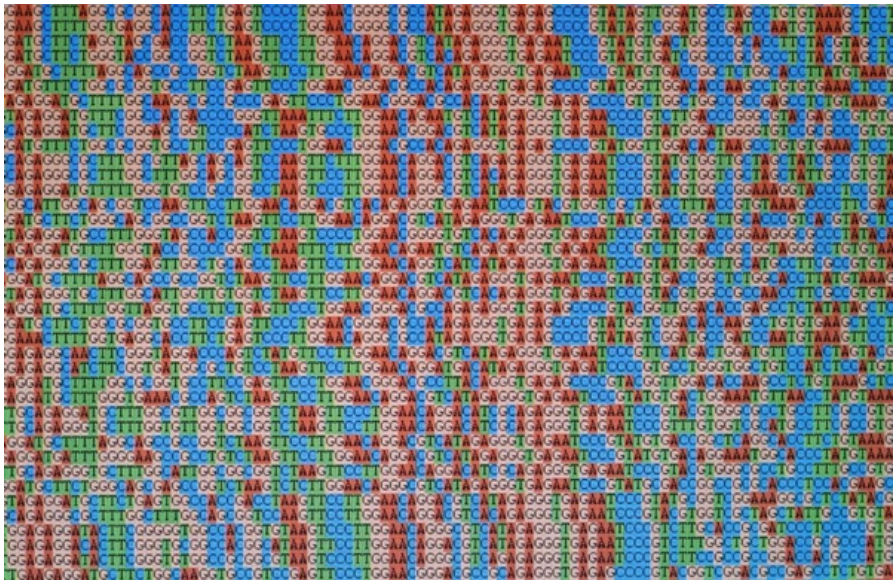
作者：宗华 来源：中国科学报 发布时间：2018/6/26 9:41:09

你到底有多少基因

科学家公布人类基因数量引发争议



在人类基因组项目完成十多年后，辨别基因仍是一项挑战。

图片来源：Alan Phillips/Getty

估测人类基因组中基因数量的最早尝试涉及喝醉酒的基因学家、美国纽约冷泉港的一个酒吧以及纯粹的臆测。

那是2000年。当时，人类基因组序列草图仍在绘制中。基因学家正在打赌人类拥有多少基因，赌注从几万个到几十万个不等。近20年后，掌握了真实数据的科学家仍无法就这一数量达成一致。在他们看来，这一知识鸿沟阻碍了发现相关疾病突变的努力。

填补这一空白的最新努力利用了来自上百个人类组织样本的数据，并于日前发表在预印本服务器BioRxiv上。它包括近5000个此前未被发现的基因，其中近1200个携带制造蛋白质的指令。2.1万余个蛋白质编码基因的总数和此前估测（认为这一数字在2万左右）相比有大幅提高。

不过，很多遗传学家仍不相信所有最新提出的基因都能经得起仔细推敲。他们的批评强调了辨别新基因甚至定义一个基因的难度。

“20年来，人们一直致力于此项研究，但我们仍未获得答案。”带领团队开展最新研究的约翰斯·霍普金斯大学计算生物学家Steven Salzberg表示。

2000年，随着基因组学界就有多少人类基因将被发现的问题展开热烈讨论，Ewan Birney发起了GeneSweep竞赛。如今身为欧洲生物信息学研究所（EBI）联合所长的Birney在一年一度的基因组学会议期间，在一间酒吧里最先下注。

这场竞赛最终吸引了1000多人参与以及3000美元的累积赌注。关于基因数量的赌注从多于31.2万个到不足2.6万个不等，平均在4万左右。当时，估测的数量范围已经缩小，但仍存在不同意见。

基因数量依据被分析的数据、利用的工具以及剔除错误信息的标准而有所不同。最新计数利用了一个更大的数据集、另一种不同于此前努力的计算方法，以及定义基因的更宽泛标准。

Salzberg团队利用了基因型组织表达（GTEx）项目的数据。该项目对从几百具尸体上采集的30多个不同组织的RNA进行了测序。RNA是DNA和蛋白质之间的“媒介”。研究人员想辨别出编码蛋白质的基因以及不编码蛋白质但仍在细胞中扮演重要角色的基因。为此，他们组装了GTEx的9000亿个微小RNA片段并将其同人类基因组进行比对。

不过，仅一段DNA被表达为RNA并不意味着它是一个基因。为此，该团队尝试利用各种标准过滤掉噪音。例如，他们将获得的结果同来自其他物种的基因组进行比较，并且推断远亲生物共享的序列可能在进化过程中被保存下来，因为它们是有用的，基因也可能如此。

研究人员获得了21306个蛋白质编码基因和21856个非编码基因——远多于两个最广泛使用的人类基因数据库中的基因数量。由EBI维护的GENCODE基因集包括19901个蛋白质编码基因和15779个非编码基

相关新闻 相关论文

1 可直接阻止病毒复制，一基因抗病毒的秘密被破解
2 花一亿元驳斥一篇转基因论文，值吗？
3 基因研究助力植物区系分区
4 抗病基因延长树木寿命
5 南京医科大找到影响女性排卵受孕的“关键基因”
6 杨树关键基因或促生物燃料大发展
7 科学家给2万年前人熊猫做线粒体基因组测序
8 中国科学家找到影响女性排卵受孕的“关键基因”

图片新闻

>>更多

一周新闻排行 一周新闻评论排行

1 关英科学家获2018年度诺贝尔化学奖
2 掌控进化：生命这样被改写
3 陈列平与诺奖失之交臂 专家：原因有三
4 今年诺奖自然科学奖“写满”两个字：续命
5 18年里18人获奖，好学术环境比诺奖更重要
6 华人女科学家曹颖获美国“天才奖”
7 科技发展40年：多项指标世界领先
8 考研人数攀升，为何推免比例还更高？
9 院士为栽培技术鸣不平：研发投入勿“跑偏”
10 “上帝粒子”之父利昂·莱德曼逝世

更多>>

编辑部推荐博文

- 热力学中一道“伟人的习题”
- 安抚牛人的最好法子是封官？
- 对文章假设、观察和解释部分进行区分的重要性
- 如释重负，这问题压了我十多年！
- 回到山南
- 该给人还是狗接种狂犬病疫苗？菲律宾的经验教训

更多>>

论坛推荐

因。由美国国家生物技术信息中心（NCBI）管理的RefSeq数据库拥有20203个蛋白质编码基因和17871个非编码基因。

NCBI基因组研究人员、RefSeq 之前的负责人Kim Pruitt表示，出现这一差异的部分原因可能是Salzberg团队分析的数据量不同。不过，还有另外一个重要差异。GENCODE和RefSeq均依赖于人工管理——有人评审每个基因的证据并且作出最终判断。Salzberg团队则完全依赖于计算机程序筛选数据。

“如果人们喜欢我们的基因目录，那么或许几年后我们将成为人类基因的仲裁者。”Salzberg说。

不过，很多科学家表示，他们需要更多证据以确信最新目录是准确的。协调GENCODE人工注释工作的EBI计算生物学家Adam Frankish介绍说，他和团队已经扫描了Salzberg团队辨别的约100个蛋白质编码基因。根据他们的估测，仅有1个看上去是真正的蛋白质编码基因。

与此同时，Pruitt小组分析了Salzberg团队公布的约十几个新的蛋白质编码基因，但并未发现任何符合RefSeq标准的基因。一些同看上去属于侵入人类祖先基因组的逆转录病毒的基因组区域重叠，剩下的则属于极少被翻译成蛋白质的其他重复性片段。

不过，Salzberg认为，一些重复序列可被视为基因。一个例子是出现在RefSeq 中并且编码在结直肠癌中过度表达的蛋白质的ERV3-1。Salzberg还承认，位于其团队目录中的新基因有待该团队和其他人确认。（宗华编译）

《中国科学报》（2018-06-26 第3版 国际）

更多阅读

《自然》相关文章（英文）

打印 发E-mail给：

以下评论只代表网友个人观点，不代表科学网观点。

2018/6/26 11:38:12 xlsd

争议很正常

2018/6/26 10:43:51 qw10805

翻译的很别扭啊

目前已有2条评论

查看所有评论

需要登录后才能发表评论，请点击 [登录]