Journal of Clinical Personalized Medicine 临床个性化医学, 2013, 1, 1-6
http://dx.doi.org/10.12677/jcpm.2013.11001 Published Online December 2013 (http://www.hanspub.org/journal/jcpm.html)

Predicting Molecule Pathways between LOXL1 and TGF-β

Xiaonan Gong[1,2], Libing Shi[1,2], Xiaohui Zou[1,2*]

[1]Department of Bioinformatics, Zhejiang University, Hangzhou
[2]School of Medicine, Zhejiang University, Hangzhou
Email: gxn_1234@163.com, Shilibing1215@gmail.com, *zouxiaohui@gmail.com

Received: Nov. 18th, 2013; revised: Nov. 30th, 2013; accepted: Dec. 3rd, 2013

Abstract: POP refers to a pelvic floor structure support disorder that visceras in the female pelvic cavity move down along its normal position. However, there is no radical cure for it. So it is a significant research to inquire the pathogenesis of POP. From the information already in hand, we know that LOXL1 knock-down mouse will always produce POP after delivery. In the meantime, the expression of TGF-β has a strong relationship with the severity of POP. On the basis of the data already known, we search for the massive data on the Internet and chose two data sets to do the analysis. As for data mining, we do correlation analysis and cluster on the microarrays that we had chosen. Finally, to show it more clearly, we set up a biology network. We found that TGF-β can regulate the expression of LOXL1 through the smad and non-smad pathways, meanwhile, LOXL1 is involved in the formation of focal adhesion and the crosslinking between elastin and collagen.

Keywords: LOXL1; TGF-β; Data Mining; Network

LOXL1 与 TGF-β 相关分子作用路径的预测

龚晓楠[1,2]，施丽冰[1,2]，邹晓晖[1,2*]

[1]浙江大学生命科学学院，杭州
[2]浙江大学医学院，杭州
Email: gxn_1234@163.com, Shilibing1215@gmail.com, *zouxiaohui@gmail.com

收稿日期：2013 年 11 月 18 日；修回日期：2013 年 11 月 30 日；录用日期：2013 年 12 月 3 日

摘　要：盆腔器官脱垂(POP)是一种盆底支持结构功能障碍性疾病，而现有的手术手段并不能根治这种疾病，因此探究 POP 的发病机制是一项比较有意义的研究。已有的资料表明，类氨酰氧化酶——LOXL1 的敲除能够导致小鼠在分娩之后发生 POP。同时 TGF-β 的表达量高低与 POP 的严重程度相关。在已经掌握资料的基础上，我们对网络上现有的海量微阵列数据进行了查找，选取了其中两个数据集进行分析。我们通过数据挖掘的方式对于 POP 相关的生物芯片进行相关性分析以及聚类，最后建立生物学网络。我们发现 TGF-β 同时通过 smad 与非 smad 通路调控 LOXL1 的表达，而 LOXL1 在细胞外基质参与黏着斑的形成、弹性纤维与胶原的交联等等。

关键词：LOXL1；TGF-β；数据挖掘；网络

1. 引言

盆腔器官脱垂(pelvic organ prolapsed, POP)指女性盆腔内脏器(如膀胱、子宫、阴道残端等)沿其正常位置下移的一种盆底支持结构功能障碍性疾病。有研究表明弹性纤维是维持盆地结构一类重要蛋白，而与

*通讯作者。

维持弹性纤维动态平衡的一类关键酶——类赖氨酰氧化酶 1(lysyl oxidase-like-1, LOXL1)是一种铜依赖性的单胺氧化酶，LOXL1 基因敲除的小鼠出现弹性纤维大量流失，原弹性蛋白增加的情况，进而发生 POP[1]。

TGF-β 是一类具有高活性和多效能细胞生长因子，具有广泛的生物活性.而 TGF-β1 可促进细胞外基质(ECM)的合成，同时可通过多条不同途径实现对 LOXL1 的调控。

本研究拟通过对 GEO、ArrayExpress 等高通量数据库已有数据的数据挖掘来实现 LOXL1 与 TGF-β 相关的分子路径的预测，从而在一定程度上了解 POP 的发病机制。

2. 数据收集

本实验主要收集高通量的生物芯片数据以及 Chip-SEQ 数据，生物芯片数据来自 NCBI 的 GEO datasets 数据库，以 pelvic organ prolapse 检索，经过挑选选择了如下两个数据集：

Gene expression profile in pelvicorgan prolapse：17 个样本(发生 POP 的样本数为 8，正常的对照组样本数为 9)，每个样本有两次重复[2]。

Endometriosis——endometrial tissue：19 个样本。(子宫内膜异位症的样本数为 10，正常的对照组样本数为 9)[3]。

生物芯片的数据主要用于表达聚类，不涉及表观修饰以及染色体的蛋白质互作关系的探究。

Chip-Seq 的数据主要来自 UCSC 的 ENCODE，这里没有用到原始数据，而是使用已经整合完毕的数据。

3. 数据处理

3.1. 线性模型

生物芯片分析常用的方式是相关性分析，这里应用的芯片数据没有取对数，而是以直接表达量作为变量，将芯片中的基因与 LOXL1 的表达量作线性相关分析，LOXL1 的表达量为因变量(即 y)，其余基因的表达量为自变量(即 x)。使用的统计软件为 R，相关系数的计算公式为皮尔逊相关系数：

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

3.1.1. Gene Expression Profile in Pelvicorgan Prolapse

该芯片与盆腔器官脱垂相关度非常高，所以将作为主要的研究材料。该芯片共涉及 32,878 个探针，经过筛选有对应基因的探针，余下 9627 个探针。利用 R[4]对这 9627 个表达进行线性相关分析，取显著值 <0.01 的结果，最后得到 163 个探针对应的基因。在这 163 个结果的基础上，我们进行了基因注释以及通路的查找，全部流程可见图 1。

发现与 LOXL1 相关性最高的是 TGF-β 通路和细胞外基质(ECM)合成通路，这个结果无疑也提示了 TGF-β 与 LOXL1 之间的调控关系。

3.1.2. Endometriosis——Endometrial Tissue

因为线性模型的应用，会使得结果相对地有较多的假阴性，所以使用该生物芯片数据作为前一个数据的对比，经过相同的分析步骤，得到最终的 332 个探针位置，将这 332 个探针相关的基因输入 DAVID[5]，发现相关的通路有：FOCAL adhension、ECM-receptor interaction、MAPK pathway、TGF-β 通路等。总体对比，主要的通路与上一芯片有重叠(如 ECM、TGF-β)，但是出现了更多新通路的注释。

3.1.3. 交集与并集结果

为了能够直观地看出两个芯片结果之间的关系，我们绘制了相关的韦恩图，见图 2。其中 15 个基因两个芯片分析结果共有，分别为 GPR88、FMOD、COL16A1、MFAP2、MMP23B、HSD11B1、COL8A1、SPARC、MTHFD1L、C7ORF10、MPP6、ITGBL1、MYH3、COL1A1、RCN3。可以说，这 15 个基因与 LOXL1 具有相似调控模式或者有相互作用的可能性

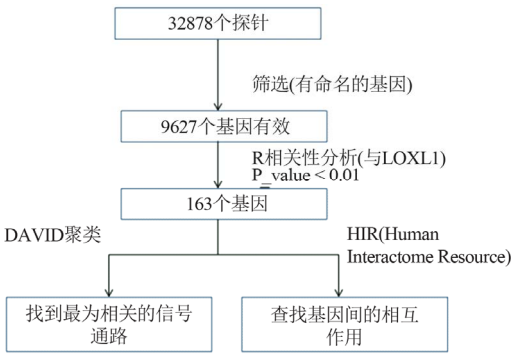

Figure 1. The flow & result of gene correlation analysis
图 1. 基因相关性分析流程及结果

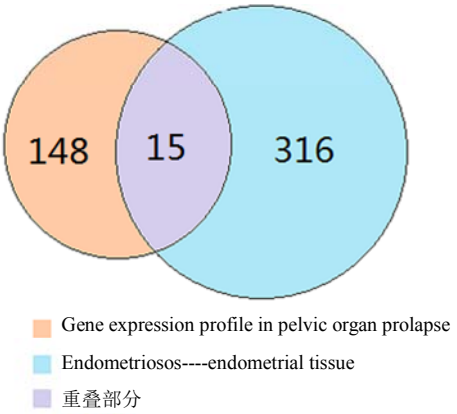

Figure 2. The Venn diagram for two microarrays
图 2. 芯片结果的韦恩图

非常高。为了将这 15 个基因与 LOXL1、LOX 家族蛋白联系起来，我们对这 15 个基因进行了注释，其结果见表 1(增加了 LOXL1 和 LOX)。

3.2. 数据挖掘与聚类

3.2.1. 数据预处理

GEO 数据库上的数据通常会遵循微阵列实验最小信息守则(MIAME)，包括原始数据、最终处理后的数据(处理如标准化)、该实验的必要注释、实验设计(包括样本数据间的关系)、该微阵列芯片的信息、必要的实验室和数据处理的信息。在以上这些中，最需要在数据分析过程中进行关注的是原始数据和最终处理后的数据，同时如果取得的是最终处理后的数据，需要明确实验者对数据做了哪些处理。

生物芯片在经过背景校正、缺失值处理、数据过滤和标记之后，最重要的就是进行数据的标准化。未经过标准化处理的芯片数据基因之间往往都呈现出很强的相关性，这些高相关性一部分是由基因表达水平变化引起的，而另外一部分是由系统偏差引起的。对芯片数据进行标准化处理的目的之一是消除系统偏差引起的高相关性，同时保留由真正生物学原因引起的基因表达水平高相关性。生物芯片的标准化有芯片内的标准化(Lowess Normalization、Print-tip Normalization)、芯片间的标准化(Quantile Normalization、Global Normalization、Median Normalization)，在进行数据预处理时，我们应该关注这方面的信息。

Table 1. Annotation result of the 15 genes
表 1. 15 个基因的注释结果

Gene name	CYTOBAND	Extracellular_region	Signal	Interal to membrane	Transmembrane	AP1 in promoter
GPR88	1p21.3			√	√	
FMOD	1q32	√	√			
COL16A1	1p35 - p34	√	√			√
MFAP2	1p36.1 - p35	√	√			√
MMP23B	1p36.3, 1p36.33	√		√	√	√
HSD11B1	1q32-q41			√	√	√
COL8A1	3q12.3	√	√			√
LOX	5q23.2	√	√			
SPARC	5q31.3 - q32	√	√			√
MTHFD1L	6q25.1					√
CYORF10	7p14.1					√
MPP6	7p15					
ITGBL1	13q33	√	√	√		√
LOXL2	15q24 - q25, 15q22	√	√			√
MYH3	17p13.1					√
COL1A1	17q21.33	√	√			√
RCN3	19q13.33		√			√

3.2.2. 聚类结果

该部分只使用 POP 相关的生物芯片进行聚类，使用的是 cluster3.0。芯片经过 Quantile Normalization，不需要再进行芯片间的标准化。

这里选择的聚类方法是 hierarchical clustering，它是最常用的一种聚类方法，其基本思想是：首先定义样本间的距离或相似系数以及类与类间的距离，一开始将 N 个样本各自看成一类，此时类间的距离与样本间的聚类是等价的。然后计算各类之间的距离，选择其中距离最小的两类合并成为一个新类。计算这一新类与其他各类之间的距离，再合并其中距离最小的两类。如此反复进行，每次减少一类，直到所有样本归成一类。

Cluster3.0[6]过程先对数据取对数，为了使得基因存在上下调的区别，需要进行基因间的 Median Normalization。选用类平均法，相似矩阵为 correlation (uncentered)。聚类后的结果可见图 3。

4. 结果分析

4.1. 文献整合及其验证

相对来说，LOXL1 在青光眼方面比较为人所知，通常它的突变被认为是导致该种疾病的一个原因。所以，有一些研究青光眼的学者已经对参与 LOXL1 调控的通路有了一些探究。如已知的 LOXL1 与 FBLN5 的蛋白质间相互作用，以及 TGF-β 通过 smad 以及非 smad 途径参与调控 LOXL1。已经有文献证实，TGF-β 表达量的变化可以影响 LOXL1 的表达，该文献同样探究了 LOXL1 的转录因子 AP1，他们用 western blot、QPCR 的方式证明了 AP1 可以调控 LOXL1 的表达[7]。尽管已经由文献指出了 LOXL1 的上游路径，但是 AP1 调控得到的结果并非最直接的证据，由此，我们着手寻找与 LOXL1 相关的 Chip-Seq 数据。

为了证明 AP1 确实能够结合 LOXL1 的启动子区域，我利用了 UCSC 上的 Chip-Seq 信息，发现在 LOXL1 上游 2000 bp 左右的位置，存在 c-Fos 和 c-Jun 的结合位点，如图 4。Western blot 结果与芯片结果相结合，我已经由较大的把握认为 AP1 是 LOXL1 的转录因子之一。

4.2. 生物芯片结果分析

本项目使用了两个从 GEO 数据库上下载的生物芯片数据，分别对两者进行了线性相关性分析以及聚

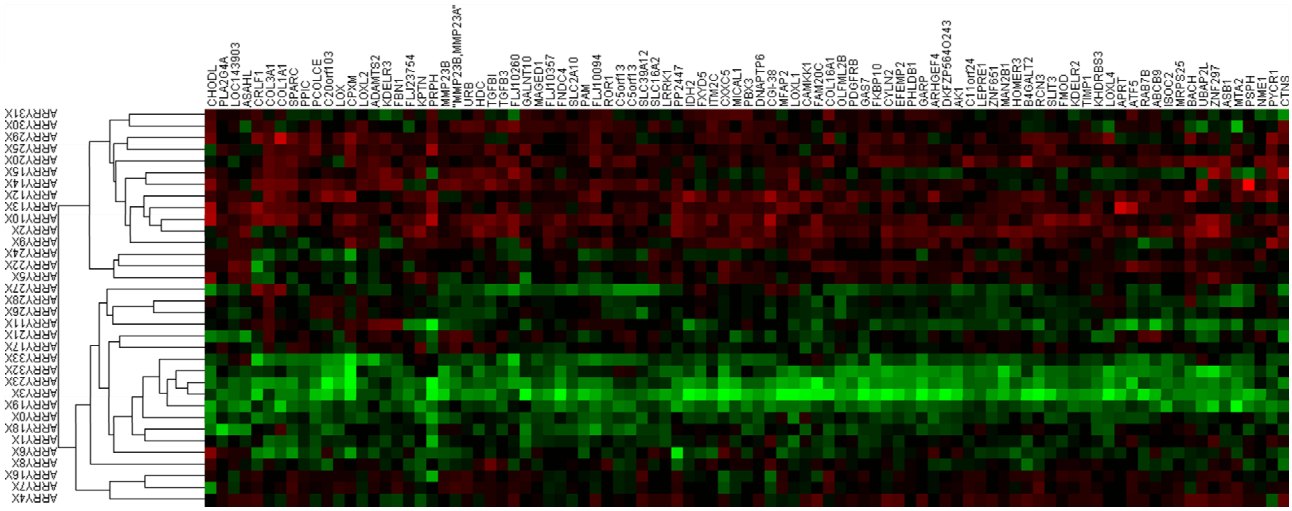

Figure 3. Hierachical clustering
图 3. 系统聚类

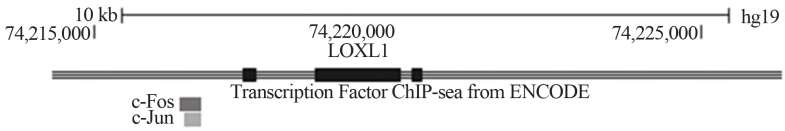

Figure 4. TF bind to promoter region of LOXL1
图 4. LOXL1 的启动子区域结合蛋白

类分析。

4.2.1. 线性相关结果分析

我们在实验中选取了两个生物芯片进行先行相关结果分析，对得出的结果分别取了交集和并集。但就并集而言，我们得到的有参考意义的结果是相关的调控通路，FOCAL adhension、ECM-receptor interaction、MAPK pathway、TGF-β 等通路，主要为之后构建分子路径提供参考。

而我们得到的交集共有 15 个基因，这 15 个基因同时存在于两个芯片中且与 LOXL1 的相关性显著值都小于 0.01。主要分析这 15 个基因之间是否可能存在未知的相互作用。

我们通常这样解读相关性高的基因：他们在很大程度上功能相关、调控同一代谢通路、受同一转录因子控制或者这些基因之间存在相互作用。

而我们在这里进行的线性相关分析，最主要的目的在于找出最为相关的关键基因，由关键的基因开始延伸构建基因相互作用的网络。

首先，我们来看是否有基因是因为受同一转录因子控制而相关。我们知道，因为位置相近而受到同一转录因子调控的基因通常在功能上也存在相似性。而同时，一个基因会存在多个不同的转录因子，已知的实验结果存在一些假阳性。所以，在这个假设下，我们先探究了有多少基因与 LOXL1 位置相近。在经过注释之后(见表 1)，可见挑选出来的 15 个基因没有与 LOXL1 位于同一条染色体的。在考察了基因位置之后，查看基因的转录因子是否存在 AP1，具体可见表 1，有 12 个基因存在 AP1 的结合，但是在经过进一步考察之后，发现对于绝大多数而言，AP1 结合位置在 5’-UTR 和 3’-UTR 之间。

再者，在经过基因注释之后，我们可以看到有 8 个与细胞外基质相关，因为 LOXL1 主要在细胞外基质起作用，所以有 8 个基因与细胞外基质有关是完全可以接受的，也从一方面验证了这 15 个基因与 LOXL1 相关的可靠性还是比较高的。

4.2.2. 聚类结果分析

在聚类分析的结果中，包含了基因的聚类和芯片的聚类。

从基因的聚类中可以查找功能相关、结构相似的基因，而从芯片的聚类中可以找出有相似基因表达谱的样本。经过与各样本状态的比对，发现 POP 的发病与否与芯片样本的聚类并无直接关系。也就是说，表达谱相似的样本同时存在 POP 病人与非 POP 病人。这就提示 LOXL1 并非 POP 发病的必要条件，甚至有些 POP 病人出现 LOXL1 的表达上调，猜测这可能是一种代偿机制。

5. 总结

在利用数据挖掘的方法探究 LOXL1 与 TGF-β 之间的分子通路之后，我们得出的结论是这样的：

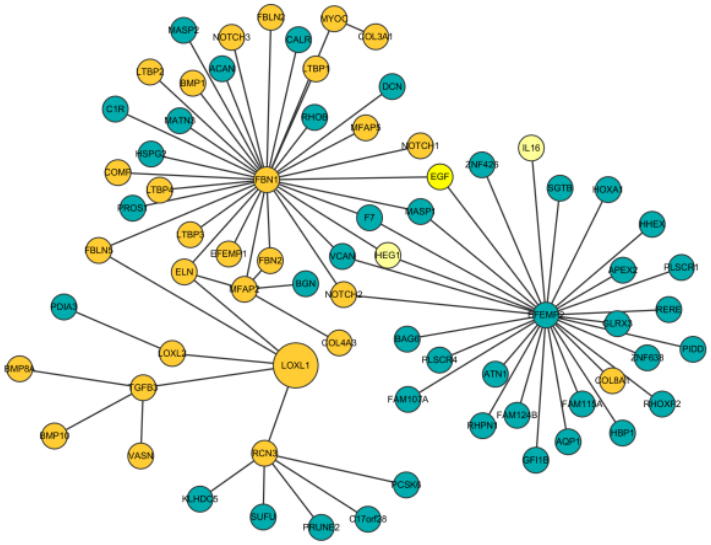

Figure 5. Biology network
图 5. 生物网络

由 TGF-β 通过 smad 和非 smad 通路调控 LOXL1 的表达，非 smad 通路包括 JNK、MAPK 路径。也就是说 TGF-β 能够诱导 LOXL1 的表达，促进细胞外基质弹性纤维和胶原的交联。而 LOXL1 分泌的过程中，比较可能起到作用的蛋白质则是内质网的网钙蛋白 RCN3。而在 LOXL1 出细胞之后，则与 FBLN5[1]一起参与促进弹性蛋白原的赖氨酸残基的去氨基化。

最后，在结合实验已有的蛋白质互作数据和计算机预测的蛋白质互作数据后[8]，我们建立了一个基因互作网络[9]，具体的生物网络可见图 5。

6. 感谢

本项目由浙江省大学生科技创新活动计划(2013R401065)资助，感谢浙江大学医学院干细胞与组织工程 B610 实验室的各位老师、同学提出意见与建议，感谢浙江大学 210 生物信息实验室提供实验平台。

参考文献 (References)

[1] Liu, X., Zhao, Y., Gao, J., et al. (2004) Elastic fiber homeostasis requires lysyl oxidase-like 1 protein. *Nature Genetics*, **36**, 178-182.

[2] Brizzolara, S.S., Killeen, J. and Urschitz, J. (2009) Gene expression profile in pelvic organ prolapse. *Molecular Human Reproduction*, **15**, 59-67.

[3] Hawkins, S.M., Creighton, C.J., Han, D.Y., et al. (2011) Functional microRNA involved in endometriosis. *Molecular Endocrinology* (*Baltimore*, *Md*.), **25**, 821-832.

[4] R. Core Team (2013) R: A language and environment for statisticalcomputing. http://www.R-project.org/

[5] Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protocols*, **4**, 44-57.

[6] Eisen, M.B., Spellman, P.T., Brown, P.O., et al. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of USA*, **95**, 14863-14868.

[7] Sethi, A., Mao, W., Wordinger, R.J., et al. (2011) Transforming growth factor-beta induces extracellular matrix protein cross-linking lysyl oxidase (LOX) genes in human trabecular meshwork cells. *Investigative Ophthalmology & Visual Science*, **52**, 5240-5250.

[8] Zhang, Q.C., Petrey, D., Garzon, J.I., et al. (2013) PrePPI: A structure-informed database of protein-protein interactions. *Nucleic Acids Research*, **41**, D828-D833.

[9] Cline, M.S., Smoot, M., Cerami, E., et al. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols*, **2**, 2366-2382.