Comment

Size doesn't matter

Gregory A Petsko

Address: Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham, MA 02454-9110, USA.
E-mail: petsko@brandeis.edu

Published: 28 February 2001

Genome ***Biology*** 2001, **2(3)**:comment1003.1–1003.2

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/3/comment/1003

So now we know. The first 'official' count of the number of genes in the human genome is in, and the total is smaller than almost anyone had imagined. Sorting out the pseudogenes from the real ones will take some time, so the number may increase a bit, but it seems clear that the genome of *Homo sapiens* contains fewer than 40,000 genes, with the final number probably being closer to 30,000. Quibble about the exact count if you will, but the total will probably not approach even half the 80,000-100,000 estimate that was widely bandied about when the Human Genome Project began.

What a blow to our collective ego as a species! Thirty thousand genes is only 50% more than the 19,000 in the genome of the nematode worm *Caenorhabditis elegans*. It is just a bit more than double the 13,000 genes in the fruitfly *Drosophila melanogaster*. And it is only five times the number of genes in a unicellular microbe, the budding yeast *Saccharomyces cerevisiae*. I teach undergraduates, so the idea that human beings are not even an order of magnitude more complex than a fungus is not incredible to me, but it still begs the question: how could we have evolved as such complex beings with such a minimalist genome? Does size truly not matter?

Well, the first lesson from the human genome sequence would seem to be that the number of coding sequences doesn't matter as much as we thought. There is still a crude relationship between the complexity of the organism and the number of genes: no bacterium whose genome has been sequenced to date has a gene count even as large as that of yeast, and metazoa all seem to have tens of thousands of genes instead of a few thousand. Yet the preliminary reports of some other vertebrate genomes, such as that of the puffer fish, suggest a gene count that may exceed that of the human genome, and some plant genomes are also shaping up to be quite large. Clearly, the correlation between complexity and gene number is a loose one.

But the size of the proteome is another thing altogether. No one knows how many different proteins make up the complete ensemble of human gene products, but current estimates range from 100,000 to several times this number (of course, we should remember that the number of human genes was once estimated to be that large too). This is in stark contrast with the situation for simpler organisms. For most bacteria, the relationship between genes and proteins is approximately 1:1, as indicated by the number of proteins that can be resolved on big two-dimensional gels. Yeast and other microbial eukaryotes would seem to be similar in this respect. But for metazoa, especially vertebrates and even more especially mammals, there is considerable expansion of the proteome relative to the genome.

One place where mechanisms for this are evident is in the immune system, where polymorphisms in the V, J, and D regions, somatic mutations, and recombination can combine to produce, potentially, billions of different immunoglobulins from a set of genes many orders of magnitude smaller in number. From this lesson we can begin to imagine some of the mechanisms by which a small number of genes can give rise to mind-boggling complexity at the level of the cell or the animal.

The first of these is by various forms of editing of the message. Cells of higher organisms clearly treat mRNA in much the same way my word-processor can treat text. Messages can be cut and pasted in many different ways (alternative splicing), and individual words and phrases can be modified or replaced (RNA editing). There are no good estimates for the number of human genes that are subjected to either of the these procedures, but whenever it happens, multiple gene products - often of quite different sequence and, presumably, function - are produced from a single coding region of DNA. Available data suggest that the frequency of such manipulation increases dramatically as one goes 'up' the evolutionary scale towards humans. One

conclusion we can draw from this is that we desperately need methods to scan a gene sequence and to know whether alternative splicing and/or editing is likely, and, ideally, what the results of such modifications will be. It seems clear that the key to how to do this will lie in understanding the role of the non-coding regions of the genome, which no one in their right mind should ever refer to again as 'junk' DNA (unless it is with tongue planted firmly in cheek).

Of course, the non-coding regions are also where much of the regulation of gene expression is controlled, through the binding of enhancers and other modulators of transcription. Relative to other eukaryotes, this part of the human genome is very large, so that the total number of base pairs is in the billions even though the number of genes is only a few tens of thousands. I suspect that expansion of the non-coding part of the genome is very important for the evolution of complexity, since it scales well with the apparent sophistication of the organism. Increasing the size of regulatory elements would allow for a greater number of combinatorial possibilities for gene expression, thus permitting a wide range of phenotypes from a smaller set of instructions.

This consideration, though obvious, has, I think, profound consequences - because it suggests to me that the real issue isn't even the number of proteins that can be produced from a single transcript. The real issue is the number of distinct protein functions that a given gene can encode. Here we are on shakier ground, but the evidence is mounting rapidly that this number could be large, and word-processing of the message is only one of several mechanisms by which functional possibilities are expanded. Post-translational modifications, such as limited proteolysis, phosphorylation and methylation, can clearly alter the function of a protein, in some cases by serving as a reversible switch. Ligand binding can do the same - the small GTPases have different cellular functions in their GTP-bound and GDP-bound states, for example. So can binding to a membrane or another protein: the resulting conformational rearrangements can cause a complete change in what a given gene product can do. The location within the cell in which a protein is found can also determine its function: witness the number of proteins that can act as transcription factors once they are translocated to the nucleus, usually after some covalent modification such as phosphorylation or following the release of some inhibitory partner. It seems clear that we cannot claim to have enumerated the functions of a gene until we have established the totality of the modifications and interactions that its protein product(s) can undergo, and the precise locations in which they occur.

Yet even this is unlikely to suffice. Recently, it has become clear to many biologists that, at least for some proteins, the concept of a single 'active site' is too simplistic. Consider the case of the extracellular cytokine neuroleukin and the housekeeping glycolytic enzyme phosphoglucose isomerase (PGI). The second enzyme in the pathway from glucose to pyruvate, PGI would appear to be an example of a simple gene product with one function: to convert glucose-6-phosphate to fructose-6-phosphate. Neuroleukin, a potent cytokine in the development of the central nervous system, would also appear to be an example of one gene - one function. But appearances, in genomics, are deceptive, for PGI and neuroleukin are the same molecule.

Leaving aside for the moment the obvious question of how an intracellular metabolic enzyme with no signal sequence gets out of the cell in the first place, the question of what it is doing out there is tough to answer. This is not a case of alternative splicing or post-translational modification, because purified PGI from a cloned gene will function just fine in a neuroleukin assay. We call this phenomenon 'moonlighting': the taking of a second job by a protein whose function we thought we knew. And PGI, astonishingly, seems to have two more jobs besides (when does it ever sleep?). It also moonlights as autocrine motility factor (AMF), a role in which it causes tumor cells to become motile, and as DMM, a mediator of the differentiation of leukemia cells. Specific receptors have been isolated for some of its functions.

PGI is not the only eukaryotic enzyme that moonlights. Thrombin, the enzyme whose action causes blood to clot, also functions as a cytokine through binding to a specific receptor. Methionine aminopeptidase doesn't only remove the amino-terminal methionine residue from newly synthesized proteins; it also serves as a specific cofactor in the translational machinery of the ribosome. In all of these cases, the non-enzymatic functions of these proteins are independent of their catalytic action and reside in regions of the protein surface distinct from the 'active site'. Many more examples of moonlighting are turning up all the time, and the phenomenon may explain a curious fact, namely that the average size of a given protein increases as one goes from bacteria to higher organisms. The grafting of non-enzymatic signaling and regulatory functions onto the polypeptide chain as organisms became multicellular would allow the genome size to remain relatively small while expanding the size of the gene products only modestly (for a further account of moonlighting, read the 1999 review by Connie Jeffery: *Trends Biochem Sci* 1999, **24:**8-11).

Thus we have at least four potential mechanisms by which a small number of genes can give rise to many times that number of functions: word-processing of the message; post-translational modification, ligand binding, and localization; combinatorial protein-protein association and regulation of expression; and moonlighting. Taken together, they easily allow 30,000 genes to produce 150,000 'different' proteins at the level of function. In fact, the more one thinks about it, the more one suspects that, for *Homo sapiens*, the number of gene functions - as distinct from genes - may be seriously underestimated. Any bids for a million?