将 4 万个黑猩猩的 DNA 序列与人类基因组进行比对，全基因组 DNA 比对相似性同一性为 86%-89%

为了提供一套全新且偏差较小的全球分析数据，我们使用 BLASTN 算法对黑猩猩和人类基因组进行了大规模的 DNA 序列比对。一组实验启用了查询序列和目标序列的低复杂度序列掩蔽，而另一组实验则禁用了掩蔽参数。每组子实验测试了三种不同字长（7、11 和 15）和五种不同 e 值（1000、10、0.1、0.001 和 0.00001）的十五种组合，总共尝试了 120 万次全基因组比对。每个 BLASTN 查询任务都使用了一个包含 40,000 条黑猩猩全基因组鸟枪法测序序列（WGSS）的数据集，这些序列来自美国国家生物技术信息中心（NCBI），并与四种不同的人类基因组组装版本（GRCH37、GRCH36、Alternate SNP Assembly 和 Celera Assembly）进行比对。

使用低复杂度序列掩蔽技术可使计算时间缩短约 5-6 倍，略微延长比对长度，减少数据库匹配数量，并略微降低核苷酸同一性百分比。根据 BLASTN 参数组合的不同，30 个独立实验中人类和黑猩猩的平均序列同一性在 86%至 89%之间。黑猩猩查询序列的平均长度为 740 个碱基，根据 BLASTN 参数组合的不同，平均比对长度在 121 至 191 个碱基之间。

随后，通过与 NCBI 工作人员的私人通信以及本研究的支持数据，黑猩猩序列被认为与人类基因组存在一定程度的同源性。然而，排除大量未比对的黑猩猩序列数据后，人类-黑猩猩 DNA 全基因组相似性的保守估计为 86-89%。本研究结果明确表明，人类和黑猩猩基因组的同一性至少比通常认为的低 10-12%。这些结果与人类和黑猩猩之间观察到的巨大解剖学和行为学差异更加吻合。
关键词：人类与黑猩猩相似性，基因组比较，DNA 比对，BLASTN 算法

介绍

一种常见的说法通过晦涩难懂的研究出版物和通俗进化科学作者传播，即黑猩猩（*Pan troglodytes*）和人类（*Homo sapiens*）的 DNA 相似度高达 98-99%。几乎所有过去的人类-黑猩猩 DNA 比较研究都存在一个主要问题：数据在进行比对、汇总和讨论之前，通常要经过多道预筛选、过滤和选择工序。无法比对的区域通常会被省略，比对结果中的空位也常常被丢弃或混淆处理。

在即将发表的一篇论文中，Tomkins 和 Bergman (2012) 逐一讨论了大多数关于人类-黑猩猩 DNA 相似性的关键研究论文，并指出，即使纳入了被舍弃的数据（如有提供），人类和黑猩猩的 DNA 相似性实际上

也不超过 80-87%，甚至可能更低。以下简要分析了三篇关于人类-黑猩猩基因组进化比较的关键论文，这些论文提供的数据与本研究的结果一致。如需更全面的文献综述，请参阅 Tomkins 和 Bergman (2012)。

最早将黑猩猩基因组大片段与人类基因组进行比较的研究之一，是布里顿实验室于 2002 年利用其自主开发的 Fortran 计算机程序完成的。该研究基于五个已知与人类同源的黑猩猩大 DNA 片段（BAC 克隆），并对其进行了全面的测序。所有五个 BAC 克隆的 DNA 序列总长度为 846,016 个碱基，但只有 92%的 DNA 与人类序列比对成功，而该论文仅报告了其中的 779,132 个碱基。包含插入和缺失（indel）的比对结果表明，人类与黑猩猩的相似度为 95%（Britten 2002）。然而，当纳入所有五个 BAC 克隆的完整序列后，黑猩猩与人类同源区域的最终 DNA 相似度仅为 87%。

2004 年，Watanabe 等人利用多种 BAC 文库筛选出代表黑猩猩 22 号染色体的克隆进行 DNA 测序。随后，他们将测序结果与人类同源区域进行比较。需要注意的是，黑猩猩 BAC 克隆的选择标准是每个克隆必须包含 6-10 个人类 DNA 标记。这种初始程度的预筛选偏差是一种常用的方法。与许多进化生物学文献一样，本文及其补充信息中均未提供整体 DNA 比对统计数据。对于比对片段，作者给出了 1.44%的核苷酸替换率，但并未

提供包含插入/缺失（indel）在内的相似性估计值。作者指出存在 82,000 个 indel，并提供了一个显示 indel 大小分布的直方图。然而，平均 indel 大小或总 indel 长度的数据明显缺失。此外，作者给出了序列缺口的数量，但未提供关于总缺口长度的具体数据。尽管比较的是测序完善的直系同源区域，但用于计算人类和黑猩猩之间精确 DNA 相似性的数据却被忽略了。基于碱基替换和插入缺失的图形数据估算，可以推断出总体相似性约为 80-85%。

关于人类-黑猩猩基因组比较的主要里程碑式出版物是 2005 年发表在*《自然》*杂志上的黑猩猩测序与分析联盟（The Chimpanzee Sequencing and Analysis Consortium）的论文。遗憾的是，该论文以高度选择性且令人困惑的方式呈现了比较数据，并且缺乏关于比对的详细表格数据。论文的大部分内容主要关注于针对不同分歧率和选择压力的各种假设性进化分析。然而，根据该论文中给出的数据，我们可以计算出人类和黑猩猩之间大致的基因组相似性。作者指出：

> *黑猩猩和人类基因组的最佳互惠核苷酸水平比对涵盖了约 2.4 吉碱基（Gb）的高质量序列（黑猩猩测序和分析联盟 2005 年，第 71 页）。*

当时，人类基因组组装估计已接近完成，大小为 2.85 Gb，错误率为十万分之一（国际人类黑猩猩基因组测序联盟，2004）。2005 年，黑猩猩基因组的作者也指出：

> *因此，基因组之间的插入缺失差异总计约为 90 Mb。这一差异相当于两个基因组的约 3%，远大于核苷酸替换导致的 1.23% 的差异（黑猩猩测序和分析联盟 2005 年，第 71 页）。*

通过将插入缺失和替换数据（4.23%）应用于 2.4 Gb 的比对人类-黑猩猩序列，并考虑未*比*对的人类序列量，可以计算出最大相似度为 80.6%。这是一个非常保守的估计，因为核苷酸 BLAST 的默认比对会掩盖大量低复杂度序列。此外，最新的黑猩猩基因组框架（“黄金路径”连续 ENSEMBL 组装；http://uswest.ensembl.org/Pan_troglodytes/Info/Index）表明，黑猩猩基因组比人类基因组大约大 8%。纳入此数据将进一步降低全基因组相似度，使其低于 74%。关于黑猩猩和人类基因组的测序方法以及理解这些技术对于解释 DNA 相似性问题的重要性，请参阅 Tomkins (2011a) 的最新综述。

自 2005 年黑猩猩测序和分析联盟报告以来，世俗文献中一直缺乏人类和黑猩猩之间全面的全基因组比较。

创造论评论与分析

总体而言，神创论者对人类-黑猩猩基因组相似性的研究主要局限于对进化研究中提出的论断进行解读，而没有充分探讨进化研究中使用的高度选择性方法或经常被忽略的无法比对的数据。尽管如此，许多重要的观点和发现仍然被揭示出来。

在黑猩猩基因组计划完成之前，分子生物学家大卫·德威特指出，尽管人类和黑猩猩的 DNA 被认为具有高度相似性，但在细胞遗传学、转座元件的类型和数量、插入和缺失事件、基因表达模式以及 mRNA 剪接等方面仍存在显著差异（DeWitt 2003）。在后来的报告中，德威特还证明，即使接受 5%的全基因组差异，这种相似性水平仍然不足以支持与进化时间线相符的各种选择和共同祖先假说模型（DeWitt 2005）。桑福德等人（2008）通过计算机模拟进一步检验了人类基因组中突变积累的速率，发现无论已报道的人类-黑猩猩基因组差异如何，该速率都对达尔文进化时间线构成了严峻挑战。

许多将人类和黑猩猩从共同祖先中分离出来的突变（DNA 序列差异)被认为发生在基因组的非编码区域，这一发现最近已得到一份进化报告的证实（Polavarapu 等人，2011）。尽管关于人类和黑猩猩之间非编码 DNA 差异的进化报告不断涌现，但这些差

异与人类基因组整个非编码区域的功能性和丰富特征之间的逻辑关联却被大大低估了。对 DNA 元件百科全书（ENCODE）的广泛研究已有力地证实了非编码 DNA 的许多关键特征（ENCODE 项目联盟，2011）。在神创论研究领域，生物学家伍德莫拉普和巴顿是最早阐明非编码 DNA 领域多样化数据如何支持各种重要非编码序列类别和 DNA 特征在基因组范围内发挥功能的神创论学者之一（巴顿，2005；伍德莫拉普，2004）。在最近一篇全面综述中，分子生物学家兼智能设计论支持者乔纳森·威尔斯（Jonathan Wells）对非编码 DNA 中的各种设计特征进行了深入探讨，并彻底驳斥了“垃圾 DNA”这一谬论（Wells 2011）。关于该主题的简要综述以及对威尔斯著作的总结，请参阅汤姆金斯（Tomkins）最近发表的文章（2011b）。

或许人类与黑猩猩之间最大的持续性矛盾，也是与所谓高度相似性论断不符之处，在于二者在行为和解剖结构上的显著差异，正如创造论者安德森（2007）、珀多姆（2006）和维兰德（2002）所总结的那样。这些显而易见的人类与黑猩猩之间的差异，似乎与二者之间近乎完全相同的 DNA 相似性的说法并不相符。事实上，一位为 BBC 撰稿的世俗科学作家最近出版了一本名为《*并非黑猩猩*》（泰勒，2009）的书，专门论述了这一悖论。

尽管许多神创论作者暂时接受了关于人类和黑猩猩 DNA 相似性的标准进化论观点，但一些报告指出，“几乎相同”的教条并非表面看起来那么简单明了。事实上，有报告指出，关于人类和黑猩猩 DNA 相似性的进化数据报告大多是预先筛选过的数据，这些数据已知在某种程度上是同源的（序列相似），例如不同物种间共享的高度相似的蛋白质编码序列（Tomkins 2009a, 2009b）。此外，最近的一项文献综述结合生物信息学研究项目，评估了两个类似黑猩猩的染色体（2a 和 2b）融合形成人类 2 号染色体的假设。该项目表明，进化灵长类动物融合范式在许多关键方面存在严重缺陷，进一步否定了几乎相同的 DNA 主张（Bergman 和 Tomkins 2011；Tomkins 2011c；Tomkins 和 Bergman 2011）。

在神创论研究群体中，鲜有大规模生物信息学研究比较人类和黑猩猩的基因组。首例此类分析报告由神创生物学家托德·伍德（Todd Wood）在其发表的关于人类和黑猩猩生物学相似性的综述中简要提及（Wood 2006）。尽管该报告主要为文献综述，但也简要描述了伍德本人为验证 2005 年黑猩猩基因组组装结果而进行的分析。伍德利用已知相似且可比对的共有基因推导出的蛋白质序列，对黑猩猩和人类进行了大规模氨基酸序列比对。然而，对已知直系同源基因（跨物

种相似基因）的 DNA 编码序列进行电子翻译后的蛋白质序列比对，并不能准确反映全基因组的 DNA 相似性，因为人类基因组中只有不到 4%的序列编码蛋白质（国际人类基因组测序联盟 2004）。更重要的是，使用电子生成的蛋白质进行比较的主要问题是，大多数人类基因会经历选择性转录和翻译、多种外显子剪接方法、基因内调控 RNA 编码片段、增强子元件以及许多其他复杂的转录剪接编码特征（Barash 等人 2010；ENCODE 项目联盟 2011；Wells 2011）。

最近，伍德在 2011 年创造生物学会年会上发表了一篇关于人类-黑猩猩全基因组比较的论文，并发布了该研究的简要摘要（Wood 2011）。伍德指出，他使用 BLASTN 算法将 4 万条随机的黑猩猩基因组序列与最新版本的人类基因组进行两两比对。然而，关于所用算法参数以及数据返回和评估方式的细节却缺失。伍德显然使用了标准的默认参数，这些参数本应包含序列掩蔽和单个查询序列的多重比对数据汇总——即黑猩猩序列在人类基因组的多个位置匹配。因此，报告的是多个查询比对结果的总和，而不是单个黑猩猩序列查询结果之间的比较。从摘要中呈现的结果来看，伍德可能选择了 Megablast——BLASTN 的一个变体——它使用包含不连续词长模板功能和评分矩阵的默认参数。Megablast 会筛选出高度相似的 DNA 序列，而忽略了

构成人类和黑猩猩基因组主体的大部分复杂度较低、相似度较低的基因组特征。因此，伍德最终的统计数据偏向于极高的序列一致性，忽略了大部分被比较的序列。

无论 Wood 使用的是 Megablast 还是标准的 BLASTN 默认值，这些方法通常仅用于检测高度相似的区域，而无法提供客观的全基因组比对数据。我们尝试使用 BLASTN 的标准默认参数（词长=11，默认空位，e 值=10）重复 Wood 研究的一小部分，结果仅获得了黑猩猩和人类之间 89%的最大 DNA 序列同一性（Tomkins 2011d）。该值与 Wood（2011）摘要中报告的 98%以上的相似性相矛盾。Wood 和 Tomkins 报告的这些初步结果清楚地表明，有必要使用更广泛、更严格控制的 BLASTN 算法变量在该领域开展进一步研究。

基因组比较的理念和方法

为了对黑猩猩和人类基因组进行一次全新且更客观的比较，我们开展了一项研究，将黑猩猩全基因组鸟枪法测序（WGSS，与 Wood 2011 年使用的方法相同）与四个已发布的人类基因组版本进行比对。理论上，黑猩猩 WGSS 应该是随机的，因为它来源于物理剪切（片段化）的基因组 DNA，这些 DNA 被克隆到质粒测序载体中。我们可以从美国国家生物技术信息中心（NCBI；www.ncbi.nlm.nih.gov）下载一个包含 40,000 条精选

WGSS 序列的压缩文件。该文件中包含一个“TRACEINFO”xml 文件，其中描述了黑猩猩序列已完全处理——去除了低质量碱基和污染的载体序列。尽管这些文件被列为“跟踪序列”，但它们并非 Wood 在其 2011 年摘要中声称的原始未处理序列。因此，这些序列可以直接使用基本局部比对搜索工具(BLAST)算法(Altschul 等人，1990 年）进行查询，无需任何额外的处理来提高质量。

对于目标数据库，我们从 NCBI 的 FTP 存档中下载了最新的 BLAST 预格式化人类基因组组装存档文件。根据与 NCBI 工作人员的私人沟通，我们了解到该存档包含四个不同的人类基因组组装版本（GRCH37、GRCH36、Alternate Assembly 和 Celera Assembly）。这些组装版本均未进行低复杂度序列的预屏蔽处理，因此可以将四个不同的完整人类基因组组装版本纳入目标数据库，并测试序列屏蔽的影响。

虽然目前已知大部分基因组具有广泛的功能（Wells 2011），但在 BLASTN 搜索中使用低复杂度序列掩蔽技术会排除许多关键的遗传特征。BLAST 的原始开发者（Altschul 等，1994）提出使用低复杂度 DNA 序列掩蔽技术的理由是，这类序列常常会干扰进化分析，他们针对这些 DNA 特征作出了如下陈述：

这些片段中的大多数通常不能逐个位置地给出有意义的比对，以反映实际的结构和突变历史：它们显然进化得相对较快。

缺乏低复杂度序列掩蔽技术会导致大量额外 DNA 序列的纳入，从而显著增加计算处理资源的需求。然而，自 1994 年以来计算机硬件性能的提升使得大规模 DNA 分析成为可能，并且低复杂度序列的纳入也变得易于测试。因此，我们采用了两组独立的实验。一组实验对查询序列和目标序列均进行了掩蔽处理，而另一组实验则完全禁用了掩蔽技术。

启发式 BLASTN 算法非常适合对超大型 DNA 数据库进行计算量巨大的搜索——其目标是识别短 DNA 序列片段（例如平均长度为 740 个碱基的黑猩猩 WGSS）的局部相似区域。BLASTN 算法的工作原理是基于预设的“词长”（相同 DNA 碱基的数量）进行初始短匹配。这些初始种子匹配会沿两个方向依次扩展，直到比对结果不再具有显著性（基于预设的 e 值）或两个比对序列中的任何一个终止。虽然已有大量文献报道了 BLAST 算法在蛋白质相关相似性搜索中的应用和机制，但关于其核苷酸版本在大规模全基因组研究中的参数利用研究却很少。最初描述 BLAST 算法的论文（Altschul 等人，1990）至今仍是最具参考价值的文献之一。如

需了解 BLAST 算法的最新和更全面的综述，请参阅 Mitrophanov 和 Borodovsky（2006）。

为了获得全面的结果，我们决定采用多种 BLASTN 算法参数组合进行独立计算实验，这种方法与 Altschul 等人（1990）先前使用的方法大致相同。具体而言，我们测试了三种词长参数（7、11 和 15）和五种 e 值参数（1000、10、0.1、0.001 和 0.00001）的组合。e 值越低，算法执行的序列匹配就越严格、越精确。Altschul 等人（1990）在最初的 BLAST 论文中指出，词长和 e 值是任何基础 BLAST 分析中需要测试的关键算法参数。

总之，我们进行了两组各包含 15 种词长和 e 值组合的实验：一组实验对查询序列和目标序列都使用了掩码，而另一组实验则禁用了掩码。在所有 30 个独立的 BLASTN 实验中，我们针对四个不同版本的组装人类基因组（每个版本约 2.85 Gb）进行了总计 120 万次比对（每个实验 4 万次查询）。

鉴于全基因组分析需要大量的累积和比较数据，因此仅返回每个数据库比对结果的最佳比对结果（如果存在）。出于多种原因，不允许使用空位。首先，Altschul 等人（1990）确定，对于旨在利用 BLAST 定位局部相似区域的比对而言，添加空位策略的影响可以忽略不计。其次，对所有查询结果进行客观比较，可以否定

该算法中使用空位策略的必要性。最后，所获得的最佳局部成对比对结果涉及多种不同的匹配参数，从非常宽松到非常严格，包括词长和 e 值。

为了计算检验统计量（该统计量用于确定查询序列是否基于预设的 e 值获得显著匹配分数），我们使用了内置的标准核苷酸替换矩阵。在 NCBI 版本的 BLASTN 中，此功能不可自定义。早期非商业 WUBLAST 软件包中的 BLASTN 替换矩阵此前已针对微阵列探针的优化进行了定制（Ekland 等人，2010）。

材料与方法

最新独立版的 BLAST 软件包（ncbi-blast-2.2.25+）从 NCBI 软件库（ http://www.ncbi.nlm.nih.gov/guide/data-software/ ）下载，并安装在配备 20 GB 内存的双四核 Intel Xeon 处理器的 Apple Mac G5 台式电脑上。操作系统为 Mac OS 10.7，BASH shell 已更新至 4.2 版本，以便使用 renice 的高级功能进行进程/作业控制。BLASTN 作业采用半自动化方式运行，路径和算法参数的设置使用了 Tomkins 编写的 POSIX shell 脚本的变体（如有需要可提供）。BLASTN 的输出参数设置为 CSV（逗号分隔值）格式，以便在标准桌面电子表格软件中进行基本分析和绘图。

我们从 NCBI（ http://www.ncbi.nlm.nih.gov/Traces/ ）下载了包含 40,000 条序列的黑猩猩查询数据集，该数据集以 tar 压缩包（chimp_traces.tar.gz）的形式提供。黑猩猩序列压缩包解压后得到 fasta 格式的单个序列，然后使用标准的 POSIX shell 命令将它们连接成一个大的 fasta 格式文件。压缩包中包含“TRACEINFO”xml 文件，其中描述了黑猩猩序列已完全处理——去除了低质量碱基和污染的载体序列。尽管文件被列为原始序列，但它们实际上是经过完全处理的高质量序列。因此，这些序列可以直接使用 BLAST 算法（Altschul 等人，1990）进行查询，无需任何额外的质量提升处理。

我们从 NCBI 的 ftp 站点 ftp://ftp.ncbi.nih.gov/blast/db/下载了最新版本的 BLASTN 预格式化人类基因组组装 tar 压缩包（共 9 个文件；human_genomic.00 至 human_genomic.08）。根据与 NCBI 工作人员的私人通信，这些压缩包包含四种不同的人类基因组组装(GRCH37、GRCH36、Alternate Assembly 和 Celera Assembly）。私人通信还证实，这些数据库未进行任何预屏蔽处理。所有九个压缩包均被解压并部署到同一个目标数据库目录中。

表 1 和表 2 分别列出了测试的各种 BLASTN 参数设置及其输出结果。用于控制查询序列掩码的参数“-dust”设置为“no”或“yes”（默认值为“20 64 1”）。目标数据库掩码参数“-soft_masking”设置为“true”或“false”。查询作业的运行时间取决于掩码、字长和 e 值参数的设置，在 renice 设置为 -10 时，每个作业大约需要 2 到 6 天才能完成。通常情况下，使用 BLASTN CPU 优化（参数“-num_threads”）同时运行多个查询作业。

结果与讨论

人类与黑猩猩序列比对的最高一致性为 86% - 89%。

有关所有 30 个 BLASTN 实验的数据摘要，请参见表 1 和表 2，其中汇总了 40,000 个黑猩猩序列与人类基因组的四个不同版本进行 120 万次比对的尝试结果。

对于人类和黑猩猩基因组可比对区域的整体序列相似性，不同实验组在是否使用低复杂度序列掩蔽方面略有差异。在未进行掩蔽的实验组中，DNA 相似性从最低的 86.4%到最高的 88.9%不等，具体数值取决于词长和 e 值参数的组合（表 1）。鉴于近期研究表明人类和黑猩猩之间的关键差异存在于基因组的低复杂度区域（Polavarapu 等人，2011），使用未掩蔽的序列是

一个重要的考虑因素。在对查询序列和目标序列均采用低复杂度序列掩蔽的实验组中，DNA 相似性从最低的 86.2%到最高的 88.8%不等，具体数值取决于词长和 e 值参数的组合（表 2）。掩蔽的使用似乎对整体序列相似性统计数据有轻微影响。然而，最显著的差异在于计算处理时间，启用掩蔽后计算处理时间迅速缩短（数据未显示）。

表 1. 基于完整查询序列和目标序列（禁用掩码）的 BLASTN 比对结果。数据来自 40,000 条 WGSS 黑猩猩痕迹档案读段，比对对象为四个人类基因组组装版本（GRCh37、GRCh36、替代组装版本和 Celera 组装版本）。返回最佳数据库匹配结果（如有）。

E 值阈值	字号	热门歌曲数量	比对碱基的同一性百分比	每个查询序列的平均碱基匹配数	每个查询序列的平均比对碱基数	查询序列平均总长度（碱基数）
1000	7	40,000	87.2	109	125	740
10	7	40,000	87.2	109	125	740
0.1	7	36,437	86.8	118	136	740
0.001	7	29,095	86.4	140	161	740

0.00001	7	26,108	86.3	152	174	740
1000	11	40,000	87.6	109	125	740
10	11	40,000	87.6	109	125	740
0.1	11	35,788	87.1	119	137	740
0.001	11	28,507	86.5	142	163	740
0.00001	11	25,736	86.4	153	176	740
1000	15	40,000	88.9	107	122	740
10	15	39,999	88.9	107	122	740
0.1	15	33,508	87.9	123	141	740
0.001	15	26,740	87.1	147	168	740
0.00001	15	24,392	86.9	159	181	740

本研究中整体 DNA 序列相似性数值与之前几篇进化研究论文的结果相符，这些论文中，对于省略的数据，序列一致性可计算为 85-87%（Tomkins 和 Bergman，2012）。根据与 NCBI 工作人员的沟通，我们认为这 40,000 条黑猩猩序列很可能经过预筛选，并且已知在某种程度上与人类同源，尽管这一点无法通过进一步的查询来证实。鉴于在几种算法参数组合下（表 1 和表 2），所有 40,000 条序列均在人类基因组中找到了匹配序列，因此黑猩猩查询序列很可能已经过与人类 DNA 同源性的预筛选。此外，算法省略了每条 WGSS 黑

猩猩序列比对区域之外的大量数据。因此，约 86-89% 的最大一致性是一个非常保守且合理的估计。这些数据有力地证实，从全基因组的角度来看，经常吹捧的人类和黑猩猩之间 98-99% 的相似度估计是完全不准确的。

表 2. 基于对查询序列和目标序列均采用低复杂度序列掩蔽的 BLASTN 比对结果。数据来自 40,000 条 WGSS 黑猩猩痕迹档案序列，比对对象为四个人类基因组组装版本（GRCh37、GRCh36、替代组装版本和 Celera 组装版本）。返回最佳数据库匹配结果（如有）。

E 值阈值	字号	热门歌曲数量	比对碱基的同一性百分比	每个查询序列的平均碱基匹配数	每个查询序列的平均比对碱基数	查询序列平均总长度（碱基数）
1000	7	40000	87.1	109	125	740
10	7	40000	87.1	109	125	740
0.1	7	36111	86.7	119	136	740
0.001	7	28294	86.2	143	164	740
0.00001	7	25280	86.2	155	178	740
1000	11	39999	87.5	108	124	740

10	11	39997	87.5	108	124	740
0.1	11	34808	86.9	121	139	740
0.001	11	26901	86.2	148	169	740
0.00001	11	24264	86.2	159	183	740
1000	15	39997	88.8	106	121	740
10	15	39985	88.8	106	121	740
0.1	15	31361	87.5	129	147	740
0.001	15	24349	86.6	158	180	740
0.00001	15	22583	86.6	167	191	740

词长和 E 值的影响

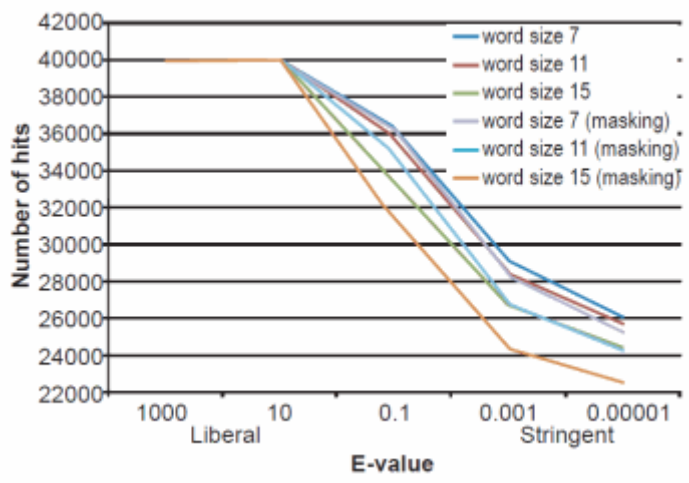

图 1. BLASTN 结果，展示了 e 值与匹配数之间的关系。最多可获得 40,000 个匹配结果。

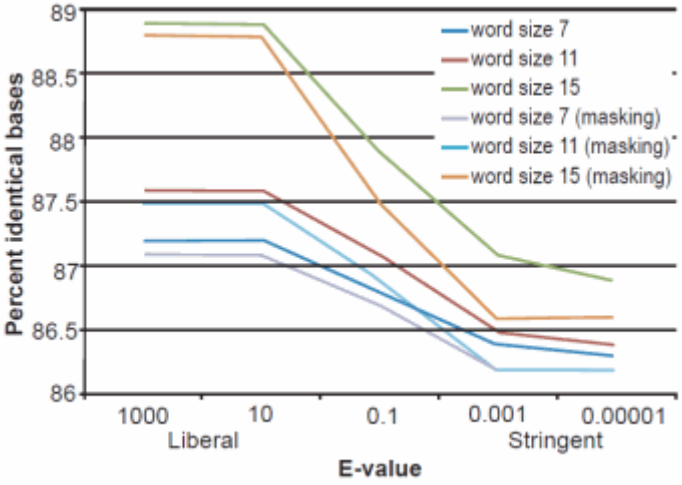

图 2. BLASTN 结果显示 e 值与平均序列一致性百分比之间的关系。

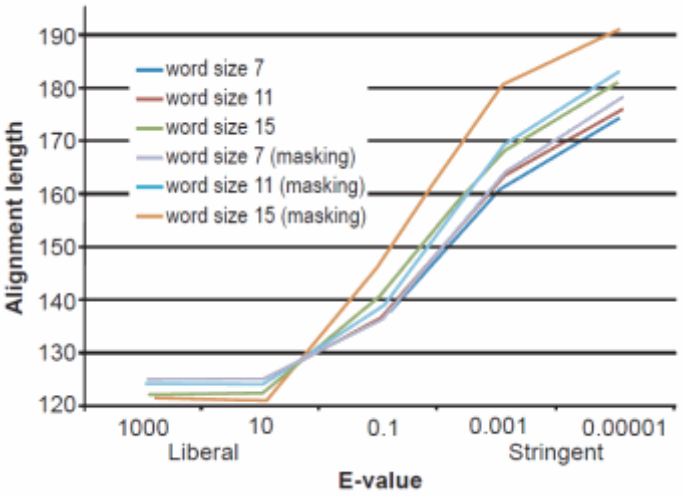

图 3. BLASTN 结果描绘了 e 值与平均比对长度之间的关系。

词长的影响相当显著，算法趋势与 Altschul 等人 (1990) 先前描述的结果高度一致。图 1、图 2 和图 3 以图形方式展示了所有实验中词长、e 值和序列掩蔽的影响。

在所有词长下，计算权衡的趋势都十分明显。随着 e 值越来越严格，匹配结果减少（表 1），比对结果的平均一致性百分比降低（图 2），平均比对长度增加（图 3）。然而，这些影响直到 e 值低于 10 时才出现。实际上，e 值为 10 和 1000 时，在所有词长下都产生了相同的结果（表 1 和表 2；图 1、图 2 和图 3）。如前所述，几乎所有 40,000 个序列在使用宽松的匹配参数时都产生了匹配结果，这表明查询序列之前已经过与人类同源性的筛选，这一观察结果也得到了与 NCBI 工作人员电子邮件沟通中提供的信息的证实。当然，一个显著的权衡是，使用宽松的匹配参数产生的比对结果长度也短得多。使用更高级别的严格性会大大延长比对时间，但也会降低序列一致性百分比和数据库中的阳性匹配数量。

BLASTN 查询实验中最有趣的方面之一是，即使在产生最长比对结果的条件下，平均也仅获得了 24%（181 个碱基，未屏蔽）和 26%（191 个碱基，屏蔽）的匹配结果（总共 740 个碱基）。而那些产生最高序列一致性和最多匹配结果的最宽松参数，也仅获得了 740 个碱基中的 16% 的比对结果。

默认 BLASTN 参数结果

NCBI 网站帮助文档中列出的 BLASTN 标准推荐默认参数针对多种搜索条件。对于标准核苷酸 BLAST，默认

词长为 11，e 值为 10，并结合序列掩蔽。本研究中的默认参数（表 1）产生了 40,000 个比对结果，与使用 e 值为 1000 的结果基本相同。在这些设置下，每个比对结果的平均序列一致性为 87.6%。然而，平均比对长度较短，仅为 740 个碱基中的 125 个。

NCBI 的帮助文档还建议，对于短序列或近乎精确的匹配，应使用 7 个碱基对的字长，e 值设为 1000（且不进行掩码处理）。这通常适用于需要精确靶序列的特定应用，例如开发基于 PCR 的实验室研究引物。总体而言，这些参数有助于所有 40,000 条序列的比对，比对一致性达到 87.2%，且比对长度最短为 125 个碱基。

需要注意的是，NCBI 上述两项默认参数建议旨在提高 NCBI 网络工具 BLAST 服务器（www.ncbi.nlm.nih.gov/BLAST）的在线搜索速度和便捷性。用户还可以通过 BLAST 在线服务器网站访问各种帮助页面的链接。

鉴于本研究中开展的 BLASTN 实验范围广泛，且所应用的查询类型多样，目前已发表的信息有限。显然，对于未来的此类研究，可以通过采用约 15 个单词的固定字长，并结合 10 到 0.00001 的 e 值范围，来获得更全面的结果，从而减少实验次数和计算资源的消耗。

用于全基因组查询的 BLAST 软件选项

本研究使用了 40,000 条平均长度约为 740 个碱基的 WGSS 黑猩猩序列，并将其与包含四种不同人类基因组组装版本的数据库进行比对。显然，这是一项计算量巨大的工作，由于 NCBI BLAST 网络服务器对作业大小和算法参数操作的限制，无法完成。此外，使用 UCSC 基因组浏览器（http://genome.ucsc.edu/ ）所采用的 BLAT 比对工具(类似 BLAST 的比对工具;Kent 2002)也因多种原因而不适用。首先，BLAT 算法使用索引数据库，其中省略了低复杂度序列。由于 BLAT 并非通过索引系统直接进行序列比对，而仅返回高度相似的匹配结果，因此本研究作者没有尝试将其安装为本地网络服务器并用于本研究。事实上，UCSC 网站在其 BLAT 服务器页面的“关于 BLAT”部分中对 BLAT 的限制作出了如下声明（http://genome.ucsc.edu/cgi-bin/hgBlat?command=start）。

> *DNA 上的 BLAT 旨在快速查找长度为 25 个碱基或更长、相似度达到 95%及以上的序列。它可能会遗漏差异较大或较短的序列比对。它可以找到 25 个碱基的完美匹配序列，有时甚至可以找到 20 个碱基的匹配序列。*

最后，要充分利用 BLASTN 算法的众多参数，并针对 DNA 序列进行实际 DNA 序列比对，最好的方法是使用 NCBI BLAST 套件的本地命令行。

总结与结论

本研究利用 BLASTN 算法，在 30 个独立的实验中，对黑猩猩和人类基因组进行了大规模的 DNA 序列比对分析。每个实验均在低复杂度序列掩蔽和非掩蔽条件下，使用不同的 e 值和词长组合，总共尝试了 120 万次比对。除了测试序列掩蔽外，还评估了三种不同词长（7、11 和 15）和五种不同 e 值（1000、10、0.1、0.001 和 0.00001）的 15 种组合。每个实验均返回了最佳比对结果及其对应的 e 值。

查询数据包含 40,000 条黑猩猩全基因组鸟枪法测序序列（WGSS），这些序列来自美国国家生物技术信息中心（NCBI）。随后，通过与 NCBI 工作人员的私人沟通以及本研究的支持数据，这些序列被确定为需要进行与人类基因组同源性预筛选的序列。这些黑猩猩序列在 30 个独立的实验中，针对四种不同的高质量人类基因组组装版本（GRCH37、GRCH36、Alternate SNP Assembly 和 Celera Assembly）进行了比对。

使用低复杂度序列掩蔽技术可使计算时间减少约 5 到 6 倍，略微延长比对长度（0 到 12 个碱基），减

少数据库命中次数（0 到 2,391 次命中），并略微降低核苷酸同一性百分比（0.1% 到 0.5%）。

根据 BLASTN 参数组合的不同，30 个独立实验中人类和黑猩猩之间的平均序列一致性在 86% 到 89% 之间变化。黑猩猩查询序列的平均长度为 740 个碱基，根据 BLASTN 参数组合的不同，平均比对长度在 121 到 191 个碱基之间变化。

排除未比对克隆的数量或未比对克隆中大量碱基的数据后，对全基因组人类-黑猩猩 DNA 相似性的保守估计值不超过 86-89%。这些估计值的保守性还体现在以下事实：所测试的 40,000 条黑猩猩序列均为预先筛选的、已知与人类基因组比对的同源序列。

读完这篇文章，你心里是否有一些触动？有没有一些新的想法，或者值得你认真思考的问题？或许，你也开始重新思考自己的信仰和人生的方向。

如果你愿意，现在就可以向上帝祷告，打开心门，成为祂的儿女。祷告不需要华丽的言辞，只要一颗真诚的心。你可以这样祷告：

天父上帝，

今天我来到你面前，愿意立定心志，宣告我相信耶稣基督是我的救主，是我生命的主。我愿意离开过去那

些不讨你喜悦的生活方式，求你赦免我的过犯。靠着你的恩典，帮助我学习顺服你、爱人如己，活出你所赐的新生命。求圣灵每天引导我、扶持我，使我一生荣耀你的名。奉主耶稣基督的名祷告，阿们。

如果你已经做了这个祷告，愿你知道，你并不孤单。信仰的道路需要陪伴和成长。鼓励你在自己居住的地方，寻找一间合适的教会，与弟兄姐妹一同聚会、学习和成长。

如果你有任何疑问，或在信仰上需要帮助，欢迎随时写信与我们联系。我们愿意倾听，也愿意与你一同前行。