Science 文章《从基因组变异模式推断世界范围的人类关系》评述
——兼论数据科学的研究范式

张　悦[1]　刘宇航[2]
[1] 浙江大学
[2] 中国科学院计算技术研究所

关键词：数据科学　基因研究

引言

数据科学（Data Science）作为一门新兴的“专业”目前受到国内外大学的普遍重视，分别做了相应的课程设置，一些顶尖大学还成立了专门的数据科学学院或专业。但是数据科学尚处于发展初期（infancy）[3]，从研究对象到研究方法都没有确定。本文解析一篇《科学》（*Science*）上发表的文章《从全基因组变异模式推断出世界范围的人类关系》[1]，结合实例透视数据科学的研究范式，为同行的后续研究提供参考。

这篇文章介绍了如何利用人类基因组多样性小组（HGDP-CEPH）51个种群中938个无关个体的65万个常见单核苷酸多态性基因座（SNP）[1]进行相关研究，以详细描述全球人类关系与人类基因变异。鄂维南院士在《数据科学的基本内容》一文中提到过这篇文章[2, 4]，但没有展开叙述。

人类遗传多样性是由人口和生物学因素共同决定的，对理解疾病的遗传基础具有重要意义。2020年9月，中国科学院古脊椎动物与古人类研究所研究员付巧妹在向习近平总书记汇报工作时，提到了这一“最古老”的研究课题背后的科学价值。在过去的30年中，对DNA序列变异的能力开展的研究极大地增加了我们对人类之间的血缘关系和历史演变的了解。尽管取得了这些进展，但是这些研究仅基于有限的基因组或人口，并不能完全了解突变、重组、迁移、人口统计、选择和随机漂移的相对重要性，这一课题还有很大的研究探索空间。

文章主要内容与相关结论

样本与数据集的选择

人类基因组多样性小组由1064名个体组成，他们分别来自51个人群。按地理位置对人群进行划分，样本中的51个人群分别来自非洲（北非和撒哈拉以南非洲）、欧洲、中东、南亚/中亚、东亚、大洋洲和美洲。与传统的研究相比，该研究一方面增加了对部分人口特征的研究，另一方面大幅增加了以往研究的基因组和人群覆盖率，可以对人类遗传变异进行全面的表征。在1064名个体中，有1043人成

[1] 指在基因组水平上由单个核苷酸的变异所引起的DNA序列多态性，是人类可遗传变异中最常见的一种。

功地进行了基因分型[2]（假设检出率大于98.5%为成功），有952个样本来自不相关[3]的祖先个体。在成功分出基因型的1043人中，有938个彼此无关的个体，其中包括615名男性和323名女性。

在原文献的支撑材料中能够找到可供分析的Excel文件。其中，数据表提供了51个人群的33项数据，包括人群名称（比如中国的若干民族）、该人群所在的国家、所属的洲际位置以及各项研究方法用到的数据特征。

哈迪－温伯格检验

基于哈迪－温伯格（Hardy-Weinberg）法则[4]，文中对51个人群进行了哈迪－温伯格平衡测试[5]。文章在这部分的精确检验中主要选取了检验的P值小于0.01与小于0.001两种情况，具体得出以下结论：当检验的P值小于0.001时，SNP的绝对数量在0~570范围内变化，相对比例低于10%；当检验的P值小于0.01时，SNP的相对比例低于1%。SNP的数量也取决于人群，并且与样本量呈正相关。在P值小于0.01的情况下，大多数SNP杂合子计数均低于预期的结论，这也表明杂合子计数的偏差与检验水平之间存在一定程度的相关性。这一部分主要是通过对数据的描述性分析得出的结论，可以对数据集中的相关数据进行分析来加以验证。

检测个人血统

通过个人血统的基因来源进行辨识

这里的分析假设每个人的基因组都起源于k个祖先（无先验信息），每个种群对这个个体基因的贡献程度由k系数描述，每个人的k系数总计为1。文献中给出了k=7时的个人祖先图像，其中每条竖线表示一个个体，每条线中的不同颜色表示来自不同祖先的基因。

通过对其他参考文献以及参考资料的查阅，我们得到了更多不同的k值对应的图像。例如，将k=7的图像与k=6的图像相比较，可以发现**新成分以最高比例出现在中东人口中（与非洲人口和欧洲人口分离开来）**。同理，当k=8或者更高时，会出现其他类别，**往往代表了某个区域中的异常种群**。另外，在许多人口中，血统主要来自其中一种推断的成分，而在其他人群中，**尤其是在中东和南亚/中亚，则有多种血统**。

基于个人血统探究人口子结构

由于属于同一公认种群的个体几乎总是显示相似的祖先比例，因此对种群之间的遗传关系进行统计学评估非常有意义。该研究使用直系黑猩猩等位基因为系统外群，通过极大似然法（Maximum Likelihood，ML）构建系统发育树[6]，以此对种群之间的遗传关系进行统计学评估。

文中提到了两种构建系统发育树的方法，分别为NJ（Neighbor Joining）法[7]与ML法[8]，并且指出用这两种方法构建的系统发育树是极为类似的。

构建系统发育树之后，可以结合地理先验信息，将个人血统与祖先遗传关系和地理先验信息结合起来，**由个人血统对应种群地理位置，依据祖先遗传信息推断时间顺序，从而可以探究不同地域人群扩**

[2] 基因分型是通过使用生物学试验检查个体的DNA序列的过程。

[3] 相关指的是数据集中有些人是另一些人的一级亲属或二级亲属。一级亲属包括父母、子女及兄弟姐妹；二级亲属包括叔、伯、姑、舅、姨、祖父母、外祖父母。

[4] 该法则是指在一个有性生殖的自然种群中，在符合以下5个条件的情况下，各等位基因的频率和等位基因的基因型频率在一代一代的遗传中是稳定不变的：（1）无限大的种群；（2）种群间雌雄个体间的交配是随机的；（3）没有基因突变发生；（4）没有任何形式的自然选择；（5）没有基因的迁入与迁出。

[5] 检验基因型分布是否符合哈迪－温伯格平衡。

[6] 系统发育树用一种类似树状分支的图形来概括各种（类）生物之间的亲缘关系。

[7] NJ法是一种自底向上（bottom-up）的聚类方法，认为进化历程中发生碱基替代次数最少的系统发育树为最优树。

[8] ML法基于碱基替代模型，认为似然值最大的系统发育树为最优树。

张的先后顺序，即可刻画出其动态的人口迁徙或扩张过程。由构建的系统发育树可以发现：**撒哈拉以南非洲的种群位于最靠近树根的地方，并且从树的内侧向外依次对应于北非、南亚 / 中亚、欧洲、大洋洲、东亚、美洲、中东。这在很大程度上与众多学者假设的人群扩张顺序一致，支持了人类起源的“Out of Africa”**[9]**假说。**

Fst 计算与人口结构主成分分析

Fst 计算

研究中通过使用所有常染色体 SNP 的人群等位基因频率，计算了所有人群对的 Fst [10]，其计算公式为：Fst =（HT-HS）/ HT。其中 HS 代表亚群体中的平均杂合度，HT 代表复合群体中的平均杂合度。

不难看出，Fst 取值从 0 到 1，用于衡量种群分化程度，取值为 0 表示两个种群间是随机交配的，基因型完全相似；取值为 1 则表示两个种群是完全隔离的，基因型完全不相似。

在两个地理区域内进行精细规模的人口结构主成分分析

因为该研究总共有 51 个人群，所以这里分析的是 51 × 51 的 Fst 矩阵，主要从以下两方面展开：

1. 捕获遗传变异的主要部分：如果 A、B 群体混合生成了 C 群体，那么在 PCA 分析结果图像中，C 群体会在 AB 的连线上。但是如果该混合事件发生在很久以前，由于混合群体经历了自己特有的遗传漂变，那么 C 群体会逐渐偏离该连线。**该研究通过这种方法将时间维度的种群变化用空间维度的图像来刻画，**文中关于这一部分得出的结论是**随机遗传漂变为导致遗传变异模式的主要因素。**

2. 研究迁移路线：用撒哈拉以南非洲的群体作为变量空间进行 PCA 分析，将其他群体映射到该变量空间，可以发现，所有其他非撒哈拉以南非洲的群体都聚在一起，**暗示了人类从南非的一次迁出。**

个人水平上的区域聚类

在研究个人水平上的区域聚类时，计算 938 个人的状态一致性（Identity-by-State, IBS）矩阵，并对该矩阵的所有样本和 7 个区域分别进行了主成分分析，使用最重要的成分说明个体之间的遗传相关性。**这些个人水平的结果都表明，尽管某些种群的样本量有限，但种群结构似乎很稳健。**

单倍型杂合性研究

文章首先获取 51 个种群中染色体数据的单倍型频率，据此来计算预期单倍型杂合度并取平均值，然后将其与“距埃塞俄比亚城市亚的斯亚贝巴（Addis Ababa，人类扩张的合理起点）的地理距离”作图并进行比较。可以得出结论：**常染色体单倍型的平均杂合度与距亚的斯亚贝巴的距离呈负相关。这一趋势与建立者效应一致（遗传漂变）：在这种情形下，人口膨胀涉及从撒哈拉以南非洲的单一起源开始，一小部分人连续迁移出先前的地点。**这也再次印证了之前的一系列结论。

重要结论在东亚 / 中国的应用

数据集中关于中国的相关数据

原数据集中的 51 个人群中，属于中国的共有 15 个，并且在数据集中按民族体现、分类，15 个人群分别对应的民族名称如下表：

表 1　数据集中中国的 15 个民族对应名称

Oroqen: 鄂伦春族	Hezhen: 赫哲族	Daur: 达斡尔族
Mongola: 蒙古族	Xibo: 锡伯族	Tu: 土族
Naxi: 纳西族	Yizu: 彝族	Han_N: 哈尼族
Han: 汉族	Tujia: 土家族	Miaozu: 苗族
She: 畲族	Dai: 傣族	Lahu: 拉祜族

[9] 非洲起源说，由达尔文在 1871 年出版的《人类起源与性的选择》中提出。

[10] Fst：群体间遗传分化指数，是种群分化和遗传距离的一种衡量方法。分化指数越大，差异越大。适用于亚群体间多样性的比较。

个人血统检测

基于前文的个人血统检测的内容与方法，可以进一步研究东亚的情况。根据文中的图像发现，在 k=2 时，画出的个人祖先图像中比较明显的两个人种是非洲人种与东亚人种。由此可以看出，**东亚地区的人种起源相对较早，并且有相对纯正的血统，这也与一直以来比较传统的两大人类起源理论相符合。**[11]

进一步观察中国各个民族的个体基因组成情况，可以发现**中国的种族血统相对古老并且纯正**。我们可以将系统发育树中涉及到中国的这些种族的部分放大，如图 1 所示。

根据图 1 我们可以得到一些关于中国种族的假设：（1）维吾尔族是中国出现相对较早的人群，其与系统发育树的起点直系黑猩猩之间的血缘关系较远。（2）雅克库族与鄂伦春族（中国起源较早的民族）有比较紧密的血缘关系，相似度高。（3）汉族与中国出现最早的人群（与原始人群相似度更高的维吾尔族）之间已经出现了一定程度的差异，并且经过了比较长时间的繁衍进化。（4）日本人与中国汉族之间有比较亲密的血缘关系。（5）傣族可能是中国出现最晚的人群之一，与最原始的中国人群之间有较大差异。该人群与柬埔寨人有比较亲密的血缘关系。以上族群出现早晚的假设已经得到验证，历史资料考察与所得结论基本一致。

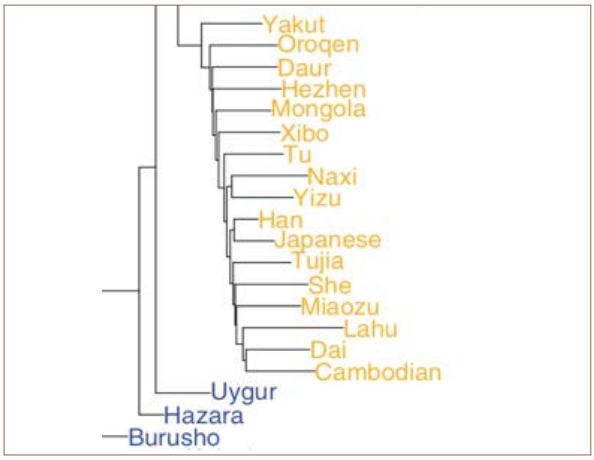

图 1　系统发育树中中国各种族部分

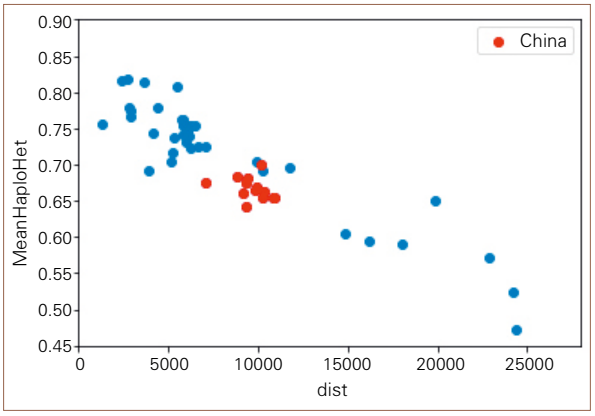

图 2　中国各民族在单倍型杂合度与到亚的斯亚贝巴距离关系图中的位置

单倍型杂合性研究

类似地，根据原始数据集，可以大致找出中国各种族在“单倍型杂合度与距人类扩张的合理起点地理距离”这一散点图中的位置（图 2 中的红色部分），并且可以大致观察出它们到亚的斯亚贝巴的绝对距离以及与其他种群相比的相对距离，这一位置与系统发育树中中国各种群的位置也大致一致。

总结

这篇文章作为一篇经典的基因研究文章，很好地展示了如何将生物专业的领域知识与数据科学的相关方法结合起来，用数据科学的视角看待与解决问题。

数据科学具有跨学科意义

该项研究获得显著成果的关键之一就在于**更加注重对变量之间相关性的研究**。数据科学的研究方法开阔了研究者的视野。相比于传统的 DNA 序列研究与基因层面的血统研究，该项研究引入了更多

[11] 达尔文在 1871 年出版的《人类起源与性的选择》中假设非洲是人类的摇篮，而另一位进化论者海格尔在 1963 年发表的《自然创造史》中主张人类起源于南亚。

宏观层面的变量（如地理迁移痕迹等），研究思路包括寻找数据之间的相关性以及所得数据与其他变量之间的相关性等，进行了大量数据的比对，更好地发挥了数据科学研究方法的能动性。

例如，传统的研究更多地专注于技术手段提升、研究仪器精细化，对于基因本身进行研究，而该项研究跳出了原本的研究范围，将基因研究得到的数据与地理距离、人种之间差异性、个体杂合度等众多变量联系在一起，广泛研究其相关性，再一一进行深入探讨，从而得出一系列具有创新意义的研究成果。

这从一方面揭示了数据科学基于数据进行探索、发现关系、提出问题的特征，另一方面也体现了**数据科学方法的跨学科意义**。对于其他学科来说，无论是在研究思路还是研究方法上，数据科学的相关方法都有借鉴意义，在专业领域利用数据科学的手段可以更好地挖掘数据背后的关系，从而进行更多的创新。

数据科学研究方法能将抽象问题具像化

人类起源、个体或种族之间的血缘关系、遗传漂变的解释等问题，本来都是比较复杂和抽象的，但是作者在研究过程中很好地选择与利用了数据，将难以比较或者难以研究的问题转化为可用数据解决的问题，将相对较为主观的问题量化为可以用计算机处理的变量。

以研究个人血统为例，本文中主要通过 k 值的假设来建立相应的模型，在此基础上再进一步将模型推广，把数据与地理位置扩张、人口迁移等相关内容联系起来。从主观角度来看，地理扩张是时间序列上的变动，而基因序列是微观层面一个个体的特征，两者看似相关性很小，甚至研究的维度也不相同，但是该项研究以个人祖先的血统为桥梁，建立起了两者之间的联系。**将截面数据与时间数据序列结合起来，**在研究某个地域个体的混合祖先时间顺序的基础上，再比对不同地域的人群个人血统，即可得到地理扩张的方向。这也启发着我们在其他研究中，**对于时间序列方面的数据，也可以通过引入中间变量将其转换为对一个时间点上截面数据的研究，从而降低数据量与研究难度。**

数据可视化技术的应用

文章中涉及到大量对于数据的探索与数据间关系的研究，**数据可视化为识别变量因果关系和相关性提供了很好的手段。**原文中有大量数据可视化应用的例子，以文献中的极大似然法为例，该项研究中用 k 值的假设建立相应的模型后，通过算法的设计将数据可视化，绘制出相应的图谱以及发育树来更好地解释结果。还有对人口迁徙路径的研究等，都是利用数据可视化来清晰、有条理地向人们展示数据。另外，数据可视化也在最初获取数据、对数据进行初步探索的阶段起着重要作用，研究者可以基于此对变量提出一些大胆的假设。

人类古基因组学、演化遗传方面的研究仍然还有许多值得研究探索的问题。这篇经典的文献给人们带来的启发不仅仅在于其所提出的结论上，更在于如何用清晰的思路与逻辑，结合新的手段与不断进步的算法进行创新性的研究。 ■

张　悦

浙江大学竺可桢学院本科生。主修统计，辅修信息管理与信息系统。研究兴趣为数据挖掘与机器学习。

刘宇航

CCF 高级会员，*CCCF* 特邀专栏作家，CCF 职业伦理与学术道德委员会常委。中科院计算所副研究员、硕士生导师。主要研究方向为计算机体系结构、高性能计算、数据密集型计算、类脑计算、存储系统。liuyuhang@ict.ac.cn

参考文献

[1] Li J Z, Absher D M, Tang H, et al. Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation[J]. *Science*, 2008, 319(5866):1100-1104.

[2] 鄂维南 . 数据科学的基本内容 . 中国计算机学会通讯 . 第 13 卷第 8 期，2017 年 8 月 . 45-48.

[3] Francine, Rob, Henrik. Realizing the Potential of Data Science. Communications of the ACM, 2018, 61(4):67-72.（实现数据科学的潜能 . 刘宇航 译，中国计算机学会通讯，第 14 卷第 8 期，2018 年 8 月 . 90-96.）

[4] 李国杰 . 发展数据学科应在何处发力？中国计算机学会通讯，第 14 卷第 8 期，2018 年 8 月，主编评语 .