Low coverage mammalian sequencing
Proposal for set 3

January 2006

Summary

To identify the functional elements in the human genome, previous reports have recommended obtaining low (2x) coverage from 24 mammals. Our ongoing studies continue to indicate that this number of genomes (yielding 4 substitutions per base) should allow detection of conserved 6-mers with a false positive rate of 1 per 10 kb.

Currently, 16 mammals have been approved for sequencing to low coverage (2x) and are in process. Eleven have been sequenced to date, of which eight have been assembled and six have been aligned to the human genome. The data are performing as predicted in terms of assembly quality, human genome coverage and element detection.

With the current data, it should be possible to identify reliably ~50% of the conserved elements in the human genome (based on analysis of the much more deeply covered CFTR region.) This both indicates that the project is on target, and that (as expected) more than 16 mammals will be needed to detect conserved 6mers at a reasonable false positive rate.

We propose sequencing a third set of 8 mammals for sequencing at 2X coverage to reach a level of resolution that should allow delineation of important functional elements, such as transcription factor binding sites.

Background

We see two main plans of attack to identify the functionally constrained elements in the human genome:

1. **Identification of conserved elements by further sequencing of 2x mammals (set 3).** This approach exploits the substitutions that occur over evolutionary time in unconstrained sequence. Previous studies cited in earlier reports have consistently shown that ~24 genomes are necessary to identify conserved 6-mers (e.g., the length of a typical binding site for a transcription factor) with a false positive rate of 1/10,000 bp. To provide further empirical evidence, we have continued to evaluate the emerging 2X data as follows:

    a. **Comparison of aligned coverage of human genome with prediction**. How closely does each genome sequence match the expected ~80% representation of the orthologous sequence in the human genome.

    b. **Use the sequenced 2x genomes to calculate current resolving power for k-mers and compare to theory.** Because mutations occur randomly in neutral DNA, a short sequence element will sometimes appear to be conserved by chance. Using current data, we can calculate the proportion of sites of a given size that would be conserved by chance (false positive rate) and calculate how many genomes are needed to reduce it to an acceptable level. Agreement between theory and practice is critical.

    c. **Confirmation of feature identification using CFTR data for many mammalian species (~4 subs/site).** By comparison of the performance of the currently available 2x data with the much deeper and higher quality ENCODE data, we can test whether the data are following predictions and infer whether we will need 24 mammals.

2. **Identification of the 'core' mammalian genome (common to all mammals) through high-coverage sequencing of 4 selected mammals.** This strategy exploits the larger scale deletions that have occurred in mammalian genomes since divergence from our most recent common ancestor. Studies of the mouse and dog genomes have shown this approach to be complementary to the 2x coverage. The details of this strategy are outlined in the accompanying proposal for generation of high coverage genomes. In the present proposal we describe the background, an update on progress and the proposed third set of mammals for 2X sequencing.

Update

1. Status of 2x genomes

Progress for low coverage mammalian sequencing is on schedule. Sequencing of all eight "Set 1" mammals and three "Set 2" mammals is complete. Assemblies have been generated for eight species (Table 1). Contig and supercontig N50 lengths are similar across the species.

Table 1: Status on 2x assemblies

Species	Genome Size (Gb)	Repeat content (Gb)	Non repetitive content (Gb)	Contig N50 (Kb)	Super N50 (Kb)	Total contig length (Gb)	human genome covered[1] (%)
Aligned							
Elephant	3.1	0.68	2.41	2.8	45.1	2.30	83
Armadillo	2.7	0.61	2.14	2.7	45.7	2.15	65
Rabbit	2.8	0.72	2.09	3.2	54.5	2.08	75
Tenrec	3.1	1.00	2.15	3.1	48.3	2.11	81
Guinea Pig	2.5	0.49	2.02	2.8	48.0	1.95	84
Common Shrew	2.8			3.2	47.8	1.83	73
Assembled							
Hedgehog	3.4			2.8	33.0	2.13	
Cat	2.5[2]			2.2[2]	75.8	1.65	
Sequenced							
Bat							
Tree shrew							
Squirrel							

[1] Coverage of the human genome as compared to coverage with near-finished mouse genome.
[2] The cat assembly is performing slightly different from other genomes. Preliminary data suggests it is due to a larger genome with segmental duplications not detected in the genome size estimate.

2. Analysis of Alignments

Preliminary analysis of the six aligned genomes, show the following results:

A. How does the aligned coverage of the human genome compare to prediction?
The coverage of the human genome by aligned sequence was calculated by comparing to the coverage by near-finished mouse sequence. As seen in table 1, that mean coverage is ~77%, which is very close to the expectation of ~80%.

B: Use the sequenced 2x genomes to calculate current k-mer resolving power and compare to theory. In the proposal for a second set of 2x mammals we showed that the false positive rate for the first three 2x mammalian genomes followed the trend predicted

by the Eddy model (Eddy, PloS, 2005). The data from six 2x mammals continue to follow the expected trend. These results argue that continued sequencing of a third set of eight additional mammals to 2x coverage will bring the false positive rate down to the eventual target false positive rate of 1 in 10 kb for sensitive and specific resolution of small functional elements such as transcription factor binding sites.

C: Confirmation of feature identification using many mammalian species (~4 subs/site). Roughly 5% of the human genome appears to be under purifying selection (MGSC, Nature, 2002 Lindblad-Toh, Nature, 2005, Cooper, Genome Res, 2005).

In an analysis of 29 mammals with a total branch length ~4subst/site (including 3 marsupials) in ENCODE region 1 (CFTR), Cooper identified conserved elements covering ~5.5% of the region using the GERP method.

We therefore performed GERP analysis on five ENCODE regions using the four high-coverage genomes of human, mouse, dog, cow (HMDC) and the six 2x mammals currently aligned to the human genome. With these genomes, we identified conserved elements covering ~2.8% of the regions (Table 2).

Of the elements identified, 45% were coding, with 81% of exons identified. The median size of detected elements was 31 bp. By contrast, the median size in the Cooper analysis was 12 bp. This confirms that the current analysis is still missing many conserved elements, due to insufficient depth of the evolutionary tree.

Table 2: GERP elements identified using HMDC + six 2x mammals

ENCODE region	Length (bp)	% region conserved	% exons overlapped	% coding elements	Minimum element length	Median element length
ENm001	1,700,000	3.2	93	0.28	12	29
ENm002	1,000,000	3.9	85	0.55	11	30
ENm003	500,000	2.8	86	0.42	13	31
ENm004	1,700,000	2.1	83	0.50	10	30
ENm005	1,700,000	3.3	79	0.42	11	28
ENm006	1,300,000	1.9	63	0.53	16	36
	Mean	2.9	81.5	0.45	12.2	30.7

To estimate the value of additional sets of 2x mammals, we performed GERP on the ENm001 region for various sets of mammals:

(A) 1.7 subst/site, reflecting current dataset;
(B) 2.2 subst/site;
(C) 2.7 subst/site (given the collection of mammals available in the Cooper dataset, this is the closest scenario to effective branch-length seen with set 1 and 2 low coverage mammals);
(D) ~4 subst/site (full Cooper dataset)

With **C** we were only able to detect 4.0% of the genome as conserved (Table 3), suggesting that with set 1 and 2 we are still missing conserved elements. By contrast, the increased branch length in set D allows detection of considerably more conserved elements than with set1+2 mammals alone.

Table 3: GERP elements identified with increasing numbers of species

Mammal set	Effective branch length (Subst/site)	% region conserved	% exons identified	% coding elements
A	2.0	3.2	93	0.28
B	2.2	3.2	92	0.29
C*	2.7	4.0	98	0.29
D	4.0	5.5	98	0.12

*Branch length close to the currently approved sixteen 2x mammals (set 1+2)

Proposal: An additional set of mammals

We propose a set of ten potential set 3 mammals (Table 4 and 5), from which eight would be used. These organisms were chosen for their long branch length, combined with representation across the evolutionary tree (Figure 1). The final choices will be made based on availability, promise as biomedical models and genome features.

Table 4: Potential mammals for set 3

Species	Added Branch length	Attributes
Elephant shrew	0.22	Long branch; Afrotherian
Flying lemur	0.11	Outgroup to primates
Dolphin	0.12	Marine mammal
Horse	0.11	perissodactyl
Llama	0.11	Basal artiodactyl
Mole	0.15	Long branch
Mouse lemur	0.12	Neurobiological model
Pika	0.19	Long branch
Kangaroo rat	0.27	Long branch
Tarsier	0.18	Long branch

Note that the average branch length in this set is 0.16 substitutions/site as compared to 0.20 in set 1 and 0.15 in set 2.

Figure 1: Phylogenetic tree for Sets 1 (red), set 2 (blue) and set 3 (purple)

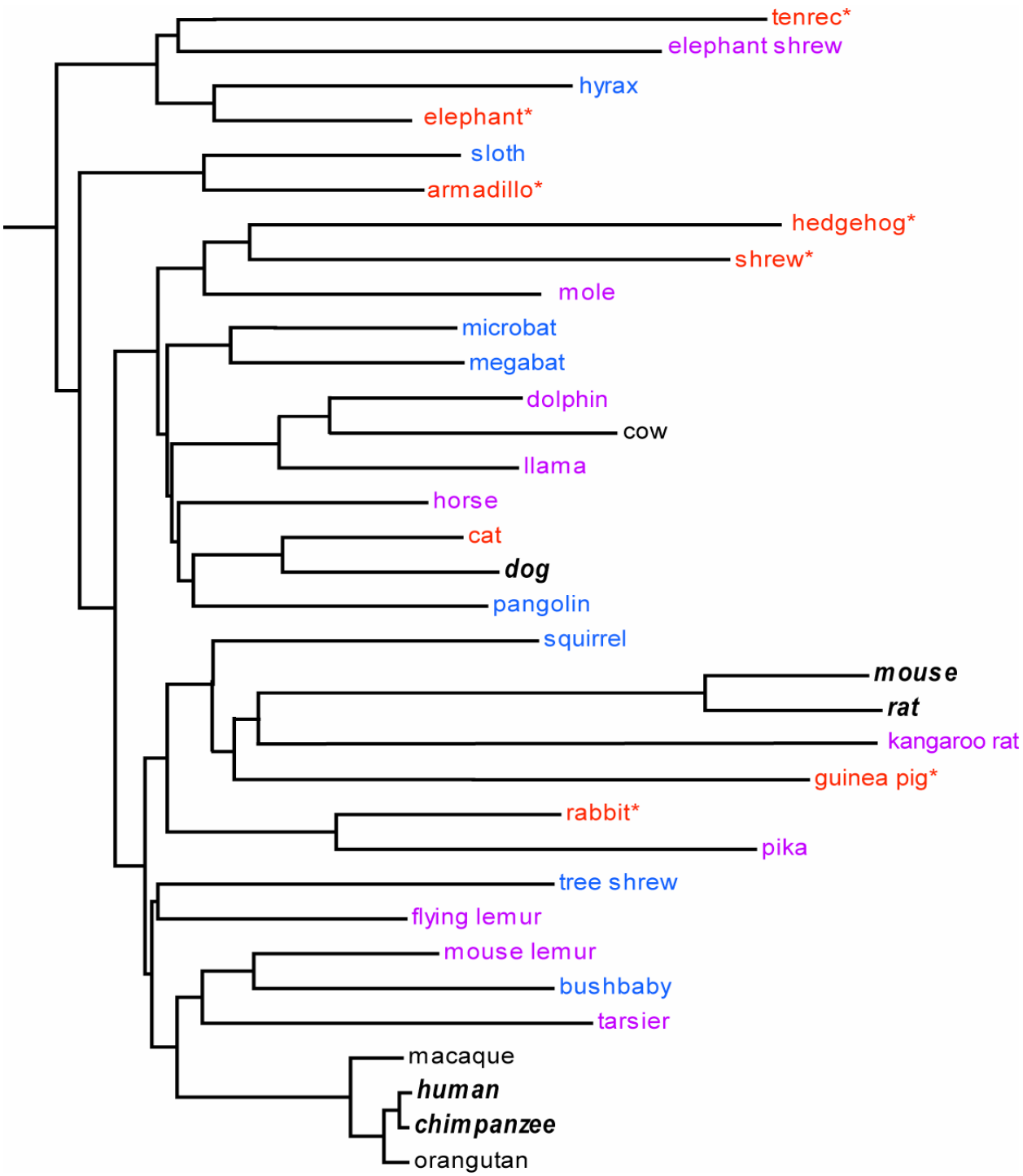

Table 5: Summary of branch length and diverge contribution for sets 1-3

3Species	Distance to human	Divergence Added subst/site	Total divergence*
Finished / in progress:			
Human	0	0	0
Mouse	0.450	0.450	0.45
Rat	0.456	0.080	0.53
Chimpanzee	0.009	0.004	0.53
Dog	0.309	0.188	0.72
Macaque	0.051	0.024	0.75
Opossum	0.946	0.812	
Cow	0.363	0.203	0.95
Orangutan	0.013	0.011	0.96
Set 1:			
Elephant	0.323	0.161	1.12
Armadillo	0.307	0.156	1.28
Rabbit	0.310	0.179	1.46
Tenrec	0.484	0.278	1.73
Guinea pig	0.423	0.262	2.00
Shrew	0.414	0.260	2.26
Cat	0.292	0.082	2.34
Hedgehog	0.438	0.242	2.58
Set 2:			
Microbat	0.290	0.131	2.71
Squirrel	0.300	0.148	2.86
Tree shrew	0.301	0.183	3.04
Bushbaby	0.278	0.137	3.18
Megabat	0.294	0.107	3.29
Hyrax	0.396	0.163	3.45
Sloth	0.324	0.116	3.57
Pangolin	0.305	0.134	3.70
Set 3:			
Pika	0.399	0.191	3.89
Kangaroo rat	0.403	0.270	4.16
Llama	0.319	0.109	4.27
Horse	0.277	0.113	4.38
Mole	0.327	0.153	4.54
Tarsier	0.296	0.178	4.71
Flying lemur	0.235	0.113	4.83
Lemur	0.225	0.119	4.95*
Elephant shrew	0.435	0.220	
Dolphin	0.316	0.082	

Tree is calibrated such that human/mouse divergence = 0.450

*Note that the total divergence for 2x mammals, is only ~80% of that listed number based on coverage, bringing the total effective branch length to about 4 substitutions per site.