BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 17 2010, pages 2116–2120
doi:10.1093/bioinformatics/btq360

Sequence analysis

Advance Access publication July 7, 2010

Genome-wide functional element detection using pairwise statistical alignment outperforms multiple genome footprinting techniques

R. Satija[1,*], J. Hein[1] and G. A. Lunter[2]

[1]Department of Statistics and [2]Wellcome Trust Centre for Human Genetics, Oxford University, Oxford, UK

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Comparative genomic sequence analysis is a powerful approach for identifying putative functional elements *in silico*. The availability of full-genome sequences from many vertebrate species has resulted in the development of popular tools, for example, the phastCons software package that search large numbers of genomes to identify conserved elements. While phastCons can analyze many genomes simultaneously, it ignores potentially informative insertion and deletion events and relies on a fixed, precomputed multiple sequence alignment.

Results: We have developed a new method, GRAPeFoot, which simultaneously aligns two full genomes and annotates a set of conserved regions exhibiting reduced rates of insertion, deletion and substitution mutations. We tested GRAPeFoot using the human and mouse genomes and compared its performance to a set of phastCons predictions hosted on the UCSC genome browser. Our results demonstrate that despite the use of only two genomes, GRAPeFoot identified constrained elements at rates comparable with phastCons, which analyzed data from 28 vertebrate genomes. This study demonstrates how integrated modelling of substitutions, indels and purifying selection allows a pairwise analysis to exhibit a sensitivity similar to a heuristic analysis of many genomes.

Availability: The GRAPeFoot software and set of genome-wide functional element predictions are freely available to download online at http://www.stats.ox.ac.uk/~satija/GRAPeFoot/

Contact: satija@stats.ox.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 20, 2009; revised on June 7, 2010; accepted on June 30, 2010

1 INTRODUCTION

Despite the wealth of currently available DNA sequence, the important task of identifying functionally important biological regions remains difficult. One popular approach is to search for regions that are putative targets of purifying selection by virtue of being conserved at a variety of evolutionary distances (Tagle *et al.*, 1988). While mutational cold spots could also explain the existence of conserved elements, population genetics studies refute this hypothesis and conclude that purifying selection maintains conserved elements in both *Drosophila* and human genomes (Casillas *et al.*, 2007; Drake *et al.*, 2005).

There exists a clear need for general methods that identify functional DNA in large mammalian and vertebrate genomes. Previous genome-wide conservation studies have estimated that 3–8% of the human genome is under purifying selection (Chiaromonte *et al.*, 2003; Lunter *et al.*, 2006; Siepel *et al.*, 2005). However, protein-coding elements have been estimated to make up only 1–2% of the genome (International Human Genome Sequencing Consortium, 2004). Non-protein-coding conserved material, thus, represents a large fraction of all functional DNA, and specialized techniques such as exon-finders cannot identify the majority of the dark matter present in the human genome.

Two general techniques have been used for identifying putative functional regions in the human genome by searching for regions with reduced rates of mutation. An example of an approach focusing solely on substitution data is the phastCons software package (Siepel *et al.*, 2005). phastCons utilizes a phylogenetic hidden Markov model (phylo-HMM) which annotates regions in a given multiple sequence alignment as conserved and non-conserved by modelling conserved regions with reduced rate of point mutations. The use of an HMM framework allows phastCons to estimate all model parameters, including the length and expected level of conservation, directly from the sequence data. Theoretical studies have concluded that comparative studies may exhibit increased statistical power when analyzing datasets with larger numbers of genomes, and phastCons has been applied to both a set of 17 and a set of 28 vertebrate genomes (Eddy, 2005; Margulies *et al.*, 2005).

A complementary approach, implemented in the consIndel software package, searches a precomputed multiple sequence alignment for regions with a reduced rate of insertion and deletion (indel) mutations (Lunter *et al.*, 2006). In this approach, the annotated set of ancestral repeats (ARs) is analyzed to determine the distribution of ungapped regions (sequences between indel events) in neutral regions. When the entire genome is examined, long ungapped regions that deviate from this null distribution were found to be highly enriched with previously annotated functional elements. Both phastCons and consIndel have dedicated tracks on the UCSC genome browser (Miller *et al.*, 2007).

While these techniques successfully identify many previously annotated functional elements, both make their predictions based on a paucity of either substitution or indel mutations. Additionally, both techniques make their predictions based on a fixed multiple sequence alignment. While this reduces the computational complexity of both methods, uncertainty or errors in the alignment will impact

*To whom correspondence should be addressed.

the downstream functional element predictions. Especially as the number of genome sequences increases, thereby exponentially increasing the number of possible multiple sequence alignments and thus the difficulty of the multiple-alignment problem, the reliance on a fixed alignment can result in flawed annotations (Margulies, 2008). Additionally, previous studies have demonstrated how standard score-based alignment methods produce biased alignments that may drastically underestimate the level of indel mutations (Lunter, 2007b; Lunter *et al.*, 2008).

Statistical alignment offers a framework for combining sequence alignment with comparative sequence annotation, thereby removing the traditional dependence on a fixed alignment (Hein *et al.*, 2000). We have previously shown how computing a probability-weighted alignment distribution can increase predictive accuracy by correctly accounting for alignment error or uncertainty during annotation (Satija *et al.*, 2008, 2009). Our method, combining statistical alignment and phylogenetic footprinting (SAPF), annotates functional elements in short genomic regions using a HMM which modeled both a reduced rate of substitution and indel events in conserved regions. However, since the number of hidden states in the model and the number of potential alignments that must be analyzed both increase exponentially with the number of sequences, expanding this full probabilistic treatment to the entire genome for multiple sequences is computationally infeasible.

It is possible to extend SAPF to perform full probabilistic annotation of two genomes by building upon the probabilistic aligner GRAPe (Lunter, 2007a; Lunter *et al.*, 2008). GRAPe has been found to produce superior alignments of the human and mouse genomes using a technique called Marginalized Posterior Decoding. While the reduced number of genomes may potentially reduce the signal available for phylogenetic footprinting, previous studies have shown how this loss of signal could be fully offset by using probabilistic alignment techniques (Lunter *et al.*, 2008). Our new software tool, GRAPeFoot, combines statistical alignment and annotation of the human and mouse genomes in order to identify regions with reduced rates of both substitution and indel mutations.

2 METHODS

2.1 Model overview

GRAPeFoot uses a pairwise HMM (Durbin *et al.*, 1998) to simultaneously align and annotate two DNA sequences. The HMM contains two sets of hidden states: one set models neutrally evolving sequences (fast evolution) and a second set models conserved elements (slow evolution). Transition and emission parameters in slow states are set to model a reduced rate of evolutionary events (substitutions, insertions and deletions). The model can switch between fast and slow states in order to annotate both neutral and functional subsequences. Since we cannot model a continuum of constraint strengths, our method uses a single category of states to represent conserved sequence as an approximation to the range of conservation levels present in functional elements. We address one of the issues related to this approximation when training our model parameters.

2.2 The GRAPeFoot HMM

GRAPeFoot models neutral sequences using the standard pairwise alignment HMM implemented in the GRAPe software package. The HMM consists of the basic three state (*match*, *insert* and *delete*) HMM with the addition of two extra indel states (*insert*2 and *delete*2), which specifically model long indel events. These extra states allow the length of insertion and deletion events to be modeled as a mixture of two geometric distributions, one modeling the length of short indel events and the other modeling long indels. This was found to result in a small but significant improvement in alignment quality (Lunter *et al.*, 2008). To complete the GRAPeFoot HMM, we add three states (*slow_match*, *slow_insert* and *slow_delete*) modeling conserved elements, and enable transitions between the *match* and *slow_match* states, thus allowing the model to alternate between annotating neutral sequence and conserved elements. We assume that there are only short indel events in conserved regions and thus there is no need for extra indel states.

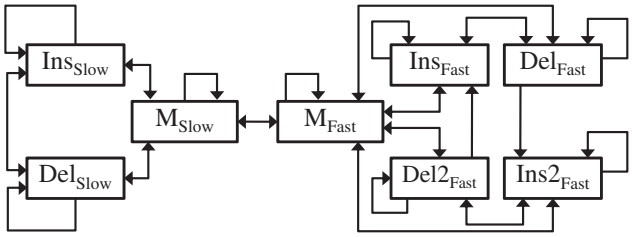

Fig. 1. Graphical representation of the GRAPeFoot HMM. Slow states are used to model conserved functional elements and fast states model background sequence. An additional pair of insertion/deletion states (*insert*2, *delete*2) model a mixture geometric indel length distribution in background sequence (Lunter *et al.*, 2006).

A path of state transitions and emissions through the GRAPeFoot HMM represents both a pairwise sequence alignment and an annotation of each alignment column as either neutral sequence or a conserved element. The posterior probabilities for each state path can be calculated after running the standard Forward and Backward algorithms (Durbin *et al.*, 1998). We integrate over all possible alignments to calculate, for each base in the human genome, the predicted posterior probability that it is part of a functional element.

2.3 The GRAPe whole-genome aligner

The HMM described in Figure 1 is the main building block of the GRAPeFoot algorithm. It is used to assess homology and compute alignments of short segments, up to 1000 bp, of DNA. In brief, the algorithm works by first aligning short segments around guides, or putatively homologous genome loci. We used BlastZ alignments as input, from which GRAPeFoot generates one guide for every 20 BlastZ alignment columns on-the-fly. The large density ensures that all BlastZ alignments are considered, but does not preclude alternative alignments from being considered. It is also possible to provide an explicit list of alignment guides, which may be generated from, for example, BLAST or BLAT searches.

Around each guide, local homology is assessed, and a local alignment is computed if the homology test succeeds. Homology testing is implemented by using a modified Fisher–Yates algorithm to permute both sequences while keeping tri-nucleotide counts fixed. The alignment log-likelihood for the permuted sequences is calculated by the Forward algorithm, and their mean and variance is estimated using a minimum of six permuted sequence pairs. When their log-likelihood exceeds that of the permutations by 2 SDs plus a threshold of log(10.0) which was determined by trial and error.

When alignments pass the homology test and extend sufficiently far, further guides are generated on either side of the aligned sequence, in such a way that when the alignment in fact extends beyond the 1000 bp window, the resulting alignments will overlap. When two alignments generated from such guides in fact overlap, and exactly agree in at least one 'match' alignment column with posterior support over 0.25 , they are combined into a single larger alignment. This process continues until no further homologous sequence is detected.

In this way, the algorithm is able to iteratively grow large aligned sequences, potentially far beyond the initial alignment guides. When the non-overlapping sequences are far apart, the alignments are output separately.

2.4 Estimating model parameters

The model transition and emission parameters control the expected lengths of fast and slow regions as well as the expected rates of substitutions, insertions and deletions in each. The transition probabilities between match and indel states set the expected rate of indel events, and the emission probabilities from the match states, represented by a substitution probability matrix, set the expected rate of substitution.

From previous work (Lunter, 2007b), we took a set of substitution probability matrices, one from each of 20 G+C bands. The data were split into different G+C bands in order to account for observed variation in substitution and indel mutation probabilities for sequences with different G+C content (Lunter, 2007b). These substitution probability matrices, denoted by $\hat{p}$ correspond to emission probabilities from the *match* state, and represent proportions of each homologous nucleotide pairing observed in neutral sequence. We used this information to estimate a neutral substitution rate matrix, denoted by $\hat{Q}$ which contains the rate of change for each base to every other base. These matrices are related by the equation $\hat{p}=e^{\hat{Q}t}$, where t is the divergence time between the two species. To calculate the slow HMM state emission probabilities for the *slow_match* state, corresponding to the matrix $p_{\hat{slow}}$, we use the formula $p_{\hat{slow}}=e^{\hat{Q}\alpha t}$. This allows us to model the reduced rate of substitution mutations in conserved regions by a single parameter α, constrained to be <1, while retaining different emission probabilities for each G+C band.

The rates of indel mutations are set by the transition probabilities from the *match* to the *indel* states. As with the emission probabilities, we obtained a set of indel probabilities for different G+C bands from previous work (Lunter, 2007b) and used these to model non-conserved regions. To calculate the indel probabilities for conserved regions, we reduced the corresponding transition probabilities in the non-conserved regions by a factor β. The scaling factors α and β are two GRAPeFoot model parameters used to generate the complete set of transition and emission probabilities. Two additional model parameters, δ and ϵ, were used to define the geometric distributions modeling the length of conserved and non-conserved regions, and correspond to the probabilities of a fast state transitioning to a fast state, or a slow state transitioning to a slow state. The final model parameter, θ, is the self-transition probability applying to both the *slow_insert* and *slow_delete* state and thus sets the expected length of insertions in slow regions. This extra parameter enables GRAPeFoot to model indel events in functional and background sequence as differing both in frequency and in length.

While phastCons estimates some HMM parameters via maximum likelihood estimation (MLE), we chose a different approach for parameter estimation. We expected that a significant proportion of the functional sequence in the genome would be highly conserved, and the MLE parameters would be set to optimally detect these elements at the expense of detecting weakly conserved elements. Thus, parameters which maximized the likelihood of our model could cause our method to miss functional elements with lower levels of conservation. In contrast, model parameters that are trained to efficiently identify weakly conserved regions allow the model to identify highly conserved elements with good power as well. Thus, we utilized a supervised learning approach on a segment of chromosome 21 corresponding to 1% of the human genome in order to obtain our parameter estimates. We trained our parameter set to maximally identify weakly conserved exons, while requiring that the false positive rate remained low. For comparison, we also computed a set of MLE parameters and found that our supervised learning parameters, while decreasing the overall likelihood, were far more tolerant to functional regions that were not perfectly conserved and allowed our model to better discriminate between annotated exons and ARs in this region. This analysis, along with the exact parameter values used by GRAPeFoot, is presented in the Supplementary Material.

2.5 Calling individual elements

The output of the Forward and Backward algorithms can be used for posterior decoding (Durbin *et al.*, 1998), a process which returns the predicted posterior probability that each individual base in the human genome has been subjected to the effects of purifying selection. While these base pair probabilities are useful, many researchers are most interested in a set of contiguous elements which are predicted to be functional. One method of producing such a list is to use the Viterbi algorithm, which calculates the single most likely state path through the HMM, and thus does not consider the influence of multiple alternative alignments. Instead, we converted individual base probabilities into a list of contiguous elements. We defined an element as a continuous stretch of DNA with a minimum length of 5 nt where the annotated posterior probability of each base was greater than a probability threshold. For example, using a probability threshold of $P=0.6$, we examined the set of all bases with annotated posterior probabilities $>60\%$, and all contiguous sequences with a length >5 nt defined the set of called elements. Using different probability thresholds alters the stringency of element calling and allows us to control for false discovery rate. We used the approach of Lunter *et al.* (2006) and created 10 different sets of called elements, each using a different probability threshold. Four of these sets can be downloaded from the GRAPeFoot web site and easily imported into the UCSC genome browser, and the remaining datasets are available from the authors upon request. The Supplementary Material also provides basic statistics, such as the percentage of the genome predicted to be functional, for each of these predicted element sets.

3 RESULTS

In order to assess the performance of GRAPeFoot, we benchmarked its performance against the phastCons and consIndel software packages. We tested the three software package on five different types of previously annotated elements in the human genome.

- Exons: we used the set of exons from the UCSC known genes track, which accumulates genes from a variety of sources based on protein data (SWISS-PROT and TrEMBL) and mRNA data (RefSeq, Genbank).
- Putative transcription regulatory regions (pTRRs): these are a set of putative transcriptional regulatory sequences that have been identified from ENCODE experimental data. The 100 bp regions were identified via data from chromatin immunoprecipitated samples hybridized to high-density microarray chips (ChIP-chip) for sequence-specific transcription factors. Set specificity was increased by requiring additional experimental evidence (chromatic activation, DNAase hypersensitivity and nucleosome depletion).
- Predicted regulatory modules (PReMods): a genome-wide set of computationally predicted human transcriptional modules. Predictions are made by searching for clusters of phylogenetically conserved and repeating transcription factor motifs using a motif database consisting of 229 TFs.
- microRNAs: we used the miRBase database, a database of published microRNA sequences and annotations.
- ARs: we used the set of repeats predicted by the RepeatMasker program, which screens DNA sequences for interspersed repeats and low complexity DNA sequences. Since ARs are presumed to evolve neutrally, we benchmark the performance in finding these sequences in order to estimate a false positive rate.

We wanted to compare the proportion of previously annotated elements detected by each of the three methods for a variety of different thresholds. To compare the three methods, we constructed receiver operating characteristics (ROC) curves for each method

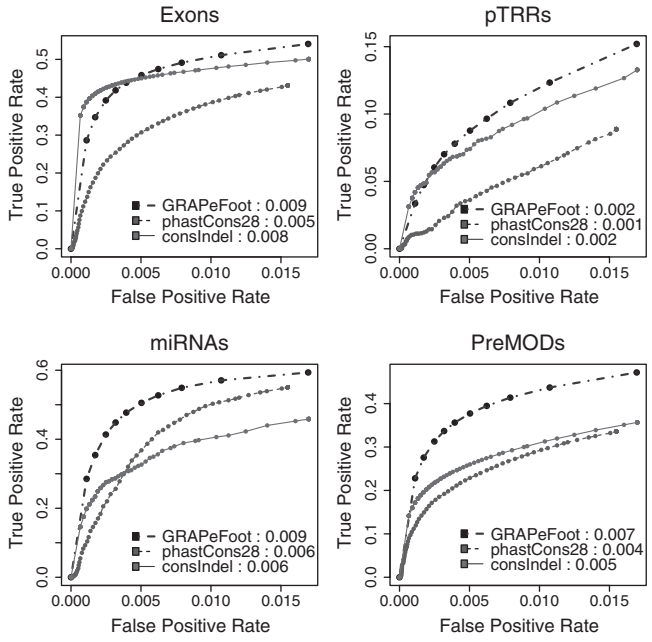

Fig. 2. ROC curves comparing the performance of GRAPeFoot, phastCons and consIndel for four classes of functional elements. The legend for each plot displays the area under each of the ROC curves.

for each of the different functional element types. The *y*-axis of the ROC curve represents the true positive rate: the percentage of nucleotides within a functional element class that are correctly annotated as conserved (true positives), and the *x*-axis represents the false positive rate: the percentage of nucleotides within ARs detected by the method (false positives) for the same threshold. The curve is constructed by placing and connecting points on the plot for a range of different thresholds. For the GRAPeFoot results, each point on the ROC curve corresponds to 1 of the 10 sets of contiguous elements described in Section 2.5. Both phastCons and consIndel predictions assign each predicted contiguous element a score ranging from 0 to 1000. When constructing ROC curves for these predictions, we used these scores for thresholding. Each point on the ROC curves can be easily reproduced using the base-pair-wise intersection feature on the UCSC genome browser after loading in our publicly available list of conserved element predictions. We demonstrate an example of this in the Supplementary Material.

Since the threshold value does not appear on either axis, the ROC curve allows us to compare method performance for a variety of different sensitivity/specificity regimes. Figure 2 displays ROC curves for four classes of functional elements: exons, pTRRs, miRNAs and preMODs.

The legend for each plot displays the area under the curve (AUC), a summary statistic that evaluates the method's performance while taking into account both sensitivity and specificity. Higher AUC values correspond to increased predictive accuracy. Each ROC curve has been modified to exclude the point (1, 1), which is present by definition in all ROC curves. This point corresponds to using a score or probability cutoff of zero in which case each method would annotate each nucleotide in the genome as functional, corresponding to both a true positive rate and a false positive rate of 100%. The inclusion of this point on the graph, however, hinders the ability to compare the ROC curves for the different methods, and thus we chose to omit the point from both the graphs and the AUC calculations.

The ROC analysis in Figure 2 shows that GRAPeFoot performs well when compared with both phastCons and consIndel for all four types of functional elements. Additionally, the AUC values demonstrate that the quality of predictions varies for the different methods. For all methods, accuracy was highest when detecting exons, followed (in decreasing order) by miRNAs, PReMods and pTRRs. This performance ordering was not unexpected given the average length and expected conservation of these functional elements. Both theoretical and genome-based studies have shown that the ability to detect functional elements via comparative genomics improves with longer elements and higher levels of conservation (Eddy, 2005; Stark *et al.*, 2007). Exons have an average length of 170 bp and are in general well conserved among vertebrates while microRNAs are significantly smaller, with an average length of 24 bp, and exhibit a reduced but similar level of conservation as exons (Bartel, 2004). In contrast, regulatory modules are expected to contain large percentages of neutral DNA interspersed with regulatory signals such as transcription factor binding sites, which have an average length of 6–15 bp and can exhibit high degrees of variability (GuhaThakurta, 2006). Thus, while all pTRRs were constrained to be 100 bp, only a small proportion of these bases (e.g. a pTRR could contain only one transcription factor binding site) may be functional and therefore subject to purifying selection. While all methods performed significantly better when detecting PReMODs, another set of putative regulatory modules, these modules were identified by searching for clusters of phylogenetically conserved motifs and therefore were assured to exhibit at least partial sequence conservation.

4 CONCLUSIONS

Our results demonstrate the potential value of comparative analyses performed on just two sequences by using integrated modeling of substitutions, indels and selection. Though GRAPeFoot had the benefit of analyzing only the human and mouse genomes, our software identified constrained elements at rates comparable with phastCons, which utilized data from 28 full vertebrate genomes. These results imply that sequencing large numbers of genomes and running heuristic methods in order to process all the data may not be the only effective way to decode the locations of functional elements in the genome. Improved modeling methods, even if they can only run on a reduced number of genomes, have the potential to squeeze more information out of the data, and can also be run on datasets where large numbers of extant species have not been sequenced.

GRAPeFoot exhibits how statistical alignment techniques can be used to extend traditional comparative analyses. While both phastCons and consIndel require a precomputed fixed alignment, GRAPeFoot simultaneously performs probabilistic alignment and conservation analysis, a process which significantly improves upon the accuracy of traditional pairwise genomic alignment. GRAPeFoot avoids relying on score-based alignment practices which often use integer penalties to represent substitution and indel mutations in order to find the single alignment with the highest score, a process that results in known alignment biases (Lunter *et al.*, 2008). In contrast, GRAPeFoot's probabilistic alignment framework utilizes parameters which characterize evolutionary models, and

GRAPeFoot sums over all possible indel configurations in order to remove these biases. These improvements enable GRAPeFoot to increase the accuracy of both the sequence alignment and the footprinting predictions.

The benchmarking results also demonstrate how both indel mutations and substitutions are crucially informative and should be considered in conservation studies. While most footprinting methods focus on finding regions with low numbers of substitutions, we found that consIndel performed similarly to phastCons despite ignoring substitution information and focusing only on the frequency of indel events. In contrast, GRAPeFoot considers both substitution and indel mutations to be informative. Additionally, the transition parameters in GRAPeFoot allowed the model to discriminate between neutral and functional sequence not only based on the frequency of indel events, but also on their length as well. This ensures that short and adjacent indel mutations (which may be offsetting) will not incorrectly cause GRAPeFoot to annotate a region as neutral.

While we would like to extend GRAPeFoot to analyze multiple sequences, it is currently infeasible to do full probabilistic alignment and footprinting of even three sequences on a genome-wide scale. We are currently attempting to develop methods to approximate the multiple alignment distribution instead of utilizing a full dynamic programming calculation. One possible future route involves approximating the multiple-alignment distribution by analyzing the full set of pairwise alignments. We are also modifying the GRAPeFoot HMM to allow for the insertion and deletion of functional elements. When analyzing data from multiple sequence data, this modification would allow GRAPeFoot to annotate the evolution of functional elements along a tree and to trace the evolutionary history of different element classes.

ACKNOWLEDGEMENTS

We thank István Miklós and Rune Lyngsø for helpful discussions. We also thank the Oxford Supercomputing Centre for providing computational resources.

Funding: Rhodes Trust, UK (R.S.).

Conflict of Interest: none declared.

REFERENCES

Bartel,D. (2004) MicroRNAs genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Casillas,S. *et al.* (2007) Purifying selection maintains highly conserved noncoding sequences in Drosophila. *Mol. Biol. Evol.*, **24**, 2222.

Chiaromonte,F. *et al.* (2003) The share of human genomic DNA under selection estimated from human-mouse genomic alignments. In *Cold Spring Harbor Symposia on Quantitative Biology*, Vol. 68, CSHL Press, NY, pp. 245–254.

Drake,J. *et al.* (2005) Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat. Genet.*, **38**, 223–227.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis*. Cambridge University Press, New York.

Eddy,S. (2005) A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.*, **3**, e10.

GuhaThakurta,D. (2006) Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Res.*, **34**, 3585.

Hein,J. *et al.* (2000) Statistical alignment: computational properties, homology testing and goodness-of-fit. *J. Mol. Biol.*, **302**, 265–279.

International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

Lunter,G. (2007a) HMMoC a compiler for hidden Markov models. *Bioinformatics*, **23**, 2485.

Lunter,G. (2007b) Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. *Bioinformatics*, **23**, i289.

Lunter,G. *et al.* (2006) Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput. Biol.*, **2**, e5.

Lunter,G. *et al.* (2008) Uncertainty in homology inferences: assessing and improving genomic sequence alignment. *Genome Res.*, **18**, 298.

Margulies,E. (2008) Confidence in comparative genomics. *Genome Res.*, **18**, 199.

Margulies,E. *et al.* (2005) An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl Acad. Sci.*, **102**, 4795–4800.

Miller,W. *et al.* (2007) 28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res.*, **17**, 1797.

Satija,R. *et al.* (2008) Combining statistical alignment and phylogenetic footprinting to detect regulatory elements. *Bioinformatics*, **24**, 1236.

Satija,R. *et al.* (2009) BigFoot: Bayesian alignment and phylogenetic footprinting with MCMC. *BMC Evol. Biol.*, **9**, 217.

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034.

Stark,A. *et al.* (2007) Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature*, **450**, 219–232.

Tagle,D. *et al.* (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.*, **203**, 439–455.