

Junk DNA

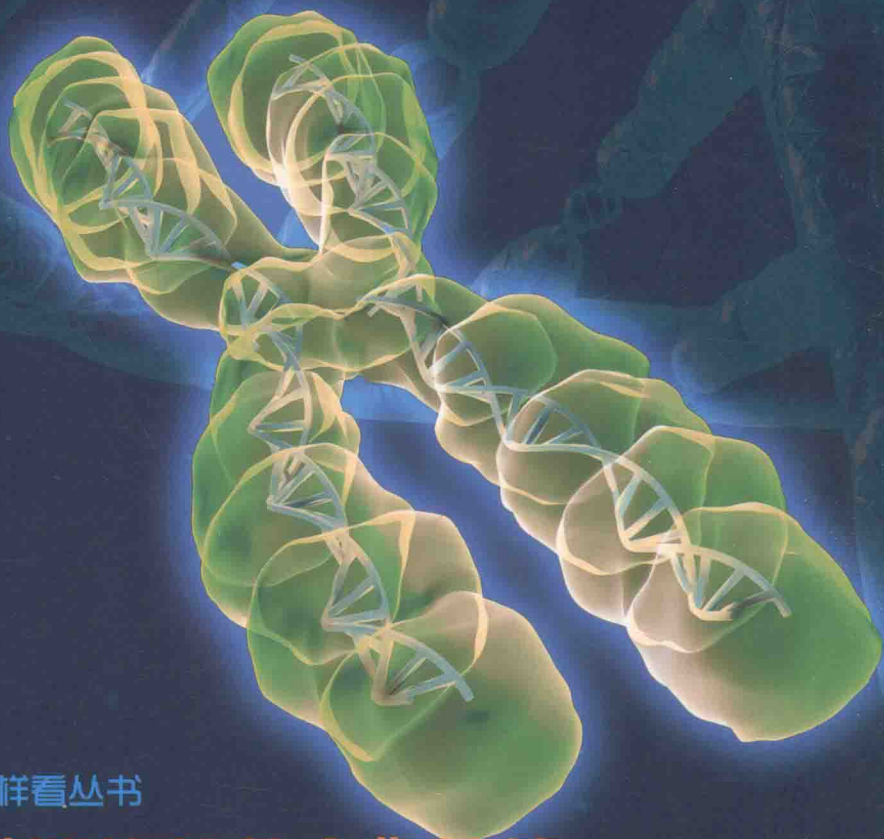
垃圾DNA

[英] 内莎·凯里 著
(Nessa Carey)
贾乙 王亚菲 译

重庆出版集团 重庆出版社
重庆文化传播公司

探索人类基因组暗物质之旅

HOW MODERN BIOLOGY IS REWRITING OUR UNDERSTANDING OF GENETICS,
DISEASE AND INHERITANCE



科学可以这样看丛书

畅销书《遗传的革命》续篇

人类基因组中98%是暗物质

探秘遗传学、门外汉都能读懂的世界科学名著

内容翔实、风趣幽默，读者们会非常喜欢内莎·凯里的《垃圾DNA》。

——莎朗·Y. R.登特(Sharon Y. R. Dent)
德克萨斯大学安德森癌症中心

科学可以这样看丛书

JUNK DNA
垃圾 DNA

探索人类基因组暗物质之旅

[英]内莎·凯里(Nessa Carey) 著
贾 乙 王亚菲 译

《遗传的革命》续篇
以非基因视角解读生命
开启基因组中暗物质的发现与探索

重庆出版集团  重庆出版社
 果壳文化传播公司

JUNK DNA: A JOURNEY THROUGH THE DARK MATTER OF THE GENOME

BY NESSA CAREY

Copyright © 2015 BY NESSA CAREY

This edition arranged with ANDREW LOWNIE LITERARY AGENT

through BIG APPLE AGENCY, INC., LABUAN, MALAYSIA.

Simplified Chinese edition copyright © 2016 Chongqing Publishing House Co., Ltd.

All rights reserved.

版贸核渝字(2015)第087号

图书在版编目(CIP)数据

垃圾 DNA / [英]内莎·凯里著;贾乙,王亚菲译. —重庆:重庆出版社, 2017.3

(科学可以这样看丛书 / 冯建华主编)

书名原文:JUNK DNA

ISBN 978-7-229-11749-8

I. ①垃… II. ①内… ②贾… ③王… III. ①脱氧核糖核酸—普及读物 IV. ①Q523-49

中国版本图书馆 CIP 数据核字(2016)第 268442 号

垃圾 DNA

JUNK DNA

[英]内莎·凯里(Nessa Carey) 著

贾 乙 王亚菲 译

责任编辑:连 果

责任校对:李春燕

封面设计:博引传媒·何华成



重庆出版集团 出版
重庆出版社

重庆市南岸区南滨路162号1幢 邮政编码:400061 <http://www.cqph.com>

重庆出版集团艺术设计有限公司制版

重庆市国丰印务有限责任公司印刷

重庆出版集团图书发行有限公司发行

E-MAIL:fxchu@cqph.com 邮购电话:023-61520646

全国新华书店经销

开本:710mm×1 000mm 1/16 印张:15.25 字数:250 千

2017年3月第1版 2017年3月第1次印刷

ISBN 978-7-229-11749-8

定价:39.80 元

如有印装质量问题,请向本集团图书发行有限公司调换;023-61520678

版权所有 侵权必究

Advance Praise for *JUNK DNA*
《垃圾 DNA》一书的发行评语

你基因组里只有 2% 的部分能够编码并产生蛋白质——生命依赖的基本分子。那么你基因组里其他的 98% 的部分是干什么的呢？

——内莎·凯里，《遗传的革命》一书作者

引人入胜、内容翔实、风趣幽默，读者们会非常喜欢内莎·凯里的《垃圾 DNA》。

——莎朗·Y. R. 登特 (Sharon Y. R. Dent)

德克萨斯大学安德森癌症中心

(University of Texas MD Anderson Cancer Center)

凯里明确了两点内容：我们（对自然）的理解总是暂时的和不断发展的，以及染色体的功能远超我们的想象。

——《出版商周刊》(*Publisher's Weekly*)

《垃圾 DNA》为变化迅速且越来越神秘的基因组提供了一个前沿且详尽的指南。

——琳达·格迪斯 (Linda Geddes)

《新科学家》(*New Scientist*)

幽默而全面的《垃圾 DNA》，为愿意付出时间和精力来阅读本书的读者完美地诠释了一个复杂的主题。

——《图书馆杂志》(*Library Journal*)

凯里为她阐述的主题付出了富有感染力的热情，对一个我们将来肯定会不断接触到的主题进行了完美的介绍。

——马克·迪斯托 (Mark Diston)

“登记”网站 (*The Register*)

《垃圾 DNA》是一本很好的与生命科学、遗传学相关的入门读物。

——《科学家》(*The Scientist*)

该领域的第一本科普概述……(凯里)对垃圾 DNA 的历史及其在基础和应用科学方面的影响都有着很好的把握。她具备将复杂的生物化学过程比喻成形象事物的天分,而这使非专业人士可以读懂这本书成为可能。

——埃尔夫·阿克塞尔·卡尔森 (Elof Axel Carlson)
《生物学评论季刊》(*The Quarterly Review of Biology*)

“与《遗传的革命》一书一样,作者以自己的研究经验为基础,用易于理解的方式带领读者对基因组中的‘暗物质’进行了一次探索之旅。尽管这些‘暗物质’在学术领域仍有巨大的争议存在,但这个快速发展的领域仍不失为我们探索基因组秘密的一个有用的新视角。”

——韦亚东,美国耶鲁大学医学院副研究员

“本书成功地对一个快速发展而且颇具学术争议的领域进行了全面的科普介绍,即使读者没有该领域的背景知识也能读懂本书。通过阅读这本书,读者会发现基因组的复杂程度远远超过了我们的想象。所以,只要对生命科学感兴趣,这本书就很适合你阅读。”

——孙明宽,美国约翰霍普金斯大学医学院博士后

*For Abi Reynolds, who is always by my side
And for Sbeldon—good to see you again*

献给

永远支持我的阿比·雷诺兹
同样献给谢尔登——很高兴再见到你

致 谢

我很幸运，我的第二本书依然受到伟大的经纪人，安德鲁·劳尼（Andrew Lownie）以及可爱的出版商的支持。对于 ICON 出版社，我特别要感谢邓肯·希斯（Duncan Heath）、安德鲁·弗洛（Andrew Furlow）和罗伯特·沙曼（Robert Sharman），但我不会忘记他们的前同事西蒙·弗林（Simon Flynn）和亨利·洛德（Henry Lord）。对于哥伦比亚大学出版社，我非常感谢帕特里克·菲茨杰拉德（Patrick Fitzgerald），布丽姬·弗兰纳里 - 麦科伊（Bridget Flannery-McCoy）和德里克·沃克（Derek Warker）。

一如既往，各种娱乐与教化都对我进行着支持。康纳·凯里（Conor Carey）、菲恩·凯里（Finn Carey）和加布里埃尔·凯里（Gabriel Carey）都参与其中，除了亲属之外，我还想感谢艾奥娜·汤姆斯 - 莱特（Iona Thomas-Wright）。还有给我无尽支持和大量饼干的令人愉快的婆婆，丽莎·多兰（Lisa Doran）。

自从我的第一本书出版后，我有了大量为非专业人士进行科学讲座的机会。已邀请我发言的各种机构太多，我无法在这里一一列出他们的名字，但他们都知道我在感谢他们，我非常享受你们给予我的特权。这非常令人鼓舞，谢谢你们所做的一切。

最后，我必须感谢阿比（Abi）。这位仁慈的人宽恕了我曾承诺过参加交谊舞课但事实上却没有参加这件事。

作者按

其实写一本关于垃圾 DNA 的书还是有些困难的，因为它的含义一直都在不断地演变。有部分原因是新的数据会不停地改变我们的旧有观念，结果就是，只要一个垃圾 DNA 被证明事实上具有一定的功能，有些科学家就会说（在逻辑上完全正确）它不再是垃圾。但使用这种方法有可能掩盖掉这些年来我们对基因组的理解产生了何等翻天覆地变化的过程。

与其花时间在这里纠缠不清，我决定采取一个快刀斩乱麻的方法，使用一个简单粗暴的定义。把任何不能编码蛋白的序列都定义为垃圾，如同它在过去最初的定义一般（20 世纪下半叶）。纯粹主义者会尖叫抗议，但就这样决定了。问 3 个不同的科学家什么是“垃圾”DNA，我们可能会得到 4 个不同的答案。因此，在开始的时候就简单一点利大于弊。

同样，我使用“基因”这个词来描述编码蛋白的那些 DNA 序列。这个定义会贯穿本书的始终。

在我的第一本书《遗传的革命》出版后，我意识到不同的读者对基因名称的需求差异很大。有些人喜欢知道我们正在讨论的是哪个基因，但对于其他读者而言，这或许会干扰阅读的流畅性。所以这次，我只在绝对必要的情况下才使用具体的基因名称。

对基因组暗物质的介绍

想象一下现在你手里有一个戏剧、一部电影、一部电视节目的剧本。当然，完全可能有人仅仅把剧本当成一本书来看。但是，如果它被用来产生某些东西的时候，它会变得更有力量。在被大声朗读出来，甚至被表演出来时，它就不再仅是页面上的一串字符而已了。

DNA 就是如此，它是最杰出的剧本。仅使用 4 个简单的字母，它就携带了生命的所有编码，从细菌到大象、从啤酒酵母到蓝鲸。但试管中的 DNA 是很无聊的，它完全没有用处。而细胞或动物开始用它进行生产的时候，DNA 就会变得令人兴奋。DNA 被作为制造蛋白的编码，这些蛋白对呼吸、进食、排泄废物、生殖和所有其他生命特有的活动都至关重要。

蛋白的功能是如此重要，以至于 20 世纪的科学家使用它们来界定基因的含义。基因就是能够编码蛋白的 DNA 序列。

让我们来怀念一下历史上最伟大的剧作家威廉姆·莎士比亚（William Shakespeare）。我们可能需要一段时间才能理解莎士比亚的著作，因为英语从他去世的那个世纪起已发生了不小的改变。但即使如此，我们仍然相信，诗人仅书写了他需要演员说的那些话。

例如，莎士比亚不会像下面这么写剧本：

Vjeqriugfrhbvruewhqoerahcxnqowhvgbutyunyhewqicxh
jafvurytnpemxoqp[etjhnuvrwwwebcxewmoipzowqmroseuiedn
rcvtycuxmqpzjmoimxdcnibyrvyrebanyhcuxqimokzqoxkmdci
fwrvjhentbubygdecftywerftxunihzxqwemiuqwjiqpodqeothe
rpowhdymrxnamehnfeicvbrgytrchguthhhhhhhgcwouldupaizm
jdpqsmellmjzufernngbyunasechuxhrtgcnionytuiongdjsio
niodefnionihiyhoniosdreniokikiniourvjcxoiqweopapqswee

JUNK DNA

twxmocviknoitrbiobeierrrrrrruorytnihgfiwoswakxdcjdrf
uhrqplwjkdhvmogmrfbvhncdjiwemxsklowe

事实上，他只写了画线部分的单词：

Vjeqriugfrhbvruewhqoerahcxnqowhvgbutyunyhewqicxh
jafvurytnpemxoqp[etjhnuvrwwwbcxewmoipzowqmroseuiedn
rcvtycuxmqpzjmoimxdcnibyrwvytebanyhcuxqimokzqoxkmdci
fwrvjhentbubygdecftywerftxunihzxqwemiuqwjiqpodqeother
powhdymrxnamehnfeicvbrgytrchguthhhhhhhgcwoulddupaizm
jdpqsmelllmjzufernngbyunasechuxhrtgcnionytuiongdjsio
niodefnionihyhoniosdreniokikiniourvjcxoiqweopapqsweet
wxmocviknoitrbiobeierrrrrrruorytnihgfiwoswakxdcjdrfu
hrqplwjkdhvmogmrfbvhncdjiwemxsklowe

就是“A rose by any other name would smell as sweet（即使给玫瑰换个称谓，它依然芳香如故）”。

但如果看看我们自己的 DNA 剧本，它并不像莎士比亚那些画了线的词语那么清晰和紧凑。相反，每个编码蛋白的区域就像是漂浮在“废话海洋”上的一个词。

多年来，科学家们并不能解释为什么那么多的 DNA 不编码蛋白。这些非编码的部分被误会成“垃圾 DNA”。但渐渐地，这个看法已经被一大堆理由逼迫得站不住脚了。

也许，转变该观点最根本的原因是源于我们细胞中垃圾 DNA 那惊人的总量。其中一个最大的冲击来自 2001 年人类基因组测序完成的时候，人们发现人类细胞中 98% 以上的 DNA 被划入了垃圾的行列。它不编码任何蛋白。上面使用的莎士比亚的比喻实际上是一个精简了的模型。在基因组里，乱七八糟的文字的比例约为前面提到的乱序剧本的 4 倍。每一个有意义的字母都对对应着超过 50 个字母的垃圾文字。

还有一个类似的比喻。请想象下我们去参观一个汽车厂，这家汽车厂也许像法拉利那么高端。如果我们看到只需要两个人就能造好一辆炫酷的红色跑车，而同时又有 98 个人在一旁无所事事的话一定很惊讶。这显然是

荒谬的，那么为什么这在我们的基因组里是合理存在的？当然，你可以说从共同祖先进化而来的生物往往不是完美的——比如我们人类就确实不需要阑尾（本书中作者多次以阑尾为例说明人体确实存在一些无功能的器官，但此观点尚存争议，因为有研究者认为阑尾在免疫和消化方面仍具有一定的功能。）——但这次似乎不完美得有点过头了。

事实上，在我们的汽车厂里，更可能的情况应该是，由两个人组装一辆汽车，其他的 98 个人做着其他一切使这家企业运转的工作。融资、记账、宣传产品、处理养老金、打扫厕所和汽车销售等，这可能是一个更好的垃圾 DNA 在我们基因组里工作的模式。我们可以把蛋白作为生命所需的最终物质，但如果没有垃圾 DNA 的话，它们永远不会被正确地生产和整合出来。确实，两人就能制造汽车，但他们不能保证公司可以出售汽车，也不能把它变成一个强大且成功的品牌。同样，如果没有汽车被生产出来，就算有 98 个员工在展厅拖地板和磨破嘴皮子也没有任何意义。只有每个组件各司其职的时候整个组织才能运转。而这，就是我们基因组的模式。

另一个来自基因组测序的令人震惊的事实是，使用经典的基因模型无法解释人类极其复杂的解剖结构、生理功能、智力和行为。在编码蛋白的基因的数量上，人体跟简单而微小的蠕虫几乎具有相同的数量（大约 20000 个）。更值得注意的是，大部分蠕虫的基因跟人类基因可以直接等效。

当研究人员想在 DNA 水平上深入分析人类与其他生物到底有什么区别的时候，很明显，基因不能提供解释。事实上，遗传物质只有一个特征与复杂性相关。这个唯一的随着动物复杂性增加而增加的就是垃圾 DNA 区域。一个生命体越复杂，垃圾 DNA 所占的百分比就越高。只有现在，科学家们才真正开始关注这个有争议的观点，就是垃圾 DNA 可能是进化复杂性的关键。

从某个角度看，这些数据所引出的问题是很明显的。如果垃圾 DNA 是如此的重要，那它到底是干什么的？如果不编码蛋白，它在细胞中的作用是什么？现在人们逐渐接受了垃圾 DNA 其实具有多种不同功能的观点，所以对下面的叙述我们并不会感到吃惊。

它们中的一些在染色体（我们的 DNA 被打包成的巨大分子）中形成特殊的结构。这种垃圾 DNA 保护我们的 DNA 不被解体和受损。随着我们年龄的增长，这些区域不断缩小，最后缩小到临界程度。之后，我们的遗

JUNK DNA

传物质就变得很容易遭遇潜在的灾难性的重排，从而导致细胞的死亡或癌变。其他结构类型的垃圾 DNA 在细胞分裂成子细胞过程中进行染色体均分时作为锚点。（“子细胞”是指由亲代细胞分裂产生的细胞。）其他的则作为绝缘结构，限制了特定区域染色体的基因表达。

但是我们很多的垃圾 DNA 并不仅仅具有上面提到的功能。它确实不编码蛋白质，但它却编码另外一种分子，我们称之为 RNA。垃圾 DNA 里面有很大一类在细胞内建造工厂，来帮助蛋白质的合成。其他类型的 RNA 分子负责将制造蛋白质所需的原料转移到工厂里面。

另外有些垃圾 DNA 是来自病毒和其他微生物的遗传物质的入侵，它们已经如遗传间谍般融入了人类染色体中。这些早已死去的生物的残留物对细胞有着潜在的危险，有些在个体中，而有的甚至存在于广泛的人群中。哺乳动物细胞已进化出多种机制来使这些病毒元件保持沉默，但有的时候这些系统可以被打破。当出现这种问题的时候，它们所产生的影响可以是相对良性的，比如特定种系小鼠毛色的改变，也可以是很严重的，比如增加罹患癌症的风险。

就在前几年，人们认识到垃圾 DNA 的一个主要作用其实是调节基因的表达。有时候，它在个体上会有巨大的、明显的效果。比如，一个垃圾 DNA 就可以决定雌性动物能否保持正确的基因表达模式。它的作用也可以是在群体中的，一个最常见的例子是虎斑猫颜色特征的控制。在极端的例子中，该机制也能解释为什么有同样遗传性疾病的同卵双胞胎女性会出现截然不同的症状。在某些情况下，情况可以极端到双胞胎中的一个罹患了严重危及生命的疾病，而另一个则是完全健康的。

成千上万的垃圾 DNA 片段被认为参与了调控基因表达的网络。它们就像是遗传剧本的舞台导演一样，只是其指导对象的复杂性是我们在剧院里无法想象的。它绝对不是“出去，被熊追赶”那么简单。而应该是类似于“如果在温哥华和珀斯的暴风雨中表演《哈姆雷特》，就应该重读《麦克白》这一行的第四音节。除非有个业余演员在蒙巴萨表演《理查三世》而且基多在下雨”。

研究人员刚刚才开始揭开垃圾 DNA 庞大网络中的奥秘和关联的一角。这个领域仍极具争议。在极端的情况下，有科学家声称有些武断的说法严重缺乏实验证据的支持。其他有些人则觉得有整整一代科学家（甚至更多）被困在一个过时的模型中而无法看到或者理解这个新领域。

有部分原因是，我们可以用来探索垃圾 DNA 功能的手段还比较落后。这有时会让研究人员很难使用实验来检验他们的假设。确实，我们对该领域的研究时间还相对较短。但有时候，我们可以从实验室的板凳跟机器旁退出来，去草坪转转。实验每天都在我们身边发生，因为自然和进化已经使用了几十亿年的时间来尝试各种变化。即使仅仅在我们这个物种出现和繁衍的时间段，也已经有足够的时间来进行大规模的实验测试。因此，我们将在本书中利用人类遗传学的火炬来探索黑暗。

可以有许多方法来开始我们的探究基因组暗物质之旅，这里，让我们用一个有些奇怪但不容置疑的事实来开始。一些遗传病是由垃圾 DNA 的突变引起的，这应该是我们进入隐藏的基因组宇宙的最好起点。

目录

1	□ 致谢
1	□ 作者按
1	□ 对基因组暗物质的介绍
1	□ 1 为何暗物质是有意义的
7	□ 2 当暗物质事实上变得更黑暗
17	□ 3 所有的基因都去哪儿了?
25	□ 4 不速之客
33	□ 5 随着我们变老一切都在减少
45	□ 6 二才是完美的数字
57	□ 7 用垃圾来涂抹
71	□ 8 玩个长时间的游戏
83	□ 9 给暗物质加点颜色
91	□ 10 为什么父母爱垃圾
105	□ 11 有任务的垃圾
115	□ 12 启动,上调
127	□ 13 无人之地
B5	□ 14 ENCODE 项目 ——走向垃圾 DNA 的大科学
147	□ 15 无头的皇后、奇怪的猫和肥胖的小鼠
161	□ 16 非翻译区中的迷失
175	□ 17 为什么积木比模型好
189	□ 18 迷你也能很强大
201	□ 19 有用的药物(有时候)
209	□ 20 黑暗中有一丝光亮
215	□ 附录:正文中出现过的与垃圾 DNA 有关的人类疾病

1 为何暗物质是有意义的

有时候，同一个家庭里面的成员会陷入不同的悲惨遭遇。如下面这个例子：一个男孩出生了，让我们称他为丹尼尔（Daniel）。他出生的时候就全身瘫软，而且不能自主呼吸。经过大量的治疗，丹尼尔活了下来，他的肌张力有所提高，从而能够进行独立呼吸和移动。但是，随着年龄的增长，丹尼尔出现了会伴随一生的学习能力障碍。

他的妈妈萨拉（Sarah）很爱丹尼尔，而且每天都照顾着他。但当她到了30岁的时候，因为自己也出现了奇怪的症状而使照顾丹尼尔变得更加困难。她的肌肉开始变得僵硬，进而，出现了抓住东西后难以放下的情况，因此，她不得不放弃了对技术要求很高的兼职工作——陶器修补工。她的肌肉开始出现了明显的消瘦，而她还是尽力克服了这些困难，直到42岁，萨拉突然死于心律失常，这是一种因平时维持心脏跳动的电信号出现异常而导致的致死性疾病。

于是，照顾丹尼尔的重任落在了萨拉的母亲，珍妮特（Janet）的身上。这对于她是项挑战，不仅仅是因为要照顾外孙和承受失去女儿的巨大悲痛。因为珍妮特在50岁的时候就得了白内障，因此她的视力非常糟糕。

看起来这个家庭好像是因为不幸而遭受了一组不相干疾病的困扰。但专家们开始注意到一些不寻常的事情。这种模式——一个人患白内障，其女儿出现肌肉僵硬和心脏缺陷，而孙子辈出现肌肉瘫软和学习障碍——在多个家庭都发生了。而这些人的家庭是分散生活在世界各地的，并且之间没有血缘关系。

科学家们意识到他们关注的是一种遗传性疾病。他们把它命名为强直性肌营养不良症（Myotonic dystrophy）。这种情况发生在罹患该病家庭的每一代。如果父母患病的话，孩子一般有50%的概率患病。没有性别差异，不论男女都能将该病遗传给下一代。

这些遗传特征是非常典型的单基因突变引起的疾病所具有的。突变是指正常 DNA 序列的改变。我们细胞中的基因一般具有两个拷贝，一个来自母亲，一个来自父亲。强直性肌营养不良症在每一代中都出现症状的这种遗传特征，一般被称为显性。在显性遗传的疾病中，两个基因拷贝中往往只有一个上面携带有突变。它就是从父母中患病的那一方遗传到的拷贝。这种突变基因在即使细胞中另一条拷贝是正常的情况下也能导致疾病的发生。该突变基因在某种程度上相对正常基因更“显性”。

但是同时，强直性肌营养不良症具有一些跟典型的显性遗传疾病完全不同的特点。首先，显性遗传疾病不会在一代一代的遗传中变得越来越严重。它没有理由会这样，因为患儿遗传到的突变跟他们的患病父母是一样的。强直性肌营养不良症患者还会随着遗传代数的增加而发病得越来越早，这也是不寻常的一点。

强直性肌营养不良症还有一个跟典型遗传病不同的地方。丹尼尔身上出现的如此严重的先天症状，仅在患病母亲生出的孩子中出现。患病父亲从不遗传这么严重的症状。

在 20 世纪 90 年代初，很多不同研究团队鉴定出了导致强直性肌营养不良症的基因改变。跟这个不寻常的疾病很匹配，这是一个不同寻常的突变。强直性肌营养不良症的基因里包含了一小段重复很多次的 DNA 序列。这个小序列由构成 DNA 序列的 4 个字母中的 3 个构成。在强直性肌营养不良症基因中，该重复序列由字母 C（胞嘧啶）、T（胸腺嘧啶）和 G（鸟嘌呤）组成，这里没有出现的基因字母是 A（腺嘌呤）。

在没有强直性肌营养不良症突变的人中，大概有 5 到 30 个前后相连的这种 CTG 序列的拷贝。子女从父母那里遗传到相同的重复数目。但是，一旦重复的数量变大，超过 35 个左右的拷贝时，该序列就会变得不稳定，可能导致在从父母传递给子女的过程中出现改变。一旦超过 50 个拷贝，该序列就变得非常不稳定。当这一切发生时，父母就会传递给子女比他们实际拥有的更多的拷贝数。随着重复数量的增加，症状会在生命早期就变得更严重和更显著。这就是为什么在本章开始提到的那个家庭中，他们罹患的疾病会一代比一代更严重。该疾病也只有在母亲将能够导致严重疾病表型的巨大重复数量的拷贝传递给下一代时，才会表现出来。

这种 DNA 重复序列的不断扩展是一种非常少见的突变机制。但对导致强直性肌营养不良症这种 DNA 重复序列不断扩展的鉴定则引出了一些更不