Progress report on 2x mammalian sequencing

Annotating the Human Genome Working Group
April 28, 2005

Summary

Sequencing and assembly of set 1 of the 2x mammals is on schedule and yielding results very similar to those predicted:

• The genome assemblies have better than the originally predicted continuity, with an N50 supercontig size of ~50 kb for the assisted assembly.

• The 2X assemblies provide coverage of the human genome that is ~70% of that obtained with deep coverage assemblies. This is ~100% of the predicted value based on simulations.

• The accuracy of distinguishing conserved k-mers from background is improving as predicted with each additional genome. Upon the completion of set 1, we expect the false positive rate for detection of a perfect 6-mer to be ~600 per 10kb, consistent with theoretical predictions by Eddy.

Accurate recognition of sequence signals such as transcription factor binding sites (6-mers) will require decreasing the false positive rate to 1 per 10 kb, as reported by Eddy (Eddy, 2005).

Given the close agreement with theoretical predictions and the need to reduce the false positive rate to the target level, the Working Group proposes continuing with the project to annotate the human genome by undertaking sequencing of the mammals in set 2. This should bring the false positive rate for signals such as transcription factor binding sites to ~50 per 10kb. Once this is achieved, we will need to assess the role that additional comparative sequencing can play in identifying the bulk of functional elements in the human genome. If theory holds, adding a third set of low coverage mammalian genomes would reduce the false positive rate to ~10 per 10 kb.

1. Sequencing and assembly status for set 1

Progress for mammalian sequencing at low coverage is proceeding on schedule. We have completed 2X sequencing for five mammalian species and expect sequencing of all "set 1" mammals to conclude by August (Table 1).

Table 1: "Set 1" mammals: sequencing status

Species	Center	Genome (Gb)	Heterozygosity	Sequencing	Status
Elephant	Broad	3.1	1/9,000	complete	assembly complete
Armadillo	Broad	2.8	1/2,600	complete	assembly complete
Rabbit	Broad	2.8	inbred	complete	assembly complete
Tenrec	Broad	3.2	1/850	complete	assembly in progress
Guinea pig	Broad	2.5	inbred	complete	sequencing complete
Cat	Agencourt	2.6	1/3,000	Jan - June	sequencing in progress
Shrew	Broad		1/1,000	May - July	ready for sequencing
Hedgehog	Broad			July - Aug	heterozygosity testing

Low coverage genome assemblies have been generated for elephant, armadillo and rabbit, and have the expected properties in terms of contiguity (Table 2). The inbred rabbit has the largest N50 contig and supercontig sizes among the unassisted assemblies, presumably a consequence of the absence of polymorphism. Assisted assemblies using read alignment to human and dog to validate single links have also been performed. The assisted step doubles or triples the N50 supercontig size and leads to incorporation of ~10% extra reads.

Table 2: Low coverage genome assembly statistics

	Rabbit *Oryctolagus cuniculus*		**Elephant** *Loxodonta africana*		**Armadillo** *Dasypus novemcinctus*	
Distance to human (subst/site)	0.310		0.323		0.307	
Genome size (Gb)	2.81		3.09		2.75	
Heterozygosity rate (1/x bp)	inbred		1/9,000		1/2,600	
Sequencing coverage (Q20)	1.95x		1.94x		1.97x	
Assembly type:	unassisted	assisted	unassisted	assisted	unassisted	assisted
Reads assembled	76%	84%	69%	80%	71%	81%
Total contig length (Gb)	1.87	2.08	1.92	2.15	2.04	2.30
Contig N50 size	3.2	3.2	2.5	2.7	2.7	2.8
Supercontig N50 (kb)	20.4	54.5	10.2	45.7	15.5	45.1
Total contig length/genome size	67%	74%	70%	78%	66%	74%
Reads aligning uniquely:						
On human	53%		51%		46%	
On dog	49%		46%		49%	
On dog or human	65%		60%		59%	
Pairs on human	23%		23%		22%	
Pairs on dog	19%		22%		20%	
Pairs on human or dog	29%		29%		27%	

2. Assessment of genome assemblies

We have evaluated the 2X assemblies for two different properties:

1) Coverage of the human genome compared to the maximum possible coverage, and
2) Power to distinguish selected from neutral bases as a function of the number of sequenced genomes.

2.1 Coverage assessment

We evaluated the 2x assemblies to ensure they provide the expected coverage of the human genome. Our previous simulations with mouse showed that the number of human bases that could be covered by a 2X assembly is 70% of the number of bases that could be covered with deep shotgun sequence. We performed this analysis for armadillo and rabbit (the elephant assembly was not available at the time), using our actual 2X assemblies and the deep coverage available in the CFTR region. The corresponding portions are 70% for armadillo and 67% for rabbit (Figure 1). Thus, both organisms are performing exactly as predicted.

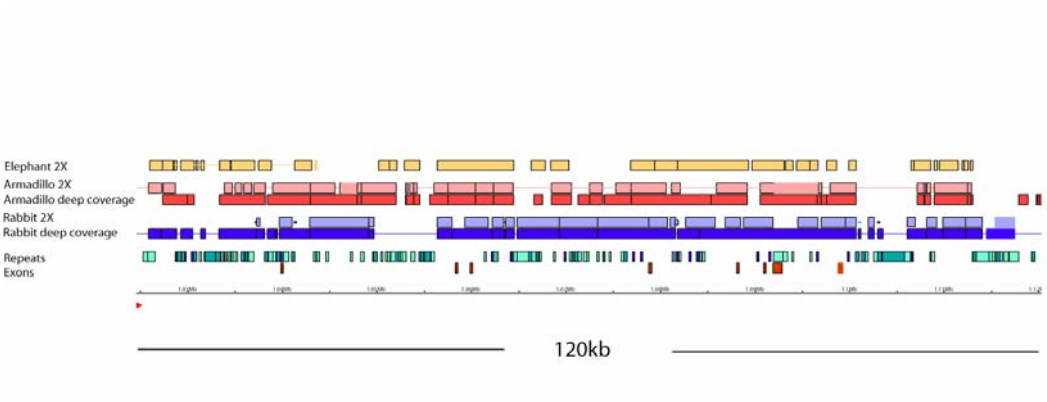


Figure 1. Comparison of coverage provided by 2x assemblies and comparative grade sequence across 120kb from the CFTR region. The 2x assemblies cover ~70% of alignable bases as expected. No comparative grade elephant sequence is available in this region.

To further investigate how well the 2X assemblies covered the human genome, we looked at the proportion of the human genome covered by the armadillo 2X assembly in the 44 ENCODE regions. Because we cannot compare the results to deep shotgun sequence from armadillo, we instead compare them to the coverage of the regions by the deep shotgun mouse sequence. (The rationale behind this is that the number of orthologous bases for each mammal will be correlated within one region but will vary widely between regions.) Results for these 44 regions are very encouraging; the inter-quartile range varied between 72% and 110% of mouse bases with a median value of 83% (Figure 1).

The median value (coverage by 2X armadillo/coverage of deep shotgun mouse) is higher than the expected value of 70% (for coverage by 2X armadillo/coverage of deep shotgun armadillo) because the armadillo is closer to human than is the mouse.

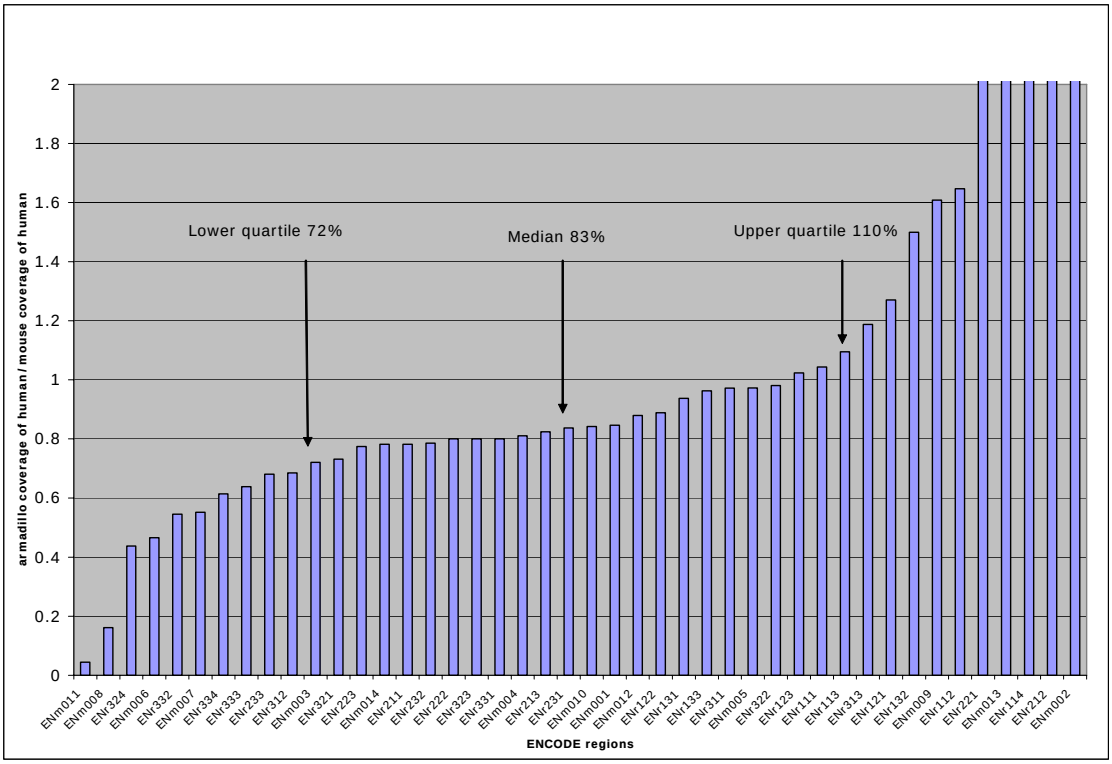


Figure 2. Fraction of armadillo coverage of human versus mouse coverage of human in the ENCODE regions. Note that five regions lack mouse coverage and two regions have very low armadillo coverage**.**

2.2 Assessment of detection of conserved k-mers

Our second assessment concentrated on how extra genomes increase the resolution of conserved k-mers in the genome. One important class of conserved elements is expected to be transcription factor binding sites. With this in mind, we decided to focus on 6-mers, as this is close to the typical transcription factor binding site size.

We based our analysis on an analytical approach from Eddy (Eddy, 2005). This approach is based on the notion that most k-mers ('random k-mers') drift at a neutral rate, but some k-mers ('functional k-mers') change at a much lower rate. As one adds genomes, the total set of 'conserved k-mers' contains a decreasing proportion of random k-mers and an increasing proportion of functional k-mers. Eddy studies the 'false positive rate', which he defines as the proportion of random k-mers among the conserved k-mers.

We used the Eddy method to calculate a theoretical curve of how the false positive rate decreases with the addition of extra genomes. Eddy proposes that a reasonable target is a false positive rate of 1/10,000 bases – that is, about one falsely predicted binding site per 10 kb, or, equivalently, ~10 per 100kb genomic locus. In agreement with his paper, the results suggest that ~25 genomes are needed with an average distance comparable to that of the dog.

With real data, we can directly estimate the empirical false positive rate as the conservation rate for k-mers in ancient repeats (assumed to be all random k-mers).

We first compared Eddy's theoretical approximations to the empirical false positive rates for deep coverage sequence data from the CFTR region, to see how the theory stood up to reality (Figure 3, dark green line). The false positive rates agreed to within an average of 17% for up to

11 extra genomes. This confirms that our treatment of the Eddy model is a reasonable approximation.

Encouraged by this, we calculated the empirical false positive rate for all currently available deep coverage mammalian assemblies (mouse, rat, dog), and the three available 2X assemblies (elephant, armadillo, rabbit) (Figure 3, light blue line). The results agree reasonably well with the model.

Importantly, the results show that the empirical data is tracking well with the theory. It indicates that we are on track for attaining ~600 false positives per 10 kb with 8 mammals at 2X coverage and ~50 false positives per 10 kb with 16 mammals at 2X coverage.

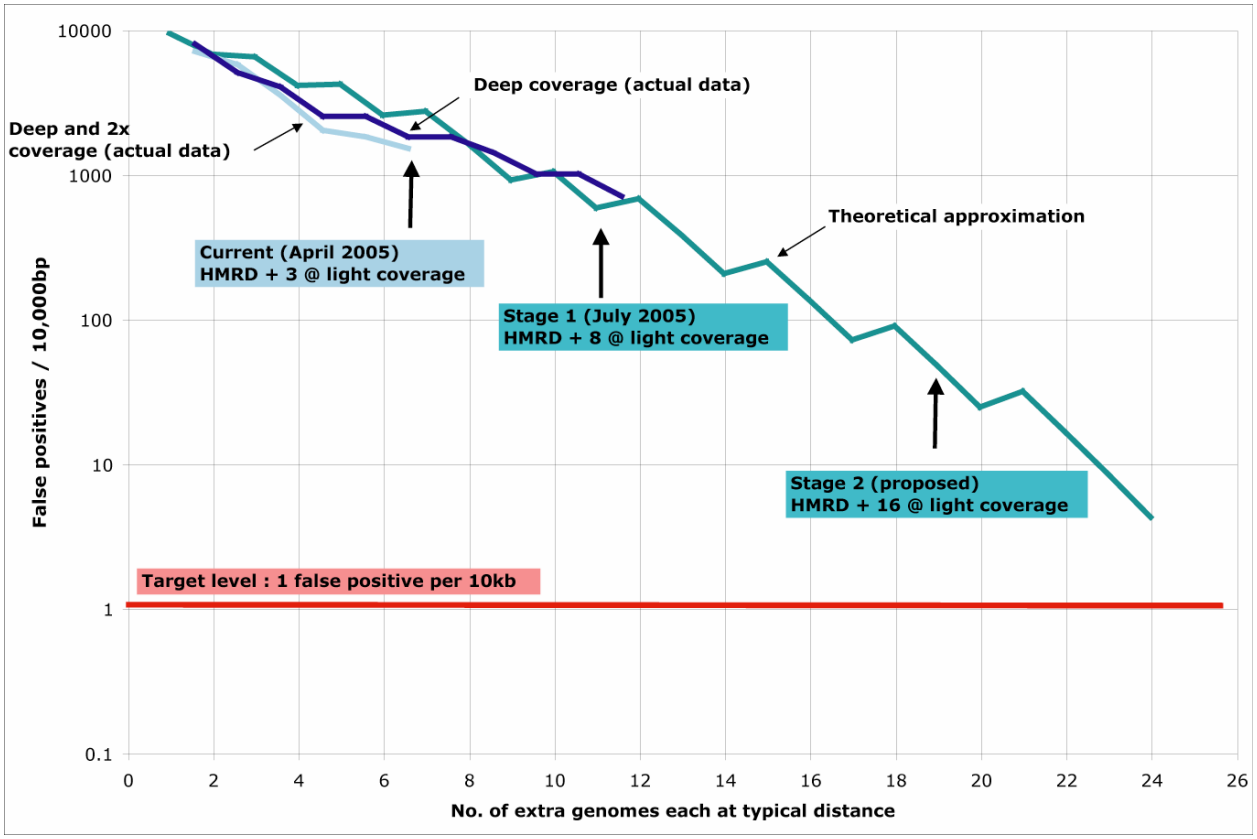


Figure 3. The false positive rate for 6-mers as a function of extra aligned genomes. The theoretical approximation (turquoise) is taken from (Eddy, 2005), and assumes that the substitution rate in functional k-mers is 20% of the neutral rate. Each extra genome is assumed to be at a typical distance of 0.19 substitutions per site. The deep coverage curve (dark blue) is calculated using alignments of ancient repeats in the CFTR region. The deep and 2X coverage curve is calculated in a similar fashion using deep coverage of mouse, rat and dog, and then 2X of elephant, armadillo and rabbit. (HMRD = human, mouse, rat, dog)

3. Progress on collection of Set 2

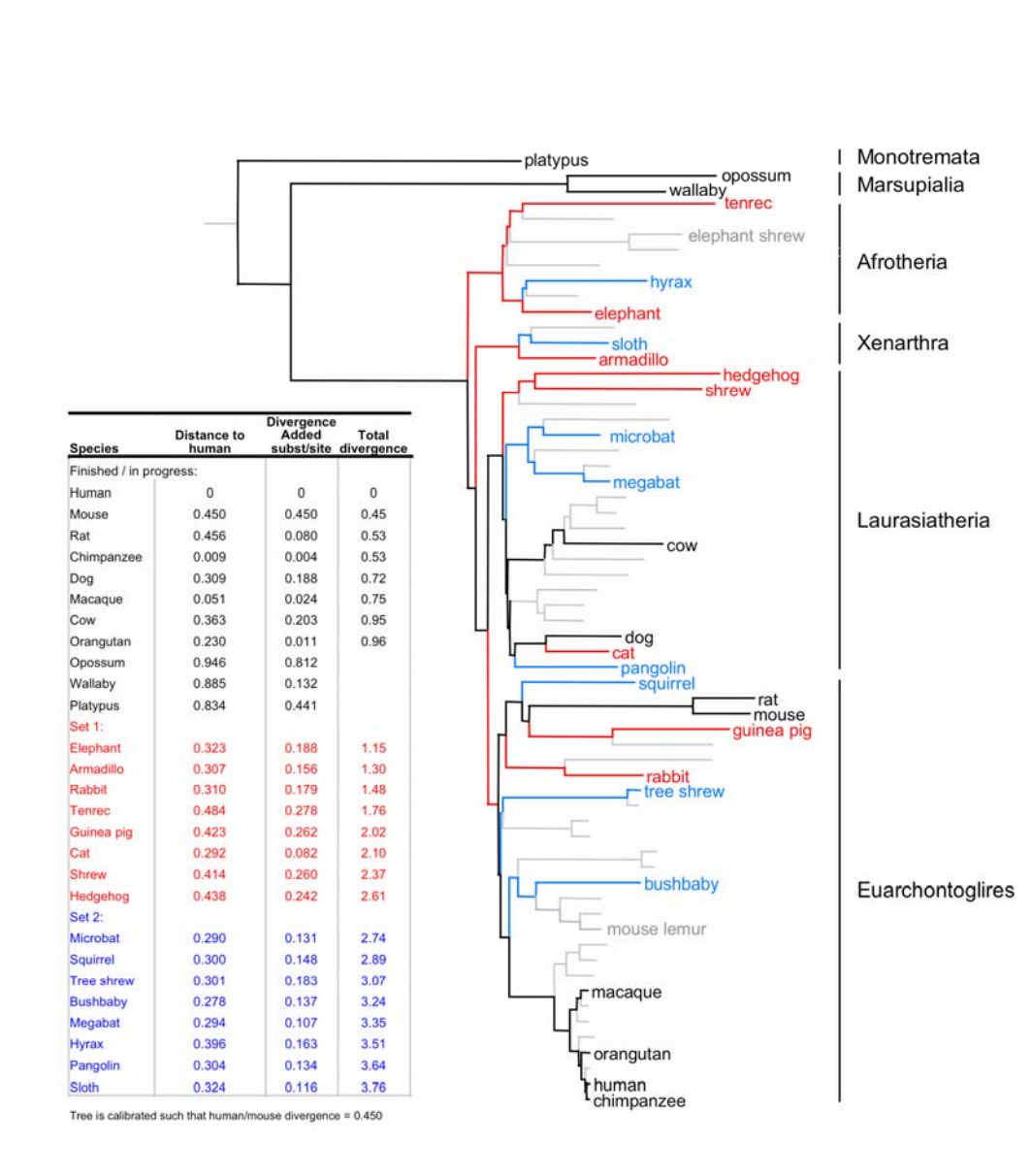
Five mammals in Set 2 are in various stages of heterozygosity assessment and will be ready for sequencing in the fall. Tentative sources have been located for the remaining three Set 2 mammalian species (Table 3).

Table 3: Set 2 mammals: sample collection

Species	Status
Microbat	heterozygosity: 1/800
Squirrel	heterozygosity testing
Bushbaby*	assessing tissue quality
Tree shrew	assessing tissue quality
Sloth*	locating tissue source
Megabat	locating tissue source
Hyrax **	locating tissue source
Pangolin	locating tissue source

* Bushbaby or sloth might be replaced by mouse lemur should issues with tissue source, DNA quality or heterozygosity rate warrant it

** Hyrax could be replaced by elephant shrew should issues with tissue source, DNA quality or heterozygosity rate warrant it



Species	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
Cow	0.363	0.203	0.95
Orangutan	0.230	0.011	0.96
Opossum	0.946	0.812	
Wallaby	0.885	0.132	
Platypus	0.834	0.441	
Set 1:			
Elephant	0.323	0.188	1.15
Armadillo	0.307	0.156	1.30
Rabbit	0.310	0.179	1.48
Tenrec	0.484	0.278	1.76
Guinea pig	0.423	0.262	2.02
Cat	0.292	0.082	2.10
Shrew	0.414	0.260	2.37
Hedgehog	0.438	0.242	2.61
Set 2:			
Microbat	0.290	0.131	2.74
Squirrel	0.300	0.148	2.89
Tree shrew	0.301	0.183	3.07
Bushbaby	0.278	0.137	3.24
Megabat	0.294	0.107	3.35
Hyrax	0.396	0.163	3.51
Pangolin	0.304	0.134	3.64
Sloth	0.324	0.116	3.76

Tree is calibrated such that human/mouse divergence = 0.450

Figure 4. Phylogenetic relationships of mammalian genomes. The genomes sequenced in set 1 (red) are shown, along with those proposed for set 2 (blue), alternates (grey) and genomes sequenced under other programs (black). The inset provides details on the branch length provided by each species.

Reference

Eddy, S. R., 2005, A model of the statistical power of comparative genome sequence analysis, *PLoS Biol* **3**(1)**:**e10.