The Evolving Definition of a Gene

KAREN HOPKIN

With the discovery that nearly all of the genome is transcribed, the definition of a "gene" needs another revision.

There was a time when a biologist could discover a new species by strolling into a field and saying, "What, ho! I've never seen one of those before!" Of course, by now biologists have seen it all—or at least all of the things that are easy to see. So a scientist who spies something seemingly new first needs to consider whether that creature belongs to a previously identified species.

> **gene\jēn*n*:** the unit of heredity that determines phenotype

But that's not as easy as it sounds. "There are people in the world of systematics who spend the entirety of their existence debating what is meant by the word 'species,'" says Laurence Hurst, of the University of Bath in the United Kingdom. "So, are two individuals members of the same or different species? It depends on how you define 'species.'"

Now the same seems to be happening with genes. "It's a slippery concept to define," says Chris Ponting, of the University of Oxford. "There's no one definition that encompasses all the objects that could be defined as being genes." In the past five years, numerous investigators using a variety of techniques have uncovered a cornucopia of ribonucleic acids (RNAs) that have excited great interest and called into question the way we think about "genes." Some RNAs, like microRNAs, regulate the expression of suites of genes. Some appear to influence the state of chromatin. Others may simply be the product of transcriptional noise—which may or may not play a role in keeping genes "readable." All lead us from our traditional genes-encode-proteins formulation of genome function.

"People have been discussing the meaning of the term 'gene' for many, many years," says Roderic Guigo, of the Center for Genome Regulation in Barcelona. "As we gain more knowledge about the molecular basis of genome activity, we should be able to more precisely define the concept of the gene. But actually it's the other way around. The more we learn, the less clear we are about what a gene is."

> **gene\jēn*n*:** the unit of heredity—located on chromosomes—that determines phenotype

Harvard's William Gelbart says, "The reality is that chromosomes are real biological objects that can be purified in a test tube.... And a 'gene' is a conceptual construct that helps us think about the individual units within that chromosome without quite knowing what they are. It's nice if people agree on the meanings of words so you know what you're talking about. But I don't think there will ever be an agreed-upon definition of a gene. I don't think there ever has been."

"When you speak with physicists, they're sometimes surprised we can use a concept that's so ill-defined," adds Laurent Duret, of the National Center for Scientific Research in France. "But in practice, when biologists talk about genes, they understand each other." What they don't yet understand, though, is what all these newly discovered transcripts are doing, and how they play into genome evolution and activity.

> **gene\jēn*n*:** the unit of heredity—located on chromosomes, that encodes one enzyme—that determines phenotype

Blame the biochemists

Genes were once defined in terms of heredity: A gene was essentially a heritable unit that produces a phenotype—the way an organism looks or behaves. "That's something we can measure," says Winston Hide, of the Harvard School of Public Health. "For example, a fungus can inherit the ability to metabolize leucine." And for population geneticists, that is definition enough: A gene conveys information about phenotype.

BioScience 59: 928–931. © Karen Hopkin. ISSN 0006-3568, electronic ISSN 1525-3244. All rights reserved. doi:10.1525/bio.2009.59.11.3

Then along came biochemists, who wanted to get physical and to attach the concept to a particular molecule and, later, to a particular stretch of nucleotides in the genome. Genes, we learned, are made of DNA. And DNA is transcribed into RNA and then translated into proteins. Genes, therefore, code for proteins. "That's the operational definition of a gene," says John Quackenbush, of the Dana-Farber Cancer Institute in Boston. "But the focus on genes being protein-coding elements was driven by a bias in our understanding of what influences phenotype. We thought that phenotype was caused by the proteins that make up the cell. I think our understanding has evolved dramatically since then."

gene\jēn*n*: the unit of heredity—located on chromosomes, made of DNA and transcribed into RNA, that encodes one enzyme—that determines phenotype

We now know that only a small percentage of the genome actually encodes protein. Yet almost all of the genome is transcribed. And many of the resulting RNAs—including the now famous microRNAs and small interfering RNAs—play a key role in regulating gene activity. "The fact that a lot of these intergenic regions and noncoding regions are active in some sense starts to make you wonder whether canonical genes are really all there is in terms of biological function," says Mark Gerstein, of Yale University.

Quackenbush agrees: "To understand how the information in the genome plays out into phenotype, the crucial element is not what encodes protein, but what regulates the complex interplay between the DNA sequence and the endpoint of phenotype." And much of that complexity—whether it's in the form of alternative splicing, epigenetic modification, or the regulatory handiwork of noncoding transcripts—involves RNA.

gene\jēn*n*: the unit of heredity—located on chromosomes, made of DNA and regulatory elements, transcribed into RNA, and translated into protein—that influences phenotype

Wild West-omics

In 2004, scientists launched the Encyclopedia of DNA Elements, or ENCODE—a project aimed at cataloging all of the functional elements in a representative 1 percent of the human genome. One of the most eye-opening findings to come from this large-scale investigation, published in *Nature* in 2007, is that "basically the whole genome is transcribed, or close to it," says Paul Flicek, of the European Bioinformatics Institute and a member of the ENCODE consortium.

"It was an absolute shocker," agrees Hurst. "Not that transcription was going on—but that so much of it was going on. The proportion of the genome being transcribed is not even 10 or 20 percent. It's way up there at 80 percent plus."

In fact, evidence for this wide-ranging transcription had already begun to accumulate. From 2000 to 2003, researchers at the RIKEN institute in Japan, as part of the FANTOM consortium, studied more than 150,000 full-length mouse cDNAs (the complementary DNA sequences that correspond to mature messenger RNA transcripts). They discovered that more than half did not have protein-coding regions and that many are present in alternative forms that begin or end in different places (published in *Nature* in 2002 and in *Science* in 2005). Others have used tiling arrays—chips spotted with probes that represent all of the nonrepetitive sequences in the genome, rather than just those containing protein-coding genes—and found a similar abundance of transcription. Several groups have started to use RNA sequencing to probe the transcriptome. "It's a much more accurate way to zoom in on the transcribed regions of the genome and to identify and quantify the messages," says Michael Snyder, of Stanford University.

These new techniques, Guigo says, "are like the microscope of this century, allowing us to look at the transcriptome at much higher resolution." Although scientists are still working to get that microscope in focus, he says, "the picture that's emerging suggests that the transcriptome is of a complexity that hadn't been anticipated. And we're seeing more and more that RNA molecules may carry out biological functions that do not involve being translated into protein."

gene\jēn*n*: the unit of heredity—located in clusters interspersed with junk DNA, made of DNA and regulatory elements, transcribed into RNA which is then spliced to produce a message that is translated into protein—that influences phenotype

The challenge, of course, is figuring out what they do. "Now that everybody accepts the existence of these noncoding RNAs, the next step is convincing people that they're not just artifacts," says RIKEN's Piero Carninci. The new RNA sequencing technologies help. "We can see the same RNAs appearing again and again, which gives us confidence that what we're seeing is not just random cleavage or degradation, but real RNAs that are made by the cell."

For example, "take small interfering RNAs and microRNAs. People used to think these noncoding RNAs were just noise," Quackenbush says. "Well, there are a lot of people who've built careers and won Nobel prizes for looking at these pieces of RNA that don't seem to encode anything. So I think we have a lot to learn about what's biological and what's artifactual. And as we continue to take these new technologies and turn them on the genome, I think we'll discover all sorts of elements that don't exist in our current catalog of genes but play an important role in biology."

"The genome is maybe not exactly the Wild West, but it certainly has elements of being a new frontier," adds Gerstein. "There's literally this vast open space of noncoding sequence. And we're using technology to get at it. Just as trains and railroads helped open up the West, new sequencing and array technologies are allowing us to explore and interrogate these uncharted expanses of the genome."

> **gene\jēn*n*:** the unit of heredity—located in clusters interspersed with junk DNA, made of DNA and regulatory elements, transcribed into RNA which is then spliced to produce a series of messages that can then be translated into a family of related proteins—that influences phenotype

Slippery "scripts"?

Genes were tricky to define even before the discovery of this rampant transcription. For example, many mammalian genes are regulated by elements that can lie a megabase away from the protein-coding portion of the sequence. Are they part of the gene? And what about alternative splicing? "If you have two transcripts and one has a longer 5-prime end than the other, do they belong to the same gene?" Hurst asks. "What if you have two transcripts that come off the same bit of DNA, but the mature transcripts yield proteins that do not share a single amino acid in common because one of the transcripts is entirely intronic to the other? Are they different versions of the same gene? I don't know. But I think you have to establish a set of standards for deciding which bits belongs to the same gene."

The problem will continue to grow more complex. "In flies, we have a gene that's expressed from both strands," says Gelbart. "So you have alternate splicing that somehow involves the opposite strand." Some researchers have even identified chimeric transcripts—RNA molecules that include sequences from different parts of the genome glued together to make something entirely new. "Is that one gene? Two genes? Who knows?" Hurst says.

Investigators responsible for annotating genomes have come up with their own working definition. "For Flybase, our rule is that any two transcriptional units that include at least one amino acid codon in common are part of the same gene," Gelbart says. "It's a totally arbitrary definition that none of us will defend in any way other than to say at least we can compute it."

The folks at Ensembl, the UK-based database of eukaryotic genomes, have also focused on transcripts. "We want to find the transcription start site and end site," Flicek says. "And we're happy to call that a gene." Two transcripts are part of the same gene when they share at least one exon. In which case, he says, "the gene begins where the first transcript begins and ends where the last transcript ends."

> **gene\jēn*n*:** the unit of heredity—located in clusters interspersed with junk DNA, made of DNA and regulatory elements, that encodes a protein or family of proteins or a functional RNA molecule—that influences phenotype

That approach, Hurst says, seems "philosophically clean. Because a transcript is a thing you can unambiguously identify, whereas a gene may have multiple transcripts, so it's a more flexible entity."

Clarifying genes' boundaries would presumably also make them easier to count. "How many genes are there in the human genome? There's no single number you can find anywhere," Guigo says. "Eight years after the completion of the human genome sequence, nobody is saying there are 19,723 genes. The fact that we don't have a number is not only because genes are difficult to find but because we don't have a clear idea what a gene is. If we don't know what we're counting, it's very difficult to count."

And that's not a problem that's unique to the human genome. "The yeast genome was sequenced in 1995," Quackenbush says. "How many genes are there in yeast? Nobody can answer that question. For protein-coding genes we can get a reasonable approximation," he says. But for RNAs, new transcripts are still being discovered. "And I think we can't ignore them, because they're something that yeast is expending energy to create. And if a yeast is making that investment, we should probably make the investment to understand why it bothers, why these things are important."

Messy or precise?

So what are all these RNAs doing? "That's a profound biological question and one which I think is absolutely core to our understanding of how genomes operate," Hurst says.

Some people still think they might be noise. "We know that the cellular machinery for gene expression makes errors," adds Duret, "and we know that these errors must be quite frequent, because cells have evolved mechanisms for detecting and degrading aberrant transcripts." For example, nonsense-mediated and nonstop decay systems get rid of transcripts that contain premature stop codons or no stop codons. So it could be that there's just a ton of low-level transcription going on throughout the genome. Or, as someone in the "just noise" camp might put it: "The transcriptional machinery exists in the cell, DNA is in the cell, so these things happen," explains Flicek, who does not actually subscribe to the philosophy.

But some variants clearly do have biological functions, as evidenced by the activities of microRNAs. "So there are two models," Hurst says. "One, the world is messy and we're forever making transcripts we don't want. Or two, the genome is like the most exquisitely designed Swiss watch and we don't yet understand its workings. We don't know the answer—which is what makes genomics so interesting."

It could also be that we're asking the wrong question, Ponting says. "If

decades ago, someone asked, 'What do proteins do?' you wouldn't expect one answer," he says. "Similarly there's likely to be a gamut of functions these RNAs are performing."

To assess those functions, scientists could fall back on some of the tried-and-true techniques used to determine what proteins do: for example, knocking out their "genes" and asking whether that perturbation has some measurable effect on phenotype. Of course, the same caveats apply to that approach as do to studying protein-coding genes. "Some knockouts don't give you a phenotype, perhaps because we're not using the right assay," Snyder says. "So just because you haven't found a function doesn't mean there isn't one."

A similar argument can be made for searching for sequence conservation. "Conservation imputes function, but lack of conservation does not impute lack of function," says John Mattick, of the University of Queensland in Australia. "And regulatory sequences evolve very quickly. That's where evolution plays. There are only so many ways to make a wheel—or an oxygen-binding pocket. And we have largely the same set of protein-coding genes as nematodes. So if you have a common set of protein components, where you see the evolutionary innovation is in the regulatory circuitry."

Flicek agrees. "If the only things that are functional in human and mouse genomes are things that are conserved, there shouldn't be much difference between humans and mice," he says. "So there must be functional regions that are not conserved."

If proteins are the skeletons of life, these noncoding RNAs "could be like the decorations that make one species different from another," Guigo adds. Or, as Mattick puts it: "Proteins are the analog components of the system. And these RNAs are the computational engine. So the genome is not just oases of protein-coding sequences in a desert of junk, but rather islands of protein-coding sequences in a sea of regulation, most of which is transacted by RNA. So it's an RNA world after all." That added level of RNA-driven regulation, he adds, "almost certainly explains the difference between more complex organisms and less complex organisms. And it almost certainly provides the infrastructure required for programming both development and cognition."

Visit these Web sites for more information:

www.nature.com/nature/focus/encode

www.genome.cshlp.org/content/17/6/682

www.genome.cshlp.org/content/17/6/669

The name game

Whether the DNA that encodes these RNAs can be considered a gene is still a matter of debate. Perhaps it depends not on whether they function but on how. "A region of DNA that is transcribed and produces an RNA molecule that is, in itself, functional—that's a noncoding RNA gene," Duret says. But in the yeast genome, there exist regions for which it seems the transcript is less important than the act of transcription. In this case, the production of transcripts, and the binding of the transcriptional machinery, appears to keep the chromatin in an open state, to the benefit of the essential genes that rest nearby. Duret prefers not to call those sequences genes, but Hurst says, "I don't care what you call them, as long as we can understand what they're doing and why."

In some cases, a little fine-tuning should help to clarify, at least at the level of semantics. "If we're talking about protein-coding genes, we need to say protein-coding genes," says Duret. Similarly, people may refer to "gene loci," which includes all the transcripts that spring from a particular stretch of DNA. They can also talk about transcripts or transcriptional units. "One can always find exceptions to any definition," notes Snyder. "That doesn't mean definitions aren't valuable. I mean, we've got to call this stuff something."

Naming the bits of DNA that encode all these newly discovered transcripts may encourage people to pay them attention. "Fifteen years ago, I remember screening a library and coming up with five protein-coding genes and one piece that was noncoding," Carninci says. "All the students and postdocs picked up the protein-coding genes, and the noncoding one was left aside." That snippet was neglected for 10 years before if was found to code for a short, functional RNA. "This is why it's important to count these noncoding genes, to classify them," Carninci says. "So students and postdocs and PIs and funding agencies will start to recognize that these things are a very important part of the genome."

gene\jēn*n*: the unit of heredity—made of DNA or RNA, that encodes a coherent set of potentially overlapping functional product molecules, either protein or RNA—that influences phenotype in ways we may or may not be able to measure

In the meantime, biologists will continue to explore, much as they've always done. "Whatever a gene is, it's vastly more complicated than something that gets transcribed and makes a protein," Hurst says. "If you started out with the hubris to think that the only bits of the genome that are important are the ones that make protein, you were wrong. That's one thing that genomics has taught us." As for what to call these things, he says, "I think we'll carry on being pleasantly ambiguous. We'll leave the semantic niceties for the philosophers, and we'll get on and do the science."

Karen Hopkin (e-mail: khopkin@nasw.org) is a freelance writer based in Somerville, Massachusetts.