Sequencing depth and coverage: key considerations in genomic analyses

David Sims, Ian Sudbery, Nicholas E. Ilott, Andreas Heger and Chris P. Ponting

Abstract | Sequencing technologies have placed a wide range of genomic analyses within the capabilities of many laboratories. However, sequencing costs often set limits to the amount of sequences that can be generated and, consequently, the biological outcomes that can be achieved from an experimental design. In this Review, we discuss the issue of sequencing depth in the design of next-generation sequencing experiments. We review current guidelines and precedents on the issue of coverage, as well as their underlying considerations, for four major study designs, which include *de novo* genome sequencing, genome resequencing, transcriptome sequencing and genomic location analyses (for example, chromatin immunoprecipitation followed by sequencing (ChIP–seq) and chromosome conformation capture (3C)).

Depth
The average number of times that a particular nucleotide is represented in a collection of random raw sequences.

Computational Genomics Analysis and Training Programme, Medical Research Council Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, Le Gros Clark Building, University of Oxford, Parks Road, Oxford OX1 3PT, UK.
Correspondence to D.S. and C.P.P.
e-mails: david.sims@dpag.ox.ac.uk; chris.ponting@dpag.ox.ac.uk
doi:10.1038/nrg3642

Genomics is extending its reach into diverse fields of biomedical research from agriculture to clinical diagnostics. Despite sharp falls in recent years[1], sequencing costs remain substantial and vary for different types of experiment. Consequently, in all of these fields investigators are seeking experimental designs that generate robust scientific findings for the lowest sequencing cost. Higher coverage of sequencing (BOX 1) inevitably requires higher costs. The theoretical or expected coverage is the average number of times that each nucleotide is expected to be sequenced given a certain number of reads of a given length and the assumption that reads are randomly distributed across an idealized genome[2]. Actual empirical per-base coverage represents the exact number of times that a base in the reference is covered by a high-quality aligned read from a given sequencing experiment. Redundancy of coverage is also called the depth or the depth of coverage. In next-generation sequencing studies coverage is often quoted as average raw or aligned read depth, which denotes the expected coverage on the basis of the number and the length of high-quality reads before or after alignment to the reference. Although the terms depth and coverage can be used interchangeably (as they are in this Review), coverage has also been used to denote the breadth of coverage of a target genome, which is defined as the percentage of target bases that are sequenced a given number of times. For example, a genome sequencing study may sequence a genome to 30× average depth and achieve a 95% breadth of coverage of the reference genome at a minimum depth of ten reads.

An ideal genome sequencing method would faultlessly read all nucleotides just once, doing so sequentially from one end of a chromosome to the other. Such a perfect approach would ensure that all polymorphic alleles within diploid or polyploid genomes could be identified, and that long identical or near-identical repetitive regions could be unambiguously placed in a genome assembly. In real-world sequencing approaches, read lengths are short (that is, ≤250 nucleotides) and can contain sequence errors. When considered alone, an error is indistinguishable from a sequence variant. This problem can be overcome by increasing the number of sequencing reads: even if reads contain a 1% variant-error rate, the combination of eight identical reads that cover the location of the variant will produce a strongly supported variant call with an associated error rate of 10^{-16} (REF. 3). Increased depth of coverage therefore 'rescues' inadequacies in sequencing methods (BOX 1). Nevertheless, generating greater depth of short reads does not cure all sequencing ills. In particular, it alone cannot resolve assembly gaps that are caused by repetitive regions with lengths that either approach or exceed those of the reads. Instead, in the paired-end read approach, paired reads — two ends of the same DNA molecule that are sequenced and which are separated by a known distance — are used to unambiguously place repetitive regions that are smaller than this distance.

Sequencing is enriching our understanding not only of genome sequence but also of genome organization, genetic variation, differential gene expression and

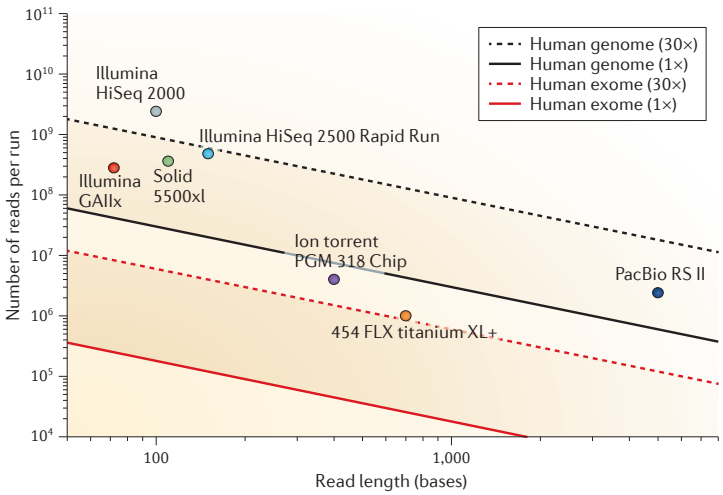
© 2014 Macmillan Publishers Limited. All rights reserved

Personalis EX2022.121

diverse aspects of transcriptional regulation, which range from transcription factor-binding sites to the three-dimensional conformation of chromosomes. As these areas of genome research often adopt markedly different sequencing depths (FIG. 1), we review this issue for each area in turn. First, we examine current best practice in *de novo* genome sequencing and assembly. We then proceed to consider genome resequencing and targeted resequencing approaches, particularly whole-exome sequencing (WES). Second, we discuss the rapidly evolving area of transcriptome sequencing, specifically the different considerations that are needed for transcript discovery compared with the analyses of differential expression and alternative splicing. Finally, we explore a range of methodologies that identify the genomic sites of transcription factor binding, chromatin marks, DNA methylation and spatial interactions that are revealed by chromosome conformation capture (3C) methods. We discuss experimental considerations that are relevant to sequence depth, which are required for the generation of high-quality, unbiased and interpretable data from next-generation sequencing studies.

Box 1 | **Sequencing coverage theory**

Much of the original work on sequencing coverage stemmed from early genome mapping efforts. In 1988, Lander and Waterman[96] described the theoretical redundancy of coverage (c) as *LN/G*, where *L* is the read length, *N* is the number of reads and *G* is the haploid genome length. The figure shows the theoretical coverage (shown as diagonal lines; c = 1× or 30×) according to the Lander–Waterman formula for human genome or exome sequencing. The coverage that is achieved by sequencing technologies according to the manufacturers' websites is also indicated (see the figure). Unfortunately, biases in sample preparation, sequencing, and genomic alignment and assembly can result in regions of the genome that lack coverage (that is, gaps) and in regions with much higher coverage than theoretically expected. GC-rich regions, such as CpG islands, are particularly prone to low depth of coverage partly because these regions remain annealed during amplification[97]. Consequently, it is important to assess the uniformity of coverage, and thus data quality, by calculating the variance in sequencing depth across the genome[98]. The term depth may also be used to describe how much of the complexity in a sequencing library has been sampled. All sequencing libraries contain finite pools of distinct DNA fragments. In a sequencing experiment only some of these fragments are sampled. The number of these distinct fragments sequenced is positively correlated with the depth of the true biological variation that has been sampled.



GAIIx, Genome Analyzer IIx; PacBio, Pacific Biosciences; PGM, personal genome machine.

De novo genome sequencing

The major factors that determine the required depth in a *de novo* genome sequencing study are the error rate of the sequencing method, the assembly algorithms used, the repeat complexity of the particular genome under study and the read length. Genomes that have been sequenced to high depths by short-read technologies are not necessarily a substantial improvement in assembly quality compared with those produced using the earlier lower-coverage Sanger sequencing technology. Although the human genome was initially assembled to high quality with 8–10-fold coverage using long-read Sanger sequencing[2], a raw coverage of ~73-fold was required to generate the first short-read-only assembly of the giant panda genome that was of lower quality than the human genome[4]. A similarly low coverage (~7.5-fold) dog genome, which is similar in size to that of the giant panda and was assembled using Sanger sequencing reads, is more complete and more contiguous than the giant panda genome[3]. These differences arise because Sanger sequencing reads are longer, are derived from larger insert libraries and can be assembled using mature assembly algorithms[3].

High-quality assemblies are now often produced using hybrid approaches, in which the advantages of high-depth, short-read sequencing are complemented with those of lower-depth but longer-read sequencing. For example, sequencing the draft assembly of the wild grass *Aegilops tauschii* was a considerable challenge owing to its large size (4.4 Gb) and to the fact that two-thirds of its sequence consists of highly repetitive transposable element-derived regions[5]. The draft genome was successfully assembled first into short fragments (that is, contigs) using 398 Gb (that is, a 90-fold coverage) of high-quality short reads from 45 libraries with insert sizes between 0.2 kb and 20 kb, and these fragments could then be linked into longer scaffolds using paired-end read information. Gaps between contigs predominantly contained repetitive sequence, the unique placement of which posed difficulties. These gaps were filled in using a subsequent addition of 18.4 Gb (that is, a fourfold coverage) of Roche 454 long reads. A recently introduced approach to sequencing repeat-rich genomes is to barcode and sequence to an average of 20× depth all reads that are derived from each of many collections of hundreds or thousands of short (6–8 kb) DNA fragments[6]. By assembling each collection separately, many otherwise confounding repetitive sequences of the *Botryllus schlosseri* tunicate genome were resolved. By applying approaches that are complementary in aspects such as read lengths and coverage biases, hybrid library and assembly methods are likely to dominate in the near future[7,8].

Twofold coverage and lower-quality assemblies have been produced using Sanger sequencing for a selection of mammalian genomes to identify sequences that are conserved in eutherian species, including humans[9]. The Lander–Waterman approach (BOX 1) predicts that ~86%

© 2014 Macmillan Publishers Limited. All rights reserved

Personalis EX2022.122

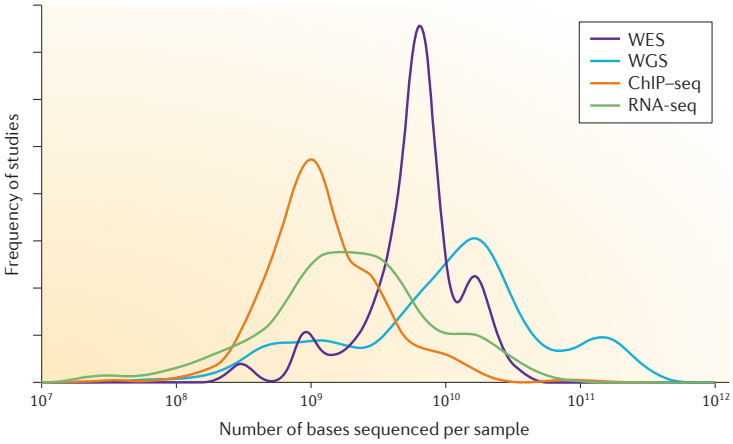


Figure 1 | **Sequencing depths for different applications.** The frequency of studies that use read counts of all runs (which are typically flow-cell lanes) and that were deposited from 2012 to June 2013 for the Illumina platform in the European Nucleotide Archive (ENA) is shown. The plot provides an overview of sequencing depths that are usually chosen for the four most common experimental strategies. Densities have been smoothed and normalized to provide an area under the curve that is equal to one. The depth and therefore the cost of an experiment increase in the order of chromatin immunoprecipitation followed by sequencing (ChIP–seq), RNA sequencing (RNA-seq), whole-exome sequencing (WES) to whole-genome sequencing (WGS). Although ChIP–seq, WES and WGS have typical applications and thus standardized read depths, the sequencing depth of RNA-seq data sets varies over several orders of magnitude. Multimodal distributions of WES and WGS reflect different target coverage. To generate this figure, runs were summed by experiment and, for each study, one experiment was chosen at random to avoid counting large studies more than once. Note that the ENA archive only contains published data sets and excludes medically relevant data sets. The plot was created from 771 studies.

Sequence capture
The enrichment of fragmented DNA or RNA species of interest by hybridization to a set of sequence-specific DNA or RNA oligonucleotides.

GC bias
The difference between the observed GC content of sequenced reads and the expected GC content based on the reference sequence.

Variant calling
The process of identifying consistent differences between the sequenced reads and the reference genome; these differences include single base substitutions, small insertions and deletions, and larger copy number variants.

(that is, $1 - e^{-2}$) of bases in such genomes are covered once by a sequencing depth of 2× although, in reality, this decreases to ~65% for mammalian genomes that are sequenced at twofold coverage[10]. In these and other studies, low coverage has two principal effects on subsequent analyses and biological interpretation. First, it is not possible to resolve whether an absence of a protein-coding gene, or a disruption of its open reading frame, represents a deficiency of the assembly or a real evolutionary gene loss. Second, and perhaps more seriously, low depth can introduce sequence errors that are in danger of being mistakenly propagated through downstream analyses and misdirecting conclusions of a study. To mitigate this possibility, two approaches are recommended. First, low-quality bases or sequences that align poorly against a closely related genome should be discarded from such analyses. Second, adjacent bases that have high-quality scores should also be discarded because they can contain a high density of residual sequence errors[11].

DNA resequencing

DNA resequencing explores genetic variation in individuals, families and populations, particularly with respect to human genetic disease. Requirements for sequencing depth in these studies are governed by the variant type of interest, the disease model and the size of the regions of interest. Resequencing can reveal single-nucleotide variants (SNVs), small insertions and deletions (indels), larger structural variants (such as inversions and translocations) and copy number variants (CNVs). Naturally, the design of a particular study depends on the biological hypothesis in question, and different sequencing strategies are used for population studies compared with those for studies of Mendelian disease or of somatic mutations in cancer. Furthermore, targeted resequencing approaches allow a trade-off between sequencing breadth and sample numbers: for the same cost, more samples can be sequenced to the same depth but over a smaller genomic region. Here, we discuss the merits of whole-genome sequencing (WGS) relative to targeted resequencing approaches, including WES, in the context of these different variant types and disease models.

WGS versus WES. High-depth WGS is the 'gold standard' for DNA resequencing because it can interrogate all variant types (including SNVs, indels, structural variants and CNVs) in both the minority (1.2%) of the human genome that encodes proteins and the remaining majority of non-coding sequences. WES is focused on the detection of SNVs and indels in protein-coding genes and on other functional elements such as microRNA sequences; consequently, it omits regulatory regions such as promoters and enhancers. Although costs vary depending on the sequence capture solution, WES can be an order of magnitude less expensive than WGS to achieve an approximately equivalent breadth of coverage of protein-coding exons. These reduced costs offer the potential to greatly increase sample numbers, which is a key factor for many studies. However, WES has various limitations that are discussed below.

SNV and indel detection. Early genome resequencing studies focused specifically on the two most common classes of sequence variation, which are SNVs and small indels. The first human genome that was sequenced using Illumina short-read technology showed that, although almost all homozygous SNVs are detected at a 15× average depth, an average depth of 33× is required to detect the same proportion of heterozygous SNVs[12]. Consequently, an average depth that exceeds 30× rapidly became the de facto standard[13,14]. In 2011, one study[15] suggested that an average mapped depth of 50× would be required to allow reliable calling of SNVs and small indels across 95% of the genome. However, improvements in sequencing chemistry reduced GC bias and thus yielded a more uniform coverage of the genome, which later reduced the required average mapped depth to 35× (REF. 15). The power to detect variants is reduced by low base quality and by non-uniformity of coverage. Increasing sequencing depth can both improve these issues and reduce the false-discovery rate for variant calling. Although read quality is mostly governed by sequencing technology, the uniformity of depth of coverage can also be affected by sample preparation. A GC bias that is introduced during DNA amplification by PCR has been identified as a major source of variation in coverage. Elimination of PCR amplification results in

© 2014 Macmillan Publishers Limited. All rights reserved

Personalis EX2022.123

Box 2 | **Genomic alignment and mappability**

The first major data processing step in sequencing studies for species with a reference genome is the alignment of sequencing reads to this reference. The choice of alignment algorithm often influences final coverage values, as different algorithms show varying false-positive and false-negative rates[99,100]. Even the best mapping algorithms cannot align all reads to the reference genome, which is perhaps due to sequencing errors, structural rearrangements or insertions in the query genome, or deletions in the reference. Indeed, analyses of unmapped reads are often used for the identification of structural variants and non-reference insertions[40,101]. Furthermore, it is not possible to unambiguously assign reads to all genomic regions, as some regions will contain low-degeneracy repeats or low-complexity sequences. The 'mappability' (also known as uniqueness) of a sequence within a genome has a major influence on the average mapped depth and is an important source of false-negative single-nucleotide variant calls[102]. Mappability improves with increased read length and generally shows an inverse correlation with genomic repeats[103]. One approach to increase coverage in regions of low mappability is to use longer reads that improve the chance of a read encompassing a unique sequence that anchors all remaining sequences. A second approach is to generate paired-end libraries with longer insert sizes, which increases the chance of one read of the pair mapping to a unique region outside the repeat sequence. It is often useful to use mappability data to normalize read depth, for example, when using depth of coverage to estimate DNA copy number.

improved coverage of high GC regions of the genome and in fewer duplicate reads[16].

In WES, differences in the hybridization efficiency of sequence capture probes, which are possibly again attributable to GC content variation, can result in target regions that have little or no coverage. Uniformity of coverage will also be influenced by repetitive or low-complexity sequences, which either restrict bait design or lead to off-target capture. Furthermore, unlike WGS, WES still routinely uses PCR amplification, which must be carefully optimized to reduce GC bias[17]. As a result of increased variation in coverage, a greater average read depth is required to achieve the same breadth of coverage as that from WGS, and an 80× average depth is required to cover 89.6–96.8% of target bases, depending on the platform, by at least tenfold[18]. Different sequence capture kits yield different coverage profiles, and designs with higher density seem to be more efficient, which provide better uniformity of coverage and better sensitivity for SNV detection[18,19]. As capture kits have improved sequence coverage, the amount of sequencing required has inevitably increased. Regardless of the capture protocol or the sequencing platform used, there has been a trend for recent exome studies to require a minimum of 80% of the target region to be covered by at least tenfold[20–22]. All WES kits are prone to reference bias, which arises from capture probes that match the reference sequence and thus tend to preferentially enrich the reference allele at heterozygous sites; such bias can produce false-negative SNV calls[23].

CNV detection. CNVs can be detected from WGS and WES[24,25] data using methods that analyse depth of coverage. These methods pile up aligned reads against genomic coordinates, then calculate read counts in windows to provide the average depth across a region. Copy number changes can then be inferred from variation in average depth across genomic regions. In WGS, reasonable specificity can be obtained with an average depth of as little as 0.1× (REF. 26). However, sensitivity, break-point detection and absolute copy number estimation all improve with increasing read depth[26,27]. Regardless of average read depth, depth-of-coverage methods are vulnerable to false positives that are being called owing to local variations in coverage even after correction for both GC bias and 'mappability' (BOX 2), and cross-sample calling is required to reduce this effect[28].

Study design for different disease models. In contrast to the high depth that is required to accurately call SNVs and indels in individual genomes, population genomics studies benefit from a trade-off between sample numbers and sequencing depth, in which many genomes are sequenced at low depth (for example, 400 samples at 4×) and their variants are called jointly across all samples[29–31]. Variant calls on individual low-depth genomes have a high false-positive rate, but this is mitigated by combining information across samples. This approach provides good power to detect common variants at a proportion of the sequencing cost of deep sequencing[29,30]. Indeed, even ultra-low-coverage sequencing (that is, sequencing at 0.1–0.5×) captures almost as much common variation (that is, variants with >1% allele frequency) as single-nucleotide polymorphism (SNP) arrays[32]. Conversely, reliable identification of variants in either highly aneuploid genomes or heterogeneous cell populations, such as those from tumours, requires greater depth of coverage than those from normal tissue[33]. Targeted enrichment and ultra-deep sequencing (that is, sequencing at 1,000×) of limited regions of interest can be used to study clonal evolution in cancer samples, in which specific variants are present in <1% of the cell population[34]. The identification of disease-causing *de novo* or recessive variants is often best served by sequencing parent–child trios. In this case, it is recommended that the same depth of sequencing is obtained for each of the family members in order to minimize false-positive calls in the proband and false-negative calls in the parents[35].

Analyses of DNA resequencing data. A typical analysis pipeline for DNA resequencing data involves the alignment of sequencing reads to a reference genome followed by variant calling. A post-alignment step to remove all but one duplicates (that is, the removal of two or more read pairs with both forward and reverse reads that map to identical genomic coordinates) is important for accurate variant calling, as it ensures that errors that are introduced and amplified during PCR do not result in erroneous calls[36]. Duplicate read removal can significantly reduce the number of high-quality mapped reads and thus the average depth of coverage (TABLE 1). Even in species with a complete reference genome, assembly approaches (reviewed and compared in REFS 37–39) offer several advantages over those using reference alignment. First, assembly can faithfully recapitulate divergent sequence, such as that of the human leukocyte antigen (*HLA*) locus, which often does not align well to a reference genome. Second, assembly

Low-complexity sequences
DNA regions that have a biased nucleotide composition, which are enriched with simple sequence repeats.

Clonal evolution
An iterative process of clonal expansion, genetic diversification and clonal selection that is thought to drive the evolution of cancers, which gives rise to metastasis and resistance to therapy.

© 2014 Macmillan Publishers Limited. All rights reserved

Personalis EX2022.124

Table 1 | **Sources of uninformative reads for different experiments**

Source of uninformative reads	WGS	WES	ChIP–seq	RNA-seq
Sequencing adaptor reads	✓	✓	✓	✓
Low-quality reads	✓	✓	✓	✓
Unmapped reads	✓	✓	✓	✓
Reads that do not map uniquely	✓	✓	✓	✓
PCR duplicates	✓	✓	✓	✓
Reads that map out with peaks, transcript models or exons	–	✓	✓	✓
Reads that map to uninformative transcripts (for example, rRNA)	–	–	–	✓

ChIP–seq, chromatin immunoprecipitation followed by sequencing; RNA-seq, RNA sequencing; rRNA, ribosomal RNA; WES, whole-exome sequencing; WGS, whole-genome sequencing.

Dynamic range
The range of expression levels over which genes and transcripts can be accurately quantified in gene expression analyses. In theory, RNA sequencing offers an infinite dynamic range, whereas microarrays are limited by the range of signal intensities.

Long non-coding RNAs
(lncRNAs). RNA molecules that are transcribed from non-protein-coding loci; such RNAs are >200 nt in length and show no predicted protein-coding capacity.

Cap analysis of gene expression
(CAGE). In contrast to RNA sequencing, CAGE produces short 'tag' sequences that represent the 5′ end of the RNA molecule. As CAGE does not sequence across an entire cDNA, it requires a lower depth of sequencing than RNA sequencing to quantify low-abundance transcripts.

Spike-in control RNAs
A pool of RNA molecules of known length, sequence composition and abundance that is introduced into an experiment to assess the performance of the technique.

Fragments per kilobase of exon per million reads mapped
(FPKM). A method for normalizing read counts over genes or transcripts. Read counts are first normalized by gene length and then by library size. After normalization, the expression value of each gene is less dependent on these variables.

can avoid the mis-mapping of reads that originate from incomplete regions of the reference genome. Third, assembly enables multiple variant types to be analysed at once, which minimizes errors around clusters of variants. The latest assembly methods, such as Cortex[40], can consider multiple eukaryotic genomes simultaneously while incorporating information about known variation. This allows variant calling against a range of different genomes rather than a single reference genome. This method required only an average depth of 16× during the assembly of human *HLA* regions to provide results that are in good agreement with laboratory-based typing[40]. However, as assembly methods are still unable to fully reconstruct entire genomes owing mainly to repeat content, they are only able to call variants in 80% of the genome.

Transcriptome sequencing

RNA sequencing (RNA-seq) allows the detection and the quantification of expressed transcripts in a biological sample. Its applications include novel transcript discovery, and analyses of differential expression and alternative splicing. RNA-seq has advantages over microarray gene expression analyses, as it provides an unbiased assessment of the full range of transcripts with a greater dynamic range[41,42]. Large numbers of RNA-seq experiments have now been carried out in many cell and tissue types across diverse conditions, yet few clear guidelines on read counts have emerged. This is because sequencing requirements are often dependent on the biological question under investigation, as well as on the size and the complexity of the transcriptome being assayed. Here, we describe the concepts that govern the coverage required in RNA-seq experiments and illustrate these with examples from the literature.

Coverage in transcriptome sequencing. Coding and non-coding transcripts can be expressed at vastly different levels — from one copy to millions of copies per cell — in different cell types and developmental stages. Consequently, in any given RNA-seq experiment, coverage varies considerably across transcripts, and read count, read length and the number of biological replicates are more important experimental considerations than transcriptome-wide coverage statistics. Furthermore, when used for differential expression analyses, RNA-seq can be considered as a tag-counting application. In this case, a sufficient number of reads are required to quantify exons and splice junctions in the sample. Therefore, the number of reads that is required in an experiment is determined by the least abundant RNA species of interest — a variable that is not known before sequencing.

The number of useful reads that is generated in a study can be optimized either by depleting the ribosomal RNA (rRNA) fraction, which constitutes ~90% of total RNA in mammalian cells, or by enriching for the RNA species of interest, such as the use of immobilized oligo-deoxythymidine to enrich for polyadenylated RNAs[43]. Total RNA that is depleted in rRNA contains reads from both non-polyadenylated transcripts and pre-processed mRNA transcripts. Consequently, many reads will align to intronic sequences, thereby decreasing the proportion of reads that map to expressed exons and reducing the power to detect splice junctions. A good indication of the performance of an RNA-seq experiment is provided by the proportion of reads that are mapped to rRNA and other highly expressed RNAs, and by the proportion that are mapped to splice junctions and coding exons. Using a poly(A) selection protocol with paired reads of lengths that are >76 bp, >80% of read pairs can be expected to map to the reference genome in experiments using human samples, and >70% of these reads can be expected to map with zero mismatches[44]. With this approach, the number of reads that map to rRNA will be minimal (that is, <1%), and ~15% of reads will map across splice junctions.

Transcript discovery. One application of transcriptome sequencing that is not possible using microarrays is the identification of novel transcripts, such as long non-coding RNAs (lncRNAs) and alternative transcripts of protein-coding genes. Many of these transcripts are expressed at low levels[45,46], and their discovery therefore requires either deep sampling of the transcriptome or mapping of transcription start sites using cap analysis of gene expression (CAGE). The power to detect a transcript depends on its length and abundance in the sequencing library, as well as on its mappability to the reference genome. The sequencing of RNA standards from the External RNA Control Consortium[47] revealed that molecules that are present at frequencies of 0.6–2.5 molecules per 10^7 molecules could not be detected using 12.4 million uniquely mapping 36-bp reads[48]. Furthermore, the accuracy of abundance estimations using spike-in control RNAs in deeply sequenced human data sets (which contain >94 million uniquely mapped 76-bp paired-end reads) showed a clear dependence on both length and GC composition of an RNA molecule[48]. Sampling of transcripts is also affected by library preparation. Sequenced reads that are generated using Illumina protocols show compositional biases at their 5′ ends owing to the nonrandomness of the hexamer primers that are used in cDNA synthesis[49]. This results in nonrandom sampling of the transcriptome and an

© 2014 Macmillan Publishers Limited. All rights reserved

Personalis EX2022.125

uneven coverage across transcripts[49]. The discovery of novel, rare transcripts is therefore dependent on multiple factors, and it is estimated that >200 million paired-end reads are required to detect the full range of transcripts, including all possible isoforms, in human samples[50].

The transcriptional capacity of a genome affects the read depth that is required for profiling. Mammalian genomes contain tens of thousands of genes, many of which consist of multiple isoforms and are transcribed pervasively across intergenic segments[51]. Some vertebrates, single-cell eukaryotes, bacteria and archaea have less complex genomes and thus lower potential transcriptional output. For example, 80% of yeast genes can be detected (that is, with more than four reads mapping at their 3′ ends) with only four million reads, and there is little increase in the number of detected genes as additional data are added[42,52]. A similar result was obtained for log-phase *Escherichia coli* K12 cultures using two million sequenced reads[53].

Differential expression analyses. Differences in gene expression over time or due to either external stimulation or experimental perturbation are often of interest, and these differences can be used to infer the involvement of specific biological pathways and to generate additional hypotheses. In RNA-seq analyses, gene or transcript abundance is frequently expressed as fragments per kilobase of exon per million reads mapped (FPKM), which provides a length and depth normalization to permit comparisons both within and between samples. Current FPKM calculations use the 75th-percentile of the read-count distribution instead of the total number of mapped reads, which can be skewed by highly expressed outliers[54]. This method improves robustness of differential expression calls for genes of low expression when few highly expressed RNAs dominate a sample. The Encyclopedia of DNA elements (ENCODE)[55] consortium have provided data to assess the number of reads that is required to accurately quantify genes across the dynamic range of FPKM values in human cells[44]. By generating 214 million 100-bp paired-end reads from H1 human embryonic stem cells, the consortium was able to carry out a saturation analysis (see Standards, guidelines and best practices for RNA-seq). Using the full data set as the benchmark, they determined that, for genes with more than ten FPKM, the abundance of 80% of genes could be accurately quantified, within 10% of the full data set, using ~36 million mapped reads[56,57]. However, genes that are expressed at low levels (that is, those with fewer than ten FPKM) could only be accurately quantified with ~80 million mapped reads. If the research question requires the accurate quantification of genes across the entire abundance range — including, for example, those encoding lncRNAs — then either samples should be sequenced at high depth (that is, >80 million reads per sample) or RNA-capture techniques[58] should be used to enrich for low-abundance transcripts. However, if the expectation is that the expression of abundant transcripts (that is, those with more than ten FPKM) changes across conditions, then 36 million reads per sample may be sufficient. Given a fixed budget, reducing the amount of sequence per sample allows the inclusion of more biological replicates per condition. Although this results in a decrease in technical precision at the level of individual RNAs, it allows more accurate estimates of biological variability and contributes to a more robust analysis of differential expression. It is noteworthy that, although few biological replicates ($n < 5$) are often used for differential expression analyses, the trade-off between the depth at which each sample is sequenced and the number of biological replicates must be carefully considered. It is clear that parametric methods — for example, DESeq[59], EdgeR[60] and CuffDiff[61] — that are used to assess differential expression rely on their ability to accurately model biological variability. This is evidenced by the observation that increasing sequencing depth of few replicates (that is, one replicate per condition across two conditions) results in an increase in the number of false-positive differential expression calls. These false positives have been attributed to either short genes that are expressed at low levels or genes with small fold changes[50]. Sequencing deeper means that transcripts that are expressed at low abundance will be detected, but their relevance in a biological context can only be assessed when biological variation can be accurately modelled through replication. Methods for calling differential expression are an active area of research, particularly complex models that attempt to resolve transcription at the level of the transcript rather than the gene[61]. A lack of existing benchmarking data sets means that it is not clear what read depth and what level of replication will be sufficient to carry out such analyses. One solution is a staged sequencing approach using a multiplexed library of all samples and replicates followed by its sequencing in stages until all transcripts of interest have been sufficiently covered and can be accurately quantified (BOX 3).

Analyses of alternative splicing. Most metazoan genes express numerous alternative transcripts (that is, isoforms) that are proposed to contribute to the complex development, organization and function of different tissues[62]. RNA-seq experiments can incorporate information from reads that span exon junctions to infer the presence of alternative isoforms. Two early alternative-splicing studies[63,64] used between 3.5 million and 4.4 million 27-bp reads and between 12 million and 29 million 32-bp reads per sample. Despite being shallow by today's standards, these depths of sequence have allowed the following conclusions to be drawn: the majority of human genes are alternatively spliced[64]; exon skipping is the major class of alternative splicing[63]; and exon usage varies substantially depending on tissue type or cell type[63,64]. A more recent study used ~30 million 80-bp single-end reads to assess differential exon use between embryonic and adult brain tissue in mice[65]. By identifying the exons with expression levels that are higher than expected relative to the overall gene expression level, the study was able to identify alternative splicing events for genes that are involved in actin cytoskeleton regulation in adults and in neuronal signal

Saturation
In the context of sequence depth, the point at which the addition of extra reads to an analysis yields no improvement in the number of significant effects identified.

Parametric methods
Methods that rely on assumptions regarding the distribution of sampled data. In RNA sequencing, differential expression analysis sampled reads are assumed to follow a Poisson or negative binomial distribution.

CLIP–seq
(Crosslinking immunoprecipitation followed by sequencing). A method for interrogating RNA–protein interactions, in which RNAs are crosslinked to proteins by ultraviolet radiation and then fragmented. After immunoprecipitation of the protein of interest, the RNA is converted to cDNA and sequenced.

iCLIP
(Individual nucleotide-resolution crosslinking and immunoprecipitation). An extension of CLIP–seq that produces base-pair resolution. It relies on the fact that most cDNA synthesis reactions terminate at the crosslinked bases of the RNA; these prematurely terminated bases are purified and sequenced.

PAR–CLIP
(Photoactivatable-ribonucleoside-enhanced crosslinking immunoprecipitation). An extension of CLIP–seq, in which the photoactivatable nucleotide uridine analogue 4SU is incorporated into RNA. Upon activation with ultraviolet radiation, these bases form covalent crosslinks with bound proteins. Following conversion to cDNA, uncrosslinked uridines become thymidines, whereas crosslinked uridines become cytosines, thus indicating the protein-binding sites in the RNA.

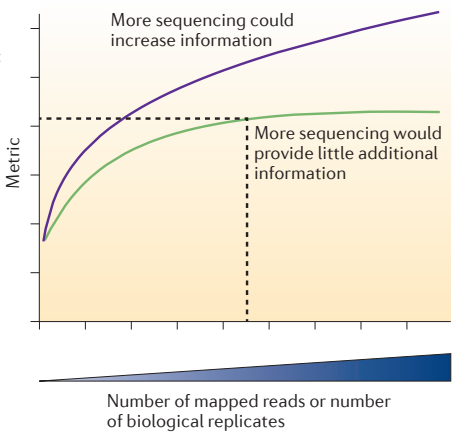
© 2014 Macmillan Publishers Limited. All rights reserved

Personalis EX2022.126

Box 3 | **Staged sequencing for predicting sequencing requirements**

Possible metrics:

- General transcriptome coverage: percentage of genes covered over 90% at a given expression level
- Differential expression: number of differentially expressed genes
- Alternative isoform detection: percentage of split reads (that is, junction that spans reads)
- ChIP–seq peak detection: number of enriched loci



Upon commencing any next-generation sequencing experiment it is difficult to predict the level at which samples should be sequenced. For example, the detection of lowly expressed transcripts and rare splice events in RNA sequencing requires very deep sequencing. Regardless of the specific interest of the experiment, it is prudent to predict the amount of sequence that is required both to answer the biological question and to prevent excessive sequencing. An initial round of sequencing of all experimental samples can be achieved through multiplexing libraries on a single lane: by adding unique DNA tags to each library, sequence reads for individual samples can be extracted after sequencing. Depending on the total number of samples in the experiment, multiple lanes each containing all libraries can be sequenced. Multiplexing each sample on a single lane removes any biases that are associated with inter-lane or inter-run variability, thus permitting data supplementation. These data can then be used to assess the sequencing requirement for the study by sub-sampling various proportions of the full data set and by carrying out saturation analyses. Experiment-specific metrics can aid in study design (see the figure). For example, if the interest is in identifying differentially expressed genes between two conditions, then it would be useful to assess the number of differentially expressed genes that are identified as a function of sequencing depth. Nevertheless, if only few biological replicates are included in the analysis, then there are likely to be false-positive differential expression calls. The number of replicates should be carefully considered in the design phase of the experiment — without appropriate replication the curve may not reach saturation until all genes are called as differentially expressed. In a chromatin immunoprecipitation followed by sequencing experiment, the number of peaks that are discovered could be used. The same concept can be applied to replicate number to determine the level of biological replication at which saturation of differentially expressed genes is reached. If these data are insufficient, then additional sequence can be generated and the process repeated until saturation is achieved. Such approaches were recently formalized using capture–recapture statistics to predict saturation of uniquely sequenced reads, enriched peaks or expressed genes from small initial sample reads[104].

transduction in embryos[65]. Nevertheless, even with this increased depth of sequencing, the transition from exon usage analyses to the assembly of complete isoforms at every expressed locus remains a substantial challenge.

Location, location, location: from ChIP–seq to Hi-C

By location-based methods we are referring to experiments that seek to map the sites of interaction between nucleic acids and other molecules. These include sites of DNA–protein interactions (using chromatin immunoprecipitation followed by sequencing (ChIP–seq)[66] and ChIP-exo[67]); RNA–protein interactions (using methods that are based on crosslinking immunoprecipitation (CLIP), including CLIP–seq[68,69], iCLIP[70] and PAR–CLIP[71]); RNA–DNA interactions (using CHART[72] and CHiRP[73]); and DNA–DNA interactions (using 3C-based methods, including circularized chromosome conformation capture (4C), chromosome conformation capture carbon copy (5C), Hi-C and chromatin interaction analysis by paired-end tag sequencing (ChIA–PET))[74,75]. Our discussion of such approaches also includes some methods that are aimed at assaying the state of the DNA, such as those that interrogate the openness of chromatin (for example, DNaseI-seq[76]) without histone precipitation and those that measure DNA methylation (for example, MeDIP–seq[77] and CAP–seq[78]).

In a typical experiment, nucleic acid fragments that are involved in an interaction are isolated and are subjected to high-throughput sequencing. The resulting reads are regarded as tags that can be used to quantify distinct molecules in the sample. In this case, the read length and the error rate only need to be sufficient to distinguish between the different molecules, for example, to unambiguously identify a location in the genome. The number of reads that map to a particular nucleotide is the primary quantity of interest and is used to estimate the abundance of molecules sequenced. Thus, the required sequencing depth depends on the number of true genomic locations. In the case of ChIP–seq experiments for transcription factor binding, such depth is often unknown at the outset, although it may be known, for example, when comparing methylation profiles between cell types.

Although the number of reads that is necessary to complete a reasonably detailed ChIP–seq experiment has been examined, similarly detailed studies are currently lacking for all other techniques. Here, we first examine the read counts that are necessary for a successful ChIP–seq experiment. We then discuss general considerations that influence the number of read counts that are required when using other techniques.

Identifying DNA–protein interactions using ChIP–seq. The original ChIP–seq study sequenced only 2–5 million reads per sample, and yet nearly all sites across the genome with a strong match to the canonical binding motif of RE1-silencing transcription factor (REST, which is the protein of interest) were found among the 1,946 peaks that were identified[66]. Subsequent studies found that, in general, by sequencing more reads a greater number of binding sites are identified[79–81]. An important factor that influences the read count that is required for a ChIP–seq experiment is whether the protein (or chromatin modification) is a point-source factor, a broad-source factor or a mixed-source factor[79] (FIG. 2). Point sources occur at specific locations in the genome. This class includes sequence-specific transcription factors as well as some highly localized chromatin marks, for example, those associated with enhancers and transcription start sites. Broad sources are generally those that cover extended areas of the genome, such as many chromatin marks (for example, histone H3 lysine 9 trimethylation (H3K9me3) marks). Mixed-source factors, such as RNA polymerase II, yield both types of peaks. As expected, broad-source and mixed-source factors

CHART
(Capture hybridization analysis of RNA targets). A method that uses biotinylated oligonucleotides to pull down complementary RNAs (which are generally long non-coding RNAs) and their associated DNA after crosslinking. The resulting DNA is then sequenced to identify sequences that are associated with the RNA.

© 2014 Macmillan Publishers Limited. All rights reserved

Personalis EX2022.127

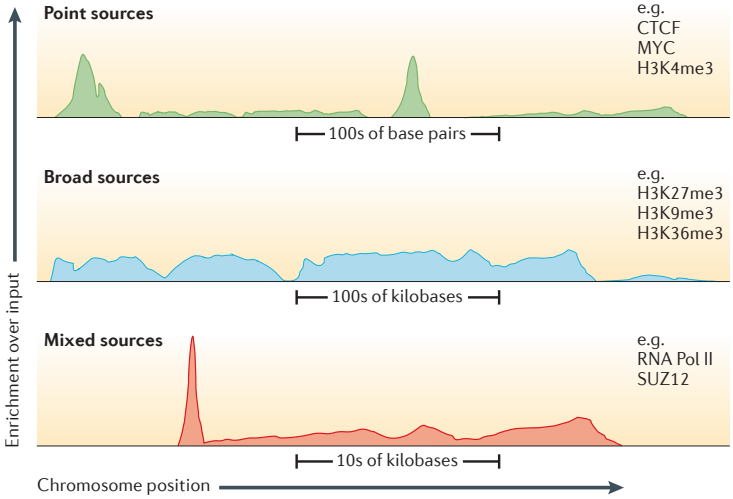


Figure 2 | **The three different types of peaks in chromatin immunoprecipitation followed by sequencing experiments.** Point sources (top panel), such as sequence-specific transcription factors, bind to specific locations in the genome and generate narrow peaks of a few hundred base pairs. Broad sources (middle panel), which include many chromatin marks (such as histone H3 lysine 27 trimethylation (H3K27me3) marks), generate large regions of enriched signal. Mixed-source factors (left panel), notably RNA polymerase II (RNA Pol II), generate enriched regions of a range of sizes. CTCF, transcriptional repressor CTCF; MYC, myc proto-oncogene protein; SUZ12, Polycomb protein SUZ12.

CHiRP
(Chromatin isolation by RNA purification). A method to capture DNA that is associated with RNA (particularly long-non coding RNAs); it is based on a similar principle to CHART.

DNaseI-seq
(DNase I hypersensitive site sequencing). A method to identify regions of open chromatin. Regions of open chromatin are sensitive to DNase I digestion, whereas those in regions of close chromatin are not. Sequencing of fragment ends after DNase I digestion thus reveals the locations of open chromatin.

MeDIP–seq
(Methylated DNA immunoprecipitation followed by sequencing). A method to identify regions of methylated DNA, in which chromatin immunoprecipitation is carried out using an antibody that recognizes methylated cytosine and the resulting immunoprecipitated DNA fragments are subjected to sequencing.

require a greater number of reads than point-source factors[79,81].

The ENCODE project's guidelines for ChIP–seq experiments suggest that point-source factor experiments should use 20 million reads per factor, summed across replicates, in mammals or two million reads per factor in organisms with smaller genomes, such as the fruitfly and the nematode worm[79]. However, at this level most of the factors assayed have not reached saturation in the numbers of peaks identified[79,57], and saturation is not achieved even at 55 million reads, or 100 million reads for some factors, in human cells. In a study of the smaller fruitfly genome, it was found that peak identification for one transcription factor started to show signs of saturation at 16.2 million reads, which is equivalent to ~327 million reads in humans[81], although the numbers of reproducible peaks between multiple replicates started to saturate at 5.4 million reads (and at ~110 million reads in humans).

For broad-source or mixed-source factors it remains unclear what an appropriate number of reads might be; as a guide, the ENCODE consortium used 40 million reads across all replicates[79]. Evaluating the saturation of the number of enriched regions for broad-source factors is complicated because the generation of more reads results in fewer regions as many smaller enriched regions combine[81]. Nonetheless, at 16.2 million reads in fruitflies (which is equivalent to 327 million reads in humans), the number of regions that are enriched in H3K36me3 shows little sign of saturation, although fewer reads are needed to saturate the calling of reproducible peaks[81].

High numbers of reads are required to identify all possible peaks. Peaks that are newly discovered as the number of reads increases tend to show a lower average enrichment relative to the control sample, which suggests that they mark either more weakly bound sites[79,80] or sites where a lower proportion of histones are modified. It should be noted that, although the enrichment of a peak compared with the control sample may provide an indication of binding strength, it is not necessarily a good measure of the probability that the site is biologically functional[82].

The number of reads in each sample must be balanced against other factors when deciding on experimental design. It is important that all ChIP-enriched samples are matched by appropriate control samples. These controls include input DNA that is not enriched, samples that are enriched by ChIP for a non-DNA-binding protein (such as immunoglobulin G) and, in the case of histone modifications, enrichment for unmodified histones. Such control samples should be acquired from the same cell type under the same conditions as the test sample and ideally be processed in parallel[79]. These samples should be sequenced to an equivalent depth to, or an even greater depth than, the ChIP-enriched sample because reads will be distributed across a larger proportion of the genome[79–81,83,84]. Although technical replicates are generally not necessary, it is important to include at least two biological replicates in any experimental design to ensure maximum sensitivity[79,83] but not necessarily accuracy. The Irreproducible Discovery Rate framework provides a means by which to select reproducible peaks across replicates[85] and is more simply applied to two replicates. Paired-end sequencing is preferred over single-end sequencing, as it allows improved identification of duplicated reads and a better estimation of the fragment size distribution, and it also increases the efficiency of mapping to repeat regions[81]. Long reads are not generally thought to be necessary, although they also assist in uniquely mapping reads to repetitive regions.

ChIP–exo extends the ChIP–seq technique by providing base-pair resolution for the binding sites of DNA-binding proteins[86]. In a ChIP–exo experiment, after immunoprecipitation of fragmented chromatin with the protein of interest and ligation of adaptor sequences, a 5′-to-3′ exonuclease is applied. Digestion of the precipitated DNA proceeds until the exonuclease is blocked by the bound protein. The point at which digestion terminates indicates the location of the protein of interest. Published ChIP–exo studies have examined samples in *Saccharomyces cerevisiae* and have used between 200,000 reads (for the sequence-specific Reb1 (REF. 86)) and seven million reads (for a study of general transcription factors[87]) per factor per replicate, which would translate to very high read numbers in a mammalian genome. Nevertheless, one successful experiment for the translational repressor CTCF in human cells used 20–40 million mapped reads per replicate and identified 93% of ~19,000 previously identified binding sites as well as a further ~17,000 locations, 99.5% of which contained a canonical CTCF-binding motif[86]. Currently, ChIP–exo experiments have not included control samples because

© 2014 Macmillan Publishers Limited. All rights reserved

Personalis EX2022.128

Table 2 | **Representative read counts for location-based approaches**

Techniques	Read counts in representative studies	Refs
DNaseI-seq and FAIRE–seq	20–50 million	79
CLIP–seq	7.5 million; 36 million	89, 90
iCLIP and PAR–CLIP	8 million; 14 million	105, 106
CHiRP and CHART	26 million	72
4C	1–2 million	92
ChIA–PET	20 million	107
5C	25 million	108
Hi-C	>100 million	94
MeDIP–seq	60 million	109
CAP–seq	>20 million	110
ChIP–seq	>10 million per sample (point source); >20 million per sample (broad source)	79

4C, circularized chromosome conformation capture; 5C, chromosome conformation capture carbon copy; CAP–seq, CxxC affinity purification sequencing; CHART, capture hybridization analysis of RNA targets; ChIA–PET, chromatin interaction analysis by paired-end tag sequencing; ChIP–seq, chromatin immunoprecipitation followed by sequencing; CHiRP, chromatin isolation by RNA purification; CLIP–seq, crosslinking immunoprecipitation followed by sequencing; DNaseI-seq, DNase I hypersensitive site sequencing; FAIRE–seq, formaldehyde-assisted isolation of regulatory elements followed by sequencing; iCLIP, individual nucleotide-resolution crosslinking and immunoprecipitation; MeDIP–seq, methylated DNA immunoprecipitation followed by sequencing; PAR–CLIP, photoactivatable-ribonucleoside-enhanced crosslinking immunoprecipitation.

CAP–seq
(CxxC affinity purification sequencing). A method to identify genomic regions that are enriched for unmethylated CpG dinucleotides on the basis of binding of the CxxC domain to such regions. A recombinant CxxC domain from the KDM2B protein is biotinylated and is bound to DNA. After fragmentation, DNA bound to the biotinylated CxxC domain is recovered and sequenced.

Peaks
Regions of the genome with an enrichment of mapped reads compared with a control track or a local background. Produced by peak callers, these are often the output of location-based experiments.

Point-source factor
A protein factor that yields narrow and localized peaks in chromatin immunoprecipitation followed by sequencing experiments, such as sequence-specific transcription factors or some modified histones that occur in localized regions.

Broad-source factor
A protein factor or modification that marks extended genomic regions, such as many modified histones.

background levels are assumed to be low, but such experiments have included three or four replicates per sample. This low background level contributes to a high signal-to-noise ratio in ChIP–exo and could partly explain its extra sensitivity.

Other location-based techniques. In recent years, a plethora of techniques for assessing the sites of interactions between a molecule and nucleic acids using high-throughput sequencing have been described[68–72,75,76,88]. These techniques are superficially similar to ChIP–seq in that nucleic acids that interact with the factor of interest are enriched and then sequenced. However, the sequencing requirements may differ from a traditional ChIP–seq experiment that uses a sequence-specific transcription factor. Representative sequencing read counts for recently published examples of these techniques are shown in TABLE 2.

Of all issues that require consideration when designing such experiments, the most important one is perhaps the complexity of the library to be sequenced (BOX 1), which is mostly influenced by the proportion of the genome that is expected to be targeted and by the amount of starting material. Experiments that target a large proportion, or even most, of the genome (for example, DNaseI-seq and MeDIP–seq) require a larger number of reads than experiments that target a small proportion of the genome (for example, iCLIP and 4C). Additionally, a library that is produced from a small amount of starting material will be of low complexity, and its sequencing will be rapidly exhausted. For example, CLIP experiments often start from small amounts of purified RNA, which cause many of the sequenced reads to be identical[69,89,90]. These identical reads are assumed to be PCR duplicates, although new techniques, such as random barcoding, are helping to ameliorate this problem[70]. A second issue for consideration is that the signal-to-noise ratio determines the number of reads that is necessary to distinguish genuine signals from background signals, and higher noise levels require a greater number of reads. Techniques that use exonucleases, such as ChIP–exo and iCLIP, are expected to show low background signals, as nonspecific nucleic acids are removed by digestion[67,70]. This does not only reduce the necessary sequencing depth but also removes the need to sequence negative-control samples.

In MeDIP–seq, the required coverage is determined by the number of CpG dinucleotides in the genome. It is suggested that 60 million reads (36-bp paired-end reads) are sufficient to interrogate the majority of methylated CpG in the human genome[74]. To assess differential methylation, window-based read-counting methods can be applied, in which the genome is segmented into regions of equal size and differential methylation is inferred if the number of reads in a region differs significantly between conditions. Methods such as DESeq and EdgeR take into account different read depth between samples, as well as the noise due to the counting process and biological variation. However, there are no current guidelines for the amount of coverage and the number of replicates that are required to accurately call differentially methylated regions.

In some experiments, only a proportion of all reads that are mapped will prove to be useful. For example, in CLIP experiments, mutations at the site of crosslinking can be used to identify the precise location of crosslinking, but these mutations only happen in a minority of the reads that map to a region[91].

Finally, some interactions will be rarer than others, and their detection requires greater numbers of reads. This is particularly apparent in experiments that involve transcripts, such as CLIP and CHART. This is because transcripts are expressed at varying levels and most reads from any experiment map to highly expressed transcripts. Thus, to confidently identify interactions that involve lowly expressed transcripts, considerably more reads are required.

3C assays. 3C is a high-throughput sequencing approach for capturing interactions between two genomic regions. The frequency by which paired reads are mapped to two regions is considered to indicate the physical proximity of these regions in the nucleus. Concepts and applications of several methods that are derived from 3C have been reviewed elsewhere[75]. One of these methods — 4C — assays the interactions from one location in the genome and requires relatively few reads (that is, one to two million reads[92,93]). A *trans* interaction is unlikely to be captured because each cell in a population can only contribute at most two ligation products to a library — one from each copy of the bait — and most of these are likely to be local interactions.

A second method — Hi-C — measures interactions between all possible sites with all other possible sites that cover the whole genome. This results

© 2014 Macmillan Publishers Limited. All rights reserved

Personalis EX2022.129

Mixed-source factor
A protein factor or modification that produces peaks which are similar to those of both point-source and broad-source factors.

Technical replicates
Replicates that are derived from the same initial biological sample (as opposed to biological replicates). The variation between two such samples will be due to the variation that is introduced by the technique used rather than the underlying variation in the biology.

PCR duplicates
Pairs of reads that originated from the same molecule in the original biological sample and that are filtered out in many analyses.

Library complexity
The number of unique biological molecules that are represented in a sequencing library.

in as many as 10^{11} possible combinations, and library complexity is therefore not a limiting factor[94]. The required read count depends on the required resolution of the results and on the expected frequency of the interactions. In experiments that have fewer reads or regions with fewer interactions, the contact frequency must be averaged over large windows to gain an accurate estimate. Thus, for *cis* interactions a resolution of 400 kb requires 16.5 million unique reads, whereas to achieve a resolution of 100 kb, >100 million reads are recommended[94]. *Trans* interactions are expected to be much rarer, and thus more reads are required for their detection: 100 million reads that allow a resolution of 100 kb for *cis* interactions only yield a resolution of 1 Mb for *trans* interactions[94].

For other approaches, including 5C (in which a large number of individually designed 3C experiments are conducted in parallel) and ChIA–PET (which combines Hi-C with a ChIP step to recover interactions that are facilitated by a protein factor), the required read count depends on what is being captured. The number of interactions that is assayed in a 5C experiment is completely at the discretion of the experimenters[95]. In ChIA–PET studies the number of interactions depends on the DNA-binding protein used; a recent protocol recommends the use of at least 20 million reads[81]. It must also be borne in mind that a proportion of the molecules in these libraries do not represent valid interactions. For example, in Hi-C, unligated fragments (that is, 'dangling ends') may be present in as much as 10–45% of a library[94]. Processing pipelines for all 3C-based methods recommend the removal of duplicate reads. This means that the total number of reads with useful information is even smaller than the initially apparent number.

Conclusions and future directions

Many factors influence the minimum read depth that is required to adequately address a biological question using sequencing. Design of these experiments requires careful consideration of issues that relate to biases in genome structure, transcriptome complexity and read mappability; to the relative abundance of reads that inform about the biological question (TABLE 1); and to the trade-off between controlled, replicated design and sequencing depth. Approaches that deplete uninformative reads or that enrich for informative reads will boost experimental power, for example, by allowing greater sampling of rare alternative protein-coding transcripts or lncRNAs. Features that are more rarely sampled by sequencing are not necessarily the least interesting because, for example, a transcription factor may carry out its critical function only at a single genomic site to which it has moderate binding affinity, and a lncRNA may be required to convey its function close to its site of synthesis before being degraded rapidly. To reveal such instances requires sequencing at greater depths.

Saturation analyses can be applied as an attempt to calculate the required depth at which sequencing must be carried out. However, such analyses presume a fixed true-positive set of transcripts or binding locations, the recovery of which is increased with increased sequencing depth. Care must be taken when dealing with heterogeneous samples, as the true set may be cell type specific.

Future experiments are likely to benefit from lower sequencing costs. The greater benefit is perhaps expected from increases in the numbers of samples — such as from individuals in WGS or WES and from single cells when studying somatic changes during tumour evolution — that can be sequenced. However, lower costs would also provide more widespread opportunities for laboratories to increase the sequencing depth that is used in their experiments. Increased depth would be expected to improve the precision by which known phenomena can be defined and to reveal new phenomena that cannot be observed using current experimental designs. Future improvements in sequencing technology, such as longer read lengths and/or reduced error rates, would lower the sequencing depths that are required for genome sequencing and resequencing experiments but not for many counting-based methods, such as RNA-seq and ChIP–seq.

1. Wetterstrand, K. A. DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP). *National Human Genome Research Institute* [online], www.genome.gov/sequencingcosts (2013).
2. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
3. Schatz, M. C., Delcher, A. L. & Salzberg, S. L. Assembly of large genomes using second-generation sequencing. *Genome Res.* **20**, 1165–1173 (2010).
4. Li, R. *et al.* The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
5. Jia, J. *et al. Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* **496**, 91–95 (2013).
6. Voskoboynik, A. *et al.* The genome sequence of the colonial chordate, *Botryllus schlosseri*. *Elife* **2**, e00569 (2013).
7. Ribeiro, F. J. *et al.* Finished bacterial genomes from shotgun sequence data. *Genome Res.* **22**, 2270–2277 (2012).
8. Schatz, M. C., Witkowski, J. & McCombie, W. R. Current challenges in *de novo* plant genome sequencing and assembly. *Genome Biol.* **13**, 243 (2012).
9. Margulies, E. H. *et al.* An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc. Natl Acad. Sci. USA* **102**, 4795–4800 (2005).
10. Green, P. 2x genomes — does depth matter? *Genome Res.* **17**, 1547–1549 (2007).
11. Rands, C. M. *et al.* Insights into the evolution of Darwin's finches from comparative analysis of the *Geospiza magnirostris* genome sequence. *BMC Genomics* **14**, 95 (2013).
12. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
This is the first study to sequence a human genome using short reads; it examines the read depth that is required for calling SNVs.
13. Ahn, S. M. *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.* **19**, 1622–1629 (2009).
14. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
15. Ajay, S. S., Parker, S. C., Abaan, H. O., Fajardo, K. V. & Margulies, E. H. Accurate and comprehensive sequencing of personal genomes. *Genome Res.* **21**, 1498–1505 (2011).
16. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G + C)-biased genomes. *Nature Methods* **6**, 291–295 (2009).
17. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).
18. Clark, M. J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nature Biotech.* **29**, 908–914 (2011).
19. Sulonen, A. M. *et al.* Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol.* **12**, R94 (2011).
20. Zhou, Q. *et al.* A hypermorphic missense mutation in *PLCG2*, encoding phospholipase Cγ2, causes a dominantly inherited autoinflammatory disease with immunodeficiency. *Am. J. Hum. Genet.* **91**, 713–720 (2012).
21. Thauvin-Robinet, C. *et al. PIK3R1* mutations cause syndromic insulin resistance with lipoatrophy. *Am. J. Hum. Genet.* **93**, 141–149 (2013).
22. Yu, T. W. *et al.* Using whole-exome sequencing to identify inherited causes of autism. *Neuron* **77**, 259–273 (2013).
23. Quail, M. A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nature Methods* **5**, 1005–1010 (2008).

© 2014 Macmillan Publishers Limited. All rights reserved

Personalis EX2022.130

24. Fromer, M. *et al.* Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am. J. Hum. Genet.* **91**, 597–607 (2012).
25. Krumm, N. *et al.* Copy number variation detection and genotyping from exome sequence data. *Genome Res.* **22**, 1525–1532 (2012).
26. Xie, C. & Tammi, M. T. CNV–seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10**, 80 (2009).
27. Medvedev, P., Fiume, M., Dzamba, M., Smith, T. & Brudno, M. Detecting copy number variation with mated short reads. *Genome Res.* **20**, 1613–1622 (2010).
28. Klambauer, G. *et al.* cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* **40**, e69 (2012).
29. Le, S. Q. & Durbin, R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.* **21**, 952–960 (2011).
30. Li, Y., Sidore, C., Kang, H. M., Boehnke, M. & Abecasis, G. R. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res.* **21**, 940–951 (2011).
31. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
32. Pasaniuc, B. *et al.* Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genet.* **44**, 631–635 (2012).
33. Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473–477 (2010).
34. Schuh, A. *et al.* Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood* **120**, 4191–4196 (2012).
35. Li, B. *et al.* A likelihood-based framework for variant calling and *de novo* mutation detection in families. *PLoS Genet.* **8**, e1002944 (2012).
36. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43**, 491–498 (2011).
37. Nagarajan, N. & Pop, M. Sequence assembly demystified. *Nature Rev. Genet.* **14**, 157–167 (2013).
38. Bradnam, K. R. *et al.* Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *GigaScience* **2**, 10 (2013).
39. Salzberg, S. L. *et al.* GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* **22**, 557–567 (2012).
40. Iqbal, Z., Turner, I. & McVean, G. High-throughput microbial population genomics using the Cortex variation assembler. *Bioinformatics* **29**, 275–276 (2013).
41. Nookaew, I. *et al.* A comprehensive comparison of RNA-seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae. Nucleic Acids Res.* **40**, 10084–10097 (2012).
42. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63 (2009).
43. Kingston, R. E. Preparation of poly(A)$^+$ RNA. *Curr. Protoc. Mol. Biol.* **21**, 4.5.1–4.5.3 (2001).
44. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
In this study, RNA-seq data from 15 deeply sequenced ENCODE human cell lines are presented. It catalogues transcribed regions of the human genome and describes expression levels, RNA processing and subcellular localization for various classes of RNAs.
45. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
46. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
47. External RNA Controls Consortium. Proposed methods for testing and selecting the ERCC external RNA controls. *BMC Genomics* **6**, 150 (2005).
48. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
This study describes the use of synthetic RNAs for assessing the performance of RNA-seq methods. The importance of benchmarking performance and the limits of detection of RNA-seq are highlighted. It also reports the dependence of transcript detection on transcript length, GC composition and abundance.
49. Hansen, K. D., Brenner, S. E. & Dudoit, S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res.* **38**, e131 (2010).
50. Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome Res.* **21**, 2213–2223 (2011).
51. Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. *Nature Rev. Genet.* **8**, 413–423 (2007).
52. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
53. Haas, B. J., Chin, M., Nusbaum, C., Birren, B. W. & Livny, J. How deep is deep enough for RNA-seq profiling of bacterial transcriptomes? *BMC Genomics* **13**, 734 (2012).
54. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics* **11**, 94 (2010).
55. ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. *Science* **306**, 636–640 (2004).
56. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111 (2009).
57. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
Using deeply sequenced human H1 embryonic stem cells, the ENCODE consortium describes the dependency of accurate transcript abundance on the number of sequenced reads and finds that 80% of transcripts that are expressed at > 10 FPKM can be accurately quantified using ~ 36 million reads.
58. Halvardson, J., Zaghlool, A. & Feuk, L. Exome RNA sequencing reveals rare and novel alternative transcripts. *Nucleic Acids Res.* **41**, e6 (2013).
59. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
60. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
61. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotech.* **31**, 46–53 (2013).
62. Kalsotra, A. & Cooper, T. A. Functional consequences of developmentally regulated alternative splicing. *Nature Rev. Genet.* **12**, 715–729 (2011).
63. Sultan, M. *et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 956–960 (2008).
This is the first study to use deep RNA-seq to assess the extent of alternative splicing in human cells. It finds that the majority of human genes are spliced and that isoform distribution is variable across different cell types.
64. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
65. Dillman, A. A. *et al.* mRNA expression, splicing and editing in the embryonic and adult mouse cerebral cortex. *Nature Neurosci.* **16**, 499–506 (2013).
66. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of *in vivo* protein–DNA interactions. *Science* **316**, 1497–1502 (2007).
67. Rhee, H. S. & Pugh, B. F. ChIP–exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr. Protoc. Mol. Biol.* **100**, 21.24.1–21.24.14 (2012).
68. Sanford, J. R. *et al.* Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res.* **19**, 381–394 (2009).
69. Licatalosi, D. D. *et al.* HITS–CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464–469 (2008).
70. Konig, J. *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Struct. Mol. Biol.* **17**, 909–915 (2010).
71. Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR–CLIP. *Cell* **141**, 129–141 (2010).
72. Simon, M. D. *et al.* The genomic binding sites of a noncoding RNA. *Proc. Natl Acad. Sci. USA* **108**, 20497–20502 (2011).
73. Chu, C., Qu, K., Zhong, F. L., Artandi, S. E. & Chang, H. Y. Genomic maps of long noncoding RNA occupancy reveal principles of RNA–chromatin interactions. *Mol. Cell* **44**, 667–678 (2011).
74. de Laat, W. & Dekker, J. 3C-based technologies to study the shape of the genome. *Methods* **58**, 189–191 (2012).
This is an introduction to a useful methods volume that contains detailed discussion of the experimental considerations (including sequence depth) and computational considerations that are required when designing high-throughput 3C-type experiments.
75. Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Rev. Genet.* **14**, 390–403 (2013).
76. Hesselberth, J. R. *et al.* Global mapping of protein–DNA interactions *in vivo* by digital genomic footprinting. *Nature Methods* **6**, 283–289 (2009).
77. Down, T. A. *et al.* A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature Biotech.* **26**, 779–785 (2008).
78. Blackledge, N. P. *et al.* Bio-CAP: a versatile and highly sensitive technique to purify and characterise regions of non-methylated DNA. *Nucleic Acids Res.* **40**, e32 (2012).
79. Landt, S. G. *et al.* ChIP–seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22**, 1813–1831 (2012).
This paper presents the ENCODE guidelines for ChIP–seq and similar experiments, which provide a baseline minimum standard for the design of new studies, including recommendations on sequencing depth, number of replicates, controls and measures to assess the quality of results.
80. 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).
81. Chen, Y. *et al.* Systematic evaluation of factors influencing ChIP–seq fidelity. *Nature Methods* **9**, 609–614 (2012).
This is a comprehensive analysis of the factors that affect the success of a ChIP–seq experiment, including sequencing depth, which is carried out to a high maximum depth.
82. Ozdemir, A. *et al.* High resolution mapping of Twist to DNA in *Drosophila* embryos: efficient functional analysis and evolutionary conservation. *Genome Res.* **21**, 566–577 (2011).
83. Rozowsky, J. *et al.* PeakSeq enables systematic scoring of ChIP–seq experiments relative to controls. *Nature Biotech.* **27**, 66–75 (2009).
84. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nature Rev. Genet.* **10**, 669–680 (2009).
85. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Statist.* **5**, 1752–1779 (2011).
86. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein–DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408–1419 (2011).
87. Rhee, H. S. & Pugh, B. F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483**, 295–301 (2012).
88. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).
89. Cho, J. *et al.* LIN28A is a suppressor of ER-associated translation in embryonic stem cells. *Cell* **151**, 765–777 (2012).
90. Eom, T. *et al.* NOVA-dependent regulation of cryptic NMD exons controls synaptic protein levels after seizure. *Elife* **2**, e00178 (2013).
91. Asan *et al.* Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol.* **12**, R95 (2011).
92. van de Werken, H. J. G. *et al.* Robust 4C–seq data analysis to screen for regulatory DNA interactions. *Nature Methods* **9**, 969–972 (2012).

© 2014 Macmillan Publishers Limited. All rights reserved

Personalis EX2022.131

93. Splinter, E., de Wit, E., van de Werken, H. J. G., Klous, P. & de Laat, W. Determining long-range chromatin interactions for selected genomic sites using 4C–seq technology: from fixation to computation. *Methods* **58**, 221–230 (2012).
94. Belton, J.-M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276 (2012).
95. Ferraiuolo, M. A., Sanyal, A., Naumova, N., Dekker, J. & Dostie, J. From cells to chromatin: capturing snapshots of genome organization with 5C technology. *Methods* **58**, 255–267 (2012).
96. Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231–239 (1988).
97. Veal, C. D. *et al.* A mechanistic basis for amplification differences between samples and between genome regions. *BMC Genomics* **13**, 455 (2012).
98. Sampson, J., Jacobs, K., Yeager, M., Chanock, S. & Chatterjee, N. Efficient study design for next generation sequencing. *Genet. Epidemiol.* **35**, 269–277 (2011).
99. Wang, W., Wei, Z., Lam, T. W. & Wang, J. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. *Scientif. Rep.* **1**, 55 (2011).
100. Hatem, A., Bozdag, D., Toland, A. E. & Catalyürek, Ü. V. Benchmarking short sequence mapping tools. *BMC Bioinformatics* **14**, 184 (2013).
101. Mijuskovic, M. *et al.* A streamlined method for detecting structural variants in cancer genomes by short read paired-end sequencing. *PLoS ONE* **7**, e48314 (2012).
102. Lee, H. & Schatz, M. C. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics* **28**, 2097–2105 (2012).
103. Derrien, T. *et al.* Fast computation and applications of genome mappability. *PLoS ONE* **7**, e30377 (2012).
104. Daley, T. & Smith, A. D. Predicting the molecular complexity of sequencing libraries. *Nature Methods* **10**, 325–327 (2013).
105. Gottwein, E. *et al.* Viral microRNA targetome of KSHV-infected primary effusion lymphoma cell lines. *Cell Host Microbe* **10**, 515–526 (2011).
106. Rogelj, B. *et al.* Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. *Scientif. Rep.* **2**, 603 (2012).
107. Zhang, J. *et al.* ChIA–PET analysis of transcriptional chromatin interactions. *Methods* **58**, 289–299 (2012).
108. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
109. Taiwo, O. *et al.* Methylome analysis using MeDIP–seq with low DNA concentrations. *Nature Protoc.* **7**, 617–636 (2012).
110. Long, H. K. *et al.* Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *Elife* **2**, e00348 (2013).

Acknowledgements
The Computational Genomics Analysis and Training Centre is funded by a UK Medical Research Council Strategic Award.

Competing interests statement
The authors declare no competing interests.

FURTHER INFORMATION
Computational genomics analysis and training: www.cgat.org
Standards, guidelines and best practices for RNA-seq: http://encodeproject.org/encode/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

© 2014 Macmillan Publishers Limited. All rights reserved

Personalis EX2022.132