PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2016, 43(1): 5~13
www.pibb.ac.cn

基因组织特异性相关研究进展 *

刘 伟 ** 孙志强 谢红卫
(国防科学技术大学机电工程与自动化学院自动控制系，长沙 410073)

摘要 研究基因的组织特异性是了解生命活动进程和组织功能的重要一步．尽管对于看家基因和组织特异基因的研究由来已久，但是对于它们仍缺少统一的定义方式和检测方法．在定义方式上，可以从基因的组织表达数和在各组织间的表达变化情况来分别定义看家基因和组织特异基因．通常将在大多数正常组织中有表达，且表达水平较稳定的基因称为看家基因，而将在一个或少数组织中优势表达的基因定义为组织特异基因或组织选择基因．在检测方法上，高通量实验技术，包括基因芯片、RNA-seq 和质谱技术等已成为检测基因组织特异性的主要方法．通过比较多个典型研究的实验结果，发现不同检测方法的覆盖度和灵敏度存在很大差异，其中 RNA-seq 技术最为灵敏，获得的看家基因数目最多，质谱技术检测出来的看家基因和组织特异基因数目较少，而基因芯片方法给出的多个检测结果间差别较大．尽管不同的定义方式和检测方法所导致的看家基因(或组织特异基因)的集合不完全一致，但不同的看家基因数据集(或组织特异基因)却展现出非常一致的功能和特性．看家基因通常实现所有组织和细胞都必须的基本功能，而看家基因与其他组织表达基因间的相互作用以及组织特异基因间的相互作用则实现了组织的特有功能．同时，基因的组织特异性与疾病之间具有密切联系，相比其他基因，看家基因更有可能成为癌基因，而组织特异基因则更有希望发展成为药物靶标．

关键词 看家基因，组织特异性，疾病
学科分类号 Q61
DOI: 10.16476/j.pibb.2015.0268

基因的组织特异性对于研究组织内的生命活动过程和蛋白质功能具有重要意义[1-3]．近年来，各种组织相关分子表达数据的大规模增长，为基因组织特异性的检测和分析提供了重要机遇[4-5]．看家基因最初定义为在所有细胞系中都有表达的基因，近年来，也用于指征那些具有稳定表达、用于维持细胞功能的基因．由于看家基因在各组织中的广泛表达，它们被认为是必要基因的候选．而组织选择基因是指那些在一个或少数几个组织类型中优势表达的基因．其中，仅在一个组织中独特表达的基因则称为组织特异基因．基因的组织特异表达往往预示着它们具有与组织相关的功能，因此更可能成为潜在的药物靶标或疾病标志物．本文对看家基因和组织特异基因的定义、功能、特性以及与疾病的关联进行了总结和阐释．尽管看家基因和组织特异基因的定义和检测方法多样，但是对其进行后续分析发现，看家基因和组织特异基因在功能、特性上都具有显著区别．这些区别对于理解器官组成和运行方式以及相关的疾病和药物研究具有一定参考价值．

1 基因组织特异性的定义

根据基因在各组织中的表达丰度和范围，可以划分基因的组织特异性．在相关研究中，人们最为关注的是看家基因和组织特异基因 / 组织选择基因，而那些展现出中间范围的表达模式基因则研究较少[6]．尽管看家基因和组织特异基因的概念被广泛使用，但是对于它们的内涵却有着多种不同的解释．

1.1 看家基因

按照基因的组织表达特性，看家基因有两种定

* 国防科学技术大学科研计划项目(JC15-03-01)，国际合作项目(2014DFB30010)和国家自然科学基金(31171266)资助.
** 通讯联系人.
Tel: 0731-84573369, E-mail: angel_nudt@126.com
收稿日期：2015-11-12，接受日期：2015-12-14

义方式：第一种是指在所有组织 / 细胞类型中都有表达的基因，这是较为常用的一种定义方式．实际上，由于测量误差和随机噪声的影响，有些基因由于表达水平很低，当低于指定阈值时就无法确定其是否为看家基因，这称为“表达泄露”．当采用基因芯片等手段检测基因在组织样本中的表达时，受噪声影响对于每个转录本总能检测到一定量的表达，因此需要人为设定阈值来排除噪声干扰．尽管大部分的看家基因相比组织特异基因的表达丰度更高，但是有些看家基因，如转录因子，可能具有较低的表达丰度，那么采用统一的阈值设定法就无法识别这些基因．鉴于此，Butte 等[7]提出了第二种定义方式，即将在不同组织中具有常数表达或稳定表达的基因定义为看家基因[8-9]．这种定义方法可以涵盖表达水平较低的看家基因，因此在最近的一些研究中得到了推广应用[10-11]．对于看家基因目前尚没有统一的、严格的定义，通常做法是将在大多数正常组织中有表达，且表达水平较稳定的基因作为看家基因．

1.2 组织特异基因

与看家基因的定义相类似，组织特异基因也可以从基因的组织表达数和在各组织间的表达变化情况来分别定义．常见的定义方式是将在一个或少数组织中有表达的基因定义为组织特异基因，或组织选择基因．另外一种定义则基于基因在各组织中表达的均衡性，将在一个或少数组织中优势表达的基因定义为组织特异基因，但实际上不同文献中提出的筛选标准不尽相同．如 Dezso 等[8]考虑了噪声的影响，将在一个组织中独特表达，且信噪比大于 10 的基因定义为组织特异基因．Yanai 等[6]提出了一种组织特异性指标 TSI，综合基因的组织表达数和表达量变化情况来考察基因的组织特异性．当基因仅在一个组织中有表达时，TSI=1；当基因在多个组织中有表达，但在某一组织中的表达量比其他组织中的表达量明显高得多时，TSI 是一个接近于 1 的值．通过设定 TSI 的阈值，可以筛选出在少数组织中有表达且表达丰度较高的基因．Pan 等[12-13]提出了三个量化指标，包括组织特异指标 SPM、散布指标 DPM 和贡献指标 CTM 以识别组织特异基因和组织选择基因． 其他类似的指标有综合检测置信度和表达丰度的加权指标 FPEI[10]、基于基因表达水平排序的 HKera[14]、基于谱分析的指标[15]等．由于缺少标准数据集，目前对于这些定义的优劣还难以定量地评估．

2 基因组织特异性的检测方法

检测基因的组织特异性是了解基因功能的重要基础．早期，研究人员主要借助于小规模的实验技术，如 RT-PCR 和 Western blot 技术来进行小范围的基因组织特异性的检测．而近年来，随着高通量实验技术，包括基因芯片、RNA-seq 和质谱技术的发展，使得人们可以从不同层面、大规模地检测和分析基因的组织特异性．

2.1 小规模实验技术

RT-PCR 和 Western blot 等实验技术可用于检测基因的组织表达情况，结果准确性较高，但由于通量限制，往往规模较小．RT-PCR (reversed transcript PCR)是一种将 cDNA 合成与 PCR 技术结合分析基因表达的快速灵敏的方法，主要用于对表达信息进行检测或定量分析，还可以用来检测基因表达差异而不必构建 cDNA 文库克隆 cDNA[16]．目前，小规模的实验技术主要用途有：通过考察看家基因在不同组织中的表达均衡性，检验其是否适合作为芯片等实验的内标[17]；大规模组织特异性实验的部分数据验证[18]；评估某种实验技术，如 RNA-seq 中转录本定量的准确性[19]；少数基因的功能研究等[20]．

2.2 基因转录组技术

自从转录组技术出现以后，研究人员一直致力于利用该技术寻找人类看家基因和组织特异基因，以了解基因组的结构和生物学过程的基本原理．相比小规模实验技术，组织相关的基因表达数据规模较大、易于获取，但受噪声影响较大，结果往往不够准确．

在利用基因芯片技术检测基因组织表达特异性的实验中，Su 等[21]的研究规模最大，应用也最广．他们采用基因芯片检测了 79 个人体组织和 61 个小鼠组织中的基因表达情况，估计约有 6%的基因是广泛表达的，而每个组织中表达所有基因的 30%～40%．其他的大规模组织特异性研究有 Ge[22]和 Dezso[8]等的工作，检测的组织数目分别为 36 和 31 个．为了充分利用现有组织特异数据集以及对不同实验室获得的芯片数据集进行比较，也有一些团队采用了荟萃分析的研究方法[10, 23-24]．如 Chang 等[10]编辑了来自 104 个基因芯片数据集中 43 个人体正常组织的 1 431 个样本，通过基因表达评估方法，筛选出 2 064 个看家基因和 2 293 个组织选择基因．考虑到样本规模、组织数目、数据质量和筛

选标准等诸多因素对最终结果的影响，荟萃分析需要采用非常严格的数据处理方法对原始数据集进行重新处理．理论上，由荟萃分析得到的基因集合相比单一来源的数据集更具有鲁棒性．

近年来，随着下一代测序技术的发展，高通量测序数据被越来越多地用于检测基因的组织特异性，以获得具有更高敏感度的结果[25-27]．在RNA-seq 实验中，样本的所有 RNA 被随机分段、反转录、连接到转接子、测序，采用 RPKM(Reads Per Kilobase of exon model per Million mapped reads)来估计基因在各组织中的表达量．相比基因芯片技术，RNA-seq 技术的实验可重复性更好，对于基因的低表达和差异表达更敏感，与蛋白质表达水平具有更好的相关性，因此检测结果更加准确，覆盖度更高[28]．Ramskold 等[19]的工作具有一定代表性，他们发现了 7 897 个基因在研究的所有 24 个组织中都有表达，占总基因数目的 42%．最近，Uhlen 等[29]通过将在组织和器官水平上的定量转录组学技术与基于组织微阵列检测的免疫组织化学相结合，更是将蛋白质空间定位精度提高至单个细胞水平．他们的研究结果表明，44%的蛋白质编码基因在所有分析的 32 个组织中都有表达，约有 12%的基因仅在一个特定组织中表达且表达水平为其他组织平均表达水平的 5 倍以上．一个有趣的现象是，他们发现在很多其他文献中描述的组织特异基因实际上在多个组织中都有表达，只是在某一组织中表达水平较高，而在其他组织中表达水平较低而已．

2.3 蛋白质检测技术

鉴于转录表达水平与蛋白质表达水平的中度相关(0.3~0.7)，研究人员一直试图从蛋白质层面获得蛋白质组织表达谱[30-32]．但受限于质谱检测技术的覆盖度问题，早期研究所能提供的组织表达蛋白质数目较少，定量不够准确[33]．近年来，随着质谱实验技术和数据处理方法的日益完善，测定大规模蛋白质组织表达谱成为可能．2014 年，《自然》(*Nature*)杂志上连续发布了两个大规模的人类蛋白质组学图谱，使得人们能够直接从蛋白质表达水平研究基因的组织特异性[34-35]．其中，Kim 等[34]利用高分辨率的傅里叶变换质谱技术，检测了 17 294 个基因(占所有已注释基因的 84%)在 30 个正常人体组织中的表达，包括 17 个成人组织、7 个胚胎组织和 6 个基本的造血细胞器．而 Wilhelm 等[35]通过整合分析 ProteomicsDB 数据库的多个质谱数据集，将研究规模扩展至 92%的已注释基因，揭示了它们在 47 个组织和体液中的组织表达谱．通过分析大规模蛋白质图谱，Liu 等[36]考察了看家基因和组织特异基因的表达特性，发现相比基因表达数据，由蛋白质表达谱数据推断获得的组织特异网络规模更大．

尽管蛋白质组学数据可以直接反映蛋白质在不同组织 / 细胞系中的表达水平，而且在预测蛋白质相互作用方面比转录组数据更加有效，但仅从大规模数据集出发确定蛋白质的组织特异性还存在一定问题．例如，Kim 等[34]对由质谱方法发现的 32 个组织特异蛋白质进行 Western blot 实验验证，结果发现仅有 8 个蛋白质被证实为组织特异表达，其余的 24 个蛋白质或者未被检测到或者在多个组织类型中被检测到．这提示我们，由高通量技术获得的蛋白质组织表达数据存在相当程度的假阳性．Ezkurdia 等[37]的研究则提示这两个大规模数据集有可能高估了已识别出的基因编码蛋白数目，因此需谨慎使用．

2.4 不同检测方法的比较

以上三类方法为基因的组织特异性检测提供了多种备选方案．由于检测方法的敏感性不同、特征提取方法的差异以及采样不完整等问题，目前对于看家基因的研究存在着相当程度的假阳性．Zhang 等[38]比较了已发表的 187 个组织和细胞类型中的 15 个看家基因数据集，出乎意料的是，仅有 1 个基因(peroxiredoxin 1，PRDX1)出现在全部的 15 个数据集中，17 个基因出现在 14 个数据集中．也就是说，不同研究获得的看家基因之间交集很小．造成这种现象的原因有两个方面：一是不同实验研究的检测技术不一样，数据处理方法也不一样；二是由于组织可能处于各种发育、病理和生理状态，如将看家基因严格地定义为在所有组织类型中都有表达，那么很可能在某些状态的组织或细胞系中无法检测到．因此，有更多的研究人员转向支持看家基因的第二种定义方式．

类似于看家基因，由不同检测技术获得的组织特异基因数据集间的差别也很大．这是由于目前研究所包含的组织样本类型通常是不完整的，人体中共有 200 个左右的组织和细胞类型，而在一次大规模基因表达实验中研究的组织样本数一般只有几十个，还不到所有组织类型的一半．这就很可能导致如下现象的出现：随着研究的组织数目越多，检测到的组织特异基因的数目就越少．也就是说，有些基因并非在一个组织内有表达，只是由于研究的样本组织类型所限，尚未检测到该基因在其他组织中

的表达. 与之类似，当研究的组织样本数增加时，检测到的看家基因的数目也将减少.

表 1 给出了近年来有代表性的基因组织特异表达数据集，相关研究得到的看家基因和组织特异基因参见附件表 S1 和表 S2. 经比较可以发现：由 RNA-seq 技术获得的看家基因数目最多，这可能是由于 RNA-seq 技术较为敏感，能够检测出基因在组织中的微弱表达；质谱技术检测出来的看家基因和组织特异基因数目较少，一方面的原因是研究中的总基因数目相对较少，另一方面则是质谱方法的检测和定量技术还不够成熟；基因芯片方法给出的检测结果差别较大，这可能源于基因芯片实验的数据重复性不够好，以及后续处理方法的多样性.

Table 1 The typical researches about gene tissue specificity

表 1 有代表性的基因组织特异性研究

作者	年份	检测技术	组织数	总基因数	看家基因数	特异基因数	参考文献
Jongeneel 等	2005	MPSS	32	18 677	4 006	4 232	[39-40]
Zhu 等	2008	EST	18	17 288	3 140	885	[41]
Su 等	2004	基因芯片	79	22 283	1 789	1 119	[21]
Ge 等	2005	基因芯片	36	22 283	7 841	2 503	[22]
Dezso 等	2008	基因芯片	31	32 878	2 374	1 381	[8]
Ramskold 等	2009	RNA-seq	24	18 805	7 897	3 375	[19]
Eisenberg	2013	RNA-seq	16	34 475	5 348	3 632	[9]
Uhlen 等	2015	RNA-seq 和组织芯片	32	20 344	8 874	2 355	[29]
Kim 等	2014	质谱	30	17 294	2 350	1 537	[34]
Wilhelm 等	2014	质谱	48	18 097	6 467	–	[35]

3 不同组织特异性基因的功能与特性

尽管不同的定义方式和检测方法所导致的看家基因(或组织特异基因)的集合不完全一致，但在分析中发现，不同的看家基因数据集(或组织特异基因)却展现出非常一致的功能和特性. 换句话说，看家基因和组织特异基因之间的差别是如此之大，以至于它们不会被检测方法的假阳性和假阴性所掩盖.

3.1 不同组织特异性基因的功能

看家基因的广泛表达说明，它们的产物是所有细胞都需要的，用于保持基本的细胞结构和功能. 通常，看家基因用于实现各种组织所需要的重要的生物学功能，主要包括如下几类：参与细胞合成的核糖体蛋白；细胞代谢和基因表达必需的酶；能量产生所需的线粒体蛋白；用于保持细胞结构完整性的蛋白. 各种研究所得到的结论基本一致，如 Prieto 等[42]对三个不同实验室获得的看家基因数据集进行功能分析，发现它们最为富集的前三项都是蛋白酶体、核糖体和氧化磷酸化.

组织特异的蛋白质与该组织要实现的功能一致，同时反映了不同组织的相似性和特异程度[8]. 一般，功能相近的器官间拥有很高比例的共有表达基因，如子宫颈和食道展现出非常相似的基因表达模式，这是由于它们的组织特异基因都与上皮发育相关. 而某些实现特殊功能的器官则拥有更高比例的专有表达基因，如胰腺和睾丸特异基因很少在其他组织中有表达，说明它们的功能是这些组织专有的[29]. 各个器官通过调整看家基因和组织表达基因的比例来实现各自独特的功能. 如扁桃体表现出上皮和免疫组织的混合基因表达模式，说明扁桃体由这两类组织构成.

3.2 不同组织特异性基因的特性

为了实现不同的功能，看家基因与组织特异基因在组织表达特性、生物物理学属性和网络连接属性上都展现出显著的区别.

作为区分看家基因和组织特异基因的主要特征，看家基因与组织特异基因在组织表达特性上具有明显差异. 由于看家基因参与最基本的细胞保持作用，一般认为看家基因在所有的细胞和条件下都保持了常数的表达水平. 因此，在多种生物学技术和基因组研究，如 RT-PCR、基因芯片、RNA 印迹和 RNase 保护芯片中，看家基因都作为内标来使用. 很多实验研究证实了这一点，Uhlen 等[29]发

现大多数的看家蛋白质在人体各组织中的表达水平相似，如细胞核膜蛋白 SUN2．但也有部分看家蛋白质表现出了组织表达的倾向性，如编码线粒体蛋白质的转录本在心肌(占所有转录本的 32%)和骨骼肌(28%)中有较高比例的表达，说明了它们对于横纹肌组织中能量代谢的重要性．因此，并非所有的看家基因都适合作为实验内标，一般在使用前需要进行实验确定．相比之下，组织特异基因在各组织中的表达差异较大，但在同一组织中的特异蛋白往往具有类似的基因表达模式．同时，不管是看家基因还是组织特异基因在不同物种的同源基因间具有非常相似的表达丰度，如在人和小鼠的肌肉组织内同源基因间的表达相关性为 0.76，肝和脑中同源基因间的表达相关性为 0.77[19].

在生物物理学属性上，看家蛋白与组织特异蛋白具有显著的差别[43-44]．相比组织特异蛋白，看家蛋白质倾向于序列更短[45]、包含更多的短重复序列[46]、包含更少的蛋白质结构域[47]、表现出更低的启动子序列保守性[48]、具有更简单的转录调控和更慢的进化速率[49-50]．也就是说，为了实现细胞内的基本功能，看家基因相比组织特异基因通常在进化上更加保守，具有更加简单的蛋白质结构，以保持其功能的稳定性．基于这些能够指征蛋白质组织特异性的属性，也有一些研究人员利用机器学习方法来构建看家蛋白与组织特异蛋白的分类器[51]．但由于看家蛋白与组织特异蛋白在部分生物物理学属性间存在很大程度的交叠[52]，此类方法的预测准确率通常不高．

通过将通路信息与组织特异表达数据相结合构建组织特异的生物网络，可以鉴别看家基因与组织特异蛋白在网络拓扑属性上的差异[5, 53]．相比网络中所有节点的拓扑属性分布，看家基因往往具有更高的连接度和更短的蛋白质间通路距离; 而相比那些广泛表达的蛋白质，组织特异性越强的蛋白质其相互作用的数目更少，更可能是进化上比较年轻的蛋白质[36]．Zhu 等[41]的研究得到了与前人基本一致的结论，他们发现超过一半的中心蛋白属于广泛表达的单元．同时，看家基因与组织特异基因在连接模式上表现出显著的差异[36]．看家基因不仅与看家基因发生相互作用，而且与组织特异基因之间存在广泛的连接; 组织特异基因则倾向于同一组织的特异基因发生相互作用．这些研究成果可以帮助我们理解组织的基本结构，在组织的蛋白质相互作用网络中，看家基因编码蛋白形成了网络的核，而组织特异基因编码蛋白则形成了一个个簇连接在核的周围[54].

3.3 基因组织特异性与疾病的关系

基因的组织特异性与疾病之间具有密切联系，基因在人体内的表达范围和表达丰度决定了当其发生异常时将引起多大范围的身体反应，从而影响了其与疾病的关联程度和成为药物靶标的可能性．

将基因的组织特异性与疾病进行关联研究，有助于发现疾病的致病机理．研究发现，已知的疾病相关基因多为组织特异基因，它们通过在某些组织中过度表达来引发病理状态[55-57]．这说明大部分疾病具有明显的组织特异性．一个例外情况是癌症．研究表明，大部分的癌基因(约 60%)为看家基因，仅有少数癌基因表现出了组织特异性[29]．鉴于大部分的癌基因参与正常的生长调节和细胞周期调控，癌基因表现得缺乏组织特异性并不让人意外．进一步，比较各基因在正常组织和对应癌症细胞系中的表达情况，可用于识别癌症细胞系中的突变基因．研究表明，70%以上在正常细胞中表达的看家基因在癌细胞系中也有表达，且表达水平相当[58]．不同癌细胞系间存在着上百个专有看家基因，即在各癌细胞系中都有表达而在正常细胞中没有表达的基因，它们多与细胞调控相关，实现对抑制增长信息不敏感、抵抗细胞凋亡、持久的血管生成和无限复制潜能等功能．而很多在正常组织中出现的组织特异基因在对应的癌症细胞系中表现为低表达或者完全不表达[29]．这些结果说明癌细胞通常是“不分化的”，即各种癌细胞系通常具有相似的特性，而缺少组织特异的表达特性．这或许可以为不同癌症间的相似性以及癌细胞的转移性提供新的解释．

研究基因的组织特异性对于筛选药物靶标具有一定参考价值．大多数的药物通过作用于靶向蛋白并调节其活性来发挥作用，而靶向蛋白的组织表达特性则决定了药物影响的身体部位．因此，看家基因和组织特异基因在发展成为药物作用靶标的前景上展现出了一定差别．Uhlen 等[29]按照组织特异性对已知的药靶基因进行分类，发现尽管约 30%的药靶蛋白是看家基因，但是大部分药靶都表现出了组织倾向性．Dezso 等[8]比较了看家基因和组织特异基因集合中药物靶标的分布情况，发现在组织特异基因中注释为治疗性靶标(即现有药物的直接作用靶点)的蛋白质数量比看家基因中要多 1 倍(比例分别为 3.3%和 1.5%)．这一现象在乳腺疾病的相关药靶中尤为突出，治疗性药靶占乳腺和胸腺特异表

达蛋白的 25%．这提示我们，在大多数的疾病中组织特异蛋白相比看家蛋白更有望发展成为药靶，作用于这些靶点的药物将具有更好的组织作用特异性和更低的毒副作用．表 2 给出了基因组织特异性与疾病相关研究的代表性结果．

Table 2 The results inferred from investigations about gene tissue specificity and diseases

表 2 基因组织特异性与疾病相关研究成果

作者	年份	检测技术	研究方法	主要结果	参考文献
Ge 等	2005	基因芯片	考察组织特异基因在癌症样本中的表达情况	肝癌样本中肝脏特异基因的表达水平与肿瘤分化程度正相关，组织特异基因可帮助定位转移肿瘤的源头	[22]
Lage 等	2008	基因芯片	比较疾病基因与非疾病基因在组织中的表达情况	疾病基因通常是组织特异的，组织特异基因的选择性过表达引发疾病	[57]
Dezso 等	2008	基因芯片	比较组织特异基因和看家基因在已知药靶中的分布	组织特异基因成为药靶的几率是看家基因的 2 倍	[8]
Chen 等	2012	RNA-seq	比较癌症样本看家基因与正常样本的看家基因	癌症看家基因比正常看家基因具有更高的表达水平，AT 富集，参与细胞周期调控功能	[58]
Ganegoda 等	2013	基因芯片	利用组织特异的基因网络预测疾病相关基因	组织特异的基因网络相比通用网络能够更好地预测疾病相关基因	[59]
Uhlen 等	2015	RNA-seq 和组织芯片	考察已知药靶和癌基因的组织特异性	大部分的已知药靶具有组织特异性，而癌基因通常不具有组织特异性	[29]

4 结论与展望

作为功能基因组研究的一个重要组成部分，近年来，关于基因组织特异性的研究层出不穷，这是因为了解基因在哪些组织中有表达对于理解基因如何发挥作用以及基因之间的关系是非常重要的．通过系列的研究工作，人们对于基因的组织表达情况有了很多新的见解，主要体现在：a．基因的表达方式．基因的组织特异性不仅体现在该基因在一个、多个或所有组织中有表达，还体现在它表达的丰度以及与其他组织中表达情况的比较差异．b．基因的表达范围．早期研究所揭示的基因在组织中的表达范围是很有限的(约占所有基因的 30%)，但是随着各种更加敏感的检测方法的出现，人们发现基因在组织中的表达范围比已知的更高(约占所有基因的 60%)，也就是说基因并非是沉默的大多数，它们在组织中的表达是相当活跃的．c．基因的作用方式．看家基因以及看家基因间的相互作用实现了所有组织和细胞都必须的基本功能，而看家基因与其他组织表达基因间的相互作用以及组织特异基因间的相互作用则实现了组织的特有功能．d．基因对疾病的影响．基因表达范围和表达丰度的异常改变是造成疾病发生的根本原因，与组织功能相关的组织特异基因表达变化时可能引发局部的组织病变，而当与细胞分裂或损伤修复相关的看家基因表达发生变化时则可能引发致命的癌症．

预期下一步的研究工作将向着标准化、分析更深入、覆盖的组织类型更全面的方向发展，包括：a．针对看家基因和组织特异基因提出能够为大家广泛接受的统一定义方式，以方便不同研究之间的比较；b．综合比较各种检测方法的优缺点，建立基因组织特异性的标准检测方法，提升检测结果的准确度和敏感性；c．对人体内所有的组织器官和细胞系进行系统的采样和定量检测，消除少数组织研究的偏性；d．对已确知的看家基因和组织特异基因进行细致的功能分析和机理解释．相关研究的深入必将有助于人们更好地描述个体、组织、器官、细胞各个层面上生命活动的详细图画，揭示生命的奥秘．

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

参 考 文 献

[1] Hsiao L L, Dangond F, Yoshida T, *et al*. A compendium of gene expression in normal human tissues. Physiol Genomics, 2001, **7**(2): 97-104

[2] Tu Z, Wang L, Xu M, *et al*. Further understanding human disease genes by comparing with housekeeping genes and other genes. BMC Genomics, 2006, **7**: 31

[3] Hwang P I, Wu H B, Wang C D, *et al*. Tissue-specific gene expression templates for accurate molecular characterization of the normal physiological states of multiple human tissues with implication in development and cancer studies. BMC Genomics, 2011, **12**: 439

[4] Lee S, Jo M, Lee J, *et al*. Identification of novel universal housekeeping genes by statistical analysis of microarray data. J Biochem Mol Biol, 2007, **40**(2): 226−231

[5] Bossi A, Lehner B. Tissue specificity and the human protein interaction network. Mol Syst Biol, 2009, **5**: 260

[6] Yanai I, Benjamin H, Shmoish M, *et al*. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics, 2005, **21**(5): 650−659

[7] Butte A J, Dzau V J, Glueck S B. Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues". Physiological genomics, 2001, **7** (2): 95−96

[8] Dezso Z, Nikolsky Y, Sviridov E, *et al*. A comprehensive functional analysis of tissue specificity of human gene expression. BMC Biology, 2008, **6**: 49

[9] Eisenberg E, Levanon E Y. Human housekeeping genes, revisited. Trends in Genetics, 2013, **29**(10): 569−574

[10] Chang C W, Cheng W C, Chen C R, *et al*. Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. PLoS One, 2011, **6**(7): e22859

[11] She X, Rohl C A, Castle J C, *et al*. Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. BMC genomics, 2009, **10**(1): 269

[12] Pan J B, Hu S C, Wang H, *et al*. PaGeFinder: quantitative identification of spatiotemporal pattern genes. Bioinformatics, 2012, **28**(11): 1544−1545

[13] Pan J B, Hu S C, Shi D, *et al*. PaGenBase: a pattern gene database for the global and dynamic understanding of gene function. PloS ONE, 2013, **8** (12): e80747

[14] Chiang A, Shaw G, Hwang M. Partitioning the human transcriptome using HKera, a novel classifier of housekeeping and tissue-specific genes. PLoS ONE, 2013, **8**(12): e83040

[15] Dong B, Zhang P, Chen X, *et al*. Predicting housekeeping genes based on Fourier analysis. PLoS One, 2011, **6**(6): e21012

[16] Erickson H S, Alber P S, Gillespie J W, *et al*. Quantitative RT-PCR gene expression analysis of laser microdissected tissue samples. Nat Protoc, 2009, **4**(6): 902−922

[17] Touchberry C D, Wacker M J, Richmond S R, *et al*. Age-related changes in relative expression of Real-time PCR housekeeping genes in human skeletal muscle. Journal of Biomolecular Techniques, 2006, **17**(2): 157−162

[18] Park S W, Kang S W, Goo T W, *et al*. Tissue-specific gene expression of silkworm by quantitative real-time RT-PCR. BMB reports, 2010, **43**(7): 480−484

[19] Ramskold D, Wang E T, Burge C B, *et al*. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput Biol, 2009, **5**(12): e1000598

[20] Zhang J, Ahn J, Suh Y, *et al*. Identification of CTLA2A, DEFB29, WFDC15B, SERPINA1F and MUP19 as novel tissue-specific secretory factors in mouse. PLoS ONE, 2015, **10**(5): e0124962

[21] Su A I, Cooke M P, Ching K A, *et al*. Large-scale analysis of the human and mouse transcriptomes. Proc Natl Acad Sci USA, 2002, **99**(7): 4465−4470

[22] Ge X, Yamamoto S, Tsutsumi S, *et al*. Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. Genomics, 2005, **86**(2): 127−141

[23] Cavalli F M, Bourgon R, Huber W, *et al*. SpeCond: a method to detect condition-specific gene expression. Genome Biology, 2011, **12**(10): R101

[24] Xiao S J, Zhang C, Zou Q, *et al*. TiSGeD: a database for tissue-specific genes. Bioinformatics, 2010, **26**(9): 1273−1275

[25] Lee J H, Park I H, Gao Y, *et al*. A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. PLoS Genet, 2009, **5**(11): e1000718

[26] Emig D, Albrecht M. Tissue-specific proteins and functional implications. J Proteome Res, 2011, **10**(4): 1893−1903

[27] Tong C, Wang X, Yu J, *et al*. Comprehensive analysis of RNA-seq data reveals the complexity of the transcriptome in Brassica rapa. BMC Genomics, 2013, **14**: 689

[28] Fu X, Fu N, Guo S, *et al*. Estimating accuracy of RNA-Seq and microarrays with proteomics. BMC Genomics, 2009, **10**: 161

[29] Uhlen M, Fagerberg L, Hallstrom B M, *et al*. Tissue-based map of the human proteome. Science, 2015, **347**(6220) : 1260419

[30] Ghaemmaghami S, Huh WK, Bower K, *et al*. Global analysis of protein expression in yeast. Nature, 2003, **425**(6959): 737−741

[31] Griffin T J, Gygi S P, Ideker T, *et al*. Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. Mol Cell Proteomics, 2002, **1**(4): 323−333

[32] Kislinger T, Cox B, Kannan A, *et al*. Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. Cell, 2006, **125**(1): 173−186

[33] Farrah T, Deutsch E W, Omenn G S, *et al*. State of the human proteome in 2013 as viewed through PeptideAtlas: comparing the kidney, urine, and plasma proteomes for the biology- and disease-driven Human Proteome Project. J Proteome Res, 2014, **13**(1): 60−75

[34] Kim M S, Pinto S M, Getnet D, *et al*. A draft map of the human proteome. Nature, 2014, **509**(7502): 575−581

[35] Wilhelm M, Schlegl J, Hahne H, *et al*. Mass spectrometry-based draft of the human proteome. Nature, 2014, **509**(7502): 582−587

[36] Liu W, Wang J, Wang T, *et al*. Construction and analyses of human large-scale tissue specific networks. PLoS ONE, 2014, **9** (12): e115074

[37] Ezkurdia I, Vazquez J, Valencia A, *et al*. Analyzing the first drafts of the human proteome. J. Proteome Res, 2014, **13**(8): 3854−3855

[38] Zhang Y, Li D, Sun B. Do housekeeping genes exist? PLoS ONE, 2015, **10**(5): e0123691

[39] Jongeneel C V, Delorenzi M, Iseli C, *et al*. An atlas of human gene expression from massively parallel signature sequencing (MPSS). Genome Research, 2005, **15**(7): 1007−1014

[40] Reverter A, Ingham A, Dalrymple B P. Mining tissue specificity, gene connectivity and disease association to reveal a set of genes

that modify the action of disease causing genes. BioData Min, 2008, **1**(1): 8

[41] Zhu J, He F, Song S, *et al*. How many human genes can be defined as housekeeping with current expression data? BMC Genomics, 2008, **9**: 172

[42] Prieto C, Risuen˜o A, Fontanillo C, *et al*. Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. PLoS ONE, 2008, **3**(12): e3911

[43] Eller C D, Regelson M, Merriman B, *et al*. Repetitive sequence environment distinguishes housekeeping genes. Gene, 2007, **390**: 153-165

[44] De Ferrari L, Aitken S. Mining housekeeping genes with a Naive Bayes classifier. BMC Genomics, 2006, **7**: 277

[45] Eisenberg E, Levanon E Y. Human housekeeping genes are compact. Trends Genet, 2003, **19**(7): 362-365

[46] Eller C D, Regelson M, Merriman B, *et al*. Repetitive sequence environment distinguishes housekeeping genes. Gene, 2007, **390**(1-2): 153-165

[47] Lehner B, Fraser A G. Protein domains enriched in mammalian tissue specific or widely expressed genes. Trends Genet, 2004, **20**(10): 468-472

[48] Farre D, Bellora N, Mularoni L, *et al*. Housekeeping genes tend to show reduced upstream sequence conservation. Genome Biol, 2007, **8**(7): R140

[49] Williams T, Yon J, Huxley C, *et al*. The mouse surfeit locus contains a very tight cluster of four "housekeeping" genes that is conserved through evolution. Proc Natl Acad Sci USA, 1988, **85**(10): 3527-3530

[50] Zhang L, Li W H. Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol, 2004, **21** (2): 236-239

[51] Paik H, Ryu T, Heo H S, *et al*. Predicting tissue-specific expressions based on sequence Characteristics. BMB reports, 2011, **44** (4): 250-255

[52] Shaw G T, Shih E S, Chen C H, *et al*. Preservation of ranking order in the expression of human housekeeping genes. PLoS One, 2011, **6**(12): e29314

[53] She X, Rohl C A, Castle J C, *et al*. Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. BMC Genomics, 2009, **10**(1): 269

[54] Lin W, Liu W, Hwang M. Topological and organizational properties of the products of house-keeping and tissue-specific genes in protein-protein interaction networks. BMC Systems Biology, 2009, **3**: 32

[55] Winter E E, Goodstadt L, Ponting C P. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. Genome Res, 2004, **14**(1): 54-61

[56] Goh K I, Cusick M E, Valle D, *et al*. The human disease network. Proc Natl Acad Sci USA, 2007, **104**(21): 8685-8690

[57] Lage K, Hansen N T, Karlberg E O, *et al*. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proc Natl Acad Sci USA, 2008, **105** (52): 20870-20875

[58] Chen M, Xiao J, Zhang Z, *et al*. Identification of human HK genes and gene expression regulation study in cancer from transcriptomics data analysis. PLoS ONE, 2013, **8**(1): e54082

[59] Ganegoda G U, Wang J, Wu F, *et al*. Prediction of disease genes using tissue-specified gene-gene network. BMC Systems Biology, 2014, **8**(Suppl 3): S3

The Progress of Gene Tissue Specificity Researches*

LIU Wei**, SUN Zhi-Qiang, XIE Hong-Wei

(*Department of Automatic Control, College of Mechanical & Electronic Engineering and Automatization, National University of Defense Technology, Changsha* 410073, China)

Abstract Investigating gene tissue specificity is an important step to understand life process and tissue functions. Despite the long history of research about housekeeping genes and tissue specific genes, their definition and detection methods are various. Housekeeping genes and tissue specific genes can be defined from the aspect of tissue expression number and expression variation across tissues, respectively. In general, housekeeping genes are usually defined as those expressed in most tissues with stable expression levels, while tissue specific or tissue selective genes are defined as those predominantly expressed in one tissue or a few tissues. High-throughput technology, such as microarray, RNA-seq and mass spectrometry have become the main methods to detect the tissue specificity of genes. By comparing the experimental results of some typical researches, we found there are significant differences between different methods in their coverage and sensitivity. In these methods, RNA-seq was the most sensitive, which can detect the most number of housekeeping genes, while mass spectrometry can only detect less tissue specific genes, and the results from different microarray experiments were various. Despite different definition and technology can lead to different housekeeping and tissue specific gene datasets, these datasets have very consistent functions and characters. Housekeeping genes usually implement fundamental functions of tissues or cells, while tissue specific genes perform most specific functions of tissues. Meanwhile, the tissue specificity of genes has close relations with diseases. Compared to other genes, housekeeping genes tend to become cancer genes, while tissue specific genes are more like to become drug targets.

Key words housekeeping gene, tissue specific, disease

DOI: 10.16476/j.pibb.2015.0268

* This work was supported by grants from Science Research Project of National University of Defense Technology (JC15-03-01), International Cooperation Project(2014DFB30010) and The National Natural Science Foundation of China (31171266).

**Corresponding author.

Tel: 86-731-84573369, E-mail: angel_nudt@126.com

Received: November 12, 2015 Accepted: December 14, 2015