癌症相关基因的系统发育地层学研究

王偲琪[1]，谷迅[2]，周展[1*](1. 浙江大学药学院，杭州 310058；2. 爱荷华州立大学，美国爱荷华州 埃姆斯市 50011)

摘要：目的 结合“癌症返祖假说”，从系统发育地层学角度分析单细胞生物到多细胞生物的演化过程，研究癌症相关基因的进化年代特征，为癌症机制研究和治疗方案开发提供参考。**方法** 利用系统发育地层追踪方法 Phylostratr 定位人类所有蛋白质编码基因的进化起源，分析管家基因、癌症驱动基因、抑癌基因、原癌基因、中性基因和分化基因的进化年代特征，通过对数优势比、超几何分布比较不同功能类别基因及人类所有蛋白质编码基因的系统发育地层分布差异。利用 TCGA 获取肿瘤样本和正常样本基因表达数据，利用转录组年龄指数方法进行比较分析。**结果** 人类 20 291 条蛋白质编码基因根据序列比对结果中最远同源序列物种所在的地层被划分到 27 个地层中。鉴定 4 159 个管家基因、527 个癌症驱动基因、87 个抑癌基因、134 个原癌基因、10 755 个中性基因和 4 274 个分化基因在系统发育地层拓扑结构中的分布特征，结果显示不同功能基因的系统发育地层分布与人类所有蛋白质编码基因相比有显著差异，其中癌症相关基因具有更古老的系统发育地层分布。胆管癌、结肠癌、肺腺癌、肝癌、头颈鳞状细胞癌、肾嫌色细胞癌的转录组年龄指数分析显示，肿瘤组织中高度保守的古老基因呈现更高表达。**结论** 癌症相关基因在系统发育地层中具有更古老的进化起源，暗示其在物种进化过程中具有更保守的功能。肿瘤组织中高度保守的古老基因表达增强的现象，可用于探索肿瘤基因表达模式，为抗肿瘤药物新靶点发现及药物研究提供新思路。

关键词：癌症返祖理论；系统发育地层学；不同功能类别基因；转录组年龄指数

中图分类号：R966　　**文献标志码：**A　　**文章编号：**1007-7693(2024)02-0177-15

DOI: 10.13748/j.cnki.issn1007-7693.20233174

引用本文：王偲琪, 谷迅, 周展. 癌症相关基因的系统发育地层学研究[J]. 中国现代应用药学, 2024, 41(2): 177-191.

Phylostratigraphy Study of Cancer-related Genes

WANG Siqi[1], GU Xun[2], ZHOU Zhan[1*](*1.College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China; 2.Iowa State University, Ames 50011, USA*)

ABSTRACT: OBJECTIVE To analyze the evolution of the transition from unicellular organisms to multicellular organisms from a phylogenetic stratigraphy perspective, combining the "cancer atavism hypothesis". To investigate the evolutionary chronology of cancer-related genes to guide research on cancer mechanisms and the development of treatment strategies. **METHODS** Phylostratr was used to identify the systematic evolutionary strata of all human protein-coding genes, housekeeping genes, cancer driver genes, tumor suppressor genes, oncogenes, neutral genes, and differentiation genes. Differential distribution of genes from different functional categories and human protein-coding genes was analyzed using log-odds ratios and hypergeometric distributions. TCGA was utilized to investigate transcriptional expression datas in cancer tumor samples and normal samples, and calculations and analysis were performed using transcriptome age index. **RESULTS** A total of 20 291 protein-coding genes were classified into 27 different strata based on the farthest homologous species in the sequence alignment results. Within the phylogenetic stratigraphic structure, the datasets of 4 159 housekeeping genes, 527 cancer driver genes, 87 tumor suppressor genes, 134 oncogenes, 10 755 neutral genes, and 4 274 differentiation genes exhibit distinct distribution patterns. The overall distribution of these genes significantly differs from that of all human protein-coding genes. Cancer-related genes exhibited a more ancient phylogenetic stratigraphic distribution. Transcriptome age index results for bile duct cancer, colon cancer, lung cancer, liver cancer, head and neck cancer, and kidney chromophobe samples showed strong expression of highly conserved and ancient genes within the tumors. **CONCLUSION** Cancer-related genes exhibit older evolutionary origins within the phylogenetic context suggesting a more conserved function during species evolution. And the phenomenon of enhanced expression of highly conserved ancient genes in tumor tissues can be used to explore tumor gene expression patterns, and provide new ideas for the discovery of new anti-tumor drug targets and drug research.

KEYWORDS: cancer atavism theory; phylostratigraphy; genes of different functional categories; transcriptome age index

基金项目：国家自然科学基金面上项目 (32370712)

作者简介：王偲琪，女，硕士生　E-mail: 3170102522@zju.edu.cn　**通信作者：**周展，男，博士，副教授　E-mail: zhanzhou@zju.edu.cn

现有基因的不断变异和重新利用以及新基因的出现帮助物种适应不断变化的新环境，增加生物多样性。前者通过改变基因序列获得新的功能基序以进化得到基因新功能[1]，而后者主要源于基因从头新生[2]，被称为孤儿基因。系统发育地层学方法于 2007 年由 Domazet[3] 引入，是一种用于测定基因和基因家族出现年代的方法。通过将基因家族祖先基因 (即功能性创始蛋白质结构域) 的进化起源与特定宏观进化演变过程结合，系统发育地层学可以推断生物体中每个基因的进化起源，揭示蛋白质编码基因的年代分层。该过程以目标物种为起点，向前按照系统进化分类逐级递推至单细胞生物时期，同时使用序列比对方法，在每个系统进化发育枝中寻找是否存在目标蛋白质序列同源物，以最古老同源物所在系统发育进化层级确认基因年代。

系统发育地层方法除能用于分析基因起源模式[4]，还被用于研究进化机制，如比较基因生命周期[5]、探究各种群发育结构和细胞进化类型[6] 等。系统发育地层分析显示，与相对较老的基因相比，相对年轻的基因进化得更快[7]，表达较低[8-9]，编码较短的蛋白质[8]，受到较弱的负选择和更强的正选择[9]，不太可能与人类疾病相关[10]，较少表达于动物胚胎发育早期阶段[11] 等。

基因组不仅在数百万年的进化历程中，也在单个生物的生命周期内发生着深刻变化，其中最极端的例子就是癌症。因此，癌症也被视为是进化的缩影[12]。癌细胞几乎没有控制地增殖，调节细胞死亡功能受损、资源垄断、分化失调和细胞外环境破坏等不符合多细胞性的欺骗表型[13] 脱离了机体正常细胞固有的基本特征和演化过程，推测存在决定性因素和遗传学基础，但从进化角度对癌症的发生和性质的解释未受到广泛关注。然而，治疗和预防的进展取决于对癌症生物学更深入的理解。

Boveri[14] 将癌细胞描述为失去了正常组织细胞保留的性质并倾向于非常原始性质的细胞。随着对癌症认识的深入，Boveri 的猜测在癌症返祖假说[15-16] 中得到认可，该假说试图追踪癌症的深层进化根源，提出癌症的发生可以追溯到大约 6 亿年前复杂的多细胞生物出现之前，生物从单细胞状态过渡到多细胞状态的早期阶段。单细胞生物向多细胞生物的过渡需要获得控制以维持组织的细胞完整性，抑制克隆欺骗和癌变倾向，当控制被解除时，细胞转向原始单细胞样表型[17]。支持该假说的证据一直以来都是观察性的，直到系统进化学研究提出了癌症基因与多细胞生命出现之间的联系[18]，并在基因、表达、调控网络等角度进行了验证，才为癌症返祖理论提供了更坚实的证据。

系统发育地层学研究方法，为理解人类基因的进化发展、适应历史提供了全新视角。按照系统进化起源对基因进行分层，可以揭示基因中重要适应性事件的足迹。多细胞生物的起源可以被视为一种宏观进化转变，它需要新的基因功能。癌症应当与多细胞生命密切相关，因为它可以被看作是多细胞生物体内细胞相互作用的失常。因此，通过对癌症基因起源进行系统发育谱系追踪，也能提供关于多细胞生命起源的见解[18]，并且提供癌症机制研究新思路，指导开发治疗方案。

1　数据来源

1.1　人类基因数据

2023 年 3 月 1 日从 UniProtKB Swiss-Prot 数据库[19](UniProtKB 数据库中 Swiss-Prot 是由 TrEMBL 经过手动注释后得到的高质量非冗余数据库) 下载用于比对的人类参考基因组 fasta 文件，共得到 20405 条人类蛋白质序列，进行删除<30 aa 的蛋白质序列等数据清洗后共得到 20291 条蛋白质编码序列，版本代号为 GRCh38。

1.2　不同功能类别基因数据

为了研究不同功能类别的基因，本研究搜集各个公开基因数据库的信息进行分析。

整理 HRT Atlas v1.0 database[20] 和 Ramsköld[21]、Uhlén[22]、Eisenberg[23] 共 4 个研究中的管家基因列表，分别标记为 HRT Altas_v1.0、Ramsköld_2009、Uhlén_2015、Eisenberg_2013，取在 4 个列表中交集≥2 的 4159 个基因作为本研究的管家基因集。

整理 COSMIC Cancer Gene Census Tier1[13]、ICGC Data Portal PCAWG[24]、IntOGen[25]、Cell Consensus Gene[26]、OncoKB[27]、ONGene[28]、The Network of Cancer Genes(NCG)[29] 共 7 个研究中的驱动基因列表，分别标记为 CGC_Tier1、ICGC、IntOGen、Cell_Consensus、OncoKB、ONGene、NCG，取在 7 个列表中交集≥4 的 527 个基因作为本研究确认的癌症驱动基因集。

整理 The Network of Cancer Genes(NCG)[29]、TSGene2.0[30] 和 Davoli[31]、Vogelstein[27] 共 4 个研究中的抑癌基因列表，分别标记为 NCG6.0_tsg、TSGene2.0_tsg、Cell2013_tsg、CancerGenomeLandscapes_tsg，取在 4 个列表中交集≥2 的 87 个基因作为本研究的抑癌基因集。

整理 The Network of Cancer Genes(NCG)[29]、DORGE[32] 和 Davoli[31]、Vogelstein[27] 共 4 个研究中的原癌基因列表，分别标记为 NCG6.0_onco、DORGE_onco、Cell2013_onco、CancerGenomeLandscapes_onco，取在 4 个列表中交集≥2 的 134 个基因作为本研究的原癌基因集。

整理 Davoli 等[31] 研究中的中性基因列表，得到 10 755 个基因作为本研究的中性基因集。

整理 Gene Ontology 等[33-35] 数据库中手动搜索“分化”词条的基因结果，得到 4 274 个基因作为本研究的分化基因集。

汇总不同功能类别基因列表后，对于各个研究中基因名称不一致的情况，使用 HGNC 数据库[36] 检索基因的官方名称，进行查找、匹配与统一命名。所有注释的人类蛋白质编码基因被用作对照。

UpSetR 包[37] 用于可视化不同数据库中基因交集情况。

1.3 癌症中基因表达数据

从 TCGA(https://portal.gdc.cancer.gov/) 中下载 RNA 测序表达数据，下载时间为 2023 年 4 月 23 日。

2 方法

2.1 定义系统发育地层进化层级

根据 NCBI 网站 Taxonomy 数据库的谱系分类方法定义人类进化层级，见表 1。根据 BLAST[38] 序列比对得到的最近共同祖先所在的谱系，可以将全基因组基因划分为不同的系统进化层级。此分析框架下，不同层级下代表物种数量、代表物种基因组装质量等为评价因素。本研究将人类的基因划分为 27 个进化层级，细胞生物被定义为系统进化层级的第一层 (PS1)，真核生物定义为系统进化层级的第二层 (PS2)，依次类推，将人类定义为系统进化层级的最后一层 (PS27)。

2.2 人类基因谱系比对

使用 R 包 Phylostratr[39]，估计人类中每个基因的谱系结构。Phylostratr 包通过以下流程确认基因进化起源的系统发育地层，见图 1。①构建进化树。基于 NCBI 数据库的物种分类学，从 UniProt 数据库中检索不同分类层下的代表物种创建进化分枝数。目标物种现代人类在 UniProt 数据库中的 ID 为“9606”，使用 UniProtKB Swiss-Prot 数据库得到的非冗余人类基因序列替代系统自动检索得到的文件。②过滤进化树。使用最大化物种多样性的算法裁剪进化树，减少选择进化距离非

表 1 Phylostratr 进行人类基因年代分析中使用的物种

Tab. 1 Phylostratr conducts species information used in human genetic dating

地层	系统进化层级	类别	代表性物种Uniprot ID
1	Cellular Organisms	Unicellular(UC)	84156|2026791|153854|1457250|1262986
2	Eukaryota	Unicellular(UC)	44689|184922|1036724|353153
3	Opisthokonta	Unicellular(UC)	766039|595528|691883|81824|559292|946362|667725
4	Metazoa	Early Metazoan(EM)	400682
5	Eumetazoa	Early Metazoan(EM)	252671|6087|27923|45351|669202|10228|287889
6	Bilateria	Early Metazoan(EM)	7227|46245|6192|2762511|34506|92179
7	Deuterostomia	Early Metazoan(EM)	133434|2904734|7668
8	Chordata	Early Metazoan(EM)	7741|7740|7719|51511|34765
9	Vertebrata	Early Metazoan(EM)	7764|7757
10	Gnathostomata	Early Metazoan(EM)	7868|137246|75743
11	Euteleostomi	Early Metazoan(EM)	7906|7917|7955|27687|7918|48698
12	Sarcopterygii	Early Metazoan(EM)	7897
13	Tetrapoda	Early Metazoan(EM)	57060|260995|445787|1415580|8355
14	Amniota	Early Metazoan(EM)	38654|75864|9031|1825980|35670|8508
15	Mammalia	Mammal-Specific(MM)	9258
16	Theria	Mammal-Specific(MM)	13616|38626|9305|29139
17	Eutheria	Mammal-Specific(MM)	185453|9785|1230840|127582
18	Boreoeutheria	Mammal-Specific(MM)	9913|9615|9796|202257|191816|59463
19	Euarchontoglires	Mammal-Specific(MM)	10029|10181|9995|10094|9986|10116|246437
20	Primates	Mammal-Specific(MM)	30608|30611|1328070|379532
21	Haplorrhini	Mammal-Specific(MM)	1868482
22	Simiiformes	Mammal-Specific(MM)	37293|9483|2715852|39432|9515
23	Catarrhini	Mammal-Specific(MM)	60711|336983|9541|591936|61622
24	Hominoidea	Mammal-Specific(MM)	61853
25	Hominidae	Mammal-Specific(MM)	9601
26	Homininae	Mammal-Specific(MM)	9598
27	Homo sapiens	Mammal-Specific(MM)	9606

常近的物种，尽可能保留系统发育中多样性的代表，减少因相关生物蛋白质组之间的依赖而造成的偏倚。同时尊重用户选择的权重约束，不同进化枝/物种类别中对应的物种数量，可以通过重新定义权重进行修改，对于可用序列太少或蛋白质组不完整的物种，系统将自动进行过滤。本研究中在对人类基因进行分析时，选择默认值不进行修改。③填充进化树。从 UniProt Proteome 数据库[40-41] 中检索蛋白质组，并可以根据需要删除或增加系统自动生成的分枝数物种。经过本步骤，在对人类基因进行地层分析时，共使用 102 个物种 (表 1)。④构建蛋白质组数据库。每个蛋白质组数据库中包括 10 个文件，用于后续分析。⑤目标基因的相似性检索。针对进化分枝中每个物种的蛋白质组进行目标物种蛋白质组的成对 BLAST 检索。根据灵敏度和准确度的需要，使用者可以对 E 值阈值进行修改，还可以用 HMMER 等[42] 算法取代 BLAST 算法。本研究中使用 HMMER3.0 和 Pfam 进行蛋白质数据库的检索和构建。⑥推断同源序列。根据每个目标物种蛋白质编码序列的“最佳命中”，即平衡了灵敏度和准确度后最符合要求的蛋白质编码序列，确认该条序列的最古老同源序列。针对蛋白质编码序列进行同源性匹配，使用 BLASTP 算法，并设置 E 值阈值 10^{-4}。⑦推断目标基因地层。基于“最佳命中”，将蛋白质序列分配到与该序列具有同源序列的最深进化枝对应的地层，即以基因能比对到的最古物种所在的进化层级定义基因年龄，在任何物种中无比对结果的基因定义为人类特有的新基因。本研究对人类基因的系统发育地层分类，共设置 27 个分类地层。

2.3 基因功能注释

Gene Ontology[33-35] 数据库提供关于基因及其产物功能结构化、可计算的知识，即定义具有特定相互作用关系的基因功能类 (GO 术语)。本体论涵盖了基因功能的 3 个不同方面：分子功能 (基因产物在分子水平上的活性)、细胞成分 (基因产物活性相对于生物结构的位置) 和生物过程 (利用基因分子功能的生物程序)。

DAVID[43] 生物信息学资源被用来理解所获得的基因列表在生物学上的重要性，在基因数据库信息的基础上可以提供关于基因功能分类、功能注释聚类、功能注释图、功能注释表、基因 ID 转换和基因名称批量查看等功能。

将人类不同功能类别基因数据导入 GO 和 DAVID 数据库，进行通路富集分析。

2.4 统计分析

比较特定功能类别基因起源于每个系统发育地层中的频率与其预期出现频率，通过计算对数几率比显示偏离情况。本研究使用 Log(odds+1)，Log2 表示分布无差异，>Log2 表示过度代表，<Log2 表示代表性不足。虚假发现率在 0.05 水平进行多重比较[44] 校正的双尾超几何检验[45] 来测试

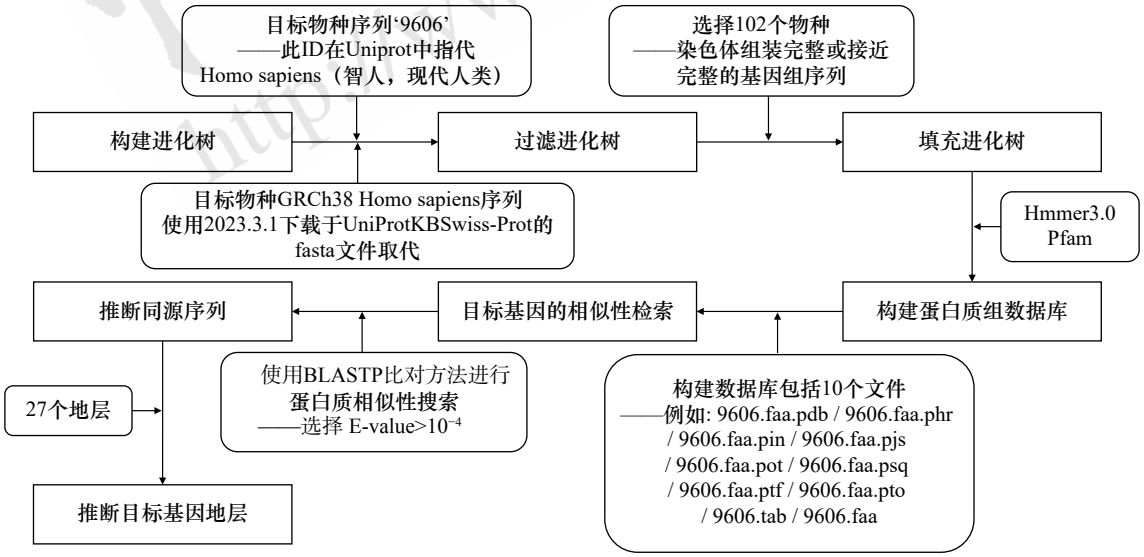

图 1 使用 Phylostratr 进行人类基因年代分析的流程

Fig. 1 Flow of human genetic dating using Phylostratr

与预期频率的偏差，P<0.05为差异有统计学意义。

2.5 转录组年龄指数 (transcriptome age index，TAI) 计算

TAI[11] 将基因年龄与其在特定发育阶段的表达水平结合考虑，根据进化年龄累积度量 (公式 1) 加权计算样本中所有基因的表达水平。使用 TAI 方法计算肿瘤和正常样本的转录组年龄：

$$\mathrm{TAI}=\frac{\sum_{i=1}^{n} ps_i \times e_i}{\sum_{i=1}^{n} e_i} \quad (1)$$

ps_i 代表基因 i 的系统发育地层，e_i 代表基因 i 的表达量，n 代表用于 TAI 计算的基因总数。通过赋予年轻地层更大的权重，弥补古老地层拥有更多基因的事实。

对于基因在肿瘤和正常组织样本中的表达数据，当 TPM>1 时计入考虑。

3 结果

3.1 蛋白质序列长度

以人类为焦点物种进行系统发育地层分析时，使用 BLASTP 比对 27 个地层中除人类外剩下 101 个物种与人类蛋白质序列的相似性。蛋白质序列长度会影响 BLAST 比对的准确性、特异性和可靠性[46]，分析 102 个物种的蛋白质序列长度，见图 2。四分值结果显示，多数物种蛋白质序列长度>150 个氨基酸，适用于系统发育地层进化分析。物种蛋白质序列长度最小值多在 50 个氨基酸以内，这些蛋白更倾向于与新进化基因有关[47]，不适用于同源序列比对，因此在研究中删除长度<30 aa 的人类蛋白质序列。

3.2 不同 E 值截断结果

使用系统发育地层分析方法 R 包 Phylostratr 对人类所有编码基因进行年代定位，不同 E 值截断会改变人类基因进化起源分布。与基因复制相比，从头基因的诞生对于新 (功能性) 基因起源的作用是进化基因组学的重要课题，并在不同分类群得到了充分证实[48-49]。系统发育地层学结果可以提供从头基因的基本概述，当序列存在高度分歧时，同源序列检测程序如 BLASTP 易出现假阴性从而低估基因年龄[50]。比对不同 E 值截断下基因起源分布差异，可以提供对基因保守性的新见解。例如，比对 E 值截断为 10^{-2} 和 10^{-5} 下人类基因起源分布 (图 3)，共 16 170 个基因进化地层结果保持一致，即接近 80% 的基因进化起源不变。一方面，说明 BLAST 比对方法的可行性，另一方面，体现平衡准确度和灵敏度的必要性。此外，剩下地层结果不一致的基因表明存在起源分歧，值得进一步细究。

3.3 人类基因系统发育地层进化起源

系统发育地层学方法用来评估人类基因中 20 291 个注释蛋白编码基因出现的时间 (图 4)，基因被映射到相应的系统发育地层 (PS)。根据具有高组装质量的蛋白质序列信息的群体共识系统发育关系，定义了 27 个系统进化地层。第一层细胞生物

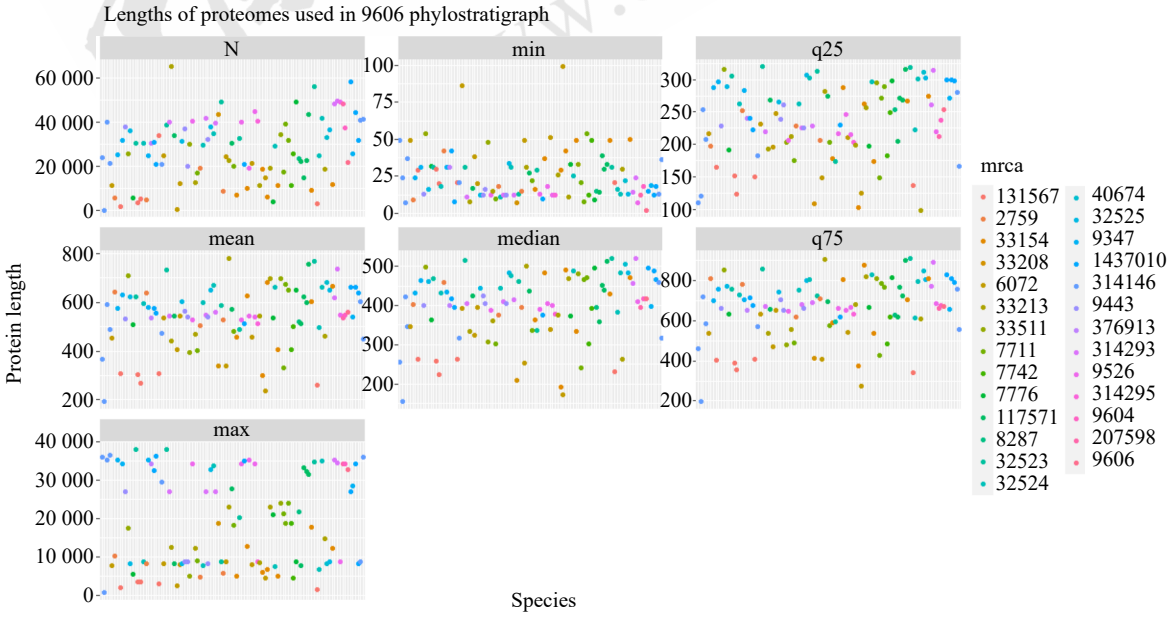

图 2 102 个物种蛋白质编码序列长度

Fig. 2 Length of protein-coding sequences for 102 species

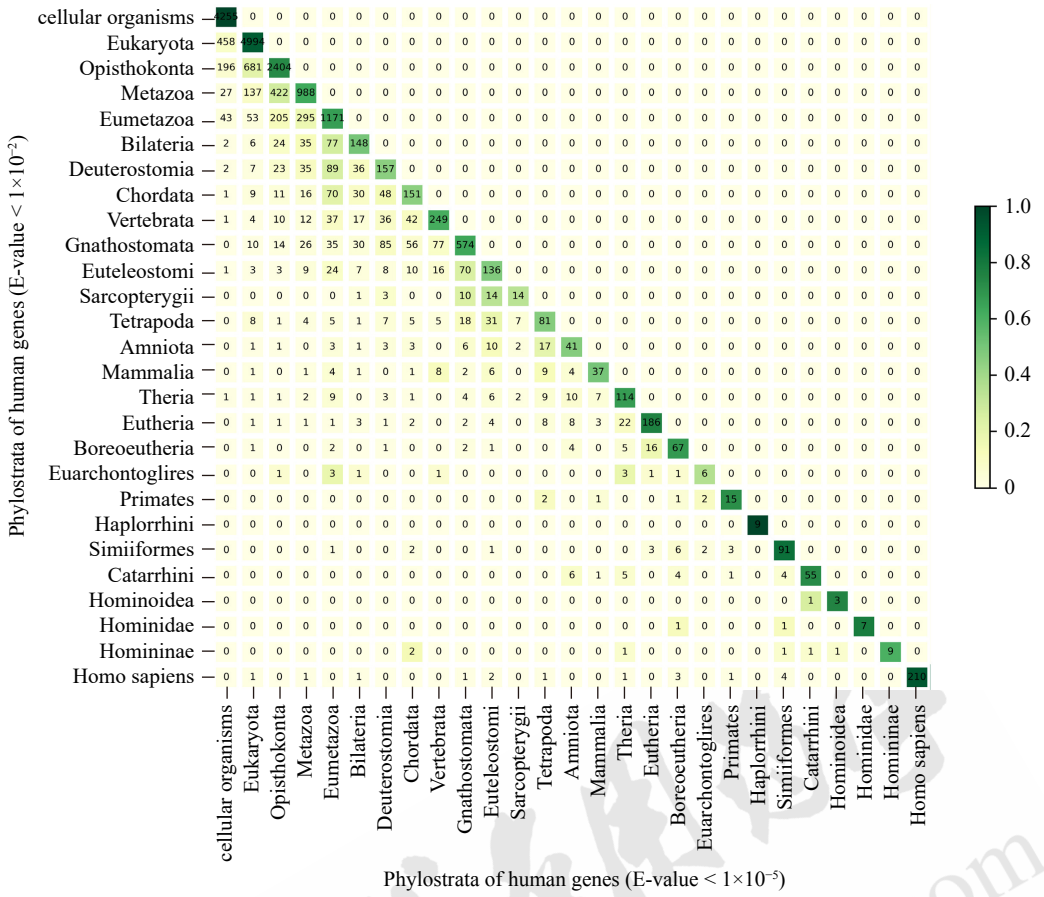

图 3 BLASTP 比对时不同 E 值截断下人类基因起源的系统发育地层分配

Fig. 3 Phylostratigraphic distribution of human gene origin under different E-value truncation during BLASTP comparison

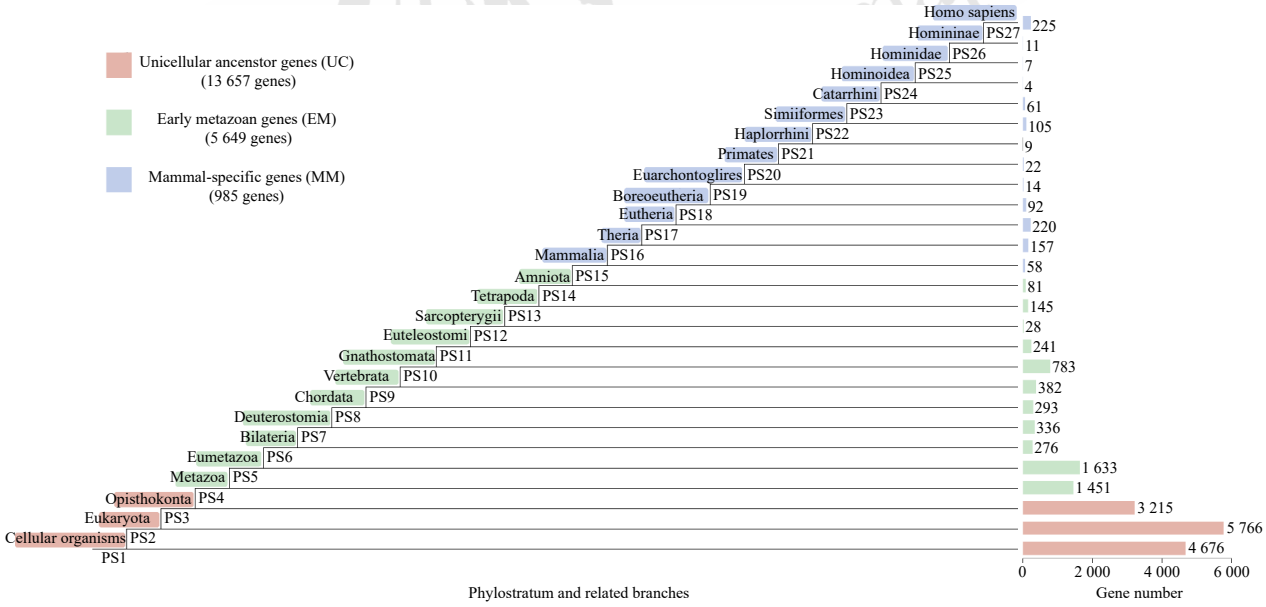

图 4 人类基因组系统发育地层起源分布

Fig. 4 Phylostratigraphic origin and distribution of human genome system

层 (PS1) 代表所有细胞生命的基础，起源于该层的基因具有最长久的发育历史，是最古老的一批基因。最后一个地层代表人类层 (PS27)，进化于该层的基因即具有最短暂的发育历史，是最年轻的一批基因。本研究在 BLASTP 搜索同源序列时，使用 10^{-4} 作为截断值，在灵敏度和准确度中进行

折中，选择较高的准确度，减少假阳性率。

人类基因组中约 67% 的注释蛋白质编码基因起源于单细胞生物 (PS1-3)(图 5)，这些基因通常与基础的细胞功能有关，如代谢过程和转录调控，见图 6。其余基因出现在系统发育历史的后期，即后生动物时期 (PS4-27)。该时期，基因起源峰值与

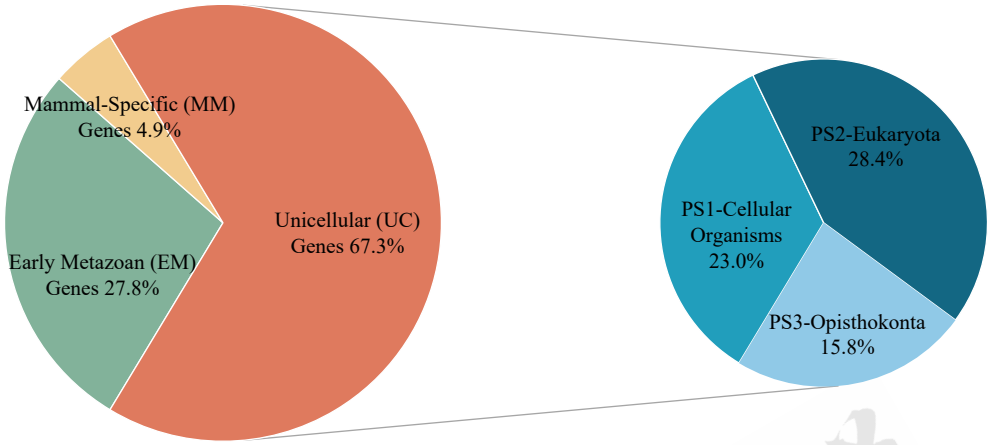

图 5　不同系统发育地层中人类基因组分布比例

Fig. 5　Human genome distribution ratio in different phylostratums

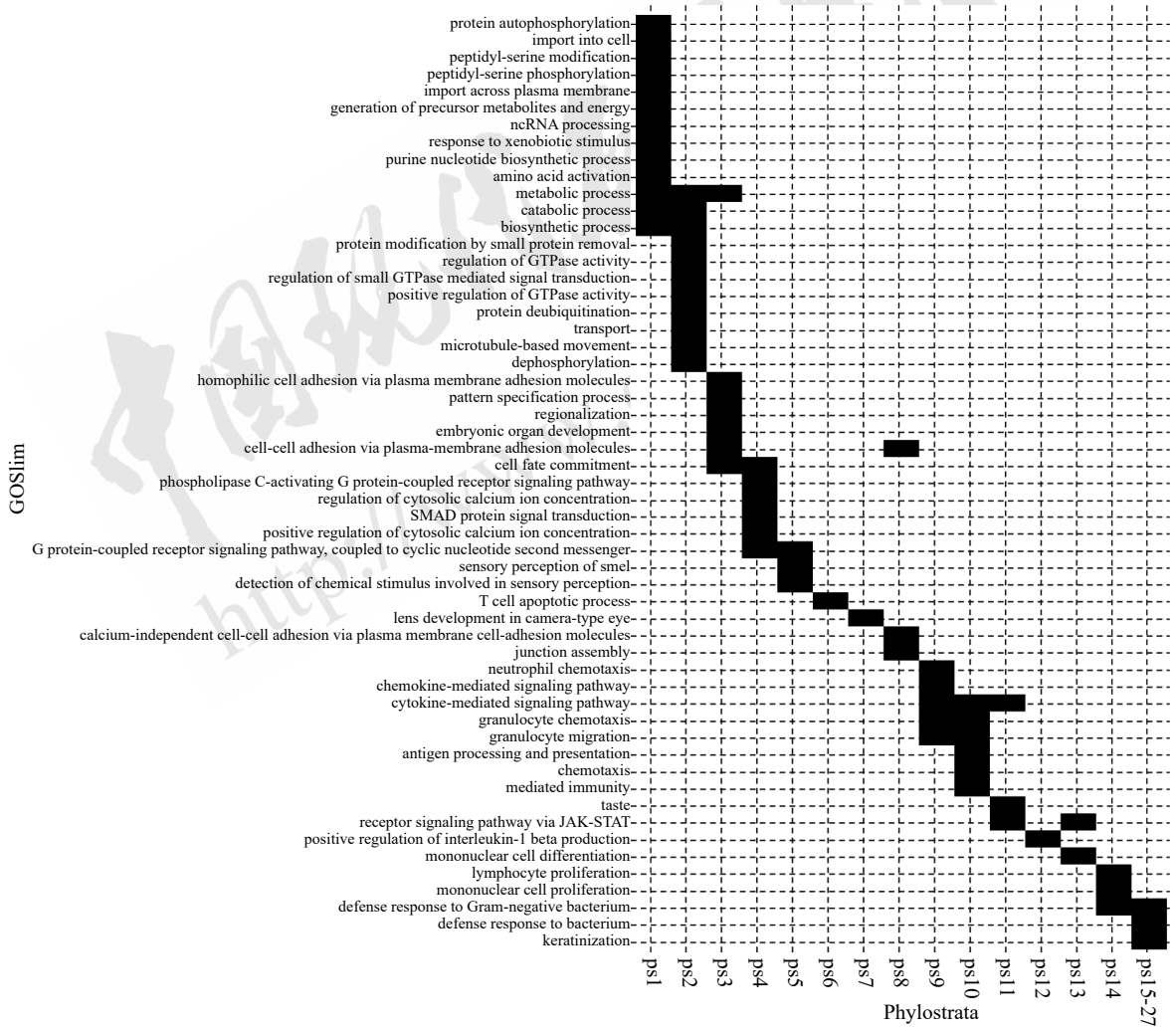

图 6　不同系统发育地层起源基因的 GO 富集结果

Fig. 6　GO enrichment results of phylostratigraphic origin genes in different phylostratums

重大生物进化历史有关。基因出现的第二个高峰PS4-5 地层，代表单细胞向多细胞生物的转变[18]，如 G 蛋白信号转导、细胞-细胞通讯和视觉等功能可以追溯到该时期。与抑癌功能相关的功能，如抑制基因无限制扩增、T 细胞凋亡等也在该时期得到进化，见图 6。PS10-11 周围的峰值代表无脊椎动物向脊椎动物的转变，PS17-18 周围的峰值代表爬行动物向哺乳动物的转变，主要与免疫、神经系统、感觉器官等功能的出现有关。PS27 的高峰代表了在人类诞生以后演化出的基因，其中许多可能代表在该类群从头进化的孤儿基因，与大脑、言语等功能的发展有关，为研究人类特异性的分子机制提供了重要视角。

3.4 不同功能类别基因的数据来源比对

为评估不同功能类别基因在系统发育地层的分布特点，本研究整理了癌症驱动基因、抑癌基因、原癌基因、管家基因、中性基因、分化基因共六类功能基因列表，并汇总前 4 种功能基因列表不同数据来源的交集情况 (图 7)，取在至少一半数据集中得到验证的基因交集用于后续分析。

3.5 不同功能类别基因的系统发育地层年代比较

癌症的发生发展可以被理解成是个体进化过程[12]，其特征让人联想到单细胞生物。由于癌症与体细胞突变，特别是热点突变间存在不可忽视的联系，驱动基因是本研究的重点。通过同源性搜索，对人类谱系中不同分类群和不同类型的基因进行系统发育地层研究，论证癌症返祖现象。

根据进化年龄，绘制人类不同功能类别基因和所有蛋白质编码基因的分布，见图 8。得到的 27 个地层与对应的年龄进行匹配，更直观体现基因进化起源特性。不同功能类别基因的平均进化年龄以百万年为单位进行数值测量，即分析不同类别基因集中 50% 基因起源对应的时间点。不同类别基因组中位年龄排序，从老到新依次为原癌基因、管家基因、抑癌基因、驱动基因、所有蛋白质编码基因、分化基因和中性基因。进一步对比发现，年龄分布上，不同功能类别基因与所有蛋白质编码基因相比具有显著统计学差异，见图 9。

系统发育地层第一、二层代表真核生物的起源，该层级涵盖结构功能和生物过程生成的重要进化转变。原癌基因、管家基因、抑癌基因、驱动基因在该地层有最广泛的分布，体现基因功能与真核生物原始生物过程和功能间的强关联性及古老基因在癌症中的高度突变和过度表达性。即在原核生物中进化的古老突变应激反应被用来保持生殖系和免疫系统的多样性，同时在癌症中还原原始表型。抑癌基因数量激增于 Opisthokonta 和 Bilateria 时期，体现抑癌作用与早期后生生物发生发展间的强关联性，在单细胞生物向多细胞

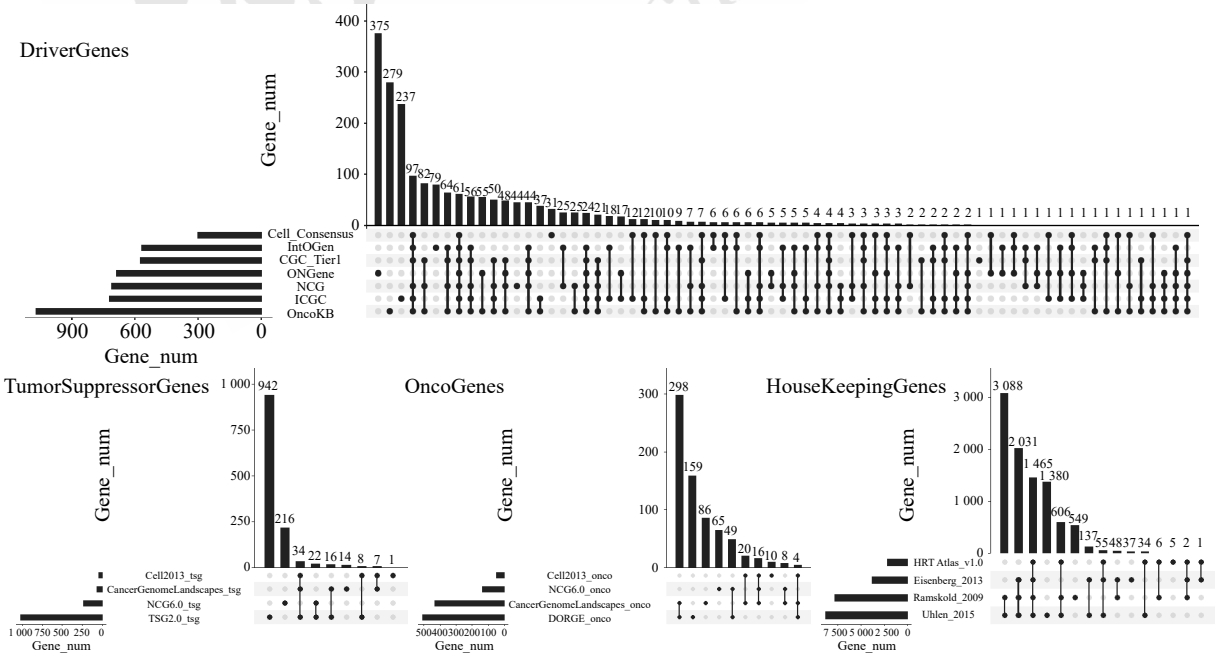

图 7 不同数据来源的癌症驱动基因、抑癌基因、原癌基因、管家基因数据集的并交集情况

Fig. 7 Intersection of cancer driver genes, tumor suppressor genes, oncogenes and house-keeping genes from different data sources

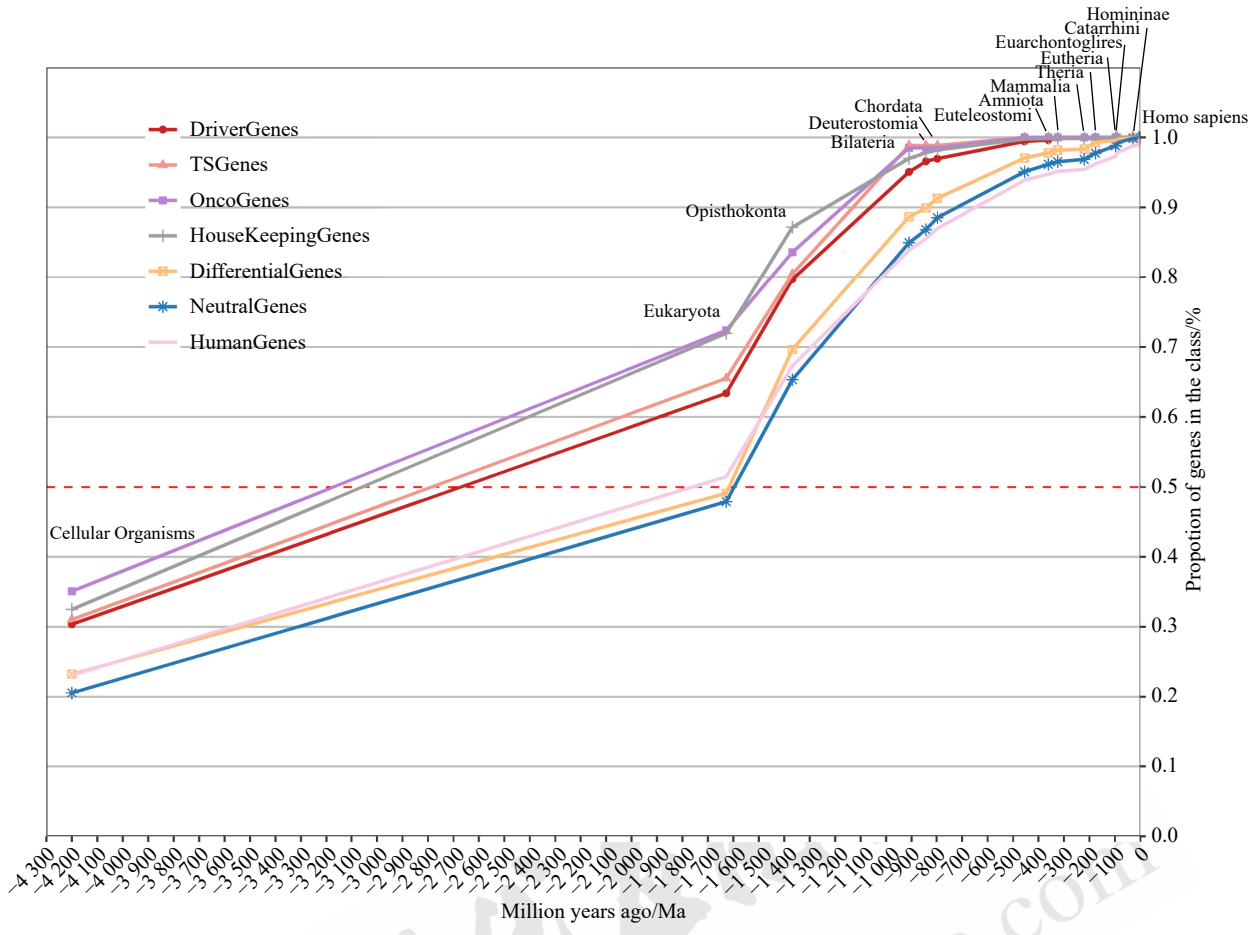

图 8 人类不同功能类别基因和所有蛋白质编码基因按进化年龄分布

Fig. 8 Distribution of different functional classes and all protein coding genes according to their evolutionary ages

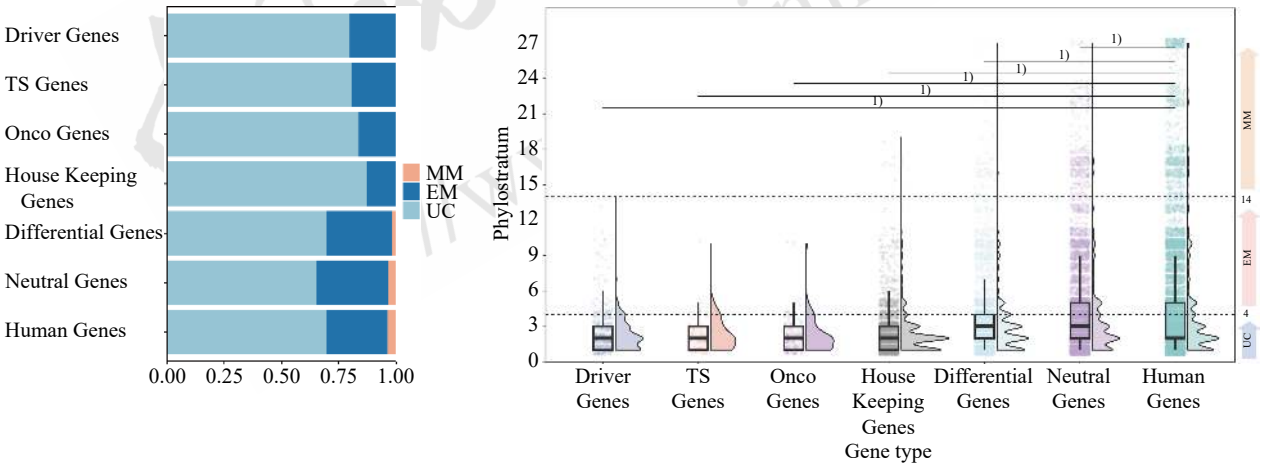

图 9 6 种功能类别基因地层起源占比分析

1)P<0.001。

Fig. 9 Analysis of phylostratigraphic origin proportion of six functional classes

1)P<0.001.

生物转变过程中，抑癌基因起到抑制单细胞性表型，并保持基因组的稳定性的作用。癌症中单细胞祖先特征的再次出现，源于抑制它们或将它们限制在胚胎发生或伤口愈合等特定环境中的调节被破坏。

3.6 超几何统计分布显示不同功能基因的分布差异

将基因起源定位到系统发育地层，通过计算

超几何统计分布，结合优势对数比计算结果，可以对每个系统发育地层中特定基因类别数量的相对过多或者过少情况进行统计分析，不同功能类别基因在与所有蛋白质编码基因组相同的拓扑结构上显示出不同的模式，见图 10。

管家基因在早期单细胞生物时期 PS1-2 显著高表达，此后在系统进化地层中均代表性不足。这与管家基因维护基因组的完整性、避免 DNA 损伤和突变的功能契合，说明确保基因组稳定性的功能机制在远早于生物多细胞性出现之前已经存在。

中性基因除在最新进化地层 PS26-27 中代表性不足外，在大多数地层中显著过度表达。中性基因本身可能没有直接功能，但在进化过程中可以提供遗传多样性，有助于提高生物适应性和进化潜力。此外，也可以用作分子生物学和遗传学研究中的控制标记，帮助研究其他基因和遗传过程。

分化基因在 PS3-4、PS9 和 PS16 中显著高表达，对应于多新兴功能出现的多细胞类、脊椎类和哺乳类生物诞生时期。

原癌基因在 PS1-2 和 PS4 中具有显著过表达，在 PS5 中代表性严重不足，体现原癌基因在物种发展中根深蒂固的历史。原癌基因在发生突变或异常活化的情况下可以促进细胞不受控制增殖和癌症发展，这种推动细胞向单细胞特征发展的作用，暗示原癌基因与单细胞生物时期不可分割的联系。

抑癌基因在早期后生动物 PS4 中显著过度表达，与多细胞生物特性相契合，对应于其抑制癌症的发展、维持细胞健康和避免不受控制的细胞增殖的作用。

驱动基因在单细胞生物时期和早期后生时期 PS1-4 中显著过度表达，自生物进化到后生动物后期开始，驱动基因在统计学上显著代表性不足。驱动基因的突变或异常活化促使肿瘤细胞的生长和分裂，系统发育地层分析验证了其与单细胞时期密不可分的关系。

与癌症有关的 3 类基因原癌基因、抑癌基因和癌症驱动基因在系统发育地层上的拓扑分布显示有 2 个强烈的高峰，一个是在单细胞生物起源时期，另一个是在多细胞生物进化开始的早期后生动物时期，体现癌症相关基因与单细胞性和多细胞性间的联系。

3.7 基因 TAI 比较

系统发育地层学的重要特性是建立了一个系统发育尺度，基因组中的每个基因都有其系统发育等级。在使用系统发育层次结构的基础上，本研究通过 TAI 将其与个体肿瘤和正常样本的基因表达数据进行关联。该方法将基因年龄与其在特定发育阶段的表达水平结合考虑，将同一地层的基因表达加权相加。TAI 值反映样本表达基因的年代情况，较低的 TAI 值说明样本更高表达古老基因，代表更古老的转录组。反之，较高的

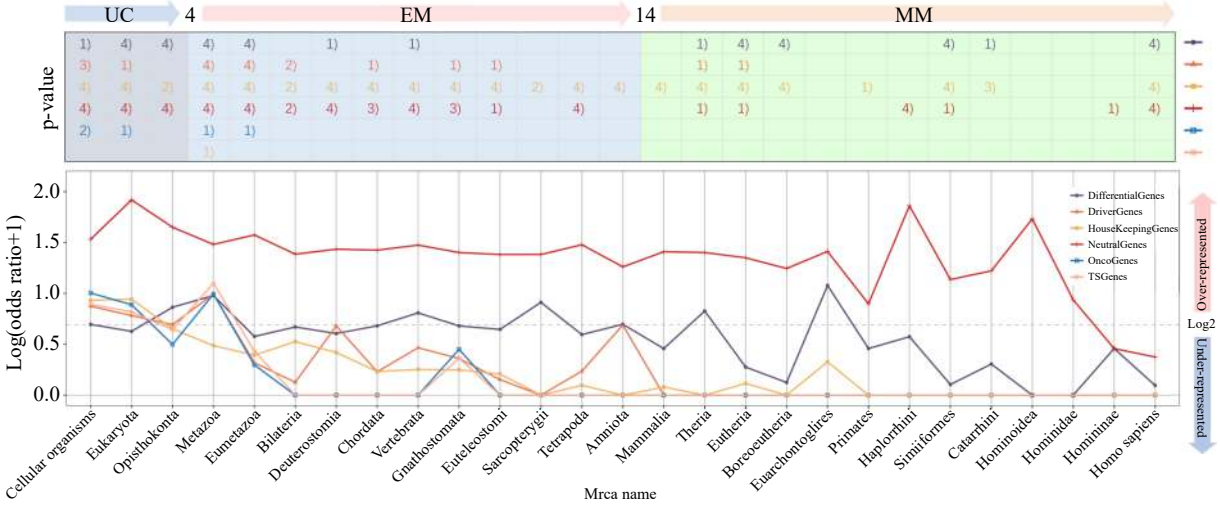

图 10 系统发育地层中不同类别基因的统计分析

[1)]P<0.05，[2)]P<0.01，[3)]P<0.001，[4)]P<0.000 1。

Fig. 10 Statistical analysis of different datasets on the phylostratigraphic map

[1)]P<0.05, [2)]P<0.01, [3)]P<0.001, [4)]P<0.000 1.

TAI 值说明转录组高表达基因多为年轻基因，代表更年轻的转录组。使用 TAI 方法分析肿瘤样本和正常组织样本 TAI 差异，发现胆管癌 (CHOL)、结肠癌 (COAD)、肺腺癌 (LUAD)、肝癌 (LIHC)、头颈鳞状细胞癌 (HNSC)、肾嫌色细胞癌 (KICH) 6 种肿瘤类型的肿瘤样本 TAI 值低于正常对照样本，见图 11。说明肿瘤转录组中高度保守的古老基因倾向高表达，通过上调源自原始单细胞祖先的基因和广泛失活更近期进化的基因，激活单细胞样表型，促使肿瘤细胞去分化回归到系统发育早期状态。

4 讨论

从时间上看，使用系统地层学方法验证癌症返祖理论经历了 3 个发展阶段：第一阶段发生在 2007—2010 年，起源于 Domazet-Lošo 等开始从事基因系统发育地层学研究[3,11,18]。Domazet-Lošo 使用系统发育框架对基因进行分组，结合果蝇的胚胎表达数据揭示进化中重要适应性事件的足迹。将系统发育地层学运用到人类基因年代分析，结合癌症数据，Domazet-Lošo 验证了与癌症发生发展有关的基因功能的古老起源。进一步拓展系统发育地层学，结合个体发育基因表达水平，Domazet-Lošo 开发了转录组年龄指数分析方法。第二阶段发生在 2017—2019 年，包括由 Trigos 等拓展系统发育地层学的分子生物学研究[51-53]。在基因表达分析基础上，通过研究肿瘤中多细胞性基因调控网络的变化，Trigos 将肿瘤发生的驱动因素描述为在早期多细胞生命进化过程中产生关键调控环节的基因，并从进化角度挖掘药物靶点。第三阶段发生在 2019—2021 年，癌症返祖理论的假设由 Bussey、Davies 和 Lineweaver[54-55] 进行完善。癌症被定义为重新部署古老的单细胞程序，以牺牲宿主为代价支持细胞的存活，并打破多细胞生命所需的合作契约。返祖理论将癌症看作是细胞回到早期进化能力的现象[56]。该古老程序在原生动物时代进化，多细胞性适应中并没有必然地抹去古老的基因程序，只是抑制了它们的活性，至少在生物体寿命的持续时间内。相关研究仍在不断完善，包括使用系统发育地层学进行人类卵巢癌的基因和基因共表达网络研究，揭示卵巢癌不同功能聚类模块进化模式[57]；使用系统地层学研究单细胞向多细胞生物发育转变中基因调控和细胞间通信功能的进化特征等[58]。

Domazet 等[18] 使用根据研究需求量身定制的 Perl 脚本分析基因系统发育起源。Ekstrom 等[59] 提出的 ORFanFinder 是一种基于网络的系统地层学工具，它将分析限制在与 NCBI 共同树的命名等级相对应的地层中，对于许多用途来说分辨率太低，并且不提供诊断、统计或定制。本研究使用的 Phylostratr R 包为判断基因起源和系统发育年代提供了高度自动化、个性定制化的系统地层管道。

系统发育地层学研究的焦点物种 (本研究中为人类) 的基因年龄，由焦点物种与其具有基因序列同源性的最远亲分类群的分化时间确定，该特性使得系统发育地层学对假阳性敏感。受到序列相似性的影响，快速进化的基因比缓慢进化的基因失去序列相似性的速度更快，因此预计前者会比后者具有更高的序列比对错误率，这将产生年轻基因更快进化的虚假模式。此外，根据定义，从头起源的孤儿基因在谱系其他物种中没有同源物，如何在基因组中识别孤儿基因并与噪声区分，避免孤儿基因未被注释或错误注释成为一大挑战。

本研究通过在各地层模式选取代表模式物种，控制系统发育地层学误差[60]。该过程需要使用同源序列比对方法，最常用的是 BLAST 比对。BLAST 使用启发式搜索策略，通过查找序列间的

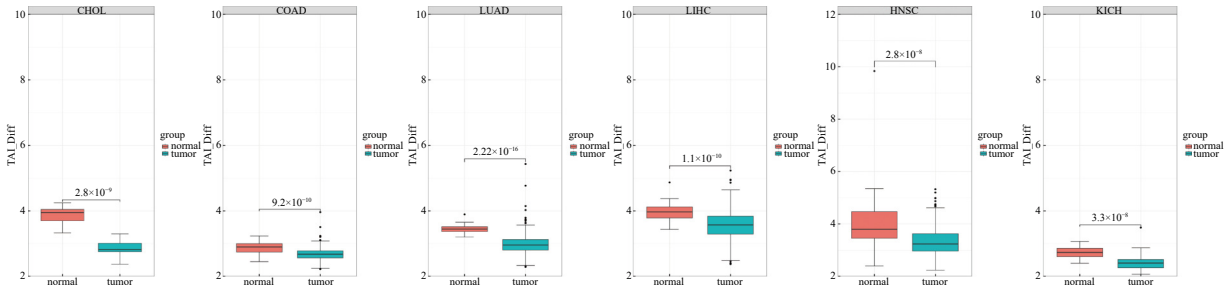

图 11 6 种癌症和正常样本的 TAI 结果

Fig. 11 Six TAI results of tumor and normal samples

短匹配片段来识别潜在的相似区域。之后，BLAST 方法扩展短序列以生成更长的局部对齐。在这个过程中，常用的是 BLOSUM(Blocks Substitution Matrix) 或 PAM(Point Accepted Mutation) 矩阵。这些矩阵是基于已知蛋白质家族的比对数据制定的，反映了不同氨基酸替代的可能性和频率。这些矩阵隐含地考虑到了不同位点的进化速率可能有所不同。受基因序列特点影响，BLAST 比对可能造成误差并产生有偏移的结果。假阳性错误是由于非同源基因被错误识别成同源基因导致基因年龄被高估，通过使用严格的 E 值截断可以避免因偶然序列相似性引起的假阳性。假阴性错误是由于同源物与查询的基因序列相似性太低而无法被检测到，广泛的计算模拟表明，这种系统发育年龄判断错误的情况是不可忽视的，至少在 5%~14% 基因中发生[61]。研究表明，BLAST 分析基因起源的能力取决于基因的进化速率[62-63] 和序列长度[64-65]，由于快速进化的基因比缓慢进化的基因失去序列相似性的速度更快，预计前者会比后者具有更高的 BLAST 错误率，这将产生年轻基因更快进化的虚假模式[62]。理解 BLAST 误差如何影响系统发育地层学的可靠性将对各种演化研究具有重要意义，Moyers 等[61] 通过现实的计算机模拟和基因组数据估计的参数来评估系统发育地层学的准确性，并调查其误差对基因组进化结果的影响。未来进一步研究可以考虑从不同比对方法出发，进行人类基因进化年代分析，比较其他同源序列检测方法，如 HMMER[42]、SMI-BLAST[66]、PSI-BLAST 等是否具有更好的表现。或者尝试使用更复杂的模型，如模拟分子钟的方法，以在分析中考虑不同蛋白及位点的进化速率差异。当然，这可能需要更多的数据和计算资源，以提高准确性。

研究中使用 BLAST 分析的误差会偏向保守，即具有更高进化速率的基因可能会被放置在比他们真实起源年龄更年轻的系统发育地层中。鉴于研究中采取了较低的 E 值截断 (10^{-4})，可以追踪基因功能模块[67]。此外，对比 Domazet 等[18] 将人类基因划分为 19 个地层年代和 Trigos 等[53] 将人类基因划分为 16 个地层，本研究根据 NCBI 分层情况，将人类基因划分到 27 个地层。在后生动物发育时期提供了更细分的系统地层，为捕捉快速进化基因的发育历史，提供更精确的基因定位。所以认为对人类基因的年代研究有追溯到大多数基因的真实起源。

癌症返祖理论将癌症被描述为细胞从适应多细胞组织生活的情况下回到单细胞行为的状态。如果是这样，可以预期，癌症中的基因突变会破坏抑制有害于多细胞性功能的机制，这些机制应该在多细胞生命出现之后或与之同时进化，并且在肿瘤中应当存在单细胞时期基因的高表达。基因表达系统地层学分析表明，癌症与单细胞生物向多细胞生物过渡时形成的返祖遗传程序最为密切相关[68]。使用 TAI 可定量比较不同地层基因间表达高低，较低的 TAI 对应于早期系统发育地层基因的较高表达。目前 TAI 指数方法已经被广泛应用于动物、植物、微生物等类群发育表达分析，在发育沙漏模型进化推演、干细胞起源进化、海洋幼虫起源进化等科研问题分析中取得突破性发现[69]。本研究从 TCGA 中挖掘肿瘤和正常组织样本最新 TPM 表达数据用于 TAI 指数分析，在 CHOL、COAD、LUAD、LIHC、HNSC、KICH 6 种肿瘤和正常组织样本中的计算结果表明，肿瘤组织具有更古老的转录组，高表达 UC 基因，较低表达年轻基因，是对返祖理论结果的验证。

研究提供的生物医学意义在于，进一步论证癌症相关基因 (驱动基因、抑癌基因、原癌基因) 存在在古老地层年代中显著过度表达，在最新进化地层中代表性不足的现象。考虑到>90% 的癌症相关基因在双侧对称动物之前就已经出现，为理解基因参与生物过程的背景和机制，可使用代表基因出现进化水平的模式生物，如使用与人类进化距离较远的昆虫类生物作为癌症基因研究功能模型似乎是一种可行方案。

未来，以系统发育地层学为角度的深入研究可以从以下方面展开。一是聚焦基因，结合地层年代深挖基因、基因位点、基因调控网络等的进化特点。二是聚焦肿瘤，结合地层年代分析癌症疾病进展模式，研究突变负担和表观遗传失调顺序，针对不同癌症疾病进展地层差异，从正常组织器官进化特点等方向展开探讨。三是聚焦单细胞到多细胞状态的相关途径，如细胞黏附和细胞通信等方面研究。从癌症演化步骤角度出发探索癌症机制研究至关重要，有望为理解生物发展及癌症进化提供新思路，发现癌症早筛及诊断新方

式，挖掘抗癌新靶点。

5 结论

本研究使用系统发育地层学，将人类基因从进化角度划分为 27 个地层，提供了更细致的人类基因地层年代图谱。从系统发育地层学角度进行基因年代分析，为理解基因进化历史提供了新思路，为多细胞起源的分析提供了新见解，并证实了癌症相关基因的古老起源，即发现癌症驱动基因、抑癌基因和分化基因多来源于人类古老的系统发育地层，在进化上具有同时性。尽管相当数量癌症相关基因的起源早于多细胞生物发现时期，但癌症相关基因进化的第二个高峰与多细胞生物出现紧密相关。通过 TAI 数据分析，发现肿瘤组织转录组转向古老基因的表达增强，即上调源自原始单细胞祖先的基因和广泛失活更近期进化的基因，验证癌症返祖理论的合理性。针对癌症相关基因的古老起源和古老基因在肿瘤组织的高表达特性，可以指导临床癌症机制研究和治疗方法开发，服务精准医学的发展。

REFERENCES

[1] KAESSMANN H. Origins, evolution, and phenotypic impact of new genes[J]. Genome Res, 2010, 20(10): 1313-1326.

[2] MCLYSAGHT A, HURST L D. Open questions in the study of *de novo* genes: What, how and why[J]. Nat Rev Genet, 2016, 17(9): 567-578.

[3] DOMAZET-LOSO T, BRAJKOVIĆ J, TAUTZ D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages[J]. Trends Genet, 2007, 23(11): 533-539.

[4] CARVUNIS A R, ROLLAND T, WAPINSKI I, et al. Proto-genes and *de novo* gene birth[J]. Nature, 2012, 487(7407): 370-374.

[5] ABRUSÁN G. Integration of new genes into cellular networks, and their structural maturation[J]. Genetics, 2013, 195(4): 1407-1417.

[6] SESTAK M S, BOŽIČEVIĆ V, BAKARIĆ R, et al. Phylostratigraphic profiles reveal a deep evolutionary history of the vertebrate head sensory systems[J]. Front Zool, 2013, 10(1): 18.

[7] ALBÀ M M, CASTRESANA J. Inverse relationship between evolutionary rate and age of mammalian genes[J]. Mol Biol Evol, 2005, 22(3): 598-606.

[8] WOLF Y I, NOVICHKOV P S, KAREV G P, et al. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages[J]. Proc Natl Acad Sci U S A, 2009, 106(18): 7273-7280.

[9] CAI J J, PETROV D A. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes[J]. Genome Biol Evol, 2010(2): 393-409.

[10] DOMAZET-LOSO T, TAUTZ D. An ancient evolutionary origin of genes associated with human genetic diseases[J]. Mol Biol Evol, 2008, 25(12): 2699-2707.

[11] DOMAZET-LOŠO T, TAUTZ D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns[J]. Nature, 2010, 468(7325): 815-818.

[12] MERLO L M, PEPPER J W, REID B J, et al. Cancer as an evolutionary and ecological process[J]. Nat Rev Cancer, 2006, 6(12): 924-935.

[13] SONDKA Z, BAMFORD S, COLE C G, et al. The COSMIC Cancer Gene Census: Describing genetic dysfunction across all human cancers[J]. Nat Rev Cancer, 2018, 18(11): 696-705.

[14] BOVERI T. Über mehrpolige Mitosen als Mittel zur Analyse des Zellkerns[J]. Verh Phys. Med Ges Würzburg, 1902(35): 67-90.

[15] VINCENT M. Cancer: A de-repression of a default survival program common to all cells? : A life-history perspective on the nature of cancer[J]. Bioessays, 2012, 34(1): 72-82.

[16] DAVIES P C W, LINEWEAVER C H. Cancer tumors as Metazoa 1.0: Tapping genes of ancient ancestors[J]. Phys Biol, 2011, 8(1): 015001.

[17] HAMMERSCHMIDT K, ROSE C J, KERR B, et al. Life cycles, fitness decoupling and the evolution of multicellularity[J]. Nature, 2014, 515(7525): 75-79.

[18] DOMAZET-LOSO T, TAUTZ D. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in Metazoa[J]. BMC Biol, 2010, 8: 66.

[19] BOUTET E, LIEBERHERR D, TOGNOLLI M, et al. UniProtKB/swiss-prot[M]. Plant Bioinformatics. Totowa, NJ: Humana Press, 2007: 89-112.

[20] HOUNKPE B W, CHENOU F, LIMA F D, et al. HRT Atlas v1.0 database: Redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets[J]. Nucleic Acids Res, 2021, 49(D1): D947-D955.

[21] RAMSKÖLD D, WANG E T, BURGE C B, et al. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data[J]. PLoS Comput Biol, 2009, 5(12): e1000598.

[22] UHLÉN M, FAGERBERG L, HALLSTRÖM B M, et al. Proteomics. Tissue-based map of the human proteome[J]. Science, 2015, 347(6220): 1260419.

[23] EISENBERG E, LEVANON E Y. Human housekeeping genes, revisited[J]. Trends Genet, 2013, 29(10): 569-574.

[24] SABARINATHAN R, PICH O, MARTINCORENA I, et al. The whole-genome panorama of cancer drivers[J]. bioRxiv, 2017. Doi: 10.1101/190330.

[25] MARTÍNEZ-JIMÉNEZ F, MUIÑOS F, SENTÍS I, et al. A compendium of mutational cancer driver genes[J]. Nat Rev Cancer, 2020, 20(10): 555-572.

[26] BAILEY M H, TOKHEIM C, PORTA-PARDO E, et al. Comprehensive characterization of cancer driver genes and mutations[J]. Cell, 2018, 173(2): 371-385. e18.

[27] VOGELSTEIN B, PAPADOPOULOS N, VELCULESCU V E, et al. Cancer genome landscapes[J]. Science, 2013,

339(6127): 1546-1558.
[28] LIU Y N, SUN J C, ZHAO M. ONGene: A literature-based database for human oncogenes[J]. J Genet Genomics, 2017, 44(2): 119-121.
[29] REPANA D, NULSEN J, DRESSLER L, et al. The Network of Cancer Genes (NCG): A comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens[J]. Genome Biol, 2019, 20(1): 1.
[30] ZHAO M, KIM P, MITRA R, et al. TSGene 2.0: An updated literature-based knowledgebase for tumor suppressor genes[J]. Nucleic Acids Res, 2016, 44(D1): D1023-D1031.
[31] DAVOLI T, XU A W, MENGWASSER K E, et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome[J]. Cell, 2013, 155(4): 948-962.
[32] LYU J, LI J J, SU J Z, et al. DORGE: Discovery of oncogenes and tumor suppressor genes using genetic and epigenetic features[J]. Sci Adv, 2020, 6(46): eaba6784.
[33] ASHBURNER M, BALL C A, BLAKE J A, et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium[J]. Nat Genet, 2000, 25(1): 25-29.
[34] GENE ONTOLOGY CONSORTIUM, ALEKSANDER S A, BALHOFF J, et al. The gene ontology knowledgebase in 2023[J]. Genetics, 2023, 224(1): iyad031.
[35] CONSORTIUM T G O. The Gene Ontology Resource: 20 years and still GOing strong[J]. Nucleic Acids Res, 2019, 47(D1): D330-D338.
[36] SEAL R L, BRASCHI B, GRAY K, et al. Genenames.org: The HGNC resources in 2023[J]. Nucleic Acids Res, 2023, 51(D1): D1003-D1009.
[37] CONWAY J R, LEX A, GEHLENBORG N. UpSetR: An R package for the visualization of intersecting sets and their properties[J]. Bioinformatics, 2017, 33(18): 2938-2940.
[38] ALTSCHUL S F, GISH W, MILLER W, et al. Basic local alignment search tool[J]. J Mol Biol, 1990, 215(3): 403-410.
[39] ARENDSEE Z, LI J, SINGH U, et al. Phylostratr: A framework for phylostratigraphy[J]. Bioinformatics, 2019, 35(19): 3617-3627.
[40] UNIPROT CONSORTIUM. UniProt: The universal protein knowledgebase in 2021[J]. Nucleic Acids Res, 2021, 49(D1): D480-D489.
[41] UNIPROT CONSORTIUM. UniProt: A hub for protein information[J]. Nucleic Acids Res, 2015, 43(Database issue): D204-D212.
[42] FINN R D, CLEMENTS J, ARNDT W, et al. HMMER web server: 2015 update[J]. Nucleic Acids Res, 2015, 43(W1): W30-W38.
[43] SHERMAN B T, HAO M, QIU J, et al. DAVID: A web server for functional enrichment analysis and functional annotation of gene lists (2021update)[J]. Nucleic Acids Res, 2022, 50(W1): W216-W221.
[44] BENJAMINI Y, HOCHBERG Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing[J]. J R Stat Soc Ser B Stat Methodol, 1995, 57(1): 289-300.
[45] RIVALS I, PERSONNAZ L, TAING L, et al. Enrichment or depletion of a GO category within a class of genes: Which test?[J]. Bioinformatics, 2007, 23(4): 401-407.
[46] WILSON B A, FOY S G, NEME R, et al. Young genes are highly disordered as predicted by the preadaptation hypothesis of *De novo* gene birth[J]. Nat Ecol Evol, 2017, 1(6): 0146-0146.
[47] JAMES J E, WILLIS S M, NELSON P G, et al. Universal and taxon-specific trends in protein sequences as a function of age[J]. Elife, 2021(10): e57347.
[48] BAALSRUD H T, TØRRESEN O K, SOLBAKKEN M H, et al. *De novo* gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data[J]. Mol Biol Evol, 2018, 35(3): 593-606.
[49] ZILE K, DESSIMOZ C, WURM Y, et al. Only a single taxonomically restricted gene family in the *Drosophila melanogaster* subgroup can be identified with high confidence[J]. Genome Biol Evol, 2020, 12(8): 1355-1366.
[50] JAIN A, PERISA D, FLIEDNER F, et al. The evolutionary traceability of a protein[J]. Genome Biol Evol, 2019, 11(2): 531-545.
[51] TRIGOS A S, PEARSON R B, PAPENFUSS A T, et al. Altered interactions between unicellular and multicellular genes drive hallmarks of transformation in a diverse range of solid tumors[J]. Proc Natl Acad Sci USA, 2017, 114(24): 6406-6411.
[52] TRIGOS A S, PEARSON R B, PAPENFUSS A T, et al. How the evolution of multicellularity set the stage for cancer[J]. Br J Cancer, 2018, 118(2): 145-152.
[53] TRIGOS A S, PEARSON R B, PAPENFUSS A T, et al. Somatic mutations in early metazoan genes disrupt regulatory links between unicellular and multicellular genes in cancer[J]. Elife, 2019(8): e40947.
[54] BUSSEY K J, DAVIES P C W. Reverting to single-cell biology: The predictions of the atavism theory of cancer[J]. Prog Biophys Mol Biol, 2021(165): 49-55.
[55] LINEWEAVER C H, BUSSEY K J, BLACKBURN A C, et al. Cancer progression as a sequence of atavistic reversions[J]. Bioessays, 2021, 43(7): e2000305.
[56] THOMAS F, UJVARI B, RENAUD F, et al. Cancer adaptations: Atavism, *de novo* selection, or something in between?[J]. Bioessays, 2017, 39(8): 1700039. Doi: 10.1002/bies.201700039.
[57] ZHANG L Y, TAN Y, FAN S J, et al. Phylostratigraphic analysis of gene co-expression network reveals the evolution of functional modules for ovarian cancer[J]. Sci Rep, 2019, 9(1): 2623.
[58] JACQUES F, BARATCHART E, PIENTA K J, et al. Origin and evolution of animal multicellularity in the light of phylogenomics and cancer genetics[J]. Med Oncol, 2022, 39(11): 160.
[59] EKSTROM A, YIN Y B. ORFanFinder: Automated identification of taxonomically restricted orphan genes[J]. Bioinformatics, 2016, 32(13): 2053-2055.
[60] CHEN F, MACKEY A J, STOECKERT C J Jr, et al. OrthoMCL-DB: Querying a comprehensive multi-species collection of ortholog groups[J]. Nucleic Acids Res, 2006,

34(Database issue): D363-D368.
[61] MOYERS B A, ZHANG J Z. Phylostratigraphic bias creates spurious patterns of genome evolution[J]. Mol Biol Evol, 2015, 32(1): 258-267.
[62] ELHAIK E, SABATH N, GRAUR D. The "inverse relationship between evolutionary rate and age of mammalian genes" is an artifact of increased genetic distance with rate of evolution and time of divergence[J]. Mol Biol Evol, 2006, 23(1): 1-3.
[63] MOYERS B A, ZHANG J Z. Toward reducing phylostratigraphic errors and biases[J]. Genome Biol Evol, 2018, 10(8): 2037-2048.
[64] MOYERS B A, ZHANG J Z. Further simulations and analyses demonstrate open problems of phylostratigraphy[J]. Genome Biol Evol, 2017, 9(6): 1519-1527.
[65] MOYERS B A, ZHANG J Z. Evaluating phylostratigraphic evidence for widespread *de novo* gene birth in genome evolution[J]. Mol Biol Evol, 2016, 33(5): 1245-1256.
[66] JIN X P, LIAO Q, WEI H, et al. SMI-BLAST: A novel supervised search framework based on PSI-BLAST for protein remote homology detection[J]. Bioinformatics, 2021, 37(7): 913-920.
[67] ALBÀ M M, CASTRESANA J. On homology searches by protein Blast and the characterization of the age of genes[J]. BMC Evol Biol, 2007, 7: 53.
[68] VINOGRADOV A E. Human transcriptome nexuses: Basic-eukaryotic and metazoan[J]. Genomics, 2010, 95(6): 345-354.
[69] 王静, 王师. 转录组年龄指数法在动物发育进化研究中的应用[J]. Bio-protocol, 2021: e1010642-e1010642.

收稿日期：2023-10-31
(本文责编：沈倩)

青年编委：周展

周展，博士，副教授，博士生导师。现担任浙江大学药学院药学系副主任、药物代谢和药物分析研究所副所长，浙江大学智能创新药物研究院院长助理。研究方向主要围绕肿瘤精准医疗，通过组学大数据和人工智能技术寻找驱动肿瘤发生发展的关键基因及突变，构建抗原识别系统筛选肿瘤特异性新抗原靶点，并利用蛋白质设计技术开展靶向生物药物研究。主持和参与多项国家自然科学基金、国家重点研发计划等科研项目，在 *Nat Commun*、*Mol Biol Evol*、*Mol Ther* 等权威期刊发表学术论文 70 多篇，授权发明专利 6 项，获得软件著作权 12 项。