A proposal to top up the remainder of the 2x mammals to high-quality draft standard

Michele Clamp, Federica diPalma, Kerstin Lindblad-Toh, Eric Lander.
The Broad Institute.
August 2008

Summary:

The initial proposal to annotate the human genome using many mammals showed that 26 species at full coverage would resolve all 6mers under constraint, whereas 2x coverage genomes where only 86% of the genome is captured would resolve ~10 bp windows. Twenty-two 2x mammals have now been sequenced and a complete analysis is ongoing. The remaining two genomes (pangolin and flying lemur) have suffered from DNA quality problems and are therefore not complete at this point.

In an intermediate analysis of a 22-way mammalian alignment (six full coverage and sixteen 2x coverage) supports our prior calculations that 5% of the genome is under selection, but only roughly half of these 5% can be resolved at present. We are now able to use 12-bp windows as being under contraint, which is a huge improvement over previous mouse-human comparisons (50 bases).

In parallel with the 2x mammals project, eleven of these species (horse, cat, armadillo, elephant, rabbit, guinea pig, bushbaby, mouse lemur, tarsier, tree shrew and microbat) have been approved for high coverage sequencing based on their biomedical relevance and sequencing is currently ongoing.

With the emerging new sequencing technologies, sequencing of large genomes will get several orders of magnitude cheaper allowing us the luxury of contemplating sequencing many mammals at high coverage for a considerable lower cost. Thus, based on the success of the 2x project, we have evaluated the need for topping up the remaining 2x genomes using new sequencing technology. We believe that it would be very important to be able to resolve transcription factors biding sites (ie 6-mers). Furthermore, certain gene and motif analysis is currently hindered by the high number of sequencing errors and indels present in low coverage data. We therefore propose topping-up the additional eleven 2x mammals to high coverage using new sequencing technology.

Background

24mammals project

To identify the functional elements in the human genome, previous proposals have recommended obtaining low (2x) coverage from 24 mammals. This recommendation initiated by the discovery that 5% of the human genome appears to be under selected constraint based on the comparison of human, mouse and dog and that these regions appear to largely be similar across mammals. The studies indicate that this number of genomes (providing 4 substitutions per base) should allow detection of conserved 6-mers with a false positive rate of 1 per 10 kb using full coverage genomes. With 2x coverage genomes, where only 86% of the genome is captured, 10-br features under constraint can be detected. For these calculations, we had available rough estimates of how constrained the regions would be and we assumed all would be under the same constraint. Obviously it's easier to detect the highly constrained regions and so we may be unable to find the more weakly constraint elements.

Twenty-two of the selected twenty-four 2x genomes have now been sequenced and a multiple alignment including all species is expected in September 2008. Two species will not be included in the current analysis, since the sequencing has been delayed due to DNA problems; pangolin and flying lemur. A 2x—mammals international consortium is working on the analysis of the 2x genomes and expect to be able to improve the gene and conserved element annotation in the human genome during the fall of 2008.

An intermediate analysis of a 22-way mammalian alignment (six full coverage and sixteen 2x coverage) supports the prior calculations that 5% of the genome is under selection and that about half of the selected portion can be resolved. We have now been able to use 12-bp windows for the analysis, which is a huge improvement over previous mouse-human comparisons. Our sensitivity is greatly increased and also we can resolve smaller regions of genome (12 bases compared to 50 bases) as being under constraint.

In 2006 a decision was made to top up eleven on the 2x mammals based on their biomedical relevance or important positions in the mammalian tree. These projects are currently ongoing with only horse and guinea pig completed at this point. Some species are being sequenced by ABI 3730s and others by new sequencing technology.

The current data is performing according to theoretical predictions

Although we don't have the full set of genomes yet we can assess how our constraint detection method performs compared to theory. The Eddy method (PLoS 2005) predicts that for 22 mammals we should be able to resolve most of the constrained **8mers**. Our analysis however shows that we can only resolve **12mers** with 22 genomes. This is entirely consistent however if we take into account two assumptions made in the theoretical prediction.

Our theoretical calculations were made with full coverage genomes. Sequencing to only 2x results in the effective genome number being about 75% of the total (as we only have 75% genome coverage). If we recalculate the Eddy predictions using the corrected genome number then we are perfectly on track. Using an effective genome number of 15 as opposed to 22 then we expect to resolve a 12mer which is currently the case (Table 1)

Secondly we made an assumption that all constrained elements would be under the same constraint as coding exons (5 times more constrained than neutral). In practice this is not the case and the majority of non-coding elements are more weakly constrained than this (90% of the most highly constrained regions are coding). The result is that although the theory predicts we could resolve 99% of constraint we can only manage 50%. If we wish to detect the weaker elements we would need more genomes or full coverage assemblies of existing genomes.

No. aligned genomes	Predicted kmer resolution (6X genome coverage)	Effective no. genomes for 2X coverage	Predicted kmer resolution (2X genome coverage)	Actual kmer resolution based on whole genome alignments.
22	8	15	**12**	**12**
26	6-7	20	10	-
29	5-6	21	6	-

Table 1. The smallest theoretically resolvable constrained kmer for different numbers of genomes : kmers for full coverage assemblies (column 2) and 2X coverage assemblies (column 4). Column 5 shows that for the currently available 22 genomes the data does indeed match the theory.

How is the current 5% different from the previous 5%?

A first glance at the amount of genome under selection seems to say that adding extra genomes has not had any effect. The previous estimate (using human, mouse, rat and dog) for the fraction of genome under selection was 5% and with 22 genomes it is still 5%. So what is the difference? A large difference is the increased resolution. The initial estimate was made using 50bp windows and only a small fraction of them could be reliably distinguished from neutral sequence. The extra genomes give us the ability to separate ~50% of the constrained regions and pinpoint them to within 12bp.

Figure 1a shows that for human,mouse,rat dog and 50bp windows the constrained portion (marked c50) is barely distinguishable from the bulk neutral sequence. In stark contrast however with 22mammals and 12bp windows (Figure 1b) the constrained sequence is clearly distinguishable from neutral (marked c12).

In addition to increased resolution of already identified constrained regions many novel non-coding elements have been revealed including 800,000 eutherian specific elements. We predict that more genomes will reveal even more innovation in mammalian constraint.

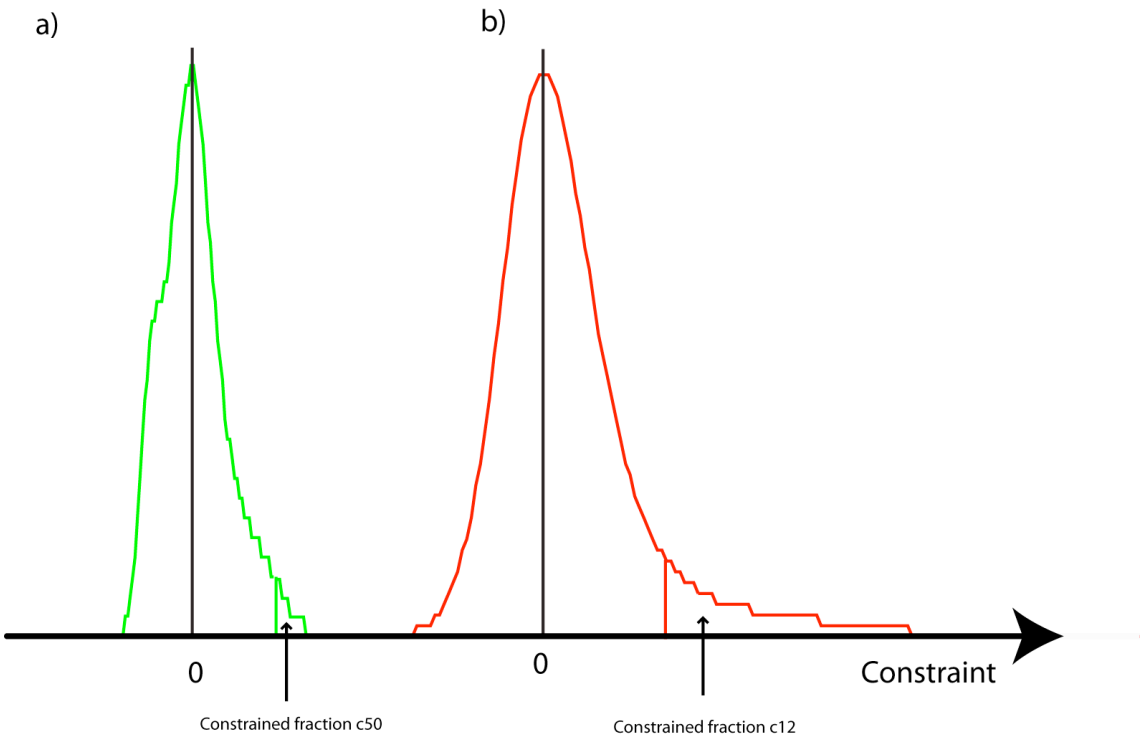

Figure 1. Constraint distributions for a) 4 species over a 50mer and b) 22 species over a 12mer. The resolvable constrained portions (5% FDR) are marked by arrows. It can easily be seen that 22 mammals enables much more of the constraint to be detected.

Resolving all constrained 6mers – we propose topping up the remaining 2X mammals

To achieve the goal of finding all constrained 6mers (a typical transcription factor binding site length) or to determine the constraint on each base in the human genome we need extra branch length. We can achieve the former either by sequencing additional 2x mammals or by 'topping up' the existing 2X mammals so they have full coverage. The 'topping up' option is attractive as it has the additional benefit of providing longer assembled contigs. This allows greater accuracy in orthologous alignment to human and also better data for the analysis of rearrangements in other mammals.

A second important benefit of topping up the 2X genomes is the increased accuracy of the assembly at the base pair level. Analysis of coding exons has revealed that, although wrong base pair calls are relatively rare, the number of indels is greatly increased compared to full coverage assemblies. For example the reading frame conservation (RFC) score detects 99% of exons at a threshold of 90 (Clamp et al 2007 PNAS). If we then measure RFC for a 2X genome this proportion plummets to only 80%. This effect is purely due to spurious indels in the assembly.

We therefore propose starting with topping up the 2x mammals to achieve the 6-bp constraint target and to later move on to additional species if we want to detect constraint at a single base level.

Proposal

The current proposal is therefore to generate draft genomes of the eleven remaining 2x genomes shown in Figure 2 in red.

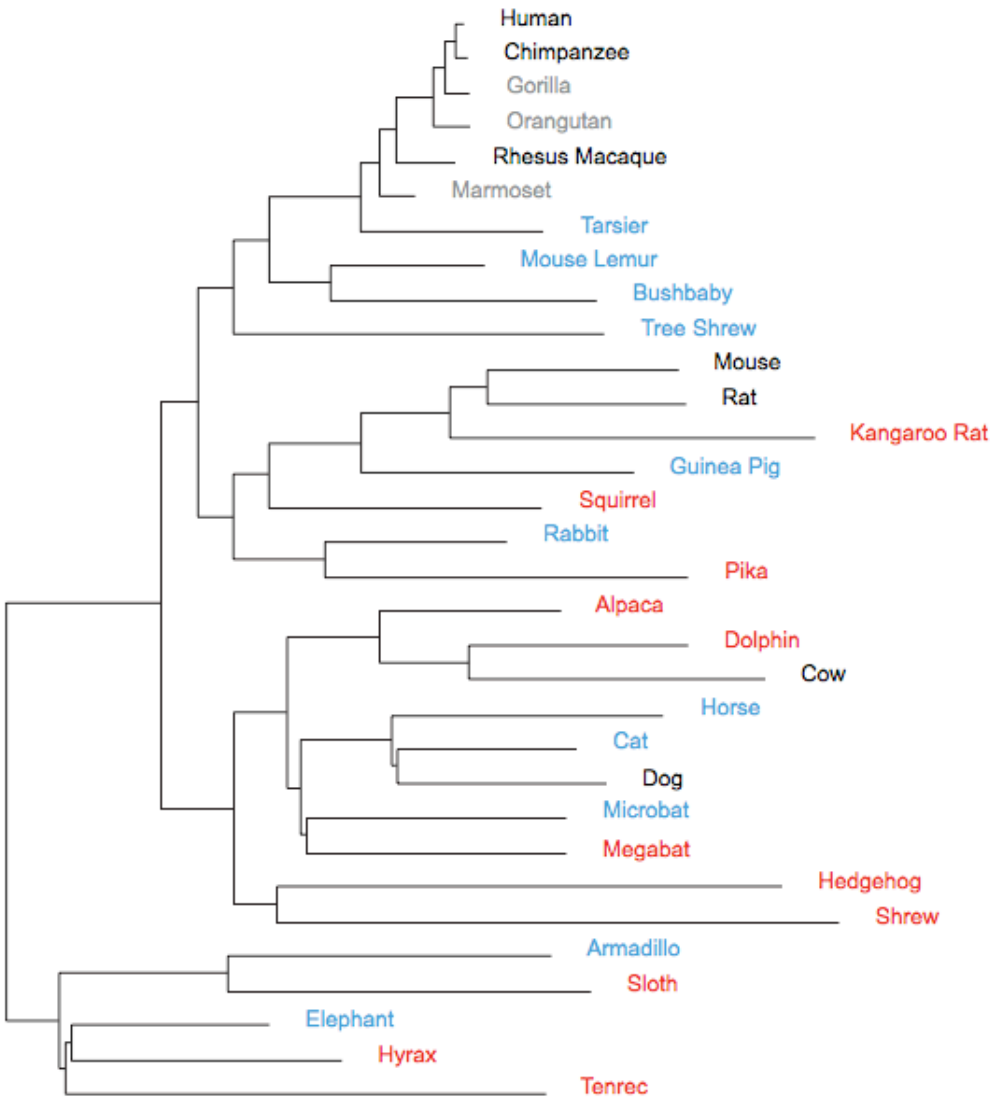

Figure 1: Eleven 2x genomes to be topped up to draft quality using sequencing technology (red). Also shown 2x genomes already being topped up (blue) and other complete genomes (black) and ongoing projects (grey).

For most of the eleven species, DNA from the same individual still exists. For others a closely related individual will be chosen if available.

Sequencing technology

To make the best use of the existing reads and the new sequencing technology ideally the 2x Sanger reads would be topped up with Solexa and/or 454 reads. It is currently unknown how well a Sanger/Solexa hybrid assembly algorithm will work, but currently 454/solexa hybrid assembly is under development for several vertebrate and mammalian projects at several centers. Both chicken and stickleback are being sequenced with new technology and can be directly compared to high coverage Sanger assemblies. The first results are expected in the fall of 2008. [WU is this true for chicken?]

Before deciding on the exact strategy for these genomes we propose examining the quality from several new technology vertebrate genomes to be able to choose the optimal data types and combinations as well as assembly methods. The goal of this analysis is to examine contiguity, base quality and structural quality of the assembly and to ensure that indels are at a frequency comparable to in a high coverage draft.

Conclusion

We believe that topping up the remaining eleven 2x genomes is a cost effective way to achieve increased resolution to the annotation of the human genome, namely reaching the level of individual transcription factor binding sites.