

How to count...human genes

Samuel A.J.R. Aparicio

University Department of Oncology, Wellcome Trust Centre for Molecular Mechanisms in Disease, Cambridge CB2 2XY, UK
e-mail: saparici@hgmpr.mrc.ac.uk

Biology occasionally mirrors human activity with unnerving irony. This year has seen the spectacular rise and fall of biotechnology stock values as at first exuberance and then sanity swept through the investor community. Based on reports¹⁻³ presented on pages 232, 235 and 239, similar sentiments should now apply to estimates from some organizations of ever-increasing values for the total number of human genes. With the near completion of the human draft sequence, mere gene counting may seem a sterile exercise—the 'real' answer will surely be known soon? The analyses in this issue throw into sharp focus the question of what should be counted as a gene. They indicate that, not only should our expectations for the full number of human genes be revised downwards, but also, that existing EST databases may contain as little as 40% of the protein-coding fraction of the human genome.

EST clustering versus direct sampling

Previous attempts at estimating the number of human gene loci have been predicated on approaches such as measuring the complexity of cellular RNA, reassociation kinetics, CpG island determination, evolutionary 'rules of thumb' or assuming that cDNA sequences represent genes⁴⁻⁶. This latter approach is quite popular, and employed by a group³ from The Institute for Genomic Research in one of the present analyses. The authors make use of the extensive public cDNA sequence databases to estimate the total number of genes. They carefully clustered the cDNA sequences (to eliminate sequence redundancy), excluding 'singleton' sequences (as these are probably artefactual), and then estimated the fraction of the clustered sequences that might represent known genes. They estimate that there are between 120,000 and 140,000 human genes. A key assumption is that clustering is sufficient to remove artefacts that arise consequent to false poly(A) priming, clone

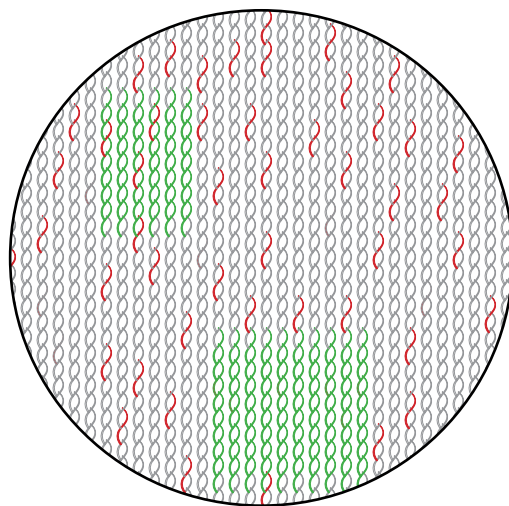
amplification, DNA contamination and other factors that would spuriously inflate the total estimate. Crucially, however, one should challenge the assumption that almost any (clustered) transcribed sequence represents a gene.

Through alternative and independent processes, groups led by Philip Green and Jean Weissenbach arrive at a quite different estimate: they conclude that the number of protein-coding genes is approximately 35,000 in total. Green and colleagues¹ use a

the first set, the authors used either the curated annotated genes from chromosome 22 (ref. 4) or a filtered set of near full-length mRNA sequences from GenBank. Comparison of these sets with a reduced set of clustered EST sequences indicates that there are either 34,700 genes (based on chromosome 22) or 33,600 genes (based on mRNA sequences) in the human genome.

Weissenbach and colleagues² used a different approach that exploits pufferfish genomic sequence. They describe the sequencing of BAC clone ends from the compact genome of a pufferfish^{7,8}; these collectively represent approximately one-third of the genome sequence of this vertebrate. By comparing this sequence set with a known, curated set of protein-encoding genes, they were able to calibrate an algorithm that detects orthologous coding sequences (called ecores) between pufferfish and human genomes. On measuring the mean number of ecores found per human gene across the calibration set of genes and then running the same analysis across the human draft sequence (approximately 60% of the human genome at the time of analysis), they arrived at an upper limit of 34,000 genes. Note that the pufferfish sequence will not pick up some human orthologues, owing to evolutionary distance or because they are not represented in the sequence set. But so long as the percentage of undetectable sequences (around 30% in this case) is not disproportionately represented in the pufferfish or human sequence sets, the calculation holds true.

They also showed that, whereas known, related and pseudogenes are easily detected, applying the algorithm to chromosome 22 pulled out only 4% of the 300 gene sequences predicted by Genscan—indicating that Genscan significantly overpredicts human genes. A similar detection sensitivity was observed with the Unigene set: the pufferfish sequence identified about 65% of the 10,500 protein-coding transcripts as ecores, but only matched 4%



Two-sample comparison method for estimating gene numbers. The schematic represents a body of sequences (grey spirals) for which a homogeneously representative sample, n_1 , is taken (red spirals). The fraction of the total sequences, G , in set 1 is given by $f = n_1/G$. A second sample is taken, which may be biased, redundant or incomplete (green spirals). Providing the sequences are of sufficient quality to correctly identify orthologous matches of the first set, then the fraction of matches for the first set in the second set will approximate to the fraction of sequences from the first set, to the total number of sequences. In other words, $n_1/G \cong m/n_2$ where m is the number of set 1 sequences matching set 2 sequences and n_2 is the total number of set 2 sequences. Thus $G = n_1 n_2 / m$.

modified form of the method applied to the genome of *Caenorhabditis elegans*, a simple and elegant method involving two samples (see figure). This approach requires a small but homogeneously sampled collection of genes from the genome and a second comparison sample which is larger but can be biased, redundant and incomplete, as long as the sequence is of sufficient quality to reliably match sequences of the first set. For

of EST clusters. The comparison indicates that the clustered EST set is redundant and probably contains no more than 40% of the coding fraction of the human genome.

Bursting the bubble

In principle, the estimates by Green and Weissenbach are subject to biases that could lead them to an underestimation of gene number. As these biases are methodologically independent, however, it seems unlikely that these estimates will prove wildly inaccurate. Agreeing with them are estimates^{9,10} of 40,000 and 45,000 genes based on chromosomes 21 and 22. As mammals have probably experienced two genome doublings since their divergence from multicellular invertebrates, the total is very unlikely to exceed about 60,000 genes in any event. So why should calculations based on EST data have resulted in such large estimates?

At the core of these findings are issues of definition and recognition—that is, how does one define a gene, and how does one recognize it? According to classical genetics, genes are the heritable units responsi-

ble for an associated phenotype. Although in some cases this relationship derives from mutation of regulatory elements or other non-coding DNA elements, in most cases it is synonymous with mutation of protein-coding DNA sequences. Although the tendency (especially in a pay-per-sequence access mode) is to assume that any transcript represents a gene, classical genetics demands some evidence of associated function. Crucially, what is not yet established (but is implied to be relatively abundant by these studies) is the extent of biological “noise” in the transcriptome of any given cell. In other words, what fraction of transcripts which can be isolated have any meaningful function? What fraction might be mere by-products of spurious transcription, spuriously fired off, perhaps on the antisense strand from promoters or CpG islands associated with protein coding genes (as seems to be the case with a number of imprinted genes)? The good news therefore, is that the human draft sequence will be a goldmine for protein-coding sequences not represented in the EST collections.

Clearly, the task of annotating genes in the human sequence will take time, and a comparative approach has much to offer. Beyond this lies the challenge of understanding gene regulation. The ability to make adequate comparisons of non-coding sequences of different species should be a rapid means by which to obtain a regulatory element ‘framework’ for the human genome. Evolution is certainly more powerful and has more to teach us here than any extant computer algorithm. Our ability to compare the genomes of many species may yet turn *in silico* biology into a true science.

1. Ewing, B. & Green, P. *Nature Genet.* **25**, 232–234 (2000).
2. Roest Crollius, H. *et al. Nature Genet.* **25**, 235–238 (2000).
3. Liang, F. *et al. Nature Genet.* **25**, 239–240 (2000).
4. Soares, M.B. *et al. Proc. Natl Acad. Sci. USA* **91**, 9228–9232 (1994).
5. Antequera, F. & Bird, A. *Proc. Natl Acad. Sci. USA* **90**, 11995–11999 (1993).
6. Fields, C., Adams, M.D., White, O. & Venter, J.C. *Nature Genet.* **7**, 345–346 (1994).
7. Dunham, I. *et al. Nature* **402**, 489–495 (1999).
8. Hattori, M. *et al. The DNA sequence of chromosome 21. Nature* **405**, 311–319 (2000).
9. Brenner, S. *et al. Nature* **366**, 265–268 (1993).
10. Aparicio, S. *et al. Proc. Natl Acad. Sci. USA* **92**, 1684–1688 (1995).

Better taste through chemistry

Peter Mombaerts

*The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA.
e-mail: mombaer@rockvax.rockefeller.edu*

Wedding genetics with genomics, two groups^{1–3} have independently identified in mammals a multigene family encoding seven-transmembrane proteins that probably represent the hitherto elusive taste receptors. Functional expression in heterologous cells confers responsiveness to bitter tastants in several cases. This discovery may lead to the design of bitter antagonists—and perhaps better butter for Betty Botter.

Mammals perceive four types of taste: sour, salty, sweet and bitter⁴. In recent years, a fifth modality has come to be recognized: umami, the taste of monosodium glutamate. Tastants interact with taste receptors that sit on the surface of taste receptor cells (TRCs). These specialized epithelial cells are grouped in taste buds, which in turn are part of taste papillae, macroscopic structures on the surface of the tongue and palate (Fig. 1). TRCs synapse with afferent neurons, whose axons project into the brain. Here, gustatory signals meld with olfactory and somatosensory stimuli, producing the unique sensations elicited by, say, ice-cream, lemon zest or chicken vindeloo.

Whereas our sense of taste provides many delights, we understand little of its molecular

mechanisms. The literature harbours evidence for almost any imaginable signalling pathway, which may be a consequence of the true complexity of this chemosensory sys-

*Betty Botter bought some butter
but, she said, the butter's bitter;
if I put it in my batter
it will make my batter bitter,
but a bit of better butter
will make my batter better
So she bought a bit of butter
better than her bitter butter,
and she put it in her batter
and the batter was not bitter.
So 'twas better Betty Botter
bought a bit of better butter.*

tem. Sour and salty tastes are probably mediated directly by the interaction of cations with ion channels, and the other modalities are thought to rely on G-protein signal-transduction pathways. (G-protein-coupled receptors are seven-transmembrane (7TM) proteins.) Supporting this hypothesis is the finding that ablation of gustducin (a well-characterized G-protein) results in mice with diminished ability to detect sweet and bitter tastants⁵. Furthermore, a truncated version of a metabotropic glutamate receptor, also a 7TM receptor, can transduce the taste of umami⁶. Last year, the groups of Ryba and Zuker identified two putative 7TM receptors⁷, termed taste receptor-1 and -2 (TR1 and TR2, respectively), and noted their expression in certain types of