[Invited Talk]
Colossal and super-colossal ultraconservation

Jonathan Miller[†1]

It is shown that the so-called ultraconserved sequences represent a small contribution to a substantially larger set of genomic sequences, a set that can be exhibited via whole-genome alignment or intersection such that their lengths fall into a "heavy," power–law tail of a sequence–length distribution. It is proposed that recombination is the primary mode of generation for this larger set of genomic elements, including the ultraconserved sequences. Evidence is presented for the scale-invariance of these putatively recombination-generated sequences.

1. Background

Modern genome sequence-based comparative genomics came of age in the mid-1990's, when Brenner and his colleagues discovered remarkably long sequences shared identically between fragments of the pufferfish and human genomes. Brenner choose fugu as a model organism because of its compact genome, which he anticipated would yield a distillation of functional sequence elements, both coding and non-coding. It was subsequently demonstrated that when introduced transgenically, certain of these sequences acted as enhancers in developing mouse embryos[1)].

The runs of perfectly-conserved sequence in question, for example the fugu *Pax6* genomic sequence encompassing Motif E (51 bases; c.f. figure 6c of[1)]) were not much longer than 50 nucleotides, but fugu is more distantly related to human than is mouse or rat, and the corresponding perfectly-conserved sequence between mouse and human encompassing the *Pax6* Motif E is 231 bases long. Thus, those sequences would today be called "ultraconserved" elements[2)]. They illustrate the principle that underlies the effectiveness of comparative genomics, namely that sequence variation is constrained by function. Nevertheless, what appears to be constraint on sequence variation can also arise by neutral processes. Pre-genomic era comparative genomics was driven in part by an intensive experimental focus on highly-expressed genes or easy-to-isolate gene-products, and on the most strongly-conserved genes and proteins that had been discovered at the time. In sheer numerical terms, only a small fraction of the genome - and its potential functionality - had been studied by the time whole-genome sequences became available, and the coverage was heavily biased both by technical practicalities and scientific fads.

Current methods for eukaryotic comparative genomics rely on an assumption, explicit or implicit, that linkage does not contribute significantly to observed patterns of sequence conservation. The conceptual justification for this assumption seems to be that linkage decays exponentially in time, so that for distantly-related organisms the effect of linkage becomes negligible. Thus, if we consider a correlation $C(i,j)$ of alleles between two loci i and j on the same chromosome, we anticipate that the magnitude of the correlation will become vanishingly small after many generations. The problem with this reasoning is that the relative magnitudes of the correlations, and not merely the absolute magnitudes, are important. Over long times, a little bit tighter linkage can have a substantial impact.

One way to address the relative correlation is to study the form of the correlation function, $C(i,j)$ as a function of i and j, subject to a fixed but possibly arbitrary normalization. For a fixed i, there may not be data sufficient to compute such a quantity with much precision, but if we are lucky, it may be possible to identify a suitable proxy, for this function.

In the following, we will adopt a certain class of distributions as our proxy. Their observed homogeneity over a genome allows us to compute the shapes of these distributions with no formal apparatus. (Here, "distribution" is just a fancy name for a histogram, a "locus" is a fixed position in some reference genome - e.g. base 7937987 of mouse chromosome 3, version 48.2 - and "allele" is the nucelotide at that position: A, G, C, or T). The histogram we study consists of a distinct bin for each length of conserved sequence. If we were studying ultraconserved

[†1] Physics and Biology Unit, Initial Research Project, Okinawa Institute of Science and Technology, 7542 Onna, Onna-Son, Kunigami, Okinawa 904-0411. E-mail: jnthnmllr@oist.jp.

sequences, we would find all contiguous runs of identical nucleotides in the whole-genome alignment of selected genomes. For each length L, we would count the number of such runs, and plot this quantity on the y-axis of a histogram with length L on the x-axis.

If substitution were the only relevant mechanism of genome evolution, linkage could only arise via some interaction (steric, perhaps) between the loci or their respective gene products. In the absence of such interaction, it is straightforward to see that the shape of this distribution would be exponential, namely $\propto \exp\{-L/\lambda\}$ where λ is the total length of the genome, divided by the total number of substitutions that have been fixated since the species diverged from one another. This represents the simplest "mean-field theory," where the fundamental length–scale λ is set by the inverse number of substitutions per base.

This particular mean-field theory is obviously a substantial idealization. Transposition, inversion, chromosome rearrangement, and variable base composition are just few of many factors not directly represented in the theory. Methods have been proposed to deal with some of these complexities, but they remain immature. Of course, all theories (or models) can at best be approximations to real-world phenomena.

As a correction to this mean-field theory, positional variation or inhomogeneity of substitution rate is just another bell or whistle: λ becomes a function of position whose variation must be accounted for and parameterized; however, the observations summarized here and elsewhere suggest that such inhomogeneity is - for the most part - merely apparent, arising instead from processes that in fact act homogeneously on the genome.

These drawbacks notwithstanding, mean-field theory generally represents the simplest approach to a quantitative problem. The process of understanding generally begins by constructing mean-field theories and applying them to calculate experimentally measurable quantities. It is an empirical question, as to whether any given theory is a suitable description of the data. Ordinarily, it is a significant accomplishment, and a crucial first step, to identify a theory that puts us into the ballpark of accounting for the data. For the current practice of comparative genomics, this mean-field theory forms the basis of all existing tools: any deviations from it are interpreted as *prima facie* evidence of selection. Does this mean-field theory at least put us into the ballpark of what is observed in whole-genome sequences?

The answer to this question, elucidated in[3], is that it fails the simplest test. Some new details, and some significant extensions, will be given below.

2. Ultraconservation.

We first construct the histogram discussed in the first section, for ultraconserved sequences in figure 1. Pairwise sequence alignments against human (hg18) were downloaded from UCSC[4], and all exactly-matching contiguous sequence runs were binned separately for each species. Any repeat-masking in the alignments was disregarded; i.e. no distinction was made between lower-case and upper-case bases. Aside from the latter, technical details of the computation have been described [3][6]. All plots in figures 1 and 2 are three-point running averages. The query genomes were, from right to left in figure 1: panTro2, ponAbe2, rheMac2, calJac1, equCab1, canFam2, bosTau4, felCat3, oryCun1, cavPor3, mm9, rn4, sorAra1, monDom4, ornAna1, taeGut1, galGal3, anoCar1, xenTro2, oryLat2, gasAcu1, fr2, danRer5, tetNig1, petMar1, braFlo1, strPur2.

This histogram, plotted on log-log axes, is strange because with the possible exception of the primates, the anticipated exponential form is absent. What one observes instead is a straight-line regime that extends for the human-mouse comparisons from lengths of around 500 bases to around 15 bases. The slope is approximately -4. There is no obvious separation of scales; in fact, the straight-line suggests that in this length span, all scales are equally weighted, a characteristic known as scale-invariance. Scale-invariance can be an indirect indication of self-similar geometry; more direct indications are discussed in later sections of this manuscript. Note that the 51 base *Pax6* sequence element referred to in the introduction is located well within the straight-line regime.

As the evolutionary distance to the target genome, in this case human, decreases, the straight-line regime acquires greater downward curvature and an increasingly better fit to an exponential distribution. These properties are quite general and largely independent of the choice of target species. Three-genome alignments and intersections also share these features - but in comparisons involving more organisms, terminating a run based upon a mismatch appearing in only one genome turns out not to be the best strategy.

Some remarks placing these plots in a more general context are in order:

1. This histogram is not a "ranked list" of the form addressed in recent scientific and layman-oriented popularizations of "heavy tails," such as a "Zipf" law or a plot of CD sales rank. The lengths here are physical and are measured in numbers of nucleic acid residues or angstroms, with the real-space geometry that these units entail. We make no judgment here on the significance, if any, of power–laws observed in ranked lists, which have to be assessed on a case–by–case basis; we merely observe that the distinction between scaling laws with and without geometry was made by Mandelbrot in the 1950s, to whom we refer the reader[10].

2. As the quality of a sequence assembly improves (usually with increasing version number) the straightness of the lines tends also to improve correspondingly; with later versions of mouse, the line translates uniformly up and to the right. Sea Urchin in particular demonstrates a marked improvement between first and second builds in this respect. The presence or absence of unresolved bases ('N') does not contribute significantly, presumably because in the assemblies studied here, they tend to appear in rare large blocks.

3. If assembly is not the source of a potential artifact, then alignment becomes the next suspect. Potential artifacts of alignment and repeat-masking have been comprehensively addressed[3], but a decisive observation reported in section 4 of this manuscript, is that intersection (exhaustive all-on-all comparison) of unmasked whole-genome sequences between, for example, human and mouse, yields the same power-law behavior. The overwhelming majority of the matching sequence runs that compose this latter set are not found in whole-genome alignments obtained from UCSC. These are the "super-colossal ultraconserved" sequences.

4. The same power-law behavior is also observed for individual chromosomes of, for example, mouse when compared to the whole human genome. The only chromosome that, at least for vertebrates, consistently yields the power-law when compared on its own between two species, is the X chromosome. Mitochondrial genomes exhibit primarily exponential behavior when compared to one another.

5. In the first instance, ultraconserved sequences were defined as perfectly-matching runs of contiguous bases exceeding some minimum length in the alignment of human, mouse and rat. The selection of three (versus two or four, say) genomes was arbitrary, as was the choice of the particular genomes themselves. The choice of the minimum length was without theoretical justification, and for any practical purpose, arbitrary. In principle it must depend upon the set of genomes and their evolutionary distances from one another. One could, as the authors implicitly did, try to select a length based upon the mean-field theory, but since the data indicate that the mean-field theory is "not in the ballpark" of even a crude approximate description, this strategy would appear to be futile.

6. Nevertheless, the sequence elements comprising these histograms for human/rat/mouse alignment correspond (above an arbitrary cutoff) are *exactly* the ultraconserved sequences obtained in [2]. For lengths above 30 bases, they coincide to within a fraction of a percent to what is obtained by human/rat/mouse intersection of repeat–masked whole genomes. (Obviously, any exact contiguous sequence match recovered by whole-genome alignment must also be recovered by intersection; the converse does not hold).

3. Colossal ultraconservation.

3.1 A=G/C=T runs.

We now loosen the stringency of the matching condition. As first described in[1], we identify A with G and C with T, so that a genome becomes a binary string. Although we are in the process of remediating this omission, currently available whole-genome alignments were performed while maintaining these distinctions, so that the best we can do with existing alignments is to terminate runs of "matching" bases at indels (deletions or insertions, denoted by "–"in each alignment block) and at A/C, A/T, G/C and G/T mismatches. Intersections, on the other hand, we can readily perform ourselves by effectively translating all G bases in the genome sequence to A, and all C bases to T.

For the example of the *Pax6* gene raised in the introduction, the effect of this A=G/C=T equivalence is to lengthen the human/fugu element encompassing Motif E from 51 bases to 131 bases, and the human/mouse element from 231 to 256 bases.

As illustrated in figure 2, the resulting histogram for the A=G/C=T runs is

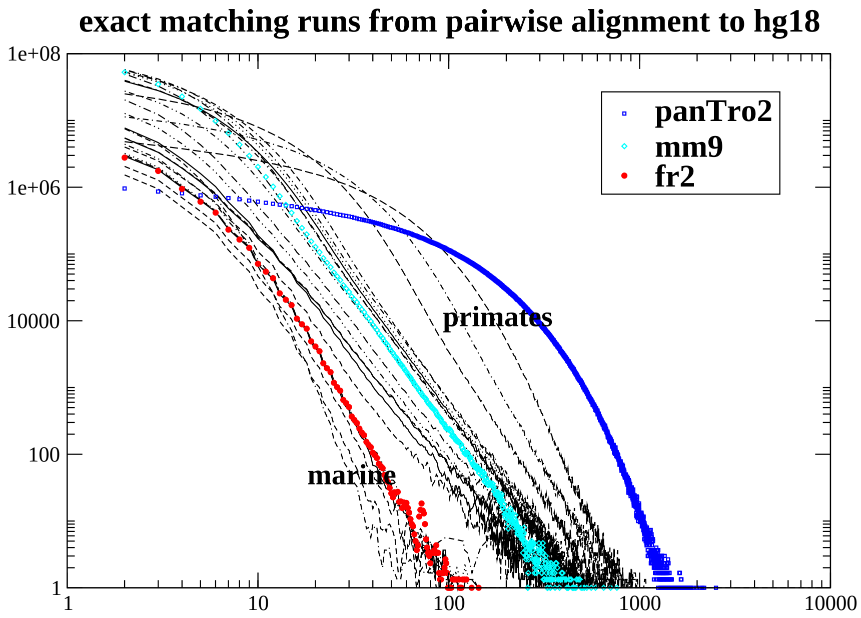

Fig. 1

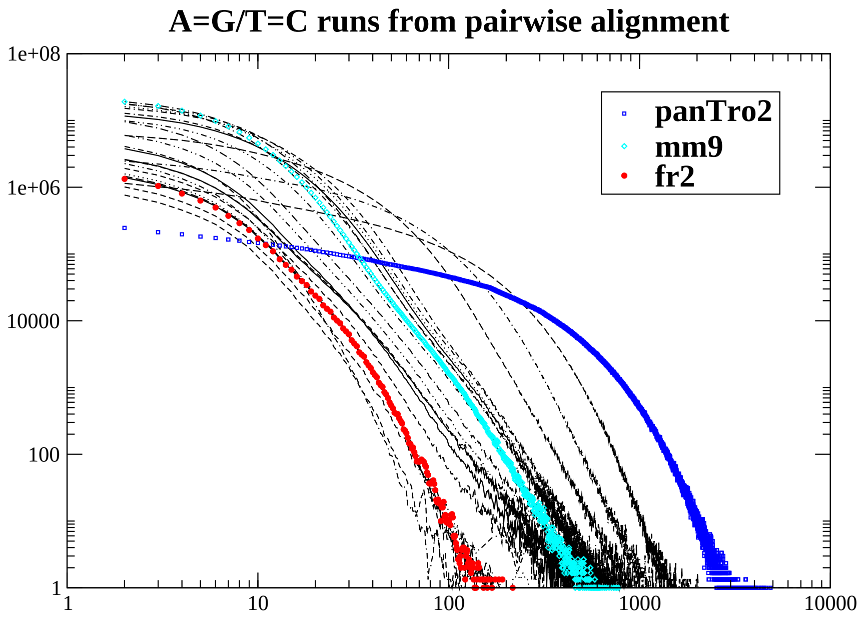

Fig. 2

nearly identical in form to the exact-match histogram of figure 1, except for what can be roughly characterized as a translation of the curves to the right, by $\log 2$. The curve is translated, which does not imply that the length of each sequence in bases is multiplied by this factor. Approximately the same curves are obtained if repeat-masked sequence in the alignments is discarded, and for around 60 bases or more, by intersection of repeat–masked whole genomes.

These observations raise questions about the significance of ultraconservation. Namely, any reasonable model for neutral evolution will, given any reasonable measure of sequence conservation, yield a set of most-conserved sequences and a set of least-conserved sequences. It is obviously in general a mistake to infer that the most-conserved sequences thus obtained are more likely to be under selection than the least-conserved sequences; rather, the differential arises wholly from the "noise" intrinsic to the process of neutral evolution.

3.2 indel-terminated runs.

The examples of subsections **3.2** and **3.3** are limited to alignment-based comparisons, because it is not yet clear how to define them in terms of an intersection. Figure 3 shows histograms of aligned sequence runs terminated by indels, which for our purposes we define as one or more deletions or insertions. More explicitly, a run of "matches" is terminated by either (a) the beginning or end of an alignment block; or (b) a "–" symbol in either one of the pair of aligned sequences.

The qualitative similarity of this set of histograms to the previous two is readily apparent, except possibly for the "bump" in the leftmost curves in the range of 100 to 200 bases. This bump originates in the failure of current whole-genome alignment methods to deal properly with indels above around 100 bases in length (data not shown), and we believe it would disappear if the alignment process were modified to compensate for this deficiency.

The corresponding lengths for sequence runs encompassing the *Pax6* Mo-

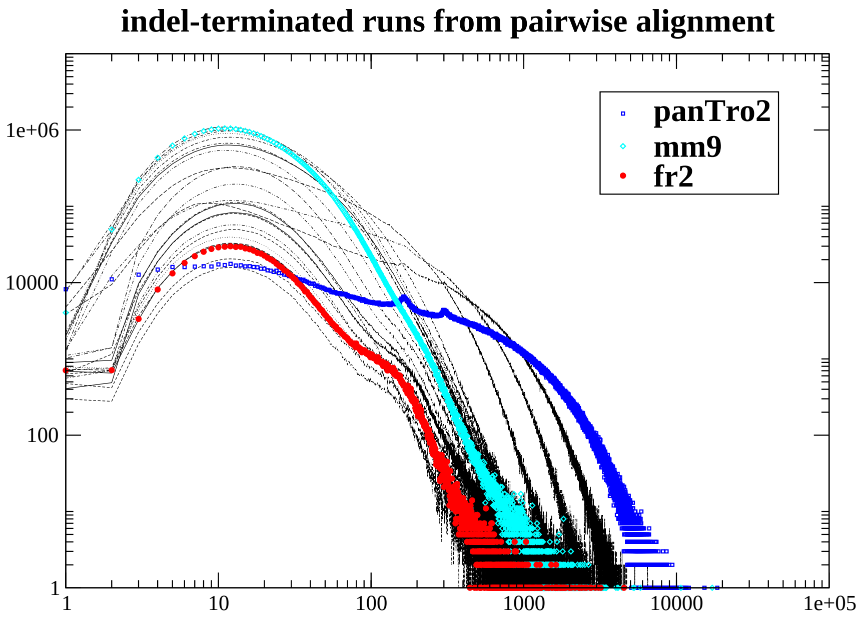

Fig. 3

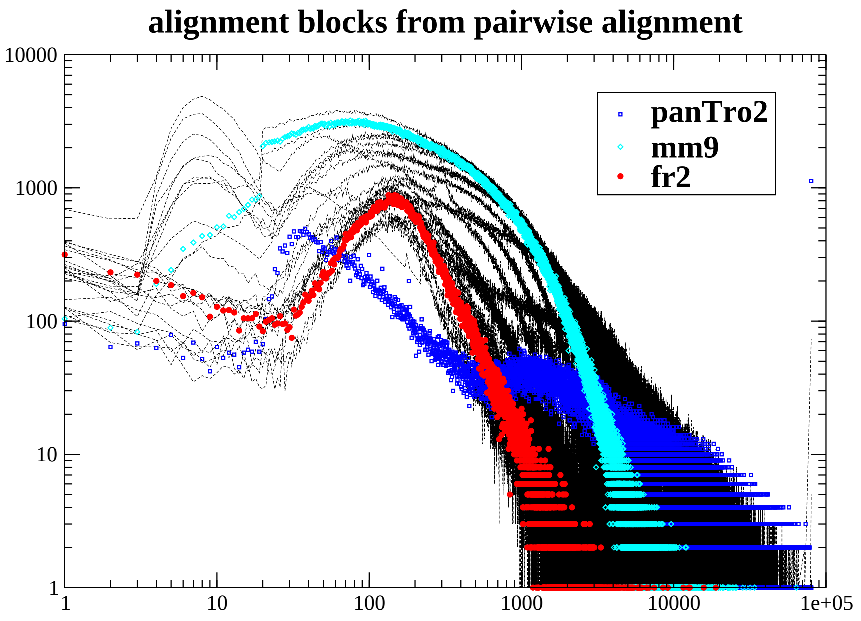

Fig. 4

tif E genome sequence neighborhood become: human/fugu: 51 (exact); 131 (A=G/T=C); and 304 (indel-terminated); human/mouse: 231 (exact); 256 (A=G/T=C); and 338 (indel-terminated).

3.3 alignment-blocks.

Our final example for this section is the raw alignment-block length histogram, figure 4. How it fits in with the plots described in the preceeding sections will become apparent in section 5.

4. Super-colossal ultraconservation.

Whole-genome alignment and whole-genome intersection yield essentially similar outcomes for exact matches and A=G/C=T matches, at least for sufficiently long sequences. Whole-genome alignment was an essential prerequisite for the examples of subsections **3.2** and **3.3**, which can't yet be produced by intersection. We now show a sequence comparison that is readily accomplished by intersection, but so far does not seem to have been obtained by alignment.

In figure 5, we exhibit A=G/C=T runs obtained from UCSC mouse/human whole-genome alignment, and from whole-genome intersection. The intersection was performed on the entire human and mouse genome sequences, irrespective of repeat-masking. The exact-match alignment histogram and the indel-terminated alignment histogram are shown in figure 5 to provide context; they are imported directly from figures 1 and 3 respectively; the A=G/C=T runs from alignment are also shown, imported directly from figure 2. Observe that the set of indel-terminated runs must by definition contain all the A=G/C=T runs obtained from alignment. Since intersections are exhaustive all-on-all comparisons, the A=G/C=T runs in the intersections needn't be subsequences of indel-terminated runs; indeed the overwhelming majority are not to be found among the indel-terminated sequence fragments. The A=G/C=T intersection histogram parallels the A=G/C=T alignment histogram, but with greater than a factor of 4 times

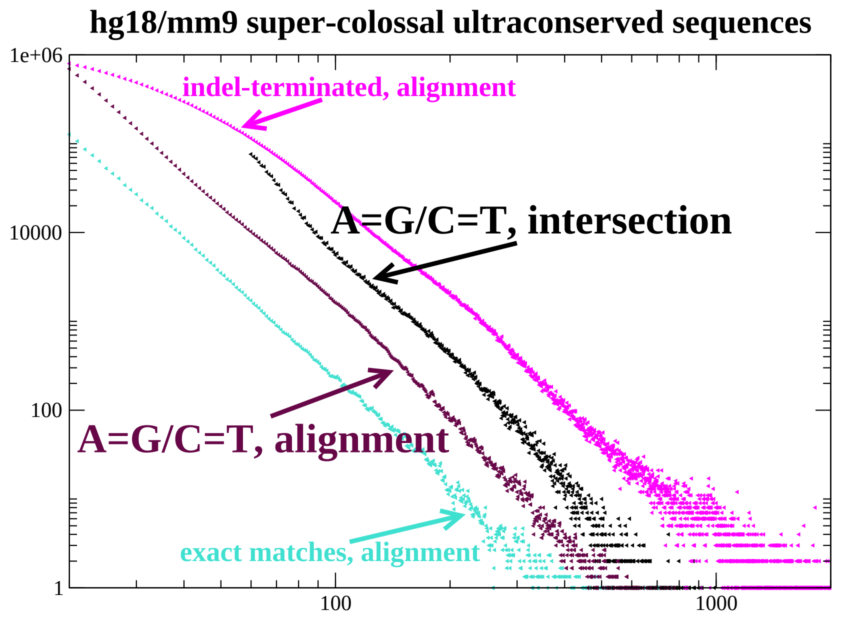

Fig. 5

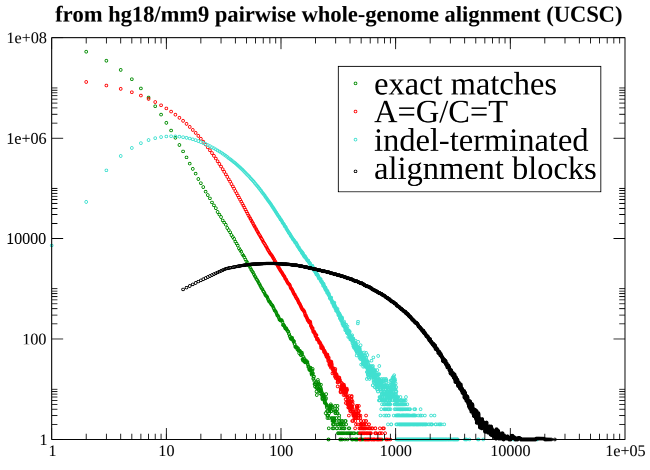

Fig. 6

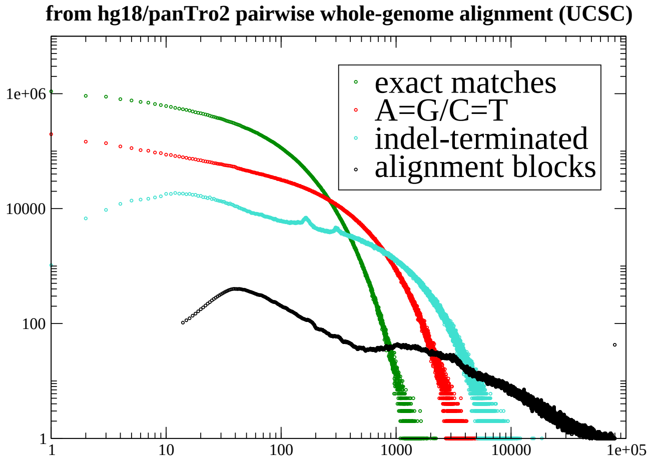

Fig. 7

the sequence at any given length.

Qualitatively similar features are observed in, for example, mouse or cow intersected with any of the primates. Efforts to use this observation to build better alignments are underway.

5. Self-similarity

In figures 6 and 7 respectively, the human/mouse and human/chimpanzee tracks have been imported from figures 1-4: exact-matches, A=G/C=T runs, indel-terminated runs, and block alignment lengths. The juxtaposition of the histograms for a single pair of species reveals clearly the parallel structure of the human/mouse plots, a structure that is far less obvious in the human/chimpanzee plots, and that would be completely absent for random sequence.

It is worth observing that for each pair of species, the alignment-derived sequences composing the exact-match plot of figure 1 are a subset of those com-

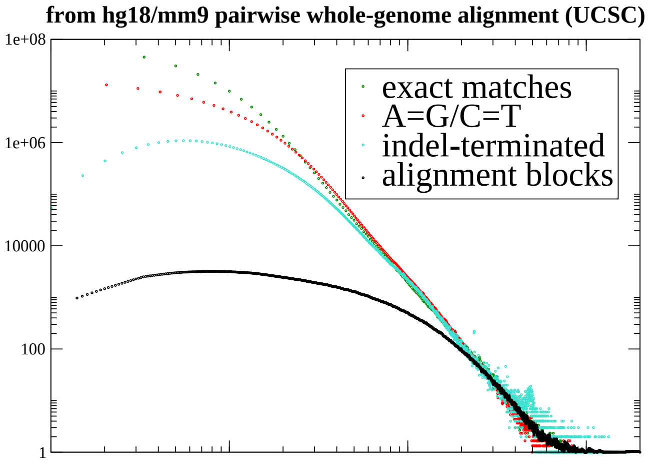

Fig. 8

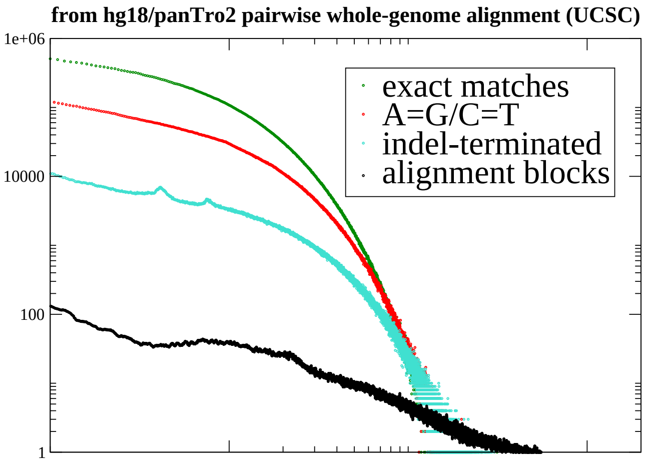

Fig. 9

posing the A=G/C=T run plot of figure 2, which are in turn a subset of those composing the indel-terminated run plot of figure 3, which once again are a subset of those composing the alignment block length plot of figure 4.

Consequently, the indel-terminated run sequences can be thought of as the result of cutting the alignment blocks at all indels. Cutting the indel-terminated run sequences at A/C, A/T, G/C, or G/T substitutions yields the A=G/C=T runs; cutting the A=G/C=T runs at A/G or C/T substitutions yields the exact matches. For random sequences, each successive step would produce an exponentially shaped histogram, $\exp\{-L/\lambda\}$ where λ becomes successively smaller in magnitude, in proportion to the total accumulated density of "defects," which for our purposes correspond to the cuts. In this process, any lines on the log-log plot would become progressively more steeply-curved.

As figure 7 shows, the last paragraph is a fairly decent characterization of the human/chimpanzee plots – but it does not apply at all to the human/mouse plots. We've tried to elucidate this distinction in figures 8 and 9, in which we have done our best to line up the individual plots by horizontal translation for human/mouse and human/chimpanzee respectively.

We do not mean to suggest that the human/chimpanzee aligned sequences are, by any means, "random;" far from it. But it should be apparent that the human/mouse aligned sequences have an additional structure not readily evident in the human/chimpanzee aligned sequences. This structure represents a more stringent reflection of self-similarity than the power-law, on which basis alone the self-similarity was first proposed[3)].

Far more explicit representations must be feasible, but self-similarity can take many forms, some of which are rather complex. The observations reported here are just another step towards that goal. What could be its origin?

6. Recombination.

In our first efforts at characterizing this structure[3)], we described the phenomenon in terms of "spatial correlations of conservation." We were reluctant to invoke the term "linkage" because linkage can occur between different chromosomes, not merely between bases on the same chromosome, and also because of the subtleties of "linkage disequilibrium;" however, these subtleties are probably unavoidable.

There is a third reason, namely that we first came upon this peculiarity when we found that runs of sequence conservation that were naively too short to be significant, turned out to correspond to microRNAs[6)]. Perhaps the most dramatic example, which can be found together with the first discovery of the power-law in the 2005 Cold Spring Harbor Genome Informatics meeting abstracts, is the enrichment for mature microRNA sequences in intersections of frog with mammals. The power-law regime is also rich in protein-coding sequence, and protein amino acid sequence too displays its own power-law distribution[8]. It was difficult to reconcile ourselves with the possibility that we could be studying the outcome of a neutral process - or more fairly, a neutral process shaped by selection. The connotation of the term "linkage" is itself quite neutral on the matter! Nevertheless, the observations we have described surely ammount to no more than the decay of linkage with separation of the loci, albeit with a rather distinctive functional form. And it is unclear how such a spatial decay of linkage can arise without recombination.

As the evolutionary distance between the species being compared becomes smaller, the fraction of sequence aligned becomes greater and the exact-match histogram, for example, becomes dominated by longer "diagonal" matches. Taken to the extreme of comparison of a genome to itself, it is clear that these diagonal matches are trivial, and their extension is reminiscent of the onset of long-range order in an Ising model. These diagonal matches are removed in creating a self-alignment[5] and we have demonstrated within the self-alignment[7] an distinct power-law that we speculate may reflect an ordered phase. The order parameter could be closely connected to the finite expectation value for an allele at a locus within a population, which must become vanishingly small for most of the genome in the limit of comparing genomes from distant species.

As the evolutionary distance between the species being compared becomes greater, it is plausible that a form of "quasi linkage equilibrium" proposed a long time ago by Kimura[11] becomes relevant and accounts for the spatial decay of linkage observed here. This possibility is under investigation.

7. Conclusion

We think it is likely that the biology of ultraconservation can be better understood within the context of the wider classes of conservation introduced here. More generally, we believe that our observations point toward a neutral theory of evolution, where the neutrality is with respect to recombination (broadly conceived) as well as with respect to mutation.

References

1) Kammandel, B., Chowdhury, K., Stoykova, A., Aparicio, S., Brenner, S., and Gross, P.: Distinct *cis*–Essential Modules Direct the Time–Space Pattern of the *Pax6* Gene Activity, *Developmental Biology*, Vol.205, pp.79–97 (1999).
2) Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., Haussler, D.: Ultraconserved elements in the human genome, *Science*, Vol.304(5675), pp.1321-5 (2004).
3) Salerno, W., Havlak, P., Miller, J.: Scale-invariant structure of whole-genome intersections and alignments, *Proc Natl Acad Sci USA*, Vol.103(35), pp.13121-5 (2006).
4) Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., et al.: The UCSC genome browser, *Nucl Acids Res*, Vol.31, pp.5154 (2003).
5) see http://genome.ucsc.edu/goldenPath/credits.html .
6) Tran, T., Havlak, P., Miller, J.: MicroRNA enrichment among short 'ultraconserved'sequences in insects, *Nucl Acids Res*, Vol.34(9), pp.65-74 (2006).
7) Miller, J.: Genomic Imprint of the Interactome: Universality of Genome Evolution, *ICSB* (2007).
8) Miller, J.: Scaling of Sequence: Impact of Interactome? *First annual q-bio conference on cellular information processing* Santa Fe, NM (2007).
9) Smit, A.., Hubley, R., Greene, P.: RepeatMasker Open-3.0, www.repeatmasker.org, (1999-2003).
10) Mandelbrot, B.B.: *The fractal geometry of nature*, N.Y: W.H. Freeman, pp.344-7 (2009).
11) Kimura, M.: *Attainment of quasi linkage equilibrium*, Genetics, Vol.52, pp.875-890 (1965).