· REVIEW · · 综 述 ·

DNA 混合分析技术的单体型频率估计方法

何柳, 唐迅, 胡永华

(北京大学医学部公共卫生学院流行病与卫生统计学系,流行病学教育部重点实验室,北京 100191)

[摘要] DNA 混合分析技术的广泛使用,能够明显降低研究的工作量和成本,然而仍然有其自身局限,例如损失了大量个体遗传信息、测量误差较大等。为了尽量克服 DNA 混合分析技术的自身问题,其检测和分析方法在不断发展与完善。在单体型频率估计方面,基于最大期望(expectation-maximization,EM)算法的新方法不断涌现,如 HaploPool 算法和 PoooL 算法,其准确性、实用性均增强。

[关键词] DNA 混合分析技术; 单体型; 遗传流行病学

DOI:10.3969/j.issn.1672-7347.2011.05.015

Estimation of haplotypes based on DNA pooling

HE Liu, TANG Xun, HU Yonghua

(*Department of Epidemiology & Biostatistics, Key Laboratory of Epidemiolog of Ministry of Education, Health Science Center, Peking University, Beijing* 100191, *China*)

Abstract: DNA pooling, a fast and economic study strategy, is widely used in areas of scientific research. In spite of various limits, researchers are making their efforts to improve DNA pooling toward a more perfect direction, including allele frequency detection and estimation of haplotypes. In haplotype estimation, more and more analyzing methods originated from the expectation-maximization algorithm have appeared, with improved accuracy and practicality, such as HaploPool algorithm and PoooL algorithm.

Key words: DNA pooling; haplotypes; genetic epidemiology

1985 年,Arnhein 等[1]首次介绍了在病例对照研究中使用混合 DNA 样本探索与 2 型糖尿病相关位点存在连锁不平衡的限制性片段长度多态性(restriction fragment length polymorphisms,RFLPs)的方法,通过此方法发现了与 2 型糖尿病相关的特异限制性多态性片段,并显示此方法经济且快速。此后,DNA 混合分析技术(DNA pooling)应运而生,该技术的主要原理是将 n 组个体的等量 DNA 样本分别混合形成 n 个 DNA 混合池(DNA pools),用检测方法对各 DNA 混合池进行某等位基因频率的估测并比较其在各池间的差异,即利用对混合 DNA 的基因频率的比较来替代对个体基因分型的比较。

随着 DNA 混合分析技术的不断成熟,基于个体遗传标志物测量的研究工作量和研究成本得以降低,以 500∶500 的病例对照研究为例,若按传统的方法对个体进行基因分型则需要 1 000 次基因分型实

收稿日期(Date of reception) 2010-04-22

作者简介(Biography) 何柳,博士研究生,主要从事遗传流行病学方面的研究。

通信作者(Corresponding author) 胡永华,E-mail:yhhu@bjmu.edu.cn

基金项目(Foundation items) 国家自然科学基金(30671807,30872173)。 This work was supported by the National Natural Sciences Foundation of China (30671807, 30872173).

验,但是如果将病例对照分别组成一个 DNA 池则只需要 2 次等位基因频率估测实验,极大地减少了检测工作量和成本。目前,DNA 混合分析技术已被广泛应用,至今已经覆盖了传染病或寄生虫病的诊断[2]、遗传性疾病和遗传倾向明显的疾病研究[3]、心脑血管疾病相关基因的分析[4]、肿瘤相关基因的研究数量性状位点(quantitative trait locus,QTL)[5]的关联研究等领域。然而,DNA 混合分析技术与基于个体基因分型研究相比仍然有其自身的局限,主要表现在:损失了大量的个体遗传信息,必须通过数学变换或各种信号转换来对不同 DNA 混合组的等位基因频率进行估计;等位基因频率估计过程中产生的测量误差较大,方差范围为 0.02 ~ 0.04[6],导致对大样本构成的 DNA 混合池进行等位基因频率估计的准确性偏低;无法直接进行中间表型、单体型和特殊遗传模型的构建分析;无法进行基因-基因、基因-环境交互作用研究;无法进行人群分层和调整[7]。因此,为了解决以上问题,各类 DNA 混合分析技术也在不断完善,其中基于 DNA 混合分析技术的单体型频率估计方法受到了越来越多的关注。

单体型是指一条染色体上紧密连锁的多个等位基因的线性排列,单体型分析被认为是一种高效低成本的分析方法。一方面,就单核苷酸多态性(single nucleotide polymorphisms,SNPs)而言,连锁不平衡的 SNPs 在物理距离上相互接近,由此在染色体某个区域构建出的单体型可以用于代表一组遗传距离接近的 SNPs;另一方面,单体型变异产生的多样性又是有限的,研究单体型与疾病性状的关联关系不会增加分析的难度[8]。因此研究者十分重视单体型的识别及分析其在疾病、性状关联研究中的重要性[9]。然而,DNA 混合分析技术作为另一种高效低成本的研究手段却几乎丢失了全部的单体型频率信息而无法直接进行单体型的构建分析。进入 21 世纪以来,有许多研究者意图取长补短将 DNA 混合分析技术和单体型分析进行结合,利用 DNA 混合分析技术所得的等位基因频率信息对单体型频率进行估算,以创造出更加高效低成本的研究设计和分析方法并尽可能地挖掘出更多的遗传信息。目前基于 DNA 混合分析技术的单体型频率估计方法主要有3 种。

1 最大期望算法

性状相似的个体具有相似的遗传背景,根据某性状可以将个体分为程度相近的若干小组,每组组成一个小样本 DNA 混合池。基于这种思想,研究者在多项研究中首先将原本用于处理缺失数据的最大期望(expectation-maximization,EM)算法引入到小样本 DNA 混合池的单体型频率估计中来[10-12]。

传统的 EM 算法用于单体型频率估计时主要基于从 DNA 混合池样本中得到的基因分型频率数据,是以通过反复迭代寻找最大对数似然估计为基本思想的算法。每次迭代计算主要分 2 个步骤:"E-step"和"M-step"。EM 算法可以通过矩阵计算同时对以 10^5 为上限的多个单体型频率进行估计。以一个单体型的频率估计为例,假设有 T 个 DNA 混合池,m_i 表示该单体型在第 i 个 DNA 混合池中的频数,M 为向量$(m_1,\cdots,m_T)$,即估算目标,其具体值无法直接观察到,需要通过基于 Hardy-Weinberg 平衡假设的完全对数似然函数(complete log-likelihood function)(公式 1)进行估计。

$$\iota(p \mid M) \propto \sum_{i=1}^{T} m_i \log P \qquad \text{(公式 1)}$$

在第 t 次迭代的"E-step",用完全对数似然函数计算给定基因分型频率数据和当前参数值 $P^{(t-1)}$(即第 $t-1$ 次迭代计算所得参数值)下的期望似然估计(expected log-likelihood);第 2 步是"M-step",即将第 t 次迭代过程中 E-step 所得的期望对数似然估计极大化,计算此时第 t 次迭代的参数值 $P^{(t)}$,用于第$t+1$ 次迭代的计算。通过反复迭代计算直至相邻两步参数值 $P^{(t)}$ 和 $P^{(t-1)}$ 的差收敛到小于事先约定的一个很小值为止。此时,最后一步迭代计算所得参数值对应的与基因分型频率数据相一致的向量 M,即 T 个 DNA 混合池中该单体型频率的估计值。

基于 EM 算法的软件常见的有 EH 软件(http://linkage.rockefeller.edu/software/eh)和 Haploview 软件[13](http://www.broadinstitute.org/haploview/haploview-downloads)。EM 算法运算量庞大,对于大样本 DNA 混合池而言,目前的计算机设备将面临很大的运行难度,而且会耗费相当大的运算时间。因此,运用 EM 算法估计单体型频率时,常把 DNA 混合池样本数量限制在 10 以下。为了克服 EM 算法的缺陷,在以后的方法学探索过程中又在 EM 算法的基础上衍生出了其他基于 DNA 混合分析技术的单体型频率的估计方法,如后文介绍的 HaploPool 算法和 PoooL 算法。

2 HaploPool 算法

Kirkpatrick 等在 EM 算法的基础上对单体型频率估计的方法进行改进,并于 2007 年公开发表了新

方法 HaploPool[14](HaploPool 软件获取地址:http://haplopool. icsi. berkeley. edu/haplopool/)。HaploPool 与以往 EM 算法的不同主要表现在:1)HaploPool 在 EM 算法的基础上加入了最优系统发生模型(perfect phylogeny model),增加了单体型频率估计的准确性;2)EM 算法主要用于染色体某区域由少量 SNPs 构建的单体型分析,对于大量 SNPs 还需引入线性回归方法;3)HaploPool 有更高的计算效率,它的运行速度比传统 EM 算法快 6 倍。例如,用传统 EM 算法对包含 2 个 DNA 样本的小样本 DNA 混合池进行染色体某个区域超过 10 个 SNPs 的扫描将耗费数小时,而同样的资料用 HaploPool 只需要几秒钟就可以完成;4)HaploPool 具有更好的处理缺失数据和基因分型错误的能力。因此,HaploPool 是 DNA 混合分析技术与单体型分析相结合的探索中的又一进步。采用 HaploPool 方法进行分析主要包括 3 个步骤:首先是通过最优系统发生模型来计算人群中潜在的单体型并用贪婪算法(greedy algorithm)来增强对单体型的推断[15];然后,将所研究染色体某个区域中的 SNPs 分成若干亚组,对各个亚组应用 EM 算法来推断单体型的频率;最后,将通过各个亚组 SNPs 推断出来的单体型频率信息用线性回归的方法进行整合,从而得到整个目标染色体区域的所有 SNPs 构成的单体型频率。HaploPool 分析方法推测单体型频率的过程包括 3 条路径:第 1 条路径只包括最优系统发生模型的拟合过程,第 2 条路径同时包括了最优系统发生模型拟合以及贪婪算法的推测过程,第 3 条路径只包括了贪婪算法的推测过程。当 SNPs 的数量增加时,第 2 条路径所给出的推测结果相比另外 2 条路径将更加准确。HaploPool 则基于 SNPs 的数量选择最合适的路径进行单体型频率推测。

总之,HaploPool 的优势集中在 3 个方面:一是它继续将 EM 算法估计小样本 DNA 混合池样本单体型频率准确率高的优势发扬光大,二是它具有综合使用最优系统发生模型和贪婪算法来鉴别常见单体型的能力,三就是它能够准确地由多个 SNPs 亚组构建的单体型频率来推断染色体区域中所有 SNPs 构建的单体型频率。Stephens 等[16]将使用 HaploPool 进行单体型频率估计的结果与公认最先进的 phasing 法(通过 Phase 软件完成,软件获取地址:http://depts. washington. edu/uwc4c/express-licenses/assets/phase/)估计的结果进行比较,发现以整个人群单体型频率作为参照时,若要使 phasing 法的估计准确度与 HaploPool 相同,则需要额外增加 45% 非混合样本的基因分型实验,从侧面印证了 HaploPool 的高效性。此外,Kirkpatrick 等[14]还证明了 HaploPool 使用的广泛性,即不仅可以使用于 DNA 混合分析设计,还可以用于基于个体基因分型的数据。

尽管 HaploPool 让 DNA 混合分析技术前景更为光明,但它还是被小样本 DNA 混合池的前提所局限。模拟误差分析提示:在使用 HaploPool 时,只有当 DNA 混合池包含的样本数量为 2 时才可以在增加准确度和减少误差两者间达到理想的平衡状态,当每个混合池中样本数量超过 2 时,误差就会增大。同时,Kirkpatrick 等[14]还认为当前支持大样本 DNA 混合池单体型频率估计的方法远远不能像 HaploPool 一样在估计效能上与传统的基于个体的基因分型结果相媲美;但是研究者不得不考虑科研成本。基于小样本 DNA 混合池的 HaploPool 虽然减少了基因分型的工作量和工作成本,但其所消耗的成本对于大多数科研小组仍是极大的负担,尤其是昂贵的全基因组关联研究(genome-wide association studies, GWAS)设计。甚至基于小样本 DNA 混合池的分析技术在准确定量个体 DNA 样本以构建若干 DNA 混合池的过程中还增加了一笔个体基因分型设计所没有的花费。

3　PoooL 算法

用大样本 DNA 混合池来进行单体型频率估计仍是研究者努力的方向。对于包含成百上千个体 DNA 的混合池,即便不能完全否定其用于估计单体型频率估计的可能性,但就目前的大部分计算机设备而言,这仍然是不可能完成的任务。2008 年,Zhang 等[17]继 Kirkpatrick 之后在 Bioniformatics 杂志上介绍了一种基于大样本 DNA 混合池估计单体型频率的新方法,取名 PoooL(PoooL 软件获取地址:http://staff. ustc. edu. cn/ ~ ynyang/poool)。其突出特点为该方法所承载的计算负荷并不取决于 DNA 混合池所包含样本的多少,这在一定程度上克服了大样本 DNA 混合池估计单体型频率的困难,使得用大样本 DNA 混合池估计单体型频率成为可能。

从方法来看,PoooL 仍然是在 EM 算法的基础上发展起来的,但在运算速度上明显优于传统 EM 算法,同时并没有降低单体型频率估计的准确度。PoooL 是一种约束 EM 算法(constrained EM algorithm),其特点就是在传统 EM 算法的基础上结合了基因分型观察值的均数以及方差-协方差矩阵的思想,并引入了对等位基因频率和成对连锁不平衡系数(coefficient of linkage disequilibrium, LD 系数)或

成对单体型频率的线性约束(linear constraints),从而简化传统 EM 算法中对条件期望值(conditional expectation)的计算。在线性约束下,EM 算法被转变成约束最大熵模型(constrained maximun entropy model),此时就可以用改进迭代算法(improved iterative scaling,IIS)来进行单体型频率的估计。PoooL 的另一个特点是使用了“重要因子(important factor)”的概念。重要因子是在 DNA 混合池中混合样本的等位基因频率估计值满足正态分布的假设下产生的,它同样受到线性约束的限制,在整个迭代过程中其值恒定不变,不需要进行重复估计。重要因子的这一特性有利于将 EM 算法“M-step”的最大化问题向约束最大熵模型转化,并可以通过 IIS 算法得到最优模型。

Zhang 等[18]使用已公开发表的 3 套数据来运用 PoooL 算法对单体型频率进行模拟估计,分别是:芬兰用于 1 型糖尿病遗传学研究的 FUSION 数据(the Finland-United States Investigation of NIDDM Genetics Study)[18-20],用于验证 PoooL 在大样本 DNA 混合池中使用的优越性;Yang 等[21]研究 *AGT* 基因 10 个多态性位点的数据,用于验证 PoooL 在含有较大数量 SNPs 数据中进行单体型频率估计的能力;Jain 等[22]的研究数据,将 PoooL 和传统 EM 算法进行比较,用于验证 PoooL 在运算效率上的优越性以及应用于小样本 DNA 混合池的能力。

通过模拟应用发现 PoooL 可以捕获大部分单体型频率信息;同时,随着 DNA 混合池数量(T)的增加和每个混合池中所含样本量(n)的增加,或者随着构建单体型的 SNPs 减少,所得预测值与真实值的差距均会减小。当用于构建单体型的 SNPs 数量为 3,各 DNA 混合池的样本量 $n \geqslant 50$ 时,误差变化的值将趋于稳定。此外,研究者还发现:PoooL 在应用于小样本 DNA 混合池($n=2$)时同样具有高准确性,其精确度与 EM 算法的精确度十分接近[17]。从计算效率来讲,PoooL 的运算效率明显优于传统 EM 算法,其运算速度不取决于混合池样本量的大小,只与混合池的数量成线性相关。

与 HaploPool 相比,PoooL 的优越性在于它可以在大样本 DNA 混合池中进行高效的单体型频率估计。综合科研经费、操作难度和仪器灵敏度等因素,实际科研过程中往往会选择几十或几百个样本组成一个 DNA 混合池,因而,只能支持小样本 DNA 混合池的 HaploPool 将不会成为首选方法,而 PoooL 更具可行性。但是,在使用 PoooL 过程中也要注意到它的局限,即只能用于构建单体型的 SNPs 数量较少的情况。当构建单体型的 SNPs 数量较多时,可以考虑结合使用滑窗迭代算法(sliding window method)[23]或者分割捆绑算法(partition-ligation method)[24]。

4 前景与展望

DNA 混合分析技术除了已经涉及到各个物种和各种疾病,在基因表达调控和基因变异层面也逐渐延伸到更新的领域,比如以染色体甲基化为代表的表观遗传学,目前基因变异研究的热点领域如拷贝数变异(copy number variations,CNVs)等。在研究设计方面,也不再局限于简单的 2 个 DNA 混合池的大样本病例-对照研究,而是倾向于选择在限制每个混合池的样本数量的前提下增加 DNA 混合池数量并对每个混合池进行多次重复测量取平均值报告等位基因频率的更为复杂的研究设计,以尽量捕捉准确的遗传信息。针对 DNA 混合分析技术难以对人群进行分层分析的缺陷,研究者一开始使用统计学方法调整研究个体的主要背景变量,后来则在研究设计阶段就注重选择与病例相匹配的对照构成对照组 DNA 混合池。近年来,随着家系设计作为一种遗传流行病学研究方法的不断发展与成熟,研究者也先后将 DNA 混合分析技术应用于以核心家系为基础的关联分析以避免人群分层的影响。此外,随着国际人类基因组单体型图(HapMap)计划初步完成和高通量基因分型技术的迅猛发展,GWAS 在全球范围内受到关注。然而,GWAS 对于个体基因组进行扫描所消耗的高昂科研成本让国内外众多的实验室望尘莫及,因此,DNA 混合分析技术凭借其优势同时结合目前推崇的两阶段研究设计[25]也成为 GWAS 中经常采用的方法。

DNA 混合分析技术作为一种个体基因分型的替代方法,在研究成本和科研工作量上均具有明显的优势,但目前仍有许多问题尚待解决,比如如何进一步提高测量值的准确度,如何进行基因-基因和基因-环境交互作用的分析,如何增加成本-效果比等。因此,仍需要更高效、更完善的检测手段和分析方法,以及更先进的分析平台。DNA 混合分析技术今后将有更广阔的发展空间。

参考文献:

[1] Arnheim N C. Strange H E. Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease:Studies of HLA class II loci[J]. Proceedings Nat Acad Sci, 1985, 82(20):6970-6974.

[2] Bharti A R, Letendre S L, Patra K P, et al. Malaria diagnosis by a polymerase chain reaction-based assay using a pooling strategy[J]. Am J Trop Med Hyg, 2009, 81(5):754-757.

[3] Bugeja M J, Booth D, Bennetts B, et al. An investigation of polymorphisms in the 17q11.2-12 CC chemokine gene cluster for association with multiple sclerosis in Australians[J]. BMC Med Genet, 2006, 7: 64.

[4] Matkovich S J, Van Booven D J, Hindes A, et al. Cardiac signaling genes exhibit unexpected sequence diversity in sporadic cardiomyopathy, revealing HSPB7 polymorphisms associated with disease[J]. J Clin Invest, 2010, 120(1): 280-289.

[5] Baro J A, Carleos C, Corral N, et al. Power analysis of QTL detection in half-sib families using selective DNA pooling[J]. Genet Sel Evol, 2001, 33(3):231-247.

[6] Sham P, Bader J S, Craig I, et al. DNA Pooling: a tool for large-scale association studies[J]. Nat Rev Genet, 2002, 3(11):862-871.

[7] Pearson J V, Huentelman M J, Halperin R F, et al. Identification of the genetic basis for complex disorders by use of pooling-based genomewide single-nucleotide-polymorphism association studies[J]. Am J Hum Genet, 2007, 80(1):126-139.

[8] De Bakker P I, Graham R R, Altshuler D, et al. Transferability of tag SNPs to capture common genetic variation in DNA repair genes across multiple populations[J]. Pac Symp Biocomput, 2006, 11:478-486.

[9] International HapMap Consortium. The International HapMap Project[J]. Nature, 2003, 426(6968):789-796.

[10] Hoh J, Matsuda F, Peng X, et al. SNP haplotype tagging from DNA pools of two individuals[J]. BMC Bioinformatics, 2003, 4: 14.

[11] Ito T, Chiku S, Inoue E, et al. Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data[J]. Am J Hum Genet, 2003, 72(2):384-398.

[12] Yang Y, Zhang J, Hoh J, et al. Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA[J]. Proc Natl Acad Sci U S A, 2003, 100(12):7225-7230.

[13] Barrett J C, Fry B, Maller J, et al. Haploview: analysis and visualization of LD and haplotype maps[J]. Bioinformatics, 2005, 21(2):263-265.

[14] Kirkpatrick B, Armendariz C S, Karp R M, et al. HaploPool: improving haplotype frequency estimation through DNA pools and phylogenetic modeling[J]. Bioinformatics, 2007, 23(22): 3048-3055.

[15] Philip E B. RECOMB'04: Proceedings of the Eighth annual international Conference on Resaerch in Computational Molecular Biology[M]. New York, USA: ACM Press, 2004:10-19.

[16] Stephens M, Smith N J, Donnelly P. A new statistical method for haplotype reconstruction from population data[J]. Am J Hum Genet, 2001, 68(4):978-989.

[17] Zhang H, Yang H C, Yang Y. PoooL: an efficient method for estimating haplotype frequencies from large DNA pools[J]. Bioinformatics, 2008, 24(17):1942-1948.

[18] Zhang H, Zhang H, Li Z, et al. Statistical methods for haplotype-based matched case-control association studies[J]. Genet Epidemiol, 2007, 31(4):316-326.

[19] Valle T, Tuomilehto J, Bergman R N, et al. Mapping genes for NIDDM. Design of the Finland-United States investigation of NIDDM Genetics (FUSION) study[J]. Diabetes Care, 1998, 21(6):949-958.

[20] Lin D Y, Zeng D. Likelihood-based inference on haplotype effects in genetic association studies[J]. J Am Stat Assoc, 2006, 101(473): 89-104.

[21] Yang Y, Hoh J, Xu F, et al. Efficiency of SNP haplotype estimation from pooled DNA[J]. Proceedings Nat Acad Sci, 2003, 100(12):7225-7230.

[22] Jain S, Tang X, Narayanan C S, et al. Angiotensinogen gene polymorphism at -217 affects basal promoter activity and is associated with hypertension in African-Americans[J]. J Biol Chem, 2002, 277(39): 36889-36896.

[23] Yang H C, Pan C C, Lin C Y, et al. PDA: Pooled DNA analyzer[J]. BMC Bioinformatics, 2006, 7:233.

[24] Niu T, Qin Z S, Xu X, et al. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms[J]. Am J Hum Genet, 2002, 70(1):157-169.

[25] Zuo Y, Zou G, Zhao H. Two-stage designs in case-control association analysis[J]. Genetics, 2006, 173(3):1747-1760.

(本文编辑　彭敏宁)