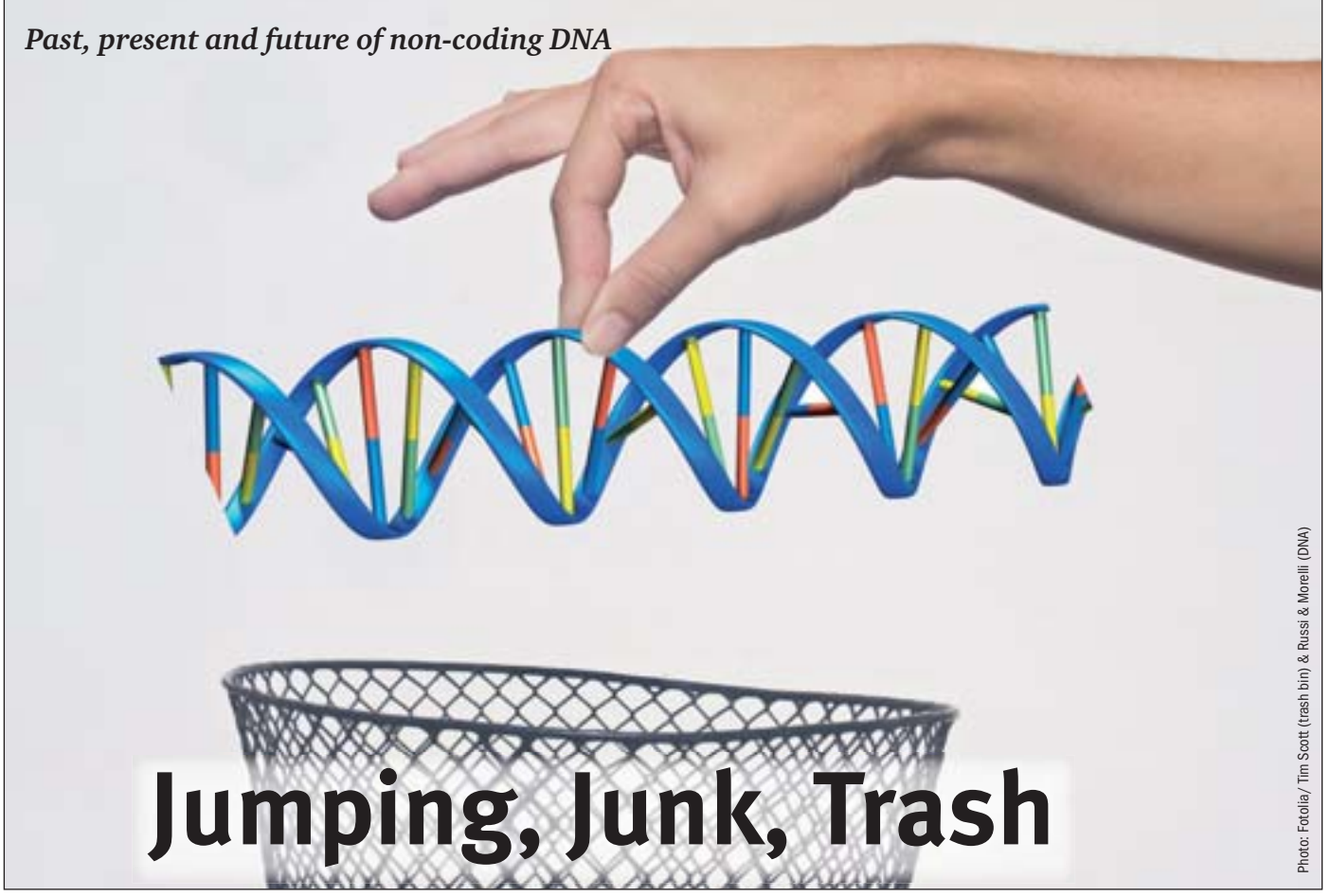
Past, present and future of non-coding DNA



Jumping, Junk, Trash

Photo: Fotolia/ Tim Scott (trash bin) & Russi & Morelli (DNA)

Decades after decoding many genomes including the human one, some big questions still remain: Why is there so much DNA that does not code for proteins? Does it have any function or is it really just junk? And speaking of junk: Are transposons junk or not?

When the first draft of the human genome finally became available, there were many surprises, one of them being the ridiculously large amount of DNA that did not code for proteins – or anything at all. It seemed like the genome was just a huge chaotic mess sprinkled with tiny nuggets!

In the genomes of higher eukaryotes, protein-coding genes are generally composed of exons containing an open reading frame: thus, their DNA sequences can be "read" into their corresponding amino acid sequences throughout their full length. However, the analysis of the human genome revealed that less than 2% of the human DNA actually codes for proteins.

So much junk

So what about this huge amount of DNA not coding for proteins? Until now, many different terms have been formed: junk DNA, non-coding DNA, intergenic regions… Some magazine articles have even wickedly paraphrased it as "The Dark Matter of the Genome".

The term 'junk DNA' was coined in Susumu Ohno's article "So Much 'Junk DNA' in our Genome" from 1972 (*Brookhaven Symposia on Biology,* 23: 366-70). However, Ohno's definition of 'junk DNA', at the time, was somewhat different to the meaning today. He referred to 'junk DNA' to all pseudogenes, which had originated by duplication and then degraded. Accordingly, he wrote, "The earth is strewn with fossil remains of extinct species; is it a wonder that our genome, too, is filled with the remains of extinct genes?"

Moreover, Ohno attributed a certain purpose to this 'junk DNA', namely to safeguard protein coding genes from harmful deleterious mutations. His basic idea was that since all genes were surrounded by duplicated, non-functional copies of themselves, most, if not all, mutations would fall into these sequences – and so the 'real' genes would be protected.

However, as so often in science, some agreed with Ohno's term 'junk DNA', some didn't and some were somewhere in between. Today, the term 'junk DNA' is no longer *en vogue* among many genome experts. They rather prefer to talk about 'non-coding DNA'. And this certainly includes much more than just pseudogenes, which actually represent less than 1% of the non-coding DNA. The remaining 99% are made of everything else, for which no involvement in any protein-coding function has been found so far including, for example, introns, repeated sequences, interspersed elements, telomers, transposons…

In the last decades, considerable research has been carried out showing indeed some – mostly regulatory – functions for a couple of these non-coding elements. For most of the 'junk', however, we still have not the slightest idea why it is there, why it is maintained over time or what it is doing.

One of the people who have been doing some crucial work on the importance of non-coding DNA is Sebastian E. Ahnert from the Cavendish Laboratory at the University of Cambridge. In a recent paper, he argues that basically all eukaryotes from yeast to humans require a certain minimum amount of non-coding DNA (*J. Theor. Biol.* 252: 587-92).

Answers needed

Ahnert and Co. observed that the minimum increases quadratically with the amount of DNA located in exons and, subsequently, he developed a mathematical framework for predicting the respectively required quantity. He, therefore, concludes, "I think 'junk DNA' is an unfortunate term, coined by scientists in a moment of overconfidence. It is likely that this 'junk' will turn out to be as important as the genes it surrounds and that it remains still poorly understood."

Indeed, after all these years, there are still two main questions that trouble genome experts. Firstly, why is there so much seemingly useless DNA anyway? Secondly, and perhaps even more confusing, why are there such large differences in the respec-

tive amounts of non-coding DNA between different organisms?

Currently, we are apparently getting some answers to the first question. We now know that about 10% of the non-coding DNA actually plays an important role at some level but what about the remaining 90%? For this DNA, potential hidden meanings still have to be unearthed, unless it is indeed just 'junk'.

Regarding the second question, there is no straight answer in sight. If you compare the amounts of non-coding DNA between yeast and humans, for example, you might be surprised to learn that in humans almost 98% of the genome is non-coding DNA, whereas in yeast it is just over 30%. How can yeast do well with so much less non-coding DNA? Or why do we have so much?

The amazing roles of TEs

You might think humans have more non-coding DNA, because, at the end of the day, we are much more complex organisms than unicellular yeast. Then let's stop for a minute and first of all do "the onion test", a term coined by evolutionary biologist T. Ryan Gregory from the University of Guelph in Canada. Basically, the test is nothing more than one simple question. In his blog "Genomicron", Gregory writes, "The onion test is a simple reality check for anyone who thinks they have come up with a universal function for non-coding DNA. Whatever your proposed function, ask yourself this question: Can I explain why an onion needs about five times more non-coding DNA for this function than a human?"

The onion is only one example. There are many other organisms that are clearly simpler than humans but have much more non-coding DNA in their genomes. Therefore, the tempting equation "the higher the complexity, the more non-coding DNA" is simply wrong.

A somewhat controversial group in this context are transposable elements (TE), also known simply as transposons or jumping genes.

Traditionally, TEs have been dismissed as useless (or even harmful) and, therefore, have been attributed to 'junk DNA'. At the same time, strictly speaking, they cannot be regarded as 'non-coding' in their entirety. In fact, functional TEs within their sequence actually code for the enzymes needed to cut and paste themselves anywhere in the genome. However, the picture is even more ambivalent: In recent years, researchers have added a considerable amount of evidence indicating that TEs might not at all be completely 'useless'. Thus, the question, whether still to assign TEs to 'junk DNA', has become ever more pressing. In the meantime, many would clearly say 'no'.

Weird names and complex structures

Let's delve a bit deeper into the world of jumping genes. TEs were first described in the 1950s by Barbara McClintock but even today we are only beginning to understand how they work. They are found throughout the genomes of all organisms, in very different proportions. In humans, for example, they represent 45% of the genome, whereas in *E. coli,* they make up only 0.3%. Of course, the genome of *E. coli* is much smaller but it's the proportion that matters. In general, they are divided into two main types – those TEs that use the DNA replication machinery to multiply and those using RNA. These latter ones are called retrotransposons, of which, again, there are two types – one containing "Long Terminal Repeats" at the ends (called LTR retrotransposons) and the other one that doesn't (called non-LTR retrotransposons). Moreover, there are a couple of different TEs with much more complex structures and even weirder names. One example lies in the "helitron elements" of the maize genome – rolling circle elements, which were found to be responsible for the duplication of segments of genes into new locations.

Photo: photocase.com/Helgi

Are you ready for "the onion test"?

TE invasion

Now, imagine millions of these sequences moving around the genome. Surprisingly enough, most of these sequences are neutral and cause no harm; but eventually some of them do land inside a protein-coding gene and then things can easily get nasty: TEs are known to disrupt gene function and can even cause chromosome rearrangements. No wonder, they have also been associated with serious diseases, particularly cancers. For example, haemophilia, tumours of reproductive organs, and breast cancers apparently can result from insertions of TEs within or near genes.

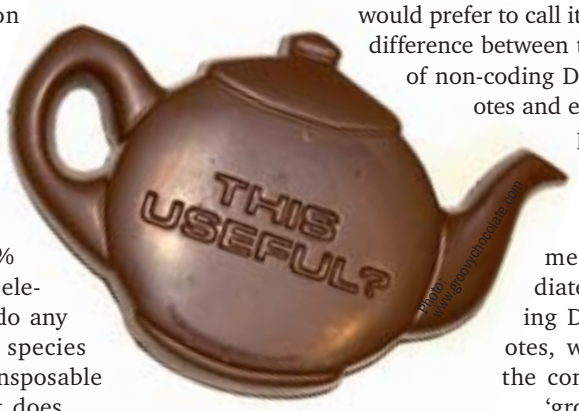
Indeed, TEs are also known to be beneficial or, at least, to have some definite, but yet unknown, meaning. First of all, the fact that many TEs are highly conserved among different species shows that they are under strong selection. Further evidence comes from a variety of fronts. For example, some TEs are differentially expressed in different tissues, yet others are respectively expressed, either in male or female reproductive cells. Certain other TEs are under suspicion to have helped in evolving RNA interference, a mechanism, by which cells destroy target RNAs from undesired DNA products. And yet another interesting group of retrotransposons from mice has been found to jump into regulatory regions of neuronal genes and alter their expression, thereby initiating the formation of different neuronal cell populations.

At the population level, TEs apparently also play important roles. By jumping around the genome at a much higher rate than any known mutation rate, they create a great deal of nucleotide diversity, which, in turn, is the raw material of evolution. Therefore, these TEs may well contribute to adaptation of species to new environments and even speciation.

More questions than answers

Christian Biémont, director of research at the National Center for Scientific Research in Lyon, France has specialised in the role of TEs in evolution and population biology. He states, "Because transposable elements can transpose at high frequency, they are more powerful as producers of the raw material of evolution than the classical base substitutions and their waves of mobilisation and loss through evolutionary times

need to be analysed in relation to speciation events. However, we are not yet able to link transposable element characteristics, such as a high copy number, high transposition or transcription activity, to the ability of populations or species to adapt better to new environmental conditions. Thus, the questions remain: Is there any relationship between evolutionary radiation and the composition of the genomes in terms of transposable elements and other repetitive sequences? Does a species with 50-70% of transposable elements actually do any better than a species with 1% of transposable elements? What does it mean for a plant or an amphibian to have more than 70% of transposable elements? Do the genomes really need transposable elements?"



Despite its reputation, 'junk DNA' is (probably) not as useless as a chocolate teapot.

More questions than answers; which is not unusual in this field of study. However, in the current age of whole genome sequencing things seem to be moving rapidly and perhaps we will see answers to these questions sooner than later.

Biémont elucidates, "Altogether, further research needs be done in order to identify the links between environmental changes, changes in host population and genome sizes, and population structuring on the one hand, and changes in epigenetic gene regulation during the early and late stages of development on the other. Understanding these mechanisms and to what extent they apply to disease is then of utmost importance. Indeed, such research, for example, is likely to have a major impact on the field of cancer, which is connected to both epigenetic processes and transposable element reactivation. Transposable elements and all the other repeated sequences could well surprise us yet again."

A modern paradox

As said before, there is obviously no correlation between the complexity of an organism and its genome size (remember the onion!); the same is also true for the number of genes (humans have less genes than some plants or even unicellular prokaryotes) as well as for the amount of non-coding DNA. There does, however, seem to be a correlation between complexity and the portion of TEs in the genome. More complex animals display a higher percentage of TEs in their non-coding DNA. But why? This question still buzzes through the minds of many scientists.

Sebastian Ahnert shares some of his ideas on the topic, "The thing I find most interesting about non-coding DNA (as I would prefer to call it) is the marked difference between the proportions of non-coding DNA in prokaryotes and eukaryotes. This probably points to the evolution of genetic regulatory mechanisms mediated by non-coding DNA in eukaryotes, which could be the consequence of a 'growth ceiling' of prokaryotic gene regulatory networks. This ceiling exists because it appears to be inefficient for prokaryotic genomes to grow above a certain size, as the regulatory overhead, which in prokaryotes is mediated entirely by regulatory genes, scales quadratically with the total amount of genetic information. In order to evolve more complex organisms, nature had to therefore devise a new way of mediating genetic regulation, through non-coding DNA."

And Christian Biémont adds, "It is now clear that transposable elements have had and have a huge impact on genome composition and genome functioning. As far as the human genome is concerned, we have mainly concentrated so far on the 1-2% of the genome comprised of protein-coding genes. We, therefore, now have to incorporate all the sequences that surround them (not only the transposable elements, which represent around 45% of the genome but 70%, if we include all sequences that derived from transposable elements that are no longer recognised as such), if we are to have an overall view of all the forces that allow genomes to define an organism, in interaction with the environment."

Too blind to see?

In conclusion, a great deal of the DNA previously described as junk DNA can be ascribed well-established functions today – there is hard evidence for it, at least. However, there are lots and lots of DNA stretches that still don't seem to be of any function at all.

Only time will tell if these sequences are, indeed, true junk or whether we are still too blind to see the real meaning of the "Dark Matter of the Genome".

Frederick Gruber

ONE FINE DAY IN THE LAB...

BY LEONID SCHNEIDER