动物学杂志 *Chinese Journal of Zoology* 2015, 50(3): 470 ~ 476

贝叶斯一致性分析法在全基因组系统发生树分析上的应用

章祎胤 徐福意 戚礼兴 李 凯* 周宇荀 肖君华

东华大学化学化工与生物工程学院 上海 201620

摘要：构建系统发生树时，其拓扑结构会在不同的基因组区域产生不一致性。对此问题，贝叶斯一致性分析法（BCA）可在全基因组规模上进行系统发生树分析，并进而对不一致性信息进行量化统计。采用此方法对由 C3H/Hu 小鼠（*Mus musculus*）和 129Sv 小鼠回交多代产生的 129S1 小鼠进行系统发生树分析，输入相应的一组序列文件，用若干生物信息学软件（如 VCFtools，RepeatMasker，PAUP*4.0，MrModelTest，MrBayes 等）对其进行屏蔽重复序列、序列比对等处理，辅以 Perl 语言脚本，最终得到全基因组范围不同区段系统发生树不一致信息。在小鼠 10 号染色体的所有 99 个基因座中，支持 129S1 和 129Sv 品系小鼠为姐妹关系的拓扑结构占了 84.7%（后验概率最高），这证明了 C3H/Hu 小鼠对 129S1 小鼠基因组的贡献程度较小。结果表明，贝叶斯一致性分析法有助于基因组不同区段进化历史的研究。

关键词：贝叶斯一致性分析法；系统发生树；小鼠基因组

中图分类号：Q811.4 **文献标识码：**A **文章编号：**0250-3263（2015）03-470-07

Utilization of Bayesian Concordance Analysis in Phylogenetic Analysis across Whole Genome

ZHANG Yi-Yin XU Fu-Yi QI Li-Xing LI Kai* ZHOU Yu-Xun XIAO Jun-Hua

College of Chemistry, Chemical Engineering and Biotechnology, Donghua University, Shanghai 201620, *China*

Abstract: During the construction of phylogenetic trees, there might be discordance of topologies in different genome regions. In order to address this issue, Bayesian concordance analysis (BCA) can be utilized to perform a phylogenetic analysis across whole genome and statistically quantification of the discordance. In this article, BCA was used to analyze the phylogenetic history of the strain of 129S1 (*Mus musculus*) which is originated from the backcross offspring of several generations between C3H/Hu strain and 129/Sv strain. Supplemented by Perl scripts, our pipeline took the genome sequence files as input and calls several bioinformatics software (e.g. VCFtools, RepeatMasker, PAUP*4.0, MrModelTest, MrBayes and so on) to

基金项目 国家自然科学基金项目（No. 31171199），上海市创新行动实验动物研究项目（No. 11140900200，13140900300），中央高校基本科研业务费专项资金，东华大学“励志计划”项目（No. B201308）；

* 通讯作者，E-mail：likai@dhu.edu.cn；

第一作者介绍 章祎胤，男，硕士研究生；研究方向：生物信息学；E-mail：dreamchaserzyy@sina.com。

收稿日期：2014-09-17，修回日期：2015-01-05 DOI: 10.13859/j.cjz.201503019

mask their repeat sequences, align sequences and so on. Then we obtained the phylogenetic discordance information of various locus across whole genome. Among all 99 loci in chromosome 10, 87.4% of loci were supported with a single topology of 129S1/129P2 (higher posterior probability), which is consistent with the hypothesis that the C3H/Hu mouse makes less contribution to the 129S1 genome. Our results indicate that BCA benefits the further studies on evolutionary histories of different genome regions.

Key words: Bayesian concordance analysis; Phylogenetic analysis; Mouse genome

随着测序技术的不断发展与成熟，大量基因组数据的获得使得研究人员能在基因组水平上构建高解析力的系统发生树，如对微生物（Comas et al. 2007）和动植物（Zou et al. 2008，Pollard et al. 2009）的研究。然而，不同群体间基因渗入与重组等生物学现象的存在，易导致不同基因区段发生系统树拓扑结构的不一致。此类由单基因构建的系统发生树（基因树）的冲突，在快速分化的近缘种分析中较为常见。不仅在无脊椎动物，如在果蝇（*Drosophila erecta*）、黑腹果蝇（*D. melanogaster*）和 *D. yakuba* 果蝇中的 9 405 个基因之间出现了显著不同（Pollard et al. 2009），即便在高等哺乳动物间，如人类（*Homo sapiens*）、大猩猩（*Gorilla gorilla*）和黑猩猩（*Pan troglodytes*），也有大量的基因树不一致性被报道（Chen et al. 2001，Wall 2003，Patterson et al. 2006，Hobolth et al. 2007）。尽管如此，这种基因组水平上的不一致性却很少被量化统计（White et al. 2009），导致至今对这种基因组范围上的不一致程度依然知之甚少。

传统构建系统发生树的方法有两种：总证据法（total evidence approach）和一致性法（consensus method）。总证据法主张将所有的数据串联起来而无视不同基因座之间不同的进化可能性，从而构建出一个系统发生树。该方法忽视了诸如杂交、不完全谱系分选和基因的水平转移等生物学现象，从而往往造成这个方法构建出的系统发生树与真实的物种进化历史不一致（Degnan et al. 2006，Kubatko et al. 2007）。而与总证据法相反，一致性法主张对于每个基因进行单独分析，对每个基因作出最适估计（Ané et al. 2007）。虽然此法保留了每个基因潜在拓扑结构的多样性，但是缺少一个客观的方法去整合单个基因树中的不确定性，并对其系统发生树进行估计；另外，由于每个基因是独立分析的，导致每个基因所支持的系统发生树的信息并没有被其他基因所共享（Ané et al. 2007），亦不能有效反应来自同一个物种不同基因的共同进化历史。

贝叶斯一致性分析法（Bayesian concordance analysis，BCA）是在一致性法的基础上加以改进，在每棵基因树的估计中加入了不确定性参数并允许基因树之间相互影响，进而在全基因组范围对物种的系统发生树进行估计（Ané et al. 2007）。具体原理上，先利用贝叶斯方法对基因组的单基因进行系统发生估计，得到所有单基因系统发生的后验分布，然后该后验分布作为第二次贝叶斯分析的先验分布，最终通过第二次贝叶斯分析得到联合后验概率分布（Weisrock 2012）。近年来，BCA 最常应用于研究独立进化的基因，以及在物种的全基因组范围上定量分析系统发生树的不一致性（Horvath et al. 2008，Cranston et al. 2009，Jacobsen et al. 2011，Wielstra et al. 2014）。

据文献报道，在构建 129 品系的小鼠过程中形成了三个亚品系（substrain），分别为 Parental（P）、Steel（S）和 *Ter*（T）。其中 S 亚品系中的 129S1 是由 C3H/Hu 与 P 亚品系的 129/Sv 回交 12 ~ 14 代产生的（Simpson et al. 1997）。由于 129P2 与 129Sv 同属 P 亚品系，C3HHeJ 与 C3H/Hu 亲缘关系较近（Simpson et al. 1997），且 129P2 与 C3HHeJ 全基因组序列已知，为了获取数据的方便性，我们利用 129P2

品系小鼠基因组序列代替 129/Sv 小鼠，C3HHeJ 品系小鼠基因组序列代替 C3H/Hu，并采用贝叶斯一致性分析法寻找 129S1 小鼠的系统发生历史以及评估 C3H/Hu 小鼠的基因组对 129S1 小鼠基因组的贡献程度。

1 材料与方法

1.1 相关软件的获取

本方法所涉及软件较多。总体分析流程如图 1 所示。获得全基因组数据后须对数据进行以下处理：(1)使用 VCFtools（Danecek et al. 2011）获得基因组一致性序列。(2)用 RepeatMasker 将参考序列及一致性序列中的重复序列去除。(3)使用 Mercator/MAVID（Bray et al. 2004）进行不同物种基因组之间的比对，产生同源的区块。(4)然后使用 MDL（Ané 2011）软件通过调用 PAUP*4.0（Swofford 2003）对各个同源区块进行基于最小描述长度原则（Minimum Description Length）的进一步分区，以划分为不同基因座。(5)使用 MrModelTest 对每一个划分出的基因座进行模型选择，然后通过 MrBayes（Huelsenbeck et al. 2001，Ronquist et al. 2003）进行分析。(6)使用 BUCKy（Ané et al. 2007，Larget et al. 2010）对物种的主要系统发生历史进行评估，并找出支持不同拓扑结构的基因组比例。上文中所提到的软件，除了 PAUP*4.0 外，均是可从网络免费下载并进行编译安装的。表 1 列出了所用的软件及其下载网址。安装及编译方法可在所下载的软件压缩包中的 README 中获得，在此不再赘述。

1.2 数据获取及相关处理

本方法所需的数据几乎都是从网络上下载而来。其中，为了获得数据的方便性，我们用与 C3H/Hu 小鼠亲缘关系更近的 C3HHeJ 小鼠的数据代替 C3H/Hu 小鼠的数据；以 129P2 小鼠代表 129P 亚品系。另外，我们将 SPRETEiJ 品系小鼠（*M. spretus*）和大鼠（*Rattus norvegicus*）作为外类群。小鼠参考序列以及相关品系小鼠的 SNP 信息（VCF 文件）下载自英国维尔康姆基金会桑格研究所（The Wellcome Trust Sanger Institute）的网站（http:// www.sanger.ac.uk/）；大鼠参考序列及大鼠和小鼠的序列注释文件（GFF 或 GTF 格式文件）下载自 Ensembl Genome Browser。其中小鼠参考序列的版本号为 GRCm38_68。通过使用 VCFtools，将从 Sanger Institute 网站上下载的 SNP 信息整合到小鼠参考序列中，以产生相关品系的一致性序列。然后用 RepeatMasker 将小鼠参考序列和产生的一致性序列中的重复序列屏蔽。最后利用 ABBlast 和 Mercator 程序包中的相关程序分别产生 hard-masked 和 soft-masked 的序列。

表 1 贝叶斯一致性分析法所需软件及其下载网址

Table 1 The required software for Bayesian concordance analysis and their download sites

软件名 Softwares	下载网址 Download sites
VCFtools	http://sourceforge.net/projects/vcftools/
RepeatMasker	http://repeatmasker.org/RMDownload.html
Mercator	https://www.biostat.wisc.edu/~cdewey/software.html
MAVID	http://bio.math.berkeley.edu/mavid/download/
MDL	http://www.stat.wisc.edu/~ane/
PAUP*4.0	http://paup.csit.fsu.edu/
MrModelTest	https://github.com/nylander/MrModeltest2
MrBayes	http://mrbayes.sourceforge.net/
BUCKy	http://www.stat.wisc.edu/~ane/bucky/index.html

1.3 全基因组间比对

由于大鼠的参考序列中不含有 Y 染色体，所以我们在比对中也将小鼠参考序列中的 Y 染色体去掉。另外，在比对中也需去除参考序列中的一些未定位的随机序列。随后利用 Mercator 产生直系同源图谱，MAVID 进行基因组间比对。整个过程不超过 1 d。随后利用 Perl 脚本将各品系的 soft-masked 序列映射（mapping）到小鼠参考序列比对完的结果中，并随即产生各个区段的 NEXUS 文件（Maddison et al. 1997），以便后续分析。

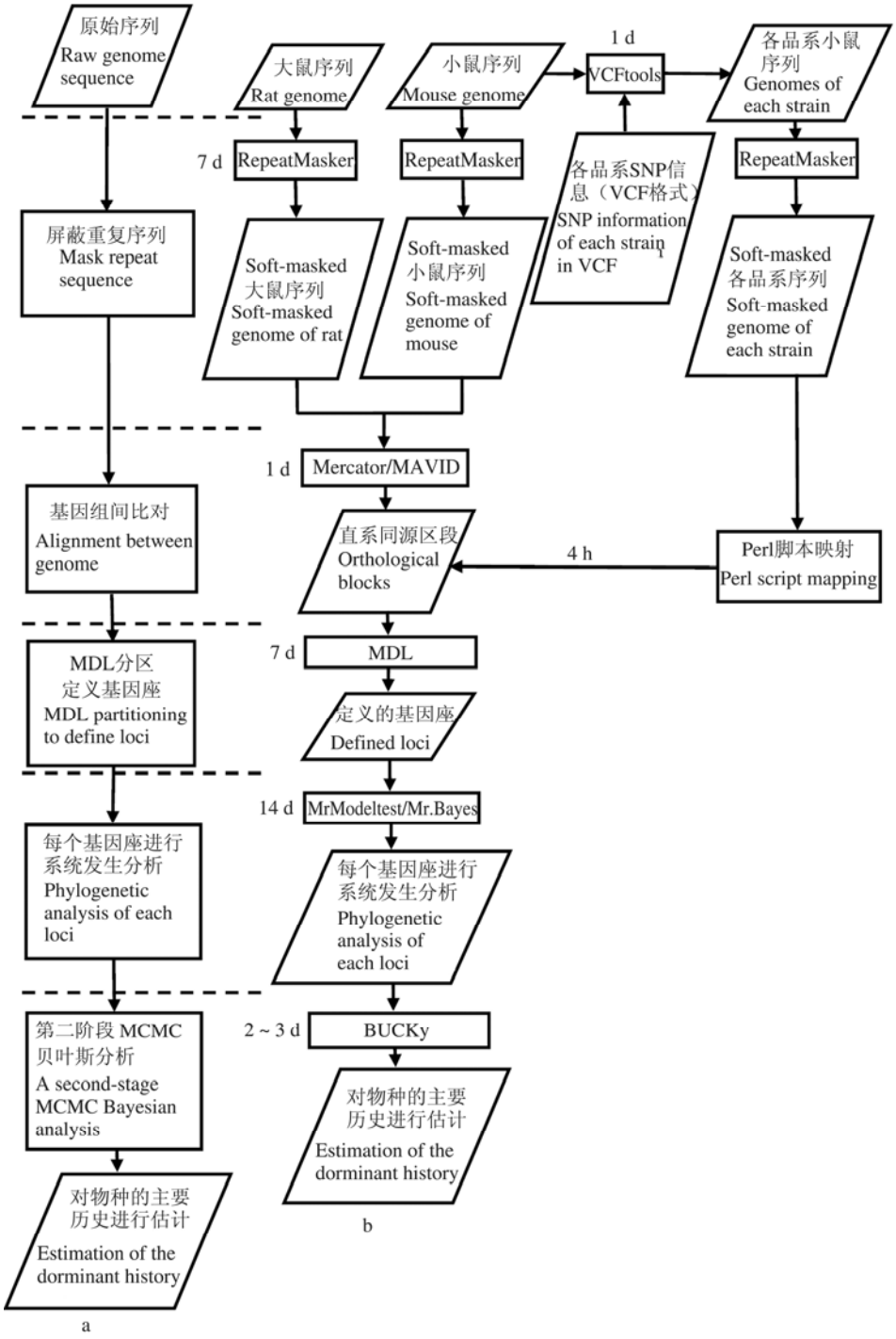

图 1　贝叶斯一致性分析法流程图

Fig. 1　The flow diagram of Bayesian concordance analysis

a 图表示的是贝叶斯一致性分析法的大致流程。b 图表示的是文中应用贝叶斯一致性分析法研究 129 小鼠系统发生历史的具体步骤。虚线划分了每一步对应的具体范围。每一步所需的大致时间都已在其旁边标注。

Fig. a shows the general procession of Bayesian concordance analysis. Fig. b shows the detailed procedure of the application of BCA in this paper to study the phylogenetic history of 129 mice. And dashed lines partition the specific realm of each step. The roughly necessary time of each step has been marked aside.

1.4 最小描述长度分区

利用 MDL 将每个所产生的直系同源区段 NEXUS 文件进行进一步的分区。MDL 首先会基于一个参数，ncharbase，对区段进行划分。如一区段长度为 60 500，ncharbase 设为 30 000，则该区段首先会被分为 6 个分区，第一个分区长度为 30 000，位置从 1 到 30 000 个碱基；第二个分区长度为 60 000，从 1 到 60 000；第三个分区长度为 60 500，从 1 到 60 500；第四个分区长度为 30 000，从 30 001 到 60 000；第四个分区长度为 30 500，从 30 001 到 60 500；第六个分区长度为 500，从 60 001 到 60 500。然后调用 PAUP*4.0 对每个分区进行计算枝长，即描述长度。这一步可能非常缓慢，随着分区长度的增加，计算的时间也会相应增长。最后 MDL 会从中挑出枝长最小的分区。ncharbase 参数默认值为 1 000。为了兼顾效率和准确性，我们对不同长度的区段设置了不同长度的 ncharbase 值。通过产生的分区文件（mb 文件）编写 Perl 脚本对 NEXUS 文件进行调整，使其适合后续的分析。

1.5 单基因座系统发生分析

将从 MDL 分区后的各个基因座分别通过 MrModelTest 进行 DNA 替换模型选择。选择基于赤池信息量准则（Akaike Information Criterion，AIC）分值最高的模型（Posada et al. 2004）。然后用 MrBayes 用 4 条马尔科夫链运行 100 000 代，取样频率为 100，舍弃前总样本数的 25%作为“老化”样本（burn-in）之后分别对参数值和各进化树进行总结。其余参数都为默认。两个程序之间需应用 Perl 脚本对每个基因座的 NEXUS 文件格式做出相应的调整。

1.6 第二阶段贝叶斯分析

每个基因座的后验分布作为使用 BUCKy 进行第二阶段 MCMC 贝叶斯分析的输入。贝叶斯一致性模型整合了基因树一致性的先验分布 α。当 α 为无穷（α = infinite）时，代表先验分布中基因树之间完全独立；当 α 非常小时（如 $\alpha = 10^{-6}$ 时），代表先验分布中的基因树几乎完全一致。我们选取在这两种极端情况之间居间的值 α = 1。最后通过 Perl 脚本对输出文件进行整理统计。

2 结果与讨论

运行完整个流程，得到以下一系列文件及结果，包括：RepeatMasker 产生的各品系和参考序列的 soft-masked 和 hard-masked 序列文件；Mercator/MAVID 比对过程产生的各个直系同源区块文件（mavid 文件）和同源图谱（map 文件）以及随后由 Perl 脚本产生的 NEXUS 文件；MDL 分区过程产生的分区文件（mb 文件）以及随后用 Perl 脚本产生的 NEXUS 文件；由 MrModelTest 模型选择产生的 mrmodel.scores 文件以及随后由 Perl 脚本产生的用于 MrBayes 分析的 NEXUS 文件；由 MrBayes 产生的 t 文件、p 文件以及其他相关文件；由 BUCKy 中的 mbsum 程序产生的总结文件和 bucky 程序产生的 concordance 文件、cluster 文件、pairs 文件以及 gene 文件。

这些文件中最值得关注的主要是最后产生的 gene 文件。它对每个基因座支持的拓扑结构做了最后的总结。根据这个文件，我们可以通过编写 Perl 脚本排除比对中的 gap 并做出相关统计，然后通过使用 R 语言作图，使数据图形化直观化。使人们能更清楚地了解不同基因片段的来源。

图 2 即为对其 10 号染色体序列进行的 BCA 统计结果。显示绝大部分基因组分区都支持 129S1 和 129P2 为主要姐妹关系的拓扑结构；少数支持 129S1 和 C3HHeJ 为主要姐妹关系的拓扑结构。说明 129S1 小鼠 10 号染色体中绝大部分区段都是源自于 129 品系的小鼠，少部分来自于 C3H/Hu 小鼠。图 3 展示是其 10 号染色体精细规模的系统发生树不一致性。少量的支持 129S1 和 C3HHeJ 为主要拓扑结构的区段分散在大片支持 129S1 和 C3HHeJ 为主要拓扑结构区段中间。说明 129S1 小鼠在培育和保存过程中渗入了少量 C3HHeJ 小鼠基因。这和

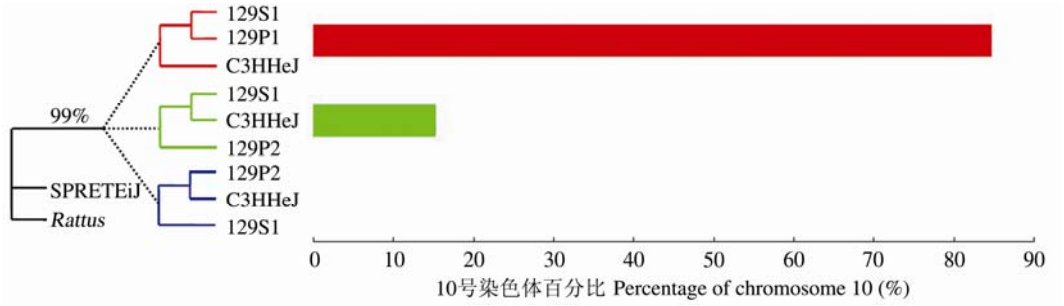

图 2　小鼠 10 号染色体的系统发生树基因组分区

Fig. 2　Genomic partioning of phylogenetic history in the chromosome 10 of mice

我们从小鼠的 10 号染色体上 99 个基因座点树中估计出贝叶斯一致性因子。横坐标表示小鼠 10 号染色体中对于三种拓扑结构支持率的百分比。约 99%的基因座都把 SPRETEiJ 品系小鼠和大鼠作为了外类群。其中，支持 129S1 和 129P2 为姐妹关系的拓扑结构占了 84.7%（后验概率最高），而其他两种拓扑结构分别占了 15.3%（129S1/C3HHeJ）和 0（129S2/C3HHeJ）。

Bayesian concordance factors were estimated from 99 individual locus trees in the chromosome 10 of mice. The X-axis represents the percentage of the loci in mouse chromosome 10 which support the topologies. About 99% of loci place SPRETEiJ and *Rattus* as the outgroup to the *M. musculus* subspecies. Within *M. musculus*, 84.7% of loci were supported with higher posterior probability of a single 129S1/129P2 topology. 15.3% of loci were supported a single 129S1/C3HHeJ topology and none of loci were supported a single 129S2/C3HHeJ topology.

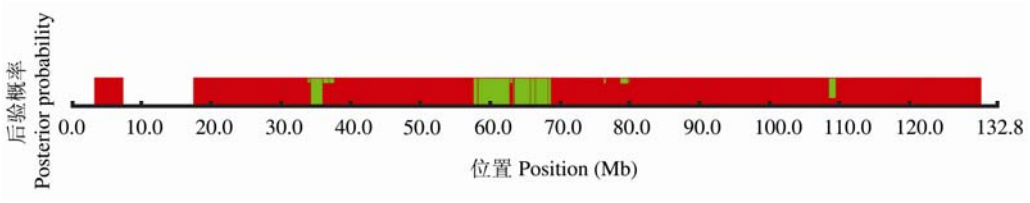

图 3　小鼠 10 号染色体精细规模的系统发生树不一致性

Fig. 3　Fine-scale phylogenetic discordance of the chromosome 10 of mice

我们将 10 号染色体上每个基因座拓扑结构的后验概率（PP）映射到其染色体上，以表征其 99 个基因座的不一致性模式。其中的颜色与图 2 中的颜色相对应。

The posterior probability (PP) of each topology is mapped onto the chromosome 10 to characterize the fine-scale patterns of discordance among the 99 loci. The colors are corresponding to those of the 3 topologies in the Fig. 2.

文献中所报道的 129S1 小鼠的育成过程，即与 C3H/Hu 小鼠杂交后再回交十几代的过程相符合。

BCA 方法整合了大量序列操作软件和系统发生分析软件，除了对操作者的软件操作和对软件的理解有较高要求以外，对操作者的 Perl 语言和 R 语言也有较高要求。深入理解每一步所涉及的软件中的各种参数有助于其更好的运行与分析，例如 MrBayes 参数的设置及调整等，这将更有助于提高分析的效率。

本方法的一个关键点在于群体内各品系的序列需一样长。本文通过各品系小鼠的 SNP 数据导入参考序列产生各品系的序列。对于其他物种也应以参考序列为标准，导入相应序列变异以有利于后期处理。

参　考　文　献

Ané C. 2011. Detecting phylogenetic breakpoints and discordance from genome-wide alignments for species tree reconstruction. Genome Biology Evolution, 3: 246–258.

Ané C, Larget B, Baum D A, et al. 2007. Bayesian estimation of

concordance among gene trees. Molecular Biology and Evolution, 24(2): 412–426.

Bray N, Pachter L. 2004. MAVID: Constrained ancestral alignment of multiple sequences. Genome Research, 14(4): 693–699.

Chen F C, Li W H. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. American Journal of Human Genetics, 68(2): 444–456.

Comas I, Moya A, Gonzalez-Candelas F. 2007. From phylogenetics to phylogenomics: the evolutionary relationships of insect endosymbiotic gamma-Proteobacteria as a test case. Systematic Biology 56(1): 1–16.

Cranston T A, Hurwitz B, Ware D, et al. 2009. Species trees from highly incongruent gene trees in rice. System Biology, 58(5): 489–500.

Danecek P, Auton A, Abecasis G, et al. 2011. The variant call format and VCFtools. Bioinformatics, 27(15): 2156–2158.

Degnan J, Rosenberg N. 2006. Discordance of species trees with their most likely gene trees. PLoS Genetics, 2(5): e68.

Hobolth A, Christensen O F, Mailund T, et al. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genetics, 3(2): e7.

Horvath J E, Weisrock D W, Embry S L, et al. 2008. Development and application of a phylogenomic toolkit: resolving the evolutionary history of Madagascar's lemurs. Genome Research, 18(3): 489–499.

Huelsenbeck J P, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics, 17(8): 754–755.

Jacobsen F, Omland K E. 2011. Species tree inference in a recent radiation of orioles (Genus Icterus): multiple markers and methods reveal cytonuclear discordance in the northern oriole group. Molecular Phylogenetics and Evolution, 61(2): 460–469.

Kubatko L S, Degnan J. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Systematic Biology, 56(1): 17–24.

Larget B R, Kotha S K, Dewey C N, et al. 2010. BUCKy: Gene tree/species tree reconciliation with the Bayesian concordance analysis. Bioinformatics, 26(22): 2910–2911.

Maddison D R, Swofford D L, Maddison W P. 1997. NEXUS: an extensible file format for systematic information. Systematic Biology, 46(4): 590–621.

Patterson N, Richter D J, Gnerre S, et al. 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature, 441(7097): 1103–1108.

Pollard D A, Iyer V N, Moses A M, et al. 2009. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. PLoS Genetics, 2(10): e173.

Posada D, Buckley T R. 2004. Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests. Systematic Biology, 53(5): 793–808.

Ronquist F, Huelsenbeck J P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics, 19(12): 1572–1574.

Simpson E M, Linder C C, Sargent E E, et al. 1997. Genetic variation among 129 substrains and its importance for targeted mutagenesis in mice. Nature Genetics, 16(1): 19–27.

Swofford D L. 2003. PAUP*: Phylogenetic analysis using parsimony, beta version 4.0b10. Sunderland, Massachusetts: Sinauer Associates. [CP/OL] [2014-08-07]. http://paup. csit.fsu.edu/ paupfaq/faq.html.

Wall J D. 2003. Estimating ancestral population sizes and divergence times. Genetics, 163(1): 395–404.

Weisrock D W. 2012. Concordance analysis in mitogenomic phylogenetics. Molecular Phylogenetics and Evolution, 65(1): 194–202.

White M A, Ané C, Dewey C N, et al. 2009. Fine-scale phylogenetic discordance across the house mouse genome. PLoS Genetics, 5(11): e1000729.

Wielstra B, Arntzen J W, van der Gaag K J, et al. 2014. Data concatenation, bayesian concordance and coalescent-based analyses of the species tree for the rapid radiation of triturus newts. PLoS One, 9(10): e111011.

Zou X H, Zhang F M, Zhang J G, et al. 2008. Analysis of 142 genes resolves the rapid diversification of the rice genus. Genome Biology, 9(3): R49.