPT：塘四平头血缘；ZI330：自330和旅大红骨血缘。

3）群体遗传结构分析

群体遗传结构指遗传变异在物种或群体中的一种非随机分布。按照地理分布或其他标准可将一个群体分为若干亚群，处于同一亚群内的不同个体亲缘关系较高，而亚群与亚群之间则亲缘关系稍远。群体结构分析有助于理解进化过程，并且可以通过基因型和表型的关联研究确定个体所属的亚群。项目采用admixture进行结构分析[3]。

该项目分析结果见图5，矢量图见附件struture.pdf。图中不同颜色对应不同的群体类型。

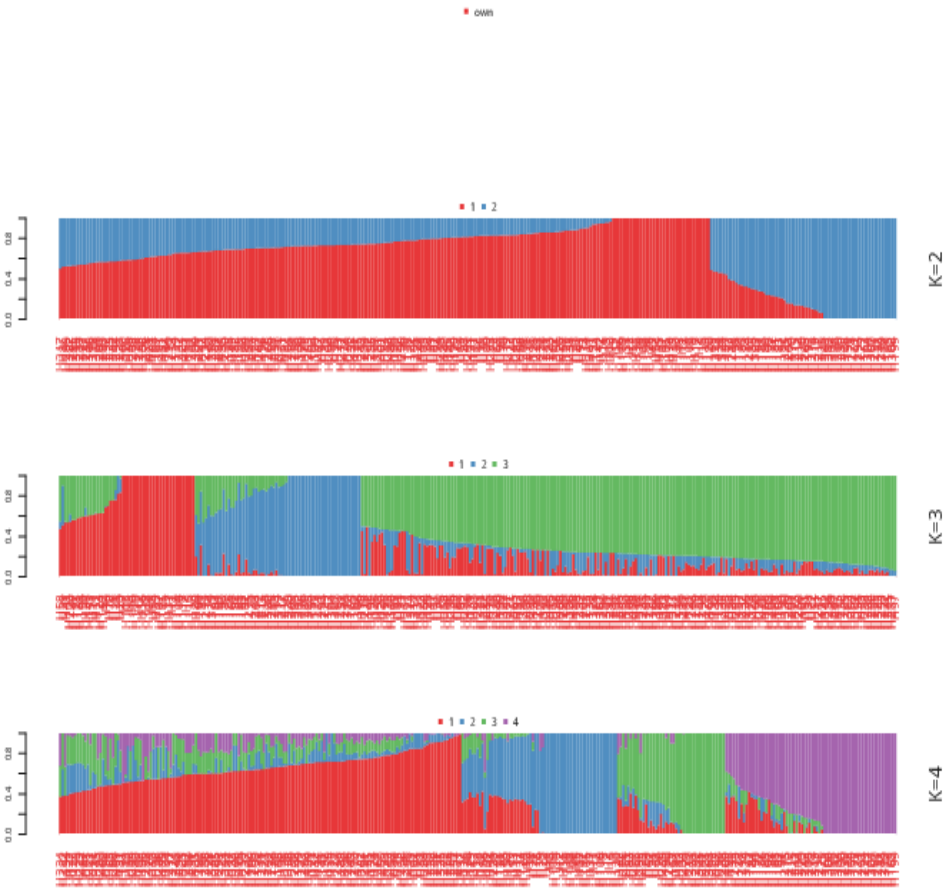

图5 structure 图

种质资源分析

1）IBD分析

IBD分析能够将材料的血缘构成细分出来。该方法按照国内的几大主要血缘（瑞德、塘四平头等）去拆分材料的血缘构成。IBD方法会取遗传比例最高的血缘作为IBD分群结果，同时也会把每个血缘所占的比重（IBD遗传背景比例）展现出来。IBD算法最初是为高密度分子标记检测设计的，在分子标记密度较低时准确性可能会受一些影响。如果在低密度标记检测时IBD算法与其他结果出现结果冲突，请以高遗传相似度骨干自交系为准。如果IBD遗传背景比例结果显示某个材料中每个血缘都占有一定比例而且比例都不大，这种结果一般代表IBD分析无法准确拆分该样本血缘，原因可能是该样本血缘太过混杂或者血缘超出了中玉金标记参照自交系的覆盖范围。

IBD分群部分结果见表4：

表4 IBD分群结果

sample	H7	H160	Group
H1	0.926	0.074	H7
H10	0.854	0.146	H7
H100	0.558	0.442	H7
H101	0.609	0.391	H7
H102	0.803	0.197	H7

sample：样品名称
Group：IBD分群结果
ZI330、SPT、REID等：群体血缘比例

2）遗传相似度分析

遗传相似度分析提供所有材料之间的遗传相似度。该分析既可以为客户进行材料测配（挑选相似度低的材料）和群体改良（挑选相似度高的材料）提供数据依据，同时又能通过查看材料与中国种质骨干自交系的相似度来为材料分群提供依据。同一血缘材料的遗传相似度一般会在60%或70%以上，50%左右或者以下的遗传相似度一般代表材料间没有明显的血缘联系。

详情见附件CGMCP2019120.genetic_similarity_matrix_nuc.csv。部分见表5：

表5 遗传相似度矩阵

	H1	**H10**	**H100**	**H101**	**H102**
H1	1.000	0.786	0.629	0.632	0.770
H10	0.786	1.000	0.624	0.614	0.698
H100	0.629	0.624	1.000	0.627	0.679
H101	0.632	0.614	0.627	1.000	0.661
H102	0.770	0.698	0.679	0.661	1.000

3）综合分群

综合分群将IBD分析、遗传相似度分析的结果进行归纳和合并，以此对材料的血缘有一个综合的判断。一般来说如果材料的血缘比较纯（混杂程度小）的话，则两种分析会对材料的杂种优势群归属有着相同的判断，这种情况下综合分群结果与两种分析方法结果相同。但是如果两种分析结果不同的话，则一般是因为材料血缘混杂所导致的，这种情况下综合分群结果将会为空，请结合IBD遗传背景比例和高遗传相似度骨干自交系结果对材料血缘进行判断。

详情见综合分群结果.csv，部分见表6：

表6 综合分群结果

sample	IBD Group	H7	H160	UNDECIDED	topgs1	topgs1group
H1	H7	0.926	0.074	0.000	CK_H7_0.9036	H7
H10	H7	0.854	0.146	0.000	CK_H7_0.8086	H7
H100	H7	0.558	0.442	0.000	CK_H7_0.6398	H7
H101	H7	0.609	0.391	0.000	CK_H7_0.629	H7
H102	H7	0.803	0.197	0.000	CK_H7_0.7669	H7

sample：样品名称
IBD Group：IBD分群
topgs1group：该样品相似性最高的CK系对应的群体
topgs1：该样品相似性最高的CK系，相似性值为第二个下划线后面的数值

参考文献

1. Vilella A J, Severin J, Uretavidal A, et al. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates.[J]. Genome Research, 2009, 19(2):327-335.
2. Yang J, Lee S H, Goddard M E, et al. GCTA: a tool for genome-wide complex trait analysis.[J]. American Journal of Human Genetics, 2011, 88(1):76-82.
3. Alexander D H, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals[J]. Genome Research, 2009, 19(9):1655-1664.