与澳大利亚国旗具有相通之处，色彩鲜明，易于接受。第二，ATR官方网站（网址：http://www.twins.org.au）。该网站覆盖ATR常规工作内容，研究者可以方便快捷查询到自己所需要的信息，界面友好，便于使用。第三，双生子研讨会。举办以双生子研究为主题的研讨会，创造机会让研究者了解已经开展的双生子研究，培训研究者如何进行设计、现场实施以及数据分析等。第四，建立双生子研究者网络。主要通过电子邮件的方式让双生子研究者了解ATR近期活动、发表文章、举办会议、合作机会甚至研究经费等。第五，旅行奖励。资助学生或者研究人员参加培训或者国际会议，每年提供2次，每次数额6～9名，资助额度每人从300～2000澳币不等。第六，ATR通讯。ATR自1983年开始坚持每18个月出版一期通讯，并且保证寄到每名双生子手中；通讯介绍一年来ATR的大事记，以及有哪些研究正在募集双生子；这份通讯保证了双生子和ATR的长期联系和良好的关系，并为需要募集双生子的研究进行了有效的宣传。第七，全国巡回的双生子节日。最近一次于2009年3月在悉尼举办。第八，媒体宣传及平面广告。实践证明，印有ATR标志的冰箱贴最为有效。第九，参加研究的双生子能够接受免费的健康体检，而且ATR要求研究者务必将双生子的检查结果反馈至双生子本人。

6. 小结：综合ATR的管理及发展，具有如下特点：第一，政府支持，资金保障：从ATR建立之初，澳大利亚政府就给予一定的资金支持，发展成为现在每5年一轮逐步增长的经费保障；第二，专门机构，长期工作：ATR以项目为基础，设立ATR办公室，聘请项目官员专职进行双生子、研究者和双生子研究信息的管理工作，长期系统，历经近40年，ATR已经逐渐发展成熟，形成具有特定目标、政策以及规范操作规程的研究平台机构；第三，定位明确，服务平台：ATR将其定位为双生子研究的平台，而非双生子研究者本身，作为双生子研究对象和研究者的桥梁，ATR在保证双生子权利的前提下促进双生子配合双生子研究，2010年ATR将进行双生子及研究者满意度调查，此项调查正是反映了ATR明确的服务机构定位；第四，内容集中，登记有效：由于明确的服务平台定位，而非某项专门的研究，因此ATR致力于双生子联络方式的及时更新，保证双生子登记信息的有效和对于双生子研究的可及性；第五，受众广泛，学科交叉：目前已有数百项研究应用双生子资源，ATR曾有助于回答覆盖心理、行为、心脑血管疾病、骨代谢、眼科等多种疾病或者性状的研究问题，有研究者甚至将ATR的资源平台作为在澳大利亚为什么要进行双生子研究的一个原因[4]；第六，注重合作，影响全球：ATR重视来自全球的合作，在笔者走访期间，慷慨提供各种资源为笔者参考学习，并且作为国际双生子协会的发源地，ATR的不断成长为全球许多国家双生子登记系统提供了可供借鉴的经验。

中国双生子登记系统（Chinese National Twin Registry，CNTR）从2001年开始，已经步入第10个年头，如何更好地服务于国内双生子研究，亦需要不断总结并借鉴国外双生子登记系统的成功经验。ATR作为一个开放的、共享的双生子资源，为公共卫生和生物医学研究的整合提供了典范。具体到CNTR，首先要继续明确定位，实现从一个研究者向一个研究平台的过度；其次，建立CNTR官方网站，并开通网络募集双生子通道；第三，继续保持和澳大利亚以及其他国际双生子同行合作，扩大双生子登记系统的国内和国际影响力并提高双生子研究水平；第四，积极争取政府支持，长期获得经费保障，实现双生子登记系统的可持续发展。

参 考 文 献

[1] Andreas Busjahn, Yoon-Mi Hur. Twin Registries: An Ongoing Success Story. Twin Res Hum Genet, 2006, 9(6): 705.
[2] Hopper JL, Treloar SA, de Klerk NH, et al. Australian twin registry: a nationally funded resource for medical and scientific research, incorporating match and WATCH. Twin Res Hum Genet, 2006, 9(6): 707-711.
[3] Hopper JL. The Australian twin registry. Twin Res Hum Genet, 2002, 5(5): 329-336.
[4] Wark JD, Nowson C. Influence of nutrition on bone health: the twin model approach//New SA and Bonjour JP, eds. Nutritional aspects of bone health. Cambridge: Royal Society Chemistry, 2003: 451-461.

（收稿日期：2010-01-12）
（本文编辑：尹廉）

基因型填补的原理与方法及其在遗传流行病学研究中的应用

莫兴波　顾东风

【关键词】 基因型填补；遗传流行病学；全基因组关联研究

Genotype imputation: principle, methods and application in studies on genetic epidemiology

MO Xing-bo, GU Dong-feng. Division of Population Genetics and Prevention, Cardiovascular Institute, Fu Wai Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100037, China
Corresponding author: *GU Dong-feng*, Email: gudongfeng@vip.sina.com

【Key words】 Genotype imputation; Genetic epidemiology; Genome-wide association study

DOI: 10.3760/cma.j.issn.0254-6450.2010.06.026
作者单位：100037 北京，中国医学科学院北京协和医学院阜外心血管病医院群体遗传与人群防治研究室
通信作者：顾东风，Email：gudongfeng@vip.sina.com

基因型填补（genotype imputation）是指依据已分型位点

的基因型对数据缺失位点或未分型位点进行基因型预测的方法。在全基因组关联研究(genome-wide association study, GWAS)中,基因芯片分型的位点大约为10万~100万个。芯片的分型缺失率较高,一定程度上降低了芯片的密度。对这些芯片产生的数据进行填补可以提高全基因组遗传标记的覆盖率及研究的效率,增加阳性关联位点的筛查成功率。基因型填补可应用于精细定位(fine-mapping),填补已确认的关联位点附近的位点,以便评价相邻SNP位点的关联证据,加快复杂性疾病易感基因的定位。同时,基因型填补还可降低直接分型的成本。此外,由于不同基因分型平台所选择的SNP位点只是全基因组中所有SNP位点的一部分,不同平台所分型的SNP位点并不相同,对这些基因型数据直接合并将会造成大量的基因型缺失,所以研究者在整合这些数据进行联合分析或对这些研究进行Meta分析时需要对各个平台产生的数据进行填补。目前已有很多成功地对几个GWAS进行Meta分析的研究,这些研究不仅可以增加已发现的关联位点证据,还可以发现新的关联位点,例如2型糖尿病[1]、血脂水平[2]、冠心病[3]、身高[4]、体重指数(BMI)[5]、克罗恩病[6]等疾病或表型的关联研究。

1. 连锁不平衡与基因型填补:

(1)连锁与连锁不平衡:连锁与连锁不平衡(linkage disequilibrium, LD)是两个重要的遗传学概念。如果同一条染色体上2个位点的位置比较近,则这2个位点上的等位基因倾向于一起传递给下一代,这一现象在遗传学中称为连锁(linkage)。LD与连锁有关,但两者是不同的概念。连锁描述的是位点位置关系,可以通过重组率来衡量,它是连锁分析的理论基础。而LD与位点上的等位基因概率有关,即如果不同位点上的等位基因不是独立出现,则称它们处于LD状态,它描述的是群体在不同位点上的等位基因关联性,是关联分析的理论基础。LD多用于研究进化和基因定位[7,8]。

考虑2个位点A、B,各自具有2个等位基因A、a和B、b。如果这2个位点联合配子概率等于边缘配子概率的乘积,即:P(AB)=P(A)P(B)、P(Ab)=P(A)P(b)、P(aB)=P(a)P(B)和P(ab)=P(a)P(b),则这2个位点处于连锁平衡状态,否则为LD。可以定义LD参数$D_{AB}=P(AB)-P(A)P(B)$,现在一般用r^2来衡量位点间LD大小,r^2定义为:

$$r^2 = \frac{D_{AB}^2}{P(A)[1-P(A)]P(B)[1-P(B)]}$$

在世代传递过程中,LD位点的等位基因倾向于一起传递给后代。同一条染色体上一起传递的等位基因构成一个单体型。新生突变或染色体重组都可以产生新的单体型[9]。一个个体在若干位点上有2个单体型,它们包含了位点之间的连锁信息。在一个人群中,单体型的频率可以代表它所涵盖区域内所有遗传标记的信息。

(2)基因型填补:基因型数据填补的理论依据是位点间的LD。LD模式在不同人群中是不同的,但是在同一人群中,LD模式相对固定。利用现有特定人群样本的单体型结构信息(如HapMap)作为参照,可以很准确地推断相应人群研究样本中未分型位点或数据缺失位点的基因型。基因型填补示意图中前6个个体的基因型数据可以作为参照面板,未分型位点和缺失位点的基因型都可以根据参照面板的完整数据来填补(图1)。填补时首先对这些位点进行单体型分析,找到个体间匹配的单体型。由于同一单体型中SNP位点之间存在LD,所以只要找到匹配的单体型,就可以在相应单体型中用已分型位点的数据来填补缺失位点和未分型位点数据。

C/G	G/G	G/T	G/G	A/T	T/A	C/C	T/T	T/T	C/G	T/T	A/T	C/G	↑
C/C	A/T	A/G	A/T	G/A	C/A	T/A	C/G	T/A	T/T	T/A	A/T	C/G	参
C/G	G/G	A/T	G/A	G/A	C/A	T/T	C/G	T/T	C/C	C/G	G/G	T/T	照
A/G	C/G	G/G	G/G	C/G	T/T	T/T	C/G	C/G	T/A	A/G	T/T	A/G	面
T/T	A/T	A/T	G/G	A/T	C/G	T/T	T/T	T/A	T/A	T/T	A/T	A/T	板
C/C	A/T	G/G	G/A	A/T	T/T	C/G	T/A	C/C	A/A	C/C	C/G	G/G	↓
T/A				G/G		C/G				T/A	A/G		
T/A				T/T		G/G				T/A	T/T		
A/A				A/T		?				C/G	C/C		
?				C/G		T/T				A/G	T/A		
T/T				G/G		?				T/T	?		
C/G				G/C		G/G				G/G	T/A		

注:图中每一行代表一个个体,每一列代表一个位点。黑体的位点是对所有个体都进行分型的位点,其他位点仅在部分个体中进行分型,"?"表示该位点基因型缺失,空白区是未分型位点

图1 基因型填补示意图

家系数据的基因型填补比较直观,家系成员中共享的染色体片段比较长(一般长达几百万碱基对,包含数千个SNP),这些片段称为同源一致性(identical-by-descent, IBD)区域。利用家系成员间共享IBD区域进行基因型填补最早由Burdick等[10]提出。在家系资料中进行基因型数据填补时,可以对家系的部分成员进行较多位点的分型,对其余成员只进行少数位点分型,如一个有三代人的家系,包括祖父母、父母和子女,则可以对祖父母和父母进行较多位点的分型,对子女则进行较少位点分型。然后利用父母和祖父母的基因型数据作为参照面板,填补子女的未分型位点。

在无关个体中进行基因型填补与家系数据的基因型填补相似,区别在于无关个体共享的IBD区域比较短(可以理解为他们的共同祖先较远),使得寻找匹配的单体型变得困难,不过这些短片段单体型仍然可以为基因型填补提供有用信息。对无关个体进行基因型填补时,首先对研究样本在适量的位点进行分型(一般为10万~100万个位点),将这些样本数据与包含较多位点含有较详细信息的参照面板进行比较,确定研究样本和参照面板样本共享的染色体片段,最后结合分型位点和共享单体型信息,填补研究样本中缺失或未分型位点。

在实际应用中并不是所有填补完成的位点都可以用于分析。基因型填补存在一定的错误率,填补错误的基因型反过来会影响GWAS分析,填补质量不好的位点需要去除。常用的填补方法都有相应的指标用于衡量填补质量,如MACH的r^2[11],现有的研究一般选择$r^2 \geq 0.3$位点用于关联分析。影响填补准确度的因素主要包括LD的强度、未分型位点最小等位基因频率、遗传标记密度和参照面板的大小。Pei等[12]的

研究显示，基因型填补的准确度随LD强度增加而增加。在相同的LD水平下，填补准确度随未分型位点最小等位基因频率的增加而降低，随参照面板样本量和遗传标记密度的增加而提高。LD程度越低填补准确度变动的幅度越大。

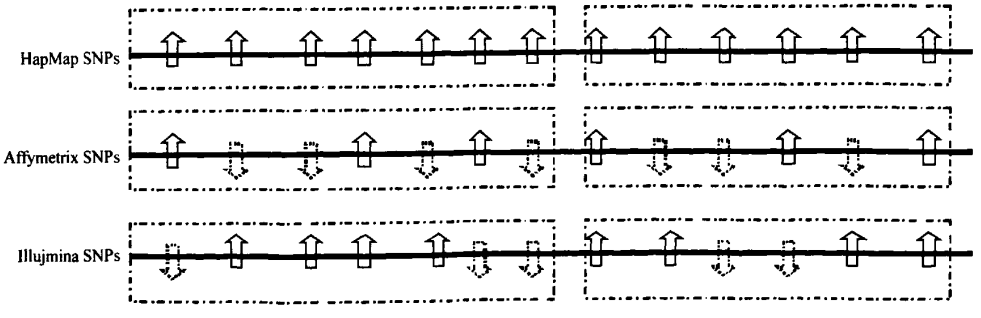
2. 质量控制和数据整理：对基因型数据进行填补之前要进行严格的质量控制，以避免或减小填补误差或偏倚。基因型填补之前对原始数据的质量控制与GWAS数据的质量控制类似。PLINK是一个常用GWAS工具——包括数据的质量控制[13]。统计分析软件R（http://cran.r-project.org/）的一些GWAS扩展包，如GenABEL等，也可以用于遗传数据的质量控制。

对于样本的质量控制包括基因分型缺失率＞3%[WTCCC（wellcome trust case control consortium）[14]对7种常见疾病进行GWAS时采用的排除标准）]、杂合率＞30%或＜23%[14]、排除重复（相似性＞99%[14]）或亲属（相似性在86%～98%之间[14]）样本（基于一般人群的研究设计）等。对于SNP位点的质量控制包括控制最小等位基因频率（MAF＞1%）、缺失率（排除缺失率＞5%的位点，MAF＜5%的位点阈值为1%[14]），检验哈迪-温伯格平衡（$P<5.7\times10^{-7}$[14]）和孟德尔错误（Mendel error）等。另外，对基于一般人群的研究还应进行人群分层分析，对于病例对照研究还应考虑病例对照组之间缺失率的差异等。

由于不同研究在研究设计和原始数据质量等方面存在差异，质量控制的阈值并没有统一标准。在具体应用中设置质量控制阈值时，需要根据研究本身的特点选择合适的阈值。对于分型率而言，如果设置的阈值较高，那么很多位点或样本将因分型率太低而被排除，这样有可能会把真正的关联信号排除，或导致假阳性率的增加；如果对分型率选择较低的阈值，那么质量控制后的数据质量仍然较差，而分型率较低的位点仍然会影响基因型填补的准确性。对于哈迪-温伯格平衡检验阈值则应根据研究的样本量及原始数据的质量等进行设置。

在对SNP位点进行基因型填补时，明确每个位点名称或者*rs*号及其在染色体上的位置和等位基因所在DNA链的方向（可以从NCBI的dbSNP数据库中查到）非常重要，这些信息对单个GWAS而言并不是必须的，但是在对多个全基因组研究进行Meta分析时却非常重要。在dbSNP中查找SNP位点的相关信息时要明确数据库的发布版本（如build 35或36），因为不同发布版本之间存在差异，比如SNP位点的*rs*号在不同的发布版本中可能不一样，等位基因链的方向也可能不一样。基因型填补时样本的等位基因和参照面板的等位基因必须在DNA双链的同一条链上，即要求都是在正链（forward/＋ strand）或者都是在负链（reverse/— strand）上。对于样本数据，芯片分型位点等位基因所在DNA链的方向可以参考芯片公司的相关说明文件，如Affymetrix的NetAffx文件。但是并不是所有的基因分型芯片公司都提供相关的说明文件，在这种情况下，可以通过比较样本数据和HapMap数据库中SNP位点等位基因频率的差异来分析等位基因的链方向，位点基因型数据不匹配时则进行正负链的转换，即把某位点负链上的A/G基因型转换成T/C基因型，于是该位点就变成了正链。但是这种方法也不能把所有错误检测出来，比如A/T或C/G基因型，因为A和T、G和C是互补的碱基，基因型A/T与T/A或者C/G与G/C无法区分。这时可以考虑将这些位点排除，如Illumina平台会把这样的位点排除。另外，对于HapMap的SNP数据，release 21a之前发布版本都是基于NCBI build 35，release 22及更高的发布版本则都基于NCBI build 36，在应用HapMap数据时也必须明确数据库之间对应的不同发布版本。

3. 参照面板：目前，大部分填补方法都以国际单体型图计划（the international hapMap project）第二阶段的数据作为参照面板（reference panel）。HapMap第二阶段数据包括了来自欧洲（CEU）、中国北京（CHB）、日本东京（JPT）和非洲（YRI）270个个体的超过310万个SNP位点信息[15]。HapMap的SNP位点数量现已近400万个。参照面板的选择主要依据研究人群来确定：欧洲人群的研究则应该选择CEU数据作为参照；中国汉族人则可选择CHB数据作为参照；至于其他一些没有相应样本数据的人群，如中东地区的一些人群、美国的土著居民等，则可以选择包含HapMap的3个样本的混合样本数据作为参照[16]。已有的基因型填补研究基本上都基于欧洲人群，这些研究都以HapMap的CEU样本作为参照面板。模拟分析和实际研究都证明通过基因型填补获得的基因型数据具有很高的准确度，即证明了使用HapMap数据做参照面板的可行性。另一种获得参照面板的方法是从研究样本中选择一个子样本，对这个子样本的个体进行较多



注：向上箭头表示已分型位点（3个数据集中SNP位点存在差异），虚线方框表示对数据进行单体型分析后确定的匹配单体型，单体型内的SNP位点间存在LD，因而可以用HapMap中的SNP信息来填补研究样本中未分型的位点。图中虚线向下箭头表示填补的位点

图2　不同分型平台产生数据间的比较与填补示意图

位点的分型，这个子样本的基因型数据就可以作为一个参照面板。如Chambers等[17]开展的一个肥胖相关研究中就采用了这种方法，这种方法比直接使用HapMap数据费用更高，但是这种方法填补的基因型具有更高的准确度。

4. 不同平台数据的比较与合并：到目前为止，GWAS的分型平台基本上都来自Affymetrix和Illumina两家公司。不同公司在设计芯片时挑选的位点不一样，来自不同公司的芯片所分型位点差异很大，同一公司不同产品之间也存在差异，所以不能直接合并和分析不同平台产生的数据，也不能直接对不同的GWAS进行Meta分析。一种保守方法是将分析限制在所有研究中都存在的SNP位点。这种方法降低了遗传标记的基因组覆盖率，所以不可取。另一种方法则将这些不同研究的数据与标准参照面板比较进行填补。目前常用方法是利用HapMap的单体型信息填补不同研究的数据，填补位点可达到两百多万个。这种方法不仅可以使不同的研究之间拥有较多相同的SNP位点，还可以提高研究的效率。

不同GWAS数据之间的比较和填补示意图见图2。图中基因型数据的填补是以HapMap数据作为参照面板，分别对来自Affymetrix和Illumina公司的分型平台产生数据进行填补。Zeggini等[1]结合WTCCC、DGI（diabetes genetics initiative）和FUSION（finland-united states investigation of NIDDM genetics）3个GWAS，分析了10 128人、约220万个位点（直接分型位点加上填补位点），重复验证了这些研究发现与2型糖尿病相关联的位点，并且额外地发现6个新的关联位点。在Zeggini的Meta分析中，WTCCC研究样本利用Affymetrix GeneChip Human Mapping 500k分型平台，分型SNP位点（经过质量控制后）为393 143个；DGI利用Affymetrix GeneChip Human Mapping 500k分型平台，分型SNP位点（经过质量控制后）为378 860个；FUSION利用Illumina HumanHap 300 BeadChip分型平台，分型SNP位点（经过质量控制后）为306 222个。这些研究所分型的SNP位点中，44 750个位点同时存在于3个研究样本中，308 628个位点同时存在于其中的某2个研究样本中，245 158个位点只在其中1个研究样本中进行分型。由此可见不同分型平台产生的数据之间存在差异，所以分析时如果将这3个研究数据直接合并将造成大量基因型数据缺失。Zeggini等用Hapmap第二阶段的60个CEU无亲缘关系样本数据作为参照面板[CEU样本包含30个核心家系（每个家系包括父母和一个孩子），每个个体分型的SNP位点有300多万个[11]，该研究只取每个家系中父母的数据]，在所有研究样本中共填补1 570 311个位点，分析时的总位点数达到2 168 847个。

5. 常用算法和一些软件：基因型填补所用的参数估计方法包括期望最大化算法（expectation maximization algorithm，EM）和马科夫链蒙特卡罗算法（Markov Chain Monte Carlo，MCMC）。

（1）EM：EM是统计中用于寻找概率模型参数的极大似然估计（maximum likelihood estimate，MLE）的迭代方法。EM算法包括期望步和最大化步。在进行基因型填补分析时，首先为每个缺失的基因型随机地选择一个值和能够发现已测基因型的似然值，然后检验这些假设值是否正确。这些缺失基因型的估计值将经过多次迭代来校正，直至估计误差最小。EM可以捕捉到局部最大值，所以它会以各种不同的起始状态（即第一步随机选择的值）多次运行以保证填补的准确性。

（2）MCMC：另外一种方法是利用贝叶斯模型（Bayesian Models）估计缺失的基因型，目前主要应用的是MCMC。MCMC是计算机随机模拟方法，它也是一种迭代算法。MCMC区别于EM的地方在于它考察的是整个参数空间，而不仅仅只是局部最大值，这样将有利于发现那些频率较小却对疾病或表型有重大影响的等位基因。这种方法需要迭代次数较多，计算时间也因而增加。

（3）常用软件：现有的基因型填补方法主要基于以下几种统计模型：单体型聚类算法[18]、隐马可夫模型[19]、马科夫链模型[11,20]。现在已有不少免费软件可用于基因型填补。这些软件都是命令行软件（大部分都是在Unix/Linux系统下运行，也有既可以在Linux下也可以在Windows下运行的命令行软件，如MACH）。Li等[16]将这些软件归为两类：第一，填补每个缺失基因型时考虑所有分型的位点，这类软件包括：IMPUTE[19]、MACH[11]和fastPHASE/BIMBAM[15,20]。第二，填补时只考虑基因型缺失位点附近的一些已分型位点，这类软件包括：PLINK[13]、TUNA[21]、WHAP[22]、BEAGLE[23]。目前MACH和IMPUTE用得最多。不同填补方法对基因型估计的准确度不同。Pei等[12]和Nothnagel等[24]分别对这几种方法进行了比较，他们的研究结果均显示MACH和IMPUTE的填补准确度相似，且这两种方法的准确度高于其他方法。

MACH（Markov Chain Haplotyping）是美国密歇根大学生物统计系开发的一个软件，这个软件是基于马科夫链模型设计的，根据个体的基因型推断单体型。这种算法做单体型分析时先随机地选择一对与已观测的基因型匹配单体型，并且为转移概率（模型的参数）估计一个初始值。在运算过程中，单体型对（Haplotype pair）会不断地通过蒙特卡罗方法迭代更新。每次迭代都利用HMM对每一个个体抽取一对新的单体型，模型参数也在每次迭代中得到更新。经过多次的重复和更新后可以得到一对真正匹配（或匹配概率最大）的单体型。基因型填补的分析过程与上述单体型分析过程类似，区别在于它在迭代过程中产生一系列的计数。这些计数记录了对每个基因型抽样的次数，反映了可能观测到某基因型的相对概率，抽样次数最多的基因型即为可能性最大的基因型。用MACH填补时的迭代次数一般是50～100次，迭代次数越多则填补越准确，但所需时间就越长。

IMPUTE是牛津大学的Bryan Howie和Jonathan Marchini开发的一个基于HMM的软件。IMPUTE在填补分析时假设每个个体之间的基因型是相互独立的。它把已知单体型对序列看作是HMM中的隐状态，同时定义一个以已知单体型（参照面板中的单体型信息）为条件的条件概率作为转移概率，用这些隐状态和转移概率建立HMM，即根据已知单体型估计缺失基因型。为方便比较，IMPUTE会输出每

个可能基因型的后验概率，具有最大后验概率值的基因型则为用于填补的基因型。

fastPHASE是华盛顿大学开发的基于HMM的单体型聚类算法的软件，这个软件应用基于聚类模型的方法估计单体型，并且假设每一个单体型都从某一个聚类中产生。fastPHASE用EM算法估计模型参数，利用基于HMM中隐变量的条件分布计算缺失基因型在已观测基因型和估计的模型参数条件下的条件概率，使这个条件概率最大的基因型则成为该位点基因型的填补基因型。

BEAGLE是奥克兰大学开发的一个软件，这个软件利用局部单体型聚类(localized haplotype clustering)方法定义一个二倍体HMM(diploid HMM)，该模型可用于推断单体型以及缺失数据。Beagle和fastPHASE都是基于HMM单体型聚类的方法，它们之间的区别在于fastPHASE在估计模型参数时使用的是EM算法，而Beagle用根据当前所估计的单体型经一步算法计算得到的经验值作为参数。这意味着fastPHASE若要在一个合理的时间内进行参数估计，它就需要对某些因素进行控制，如固定聚类的数量。正因为这样，在建立模型时fastPHASE在建立模型时聚类数目是固定的，而Beagle建立模型时在每个位点都可以改变聚类的数目。

基因型填补对计算机性能要求较高，而且不同软件工作时对计算机硬件需求和平均计算时间存在较大差异。Nothnagel等[24]的研究中显示，IMPUTE在工作时占用了16 GB的计算机内存，MACH和BEAGLE等方法占用的内存较少。但另一方面，IMPUTE使用的时间比MACH少。BEAGLE工作时占用的内存很小(2 GB)，而且填补速度较快，但是填补的准确度比不上MACH和IMPUTE。因此在实际应用中，研究者应该考虑填补准确度、计算机硬件条件和工作时间等多方面的因素以选择合适的软件。对染色体逐条进行填补或者一次只对部分样本分批进行填补可以降低填补对计算机性能需求。如果计算机性能偏低或者受到时间的限制，研究者可以考虑只对感兴趣的染色体或区域进行填补。

6. 未来的基因型填补方法：目前的基因型填补方法主要是针对SNP的填补，在GWAS中对SNP进行填补可以使研究的效率提高，但是对那些频率较小且对表型有重要影响的位点发现所起作用并不大。随着人类遗传学研究技术的不断发展，遗传数据填补方法也会相应地发生改变。首先参照面板会发生改变，它将包含比HapMap更丰富的人类基因组变异信息。其次基因型填补的算法也将相应改变，如用于填补拷贝数变异的方法。千人基因组计划(1000 Genomes Project，参考http://www.1000 genomes.org)对大约1200人进行全基因组测序，这个计划的完成将会带来一个包含更详细更有医学价值的人类基因组变异的图谱。这个面板中将包含更多罕见变异和结构变异。这些数据将替代HapMap的数据作为参照在研究中使用，而基于填补方法的研究也将以更高的精度来筛查出更多的与疾病和表型相关联的位点。

参 考 文 献

[1] Zeggini E, Scott LJ, Saxena R, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet, 2008, 40(3): 638-645.

[2] Kathiresan S, Melander O, Guiducci C, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat Genet, 2008, 40(2): 189-197.

[3] Willer CJ, Sanna S, Jackson AU, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet, 2008, 40(2): 161-169.

[4] Sanna S, Jackson AU, Nagaraja R, et al. Common variants in the GDF5 region are associated with variation in human height. Nat Genet, 2008, 40(2): 198-203.

[5] Loos RJ, Lindgren CM, Li S, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. Nat Genet, 2008, 40(6): 768-775.

[6] Barrett JC, Hansoul S, Nicolae DL, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet, 2008, 40(8): 955-962.

[7] Abecasis GR, Cookson WO. GOLD—Graphical Overview of Linkage Disequilibrium. Bioinformatics, 2000, 16(2): 182-183.

[8] Zhao H, Nettleton D, Dekkers JC. Evaluation of linkage disequilibrium measures between multiallelic markers as predictors of linkage disequilibrium between single nucleotide polymorphisms. Genet Res, 2007, 89(1): 1-6.

[9] Akey J, Jin L, Xiong M. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet, 2001, 9(4): 291-300.

[10] Burdick JT, Chen WM, Abecasis GR, et al. In silico method for inferring genotypes in pedigrees. Nat Genet, 2006, 38(9): 1002-1004.

[11] Li Y, Ding J, Abecasis GR. Mach 1.0: rapid haplotype reconstruction and missing genotype inference. Am J Hum Genet, 2006, 79: S2290.

[12] Pei YF, Li J, Zhang L, et al. Analyses and comparison of accuracy of different genotype imputation methods. PLoS One, 2008, 3(10): e3551.

[13] Purcell S, Neale B, Todd-Brown K, et al. PLINK: a toolset for whole genome association and population-based linkage analyses. Am J Hum Genet, 2007, 81(3): 559-575.

[14] Wellcome Trust Case Control Consortium. Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. Nature, 2007, 447(7145): 661-678.

[15] The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature, 2007, 449(7164): 851-861.

[16] Li Y, Willer C, Sanna S, et al. Genotype Imputation. Annu Rev Genomics Hum Genet, 2009, 10: 387-406.

[17] Chambers JC, Elliott P, Zabaneh D, et al. Common genetic variation near MC4R is associated with waist circumference and insulin resistance. Nat Genet, 2008, 40(6): 716-718.

[18] Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet, 2006, 78(4): 629-644.

[19] Marchini J, Howie B, Myers S, et al. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet, 2007, 39(7): 906-913.

[20] Servin B, Stephens M. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet, 2007, 3(7): e114.

[21] Nicolae DL. Testing untyped alleles (TUNA) -applications to genome-wide association studies. Genet Epidemiol, 2006, 30(8): 718-727.

[22] Zaitlen N, Kang HM, Eskin E, et al. Leveraging the HapMap correlation structure in association studies. Am J Hum Genet, 2007, 80(4): 683-691.

[23] Browning SR. Multilocus association mapping using variable-length Markov chains. Am J Hum Genet, 2006, 78(6): 903-913.

[24] Nothnagel M, Ellinghaus D, Schreiber S, et al. A comprehensive evaluation of SNP genotype imputation. Hum Genet, 2009, 125(2): 163-171.

(收稿日期：2009-12-14)
(本文编辑：万玉立)