Comparative assessment of methods for aligning multiple genome sequences

Xiaoyu Chen & Martin Tompa

Multiple sequence alignment is a difficult computational problem. There have been compelling pleas for methods to assess whole-genome multiple sequence alignments and compare the alignments produced by different tools. We assess the four ENCODE alignments, each of which aligns 28 vertebrates on 554 Mbp of total input sequence. We measure the level of agreement among the alignments and compare their coverage and accuracy. We find a disturbing lack of agreement among the alignments not only in species distant from human, but even in mouse, a well-studied model organism. Overall, the assessment shows that Pecan produces the most accurate or nearly most accurate alignment in all species and genomic location categories, while still providing coverage comparable to or better than that of the other alignments in the placental mammals. Our assessment reveals that constructing accurate whole-genome multiple sequence alignments remains a significant challenge, particularly for noncoding regions and distantly related species.

With the rapid sequencing of many related genomes, comparative sequence analysis has emerged as one of the most important areas of computational biology. The fundamental tool of comparative sequence analysis is multiple sequence alignment.

As an example of alignments that are intended for comparative sequence analysis, consider the whole-genome multiple sequence alignments of the UCSC Genome Browser[1]. Sophisticated analyses rely implicitly on the correctness of such alignments. For instance, it is standard practice to search for regulatory elements by scanning the regulatory regions of such whole-genome alignments to identify short windows that are well conserved across the species[2,3]. Similar conservation-based applications include gene prediction[4,5], noncoding RNA prediction[6,7] and, more generally, predicting genomic elements that are under purifying selection[8–13]. In regions where the sequences are misaligned, these methods may fail to find conserved sites that exist.

Downstream applications of genomic multiple sequence alignments are not limited to identifying regions under purifying constraint. Other important applications include inference of phylogeny[14,15], estimates of substitution rates[15,16], understanding of evolutionary mechanisms[17,18] and identification of regions under positive selection[11,19–23]. Because misaligned sequences could easily produce false signals of evolutionary change, these downstream applications are at greater risk of a loss of accuracy when sequences are misaligned[15].

The many existing multiple-alignment tools often produce quite different alignments when applied to the same set of input sequences[10,15,24], leading users to wonder which alignment, if any, is 'right'. Because of this, a number of recent reviews and articles[10,25–30] have made compelling pleas for methods to assess the accuracy of genomic multiple sequence alignments and to compare the alignments produced by different tools. We address this issue here.

Recently, the ENCODE Multi-Species Sequence Analysis team used four different pipelines to align 1% of the human genome to 27 other vertebrate genomes[10]. The four alignment tools are TBA[31], MAVID[32], MLAGAN[33] and Pecan[34]. The four ENCODE alignments provide a rich resource for comparison of whole-genome alignment tools. What makes these alignments an unprecedented test bed for comparison is that four expert teams used four different alignment methods to align the same 28 vertebrate sequences, spanning 554 Mbp of sequence in total. What makes such a comparison a challenge is the number of dimensions to be taken into account: how much agreement is there among the alignments? Which method is most accurate in aligning distantly related species? How do the methods compare in accuracy in coding and noncoding regions? Which methods align more input sequence than the others? When one method aligns more input sequence than the others, how accurate are these additional aligned regions?

Margulies *et al.*[10] performed the first comparative analyses of these four alignments. They compared estimates of sensitivity, which is the fraction of orthologous residues that are correctly aligned (using as proxies coverage of human coding sequences and ancestral repeats) and estimates of specificity, which is the fraction of aligned residues that are truly orthologous (using as proxies coding sequence periodicity and nonalignment to human *Alu* sequences). Our comparative assessment is more comprehensive than the initial assessment of Margulies *et al.* They estimated the alignment coverage and accuracy by extrapolating from coding regions and repeats. In contrast, we compare alignment coverage and accuracy at all sites, broken down by location into four categories: coding, UTR, intronic and intergenic. Margulies *et al.* restricted their analyses to mammalian alignment, omitting chicken, *Xenopus*, tetraodon, fugu and zebrafish. We include all aligned vertebrate species, and discover that some of the most dramatic differences occur in these distant species.

Our analyses are divided into three types.

1. We measure precisely the agreement and disagreement among the alignments. The purpose of this analysis is to establish that differences among the alignments are substantial; it is not

Department of Computer Science and Engineering, Department of Genome Sciences, University of Washington, Seattle, Washington, USA. Correspondence should be addressed to M.T. (tompa@cs.washington.edu).

Received 21 December 2009; accepted 27 April 2010; published online 23 May 2010; doi:10.1038/nbt.1637

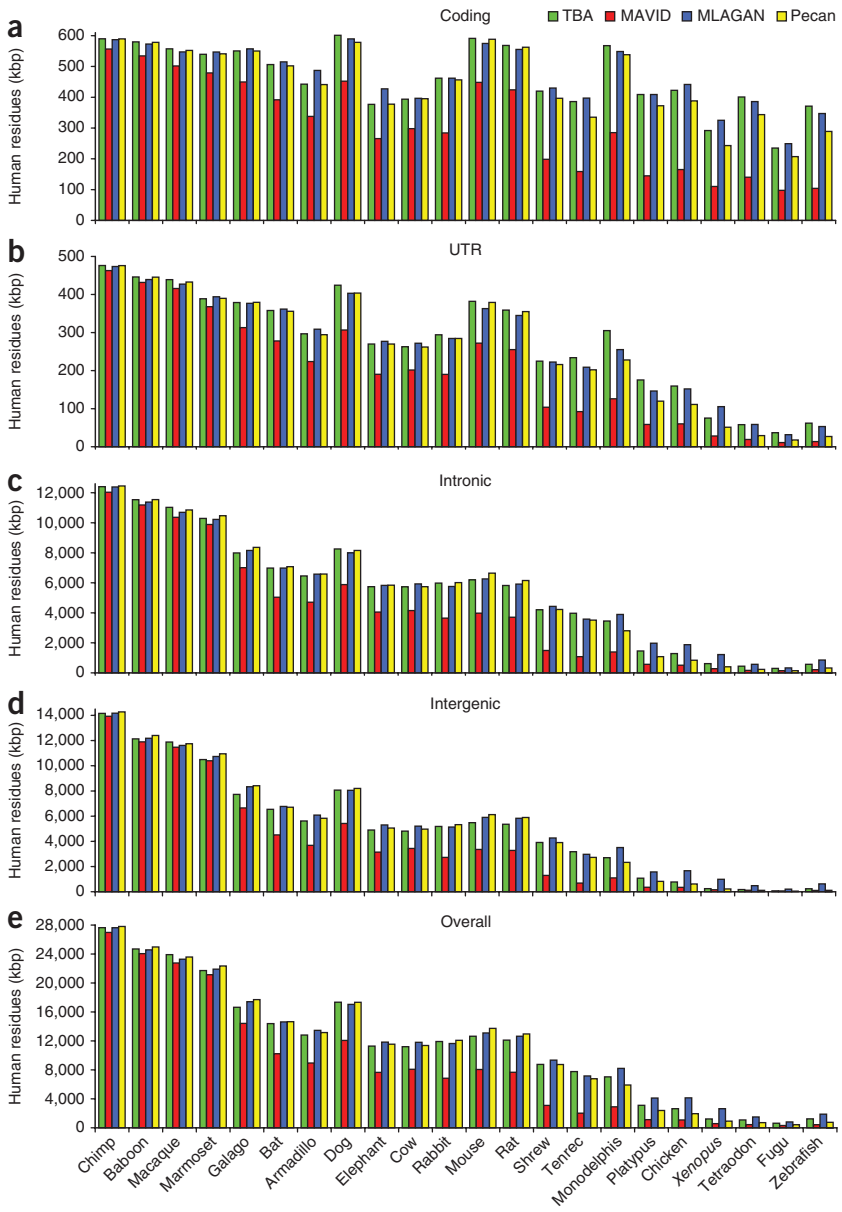

Figure 1 Comparison of coverage of the alignments. The comparison is broken down by species and by location category; also provided is an overall chart that aggregates all four location categories. Species are displayed on the horizontal axis in order of increasing total branch length from human, according to a phylogeny estimated from fourfold-degenerate sites of third codon positions in the ENCODE regions[10]. The vertical axis represents the number of human residues aligned to each species given on the horizontal axis, in units of kilobase pairs (Kbp). Note that the vertical scales are different in each of the charts. The figure shows that TBA, MLAGAN and Pecan all have comparable coverage in all the placental mammals (chimp through tenrec) across all location categories. For all alignments, note that the coverage decreases approximately as species distance from human increases, particularly in the noncoding location categories.

intended to expose relative advantages and disadvantages of the alignments.

2. We compare the coverage of each alignment, which is the number of human residues aligned to each species.
3. We compare the accuracy of each alignment. To estimate accuracy, we use a statistical method called StatSigMA-w[35], which identifies suspiciously aligned portions of each.

Each analysis is broken down by species and by location with respect to annotated human genes. This provides the most comprehensive comparison of large-scale alignments to date and suggests a methodology for future comparisons. Finally, we exploit the availability of alternative alignments by demonstrating how often the alignment of a region identified as suspicious can be improved by some alternative alignment.

We use StatSigMA-w[35] to measure the accuracy of genome-size alignments. In the past, two other approaches were used to measure accuracy. The first uses sequences constructed by simulating evolution[31,34]. The strength of this is that the correct alignment is known, so that alignment sensitivity and specificity can be measured accurately. The drawback is its sensitivity to assumptions in the simulation about underlying evolutionary processes, particularly genomic rearrangements, that are not well understood. The second approach measures the accuracy with which known homologous features are aligned. For example, known orthologous exons are often used[10,32,33,36], as are known repeat families[10,34]. Such features represent a narrow spectrum of the genome, and evaluations based on them may not extrapolate well to other genomic regions. In particular, the use of orthologous coding exons has the drawback that they are usually well conserved and most tools align them accurately. In contrast, StatSigMA-w allows direct evaluation of the accuracy at all aligned sites, rather than being limited to a small number of genomic features.

There are many alignment scoring functions that measure conservation and cannot serve as measures of alignment accuracy, including sum of pairs, percent identity, entropy, binCons[9], phastCons[13], GERP[8], Gumby[12] and phyloP[11]. In a perfectly accurate alignment, where the measure of alignment accuracy should be high throughout, conservation scores will be high in regions under purifying selection and low in regions evolving neutrally or under positive selection. Conversely, in an alignment that is not perfectly accurate, there can be regions that have high conservation across nearly all sequences, with the remaining sequences misaligned (Fig. 2 and Table 1 of ref. 35). In such regions, the alignment accuracy will be low, but conservation scores will be high. These facts together suggest that any conservation score is a poor measure of alignment accuracy.

RESULTS

Alignment coverage

Given alignment *A* and nonhuman species *S*, *A*'s "coverage" by *S* is the number of human residues aligned by *A* to a residue or gap in *S* (after removing gaps longer than 20 bp; see Online Methods). **Figure 1** compares the alignment coverage for all species in the four location categories. (See also **Supplementary Coverage Spreadsheet**.)

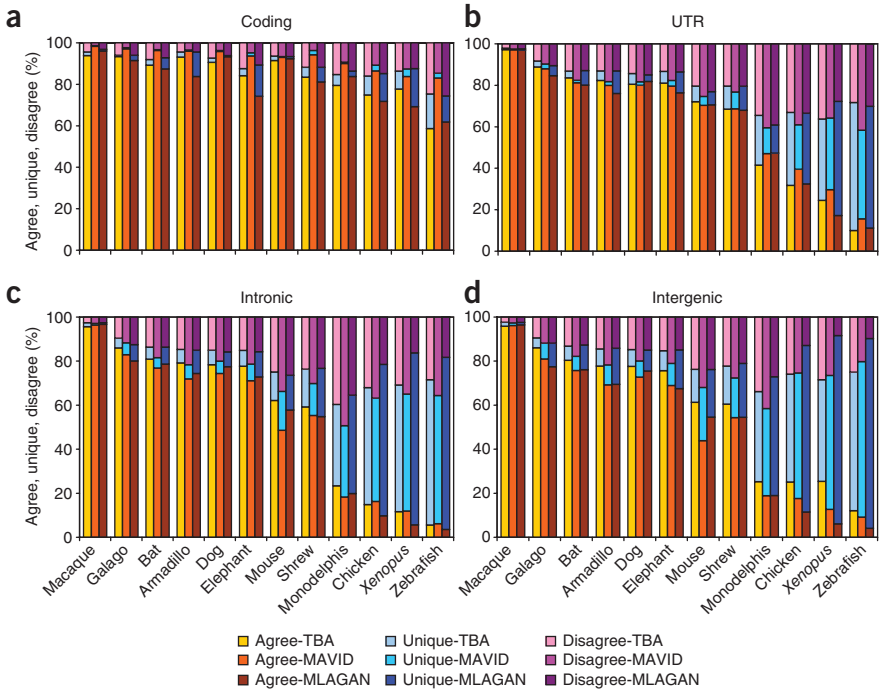

Figure 2 Comparison percentages agree%, unique% and disagree% for TBA, MAVID and MLAGAN. (See Online Methods for the explanation of why Pecan is excluded.) The comparison is shown for 12 representative species and broken down by location category. Species are displayed on the horizontal axis in order of increasing total branch length from human. Note the trend that agree% decreases and unique% increases as the species distance from human increases and also as one moves from coding to UTR to intronic/intergenic categories.

For all alignments, the coverage decreases approximately as species distance from human increases, particularly in noncoding location categories. Minor exceptions are seen for dog, mouse, rat and monodelphis. For mouse, rat and monodelphis, the explanation may be that more sequence was available to the aligners than for any other nonprimate[10].

MAVID consistently has the lowest coverage in nearly all species and location categories. For distant species, MAVID often has only half the coverage of other alignments, even in coding regions. The other alignments have comparable coverage in all placental mammals (chimp through tenrec) and location categories. These observations are consistent with earlier findings[10]. In the intronic and intergenic regions of more distant species, MLAGAN has the highest coverage, followed in order by TBA, Pecan and MAVID. The most extreme case occurs in *Xenopus* intergenic regions, where MLAGAN has over four times the coverage of any other.

Level of agreement among alignments

Next we investigate the level of agreement among alignments. For each alignment *A* and nonhuman species *S*, consider the following measures (defined precisely in Online Methods):

1. "agree%," the percentage of human residues aligned by *A* to *S* that are aligned to the same coordinate in *S* by some other alignment.
2. "unique%," the percentage of human residues aligned by *A* to *S* that are not aligned to *S* by any other alignment.
3. "disagree%" = 100% – agree% – unique%. This is the percentage of human residues aligned by *A* to *S* that are aligned to *S* differently by some other alignment and not aligned to the same coordinate in *S* by any other alignment.

Note that these are percentages of *coverage*, defined in the previous section.

Figure 2 illustrates these comparison percentages for three ENCODE alignments. (See also **Supplementary Comparison Percentage Spreadsheet**.) The first observation is that there are no major differences in comparison percentages among the alignments. MLAGAN has somewhat greater unique% in the intronic and intergenic regions of nonmammals, consistent with its higher coverage in these regions.

There are clear trends relating the location categories. Firstly, the intronic and intergenic categories have similar comparison percentages. If the species is kept fixed, agree% decreases and unique% increases as one moves from the coding to UTR to intronic and intergenic categories, reflecting the increased difficulty of aligning noncoding regions.

Fixing next the location category, there are clear trends relating the species. In all noncoding categories, as the species distance from human increases, agree% decreases and unique% increases, reflecting increased difficulty of aligning more diverged sequences[25,31]. Compared to placental mammals, the more distant species have sharply decreased agree% and increased unique% in noncoding location categories. Most nonmammals have agree% < disagree% < unique% in intronic and intergenic regions, demonstrating little agreement among alignments.

Because mouse is an important model organism, and because human-mouse alignments are widely used in research, the level of agreement for mouse is of particular interest. Intronic and intergenic regions account for 95% of the human sites aligned to mouse (**Fig. 1**). In these categories combined, agree% for mouse is disturbingly low, ranging from MAVID's 46% to TBA's 62% (**Fig. 2**). The situation is even worse in the distant species, which have much lower agree% values. Such low levels of agreement indicate that constructing a reliable whole-genome multiple sequence alignment remains a significant challenge, particularly for noncoding regions and distantly related species.

Alignment accuracy

Wherever alignments do not agree, which alignment, if any, is correct? This is difficult to assess because the true alignment (the one that aligns all and only orthologous residues) is inherently unknown.

We use StatSigMA-w[35] to estimate alignment accuracy. Given an alignment *A* and a nonhuman species *S*, StatSigMA-w identifies "suspiciously aligned regions for *S*," which have at least 50 columns and statistical evidence that S is no better aligned in this region than a random sequence (see Online Methods for details). The percentage of aligned sites of species *S* that fall in suspicious regions for *S* is denoted "suspicious%." **Figure 3** compares suspicious% values of the four alignments for all species. (See also **Supplementary Suspicious Percentage Spreadsheet**.)

When we first compare the four alignment methods for fixed species and fixed location category, MLAGAN has the highest (or nearly highest) suspicious% and Pecan the lowest (or nearly lowest) suspicious%

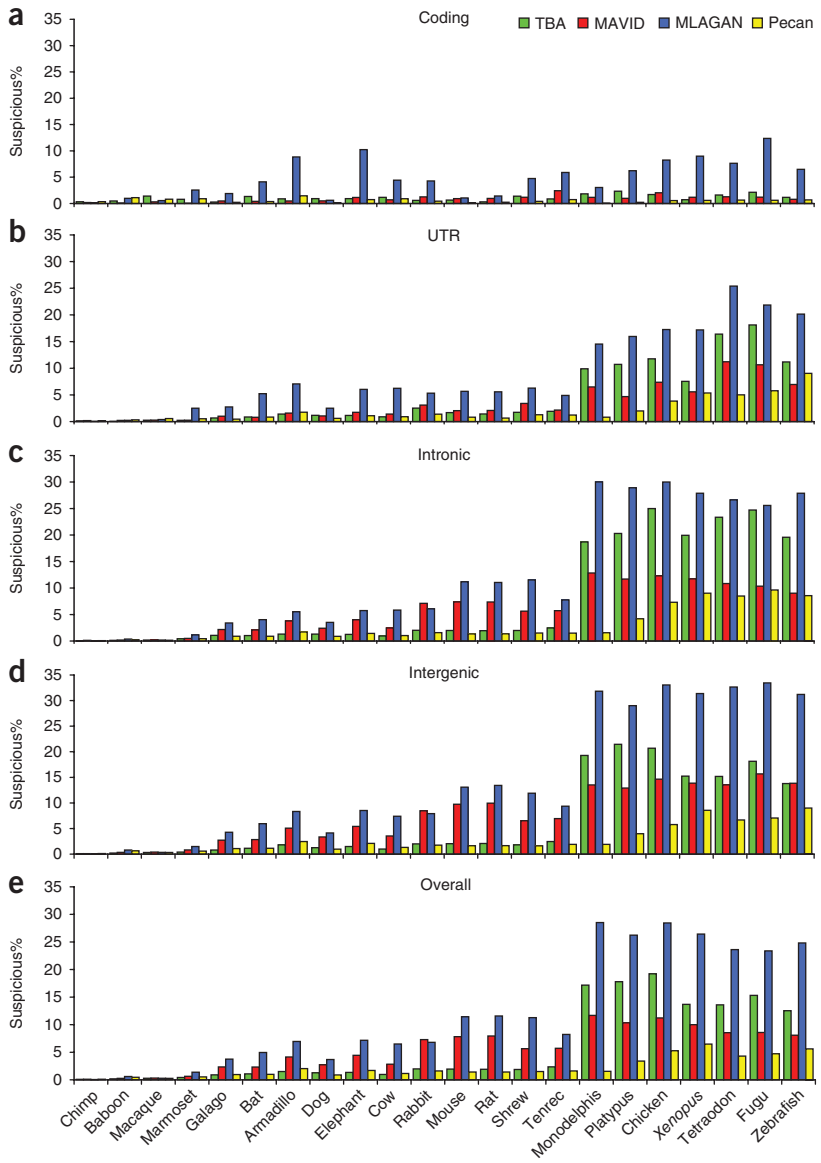

Figure 3 Comparison of accuracy of the alignments, as measured by suspicious%. The comparison is broken down by species and by location category, plus an overall chart that aggregates all four location categories. Species are displayed on the horizontal axis in order of increasing total branch length from human. For each alignment and each noncoding category, suspicious% generally increases as species distance from human increases, with a noticeable jump between the placental mammals and more distant species. Note that Pecan has the lowest or near lowest suspicious% for every species and location category.

for every species and location category. For Pecan, suspicious% is <10% in every species and category, whereas MAVID's is as large as 16% (fugu intergenic), TBA's as large as 25% (chicken intronic) and MLAGAN's as large as 33% (fugu intergenic). TBA's and Pecan's suspicious% values are comparable for all placental mammals and location categories: both have suspicious% ≤2.5% in all such categories. However, TBA's suspicious% rises precipitously in noncoding regions of more distant species, up to 14–25% depending on the species.

Turning to the trends in noncoding regions as the species varies, with alignment and location category fixed, suspicious% generally increases as species distance from human increases. There is a jump in suspicious% when one moves from placental mammals to more distant species, which is particularly noticeable in TBA and MLAGAN. These trends again reflect the increased difficulty of correctly aligning distant species.

We turn finally to the trends as location category varies. As in comparison percentages, there is little difference in suspicious% values between intronic and intergenic categories. Generally, suspicious% increases as one moves from coding to UTR to intronic and intergenic categories, reflecting increased difficulty of aligning noncoding regions correctly. In coding regions, MLAGAN has greater suspicious% than the other alignments, sometimes exceeding 10%. Each of the other three has suspicious% <2.5% in the coding regions of every species.

Our accuracy comparison disagrees sharply with that of Margulies *et al.*[10] on the nonplacental mammals monodelphis and platypus. As **Figure 3** illustrates, suspicious% increases in these species in the order Pecan, MAVID, TBA, MLAGAN. In intronic and intergenic regions, TBA's suspicious% is 5 times that of Pecan in platypus and 10–12 times that of Pecan in monodelphis (**Fig. 3**). In contrast, in terms of *Alu* exclusion for these two species, Margulies *et al.*[10] showed that TBA is best, with Pecan and MAVID close behind. In their analysis, unlike ours, monodelphis and platypus do not show patterns of alignment accuracy significantly different from those of cow, dog, armadillo, elephant, tenrec, shrew, bat and rabbit. See **Supplementary Text Section 1** for further discussion.

Taken together, the results of this section suggest that the accuracy of all alignments decreases in more distantly related species and in noncoding regions. Pecan appears to be most accurate overall.

Improving suspicious alignments

The ENCODE alignments provide an interesting test bed for determining whether suspiciously aligned regions can be improved, also adding evidence supporting StatSigMA-w's predictions of misalignment. (Evidence given in previous work[35,37] includes poor protein BLAST *E*-values in suspicious coding regions and results on simulated data. Additional evidence in **Supplementary Figs. 1** and **2** shows that suspicious regions are highly depleted in alignment-agreeing coordinates and enriched in alignment-unique coordinates.)

As a first step towards improving suspiciously aligned regions, we plotted pairwise alignment scores of suspicious alignments versus alternative alignments. **Figure 4** shows scatter plots of pairwise alignment scores for suspicious MLAGAN alignments versus nonsuspicious alternative alignments of the same human region (details in Online Methods). Scatter plots with each of the other alignments replacing MLAGAN have similar patterns (**Supplementary Fig. 3**). In the baboon plot, nearly every point lies above $y = x$, suggesting that suspiciously aligned baboon regions can be improved by an alternative alignment. In the mouse and zebrafish plots, the majority of points lie above $y = x$, suggesting that most suspiciously aligned regions can

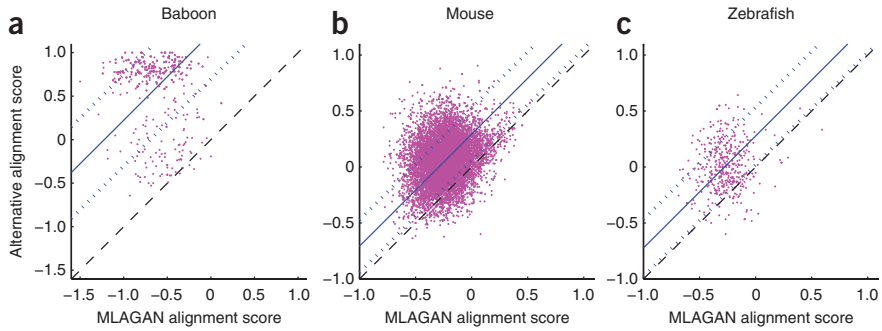

Figure 4 Pairwise alignment scores of suspicious regions versus those for alternative alignments of the same human region. For three representative species *S* (baboon, mouse and zebrafish) and one representative target alignment (MLAGAN), scatter plots show all points (x', y'), where x' is the pairwise human-*S* alignment score of an MLAGAN alignment region that is suspicious for species *S*, and y' is the pairwise human-*S* alignment score of one of the other three alignments for the same human region that is not suspicious for *S*. (See Online Methods for the scoring function and **Supplementary Fig. 3** for other target alignments.) Alignment scores are normalized by alignment length. The dashed black diagonal line has equation $y = x$. The solid blue line has equation $y - x = \mu$, where μ is the mean value of $y' - x'$ for all points (x', y') in the plot. The dotted blue lines have equations $y - x = \mu \pm \acute{o}$, where $\acute{o}$ is the standard deviation of $y' - x'$ for all points (x', y') in the plot. Note that most points lie above the line $y = x$, suggesting that most of the suspiciously aligned regions can be improved by one of the alternative alignments.

be improved by an alternative alignment. Points lying below this line can be explained by two possibilities: (i) StatSigMA-w's prediction of misalignment is incorrect, or (ii) StatSigMA-w's prediction of misalignment is correct, but the alternative alignment is no better, either because the human sequence has no ortholog in the target species or because it is difficult to identify and align the correct ortholog. The fact that nearly all points in the baboon plot lie above the diagonal supports the latter explanation, because human sequences are more likely to have orthologs (that are not difficult to align) in baboon than in mouse or zebrafish.

With either explanation for the points below $y = x$, it is natural to ask why StatSigMA-w does not identify the alternative alignment as suspicious as well. The explanation is that StatSigMA-w makes conservative calls of suspicious regions (details in Online Methods), which suggests that there are other misalignments besides the regions StatSigMA-w labels suspicious.

Taken together, the results of this section suggest that most suspiciously aligned regions can be improved by an alternative alignment method.

DISCUSSION

For four multi-vertebrate alignments of the 30-Mbp human ENCODE regions, we performed three comprehensive analyses: we measured the level of agreement among alignments, we compared their coverage, and we compared their accuracy.

In the first of these analyses, we found a surprisingly low level of agreement among the alignments of human noncoding regions to nonplacental mammals and more distant species. Even for mouse, an important model organism, only about half the sites aligned in one alignment agree with some other alignment. This suggests caution for users of whole-genome alignments. (Even though Pecan could not be included in this first analysis due to missing information, this analysis was not intended to compare the quality of the alignments, which is determined by the comparisons of coverage and accuracy below. The intent of this first analysis is rather to appreciate the lack of agreement that exists among alignment methods.)

In a comparative assessment, the goal is to learn which method is best. To answer this, alignment coverage and accuracy must be considered together. Because we have used the suspicious% measure of StatSigMA-w[35] to estimate alignment inaccuracy, the ideal alignment is one with high coverage (**Fig. 1**) and low suspicious% (**Fig. 3**). **Figure 5** summarizes suspicious% versus coverage, but only for the aggregation of all location categories. MAVID has the lowest coverage for nearly all species and location categories. The other alignments have comparable coverages in all species and location categories (with one exception, discussed below). Pecan has the lowest or nearly lowest suspicious% for all species and location categories, less than 10% in each of these 22 × 4 categories. Taken together, these results suggest that the Pecan alignment is the best among the four ENCODE alignments. Given the number of dimensions for alignment comparison, it is surprising that any one alignment appears best in nearly every category.

The exception to comparable coverage for TBA, MLAGAN and Pecan occurs in intronic and intergenic categories of nonplacental species (monodelphis, platypus, chicken, *Xenopus*, tetraodon, fugu, and zebrafish), where MLAGAN has the highest coverage, TBA the next highest and Pecan the lowest. Among these seven species and two location categories, MLAGAN has up to 4 times the coverage of TBA (averaging 1.9 times the coverage over all 14 categories), and TBA has up to 2 times the coverage of Pecan (averaging 1.5 times the coverage over all 14 categories). However, in each of these 14 categories, MLAGAN's suspicious% is greater than TBA's, which is greater than Pecan's, and in most categories these differences are great. This suggests that the additional coverage in these categories may not be worth the decreased accuracy. For example, averaged over these 14 categories, TBA's suspicious% is 4 times Pecan's, whereas TBA's coverage is only 1.5 times Pecan's.

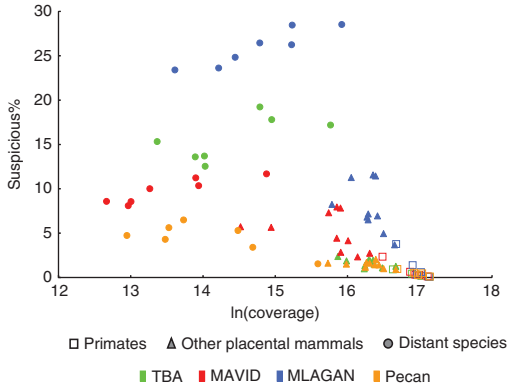

Figure 5 Summary plot of suspicious% vs. coverage, aggregated over all four location categories. The horizontal axis is on a logarithmic scale. For a given species, points that are lower and farther right represent better performance. Note the comparable performance of Pecan and TBA on placental mammals (orange and green triangles) and the superior accuracy of Pecan on the distant species (orange circles).

For placental mammals, TBA's coverage and suspicious% are comparable to Pecan's in every species and location category. In this realm, TBA and Pecan emerge together as best.

Focusing finally on coding regions, all alignments except MLAGAN seem very accurate, with suspicious% <2.5% for every species. MAVID's lower coverage suggests that TBA and Pecan are best in coding regions.

TBA's overall suspicious% values in **Figure 3** are consistent with those reported in ref. 35 for the 17-vertebrate MULTIZ alignment of human chromosome 1. In particular, both demonstrate the same precipitous rise in suspicious% as one moves from placental mammals to more distant species. For the 14 nonprimates present in both alignments, the overall TBA ENCODE suspicious% values range from 0.4 to 1.16 times those of the MULTIZ whole-chromosome alignment, depending on the species, with an average ratio over all 14 species of 0.7. One reason why ENCODE suspicious% values may be less than those of the MULTIZ whole-chromosome alignment is that each ENCODE region is so much shorter than chromosome 1, and the orthologous sequences for each individual ENCODE region were prepared and supplied to the aligners.

In conclusion, we provide the most comprehensive comparison of large-scale alignment methods to date, and we propose a methodology for future comparisons of whole-genome multiple alignments. These comparisons provide critical accuracy feedback to alignment tool designers. Our assessment reveals that constructing accurate whole-genome multiple alignments remains challenging, particularly for noncoding regions and distant species. Users should exercise caution when assuming alignment correctness in these situations.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank P. Green, W. Noble, W.L. Ruzzo and especially A. Prakash for helpful discussions and technical advice. We thank the US National Institutes of Health and the Natural Sciences and Engineering Research Council of Canada for financial support.

AUTHOR CONTRIBUTIONS

X.C., design, implementation, experimentation, analysis; M.T., design, analysis.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Kent, W. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
2. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**, e7 (2005).
3. Xie, X. *et al.* Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl. Acad. Sci. USA* **104**, 7145–7150 (2007).
4. Gross, S.S. & Brent, M.R. Using multiple alignments to improve gene prediction. *J. Comput. Biol.* **13**, 379–393 (2006).
5. Siepel, A. *et al.* Targeted discovery of novel human exons by comparative genomics. *Genome Res.* **17**, 1763–1773 (2007).
6. Pedersen, J.S. *et al.* Identification and classification of conserved RNA secondary structures in the human genome. *PLOS Comput. Biol.* **2**, e33 (2006).
7. Washietl, S., Hofacker, I.L., Lukasser, M., Hüttenhofer, A. & Stadler, P.F. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.* **23**, 1383–1390 (2005).
8. Cooper, G.M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).
9. Margulies, E. *et al.* Identification and characterization of multi-species conserved sequences. *Genome Res.* **13**, 2507–2518 (2003).
10. Margulies, E.H. *et al.* Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res.* **17**, 760–774 (2007).
11. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
12. Prabhakar, S. *et al.* Close sequence comparisons are sufficient to identify human cis-regulatory elements. *Genome Res.* **16**, 855–863 (2006).
13. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
14. Felsenstein, J. *Inferring Phylogenies* (Sinauer Associates, 2004).
15. Wong, K.M., Suchard, M.A. & Huelsenbeck, J.P. Alignment uncertainty and genomic analysis. *Science* **319**, 473–476 (2008).
16. Siepel, A. & Haussler, D. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* **21**, 468–488 (2004).
17. Murphy, W.J. *et al.* Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**, 2348–2351 (2001).
18. Nikolaev, S. *et al.* Early history of mammals is elucidated with the ENCODE multiple species sequencing data. *PLoS Genet.* **3**, e2 (2007).
19. Bird, C.P. *et al.* Fast-evolving noncoding sequences in the human genome. *Genome Biol.* **8**, R118 (2007).
20. Kim, S. & Pritchard, J. Adaptive evolution of conserved non-coding elements in mammals. *PLoS Genet.* **3**, e147 (2007).
21. Nielsen, R. *et al.* A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, e170 (2005).
22. Pollard, K.S. *et al.* An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**, 167–172 (2006).
23. Prabhakar, S., Noonan, J.P., Pääbo, S. & Rubin, E.M. Accelerated evolution of conserved noncoding sequences in humans. *Science* **314**, 786 (2006).
24. Dewey, C.N., Huggins, P.M., Woods, K., Sturmfels, B. & Pachter, L. Parametric alignment of *Drosophila* genomes. *PLOS Comput. Biol.* **2**, e73 (2006).
25. Blanchette, M. Computation and analysis of genomic multi-sequence alignments. *Annu. Rev. Genomics Hum. Genet.* **8**, 193–213 (2007).
26. Kumar, S. & Filipski, A. Multiple sequence alignment: in pursuit of homologous DNA positions. *Genome Res.* **17**, 127–135 (2007).
27. Lunter, G. *et al.* Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res.* **18**, 298–309 (2008).
28. Margulies, E.H. Confidence in comparative genomics. *Genome Res.* **18**, 199–200 (2008).
29. Margulies, E.H. & Birney, E. Approaches to comparative sequence analysis: towards a functional view of vertebrate genomes. *Nat. Rev. Genet.* **9**, 303–313 (2008).
30. Rokas, A. Lining up to avoid bias. *Science* **319**, 416–417 (2008).
31. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).
32. Bray, N. & Pachter, L. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res.* **14**, 693–699 (2004).
33. Brudno, M. *et al.* LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**, 721–731 (2003).
34. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* **18**, 1814–1828 (2008).
35. Prakash, A. & Tompa, M. Measuring the accuracy of genome-size multiple alignments. *Genome Biol.* **8**, R124 (2007).
36. Dubchak, I., Poliakov, A., Kislyuk, A. & Brudno, M. Multiple whole-genome alignments without a reference organism. *Genome Res.* **19**, 682–689 (2009).
37. Prakash, A. & Tompa, M. Assessing the discordance of multiple sequence alignments. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* **6**, 542–551 (2009).

ONLINE METHODS

Alignments. Many computational tools are available for computing multiple sequence alignments. The reason so many diverse tools exist is that computing the optimal multiple sequence alignment is inherently an intractable computational problem[38]. This is true even for relatively short alignments consisting of hundreds or thousands of columns, such as alignments of proteins or genomic promoter regions. The problem becomes even harder when one wants to compute a whole-genome multiple sequence alignment, which may consist of millions or billions of alignment columns and, in addition, must contend with the complication of arbitrary genome rearrangements such as translocations, duplications, inversions and so on[29].

Each of the four whole-genome alignment programs studied here is actually integrated with other programs to form a pipeline for building the alignments[10]. For convenience, we use the names TBA, MAVID, MLAGAN and Pecan throughout to represent their respective pipelines.

Although the ENCODE alignments contain sequence from 28 vertebrate genomes, we omitted from all results the alignments of five species, colobus monkey, dusky titi, owl monkey, mouse lemur and hedgehog, because for each of them less than 3.5 Mbp of sequence was available for alignment to the 30-Mbp human ENCODE sequence. For other mammals at least 17 Mbp of sequence was available[10]. This left six primates (human, chimpanzee, baboon, macaque, marmoset and galago), ten other placental mammals (bat, armadillo, dog, elephant, cow, rabbit, mouse, rat, shrew and tenrec), two nonplacental mammals (monodelphis and platypus) and five nonmammals (chicken, *Xenopus*, tetraodon, fugu and zebrafish) for all the analyses.

For the TBA, MAVID and MLAGAN alignments, any column containing the gap character in human was removed. This was done for consistency with the Pecan alignment, which contains no gaps in human.

Multiple sequence alignments based on the September 2005 sequence freeze of the ENCODE Multi-Species Sequence Analysis[10] were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg17/encode/alignments/SEP-2005/alignments/ for TBA, MAVID and MLAGAN and from http://www.ebi.ac.uk/~bjp/pecan/encode_sept_pecan_mfas_proj.tar.bz2 for Pecan. The phylogeny with branch lengths that was input to StatSigMA-w and used to determine the species order in all figures was downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg17/encode/alignments/SEP-2005/phylo/tree_4d.tba.v2.nh.

Preprocessing long gaps. There are great differences in gap length distribution among the alignment programs: MAVID, MLAGAN and Pecan tend to use very long gaps, whereas TBA prefers to simply omit the species from that portion of the alignment rather than assign it a very long gap. **Supplementary Figure 4** shows the length distribution of gaps for these four alignments in ENm003, a representative ENCODE region. In the TBA alignment, most of the gaps have length 1–5 bp, and there is no gap longer than 200 bp. The other alignments, in contrast, have a much greater fraction of gaps longer than 50 bp.

The aligner's decision of whether to omit a species *S* from a given region or assign a long gap to *S* in that region is arbitrary, and this arbitrary decision could have affected our comparisons. For example, in our measurement of level of agreement among the alignments, it could make the difference between a human coordinate *h* being labeled disagree (if species *S* has a long gap at *h* in some other alignment) or unique (if species *S* is omitted at *h* in the other alignments). More importantly, it would affect the accuracy assessment, as StatSigMA-w treats gaps very differently from the way that it treats absent species[35].

Therefore, to put the alignments on equal footing for comparison, we preprocessed all the alignments to remove gaps longer than 20 bp, as though the species containing such a long gap is simply absent from that alignment region. For example, if a human subsequence was aligned to a gap of length 30 bp in mouse, we treated this human subsequence as unaligned in mouse after the removal of long gaps, though of course still possibly aligned to other species. The threshold of 20 bp was chosen so as to make the gap length distributions of the alignments much more comparable (see **Supplementary Fig. 4**).

Genomic location categories. Our comparisons are all broken down into the four distinct location categories of coding, UTR, intronic and intergenic. Human sites are categorized according to annotated UCSC Known Genes downloaded from the UCSC Genome Browser (assembly July 2007). Because two known genes may overlap (for example, because of alternative splicing), a single human site may fall into more than one category. In order to assign exactly one category to each human site, we use the following priority order for the four location categories, listed from highest to lowest priority: coding, UTR, intronic, intergenic. For example, if a site is contained in a coding exon in one isoform and in an intron in another, that site will be categorized as coding.

Comparison percentages. The ENCODE alignments are human-centric and represented in the coordinates of the human sequence. We therefore use the human sequence as our reference when measuring the level of agreement of the alignments. We compare, for each coordinate *h* in the human sequence and each nonhuman species *S*, the coordinate in *S* (if any) that is aligned to *h* by the alignments. By "coordinate" we mean the chromosome (or scaffold) name together with the position within that chromosome (or scaffold). Given an alignment *A* (for example, TBA), a nonhuman species *S* (for example, mouse), and a human coordinate *h*, when *A* aligns human coordinate *h* to coordinate *s* in *S*, there are three possible cases:

1. If there is at least one other alignment that also aligns *h* to *s*, we say that A "agrees."
2. If *A* is the only alignment that aligns human coordinate *h* to the target species *S*, we say that *A* is "unique."
3. If there is some other alignment that aligns human coordinate *h* to something in *S*, but the aligned coordinate in *S* is not *s* for any other alignment, we say that *A* "disagrees."

Another case that must be considered is when *A* aligns human coordinate *h* to a gap "–" in *S*. If two alignments both align *h* to a gap, we do not simply conclude that these two alignments agree on *h*. Instead, we take into consideration the contexts of the aligned gaps. Suppose, for example, that both TBA and MAVID align a human coordinate *h* to a gap in mouse. Let the first aligned mouse coordinate to the left and right of the gap be m_L and m_R, respectively, for TBA, and m'_L and m'_R, respectively, for MAVID. For the human coordinate *h*, TBA and MAVID will be considered to agree if and only if $m_L = m'_L$ and $m_R = m'_R$. Otherwise, the two alignments will be considered to disagree, because TBA's gap and MAVID's gap are actually inserted between different pairs of mouse coordinates.

For a given target species *S*, the absolute number of human coordinates for which alignment *A* agrees, disagrees or is unique is significantly influenced by the coverage of *S* in *A*, as defined in the section on "Alignment coverage." Instead, we calculate the *percentage* of the human coordinates in each of those three categories among all human coordinates aligned to *S* by *A*. These comparison percentages are denoted "agree%," "unique%" and "disagree%," respectively. Note that (i) the sum of these three percentages is 100% for any fixed alignment *A* and species *S*; (ii) the comparison percentages may differ for varying alignments *A*; and (iii) the definition of agree% only requires the target alignment to agree with one other alignment.

All the ENCODE alignments except Pecan's provide coordinates for the aligned species. Because this information is absent from the Pecan alignment, we must omit Pecan from the agreement comparisons. Because our goal is to measure the extent to which different large-scale genomic alignments agree and disagree with each other, the trends are clear enough even without Pecan.

It is worth noting that the comparison labels assigned to different alignments for the same human coordinate are not independent. For example, if a particular human coordinate is labeled *agree* for one alignment, this coordinate must be labeled the same way for at least one of the other alignments.

StatSigMA-w. Given any multiple sequence alignment and a phylogeny of the aligned sequences, StatSigMA-w[35] assesses the accuracy of the alignment and identifies suspiciously aligned regions. It is based on a statistical model that generalizes the Karlin-Altschul theory[39] from pairwise to multiple sequence alignment.

More specifically, StatSigMA-w assigns a "discordance score" to every site of the alignment and identifies a set of worst-aligned species for that site. (See ref. 35 for details. StatSigMA-w actually identifies a branch of the phylogeny whose removal would separate the species aligned at that site into two subsets that may be misaligned to each other, depending on the value of the discordance score; any species separated from human by this branch is referred to as a 'worst-aligned species'.) The discordance score, much like a *P* value, ranges between 0 and 1 and measures how likely it is that a worst-aligned species at that site is misaligned to the human sequence, with higher score indicating greater likelihood of misalignment. In practice, discordance scores show a bimodal behavior, with nearly all alignment columns having score either >0.1 or $<10^{-4}$ (see Fig. 1 of ref. 35). This bimodality makes discordance values somewhat more intuitive, classifying alignment columns neatly into those that appear well aligned (score $<10^{-4}$) and those that suggest poor alignment (score >0.1).

For each nonhuman species *S*, StatSigMA-w next identifies "suspiciously aligned regions," which are regions of the alignment (i) that have at least 50 sites, (ii) in which all sites have discordance score at least 0.1, with *S* being a worst-aligned species at each site, and (iii) that do not contain too many gaps. (See ref. 35 for details.) Given the bimodality described above, if the threshold of 0.1 were changed to 10^{-2} or 10^{-4}, the suspicious-region predictions would hardly change. The threshold of 50 sites focuses attention on those moderate to long regions where *S* appears to be misaligned.

Using the phylogeny generated for the ENCODE regions[10], we ran StatSigMA-w on all four ENCODE alignments. The identified suspicious regions are available as UCSC Genome Browser custom tracks, for all four alignments and all 22 species, at http://bio.cs.washington.edu/encode-msa/. As a summary figure, the percentage of aligned sites of species *S* that fall in StatSigMA-w suspicious regions for *S* is denoted "suspicious%." We use the suspicious% values of each species to compare the accuracy of the four ENCODE alignments.

Comparing suspicious and alternative alignments. This section describes the procedure that was used to create the scatter plots of **Figure 4** and **Supplementary Figure 3**. Given a target alignment (say, MLAGAN), a target species (say, baboon), and an alternative alignment (say, Pecan), we performed the following analysis for each region of the MLAGAN alignment that StatSigMA-w identified as suspiciously aligned for baboon. Let *h* be the human genomic region in this suspicious alignment. If Pecan, the alternative alignment, does not align *h* to some sequence in baboon, or if this region overlaps a suspicious region for baboon in the Pecan alignment, discard *h* and go on to the next suspicious region. Otherwise, let A_M and A_P be the human-baboon alignments of MLAGAN and Pecan, respectively, for the human region *h*, and let B_M and B_P be the baboon sequences aligned to *h* by MLAGAN and Pecan, respectively. If either of B_M or B_P is a substring of the other, discard *h* and go on to the next suspicious region. (This is a proxy for MLAGAN and Pecan agreeing on part of their alignment, because Pecan does not supply nonhuman genomic coordinates.)

At this point we are left with a suspicious MLAGAN human-baboon alignment A_M and a nonsuspicious Pecan human-baboon alignment A_P that do not agree. We then compute pairwise alignment scores S_M and S_P, respectively, of these two alignments using the following BLASTN scoring function: for mouse and zebrafish, the scoring function is +1 for match and −1 for mismatch or gap; for baboon, the mismatch score is −2 to reflect the smaller divergence between human and baboon[40]. Add a length-normalized point $(S_M/L_M, S_P/L_P)$ to the scatter plot of **Figure 4**, where L_i is the length of the alignment A_i. Repeat the procedure with TBA and MAVID replacing Pecan as the alternative alignment.

To ensure that the differences between the baboon scatter plots and those of mouse and zebrafish are not due to differences in the alignment scoring function, we also created scatter plots for baboon using the same scoring function as those used for mouse and zebrafish. This had little noticeable effect on the patterns of the scatter plots (**Supplementary Fig. 3**).

38. Wang, L. & Jiang, T. On the complexity of multiple sequence alignment. *J. Comput. Biol.* **1**, 337–348 (1994).
39. Karlin, S. & Altschul, S.F. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci. USA* **90**, 5873–5877 (1993).
40. States, D.J., Gish, W. & Altschul, S.F. Improved sensitivity in nucleic acid database searches using application-specific scoring matrices. *Methods: A Companion to Methods in Enzymology* **3**, 66–70 (1991).

doi:10.1038/nbt.1637