科学通报 2017 年 第 62 卷 第 7 期: 619 ~ 625
《中国科学》杂志社
SCIENCE CHINA PRESS
Science125个科学前沿问题系列解读(XXIII)
Why do humans have so few genes?

人类基因知多少

刘顺, 屈良鹄*

中山大学有害生物控制与资源利用国家重点实验室, 广州 510275
* 联系人, E-mail: lssqlh@mail.sysu.edu.cn

2016-08-23 收稿, 2016-12-05 修回, 2016-12-07 接受, 2017-01-20 网络版发表
国家自然科学基金(31471223)资助

摘要 2003年, 人类基因组计划的完成, 宣告了后基因组时代的到来. 研究人员惊奇地发现, 组成人类基因组的基因只有25000个左右, 比之前估计的要少得多. 2005年, 美国*Science*在其创刊125周年之际, 把"为什么人类基因会如此之少"列为21世纪125个最具挑战性的科学前沿问题的第3位, 并进行了专门评述. 本文结合蛋白编码基因的表达调控及非编码RNA研究新进展对其进行解读.

关键词 人类基因组, 蛋白编码基因, 基因表达调控, 非编码 RNA, 调控网络

自孟德尔发现遗传规律以来, 人们逐渐认识到生物体内存在一种遗传因子主导着生命的生长与繁殖. 这种遗传因子就是日常所说的"基因"(gene). 基因是指能够编码功能性RNA或者蛋白产物的一段DNA序列, 它是控制生物性状的基本遗传单位. 20世纪60年代, 在"遗传密码"(genetic code)被破译之后, 人们开始了一场基因的探索与发现之旅. 人类等高等生物是否具有更多的基因? 本文将从人类基因组中基因的鉴定出发, 讨论该如何理解人类基因的种类、数目及其结构与功能.

1 人类基因有多少?

1.1 人类基因数量的评估

历史上, 人类基因数目的估算及测定经历了较大的变化. 20世纪60年代, 科学家通过构建随机布尔网络(random boolean networks)模型预测人类基因约有两百万[1], 而在20世纪70年代早期, 研究人员通过计算由有害突变引起的遗传负荷(genetic load)推测人类基因组约有四万个功能座位[2]. 在1995年, 有学者借助cDNA和表达序列标签技术(expressed sequence tags, EST)比对技术估计人类编码蛋白质的基因(蛋白质基因)总数约为十万个[3]. 这样的数字波动性变化直到2003年人类基因组计划(human genome project, HGP)的完成才得以稳定下来. 人类基因组计划实现了基因的终极定位, 它与曼哈顿原子弹计划和阿波罗登月计划一起被称为20世纪的三大科学工程. 出乎意料的是, 人类基因组的初始草图版本揭示了人类蛋白基因只有26000~30000[4,5], 最后完成版本又将这一数目修改至25000左右[6]. 并且, 随着基因鉴定以及分析方法的不断发展, 这一数字还在继续减少. 例如, "GENCODE计划"[7]致力于提供人类和小鼠(*Mus musculus*)基因组高质量的参考基因注释, 在查询其各版本的基因数目的统计信息时, 发现蛋白质基因的数目在不断下降, 尤其是人类基因组的最新版本(GRCh38/hg38)发布之后, 人类蛋白质基因的数目就未超过20000. 而较新的一项研究通过分析蛋白质组学实验以及整合来自7项大规模质谱研究、五十多份人体组织的数据, 提出了人类蛋白质基因的数目只有大约19000[8].

引用格式: 刘顺, 屈良鹄. 人类基因知多少. 科学通报, 2017, 62: 619–625
Liu S, Qu L H. What don't we know about human genes? (in Chinese). Chin Sci Bull, 2017, 62: 619–625, doi: 10.1360/N972016-00761

1.2 人类与其他模式生物基因数量的比较

与其他模式生物相比，人类蛋白质基因的数目并不算多，甚至比一些较低级的生物如仅有1 mm长的秀丽线虫(*Caenorhabditis elegans*)还要少(表1). 有研究还提出假设，有超过90%的人类蛋白质基因起源于亿万年前动物界的后生动物或者多细胞生物，并且在这些基因中，超过99%的比灵长类动物的起源还要早五千万年[8]. 该研究还指出，区分人类和小鼠的基因甚至还不超过10个. 这样看来，高等生物并不一定拥有更多的基因. 在人类基因数目较少这一事实的背后，隐藏着极其重要的生命奥秘有待解读.

2 人类基因为什么这么少?

人类基因组计划表明，蛋白质基因数量的多少并不能决定一个物种的复杂性，那么数量有限的人类基因是怎样创造出结构如此复杂的生物个体呢? 早在1975年, King和Wilson[11]在研究人类和黑猩猩的进化关系时，发现了基因调控区域的变化是导致物种差异性的关键因素，而非基因本身，正是这种变化造就了人类如今独有的基因表达调控模式. 与物种间基因数目及基本功能的差异比起来，基因表达模式对造成物种间生理和发育差异的影响更显著. 基因只是遗传功能的一个基本单位，大量基因的不同组合表达可以产生巨大的生物多样性. 基因需要转录出功能性产物来发挥作用，因此基因的表达调控主要分为转录调控和转录后调控两个层次.

2.1 精细缜密的基因转录调控

人类所有的基因并不是同时表达的，相反，它们的表达具有显著的时空特异性. 不同的组织细胞，不同的发育阶段，人体体内所表达的基因数目以及程度都会不同，而控制这些基因表达开和关的主要因素是一些表观遗传调控蛋白、转录因子蛋白和DNA序列上的调控元件. 表观遗传调控蛋白针对基因邻近的DNA或者组蛋白进行甲基化、乙酰化或者磷酸化等共价修饰. 不同的修饰标记及数目对基因的转录控制会产生不同的影响. 如组蛋白H3尾的多个赖氨酸残基均可发生甲基化，H3第4位赖氨酸(K4)和第79位赖氨酸(K79)的甲基化与基因的转录激活有关，而第9位赖氨酸(K9)和第27位赖氨酸(K27)的甲基化与基因的转录沉默相关联. 这些位点上的去甲基化则与甲基化的效果相反. 这些不同模式的修饰状态及组合能作为一种语言进行“阅读”，从而调控基因表达的开与关. 转录因子是另一类调控基因转录的蛋白，它通过结合在特定的DNA序列上，促进或者抑制RNA聚合酶招募到目的基因上. 目前推测的人类转录因子约有1500个，它们常与其他蛋白形成复合物来发挥正调节或者负调节的作用. 正调节的

表 1 部分模式物种的基因组大小及蛋白质基因数量比较(统计信息来自GENCODE v24[7], Ensembl v84[9], Ensembl Plants v31[10], Ensembl Fungi v31[10]及Ensembl Bacteria v31[10])

Table 1 The genome size and protein-coding gene number comparison among some model species. (Statistical data from GENCODE v24[7], Ensembl v84[9], Ensembl Plants v31[10], Ensembl Fungi v31[10] and Ensembl Bacteria v31[10])

物种	拉丁学名	基因组大小(bp)	基因组版本号	蛋白基因数量(个)	初始基因组公布时间
酿酒酵母	*Saccharomyces cerevisiae*	12157105	R64-1-1	6692	1997.05
大肠杆菌K-12	*Escherichia coli K-12*	5277676	HUSEC2011CHR1	5494	1997.09
秀丽线虫	*Caenorhabditis elegans*	100286401	WBcel235	20447	1998.12
黑腹果蝇	*Drosophila melanogaster*	143725995	BDGP6	13918	2000.03
拟南芥	*Arabidopsis thaliana*	119667750	TAIR10	27416	2000.12
人类	*Homo sapiens*	3209286105	GRCh38	19815	2001.02
裂殖酵母	*Schizosaccharomyces pombe*	12631379	ASM294v2	5145	2002.02
粳稻	*Oryza sativa Japonica*	374424240	IRGSP-1.0	35679	2002.04
小鼠	*Mus musculus*	2730871774	GRCm38	21971	2002.12
大鼠	*Rattus norvegicus*	2870184193	Rnor_6.0	22277	2004.04
鸡	*Gallus gallus*	1046932099	Galgal4	15508	2004.12
黑猩猩	*Pan troglodytes*	3309577922	CHIMP2.1.4	18759	2005.09
狗	*Canis familiaris*	2410976875	CanFam3.1	19856	2005.12

转录因子能够招募一些共激活剂(coactivator)等因子来增强基因的表达，而负调节的转录因子则与一些辅阻遏物(corepressor)等因子互作来减少基因的转录. 这两种类型的转录因子可以通过竞争结合位点行使自己的功能，它们能够对细胞间的信号转导和环境等作出不同的应答，从而实现高度动态的基因表达调控过程. 几乎所有基因的上游或者附近都存在一段调控序列，它们的突变与否对基因的表达至关重要[11]. 这些调控序列根据它们所处的位置、在转录中的功能及作用方式可分为启动子、增强子、沉默子和绝缘子等. 它们通过与各种调控蛋白(如转录因子)相互识别和作用，可以对基因转录的起始、延伸速率等进行更加细致的调节. 一个基因能有超过一个启动子，启动子具有强弱之分，而增强子可以对不止一个基因发挥作用，调控序列元件这种对基因的多重调节机制大大增加了基因转录调控的灵活性. 此外，一个基因还可以有多个转录起始位点，产生不同的转录本，这种泛转录现象在人类基因转录中是广泛存在的[12].

2.2 复杂多变的基因转录后调控

基因在转录出编码蛋白质的信使RNA产物之后，往往需要做进一步的加工才能发挥其功能. 这些加工程序包括RNA的剪接与可变剪接、RNA编辑和RNA修饰等. 一个完整的人类基因有一些区段(称为外显子)有编码蛋白质的功能，而另一些区段(称为内含子)无编码功能，外显子和内含子的交替排列使基因的蛋白质编码序列不连续，因此基因的初始转录产物需要对内含子进行剪接才能形成成熟的RNA分子. 一个基因的初始转录产物在细胞不同的分化和人体不同的发育阶段，甚至是不同的生理状态下，通过不同的剪接方式，使基因内部的外显子和内含子数目、位置及长度发生变化，从而得到不同的成熟RNA产物并进一步翻译成具有不同功能的相关蛋白产物. 在人类中，约有95%的多外显子基因能发生可变剪接，并且大部分都不止一种可变剪接形式[13]，比黑腹果蝇(60.7%)[14]和拟南芥(42%)[15]都要多. RNA的可变剪接使基因实际上成为了一个复杂的转录单位，它让不到20000个人类蛋白质基因转录和加工出大量的不同的蛋白产物，以适应细胞、组织和发育特异性的需要. 目前，研究较多的含有大量变体的基因是来自黑腹果蝇的*Dscam*基因，它有24个外显子，根据其中4个外显子的多种可变形式推测，能够产生超过38000个变体，远超出了自身蛋白质基因的总数[16]. *TTN*基因是目前发现的拥有外显子数目最多(362个)的人类基因，根据可变剪接的加工方式，它产生不同变体的潜力将是非常巨大的[17]. 可变剪接在各种生物中都具有重要的生理意义，如果蝇的性别决定系统；异常可变剪接也是疾病发生的重要原因，如人类脊髓性肌肉萎缩症. 由此可见，mRNA的可变剪接模式是决定人类蛋白质组多样性的关键因素.

RNA的可变剪接只影响RNA分子本身的结构特性，与编码序列的内容无关，而RNA另一种重要的加工方式——RNA编辑可以通过改变RNA编码序列的方式来改变转录产物的信息特性，导致编码蛋白质的氨基酸序列、密码子的含义甚至是整个可读框的改变，从而使翻译出来的蛋白质与原基因编码的蛋白质不同，增加了基因产物的可变性和多样性. 例如，人载脂蛋白B(基因*APOB*)在肝脏翻译成全长的蛋白质，而在小肠里经过RNA编辑后翻译成截短的多肽[18]. 对RNA转录产物稳定性的调节也是基因转录后调控中非常重要的一环，它能够影响基因是否能表达更多的蛋白质或者只发生短暂的调控作用. 近年来，RNA尤其是信使RNA(mRNA)修饰的突破性进展揭示了mRNA上不仅存在不同类型的表观修饰，而且这些修饰与mRNA的半衰期有着非常密切的联系. 例如，除了mRNA的5′帽子结构的甲基化之外，mRNA其他部位的假尿嘧啶化和m6A修饰也能够调控其稳定性. 这些修饰部分是动态可变的，它们对细胞的发育分化及应激反应均发挥重要作用[19,20].

3 人类基因组的“暗物质”——非编码基因

非蛋白质编码RNA基因(非编码基因)，是一类以非编码RNA为终产物的基因，例如，tRNA和rRNA都是由非编码基因转录而来. 长期以来，人们一直认为生命本质和遗传多样性取决于蛋白质的种类及数目，RNA仅是遗传信息传递的中间分子，所以人们最初关注的基因只是蛋白质基因. 然而研究人员在解读人类基因组时，发现这类基因只占整体基因组序列的2%. 而98%基因组序列都是非蛋白质编码区，主要包含DNA复制和基因表达调控元件、转座子等重复序列以及大量的非编码基因. 由于非编码基因没有经典的蛋白质可读框，在基因组中难以识别和鉴定，也被称为基因组中的“暗物质”. 人类基因组中有

多少非编码基因, 它们具有哪些与蛋白质基因不同的生物学功能, 它们的重要性如何? 一直是生命科学的未解之谜.

1993年以来, 微RNA(microRNA)和小分子干涉RNA(siRNA)的发现, 揭示了RNA介导的基因表达调控新机制及生命细胞内普遍存在的RNA干涉体系; 核仁小分子RNA(snoRNA)的大量发现及其介导的RNA修饰功能阐明, 被誉为"核仁风暴"; 多种大分子非编码RNA及其表观调控功能也被相继报道. 在人类基因组计划进展的同时, 新的非编码RNA及其新的功能不断被发现, 引起了人们高度关注. 我国科学家在1998年"面向21世纪的RNA研究"的第109次香山学术会议上, 根据蛋白质基因内含子能够编码snoRNA的规律, 提出了高等生物"RNA基因数目与蛋白基因数目相当"的观点[21], 并建议迅速开展发现新的RNA基因及其功能的RNA研究计划. 2003年9月, 美国国家人类基因组研究所启动了"人类DNA元件百科全书计划"(The Encyclopedia of DNA Elements Project), 旨在解析人类基因组中98%的非蛋白质编码序列. 该计划的初步结果显示, 大约80%的人类基因组序列可以被转录, 而非像之前的认为它们仅是"垃圾DNA", 并从中鉴定出18400个非编码基因[22]. 近年来, 随着新一代高通量测序技术的快速发展, 越来越多的不同种类的非编码RNA基因被鉴定出来, 除了早期已知的rRNA, tRNA, snRNA和snoRNA之外, 近年来研究比较热门的小非编码RNA(sncRNA)、长非编码RNA(lncRNA)和环状RNA(circRNA)等都得到了大规模的鉴定. 在sncRNA中, 研究较为透彻的包括microRNA, siRNA和与Piwi蛋白互作的RNA(piRNA)三大类. miRBase数据库[23](v21)收录的人类microRNA前体基因共有1881个, 而最近的研究[24]又鉴定了3494个新的microRNA前体, 其中大部分都是人类特异表达的或具有组织特异性的. 内源性的siRNA和piRNA主要表达于胚胎干细胞和生殖细胞, 而piRNA是动物界特有的一类sncRNA, 它主要以成簇的方式存在于基因组中. 有研究表明人类piRNA的数目约有4万个[25], 而小鼠piRNA多达80万以上[26], 根据人类和小鼠piRNA生成方式的相似性, 研究人员推测人类piRNA的数目也许远不止4万个.

lncRNA是非编码RNA家族中的一大类, 它的鉴定及功能研究近年来呈现爆发式的发展. 与蛋白质基因相比, 大部分lncRNA序列保守性不高. 它倾向于时空特异性表达, 表达水平相对较低. 目前lncRNA鉴定的方法及筛选过滤的条件较多, 不同的研究得到的lncRNA数据集也不尽相同. 如deepBase v2.0数据库中所鉴定的人类lncRNA数量约有19000个[27], 而另一项通过使用大样本的研究将这一数字提高至将近60000个[28]. circRNA是一类比较特殊的非编码RNA, 它采用共价闭合的方式将其两端连接起来. 它高度富集于大脑, 有研究人员从该组织中发现人类基因组中约有18000个座位能产生超过65000个circRNA变体[29]. 假基因(pseudogene)也是非编码RNA家族中的重要成员, 它是在进化过程中丧失了原有基因的功能如蛋白编码能力, 但其核酸序列与有功能的蛋白基因相似的一类非编码RNA. 基因组序列上的点突变、插入缺失、逆转录转座等现象是假基因生成的重要原因, 它们的数目约在13000~14500[30]. 假基因具有调控作用, 包括作为前体产生新的小分子非编码RNA的功能正在被揭示. 例如, 假基因来源的siRNA参与基因表达调控在各种真核生物中普遍存在[31,32].

由此可见, 人类非编码基因的数目远大于编码蛋白质基因的数目. 由于人类等高等生物庞大的基因组及其难以想象的编码能力, 目前已发现的人类基因并不完全, 甚至可能还是"冰山一角". 最新的研究表明, 许多lncRNA(包括一些circRNA)具有编码寡肽的能力[33], 也可以成为sncRNA的前体[34], tRNA等组成型表达RNA在应激条件下可以被切割加工产生新的具有调控功能的非编码RNA[35]; 而蛋白质基因的内含子和UTR区域也能产生新的调控非编码RNA[36]; circRNA大部分来自蛋白基因, 这些基因在转录加工成环后就失去了翻译的能力, 变成了新的非编码RNA[37]. 这两种不同类型基因的相互转化说明了基因组中的一段DNA序列具有编码多重功能产物的潜力, 这不仅增加了对基因含义理解的难度, 也模糊了基因组中功能序列区域的界限. 人类不同组织和细胞是怎样根据自身需要来阅读基因组序列并对其进行功能区域的划分目前仍是一个谜.

为什么人类会有数量如此庞大的非编码基因呢? 最新的研究表明, 非编码RNA虽然不编码蛋白, 但是却以蛋白机器的组织者或调控分子的身份, 参与了几乎所有的细胞活动. 非编码RNA通过相互作用或者与蛋白形成不同功能的生物大分子复合物来操纵着基因的功能, 构成了一个由非编码RNA与蛋白

协同作用的遗传信息表达调控网络. 大量的非编码RNA已被证实在细胞中行使不同的功能, 包括基因的沉默和表达、RNA的剪接和可变剪接、RNA修饰和编辑、蛋白质的翻译、信号转导、基因组稳定性、表观与获得性遗传、细胞分化、免疫应答等绝大多数的生理和病理过程. 从进化的角度看, 许多非编码基因具有的保守性较低, 呈现明显的物种特异性, 提示了非编码基因的变化与生物本质和多样性有着非常紧密的联系. 人类等高等生物拥有比其他物种更多的非编码基因, 暗示了非编码基因对于增加生物调控复杂性、精密性以及物种特异性方面起着关键的作用.

4 结语

虽然人类基因组中目前鉴定的蛋白质基因仅有不到两万个, 甚至比秀丽线虫的还要少, 但人类机体的复杂性并不仅取决于蛋白质基因数目, 而是主要依靠遗传信息的表达调控. 这一调控主要有两种策略: (ⅰ) 基因的转录和转录后调控. 控制这些基因转录的主要因素是一些遗传和表观遗传调控因子等. 人类基因中含有数目众多的内含子序列, 机体通过对基因转录的mRNA前体中内含子和外显子的选择性剪接加工, 从一个基因可以产生几种至几十种蛋白同源异形体, 大大增加了人类蛋白质组的多样性. (ⅱ) 在人类基因组的非蛋白质编码区中, 存在着大量的非编码RNA基因, 它们占人类基因组转录产物的98%以上, 参与了整个遗传信息的维持、基因表达调控及细胞功能复合物构成.

总而言之, 人类基因组是一个高度结构化的蛋白质-RNA基因机器. 近年来, 新的非编码基因的大量发现及其在各种生物学过程中的重要功能, 不断开拓和更新现有分子生物学知识体系, 从一种不同于经典的蛋白质基因的角度来诠释人类基因组的结构和功能. 人们对自身的认识, 不仅需要知道基因的种类和数目, 而且还必须深入地了解基因的功能及其表达调控的规律. 21世纪生命科学的一项重大任务就是全面系统地挖掘人类等高等生物基因组中新的非编码基因及其功能网络解析.

参考文献

1 Kauffman S A. Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol, 1969, 22: 437–467
2 Ohno S. An argument for the genetic simplicity of man and other mammals. J Hum Evol, 1972, 1: 651–662
3 Goodfellow P. A big book of the human genome: Complementary endeavours. Nature, 1995, 377: 285–286
4 Venter J C, Adams M D, Myers E W, et al. The sequence of the human genome. Science, 2001, 291: 1304–1351
5 Lander E S, Linton L M, Birren B, et al. Initial sequencing and analysis of the human genome. Nature, 2001, 409: 860–921
6 The International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature, 2004, 431: 931–945
7 Harrow J, Frankish A, Gonzalez J M, et al. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res, 2012, 22: 1760–1774
8 Ezkurdia I, Juan D, Rodriguez J M, et al. Multiple evidence strands suggest that there may be as few as 19000 human protein-coding genes. Hum Mol Genet, 2014, 23: 5866–5878
9 Yates A, Akanni W, Amode M R, et al. Ensembl 2016. Nucleic Acids Res, 2016, 44: D710–D716
10 Kersey P J, Allen J E, Christensen M, et al. Ensembl Genomes 2013: Scaling up access to genome-wide data. Nucleic Acids Res, 2014, 42: D546–D552
11 King M C, Wilson A C. Evolution at two levels in humans and chimpanzees. Science, 1975, 188: 107–116
12 Forrest A R, Kawaji H, Rehli M, et al. A promoter-level mammalian expression atlas. Nature, 2014, 507: 462–470
13 Pan Q, Shai O, Lee L J, et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet, 2008, 40: 1413–1415
14 Graveley B R, Brooks A N, Carlson J W, et al. The developmental transcriptome of *Drosophila melanogaster*. Nature, 2011, 471: 473–479
15 Filichkin S A, Priest H D, Givan S A, et al. Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. Genome Res, 2010, 20: 45–58
16 Schmucker D, Clemens J C, Shu H, et al. *Drosophila* DSCAM is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell, 2000, 101: 671–684

17 Li S, Mason C E. The pivotal regulatory landscape of RNA modifications. Annu Rev Genomics Hum Genet, 2014, 15: 127–150
18 Chiu Y L, Greene W C. The APOBEC3 cytidine deaminases: An innate defensive network opposing exogenous retroviruses and endogenous retroelements. Annu Rev Immunol, 2008, 26: 317–353
19 Schwartz S, Bernstein D A, Mumbach M R, et al. Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. Cell, 2014, 159: 148–162
20 Aguilo F, Zhang F, Sancho A, et al. Coordination of m(6)A mRNA methylation and gene transcription by ZFP217 regulates pluripotency and reprogramming. Cell Stem Cell, 2015, 17: 689–704
21 Jin Y X. The 109th Xiangshan conference: Brief introduction of RNA researches towards the 21st century. Acta Bioch Bioph Sin, 1999, 31: 119–123 [金由辛. 109 次香山学术讨论会——“面向 21 世纪的 RNA 研究”简况. 生物化学与生物物理学报, 1999, 31: 119–123]
22 The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature, 2012, 489: 57–74
23 Kozomara A, Griffiths-Jones S. miRBase: Annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res, 2014, 42: D68–D73
24 Londin E, Loher P, Telonis A G, et al. Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. Proc Natl Acad Sci USA, 2015, 112: E1106–E1115
25 Girard A, Sachidanandam R, Hannon G J, et al. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature, 2006, 442: 199–202
26 Leslie M. Cell biology. The immune system’s compact genomic counterpart. Science, 2013, 339: 25–27
27 Zheng L L, Li J H, Wu J, et al. deepBase v2.0: Identification, expression, evolution and function of small RNAs, LncRNAs and circular RNAs from deep-sequencing data. Nucleic Acids Res, 2016, 44: D196–D202
28 Iyer M K, Niknafs Y S, Malik R, et al. The landscape of long noncoding RNAs in the human transcriptome. Nat Genet, 2015, 47: 199–208
29 Rybak-Wolf A, Stottmeister C, Glazar P, et al. Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. Mol Cell, 2015, 58: 870–885
30 Pei B, Sisu C, Frankish A, et al. The GENCODE pseudogene resource. Genome Biol, 2012, 13: R51
31 Wen Y Z, Zheng L L, Liao J Y, et al. Pseudogene-derived small interference RNAs regulate gene expression in African *Trypanosoma brucei*. Proc Natl Acad Sci USA, 2011, 108: 8345–8350
32 Tam O H, Aravin A A, Stein P, et al. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature, 2008, 453: 534–538
33 Ji Z, Song R, Regev A, et al. Many lncRNAs, 5′UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife, 2015, 4: e08890
34 Wilusz J E, Freier S M, Spector D L. 3′ end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. Cell, 2008, 135: 919–932
35 Li Y, Luo J, Zhou H, et al. Stress-induced tRNA-derived RNAs: a novel class of small RNAs in the primitive eukaryote *Giardia lamblia*. Nucleic Acids Res, 2008, 36: 6048–6055
36 Chirn G W, Rahman R, Sytnikova Y A, et al. Conserved piRNA expression from a distinct set of piRNA cluster loci in eutherian mammals. PLoS Genet, 2015, 11: e1005652
37 Guo J U, Agarwal V, Guo H, et al. Expanded identification and characterization of mammalian circular RNAs. Genome Biol, 2014, 15: 409

Summary for “人类基因知多少”

What don’t we know about human genes?

LIU Shun & QU LiangHu*

State Key Laboratory of Biocontrol, Sun Yat-sen University, Guangzhou 510275, China
* Corresponding author, E-mail: lssqlh@mail.sysu.edu.cn

In the past fifty years, biologists have begun to estimate protein-coding capacity of human genomes and the estimated human gene number fluctuated in a shrunken trend, ranging from two million to 25000 recognized by the Human Genome Project. This number fell to 19000 in recent studies, which suggested that human genes were even less than the nematode worm *Caenorhabditis elegans*. Apparently, the complexity and flexibility of higher mammal genomes are far more underestimated than they were once considered, which cannot be merely interpreted as the protein-coding gene counts. Scientists now hold the belief that the widening differences among higher organisms are primarily caused by the regulation of gene expression at the molecular levels, including transcriptional regulation and post-transcriptional regulation. With regard to human genome, two major strategies are for these processes. One is through the alternative splicing of exons and introns of pre-mRNAs transcribed from human genome, one gene may produce multiple protein isoforms, thus greatly increased the complexity of proteome. The phenomenon, over the past years, has unambiguously become one of the main reasons why human genome manifests such complexity with so few protein-coding genes. The second, there actually exist an enormous amount of active non-coding RNAs (ncRNAs) from non-protein coding regions that account for approximately 98% of the human genome, which form a highly intricate RNA regulatory network to make human genome more complicated. With the implementation of the encyclopedia of DNA elements (ENCODE) project, biologists surprisingly find that the ncRNA species are diverse, including snoRNAs, microRNAs, piRNAs, lncRNAs and circRNAs. They take part in maintaining the whole genetic information, regulating gene expression and constituting functional complexes in cells. Besides, novel classes of ncRNAs and various *cis*-RNA elements are expected to be discovered and identified. All together raise the fact that the human genome can be divided into many DNA regions which harbor potentials of transcribing multi-functional RNA products. But how transcription machinery determines which section of DNA sequences to read as multi-functional at particular time point still remains a mystery. Although the task of perceiving the significance of ncRNA’s role that ncRNAs play is just beginning, further studies on the structure and function of these RNAs will facilitate the understanding human genes. In summary, human genome is operated as highly sophisticated protein and RNA-producing machinery, which contains huge amount of ncRNA genes besides the protein-coding genes. To understand the operation of human genome, we not only need to clarify the variety and counts of genes, but also need to explore their function and expression regulation.

human genome, protein-coding gene, regulation of gene expression, non-coding RNA, regulatory network

doi: 10.1360/N972016-00761

屈良鹄

博士，国家杰出青年科学基金获得者，教育部“长江学者奖励计划”特聘教授，中山大学“有害生物控制与资源利用国家重点实验室”主任，中国生化与分子生物学学会 RNA 专业委员会主任. 主要从事 RNA 信息学、RNA 生物学及非编码基因资源与技术等方面研究. 先后主持国家自然科学基金重点项目、国家重点基础研究发展计划(“973”)重要科学前沿项目以及中美和中法等国际合作等项目，在国际重要杂志上发表论文 160 余篇，获得国家自然科学奖二等奖 1 项.