PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2017, 44(11): 1041~1043
www.pibb.ac.cn

LINCS——面向转化医学的细胞反应大数据计划 *

黄 昕** 何 松** 刘 阳 白 卉 伯晓晨***
(军事科学院军事医学研究院辐射医学研究所，北京 100850)

DOI: 10.16476/j.pibb.2016.0372

近十几年来，基因芯片、转录组测序、蛋白质组等技术的发展是极大地推动了生物医学研究的重要手段．通过基因和蛋白质表达谱分析可以发现疾病发生与发展的关键分子、辅助辨识诊断标志物和药物治疗靶标．然而，传统的组学分析主要关注显著差异表达基因，由于噪声的影响，差异表达分析往往带来较多的假阳性．

近年来，基因表达谱关联图谱分析逐渐成为基因表达谱分析的另一重要途径．基因表达谱关联图谱分析将细胞在不同扰动下的全基因组表达谱进行整体相似性比较，从而发现不同细胞扰动之间的关联．2006 年， Lamb 等[1]利用基因芯片技术测定了 1 309 种药物作用于 5 种人类肿瘤细胞系后的全基因组表达谱，并构建了关联图谱 CMap(connectivity map)．基于 CMap 和基因集合富集分析方法(gene set enrichment analysis，GSEA)[2]，通过与不同细胞状态下的基因表达谱印记对比分析，成功发现了新的药物作用模式[3-4]，在药物重定位[5]、疾病相关基因发现(发现上皮细胞分化抑制剂可以预防癌症转移)[6]、疾病机理研究[7]等方面得到了成功应用．

考虑到基于传统的转录组和蛋白质组分析技术构建关联图谱成本较高，很难扩大研究规模，低成本转录组检测技术 L1000(http://www．lincscloud．org/l1000/)和低成本蛋白质修饰组检测技术 P100 在最近几年得到了充分发展．在 L1000 和 P100 的技术基础上，美国国立卫生研究院(NIH)于 2010 年启动了基于网络的细胞反应印记整合图书馆(Library of Integrated Network-based Cellular Signatures，LINCS)计划，旨在全面描述不同小分子化合物、配体以及基因沉默扰动下的多层次细胞反应(如转录物表达水平、蛋白质表达水平、细胞表型等)．LINCS 计划将形成丰富的、高价值的细胞反应大数据资源，将基础医学和临床研究有机联系起来，为转化医学研究提供重要基础．

1 LINCS 计划的基础——L1000 技术和 P100技术

全基因组表达谱的数据需要测量人类 20 000 多个基因的转录表达水平，超高通量的检测造成较高的成本．事实上，人类基因表达之间存在高度的关联，不同基因的表达水平之间存在相互推算的可能．LINCS 计划所用的 L1000 技术就利用了基因表达之间的相关性，基于大规模的统计分析辨识出 978 个基因作为全基因组的标志基因(landmark genes)，通过测量标志基因的表达量，可以推算出其余 20 000 多个基因的表达量．实验结果表明，L1000 技术所采用的近 1 000 个基因的变化能够表征人体将近 80%的基因变化信息[8]．L1000 采用 Luminex 1000-plex Assay 液相芯片技术，一次可测量 1 000 个基因表达量，使成本降低到全基因组芯片成本的 1%，使得大规模测量细胞扰动表达谱成为可能．Luminex 1000-plex Assay 技术首先对含有荧光微珠标记的 384 孔板进行扫描，每孔有约 500 种不同颜色的荧光微珠，每种颜色有 100 个微珠，每种颜色按固定数量大小之比的微珠标记 2 种 mRNA，因而可以依据荧光颜色和荧光强度同时检测 2 种基因，最终获得 978 个基因表达强度的数据．

对于蛋白质修饰组，由于一小部分蛋白激酶和

* 国家自然科学基金重点项目(U1435222)资助．
** 并列第一作者．
*** 通讯联系人．
Tel: 010-66931207, E-mail: boxiaoc@163.com
收稿日期：2017-10-10，接受日期：2017-10-26

磷酸酶可以修饰各种成千上万种蛋白质上的氨基酸位点，LINCS 计划所采用的 P100 实验便基于此特性，结合质谱的靶标蛋白质组学实验找寻了 100 个具有代表性的磷酸肽探针，可以推算超过 1 000 种磷酸化位点[9]，从而使高通量、标准化、低成本的蛋白质修饰组学分析成为可能．P100 谱可用于生成分子印记，通过细胞扰动前后分子印记的比较建立细胞中的分子作用网络．

L1000 和 P100 技术的日趋成熟为 LINCS 计划的实施奠定了重要的技术基础．

2 LINCS 计划的数据产生与质量控制

LINCS 计划第一阶段(测试实验阶段)从 2010 年进行到 2013 年，重点是开发技术方法．测试实验阶段数据的产生及分析由哈佛医学院和布罗德研究所共同承担，并通过 Broad-HMS 合作项目来确保产生数据的一致性，该阶段主要研究高通量实验 L1000 并且检测不同细胞系在不同扰动下(小分子、基因沉默、蛋白质试剂、抗体试剂)基因表达谱的变化．除了不断优化技术流程、改善测试技术之外， LINCS 还开发了一系列计算技术来集成、分析和利用所生成的数据．

LINCS 计划第二阶段从 2014 年进行至今，由 HMS LINCS、 CMT MGH、 MEP LINCS、 NeuroLINCS、 BroadT LINCS、 DtoxS、 PCCSE、 Yale LINCS、 Broad Therapeutics、 CU LINCS、 ASU LINCS 和 DCIC 这 12 个数据生成中心分别获取特定细胞系在不同小分子化合物、配体和基因沉默等多个因素扰动下的细胞反应，共涉及了 1 190 个细胞系、25 581 个小分子，生成 350 个数据集．LINCS 数据库涉及的领域有激酶和小分子结合、荧光成像、转录组学、蛋白质组学、表观基因组学和代谢组学．除了 L1000 和 P100 技术外，LINCS 计划还采用了荧光成像分析、SWATH-MS、ATAC-seq、KINOMEscan 等技术对扰动后的细胞状态进行了多维度的检测．基于 L1000 的转录组数据是 LINCS 计划数据的主体，涉及 6 个数据集，包含细胞系 53 种，小分子数目 25 581 种．L1000 数据产生流程为扫描、反卷积、标准化、外推至所有基因和印记计算．因此数据库结构有四层，第一层为扫描后未经处理的流式细胞仪原始数据，第二层为从原始值反卷积后得到的近 1 000 个基因的表达量，第三层为经标准化后的实验测量基因的表达谱和由其外推估算的所有基因的表达谱，第四层为基因差异表达量的印记数据．表 S1 归纳了 LINCS 计划的检测实验类型及承担的研究中心名称．

在第二阶段，LINCS 计划的数据协调与整合中心 (Data Coordination and Integration Center，DCIC)对数据的注释信息、获取方式、分析工具等方面进行了进一步的规范和发展．LINCS 计划重视数据质量和可重复性，并制定了扰动试剂、细胞系和其余实验因素的标准，确保各个数据产生中心之间可以资源共享，最终每个数据集的质控(QA/QC)过程与数据本身一起发布．

在 L1000 检测的数据质量控制方面，LINCS 应用了 Luminex 不变集标准化和分位数标准化来去除实验所用仪器之间的系统误差．对原始值做对数处理后，首先 LINCS 挑选出 80 个在各种条件和环境下表达量基本不变的基因作为参照，与昂飞(Affymetrix HG-U133A)平台所相应的基因建立函数关系，然后依据该函数关系分析其余基因的荧光强度在昂飞平台的表达量．最终，对用同一实验仪器的不同样本测出的表达谱实行分位数标准化，使其分布相同，再对不同实验仪器测出的表达谱进行分位数标准化，从而可统一分析所有样本．

在 P100 检测的数据质量控制方面，LINCS 首先将所有比值数据进行对数化处理，并用较低的探针检测率来筛选样本，同时用较低的样本检测率来筛选探针，然后使用均值偏离度(distance-from-mean)来检测并去除不需要的样本．最终将不同实验的数据调整到同一水平，并将各组实验数据的对数比(log ratio)中位数或平均数调整为零．

3 LINCS 计划的数据获取及分析工具

为了方便研究者进行数据搜索，LINCS 计划参研单位和第三方机构还提供了丰富的数据检索和分析工具(表 S2)．

LINCS 计划的综合网站为 DCIC (http://lincs-dcic.org/#/)，其提供的数据入口为 LINCS Data Portal．在 iLINCS 网站中，用户可以基于机构名称和实验类型获得相关数据；在 LDR 工具中，用户可以按数据提交机构(如“Broad”)和实验条件(如“KINOMEscan”)来搜索已注册的数据集；在 HMS LINCS Database 中，可以通过筛选小分子、细胞系、蛋白质、抗体、试剂、数据集、数据库等条件来获取 LINCS 数据集和实验试剂的信息．如果用户想获得与某些基因集合相关的数据，可在 LINCS Canvas Browser 中输入上、下调基因列表，从而获得与之相似或者相反的前 50 个 L1000 实验

数据，或在 LIFE(http://life.ccs.miami.edu/life/)工具中输入检索项后按实验成分和扰动进行筛选．用户可使用 Harmonizome 工具获得特定过程相关的基因信息，或在 Enrichr 中依据基因名称进行搜索．若用户想获取基因表达印记，则可在 L1000CDS2 工具中上传自己的表达印记，从而获得 LINCS L1000 数据集中预测的与上传印记相似或者相反的印记列表，也可以在 CREEDS 工具中通过输入相关术语，如“Imatinib”、“TP53”、“Breast cancer”或者输入上、下调基因集来获得相反或相似的印记．

4 LINCS 数据的初步应用

LINCS 计划虽然实施不久，但其所产生的细胞反应大数据已经显示出在基因调控关系挖掘、疾病发生和药物作用机理辨识、药物新用途发现等方面的广泛用途．

Young 等[10]用 LINCS 中 L1000 里基因沉默扰动数据，采用了线性回归模型并结合了先验概率和后验概率，来推断基因间的调控关系，并且从转录因子数据库 TRANSFAC 和 JASPAR 里识别的关系得到验证．Wang 等[11]所在研究小组应用 L1000 数据集，通过组合化学结构和基因表达特征提出一种机器学习分类器来探索药物不良反应，并用该分类器分析了超过 20 000 个小分子，最终开发了浏览、搜索、预测小分子药物不良反应的工具．Xie 等[12]利用 L1000 药物扰动下转录组水平的数据，基于适用于多分类问题的机器学习算法 Softmax，系统发掘并预测了 480 种已上市药物重定位于其他治疗属性的潜力．Chen 等[13]用 L1000 数据集系统地研究并量化了药物与疾病扰动下基因表达特征之间的逆转关系，最终预测出 4 种化合物可有效逆转肝癌细胞中的基因表达，并在 5 种肝癌细胞系中加以验证．

5 结语与展望

LINCS 计划通过在一系列扰动下的基因表达及其他细胞过程层面的变化，建立了一个基于整合网络的细胞反应数据库，从而阐明在不同的基因和环境压力下细胞是如何做出反应的．目前已发布的 LINCS 数据十分丰富，描述了多种扰动状态下多层次的细胞反应，在基因调控关系推断、新药不良反应发现、药物作用机理挖掘等方面具有广泛的应用前景．

对于 LINCS 数据的应用刚刚起步，对其数据集本身的挖掘及其与其他大数据集之间的关联分析还急需要发展相关的方法和工具，比如通过细胞反应对 LINCS 小分子进行大规模聚类分析，发现小分子的潜在用途等．相信随着相关生物信息学研究的快速发展，LINCS 计划所产生的大数据在转化医学研究中的地位将越来越突出．

附件 表 S1，S2 见本文网络版附录(http:// www.pibb.ac.cn)

参 考 文 献

[1] Lamb J, Crawford E D, Peck D, *et al*. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science, 2006, **313**(5795): 1929−1935

[2] Subramanian A, Tamayo P, Mootha V K, *et al*. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA, 2005, **102**(43): 15545−15550

[3] Iorio F, Tagliaferri R, Di Bernardo D. Identifying network of drug mode of action by gene expression profiling. Journal of Computational Biology: A Journal of Computational Molecular Cell Biology, 2009, **16**(2): 241−251

[4] Iorio F, Bosotti R, Scacheri E, *et al*. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci USA, 2010, **107**(33): 14621−14626

[5] Chang M, Smith S, Thorpe A, *et al*. Evaluation of phenoxybenzamine in the CFA model of pain following gene expression studies and connectivity mapping. Molecular Pain, 2010, **6**(1): 56−69

[6] Reka A K, Kuick R, Kurapati H, *et al*. Identifying inhibitors of epithelial-mesenchymal transition by connectivity map-based systems approach. Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer, 2011, **6**(11): 1784−1792

[7] Wei G, Twomey D, Lamb J, *et al*. Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. Cancer Cell, 2006, **10**(4): 331−342

[8] Duan Q, Flynn C, Niepel M, *et al*. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. Nucleic Acids Research, 2014, **42**(Web Server issue): W449−460

[9] Abelin J G, Patel J, Lu X, *et al*. Reduced-representation phosphosignatures measured by quantitative targeted MS capture cellular states and enable large-scale comparison of drug-induced phenotypes. Molecular & Cellular Proteomics: MCP, 2016, **15**(5): 1622−1641

[10] Young W C, Raftery A E, Yeung K Y. A Posterior probability approach for gene regulatory network inference in genetic perturbation data. Mathematical Biosciences & Engineering, 2017, **131**(6): 1241−1251

[11] Wang Z, Clark N R, Ma'ayan A. Drug-induced adverse events prediction with the LINCS L1000 data. Bioinformatics, 2016, **32**(15): 2338−2345

[12] Xie L, He S, Wen Y, *et al*. Discovery of novel therapeutic properties of drugs from transcriptional responses based on multi-label classification. Scientific Reports, 2017, **7**(1): 7136−7147

[13] Chen B, Ma L. Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. Nature Communications, 2017, **8**: 16022−16034

附　录

Table S1　LINCS data generation

表 S1　LINCS 计划的数据生成

检测实验	实验描述	承担机构
基于荧光成像检测的实验	基于荧光成像进行细胞计数、细胞活性测定、细胞生长抑制实验、细胞凋亡实验、蛋白质磷酸化水平的测定、反相蛋白质阵列实验、细胞周期状态测定、药物协同作用测定、细胞表型测定、多细胞分析实验.	HMS LINCS、CMT MGH、CU LINCS、Broad Therapeutics
激酶小分子结合检测	采用 KINOMEscan 平台，测定待测化合物与激酶的相互作用	HMS LINCS
基于微环境微阵列的细胞生长测定	采用基于微环境微阵列的平台测定不同微环境的扰动对生物反应节点的影响	MEP LINCS
酶联免疫吸附剂测定	采用酶联免疫吸附剂测定实验进行蛋白质分泌的分析和蛋白质状态的分析	Yale LINCS
基于反相蛋白质阵列的实验	基于反相蛋白阵列技术进行蛋白质状态测定和细胞凋亡测定	HMS LINCS
基于磁珠的蛋白质状态的免疫分析	通过抗体包裹的磁珠捕获待分析物，实现多种蛋白的定量分析，测量原代人纤维细胞滑膜在刺激和小分子抑制剂下的反应	HMS LINCS
ATAC-seq 表观遗传分析	基于 ATAC-seq 方法通过高通量测序识别基因组上的“可接近”区域	NeuroLINCS
全染色质表观遗传分析	通过质谱检测小分子扰动多种细胞的表观调控变化，分析组蛋白翻译后修饰水平	PCCSE
激酶与小分子结合检测	采用 KiNativ 方法，检测激酶与小分子的结合程度	HMS LINCS
RNA-seq 基因表达谱测定	通过给诱导多能干细胞施加有副作用的单个药物或者能减缓副作用的组合药物，采用 RNA-seq 技术测量给药前后的基因表达量	DTOXS、NeuroLINCS
P100 蛋白质修饰测定	通过 100 个具有代表性的磷酸肽探针，检测、推断超过 1000 种的磷酸化位点	PCCSE
基于质谱法测定蛋白质状态	通过测量离子的质荷比，从而在混合物中鉴定和量化特定分子	DtoxS
SWATH-MS 蛋白质定量测定	基于 SWATH-MS 采集技术在单次分析中定量测定样品中的蛋白质	NeuroLINCS
L1000 mRNA 表达谱检测	通过 978 个基因印记检测、推算全基因组 mRNA 表达谱	Transcriptomics
代谢分析实验	通过高通量的活细胞微阵列筛选技术动态测定细胞代谢过程的表型	ASU LINCS

Table S2 LINCS data acquisition and analysis tools in LINCS

表 S2 LINCS 提供的数据获取和分析工具

工具	网址	功能
LINCS Data Portal	http://lincsportal.ccs.miami.edu/dcic-portal/	基于小分子、实验、基因、和细胞类型搜索、下载原始实验数据
HMS LINCS Database	http://lincs.hms.harvard.edu/db/	基于扰动类型、细胞系、抗体等搜索、下载 HMS LINCS 数据集
LIFE	http://life.ccs.miami.edu/life/	基于化合物、细胞系、基因名称等搜索所有 LINCS 数据，具有饼状图的数据导航
Harmonizome	http://amp.pharm.mssm.edu/Harmonizome/	包括 L1000 检测数据在内多个大数据计划的综合检索
LDR	http://amp.pharm.mssm.edu/ldr/#/	基于机构或实验条件的数据集搜索
LINCScloud	http://www.lincscloud.org/	检索、下载和分析第一阶段 L1000 检测数据
LINCS Canvas Browser	http://www.maayanlab.net/LINCS/LCB/#.V5GIzqLfqx8	基于小分子化合物、时间点或细胞系检索、展示 L1000 检测数据集
Slicr	http://amp.pharm.mssm.edu/slicr/#/search	检索、下载存储在 GEO 数据库中的 L1000 检测数据集
SEP-L1000	http://maayanlab.net/SEP-L1000/	基于药物反应数据分析药物副作用
L1000CDS2	http://amp.pharm.mssm.edu/L1000CDS2/#/index	根据提交的上调和下调基因列表识别出相似或相反细胞反应的扰动
CREEDS	http://amp.pharm.mssm.edu/CREEDS/	基于上、下调基因列表搜索、下载细胞反应印记，并进行可视化
iLINCS	http://www.eh3.uc.edu/ilincs/#/	基于分子扰动和细胞系等条件搜索 LINCS 数据集，并提供多种统计分析工具
piLINCS	http://eh3.uc.edu/pilincs/#/	蛋白质磷酸化位点印记的浏览、检索和可视化
Geo2Enrichr	http://amp.pharm.mssm.edu/g2e/	基于特征向量方法分析基因表达
PAEA	http://amp.pharm.mssm.edu/PAEA/	基于特征向量的改进方法进行基因集富集分析
Network2Canvas	http://www.maayanlab.net/N2C/#.V5GLLqLfqx8	基因网络的交互式可视化
P100/GCP Mosaic	http://amp.pharm.mssm.edu/p100mosaic	磷酸化位点和组蛋白修饰的聚类分析
Drug/Cell-line Browser	http://www.maayanlab.net/LINCS/DCB/	细胞系和小分子药物聚类的可视化分析
Docent	http://amp.pharm.mssm.edu/milestones/grid.html	浏览、比较特定细胞系和实验的数据
Gen3va	http://amp.pharm.mssm.edu/gen3va/	整合了 LINCS、TCGA、BioGPS 和 GEO 的印记，提供针对 LINCS 药物反应数据的多种分析工具