Published in final edited form as:
Nat Methods. 2009 November ; 6(11 0): S22–S32. doi:10.1038/nmeth.1371.

Computation for ChIP-seq and RNA-seq studies

Shirley Pepke[1], **Barbara Wold**[2], and **Ali Mortazavi**[2,3]

[1]Center for Advanced Computing Research, California Institute of Technology, Pasadena, CA, 91125

[2]Division of Biology, California Institute of Technology, Pasadena, CA, 91125

Abstract

Genome-wide measurements of protein-DNA interactions and transcriptomes are increasingly done by deep DNA sequencing methods (ChIP-seq and RNA-seq). The power and richness of these counting-based measurements comes at the cost of routinely handling tens to hundreds of millions of reads. While early-adopters necessarily developed their own custom computer code to analyze the first ChIP-seq and RNA-seq datasets, a new generation of more sophisticated algorithms and software tools are emerging to assist in the analysis phase of these projects. This review describes the multilayered analyses of ChIP-seq and RNA-seq datasets, discusses the software packages currently available to perform tasks at each layer, and describes some upcoming challenges and features for future analysis tools. We also discuss how software choices and uses are affected by specific aspects of the underlying biology and data structure, including genome size, positional clustering of transcription factor binding sites, transcript discovery, and expression quantification.

Introduction

A longstanding goal for regulatory biology is to learn how genomes encode the diverse patterns of gene expression that define each cell type and state. Genome-wide measurements of protein-DNA interaction by chromatin immunoprecipitation (ChIP) and quantitative measurements of transcriptomes are increasingly used to link regulatory inputs with transcriptional outputs. Such measurements figure prominently, for example, in efforts to identify all functional elements of our genomes, which is the raison d'être of the ENCODE project consortium[1]. Although large-scale ChIP and transcriptome studies first used microarrays, deep DNA sequencing versions (ChIP-seq and RNA-seq) offer distinct advantages in increased specificity, sensitivity and genome-wide comprehensiveness that are leading to their wider use[2].

The overall flavor and objectives of ChIP-seq and RNA-seq data analysis are similar to those of the corresponding microarray-based methods, but the particulars are quite different. These data-types therefore require new algorithms and software that are the focus of this piece. We view the data analysis for ChIP-seq and RNA-seq as a bottom-up process that begins with mapped sequence reads and proceeds upward to produce increasingly abstracted

[3]Corresponding author. alim@caltech.edu.

layers of information (Fig. 1). The first step is to map the sequence reads to a reference genome and/or transcriptome sequence. It is no small task to optimally align tens or even hundreds of millions of sequences to multiple gigabases for the typical mammalian genome[3], and this early step remains one of the most computationally intensive in the entire process. Once mapping is completed, users typically display the resulting population of mapped reads on a genome browser. This can provide some highly informative impressions of results at individual loci. However these browser-driven analyses are necessarily anecdotal and, at best, semi-quantitative. They cannot quantify binding or transcription events across the entire genome nor find global patterns.

Considerable additional data processing and analysis are needed to extract and evaluate the genome-wide information biologists actually want. While there are now multiple algorithms and software tools to perform each of the possible analysis steps (Fig 1), this is still a rapidly developing bioinformatics field. Our purpose here is to give a sense of the tasks to done at each layer, coupled with a reasonably current summary of tools available. We explicitly do not attempt any software “bake-off” comparisons, aiming instead to provide information to help biologists to match their analysis path and software tools to the aims and data of a particular study. Finally, we try to focus attention on some pertinent interactions between the molecular biology of the assays, the information-processing methods, and underlying genome biology.

General features of ChIP-seq

The success of genome-scale chromatin immunoprecipitation experiments depends critically on 1) achieving sufficient enrichment of factor-bound chromatin relative to nonspecific chromatin background, and 2) obtaining sufficient enriched chromatin so that each sequence obtained is from a different founder molecule in the ChIP reaction (i.e. that the molecular library has adequate sequence complexity). When these criteria are met, successful ChIP-seq datasets typically consist of 2-20 million mapped reads. In addition to the degree of success of the immunoprecipitation, the number of occupied sites in the genome, the size of the enriched regions, and the range of ChIP signal intensities all affect the read number wanted. These parameters are often not fully known in advance, which means that computational analysis for a given experiment is usually performed iteratively and repeatedly, with results dictating whether additional sequencing is needed and cost-effective. This means that the choice of software for running ChIP-seq analysis favors packages that are simple to use repeatedly with multiple datasets.

Mapped reads are immediately converted to an integer count of “tags” at each position in the genome that is “mappable” under the mapping algorithm selected and its parameters (i.e. read length can be fixed or variable; reads mapped can be restricted to those that map to a unique position in the genome or can include “multireads” that map to multiple sites). These early choices in the analysis affect sensitivity and specificity, and their effects vary based on the specifics of each genome. If only uniquely mapping reads are used, some true sites of occupancy will be invisible, because they are located in repeats or recent duplicated regions. Conversely, allocating low-multiplicity multireads will capture and improve some true signals, but will also likely create some false positives. The choice of mapping algorithm

can thus be made with eye toward increasing specificity (unique reads only) or increasing sensitivity (multireads used).

It is relevant to data processing and interpretation that ChIP reactions are enrichments, not purifications. This is especially true for current protocols that use a single antibody reagent, because the majority (~60-99%) of DNA fragments (and therefore of sequence reads) in a ChIP reaction are background, while the minority corresponds to DNA fragments to which the transcription factor or histone mark of interest was crosslinked at the beginning of the experiment. These substantial levels of impurity are expected for a one-round enrichment, and discriminating background sequence reads from true signal must be dealt with in the analysis phase. “Background” read distributions will be different depending on the composition and size of the genome. In ChIP-seq datasets from larger mammalian genomes, most nucleotides have no mapped tags since the overall mapped sequence coverage is much less than the total genome size (i.e. less than 0.1X coverage). In smaller genomes such as *Drosophila* or *C. elegans*, a typical ChIP-seq assay performed at similar 2-20M read depths will place read-tags over most of the genome at increasing densities (roughly 1X-10X coverage), and ChIP-positive signals will be compressed along the chromosome, since there is much less intergenic space per gene in the smaller genomes.

The strongest ChIP-enriched positions can have hundreds of overlapping reads for DNA binding factors that are highly efficient targets for ChIP. These strongest signals are not, however, the only biologically meaningful ones. Statistically robust and reproducible ChIP signals that have modest read counts (in absolute terms and by comparison with empirically determined background read distributions) have been observed for locations known to have high biological regulatory activity by independent criteria[4]. This means that a key challenge for ChIP-seq algorithms is to identify reproducibly true binding locations while including as few false positives from the background as possible. The background distribution of reads in ChIP-seq is often determined empirically from a control reaction, but some algorithms model the background from the ChIP sample itself. Whichever approach is taken, the background read-tag distribution is not reliably uniform, nor is it identical for all cell types and tissues of the same organism. It is also not expected to be identical from one specific ChIP protocol to another. Various artifacts can cause different chromosomal areas to be systematically underrepresented (extremes of base composition that affect library making and or sequencing itself, for instance) or over-represented (sites of preferential chromosomal breakage in the cell or during the workup). The current algorithms have each been designed to ignore a variety of false positive read-tag aggregations that are judged unlikely to be due to immuno-enriched factor binding, but they are not identical to each other and users should expect different packages and different parameters to eliminate as background some overlapping and some novel.

Classes of ChIP-seq signals

Consistent with previous ChIP-chip results, ChIP-seq tag enrichments or “peaks” generated by typical experimental protocols, can be usefully classified into three major categories: punctate regions covering a few hundred base pairs or less; localized but broader regions of up to a few kilobases; and broad regions up to several hundred kilobases. Punctate

enrichment is a signature of classic DNA-sequence specific binding of transcription factors such as NRSF or CTCF to an exact source such as their cognate motif (Fig. 2a). A mixture of punctate and broader signals is associated with proteins such as RNA Polymerase II that bind strongly to specific transcription start sites in active and stalled promoters (in punctate fashion), but RNApol2 can also be detected more diffusely over the body of actively transcribed genes[5-6] (Fig. 2b). ChIP-seq signals that come from most histone marks and other chromatin domain signatures are not point sources as described above but range from nucleosome-sized domains to very broad enriched regions that lack a single source entirely such as H3K27 trimethylation in repressed areas[7-8] (Fig. 2c).

These different categories of ChIP enrichment have distinct characteristics that algorithms can use to predict true signals optimally. Punctate events offer the greatest amount of discriminatory detail to model the source point down to the nucleotide level. To date, most algorithms have been developed and tuned for this class of binding, though specific packages can work reasonably well for mixed binding, typically requiring use of non-default parameters.

Peak-finders, regions, summits, and sources

The first step in analyzing ChIP-seq data is to identify regions of increased sequence read tag-density along the chromosome relative to measured or estimated background. After these “regions” are identified, further processing ensues to identify the most likely source-point(s) of cross-linking and inferred binding (called “sources”). The source is related, but not identical to the “summit”, which is the local maximum read density in each region When there is no single point source of cross-linking, as for some dispersed chromatin marks, the region-aggregation step is appropriate, but the “summit finding” step is not. Software packages for ChIP-seq are generically and somewhat vaguely called “peak-finders”. They can be conceptually subdivided into following fundamental components: (1) a signal profile definition for a ChIP region, (2) a background model, (3) peak call criteria, (4) post call filtering of artifactual peaks, and (5) significance ranking of called peaks. (Fig. 3). A summary of the components for twelve published software packages is given in Table 1.

The simplest approach for calling enriched regions in ChIP-seq data is to take a direct census of mapped tag sites along the genome and allow every contiguous set of base pairs with more than a threshold number of tags covering them to define an enriched sequence region. While this can be effective for highly defined point source factors with strong ChIP enrichment, it is not satisfactory overall due to inherent complexities of the signals as well as experimental noise and/or artifacts. Additional information present in the data is now used to help discriminate true positive signals from various artifacts. For example, the strand-specific structure of the tag distribution is useful to discriminate the punctate class of binding events from a variety of artifacts[9]. Because immunoprecipitated DNA fragments are typically sequenced as single-ended reads, i.e. from one of the two strands in the 5′ to 3′ direction, the tags are expected to come on average equally frequently from each strand, thus giving rise to 2 related distributions of stranded reads. The corresponding individual strand distributions will occur upstream and downstream, shifted from the source point (“summit”) by half-the average sequenced fragment length, which is typically referred to as the “shift”

(Fig. 4a). Note that the average observed fragment length can differ considerably from the "expected" fragment length derived from agarose gel cuts made during Illumina library preparation; short fragments are further favored by Illumina's solid-state PCR. For this reason, the shift is now mainly determined computationally from the data, rather than imposed from the molecular biology protocol. The shift will be smaller and the two strand distributions will come closer together in experiments in which the fragment length, read-length and recognition site length converge.

Building a signal profile

The signal profile is a smoothing of the tag counts to allow reliable region identification and better summit resolution. The simplest way to define a signal profile is to slide a window of fixed width across the genome, replacing the tag count at each site with the summed value within the window centered at the site. Consecutive windows exceeding a threshold value are merged. This is what cisGenome[10] does. SiSSRs[11], and spp[12] count tags within a window in a strand-specific fashion. Other programs also using sliding window scans, but compute various modified signal values. The program MACS[13] performs a window scan, but only after shifting the tag data in a strand specific fashion to account for the fragment length. F-Seq[14] performs kernel density estimation (KDE) with a Gaussian kernel. QuEST[9] creates separate KDE profiles for the two strands. SICER[15] computes probability scores in non-overlapping windows, then aggregates windows into "islands" of sub-threshold windows separated by gaps in order to capture broad enrichment regions. An alternate approach is to extend the ChIP-seq tags along their strand direction (called an 'XSET') and to count overlaps above a threshold as peak regions[16]. Tag extension prior to signal calculation serves the dual purpose of correcting for the assumed fragment length and also smoothing over gaps that were not tagged due to low sampling or read mappability issues. GLITR[17] uses this algorithm. PeakSeq[5] combines tag extensions with tag aggregation. ERANGE[4,18] aggregates tags within a fixed distance of one another into candidate peak regions.

Strand-specific read shifting can yield significantly improved summit resolution as well as greater sensitivity for punctate source calls, if the shift distance is accurate. If the shift is badly misestimated, some true ChIP sites will not be called. Experiments with longer average fragment lengths benefit more from read shifting because the effect is greater. The read-shift distance used is generally either fixed to a user-specified value or it is estimated from ChIP data; generally the latter is based upon high quality peaks only (those with very large enrichment relative to background). MACS, QuEST, SiSSRs, and spp perform mandatory tag shifting prior to generating a set of peak calls. ERANGE and FindPeaks[19] offer it as an option, while cisGenome shifts tags only as a post-processing step to refine binding site locations. F-Seq, GLITR, and SICER shift tags by a user-specified distance. Tag extension can accomplish the same goals as tag shifting in many cases.

Handling the background

The background model consists of an assumed statistical noise distribution or a set of assumptions that guide the use of control data to filter out certain types of false positives in the treatment data. In the absence of control data, the background tag distribution is typically

modeled with a Poisson or negative binomial distribution. When control data is available, it may be used to determine parameters for these distributions. Alternatively, the control data may be subtracted from the signal along the genome or the signal may be thresholded by its enrichment ratio relative to the control. Using experimental control data is thought important, because it significantly reduces false positive regions that come from DNA shearing biases or sequencing artifacts. CisGenome, ERANGE, GLITR, MACS, PeakSeq, QuEST, SICER, SiSSRs, spp, and USeq[20] all use control data when it is available. FindPeaks,F-Seq, XSET, and the approach of Mikkelsen et al.[8] do not.

Peak call criteria

Once the signal profile has been generated and tags allocated to regions, those for which the signal satisfies certain quality criteria are considered candidate peaks. The main quality criterion is either an absolute signal threshold or a minimum enrichment relative to the background or both. Specifics for various software implementations are given in Table 1. Default values for these are provided, but users will need to consider whether their data is similar enough to those on which a specific algorithm was tuned to justify using the defaults. Some exploration of the parameter space may be helpful. Ideally, an end user would specify a desired FDR, with parameters then set to achieve it for a given algorithm and dataset. A few packages implement some version of this (see significance ranking below), but there is no consensus yet on how to best estimate the FDR for ChIP-seq, and different methods produce different outcomes. This is discussed further in the context of significance ranking.

Post-filtering

After the initial peak calling step, simple filters are optionally available to eliminate artifacts. Two popular filtering criteria are based on the distributions of tags between the DNA strands (directionality) and single site duplicates. Directionality criteria include: fraction of plus and minus tags, fraction of plus(minus) tags occurring to the left(right) of the putative peak, and the presence of a partnered plus(minus) peak for each minus(plus) peak. Note that default values for the directionality filtering may be too stringent if data is noisier than was seen in the first generation of experiments used to develop the algorithms. Also, this filter may incorrectly reject complex peak regions, i.e. those that contain more than one summit. QuEST, FindPeaks, and PeakSeq attempt to subdivide regions into more than one summit call (multiple overlapping sources), however this remains an active area of research. Duplicate filters are relatively straightforward and eliminate tags at single sites that exhibit counts much greater than that expected by chance.

Significance ranking

Called peak regions encompass a wide range of quantitative enrichments, thus an assessment of the relative confidence one should place in a given set of peaks or, if possible, each individual peak is informative. Most of the algorithms currently compute p-values either after the fact or as part of the peak calling procedure and these are provided with the output peak list. As seen in Table 1, the packages that provide p and/or q values are: CisGenome, ERANGE, GLITR, MACS, PeakSeq, SICER, spp, and USeq. A few callers do not provide p-values, in which case the use of the peak height or fold-enrichment may be used to provide a peak ranking, though not statistical significance. From an end user perspective, the false

discovery rate is often of paramount interest and one can compute a p-value from a false discovery rate or vice versa for a known distribution. Generally, however it is not known *a priori* whether the distribution assumption made in calculating the p-value is appropriate, thus the correct false discovery rate may be far different from the one based on the p-value threshold. Therefore some programs (ERANGE, MACS, QuEST, spp, USeq) instead compute an empirical FDR by calling peaks in a portion or all of the control data. The FDR in this case is given as the ratio of the number of peaks called in the control to the number of peaks called for the ChIP data.

Specialized software to analyze histone modification ChIP-seq data that start to address higher-level analyses include ChIPDiff[21] and ChromaSig[22]. ChIPDiff uses an HMM to assess the differences in the histone modifications from the ChIP-seq signal between two libraries, for example from different cell types. ChromaSig performs unsupervised learning on ChIP-seq signals across multiple experiments to determine statistically significant patterns of chromatin modifications.

Further subtleties in the ChIP-seq signal present challenges for both computation and interpretation of downstream results. Some ChIP-seq peak regions are spatial or temporal convolutions of multiple biologically true sources. In such cases, the highest density of reads does not always correspond to a source point (Fig. 4b). This complexity can be magnified as one moves from relatively large mammalian genomes with long stretches of intervening DNA isolating regulatory modules from each other, to smaller genomes with potentially higher densities of binding sites compressed in complicated modules. Computationally, this turns the problem from one of peak identification to peak deconvolution. In regions where this occurs the signal to noise characteristics usually determine whether it is feasible to discriminate occupancy among the different individual sites. In the temporal case, a transcription factor binding site that is bound in an undifferentiated cell type, for instance, and not bound in a differentiated cell type, will be diluted relative to sites that are bound in both states whenever the starting cell population is of mixture of the two cell type. In an embryo or whole organism, a given factor may bind partly or entirely non-overlapping regulatory modules, thus mixing signals that would otherwise be spatially and/or temporally distinct in defined cell subpopulations.

Last but not least, the stochastic sampling of the DNA fragments means that, as more sequencing is done in a given sample, additional weak but potentially significant signals will continue to be discovered. How many of these are functionally important is not *a priori* clear, without explicit testing. This uncertainty will affect how these weaker features are used (or eliminated) for input into higher-level integrative analysis. Although weak sites can be confirmed by different readouts from ChIP (QPCR; ChIP-chip), supported by *in vitro* binding to the sequence, and by computational presence of binding motifs in the DNA, utterly independent evidence of occupancy, such as that provided by *in vivo* foot-printing or site-mutation in transfection assays, has yet to be marshaled for a convincingly large sample of such “cryptic” ChIP-positive sites. What is certain, however, is that the complexity of the ChIP library (how many different founder DNA fragments are captured for sequencing) and the depth of sequencing must be properly adjusted to match the experimental goal and the underlying biology. Thus chromatin marks that cover large areas of the genome call for

deeper sequencing or for additional algorithmic inferences to define large signal domains, compared with point source binding.

Transcriptome analysis of RNA-seq data

Transcriptome analysis has multiple functions, broadly divided between transcript discovery and mapping on one hand and RNA quantification on the other. The software sub-tasks needed for analysis depend on which of those two aspects are paramount in a given study. The first generation of RNA-seq studies published in 2008[17,23-28] used very short, unpaired reads (25-32 NT) of cDNA made by reverse-transcription of poly-A selected RNA (Fig. 5). As longer read-lengths and larger numbers of reads have become routine in some platforms, and as “mate-paired or paired-end” format have been added, the bioinformatics tools are evolving to handle the changing data. Experimental protocol choices also affect the downstream data analysis. For example, RNA fragmentation and size selection steps of 200bp fragment in current RNA-seq protocols will likely result in under-representation of the shortest transcripts, as has already been noted[29-30]. Given keen interest in RNA-seq, it is natural that platform vendors such as ABI and Illumina, and commercial software ventures, are beginning to provide commercial packages, but we limit this overview to publicly available packages connected to published papers (Table 2).

For a subset of RNA-seq users who work on organisms without a reference genome sequence or aim to detect chimeric transcripts from chromosomal rearrangements such as those found in tumors, analyzing the transcriptome involves assembling ESTs *de novo* using short-read assembly programs such as Velvet[31], which assemble sequences by assembling reads that overlap by a pre-selected k-mer, i.e. by a minimum number of bases. Typically, a finite range of k-mers are tried to find the optimal k-mer that will give the best assembly in terms of both number and sizes of contigs/ESTs. Since short read assemblers are primarily designed to assemble genomic sequence with relatively even depth of coverage, the five orders of magnitude of prevalence in transcriptomes represent a difficult challenge[32]. A recent study using ABySS[33] assembled 764,365 ESTs from 194 million 36 bp reads from a human follicular lymphoma transcriptome with k=28 bp; half of the 30 Mb of unique sequence is found on contigs larger than 1.1 kb. At lower sequencing depths, *de novo* assembly will work best for genes that are highly expressed enough to be tiled by reads that overlap at the selected k-mer (Fig. 6a).

Mapping splices and multireads

For all other RNA-seq analyses with 10-100 million reads and where a reference genome is known, the reads can be mapped as in ChIP-seq, but with the added opportunity to map reads that cross splice junctions (Fig. 6b-c). Known splice junctions, based on gene models and ESTs can be handled by incorporating them informatically in the primary read-mapping, while newly inferred junctions are considered later. Once the reads are mapped, the question of their correspondence with gene and transcript models arises, since it is common to have more than one transcript type from a single gene, with alternate splicing, alternate promoter use and different 3′ poly(A) addition sites all contributing diversity. More sophisticated questions follow concerning the respective prevalences of each transcript isoform, and the relative prevalence of RNAs within a given transcriptome. A final goal in a majority of

transcriptome studies is to quantify differences in expression across multiple samples in order to capture differential gene expression.

The main challenges of mapping RNA-seq reads center around the handling of splice junctions, paralogous gene families and pseudogenes. Nearly all RNA-seq packages are built on top of short read mappers such as bowtie[34] and SOAP[35] – and may require multiple runs to map splice-crossing reads. The primary approach is to simply map the ungapped sequence reads across sequences representing known splice junctions, which can also be supplemented with any set of predicted splice junctions from spliced ESTs or gene finder predictions as implemented by ERANGE or RNA-MATE[36]. However, all of these approaches are ultimately limited to recovering previously documented splices. Alternatively, packages such as TopHat[32] and G-Mo.R-Se[37] first identify enriched regions representing transcribed fragments (transfrags) and build candidate exon-exon splice junctions to map additional reads across, whereas QPALMA[38] attempts to predict whether a read is spliced as part of the mapping process.

Multireads, i.e. reads that map equally well to multiple genomic locations, arise predominantly from conserved domains of paralogous gene families and repeats. Another confounding problem is the prevalence SINEs and LINEs in the UTRs of genes as well as the abundance of retroposed pseudogenes for highly expressed housekeeping genes in large genomes. Both of these vary from one genome to the next[39]. For example, several GAPDH retroposed pseudogenes in the mouse genome differ by less than 2 nucleotides (0.2%) from the mRNA for GAPDH itself, making it difficult to map reads correctly to the originating locus based on RNA-seq data alone. Orthogonal data such as RNApol2 occupancy and ChIP-seq measurements can later be brought to bear in some cases, but different software and use-parameters make starting choices based on the RNA data alone. While the algorithms are generally sensible, specific cases can be insidious, and are worth being aware of. For example, a minority of reads from one paralog can map best to other sites (usually another paralog or pseudogene) because of the error rate in sequencing, which is quite substantial on current platforms (typically around 1%). For highly expressed genes, this can cause a shadow of expression at these pseudogenes – which may then be called as transfrags. Similarly, reads that are intron-spanning from a source gene may map instead perfectly and uniquely to a retroposed pseudogene. The ERANGE package avoids such mis-assignment by mapping reads simultaneously across the genome and splice junctions, thus turning them into multireads that are subsequently handled separately.

Assigning reads to known and new gene models

The next level of RNA-seq analysis associates mapped reads with known or novel gene models. Given a set of annotations, all tools can tally the reads that fall on known gene models, and several tools like RSAT[40] and BASIS[41] deal primarily with the annotated models. However, a substantial fraction of reads fall outside of the annotated exons, above the “noise” level generated by mismapped reads or intronic RNA from incompletely spliced hnRNA. In mouse and human samples, we have especially noticed that prominent read densities often extend well beyond the annotated 3’UTRs, or as alternatively spliced 5’UTR, internal exons, or retained introns. ERANGE, G-Mo.R-Se, and TopHat first aggregate reads

into transfrags. Whereas G-Mo.R-Se and TopHat rely primarily on spliced reads to connect transfrags together, ERANGE uses two different strategies depending on the availability of paired-reads. In currently conventional unpaired sequence read case, ERANGE assigns transfrags to genes based on an arbitrary user-selected radius, whereas in the paired-end case, it will bring together transfrags only when they are connected by at least one paired-reads. Both strategies work much better with data that preserve RNA strandedness.

Quantifying gene expression

Given a gene model and mapped reads, one can sum the read counts for that gene as one measure of the expression level of that gene at that sequencing depth. However, the number of reads from a gene is naturally a function of the length of the mRNA as well as its molar concentration. A simple solution that preserves molarity is to normalize the read-count by the length of the mRNA and the number of million mappable reads to obtain Reads Per Kb per Million (RPKM) values[18]. RPKMs for genes are then directly comparable within the sample by providing a relative ranking of expression. While straightforward, RPKM values have several substantive detail differences between software packages, and there are also some caveats in using them. Whereas ERANGE uses a union of known and novel exon models to aggregate reads and determine an RPKM value for the locus, TopHat and RSAT restrict themselves to known or pre-specified exons. ERANGE will also include spliced reads and can include assigned multireads in its RPKM calculation, whereas other packages limit themselves to uniquely mappable reads.

Several experimental issues influence the RPKM quantification, including the integrity of the input RNA, the extent of ribosomal RNA remaining in the sample, size selection steps, and the accuracy of the gene models used. RPKMs reflect the true RNA concentration best when samples have relatively uniform sequence coverage across the entire gene model, which is usually approached by using random priming or RNA-ligation protocols, although both currently fall short of desired uniformity. Poly(A) priming has different biases (3′) from partial extension or when there is partial RNA degradation. Resulting ambiguities in RPKMs from an RNA-seq experiment are akin to microarray intensities that need to be further post-processed before comparison to other RNA-seq samples using any number of well-documented normalization methods such as variance stabilization[42], for example.

More sophisticated analyses of RNA-seq data allow users to extract additional information from the data. One area of considerable interest and activity is in transcript modeling and quantifying specific isoforms. BASIS calculates transcript levels from coverage of known exons by taking advantage of specifically informative nucleotides from each transcript isoform. A second area is sequence variation. The RNA sequences themselves can be mined to identify positions where the base reported differs from the reference genome(s), identifying either a single-nucleotide polymorphism or a private mutation[25,43]. When these are heterozygous and phased or informatively related to the source genome, RNA SNPs can be used to detect allele-specific gene expression. Yet another source of observed sequence differences between the transcriptome and genome are changes due to RNA-editing[44]. In general, bioinformatics tools are evolving to match changes in sequencing technology. Longer and more informative reads produce a higher fraction of uniquely mappable reads

that cross one or more splice junctions, which calls for changes in transcript mapping and assembly. Paired reads with good control over insert size distribution (i.e. tight size distributions) will provide a superior substrate for determining long-range isoform structure and quantifying them. We also expect that strand-reporting protocols[45] will be more widely used and that they will help to disambiguate instances where both strands are represented or where the strand of origin is entirely unknown.

Future opportunities and challenges

A virtue of sequence-based RNA and ChIP datasets is that the raw unmapped reads can be re-analyzed to gain the benefits of ongoing algorithmic improvements and updated genome references and gene models, including SNP anotations and, eventually, source DNA sequences from the same individuals or cell lines used for RNA and ChIP experiments. Beyond these incremental changes, major improvements are anticipated for both ChIP-seq and RNA-seq that will require substantial algorithmic advances. Variations on chromatin conformation capture (3C)[46] and their combination with ChIP-seq in genome-wide formats promise to provide physical linkages between distal (even transchromosmal) regulatory elements and the genes that they regulate[47]. They call for new algorithms and software to find, cull, quantify and ultimately integrate longer-range physical interactions in the nucleus with the kind of occupancy and chromatin state information now being gathered. The current forms of RNA-seq will likely transition to a more quantitative form of “universal” RNA-seq that captures short and long RNAs while preserving strand origin without poly(A) selection[48]. Whereas ChIP-seq is less likely to benefit from the substantially longer reads promised by the upcoming generation of DNA sequencers, these will be invaluable to RNA-seq as most transcripts will be unambiguously sequenced as a single “read”.

Growth of publicly available ChIP-seq and RNA-seq datasets will increasingly drive integrated computational analysis that aims to address basic questions about how the chemical code of *in vivo* DNA binding for multiple factors relates to transcription output. ChIP-seq experiments, just as ChIP-chip experiments before them, reveal thousands of reproducible binding events that do not follow the simplest possible logic of a predictable positive or negative effect on the nearest promoter. What is the logic? How can functionally important sites of occupancy can be discerned computationally and discriminated from others that are inactive or differently active sites? Computational integration of factor binding, histone marks, polymerase loading, methylation and other genome-wide data will be pursued to learn if highly combinatorial models of inputs can predict regulatory output. Finally, further integrative analyses that draw on data from RNAi perturbations and high-throughput functional element assays will likely be needed to extract functionally the important connections and relationships of a working regulatory code.

Acknowledgments

This work was supported by The Beckman Foundation, The Beckman Institute, The Simons Foundation and US National Institutes of Health (NIH) grant U54 HG004576 to B.W., Fellowships from the Gordon and Betty Moore Foundation, Caltech’s Center for the Integrative Study of Cell Regulation, and the Beckman Institute to A.M, and support from the Gordon and Betty Moore foundation to SP. The authors would like to especially thank G. Marinov, and P. Sternberg for many helpful discussions of this manuscript.

References

1. ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447(7146):799–816. [PubMed: 17571346]
2. Wold B, Myers RM. Sequence census methods for functional genomics. Nat Methods. 2008; 5(1): 19–21. [PubMed: 18165803]
3. Trapnell C, Salzberg SL. How to map billions of short reads onto genomes. Nat Biotechnol. 2009; 27(5):455–7. [PubMed: 19430453]
4. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide Mapping of in Vivo Protein-DNA Interactions. Science. 2007; 316:1497–1502. [PubMed: 17540862]
5. Rozowsky J, Euskirchen G, Auerbach RK, Zhang ZD, Gibson T, Bjornson R, Carriero N, Snyder M, Gerstein MB. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nature Biotechnology. 2009; 27(1):66–75.
6. Baugh LR, Demodena J, Sternberg PW. RNA Pol II accumulates at promoters of growth genes during developmental arrest. Science. 2009; 324(5923):92–4. [PubMed: 19251593]
7. Barski A, Cuddapah S, Cui K, Roh TY, SChones DE, Wang Z, Wei G, Chepelev I, Zhao K. High-resolution profiling on histone methylations in the human genome. Cell. 2007; 129(4):823–37. [PubMed: 17512414]
8. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim T, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE. Genome-wide maps of chromatin state in pluripotent and linearge-committed cells. Nature. 2007; 448:553–560. [PubMed: 17603471]
9. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow. A. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nature Methods. 2008; 5(9):829–834. [PubMed: 19160518]
10. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nature Biotech. 2008; 26(11):1293–1300.
11. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Research. 2008; 36(16):5221–5231. [PubMed: 18684996]
12. Kharchenko PV, Tolstorukov MY, Park PJ. Design and anlysis of ChIP-seq experiments for DNA-binding proteins. Nature Biotech. 2008; 26(12):1351–1359.
13. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). Genome Biology. 2008; 9:R137.1–.9. [PubMed: 18798982]
14. Boyle AP, Guinney J, Crawford GE, Furey TS. F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics. 2008; 24(21):2537–2538. [PubMed: 18784119]
15. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics. 2009; 25(15):1952–8. [PubMed: 19505939]
16. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature Methods. 2007; 4(8):651–657. [PubMed: 17558387]
17. Tuteja G, White P, Schug J, Kaestner KH. Extracting transcription factor targets from ChIP-Seq data. Nucleic Acids Research. Advance Access published June 24, 2009.
18. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. Nature Methods. 2008; 5(7):621–8. [PubMed: 18516045]
19. Fejes AP, Robertson G, Bilenky M, Varhol R, Bainbridge M, Jones SJM. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. Bioinformatics. 2008; 24(15):1729–1730. [PubMed: 18599518]
20. Nix DA, Courdy SJ, Boucher KM. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. BMC Bioinformatics. 2008; 9:523. [PubMed: 19061503]

21. Xu H, Wei C, Lin F, Sung WK. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. Bioinformatics. 2008; 24(20):2344–2349. [PubMed: 18667444]
22. Hon G, Ren B, Wang W. ChromaSig: A probabilistic approach to finding common chromatin signatures in the human genome. PLoS Computational Biology. 2008; 4(10)
23. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008; 320(5881):1344–9. [PubMed: 18451266]
24. Wihelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bähler J. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 2008; 453(7199):1239–43. [PubMed: 18488015]
25. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods. 2008; 5(7):613–9. [PubMed: 18516046]
26. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008; 18(9):1509–17. [PubMed: 18550803]
27. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science. 2008; 321(5891):956–60. [PubMed: 18599741]
28. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; 456(7221):470–6. [PubMed: 18978772]
29. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. Biology Direct. 2009; 4:14. [PubMed: 19371405]
30. Bullard JH, Purdom EA, Hansen KD, Durinck S, Dudoit S. Statistical inference in mRNA-Seq: exploratory data analysis and differential expression. UC Berkeley Division of Biostatistics Woring Paper Series. 2009; 247
31. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008; 18(5):821–9. [PubMed: 18349386]
32. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-seq. Bioinformatics. 2009; 25(9):1105–11. [PubMed: 19289445]
33. Birol I, Jackman SD, Nielsen C, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra MA, Jones SJ. De novo transcriptome assembly with ABySS. Bioinformatics. 2009
34. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10(3):R25. [PubMed: 19261174]
35. Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics. 2008; 24(5):713–4. [PubMed: 18227114]
36. Cloonan N, Xu Q, Faulkner GJ, Taylor DF, Tang DTP, Kolle G, Grimmond SN. RNA-MATE: A recursive mapping strategy for high-throughput RNA-sequencing data. Bioinformatics. 2009
37. Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C, Jaillon O, Artiguenave F. Annotating genomes with massive-scale RNA sequencing. Genome Biol. 2009; 9(12):R175. [PubMed: 19087247]
38. De Bona F, Ossowski S, Schneeberger K, Rätsch G. Optimal spliced alignments of short sequence reads. Bioinformatics. 2008; 24(16):i175–80.
39. Zhang Z, Carriero N, Gerstein M. Comparative analysis of processed pseudogenes in the mouse and human genomes. Trends Genet. 2004; 20(2):62–7. [PubMed: 14746985]
40. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-seq. Bioinformatics. 2009; 25(8):1026–32. [PubMed: 19244387]
41. Zheng S, Chen L. A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. Nucleic Acids Res. 2009; 37(10):e75. [PubMed: 19417075]

42. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential gene expression. Bioinformatics. 2002; 18(Suppl 1):S96–104. [PubMed: 12169536]
43. Chepelev I, Wei G, Tang Q, Zhao K. Detection of single nucleotide variations in expressed exons of the human genome using RNA-seq. Nucleic Acids Res. 2009 Published online: 15 June 2009.
44. Li JB, Levanon EY, Yoon JK, Aach J, Xie B, Leproust E, Zhang K, Gao Y, Church GM. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. Science. 2009; 324(5931):1210–3. [PubMed: 19478186]
45. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008; 133(3):523–36. [PubMed: 18423832]
46. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, Green RD, Dekker J. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. 2006; 16(10):1299–309. [PubMed: 16954542]
47. Fullwood MJ, Wei CL, Liu ET, Ruan Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genomes analyses. Genome Res. 2009; 19:521–532. [PubMed: 19339662]
48. Armour CD, Castle JC, Chen R, Babak T, Loerch P, Jackson S, Shah JK, Dey J, Rohl CA, Johnson JM, Raymond CK. Digital transcriptome profiling using selective priming for cDNA synthesis. 2009 Published online: 9 August 2009.

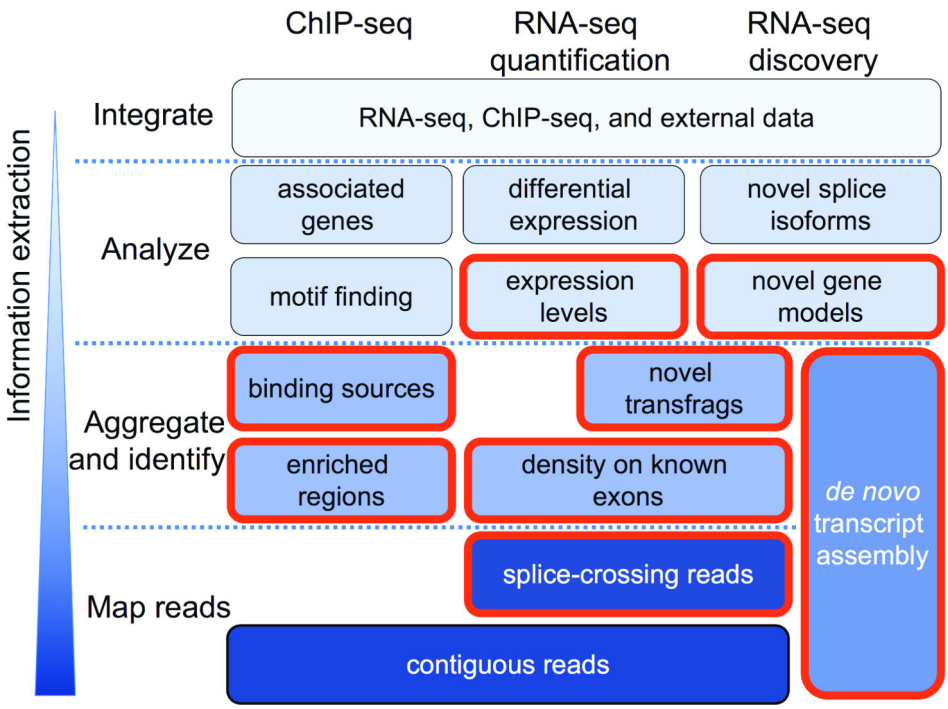

Figure 1. A hierachical overview of ChIP-seq and RNA-seq analyses

The bottom-up analysis of ChIP-seq and RNA-seq data typically involves the use of several software packages whose output serves as the input of the higher level analyses, with the subsections covered by this review circled in red. Apart from *de novo* transcript assembly for organisms without a reference genome, all sequence-counting packages build upon the output of read mappers onto a reference sequence, which serves as the input of programs that aggregate and identify these reads into enriched regions, density of known exons; many of these programs will further try to identify the sources (ChIP-seq) or novel RNA-seq transcribed fragments (transfrags). These regions and sources can then be analyzed to identify motifs, genes, or expression levels that are typically considered the biologically relevant output of these analyses. As the amount of RNA-seq and ChIP-seq data rapidly accumulates, the need for packages supporting integrative analyses is becoming increasingly pressing.

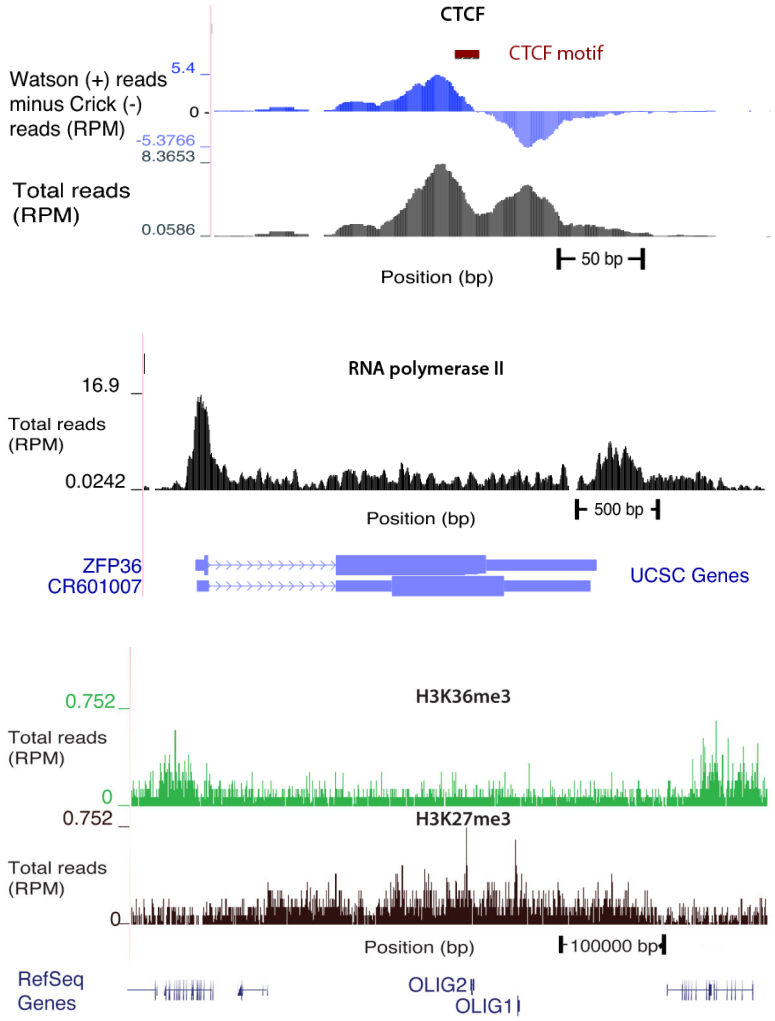

Figure 2. ChIP-seq Peak Types From Various Experiments and Peak Calling

Data shown in (**a-c**) are from remapping of a previously published human ChIP-seq dataset[7]. (**a**) Proteins that bind DNA in a site-specific fashion such as CTCF form narrow peaks 100's bp wide. The difference of plus and minus read counts is generally expected to cross zero near the signal source, the source in this example being the CTCF motif indicated in red. (**b**) Signal from enzymes such as RNA Polymerase II may show enrichment over regions up to a few kb in length. (**c**) Experiments that probe larger scale chromatin structure such as the repressive mark for H3K27me3 may yield very broad "above"-background regions spanning several 100 kb's.

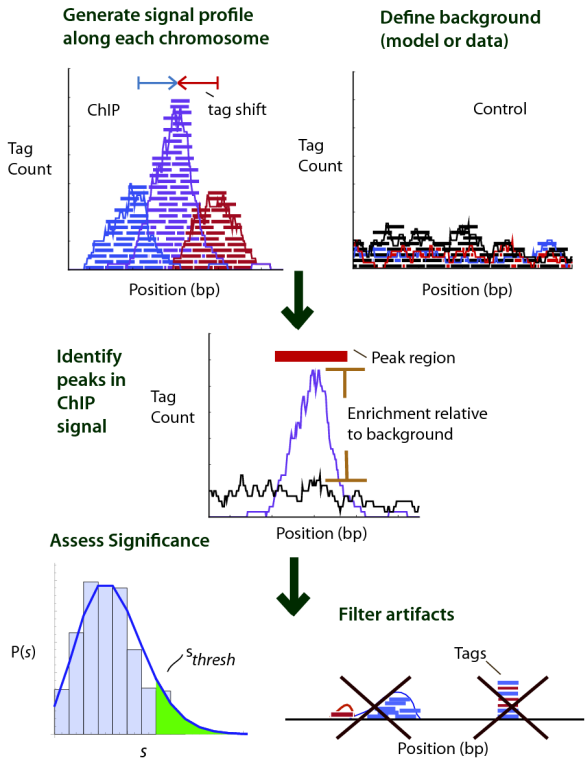

Figure 3. ChIP-seq Peak Calling sub-tasks

Sequence reads are first aligned to the genome. A signal profile that takes on a value at each bp is formed via a census algorithm, e.g. counting the number of reads overlapping each base pair along the genome (upper left plot). In the figure, blue represents '+' strand reads, red represents '−' strand reads, and purple shows the combined distribution after shifting the '+' and '−' reads toward the center by the read shift value. Further processing is sometimes performed prior to evaluating the signal (strand-specific tag shifting or smoothing for example). If experimental control data is available (brown), the same processing steps are applied to it to form a background profile (upper right plot); otherwise, a random genomic background may be assumed. The signal and background profiles are compared in order to define regions of enrichment. Finally, peaks are filtered to reduce false positives and ranked according to relative strength or statistical significance. In the lower left figure, P(*s*) refers to the probability of observing a location with *s* reads covering it. The bars represent the control data distribution. A hypothetical Poisson distribution fit is shown with sthresh indicating a cutoff above which a ChIP-seq peak might be considered significant. The lower right is a schematic representation of two types of artifactual peaks that may be filtered separately: single strand peaks and peaks formed by multiple occurrences of only one or a few reads.

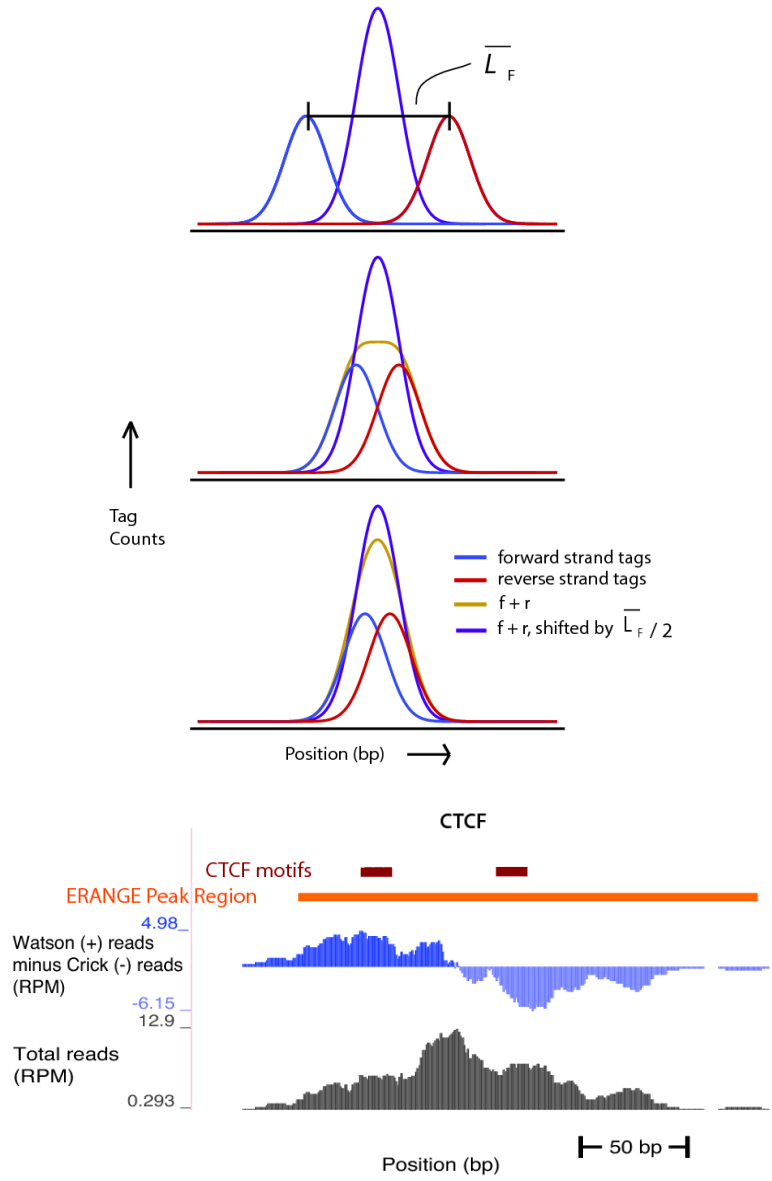

Figure 4. The impact of Fragment Length and Complex Peak Structures in ChIP-seq

(**a**) The average DNA fragment length can affect resolution with respect to binding site determination. A ChIP-seq experiment yields distributions for tags sequenced from the forward and reverse strands, the maxima of which should be separated by the average fragment length. In real experimental data, an overlap of the two distributions is often observed. If the average fragment length is much longer than the width of the strand distributions, the binding site will fall in between the two distributions. Tag locations are shifted toward the middle will result in a single summit (top illustration). Intermediate fragment lengths yield a single broadened peak in the unshifted aggregate distribution, and tag shifting may improve resolution a small amount by more precisely locating the summit (middle illustration). Very short fragments, such that the strand-specific densities are separated by a distance significantly less than the width of the individual distributions, can yield good binding site resolution without tag shifting tag. (**b**) Overlapping tag distributions

are observed for clusters of nearby peaks such as the pictured double for a CTCF peak region in human[7]. Motif mapping reveals two CTCF binding sites (in red), though ChIP-seq signal suggests a single binding site call lying between the two motifs. As an example, the ERANGE region call (orange) is shown to cover both motifs. The problem of reliably discriminating multiple binding sites with very closely overlapping signals is an ongoing area of research.

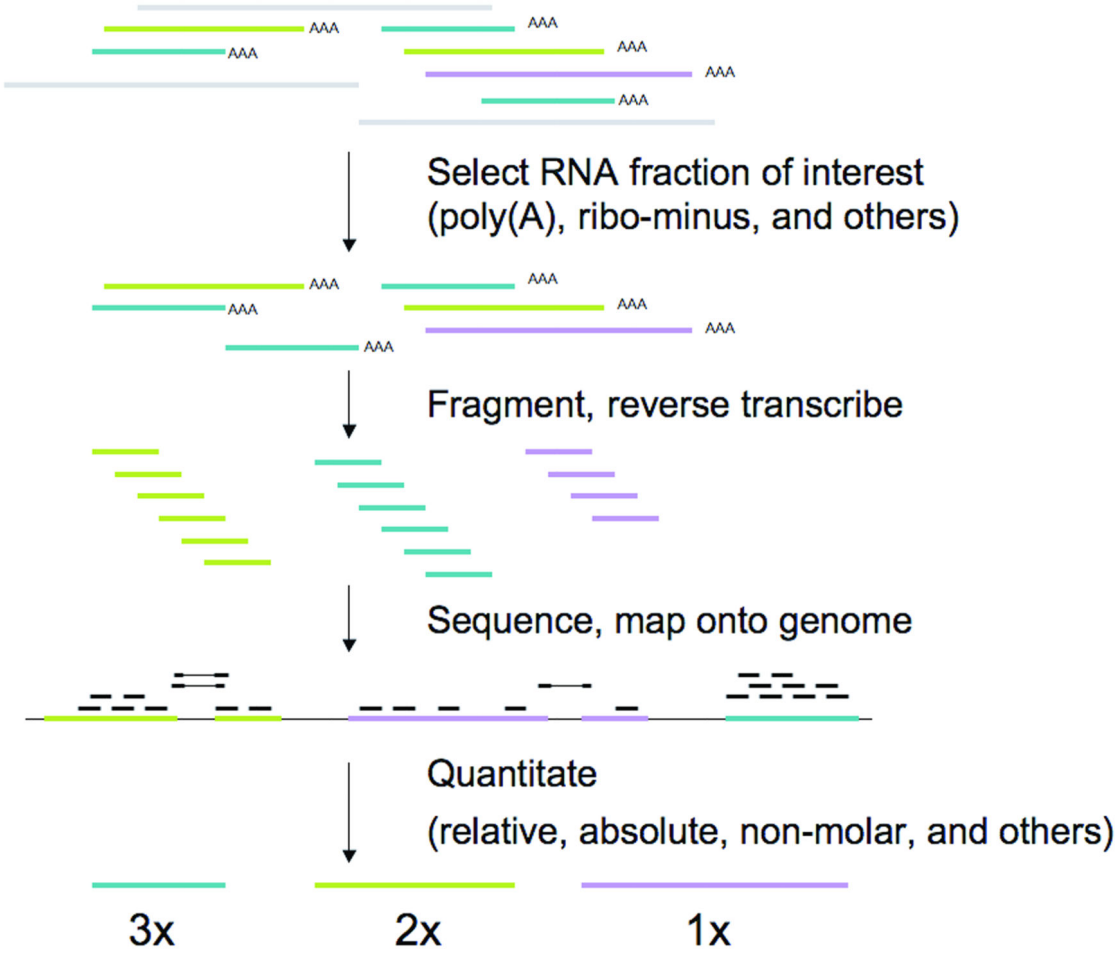

Figure 5. Overview of RNA-seq

A RNA fraction of interest is selected, fragmented, and reverse transcribed. The resulting cDNA can then be sequenced using any of the current ultra-high-throughput technologies to obtain ten to a hundred million reads, which are then mapped back onto the genome. The reads are then analyzed to calculate expression levels.

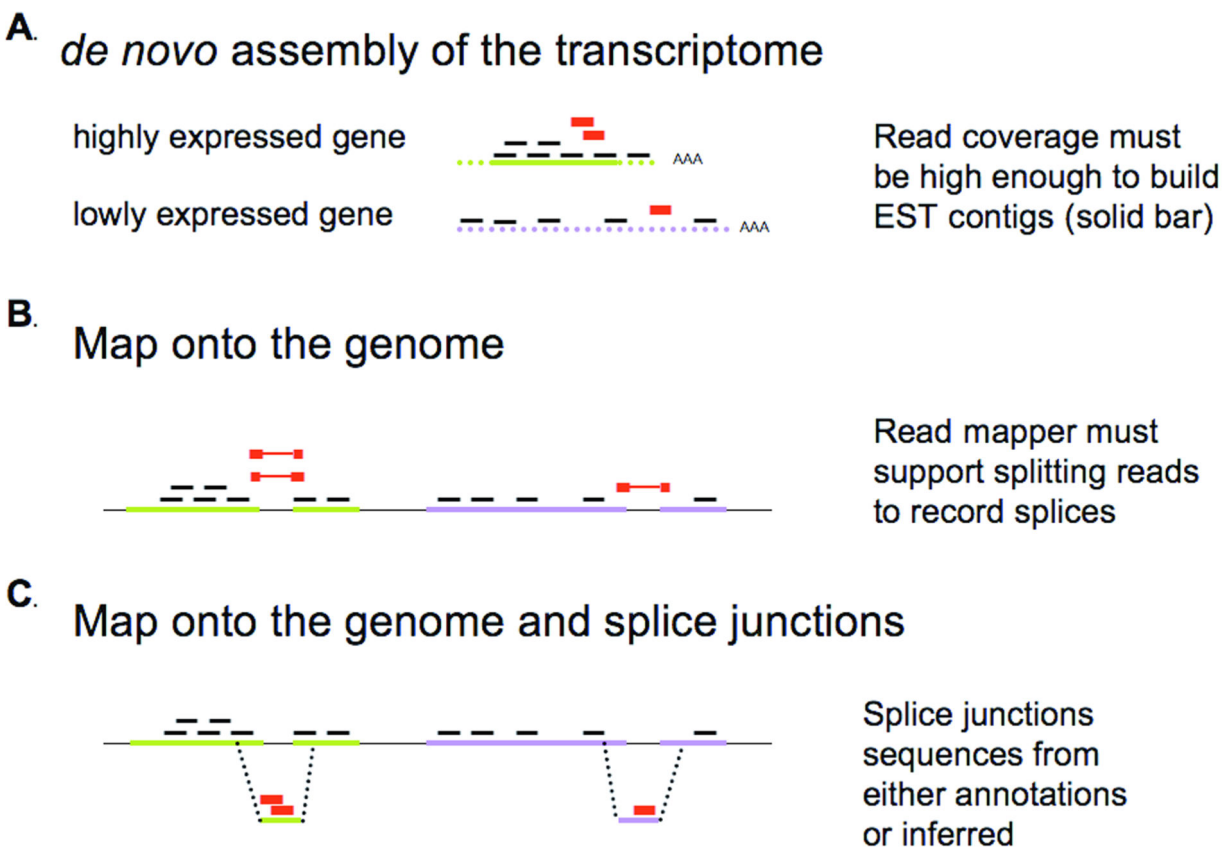

Figure 6. Approaches to handling of spliced reads

(**a**) In *de novo* transcriptome assembly, splice-crossing reads (red) are no different than any other reads, but will only contribute to a contig (solid green), when the reads are at high enough density to overlap by more than a set of user-defined assembly parameters. Parts of gene models (dotted green) or entire gene models (dotted magenta) can be missed if expressed at sub-threshold. (**b**) Splice crossing reads can be mapped directly onto the genome if the reads are long enough to make gapped-read mappers practical. (**c**) Alternatively, regular short read mappers can be used to map spliced reads ungapped onto supplied additional known or predicted splice junctions.

Table 1

Publicly available ChIP-seq software packages discussed in this review

	Profile	Peak Criteria[1]	Tag Shift	Control Data[2]	Rank By	FDR[3]	User Input Parameters[4]	Artifact Filtering: Strand-based / Duplicate[5]	Refer-ence
CisGenome v1.1	Strand-specific window scan	1: Number of reads in window, 2: Number of reads in window – control reads	Average for highest ranking peak pairs	Conditional binomial used to estimate FDR	No. of reads under peak	1: Negative binomial, 2: conditional binomial	Target FDR, optional window width, window interval	Yes / Yes	10
ERANGE v3.1	Tag aggregation	1: Height cutoff, 2: Height and fold enrichment over control counts in region	Hiqh quality peak estimate, per-region estimate, or input	Used to calculate fold enrichment and optionally p-values	p-value	1: None 2: *# control* / *# ChIP*	Optional peak height, ratio to background	Yes / No	4,18
FindPeaks v3.1.9.2	Aggregation of overlapped tags	Height threshold	Input or estimated	N/A	N	1: Monte Carlo simulation 2: N/A	Minimum peak height, subpeak valley depth	Yes / Yes	19
F-Seq v1.82	Kernel density estimation	s Standard deviations above kde for 1: random background, 2: control	Input or estimated	Kde for local background	Peak Height	1: None 2: None	Threshold standard deviation value, kde bandwidth	No / No	14
GLITR	Aggregation of overlapped tags	Classification by height and relative enrichment	User input tag extension	Multiply sampled to estimate background class values	Peak height and fold enrich-ment	2: *# control* / *# ChIP*	Target FDR, number nearest neighbors for clustering	No / No	17
MACS v1.3.5	Tags shifted then window scan	Local region Poisson p value	Estimate from high quality peak pairs	Used for Poisson fit when available	p-value	1: None 2: *# control* / *# ChIP*	p-value threshold, tag length, mfold for shift estimate	No / Yes	13
PeakSeq	Extended tag aggregation	Local region binomial p value	Input tag extension length	Used for significance of sample enrichment w/ binomial	q-value	1: Poisson background assumption 2: From binomial for	Target FDR	No / No	5

	Profile	Peak Criteria[1]	Tag Shift	Control Data[2]	Rank By	FDR[3]	User Input Parameters[4]	Artifact Filtering: Strand-based / Duplicate[5]	Refer-ence
				distribution		sample + control			
QuEST v2.3	Kernel density estimation	2: Height threshold, background ratio	Mode of local shifts that maximize strand cross correlation	Kde for enrichment and empirical FDR estimation	q-value	1: N/A 2: *# control* / *# ChIP* as a function of profile threshold	Kde bandwidth, peak height,subpeak valley depth,ratio to background	Yes / Yes	9
SICER v1.02	Window scan with gaps allowed	P value from random background model, enrichment relative to control	Input	Linearly rescaled for candidate peak rejection and p-values	q-value	1: None 2: From Poisson p-values	Window length, gap size, FDR (w/ control) or E-value (no control)	No / Yes	15
SiSSRs v1.4	Window scan	N_+-N_- sign change, N_++N_- threshold in region	Average nearest paired tag distance	Used to compute fold-enrichment distribution	p-value	1: Poisson 2: control distribution	1: FDR 1,2: N_++N_- threshold	Yes / Yes	11
spp v1.0	Strand specific window scan	Poisson p-value (paired peaks only)	Maximal strand cross-correlation	Subtracted before peak calling	p-value	1: Monte Carlo simulation 2: *# control* / *# ChIP*	Ratio to background	Yes / No	12
USeq v4.2	Window scan	Binomial p-value	Estimated or user specified	Subtracted before peak calling	q-value	1, 2: Binomial 2: *# control* / *# ChIP*	Target FDR	No / Yes	20

[1] Throughout the table 1: and 2: refer to one sample and two-sample experiments, respectively.

[2] The ‘Control Data’ column is intended to give a rough idea of how control data is used by the software. ‘N/A’ means that control data is not handled.

[3] The “FDR’ column describes how the FDR is or optionally may be computed. Note that ‘None’ indicates an FDR is not computed, however the experimental data may still be analyzed; ‘N/A’ indicates the experimental setup (1 sample or 2) is not yet handled by the software.

[4] The lists of ‘User Input Parameters’ for each program are not exhaustive but rather comprise a subset of greatest interest to new users.

[5]'Strand-based' artifiact filtering rejects peaks if the strand-specific distributions of reads do not conform to expectation, for example by exhibiting extreme bias of tag populations for one strand or the other in a region. 'Duplicate' filtering refers to either removal of reads that occur in excess of expectation at a location or filtering of called peaks to eliminate those due to low complexity read pileups that may be associated with, for example, microsatellite DNA.

Table 2

List of publicly available RNA-seq software packages discussed in this review

	Primary Category	Discovery	Need genomic assembly	Associated Read Mapper	Splice junctions	Quantitation	Reference
ABySS v1.0.11	Short-read assembler	Yes	No	N/A	- assembled	Read Coverage	33
BASIS V1	Existing transcript quantitation	No	Yes	External	- from existing models	Read coverage	41
ERANGE v3.1	Existing and novel gene quantitation	Yes	Yes	Blat Bowtie Eland	- from existing models - novel with blat	RPKM from gene annotations and novel transfrags	18
G-Mo.R-Se v1.0	Novel gene model annotation	Yes	Yes	SOAP	- predicted from transfrags	No	37
QPALMA v0.9.9.2	Spliced read mapper	Yes	Yes	Integrated	- predicted from transfrags	No	38
RNA-mate v1.1	Existing and novel gene quantitation	Yes	Yes	mapreads	- from existing models	Deprecated in v1.1	36
RSAT v0.0.3	Existing transcript quantitation	No	No – requires transcript sequences	Eland SeqMap (bundled)	- from supplied transcript sequences	RPKM from transcript sequences	8
TopHat v1.0.10	Existing and novel gene quantitation	Yes	Yes	Bowtie	- predicted from transfrags - from existing models	RPKM from supplied annotations	32
Velvet v0.7.47	Short read assembler	Yes	No	N/A	- assembled	Fold coverage	31