Table of Contents

[dNet方法解读与使用总结 1](#_Toc500344302)

[方法文献概要 1](#_Toc500344303)

[Dnet pipeline流程 2](#_Toc500344304)

[1. Cox regression生存相关性分析 ( ‘survival’包) 4](#_Toc500344305)

[2. 构建core-net: 病人-生存期相关基因互作网络图 ( ‘dnet’包; ‘igraph’包) 4](#_Toc500344306)

[(1) Robustness(via data removal) 4](#_Toc500344307)

[(2) Significance( via randomization) 4](#_Toc500344308)

[(3) Core-net内部Communities构成分析 5](#_Toc500344309)

[3. 基于core-net概览不同癌种之间的关系 (Tumour type landscape, ‘supraHex’包) 7](#_Toc500344310)

[4. dnet包‘dGSEA’函数的功能: 不同core-net生存基因集预测效果评价、瘤间相似性和瘤内异质性分析 10](#_Toc500344311)

[(1) 比较不同方法构建的core-net生存基因集的预测表现 10](#_Toc500344312)

[(2) 肿瘤相似性和异质性研究(Tumour similarity & hetetogeneity) 12](#_Toc500344313)

[(3) dnet包‘dGSEA’函数的应用总结和使用方法-GSEA分析 13](#_Toc500344314)

[(4) dnet包‘dGSEA’函数的应用参考文献 14](#_Toc500344315)

[5. dnet包‘dEnricher’函数的功能：各种ontologies富集分析、基因进化溯源分析 14](#_Toc500344316)

[(1) core-net生存基因集生存期预测潜力以及药物相关性预测潜力 14](#_Toc500344317)

[(2) core-net基因集进化相关性分析 16](#_Toc500344318)

[(3) Other ontologies 注释 18](#_Toc500344319)

[(4) dNet包dEnricher函数的使用方法 19](#_Toc500344320)

[(5) Dnet包dEnricher函数的应用参考文献 20](#_Toc500344321)

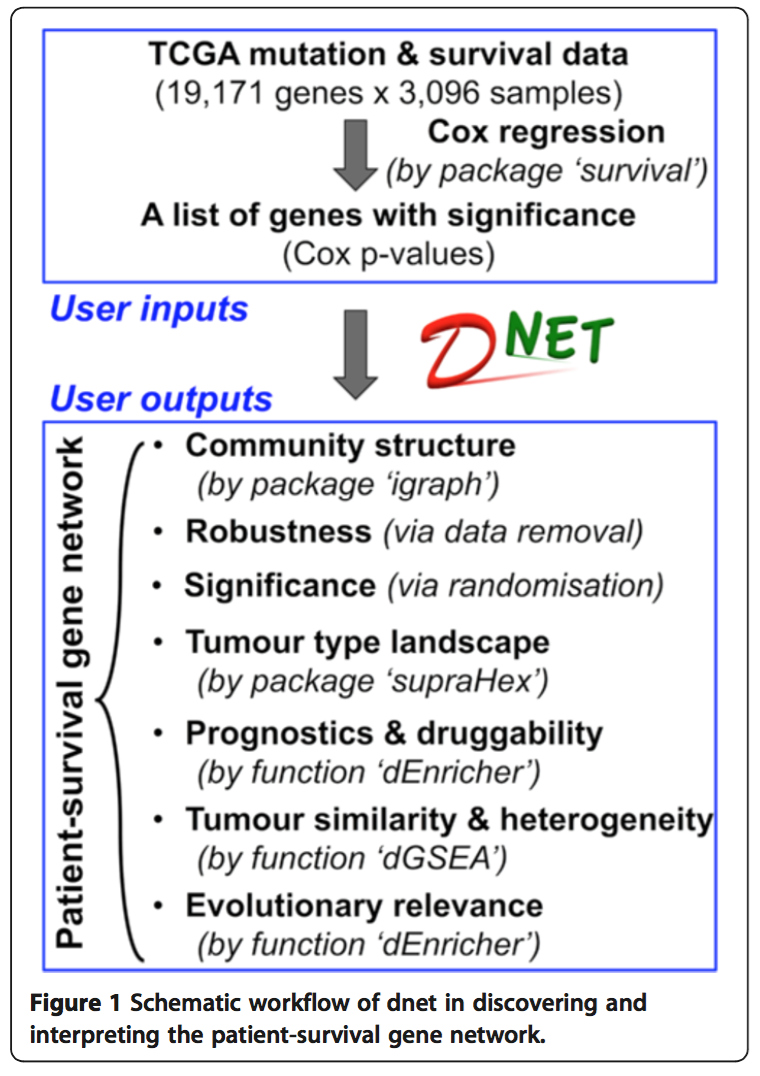
[6. Dnet构建core-net基因互作网络核心代码解读 20](#_Toc500344322)

# **dNet方法解读与使用总结**

## **方法文献概要**

基于TCGA数据库12个癌种(共3,096个癌症病人)的19,171个基因的体细胞突变信息，结合每个病人的生存期数据，发现了一个能至少部分程度上控制癌症病人生存期的基因互作网络（扣除肿瘤origin和肿瘤类型的影响后，该互作网络基因集的突变数目与病人生存期显著关联），该网络包含的基因集(以下称core-net生存基因集)有临床靶向药物治疗的潜力，通过对该网络的基因集的溯源研究，发现了起源于后口动物(Deuterostomia)的两个影响癌症病人生存期的重要基因: PTK2 and VAV1，这两个基因之前研究的较少。

## **Dnet pipeline流程**



1. 导入输入数据集矩阵：来自12个癌种的19,171个基因在3,096个癌症病人样本中的体细胞突变数量分布矩阵、3,096个癌症病人样本的临床表型信息（包括生存期信息、肿瘤原发位置信息、肿瘤类型信息等），该数据已经被整合到TCGA\_mutations数据集中，通过library("Biobase")导入。
2. 通过Cox regression生存相关性分析 ( ‘survival’包)对19,171个基因的生存期相关性进行预测，得到每个基因的Cox HR值(生存期相关性)和Cox P values(生存期相关性显著性)。
3. 基于每个基因的基因名称和生存相关显著性P值(Cox P values)，从已有蛋白互作网络(STRING)中获得生存期相关基因互作网络图(dNet包的dNetInduce函数和dNetPipeline函数)，即core-net生存基因集，并对core-net基因集进行了robust分析(via data removal)、significance分析(via randomisation)以及基因集内部的communities结构分析(‘igraph’包) 。
4. 基于core-net生存基因集对12个癌种之间的关系进行了概览，即根据core-net网络图的连接关系构建每条edge在不同类型肿瘤中的信息矩阵(edges × tumour)，通过dnet包的‘dNetReorder’函数(扩展自‘supraHex’包)将不同类型肿瘤之间的关系自组织到一张2D图中进行展示。
5. 对core-net生存基因集的生存期预测效力和药物基因富集情况(dnet包的‘dEnricher’函数：DGIdb ontology富集分析)进行了分析。
6. 对core-net生存基因集进行了多癌种突变频率高均一性基因富集程度、单癌种高突变频率基因富集、病人个体高突变基因富集程度的分析(dnet包的dGSEA, Gene Set Enrichment Analysis)，从生存期的角度揭示了肿瘤之间的相似性和肿瘤内部的异质性。
7. 对core-net生存基因集进行了进化溯源分析(dnet包的‘dEnricher’函数)。

## **Cox regression生存相关性分析 ( ‘survival’包)**

输入数据集：

1. 基因somatic突变数量矩阵：19171 genes X 3096 samples；各基因在各病人样本中发生somatic突变的数目矩阵。
2. 样品表型信息矩阵：包括每个病人的临床指标/生存期信息，如"time" (i.e. survival time in days), "status" (i.e., survival status: 0=alive; 1=dead), "Age" (the patient age in years), "Gender" (the patient gender: male/female), "TCGA\_tumor\_type", "Tumor\_stage", "Tumor\_grade等等。

分析方法：构建3个临床共变量(年龄、性别、肿瘤类型)的基线回归，构建包含全部四个解释变量的全回归(年龄、性别、肿瘤类型、基因突变数量)，通过LRT(Likelihood ratio tests, 与the sequential ANOVA同义)对全回归和基线回归进行比较，计算得到每个基因的生存期相关性（Cox HR值）、生存期相关性的显著性（Cox p-values）。Cox HR 和 P value可以作为生存期预测的指标，用来衡量每个基因的突变状态与病人生存优势的关系程度（校正每个样本的年龄、性别、肿瘤类型后）。该步得到的基因列表（包含基因名称和基因的Cox p-values）可作为下一步core-net构建的输入。

## **构建core-net: 病人-生存期相关基因互作网络图 ( ‘dnet’包; ‘igraph’包)**

基于已有的基因列表（基因名称＋基因显著性P值，本文为Cox p-values），从已有蛋白互作网络(STRING)获取到该列表基因的网络图，设定一个P值作为阈值，用dnet pipeline从该网络图中寻找到最大得分的subgraph，即得到病人-生存期相关基因互作网络图(以下称core-net)。

对core-net的Robustness和Significance进行了验证, 对core-net内部的communities构成进行了分析。

### Robustness(via data removal)

通过一次移除一个癌种的数据来重新构建core-net，计算原core-net中每条edge的置信值(confidence score)，即bootstrap值，即衡量移除每个癌种的数据后，每条edge重新出现在core-net中的几率。

### Significance( via randomization)

保留core-net节点的连接关系，对节点(基因)的label进行randomisation，以这种方式构建100个与原core-net同等大小的新core-net作为背景，估计原core-net每条边的显著性，通过dnet包的‘dPvalAggregate’函数可以将100个P值整合为一个P值。

### Core-net内部Communities构成分析

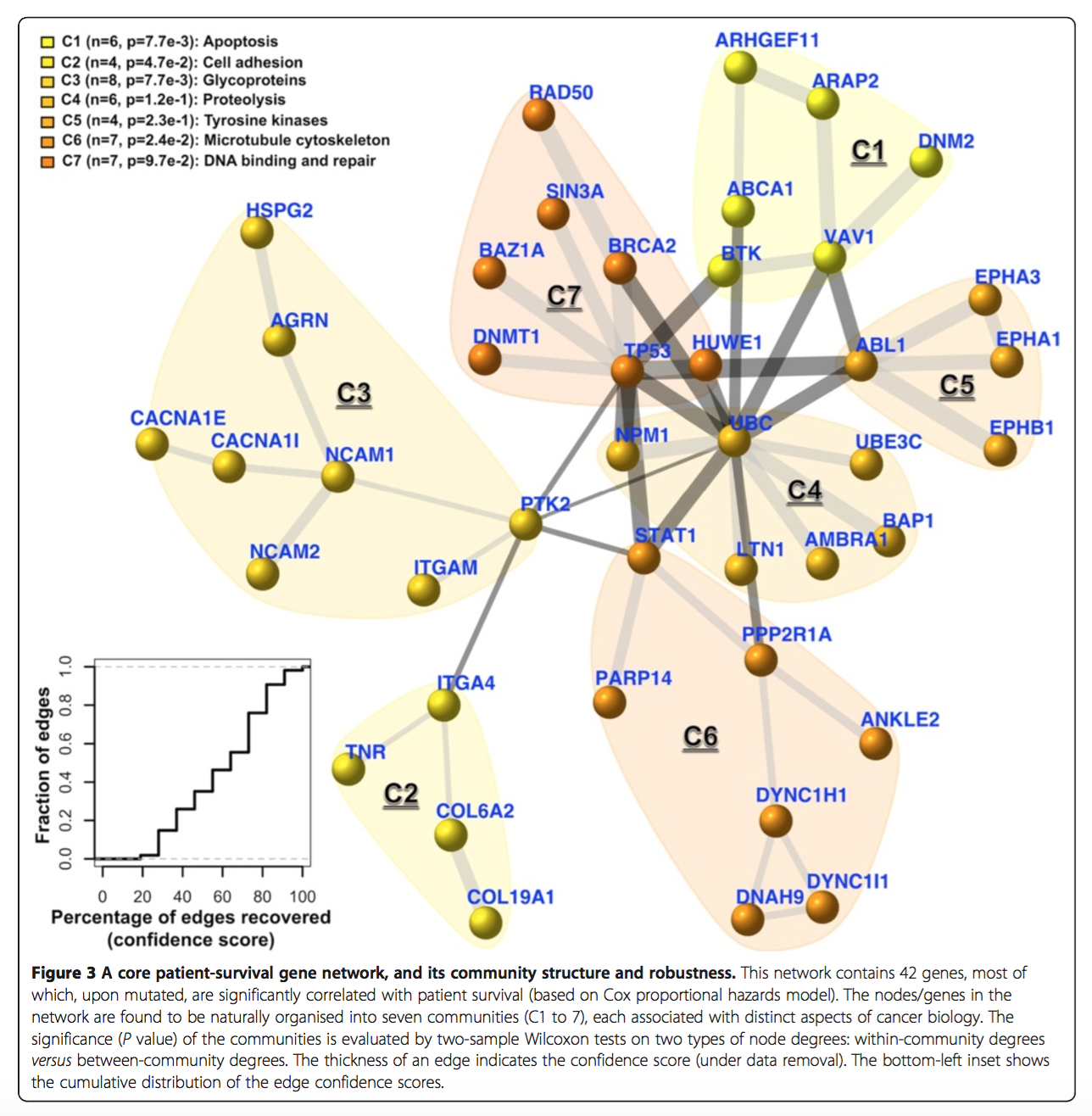
为了理解core-net的内部结构，通过‘igraph’包的spin-glass模型和simulated annealing功能进一步发现net-work的communities构成。通过两样本Wilcoxon 检验对Communities内部的节点和不同Communities之间的节点进行显著性P值检验。

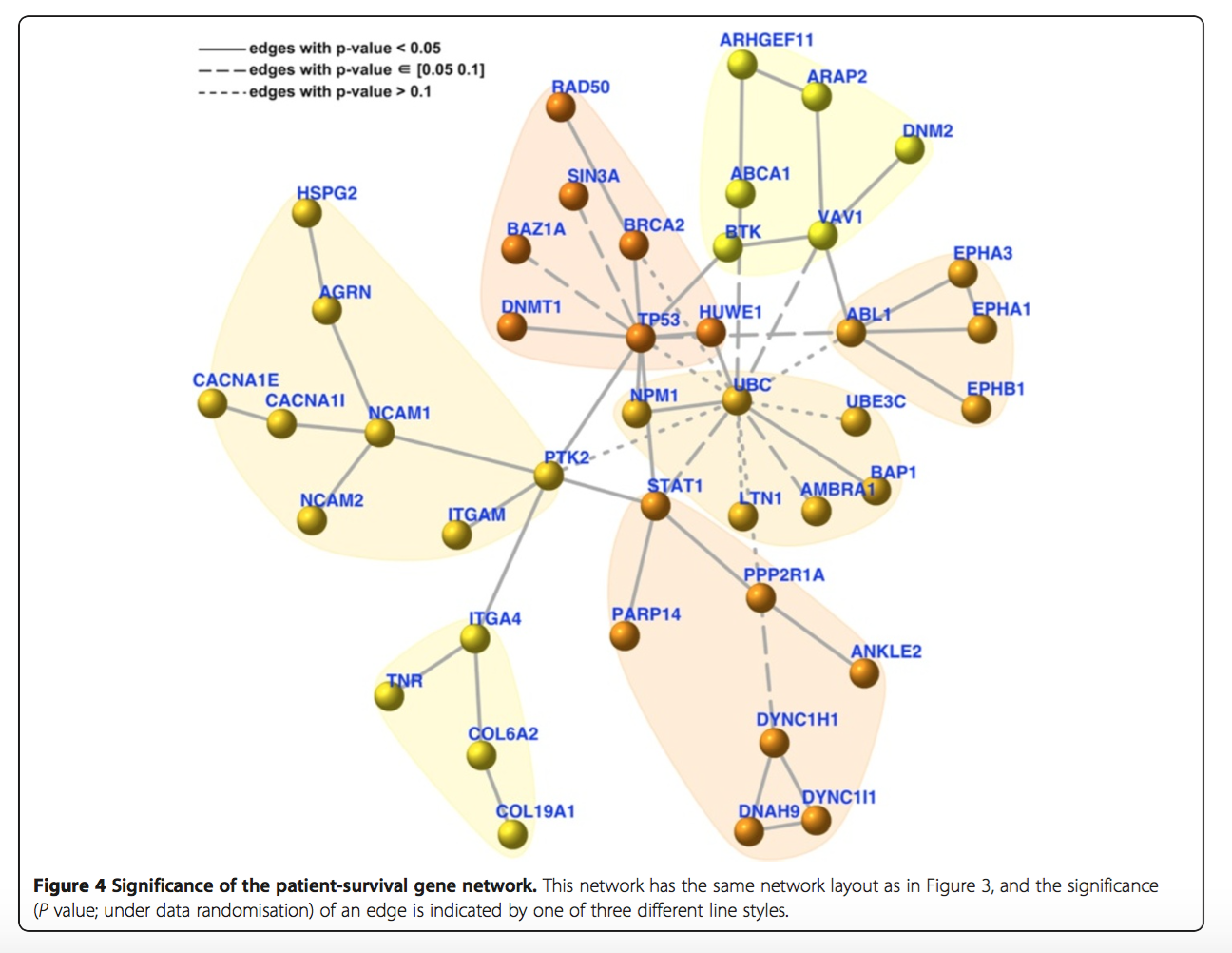
以上分析结论如下图Figure 3所示：

通过 Robustness分析，验证了每条edge被重新恢复的可能性很高，尤其那些属于核心communities(C7,C1,C5,C4)的edges(Figure 3A)；

通过Significance分析，验证了大部分edge的出现都是真阳性，即该edge在core-net中的出现不是随机的(Figure 3B)；

通过对core-net的communities构成进行分析，在没有先验知识的前提下，发现了7个明显代表不同生物学过程的communities(C1-C7)：the apoptosis community (C1), the cell adhesion community (C2), the community containing genes all encoding glycoproteins (C3), the proteolysis community (C4), the tyrosine kinase community (C5), the microtubule cytoskeleton community (C6), the DNA binding and repair community (C7)。其中C7的TP53基因直接与5个communities(C1、C5、C4、C6、C3)相连，TP53基因通过C3的PTK2基因与C2间接相连(像PTK2这种连接两个communities的中间基因，往往与病人生存期关联不大)。

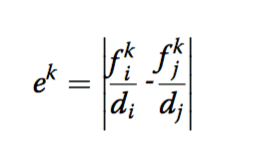




## **基于core-net概览不同癌种之间的关系 (Tumour type landscape, ‘supraHex’包)**

基于core-net的基因集以及这些基因在不同类型肿瘤样本中的突变频率来研究不同类型肿瘤之间的关系。

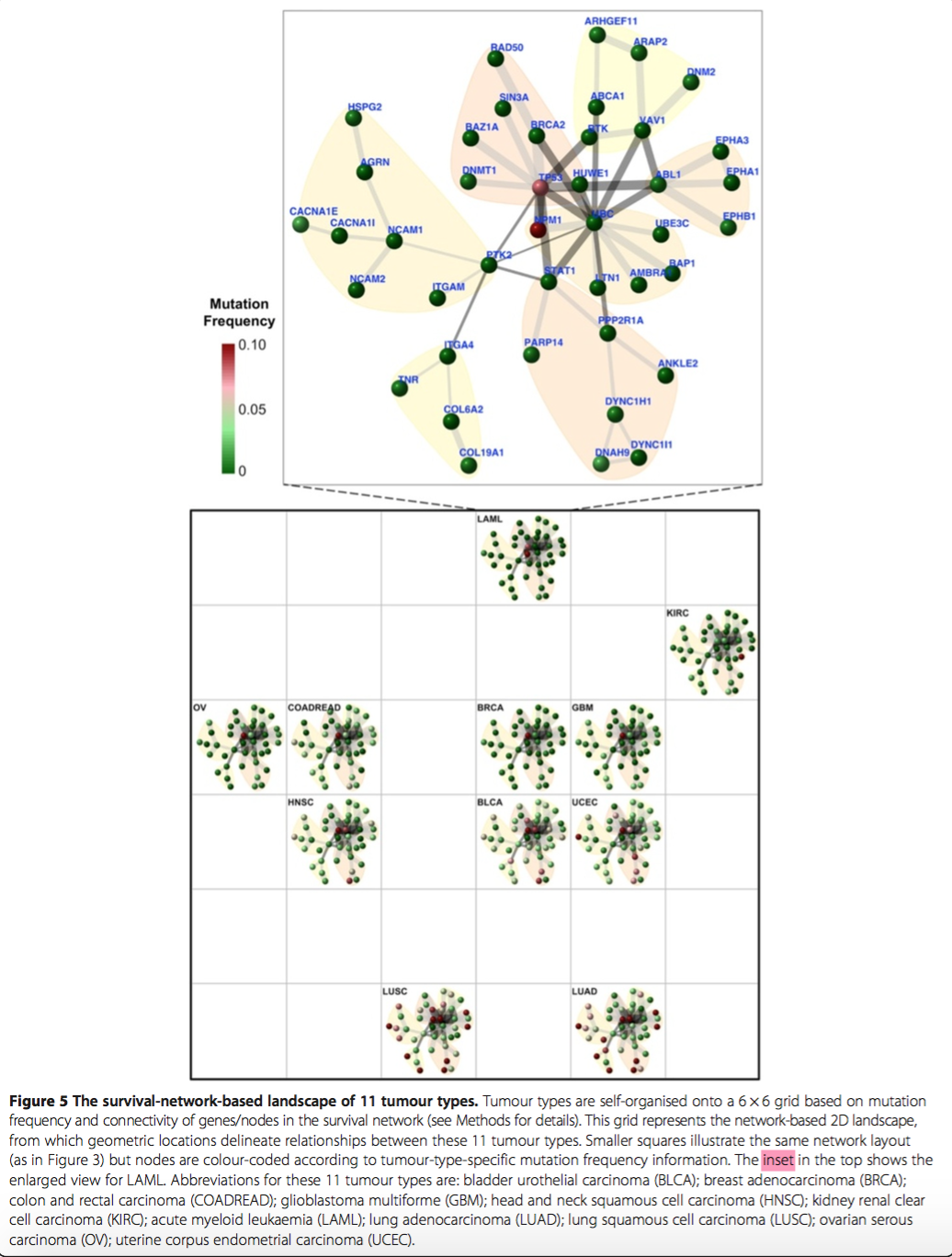
Dnet提供了两种研究方法：1. 通过构建基因在不同类型肿瘤中的发生频率矩阵(genes × tumour)，构建nj树来查看不同类型肿瘤的关系；2.利用core-net网络图的连接关系：构建每条edge在不同类型肿瘤中的信息矩阵(edges × tumour)，通过dnet包的‘dNetReorder’函数(扩展自‘supraHex’包)将不同类型肿瘤之间的关系自组织到一张2D图中进行展示。



公式说明：对于每条edge的信息 ek， di 和dj 分别代表该edge两端节点i、j的连接度，f ki 和 f kj分别代表i、j两节点(基因)在癌种k中的突变频率。ek对网络图的信息进行了充分利用：综合考虑了网络图中每个节点(基因)的突变频率以及节点的连接度。

基于core-net揭示不同类型肿瘤的关系如下图Figure 5所示：

该图不仅直观展示了不同类型肿瘤之间的关系，并且对其背后的原因(该类型肿瘤的生存基因突变频率)进行了直接揭示。例如LUAD和LUSC两种源于肺组织的癌症由于包含最高比例的高频突变生存基因，在图中的位置并列排布到了最边缘。而LAML由于包含独有的高频突变NPM1和TP53基因，使得其在图中的位置排布明显偏离于其他癌种(Figure 5上方对LAML单独进行了放大展示)。

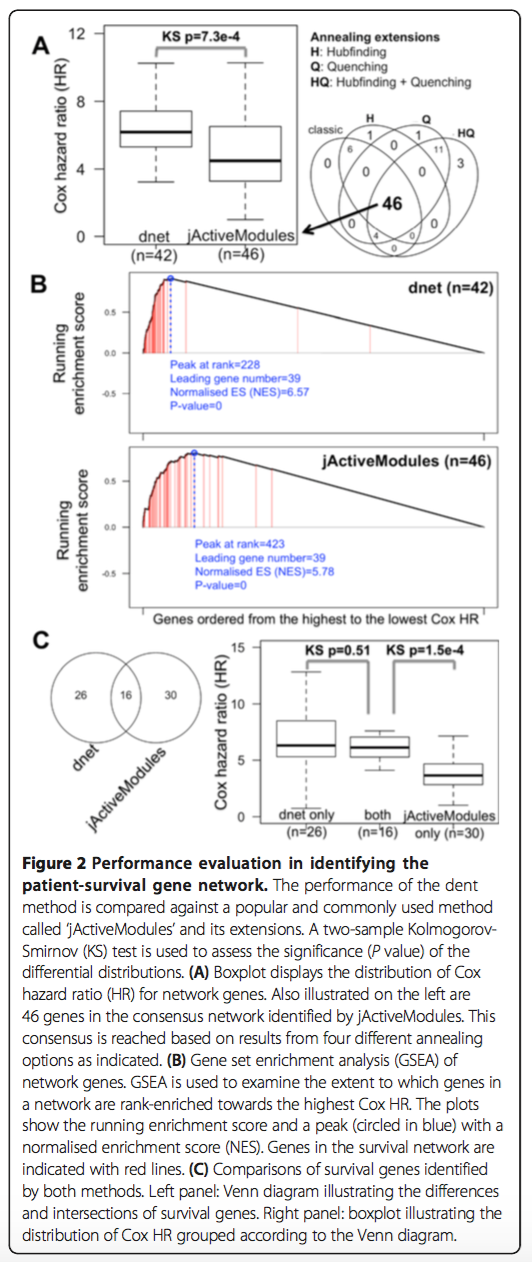


## **dnet包‘dGSEA’函数的功能: 不同core-net生存基因集预测效果评价、瘤间相似性和瘤内异质性分析**

### 比较不同方法构建的core-net生存基因集的预测表现

dnet 和 jActiveModules (Cytoscape的插件)都可以构建core-net。评估不同core-net基因集预测生存期的表现等同于评估高Cox HR值(或者低 Cox P values)基因在core-net中富集的趋势。通常有两种策略：1. 通过Kolmogorov-Smirnov (KS, KS P value) 检验不同Core-net基因集Cox HR值的分布差异(Figure 2A)；2. 通过dnet包的’GSEA’函数计算core-net基因集富集高Cox HR值基因的程度，即在一个已经事先按照Cox HR由高到低排好序的基因列表中，core-net基因集富集top Cox HR基因的程度(Figure 2B，看NES值，该值越高，高Cox HR值基因的富集程度也越高))。

Figure 2A、Figure 2B、Figure 2C均表明dnet构建的core-net基因集预测生存期(Cox HR值)的效果优于jActiveModules构建的core-net基因集。

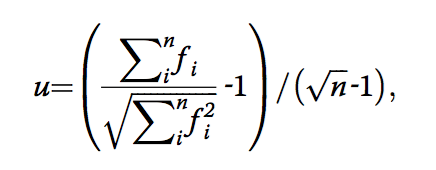


### 肿瘤相似性和异质性研究(Tumour similarity & hetetogeneity)

通过GSEA可以对core-net基因集在多癌种突变频率均一性、单癌种高突变频率、病人个体高突变数目三个不同指标的富集程度进行分析(Figure 7)。

单癌种突变频率(within-tumour-type mutation frequency)：某个基因的癌种内突变频率即为某癌种携带该突变基因的病人数目占该癌种所有病人数目的比例。

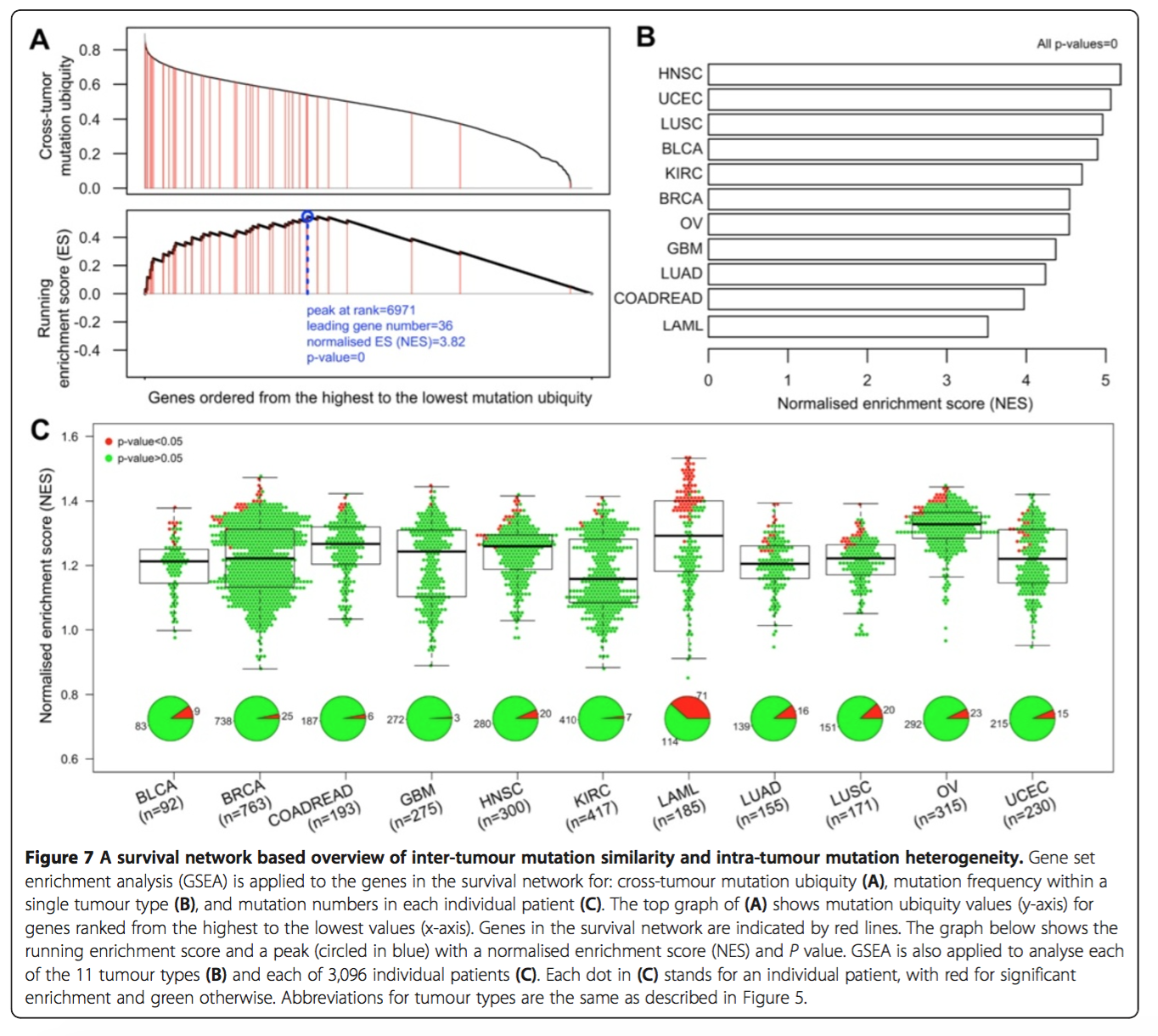
多癌种突变频率均一性(cross-tumour mutation ubiquity):基于单癌种突变频率进行定义，即某基因的单癌种突变频率在不同癌种中分布的普遍性。某基因的突变普遍性u定义为：



其中fi为某基因在癌种i的突变频率，n为癌种的数目。 u值分布范围为[0-1]，当一个基因在所有癌种中的突变频率几乎相同时，该基因的u值趋近于1，当一个基因只在一个癌种中发生突变时，该基因的u值趋近于0，其他情况介于0-1之间。

如图Figure 7A所示，core-net基因集的基因均呈现较高的u值，如图Figure 7B所示，core-net基因集在单个癌种中也呈现高频突变基因富集的趋势(共12个癌种)。相比之下，如Figure 7C所示，core-net基因集在单个肿瘤内部不同病人样本的高突变基因富集程度呈现高度的异质性(共12个癌种，3,096个病人样本)。

这些结果为解释肿瘤之间的相似性和肿瘤内部的异质性提供了新的分析维度：生存期相关。



### dnet包‘dGSEA’函数的应用总结和使用方法-GSEA分析

应用总结：

1. 基因集富集分析 (Gene Set Enrichment Analysis, GSEA) 的基本思想是使用预定义的基因集（通常来自功能注释或先前实验的结果），将基因按照在两类样本中的差异表达程度排序，然后检验预先设定的基因集合是否在这个排序表的顶端或者底端富集。基因集合富集分析检测基因集合而不是单个基因的表达变化，因此可以包含这些细微的表达变化，预期得到更为理想的结果。从中可以看到GSEA分析有三个特点：(1)分析的基因集合而不是单个基因；(2)将基因与预定义的基因集进行比较；(3)富集分析。
2. GSEA的优势：以差异富集分析为例，一般的富集分析（如GO和Pathway分析）往往侧重于比较两组间的基因表达差异，集中关注少数几个显著上调或下调的基因，这容易遗漏部分差异表达不显著却有重要生物学意义的基因，忽略一些基因的生物特性、基因调控网络之间的关系及基因功能和意义等有价值的信息（<http://www.360doc.com/content/16/1026/18/19913717_601568937.shtml> ）。

使用方法：

eTerm <- dGSEA(data=rankdata, identity="symbol", genome="Hs", ontology="Customised", customised.genesets=geneset, weight=0, nperm=2000)

只要给定一个已排序的基因列表(rankdata：第一列基因名，第二列指标值；可以按照各种指标排序，如按照Cox HR值由高到低排好序的基因列表，或者按照多癌种频率发生普遍性u值由高到低排好序的基因列表，或者按照单癌种发生频率由高到低排好序的基因列表，或者按照单个病人突变数目由高到低排好序的基因列表，或者按照分组间差异表达程度由高到低排好序的基因列表……)，就可以基于该列表计算属于该列表任意基因集(geneset)相应指标的富集指数(NES, normalised enrichment score)。

### dnet包‘dGSEA’函数的应用参考文献

1. Human primary liver cancer–derived organoid cultures for disease modeling and drug screening. Nature medicine 2017.通过对159个基因集进行GSEA分析，发现与肝癌的发生关联最大的基因集。
2. Broad研究所GSEA分析网址：<http://software.broadinstitute.org/gsea/index.jsp>

## **dnet包‘dEnricher’函数的功能：各种ontologies富集分析、基因进化溯源分析**

### core-net生存基因集生存期预测潜力以及药物相关性预测潜力

通过分析core-net生存基因集的生存期预测潜力和药物基因富集情况来评估core-net生存基因集的应用潜力。

1. 生存期预测潜力

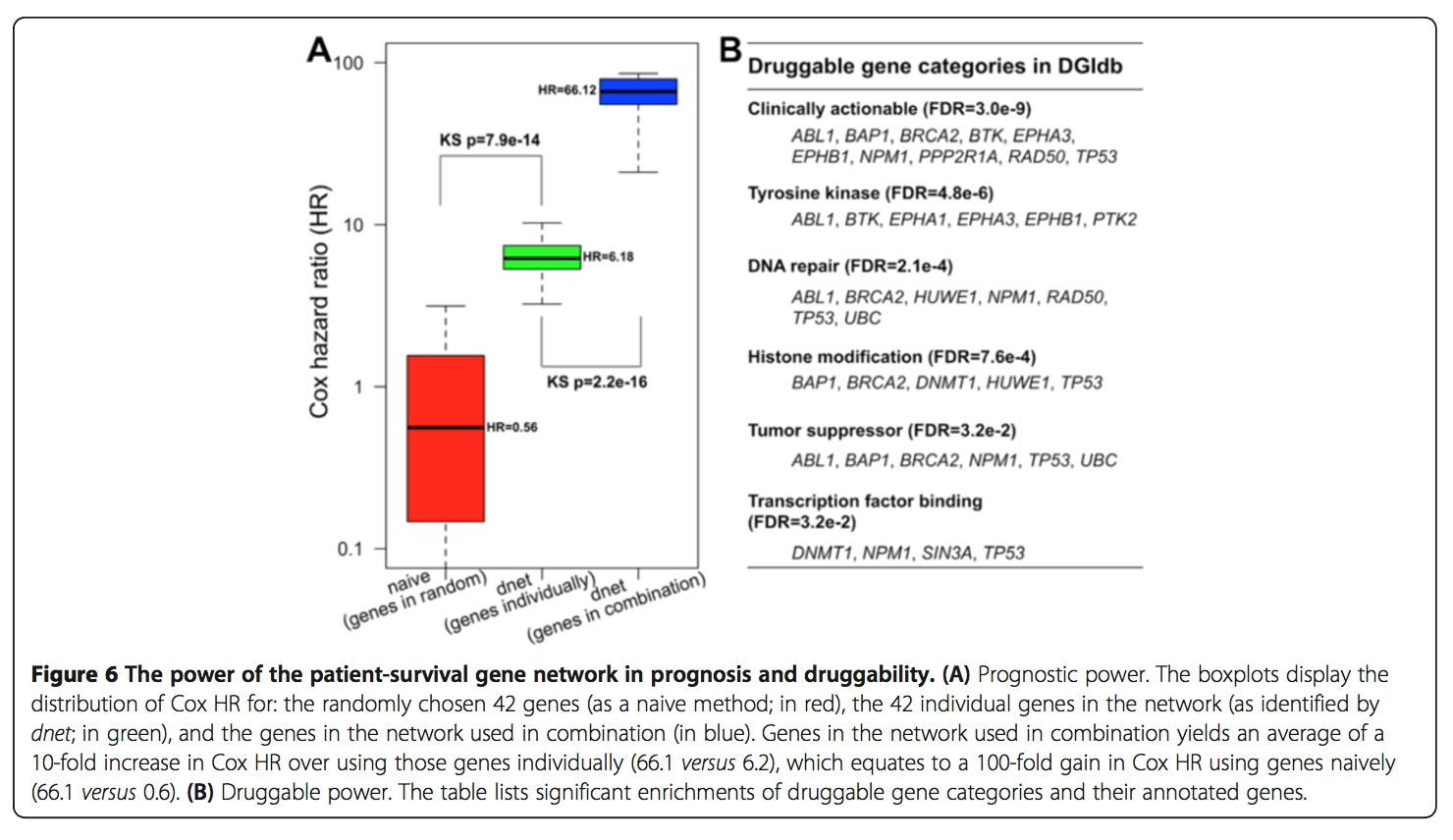
通过两种方式进行生存期预测效果比较：1. 随机选取相同数目的基因(naive baseline); 2. 选取top Cox HR值的相同数目的基因(ideal baseline)。通过两样本KS检验评估不同分布之间的差异显著性。

由Figure 6A可以发现，与naïve基因集(随机选取相同数目的基因)相比，用core-net基因集中单个基因进行生存期预测的HR平均值提高了10倍，而用用core-net基因集的组合(将top Cox HR值的基因逐个添加到组合中)进行预测提高到了100倍。由于core-net基因集中的基因并不总是高HR值的基因，这说明core-net基因集中的基因是作为整体相互作用对病人生存期产生影响的，即使这样，core-net基因集中的基因与Cox HR值最高的同等数据的基因对病人生存期预测的效力也是相当的。这些结果一起说明了core-net生存基因集作为整体能更好地进行病人生存期预测的潜力。

1. 药物相关性预测

通过药物基因分类的富集情况，评估core-net生存基因集的药物相关性。通过基于富集分析的超几何分布(dnet的‘dEnricher’函数)，发现core-net富集的DGIdb药物基因分类(富集显著性：经过多重假设检验校正后的FDR≤0.05) 。如Figure 6B所示，top1富集的药物分类是‘Clinically actionable’分类。富集在‘Clinically actionable’分类的基因包括：编码酪氨酸激酶的基因(ABL1, BTK, EPHA3, EPHB1), 参与DNA修复的基因(ABL1, BRCA2, NPM1, RAD50, TP53), 组蛋白修饰基因(BAP1, BRCA2, TP53), 抑癌基因(ABL1, BAP1, BRCA2, NPM1, TP53)。这些说明core-net生存基因集具有临床应用的潜力。

除了基于多癌种进行core-net构建，也对单个癌种进行了core-net构建，12个癌种各自构建的core-net基因集数目不等，LAML最低，LUSC最高，与每个癌种的突变数量一致。对单癌种的core-net进行药物基因富集分析发现，大部分单癌种的core-net基因集与多癌种core-net基因集富集的药物分类相类似，但是不同癌种之间的core-net基因集组成差异很大，说明不同癌种的core-net基因集行使着某些共同的功能，同时暗示了研究不同癌种共性的可能性。



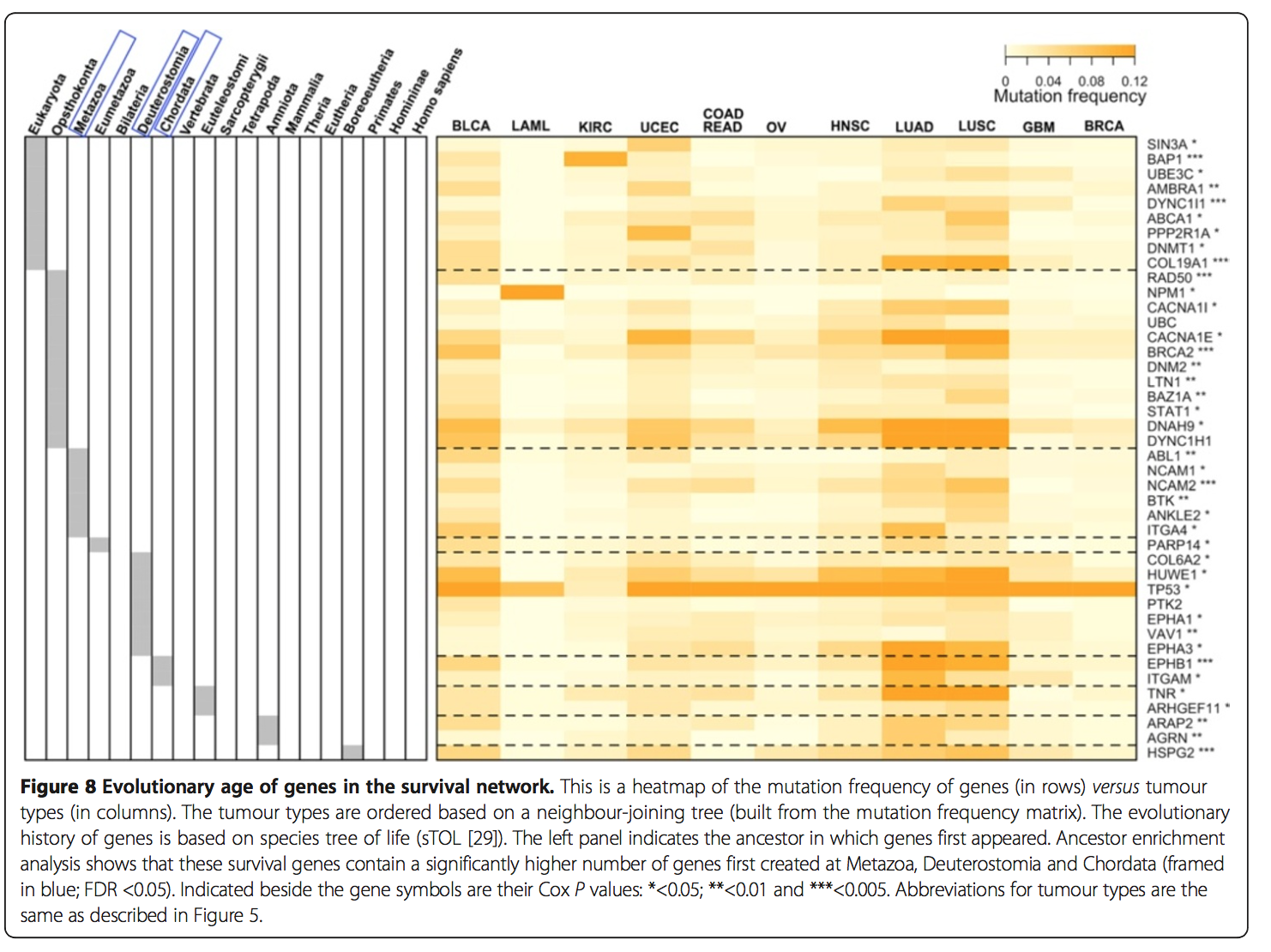
### core-net基因集进化相关性分析

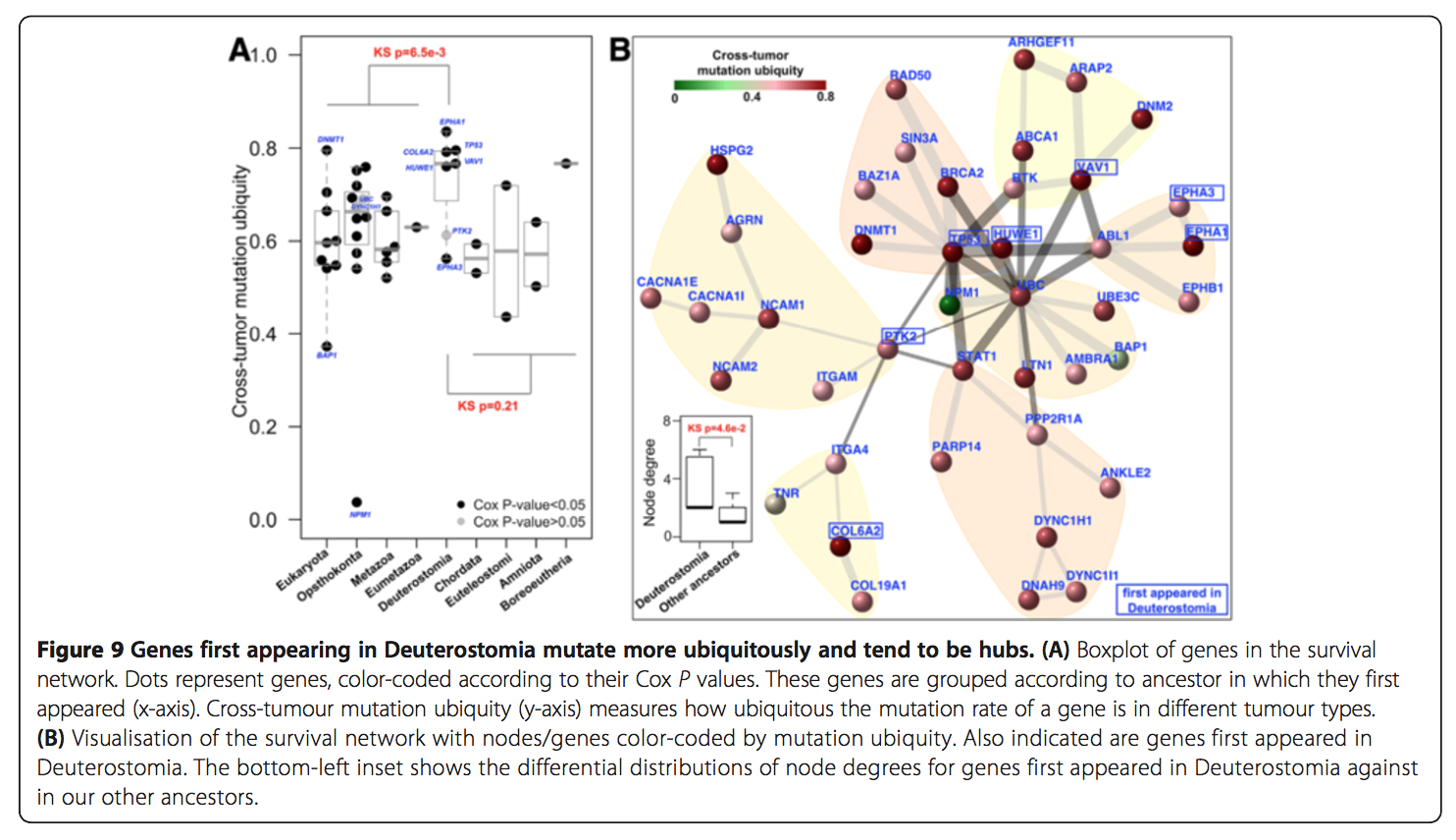
由于core-net基因集控制着癌症病人的生存期，有必要对core-net基因集进行进化溯源。基于sTOL树进行core-net基因集的物种起源定位，某个基因第一次出现的祖先决定了该基因的年龄。分析发现，半数core-net基因创建于动物-真菌分界线(后鞭毛生物，Opisthokonta)或更早，65%的core-net基因创建于动物祖先- Metazoa(后生动物)或更早，高达83%的core-net基因创建于人类祖先-Deuterostomia(新口动物)或更早。为了检验每个生存基因的溯源时间不是随机的，以每个祖先的基因集为背景进行了富集分析，结果证明了生存基因出现时间的非随机性并且倾向首次创建于Metazoa, Deuterostomia 和 Chordata(脊索动物类)这三个祖先(FDR<0.05)。

通过Figure8同时对core-net基因的溯源祖先及其在各个癌种中的突变频率进行展示发现，起源于Deuterostomia的基因倾向于在各个癌种中具有突变频率均一性。为了验证这一发现，对不同溯源祖先包含的生存基因集进行肿瘤间突变频率均一性分析发现，起源于Deuterostomia的生存基因与起源于更早祖先的生存基因相比，在肿瘤间突变频率均一性的分布上存在显著差异(Figure 9A, KS tests; P = 6.5e-3)，并且起源于Deuterostomia的生存基因与起源于Deuterostomia之后祖先的生存基因相比，在肿瘤间突变频率均一性的分布上不存在显著差异(Figure 9A, KS tests; P = 0.21)。

Figure 9B 对生存基因的突变均一性、溯源祖先和连接度(degree)进行了全面展示，从中可以发现起源于Deuterostomia的基因和起源于其他祖先的基因的节点连接度显著不同。

之前的癌症基因进化溯源分析报导过Metazoa对于原癌基因和肿瘤抑癌基因出现的重要性，与本文研究相一致：许多癌症患者的生存基因首次创建于Metazoa。除此之外，本文研究发现更晚出现的Deuterostomia在塑造癌症患者生存期相关基因方面有重要进化意义。本文研究为癌症进化分析增加了一个新的视角：生存期相关。癌症生存期相关的基因在Metazoa之后持续出现在Deuterostomia中并非巧合。Deuterostomia以其决定细胞命运的可塑性著称，这一特性在Deuterostomi之前的祖先中未曾出现。对Deuterostomia具有这一特性最保守的解释：许多广泛影响细胞命运的高频突变基因首次在该物种中得到创建。创建于Deuterostomia的生存基因对癌症病人生存期有着重要的影响，一方面由于其广泛突变于多个不同细胞/癌种，另一方面由于其进化上的必要性。这些基因有一些得到了广泛研究，像TP53及其降解物HUWE1，蛋白­-酪氨酸激酶信号(EPHA1、EPHA3)，而其他基因像PTK2、VAV1则研究的较少，本研究揭示这些基因对癌症研究的意义远大于当前的认识。



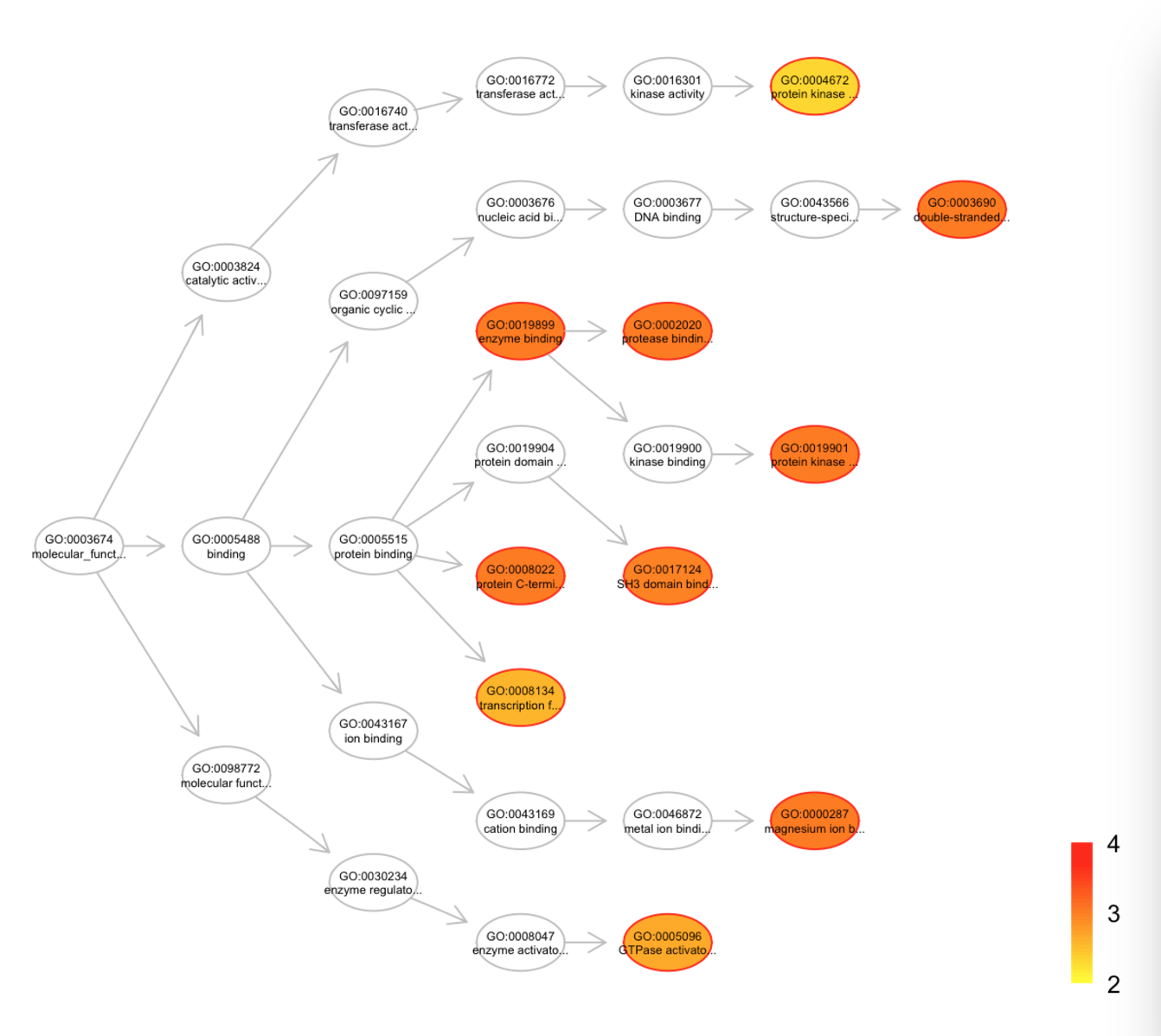


### Other ontologies 注释

除了上面提及的DGIdb ontology(DGIdb 药物基因分类富集分析)、PS2 ontology(基因进化分析)外，dnet收录了许多其他ontologies，包括GOBP（GOBP 富集分析） 、GOMF(GOMF 富集分析)、MP(Mammalian\_phenotype)、DO(Disease\_Ontology)、MsigdbC2CP(通路富集分析-CP)、MsigdbC2KEGG(通路富集分析-KEGG)、MsigdbC2REACTOME(通路富集分析-REACTOME)、MsigdbC2BIOCARTA(通路富集分析-BIOCARTA)、SF(SCOP 超家族结构阈富集分析)等等。

以上数据集均已下载至集群路径：/PUBLIC/software/CANCER/Software/dNet.更多数据集可以去<https://github.com/hfang-bristol/RDataCentre/blob/master/dnet/1.0.7/>自行下载。

展示结果均如下图所示：



### dNet包dEnricher函数的使用方法

dEnricher is supposed to conduct enrichment analysis given the input data and the ontology in query. It returns an object of class "eTerm". Enrichment analysis is based on either Fisher’s exact test or Hypergeometric test. The test can respect the hierarchy of the ontology.

dEnricher函数用法：

以GOBP富集分析为例，V(net)$name表示基因集的基因名，"Hs"表示人类基因组，ontology="GOBP"表示使用GOBP ontology数据集，RData.location指定了GOBP ontology数据集所在的路径：

eTerm <- dEnricher(data=V(net)$name, identity="symbol", genome="Hs", ontology="GOBP", RData.location="/PUBLIC/software/CANCER/Software/dNet")

### Dnet包dEnricher函数的应用参考文献

1. Human primary liver cancer–derived organoid cultures for disease modeling and drug screening. Nature medicine 2017.

基于dnet包的dEnricher功能对目标基因集的疾病相关性进行了the Disease Ontology terms富集分析，从而揭示目标基因集能够作为肿瘤标签的一些原因。

1. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature 2016.

用dnet包dEnricher函数进行了Ontology 和通路富集分析。

1. Whole-genome landscape of pancreatic neuroendocrine tumours. Nature 2017.

用dnet包dEnricher函数进行了Ontology 和通路富集分析。

**另外文献2、文献3均用dnet进行了显著差异表达基因集互作网络的构建。**

## **Dnet构建core-net基因互作网络核心代码解读**

<输入文件> 第一列基因名；第二列基因P值

# 从STRING database (version 10)数据库导入人类蛋白互作网络(igraph对象，该对象只保留了score>=400的连接关系)。

library(‘dnet’)

library(‘igraph’)

load("/PUBLIC/software/CANCER/Software/dNet/org.Hs.string.RData")

# 筛选保留score>=700的连接关系

network <- subgraph.edges(org.Hs.string, eids=E(org.Hs.string)[combined\_score>=700])

#从network中分离出只包含P值列表基因的互作网络

ind <- match(V(network)$symbol, names(pvals))

nodes\_mapped <- V(network)$name[!is.na(ind)]

network <- dNetInduce(g=network, nodes\_query=nodes\_mapped, knn=0, remove.loops=F, largest.comp=T)

V(network)$name <- V(network)$symbol

# 根据P的设定阈值得到最终需要的互作基因集-core-net

net <- dNetPipeline(g=network, pval=pvals, method="customised", significance.threshold=2.5e-02)

**基于任意基因列表(基因名称+基因显著性P值)，如生存期相关性基因列表，差异表达基因列表…均可进行相关core-net互作基因集的构建，对于构建好的core-net基因集可以用dNet包自带的函数进行各种分析，如core-net基因集communities构成分析(spinglass.community函数)、基于core-net分析不同肿瘤之间的关系(dNetReorder函数)、分析core-net基因集富集某指标的程度 (dGSEA函数)、对core-net基因集进行各种ontologies富集分析(dEnricher函数)等。**

dNet demo网址:

<http://supfam.org/dnet/demo-TCGA.html>