Published in final edited form as:

Stat Appl Genet Mol Biol. 2004 ; 3: Article23.

Mammalian Genomes Ease Location of Human DNA Functional Segments but Not Their Description*

Lee A. Newberg* and Charles E. Lawrence†

*New York State Department of Health Wadsworth Center & Rensselaer Polytechnic Institute
Department of Computer Science, Lee.Newberg@wadsworth.org

†Brown University Division of Applied Mathematics, Lawrence@DAM.Brown.EDU

Abstract

Under the assumption that a significant motivation for sequencing the genomes of mammals is the resulting ability to help us locate and characterize functional DNA segments shared with humans, we have developed a statistical analysis to quantify the expected advantage. Examining uncertainty in terms of the width of a confidence interval, we show that uncertainty in the rate of nucleotide mutation can be shrunk by a factor of nearly four when nine mammals; human, chimpanzee, baboon, cat, dog, cow, pig, rat, mouse; are used instead of just two; human and mouse. Contrastingly, we show confidence interval shrinkage by a factor of only 1.5 for measurements of the distribution of nucleotides at an aligned sequence site. These additional genomes should greatly help in identifying conserved DNA sites, but would be much less effective at precisely describing the expected pattern of nucleotides at those sites.

Keywords

Comparative genomics; Sequence Conservation; Sequence pattern identification

1 Introduction

Comparisons of multiple related genomes have already produced a number of interesting findings, and sequencing resources are available to obtain the genomes of many more species. For studies of human disease, there is naturally a strong interest in the genomes of vertebrates, especially mammals. Decisions concerning the particular species to sequence depend on a number of important factors. While much useful and constructive discussion about these choices has ensued, there have been few quantitative analyses addressing this issue.

Here we consider two of these factors: 1) pattern discovery of functional elements, such as transcription factor binding site models, and 2) identification of unusually conserved sequence fragments. To address these issues, we examined the evolutionary/phylogenetic tree of nine mammals; human, mouse, dog, cow, pig, rat, cat, baboon, and chimpanzee; the last seven of which are being sequenced in the NISC Comparative Sequencing Program (Thomas et al., 2003), and statistically ask how these would improve confidence intervals for estimates of meaningful statistics, relative to estimates made from just human and mouse.

*The authors gratefully acknowledge the use of the Computational Molecular Biology Core of the Wadsworth Center. C.E.L. is supported in part by NIH (NHGRI) grant 5R01HG00125707 and DOE grant DEFG0201ER63204. The authors wish to thank the anonymous referees for their constructive criticism.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, bepress. *Statistical Applications in Genetics and Molecular Biology* is produced by The Berkeley Electronic Press (bepress). <http://www.bepress.com/sagmb>

State-of-the-art statistical phylogenetic models are quite advanced and there exist several good approaches to modeling the evolution of biopolymer sequences. Many are based upon the mathematical model of Neyman (1971) and algorithm of Felsenstein (1981) and include ever more sophisticated models for nucleotide substitution (Jukes & Cantor, 1969; Kimura, 1980; Felsenstein, 1981; Lanave et al., 1984; Hasegawa et al., 1985; Barry & Hartigan, 1987; Kishino & Hasegawa, 1989; Rodríguez et al., 1990; Tamura & Nei, 1993; Yang, 1994a,b, 1995; Gu et al., 1995; Felsenstein & Churchill, 1996; Arvestad & Bruno, 1997; Waddell & Steel, 1997; Arvestad & Bruno, 1997; Friedman et al., 2002; Holmes & Rubin, 2002; Siepel & Haussler, 2004).

For any given sequence position, we are interested in estimating confidence interval shrinkage for two parameters, (1) the value γ , which is the expected rate of nucleotide mutation at that position and (2) the vector $(\theta_A, \theta_T, \theta_C, \theta_G)$, which is the expected equilibrium distribution of nucleotides at that sequence position. As such we will employ a statistical model that is parameterized by only γ , θ , and the phylogenetic tree edge lengths; it is a simple extension of the early nucleotide substitution model of Felsenstein (1981). See Equation 1 in the Methods Section for more information.

For the set of tree edge lengths we use the result of Siepel & Haussler (2004) for non-coding sequence, which is depicted in Figure 1. Each edge in this tree represents the expected number of mutations per sequence position (for non-coding sequence) between an ancestral and descendant individual.

We combine these ingredients in an analytical approach to evaluate the precision to which the rate of mutation and nucleotide distribution can be determined. Specifically we imagine the following thought experiment: We generate sequence data according to our model, as parameterized by some γ^* and θ^* . Ignoring the parameters used to generate the data, but examining this instance of randomly generated data we try to rediscover the parameter values from which the data was generated. As estimates we chose the “maximum likelihood” values of γ and θ , which comprise the parameterization of the model that best explains this generated data. We imagine that this thought experiment is repeated many times and we ask how close the estimates (γ, θ) are to the true values (γ^*, θ^*) . The closer these estimates bunch about the true values, the more confidence we will have in estimates that we derive from a single instance of real data.

Note that instead of actually generating random sequence we use an asymptotic approach via the Fisher information Matrix. (See the Methods Section 2.6 for more details.) One significant advantage of using this asymptotic approach is the Cramer-Rao Theorem (Stuart et al., 1999); any estimates of the parameters cannot have a better standard deviation from the true value than that standard deviation indicated by the Fisher information matrix approach.

2 Methods

2.1 Phylogenetic Tree

Using ancestral repeats data from the NIH Intramural Sequencing Center (NISC) Comparative Sequencing Program, (Thomas et al., 2003), a phylogenetic tree for nine mammals was determined, (Siepel & Haussler, 2004), which we depict in Figure 1. The tree was constructed with the model described as “U3S with rate variation”; when used to evaluate 162,743 bp of non-coding sequence data, this tree gives the maximum likelihood among all phylogenetic models that employ a not-necessarily-reversible, strand-symmetric, trinucleotide model of mutation, with rate variation according to the approach of Yang (1994b). See Siepel & Haussler (2004) for more information. The estimated overall distribution of nucleotides for this tree is:

$\theta_A = 0.2967$, $\theta_T = 0.3122$, $\theta_C = 0.1949$, and $\theta_G = 0.1962$. The model of Siepel & Haussler (2004) assigns this equilibrium distribution to each species in the tree.

2.2 Nucleotide Substitution Model

We use a nucleotide substitution model parameterized by γ , θ , and a tree edge's length x .

$$M_x = \begin{pmatrix} \Pr[A|A] & \Pr[T|A] & \Pr[C|A] & \Pr[G|A] \\ \Pr[A|T] & \Pr[T|T] & \Pr[C|T] & \Pr[G|T] \\ \Pr[A|C] & \Pr[T|C] & \Pr[C|C] & \Pr[G|C] \\ \Pr[A|G] & \Pr[T|G] & \Pr[C|G] & \Pr[G|G] \end{pmatrix} = e^{-k\gamma x} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + (1 - e^{-k\gamma x}) \begin{pmatrix} \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \\ \theta_A & \theta_T & \theta_C & \theta_G \end{pmatrix} \quad (1)$$

where

x = the evolutionary distance between an ancestral and a descendant individual,

$\Pr[d|a]$ = the probability that a descendant individual shows the nucleotide d
when the ancestral individual shows the nucleotide a ,

θ_b = the equilibrium probability of nucleotide b

γ = the relative rate of mutation, and

$$k = \frac{1}{1 - \theta_A^2 - \theta_T^2 - \theta_C^2 - \theta_G^2}.$$

This model has the necessary features that, when $\gamma = 1$, the expected number of substitutions is x , because of the choice for k ; that as $x \rightarrow 0^+$, the substitution matrix is the identity matrix; that as $x \rightarrow +\infty$, the substitution matrix gives an equilibrium distribution independent of which nucleotide we started with (*i.e.*, all of the rows are equal); and that $M_{a+b} = M_a M_b$, correctly modeling that the substitution resulting from evolution described by an evolutionary distance a , followed by evolution described by an evolutionary distance b , is equal to the evolution described by the sum of the evolutionary distances.

2.3 Likelihood of a Phylogenetic Tree

The algorithm of Felsenstein (1981) is used calculate the probability of observing aligned sequence data at a single alignment position across all leaf species. Writing D_r for this aligned set of nucleotides (one nucleotide for each leaf species in the tree), the algorithm allows us to calculate $\Pr[D_r]$ given specific values for γ , θ , and the tree edge lengths T , recursively. Briefly:

$$\Pr[D_r] = \sum_{b_r \in \{A, T, C, G\}} \Pr[D_r | b_r] \theta_{b_r} \quad (2)$$

$$\Pr[D_p | b_p] = \begin{cases} \prod_{c \in \{p \text{ s children}\}} \Pr[D_c | b_p] & \text{if } p \text{ has children} \\ 1 & \text{if } p \text{ is a leaf} \end{cases} \quad (3)$$

$$\Pr[D_c | b_p] = \sum_{b_c \in \{A, T, C, G\}} \Pr[D_c | b_c] \Pr[b_c | b_p] \quad (4)$$

where

r = the root species of the phylogenetic tree,
 p = any species of the tree,
 c = any immediate child species of p ,
 b_p = a nucleotide for species p ,
 D_p = the subset of D_r corresponding to leaf species descendants of p ,
 $I(D_p = b_p)$ = The indicator function which is 1 if, in D , leaf species p
 has nucleotide b_p , and is 0 otherwise, and
 $\Pr[b_c \mid b_p]$ = the entry from M_x (Equation 1) for x
 equal to the length of the edge connecting species p and c .

For more information see, *e.g.*, Durbin et al. (1998) or Swofford et al. (1996).

2.4 Asymptotic y Estimator Variance

Following the approach described in Stuart et al. (1999) we measure our ability to recover the γ parameter, from observed data via an expected log likelihood function $LL(\gamma|\gamma^*)$:

$$\text{LL}(\gamma \mid \gamma^*) = \sum_{\mathbf{D}} \log(\text{Pr}[\mathbf{D} \mid \gamma]) \text{Pr}[\mathbf{D} \mid \gamma^*]. \quad (5)$$

Intuitively, we are measuring how well data generated with a model parameterized by γ^* can be explained by the model parameterized with γ . Our confidence interval is tight if this function falls off quickly as γ deviates from its mode at γ^* . This rate of decline is measured by the second derivative of $\text{LL}(\gamma|\gamma^*)$ with respect to γ , evaluated at $\gamma = \gamma^*$, and the negative of the reciprocal of this second derivative is the asymptotic variance for the γ estimator.

To calculate the variance reduction factor for all nine mammals relative to just human and mouse, we perform the following operations twice, once for the tree of nine mammals and once for the tree of just human and mouse. For each possible tuple of nucleotides, one nucleotide for each leaf species in the tree, using the algorithm of Felsenstein (1981) we evaluate $\Pr[D|\gamma]$ for three values of γ , specifically, $\gamma = 0.99999$, 1.0 , 1.00001 . For each of these three values for γ we compute the sum of Equation 5 assuming $\gamma^* = 1.0$, and we compute the second derivative with respect to γ , numerically. In these calculations, we fix ϑ to the values given in Section 2.1.

2.5 Combining θ Estimator Variances

In Section 2.6 we will describe how we compute the asymptotic variances and covariances for the components of ϑ but first we describe in this section why we have made the somewhat arbitrary decision to use the sum of the four pure variances in the computation of variance reduction factor ratios. The product of the (pure) variances and the determinant of the covariance matrix (i.e., the volume of the confidence ellipsoid) are two obvious alternatives. (For more see, e.g., Chapter 2 in Silvey (1980).) We settled upon the sum of the individual variances for several reasons:

- Unlike the case for some of the alternatives, for the sum of variances approach, a zero variance in one of the dimensions does not hide the uncertainties in the other dimensions.
- Because it is the matrix trace of the covariance matrix, and because the trace is a characteristic that is unchanged when we perform an orthogonal change of basis, the metric does not depend on the choice of orthogonal basis used to describe the equilibrium $\emptyset$.

- As explained in Section 2.7, variance reduction factors calculated using this sum of four estimator variances can be identified with the intuitively appealing “effective species count.”
- This sum of variances is also the variance of the quantity $\sqrt{\sum_b (\theta_b - \theta_b^*)^2}$, the Euclidean distance (l^2 -norm) between the model vector parameter θ^* and its estimate $\hat{\theta}$.

2.6 Asymptotic $\hat{\theta}$ Estimator Variances

In this case we are estimating a vector. If we could ignore the restriction that the components of $\hat{\theta}$ must sum to 1, we would proceed via the Hessian, the pure and mixed second derivatives of

$$LL(\hat{\theta} | \theta^*) = \sum_D \log(\Pr[D | \hat{\theta}]) \Pr[D | \theta^*]. \quad (6)$$

with respect to the vector components θ_A , θ_T , θ_C , and θ_G . It would be the negative of the matrix inverse of the Hessian that is the asymptotic covariance matrix for the $\hat{\theta}$ estimator. We would use the sum of the pure variances, the diagonal entries of the covariance matrix, in the ratio used to determine a variance reduction factor.

However, because the components of $\hat{\theta}$ are constrained to sum to 1, there are three degrees of freedom in $LL(\hat{\theta} | \theta^*)$, rather than four. A possible choice for the degrees of freedom is given by the set of equations:

$$\theta_A = \psi_1, \quad \theta_T = \psi_2, \quad \theta_C = \psi_3, \quad \text{and} \quad \theta_G = 1 - \psi_1 - \psi_2 - \psi_3. \quad (7)$$

where each ψ_i value is nonnegative, and their sum is not more than 1. We have used this set of non-degenerate parameters, although any set of three linearly independent (not necessarily orthogonal) parameters will do. With this choice, the matrix of variances and covariances of the θ_b estimators is

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{pmatrix} \left(- \frac{\partial^2 LL(\vec{\psi} | \vec{\psi}^*)}{\partial \psi_i \partial \psi_j} \right) \bigg|_{\vec{\psi} = \vec{\psi}^*}^{-1} \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}. \quad (8)$$

The trace of this matrix, i.e., the sum of the $\hat{\theta}$ components' pure variances, can be computed as the sum of all of the off-diagonal elements of $\left(- \frac{\partial^2 LL}{\partial \psi_i \partial \psi_j} \right) \big|_{\vec{\psi} = \vec{\psi}^*}^{-1}$, plus twice the sum of its diagonal elements.

Much as with the γ estimator variance, this Hessian is evaluated numerically, at $\gamma^* = 1$ and $\hat{\theta}$ as given in Section 2.1.

2.7 Effective Species Count

When two species are closely related there has been little evolutionary time for their genomes to diverge and we expect a high correlation in their sequences. In particular, a pair of fragments of DNA, one from each species, that have descended from a common ancestor fragment are likely to be quite similar to each other, even if they are non-functional “junk” DNA. In the other extreme, two species which are very distant evolutionarily are likely to have genome sequences which are nearly statistically independent. That is, fragments of DNA in each species that share a common ancestor will have garnered so many mutations *en route* to the current-

day species that we expect the descendent species' fragments to have mostly “forgotten” their common heritage.

When sequences from several species are statistically independent, (*i.e.*, in the limit as we grow every branch length in the evolutionary tree to an infinite amount of time), the mathematics tells us that the θ variance reduction factor will be proportional to the number of sequenced species. In particular, when measured relative to a tree of just one species, in this limit the variance reduction factor exactly equals the number of sequenced species in the tree. Thus it is natural to describe the θ variance reduction factor relative to a single species as the “effective number of independent species,” “effective sample size,” or simply, “effective species count.”

This easy identification is a primary reason that we focus on the variance of the estimators rather than, *e.g.*, their standard deviations.

3 Results

3.1 Estimating the Rate of Mutation

For reasons described in the Methods Section 2.7 we describe confidence interval shrinkage for the rate of mutation parameter γ in terms of the *variance* of the estimates of γ about its true value. We say that the variance reduction factor is f when adding additional sequenced species to a phylogenetic tree yields a variance for γ estimates that is $1/f$ of the original value. In our analysis, the estimator variance to which all γ variance estimates are compared is that arising from just human and mouse. The top curve graphed in Figure 2 shows the variance reduction factors achieved if mammals are added to human and mouse in the order dog, cow, pig, rat, cat, baboon, and chimpanzee.

This order of species is the “greedy” ordering. More specifically, under the assumption that human and mouse are already sequenced, we determine that the subsequent species that will most increase the variance reduction factor for γ is dog. That is, the subtree of Figure 1 containing just human, mouse, and dog has the best variance reduction factor among all subtrees that are composed of human, mouse, and exactly one other species. Once dog has been sequenced, the most useful subsequent species is cow. Proceeding in this greedy fashion we determine the order in which to sequence the remaining mammals as shown in Figure 2.

When taken together, the aligned sequence data from these nine mammals are 3.5 times as effective for conservation identification as the data from human and mouse alone. That is, a conserved DNA fragment that is identified by examination of aligned data from all nine species is as statistically easy to identify as a similarly conserved fragment that is 3.5 times as long, but is identified using aligned human and mouse sequence data only. In other words, many fewer sequence positions are required when aligned sequence information from all nine species is available.

3.2 Estimating the Distribution of Nucleotides

With the probability distribution θ we have four variances, one for each component of θ , as well as six covariance terms. For reasons described in the Methods Section 2.5 we describe confidence interval shrinkage for the nucleotide distribution θ in terms of the single value which is the sum of the four variances. We say that the variance reduction factor for θ estimates is f when adding additional sequenced species to a phylogenetic tree yields a sum of the four estimator variances that is $1/f$ of the original value. The lower curve graphed in Figure 2 shows the variance reduction factors for θ (relative to just human and mouse) as the additional mammals are added in greedy order. The greedy ordering for γ estimation and θ estimation are the same in this case.

With θ estimation it is sensible to measure the sum of the four estimator variances even with a single species. The variance reduction factors relative to human alone are given in Figure 3.

The nine mammals are only 1.5 times as effective for pattern identification as the data from human and mouse alone, or equivalently, 2.1 times as effective for pattern identification as the data from human alone. That is, a collection of sequence positions believed to have the same equilibrium distribution of nucleotides (because, for instance, they are each believed to be the first position of a functional pattern which is repeated several times in the genome) that is characterized by examination of aligned data from all nine species is as statistically useful as a similar set of aligned sequence positions that is 1.5 times as large, but is analyzed using aligned human and mouse sequence data only. In other words, having information about all nine species provides some reduction in the number of sequence positions needed to achieve a given precision, but not as much as these authors had hoped.

4 Discussion

In searching for the patterns describing transcription factor binding sites in prokaryotes, McCue and colleagues were able to identify transcription factor binding sites and sets of co-regulated genes, or regulons, using only the genome sequences of multiple related species, (McCue et al., 2001, 2002). They achieved these genomic-scale results because they could identify patterns likely to be transcription factor binding sites, using a single gene and its orthologs in related species. The results of Figure 3 do not encourage a similar approach among the mammals. It seems that significant progress in transcription factor binding site pattern identification for mammals will require additional data that identify multiple genes, for which multiple observations of each will be necessary if we are to discern a pattern of interest. For example, data from a carefully controlled expression array study may be appropriate.

The addition of non-mammalian species would likely improve the variance reduction factor substantially, but with the tradeoff of a loss of specificity. Thus, DNA patterns associated with broadly conserved functions (such as the biosynthesis of fundamental cellular components) may be identifiable with such data sets, but the addition of non-mammal species will not aid in the identification of those patterns that are specific to specialized mammalian functions.

Contrastingly, the results for the identification of conserved regions of DNA are encouraging. With sequences from additional mammals, we will likely attain significant improvement in our ability to locate regions of conserved DNA. In particular, researchers have used conservation as a step in locating transcription factor binding sites in γ -proteobacteria (Rajewsky et al., 2002), metazoans (Lenhard et al., 2003), monkeys (Boffelli et al., 2003), and yeast (Lee et al., 2002; Kellis et al., 2003; Cliften et al., 2003) and we can expect this technique to become more useful in human studies with the availability of mammalian genomes. Unfortunately, as indicated above, it appears that subsequent computational identification of the sequence patterns of functional sites will not be as easy.

The results do not appear to be significantly dependent on our choice of nucleotide substitution model (Equation 1 in the Methods Section 2.2). For instance, the model of Hasegawa et al. (1985) allows us to specify the rate of transitions ($A \leftrightarrow G, T \leftrightarrow C$) relative to transversions (all other nucleotide substitutions). Using a transition/transversion ratio of 3.0 gives results similar to those in Figures 2 and 3; for instance, the nine mammal numbers in Figure 2 become 3.528 and 1.468, and the nine mammal number in Figure 3 becomes 2.051.

Note that with this set of species the measure of the quality of any subset of the species appears monotonically related to the total of the edge lengths in the subtree spanned by those species. In practice, choosing genomes to sequence based upon this “total of edge lengths criterion”

may prove effective in prioritizing sequencing efforts, even though it does not give us the variance reduction factor obtained from the additional sequenced genomes.

5 References

- Arvestad L, Bruno WJ. Estimation of reversible substitution matrices from multiple pairs of sequences. *J Mol Evol* 1997;45(6):696–703. [PubMed: 9419247]PubMed 9419247
- Barry D, Hartigan JA. Statistical analysis of hominoid molecular evolution. *Statistical Science* 1987;2(2):191–210.
- Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* 2003;299(5611):1391–1394. [PubMed: 12610304]PubMed 12610304
- Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* 2003;301(5629):71–76. [PubMed: 12775844]PubMed 12775844
- Durbin, R.; Eddy, S.; Krogh, A.; Mitchison, GJ. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press; Cambridge, United Kingdom: 1998. Building phylogenetic trees; p. 160–232.
- Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 1981;17(6):368–376. [PubMed: 7288891]PubMed 7288891
- Felsenstein J, Churchill GA. A hidden Markov model approach to variation among sites in rate of evolution. *Mol Biol Evol* 1996;13(1):93–104. [PubMed: 8583911]PubMed 8583911
- Friedman N, Ninio M, Pe'er I, Pupko T. A structural EM algorithm for phylogenetic inference. *J Comput Biol* 2002;9(2):331–353. [PubMed: 12015885]PubMed 12015885
- Gu X, Fu Y-X, Li W-H. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol Biol Evol* 1995;12(4):546–557. [PubMed: 7659011]PubMed 7659011
- Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 1985;22(2):160–174. [PubMed: 3934395]PubMed 3934395
- Holmes I, Rubin GM. An expectation maximization algorithm for training hidden substitution models. *J Mol Biol* 2002;317(5):753–764. [PubMed: 11955022]PubMed 11955022
- Jukes, TH.; Cantor, C. Evolution of protein molecules. In: Munro, HM., editor. *Mammalian Protein Metabolism*. 3. Academic Press; New York, NY: 1969. p. 21–132.
- Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 2003;423(6937):241–254. [PubMed: 12748633]PubMed 12748633
- Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 1980;16(2):111–120. [PubMed: 7463489]PubMed 7463489
- Kishino H, Hasegawa M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol* 1989;29(2):170–179. [PubMed: 2509717]PubMed 2509717
- Lanave C, Preparata G, Saccone C, Serio G. A new method for calculating evolutionary substitution rates. *J Mol Evol* 1984;20(1):86–93. [PubMed: 6429346]PubMed 6429346
- Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne J-B, Volkert TL, Fraenkel E, Gifford DK, Young RA. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 2002;298(5594):799–804. [PubMed: 12399584]PubMed 12399584
- Lenhard B, Sandelin A, Mendoza L, Engström P, Jareborg N, Wasserman WW. Identification of conserved regulatory elements by comparative genome analysis. *J Biol* 2003;2(2):13. [PubMed: 12760745]PubMed 12760745
- McCue LA, Thompson W, Carmack CS, Lawrence CE. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res* 2002;12(10):1523–1532. [PubMed: 12368244]PubMed 12368244

- McCue LA, Thompson W, Carmack CS, Ryan MP, Liu JS, Derbyshire V, Lawrence CE. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res* 2001;29(3):774–782. [PubMed: 11160901]PubMed 11160901
- Neyman, J. Molecular studies of evolution: a source of novel statistical problems. In: Gupta, SS.; Yackel, J., editors. *Statistical Decision Theory and Related Topics*. Academic Press; New York, NY: 1971. p. 1-27.
- Rajewsky N, Socci ND, Zapotocky M, Siggia ED. The evolution of DNA regulatory regions for proteogamma bacteria by interspecies comparisons. *Genome Res* 2002;12(2):298–308. [PubMed: 11827949]PubMed 11827949
- Rodríguez F, Oliver JL, Marín A, Medina JR. The general stochastic model of nucleotide substitution. *J Theor Biol* 1990;142(4):485–501. [PubMed: 2338834]PubMed 2338834
- Siepel A, Haussler D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol* 2004;21(3):468–488. [PubMed: 14660683]PubMed 14660683
- Silvey, SD. *Optimal Design: An Introduction to the Theory for Parameter Estimation*. Chapman and Hall; 1980.
- Stuart, A.; Ord, JK.; Arnold, S. *Classical Inference and the Linear Model* vol. 2A, of Kendall's *Advanced Theory of Statistics*. sixth. Arnold; London: 1999. p. 17.13-17.27.
- Swofford, DL.; Olsen, GJ.; Waddell, PJ.; Hillis, DM. Phylogenetic inference. In: Hillis, DM.; Moritz, C.; Mable, BK., editors. *Molecular Systematics*. second. Sinauer Associates, Inc; Sunderland, MA: 1996. p. 407-514.
- Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 1993;10(3):512–526. [PubMed: 8336541]PubMed 8336541
- Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, McDowell JC, Maskeri B, Hansen NF, Schwartz MS, Weber RJ, Kent WJ, Karolchik D, Bruen TC, Bevan R, Cutler DJ, Schwartz S, Elnitski L, Idol JR, Prasad AB, Lee-Lin SQ, Maduro VV, Summers TJ, Portnoy ME, Dietrich NL, Akhter N, Ayele K, Benjamin B, Cariaga K, Brinkley CP, Brooks SY, Granite S, Guan X, Gupta J, Haghighi P, Ho SL, Huang MC, Karlins E, Laric PL, Legaspi R, Lim MJ, Maduro QL, Masiello CA, Mastrian SD, McCloskey JC, Pearson R, Stantripop S, Tiongson EE, Tran JT, Tsurgeon C, Vogt JL, Walker MA, Wetherby KD, Wiggins LS, Young AC, Zhang LH, Osoegawa K, Zhu B, Zhao B, Shu CL, De Jong PJ, Lawrence CE, Smit AF, Chakravarti A, Haussler D, Green P, Miller W, Green ED. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 2003;424(6950):788–793. [PubMed: 12917688]PubMed 12917688
- Waddell PJ, Steel MA. General time-reversible distances with unequal rates across sites: mixing γ and inverse gaussian distributions with invariant sites. *Mol Phylogenet Evol* 1997;8(3):398–414. [PubMed: 9417897]PubMed 9417897
- Yang Z. Estimating the pattern of nucleotide substitution. *J Mol Evol* 1994a;39(1):105–111. [PubMed: 8064867]PubMed 8064867
- Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* 1994b;39(3):306–314. [PubMed: 7932792]PubMed 7932792
- Yang Z. A space-time process model for the evolution of DNA sequences. *Genetics* 1995;139(2):993–1005. [PubMed: 7713447]PubMed 7713447

Figure 1.
Phylogenetic Tree of Mammals, which is Figure 4a of Siepel & Haussler (2004), showing the expected number of mutations per non-coding sequence position, for each tree edge.

Figure 2.
Comparison of Variance Reduction Factors in the identification of patterns in DNA sequence vs. the identification of DNA conservation. The mammals are added to human and mouse in the optimal greedy order: dog, cow, pig, rat, cat, baboon, and chimpanzee. Both variances are normalized, so that the variance of the human and mouse pair is equal to 1.

Figure 3.
Identification of Patterns in DNA Sequence: the variance reduction factor relative to human alone (a.k.a. effective sample size; see Section 2.7) starting with human and mouse, and greedily adding the seven remaining mammals in the order: dog, cow, pig, rat, cat, baboon, and chimpanzee.