分类号_____ 密级_____

UDC _____ 编号_____

中国科学院研究生院

博士学位论文

混合人群和疾病基因的自然选择研究

靳文菲

指导教师 金力教授 徐书华研究员

中国科学院上海生命科学研究院计算生物所

申请学位级别 博士 学科专业名 计算生物学

论文提交日期 2011 年 11 月 论文答辩日期 2011 年 11 月 17 日

培养单位 中国科学院上海生命科学研究院计算生物所

学位授予单位 中国科学院研究生院

答辩委员会主席 钟扬教授

Signatures of Natural Selection in Admixed Populations and on Disease Genes

A Dissertation Submitted to Graduate University of the Chinese Academy of Sciences

For the Degree of Doctor of Philosophy

By

Wenfei Jin

Supervisors:

Professor Li Jin

Professor Shuhua Xu

November, 2011

混合人群和疾病基因的自然选择研究

中国科学院研究生院博士学位论文

靳文菲

导师

金力教授

徐书华研究员

2011 年 11 月

献给我的父亲

目录

混合人群和疾病基因的自然选择研究

靳文菲（计算生物学）

指导老师：金力教授和徐书华研究员

摘要

虽然人类基因组上大多数的遗传多态性位点是中性进化的，但是有一些位点和性状被自然选择重塑并且在各地人群适应当地环境中起着非常重要的作用。大量的高密度单核苷酸多态性数据为我们在基因组水平检测自然选择提供了必要的条件。我们重点研究美国黑人受到的自然选择是因为他们的非洲祖先人群在历史上经历过非常残酷的环境和极高的死亡率。我们收集了 1890 个美国黑人样本及其可能祖先人群的 3320 个个体的常染色体上的 491,526 个单核苷酸多态性位点。我们通过本文发展的祖先染色体片段分布法揭示美国黑人的形成过程，发现美国黑人的祖先染色体片段分布与 14 个世代的持续基因流模型（continuous gene flow model）最匹配，即欧洲人群持续贡献基因流到美国黑人中的模型能够较好的解释美国黑人的形成过程。

我们先用祖先贡献偏离法发现美国黑人基因组上有几个区域表现出特别高的非洲成分或欧洲成分，这可能是美国黑人在发生群体混合后自然选择留下的痕迹。随后我们发展了一种检测自然选择新方法：我们利用美国黑人基因组里的非洲祖先成分人工重构了一个的非洲祖先群体 (ancestral African population; AAF)，我们再比较这个重构的非洲祖先群体与非洲土著人群在基因组上的差异，这种方法可以检测到美国黑人离开非洲后受到的总的自然选择，包括混合前和混合后。我们发现用这种新方法检测到的很多受自然选择的基因属于美国黑人特异性高发疾病如前列腺癌和高血压相关基因。我们因此推测这些疾病基因可能在美国黑人适应北美新环境中起了重要作用。非洲祖先群体与非洲土著人群差异很大的区域也包含了多个通过突变增加携带着抵抗疟疾能力的基因，包括对 *HBB* 和 *CD36*。我们对这些基因进行分析，发现其抵抗疟疾的等位基因在重构的非洲祖先群体中的频率比非洲土著人群

低，这证实了来自疟疾的选择压力在新大陆得到释放。我们用这两种不同的方法筛选出来的基因基本完全不一致，这可能也意味着美国黑人在混合前后面临的主要环境压力不一样。最后我们认为本文检测自然选择的新方法有很好的统计效果，并且可以应用到其它混合人群如拉美人群和维吾尔族人群。

通过对混合人群的自然选择研究，我们意识到人类疾病相关基因的进化与自然选择有紧密的联系，于是我们随后做了一系列比较分析来揭示不同类型的疾病基因受到的自然选择。我们首先发现同时与单孟德尔遗传病和复杂疾病相关联的基因比预期数目显著性的多很多，我们姑且把这类基因简称为双联基因（Mendelian and Complex disease gene; MC gene）。这样我们把人类的全部基因分为必需基因，双联基因，单孟德尔遗传病基因（去除与复杂疾病相关基因），单复杂疾病基因（去除与孟德尔遗传病相关基因）和其它基因。对人类的多态性数据分析后，我们发现双联基因和复杂疾病基因都受到了近期的正选择，而必需基因和 单孟德尔遗传病基因受到较强的负选择。对物种间差异数据分析表明必需基因总是最保守，这支持必需基因在长期的进化史中总是受到最强的负选择。而双联基因一般是第二保守，意味着在进化中也受到较强的负选择。拷贝数变异在不同类基因类型中的富集分析也支持双联基因同时受到较强的正选择与负选择。除此之外，我们也比较了各类基因在基因表达模式，基因结构，蛋白蛋白相互作用，群体分化等方面的差异。总的来看，每类疾病基因都有一些不同的特征。我们推测双联基因的很多特征和他们在复杂疾病和单孟德尔遗传病中的双重作用有关。据我们所知，该工作是第一个对双联基因特征进行系统分析的研究，我们在比较过程中也对其它四类基因有了新的认识。我相信，这样的一个系统的比较分析，有助于人们深入了解疾病基因的产生和进化机制。

关键词：自然选择；群体混合； 美国黑人；单核苷酸多态性；疟疾； 孟德尔遗传病； 复杂疾病； 疾病基因；进化；拷贝数变异； 基因表达；基因结构；组织特异性；智能通路分析

Signatures of Natural Selection in Admixed Populations and on Disease Genes

Wenfei Jin (Computational Biology)
Directed by Prof. Li Jin and Shuhua Xu

Abstract

Although the vast majority of human genetic diversity evolves neutrally, some loci have been shaped by natural selection, which plays an important role in adapting different populations to their local environments. The availability of high-density SNPs has provided the essential resources for genome-wide scanning signatures of natural selection. We focused on the natural selection in African Americans (AfA) due to their African ancestors experienced very high mortality in history. We firstly proposed that genome-wide distribution of ancestral chromosomal segments could reveal the population admixture dynamics. Based on 491,526 autosomal SNPs genotyped in 1,890 AfA and 3,320 individuals from their potential ancestral populations, we implemented this novel approach and found that a continuous gene flow (CGF) model, in which AfA continuously received gene flow from European populations over about 14 generations, best explained the admixture dynamics of AfA among several putative models.

We found several genomic regions showing excess of African or European ancestry, which might only reflect signatures of the natural selection since population admixture. Then we developed a new strategy to detect natural selection both pre-and post-admixture by reconstructing an ancestral African population (AAF) from inferred African components of ancestry in AfA and comparing it with indigenous African populations (IAF). Interestingly, many selection-candidate genes identified by the new approach were associated with AfA specific high-risk diseases such as prostate cancer and hypertension, suggesting an important role these disease-related genes might have played in adapting to new environment. *CD36* and *HBB*, whose mutations confer a degree of protection against malaria, were also located in the highly differentiated regions between AAF and IAF. Further analysis showed that the frequencies of alleles protecting against malaria in AAF were lower than that in IAF, which consists with the relaxed

selection pressure of malaria in the New World. There is no overlap between the top candidate genes detected by the two approaches, indicating the different environmental pressures AfA experienced pre-and post-population-admixture. We suggest that the new approach is reasonably powerful and can also be applied to other admixed populations such as Latinos and Uyghurs.

Since we found many disease genes have been subjected to natural selection, we performed a series of comparative analyses to reveal the evolution of these genes. There were unexpected large number of genes associated with both complex diseases and Mendelian diseases, which were referred to as Mendelian and complex diseases (MC) genes. Thus we classified human genes into five categories: MC genes, Mendelian but not complex diseases (MNC) genes, complex but not Mendelian diseases (CNM) genes, essential genes and OTHER genes. Analysis of the human polymorphic data showed that both MNC and essential genes were under stronger purifying selection, while MC and CNM genes have been subjected to recent positive selection. Analysis of divergence data also showed that essential genes were always the most conserved, indicating the strongest purifying selection on these genes. MC genes were the second most conserved, suggesting a strong purifying selection as well. Over-representation analysis of copy number variations (CNVs) in different gene categories also suggested that MC genes have been subjected to both strong purifying and positive selection. In addition, we also compared the gene expression pattern, gene structure, protein-protein interaction, population differentiation and many other characteristics among these gene categories. It seems that each kind of genes has some specific characteristics. We think that many characteristics of MC genes could be attributed to their double identities as both Mendelian and complex disease genes. To our knowledge, this study is the first effort to characterize the features of MC genes; we also gained many new insights on the other four kinds of genes in these comparative analyses.

Keywords: Natural Selection; Population Admixture; African Americans; Single nucleotide polymorphism (SNP); Malaria; Mendelian disease; Complex disease; Disease gene; Evolution; Gene ontology; Copy number variation (CNV); Gene expression; Gene structure; Tissue specificity; Ingenuity pathway analysis (IPA)

引 言

当外界环境压力使得具有不同表型的个体的生存率或繁殖力不同时，自然选择开始发挥作用，并最终影响表型。自然选择是生物本能和生理活动无法适应外界环境压力的必然结果。例如，一般动物会本能的改变其行为如日常活动以适应环境的变化。如果生物行为的改变仍不足以应付其所处环境的挑战，随后伴随着基因表达水平改变的一系列生理机制将进一步缓解环境压力，如降解体内脂肪以应对食物短缺。如果上述机制仍然不能有效应对环境变化，不同表型和基因型的个体将会表现出不同的生存率与繁殖力。含优势等位基因的个体一般表现出较高的存活率并繁殖更多的后代，这意味着自然选择和适应性进化发生。这样，自然选择使得优势表型及其潜在的基因突变在种群内变的更加普及。经过一定时间，生物体将能够更好的适应其所处环境以保证其正常的生存和繁殖----即适应性进化。

被科学所证实的自然选择，作为现代进化理论的关键机制，目前已经被社会大众特别是无神论者广泛接受。虽然很多人知道各种动植物在外界环境中会受到自然选择，但是很少有人会感觉现代人类也受到很强的自然选择。可能由于很强的负向自然选择，大多数动植物还只能在特定的环境中生存，就连与我们亲缘关系最近的黑猩猩也只能在非洲大陆的特定区域生存。但是自从约 10 万年前第一支现代人类走出非洲大陆以来，人类几乎已经成功的占领了世界的每个角落。人类迁移历史和人群历史在某种程度上可以解释当前的人类遗传学和表型多态性，同时人类在分散到世界各地的过程中会遇到各种在温度、湿度、光照等方面差异性明显的环境，包括很多人类不可抗拒的环境如高原缺氧等环境。这样可以推测人类分布到世界各地

的过程必然必然伴随着适应各地环境的过程。由于欧洲宗教因素，人类进化并没有在 1859 年达尔文发表的《*物种起源*》中被提及，之后进化理论逐渐随被用来解释人类的起源和进化，人群差异和多态性。现代遗传学不仅发现了各地人群与当地环境相关的一些性状，也发现这些性状背后的一些基因。可以说，没有人类对各地环境的适应性进化，人类也不可能分散到世界的各个角落。

但是人类始终还没有能够很好的解释大多数自然选择背后的遗传突变及其生物学机制。数十年前，鉴定正选择的靶点主要依赖于候选基因的方法，而候选基因的方法不仅效率低下，还受制于我们对基因功能的认识。近来，伴随着高密度基因型数据和测序数据的快速增长，基于全基因组数据的自然选择研究越来越多。鉴定出受自然选择的基因不仅能够解释人类如何适应各地的病原体、气候、饮食，认知等各种环境压力。这些发现不仅有助于我们更加深入地理解人类的起源和历史，并为鉴定具有特定重要功能的特定基因提供了可能，从而阐明一些人类疾病的遗传学基础。

由于医学和生活条件的巨大进步，很多人（特别是现代西方社会）认为自然选择对人类不再起重要作用。我们选取近期形成的混合人群美国黑人群体和墨西哥人群体做为研究对象，来重新探讨这个问题。首先，我们对美国黑人群体的分析发现其受到过明显的自然选择并筛选到这些受到自然选择的基因。同时，我们发现很多受过自然选择的区域包含有美国黑人特异性高发疾病的相关基因。非洲和美洲的环境差异，特别是病原体种类如疟疾的巨大差异，在美国黑人群体的基因组上留下很明显的痕迹。我们初步的研究分析也发现墨西哥人群种中存在的有趣的选择信号。基于当前的研究结果，我们可以推论出，自然选择在人类基因中仍然起作用，人类

还在继续进化。一方面，各种疾病致死或致残可以被认为是现代自然选择的一种形式。 另一方面，对于人类疾病基因进化的阐述有望阐明人类疾病基因的起源。我们发现同时在单基因遗传病和复杂疾病中起作用的基因数目巨大。我们将人类疾病基因分成不同的亚类后发现，不同类型的基因受到的自然选择方式完全不一样。同时这些研究将提供关于自然选择过程和机制的很多新见解，最终必将充实现代进化理论。

第一章：自然选择和人类适应性进化概述

现代人类的起源和走出非洲学说

大约 230-250 万年之间，人属的祖先在非洲与南方古猿（Australopithecus）分离开来[1]并形成巧人。随后的历史上，人属曾经发展出多个不同的物种如直立人，海德堡人和尼安德特人等。现在， 现代人类（也称智人，*Homo sapiens)*之外的其它人属物种都已经灭绝。关于解剖学上的现代人类(anatomically modern human)起源有两种相互竞争并相互矛盾的学说：晚近非洲起源学说和多地区演化说。两个学说的争议在于现代人类是否只起源于非洲，前者主张现代人类起源于非洲随后扩散到世界各地，后者认为现代人是由各地的古人类分别进化而来的。 20 世纪 80 年代, Cann *et al.* [2]利用线粒体 DNA 的多态性对人类进化做的先驱性研究对于理解现代人类的起源和迁移具有划时代意义。 最近 30 年来，由于对线粒体 DNA 和 Y 染色体的多态性分析都支持晚近非洲起源学说[3,4]，其逐渐被人们接受并变为现代人类起源的主流观点。根据晚近非洲起源学说，大约在 20 万年，古人类只在非洲进化成解剖学意义上的现代人。随后一支现代人在 12.5 万年到 6 万年之间走出非洲，随后他们取代了各地的古人类如欧洲的尼安德特人和亚洲的直立人。

但是，线粒体和 Y 染色体分别只代表一个基因座，这样它们提供的信息受到严重的基因漂变影响。 最近几年，几个已经灭绝的古人类的 DNA 已经被全基因组测序[5,6]，这些全基因组多态性位点能够提供更全面的信息，这样进一步分丰富了我们对人类起源和迁徙的认识。第一个被全基因组测序的古人类是与现代人类有着最近共同祖先祖先的尼安德特人。在灭绝以前（大约 3 万年前），尼安德特人曾经广泛的分布在欧亚大陆的西部。对尼安德特人和现代人的基因组进行比较分析进一

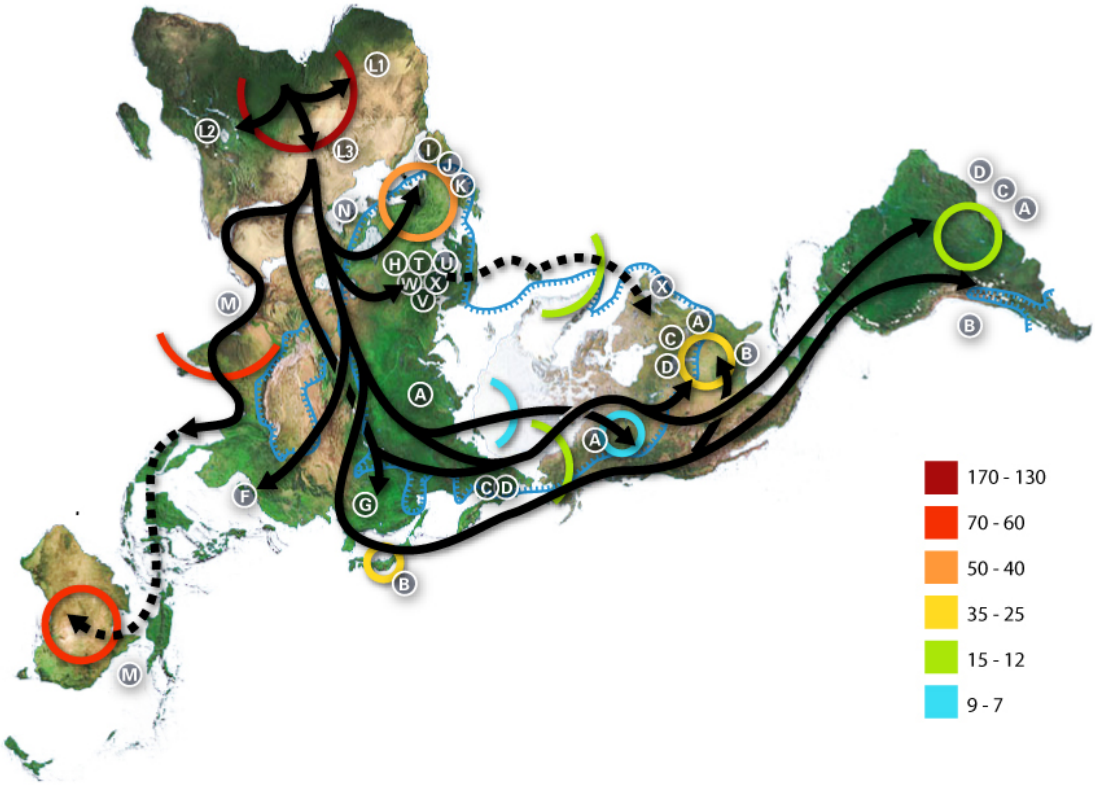


图 1. 线粒体 DNA 揭示的现代人类迁徙图。

步支持晚近非洲起源学说[5]。但是与现代非洲人相比，尼安德特人与现代欧亚大陆人共享更多的基因变异，这可能表明尼安德特人的基因流对欧亚大陆现代人群有贡献，既尼安德特人和现代欧亚大陆人的祖先发生过混合。并且这次人群混合发生在欧洲和亚洲大陆祖先人群分离之前，尼安德特人总的贡献了欧亚大陆人群基因组的 1-4%[5]。第二个全群基因组测序的古人类被称之为丹尼索瓦人，其命名源自发现其化石的南西伯利亚的丹尼索瓦洞穴。虽然丹尼索瓦人与尼安德特人处于相同的世系，但是丹尼索瓦人并没有参与到尼安德特人与现代人发生混合的事件中。进一步分析表明来自丹尼索瓦人的基因流贡献了东南亚现代美拉尼西亚人的基因组的 4-6%[6]。随后的研究表明来自丹尼索瓦人的基因流对新几内亚人群，澳大利亚土著人群和 Mamanwa 人群有贡献，但是对现代东亚人群，西印度尼西亚人，Jehai 和

Onge 没有贡献[7].

最近，对一个澳大利亚土著全基因组测序发现他们的祖先可能在 6.2-7.5 万年之间到达亚洲，这次迁徙与现代东亚人 2.5-3.8 万年的迁徙属于两个不同的事件[8]。但这两次迁徙的祖先人群在美洲印第安人与现代东亚人的祖先分离之前已经有基因交流。同时研究表明，澳大利亚土著人群可能是非洲之外持续繁衍的最古老的现代人类群体。

总之，现代遗传学证据表明：现代人类的祖先大约 20 万年起源于非洲，大约最近 10 万年走出非洲并逐渐扩散到了世界各地。虽然现代人类最终取代了各地的古人类，但并不是直接的完全取代。各种古人类通过与现代人类通婚使得其少量遗传信息得以保留。而现代人类走出非洲的过程可能由一系列复杂的事件构成，至少包括两次大的迁徙。同时，现代人类在分散到世界各地的过程中要面对各地完全不同环境的挑战，这样人类分布到世界各地的这个过程可以看成人类适应各地环境的过程，即现代人类经历各种自然选择的过程 [9,10]。

现代人对高纬度气候的适应

人类走出非洲后面对的第一个挑战是变冷的气候，以及减少的日照。这种环境变化在人的形态如身高，体重，体重指数，外鼻孔的大小、嘴唇的厚薄和凹凸、毛发的分布上留下很明显的痕迹[11,12]。具体来说，个体的体重和身高随着气候变冷分别变高和变大。因为个体变大有利于产热，而相对体表面积减少有利于保温。 在这些特征中，肤色的变化是最明显的：现在我们可以看到深色皮肤的人群集中在热带，浅色皮肤的人群集中在高纬度[11,12]。这种肤色的差异是由于个体

本身的黑色素类型及数量在人群中的的分布不同导致的。而黑色素的全球分布可以用逃避紫外辐射伤害和维生素 D3 合成的平衡选择来解释[13, 14]。在热带地区，深色的黑色素可以使紫外线散射并吸收紫外辐射，从而保护皮肤不受阳光灼伤，防止营养如叶酸被阳光破坏。在这种情况下，任何影响色素生成的突变都是有害的，参与色素合成的基因如 *MC1R* 在非洲受到很强的负选择。但在欧洲和东亚人群中，*MC1R* 是高度多态的，其中包括很多非同义突变。在高纬度地区，阳光不是很充足。紫外线透射皮肤是合成维生素 D3 所必需的，因此皮肤颜色变淡变浅才受大自然偏爱。虽然 *ASIP* 和 *OCA2* 的多态性可能在全球人群的肤色淡化中起着一定作用。但是欧亚大陆西部人群和东亚人群的皮肤淡化可能是两个相互独立的事件。利用高密度的单核苷酸多态性数据在全基因组水平进行延长的单体型纯合度和群体差异性分析，发现两大人群的受到选择的肤色相关基因很多是不一样的。*SLC45A2* 和 *SLC24A5* 在欧亚大陆西部人群中受到很强的自然选择（Figure 1a）；而 *EADR* 和 *ED2R* 在东亚人群和美洲人群中受到很强的自然选择（Figure 1b） [15]。同时，我们的祖先在热带非洲进化出一套适应当地环境的机制——热适应。虽然出汗降温，但出汗的同时也伴随着盐分的丢失。当环境中的盐分很少时，这样在大量失水后保持动脉弹性和收缩压的个体（能很好保持盐分）才更有可能幸存下来。当现代人类走出非洲后，这种适应优势反而变成劣势——导致较高的血压。这种现象得到现代遗传学的支持，特别是 *GNB3* 825T 能够解释全球血压差异的 64%。

在现代人类走出非洲后，一般人群在某种程度上已经适应了的当地环境。由于人类最近的迁徙和殖民，很多人生活的环境和他们祖先生活的气候环境完全不同，这种基因组和生活环境的不一致已经导致了很多健康问题。比如，浅肤色的人

生活在低纬度很容易患皮肤癌；而深肤色的人生活在高纬度很容易患维生素 D 缺乏症。澳大利亚白人患一些皮肤癌的概率比澳洲土著高 10 倍[14]。而美国黑人患维生素 D 缺乏症是美国最高的。另外，从热带迁移到温带的个体很容易患高血压，最典型的是美国黑人极高的高血压患病率。

现代人的高原适应

虽然人的体质特征随着纬度的渐变是很明显的，但最奇特和给人印象很深的是人的高原适应。人类在高原上面对的生存和生殖挑战包括外周氧降低（deceased ambient oxygen tension），强的紫外辐射，很大的昼夜温差和干燥的环境。但是全球却有 1.4 亿人永久的居住在亚洲，美洲和东非的高原上（海拔>2500 米）。世代居住在高原上的人群与平地人群之间，以及他们相互之间表现出一些独特的生理特征。比如：藏族模式（低血红蛋白伴动脉低氧血症）；安第斯山模式（红细胞增多伴动脉低氧血症）；埃塞俄比亚模式（血红蛋白浓度和动脉氧饱和度均处于海平面正常水平）。

虽然对高原人群的生理特征的描述已经有上百年的历史，直到最近这些特征的遗传分子基础才逐渐被揭示出来。近期几个研究小组在全基因水平对藏族人群进行的分析分别独立的发现 *EPAS1*（又名 *HIFPH2*,位于 1q42.2)和 *EGLN1*（又名 *HIF2A*,位于 2p21)在藏族的高原适应中起着重要作用 （Figure 2）,这两个基因都在缺氧诱导因子（Hypoxia inducible factor，HIF）通路中起着重要作用[16-19]。并且这两个基因的遗传多态性与血红蛋白含量显著相关。徐书华等[16]分析了这两个基因的连锁不平衡并分别找到了藏族的优势单体型。根据藏族在基因 *EPAS1* 和 *EGLN1*

的优势单体型显著增多的现象，他们提出优势单体型携带者模型来揭示这两个基因在藏族高原适应中的作用。另一方面，对安第斯高原人群的研究发现 HIF 通路中的基因 *EGLN1*，*ENDRA*，*PRKAA1* 和 *NOS2A* 可能在其高原适应中具有重要作用。因此，HIF 通路有可能在不同人群的高原适应中都起着重要作用。

人类对食物变化的适应

早在 1859 年的《物种起源》中，达尔文已经着重强调了食物在生物进化中的重要作用。自从我们的祖先开始使用石器等工具和掌握火的使用开始，人类的饮食结构发生了重大变化。随后，人类从采集狩猎型社会向农耕社会过渡，以及后来从农耕社会向工业社会过渡都极大了影响了人类的生活方式和饮食。这种遗传的生物特征与人类当前生活饮食的不一致为自然选择地发生提供了靶点。

乳糖耐受症（Lactase persistence）是现在研究最系统最完善的人类食物适应性自然选择。 人的乳糖酶（lactase-phlorizin hydrolase）主要在小肠中表达，并在那里把乳糖分解成容易吸收的葡萄糖和半乳糖。人类与其他哺乳类动物一样，在哺乳期主要靠母乳喂养，断奶后乳糖酶表达量下降并且乳糖代谢能力也随之下降。对于以牛奶及乳制品做为日常饮食的群体来说，能够消化食物中的乳糖会为个体带来生存上的优势。流行病学研究已经发现世界范围内乳糖耐受人群与食用牛奶及乳制品人群有很高的相关性。通过对欧洲人群的连锁不平衡和单体型分析已经发现了编码乳糖酶的 LCT 基因上游的调控突变 C/T-13910 导致了乳糖耐受[20]。并且欧洲人群中携带 13910-T 的单体型的连锁不平衡延伸到 1Mb 以上，近一步证实了 LCT 受到最近的自然选择。非洲的乳糖耐受人群携带的 LCT 调控突变（G/C-14010,

T/G-13915 和 C/G-13907)与欧洲人不同，并且非洲内部各群体之间也相互不同，说明乳糖耐受在各地独立起源，即发生了趋同进化[21]。 另外，相对于游牧或畜牧文明，农耕社会饮食的显著特征是高淀粉消耗。有趣的是，人 α 唾液唾液淀粉酶（AMY1）基因已经被报道在种植业为主的人群中受到最近的正向自然选择 [22]。AMY1 基因在个体中的的拷贝数与唾液淀粉酶的表达水平高度相关，同时与人群的淀粉消耗成正相关。这样，携带更多的 AMY1 拷贝数的个体可以从淀粉为主的饮食中获得更多的营养，从而获得生存上的优势。

人类对病原体的适应及其对免疫相关基因的影响

长期以来，传染性疾病一直是人类进化史上的一个沉重的负担。在巴斯德的微生物病理学说建立之前，人的平均寿命只有 25 岁左右。随后，伴随着卫生水平的提高,以及疫苗和抗体的推广，传染性疾病逐渐得到了控制。作为曾经的人类第一致命杀手，病原体在人类的基因组上留下了很深的痕迹。但是病原体和自然选择的关系在很长一段时间内并不为人知，直到约翰.霍尔丹在研究地中海贫血病人的疟疾感染的时候意识到两者之间的联系。因为宿主与病原体之间的复杂竞争关系（适应与反适应，逃避与反逃避），使得病原体适应比其它自然选择要复杂的多。同时在人与病原体竞争的过程中，人类不断的提高自身的免疫系统来抵抗各种病原体感染。虽然人类和黑猩猩才分离三百万年，两者对很多传染疾病如艾滋病，疟疾，乙肝，丙肝和 A 型流感的感染力和患病程度已经差异很大。 对全基因组受到近期自然选择的区域进行分析，发现免疫相关基因显著富集，说明免疫相关基因更容易受到自然选择 [23,24]。对全球人群的分析发现环境中的病原体类型及丰度已

经影响了人类的遗传多态性，特别是人白细胞抗原（Human leukocyte antigen）的多态性与居住环境中的病原体种类成正相关。

全基因组关联分析（genome wide association studies）通过关联分析在全基因组范围内筛选与疾病/性状相关的遗传变异(单核苷酸多态性和拷贝数)，现在也广泛用来检测传染疾病感染程度和患病程度相关的遗传变异。这些检测到的基因可能是病原体感染的的靶点，载体或参与了病原体感染通路的其它蛋白。这样仔细分析这些病原体关联基因将有利于我们阐明病原体感染过程及自然选择的靶点。虽然免疫相关基因一般能够帮助宿主逃避感染，但如果病原体发展了利用该基因产物进入细胞或生存的机制，那么基因失活突变对于宿主来说是有利的。由于很强的选择压，一些基因如 *CCR5, DARC, CASP12, SERPINA2* 和 *SIGLEC12* 的丧失功能性突变的频率已经升的很高。这种现象为我们认识免疫相关基因的冗余提供了一个新的视角[25]。人类对病原体的适应研究中已经有很多的经典范例，其中最典型的是疟疾所形成的选择压对人类基因组的影响。疟疾曾经并且现在还是热带地区儿童的第一大杀手 [26]。由于疟疾带来的极高死亡率，已经报道的受到疟疾影响的基因已经多达几十个，并且疟疾驱动产生了我们最常见的几种单基因疾病如镰刀状细胞贫血症，α-地中海贫血，β- 地中海贫血和葡萄糖-6-磷酸酶缺失症。但是，我们现在看到红细胞相关的变异可能只占疟疾相关变异的很小一部分，很多免疫相关基因及炎症相关还有待于近一步发现。

在一般人群中检测自然选择的方法

虽然基因组上的大多数区域是中性进化的，一些区域还是受到了自然选择的

影响。 正向选择的基本形式是选择扫荡（selective sweep)，该过程使得优势等位基因的频率迅速上升。由于搭车效应（hitchhiking)，正向自然选择会影响很大一个区域的遗传多样性。 基于受选择的区域与中性预期或与基因组平均水平的不一致，很多检测自然选择的方法已经被发展出来，包括利用等位基因频谱的变化，增多的高频突变，群体差异，延长的单体型纯和度等[9,27,28]。

1）等位基因频谱的变化

衡量 DNA 多态性最常见的参数有：分离位点（segregating sites)，表示为 K，指所取 DNA 样本中具有不同碱基状态的位点数；任意两 DNA 序列之间核苷酸的平均差异数目(mean number of nucleotide differences)，表示为Π. 等位基因频谱（allele frequency spectrum)的变化是指时间上等位基因频谱的分布差异。在实际应用中，一般用几个衡量 DNA 多态性的参数对群体遗传参数 θ 进行估算。在中性状态下，不同的方法计算出的 θ 应该相等。当几个参数估算出来的 θ 差异很大时，可以认为不符合中性，即受到自然选择。Tajima 是第一个提出这个原理并根据这个原理提出的检测自然选择的方法 Tajima’s D[29]，随后一系列类似的方法被提出来，包括 Fu and Li’s D[30] 和 Fay and Wu’s H[31]。

事实上，人口统计学历史的变化(demographic history)也会改变遗传多态性的模式，并产生类似于自然选择一样的痕迹。比如，群体经历瓶颈事件将增加中间频率的多态性位点，这种特种类似于平衡选择留下的痕迹。群体扩张增加低频变异，很像负向自然选择和正向自然选择留下的痕迹 [32]。但是人口统计学历史的变化对全基因组每个位点的影响是相似的，但是自然选择只影响基因组上的一些区

域。这样， 比较一个区域的等位基因频谱与基因组上其它区域之间的差异可以帮我们寻找自然选择的信号。

2）衍生等位基因频率的增高（increased derived allele frequencies）

相对于祖先等位基因，衍生等位基因（即新突变）一般只有很短的历史，其频率一般也很低。由于搭车效应，发生过选择扫荡的区域将会有很多高频的突变。对黑猩猩和其他灵长类动物的全基因组测序极大地方便了我们寻找人类多态性位点的祖先等位基因。虽然很少有研究仅仅根据增高的衍生等位基因来判断其是否受到自然选择，但是衍生等位基因的频率被很多研究整合到其它统计量中，用来判别综合判断一个基因是否受到自然选择 [30,31,33]。

3）多态性变异偏离种间差异

在中性状态下，种内多态性和种间差异大多是相同的突变导致的，它们的相关性会很高。比如，如果一个区域的突变率很高，这个区域的多态性会增高，同时高的突变率决定了物种间高的碱基替代率，最终也会导致高的物种差异。与此相反，较低的突变率会导致较低的种内多态性和种间差异。种内多态性和种间差异的不一致反映了其偏离了中性期望，既有可能受到过自然选择。相对于一定的种间差异，较低的多态性可能反映了最近的正向自然选择。相对升高的多态性反映了负向自然选择。 Hudson-Kreitman-Aguade (HKA) 检验 [34] 和 McDonald-Kreitman (MK) 检验[35] 是根绝这个原理发展起来的自然选择检测方法，其中 HKA 检验比较分析多个基因座，而 MK 检验比较分析一个基因座上的非同义突变和同义突变。

4）延长的单体型纯和度（Extended haplotype homozygosity）

染色体上处于连锁不平衡的核苷酸的特定排列组合称为单体型。在中性条件下，新的突变要经过很长时间才能漂变到很高的频率。在这个过程中，重组将会不停的打断携带该突变的单体型并使得其延伸长度变短。而选择扫荡使得优势等位基因频率在很短的时间内迅速上升，这样重组来不及把受选择的优势单体型打断。所以，如果某单体型在群体中的连锁不平衡比与其相应等位基因的其它单体型延伸长，那么这个单体型很可能经历了正向自然选择[36]。根据这个原理发展的统计量包括 LRH[36]， iHS[24]， XP-EHH 和 LDD 等[15, 24, 37-40]。

5）群体差异

任意两个群体总的差异是由群体历史和遗传漂变决定的，这种影响对每个点都是相同或相似的[41, 42]。但是，各地完全不同的环境可能会形成不同的选择压力，这种不同的选择压力可能作用于基因组的某个区域，使得该区域的群体差异相对变大[43-45]。第一个根据群体差异在全基因组水平寻找正向自然选择信号的研究比较了欧亚非三大洲的群体差异系数 F_{ST}[44]，该方法可以检测到群体分离后不同环境导致的群体差异。纯和单体型延伸长度在群体之间的差异（cross-population extended haplotype homozygosity; XP-EHH）也可以检测近期的正向自然选择[15, 38]，该方法有很高的统计效力来检测最近一千个世代内的自然选择。

6）似然比检验（Likelihood-ratio test）

似然比检验是一种拟合优度检验(Good of fitness)，通过计算样本数据在搭车效应和中性进化下的似然函数比值，来判断选择扫荡是否发生。其理论依据是搭车效应会在选择靶点附近形成特异的多态性低谷，通过似然函数就可以检验 DNA 上出现的低谷是否违背中性假说。

7）复合检验

利用多种统计量来检测自然选择称为复合检验，这种策略能验证发现的自然选择信号是否可靠，即降低假阳性。最近， Grossman 等[33]分析了 5 个检测自然选择的统计量（F_{ST},XP-EHH,iHS,iHH,DAF），他们发现统计量的值在选择靶点上高度相关，而在其它位点上的信号是不相关的。他们根据这种现象提出了检测选择靶点的一种混合似然检验——多信号混合检验（CMS；composite of multiple signals）：

$$CMS=\prod_{i=1}^{n}\frac{P(s_i\mid selected)\times\pi}{P(s_i\mid selected)\times\pi+P(s_i\mid unselected)\times(1-\pi)}$$

其中，π是均匀的先验概率；对于每一个统计量 i,它的概率 P 是根据其值 s_i 在模拟数据中是否受到自然选择来计算。当把 CMS 应用到国际人类单体型计划的数据（HapMap）中，能够把受选择靶点缩小到 55kb 的范围内，并且能够直接检测到一些受选择的靶点，其中一些是没有被报道过的。

8）在全基因组水平检测自然选择

在一定群体参数下，我们可以根据中性进化理论得到一个自然选择统计量的预期值域，随后通过比较检测位点的值是否偏离了中性进化来检测自然选择。但是一般人群会经历很复杂的群体历史，即我们很难推测中性预期是多少，这样我们就很难检测单个位点是否受自然选择。但是全基因组数据的出现为我们提供了一个检测自然选择的新理念：我们只要在全基因组范围内计算出一个统计量在每个位点上或每个区域内的值，其值处于两个极端的位点即有可能受到自然选择（Figure 3）。

混合人群及检测其受到的自然选择

混合人群一般是由彼此隔离的祖先人群在近期发生大规模基因交流产生。对于研究适应性进化，人不是一个好的研究对象，至少是一个不容易操作更不能任意设计实验的研究对象。然而，混合人群为我们的研究提供了难能可贵的材料。人群在经历了长期隔离后发生大规模的融合，不但显著地增加了个体杂合度和人群遗传多样性，而且对人体基因表达模式和性状变异都产生了深远的影响。因为新人群面临的一切环境都是新的，混合人群从一开始就面临着对环境适应的压力。在个体水平上，体现在杂合度增高，遗传多样性增加。人群的混合导致风险等位基因在同一个个体中显著富集，对个体的表型性状或疾病状态产生重要影响。

混合人群的自然选择研究始于 1963 年[46]等对美国黑人的研究。他们在 1287 个美国黑人里面检测了 12 个基因，发现欧洲遗传成分大约占美国黑人的 10%。有几个位点欧洲的贡献很高，作者就认为可能是自然选择导致的。但是这些信号在以后的研究中没有被重复出来。由于统计方法和数据的限制，该领域在以后的很多年一直没有多少进展。直到 2007 年，Tang 等[47]用自己发展的新方法对波

多黎各人的祖先成分进行估算，发现 HLA 等几个区域的祖先贡献显著偏离平均值。该研究从新激起人们对混合人群自然选择研究的兴趣并引发了一些列讨论。

居住在中国西北的人群如新疆维吾尔族和哈萨克族历史和来源复杂、遗传多样性高，大多数人群都表现出不同程度的中西方混合。考古学证据表明，有西方人特征的人群，大约在 4000 年前定居于如今的新疆地区。同时，也已经有遗传学证据表明，已经分化数万年的东西方人群在此汇聚，形成新的人群[48-50]。这种新的混合人群不仅在体质外貌上表现出东西方融合的特征，其基因组上也表现出更高的遗传多样性。近几年我们对西北混合人群进行了系统地分析，如发展了单倍型共享（haplotype-sharing），重构维吾尔族人群形成的遗传模型，通过比较真实数据与计算机模拟数据的单倍型共享模式，揭示了新疆维吾尔族不是在欧亚人群分开之前就已经存在的古老人群，而是由具有白人血统和东亚血统的祖先混合产生，他们各自对现代维吾尔族贡献了约 50%的遗传成分 [48,50]。以单次混合模型为基础，推测混合发生在大约 2000 年前。既然是彼此分化几万年的东西方人群的遗传混合，必然意味着至少一方，甚至双方的祖先是在近期涉足这片新的土地，这里的一切，如果有不同于祖先人群的原居住地，则意味着面临着新的环境。长期隔离的人群汇合在一起，已经各自固定的等位基因在个体基因组中重新组合，对融合人群的基因组遗传多样性以及代谢和多种表型可能会产生显著影响。新形成的混合人群在西北地区从新繁衍了一两百个世代，足以让一个新的人群适应新地域的自然地理和气候环境，在遗传结构上与祖先人群可能有了很大改变，为在现代人群中研最近的自然选择提供了天然素材。

图示

Figure 1. Global distributions of candidate polymorphisms for skin characteristics, which are adapted from Sabeti [15]. a, *SLC24A5* A111T is common in Europe, Northern Africa and Pakistan, but rare or absent elsewhere; b, *EDAR* V370A is common in Asia and the Americas, but absent in Europe and Africa.

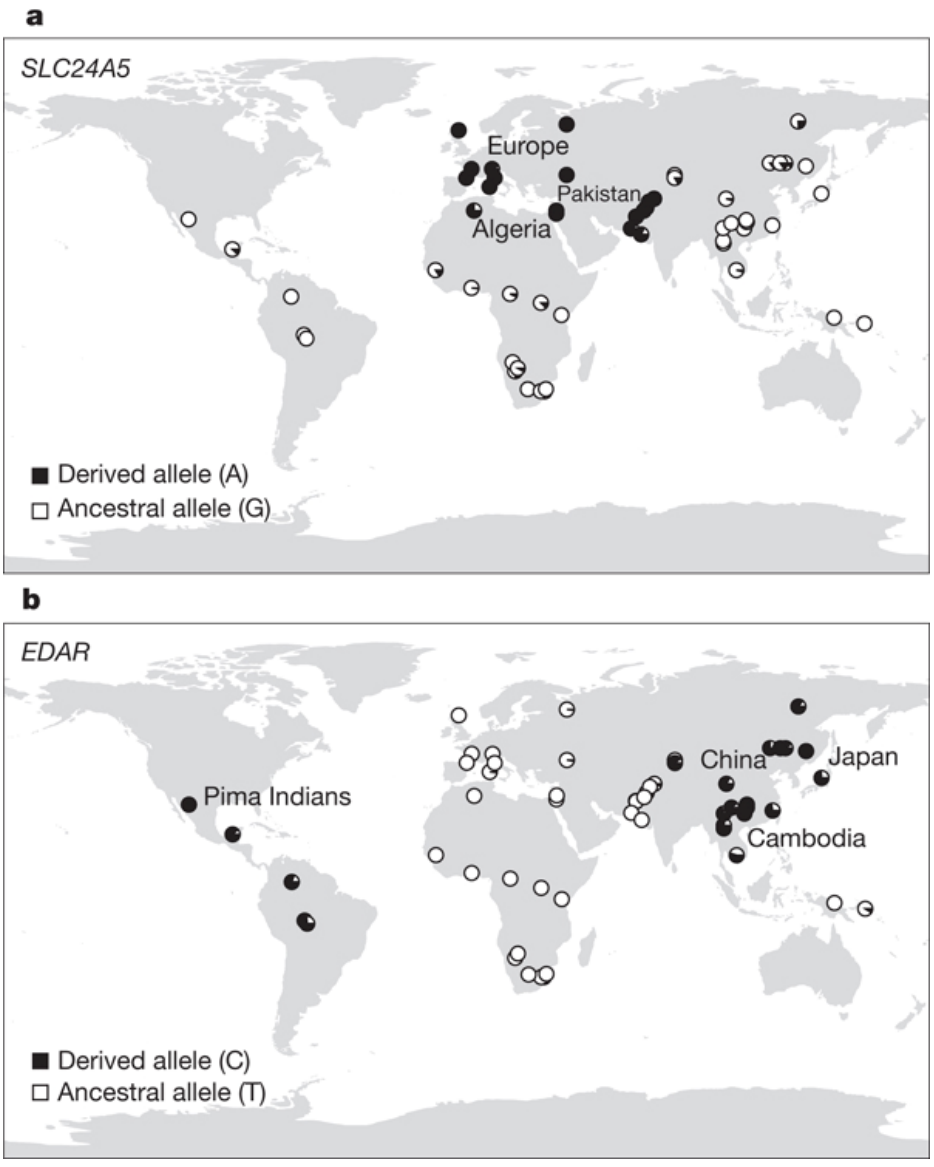


Figure 2. ***EPAS1* and *EGLN1* are the strongest signals for Tibetan adaptation to high altitude.** The figure is obtained from *Xu et al.*[16]. (A) Genomic distribution of locus-specific F_{ST} between Tibetan and Han Chinese, (B) Genomic distribution of XP-CLR score. *EGLN1* and *EPAS1* were identified showing significant high F_{ST} values (>0.30, top 0.01%) and high XP-CLR scores (>100, top 0.01%).

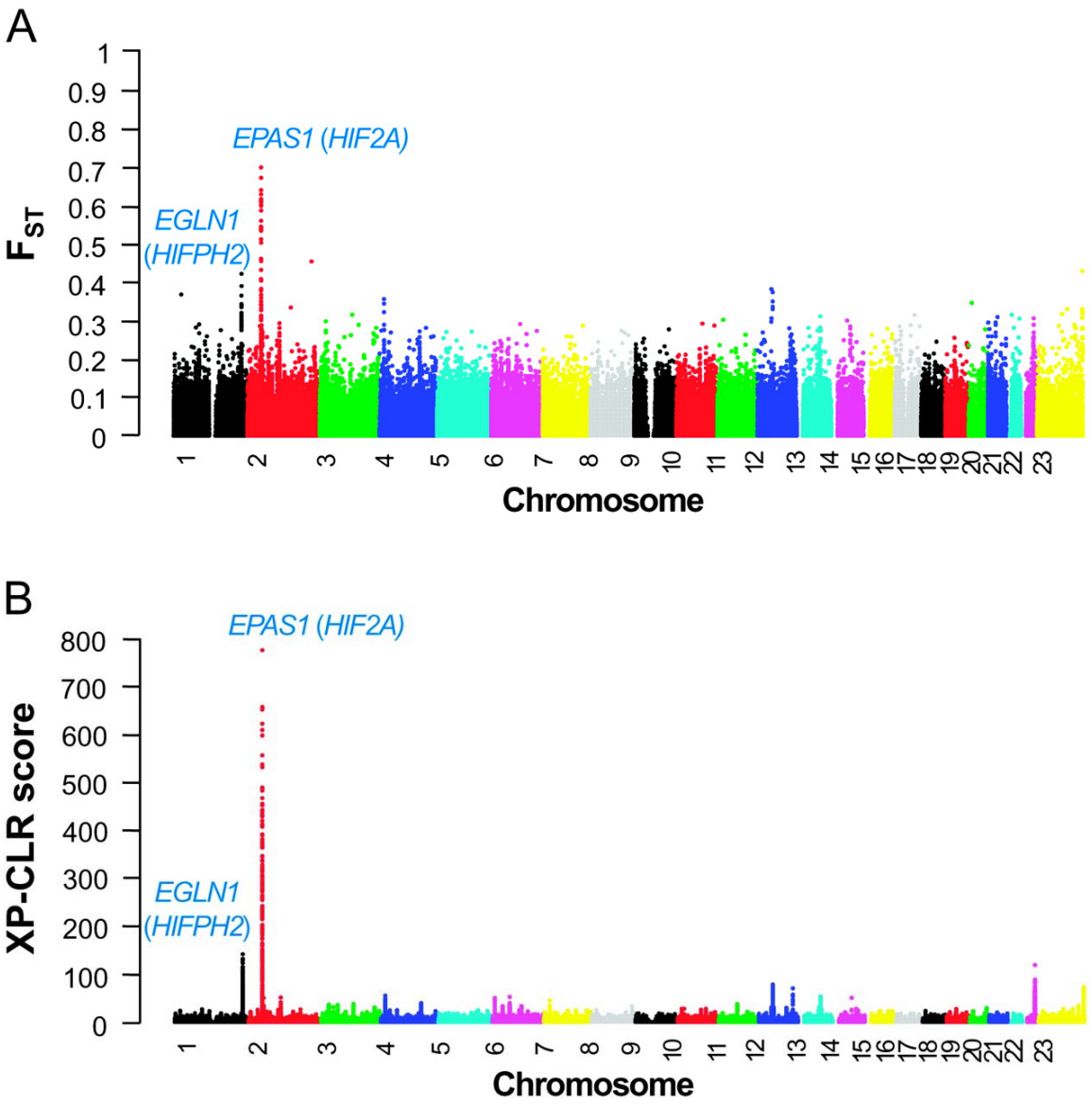
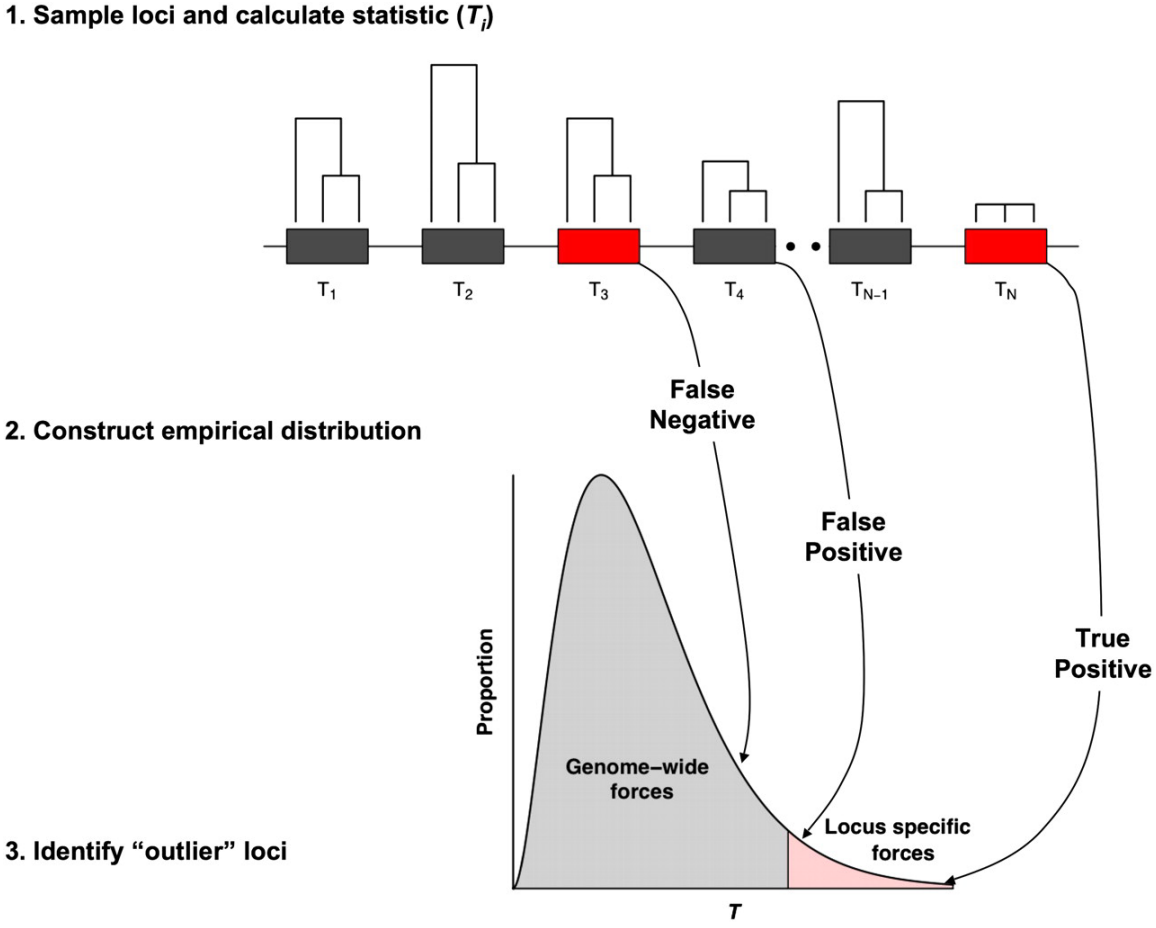


Figure 3. **Schematic of "outlier of summary statistics" for detecting natural selection, obtained from study by Akey [51].** This approach usually set a threshold at the tail of empirical distribution and is commonly used when genome-wide data are available.



参考文献

1. McHenry HM (2009) Human Evolution. In: Travis MRJ, editor. Evolution: The First Four Billion Years. Cambridge, Massachusetts: The Belknap Press of Harvard University Press. pp. p. 265.

2. Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325: 31-36.

3. Ke Y, Su B, Song X, Lu D, Chen L, et al. (2001) African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes. Science 292: 1151-1153.

4. Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC (1991) African populations and the evolution of human mitochondrial DNA. Science 253: 1503-1507.

5. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, et al. (2010) A draft sequence of the Neandertal genome. Science 328: 710-722.

6. Reich D, Green RE, Kircher M, Krause J, Patterson N, et al. (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468: 1053-1060.

7. Reich D, Patterson N, Kircher M, Delfin F, Nandineni MR, et al. (2011) Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. Am J Hum Genet 89: 516-528.

8. Rasmussen M, Guo XS, Wang Y, Lohmueller KE, Rasmussen S, et al. (2011) An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia. Science 333: 94-98.

9. Xu S, Jin W (2012) Population Genetics in the Genomic Era Population Genetics. Winchester, United Kingdom: InTech

10. 季林丹, 徐进, 张亚平 (2012) 人类群体环境适应性进化研究进展 科学通报 57: 112-119.

11. Balaresque PL, Ballereau SJ, Jobling MA (2007) Challenges in human genetic diversity: demographic history and adaptation. Hum Mol Genet 16 (R2): R134-139.

12. Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20: R208-215.

13. Jablonski NG, Chaplin G (2010) Human skin pigmentation as an adaptation to UV radiation. Proceedings of the National Academy of Sciences of the United States of America 107: 8962-8968.

14. Parra EJ (2007) Human pigmentation variation: Evolution, genetic basis, and implications for public health. American Journal of Physical Anthropology: 85-105.

15. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-918.

16. Xu S, Li S, Yang Y, Tan J, Lou H, et al. (2011) A genome-wide search for signals of high-altitude adaptation in Tibetans. Mol Biol Evol 28: 1003-1011.

17. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, et al. (2010) Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329: 75-78.

18. Simonson TS, Yang Y, Huff CD, Yun H, Qin G, et al. (2010) Genetic evidence for high-altitude adaptation in Tibet. Science 329: 72-75.

19. Aggarwal S, Negi S, Jha P, Singh PK, Stobdan T, et al. (2010) EGLN1 involvement in

high-altitude adaptation revealed through genetic analysis of extreme constitution types defined in Ayurveda. Proc Natl Acad Sci U S A 107: 18961-18966.

20. Peltonen L, Enattah NS, Sahi T, Savilahti E, Terwilliger JD, et al. (2002) Identification of a variant associated with adult-type hypolactasia. Nature Genetics 30: 233-237.

21. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. Nature Genetics 39: 31-40.

22. Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, et al. (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet 39: 1256-1260.

23. Moyzis RK, Wang ET, Kodama G, Baidi P (2006) Global landscape of recent inferred Darwinian selection for Homo sapiens. Proceedings of the National Academy of Sciences of the United States of America 103: 135-140.

24. Voight BF, Kudaravalli S, Wen XQ, Pritchard JK (2006) A map of recent positive selection in the human genome. Plos Biology 4: 446-458.

25. Casanova JL, Quintana-Murci L, Alcais A, Abel L (2007) Immunology in natura: clinical, epidemiological and evolutionary genetics of infectious diseases. Nature Immunology 8: 1165-1171.

26. Snow RW, Guerra CA, Noor AM, Myint HY, Hay SI (2005) The global distribution of clinical episodes of Plasmodium falciparum malaria. Nature 434: 214-217.

27. 林栲，李海鹏 (2009) DNA 水平上检测正选择方法的研究进展. 遗传 31: 896-902.

28. 周琦, 王文 (2004) DNA 水平自然选择作用的检测. 动物学研究 25: 73-80.

29. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

30. Fu YX, Li WH (1993) Statistical Tests of Neutrality of Mutations. Genetics 133: 693-709.

31. Wu CI, Fay JC (2000) Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413.

32. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, et al. (2005) Genomic scans for selective sweeps using SNP data. Genome Research 15: 1566-1575.

33. Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, et al. (2010) A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. Science 327: 883-886.

34. Hudson RR, Kreitman M, Aguade M (1987) A Test of Neutral Molecular Evolution Based on Nucleotide Data. Genetics 116: 153-159.

35. Mcdonald JH, Kreitman M (1991) Adaptive Protein Evolution at the Adh Locus in Drosophila. Nature 351: 652-654.

36. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-837.

37. Zhang C, Bailey DK, Awad T, Liu GY, Xing GL, et al. (2006) A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations. Bioinformatics 22: 2122-2128.

38. Tang K, Thornton KR, Stoneking M (2007) A new approach for using genome scans

to detect recent positive selection in the human genome. Plos Biology 5: 1587-1602.

39. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19: 826-837.

40. Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, et al. (2007) Localizing recent adaptive evolution in the human genome. PLoS Genet 3: e90.

41. Kimura M (2003) The neutral theory of molecular evolution. (United Kingdom: Cambridge University Press, Cambridge).

42. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38: 1358-1370.

43. Wright S (1951) The Genetical Structure of Populations. Annals of Eugenics 15: 323-354.

44. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome research 12: 1805.

45. Lewontin RC, Krakauer J (1973) Distribution of Gene Frequency as a Test of Theory of Selective Neutrality of Polymorphisms. Genetics 74: 175-195.

46. Workman PL, Blumberg BS, Cooper AJ (1963) Selection, Gene Migration and Polymorphic Stability in a U. S. White and Negro Population. Am J Hum Genet 15: 429-437.

47. Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, et al. (2007) Recent genetic selection in the ancestral admixture of Puerto Ricans. Am J Hum Genet 81: 626-633.

48. Xu S, Huang W, Qian J, Jin L (2008) Analysis of genomic admixture in Uyghur and its implication in mapping strategy. Am J Hum Genet 82: 883-894.

49. Xu S, Jin L (2008) A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. Am J Hum Genet 83: 322-336.

50. Xu S, Jin W, Jin L (2009) Haplotype-sharing analysis showing Uyghurs are unlikely genetic donors. Mol Biol Evol 26: 2197-2206.

51. Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19: 711-722.

第二章：祖先染色体片段分布揭示的人群混合历史

摘要

人群混合的历史和过程将决定混合人群的单体型结构和连锁不平衡模式，而连锁不平衡模式对于包括筛选疾病风险位点和自然选择信号在内的多种医学和进化研究具有重要作用。但是，很多之前的研究都是基于非常简单的模型以至于根本没有考虑混合过程本身的复杂性。在这里，我们提出两种研究群体混合动力和过程的方法：祖先染色体片段分布法和个体祖先比例分布法。通过对全基因组的模拟数据和经验数据进行分析，我们发现来自于不同祖先的染色体片段分布比个体祖先比例分布有更好的统计效力，较少的受到抽样误差的影响，能追溯更久远的历史。对1890 个美国黑人的分析表明，美国黑人在 14 个世代内连续接受欧洲人群基因流模型 （continuous gene flow model）比其它模型能更好的解释观察到的祖先染色体片段分布。我们还发现一些美国黑人的欧洲祖先成分比任何模拟的样本要高，表明美国黑人人群存在亚结构，这种亚结构必然由一些个体持续的与欧洲人群婚配导致的。相对于其它模型，我们还发现墨西哥人的混合更加符合 24 个世代的渐混模型（gradual admixture model）。总之，这个研究不仅提供了研究群体混合动力的方法，也丰富了我们对混合人群如美国黑人和墨西哥人的人群历史的认识。

前言

当之前相互隔离的人群相遇并通婚，就会产生新的混合人群，这种现象在人类历史中可能是一种常见现象[1-5]。在过去的几个世纪里，人类的大规模迁徙导致了不同大洲人群之间的通婚和混合。现在遗传学界界对近期混合人群如美国黑人和拉美混血人群研究越来越多，主要是因为他们在疾病基因的筛选中有很重要的优势，特别是基于混合人群筛选疾病基因的方法混合作图（admixture mapping）已经被发展出来[6-9]。混合作图的统计效力在于祖先人群等位基因频率不同产生了延长的和增强的连锁不平衡[6, 10, 11]。但是群体混合过程决定了混合人群的连锁不平衡模式。这意味着群体混合动力对混合作图的统计效力产生重大影响[11-14]。

事实上，精确地估算群体混合动力不仅仅有利于混合作图，也会促进很多其它人群相关研究，如在混合人群中筛选受自然选择位点[12, 15]和阐明人群历史[16]。虽然已经有些研究分析了模拟数据[2, 17]和低密度的实验数据[11, 14]，但是我们对混合人群的精细混合动力仍然不清楚。近期全基因组高密度核苷酸多态性位点的研究极大地丰富了我们对混合人群精细遗传结构的认识[18-25]。但是，这些研究并没有考虑人群混合过程本身的复杂性。更可惜的是，由于分析上的挑战，重组塑造的混合人群的单体型和祖先染色体片段却被故意忽略[18]。

对于来自有多个世代的混合群体的个体，他们的染色体就像来至不同祖先群体的染色体片段组成的马赛克。在每一个世代，重组将会重塑混合人群中的祖先染色体片段，这些祖先染色体片段的分布将能提供更详细的群体历史信息。详细的

说，随着混合代数的增加，祖先染色体片段会被切的更小。而近期混合个体的染色体将包含很多很长的祖先染色体片段。很多研究已经利用平均的祖先染色体片段估算群体混合的世代数[19-23]。祖先染色体片段的分布可能能够提供群体混合历史和混合动力的更多信息，但之前并未有人涉足。

在本研究中，我们首先进行向前模拟来分析三个完全不同的模型下混合动力对祖先染色体片段分布的影响。我们发现祖先染色体片段分布能够提供群体混合动力的精细信息。并且该方法比较稳健，尤其对各种干扰因数如遗传标记密度，混合人群群人口统计学的变化和其祖先人群人口统计学的变化不敏感。随后我们把该方法应用到几个近期混合人群如美国黑人和墨西哥人。我们不仅揭示这些群体和混合动力学，还发现它们在群体一些群体亚结构。

材料和方法

数据和群体样本

本研究分析了美国黑人人群和墨西哥人人群两个混合人群及其祖先群体的3398 个个体的基因型。这些基因分型数据分别从国际人类单体型计划（International Haplotype Map Project; HapMap, http://www.hapmap.org）[24,25]， 人类基因组多样性计划（Human Genome Diversity Project; HGDP, http://www.cephb.fr/en/hgdp）[26]，美国国立普通医学科学研究所的人类变异参照组 (Human Variation Panels from National Institute of General

Medical Sciences; HVP），墨西哥遗传多样性计划（Mexican Genetics Diversity Project; MGDP, ftp://ftp.inmegen.gob.mx/）[27] 和 Illumina 病例对照数据库（iControlDB; http://www.illumina.com）获得。我们最终获得 5 个 HapMap 人群的 580 个样本：包括 87 个来自美国西南的美国黑人（ASW），167 个采自尼日利亚的伊巴丹的约鲁巴人（YRI），165 祖先来自北欧和西欧的美国犹他州居民（CEU），77 个墨西哥裔洛杉矶居民（MEX）和 84 个北京汉族（CHB）[24,25]。除此之外，我们还获得了 iControlDB 中的 2161 个美国黑人 illumina 550K 微珠芯片基因分型数据，HVP 中的 100 个美国墨西哥人 Affymetrix SNP 6.0 基因芯片数据，MGDP 中的 300 个墨西哥混合人群样本和 30 个印第安人的 Affymetrix 100K 基因芯片数据，以及 64 个印第安人的 illumina 650K 微珠芯片基因分型数据。为了在研究中保持保持尽可能多的遗传标记，我们没有把所用样本放在一起作分析。而是把每一个混合人群和其祖先人群的基因型合并到一起做分析。对于每一组数据，我们去除有血缘关系的个体，去除数据缺失大于 10%的个体和数据缺失大于 10%的位点。

混合模型和数据模拟

实际的群体混合过程可能很复杂，我们可能根本不知道也很难直接进行分析。在本研究中，我们首先分析了三种典型的简化的混合模型（Figure 1）：混合隔离模型（Hybrid isolation model; HI model）[6]，渐混模型（gradual admixture model; GA model）[13]和持续基因流模型（continuous-gene-flow model; CGF model）[11,12]。这三种混合模型基本上代表了所有主要的混合方

式，其形成的遗传结构和连锁不平衡模式已经在之前的研究中有探讨[11-13, 17]。在这三种模型中，m_1 和 m_2 分别代表两个祖先人群 pop1 和 pop2 对混合人群的遗传贡献。t 代表了开始混合到现在所经历的世代。在混合隔离模型中，混合只发生在第一个世代，之后两个祖先人群对混合人群不再有任何直接贡献。混合人群只受到之后的重组和遗传漂变的影响。在渐混模型中，两个祖先人群 pop1 和 pop2 的遗传成分在每一个世代分别以固定的速率 m_1/t 和 m_2/t 贡献给混合人群，同时混合人群的上一代贡献了其它遗传成分。在 G_i 代，混合人群的上一代即 G_{i-1} 对混合人群的贡献为$\beta = (G_{i-1})/t$。持续基因流模型可以被看做渐混模型的一种扩展。在这个模型中，一个祖先人群将持续的接受另一个祖先人群持续的恒定的单向基因流$\alpha = 1-(m_1)^{1/t}$。我们将试图通过分析这三个模型下的祖先染色体片段分布来探讨其群体混合动力。

我们根据上面提到的三个模型发展了模拟常染色体数据的程序。我们用相位已知的 YRI 和 CEU 基因型作为祖先人群的基因型。根据它们在每一代对混合人群的贡献分别抽取相应的数量的单体型，再用这些抽取的单体型重构出混合人群。 因为模拟的群体历史较短，我们忽略群体混合之后的突变。在这个模拟中，我们根据 HaMap (release #22)的重组图提供的重组率引入重组。每个群体的有效群体大小（N_e）都设为 10000。 我们在模拟数据时对混合人群的每个祖先染色体片段进行标记。在持续基因流模型中，对于一个给定的混合比例，祖先人群接受基因流和提供基因流经历的群体历史是不一样的。这样，我们把提供基因流的祖先人群称之为 CGFD；把接受基因流的祖先人群称之为 CGDR。每个模型混合的世代都分别设置成 10, 20, 50, 和 100。为了分析各个因素对祖先染色体片段分布的影响，我们又做

了很多相应的模拟。

群体遗传学分析和祖先染色体片段分布的推断

主成分分析（principal component analysis; PCA）是将很多变量通过线性变换为少数几个重要变量的一种多元统计分析方法。为了减弱位点之间连锁不平衡对主成分分析的影响，我们在基因组上平移一个包含50个单核苷酸多态性位点的窗口，每次把窗口内r^2 >0.5的位点去除并向前平移5个位点。对全基因组位点进行过滤之后，我们利用EIGENSOFT[28]对所有样本在个体水平做主成分分析。因为采用期望最大化法（expectation-maximization; EM)的FRAPPE[29]能够利用所有的位点，我们利用它估算每个个体的祖先成分和混合比例。另外我们也使用了一个综合了贝叶斯（Bayesian）和马尔科夫蒙特卡洛（Markov chain Monte Carlo；MCMC）来对个体进行群体聚类的方法，这个方法可以在现成的软件STRUCTURE中实现[30]。对于STRUCTURE分析，因为连锁不平衡的位点可能会被认为人群混合所致的，我们把任意两个位点之间距离小于一百万bp的位点去除。

自从高密度的单核苷酸多态性数据大量出现以后， 很多基于这种数据推断祖先来源的方法发展起来。比如 SABER [22], LAMP与LAMP-ANC [32], HAPAA [30] 和 HAPMIX [21]。我们首先依据我们已有的基因分型数据模拟出一套数据，然后用这一套模拟数据来评估各个方法的准确度。我们发现HAPMIX比其他几种方法有更高的准确度。这样，我们在以后的分析中使用HAPMIX对每个位点的祖先来源进行推断。对于祖先人群的单体型， 我们直接从HapMap下载或用fastPHASE推断[33]。

分布差异的度量

柯尔莫哥洛夫 - 斯米尔诺夫检验（Kolmogorov - Smirnov test）[34]是用累计次数或累计频数来判断两个分布函数之间是否纯在显著差异的一种非参数检验。我们用柯尔莫哥洛夫 - 斯米尔诺夫检验评估不同模型下及实际数据的祖先染色体片段分布是否有显著不同。同时，为了量化两个分布之间的差异，我们计算了两个分布之间的推土机距离（earth mover’s distance; EMD）[35,36]。推土机距离是评估两个概率分布相似度的一种量化方法。形象的说，如果有两个分布，一个分布可以看成空间上的一堆土，另外一个分布可以看成是在相同空间上的一个洞穴。推土机距离就是计算用这堆土填充空间上的洞穴所需要的最少工作量。这样，推土机距离的值对应着完成这两个分布转换所需要的运输量乘以运输距离。推土机距离较小意味着两个分布相似度较高。

对美国黑人样本的分析

虽然欧洲人群和非洲人群贡献了美国黑人基因库的的绝大多数遗传成分，但是美洲印第安人和东亚人对某些美国黑人的基因成分也有少量贡献。为了把美国黑人的混合分析简化成两个祖先人群的混合，我们首先去除了包含明显印第安人或东亚人成分的样本。并对样本进行了一系列过滤，我们最终获得1890个美国黑人样本和其可能祖先人群的354参考样本，每个样本包含491557个常染色体单核苷酸多态性位点。在这1890个美国黑人样本中，52个来自于人类单体型计划的ASW,1838个来自于iControlDB。四个分别来自四个大陆的一般人群（112 CEU, 112 YRI, 84

CHB, and 44 AMI）代表了美国黑人的可能祖先人群。

我们用PLINK[37]过滤高度连锁的单核苷酸多态性位点，把最初的数据减少到341672个单核苷酸多态性位点。我们利用过滤剩下的位点对样本在个体水平进行主成分分析。并用FRAPPE进行群体结构和个体祖先成分的估计。为了用HAPMIX推断美国黑人的祖先染色体片段，我们分别用88个YRI和88个CEU的单体型数据作为美国黑人的非洲祖先人群和欧洲祖先人群。根据FRAPPE的结果，我们设定欧洲人群对美国黑人的贡献为21.65%。根据美国黑人的历史，我们设定一系列群体混合世代，我们把取似然值最大的混合世代作为美国黑人估算的混合世代。

我们根绝估算的参数用自己的程序模拟美国黑人的群体混合过程，通过比较实际观察到的祖先染色体片段分布与模拟的分布来寻找最合适的模型。我们不是对每个祖先人群的祖先染色体片段分别单独进行考虑，而是对非洲人群和欧洲人群祖先的祖先染色体片段进行综合考虑。对于持续基因流模型，我们把欧洲人群持续贡献基因流给非洲人群形成美国黑人的模型称之为CGF1模型。非洲人群持续贡献基因流给欧洲人群称之为CGF2。我们根据国际人类单体型计划给定的有效群体来设定各个群体的有效群体大小。具体来说，我们分别设定非洲人群，欧洲人群和美国黑人的有效群体大小为17094, 11418 和17094。根据在这1890个样本中观察的结果，我们设定欧洲人群对美国黑人的贡献为21.65%。我们设定美国黑人混合的世代从10到17，每间隔一个世代做一次模拟。无论对于模拟数据还是实际数据，我们分别去除祖先染色体片段长度<0.5 cM的非洲祖先片段和<0.8 cM的欧洲祖先染色体片段。我们计算了观察数据的祖先染色体片段分布与各个模型的分布之间的推土机距离。当

模拟数据的分布与观察数据的分布之间的推土机距离达到最小时，我们认为这个模型与观察数据最吻合。

对墨西哥人群的分析

墨西哥人群是在混合作图中最常用另一个混合人群。我们获得了 458 墨西哥人群样本，包括 300 个墨西哥遗传多样性计划的样本，100 个人类变异参考样本，和 58 个人类单体型计划的样本。 我们对墨西哥人群分析的方法和对美国黑人的分析方法是一致的，只是使用了与墨西哥人群相对应的参数。

结果

祖先染色体片段分布可以揭示群体混合动力

我们首先分析了在三个不同的混合模型下（混合隔离模型，渐混模型和持续基因流模型）[11-13,17]祖先染色体片段分布的不同（Figure 1）。在祖先贡献比例一定的持续基因流模型中，我们称在每个世代持续贡献遗传成分给混合人群的祖先人群为 CGFD。 只在第一代提供了遗传成分的祖先人群称为 CGFR。当我们设定祖先人群对混合人群的贡献为 50%时，我们发现每个模型下的祖先染色体片段分布在经历相同的世代之后都显著不同($P < 2.2 \times 10^{-16}$, two-sample K-S test)。我们同时发现混合隔离模型下的祖先染色体片段相对较小。这使得其分布与其它模型

下的分布差异较大。我们推测祖先染色体片段分布服从对数正太分布，随后对所有祖先染色体片段长度取对数。经过这个转化后，各种模型下的祖先染色体片段分布（Figure 2）仍然显著不同($P < 2.2 \times 10^{-16}$, two sample K-S test)。根据任意两个分布之间的推土机距离，我们知道在所有的模拟分布中渐混模型下的祖先染色体片段分布和 CGFR 下的分布最相似；混合隔离模型下的分布和 CGFR 下的分布差异最大。在每个模型下，不同世代的两个分布之间的推土机距离随着混合世代的增加而增加。当我们不考虑很短的祖先染色体片段，各个模型下的分布仍然有显著差异。

我们在随后的模拟中改变祖先人群对混合人群的贡献比例，观察祖先染色体片段分布的变化。我们发现祖先人群的贡献比例很低时，相同世代的渐混模型祖先染色体片段分布与 CGFD 的分布很相似。但是当祖先人群的贡献比例很高时，渐混模型的分布更像 CGFR 的分布。当祖先人群持续的贡献遗传成分到混合人群中（渐混模型和 CGFD），混合人群中会保留一些很长的祖先染色体片段。在相同的世代下，混合隔离模型下的祖先染色体片段总是比其它模型包含更多的短片段。简言之，这些观察表明祖先染色体片段分布包含很多群体混合动力的信息。

个体混合比例反映的群体混合动力

混合人群的个体混合比例是群体遗传中最常用参数之一，并且可以通过很多软件直接计算出来[23, 38-40]。我们首先观察在各个模型下个体混合比例的分布。当设定祖先人群对混合人群的贡献为 50%时，我们发现在混合世代等于 10 时，个

体混合比例的分布在每一个模型下都显著不同 (Figure 3A; $P < 2.2 \times 10^{-16}$, two-sample K-S test)。但是随着世代的增加，任意两个分布之间的差异逐渐减少(Figure 3)。最终，任意两个分布之间的差异都不再显著(Figure 3D)。对任意两个分布之间的推土机距离的分析支持我们的观察到的结果。当我们改变祖先人群对混合人群的贡献比例，我们发现个体混合比例的分布总是能揭示近期混合人群的混合动力；但不能揭示经历了几十个世代以上的人群混合动力。理论上，虽然这个方法简单可行，但是这个方法过度依赖大的样本量。

祖先染色体片段分布法的稳健性

在把祖先染色体片段分布法法应用到真实数据之前，我们先对其性质和影响因素进行分析。首先，我们发现混合人群及其祖先人群的有效群体大小对祖先染色体片段分布没有影响。同时群体扩张和瓶颈事件也不影响分布。这说明祖先染色体片段分布不受一般人口统计学数据的影响。但是祖先染色体片段分布受到混合群体亚结构（非随机婚配）的影响，这种亚结构使得分布变得更扁平一些。当连续的遗传标记之间的距离小于祖先染色体片段的长度时，祖先染色体片段分布不会受到影响。我们发现虽然 HAPMIX 统计上的错误对片段的平均长度影响很小，但是会使得分布变得平缓一些。为了降低局部祖先推测法本身的统计误差对结果的影响，我们对模拟数据和实际数据的祖先染色体片段一起进行推断。祖先染色体片段分布法利用重组产生的信息，其分布基本不受祖先人群的等位基因频率影响。由于每个混合人群的个体都包含很多祖先染色体片段，该方法只需要较少的样本就可以获得一个

比较可靠的分布。总之，该方法较少的受到群体人口历史统计学本身事件的影响，是个相对稳健的方法。

美国黑人群的混合动力和历史

因为美国黑人在混合作图中广泛使用，研究美国黑人的混合动力对于提高混合作图的效果具有重要作用。我们用通过质控的 1890 个美国黑人样本对其混合动力进行分析（Figure 4）。设定欧洲人群贡献了美国黑人 21.65%的遗传成分，直接根据似然估计算出美国黑人的平均混合世代是 7 代，这与之前的结果很相似[19,21,41-43]。但是自从 18 世纪奴隶贸易把美国黑人的非洲祖先带到美洲以来，据现在已经有 300 年（假设 20 年一个世代，已经有 15 个世代）。这样，根据基因组数据计算出的混合世代数似乎与历史记载相互矛盾。

为了解决基因组估算世代与历史记载的不一致，我们模拟了一系列美国黑人群体。在每个模型下，我们设定欧洲人群贡献了美国黑人 21.65%的遗传成分，设定的混合世代从 10 到 17，每次增加一个世代。实际数据中的欧洲祖先染色体片段分布和和非洲祖先染色体片段分布是根据 HAPMIX 的输出结果获得。我们比较了实际数据的祖先染色体片段分布与四个模型之间的差异：混合隔离模型，渐混模型，CGF1（欧洲人群持续的贡献基因流到非洲祖先人群/美国黑人人群中）和 CGF2（非洲人群持续的贡献基因流到欧洲祖先人群/美国黑人人群中）。在所有模拟数据中，14 个世代 CGF1 下非洲祖先染色体片段分布和欧洲祖先染色体片段分布与实际数据之间的推土机距离分别达到最小：分别是 0.0204 和 0.0239 (Figure 5)。因

此我们认为 14 个世代的 CGF1 和实际人群历史最匹配（Figure, 5A, 5C）。即欧洲人群持续的贡献基因流给美国黑人大约 14 代能很好的解释美国黑人的混合历史和过程。

虽然美国黑人实际的混合过程可能比我们模拟的要复杂很多，但是 14 个世代的 CGF1 已经捕获到了这其中最主要的模式。直接观察 14 个世代 CGF1 下的分布与实际数据的分布，虽然发现两者差异确实已经很小，实际数据的分布比模拟数据仍要平缓一些（Figure 5B, 5D）。这可能是由于实际数据中存在非随机婚配或较高的统计误差。考虑到最近 200 年很少有非洲人能够移民到美国，且美国白人在历史上不断地与美国黑人通婚的事实，CGF1 模型也很符合社会现实。除此之外，美国欧洲后裔与美国黑人通婚所生子女一般都归为美国黑人，相当于美国欧洲人群持续贡献基因流到美国黑人。既然欧洲遗传成分持续贡献到美国黑人中，美国黑人中的欧洲遗传成分在未来还会持续增长。

美国黑人的个体混合比例分布不能很好地匹配上任意一个模型（Figure 6）。这可能是由于美国黑人的样本量不够大，抽样有偏差或美国黑人的亚结构导致。仔细分析美国黑人的个体混合比例后，我们发现少量美国黑人的欧洲祖先成分比任何模拟的比例都多（Figure 6）。我们推测这可能由于一些美国黑人个体进入了美国白人社会，他们持续与美国欧洲后裔或欧洲成分为主的个体通婚导致的。其次，我们发现包含的欧洲成分或非洲成分越高（接近于 1）其混合世代越久远，欧洲成份和非洲成分的混合比例越接近 1:1 其混合世代越短（Figure 7）。

墨西哥人群的混合历史

墨西哥混合人群也是混合作图使用最多的群体之一。一般遗学传数据估算的墨西哥人人群平均混合世代是 15 个世代或更少[3, 44, 45]，但是实际上从第一个欧美混血儿到现在已经有 500 年的历史（25 个世代）。我们希望通过分析其混合动力来解决基因组数据估计的群体混合世代与记载的墨西哥人群混合历史之间的差异。我们去除含有明显非洲成分的的个体，对主要包含欧洲和美洲祖先成分的 413 个墨西哥人进行分析。在主成分分析中，4 个大洲人群（YRI, CHB, CEU 和 AMI）位于梯形图像的四个角上；墨西哥人群分散于欧洲人群和美洲印第安人之间的空间上（Figure 8）。我们估算欧洲人群对墨西哥人群的贡献是 49.2%。

我们模拟每个模型下群体混合分别经历 15-25 世代后的基因型，假设欧洲人群对墨西哥人群的贡献为 49.2%。我们用 HAPMIX 推出所有模拟数据和实际墨西哥人群的祖先染色体片段。我们发现混合隔离模型的祖先染色体片段分布与实际数据分布之间的推土机距离随着世代增大而增大，并且混合隔离模型下的分布应对于实际数据缺少长的染色体片段（Figure 9A, 9C）。对于 CGF1（欧洲人群持续的贡献基因流到墨西哥人群）或 CGF2（美洲印第安人群持续的贡献基因流到墨西哥人群），欧洲祖先染色体片段分布的最小推土机距离与美洲祖先染色体片段的最小推土机距离达到最小值的世代不一致。渐混模型下两个祖先人群的祖先染色体片段分布与观察分布的推土机距离比其它模型的都低，并且几乎同时达到最低，这意味着渐混模型与实际数据最匹配（Figure 9）。欧洲祖先成分和美洲祖先成分的最小推土机距离分别在渐混模型的 24 代和 23 代（EMD=0.0076 和 EMD=0.0163）。

讨论

跨种族婚姻受到各种社会，文化，经济，政治和地理因素的影响，如人群迁徙，重殖民，种族歧视，种族冲突和种姓制度等，各种因素交织使得群体混合过程变得极为复杂。这样，我们不能期待几个极其简单的混合模型能够很好的解释实际的混合过程。但是，为了方便混合作图，进化和其他医学研究，我们建议用简化的混合模式来概括群体混合过程。在本研究中，我们提出两种研究群体混合过程和动力的方法。从理论上看，个体混合比例分布有很高的统计效力来揭示近期混合群体的混合历史。但是，这种方法需要很大的样本量，并且极易受到采样误差的影响。祖先染色体片段分布只需要较少的样本量，能够解析各种历史背景下的群体混合动力。这种方法利用群体混合之后的重组信息，不受祖先等位基因频率和人口统计的变化影响。

通过混合作图，新大陆的混合人群如美国黑人已经被广泛的应用于疾病基因筛选。同时美国黑人的人群混合过程和历史对混合作图的统计效力具有极其重要的影响[11, 17, 46, 47]。很多以前的研究没有考虑群体混合动力学，大多通过简单的 6-8 代混合隔离模型对美国黑人的混合历史进行模拟[19, 42]。但是我们发现美国黑人的混合历史能够更好地被 14 个世代的持续基因流模型来解释，这样产生的连锁不平衡模式与之前的假设必定不同。我们建议以后的研究应该根据持续基因流模型对美国黑人进行模拟，这样才能真实反应美国黑人的连锁不平衡模式。我们同时对墨西哥人群的混合历史进行分析，发现其混合过程更接近于渐混模型。

毫无疑问，祖先染色体片段分布法的表现依赖于我们推测的祖先染色体片段的准确度。在本研究中，我们采用现存的方法对混合群体的祖先人群和祖先染色体片段进行推算。虽然我们用的 HAPMIX 极为准确，我们还是发现其中包含了过多的短的染色体片段，这些片段可能由于有来自其他祖先人群的祖先染色体片段，或者由于重组界限的分辨率有限导致的。虽然 HAPMIX 只适合做两个祖先人群的混合，但是我们可以通过合并祖先人群把多祖先人群混合转变成两个祖先人群混合来分析。比如，当我们分析墨西哥人的混合时，我们可以把欧洲祖先人群和非洲祖先人群合并当做一个祖先人群，然后把美洲印第安人当做另一个祖先人群，这样我们就把三个祖先人群的混合变成两个祖先人群的混合。该工作是第一个根据祖先染色体片段分布分析群体混合动力和混合过程的研究。随着局部祖先推测方法的提高，我们将能揭示更精细的群体混合动力和历史。

图示

Figure 1. Admixture models used to simulate the population admixture process.

Hybrid isolation (HI) model and continuous-gene-flow (CGF) model were adapted from Long[12], Graduate admixture (GA) model was adapted from Ewens and Spielman[13]. In each model, the genetic contributions of pop1 and pop2 are m_1 and m_2, respectively. The admixed population experienced G_i generation, which range from 1 to t generation.

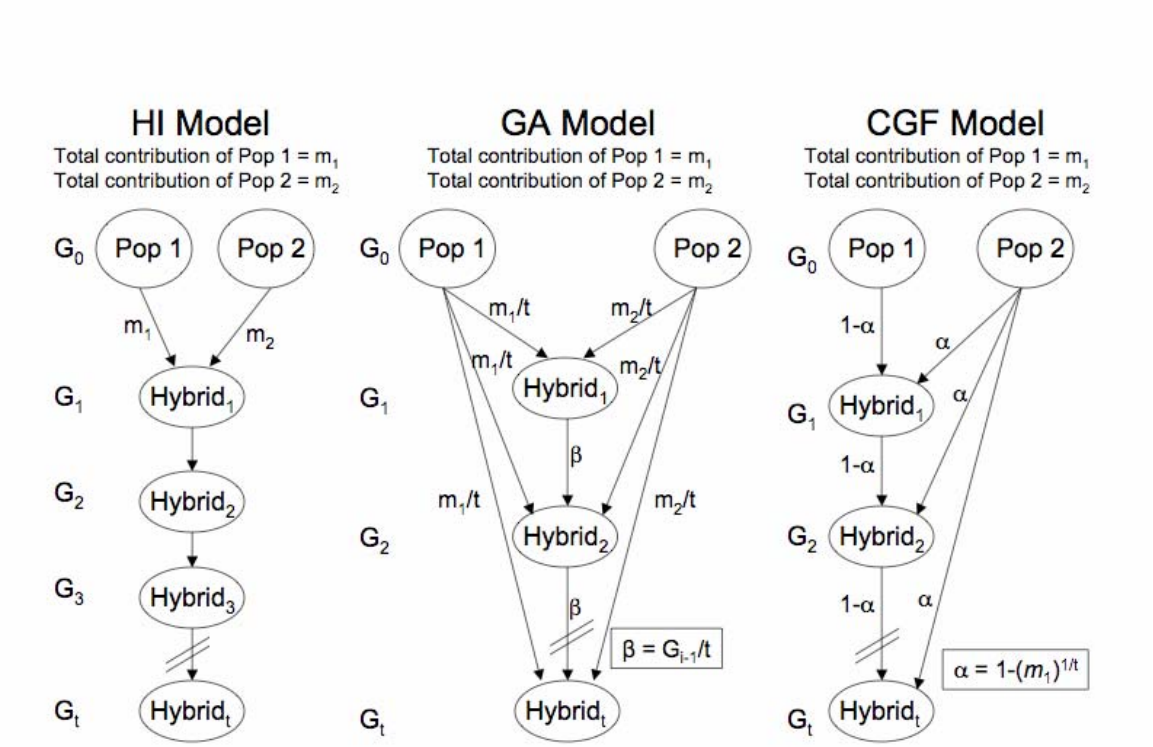


Figure 2. Distributions of chromosomal segments of distinct ancestry (CSDA) length when genetic contribution of the ancestral parental population to the admixed population is 50%.

G = "number of generations since admixture." (A) Number of generations since admixture was set to 10. (B) Number of generations since admixture was set to 20. (C) Number of generations since admixture was set to 50. (D) Number of generations since admixture was set to 100.

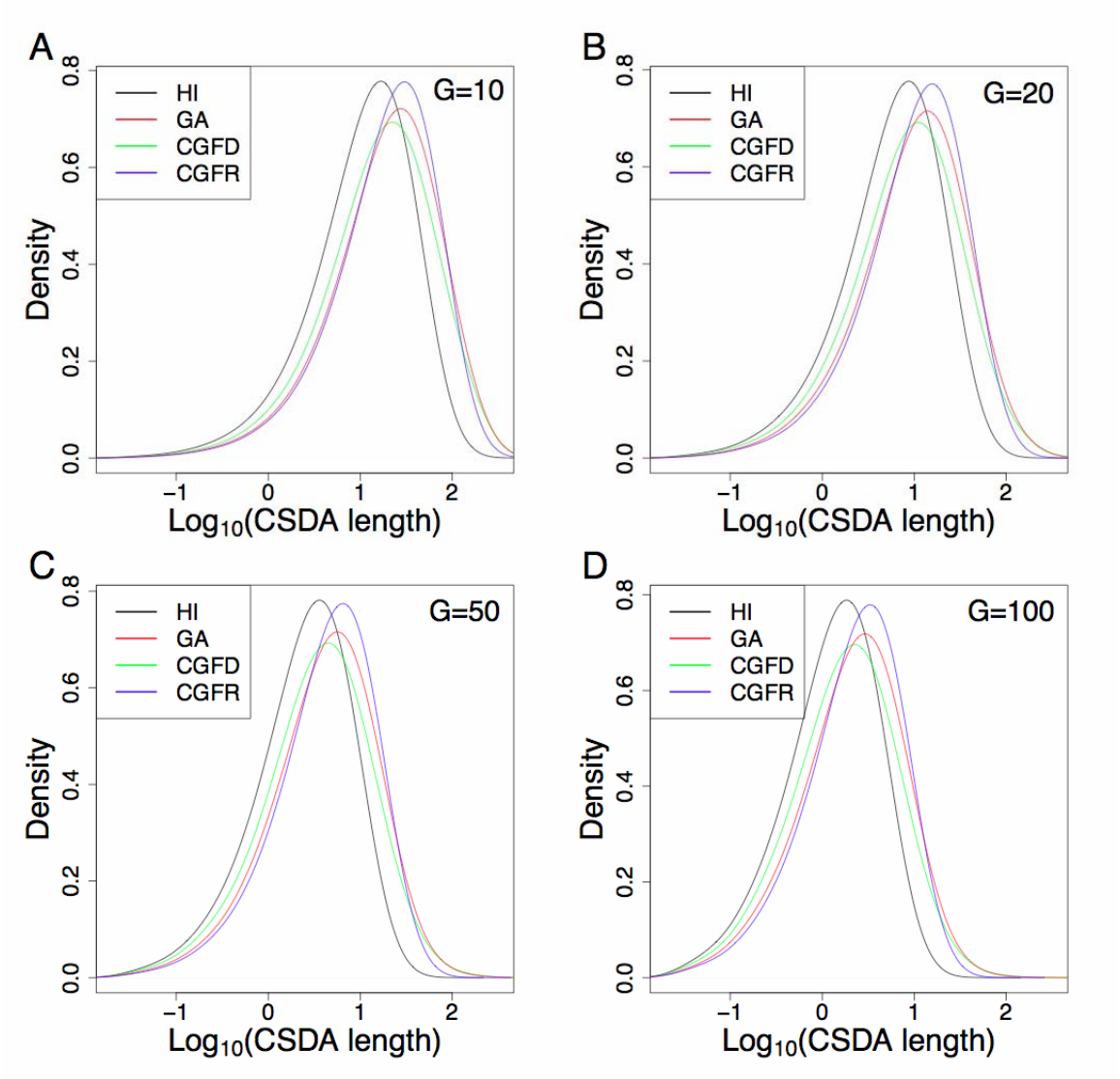


Figure 3. Distributions of individual ancestral proportion when genetic contribution of the ancestral parental population to the admixed population is 50%.

G = "number of generations since admixture." (A) Number of generations since admixture was set to 10. (B) Number of generations since admixture was set to 20. (C) Number of generations since admixture was set to 50. (D) Number of generations since admixture was set to 100.

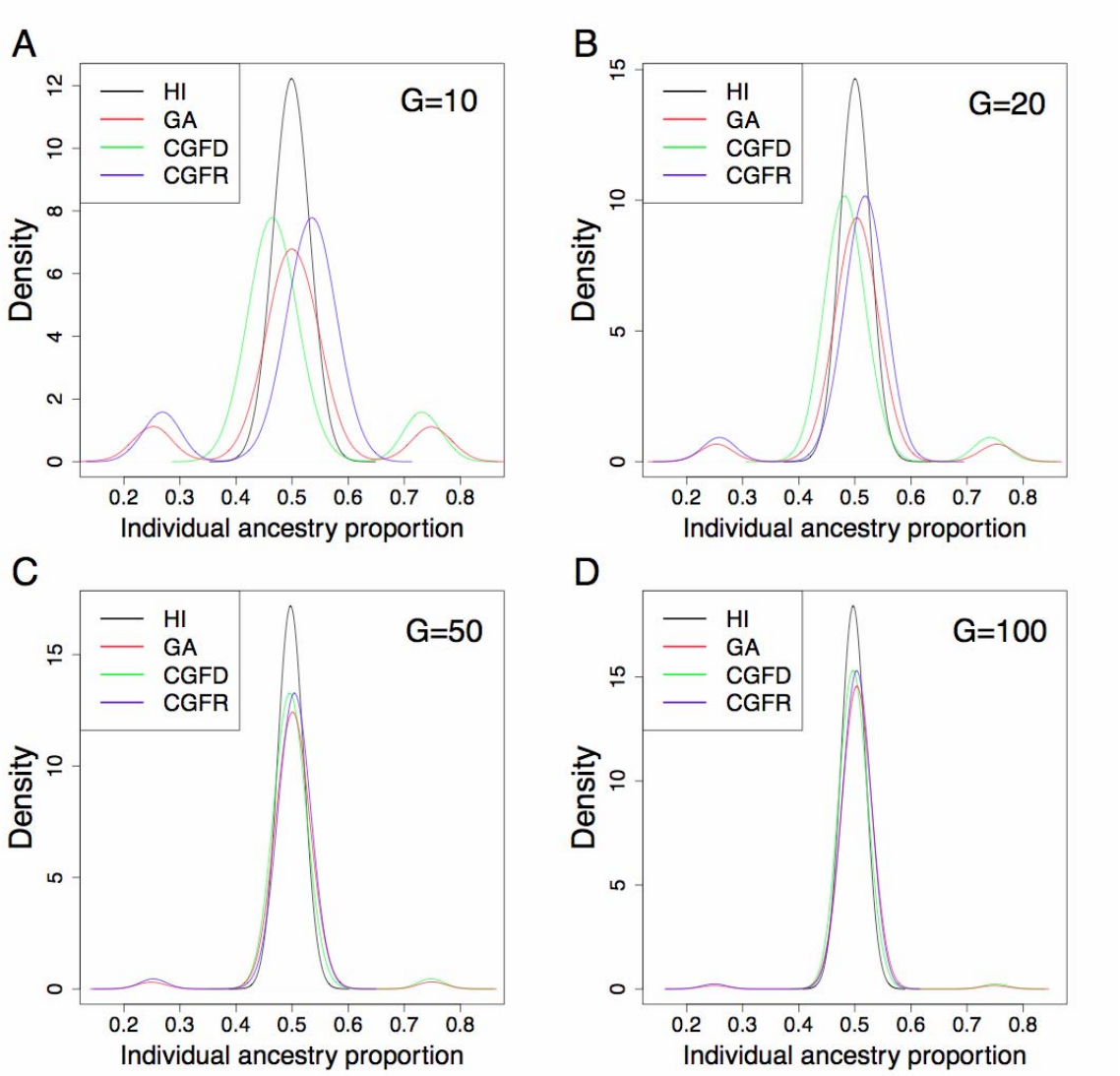


Figure 4. Principal component analysis of African-American samples and samples from the putative parental populations.

"AfA" represents African-Americans. The number in parentheses is the percentage of the total variance in the top ten PCs. It is shown that the 1,890 filtered African-Americans dispersed between YRI and CEU.

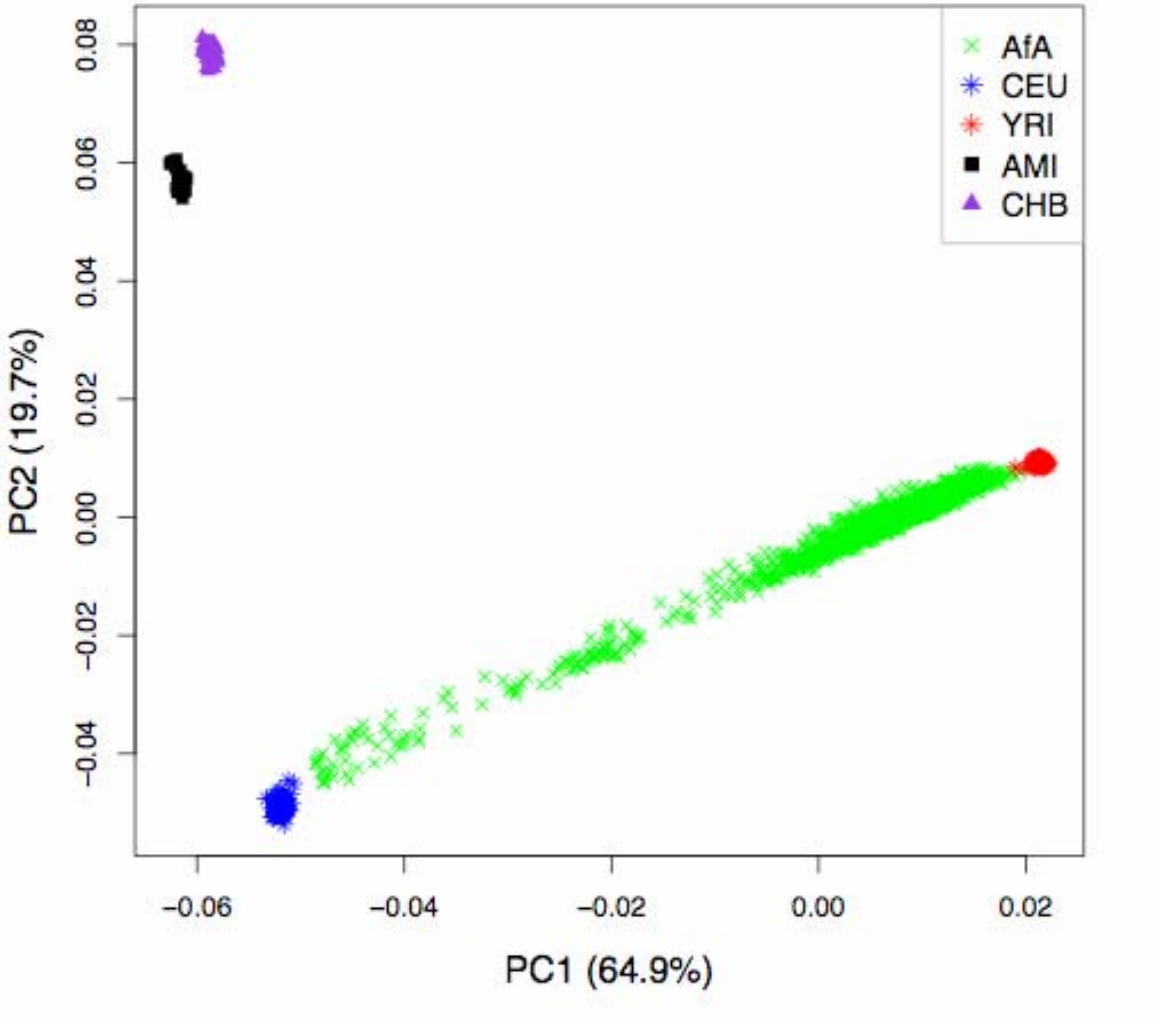


Figure 5. Admixture dynamics of European and African ancestry in African-Americans.
For the CGF model, the case in which Europeans continually served as genetic donor was considered as CGF1 model, while Africans as genetic donor was considered as CGF2 model. To find the model fit the empirical distribution best, earth mover's distance (EMD) between empirical data and that of each model was calculated. The model showing the lowest EMD with the empirical data was considered fit best. (A) Distribution of EMDs for African ancestral components between empirical data and each model. (B) Empirical distribution of CSDA length for African ancestral components and simulated distributions when the number of generations was set to 14. (C) Distribution of EMDs for European ancestral components between empirical data and each model. (D) Empirical distribution of CSDA length for European ancestral components and the simulated distributions when the number of generations was set to 14.

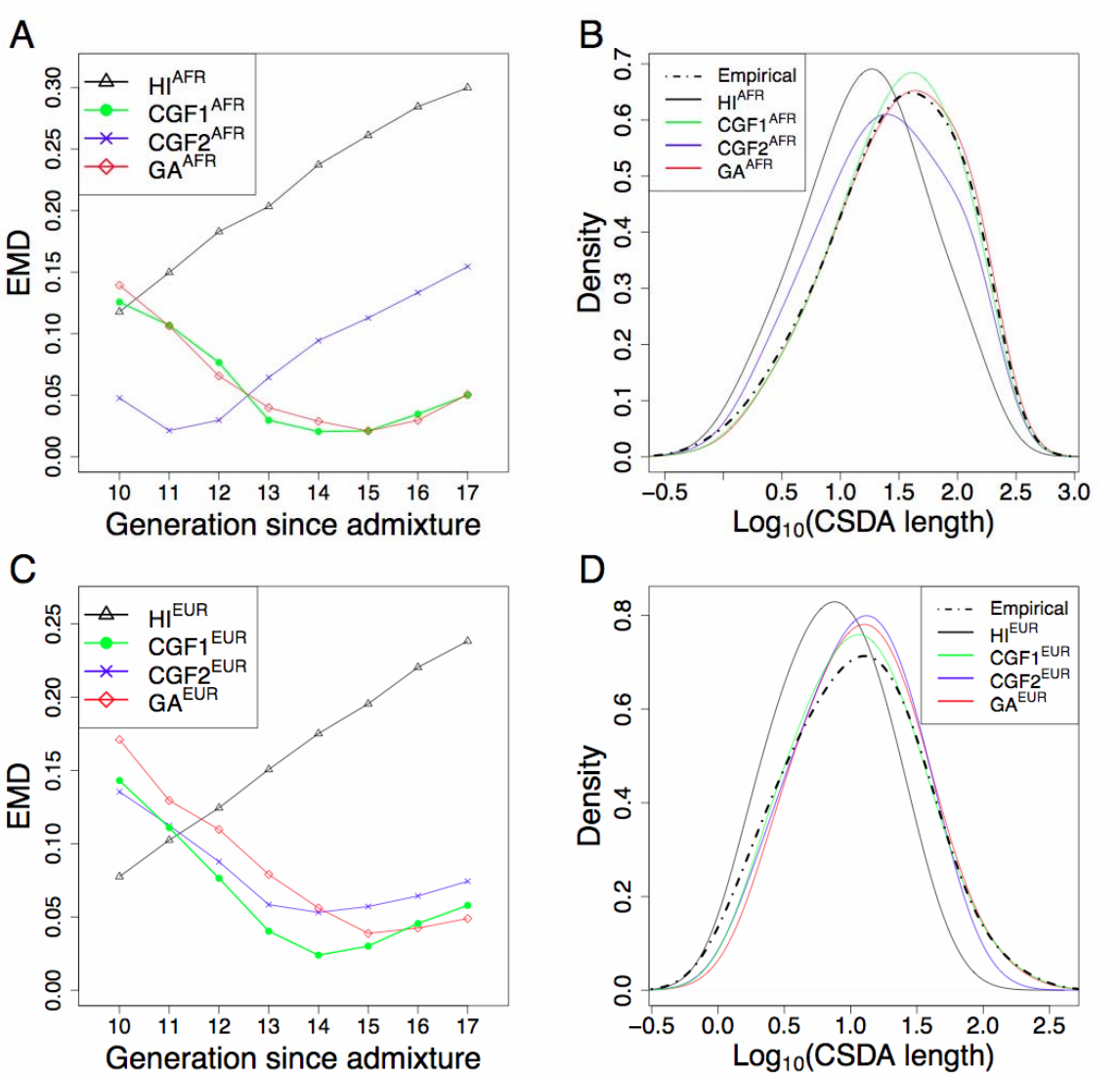


Figure 6. Empirical distribution of individual ancestry proportion for African ancestry components and simulated distribution when the number of generations is set at 14.

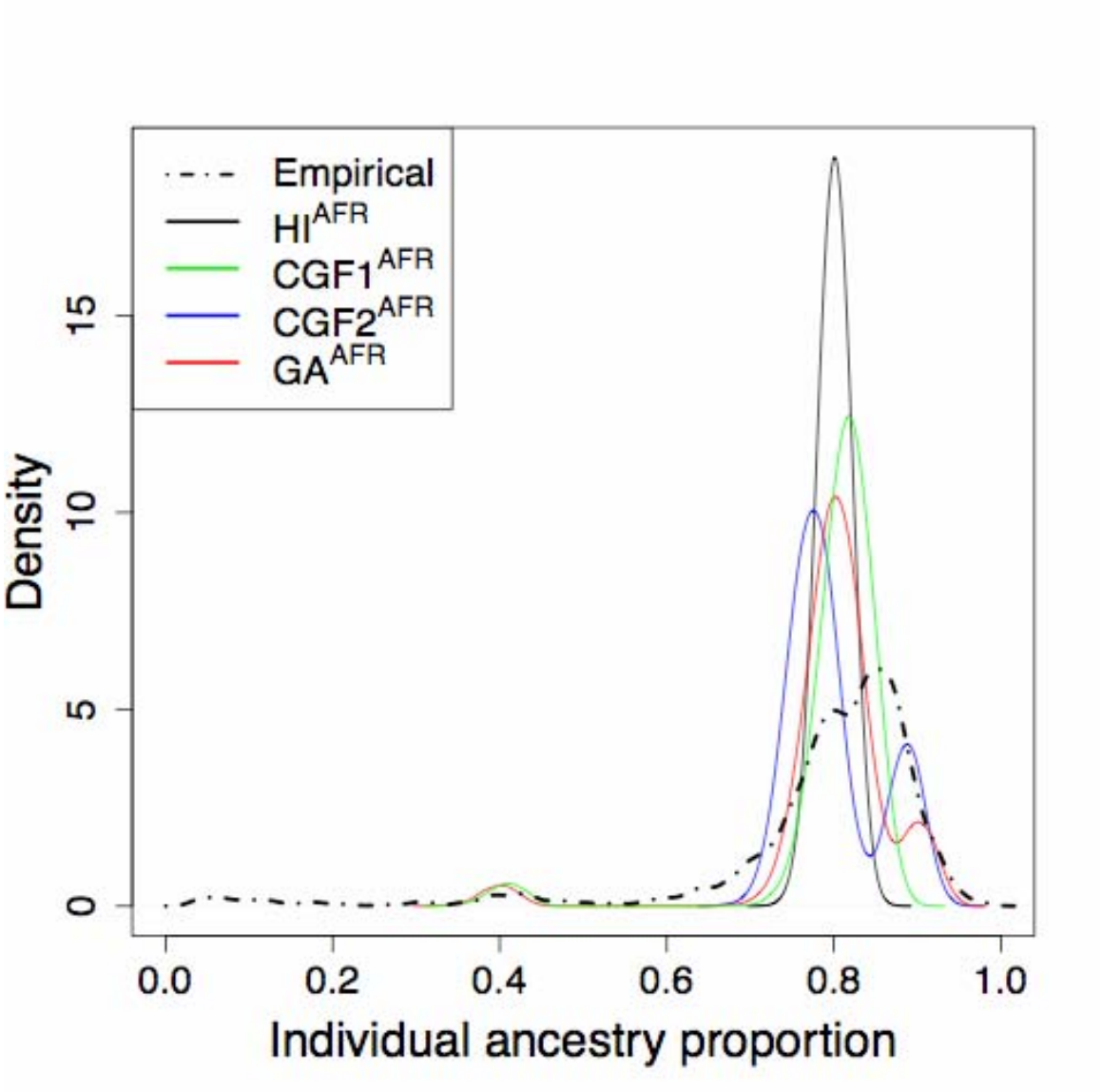


Figure 7. Relationship between ancestry proportion and estimated number of generations since admixture for each African-American individual.

Each black points correspond to one of the 1890 African-Americans studied here.

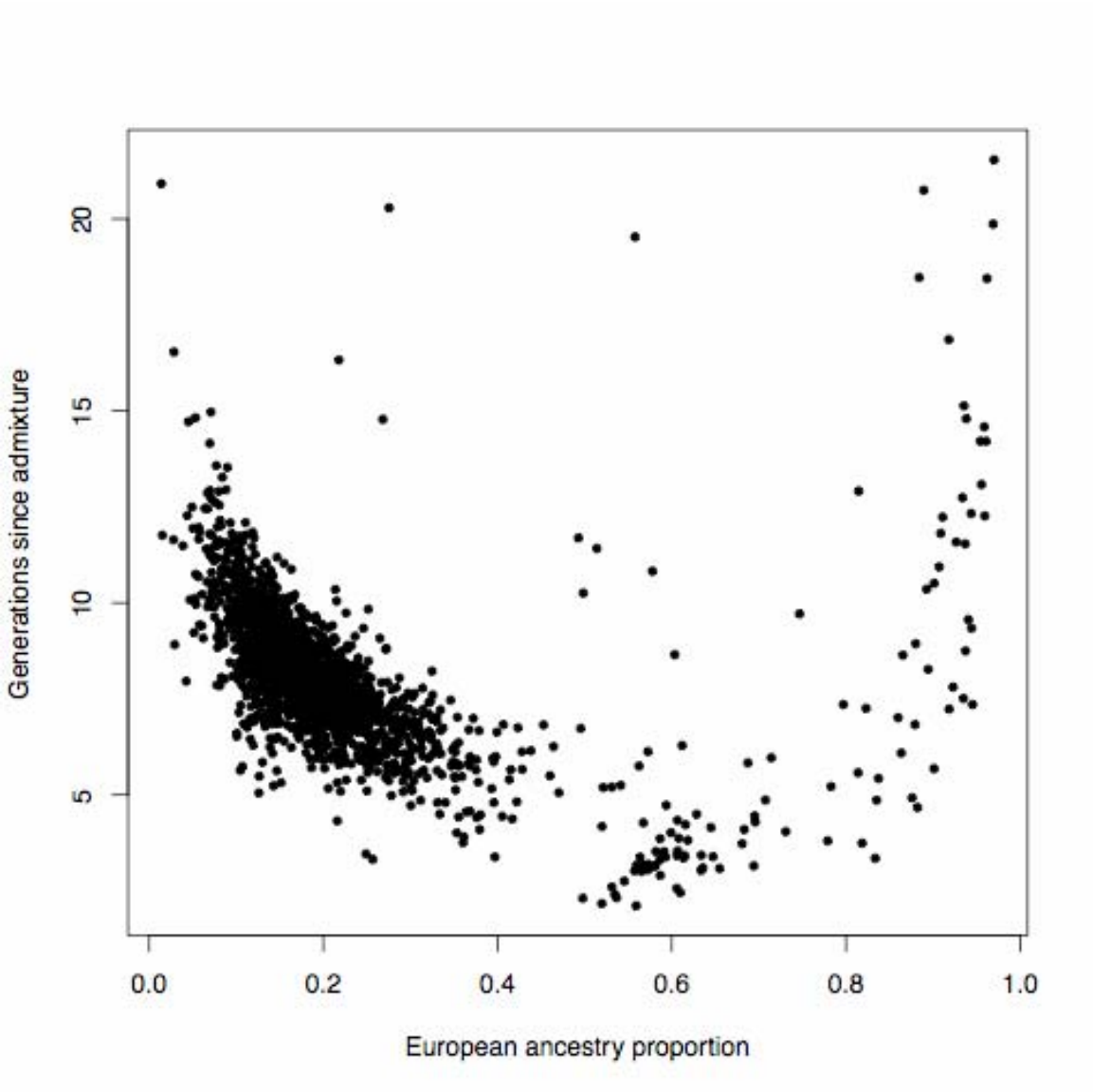


Figure 8. Principal component analysis of Mexican samples and samples from the putative parental populations.

The number in parentheses is the percentage of the total variance in the top ten PCs. It is shown that the filtered 413 Mexicans dispersed between CEU and AMI.

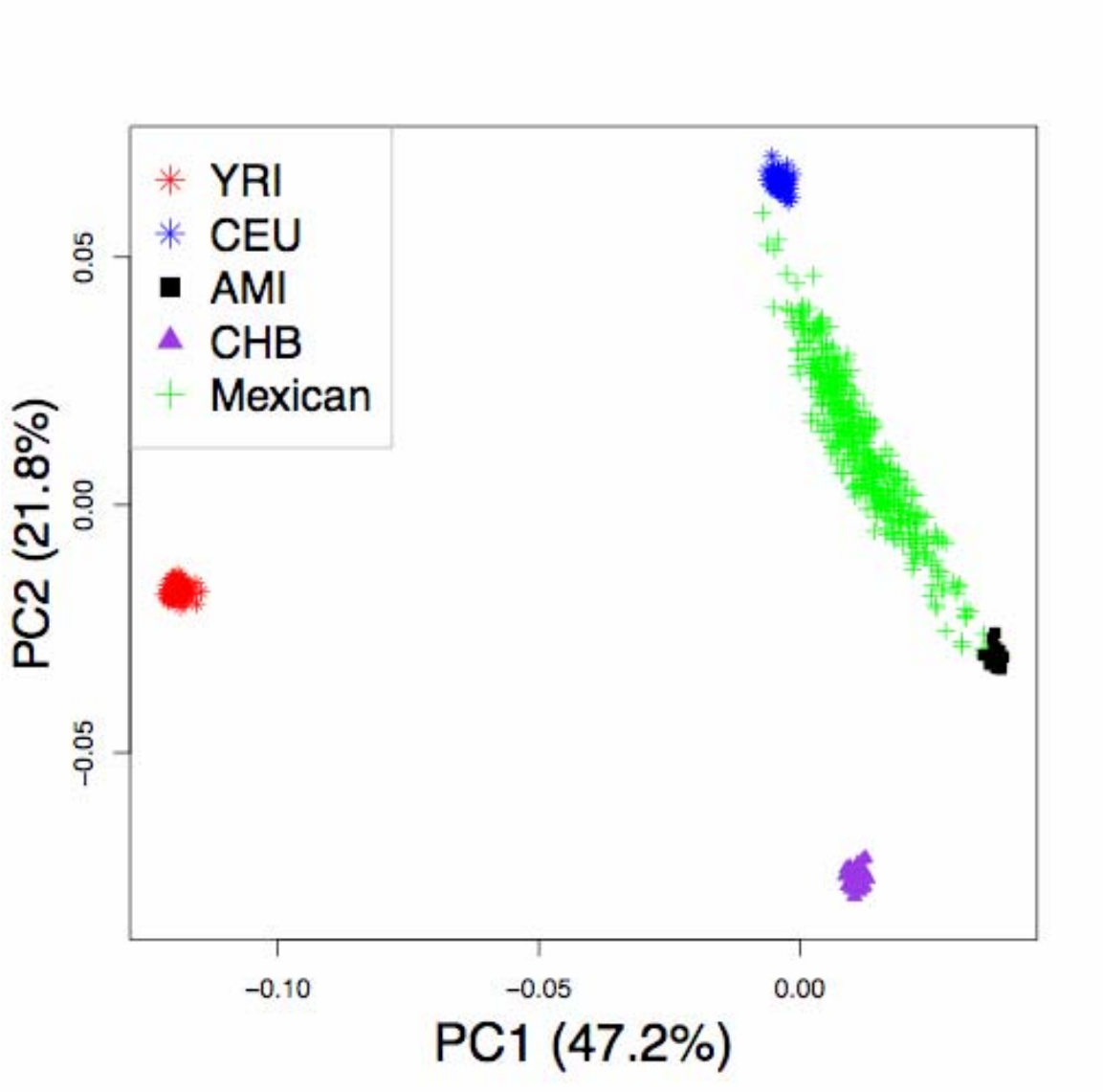
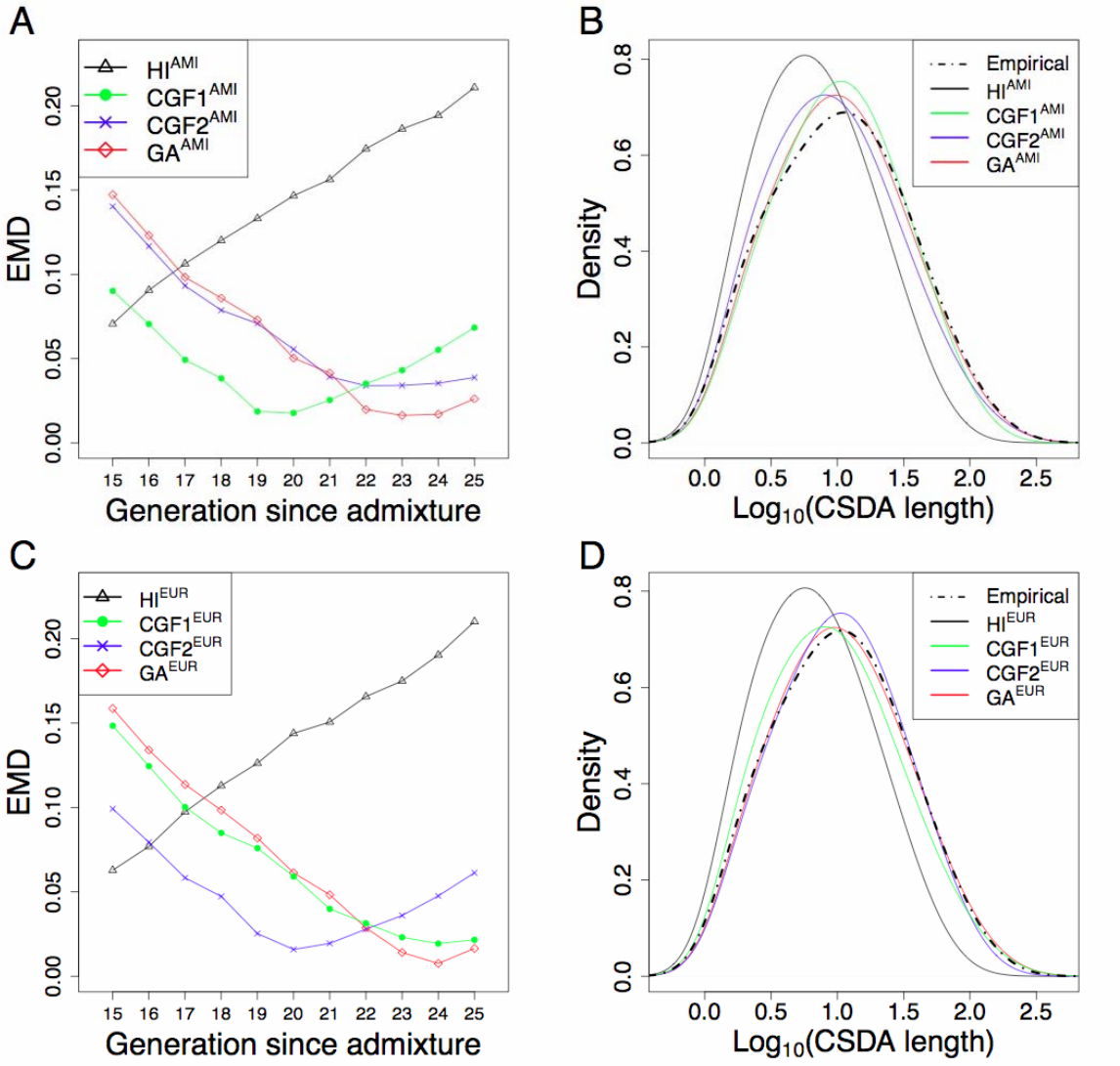


Figure 9. Admixture dynamics of European and Amerindian ancestry in Mexicans.

The model showing the lowest EMD with the empirical data was considered as fit best. The GA model, in which both European and Amerindian populations continuously contributed to the Mexican gene pool over about 24 generations, fit the empirical data best. (A) Distribution of EMDs for Amerindian ancestral components between empirical data and each model. (B) Empirical distribution of CSDA length for Amerindian ancestral components and the simulated distributions when the number of generations was set to 24. (C) Distribution of EMDs for European ancestral components between empirical data and each model. (D) Empirical distribution of CSDA length for European ancestral components and the simulated distributions when the number of generations was set to 24.



参考文献

1. Seldin MF, Pasaniuc B, Price AL (2011) New approaches to disease mapping in admixed populations. Nat Rev Genet 12: 523-528.

2. Verdu P, Rosenberg NA (2011) A general mechanistic model for admixture histories of hybrid populations. Genetics 189: 1413-1426.

3. Wang S, Ray N, Rojas W, Parra MV, Bedoya G, et al. (2008) Geographic patterns of genome admixture in Latin American Mestizos. PLoS Genet 4: e1000037.

4. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, et al. (2009) The genetic structure and history of Africans and African Americans. Science 324: 1035-1044.

5. HUGO Pan-Asian SNP Consortium, Abdulla MA, Ahmed I, Assawamakin A, Bhak J, et al. (2009) Mapping human genetic diversity in Asia. Science 326: 1541-1545.

6. Chakraborty R, Weiss KM (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci U S A 85: 9119-9123.

7. McKeigue PM (1997) Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. Am J Hum Genet 60: 188-196.

8. McKeigue PM (1998) 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 63: 241-251.

9. Montana G, Pritchard JK (2004) Statistical tests for admixture mapping with case-control and cases-only data. Am J Hum Genet 75: 771-789.

10. Stephens JC, Briscoe D, O'Brien SJ (1994) Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. Am J Hum Genet 55: 809-824.

11. Pfaff CL, Parra EJ, Bonilla C, Hiester K, McKeigue PM, et al. (2001) Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. Am J Hum Genet 68: 198-207.

12. Long JC (1991) The genetic structure of admixed populations. Genetics 127: 417-428.

13. Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 57: 455-464.

14. Parra EJ, Kittles RA, Argyropoulos G, Pfaff CL, Hiester K, et al. (2001) Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. American Journal of Physical Anthropology 114: 18-29.

15. Adams J, Ward RH (1973) Admixture studies and the detection of selection. Science 180: 1137-1143.

16. Pool JE, Nielsen R (2009) Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics 181: 711-719.

17. Guo W, Fung WK (2006) The admixture linkage disequilibrium and genetic linkage inference on the gradual admixture population. Yi Chuan Xue Bao 33: 12-18.

18. Xu S, Jin L (2011) Chromosome-wide haplotype sharing: a measure integrating recombination information to reconstruct the phylogeny of human populations. Ann Hum Genet 75: 694-706.

19. Seldin MF, Morii T, Collins-Schramm HE, Chima B, Kittles R, et al. (2004) Putative ancestral origins of chromosomal segments in individual african americans: implications for admixture mapping. Genome Res 14: 1076-1084.

20. Xu S, Huang W, Qian J, Jin L (2008) Analysis of genomic admixture in Uyghur and its implication in mapping strategy. Am J Hum Genet 82: 883-894.

21. Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, et al. (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet 5: e1000519.

22. Tang H, Coram M, Wang P, Zhu X, Risch N (2006) Reconstructing genetic ancestry blocks in admixed individuals. American Journal of Human Genetics 79: 1-12.

23. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567-1587.

24. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.

25. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. (2010) Integrating common

and rare genetic variation in diverse human populations. Nature 467: 52-58.

26. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100-1104.

27. Silva-Zolezzi I, Hidalgo-Miranda A, Estrada-Gil J, Fernandez-Lopez JC, Uribe-Figueroa L, et al. (2009) Analysis of genomic diversity in Mexican Mestizo populations to develop genomic medicine in Mexico. Proc Natl Acad Sci U S A 106: 8611-8616.

28. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2: e190.

29. Sankararaman S, Kimmel G, Halperin E, Jordan MI (2008) On the inference of ancestries in admixed populations. Genome Res 18: 668-675.

30. Sundquist A, Fratkin E, Do CB, Batzoglou S (2008) Effect of genetic divergence in identifying ancestral origin using HAPAA. Genome Res 18: 676-682.

31. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38: 1358-1370.

32. Sankararaman S, Sridhar S, Kimmel G, Halperin E (2008) Estimating local ancestry in admixed populations. Am J Hum Genet 82: 290-303.

33. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629-644.

34. Lilliefo.Hw (1967) On Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. Journal of the American Statistical Association 62: 399-&.

35. Hitchcock FL (1941) The distribution of a product from several sources to numerous localities. J Math Phys 20: 224-230.

36. Rubner Y, Tomasi C, Guibas LJ (2000) The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision 40: 99-121.

37. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-

575.

38. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

39. Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: analytical and study design considerations. Genetic epidemiology 28: 289-301.

40. Wang J (2003) Maximum-likelihood estimation of admixture proportions from genetic data. Genetics 164: 747-765.

41. Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, et al. (2004) A high-density admixture map for disease gene discovery in african americans. The American Journal of Human Genetics 74: 1001-1013.

42. Tian C, Hinds DA, Shigeta R, Kittles R, Ballinger DG, et al. (2006) A genomewide single-nucleotide-polymorphism panel with high ancestry information for African American admixture mapping. Am J Hum Genet 79: 640-649.

43. Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, et al. (2004) A high-density admixture map for disease gene discovery in african americans. Am J Hum Genet 74: 1001-1013.

44. Tian C, Hinds DA, Shigeta R, Adler SG, Lee A, et al. (2007) A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping. Am J Hum Genet 80: 1014-1023.

45. Price AL, Patterson N, Yu F, Cox DR, Waliszewska A, et al. (2007) A genomewide admixture map for Latino populations. Am J Hum Genet 80: 1024-1036.

46. Pfaff CL, Kittles RA, Shriver MD (2002) Adjusting for population structure in admixed populations. Genet Epidemiol 22: 196-201.

47. Smith MW, O'Brien SJ (2005) Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nat Rev Genet 6: 623--632.

第三章：美国黑人受到自然选择

摘要

鉴于美国黑人的非洲祖先在历史上经历过极高的死亡率，选择美国黑人群体进行自然选择研究具有重要意义。本研究收集了 1890 个美国黑人和它们可能祖先人群的 3320 个个体的常染色体上的 491,526 个单核苷酸多态性位点。我们利用传统的祖先贡献偏离法发现美国黑人基因组上有几个区域表现出特别高的非洲成分或欧洲成分，这些区域可能是群体混合后自然选择在美国黑人基因组上留下的痕迹。随后我们发展了一种在混合人群中检测自然选择的新方法：我们利用美国黑人基因组里的非洲祖先成分人工构造了一个非洲祖先群体 (ancestral African population; AAF)，比较这个非洲祖先群体与非洲土著人群在基因组上的差异可以检测到美国黑人离开非洲后受到的总的自然选择，包括混合前和混合后。我们发现用该方法检测到的很多受自然选择的基因与美国黑人的高发疾病如前列腺癌和高血压相关。我们因此推测这些疾病基因可能在美国黑人适应北美的新环境中起着重要作用。非洲祖先群体与非洲土著差异很大的区域也包含了多个抵抗疟疾的基因。通过对 *HBB* 和 *CD36* 的分析，我们发现抵抗疟疾的等位基因在非洲祖先群体中的频率比非洲土著低很多，这证实了来自疟疾的选择压力在新大陆得到释放。我们用这两种不同的方法筛选出来的受选择的基因基本完全不一样，这可能也意味着美国黑人在混合前后面临的环境不一样。最后，我们认为本研究提出的新方法有很好的统计效果，并且可以应用到其他混合群体，比如拉美人群和维吾尔族人群。

前言

尽管大部分人类遗传多态性是中性进化的，一些遗传多态性在人群适应当地环境过程中还是被自然选择重塑 [1, 2]。近期自然选择研究已经发现由于病原体、饮食、气候和其他环境不同导致了一些基因有明显的群体差异。这些研究发现极大地增加了我们对人类起源及进化的认识，有助于鉴定出具有重要生物功能的基因并进一步阐明一些人类疾病的遗传学基础 [1, 3-6]。近来大量高密度单核苷酸多态性数据为我们利用典型的一般人群在全基因组范围内筛选自然选择的印迹提供了必要的条件 [3, 4, 7, 8]。虽然已经有几个研究以近期混合人群为研究对象 [9-11]，但是还没有研究比较混合人群中的祖先成分与其相应祖先人群的差异，这种差异可能反映了两者分离后受到的自然选择。美国黑人作为经典的混合人群，为我们进行这样的分析提供了独一无二的条件。

美国黑人是指具有近期撒哈拉以南非洲祖先成分的美国居民。大多数美国黑人的非洲祖先成分是从奴隶贸易时期被运到北美的大概 50-65 万非洲奴隶那里遗传获得的 [12, 13]。由于早起奴隶极其糟糕的生活条件并会感染他们从未接触过的病原体，很多非洲奴隶死于大西洋航行途中或刚到达美洲的一段时间。虽然在整个奴隶贸易时期丧失的生命已经不可确切统计，但这个数量可能和当时的奴隶数量相当，或者更多 [14]。这样奴隶贸易时期美国黑人的高死亡率可以归结于巨大的环境压力如未接触过的病原体。这种持续的选择压可能使得优势等位基因的频率持续上升，最终导致美国黑人的非洲祖先成分与非洲土著人群在这些位点上的差异变

大。

同时，在非洲移民和欧洲移民到达美洲之前，他们各自的祖先已经在完全不同的环境中分别进化了几万年 [10, 11]。美洲全新的环境对这些人群及他们的混合人群来说都是一个挑战。 发生群体混合后，携带优势等位基因的祖先人群在这个位点的遗传贡献会逐渐上升，这可以作为自然选择发生的一种信号 [10, 11]。比如，曾经有研究利用美国黑人基因组上一些位点祖先贡献差异来检测自然选择 [15-17]。最近，Bryc 等 [9] 发现分析了 365 个美国黑人样本基因组上接近 50 万个单核苷酸多态性位点。他们发现有三个区域显示出特别高或或特别低的非洲成分（大于 3 倍的标准差），他们认为这可能是自然选择留下的痕迹。

在这项研究中，我们分析了 1890 个美国黑人和 3320 个可能的祖先个体的近 50 万个单核苷酸多态性位点。我们根据美国黑人经历的不同环境和历史时期把他们受到的自然选择分为混合前和混合后 (Figure 1)。首先，我们根据美国黑人祖先贡献的全基因组分布来检测美国黑人祖先在群体混合后可能受到的自然选择。然后，我们发展了一种新的方法：利用美国黑人基因组里的非洲祖先成分重构一个非洲祖先群体 (ancestral African population; AAF)，通过比较这个重构的非洲祖先群体与非洲土著人群在基因组上的差异来检测美国黑人离开非洲后受到的总的自然选择，包括混合前和混合后 (Figure 2)。我们发现用新的方法检测到的基因能够很好的解释美国黑人经历的历史。

材料和方法

数据组装与质量控制

基因分型数据分别从国际人类单体型计划（International Haplotype Map Project; HapMap, http://www.hapmap.org），人类基因组多态性计划（ Human Genome Diversity Project; HGDP, http://www.cephb.fr/en/hgdp）和 Illumina 病例对照数据库（iControlDB; http://www.illumina.com）获得。我们最终获得 588 个 HapMap 样本（包括 87 个 ASW, 167 个 YRI, 165 个 CEU, 85 个 CHD, 84 个 CHB）[18], 300 个 HGDP 样本（156 欧洲人个体，102 个非洲人个体和 42 个美洲印第安人个体）[19]。5455 个 iControlDB 病例对照样本（2161 个美国黑人个体和 3294 个高加索人个体）。这些病例对照样本均采用的 illumina 550K 微珠芯片技术进行基因分型并且通过了很严格的质控。

来至 HapMap 的美国黑人样本（ASW）和来至 iControlDB 的美国黑人样本之间并没有显著差异，我们把它们当做一个美国黑人群体来考虑。我们将去除不符合以下条件的样本：（1）已知有血缘关系的个体和 IBD>0.2 的个体（利用 PLINK [20]）；（2）对每个人群进行主成分分析时，主成分 SD>6 的异常个体（只考虑前两个主成分）；（3）对于每一个美国黑人样本，有 2%的祖先成分不是来至欧洲和非洲（可能来至于美洲印第安人或亚洲人）；（4）欧洲祖先成分的贡献>99%的个体，可能是欧洲移民的后裔；（5）非洲祖先成分的贡献>99%的个体，可能是近期非洲移民的后裔 [21] 。

我们把通过质控的样本放在一起，去除无法区分正负链的核苷酸多态性位

点；同时去除同一个人群内不同来源个体之间有很大差异的位点，以此来消除或减弱批次效应的影响。我们进一步去除位点数据缺失大于 10%的个体和数据缺失大于10%的位点。去除每个祖先人群中不符合哈迪-温伯格平衡的位点（$P<2\times10^{-6}$）。最终的数据包括来至 21 个人群的 5210 个个体，这些样本每个个体包括 503694 个单核苷酸多态性位点，其中常染色体位点 491,526 个。

群体和样本信息

我们共收集了 1890 个美国黑人样本，这些样本的基因组基本完全由欧洲成分和非洲成分构成。其中 1838 个样本来自 iControlDB 病例对照数据库，52 个来自 HapMap [18]。此外，取自 HapMap 的 113 个 YRI 及取自 HGDP 的 7 个非洲人群的 102 个样本 [19] 代表了美国黑人的可能非洲祖先人群。同时，取自 HapMap 的 113 个 CEU 及 取自 HGDP 的 8 个欧洲人群的 156 个样本 [19] 代表了美国黑人的可能的欧洲祖先人群。我们另外收集了 2648 个高加索人种病例对照样本代表更广泛的欧洲人群。此外，84 个 CHB 和 85 个 CHD 代表了东亚人群，24 个很纯的印第安人样本代表了他们的可能的美洲祖先人群。

群体遗传结构分析

主成分分析（principal component analysis; PCA）是将很多变量通过线性变换转化为少数几个重要变量的一种多元统计分析方法。为了减弱位点之间连锁不平衡对主成分分析的影响，我们在基因组上移动一个包含50个单核苷酸多态性位点的窗口，每次把窗口内r^2 >0.5的位点被去除并平移5个位点。这样我们把全基因组

的数据降到341,672个不是紧密连锁的位点，我们利用这些位点对所有样本在个体水平做主成分分析 [21]。聚类的方法能够直接把两个祖先群体分成两个主要成分，同时混合人群会表现出同时包含两种成分的特征。因为基于期望最大化算法（expectation-maximization; EM）的FRAPPE [22]能够利用所有的位点。我们利用它估算每个样本的祖先成分及混合比例。另外我们也用了一个基于贝叶斯（Bayesian）和马尔科夫蒙特卡洛方法（MCMC）来进行个体聚类的方法，这个方法已经在软件STRUCTURE中实现 [23,24]。在FRAPPE分析中，我们对491526个位点迭代分析1万次。对于STRUCTURE分析，因为连锁不平衡的位点可能被认为人群混合所致，我们把任意两个位点之间距离小于1百万的位点去除。

群体差异分析

当一个种群由于非随机交配或地理隔离等产生两个或多个亚群体，各个亚群体的等位基因频率将会表现出不同。假设一个大的种群由n个亚群体构成，每个亚群的等位基因频率为p_i。该种群的等位基因频率为各个亚群的平均值$\bar{p}$。这个等位基因反映出来的群体差异为：

$$Var(p)=\frac{1}{n}\sum_{i=1}^{n}(p_i-\bar{p})^2$$

$$F_{ST}=\frac{Var(p)}{\bar{p}(1-\bar{p})}$$

这里，F_{ST} 称为群体分化系数，是度量群体差异的最重要的统计量。群体分化系数越大，说明群体之间的差异越大。相反，说明群体差异越小。但是样本量的波动会

对F_{ST}的值产生重要的影响，为了使得不同样本量的群体之间的群体差异系数有可比性。Weir和Cockerham [25] 通过考虑样本量对F_{ST}进行了改进，称之为无偏群体分化系数（unbiased F_{ST}）。本研究计算的群体差异和单点F_{ST}均为无偏F_{ST}。

推断位点的祖先来源

自从高密度单核苷酸多态性数据大量出现以后，很多基于这种都密度数据推断祖先来源的方法发展起来。比如ANCESTRYMAP [26], SABER [27], LAMP与LAMP-ANC [28], HAPAA [29] 和 HAPMIX [30]。我们首先依据获得的非洲和欧洲祖先人群基因分型数据模拟出混合人群的基因分型数据，然后用这一套模拟数据来评估各个方法的准确度。我们发现HAPMIX比其他几种方法有更高的准确度。这样，以后的分析都使用HAPMIX对每个位点的祖先来源进行推断。同时，因为LAMP-ANC的表现也比较好（仅次于HAPMIX），我们也使用LAMP-ANC对每个位点的祖先来源进行推测。

我们直接从HapMap网站下载HapMapIII期定相基因分型数据（phased data）[18]。我们使用88个CEU个体和88个YRI个体（这些个体都来自小孩和父母基因型已知的三人家系数据中的的无关父母个体）分别代表美国黑人的非洲祖先群体和欧洲祖先群体。我们采用FRAPPE计算出来的欧洲祖先人群对美国黑人基因组的平均贡献作为HAPMIX需要的混合比例。我们设定不同的群体混合代数，当取得最大似然值时的混合代数将被认为美国黑人的混合代数。在双倍体模型下运行HAPMIX, 我们能直接获得每个个体的单体型和每个染色体片段的祖先来源信息。我们利用HAPMIX推出来的美国黑人非洲成分和欧洲成分分别构建了美国黑人的非洲祖先人群(ancestral African population; AAF)和欧洲祖先人群（ancestral European population;

AEU)。简单的说，对于每一个核苷酸多态性位点，美国黑人的非洲祖先片段上或欧洲祖先片段上的等位基因频率即为非洲祖先人群或欧洲祖先人群的等位基因频率。

模拟美国黑人及其祖先群体

根据中性进化学说，混合人群及其祖先群体的遗传漂变会导致混合比例和群体差异在不同位点上有明显差异 [25,31]。这样，我们模拟了一系列数据来探索遗传漂变可能对美国黑人非洲祖先人群和现在非洲人群之间群体差异的影响，以及美国黑人欧洲祖先人群和现代欧洲人群之间群体差异的影响。在这个模拟过程中，我们根据 HapMap 第 22 次发布的数据提供的遗传图谱来引入重组。因为我们只模拟了很短的历史，所以我们不考虑这个过程中的突变。各个群体的有效群体（effective population sizes; N_e）大小参照了 HapMap 提供的数据。具体来说，设定非洲人，欧洲人和美国黑人的有效群体大小分别为 17,094, 11,418 和 17,094。根据我们已有的分析，设定美国黑人的形成过程为连续基因流模型（Continuous-gene-flow model; CGF model）。

我们利用前面提到的 88 个 YRI 个体和 88 个 CEU 个体的定相基因分型数据来分别代表混合之前的非洲人群和欧洲人群。根据我们上一章对美国黑人混合动力学研究的结果，我们设定群体混合代数为 14。美国黑人每一代接受的基因流可以根据公式$\alpha = 1-(m_1)^{1/t}$来计算，m_1是美国黑人平均携带的欧洲成分 21.65%，这样每代欧洲人贡献大约 1%的遗传成分到美国黑人中。同时，欧洲人群和非洲人群分别进化 14 个世代。最后，为了使模拟数据和和我们分析数据的样本量相匹配，我们

模拟输出 113 个欧洲个体，113 个非洲个体和 1890 个美国黑人个体，并输出每个美国黑人的每个位点的祖先状态。因为我们知道美国黑人的非洲祖先从非洲到美洲相当于经历了一个瓶颈事件，我们把瓶颈事件设置在美国黑人的第一个世代。其强度分别设为美国黑人的非洲祖先人群有效群体大小减少到 8,000, 5,000, 3,000, 2,000 和 1,000。

功能注释和智能通路分析（ingenuity pathway analysis; IPA）

我们根据 HapMap 网站 [32] 提供的注释信息对祖先贡献和群体差异的异常区域进行功能注释。我们同时利用智能通路分析对这些群体差异较大区域进行基因功能和网络整合分析。智能通路分析 IPA 将搜索高度结构化的智能知识库（Ingenuity Knowledge Base），这个数据库包括人工阅读提取的两百六十多万条公开发表的科研成果和报告构建出来的，包括分子相互作用，基因调控，基因环境互作，代谢物组学和蛋白质组学的实验数据等。智能通路分析的应用范围包括靶标的发现及验证、代谢组学研究、先导化合物的验证及作用机理研究、毒性及安全性评估、生物网络模拟及分析、生物标记物研究等等。IPA 强大的数据分析和搜索能力能够帮助发现导致群体差异的重要机制，在更高的分子网络和细胞成面理解这种差异。

统计分析

在本研究中，所有的统计计算和作图都使用 R 完成。我们使用柯尔莫诺夫-

斯米尔诺夫检验（Kolmogorov－Smirnov test)来检验全基因组单点 F_{ST}的经验分布和模拟分布之间是否有显著差异。用卡方检验（Chi-squared test;χ^2 test）来检验各类功能性单核苷酸多态性位点是否相对于非基因区的位点在群体差异大的单核甘酸多态性位点集合中有显著富集。

结果

美国黑人的欧洲和非洲祖先人群

毫无疑问，来自于西非和欧洲的人群一定贡献了美国黑人的基因库。在这项研究中，我们选取他们作为美国黑人的祖先人群。但是，西非和欧洲的人群在最近300 年受到遗传漂变(genetic drift)和各种人口统计学历史(demographic history)的影响，使得选择恰当的代表性人群变得复杂。西非和欧洲的多个人群对美国黑人有贡献使得选择恰当的人群变得更为复杂。一方面我们拥有很高质量的YRI 和 CEU 的定相基因分型数据（phased data)，在这项研究中我们选择 YRI 和CEU 分别代表美国黑人的西非祖先人群和欧洲祖先人群。另一方面，我们分析了人类基因组多样性计划 (Human Genome Diversity Project; HGDP) [19] 中的多个人群，我们发现 YRI 和 CEU 比其它基因型已知的人群有更好的代表性(Figure 3)。

当我们设定 K=2, 利用 FRAPPE [22] 对美国黑人及其可能祖先人群的全基因组单核苷酸多态性位点进行分析，我们发现在群体水平欧洲人群贡献了美国黑人基因组的 21.65% (Figure 4)。STRUCTURE [23,24] 分析的结果和 FRAPPE 的结果是

基本一直的。然后对连锁不平衡过滤后的 341，672 个单核苷酸多态性位点进行个体水平的主成分分析 [33]。根据每个样本投影到第一主成分上的坐标，我们推出欧洲人群的贡献是 21.61% (Figure 3D)。这些估计值与之前的报道基本一致，虽然我们使用的数据完全不同 [30,34]。

美国黑人基因组上祖先贡献异常区域

为了估计欧洲人群和非洲人群对美国黑人基因组的贡献，我们利用前面提到的88个YRI个体和88个CEU个体的定相基因分型数据分别代表美国黑人的欧洲人群和非洲人群。在CEU和YRI中都是同一种单态的单核苷酸多态性位点基本不能提供美国黑人的祖先信息，我们在随后的分析中去除这种位点。每个样本的混合代数基本分布在1到12之间，基本符合持续基因流模型（continuous-gene-flow model)的预期。我们利用这些已知和推测参数作为HAPMIX的信息参数来推算每个美国黑人的染色体片段祖先来源。我们得出全基因组范围内欧洲人群对美国黑人每个位点的贡献是21.68% ± 0.75% (平均值 ± 标准差)。我们发现本研究中欧洲人群对美国黑人的贡献在位点之间差异比之前新大陆混合人群研究报道的都要小 [9-11,15-17]，这可能是因为我们采用了更大的样本。

混合人群基因组上表现出特别高或低的祖先贡献区域可能是自然选择留下的痕迹 [9-11,35]。在本研究中，我们把美国黑人基因组上祖先贡献差异超过3倍标准差的区域定义为自然选择的候选区域。我们在美国黑人基因组中发现四个区域（2p22, 3q13, 6q26, 16q21）表现出特别高的欧洲成分，两个区域（1p36,

2q37）表现出特别高的非洲成分(Figure 5)。并且每一个区域的祖先贡献都与全基因组的贡献显著不同(P < 2.2×10^{-16}, t test)，我们对每个区域进行了详细的注释（Table 1）。很多候选基因区域都可以用LAMP-ANC重复出来，虽然它的准确度要比HAPMIX低很多。

美国黑人全基因组祖先贡献偏离最大的区域在1p36，非洲祖先成分在这个区域显著增高。 *IGSF21*和*AKR7A2*分别落在这个区域中祖先贡献偏离最大的单核苷酸多态性位点附。*IGSF21*属于免疫球蛋白超级家族，而*AKR7A2*参与了醛和酮的解毒过程，可能是多种癌症如胰腺癌的风险因素 [36,37]。另外一个表现出很高非洲成分的区域是2q37，这个区域的*PDCD1*参与了重要的免疫通路并是多种疾病的风险因素。美国黑人欧洲成分最高的区域在3q12，*LSAMP*是这个区域唯一的基因，它是骨肉瘤中的抑癌基因并被报道与冠状动脉疾病相关联 [38,39]。有趣的是，欧洲成分很高的2p22区域附近的的*EIF2AK2*参与了流感感染通路，这可能意味着欧洲祖先成分能够更好的抵抗流感。也意味着欧洲祖先人群和非洲祖先人群可能在抵抗流感能力方面表现出不同 [40,41]。

美国黑人的欧洲和非洲祖先成分

HAPMIX推出来的美国黑人非洲祖先染色体片段和欧洲祖先染色体片段即为其非洲祖先成分和欧洲祖先成分，我们把这种祖先成分当做群体来看待，分别称为非洲祖先人群（ancestral African population; AAF）和欧洲祖先人群（ancestral European population; AEU）。因为HAPMIX在这种在模拟数据中有

98%的准确度，我们相信我们重构的祖先人群是高度可信的。

我们比较了欧洲祖先人群与各个现代欧洲人群之间的群体差异。我们发现CEU与欧洲祖先人群的差异性最小($F_{ST\ [AEU-CEU]}$ = 0.0005)。高加索人种病例对照人群（CAU-GWAS）与欧洲祖先人群的群体差异次之（$F_{ST\ [AEU-CAU-GWAS]}$ = 0.0006)。 在各个非洲土著人群中，YRI与非洲祖先人群的群体差异最小($F_{ST\ [AAF-YRI]}$ = 0.0007)。当我们把这种观察到的人群差异与其相应的模拟数据的人群差异进行比较，我们发现非洲人群和欧洲人群表现出完全不同的模式。观察到的欧洲祖先人群与CEU的群体差异(观察的$F_{ST\ [AEU-CEU]}$ = 0.0005)比模拟的群体差异(模拟的 $F_{ST\ [AEU-CEU]}$ = 0.0006)略小，并且观察到全基因组F_{ST}分布和模拟的分布没有很大的差异（P = 0.042, 柯尔莫诺夫-斯米尔诺夫检验）。

但是，对于非洲人群，我们发现实际观察到的非洲祖先人群与YRI($F_{ST\ [AAF-YRI]}$ = 0.0007)的群体差异比模拟的值要大(模拟的$F_{ST\ [AAF-YRI]}$ = 0.0006)。并且观察到的全基因组F_{ST}分布和模拟的分布有很明显的显著不同 （$P < 2.2\times10^{-16}$, 柯尔莫诺夫-斯米尔诺夫检验）。从Q-Q图上看出，全基因组实际的F_{ST}分布中有富集了很多F_{ST}很大的位点(Figure 6)，这意味着美国黑人的非洲成分受到明显的自然选择。观察到的这种现象基本不受美国黑人非洲祖先人群所经历的瓶颈事件的影响。

美国黑人非洲祖先人群和非洲土著高群体差异区域

根据中性学说，人群历史对基因组上每个位点群体差异的影响是一致的，至

少是相似的 [13, 14, 42]。但是，正向自然选择会作用于基因组的某些区域，使得这些区域的群体差异相对于中性漂变增大 [2, 25]。在全基因组筛选群体差异大的区域已经被作为一种寻找正向自然选择的方法 [35, 43]。自从非洲移民离开非洲踏上去美洲的行程，他们将经历与非洲土著人群完全不同的群体历史。美国黑人在奴隶贸易时期的高死亡率意味着他们在当时受到很强的自然选择。由于低频等位基因（minor allele frequency; MAF）在祖先估计中受到更严重的取样误差和统计误差，我们把YRI或非洲祖先人群中MAF<0.05的多态性位点去除。

虽然非洲祖先人群和 YRI 之间总的群体差异很小，我们还是看到了一些区域的群体差异异常的高（Figure 7A）。这些群体差异特别大的区域即我们之前看到的比中性模拟数据群体差异大的区域。我们对全基因组群体差异特别大的区域 (99.99^{th} percentile; F_{ST}>0.0452)进行了功能注释（Table 2）。虽然之前的报道说单个位点的 F_{ST} 变异太大，作为自然选择的信号不是很可靠 [44]。但是这四个区域(7q21, 6p21-22, 1q22 和 11p15)各有多个位点的 F_{ST} 差异很大，说明这些信号应该可靠的。7p21 区域有 CD36 和 SEMA3C 两个基因。CD36 直接参与了疟原虫感染和寄生红细胞的附着过程。同时，CD36 能结合长链脂肪酸，意味着它参与了脂肪酸的转运或调控了脂肪酸的转运 [45-47]。CD36 已经被发现受到很强的正向自然选择，其选择压力可能来自于疟疾或其它未知因数 [48-51]。SEMA3C 能被 ADAMTS1 诱导表达，能促进癌症细胞的迁移 [52]。6p21-22 区域包含了人类主要组织相容性复合体（major histocompatibility complex; MHC），已经被各种研究报道受到自然选择 [10, 53]。并且该区域的非洲祖先成分较高，虽然没有达到 3 倍的标准差。MUC1 基因位于 1q22 区域，参与由 PDGF 介导的信号通路并通过与病原体结合

提供保护的功能[54,55]。HBB 和 HBD 位于 11p15 区域， 因为他们的基因突变使得个体能够抵抗治疟疾，一直受到很强的平衡选择 [56,57]。

智能通路分析能够在较高的细胞层次和分子机制上探索自然选择候选基因之间的功能和信号通路。我们对群体差异最高的 402 个单核苷酸多态性位点(99.90th percentile; F_{ST} >0.0287)进行智能通路分析，我们发现代谢疾病相关基因在其中有显著的富集(P = 1.51×10^{-16})。随后依次是内分泌疾病相关基因(P = 2.23 ×10^{-16})，免疫疾病相关基因(P = 9.30×10^{-12})和遗传疾病致病基因(P = 9.30×10^{-12})。在所有信号通路中，抗原展示通路(P = 1.95×10^{-4})，移植排斥反应信号通路(P = 4.69×10^{-3})，移植物抗宿主病信号通路(P = 4.69×10^{-3})和自身免疫性甲状腺疾病信号通路(P = 5.35×10^{-3})中的基因都有显著富集。上述四种信号通路都涉及到免疫系统，这可能是由于撒哈拉以南的非洲和北美之间巨大的环境和病原体等差异不同导致的。我们还对群体差异较高的 4,011 单核苷酸多态性位点 $_{T}$ (99.00th percentile; F_{ST} > 0.0162)进行了智能通路分析，其结果和 99.9 百分位的结果相似。但是 IL-9 信号通路（p =8.01×10-3）和表皮生长因子信号通路（p =6.38×10-3）相关基因有显著富集。

重构美国黑人基因组及其与美国黑人的差异

我们根据 YRI 和 CEU 的基因型重构了一个新的美国黑人群体（reconstituted African Americans; rAfA)，通过比较其与现代美国黑人基因组的差异来检测美国黑人受到的自然选择。其合理性在于我们假设 rAfA 是中性的不

受自然选择影响，而美国黑人由于经历环境巨变而受到正向自然选择。rAfA 每个位点上的等位基因频率是根据非洲和欧洲混合比例（21.68%)及他们的代表人群 CEU 和 YRI 的等位基因频率算出。用这种方法筛选自然选择候选基因可以避免局部祖先来源估算引入的误差。美国黑人非洲祖先人群与 YRI 之间差异最大的 42 个核苷酸多态性位点中的 34 个都出现在 rAfA 与美国黑人差异最大的区域（最高的 1%）。我们进一步比较发现非洲祖先人群与 YRI 之间差异最大的几个区域和 rAfA 与美国黑人差异最大的几个区域是一致的（Figure 7B），说明美国黑人受到的自然选择基本上都发生在其非洲祖先成分上。当我们使用 CAU-GWAS 替代 CEU 作构建 rAfA。发现结果与之前的观察基本一致。

随后，我们利用一组可能贡献了美国黑人非洲祖先成分的土族人群来代替单一非洲人群 YRI 来重构 rAfA。根据之前的报道 [12]，我们首先选取 64%的约鲁巴人，19%的 Mandenka 和 14%的班图人重构了一个可能贡献了美国黑人非洲祖先成分的纯非洲混合人群。我们用这个重构的纯非洲混合人群和 CAU-GWAS 重构了 rAfA。我们发现这个重构的 rAfA 与美国黑人全基因组的 F_{ST} 分布和之前的分布基本一致(Figure 8)。

美国黑人非洲祖先成分受到正选择的另一个证据

因为正向自然选择一般作用于特定群体的特定位点，结果会使得受选择位点与其它群体的群体差异相对增大即 F_{ST} 变大 [58]。假设正向自然选择倾向于作用在有重要功能的多态性位点，这样高群体差异性的单核苷酸多态性位点中会富集很多

功能性位点 [7]。我们首先设定在非洲祖先人群和 YRI 之间群体差异最高的 1%的多态性位点（F_{ST} > 0.0164)为高群体差异多态位点，然后评估各类多态性位点在其中的富集程度。

根据多态性位点在基因组上的位置和功能，我们把全基因组的多态性位点分为基因间区多态，基因区多态，内含子多态，3' UTR 多态, 5' UTR 多态, 同义多态，非同义多态，编码区多态，转录区多态，近基因 3'端多态和近基因 5'端多态。我们发现每一类多态性位点的 F_{ST} 分布本身没有显著差异。但是，基因区多态在高群体差异多态中的比例比基因间区多态显著增多(χ^2 test, P=0.046; Figure 9)。转录区多态的比例相对于基因间区更显著(χ^2 test, P=0.004; Figure 9)。因为基因间区多态的功能相对接近于中性，我们可以认为比基因间区显著增多的功能多态是正向自然选择导致的。但是，同义多态在高群体差异多态位点中的比例比基因间区多态的比例高出 1.22 倍，这可能是由于它们与受选择的功能多态位点相连锁导致的（即搭车效应）。当我们把群体差异最高的 5%的位点或 0.1%的位点设为高群体差异性位点时，所得结果基本不变。

讨论

美洲大陆新混合人群如美国黑人的两个祖先人群在混合之前已经隔离了几万年，在美洲新环境中又重新混合到一起，这为我们研究近期自然选择的提供了独特的条件。通过估算混合人群祖先贡献增多或减少来检测自然选择留下痕迹的方法已经在之前的研究中多次使用 [9-11]，但是该方法无法检测群体混合发生之前的自

然选择。因此，我们发展了一种能够检测到群体混合之前的自然选择的方法。通过比较非洲祖先人群和非洲土著之间的差异，我们够检测美国黑人的非洲祖先离开非洲之后受到的自然选择（包括混合前后两个时期）。我们检测到的自然选择的信号大部分能够被能够通过比较美国黑人和重构美国黑人 rAfA 的群体差异重复出来。

我们通过第一种方法检测到 6 个区域的祖先贡献有偏离。但是，祖先偏离最大的区域的祖先偏离也<0.026，这比之前对混合人群的研究报道的都要低 World [9-11,15-17]。例如，波多黎各人群中的非洲祖先成分比平均值高出 0.14，比本研究的值高出 5 倍还多 [10]。我们认为本研究中的祖先贡献变异很小可能是由于本研究用了更大的样本，更高的位点密度，高度精确的定相基因分型数据和更有效的统计方法。按照最大的祖先贡献偏离，我们估算群体混合后最大的自然选择系数应该低于 0.002，（假设在连续基因流模型下遗传 14 代）。 考虑到各种统计误差和遗传漂变的影响， 真实的自然选择系数会比 0.002 小很多。这个结果也基本符合美国黑人在奴隶贸易结束之后近两百多年较低的死亡率。

我们使用新方法检测到多个非洲祖先人群和 YRI 之间有很大群体差异的区域。这些区域有些位点的等位基因频率相差 0.1 以上，这说明非洲祖先人群离开非洲之后受到的自然选择还是很大的。这些新鉴定出来的区域和第一个方法鉴定出来的区域基本上不重叠（MHC 区域除外），这可能反映了美国黑人的非洲祖先人群在混合前后经历的环境不一样。智能通路分析高群体差异的多态位点发现他们富集了很多疾病相关基因，比如代谢疾病相关基因($P = 1.51\times10^{-16}$)，内分泌相关基因($P = 2.23 \times10^{-16}$)，免疫相关基因($P = 9.30\times10^{-12}$)和遗传疾病相关基因($P = 5.67\times10^{-}$

[11])。其中，自然选择候选基因 PSCA, ZP4 和 AKAP12 与美国黑人特异性高发疾病高血压和前列腺癌相关 [59,60]。另外五个处于高群体差异区的基因（*CD36*, *HBB*, *HBD*, *HLA-B*, *HLA-DR*)与疟疾的感染相关。

相对于高加索人种，美国黑人的几个主要癌症的患病率和总的癌症患病率及死亡率都比较高 [61]，同时肥胖相关疾病比如糖尿病，高血压和前列腺癌的患病率也比较高 [59]。有趣的是，我们发现很多使用新方法检测到的自然选择候选基因与美国特异性高发疾病如高血压，前列腺癌和系统性硬化病等相关联。特别是，其中一个高群体差异性多态性位点（rs2294008; F_{ST} =0.04561）是基因 *PSCA* 中的一个错义突变 c.57T>C（蛋氨酸变成苏氨酸；p.Met1Thr)。该位点位于 8q24 区域，已经被多个研究发现与胃癌和膀胱癌相关联 [62-64]。同时，很多研究发现 8q24 区域有多个相对独立位点与美国黑人中的前列腺癌发病相关联 [65-67]。我们根据观察到的这种现象提出一个假说：美国黑人高发疾病相关联的基因可能在美国黑人的非洲祖先人群适应新的环境中起了很大作用，所以他们才会在非洲祖先人群和 YRI 之间表现出很高群体差异。我们认为进一步分析 8q24 区域能够对美国黑人的前列腺癌高发病率和发病机制有进一步了解，同时也有利于发现胃癌和膀胱癌的在日本发病的分子遗传机制。

在所有通过群体差异鉴定的自然选择候选基因中，有五个基因(*CD36*, *HBB*, *HBD*, *HLA-B*, *HLA-DR*)可是由于疟疾受到了自然选择 [68,69]。在热带非洲，疟疾高发带来的高死亡率形成很高的选择压。当一些基因的功能丧失或部分丧失突变能够为个体提供一定抵抗疟疾的能力时，这种突变的等位基因频率会在疟疾高发的非

洲升的很高，虽然这些抵抗疟疾的突变本身是有害的。但是，当美国黑人的非洲祖先到达美洲后，疟疾已经不再是一个重要的选择压。之前能够保护个体逃避或减弱疟疾病情的等位基因不再有任何优势，其负面效应可能会逐渐显示出来 [70]。这样，抵抗疟疾的这些等位基因的频率在美国黑人的非洲祖先人群中的会逐渐下降？

随后，我们试图对非洲祖先人群和现代非洲人群中的抵抗疟疾等位基因频率进行比较，但是本研究使用的的基因型数据中并没有包含之前报道的直接抵抗疟疾的功能性等位基因或突变位点。我们通过分析与这些功能性突变高度连锁的多态性位点来推断抵抗疟疾的等位基因频率变化。多态性位点 rs3211938 是基因 CD36 上的一个无义突变 c.1389T>G (p.Tyr325X)，该突变使得蛋白翻译提前终止，基因功能基本丧失。因为疟疾或其它未知因数，CD36 被报道在非洲人群中受到很强的自然选择[47-50]。我们发现三个位于 7q21 区域的在非洲祖先人群和 YRI 之间高度群体差异的多态位点(Table 3)均与 rs3211938 高度连锁(在 YRI 中，每一个位点与 rs3211938 的 r^2 大于 0.4)。令人兴奋的是，我们确实发现这些与抵抗疟疾的等位基因（rs3211938:G)连锁的等位基因在美国黑人非洲祖先人群中的频率比现代非洲人群 YRI 的频率要低很多(Table 3)。多态位点 rs334 (c.70A>T; p.Glu7Val)是一个谷氨酸突变成缬氨酸的错义突变，该突变导致了研究最彻底的隐性遗传病镰刀性贫血（sickle cell anemia [MIM 603903]) [56,57,71]，同时该突变也为一般个体提供了抵抗疟疾的能力。高群体差异的多态位点 rs7952293 与 rs334 有很强的连锁，特别是，rs7952293:A 和 rs334:T 组成的单体型 AT 占了 rs334:T 构成的单体型的 87%。我们发现 rs7952293:A 在非洲祖先人群中的频率（0.2261）比 YRI 中的频率（0.3172）低很多。这同样说明非洲祖先人群中抵抗疟疾的等位基因频率相对

于现代非洲人群已经下降了很多。同时这类现象也表明不同地区的不同病原体在能够重新塑造人的等位基因频率，病原体是人类进化的一个重要动力。

本研究利用了非常大的样本量和高密度的单核苷酸多态性位点来降低随机统计误差和提高研究的可靠性。但是我们发现美国黑人全基因组祖先贡献偏离最大的区域也小于 0.026 。考虑到遗传漂变和统计误差，在混合人群中检测到这样弱的自然选择信号仍然是一个巨大的挑战。我们建议以后利用近期混合人群检测自然选择至少要收集上千个样本，甚至更多的样本来降低统计误差。我们用新的方法检测到很多与美国黑人特异性高发疾病如高血压和前例腺癌相关联的基因，也检测到五个与疟疾相关联的基因。因此我们的新方法对于检测混合人群混合前后的自然选择有很高的统计效力，可以推广到其它混合人群如拉美人群和中国维吾尔族人群中的自然选择研究中 [72-74]。

表格

Table 1. Regions Showing Excess of European or African Ancestry.

Regions	Position	Excess ancestry	Size (bp)	SNPs	Highest deviation	Genes	Pathways	Related diseases
1p36	chr1:17409539..21604321	African	4194783	489	0.0253	*AKR7A2**, *IGSF21*, *DDOST**, HTR6 *et al.*	Diabetes pathways, signaling by GPCR, metabolism of amino acids	Diabetes, pancreatic cancer
2q37	chr2:241750403..242568618	African	818216	16	0.0231	*SEPT2**, *HDLBP**, *PDCD*1*, *FARP2 et al.*	Signaling in immune system, Axon guidance, metabolism of nucleotides	Bladder cancer, lung cancer, coronary atherosclerosis
2p22	chr2:37451925..37508581	European	56657	9	0.0230	*QPCT*, (*EIF2AK2** 222kb)	Influenza infection	Influenza infection
3q13	chr3:116930811..118313302	European	1382492	216	0.0253	*LSAMP**	Homophilic adhesion	Osteosarcoma
6q26	chr6:163653158..163653428	European	271	2	0.0225	*PACRG**	Mediate proteasomal degradation	Juvenile Parkinson's disease
16q21	chr16:61214438..61242497	European	28060	9	0.0229	NA	NA	NA

NA: not available. * Denotes genes associated with diseases. Genes in parentheses are strong candidates out of the chromosome location but closest.

Table 2. Regions with highly differentiated allele frequency between AAF and YRI (F_{ST} >0.0452).

Regions or SNPs	Position	Size (bp)	SNPs	Highest F_{ST}	Genes	Pathways	Related disease
1p21	chr1:100125058..100183875	58817	2	0.0562	*AGL**	Metabolism of carbohydrate	Glycogen storage disease
1q22	chr1:153401959..153464086	62127	4	0.0692	*THBS3**, *MUC1**, *MTX1*, *TRIM46*, *KRTCAP2*	Signaling by PDGF	Stomach cancer, breast cancer, osteosarcoma
rs12094201	chr1:236509336	1	1	0.0561	(*ZP4** 389kb)	NA	Hypertension, Non-alcoholic fatty liver
rs7642575	chr3: 31400165	1	1	0.0453	(*STT3B, OSBPL10** 149 kb)	NA	Peripheral arterial disease
6p21-p22	chr6:26554684..33961049	7406365	11	0.0711	*HLA-B*, HLA-C, EHMT2*, HLA-DPA1*, HLA-DRB5, EHM, BTN3A3, et al*	Signaling by GPCG, signaling in immune system, HIV infection, Diabetes pathway	HIV, Crohn's disease, rheumatoid arthritis, juvenile idiopathic arthritis, colorectal cancer, systemic sclerosis
6q25	chr6:151555551..151569258	13707	2	0.0545	(*AKAP12** 40kb)	Cell growth	Hypertension, hemorrhagic stroke
rs10499542	chr7: 22235870		1	0.04606	*RAPGEF5**	GTP/GDP-regulation	Thyroid stimulating hormone
7q21	chr7:79768487..80482597	714110	10	0.0946	*CD36*, SEMA3C*	Metabolism of lipids and lipoprotein	Metabolic syndrome, malaria
8q24	chr8:143754039..143758933	4894	2	0.04679	*PSCA**	NA	Prostate cancer, bladder cancer, gastric cancer
11p15	chr11:5034229..5421456	387227	3	0.0617	*HBB*, HBD*, HBE1*, HBG2, OR51I1, et al*	Signaling by GPCR	Sickle cell disease, beta-thalassemia, malaria
rs4883422	chr12:7189594	1	1	0.04721	*CLSTN3*	NA	NA
rs6491096	chr13:25488362	1	1	0.04716	*ATP8A2*	NA	NA
rs1075875	chr16: 47595721	1	1	0.0766	(*CBLN1* 277kb)	NA	NA
rs6015945	chr20:59319574	1	1	0.0627	*CDH4**	Cell junction organization	Alzheimer's Disease

NA: not available. * Denotes the genes associated with diseases. Genes in parentheses are strong candidates out of the chromosome location but closest.

Table 3. SNPs showing strong linkage with rs3211938(G) in CD36.

dbSNP_id	Alleles	Allele linked with rs3211938(G)	Frequency (YRI)	Frequency (AAF)	F_{ST} (AAF-YRI)	r^2 (YRI)
rs10216027	C/T	T	0.29	0.16	0.0569	0.514
rs1404315	A/C	C	0.32	0.16	0.0693	0.489
rs1722504	C/T	C	0.38	0.22	0.0652	0.403

图示

Figure 1. Schematic of the two selection events (pre-and post-admixture) in African Americans.

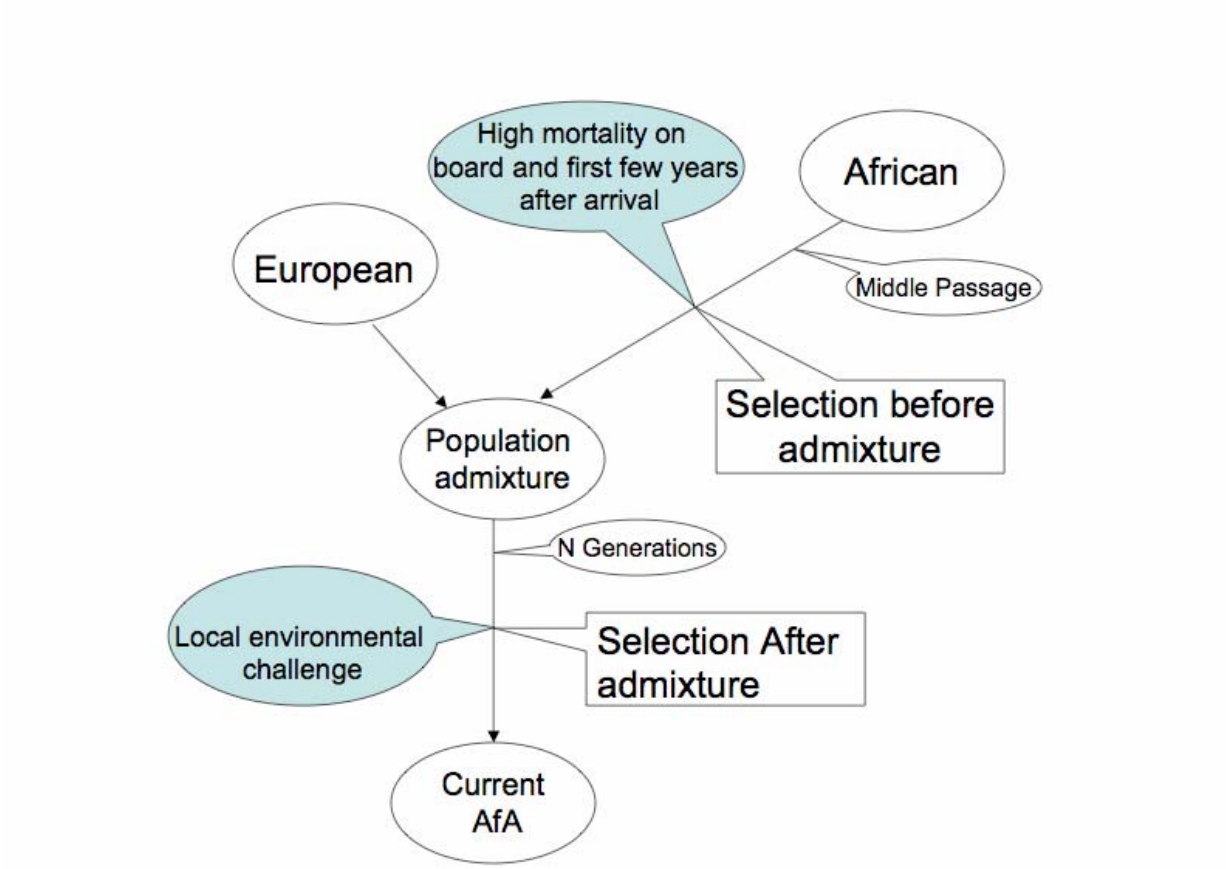


Figure 2. Schematic of the strategy for detecting natural selection by comparing the African components of ancestry with indigenous African populations.

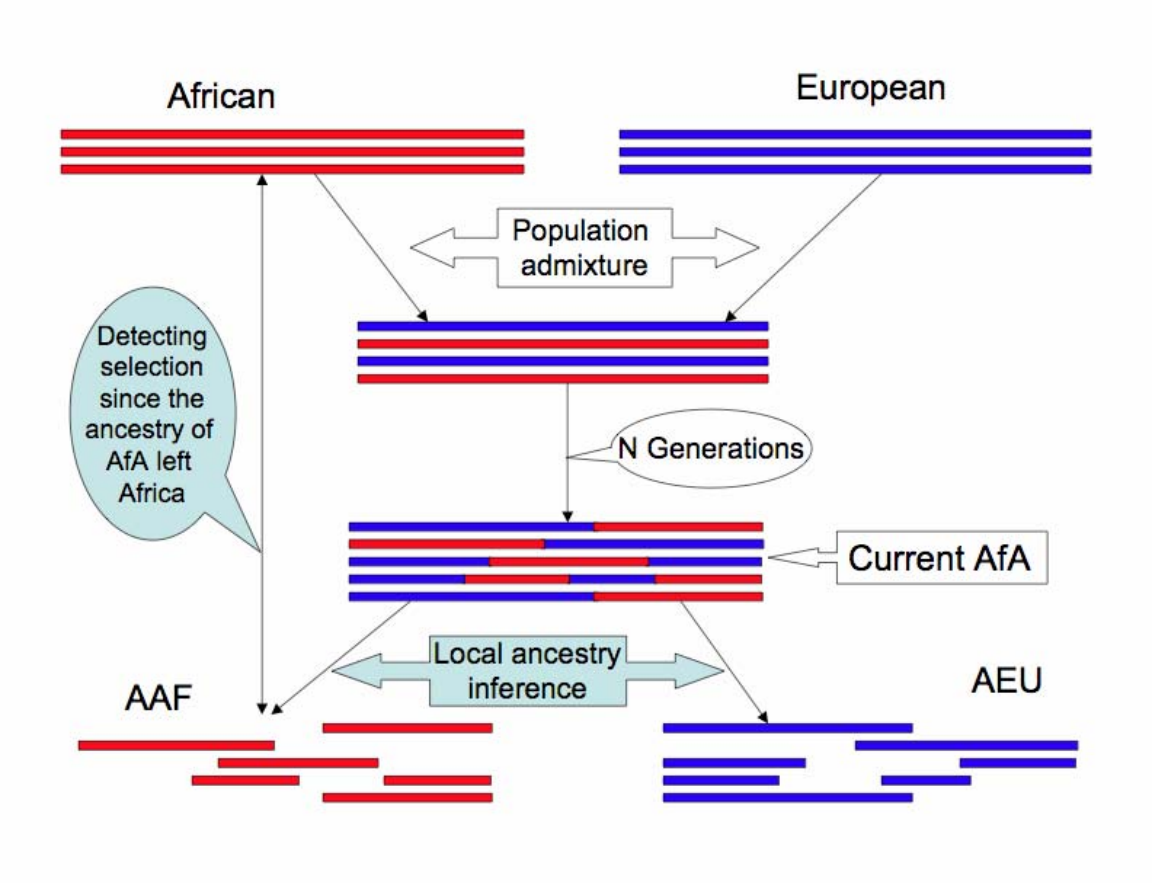


Figure 3. Analyses of the first two principal components. The % Eigenvalue is the percentage of the total variance in the Top ten PCs. (A) 1,890 AfA and 3320 samples from 19 other populations represent global-wide populations. (B) 1,890 AfA and all possible European or African ancestral populations. (C) 1,890 AfA, YRI and all possible European ancestral populations. (D) 1,890 AfA and two putative parental populations (YRI and CEU).

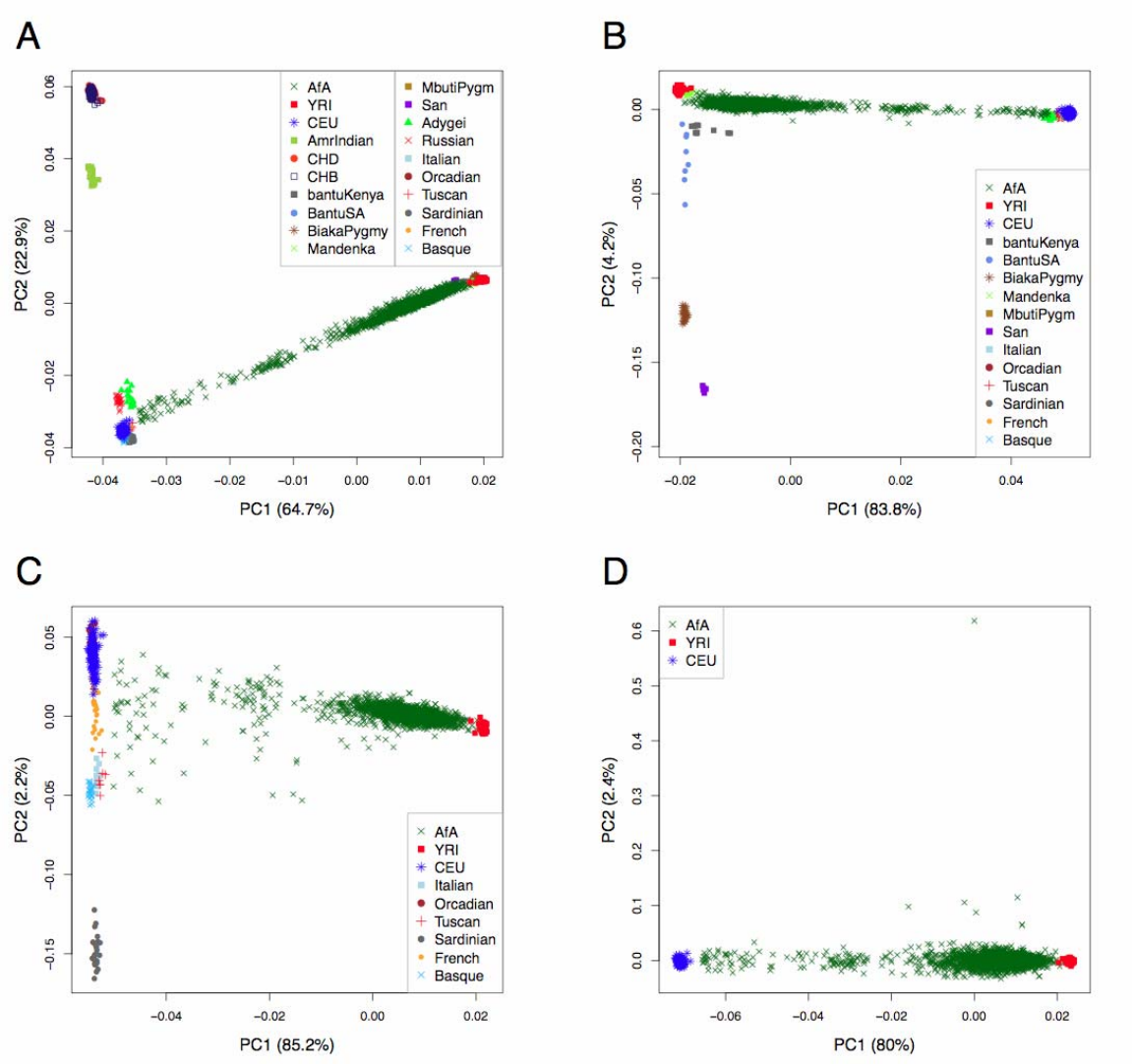


Figure 4. FRAPPE analysis of 8 African populations, 9 European populations and AfA when K=2. Each individual is represented by a vertical line, which is partitioned into two segments corresponding to the inferred membership of the genetic clusters indicated by the colors.

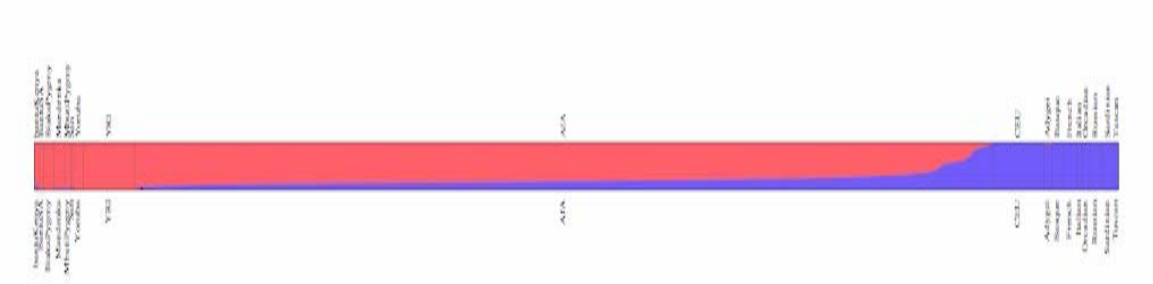


Figure 5. Genome-wide Distribution of European Ancestral Contributions. Mean European ancestral contribution across 1,890 African American individuals at each SNP. Green line is the estimated genome-wide mean European ancestral contribution (21.68%). Blue bands indicate +2 and -2 SDs from the mean ancestral contribution and red Bands indicate +3 and -3 SDs from the mean ancestral contribution.

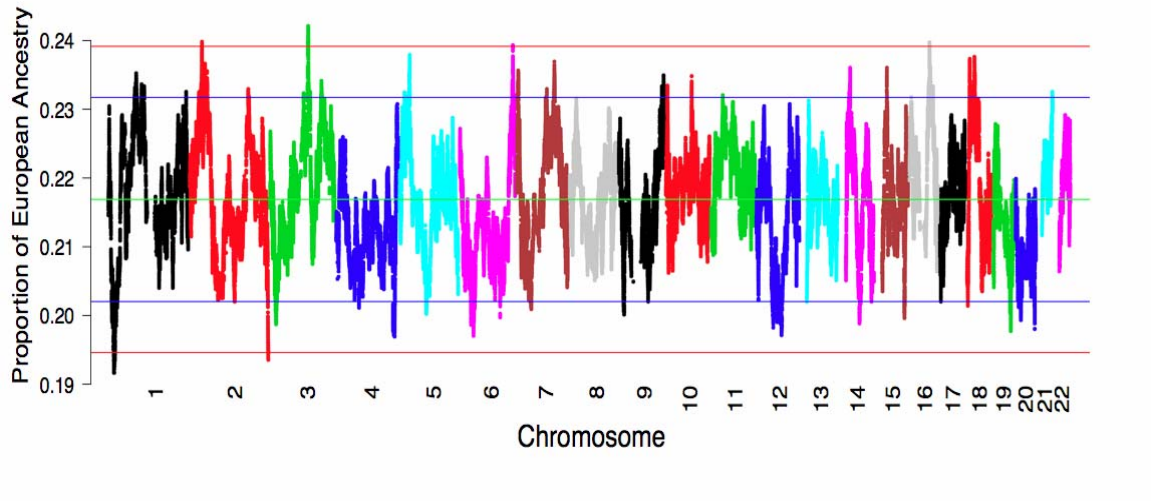


Figure 6. Q-Q plot of locus-specific F_{ST} between AAF and YRI. This Q-Q plot compares empirical values on the vertical axis to simulated values on horizontal axis. Red line shows null distribution that F_{ST} of empirical data are the same as that of simulated.

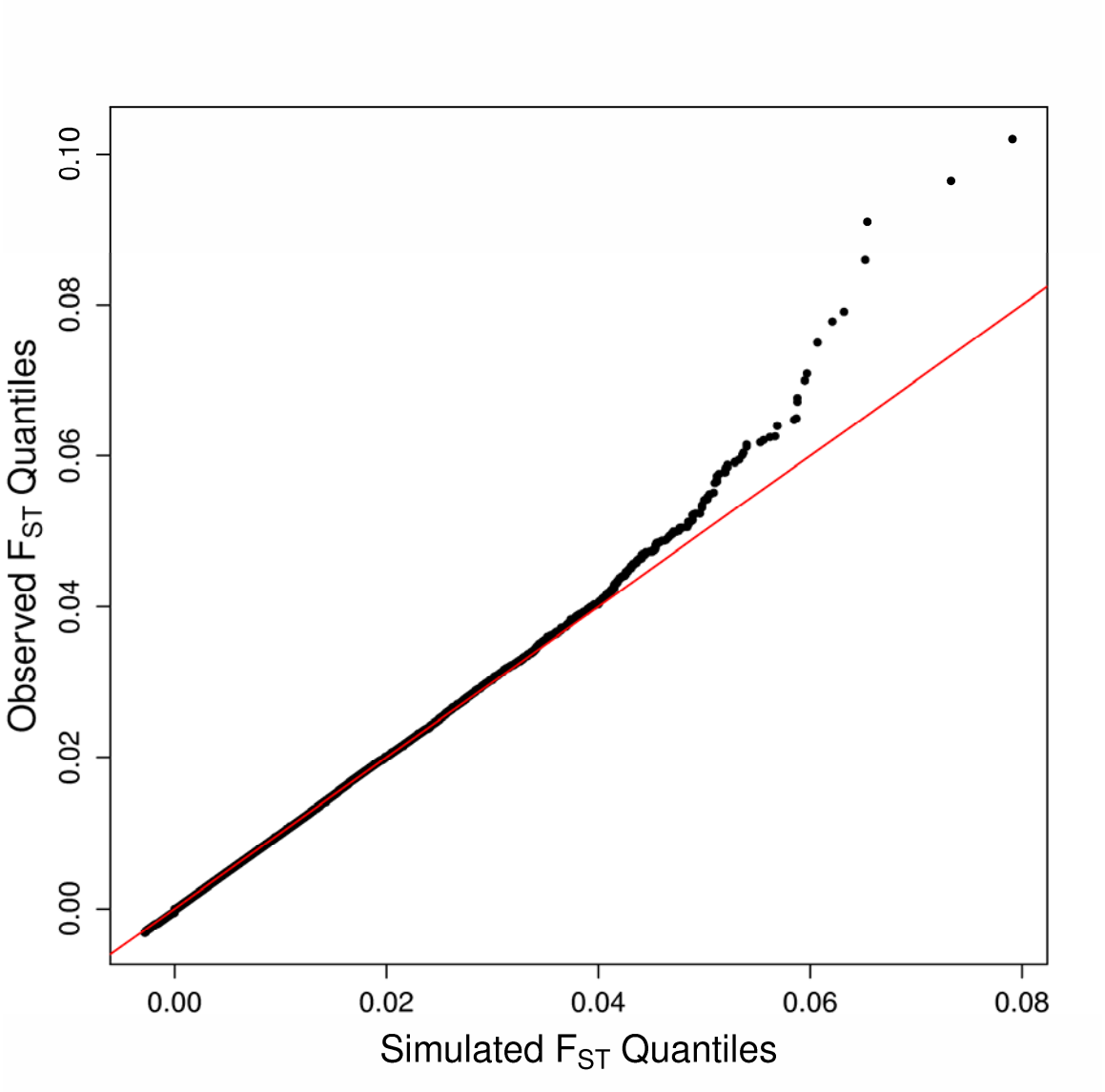


Figure 7. Genomic Distribution of F_{ST} between AAF and YRI (A) and Genomic Distribution of F_{ST} between African American and rAfA (B). The dashed red horizontal line indicates the cutoff threshold (99.99th percentile). Locus-specific F_{ST} between YRI and CEU were calculated when MAF >0.05 in both populations. The rAfA was constituted according to the ancestry proportion of CEU and YRI under neutrality.

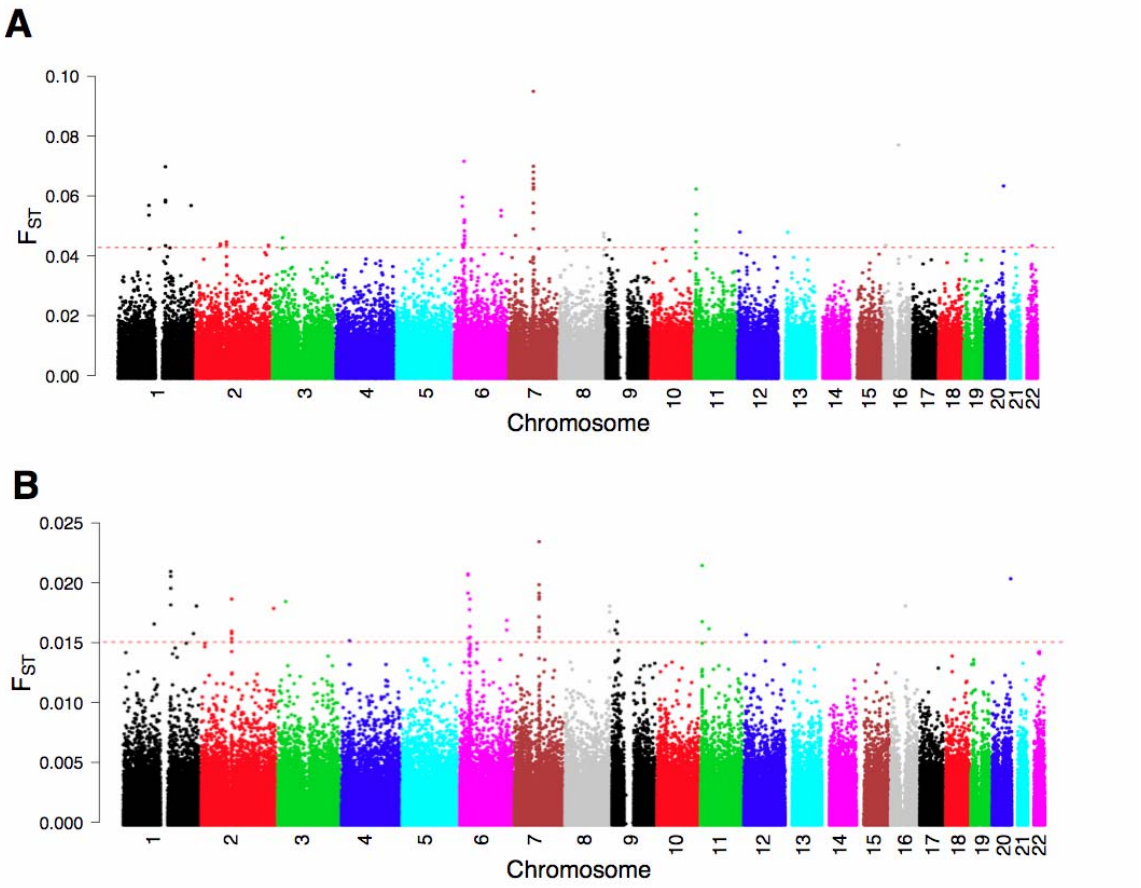


Figure 8. Genome-wide distribution of F_{ST} between AfA and rAfA. rAfA is construct by CAU-GWAS and APP (African parental population of AfA constructed by YRI, Mandenka and Bantu) according to estimated admixture proportion. The dashed red horizontal line indicates the cutoff threshold (99.99th percentile).

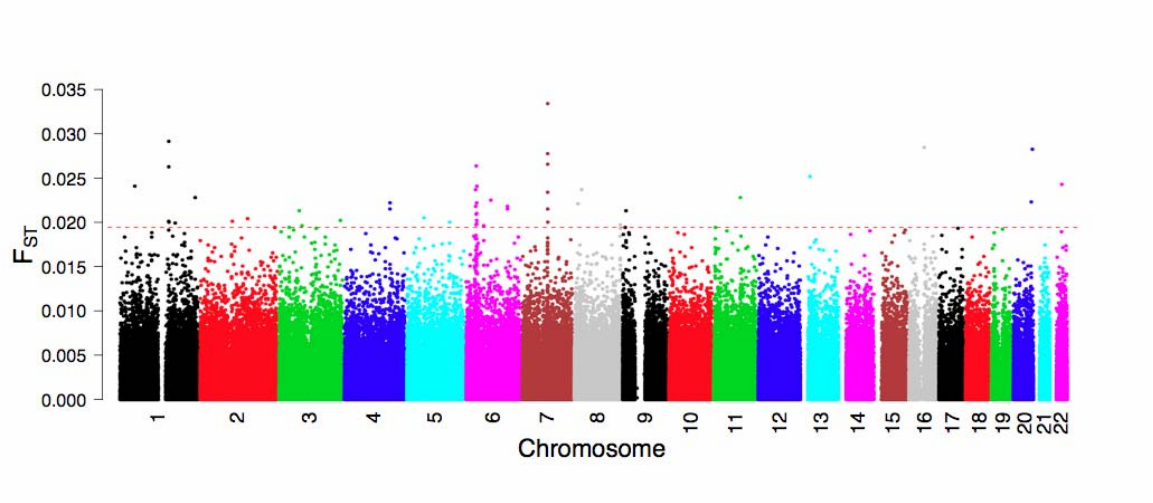
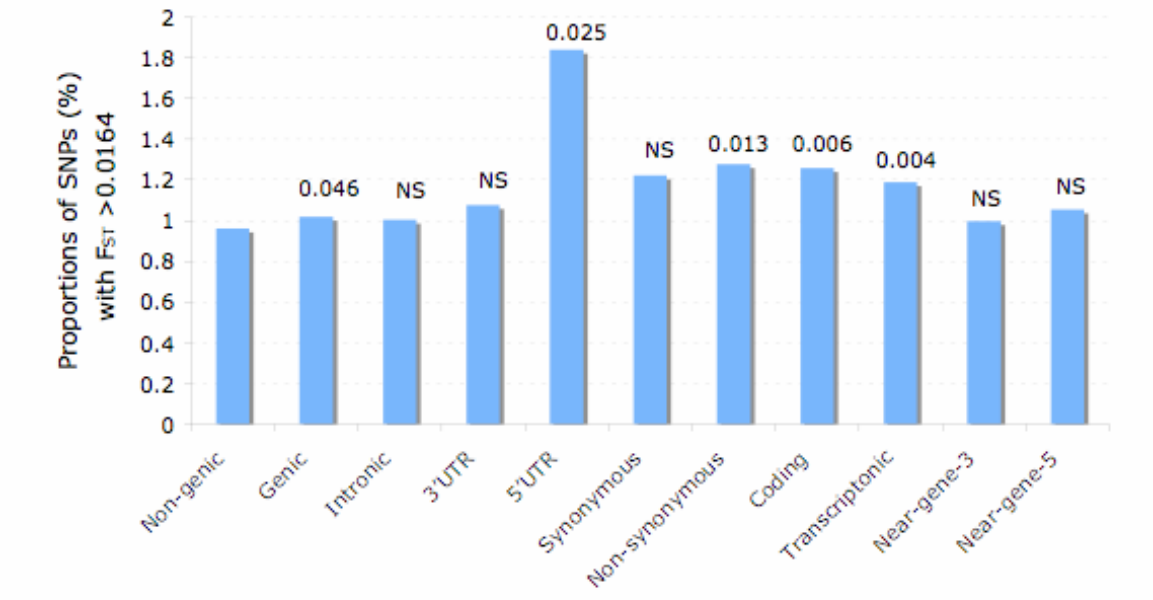


Figure 9. Enrichment of high F_{ST} loci for different SNP categories. Observed excess of high F_{ST} loci in different SNP classes, with respect to nongenic class, among high F_{ST} bin (99th percentile; F_{ST} >0.0164). The values on the bar are p-values of χ^2 tests. "NS" stands for "not significant".



参考文献

1. Balaresque PL, Ballereau SJ, Jobling MA (2007) Challenges in human genetic diversity: demographic history and adaptation. Hum Mol Genet 16 (R2): R134-139.

2. Kimura M (2003) The neutral theory of molecular evolution. (United Kingdom: Cambridge University Press, Cambridge).

3. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-918.

4. Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19: 711-722.

5. Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG (2007) Recent and ongoing selection in the human genome. Nat Rev Genet 8: 857-868.

6. Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, et al. (2008) Adaptations to climate in candidate genes for common metabolic disorders. PLoS Genet 4: e32.

7. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. Nat Genet 40: 340-345.

8. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19: 826-837.

9. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, et al. (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc Natl Acad Sci U S A 107: 786-791.

10. Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, et al. (2007) Recent genetic selection in the ancestral admixture of Puerto Ricans. Am J Hum Genet 81:

626-633.

11. Basu A, Tang H, Zhu X, Gu CC, Hanis C, et al. (2008) Genome-wide distribution of ancestry in Mexican Americans. Hum Genet 124: 207-214.
12. Zakharia F, Basu A, Absher D, Assimes TL, Go AS, et al. (2009) Characterizing the admixed African ancestry of African Americans. Genome Biol 10: R141.
13. Thomas H (1999) The Slave Trade: The Story of the Atlantic Slave Trade. Simon & Schuster 1440-1870.
14. Stannard D (1993) American Holocaust. Oxford University Press.
15. WORKMAN PL, BLUMBERG BS, COOPER AJ (1963) SELECTION, GENE MIGRATION AND POLYMORPHIC STABILITY IN A U. S. WHITE AND NEGRO POPULATION. Am J Hum Genet 15: 429-437.
16. Blumberg BS, Hesser JE (1971) Loci differentially affected by selection in two American black populations. Proc Natl Acad Sci U S A 68: 2554-2558.
17. Reed TE (1969) Caucasian genes in American Negroes. Science 165: 762-768.
18. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58.
19. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100-1104.
20. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.
21. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2: e190.
22. Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: analytical

and study design considerations. Genetic epidemiology 28: 289-301.

23. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567-1587.

24. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

25. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38: 1358-1370.

26. Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, et al. (2004) Methods for high-density admixture mapping of disease genes. Am J Hum Genet 74: 979-1000.

27. Tang H, Coram M, Wang P, Zhu X, Risch N (2006) Reconstructing genetic ancestry blocks in admixed individuals. American Journal of Human Genetics 79: 1-12.

28. Sankararaman S, Sridhar S, Kimmel G, Halperin E (2008) Estimating local ancestry in admixed populations. Am J Hum Genet 82: 290-303.

29. Sundquist A, Fratkin E, Do CB, Batzoglou S (2008) Effect of genetic divergence in identifying ancestral origin using HAPAA. Genome Res 18: 676-682.

30. Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, et al. (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet 5: e1000519.

31. Long JC (1991) The genetic structure of admixed populations. Genetics 127: 417-428.

32. Thorisson GA, Smith AV, Krishnan L, Stein LD (2005) The international HapMap project web site. Genome research 15: 1592.

33. McVean G (2009) A Genealogical Interpretation of Principal Components Analysis. Plos Genetics 5.

34. Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, et al. (2004) A

high-density admixture map for disease gene discovery in african americans. The American Journal of Human Genetics 74: 1001-1013.

35. Oleksyk TK, Smith MW, O'Brien SJ (2010) Genome-wide scans for footprints of natural selection. Philos Trans R Soc Lond B Biol Sci 365: 185-205.

36. Praml C, Schulz W, Claas A, Mollenhauer J, Poustka A, et al. (2008) Genetic variation of Aflatoxin B1 aldehyde reductase genes (AFAR) in human tumour cells. Cancer Lett 272: 160-166.

37. Cui Y, Tian M, Zong M, Teng M, Chen Y, et al. (2009) Proteomic analysis of pancreatic ductal adenocarcinoma compared with normal adjacent pancreatic tissue and pancreatic benign cystadenoma. Pancreatology 9: 89-98.

38. Yen CC, Chen WM, Chen TH, Chen WY, Chen PC, et al. (2009) Identification of chromosomal aberrations associated with disease progression and a novel 3q13. 31 deletion involving LSAMP gene in osteosarcoma. International journal of oncology 35: 775.

39. Kresse SH, Ohnstad HO, Paulsen EB, Bjerkehagen B, Szuhai K, et al. (2009) LSAMP, a novel candidate tumor suppressor gene in human osteosarcomas, identified by array comparative genomic hybridization. Genes, Chromosomes and Cancer 48.

40. McAllister CS, Toth AM, Zhang P, Devaux P, Cattaneo R, et al. (2010) Mechanisms of protein kinase PKR-mediated amplification of beta interferon induction by C protein-deficient measles virus. J Virol 84: 380-386.

41. Pereira RM, Teixeira KL, Barreto-de-Souza V, Calegari-Silva TC, De-Melo LD, et al. (2010) Novel role for the double-stranded RNA-activated protein kinase PKR: modulation of macrophage infection by the protozoan parasite Leishmania. Faseb J 24: 617-626.

42. Meltzer M (1993) Slavery: A World History. Da Capo Press.

43. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome research 12: 1805.

44. Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005) Measures of human population structure show heterogeneity among genomic regions. Genome research 15: 1468.

45. Baruch DI, Gormely JA, Ma C, Howard RJ, Pasloske BL (1996) Plasmodium falciparum erythrocyte membrane protein 1 is a parasitized erythrocyte receptor for adherence to CD36, thrombospondin, and intercellular adhesion molecule 1. Proc Natl Acad Sci U S A 93: 3497-3502.

46. Oquendo P, Hundt E, Lawler J, Seed B (1989) CD36 directly mediates cytoadherence of Plasmodium falciparum parasitized erythrocytes. Cell 58: 95-101.

47. Erdman LK, Cosio G, Helmers AJ, Gowda D, Grinstein S, et al. (2009) CD36 and TLR Interactions in Inflammation and Phagocytosis: Implications for Malaria. The Journal of Immunology.

48. Fry AE, Ghansa A, Small KS, Palma A, Auburn S, et al. (2009) Positive selection of a CD36 nonsense variant in sub-Saharan Africa, but no association with severe malaria phenotypes. Hum Mol Genet 18: 2683-2692.

49. Aitman TJ, Cooper LD, Norsworthy PJ, Wahid FN, Gray JK, et al. (2000) Malaria susceptibility and CD36 mutation. Nature 405: 1015-1016.

50. Omi K, Ohashi J, Patarapotikul J, Hananantachai H, Naka I, et al. (2003) CD36 polymorphism is associated with protection from cerebral malaria. Am J Hum Genet 72: 364-374.

51. Omi K, Ohashi J, Patarapotikul J, Hananantachai H, Naka I, et al. (2002) Fcgamma receptor IIA and IIIB polymorphisms are associated with susceptibility to cerebral malaria. Parasitol Int 51: 361-366.

52. Esselens C, Malapeira J, Colome N, Casal C, Rodriguez-Manzaneque JC, et al. (2010) The cleavage of semaphorin 3C induced by ADAMTS1 promotes cell migration. J Biol Chem 285: 2463-2473.

53. Garrigan D, Hedrick PW (2003) Perspective: detecting adaptive molecular polymorphism: lessons from the MHC. Evolution 57: 1707-1722.

54. Davila S, Froeling FE, Tan A, Bonnard C, Boland GJ, et al. (2010) New genetic associations detected in a host response study to hepatitis B vaccine. Genes Immun 11: 232-238.

55. Li Y, Dinwiddie DL, Harrod KS, Jiang Y, Kim KC (2010) Anti-inflammatory effect of MUC1 during respiratory syncytial virus infection of lung epithelial cells in vitro. Am J Physiol Lung Cell Mol Physiol 298: L558-563.

56. Ashley-Koch A, Yang Q, Olney RS (2000) Sickle hemoglobin (HbS) allele and sickle cell disease: a HuGE review. Am J Epidemiol 151: 839-845.

57. Wood ET, Stover DA, Slatkin M, Nachman MW, Hammer MF (2005) The beta -globin recombinational hotspot reduces the effects of strong selection around HbC, a recently arisen mutation providing resistance to malaria. Am J Hum Genet 77: 637-642.

58. Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39: 197-218.

59. Goran MI (2008) Ethnic-specific pathways to obesity-related disease: the Hispanic vs. African-American paradox. Obesity (Silver Spring) 16: 2561-2565.

60. Smith MW, O'Brien SJ (2005) Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nat Rev Genet 6: 623--632.

61. Jemal A, Siegel R, Ward E, Murray T, Xu J, et al. (2006) Cancer statistics, 2006. CA Cancer J Clin 56: 106-130.

62. Sakamoto H, Yoshimura K, Saeki N, Katai H, Shimoda T, et al. (2008) Genetic variation

in PSCA is associated with susceptibility to diffuse-type gastric cancer. Nat Genet 40: 730-740.

63. Matsuo K, Tajima K, Suzuki T, Kawase T, Watanabe M, et al. (2009) Association of prostate stem cell antigen gene polymorphisms with the risk of stomach cancer in Japanese. Int J Cancer 125: 1961-1964.

64. Wu X, Ye Y, Kiemeney LA, Sulem P, Rafnar T, et al. (2009) Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. Nat Genet 41: 991-995.

65. Freedman ML, Haiman CA, Patterson N, McDonald GJ, Tandon A, et al. (2006) Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proc Natl Acad Sci U S A 103: 14068-14073.

66. Al Olama AA, Kote-Jarai Z, Giles GG, Guy M, Morrison J, et al. (2009) Multiple loci on 8q24 associated with prostate cancer susceptibility. Nat Genet 41: 1058-1060.

67. Yeager M, Chatterjee N, Ciampa J, Jacobs KB, Gonzalez-Bosquet J, et al. (2009) Identification of a new prostate cancer susceptibility locus on chromosome 8q24. Nat Genet 41: 1055-1057.

68. Kwiatkowski D (1999) The molecular genetic approach to malarial pathogenesis and immunity. Parassitologia 41: 233-240.

69. Kwiatkowski DP (2005) How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet 77: 171-192.

70. Platt OS, Brambilla DJ, Rosse WF, Milner PF, Castro O, et al. (1994) Mortality in sickle cell disease. Life expectancy and risk factors for early death. N Engl J Med 330: 1639-1644.

71. Winichagoon P, Fucharoen S, Chen P, Wasi P (2000) Genetic factors affecting clinical severity in beta-thalassemia syndromes. J Pediatr Hematol Oncol 22: 573-580.

72. Xu S, Huang W, Qian J, Jin L (2008) Analysis of genomic admixture in Uyghur and its implication in mapping strategy. Am J Hum Genet 82: 883-894.

73. Xu S, Jin L (2008) A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. Am J Hum Genet 83: 322-336.

74. Xu S, Jin W, Jin L (2009) Haplotype-sharing analysis showing Uyghurs are unlikely genetic donors. Mol Biol Evol 26: 2197-2206.

第四章：疾病基因受到的自然选择

摘要

传统上，遗传相关疾病可以分为少见的孟德尔遗传病（Mendelian Diseases）和较常见的复杂疾病（Common Diseases or Complex Diseases），这种分类极大地方便了遗传咨询和疾病基因鉴定。但是，通过分析两个常用的疾病数据库，我们发现54%的孟德尔遗传病关联基因出现在复杂疾病基因数据库里，这比预期的重叠基因数目显著增多（$P < 2.2\times10^{-16}$, x^2 test)。我们姑且把这类在两种疾病中都起作用的基因简称为双联基因（Mendelian and Complex disease gene; MC gene）。这样我们把人类的全部基因分为必需基因，双联基因，孟德尔遗传病基因（去除与复杂疾病相关基因），复杂疾病基因（去除与孟德尔遗传病相关基因）和其它基因。对人类的多态性数据分析后，我们发现双联基因和复杂疾病基因都受到了近期的正选择，而必需基因和孟德尔遗传病基因受到较强的负向选择。对物种间差异数据分析表明必需基因总是最保守，这支持必需基因在长期的进化史中总是受到强的负向选择。而双联基因一般是第二保守，意味着在进化中也受到较强的负向选择。拷贝数变异在不同类基因类型中的富集分析也支持双联基因受到较强的正选择与负向选择。除此之外，我们也比较了各类基因在基因表达模式，基因结构，蛋白蛋白相互作用，群体分化等方面的差异。虽然双联基因的表达水平不比其它疾病基因显著增高，但是双联基因的组织表达特异性比其它任何类型的基因都显著性的高($P<7.5\times10^{-5}$)。相比其它疾病基因，双联基因还表现出与较多的表型和疾病种类相关联，参与了更复杂的网路调控。我们以此推测双联基因的很多特征与他们在复杂疾病和孟德尔遗传病中的双重作用有关。同时，我们在该研究中发现拷贝数变异在很

多复杂疾病产生过程中可能起着很重要的作用。据我们所知，该工作是第一个对双联基因特征进行系统分析的研究，我们在这个比较分析过程中也对其它四类基因有了新的认识。我相信，这样的一个系统的比较分析，有助于人们深入了解疾病基因的产生和进化机制。

前言

遗传几乎在所有的人类疾病中都起着一定作用。对于一些过去认为由环境因素引起的疾病，疾病感染的易感性和疾病的患病程度也已经被发现与遗传多态相关联[1, 2]。在一些情况下，基因组上的一个突变或变异即可引起一种疾病。但是在另一些情况下，一个突变或变异必须和其它多态位点相互作用，并在特定环境下才会引起疾病。这样，遗传相关疾病可分为孟德尔遗传病和复杂疾病[3, 4]。孟德尔遗传病如镰刀型细胞贫血和囊肿性纤维化是单个位点变异引起的，它们在人群中的遗传传递符合孟德尔遗传定律。但是复杂疾病如常见糖尿病和癌症是由多位点突变及与环境互作引起的。流行病学研究已经发现它们的发病有家族倾向性但又不遵循孟德尔遗传模式。

这种疾病分类方法有利于疾病的遗传咨询和疾病基因鉴定方法的发展[3-6]。虽然孟德尔遗传病在一般人群中的发病率极低，但是一代代的人类遗传学家在孟德尔遗传病的研究上面付出了巨大的心血并取得了某些成功[3, 7]。这是因为孟德尔遗传病和特征的遗传背景单一并有着非常高的外显率，更容易用来发现致病位

点及其发病机理。在这个过程中揭示的功能通路或网络可能也在复杂疾病的发病过程中起着重要作用[3,7]。甚至在人类基因组计划完成之前（Human Genome Project; HGP)，很多孟德尔遗传病的致病基因已经通过连锁定位 （linkage mapping）确定下来[8,9]。对孟德尔遗传病的研究极大地增加了我们对致病性突变，基因功能和基因调控的认识，同时也有助于发展针对这些疾病的诊断方法和治疗方法[7]。

但是，对于孟德尔遗传病的研究并没有太多的增加我们对复杂疾病遗传机制的认识[3,6]。在过去的的六年中，基因分型技术的进步诱使很多科学家重新关注复杂疾病。全基因组关联研究（Genome-wide association studies; GWAS）一次分析成千上万个样本上的几十万或上百万个多态性位点，并且已经可重复性的鉴定出近千个复杂疾病相关联的基因[10,11]。这些研究不仅发现了很多与复杂疾病相关联的新位点，同时使得我们对很多基因的功能和生物学意义有了进一步的了解。但是复杂疾病关联位点只能解释已知遗传力的很小一部分，即使对于已知的关联性，要想阐明表型和遗传多样性联系的机制也是一个巨大的挑战[12]。

虽然复杂疾病关联基因和孟德尔遗传病的致病基因在大多数情况下是不一致的，但是一些基因被发现在两类疾病中都是风险因素[13-15]。如果一个基因在复杂疾病和孟德尔遗传病中都是风险因素，那么这个基因会不会有什么特殊性的特征？除此之外，阐明疾病基因的进化有利于我们认识人类疾病的起源，这对于医学遗传学有重要参考意义[16,17]。过去有多个研究已经比较分析了疾病基因和非疾病基因的特征[17-23]，并且也有研究比较分析了孟德尔遗传病连锁基因和复杂疾

病关联基因的异同[24-27]。但是，还没有研究系统分析在孟德尔遗传病和复杂疾病中都是风险因素的基因具有的特征。

在本研究中，我们首先检索一个孟德尔遗传病数据库和一个复杂疾病数据库，由此构建在复杂疾病和孟德尔遗传病中都起作用的基因列表，我们称这些基因为双联基因。我们把人类的基因组上的基因分为五类：双联基因，复杂疾病基因（去除双联基因），孟德尔遗传病基因（去除双联基因），必需基因和其它基因。随后我们比较了各类基因在基因结构，基因表达模式，蛋白蛋白相互作用，群体分化，自然选择，进化等方面的差异。我们的分析表明双联基因在很多方面和其它种类的基因不同。比如，相对于其它疾病基因，双联基因与更多的疾病和表型相关联，参与了更复杂的蛋白-蛋白互作网络，有更高的组织特异性，编码更长的蛋白，有更多的转录物种类。我们还发现双联基因受到更强的正向自然选择和长期的负向选择。

材料和方法

孟德尔遗传病和复杂疾病的相关基因列表

本研究分析了两组孟德尔遗传病相关基因。第一组基因从人类孟德尔遗传网站（Online Mendelian Inheritance in Man database; OMIM）下载[8,9]。OMIM是目前最全面，最详尽，最权威的人类孟德尔遗传病表型和相关基因的公共数据库。我们只从中选取在两个以上实验室或多个家系中能够重复检测出来的基因，这

总共有 3261 个基因。第二组孟德尔遗传病相关基因来至于一个叫做 hOMIM 的数据库，该数据库是 Blekhman 等 [25] 在 2008 年以前对 OMIM 数据库手工整理的出来的基因列表。该数据库只收集了高度可信的的孟德尔遗传病及其致病基因，共 968 个基因。该数据库已经被最近的多个研究使用和引用[25,28,29]。

本研究分析使用了两组复杂疾病关联基因。第一组基因来自于遗传关联研究数据库（genetic association database; GAD）[30,31]。这个数据库收集了几乎所有的遗传关联分析研究结果。我们仅选取数据库中那些只有阳性关联报道的基因，总共有 3117 个基因。另一组复杂疾病关联基因来至于美国人类基因组研究所(National Human Genome Research Institute; NHGRI)的全基因组关联研究数据库（catalog of published genome-wide association studies; GWAS catalog），该数据库只收录采用大于 10 万个单核苷酸多态性位点的全基因组关联分析，共 1799 个基因。这 1799 个基因被包含在第一组的 3117 个复杂疾病关联基因中。

我们用 UCSC 数据库提供的 21,526 个参照基因（reference genes)代表人类基因组上所有的已知基因。为了研究在孟德尔遗传病和复杂疾病中都是风险因素的基因，我们把在 GAD 数据库和 hOMIM 数据库重叠的 524 个基因称为双联基因（Mendelian and complex disease genes; MC genes)；把仅在 hOMIM 数据库但是没有在 GAD 数据库中的 444 个基因称为孟德尔遗传病基因（Mendelian but not complex disease genes; MNC genes）；把仅在 GAD 数据库但是没有在 hOMIM 数据库中的 2,593 个基因称为复杂疾病基因（complex but not Mendelian disease

genes; CNM genes）。我们把既不在 hOMIM 数据库又不再 GAD 数据库的参照基因称为非疾病基因（non-disease genes; ND genes)。我们随后把非疾病基因分成必需基因（essential genes）和其它基因（OTHER genes)。虽然很多研究利用基因表达数据推测看家基因（housekeeping genes)并用来替代必需基因作分析[17,26,32,33]。我们对之前四篇研究报道中的看家基因进行分析后发现只有 13 个基因在这四个研究中是都重复出来。这意味着用基因表达数据推测看家基因来代替必需基因可能不是那么恰当。在本研究中，我们使用了在小鼠中敲除致死或致不孕的 1520 个基因的人类同源基因来代替人类的必需基因[34]。我们把所有不属于上面提到的基因类别的参照基因称为其它基因。

蛋白-蛋白相互作用数据

我们使用了两个高质量的蛋白-蛋白互作数据库。第一个数据库是人类蛋白参照数据库（Human Protein Reference Database; HPRD）第九版[35]，这个数据库是专业的生物学家根据之前的高质量实验数据手工整理得到的。这个数据库包含了9，673个蛋白，收录的这些蛋白至少与另外一个蛋白有相互作用，其总的蛋白相互作用对数是39，240个。这个数据库是现在最全面的蛋白相互作用公共数据库，已经有多个研究分析使用了这个数据库的数据[26,36]。第二个蛋白互作数据库收集了两个高通量酵母双杂交实验(yeast two-hybrid experiment)的结果，包括7，533个基因的22，052个相互作用[37]，我们称为酵母双杂交数据库（YTH）。实际上人类蛋白参考数据库和酵母双杂交数据库可以看成两个互补的数据库。除此之

外，我们把这两个数据库合并到一起做成一个更全面的数据库，我们称为复合数据库（HPRD-YTH）。

基因表达数据分析

生物学基因门户系统 （Gene portal system; BioGPS）是一个可扩展可扩充的基因注释门户网站[38]。我们下载了该数据库（http://biogps.gnf.org）中的所有人类基因表达数据（Human Gene Atlas）。这套数据采用了有44 ,776个探针的U133A/GNF1H 芯片检测了84人类组织样本。为了使得不同基因的表达信息清晰明了，我们去除了所有能匹配到多个基因上的探针。我们用多个探针的平均基因表达信号强度（signal intensity: S）来度量单个基因的表达水平。在本研究中，我们分析了三个基因表达统计量：基因表达水平，最高表达强度和组织特异性。基因表达水平是每个基因在84个组织中表达水平的平均值。最高表达强度（highest expression signal; S_{max}）是指每个基因在表达量最高的组织中的表达水平。我们用组织特异性指数（tissues specificity index; τ ）[39] 来衡量基因的组织表达特异性。组织特异性指数衡量的是该基因的表达水平在所有组织中的的不均一性或表达水平的杂合性。组织特异性指数的值分布在0到1之间，越高代表组织特异性越高。其计算方法为：

$$\tau = \frac{\sum_{j=1}^{n}\left(1-\frac{\log_2 S(j)}{\log_2 S_{\max}}\right)}{n-1}$$

其中 n 代表的是统计的组织个数。$S(j)$代表第 j 个组织的表达水平。我们最后对

15，246 个基因的表达数据进行了系统的分析。

用DAVID对基因进行GO（Gene ontology)注释

为了更好的理解具有某些特征的基因列表背后的生物学意义，DAVID 数据库（注释，形象化，系统数据挖掘数据库） [40,41] 提供了一系列基因功能注释工具。我们用 DAVID 对双联基因，孟德尔遗传病基因，复杂疾病基因和必需基因进行 GO 注释和富集分析，试图找到每类基因的功能特征。

人类基因的结构和进化分析

人类基因的结构，每对同源基因在人类和其它物种之间的同义突变率（synomymous; d_S)及非同义突变率（non-synonymous ； d_N） 都从 Endembl 数据库获得[42,43]。我们获得基因结构注释包括染色体位置，编码序列长度（coding sequence; CDS)，转录物的种类，5’ UTR 起始，5‘UTR 终止，3’ UTR 起始， 3’ UTR 终止等。我们检索的全基因组同义突变率和非同义突变率包括人与黑猩猩，人与猩猩，人与恒河猴，人与小鼠，人与牛，人与熊猫，人与猪，人与兔子，及人与大鼠。这些基因的同义突变率和非同义突变率是采用最大似然法的 PAML[44]对编码区序列比对得到的。我们用 DAVID 提供的编码转换工具把 Ensembl ID [40,41]转换为正式的基因名（official gene symbol)。

对于一个基因在两个比较的物种间存在多个相似度很高的可能同源基因，我

们取其中相似度最高的一组作为同源基因。同时，当一组基因的同义突变率为零，我们将在以后的分析中移除该同源基因。最后，我们获得了人-黑猩猩同源基因 17044 个，人-猩猩同源基因 16250 个，人-恒河猴同源基因 16424 个，人-小鼠同源基因 16137 个，人-牛同源基因 15802 个，人-熊猫同源基因 16084，人-猪同源基因 13065 个，人-兔同源基因 14865 个和人-大鼠同源基因 15，576 个。我们对同源基因相似度大于 80%的基因对进行进化分析，同时也对相似度大约 90%的做相似的分析。

人类多态性数据分析

我们用HapMap的单核苷酸多态性数据(公共数据释放第#27次) [45,46] 做群体差异分析。我们首先去除这每个群体中有近亲血缘关系的个体（根据其样本信息和IBD>0.2)，然后取他们共有的常染色体位点。最后，我们得到113个YRI无关个体，113个CEU无关个体，和84个CHB无关个体代表欧亚非三大洲人群，每个个体有3,836,272常染色体多态性。我们利用Weir和Cockerham提出的公式 [47]计算了三大人群之间的无偏群体差异系数。

随后，我们利用单体型综合得分（integrated haplotype score; iHS）[48]来检测人群最近受到的自然选择。我们从HapMap网站（第22次释放）[46]获得定相的CEU基因分型数据，并从UCSC网站 [49]获得每个单核苷酸多态性位点的祖先信息。因为X染色体经历的群体历史和常染色体不一样，且只有较小的样本量，我们在这里只考虑常染色体。我们把每个位点依次当做核心位点，计算每个低频等位基

因（MAF)大于5%的多态性位点的的iHS。同时只考虑包含SNP个数大于5的基因，我们把每个基因内|iHS|>2的SNP占基因内所有SNP的比列当做这个基因的iHS。除此之外，我们分析了Applera数据[21]中对欧洲人重测序的8427个基因。计算了每个基因的Tajimm’s D和Fu和Li’s D*。

拷贝数变异（Copy number variations; CNVs）和拷贝数变异图谱

全基因组重测序研究表明基因组上大多数的差异来自于长度大于1kb的拷贝数变异[50]，这些拷贝数变异可以通过个体之间和群体之间的比较发现[51,52]。拷贝数变异影响到很多生物过程比如基因表达，人类进化和人类疾病。在本研究中，我们分析Conrad等 [52]检测到的11700拷贝数变异（其中8599已被试验验证）在来检测各类基因在拷贝数变异区的富集程度。

结果

孟德尔遗传病连锁基因更有可能导致复杂疾病

我们首先分析了 hOMIM 数据库[25]中的 968 个孟德尔遗传病连锁基因，这些疾病基因都是很典型的单突变致病。遗传关联数据库（genetic association database; GAD) [30,31]中有 3117 个复杂疾病关联基因。我们发现这 968 个孟德尔遗传病连锁基因中有 524 是复杂疾病关联基因，这比预期的重叠基因数目高出 8.2 倍(Figure 1; $P < 2.2\times10^{-16}$, x^2检验)。大量的孟德尔遗传病连锁基因和复杂

疾病关联基因重叠意味着孟德尔遗传病基因容易导致复杂疾病，或复杂疾病基因的突变更容易导致孟德尔遗传病。我们把这 524 个在孟德尔遗传病和复杂疾病中同时起作用的基因称为双联基因（MC genes)。

当我们用其它数据库做相同分析时，我们发现双联基因的数目总是比预期的多。特别是，我们分析了全基因组关联研究数据库（GWAS catlog) [10]中的 1799 个复杂疾病关联基因，这些基因都是用位点数大于 10K 的芯片做关联分析发现的。我们也分析了 OMIM 数据库 [8,9]中只有阳性报道的基因，这包括 3261 个在两个以上的研究中重复发现的关联基因。

双联基因平均关联的疾病和表型更多

因为每个双联基因至少关联两种以上的疾病（一个孟德尔遗传病和一个复杂疾病），我们随后分析每个双联基因在 hOMIM 或 GAD 中关联的疾病和表型个数是否与其它类型的基因有所不同。 首先，我们把 968 个孟德尔遗传病连锁基因分为 524 个双联基因和 444 孟德尔遗传病基因。我们发现每个双联基因平均关联 1.77 ± 0.059 (均值 ± 标准误)种孟德尔遗传病，显著性的比每个孟德尔遗传病基因关联的疾病要多(1.35 ± 0.043, 均值 ± 标准误; P = 2.8×10^{-6}, Wilcoxon test)。 我们进一步把全基因组的 3117 个复杂疾病关联基因分为 524 个双联基因和 2593 个复杂疾病基因。我们发现每个双联基因平均关联 5.83 ± 0.50(均值 ± 标准误)种复杂疾病。显著性的比复杂疾病基因平均关联的的复杂疾病多(2.89 ± 0.14, 均值 ± 标准误; P = 2.8×10^{-6}, Wilcoxon test)。总之，平均每个双

联基因关联了 5.83 种复杂疾病和 1.77 种孟德尔遗传病，比孟德尔遗传病基因和复杂疾病基因关联的疾病都要多。当我们对每个基因关联的表型性状进行分析，我们同样发现每个双联基因关联的表型比其它种类的基因多。

双联基因平均互作蛋白数比其它类型疾病基因更多

每个双联基因关联了更多的疾病和表型，它们是不是参与了更加复杂的蛋白调控网络？我们根据人类蛋白参照数据库(HPRD) [35]中的蛋白相互作用关系构建了人类的蛋白相互作用网络。我们发现与双联基因相互作用的蛋白数目（12.00 ± 1.02；均值 ± 标准误）比与孟德尔遗传病基因($P = 3.43\times10^{-7}$, Wilcoxon test)和非疾病基因（non-disease genes, ND genes; $P < 2.2\times10^{-16}$, Wilcoxon test）相互作用的蛋白数目显著增多 (Figure 2)。并且也比复杂疾病基因作用的蛋白多（$P = 0.058$, Wilcoxon test）。同时，与复杂疾病基因相互作用的蛋白数目（10.96）比与孟德尔遗传病基因相互作用的蛋白数目显著多 ($P = 9.57\times10^{-6}$, Wilcoxon test)，这与之前的研究报道复杂疾病基因比孟德尔遗传病基因参与更加复杂的网络基本一致[26,53]。

我们进一步把非疾病基因分为必需基因和其它基因。我们发现与必需基因相互作用的蛋白数目是所有基因类别中最多的(16.00 ± 0.67；均值 ± 标准误)，而与其它基因相互作用的蛋白数目是所有基因类别中最少的(5.74 ± 0.14；均值 ± 标准误)，其他几类基因处于他们中间，依次是双联基因(12.00)，复杂疾病基因 (10.96) 和孟德尔遗传病基因 (7.40) (Figure 2)。并且每类基因相互作用的

蛋白数目的中位数的与其平均数的顺序是一直的，其中位数依次是必需基因（8），双联基因(6)，复杂疾病基因(5)，孟德尔遗传病基因 (4)和其它基因(3)。这些结果和之前的研究报道基本上是一致[37,53]，但是和最近 Podder 和 Ghosh [26]采用 83 个复杂疾病基因所作的研究结果不一直。在他们的研究中，他们发现复杂疾病基因比看家基因（必需基因）参与了更复杂的蛋白网络。

在酵母中，必需基因一般会编码参与复杂网络的蛋白，这些基因更像蛋白网络的中心[54]。对于人类疾病基因是否倾向于编码网络中心蛋白仍然有争议[37]。我们的研究表明网络中心蛋白在疾病基因中的比列比其它基因高，但是比必需基因低。同时，双联基因和复杂疾病基因比孟德尔遗传病基因包含有更多的网络中心蛋白。当我们用酵母双杂交数据库[37]和复合数据库对数据进行分析，发现观察到的趋势与前面的观察基本一致。

双联基因编码更长的蛋白

基因结构的特征曾经被用来预测疾病基因[19,23]。我们在这里对这五类基因的编码蛋白长度（CDS)，基因本身长度，转录物的种类数，3’-UTR 长度和 5’-UTR 长度进行分析. 我们发现每一类疾病基因的编码蛋白长度都显著性的比其它基因长($P < 3.03\times10^{-12}$, Wilcoxon test) (Figure 3A, 3B)，这进一步支持了疾病基因编码比非疾病基因长的蛋白[19,23]。进一步分析发现双联基因编码的蛋白长度(2,396 ± 237nt; 均值 ± 标准误)在所有基因类型中最长，同时也比必需基因编码的蛋白长($P = 0.0084$, Wilcoxon test)。随后我们根据每个基因在人类参

考基因组（human reference genome)上的物理位置（从开始到终止）计算其长度。我们发现每一类疾病基因的平均长度都比其它基因的长度长（$P < 9.64 \times 10^{-7}$, Wilcoxon test）(Figure 3C, 3D)。但是，复杂疾病基因的平均长度在所有类型的基因中是最长的，这意味着双联基因的紧凑性（compactness)比复杂疾病基因要高。双联基因的平均长度(80,476nt ± 6,906nt; 均值 ± 标准误)并不比必需基因显著性的长，但是要比孟德尔遗传病基因和其它基因显著的长(Figure 3C, 3D)。

当一个基因有多种异构体时，我们在上面的分析中随机地选择了一个异构体。当我们只选用每个基因最长的异构体，或去除所有有异构体基因重复上面的分析后，我们看到趋势基本不变。进一步分析发现每一类疾病基因的异构体数目都显著性的比其它基因多(P < 1.23×10^{-10}, Wilcoxon test)，这可能能够解释为什么很多改变选择性剪切的突变都是致病性的[55]。更有趣的是，双联基因的的平均异构体数目（7.62 ± 0.30; 均值 ± 标准误）在所有类别中是最多。必需基因的5’UTR 长度(229nt ± 6nt; 均值 ± 标准误）在所有类别中是最长的(P < 7.4×10^{-5}, Wilcoxon test)。而双联基因的 5’UTR 长度除了和必需基因有显著差别，和其它类型的基因没有任何差异。双联基因的 3’UTR 长度 440nt ± 216nt; 均值 ± 标准误)比其它任何类型的基因都短，如双联基因的长度比孟德尔遗传病基因的长度(1158nt ± 353nt; 均值 ± 标准误）显著性的短 (P=0.045, Wilcoxon test)。

双联基因的组织特性最高

基因表达的差异是可能是引起复杂疾病的一个重要机制[56, 57]。虽然有很多研究对健康和病例组织的基因表达模式进行了分析[57, 58]，还没有研究系统分析疾病基因在各种正常组织中的表达模式。我们在这个研究用 44776 个探针对 84 个人类组织进行分析。我们的分析表明三类疾病基因和必需基因的表达水平都显著性的比其它基因的表达水平高(P < 7.1 $\times 10^{-9}$, Wilcoxon test)，但这四类基因之间的表达水平并没有显著性差异(Figure 4A)。S_{max} 代表一个基因在所有组织中表达量最高的那个值。分析表明双联基因的平均 S_{max} 值(1568S ± 177S; 均值 ± 标准误)在所有基因类型中是最高的(Figure 4B)，甚至比其它基因的 S_{max} 值高出 4 倍。

我们用组织特意系数τ（tissues specificity index） [39] 来衡量基因表达的组织特异性，τ的取值范围在 0 和 1 之间。我们的分析表明三类疾病基因的组织特异性显著性的比必需基因和其它基因的组织特异性高(P < 1.3×10^{-5}, Wilcoxon test; Figure 4C, 4D)，这支持之前报道的差异性表达基因更有可能和疾病相关联[59, 60]。虽然双联基因的表达水平和必需基因基本一致，但是双联基因的组织特异性(0.57τ ± 0.0093; 均值 ± 标准误）比任何其它类别基因的组织特异性都显著性的高(P < 3.5 $\times 10^{-5}$, Wilcoxon test; Figure 4C, 4D)。孟德尔遗传病基因的组织特异性和复杂疾病基因的组织特异性的非常相似，它们的值都显著性的比双联基因的组织特异性低，同时比必需基因和其它基因的组织特异性高(P < 1.3×10^{-5}, Wilcoxon test)。双联基因和复杂疾病基因有很高的组织特异性和异构体，这可能意味着复杂疾病基因的表达需要更复杂的调控，这样我们能够解

释为什么很多基因组关联分析发现的单核苷酸多态性位点是数量表达性状（expression quantitative trait loci; eQTLs）[61]。

当我们对 Tu 等[17]报道的 1534 个看家基因（housekeeping genes）进行分析，我们发现这些看家基因的平均基因表达水平和组织特异性在所有基因类型中都是最高的(Figure 5)，这与之前的报道基本一致[20,62]。虽然很多人认为看家基因和必需基因有相似的特征并且很多研究使用看家基因代替了必需基因[17,33]，但是看家基因和必需基因和的表达水平和组织特异性是完全不一样的。

不同类型基因的群体差异

已经有一些研究对复杂疾病相关联的多态性位点的群体差异进行过一些分析[63,64]。我们利用 HapMap 中的 3 大洲人群（YRI, CEU, CHB）的单核苷酸多态性数据[45,46]计算各类基因的群体差异，即根据每个基因中的单核苷酸多态性位点计算出基因的群体差异。每一类基因的群体差异 F_{ST} 的基本信息已经被表示出来（Table 1）。其它基因的群体差异比必需基因的群体差异(P = 0.027, Wilcoxon test)和双联基因的群体差异(P = 0.022, Wilcoxon test) 显著性的高，虽然它们数值上的差异不是那么明显。这与之前报道的复杂疾病关联多态位点的群体差异不比其它位点的群体差异高的结果是一致的[63,64]。因为有群体特异性自然然选择，曾经有假说认为个体的民族来源可以作为疾病风险性的预测[65]。因为每类疾病基因的群体差异性都不比其它基因高，因此我们认为：在多数情况下民族来源不是疾病预测的一个好的指标。

不同类型基因受到的自然选择

因为孟德尔遗传病基因和必需基因上的很多突变都是致死或致残的，所以这些基因受到很强的负向选择（purifying natural selection）[19,66,67]。但是复杂疾病基因受到的自然选择的类型和强弱现在还没有很明确的定论[68,69]。Blekhman 等 [25] 比较分析了多种基因类型并发现复杂疾病基因在他们分析的基因类型里最不保守，因此他们认为很多复杂疾病基因可能受到最近的正向自然选择。在本研究里，我们使用两类完全不同的统计量对每个基因受到的自然选择进行分析。首先，我们利用 HapMap 数据计算每个基因的单体型综合得分（iHS）[48]来推测每个基因受到的正向自然选择。然后，我们分析重测序数据计算其 Tajima’s D 并估算其受到的总的自然选择模式。

单体型综合得分[48]根据同一个位点受到正向自然选择的等位基因相对于另一个等位基因表现出来的延长的连锁不平衡（linkage disequilibrium; LD）来寻找自然选择的信号。我们发现双联基因和复杂疾病基因的平均单体型综合得分非常相近（iHS 平均值分别为 0.075, 0.077），它们的值比其它三类基因的值都显著性的高（Figure 6A; $P < 0.02$, Wilcoxon test）。对必需基因和双联基因的 iHS 值作 Q-Q 图表明基本所有的 iHS 值都位于双联基因一侧(Figure 6B)，表明双联基因相对于必需基因受到更强的正向自然选择。在 Q-Q 图右端，实际观察值与期望值偏离更远，这意味着这部分双联基因相对于必需基因受到特别强的正向自然选择。复杂疾病基因受到的正向自然选择和双联基因很相似(Figure 6C)。

除此之外，我们还计算了重测序数据 Applera 中每个基因 Tajima’s D[70]

的值。虽然 Tajima’s D 只是区分基因序列是否是中性进化的，因为一个群体内每个基因有着相似的历史，极端的 Tajima’s D 值就可能是自然选择留下的印迹。我们发现双联基因和复杂疾病基因 Tajima’s D 非常相似（Figure 6D)。除了必需基因和复杂疾病基因之间的差异显著外（P = 0.029, Wilcoxon test），所有其它任意两类基因之间都没有显著差异(Figure 6D)。这些分析基本上支持一些复杂疾病基因更有可能受到正向自然选择 [25].。但是我们更多的基因表明复杂疾病基因受到的自然选择没有 Blekhman 等[25]报道的那么明显。

这些根据 iHS 和 Tajima’s D 做的分析表明双联基因和复杂疾病基因在近期经历了非常相似的进化历史。虽然它们相对于必需基因和孟德尔遗传病基因受到较强的正向自然选择，只有很小的一部分受到很强的正向自然选择。我们也发现双联基因和复杂疾病基因中富集了更多之前报的的受到正向自然选择的基因（Figure 7)。

不同时间尺度上相对进化率的变化

阐明疾病基因的进化有利于我们认识和理解人类疾病的起源和进化[16,26]，为了这个目标，已经有很多研究对疾病基因和非疾病基因做了详细的比较性研究[17-21,29]。近期也有多个研究对孟德尔遗传病基因和复杂疾病基因进行了比较[24-26]。 虽然之前的研究已经知道疾病基因没有必需基因/看家基因那么保守，但是在其它方面还有很多争议 [17,18,22-26]。我们推测各种争议可能是由于不同的研究使用不同的数据导致的，特别是不同研究使用的数据的进化尺度不一样。我们对两个大的时间尺度上的 9 种哺乳动物与人的同源基因经行比较分析。

我们首先分析了人类与 3 个其它灵长类动物全基因组同源基因的进化速率（evolutionary rate; d_N/d_S）。在人-黑猩猩同源基因中，我们发现必需基因的进化速率是最慢的，即必需基因最保守，而其它基因的进化速率是最快(Figure 8A)。每一类疾病基因的进化速率都显著性的比其它基因的进化速率慢（$P < 0.016$; Wilcoxon test），同时也显著性的比必需基因的进化速率快（$P < 1.93\times10^{-5}$; Wilcoxon test），这表明疾病基因比其它基因保守但比必需基因进化速度快。进一步分析表明双联基因的进化速率（0.26 ± 0.013; mean ± SEM）在三类疾病中的进化速度最慢，说明双联基因在所有疾病基因中最保守。最有趣的是，在每一类基因中都存在进化速率 $d_N/d_S > 1$ 的基因（Figure 8B），这表明人类和黑猩猩分离以来每一类基因中都有一些基因受到正向自然选择。在所有基因类别中，双联基因和必需基因包含的进化速度 $d_N/d_S < 1$ 的基因的比例是最高的，这说明这两类基因包含的受到负向选择的基因比例是最高的。当我们对人-猩猩，或人-恒河猴之间的差异进行研究，虽然他们的进化速率的绝对值是不一致的，但是我们观察到的差异模式基本不变。

我们在更大的时间尺度上比较分析了人类与 6 个非灵长类哺乳动物之间的分子进化速率。对人类与小鼠的同源基因比较分析发现必需基因的平均分子进化速率仍然是最低的，即最保守的 (Figure 8C)。在这个时间尺度上，我们发现不是每一种疾病基因都比其它基因保守（比如孟德尔遗传病基因比其它基因的进化速度还快），这与人类-灵长类比较分析观察到的结果不一致。在所有基因中，没有任何基因的分子进化速率大于 1（Figure 8D），这表明从长期来看所有基因都受到负向选择，对于一个基因来说，其受到的正向自然选择只是临时性的。当我们分析了

其它非灵长类哺乳动物如大鼠，牛，猪，兔子和熊猫与人类的同源基因后，发现它们与小鼠表现出很类似的特征。当我们只对非同义突变的突变率进行分析，我们发现其结果和分子进化速率的结果一致。我们只对同源性大于 90%的同源基因进行分析，结果也与之前的观察一致。

我们提出一个假说来解释我们观察到的相对分子进化速率随时间改变而改变的现象。人类疾病基因可能在所有灵长类动物中起着重要的作用，这使得它们相对于其它基因更保守。但是在与人类差异较大的非灵长类哺乳动物中，有些人类疾病基因的同源基因功能已经不同于其在人体内的功能或者功能部分消失，从而使得这些基因的保守性降低，甚至低于其它基因。简言之，一些导致人类疾病的基因可能在非灵长类哺乳动物中发挥着不同于其在人体内的作用。这样，通过研究模式生物如小鼠和大鼠来推测人类疾病基因的功能和参与的信号通路就要特别小心，因为一些在模式生物中得出的结论有可能对人类不适用。 我们也发现一些基因在在灵长类动物的进化中受到明显的正向自然选择，而人与非灵长类哺乳动物的同源基因中没有发现受到正向自然选择的基因。这表明作用于某个基因的正向自然选择只是临时的，从长期的进化来看，只有纯化选择是长期的，持续的。

不同类型基因与拷贝数变异的关系

拷贝数变异能够影响基因功能，当拷贝数变异区域与基因区域重叠时这种效果更明显[52, 71]。另一方面，拷贝数变异的命运如拷贝数等位基因频率的变化也受到其所处的基因的功能的影响。例如，当拷贝数变异使得基因功能丧失以至于

导致个体死亡时，该拷贝数一般不能够在群体内扩散和维持。并且已经有报道发现落于基因区的拷贝数变异数目比落于非基因区的数目显著性的相对减少[52]。为了研究各类基因分布和拷贝数变异分布的相对关系，我们对每一类基因包含的拷贝数变异的富集程度进行分析。我们发现位于复杂疾病基因区的拷贝数变异相对数目比其它基因区域显著性的增多（$P = 5.16\times10^{-6}$, Fisher's exact test; Table 2），意味着拷贝数变异更有可能导致复杂疾病。当只考虑常见删除性变异时，这种富集更加明显（$P = 2.59\times10^{-16}$, Fisher's exact test）（Table 2）。与此相反，位于必需基因区域的拷贝数变异相对数目比其它基因区域显著性的少（$P = 8.15\times10^{-4}$, Fisher's exact test）（Table 2）。同时，位于孟德尔遗传病基因区域的拷贝数变异数目也比其它基因区域显著性的减少 （$P = 0.025$, Fisher's exact test），多等位基因区域的相对数目少的更加明显（$P = 4.26\times10^{-4}$, Fisher's exact test）。位于双联基因区域的拷贝数变异即不显著增多，也不显著减少。当我们对 Conrad 等[52]检测出的所有拷贝数变异进行分析得出相似的结果。

拷贝数变异的突变率对每一类基因应该是一样的，或者应该没有显著区别。但为什么必需基因和孟德尔遗传病基因区显著性的缺少拷贝数变异而复杂疾病基因区域显著性的富含拷贝数变异？一个可能的解释是发生在必需基因和孟德尔遗传病基因区的拷贝数变异突变可能是致死性的或严重影响携带者。这种很强的纯化选择可能导致这些拷贝数变异很快消失，从而导致这些基因区域的拷贝数变异相对数目很少。但是，很多发生在复杂疾病区域的拷贝数变异突变可能不会直接导致很严重后果，甚至在一些时候带来某种优势---即受到正向自然选择。这样最

终导致了拷贝数变异在复杂疾病基因区域富集。既然拷贝数变异更有可能导致复杂疾病，这意味着仔细对与复杂疾病关联的单核苷酸多态性位点附近的拷贝数变异进行分析更有可能找到疾病背后的遗传机制。因为双联基因同时参与了复杂疾病和孟德尔遗传病，它们同时受到正向和负向自然选择，这与通过多态性数据和种间差异检测到的结果基本一致。

不同类型基因的功能特征

利用 DAVID，我们发现疾病相关突变在双联基因，孟德尔遗传病基因和必需基因中显著富集 ($P < 1.65\times10^{-122}$, FDR)。更重要的是，我们发现每一类基因都有自己独特的功能特征。我们首先发现胚胎发育和细胞活动相关的基因在必需基因中富集，比如，在这 1，520 个必需基因中，179 个基因参与了卵孵化和出生以前的胚胎发育 （$P = 5.50\times10^{-81}$, FDR），166 个基因参与了胚胎的形态发生 ($P = 1.92\times10^{-69}$, FDR)，363 个基因参与了 DNA 结合 ($P = 1.46\times10^{-61}$, FDR) 和 469 个基因参与了转录调控 ($P = 4.92\times10^{-31}$, FDR)。双联基因更多的富集了与体液水平和体液动态平衡相关联的基因。比如，117 个基因参与了体液稳态过程（$P = 9.72\times10^{-38}$, FDR），98 个基因参与了化学稳态过程 ($P = 3.44\times10^{-35}$, FDR)，44 个基因参与了体液水平调节 ($P = 2.57\times10^{-24}$, FDR)，49 个参与了血液循环 ($P = 8.73\times10^{-24}$, FDR)。对于孟德尔遗传病基因，它们富集的基因功能类别包括视觉感知（$P = 4.35\times10^{-17}$, FDR），光刺激的感官知觉（$P = 4.35\times10^{-17}$, FDR），外胚层发育（$P = 5.40\times10^{-9}$, FDR）和表皮发育（$P = 2.63\times10^{-8}$, FDR）。复杂疾病基因

富集了参与生理调控和免疫相关的基因如细胞增殖调控（$P = 9.83\times10^{-33}$, FDR），受伤应激反应（$P = 1.11\times10^{-30}$, FDR），防御应答（$P = 3.05\times10^{-30}$, FDR）和免疫反应（$P = 1.89\times10^{-28}$, FDR）。

讨论

在过去的一个世纪里，孟德尔遗传病和复杂疾病被当做人类疾病频谱上的两个极端----即两大类完全不同的疾病。虽然之前的研究发现一些基因与这两类疾病都有关联[4, 13, 15, 72]，但这些基因并没有被系统性的收集起来并作分析。比如，基因 BRCA1，BRCA2, TP53 和 CDKN2A 中的多态性是一些常见癌症的风险因素，同时在这些基因中的突变也导致了某些孟德尔遗传癌症[13]。一些导致孟德尔隐性遗传疾病的等位基因如 *CFTR* 和 *GBA* 处于杂合态时也会增加复杂疾病风险[4]。据我们所知，在该研究之前没有研究系统分析过双联基因的特征。特别是，我们发现把双联基因和其它疾病基因分离开来还帮助我们发现一些疾病基因隐藏的独特特征。比如，如果我们不把双联基因和孟德尔遗传病基因分离开来，我们就不能发现孟德尔遗传病基因落入拷贝数变异区域的数目显著性的减少。我们在这里进一步强调我们发现孟德尔遗传病基因更可能是复杂疾病的风险因数，这提供了很强的证据说明我们应该选择孟德尔遗传病基因作为关联分析的候选基因。

多角度的分析已经表明双联基因有很多独特的特征。我们在下面对双联基因的这些特征之间的关系进行讨论，以便获得对这些特征更全面的认识。首先，双联基因比孟德尔遗传病基因和复杂疾病基因平均关联更多的表型和疾病。我们对所有基

因的蛋白-蛋白相互作用网络进行分析，发现双联基因平均的作用蛋白数目比孟德尔遗传病基因和复杂疾病基因都多，即双联基因参与了更复杂的蛋白作用网络。这种现象可能说明与较多蛋白相互作用的基因可能参与和调控了更多的表型和疾病。其次，在蛋白-蛋白相互作用中与必需基因直接相互作用的蛋白数量最多，同时必需基因在五类基因中最保守。这支持基因的相互作用蛋白数目越多越保守的假说[73, 74]。第三， 在五类基因中，双联基因编码的蛋白长度最长，其转录物种类也是最多的，这可能能够说明为什么它们比其它两类疾病基因连接更多的蛋白。虽然双联基因的表达水平与孟德尔遗传病基因和复杂疾病基因的表达水平并没有显著不同，但是双联基因的组织特异比任何其它类型的基因都高。除此之外，我们发现每一类基因都有一些独特的功能特征。简言之，双联基因编码最长的蛋白，有最高的组织特异性和最高的转录物种类，参与了比较复杂的蛋白-蛋白相互作用网络并与多种疾病相关联。

毫无疑问，人类对环境因数如病原体，气候和饮食的适应对于人类的繁衍起了重要作用[75, 76]。在本研究中，我们不仅利用人类多态性数据揭示人类最近受到的自然选择，同时也用种间差异数据分析了可能发生在人类祖先中的自然选择。我们对于发生在种间和种内两个时间尺度上的自然选择的一致性特别感兴趣。根据欧洲人群的单体型数据来分析现代人群最近几万年受到自然选择，我们发现双联基因和复杂疾病基因相对于孟德尔遗传病基因和必需基因受到更强的正向自然选择[77]。我们知道受到正向自然选择的基因比受到负向自然选择的基因进化快；与基因相互作用的蛋白越多该基因进化越慢[73, 74]。依此推断，必需基因在物种间也是最保守的。双联基因应该比复杂疾病基因和孟德尔遗传病基因都要保守。事实

上，我们确实发现各类基因在灵长类动物间的差异符合我们的预期。总的来说，必需基因在我们研究的任何时间尺度上都是最保守的。孟德尔遗传病基因在人群中受到负向自然选择，其在较短的时间尺度（灵长类动物进化）上是相对保守的。但是在更大尺度上其进化最快或者接近最快的一类基因，这与之前报道的与基因相互作用的蛋白越少其进化速度越快相一致[73, 74]。双联基因和复杂疾病基因受到的自然选择表现出很复杂的模式，这可能是由于其在不同时期特别是近期受到正向自然选择有关。对拷贝数变异的富集分析支持双联基因受到正负两种自然选择。

孟德尔遗传病基因和复杂疾病基因之间有大量的重叠，所以这两类致病基因并没有明显的界限。虽然双联基因表现出一些独特的特征，其中的一些特征可能源于它们在孟德尔遗传病和复杂疾病中的双重作用。因为孟德尔遗传病基因和复杂疾病基因的数量都在快速增长，我们可以预测双联基因的数目会持续增长。一方面，全基因组关联分析总是会对基因功能提出新的见解，这有助于我们重新审视孟德尔遗传病的致病机理。另一方面，通过研究双联基因导致的孟德尔遗传病，我们能够对这些基因在复杂疾病中功能和网络有更加清晰地认识。这样，在孟德尔遗传病和复杂疾病研究中取的的进展在将来的研究中互相促进。简言之，我们从多个方面系统的分析了双联基因的特征，我们在这个比较分析中也对其它 4 类基因的特征有了新的认识。既然双联基因在人类疾病网络中起着这么重要的作用，仔细对这些基因进行分析能够帮助我们理解致病突变和基因调控，有利于我们发展有效地治疗和诊断工具。

表格

Table 1. Comparison of F_{ST} among the five gene categories. 'Sign.Diff' = significantly different; 'NA' = no significant.

	Mean	Median	SEM	Max	Min	Sign.Diff
MC	0.1153	0.1066	0.0026	0.4506	-0.0044	OTHER
Essential	0.1188	0.1083	0.0016	0.6411	-0.0029	OTHER
CNM	0.1196	0.1108	0.0013	0.5919	-0.0090	NS
MNC	0.1225	0.1114	0.0032	0.3976	-0.0005	NS
OTHER	0.1256	0.1120	0.00064	0.8557	-0.0088	MC, essential

Table 2. Enrichment/impoverishment of genes for each gene category in validated CNVs. P-values were calculated using Fisher's exact test based on the number of genes located in CNVs or not. The red implicated significant enrichment and green implicated significant impoverishment. 'Gen' = genotyped; 'Common' = MAF ≥ 10%, 'Rare' = MAF < 10%.

CNV / Genes	All CNVs	Gen.Dups	Gen.Dels	Gen.Multi	Common CNVs	Common Dups	Common Dels	Rare CNVs	Rare Dups	Rare Dels
Essential	8.15×10^{-4}	1.13×10^{-7}	0.722	7.22×10^{-6}	1.27×10^{-4}	9.47×10^{-7}	0.213	0.046	2.31×10^{-5}	0.80
CNM	5.16×10^{-6}	0.965	2.31×10^{-15}	0.339	3.79×10^{-7}	0.91	2.59×10^{-16}	1.04×10^{-4}	0.616	6.63×10^{-9}
MC	0.743	0.223	0.614	0.563	0.599	0.337	0.819	0.934	0.122	0.344
MNC	0.025	0.041	0.358	4.26×10^{-4}	0.021	0.049	0.26	0.241	0.072	0.82

图示

Figure 1. Mendelian disease genes are more likely to be involved in complex diseases. Overall 54.1% of (524 of 968) the Mendelian disease genes are also genetic risk factors for complex diseases. χ^2 test was performed to examine whether Mendelian disease genes were significantly enriched in complex diseases genes.

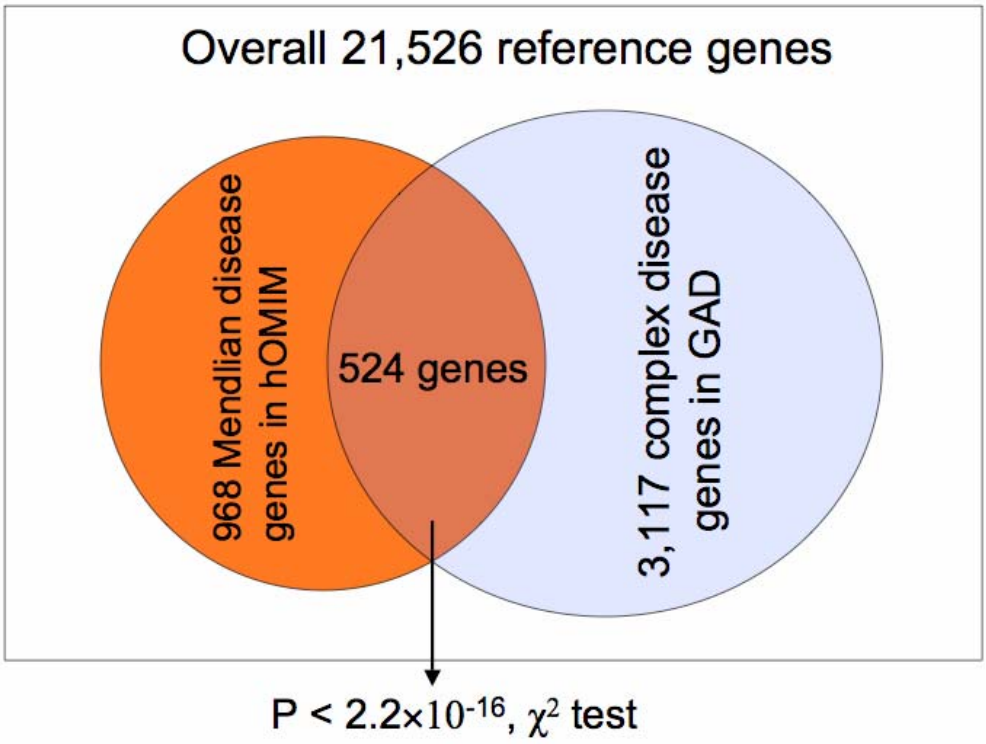


Figure 2. Comparison of connectivity among different gene categories. The connectivity of each gene was calculated based on HPRD network, the protein-protein interaction network constructed from human protein reference database (HPRD). Error bars represent the standard error of the mean.

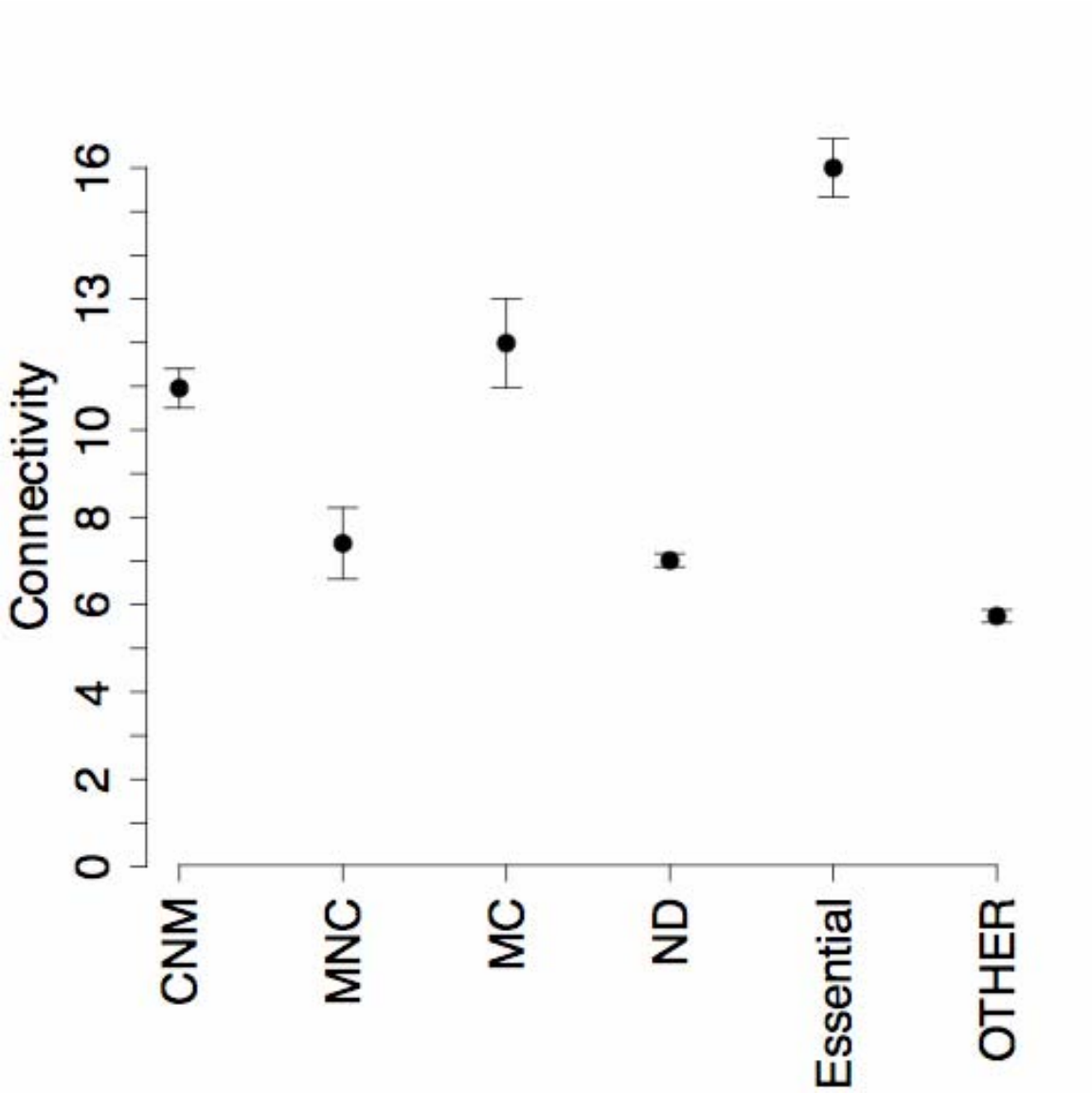


Figure 3. Gene structures among different gene categories. Error bars represent the standard error of the mean. (A) Comparison of CDS length. (B) Cumulative distribution of CDS length. (C) Comparison of gene length. (D) Cumulative distribution of gene length.

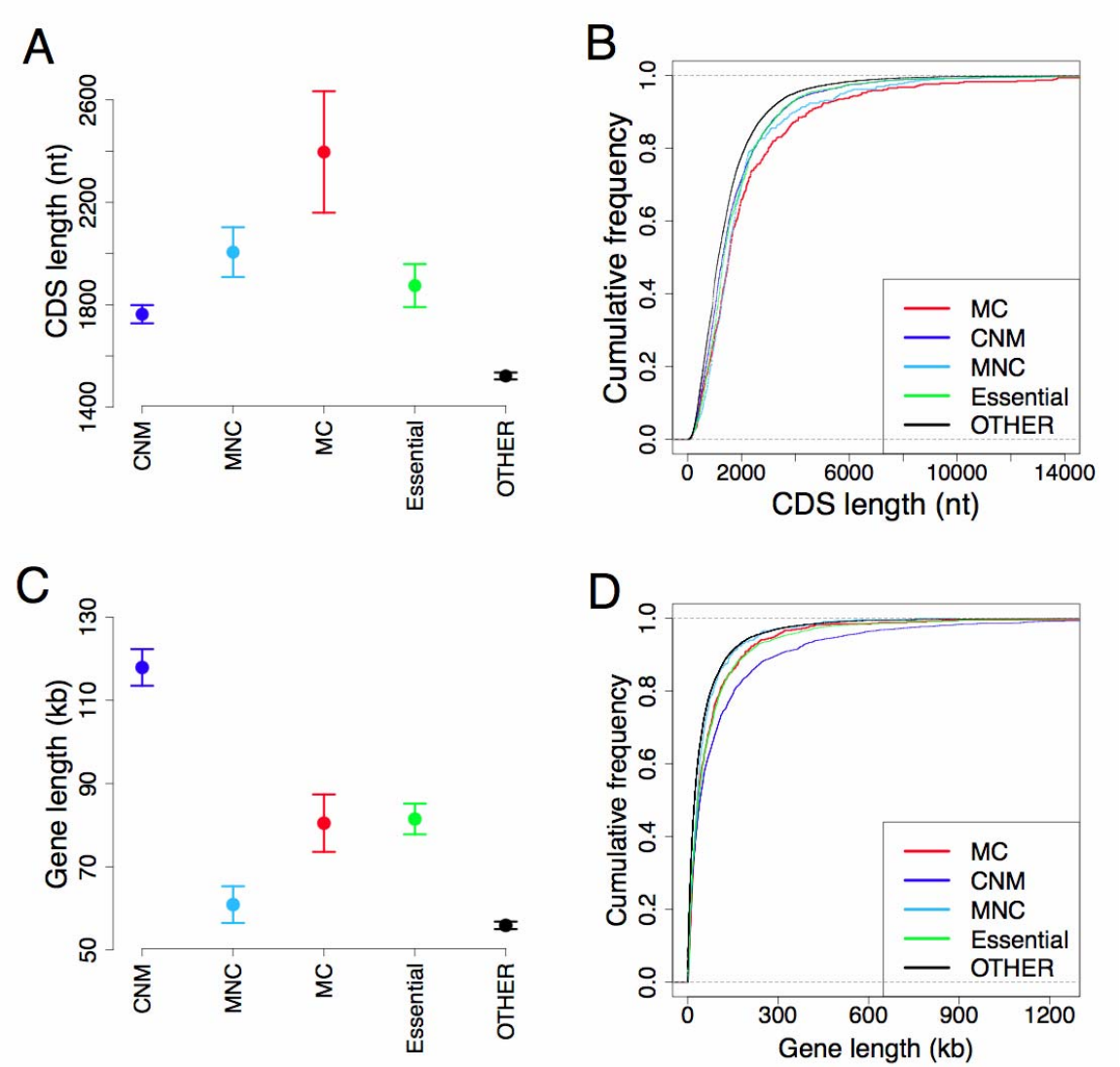


Figure 4. Gene expression patterns showing by different statistics. Error bars represent the standard error of the mean. (A) Comparison of gene expression level. (B) Comparison of S_{max} (the highest expression signal of each gene across all tissues). (C) Comparison of tissue specificity. (D) Cumulative distribution of tissue specificity among different gene categories.

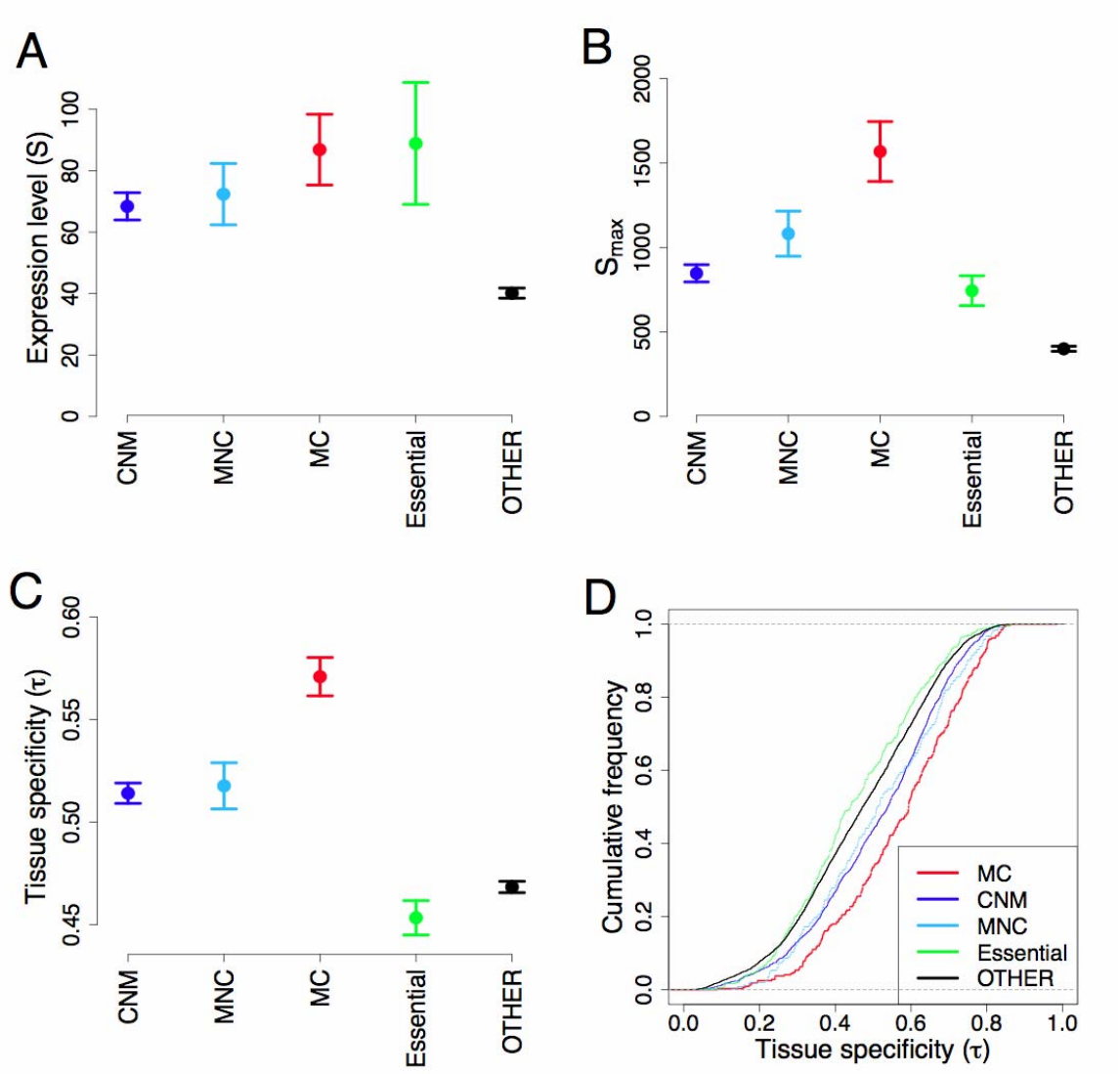


Figure 5. Gene expression patterns including housekeeping genes. Error bars represent the standard error of the mean. (A) Comparing gene expression level. (B) Cumulative distribution of expression level in different gene categories. (C) Comparing tissue specificity. (D) Cumulative distribution of tissue specificity in different gene categories.

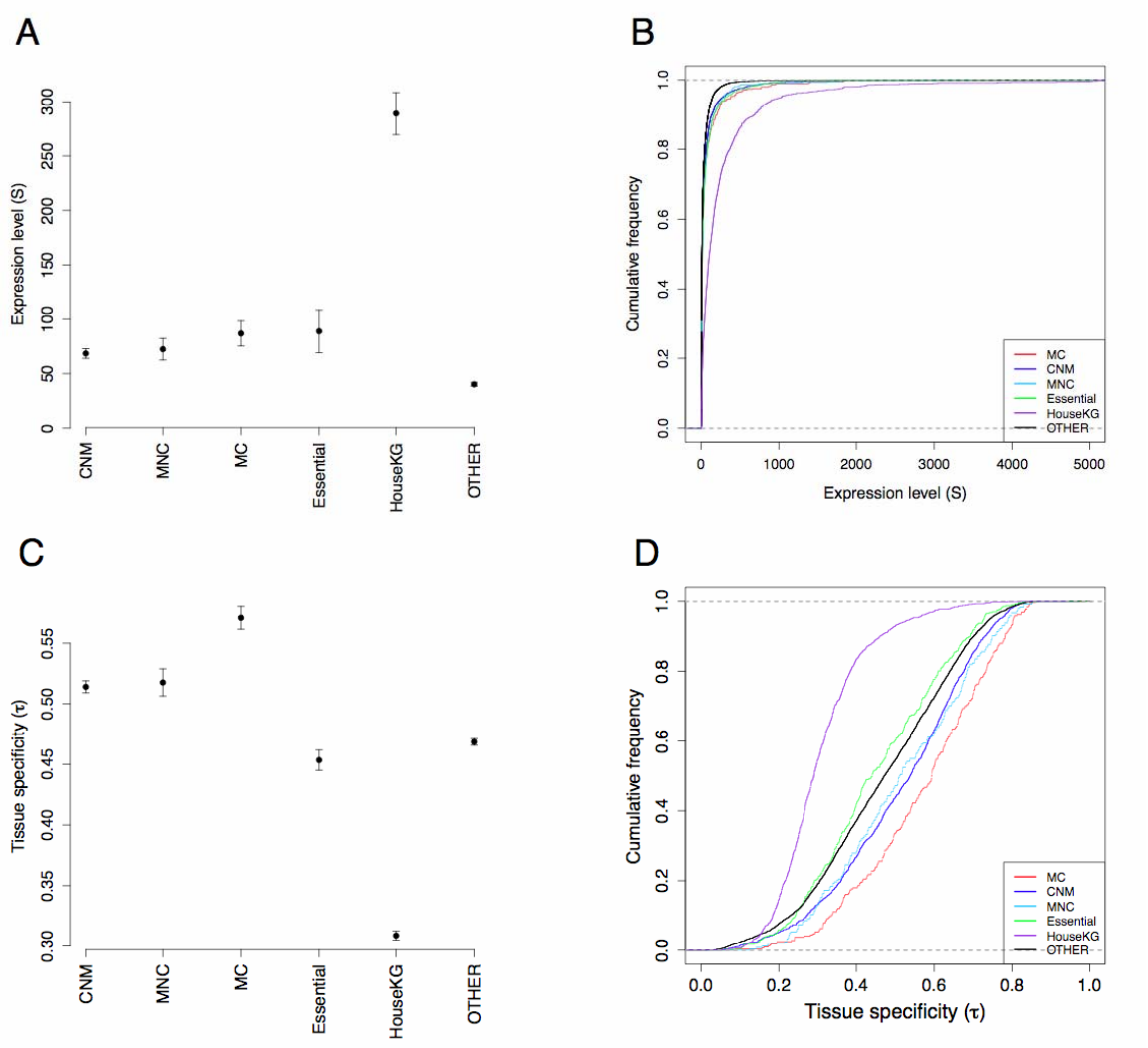


Figure 6. Natural selection estimated from polymorphism data. (A) Comparison of iHS based on CEU in HapMap. (B) Q-Q plot of iHS between MC genes and essential genes. Red line shows null distribution that the values of MC genes are the same as those of essential genes. (C) Q-Q plot of iHS between CNM genes and essential genes. Red line shows null distribution that the values of CNM genes are the same as those of essential genes. (D) Comparison of Tajima's D based on European-American in Applera dataset. Error bars represent the standard error of the mean.

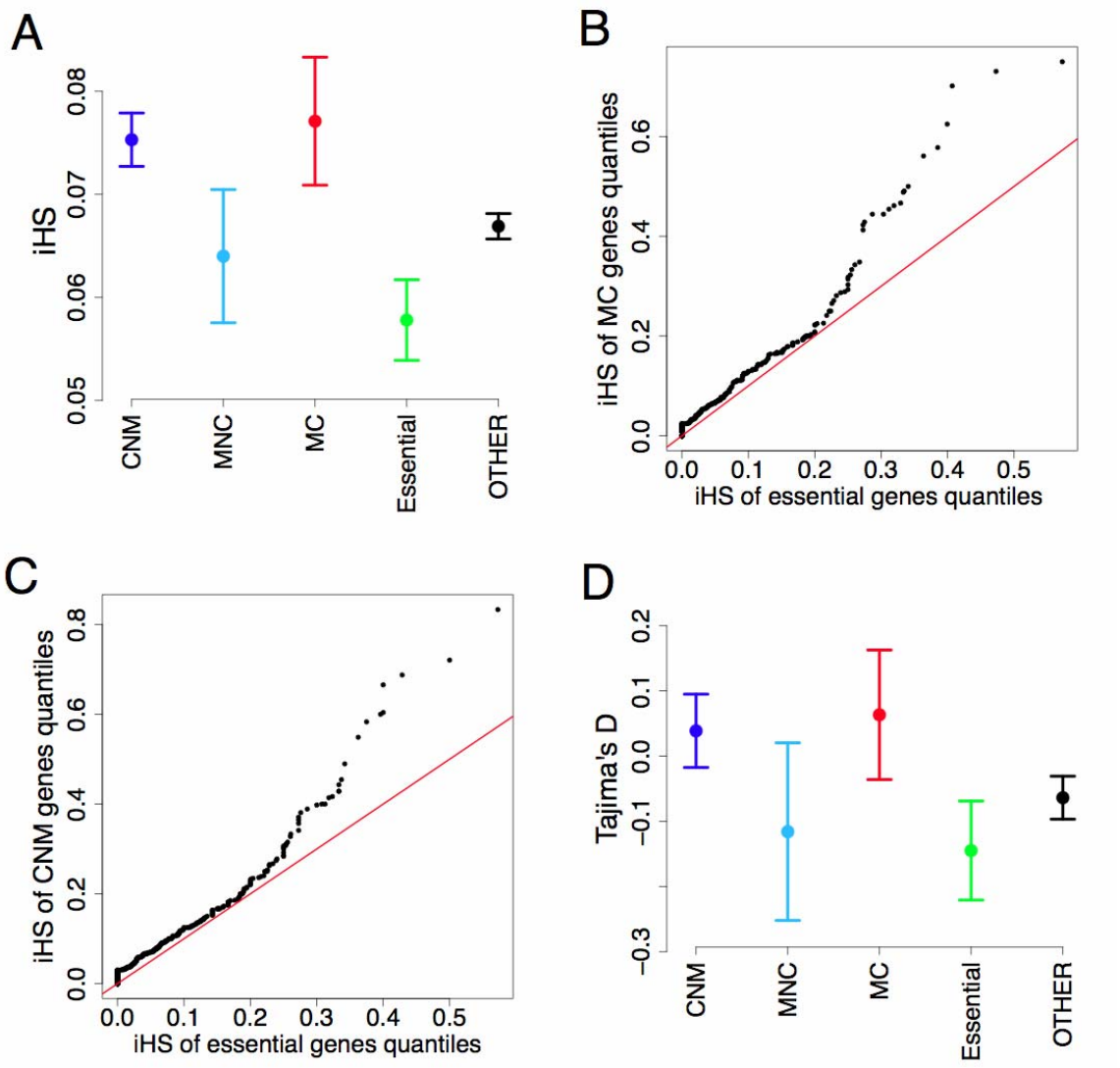


Figure 7. Proportions of positive selection candidates in each gene categories. Positive selection candidates collected and evaluated by Sabeti *et al* (from their Tables S4)[77].

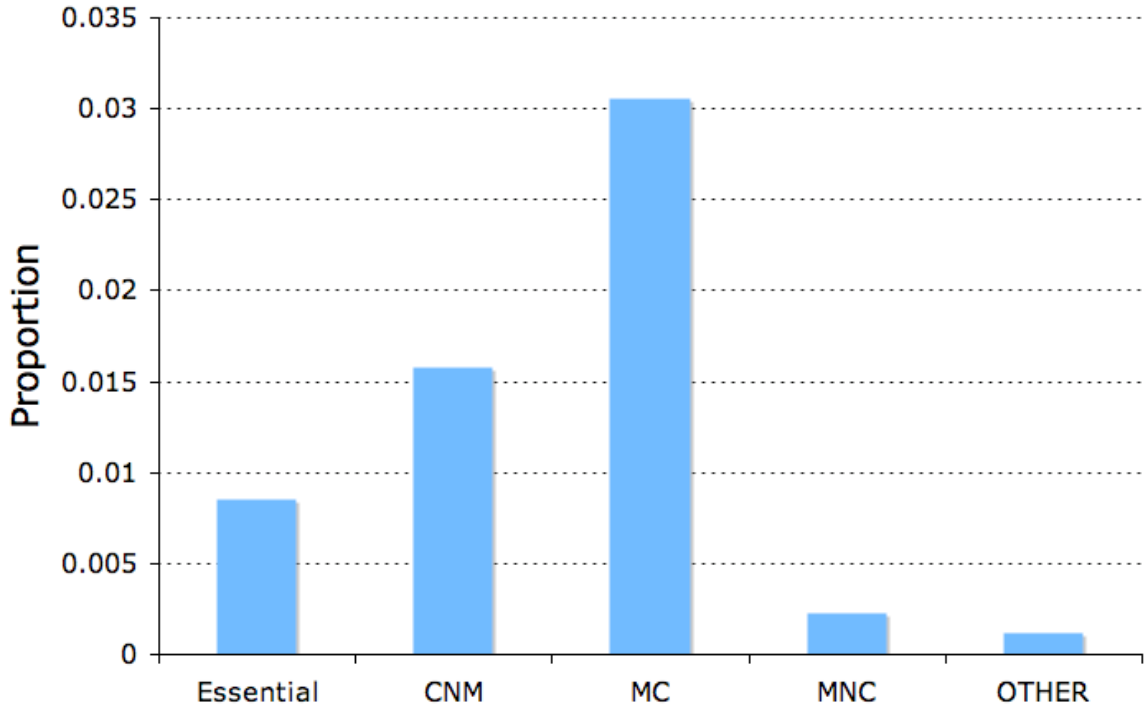
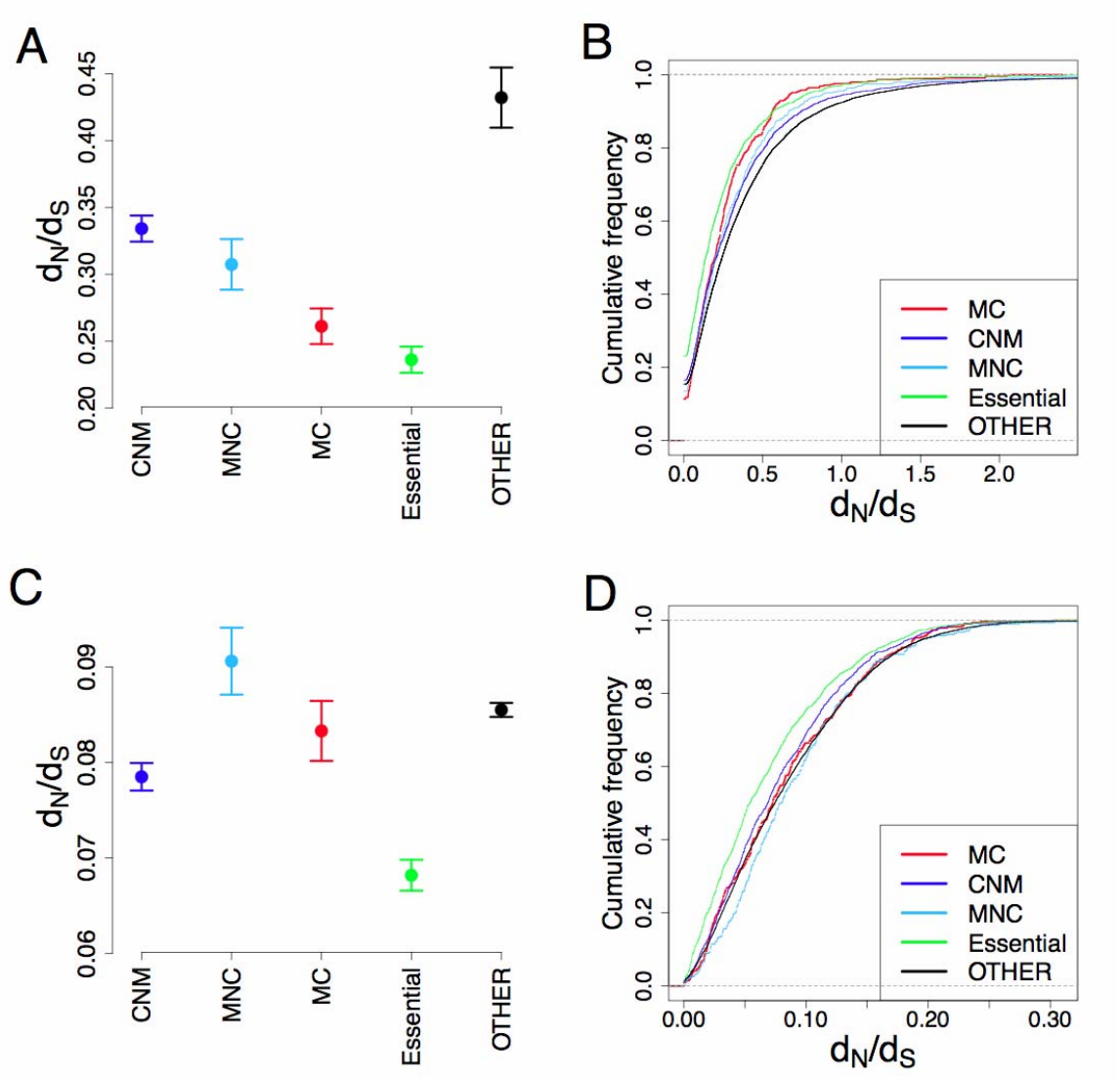


Figure 8. Evolutionary rates of each gene category estimated from divergence data. Error bars represent the standard error of the mean. (A) Comparison of d_N/d_S based on human-chimpanzee orthologous gene pairs. (B) Cumulative distribution of d_N/d_S based on human-chimpanzee orthologous gene pairs. (C) Comparison of d_N/d_S based on human-mouse orthologous gene pairs. (D) Cumulative distribution of d_N/d_S based on human-mouse orthologous gene pairs.



参考文献

1. Davila S, Hibberd ML, Hari Dass R, Wong HE, Sahiratmadja E, et al. (2008) Genetic association and expression studies indicate a role of toll-like receptor 8 in pulmonary tuberculosis. PLoS Genet 4: e1000218.

2. Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, et al. (2007) A whole-genome association study of major determinants for host control of HIV-1. Science 317: 944-947.

3. Antonarakis SE, Chakravarti A, Cohen JC, Hardy J (2010) Mendelian disorders and multifactorial traits: the big divide or one for all? Nat Rev Genet 11: 380-384.

4. Sidransky E (2006) Heterozygosity for a Mendelian disorder as a risk factor for complex disease. Clin Genet 70: 275-282.

5. Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet 33 Suppl: 228-237.

6. Altshuler D, Daly MJ, Lander ES (2008) Genetic Mapping in Human Disease. Science 322: 881-888.

7. Antonarakis SE, Beckmann JS (2006) Mendelian disorders deserve more attention. Nat Rev Genet 7: 277-282.

8. McKusick VA (2007) Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet 80: 588-604.

9. Amberger J, Bocchini CA, Scott AF, Hamosh A (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res 37: D793-796.

10. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362-9367.

11. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics 9: 356-369.

12. Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10: 241-251.

13. Antoniou AC, Chenevix-Trench G (2010) Common genetic variants and cancer risk in Mendelian cancer syndromes. Curr Opin Genet Dev 20: 299-307.

14. Hinney A, Vogel CI, Hebebrand J (2010) From monogenic to polygenic obesity: recent advances. Eur Child Adolesc Psychiatry 19: 297-310.

15. Bultron G, Kacena K, Pearson D, Boxer M, Yang R, et al. (2010) The risk of Parkinson's disease in type 1 Gaucher disease. J Inherit Metab Dis 33: 167-173.

16. Keller MC, Miller G (2006) Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best? Behav Brain Sci 29: 385-404; discussion 405-352.

17. Tu Z, Wang L, Xu M, Zhou X, Chen T, et al. (2006) Further understanding human disease genes by comparing with housekeeping genes and other genes. BMC Genomics 7: 31.

18. Smith NG, Eyre-Walker A (2003) Human disease genes: patterns and predictions. Gene 318: 169-175.

19. Kondrashov FA, Ogurtsov AY, Kondrashov AS (2004) Bioinformatical assay of

human gene morbidity. Nucleic Acids Res 32: 1731-1737.

20. Winter EE, Goodstadt L, Ponting CP (2004) Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. Genome Res 14: 54-61.

21. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, et al. (2005) Natural selection on protein-coding genes in the human genome. Nature 437: 1153-1157.

22. Huang H, Winter EE, Wang H, Weinstock KG, Xing H, et al. (2004) Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes. Genome Biol 5: R47.

23. Lopez-Bigas N, Ouzounis CA (2004) Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res 32: 3108-3114.

24. Thomas PD, Kejariwal A (2004) Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. Proc Natl Acad Sci U S A 101: 15398-15403.

25. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, et al. (2008) Natural selection on genes that underlie human disease susceptibility. Current Biology 18: 883-889.

26. Podder S, Ghosh TC (2010) Exploring the Differences in Evolutionary Rates between Monogenic and Polygenic Disease Genes in Human. Molecular Biology and Evolution 27: 934-941.

27. Cooper DN, Mort M (2010) Do Inherited Disease Genes Have Distinguishing Functional Characteristics? Genetic Testing and Molecular Biomarkers 14: 289-291.

28. Huang N, Lee I, Marcotte EM, Hurles ME (2010) Characterising and predicting haploinsufficiency in the human genome. PLoS Genet 6: e1001154.

29. Cai JJ, Borenstein E, Chen R, Petrov DA (2009) Similarly strong purifying selection acts on human disease genes of all evolutionary ages. Genome Biol Evol 2009: 131-144.

30. Becker KG, Barnes KC, Bright TJ, Wang SA (2004) The genetic association database. Nat Genet 36: 431-432.

31. Zhang Y, De S, Garner JR, Smith K, Wang SA, et al. (2010) Systematic analysis, comparison, and integration of disease based human genetic association data and mouse genetic phenotypic information. BMC Med Genomics 3: 1.

32. Haverty PM, Weng Z, Best NL, Auerbach KR, Hsiao LL, et al. (2002) HugeIndex: a database with visualization tools for high-density oligonucleotide array data from normal human tissues. Nucleic Acids Res 30: 214-217.

33. Eisenberg E, Levanon EY (2003) Human housekeeping genes are compact. Trends in Genetics 19: 362-365.

34. Liao BY, Scott NM, Zhang J (2006) Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol Biol Evol 23: 2072-2080.

35. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. (2009) Human Protein Reference Database--2009 update. Nucleic Acids Res 37: D767-772.

36. Kandasamy K, Keerthikumar S, Goel R, Mathivanan S, Patankar N, et al. (2009) Human Proteinpedia: a unified discovery resource for proteomics research.

Nucleic Acids Res 37: D773-781.

37. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, et al. (2007) The human disease network. Proc Natl Acad Sci U S A 104: 8685-8690.

38. Wu C, Orozco C, Boyer J, Leglise M, Goodale J, et al. (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol 10: R130.

39. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, et al. (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21: 650-659.

40. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4: P3.

41. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.

42. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, et al. (2009) Ensembl 2009. Nucleic Acids Res 37: D690-697.

43. Haider S, Ballester B, Smedley D, Zhang J, Rice P, et al. (2009) BioMart Central Portal--unified access to biological data. Nucleic Acids Res 37: W23-27.

44. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555-556.

45. Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, et al. (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58.

46. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.

47. Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38: 1358-1370.

48. Voight BF, Kudaravalli S, Wen XQ, Pritchard JK (2006) A map of recent positive selection in the human genome. Plos Biology 4: 446-458.

49. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, et al. (2010) The UCSC Genome Browser database: update 2010. Nucleic Acids Res 38: D613-619.

50. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452: 872-876.

51. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, et al. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40: 1166-1174.

52. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, et al. (2010) Origins and functional impact of copy number variation in the human genome. Nature 464: 704-712.

53. Feldman I, Rzhetsky A, Vitkup D (2008) Network properties of genes harboring inherited disease mutations. Proc Natl Acad Sci U S A 105: 4323-4328.

54. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, et al. (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet 38: 285-293.

55. Lopez-Bigas N, Audit B, Ouzounis C, Parra G, Guigo R (2005) Are splicing

mutations the most frequent cause of hereditary disease? Febs Letters 579: 1900-1903.

56. Dermitzakis ET (2008) From gene expression to disease risk. Nat Genet 40: 492-493.

57. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009) Mapping complex disease traits with global gene expression. Nat Rev Genet 10: 184-194.

58. Kilpinen S, Autio R, Ojala K, Iljin K, Bucher E, et al. (2008) Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. Genome Biol 9: R139.

59. Chen R, Morgan AA, Dudley J, Deshpande T, Li L, et al. (2008) FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease. Genome Biol 9: R170.

60. Lage K, Hansen NT, Karlberg EO, Eklund AC, Roque FS, et al. (2008) A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proc Natl Acad Sci U S A 105: 20870-20875.

61. Nicolae DL, Gamazon E, Zhang W, Duan SW, Dolan ME, et al. (2010) Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. Plos Genetics 6: -.

62. Zhang L, Li WH (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol 21: 236-239.

63. Lohmueller KE, Mauney MM, Reich D, Braverman JM (2006) Variants associated with common disease are not unusually differentiated in frequency across populations. American Journal of Human Genetics 78: 130-136.

64. Myles S, Davison D, Barrett J, Stoneking M, Timpson N (2008) Worldwide

population differentiation at disease-associated SNPs. Bmc Medical Genomics 1: -.

65. Bamshad M, Wooding S, Salisbury BA, Stephens JC (2004) Deconstructing the relationship between genetics and race. Nat Rev Genet 5: 598-609.

66. Wilson AC, Carlson SS, White TJ (1977) Biochemical evolution. Annu Rev Biochem 46: 573-639.

67. Hirsh AE, Fraser HB (2001) Protein dispensability and rate of evolution. Nature 411: 1046-1049.

68. Kryukov GV, Pennacchio LA, Sunyaev SR (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet 80: 727-739.

69. Di Rienzo A, Hudson RR (2005) An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet 21: 596-601.

70. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

71. Hurles ME, Dermitzakis ET, Tyler-Smith C (2008) The functional impact of structural variation in humans. Trends Genet 24: 238-245.

72. Singleton A, Myers A, Hardy J (2004) The law of mass action applied to neurodegenerative disease: a hypothesis concerning the etiology and pathogenesis of complex diseases. Hum Mol Genet 13 Spec No 1: R123-126.

73. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. Science 296: 750-752.

74. Fraser HB, Wall DP, Hirsh AE (2003) A simple dependence between protein evolution

rate and the number of protein-protein interactions. BMC Evol Biol 3: 11.

75. Balaresque PL, Ballereau SJ, Jobling MA (2007) Challenges in human genetic diversity: demographic history and adaptation. Hum Mol Genet 16 (R2): R134-139.

76. Oleksyk TK, Smith MW, O'Brien SJ (2010) Genome-wide scans for footprints of natural selection. Philos Trans R Soc Lond B Biol Sci 365: 185-205.

77. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, et al. (2006) Positive natural selection in the human lineage. Science 312: 1614-1620.

第五章：总结和展望

人类适应当地环境的遗传学基础和人类疾病的遗传学基础是群体遗传学的两个重要研究方向。在本文中，我们首先研究了近期混合人群特别是美国黑人的历史。根据我们揭示的群体历史，我们在中性条件下对美国黑人的形成进行模拟。通过比较真实数据和模拟数据，我们发现美国黑人在离开非洲之后受到很强的自然选择。这表明在人类在近期仍可以受到较明显的自然选择，这也意味这人类的生物进化仍在进行。因为发生在美国黑人群体中的自然选择导致了一些美国黑人的特异性高发疾病，所以这表明近期人类迁移导致的环境变化会诱发一些人类疾病，这对于人类的迁徙有一定的指导意义。疟疾导致的选择压在的北美得到释放，其在基因组上留下很明显的痕迹。这将是自然选择如何塑造人类基因组以及自然选择方向改变后人类基因组又如何被重塑的一个经典范例。

人类疾病的致死性和致残性使得研究和讨论疾病基因受到的自然选择变得具有重要意义。我们根据基因是否参与了孟德尔遗传病和复杂疾病对疾病基因进行简单分类。我们发现复杂疾病基因更多的的受到最近的正向自然选择。但是孟德尔遗传病基因受到负向自然选择。同时参与了复杂疾病和孟德尔遗传病的基因受到正向和负向两种选择压。但是这种现象是对很多基因一起分析发现的一种趋势，具体到每个基因可能不完全遵守这种规律。

既然自然选择作用于人类疾病基因，这些疾病基因内的等位基因频率必然会受到影响。而疾病基因中的等位基因频率分布又将决定发现疾病位点的策略和方法。当致病位点的遗传基础很简单时，设计诊断和治疗策略也会很简单。传统的群体遗传学简单的假设致病位点在理想群体中的频率，然后设计一定的统计量和方法去检测这个致病位点。但是这些统计量和方法基本上都没有考虑自然选择对致病位点的

等位基因频率的影响。

另一方面，不同类型的疾病基因可能受到方向完全不同的自然选择压，同一个基因在不同的时期也会受到不同方向的选择压，这使得不同疾病风险位点的等位基因频率变得更为复杂。特别是人类近期历史经历了一系列奠基者效应和迅速的群体扩张，这种人群历史对致病等位基因频率产生什么样的影响仍然不清楚。这样，我们建议以后筛选致病位点的方法能够考虑更宽广的等位基因频率，而不是仅仅简单考虑中间频率变异和低频变异。

专有名词缩写

LD	Linkage disequilibrium
EHH	Extended haplotype homozygosity
iHS	integrated haplotype score
HGDP	Human genome diversity project
HapMap	International HapMap Project
AfA	African Americans
AAF	Ancestral African population
IAF	Indigenous African populations
PCA	Principle component analysis
CNV	Copy number variation
SNP	Single nucleotide polymorphism
IPA	Ingenuity pathway analysis
MC	Mendelian and complex diseases
MNC	Mendelian but not complex diseases
CNM	Complex but not Mendelian diseases
SEM	Standard error of the mean
GAD	Genetic association database
OMIM	Online Mendelian Inheritance in Man database
hOMIM	hand-curated OMIM database

网络资源

The URLs for data presented herein are as follows:

Coriell Cell Repositories, http://ccr.coriell.org

The International HapMap Project, http://www.hapmap.org/

The HGDP-CEPH project, http://www.cephb.fr/en/hgdp

Mexican Genetic Diversity Project (MGDP), ftp://ftp.inmegen.gob.mx/

Illumina iControl Database (iControlDB), http://www.illumina.com/science

PLINK v1.06, http://pngu.mgh.harvard.edu/purcell/plink/

EIGENSOFT, http://genepath.med.harvard.edu/~reich/Software.htm

FRAPPE and SABER, http://med.stanford.edu/tanglab/software/

fastPHASE, http://stephenslab.uchicago.edu/software.html

PHASE program, http://stephenslab.uchicago.edu/software.html

STRUCTURE, http://pritch.bsd.uchicago.edu/software.html

HAPMIX, http://www.stats.ox.ac.uk/~myers/software.html

Annotation of SNPs, http://genome.ucsc.edu/

HAPMIX, http://www.stats.ox.ac.uk/~myers/software.html

HAPAA, http://hapaa.stanford.edu/HAPAA%20Project.html

LAMP and LAMP-ANC, http://lamp.icsi.berkeley.edu/lamp/

Ensembl Biomart, http://www.ensembl.org/biomart

OMIM genes, http://www.nslij-genetics.org/search_omim.html

BioGPS, http://biogps.gnf.org

The Genetic Association Database (GAD), http://geneticassociationdb.nih.gov/

Human Protein Reference Database (HPRD), http://www.hprd.org/

NIEHS SNPs, http://egp.gs.washington.edu/

GWAS catalog, http://www.genome.gov/admin/gwascatalog.txt

OMIM genes, http://www.nslij-genetics.org/search_omim.html

R, http://www.r-project.org/

Mexican Genetic Diversity Project (MGDP), ftp://ftp.inmegen.gob.mx/

后记

Firstly, I would like to say that life in the late of my PhD program was really tough. In those days, the pressure of graduation nearly drove me crazy, and the life is boring only with my computer stay nearby. Looking back now, I was really surprised that I had gone through. Luckily, the results turned out to be good after all these hard work. I am proud that three or four of my works can be published on high reputation journals.

What I have presented in this thesis is just a tip of the iceberg. In fact, I have explored tens of research projects in the past few years, although most of which have ended up unfinished or unpublished. I do not regret them because I have learned much in the process of doing research. Although many of them might be negligible or too difficult to complete, the ideas underlying each of them used to be thought as great and glorious. I really miss the life at the beginning of my PhD project when there is little pressure. I could still feel the excitement when I started my first project in Felix's lab, under the guidance of Dr. Xu, and the frustration when the programming problem obstructed the way of my project, and the happiness when I completed the project at the end of my first rotation.

In fact, my style in doing research might be inherited from my supervisors Felix and Dr. Xu. Felix always made science fun for me, with his enthusiasm and inspiration. I still remember many interesting stories about his experiences in research and science, which are enlightening, especially the one he told in his office on Liren Biology Building when I did my first rotation. I really appreciate Felix for his ideas, patience, kindness and academic experience, which are invaluable to me and may benefit me in my whole

scientific life. I still remember the first meeting with Dr. Xu in which he told me that I could join with him only if I was interested in his work. Room 312, where he met me and used to be his office, is the place where I devoted most of my time. I also thank him for his guidance and supervision. I am also thankful for the excellent examples both of the two supervisors have set themselves as successful population geneticists. Finally, I really appreciate their patience for revising my manuscripts, which could not be published without their hard work.

I have been in Shanghai for almost six years since I did my undergraduate dissertation in institute of biochemistry and cell biology, SIBS. I really favor this beautiful dynamic city with various cultures and people, during which time I experienced many novel things. Firstly, I appreciated Academician Enduo Wang in institute of biochemistry and cell biology, who provided the opportunity and basic living conditions for me being an intern in her lab. I also thank Prof. Mofang Liu for her recommendation and her guidance. I might not come and stay in Shanghai without their help. The six years in Shanghai is the most important period in my life, during which I experienced a fateful change from an ignorant college student to a young scientist, made a lot of friends that had similar dreams, attended many conferences, met lots of famous scientists, experienced my first real love and had a life direction. I think there will never be a second chance for me to get so many things in such a short period.

I also really enjoy the academic environment in SIBS, which maybe the best in the world according to my limited experiences. The campus is always very quiet and peaceful even in the central area of the modern and dynamic city. I still remember the good days I walked with Jing around the campus. Since there may be still a few months before I leave,

maybe I can fully enjoy a small period of wonderful life in the following days after the thesis defense. I think I will miss everything here no matter where I will be in the future.

Wenfei Jin

Nov 2, 2011

SIBS main building, Shanghai

发表文章及所受奖励

A. 已发表和待发表文章

[1] ***Wenfei Jin***, *Sijia Wang, Haifeng Wang, Li Jin, Shuhua Xu*. 2011. Exploring Population Admixture Dynamics using Empirical and Simulated Genome-Wide Distribution of Ancestral Chromosomal Segments. American Journal of Human Genetic. (Under Review).（本毕业论文第二章）

[2] ***Wenfei Jin***, *Shuhua Xu, Haifeng Wang, Yongguo Yu, Yiping Shen, Bailin Wu, Li Jin*. 2011. Genome-Wide Detection of Natural Selection in African Americans Pre-and Post-Admixture. Genome Research. (Accepted) （本毕业论文第三章）

[3] ***Wenfei Jin***, *Pengfei Qin, Haiyi Lou, Li Jin, Shuhua Xu*. 2011. A Systematic Characterization of Genes Underlying Both Complex and Mendelian Diseases. Human Molecular Genetics. (Accepted) （本毕业论文第四章）

[4] *Shuhua Xu,* ***Wenfei Jin***, *Li Jin*. 2009. Haplotype Sharing Analysis Showing Uyghurs Are Unlikely Genetic Donors. Mol. Biol. Evol. 26:2197-2206.

[5] *Shuhua Xu, Xianyong Yin, Shilin Li,* ***Wenfei Jin***, *Haiyi Lou, Ling Yang, Xiaohong Gong, Hongyan Wang, Yiping shen, Xuedong Pan, Yungang He, Yajun Yang, Yi Wang, Wenqing Fu, Yu An, Jiucun Wang, Jingze Tan, Ji Qian, Xiaoli Chen, Xin Zhang, Yangfei Sun, Xuejun Zhang, Bailin Wu, Li Jin*. 2009. Genomic Dissection of Population Substructure of Han Chinese and Its Implication in Association Studies. Am. J. Hum. Genet. 85:762-774.

[6] *Shuhua Xu, Shilin Li, Yajun Yang, Jingze Tan, Haiyi Lou,* ***Wenfei Jin***, *Ling Yang, Xuedong Pan, Jiucun Wang, Yiping Shen, Bailin Wu, Hongyan Wang, Li Jin.* 2010. A

Genome-Wide Search for Signals of High Altitude Adaptation in Tibetans. Mol. Biol. Evol. 28, 1003-1011.

[7] *Haiyi Lou, Shilin Li, Yajun Yang, Xin Zhang,* ***Wenfei Jin****, Bailin Wu, Li Jin, Shuhua Xu*. 2011. A Map of Copy Number Variations in Chinese Populations. PLoS ONE 6: e27341.

[8] ***Wenfei Jin****, Shuhua Xu, Li Jin*. Genome-Wide Detection of Natural Selection in Amerindians and Mexicans. (In preparing).

[9] *Pengfei Qin,* ***Wenfei Jin****, Dongsheng Lu, Haiyi Lou, Jiucun Wang, Huji Xu, Li Jin, Shuhua Xu.* A Panel of Ancestry Informative Markers for Estimating and Correcting for the Potential Effects of Population Stratification in Han Chinese. (In preparing).

[10] ***Wenfei Jin****, Shuhua Xu, Li Jin*. Biased Gene Conversion or Genotyping Bias: Implication from HapMap Data. (In preparing).

B. 所受奖励

2011年 国际人类遗传学大会旅行奖 （International Congress of Human Genetics; ICHG）

2011年 上海生科院辉瑞奖学金

2010年 中国科学院地奥奖学金

C. 主要会议摘要

Wenfei Jin*, Shuhua Xu, Pengfei Qin, Li Jin*. 2011. Characterizing Genes Associated with both Complex and Mendelian Diseases. The 12th International Congress of Human Genetics/61th Annual Meeting of ASHG: October 11-15, 2011, Montreal, Quebec, Canada.

D. 图书章节

Shuhua Xu, ***Wenfei Jin****. 2011. Population Genetics in the Genomic Era. Intech. (Revised)*

致谢

It is a great honor to be one of the first PhD students enrolled and graduated from PICB. I would like to express my sincere appreciation to those people who made this thesis possible with their support and assistance.

First and foremost, I want to thank my supervisors Prof. Felix Li Jin and Dr. Shuhua Xu. With their enthusiasm, inspiration and endeavor, they have made population genetics fun for me. I still vividly remember many of the meetings and conversations with them. I appreciated both of them for their guidance and suggestions, listening to my boring reports and reading pages of drafts. Particularly, I thank Felix for his ideas, patience, kindness and academic experience, which are invaluable to me. I still remember many interesting stories about his experiences in research and science, which are enlightening. Dr. Xu is the guide that led me into the kingdom of science and made my first publication come true. I also appreciated him for paying so much attention to me and designing my initial PhD project. I am also thankful for the excellent examples both supervisors have set themselves as successful population geneticists. Finally, I really appreciate their patience for revising my manuscripts, which could not have been published without their help and hard work.

The members of the Computational Genomics Group and Population Genomics group have contributed immensely to my personal and professional time. Both groups always provide source of good advice and collaboration, as well as friendship. Firstly, I am grateful for people who have given me constructive suggestions including Zhengwen Jiang, Yungang He, Shi Yan, Wei Wang and Erli Wang. I also would like to acknowledge

other members in the group including Ling Yang, Haiyi Lou, Pengfei Qin, Hongyang Xu, Dongsheng Lu, Ran Li, Minxian Wang, Meng Shi, Ying Zhou and Ruiqing Fu. The previous group members including Zhifei Ma, Zhijun Wang, Gu Lei were also appreciated.

The encouragements from many friends have been indispensable, and I would like particularly to acknowledge: Guofeng Meng, Chunxuan Shao, Yuling Liu, Chaofeng Wang, Kai Weng, Zhongshan Li and Niyi Shao.

I thank Miss Jing Pu for her assistance in editing my manuscripts. I also acknowledge Ms. Lisa Shuqin Li, who has provided so much help in both life and study that I can concentrate on my study and research.

I also appreciate my brothers, sisters, aunts and uncles for their supports. Especially, I thank my second elder brother for his strong support in recent ten years and hope he has a more bright future.

Last but most important, I want to appreciate my father, who has devoted so much to me that I could not overstate my gratitude to him. I feel proud that I have such a great father.

中科院上海生命科学研究院

研究生学位论文声明

本人郑重声明：

1） 所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容或属合作研究共同完成的工作外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。

2） 所呈交的学位论文，实验结果均由相应的实验数据分析得出，实验数据真实可靠。

该声明的法律结果由本人承担。

学位论文作者签名：

日期： 年 月 日

研究生学位论文版权使用授权声明

本人完全了解并同意遵守中科院上海生命科学研究院有关保留、使用学位论文的规定，即：生科院有权保留送交论文的复印件和电子文件，并提供论文的目录检索及借阅、查阅；生科院可以公布论文的全部或部分内容，可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其它复制手段保存和汇编本学位论文。保密的论文在解密后遵守此规定。

作者签名：_____ 导师签名：_____ 日期：