Genomic Imprint of the Interactome: Universality of Genome Evolution

Jonathan Miller[1]
1. Human Genome Sequencing Center, Baylor College of Medicine, Houston, USA
*email: jnthnmllr@gmail.com

Because genomes encode the components of interacting networks of proteins and nucleic acid sequences, including both DNA and RNA molecules, it is well-appreciated that these interactions are reflected in the *local* structure of genomic sequence. Developments in recent years are converging on a more unexpected conclusion: that these networks have driven the evolution of *global* and *universal* genomic architectures.

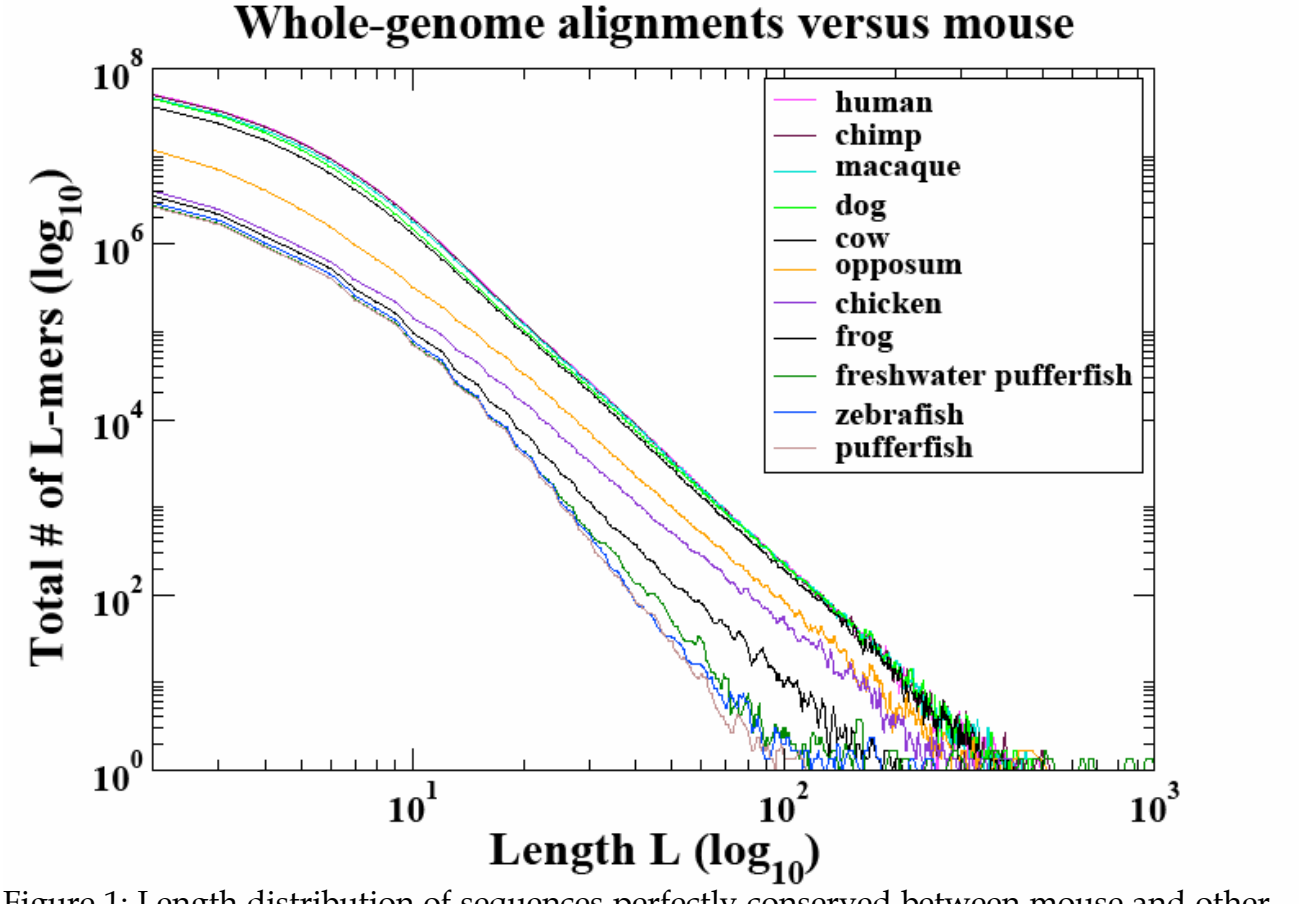

Figure 1: Length distribution of sequences perfectly conserved between mouse and other vertebrates from pairwise whole-genome alignments. Repeat-masked whole genome alignments were obtained from UCSC; all contiguous runs of identical sequence (aligned with no substitutions, insertions, or deletions) were identified, and their lengths plotted on a log-log plot. For human/mouse/rat ternary alignments (not shown), **the sequences above 200 bases in length correspond exactly to the so-called "ultra-conserved" sequences**, and exhibit a comparable algebraic distribution over a comparable range of scales. Observe that the "ultra-conserved sequences" [1] don't stand out, suggesting that they are no more strongly conserved than much of the rest of the genome [2].

These developments themselves have been driven by the new possibilities that whole-genome sequences have opened for the classical field of comparative genomics. The 2002 human-mouse whole-genome comparison revealed on the order of three times as much strongly-conserved sequence as would have been expected based on conservation of protein-coding genes alone. The discovery of "*ultra*conservation" [1] in 2004 heralded the application of model-independent inference of functional sequence on genome-wide scales, by revealing the longest contiguous runs of sequence common to multiple diverse organisms. No modeling - indeed only minimal argument based on the evolutionary distances among selected vertebrates averaged over their entire genomes - was needed to justify the inference that "ultra-conserved" sequences, defined approximately as those exceeding 200 bases in length, were likely to be biologically functional, although assignment of any specific functional mechanisms could at the time be only speculative.

In 2006, these "ultra-conserved" sequences were demonstrated to yield enhancer activities for neural differentiation in mouse embryos [3] at validation rates close to 50%. How might their role as enhancers have led to conservation that appears so extraordinary, genome-wide? One possible answer to this question has been proposed by Mattick [4], who argues that *some* enhancers consist of multiple and overlapping transcription-factor [TF] binding sites. Thus, the effect of a mutation on the binding of a single TF is multiplicatively increased because it affects the binding of multiple TFs encoded elsewhere in the genome. To compensate the mutation in the enhancer without adverse impact on fitness, either the enhancer would have to be duplicated, or all the TFs would simultaneously have to be mutated – a wildly improbable confluence of events.

Thus, this explanation suggests that "ultraconservation" directly reflects non-local interaction of multiple DNA and/or gene products – interactions that also must be strongly-conserved in evolution. Further evidence for this mechanism emerged with the discovery, first in insects [5] and later in vertebrates [2], of *micro*conservation– that *short* (20-50 base) sequences perfectly conserved among diverse organisms are enriched for mature microRNAs by factors of 100,000 over whole genome. MicroRNAs act to repress translation by very rough protein-mediated complementary base-pairing to 3'UTRs of their target mRNAs, and therefore would not be expected to be strongly conserved *unless* they simultaneously regulate many targets. Indeed, the perfectly-conserved microRNAs while currently believed to constitute only a relatively small minority of all microRNAs, are believed to be those that regulate many distinct targets, precisely by the reasoning outlined above for enhancers. Therefore, this subset of microRNAs would be expected to represent an exceptionally strongly-conserved class of sequence.

At the purely genomic level, perfectly-conserved enhancer and mature microRNA sequences differ only in scale - perfectly conserved microRNA sequences are typically in the few tens of base range, and the perfectly conserved enhancers so far discovered are typically a few hundred bases in length. Nevertheless, the mechanism for their strong functional selection is identical, and this mechanism is not scale-specific. Remarkably, as first observed in [2], sequence conservation also has no intrinsic scale. The proposed mechanism is universal and scale-independent, and as demonstrated for the first time in

this extended abstract – figure 1 for vertebrates and figure 2 for yeast - the character of the scaling itself universal, and covers close to 15% of the human genome. The prediction then is that this mechanism – simultaneous constraint by multiple system-wide interactions mediated by DNA and gene products – applies not only to a subset of enhancers and microRNAs, but also to the remainder of this 15%.

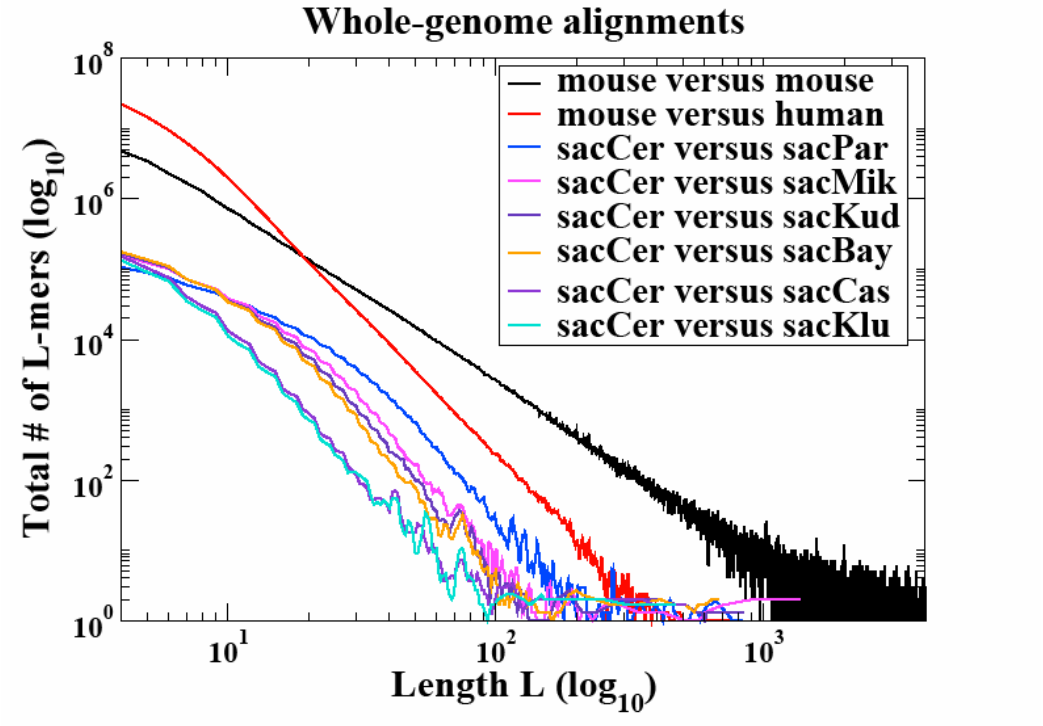

Figure 2: Length distribution of sequences perfectly conserved between the genomes of yeast Saccharomyces cerevisiae and other yeast sub-species. Mouse versus human is shown on the same plot for comparison; note the regimes of parallel slope, especially as the evolutionary distance between yeast subspecies increases. Also shown is the length distribution of runs of identical sequence obtained when the mouse genome is aligned against itself, and the diagonal (trivial self-alignments) removed.

1. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. **Ultraconserved elements in the human genome.** Science. 2004 May 28;304(5675):1321-5.

2. Salerno W, Havlak P, Miller J. **Scale-invariant structure of strongly conserved sequence in genomic intersections and alignments.** Proc Nat Acad Sci USA. 2006 Aug 20;103(35):13121-5.

3. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM. **In vivo enhancer analysis of human conserved non-coding sequences**. Nature. 2006 Nov 23;444(7118):499-502.

4. Mattick JS. **RNA regulation: a new genetics?** Nat Rev Genet. 2004 Apr;5(4):316-23.

5. Tran T, Havlak P, Miller J. **MicroRNA enrichment among short 'ultraconserved' sequences in insects.** Nucl Acids Res. 2006 May 12;34(9):e65-74.