

利用群体基因组学策略扫描人群特异的自然选择信号

邓立彬^{①②} 汤晓丽^① 康 健^③ 王青芸^④ 曾长青^{①*}

(① 中国科学院北京基因组研究所, 北京 101300; ② 中国科学院研究生院, 北京 100049; ③ 清华大学理学院数学系, 北京 100084;

④ 北京师范大学数学科学学院, 北京 100875. * 联系人, E-mail: czeng@genomics.org.cn)

摘要 自然选择是人类进化的动力, 直接影响人群分化. 目前对于自然选择在人群分化中的作用及其功能还知之甚少. 为了探讨这一问题, 本研究利用人类基因组国际单体型图计划的数据分析人群的遗传分化. 通过群体基因组学策略扫描常染色体区, 共发现了 12669 个高人群分化的 SNPs 和 1853 个人群特异的自然选择“候选基因”, 并通过基因注释确定了 121 种受到强选择压力的生物学过程. 多层次分析显示, 人类基因组中存在普遍的正选择信号, 这些信号聚集于特定的染色体区域; 且多数“候选基因”和相应的生物学过程均局限于特定人群中, 提示自然选择在人群分化中发挥了重要作用. 本研究为从群体遗传和功能学角度深入研究自然选择和人群分化提供了新的线索.

关键词 自然选择 人类基因组单体型图计划 单核苷酸多态性 群体基因组学 基因注释

自然选择(natural selection)是达尔文进化的动力, 可对生物表型产生多种影响. 利用遗传学方法研究自然选择一直是生物学的重要方面, 不但可增进对生物进化、表型多样性的认识, 还将有助于鉴定基因组中重要的功能区域. 正选择(positive selection)是人类历史中自然选择的主要类型, 在人群分化和表型差异形成中发挥了重要作用^[1]. 近来许多基因组范围的自然选择研究均提示, 在人类基因组中正选择信号比先前所认为的更加普遍^[2-6]. 据此推测, 人类群体在过去 10 万年急剧扩增中受到了较强的地区性选择压力(如独特的饮食、疾病和气候的变化等)^[7]. 但目前对于正选择在人群分化中的作用及功能背景还缺乏系统了解.

传统的自然选择研究由于缺乏遗传多态数据仅限于分析少数基因, 难以在基因组的层次上进行系统探索. 近年来随着单核苷酸多态性(single nucleotide polymorphism, SNP)研究的深入, 特别是全基因组规模 SNP 分型计划的完成和高通量 SNP 芯片的推出^[2,8,9], 群体基因组学(population genomics)作为一种新的分析策略被应用于自然选择研究^[10,11]. 近期多项研究均验证了该策略的可靠性^[3,4,7,12]. 与传统的单基因分析比较, 基因组水平的群体遗传分析能更精确地发现自然选择. 此外, 利用基因组数据系统地扫描自然选择的“候选基因”还可发现基因的新的生物学功能^[3,4], 为基因功能和进化研究提供线索.

2002 年, Akey 等人^[6]首先根据群体基因组学策

略对自然选择进行全基因组扫描. 通过分析 26530 SNP 位点在人群间的频率差异, 确定基因区域 F_{ST} 的“离群值”(outliers), 发现了 156 个正选择“候选基因”. 后续研究还提示正选择的地区性(geographic restriction)可能推动了人群分化^[1,5]. 然而, 目前这类分析只利用了少量遗传标记, 不但提供的基因信息非常有限, 而且难以消除测量偏倚(ascertainment bias)(如中高频位点过多)对结果的影响. 人类基因组国际单体型图计划(the International Haplotype Map Project, HapMap)的完成^[12], 不仅为基因组范围的疾病相关分析提供了高密度的 SNP 分型数据, 还为系统分析自然选择提供了新的契机. 最近, Wang 等人^[13]和 Voight 等人^[14]利用 SNPs 间的连锁不平衡(linkage disequilibrium, LD)分布发现了基因组中一系列近期正选择(recent positive selection)信号, 并印证了正选择的地区性, 提示人群间自然选择“候选基因”的功能学背景存在差异.

为了在基因组范围系统探讨正选择在人群分化中的作用及其功能背景, 我们利用 HapMap 数据分析亚、非、欧洲人群的遗传分化, 在常染色体区域通过群体基因组学策略发现高人群分化的 SNP 以及人群特异的自然选择(population-specific natural selection)“候选基因”, 从而在基因组水平评价自然选择对人群分化的影响; 并以此建立多层次的自然选择图谱; 结合基因注释信息分析正选择在人群分化过程中的功能基础; 通过比较 HapMap 人群间“候选基因”

的基因组分布和功能背景的差异探讨自然选择在人群分化中的作用。

1 材料与方法

(i) HapMap样品的分型数据. 全常染色体区域的 SNP 分型数据从 HapMap 网站下载 (HapMap Phase I/rel#16c, <http://www.hapmap.org/>)^[13]. HapMap 样品包括亚、非、欧三个种群, 由 89 个亚洲随机个体(45 个中国汉族人(北京)和 44 个日本人(东京))、30 个非洲(尼日利亚-伊巴丹的约鲁巴部落)和 30 个欧洲(祖先来自北欧和西欧的犹他州居民)三联家系组成^[14]. 本研究分析了其中的 209 个无关个体, 包括 89 个亚洲、60 个欧洲和 60 个非洲样本, 根据 HapMap 定义和国际常用惯例, 分别用 ASN, CEU 和 YRI 表示。

本研究采用的分型数据质量控制标准为: 在各 HapMap 人群中, (1) 最小等位基因频率(minor allele frequency, MAFs)大于 1%; (2) 达到 Hardy-Weinberg 平衡($P > 0.001$); (3) 分型成功率大于 80%. 在 HapMap 项目 I 期的约 1 百万个分型数据中, 常染色体区域共有 629958 个位点通过质控标准(各常染色体的分布见表 1)。

(ii) 单位点群体遗传学分析. 每个 SNP 在两两人群间的等位基因频率差异采用单位点配对 F_{ST} (pairwise F_{ST}) 衡量^[15]:

$$\hat{\theta} = \frac{\sum_i \left\{ \frac{1}{2} \sum_u (\tilde{p}_{lu_i} - \tilde{p}_{lu_2})^2 - \frac{1}{(2(2n-1))} [2 - \sum_u (\tilde{p}_{lu_i}^2 + \tilde{p}_{lu_2}^2)] \right\}}{\sum_i (1 - \sum_u \tilde{p}_{lu_i} \tilde{p}_{lu_2})}, \quad (1)$$

其中, l 表示所分析的 SNP 位点, u_i 表示 SNP 在第 i 个人群中第 u 个等位基因; 则 p_{lu_i} 表示 SNP 的第 u 个等位基因在第 i 个人群中的频率. 当仅考虑一个位点时, 公式(1)可简化为:

$$F_{ST} = \frac{\frac{2n}{2(2n-1)} \sum_u \tilde{p}_{u_1}^2 + \frac{2n}{2(2n-1)} \sum_u \tilde{p}_{u_2}^2 - \sum_u \tilde{p}_{u_1} \tilde{p}_{u_2} - \frac{1}{2n-1}}{1 - \sum_u \tilde{p}_{u_1} \tilde{p}_{u_2}}. \quad (2)$$

单位点 F_{ST} 的计算在 Matlab 7.0(The Mathworks, Natick, MA)中进行。

(iii) 单位点 F_{ST} “离群值”的确定. 为了确定单位点 F_{ST} 的“离群值”, 采用 boxplot 策略^[16] 确定各指标的上、下分界值(threshold). 首先计算各染色体 F_{ST} 经验分布的四分位区间(inter-quartile range, IQR):

$$Q = F^U - F^L, \quad (3)$$

其中, F^L 和 F^U 分别是上、下四分位值. 则上、下分界值为:

$$UL = F^U + 1.5Q, \quad (4)$$

$$LL = F^L - 1.5Q. \quad (5)$$

所有大于上分界值(UL)或小于下分界值(LL)的位点被定义为“离群”位点。

(iv) SNP 位点的基因定位. 根据 EntrezGene (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene>) 的搜索结果, 将 629958 个位点定位到基因区域(包括 5'UTR 区、编码区、内含子序列或 3'UTR 区). 在基于人类基因组第 16 版拼接图谱(2003 年 7 月)的已知基因数据库中, 17422 个已知基因存在至少一个已分型的位点。

(v) 基因诠释分析. 根据 PANTHER 基因诠释数据库^[17] (<http://www.pantherdb.org/>) 的搜索结果对 17422 个已知基因的生物学过程(biological process)进行诠释. 其中每个基因可以同时归属于不同的生物学过程。

(vi) 中性溯祖模拟. 根据溯祖理论(coalescent theory)模拟特定中性进化条件下的 SNPs 分型数据. 利用 COSI (<http://www.broad.mit.edu/~sfs/cosi>) 在最适参数条件下模拟 50 Mb 区域内 SNPs 分型数据^[18].

(vii) 统计学分析. 本研究中所有的统计学分析均在 Matlab 7.0 (The Mathworks, Natick, MA) 中进行, 统计学显著性精确到 10^{-12} .

2 结果与分析

2.1 常染色体区域的自然选择信号

在人类基因组常染色体区域, 我们利用 F_{ST} 衡量 SNP 在 HapMap 人群间的频率差异, 并构建 F_{ST} 的经验分布(图 1). 在 3 对人群比较(亚洲与欧洲、亚洲与非洲以及欧洲与非洲人群)中, 单位点 F_{ST} 经验分布与自由度为 2(或 3)的 χ^2 分布很相似, 在高 F_{ST} 值区域存在明显拖尾. 在亚洲与欧洲、亚洲与非洲以及欧洲与非洲人群比较中, 分别有 4%, 11%, 9% 的 F_{ST} 值大于 0.4, 极高 F_{ST} ($F_{ST} > 0.4$) 的 SNP 比例明显多于中性溯祖模型模拟的数据(数据未列出), 提示 HapMap 数据受到地区性正选择的影响, 已偏离了中性假设. 为进一步验证这种推测, 我们分析了单位点 F_{ST} 在各常染色体上的分布情况(网络版附表 1)。

虽然各常染色体上 F_{ST} 的数值显示出极大的差异(标准差较大)(网络版附表 1), 但间隔越小的位点间 F_{ST} 数值越相近, 提示邻近位点比间隔更远的位点共享更多的进化历史. 因此, 我们计算了不同多态位点间隔的 SNP 对之间 F_{ST} 数值的相关性. 如图 2 所示,

表 1 22 条常染色体上极大 F_{ST} “离群值”的分界值和位点数目

染色体	上分界值 ^{a)}			“离群”位点个数 ^{b)}		
	亚洲/欧洲	亚洲/非洲	欧洲/非洲	亚洲/欧洲	亚洲/非洲	欧洲/非洲
1	0.350	0.585	0.510	2372*	1556**	1909**
2	0.375	0.595	0.535	2857**	1978**	1801*
3	0.350	0.595	0.510	1922**	1287*	1613*
4	0.325	0.545	0.535	1815*	1636*	1224*
5	0.350	0.560	0.485	1628*	1390*	1461**
6	0.325	0.585	0.460	2451**	1342**	1945**
7	0.350	0.535	0.485	1209*	1198*	1230*
8	0.350	0.595	0.520	2095**	1449*	1849**
9	0.350	0.535	0.485	1892*	1487*	1547*
10	0.375	0.560	0.510	1475*	1152*	1093*
11	0.325	0.560	0.485	1447*	942*	1102*
12	0.350	0.560	0.485	1295*	1005*	1219*
13	0.325	0.545	0.460	1395*	825*	961*
14	0.375	0.535	0.535	770*	718*	641*
15	0.385	0.585	0.570	826*	478*	505*
16	0.350	0.545	0.535	753*	622*	394
17	0.350	0.620	0.535	675*	467*	565*
18	0.300	0.535	0.485	1256*	773*	880*
19	0.760	0.980	0.955	550	18	37
20	0.695	0.980	0.805	429	20	222
21	0.735	0.905	0.905	711	114	82
22	0.350	0.585	0.485	499*	374*	439
合计	0.350	0.570	0.540	30322	20831	22719

a) 各常染色体上确定高人群分化离群位点的上分界值. b) *示 $P < 10^{-3}$, **示 $P < 10^{-12}$, 各常染色体上离群位点分布的随机性检验的显著性水平

当 SNP 对的间隔为 1~50 个位点时, F_{ST} 数值的相关程度(r^2)随着 SNP 间隔增大而降低. 在其他常染色体中我们也发现了类似的相关模式(数据未列出), 间隔小于 10 个 SNP(物理距离 30 kb 左右)的位点间 F_{ST} 数值高度相关($P < 0.001$), 且当间隔扩大到 100 个位点以上时(物理距离超过 200 kb)这种显著相关性($P < 0.01$)依然存在. 这种 SNPs 之间 F_{ST} 数值的统计学相关性反映了基因组中位点间的连锁不平衡, 而长距离的强相关性可能由正选择的“搭便车”效应(hitchhiking)和/或平衡选择所引起. 这些结果提示 HapMap 计划第 I 期的分型密度(1SNP/3kb)能确保在人类基因组中扫描自然选择的信号.

2.2 高人群遗传分化的 SNP 位点

在常染色体区域 F_{ST} 经验分布的基础上(图 1), 我们进一步运用 boxplot 方法鉴定自然选择的“候选位点”. 在上、下分界值的设定上, 考虑到各染色体的单位点 F_{ST} 与整个常染色体区的分布存在显著差异(网络版附表 1). 为了获得稳健(robust)结果, 我们采用了各染色体特定的分界值来确定相应的“离群”位点.

在人群遗传分化分析中, 地区性正选择可以导致目标位点 F_{ST} 增大, 而平衡选择可以导致 F_{ST} 减小 [7]. 表 1 显示极高 F_{ST} 的“离群”位点在各染色体上的详细分布, 这些位点可能受到了地区性正选择的影响. “离群”位点分界值的经验性显著水平约为 0.03(0.0015 ~ 0.0593). 在常染色体区域我们共确定了 61203 个“离群”位点, 约占总位点数的 10%. 这些候选位点在 3 对比较中至少有一个极高的 F_{ST} 值, 其中在亚洲与欧洲人群比较中离群位点最多, 共 30322 个 SNPs; 亚洲与非洲人群比较中有 20831 个, 欧洲与非洲人群比较中有 22719 个位点.

同时, 这些离群位点在染色体中有聚集分布(clustering)的趋势. 为了检验这种趋势, 在每条染色体上我们假设“离群值”出现的位置是随机的, 那么相邻“离群值”之间的距离大于或小于所有相邻“离群值”之间距离平均值的几率相等. 因此, 如果小于距离均值的相邻“离群值”对的个数显著增多(χ^2 检验, $df = 1$, $P < 10^{-3}$), 就表明这些“离群值”有聚集的趋势. 基于这种分布随机检验(randomness test), 我们发现“离群

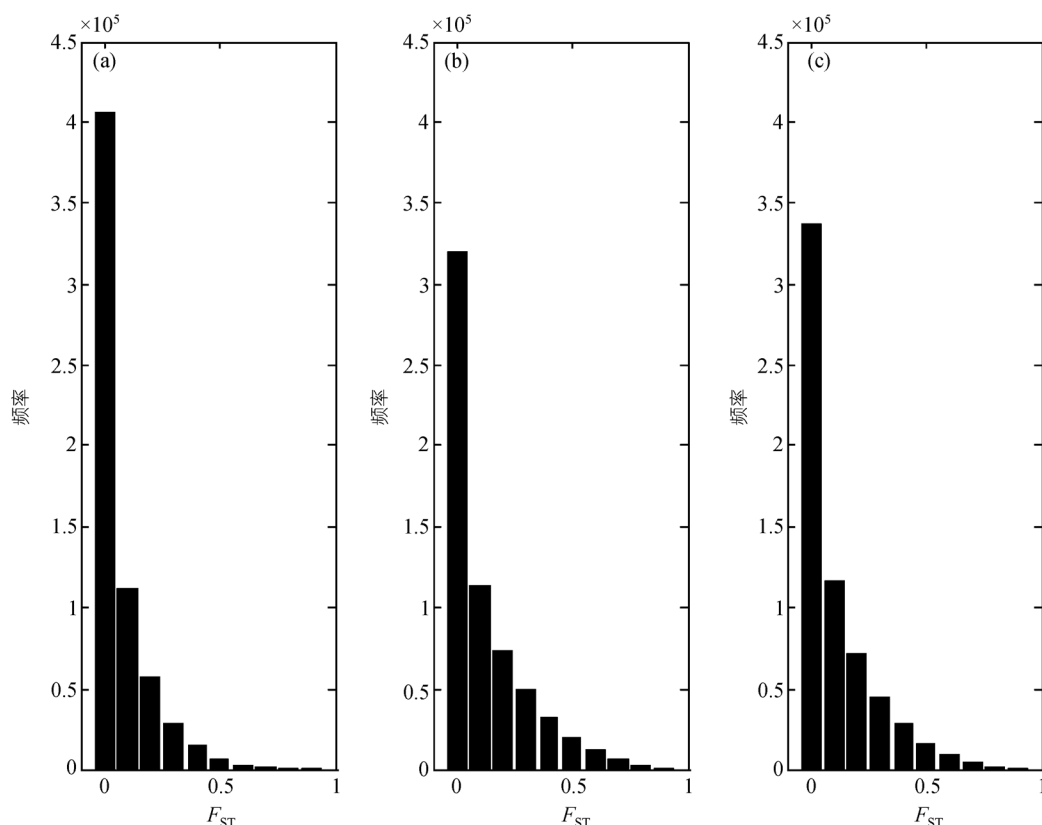


图1 单位点配对 F_{ST} 在所有常染色体区的经验分布
(a)~(c)分别为亚洲与欧洲、亚洲与非洲和欧洲与非洲人群的比较

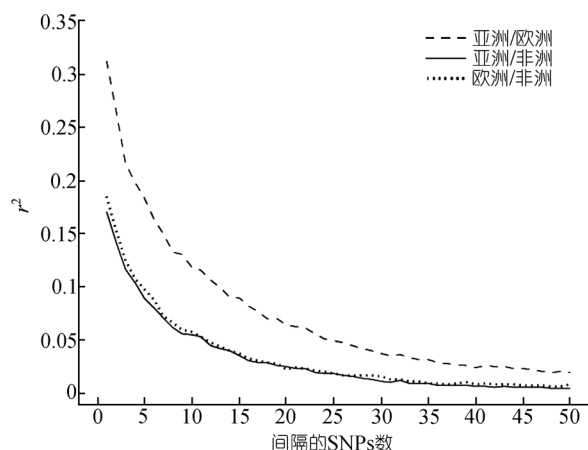


图2 8号染色体上不同间隔的SNP对之间 F_{ST} 数值的相关性
横轴代表SNP对的间隔从1~50个位点, 纵轴为特定间隔的SNP对之间 F_{ST} 数值相关性(r^2)

值”在大部分常染色体上并不随机。尤其是在6号染色体上, 在3对人群比较中“离群”位点的分布明显偏离了随机假设, 呈现出极显著的聚集趋势($P < 10^{-12}$, 表1)。这个结果说明, 正选择可以通过“搭便车”效应而影响与优势突变(位点、单体型或区域)相连锁的中

性遗传标记, 使得聚集的 F_{ST} 值呈现异常。

虽然没有“离群”位点同时在3对人群配对比较中均体现出极高 F_{ST} 值, 我们发现12669个“离群”位点在两对人群配对比较中同时呈现阳性(约占总“离群”位点数的1/5)。这种成对的高 F_{ST} 值表明, 特定人群(如ASN)等位基因频率不同于另两个人群, 很大可能是人群特异的正选择所造成。在亚洲人群中有3383个位点的等位基因频率明显不同于其他两个人群, 显著大于各人群比较相互独立假设下的预期个数(1003个, χ^2 检验, $df = 1$, $P < 10^{-12}$)。这个结果提示, 亚洲人群走出非洲后, 可能的环境改变造成特定位点的优势基因型频率增大, 导致与其他两个人群相比遗传差异均增大。而在HapMap的欧洲和非洲样品中, 这类高人群分化位点分别有2541和6745个(显著大于预期, χ^2 检验, $df = 1$, $P < 10^{-12}$), 提示这两个人群中也存在不同程度的人群特异的自然选择。

2.3 人群特异性自然选择的“候选基因”

为了分析自然选择在人群分化中的相关功能基础, 我们进一步在全基因组范围分析了人群特异的

自然选择“候选基因”. 根据EntrezGene的搜索结果我们将 12669 个高人群分化的位点进行基因定位. 若基因区域存在这类候选位点, 则该基因被认为是“候选基因”. 在有分型数据的 17422 个已知基因中, 共确定了 1853 个“候选基因”(~10%), 这些“候选基因”中的大部分为首次发现. 同时这些“候选基因”包括了一些熟知的受到自然选择的基因(如 *ALDH2*, *APOBEC3G*, *ASPM*, *BRCA1*, *CYP1A2*, *CLSPN*, *CYP3A4* 和 *VAV3* 等)^[19-26]; 而与其他全基因组自然选择分析对比, 51 个“候选基因”已证明存在极高的人群分化^[6], 211 个“候选基因”在相应的人群中存在密集的近期正选择信号^[4].

这些“候选基因”在 HapMap 人群中有不同的分布. 在亚洲、欧洲以及非洲人群中分别有 556, 482 和 1011 个“候选基因”. 大部分基因(1675 个, 约占 90.4%) 仅在一个 HapMap 人群中有自然选择的信号, 提示人群特异正选择的地区特异性(图 3). 同时“候选基因”在基因组中聚集分布(图 4, 网络版附图 1 和 2), 而且聚集的区域在不同人群之间也呈现出特异性. 例如在亚洲人群中 8p23.1 区域(8.1 ~ 11.7 Mb)聚集了 6 个正选择“候选基因”(*THEX1*, *MSRA*, *XKR6*, *C8orf16*, *MTMR9* 和 *BLK*), 约占该区域已分型基因(27 个)的 1/5. 而在欧洲和非洲人群中该区域的正选择“候选基

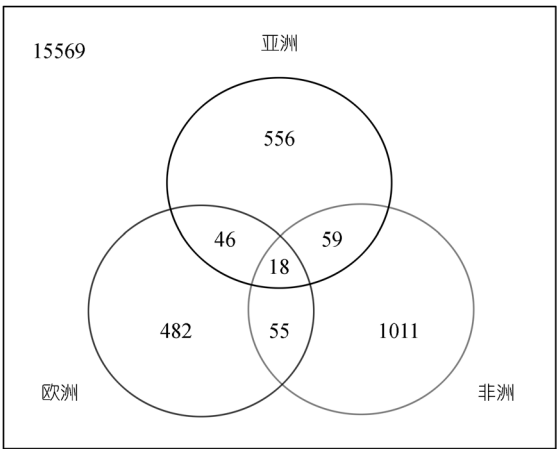


图 3 人群特异正选择的“候选基因”人群间分布模式
圆圈中黑体数值代表各人群中“候选基因”个数; 交叉区域的数值为同时在两个或以上人群中出现的“候选基因”个数. 左上角数值为非“候选基因”个数

因”分别为 1 和 3 个.

另一方面, 在 1853 个“候选基因”中, 有 178 个“候选基因”在至少两个人群中均显示极高的人群分化, 是独立假设下预期值(74 个共享基因)的两倍以上. 特别是在亚洲和欧洲人群之间, 我们发现了 64 个共享基因, 约是预期值(17 个)的 4 倍, 显著偏离了独立性假设(χ^2 检验, $df=1$, $P < 10^{-12}$). 虽然这些共享基因

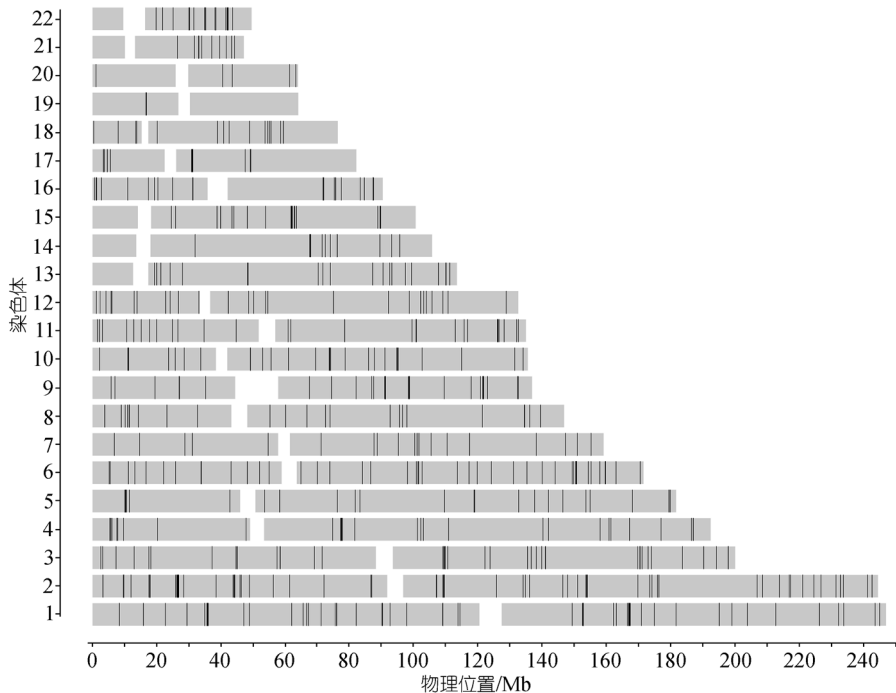


图 4 亚洲人群中正选择“候选基因”的基因组分布
灰色区域代表染色体, 白色区域代表着丝粒, 黑线代表“候选基因”的物理位置

在不同人群中正选择的目标位点并不一致,但可能提示人群间正选择具有相似的生物学功能背景.

2.4 人群特异正选择“候选基因”的功能基础

通过基因诠释(gene ontology, GO)分析,我们将17422个已知基因中的15149个归类到242种生物学过程.其中的210种生物学过程存在正选择“候选基因”,121种生物学过程的“候选基因”比例明显高于基因组背景(χ^2 检验, $df=1$, $P<0.01$)(网络版附表2).与前期的分析相似,与细胞通讯和黏附、发育过程(包括发育过程、心脏发育及中胚层发生)、信号通路(包括钙调信号、细胞黏附信号及受体蛋白酪氨酸介导的信号通路)、化合物代谢(包括糖旁路代谢、环核苷酸代

谢)以及转录起始等生物学过程相关的基因在人类进化中受到较强的正选择压力^[4,6].同时,我们的结果还提示与神经系统相关的生物学过程,如神经发生、神经元活动、旁路神经元活动、神经递质释放等也可能受到了正选择的压力,这些生物学过程共包括了169个正选择“候选基因”(网络版附表2).

虽然有些生物学过程(如细胞黏附、细胞黏附介导的通路等)在HapMap三个人群中均可能受到较强的正选择压力.但我们的分析仍显示不同人群受到正选择的生物学模式有差异(表2).亚洲人群中受到正选择的基因主要集中于信号通路以及与神经系统相关的生物学过程(各包括了55和56个自然选择“候选基因”).而欧洲人群中与发育过程相关的生物学过程包括了23个自然选择“候选基因”,提示了较强的正选择压力.并且如表4所示,多数生物学过程(13/25)的正选择信号仅存在于一个群体中,显示这些过程在不同人群中受到的不同选择压力.

3 讨论

3.1 利用群体基因组学策略确定自然选择信号

用传统的单基因遗传分析策略研究自然选择所面临的难题是如何有效排除人口动态历史(如建群者效应或瓶颈效应)的混杂作用.全新的群体基因组学策略认为,人口动态历史将影响基因组中多数位点的变异模式,而自然选择仅作用于特定的基因或位点^[7].因此,在全基因组范围构建某一群体遗传指标的经验分布,即可以反映特定人口动态历史对于人类基因组的整体影响;同时,处于分布边缘的位点则提示受到了自然选择.近期的许多自然选择分析均表明,基于该策略的outlier方案能有效排除混杂因素确定自然选择的候选位点.而Thornton等人^[27]的研究表明,在群体基因组学分析中, F_{ST} 是最简单且有效的群体遗传指标之一.

在中性条件下,人群间 F_{ST} 是由遗传漂变(genetic drift)决定的,而人群特异正选择将使候选基因及其邻近遗传标记的 F_{ST} 值增大^[10,11,15].例如,Beaumont等人^[28]发现人群遗传分化分析有足够的效力检测正选择.而我们针对人类历史中可能的人口动态历史进行了针对性的溯祖模拟(数据未列出);结果与Akey等人^[6]的研究相似,单位点 F_{ST} 能有效发现HapMap数据中的选择信号.同时我们还利用多种群体遗传指标(F_{ST} , Tajima's D等)进行分析.不但在常染色体区

表2 各人群中“候选基因”个数明显高于基因组背景的基因诠释类别

基因诠释类别	亚洲	欧洲	非洲
胚胎前/后轴发育模式	—	—	0.0167
钙调信号	0.0397	—	—
阳离子运输	0.0048	—	0.0002
细胞黏附 ^{a)}	0.0025	0.0026	0.0009
细胞黏附介导的信号	0.0005	0.0006	0.0046
细胞通讯 ^{a)}	—	—	0.0009
细胞运动	—	0.0394	—
发育过程 ^{a)}	—	0.0019	—
心脏发育	—	—	0.0247
中胚层发生	—	—	0.0018
环核苷酸代谢	—	—	0.0137
转录起始 ^{b)}	—	—	0.0271
肌肉收缩	0.0001	0.0460	—
神经-神经间隙传导	0.0097	—	0.0393
神经发生	0.0006	—	0.0001
神经元活动	—	—	0.0016
神经递质释放	0.0001	—	—
癌基因	—	—	0.0250
糖旁路代谢 ^{b)}	0.0233	—	—
旁路细胞间信号层叠	0.0359	—	0.0286
旁路神经元活动	0.0246	0.0009	—
蛋白磷酸化	—	—	0.0001
受体蛋白酪氨酸介导的信号通路	0.0164	—	0.0001
胞外分泌调节	—	—	0.0431
突触传导	0.0019	0.0001	0.0001

a) 文献^[6]报道的在人类进化中受到较强的正选择压力生物学过程; b) 文献^[4]报道的在人类进化中受到较强的正选择压力生物学过程. —, 该基因诠释类别的“候选基因”个数不高于基因组背景

鉴别出了较高比例的高 F_{ST} “离群”位点(~10%),而且发现在基因组层次上这些结果具备明显的可比性(数据未列出)。这些结果与前期研究一致,表明人类基因组中存在广泛分布的正选择信号 [2~6]。

3.2 鉴定“离群值”的方法

通过群体遗传指标的“离群值”鉴定自然选择信号需要首先确定用于判断的分界值。只有分布于分界值之外的数据点才被认为是自然选择的候选位点。在目前研究中,两种统计学方案被应用于确定离群位点:一种方案是“超出平均值两个或三个标准差的数据点”;另一种方案是“超出特定百分位数的数据点” [29]。这些方法需要利用整个数据集(包括可能的“离群值”)来估计分界值所用的参数,而前期的统计学分析显示,只要数据集中存在少量的离群数据就会大大影响这些参数的估计,因此两种方法在实际应用中尚存在缺陷 [30]。

此外,这两种方法一般需要假设所分析的数据集服从正态分布或其他对称的钟形分布。然而,大部分数据的经验分布都是多因素混合的结果,并不服从正态,甚至可能存在很大的偏斜(skewness)。因此,我们采用了 boxplot 策略,不用对数据的分布进行附加假设,也不需要估计分布参数(均值或标准差)。这种简单的策略对于复杂程度高的分布也可获得稳健(robust)的结果,尤其适用于处理大型数据集。

3.3 人群特异的正选择

前期的许多研究提示正选择具有人群特异性 [3,4,6],并可能导致人群间表型分化。本研究中多层次分析均证实了正选择的人群特异性。

虽然多数研究显示与非洲人群相比在其他人群中发现的自然选择信号较多 [1,5,31,32],并据此推测走出非洲后,人群为了适应新的生存环境可能经历了更大的正选择压力。但一些最新的研究结果却与这种观点相矛盾,不但在非洲人群中确定了更多的近期正选择信号,而且提示非洲人群较高的群体重组率是造成前期阳性信号太少的原因 [3,4]。在本研究也出现类似情况。亚洲样本和欧洲样本具备相似的正选择压力的程度,高人群分化的SNP及人群特异正选择“候选基因”数目相近。与此相反,在非洲人群中受到特异正选择的“候选位点”、“候选基因”和生物学过程均较多。尽管还需要其他研究的支持以排除始祖人群较大的有效群体(N_e)和(或)较高的群体重组率所

造成的偏差,这些结果提示了非洲人群独特的人口动态历史(或自然选择模式)。

致谢 感谢 HapMap 合作组所提供的分型数据,感谢北京华大基因研究中心基因分型平台在 HapMap 计划中所作的工作,感谢张大可、郝向稳、陈伟对本文所做的校正。

参 考 文 献

- 1 Storz J F, Payseur B A, Nachman M W. Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. *Mol Biol Evol*, 2004, 21: 1800—1811[DOI]
- 2 The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 2005, 437: 1299—1320[DOI]
- 3 Wang E T, Kodama G, Baldi P, et al. Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc Natl Acad Sci USA*, 2006, 103: 135—140[DOI]
- 4 Voight B F, Kudaravalli S, Wen X, et al. A map of recent positive selection in the human genome. *PLoS Biol*, 2006, 4: e72[DOI]
- 5 Akey J M, Eberle M A, Rieder M J, et al. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol*, 2004, 2: e286[DOI]
- 6 Akey J M, Zhang G, Zhang K, et al. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res*, 2002, 12: 1805—1814[DOI]
- 7 Ronald J, Akey J M. Genome-wide scans for loci under selection in humans. *Hum Genomics*, 2005, 2: 113—125
- 8 Hinds D A, Stuve L L, Nilsen G B, et al. Whole-genome patterns of common DNA variation in three human populations. *Science*, 2005, 307: 1072—1079[DOI]
- 9 Steemers F J, Gunderson K L. Whole genome genotyping technologies on the BeadArray platform. *Biotechnol J*, 2007, 2: 41—49[DOI]
- 10 Cavalli-Sforza L L. Population structure and human evolution. *Proc R Soc Lond B Biol Sci*, 1966, 164: 362—379
- 11 Lewontin R C, Krakauer J. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, 1973, 74: 175—195
- 12 Kelley J L, Madeoy J, Calhoun J C, et al. Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res*, 2006, 16: 980—989[DOI]
- 13 Thorisson G A, Smith A V, Krishnan L, et al. The International HapMap Project Web site. *Genome Res*, 2005, 15: 1592—1593[DOI]
- 14 The International HapMap Consortium. Integrating ethics and science in the International HapMap Project. *Nat Rev Genet*, 2004, 5: 467—475[DOI]
- 15 Weir B S, Cockerham C C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 1984, 38: 1358—1370[DOI]

- 16 McGill R, Tukey J W, Larsen W A. Variations of box plots. *Am Stat*, 1978, 32: 12—16
- 17 Thomas P D, Kejariwal A, Campbell M J, et al. PANTHER: A browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res*, 2003, 31: 334—341[DOI]
- 18 Schaffner S F, Foo C, Gabriel S, et al. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res*, 2005, 15: 1576—1583[DOI]
- 19 Oota H, Pakstis A J, Bonne-Tamir B, et al. The evolution and population genetics of the ALDH2 locus: Random genetic drift, selection, and low levels of recombination. *Ann Hum Genet*, 2004, 68: 93—109[DOI]
- 20 Zhang J, Webb D M. Rapid evolution of primate antiviral enzyme APOBEC3G. *Hum Mol Genet*, 2004, 13: 1785—1791[DOI]
- 21 Zhang J. Evolution of the human ASPM gene, a major determinant of brain size. *Genetics*, 2003, 165: 2063—2070
- 22 Huttley G A, Easteal S, Southey M C, et al. Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Australian Breast Cancer Family Study. *Nat Genet*, 2000, 25: 410—413[DOI]
- 23 Wooding S P, Watkins W S, Bamshad M J, et al. DNA sequence variation in a 3.7-kb noncoding sequence 5' of the CYP1A2 gene: Implications for human population history and natural selection. *Am J Hum Genet*, 2002, 71: 528—542[DOI]
- 24 Carlson C S, Thomas D J, Eberle M A, et al. Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res*, 2005, 15: 1553—1565[DOI]
- 25 Thompson E E, Kuttub-Boulos H, Witonsky D, et al. CYP3A variation and the evolution of salt-sensitivity variants. *Am J Hum Genet*, 2004, 75: 1059—1069[DOI]
- 26 Sabeti P C, Schaffner S F, Fry B, et al. Positive Natural Selection in the Human Lineage. *Science*, 2006, 312: 1614—1620[DOI]
- 27 Thornton K R, Jensen J D. Controlling the false-positive rate in multilocus genome scans for selection. *Genetics*, 2007, 175: 737—750[DOI]
- 28 Beaumont M A, Balding D J. Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol*, 2004, 13: 969—980[DOI]
- 29 Hoaglin D C, Mosteller F, Tukey J W. Understanding Robust and Exploratory Data Analysis. New York: John Wiley & Sons, 1983
- 30 Wayne W A, Stephan S R. Testing for outliers from a mixture distribution when some data are missing. *Comput Stat Data Anal*, 2003, 44: 193—210[DOI]
- 31 Stajich J E, Hahn M W. Disentangling the effects of demography and selection in human history. *Mol Biol Evol*, 2005, 22: 63—73[DOI]
- 32 Kayser M, Brauer S, Stoneking M. A genome scan to detect candidate regions influenced by local natural selection in human populations. *Mol Biol Evol*, 2003, 20: 893—900[DOI]

附录

附表 1 单位点 F_{ST} 值在 22 条常染色体上的差异分布参数

染色体	位点数目 ^{a)}	单位点 F_{ST} ^{b)}		
		ASN/CEU ^{c)}	ASN/YRI ^{c)}	CEU/YRI ^{c)}
1	49221	0.09±0.12*	0.15±0.18	0.13±0.16*
2	57155	0.10±0.13*	0.16±0.18*	0.14±0.16
3	41720	0.09±0.12*	0.16±0.18	0.14±0.16
4	37514	0.08±0.11**	0.15±0.17	0.14±0.16
5	37492	0.09±0.12	0.15±0.17*	0.13±0.15**
6	42700	0.09±0.12**	0.15±0.17*	0.13±0.15**
7	28413	0.09±0.12*	0.14±0.17**	0.13±0.16**
8	48863	0.09±0.12*	0.16±0.18	0.14±0.16*
9	38271	0.09±0.12	0.14±0.17**	0.13±0.15*
10	31131	0.10±0.13	0.15±0.17*	0.13±0.16*
11	28238	0.09±0.11**	0.15±0.17*	0.13±0.15*
12	26652	0.10±0.12	0.15±0.17*	0.13±0.16*
13	23514	0.09±0.12*	0.15±0.17*	0.12±0.15**
14	18453	0.09±0.12	0.14±0.17*	0.14±0.16
15	16200	0.11±0.14**	0.15±0.17	0.15±0.17**
16	14949	0.09±0.12	0.15±0.17	0.14±0.16
17	15001	0.09±0.12*	0.17±0.19*	0.14±0.17
18	25871	0.08±0.10**	0.14±0.16**	0.13±0.15*
19	10837	0.20±0.25**	0.25±0.26**	0.24±0.26**
20	12720	0.18±0.21**	0.25±0.25**	0.21±0.22**
21	13143	0.20±0.24**	0.23±0.25**	0.23±0.24**
22	11900	0.09±0.12*	0.15±0.18	0.13±0.15*
All	629958	0.10±0.13	0.16±0.18	0.14±0.16

a) 染色体上 HapMap 分型 SNP 总数; b) 各染色体上单位点 F_{ST} 的平均值和标准差; c) ASN/CEU, ASN/YRI, CEU/YRI 分别代表亚洲与欧洲、亚洲与非洲、欧洲与非洲人群之间的比较; *示 $P < 10^{-3}$, **示 $P < 10^{-12}$, 各常染色体均值与全常染色体均值的 Mann-Whitney's U test 检验的显著性水平

附表 2 自然选择“候选基因”个数高于基因组背景的基因注释类别

基因注释类别	基因个数	P 值	亚洲	欧洲	非洲
Amino acid transport	7	3.08E-04	—	—	—
Angiogenesis	8	1.70E-04	—	—	—
Anion transport	8	3.35E-03	—	—	—
Anterior/posterior patterning	6	9.14E-04	—	—	0.0167
Apoptosis	17	5.38E-05	—	—	—
B-cell- and antibody-mediated immunity	7	9.71E-03	—	—	—
Calcium ion homeostasis	5	4.26E-03	—	—	—
Calcium mediated signaling	18	1.99E-09	0.0397	—	—
Carbohydrate metabolism	13	1.98E-05	—	—	—
Carbohydrate transport	7	4.25E-04	—	—	—
Cation transport	71	<1.00E-12	0.0048	—	0.0002
Cell adhesion	87	<1.00E-12	0.0025	0.0026	0.0009
Cell adhesion-mediated signaling	61	<1.00E-12	0.0005	0.0006	0.0046
Cell communication	46	<1.00E-12	—	—	0.0009
Cell cycle	11	2.91E-04	—	—	—
Cell cycle control	45	<1.00E-12	—	—	—
Cell motility	41	<1.00E-12	—	0.0394	—
Cell proliferation and differentiation	82	<1.00E-12	—	—	—

续附表 2

基因注释类别	基因个数	P 值	亚洲	欧洲	非洲
Cell structure	58	<1.00E-12	—	—	—
Cell structure and motility	24	4.00E-10	—	—	—
Cell surface receptor mediated signal transduction	15	4.43E-05	—	—	—
Cholesterol metabolism	10	2.74E-05	—	—	—
Chromatin packaging and remodeling	18	2.70E-07	—	—	—
Cytokine and chemokine mediated signaling pathway	15	9.28E-04	—	—	—
Detoxification	6	8.26E-03	—	—	—
Developmental processes	56	<1.00E-12	—	0.0019	—
DNA recombination	5	8.99E-03	—	—	—
DNA repair	16	7.37E-06	—	—	—
DNA replication	22	1.80E-07	—	—	—
Electron transport	14	1.10E-04	—	—	—
Embryogenesis	12	1.50E-04	—	—	—
Endocytosis	10	8.15E-04	—	—	—
Endoderm development	4	3.65E-04	—	—	—
Exocytosis	10	5.67E-05	—	—	—
Extracellular matrix protein-mediated signaling	12	2.04E-07	—	0.0208	—
Extracellular transport and import	9	9.01E-04	—	—	—
General vesicle transport	19	4.91E-05	—	—	—
G-protein mediated signaling	66	<1.00E-12	—	—	—
Hearing	4	9.59E-03	—	—	—
Heart development	9	3.10E-06	—	—	0.0247
Hematopoiesis	6	9.90E-03	—	—	—
Homeostasis	11	9.09E-07	—	—	—
Immunity and defense	26	1.97E-08	—	—	—
Induction of apoptosis	13	4.00E-04	—	—	—
Inhibition of apoptosis	9	5.11E-03	—	—	—
Intracellular protein traffic	15	2.73E-04	—	—	—
Intracellular signaling cascade	30	1.00E-11	—	—	—
Ion transport	5	8.03E-03	—	—	—
JAK-STAT cascade	7	5.99E-03	—	—	—
Ligand-mediated signaling	33	1.86E-09	—	—	—
Lipid and fatty acid transport	14	5.73E-06	—	—	—
Lipid metabolism	20	1.08E-08	—	—	—
MAPKKK cascade	21	8.73E-09	—	—	—
Meiosis	7	4.18E-03	—	—	—
Mesoderm development	21	3.00E-11	—	—	0.0018
Metabolism of cyclic nucleotides	8	7.39E-05	—	—	0.0137
Miscellaneous	13	4.74E-05	—	—	—
Mitosis	11	2.85E-03	—	—	—
mRNA capping	2	3.41E-03	—	—	—
mRNA splicing	19	1.41E-06	—	—	—
mRNA transcription	28	2.40E-10	—	—	—
mRNA transcription initiation	7	1.13E-03	—	—	0.0271
mRNA transcription regulation	112	<1.00E-12	—	—	—
Muscle contraction	31	<1.00E-12	0.0001	0.0460	—

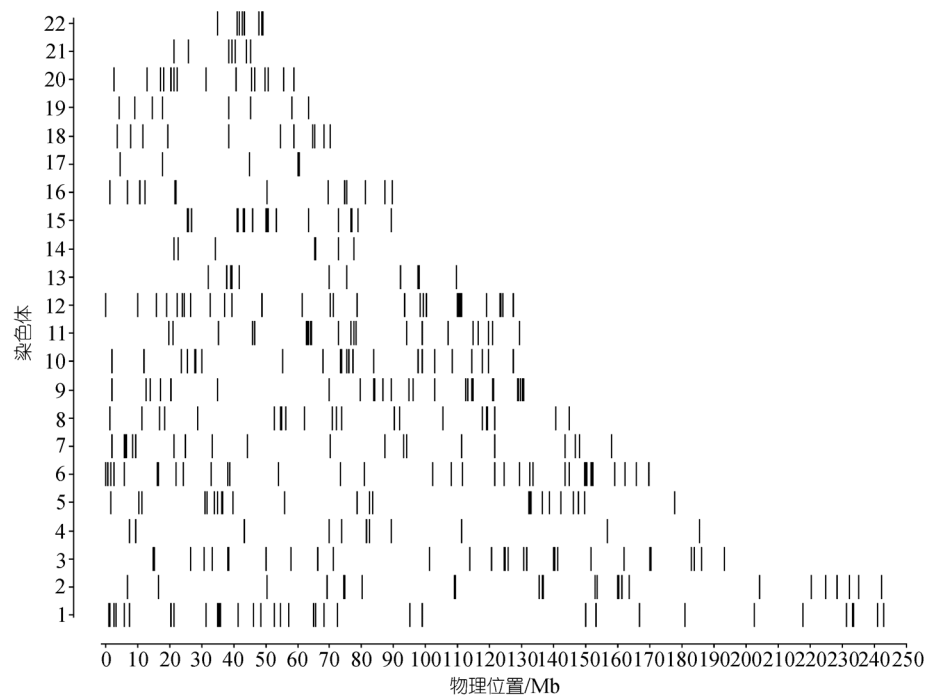
续附表 2

基因注释类别	基因个数	P 值	亚洲	欧洲	非洲
Muscle development	13	1.39E-05	—	—	—
Nerve-nerve synaptic transmission	14	8.14E-08	0.0097	—	0.0393
Neurogenesis	81	<1.00E-12	0.0006	—	0.0000
Neuromuscular synaptic transmission	6	2.85E-05	—	—	—
Neuronal activities	23	<1.00E-12	—	—	0.0016
Neurotransmitter release	20	<1.00E-12	0.0000	—	—
Nuclear transport	8	2.57E-03	—	—	—
Nucleoside, nucleotide and nucleic acid metabolism	23	7.68E-08	—	—	—
Oncogene	15	2.80E-08	—	—	0.0250
Oncogenesis	17	1.84E-06	—	—	—
Oogenesis	7	2.05E-03	—	—	—
Other amino acid metabolism	5	7.38E-04	—	—	—
Other carbohydrate metabolism	8	1.97E-04	0.0233	—	—
Other carbon metabolism	8	1.44E-03	—	—	—
Other developmental process	8	2.57E-03	—	—	—
Other immune and defense	16	1.58E-06	—	—	—
Other intracellular protein traffic	11	1.37E-06	0.0340	—	—
Other intracellular signaling cascade	33	<1.00E-12	0.0359	—	0.0286
Other metabolism	31	3.38E-06	—	—	—
Other neuronal activity	25	<1.00E-12	0.0246	0.0009	—
Other oncogenesis	6	1.92E-03	—	—	—
Other polysaccharide metabolism	13	6.41E-05	—	—	—
Other protein targeting and localization	4	1.76E-03	—	—	—
Other sensory perception	3	8.75E-03	—	—	—
Other signal transduction	6	8.26E-03	—	—	—
Oxidative phosphorylation	6	9.06E-03	—	—	—
Phagocytosis	5	6.32E-03	—	—	—
Phospholipid metabolism	13	1.63E-04	—	—	—
Protein acetylation	7	7.89E-06	—	—	—
Protein biosynthesis	21	1.73E-04	—	—	—
Protein disulfide-isomerase reaction	4	1.76E-03	—	—	—
Protein glycosylation	23	3.63E-09	—	—	—
Protein metabolism and modification	8	4.55E-04	—	—	—
Protein phosphorylation	92	<1.00E-12	—	—	0.0001
Protein targeting	13	3.79E-06	—	—	—
Protein targeting and localization	7	1.13E-03	—	—	—
Protein-lipid modification	4	2.24E-03	—	—	—
Proteolysis	71	<1.00E-12	—	—	—
Purine metabolism	6	8.26E-03	—	—	—
Pyrimidine metabolism	6	1.45E-03	—	—	—
Receptor mediated endocytosis	9	3.06E-03	—	—	—
Receptor protein serine/threonine kinase signaling pathway	5	8.03E-03	—	—	—
Receptor protein tyrosine kinase signaling pathway	34	<1.00E-12	0.0164	—	0.0001
Regulated exocytosis	9	9.56E-06	—	—	0.0431
Regulation of carbohydrate metabolism	3	5.31E-03	—	—	—
Sensory perception	4	6.01E-03	—	—	—
Signal transduction	31	4.00E-11	—	—	—

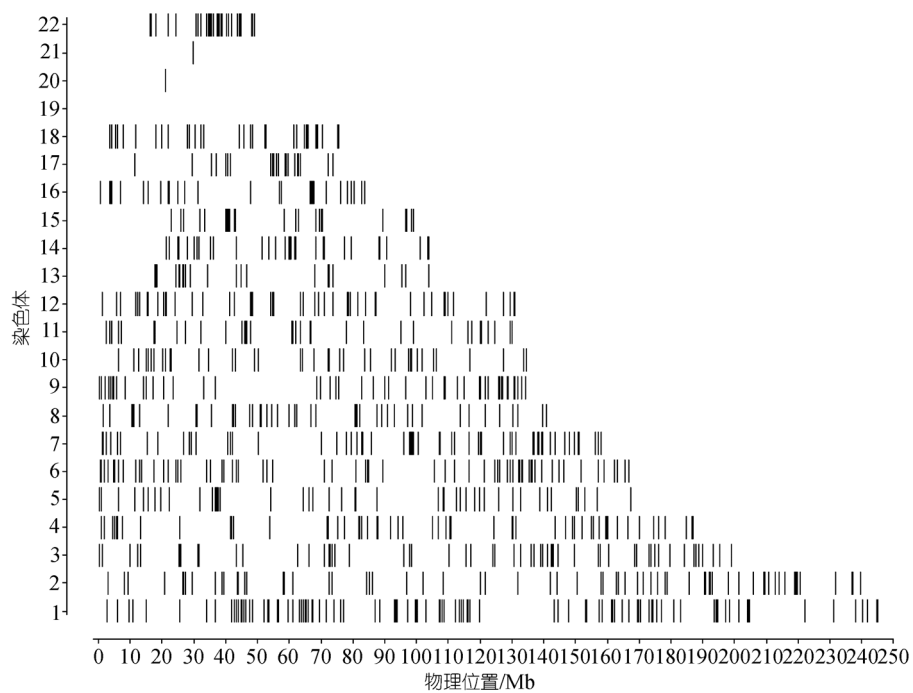
续附表 2

基因诠释类别	基因个数	<i>P</i> 值	亚洲	欧洲	非洲
Skeletal development	14	4.47E-06	—	—	—
Small molecule transport	15	6.04E-06	—	—	—
Spermatogenesis and motility	8	5.03E-03	—	—	—
Steroid hormone metabolism	5	2.33E-03	—	—	—
Stress response	21	7.86E-08	—	—	—
Synaptic transmission	22	<1.00E-12	0.0019	0.0000	0.0000
T-cell mediated immunity	11	1.06E-03	—	—	—
Transport	44	<1.00E-12	—	—	—
Tumor suppressor	10	6.50E-05	—	—	—
Vision	12	3.44E-03	—	—	—

—代表该基因诠释类别的“候选基因”个数不高于基因组背景



附图 1 欧洲人群中自然选择“候选基因”的基因组分布
灰色代表染色体区域，白色区域代表着丝粒，黑线代表“候选基因”的物理位置



附图 2 非洲人群中自然选择“候选基因”的基因组分布
灰色代表染色体区域, 白色区域代表着丝粒, 黑线代表“候选基因”的物理位置