

基因是什么？

基因是 DNA 链上的串珠的概念正在迅速变得过时。编码蛋白质的 DNA 序列没有始端，也没有结尾。
RNA 是遗传信息组合的关键部分。

编者按：2006 年 5 月 25 日出版的《自然》杂志 (Nature) 在 New Feature 栏目刊登了题为“什么是基因？”的短文 (Helen Pearson, Nature, 2006, 441: 399~401)。该文结合近年的研究成果，对基因的概念作了新的诠释，值得引起重视。现特邀中国科学院生物物理研究所唐捷研究员将此文编译发表于此，以飨读者。

在经典遗传学中，基因是一个抽象的概念——一个能够把父母的性状传给孩子的遗传单位。从生物化学家的角度，那些性状与酶和蛋白质相关，一个基因决定一个蛋白质。在分子生物学家看来，基因是一种实体——能够转录成信使 RNA，再以此为基础合成蛋白质的一种 DNA 序列。染色体上的螺旋状 DNA 像是绳子，而基因就是绳子上的一颗颗珠子。这种看法虽然还被许多科学家认可，但分子遗传学家研究做得越深入，对基因的确切定义就越不肯定。染色体上信息的分布比以前想象的复杂得多。RNA 不是被动地将基因中的信息传递下去，而是主动地调控细胞的活动。在一些情况下，RNA 还能将信息传给下一代。一个令人惊讶的研究发现，植物可以根据从上一代遗传下来的 RNA 来改写自己的 DNA，最近《自然》杂志的文章则提示这一现象在小鼠中同样存在。英国 Bath 大学的进化遗传学家 Laurence Hurst 说：“如果这类现象确实广泛存在，其影响将是巨大的”。

热点问题 1977 年在病毒中发现的选择性剪切，使一个 DNA 序列编码一个蛋白质分子这一生物学概念第一次受到挑战。大多数编码人类蛋白的 DNA 序列，有多个合成蛋白质的外显子，被不编码的内含子分隔开。选择性剪切时，细胞剪去内含子，把外显子按不同顺序结合起来，形成能编码不同蛋白质的信息。遗传学家随后也记载了重叠基因，基因内基因，和无数其他怪异的组合。

然而，选择性剪切本身并不需要对基因观念进行太大的重新评价。它仅仅显示有些 DNA 序列可以编码不止一个蛋白质。今天人们对 RNA 前所未有的认识对基因定义的冲击要深入得多。一个基因一个蛋白质的说法，正受到那些大量提取、分析 RNA 信息及人、鼠基因组转录的研究者的攻击。比如 Affymetrix 公司由 Thomas Gingeras 领导的学者

们，最新研究了 8 个人类细胞系的 10 条染色体的全部转录，精确分析出了每个转录是从染色体上的何处得来。这些研究得出的图谱复杂得令人难以置信。并不是分散的基因尽职地各自生产与之相同的 RNA 转录物，而是相当多的转录把基因组的许多片段转变成不同长度的多个 RNA 链。这些 RNA 链可以由 DNA 的两条链得来，而不是传统认为的只由 DNA 的一条链转录。有些转录是由含有蛋白质编码基因的 DNA 区域得来，但许多转录不是。Gingeras 的同事 Philip Kapranov 说：“这是带有革命性的发现。我们开始认识到基因组是充满重叠转录的”。

另一些研究提出了产生大量转录的原因。人们发现，转录可以由编码某个蛋白质的 DNA 序列开始，而一直进入到另一个编码完全不同的蛋白质的基因，产生一个融合转录。通过对人类 RNA 转录数据库的研究，Gingeras 小组估计，在传统意义上的基因区域，约有 4%~5% 的 DNA 以这种方式转录。产生融合转录可以是细胞从有限的外显子生成大量的不同蛋白质的一种方式。许多科学家开始认识到，在 DNA 中蛋白质的编码是没有界线的。每个序列可进入下一个以致再下一个。这个观点将成为今年将发表的 ENCOPE 项目的中心论点之一。Kapranov 组报告说，他们记录了许多转录的例子，编码蛋白质的外显子与基因组另一处的外显子结合。相结合的外显子可以在几十万碱基对以外，中间隔着几个其他“基因”。去年，耶鲁大学医学院的 Richard Flavell 记载了人的免疫系统的基因被另外染色体上的调节区域控制的现象。证明这种基因的延续甚至可以跨越染色体的界限。

需要新概念 对转录的大量调查发现，鼠和人基因组产生的 RNA 有很多并不编码蛋白质。比如去年日本的研究者联盟估计，有 63% 之多的鼠基因组被转录了，而只有 1%~2% 的基因中含有常见的外

显子序列. RNA 序列不仅仅是 DNA 和蛋白质合成机制的中间步骤, 这个发现本身并不新鲜. 细胞的蛋白质合成机器需要一些 RNA 分子和蛋白质来运作. 但发现“微 RNA”和其他 RNA 分子控制许多动植物的细胞过程, 还有新认识到的 RNA 转录的过度, 都支持 RNA 可以活跃地运作并执行基因组的指令的观点. 即使不叫基因这个名字, 产生非编码 RNA 的区域也应该被赋予基因的身份. “我认为人们该退后一步好好想了想,” Queensland 大学的分子生物学家 John Mattick 说, “系统中的许多信息其实是由 RNA 传递的”.

现在争论的焦点在于这些新发现的 RNA 在基因组中起的作用到底有多大. 可以想象, 一个产生过度转录, 然后忽略无用转录的体系, 比多个只产生特定的有用转录的体系要容易形成. 但去年的一项研究表明, 大量的过剩 RNA 中, 至少有一些是有用的. 在 Novartis 研究基金基因组研究所工作的 John Hogenesh 和他的同事们, 系统性地灭活了人细胞中 500 个以上非编码 RNA 的活性, 发现其中 8 个参与细胞信号传导和细胞生长. 但 Hogenesh 和许多其他科学家依然相信, 非编码的 RNA 在功能上远不如那些编码蛋白的 RNA 重要. 以前, 当科学家们寻找疾病的遗传基础或其他特性时, 他们发现, 绝大多数与疾病相关的突变是发生在编码蛋白质的基因内, 而不是别的区域内. “有绝对多的证据表明, 那些编码蛋白质的基因的地位是不会被动摇的”, Hogenesh 说.

最近这些有关非编码 RNA 分子的发现, 对分子生物学以外的人们没有构成太大的问题. 种群遗传学家可以检查一种特性如何传下去和进化, 而不管其背后精确的分子机制. 比如, 遗传学家可以建立模型来表现一种突变如何遗传的, 不管它是影响了蛋白质, 非编码 RNA 还是调节区. “我并不关心它合成蛋白质与否”, Hurst 说, “公式还是一样的”. 但对于基因组外遗传模式的研究来说, 情况就不同了. 最近一些年, 许多研究者致力于表观遗传. 其遗传信息从父母传到子女, 并不依赖 DNA 序列. 法国健康医学研究院的 Minoo Rassoulzadegan 研究组报导, RNA 有时会使传统的遗传模式复杂化. 例如在小鼠中有一个正常的和一个突变的 Kit 基因, 就可以引起尾巴和脚的白斑. 奇怪的是, 这些鼠的一些后代, 虽然遗传了两个正常的 Kit 基因, 依然出现白的尾巴. 法国研究小组认为突变的 Kit 基因产生了异常的 RNA 分子, 积聚在精子里, 传给卵子. 这少量的 RNA 不知通过

何种途径使下一代的正常 Kit 基因沉默了, 而产生带白斑的尾巴. 合作者 Francois Cuzin 说, “我们相信, 这是一种普遍的现象.”

如果认为这一现象很奇怪, 那去年普渡大学的 Robert Pruitt 和同事在十字花科植物拟南芥上的成果就更怪了. 被研究的基因是 HOTHEAD. 他们的分析表明, 一些植物没有携带他们父母具有的突变型的 HOTHEAD, 而用更早祖先的正常编码替换了异常的 DNA 序列. Pruitt 说, “这一发现完全改变了我对遗传的看法.” Pruitt 正致力于研究植物怎么可以表现这样的特性. 一种解释认为, 它们携带了祖父母的备份拷贝的遗传信息. 这些遗传信息编码在 RNA 里, 与正常的 DNA 一起传给种子, 而后用作模板来“纠正”某个基因. 可以想象, 一些神秘的非编码转录物可能起这种作用. Pruitt 说: “我认为, 有一些传统 DNA 基因组以外的物质被遗传下去了.” 这些发现如何影响我们对进化的理解, 还有待更进一步探讨. 但是对 RNA 跨世代携带遗传信息的研究, 肯定会更进一步丰富基因的定义.

变化中的观念 对于许多不直接涉及分子机制的科学家来说, 继续以简单概念认识遗传学会有问题吗? 一些遗传学家认为答案是肯定的. 他们担心, 研究者持有过于简单的基因观念, 可能会忽视与之不符的重要实验结果. 比如一个医学研究人员可能就会忽视在一个位点上某序列产生的许多不同的转录物. 而且对“基因是什么”缺乏一个清楚的共识, 也可能妨碍合作. 哈佛大学的发育遗传学家 William Gelbert 说: “我发现有时谈论基因时, 很难理解一些人的意思, 因为我们双方的定义不一样.” 对于生物信息学家来说, 没有一个清楚的基因定义也是很困难的. 因为他们正试图用计算机程序来找出 DNA 上一个基因开始和结束的标记. 但达成统一的定义基本也是不可能的. 例如 25 个科学家花了两天时间开会, 才找到一个大家都能同意的松散的基因定义: 一个与调控区域、转录或功能序列相关联的, 在基因组序列中可找到的, 对应于一个遗传单位的区域.

与其挣扎着达成一种定义, 大多数遗传学家倾向于用比较明确的形容词来修饰那些基本名词. 例如, 在基因前面加上“编码蛋白质的”或其他修饰.

正是基因定义的非唯一性引发了遗传学家持续的好奇心. 如果一个现象能被一个简单的四字母词很好地解释, 也就不那么神奇了.

(唐捷编译)