Published in final edited form as:

Nat Immunol. 2012 September ; 13(9): 802–807. doi:10.1038/ni.2407.

Technical considerations for functional sequencing assays

Weihua Zeng and Ali Mortazavi

Department of Developmental and Cell Biology, University of California, Irvine, California, USA
and at the Center for Complex Biological Systems, University of California, Irvine, California, USA

Ali Mortazavi: ali.mortazavi@uci.edu

Abstract

Next-generation sequencing technologies need careful design of experiments and evaluation of results to meet field requirements. Here we discuss technical considerations for these high-throughput assays, together with criteria to assess the quality of the results and the necessary validation.

The application of sequencing technologies to answer functional questions in biological systems is now standard. This shift from array-based assays was spearheaded by the widespread adoption starting in 2007 of chromatin immunoprecipitation (ChIP) coupled with high-throughput sequencing (ChIP-seq) that has made genome-wide, high-resolution mapping of protein DNA-binding highly efficient and affordable with at least ten million mapped sequence tags ('reads') (technical considerations for ChIP-seq are discussed in Kidder *et al.*¹).

However, patterns of protein-DNA binding alone do not account for all changes in gene expression, which is the ultimate goal of many functional studies. It is still necessary to measure global changes in gene expression levels, such as changes in transcription (for example, alternative splicing, promoter use or polyadenylation sites) within the same gene, to characterize the activity of regulatory elements of these genes besides those bound by known factors, and to identify the genes that are affected by distal regulatory elements. As short-read sequencing technologies increased throughput to 20–40 million reads per sequencing run in 2008, it became practical to apply them to transcriptome quantification and transcript discovery by sequencing cDNA derived from RNA (RNA-seq)^{2,3}, which has driven the widespread adoption of functional sequencing. High-throughput sequencing also enabled the genome-wide identification of open chromatin regulatory elements such as enhancers and promoters, through characterization of DNase I-hypersensitive regions (DNase-seq)^{4,5}. As regulatory elements that control gene expression can be 1 megabase or farther away from their target promoters, or on another chromosome altogether, the genome-wide mapping of long-range interactions using assays such as carbon-copy chromosome conformation capture (5C)⁶ and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET)⁷ are becoming increasingly important.

In this Commentary we discuss the experimental design of each of these assays individually before addressing sequencing-related issues that are common to all of them.

RNA-seq

The measurement of RNA expression is a foundation of many experiments done in biomedical research. It is therefore natural that the sequencing of long and short RNA both for quantification and discovery is the most popular functional sequencing assay (Fig. 1a). Quantification of mRNA transcripts in RNA-seq is performed by calculating values reported in units of reads per kilobase per million mapped reads (RPKM²) with a paired-end fragment equivalent, fragments per kilobase per million reads (FPKM)⁸, also commonly used for each gene (Fig. 1b). RPKM normalizes for differences in gene size and makes the comparison of genes within the same sample meaningful in terms of molar equivalents (Fig. 1b). As RNA-seq is not based on predetermined DNA probes to known genes, it is a powerful tool for the discovery of new exons, splice junctions, transcripts and genes⁸ as well as new small RNAs⁹ (Fig. 1c). The reads can be used to assemble transcripts that result from gene rearrangements and can also help to identify disease-associated genomic abnormalities^{10,11}. Properly filtered RNA-seq reads can be mined for sequence variants and RNA-editing events with tuned analysis and filtering pipelines¹² (Fig. 1d).

RNA sample preparation

As with all high-throughput sequencing assays, the quality of RNA-seq results depends on the quality of the starting material, method-specific library construction and data processing. Major factors that contribute to quality are (i) the enrichment and homogeneity of the samples, and (ii) the integrity of the recovered RNA. Whereas some experimental protocols are amenable to analyses of mixed populations, the best results are obtained when homogenous cell populations are isolated using a variety of techniques such as cell sorting because the use of mixed samples could lead to the misassembly of transcripts. Total RNA is extracted and its quality should be measured using the Agilent Bioanalyzer RNA integrity number (RIN)¹³ to assess the extent of RNA degradation. RNA samples with RINs greater than 8 (out of 10) are suitable for library construction. One to ten micrograms of total RNA is sufficient for robust RNA-seq library construction, and this amount can be extracted from 0.3–3 million primary T cells or B cells. In cases when acquiring cells for at least 10 nanograms of total RNA per sample is challenging (for example, rare cell populations), performing RNA-seq with a few or even single cells requires the use of specially adapted PCR amplification strategies to increase the amount of starting material linearly before or as part of library construction¹⁴. The quality of highly amplified samples should be monitored by real-time quantitative PCR (QPCR) with both positive and negative controls to measure distortions caused by this extra amplification. Monitoring can be accomplished by adding known quantities of exogenous transcripts ('spike-ins') to the sample.

Total RNA consists of >90% rRNA as well as considerable amounts of tRNA that are normally not interesting targets for sequencing. There are several strategies to remove rRNA from the total RNA samples to enrich for mRNA. As most mature mRNAs transcribed by RNA polymerase II have a poly(A) tail (although several histone transcripts are well-known exceptions), mRNA can be highly enriched by hybridization capture using oligo(dT) beads.

Another benefit of poly(A) selection is to remove immature RNA molecules that are not yet completely processed, and two rounds of selection produce a cleaner result than one round does. An alternate strategy consists of removing the rRNA using hybridization capture with beads bearing sequences complementary to rRNAs and thereby enriching the other RNA molecules in the flow-through. Although the hybridization efficiency is respectable, the potentially higher percentage of rRNA in total RNA samples (5–10%) always makes the remaining rRNA amount a concern for such ‘ribo-minus’ strategies¹⁵. Whereas reads from rRNA can be filtered out computationally, reads containing sequencing errors can map to ancient ribosomal repeats, which can mislead downstream analyses to predict a novel transcript. Several strategies can be applied to enrich for small RNAs such as performing size selection through polyacrylamide gel electrophoresis or using commercially available kits designed for this purpose. Small RNA library construction typically involves RNA adaptor ligation to the 5′ and 3′ end of these molecules before reverse transcription⁹.

cDNA synthesis

Nearly all commonly used sequencing technologies require DNA libraries; therefore RNA must be converted into cDNA. Typically, a conventional protocol for synthesis of double-stranded cDNA is used to maximize the efficiency and fidelity of reverse transcription. To ensure even coverage of whole transcripts, RNA is first fragmented to minimize secondary structure formation and to reduce end biases, before first-strand synthesis with random primers². First-strand cDNA synthesis with the oligo(dT) primer, followed by sonication of double-strand cDNA into small fragments, is still used at times. Although priming with oligo(dT) introduces 3′ bias, and cDNA fragmentation contributes to both 3′ and 5′ bias, this strategy is still viable for single-cell RNA-seq assays to obtain enough starting material¹⁴.

Strand information is particularly helpful for distinguishing overlapping transcripts on opposite strands, especially for *de novo* transcript discovery¹⁶. One strategy to preserve strandedness is to ligate adapters in a predetermined direction to the RNA or first-strand cDNA¹⁷. However, this method is time-consuming, labor-consuming and causes high bias to both 5′ and 3′ ends of cDNA molecules (these biases are also present in small RNA libraries). A more robust strategy is to integrate a chemical label such as dUTP to the second-strand cDNA and thus make it possible to distinguish the second-strand cDNA from the first strand during library construction¹⁶. A small amount of reads are observed from the opposite strand at a low level (~1%). Thus, the validity of antisense transcripts on highly expressed genes must be considered carefully.

In addition to conventional work flows of RNA-seq library construction, new methods are being developed to enable scaling up experiments in parallel. One such method is based on the Nextera strategy¹⁸, which integrates the adaptor sequences into the target cDNA with a transposase, followed by direct PCR amplification to generate the library. The Nextera strategy provides substantial time and labor savings by eliminating the steps of adaptor ligation and size selection, with the potential downsides of biases at both 5′ and 3′ ends and the risk of over-representing genomic contamination.

It is critical that the distribution of fragment sizes in RNA-seq libraries be tightly controlled and unimodal (shown as a single peak in Agilent Bioanalyzer result; Fig. 1a), as it is an

important parameter for bioinformatic analysis and accurate transcript assembly. A narrow band at 50 base pairs (bp) corresponding to the cDNA sample with ligated adapters using size selection through gel electrophoresis² or beads in automated pipelines works best. The distribution of fragment sizes can be checked by Agilent Bioanalyzer before sequencing and should be verified after sequencing computationally. The 5'- and 3'-end biases are common issues that plague RNA-seq samples are caused typically by inappropriate sample preparation, fragmentation, reverse transcription, size selection or adaptor ligation. As poly(A) selection of degraded RNA results in substantial 3'-end bias, samples should be assessed after sequencing for relative evenness of coverage along the length of highly expressed transcripts. RNA-seq samples in which more than 50% of the reads fall on the last 300 bp of annotated genes are too degraded to quantify using standard tools.

Analysis

RNA-seq data analysis is more computationally challenging than that for other sequencing-based methods because a considerable fraction of the reads map across splice junctions. As reads become longer, the fraction of reads crossing one or more splice junctions grows. Furthermore, alternative splicing increases the complexity of transcript modeling and quantification. It is easier to do quantification using only known, high-quality gene models. However, the search for cell type-specific new isoforms and long noncoding RNAs continues unabated even in well-annotated genomes such as human and mouse. It is critically important to filter transcript-discovery artifacts from genuine novel gene models before combining them with existing gene annotations to more accurately quantify gene expression. One strategy involves filtering new predicted transcript models that have extremely low RPKM (for example, <0.1 RPKM) or extremely high RPKM ($>10^5$ RPKM) as well as transcripts not found in more than one biological replicate with more sophisticated filtering algorithms in active development.

A variety of programs exist to analyze RNA-seq data when a reference genome is available that follow similar overall strategies. Reads are aligned to both the reference genome and across splice junctions. The mapped reads are then assembled into transcripts and compared to known gene annotations to identify the novel transcripts. Differences in expression are analyzed using one of several statistical methods. A popular RNA-seq analysis pipeline is the combination of Tophat and Cufflinks¹⁹. Reads are mapped with Tophat and assembled into transcripts with FPKM values using Cufflinks. These transcripts are combined with the reference gene annotations using Cuffmerge and 'differential' gene or transcript expression in multiple conditions or time points is called using Cuffdiff. Published alternatives to Cufflinks include Scripture²⁰ and RSEM²¹. There are several software packages designed specifically to detect differential gene expression such as edgeR²², DESeq²³ and DEGseq²⁴, which work from read-count data on existing gene annotations.

Replicates are crucial to estimate the variability within the same condition and between different treatments. Although the use of replicates is optional with some packages, it is required by others and is a necessity for reducing false positives. Biological replicates (different growths of cells processed at different times) are preferred because technical replicates (different lanes of the same library) show little variation. However, there will be

systematic variation when comparing the same biological sample sequenced with different RNA-seq protocols, platforms or read lengths, which would justify technical replicates in addition to biological replicates.

Sequence differences from the reference genome in RNA-seq reads can be mined for known single-nucleotide polymorphisms (SNPs) and new single nucleotide variants (SNVs). The identified SNPs and SNVs can serve as markers to quantify allele-specific expression^{25,26} as well as filtered against the reference to characterize potential RNA editing events when a subset of SNVs are not found in the genomic sequence. Deep whole-genome sequencing of an individual genome is the most reliable method of detecting SNVs. However, whole-genome sequencing requires billions of reads to achieve the 100× average coverage of exome sequencing that would lead to acceptable genome-wide coverage in the difficult-to-sequence portions of the genome. Searching for explanatory SNVs in RNA-seq data is informative because the most easily understood disease-related mutations are those in the coding sequence of expressed genes, and RNA-seq can result in very high coverage in medium to highly expressed genes that match or exceeds coverage attained with exome sequencing. For example, a gene expressed at 10 RPKM in a RNA-seq sample sequenced to a depth of one hundred million 100-bp mapped reads will have 100× average coverage. RNA-seq reads with one or two mismatches to the reference (or resequenced) genome are typically filtered under strict criteria to remove sequencing and mapping errors resulting from known artifacts introduced by library preparation, sequencing technologies and spliced read mapping. Candidate SNVs can then be compared to known variation databases such as dbSNP using ANNOVAR²⁷ or other packages designed to annotate sequence variants. Important candidates should be validated by Sanger sequencing genomic DNA PCR products covering the target regions.

SNVs can serve as markers to quantify allele-specific expression, which provides important clues to assessing the *cis*-regulatory differences between different haplotypes as well as the occurrence of imprinting or random mono-allelic expression^{25,26}. RNA-seq reads can be mined to discover RNA editing events, which are post-transcriptional changes to transcripts that can affect coding sequence or untranslated regions¹². The challenge here is to identify and filter out genomic variants specific to the sample, but that may not be found in public databases such as dbSNP, as well as to minimize the systematic impact of mismatched reads from uncharacterized splice junctions.

DNase-seq

A well-known characteristic of active DNA elements (promoters, enhancers and even insulators) is that they are more accessible to DNase I digestion than the rest of the genome²⁸. DNase-seq is used to measure the global distribution pattern of DNase I cleavage and identify open chromatin associated with active DNA elements. DNase-seq is particularly attractive for discovery of candidate regulatory regions without relying on availability and specificity of antibodies. In this assay, nuclei of samples are isolated and digested with DNase I for a short period of time. The released DNA fragments are isolated and sequenced to identify the DNase I-hypersensitive regions⁴ (Fig. 2). DNase I hypersensitivity assays work reliably only in fresh samples and thus preclude snap-frozen or

otherwise fixed samples from consideration. The cell lysis time and the applied concentration of detergent should be optimized to achieve maximal lysis efficiency while still keeping the majority of nuclei intact. The number of recovered nuclei after lysis should be compared with the cell count before treatment; a recovery rate of 80–100% indicates a successful lysis procedure (Fig. 2).

There are two different strategies to obtain the DNase I-hypersensitive fragments after the isolation of nuclei: either by one DNase I cut⁵ or two DNase I cuts⁴. Regardless of the exact method, a critical part of DNase-seq involves fine-tuning the amount of DNase I necessary to cut the hypersensitive regions without overdigestion. DNA fragments generated by DNase I-hypersensitive digestion are purified by either gel electrophoresis or sucrose gradient ultracentrifugation, and the optimal size-selection ranges are different for the one-cut and two-cut strategies^{4,5}. Size selection before and during library construction directly impacts the representation of both strong and weak DNase I digestion sites. Therefore, quality control using QPCR or Southern blot is important to ensure the enrichment of known open chromatin elements such as the globin locus control region compared to control, non-hypersensitive regions. Samples with greater than 20-fold enrichment between positive and negative control regions should make good libraries. Greater fold enrichment typically correlates with better sample quality.

After read mapping, the background noise resulting from random digestion with DNase I—which is usually much weaker than the signal from authentic DNase I-hypersensitive sites—is removed by comparing the signal at a given location ('bin') to the signals from a large flanking region and the expected background signal intensity based on the false discovery rate. The sharp boundaries of genuine DNase I-hyper-sensitivity signals between adjacent bins are further dampened with a 'smoothing' function. Software such as F-seq²⁹ and algorithms such as HotSpot³⁰ are optimized to work on data from the different protocols for the assessment of DNase I-hypersensitivity and can be used to identify peaks of DNase I hypersensitivity. To correct for the bias in the efficiency of DNase I digestion in different regions with extreme base composition or copy-number variation, a comparison to undigested DNA (similar to what is done for ChIP-seq) should be included to the data analysis⁴.

Bound transcription factors protect their binding sites from digestion, whereas the sequences immediately flanking the binding sites are still accessible to DNase I. Therefore, a current area of active research focuses on the deep sequencing of DNase-seq libraries to reveal the 'footprint' of these factors⁵. Hypersensitive regions and footprints can be mined for sequence motifs using a variety of programs⁴, such as MEME. The identified motifs can also be compared with those from the transcription factor motif databases such as JASPAR and TRANSFAC. These results are often compared to ChIP-seq data for specific factors in the same or related cell types. We refer readers to reference 31 for a discussion of the analysis of candidate regions for *cis*-regulatory elements.

Long-range interaction assays

One of the best-characterized mammalian long-range acting regulatory elements is the locus control region of the β -globin locus, which is 50 kilobases (kb) away from the genes that it regulates. Some human enhancers are known to be up to a megabase away from their target promoters. This has led to the development of several assays to detect interactions between distal DNA regions, whether they are promoters, enhancers or have any other functional roles.

Direct long-range physical contacts can be detected using chromosome conformation capture (3C) assay³². The 3C assay can be used to measure the interaction between two chromatin regions in a 'one-to-one' fashion by detecting PCR products that are present only when two distal, digested regions are ligated together. Several 3C-derived strategies have been introduced that allow higher throughput. The circular chromosome conformation capture (4C) assay is based on divergent primers covering the bait sequence to detect all regions interacting with it³³. The 5C assay requires probes for thousands of restriction digestion fragments that, in combination, can be used to assay millions of interactions⁶.

In 2009, two strategies to detect the global chromatin interactions were introduced (Fig. 3). Hi-C³⁴ is a global extension of 3C-based methods that relies on restriction enzymes to fragment formaldehyde-fixed genomic DNA. The sticky DNA ends generated by restriction digestion are repaired to blunt ends, and biotinylated nucleotides are incorporated at the ends of these fragments. After blunt-end ligation, the DNA is sheared by sonication and only the products containing ligation junction fragments are enriched using streptavidin beads. The purified products are subjected to library construction and paired-end sequencing (Fig. 3a).

The second method, ChIA-PET, is an extension of ChIP-seq that can be used to capture distal interactions mediated by a specific protein of interest⁷. Instead of restriction digestion, a ChIA-PET requires sonication to shear chromatin. The interacting DNA fragments bound by the protein of interest are enriched by ChIP. Biotinylated half linkers (adapters with sticky ends) are ligated to the DNA fragment ends and the juxtaposed chromatin regions are joined together via the sticky ends of the half linkers. The half linkers also contain a recognition site for the type II restriction enzyme MmeI, which cuts 20 bp downstream of its recognition site. Therefore, after MmeI digestion, the ends of the two interacting regions, called paired-end tags will be released with the ligated half linkers. The paired-end tags with the biotinylated half linkers are isolated using streptavidin beads and processed into a library for paired-end sequencing (Fig. 3b).

Hi-C requires much less starting material (usually one-fifth to one-tenth of that for ChIA-PET), can be used to detect potentially all interactions mediated by different protein factors and is easier to perform than ChIA-PET. However, Hi-C is limited by the availability of restriction sites and its resolution suffers in mammalian genomes: 1 megabase with ten million reads³⁴ and 50–100 kb when using at least one billion reads³⁵. ChIA-PET generates much higher resolution chromatin interaction maps, down to 1–10-kb scale from just tens of million reads⁷, but the quality of the antibody strongly affects the outcome of the assay, and

the assay can be difficult to optimize. The large amount of starting material (>100 million cells) makes ChIA-PET impractical for the analysis of rare cell populations.

Major concerns for all long-range chromatin interaction assays are maximizing the detection efficiency of genuine interactions while minimizing the frequency of nonspecific random ligations between unrelated DNA fragments. For 4C and 5C, the restriction enzyme, bait sequence and probes need to be carefully selected to enhance assay efficacy^{6,33}. For Hi-C, the completeness of restriction enzyme digestion and blunt-end ligation are important for efficiency of the detection. The success of the digestion and the extent of blunt-end ligation are evaluated by a carefully designed PCR digestion control assay³⁴. ChIA-PET results are affected by the size of fragments after sonication and the efficiency of crosslinking. The size range requirement for the sonication step for a ChIA-PET sample is larger than the normal size range for ChIP-seq analysis of the same factor and needs careful optimization.

Ligation is performed in all long-range interaction assays under conditions of extremely low DNA concentrations to reduce nonspecific ligation events. For 3C, 4C, 5C and Hi-C, the restriction enzyme digestion is performed under optimized SDS concentrations, which helps to remove proteins not directly crosslinked to DNA and reduce the random ligation background³⁴. Recently, a modified version of Hi-C named tethered conformation capture (TCC) has been reported³⁶ in which proteins mediating long-range DNA interaction are biotinylated, and these protein-DNA complexes are immobilized at low surface density on streptavidin beads. The interacting DNA fragments are ligated on this solid surface instead of in solution (Fig. 3a), and this modification has been shown to substantially reduce nonspecific random ligation. Therefore, TCC enhances the detection fidelity for interchromosomal interactions, which usually have much lower signal intensities and are masked by high background noise from nonspecific ligation. In contrast, ChIA-PET sequencing uses the frequency of chimeric ligation between two different half linkers to determine the probability of random ligation⁷.

To confirm new interactions identified with these high-throughput assays, 3C can be performed to analyze specific interacting regions, or an orthogonal assay such as fluorescence *in situ* hybridization should be used to validate important long-distance (>100 kb) or interchromosomal interactions. Software for analysis of data from long-range interaction assays is still under active development, including a method for Hi-C data normalization³⁷ and a ChIA-PET analysis pipeline³⁸.

Common sequencing considerations

Library complexity is an important determinant of result quality, and it is mainly affected by the amount of starting material going into the assay and the amount of recovered DNA after the assay. Sequencing library construction of RNA-seq and DNase-seq samples typically uses one or more size-selection steps. These size-selection steps are critical because enrichment efficiency, which is usually affected by the selected sample size range, needs to be balanced with the amount of DNA recovered. Libraries with good enrichment but too little DNA (typically less than 1 nanogram) may have complexity issues that do not yield more than a few million distinct reads. High number of PCR cycles during the amplification

step of library construction will lower library complexity. A helpful quality control assay is QPCR using the isolated material such as DNase I-treated fragments³⁹ before library construction and using the constructed libraries before sequencing with positive and negative controls. One hallmark of overamplification is the shift in size of the PCR products compared to the expected size range¹. QPCR is also the conventional way to validate the data from high-throughput sequencing assays, but it is likely to suffer from amplification artifacts, too. Finding consistent results in true biological replicates is more convincing.

Universal questions of functional sequencing are whether a library should be sequenced using single-end or paired-end reads, the read-length to use and the appropriate sequencing depth (Table 1). Although paired-end reads are twice as expensive, they deliver more than twice the information, namely the read sequences plus their connectivity. Some samples such as those used in DNase-seq and small RNA-seq are nearly always sequenced using single-end reads, whereas long-range interactions always require paired-end sequencing. mRNA is sequenced with either single- or paired-end reads depending on the objectives. If the goal is to quantify gene and transcript expression without performing discovery, then short 50-bp reads are sufficient. If the goal is to discover and assemble new transcripts, then longer paired reads are necessary to obtain as much information as possible about connectivity and splicing⁴⁰.

Another choice to make is whether to use multiplex libraries in which each sample is sequenced to a shallow to medium depth compared to deep sequencing of a single library per lane. Multiplex libraries are cost-efficient and are ideally suited for libraries with low to medium complexity that do not benefit from deeper sequencing such as small RNA libraries. Given that a lane of Illumina HiSeq can yield in 2012 about 200 million raw (unmapped) reads, investigators must decide how to allocate those reads based on whether they are interested in discovery that requires deep sequencing. For example, transcript discovery and deep DNase I footprinting require deep sequencing of 100–200 million mapped reads per replicate, whereas more shallow surveys of RNA-seq expression and open chromatin are adequately performed with 20–40 million mapped reads. Libraries for the analysis of long-range interactions that are complex enough should be sequenced as deeply as possible to identify weak or infrequent interactions.

There will undoubtedly be more changes in technology in the next few years that will alter the state of the art in functional sequencing. In particular, the promise of long reads (>1 kb) from an upcoming generation of sequencers, such as Pacific Biosciences and Oxford Nanopore, could dramatically simplify RNA-seq analysis once these platforms can sequence the transcriptome deeply enough. By contrast, DNase-seq and digital footprinting will benefit from deeper sequencing of short reads at a lower cost. Long-range chromatin interactions protocols will likely evolve further to assay smaller populations of cells at higher resolutions. Finally, computational analysis packages will continue to evolve at a fast pace as data sets proliferate. However, all new technologies and protocols will have artifacts and biases that risk being taken as a discovery if one is not careful. Last but not least, sequencing technologies have not abolished the laws of statistics nor the need for good experimental design and built-in controls. Reproducibility in biological replicates should be the *sine qua non* for accepting rare results from sequencing.

Acknowledgments

The work of A.M. on this manuscript was supported by the University of California Irvine Center for Complex Biological Systems and US National Institutes of Health P50 GM076516. We thank I. Blitz, S. Sandmeyer, M. Ho, R. Murad, R. Ramirez, M. Macchietto and E. Park for many helpful discussions of this manuscript.

References

1. Kidder BL, Hu G, Zhao K. *Nat. Immunol.* 2011; 12:918–922. [PubMed: 21934668]
2. Mortazavi A, Williams BA, Mccue K, Schaeffer L, Wold B. *Nat. Methods.* 2008; 5:621–628. [PubMed: 18516045]
3. Cloonan N, et al. *Nat. Methods.* 2008; 5:613–619. [PubMed: 18516046]
4. Hesselberth JR, et al. *Nat. Methods.* 2009; 6:283–289. [PubMed: 19305407]
5. Boyle AP, et al. *Genome Res.* 2011; 21:456–464. [PubMed: 21106903]
6. Dostie J, Dekker J. *Nat. Protoc.* 2007; 2:988–1002. [PubMed: 17446898]
7. Fullwood MJ, et al. *Nature.* 2009; 462:58–64. [PubMed: 19890323]
8. Trapnell C, et al. *Nat. Biotechnol.* 2010; 28:511–515. [PubMed: 20436464]
9. Linsen SEV, et al. *Nat. Methods.* 2009; 6:474–476. [PubMed: 19564845]
10. Grabherr MG, et al. *Nat. Biotechnol.* 2011; 29:644–652. [PubMed: 21572440]
11. Robertson G, et al. *Nat. Methods.* 2010; 7:909–912. [PubMed: 20935650]
12. Bahn JH, et al. *Genome Res.* 2012; 22:142–150. [PubMed: 21960545]
13. Schroeder A, et al. *BMC Mol. Biol.* 2006; 7:3. [PubMed: 16448564]
14. Tang F, et al. *Nat. Protoc.* 2010; 5:516–535. [PubMed: 20203668]
15. Cui P, et al. *Genomics.* 2010; 96:259–265. [PubMed: 20688152]
16. Parkhomchuk D, et al. *Nucleic Acids Res.* 2009; 37:e123. [PubMed: 19620212]
17. Lister R, et al. *Cell.* 2008; 133:523–536. [PubMed: 18423832]
18. Gertz J, et al. *Genome Res.* 2012; 22:134–141. [PubMed: 22128135]
19. Trapnell C, et al. *Nat. Protoc.* 2012; 7:562–578. [PubMed: 22383036]
20. Guttman M, et al. *Nat. Biotechnol.* 2010; 28:503–510. [PubMed: 20436462]
21. Li B, Dewey CN. *BMC Bioinformatics.* 2011; 12:323. [PubMed: 21816040]
22. Robinson MD, McCarthy DJ, Smyth GK. *Bioinformatics.* 2010; 26:139–140. [PubMed: 19910308]
23. Anders S, Huber W. *Genome Biol.* 2010; 11:R106. [PubMed: 20979621]
24. Wang L, Feng Z, Wang X, Wang X, Zhang X. *Bioinformatics.* 2010; 26:136–138. [PubMed: 19855105]
25. Montgomery SB, et al. *Nature.* 2010; 464:773–777. [PubMed: 20220756]
26. Pickrell JK, et al. *Nature.* 2010; 464:768–772. [PubMed: 20220758]
27. Wang K, Li M, Hakonarson H. *Nucleic Acids Res.* 2010; 38:e164. [PubMed: 20601685]
28. Gross DS, Garrard WT. *Annu. Rev. Biochem.* 1998; 57:159–197. [PubMed: 3052270]
29. Boyle AP, Guinney J, Crawford GE, Furey TS. *Bioinformatics.* 2008; 24:2537–2538. [PubMed: 18784119]
30. Sabo PJ, et al. *Proc. Natl. Acad. Sci. USA.* 2004; 101:16837–16842. [PubMed: 15550541]
31. Hardison RC, Taylor J. *Nat. Rev. Genet.* 2012; 13:469–483. [PubMed: 22705667]
32. Dekker J, Rippe K, Dekker M, Kleckner N. *Science.* 2002; 295:1306–1311. [PubMed: 11847345]
33. Göndör A, Rougier C, Ohlsson R. *Nat. Protoc.* 2008; 3:303–313. [PubMed: 18274532]
34. Lieberman-Aiden E, et al. *Science.* 2009; 326:289–293. [PubMed: 19815776]
35. Dixon JR, et al. *Nature.* 2012; 485:376–380. [PubMed: 22495300]
36. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. *Nat. Biotechnol.* 2012; 30:90–98. [PubMed: 22198700]
37. Yaffe E, Tanay A. *Nat. Genet.* 2011; 43:1059–1065. [PubMed: 22001755]
38. Li G, et al. *Genome Biol.* 2010; 11:R22. [PubMed: 20181287]

39. Dorschner MO, et al. Nat. Methods. 2004; 1:219–225. [PubMed: 15782197]
40. Katz Y, Wang ET, Airoidi EM, Burge CB. Nat. Methods. 2010; 7:1009–1015. [PubMed: 21057496]

Figure 1. RNA-seq work flow

(a) Schematic diagram of RNA-seq library construction. Total RNA is extracted from 300,000 cells to 3 million cells, and a small aliquot is used to measure the integrity of the RNA. rRNA is then depleted through one of several methods to enrich subpopulation of RNA molecules, such as mRNA or small RNA. mRNA is fragmented into a uniform size distribution and the fragment size can be monitored by RNA gel electrophoresis or Agilent Bioanalyzer. The cDNA is then built into a library. The size distribution pattern of the library can be checked by Agilent Bioanalyzer; this information is important for RNA-seq

data analysis. **(b)** Mapping programs align reads to the reference genome and map splice junctions. Gene expression can be quantified as absolute read counts or normalized values such as RPKM. **(c)** If RNA-seq data sets are deep enough and the reads are long enough to map enough splice junctions, the mapped reads can be assembled into transcripts. **(d)** The sequences of the reads can be mined by comparing the transcriptome reads with the reference genome to identify nucleotide variants that are either genomic variants (for example, SNPs) or candidates for RNA editing.

Figure 2. DNase-seq work flow

Nuclei are isolated from 10 million cells, and the isolation efficiency can be determined by calculating the ratio of recovered nuclei to the starting cell number. The nuclei are subjected to DNase I digestion for a short time, and the active DNA elements are enriched by size selection. The enrichment efficiency and specificity can be examined by Southern blot or by QPCR on presumed active and inactive regulatory elements. Based on this quality control, the amount of DNase I used can be adjusted to achieve the optimal enrichment of DNase I–hypersensitive sites.

Figure 3. Hi-C and ChIA-PET work flow

(a) Chromatin from 20 million cells is collected for Hi-C. After restriction digestion, sticky ends are repaired into blunt ends and biotin (asterisks) is integrated. In conventional Hi-C, the blunt-end ligation is carried out in solution. In TCC, ligation is performed on a solid surface. After the biotin on the unligated ends is removed, the DNA samples are sheared into small fragments, and the ligation products are enriched by streptavidin pull-down and then built into libraries for sequencing. (b) At least 100 million cells are required for one ChIA-PET assay. The long-range DNA interactions that a protein factor of interest participates in are enriched by ChIP and the quality of this starting material is verified by QPCR. The frequency of random ligations between the two different half linkers (green and yellow bars) can be used to estimate the frequency of nonspecific ligation. The pair-end tags produced by MmeI digestion is purified and used in library building. The Agilent Bioanalyzer profile of the library can be used to ascertain the sample quality, as there should only be one sharp clean peak.

Table 1

Recommended sequencing formats

Assay	Biological replicates	Read length (bp)	Single-end	Paired-end	Sequence depth
RNA-seq	Yes	36–150	Yes	Yes	20–40 million for quantification >100 million for transcript discovery
DNase-seq	Yes	36–50	Yes	No	20–40 million for hypersensitive peaks >200 million for digital footprinting
Long-range (5C, Hi-C or ChIA-PET)	Yes	36 (5C and ChIA-PET) 75 (Hi-C)	No	Yes	20 million for strong interactions 100–1,000 million for weaker interactions