The digital generation

Next-generation sequencing is pushing gene-expression profiling further into the digital age. But analog methods still have plenty of wind left. Nathan Blow looks at the looming battle over the cell's transcriptome.

Could it be only a matter of time before all gene-expression analysis goes digital? The answer is less straightforward than you might think. When it comes to studying a cell's transcriptome — the collection of all RNA transcripts produced at a specific time — next-generation sequencing promises insights into the way genes are expressed and regulated in cells. Compared with analog methods such as microarrays and real-time PCR, the technology is still in its infancy. But with availability increasing and costs declining, more researchers are turning a curious eye to sequencing's high-throughput digital readouts.

The future of analog techniques — and their possible obsolescence — was one of the questions tackled in a conference panel session held before a packed audience at last year's conference on issues in parallel sequencing at Harvard University Medical School in Boston, Massachusetts. The panel explored the changed landscape of gene-expression analysis since the arrival of next-generation sequencing, and feelings on the subject in both panel and audience were mixed.

"Sequencing allows you to ask many different types of questions and look at transcription from new angles," says Shirley Liu, an associate professor at the Harvard School of Public Health and Dana-Farber/Harvard Cancer Center in Boston, who organized the conference and led the panel. But she thinks sequencing has a long way to go before it reaches the level of adoption that microarrays have among researchers. "My feeling is that if people have never done this kind of work and are trying things for the first time or are asking traditional questions, they are better off with arrays now," she says.

By the numbers

"These days researchers are stepping into a new world when looking at sequencing," says Shawn Baker, a senior product manager for RNA analysis at Illumina in San Diego, California. Although the 'big splash' for most new next-generation sequencing systems comes from deciphering eukaryotic genomes at a breakneck pace, Baker says that digital gene expression and RNA profiling are increasingly popular. He suggests that the company's Genome Analyzer is used for this type of work 30–40% of the time — although he adds that when the major genome centres are subtracted from the user mix that percentage is likely to be much higher.

Four companies now provide next-generation sequencing platforms that support, or plan to support, digital gene-expression applications, but that number is widely predicted to rise in the coming years as developers work on 'next-next-generation' single-molecule sequencing systems. For now, Applied Biosystems (part of Life Technologies) in Foster City, California, recently released a whole transcriptome-expression kit along with a small RNA expression kit for use on its SOLiD 3 platform. Illumina has developed a digital gene-expression kit for RNA analysis and profiling for use on its Genome Analyzer; and 454 Life Sciences, a Roche company located in Branford, Connecticut, uses a serial analysis of gene expression (SAGE) tag-based approach for expression profiling on its GS FLX system.

This level of development and interest among researchers should come as no surprise: gene-expression analysis is an application that seems ideally suited to next-generation sequencing. "For digital gene expression you are just counting the number of times you hit a gene and then assuming that that represents the number of copies of the transcript that you have in your population," says Chad Nusbaum, co-director of the genome sequencing and analysis programme at the Broad Institute and Harvard University, both in Cambridge, Massachusetts. So the more you can count — and next-generation systems can count a lot — the better the measure of copy number for even those rare transcripts in a population.

J. RINN

John Rinn is working on the biological functions of large non-coding RNAs.

"People are sequencing anything from 10 million to 40 million reads in a run," says Baker, "and they are getting a phenomenal level of data." Last year, short-read high-throughput sequencing, or 'RNA-seq', was used to obtain global views of gene expression in human embryonic kidney and B cells[1], to profile transcription in mouse embryonic stem cells[2] and to quantify whole mouse-cell transcriptomes[3].

Those studies also highlight the other benefit that comes with a sequencing-based approach. "Sequencing is the best way to truly profile all aspects of sequence variation," says Roland Wicki, director of Applied Biosystems' SOLiD application development. In addition to looking at RNA expression patterns, RNA-seq can allow researchers to discover new classes of RNA, detect point mutations in expressed transcripts, identify fusion transcripts or uncover new alternative splicing events — discovery applications not possible with other technologies. Using the SOLiD 3 platform, Nicole Cloonan of the Institute for Molecular Bioscience at the University of Queensland in St Lucia, Australia, and her colleagues identified more than 2,000 expressed single nucleotide polymorphisms (SNPs) from embryonic stem cells and determined that the RNA-seq approach could detect 25% more genes than analog microarrays. Ali Mortazavi of the California Institute of Technology, Pasadena, and his

ILLUMINA

Next-generation sequencing platforms such as the Illumina Genome Analyzer are allowing scientists to explore the transcriptome in greater detail than ever before.

APPLIED BIOSYSTEMS

Applied Biosystems has released a whole-transcriptome-expression analysis kit for the SOLiD 3 platform.

colleagues used Illumina's Genome Analyzer to find 3,500 genes that showed one or more alternative splice forms.

Wicki says that when it comes to the identification of point mutations in transcripts or analysis of allele-specific expression, next-generation sequencing platforms such as the SOLiD 3 system, which uses a unique ligation-based sequencing approach that allows for two-base encoding, tend to be highly accurate. Still, the real value of such systems lies in the ever-increasing numbers of sequencing reads they can generate.

Enriching discovery

Sequencing today is allowing researchers to increase the dynamic range of their investigations simply by increasing the number of sequencing reads analysed. There is a cost to this, though. "High-throughput sequencing of RNA does take a lot of reads to get depth," says John Rinn, an assistant professor at Beth Israel Deaconess Medical Center in Boston, Massachusetts. "So you end up sequencing highly abundant RNAs over and over again before getting to low-abundance stuff." Although this is necessary when it comes to profiling gene-expression patterns, it has led several researchers, including Rinn's group, to develop new methods to enrich for specific populations of RNAs from the transcriptome (see 'Rethinking junk DNA') when it comes to discovery applications.

Several new enrichment procedures focus on eliminating the abundant amounts of ribosomal RNA from a pool of total RNA before the library-generation step of RNA-seq. Invitrogen, part of Life Technologies in Carlsbad, California, recently introduced the RiboMinus kit for RNA-seq applications. The kit depletes ribosomal RNA from a sample by binding the ribosomal RNA to probes containing locked nucleic acids, which are then separated from the sample using magnetic beads.

Another enrichment approach was developed by Evrogen of Moscow, Russia. This uses a duplex-specific nuclease for normalization of RNA transcript levels. Following complementary DNA generation, the templates are denatured and a duplex-specific nuclease is added to the reaction. Although abundant transcripts find matches and become double-stranded, thereby acting as targets for the nuclease, less abundant transcripts take longer to find their partners and so are degraded less frequently.

With or without enrichment for specific RNA populations, the analysis of tens of millions of sequence reads can be a daunting task for most researchers, especially as analysis tools in the digital world are not as advanced as their analog counterparts.

"High-throughput sequencing data analysis is totally different from using arrays," says Liu,

RETHINKING JUNK DNA

When the complete sequence of human chromosome 22 was first published in 1999 (ref. 4), John Rinn, an assistant professor at Beth Israel Deaconess Medical Center and an associate member of the Broad Institute in Cambridge, Massachusetts, got very excited. He was not interested in looking at the map of known protein-coding genes on the chromosome, but rather everything else. "We wanted to see if we could find biologically active molecules in the human genome that no one previously knew about," he says.

Armed with the sequence of an entire chromosome — and a year later the whole human genome — researchers and developers began to create genome-wide tiling microarrays. "By probing these tiling arrays we found out that there are tonnes of biologically active regions by proxy of RNA being made," says Rinn — results he and his colleagues reported in 2003 (ref. 5). Since then, Rinn has focused his efforts on understanding a collection of these RNAs known as large intervening non-coding RNAs (lincRNAs).

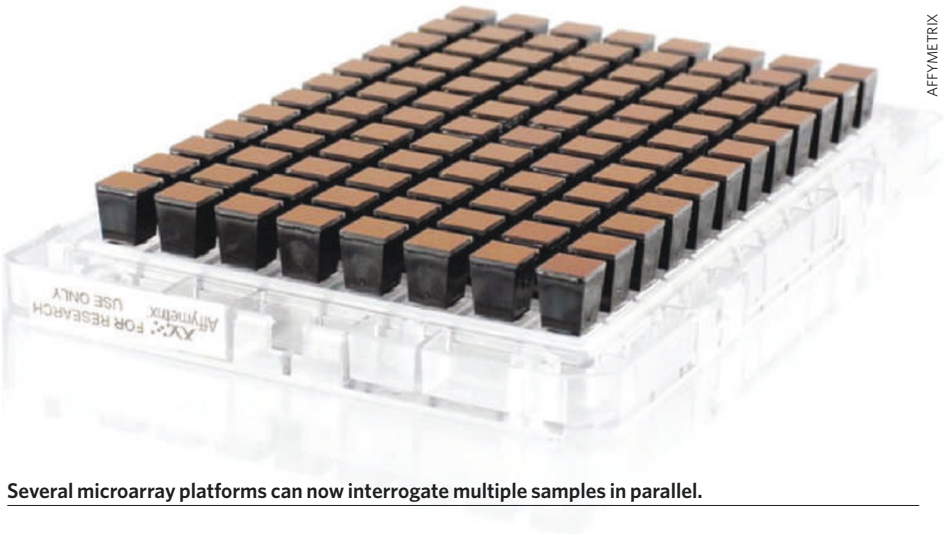
"Initially many people thought that this had to be an artefact of the technology: how could there be so many RNA molecules that we have never seen before?" says Rinn. Arguments against a true biological purpose for lincRNAs came largely from the lack of evolutionary conservation within their sequences — conservation implies function, whereas lack of conservation can often imply noise.

As so few functional lincRNAs had been described, Rinn and his colleagues set out to find more. In 2007 they reported the identification of a new 2.2-kilobase large non-coding RNA, which they called HOTAIR. It played a role in the guiding of chromatin complexes within the cell[6]. Although only a single new functional lincRNA — and still only one of four known to be functional at the time — the discovery gave Rinn an idea on how to enrich for functional lincRNAs from the genome.

"What we did next was to go after things that looked like HOTAIR," he explains. Instead of using an RNA-based approach, the group decided to look at chromatin structure. Histones have clear indications of where active genes start and stop. Using high-throughput chromatin

adding that at the moment there are few standard tools for analysing the digital gene-expression data sets generated with next-generation sequencing platforms.

"I think in the middle of last year we realized that there was a shift coming from the analog to the digital platform," says Roald Forsberg, director of scientific software solutions at CLC bio in Aarhus, Denmark, which specializes in analysis tools for high-throughput sequencing data sets. This led CLC bio to update its main and genomics software packages to support the analysis of both digital and analog gene-expression data sets. Although Forsberg suspects the shift to digital will take time, he thinks analysis tools with the ability to interrogate both sequencing and microarray data sets will remain critical to researchers for even longer. "There has been much investment in microarrays in both the academic and pharma worlds, using unique tissue samples in a lot of cases, which would be a shame to ignore or just throw away," he says.



AFFYMETRIX

Several microarray platforms can now interrogate multiple samples in parallel.

Higher-level complications

Currently, the biggest challenge for researchers looking at next-generation sequencing approaches is probably the sheer volume of data. "For serial analysis of gene-expression experiments there are some tools, but a lot of them are not well equipped to deal with the amount of tags that come out from high-throughput sequencing," says Liu. This creates the need to perform most data analysis on more powerful computer clusters. "It is not something where you can download a program to run on your laptop," says Liu.

Then there is the challenge of determining where those millions of tags or short sequences come from in the genome. "It is not uncommon for a sequencing platform to generate data and only 40–50% of the data are mappable," says Liu — meaning that the remaining could not be mapped at all or mapped to regions of the genome once considered 'junk'. And for Liu and others the question then becomes, are these biological relevant sequences — for instance, unannotated genes or antisense transcripts — or merely artefacts of the sequencing process?

"I think it is hard to say, but it is probably going to end up being a bit of both," says Jay Shendure, an assistant professor in the department of genome sciences at the University of Washington in Seattle. Shendure says that his group has performed runs on the Illumina Genome Analyzer using genomic DNA where 95% of the reads were mappable, implying that for human DNA libraries with similarly sized read-lengths, the technical artefacts are not so great as to result in only 50% mappable reads. He cautions, however, that there are additional steps in making a sequencing library from RNA that could introduce some artefacts.

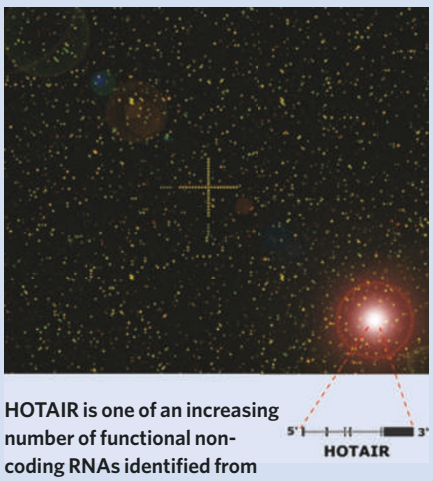
Developers at CLC bio have also experienced problems in mapping short-read RNA-seq data sets. Forsberg hopes that the future use of 'paired-end' reads — in which sequencing information is obtained from both ends of a DNA fragment — might help when mapping back to the genome. He has noticed reservations among researchers when it comes to generating either longer reads or using paired-end protocols for RNA-seq, because this increases the time and cost of sequencing.

Economic advantage

Although next-generation sequencing provides a high-throughput option to look at gene-expression profiles from a small number of samples, it turns out that in the analog world microarrays provide their own high-throughput advantage to researchers. "With microarrays you can profile RNA from a hundred different samples, which would be incredibly expensive to do *de novo* with sequencing," says Rinn.

Lower costs alongside increasing probe density on whole-genome tiling arrays for transcript-mapping applications are keeping microarrays from being lost in the blur of sequencing advances. "I think the array platform is well suited to screening studies for a quick look at the transcriptome or the general clustering of samples," says Baker. He is quick to add that the per-sample cost for microarrays has dropped significantly in recent years, with some genome-wide tiling arrays now costing

immunoprecipitation (ChIP) sequencing on the Illumina Genome Analyzer to look for these marks, Rinn and his colleagues at the Broad Institute developed genome-wide chromatin state maps. Then, just as with his analysis of chromosome 22 almost ten years ago, Rinn says he threw out the known protein-coding genes and looked at what was left. He identified 1,600 other RNAs located by themselves in the middle of nowhere in the genome that look just like HOTAIR[7].



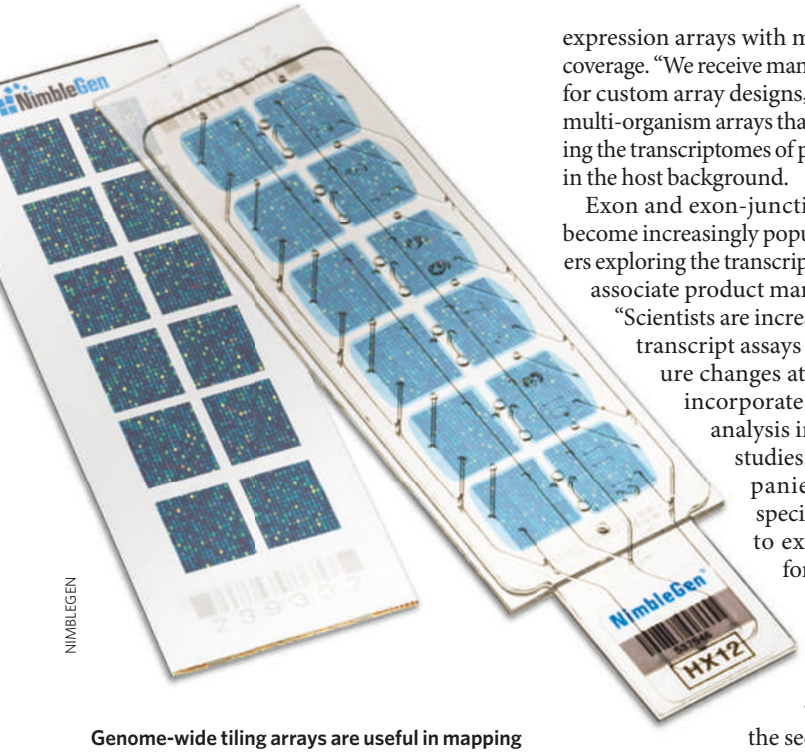
J. RINN

HOTAIR is one of an increasing number of functional non-coding RNAs identified from the human genome.

To determine if some of their newly discovered RNAs were functional, the team took a 'guilt by association' approach, using microarrays to profile a number of the newly identified lincRNAs in 21 different tissue samples while at the same time profiling protein-coding genes in the same tissue samples.

Then they asked the question: which RNAs had similar profiles to protein-coding genes of known function? Their initial analysis was followed by further validation using independent systems. "This has turbo-charged the field, as not only can we identify these things now but we can get a good idea of what they might be doing to test functional relationships," says Rinn.

For Rinn and his colleagues it is now time to muster all the force they can to explore these RNAs. "We are going to throw the Broad kitchen sink at them," says Rinn, who is teaming up with a number of scientific platforms at the Broad Institute to look at the effects of knocking down each newly discovered lincRNA. N.B.



NIMBLEGEN

Genome-wide tiling arrays are useful in mapping RNAs to specific genomic locations.

less than $100 per sample.

"Resolution is essential for tiling arrays to achieve sensitive and reproducible data, and the more closely spaced the probes are, the more likely they will detect small exons and small RNAs," says Rohaizah James, expression product manager at Roche NimbleGen in Madison, Wisconsin.

High density seems to be the order of the day. Roche NimbleGen now offers a 2.1-million-feature whole-genome tiling array to explore both coding and non-coding parts of the genome. Affymetrix of Santa Clara, California, provides tiling arrays with 6.4 million features consisting of 25 base-pair probes spaced 10 bases apart across the entire human genome for transcript mapping. And Agilent Technologies, also in Santa Clara, California, offers custom tiling arrays with several hundred thousand features.

James says that tiling arrays can be used for quick and accurate empirical annotations of transcriptomes, information that can then be applied to an expanding area of interest among researchers: building custom differential-expression arrays with more comprehensive coverage. "We receive many interesting requests for custom array designs," says James, such as multi-organism arrays that are useful for studying the transcriptomes of pathogenic organisms in the host background.

Exon and exon-junction arrays have also become increasingly popular among researchers exploring the transcriptome, says Joel Fellis, associate product manager at Affymetrix. "Scientists are increasingly using whole-transcript assays and arrays to measure changes at the exon level, and incorporate alternative splicing analysis into their expression studies." A number of companies now offer exon-specific arrays in addition to exon-junction arrays for these studies.

Having been around for some time now, microarrays also tend to offer researchers the security of a standardized technique. "I think that is a very mature tool pack, where people know what to expect," says Liu. The tools to analyse microarrays are more widely available and better developed, making analysis less complicated and easier for even novice researchers. Analog measurements come with a downside, however — one that has hounded microarrays from the start. "There is a frequent lack of inter-platform reproducibility," says Shendure, something that digital, direct-counting platforms could overcome.

J. SHENDURE

Jay Shendure thinks microarrays could migrate towards preparative uses.

Others argue the real challenge for microarrays in the coming years will be to remain up to date. "Our understanding of the transcriptome is constantly evolving, making it difficult for microarrays to stay current," says Wicki. For this reason a number of researchers expect microarrays to migrate towards more targeted applications in the future, perhaps associated with biomarker validation or other diagnostic applications. "I think in terms of basic science research or mechanistic understanding it could be that sequencing will gradually phase out arrays," Liu says.

Although microarrays continue to be the tool of choice for most researchers for gene-expression analysis, looking at copy-number variations within the genome and genotyping, some scientists see even bigger changes for this technology on the horizon. "I think sequencing will penetrate all those markets with increasing force as the cost drops," says Shendure. Although he thinks this will lead to a steady decline in the analytical role of microarrays over time, he strongly suspects their use in preparative applications will expand in the coming years.

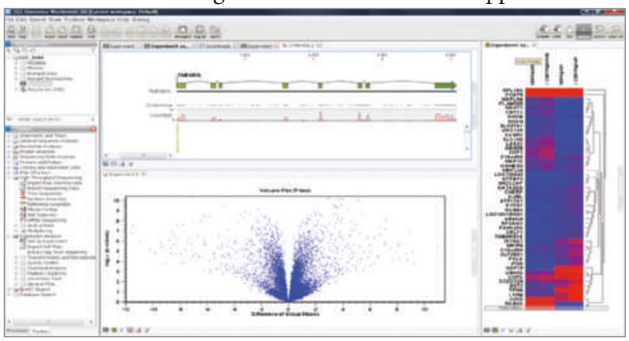
The way ahead

"Arrays will have a role as a super-cheap way of making DNA," says Shendure. At the 2009 Advances in Genome Biology Technology meeting in February at Marco Island in Florida, Shendure set up a workshop on methods to selectively capture subsets of the genome or transcriptome for high-throughput sequencing. Three of the four methods he highlighted relied on microarrays to isolate specific fragments of DNA either by hybridization of genomic libraries directly to the arrays, or by solution-phase hybridization to oligonucleotides cleaved from arrays, or by an approach that converts oligonucleotides released from arrays to inversion probes to capture exons. "We are now able to capture the whole coding genome on a single array," he says.

Developers have recently started supplying arrays for preparative applications as well. Agilent Technologies has partnered with the Broad Institute to provide the SureSelect Target Enrichment System for the Illumina Genome Analyzer, with a version in development for Applied Biosystems' SOLiD platform. The current system uses up to 55,000 biotinylated RNA probes in a single tube to capture particular segments of the genome, but future plans include an array-based partitioning tool for smaller-scale experiments. Last month, Roche NimbleGen introduced a human exome sequence capturing array to prepare samples for sequencing on the GS FLX systems from 454 Life Sciences.

In the end, most researchers think advances in next-generation sequencing for digital gene-expression analysis will continue to push microarrays towards more specific applications. Just how soon researchers leave their analog roots and take that digital plunge, however, will depend on how quickly developers can standardize and support a digital lifestyle. ■

Nathan Blow is technology editor for *Nature* and *Nature Methods*.



CLC BIO

Software programs such as CLC bio's Genomic Workbench are starting to simplify digital gene-expression analysis.

1. Sultan, M. *et al. Science* **321,** 956–960 (2008).
2. Cloonan, N. *et al. Nature Methods* **5,** 613–619 (2008).
3. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. *Nature Methods* **5,** 621–628 (2008).
4. Dunham, I. *et al. Nature* **402,** 489–495 (1999).
5. Rinn, J. L. *et al. Genes Dev.* **17,** 529–540 (2003).
6. Rinn, J. L. *et al. Cell* **129,** 1323–1329 (2007).
7. Guttman, M. *et al. Nature* doi:10.1038/nature07672 (2009).