Defining transcribed regions using RNA-seq

Brian T Wilhelm[1,4], Samuel Marguerat[2,4], Ian Goodhead[3] & Jürg Bähler[2]

[1]Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, Montréal, Québec, Canada. [2]Department of Genetics, Evolution & Environment and UCL Cancer Institute, University College London, London, UK. [3]Unit for Functional and Comparative Genomics, School of Biological Sciences, University of Liverpool, Liverpool, UK. [4]These authors contributed equally to this work. Correspondence should be addressed to J.B. (j.bahler@ucl.ac.uk).

Published online 21 January 2010; doi:10.1038/nprot.2009.229

Next-generation sequencing technologies are revolutionizing genomics research. It is now possible to generate gigabase pairs of DNA sequence within a week without time-consuming cloning or massive infrastructure. This technology has recently been applied to the development of 'RNA-seq' techniques for sequencing cDNA from various organisms, with the goal of characterizing entire transcriptomes. These methods provide unprecedented resolution and depth of data, enabling simultaneous quantification of gene expression, discovery of novel transcripts and exons, and measurement of splicing efficiency. We present here a validated protocol for nonstrand-specific transcriptome sequencing via RNA-seq, describing the library preparation process and outlining the bioinformatic analysis procedure. While sample preparation and sequencing take a fairly short period of time (1–2 weeks), the downstream analysis is by far the most challenging and time-consuming aspect and can take weeks to months, depending on the experimental objectives.

INTRODUCTION

Surveying the fission yeast transcriptome at single-nucleotide resolution with RNA-seq

Recent data from several organisms indicate that the transcribed portions of genomes are much larger and complex than has previously been anticipated[1]. In addition, these data suggest that many functional properties of transcripts are based not on coding sequences but on regulatory sequences in untranslated regions[1]. Moreover, long and short noncoding transcripts are now recognized to have a major role in the regulation of gene expression[2,3]. Alternative start and polyadenylation sites as well as post-transcriptional regulatory processes, such as splicing, add an additional dimension to the rich transcriptional output. Genome-wide transcriptome structures have been sampled mainly using hybridization-based methods under one or a few experimental conditions. The recent development of next-generation sequencing technologies (NGS)[4] has opened new horizons for the analysis of transcriptional complexity. These new sequencers produce gigabases worth of sequence data in less than a week and for only a fraction of the costs incurred by classic Sanger sequencing. Although initially applied to the sequencing of genomic DNA, these techniques have more recently been used to sequence cDNA (RNA-seq). To date, the transcriptomes of about a dozen organisms have been sequenced in various organs, cell lines or physiological conditions[4,5].

We have previously applied RNA-seq, supplemented with data from high-density tiling arrays, to globally sample transcripts of the fission yeast *Schizosaccharomyces pombe* under multiple conditions, including rapid proliferation, meiotic differentiation and environmental stress, as well as in RNA processing mutants[6]. RNA-seq is proving to be a powerful quantitative method that shows little, if any, background noise. Besides providing reliable measurements of transcript expression levels independently of any genome annotation, RNA-seq is sensitive enough to detect widespread transcription in >90% of the *S. pombe* genome during vegetative growth. This includes traces of transcripts restricted to precise physiological conditions, and which, therefore, are not robustly transcribed or are rapidly degraded during vegetative growth. In addition, sequencing deep into the fission yeast transcriptome has uncovered many new noncoding transcripts, some of which are regulated in different physiological conditions. Complementary tiling array analysis has demonstrated that some of these transcripts were transcribed in the anti-sense direction, overlapping protein-coding genes. Besides discovering new transcriptional units, our combined analysis has improved the genome annotation by characterizing untranslated regions and alternative start and end sites of transcripts. Whole transcriptome sequence data contain information about post-transcriptional rearrangements such as splicing. For every intron in the genome, sequence reads spanning exon–exon junctions can be compared with reads spanning exon–intron junctions, thus providing an estimate of splicing efficiency. In fission yeast, such analysis has uncovered a surprising variability in splicing efficiency across introns, genes and conditions, indicating that even a simple eukaryote exploits regulation of RNA processing to shape a dynamic transcriptome. It is interesting to note that splicing efficiency is also largely coordinated with transcript levels, suggesting cross-talk between the transcriptional and post-transcriptional regulatory apparatus. Finally, this approach has also allowed the discovery of previously unannotated splicing events. Here, we describe in detail the RNA-seq protocol as well as data analysis strategies that we have developed for the in-depth investigation of the fission yeast transcriptome[6].

The Illumina sequencing platform

At present, three platforms dominate the NGS market. The FLX system from 454 (a Roche company, Basel, Switzerland), the SOLiD system from ABI, Foster City, California, USA and the Genome Analyser commercialized by Illumina, San Diego, California, USA. The technical principles behind these three systems have been described in detail elsewhere[7]. The protocol presented here is designed for the Illumina Genome Analyser, which is currently the most widely available platform (http://www.illumina.com/technology/sequencing_technology.ilmn). This platform allows the parallel analysis of millions of DNA molecules producing short sequence reads of 37 to over 70 nucleotides in length, depending on the sequencing conditions and machine version used. To achieve this, a library of short DNA molecules with specific adaptors ligated at both ends is applied to a sealed glass device called a flow cell. The surface of the flow cell is coated with oligonucleotides specific for the

library adaptors, allowing single DNA molecules to bind randomly to its surface. In a second step, each isolated DNA molecule is amplified locally on the slide (bridge amplification), creating distinct clusters of identical DNA molecules. These clusters are then sequenced by synthesis, in parallel, on the surface of the flow cell. For each round of sequencing, the four nucleotides coupled to four different fluorophores are added at the same time to the flow cell. Analogous to the use of standard chain terminator ddNTPs, the fluorophore-tagged dNTPs are reversibly chemically blocked to ensure that only a single incorporation occurs. After each round of dNTP incorporation, the slide is then imaged to record the nucleotide that has been added to each cluster. Blocking groups and fluorophores are finally removed before continuing with the next round. Each run produces terabytes of image data that are analyzed for the fluorescence emitted by each cluster at each cycle, thus allowing deduction of the cluster DNA sequence.

Experimental design

We describe below the procedure followed to obtain fission yeast cDNA libraries compatible with the Illumina Genome Analyser. It is important to note that this approach is easily transferable to any type of organism, except for the culture conditions and RNA extraction protocol. An outline of this protocol, with estimates of time required for each step, is shown in **Figure 1**. To provide sufficient RNA for the protocol below, a 100 ml culture of fission yeast cells were grown in liquid medium to exponential phase. Total RNA was then extracted from the frozen pellet using the hot phenol technique[8], which we recommend if working with fission yeast. However, the successful creation of sequencing libraries relies more on the quality of the starting RNA material than on the technique used for its isolation. Indeed, because of how the sequence data is processed (described below), RNA from any strain or growth condition can be directly compared as long as sufficient high-quality RNA can be extracted. Ribosomal RNA (rRNA) represents the vast majority of the total cellular RNA. To maximize the diversity of sequences retrieved from RNA-seq libraries, it is therefore advisable to reduce the quantity of rRNA present in the sample. This can be achieved either by enrichment for polyadenylated RNA or by depletion of rRNA species. This protocol describes an approach based on the selection of poly(A)+ RNA, but depletion of rRNA has been found to be an attractive alternative[9] (see 'Alternative experimental design' below).

Because of the huge depth of sequencing data generated by NGS, small amounts of contaminating genomic DNA may appear to be transcribed RNA when the sequencing results are analyzed. It is therefore important to treat the enriched RNA samples with DNase to completely remove any remaining amounts of DNA. After a column purification (Qiagen, Hilden, Germany) of the DNase digests, the enriched RNA samples are then converted to cDNA using poly(dT) or random primers. Although priming reverse transcription with poly(dT) further enriches the library for sequences derived from poly(A)+ RNA, it also reduces the representation of the 5′ ends of long transcripts. The use of random primers represents a good alternative, ensuring a more uniform sequence representation across transcripts. Finally, second-strand cDNA synthesis is carried out to obtain double-stranded cDNA, which can in turn be processed using the Illumina library creation protocol originally developed for genomic DNA. In this procedure, the cDNA sample is first fragmented into pieces of a size compatible with

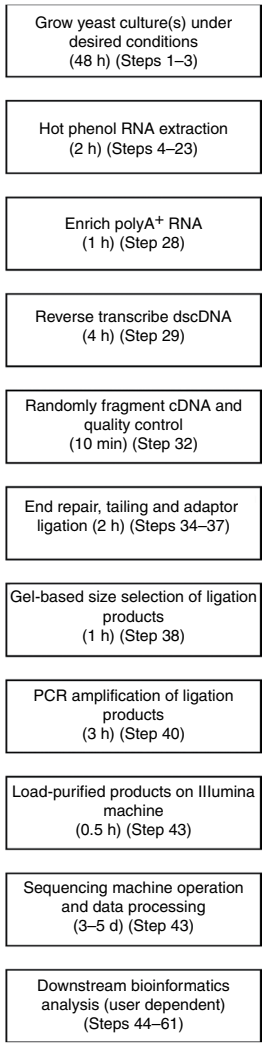

Figure 1 | Flowchart of experimental procedure. An overview of the main steps of the RNA-seq protocol is shown in a schematic flow chart. Each step indicates the approximate amount of time required in minutes (min), hours (h) or days (d) and refers to the appropriate step in the protocol.

downstream cluster formation and sequencing. This is usually carried out by nebulization, but other techniques have been explored[10]. Nebulization (or other fragmentation methods) generates double-stranded cDNA fragments with a mix of blunt ends as well as 3′/5′ overhangs. The library needs, therefore, to be treated with Klenow and T4 DNA polymerases to obtain fragments with uniform blunt ends. Next, adenine residues are added to the fragment extremities using the Klenow enzyme to prepare the library for efficient adaptor ligation. Finally, the DNA adaptors necessary for cluster formation, amplification and sequencing are ligated at both ends of the fragments using a T4 DNA ligase. After PCR amplification of the library for a limited number of cycles, fragments of 120–170 bp in length are size-selected on an agarose gel. This step will remove unincorporated adaptors and ensure that the average fragment size of the library is optimal for attachment to the flowcell and cluster formation. Such RNA-seq libraries are ready for cluster formation and loading on the Genome Analyser.

Alternative experimental design

The protocol described here has the advantage of being robust and of being partially based on the genomic DNA library creation

protocol initially developed by Illumina. These features have the advantage that most new technological developments optimized for genomic DNA sequencing should be easily transferable to this protocol and, therefore, to the analysis of cDNA. Paired-end analysis, for instance, is a variation of the classical Illumina sequencing protocol, where both ends of each library fragment are sequenced[11]. Applied to RNA-seq, this approach could help in deciphering the structure of complex transcriptomes rich in repetitive sequences. Another variation, called multiplexing, allows sequencing of pools of samples in a single reaction (each library being tagged with a different DNA sequence index). This approach permits reduction of the amount of resources needed per sample at the cost of sequence depth. It could be an attractive option for small transcriptomes.

A limitation of the protocol described here is its inability to determine from which DNA strand transcripts are derived. This could be problematic for genomes where overlapping transcription is prevalent[1]. Several alternative RNA-seq protocols can be found in the literature. They mostly differ from the approach described here by the way in which the oligonucleotide adaptor sequences, required for sequencing, are joined to the cDNA library to be sequenced. It is important to note that these alternative approaches retain information about the strand of the genome from which transcripts are expressed. We will briefly describe below six methods published so far. A protocol developed initially for sequencing of small RNAs has been adapted for the analysis of total RNA[12]. In this case, the RNA sample is fragmented to reach a size compatible with sequencing. The adaptor sequences are then ligated directly to the RNA molecules using a T4 RNA ligase. This RNA library can then be reverse transcribed and sequenced. Fragmenting RNA, rather than DNA, has the possible advantage of reducing RNA secondary structures that can lead to heterogeneity in coverage. In a second approach, fragmented RNA is polyadenylated *in vitro* and reverse transcribed using a poly(dT) primer containing both adaptor sequences that are separated back-to-back by an endonuclease site. The cDNAs are then ligated intramolecularly and recleaved at the endonuclease site, leaving single-stranded cDNA containing the adaptors at both ends[13]. This approach appears to be leading to uniform sequence coverage across transcripts. A third protocol uses double random priming to produce adaptor-containing cDNA. First-strand cDNA is synthesized using random primers containing one biotinylated adaptor sequence and purified on streptavidin beads. Second-strand cDNA is then synthesized from those purified templates with random primers containing the second adaptor sequence[14]. In a fourth approach, initially developed for the SOLiD platform, a peculiarity of certain reverse transcriptases is exploited. These enzymes add short poly(C) tracts at the end of the cDNA they synthesize. If primers containing an adaptor sequence coupled to three guanines are added to the reverse transcription reaction, they will hybridize to the poly(C) tracts of the newly synthesized cDNA, thereby allowing the reverse transcriptase to 'switch' from its template RNA to the primer and to incorporate the adaptor sequence at the 3′ end of every synthesized cDNA. The 5′ adaptor sequence is appended to the primer used for first-strand synthesis[9]. This approach has the advantage of avoiding RNA or DNA ligation steps. The next protocol uses dUTP as a surrogate for dTTP during second-strand synthesis. This permits selective degradation of the second cDNA strand after adaptor ligation using a uracil-*N*-glycosylase. This treatment and the use of engineered DNA adaptors ensures that only the cDNA strand corresponding to the actual transcript is used for library amplification and sequencing[15]. In a last protocol, sequencing adaptors are ligated directly to single-stranded cDNA, which has the advantage of being simple and close to the original Illumina protocol[16]. No doubt, comparing these different approaches will be of great use in understanding the impact of library preparation on RNA-seq data.

Different strategies have also been used to reduce the proportion of rRNAs in RNA-seq libraries, based on either the enrichment of polyadenylated transcripts or depletion of rRNAs. rRNAs represent over 90% of the total RNA in the cell. Therefore, when total RNA is analyzed by RNA-seq, the great majority of sequence reads will be derived from transcripts encoding rRNAs, reducing dramatically the coverage of the rest of the transcriptome. This bias in coverage represents a waste of resources. In this protocol, we have used a commercially available kit for the affinity purification of poly(A)+ transcripts. This approach is quite robust, allowing a substantial reduction of the proportion of rRNAs present in the sample. It is important to note that poly(dT) primers not only exclusively bind to polyadenylated transcripts but also interact with nonpolyadenylated transcripts, possibly because of internal poly(A) tracts[17]. Several poly(A)+ purification systems are at present commercially available, but a detailed comparison of their respective performances is outside the scope of this protocol. Selective depletion of rRNA is an alternative approach to enrich for nonribosomal RNA. This approach has the advantage of not restricting the enrichment of transcripts to only those that bind to a poly(dT) primer. This is an attractive alternative as many transcripts have been reported to lack a poly(A) tail and may, therefore, be missed by analyses based on poly(dT) enrichment[1]. In our hands, however, depletion of rRNA is less efficient than the selection of poly(A)+ RNA. A simple approach to this end is to purify rRNA out of the total RNA sample with magnetic beads (Ribominus, Invitrogen, Carlsbad, California, USA) coupled to oligonucleotides whose sequences are complementary to rRNA sequences. A second approach makes use of an exonuclease that specifically degrades RNA molecules bearing a 5′ phosphate (Epicentre, mRNA-ONLY kit, Madison, Wisconsin, USA), which is a characteristic of rRNA but not of capped mRNAs. Therefore, rRNAs are selectively degraded, leaving the sample enriched for mRNAs though possibly also for other RNA species resistant to the enzyme.

Data analysis

Data analysis is a critical aspect and often the most time consuming aspect of RNA-seq. There are two main challenges in carrying out data analysis. First, the enormous amount of data produced by NGS technologies can make downstream computational analysis difficult without appropriate infrastructure. For this reason, important efforts are currently being dedicated to the development of computational tools, ideally allowing the analysis of NGS data using a standard desktop computer. However, many applications still do require high-end computing resources, and we recommend that this issue be taken into account before embarking on large NGS-based experiments. Second, the huge amount of data produced by each sequencing run requires substantial storage and backup capacity, which can add considerably to the experimental costs.

A graphical overview of several essential aspects of the data analysis portion of the protocol is shown in **Figure 2**. These steps include matching the reads to the reference genome (**Fig. 2a**), defining transcribed regions and identifying novel introns, genes and so on (**Fig. 2b**), and comparing sequence expression scores between

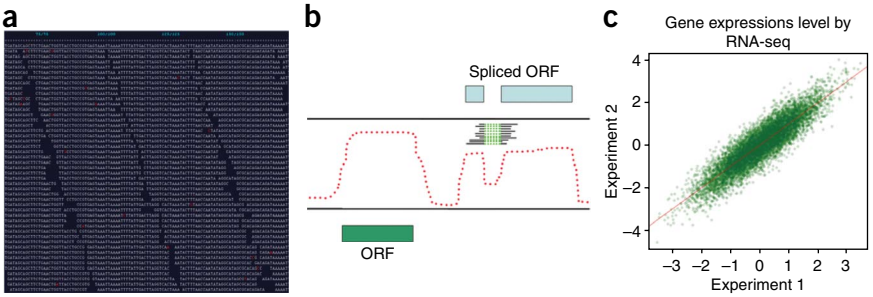

Figure 2 | Diagrammatic view of analysis steps in a RNA-seq experiment. Example data of the output of several critical steps common to most RNA-seq experiments are shown including (**a**) the alignment of sequence reads to a reference genome, (**b**) plotting of single nucleotide signal scores and spliced reads with known genome annotation and (**c**) a comparison of sequence expression scores for features in two different growth conditions.

growth conditions (**Fig. 2c**). The details of all the steps of the data analysis are described below in detail.

The raw outputs of the Illumina Genome Analyser are images of the flow cell taken in each of four wavelengths after every sequencing cycle. Signal intensities of the four nucleotides are extracted from these images and used for base calling. The Illumina image analysis software (Firecrest) collects feature information that the base-calling software (Bustard) will subsequently convert into IUPAC-coded DNA sequences with associated intensity and quality scores. The last stage of 'onboard software', Gerald, will perform quality calibration and data filtering, ultimately producing a text file of short read sequences in FASTQ format. An example of the FASTQ format is shown below:

```
@SEQUENCE_IDENTIFIER
GTATAGCGATAGAATTTCGATATTGGATTTAGATCTTCAA
+SEQUENCE_IDENTIFIER
%%%+++)(%%%%).1***-+*''))**55CCF>%%
%>>CCC
```

Working on this preprocessing pipeline can substantially increase the amount and the quality of data[18–20]. A filtering step can be applied at this point, allowing the removal of poor quality sequence reads that can reduce computational time and effort for further analysis. The FASTQ data format has the advantage of compactly storing a quality score for each base called in analogous, although not identical, manner to a Phred score[21] and so this can be used to filter individual sequences. Illumina FASTQ scores can be converted to Q_{phred} score using the PERL code: $Q = 10 \times \log(1 + 10\ (\text{ord}(\$sq) - 64) / 10.0)) / \log(10)$. The quality scores for each base pair position in each sequence can therefore be converted, and reads whose sum of Phred scores falls below an arbitrary threshold can be excluded—e.g., a 35-bp read would need a sum of quality scores of 1,400 (35×40) to have an average Phred score of 40 (probability of 1:10,000 of incorrect base call).

When a satisfactory base call cannot be made by the Illumina pipeline, a base will be reported as an 'N'. Those reads which contain only bases called as Ns should be removed from the dataset, as they are noninformative. Similarly, sequence reads which contain <15 bp of sequence before the first N is called should be removed from the dataset, as they are too short to map efficiently. A common artifact of Illumina sequence data is the presence of aberrant homopolymeric A/T stretches, which can result from oligo-dT priming too far down the poly(A) tail for the read to reach the actual transcript coding sequence as well as from platform specific artifacts. Those reads should be filtered out. However, it is important to keep in mind that reads containing a mixture of poly(A) or poly(T) stretches and alignable sequences can contain true biological information, such as positions of polyadenylation sites.

Filtered sequence reads can then be mapped back to a reference genome to derive expression values for annotated and unannotated genomic features. There are published examples of RNA-seq efforts that do not use a direct reference genome but build up models of expressed features *de novo* or from related species[22–25]. Although this possibility exists, the details involved are outside the scope of this protocol. The optimization of the process of mapping reads, which often represents a bottleneck in the analysis, is an extremely active field of research and many different mapping/alignment algorithms designed to process millions of small reads are now available[4,5,26]. The mapping strategy used in our initial study used a program called BLAT[27] and proved to be robust but computationally intensive. The general principles described in this protocol can be easily emulated to other mapping software packages. We recommend trying different mapping approaches as individual programs may show differing performance depending on factors such as the size and complexity of reference genomes used.

When using BLAT on a computer cluster with LSF as a scheduling system, jobs matching FASTA short read sequence files to reference sequences should be submitted as a batch. The command syntax may vary by cluster set-up, but should be similar to:

```
bsub    -J 'myBLAT[1–50]'
        -o results_BLAT-%I.out
        -R 'select[mem > 2,000] rusage[mem = 2,000]'
        -q long
        -P pombe
        ./BLAT_seq.sh
```

where the options shown represent job name (-J), output file name (-o), resource requirements (-R), priority queue (-q) and project name (-P), and the shell script to be run on batch is 'BLAT_seq.sh'. The contents of BLAT_seq.sh, which contains the BLAT command, are shown below:

BLAT FASTA_genome.txt FASTA_sequences.txt Output.bsl -out = psl -oneOff = 1 -noHead -tileSize = 8

where the files for the reference genome (Genome.txt), short read files (FASTA_sequences.txt) and output file (Output.bsl) are specified after the BLAT command, along with the output style (-out = psl), the number of mismatches allowed in a tile used to trigger an alignment (-oneOff = 1), whether a header is printed for the results (-noHead) and the size of match which will trigger an alignment (-tileSize = 8). The 'bsub' command can pass an environmental variable representing the batch job number (range shown in square brackets after -J option) to the shell script to be run, which can be used as a filename appendix to assign input from each of the correspondingly numbered FASTA sequence files and produce identically numbered output files.

The last step of this part of the analysis is to identify reads with multiple matches in the genome and to separate them from reads with unique matches. This is important because reads with multiple hits will significantly affect any downstream analysis. In the BLAT output file, reads which match multiple times in the genome will have the matches listed sequentially by query name ('qName'), so a regular expression search for a qName which appears on two consecutive lines indicates a read which matches more than one location. Unique reads can be written to one output file, whereas reads with multiple matches can be written to another file. Although partitioning unique reads, other values in the results file (i.e., mismatches, Ns) can be used to calculate a percent identity for each read. An identity value for reads that match multiple locations can be used to distinguish between a *bona fide* match (high identity) and weak matches to the related sequence (lower identity). Appropriate cut-offs to use for this analysis will be genome/species specific, but for the relatively unrepetitive genome of *S. pombe* few reads tend to fall into this class.

Beside the software described above, a variety of open-source mapping programs now exist, including MAQ[28], BOWTIE/TOPHAT[29,30], SHRiMP[31] and SOAP[32]. All of these programs increase the speed of read mapping by orders of magnitude while also improving sensitivity. In addition, these programs can be used to directly determine SNPs (MAQ, SOAPsnp), map sliced reads (BOWTIE/TOPHAT) or directly utilize the quality scores in the FASTQ files (MAQ). The use of these more efficient mapping programs will change the protocol above as the output files will have a different format, but the minor adjustments in the later analysis should be more than compensated for by removing the need for a large compute farm for read mapping.

Once the data have been mapped back to a reference genome, a series of straightforward analyses can be carried out, which will be common to most RNA-seq experiments, independent of the species used. First, an expression score can be computed for every position in the genome by simply counting how many times each base has been sequenced. Expression scores can be viewed as being equivalent to microarray data with single base-pair resolution. It is therefore tempting to take advantage of existing analysis tools developed for microarray analysis. It is important, however, to keep in mind that RNA-seq data, because of their digital nature, are in essence different from microarray signal intensities. Therefore, tools developed for microarray analysis should not be used blindly for the treatment of RNA-seq data. Assignment of expression scores for annotated genomic elements can be carried out by using a PERL script to parse all mapped reads (with or without reads matching in multiple locations) and to assign scores for the elements where these reads fall into. This can be done either by scoring each sequence read in an element once (regardless of length) or by taking the start and end of the read match and scoring each base pair once. Both approaches will yield fairly similar results. Once completed, the expression values can be median-centered and differentially expressed genes can be identified.

Second, RNA-seq is a powerful tool for the study of the regulation of alternative splicing. Analyzing sequence reads spanning exon–exon junctions ('transreads') provide qualitative and quantitative information about any splicing event. Identification of transreads can be carried out by parsing the BLAT output file for reads with large values for gaps inserted into the target sequence ('tBaseInsert'). Counting the number of transreads spanning a given exon–exon junction (TR) and comparing it to the average number of reads that cover the corresponding exon–intron junctions at either end of the intron (OR) provide a measurement of splicing efficiency (i.e., SE = TR: $[(OR_{5'} + OR_{3'}) / 2]$). Splicing efficiency ratios within growth conditions or between different growth conditions can be used to identify those exons which exhibit regulated splicing.

Third, previously unannotated sequence features such as new exons, introns or 5′ or 3′ untranslated regions (UTRs) can be identified. Identification of novel exons can be achieved by writing a PERL script to find transreads that have one end of the spliced read in a known exon, but where the other end is not in an annotated exon. The coordinates of the end of the spliced read that represent a putative new exon can be used to rescreen all reads to find others that overlap this new exon. Reads which do overlap the new exon can be used to increase the size of the exon in an iterative process until no more overlapping reads are found. Expression scores for the new exons should be comparable with upstream/downstream exons and should be visually inspected for proper splice site sequences. Assignment of expression values is often limited to the exons of a transcript or annotated open reading frame (ORF), because in many cases the true extent of the 5′ and 3′ UTRs are undetermined. Estimating the ends of the UTRs of a transcript can be difficult using nonstrand specific data from oligo-dT-primed RNAs such as those described in this protocol. Much less frequent reads towards the 5′ end of a transcript (due to the RT enzyme falling off) can also make accurate determination of 5′ UTRs difficult. In addition, ORFs in close proximity may have strong signals from their UTRs, which may overlap, thus preventing determination of one end. Bearing in mind these limitations, it is possible to estimate the ends of UTRs. A minimum signal threshold can be determined by looking at regions of the genome known not to contain transcribed elements (i.e., an intergenic region with little or no signal). Once this value is determined, a PERL script can start from the end of the last known exon to determine the point at which the signal drops below the threshold. The same PERL script can then carry out the same determination at the 3′ end.

MATERIALS

REAGENTS

- Distilled water
- Yeast strain of interest
- Diethyl pyrocarbonate (DEPC) (Sigma, cat, no. D5758) **! CAUTION** Extremely poisonous by inhalation or skin contact and causes irritation to the eye and the skin. Use glove and safety glasses and work under a fume hood.
- Bacto–yeast extract (BD, cat. no. 212750-500G)
- 3 M Sodium acetate (NaAc) buffer solution, pH 5.2 (Sigma, cat. no. S7899)
- Sodium hydroxide (Sigma, cat. no. 221465) **! CAUTION** Extremely corrosive and poisonous. Causes burns along the digestive and the respiratory tracts. Use glove and safety glasses and work under a fume hood.
- 1 M Tris-HCl, pH 7.6 (Sigma, cat. no. T-2788)
- SDS, 10% (wt/vol) solution (Ambion, cat. no. AM9822) **! CAUTION** May cause severe irritation to the eye, the skin, the respiratory and the digestive tracts. Use gloves and safety glasses.
- Ethylene diamine tetraacetic acid (EDTA) disodium dihydrate (Sigma, cat. no. E4884-100G)

- Glucose (Sigma, cat. no. G8270-1KG)
- Difco agar (BD, cat. no. 214530-500G)
- Histidine (Sigma, cat. no. H3751-5G)
- Leucine (Sigma, cat. no. 855448-10G)
- Adenine (Sigma, cat. no. A2786-25G)
- Uracil (Sigma, cat. no. U1128-100G)
- Lysine (Sigma, cat. no. L8021-5G)
- Phenol–chloroform, acidic (refrigerated, Sigma, cat. no. P-1944) **! CAUTION** May cause irritation to the eye and the skin. May cause depression of the central nervous system, cardiac disturbances, reproductive and fetal effects, cancer based on animal studies and irritation to the respiratory and the digestive tracts. Use gloves and safety glasses and work under a fume hood.
- Qiagen RNeasy Spin mini kit (Qiagen, cat. no. 74104)
- Qiagen Oligotex mRNA purification kit (Qiagen, cat. no. 72022)
- QIAquick PCR clean-up kit (Qiagen, cat. no. 28104)
- SuperScript II double-stranded cDNA synthesis kit (Invitrogen, cat. no. 11917-010)
- Chloroform:isoamyl alcohol (24:1)(Sigma, cat. no. C-0549) **! CAUTION** May cause irritation to the eye and the skin. May cause depression of the central nervous system, cardiac disturbances, reproductive and fetal effects, cancer based on animal studies and irritation to the respiratory and the digestive tracts. Use gloves and safety glasses and work under a fume hood.
- EtOH, absolute (Sigma, cat. no. 7023-500ML) **! CAUTION** Highly flammable and may cause irritation to the respiratory tract, the eye and the skin. Keep the container tightly closed. Keep away from sources of ignition.
- TURBO DNase (Ambion, cat. no. AM2238)
- QIAquick gel extraction kit (Qiagen, cat. no. 28704)
- Illumina machine reagents
- Klenow (3′→5′ exo$^-$)(NEB, cat. no. M0212L)
- T4 DNA polymerase (NEB, cat. no. M0203S)
- T4 Quick ligation kit (NEB, cat. no. M2200L)
- 50× TAE buffer (Bio-Rad, cat. no. 161-0743)
- Anchored oligo(dT)20 primer (Invitrogen, cat. no. 12577011)
- Agarose (Invitrogen, cat. no. 16500500)
- Phusion Taq (Finnzymes, cat. no. F-540S)
- 10 mM dNTP mix (Invitrogen, cat. no. 18427-013)

EQUIPMENT

- Table-top centrifuge for 1.5 ml Eppendorf tubes
- Table-top centrifuge for 50 ml Falcon tubes
- Heating block for 1.5 ml tubes
- Thermometer
- 250 ml Erlenmeyer flasks
- Nanodrop 2000 (Thermo Scientific)
- Shaker/incubator
- 50 ml Falcon tubes
- Razor blade
- 2 ml Phase-lock tubes (heavy) (Eppendorf)
- 1.5 ml Eppendorf tubes (Eppendorf)
- 2 ml Eppendorf tubes (Eppendorf)
- Illumina Genome Analyser (Illumina)
- Agilent Bioanalyzer (Agilent)
- Agilent RNA Nano 6000 kit (Agilent)
- Agilent Bioanalyzer DNA 1000 kit (Agilent)
- Computer workstation (>4 Mb RAM)
- PCR machine
- Agarose gel electrophoresis tank (VWR, cat. no. 89032-292)
- Adjustable desktop power supply (VWR, cat. no. 93000-744)
- Nebulizer (Invitrogen, cat. no. K7025-05)

REAGENT SETUP

DEPC-treated water To make 0.1% (vol/vol) solution, add 1 ml of DEPC to 1l of ddH_2O and leave at room temperature (RT, 20 °C) for 2 h. Autoclave water and allow to cool before use.

0.5 M EDTA Add 186.1 g EDTA to 800 ml of ddH20. Add ~20 g of NaOH pellets while stirring until pH reaches 8.0 and then autoclave. Can be stored indefinitely at RT.

Tris-EDTA (TE) Dilute 1 ml of 1 M Tris (pH 7.6) and 0.2 ml of 0.5 M EDTA into 97 ml of DEPC-treated water.

TES solution Dilute 1 ml of 1 M Tris pH 7.6, 2 ml of 0.5 M EDTA and 5 ml of 10% (wt/vol) SDS into 92 ml of DEPC-treated water. **! CAUTION** DEPC is highly toxic. Use goggles and work in a fume hood. Can be stored indefinitely at RT.

Supplement solutions Prepare 7.5 mg ml^{-1} stock solutions of histidine, leucine, adenine, lysine and a 3.75 mg ml^{-1} stock solution for uracil in dd$H_2$0 and then autoclave. Can be stored indefinitely at RT.

YES broth Add 5 g of yeast extract and 30 g of glucose to 1 liter of ddH_2O and stir until dissolved. Add supplement solutions above to final concentration of 250 mg litre^{-1}. Autoclave. Can be stored indefinitely at RT.

YES plates Prepare recipe for broth as above but add 20 g of agar per liter of broth. Autoclave. Can be stored indefinitely at RT.

EQUIPMENT SETUP

Shaker incubator The shaker incubator should be set at 32 °C (or appropriate temperature for mutant strains).

Heating block The heating block should be set to 65 °C, with temperature verified by thermometer.

Illumina machine Illumina machine should be installed and tested using standardized controls well before running test samples. In addition, informatics issues related to raw data storage, processing transfer and archiving should be addressed well in advance.

Analysis workstation Analysis workstation should be setup in advance with latest distribution of appropriate software (ActiveState PERL, R, Bioconductor and so on).

PROCEDURE

Yeast culture

1| Start preculture of *S. pombe* growing at 32 °C in 50 ml flask overnight by inoculating 30 ml of YES media with cells picked from a YES agar plate. For timing estimates of the different steps, see **Figure 1.**

2| From the exponentially growing preculture (OD 595 <0.5), transfer a sufficient volume of precultured yeast to 100 ml of fresh YES media such that the final OD of the harvested material will be between 0.2 and 0.5 (i.e., inoculate 100 ml YES with a small amount of exponential-phase preculture and incubate at 32 °C until it reaches OD 595 of 0.2–0.5).

3| Separate the culture into two 50 ml Falcon tubes and centrifuge for 2 min at 1,500*g* at 25 °C to form cell pellets.

▲ **CRITICAL STEP** Upregulation of stress genes can occur very rapidly in cells, so cells should be centrifuged and lysed as quickly as possible.

RNA extraction

4| Discard the supernatant, add 750 μl of TES to pellet and resuspend cells with pipette.

5| Immediately add 750 μl acidic phenol–chloroform and vortex the sample for 15 s.

! CAUTION May cause irritation to the eye and the skin. May cause depression of the central nervous system, cardiac disturbances, reproductive and fetal effects, cancer based on animal studies and irritation to the respiratory and the digestive tracts. Use gloves and safety glasses and work under a fume hood.

6| Place the sample into a preheated 65 °C heat block.
▲ CRITICAL STEP If multiple samples are being prepared, then carry out Steps 4–6 for each sample before proceeding with the subsequent samples.

7| Incubate all the samples at 65 °C for 1 h. Every 10 min, vortex each sample for 10 s.

8| Place the samples on ice for 1 min, vortex for 20 s and centrifuge for 15 min at 14,000*g* at 4 °C.

9| Prespin 2 ml yellow phase-lock tubes for 10 s at 14,000*g* at RT.

10| Add 700 μl of acidic phenol–chloroform to each phase-lock tube.

11| Take 700 μl of the aqueous phase from Step 8 and add to the phase-lock tubes from Step 10.

12| Mix the tubes thoroughly by inverting (do not vortex the tubes) and then centrifuge for 5 min at 14,000*g* at 4 °C.

13| Prespin 2 ml phase-lock tubes as in Step 9.

14| Add 700 μl of chloroform:isoamyl alcohol (24:1) to each phase-lock tube from Step 13.

15| Take 700 μl of the aqueous phase from Step 12 and add to the phase-lock tubes from Step 13.

16| Mix the tubes thoroughly by inverting (do not vortex tubes) and then centrifuge for 5 min at 14,000*g* at 4 °C.

17| Prepare normal 2 ml Eppendorf tubes with 1.5 ml of ice cold 100% EtOH and 50 μl of 3 M NaAc pH 5.2.

18| Transfer 500 μl of aqueous phase from Step 16 to the tubes from Step 17 and vortex tubes for 10 s. Samples can be precipitated at −20 °C overnight (or at −70 °C for 30 min).
■ PAUSE POINT Samples can be left overnight at −20 °C.

19| Centrifuge for 10 min at 14,000*g* at RT. Discard the supernatant and add 500 μl 70% (vol/vol) EtOH (4 °C, prepared with DEPC water). Do not vortex the samples.

20| Spin the tubes for 1 min with the tubes in same orientation as in Step 19. When finished, aspirate most of the supernatant and spin the tubes again for 5 s and remove rest of the liquid with pipette.

21| Air dry pellets for 5 min at RT.

22| Add 100 μl of DEPC water and incubate 1 min at 65 °C (or 10 min at RT). Dissolve pellet first by pipetting up and down (~30×) until no particles are left, then gently vortex for 10 s.

23| Measure the RNA concentration by Nanodrop (or alternative UV spectrophotometer) with DEPC water as reference. Expect 1.6–4 mg of RNA (OD = 0.2–0.5), but the yield may be less for RNA isolated under some growth conditions (e.g., meiosis).
■ PAUSE POINT RNA can be stored indefinitely at −70 °C.
? TROUBLESHOOTING

RNA purification
24| If required because of residual organics (low 260:230 ratio), further purify 100 μg of RNA from Step 23 on an RNeasy spin column according to the manufacturer's instructions.

25| Incubate RNA with RNase-free DNase for 30 min at 37 °C according to manufacturer's instructions.

▲ CRITICAL STEP The DNase must be RNase free with high activity. Although RNeasy columns will preferentially bind RNA, after this step any trace amounts of genomic DNA will be impossible to distinguish from products of bona fide transcription.

26| Purify samples using Qiagen RNAeasy columns as in Step 24.
■ PAUSE POINT RNA can be stored indefinitely at −70 °C.

27| Assess the quality of the purified RNA using an RNA Nano chip on the Agilent Bioanalyzer. Profile and electrophoretogram should look like the one shown in **Figure 3**. The RNA integrity number (RIN) should be >9 to continue.

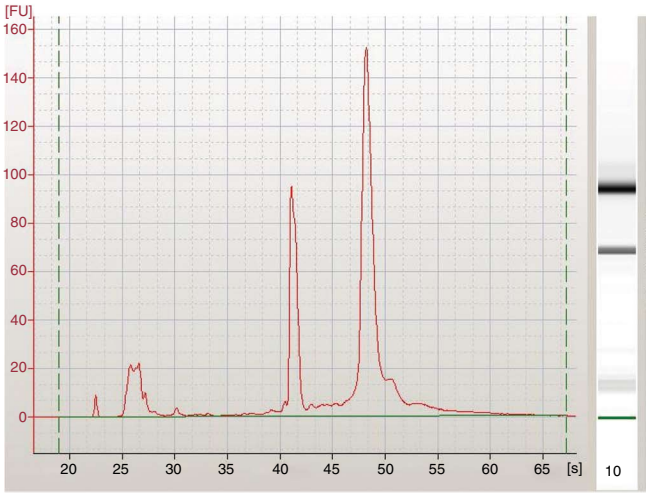

Figure 3 | Example of acceptable Bioanalyzer trace for total RNA. Total RNA samples run on a bioanalyzer will produce a profile plot, where the *y* axis represents fluorescence units (FU) and the *x* axis represents the runtime in seconds (s). The two peaks of the 18S (~41 s) and 28S (~49 s) ribosomal RNAs are used to calculate RNA integrity number (RIN) as a proxy for RNA quality.

Poly(A) enrichment of RNA

28| Enrich the RNA sample for mRNA by using a poly(A)+ RNA selection kit, such as the Oligotex mRNA purification kit from Qiagen. Carry out two sequential rounds of poly(A) enrichment on each sample according to the manufacturer's instructions and measure RNA quantity by Nanodrop.
▲ CRITICAL STEP Elution of the poly(A)+ mRNA should be done in large volume (~100 µl) to ensure maximum yield. Care should be taken to ensure that the elution buffer does not cool while resuspending Oligotex beads. Typical yields of poly(A)-enriched material represent ~25% of input material.
? TROUBLESHOOTING

cDNA preparation

29| Carry out reverse transcription with SuperScript double-stranded cDNA synthesis kit or similar system using the manufacturer's protocol for mRNA, with an oligo-dT_{20} primer. Carry out three independent RT reactions, purify the cDNAs using QIAquick PCR purification kit and pool them.
■ PAUSE POINT Synthesized cDNA can be stored indefinitely at −20 °C.

30| Assess the quality of the prepared cDNA by using a DNA 1000 kit on the Agilent Bioanalyzer.
▲ CRITICAL STEP In this section, we refer to Illumina protocols, which are frequently updated and refined, so it is best to refer to current Illumina protocols for these steps. Illumina proprietary reagents also change frequently; the most current should be used.
? TROUBLESHOOTING

Sample preparation for Illumina 1G machine

31| Dilute the pooled cDNA sample to a volume of 50 µl with TE and 700 µl of nebulization buffer (Illumina).

32| Fragment samples at 32–35 psi for 6 min on ice using nebulizer.

33| Purify the fragmentation products using QIAquick kit eluting in 50 µl of water. Qiagen PB buffer can be pipetted around the nebulizer to recover the maximum amount of sample.

34| Incubate fragments with 50 U of Klenow DNA polymerase and 30 units of T4 DNA polymerase with 20 µl of a mix of 10 mM dNTPs for 30 min at 16 °C to blunt fragment ends.

35| Purify blunt end-repair fragments (Step 34) using a QIAquick PCR purification column with sample eluted in 32 µl of Qiagen elution buffer.

36| Add 15 units of exo-Klenow, 10 µl of 1 mM dATP and 5 µl of 10× Klenow buffer and incubate at 37 °C for 30 min to add a 3′ adenine overhang to the blunt-ended double-stranded cDNA fragments.

37| Ligate the tailed cDNA fragments to the two proprietary oligonucleotide adaptors (Illumina) that will permit the nonspecific amplification of cDNA fragments. Add adaptors and T4 DNA ligase (NEB, Ipswitch, Massachusetts, USA) according to the current Illumina protocols and incubate at RT for 15 min.

38| Separate ligation products on a 2% (wt/vol) Tris-acetate-EDTA (TAE)-agarose gel. Using a razor blade, excise the specific region of the gel desired, according to the size range for the insert (generally 120–170 bp in size).

39| Use Qiagen Gel purification kit to extract ligation products according to the manufacturer's protocol, eluting fragments in 30 μl of water.

40| Set up PCR reactions using 1 μl of purified cDNA fragments (Step 39), 22 μl of water, 25 μl of 2X Phusion Taq master mix and 1 μl each of 1.1 and 2.1 primers (Illumina)—these primers are complementary to the ligated adaptors (Step 37) and compatible with oligonucleotides attached to the Illumina flowcell. Set up cycling program as

Time (s)	Temperature (°C)	Cycling
30	98	Hold
10	98	18 cycles
30	65	
30	72	
300	72	Hold
Infinite	4	Hold

■ **PAUSE POINT** PCR reactions can be left at 4 °C overnight.

41| Following amplification of ligation products, purify the PCR products using QIAquick PCR kit or equivalent and measure the concentration with a Nanodrop.

42| Dilute the cDNA to a working concentration of 10 nM in TE.
■ **PAUSE POINT** Fragment library can be stored indefinitely at −20 °C.

Illumina genome analyzer analysis
43| Using Illumina cluster station, run the appropriate .xml script to operate syringe pump to load cDNA fragments onto an Illumina flowcell.
▲ **CRITICAL STEP** The Illumina Genome Analyzer should be set up, tested and prepared to run well in advance of sample loading. Manufacturer's protocols for this should be followed precisely.

44| Run the 'on board' Illumina analysis pipeline (Firecrest, Bustard and Gerald).

45| Parse through all FASTQ files to strip quality scores and temporarily store data in FASTA format.

46| Identify sequence reads which contain only Ns as base-calls and remove from the dataset.
? TROUBLESHOOTING

47| Identify sequence reads which contain <15 bp of sequence before the first N base is called and remove these from the dataset.

48| Identify sequence reads that are homopolymeric A/T stretches and remove these from the analysis.

Mapping of reads
49| Align filtered FASTA reads to reference genomes using the BLAT program with parameters set as 'BLAT FASTA_genome.txt FASTA_sequences.txt Output.bsl -out = psl -oneOff = 1 -noHead -tileSize = 8'. Speed of analysis will be highly dependent on computer hardware and number of reads to match.

50| Optional. Use a PERL script to parse filtered FASTA reads and split these into an arbitrary number of files depending on available nodes in computer cluster.

51| Optional. Use LSF to submit batch job for BLAT matching of all FASTA sequence files generated in Step 41. The command syntax may vary by cluster setup, but should be similar to

```
bsub    -J 'myBLAT[1-50]'
        -o results_BLAT-%I.out
        -R 'select[mem>2000] rusage[mem=2000]'
        -q long
        -P pombe
        ./BLAT_seq.sh
```

52| Optional. Use a PERL script to parse through output files to combine individual results files from batch job into a single result file.

53| Using a PERL script, parse through the BLAT result files to group read matches based on uniqueness.

Preliminary analysis of RNA-seq data

54| Identify spliced reads by parsing the BLAT output file for reads with large values for gaps inserted into the target sequence ('tBaseInsert').

▲ **CRITICAL STEP** Set a minimum threshold larger than the number of base pairs in the smallest intron to remove sequences with very small gaps that likely represent artifacts in the sequence reads, or errors/differences with the published genome. Spliced reads can be written to a separate output file for further analysis. For *S. pombe*, a cut-off of 20 bp is acceptable.

▲ **CRITICAL STEP** Sequence reads starting and ending with homopolymeric tracts may be split by BLAT to match distant homopolymeric tracts in the reference genome. Such 'artifactual introns' can be hundreds of kilobases in size and can be filtered by setting a maximum value for 'tBaseInsert' values. Before exclusion of these reads as artifacts, they should be examined for the presence of homopolymeric tracts.

55| Assign expression scores for annotated genomic elements using a PERL script to parse all mapped reads.

56| Length-normalize the expression scores calculated in Step 55 by dividing either the number of reads in the element, or the sum of scores for each base-pair position in the element, by the length of the element.

57| If using multiple growth conditions, expression scores calculated can be normalized for the total amount of mappable sequence in each run.

58| Identify novel exons using PERL script to find spliced reads (output file from Step 54) that have one end of the spliced read in a known exon, but where the other end is not in an annotated exon.

59| Calculate the splicing efficiency for known and unknown introns.

60| Extract transcript boundaries using PERL scripts.

61| Visualize RNA-seq data along genome annotation data using R and a combination of various R/Bioconductor packages (i.e., tilingArray, GRID graphics and so on) on a computer with at least 4 GB of RAM.

? TROUBLESHOOTING

● TIMING

The timing of this protocol is highly dependent on the research goals and on the model organism being used. The time requirements of different steps are indicated in **Figure 1**. With a simple model organism (i.e., yeast), it is possible to prepare the cDNA required for sequencing in 2–3 d. The machine time is a fixed period of time, often 3–7 d, although improvements in next-generation sequencers continually reduce this time. The most substantial amount of time required is for the data analysis stage of the protocol. Even the most trivial aspects of the protocol, such as mapping reads to a reference genome, can take days, depending on the computational resources available and the amount of data generated. It would be highly advisable for researchers to ensure that sufficient computation resources and expertise are available to maximize the use of the data obtained.

? TROUBLESHOOTING

Troubleshooting advice can be found in **Table 1**.

TABLE 1 | Troubleshooting table.

Step	Problem	Possible reason	Solution
23	Poor quality RNA	Contamination of reagents with RNase	Prepare new buffers and reagents taking care to prevent contamination with RNases. Use only diethyl pyrocarbonate (DEPC)-treated water for preparing solutions
		Degradation of RNA due to growth conditions during preparation	Some *Schizosaccharomyces pombe* growth conditions (e.g., meiosis) make it difficult to isolate large amounts of high quality RNA during cell lysis. In these cases, Step 24 should not be considered optional. It may also be necessary to test other, more rapid, techniques for RNA collection
28	Poor yield of mRNA from enrichment kit	Elution buffer was allowed to cool below 70 °C before elution	Collection tubes can be placed into a 70 °C heating block while applying the elution buffer. A larger volume of elution buffer (which will cool more slowly) can also be used
30	Poor yield/quality of cDNA	Reverse transcriptase kit reagents are suboptimal or protocol was not followed exactly	Large fragment (Klenow) of *Escherichia coli* DNA polymerase I is highly heat labile. It should always be stored in a nonfrost-free freezer. Precooling of incubation blocks is recommended
46	High error rate in sequence reads	Possible hardware issues with Genome Analyzer	Verify with GA operations manual that all parameters are within normal tolerances
			If machine operation is normal, consideration should be given to divergence of reference genome and strain being used. Many initial sequencing errors may not have been detected and individual lab strains may have diverged over time
61	Very low level of coverage of 5′ end of long genes	Insufficient depth of sequencing	Oligo dT priming of the room temperature reaction will have the disadvantage of providing low coverage at the 5′ ends of genes as the reverse transcriptase enzyme will fall off before fully extending. A more processive enzyme may help reduce this effect, otherwise random priming of cDNA should be done after removal of rRNA

ANTICIPATED RESULTS

The RNA-seq protocol outlined here is an extremely sensitive and comprehensive technique for characterizing transcriptomes. For a simple model organism such as fission yeast, a single run of current NGS machines will generate several 100-fold coverage of the transcribed portion of the genome. This high sequencing depth actually obviates the need to do any subtractive normalization of the RNA to overcome inherent expression differences in genes to characterize genes expressed at very low levels. In addition to allowing accurate and quantitative measurement of gene expression, which have been shown to compare well with microarray measurements, RNA-seq also allows simultaneous discovery of novel exons and genes[1]. Finally, the RNA-seq data allows the validation of known introns, discovery of novel introns as well as quantitative measurements of splicing efficiency. **Figures 2–4** show examples of data based on the output of the protocol described above.

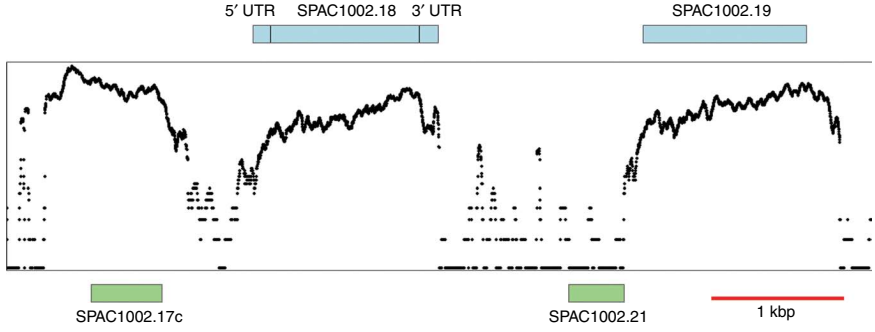

Figure 4 | Example of annotated output from data analysis steps of the protocol. Sequence expression scores for each nucleotide position for an ~6 kb region on chromosome I of *Schizosaccharomyces pombe* are shown. Strand specific annotation is shown above and below the nonstrand specific sequence scores indicated in the middle panel. A 1 kb line in the bottom right indicates the scale for the figure.

Given the rapid evolution of next-generation sequencing technologies, the exact approach described above will soon likely be supplanted by more efficient methodologies, but the output of the process and overall framework for analysis will likely remain unchanged.

ACKNOWLEDGMENTS We thank Dr. J.-R. Landry for critical reading of the manuscript. Research in the Bähler laboratory is funded by Cancer Research UK and by PhenOxiGEn, an EU FP7 research project.

AUTHOR CONTRIBUTIONS All authors contributed extensively to the work presented in this paper.

Published online at http://www.natureprotocols.com/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Kapranov, P., Willingham, A.T. & Gingeras, T.R. Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.* **8**, 413–423 (2007).
2. Mercer, T.R., Dinger, M.E. & Mattick, J.S. Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* **10**, 155–159 (2009).
3. Carthew, R.W. & Sontheimer, E.J. Origins and mechanisms of miRNAs and siRNAs. *Cell* **136**, 642–655 (2009).
4. Marguerat, S. & Bähler, J. RNA-seq: from technology to biology. *Cell Mol. Life Sci.* published online, doi:10.1007/s00018-009-0180-6 (27 October 2009).
5. Wilhelm, B.T. & Landry, J. RNA-seq—quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48**, 249–257 (2009).
6. Wilhelm, B.T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239–1243 (2008).
7. Mardis, E.R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402 (2008).
8. Lyne, R. *et al.* Whole-genome microarrays of fission yeast: characteristics, accuracy, reproducibility, and processing of array data. *BMC Genomics* **4**, 27 (2003).
9. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008).
10. Quail, M.A. *et al.* A large genome center's improvements to the Illumina sequencing system. *Nat. Methods* **5**, 1005–1010 (2008).
11. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
12. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**, 523–536 (2008).
13. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S. & Weissman, J.S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
14. Li, H. *et al.* Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *Proc. Natl. Acad. Sci. USA* **105**, 20179–20184 (2008).
15. Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37**, 123 (2009).
16. Croucher, N.J. *et al.* A simple method for directional transcriptome sequencing using Illumina technology. *Nucleic Acids Res.* published online, doi:10.1093/nar/gkp811 (8 October 2009).
17. Furuno, M. *et al.* Clusters of internally primed transcripts reveal novel long noncoding RNAs. *PLoS Genet.* **2**, e37 (2006).
18. Quinlan, A.R., Stewart, D.A., Strömberg, M.P. & Marth, G.T. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat. Methods* **5**, 179–181 (2008).
19. Rougemont, J. *et al.* Probabilistic base calling of Solexa sequencing data. *BMC Bioinformatics* **9**, 431 (2008).
20. Whiteford, N. *et al.* Swift: primary data analysis for the Illumina Solexa sequencing platform. *Bioinformatics* **25**, 2194–2199 (2009).
21. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
22. Denoeud, F. *et al.* Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* **9**, R175 (2008).
23. Hahn, D.A., Ragland, G.J., Shoemaker, D.D. & Denlinger, D.L. Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly Sarcophaga crassipalpis. *BMC Genomics* **10**, 234 (2009).
24. Yassour, M. *et al.* Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc. Natl. Acad. Sci. USA* **106**, 3264–3269 (2009).
25. Toth, A.L. *et al.* Wasp gene expression supports an evolutionary link between maternal behavior and eusociality. *Science* **318**, 441–444 (2007).
26. Trapnell, C. & Salzberg, S.L. How to map billions of short reads onto genomes. *Nat. Biotechnol.* **27**, 455–457 (2009).
27. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
28. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
29. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
30. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
31. Rumble, S.M. *et al.* SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol.* **5**, e1000386 (2009).
32. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).