遗传 HEREDITAS (Beijing) 2010 年 9 月, 32(9): 921－928
ISSN 0253-9772 www.chinagene.cn

研究报告

DOI: 10.3724/SP.J.1005.2010.00921

随机SNP在全基因组关联研究人群分层分析中的应用

曹宗富[1, 2], 马传香[1, 2], 王雷[1, 2], 蔡斌[1, 2]

1. 生物芯片北京国家工程研究中心, 北京 102206;
2. 博奥生物有限公司, 北京 102206

摘要: 在复杂疾病的全基因组关联研究中, 人群分层现象会增加结果的假阳性率, 因此考虑人群遗传结构、控制人群分层是很有必要的。而在人群分层研究中, 使用随机选择的 SNP 的效果还有待进一步探讨。文章利用 HapMap Phase2 人群中无关个体的 Affymetrix SNP 6.0 芯片分型数据, 在全基因组上随机均匀选择不同数量的 SNP, 同时利用 *f* 值和 Fisher 精确检验方法筛选祖先信息标记(Ancestry Informative markers, AIMs)。然后利用 HapMap Phase3 中的无关个体的数据, 以 *F*-statistics 和 STRUCTURE 分析两种方法评估所选出的不同 SNP 组合对人群的区分效果。研究发现, 随机均匀分布于全基因组的 SNP 可用于识别人群内部存在的遗传结构。文章进一步提示, 在全基因组关联研究中, 当没有针对特定人群的 AIMs 时, 可在全基因组上随机选择 3 000 以上均匀分布的 SNP 来控制人群分层。

关键词: 全基因组关联研究; 人群分层; 祖先信息标记; 随机SNP; Affymetrix SNP 6.0 芯片

Analysis of population stratification using random SNPs in genome-wide association studies

CAO Zong-Fu[1, 2], MA Chuan-Xiang[1, 2], WANG Lei[1, 2], CAI Bin[1, 2]

1. *National Engineering Research Center for Beijing Biochip Technology*, *Beijing* 102206, *China*;
2. *CapitalBio Corporation*, *Beijing* 102206, *China*

Abstract: Since population genetic STRUCTURE can increase false-positive rate in genome-wide association studies (GWAS) for complex diseases, the effect of population stratification should be taken into account in GWAS. However, the effect of randomly selected SNPs in population stratification analysis is underdetermined. In this study, based on the genotype data generated on Genome-Wide Human SNP Array 6.0 from unrelated individuals of HapMap Phase2, we randomly selected SNPs that were evenly distributed across the whole-genome, and acquired Ancestry Informative Markers (AIMs) by the method of f value and allelic Fisher exact test. *F*-statistics and STRUCTURE analysis based on the select different sets of SNPs were used to evaluate the effect of distinguishing the populations from HapMap Phase3. We found that randomly selected SNPs that were evenly distributed across the whole-genome were able to be used to identify the population structure. This study further indicated that more than 3 000 randomly selected SNPs that were evenly distributed across the whole-genome were substituted for AIMs in population stratification analysis, when there were no available AIMs for specific populations.

Keywords: genome-wide association study; population stratification; ancestry informative markers; random SNP; Affymetrix SNP 6.0 array

收稿日期: 2009–11–12; 修回日期: 2010–03–10
基金项目: 国家高技术研究发展计划项目(863 计划)(编号：2009AA022708)资助
作者简介: 曹宗富(1978–), 男, 硕士, 专业方向：统计遗传学。E-mail: zfcao@capitalbio.com
通讯作者: 蔡斌(1976–), 男, 硕士, 研究方向：疾病的遗传机制, 芯片技术在疾病研究、物种检测领域的应用。Tel: 010-80715888; E-mail: bcai@capitalbio.com

在复杂疾病的全基因组关联研究中，人群分层现象会增加结果的假阳性率[1, 2]，考虑人群遗传结构将有助于降低这种假阳性。因此，在大多疾病的全基因组关联研究中，针对人群分层问题，利用主成分分析、基因组对照、STRUCTURE分析等方法，来检出并校正研究人群中可能潜在的人群亚结构。WTCCC(Wellcome Trust Case Control Consortium)对7 种常见疾病的全基因组关联研究中，用多维尺度分析和主成分分析等多种方法来处理人群分层，并剔除了 153 个非欧洲祖先的个体[3]。张学军等[4, 5]在对银屑病和系统性红斑狼疮的全基因组关联研究中，用主成分分析的方法对中国汉族样本进行人群分层分析，识别并剔除离群个体，然后对剩余的样本进行分析，没有发现人群分层的证据。Gudmundsso等[6]在前列腺癌的全基因组关联研究中，利用基因组对照的方法对冰岛的人群分层进行校正。Papassotiropoulos等[7]在对记忆的全基因组关联研究中，对 351 个瑞士个体，基于 318 个不连锁的SNP，用STRUCTURE分析识别出 10 个祖先不同的个体并进行剔除。

祖先信息标记(Ancestry informative markers, AIMs)对检出人群中可能存在的遗传结构具有极大价值。Froguel等[8~10]在进行 II 型糖尿病的全基因组关联研究中，对法国DESIR队列中的 658 个个体，用HapMap人群作参考人群，基于 328 个不同大陆人群的AIMs用STRUCTURE来识别研究人群潜在的遗传结构，并根据每一个个体归属于欧洲人群的祖先系数(Ancestry coefficient)，剔除了 43 个非欧洲祖先个体来控制人群分层。

由于 STRUCTURE 的方法易于操作，结果直观性强，为大多研究者所使用。然而, STRUCTURE 不能用全基因组的 SNP 位点进行分析，目前通常选择一定数目的 AIMs 来寻找人群亚结构。而在实际研究中，很多情况下并没有针对特定研究人群的 AIMs 可供使用，使得 STRUCTURE 应用受到很大的限制。虽然多数 SNP 的多态性在不同祖先人群之间的差异很微弱，然而，足够数量的 SNP 的累积效应却不容忽视。假如基于一定数量随机选择 SNP 的 STRUCTURE 分析可以发现人群亚结构，那么将为 STRUCTURE 在全基因组关联的分层分析提供便利。然而，随机选择的 SNP 在人群分层研究中的效果如何却有待进一步探讨。基于此，我们利用 HapMap 数据，分别选择随机均匀分布于全基因组的不同数目 SNP，并与用 f 值和 Fisher 精确检验方法选择的 AIMs 进行比较，观察它们在寻找遗传结构方面的效果差异，进一步评估随机均匀分布于全基因组的 SNP 用于全基因组关联研究分层分析的可行性。

1 材料和方法

1.1 样本及数据来源

所用样本来自 HapMap[11, 12]，包括 HapMap Phase2 样本和HapMap Phase3 样本。HapMap Phase2 共 4 个人群, 270 个样本，其中 90 个欧洲祖先(CEU)样本、45 个中国北京汉族(CHB)样本、45 个日本东京(JPT)样本、90 个非洲祖先(YRI)样本；HapMap Phase3 中选取同样 4 个人群的样本共 541 个，其中 180 个CEU样本、90 个CHB样本、91 个JPT样本、180 个YRI样本。

HapMap Phase2 样本的原始分型数据由 Affymetrix 公司提供，用 Affymetrix SNP 6.0 芯片进行基因分型。HapMap Phase3 数据来自于 HapMap 网站，所有个体用 Illumina Human1M 和 Affymetrix SNP 6.0 两个平台共同分型。CEU 个体家系信息从 Coriell Cell Repositories 获得。

1.2 方法

1.2.1 数据预处理

HapMap Phase2 样本预处理：过滤掉 CEU 样本和 YRI 样本中具有亲缘关系的子代样本(60 个)，过滤掉常染色体上分型成功率小于 98%的样本(0 个)，最终剩余 210 个无关个体。对 Affymetrix SNP 6.0 芯片分型数据预处理包括：过滤掉 X、Y 和线粒体的 SNP 位点(37 122 个)，过滤掉 210 个样本中分型成功率小于 95%的位点(1 909 个)、MAF(minor allele frequency)小于 0.005 的位点(1 844 个)、无信息量位点(630 个)及在单个人群中偏离 Hardy-Weinberg 平衡(HWE，P＜0.001)的位点(10 311 个)，最终常染色体上剩余 854 749 个 SNP 位点。基于 210 个样本在 854 749 个常染色体位点上的分型数据筛选人群之间的 AIMs。

HapMap Phase3 数据在 HapMap 发布时已做过预处理。首先，剔除与 Phase2 重复的样本，然后根据 CEU 家系图剔除 31 个子代个体和 30 个 YRI 子代

个体, 剩下样本中有 4 个 CHB 个体、2 个 JPT 个体、3 个 CEU 个体、2 个 YRI 个体在 HapMap 的预处理中被剔除, 最终剩余 199 个无关样本的分型数据, 其中包括 56 个 CEU 样本、41 个 CHB 样本、44 个 JPT 样本、58 个 YRI 样本。用这些和 Phase2 完全不同的样本数据评估所选出的 AIMs 区分人群的效果。

1.2.2　AIMs 的筛选方法

分别采用 f 值和 Fisher 精确检验方法筛选人群之间的 AIMs。

利用f值筛选人群的AIMs[13], 首先计算出每个位点在 4 个人群两两之间的f值, 观察f值大于 0.3 的SNP位点情况, 并进一步选取AIMs。f的计算公式为:

$$f=\frac{(u_x-u_y)^2}{4\bar{u}(1-\bar{u})},\quad \bar{u}=\frac{u_x+u_y}{2}$$

u_x和 u_y分别为同一个等位基因在 x 和 y 两个人群中的频率。

利用 Fisher 精确检验方法筛选 AIMs, 首先对每个位点进行自由度为 1 的 Fisher 精确检验, 观察人群两两之间 P 值小于 5.85×10^{-8} (Bonferroni 校正后约为 0.05)的 SNP 位点情况, 然后进一步挑选 AIMs。

1.2.3　AIMs 集合和随机 SNP 集合的构成

根据 4 个人群 f 值和 Fisher 精确检验 P 值两两比较结果, 分别获得 6 组数值。将位点按 f 值从大到小排序, 分别取前 10、25、50、100、200 个 SNP 位点, 各组分别合并。将位点按 Fisher 精确检验 P 值从小到大排序, 然后做相同处理。然后将两种方法获得的 SNP 的对应组合分别取交集, 得到不同数目的 AIM 集合, 分别记为 Top10、Top25、Top50、Top100、Top200, 作为所研究的 HapMap 4 个人群的 AIMs。

对 CHB 和 JPT 人群, 分别取 f 值和 Fisher 的精确检验 P 值前 500、1 500、3 000、5 000、10 000 个SNP位点, 然后相同数目的SNP组合分别取交集, 分别记为 Top500、Top1500、Top3000、Top5000、Top10000, 作为 CHB 和 JPT 这两个人群的 AIMs。

根据筛选的 AIM 数目, 在全基因组上选取不同的随机 SNP 集合, 其数目和 AIM 数目大致相同。随机 SNP 的选择方法是, 把所有常染色体基因组均匀划分为多个不同的区域, 然后在每个区域随机选择一个 SNP, 得到一个随机 SNP 集合。对每个随机数目都重复 10 次。

选择文献报道的AIMs包括: Kosoy等[14]报道基于In筛选的区分不同大陆起源洲际人群的 128 个AIMs组成AIMS128; Tian等[15]报道基于In筛选的区分东亚人群的EASTASAIMS, 标记为EASTASAIMS。

1.2.4　效果评估和验证

基于 HapMap Phase3 预处理后的数据, 分别用 F-statistics和STRUCTURE软件分析两种方法评估随机选择的 SNP 组合和筛选的 AIMs 对人群的区分效果。

1.2.4.1　用 F-statistics 方法评估

F-statistics(Fst)方法是基于Weir和Cockerham算法计算的[16]。Fst是表征亚群体间的遗传分化尺度, 可以对不同人群之间遗传关系的远近进行量化。

基于 HapMap Phase3 数据, 用 Fst 方法在 4 个人群中分别评估筛选的 AIMs 组合和随机选择的 SNP 组合区分人群的效果。基于随机选择的 SNP 集合的 Fst 为 10 次重复抽样的平均值。评估的 AIMs 组合包括: Top10、Top25、Top50、Top100、Top200; 评估的随机选择的 SNP 组合包括: Random50、Random120、Random250、Random500、Random1500、Random3000、Random5000、Random10000。随机 SNP 组合名称中的数值表示随机 SNP 的数目。然后分别计算随机 SNP 组合和 AIMs 组合在不同人群之间多个 Fst 的均值, 根据 Fst 均值大小判断不同策略的 SNP 组合识别人群遗传关系的效果。

1.2.4.2　STRUCTURE 软件评估

STRUCTURE 2.3.1[17,18]软件进行人群遗传结构分析是基于Bayesian的聚类方法进行的。本研究利用此软件, 采用混合模型, 设置参数Burn-in≥10000、MCMC≥10000, 假定所有个体都来自于K个人群, K=2~6, 每个K值运行 4 次, 观察结果一致性。所有结果都获得 3 次以上的一致性。用distruct1.1 软件对STRUCTURE的输出结果进行画图。综合人群聚类图和每个人群的祖先系数小于 0.8 的个体所占的比例, 观察区分人群的效果。

基于 HapMap Phase3 数据, 针对 4 个不同大陆起源的洲际人群, 评估 AIMs 组合: Top10、Top25、Top50、Top100、Top200、AIMS128 和随机选择的 SNP 组合: Random50、Random120、Random250、Random500、Random1000、Random3000、Random5000 区分洲际人群的效果。针对 CHB 和 JPT 两个东亚人群, 评估 AIMs 组合: Top500、Top1500、Top3000、

Top5000、Top10000、EASTASAIMS 和随机选择的 SNP 组合: Random500、Random1500、Random3000、Random5000、Random10000 区分洲内人群的效果。

2 结果与分析

2.1 *f*值和Fisher精确检验

基于 HapMap Phase2 预处理后的数据, 对 4 个人群两两之间分别计算 *f* 值并进行 Fisher 精确检验。人群两两比较之间的 *f* 值和 Fisher 精确检验的 *P* 值的负对数高度相关, 相关系数都在 0.98 以上(图 1), 提示 *f* 值和 Fisher 精确检验两种方法具有较强的一致性。*f* 值结果显示, 在 YRI 和其他 3 个人群之间, 全基因组中有 3.50%以上 SNP 的 *f* 值大于 0.3, 其次是 CEU 与亚洲两个人群之间, 其比例在 1.26%以上, 而东亚人群 CHB 和 JPT 之间, 所有 SNP 的 *f* 值都在 0.3 以下。Fisher 精确检验结果显示, 在 YRI 和其他 3 个人群之间达到全基因组显著性水平的 SNP 数目超过 185 000, 占全基因组 SNP 总数的 1/5 以上, CEU 与亚洲两个人群之间则超过 80 000 个, 东亚人群 CHB 和 JPT 之间仅有 7 个 SNPs 达到全基因组显著性水平(表 1)。这两种方法均提示, 在全基因组水平, 有相当大比例 SNP 的等位基因频率在不同大陆起源的洲际人群之间具有较大差异, 而在东亚起源的 CHB 和 JPT 之间差异则相对较小。

表 1 人群之间等位基因频率较大差异的 **SNP** 数量统计

人群比较	$f \geq 0.3$		$^{*}P < 5.85\times10^{-8}$	
	N	比例(%)	*N*	比例(%)
YRI vs. JPT	44843	5.2463	188587	22.0634
YRI vs. CHB	44564	5.2137	186886	21.8644
YRI vs. CEU	29959	3.5050	185865	21.7450
CEU vs. JPT	11264	1.3178	86339	10.1011
CEU vs. CHB	10854	1.2698	83976	9.8246
CHB vs. JPT	0	0.0000	7	0.0008

注: *Fisher 精确检验的 *P* 值; YRI、JPT、CEU、CHB 的含义见图 1。

2.2 AIMs集合和随机SNP集合的构成

4 个人群不同 SNP 数目的 AIM 集合为 Top10、Top25、Top50、Top100 和 Top200, 它们分别包含了 51、113、226、448、855 个 SNP。

CHB 和 JPT 人群的 AIM 集合是 Top500、Top1500、Top3000、Top5000、Top10000, 它们分别包含 459、1 413、2 933、4 904、9 875 个 SNP。

根据筛选的 AIM 数目, 在全基因组上分别选取大致相同数目随机均匀的 SNP 组合 Random50、Random120、Random250、Random500、Random1000、Random1500、Random3000、Random5000、Random10000, 其中组合名称中的数值表示随机 SNP 的数目。对每个随机数目, 均有 10 个由不同 SNP 组成的集合。

2.3 效果评估和验证

2.3.1 用 *F*-statistics 方法评估

图 2 显示了随机 SNP 组合和 AIMs 的 *Fst* 比较。在相同的人群之间, 基于 *f* 值和 Fisher 精确检验筛选的 AIMs 组合的 *Fst*, 高于文献报道计算的 *Fst*(AIMS128), 基于随机选择 SNP 组合的 *Fst* 则要低于所有 AIMs 组合的 *Fst*。同时, 比较每一个 AIMs 或 SNP 组合的 *Fst* 发现, 每一个组合均提示了 4 个人群之间相同的遗传关系, 即 YRI 与其他人群具有最远的遗传祖先, 而东亚人群 CHB 和 JPT 则具有最近的遗传祖先。这些结果共同提示, 一定数目随机均匀分布于全基因组的 SNP 可以识别人群之间存在的遗传结构, 但是其区分人群分层的效果可能不如 AIMs。

2.3.2 STRUCTURE 软件评估

针对 4 个洲际人群的分析发现, 当 k=3~6 时, 基于 AIMs 和全基因组上随机均匀分布的 SNP 都具有区分洲际人群的能力, 洲际人群被聚为与地理位置相对应的 3 个人群, 而 CHB 和 JPT 则不能区分(图 3, k=2、4、5、6 的结果没有显示)。图 4A 显示, 随着 SNP(或 AIMs)数目的增多, 区分人群结构的效果越来越好, 当 AIMs 的数目大于 448(top100)或随机 SNP 数目大于 500 时, 所有个体的最大祖先系数都在 0.8 以上。当 k=4 时, 把随机 SNP 数目增大到 5 000 时, 仍然不能识别 CHB 和 JPT 内部存在的遗传结构(图 5)。

针对 CHB 和 JPT 人群的分析发现, AIMs 和全基因组上随机均匀分布的 SNP 都具有区分祖先起源较近人群的能力。当 k=2 时, 除 Random500 以外的其他组合都能识别两个人群内部存在的遗传结构, 但是区分能力有着较大差异(图 6)。图 4B 显示, 当 SNP 数目超过 1 500 时, 随机选择的 SNP 要比筛选的 AIMs 识别人群结构的效果更好。当随机选择的 SNP

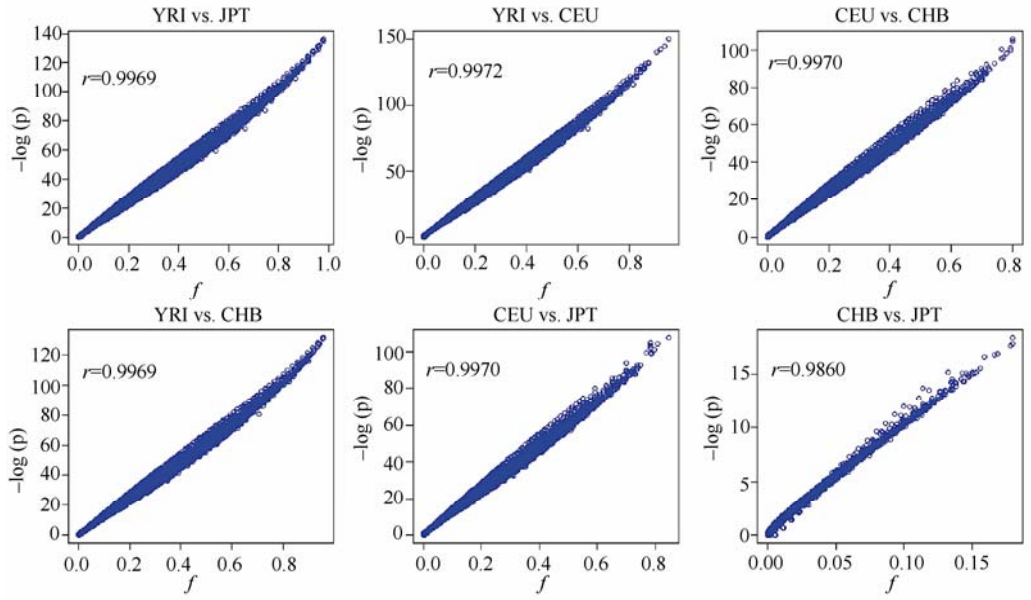

图 1　*f* 值和 Fisher 精确检验两种方法的一致性比较

YRI: 非洲祖先样本; JPT: 日本东京样本; CEU: 欧洲祖先样本; CHB: 中国北京汉族样本。

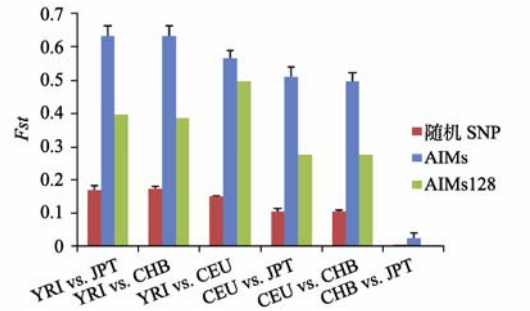

图 2　随机 SNP 和 AIMs 的 *Fst* 比较

随机 SNP 表示随机均匀选择的不同数目的 SNP 组合, AIMs 表示根据 *f* 值和 Fisher 精确检验筛选的不同数目 AIMs 组合，AIMs128 表示文献报道的 AIMs 组合。横轴为 4 个人群的两两比较, YRI、JPT、CEU、CHB 的含义见图 1; 纵轴为 *Fst* 大小。对 AIMs128 系列显示为 *Fst* 大小，而其他两个系列显示为多个 SNP 或 AIM 组合 *Fst* 的平均值，并以标准差表示多个 *Fst* 的变异大小。

为 3 000 时，人群中 92.4%的个体的最大祖先系数都大于 0.8；当 SNP 数目增大到 10 000 时，则 99.7%的个体的最大祖先系数都大于 0.8。

3　讨 论

Fst 和 STRUCTURE 的结果发现，对不同大陆起源的洲际人群，利用本研究筛选的 AIMs 可以很好地识别人群遗传结构。对祖先起源较近的东亚人群 CHB 和 JPT，增大 AIMs 的数目，也可以识别内部的遗传结构。而随机均匀分布于全基因组的 SNP，在区分不同大陆起源的人群以及祖先较近的人群，也可以识别人群内部的遗传结构。当区分不同大陆起源的人群时, 500 个以上的 SNP 就可以获得所有个体

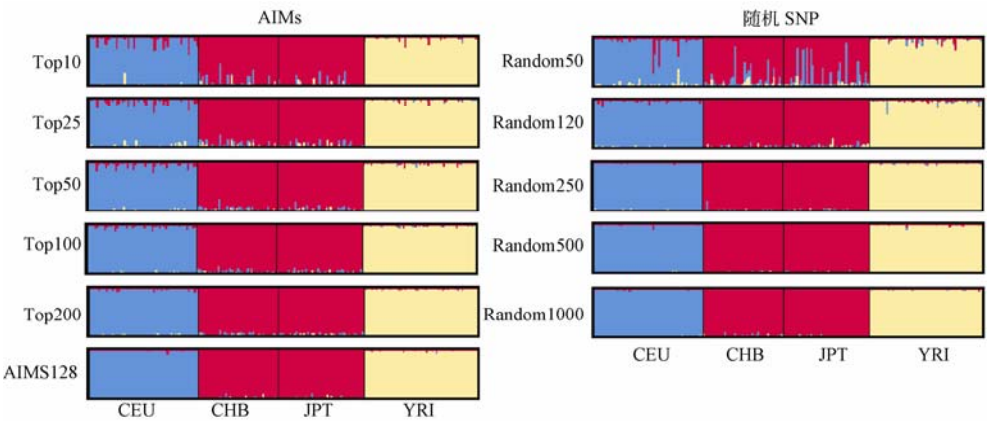

图 3　4 个人群基于不同 SNP 集合的 STRUCTURE 聚类图(k=3)

YRI、JPT、CEU、CHB 的含义见图 1。

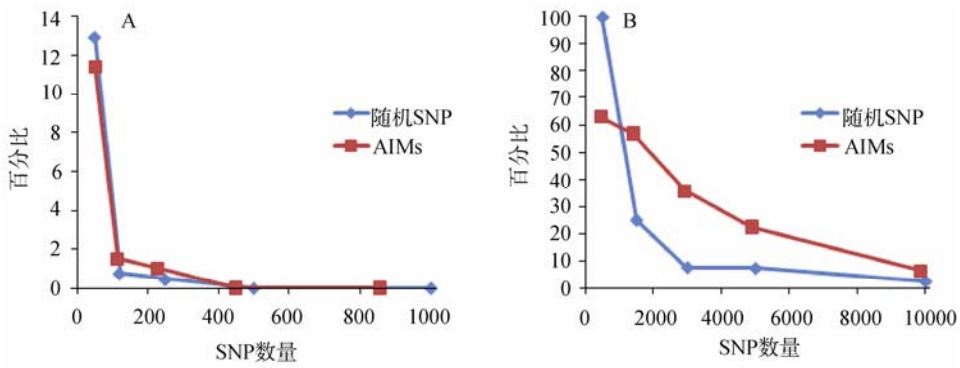

图 4　SNP 数量和 STRUCTURE 个体祖先推断关系图

A: 反映不同数目的 AIMs 和随机分布于全基因组的 SNP 对 4 个洲际人群的识别效果。B: 反映不同数目的 AIMs 和随机分布于全基因组的 SNP 对 CHB(中国北京汉族样本)和 JPT(日本东京样本)两个祖先较近的洲内人群的识别效果。纵轴的百分比是指该人群的祖先系数小于 0.8 的个体所占的百分比。

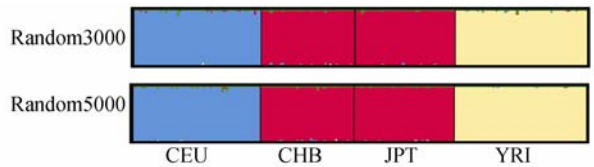

图 5　4 个人群基于不同 SNP 集合的 STRUCTURE 聚类图(k=4)

YRI、JPT、CEU、CHB 的含义见图 1。

的最大祖先系数大于 0.8 的效果; 而当区分祖先起源较近的人群时, 3 000 个 SNP 就可以使 92.4%的个体的最大祖先系数大于 0.8, 10 000 个 SNP 就可使 99.7%的个体的最大祖先系数大于 0.8。这些结果提示, 在全基因组关联性研究中, 无论是祖先起源较远的洲际人群, 还是祖先起源较近的人群之间, 都可以用随机均匀分布的位点来寻找人群潜在的遗传结构。基于随机选择 SNP 组合计算的 *Fst* 要小于基于 AIM 计算的 *Fst*, 提示随机选择 SNP 识别人群内部遗传结构的效果可能不如 AIMs。但从 STRUCTURE 分析的结果来看, 随机选择的 SNP 同样能够识别人群内部的遗传结构。因此, 当有针对特定人群可用的 AIMs 时, 优先选择 AIMs 来识别遗传结构; 当针对特定研究人群并没有可用的 AIMs 时, 随机均匀分布于全基因组的 SNP 则是一种较好的选择。特定研究识别遗传结构时, 需要的 SNP 数目与人群之间遗传分化的大小有关, 祖先起源越近, 需要的 SNP 数目越多。对

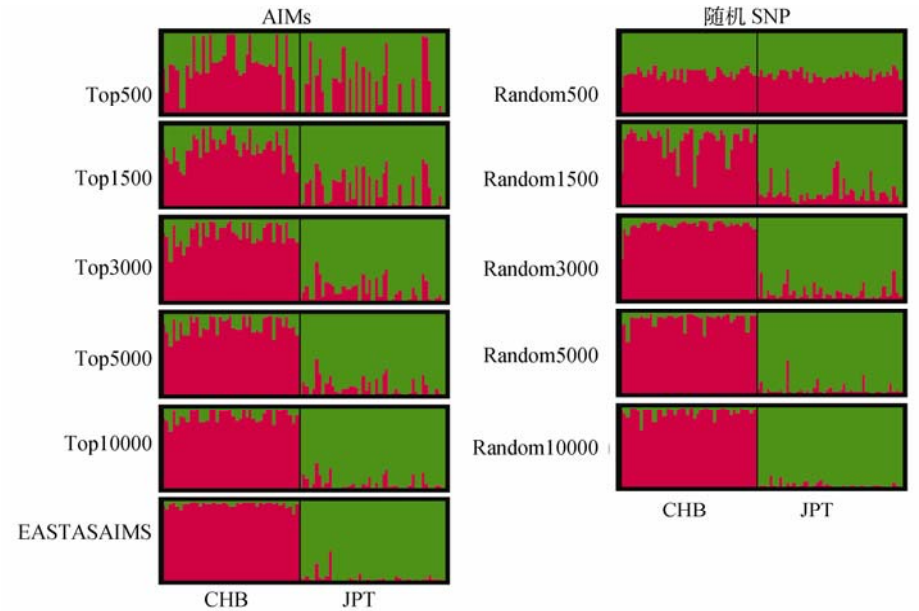

图 6　CHB(中国北京汉族样本)和 JPT(日本东京样本)基于不同 SNP 集合的 STRUCTURE 聚类图

于不同大陆起源的人群, 需要 500 个以上的 SNP; 对祖先起源较近的人群, 至少需要 3 000~10 000 个 SNP。同时, 为了保证结果的可靠性, 可增加不同 SNP 数目进行重复验证, 以观察结果的一致性。

*f*值和Fisher精确检验结果提示, CHB和JPT之间等位基因频率具有较大差异的SNP较少, 同时, 本研究对不同SNP组合的*Fst*分析结果均提示, CHB和JPT具有较近的祖先, 该结果也被Rosenberg等[19]用全基因组的微卫星数据聚类得到证实。当把 4 个人群进行STRUCTURE分析时, 任何SNP组合都不能识别东亚人群CHB和JPT间的遗传结构。然而, 当把CHB和JPT单独分析时, 增加AIMs或随机选择的SNP数目, 或者利用EASTASIAAIMS, 却能够发现内部存在的遗传结构。Miao等[20]利用Y-STR单体型和HapMap Phase2 的常染色体SNP数据进行STRUCTURE分析也获得了相似的结果。这些结果说明, 虽然CHB和JPT具有较近的遗传祖先, 但其内部存在的遗传结构仍然可以识别, 只是与祖先较远的人群一起进行分析时, 难以识别内部存在的遗传结构。这些结果提示, 人群之间较大的遗传分层使得较小的遗传分层难以识别。在全基因组关联性研究中, 当对祖先较近人群进行人群分层研究时, 目前常常采用不同大陆起源的 HapMap 人群做参考, 需要慎重对待, 还需要进一步选用祖先较近的人群做参考人群单独进行分层分析。

比较 *Fst* 和 STRUCTURE 两种方法的结果, 可以发现两种方法的结果有一定的差异。譬如, *Fst* 结果显示 AIM 识别人群结构的效果要比随机均匀选择的 SNP 效果更好, 但是 STRUCTURE 的结果显示, 随机均匀选择的 SNP 识别人群结构的效果并不比筛选的 AIM 效果差。这些结果反映了两种方法在算法上的差异, 每种算法都可能有其局限性, 因此识别人群分层时常需要多种方法相互验证。尤其是单独对CHB和JPT分析时, 当SNP数目超过1 500时, 随机选择的 SNP 要比筛选的 AIMs 识别人群结构的效果更好, 可能是由于 STRUCTURE 软件算法或抽样样本量太小引起的, 该问题有待进一步探讨。

Fisher 精确检验结果显示在不同祖先人群之间等位基因频率具有显著差异的位点最高可占全基因组的20%以上, 提示遗传结构对关联分析的影响是巨大的。同时, 对祖先较近人群 CHB 和 JPT 之间, 达到全基因组显著性的也有 7 个 SNPs, 对关联研究也具有很大影响。这些结果提示关联研究中对不同大陆起源的洲际人群的遗传结构检出非常重要, 即使祖先起源较近人群, 其遗传亚结构的影响也依然不容忽视。

同时, 本研究筛选出来的人群之间具有全基因组显著性的位点, 可为关联研究提供参考。在不同祖先起源的全基因组关联研究中, 当检出与疾病关联的SNP中含有与本研究筛选出的位点时, 应谨慎对待。在GWAS中, 国际合作越来越多, 同时也有更多的GWAS数据可以共享使用[3,21]。当多个不同大陆起源洲际人群的复杂样本联合使用时, 人群遗传结构是必然存在的。因此, 本研究对复杂样本的关联研究也具有参考价值。

参考文献**(References)**:

[1] Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet*, 2005, 6(2): 109–118.[DOI]

[2] Balding DJ. A tutorial on statistical methods for population association studies. *Nat Rev Genet*, 2006, 7(10): 781–791.[DOI]

[3] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14 000 cases of seven common diseases and 3 000 shared controls. *Nature*, 2007, 447(7145): 661–683.[DOI]

[4] Zhang XJ, Huang W, Yang S, Sun LD, Zhang FY, Zhu QX, Zhang FR, Zhang C, Du WH, Pu XM, Li H, Xiao FL, Wang ZX, Cui Y, Hao F, Zheng J, Yang XQ, Cheng H, He CD, Liu XM, Xu LM, Zheng HF, Zhang SM, Zhang JZ, Wang HY, Cheng YL, Ji BH, Fang QY, Li YZ, Zhou FS, Han JW, Quan C, Chen B, Liu JL, Lin D, Fan L, Zhang AP, Liu SX, Yang CJ, Wang PG, Zhou WM, Lin GS, Wu WD, Fan X, Gao M, Yang BQ, Lu WS, Zhang Z, Zhu KJ, Shen SK, Li M, Zhang XY, Cao TT, Ren W, Zhang X, He J, Tang XF, Lu S, Yang JQ, Zhang L, Wang DN, Yuan F, Yin XY, Huang HJ, Wang HF, Lin XY, Liu JJ. Psoriasis genome-wide association study identifies susceptibility variants within LCE gene cluster at 1q21. *Nat Genet*, 2009, 41(2): 205–210.[DOI]

[5] Han JW, Zheng HF, Cui Y, Sun LD, Ye DQ, Hu Z, Xu JH, Cai ZM, Huang W, Zhao GP, Xie HF, Fang H, Lu QJ, Xu JH, Li XP, Pan YF, Deng DQ, Zeng FQ, Ye ZZ, Zhang XY, Wang QW, Hao F, Ma L, Zuo XB, Zhou FS, Du WH, Cheng YL, Yang JQ, Shen SK, Li J, Sheng YJ, Zuo XX, Zhu WF, Gao F, Zhang PL, Guo Q, Li B, Gao M, Xiao FL, Quan C, Zhang C, Zhang Z, Zhu KJ, Li Y, Hu DY, Lu WS, Huang JL, Liu SX, Li

H, Ren YQ, Wang ZX, Yang CJ, Wang PG, Zhou WM, Lv YM, Zhang AP, Zhang SQ, Lin D, Li Y, Low HQ, Shen M, Zhai ZF, Wang Y, Zhang FY, Yang S, Liu JJ, Zhang XJ. Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat Genet*, 2009, 41(11): 1234–1237.[DOI]

[6] Gudmundsson J, Sulem P, Manolescu A, Amundadottir LT, Gudbjartsson D, Helgason A, Rafnar T, Bergthorsson JT, Agnarsson BA, Baker A, Sigurdsson A, Benediktsdottir KR, Jakobsdottir M, Xu J, Blondal T, Kostic J, Sun J, Ghosh S, Stacey SN, Mouy M, Saemundsdottir J, Backman VM, Kristjansson K, Tres A, PartinAW, Albers- Akkers MT, Godino-Ivan Marcos J, Walsh PC, Swinkels DW, Navarrete S, Isaacs SD, Aben KK, Graif T, Cashy J, Ruiz-Echarri M, Wiley KE, Suarez BK, Witjes JA, Frigge M, Ober C, Jonsson E, Einarsson GV, Mayordomo JI, Kiemeney LA, Isaacs WB, Catalona WJ, Barkardottir RB, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat Genet*, 2007, 39(5): 631–637.[DOI]

[7] Papassotiropoulos A, Stephan DA, Huentelman MJ, Hoerndli FJ, Craig DW, Pearson JV, Huynh KD, Brunner F, Corneveaux J, Osborne D, Wollmer MA, Aerni A, Coluccia D, Hänggi J, Mondadori CR, Buchmann A, Reiman EM, Caselli RJ, Henke K, de Quervain DJ. Common Kibra alleles are associated with human memory performance. *Science*, 2006, 314(5798): 475–478.[DOI]

[8] Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 2007, 445(7130): 881–885.[DOI]

[9] Bouatia-Naji N, Bonnefond A, Cavalcanti-Proença C, Sparsø T, Holmkvist J, Marchand M, Delplanque J, Lobbens S, Rocheleau G, Durand E, De Graeve F, Chèvre JC, Borch-Johnsen K, Hartikainen AL, Ruokonen A, Tichet J, Marre M, Weill J, Heude B, Tauber M, Lemaire K, Schuit F, Elliott P, Jørgensen T, Charpentier G, Hadjadj S, Cauchi S, Vaxillaire M, Sladek R, Visvikis-Siest S, Balkau B, Lévy-Marchal C, Pattou F, Meyre D, Blakemore AI, Jarvelin MR, Walley AJ, Hansen T, Dina C, Pedersen O, Froguel P. A variant near MTNR1B is associated with increased fasting plasma glucose levels and type 2 diabetes risk. *Nat Genet*, 2009, 41(1): 89–94.[DOI]

[10] Bouatia-Naji N, Rocheleau G, Van Lommel L, Lemaire K, Schuit F, Cavalcanti-Proença C, Marchand M, Hartikainen AL, Sovio U, De Graeve F, Rung J, Vaxillaire M, Tichet J, Marre M, Balkau B, Weill J, Elliott P, Jarvelin MR, Meyre D, Polychronakos C, Dina C, Sladek R, Froguel P. A polymorphism within the G6PC2 gene is associated with fasting plasma glucose levels. *Science*, 2008, 320(5879): 1085–1088.[DOI]

[11] The International HapMap Consortium. The International HapMap Project. *Nature*, 2003, 426(6968): 789–796.[DOI]

[12] Thorisson GA, Smith AV, Krishnan L, Stein LD. The International HapMap Project Web site. *Genome Res*, 2005, 15(11): 1592–1593.[DOI]

[13] McKeigue PM. Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. *AM J Hum Genet*, 1998, 63(1): 241–251 [DOI]

[14] Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, De La Vega FM, Seldin MF. Ancestry informative marker sets for determining continental origin and admixture proportions in common populationns in America. *Hum Mutat*, 2009, 30(1): 69–78.[DOI]

[15] Tian C, Kosoy R, Lee A, Ransom M, Belmont JW, Gregersen PK, Seldin MF. Analysis of East Asia genetic substructure using genome-wide SNP arrays. *PLoS One*, 2008, 3(12): e3862.[DOI]

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

[17] Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*, 2000, 155(2): 945–959.

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

[19] Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. Genetic structure of human populations. *Science*, 2002, 298(5602): 2381–2385.[DOI]

[20] He M, Gitschier J, Zerjal T, de Knijff P, Tyler-Smith C, Xue YL. Geographical affinities of the HapMap samples. *PLoS ONE*, 2009, 4(3): e4684.[DOI]

[21] Myocardial Infarction Genetics Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet*, 2009, 41(3): 334–341.[DOI]