PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2015, 42(7): 674~684
www.pibb.ac.cn

利用 AIMs 分析亚洲 9 个绵羊群体的群体结构 *

张媛媛[1] 韩德平[1] 邓卫东[3] 毛华明[3] 邓学工[2]** 邓学梅[1]**
([1] 中国农业大学畜禽育种国家工程实验室，北京 100193；[2] 东北大学理学院，沈阳 110819；
[3] 云南农业大学动物科技学院，昆明 650201)

摘要 祖先信息标记(ancestry informative makers，AIMs)可用来分析群体的遗传结构．本研究利用 Illumina OvineSNP50 芯片上的 SNP 位点，在云南乌骨绵羊和其他 8 个亚洲绵羊群体中以 Rosenberg 等定义的 Informativeness 统计量为筛选方法，选取 Informativeness 值最高的前 20、50、100、500 个 SNP 位点与相应数目的随机 SNP 位点分别用来推断群体的遗传结构．通过主成分分析和用 fastSTRUCURE 推断祖先成分的方法，评价 AIMs 在推断亚洲绵羊群体遗传结构中的作用．研究显示，利用筛选到的高信息含量标记 AIMs，可减少群体结构研究中需要的 SNP 位点数目．前 50 个 AIMs 可有效地将绵羊群体分为 4 个大类，这与利用全基因组 SNPs 分析得到的群体结构是一致的，即乌骨绵羊群体 blackbone 单独为一类、西藏群体 changthangi 和 tibetan 归为一类，孟加拉的 banglandeshi、banglandeshiGarole 群体和印度的 IndianGarole 群体归为一个类群，其余三个群体(印度尼西亚 sumatran、garut 群体和印度的 deccani 群体)近似归为一类．这 4 大类群在 AIMs 上存在显著的分化，利用这些位点信息可以为研究群体特征和进化关系提供线索．

关键词 祖先信息标记，绵羊，群体结构，主成分分析，祖先成分推断，遗传分化系数
学科分类号 Q38 **DOI**: 10.16476/j.pibb.2015.0062

群体结构是在进行遗传关联研究中值得关注的一个重要问题，实验群体中混入不同祖先来源的个体，会导致 case 和 control 之间的等位基因频率出现偏差，因而造成表型与基因型之间出现异常的关联，或者检测不到正确的关联．特别是在大群体中对低风险度的遗传变异位点的关联分析中，群体结构导致错误关联的问题更为严重[1-5]．不同祖先来源群体的等位基因频率的差异虽然会引起群体分层现象，但这些遗传标记能够反映群体的祖先来源和进化历史，往往也代表了这些群体的遗传特征．例如，乳糖酶(lactase，LCT)基因上的一个 SNP 在欧洲人群中由于受到正向选择而具有较高的频率，因而在 European American 中研究体高与基因型的关系时，会造成该位点与体高表型强相关[6]．此外，与酒精成瘾高度相关的 SNP 标记 rs1799971 在高加索人群中频率较高，但其可能并不是导致易成瘾的致因突变[7-8]．这些遗传标记，如果在全基因组关联研究(GWAS)中忽略实验群体的祖先来源，则很容易被认为是与研究表型相关联的标记；但另一方面，这些标记在群体结构的研究中具有重要意义．类似于这些在不同群体之间基因频率差异非常大的多态性位点被称为祖先信息标记(ancestry informative markers，AIMs)，AIMs 可用于研究群体结构，推断祖先个体及分析遗传进化关系[9-11]．

目前，已报道有许多方法可以用来矫正群体结构，减少对关联分析的影响，比如基因组控制(genomic control，GC)[12-13]，多维变量方法——主成分分析(principal component analysis，PCA)[14]，以及混合线性模型中加入可变方差组分估计的亲缘关系矩阵进行群体结构校正的方法(EMMA/EMMAX)

* 国家自然科学基金(U1136605), 教育部博士点基金(20120008110049)和国家转基因重大专项(2011ZX08009-001)资助项目．
** 通讯联系人．Tel: 010-62733933
邓学梅．E-mail: deng@cau.edu.cn
邓学工．E-mail: dengxuegong@tom.com
收稿日期：2015-03-10，接受日期：2015-06-03

等[15]．PCA 的方法在人类及动物的群体结构研究中应用广泛，而且 PCA 的结果可以用来作图，直观反映群体之间的聚类情况，因而 PCA 方法评价群体结构直观方便．对群体的遗传成分进行推断，也是常用的群体结构分析方法，常用 STRUCTURE 软件[16]进行分析，通过设定祖先成分数目 K，对群体中的个体的祖先成分进行推断，从而判断群体之间的遗传关系．

本研究利用绵羊 Illumina OvineSNP50 的 SNP 数据，分析了已知群体来源的 9 个绵羊群体的群体聚类情况，并检测了群体中祖先信息标记位点(AIMs)．利用检测到 AIMs 集合和相同数目的随机 SNPs 集合分别对这 9 个绵羊群体进行 PCA 分析和祖先成分推断，比较两类 SNP 集合在检测群体结构上的差异，同时比较由 AIMs 得到的群体结构与全基因组 SNP 分析得到的群体结构是否相符合．研究发现，AIMs 可以有效地推断亚洲 9 个绵羊群体的群体结构，AIMs 推断群体结构的效果显著优于随机 SNP．AIMs 在不同群体之间的频率差异较大，可能为研究这几个绵羊群体遗传进化特征提供线索．

1 材料与方法

1.1 样本采集与 SNP 数据整理

在云南省兰坪县，随机采集 60 只乌骨绵羊的颈静脉抗凝血．取抗凝血 20 μl，按常规酚仿法抽提血液 DNA．经检验合格后，利用 Illumina OvineSNP50 对乌骨绵羊个体进行全基因组的 SNP 分型．SNP 数据质控参考文献[17]的质控方法：a．去除检出率小于 0.99 的 SNP 位点；b．去除分型结果存在质疑的 SNP 位点；c．去除无多态性的 SNP 位点；d．去除不符合哈代温伯格平衡的位点．从 ISGC(international sheep genomic consortium)下载 8 种亚洲绵羊群体的基因组 SNP数据．本研究中选取的不同绵羊群体，以及每个群体的样本数量见表 1．通过合并乌骨绵羊，获得 9 种亚洲绵羊的数据，将质控后得到 SNP 重新比对到绵羊参考基因组 OAR3.1 上，其中常染色体上的 SNP 为 41 753 个．SNP 位点的期望杂合度，群体近交系数应用 PLINK[18](--het)计算．通过 PLINK[18]方法，去除其中较高 LD 的 SNP(--indep-pairwise 50 5 0.5)，经过滤后得到 39 158 个 SNP，用于后续 AIMs 筛选．不同绵羊群体的具体筛选结果见表 1．

Table 1 Genetic diversity and population size of each population

Population	District	Abbrev	*n*	*F*	*He*
BangladeshiBGE*	Bangladesh	BGE	24	0.32	0.31
BangladeshiGarole*	Bangladesh	BGA	24	0.28	0.30
Changthangi*	Changthang, China	CHA	29	0.17	0.35
Deccani*	Deccan, India	DEC	24	0.16	0.33
Garut*	Indonesia	GAR	22	0.17	0.32
IndianGarole*	India	IGA	26	0.28	0.29
Sumatran*	Western Indonesia	SUT	24	0.24	0.31
Tibetan*	Tibet, China	TIB	37	0.20	0.34
Blackbone	Yunnan, China	BLB	60	0.22	0.32

Genetic diversity indices measured within population. *n* gives the number of individuals used to calculate the AIMs, expected heterozygosity or gene diversity (*He*), the inbreeding coefficient (*F*). District, from where the samples were collected. Abbrev, abbreviation of each population. *Data of these populations were downloaded from International Sheep Genomics Consortium(ISGC, www.sheephapmap.org)

1.2 AIMs 位点筛选及 AIMs 集合构建

AIM 位点的筛选参考 Rosenberg 等[10]的方法．定义每个 SNP 位点的信息量值(informativeness，I_n)，具有较大 I_n 值的位点为 AIM 位点．具体计算公式如下：

$$I_n=\sum_{j=1}^{N}(-p_j\ln p_j+\sum_{i=1}^{K}\frac{p_{ij}}{K}\ln p_{ij})$$

其中，p_j 为等位基因 j 在所有群体的平均频率．p_{ij} 是第 i 个群体中等位基因 j 的频率．K 为群体总数，N 为该位点的等位基因数目．ln 为自然对数，且定义 0ln0=0．

将上述经过 PLINK 方法过滤得到的 39 158 个 SNP 位点的 I 值，按照从大到小的顺序进行排序．选取 I 值较高的前 20、50、100 和 500 个 SNP 位点，分别组成不同的 AIMs 集合，标记为 Top20，Top50，Top100 和 Top500，用于后续的群体结构推断．

1.3 随机 SNP 集合构建

对筛选获得的 39 158 个 SNP，随机抽取 20、50、100 和 500 个 SNP 位点，组成随机 SNP 集合，分别记为 random20，random50，random100 和 random500，用以进行后续的群体结构推断．为增加结果准确度，每个随机 SNP 集合均重复构建 3 次．

1.4 P 群体遗传结构分析

利用 smartpca 软件[19]，对基因组上 SNP 标记进行 PCA 分析．得到的 eignvec 文件中包含每个个体的 PCA 的特征向量，抽取前两个维度向量作为群体结构的二维指示坐标，利用 R 软件进行作图分析．PCA 是主要的非参数方式估计群体结构的方法，另外通过假设群体具有 *K* 个祖先来源的方法估计群体之间的祖先成分差异也可以用来评价群体的遗传结构．因而以 STRUCTURE 方法结合 PCA 方法对获得的 AIMs 集合进行验证．STRUCTURE 分析采用 fastSTRUCTURE[20]分析软件完成．

1.5 群体 *Fst* 值计算

以 Wright's *Fst* 值[21]作为群体遗传结构的参考指标．*Fst* 用于表征亚群体间的遗传分化尺度，可以对不同群体之间的遗传关系远近进行量化．本研究中，9 个亚洲绵羊群体的两两之间的 *Fst* 值通过 Genepop 软件[22]计算获得．

根据群体的 *Fst* 值，通过 neighbor-joining 算法构建关系树，直观反映群体之间的遗传关系远近．

2 结果与分析

2.1 9 个绵羊群体的群体结构分析

Fst 常用来表示群体之间的分化程度，可以用来推断群体结构、迁移、扩张等[21, 23]．9 个绵羊群体两两之间的 *Fst* 值见表 2．结果表明，各个群体间都存在中等以上程度($Fst > 0.05$)的分化；所处地域相近的绵羊群体之间的 *Fst* 较小，如孟加拉的 BGE 和 BGA 群体，藏北区域 CHA 群体与 TIB 群体．乌骨绵羊 BLB 群体分布在云南兰坪地区，与西藏的 TIB 群体和 CHA 群体地理位置接近，群体间的 *Fst* 值最小，BLB 群体与其他群体的分化程度较高(Fst=0.12～0.17)．为了观察群体之间的遗传距离关系，利用 *Fst* 值构建了 neighbor-joining tree (图 1)．

Table 2 *Fst* statistics between populations reveal the genetic relationship among sheep populations

	BGE	BGA	CHA	DEC	IGA	SUM	GAR	TIB
BGA	0.06728							
CHA	0.09656	0.1045						
DEC	0.08755	0.09508	0.06551					
IGA	0.09283	0.09659	0.1208	0.1094				
SUM	0.1084	0.1149	0.1087	0.09923	0.1305			
GAR	0.117	0.1226	0.09652	0.09781	0.1412	0.1207		
TIB	0.1182	0.1258	0.03239	0.08767	0.142	0.1294	0.1168	
BLB	0.1499	0.1577	0.07084	0.1215	0.1722	0.1589	0.1472	0.06829

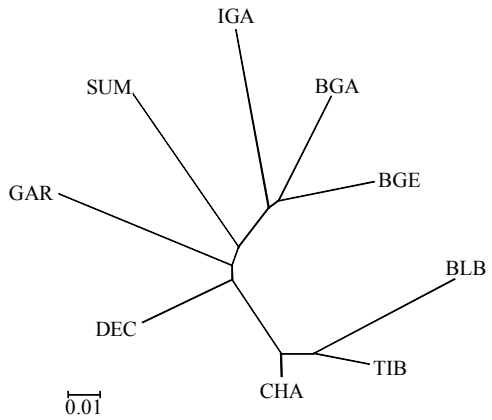

Fig. 1 Genetic relationship between sheep populations based on Fst statistics

Fst values were estimated by a subset of 39 158 SNP identified by LD-based SNP pruning through PLINK (--indep-pairwise 50 5 0.5). Neighbor-joining tree was generated to show the genetic relatedness between populations.

PCA 和 STRUCTURE 方法是常用来研究群体遗传结构的非参数和参数方法．本研究中，我们对 OvineSNP50 芯片上的 SNP 位点进行质控和降低连锁程度处理后，剩余的 39 158 个常染色体上的 SNP 位点用来对乌骨绵羊和其他 8 个亚洲绵羊群体进行群体结构推断，PCA 和 TRUCTURE 结果分别由如图 2、图 3 展示．

在图 2 的 PCA 结果中，每一个点代表一个个体，被标记上不同颜色来代表所属群体，能直观反映个体之间的聚集情况．C1-C2、C1-C3、C2-C3 这三个组合都能很好地展示群体聚类和分层的现象．为了统一，后续的分析讨论中，仅以 C1-C2 组合来展示群体的遗传结构特点．显然，同一个群体中的绵羊个体聚类在一起，不同群体之间较为分

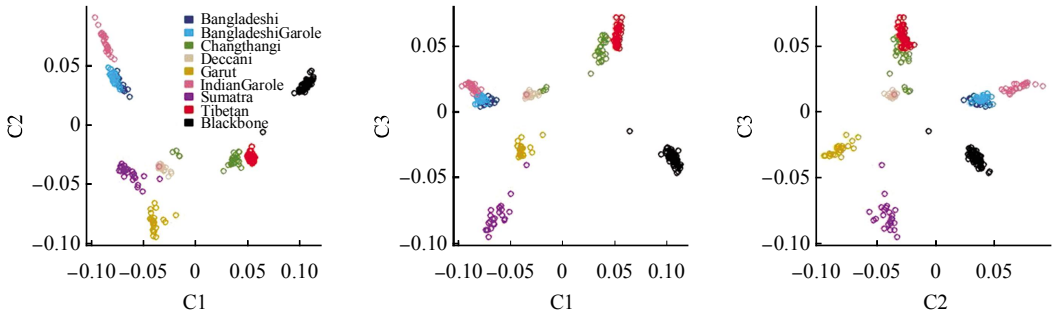

Fig. 2　Population structure within the sheep populations

Principal component analysis of genetic distance was performed using a subset of 39 158 SNPs identified by LD-based SNP pruning through PLINK (--indep-pairwise 50 5 0.5). Individuals are color coded to present their populations.

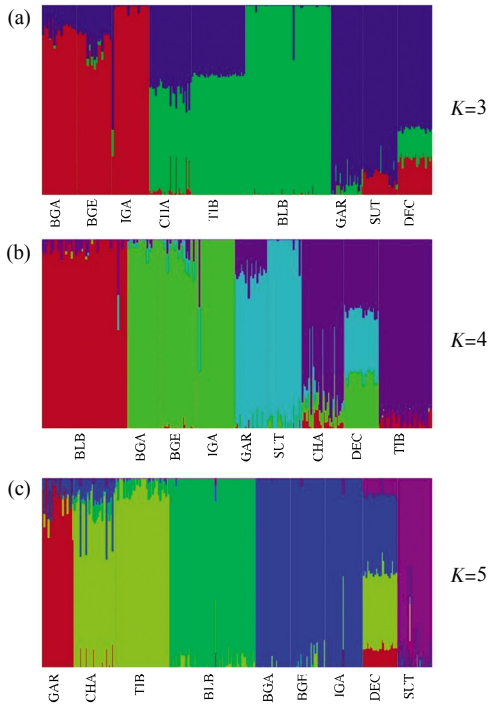

Fig. 3　Individual ancestry infered using a subset of 39 158 filtered SNP

Regional ancestry inferred with the fastSTRUCTURE programs and plotted with the Distruct program. Each individual is represented by a vertical line partitioned into colored segments whose lengths correspond to the ancestry coefficients in up to K inferred ancestral groups. (a) K=3. (b) K=4. (c) K=5.

散．彼此所处地理位置相近的群体反映在 PCA 图上的距离也较近，这与 *Fst* 的结果(图 1 和表 1)是类似的．

STRUCTURE 方法得到的结果(图 3)与 PCA 的结果是一致的．STRUCTURE 方法通过假定不同的祖先来源数目 K，可以将不同群体之间是否含有相同祖先成分表示出来．K=3～5 时，BLB 群体都能被单一的推断组分所代表，且随着 K 增加，BLB 的推断组分几乎不存在于其群体中，说明 BLB 群体与其他群体有不同的祖先来源，但在 CHA 和 TIB 这两个与 BLB 地理位置较近的群体中，也含有少量的 BLB 的推断组分，推测 CHA 和 TIB 群体可能曾经与 BLB 群体发生基因交流．另一方面，具有相似的推断祖先成分的群体，其祖先来源可能相同，且往往是地理位置比较接近，比如位于西藏地区的 TIB 和 CHA 群体之间的 *Fst* 较小(0.032)，在 K=3～5时，都表现出具有相似的推断组分的群体结构特点，印度的 Garole 绵羊群体 IGA 和邻国孟加拉的 Garole 绵羊群体 BGA，彼此之间的 *Fst* 比与其他群体之间的 *Fst* 较小，因而在 STRUCTURE 分析中也表现为相似的群体结构，但孟加拉本土绵羊群体 BGE 和 BGA 的群体结构相似度更高．DEC 是典型的具有多个混合祖先来源的群体．

2.2　利用不同的 SNP 集合推断绵羊群体遗传结构

根据计算得到的 SNP 的 I_n 值，挑选 I_n 值最大的前 20，50，100 和 500 个 SNP 构成 top20，top50，top100 和 top500 的 top SNPs 集合；另随机抽取相应数目的 SNP 构成 random SNPs 集合 random20，random50 和 random100，random500．由 top SNPs 集合计算两两群体之间的 *Fst* 值高于 random 集合，也高于全基因组 SNP 计算的 *Fst* 值，说明这几个绵羊群体在筛选到的 AIMs 上出现了明显的分化．这是因为根据 I_n 统计量筛选出来的

AIMs，在不同群体中存在明显的等位基因频率差异．图 4i 比较了前 50 个 AIMs 在 9 个绵羊群体的频率差异．

从图 4 PCA 结果反映的群体聚类和离散情况来看，利用 AIMs 分析群体结构的效果好于采用随机 SNPs．采用 20~100 个 AIMs 进行 PCA 分析时，9 个绵羊群体在主成分一上主要分为三大类群体，分别是乌骨绵羊群体 BLB、西藏群体 CHA 和 TIB，其余群体归为一大类(图 4a～c)．而相应数目的随机 SNP 集合的 PCA 结果均不能明显区分各个群体(图 4e～g)．

随着采用的随机 SNP 数目增多，PCA 所展示的群体聚类的结果趋于明显．但是利用 AIMs 分析群体结构的效果明显优于利用相同数目随机 SNP 分析的结果．

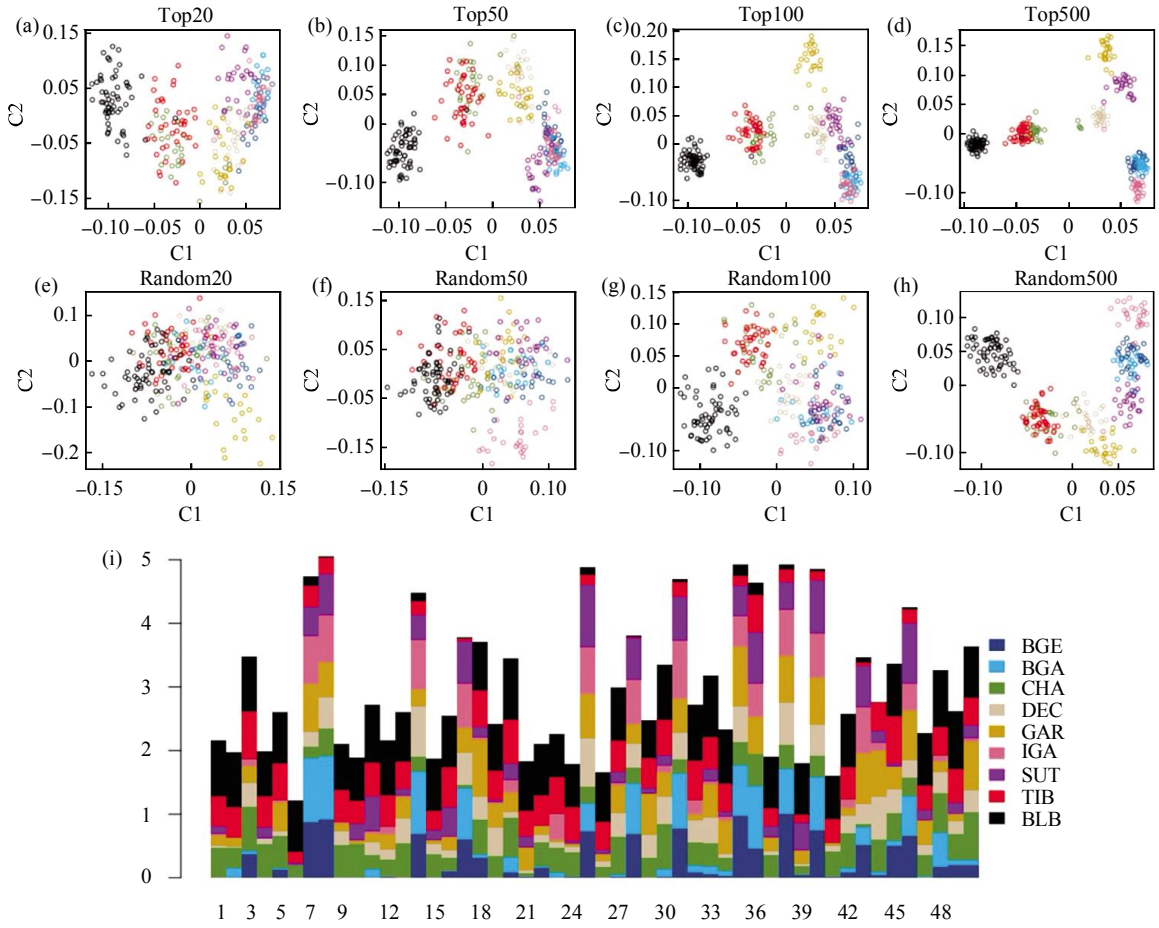

Fig. 4　Population structure PCA plots by different SNP subsets and top50 SNPs frequencies within 9 sheep populations

Principal component analysis of genetic distance was performed using different subsets of SNPs. Individuals are color coded to present their populations. (a～h) PCA population structure generated by SNP subsets top20, top50, top100, top500, random20, random50, random100 and random500, respectively. (i) Alleles frequencies of top50 SNPs within different sheep populations.

2.3　利用不同的 SNP 集合推断绵羊群体的祖先成分

利用 fastSTRUCTURE 软件，通过设置 K=3～6 分别计算和推断不同绵羊群体的祖先成分．K=4 时群体祖先成分的推断结果如图 5 所示．

结果表明，应用前 50 个 AIMs(图 5f)即可以获得与全基因组 SNP(图 3b)相似的祖先成分推断结果：BLB 群体有较为独立的成分，CHA 和 TIB 群体的祖先成分结构一致，BGA、BGE 和 IGA 群体祖先成分结构一致；利用全基因组 SNP 推断群体祖先成分时，GAR 和 SUT 具有非常相似的祖先来源，但 GAR 还含有另一祖先成分(图 3b)．而在利用前 50 个 AIMs 的推断中，GAR 和 SUT 祖先成分来源几乎一致．

利用前 20 个 AIMs，可以将 BLB 群体、CHA 和 TIB 群体的祖先来源差异凸显出来，但是对于其他群体的祖先成分推断不是很明显.

由横向比较相同数目 random 和 top 集合的 SNP 群体祖先成分的推断效果来看，AIMs(图 5e～h)显著优于随机 SNPs(图 5a～d)，而且随着 SNP 数目增加，群体祖先成分推断趋于明确.

AIMs 推断群体祖先成分时，将群体之间 *Fst* 值较小的群体判断为具有相似的祖先成分，从而分为几个类群，例如 CHA 和 TIB 群体之间的 *Fst* 值为 0.032，且被推断出具有相似的祖先成分(图 3，图 5f～h)，在 PCA 分析中容易归为一类(图 4a～d)；BGE、BGA 和 IGA 群体也有类似的情况.

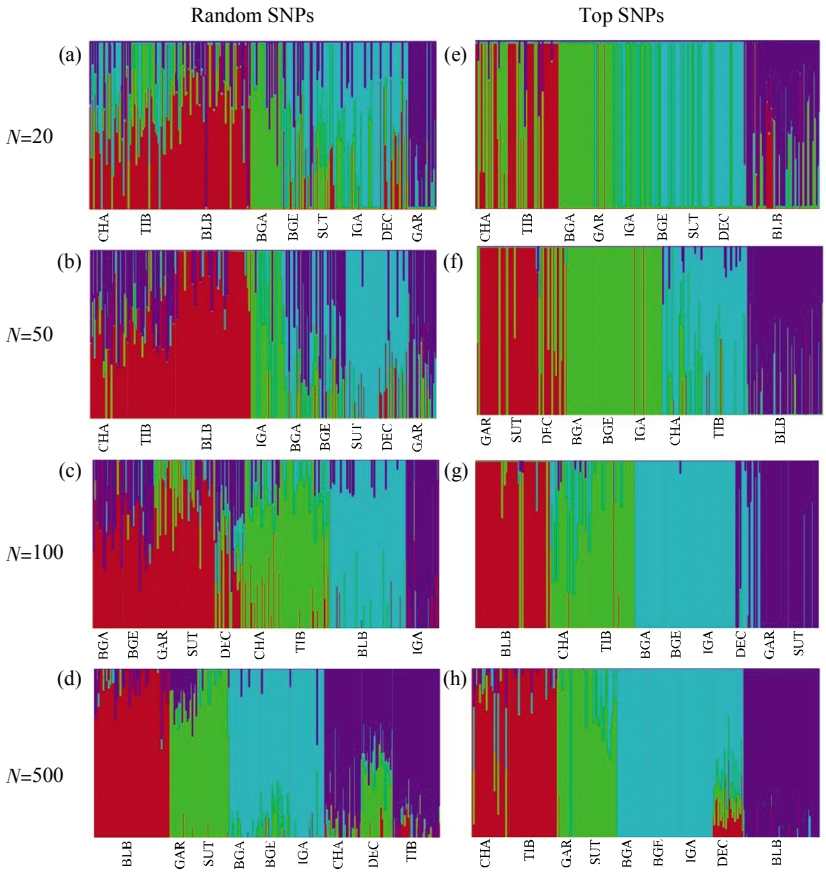

Fig. 5　Individual ancestry inferred by random SNPs subsets(a～d) and top SNPs subsets (e～h)

Regional ancestry inferred with the fastSTRUCTURE at K=4 and plotted with Distruct program. Each individual is represented by a vertical line partitioned into colored segments whose lengths correspond to the ancestry coefficients in up to four inferred ancestral groups. N means number of SNPs used in individual ancestry inferring.

2.4　AIMs 注释

选用 20～50 个随机 SNP 并不能推断出群体结构，而 20～50 个 AIMs 的 PCA 结果仅在一个维度就可将 9 个绵羊群体聚为几个大的类群. 图 4i 的 AIMs 频率分布也显示绵羊群体在这些位点的频率存在明显差别. 这些位点可能与不同绵羊群体的遗传特异性有关，BLB 群体分布在云南兰坪地区、CHA 和 TIB 群体都是西藏绵羊群体，BGA 和 BGE 群体是孟加拉群体，BGA 和 IGA 都是 Garole 绵羊[24]. 采用 Ensembl 的 VEP 软件对 I_n 值较高的前 50 个 AIMs 进行注释，一共注释到 26 个基因，GO 的生物学过程注释结果如表 3 所示，生物学过程的注释大多与心肌和骨骼肌细胞发育和分化有关. 这些基因是否与不同品种之间的遗传差异有关还需进一步研究.

Table 3 GO Annotations

Category	Term	*P* Value	Genes
Biological Process	GO:0055007 cardiac muscle cell differentiation	0.03183	ATG5, FOXP1
Biological Process	GO:0007517 muscle organ development	0.03478	ATG5, FOXP1, SERP1
Biological Process	GO:0035051 cardiac cell differentiation	0.03862	ATG5, FOXP1
Biological Process	GO:0048738 cardiac muscle tissue development	0.07714	ATG5, FOXP1
Biological Process	GO:0006412 translation	0.07765	ZNFX1, EIF4A1, EIF2A, SNORA67
Molecular Function	GO:0005524 ATP binding	0.00626	DDX27, KCNT2, PIK3CB, ZNFX1, EIF4A1, SNORA67, PRKG2, ATP8B4
Molecular Function	GO:0032559 adenyl ribonucleotide binding	0.00669	DDX27, KCNT2, PIK3CB, ZNFX1, EIF4A1, SNORA67, PRKG2, ATP8B4
Molecular Function	GO:0030554 adenyl nucleotide binding	0.00863	DDX27, KCNT2, PIK3CB, ZNFX1, EIF4A1, SNORA67, PRKG2, ATP8B4
Molecular Function	GO:0001883 purine nucleoside binding	0.00929	DDX27, KCNT2, PIK3CB, ZNFX1, EIF4A1, SNORA67, PRKG2, ATP8B4
Molecular Function	GO:0001882 nucleoside binding	0.00961	DDX27, KCNT2, PIK3CB, ZNFX1, EIF4A1, SNORA67, PRKG2, ATP8B4
Molecular Function	GO:0000166 nucleotide binding	0.01226	DDX27, KCNT2, PIK3CB, ZNFX1, EIF4A1, MSI2, SNORA67, PRKG2, ATP8B4
Molecular Function	GO:0032555 purine ribonucleotide binding	0.01786	DDX27, KCNT2, PIK3CB, ZNFX1, EIF4A1, SNORA67, PRKG2, ATP8B4
Molecular Function	GO:0032553 ribonucleotide binding	0.01786	DDX27, KCNT2, PIK3CB, ZNFX1, EIF4A1, SNORA67, PRKG2, ATP8B4
Molecular Function	GO:0017076 purine nucleotide binding	0.02188	DDX27, KCNT2, PIK3CB, ZNFX1, EIF4A1, SNORA67, PRKG2, ATP8B4
Molecular Function	GO:0042623 ATPase activity, coupled	0.04325	DDX27, EIF4A1, SNORA67, ATP8B4
Molecular Function	GO:0016887 ATPase activity	0.06245	DDX27, EIF4A1, SNORA67, ATP8B4
Molecular Function	GO:0003743 translation initiation factor activity	0.07262	EIF4A1, EIF2A, SNORA67

3 讨 论

3.1 全基因组 SNP 标记反映的绵羊群体结构

本研究中采用的是 Illumina OvineSNP50 芯片上的 SNP 位点，该芯片在设计过程中评价了 3 个检测平台得到的 SNP 数据的多态性和准确性(Illumina GA 测序平台得到了 76K 的 SNPs、454-PLX 270K 的 SNPs、Pilot chip 的 1 536 个 SNPs)，非常适合应用于群体遗传学的研究(http://www.sheephapmap.org/genseq.php). Kijas 等[25]根据 Pilot chip 的 SNP 数据，对 22 个绵羊品种和 2 个野生绵羊群体进行群体结构分析，后来又应用 OvineSNP50 芯片研究了全球 74 个绵羊品种共 1 612 个个体的遗传结构和群体历史[17]. 结果表明，各个绵羊群体能根据地理位置分别聚类，绵羊群体的遗传结构差异与群体之间地理位置远近有关. 本研究中亦采用 OvineSNP50 芯片上的 SNP 位点，对中国乌骨绵羊群体、中国西藏地区、印度和孟加拉的 9 个绵羊群体结构分析，也得到绵羊群体遗传结构与地域有关的结果. 另外，对 9 个绵羊群体进行 *Fst* 计算，得到的群体遗传结构结果与 PCA 分析和 STRUCTURE 分析的群体结构是一致的.

此外，*Fst* 和 PCA 的结果均表明乌骨绵羊群体与其他绵羊群体之间的遗传距离较远，由 STRUCTURE 的结果看出，乌骨绵羊遗传背景单纯，几乎只含有一个祖先成分(少部分个体含有其他群体的祖先成分)，说明乌骨绵羊群体在很长的一段历史中几乎没有其他绵羊群体的血缘渗入，乌骨绵羊的独特品种特点可能是其在漫长的进化历史中独立产生的. 西藏的两个藏羊群体 TIB 和 CHA 含有部分乌骨绵羊成分，说明乌骨绵羊的祖先可能曾有部分向藏区迁入，造成这两个西藏绵羊群体含有少量的乌骨绵羊祖先群体的血缘.

同一地区或地理位置相近的绵羊群体一般遗传结构相近或有共同的祖先来源，如西藏的 CHA和 TIB 群体，孟加拉的 BGE、BGA 群体与印度的

IGA 群体．印度德干地区的 DEC 是混合祖先来源群体，它与多个临近的群体之间含有相同的祖先成分，说明 DEC 群体与周围地区的多个绵羊群体存在广泛的基因交流．

3.2 *Fst* 与 I_n 的关系

Fst 目前被广泛应用于分析多个物种的群体遗传结构和分化位点的筛选[26]．Weir 等[21](1984 年)提出 *Fst* 的简化定义式(MSP-MSG)/[MSP +(n_c-1)MSG]应用较为普遍，其中，MSG 和 MSP 分别为群体内和群体间观测到的位点频率均方差，n_c 指校正后的群体间平均样本大小．*Fst* 的理论取值范围应为 0～1，但是在计算中会出现一些位点的 *Fst* 取值小于 0，这是因为抽样误差造成的，低于 0 的 *Fst* 值可以认为取值为 0.

不同群体经历不同的遗传漂变、迁移和选择等进化过程，导致某些位点上的基因频率存在很大差异，这些位点的 *Fst* 值会显著高于平均水平．因而 *Fst* 值也常用来进行群体间分化位点的筛选．Rosenberg 等[10]认为，理想的 AIMs 应该是在一些群体中已经固定，但在其他群体中仍有多态性的标记，据此发展了 Informativeness(I_n)的计算公式．根据其算法，I_n 最小取值为 0，此时所有群体中的等位基因频率相同；$N>=K$ 且等位基因只出现在一个群体中，I_n 可以取到理论最大值 log$K$．Rosenberg 等对几种常用来筛选祖先信息标记的方法进行比较，认为 I_n 统计量更适合用来筛选 AIMs位点[10].

本研究在 9 个绵羊群体中，分别计算每个 SNP 位点的 *Fst* 和 I_n 值，并绘制散点图和密度分布图，*Fst* 与 I_n 有很好的相关性，Pearson 相关系数为 0.95 (图 6a)，概率密度分布相似(图 6b, 6c)．I_n 值较高的位点其 *Fst* 取值一般也较高，但位点在两种统计量中的排序存在差异．比如图 4i 中，以 I_n 值从大到小排序的第 29 个 SNP(rs413874690)，其按 Fst 大小排序为 422 位，孟加拉的 BGE、BGA 和印度的 IGA 群体在该位点上的等位基因频率为 0；而在乌骨绵羊 BLB 群体和印度尼西亚 GAR 群体中的等位基因频率较高，分别为 0.66 和 0.60．I_n 比 *Fst* 统计量更适合筛选在某些群体中已经固定，而在另一些群体中具有多态性的分化标记位点．因而在多个群体间筛选遗传分化位点时，I_n 方法可以作为 *Fst* 方法的补充．

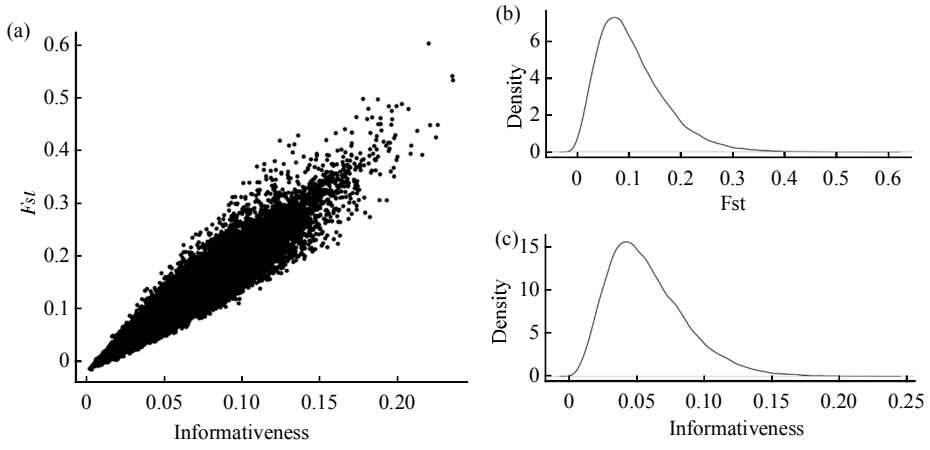

Fig. 6　Comparison of distribution of informativeness (*I*)　and *Fst* statistics

(a) The correlation relationship between I_n and Fst, Pearson's correlative coefficient =0.95. Density distributions of *Fst* (b) and I_n (c) are similar.

3.3 应用 AIMs 分析绵羊群体结构

利用 OvineSNP50 芯片对个体进行全基因组 SNP 分型，进而估计群体结构，所需的成本较高．如果在某些遗传问题的研究中，通过检测较少的标记而推断出群体效应的影响，就可以利用较少的研究成本而完成研究要求．Rosenberge 等[10]利用 I_n 统计量筛选了数十个 AIMs，通过对已知分类的人群数据进行聚类分析，得到了人群以各个大洲分别聚类的效果．Kosoy 等[27]利用 I_n 统计量构建了一个 128 个 AIMs 的集合，用以分析不同人群的群体结构，进一步验证了在关联分析中利用少量的 AIMs 可以校正群体结构．

利用筛选到的高信息含量标记 AIMs，可减少群体结构研究中需要的 SNP 位点数目．本研究中，50 个 AIMs 就可以将 9 种绵羊群体归为 4 大类群，乌骨绵羊群体 BLB 单独归为一类，西藏绵羊群体 TIB 和 CHA 归为一类，孟加拉 BGE、BGA 和印度的 IGA 归为一个类群(这三个群体地理位置接近，BGA 和 IGA 都属于 Garole 品种)，以及印度尼西亚 SUT、GAR 和印度的 DEC 群体.

通过少量 AIMs 推断群体结构的结果与利用全基因组 SNP 位点推断的结果基本吻合，但在推断群体精细结构特点时，需要采用更多的 SNP 标记．通过选取更多的 SNP(上千甚至上万个)位点，才可以得到较为准确的群体结构．随着 SNP 数目增加，无论是 AIMs 还是随机 SNP 集合，其区分群体结构的效力都显著提高，但 AIMs 区分不同群体的能力好于相同标记数目的随机 SNP 集合．由 PCA 和 fastSTRUCTURE 的结果可以看出，AIMs 在 SNP 数目达到 50 的时候，就已经能够很明确区分乌骨绵羊群体与其他的绵羊群体，而用 500 个随机的 SNP 仍未达到理想的区分效果．对于这些群体的其他样本进行群体遗传分析时，仅对这 20～50 个 AIMs 进行检测即可推断群体之间的关系.

通过对 AIMs 的筛选可能会获得一些群体特有的变异．本文对 SNP 位点按照 I_n 值排序，统计每个群体的 AIMs 的频率(图 4i)，乌骨绵羊群体在许多 AIMs 上与其他绵羊群体存在很大频率差异．例如，第 6 位的 SNP(rs403858508)在乌骨绵羊群体中有很高的频率(freq=0.81)，而在其他绵羊群体中频率很低，第 7 位的 SNP(rs408399345)，在乌骨绵羊群体中频率仅为 0.15，但在其他群体中频率很高．这些 AIMs 中也许反映了乌骨绵羊独特的遗传特点，但仍需进一步验证.

4 总　　结

已知群体类别的混合群体中仍然包含有复杂的遗传关系，地理位置接近的群体间一般遗传关系较近．利用 Rosenberg 的 I_n 统计量很有效地筛选高信息含量的位点，以求利用较少的 SNP 位点满足群体遗传结构分析的要求．本研究从 Illumina OvineSNP50 芯片上 5 万多个 SNP 位点中筛选高信息含量的位点，分别构成 20、50、100 及 500 个位点的 AIMs 集合分别用来分析 9 个绵羊群体的遗传结构．通过主成分分析和 STRUCTURE 等分析手段，对本实验中筛选出的不同 AIMs 数目的集合是否能用于群体结构分析进行验证．结果表明，利用 AIMs 分析群体结构与利用全基因组 SNPs 的结果是一致的，且明显优于随机 SNP 集合的群体结构分析效果，说明少量的 AIMs(I_n 值最高的前 20～50 个 AIMs)可以有效地解释群体结构．利用 I_n 统计量筛选 AIMs，更容易筛选到在某些群体中已经固定而在另一些群体中仍有多态性的分化位点，这些 AIMs 也许反映了不同绵羊群体的遗传特点，但仍需进一步验证．本实验中实施的 AIMs 筛选与验证方法同样适用于其他绵羊或自然形成的畜禽群体 AIMs 的筛选工作.

参 考 文 献

[1] Lander E S, Schork N J. Genetic dissection of complex traits. Science, 1994, **265**(5181): 2037-2048

[2] Pritchard J K, Rosenberg N A. Use of unlinked genetic markers to detect population stratification in association studies. American Journal of Human Genetics, 1999, **65**(1): 220-228

[3] Pritchard J K, Stephens M, Rosenberg N A, *et al*. Association mapping in structured populations. American Journal of Human Genetics, 2000, **67**(1): 170-181

[4] Marchini J, Cardon L R, Phillips M S, *et al*. The effects of human population structure on large genetic association studies. Nature Genetics, 2004, **36**(5): 512-517

[5] Helgason A, Yngvadottir B, Hrafnkelsson B, *et al*. An Icelandic example of the impact of population structure on association studies. Nature Genetics, 2005, **37**(1): 90-95

[6] Campbell C D, Ogburn E L, Lunetta K L, *et al*. Demonstrating stratification in a European American population. Nature Genetics, 2005, **37**(8): 868-872

[7] Van Den Wildenberg E, Wiers R W, Dessers J, *et al*. A functional polymorphism of the mu-opioid receptor gene (OPRM1) influences cue-induced craving for alcohol in male heavy drinkers. Alcoholism, Clinical and Experimental Research, 2007, **31** (1): 1-10

[8] Haerian B S, Haerian M S. OPRM1 rs1799971 polymorphism and opioid dependence: evidence from a meta-analysis. Pharmacogenomics, 2013, **14**(7): 813-824

[9] Falush D, Stephens M, Pritchard J K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics, 2003, **164**(4): 1567-1587

[10] Rosenberg N A, Li L M, Ward R, *et al*. Informativeness of genetic markers for inference of ancestry. American Journal of Human Genetics, 2003, **73**(6): 1402-1422

[11] Kidd J R, Friedlaender F R, Speed W C, *et al*. Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. Investigative Genetics, 2011, **2**(1): 1-13

[12] Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. Theoretical

Population Biology, 2001, **60**(3): 155−166

[13] Devlin B, Roeder K. Genomic control for association studies. Biometrics, 1999, **55**(4): 997−1004

[14] Price A L, Patterson N J, Plenge R M, *et al*. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics, 2006, **38**(8): 904−909

[15] Kang H M, Sul J H, Service S K, *et al*. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics, 2010, **42**(4): 348−354

[16] Pritchard J K, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics, 2000, **155**(2): 945−959

[17] Kijas J W, Lenstra J A, Hayes B, *et al*. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. PLoS Biology, 2012, **10**(2): e1001258

[18] Purcell S, Neale B, Todd-Brown K, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics, 2007, **81**(3): 559−575

[19] Patterson N, Price A L, Reich D. Population structure and eigenanalysis. PLoS Genetics, 2006, **2**(12): e190

[20] Raj A, Stephens M, Pritchard J K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics, 2014, **197**(2): 573−589

[21] Weir B S, Cockerham C C. Estimating F-statistics for the analysis of population-structure. Evolution, 1984, **38**(6): 1358−1370

[22] Rousset F. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Molecular Ecology Resources, 2008, **8**(1): 103−106

[23] Jakobsson M, Edge M D, Rosenberg N A. The relationship between FST and the frequency of the most frequent allele. Genetics, 2013, **193**(2): 515−528

[24] Sharma R C, Arora A L, Narula H K, *et al*. Characteristics of Garole sheep in India. Animal Genetic Resources, 1999, **26**: 57−64

[25] Kijas J W, Townley D, Dalrymple B P, *et al*. A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. PloS One, 2009, **4**(3): e4668

[26] Weir B S, Hill W G. Estimating F-statistics. Annual review of Genetics, 2002, **36**: 721−750

[27] Kosoy R, Nassir R, Tian C, *et al*. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. Human Mutation, 2009, **30**(1): 69−78

Population Structure of 9 Asian Sheep Populations Inferred by AIMs*

ZHANG Yuan-Yuan[1)], HAN De-Ping[1)], DENG Wei-Dong[3)], MAO Hua-Ming[3)],
DENG Xue-Gong[2)**], DENG Xue-Mei[1)**]

([1)] *National Engineering Laboratory for Animal Breeding, China Agricultural University, Beijing* 100193, China;
[2)] *College of Science, Northeastern University, Shenyang* 110819, China;
[3)] *College of Animal Science and Technology, Yunnan Agricultural University, Kunming* 650201, China)

Abstract Ancestral information Markers (AIMs) can be utilized for analysis of population genetic structure. In this study, AIMs were selected from Illumina OvineSNP50 chip in blackbone sheep and other eight kinds of Asian sheep populations by Informativeness-statistic defined by Rosenberg. Then subsets of 20, 50, 100, 500 SNPS loci with higher Informativeness value and the corresponding number of random SNP loci were used to infer the population genetic structure, respectively. Principal component analysis (PCA) and fastSTRUCURE methods evaluated AIMs' role in distinguishing the population structure of these 9 Asian sheep population structure. AIMs screened in our study are usefulness in assigning samples to different genetic groups, helping reducing the number of SNPs in sheep genetic researches. The top 50 AIMs can be used effectively to cluster sheep populations into 4 groups, which is consist with the result by the genome-wide SNPs: 1)blackbone group, 2)changthangi and tibetan group; 3)group of banglandeshi, banglandeshiGarole and IndianGarole, 4) group of sumatran, garut and deccani. Alleles Frequncies of AIMs are significantly different among these sheep populations, inferring these markers may be useful in the genetic evolution analysis of these 9 sheep population.

Key words AIMs, sheep, population structure, principal component anlysis, ancestry inference, *Fst*
DOI: 10.16476/j.pibb.2015.0062

*This work was supported by grants from The National Natural Science Foundation of China(U1136605), Ph.D. Programs Foundation of Ministry of Education of China (20120008110049) and Programs of the Major Project for Cultivation Technology of New Varieties of Genetically Modified Organisms of the Ministry of Agriculture(2011ZX08009-001).
**Corresponding author. Tel: 86-10-62733933
DENG Xue-Mei. E-mail: deng@cau.edu.cn
DENG Xue-Gong. E-mail: dengxuegong@tom.com
Received: March 10, 2015 Accepted: June 3, 2015