Review

Characterizing natural variation using next-generation sequencing technologies

Yoav Gilad[1], Jonathan K. Pritchard[1,2] and Kevin Thornton[3]

[1] Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA
[2] Howard Hughes Medical Institute, University of Chicago, Chicago, IL 60637, USA
[3] Department of Ecology and Evolutionary Biology, University of California Irvine, Irvine, CA 92697, USA

Progress in evolutionary genomics is tightly coupled with the development of new technologies to collect high-throughput data. The availability of next-generation sequencing technologies has the potential to revolutionize genomic research and enable us to focus on a large number of outstanding questions that previously could not be addressed effectively. Indeed, we are now able to study genetic variation on a genome-wide scale, characterize gene regulatory processes at unprecedented resolution, and soon, we expect that individual laboratories might be able to rapidly sequence new genomes. However, at present, the analysis of next-generation sequencing data is challenging, in particular because most sequencing platforms provide short reads, which are difficult to align and assemble. In addition, only little is known about sources of variation that are associated with next-generation sequencing study designs. A better understanding of the sources of error and bias in sequencing data is essential, especially in the context of studies of variation at dynamic quantitative traits.

The study of natural variation at multiple levels

In the post-genome era, one of the central goals of evolutionary biologists is to understand the evolutionary histories of genetic and regulatory mechanisms that underlie organismal diversity. Using genomic technologies, such as high-throughput Sanger sequencing and DNA microarrays, comparative studies of natural variation at the molecular level have yielded important insights into population histories, as well as into the genetic mechanisms underlying adaptation and speciation. For example, studies of the patterns of nucleotide diversity (namely, studies of variation at the genomic level) have been used to infer the nature of selective forces acting on specific loci [1], as well as to perform genome-wide scans for the signatures of natural selection [2,3]. Similarly, studies of regulatory variation within and between species were used to characterize the selective pressures that shape gene regulatory processes [4–6].

Indeed, over the past decade, genomic technologies provided first glimpses into the extent of natural variability at the molecular level and the evolutionary forces that shape this variation. Moreover, in several cases, patterns of variation at the molecular level (either structural or regulatory) were used to explain variation in ultimate physiological and morphological phenotypes (reviewed in Refs [7,8]). Although the insights revealed by these studies of natural variation are considerable, until recently, these studies were strictly limited in scope by available technology.

For example, at the genomic level, most studies of natural variation focused on populations of species for which a sequenced reference genome was available, because it was nearly inconceivable for single laboratories to sequence entire genomes. Moreover, owing to the cost of traditional sequencing methods, most large-scale studies of genetic diversity (especially in species with large genomes) chose to genotype previously identified polymorphisms rather than discover new mutations by direct sequencing. The genotype approach typically relies on the assumption that for many questions it is sufficient to genotype only a subset of variants (e.g. because un-typed variation is in linkage disequilibrium with typed variation [9]). However, even in cases where this assumption is mostly valid, these studies inevitably result in a low-resolution description of genetic variation (e.g. in the case of mapping the genetic basis for a trait, such as susceptibility to a disease, genotyping studies usually result in the identification of a genomic region associated with the trait, not a specific functional variant). By contrast, with the development of next-generation sequencing technologies (reviewed in Ref [10]), complete surveys of all genetic variation in a large number of individuals have become feasible. Moreover, individual laboratories or small consortia will soon be able to sequence entire genomes, effectively allowing any species to be developed as a 'model system'.

Next-generation sequencing technologies also enable one to have a more complete picture of regulatory variation than was previously possible. Indeed, most studies of regulatory variation to date have used microarrays, which rely on hybridization to specific probes. Thus, microarray-based studies can survey variation only in genomic regions that are probed by the array (and while genome-wide tiling arrays are available for a few species, these are extremely expensive for species with large genomes). Moreover, for

Corresponding authors: Gilad, Y. (gilad@uchicago.edu); Thornton, K. (krthornt@uci.edu).

0168-9525/$ – see front matter doi:10.1016/j.tig.2009.09.003 Available online 2 October 2009

Box 1. Applications of next-generation sequencing

Beyond the direct sequencing of genomes and transcripts, applications that we discuss in detail in the main text, evolutionary biologists have begun utilizing next-generation sequencing to learn more about variation at regulatory mechanisms by using empirical protocols such as ChIPseq, sequencing of DNA hypersensitive sites, and specific assays designed to explore variation in epigenetic modifications. We describe these applications briefly.

- Chromatin immunoprecipitation followed by sequencing (ChIP-seq) aims to identify interactions between proteins and DNA [67]. By cross-linking proteins and DNA before precipitation using specific antibodies, one can enrich the sample for chromatin bound by specific transcription factors. The enriched chromatin is then visualized by direct sequencing to identify, at high resolution, the genomic positions to which the transcription factor was bound. The number of reads that map to any particular location is a rough measure of the abundance of sequences from that location among the ChIP-bound fragments. Emerging studies of variation in transcription factor binding to specific promoters within population and between species can reveal mechanisms underlying ultimate variation at transcript levels.
- DNA hypersensitive sites are genomic loci that are not bound by proteins and are therefore exposed to digestion by nuclease enzymes. By sequencing the products of genomic DNA digestion with DNase I, one can characterize the genomic regions bound by proteins such as transcription factors and histones, and study the positions of nucleosomes [68]. Studies of variation in the location of DNA hypersensitive sites between individuals and across species are now emerging and might provide additional insight into the mechanism underlying expression quantitative trait loci.
- Finally, next-generation sequencing facilitates genome-wide studies of epigenetic modifications [69]. In particular, bisulfite sequencing to identify methylated CpG sites can now be performed in high throughput and in many individuals simultaneously. Uncovering natural variation in epigenetic modifications will add another layer to our understanding of the mechanisms underlying morphological and physiological diversity, as well as mechanisms by which genes and environments interact to activate particular regulatory programs.

multi-species studies of regulatory variation, specific dedicated microarray platforms had to be developed for each combination of species [11]. By contrast, using next-generation sequencing, one can directly study variation at all transcripts, as well as variation in regulatory interactions (such as transcription factor binding) within and between species (Box 1).

Thus, next-generation sequencing approaches have the potential to allow one to work on any species, collect genome-wide natural variation data at unprecedented resolution, and provide considerable additional insight into the mechanisms of regulatory evolution. However, these technologies are still being developed, and the different properties that affect inference made using sequence data are still being actively explored. For example, there is uncertainty about the precise relationship between sequence coverage and the false positive and negative rates associated with detecting mutations. In addition, we know very little about the effects of base composition on quantitative inference of gene expression levels using sequence data. In what follows, we discuss what recent studies have taught us about such properties of next-generation sequencing data and we provide several recommendations regarding study designs.

Sequencing entire genomes

Next-generation sequencing platforms provide unprecedented sequencing capacity (Figure 1). The most immediate impact of these technologies on evolutionary studies is the ability to sequence entire genomes. Currently, however, most individual-genome sequencing studies were performed in traditional model systems, whereas the sequencing of new species has largely occurred in microorganisms, probably because the task of assembling a new genome is easier in microorganisms, which have generally smaller and less-repetitive genomes. Indeed, the short length of current sequence reads and, in particular, the short inset size of the sequenced segments [12], make it extremely difficult to assemble large genomes with many repetitive elements de novo. Future developments in sequencing technologies, such as larger inset sizes (which may soon be available by single-molecule sequencing platforms), or a much greater read length (which may soon be

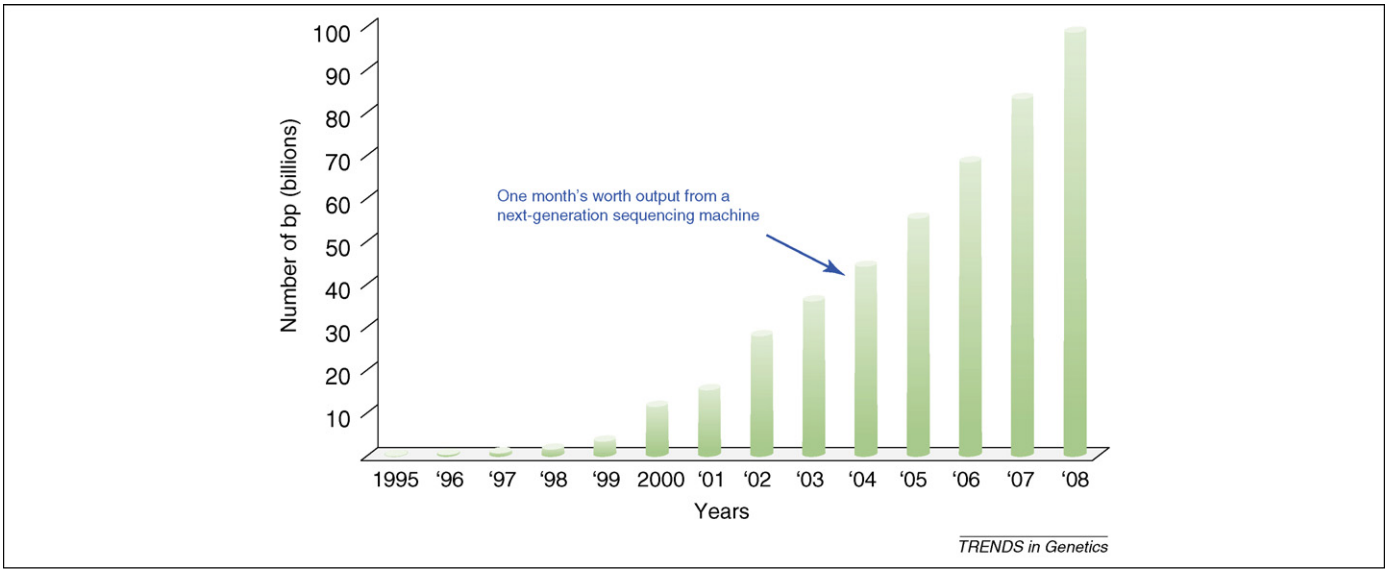

Figure 1. A cumulative plot of the amount of DNA sequence (in bp) deposited in GenBank since 1995. The data were taken from http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html (data downloaded on 10 August 2008; data for 2009 was not available).

available both by 454 and Illumina/Solexa platforms) should facilitate de novo assembly of complex genomes.

To date, whole-genome sequencing studies of individuals from a single species have been published for the model plant *Arabidopsis thaliana* [13], the worm *Caenorhabditis elegans* [14], the fruit fly *Drosophila melanogaster* [15], the yeasts *Pichia stipitis* [16] and *Saccharomyces cerevisiae* [17] and in humans [18–23].

Results from these relatively few individual-genome sequencing studies highlight the important impact that next-generation sequencing technologies are expected to have on studies of variation at the genomic level. For example, the study of yeast by Smith *et al*. [16] illustrated how the new technology allows one to interrogate results from traditional experiments at unprecedented depth. Indeed, it is now possible to identify every mutation that arises during experimental evolution, and to compare accurately the entire genomes of ancestral and evolved strains. Thus, one can fully sequence frozen and ancestral strains for model systems in which experimental evolution in the laboratory has been progressing for many years, thereby facilitating comparative genomic analyses that probably could not have been foreseen when some of these experiments were originally designed. Similarly, in the context of mutation accumulation experiments [24], next-generation sequencing allows one to identify every substitution in the mutation accumulation lines. These studies provide invaluable insights into the patterns of spontaneously-arising mutations and enable evolutionary geneticists to estimate mutation rates and the rates associated with the nucleotide substitution matrix (two fundamental parameters that are used to model both genomic and regulatory evolution [25–27]).

In addition to comparisons of genomes within a species, next-generation sequencing provides opportunities for new discoveries using comparative genomic approaches. For example, by comparing fully sequenced yeast genomes, Doniger *et al.* identified novel introgressed regions between *S. cerevisiae* and *Saccharomyces paradoxus* [17]. In turn, Kulathinal *et al.* studied patterns of divergence in fruit flies by using sequence data from the 454 platform to obtain partial assemblies of two subspecies of *Drosophila pseudoobscura* and *Drosophila miranda* [28]. They assembled the sequence reads by using the available *D. pseudoobscura* reference genome, and demonstrated that divergence between the Drosophila species is elevated near inversions that are fixed between species. This observation suggests a role for inversions in maintaining species integrity in the face of ongoing hybridization. Thus, even though there are relatively few comparative genomic analyses published, there is clearly a great potential for evolutionary insights into the processes of genome evolution and speciation.

What have we learned from fully sequenced individual genomes?

The genome-wide sequencing studies to date have included, at most, a few individuals from each species. While the sample sizes are still too small for detailed population genetic analysis, these studies have led to some insights into the abundance of deleterious mutations in individual genomes. For example, the published human genomes suggest that each individual carries mutations known to underlie both Mendelian and complex disorders [18–21]. Indeed, by analysing individual human genomic sequences, Chun and Fay have estimated that each individual human carries ~800 deleterious mutations [29]. Similarly, a whole-genome sequencing study of two *S. cerevisiae* strains suggests an abundance of segregating deleterious genetic variation, including amino acid polymorphisms at sites that are otherwise conserved between species, large insertion/deletion mutations, and polymorphic premature stop codons [17]. Beyond these insights, recent genome-wide sequencing studies revealed that accurate calls of polymorphic sites are possible with moderate sequence coverage (Box 2) and suggest that rare polymorphic sites will be uncovered shortly in large samples.

Indeed, in species such as *A. thaliana*, maize, and *D. melanogaster*, where polymorphic sites are frequent (roughly once every 100 bp), most data on genetic variation have been collected using direct Sanger sequencing and thus include rare variants. However, the situation has been quite different in species with lower genetic diversity and large genomes (such as humans). The low rate of polymorphism and the large genome size made it far more efficient to type known polymorphic sites in large panels of individuals rather than perform direct sequencing. This approach results in a bias towards intermediate frequency alleles [30], a property that has the potential to obscure the signature of recent selective pressures or demographic fluctuations [31,32].

In contrast, next-generation sequencing of whole genomes should provide a nearly unbiased picture of genetic variation, including the identification of rare variants. The characterization of rare polymorphic sites will be particularly useful in studies of natural selection and in studies aimed at inferring the demographic history of a population (for a recent review of these topics, see Ref. [33]). In this context, the 1000 Genomes Project (http://www.1000genomes.org) seeks to identify all single-nucleotide polymorphisms in humans, with a minor allele frequency of 1% or higher. The identification of rare variants should also improve our ability to identify loci contributing to variation in quantitative traits and complex disease susceptibility.

RNA sequencing facilitates high-resolution transcript profiling

Over the past decade, genome-wide comparative studies of gene expression levels have provided insight into gene regulatory processes in many different contexts, including studies of gene regulation during development, altered gene regulatory patterns associated with disease or pharmacological responses, and of the evolution of gene regulation [34–37]. Most of these studies, however, estimated gene expression levels using microarrays, which rely on hybridization to specific probes. The probes are typically designed to complement only a small proportion of each gene; therefore, estimates of overall gene expression levels based on microarray data may often be inaccurate.

Microarray platforms that allow one to estimate the relative expression levels of individual exons have been

Box 2. Aligning short reads and detecting mutations

Because most next-generation sequencing platforms produce short (30–100 bp) reads, the assembly of genomes is the first challenge in analysing such data. To date, most studies used assembly-by-reference, where short sequence reads are mapped back onto a published genomic sequence. The challenge in these cases is to quickly (computational time becomes an important issue to consider when dealing with billions of short reads) and accurately map reads onto the reference sequence. Several aligner algorithms to do this were published recently [70,71], and others were made available online (e.g. Mosaik, http://bioinformatics.bc.edu/marthlab/Mosaik). These aligners typically preprocess the reference genomic sequence to speed mapping, and then attempt to identify a unique map position for each read, allowing for a certain number of mismatches. The aligners differ in the way they take into account mismatches or close secondary matches in the reference genome. Some algorithms, such as Mosaik and SHORE [13], allow for small insertions or deletions with respect to the reference genomic sequence. In addition, a growing number of aligners now support paired-end data, and several, including Mosaik, can often resolve cases where one mate of paired-end reads maps uniquely, and the other does not. To our knowledge, there is no consensus at the moment on which might be the 'best' aligner.

Once the short sequence reads are aligned and assembled, the next challenge is to identify mutations or polymorphisms between sequences. In this respect, the questions on everyone's mind are (i) which platform provides more accurate results? and (ii) what is the tradeoff between overall sequence coverage and false negative and positive rates associated with discovering new polymorphisms? To address such issues, Smith *et al.* compared the performance of three next-generation sequencing platforms (Illumina/Solexa, Roche/454 and ABI SoLiD) in detecting mutations that arose during selection of a yeast strain (*Pichia stipitis*) for increased ethanol production [16]. Following data collection, the Illumina and 454 reads were aligned to a reference genome using Mosaik, and the SoLiD data were aligned using the software provided by ABI. With respect to detecting mutations, Smith *et al.* found that inference based on ABI/SoLiD data produced the lowest false positive rate (in fact, the false positive rate in the SoLiD data of Smith *et al.* was zero regardless of overall coverage). In turn, inference based on the 454 data produced the highest false positive rate. All three technologies were able to identify all the single-nucleotide mutations given at least 10–15-fold nominal sequence coverage. It should be noted, however, that Smith *et al.* sequenced a haploid species and therefore side-stepped the challenge of detecting heterozygote mutations [16].

Using human sequences, Bentley *et al.* [18] and Wang *et al.* [19] compared genotype calls from the resequencing of individual human samples using next-generation sequencing platforms to genotype calls from Affymetrix 500K chips for the same individuals. These studies found that higher coverage was required to accurately call heterozygous genotypes (~30 ×) than was necessary for homozygous genotypes (~15 ×).

developed (e.g. Affymetrix exon arrays), and custom microarrays that include probes for a subset of exon–exon boundaries are also in use [38,39]. However, it is impractical to include on a microarray probe-sequences for all possible transcripts (namely, probes that complement all alternative exon definitions and all possible combinations of exon junctions). Therefore, microarrays are ultimately not well suited for the characterization of differences in alternative splicing, alternative transcription start or end sites, or overall gene structure differences between samples (e.g. in a comparison of gene expression levels across species).

The recent developments in sequencing technology have made it possible to use sequence-based approaches for expression profiling (using RNAseq protocols [40–42]). These new approaches do not rely on specific predesigned probes, and thus can provide a more detailed picture of gene regulatory variation compared with microarray data (Figure 2). In particular, RNAseq data enable one to easily explore transcription of genomic regions that are not functionally annotated, estimate exon and gene expression levels, as well as study differences in exon usage. Moreover, by mapping sequence reads that span exon–exon junctions, RNAseq data can be used to characterize exon-skipping events; namely, cases of alternative splicing. Indeed, recent RNAseq studies found numerous transcripts from genomic regions that were not previously known to be transcribed, and revealed a much higher diversity of alternative splice variants than previously recognized [43–45].

That said, the analysis of RNAseq data to identify entire transcripts is still rather limited as currently all published methods to infer transcript expression levels rely on a prior description of existing transcripts (e.g. by using EST databases). However, we expect that the true diversity of alternative spliced transcripts far exceeds the current collection (which is based mainly on relatively low-throughput sequencing of cDNA products). Thus, there is a need for the development of a statistical method that will enable us to infer the structure of whole transcripts and their relative expression levels, directly from RNA sequencing data.

RNAseq study design and analyses

Beyond the ability to characterize transcription levels regardless of previous annotation, another advantage of RNA sequencing over expression profiling using microarrays is that, in principle, it should be easier to compare results across studies. The absolute intensity values from microarray hybridizations are difficult to interpret beyond the context of the original comparative study, because absolute intensity values are affected by a large number of confounding physical and environmental variables [46]. By contrast, RNAseq-derived estimates of absolute gene expression levels, based on the number of mapped sequence reads, are expected to be more easily interpretable. In other words, while microarrays can typically be used only to estimate relative expression levels (i.e. to compare expression levels across samples), it is expected that RNAseq can provide more reliable estimates of absolute expression levels, which in turn can be compared easily across studies. In practice, however, this may not be the case.

While the determinants of biases in gene expression estimates based on next-generation sequencing data are not fully understood, it is becoming clear that base composition has a major role in such biases. In this sense, RNAseq data are rather similar to microarray data, where estimates of absolute expression levels are biased by the effects of base composition on the performance of probes [47]. Moreover, we have found that differences in the library preparation protocols and differences in concentration between RNA samples will bias gene expression estimates based on RNAseq data [42]. These are not

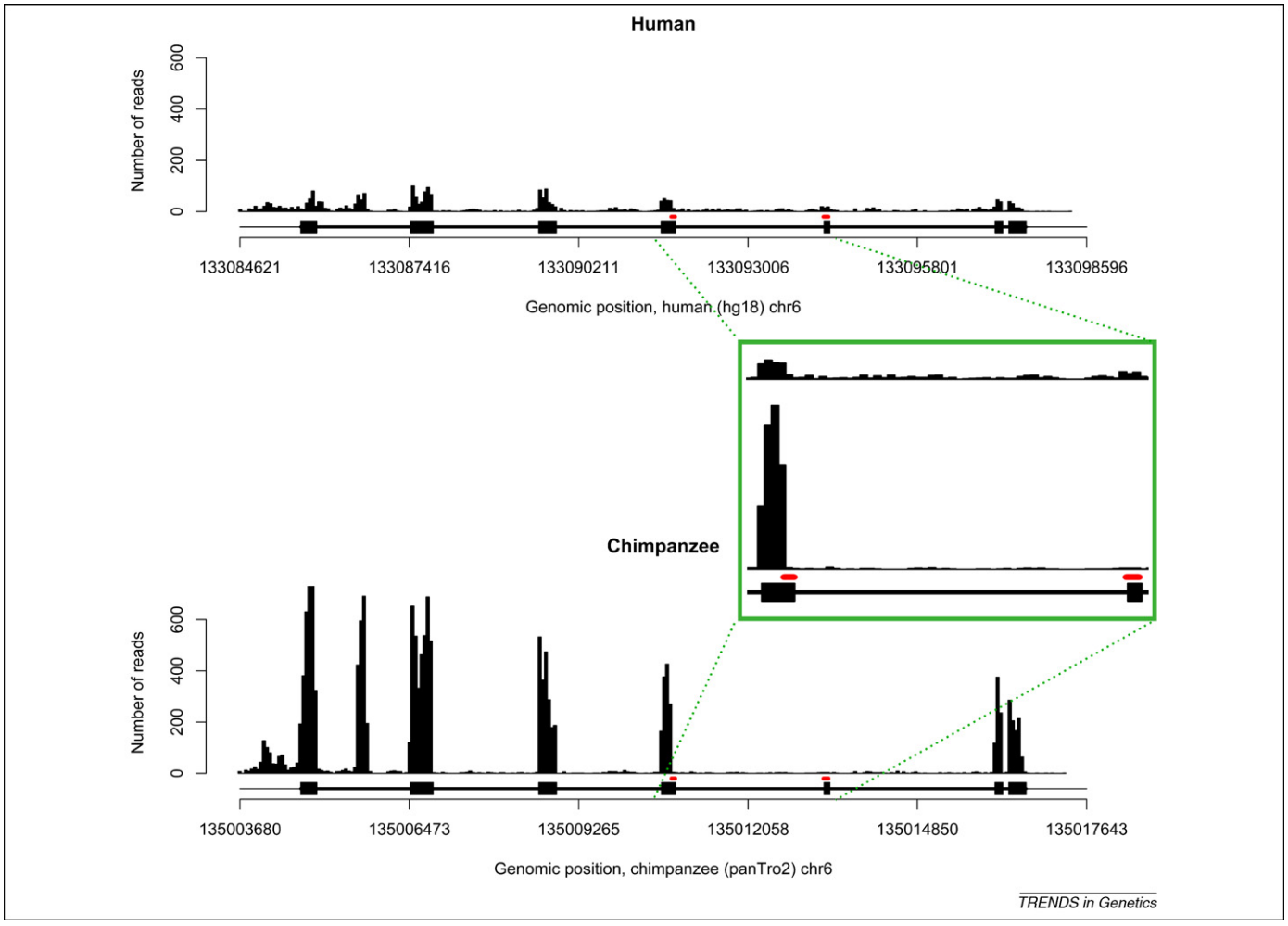

Figure 2. RNAseq data from human and chimpanzee liver samples are plotted along the Vanin-family protein 3 (*VNN3*) gene region. As can be seen, the *VNN3* gene is highly expressed in chimpanzee compared to human. However, using data collected with a multi-species microarray [4], we previously inferred that the *VNN3* gene is highly expressed in humans compared with chimpanzee. The reason for the discrepancy between the two datasets is that the microarray probes were designed to complement a limited exonic region of the gene, which, in contrast to the entire gene, seems to be highly expressed in humans (the red lines represent the positions of the array probes for this gene – see expanded section).

intuitive effects, because differences in concentration and library preparation can be seen as systematic differences across samples, which in principle could be expected to be well adjusted for by normalization techniques (such as those typically applied to microarray data [48]). However, in the case of RNAseq data, genomic segments with different base composition are affected to different extents by the quality of library preparation and sample concentration. Hence, a simple global normalization cannot adjust for these effects. RNA sequencing study designs should take these effects into account (Box 3) but, in any case, owing to these technical effects, it is unclear how valid direct comparison of results across multiple RNAseq studies might be.

Another important aspect of RNAseq data is that the number of sequence reads that map to a particular mRNA tends to be roughly proportional to the mRNA concentration multiplied by the mRNA length. Thus, long genes tend to be represented by more sequence reads than short genes expressed at the same level. As a result, estimates of expression levels (normalized by gene length) tend to be less variable for long exons or long genes, than for shorter exons or genes. For example, Oshlack and Wakefield showed recently that the ability to identify differentially expressed genes between samples is strongly associated with the length of the transcript [49]. Moreover, when overall sequence coverage is increased, the corresponding increase in the power to detect differences in expression levels across samples is also associated with gene length [49]. Microarray data are not susceptible to this complex interaction between gene length and the power to detect differences in expression levels, because all probes on the array are typically of the same length.

Importantly, the association between gene length and the power to detect gene expression differences using RNAseq can bias many downstream analyses [49]. For example, ranking or testing for functional enrichments among differentially expressed genes can result in the spurious identification of pathways or functional annotations that include mainly longer genes.

Using next-generation sequencing to measure allele-specific effects

Another advantage of gene expression profiling by RNA sequencing compared to array-based approaches is that, regardless of the study design, it is possible to test directly

Box 3. On next-generation sequencing study designs

Most of the issues pertaining to sequencing study designs are not specific to next-generation sequencing technologies, but rather common to all large-scale genomic studies. Nevertheless, when a new and exciting technology becomes available and widely used, it seems appropriate to review briefly a few principles of effective study designs.

The most important aspect of an effective design is to ensure that the study will produce data that are suitable to address the questions that are being asked. For example, if the goal of the study is to compare alternative splice variants across human tissues, the design must allow the researcher to separate tissue-specific effects from other possible confounding effects, such as individual-specific effects. For example, studies that compare inter-tissue gene expression profiles by sampling each tissue from a different individual might spuriously interpret inter-individual differences in overall gene expression levels or in alternative splicing as regulatory differences between tissues.

Similarly, effective study designs should allow the researcher to interpret the effects of interest in the correct context of possible confounding sources of variation. For example, if the goal of the study is to estimate the overlap in binding locations of a specific transcription factor across individuals, the design must allow the researcher to independently estimate variation across replicate experiments within an individual. Indeed, studies that estimate inter-individual overlap in transcription factor binding sites using one replicate per individual might spuriously interpret poor technical replicability as low overlap across individuals.

Finally, effective study designs must allow the researcher to independently estimate salient sources of variance, or be balanced with respect to effects that could bias the results. For example, sequencing RNA libraries on different sequencing platforms can result in biased estimates of individual gene expression levels. Studies that do not balance their design with respect to this effect, or include replicates that will allow one to independently estimate the 'platform effect', might spuriously interpret bias due to the association between samples and the platform on which the samples were sequenced as true biological differences in gene expression levels.

for allele-specific effects in heterozygotes. For example, a heterozygous polymorphic locus in the exon of a gene can be used to test whether the two copies of a diploid gene are expressed at different levels (e.g. due to genetic imprinting [50,51], or functional differences in *cis*-acting regulatory alleles that affect gene expression levels [52,53]). Similarly, if a heterozygous polymorphic locus lies within a putative transcription-factor binding site, then one can test whether the alternative alleles cause differential binding using a ChIPseq experiment (Box 1). These types of tests are very attractive because the two alternative alleles should be exposed to exactly the same set of *trans*-acting factors. Performing such investigations using microarray technologies is much more difficult because sequence differences between the alleles at the location of the probe can affect hybridization kinetics. Thus, using microarrays, differences in hybridization properties have to be taken into account before estimates of differential abundance between the alleles can be obtained.

Although the study of allele-specific expression patterns will undoubtedly be very valuable, we have recently found that read-mapping biases can lead to spurious estimates of allele-specific expression levels [72]. The basic problem was illustrated by an RNAseq study of allele-specific expression levels using ~32 million sequence reads of 35 bp each from a library of human lymphoblastoid cell line mRNA. When these reads were mapped to a standard reference human genome sequence, 52% of all reads that overlapped known heterozygous polymorphisms matched the reference allele (namely, the allele shown in the reference human genome). However, among genes that showed significant difference in allelic expression levels, a much higher proportion (72%) of reads that overlapped known heterozygous polymorphisms matched the reference allele.

Using simulations of the sequencing and read-mapping process we found that this skew towards reference alleles can be explained by mapping biases. Specifically, we found that a fraction of reads that include the alternative (non-reference) allele align better to an alternative genomic location. As a result, the estimated expression level of the alternative allele at the original genomic location is artificially low. Aligning reads to a masked reference genome with respect to all known polymorphisms removed the bias towards the reference allele, but did not solve the principle problem because one of the alleles can still match better to an alternative position in the genome. In other words, while we did not observe a bias towards the reference allele once we used a masked genome, the number of falsely identified differences in allelic expression levels did not decrease substantially.

Our current preferred method for dealing with this problem is to identify (and exclude) reads that overlap polymorphic sites with an inherent read-mapping bias. In our data, we thus excluded 40% of the cases that originally showed differences in expression levels between alleles. Our simulation-based method is helpful for identifying genomic regions in which mapping problems might affect allele-specific expression level estimates; however, there is clearly scope for statistical approaches to this problem that will be more robust and yield better power. Additionally, we expect that longer reads and paired-end reads will help to ameliorate these mapping biases.

Studying the mechanisms of regulatory variation

It has long been thought that evolution acts on 'two levels', on protein-coding sequences and on sites that impact gene regulation [54,55]. In the past few years, there have been numerous genome-wide studies of the evolutionary forces that shape genetic diversity at protein-coding sequences [2,56]. A common goal has been to identify genes that show signatures of the action of natural selection on particular lineages.

However, studies of the evolution of gene regulation have generally lagged behind, largely because, unlike protein-coding sequences, the annotation of gene regulatory sequences remains highly incomplete. We still have limited information about which locations in the genome have regulatory roles, and what their functional roles are. Thus, for the most part, it has been challenging to study the evolution of regulatory elements at a genomic level (but see Refs [57–59]).

Fortunately, a variety of experimental methods have been developed recently for measuring aspects of gene regulation on a genome-wide scale. These methods include measurement of chromatin accessibility, histone modifications, nucleosome positioning, methylation patterns,

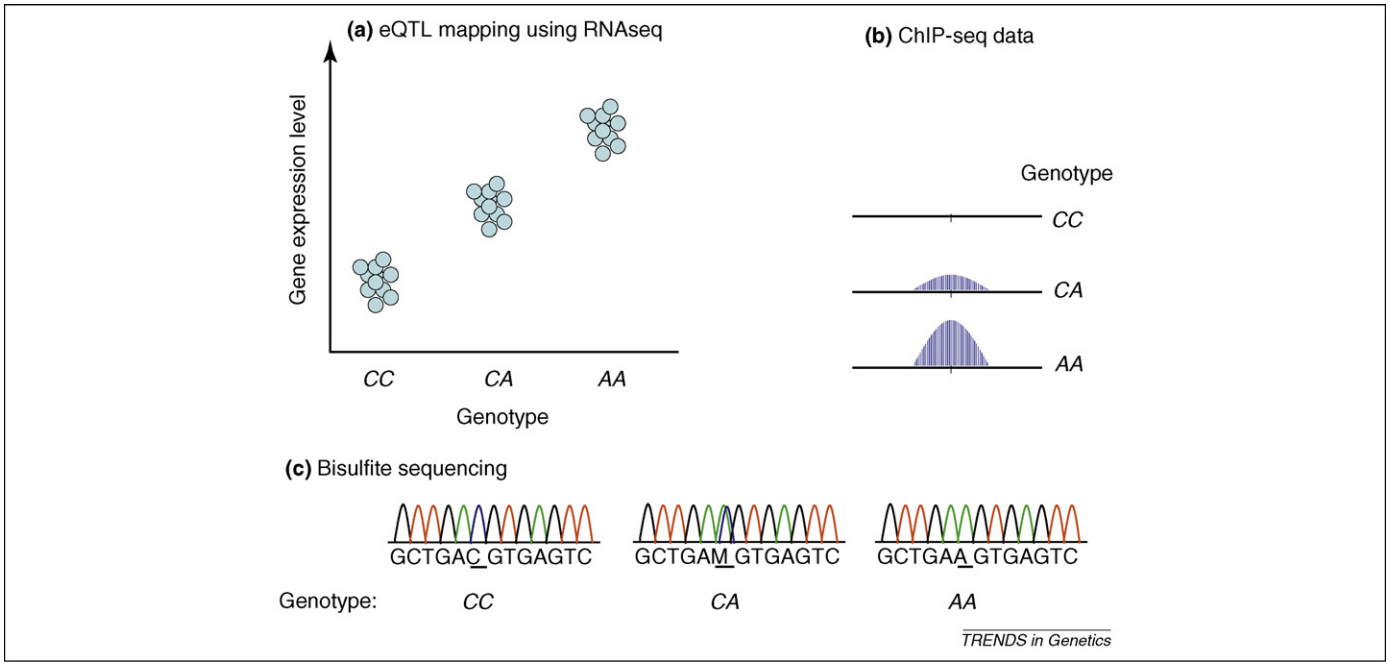

Figure 3. A heuristic example of the inference that can be made from a combination of different sources of genomic data. In this example, we make the assumption that the resolution of all three approaches is sufficient to identify a specific causative nucleotide. In reality, at the moment at least, the resolution of these techniques is sufficient to identify only candidate small (few kb) genomic regions. **(a)** Using RNAseq and genotyping data, an expression quantitative trait locus (eQTL) is identified. **(b)** ChIPseq data indicate that the same eQTL underlies differential binding of a transcription factor to the locus, thereby suggesting a regulatory mechanism that can explain the observed association between genotype and gene expression level. **(c)** Bisulfite sequencing suggests that the mechanism that underlies difference in transcription factor binding across genotypes, and ultimately differences in gene expression levels, is DNA methylation.

and directed assays to determine the binding locations of DNA- or RNA-binding proteins. For most of these techniques, next-generation sequencing has become a critical tool for obtaining high-resolution, genome-wide read-outs of the results.

For example, chromatin immunoprecipitation (ChIP) assays use antibodies to extract fractions of the genome that are bound by a particular transcription factor, or that are attached to histones with a particular modification. Until recently, ChIP studies used microarrays to measure the abundance of chromatin from any given location in the genome. These measurements can now be achieved with far greater sensitivity and specificity using next-generation sequencing [60]. Importantly for evolutionary studies, sequencing approaches also facilitate studies across different species, whereas traditional approaches using microarrays are more limited due to the effect of sequence mismatches on hybridization kinetics [61]).

By using ChIPseq as well as associated approaches to characterize regulatory interactions (Box 1), we hope to eventually develop a much richer understanding of gene regulatory processes. Indeed, projects are underway to catalogue the regulatory sequences in humans, flies, and worms on genome-wide scales. These projects will provide a map of regulatory sites in each genome, and should lead to a deeper understanding of regulatory mechanisms.

From an evolutionary perspective, the real power of these new approaches will come as we move beyond the initial projects to survey regulatory variation within and between species (e.g. by assessing the extent to which transcription factor binding sites are shared across species [62–64]). As our understanding of gene regulation improves, we can start asking more detailed mechanistic questions. For example, we know that within humans there is an enrichment of high-F_{ST} polymorphisms at non-synonymous sites [65], implying that some of these sites have been targets of positive selection in certain populations. At present, it is difficult to ask similar questions for regulatory elements, such as transcription factor binding sites, simply because these are currently poorly annotated. Hopefully, in the near future, we will have a much more complete picture of the types of sites that are targets of selection within species, as well as the phenotypic consequences of selection on those sites. Similarly, we can ask these questions regarding variation across species. For example, what are the functional consequences of turnover in transcription factor binding sites in different lineages? Indeed, by combining information from multiple sources we can characterize specific regulatory mechanisms, which might enable the identification of the molecular basis for adaptations of complex physiological phenotypes (Figure 3). Combination of different types of genomic data, such as hypersensitive sites and ChIPseq (Box 1), can also enable high-throughput characterization of mechanistic variation in the usage of regulatory elements.

Concluding remarks

Next-generation sequencing platforms allow us to survey multiple levels of natural variation at unprecedented resolution and depth. Although most sequencing studies have focused on the major model systems for which high-quality reference genomes are available, the future holds great promise for non-model systems. As sequencing costs decrease, and both laboratory and computational protocols improve, it is likely that the barrier to entry into genomics

research will be lowered for many systems. Currently, the technology is in place to sequence cDNA libraries at a fraction of the cost compared with just a few years ago, and the sequencing capacity available to individual laboratories might already be sufficient to sequence complete mammalian genomes at moderate coverage (e.g. tenfold). Cancer laboratories, for example, already utilize these technologies to characterize all the mutations in tumor genomes [66].

For systems with genomes close to 100 MB, the $1000 genome is already possible. What is lacking are efficient ways to assemble the sequence and a standard set of validated analysis tools. At this time, both the sequencing technologies and the tools to analyse the data are evolving quickly. New single-molecule sequencing platforms, which offer the ability to sequence minute amounts of nucleic acids without amplification, should improve our ability to obtain quantitative measures. Additional applications that utilize the vast amounts of available sequencing capacity are also evolving rapidly. Consequently, we believe that the full realization of the effects of next-generation sequencing technologies on the study of natural variation is still to come.

Acknowledgments

We thank T. Long, C. Ober, Z. Gauhar and J. Marioni for helpful discussion and/or comments on the manuscript. Y.G. is supported by the Sloan foundation and NIH grants HL092206 and GM077959; K.T. is supported by the Sloan Foundation and by NIH grant GM085183; J.K.P. is supported by the HHMI and by NIH grant MH084703.

References

1 Kelley, J.L. and Swanson, W.J. (2008) Positive selection in the human genome: from genome scans to biological significance. *Annu. Rev. Genomics Hum. Genet.* 9, 143–160
2 Clark, A.G. *et al.* (2003) Positive selection in the human genome inferred from human-chimp-mouse orthologous gene alignments. *Cold Spring Harb. Symp. Quant. Biol.* 68, 471–477
3 Voight, B.F. *et al.* (2006) A map of recent positive selection in the human genome. *PLoS Biol.* 4, e72
4 Blekhman, R. *et al.* (2008) Gene regulation in primates evolves under tissue-specific selection pressures. *PLoS Genet.* 4, e1000271
5 Khaitovich, P. *et al.* (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* 309, 1850–1854
6 Oleksiak, M.F. *et al.* (2002) Variation in gene expression within and among natural populations. *Nat. Genet.* 32, 261–266
7 Carroll, S.B. (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134, 25–36
8 Hoekstra, H.E. and Coyne, J.A. (2007) The locus of evolution: evo devo and the genetics of adaptation. *Evolution Int. J. Org. Evolution* 61, 995–1016
9 McCarthy, M.I. *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9, 356–369
10 Mardis, E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141
11 Gilad, Y. and Borevitz, J. (2006) Using DNA microarrays to study natural variation. *Curr. Opin. Genet. Dev.* 16, 553–558
12 Chaisson, M.J. *et al.* (2009) De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Res.* 19, 336–346
13 Ossowski, S. *et al.* (2008) Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res.* 18, 2024–2033
14 Hillier, L.W. *et al.* (2009) Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. *Genome Res.* 19, 657–666
15 Daines, B. *et al.* (2009) High-throughput Multiplex Sequencing to Discover Copy Number Variants in Drosophila. *Genetics* 182, 935–941
16 Smith, D.R. *et al.* (2008) Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res.* 18, 1638–1642
17 Doniger, S.W. *et al.* (2008) A catalog of neutral and deleterious polymorphism in yeast. *PLoS Genet.* 4, e1000183
18 Bentley, D.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59
19 Wang, J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature* 456, 60–65
20 Ahn, S.M. *et al.* (2009) The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. *Genome Res.* 19, 1622–1629
21 Wheeler, D.A. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452, 872–876
22 McKernan, K.J. *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* 19, 1527–1541
23 Pushkarev, D. *et al.* (2009) Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.* 27, 847–850
24 Keightley, P.D. *et al.* (2009) Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Res.* 19, 1195–1201
25 Gilad, Y. *et al.* (2006) Natural selection on gene expression. *Trends Genet.* 22, 456–461
26 Denver, D.R. *et al.* (2005) The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. *Nat. Genet.* 37, 544–548
27 Rifkin, S.A. *et al.* (2005) A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. *Nature* 438, 220–223
28 Kulathinal, R.J. *et al.* (2009) The genomics of speciation in Drosophila: diversity, divergence, and introgression estimated using low-coverage genome sequencing. *PLoS Genet.* 5, e1000550
29 Chun, S. and Fay, J.C. (2009) Identification of deleterious mutations within three human genomes. *Genome Res.* 19, 1553–1561
30 Marth, G.T. *et al.* (2004) The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics* 166, 351–372
31 Clark, A.G. *et al.* (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* 15, 1496–1502
32 Nielsen, R. (2004) Population genetic analysis of ascertained SNP data. *Hum. Genomics* 1, 218–224
33 Thornton, K.R. *et al.* (2007) Progress and prospects in mapping recent selection in the genome. *Heredity* 98, 340–348
34 White, K.P. (2001) Functional genomics and the study of development, variation and evolution. *Nat. Rev. Genet.* 2, 528–537
35 Whitehead, A. and Crawford, D.L. (2006) Neutral and adaptive variation in gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 103, 5425–5430
36 Scherf, U. *et al.* (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* 24, 236–244
37 Wray, G.A. (2007) The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* 8, 206–216
38 Calarco, J.A. *et al.* (2007) Global analysis of alternative splicing differences between humans and chimpanzees. *Genes Dev.* 21, 2963–2975
39 Stolc, V. *et al.* (2004) A gene expression map for the euchromatic genome of Drosophila melanogaster. *Science* 306, 655–660
40 Wang, Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63
41 Mortazavi, A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628
42 Marioni, J.C. *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18, 1509–1517
43 Wilhelm, B.T. *et al.* (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453, 1239–1243
44 Nagalakshmi, U. *et al.* (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320, 1344–1349
45 Sultan, M. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321, 956–960
46 Draghici, S. *et al.* (2006) Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.* 22, 101–109

47 Oshlack, A. *et al.* (2007) Using DNA microarrays to study gene expression in closely related species. *Bioinformatics* 23, 1235–1242
48 Bolstad, B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193
49 Oshlack, A. and Wakefield, M.J. (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* 4, 14
50 Babak, T. *et al.* (2008) Global survey of genomic imprinting by transcriptome sequencing. *Curr. Biol.* 18, 1735–1741
51 Wang, X. *et al.* (2008) Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain. *PLoS One* 3, e3839
52 Serre, D. *et al.* (2008) Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genet.* 4, e1000006
53 Wittkopp, P.J. *et al.* (2008) Independent effects of cis- and trans-regulatory variation on gene expression in Drosophila melanogaster. *Genetics* 178, 1831–1835
54 Britten, R.J. and Davidson, E.H. (1971) Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *Q. Rev. Biol.* 46, 111–138
55 King, M.C. and Wilson, A.C. (1975) Evolution at two levels in humans and chimpanzees. *Science* 188, 107–116
56 Kosiol, C. *et al.* (2008) Patterns of positive selection in six Mammalian genomes. *PLoS Genet.* 4, e1000144
57 Pollard, K.S. *et al.* (2006) An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443, 167–172
58 Prabhakar, S. *et al.* (2008) Human-specific gain of function in a developmental enhancer. *Science* 321, 1346–1350
59 Jeong, S. *et al.* (2008) The evolution of gene regulation underlies a morphological difference between two Drosophila sister species. *Cell* 132, 783–793
60 Jothi, R. *et al.* (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.* 36, 5221–5231
61 Gilad, Y. *et al.* (2005) Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Res.* 15, 674–680
62 Wilson, M.D. *et al.* (2008) Species-specific transcription in mice carrying human chromosome 21. *Science* 322, 434–438
63 Odom, D.T. *et al.* (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat. Genet.* 39, 730–732
64 Moses, A.M. *et al.* (2006) Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput. Biol.* 2, e130
65 Barreiro, L.B. *et al.* (2008) Natural selection has driven population differentiation in modern humans. *Nat. Genet.* 40, 340–345
66 Ley, T.J. *et al.* (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 456, 66–72
67 Farnham, P.J. (2009) Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.* 10, 605–616
68 Boyle, A.P. *et al.* (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132, 311–322
69 Varley, K.E. *et al.* (2009) Intra-tumor heterogeneity of MLH1 promoter methylation revealed by deep single molecule bisulfite sequencing. *Nucleic Acids Res.* 37, 4603–4612
70 Li, H. *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18, 1851–1858
71 Li, R. *et al.* (2008) SOAP: short oligonucleotide alignment program. *Bioinformatics* 24, 713–714
72 Degner, J.F., *et al.* (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* in press.