229－242.
[30] Li X.B., et al.,1995, *Plant Cell Rep.*,**15**:97－101.
[31] Puddephat I.J., et al., 1999, *J-Hotic-Sci-Biotech.*, **24**: 714－720.
[32] 张海燕等, 1999,中国激光,**26**(11):1053－1056.
[33] Halfhill M.D., et al., 2001, *Theo. And Appl. Gene.*, **103**:659－667.
[34] SeⅡ J.,et al.,1995, *Plant Cell Rep.*,**14**:620－625.
[35] SHEN Y.G., et al., 2002, *Acta Botanice Sinica*, **44**: 956－962.
[36] Michael W.L., et al., 1995, *Plant Mole. Biol.*,**109**: 1389－1394.
[37] Henzi M.X., et al., 2000, *Plant Cell Rep.*, **19**: 994－999.
[38] Pius P.K.,et al.,2000, *Plant Cell Rep.*,**19**:888－892.
[39] 徐淑平等,2002, 植物生理和分子生物学报, **28**(4):253－260.
[40] Block D.,et al., 1991, *Theor. Appl. Genet.*,**82**:257－263.
[41] Neve D.M.,et al., 1997, *Plant J.*, **11**:15－29.

科学史简介

基因回眸与展望

崔 映 宇
(中山大学生命科学学院 广州 510275)

摘 要 回顾基因由萌芽、产生到不断完善的过程,阐述其在不同历史时期的内涵与外延,结合遗传学的发展谈人们对其认识的不断深化,并联系学科发展前沿和最新研究成果展望其发展趋势。

基因是遗传学最基本的概念,1909年,Johannsen W.L.首次提出,泛指控制生物性状、又按孟德尔规律传递的遗传因子。分子遗传学建立后,又被定义为"具有特定遗传效应的DNA片段"。近一个世纪来,随着技术进步和研究的深入,基因在内容和形式上不断完善,尤其是人类基因组计划实施以来,又被赋予新的内涵。因此回顾其产生和发展的历史,把握其与时俱进的脉络,并展望其前景颇有意义。

一、萌 芽

基因的产生和发展离不开生产实践。19世纪中叶,孟德尔(Mendel G.)通过对豌豆杂交长达8年的观察和总结,于1865年报告了其结果,翌年发表了题为"植物杂交实验"的论文,首次提出遗传因子控制生物性状假说,并揭示其传递规律——孟德尔定律[1]。他所指遗传因子,即基因的萌芽。

二、产 生

1900年,H. de Vires、C. Correns和E. Tschermak Von分别在月见草、玉米和豌豆的杂交实验中证实孟德尔定律,标志着遗传学的诞生。1906年,Bateson W.正式提出"遗传学"概念。1909年,Johannsen W.用"基因"取代"遗传因子",提出"基因型"和"表现型",强调表现型是基因型和环境相互作用的结果[2]。这表明自"基因"问世起,人们就注意从事物间的互作和联系来研究它。此时,"基因"仅是一个抽象的符号。

三、经典基因概念的确立[3]

基于杂交实验中遗传因子的行为与配子形成和受精过程中染色体的行为完全平行,1903年,Sutton W.和Boveri T.大胆假设,认为遗传因子位于染色体上,催生了细胞遗传学(cytogenetics)。

1910年,Morgan T. H.等发现控制果蝇白眼突变性状的w基因X连锁现象,提出"特定基因与特定染色体上的特定位置相连锁"的观点,即"摩尔根连锁规律",证明"基因线性排列在染色体上,并占据一定位置",绘制出标示基因在染色体上位置的"染色体图",正式建立"染色体遗传理论",细胞遗传学

E-mail:yingyu－cui@.sohu.com

正式诞生。

1926年，Morgan发表《基因论》，首次提出“三位一体”的基因概念，认为：“基因首先是一个功能单位，能控制蛋白质的合成，从而控制性状发育；其次是一个突变单位，一定条件下，野生型基因能突变成相应的突变型基因，从而表现出变异类型；再就是一个重组单位，两不同基因可重组，产生与亲本不同的新类型；基因在染色体上按一定顺序、间隔一定距离线性排列，各自占有一定的区域。”一句话，基因是染色体上集“功能单位”、“突变单位”和“重组单位”于一身，占据一定空间的实体，也就是说，基因不再是抽象的符号，而是稳定、不可分割的客观存在，标示着一个个有机的化学实体。基因物质性的确立，实现了基因概念由抽象到具体的首次突破，为基因结构和功能的研究奠定了基础。

四、经典基因概念的发展[3]

1928年Griffith F.的肺炎双球菌转化实验和1944年Avery O. T.等人的研究证明DNA是遗传物质，基因的化学本质是DNA，基因是DNA上的功能单位。随后，基因得以进一步发展。

1.“一个基因一个酶”

1941年，Beadle G. W.和Tatum E. L.等以红色链孢霉为材料，研究基因的生理、生化功能，证明基因通过酶起作用，提出“一个基因一个酶”假说，认为生物的性状可分为许多单位性状，每个单位性状均受一种酶影响，而酶决定于基因。后来，发展成“一个基因一条多肽链”。

2. 顺反子[4]

1953年，Watson和Crick提出DNA双螺旋模型，对其分子结构、自我复制、相对稳定性、变异性以及其如何贮存和传递遗传信息等问题进行了合理解释，明确了基因是DNA分子中的一个片段，这标志着分子生物学时代的来临和分子遗传学的诞生。于是，基因结构和功能的分子水平研究成为时尚。

1955年，S. Benzer通过“互补实验”分析E. *coli* T4噬菌体rⅡ区基因的精细结构，认为基因可进一步再分，提出顺反子、突变子和重组子概念：(1)顺反子是一个遗传功能单位，一个顺反子决定一条多肽链，其平均长度约1000 bp，包含许多突变子和重组子；(2)突变子是能发生突变的最小单位，可以是一个或几个核苷酸，其中任一核苷酸的改变都可形成一个突变子；(3)重组子是能够交换的最小单位，有起点和终点，各个重组子之间均有一定的距离，彼此间能发生交换，可以是单核苷酸的互换，也可是几个核苷酸的重组，若为前者，重组子就是突变子。

总之，顺反子是基因的同义语，不同之处在于它含许多突变子和重组子，而后两者实际上都可以是一个核苷酸对。顺反子把基因具体化为DNA分子的一段核苷酸序列，能储存和传递遗传信息，是决定一条多肽链的完整功能单位，但又是可分的，其核苷酸可以独自发生突变。顺反子否定了“基因是突变、重组的最小单位”的概念，标志着人类对基因认识上“由不可分到可再分”的第二次突破，为基因的表达调控研究奠定了理论基础。

3. 操纵子[4]

1961年，Jacob F.和Monod J. 研究E. *coli* 乳糖代谢机制，提出乳糖操纵子模型，首次揭示了原核基因表达的调控，为基因表达调控机理的深入研究奠定了基础，同时丰富了基因内涵。他们认为基因不仅是传递遗传信息的载体，还具有调控其他基因表达活性的功能。

他们将乳糖操纵子的基因分为结构基因、调节基因、操纵基因及启动基因，认为：(1)结构基因直接控制蛋白质合成，是决定一条多肽链的功能单位，在一个操纵子中往往多个结构基因呈连锁关系；(2)操纵基因居结构基因之前，能同阻遏物结合，可间接控制结构基因的表达；(3)启动基因又居操纵基因之前，为RNA多聚酶结合位点，RNA多聚酶必须越过操纵基因才能对结构基因进行转录，继而翻译成相应蛋白；(4)调节基因编码的可扩散性阻遏蛋白，通过与操纵基因结合与否调控整个操纵子。显然，“操纵基因”和“启动基因”与顺反子相悖，因其不编码可在细胞内扩散的产物，其突变拷贝无法用反式的野生拷贝进行互补。后来，人们发现乳糖操纵子“操纵基因”的跨度仅约26bp[5]。基于此，分子遗传学家剥夺了“操纵基因”和“启动基因”的基因冠名权，分别代之以“操作子”(operator)和“启动子”(promoter)。

综之，无论结构上，还是功能上，“基因都是可分的”；基因不仅能单独起作用，而且相互间还存在彼此制约的调控网络，每个基因都在该系统中发挥各自功能；基因可以有自身产物，也可以没有。一言以蔽之，操纵子是原核生物遗传信息传递、表达和调控的统一体。

4. 跳跃基因

跳跃基因是指能够改变自身位置的一段DNA序列，也叫“转座元”，可在染色体内或不同染色体间

移动，其本身可编码转座酶(transposase)，也含有非蛋白编码序列。

1951年，“玉米夫人”McClintock B.首次提出在染色体上可移动的“控制元件”，发现玉米的“激活-解离”(Ac-Ds)系统，奠定了“基因可动论”的基础。可惜，当时学术界因受经典基因概念的禁锢而未达成共识！直至陆续在大肠杆菌、噬菌体、酵母、果蝇以至哺乳动物体内发现可动遗传元件，其“基因可动论”才得到公认，并荣获1984年度诺贝尔奖。

跳跃基因的发现使人们认识到基因不是稳定、静止不动的实体，而是一段在结构上有明确界限的DNA序列，可通过自身运动调节相关基因活性。跳跃基因概念的确立标志着人类对基因认识上由“不可动”到“可移动”的第三次突破。

至此，“三位一体”的经典基因概念框架基本被打破，仅保留其独立的功能单位实体未变。基因由“由不可分到可再分”、“由不可动到可移动”观念上的突破，反映了人类对基因认识的不断深化。

五、近代基因概念的多元化[3,4,6]

20世纪70年代，DNA体外重组技术和基因工程技术成熟，人们对基因研究更加深入，涌现出断裂基因、重叠基因、假基因、重复基因、超基因和印记基因等多元概念。

1. 断裂基因

断裂基因，又称不连续基因，由编码序列和非编码序列相间排列构成，编码序列叫外显子(exon)、非编码序列为内含子(intron)，因为编码序列存在于成熟的RNA中，翻译成蛋白质后，可呈现出一定性状，而非编码序列在RNA剪接过程中被切除。

1977年，Bergets M.，Moore C.和Shap P.A.等在研究腺病毒时，发现其头部蛋白的编码基因中存在非编码序列，首次提出断裂基因概念；随后，Flavell & Jeffreys和Chamon分别在进行兔的β-球蛋白基因和鸡的输卵管卵清白蛋白基因研究时，证实这两个基因中也存在非编码序列。近年研究发现，少数细菌基因中也有内含子序列，这为研究基因演化提供了启迪。外显子保守性要远高于内含子，由于内含子突变最终不出现在mRNA中，不受选择压力的影响，在基因进化中的作用可能更大，积累了突变的内含子可能演化成新基因，而且内含子的存在有利于不同基因外显子间的重组，使RNA顺式剪接(cis-splicing)和反式剪接(trans-splicing)成为可能，更增加了真核基因组本来已经巨大的编码潜能，可见内含子的存在对生物进化可能具有重大意义。

2. 重叠基因

重叠基因，指共用一段DNA序列的两个或两个以上结构基因的互称，这与“互不沾染、单个分离”的传统基因概念相悖。

1973年，Weiner等在研究大肠杆菌的RNA病毒Qβ时发现，编码其外壳小蛋白的基因与编码决定其感染能力的大蛋白的基因在指导蛋白质的合成时开始于同一起始点，即编码大蛋白的基因包含了编码小蛋白的基因，首次提出重叠基因概念；1977年，Sanger等在研究噬菌体ΦX174的DNA核苷酸序列和基因排序时，再次证实并完善重叠基因概念。

目前，在果蝇、线虫以及人体内也有发现。重叠基因大致有四种情况：(1)套叠基因(nested gene)，即一个基因的序列完全落在另一个基因序列之中，其读码框可以相同，也可不同。如噬菌体ΦX174的DNA中B基因包含在A基因内，E基因包含在D基因内，但其读码框不同；(2)两个基因仅有一个核苷酸重叠，即前一个基因的终止点和后一个基因的起始点重叠。如ΦX174 DNA中D基因的终止码的第三个核苷酸是J基因起始码的第一个核苷酸；(3)三层重叠基因，即三个不同基因共用一段核苷酸序列。如G_4病毒编码K蛋白的基因k的核苷酸序列有两个位置重叠着三个基因，第一个位置为五核苷酸TGATG，分别为K、A、B三个基因编码，另一个位置为四核苷酸ATGA，分别为K、A、C三个基因编码；(4)双链DNA分别作模板，按不同方向转录出独特的mRNA。例如人的Ⅰ型神经纤维瘤(neurofibromatiosis typeⅠ，NFⅠ)基因，约300kb，其第一个内含子中发现三个编码蛋白质的基因，但其转录方向恰与NFⅠ基因的转录方向相反，即NFⅠ基因的无意义链却是这三个基因的有意义链，可见内含子、外显子的划分是相对的、有条件的；这种反向转录的情况表明，重叠基因的转录各自独立、互不依赖。

基因组较小的病毒，核苷酸数目不多，重叠基因的存在使较小的空间容纳尽可能多的信息，高效经济地利用核苷酸，指导合成尽可能多的蛋白质，这对其生存无疑有利；而高等生物基因组中，尽管有大量的非蛋白编码序列的存在，依然利用基因的内含子编码另一些蛋白，很可能具有人类尚未理解的进化意义。

3. 假基因

假基因，看似正常基因，却不能表达任何RNA

或蛋白质，包括已知功能基因的残存拷贝、散在分布的长细胞核因子(long interspersed nuclear elements, LINEs)和散在分布的短细胞核因子(short interspersed nuclear elements, SINEs)[7,8]，后两者为哺乳动物基因组中的重复序列，其中 LINE 源于 RNA 聚合酶Ⅱ的转录产物，是一类可自主转座的反转录转座子，而 SINE 源于 RNA 聚合酶Ⅲ的转录产物，则是一类非自主转座的反转录转座子。假基因通常以ψ表示，其核苷酸序列与相应正常基因的同源度可达 75% - 80%，由于突变阻碍了其自身表达，从而失去正常转录功能。1977 年，Jacq C. 等在研究非洲爪蟾的 5S RNA 基因时，首先发现假基因[9]，后来人们陆续发现珠蛋白基因簇、免疫球蛋白基因簇、组织相容性抗原基因簇中也存在，而且通常散布于有活性的功能基因之间。此类基因仅限于真核生物，且大部分位于染色体上正常基因的附近，也有的位于不同的染色体上。假基因的结构特点有：(1)不同部位有不同程度的缺失或插入；(2)缺少正常基因的内含子和启动子；(3) 5' 端都有真核 mRNA 分子特有的 AATAAA 信号，造成转录启动区的缺陷；(4)两侧有顺向重复序列。可见，假基因是正常基因转录出的 mRNA 经加工以后，再反转录成 cDNA，然后整合到染色体上的基因组中去的结果，故又称"加工基因"(processed gene)，基于此，也常被视为一种非自主的反转录转座元件。但近年研究表明，假基因可以分为"加工的"和"未加工的"两类[10]，前者从 mRNA 反转录而来，不含任何内含子结构，而后者可能源于基因重复。2002 年，Harrison et al. 等发展了一种筛查和识别人类基因组中假基因的新方法，依据多聚腺苷化的有无以及与功能基因除去内含子后的序列连续同源度是否大于 70%，发现人类基因组最先完整测序的 21 和 22 号染色体总共有 189 个加工假基因，195 个未加工假基因，还有 70 个假基因性的免疫球蛋白基因片段，并据以推断整个人类基因组可能大约有近 20000 个假基因，其中一半以上为加工假基因。他们还发现在这两条染色体的着丝点附近有显著冗余的相关基因的假基因，提示人类基因组中假基因热点的存在；而加工假基因群分布则相当均匀，其中 22 号染色体假基因群以免疫球蛋白基因片段为主[10]。先前，人们认为假基因无生物学功能，不受进化的选择压力，可积累突变，常常同时存在三种终止码序列，由于缺乏选择压力，它可能会由于随机突变的积累而变得面目全非，因而被视为垃圾。但近来科学家们认识到，这对假基因有失公平，因为实验中观察不到其编码产物并不能证明它在生物体内从未编码功能产物，而且不编码功能产物也并不能排除其可能拥有的其它功能；同时，排除仅仅基于序列信息的蛋白表达也是不客观的，因为 DNA 信息可被 RNA 编辑所改变。1996 年，*N. Trabesinger-Ruef* 等报道基因转换可能参与了核糖核酸酶进化中假基因的形成，认为假基因曾被修补和表达，是产生新的生物功能大分子的一种原始资料[11]；2000 年，Woodmorappe 提出功能假基因的新概念[7]，对传统假基因无功能概念提出挑战！假基因由功能基因演变而来，其存在本身就是其功能活性的明证，若是无用的，自然选择早已将其淘汰，因为细胞合成 DNA 的能量代价极其昂贵。此外，近来研究表明，哺乳动物基因组中 SINE 家族的一员 Alu 序列能调控基因活性的增强和静息，或作为一个受体结合位点发挥自身作用。细胞遗传学水平观察，Alu 序列集中在基因转录最活跃的染色体区段内，目前在所有已知的基因内含子中，几乎都发现了 Alu 序列，此乃其他假基因功能性的一个先兆。研究还发现中国仓鼠 Alu 类家族(Alu-equivalent family)的一些成员，位于其他转录单位附近时，能被转录成单独的 RNA 分子。最近，S. Hirotsune 等就利用转基因小鼠发现一个表达的假基因能调控其同源编码基因的 mRNA 稳定性[12]。可以预见对人类基因组中散布的一亿个 Alu 拷贝的深入研究将进一步揭示这类元件的调控功能。鉴于不断有假基因功能被实验证实，这类在不同机体基因组中广布的遗传元件似乎不是无意被创造的，但其具体功能及存在的意义有待深入探讨，并且可能会成为基因组研究中的一个潜在热点。

4. 重复基因

重复基因，指在真核基因组中具有一份以上拷贝的基因，这些拷贝或在一条染色体上串联排列，或分散到多条染色体上，包括寡拷贝基因和多拷贝基因，前者有人的珠蛋白基因、癌基因及某些假基因，后者包括组蛋白基因、rRNA 基因等，其结构基础分别是 DNA 的低度重复序列和中度重复序列。

重复基因的存在可增加基因剂量，提高基因表达效率，而且不同时空表达的具有一定结构差异的产物可以满足生物个体发育不同时期的需要。

5. 超基因

超基因，指在真核生物中，作用于一种或一系列性状的几个紧密连锁的基因，类似于原核生物中的操纵子，如人类的血红蛋白基因簇。功能相同或相

关的许多基因聚集而成基因簇，可以是基因重复产生的两个相邻的相关基因，也可以是许多个相同的基因首尾衔接的串联排列，如 组蛋白基因和 rRNA 基因，其中也可以有假基因。

一个祖先基因经过重复和变异而产生的一组基因，称基因家族，结构基因家族中各成员通常具有相关甚或相同的功能。共同的祖先基因通过各种变异产生的结构大致相同，而功能不尽相似的一大批基因，虽属不同基因家族，可总称为一个超家族。在成簇的基因家族中，通过染色体重排而分散到其他位置上的成员，称为孤儿基因。

6．印记基因

印记基因(imprinted gene)，指功能受到双亲基因组的影响而被打上雌雄亲本特异标记的基因，为哺乳动物的基因组特有，实质是双亲相应基因的甲基化程度不同。印记产生于生殖细胞发育期，恒定于胚胎发育期，消除于性腺形成期并同时产生自身新的印记。在胚胎和成体的二倍体体细胞中，源于父本或母本的等位基因有选择性表达特征。印记基因表现为单亲依赖性遗传，与“无论遗传物质来自双亲中的哪一方，均具相同的表型效应，等位基因不会因为位于不同亲代来源的染色体上而产生不同的效应”的孟德尔遗传规律相悖，似乎能作为不同亲源等位基因的可识别标记，故称。印记基因的概念萌生于杂交双亲遗传效应的不同，最早可追溯到数世纪前，伊拉克人发现驴和马正反交，其子代有驴骡和马骡的差异[13]。但直到 20 世纪 50 年代末，才首次发现果蝇白眼基因座的一些等位基因在子代中的不同表达，取决于该等位基因是来源于父本还是母本；80 年代中期，科学家研究小鼠胚胎发育时发现雌雄个体的基因组对胚胎发育的作用有差异；1991 年，第一个印记基因 IGF_2 在小鼠基因组中被发现[14,15]。

印记基因主要有以下特点：①印记基因遍布于整个基因组中，人类基因组中约有 100 个印记基因；②有些印记基因聚集成簇，形成染色体印记区；③尽管有些印记基因紧密连锁，但却表现出不同的印记效应。如：小鼠 7 号染色体远端的 IGF_2 基因和 H_{19} 基因连锁，但 IGF_2 是母亲印记失活基因，H_{19} 则是父亲印记失活基因；④印记基因的内含子一般均较小，“内含子/外显子”长度之比也较小，雄性印记基因的重组率高于雌性印记基因；⑤表达具有组织特异性。如：小鼠父本染色体上的 Ins1 和 Ins2 两个基因在卵黄中是单个等位基因表达，而在胰腺组织中则是双等位基因表达。研究表明印记基因的异常甲基化会引起人类胚胎发育畸形及肿瘤等疾病的发生[16]。

此外，病毒基因有早期基因和晚期基因之分；据基因表达受环境影响与否，有持家基因和奢侈基因之别；据基因在癌症发生中的作用，又区别癌基因和抑癌基因；甚至还有人出于对胚胎发育过程的全面理解而提出时序基因(temporal gene)、格局基因(pattern gene)、同源异形基因(homeobox gene)、选择基因(selection gene)、开关基因(switch gene)和后成基因(epigene)等等。这些冠名基因概念的提出，虽然未对基因作出创新性贡献，却从不同的视角对基因进行分类、探索和功能研究，反映了 DNA 双螺旋模型建立以来基因概念外延的扩大和人们对基因结构和功能认识的不断深化，基因概念多元化的事实说明基因并非“顺反子”一词所能简单概括。基于此，分子生物学将“基因”定义为：“产生一条多肽链或功能 RNA 所必需的全部核苷酸序列。”[17]

总之，近代基因概念强调“基因应该是能够表达和产生基因产物(蛋白质或 RNA)的 DNA 序列”，根据产物类别可分为蛋白质基因和 RNA 基因，根据产物功能又可分为结构基因和调节基因。

六、现代基因概念剖析

1990 年“人类基因组计划”的实施标志着生命科学步入基因组时代。基因组是生物体内所有基因的总称，基因组计划的主要任务是“DNA 测序和基因鉴定”，目前已由测序为主的“结构基因组”逐渐过渡到以基因及其功能认定为主的“功能基因组”研究，相关的“蛋白质组”研究也方兴未艾。这些使人类成功地将基因与特定的 DNA 片段及其产物结合起来。现代基因概念至少包括基因产物的表达、功能活性的具备、编码区和调控区的涵盖三层逻辑含义[18]。

1．基因组时代的基因定义

2001 年，人类基因组序列框架草图既已绘就，但人类基因总数至今未能确定。目前倾向认为 3－4 万条，一度曾被估计到 10 万条或更多，这种状况除估算基因的方法不同外，至少说明人们对基因概念的界定还有待商榷。

基因组时代，基因的定义主要基于“三种方法”[19]和“五个标准”[18]。三种方法是(1)cDNA 克隆和 poly(A)$^+$mRNA 的表达序列标签(EST)测序，(2)比较基因组分析鉴定保守的编码区，(3)计算机预测。

这些对高丰度、高表达和进化上保守的蛋白编码基因非常有效，但也几乎必然低估其他基因的数目，比如“非编码 RNA(ncRNA)基因”。五个标准指：(1)开放阅读框(ORFs)，通过基因组中大的开放阅读框的鉴定来发现蛋白编码基因；(2)序列特征，密码偏爱(codon bias)和剪接位点等特异序列特征助于锁定基因，运用 DNA 序列特征，计算机程序即可预测近50%的 ORFs 和20%的完整基因；(3)序列保守性，通过不同生物的多序列比对鉴定基因，物种间 DNA 序列的保守性是估计基因重要性的一种好方法，当然保守序列也有可能是调控元件；(4)转录实况，RNA 或蛋白质的表达搜索，是一种非序列基础的基因鉴定，通过微阵列杂交、基因表达的系列分析(SAGE)、cDNA 作图或表达序列标签(EST)作图来完成。目前，通过标记 cDNA 与包含人类全部染色体序列的微阵列杂交结果表明“染色体相当大的区段均能稳定表达”，然而人类对这些转录区的功能却不清楚[20]；(5)活性丧失，突变基因使它的产物失去活性也是鉴定基因的一种方法，通过基因干扰或 RNAi 实现，但许多编码序列产物的失活并不导致明显的表型改变。这些都可看作对基因内涵的进一步丰富。此外，基因鉴定中还存在部分重叠、可变剪接和假基因等问题，也影响着基因的准确计数。

2．蛋白质组时代的基因定义

1994 年，澳大利亚的 Wilkins 和 Williams 等提出“蛋白质组”的概念，将其定义为“基因组编码的全部蛋白质”，为在细胞和生物体整体水平上阐明生命现象的本质和活动规律奠定了基础，也引发了人们对基因的反思。

早在 1990 年，Kane P. M. 和 Hirata R. 等在研究单细胞真核生物酿酒酵母的 TFP1 基因时就发现了蛋白质剪接现象[21]，随后在细菌和古细菌的一些基因表达中也有发现。1994 年，Davis E. O. 等提出“切除肽”(intein)和“显现肽”(extein)概念，认为这些蛋白在翻译后自动删除 inteins，连接 exteins，才形成功能蛋白[22]。Inteins 的出现使得基因指导合成的蛋白与最终的功能蛋白不一致。值得注意的是，inteins 并不等于 introns，introns 是 DNA 中的非编码序列，而 inteins 是 DNA 中的编码序列，属于 exons 范畴，是遗传信息的外现。inteins 从蛋白质中的剪除，可视为遗传信息的再分配。在“伴刀豆球蛋白 A 原”(ConA 原)到 伴刀豆球蛋白 A(ConA)的成熟过程中，肽链的起点与终点发生改变，导致氨基酸序列的彻底重排，DNA 的遗传信息也被大幅度地剪接重排得面目全非。在此过程中，伴刀豆球蛋白 A 基因，与其说是伴刀豆球蛋白 A 的模板，倒不如说是编辑伴刀豆球蛋白 A 的一份底稿，充其量是遗传信息进行加工的一个遗传单元，在对遗传信息的加工改造过程中，DNA 的核苷酸和肽链的氨基酸序列之间的共线性被完全破坏。

蛋白质剪接与 RNA 剪接极其相似，Inteins 的存在不仅促进蛋白质分子自剪接(cis-splicing)，而且还能把两个蛋白质分子连接成一个新的蛋白质分子(trans-splicing)[23]，这就在蛋白质水平上大大增加了基因表达的多样性。基于 RNA 剪接和蛋白质剪接，一个基因两个多肽链，两个基因一条多肽链，已是不争的事实。

鉴于从基因到蛋白质要经过各种形式的修饰与加工，如断裂基因的 RNA 剪接，模糊基因(cryptogenes)的 RNA 编辑(RNA editing)[24]等，即使稳定的基因类型，在 DNA 模板与其相应 RNA、蛋白质序列之间也往往不尽一致，因为它们都要经过一定程度的 RNA 剪裁或翻译后修饰。所有这些都说明 DNA 模板仅是一份非常粗糙的初稿，它不是一个僵化的铸模，更象一个活字版。准确地说，现在看来，中心法则不是精细、逐一地传递序列信息的通道，而是一个动态的分子遗传信息的加工流水线。基于此，基因可以定义为“进行遗传信息储存与加工的单元”[25]。

七、展　　望

21 世纪，生物遗传信息的范畴在扩大。传统遗传信息以 DNA 语言写就，DNA 上特异的碱基排列顺序即遗传信息。研究表明，人类基因组含有两类遗传信息，一类提供生命必需蛋白质的模板，称编码遗传信息，另一类提供何时、何地和如何应用编码信息的指令，为后成遗传信息(epigenetic information)。基于此，有人将 DNA 称为遗传信息的终极模板 (ultimate template)，染色质为其生理模板(physiological template)。染色质 DNA 的甲基化是后成遗传信息的主要形式，据称人单倍体基因组有5千万个 CpG 位点，有“甲基化”和“未甲基化”两种形式，这样就存在巨大的 DNA 甲基化可能的组合，可贮存大量的信息；此外，染色质另一组分组蛋白的氨基端发生多组合修饰，可调控其本身进入 DNA 的通道，不同的组蛋白氨基端修饰，对染色体结合蛋白产生协同或拮抗作用，从而调控染色质转录活

动或沉寂状态的动力学转换。研究表明，一个细胞的甲基化形式大致代表了该细胞表达特征的蓝图，组蛋白氨基端修饰的组合，则显示出一种“组蛋白密码”[26-28]，这些将显著扩大遗传密码的信息贮存。同时随研究深入，人们逐渐认识到蛋白质空间结构的特异性和重要性，基于蛋白质和酶的空间结构特异性决定其作用物、产物的特异性，有人提出“空间密码理论”，认为蛋白质也能体现并继续传递核遗传信息，谓之“蛋白质遗传”[29,30]。此外，尚有“糖密码”(sugar code or Glycome)的报道[31-33]等。这些都提示，遗传信息并非仅仅是中心法则所说的信息大分子中的一级序列，大分子“构型”(conformation)本身也是一种信息，而且可以是一种遗传信息[30]。基于此，若承认基因是遗传信息的载体，那么基因概念的外延似乎就应随着遗传信息范畴的扩大而有所拓展。

长期以来，三大遗传定律、基因突变和重组诠释着生物亲子代间表现型传递的规律。然而，近年来后成遗传的再发现打破了人们这一思维定式。后成遗传学(Epigenetics)可谓一门新兴学科，研究可遗传的、没有 DNA 序列变化的基因表达改变，为后成论(epigenesis)的发展，又称“表观遗传学”。1942 年，由 C. Waddington 首次提出，旨在研究由基因型产生表型的过程；1987 年，R. Holliday 提出，高等生物的遗传特性可从两个水平进行研究，首先是亲子代间基因传递的机制，其次是有机体从受精卵向成体发育过程中基因作用的模式。1994 年，Holliday 又指出基因表达的改变不仅发生在生物个体发育过程中，成体阶段依然存在，基于此，他认为“后成遗传学”研究细胞分化中基因表达的改变，以及基因表达既定形式的有丝分裂遗传，包括 DNA 和蛋白质相互作用的各种形式、DNA 水平的变化等。Holliday 提醒我们，发育过程中，DNA 序列可能发生永久性变化，并通过细胞分裂而传代，这是不可逆转的；但基因表达的可遗传改变在发育的后期能够逆转，有时就在减数分裂后，属非 DNA 序列差异的核遗传。目前，多数人认同后成遗传学是研究没有 DNA 序列变化、可遗传的基因表达（活性）的改变[34,35]。还有人基于 DNA 序列信息(质变)的遗传学研究，定义“后成遗传学”为研究基因表达水平（量变）信息的遗传[26]。也有人把后成遗传定义为非孟德尔遗传，或没有 DNA 序列改变的核遗传[34]。现代通常将“后成遗传学”定义为“研究可通过有丝分裂/减数分裂遗传、无需 DNA 序列改变的基因功能活性变化”的一门学科，可视为 Genetics 的姊妹学科，由 gene 到 Genetics，再到 Epigenetics，体现出人们对生物遗传本质和基因概念及属性认识的不断深化。

广义上，DNA 甲基化(DNA methylation)、基因沉默(gene silencing)、基因组印记(genomic imprinting)、RNA 剪接(RNA splicing)、RNA 编辑(RNA editing)、RNA 干扰(RNA interference)、组蛋白乙酰化(histone acetylation)、蛋白质剪接(protein splicing)等皆可归为“后成遗传”范畴，准确而言应为“后成遗传修饰”(epigenetic modification)，此类表达水平的可遗传变化，会造成基因产物的改变，最终导致表型变异，但其遗传不遵循孟德尔规律。近几年来，DNA 甲基化、组蛋白乙酰化、RNA 干扰、RNA 编辑等后成修饰机制被认为在基因激活与失活、个体发育和表型传递过程中的作用更大，于是，后成遗传学成为许多生命学科的研究前沿，更是当今遗传学和基因研究的一个热点，具有重要的理论和实际意义，具体涉及基因转录调节、染色质结构、基因完整性、动物克隆、肿瘤发生和防治等[36,37]。

后成遗传修饰提供了改变基因表达状态的方法，并可通过特定甲基化形式的拷贝遗传。尽管其稳定性机制不如 DNA 完善，但由于在自然界中，通常不是蛋白质序列，而是基因表达水平决定变异的表型[26,27,37,38]；这样如果把决定性状传递的功能单位——“基因”定义为“一段 DNA 序列”就显得过于简单化[34]，故有人认为，作为遗传功能单位的基因应超过 DNA 序列及至后成遗传修饰[35]。若用化学成分来描述，也不应定义为所有成分，而是在性状传递中起作用的那一部分[34]。据此，似乎可将基因的组分限定在 DNA 及其包装蛋白（组蛋白）水平，因为它们是基因组范围内遗传信息贮存和后成遗传信息恢复的物质基础[27, 28, 39]。

总之，后成遗传学是一个发展中的研究领域。在分子水平，后成遗传学对基因活性调控机制的研究刚刚起步，相应于遗传学研究由 gene→genome→genomics，可以预见后成遗传学将从以下几个层次和方向拓展基因功能研究领域：① Methylation→Epigenome→Epigenomics；② Glucide→Glycome→Glycomics；③ RNA→Transcriptome→Transcriptomics；④NcRNA→Ribonome→Ribonomics；⑤Protein→Proteome→Proteomics；⑥ Phenotype→Phenome→Phenomics。随着后成遗传学的研究深入，人们对基因的认识将会更加全面深刻。

八、结 语

20世纪是基因发育成熟的百年。初叶,“三位一体”的经典要义屡遭挑战;中叶,其化学本质的揭示及其双螺旋结构和半保留复制机制的确立,使“基因是具有一定遗传效应的DNA片段”成为共识,兼之由“不可分”到“可再分”、由“不可动”到“可移动”的理论突破,使经典遗传学不能解释的一些问题迎刃而解;末叶,人类基因组计划的实施和基因测序技术的进步,极大地丰富和拓展了近代基因概念的内涵和外延。本世纪初,蛋白质组研究的深入,使人们对基因概念再度反思,认识到“基因不仅仅是遗传的基本功能单位,更应该是遗传信息贮存和加工的单元”。21世纪,基因概念的外延将有可能随“后成遗传学”的发展而进一步拓展,其内涵也将随着纳米生物学(Nanobiology)[40]和量子生物学(Quantum biology)[41]的发展而在量子水平上充实完善,人们也将能更准确、更全面地揭示生物遗传变异的本质规律。

参 考 文 献

[1] 刘祖洞,1998,《遗传学》(上册,第二版)pp. 2-42,高等教育出版社,北京.

[2] 郭德栋,1991,《遗传学》pp. 1-6,东北师范大学出版社,长春.

[3] 赵寿元、乔守怡,2001,《现代遗传学》pp. 36-95,高等教育出版社,北京.

[4] 刘祖洞,1998,《遗传学》(下册,第二版)pp. 140-165,高等教育出版社,北京.

[5] 孙乃恩、孙东旭、朱德煦,2000,《分子遗传学》pp. 288-290,南京大学出版社,南京.

[6] 丁逸之,1999,《遗传工程词典》pp. 89-120,湖南科学技术出版社,长沙.

[7] Woodmorappe J., 2000, *CEN Tech. J.* **14**(3): 55-71.

[8] Walkup L.K., 2000, *CEN Tech. J.* **14**(2): 18-30.

[9] Jack C., et al., 1977, *Cell*, **12**(1): 109.

[10] P.M. Harrison, et al., 2002, *Genome Res*, **12**(2): 272-280.

[11] N. Trabesinger-Ruef et al., 1996, *FEBS Letters*, **382** (3):319-322.

[12] S. Hirotsune, et al., 2003, *Nature*, **423**(6935): 91-96.

[13] Savory T.H., 1970, *Sci. Am.*, **223**(6): 102-109.

[14] De Chiara T.M., et al., 1991, *Cell*, **64**(4): 849-859.

[15] I.M. Morison, et al., 2002, *Nucleic Acids Research*, **29** (1): 275-276.

[16] Vogelstein B., et al., 1998, *The Genetics Basis of Human Cancer*, pp. 95-107, Mc Gra-Hill, New York.

[17] 朱玉贤,李毅,1998,《现代分子生物学》pp. 185,高等教育出版社,北京.

[18] M. Snyder, et al., 2003, *Science*, **300**(5617): 258-260.

[19] S R Eddy, 2001, *Nature Rev. Genet.*, **2**(12): 919-929.

[20] J. L. Rinn, et al., 2003, *Genes Dev.*, **17**(4): 529-540.

[21] Kane P., et al., 1990, *Science*, **250**(4981): 651-657.

[22] Colston MJ, et al., 1994, *Mol Microbiol*, **12**(3): 359-363.

[23] Henry Paulus, 2000, *Annu. Rev. Biochem*, **69**: 447-496.

[24] Simpson L, et al., 2000, *Proc Natl Acad Sci U S A*, **97** (13): 6986-6993.

[25] Li Zhengang, 2002, *Biological progress*, **6**(1): 21-25.

[26] Jones P.A., 1999, *Nature Genetics*, **21**(2): 163-187.

[27] R.A. Martienssen, et al., 2001, *Science*, **293** (5532): 1070-1074.

[28] T. Jenuwein et al., 2001, *Science*, **293**(5532): 1074-1080.

[29] R. B. Wickner et al., 1999, *Microbiol. Mol. Biol. Rev.*, **63**:844-861.

[30] Li Zhengang, 2002, *Biological progress*, **6**(2): 12-17.

[31] R. D. Knight, et al., 2001, *Nature Review Genetics*, **2** (1): 49-58.

[32] D. N. Hebert, et al., 2003, *Nature Structural Biology*, **10**(6): 412.

[33] Sabine L., et al., 2003, *Nature*, **421**(6920): 219-220.

[34] C.-t. Wu, et al., 2001, *Science*, **293**(5532): 1103-1105.

[35] E. Pennisi, 2001, *Science*, **293**(5532): 1064-1067.

[36] Wolff A.P., 2001, *Oncogene*, **20**(24): 2988-2990.

[37] Robertson K.D., 2001, *Oncogene*, **20** (24): 3135-3155.

[38] Malik R., 2000, *Br. J. Cancer*, **83**(12): 1583-1588.

[39] S. Henikoff, et al., 2001, *Science*, **293** (5532): 1089-1102.

[40] Zhao W, et al., 2002, *Di Yi Jun Yi Da Xue Xue Bao*, **22** (5): 461-463.

[41] M E Núñez, et al., 1999, *Chemistry & Biology*, **6** (2): 85-97.