BRIEFINGS IN FUNCTIONAL GENOMICS AND PROTEOMICS. VOL 8. NO 3. 174–183 doi:10.1093/bfgp/elp013

Advantages of next-generation sequencing versus the microarray in epigenetic research

Paul J. Hurd and Christopher J. Nelson**

Advance Access publication date 25 June 2009

Abstract

Several recent studies from the field of epigenetics have combined chromatin-immunoprecipitation (ChIP) with next-generation high-throughput sequencing technologies to describe the locations of histone post-translational modifications (PTM) and DNA methylation genome-wide. While these reports begin to quench the chromatin biologists thirst for visualizing where in the genome epigenetic marks are placed, they also illustrate several advantages of sequencing based genomics compared to microarray analysis. Accordingly, next-generation sequencing (NGS) technologies are now challenging microarrays as the tool of choice for genome analysis. The increased affordability of comprehensive sequence-based genomic analysis will enable new questions to be addressed in many areas of biology. It is inevitable that massively-parallel sequencing platforms will supercede the microarray for many applications, however, there are niches for microarrays to fill and interestingly we may very well witness a symbiotic relationship between microarrays and high-throughput sequencing in the future.

Keywords: *next-generation sequencing; microarray; Solexa/Illumina; Roche 454 pyrosequencing; ABI SOLiD; ChIP-Seq*

INTRODUCTION

Today biologists are offered a new tool to query genomes. Several next-generation sequencing (NGS) platforms are harnessing the power of massively-parallel short-read DNA sequencing to digitally interrogate genomes on a revolutionary scale. We are witnessing a paradigm shift in nucleic acid analysis: the ability to sequence genetic material at full-genome depth will change the types of questions that we ask in many disciplines of biology.

Several excellent reviews have thoroughly described the chemistry and technology behind the three leading NGS platforms [1, 2]. We therefore only briefly outline these technologies, their capabilities and their limitations before we discuss their utility in genomic analysis.

NEXT-GENERATION SEQUENCING TECHNOLOGIES

454 Pyrosequencing (Roche)

The first high-throughput sequencing platform to be commercially available uses emulsion PCR of DNA library fragments affixed to micro-beads. Using the new GS FLX Titanium platform, up to one million beads, each coated with a clonally amplified DNA molecule, are pyrosequenced in parallel. With individual sequence read lengths of up to 500 bases, a single run can generate 500 Mb of sequence. Of all next-generation platforms 454 sequencing provides the longest sequence reads, making it well suited to *de novo* genome assemblies. However, a drawback of the system's chemistry is inaccuracies in calling homopolymeric stretches of sequence (i.e. AAAAA, CCCCC).

*These authors contributed equally to this work.

Corresponding author. Paul J. Hurd, School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London, E1 4NS, UK. Tel: +44-207-882-8884; Fax: +44-208-983-0973; E-mail: p.j.hurd@qmul.ac.uk

Paul J. Hurd is a Lecturer, School of Biological and Chemical Sciences, Queen Mary University of London, Mile End Road, London, E1 4NS, UK.

Christopher J. Nelson is an Assistant Professor, Department of Biochemistry & Microbiology, University of Victoria, Victoria, British Columbia, V8W 3P6, Canada.

Solexa/Illumina Genome Analyzer

The Solexa/Illumina Genome Analyzer has been the tool of choice for many recent forays into genome-wide mapping of epigenetic marks. The massively-parallel nature of sequencing millions of DNAs is similar to the 454 platform. However in the Solexa/Illumina system, clonal DNA clusters are generated by bridge amplification on a glass surface rather than on agarose beads which enables increased densities (and numbers) of DNAs to be monitored simultaneously. The Solexa/Illumina Genome Analyzer uses reversible terminator chemistry to sequence up to 100 million clonal DNA clusters in parallel. At reads of 30–75 base pairs, this yields 3–7.5 Gb of raw sequence per run. Projected improvements to the technology, including read lengths of up to 125 bases, and sub-micron ordering of DNA clusters on glass slides (currently DNAs are randomly distributed), will boost coverage to greater than 100 Gb per run. The reversible terminator chemistry of Solexa/Illumina sequencing overcomes problems in quantifying the number of bases present in homopolymer stretches that is intrinsic to pyrosequencing. Although the total throughput of bases is higher, the shorter read length makes the current form of this platform most appropriate for well-annotated genomes. However, improved computational methods [3] and longer sequence reads will facilitate *de novo* genome assembly from Solexa/Illumina based sequence tags.

SOLiD (Applied Biosysytems)

Applied Biosystems offers a third approach to high-throughput DNA sequencing. With its SOLiD 3 system, DNA libraries are amplified on beads by emulsion PCR (as in 454 pyrosequencing) and the clonal sequence represented on each bead is determined by sequential rounds of ligation to a collection of dinucleotide-encoded adapters. The SOLiD achieves high sequence accuracy (purported to be >99.94%) because each base is interrogated twice in sequential rounds of ligation to dinucleotide-encoded adaptors: once as the first base of the dinucleotide, and again as the second base of the dinucleotide. This re-reading of sequence minimizes base-calling errors and makes the SOLiD well suited to high-accuracy sampling applications such as genome resequencing and polymorphism analysis. A single run on the SOLiD 3 sequences up to 400 million DNA tags. At 35 basepairs each, this yields nearly 15 Gb of total sequence.

MICROARRAYS

Contemporary microarrays emerged in the wake of genome sequencing projects for one obvious reason: arrays require *a priori* knowledge of the query genome. The 'array' has been instrumental in transposing functional information onto the raw sequence of the genome. Microarrays have evolved from measuring the expression of thousands of genes in yeast [4, 5] to comprehensive tiling arrays that systematically probe features across whole chromosomes. This tiling incarnation of the array was first utilized to comprehensively map transcription factor binding sites in the relatively small genome of budding yeast [6]. While the entire non-repetitive fraction of the human genome can now be represented on a set of seven high-density microarrays (GeneChip® Human Tiling 2.0R Array Set, Affymetrix), the cost of such an analysis can be prohibitive for many labs, especially if an experiment involves multiple samples. For this reason, researchers are offered choices of numerous commercially available sub-genome array formats. Microarrays representing collections of promoters, coding regions, transcript 3′ ends, alternative spliced exons, SNPs, and disease-gene arrays are all commonplace. Furthermore, if no commercial assortment of probes is appropriate, custom arrays are more affordable than ever before. So if microarrays offer such flexibility and coverage, what is all the fuss about sequenced-based approaches?

While powerful, microarrays do have several intrinsic limitations:

(i) As already mentioned, microarray design requires *a priori* knowledge of the genome or genomic features. This directly affects array effectiveness in cases of incomplete, incorrect, or outdated genome annotations. Furthermore, metagenomic approaches (where the genetic content from undefined mixtures of organisms in an environment is sampled *en masse*) are severely hampered by this restriction of microarrays.

(ii) A major obstacle in microarray analysis is cross-hybridization between similar sequences. This restricts microarray analysis to the non-repetitive fraction of genomes and complicates analysis of related genes (or features), alternatively spliced transcripts, allelic gene variants, and SNPs.

(iii) High signal to noise ratios and the fact that competitive hybridization on microarrays is a relative measure, limits the dynamic range of high-confidence data. This makes the detection of low-abundance sequences difficult and quantitative resolution of changes in highly represented sequences equally challenging.
(iv) Micrograms of DNA are needed to hybridize to arrays (particularly for whole genome tiling arrays represented on multiple slides) and a reliance on PCR-based amplification of material can introduce bias into samples.
(v) Finally, there is concern that the variety of available microarray formats, preparative methodologies and analytical approaches may limit the reproducibility of microarray data [7].

NGS-based approaches offer remedies to the above problems:

(i) Knowledge of genome annotation is helpful, but not required. In fact, data from next-generation platforms has recently been assembled into a genome *de novo* [8].
(ii) The fact that material is directly sequenced and not interrogated by hybridization to user-defined sequences removes experimental bias and cross-hybridization issues from the analysis. Since NGS offers single-nucleotide resolution one can monitor expression from gene alleles that differ in as little as one nucleotide of sequence. As the length of NGS reads increases, so too will our ability to probe repetitive regions of the genome.
(iii) Quantification of signal from sequence-based approaches is based on counting sequence tags rather than relative measures between samples: the result is unlimited fully-quantitative dynamic range of signal. NGS approaches are equally adept at detecting changes in rare and highly expressed sequences present in the same sample.
(iv) In terms of experimentation, nanograms of material are sufficient for NGS, reducing or eliminating the reliance on PCR amplification of material.
(v) Since data is collected genome-wide, investigators can simultaneously monitor RNAs from, or factor binding to, all known and undefined genomic features (i.e. promoters, exons, non-coding RNAs (ncRNA) and enhancers). Because all next-generation platforms have the same data output (a tally of sequences) it is hoped that the reproducibility of experimentation, and simplicity of bioinformatics analysis, will be much improved over a variety of microarray platforms with unique oligonucleotide probes and hybridization conditions.

We discuss here the recent flurry of studies that have utilized NGS to map the epigenome. These exciting articles are selected as examples of what the latest sequencing technologies can offer in terms of shear power and throughput, but more importantly they are illustrative examples of how the richness of high-resolution sequence-based data permits insight into biological phenomena that was not afforded by previous hybridization-based microarray studies.

NGS HAS ACCELERATED EPIGENETIC ANALYSIS

The interaction between DNA and proteins plays a fundamental role in the regulation of gene expression and control of DNA accessibility for transcription, replication, DNA repair and other DNA-based processes. These interactions can be studied using a technique called ChIP [9]. In ChIP, DNA and associated proteins are chemically cross-linked (typically with formaldehyde) and the DNA is fragmented by sonication or digestion with micrococcal nuclease. Proteins cross-linked to DNA are then immunoprecipitated using an antibody specific to the protein of interest. Cross-links are then reversed and the associated DNA purified. Determining which DNAs are enriched in the sample reflects where in the genome a factor was bound. Historically, quantitative PCR has been employed to query if specific regions of DNA were co-purified with the protein factor or modification of interest. More recently, ChIP enriched DNA has been combined with microarray platforms (ChIP-on-chip or ChIP-chip) in order to define *en masse*, thousands of *in vivo* binding sites of a number of factors and 'epigentic' chromatin features ([10] and references therein).

The epigenetic component of chromatin is comprised of histone post-translational modifications (PTM) and methylated DNA. Nearly 100 modifications of histones have been described and advances in mass spectrometry continue to discover more. While a catalogue of all existing chromatin

modifications provides the alphabet of the epigenetic language, dissecting how epigenetic units define the functionality of chromatin environments also requires an understanding of their distribution across sequence features such as genes. For this reason, epigeneticists have been particularly interested in mapping the locations of nucleosomes carrying specific histone modifications [11, 12].

The ChIP-on-chip approach has proved to be productive for the genome-wide mapping of DNA-binding proteins, nucleosomes and histone modifications. However the costs incurred for complete tiling arrays of the human genome, or the requirement for custom arrays, has meant that the use of genome-wide ChIP-on-chip has been limited. In the analysis of mammalian genomes, ChIP-on-chip has been mainly restricted to chromosome wide analyses or promoter-based arrays. Alternatively organisms with small genome size such as yeast have been extensively studied since only single arrays are required to encompass all non-repetitive genome features [11, 12].

ChIP-Seq detects novel binding sites for transcription factors

A new approach, ChIP-Seq, combines ChIP with massively-parallel direct sequencing. ChIP enriched DNA is directly sequenced, using the Solexa/Illumina platform and the reads are mapped to the reference genome. The frequency of times a sequence is found is directly proportional to the amount of that sequence in the enriched ChIP. This quantitative approach was first employed to profile both histone modifications and the transcription factor/insulator binding protein CTCF (CCCTC binding factor) genome-wide [13]. Subsequently, ChIP-Seq has been employed to identify transcription factor binding sites in the human genome for neuron-restrictive silencing factor (NRSF) [14] and signal transducer and activator of transcription 1 (STAT1) [15].

The 50 bp resolution achieved in the study of NRSF facilitated subsequent identification of associated binding motifs and also allowed identification of non-canonical binding sites. To compare the sensitivity of NGS with microarray, Robertson *et al.* [15] also mapped STAT1 binding sites on chromosomes 20–22 and X using ChIP-on-chip. This enabled a direct comparison between the sensitivity of ChIP-Seq and microarrays. Consistent with ChIP-Seq providing increased sensitivitiy, the authors found 803 STAT1 binding sites in this fraction of the genome using tiling arrays, and 3090 using ChIP-Seq. Together these studies demonstrate that, compared to ChIP-on-chip, ChIP-Seq offers improved resolution and increased sensitivity, which enables a more comprehensive identification of transcription factor binding sites *in vivo*.

It has become increasingly apparent that transcription factors have the ability to regulate gene activity over large distances from the transcriptional start site (TSS). ChIP-Seq has enabled a comprehensive identification of transcription factor binding sites *via* interrogation of regions outside known promoter sequences. This has enabled identification and delineation of previously unknown transcriptional networks. A recent study of the transcription factor FoxA2 (Forkhead box A2) in mouse adult liver identified that one fifth of all binding sites were over 50 kb away from any annotated gene [16].

NGS datasets can be merged to uncover functional relationships

ChIP-Seq analysis of a number of key embryonic stem cell transcription factors (Nanog, Oct4, Sox2 and Tcf3) has revealed that nearly half of all binding sites for these factors are located in intergenic regions distal from annotated start sites [17]. In this study, Marson *et al.* have combined ChIP-Seq with NGS deep-sequencing of microRNAs (miRNA). This strategy allowed the discovery of new pathways mediating miRNA transcription. Another multiple factor approach in murine ES cells profiled the genomic locations of 13 transcription factors and 2 transcriptional co-regulators [18].

Such multifactorial approaches are prohibitively expensive *via* microarray, and subsequent downstream comparison of datasets, difficult. The massively-parallel NGS technologies combined with ChIP-Seq are perfect for such combinatorial analyses. For similar reasons, a NGS ChIP-Seq approach allows the analysis of dynamic changes in transcription factor location that take place during physiological perturbation or cellular differentiation. For instance, the genomic distribution of STAT1 in interferon-γ stimulated and unstimulated human HeLa S3 cells has been examined [15]. Furthermore, the relationship between transcription factor location and gene expression can be examined by a combination of RNA deep-sequencing using NGS in parallel to ChIP-Seq analysis of the same biological sample [16]. In addition, ChIP-Seq has also been

employed in order to map the genome-wide distribution of other non-promoter associated DNA-binding factors such as the enhancer-associated coactivator p300 in ES cells and various tissues [18, 19] and the insulator-binding protein CTCF [13, 20]. ChIP-Seq analysis of RNA polymerase II [13, 21] has revealed occupancy at inactive promoters and many unannotated regions of the human genome, most likely at putative intergenic ncRNA transcriptional units. It is only by taking a truly parallel genome-wide approach, at high-resolution and in combination with transcriptome data, that such relationships can be elucidated.

Throughput of NGS can map the positions of all nucleosomes

The accessibility of transcription factors and the basal transcription machinery to promoters and TSSs is governed in large part by nucleosome occupancy ([22] and references therein). Accordingly, nucleosome positioning and phasing has been shown to play a key role in determining gene activity [22].

Previous genome-wide analyses of nucleosome positioning by microarray have been restricted to yeast at 4-5 bp resolution. In mammals only low-resolution analyses have been performed on limited genomic regions due to the cost and complexity of suitable microarrays [22].

Nucleosome free regions (DNase I hypersensitive sites) have recently been mapped to basepair resolution, using a combination of microarray, 454 pyrosequencing and Solexa/Illumina-based approaches in human T cells [23]. Furthermore, the dynamic remodeling/repositioning of nucleosomes in promoters in response to transcriptional activation has now been determined at high-resolution using NGS in yeast [24] and human T-cells [25]. Nucleosome maps of similar resolution on a genome-wide scale have been determined for *Drosophilia melanogaster* and *Caenorhabditis elegans* [26]. Further analyses look certain to reveal how the underlying DNA sequence influences nucleosome positioning and therefore gene regulation, through accessibility to promoters and TSSs [26].

The compositional differences of nucleosomes are also being investigated. Nucleosomes containing histone variants have been described in which core histones are replaced during transcriptional activation or repression ([22], and references therein). The histone H2A variant, H2A.Z has been mapped to high-resolution genome-wide using both 454 pyrosequencing in yeast [27] and a short-read ChIP-Seq approach in human T-cells [13]. However, it is not only nucleosomal positioning and compositional modifications that are key to understanding the biological function of nucleosomes and therefore chromatin - the chemical modification of specific residues in histones is also important.

NGS delivers data-rich epigenomic maps

The PTMs of histones are implicated in influencing gene expression and genome function by establishing and orchestrating DNA-based biological processes [28]. PTMs can either directly affect the structure of chromatin or recruit co-factors that recognise histone marks and subsequently adjust local chromatin structure and output. A comprehensive and high-resolution co-localisaton analysis of histone modifications for the human genome is required in order to understand the functional correlation of various PTMs in processes such as transcription, DNA repair and DNA replication [22, 29]. Use of modification-specific antibodies in ChIP has revolutionised the ability to ascribe biological function to histone modifications. ChIP on chip has allowed elucidation of the global distribution and dynamics of various histone modifications [12]. However, prior to NGS, it had not been practical to map multiple modifications in an unbiased genomic fashion.

Unsurprisingly, one of the first applications of ChIP-Seq was in the analysis of the genome-wide distribution of histone modifications. This study and others that followed, exemplify the newfound feasibility, and utility, of obtaining collections of comprehensive genomic datasets. Barski *et al*. [13] mapped the sites of 20 histone methylations in human T-cells and another study mapped the distribution of 5 histone methylation patterns in pluripotent and lineage-committed mouse cells [21]. Such genome-wide analyses have revealed associations between specific modified histones and gene activity as well as the spatial and combinatorial relationship between different types of histone modifications. Moreover, the study by Mikkeleson *et al*. [21] revealed dynamic changes in histone modification patterns during cellular differentiation and impressively, allele-specific histone modifications.

These seminal ChIP-Seq studies, in combination with more recent analyses examining the distribution of other types of histone modifications [30, 31], have revealed that specific genomic features are associated

with distinct types of chromatin signatures. Such genome-wide chromatin landscape maps have subsequently been exploited as a tool for defining and predicting novel transcription units, enhancers, promoters, and most recently ncRNAs in previously unannotated regions of the human genome [32]. By combining ChIP-Seq analyses of histone modifications along with FoxA2 and STAT1 (with and without interferon-γ stimulation) transcription factors, Robertson *et al*. [31] were able to examine the spatial distribution and relationship of histone marks with transcription factor occupancy. Surprisingly a single factor, STAT1, was associated with 25% of all histone H3K4 monomethylated genomic regions. Similarly, the colocalisation of CTCF with histone marks has been examined [13, 20]. Here CTCF was found to demarcate repressive and active regions of the genome, reinforcing its role in chromatin domain barrier function. This was typified by CTCF binding sites located between chromatin regions marked on one side by repressive H3K27 trimethylated histones, and on the other by the 'active' acetylated H2AK5 mark.

Single-nucleotide sequence resolution offers insight

DNA methylation at the C5 position of cytosine residues in CpG dinucleotides is essential for normal development and has been implicated in genome defence, genomic imprinting, X chromosome inactivation and carcinogenesis [33]. Perturbations of DNA methylation profiles in various human cancers reinforce the need for high-resolution maps of the DNA methylome in both normal and transformed cells and tissues. Historically, a number of approaches have been developed in order to study DNA methylation profiles ([34] and references therein). The most significant of which rely on either the direct sequencing of bisulphite-treated DNA or methylated DNA immunoprecpitation (MeDIP). In the presence of sodium bisulphite, unmethylated cytosines are converted to uracil (and after amplification, to thymine) whilst methylated cytosines remain unmodified. Combining bisulphite treatment with custom-made microarrays that contain probes to discriminate converted versus unconverted cytosines at a given CpG site, allows the original DNA methylation status to be determined. Whilst this approach has proved popular, it has been limited to gene- or region-specific DNA methylation patterns [34]. A more productive approach has been the use of MeDIP in combination with microarray analysis. Similar to ChIP-on-chip, this technique relies on the use of a 5-methylcytosine-specific monoclonal antibody to enrich cytosine-methylated DNAs, followed by subsequent amplification and hybridization on a microarray. Such approaches to analyze DNA methylation profiles in humans have been restricted to specific loci, promoter-regions or larger domains of individual chromosomes and have required custom-built arrays [35]. Nonetheless, the first complete genome DNA methylation profile was reported for *Arabidopsis thaliana* using this technique [36, 37]. However, the resolution using MeDIP combined with microarray was hundreds of bases in this study, potentially containing many individual CpG dinucleotides. In order to establish precisely which CpG dinucleotides are methylated, single-base resolution is required.

The advantage of having single-base resolution is demonstrated in several papers. For example, such an approach allows the precise sequence context of the methylation sites to be determined. A DNA methylome analysis on a genome-wide scale has recently been reported in *Arabidopsis thaliana* by combining various bilsuphite treatments with Solexa/Illumina-based sequencing. Moreover, Lister *et al*. [38] combined their DNA methylome study with deep-sequencing of the transcriptome. Performing this analysis in mutant backgrounds that lacked certain types of DNA methyltransferases allowed for a comprehensive analysis of subsets of genomic targets for each class of DNA methyltransferase. A similar approach using a reduced-representation bisulphite sequencing procedure [34] with Solexa/Illumina-sequencing has recently been reported in the analysis of the non-repetitive DNA methylome of pluripotent and differentiated mammalian cells at nucleotide resolution [39]. Since most DNA methylation resides in repeat regions, this study is not truly genome-wide but at this resolution, changes in DNA methylation patterns during differentiation were analysed over 5.8 Gb. Future developments using this technique offer much promise. A bisulphite Solexa/Illumina-based approach has also been employed to map the DNA methylome of *Neurospora crassa* [40]. In addition, combined use of MeDIP and Solexa/Illumina sequencing was also reported [40].

A second example of the utility of single-base sequence accuracy involves the exploitation of the single-nucleotide polymorphisms. Heterozygosity of SNPs can be used to define the mono-allelic

distribution of histone modifications and gene activity. Mikkelson *et al.* [21] used a catalogue of over 3.5 million SNPs in mouse embryonic stem cells to identify allele-specific chromatin signatures and gene expression status: NGS coupled with genetic crosses will therefore permit direct detection of novel imprinted genes.

The power of resolving information at the sequence level has also been harnessed by several groups who have used NGS of total cDNA (RNA-Seq) [41] to identify novel splice sites genomically [42, 43]. Furthermore, NGS of cancer genomes has uncovered several novel translocations [44, 45]. Neither of these findings would be possible using DNA-based microarrays.

Bar-coding allows multiplexing of samples for NGS

The advantages of using NGS to interrogate large genomes at high-resolution and in a massively-parallel fashion are demonstrated by many of the studies highlighted in this review. However the utility of NGS to interrogate small genomes is not immediately obvious, since high-density genome-wide microarrays are affordable. Hence, in most instances when analyzing smaller genomes, the use of NGS could best be described as 'overkill' and the shear number of reads generated unnecessary for most applications. Furthermore, there are also limitations on the number of different samples that can be processed in parallel at any one time. Two recent reports offer exciting adaptations to NGS protocols that should extend the appeal of NGS to researchers using organisms with smaller genomes. Both methods employ a multiplex strategy combined with either the 454 pyrosequencing platform [46, 47] or the Solexa/Illumina platform [48]. Each uses a bar-coding system, whereby unique 'tags' are included in the oligonucleotide adapters ligated to DNA library fragments for sequencing. After individual library generation, libraries can be pooled and sequenced in parallel on either platform. After sequencing the tags are used to identify the sample origin of individual reads.

The power of NGS in epigenetic research

Extensive and combinatorial ChIP-Seq analyses of multiple transcription factors concurrent with transcriptional co-regulators, boundary elements, numerous types of histone modifications, histone variants, nucleosome occupancy, DNA methylation patterns and gene transcription data are beginning to allow the elucidation and demarcation of complex transcriptional regulatory networks. Furthermore, chromatin signatures are allowing genome annotation based on the predictive power of certain combinatorial histone modification marks and the overall landscape of the epigenome in human cells. In addition, the elucidation of specific chromatin signatures associated with genomic features such as enhancer, insulator and boundary elements and promoters, offers a new approach to the annotation of complex genomes. The ability to query multiple genomic features was previously both technically challenging and monetarily demanding. NGS has made such efforts attainable, as demonstrated by the ambitious goals of epigenetics consortia (ENCODE, NIH Epigenetics Roadmap).

FUTURE ROLES FOR MICROARRAYS AND HIGH-THROUGHPUT NGS

While several papers discussed here have realized the advantages and massive throughput of NGS technologies, microarrays may still predominate certain applications. After all, microarrays are established tools that the research community is familiar with, plus the bioinformatics pipelines for array data analysis are mature. Importantly, the refinement of microarray technology has reduced the cost of asking certain genomic questions: arrays tiling smaller genomes are available at relatively little cost. Since a single microarray can comprehensively monitor either gene expression or genome-wide localization of factors across small genomes, such as that of yeast, this niche may represent a market that microarrays will continue to dominate.

If low cost is an advantage microarrays currently provide, then additional affordable DNA chips will continue to service specific research areas. If an investigator is specifically interested in gene expression changes, the query of the thousands of genes or promoters on a single array may provide sufficient information. Microarrays may then prove useful as a screening tool when either a low-cost 'quick-look' is warranted, or when the DNA or RNA of large numbers of samples, such as clinical isolates, need to be probed. Such multiplex arrays are already commercially available, and permit parallel analysis of collections of samples.

One can envision a powerful symbiosis between microarrays and NGS technologies. Arrays may be

best suited in classifying cohorts of samples, such as tumor tissues. Once samples of interest are defined, NGS could be used to provide comprehensive deep-sequence analysis of either genomic DNA to identify mutations, or RNA to report differences at the transcriptome level. NGS has been successfully used to genetically define cancers [45] by discovering new-disease associated tranlocations [49] and alterations in miRNA abundance [50].

Another exciting merger between the microarray and NGS is currently being offered by Roche, who interestingly acquired Nimblegen (microarray provider) in 2007 after acquiring 454 Sequencing (high-throughput pyrosequencing). Roche's 'Exome tiling arrays' are being offered as a sequence-capture tool. The 2.1 million-feature tiling array has long 60-mer probes designed to capture, and release, all annotated exons in the human genome. The use of such an array permits enriching protein-coding regions prior to NGS. Such preparative arrays provide a method for researchers to enrich genomic fractions before NGS to focus deep sequencing towards regions of interest. This pipeline may be extremely useful for identifying novel RNA splice junctions. Alternately, one could imagine using this approach to subtract highly expressed RNAs (or even all known RNAs), in order to more efficiently use NGS to uncover novel intergenic transcripts. Expect to see a range of preparative arrays being marketed to concentrate the power of NGS to subpopulations of nucleic acids.

With NGS manufacturers predicting increased read lengths, reduced costs and faster sequencing from existing platforms, the future of NGS technology appears to be both promising and routinely affordable for most researchers. Such is the power of NGS technology that therein lays its problem: NGS experiments generate huge volumes of data, which currently present challenges for data management, storage and importantly, analysis. As the cost of NGS lowers and access to NGS machines increases, these issues will become ever more prominent. The lag between the development of data analysis tools and the speed with which NGS technology is advancing is already creating a data bottleneck for many users. This was equally true during the early days of microarrays, but with time and with an ever-increasing user-base, bioinformatics and data analysis support was forthcoming. A number of bioinformatics tools are already available for routine but important computational aspects of NGS, including base-calling and reference genome alignment ([2] and references therein).

Perhaps the most conclusive and manifest demonstration of NGS in terms of cost, speed and the impact that NGS will undoubtedly have, is to consider that in 2003, the first human genome was sequenced after a 13-year effort at an estimated cost of $2.7 billion. In 2008, the first human genome using NGS technology was published, after a 2-month period at 1% of the cost [51]. Even though we are only at the genesis of NGS and its application to biology and medicine, third-generation technologies are being developed which promise to advance DNA sequencing to a remarkable level. One such method is being developed by Pacific Biosciences, Inc. Based on single-molecule, real-time (SMRT) sequencing, such third generation methodologies offer massively decreased sequencing times. They are extremely high throughput and exploit the high intrinsic rates, fidelity and processivity of DNA polymerases [52]. A single molecule of DNA polymerase with template bound, is immobilized on the bottom of a nano-chamber and base incorporation monitored. With each reaction capable of producing sequence at the rate of 400 kb per day, only 14 000 nanochambers would be required to produce DNA sequence equivalent to one diploid human genome per day.

In a manner analogous to the microarray, the influence and utilization of NGS technologies will surely find widespread use and relevance in many different areas of biology, advancing far beyond the test-bed of epigenetics. The numerous opportunities afforded by rapid advancements in DNA sequencing technologies hold much promise; the age of the much-vaunted '$1000 genome' is surely within our grasp.

Key Points

- NGS permits comprehensive interrogation of genomes without prior knowledge of sequence or annotation.
- ChIP-Seq allows the genome-wide mapping of DNA binding proteins and epigenetic marks.
- The lowered cost of NGS makes comprehensive mapping of multiple features possible.
- Accurate single-nucleotide resolution permits the discrimination between highly related sequences.
- The sequence-based digital data format simplifies comparison between datasets and permits unlimited quantitative range within a sample.
- NGS offers increased sensitivity to detect rare sequences in complex genomic samples.

We apologize to researchers whose work could not be cited due to space constraints.

References

1. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 2008;**9**:387–402.
2. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008;**26**:1135–45.
3. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008;**18**:821–9.
4. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* 1996;**6**: 639–45.
5. Spellman PT, Sherlock G, Zhang MQ, *et al*. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 1998;**9**:3273–97.
6. Ren B, Robert F, Wyrick JJ, *et al*. Genome-wide location and function of DNA binding proteins. *Science* 2000;**290**: 2306–9.
7. Ioannidis JP, Allison DB, Ball CA, *et al*. Repeatability of published microarray gene expression analyses. *Nat Genet* 2009;**41**:149–55.
8. Margulies M, Egholm M, Altman WE, *et al*. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005;**437**:376–80.
9. Solomon MJ, Larsen PL, Varshavsky A. Mapping protein–DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* 1988;**53**:937–47.
10. Bulyk ML. DNA microarray technologies for measuring protein-DNA interactions. *Curr Opin Biotechnol* 2006;**17**: 422–30.
11. Barrera LO, Ren B. The transcriptional regulatory code of eukaryotic cells—insights from genome-wide analysis of chromatin organization and transcription factor binding. *Curr Opin Cell Biol* 2006;**18**:291–8.
12. Rando OJ. Global patterns of histone modifications. *Curr Opin Genet Dev* 2007;**17**:94–9.
13. Barski A, Cuddapah S, Cui K, *et al*. High-resolution profiling of histone methylations in the human genome. *Cell* 2007;**129**:823–37.
14. Johnson DS, Mortazavi A, Myers RM, *et al*. Genome-wide mapping of in vivo protein-DNA interactions. *Science* 2007; **316**:1497–502.
15. Robertson G, Hirst M, Bainbridge M, *et al*. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 2007;**4**:651–7.
16. Wederell ED, Bilenky M, Cullum R, *et al*. Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. *Nucleic Acids Res* 2008;**36**: 4549–64.
17. Marson A, Levine SS, Cole MF, *et al*. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* 2008;**134**:521–33.
18. Chen X, Xu H, Yuan P, *et al*. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* 2008;**133**:1106–17.
19. Visel A, Blow MJ, Li Z, *et al*. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 2009;**457**:854–8.
20. Cuddapah S, Jothi R, Schones DE, *et al*. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* 2009;**19**:24–32.
21. Mikkelsen TS, Ku M, Jaffe DB, *et al*. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 2007;**448**:553–60.
22. Li B, Carey M, Workman JL. The role of chromatin during transcription. *Cell* 2007;**128**:707–19.
23. Boyle AP, Davis S, Shulha HP, *et al*. High-resolution mapping and characterization of open chromatin across the genome. *Cell* 2008;**132**:311–22.
24. Shivaswamy S, Bhinge A, Zhao Y, *et al*. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol* 2008; **6**:e65.
25. Schones DE, Cui K, Cuddapah S, *et al*. Dynamic regulation of nucleosome positioning in the human genome. *Cell* 2008;**132**:887–98.
26. Jiang C, Pugh BF. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* 2009;**10**: 161–72.
27. Albert I, Mavrich TN, Tomsho LP, *et al*. Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature* 2007;**446**: 572–6.
28. Kouzarides T. Chromatin modifications and their function. *Cell* 2007;**128**:693–705.
29. Groth A, Rocha W, Verreault A, *et al*. Chromatin challenges during DNA replication and repair. *Cell* 2007;**128**: 721–33.
30. Wang Z, Zang C, Rosenfeld JA, *et al*. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet* 2008;**40**:897–903.
31. Robertson AG, Bilenky M, Tam A, *et al*. Genome-wide relationship between histone H3 lysine 4 mono- and tri-methylation and transcription factor binding. *Genome Res* 2008;**18**:1906–17.
32. Guttman M, Amit I, Garber M, *et al*. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 2009;**458**:223–7.
33. Goll MG, Bestor TH. Eukaryotic cytosine methyltransferases. *Ann Rev Biochem* 2005;**74**:481–514.
34. Beck S, Rakyan VK. The methylome: approaches for global DNA methylation profiling. *Trends Genet* 2008;**24**: 231–7.
35. Weber M, Davies JJ, Wittig D, *et al*. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet* 2005;**37**:853–62.
36. Zhang X, Yazaki J, Sundaresan A, *et al*. Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. *Cell* 2006;**126**: 1189–201.
37. Zilberman D, Gehring M, Tran RK, *et al*. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet* 2007;**39**:61–9.
38. Lister R, O'Malley RC, Tonti-Filippini J, *et al*. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* 2008;**133**:523–36.

39. Meissner A, Mikkelsen TS, Gu H, *et al*. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* 2008;**454**:766–70.
40. Pomraning KR, Smith KM, Freitag M. Genome-wide high throughput analysis of DNA methylation in eukaryotes. *Methods* 2009;**47**:142–50.
41. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009; **10**:57–63.
42. Nagalakshmi U, Wang Z, Waern K, *et al*. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 2008;**320**:1344–9.
43. Mortazavi A, Williams BA, McCue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;**5**:621–8.
44. Maher CA, Kumar-Sinha C, Cao X, *et al*. Transcriptome sequencing to detect gene fusions in cancer. *Nature* 2009; **458**:97–101.
45. Simpson AJ. Sequence-based advances in the definition of cancer-associated gene mutations. *Curr Opin Oncol* 2009;**21**: 47–52.
46. Meyer M, Stenzel U, Myles S, *et al*. Targeted high-throughput sequencing of tagged nucleic acid samples. *Nucleic Acids Res* 2007;**35**:e97.
47. Meyer M, Stenzel U, Hofreiter M. Parallel tagged sequencing on the 454 platform. *Nat Protoc* 2008;**3**:267–78.
48. Lefrancois P, Euskirchen GM, Auerbach RK, *et al*. Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. *BMC Genomics* 2009;**10**:37.
49. Chen W, Kalscheuer V, Tzschach A, *et al*. Mapping translocation breakpoints by next-generation sequencing. *Genome Res* 2008;**18**:1143–9.
50. Kuchenbauer F, Morin RD, Argiropoulos B, *et al*. In-depth characterization of the microRNA transcriptome in a leukemia progression model. *Genome Res* 2008;**18**: 1787–97.
51. Wheeler DA, Srinivasan M, Egholm M, *et al*. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 2008;**452**:872–6.
52. Eid J, Fehr A, Gray J, *et al*. Real-time DNA sequencing from single polymerase molecules. *Science* 2009;**323**:133–8.