The personal side of genomics

Innovations in DNA sequencing and genotyping are opening doors for personal genomics. Nathan Blow explores these technological advances and their implications.

The era of personal genomics is upon us, with advances in technologies such as DNA sequencing and genotyping fuelling the fires. Personal genomics is a story of researchers looking for genetic clues to our most common diseases, of dazzling advances in genetic analysis technology and of lingering questions about how the public will view and use the information.

DNA sequencing is clearly driving much of this revolution in personal genomics. In late May 2007, 454 Life Sciences in Branford, Connecticut, and the Human Genome Sequencing Center at the Baylor College of Medicine in Houston, Texas, made headlines around the world with the announcement that they had sequenced James Watson's entire genome using 454 Life Sciences' next-generation sequencing technology. And just four months later researchers at the J. Craig Venter Institute in Rockville, Maryland, along with collaborators at The Hospital for Sick Children in Toronto, Canada, and the University of California, San Diego, published the first full genome sequence of a single individual — Craig Venter[1]. This analysis, though, relied on the traditional approach of Sanger sequencing.

Now, some groups are looking to take DNA sequencing and personal genomics to even higher levels. "We want to look at 100,000 genomes and rather than just look at the genomics, in which you get an idea of variation like with the HapMap, we want to actually look at the trait connected with the variation and the environment," says George Church of Harvard Medical School in Boston, Massachusetts, and founder of the Personal Genome Project (see 'Being well informed').

M. NEMCHUK, BROAD INST.

The Broad Institute's Chad Nusbaum uses several next-generation sequencing systems in his research.

When first conceived in 2003, the Personal Genome Project faced numerous challenges, not least that the technology required to meet its goals was not even available. But technology is catching up with ambition, and advances in DNA sequencing are making it possible to decode individual genomes much faster, making endeavours such as the Personal Genome Project more feasible.

Sequencing's new wild west

A new generation of faster DNA-sequencing systems has exploded onto the genetic-analysis scene, with at least five companies offering or preparing to offer sequencers that boast amazing output. Several sequencers produce upwards of a billion bases of raw data per run — the equivalent of one-third of the human genome.

But Chad Nusbaum, co-director of the genome-biology programme at the Broad Institute in Cambridge, Massachusetts, is quick to point out that these new systems

BEING WELL INFORMED

G. CHURCH

George Church in his Personal Genomics Project picture with a ruler to measure facial features.

George Church, of Harvard Medical School in Boston, Massachusetts, is working on a project that he thinks could change the landscapes of both genomics and medicine. His Personal Genome Project aims to integrate data for genomics, environment and phenotype in more than 100,000 volunteers.

Perhaps the biggest issue for the project is how to acquire informed consent from so many participants, who will have their data become publicly available. This is something that Church hopes the Personal Genome Project will be able to address. He comments that some researchers think that this level of scale-up is not compatible with the current rules of informed consent.

From the start, participants will need to know what they are volunteering to do. "We are trying to emphasize to participants that rich holistic genetic and trait data are going to be obtained," says Church. Therefore, he says, participants must be aware of the risks and benefits of modern genetics and of the fact that modern digital information has various ways to get into the public domain.

To make certain that everyone involved clearly understands these points, the project will establish an online educational system targeted to these topics. All participants will need to pass a test to see how well they understand them. After this education and evaluation, the project will allocate participants 'set points', which will determine how much information should be released to them.

Additionally, Church says, participants will have access to data only for validated genes for which we know something about their action or disease risks or for which treatment is possible. Although that list is short now, he expects that it will grow quickly.

Still, some researchers are concerned about the risks of providing patients with genetic data. "I think there is a danger with a lot of the new and best technologies that it is tempting to provide sequence information to patients before the biological implications of those data are known," warns Bert Vogelstein of Johns Hopkins University in Baltimore, Maryland.

In August 2005 the project received approval from Harvard Medical School's institutional review board to start enrolling its first ten participants. The criteria were stringent: participants had to have at least a master's-level education in genetics or an equivalent understanding of genetics research. Church and others involved in the field were among the first to volunteer because they were thought to be the best informed to give consent.

Church thinks that now is the time to resolve the issues of informed consent because the technology has arrived to make personal genome sequencing a reality. He points out that companies such as 23andMe in Mountain View, California, and DNAdirect in San Francisco, either already offer personal genetic testing services or plan to in the future. "We have to get this in place before everything just goes crazy," says Church. N.B.

are in fact quite different from one another and at various stages of maturity. "The principle that we use when applying these new technologies is that there is a lot of expensive sequencing that we do with Applied Biosystem's 3730xl system and anything that we can move over to the new technologies, as long as it is effective, is bound to be cheaper." The systems available now from Roche, Illumina and Applied Biosystems do seem to be effective, as the Broad Institute and other organizations are using them for various sequencing-based applications.

Assembling the future

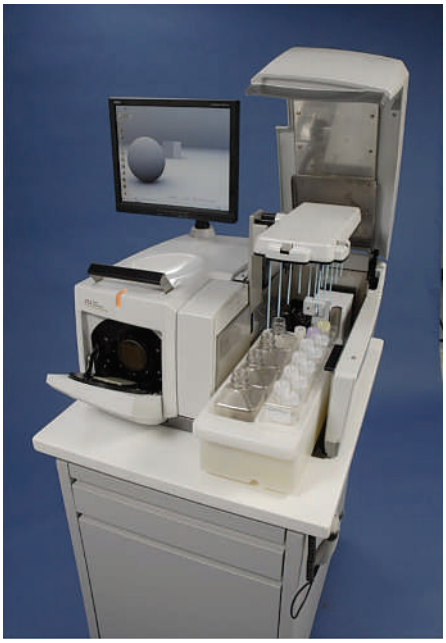
By the end of last year, 454 Life Sciences, which was founded in 2000 and was recently acquired by Roche, had more than 60 of its sequencing systems placed around the world. "Our technology is in all major US genome centres and some of the international centres," says Michael Egholm, vice-president of research and development at 454 Life Sciences.

The technology developed by 454 Life Sciences is based on two fundamental principles: emulsion PCR and pyrosequencing. Emulsion PCR side-steps the conventional process of bacterial cloning by attaching fragments of DNA 300 to 500 base pairs long to beads *in vitro,* then amplifying them with PCR to make millions of identical copies. Pyrosequencing allows for a massive parallel reaction format done in 1.6 million wells on a PicoTiterPlate. "Right now, day in and day out, we can perform 400,000 reads of 250 bases each with an accuracy of 99.5% or better," says Egholm. Although the 454 Life Sciences system is not as accurate as conventional Sanger sequencing, Egholm notes that it is an order of magnitude more productive (see 'Truth and accuracy').

This upgraded Genome Sequencer FLX System allows more sequencing cycles and therefore longer reads than the previous Genome Sequencer 20 System. Longer reads help in whole-genome sequencing and assembly applications. "We believe that shortly there will be many more *de novo* assembled genomes due to our technology," says Egholm. He notes that the genomes of several microorganisms have been assembled from scratch by use of 454 sequencing, and the technology has also been used to supplement Sanger sequencing on a few projects involving larger genomes.

At the Broad Institute, where researchers use two FLX systems and one Genome Sequencer 20, Nusbaum appreciates the ease of the 454 sequencing process. "It is nicer than Sanger sequencing because it is a faster and simpler process." He points out that at the Broad Institute, sequencing a bacterium can take a month with Sanger methodology, whereas with 454 technology it can be done in a week and without the high degree of clone tracking associated with Sanger sequencing.

Still, for *de novo* sequencing and to assemble larger genomes, such as those of mammals, longer paired reads — that is, two reads that are a known distance apart — will be necessary — an issue that Roche and 454 are trying to address. "Whether [454 sequencing] will work with a mammalian genome is a good question, and it is a little way off," says Nusbaum. But he optimistically notes that 454 Life Sciences has exceeded his expectations in surmounting several other technical hurdles. Egholm, however, is much true biological diversity, or just errors caused by the new sequencing technology. "We had to explore just how good the sequencing technology actually was," he says.



454 LIFE SCIENCES

The Genome Sequencer FLX, developed jointly by 454 Life Sciences and Roche Applied Sciences, is based on 454 sequencing technology.

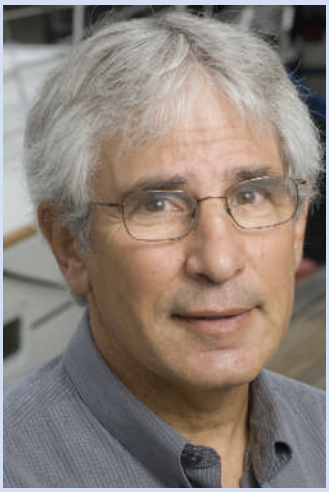
TRUTH AND ACCURACY

Mitch Sogin, director of the Josephine Bay Paul Center for Comparative Molecular Biology and Evolution at Woods Hole Marine Biological Laboratory in Massachusetts, performs environmental sampling of nucleic-acid sequences. "Every sequence has the potential to tell us an important story," says Sogin, so highly accurate analysis techniques are needed.

But when Sogin's lab switched over from traditional Sanger-based sequencing to the next-generation sequencing system of 454 Life Sciences in Branford, Connecticut, to study these environmental samples, something strange happened. "The diversity was between ten- and a hundred-fold more divergent than we expected," recalls Sogin.

So Sogin and his colleagues needed to determine whether the unexpected findings represented

M. SOGIN



Mitch Sogin tested the accuracy of 454 sequencing.

Sogin and his colleagues did a straightforward experiment in which they resequenced more than 50 templates and cloned sequences on a 454 Life Sciences Genome Sequencer 20 that they had sequenced previously with the Sanger methodology. The work showed that the 454 system was 98% accurate if no culling was used to remove bad bases or reads[6].

However, by using a very simple set of rules, which caused somewhere between 10% and 20% of the data to be discarded, the accuracy could be pushed up to 99.75%. And discarding up to 20% of the data for this level of accuracy is a trade-off that is fine with Sogin because the latest 454 system — the Genome Sequencer FLX — can produce up to 400,000 reads per run.

Others agree that for some applications the large amount of data generated by the next-generation sequencing systems could trump the accuracy produced through Sanger sequencing. "For applications such as CHIP-sequencing you can use the 454 or Solexa 1G data even though they have lower base accuracy because you do not need it. What you need is the volume for the experiment," says Chad Nusbaum of the Broad Institute in Cambridge, Massachusetts.

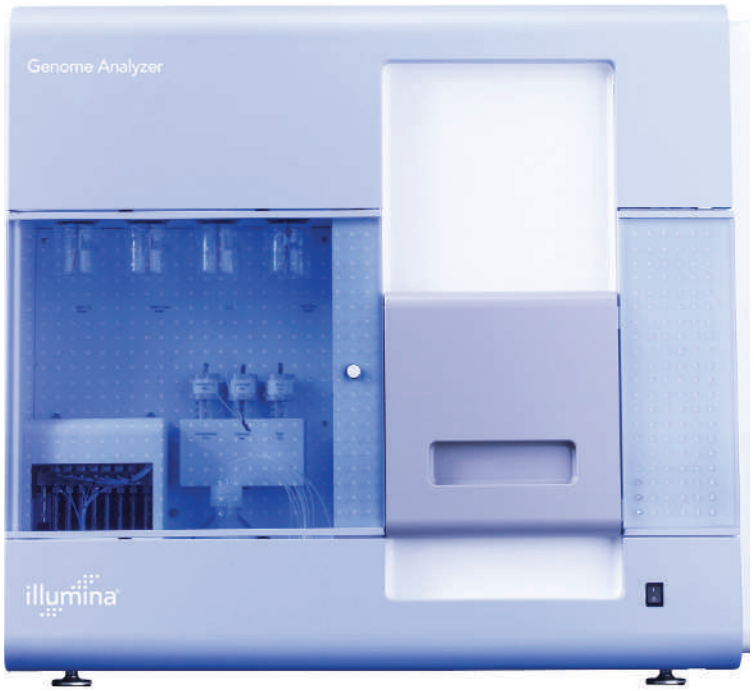
Sogin now thinks that traditional sequencing methods had been underestimating the biological diversity of the environmental nucleic-acid samples. "Turns out that the diversity is coming from low-abundance nucleic-acid populations that you are not likely to encounter if you sequence only a few hundred molecules. You see these low-abundance molecules only if you sequence many tens of thousands of molecules," he says. N.B.

more direct in his vision for the future of 454 sequencing. "My goal is simple, I want to displace Sanger sequencing for *de novo* sequencing."

Counting games

Like the Genome Sequencer FLX system, both Illumina and Applied Biosystems have used emulsion PCR as a starting point for their next-generation sequencing systems. But from there the methods of sequencing are quite different from each other.

In January this year, Illumina, located in San Diego, California, acquired the Hayward-based firm Solexa. Solexa's key technology, previously called the Solexa 1G and now named the Genome Analyzer, is a next-generation sequencing system that can sequence the equivalent of a third of the entire human genome in a single run. The Broad Institute now uses 16 Genome Analyzers for various projects. "Any application that is counting-related is a very good one to perform using Illumina's system," says Nusbaum.



ILLUMINA

The Illumina Genome Analyzer has been used for several CHIP-sequencing projects.

Nusbaum and his colleagues, along with other groups, have already demonstrated the usefulness of the Genome Analyzer in looking at patterns of chromatin structure by using chromatin immunoprecipitation[2–5]. "It is an incredibly powerful application of the technology," says Nusbaum. By pulling down DNA bound to histones carrying specific modifications, sequencing it and mapping it back to the genome, they could map the status of chromatin across the genome and throughout development. Nusbaum adds that it is also an incredibly easy application of the technology and anticipates that Illumina instruments will be used for other applications such as transcriptional profiling or microRNA and small RNA discovery. "It is also a great way to identify polymorphisms in genomes that are not extremely different."

Applied Biosystems in Foster City, California, is rolling out its new sequencing by oligonucleotide ligation and detection, or SOLiD, system in October 2007. The target is to cover a whole human genome in one run, says Kevin McKernan, senior director of scientific operations at Applied Biosystems. McKernan says that in-house, the SOLiD system has been obtaining around a gigabase more data than their target, achieving 4 gigabases of sequence per run that aligns to the target genome, and 8 gigabases overall. McKernan thinks that with advances in the PCR process over time, this will turn into 8 gigabases of sequence that aligns to the target genome. The SOLiD system differs from other next-generation sequencing systems

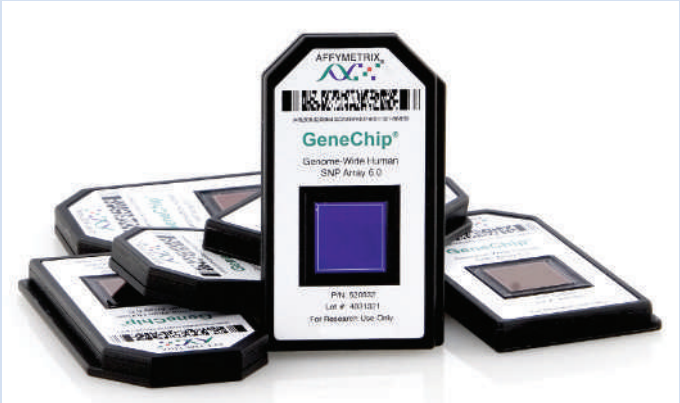
CHIPPING OUT OUR DIFFERENCES

Single nucleotide polymorphism (SNP) genotyping is a method for determining genetic variation. As more and more SNPs have been identified from the genome in recent years, the power of this technique has steadily increased.

Affymetrix, located in Santa Clara, California, is one company that is taking advantage of SNP-discovery projects, such as the International Hap Map effort, to generate SNP arrays for whole-genome association studies. In May, Affymetrix launched its next-generation array, called the Genome-wide Human SNP Array 6.0, or SNP6.0. "This chip allows us to look simultaneously at more than 1.8 million markers of genetic variation," says Keith Jones, vice-president of assay and application product development at the company.

The SNP6.0 not only offers genome-wide SNP coverage, but also contains more than 900,000 probes that target copy-number variants (CNVs) in the genome. "When walking down the path in designing the SNP6.0, we took the biochemistry that we used to generate targets to hybridize to the arrays and empirically identified probes that responded in a dose-dependent manner to changes in copy number," says Jones.

Other companies have also placed CNV probes onto their genotyping chips. Illumina, based in San Diego, California, now offers the Human 1M BeadChip, which boasts more than one million SNP and CNV probes, for whole-genome genotyping applications. Nimblegen, located in Madison, Wisconsin, and recently acquired by Roche Applied Sciences, also offers several whole-genome and custom-tiling array comparative genome hybridization products for examining copy-number variation across the entire genome. These arrays contain more than 385,000 probes at a median spacing of 6,000 base pairs. Agilent Technologies, located in Santa Clara, California, provides several array comparative genome hybridization products for analysis of copy-number variation in humans, mice and rats.



AFFYMETRIX

Affymetrix now offers the Genome-wide Human SNP Array 6.0 containing more than 1.8 million markers of genetic variation.

Illumina also offers several more-focused SNP arrays, including a cancer panel and one that targets the major histocompatibility complex, an area that Affymetrix also seems to be moving into. "I think it is also fair to say that you will see more application-specific SNP panels in the future," says Jones. N.B.

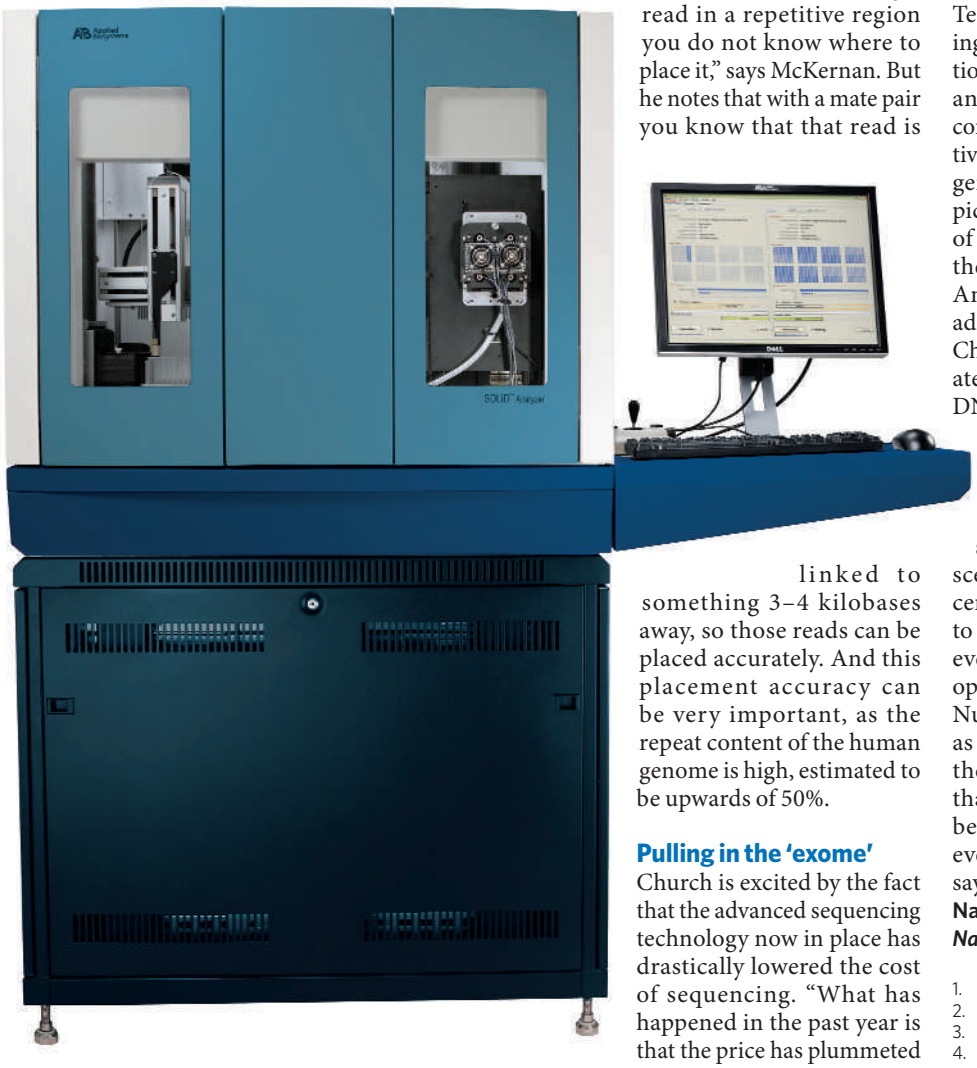
by performing sequencing by ligation rather than by synthesis, as conventional systems do. "Traditionally, people probably don't think of ligases as being more accurate than polymerases," says McKernan. But he claims that the SOLiD system achieves its accuracy by the new way in which the probes are encoded.

By doing successive rounds of ligation and looking at particular probe colour, the SOLiD sequencer obtains information not from just one base, but also from adjacent bases. So after ligation every colour has more than one base of information, which permits multiple colour calls for each base location. By using this redundant information, McKernan says a tremendous amount of error correction can be performed when making base-call decisions. "What we have been seeing is that this is giving us a ten to twenty times improvement over polymerase-based systems in terms of raw accuracy," says McKernan.

Applied Biosystems views the SOLiD system as particularly well suited for the cancer-research community and for applications in personal genomics. "Folks in the cancer community are going to gravitate towards it because they are looking for low-frequency mutations and the higher accuracy the system delivers through our error-correction scheme will be beneficial," says McKernan. He notes that cancer genomes tend to be very complex, with copy-number changes or copy-number neutral changes, such as translocations. Using sequencing systems such as SOLiD, these changes can be visualized by watching when pair reads, or 'mate pairs', break, quickly providing information about translocations. These data are currently analysed with copy-number chip assays (See 'Chipping out our differences').

Both the Genome Analyzer and the SOLiD system have some limitations because of the short read lengths of around 35 base pairs per run. "For sequencing larger genomes, 35 bases or fewer per run just won't cut it. We tried with 100 bases for a long time and had problems," says Egholm. McKernan says that the way Applied Biosystems is attempting to resolve this issue is by using mate pairs. "Whenever there is a single read in a repetitive region you do not know where to place it," says McKernan. But he notes that with a mate pair you know that that read is linked to something 3–4 kilobases away, so those reads can be placed accurately. And this placement accuracy can be very important, as the repeat content of the human genome is high, estimated to be upwards of 50%.

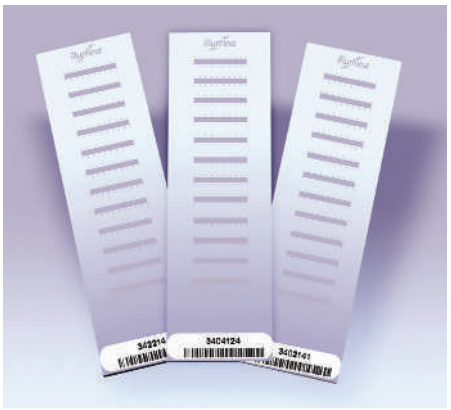
APPLIED BIOSYSTEMS



The next-generation SOLiD system from Applied Biosystems uses ligation-based sequencing methods.

Pulling in the 'exome'

Church is excited by the fact that the advanced sequencing technology now in place has drastically lowered the cost of sequencing. "What has happened in the past year is that the price has plummeted by a factor of 100," he says, making a project of the scope of the Personal Genome Project realistic from a financial perspective. The tricky part now is getting the most useful information from the human genome in a similarly cost-effective manner.



ILLUMINA

Illumina now offers several SNP and copy-number-variation probe chips for genotyping applications.

For the Personal Genome Project and other groups, the challenge is to obtain just exons, or the 'exome', from the human genome. Technically, the simplest method, performing in excess of 200,000 individual PCR reactions, would also be the most labour intensive and costly. But groups of researchers and companies are now working on ways to selectively amplify or capture different parts of the genome. "Soon you will be able to cherry-pick all the way along and identify the parts of the genome that are most likely to yield the maximum information," says Church. And these methods will be very welcome additions to the genomics world because, as Church notes, not all pieces of DNA are created equal. Although he points out that no DNA is 'junk', he contends that by examining only 1% of the genome you can get about 98% of the information about positions that cause changes in traits.

The technology necessary for a personal-genomics revolution is here on the scene. Most people say that the major concern for personal-genomics projects is how to deal with the data from participants. And even on that front there seems to be a lot of optimism within the genomics community. Nusbaum is encouraged that people such as Church have taken up such initiatives as the Personal Genome Project. "I am glad that someone like George is taking this on because he has charisma and clout, so that even if people don't want to hear what he is saying, they have to listen." ■

Nathan Blow is Technology Editor for *Nature* and *Nature Methods*.

1. Levy, S. *PloS Biol.* **5,** e254 (2007).
2. Mikkelsen, T. S. *et al. Nature* **448,** 553–560 (2007).
3. Robertson, G. *et al. Nature Methods* **4,** 651–657 (2007).
4. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. *Science* **316,** 1497–1502 (2007).
5. Barski, A. *et al. Cell* **129,** 823–837 (2007).
6. Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. & Welch, D. M. *Genome Biol.* **8,** R143 (2007).