# 基于基因保守模型整体发现的基因共表达网络

2016-4-8

（10）小组成员及分工：马子冬：文献阅读、文献讲解
范英旭：文献阅读、PPT制作
范　宇：文献阅读、PPT制作

# 概述：

- 为了阐明在全基因组水平上的基因功能，我们通过了使用人类、果蝇、蠕虫和酵母四个物种的DNA芯片，指出这些物种之间存在共表达基因。我们在其中发现了多种共表达关系，而每种关系在整个进化中保守。这种保守性也意味着这些基因对的共表达赋予了一个选择的优势，故而这些基因是功能相关的。许多这些关系为核心的生物学功能，如细胞周期、 分泌和蛋白质表达中的新基因的参与提供了强有力的证据。我们证实了这些物种的基因之间存在了关联，指出了这些关联，并证实了这些基因细胞增殖功能，我们将这种关联组装成一个基因共表达网络，我们发现了在新进化后的和原始模块之间相互关联的动物特有基因的组分。

# 背景介绍：

•人类和几种模式生物的基因组序列已经建立了这些生物体 中的细胞、 发展和行为过程所需的基因几乎完全列表。下一个重大挑战是澄清目前未知基因组中基因职能是的大分数的功能，发现基因之间如何互作以执行特定生物学过程。基因芯片技术则为我们提供了在全基因组水平上揭示基因功能的一种方法。

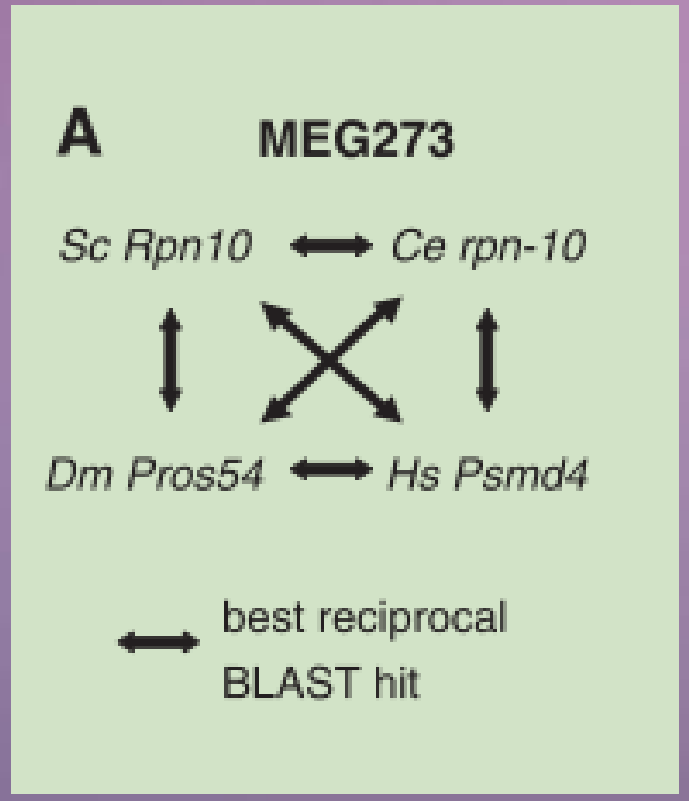
# 思路：

•基因表达谱分析所采用的常用方法是聚类，其目的就是将基因分组。组内基因的表达谱相似，它们可能有相似的功能。

•然而，产物有相同功能的编码基因不一定共享相似的转录模式。相反，有不同功能的基因可能因为巧合或随机扰动而有相似的表达谱。

•尽管有许多意外的情况存在，大量功能相关的基因的确在相关的一组条件下有非常相似的表达谱，特别是被共同的转录因子共调控的基因，或者产物构成同一个蛋白复合体，或者参与相同的调控路径。因此，在具体的应用中，可以根据对相似表达谱的基因进行聚类，从而指派未知基因的功能。

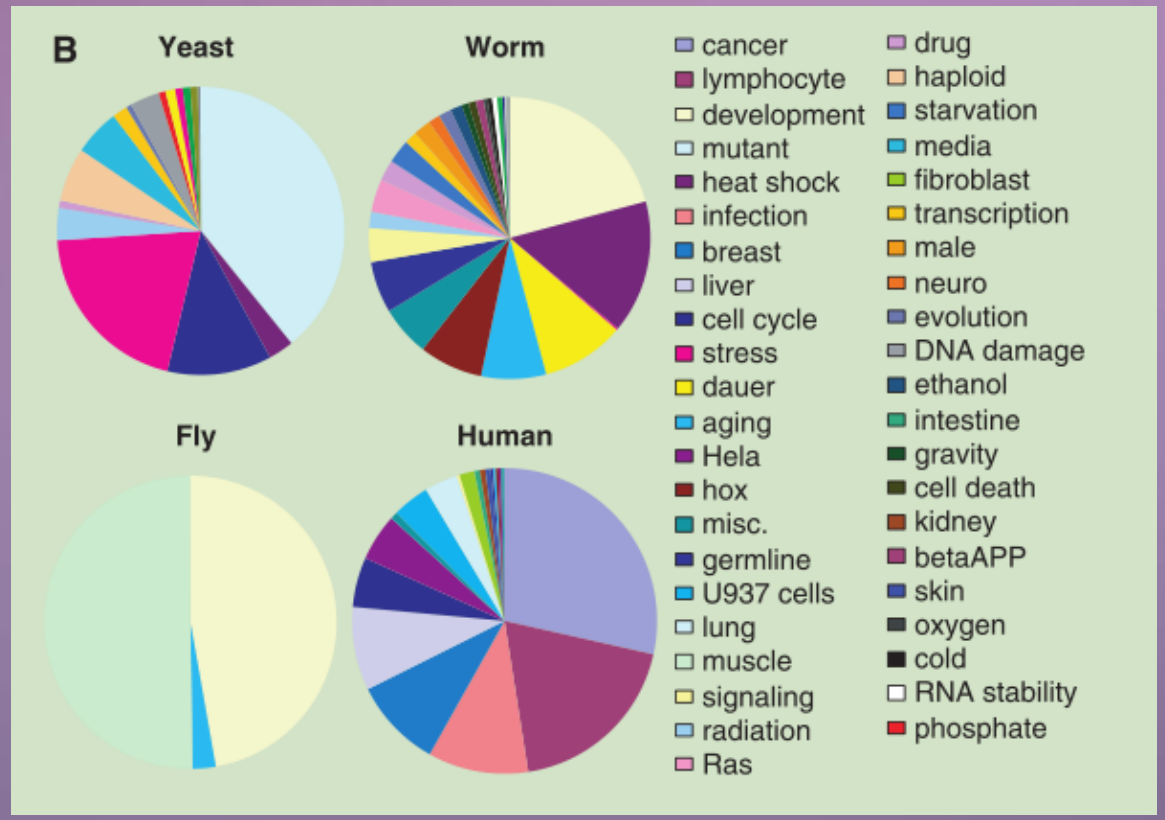
# 基因共表达网络化建设（Construction of a gene-coexpression network ）




•要确定多个生物的共表达基因，我们首先从一个生物体的基因在其他物种中相对应的直系同源基因。通过对各物种的蛋白质进行all-against-all BLAST，我们将BLAST后的得分最好的相对应的一系列基因整合，叫做metagene.

Fig. 1. (A) Example of a metagene (MEG273).An arrow points from gene X to gene Y if the protein sequence of Y had the most significant BLAST score to X's protein sequence when compared with all of the protein sequences in Y's database.

# 基因共表达网络化建设（Construction of a gene-coexpression network ）




通过不同物种，不同的实验条件（发展阶段、生长条件、压力、疾病、突变等）下的基因芯片的表达模式来指出一系列基因是功能相关的。

Fig1.(B)Compendiums of microarray expression data from four organisms included in the analysis. Color showsthe type of DNA microarray experiment.

# 基因共表达网络化建设（Construction of a gene-coexpression network ）

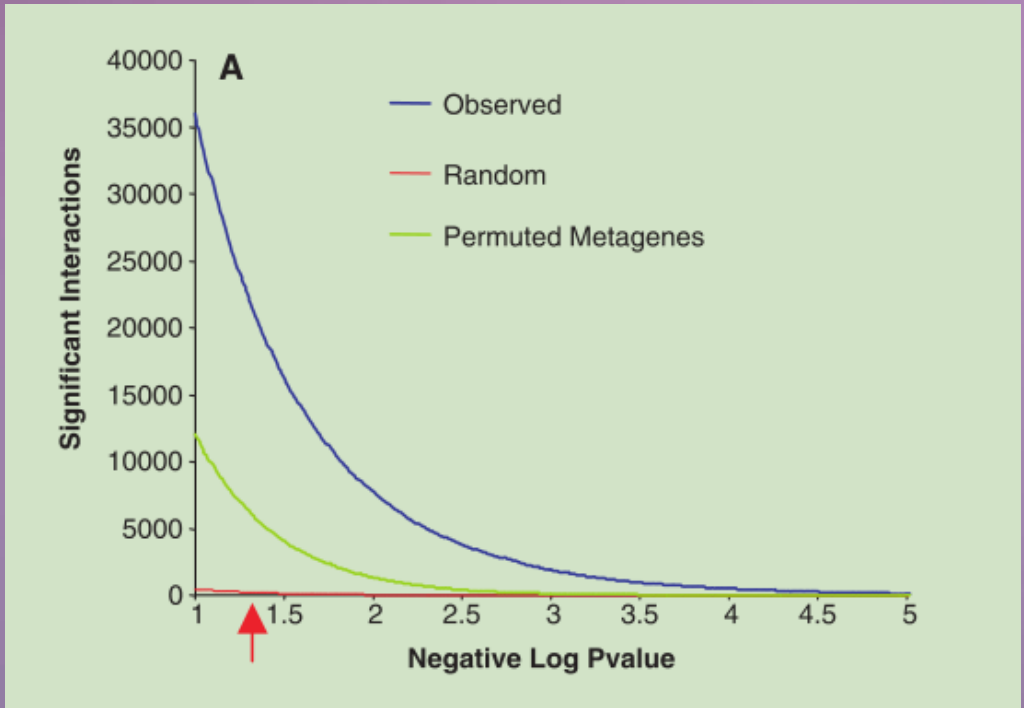
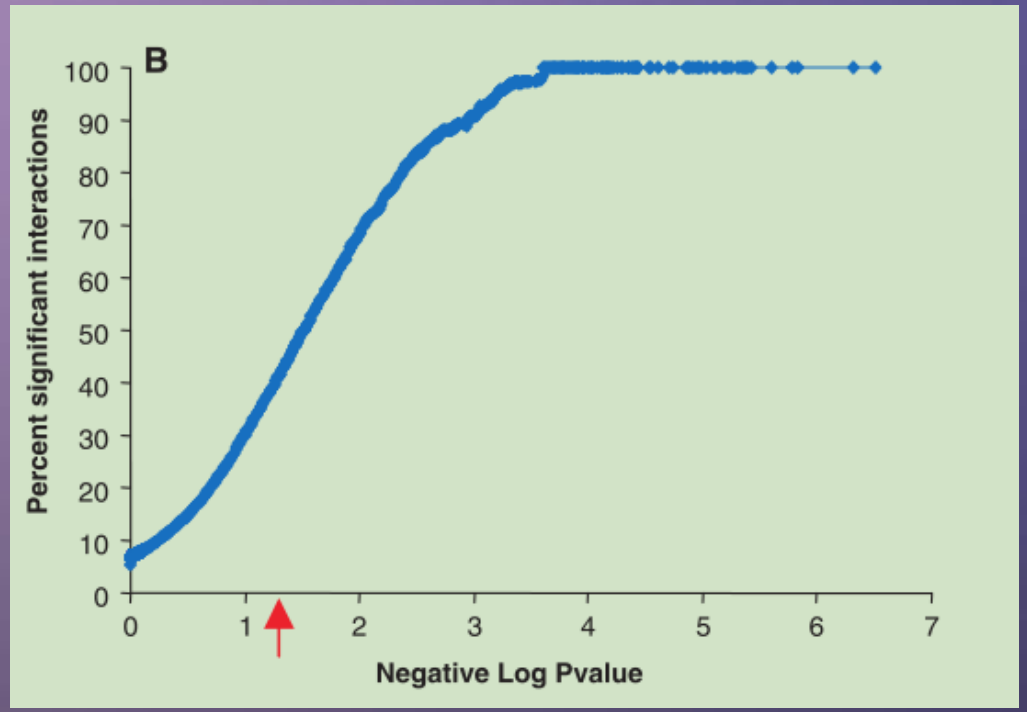



Fig.2(A) The number of metagene interactions ( y axis) exceeding a P-value cutoff (x axis) in networks constructed from real metagenes (blue curve), a random distribution (red curve), and randomly permuted meta- genes (green curve). P values are shown in log 10 scale. Red arrow marksP <0 .05, the cutoff used in the gene-coexpression network.

Fig.2(B) We randomly divided the databases of each species into two equally sized sets and then generated new networks derived from each half of the data for a series of P values. Shown is the percent of metagene pairs with P <p in the first half that have P < 0.05 in the second half, for each P value p. P values are shown in logarithmic scale.



# 基因共表达网络化建设（Construction of a gene-coexpression network ）

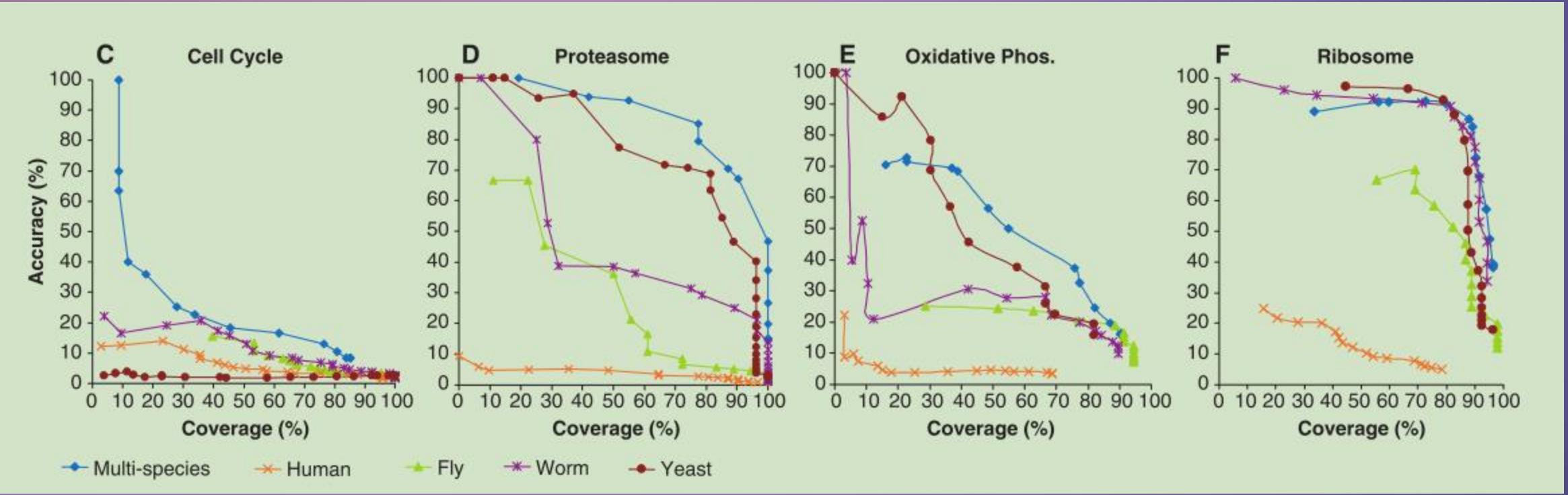



Fig.2(C to F) Comparison of multiple-species and single-species expression networks. We then calculated the percentage of links connecting two members of the category ( y axis; accuracy) and plotted this against the percentage of metagenes that are connected to at least one other metagene in the category (x axis; coverage).

# 保守基因表达链的生物学功能

# 基于多物种的基因共表达网络与单物种网络不同之处：

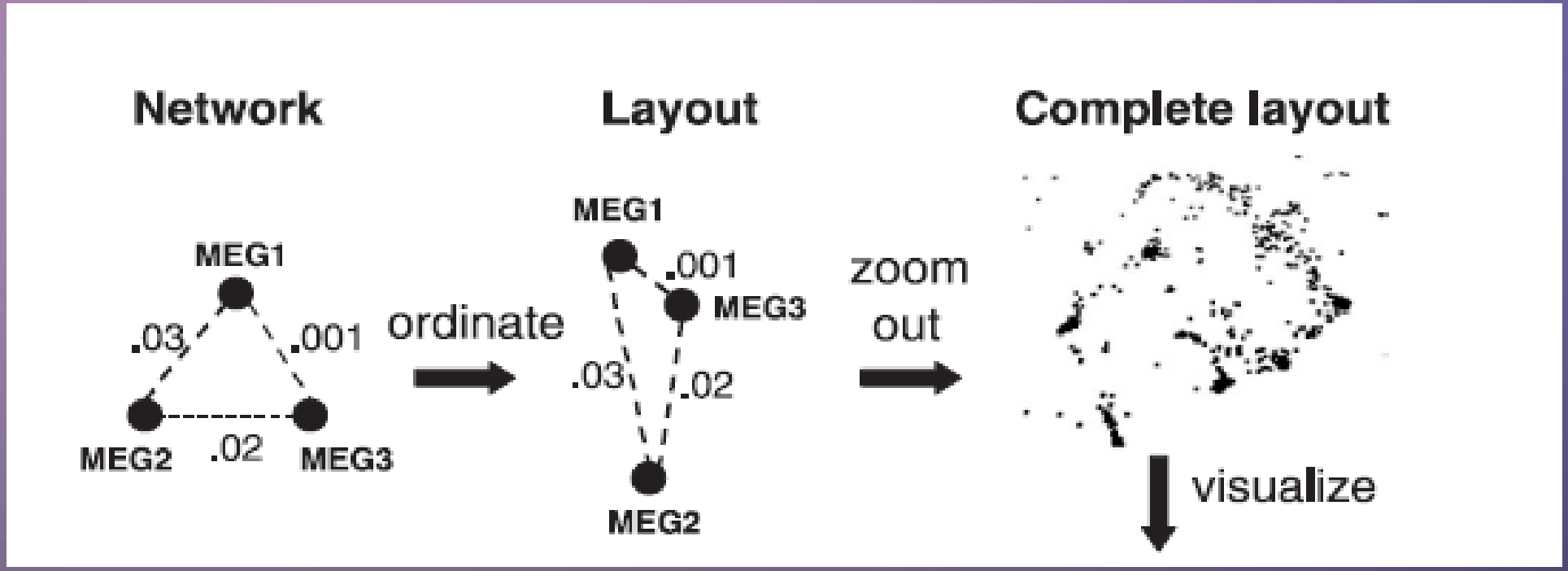
1.多物种网络只研究不同物种之间的同源基因，主要关注那些重要的，保守的生物学过程

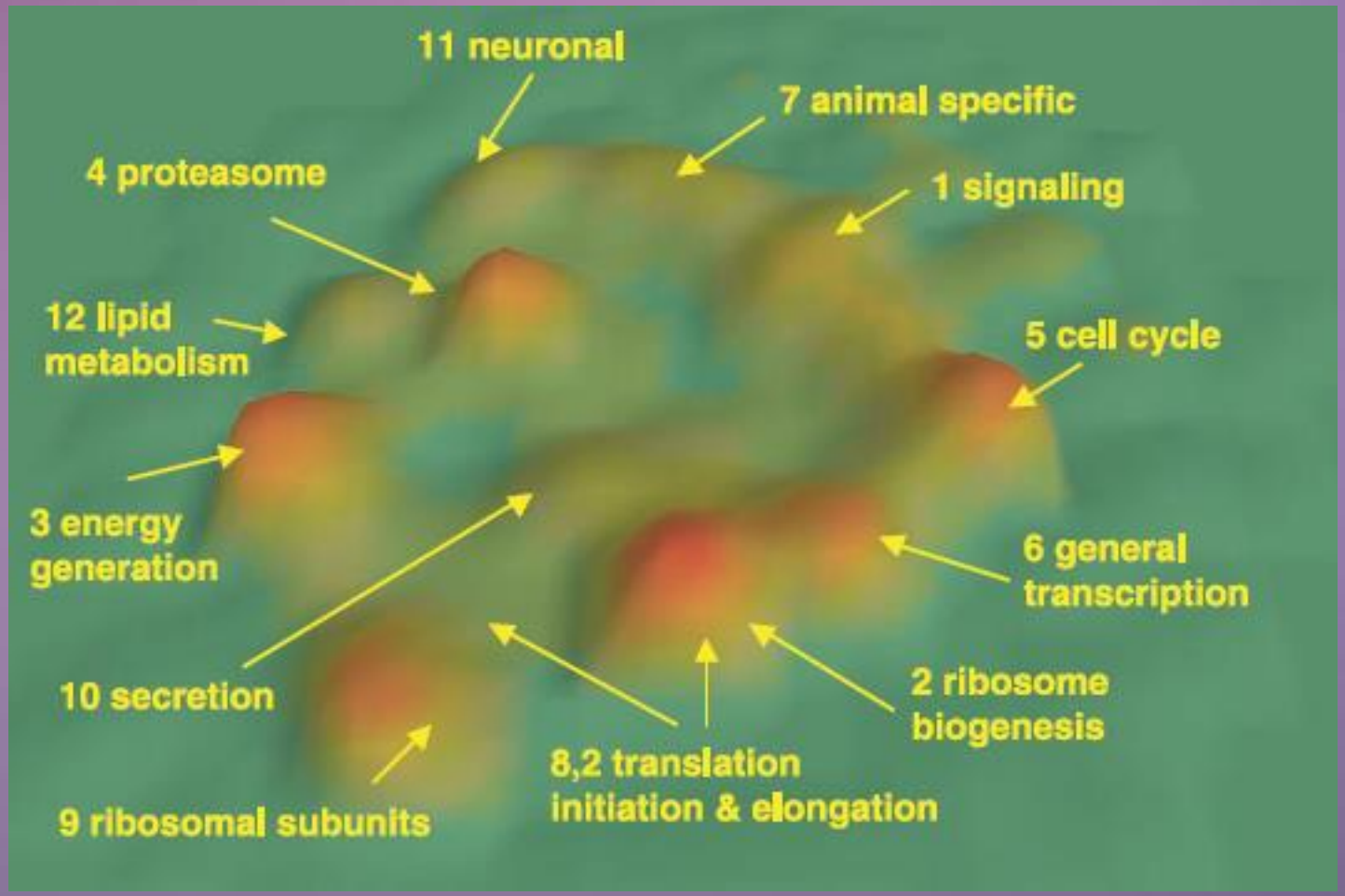
2.多物种网络中的互作是基于进化保守性的功能之间的关系，而单物种网络只反映了相关的基因表达。

# 可视化网络

通过可视化网络中的连接去了解物种间的进化保守的共表达的共同模式

选择3D布局模式，叫做VxInsight。根据P value负对数值大小，将megagene分布在XY平面，基因的密度作为Z轴。






通过K-means聚类将平面分为12个区域，称之为components，大部分组分内部都有相似的生物学功能，包括蛋白质的降解，核糖体的功能，细胞周期，代谢途径和神经元功能等。

**Table 1.** Network components.

| Component | Size* | Biological function† | Genes in component‡ | Enrichment; *P* value§ |
|---|---|---|---|---|
| 1 | 353 | Cellular cortex | 16/57 | $2.7; 10^{-6.1}$ |
| | | Signaling | 44/321 | $1.3; 10^{-5.8}$ |
| | | Animal-specific | 195/1441 | $1.3; 10^{-7.2}$ |
| 2 | 349 | Ribosome biogenesis | 102/125 | $8.0; 10^{-83}$ |
| 3 | 320 | Energy generation | 77/147 | $5.6; 10^{-42}$ |
| 4 | 271 | Proteasome | 31/32 | $12; 10^{-32}$ |
| 5 | 241 | Cell cycle | 110/202 | $7.7; 10^{-85}$ |
| 6 | 201 | General transcription | 47/142 | $5.6; 10^{-24}$ |
| 7 | 167 | Animal-specific | 124/1441 | $1.8; 10^{-17}$ |
| 8 | 156 | Translation initiation, elongation, and termination | 20/110 | $4.0; 10^{-7.3}$ |
| | | Aminoacyl transfer RNA biosynthesis | 14/31 | $9.9; 10^{-11}$ |
| 9 | 139 | Ribosomal protein subunits | 74/78 | $23; 10^{-107}$ |
| 10 | 92 | Secretion | 37/85 | $16; 10^{-38}$ |
| 11 | 65 | Neuronal | 17/42 | $21; 10^{-19}$ |
| | | Animal-specific | 58/1441 | $2.1; 10^{-15}$ |
| 12 | 57 | Lipid metabolism | 6/16 | $22; 10^{-7}$ |
| | | Peroxisome | 14/32 | $26; 10^{-17}$ |

*The total number of metagenes in the component. †Biological functions were based on edited terms from Gene Ontology (*15*) and the KEGG database (*22*). ‡The number of metagenes in the biological function group and in the component divided by the total number of metagenes in the biological function group that were also in the network. §The ratio between the number of observed metagenes in a category and the number expected by chance. The *P* value was computed as the probability of obtaining the observed number of overlaps by chance under a hypergeometric distribution.

组分5是富集相同生物学功能的集合基因的典型代表组分。

共有241个metagenes，202个参与生物学功能（参与细胞周期的功能），这个数量是利用超几何分布预测数量的7.7倍。

其中110个的功能以前就了解了。有30个调节细胞周期，例如MEG2742 (编码 cyclin E) and MEG5621(编码 Wee1)，80个终止细胞周期，例如MEG1092 (编码 DNA polymerase-_x0006_)。其余的131个基因功能是未知的，因此我们可以根据建立的共表达网络来推测这些基因可能参与细胞周期新的功能，或已知的功能。

# 预测基因功能

检测表达量

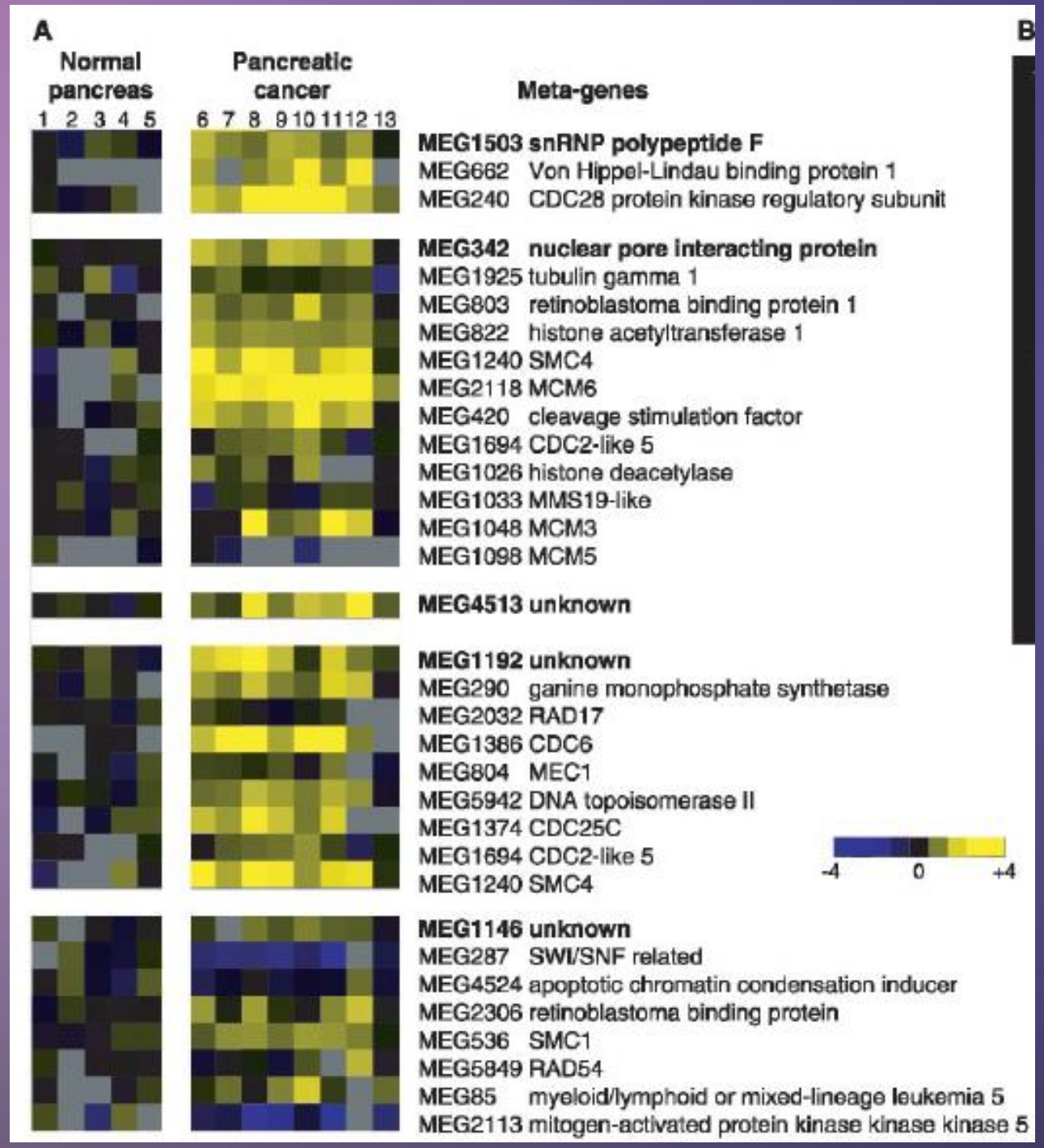
选择五个联系紧密的集合基因，分别是：

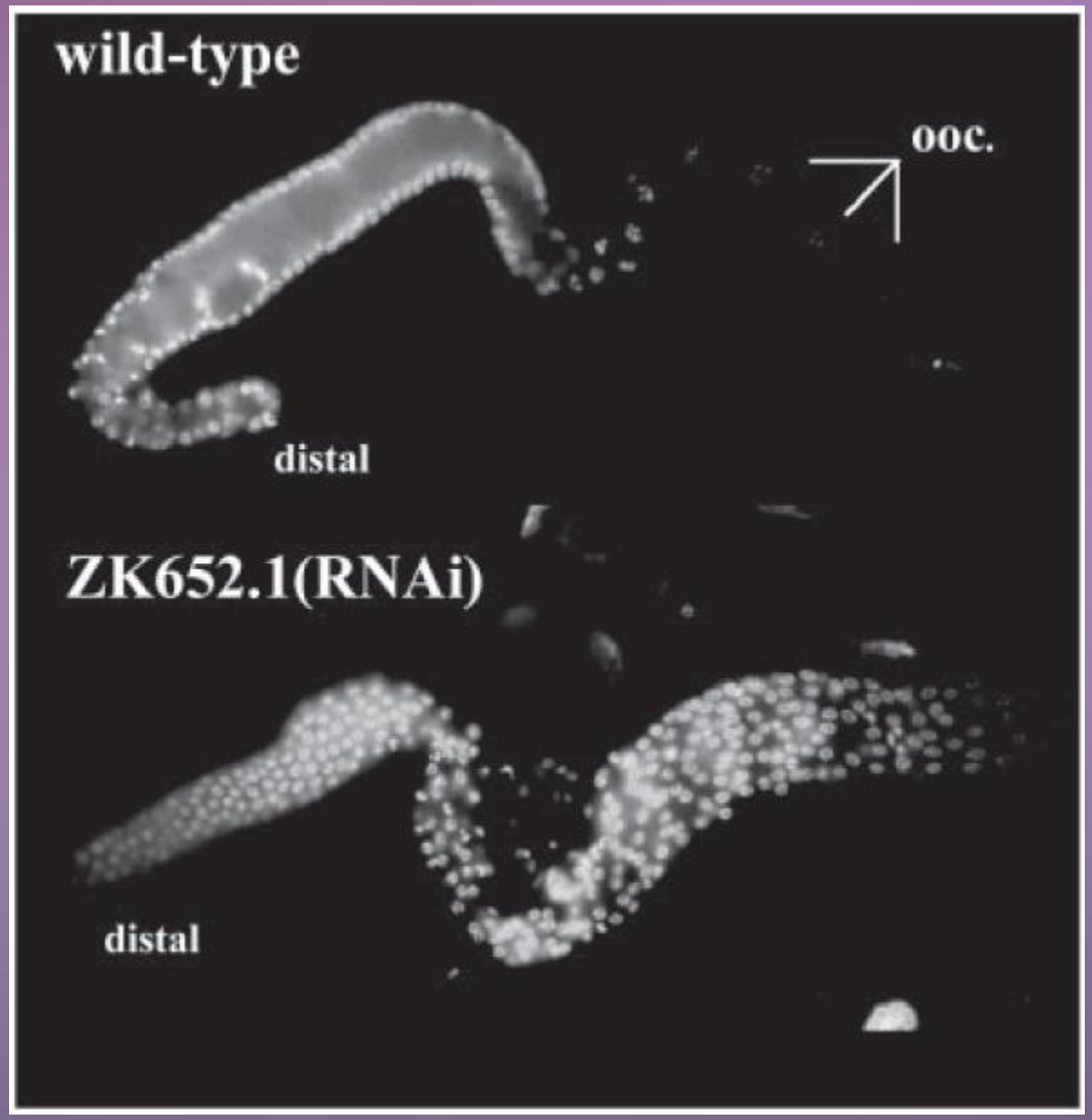
MEG1503 (which encodes an snRNP protein involved in splicing)

MEG342 (which encodes a nucleoporin-interacting component)

最后三个基因功能未知

MEG4513
MEG1192
MEG1146

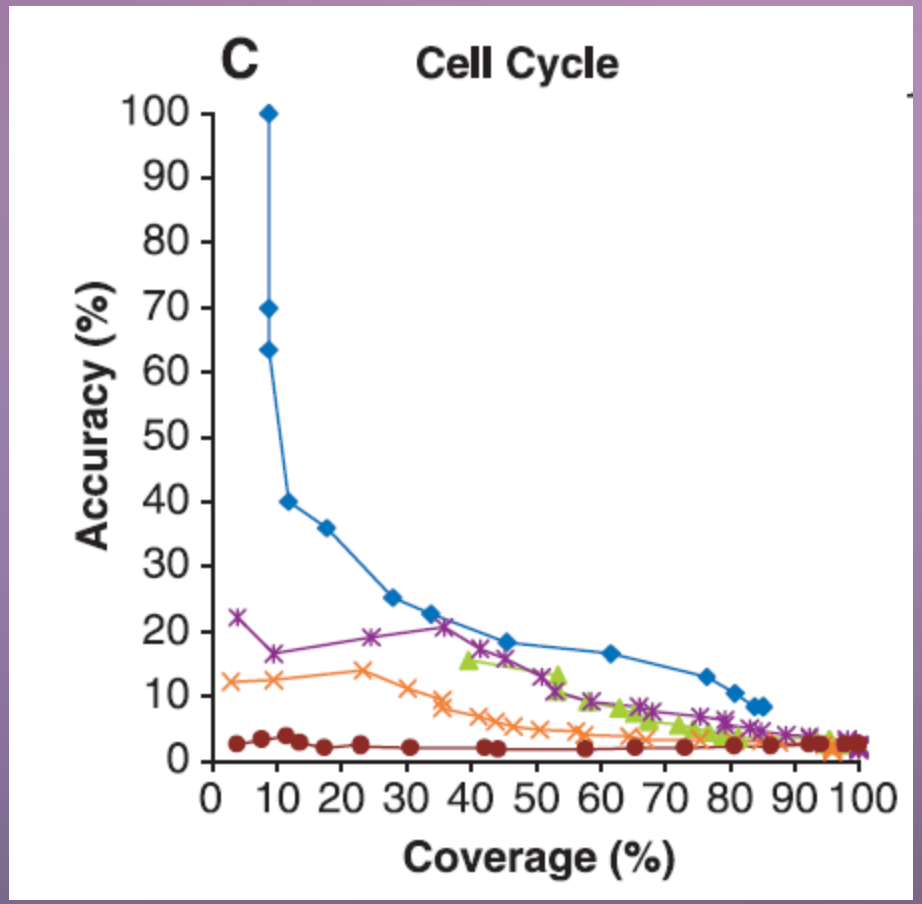





第二个实验是利用RNAi来做的。MEG1503包括线虫ZK652.1基因，通过加入双链ZK652.1 RNA，诱导突变得到突变体。

发现在突变体中，生殖腺中有大量的细胞核，表明了该基因的野生型功能是是抑制生殖细胞的增殖。

接下来对基于多物种数据的网络和单物种数据的网络进行更大范围的评估，评估的依据是利用那些已知功能的基因。
对于KEGG数据库中定义的功能，多物种构建的网络比单物种网络表现的更好。例如在细胞周期功能中，多物种网络更有优势。如下图：




对于其他的功能类，如核糖体的功能，使用单一物种的数据（如酵母）和使用多物种数据表现情况是类似的，两种网络差别不大

基于多物种的网络往往表现得更好，一个可能的原因是在建立多物种的网络时使用的DNA芯片数据更多，因此可能会更好地反映实际情况。

为了排除这种可能性，我们重新建立多物种网络，只使用了979个DNA芯片数据，这个数据量与单独的蠕虫数据建立的网络等同，发现使用较少数据的多物种网络在功能表现上（预测基因功能）与以前的网络表现没有多大差别。

因此我们认为多物种网络在功能预测上表现得更好的原因是该网络的建立是利用了生物进化的相关信息来过滤掉那些非功能相关的基因互作。

多物种网络包括570个集合基因，这些基因编码的蛋白质功能是未知的（是进化保守的同源基因，但在任何物种对其功能知之甚少），这些基因与其它基因一共有3943个交互，其中一些功能已知，因此未知的基因功能是有可能被发现的。

# 遗传模块的互作与保守性

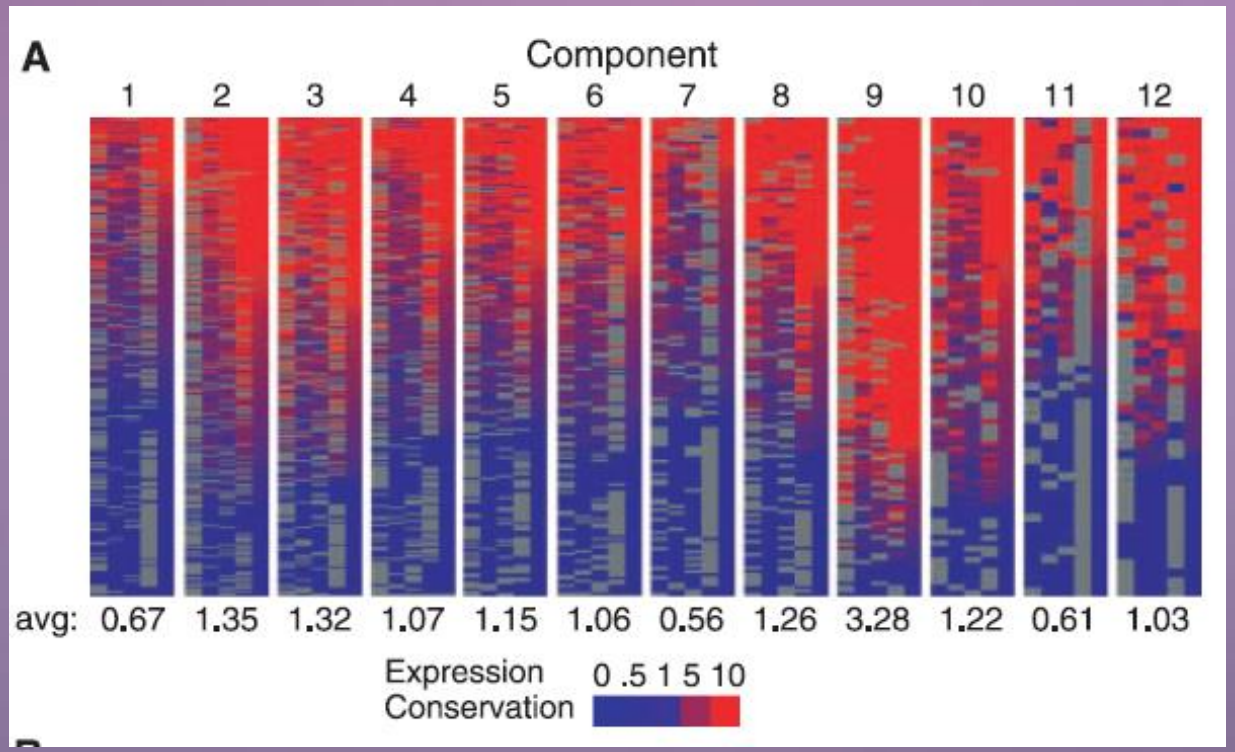
除了了解单个基因的功能，我们还需要将多个基因看做整体了解其功能。考虑三种类型的遗传模块：

1.原始的专用模块：有高度保守的编码区域，例如与核糖体有关的功能基因，从酵母到人类都是保守的；

2.进化中的模块；该模块基因功能在四个物种中有相对较大变化的，例如有神经元功能有关的基因；

3.可互换元件的模块；在不同物种中有不同的联系，如sir-2在酵母和人中是高度保守的，编码调控染色质结构和基因表达的基因，但在不同物种中有不同的下游靶位点。

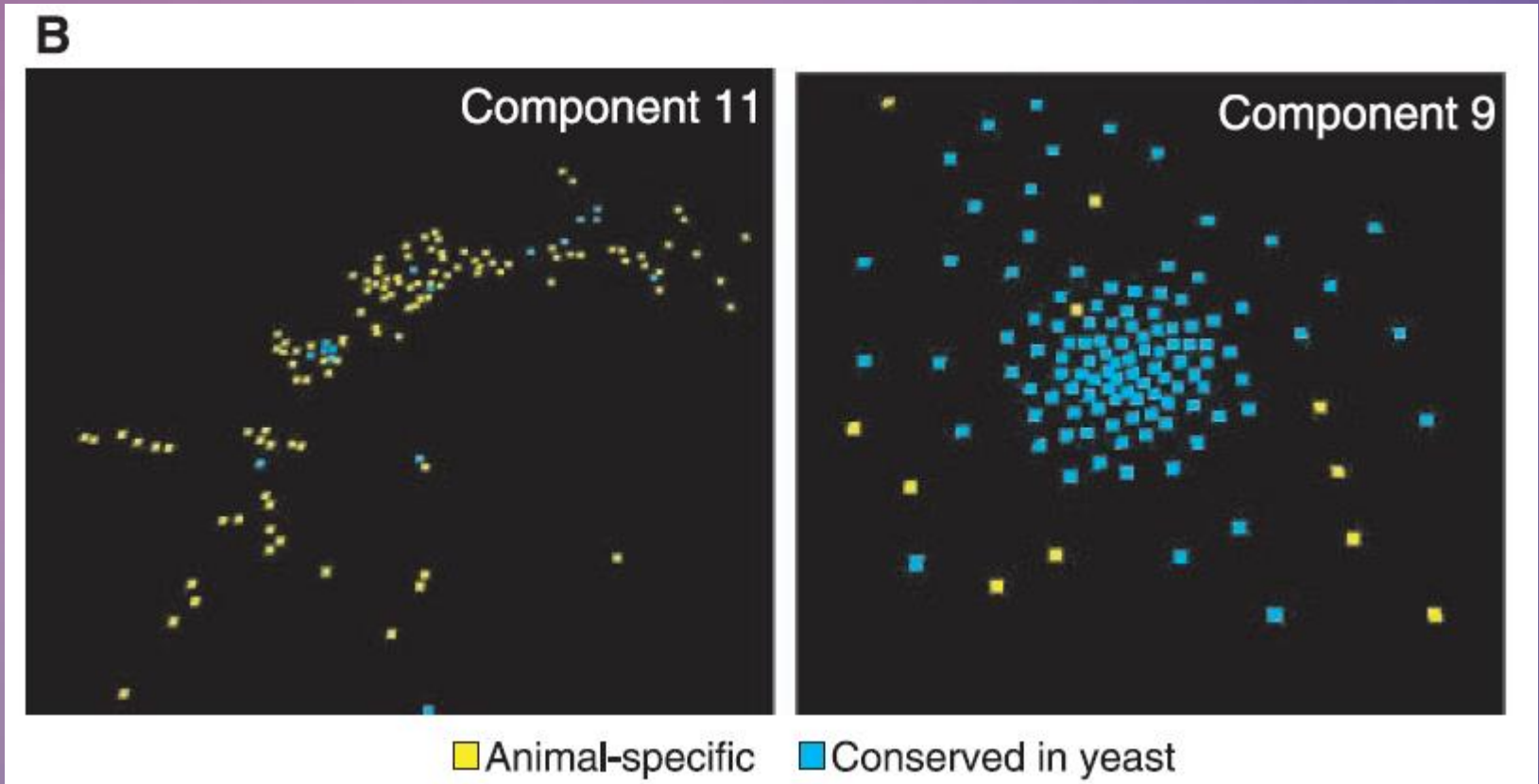
我们检测了不同模块在我们的基因共表达网络中的出现情况，首先将集合基因分为两部分，一部分是酵母基因，其余的为动物特有的基因，然后通过表达保守性指数（ECI）来衡量每个基因的保守性（数值越大，保守性越强）。




在组分1,7,11中，动物特有基因富集度最高，同时也说明他它们进化保守程度最低（得分最低）。

组分1主要是信号功能途径，属于动物特有的，在不同物种中调控不同的下游基因，预测的与实际相吻合。

组分7还没有被关联到任何一个生物学功能，但编码区域和基因互作具有很低的保守性，表明可能与进化过程有关。




与组分1,7,11相比，组分9含有动物特有基因最少，进化上具有较高的保守性，与实际情况相符合，因为该组分功能与核糖体有关。

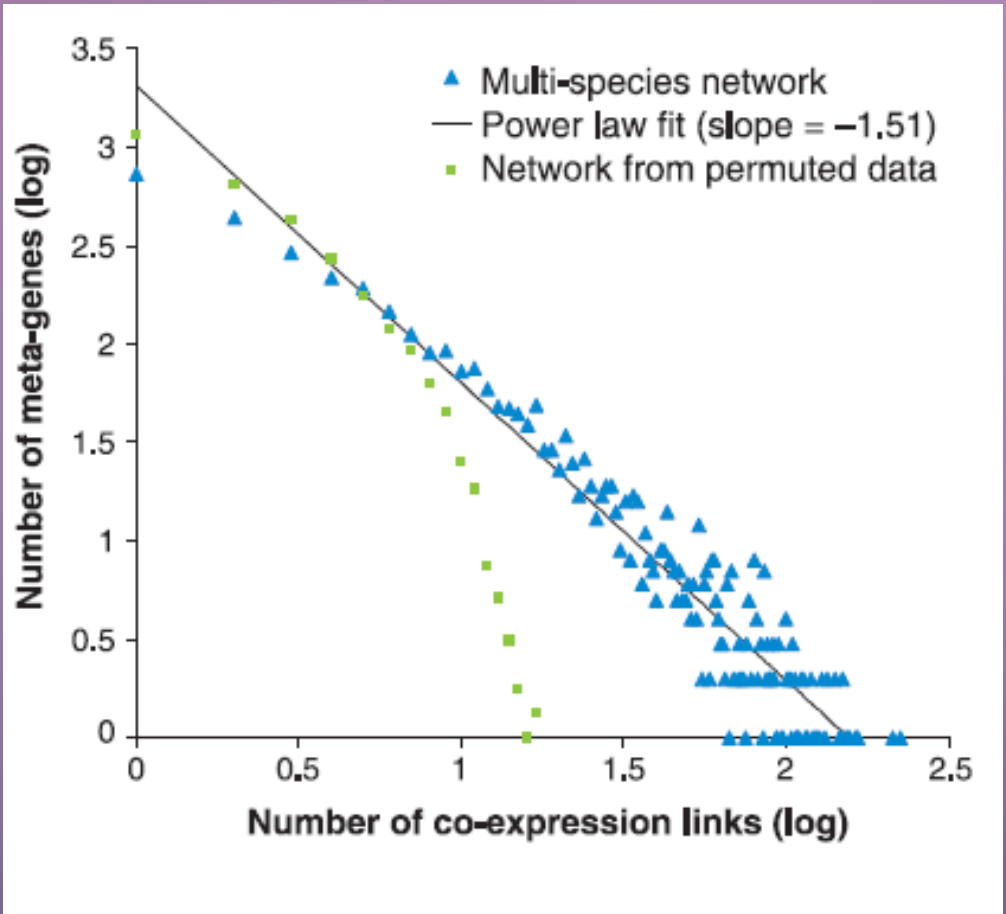
通过研究网络的连接情况（基于多细胞参与的过程以及重要的细胞生物学过程），发现许多过程是交织在一起的。

如组分3参与代谢途径，与糖酵解和三羧酸循环有关，76个是动物特有的基因，保守性较低，含有48个动物特有基因，4个基因与肌肉功能有关，消耗大量能量。

组分4主要是参与蛋白降解，含有蛋白酶体，泛素连接酶基因，有92个动物特有基因，其中有3个参与细胞凋亡，表明核心的细胞生物学过程蛋白降解与细胞程序性死亡有一定的功能联系。

# 遗传网络的连通性

一些生物学功能由许多个基因协调作用，为刻画基因共表达网络的连通性，我们计算了每个集合基因的邻居节点个数，并与置换数据的共表达网络进行了比较。我们发现基因表达链的分布是非随机的，显著性的包含更多的集合基因。




网络的连接度符合幂律分布。幂律分布函数在其它类似自然的，社会的现象中也都存在，如美国企业规模分布，万维网的分布。在一些生物学网络如蛋白互作网络中也都存在。

这一结果表明为保持一个高度相连的基因的遗传途径在整体设计中的存在一种选择性力量。

Thank you!