GENOMICS RESEARCH

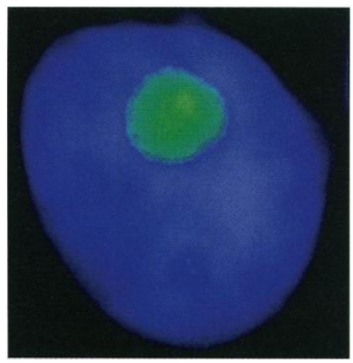
Sifting Through and Making Sense of Genome Sequences

COLD SPRING HARBOR, NEW YORK—The genomics community's annual meeting on Genome Mapping, Sequencing, and Biology once focused mostly on developing better mapping and sequencing technologies. But at this year's meeting, held here from 13 to 17 May, participants were also caught up in trying to make sense of the sequence data being produced.

Tracking Down Invisible Genes

To function normally, the cell needs genes that code for more than proteins. Also scattered throughout the genome are genes encoding RNA molecules that play a variety of roles in the cell, including assembling and maintaining the ribosomes, small particle-like structures where proteins are synthesized. These RNA genes don't have the telltale resemblances that mark protein-coding genes, making them hard to spot. New results presented at the genome meeting show that computer algorithms are getting much better at picking these genes out.

Programs called BLAST or FASTA routinely identify new protein-coding genes by comparing them with known genes, looking for sequence similarities. But even related RNA genes can have very different sequences, because RNA functions depend on how they bend and twist, not on their exact base sequences. The bending does depend, however, on which of the RNA's bases pair up—and recognizing the pairing pattern was key to the new computer program, devised by Sean Eddy and graduate student Todd Lowe of Washington University School of Medicine in St. Louis. With it, they tracked down almost all of the 40-plus yeast genes for a group of RNAs involved in building the ribosome, the so-called small nucleolar RNAs (snoRNAs).



G. MATERA AND M. FREY/CASE WESTERN RESERVE UNIV.

SnoRNA visualized. A computer program has tracked down the genes for snoRNA molecules (green) that congregate in the nucleolus, a compartment within the nucleus (blue).

The basic strategy dates to 1994, when Eddy and, independently, David Haussler from the University of California, Santa Cruz, published two mathematical formulae capable of recognizing RNA pairing patterns. At the time, "that algorithm was a theoretical proof of principle and had no practical problem to be applied to," Eddy recalls.

But in 1996, Lowe became intrigued with the snoRNAs because his journal club had discussed several research papers about these then hard-to-identify molecules. For example, one group of snoRNAs in yeast, humans, and other organisms chemically modify ribosomal RNA by adding methyl groups to particular bases. Researchers don't know what these methyl groups do, but based on the number that get attached, they thought about 55 different snoRNAs should be involved in the modifications. At the time, however, they had confirmed the existence of only one.

Eddy advised Lowe against using the algorithm for finding the rest, thinking that snoRNA researchers already had a head start using other computational and experimental approaches. Indeed, in the 2 years before Lowe undertook the project, other researchers went on to identify 20 apparent snoRNA genes. But he went ahead anyway, using the sequences of those genes to "teach" the computer program what to look for.

Lowe turned up 41 snoRNA genes, 22 of which were new. By knocking out those genes in yeast and determining which ribosomal RNA bases did not get methylated, Lowe and Eddy linked specific snoRNAs to 51 methylation sites. With all these genes and their methylation sites in hand, snoRNA researchers can now figure out "why all these genes are dedicated to this," Eddy says.

Six of the snoRNAs proved to be located in the introns—the non–protein-coding stretches of DNA that are interspersed between the coding regions of many genes and are often regarded as "junk." The existence of these RNA genes may explain why the intron DNA is often conserved, says Eddy: "Most people don't think about non–protein-coding DNA. But there's other stuff going on in the genome besides the protein [genes]."

Indeed, other researchers say the work highlights the potential of mathematical approaches for teasing meaning out of these mysterious parts of the genome. It demonstrates "a growing up of computational biology," says David Lipman, a computational biologist at the National Center for Biotechnology Information in Bethesda, Maryland. "[Eddy] is producing real data. Just think what [the results] will be from mouse and human."

More Ways to Score SNPs

No two human genomes are alike, and in the past few months the quest to catalog slight individual differences in the genetic code has turned into a heated race. Companies and academic labs are looking for single-base changes, also called single-nucleotide polymorphisms or SNPs. By studying these differences, they expect to nail down which genes contribute to the development of diseases such as diabetes or schizophrenia, which involve many genes. With those genes in hand, clinicians can begin to assess an individual's predisposition to those diseases based on his or her SNP repertoire.

For both these applications, researchers want fast and sensitive methods of screening thousands of DNA samples, and at the meeting several teams reported progress in achieving that goal. For a few years, researchers have been developing DNA chips that they hope will eventually be able to evaluate whole genomes at once. They now say that their current chips can look at on the order of 10,000 genes at once. In addition, some newer methods are emerging.

One team, for example, is harnessing a standard analytical tool—mass spectrometry—while another has a modern twist on an old technique for detecting DNA variations: using the "melting temperature" of DNA hybrids as an indicator of how closely related the hybridized strands are. "What we're seeing is an explosion of ideas about how to do that [detect SNPs]," says Eric Lander, who directs the genome center at the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts. "It's a sign of a pretty rich field."

The most high-tech of the new SNP-detection methods comes from GeneTrace Systems Inc., a biotech firm based in Alameda, California. The company is using so-called MALDI-TOF mass spectrometers, which can accurately detect small differences in the masses of relatively large molecules (*Science*, 27 March, p. 2044). GeneTrace's method exploits this sensitivity by transcribing the DNA sequences to be tested into copies of a fixed length. The copies should contain exactly the same sequence of bases except at the SNP. The single-base difference there gives rise to a mass difference, which the mass spectrometer can detect.

GeneTrace molecular biologist Yuping Tan reported that the company can analyze 10,000 samples a day, with an error rate of about one in 10,000—a fraction of the rate seen with DNA chips, says GeneTrace's Joe Monforte. Based on the results thus far, mass spectrometry "looks very good," says Aravinda Chakravarti, a human geneticist at Case Western Reserve University in Cleveland.

Ultimately, Monforte says, GeneTrace hopes to scale up its method to the point where the company can analyze hundreds of thousands of samples a day, the capacity needed to screen large groups of people to find all the genes involved in complex diseases such as diabetes or heart disease. Mass spectrometers are, however, very expensive, costing roughly $100,000 each. In contrast, the method devised by geneticist Anthony Brookes's team at Uppsala University in Sweden does not require such a major capital investment, although it is slower and limited to hunting thousands of SNPs at a time.

In the Uppsala team's approach, either a robot or a person places a single strand of the DNA sequence to be tested into a small well and then adds a short, single-stranded piece of DNA whose sequence is complementary to the target sequence and will thus bind—or hybridize—to it. The reaction also contains a dye that fluoresces when the two strands stick together. Then the well is heated, causing the two strands to separate and the fluorescence to disappear. The more perfectly matched the strands are, the more resistant they are to this denaturization. This allows researchers to distinguish which base is at the SNP location by monitoring the temperature at which the fluorescence fades.

"It's very simple and very robust," Brookes says. "Any lab could use it." He and his colleagues call this approach DASH, for dynamic allele specific hybridization. They are working with the British company Hybaid Ltd. to commercialize the technology.

"[The] method looks quite promising and is really clever," says Chakravarti. But he and Lander also realize the race to build the best SNP-scoring technology has only just begun. Says Lander: "It's too early to tell what the best way to do it will be."

A Mosaic of Gene Duplications

As the archive of life, the genome has long been considered a relatively permanent record. But increasingly, molecular biologists are realizing that the genome undergoes constant remodeling as extra copies of genes and surrounding DNA quietly sneak in and out. Two presentations at the genome meeting, one by Evan Eichler of Case Western Reserve University and the other by Julie Korenberg of Cedars-Sinai Hospital in Los Angeles, showed just how common these extra gene copies are.

Both researchers have found that the genome is larded with extra copies of small chromosomal segments. "We are a mosaic of duplications," Korenberg concludes. The cause of such duplications is still a mystery, but these extra genes and their surrounding DNA provide fodder for evolutionary change, because natural selection can ultimately put them to new uses. Eichler and Korenberg also suggest that it may be possible to get a handle on difficult problems in evolutionary biology, such as the nature of the primate family tree, by comparing the patterns of gene duplications in different species.

Eichler first began to appreciate the dynamic nature of DNA about 4 years ago when he and his colleagues tried to label a specific piece of the X chromosome. They were using a fluorescent dye attached to a DNA sequence that they expected would bind specifically to that region. But they found that the label also wound up on a spot on chromosome 16. When they took a closer look at the labeled chromosome 16 DNA, they discovered the explanation: It contained a copy of an X chromosome gene. "The entire gene, including the regulatory and structural sequence [surrounding it], had moved to a new location," he recalls.

E. EICHLER

Surprising duplications. The genome proved to contain many copies of the adrenoleukodystrophy gene (red).

That wasn't the only patch of chromosome 16 that appeared to be copied from elsewhere in the genome. When Eichler and his colleagues sequenced about 845 kilobases of the DNA, they found 17 segments that seemed to have come from other chromosomes, ranging in size from 5000 to 50,000 bases. Among these were a creatine-transporter gene, the gene that is mutated in a rare human hereditary disease called adrenoleukodystrophy—made famous in the movie *Lorenzo's Oil*—and parts of the immunoglobulin genes, he reported.

These aren't one-time duplications, the Case Western group has found. The adrenoleukodystrophy gene, for example, appears in four other places in the human genome as well. "Evolutionary biologists talk a lot about [the role of] gene duplication," says Howard Jacob, a molecular geneticist at the Medical College of Wisconsin in Milwaukee. "[Eichler] has evidence of [such] duplication." By comparing 150,000-base chunks of DNA from different parts of the human genome, Korenberg too has observed that such duplications are common.

Eichler finds they often turn up near the chromosome's centromeres, the pericentromeric region. To him, the work suggests the importance of these regions as "both graveyards and factories of evolution."

Korenberg's survey has examined other parts of the chromosomes as well. She has noticed that often the duplicated patches exist at inversion sites, where a piece of DNA breaks out of the chromosome and then sneaks back in upside down at that same site or at a different site. Once a duplication or inversion has occurred, she suggests, the DNA "is predisposed to further [changes] at those points," changes that may lead to human disease. Often she finds two or more copies of that piece there.

Whatever causes the duplications, they may enable scientists to chart evolutionary relationships. For example, because many primates are so closely related, molecular evolutionists have been unable to sort out the details of the primate family tree based on single-base differences in their genes. But both Eichler's and Korenberg's work suggest that tracking these much larger DNA changes may help. Eichler and his colleagues found that chimps have two copies of the adrenoleukodystrophy gene, neither of them on the same chromosomes as the human genes, while gorillas have several copies, some of which correspond to the chimp duplications while others resemble the human ones. In accordance with the current leading model, this suggests that a gorillalike ancestor gave rise to both chimps and people, with both descendants losing some of the duplicated sections of DNA as they evolved.

Similarly, Korenberg has been studying differences between the primates at the DNA inversion sites, looking to see how many copies of duplicated patches exist in different species. Because these duplicated genes are more likely to evolve, she expects to find some of the genes that distinguish chimps from people, say, or gorillas from chimps. Indeed, says evolutionary biologist James Lake of the University of California, Los Angeles, Eichler's and Korenberg's data "may solve this long-standing problem of the relationship of humans, chimps, and gorillas."

–Elizabeth Pennisi