Allele-specific and heritable chromatin signatures in humans

Ewan Birney[1,*], Jason D. Lieb[2], Terrence S. Furey[3], Gregory E. Crawford[3,4] and Vishwanath R. Iyer[5]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]Department of Biology, Carolina Center for Genome Sciences and Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, [3]Institute for Genome Sciences and Policy (IGSP) and [4]Department of Pediatrics, Division of Medical Genetics, Duke University, Durham, NC 27708, USA and [5]Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, Section of Molecular Genetics and Microbiology, University of Texas, Austin, TX 78712, USA

Next-generation sequencing-based assays to detect gene regulatory elements are enabling the analysis of individual-to-individual and allele-specific variation of chromatin status and transcription factor binding in humans. Recently, a number of studies have explored this area, using lymphoblastoid cell lines. Around 10% of chromatin sites show either individual-level differences or allele-specific behavior. Future studies are likely to be limited by cell line accessibility, meaning that white-bloodcell-based studies are likely to continue to be the main source of samples. A detailed understanding of the relationship between normal genetic variation and chromatin variation can shed light on how polymorphisms in non-coding regions in the human genome might underlie phenotypic variation and disease.

INTRODUCTION

The advent of next-generation sequencing (1) has revolutionized research in gene regulation. The low cost of obtaining genome-wide information for transcription factor binding, histone modifications and chromatin status has enabled the extensive study of various regulatory features of different cell types in a variety of organisms (2–4). Systematic projects such as ENCODE (5), the human Epigenome Roadmap (now expanding into the International Human Epigenome Consortium) (6) and the fly and worm modENCODE projects (7) have focussed on defining the chromatin state, how it varies between different cell types, how such stable 'cellular memory' endpoints differ and how, during development, such cellular memory is determined.

Most of these studies to date have characterized the chromatin state for a series of specific cell types, each from only a single individual and averaged over both chromosomes in the cell. However, individuals are likely to differ at some level in their chromatin state for a particular cell type, and each cell contains pairs of homologous chromosomes whose chromatin structure and expression status are not necessarily identical. Heterozygous base pairs present between homologous chromosomes enable the detection of allele-specific signals in sequence-based assays (Fig. 1). Recent studies have found that gene expression and chromatin structure do indeed differ between homologous chromosomes in the same cell, and accordingly, between the same cell types obtained from different individuals (8–10). Learning the relationship between genetic variation and variation in chromatin offers the potential to bridge the gap between genome-wide association studies (GWASs) that have linked many diseases to single nucleotide polymorphisms (SNPs) and an understanding of how such polymorphisms, most of which are found in non-coding regions, can underlie phenotypic variation. Here, we review the relevant studies to date, focussing on individual and allele-specific chromatin rather than gene expression at the level of RNA, and outline key technical considerations and potential future directions for detecting individual and allele-specific differences in chromatin organization.

*To whom correspondence should be addressed. Tel: +44 1223494420; Fax: +44 1223496611; Email: birney@ebi.ac.uk
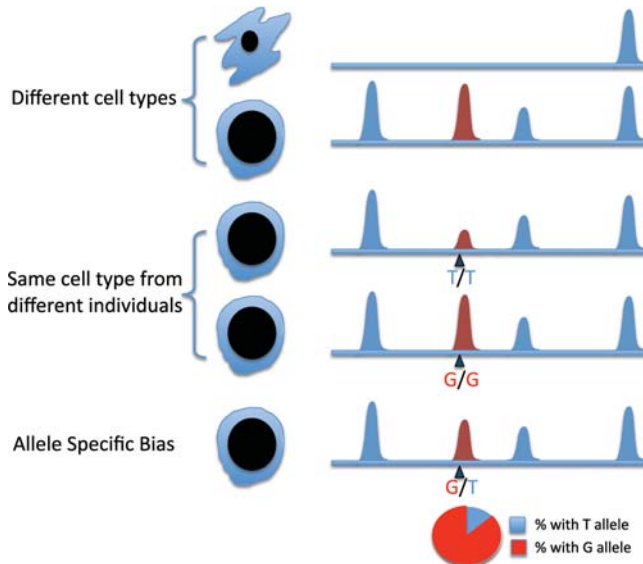
Figure 1. This figure shows three different types of experiments for assaying differences in chromatin state in equivalent and different cell types. The blue lines represent the genome sequence in a region; the blue and maroon peaks show regions of enrichment from a chromatin-based assay, such as DNaseI, FAIRE or ChIP-seq. The top panel shows results from an idealized experiment between two different cell lines, often with considerable differences. The middle panel shows an experiment between two different individuals which differ in a smaller proportion of peaks, and often, though not always, this is associated with local genetic changes, such as an SNP difference (shown as black arrows with the genotypes G/G and T/T nearby). The bottom panel shows an experiment on a single individual in which the presence of a heterozygote SNP allows the relative proportion of reads contributed by each allele to be known, as illustrated in the pie chart.

ASSAYS USED TO DETECT INDIVIDUAL-SPECIFIC AND ALLELE-SPECIFIC CHROMATIN SIGNATURES

A number of assays analyze chromatin state and identify active gene regulatory elements genome-wide including mapping DNaseI hypersensitive sites (DHSs) (11,12), formaldehyde-assisted isolation of regulatory elements (FAIRE) (13) and chromatin immunoprecipitation (ChIP) (reviewed in 14). Figure 2 schematically shows how these assays work. DHSs and FAIRE identify active regulatory elements through detection of nucleosome-free regions, whereas ChIP identifies specific transcription factor-binding sites and presence of specific histone variants and histone tail modifications. One can also study DNA methylation using a number of methods (15), but that is beyond the scope of this review and will not be discussed here. DHSs, FAIRE and ChIP are distinct methods and the strengths and limitations of using these assays in large population studies are further discussed below.

DHSs represent regions of the genome where nucleosomes have been displaced by transcription factors, making them hypersensitive to DNaseI digestion. These regions are commonly described as 'open' chromatin, whereas the remaining regions are 'closed'. DHSs can robustly identify all different types of active regulatory elements, including promoters,

enhancers, silencers, insulators and locus control regions. While DHSs do not directly reveal which transcription factor(s) are binding to each region, it does identify in a general sense where the functional regulatory elements of the genome are and whether they are open or closed across diverse cell types, as well as within the same cell type across many individuals.

FAIRE uses formaldehyde to biochemically separate DNA that is packaged in nucleosomes from DNA that is bound by non-nucleosomal proteins like transcription factors. Although FAIRE is also enriching for open chromatin regions, it is methodologically independent from DNaseI experiments and therefore complementary. It is also comprehensive in that it is an inherently genome-wide method that enriches for all known classes of regulatory elements. Since FAIRE uses for-maldehyde, a distinct advantage of this method is that it can readily be used on fixed frozen tissues.

ChIP precisely determines the location of specific DNA-associated proteins, histone variants and histone modifications within the genome, which is more informative than general open chromatin data generated by DNaseI and FAIRE. Specific factors, variants and/or modifications can be targeted for analysis depending on the disease or suspected gene involvement. While ChIP provides very specific information about factor location, this assay is limited to factors that have high-quality ChIP-grade antibodies and only one factor is tested per experiment. Tagged versions of proteins are an alternative option for cultured cells, but not suitable for studying primary cell types or tissues.

The original implementations of these methods to study chromatin involved detection of specific signals using Southern blots, PCR or microarray hybridization, but all of these methods have now been adapted to use next-generation sequencing (DNase-seq, FAIRE-seq and ChIP-seq), which in addition to providing a genome-wide readout, also offers the opportunity to resolve allele-specific signals. DNaseI, FAIRE and ChIP experiments generate libraries of DNA fragments that are enriched in genomic regions corresponding to open chromatin (DNaseI and FAIRE), or that were cross-linked to targets of specific antibodies (ChIP). These fragments will vary in size depending on the protocol, but all are amenable to construction of sequencing libraries using any of the currently available platforms. Each sequencing experiment generates tens of millions of short sequence reads that provide a sampling of the DNA in the constructed library.

To determine regions of open chromatin, ChIP targets and allele-specific biology, sequence reads must be aligned to a reference genomic sequence. A large and growing number of software packages are available to align reads and further process these data (16–18). The short length of the sequence reads, the repetitive nature of large mammalian genomes and the incompleteness of specific types of regions within the reference genome create challenges that must be carefully considered, particularly for detection of allele-specific signals at heterozygous SNPs because of the effect of apparent mismatches on shorter reads as described below. Paired-end sequencing of both ends of an enriched DNA fragment can alleviate some of the inherent uncertainty of aligning reads to the reference genome.
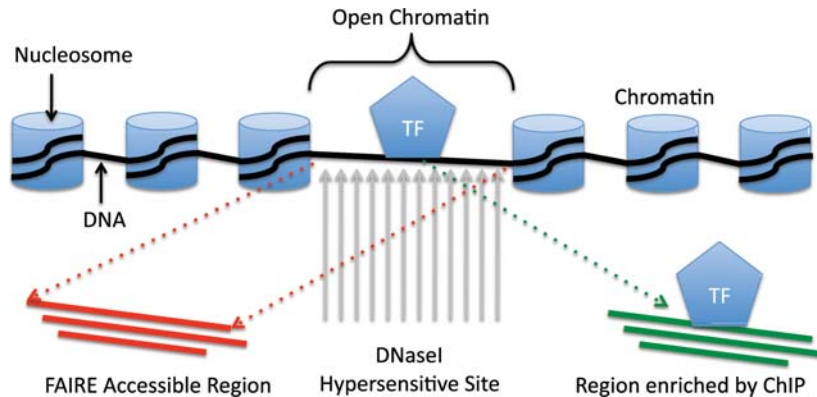
Figure 2. A schematic diagram of three chromatin assays. The blue cylinders represent the core histones, around which the DNA is wrapped. The pentagon represents a bound transcription factor (TF) which has displaced histones in a particular region; this region is described as 'open chromatin'. The FAIRE assay enriches for such open chromatin regions by differential solubility in phenol. The DNaseI assay utilizes the fact that regions of the open chromatin are much more susceptible to DNaseI digestion. The ChIP assay uses a specific antibody to enrich for DNA regions binding to a specific transcription factor or a modified histone.

CELL TYPES AMENABLE TO STUDYING DIFFERENCES IN CHROMATIN AMONG INDIVIDUALS

Chromatin signatures vary considerably between different cell types, so when exploring inter-individual differences, it is important to acquire the identical cell type from each individual. Furthermore, for disease-related studies, it would be ideal to study disease-relevant tissues. However, the heterogeneity of cells within intact tissues, and even sub-compartments within tissues, makes acquiring such pure cell types especially challenging. Obtaining intact tissue from humans also often requires invasive procedures that typically take place during medical intervention, meaning that the tissue obtained may be affected by an unrelated medical condition, further complicating downstream analyses. Organ donations from cadavers are a valuable source of tissues, but often these tissues are sub-optimal for analysis due to the degradation that may take place between death and tissue acquisition.

Because of these constraints, many studies have focussed on readily accessible white blood cell lineages. While these cells are not ideal for studying diseases unrelated to immune cell function, one distinct advantage is that they can be sorted using cell surface markers to isolate relatively pure sub-populations of cells in high numbers. However, there are challenges with this procedure as well. For example, sorting cells using positive selection often activates cells, which may confound analyses depending on the immune cell type, health of the individual at the time of blood draw and previous environmental exposures. Using negative selection, which avoids activating the cells, to isolate cell populations is sometimes possible, but cell numbers and purity are often sub-optimal. Even populations of blood cell lineages considered 'pure' by sorting are often made up of additional known and unknown sub-populations of cells, which likely fluctuate within individuals on a daily basis. Further studies are required to determine whether other more homogenous primary blood cell populations, for example neutrophils, might be better suited for

large population studies. Regardless, it remains a challenge to find a readily accessible, perfectly matched cell type from a large population of individuals.

Currently, the largest set of accessible cells from different individuals is based on Epstein–Barr-virus-transformed white blood cells, called 'lymphoblastoid lines'. One key advantage of these cells for analyzing variation in chromatin is that they have been extensively genotyped by the HapMap and 1000 Genomes project, and archived lines are readily available. They are derived from a mixed population of naive and memory B-cells that likely differ between individuals at the time of sampling. Despite this, the studies described below show that there are clear correlations between specific DNA sequence variants and the chromatin readouts. This indicates that at least some of the differences are due to individual-specific genetics, rather than variations in isolation techniques or cell-type heterogeneity. On the other hand, the effects of genetic variation and allele-specific gene expression can occur in a tissue- or cell-type-specific manner (8,19,20), indicating that results obtained from studying a single cell type may not be universally applicable to all cell types. The other readily accessible cell type is skin fibroblasts, which from a small skin biopsy can be expanded in culture to large cell numbers.

Another potential approach for studying alternative cell types is to use induced pluripotent stem cells (iPSCs) (reviewed in 21), where cells from specific individuals can be reprogramed into pluripotent cells that can be differentiated into many cell types. However, it is possible that a residual epigenetic memory of the parental cell type is retained through reprograming (22). Matched primary and iPSC-derived cell types analyzed in reasonably large numbers will be needed before one can assess how well, if at all, iPSC-based cell types could be used for exploring inter-individual chromatin differences. In addition, advances in iPSC manipulation, such as differentiating iPS cells into pure populations of disease-relevant cell types will be needed for this technology to be used routinely.

STUDIES OF VARIATION IN HUMAN CHROMATIN SIGNATURES AMONG ALLELES AND INDIVIDUALS

For assaying individual-specific variation, data from the same cell or tissue type from different individuals are required, ideally processed under the same conditions. For assaying allele-specific variation, the underlying genotype must additionally be known so that it can be associated with any observed allele-specific differences in binding or chromatin.

Allele-specific transcription factor binding has been assayed by reading out the results of ChIP by microarray hybridization for RNA pol II and various histone modifications on SNP genotyping arrays (23,24). One limitation of these studies is that only heterozygous sites that were preselected for inclusion on the array could be assayed, and of these, only the polymorphisms that overlapped a binding site were informative. Thus, only a small number of allele-specific-binding sites could be identified. ChIP-seq can be used to detect allele-specific differences in factor binding or chromatin at all heterozygous sites in a single individual by analyzing the two alleles separately. Indeed, the earliest ChIP-seq studies recognized this feature of data from these experiments and identified several instances of histone modifications where the signal from the two alleles covering a heterozygous site was significantly different (25).

These early studies were performed in cell lines that were not comprehensively and independently genotyped, thus limiting the number of sites at which the allele-specificity of each histone modification could be assayed. The availability of comprehensive genotyping data from the 1000 Genomes Project for a large set of lymphoblastoid cell lines vastly increases the number of sites with an informative underlying genotype. Two recent studies have taken advantage of the combination of next generation sequencing and genotype information to examine the extent and nature of allele-specific and individual-specific transcription factor binding and chromatin in lymphoblastoid cell lines. McDaniell *et al.* (9) measured allele-specific binding of the multifunctional transcription factor CTCF as well as individual-specific DHS open chromatin sites for six individuals from two parent–child trios. About 10% of DHS sites were found to be individual-specific, with patterns of occurrence that were consistent with inheritance. In the analysis of allele-specific signals, they observed that, in general, sequence reads that contained the reference genome allele tended to align at higher rates compared with sequence reads containing the alternate allele, generating an artifactual bias that can have the appearance of allele-specific binding. This reference bias has also been noted in similar analyses performed with RNA-seq data (26) and precautions must be taken to account and correct for it during the sequence alignment process. After the appropriate corrections, McDaniell *et al.* found that approximately 11% of assayable CTCF-binding sites in the human genome were allele-specific. Importantly, the direction of the allele-specificity was highly correlated across individuals that shared the same heterozygous genotype, and the relative strengths of the signal in homozygous parents were generally concordant with the allele-specificity at the corresponding heterozygous site in the child. Moreover, polymorphisms that were most likely to show allele-specific

binding generally corresponded to highly conserved nucleotide positions in the CTCF-binding motif. These observations indicated that allele-specificity of CTCF binding was genetic rather than epigenetic or stochastic in origin and could be inherited.

In a contemporaneous study, Kasowski *et al.* (10) examined individual-specific and allele-specific differences in the binding of the transcription factor NF-κB and RNA polymerase II across 10 lymphoblastoid cells that were also genotyped by the 1000 Genomes Project. Their approach was to first identify individual-specific binding events, and then relate these to the underlying genotype, thereby avoiding alignment bias issues. This study found that a significant proportion of individual differences in binding was due to underlying SNPs or structural variation such as deletions affecting the NF-κB-binding motif or the TATA element in the promoter in the case of variation in RNA pol II binding. Interestingly, the number of SNPs in a factor's defined binding regions was proportional to the extent of the observed binding differences. They also showed that this variation in binding was correlated with differences in gene expression, indicating a direct functional outcome of this genetic variation. Additionally, Kasowski *et al.* analyzed differences in binding between humans and chimpanzees and showed that the inter-individual differences between humans were less than inter-species differences.

Perhaps, the most important contribution of these studies is the demonstration that such individual-specific and allele-specific differences exist in transcription factor binding and open chromatin, that they can be reliably measured, that they have consequences on downstream events such as expression and that at least some proportion of these differences is due to heritable genetic variation.

RELEVANCE OF ALLELE-SPECIFIC AND INDIVIDUAL-SPECIFIC CHROMATIN STUDIES TO HUMAN DISEASE

GWASs have linked hundreds of specific SNPs to disease risk, but in most cases the causal connection between the SNPs and the phenotype is unknown, partly because most of these SNPs are non-coding. These SNPs are often assumed to be regulatory in function, but our annotation of regulatory elements is far from complete. Moreover, the effect of nucleotide variation on the activity of those regulatory elements has been far harder to understand, leading to difficulties in designing further experiments. It is likely that much of the nucleotide variation causally linked to human disease (and a multitude of other phenotypic traits) does indeed occur in regulatory elements. We hypothesize that these causal variants are manifested as changes in chromatin organization caused either by the influence of nucleotide sequence on DNA packaging directly, or by affecting the binding of regulatory factors to DNA. For example, a polymorphism associated with type 2 diabetes differentially affects open chromatin in pancreatic islet cells as detected by the FAIRE assay (27). A systematic effort to uncover the relationship between genetic variation in humans on the one hand and variation in chromatin structure and transcription factor binding on the other can dramatically narrow our search for the genetic cause of increased disease

risk, and simultaneously will provide insight into disease mechanism. Detailed analysis of allele-specific and individual-specific chromatin signatures across a broad range of genotyped individuals, as well in disease patients and normal controls, can help bridge the gap between our ability to detect genetic variation linked to disease and our ability to explain how that variation causes disease.

FUTURE PROSPECTS

The ability to cost-effectively analyze a variety of chromatin-related features is already fueling a revolution in chromatin and gene regulation studies, in particular associated with development and cell identity. The recent studies described above show that such differences exist and are biologically relevant. However, there are still considerable hurdles to overcome before such profiling can be deployed on a large scale, in particular in the context of disease studies.

A major issue will be sample acquisition, handling and purity. The prevalence and easy access of lymphoblastoid cell lines, which are present in many disease-relevant cohorts and a number of prospective studies [for example, ALSPAC (28)] means that differences measured in these cell lines would be applicable to a large set of existing studies. Owing to the long-term logistical planning for disease cohorts, it is worth considering banking other cell types (e.g. fibroblasts) and potentially primary cultures or stable derivatives of primary cultures, such as frozen, fixed chromatin preparations in current or future cohorts. Cell types most relevant to a disease would be ideal, but that must be balanced with cost, accessibility of cells and, in many cases, a requirement for invasive procedures to obtain the appropriate tissue. It seems likely that as with eQTLs, a proportion of chromatin-specific events will be detectable in many tissues, and a proportion detectable in only a restricted set. Thus, an 'inappropriate' cell line might still be informative, though clearly not as desirable as the actual disease tissue.

Assay development will continue to be critical. The relatively large number of cells currently required (1–10 m per assay) increases the logistical and purification challenges. It seems likely that improvements to the chromatin assays and to DNA sequencing will allow lower cell numbers to be used. Currently proof-of-principle experiments have worked for between 10 000 and 100 000 cells (29) for histone-modification-based ChIP-seq. Computational analysis routines are also likely to develop considerably over the coming years, and the decreases in sequencing costs will allow deeply sequencing libraries to become more routine. This will both allow better resolution of weaker enrichment signals, and allow more sites to be assessed for allele-specific biases.

Animal systems will continue to be informative since they do not have the limitations of human studies surrounding tissue accessibility, long generation times and a lack of structured genetic crosses to study potential inheritance. Next-generation sequencing allows a far deeper genetic understanding in model organisms, and the increasing number of outbred or pseudo-outbred populations (30,31) will allow experiments analogous to human disease cohorts.

It is still an open question which chromatin assay will be the most informative for understanding individual differences. Each assay's relationship to other assays, both standard chromatin assays and other informative, related assays, such as RNA-seq, still needs to be determined for individual-level differences. More general assays such as DNaseI, FAIRE and histone-modification-based ChIP capture more regulatory sites, but their signal characteristics are less appropriate for allele-specific analysis because of the more diffuse signal around enriched regions. ChIP-seq of transcription factors is highly focussed on a single protein and might miss important biological phenomena in a cell, but it often can be better integrated, for example with DNA motif analysis, and has stronger allele-specific signals. In the short term, local expertise and practicalities are likely to drive assay choice at least initially. Another important aspect to explore is the difference of lymphoblastoid lines from their parent B cells in the context of individual variation. A well-structured small 'normal' cohort, in which both many primary cells and lymphoblastoid cell lines were derived and on which many chromatin assays could be performed, would answer many of these questions. We should consider creating such a baseline resource for the correlation between chromatin patterns as being analogous to the HapMap project creating a baseline SNP and a correlation pattern between genetic variants. From such a baseline resource, we would be able to determine the optimal assay combination, at least for the studied cell type(s), and an initial set of genetically influenced variable chromatin sites. These studies would be invaluable for the planning and analysis of any future disease study on chromatin effects, though it is worth noting that unlike the genome-wide association scenario, where the use of pooled controls have become commonplace, it is likely that study-specific control samples would continue to be needed for these more experimentally demanding assays.

Conflict of Interest statement. None declared.

FUNDING

REFERENCES

1. Bentley, D.R. (2006) Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.*, **16**, 545–552.
2. Hawkins, R.D., Hon, G.C., Lee, L.K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L.E., Kuan, S., Luu, Y., Klugman, S. *et al.* (2010) Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell. Stem Cell.*, **6**, 479–491.
3. Turck, F., Roudier, F., Farrona, S., Martin-Magniette, M.L., Guillaume, E., Buisine, N., Gagnot, S., Martienssen, R.A., Coupland, G. and Colot, V. (2007) Arabidopsis TFL2/LHP1 specifically associates with genes marked by trimethylation of histone H3 lysine 27. *PLoS Genet.*, **3**, e86.
4. Akkers, R.C., van Heeringen, S.J., Jacobi, U.G., Janssen-Megens, E.M., Francoijs, K.J., Stunnenberg, H.G. and Veenstra, G.J. (2009) A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in xenopus embryos. *Dev. Cell*, **17**, 425–434.
5. ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia of DNA elements) project. *Science*, **306**, 636–640.

6. American Association for Cancer Research Human Epigenome Task Force and European Union, Network of Excellence, Scientific Advisory Board (2008) Moving AHEAD with an international human epigenome project. *Nature*, **454**, 711–715.

7. Celniker, S.E., Dillon, L.A., Gerstein, M.B., Gunsalus, K.C., Henikoff, S., Karpen, G.H., Kellis, M., Lai, E.C., Lieb, J.D., MacAlpine, D.M. *et al.* (2009) Unlocking the secrets of the genome. *Nature*, **459**, 927–930.

8. Gregg, C., Zhang, J., Weissbourd, B., Luo, S., Schroth, G.P., Haig, D. and Dulac, C. (2010) High-resolution analysis of parent-of-origin allelic expression in the mouse brain. *Science*, **329**, 643–648.

9. McDaniell, R., Lee, B.K., Song, L., Liu, Z., Boyle, A.P., Erdos, M.R., Scott, L.J., Morken, M.A., Kucera, K.S., Battenhouse, A. *et al.* (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*, **328**, 235–239.

10. Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S.M., Habegger, L., Rozowsky, J., Shi, M., Urban, A.E. *et al.* (2010) Variation in transcription factor binding among humans. *Science*, **328**, 232–235.

11. Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S. and Crawford, G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.

12. Hesselberth, J.R., Chen, X., Zhang, Z., Sabo, P.J., Sandstrom, R., Reynolds, A.P., Thurman, R.E., Neph, S., Kuehn, M.S., Noble, W.S. *et al.* (2009) Global mapping of protein–DNA interactions *in vivo* by digital genomic footprinting. *Nat. Methods*, **6**, 283–289.

13. Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. and Lieb, J.D. (2007) FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Res.*, **17**, 877–885.

14. Park, P.J. (2009) ChIP-seq: Advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.

15. Butcher, L.M. and Beck, S. (2008) Future impact of integrated high-throughput methylome analyses on human health and disease. *J. Genet. Genomics*, **35**, 391–401.

16. Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, 1135–1145.

17. Flicek, P. and Birney, E. (2009) Sense from sequence reads: methods for alignment and assembly. *Nat. Methods*, **6**, S6–S12.

18. Pepke, S., Wold, B. and Mortazavi, A. (2009) Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, **6**, S22–32.

19. Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M. *et al.* (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, **325**, 1246–1250.

20. Gregg, C., Zhang, J., Butler, J.E., Haig, D. and Dulac, C. (2010) Sex-specific parent-of-origin allelic expression in the mouse brain. *Science*, **329**, 682–685.

21. Yamanaka, S. and Blau, H.M. (2010) Nuclear reprogramming to a pluripotent state by three approaches. *Nature*, **465**, 704–712.

22. Polo, J.M., Liu, S., Figueroa, M.E., Kulalert, W., Eminli, S., Tan, K.Y., Apostolou, E., Stadtfeld, M., Li, Y., Shioda, T. *et al.* (2010) Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. *Nat. Biotechnol.*, **28**, 848–855.

23. Kadota, M., Yang, H.H., Hu, N., Wang, C., Hu, Y., Taylor, P.R., Buetow, K.H. and Lee, M.P. (2007) Allele-specific chromatin immunoprecipitation studies show genetic influence on chromatin state in human genome. *PLoS Genet.*, **3**, e81.

24. Maynard, N.D., Chen, J., Stuart, R.K., Fan, J.B. and Ren, B. (2008) Genome-wide mapping of allele-specific protein–DNA interactions in human cells. *Nat. Methods*, **5**, 307–309.

25. Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.

26. Degner, J.F., Marioni, J.C., Pai, A.A., Pickrell, J.K., Nkadori, E., Gilad, Y. and Pritchard, J.K. (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, **25**, 3207–3212.

27. Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., Bosco, D. *et al.* (2010) A map of open chromatin in human pancreatic islets. *Nat. Genet.*, **42**, 255–259.

28. Golding, J., Pembrey, M. and Jones, R. and ALSPAC Study Team (2001) ALSPAC—the avon longitudinal study of parents and children. I. Study methodology. *Paediatr. Perinat. Epidemiol.*, **15**, 74–87.

29. Adli, M., Zhu, J. and Bernstein, B.E. (2010) Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. *Nat. Methods*, **7**, 615–618.

30. Valdar, W., Solberg, L.C., Gauguier, D., Burnett, S., Klenerman, P., Cookson, W.O., Taylor, M.S., Rawlins, J.N., Mott, R. and Flint, J. (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.*, **38**, 879–887.

31. Ayroles, J.F., Carbone, M.A., Stone, E.A., Jordan, K.W., Lyman, R.F., Magwire, M.M., Rollmann, S.M., Duncan, L.H., Lawrence, F., Anholt, R.R. *et al.* (2009) Systems genetics of complex traits in *Drosophila melanogaster*. *Nat. Genet.*, **41**, 299–307.