APPLICATIONS OF NEXT-GENERATION SEQUENCING

Next-generation genomics: an integrative approach

R. David Hawkins, Gary C. Hon* and Bing Ren*

Abstract | Integrating results from diverse experiments is an essential process in our effort to understand the logic of complex systems, such as development, homeostasis and responses to the environment. With the advent of high-throughput methods — including genome-wide association (GWA) studies, chromatin immunoprecipitation followed by sequencing (ChIP–seq) and RNA sequencing (RNA–seq) — acquisition of genome-scale data has never been easier. Epigenomics, transcriptomics, proteomics and genomics each provide an insightful, and yet one-dimensional, view of genome function; integrative analysis promises a unified, global view. However, the large amount of information and diverse technology platforms pose multiple challenges for data access and processing. This Review discusses emerging issues and strategies related to data integration in the era of next-generation genomics.

Next-generation sequencing
Here, we define this as the use of established sequencing platforms, including the Illumina/Solexa Genome Analyzer, Roche/454 Genome Sequencer and Applied Biosystems SOLiD platforms, as well as newer platforms, such as the Helicos and Pacific Biosciences platforms.

Ludwig Institute for Cancer Research, Department of Cellular and Molecular Medicine, University of California, San Diego School of Medicine, 9500 Gilman Drive, La Jolla, California 92093-0653, USA.
Correspondence to B.R.
e-mail: biren@ucsd.edu
**These authors contributed equally to this work.*
doi:10.1038/nrg2795
Published online 8 June 2010

Driven by technological advances, recent years have witnessed a deluge of new methods for interrogating different properties of a cell on a genome-wide scale. Each offers a unique, although complementary, view of genome organization and cellular function. It is expected that integrating these data sets will provide more biological insights than using one data set alone. Thanks to the development of next-generation sequencing (NGS) technologies, the human genome has been mapped in many individuals; the challenge we now face is to understand this blueprint and to determine how errors lead to disease. The traditional approach of isolating individual genes and studying them in a model system is being rapidly replaced by data sets generated by both individual laboratories and large consortia using new high-throughput technologies.

Although individual data sets — including genomic, epigenomic, transcriptomic and proteomic information — are highly informative, integrating them together offers the exciting potential to answer many long-standing questions. For example, what are the functional variants of gene-distal loci identified by association studies? Where are the regulatory elements? And to what extent does the activity of regulatory elements contribute to disease phenotypes or to individual gene expression variation? Therefore, integrative analysis has become an essential part of experimental design in the era of next-generation genomics and is no longer the preserve of bioinformaticians. However, with the diversity of the high-throughput data and the seemingly endless analyses that can be performed, data integration is posing challenges for both bench scientists and computational biologists.

In this Review, we first briefly introduce the main high-throughput approaches. We then consider the types of biological questions that can be addressed through integrative analysis and insights that are starting to emerge, followed by discussion of commonly used data-integration strategies. We also consider the need for unified next-generation tools for data visualization, manipulation and analysis.

What types of genomic data sets are available?

In recent years, many high-throughput technologies have been developed to interrogate various aspects of cellular processes, including sequence and structural variation and the transcriptome, epigenome, proteome and interactome. Several recent reviews[1–7] have provided in-depth discussion of various platforms, so we only briefly introduce them below. Large collaborative projects are notably involved in using and developing genome-scale techniques, as discussed in BOX 1.

Sequence variation data. The ultimate goal of human genetics is to map every genetic variant and link each to phenotype. Currently, two high-throughput approaches are used to catalogue genetic variants: SNP genotyping arrays and resequencing. SNP arrays are cost-effective,

Box 1 | **Collaborative projects and technology development**

Over the next few years, technologies such as next-generation sequencing will generate a massive quantity of scientific data. Because of this, the scientific community must call for analytical tools to be developed alongside large-scale data production. For projects such as the Roadmap Epigenomics Project, the ENCODE Project and The Cancer Genome Atlas, data analysis and integration are clearly defined aims.

There is a broad selection of genome-scale approaches available, some of which might be redundant or might answer a different need. For example, for mapping DNA methylation on a large scale, some approaches, including reduced representation bisulphite sequencing (RRBS) and MeDIP–seq, provide cheaper but less comprehensive alternatives to full genome methylation mapping (MethylC–seq)[25]. The US National Institutes of Health Epigenome Roadmap Consortium has undertaken the task of a comparative analysis to determine how much pertinent information is gathered from different approaches. This comparative analysis will benefit the scientific community and could be of particular value to groups studying the role of DNA methylation across a cohort of patients — studies in which large numbers of samples necessitate cost efficiency. It is anticipated that such collaborative projects will lead to the first epigenome-wide association (EWA or epiGWA) studies.

In a similar way to the sequencing of the human genome itself, the mapping of the human epigenome and the cataloguing of human regulatory elements are not being left to individual laboratories. Collaborative efforts that result in a shared resource in which regulatory elements are consistently defined across the cohort of all experiments are being undertaken — for example, through the Roadmap Epigenome Consortium. This project will generate epigenomic maps for over 100 human cell types within the next several years. Similarly, the mapping of histone modifications and transcription factors in human cells by the ENCODE Consortium will provide additional insights into distal regulatory elements. Recently, several experiments using chromatin immunoprecipitation followed by sequencing (ChIP–seq) to search for such factors and modifications have been made publicly available, giving the scientific community the opportunity to begin using this resource. For model systems, *Drosophila melanogaster* and *Caenorhabditis elegans* are being investigated by the ModENCODE Consortium[116], and efforts are being made to develop a mouse ENCODE project. Maps of regulatory elements in multiple species will enable the investigation of specific questions and improve understanding of what is conserved among species.

and this strategy has been instrumental in the identification of disease-associated genes by groups such as the International HapMap Consortium[8]. More recently, NGS has reduced the cost of DNA sequencing, so it is feasible to directly sequence the exomes of an individual using methods such as sequence capture[9,10] or to sequence individual genomes, as is being performed in the 1000 Genomes project. NGS can also detect copy-number variants and gene-fusion events[11,12], and in the future NGS will probably overtake array-based detection methods owing to its superior coverage and resolution.

Transcriptomic data. NGS is also driving advances in transcriptomics[2,13]. For example, RNA sequencing (RNA–seq) can detect alternative splice variants using paired-end, relatively short reads (on the Illumina and Applied Biosystems platforms) or longer reads (using the Roche platform). In addition, RNA–seq can identify transcripts arising from gene fusion events (which are typical in cancer[14]) and can detect novel classes of non-coding RNAs (ncRNAs). For example, new classes of short RNAs have been identified that originate from promoters and gene termini[15], and many more large intergenic non-coding RNAs (lincRNAs) have been found[16]. In addition, a method that combines nuclear run-on with RNA–seq has been developed, which enables transcriptional rates in cells to be monitored[17].

Epigenomic data. DNA methylation and covalent modifications of histone proteins have been broadly defined as epigenetic modifications[18,19] and are important for transcriptional control[20–22]. High-throughput technologies now allow genome-scale mapping of these modifications[23–25]. Several large-scale analysis techniques are available that enable the survey of DNA methylation status at nucleotide resolution throughout the genome[6,26–29], including NGS coupled with bisulphite treatment of DNA. Chromatin immunoprecipitation followed by microarray (ChIP–chip) or, more recently, by sequencing (ChIP–seq)[3,4] can determine the genome-wide localization of histone modifications[30,31]. In addition, DNase I hypersensitivity site footprinting coupled with genomic arrays (DHS–chip) or NGS (DHS–seq, also known as DNase–seq)[32–36] defines regions of open chromatin structure, which can indicate potential regulatory sequences[33].

Interactome data. Interactions — both physical and functional — are an important layer of information for functional genomics. ChIP–chip and ChIP–seq are able to provide genome-scale information on DNA–protein interactions, and high-throughput sequencing of RNAs isolated by crosslinking and immunoprecipitation (HITS-CLIP, also known as CLIP–seq) is emerging as an important method for understanding RNA–protein interactions[37]. High-throughput dissection of protein–protein interaction networks has proved a greater challenge. It is largely done by the two-hybrid system[38], and in yeast this has been expedited by the cloning of all genes. However, in mammalian systems we are much further away. At a lower throughput, immunoprecipitation followed by mass spectrometry is becoming more widely available[39].

Technologies based on chromosome confirmation capture (3C) provide a snapshot of long-range interactions[40] among regions of DNA, which can be mediated through protein interactions. Circularized chromosome confirmation capture (4C)[41] and carbon-copy chromosome confirmation capture (5C)[42] provide large-scale analyses but are still limited to selected sites of interrogation[43,44]. However, recently developed methods have demonstrated the identification of long-range genomic interactions at a genomic scale through high-throughput, paired-end sequencing of the DNA fragments generated by the 3C method[45–47]. One method, Hi-C, maps numerous interactions in an unbiased fashion, whereas another, chromatin interaction analysis by paired-end tag sequencing (ChIA-PET), identifies interactions mediated by a particular protein through a ChIP step.

In addition, high-throughput methods are being used to define genetic and signalling pathways. For example, through large-scale RNAi screens, a number of key genes were linked to pathways regulating metastasis, apoptosis and senescence[48–53], and this provided new insights into cancer biology. In yeast, genetic interaction pathways

Reduced representation bisulphite sequencing
This technique cuts genomic DNA with restriction enzymes to enrich for CG-rich regions, which are then converted through bisulphite treatment and sequenced with next-generation sequencing. Bisulphite treatment converts unmethylated C to uracil — which appears as T in sequencing reads — while leaving methylated C intact.

MeDIP–seq
Methylated DNA is immunoprecipitated with an antibody against methylated cytosine and then sequenced by next-generation sequencing.

MethylC–seq
(Also known as bisulphite conversion followed by sequencing (BS–seq).) Methylated DNA is identified by shotgun sequencing of bisulphite-converted DNA.

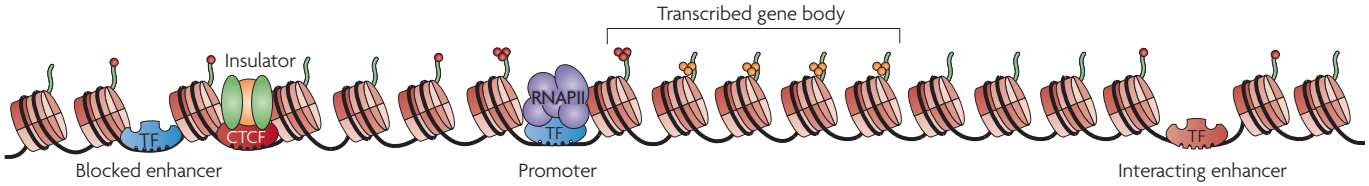

Figure 1 | **Annotating the genome through detecting transcription-factor binding sites and histone-modification states.** Promoters can be mapped by the localization of general transcription machinery and transcription factors (TFs), such as RNA polymerase II (RNAPII) or transcription initiation factor TFIID-associated factor 1 (TAF1), or by the localization of histone 3 lysine 4 trimethylation (H3K4me3). The bodies of transcribed genes and non-coding RNAs are marked by H3K36me3. Enhancers can be found by distal TF binding sites or by H3K4me1. This modification often coincides with H3K4me2, which has been shown to be necessary to recruit pioneering TFs to enhancer elements[121]. In addition, H3K4me1 sites overlap acetylated histone lysines, in agreement with acetylation islands outside promoters identifying functional enhancer elements[122,123]. Insulators are bound by CCCTC-binding factor (CTCF). Nucleosomes are shown as cylinders and example histone tails are in green. Various TFs are shown as coloured shapes. TFs bound to the insulator include CTCF and subunits of cohesin.

Sequence capture
This uses oligonucleotide microarrays or oligonucleotide-coupled beads to select for regions of the genome, such as all exons (exome sequencing) for targeted sequencing.

RNA sequencing
(RNA–seq.) RNA isolated from cells are sequenced by next-generation sequencing after conversion to cDNA.

Nuclear run-on
An assay that directly measures the transcriptional activity of a gene by incorporation of labelled UTP into its mRNA.

Histones
Small, highly conserved basic proteins, found in the chromatin of all eukaryotic cells, which associate with DNA to form a nucleosome. The amino-terminal tails of histones are subject to various post-translational modifications.

Chromatin immunoprecipitation
A technique used to identify potential regulatory sequences by isolating soluble DNA chromatin extracts (complexes of DNA and protein) using antibodies that recognize specific DNA-binding proteins.

DNase I hypersensitivity site footprinting
An assay that identifies regions of the genome that lack nucleosome structure and are therefore readily degraded by the enzyme DNase I. Such regions tend to be associated with transcriptional activity. When coupled with sequencing, the ends of DNA fragments generated by treatment of chromatin with DNase I are sequenced.

are being identified through large-scale epistasis screens (epistatic miniarray profiles (E-MAPs))[54,55], and soon such approaches might be applied to other model organisms or human cells. The power of such maps was recently shown by combining the information they provide with genome-wide association (GWA) studies in yeast to illustrate how single mutations are mechanistically relevant to key pathways[56].

Why perform integrative genomic analysis?

This broad array of data provides unprecedented opportunities for investigators to address some long-standing questions related to fundamental mechanisms of genome function and disease. For example, how might particular risk-associated SNPs affect cellular function and lead to specific diseases? What functional sequences exist in the human genome? And how are key developmental pathways regulated by epigenetic mechanisms? In this section we introduce some of the questions that integrative analysis is being used to answer; the methods for such integration are discussed in the following section.

Annotating functional features of the genome. A major challenge of understanding transcriptional control in higher eukaryotes is the incomplete catalogue of regulatory elements, particularly long-range regulatory elements, such as enhancers and insulators. As the characteristics of known regulatory elements are determined, these features can be used to identify novel elements. For example, the chromatin 'signature' of enhancers (FIG. 1) was determined and integrative analysis of histone modifications and localization profiles of the transcriptional co-activator p300 in human cells was used to find new enhancers[57,58]. Enhancer locations were confirmed by DHS analysis and functional assay, which is an important step for validating large-scale findings.

Although chromatin signatures define general classes of regulatory elements, their specific functions are dictated by transcription factors that bind the elements. For the human genome, the ENCODE Consortium members and others have used genome-wide localization of key factors to define regulatory elements, such as RNA polymerase II (RNAPII) and transcription initiation factor TFIID-associated factor 1 (TAF1) for promoter elements[59], CCCTC-binding factor (CTCF) for insulator elements[60], signal transducer and activator of transcription 1 (STAT1) and p300 for enhancers[58,61–63], and the transcriptional repressors KRAB-associated protein 1 (KAP1), suppressor of zeste 12 (SUZ12) and neural-restrictive silencer factor (NRSF, also known as REST) for silencing or repressor elements[24,64,65] (FIG. 1). These results support the feasibility for genome-wide identification of *cis*-regulatory elements, but additional functional studies are necessary for specific sites of interest. However, the activities of *cis*-regulatory elements are often restricted to specific cell types or development stages and so a comprehensive and precise catalogue of all *cis*-regulatory sequences will necessitate a thorough investigation of a multitude of transcription factors in various physiological conditions.

Inferring the function of genetic variants. GWA studies have revealed numerous SNPs that are linked to disease risk[66]. But one major obstacle is that if these SNPs fall within non-coding regions of the genome, our ability to assign functional roles to them is limited because functional features in the genome are still poorly defined in humans and other higher eukaryotes.

Recently, it was shown that SNPs could be called from short sequenced tags acquired from Illumina sequencing during ChIP–seq[67,68]. It would be highly informative to know whether transcription-factor-binding sites or chromatin-marked regulatory elements (see below) contain single-nucleotide variants (SNVs), which might be used to determine regulatory SNPs[69–71] (FIG. 2). For example, a study by Snyder and colleagues showed that SNPs found in binding regions for RNAPII and nuclear factor-κB (NF-κB) accounted for individual variability in gene expression levels[72]. Studies that identify open chromatin structures have also recovered known diabetes risk-associated SNPs[73]. Some algorithms that are used to find peaks of binding in ChIP–seq data have built-in SNP detection[74], so identifying variants could become a standard part of ChIP–seq analysis. However, it should be noted that in all efforts to identify SNPs there is an

inherent bias in mapping to the reference genome[75]. Therefore, additional measures should be taken to maximize mapped tags (for example, see REF. 72).

Calling variants in sequence-based assays will also provide important information beyond the SNP itself, as the presence of a SNP or SNV may enable detection of allele-specific expression. In the case of RNA–seq, if the transcriptional output of a heterologous locus contains a variant at or near 100% frequency, it is indicative of monoallelic expression. Allele-specific ChIP signals for transcription factors or RNAPII might offer a regulatory explanation for such allele-specific expression. For example, our group has previously demonstrated this with SNP arrays coupled with ChIP (SNP-ChIP)[76]. More recently, allele-specific regulatory regions in humans were identified through mapping DHS regions with CTCF colocalization[77]. Allele-specific DNA methylation, which can now be assayed at genome-scale, can also suggest potential mechanisms for monoallelic expression or repression, such as imprinting (see also below)[78]. Therefore, integrative analysis of allelic-specific transcription factor binding, epigenomic information and large-scale phenotypic read-outs, such as allelic-specific RNA expression data, will be key to identifying genetic or epigenetic mechanisms of gene expression. The extension of functional studies to structural variants will also be an important aim for future studies.

Understanding mechanisms of gene regulation. Because epigenetic features can control transcriptional output, and therefore traits, correlating epigenomic information and transcriptomic information can be highly informative. A classic example is genomic imprinting. Individual examples of imprinted loci — such as the *H19* locus in mammals — have been studied in detail[79] and show the complexity of transcriptional regulation, including the combined action of insulators, enhancers, chromosome looping and epigenetic marks. Genome-scale integrative analyses will enable broader questions to be answered, such as how many imprinted genes are there? How common is dysregulation of imprinting in disease? When does DNA methylation alter transcription factor binding? And what range of factors can be affected?

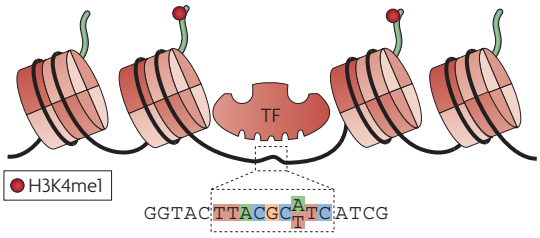

Figure 2 | **Identification of regulatory SNPs.** The sequence of a transcription factor (TF) binding site is shown with the position of an A/T polymorphism. By integrating chromatin signatures of enhancers or TF binding sites with SNP data, SNPs falling with the region would be predicted as regulatory SNPs. These could then be correlated to changes in gene expression. H3K4me1, histone 3 lysine 4 monomethylation.

Coupling histone modification data to transcriptomic data can also be valuable for the annotation of ncRNAs. Young and colleagues[80] identified microRNA (miRNA) transcription start sites by mapping the promoter-specific modification histone 3 lysine 4 trimethylation (H3K4me3) and comparing regions outside known promoters with annotated miRNAs, conserved regions, CpG islands and histone modifications (H3K36me3 and H3K79me2) that are associated with transcription elongation. Rinn and colleagues[16] mapped the location of thousands of lincRNAs by integrating these same chromatin modifications with RNA–seq data for expressed ncRNAs. It is now thought that many of these lincRNAs can influence histone modification or chromatin structure or subsequent methylation of DNA[81–83].

Integration of epigenomics with genomics and transcriptomics can also provide insights into transcription-coupled RNA processing. Recently, several groups found a correlation between exon expression and levels of H3K36me3 (REFS 84–89), and a subsequent study suggested a direct role for this modification in splicing control[87]. Further analysis of histone modifications in relation to splicing may provide additional insights into exon usage across genes[31,90]. Integration of exon expression data with HITS-CLIP data on the interaction of splicing factors with mRNA can also help to map splicing sites precisely[91]. In addition, integration of data on the promoter histone modification H3K4me3 (FIG. 1) with methods for the capture of the 5′ ends of genes (such as cap analysis of gene expression (CAGE) tags[92], which can be readily adapted to NGS) will improve annotation of transcription start sites (TSSs).

To understand what controls the spatial organization of gene expression and how regulatory elements and proteins interact with their targets, it is useful to integrate interaction data with other data sets. For example, nuclear architecture is, at least in part, defined by how chromosomes attach to the nuclear envelope. Nuclear-membrane-attached loci are typically marked by H3K9 methylation, and this modification is decreased in the laminin-associated diseases Hutchinson–Gilford progeria syndrome and facioscapulohumeral dystrophy[93,94]. The nuclear-membrane-attached regions often lie outside genes, so structural variants in unannotated genomic regions may be informative for understanding three-dimensional architecture. Future studies that integrate histone-modification profiles, transcriptomes, structural variations and chromosomal interaction data will expedite our understanding of nuclear architecture and define new mechanisms of disease.

Approaches to an integrative analysis

Several consortia are systematically interrogating genetic variation, the transcriptome, the epigenome and the interactome on a genomic scale. Each experiment adds another dimension of data to the genome, so there are now hundreds of dimensions of experimental data tethered onto the human genome (and other genomes), and this number is growing rapidly. The key to fully exploiting these data is integrating them. There are many ways to approach the challenge of data

HITS-CLIP
A technique similar to ChIP–seq in which proteins bound to RNA — such as splicing factors — are immunoprecipitated and the RNA fragments are sequenced.

Two-hybrid
An assay system in which one protein is fused to an activation domain and the other to a DNA-binding domain, and both fusion proteins are expressed in cells. Expression of a reporter gene indicates that the two proteins physically interact.

Epistatic miniarray profiles
These are created by screening the fitness of double mutants in a high-throughput manner. The results, when analysed as a whole, can reveal both positive and negative genetic interactions between genes and provide insights into biological pathways and protein–protein complexes in the cell.

Single-nucleotide variant
Sequence variations that include insertions and deletions in addition to base substitutions (which are known as SNPs).

Genomic imprinting
The epigenetic marking of a gene on the basis of parental origin, which results in monoallelic expression.

Cap analysis of gene expression
(CAGE.) The high-throughput sequencing of concatamers of DNA tags that are derived from the initial nucleotides of 5′ mRNA.

Box 2 | **Clustering**

Clustering is an integral bioinformatics tool for partitioning a large data set into more easily digestible, conceptual pieces. It can be applied to a wide variety of data, but traditionally has been applied to gene-expression profiles. Here, each gene is represented by a list of expression values in various cell types or conditions, and clustering identifies sets of co-expressed genes. In general, conventional clustering works well when the experimental values can be easily discretized into the clustered entities — for example, RNA–seq reads spanning exons in gene models are discretized to a single number (RPKM-normalized expression of genes).

However, for other applications, discretization is not possible or not desired. One example is for histone-modification data derived from chromatin immunoprecipitation followed by sequencing (ChIP–seq), in which the profile of experimental values over a contiguous region is informative. Conventional clustering can be applied to this data, provided that the profiles are well aligned. For example, to enumerate commonly occurring chromatin signatures in an unbiased way, conventional clustering can be applied to a subset of genomic regions, such as promoters. If a predefined number of clusters *k* is expected, then *k*-means clustering can be applied, in which each promoter is assigned to the most similar cluster. Alternatively, hierarchical clustering (in which each promoter is related to all other promoters, as represented in a 'tree-like' pattern) can be used to offer more flexibility. Clearly, conventional clustering can be applied to a wide variety of genomic data sets, including genomes, epigenomes[57], transcriptomes[16] and interactomes[117]. However, this method gives the best results when the set of loci examined are well-aligned, which is the case for gene definitions for which excellent annotations exist.

To cluster loci with poorly aligned or asymmetric chromatin signatures, or for poorly annotated loci such as gene-distal regulatory elements, our laboratory has developed an approach called ChromaSig[89,99]. Given a set of genomic loci, ChromaSig aligns and orients the epigenetic profiles around the loci, outputting clusters of loci that share similar profiles. Alternatively, given the genome-wide nature of epigenetic data, another clustering approach is to assign a cluster to every part of the genome. To accomplish this task, Jaschek *et al.*[118] used a hidden Markov model approach to ascertain the most likely epigenetic states given the data.

integration, and we discuss three important — although not mutually exclusive — approaches below.

Data complexity reduction. For a growing number of sequencing-based assays, such as ChIP–seq, DHS–seq, formaldehyde-assisted isolation of regulatory elements followed by sequencing (FAIRE–seq), RNA–seq and Hi-C, the result of each experiment is millions of short sequence reads, which essentially give a continuous signal of enrichment across the genome. A simple approach to reducing the complexity of this data set from millions of data points to a more manageable hundreds or thousands of sites is to summarize each experiment as a collection of genomic regions with strong enrichment of signal. For ChIP–seq, peak-finders discretize the genome-wide profiles into regions with enrichment and those without. Therefore, a commonly used method of data integration is to perform intersection analysis on enriched regions from different experiments. For example, Chen *et al.*[95] mapped a collection of 13 transcription factors using ChIP–seq in mouse embryonic stem cells, used a custom peak-finder to call regions of enrichment and observed significant co-binding of transcription factors.

Although intersection analysis on discretized data sets is straightforward to perform, special attention must be paid to the underlying assumptions of different data discretizers. For example, blanket application of a peak-finding method and set of parameters to different types of data — such as histone modifications, transcription-factor binding and open chromatin — is often ill-advised for several reasons. Firstly, the type of experiment usually dictates a specific kind of data analysis. For instance, transcription factors often bind discrete, specific sites and so ChIP–seq tags at the point of binding have a biased distribution between positive and negative strands, which can be used by peak-finders to obtain excellent precision[74,96]. However, this assumption is less suitable when binding or enrichment occurs contiguously across large stretches of DNA or in clusters, as is the case for certain chromatin modifications[30,97]. Therefore, one must be mindful of the underlying assumptions and limitations of peak-finders before applying them. Secondly, even among the same type of data, variability in data quality may necessitate calling peaks with different thresholds and/or data-normalization methods. This is especially true for ChIP–seq experiments, in which variable quality of antibodies or suboptimal ChIP conditions can lead to variable ChIP enrichment, which will require the adjustment of significance thresholds individually to achieve high sensitivity and specificity.

It is important to note that the inherently noisy nature of genome-wide data means that a perfect peak-finder cannot exist: in calling regions of enrichment, one can only hope to minimize, but not eliminate, false-positives and false-negatives. Realizing this, it is evident that we cannot simply trust peak-finders blindly and that it is especially important to inspect at least some of the results by eye. Thus, if we are to perform meaningful analysis, we cannot be far removed from the original data and should validate computational analyses experimentally.

Unsupervised integration. A more scalable method for integrating data is unsupervised learning, in which the data is approached with no prior biases, knowledge or hypotheses. To summarize a large data set into smaller groups that can be more easily conceptualized, an unsupervised approach simply asks the question: what kinds of patterns exist in a data set? One common assumption made by unsupervised approaches is that the interesting features of the data are the ones that occur frequently, and therefore the goal is to find common patterns. As diverse experimental methods equate frequency of genomic mapping with activity, an unsupervised analysis can treat these data sets equally and need not know the nature of the measurement. For example, Zhao and colleagues[30,31] profiled 37 histone modifications in human $CD4^+$ T cells. Although the number of different possible combinations of modifications is a staggering $2^{37} \approx 137.4$ billion, it is likely that most combinations do not exist or occur very infrequently. To enumerate commonly occurring chromatin signatures or other patterns, clustering approaches can be applied (BOX 2).

The genome serves as a scaffold on which high-throughput data are assembled, and from a genome-centric perspective, clustering can be seen as a way of classifying genomic loci into conceptual groups with shared attributes. Clustering data from different kinds of experiments gives distinct types of conceptual groups, and the first phase of data integration can be seen as

Formaldehyde-assisted isolation of regulatory elements followed by sequencing
(FAIRE–seq.) This technique isolates nucleosome-free regions of DNA from chromatin during phenol:chloroform extraction.

Discretization
The conversion of a continuous signal to a discrete signal.

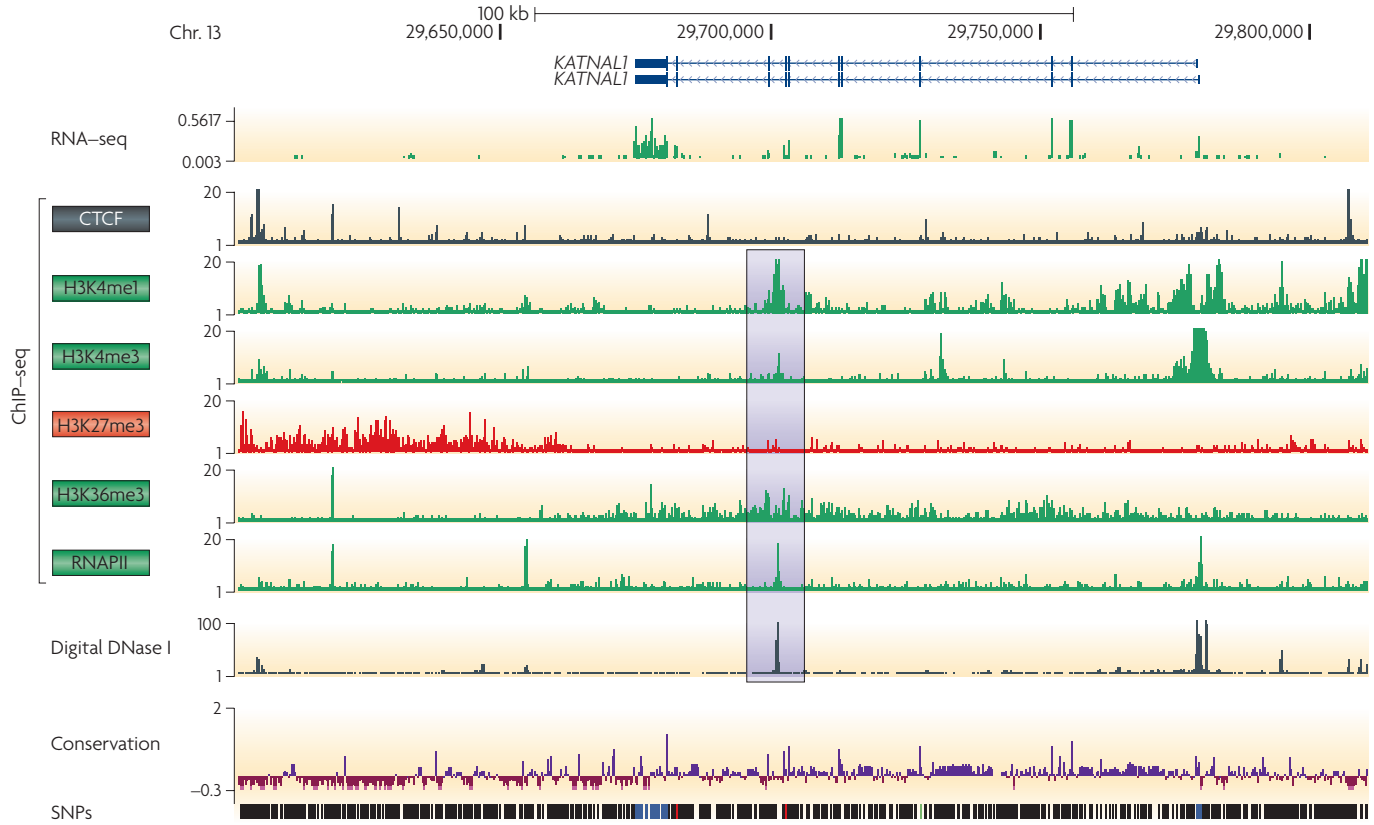

Figure 3 | **Data visualization.** The University of California-Santa Cruz (UCSC) Genome Browser is a tool for viewing genomic data sets. A vast amount of data is available for viewing through this browser. This example from the browser shows numerous data types in K562 cells from the ENCODE Consortium. A random gene was selected — katanin p60 subunit A-like 1 (*KATNAL1*) — that shows several points that can be identified by using this tool. The promoter has a typical chromatin structure (a peak of histone 3 lysine 4 trimethylation (H3K4me3) between the bimodal peaks of H3K4me1), is bound by RNA polymerase II (RNAPII) and is DNase hypersensitive. The gene is transcribed, as indicated by RNA sequencing (RNA–seq) data, as well as H3K36me3 localization. The gene lies between two CCCTC-binding factor (CTCF)-bound sites that could be tested for insulator activity. An intronic H3K4me1 peak (highlighted) predicts an enhancer element, corroborated by the DNase I hypersensitivity site peak. There is a broad repressive domain of H3K27me3 downstream, which could have an open chromatin structure in another cell type.

enumerating the conceptual modules of each data set. For example, clustering of RNA expression reveals co-expressed genes[98], clustering of histone modifications gives loci that share similar chromatin structure[57,89,99], protein–protein interaction clustering finds proteins in the same complex[100], and genetic interaction clustering reveals members of the same or similar pathways[55].

Although all modules are tethered to the genome, modules from one experiment are not linked to those from others. Thus, the next task in data integration is to connect these modules. One approach is to examine a module from one data type — for example, chromatin signatures — in the context of another data type — for example, DNA methylation[25,101,102]. Alignment of data sets on a browser, such as the University of California-Santa Cruz (UCSC) Genome Browser[103], might be useful in this regard (FIG. 3). Furthermore, the Genome Browser also contains annotations, such as gene definitions, evolutionary conservation and disease associations[104]. Therefore, co-clustering of new experimental data with known annotations can provide an easy bridge to hypothesis generation. In the past, when genomics consisted only of global gene-expression analysis, annotation libraries such as Gene Ontology[105] and the more sophisticated Gene Set Enrichment Analysis[106] were developed to provide an easy way to assess the biological significance of gene hits. As data sets are now extending to include ncRNAs, disease-associated SNPs and regions of transcription-factor binding, it seems that 'Locus Set Enrichment Analysis' will be an important part of genomics. Sets of loci that share factor binding, epigenetic modifications or disease association will provide efficient ways to form hypotheses regarding function outside of coding regions.

Another approach to connecting conceptual modules involves network biology, which leverages high-throughput techniques to find relationships that connect genomic loci and conceptual groups. Such approaches include: Hi-C, which maps how chromosomal interactions connect genomic loci to each other; E-MAPs, which use genetic interactions to connect proteins to pathways; and ChIP–seq, which links transcription factors with

regulated genes. This second level of integration — linking different kinds of experiments — can form a knowledge base that can be used to provide biological insights or to formulate hypotheses for further study.

As a hypothetical example, suppose we used ChIP–seq to map a novel transcription factor genome-wide and wanted to know the significance of its binding profile. Complicating matters, most of the binding sites are distal to promoters. Clustering reveals that a subset of binding sites share a similar chromatin environment, which suggests that these sites may function similarly. Hi-C data then links this subset of binding sites with their target genes and RNA–seq data reveals that these genes are highly expressed. Finally, protein–protein and genetic interaction data reveal that some of these expressed genes belong to related but distinct protein complexes that regulate RNA splicing. Thus, data integration would allow us to efficiently propose the hypothesis that the binding of this new factor to DNA regulates the process of RNA splicing.

Often, the scope of genomic experiments performed is so diverse that it is not immediately clear how, or even if, one experiment relates to another. It is in such cases that unsupervised, data-driven approaches to integration are most useful. Unsupervised integration is a discovery tool for finding correlations between two or more experiments. Novel associations lead to hypotheses of function, which can be followed up by supervised integration and by direct experimental validation (see below). In this way, high-throughput experiments are screens for identifying interesting, unexpected associations. Because of the power of the approach and because the inputs required are minimal, unsupervised integration is arguably the first tool that should be applied to a new data set, and it should be constantly run as new experiments are added to an existing data set to find additional associations.

Supervised integration. The discovery of patterns is one output of unsupervised integration, but the patterns alone do not advance our understanding of biology or disease. Like most systems biology approaches, unsupervised integration excels at generating hypotheses. Therefore, a novel pattern is simply an observation from which we must make and test predictions of function, often by incorporating external data sets or new experiments. This is the realm of supervised integration. Supervised integration is driven by testable hypotheses and so often relies on only a few dimensions of a full data set.

It is important to note that the choice of data to include in supervised integration and the specific method used depend crucially on the question posed. For example, using an unsupervised clustering approach, we recently observed a set of distinct histone modifications at exons, which led to the hypothesis that these modifications mark alternatively expressed exons[89]. To test this hypothesis, we needed to examine these chromatin modifications in the context of expression at the exonic level, and we were able to use previously published exon expression array data from the same cell type[107].

However, in most instances the impetus for supervised integration is anecdotal evidence obtained from observations of genome-scale data on a browser or from previously published studies. For example, Guttman *et al.*[16] took advantage of previous observations that RNAPII-transcribed genes are marked by H3K4me3 at promoters and by the spread of H3K36me3 into the transcribed region, and they used this chromatin signature to identify RNAPII-transcribed lincRNAs. Thus, supervised integration starts with a prediction based on an observation and ends with a test of this prediction. This is arguably how our biological understanding is advanced most: the more predictive the hypothesis, the more biological insights are gained. Therefore, observation and data integration cannot be independent from each other and there is no substitute for seeing the data with one's own eyes. For example, our opinion that it is necessary to see raw data using a browser is consistent with the current trend in data visualization towards replacing traditional averaged plots with more information-rich heatmaps, which provide experimental profiles for thousands of loci simultaneously (for example, genome-wide heatmaps of ChIP–chip data[58]).

As there are now tens of thousands of high-throughput experiments linked to the human genome, finding dependence relationships among the many dimensions of experimental data is essential to increasing our knowledge. In the simplest case, relationships can be discovered by correlation analysis. For example, a strong, positive correlation between the binding profiles of two transcription factors indicates that one may be dependent on the other. Additionally, for genetic interactions, finding positive and negative correlations for a mutant under different conditions can allow the systematic discovery of condition-dependent relationships (S. Bandyopadhyay, personal communication).

Although informative, correlation analysis can become unwieldy as the number of data sets grows — doubling a data set would effectively quadruple the number of computations necessary and the number of visualizations required. Luckily, machine learning techniques, notably Bayesian networks[108], offer a supervised approach for discovering relationships among data entities. Using a probabilistic framework, Bayesian networks can find dependence relationships, as van Steensel *et al.*[109] did for a panel of chromatin modifications and chromatin-associated proteins and modifiers. Bayesian networks can also readily integrate data from different kinds of experiments. For example, Yu *et al.*[110] modelled the interdependence of histone-modification profiles with the binding of transcription factors, together with their relationship to gene expression. However, it is important to note that the types of prediction that are the output of a Bayesian network crucially depend on how the network is designed, which in turn depends on the question asked. For example, Jansen *et al.*[111] designed a Bayesian network to predict protein complexes by integrating diverse data sources, including protein–protein interactions, expression and gene annotation. In summary, Bayesian networks can find relationships among diverse kinds of data and thereby create hypotheses that can be tested experimentally.

Box 3 | **Online tools for integrative analysis**

Galaxy is an online genomics analysis tool that allows users to perform a number of integrative data analyses on genomic data sets. Although not a database itself, it is directly linked into many genomic resources, such as the University of California-Santa Cruz (UCSC) Genome Browser. Galaxy allows users to upload data, parse it, reorder columns and change file formats for browser compatibility. Galaxy also provides several tools for data integration. For example, it has tools for data-set intersection and union analysis, which enables users to compare their data sets with annotated genomic loci and view the output directly on the Genome Browser. In the process, users can create and save not just new files but entire workflows that can be reused and shared with others. Best of all, Galaxy provides a platform for running tools developed by the community. In the near future, tools like Galaxy will provide bench scientists with a one-stop-shop for data analysis: given sequencing reads, add-ons will map these reads and call peaks, allowing for subsequent analyses.

Another popular online tool is DAVID[119], which is used for Gene Ontology analysis (for a step-by-step protocol, see REF. 120). Therefore, using the range of tools available online, with a few clicks one can map reads from chromatin immunoprecipitation followed by sequencing (ChIP–seq) using Galaxy, call peaks with CisGenome, use Galaxy's intersection tool to find overlapped genes and, finally, upload the transcription-factor-bound gene list to DAVID for Gene Ontology annotation (FIG. 4). Although not as efficient as a single tool, this method allows a substantial amount of analysis to be done without the need to write new software.

It is also important to note that known and novel motif finding for peaks or promoters can be done online using CEAS (Cis-regulatory Element Annotation System) and The MEME Suite. In addition to Gene Ontology annotations, understanding gene functions, pathway interactions or protein–protein interactions might be of interest for key genes. A number of online tools can now assist in this (STRING, Cytoscape and mouseNET are a few examples).

Using large-scale data sets for integrative analysis

One of the greatest challenges that comes with high-throughput technologies is the vast amount of data that they produce. The sheer amount of data produced can be difficult to manage, especially for experiments involving NGS methods. For example, Lister *et al.*[25] recently sequenced the human methylome using bisulphite shotgun sequencing, which generated 90 gigabases of sequence reads, representing 30× coverage of the human diploid genome. Transferring this amount of data to the National Center for Biotechnology Information (NCBI) public database servers took one full week. The question is how can investigators efficiently use data of this scale for comparative analyses? This challenge can be broadly divided into two: how can bench scientists look to see how one data set fits with others (from their own or other laboratories), and how can bioinformaticians provide better tools for integrated analyses?

For the bench scientist. To make strides in the era of NGS, we need tools for the bench scientist to analyse their own data in an efficient and straightforward manner. We propose that a solution would be similar to an open-source web browser, such as FireFox. It would have a series of 'add-ons', and a core group of programmers would maintain the browser code and listen to the community for ways of updating it. Importantly, the programmers would allow the community to build individual tools to enhance the browser's capabilities. The 'gatekeepers' would ensure that the tools produced are safe and work with the browser, and users could decide which add-ons are suited to their needs. Users would also see previews and read reviews and ratings for each add-on. A tool along these lines — Galaxy[112,113] — has been in development for many years and is described in BOX 3, along with other popular online tools.

One potential downside of an online analytical tool, such as Galaxy, is computational load. If the majority of scientists conducting RNA–seq or ChIP–seq experiments begin running Galaxy on a regular basis, will the whole system creep to a halt? Also, to prevent inefficient computation, add-ons would need to meet specific benchmarks for performance, such as time complexity and storage space, as the system cannot tolerate inefficient computation. Therefore it can be argued that it may be advisable to have a stand-alone analytical system. One example of such a tool is CisGenome[114], which is downloadable and compatible with several operating systems. Designed for the analysis of ChIP–chip and ChIP–seq data, it includes a browser, file-conversion tools and tools to call peaks of ChIP enrichment and to perform motif analysis. These features enable a basic workflow that is needed by many scientists. An example workflow using a range of tools is shown in FIG. 4.

Resources such as genome browsers are still some of our best tools. A good browser can distinguish good-quality from poor-quality data sets and can show trends and patterns in the data without the need for statistical measures. Such anecdotal observations can spur questions that require more sophisticated analysis. Several browsers are available, including Entrez Genome, Ensembl and the UCSC Genome Browser. Although the amount of data available on the UCSC browser, including many large-scale data sets[103], makes it very valuable, it can be slow when attempting to browse through several data sets at various locations. Other browsers, such as Anno-J, which was used for visualizing the *Arabidopsis thaliana* and human methylomes at nucleotide resolution[25,27], are much more dynamic. Scrolling through the genome is very rapid and tracks can be zoomed, scaled, re-ordered and removed almost instantly.

Bioinformatic hurdles. There are still a number of key issues in analysing NGS data, several of which have been touched on in previous reviews[4,6]. For example, it remains unclear how RNA–seq data from platforms

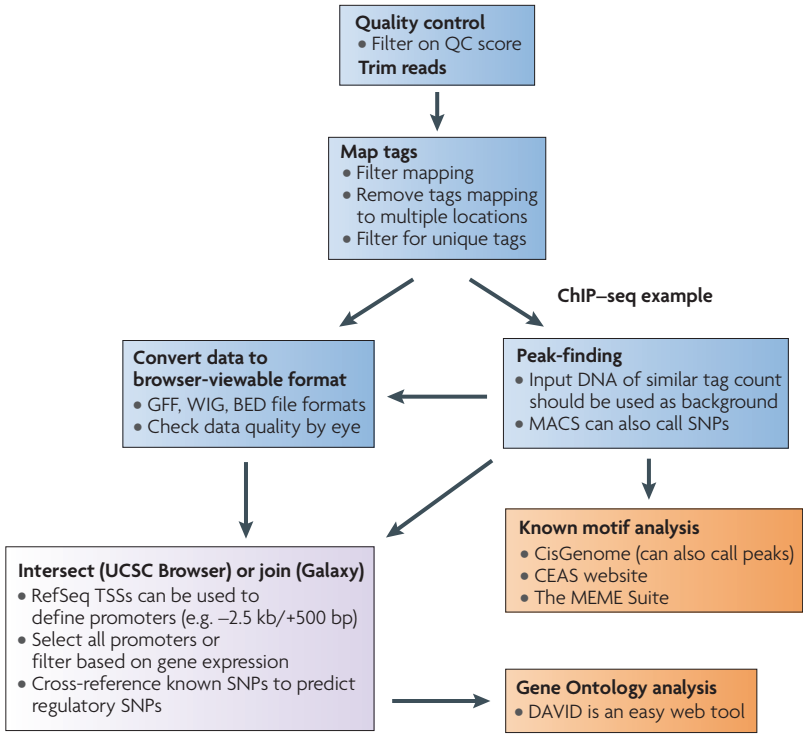

Figure 4 | **Flow chart for data analysis.** This example shows a workflow for the analysis of data from chromatin immunoprecipitation followed by sequencing (ChIP–seq). This analysis can be done by a bench scientist using current resources, and a similar strategy could be used for other types of next-generation sequencing data. Blue boxes show steps that can be performed using Galaxy. Integration or cross-sectioning of data can often be done in the University of California-Santa Cruz (UCSC) Genome Browser or by joining lists in Galaxy (purple box). Downstream steps, such as known motif analysis and Gene Ontology analysis, can be achieved with online or stand-alone tools (orange boxes). Galaxy can also be used to establish analytical pipelines for calling SNPs that could then be integrated into sequencing-based data, such as reads from ChIP–seq. CEAS, Cis-regulatory Element Annotation System; MACS, Model-based Analysis of ChIP–Seq; TSS, transcription start site.

that sequence short tags will be normalized against data from longer read platforms. Also, will RNA–seq methods be as universal as Affymetrix microarrays? Most scientists feel comfortable comparing their own and published Affymetrix platform data. It is still unclear in these early stages of data processing and normalization of RNA–seq how relative levels of expression can be compared, especially if there is a variation in the number of reads sequenced.

To address these questions more thoroughly, it will be important to revisit data normalization. Because NGS-based assays provide a digital read-out, the data is often used 'as is'. However, different experiments are sure to provide slightly varying degrees of enrichment, possibly due to antibody differences (for ChIP–seq or HITS-CLIP) or experimental variation. Therefore, two data sets used in a comparative analysis should first be normalized to each other. This applies to samples from different research groups, as well as to samples from within a data set. For example, if one experiment has a uniform reduction in peak height, non-normalized peak-finding may result in calling a cell-type-specific peak at a site that is actually shared. Normalization is therefore imperative in experiments that examine time points of differentiation or stages of disease progression in which the changes may be subtle between neighbouring stages[115]. In this regard, we will probably benefit from the numerous normalization methodologies that have been developed for microarray analysis. However, like gene-expression analysis, we are sure to find that one method does not fit all data sets and that Loess, quantile and rank-order normalizations will all be useful.

Future perspectives

Data integration itself is not an end: it is designed to generate novel hypotheses and help to test them. If a hypothetical 'data integrator' existed, its most important input would not be the data to be analysed but a specific question to answer. Depending on the question posed, analyses of the data — from what data sources are chosen to how normalization is performed, how controls are selected and precisely what is being calculated — can vary dramatically. A frequent misconception is that a data integrator is a black box that takes in data as input and generates interesting observations (or better, papers) as output. Unbiased integration strategies focus on a single question, whereas supervised integration can address any number of questions, so the scope of the types of analyses possible with supervised integration is much greater, and arguably endless. For this reason, it is unfeasible to automatically perform all possible integrated analyses, as if the data integrator were seeking both a question and its answer simultaneously. The choice of interesting questions must always be left to the researcher, and supervised integration must be tailored to each hypothesis. It is our opinion that, although unsupervised approaches can excel at finding patterns, it will be the supervised integrative methods stemming from either unsupervised methods or simple observations that will further our understanding of biology most effectively.

The future of genomic technologies holds great promise, but for genomic data and its integration to have a more meaningful impact on our understanding of biology, we must make an effort to link together all of the information that is being generated. This may require a community-wide effort, akin to Wikipedia, in which information can be updated by all but monitored for the correct citations that directly link to PubMed and the NCBI website. Each gene entry would be linked to a browser for visualizing all genomic and epigenomic information in a manner similar to viewing Gene Expression Omnibus (GEO) profiles on the NCBI website. All of the related information should be searchable with Google-like capabilities. That is, a search engine examines the entire text for terms and phrases and finds related information, even if it does not contain the exact key words. For example, NextBio currently provides a similar approach when searching for genes. This integration of knowledge will make each of us a better scientist through a greater understanding of the information around us.

1. Licatalosi, D. D. & Darnell, R. B. RNA processing and its regulation: global insights into biological networks. *Nature Rev. Genet.* **11**, 75–87 (2010).
2. Wang, Z., Gerstein, M. & Snyder, M. RNA–Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63 (2009).
3. Farnham, P. J. Insights from genomic profiling of transcription factors. *Nature Rev. Genet.* **10**, 605–616 (2009).
4. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nature Rev. Genet.* **10**, 669–680 (2009).
5. Metzker, M. L. Sequencing technologies — the next generation. *Nature Rev. Genet.* **11**, 31–46 (2010).
6. Laird, P. W. Principles and challenges of genome-wide DNA methylation analysis. *Nature Rev. Genet.* **11**, 191–203 (2010).
7. Beyer, A., Bandyopadhyay, S. & Ideker, T. Integrating physical and genetic maps: from genomes to interaction networks. *Nature Rev. Genet.* **8**, 699–710 (2007).
8. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
9. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
10. Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. *Nature Methods* **6**, 315–316 (2009).
11. Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature Methods* **6**, 99–103 (2009).
12. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nature Genet.* **41**, 1061–1067 (2009).
13. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA–Seq. *Nature Methods* **5**, 621–628 (2008).
14. Maher, C. A. *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97–101 (2009).
15. Gingeras, T. R. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009).
16. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
This study demonstrates the integration of epigenetic data with the human genome to annotate novel RNAs.
17. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
18. Bernstein, B. E., Meissner, A. & Lander, E. S. The mammalian epigenome. *Cell* **128**, 669–681 (2007).
19. Goldberg, A. D., Allis, C. D. & Bernstein, E. Epigenetics: a landscape takes shape. *Cell* **128**, 635–638 (2007).
20. Jones, P. A. & Baylin, S. B. The epigenomics of cancer. *Cell* **128**, 683–692 (2007).
21. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).
22. Li, E. Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Rev. Genet.* **3**, 662–673 (2002).
23. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
24. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* **316**, 1497–1502 (2007).
25. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).
In addition to providing the first human methylomes, this study conducts an integrative analysis of DNA methylation, histone modifications and RNA–seq.
26. Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* **452**, 215–219 (2008).
27. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
28. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766–770 (2008).
29. Pomraning, K. R., Smith, K. M. & Freitag, M. Genome-wide high throughput analysis of DNA methylation in eukaryotes. *Methods* **47**, 142–150 (2009).
30. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823–837 (2007).
31. Wang, Z. *et al.* Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genet.* **40**, 897–903 (2008).
32. Crawford, G. E. *et al.* DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nature Methods* **3**, 503–509 (2006).
33. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
34. Sabo, P. J. *et al.* Genome-scale mapping of DNase I sensitivity *in vivo* using tiling DNA microarrays. *Nature Methods* **3**, 511–518 (2006).
35. Dorschner, M. O. *et al.* High-throughput localization of functional elements by quantitative chromatin profiling. *Nature Methods* **1**, 219–225 (2004).
36. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nature Methods* **6**, 283–289 (2009).
37. Chi, S. W., Zang, J. B., Mele, A. & Darnell, R. B. Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. *Nature* **460**, 479–486 (2009).
38. Walhout, A. J. & Vidal, M. Protein interaction maps for model organisms. *Nature Rev. Mol. Cell Biol.* **2**, 55–62 (2001).
39. Hutchins, J. R. *et al.* Systematic analysis of human protein complexes identifies chromosome segregation proteins. *Science* **328**, 593–599 (2010).
40. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
41. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genet.* **38**, 1348–1354 (2006).
42. Dostie, J. *et al.* Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16**, 1299–1309 (2006).
43. Fullwood, M. J. & Ruan, Y. ChIP-based methods for the identification of long-range chromatin interactions. *J. Cell. Biochem.* **107**, 30–39 (2009).
44. Vassetzky, Y. *et al.* Chromosome conformation capture (from 3C to 5C) and its ChIP-based modification. *Methods Mol. Biol.* **567**, 171–188 (2009).
45. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
46. Fullwood, M. J. *et al.* An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
47. Duan, Z. *et al.* A three-dimensional model of the yeast genome. *Nature* **465**, 363–367 (2010).
48. Gobeil, S., Zhu, X., Doillon, C. J. & Green, M. R. A genome-wide shRNA screen identifies *GAS1* as a novel melanoma metastasis suppressor gene. *Genes Dev.* **22**, 2932–2940 (2008).
49. Gazin, C., Wajapeyee, N., Gobeil, S., Virbasius, C. M. & Green, M. R. An elaborate pathway required for Ras-mediated epigenetic silencing. *Nature* **449**, 1073–1077 (2007).
50. Bric, A. *et al.* Functional identification of tumor-suppressor genes through an *in vivo* RNA interference screen in a mouse lymphoma model. *Cancer Cell* **16**, 324–335 (2009).
51. Meacham, C. E., Ho, E. E., Dubrovsky, E., Gertler, F. B. & Hemann, M. T. *In vivo* RNAi screening identifies regulators of actin dynamics as key determinants of lymphoma progression. *Nature Genet.* **41**, 1133–1137 (2009).
52. Luo, J. *et al.* A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. *Cell* **137**, 835–848 (2009).
53. Zender, L. *et al.* An oncogenomics-based *in vivo* RNAi screen identifies tumor suppressors in liver cancer. *Cell* **135**, 852–864 (2008).
54. Schuldiner, M. *et al.* Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123**, 507–519 (2005).
55. Roguev, A., Wiren, M., Weissman, J. S. & Krogan, N. J. High-throughput genetic interaction mapping in the fission yeast *Schizosaccharomyces pombe*. *Nature Methods* **4**, 861–866 (2007).
56. Hannum, G. *et al.* Genome-wide association data reveal a global map of genetic interactions among protein complexes. *PLoS Genet.* **5**, e1000782 (2009).
57. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genet.* **39**, 311–318 (2007).
58. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
59. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880 (2005).
60. Kim, T. H. *et al.* Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231–1245 (2007).
61. Hartman, S. E. *et al.* Global changes in STAT target selection and transcription regulation upon interferon treatments. *Genes Dev.* **19**, 2953–2968 (2005).
62. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods* **4**, 651–657 (2007).
63. Visel, A. *et al.* ChIP–seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
64. O'Geen, H. *et al.* Genome-wide analysis of KAP1 binding suggests autoregulation of KRAB-ZNFs. *PLoS Genet.* **3**, e89 (2007).
65. Lee, T. I. *et al.* Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301–313 (2006).
66. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9**, 356–369 (2008).
67. Marks, H. *et al.* High-resolution analysis of epigenetic changes associated with X inactivation. *Genome Res.* **19**, 1361–1373 (2009).
68. Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).
69. Pomerantz, M. M. *et al.* The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nature Genet.* **41**, 882–884 (2009).
70. Wright, J. B., Brown, S. J. & Cole, M. D. Upregulation of c-MYC in *cis* through a large chromatin loop linked to a cancer risk-associated single-nucleotide polymorphism in colorectal cancer cells. *Mol. Cell. Biol.* **30**, 1411–1420 (2010).
71. Tuupanen, S. *et al.* The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nature Genet.* **41**, 885–890 (2009).
72. Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328**, 232–235 (2010).
This study shows that individual binding variability for RNAPII and NF-κB is linked to SNPs and structural variants that alter individual gene expression levels. The binding data enables functional annotation of regulatory SNPs.
73. Gaulton, K. J. *et al.* A map of open chromatin in human pancreatic islets. *Nature Genet.* **42**, 255–259 (2010).
These authors used open chromatin maps to recover a type 2 diabetes-associated SNP in the intron of transcription factor 7-like 2 (*TCF7L2*). Functional assays confirmed its role in enhancer activity.
74. Zhang, Y. *et al.* Model-based Analysis of ChIP–Seq (MACS). *Genome Biol.* **9**, R137 (2008).
75. Degner, J. F. *et al.* Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* **25**, 3207–3212 (2009).
76. Maynard, N. D., Chen, J., Stuart, R. K., Fan, J. B. & Ren, B. Genome-wide mapping of allele-specific protein–DNA interactions in human cells. *Nature Methods* **5**, 307–309 (2008).
77. McDaniell, R. *et al.* Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328**, 235–239 (2010).
This study shows that nucleotide sequences in human regulatory elements are variable, which suggests that these elements may contain regulatory SNPs.
78. Hellman, A. & Chess, A. Gene body-specific methylation on the active X chromosome. *Science* **315**, 1141–1143 (2007).
79. Edwards, C. A. & Ferguson-Smith, A. C. Mechanisms regulating imprinted genes in clusters. *Curr. Opin. Cell Biol.* **19**, 281–289 (2007).

80. Marson, A. *et al.* Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134**, 521–533 (2008).
81. Pandey, R. R. *et al. Kcnq1ot1* antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol. Cell* **32**, 232–246 (2008).
82. Nagano, T. *et al.* The *Air* noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* **322**, 1717–1720 (2008).
83. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA* **106**, 11667–11672 (2009).
84. Kolasinska-Zwierz, P. *et al.* Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature Genet.* **41**, 376–381 (2009).
85. Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C. & Komorowski, J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res.* **19**, 1732–1741 (2009).
86. Schwartz, S., Meshorer, E. & Ast, G. Chromatin organization marks exon–intron structure. *Nature Struct. Mol. Biol.* **16**, 990–995 (2009).
87. Luco, R. F. *et al.* Regulation of alternative splicing by histone modifications. *Science* **327**, 996–1000 (2010).
88. Spies, N., Nielsen, C. B., Padgett, R. A. & Burge, C. B. Biased chromatin signatures around polyadenylation sites and exons. *Mol. Cell* **36**, 245–254 (2009).
89. Hon, G., Wang, W. & Ren, B. Discovery and annotation of functional chromatin signatures in the human genome. *PLoS Comput. Biol.* **5**, e1000566 (2009).
90. Schubeler, D. *et al.* The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev.* **18**, 1263–1271 (2004).
91. Licatalosi, D. D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464–469 (2008).
92. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA* **100**, 15776–15781 (2003).
93. Shumaker, D. K. *et al.* Mutant nuclear lamin A leads to progressive alterations of epigenetic control in premature aging. *Proc. Natl Acad. Sci. USA* **103**, 8703–8708 (2006).
94. Zeng, W. *et al.* Specific loss of histone H3 lysine 9 trimethylation and HP1γ/cohesin binding at D4Z4 repeats is associated with facioscapulohumeral dystrophy (FSHD). *PLoS Genet.* **5**, e1000559 (2009).
95. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106–1117 (2008).
96. Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP–seq experiments for DNA-binding proteins. *Nature Biotech.* **26**, 1351–1359 (2008).
97. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887–898 (2008).
98. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
99. Hon, G., Ren, B. & Wang, W. ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. *PLoS Comput. Biol.* **4**, e1000201 (2008).
100. Spirin, V. & Mirny, L. A. Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA* **100**, 12123–12128 (2003).
101. Mikkelsen, T. S. *et al.* Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454**, 49–55 (2008).
102. Hawkins, R. D. *et al*. Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479–491 (2010).
103. Rosenbloom, K. R. *et al.* ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res.* **38**, D620–D625 (2010).
104. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
105. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
106. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
107. Oberdoerffer, S. *et al.* Regulation of CD45 alternative splicing by heterogeneous ribonucleoprotein, hnRNPLL. *Science* **321**, 686–691 (2008).
108. Needham, C. J., Bradford, J. R., Bulpitt, A. J. & Westhead, D. R. Inference in Bayesian networks. *Nature Biotech.* **24**, 51–53 (2006).
109. van Steensel, B. *et al.* Bayesian network analysis of targeting interactions in chromatin. *Genome Res.* **20**, 190–200 (2010).
This is an excellent example of using supervised integration with a Bayesian network to predict interactions among chromatin-associated proteins, then validating the findings experimentally.
110. Yu, H., Zhu, S., Zhou, B., Xue, H. & Han, J. D. Inferring causal relationships among different histone modifications and gene expression. *Genome Res.* **18**, 1314–1324 (2008).
111. Jansen, R. *et al.* A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* **302**, 449–453 (2003).
112. Taylor, J., Schenck, I., Blankenberg, D. & Nekrutenko, A. Using Galaxy to perform large-scale interactive data analyses. *Curr. Protoc. Bioinformatics* **Chapter10**, Unit 10.5 (2007).
113. Blankenberg, D. *et al.* A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Res.* **17**, 960–964 (2007).
114. Ji, H. *et al.* An integrated software system for analyzing ChIP–chip and ChIP–seq data. *Nature Biotech.* **26**, 1293–1300 (2008).
115. Taslim, C. *et al.* Comparative study on ChIP–seq data: normalization and binding pattern characterization. *Bioinformatics* **25**, 2334–2340 (2009).
116. Celniker, S. E. *et al.* Unlocking the secrets of the genome. *Nature* **459**, 927–930 (2009).
117. Collins, S. R. *et al.* Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446**, 806–810 (2007).
118. Jaschek, R. & Tanay, A. Spatial clustering of multivariate genomic and epigenomic information. *Lect. Notes Comput. Sci.* **5541**, 170–183 (2009).
119. Dennis, G. Jr *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4**, P3 (2003).
120. Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protoc.* **4**, 44–57 (2009).
121. Lupien, M. *et al.* FoxA1 translates epigenetic signatures into enhancer-driven lineage-specific transcription. *Cell* **132**, 958–970 (2008).
122. Roh, T. Y., Cuddapah, S. & Zhao, K. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.* **19**, 542–552 (2005).
123. Roh, T. Y., Wei, G., Farrell, C. M. & Zhao, K. Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. *Genome Res.* **17**, 74–81 (2007).

Acknowledgements
We apologize to those authors whose work we were unable to reference owing to limitations of space. R.D.H is supported by a postdoctoral fellowship from the American Cancer Society. We acknowledge generous funding from the Ludwig Institute for Cancer Research, the US National Institutes of Health, the California Institute of Regenerative Medicine and the Juvenile Diabetes Research Foundation. We thank the anonymous reviewers for their valuable comments on earlier versions of this Review.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Bing Ren's homepage: http://licr-renlab.ucsd.edu
1000 Genomes: http://1000genomes.org
Anno-J: http://www.annoj.org
The Cancer Genome Atlas: http://tcga.cancer.gov
CEAS: http://ceas.cbi.pku.edu.cn
CisGenome: http://www.biostat.jhsph.edu/~hji/cisgenome
Cytoscape: http://www.cytoscape.org
DAVID: http://david.abcc.ncifcrf.gov
ENCODE Project: http://www.genome.gov/10005107
Entrez Genome: http://www.ncbi.nlm.nih.gov/sites/genome
Ensembl: http://www.ensembl.org
Galaxy: http://galaxy.psu.edu
Gene Ontology: http://www.geneontology.org
Gene Set Enrichment Analysis: http://www.broadinstitute.org/gsea
International HapMap Project: http://hapmap.ncbi.nlm.nih.gov
MACS: http://liulab.dfci.harvard.edu/MACS
The MEME Suite: http://meme.sdsc.edu/meme
mouseNET: http://mousenet.princeton.edu
NCBI: http://www.ncbi.nlm.nih.gov
NextBio: http://www.nextbio.com
PubMed: http://www.ncbi.nlm.nih.gov/pubmed
Roadmap Epigenomics Project: http://www.roadmapepigenomics.org
STRING: http://string-db.org
UCSC Genome Browser: http://genome.ucsc.edu

ALL LINKS ARE ACTIVE IN THE ONLINE PDF