

重构家族血缘树

史定华

(上海大学数学系 200444)

DOI: 10.13405/j.cnki.xdwz.2015.03.002

这篇短文从人类寻根谈起, 介绍线粒体与 Y 染色体知识, 分子人类学的重要成果, 历史人类学的研究进展, 我们基于血缘距离网络的研究尝试, 直到采用大数据确定曹操遗传标记的社会物理学探索。因为是科普文章, 引用材料不另外注明。

一、线粒体与 Y 染色体

半个多世纪以前, DNA 双螺旋结构的发现和所有生命体都含有 4 种不同的碱基, 揭开了科学新篇章。4 种碱基分别是: 腺嘌呤 (Adenine)、鸟嘌呤 (Guanine)、胞嘧啶 (Cytosine) 和胸腺嘧啶 (Thymine)。分子生物学发现人类含有 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)。在传代过程中, 父母双方的染色体会打断而后重新拼接, 即重组。所以不同位置的突变之间毫无关系, 其组合完全没有意义。有没有不重组的区段? 有: 其一是线粒体 (mitochondria), 其二是 Y 染色体

(chromosome)。男性的线粒体无法传给后代 (线粒体位于细胞质里, 男女交配授精过程中, 精子细胞只有细胞核和卵细胞核融合, 精子的细胞质包括线粒体被排除在卵细胞之外), 所以人的线粒体都来自其女祖先, 自然不存在重组。Y 染色体只有男性独有, 更无法重组。它由父传子, 子传孙, 世代遗传决定了血缘关系。线粒体突变过快而且长度太短, 仅有 16569 个碱基对, 提供的信息量有限。然而, Y 染色体长达约 6000 万碱基对, 其中不重复的碱基对有 1000 万。

在 Y 染色体中, 一个碱基发生突变的结果, 称为单核苷酸多态 SNP (Single Nucleotide Polymorphism), 见图 1 上面序列碱基 A 变成 G, 这些突变构成的组合叫做单倍型 (haplotype) 或单倍



图 1 上面是 SNP 碱基突变; 下面是 STR 碱基重复

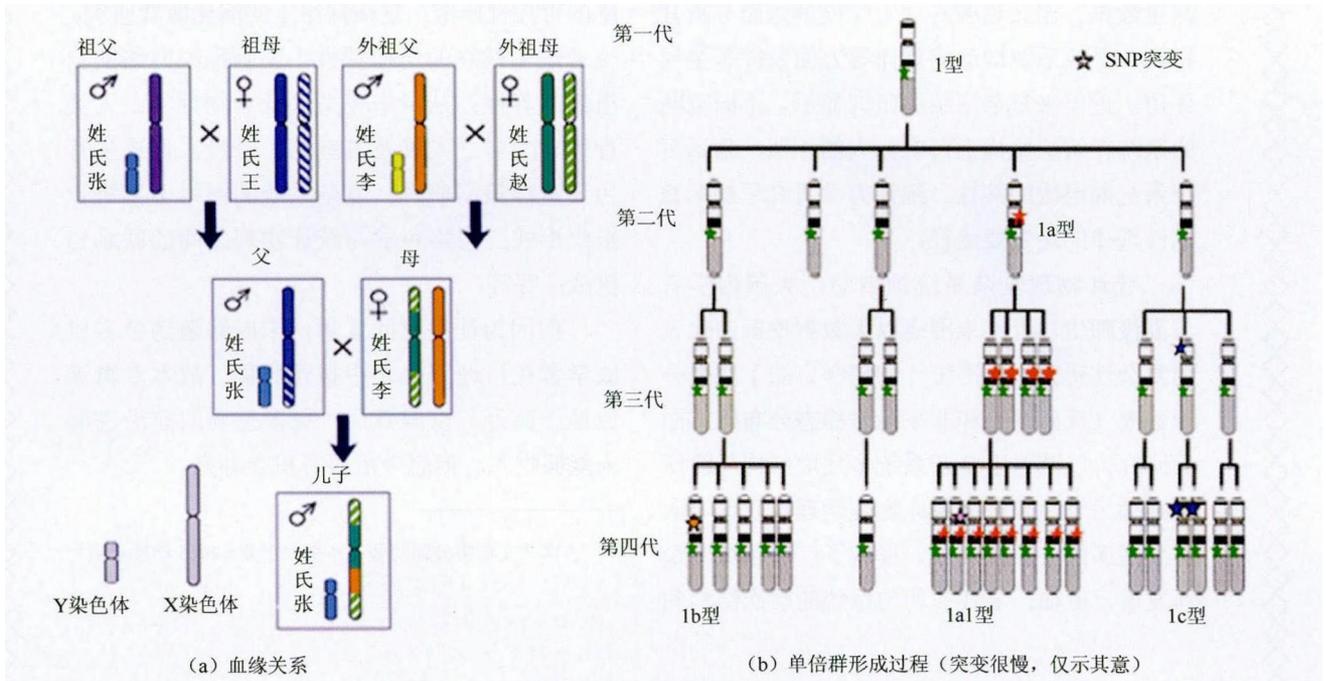


图 2 父系 Y 染色体遗传

群 (haplogroup)，可看作为遗传标记。而一个短串重复 STR (Short Tandem Repeat) 序列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，见图 1 下面序列 4 个碱基 gtat 组成的单位共重复了 12 次。不同的序列单位用英文加数字命名，如 DYS448, YGA-H4 等。同名的 STR 拷贝数不同则说明发生了突变。目前发现的 STR 不少，随着测序技术进步，测试的不同 STR 数目已由 11 个到 23 个甚至更多。某些 STR 由于突变快，可选择作为遗传时钟。

分子生物学告诉我们 Y 染色体传递决定了姓氏与血缘的关系，如图 2 (a) 所示，此图说明在男性后代中 Y 染色体是直接复制其父亲的，X 染色体则发生了重组。而 SNP 在 Y 染色体中传递形成单倍群的过程，其示意图如图 2 (b) 所示。由于 Y 染色体在复制过程中偶尔会发生突变，如图 2 (b) 中的第二代红色五角星单倍型 1 (a) 就是由一个碱基突变引起的，那么他的后代都会保持这个单倍型。在第四代中，又一个碱基突变，该单倍型就被命名为 1a1。

这样，自从人类进入文明社会有了姓氏之后，就存在两套家谱：纸质家谱和血缘家谱。然而，纸质家谱和血缘家谱并非完全一一对应，这是一个典型的复杂性问题。

二、分子人类学的成果

圣经故事记载：上帝用泥土造人，给他起名叫亚当 (Adam)。为了避免亚当孤独，又取下他一根肋骨为他造了一个配偶，取名叫夏娃 (Eve)。然后上帝在东方的伊甸，为亚当和夏娃营造了一个乐园。园中有许多动植物，其中还有一棵智慧树，并告诫他们智慧树上的苹果他们是不能吃的。然而，在蛇的诱惑下，夏娃偷吃了智慧果，并给亚当也吃了。从此他们有了智慧，明白男女身体有别，要与对方结合，二人成为一体。于是亚当和夏娃就成为了人类的始祖。

上帝为了惩罚亚当和夏娃偷吃智慧果，把他们赶出了伊甸园。后来，一个智慧苹果砸在牛顿 (Newton) 头上开创了现代科学，特别是物理学。不久前，乔布斯 (Jobs) 又啃了一口智慧苹果，发明了智能手机。终于在互联网时代人类开始寻找科学的“亚当与夏娃”。

科学的亚当与夏娃

1987 年，夏威夷大学的坎恩 (Cann) 等发现线粒体都源于非洲，这一学说被称为“非洲夏娃说”。这项研究揭开了用分子遗传学方法探索人类史前历史的序幕。从现代人的线粒体发现，现代人类起源于十多万年前非洲，那么其他大陆上更早时期的古人类必然是已灭绝的旁支。这与古人类学界原有的普遍观点不合，因此引发了很多争论。但是，对欧洲尼安德特人骸骨的古 DNA 分析，证实了他们确实与现代人没有传承关系，两个学界的认识也渐趋统一。

虽然母亲的线粒体能够传递给儿子和女儿，但传给儿子之后，就无法传递下去，所以只有母系才能一直保持。从现代女性追踪母系祖先的示意图见图 3，图中最低层是现在活着的人类，他们的线粒体都来自母亲，因此可以按照母系追踪上去，直到某个共同祖先，如图中第二排加黑圈的蓝色圆点就是追踪到的最近祖先“线粒体夏娃”。而红色、绿色和橘黄色圆点没有传承到现在，则表示灭绝。

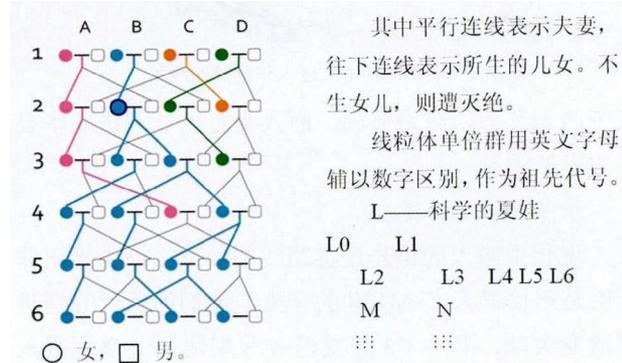


图 3 母系线粒体遗传

由于线粒体的特殊性质，它在群体间的分布格局较乱。但 Y 染色体上会不断产生新的 SNP 突变，2001 年，斯坦福大学的昂德希尔 (Underhill) 等在全世界人群中发现了 218 个 SNP 位点，构成了 131 种单倍群。这些单倍群构成的谱系与线粒体谱系的共同之处，也证明了现代人十多万年前发源于非洲。这种在 Y 染色体谱系中的结构表现得比线粒体清晰和稳定得多。由于 Y 染色体是纯粹父系遗传的，所以这种非洲起源假说被称为“非洲亚当说”。

根据单倍群编号 (英文字母加数字，可作为祖先名讳的代号)，可描述出现代人类的大致谱系，见图 4，

表 1 与杰斐逊问题有关检测报告

杰斐逊家谱	单倍群	STR 数据
Peter—President Thomas Jefferson	0000001	15.12.4.11.3.9.11.10.15.13.7
Thomas Jefferson II		
Field		
J5—J12—J20	0000001	15.12.4.11.3.9.11.10.15.13.7
J29—J35—J41	0000001	15.12.4.11.3.9.11.10.15.13.7
J30—J36—J42	0000001	15.12.4.11.3.9.11.10.15.13.7
J13—J21—J31—J37—J43—J47	0000001	15.12.4.11.3.9.11.10.15.13.7
J45—J49	0000001	15.12.4.11.3.9.11.10.15.13.7
J14—J23—J33—J39	0000001	15.12.4.11.3.9.11.10.15.13.7
J46—J50	0000001	15.12.4.11.3.9.11.10.15.13.7
Sally—Eston—H10—H15—H17	0000001	15.12.4.11.3.9.11.10.15.13.7
Hemings		
H21	0000001	15.12.4.11.3.9.11.10.15.13.7

注：这里单倍群用 0, 1 数字编码，完全相同；只测有 11-STR 数据，仅数字 16 不同。

了她家大儿子 Tom 的 5 个后代血样和最小儿子 Eston 的 1 个后代血样。另外又找到两个杰斐逊外甥祖父 John Carr 的 3 名后代取了血样，总共 14 个样本。

经过一批科学家的通力合作，测定终于取得成功。杰斐逊叔叔的 5 名后代和 Eston 后代的血样，其 Y 染色体遗传标记几乎完全相同，见表 1 测试报告。

结论：海明斯最小儿子 Eston 应是杰斐逊家族的后裔。不过，只有 Eston 后代一个样本，还可能存在其他变数。

理查三世的遗骸鉴定——母系血缘分析

2012 年 8 月，在英格兰中部莱斯特市中心的一个破败停车场地下，英国考古人员发现了一具人类遗骸。考古测定的数据与历史记载金雀花 (Plantagenet) 王朝 (1154 ~ 1485) 中的理查三世 (Richard III) 极为相似。为了科学地鉴定这具遗骸是谁？英国莱斯特大学研究人员从遗骸中提取了 DNA，并挑选了 7 名在世血亲族人进行血缘比对分析。2014 年 12 月 2 日，这项研究的结果发表在权威刊物《自然通讯》(Nature Communications) 上。

对在世父系后裔，由于理查三世没有子嗣，只得寻找与其拥有相同 Y 染色体的父系祖先亨利·萨默赛特 (Henry Somerset) 第五代博福特公爵 (5th Duke of Beaufort) 的后人。研究人员从亨利·萨默赛特在世后裔中选取了 5 个样本，分别命名为 SOM 1 至 SOM 5。

表 2 父系后裔 S-1 至 S-5 与遗骸 (Sk1) Y 染色体 23-STR 的测定结果

S-1	14	11	14	13	29	23	11	13	13	14	12	12	20	15	18	23	12	22	12	13	19	19	10
S-2	14	11	14	13	29	23	11	13	13	14	12	12	20	15	18	23	12	22	12	13	19	18	10
S-3	14	11	14	13	29	23	11	13	13	14	12	12	20	15	18	23	12	22	12	13	18	19	10
S-4	14	11	14	13	29	23	11	13	13	14	12	12	20	15	18	23	12	22	12	13	18	19	10
S-5	15	12	15	14	30	23	10	11	13	14	10	11	18	14	20	21	10	26	13	12	17	17	17
Sk1	15	13	14	13	30	22	10	11	14	16	10	12	22	15	18	21	11	21	10	12	16	15	12?

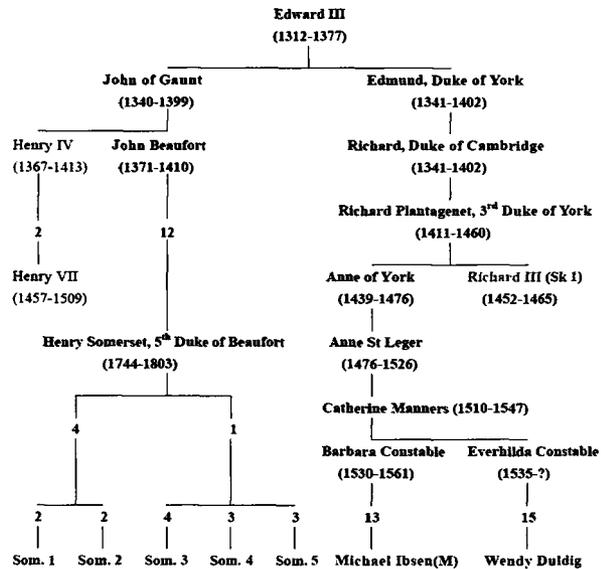


图 7 英国王室部分家谱 (其中数字是省略的世代数目)

对在世母系后裔，理查三世有一个同胞姐姐，名叫安妮约克 (Anne of York)，幸运的是安妮母系后裔家谱记录完整。为了比较遗骸与其母系后裔的线粒体 DNA，他们找到安妮的在世第 17 代后裔男子迈克尔·易卜生 (Michael Ibsen) 和第 19 代后裔女子温蒂·杜狄 (Wendy Duldig)，见图 7。

Y 染色体测试共出现了三种单倍群：S-1 至 S-4 是 R1b-U152，S-5 是 I-M170，Sk1 (遗骸) 是 G2-P287，说明遗骸与博福特公爵后人并不共祖。这里检测的是 23-STR，比通常的 17-STR 多检测了 6 个位点。博福特公爵 4 个后人 S-1 至 S-4 的 17-STR 完全相同，考虑 23-STR 也只有一个数字不同。因此这 4 人共祖没有问题，祖先单倍群为 R1b-U152，见表 2。博福特公爵后人 S-5 发生了血缘扰动，这与鉴定遗骸无关可以不考虑。另一个单倍群不同有两种可能：遗骸不是理查三世；是理查三世但发生了血缘扰动。因此需要首先确定遗骸是不是理查三世，这可以通过比较其母系后裔的线粒体 DNA 来确定。检测结果显示，只有一个碱基在温蒂身上发生了改变。根据统计分析，有

99.999% 的把握可以确定遗骸身份为理查三世国王。

现在的问题是，另一个血缘扰动发生在理查三世与博福特公爵中的哪一代上不清楚。这一研究引发了人们对英国君主血统是否纯正的怀疑声，我们将拭目以待。

四、血缘距离网络研究

上述美英两国的研究，家谱世系都不超过10代(约300年)，因此17-STR理论上不会突变，有突变也仅一个，直接比对即可。然而，中国家族历史源远流长，往往千年以上。复旦大学研究的魏武帝曹操距今近两千年，我们2012年开始研究的南宋四明史氏距今千年。

忠定王史浩的先祖辨——借助血缘距离

四明史家由郡里当差史简的遗腹子史诏开始发扬光大，在南宋王朝出过七十二进士(据考证灌水不少)，有着“一门三宰相，父子两封王”之美誉(见家谱中

表3

STR	DYS 19	DYS 389I	DYS 389II	DYS 390	DYS 391	DYS 392	DYS 393	DYS 437	DYS 438	DYS 439	DYS 448	DYS 456	DYS 458	DYS 635	Y-GA -H4	DYS 385a	DYS 385b
美德	15	13	18	23	10	14	13	14	10	11	18	16	15	20	11	13	13
MH	15	13	18	23	10	14	13	14	10	11	18	16	15	20	11	13	
JX	15	13	18	23	10	14	13	14	10	11	18	16	15	20	11	13	
WG	15	12	18	23	10	14	13	14	10	11	18	16	15	20	12	13	
CL	15	12	18	23	10	14	13	14	10	11	18	15	15	20	11	13	13
台湾	15	13	16	23	10	14	13	14	10	11	18	16	15	20	12	13	
南昌	15	12	17	23	10	14	13	14	10	11	18	16	15	19	12	13	
史家码	15	13	18	22	10	14	13	14	10	12	18	14	15	20	13	13	14
江苏	15	12	17	23	11	14	13	14	10	11	18	18	16	19	12	12	13
海南	16	13	17	23	11	14	14	14	10	13	18	15	17	21	11	13	18
湖南	15	12	16	24	10	15	14	14	11	12	18	14	16	21	12	13	14

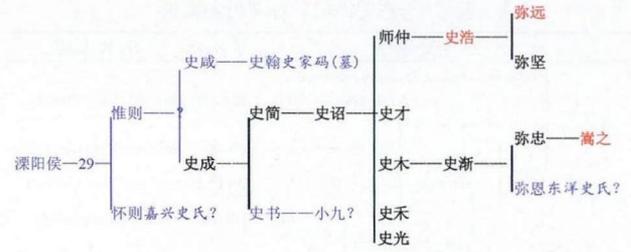
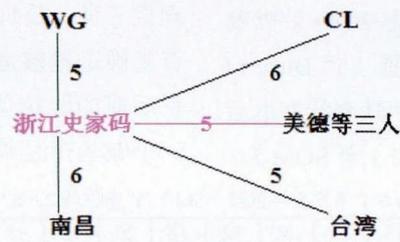
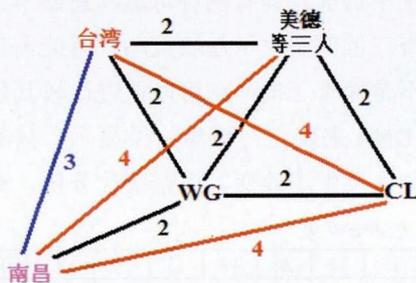
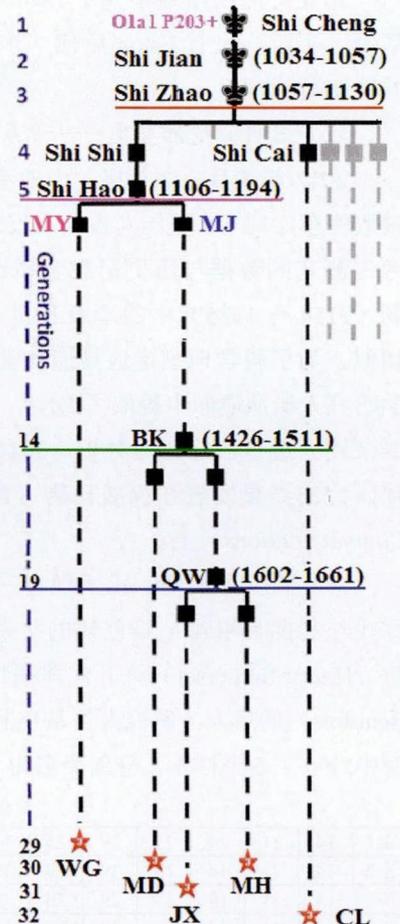


图8 四明史氏部分家谱

红色名字)。先贤们的墓葬群留下了“北有兵马俑，南有石刻群”的宝贵文化遗产，具有很高的艺术价值。

这个望族的始祖不能如此普通，于是四明史氏后代在续修族谱时就以五代后唐集贤院待制直学士史惟则为四明史氏始祖，并且认定惟则是溧阳侯30世。这引来后嗣不断的世系辩论而被称之为“千古谜团”。《古藤史氏宗谱》记载：“惟则于后晋天福四年(公元939年)辞官归隐，与怀则同时从终南山迁



注：这里 DYS389II 的数值减去了 DYS389I 的数值。11 个样本之间的血缘距离，即他们 STR 不同的数目，见两个网络图。谱系图中右边数字为代数，红色五角星为 5 名四明史氏的后代，他们分别是遗腹子史诏的第一支史师仲和第二支史才的后人。

浙江，后惟则又迁至鄞县。惟则生史成，史成生史简（1034～1057）；惟则的另一个孙子史翰成为史家码（墓）的始祖”。史成 980 年曾续修宗谱，他多大年纪才生史简？而续修族谱时则添枝加叶（见家谱中蓝色名字）让几乎江南史氏都出四明。

南宋史家作为千年望族，STR 既会有突变，又有完整家谱，因此这里采用 Y 染色体检测比对，计算血缘距离网络，再结合家族世系进行分析。但由于随机性，STR 可能会发生增减，所以 STR 如果超过 4 个不同，1000 年内就不太可能共祖。

四明史氏 5 名后代（为了保护隐私，除美德外，其余用拼音缩写）和 6 名疑似样本（采用出生地名）的单倍群都是 O1a1-P203，家谱和血缘比对分析如表 3。

结论：确定四明史氏 Y 染色体单倍群为 O1a1-P203，史家码与四明史氏不可能在南宋期间共祖，排除了江西、湖南、海南等地史氏来自四明的传说。从血缘上看，四明史氏也不可能是溧阳史氏（O3）的后代。史简的父亲史成和惟则的儿子史成，可能此成非彼成。

为了验证分析结果的可信性，又对 4 名四明史氏后代和 4 名疑似样本进行了全测序，即测试 Y 染色体中不重复 1 千万碱基对的 SNP 突变。全测序结果显示：4 名四明史氏后代共祖时间不超过 800 年，而 4 名单倍群相同的疑似样本共祖时间均都超过 1500 年以上。

南宋宰相史嵩之夫妇的墓地 2012 年先后被发现。3 月宁波文保所对史嵩之墓进行了考古发掘，发现遗骨在金丝楠木的棺木中，头骨保存完好。但由于对科学的认识所限，没有进行古 DNA 检测。史嵩之是遗腹子史诏的第三支史木的后人，若遗骸测试数据能够一致，则通过比对分析在世后裔血缘距离辅以 Y 染色体全测序方法意义重大。错失良机实在遗憾！

五、大数据推断单倍群

曹操不仅是“挟天子以令诸侯”的一代风云人物，曹氏父子还是中国文坛举足轻重的名人。曹操的身世一直众说纷纭：陈寿在《三国志》记载，曹操是西汉第二任相国曹参的后代，曹操自称其祖上可远溯至古曹国。因曹操的祖父曹腾在东汉为宦官之首，曹操的父亲曹嵩是曹腾的养子，故曹操的政敌袁绍在攻击曹

操的檄文中说“父嵩乞丐携养”。还有人说，曹操的父亲曹嵩是从夏侯氏家抱养来的。晋灭魏时曹氏遭追杀，有操姓认为是从曹姓避祸改姓而来，这些历史疑案如何科学回答？

魏武帝曹操的单倍群——父系血缘推断

2009 年，河南安阳对外宣布发现曹操墓，此消息一出即引发轰动，亦引起争议。随后，复旦大学现代人类学教育部重点实验室宣布，拟用 DNA 技术开展对曹操家族的血缘研究。

曹操是生活在近两千年前的历史人物，寻找他的遗传标记似乎遥不可及。课题组随即开始另辟蹊径：能否用现在人的遗传标记反推曹操的，从而破解曹操的身世之谜呢？

要把曹操后人与两千年前的曹操进行“亲子鉴定”，关键是要有可靠的样本。首先，课题组专家对全国各地 258 个曹姓族谱做了全面的梳理研究，并与史书和地方志对照，找到曹氏迁徙的可能线索。再从中筛选出 8 支持有族谱、经过史料分析有一定可信度的曹氏族群。然后，课题组在全国范围内招募志愿者，采集了 79 个曹姓家族的 280 名男性和 446 个包括夏侯、操等姓氏男性的静脉血样本，后来得得到样本总量超过 1000 例。通过对这些样本进行 SNP 检测，最终发现其中 6 个家族具有同样的 O2*-M268 型单倍群，他们分别为安徽舒城县（仪壹堂）、安徽舒城县（七步堂）、江苏盐城（绣虎堂）、辽宁、安徽、湖南曹氏族群。其交汇点在 1800～2000 年前——这正是曹操生活的年代。这个单倍群在全国人口中只占到 5% 左右。这就意味着他们假冒的可能性只有千万分之三，因此从法医学上可认定他们是曹操的后代。该项研究结果 2012 年发表在《人类遗传学报》上，下面是针对宣称曹操后裔和非曹姓后裔比对的统计分析报告。

表 4 其中 15 人宣称是曹操后裔和 446 名其他姓氏后裔的单倍群分析

	宣称曹操后裔	其他姓氏后裔	合计
O2*-M268	$n_{11}=6$	$n_{12}=22$	$n_{1.}=28$
其他单倍群	$n_{21}=9$	$n_{22}=424$	$n_{2.}=433$
合计	$n_{.1}=15$	$n_{.2}=446$	$n=461$

从这张表格利用统计检验方法能够获得如下有用信息：单倍群 O2*-M268 显著地高频出现，它是曹操的单倍群的可能性概率高。

光从后代 SNP 组成的单倍群分析, 还不能把曹操的遗传标记类型完全确定下来, 于是再做曹操祖先的研究, 让两头都能对接上。原来, 在二十世纪六七十年代对曹氏家族墓葬曾有挖掘。安徽亳州是曹操的老家, 它的宗族墓都埋在那里, 其中有一个“元宝坑一号墓”, 里面发现了两颗牙齿。结合墓室内中央位置的墓砖铭文“河间明府”等, 可以肯定两个牙齿确实来源于曹操叔祖父河间相曹鼎。当时这两颗牙齿被认为大有用处, 所以被放在库房里等待日后研究, 现在是时候了。课题组带回一颗保存较好的牙齿到实验室开展古 DNA 测试。通过严密的科学程序提取该牙齿中的古 DNA, 经过 6 次同样严格的反复测试, 每一次间隔一个月, 发现该牙齿的古 DNA 中 Y 染色体类型以大概率就是之前发现的 O2*-M268。

在全国 700 余万曹姓中, 除了曹操后代, 还有另外四大分支: 一是汉代宰相曹参的后代, 二是先秦古国邾国的后代, 三是江东越国曹姓后代, 四是满族入关后有部分满族人改姓为曹。

在确认曹操后代的同时, 他们还用同样方法验证了汉代丞相曹参的家族遗传基因与曹操家族遗传基因

没有关系, 从而证明曹操是曹参后人说法有误; 同时对民间传说操姓是曹操后代避祸改姓而来、曹操是夏侯氏抱养而来等说法, 研究证明曹操家族与这两个姓的血缘没有明确关系、家族遗传标记不一致, 因此说法也不准确。河南安阳发现的曹操墓, 因为不愿意做古 DNA 检测, 至今没有定论。

复旦大学这项研究在遗传学和历史学相关的理论和研究方法方面, 都有很多新的突破。他们的研究似乎可看成是大数据与社会物理学的一个探索方向。

六、结语

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, 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, 为解开历史疑案另辟蹊径。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。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。希望这个介绍能够引起我国研究复杂性科学的年轻学者们关注, 不妥之处还望赐教。

DOI: 10.13405/j.cnki.xdwz.2015.03.003

科苑快讯

为什么水星如此黑暗

水星荒凉沉寂的表面与月球类似, 而科学家一直困惑的是, 为什么水星的反射光远远暗于月球。以前的研究表明, 彗星、小行星和其他小天体在相当近期撞击水星表面后, 物质的反射光亮度是月球的 2/3。研究者说, 对于这种低反射率, 最初的解释是因为这些矿物中富含铁元素, 强烈吸收某些特定波长的光, 但水星却并非如此。这是因为水星在一个特定波长的亮度表明其表面岩石中只有不到 3% 的铁。一个研究组最近提出, 这其实归因于另一种元素——碳。

据估计, 彗星质量的约 18% 是碳, 它是这些碳元素的主要来源。但更大的来源则可能是微小的富碳陨石, 它们来自彗星尘埃, 水星受到的类似碰撞常常大约是月球的 50 倍。研究者的论文已发表于《自然地球科学》(Nature Geoscience)。撞击的高温使这些物质融化成玻璃样, 碳从这些微小陨石中蒸发出来,



形成无定形碳(类似于烟灰)、石墨和纳米钻石, 尽管阳光所及之处炽热如炼狱一般, 这些物质还是在这个没有空气的星球表面沉积下来。所以不难想象, 一个被烟灰覆盖的星球远比想象的更为黑暗。

(高凌云编译自 2015 年 3 月 30 日 www.sciencemag.org)