追踪曹操的基因

李 辉

从分子人类学方面来追寻曹操的基因有一个前提，就是我们已经知道曹操有很多后代，而且知道他的后代分布在什么地方，我们就可以按图索骥地去追踪曹操的后代，人海茫茫中进行“人肉搜索”。到目前为止，我们已经得到非常可喜的成果。这个研究过程具有什么原理？为什么我们能够追踪？今天讲座主要就是为大家介绍一下这方面的内容。

首先介绍一下，我是 1997 年加入金力教授的（复旦大学生命科学学院）人类学实验室的，1998 年开始项目的工作，到 2005 年，我调查了 300 多个民族群体。在 2005 年获得了人类生物学的博士学位，之后四年一直在耶鲁大学从事人类遗传学的研究，2009 年我回到了复旦，回到了人类学的岗位上。

回国后，我做的第一件事情就是改造了古 DNA 实验室。近年来古 DNA 技术突飞猛进，它的实验要求在不断地变化提高，所以我们必须与时俱进，跟上国际上的技术发展变革。所以我回来以后做的第一件事情，就是将这个实验室改造成了国际一流的古 DNA 实验室。我们复旦已经有了十几年的古 DNA 研究经验，“楼兰美女”

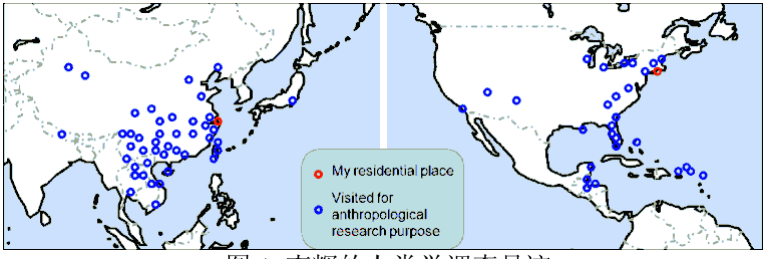


图 1 李辉的人类学调查足迹

的采样分析就是在这里做的。后来，国外几位著名的考古人类学家来参观考察了我们的古 DNA 实验室，确认我们的实验室达到了国际一流的水平。

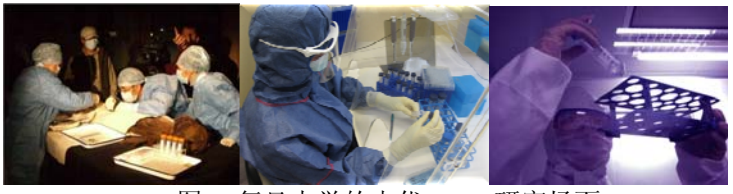


图 2 复旦大学的古代 DNA 研究场面

第二件事情，巧了，曹操的 DNA 鉴定问题来了。所以新的实验室就拿曹操“开刀”，所以我们就开始了曹操 DNA 鉴定的项目。问题很多，就像韩昇老师讲的，为什么会有两个女子尸骨存在？为什么不在主墓室中？他们到底是谁？盗墓贼是不是把骨头盗走了？

那么古代墓葬里面的 DNA 到底能够鉴定么？这是一个很大的问题。我们刚刚发布新闻说我们准备鉴定曹操遗骨真假的时候，很多网友、很多专家都出来问。甚至我们还没有鉴定，他们就出来“辟谣”。有考古专家表示：“考古不必要以 DNA 为佐证”，骨骼现在也是重要文物，为了鉴定他身份在上面钻个洞取 DNA 本身就是一个

图 3 “曹操墓”中发现的人骨遗骸

破坏文物的事情，好像完全没必要。如果为了鉴定他身份，完全可以从墓的形制等等鉴定。所以考古学不需要用这些新的技术新的方法作无谓的事情。但是大多数网友不买账，表示“DNA 是最权威可信的证据，如果确认这个墓是曹操的话，为什么不作（鉴定）呢？”所以要解决遗骨的身份问题，只能通过 DNA 的手段，没有什么其他手段可以来做。当然，证明了这个骨头不是曹操的，也不能确认这个墓不是曹操的。比如这个墓里面随便躺一个人，或许是盗墓贼。反过来，如果证明了这个骨头是曹操的，那么这个墓是曹操的，就基本上没有问题了。当然这个墓里面现在躺的人太多了，有了三个人了，他/她们分别是谁，都是需要 DNA 解决的问题了。

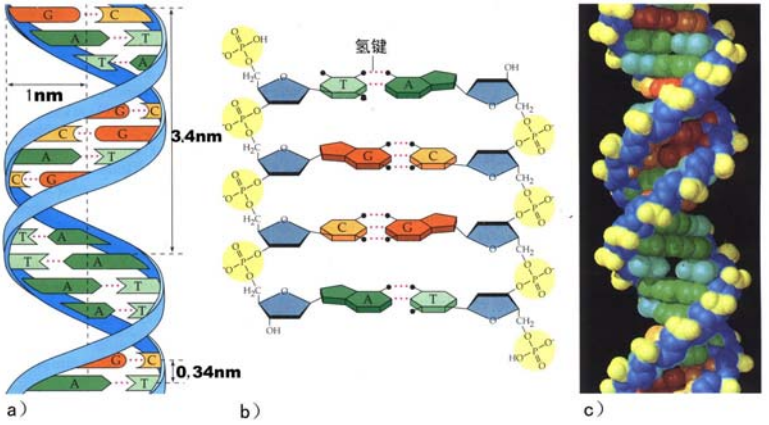
实际上通过我们的 DNA 宣传，现在有部分考古学专家有了针对性的回应，他承认 DNA 鉴定是一个很好的方法，但称 DNA 鉴定曹操有两大难题。一个是只有在专业的实验室里面测试之后才能有科学的、客观的答案，才能知道这骨骼能不能作 DNA 鉴定。当然我们从照片上看到，这个骨骼保存得很好，没有受酸性物质的污染，没有被什么金属锈迹浸染过，而且一直保存在比较阴冷的环境下，所以做古 DNA 的鉴定条件还是相当成熟的。另外一点，必须要找到确定的曹操后裔，成功地提取遗传数据作为参考。这两点好像都

是给我们的工作做铺垫用的。实际上这两大难题在其他学校可能是难题，但是在复旦绝不是难题，因为我们有国际一流的专业古 DNA 实验室，这解决了第一个难题。第二个，如何找曹操的后代，魏晋南北朝史的优秀历史学家韩昇教授已经用历史学方法找到了。曹操后裔在哪里，我们都抽一遍血就都可以验证了。

如果要鉴定曹操的遗骸，确定他的身份，必须要有一个参照。我们知道，就算作亲子鉴定，首先必须要有亲和子互相作参照。所以鉴定曹操的骨骼，必须要有一个 DNA 的参照。那么这个参照在哪里？当时就有人提出，用曹植墓中的骨头作为参照。1951 年，在山东东阿发现了曹植墓，出土了几十块曹植的遗骨，那我们用这些曹植遗骨去对比疑似曹操的骨骼的话，是不是就能确定曹操的身份？实际上这种想法是建立在亲子鉴定的法医学方法基础上面，亲子鉴定都是父子之间遗传关系的鉴定。这个其实是法医学做的事情，那法医学如何作亲子鉴定，我们必须从 DNA 的基本概念说起。

DNA 是什么？我们知道，DNA 它是一种大分子的双螺旋链，由两条很长的分子链以配对的方式结合在一起。分子链上面的基本单位是四种碱基。碱基到底是怎么构成，比较复杂，我们这里就不详细展开了。我们只要知道 DNA 是一种密码链，这种密码链上面有四种单位，而这四种单位通过排列组合成为我们的生命密码，这种密码链在我们的细胞里面以两种方式存在。

一种叫染色体。在我们细胞核里染色体一共有 46 条。染色体的空间结构极其复杂，由 DNA 的链通过双螺旋，形成二级结构；再通过缠绕在组蛋白上，形成第三级结构；缠绕完以后，再进行一次螺旋，成为四级结构；之后再进行一次螺旋，最后扭曲成一团，形成染色体。所以染色体是一种高度浓缩、高度缠绕的复杂的分子结构。



注：a）蓝带表示DNA双螺旋的磷酸骨架，双链以右旋方式互旋。b）DNA双螺旋的两条单链反向平行，碱基对以氢键相连。c）碱基对的堆积力和Van der Waals静电引力维持DNA的稳定结构。

图 4　DNA 的分子结构模型和四种碱基结合方式

每条染色体没有复制过的话，都有两条链，但两条都只有一套密码。在细胞里面，它的长度被压缩了 8000 倍到 10000 倍。所以每条染色体都非常长，里面的密码非常多，所有密码加起来总共有 30 亿个。

另外一种 DNA 存在于细胞里的线粒体，线粒体是一种单独的细胞器，它有自己的编码，它的编码形成一个环，形成环以后它的结构就非常稳定。古 DNA 研究最早从线粒体 DNA 开始，为什么？就是因为它比较稳定，保存的时间可就非常久。

DNA 编码的基本结构叫做碱基对，我们整个基因组有 30 亿个碱基对，组合成 23 对染色体和一个线粒体，线粒体是 1.6 万个碱基对。人和人之间 DNA 编码的差异占整个基因组的比例非常小，但总量加起来非常大，所以造成了人和人之间的巨大差异，当然人科

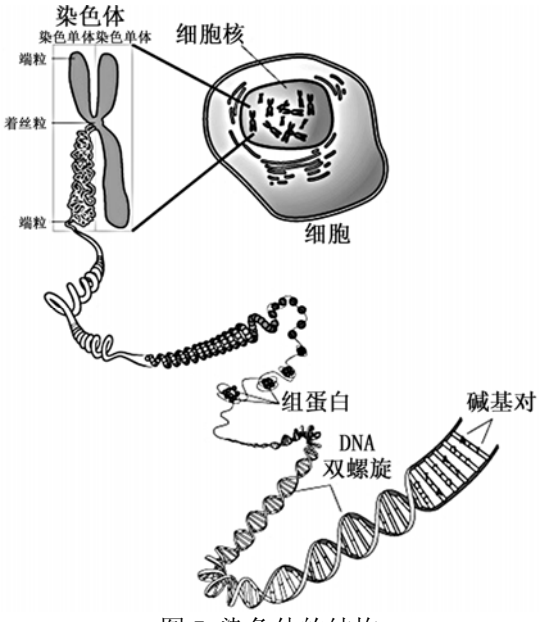


图 5 染色体的结构

现存的两个物种——现代人和黑猩猩之间的差异更大。

DNA 上的差异有些什么类型？差异主要有两大类。有一种叫做短串联重复，缩写叫 STR，如果大家看过分子生物学的有关书籍，经常会看到。在基因组里面一个固定的区段，由几个碱基组成一个单位，不断地重复，重复数在人和人之间是不同的。打个比方，就像一辆火车，它有不同的车皮数，每节车皮都是一样的。这辆火车可以装 10 节车皮，也可以装 12 节车皮，每个人对应的区段上装的数量都不一样，这就构成了重复数的差异。另外一种叫单核苷酸多态，SNP，这个更流行。我们研究曹操的基因，关键就是要找到相应的 SNP。SNP 是固定的点上面的一种碱基类型的差异，从一种类

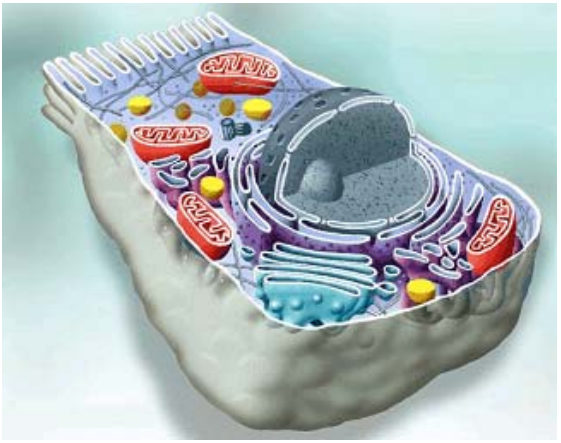


图 6 细胞中有多个线粒体（扁核桃状，视图中有四个）

图 7 两种 DNA 差异类型：STR 短串联重复和 SNP 单核苷酸多态

型变到另外一种类型，这个点的旁边的其他序列都是一样的，但就这一个点上会发生变化，比如人类基因组中这个点原始型是 A，有些人突然变成了 G，这种变化会造成特定点上的性质的彻底改变。打个比方，这就像两辆卡车。一辆卡车装的是可口可乐，一辆卡车装的是七喜，这是完全不同的两个类型。只要看一个点，就知道这两辆车完全不同。所以 SNP 这种一个碱基的差异，就是非常明显的一个变化。STR 和 SNP，这就是我们基因组中差异的最主要的两大类型。

那么法医学用什么基因组差异来分析呢？法医学用 STR 作为亲子鉴定的材料，为什么呢？因为 SNP 在一个点上有两种类型，要么是 A，要么是 B，两个人比较这个点只有两种可能性，在人群中有很多人是 A，有很多人是 B，要区分个体就很不容易，需要检测数百个 SNP 才能把每一个人都区分出来。如果用 STR 的话，情况就不同了。同一个 STR 点上有不同的拷贝数，就有不同的类型，可能有几十种类型。如果用数个 STR 来定义一个人，就很难有两个人在所有的 STR 上都一模一样。我们看这个表，法医学经常用作亲子鉴定的 STR 的点一共有 15 个，这 15 个点就是很多年前金力老师确立的。当时发了文章之后根本没有想到会被作为国际法医学标准。结果美国联邦调查局看到金老师的文章后就作为法医学的标准，然后就成为国际流行的标准了。人和人之间，一个点就会有很多很多类型，两个人如果要在 15 个点上面类型完全一样的话，这个概率是非常小的，小到一千万亿分之一。每个人每个点都有两种类型，因为每个人的基因有一半来自父亲，一半来自母亲，所以同样的一个点上每个人都有两种类型。那么亲子鉴定怎么做呢？把 15 个点的拷贝都测出来。（见图 8）举个例子，父亲在第一个点上有两种类

表 1　拉萨藏族中法医学 STR 类型的频率分布

重复数	D19S433	vWA	TPOX	D18S51	D7S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D5S818	FGA
6	-	-	-	-	-	-	-	-	-	0.104	-	-	-	-	-
7	-	-	-	-	-	-	-	-	-	0.307	-	0.025	-	0.010	-
8	-	-	0.599	-	-	-	0.163	0.005	-	0.055	0.168	0.267	-	-	-
9	-	-	0.144	-	-	-	0.099	0.545	-	0.475	0.069	0.129	-	0.056	-
9.3	-	-	-	-	-	-	-	-	-	0.059	-	-	-	-	-
10	-	-	0.005	-	0.114	-	0.193	0.203	-	-	0.223	0.257	-	0.147	-
11	-	-	0.233	-	0.010	-	0.287	0.015	-	-	0.218	0.253	-	0.409	-
12	0.045	-	0.020	0.015	0.119	-	0.223	0.243	-	-	0.238	0.064	-	0.232	-
13	0.005	-	-	0.287	0.287	-	0.030	0.416	0.005	-	0.064	0.005	-	0.141	-
13.2	-	-	-	0.005	-	-	-	-	-	-	-	-	-	-	-
14	0.297	0.208	-	0.183	0.198	-	0.005	0.055	0.020	-	0.020	-	-	0.005	-
15	0.025	0.025	-	0.119	0.208	-	-	0.010	0.371	-	-	-	-	-	-
16	0.327	0.203	-	0.119	0.059	-	-	-	0.337	-	-	-	0.010	-	-
17	0.139	0.238	-	0.069	0.005	-	-	-	0.228	-	-	-	0.035	-	-
17.2	-	-	-	0.005		-	-	-	-	-	-	-	-	-	-
18	0.050	0.208	-	0.050	-	-	-	-	0.040	-	-	-	0.124	-	0.030
19	0.074	0.114	-	0.050	-	-	-	-	-	-	-	-	0.134	-	0.051
19.2	-	-	-	0.005	-	-	-	-	-	-	-	-	-	-	-
20	0.010	0.005	-	0.059	-	-	-	-	-	-	-	-	0.144	-	0.030
21	0.030	-	-	0.025	-	-	-	-	-	-	-	-	0.074	-	0.005
21.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.056
22	-	-	-	-	-	-	-	-	-	-	-	-	0.045	-	-
22.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.147
23	-	-	-	0.010	-	-	-	-	-	-	-	-	0.262	-	0.005

重复数	D19S433	vWA	TPOX	D18S51	D7S1179	D21S11	D7S820	CSF1PO	D3S1358	TH01	D13S317	D16S539	D2S1338	D5S818	FGA
23.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.237
24	-	-	-	-	-	-	-	-	-	-	-	-	0.119	-	0.015
24.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.212
25	-	-	-	-	-	-	-	-	-	-	-	-	0.040	-	0.010
25.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.131
26	-	-	-	-	-	-	-	-	-	-	-	-	0.005	-	-
26.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.035
27	-	-	-	-	-	-	-	-	-	-	-	-	0.010	-	0.025
28	-	-	-	-	-	0.030	-	-	-	-	-	-	-	-	0.005
28.2	-	-	-	-	-	0.030	-	-	-	-	-	-	-	-	-
29	-	-	-	-	-	0.302	-	-	-	-	-	-	-	-	-
30	-	-	-	-	-	0.223	-	-	-	-	-	-	-	-	-
30.2	-	-	-	-	-	0.055	-	-	-	-	-	-	-	-	-
31	-	-	-	-	-	0.089	-	-	-	-	-	-	-	-	-
31.2	-	-	-	-	-	0.069	-	-	-	-	-	-	-	-	-
32	-	-	-	-	-	0.025	-	-	-	-	-	-	-	-	-
32.2	-	-	-	-	-	0.124	-	-	-	-	-	-	-	-	-
33	-	-	-	-	-	0.005	-	-	-	-	-	-	-	-	-
33.2	-	-	-	-	-	0.040	-	-	-	-	-	-	-	-	-
34	-	-	-	-	-	0.005	-	-	-	-	-	-	-	-	-
34.2	-	-	-	-	-	0.005	-	-	-	-	-	-	-	-	-

型，12 个拷贝和 18 个拷贝，第二个点上有 16 个拷贝，16 个拷贝，依次把父亲的 15 个点的数据都读出来。他儿子第一个点上是 18、15，我们猜测这 18 是从父亲来的，而第二个点上是 16、14，那这个 16 就是来自于父亲的 16，这个位点也对上了，另外一个 14 就是从母亲来的，当然如果我们有母本的话就更清楚了。所以两个人对比就是这样做的，那第三个点也对上了。往后数，注意这里有一个点没有对上，13 和 14。但 13 和 14 就差一个数，很可能是新的突变造成的。因为 STR 的突变率非常高。但 STR 的突变都是一步一步变的，每次加一个或者减一个拷贝，所以这个点完全有可能是儿子的一个偶然的基因突变，但这种突变是偶然的低频率的，所以不能在 15 个点里面出现好几个，那就不对了。因为突变的概率还是非常小的，每个点上大概是三百分之一，所以不能每个点都突变了。这样我们通过比较的话，就可以很明确地看出来这两个人之间的关系，这样法医学就确定了这两个人的亲子关系。

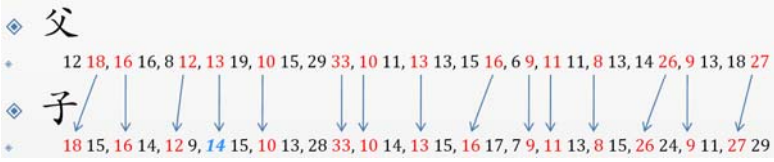


图 8　父子关系的法医学 STR 确认

那么我们能不能拿这种原理来分析曹操和曹植之间的父子关系呢？看上去好像可以，但是问题在于他们的样本是古 DNA，是古代的东西，与现代人的 DNA 性质不同。古 DNA 存在一个问题，它们是破碎的、断裂的；而现代人活体中的 DNA 的完整形态，每条染色体是连续的。古 DNA 为什么会断裂，断裂会造成怎么样的后果呢？我们看看古 DNA 是怎么断裂的。在活体细胞里面，我们有细胞作为保护，如果断裂的话，细胞中还有机制可以把它修复，把它补齐，所以不会受到氧化水解反应的影响，不会受到细菌的侵袭。

但是人一旦死亡之后，细胞破掉了，DNA 暴露出来了以后，就不断受到细菌侵袭，细菌会把 DNA“咬”断掉；或者受到光化作用、氧化作用、水解作用之后，DNA 不断地裂开。所以我们看到在古 DNA 里面，存在这样一种破碎的情况。

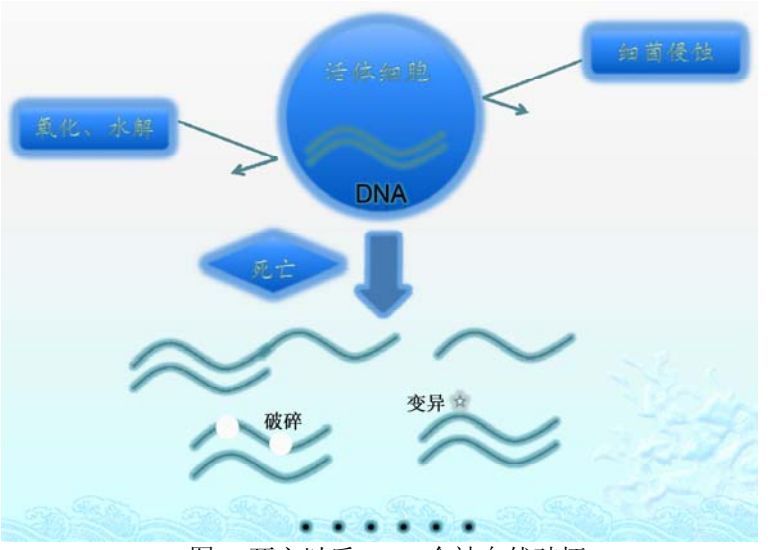


图 9 死亡以后 DNA 会被自然破坏

图 10 来自 George Poinar 在 2006 发表的一篇文章，他对古 DNA 里面的基因片段进行分析，发现大部分片段是大约 84 个碱基对的长度，还有很多片段是 151 个碱基对的长度，超过 500 个碱基对的长度的片段非常非常少。所以我们知道，古 DNA 都是非常短小的片段，都是破碎的、断裂的。

那么破碎之后，STR 就变成什么样子了呢？还能够检测清楚

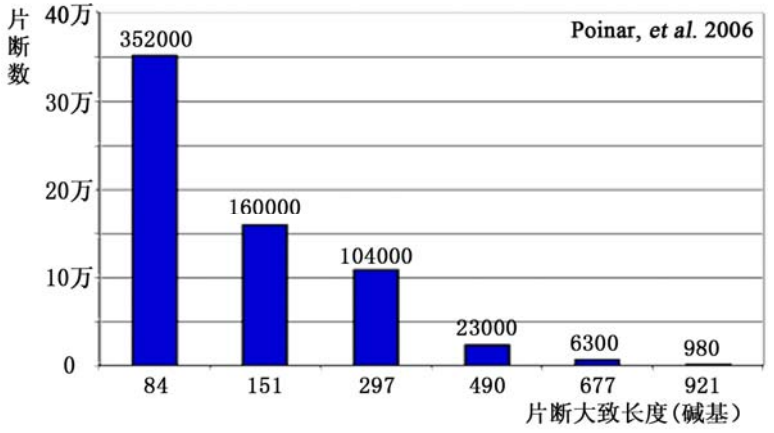


图 10 一份古代遗骸中得到的不同片段长度的 DNA 的数量

么？打个比方，原来这几列 STR 火车的车厢是完整地连起来的，现在全部断掉，这列火车它原来到底有几节车厢，我们就不知道了，那列火车到底有几节车厢我们也不知道。也就是说，破碎的古 DNA 中分析 STR 读不出 STR 的正确长度，所以用 STR 来做古 DNA 检测相对来说是不可靠的。虽然有很多文章对古 DNA 的 STR 做过尝试，但是结果还是不能令人信服。

那么我们能不能对比古代 DNA 样本的 SNP 呢？我们知道，古 DNA 断裂的话，SNP 还是能够看到的。因为每一个 SNP 是一个点上的差异，只要这个点存在，我们就能看出差异。就像前面的图中比喻的那样，知道这个车是哪个公司的，这很容易看。比如这条 DNA，这个地方是 A，那么只要这个点在，它的类型就存在。无论怎么断，这个点还是能想办法找到的。所以用 SNP 来检测古 DNA 的话，其可信度非常高。但是问题是 SNP 只有两种类型，我们要确定两个人的父子关系的话，只检测一个 SNP 或者检测十几个 SNP

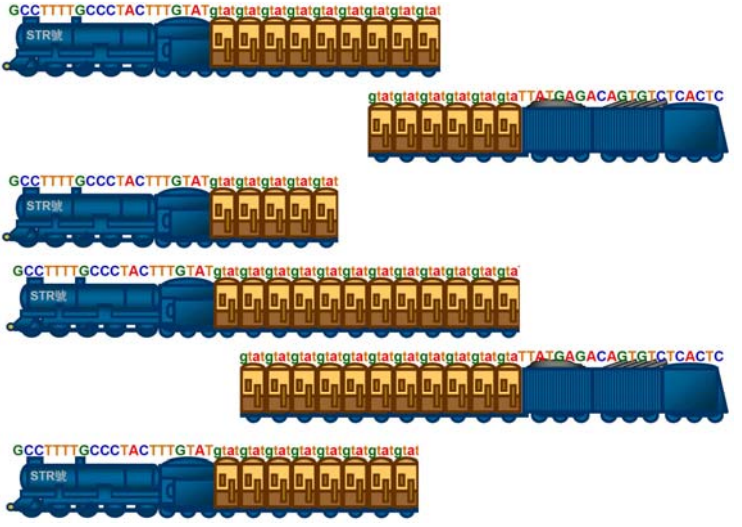


图 11　破碎的古代 DNA 中的 STR 的样子

是不够的。因为每个 SNP 只有两种类型，任何人在一个常染色体上十几个 SNP 相同的概率是非常非常大的，他们不是父子关系而能对上的概率也是非常大的。所以我们必须检测数百个 SNP 位点才可能确定这两个人是不是父子关系。但是这个成本非常高，谁来支付检测所需要的几十万元钱呢？

更何况曹植的骨头不知道去哪里了，因为当时曹植墓所属于的地区是平原省，解放时的省区划分跟现在不一样，当时骨骼挖掘出来之后都放在了新乡县，当时平原省的省会。当平原省撤销以后，很多文物都转到国家博物馆。但是这个骨骼放到哪里就没有人知道。所以这个亲子鉴定就没有办法做了。

所以寻找曹操的基因对照，鉴定这是曹操的骨头，只能是通过活着的曹操的后裔去找这些基因对照。但是我们发出了“人肉搜索”

图 12　破碎的古代 DNA 中的 SNP 的样子

追踪曹操的基因的消息之后，有很多“专家”提出质疑，包括北大生命学院的一个研究生提出了非常“专业”的质疑：曹操的骨骼经过一千多年的侵蚀，它是不是还能够检测出 DNA？按照他的想法，那应该是不可能的，是天方夜谭。那对于这个问题我估计他对于国际上古 DNA 检测的文章就从来没有看过。然后第二个疑问是每个人的基因组都来自于父亲和母亲双方基因组，基因重新组合，混一混再遗传下去，经过一千多年将近两千年的稀释或者混合，现在混成什么样子了，应该检测不出来了。这个问题没有考虑最基础的遗传学知识。我想如果他是一个复旦生命学院学遗传学的研究生就不会问

这个问题，因为我们的学生都学过遗传学。

现在，我们来解释一下什么是他说的重组和混血。我们知道人类的基因组里面有三大类遗传方式不同的东西组成：常染色体、Y染色体和线粒体。打个比方（图 13），我们建立一个家系，某两个

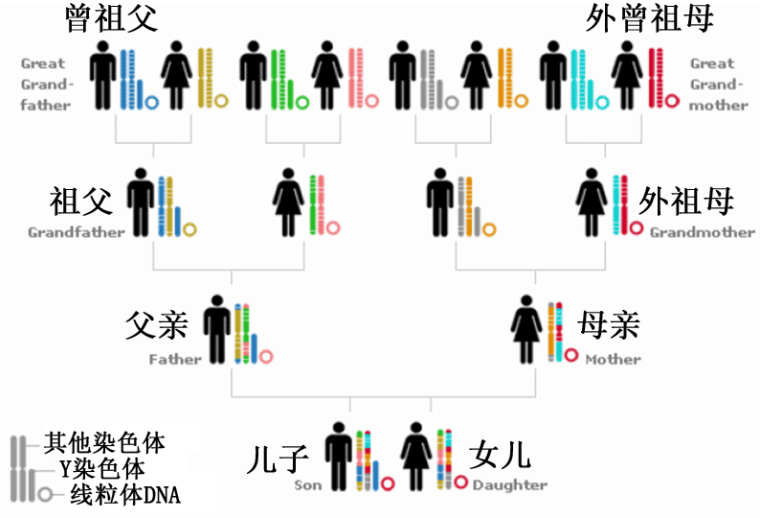


图 13 不同遗传物质的传代方式

人，一男一女兄妹俩，他们有他们的父母，他们的父母亲各自有父母，即祖父母和外祖父母，然后这四个祖辈的人又各自有他们的父母。我们从曾祖父母这一辈开始，用最长的棒形表示常染色体，用短棒表示 Y 染色体，用圈表示线粒体。曾祖辈的八个人用八种不同的颜色表示。我们看曾祖父母，他们结婚后生下的祖父。肯定是从曾祖父的常染色体中传下其中的一半，从曾祖母中传下另外一半，形成了祖父的常染色体；但是 Y 染色体只从曾祖父那里遗传下来，

因为女性没有 Y 染色体，所以生出男孩子当然就得到了父亲的 Y 染色体。但是他的线粒体是从母亲那里得到的。在传代的过程中，母亲的卵细胞是含有线粒体的，因为线粒体是一个产能机构，所有细胞运作的能量都是线粒体造的，所以卵细胞里面肯定有线粒体。那么精子里面有没有线粒体呢？精子里面也有，但是存在于精子的尾巴上。精子的头上含有染色体基因组，但是不存在线粒体的，尾巴上才有线粒体。精子的所有线粒体造的能量都用来作运动的能量。等精子钻到卵细胞里面去的时候，尾巴被扔掉，是进不去的。所以精子和卵子两者结合形成受精卵之后，受精卵里的线粒体都是来自于母亲的。所以孩子的线粒体 DNA 都是永远来源于母亲的，从来没有发现线粒体是来自于父亲的。所以线粒体永远是母系遗传。我们这样看一下图中家系中的父亲的一对常染色体，到祖父传到父亲的时候，祖父来自于他的父亲和母亲的常染色体经过了重组形成了父亲的一条常染色体，父亲的另外一条常染色体来自于祖母的父母亲。我们用了不同的颜色表示曾祖父辈的染色体，所以当传到最后这一代时，常染色体就花花绿绿了，八个祖先的类型全部在他的常染色体中出现。但是他们的 Y 染色体永远是一种颜色，永远不变，只有一个来源。而线粒体也是类似的，都来源于母系一方，来自于她/他的妈妈的妈妈的妈妈的妈妈的妈妈，不管多少代，可以一直往前推，一直可以追溯到来自于 20 万年前非洲的一个女性，我们把她叫做夏娃。《夏娃的七个女儿》这本书就是说夏娃在欧洲传下了七个“女儿”，有非洲最古老的夏娃祖先传下来的七个女性分支祖先跑到了欧洲去。

这样的话，我们就把基因组的各种遗传方式搞清楚了，常染色体是父母双方双系遗传的，Y 染色体是来自于父亲，线粒体只来自于母亲。那么我们就有了这个遗传学基础来追寻男性祖先的基因。曹操如果

有男性后代的话，他的 Y 染色体就可以一代一代传下去，只要中间没有男性的断裂，每一代都有儿子产生，他的 Y 染色体就会一直传下去，他的基因就会一直留到现代。那样子的话，我们通过人肉搜索可以找到曹操传下来的 Y 染色体。

线粒体可以传下来，Y 染色体也可以传下来。那么我们到底去研究线粒体还是 Y 染色体呢？我们刚才知道，线粒体 DNA 它是一个小小的环状的结构，环状结构比较稳定，它没有暴露的端点，所以很多对 DNA 的损伤对线粒体没有常染色体强。即便在机体死亡以后一段时间线粒体 DNA 也破碎了，它还有一个优点可以保证检测的便利，那就是拷贝数比较多。一个细胞里面，核基因组染色体只有一套，有的时候会两套，但是线粒体会有好几套。有的细胞，像肌细胞里面，特别需要运动能量的这些细胞里面，线粒体甚至有上千个。那样在样本中线粒体 DNA 的拷贝数很多，大量保留在骨骼里留在肌体里。它的拷贝数比染色体的拷贝数多得多，所以对于检测古代 DNA 来说，就是非常大的优势。所以无论什么样的古 DNA 实验室，最起码线粒体 DNA 还是能检测。但是线粒体它是母系遗传的，那曹操的卞皇后的女儿嫁人了，她的女儿的女儿也嫁人了，那她们嫁到哪里去了，一直到现在，通过家谱分析根本没有办法搜索，也没有办法在人群中寻找。所以检测古 DNA 的线粒体来确定曹操的身份这件事情没有办法做了。哪怕我们把疑似卞皇后的骨骸的线粒体 DNA 已经找出来了，在现代人中间找标准参照，我们要去人肉搜索，谁知道哪里去找？当然可能有很偶然的机会，说不定能找到现代某个人的线粒体 DNA 可以与骨骸的 DNA 对上，但是说他是曹操的后代，但这个太难了，因为母系的脉络没有什么文字依据，也没有什么办法去严格确定。

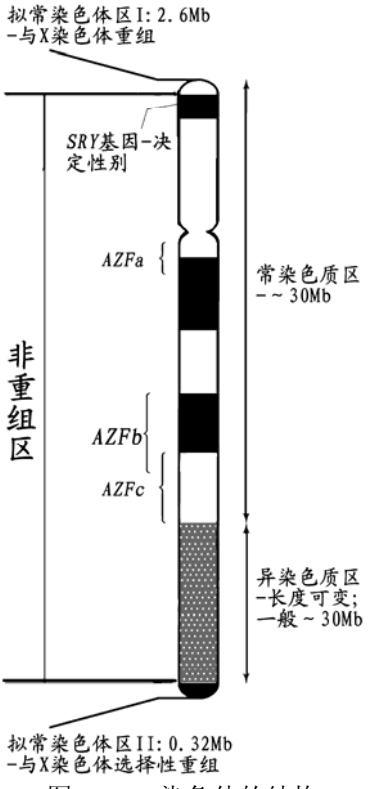


图 14　Y 染色体的结构

但是 Y 染色体就不一样了，我们很可能在后代里面找到相关的类型。我们知道这些后代有家谱记载，他们说是曹操的后代，分析

这些后代的 Y 染色体类型就可以来推导曹操的 Y 染色体类型，可以逆推。那么 Y 染色体的结构是怎么样的？它是不是真得完全不重组呢？它也并不是完全不重组。Y 染色体大概有六千万个位点，即碱基对。这六千万个碱基对的两头叫做端粒，实际上两头的端粒部分很短，只有这两头部分可以和 X 染色体进行重组。但是好在中间的绝大部分成分都是不重组的，所以我们分析的时候把两头的数据都去掉，都不去管它，只看中间的部分，就是严格地父系遗传的。我们只要看中间的片段，就可以辨认出来 Y 染色体的类型。中间的这一部分其中有一半是异染色质区。异染色质区的结构不太稳定，但是还有一半是在常染色质区，有三千万个位点。常染色质区就是我们重点研究的这一段片段，所有的 Y 染色体典型突变都是在常染色质区发现寻找的。所以这三千万个碱基对就是我们用来研究父系脉络的材料。

很重要的一点就是确定姓氏和 Y 染色体是不是能够真得对应起来。我们知道 Y 染色体它是父系遗传的，姓氏也是父系遗传，我们大部分人的姓跟父亲的是一样的。父系遗传的姓氏与 Y 染色体 DNA 的关系，我们可以从图 15 中看到。假设这个男的姓张，女的姓王，他们的小孩不管男的女的都姓张，如果是小男孩的话，他的 Y 染色体就会传下去，他的儿子还会姓张，正常情况下甚至可以说 Y 染色体永远姓张。所以 Y 染色体实际上也是“有姓”的。Y 染色体实际上不但使性染色体，还是“姓染色体”，只要不改姓、不收养、不红杏出墙。Y 染色体就永远跟着姓传下去。当然如果出现这三种问题的话就没有办法把 Y 染色体与姓紧密连在一起了。所以我们基本上认为 Y 染色体是跟着姓氏一代一代传下去。

那么 Y 染色体上面的祖先信息可以保留么？我们知道 Y 染色

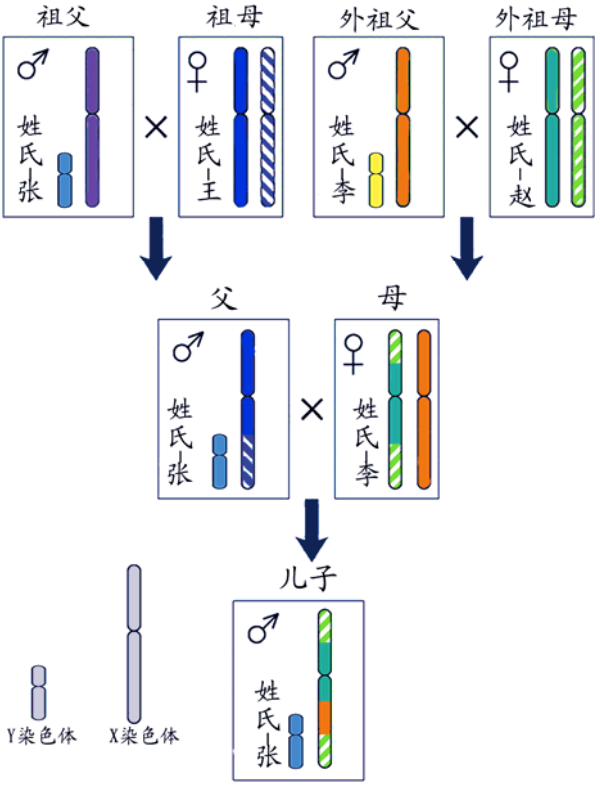


图 15　Y 染色体与姓氏有一样的传代方式

体不断地在突变，会不会祖先突变出的类型在后代中又突变没了呢？那样的话，祖先的信息就会丢失，这是一个很大的问题，很多人都在问这个问题。我们要解答这个问题，就要关注 SNP 这种突变的突变率，也就是它的突变速度有多快，每次传代的突变概率有多少。我们通过家系的调查，在家系里面两个个体隔开多少代，他们

就会有多少个 SNP 的差别，就可以估算出一个粗略的突变率。还有一个分析途径，我们知道猩猩、黑猩猩和人之间的大致分化时间，根据他们之间的 Y 染色体序列差异比较，把差异除以分化时间，我们也可以推算出了一个大概的突变率。这个突变率大约是每个人每个位点上会有三千万分之一。这个概率非常小。比如你生个孩子，他的每个位点上有三千万分之一的概率发生突变而与你不同。在整个基因组的大约 30 亿个碱基里面，每传一次代，发生突变的点的总数就有：30 亿乘以三千万分之一，等于一百。也就是说每个人跟他父亲在整个基因组里面大概有一百个点不同，所以每个人的基因组都是不一样的。所以人和人总是长得不一样，不可能一模一样。很多人怀疑祖先的突变会不会丢失，怎么样的情况会导致丢失呢？那就是已经发生突变的这个点，比如说 A 到 G 的这个突变，再变回去，就变没了。打个比方，比如说这个点上面，原来是 A，是祖先型，就是没有变化之前，猩猩也是 A；而发生突变了就是 G，那么这个点上是 G 的话，它就是突变的特征，它就有所有 G 类型的个体的共同祖先的信息在上面。但是如果后来有些 G 类型的个体在这个点上再发生一次突变，G 变成 A，那么这个突变特征就没有了，又变成原始型了，跟猩猩一样了。所以这就是突变又变没了，祖先的信息丢失了。这叫做回复突变，是基因组中的“返祖现象”。但是这种回复突变的概率有多少呢？我们来讲一下概率统计的最基本的原理，最简单的概率分析。在 Y 染色体上分析，我们在 Y 染色体上稳定片段里面选择三千万个碱基的片段，第一次突变是随机的，总是能够找到一些点突变，那是无所谓概率的。那么在一个特定的点上发生突变，再把它变没了的概率，就是三千万分之一，这是很容易理解的。固定的这个点让它发生突变的概率是三千万分之一。那么也就是说我们要传三千万代，才会把这个点变没掉，所以我们且等着吧。我们现代人从产生到现在也没有三千万代。所以我们基本上认为祖先的突变信息不会在后代中丢失的。

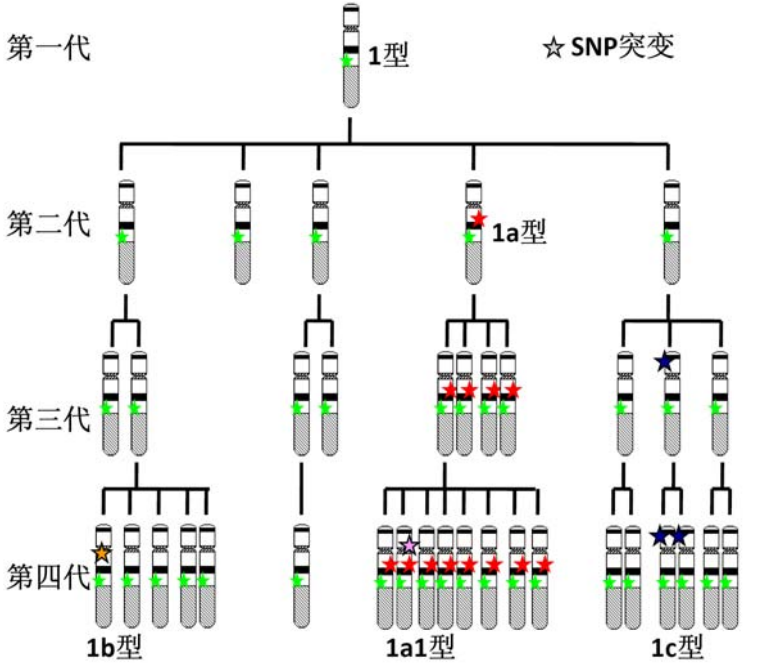


图 16　SNP 的突变发生构成不同的 Y 染色体类型

那样的话，后代跟祖先是不是没有什么差别呢？当然有差别。我们看这个图（图 16），我们在 Y 染色体上研究三千万个碱基，每个碱基每个人有三千万分之一突变的概率，所以每个男人平均总是有一个碱基，也就是一个点，是跟父亲不一样的。所以世代传递中，不断地在产生新的突变加到突变谱序上去，新发生的突变随机地出现在三千万个点当中。目前，我们在 Y 染色体上研究的点还比较少，是通过全世界范围内少量的样本中，在 Y 染色体上这里找一个，那里找一个，随机找的。所以在我们目前普遍分析的点上，当然不存

在每代有一个突变。但这些点也是我们可以区分 Y 染色体大类的基本信息，它们构成了一个基本的 Y 染色体进化树。比如图 16 中这个祖先有一个突变，我们画个五角星，一种颜色代表一种类型的突变，那么他所有的后代都会有这个突变，不会丢失的，代表第一种类型；然后他的某一个后代里面（第二代的后代里面）突然出现了第二个类型的突变，这个突变呢在他的后代里面永远也会传下去，永远不会丢，那么就形成了第二种类型；这个类型在后代中又产生了一个突变的话，就形成了第三种类型；第三种类型是第二种类型的亚型，所以这种 Y 染色体型就一个一个分下去，形成了不同的型。比如说我们把图中只有绿色突变的类型叫做 1 型；它下面出现的第二种类型叫做 1a 型，1 型的亚型；然后 1a 又产生了一个亚型 1a1 型，1 型中 a 亚型的第一个小亚型；就不断这样分下去。所以我们知道 1a1 型它是 1 型的后代型。这是很明确的一个谱系分析。那边还有 1b 型，是另外一个完全不同的突变。所以不同的分支上面的后代，他们的突变谱序就完全不同，也形成了完全不同的亚型。亚型跟亚型之间又有远近关系，1a 型和 1a1 型他们之间的差距比 1b 型肯定更小。这就是后代跟祖先的关系，祖先的信息传给后代，但是后代在祖先的信息上不断追加信息。这样的话，我们就可以通过多个分支的后代的类型追溯祖先的类型。

我们在全世界的人类群体中调查了 Y 染色体的类型后，发现了无数的 SNP 的点，现在我们一般用来分类的，有六百多个 SNP 突变点。这六百多个 SNP 突变的点，根据他们的突变谱系，就把它们分成了接近四百多个单倍型。四百多个单倍型，分别属于很多大型，从 A 型到 T 型，大型下面再分为小型（图 17）。观察这些大型的地理分布，我们发现各大型在各个地方地理分布都不同。

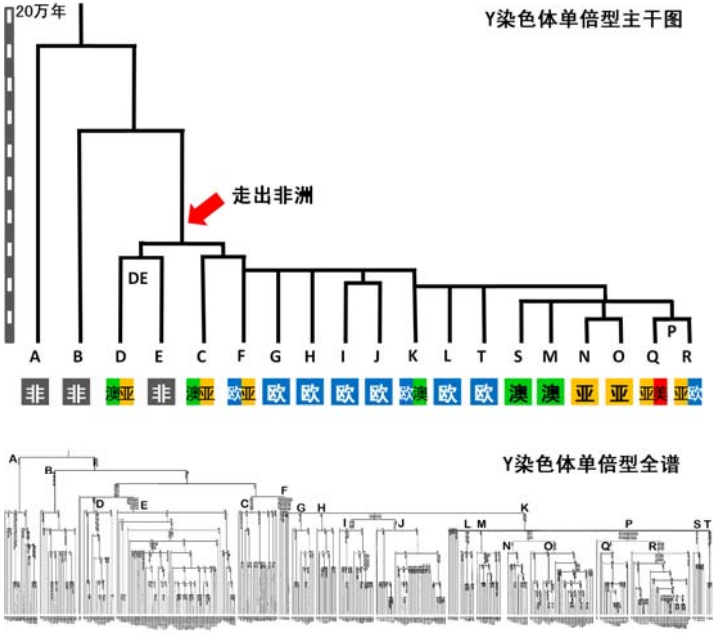
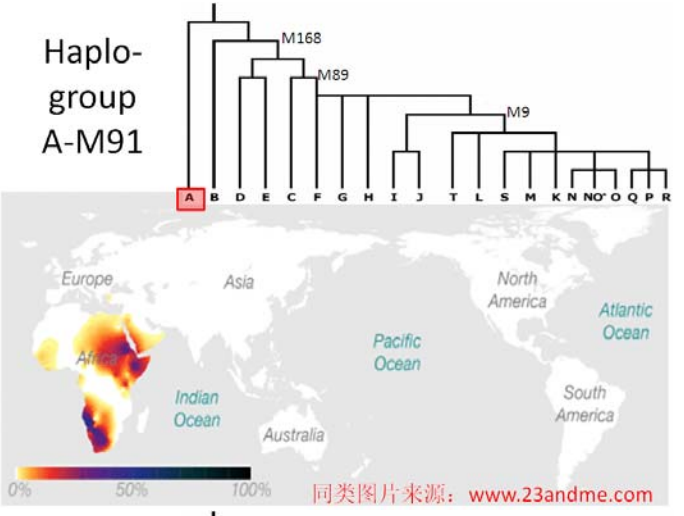
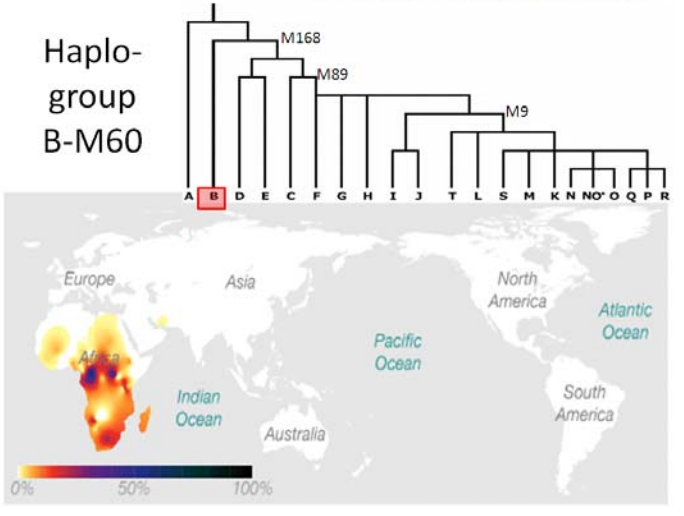
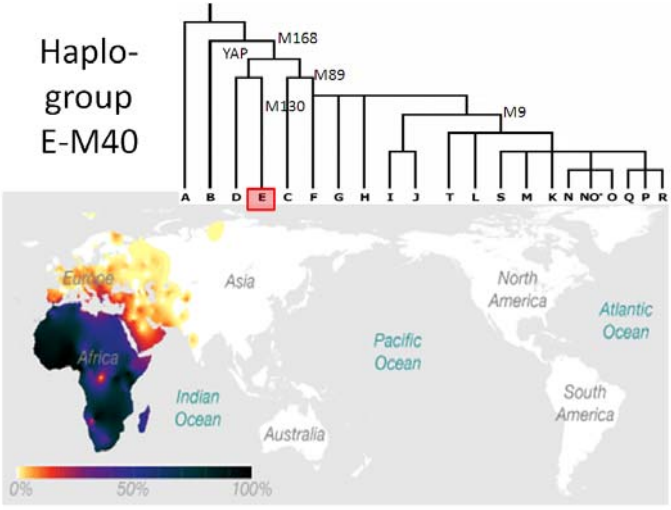


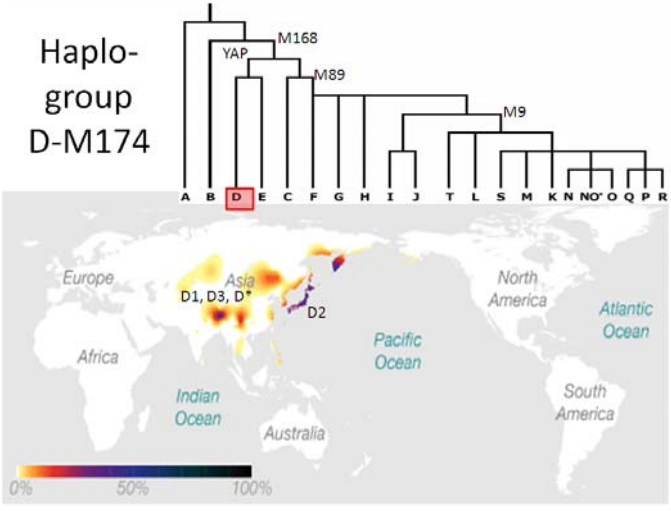
图 17　Y 染色体类型构成的进化树

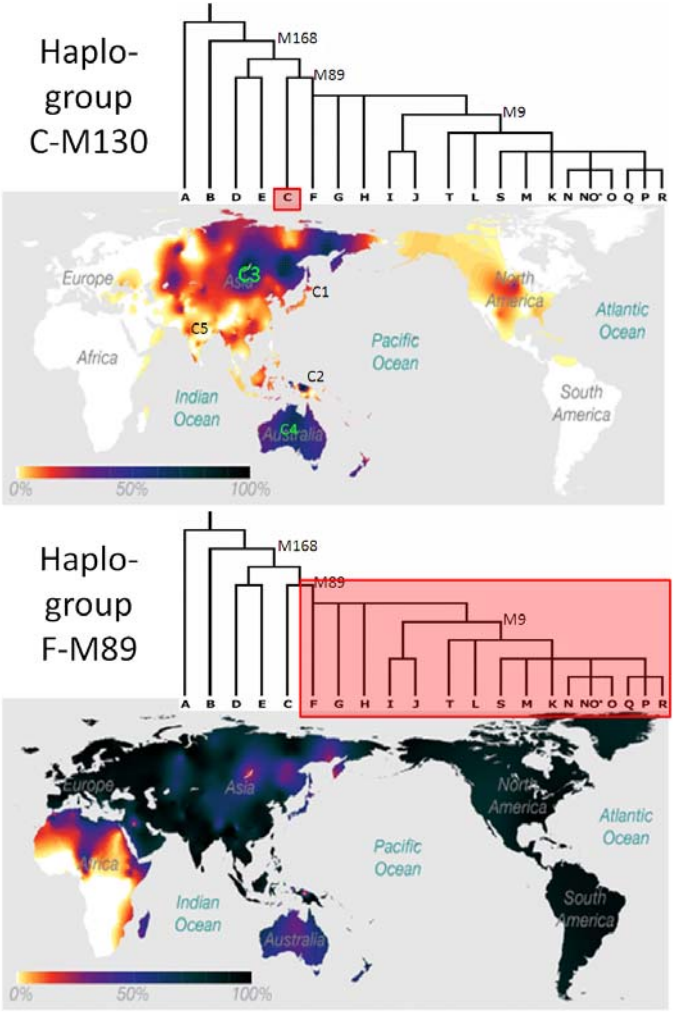
我们知道 A 和 B，还有一部分的 E 型，都在非洲出现。D 型在亚洲东部出现，在黄色人种中。C 型也是。F 型在白种人、黄种人中出现。G、H、I、J 都在白种人中出现。K 在白种人、黄种人中都有。L、T 在白种人中出现。S、M 只在棕色人种中出现。N、O、Q 都是黄种人中出现。所以整个 Y 染色体的进化结构我们就知道了，最老的类型 AB 都是在非洲，所以说我们从 Y 染色体上得出了人类走出非洲的这个概念。按照时间分析的话，我们把他追溯到了二十多万年前，把他比喻成“亚当”。下面我们看看这些类型具体怎么分布的：

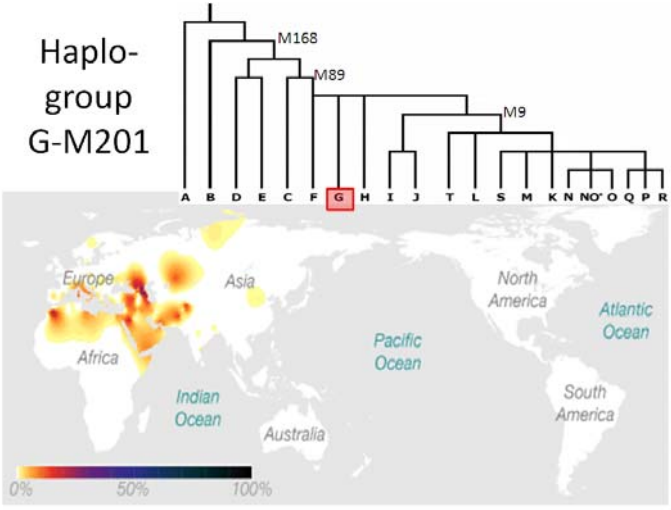

Haplo-
group
A-M91
M168
M89
M9
A B D E C F G H I J T L S M K N NO O Q P R
Europe
Asia
North
America
Atlantic
Ocean
Pacific
Ocean
Africa
Indian
Ocean
South
America
Australia
0%
50%
100%
同类图片来源：www.23andme.com

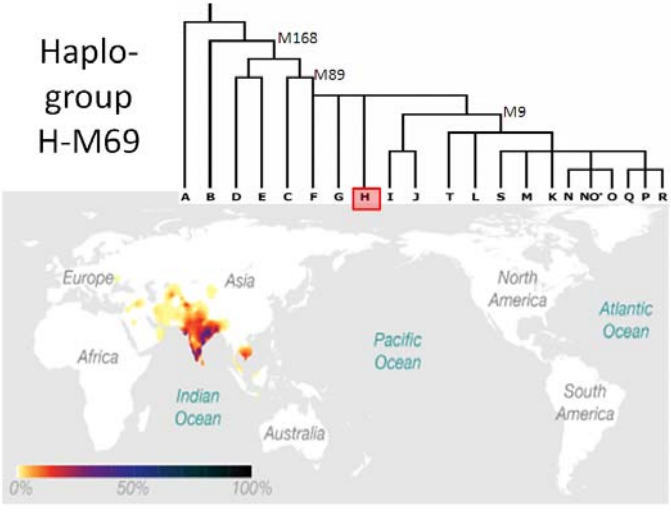

Haplo-
group
B-M60
M168
M89
M9
A B D E C F G H I J T L S M K N NO O Q P R
Europe
Asia
North
America
Atlantic
Ocean
Pacific
Ocean
Africa
Indian
Ocean
South
America
Australia
0%
50%
100%

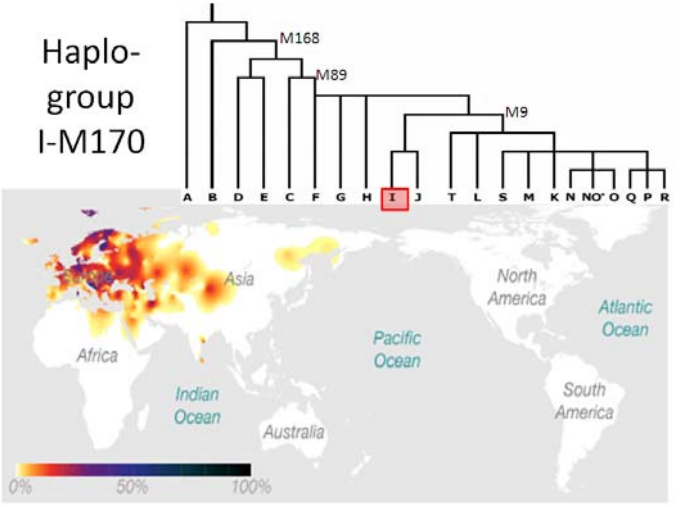

Haplo-
group
E-M40
M168
YAP
M89
M130
M9
A B D E C F G H I J T L S M K N NO O Q P R
Europe
Asia
North
America
Atlantic
Ocean
Pacific
Ocean
Africa
Indian
Ocean
South
America
Australia
0%
50%
100%

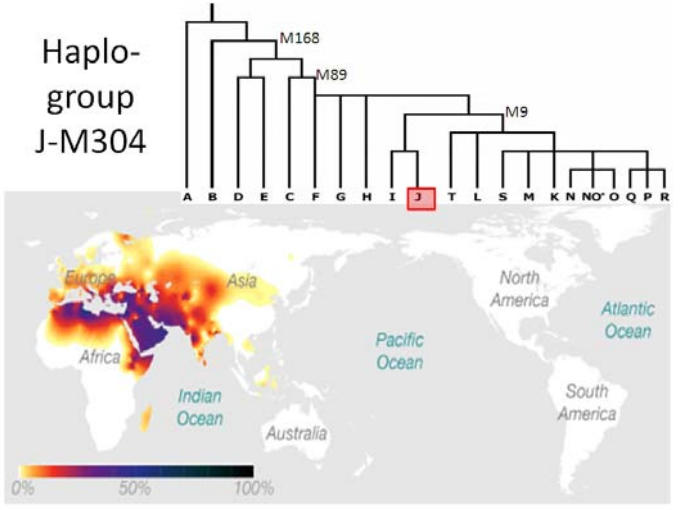

Haplo-
group
D-M174
M168
YAP
M89
M9
A B D E C F G H I J T L S M K N NO O Q P R
Europe
Asia
D1, D3, D*
D2
North
America
Atlantic
Ocean
Pacific
Ocean
Africa
Indian
Ocean
South
America
Australia
0%
50%
100%

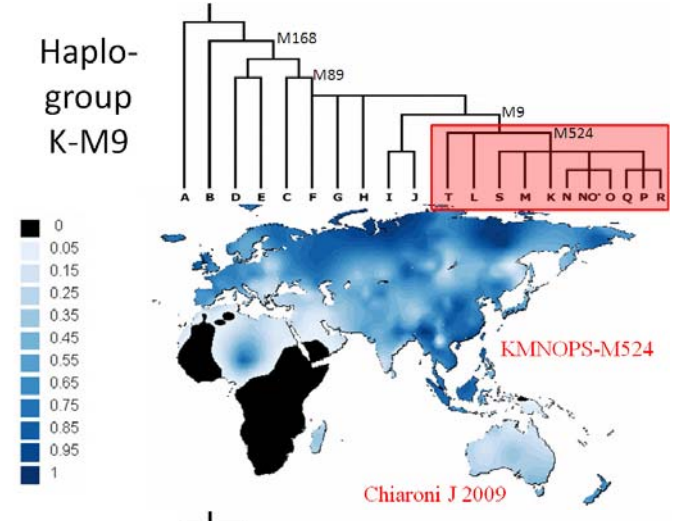

Haplo-
group
C-M130
M168
M89
M9
A B D E C F G H I J T L S M K N NO* O Q P R
Europe
Asia
C3
C1
C5
C2
C4
Africa
Indian
Ocean
Pacific
Ocean
North
America
Atlantic
Ocean
South
America
Australia
0%
50%
100%
Haplo-
group
F-M89
M168
M89
M9
A B D E C F G H I J T L S M K N NO* O Q P R
Europe
Asia
Africa
Indian
Ocean
Pacific
Ocean
North
America
Atlantic
Ocean
South
America
Australia
0%
50%
100%

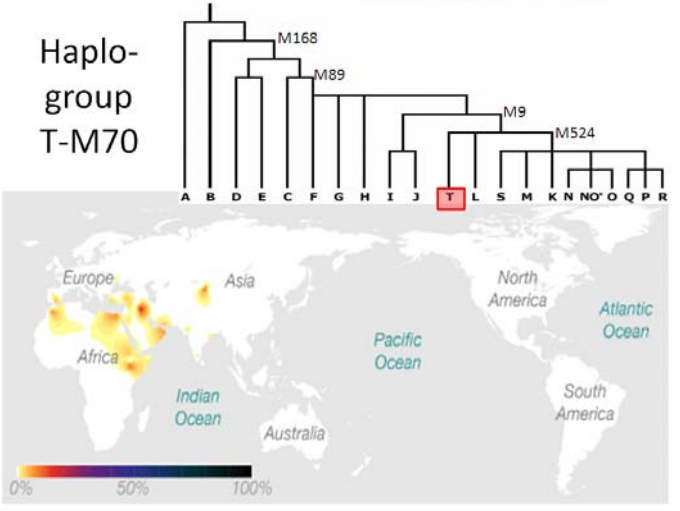

Haplo-
group
G-M201
M168
M89
M9
A B D E C F G H I J T L S M K N NO O Q P R
Europe
Asia
North
America
Atlantic
Ocean
Africa
Pacific
Ocean
Indian
Ocean
South
America
Australia
0%
50%
100%

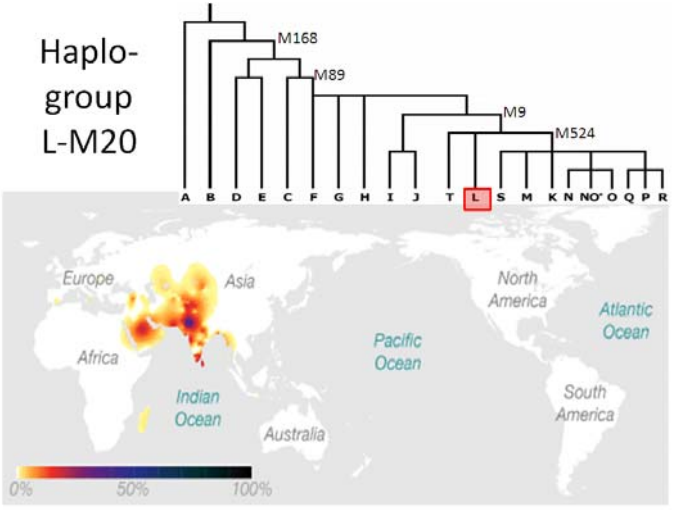

Haplo-
group
H-M69
M168
M89
M9
A B D E C F G H I J T L S M K N NO O Q P R
Europe
Asia
North
America
Atlantic
Ocean
Africa
Pacific
Ocean
Indian
Ocean
South
America
Australia
0%
50%
100%

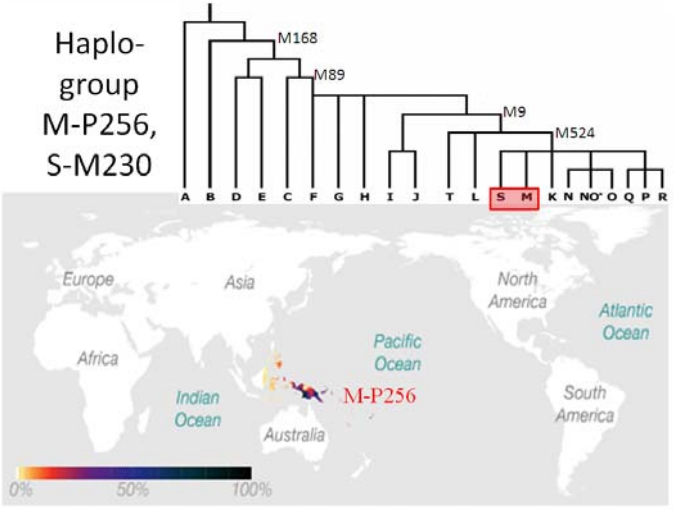

Haplo-
group
I-M170
M168
M89
M9
A B D E C F G H I J T L S M K N NO O Q P R
Asia
Africa
Indian
Ocean
Pacific
Ocean
North
America
Atlantic
Ocean
South
America
Australia
0%
50%
100%

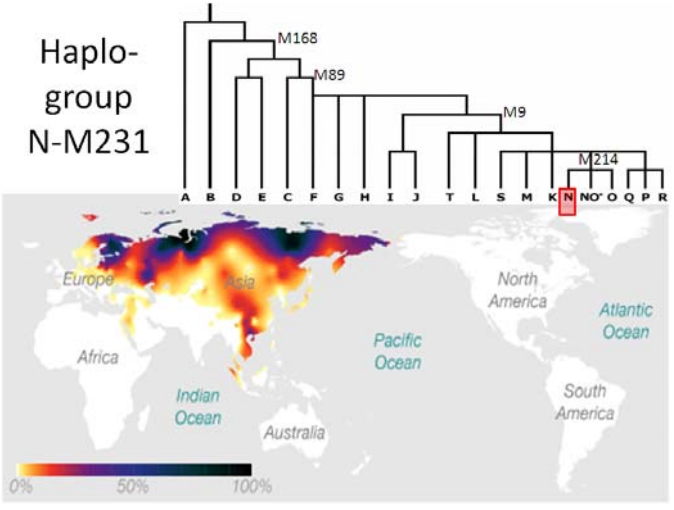

Haplo-
group
J-M304
M168
M89
M9
A B D E C F G H I J T L S M K N NO O Q P R
Europe
Asia
Africa
Indian
Ocean
Pacific
Ocean
North
America
Atlantic
Ocean
South
America
Australia
0%
50%
100%

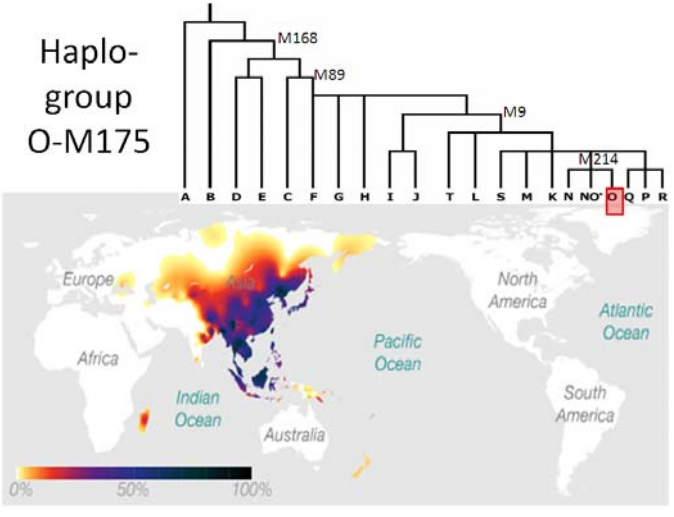

Haplo-
group
K-M9
M168
M89
M9
M524
A B D E C F G H I J T L S M K N NO O Q P R
0
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
1
KMNOPS-M524
Chiaroni J 2009

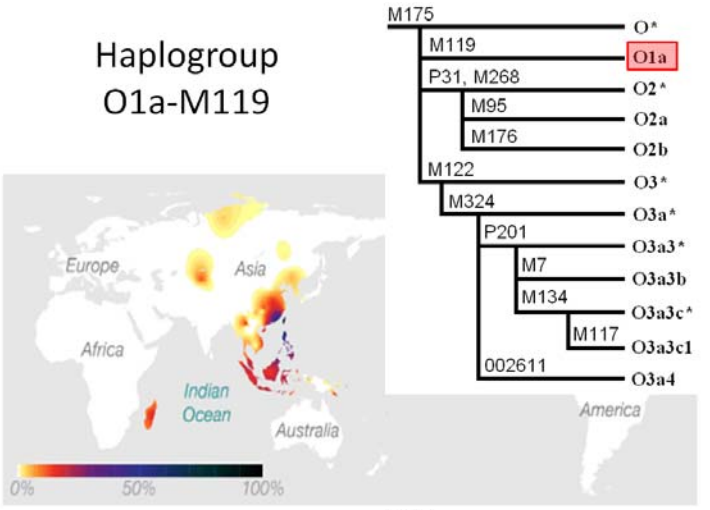

Haplo-
group
T-M70
M168
M89
M9
M524
A B D E C F G H I J T L S M K N NO O Q P R
Europe
Asia
Africa
Indian
Ocean
Australia
Pacific
Ocean
North
America
Atlantic
Ocean
South
America
0%
50%
100%

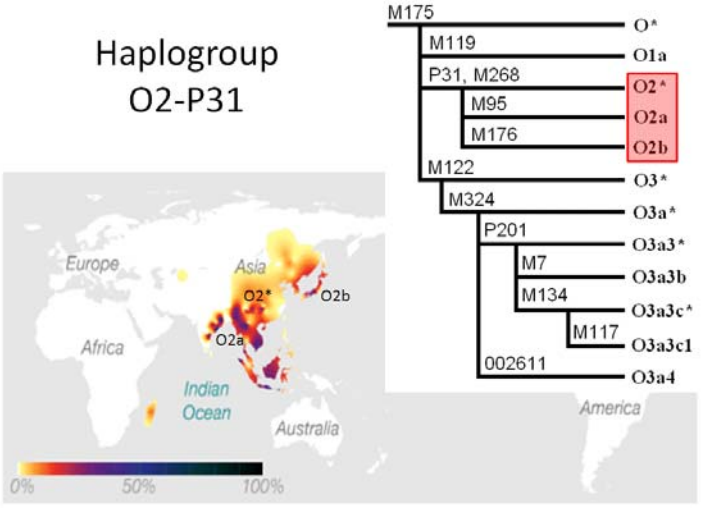

Haplo-
group
L-M20
M168
M89
M9
M524
A B D E C F G H I J T L S M K N NO O Q P R
Europe
Asia
Africa
Indian
Ocean
Pacific
Ocean
North
America
Atlantic
Ocean
South
America
Australia
0%
50%
100%

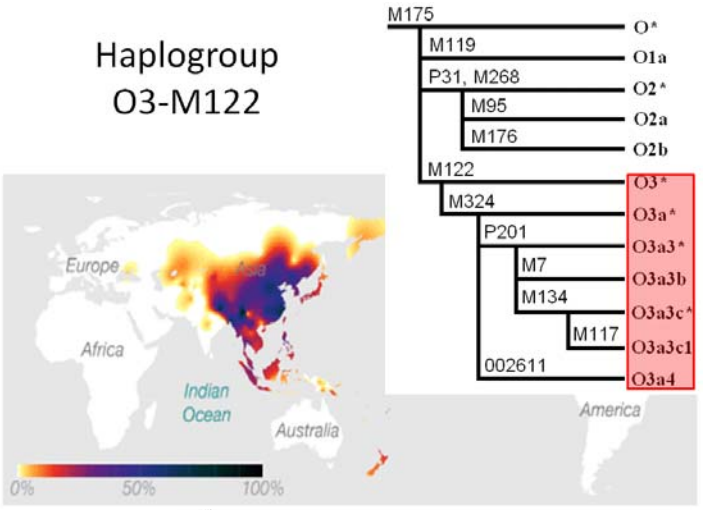

Haplo-
group
M-P256,
S-M230
M168
M89
M9
M524
A B D E C F G H I J T L S M K N NO O Q P R
Europe
Asia
Africa
Indian
Ocean
Pacific
Ocean
M-P256
North
America
Atlantic
Ocean
South
America
Australia
0%
50%
100%

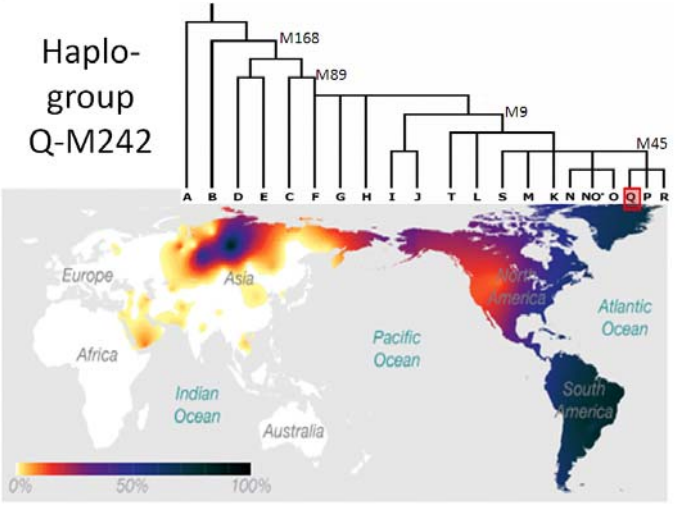

Haplo-
group
N-M231
M168
M89
M9
M214
A B D E C F G H I J T L S M K N NO O Q P R
Europe
Asia
North
America
Atlantic
Ocean
Pacific
Ocean
Africa
Indian
Ocean
South
America
Australia
0%
50%
100%


Haplo-
group
O-M175
M168
M89
M9
M214
A B D E C F G H I J T L S M K N NO O Q P R
Europe
Asia
North
America
Atlantic
Ocean
Pacific
Ocean
Africa
Indian
Ocean
South
America
Australia
0%
50%
100%


Haplogroup
O1a-M119
Europe
Asia
Africa
Indian
Ocean
Australia
America
0%
50%
100%
M175
M119
P31, M268
M95
M176
M122
M324
P201
M7
M134
M117
002611
O*
O1a
O2*
O2a
O2b
O3*
O3a*
O3a3*
O3a3b
O3a3c*
O3a3c1
O3a4


Haplogroup
O2-P31
Europe
Asia
O2*
O2b
O2a
Africa
Indian
Ocean
Australia
America
0%
50%
100%
M175
M119
P31, M268
M95
M176
M122
M324
P201
M7
M134
M117
002611
O*
O1a
O2*
O2a
O2b
O3*
O3a*
O3a3*
O3a3b
O3a3c*
O3a3c1
O3a4


Haplogroup
O3-M122
M175
M119
P31, M268
M95
M176
M122
M324
P201
M7
M134
M117
002611
O*
O1a
O2*
O2a
O2b
O3*
O3a*
O3a3*
O3a3b
O3a3c*
O3a3c1
O3a4
Europe
Asia
Africa
Indian Ocean
Australia
America
0%
50%
100%


Haplo-
group
Q-M242
M168
M89
M9
M45
A B D E C F G H I J T L S M K N NO O Q P R
Europe
Asia
North America
Atlantic Ocean
Africa
Pacific Ocean
Indian Ocean
South America
Australia
0%
50%
100%

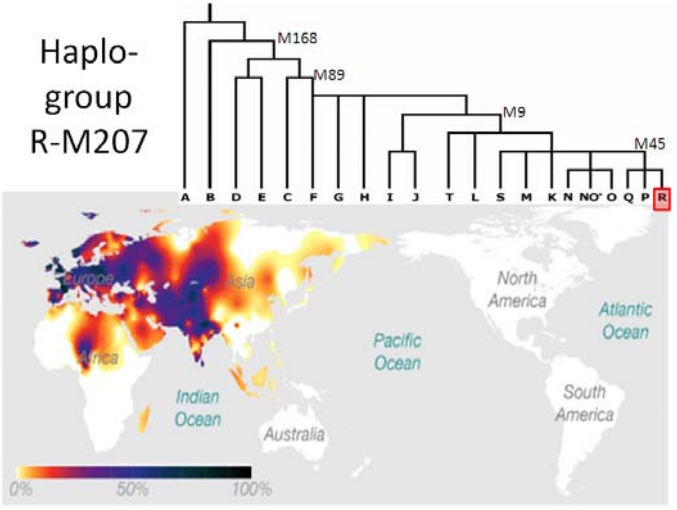


图 18 Y 染色体各种类型的地理分布

A 型大致分布在东北非和西南非，布须曼人和撒哈拉人中。B 型出现在中非部分和东非部分，可能是班图黑人的扩张造成的。E 型分布的地方就比较多，在西非洲和东南非很多，这是大部分黑人的主要类型，在欧洲和东非也有一部分的分布，甚至分布到亚洲的一些地方，分布到北亚的一些地方。前面几种类型在中国人里面一般是找不到的。D 型在中国比较多，在中国的西藏以及西北方的一些地方，一些亚型出现在了日本和东南亚，所以 D 型也是分布零散、历史很传奇的一个类型，年代估算发现它可能是在中国地区最古老的一个类型，在整个东亚地区也属于很古老的一批类型，在一般人群里面很少出现。C 型也是出现在远东地区，在澳洲非常多，在北亚也非常多，它呈现出被另外一种类型挤压以后出现两头多中间少的分布方式，还有一部分出现在北美的纳丁尼人中。

F 型是一个上游大型，它包括了 F 下游的所有类型。它的分布是这样的，整个欧亚大陆和美洲大陆基本上都是 F 型。F 型下面有 G 到 T 型。其中 G 型出现在高加索山区和小亚细亚，然后以此为中心向四周扩散。H 型出现在印度、巴基斯坦和伊朗地区，它是雅利安人扩张造成的结果，但也可能是更早的居民荼拉维达人留下的，也可能是比荼拉维达人更早的人留下的。I 型出现在欧洲的大部分地区，但是比例也不高，应该是印度到欧洲的一种土著型。J 型出现在亚非语系、阿拉伯世界和欧洲的很多地方，是白种人的类型，犹太人里面大部分都是 J 型。

K 型也是一个大的类型，包括 K 字母以后很多类型，所以它也是一个上游型，它的分布在亚洲很常见，但是把下面类型去除掉之后的纯 K 型是非常非常罕见的。T 型是一个比较少见的型，也是分布在阿拉伯世界。然后 L 型分布在以巴基斯坦为中心的一圈地区。S 和 M 就只分布在新几内亚土著人之间，传说中的食人部落里。

N 型很有意思，它的起源是在中国，但是居然分布到了整个北亚地区的很多民族中间，并且在乌拉尔民族的群体里面占到了绝大多数，在芬兰、爱沙尼亚这些国家里面，占到了绝大多数。所以这部分欧洲人的祖先居然大多是从中国出去的，部分欧洲人在欧洲号称自己是黄种人，但是我们怎么看都不像，至少比新疆人白多了，但他们还是号称自己是黄种人，所以这是一种民族心理上的错觉。我们看人的时候会更多地关注一些与周边人的差异，他们在欧洲觉得自己跟周围的白种人差别太大，因为他们关注这些差异，并且在大脑里面把这些差异给放大了。

O 型对于我们中国人来说是最重要的，因为它是东亚人最主要

的类型，在东亚人里面，几乎占到了百分之七十到百分之八十，在中国人里面甚至更多。O 型是黄种人的关键型，但是刚才说的跑到欧洲去的 N 型跟 O 型是兄弟型，非常接近的。O 型下面分很多亚型，分 O1、O2、O3 这三大支。O1 出现在中国东南的百越民族的后代中间，包括侗傣语系与南岛语系的这些民族，台湾高山族里面有些民族比如阿美族占到百分之百。O2 型非常有意思，一部分叫做 O2a，下面分两部分，一个分支出现在东南亚，从湖南湖北的苗族开始往西南地方绝大部分都是 O2a；另外一个分支 O2b 出现在日本和朝鲜，从朝鲜往东北的地方的 O2 大部分都是这个类型。O3 是一个大头，在中国大部分民族里面都有 O3，而且 O3 在大部分中国群体中都是占多数，在汉族人里面更是多数，占到了大概百分之六十，O3 下面有很多很多分支，这些分支能够让我们辨认汉族人里面到底有多少差异。

Q 型是出现在北亚的叶尼塞语系，进入到楚科奇-堪察加语系，然后进入到美洲印第安人，在美洲印第安人中越往南 Q 型越多，美洲南方很多民族百分百都是 Q 型。因为随着迁徙，他们一路走一路丢类型，到最后只剩下一种类型，所以迁徙得越到后面，类型就越少，因为不可能所有的人都能够成功迁徙而把各种类型都带到最后。Q 型的源头很可能就是在北亚叶尼塞语系分布的地方。叶尼塞语系的民族古代有很大一批，分布在西伯利亚的叶尼塞河流域，后来在俄国东进的过程中被消灭到只剩下现在 Ket 一个民族。我们一直怀疑匈奴人的主要语言就属于叶尼塞语系，跟现在大部分人群都不一样。最后一个类型是 R 型，R 型跟 Q 型、P 型都是同源的兄弟型，但是它大部分分布在欧洲。所以 Q 和 R 是一个往东跑，一个往西跑。R 型在欧洲也占到很大一部分，但是我们一直怀疑它是从亚洲跑过去的，所以很多白种人可能是从亚洲起源的。

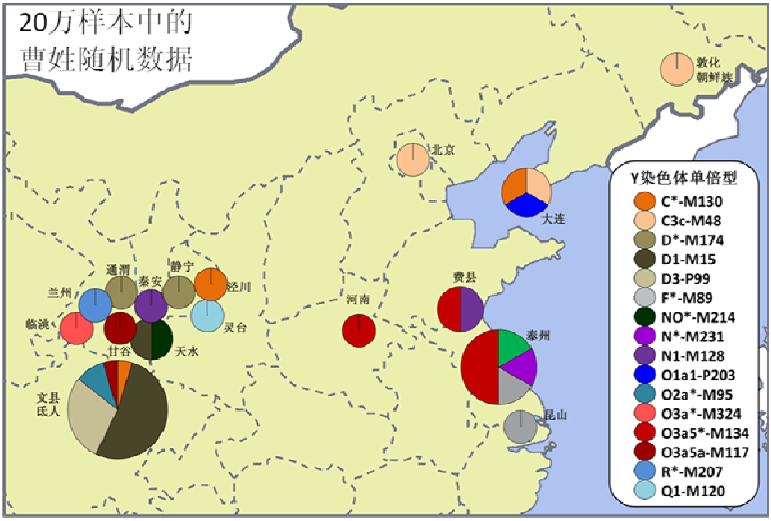


图 19　曹姓随机样本的 Y 染色体类型

但是 Y 染色体有这么多型，曹姓到底是属于哪个型呢？我们在现代人类学中心的 20 多万人的数据库里面搜索了一下，发现了 48 个汉族姓曹的人；还有一些少数民族姓曹的人，比如甘肃和四川边界上的一些氐人，现在民族成分是藏族里面，叫白马藏族，也有部分叫做蟹螺藏族的，这个民族有很多人姓曹，我还不知道为什么。我们发现，这些随机的曹姓样本里面，Y 染色体五花八门，什么类型都有。类型最复杂的是甘肃的这一群曹姓样本，一个人一种 Y 染色体类型，谁跟谁都不一样。为什么会这样？我们查阅了一下史籍，发现他们的祖先原来是唐朝的时期，从中亚迁徙过来的，叫做昭武九姓。昭武九姓实际上是有九个国家，其中有一个国家音译为曹国。当时中亚地区两河流域是阿拉伯人和唐朝相互争夺的地方。

中亚两河流域长期以来是唐朝的属地，很多国家是唐朝的属国。当阿拉伯人扩张到这个地方以后，就跟唐朝发生冲突。后来唐玄宗时，在高仙芝将军指挥的怛罗斯战役中，唐朝打败了，撤除了这个地区。于是阿拉伯人占领了这个地方，对当地居民进行伊斯兰教化。当地有很多民族不愿意接受伊斯兰教，就随着唐朝的政府官员和军队逃回来唐朝内地。因为他们是唐朝属国，所以地方丢了，就跟着逃到内地去。昭武九姓就是这么过来的，过来以后因为属国没有了，要进行正式的“改土归流”，把他们登记成汉籍，在户口上面要全部写上汉名，写汉名的时候总归要有汉姓，就根据原来的属国是什么就姓什么。鱼国的姓鱼，曹国的姓曹。所以一个国家的人都姓曹了，不管他原来姓什么了。现在我们去看昭武九姓的后代，姓曹的一个人就一个样，因为他们本来就不一样。

另外一些人，比如说中原地区，山东、河南、江苏这些地方找到的类型，大部分类型都是 O3a5*型。这个类型在汉族里面太多了，在汉族里面几乎是最大的一个类型，如果不算它的下游类型的话。另外还有 N1 型，属于 N 型的亚型；还有 D 型。这样一看的话，中原地方什么类型都有，而且跟汉族的类型比例分布基本是吻合的，所以汉族人有什么类型，曹姓也是有什么类型，比例差不多都是一样的。在东北的曹姓，大多数是 C 型的类型，也有一部分 O1。东北人中 C 很多，C 的地理分布是两头多、中间少，从东北开始往西北利亚基本都有很多 C 型，蒙古族满族里几乎占到一半以上。这样的话，我们看到曹姓 Y 染色体类型的分布跟当地人群没什么差别，哪个地方某种类型多，那么曹姓也是这种类型多，也就说普通的曹姓没有一个共同的起源。那这个现象对于我们寻找曹操的基因是有好还是不好呢？

实际上这是好的，对我们有利，为什么？我们先来讲一下曹姓

的 Y 染色体为什么有那么多不同。我们知道，还是有很多原因可以让姓氏变迁和 Y 染色体变化不吻合，会造成很多姓氏内部的多样性。有一种可能原因是基因的突变，每个人会产生一个突变点。但是我们知道，我们分的这些 Y 染色体类型都是一万年前、两万年前就形成的，它下面那些新加的 SNP 突变类型不会影响 Y 染色体大的分型，所以我们根本就不去检查。我们检查的那些点都是老早就有的，所以不是基因的突变造成的曹姓有那么多的不同 Y 染色体类型。还有一个可能是不同的血统来源取同姓。比如夏朝封了一个曹国，后代姓曹；周朝封了一个曹国，后代也姓曹；有些人在曹家的大院里面做长工，过了几年也姓曹。所以鲁迅笔下说阿 Q 姓赵，谁知道是不是真的。所以一个姓可能有不同的来源，自然 Y 染色体就会多种多样。还有一个可能是改姓。包括赐姓。包括原来的“氏”也用成姓。有的是因为避讳所以改姓。有的是因为避祸，比如有传说因为曹操的后代被追杀，所以部分人不姓曹姓操。还有少数民族改汉姓的情况，都有可能。还有入赘随母姓到底有没有？没有人知道，因为族谱里面不会说。收养的情况很多谱里面也不会说。当然儿子非父亲亲生，红杏出墙的可能就更不容易知道了。所以这些都是有可能造成曹姓的血统五花八门的原因，这就没有办法了。

这个现象对于我们寻找曹操的 Y 染色体类型是有利的，为什么呢？如果随机人群中曹姓本来就很一致，那么曹操的后代就很不容易从曹姓中区分出来。但是在随机人群中既然曹姓是这么复杂多样，那么如果曹操的后代基因类型比较一致，我们就可以很好地把曹操后代与曹姓普通人分开了。详细的分析一下：我们现在找到了数十份家谱都是说曹操是他们的祖先，我们先假设他们都是假冒的。如果他们是假冒的，实际上他们就是各地的曹姓自己编一个祖

先是曹操的说法。那么各地曹姓我们知道他们 Y 染色体类型都是不同的，他们如果是编造的共同祖先，我们只要检测他们的 Y 染色体，应该都是不同的，跟各地普通曹姓一样，谁跟谁都不挨着，跟普通的汉族频率分布也应该是一样的。所以我们去检测这些有家谱记载的曹操后裔，如果他们的类型是一致的话，这样就与随机的曹姓不一样，有家谱的曹姓确实有一个共同祖先，我们才能够判定他们是曹操的后裔。如果曹姓的基因本来就是一样的，我们再去检测家谱记载的曹操后代类型多半也一样，就是普通曹姓的类型，可能是普通曹姓冒认祖先，谁知道这个类型是不是曹操的类型。因为曹姓本来就只有一个类型，我们再用这个类型去测曹操的骨头的话，最多能证明他姓曹，谁知到他是哪个曹。所以随机人群不一样，反而更好，给我们一个差异很大的对比，很好的参照，如果到最后我们得出曹操后代都一样，才更有意义。

所以如果我们在曹操家族里面验证出来 Y 染色体都一样的话，就大致可以确定他们真的是曹操的后代。那么如何来估计我们得到的曹操 Y 染色体类型的可靠性呢？这需要一些统计分析。我们刚才看到在汉族普通群体里面没有一种类型超过 10%，所以我们假定检测出来的曹操后代类型频率最高是 10%。我们假设检测了若干个家族，比如在安徽的家族 1 里面检测出 X 型，X 型在汉族里面的频率假设是 10%，当然所有的类型在普通曹姓里面的频率都没有超过 10%，因为我们把分型分得很细，分得很细就把类型分得很多，所以每一种类型的个体都很少，没有一个类型的频率超过 10%，那么实际上 X 型的频率还要低。假设说第二个家族在辽宁，与第一个家族完全不认识的一个家族，两个家族不相干的，他们也说是曹操的后代，也检测出来是 X 型，那么这个巧合的概率有多大呢？如果家族 2 与家族 1 的家谱是伪造的话，那么他们是不同源的。家族 2 作为家族 1 不相干的群体，应该和其他的普通的曹姓一样，检测出 X

型的可能性就是频率 10%。但是如果再冒出一个家族 3，比如在四川，他们也检测出来 X 型，假设他们也是假冒的曹操后代，那么这个巧合的概率是多少呢？也是 10%。那么这三个家族都巧合而是一样的 Y 染色体类型的概率如何计算？按照统计学原理，两个事件叠加，就是两个概率相乘。比如说一个事件有二分之一的可能性，另外一个事件也有二分之一可能性，两个事件同时发生，它就只有 1/2 乘以 1/2 等于四分之一的可能性。所以三个家族都一样，换句话说，我们找到的后面两个家族类型跟第一个家族类型相同的概率只有 10%乘以 10%等于百分之一的概率。当然三个家族都一样的话，他们作为曹操后代的可靠性还是蛮大的，不可靠性只有百分之一。但是如果我们在十个家族中间调查，十个完全不相干的家族，如果都谎说自己是曹操后代，但是都检测出是 X 型，这种巧合的概率是多少呢？是百分之十乘以百分之十乘以百分之十……，乘以九次，那么这个概率只有 0.000000001，9 个 0。所以这种情况下，他们假冒的可能性是非常小的，反过来说，我们在号称曹操后裔中检测出来他们是真的曹操后裔的可靠性就是 99.999999999%。假定的 X 型频率 10%的时候，十个家族里面检测出来的结果可靠性就这么大。如果我们在曹操家族里面检测出来的类型，在汉族里面普遍地存在频率只有百分之一呢？那么这个可靠性是不是更大？最后的可靠性结果要增加好几个九，要增加一倍的九。

所以我们知道，要提高这个可靠性，要检测出来曹操的基因类型的可靠性变大，有两种办法：一种就是把这个类型的频率数值变小，或者我们要找到越来越多的家谱，多几个乘号，找到相同类型的家族越多，那么可靠性也就越大。但是曹操的类型不是我们定的，所以我们唯一的办法就是找越来越多的家谱，号称曹操后代的家谱我们找得越多越好，这样的话，可信度就越大。为了这个事，韩昇老师就给我们找出了两百多份家谱，然后一一辨别，看谁记的是曹

图 20　曹姓的采样调查

操的后代，一个一个让我们按图索骥地去找。所以我们就跑了很多地方，到处去采样，我们定了很多点。从辽宁到广东，这些点有的是有家谱，有的在文革时候破四旧把家谱都烧了。因为当时编家谱是反动的，没有人敢收藏家谱，所以就烧了，但是他们记得家谱上面说是曹操的后代，所以我们就姑且都算。我们发布了一个人肉搜索的新闻之后，很多人都来了，各种各样的人。有的地方人少，他们就自己过来，人多的地方我们就过去拜访。在调查中发现了很多有趣的现象。有一个家谱非常离谱，说自己是周文王封的曹姓始祖曹叔振铎的后代，传到他已经有八九十代，比孔子家的家谱还要完整，到底该信还是不信？我们都不管，让 DNA 说话。

那如何根据这么多样本来推断曹操 Y 染色体的序列呢？我们

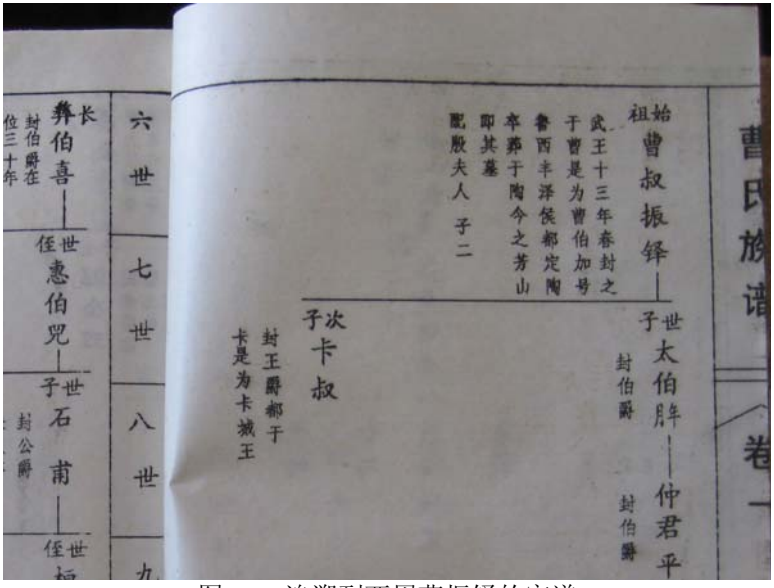


图 21　追溯到西周曹振铎的家谱

知道，光定一个型，比如说 X 型，再去验曹操遗骨，只能知道这个遗骨是出自曹操这个家族，因为曹操家族很多人，这些人一般都是这个 X 型。只能判断这个骨头是不是曹操家族的，不能判定这骨头是曹丕的、曹植的还是曹操的，或者是曹操的叔叔、伯伯。怎么准确地判断这个骨头就是曹操呢？那就需要从后代的 Y 染色体的谱序中推断曹操的完整的 Y 染色体突变的谱序。我们知道每个人 Y 染色体都不同，我们要找到曹操一个人的 Y 染色体突变谱的结构，他个人特别的突变点，才能判断这个人是不是真的曹操。要了解曹操的个人特有的突变的寻找原理，我们就要先了解 Y 染色体为何人人不同。前面提到 Y 染色体每个人每个点的突变概率是三千万分之一，非常小，落到我们可研究的稳定的这三千万个点上的话，每个

图 22 牌匾和家谱上注明的家族起源

人都会有一个新的突变，这是每个人跟他的父亲不同的一个新的点，跟其他所有的人都不同。所以 Y 染色体几乎是人人不同的。但是新的突变随机地出现在三千万个点中，我们要找到它并不容易。我们必须要把每个人的 Y 染色体从头到尾测一遍，把它的序列完全测出来，然后再比对一下，看到每个人新发生的这个点到底在哪里。三千万个点中找到一个点，大海捞针一样，这个成本非常高。在家系中间，我们知道，每一代都会在祖先的突变上面加一个点，平均地一代一代加新突变点，有几代就会大约有几个突变点累加。所以我们也就可以用这个方法来推断这个家族到底是传了几代。比如孔子的家族到了现在到底是几代孙，有没有搞错掉，都可以通过这原理来验一下。另外，每个分支的突变谱不一样，曹植的分支突变下

来，从曹植开始就与其他分支不同了，每个分支都开始走上分歧的路。曹丕后代的序列跟曹植后代的肯定也不一样，共有的东西就是曹操的东西。我们在那么多家系里面找到了号称曹植的后代有一大批，号称曹丕的后代也有一大批，这样子的话，分析起来那就很好办。

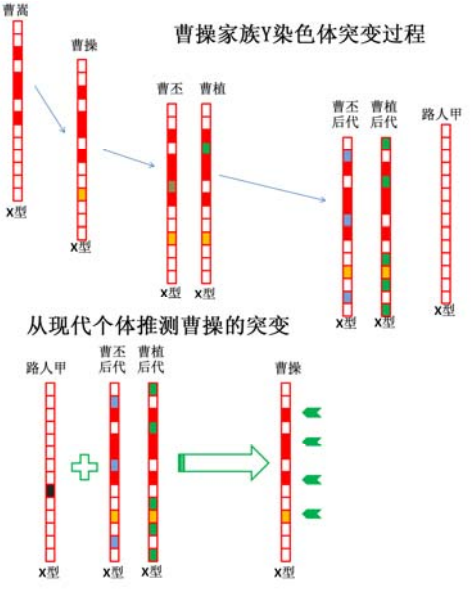


图 23　曹操家族的 Y 染色体变化过程以及从现代个体逆推曹操 Y 染色体的原理

图 23 中我们可以看到，根据突变序列，曹操的父亲是曹嵩，曹嵩和曹操都是 X 型，但是曹操可能会比曹嵩多一个突变点。曹操的后代曹丕和曹植，曹丕是在曹操的基础上加一个点，曹植在曹操的基础上也加一个点，但这两个点肯定是不同的，因为新增加的点

是随机地出现在三千万个点中的。从曹丕一代一代传下去，曹丕的后代加了很多很多点，我们在图上先标三个。传几代就标几个点，曹植的后代也是穿几代加几个，加在完全不同的地方上。所以我们知道曹植后代和曹丕后代的 Y 染色体都有很多不同。我们又要找一个完全不同的参照，一个 X 型的路人甲，他与曹植曹丕的后代序列肯定也是完全不同的。曹操有的点，路人甲肯定都没有，所以曹植、曹丕的点在路人甲里肯定是没有的，虽然都是 X 型。有了这样的发展脉络，我们反过来去看，通过路人甲、曹丕后代、曹植后代的类型去推曹操的类型，那就好办了。

路人甲他有一些突变点，是与曹植后代和曹丕后代完全不同的；路人甲没有的突变点，而曹植和曹丕的后代里面都有的，那就属于曹操家族的突变点。所以路人甲的序列是我们鉴别曹操家的新突变的对照。在分析的时候，曹植和曹丕后代之间不同的那些点我们全部剔除掉，我们剔除掉这些点以后，到最后留下的点就是曹操的类型，就是曹操的突变谱。我们就在曹操的突变谱当中挑几个点，去验证遗骨上面有没有这几个点，就能够明确地判定他是不是曹操。当然我们不能光验这几个点，要验一下曹植和曹丕特有的点在这个遗骨上有没有，以判断他不是曹丕而且不是曹植。当然这些分析可能都不需要做，因为这个骨头一旦是曹操家族的话，按照这个墓的形制，多半就是曹操的。说不定用不着花这么多钱研究这么多点，只要检验一个点就够了。这个成本实际上是很低的，而且可行性非常高。从古 DNA 里面验一个点，这个难度非常小。只要 DNA 抽取出来，多半就能做。

如果需要否定他是曹丕或者曹植的话，怎么判断曹丕或者曹植应有的点呢？我们知道曹丕和曹植传下来很多后代。图 24 中这些

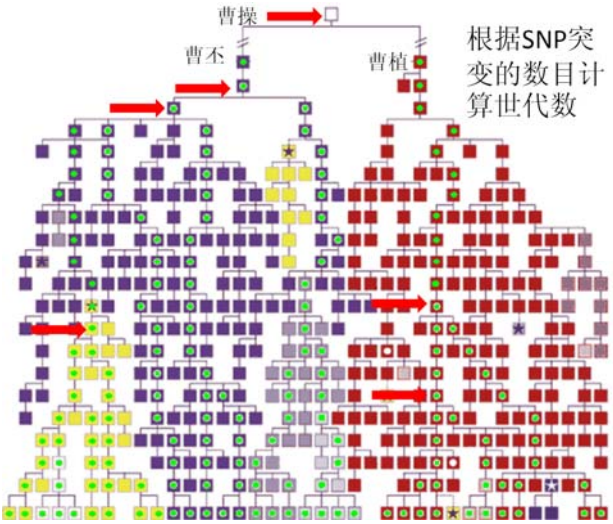


图 24 复杂的传代过程中可以推测的祖先位置

绿色的点表示后代还有现存的，他们的 Y 染色体传下来了。通过推断，我们知道通过两个分支的末端能够推断分叉点处的个体的类型，知道他的 Y 染色体是什么类型的。只要有两支后代保留到现代，我们就可以推断他们的最近的共同祖先的类型。实际上曹丕有好几个儿子，都有可能传下后代，所以推到曹丕都有可能。我们知道曹丕和曹植都有后代，所以推断出曹操就会很方便了。所以基于这种原理，我们就能够找到曹操独一无二的那些突变标记，能够非常准确地依此来推断疑似曹操遗骸到底是不是他本人，而不是曹操家族的其他人。

最后还有一个问题是很多人一直在问的，近两千年前留下的骨骼还能分析 Y 染色体 DNA 么？我们知道父系的 Y 染色体的分析要比母系的线粒体难得多。线粒体染色体是个环形结构，结构稳定，

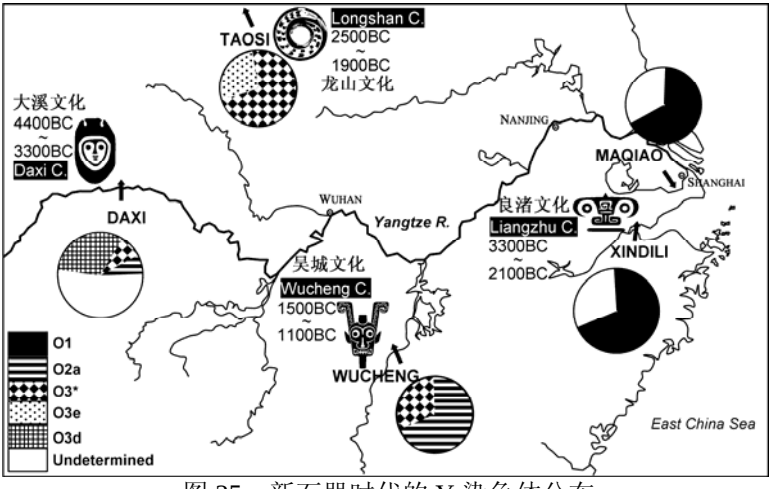


图 25　新石器时代的 Y 染色体分布

拷贝数多，所以一般古 DNA 实验室都能够做线粒体检测。但是线粒体不能够解决我们的问题，因为她是母系遗传的。现代古 DNA 技术突飞猛进，我们早就可以分析古人的 Y 染色体，分析古代 Y 染色体不是今天才想到的问题。2007 年的时候，我们在德国出版的科学期刊《Human Genetics》杂志上发表了一篇文章，把长江流域的各种考古文化，新石器时代的各种人骨遗骸分析了 Y 染色体，这些考古文化包括有良渚文化的两个点，有吴城文化的，有大溪文化的，很多样品距今都有五千年到六千年了。所以五六千年前的样本都能做，更别说曹操这种两千年前的样本了。在分析新石器时代的样本的时候，汉代的样本我们当时都是当作阳性对照的。阳性对照是什么？就是肯定能出结果的东西，所以曹操的骨头就更不在话下了。实际上古 DNA 的技术发展是非常非常快的。从八十年代初开始兴起，通用分子克隆的方法；从八十年代中期到九十年代

中期的时候，开始流行用 PCR 方法，这个技术出现了之后，古 DNA 分析技术开始流行了；到九十年代以后，就开始更大规模的发展，标志事件是尼安德特人的标准序列的建立，而且很多严格的控制手段都建立起来了，有了一个行规；到了现在，就进入了基因组时代，基因组时代对于古 DNA 来说就是检测全序列，全部都是全测序的方式得到结果。古 DNA 技术的发展都是得益于分子生物学的技术发展。

古 DNA 分析的技术发展有三次革命。实际上研究古 DNA 最早的是中国人。中国 1980 年的时候从长沙马王堆的尸体里面就提取出了 DNA，这是最早的古 DNA 的研究。虽然我们当时没有技术进行 DNA 的测序，但是我们知道可以从古代尸体里面拿到 DNA，就证明了古 DNA 是可以研究的。这是一个开创性的事件。后来通过分子克隆，就是通过细菌培养或者载体培养，把古 DNA 转载到载体里面，进行扩增。这样的方法就引发了古 DNA 研究的最早的一次高潮。期间得到的成果包括斑驴的 DNA 鉴定。斑驴在 140 年前灭绝的，博物馆里有保存了 140 年的斑驴的标本，在 1984 年进行了分析，得到了它的线粒体的 DNA，经过对比得知斑驴实际上跟斑马很接近，跟驴子不接近，非常有意思。85 年的时候，又从一个 2400 年前的木乃伊上克隆到了一段片段，分析出来很多结果。

第二次革命叫做 PCR 技术的革命，特别是后来发展出来的多重 PCR 技术和微测序技术，解决了很多问题。成功的案例就像从猛犸象里面得到的线粒体的全序列；从安达曼人（一种亚洲小黑人）很久以前在墓葬里面得到的遗骸里面得到了线粒体的 DNA。安达曼人分布在缅甸外海属于印度的一个小群岛上，离大陆非常远，他们的人种跟大陆上的人种完全不同，长得又矮又小，皮肤特别黑，但是他们又是在亚洲，他们的来源当时又很不清楚。但是可以肯定的是，安达曼人与我们的差异根本就是个人种的差异，不是民族的差异。安达曼人原来有很多民族，讲很多种语言，自从外界的人员介入他们以后，带给他们很多病菌、病毒，使大部分民族都死绝了。现在

还剩下两三个民族还活着，十几个民族都死绝了。对他们的古 DNA 研究对那些已经灭绝的安达曼人民族的数据是一种重建，所以价值也非常大。

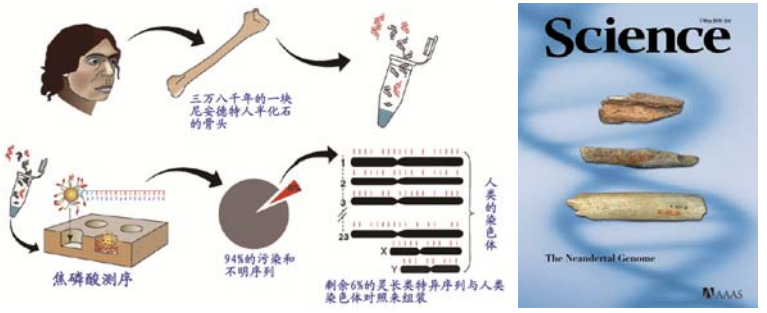


图 26　《科学》封面文章发表了尼安德特人全基因组的研究成果

最新的一个测序技术叫做下一代测序技术，这就引发了古 DNA 的第三次革命。这个新测序方法不是通过对模板破坏性的毛细管电泳来测序，而是通过序列的合成过程来进行测序。这种新的测序过程就使得我们不需要依靠 DNA 的拷贝数量来进行测序，而是我们只要有一个模板，在合成的过程中，边合成边测序。也就是说，古 DNA 的拷贝数量虽然一般都很少，但是用新方法也可以进行测序。下一代测序技术有很多种方法，比如焦磷酸测序方法、SOLiD 法。好几个公司都发明了各自的方法进行测序。分子生物学的技术真是一天一个样，所以用在我们的领域内，古代 DNA 的研究也得到了大量惊人的结果。今年的 5 月 7 日，在《Science》上面作为封面文章发表了尼安德特人的全基因组数据，什么概念？我们人是 30 亿个碱基，尼安德特人差不多也 30 亿个，那么大一个数量的数据全部都公布出来了。而这个样本距今多少年呢？三万年。三万年前的标本我们能够得到它的全基因组的数据，更何况曹操两千年都不到的 Y 染色体，又有何难？在古 DNA 技术上是没有难度的。当然科

学研究都是探索的过程，不是那么确定的，虽然在技术原理上没有难度，但在实际操作上，这骨头到底是保存条件怎么样，我们到现在为止我们都没有亲眼看到。我们不能打包票说结果肯定能出来，但是至少从目前的技术手段上，应该是没有什么难度的。

但是很多考古学专家一直在说我们国内没有这么好的技术，只能测母系线粒体。国内的技术怎么可能只停留在这一点？华大基因公司能够测出很多年前冰冻的爱斯基摩人的全基因组序列。我们复旦的技术更不可能仅停留在这一点上，只能测母系线粒体这是十几年前的事了。

但是另外还有一个问题是，如果我们在曹操的后代中间把曹操的类型的可信度已经提高到了非常高的程度，比方说 24k“足金”，百分之 99.99…9%，24 个 9，那我们还要不要去测曹操的骨头呢？我们都已经知道曹操的类型了。

（王涛 整理）