

NEXT-GENERATION SEQUENCING IN EPIGENOMICS RESEARCH

Ian K. Greaves^{1,3}, Stephen J. Ohms², Liz Dennis³, Frances Shannon^{1,4} and Jun Fan^{1*}

¹Department of Genome Biology, John Curtin School of Medical Research,
Australian National University, ACT 0200

²ACRF Biomolecular Resource Facility, John Curtin School of Medical Research,
Australian National University, ACT 0200

³CSIRO Plant Industry, Canberra ACT 2601

⁴University of Canberra, Canberra, ACT 2601

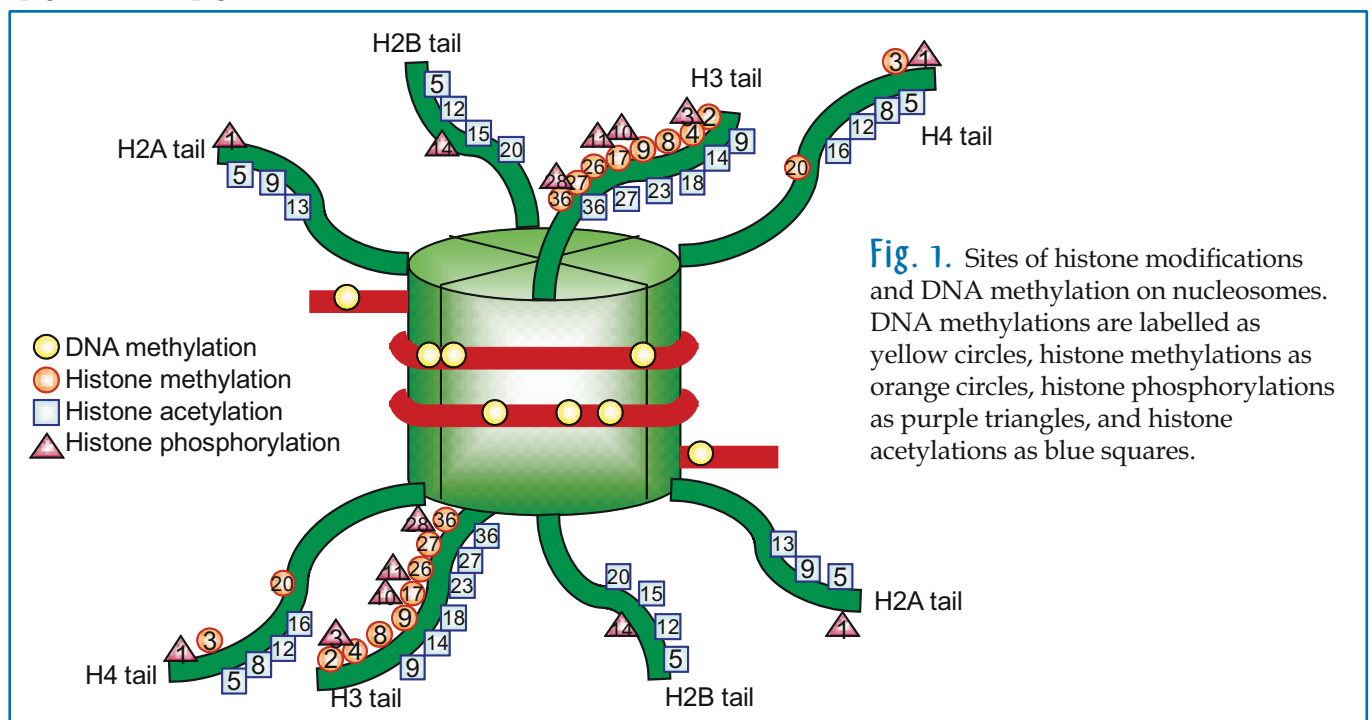
*Corresponding author: jun.fan@anu.edu.au

Introduction

It is exactly one century since Albrecht Kossel won a Nobel Prize for his work on the recognition of histone proteins. Half a century later, Watson, Crick, and Wilkins were awarded the Nobel Prize for their discoveries of the molecular structure of DNA. These two discoveries came together in our understanding of the packaging of the DNA double helix into chromatin in the eukaryotic nucleus. The 'histone code' theory was proposed by David Allis in 2000, suggesting that distinct histone modifications on one or more tails of histone proteins act sequentially or in combination to form a histone code that is read by other proteins to bring about distinct downstream events such as gene expression (1,2). This new concept, in combination with the much earlier proposal of an 'epigenetic landscape' by Conrad Waddington (3), has accelerated a merging of the fields of developmental biology and genetics, and has provided researchers a strong foundation for today's epigenetic and epigenomic research.

The term epigenetics refers to the modification of molecules that can influence the phenotype through changing gene expression without altering the nucleotide sequence of DNA. There are at least four distinct classes of epigenetic information so far identified: **1.** DNA methylation, which is a chemical modification on the DNA molecule itself at position 5 of the cytosine base (**Fig. 1**); **2.** chemical modifications of histone proteins, such as acetylation, methylation, phosphorylation, ubiquitylation and sumoylation, mainly located at the N- and C-terminal tail domains of major histones (**Fig. 1**); **3.** non-coding small RNAs; and **4.** histone variants. As next-generation DNA sequencing technology has become available at affordable prices, the mapping of these epigenetic markers in eukaryotic genomes has become a reality and has led to the new field of epigenomics.

Epigenetic mechanisms are now known to have a major role in normal development, aging and memory and also in the pathogenesis of many important diseases,



including cancer, diabetes, autoimmune disorders and neuropsychiatric conditions such as bipolar disorder (4). Understanding the patterns of epigenetic modifications at a molecular level may hold the key to determining the molecular basis of many of these disease states. Different cell types have different functional characteristics controlled by the expression of key genes, and the epigenomic landscape is likely to be a major factor in controlling this cellular differentiation and also the development of disease in response to environmental signals. One example of a hereditary disorder known to involve epigenetic mechanisms is Rett syndrome, a neurodevelopmental disorder affecting grey matter (5). In the early 1980s, cancer was one of the first human diseases found to be linked to epigenetics, but only a limited selection of specific genes were studied, and DNA isolated from primary colorectal cancers was found to be less methylated at the promoter regions of these genes than that isolated from normal tissues of the same patients (6). It is now known that alterations in methylation play a fundamental role in neoplasia, and in particular, that neoplasia appears to be characterised by genome-wide hypomethylation (7) interspersed with regions of localised hypermethylation (8), leading to chromosomal instability and increased mutability of the genome.

Our knowledge of the role of the epigenetic mechanisms involved in the disorders listed above is still elementary, but genome-wide chromatin immunoprecipitation (ChIP) sequencing and other forms of next-generation sequencing are likely to provide powerful tools over the next few decades for understanding these mechanisms.

Next-generation DNA Sequencing

Since the release of Roche's 454 in 2004, a number of high-throughput DNA sequencing platforms have been developed of which the most established are: 1. Roche's 454 that utilises a pyrosequencing platform; 2. Illumina's Genome Analyzer using bridge-PCR-based amplification followed by sequencing-by-synthesis; 3. Applied Biosystems' SOLiD (Sequencing by Oligo Ligation and Detection); and recently, 4. several single-molecule sequencing platforms.

Each instrument essentially sequences millions of different DNA fragments at the same time, resulting in millions of distinct DNA sequence reads of between 50–400 base pairs in length, depending on the platform. All next-generation sequencing machines share common features in their library preparation steps (Fig. 2). The target genomes to be sequenced must first be sheared into short fragments and have adapters ligated onto the ends of the fragments. From this point onwards, each platform differs slightly in the method of library preparation and sequencing. These methods are all well described in recent articles (9,10).

Applications in Epigenomics

In the Australian Cancer Research Foundation Biomolecular Resource Facility at the Australian National University, we have both the Roche GS FLX and Illumina GAIIx platforms. Services are available for both internal and external users, and applications cover a wide range

of techniques for epigenetic studies, including MethylC-seq, MeDIP-seq, ChIP-seq, and small RNA-seq. However, many of the descriptions in this article are limited to applications of Illumina's GAIIx (Fig. 2). Below, we describe the use of this sequencing platform to map DNA methylation across genomes, to map histone modifications using chromatin immunoprecipitation and to identify the small RNA composition of a cell or tissue.

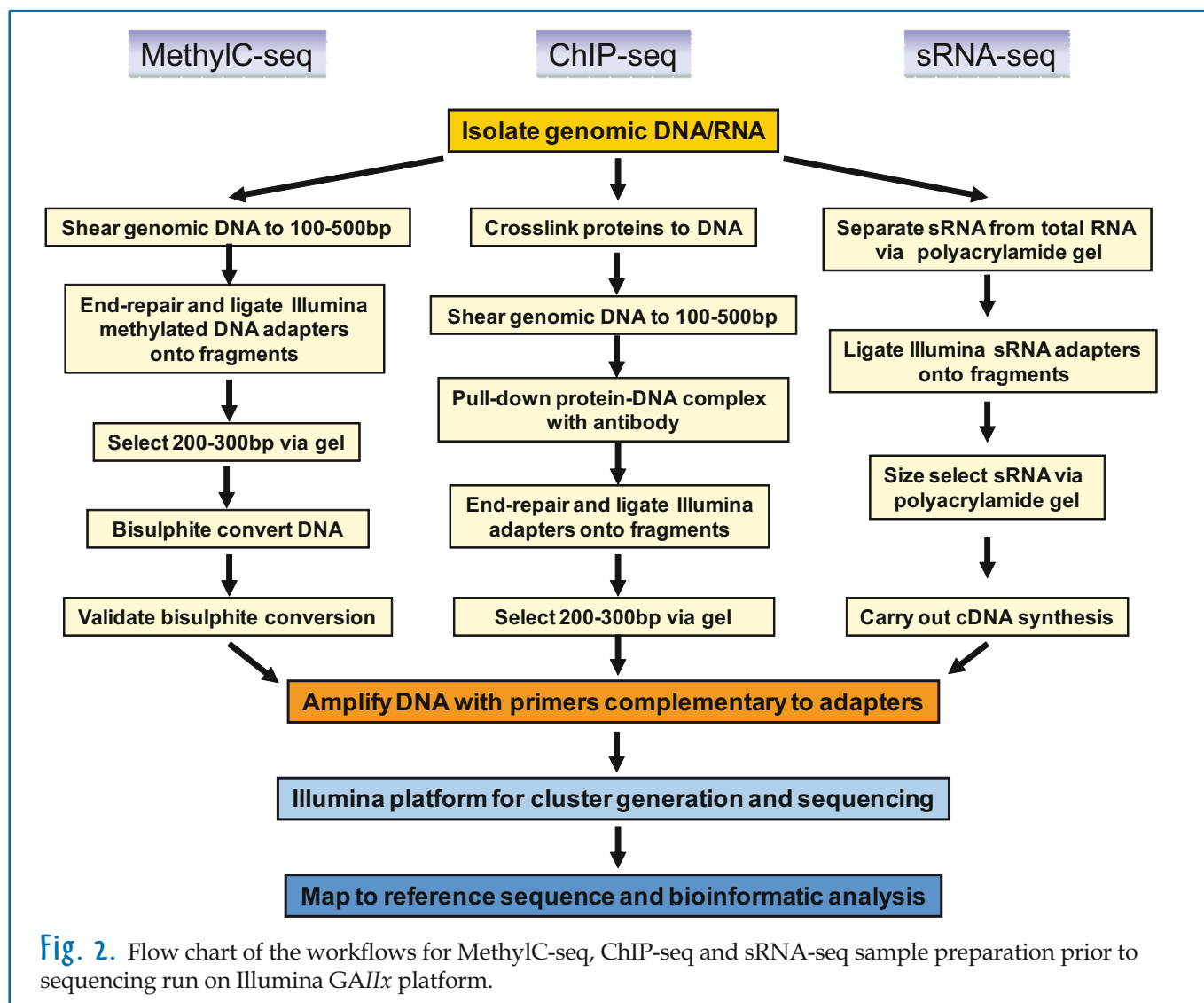
DNA Methylation

A key regulator in gene expression is the chemical modification of the cytosine base of DNA at the 5-position with a single methyl group, known as DNA methylation. These methylcytosines account for 1–6% of all nucleotides in mammalian and plant genomes. In animals and plants, the majority of DNA methylation is found in a CG context and plays a crucial role as a negative regulator of gene expression. In plants, DNA methylation is also found in a CHG and CHH context (where H=A, C or T) and is closely linked with the regulation and silencing of regions of the plant genome through the RNA-dependent DNA methylation pathway. The importance of DNA methylation as a negative regulator is best demonstrated at transposons, which are heavily methylated, thereby inhibiting their expression. The removal of DNA methylation leads to transposon activity and genomic instability, detrimentally affecting normal cellular function.

The development of next-generation sequencing is, for the first time, enabling researchers to study genome-wide DNA methylation patterns across many different biological systems. There are a number of methods available, such as MethylC-seq, MeDIP-seq and MIRA-seq, but the gold standard so far is MethylC-seq.

MethylC-seq is based on the bisulphite conversion of DNA. This procedure works through the conversion of unmethylated cytosines to uracil by the treatment of denatured single-stranded DNA with sodium bisulphite followed by sulfonation, deamination and desulfonation steps. Methylated cytosines are protected from this chemical conversion, remaining as cytosines within the bisulphite-converted genome. Upon PCR amplification and sequencing of bisulphite-converted DNA, methylated cytosines are read as cytosines while unmethylated cytosines will be converted into thymine. Upon matching the bisulphite-converted sequence to the normal genomic reference sequence, it is easy to recognise sites of methylation (C) compared to sites of non-methylation (T).

The method is outlined in Fig. 2 and certain modifications to the basic method needed to be made for use on the Illumina system. One key modification to the procedure is the use of methylated adapters. The bisulphite conversion of the sample library occurs towards the end of the procedure after Illumina's custom adapters have been ligated to the sample DNA. The adapters must have all the cytosines in their sequence methylated to prevent them from being converted. If not, the converted adapter would prevent annealing and sequencing by the sequencing primer during the final stages of sequencing. Another important step in bisulphite sequencing is checking that the library has been completely converted.



Non-converted DNA will give a large proportion of false methylation of cytosines and an inaccurate representation of methylation levels within the sample. The major factor contributing to incomplete conversion is failure to completely denature the DNA at the beginning of the bisulphite process. Chemical and protein contamination can inhibit the sodium hydroxide denaturation step, and, therefore, it is crucial that clean, high-quality DNA is used for the bisulphite conversion. Regions of the genome that are known to be unmethylated can be cloned and sequenced from the library by traditional methods. In plants, for example, regions of the chloroplast genome (known to have no cytosine methylation) are cloned and sequenced to check conversion rates.

Two issues with MethylC-seq are the reduced complexity of the bisulphite-converted genome (from a 4-base to 3-base genome) along with the depth of coverage required to get an accurate representation of methylated cytosines. Upon PCR amplification of the bisulphite-converted DNA, the uracils become thymine, reducing the majority of the genome to an A-G-T 3-base genome. This means that mapping the bisulphite-converted fragments to unique locations within the genome becomes more difficult. Longer read lengths along with more sophisticated mapping software are helping to alleviate this issue. The

second issue with MethylC-seq is the depth of coverage required to be confident of an individual cytosine's propensity for methylation. This issue makes the cost of MethylC-seq prohibitive for many research groups at present. *Arabidopsis thaliana*, with a small, 120-megabase genome, requires 2-3 gigabases of uniquely mapped data to cover the majority of the genome (11,12). The human genome at 3 gigabases required roughly 87 gigabases of data (~1.15 billion reads) (13). However, the projected dramatic improvement in sequencing output and reduced costs will soon make MethylC-seq a feasible option for many labs.

In the pioneering study mentioned above, MethylC-seq was used to map methylation throughout the entire *A. thaliana* genome at single base-pair resolution (11,12). This provided an unprecedented depth of information about DNA methylation, including previously unknown regions of methylation. It also provided insights into roles for the different methylation contexts, with the three different contexts of methylation (CG, CHG and CHH methylation) showing slightly different frequencies across gene bodies. A dramatic decrease in CHG and CHH methylation was observed in the region 1kb upstream of the transcription start site, whereas CG methylation was seen across the gene body.

Similar methylation mapping has been carried out in human embryonic stem cells. Significant advances were made in our understanding of the human methylome, including the first conclusive evidence of non-CG methylation in mammals, where 25% of all methylation sites in the embryonic cell line, H1, were in a non-CG context. Upon differentiation of these embryonic stem cells, non-CG methylation is lost, indicating that non-CG methylation is a characteristic of non-differentiated embryonic stem cells. A comparison of different embryonic stem cell lines also demonstrated widespread differences in methylation patterns (13), demonstrating the dynamic role of DNA methylation in regulating differential expression between cell types.

Once thought to be a static inherited epigenetic mark, recent genome-wide epigenomic studies are enhancing our understanding of the importance of DNA methylation in normal cellular functioning. This trend will only continue as research groups worldwide begin to study methylation patterns in more diverse biological systems, such as disease states and different developmental stages.

Small RNA (sRNA)

A powerful mechanism for silencing genomic expression is based on RNA silencing. This silencing is mediated by short RNA molecules that allow the sequence-specific silencing of target regions. Many size classes of sRNAs can be found, however, the most commonly occurring small RNAs are 21–22nt microRNAs (miRNA) and longer 24nt short interfering RNAs (siRNA). miRNAs play an essential role in development and act by directing the cleavage and/or translational inhibition of a target mRNA. This negative regulation enables specific control of gene expression within the organism. Both miRNAs and siRNAs are processed from an initial RNA molecule, which is cleaved by RNase III-like enzymes known as DICERs. In plants, different DICERs are responsible for processing miRNAs and siRNAs into their 21nt and 24nt sizes. These processed siRNAs are then loaded into a family of proteins known as Argonautes (AGO) that form effector complexes with other proteins, leading to negative regulation of the target sequence.

siRNAs are normally expressed from repetitive regions and play an important role in transposon silencing and genomic stability. Of particular interest are 24nt siRNAs in plants, as they play a crucial role in the RNA-dependent DNA methylation (RdDM) pathway. A plant-specific RNA polymerase, Pol IV, transcribes a target region. This transcript is made double-stranded by RNA-dependent RNA polymerase 2 (RDR2), which is then cleaved into 24nt siRNA by DCL3. These siRNAs are loaded into AGO4/5, which then targets regions of the genome for DNA methylation through DNA methyltransferases (DRM2/CMT3). In this case, both sRNA and DNA methylation play an essential role in maintaining genomic stability and gene regulation.

Deep sequencing of sRNA is carried out through the gel separation and extraction of sRNA from total RNA. The extracted sRNA has 5' and 3' RNA adapters ligated to it and is then converted to cDNA, which undergoes high-throughput sequencing. A particular advantage

of sRNA-seq is the relatively small number of reads needed to document experimentally valid levels of sRNA expression. A single lane of sequencing on the current Illumina platform will provide 20–25 million reads. This level of coverage is more than sufficient to obtain a clear understanding of the sRNA populations within a cell. To get the most out of one sRNA run, a potential improvement is the addition of barcodes to each sample. The addition of a small base sequence within the adapter of each sample enables researchers to run several samples in the one lane and the sequence output of the samples can then be separated bioinformatically.

One potential issue of sRNA-seq is the errors that may occur when comparing samples with different starting amounts of sRNA in the total RNA. As the sequencing machines require a standard concentration, any initial differences in the ratio of sRNA to total RNA between the two samples may be lost. One way to resolve this issue is to 'seed' the initial total RNA with a synthetic 21–24nt RNA. This seeded RNA sequence can then be used as an internal control within each sample and can directly allow the observation of any ratio differences in sRNA from the total RNA between samples (14).

Initial studies using next-generation platforms were carried out by using massively parallel signature sequencing (MPSS, Roche 454) on sRNAs from *A. thaliana* (15). This enabled the characterisation of hundreds of thousands of sRNAs within the genome and enabled the direct comparison of all sRNA populations between different tissues, developmental time points and species. For example, MPSS demonstrated a dramatic difference in the expression of miRNAs and siRNAs between inflorescences and seedlings from *A. thaliana* (15). Genome-wide sequencing has also been used to understand the role and mechanism of sRNA pathways by deep sequencing sRNA populations from plants with mutations in the RdDM pathway (16,17). Such experiments have highlighted the important impact sRNA has in regulating a large number of genes (18,19). Deep sequencing of plants with null mutations for *rdr2* and *ddc* (*drm1*, *drm2* and *cmt3* triple mutant) led to 70 and 226 upregulated transcripts, respectively, while 97 and 154 transcripts were downregulated, respectively (19).

One major advantage of deep sequencing is the flexibility it gives researchers in tackling biological problems. One example of this was work undertaken to understand how sRNAs are loaded into different AGO proteins (20). In *Arabidopsis*, there are 10 different AGO proteins, which can all bind sRNA and help regulate gene expression. To examine the subsets of sRNAs loaded into different AGO proteins, AGO proteins were purified by pull-down and the bound RNA from each AGO protein extracted. This RNA then underwent standard next-generation processing and sequencing. Consistent with the known literature, 82% of the AGO1-loaded sRNAs were annotated as miRNAs, while 57% of the AGO4-loaded sRNA matched repetitive regions consistent with 24nt siRNAs. However, one of the most striking findings was the selective loading of sRNA into different AGO proteins based on their 5' terminal nucleotides (20). For example, AGO1-loaded sRNAs were strongly biased for a

5' U (86%), whereas AGO2-loaded sRNAs were strongly biased for a 5' A (93%). The substitution of U with A in miR391 and miR393b led to these miRNAs being loaded into AGO2 instead of the normal AGO1, demonstrating conclusively the importance of the 5' terminal base in AGO loading (20).

Experiments such as these show the impact next-generation sequencing can have on our understanding of the epigenome and its influence over gene expression.

Mapping Histone Modifications Across the Genome

A large number of histone modifications have been implicated in the regulation of gene expression. The coloured shapes in **Fig. 1** represent some of the many post-translational modifications on the tails of the four major histones that can now be mapped across entire genomes. The modifications include methylation at lysines and arginines (orange circles), phosphorylation at serines or threonines (purple triangles), and acetylation (blue squares) at lysines (**Fig. 1**). Those modifications have the potential to form a complex combinatorial regulatory code. For example, tri-methylation of H3K(lysine)⁹ and the lack of H3 and H4 acetylation correlate with transcriptional repression in higher eukaryotes. In contrast, active chromatin marks, including the acetylation and methylation of H3K4 and H3K79, are positively correlated with levels of gene expression. Overall, histone modifications hold a key role in controlling and regulating genome structure and function.

Mapping the position of each histone modification and variant across specific DNA sequences can provide valuable insights into how these modifications function in a chromatin context. Chromatin immunoprecipitation is a powerful tool for this purpose and can be used to identify the specific histone associated with regions of the genome by using antibodies that recognise a specific histone modification. The method is based on the principle that formaldehyde reacts with primary amines located on amino acids of histones and the bases in DNA molecules, forming a covalent crosslink between the specific histone and the DNA on which it is situated. Once the cells are lysed, the DNA:histone complexes can be extracted and sheared into 0.5–2 kilobases in size. This complex is then immunoprecipitated using an antibody against the histone marker of interest. The DNA:histone cross-links are reversed by heating and the histones are removed by treatment with proteinase K. The DNA portion of the complex is then purified and processed to generate a library for next-generation sequencing. Compared to ChIP-on-chip experiments, ChIP-seq quantification is potentially more accurate and sensitive and overcomes the problem that commercial or custom-design microarrays cover only limited genomic regions.

Genome-wide mapping approaches provide new opportunities to decipher histone modifications. Although ChIP-seq data are less prone to error, systematic biases generated from all of the experimental procedures mentioned above create challenges for analysing ChIP-seq data. One way to minimise bias is to improve experimental procedures. For example, control ChIPs of modification-

free major histones may help to solve the potential bias problems associated with variation in nucleosome density within the genome. Alternative experimental conditions, such as the type of cross-linker, the extent of cross-linking, fractionation of chromatin, qualities of antibodies and experimental conditions that may cause cellular stress responses, should all be considered prior to ChIP-seq. Furthermore, data normalisation by comparing results across multiple biological samples is also critical in identifying whether biases are due to systematic error or true biological signal.

ChIP-seq technology can be extended for the characterisation of other DNA-binding proteins, such as mapping the binding sites of transcription factors where antibodies against the protein of interest can be used to pull down their target DNA.

We and others have successfully used ChIP-seq to map the distributions of various histone modifications and RNA polymerase II across the mammalian genome (21,22; Fan *et al.*, unpublished data). Zhao's lab used ChIP-seq to examine the status of histone occupancy, 37 different histone modifications, and the occupancy of histone variant H2A.Z and RNA polymerase II in resting human primary T-cells. They showed that typical patterns of histone methylation occurred at promoters, insulators, enhancers, and transcribed regions. It was also shown that histone modifications may act cooperatively to prepare chromatin for transcriptional activation (21,22).

Bioinformatics Analysis

The use of bioinformatics and computer methods of data analysis for next-generation sequencing related to epigenomics data have been reviewed previously (23–25). Computer-intensive methods are essential but are also the bottleneck in the analysis of high-throughput DNA ChIP-seq, MethylC-seq and sRNA-seq experiments. The basic problem is the very high number of reads coming off next-generation sequencing platforms (10–50 million reads per sequencing run) combined with the short length of the reads ranging from 36–76 nucleotides (for short-read platforms like the Illumina *GAIIX*). In addition, the reads themselves are far more likely to contain errors than those from traditional Sanger sequencing. These problems are compounded when aligning reads from mRNA sequencing. Alternative splicing of mRNAs can result in many possible versions of an mRNA transcript, which cannot be readily mapped to genomic DNA or to transcriptomes in specialised RNA databases without specialised alignment algorithms. Existing algorithms and software such as BLAST and its successor BLAT are orders of magnitude too slow for mapping next-generation sequencing data. Problems like these have led to the development of a new generation of alignment algorithms and software, including Eland, RMAP, SOAP, SHRiMP, Maq, iSSAKE, SOCS, QPalma, BowTie and TopHat (<http://seqanswers.com/wiki/Software/list>).

Following mapping of the reads, the next step in the analysis of ChIP-seq data is to take the output files from the mapping program (e.g., Maq) and input them into a peak-calling program like MACS or BayesPeak to find

immuno-enriched regions (26). Peak calling refers to the process of counting the number of reads that map to a genomic interval and subtracting a background derived from the sample itself or a second (control) sample so as to identify (or call) regions where large numbers of reads overlap to form peaks. Further steps may include smoothing peaks to more clearly distinguish signal from background noise, resolving partially merged or double peaks, and calculating a statistic to determine if a peak is significantly greater than background. If a control sample has not been run to provide the background, the experimental sample itself can be used as its own background. In this case, the background distribution of read counts of the sample is modelled by a Poisson or negative binomial distribution. The Poisson distribution is often used in the analysis of points or counts randomly distributed along a line (e.g., a chromosome) in a manner similar to analysing events in the time domain (e.g., an unexpectedly high number of phone calls within a certain time frame at a call centre). There has been a rapid proliferation of peak-calling programs besides MACS and BayesPeak, including ChipSeq Peak Finder, eRange, XSET, PeakSeq, QuEST, FindPeaks, SiSSR, and each of these uses a different statistical method (26).

In the next stage, the output from the peak-finding step in MACS (for example) can be exported in BED or WIG format and uploaded to and viewed directly in a genome browser such as the UCSC Genome Browser. The BED format is a simple text file that contains chromosomal start and end positions. Wiggle or WIG format allows the display of continuous-valued data in a track format and can be used for examining GC percent, probability scores, and transcriptome data, for example. BED format allows better visualisation of data compared to the WIG format, but there are limitations on the file sizes that can be uploaded to the UCSC Browser. The WIG format is far more compact than the BED format and may be useful when larger datasets are examined.

Alternatively, the output from the peak-calling step can be further imported into another program, such as PeakAnalyzer, in order to perform annotation and motif analysis on the predicted binding regions. In this context, annotation refers to determining the locations of the peaks in relation to known features such as genes and other annotated elements, and importantly, the transcription start sites of genes.

In a further step, a program like MEME can be used for motif analysis. The input for MEME is a file containing the sequences of interest for motif detection. MEME can search for common motifs in the detected peaks and the output from MEME can be piped into another program called TOMTOM, which can be used to compare these motifs with those in the TRANSFAC database.

Although high-throughput sequencing technologies are increasingly being used to investigate small RNA transcriptomes, the computational methods of data analysis are at an early stage of development. Non-coding (nc) RNAs appear to be a significant part of many transcriptomes. Many of these ncRNAs are processed by the cellular post-transcriptional machinery to yield shorter RNA products whose splicing patterns depend

on their secondary structure. This results in characteristic patterns of short reads that can be detected after mapping the read sequences to a reference genome. Based on this idea, a number of stand-alone and web-based computational tools now exist for the analysis of small RNA transcriptome sequencing data, including miRDeep, seqBuster, miRanalyzer and mirTools.

Finally, there are a number of online databases specialising in epigenetics information, including methylation-specific databases such as MethDB (<http://www.methdb.de>), PubMeth (<http://www.pubmeth.org>), MeInfoText (<http://mit.lifescience.ntu.edu.tw>) and the MethyLogiX DNA methylation database (<http://www.methylogix.com/genetics/database.shtml.htm>).

Histone and chromatin-related databases include the Histone Database (<http://genome.nhgri.nih.gov/histones>), ChromDB (<http://www.chromdb.org>) and CREMOFAC (<http://www.jncasr.ac.in/cremofac>).

Applications in Biology and Medicine

As mentioned in the introduction, next-generation DNA sequencing technologies have the potential to provide insights into key epigenetic mechanisms in development and disease. Following the sequencing of many eukaryotic genomes, including that of human, there have been focused efforts to explore epigenetic markers and their association with human health and disease. Since the launch of the International Roadmap Epigenomics program less than two years ago, much data has already been made publicly available (<http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics>).

There are many studies indicating that epigenetic changes occur throughout life in response to environmental, behavioral, physiological and pathological signals, and epigenomic information is providing a platform whereby the impact of environmental factors, such as smoking, toxins, drug abuse, medications, infections and diet, on the genome and gene expression beyond simple mutations can be measured and understood. Moreover, epigenetic factors may be pharmacologically reversible, alterable and controllable; thus the potential for therapeutic interventions for human disease or overcoming other epigenetic-related problems is profound.

Other examples of biological applications include understanding the role epigenetics plays in plant growth and development. Plant hybrid vigour and response to stress, drought and soil types are likely to have epigenetic components. Next-generation DNA sequencing technologies may enable us to study the complex molecular interactions that occur in hybrid genomes and to develop solutions for more efficient and effective food production.

References

1. Strahl, B.D., and Allis, C.D. (2000) *Nature* **403**, 41-45
2. Jenuwein, T., and Allis, C.D. (2001) *Science* **293**, 1074-1080
3. Goldberg, A.D., Allis, C.D., and Bernstein, E. (2007) *Cell* **128**, 635-638
4. Rodenhiser, D., and Mann, M. (2006) *CMAJ* **174**, 341-348

References continued on page 21



References continued from page 27

5. Amir, R.E., Van den Veyver, I.B., Wan, M., Tran, C.Q., Francke, U., and Zoghbi, H.Y. (1999) *Nat. Genet.* **23**, 185-188
6. Feinberg, A.P., and Vogelstein, B. (1983) *Nature* **301**, 89-92
7. Baylin, S.B., Herman, J.G., Graff, J.R., Vertino, P.M., Issa, J.P. (1998) *Adv. Cancer Res.* **72**, 141-196
8. Chen, R.Z., Pettersson, U., Beard, C., Jackson-Grusby, L., and Jaenisch, R. (1998) *Nature* **395**, 89-93
9. Ansorge, W.J. (2009) *New Biotechnology* **25**, 195-203
10. Shendure, J., and Hanlee, J. (2008) *Nat. Biotechnol.* **26**, 1135-1145
11. Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008) *Cell* **133**, 523-536
12. Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008) *Nature* **452**, 215-219
13. Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B., and Ecker, J.R. (2009) *Nature* **462**, 315-322
14. Fahlgren, N., Sullivan, C.M., Kasschau, K.D., Chapman, E.J., Cumbie, J.S., Montgomery, T.A., Gilbert, S.D., Dasenko, M., Backman, T.W., Givan, S.A., and Carrington, J.C. (2009) *RNA* **15**, 992-1002
15. Lu, C., Tej, S.S., Luo, S., Haudenschild, C.D., Meyers, B.C., and Green, P.J. (2005) *Science* **309**, 1567-1569
16. Zhang, X., Henderson, I.R., Lu, C., Green, P.J., and Jacobsen, S.E. (2007) *Proc. Natl. Acad. Sci. USA* **104**, 4536-4541
17. Nobuta, K., Lu, C., Shrivastava, R., Pillay, M., De Paoli, E., Accerbi, M., Arteaga-Vazquez, M., Sidorenko, L., Jeong, D.H., Yen, Y., Green, P.J., Chandler, V.L., and Meyers, B.C. (2008) *Proc. Natl. Acad. Sci. USA* **105**, 14958-14963
18. Zhang, B., Pan, X., Cannon, C.H., Cobb, G.P., and Anderson, T.A. (2006) *Plant J.* **46**, 243-259
19. Kurihara, Y., Matsui, A., Kawashima, M., Kaminuma, E., Ishida, J., Morosawa, T., Mochizuki, Y., Kobayashi, N., Toyoda, T., Shinozaki, K., and Seki, M. (2008) *Biochem. Biophys. Res. Commun.* **376**, 553-557
20. Mi, S., Cai, T., Hu, Y., Chen, Y., Hodges, E., Ni, F., Wu, L., Li, S., Zhou, H., Long, C., Chen, S., Hannon, G.J., and Qi, Y. (2008) *Cell* **133**, 116-127
21. Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007) *Cell* **129**, 823-837
22. Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q., and Zhao, K. (2008) *Nat. Genet.* **40**, 897-903
23. Bock, C., and Lengauer, T. (2008) *Bioinformatics* **24**, 1-10
24. Huss, M. (2010) *Brief Bioinform.* doi:10.1093/bib/bbq014
25. Lim, S.J., Tan, T.W., and Tong, J.C. (2010) *Bioinformatics* **4**, 331-337
26. Spyrou, C., Stark, R., Lynch, A.G., Tavaré, S. (2009) *BMC Bioinformatics* **10**, 299