

# Genome-wide nucleotide level mammalian ancestor reconstruction Supplementary Material

Benedict Paten<sup>\*1</sup>, Javier Herrero<sup>3</sup>, Stephen Fitzgerald<sup>3</sup>, Kathryn Beal<sup>3</sup>, Paul Flicek<sup>3</sup>, Ian Holmes<sup>2</sup>, Ewan Birney<sup>\*3</sup>

<sup>1</sup> Department of Engineering, University of California, Santa Cruz CA, USA <sup>2</sup> Department of Bioengineering, University of California, Berkeley CA, USA <sup>3</sup> The Ensembl Group, EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Email: Benedict Paten<sup>\*</sup>- [benedict@soe.ucsc.edu](mailto:benedict@soe.ucsc.edu); Ewan Birney<sup>\*</sup>- [birney@ebi.ac.uk](mailto:birney@ebi.ac.uk);

<sup>\*</sup>Corresponding author

**Keywords** Large-scale genomics, Ancestor-reconstruction, Phylogenetic alignment.

## Supplemental Ortheus Results

This document provides six supplemental results to the main paper: Genome-wide nucleotide level mammalian ancestor reconstruction. Firstly, a figure (Figure S1) showing a combined three branch transducer model, as used by Ortheus v1.0. Secondly, an analysis of the stability of Ortheus reconstructions between different runs of the program. Thirdly, a description and instructions for using the simulation models used in this paper with the GSimulator program. Fourthly, an experiment to show that the memory saving heuristic used by Ortheus has little significant effect on its output. Fifthly, a figure (Figure S3) showing the lineage specific insertion and deletion rates observed in a reconstruction of the CFTR region using data from the ENCODE consortium. Finally, a description of a supplemental spread-sheet containing the complete set of data on the discovery of AGRs.

### Testing The Stability Of Ancestor Reconstructions

To test Ortheus we used sequences from the ENCODE sequence dataset [Margulies et al., 2007] [Consortium et al., 2007], restricting our attention to the well studied cystic fibrosis transmembrane conductance region (CFTR) locus, covering 1.87 megabases of the Human genome sequence. See the main text for a description of this reconstruction.

The stability of predicted ancestor sequences was tested using a similar methodology to that described in Blanchette et al. [Blanchette et al., 2004]. For each ancestor node within the tree we examined two ancestor sequences, generated from separate runs of the program, each time seeding the random number generator with a different input. To estimate the ancestral bases of the sequences we used the maximum likelihood (ML) estimates of the ancestral bases. To compare the ancestor sequences we used Pecan [Paten et al., 2008] to first align them, and then calculated two metrics from the resulting alignment:

1. Residue disagreement: the number of matched pairs in the alignment with a substitution, divided by the average length of the two sequences.
2. Gap disagreement: the number of positions in both sequences aligned to gaps, divided by the combined length of the two sequences.

Subtracting these two values from one we arrive at the total agreement, the number of matched pairs in the alignment without a substitution, divided by the average length of the two sequences.

Figure S2(a,c,e) show the total agreement, gap disagreement and residue disagreement metrics, respectively, for all the ancestors on the lineage leading to humans as functions of the sample rate.

The total agreement quickly converges to greater than 95% for almost all ancestral nodes in the complete tree. The only exceptions to this being the root node and its direct descendent nodes, where a lack of out-groups and longer total branch lengths makes predictions less certain. Examining the disagreement between predictions (log-scales shown), the gap disagreement falls for most ancestral nodes, on average by about 0.5 to about 0.8 orders of magnitude between 1 and 200 samples. Gap disagreement appears to stabilise at a constant level for most ancestral sequences at around 50-100 samples/node. Residue disagreement falls more steeply, by about 0.8 to 1.6 orders of magnitude between one and two hundred samples. The visible separation of curves into two groups on the residue disagreement chart shows the distinct differences in evolutionary distance between the primate sub-tree and rest of the phylogeny. The residue disagreement for non-primate ancestral nodes appears consistently to continue to fall by a small amount between 100 and 200 nodes, though this difference represents on average only a 0.8% total change. Figure S2(b,d,f) show how constraint relaxation affects the stability of the predicted ancestors. Interestingly, the total agreement seems to decrease very slightly but consistently (average 0.8%) as the constraints are relaxed. It appears almost none of this difference comes from changes in the amount of gap disagreement, but in fact nearly all of the difference comes from changes in the amount of residue disagreement. Given that the alignment of the ancestor sequences to calculate these agreement and disagreement metrics is trivial, we believe that this source of residue disagreement is real, rather than an alignment artifact. However, we consider it encouraging, though definitely not a proof of global optimality, that as the constraints from the input alignment are relaxed the stability of the reconstructions does not change dramatically, despite the increase in size of the search space, while the log-probability and observed numbers of events appears to converge towards a limit.

## Simulation models

To generate simulated alignments of DNA, we used the forthcoming GSimulator program [Varadarajan et al., 2008], for generating synthetic DNA alignments. This program simulates local sequence-dependent fluctuations in substitution and indel rates, modeling effects such as CpG aversion or microsatellite expansion and contraction. Specifically, the tool generates a root sequence using a Markov model, then evolves the sequence along each branch of a phylogenetic tree using a finite-state transducer. Both the Markov model at the root, and the transducers on the branches, are context-dependent; that is, the emission and transition probabilities of the state machine depend on the last few absorbed and emitted nucleotides. (Note that the transducers used for simulation are, therefore, more parameter-rich than the

transducers used for reconstruction.) GSimulator can be ‘trained’ directly on pairwise alignment data; for the simulations described here, the program was trained on a random subset human chromosome 1 to chimpanzee Blastz [Schwartz et al., 2003] alignments downloaded from Ensembl [Flicek et al., 2007] version 49 and totalling just over 20 megabases.

We provide the root and branch transducer simulation models used to test Ortheus in the accompanying zipped directory. The models are identified as follows:

1. humanChumpChr1-Gotoh0.tra: Branch model with a set of single affine gap states and zero-contextual nucleotides.
2. humanChumpChr1-Gotoh1.tra: Branch model with a set of single affine gap states and one-contextual nucleotide.
3. humanChumpChr1-GotohMixture0.tra: Branch model with a mixture of two sets of affine gap states and zero-contextual nucleotides.
4. humanChumpChr1-GotohMixture1.tra: Branch model with a mixture of two sets of affine gap states and one-contextual nucleotide.
5. humanChimpChr1-Singlet0.tra: Root model with zero contextual nucleotides.
6. humanChimpChr1-Singlet1.tra: Root model with one contextual nucleotide.

We combined the zero-contextual nucleotide models together, and similarly the one-contextual nucleotide models together. The phylogeny used is contained in a newick tree in the same directory.

### **Testing the effects of breaking up the Ortheus computation into overlapping fragments**

Table S1 shows how the memory saving heuristic to chop up the reconstruction into overlapping fragments affects the stability of the reconstructions (see methods of main paper). Compared are different degrees of overlap, all compared with a maximum overlap of 500 columns. Clearly the disagreement between the reconstructions falls as the overlap is increased, but even with no overlap the increase in disagreement is very small because the number of fragments is small in the context of the entire alignment.

### **Ancestral Genic Region discovery summary**

Highly confident Ancestral Genic Regions (AGRs) were discovered as described in the main text. Briefly the Ensembl human protein set was used to search ancestral DNA sequence using Exonerate [Slater and

Birney, 2005] with the protein2genome model. The match from the orthologous extant human protein in an ancestral region was masked out, as were any matches which overlapped these exons, removing paralogous hits. The resulting set of matches were clearly enriched in complex gene prediction errors, for example, low complexity region matches and cryptic transposon matches; matches which hit > 50 times were removed as were matches which overlapped regions annotated by Seg [Wootton and Federhen, 1996]. This gave rise to 1,658 potential matches. These matches were then compared to the human extant protein using genewise [Birney et al., 2004] with permissive parameters. Only matches where the ancestral sequence had a match greater than 10bits to all extant sequences were kept. The resulting 32 sequences are shown in the attached excel spreadsheet. The columns are:

1. The HGNC name of the putative source gene
2. The Ensembl Gene id of the putative source gene
3. The Ensembl protein id of the putative source gene
4. The bits score for the genewise match of the source gene to the ancestor sequence
5. The best extant score of the source gene in this region
6. The difference between these scores
7. The best species for the extant match
8. The number of introns predicted in the ancestor
9. The number of introns in the source gene
10. Source chromosome
11. Source start
12. Source end
13. Ancestral sequence region in human chromosome
14. Ancestral sequence region in human start
15. Ancestral sequence region in human end
16. List of all extant regions, in the format (species,chromosome,start,end)

17. Internal tracking id for this analysis

## **Figures**

### **Figure 1 - 3-Branch Model**

A three branch transducer model that combines two affine branch transducers (X,Z branches) and one root transducer.

### **Figure 1 - The effects of sample rate and constraint relaxation on agreement of generated alignments**

The effects of changing the sample rate and diagonal constraint relaxation on the (b,c) total agreement, (d,e) residue disagreement and (f,g) gap disagreement values. Uses the same colour coding of the ancestors as in the main text.

### **Figure 3 - Trees showing lineage specific indel rates**

Phylogenetic trees of the reconstructed phylogeny showing the observed indel rates of each lineage.

**Table 1 - The effects of cutting up the reconstruction into overlapping fragments**

The effects of cutting up the reconstruction into overlapping fragments. Complete alignments of the sequences with different overlap distances (left column) were compared to a traceback distance of 500 columns. The total number of fragments was approximately the same in each. The total disagreement was calculated as the sum for all ancestors of one minus the total agreement score. Max-deviation is the greatest absolute difference observed from the average numbers during 5 repeats. The row showing an overlap of 500 columns shows the base-line disagreement between the programs during separate runs of the program with the same parameters.

## Bibliography

- Birney, E., Clamp, M., and Durbin, R., 2004. GeneWise and Genomewise. *Genome Res*, **14**:988–995.
- Blanchette, M., Green, E. D., Miller, W., and Haussler, D., 2004. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res*, **14**:2412–2423.
- Consortium, E. P., Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., *et al.*, 2007. Identification and analysis of functional elements in 1genome by the encode pilot project. *Nature*, **447**(7146):799–816.
- Flicek, P., Aken, B., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., *et al.*, 2007. Ensembl 2008. *Nucleic Acids Res*, .
- Margulies, E. H., Cooper, G. M., Asimenos, G., Thomas, D. J., Dewey, C. N., Siepel, A., Birney, E., Keefe, D., Schwartz, A. S., Hou, M., *et al.*, 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1human genome. *Genome Res*, **17**:760–774.
- Paten, B., Herrero, J., Fitzgerald, S., Beal, K., and Birney, E., 2008. Enredo and pecan: Genome-wide mammalian consistency based multiple alignment with paralogs. *Submitted*, .
- Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., and Miller, W., 2003. Human-mouse alignments with BLASTZ. *Genome Res*, **13**:103–107.
- Slater, G. S. and Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**:31–31.
- Varadarajan, A., Bradley, R., and Holmes, I., 2008. Trainable transducers and flexible phylogrammars for simulating genome evolution. *In preparation*, .
- Wootton, J. C. and Federhen, S., 1996. Analysis of compositionally biased regions in sequence databases. *Meth Enzymol*, **266**:554–71.



<i>Column Overlap</i>	<i>Avg. Total Disagreement</i>	<i>Max. Deviation</i>
500	0.450	0.026
200	0.462	0.041
100	0.466	0.041
50	0.470	0.032
0	0.478	0.041

Table 1:

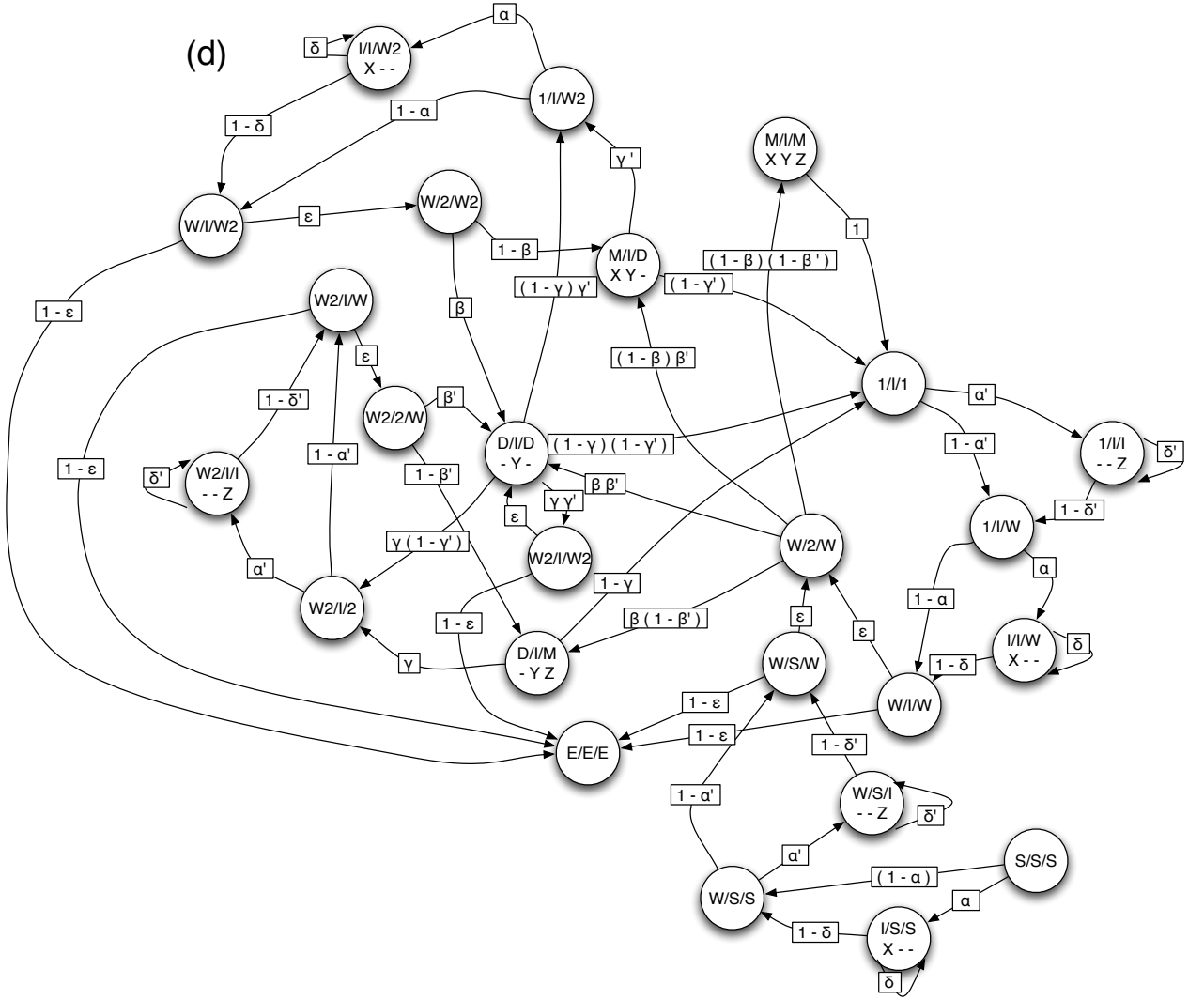


Figure 1:

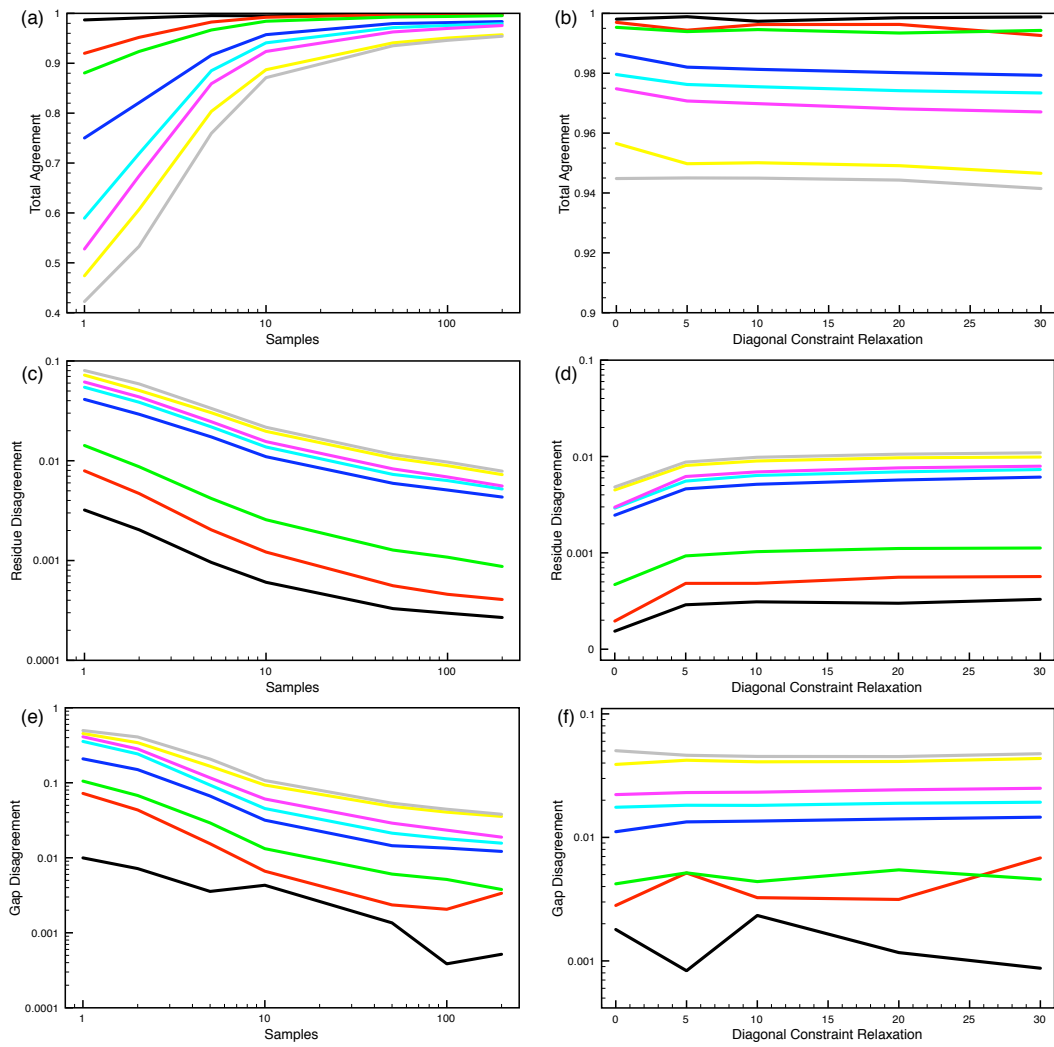


Figure 2:

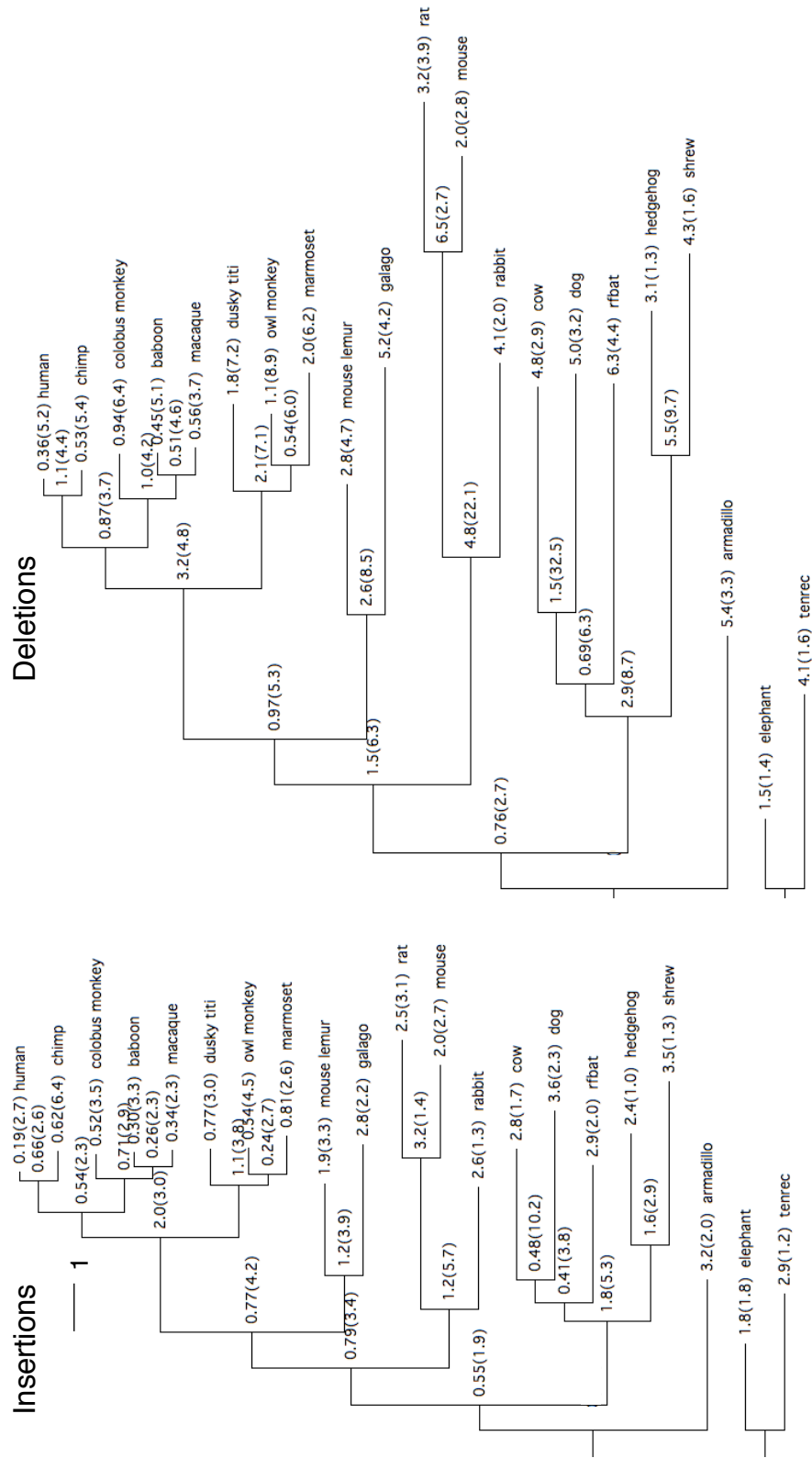


Figure 3: