# 如何研究一个宗族遗传网络

史定华[1]，阎春宁[2]

(1. 上海大学数学系　上海 宝山区　200444;　2. 上海大学管理科学与工程系　上海 宝山区　200444)

【摘要】社会人群都离不开姓氏符号，它记载着人与人之间亲疏关系的演化与博弈。该文以史姓为例，阐述姓氏、血缘、族谱之间的区别，探寻如何构建具有科学意义的宗族遗传网络，以及其中涉及的某些科学问题。

关　键　词　遗传网络;　血缘;　族谱;　姓氏;　Y染色体
中图分类号　N94　　文献标志码　A　　doi:10.3969/j.issn.1001-0548.2013.03.001

# How to Study A Clan Genetic Network

SHI Ding-hua[1] and YAN Chun-ning[2]

(1. Department of Mathematics, Shanghai University　Baoshan Shanghai　200444;

2. Department of Management Science and Engineering, Shanghai University　Baoshan Shanghai　200444)

**Abstract**　Social groups are inseparable from the surnames, which recorded the relationship between person and person in the process of evolution and gaming. Taking the family name “SHI” as an example, this paper intends to trace the history of a clan, explains the difference between surname, kinship, and pedigree of the clan, and shows how to construct the scientific significance of the clan genetic network. At last we conclude with some scientific problems involved.

**Key words**　genetic network;　kinship;　pedigree of a clan;　surname;　Y chromosome

虽然没有数据验证，但直觉告诉我们：科学重大发现并不完全随机，而往往是阵发出现的。世纪之交，为了解答人际关系的“六度分隔实验”，1998年《自然》刊登了小世界网络模型的论文[1]，开创了网络科学新纪元。同年，为了破解“美国总统杰斐逊私生子疑案”，《自然》发表了Y染色体检测报告[2]，为两个多世纪的疑案画上了句号。

可见，社会网络研究的人际关系，并不限于朋友之间的通信，亲缘关系可能更为重要，但复杂网络圈子很少将上述两篇论文联系在一起。生物网络研究虽然兴旺发达，但较多集中在基因调控、新陈代谢、蛋白质相互作用等方面。宗族遗传网络可以看作两者的结合，由于涉及跨学科的原因，这个新方向的研究似乎刚刚起步。

最近，国人在国际学术期刊上发表了两篇重要文章，一篇通过Y染色体检测揭开了曹操身世之谜[3]；另一篇则研究了我国12.8亿人口的姓氏分布情况[4]。两篇文章都采用自然科学结合社会科学的方法去研究姓氏与血缘，才开始引起国内学者对宗族遗传网络研究的关注。

两个同姓人相遇，不约而同都会说，500年前咱们是一家，你认为果真如此吗？因为姓氏世代相传，此话有一定的道理，说明姓氏能反映某种亲疏关系。可是仔细研究后，却发现谬之千里，可能5000年前都不是一家，倒是不同姓的人血缘更亲近。

亲疏关系是社会网络的一种重要关系，这里涉及姓氏的起源，血缘的遗传，族谱的甄别，世系的记载。本文打算以史姓为例，阐述姓氏、血缘、族谱之间的区别，应该如何构建具有科学意义的宗族遗传网络，以及其中涉及的某些科学问题。

## 1　姓氏起源

姓氏符号乃皇家所赐。史美珩教授于2002年为《中华姓氏谱》丛书撰写了“史”姓一册[5]，介绍了史姓的起源。大致如下：史皇氏仓颉，汉字创造者，黄帝史官；西周的史佚，伯邑考之子，史官；周朝其他史官；唐朝的大奈，突厥阿史那氏，赐姓史；唐朝李继先，本夏侯氏，也赐姓史，昭武九姓中的史姓，以及少数民族史姓等。可见，两个史姓人相遇，除非知道是同宗，否则，不仅500年前不是

收稿日期：2012－08－10；修回日期：2013－01－07
基金项目：国家自然科学基金(61174160)
作者简介：史定华(1941－)，男，教授，主要从事生物信息和复杂网络方面的研究.

一家，甚至5 000年前都可能不共祖。

北京师范大学系统科学系与中国科学院遗传所的研究人员，针对我国2008年12.8亿人口姓氏的情况分布有详细研究[4]。如中国人口姓氏的分布也像网络度分布一样服从幂律，在近万个姓氏中，使用前100个大姓的人口占中国总人口的85%。早在2001年，文献[6]就曾进行过“人口姓氏分布幂律行为的动力学模型探索”。

根据统计现在史姓人口大约有300万，史姓有幸进入前100个大姓中。史姓人口数排前三甲的省份为：河南、河北和山东，占当地人口比例数前三甲的省是：山西、河北和河南。史姓人口在全国各个省市的分布，以及占当地人口的比例如图1所示。

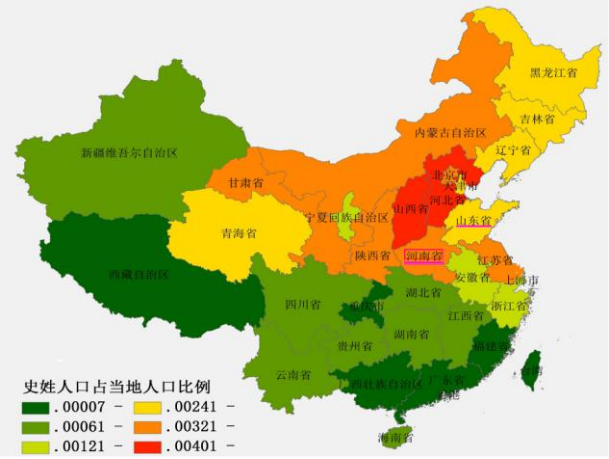


图1　史姓人口占当地人口的比例[4]

从图1大致可以看出，史姓是一个古老的姓氏，开始主要在北方繁衍，然后逐步向南方迁徙。更详细的县区史姓人口分布图也显示与全国史氏族谱的记载基本吻合。

在中华历史长河中，史姓最具有代表性的三大宗是：2 000年前汉代陕西杜陵史家，先祖史恭，因其妹妹史良娣是汉宣帝的祖母，曾追封杜陵侯；史恭四传至史崇，为现今江苏溧阳侯一世祖。1 000年前南宋浙江四明史家，先祖史浩，宋封忠定王；900年前元朝河北真定史家，先祖史天泽，元封镇阳王。个个文治武功，可谓政绩辉煌。

## 2　血缘关系

血缘关系由父母确定。诞生于半个世纪以前的分子生物学，特别是DNA双螺旋结构的发现，揭开了科学新篇章。地球生命体DNA分子都含有4种不同的碱基，分别是：腺嘌呤(adenine)、鸟嘌呤(guanine)、胞嘧啶(cytosine)和胸腺嘧啶(thymine)。

人类含有23对染色体，大约由30亿个碱基对组成，其中1对性染色体决定性别(男XY，女XX)。可见Y染色体只能在男性中传递，由父传子，子传孙，世代遗传决定了血缘关系，如图2a所示。根据文献[7]介绍，Y染色体长大约6千万碱基，两端1.5%是与X染色体重组区，余下为非重组区，其中常染色质区占一半。Y染色体中单核苷酸多态(SNP)是指一个碱基发生突变的结果，突变率为三千万分之一，发生重复突变的概率约百万亿分之一。这些突变构成的组合叫做单倍(型)群，称为遗传标记。而短串重复序列(STR)是指由几个碱基组成一个单位重复出现，它们往往有不同的拷贝数目。由于突变快，可作为遗传时钟。SNP在Y染色体中传递形成单倍群的过程，其示意图如图2b所示。

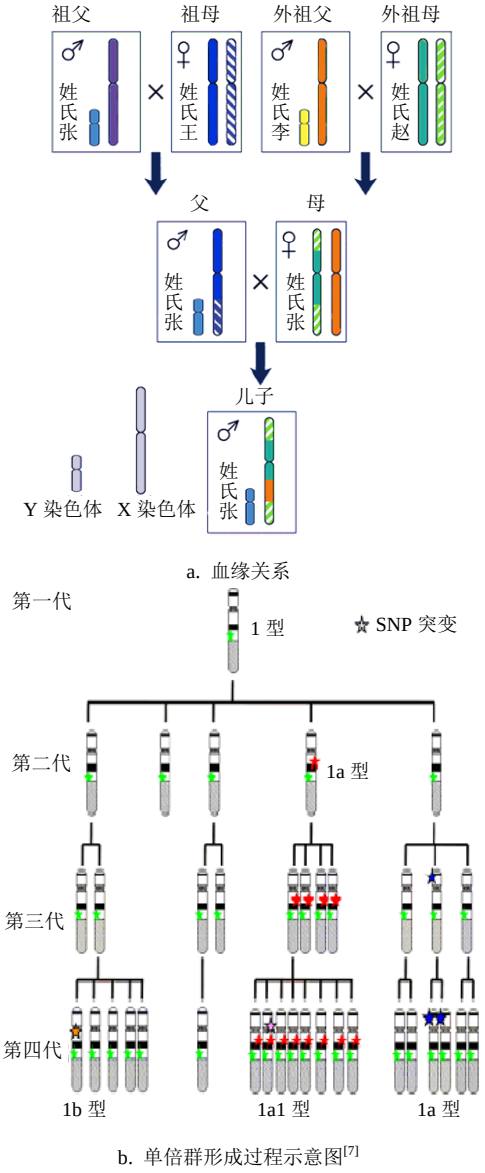


图2　Y染色体遗传

根据Y染色体检测结果分析，华夏族群90%以上的男性主要单倍群是NO-M214。这个华夏族群的共同始祖NO-M214出现在3.5～4万年前，然后分成两

大支系。O-M175出现在大约3万年前后，它先后产生的3个分支是：百越系O1-M119、三苗(百濮)系O2-M95和汉藏系O3-M122。N-M231出现在大约2万年前后的中亚地区，它的3个分支中N1稍后来到中华。除此之外，华夏族群还有更古老的C-M130后裔以及其他单倍群。

由复旦大学现代人类遗传学教育部重点实验室检测报告中提供的汉族O单倍群树如图3所示，其中做了标记的单倍群已在史氏族人中被检测到。

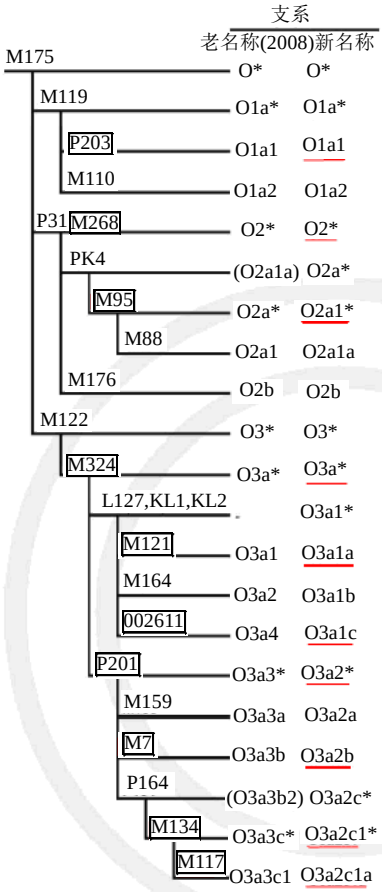


图3 汉族O单倍群树(2011年) (取自复旦检测报告)

## 3 族谱分类

族谱起着连接姓氏与血缘的桥梁。按照约定俗成，族谱主要记载某一同宗共祖的家族世系繁衍，入谱者必须是同宗共祖，即使同姓，若不同祖，也不能录入其中。

史氏族谱，根据史美珩教授“史”姓一书的论述，大约有千余年历史。代表性的有纸谱，如《古藤史氏宗谱》和《庆传录》；石谱，如《史氏庆源之碑》，另外还有《史氏谱录合编》等等。由于年代久远，加之沧桑巨变，老谱留存不多，况且残缺不全。现代分子生物学已经证实：男性Y染色体具有代代相传的特性，祖先的遗传标记(单核苷酸多态和短串重复序列)能够在直系后代中找到，可视为血谱，它为族谱研究提供了新的契机。

史氏纸谱比较完整可靠的是四明史氏族谱，它的字辈：师水弥之卿，孙公祖必仕，本立自元孝，起宗在节义，积善致悠久，济美习庭训，像一首史诗，具有很高的文化价值。为了区分宗族子孙的世次辈分，四明史家在中华姓氏中是较早使用字辈的家族。然而，有的纸谱，如江西城南(南昌郊区)史氏族谱，它的字辈：本卿孙公始，必仕德立自，元孝起宗在，节义纪汝应，国中祖述宪，章裔世贵端，良彦贤美家，则有点乱。由于种种原因，这里可能存在抄袭、假托、乱接等等，需要认真甄别，这是各地现存纸谱中的普遍缺陷。

城南史氏族谱最后一次修订是公元1919年，部分世系如图4所示。这份世系显示：祖先用了若干四明史氏族谱字辈，知道某些真定史家先祖名讳，如必澤的8子：格、樟、棣、杞、杜、梓、楷、彬，除了杜是杠的笔误外与史天泽的8子完全相同。清朝康乾年间修谱蔚然成风，特别是溧阳史贻清搞了个全国性的《史氏谱录合编》。城南史氏的祖先为了应对设下了伏笔：“来自四明，始于溧阳，堂号秉直，城守豫章”。在当时的社会背景和科技水平下，这一伏笔充满了政治智慧，随着科技发展，也给后代留下了探寻空间。

史氏石谱最俱价值的是真定天泽父亲秉直公1240年所刻《史氏庆源之碑》。碑中有“史氏得姓源于史佚。至唐有太师者始占籍为永清之兴隆里人，裔孙仕辽静江节度使又为尚书左司，其子荫补为县令。尚书家世胚胎皆此”。这段文字说明了真定史氏先祖源流的概况，因是祖先亲自撰写，比后代修撰的纸谱可靠性高。另段文字“尚书公曾祖，岁久失其名讳。生子四人，伯、仲、叔、季。季则尚书公之祖也，讳伦妣鄂氏，生一子成珪，是为尚书公之父也。”说明真定史氏应以史伦为始祖。另段文字“尚书生三子。长子天倪，为河北东西路兵马都元帅。次天安，见(xian)任宣权五路万户，知中山府事。次天泽，见任真定、河间、东平、大名、济南五路万户，知真定府事”。石谱因为种种限制包含信息不多。

史氏血谱研究刚刚起步，2012年5、6、11月来自全国各地60余位(其中中国台湾4人)史氏后人，分四批次在复旦大学做了Y染色体检测，获得的遗传标记(单倍群)结果归纳如下：O1a1-P203，O2*-M268，O2a1*-M95，O3a*-M324，O3a1a-M121，O3a1c-002611，O3a2*-P201，O3a2b-M7，O3a2c1*-M134，O3a2c1a-M117，C3*-M217，N1-M231，M-P256。

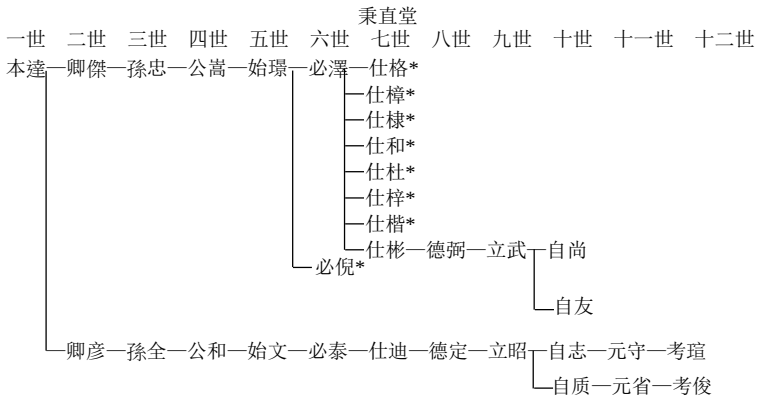


图4　1919年城南史氏族谱世系截图

解读所得结果不难发现：5～10代左右的堂兄弟完全相同，检测的准确性应该没有问题；11个遗传标记为O1a1-P203从中能够区分出6个四明史氏后人，检测的有效性可以初步肯定；同一分支，如城南史氏，遗传标记基本相同，结果的一致性揭示血缘纯正；不同分支遗传标记相距甚远，因为史姓起源多种多样，来自全国各地史氏后人的血缘怎能相同？结果的多样性正好给了印证。由于样本太小，不足万分之一，取样也欠科学，检测结果还比较粗糙，但已经获得许多有价值的信息，为破解史氏族谱中的千古谜团提供了重要的科学证据。如通过这次检测初步确定了四明遗腹子的遗传标记为O1a1-P203(下游有待确定)；还可以肯定城南史氏不是“来自四明，始于溧阳”，而是真定史家军的后裔。

现代人对家族的历史超过三代就难以说清。自己的姓，起源何时何地，有何名人？探索“姓”的起源，成了寻根的情节。近年来，为满足后人寻根的情节，全国兴起了一股修撰族谱的热潮。纸谱、石谱、血谱是族谱的3种重要类型。纸谱由后人修撰，石谱为祖先记述，血谱则是大自然所馈赠，应该充分重视，加以认真发掘。

## 4　遗传网络

2004年，史氏在全国率先创办了《史氏春秋》网站，建立起第一个史姓历史研究交流的信息平台。随后各地《史氏历史研究会》相继成立，极大地推动了史氏族谱的研究，取得了许多丰硕的成果。2008年，溧阳“侯庙”重建成功，为史氏文化研究搭起了一座新的平台。同时《中华史氏大同谱》编辑委员会的成立，为深入研究史氏族谱吹响了新的进军号角。族谱修撰，只争朝夕固然重要，科学方法不可或缺。虽然寻找旧谱和查阅史料能提供许多有价值的线索，但前面几节的论述说明，单一以历史人物和地理位置作为族谱研究的思路无法维继，需要抛弃封建大一统的思想，辅以通过Y染色体检测，确定各地史氏分支的关键遗传标记，建立史氏遗传网络，才能将族谱研究推向新的阶段。

《自然》发表的Y染色体检测报告[2]，成功地破解了“美国总统杰斐逊与黑人女奴海明斯私生子疑案”，为判断两个世系明确的分支有否血缘关系提供了成功的范例。针对多个宣称但不明确的曹操后裔分支，通过Y染色体检测利用统计方法揭开了曹操身世之谜，即找到了曹操较显著的遗传标记，文章发表在《Journal of Human Genetics》杂志上。这些研究成果说明Y染色体检测可为族谱研究提供新的科学手段，有望解决族谱研究中的始祖与外迁难题。

由于祖先的遗骨大多无法找到，只能依靠还活着的后人去统计推定。这将会涉及许多科学问题，必须通过考古学、气候学、历史学、谱牒学、遗传学、统计学、网络学等多学科交叉研究才可望取得突破。下面仅以概率统计问题做一个简单介绍。

首先，人们会遇到如不同分支传承快慢等问题。这里正态分布扮演着重要角色，过快过慢都是小概率事件，需要认真对待。史氏族谱中的许多数字谜团都与此有关。

其次，生物群体的世代繁衍和灭绝在概率论中早有研究，分支过程[8]提供了基本的数学模型。令 $X_n$ 是第$n$代群体数目，$p_k$ 是有$k$个后代的概率，具有分布 $\{p_k\}$ 的随机变量记为 $\xi$。假设 $X_0 \equiv 1$，于是 $X_{n+1}=\sum_{r=1}^{X_n}\xi_r$，引进母函数 $\varphi(s)=\sum_{k=0}^{\infty}p_k s^k$，$\varphi_n(s)=\sum_{k=0}^{\infty}P\{X_n=k\}s^k$，则有 $\varphi_{n+1}(s)=\varphi_n[\varphi(s)]$。因为 $\varphi_0(s)=s$，还有 $\varphi_{n+1}(s)=\varphi[\varphi_n(s)]$。

假设$\xi$的二阶矩存在，均值$m$，方差$\sigma^2$，则得第$n$代群体平均数 $E[X_n]=m^n$，据此，可以推算出史姓溧阳侯的后代人数，误差由 $\mathrm{Var}[X_n]=$

$$\begin{cases}\sigma^2 m^{n-1}\frac{m^n-1}{m-1} & m\neq 1\\ n\sigma^2 & m=1\end{cases}$$ 估算。

考虑到战争、生存环境等多种因素的影响，同代中某些个体延续若干代就可能会绝后。为了保证种群不灭绝，假设 $p_0+p_1<1$，这时绝后概率 $\pi$ 是方程 $\varphi(s)=s$ 的最小正根。知道了绝后概率 $\pi$，则一个个体繁衍$n$代后，存活的后代数为$[m(1-\pi)]^n$。

分支过程是数学家提出的理想概率模型，然而我们面临的却是统计问题，即从还活着的后代去推断祖先。例如，为了破解“美国总统杰斐逊与黑人女奴海明斯私生子疑案”，需要比较杰斐逊与海明斯各自男性后代的Y染色体。因为杰斐逊和他弟弟都已经绝后，改用杰斐逊叔叔5名后代的血样，海明斯的后代人丁兴旺，采集了她家Tom的5名后代和Eston的1名后代的血样。杰斐逊的传记作者、耶鲁大学历史学教授哈韦尔经过大量的考证，认为海明斯的子女其实是她和杰斐逊总统的两个外甥所生。为此，又挑选了两个外甥祖父John Carr的3名后代的血样。总共14名男子的Y染色体检测结果如图5所示。

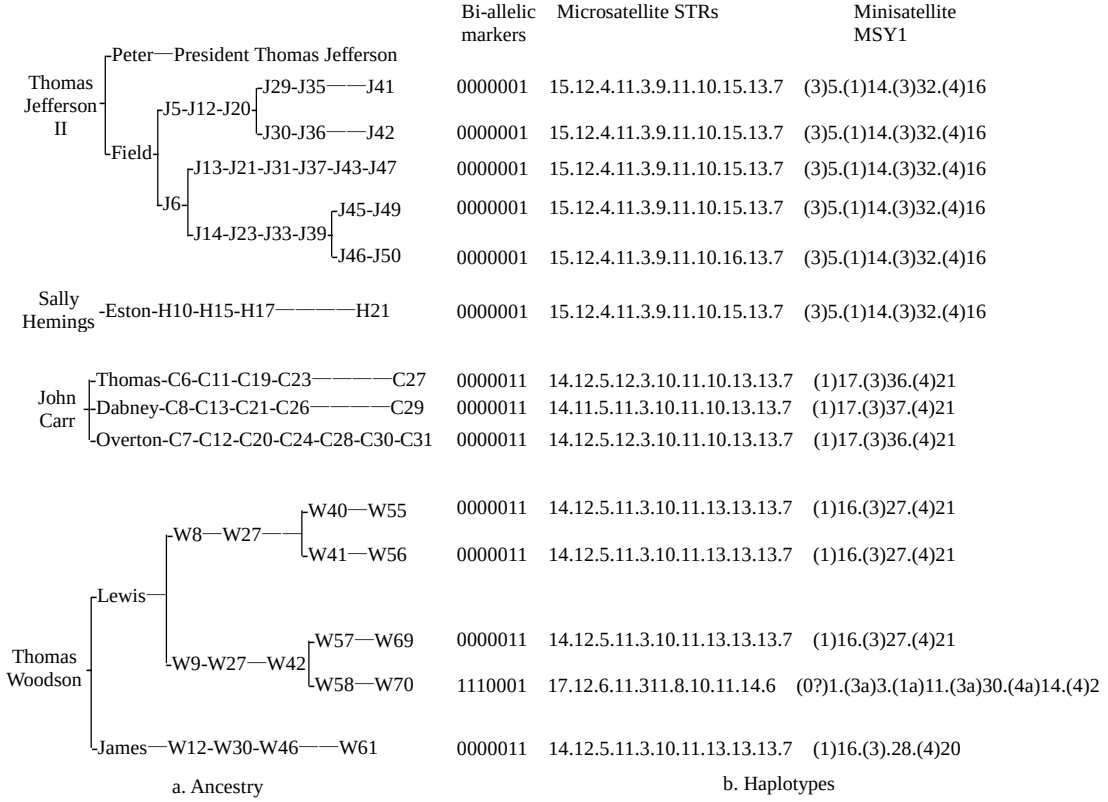


图5 相关14个后人的单倍群[2]

从检测结果不难看出：Eston后代和杰斐逊家族的遗传标记完全相同。也就是说Eston应该是杰斐逊家族的后裔，至于是不是杰斐逊的儿子，Y染色体检测并不能给出答案。因为两位的常染色体无法找到，自然也就没法做进一步的亲子鉴定。检测结果还告诉我们Tom确实是杰斐逊总统外甥的后裔，而且在Tom的一个后代中，因移花接木而改变了血缘。

作为发表在顶级权威刊物《自然》上的首篇血谱论文，这份报告说明科学研究应如何对待实验数据，不能随意删除不利数据，需要实事求是去反映世系传承中的各种“故事”。面对总统这样的名人，此举确实是勇气非凡！值得研究者认真学习。

上面是知道杰斐逊和海明斯家族的世系，比较他们之间有否亲缘关系，通过Y染色体检测就可确定。然而，对于不能完全确定家族世系的祖先，要获得祖先的遗传标记就会困难许多。如为了确定三国魏武帝曹操的单倍群，文献[3]从宣称是魏武帝曹操后裔的15个样本中检测到6个样本有单倍群O2-M268，而从其他姓氏(包括曹姓)后裔的446个样本中只检测到22个样本有。这样，就可以得到一张如表1所示的四格表[9]。

**表1 四格表**

| | 宣称曹操后裔 | 其他姓氏后裔 | 合计 |
|---|---|---|---|
| 单倍型O2-M268 | $n_{11}$=6 | $n_{12}$=22 | $n_1$=28 |
| 其他单倍型 | $n_{21}$=9 | $n_{22}$=424 | $n_2$=433 |
| 合计 | $n_{\cdot 1}$=15 | $n_{\cdot 2}$=446 | $n$=461 |

如何从这四格表获得有用信息？能否确定单倍群O2-M268显著高频出现？可用四格表的Fisher精确检验方法，以$n_{11}$为统计量，它服从超几何分布，可计算检验的$P=\frac{n_{1\cdot}!n_{2\cdot}!n_{\cdot 1}!n_{\cdot 2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}=9.323\times10^{-5}$值。又能否估算O2-M268是曹操单倍群的可能性？这就需要采用四格表的优比(Odds Ratio)检验方法，计算$\mathrm{OR}=\frac{n_{11}n_{22}}{n_{12}n_{21}}=12.72$值，由此得到可能性为92.71%(=12.72/(12.72+1)×100%)。注意，这只是统计结论，而且计算可能性的方法还值得商榷。毕竟文献[3]上没有，他们只在中文媒体上做了报道。

因此，为了做好《中华史氏大同谱》，本文认为更科学的提法是《中华史氏遗传网络》的编纂工作，首先需要基于各地纸谱，经过充分认证确定各地家族世系的可靠性。在此基础上，通过Y染色体检测，确定一支是否为另一支的外迁分支。对于时间跨度大，由于断代不能完全确定世系的分支，则需要用统计方法从多个宣称是其后代的分支确定祖先的单倍群。对于史氏宗族，除了杜陵史氏、四明史氏和真定史氏3大宗外，还有阳泉史氏、偃师史氏、乐陵史氏、城南史氏等等数百个分支，都需要确定当地祖先的单倍群，然后再将各地祖先的单倍群绘制成遗传网络图谱，方能得到正确的亲疏关系。显然，这是一项巨大的工程，不是短期内可以完成的工作。急于求成或急功近利都不是科学的态度。

## 5　结　语

从网络科学角度看，姓氏和血缘都是复杂的演化网络。如中华姓氏，先秦时期的《世本》只收入18姓，唐代初年编修的《大唐氏族志》收录293姓，北宋时编写的幼儿启蒙读物《百家姓》收录438姓，明朝人陈士元所著《姓觿》收录3 625姓，文献[4]统计的结果是7 327姓，而我国实际使用过和正在使用的姓氏约有20 000个。至于中华血缘，由于研究时间短，图3仅给出了单倍群主干，下游还有待发掘。

姓氏网络和血缘网络虽是两个不同的网络，但它们有着密切的相依关系，而要理清其中关系，则离不开族谱的收集、甄别，及Y染色体检测。一个宗族遗传网络既是姓氏网络和血缘网络的子网络，也是它们相依关系的完整记录，相对而言易于研究。

不过，这里介绍的血谱概念还有一个接受过程，目前检测的费用比较昂贵，检测精度也有待提高。然而在欧美，血谱研究正方兴未艾，每年都有上百万人接受基因测试。因其巨大医学价值和能够提供准确信息，国内很快就会赶上，不可小觑。

曾几何时，古人按手印，后来取指纹，现在已经录入身份证中。类似地，古人滴血认亲，后来输血需验血型，不久的将来人的血谱定会进入日常生活中。只要看看我们的儿孙玩起苹果产品那个熟练劲就可想而知，我们修族谱不要再贻笑大方了。

自从宇宙大爆炸，经历亿万年进化：形成物质，出现生命，演绎人类社会，涌现海量信息；与此同时诞生了：自然科学→生命科学→社会科学→信息科学→复杂性科学(理论，数学与系统科学；应用，工程与技术科学)的科学链条。在这些学科里，例如，星系起源、生命起源、人类起源、宗族起源都是复杂性问题，从中也可以看出问题的重要性。

另外，本文还参考了2012年60余个史氏后人的Y染色体检测结果，在此向这些先行者表示由衷敬意！

## 参　考　文　献

[1] WATTS D J, STROGATZ S H. Collective dynamics of small-world networks[J]. Nature, 1998(393): 440-442.

[2] EUGENE A, FOSTER M A, JOBLING P G, et al. Jefferson fathered slave's last child[J]. Nature, 1998(396): 27-28.

[3] WANG C, SHI Y, LI H, et al. Present Y chromosomes reveal the ancestry of Emperor CAO Cao of 1800 years ago[J]. Journal of Human Genetics, 2012(57): 216-218.

[4] LIU Y, CHEN L, YUAN Y, et al. A study of surnames in China through isonymy[J]. American Journal of Physical Anthropology, 2012(148): 341-350.

[5] 史美珩. 中华姓氏谱·史姓卷[M]. 北京: 现代出版社, 华艺出版社, 2002.
SHI Mei-heng. Chinese surname spectrum-SHI[M]. Beijing: Modern Press, Huayi Press, 2002.

[6] 王川, 汪秉宏, 许伯铭. 人口姓氏分布幂律行为的动力学模型探索[C]//第二届复杂性科学与经济物理学国际研讨会论文集. 桂林: [s.n], 2001.
WANG Chuan, WANG Bing-hong, XU Bo-ming. To explore the dynamics model of power-law behavior of surname distribution in population[C]//Proceedings on the Second International Symposium on Science of Complexity and Economic Physics. Guilin: [s.n.], 2001.

[7] 王传超, 严实, 李辉. 姓氏与Y染色体[J]. 现代人类学通讯, 2010(4): 26-33.
WANG Chuan-chao, YAN Shi, LI Hui. Surnames and the Y chromosomes[J]. Commun Contemp Anthropol, 2010(4): 26-33.

[8] KARLIN S, TAYLOR H M. 随机过程初级教程[M]. 2版. 庄兴元, 陈宗洵, 陈庆华, 译. 北京: 人民邮电出版社, 2007.
KARLIN S, TAYLOR H M. A first course in stochastic processes[M]. 2nd ed. Translated by ZHUANG Xing-yuan, CHEN Zong-xun, CHEN Qing-hua. Beijing: POSTS & TELECOM PRESS, 2007.

[9] 茆诗松, 王静龙, 史定华, 等. 统计手册[M]. 北京: 科学出版社, 2003.
MAO Shi-song, WANG Jing-long, SHI Ding-hua, et al. Statistics handbook[M]. Beijing: Science Press, 2003.

编　辑　蒋　晓