·综述·

DNA混合池技术在遗传及分子流行病学中的应用进展

胡洋　胡永华

【关键词】 遗传流行病学；分子流行病学；DNA混合池

Advance in DNA pooling application in genetic and molecular epidemiology

HU Yang[1], *HU Yong-hua*[2]. 1 *Basic Medical College*, 2 *School of Public Health*, *Peking University Health Science Center*, *Beijing* 100191, China

Corresponding author: *HU Yong-hua*, Email: yhhu@bjmu.edu.cn

This work was supported by a grant from the National Natural Science Foundation of China (No. 30671807, No. 30872173).

【Key words】 Genetic epidemiology; Molecular epidemiology; DNA pooling

许多慢性非传染性疾病的病因机制复杂，其发病与环境暴露因素有关，也与遗传易感性及交互作用关系密切。在应用遗传及分子流行病学对复杂疾病的病因研究中，常常需要应用多种现代分子生物学技术对样本进行检测。其中反转录-聚合酶链反应（RT-PCR）、微卫星（micro satellite，MS）、等位基因特异性杂交（allele-specific hybridization，ASH）等方法被广泛应用。基因芯片（gene chip）检测技术与全基因组关联性研究（genome-wide association study，GWAS）的方法已开始被大量运用到复杂性疾病的研究中。流行病学的研究通常涉及大规模的人群，需要对成百上千的个体样本进行几十、几百甚至更多遗传标记的测量，工作量巨大，资金消耗也十分可观，一般实验室是难以承担的。DNA混合池（DNA pooling）技术被认为是解决此问题的一种有效方法[1]。

DNA混合池技术是根据抽样原理设计的一种方法。以病例对照设计的关联研究为例，基本步骤是：①分别构建建立病例与对照样品的DNA混合池，即每个池包含N个个体的DNA（要求混合时每个样本等量、等浓度）；②分别对病例DNA混合池与对照DNA混合池进行高通量检测仪器（如基因芯片）中进行检测；③对检测结果进行统计分析，筛出可疑单核苷酸多态性（SNP）位点；④针对筛出的可疑SNP位点，对所有研究对象的个体DNA样本进行检测（或抽取部分个体样本进行检测），统计分析检测结果，比较病例与对照的差异，确定致病基因位点。

为提高应用DNA混合池技术所获结果的真实性与效率，近来出现了多种构建DNA混合池的设计方法，如多种类型病例与对照的混合池构建，多阶段设计，新的统计分析等。DNA混合池技术的特点是在不明显减小检测效力的前提下提高研究效率，节省研究经费，缩短研究周期。近年来，在DNA混合池技术的设计理论与实际应用方面有着较多新的进展，本文综述如下。

1. SNP位点检测效力：长期以来检测效力偏低一直是DNA混合池技术受限的原因。以往研究表明，只有当病例与对照组间等位基因频率差别在7.5%以上时才可检测出[2]。而近年发展的方法已使得该技术的检测效力明显提高。2006年Zuo等[3]证明，用两阶段DNA混合池技术检测大样本（病例及对照数量＞1000）时，只要病例和对照组间等位基因频率的差异不小于5%，且第二阶段选取适当比例的遗传标记进行分型，就有很高的概率（＞80.0%）检测出与疾病高度相关的遗传标记。实验结果还表明，在大样本条件下，即使当等位基因频率的差异为3.0%时，如果选取相对高的第二阶段遗传标记的分型比例，也很可能检测出显著差异大的遗传标记[3]。

2006年Bierut等[4]对导致尼古丁依赖的易感基因进行了全基因组扫描，以对尼古丁依赖的吸烟者做实验组（1050人），对尼古丁非依赖的吸烟者为对照组（879人）。为了能够快速有效且节约成本完成扫描，他们采用的是两阶段的DNA混合池设计。第一阶段选取约50%的研究对象构建了16个DNA混合池，实验组和对照组各8个，每个池包含约60份个体样本。使用了49个芯片对240万个SNP进行基因分型，经统计分析后，筛选出实验组和对照组中等位基因频率差异最显著的31 960个SNP。第二阶段则对所有的实验组和对照组样本进行个体分型，扫描由第一阶段得到的具有显著差异的31 960个SNP。研究中运用回归分析的方法分析个体分型的结果，还考虑了年龄、性别等协变量与基因型交互作用，最终发现35个SNP的P值$<10^{-4}$；其中包含一些新发现的基因，如Neurexin1、VPS13A、TRPC7、CTNNA3、CLCA1等[4]。2009年Knight等[5]沿用Bierut的样本和方法，通过个体分型和混合池2种方法对近1000个样本中50 000个SNP检测，进行病例对照研究；目的是比较混合池设计和个体分型的检测效力。他们将用2种方法检测得到的数据构建成一混合的二元正态分布（X_{pool}, X_{ind}），并且证明大部分能够被DNA混合池正确估计的数据，在发现疾病相关变量方面与个体分型所获得的数据有同等的效力。

Chew和Ramani[6]对用DNA混合池的方法和个体分型方法得到的等位基因频率的差异进行了比较。在500 K基因芯片平台上，实验组混合池和对照组混合池等位基因频率与已知通过个体分型估计的频率相关系数均＞0.97，有的混合池相关度甚至达到0.987。Ji等[7]将疾病遗传模型中的参数整合到混合池的关联性研究中，发现检测效力随基因分型数量的增加而显著提高。他们得到4个影响关联性研究检测效力的参数：①基因型相对风险；②遗传模型；③样本大小；④疾病与SNP遗传标记等位基因频率之间的相互作用关系。

DOI: 10.3760/cma.j.issn.0254-6450.2010.07.024

基金项目：国家自然科学基金（30671807, 30872173）

作者单位：100191 北京大学基础医学院（胡洋），公共卫生学院（胡永华）

通信作者：胡永华，Email: yhhu@bjmu.edu.cn

检测效力的提高使得混合池的方法在近年来被广泛应用，如2009年Bosse等[8]用DNA混合池的方法通过对2组2型糖尿病和慢性鼻窦炎患者进行病例队列研究，检测了55 000多个SNP，证实了混合池技术在高密度基因实验中可以准确的确定微小等位基因频率，并发现了一些在实验组和对照组间有显著差异的SNP位点，如糖尿病易感基因TCF7L2和HHEX的SNPs，与由GWAS得出的结果完全相符。

2. 高通量DNA测序方法：该方法的应用使DNA混合池技术成本效益更加明显。早在20世纪40年代就有了将许多样本混合到一起进行研究的方法，在大规模的流行病学研究中，这种方法切实有效，大大减少了工作量，而且有良好的成本效益。采用DNA混合池技术与高通量检测仪器联合应用则能更显著降低复杂疾病病因研究的成本，而且样本量越大其相对于单纯高通量个体分型的优势越大，成本效益更加明显。如2009年Craig等[9]通过其开展的一项年龄相关性黄斑变性的GWAS研究证实，当病例与对照各2000人时，采用构建混合池的方法，通过Illumina测序仪来对混合样本进行检测，费用仅需10万美元；若用个体分型的方法费用为180万美元，且需数月时间。经方法学对比研究，混合池方法不亚于个体分型方法，检测效力很高，而且能节省大量时间、工作量、实验次数和试剂[9]。再如，2009年Lu等[10]将DNA混合池技术与DNA探针结合衍生出一种叫做DNA探针混合池的方法，用于肿瘤细胞的外周血检测、产前诊断和移植前遗传诊断，可以提高测绘染色体异位断裂点的速度。通过对肿瘤和体外受精样本的研究发现，从配子的克隆选择到染色体断裂点的测绘所用时间可缩短至3～4 d。

3. 新的DNA混合池设计与分析方法：2009年Zhao和Wang[11]通过数理统计的方法分析DNA混合池技术设计中的误差特点、成本估计与优化设计；研究详细论述了应用两阶段DNA混合池技术开展病例对照设计关联研究时，优化成本效益的3个问题：①DNA混合池特异误差规律(pool-specific errors)的大小：在一般情况下，第一阶段DNA混合池测量成本明显低于第二阶段个体基因分型的成本，此时的两阶段DNA混合池设计具有良好的成本效益；而当混合池特异误差较大并且风险等位基因频率很低，或者混合池特异误差不很大而混合池规模较大，以上两种情况均无好的成本效益。②第二阶段基因分型与第一阶段基因分型成本的相应比值(R)的影响。以往研究报道，理想的两阶段个体基因分型设计的R值为15～20；他们通过数理分析表明，在两阶段DNA个体分型设计，当R值由1增至15时，成本将增加17.5%，而采用混合池设计，成本将仅增加1.4%；无论R值如何变化，无论风险等位基因频率的高低，两阶段DNA混合池分型设计的成本要比两阶段DNA个体分型设计的成本低10倍。③DNA混合池的规模：即第一阶段所有混合到一起的DNA样本量，该研究表明，为达到更高的成本效益，在进行GWAS的DNA混合池设计时，第一阶段样本量至少为研究对象的70.0%，而第二阶段进行个体分型时，测量遗传标记的比例应小于总量的0.5%。

与DNA个体分型相比，采用DNA混合池设计得到的数据存在信息丢失问题，如缺少关于在每个池中的每个遗传标记内偏离H-W平衡(Hardy-Weinberg equilibrium)比例的信息；缺少在每个池中遗传标记间连锁不平衡的信息等。为弥补这一缺陷，2007年Johnson[12]提出了一种基于经典模型McPeeK和Strahs的贝叶斯统计方法，用以分析通过混合池得到的数据。该方法通过试验被证明在统计学上和计算上都很有效，在大规模关联性研究中，对通过混合池得到的数据能够同时对数量性状位点(quantitative trait loci, QTL)进行分析。

DNA混合池设计原有的一大缺陷是不能提供单倍体型的信息，无法进行单倍体分析。为弥补此缺陷，2008年Zhang等[13]提出了一个叫做PoooL的计算方法来估算DNA混合池的SNP位点单倍体型频率。PoooL是一种约束性EM(expectation-maximum)算法，他们引进了一个被称为重要性因子(importance factor)的量，用以衡量一个单倍体型的可能性。基于等位基因频率呈渐近正态，并且单倍体型频率处于一个线性约束系统的假设，在迭代最大化过程中，重要性因子为一常数。他们通过模拟研究证实PoooL这种算法可以有效估计任意样品量的样品池中的单倍体型频率。该算法既能检测成百上千样品的大混合池，对少到一两个的情况也同样适用。PoooL计算的复杂性与混合池的大小无关，因此它对大混合池的计算效率比既有的估计方法有显著提高。模拟的结果还显示，这种方法还可以减小基因分型误差及人群混杂的影响。

4. DNA混合池技术误差分析：DNA混合池技术可以降低大范围遗传关联性研究的成本，但不可避免出现实验误差，这些误差可能会提高假阳性率及降低检测效力。对于DNA混合池的误差分析及误差对于其检测效力影响的相关探讨由来已久，最近的一些研究表明，与个体分型相比，影响混合池测量真实性的主要因素是对大量样本混合后产生的各种误差的大小。2007年Macgregor[14]检验了混合池内及混合池间的误差，发现大多数误差来自基因芯片上反应过程，而不是混合池的构建。当用基因芯片对384个样本的56 494 SNPs检验时发现微阵列特异性误差比混合池构建引起的误差高7倍。这提示在拥有构建良好的混合池的前提下，为了尽量减少误差，应增加每个样本对应的微阵列的数量而不是构建多个混合池。2009年Jawaid和Sham[15]的病例对照研究也证明了混合池技术的主要误差来源在于等位基因的扩增过程，即误差主要存在于反应过程中。他们应用来自DNA混合池实验和相应个体基因分型的数据，对于多种不同来源的实验误差用线性随机效应模型和回归2种统计方法进行定量研究；证明混合池具有准确估计等位基因频率的潜力。通过分析误差的组成发现，DNA混合池设计误差的来源主要是：①等位基因的差异扩增；②等位基因频率的测量；③混合池的构建误差。

5. 在不同基因检测领域的应用：根据不同的研究目的，研究者会选择相应的基因结构进行检测。已被发现的常见基因结构包括微卫星、SNP、数量性状位点、甲基化修饰、拷贝数量变异(copy number variation, CNV)等。DNA混合池的方法现已被广泛地应用到这些特点结构的检测当中，以降

低成本，提高效率。

CNV是人类基因组中一个重要结构，它们影响基因的表达水平并进而引起表型的不同。与SNP相比，CNV对于基因表达的影响更大。一些CNV被认为与某些疾病相关，如Prader-Willi综合征，DiGeorge综合征和孤独症样谱系障碍。2008年Lin等[16]用混合池和100 K基因芯片结合对2组分别为10人和30人的群体进行检测，发现12个常见的拷贝数量变异性区域，其中有10个CNV经实时定量PCR和个体基因芯片分型确认；另2个发现的CNV可以在基因组变异数据库中查到。这一实验结果提示DNA混合池可以有效地被用于检测常见的拷贝数量变异性。

DNA甲基化是最早发现的DNA修饰途径之一，大量研究表明，DNA甲基化能引起染色质结构、DNA构象、DNA稳定性及DNA与蛋白质相互作用方式的改变，从而控制基因表达。2009年Docherty等[17]用混合池的方法对89个DNA样本进行甲基化检测，与个体检测的样本比较，结果有高度一致性，证明混合池技术可以在大规模人群研究中提供准确定量的DNA甲基化平均水平，同时节省大量成本。

DNA混合分析技术还可用于QTL的关联研究。2007年Korol等[18]在对QTL研究中，提出一种DNA混合池分池的方法（fractioned DNA pooling，FDP），即将位于人群性状分布两个极端的部分随机分为数个独立的混合池，以此来提高选择性基因分型的可靠性。FDP是在概念上和结构上对选择性DNA混合池（selective DNA pooling，SDP）的一种改良，它可以对QTL检测进行置换检验而不再依赖统计学检验中假定的渐近分布。在家系研究和交叉测绘（cross mapping）设计中，FDP提供了一系列原先只能通过个体基因分型实现描绘QTL的方法；包括在被标记的等位基因在家系中仅部分共享的情况下，将多个家系和染色体上的多个遗传标记联合起来分析并检测家系中QTL的杂合子、QTL位置的置信区间估计和多关联QTL分析。

2009年Chi等[19]提出了一个将DNA混合池与选择性重组体基因分型（selective recombinant genotyping，SRG）技术结合的方法。他们假设任何分开的标记等位基因可在小规模混合池中被检测出（例如20人的混合池），并且等位基因的相对频率可以很好地在大规模样本中被估计。实验的第一阶段用两阶段DNA混合池的方法筛选存在于一对侧翼标记间的重组体；第二阶段以大量DNA分析法（bulked DNA analysis，BA）和两阶段混合池对重组体进行基因分型。实验结果提示DNA混合池和SRG的联合应用可以有效降低传统重组体基因分型的费用。

6. 展望：DNA混合池技术在一定程度上克服了传统方法精度差和精密仪器分析耗费高的问题，能够以相对低的成本进行有效检测。它同时开拓了一种研究方法的新思路。研究者不仅可以将待测样本的DNA进行混合，还可以进一步混合RNA、cDNA、血液甚至蛋白质等。这种将样本混合以化零为整降低成本的思想方法将可能会有进一步发展，其他类型的混合池技术可能会陆续出现。另外，在遗传及分子流行病学中，DNA混合池技术与其他的仪器分析手段相结合对大规模人群研究可能成为一个新的趋势。

虽然DNA混合池技术在对大规模人群的基因检测中有很高的检测效力和成本效益，但与其他任何一种检测技术一样，有其本身固有的缺点。首先，因为混合池技术得到的数据须经数理统计分析，所以处理数据时对于误差的分析是相当复杂的，而误差本身对于实验的检测效力来说至关重要。其次，由于DNA混合池技术常常用于大样本的研究，DNA样本的质量控制难度很大。另外还有文献报道，混合池技术在对弱连锁不平衡的检测方面效力不高[20]。

尽管许多方面还有待进一步完善，但DNA混合池技术目前仍不失为一种对于大规模人群基因研究的成本效益很高的检测手段，可以预期它将在分子及遗传流行病学领域中有更广泛的应用。

参 考 文 献

[1] Sham P, Bader JS, Craig I, et al. DNA pooling: a tool for large-scale association studies. Nat Rev Genet, 2002, 3(11): 862-871.
[2] Lv J, Li LM. DNA pooling technology. China Public Health, 2002, 18(12): 1517-1519. (in Chinese)
吕筠，李立明. DNA混合分析技术. 中国公共卫生，2002，18(12)：1517-1519.
[3] Zuo YJ, Zou G, Zhao H, et al. Two-stage designs in case-control association analysis. Genetics, 2006, 173(3): 1747-1760.
[4] Bierut LJ, Madden PA, Breslau N, et al. Novel genes identified in a high-density genome wide association study for nicotine dependence. Hum Mol Genet, 2007, 16(1): 24-35.
[5] Knight J, Saccone SF, Zhang Z, et al. A comparison of association statistics between pooled and individual genotypes. Hum Hered, 2009, 67(4): 219-225.
[6] Chew FT, Ramani A. Validation of pooled genotyping on the Affymetrix 500 k and SNP6.0 genotyping platforms using the polynomial-based probe-specific correction. BMC Genet, 2009, 10: 82.
[7] Ji F, Finch SJ, Haynes C, et al. Incorporation of genetic model parameters for cost-effective designs of genetic association studies using DNA pooling. BMC Genomics, 2007, 8: 238.
[8] Bosse Y, Bacot F, Montpetit A, et al. Identification of susceptibility genes for complex diseases using pooling-based genome-wide association scans. Hum Genet, 2009, 125(3): 305-318.
[9] Craig JE, Hewitt AW, McMellon AE, et al. Rapid inexpensive genome-wide association using pooled whole blood. Genome Res, 2009, 19(11): 2075-2080.
[10] Lu CM, Kwan J, Baumgartner A, et al. DNA probe pooling for rapid delineation of chromosomal breakpoints. J Histochem Cytochem, 2009, 57(6): 587-597.
[11] Zhao YH, Wang S. Optimal DNA pooling-based two-stage designs in case-control association studies. Hum Hered, 2009, 67(1): 46-56.
[12] Johnson T. Bayesian method for gene detection and mapping, using a case and control design and DNA pooling. Biostatistics, 2007, 8(3): 546-565.
[13] Zhang H, Yang HC, Yang Y, et al. PoooL: an efficient method for estimating haplotype frequencies from large DNA pools. Bioinformatics, 2008, 24(17): 1942-1948.
[14] Macgregor S. Most pooling variation in array-based DNA pooling is attributable to array error rather than pool construction error. European J Hum Genet, 2007, 15(4): 501-504.
[15] Jawaid A, Sham P. Impact and quantification of the sources of error in DNA pooling designs. Ann Hum Genet, 2009, 73: 118-124.
[16] Lin CH, Huang MC, Li LH, et al. Genome-wide copy number analysis using copy number inferring tool (CNIT) and DNA pooling. Hum Mutat, 2008, 29(8): 1055-1062.
[17] Docherty SJ, Davis OS, Haworth CM, et al. Bisulfite-based epityping on pooled genomic DNA provides an accurate estimate of average group DNA methylation. Epigenetics Chromatin, 2009, 2: 3.
[18] Korol A, Frenkel Z, Cohen L, et al. Fractioned DNA pooling: a new cost- effective strategy for fine mapping of quantitative trait loci. Genetics, 2007, 176(4): 2611-2623.
[19] Chi XF, Lou XY, Shu QY, et al. Combining DNA pooling with selective recombinant genotyping for increased efficiency in fine mapping. Theor Appl Genet, 2010, 120(4): 775-783.
[20] Xu J, Yang Y, Ying Z, et al. Testing linkage disequilibrium from pooled DNA: a contingency table perspective. Stat Med, 2008, 27(28): 5801-5815.

（收稿日期：2010-01-25）
（本文编辑：尹廉）