Meeting report

Getting positive about selection

Eugene V Koonin and Igor B Rogozin

Address: National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20894, USA.

Correspondence: Eugene V Koonin. E-mail: koonin@ncbi.nlm.nih.gov

Published: 23 July 2003

Genome ***Biology*** 2003, **4:**331

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/8/331

A report on the 68th Symposium on Quantitative Biology, 'The Genome of *Homo Sapiens*', Cold Spring Harbor, USA, 28 May - 2 June 2003.

The 68th Cold Spring Harbor Symposium was a celebration of the completion of sequencing the euchromatic portions of the human genome. There are still some gaps (around 1% of the genome), for example in the regions of recent duplications that are a characteristic feature of the human genome as discussed by many speakers, including Francis Collins (National Institute of Human Genome Research, National Institutes of Health (NIH), Bethesda, USA), Jane Rogers (Wellcome Trust Sanger Institute, Hinxton, UK), Robert Waterston (University of Washington, St Louis, USA) and David Page (Howard Hughes Medical Institute and Whitehead Institute, Cambridge, USA); more details of their talks are discussed later. This creates pleasing opportunities for more celebrations in the future as the 'true' completion is asymptotically achieved. In general, however, the current sequence is a (nearly) complete and accurate representation of the human genome and the time is ripe for a thorough genome analysis.

Big open questions remain. To mention just the obvious ones, how many protein-coding and RNA-coding genes are there in the human genome and other genomes, and how many pseudogenes; what is the extent of alternative splicing, and how important are antisense RNAs? In one answer to a long-standing question, the winner of the notorious bet on the number of human protein-coding genes has been announced: Lee Rowen (Institute for Systems Biology, Seattle, USA), who placed her wager on around 25,000 genes. Nevertheless, new genes keep emerging through computational and experimental analyses, and the important question remains: are there many genes in the genome that are expressed only for a short time and in specialized tissues, and so do not show up in EST libraries, and that are poorly conserved in evolution and so are not readily detectable in database searches?

The answers to these and other, even more fundamental, open questions are to be sought in evolutionary genomics. This simple concept has won over the mainstream genomics community, as this year's Symposium showed with crystal clarity. The recognition is probably data-driven: with the completion of the advanced draft of the mouse genome, the rapid progress in sequencing the rat and chimpanzee genomes, and a considerable amount of sequencing of other mammalian genomes, mammalian comparative genomics can now be pursued in earnest. Discussion of genome comparison methods and results dominated half of the presentations at the Symposium, if not more. Issues that, just two or three years ago, would have been considered arcana of evolutionary biology - such as the use of the K_a/K_s ratio (the ratio of the rate of non-synonymous nucleotide substitutions, which lead to a change in the encoded amino acid, to the rate of synonymous ones) to distinguish between purifying and positive selection - were addressed in many talks and vigorously debated. Purifying selection is selection acting against deleterious mutations, which are eliminated from the population. This is by far the predominant form of selection operating in evolution, the result being the largely neutral character of molecular evolution and preservation of the *status quo* in terms of fitness. A small minority of mutations significantly increases the relative fitness of their carriers, however. The frequency of these beneficial alleles increases and, ultimately, they are fixed in the population by positive selection.

A number of fundamental evolutionary issues were raised at the symposium. For example, what fraction of nucleotides in the human genome is constrained by purifying selection; this was discussed, in particular, by Collins, David Haussler (University of California-Santa Cruz, USA), and Ross Hardison (Pennsylvania State University, University Park, USA).

The prevailing view, reported in the mouse genome publication and largely agreed with by all three speakers, seems to be that, of the around 40% of nucleotides in the human genome that could be aligned with counterparts in the mouse genome, about 5% are conserved as a result of selection, whereas the remaining 35% simply have not had time to change. To phrase it somewhat differently, 95% of our DNA seems to be 'junk'. The margins of error on these numbers seem to be quite large, however. Firstly, delineation of short conserved 'islands' in long stretches of non-coding sequence is far from straightforward, and the approximately 40% value for sequence conservation might require revision, most likely downward. Secondly, the fraction of nucleotides conserved simply because of insufficient time for divergence obviously depends on the rate of neutral substitutions. The current best 'guesstimate' for human and mouse is a K_s value of around 0.6 substitutions per site, which means that the great majority of the conserved nucleotides have nothing to do with selection. This value depends on alignment procedures and statistical models of sequence evolution, however, and these are currently far from perfect. Again a revision might be due, most likely upwards. Should it be concluded that the human and mouse sequences are actually saturated with respect to neutral substitutions, any observed sequence conservation will have to be attributed to purifying selection.

Thus, we really do not know what fraction of our genome is subject to selection and hence, presumably, is functionally important. We are reasonably sure about the approximately 1.5% of the genome that consists of protein- and RNA-coding sequences: with the exception of the positions of synonymous substitutions, the great majority of these sequences - amounting to around 1% of the genome - are indeed under purifying selection. Of course, there are also some important stretches among non-coding sequences, such as transcriptional promoters and enhancers. The uncertainty about the amount of functionally constrained non-coding sequence is huge, however, ranging from perhaps as little as 2% of the genome to as much as 20%. Knowing the true extent of selective constraint in the genome is important not only from a purely theoretical standpoint, but for the practical task of identifying non-coding functional elements in the genome or, perhaps, in part, sequences coding for unknown forms of RNA. Although it is our impression that the uncertainty in the estimates of the fraction of the genome that is subject to selection was not fully exposed at the Symposium, at least one crucial aspect of the solution was stressed repeatedly: sequences, sequences, and more sequences will be the key. Indeed, comparison of genome sequences that are definitely far from saturation (for example human against chimpanzee) and saturated ones (for example human against monotreme) will help us reach certainty. Another part of the solution is, undoubtedly, a better theory that would allow us to estimate the K_s value.

Positive (Darwinian) selection at the molecular level appears to be rare but is critical for adaptation and for the 'invention' of new functions. These innovations are likely to be crucial for speciation, including the origin of modern humans. Detection of positive selection is far from being straightforward, however, as emphasized in several presentations at the Symposium and the ensuing discussions. Positive selection is measured by comparing the rate of non-synonymous to the rate of synonymous mutations in protein-coding sequences. Most often, the actual value determined is the K_a/K_s ratio: $K_a/K_s > 1$ is taken as evidence of positive selection, whereas the much more common case of $K_a/K_s \ll 1$ reflects strong purifying selection. Yoshiyuki Sakaki (RIKEN Genomic Science Center, Yokohama, Japan) presented one of the first large-scale sequence comparisons of human and chimpanzee genes, which yielded a surprisingly large number of pairs with $K_a/K_s > 1$. During the discussion, however, Ewan Birney (European Bioinformatics Institute, UK) suggested that many, if not most, of these are likely to be statistical fluctuations exacerbated by the small number of nucleotide substitutions (typically, human and chimp genes are approximately 98% identical). It seems that, without a robust statistical analysis, K_a/K_s values for closely related genomes could be more misleading than illuminating. In fact, it has been shown previously that tests for selection developed for large samples tend to be too liberal when applied to small samples. A small-sample test has been developed and might be a more reliable alternative.

Andy Clark (Cornell University, Ithaca, USA) discussed elegant maximum-likelihood approaches for testing the human-chimp comparisons for neutrality and accelerated substitution rate. Application of these methods to numerous alignments produced with chimp sequences from Celera Genomics yielded a list of candidates for positively selected human genes, which seemed to be plausible from the biological standpoint and compatible with the results from other studies - the list includes for example, some olfactory receptor genes and a number of genes involved in immunity and reproduction.

A true highlight of the Symposium was the presentation by Svante Pääbo (Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany) who also described a comparative-genomic study, in this case, of large random samples of the olfactory receptor genes sets from humans, great apes, several species of monkeys, and non-primate mammals. Pääbo and coworkers discovered two dramatic phases of pseudogenization of the olfactory receptor genes: the first occurred early in the evolution of monkeys, whereas the second took place after the divergence of the human lineage from that of chimpanzee and might still be ongoing in humans. Strikingly, the first wave of pseudogenization of olfactory receptor genes showed a perfect correlation with the advent of full trichromatic vision: once you see the world in color, smelling it precisely seems to be less of an issue. It

was also suggested during the discussion that the second wave might have been linked to the increasing crowding of human populations, as a result of which smelling one's neighbor too keenly could be a disadvantage, but the true cause remains to be understood. This study was met with considerable interest and even enthusiasm, perhaps as a glimpse of things to come with the dawn of genuine biological understanding of genomics.

In conclusion, genomics is going comparative in earnest. The importance of robust theory for further success of evolutionary genomics cannot be overemphasized. Many ready-made solutions can be found in the annals of molecular evolution and population genetics but there is no doubt that new concepts and methods will also emerge.