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).

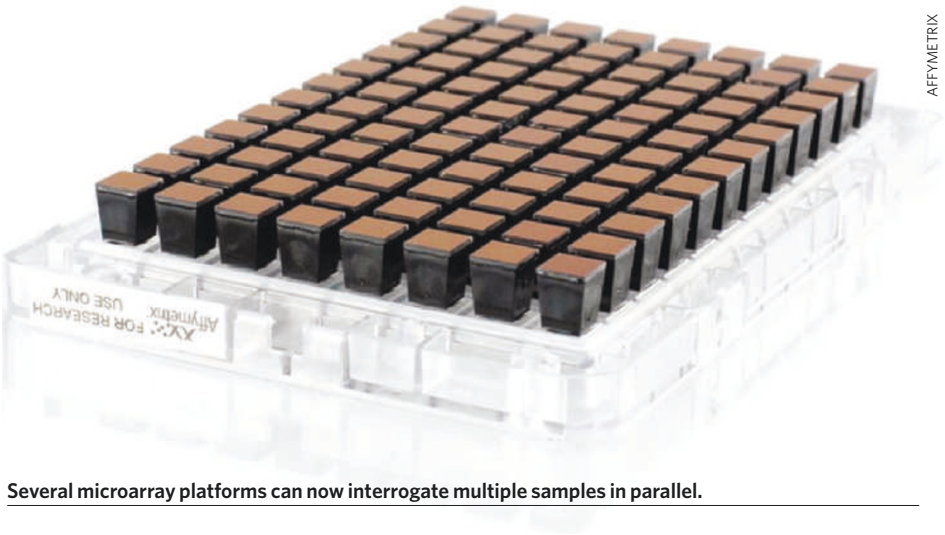
"Initially many people thought that this had to be an artefact of the technology: how could there be so many RNA molecules that we have never seen before?" says Rinn. Arguments against a true biological purpose for lincRNAs came largely from the lack of evolutionary conservation within their sequences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

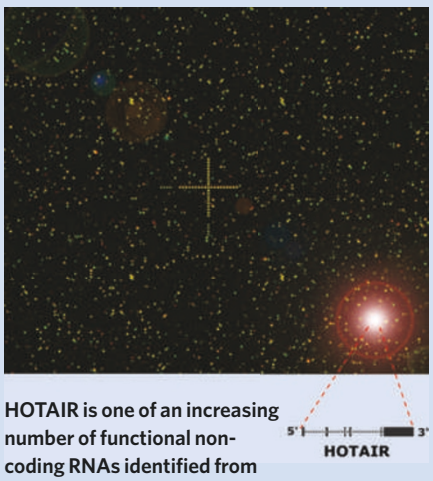
Developers at CLC bio have also experienced problems in mapping short-read RNA-seq data sets. Forsberg hopes that the future use of 'paired-end' reads — in which sequencing information is obtained from both ends of a DNA fragment — might help when mapping back to the genome. He has noticed reservations 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.