Characterizing Transcriptomes from High-Throughput Sequencing Data

Thesis submitted for the degree of "Doctor of Philosophy"
by
Moran Yassour

Submitted to the Senate of the Hebrew University
July 2012

This work was carried out under the supervision of
Prof. Nir Friedman and Prof. Aviv Regev

Acknoledgements

First and foremost I wish to thank my great advisers Nir Friedman and Aviv Regev. I couldn't have chosen better "academic parents", who were always there for me with good advice, and knew how to lead me to the next step. I would like to thank Nir for teaching me scientific thinking, and how to understand (and criticize) papers, from merely reading their abstract and figures... I would like to thank Aviv for teaching me how to work, juggle multiple projects at the same time, while still remembering exactly what is going on where, and how to finalize them all. I thank Nir for teaching me to ask many questions, and Aviv for letting me know when to stop.

I thank all the past and current members of both the Friedman lab at the Hebrew University and the Regev lab at the Broad Institute. Specifically, I would like to thank Tommy Kaplan who took me under his wing, and taught me much of what I know today. Ariel Jaimovich and Naomi Habib for being great colleagues, but more importantly great friends along the way. Jenna Pfiffner and Dawn Thompson, for teaching me the experimental side of our work, and bearing with me on all my questions during this process.

I thank Joshua Levin, Brian Haas and Manfred Grabherr for being great partners to work with. Leslie Gaffney and Lauren Solomon for tremendous graphic assistance and teaching.

I thank Nati Linial and Rotem Sorek, who were always there when I needed them; Hanah Margalit, who throughout my studies treated me as her student and helped in any way possible; Chad Nusbaum, who has only the sky for technology limitations; and Oliver Rando for many intriguing discussion early on, which helped me understand how a true collaboration works.

I wish to acknowledge the great financial help I have received during my PhD from the Canadian friends of the Hebrew University, the Larry H. Summers scholarship, Google scholarships and mostly the Clore Foundation scholarship.

Finally, I would like to thank my family: my parents and sisters who believed in me and supported me in many ways during this period. To Liron my loved one, for endless love and support, and for truly being interested in all that is going on in my side of the academia, and of course, for begin a great father to Oren, my greatest accomplishment during this period.

Abstract

The DNA in our cells contains our genetic hereditary information, and is literally the blueprint of our body. The functional units of the genome are regions of continuous DNA sequence, and are called *genes*. According to the Central Dogma of Biology, the DNA sequence of the gene is transcribed into a *messenger RNA* (mRNA), which is in turn translated to proteins, which perform most tasks in the cell.

All the cells in an organism share the same DNA, but there are dramatic morphological and functional differences between cells in various tissues and under different conditions. Many of these differences are mediated by regulation that determines which genes are "turned on".

Classically, regions in the DNA were considered as genes only if they encode proteins. Today, regions in the DNA that are transcribed to mRNA but do not encode proteins, and function at the RNA level are also considered genes, and are called *non-coding RNAs*. Antisense transcripts are a specific type of non-coding RNAs, that overlap a protein-coding gene on the opposite DNA strand. In this context, these are called the *antisense* and *sense* transcript, respectively. When the antisense gene is transcribed, it can down-regulate the expression of the sense gene.

One of the first steps in understanding a newly sequenced organism is to annotate its genes, which will enable us to predict its repertoire of proteins. Ultimately we would like to annotate the genes, find their genomic position, and understand when, why and how they are turned on and off. The simplest task is to first identify their genomic position. In some simple eukaryote organisms (like the budding yeast), the genome is very dense with genes, and the vast majority of them are not spliced. In mammals, however, the genes comprise an extremely small part of the genomic sequence. For example, in humans only 2% of the genomic sequence is protein coding, making the task of finding the genes in the sea of the genomic sequence far from trivial. Thus, sequencing the genome is only the first step in our journey, and additional steps are required for better understanding an organism. One approach to characterize all transcribed genes is to examine the collection of mRNA molecules in the cell (also known as the cell's *transcriptome*).

Experimentally defining the complete transcriptome of eukaryotic organisms has traditionally been a challenging task, most commonly using tiling microarrays or sequencing *Expressed Sequenced Tags* (ESTs). In recent years new sequencing technologies (*"next generation sequencing"* or *"high-throughput sequencing"*) have emerged. These technologies allow us to take a single sample and sequence tens of millions of short *reads*, at unprecedented high speed and low cost. These technologies open up intriguing possibilities in studying other aspects of the genome, like sequencing the entire transcriptome (an assay called *RNA-Seq*). Most studies have used RNA-Seq to quantify the expression levels of known genes, identify splice isoforms and refine gene boundaries. However, many studies depend on an existing annotation or sequenced genomes, limiting the ability of discovering novel transcripts and studying diverse organisms.

In my dissertation I present a series of studies on the development of technologies and tools for RNA-Seq analysis and their application in organisms ranging from yeast to mouse. I focus on different approaches I have developed for transcriptome reconstruction, from mapping-first ones that rely only on an available genome sequence, to Trinity, a method for *de novo* assembly of full-length transcripts without requiring a sequenced genome. In addition, I describe systematic approaches to assess the quality of RNA-Seq experiments for annotation and expression quantification, and how I use them in a comparative study on library construction methods for strand specific RNA-Seq.

Finally, I show how these approaches scale to organisms from yeasts to vertebrates, helping in genome annotation of newly discovered organisms from the *Schizosaccharomyces* clade, the identification of extensive regulated long antisense transcripts that are conserved across yeast species, transcriptome analysis in the *Bemisia tabaci* whitefly, for which the genome sequence is not available, and for the discovery of alternatively spliced isoform in mouse.

Contents

Chapter 1

Introduction

1.1 Central Dogma of Biology

The DNA in our cells contains our genetic hereditary information, and is literally the blueprint of our body. The *DNA* is a polymer of four types of *nucleotides*: adenine, cytosine, guanine and thymine (marked briefly “A”,“C”,“G” and “T” respectively). The functional units of the genome are regions of continuous DNA sequence, and are called *genes*. According to the Central Dogma of Biology, the DNA sequence of the gene is transcribed into a *messenger RNA* (mRNA), which is in turn translated to proteins. The DNA has two strands, of complementary sequence, and the location of each gene in the genome includes both its strand and its position in the genome. This will determine the sequence that encodes the protein. The proteins perform most tasks in the cell. For example, they detect extra-cellular signals, replicate the DNA in preparation for cell division, regulate which genes will be “turned on”, transcribe the genes, and so on.

All the cells in an organism share the same DNA, but there are dramatic morphological and functional differences between cells in various tissues and under different conditions. Many of these differences are mediated by regulation that determines which genes are “turned on”.

The DNA region of a gene has a few defined characteristics, and can be divided into several segments (**Fig. 1.1**). The region that encodes the protein is called the *Open Reading Frame* (or ORF) of the gene. During translation, each DNA triplet (a *codon*) is translated into one amino acid. The sequence of amino acids makes a protein. Although the gene is a consecutive region on the genome, there are regions in it that following transcription are cleaved out, and are not part of the ORF. This process is called *splicing*, the regions that are spliced out are

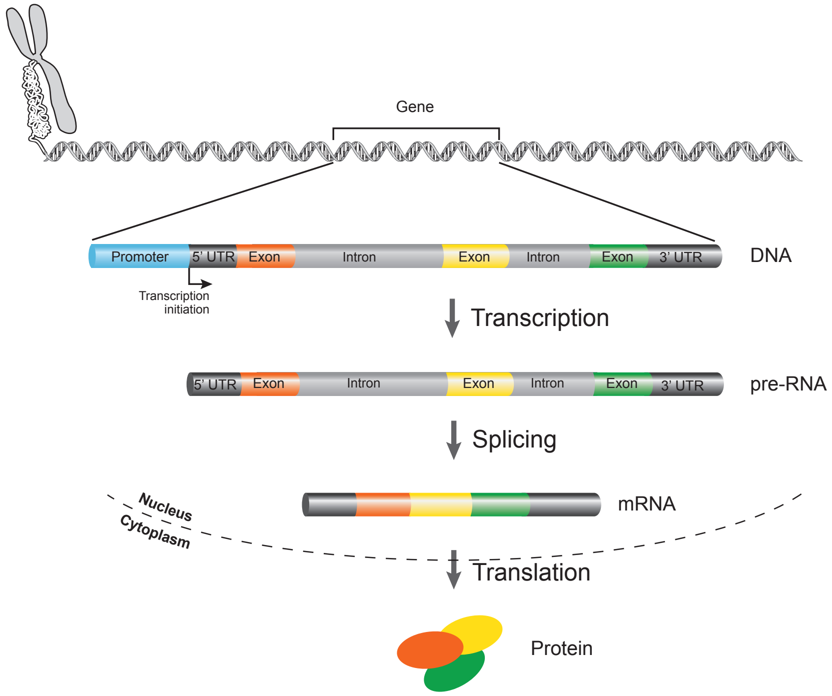


Figure 1.1: **The Central Dogma of Biology and the gene structure.** According to the Central Dogma of Biology, the DNA sequence of a gene is first transcribed to mRNA, then translated into a protein. In eukaryotes it is exported out of the nucleus after transcription and translated into a protein. The gene is comprised of several segments. The promoter (light blue) is the regulatory sequence to which transcription factors, and the RNA PolII bind. The Open Reading Frame is divided between a few exons (here in orange, yellow and green), interleaved by introns (light gray) which are spliced out. Flanking the ORF there are untranslated regions (UTRs, dark gray), which are transcribed to mRNA and exit the nucleus, but are not translated.

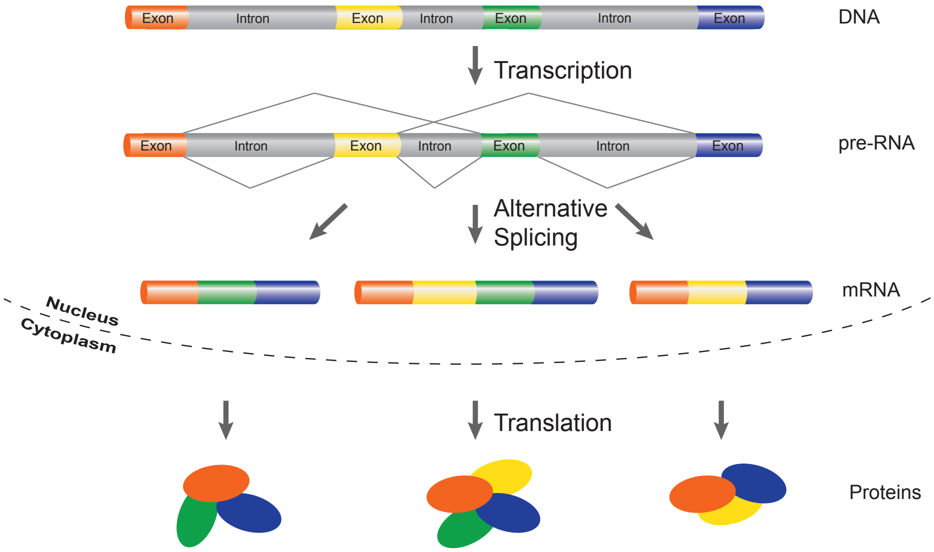


Figure 1.2: **Alternative splicing.** A single gene can give rise to several proteins using the alternative splicing process. This is done by choosing which exons (here in orange, yellow, green and blue) will be part of the final mRNA sequence, thus translated. Here I show three protein variants from a single gene with four exons.

called *introns*, and the regions that remain and will be a part of the ORF are called *exons*. The process of choosing which gene segments will be spliced out, thus which will compose the ORF, enables several proteins to be encoded in a single gene. This process is called *alternative splicing*, and each such alternative transcript is called a *splicing isoform* (**Fig. 1.2**).

In addition to the ORF, the transcribed region of the gene also contains additional flanking sequences, which are not translated, but carry some regulatory information. These regions are called the *Un-Translated Regions* of the gene, or the UTRs (**Fig. 1.1**).

1.2 Transcription Regulation & DNA Packing

The process of expressing a gene to its protein product is regulated at many points, including how accessible is the gene for transcription, the rate of transcription, mRNA degradation rate, initiation of translation, translation rate, post-translation modifications, etc. Keeping this complex picture in mind, transcription initiation plays a major role in the regulation of gene expression. For this to occur, a complex of proteins, known as the *RNA polymerase II*, has to bind to the transcription initiation site of the gene, and then to change to the right conformation to initiate the transcription. These steps are regulated by several processes.

Transcription factors are typically DNA-binding proteins that recognize specific sites in the DNA sequence (usually based on specific words that occur there). These factors either serve to recruit the RNA polymerase complex to the gene, inhibit such recruitment, or affect the rate by which the bound complex initiates transcription. Transcription factors are essential elements in modulating the expression of genes. Changes in the protein levels of transcription factor, or their state (i.e., post-translation modifications) can lead to changes in the expression of their target genes.

DNA Packing

The chromosomal DNA molecule of eukaryotic organisms is organized at several structural (3D) scales (**Fig. 1.3**). At the primary structure, chromosomal DNA is packed around *nucleosomes*, protein complexes that serve as beads around which the DNA is wrapped. A prototypical nucleosome is a complex of eight histone proteins, containing two copies of histones H2A, H2B, H3, and H4. About 147bp of DNA are wrapped around a single nucleosome forming slightly less than 2 turns (Luger et al., 1997). The position of the nucleosomes can serve a regulatory role by influencing the accessibility and hence activity of other proteins, most notably transcription factors and the transcription machinery (Ehrenhofer-Murray, 2004). Nucleosome positions can thus have a critical impact on transcriptional regulation and gene expression. Extensive recent work showed that nucleosome locations are determined by combination of several forces. First, certain DNA sub-sequence are preferable for wrapping around nucleosomes while others are rigid and exclude nucleosomes (Lowary and Widom, 1998; Anderson and Widom, 2001; Segal et al., 2006; Field et al., 2008; Mavrich et al., 2008). These constraints combined with the minimal distance between adjacent nucleosomes determine much of the nucleosome organization. Second, there are chromatin remodeling proteins that actively move nucleosomes to less preferable locations or evict them (Rando and Ahmad, 2007; Whitehouse et al., 2007). Finally, other proteins, such as transcription factors, compete with nucleosomes on the binding in particular sites.

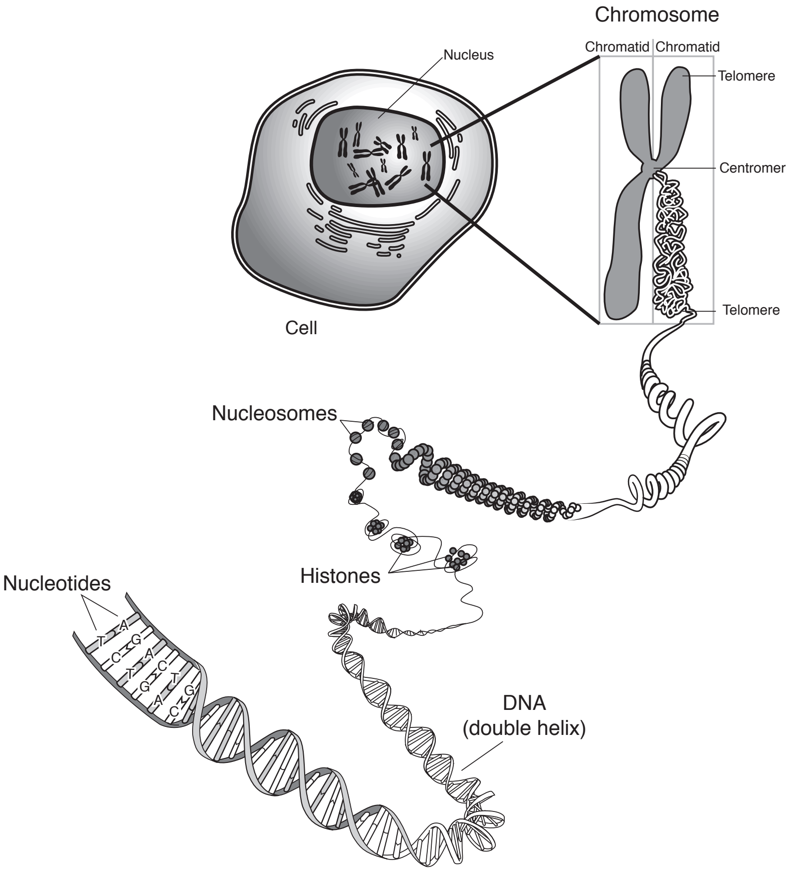


Figure 1.3: **The DNA sequence and Chromosomal packing.** Adapted from the National Human Genome Research Institute. The DNA sequence comprises of four nucleotides (adenine, cytosine, guanine and thymine), and has the 3-dimensional structure of a double helix. At the basic level, the DNA is packed around nucleosomes, which are histone protein complexes. The structure of the DNA wrapped around the nucleosomes is then further compacted and condensed to fit the small volume of the cell's nucleus.

1.3 Coding and non-coding genes

Classically, regions in the DNA were considered as genes only if they encode proteins. Until the early 80s there were two main additional classes of genes, that do not encode proteins and are functional at the RNA level. These are the *transfer RNA* (tRNA) and *ribosomal RNA* (rRNA) genes, involved in the translation process. In the recent decade there has been a tremendous increase in the discovery of functional RNA molecules, which are called in general *non-coding RNAs*, as they do not encode proteins (Bertone et al., 2004; Carninci et al., 2005; Rinn et al., 2007; Guttman et al., 2009). These RNA molecules can function through a variety of mechanisms, and can be folded into a three dimensional structure, which facilitates their function, and also increases their stability. For example, tRNAs carry amino acids to their corresponding codons in the mRNA. Each tRNA molecule folds into a structure that binds the codon on one side, and the relevant amino acid on the other. Some non-coding RNAs act as scaffolds to recruit the assembly of proteins, like the TERC RNA that serves as the template for the telomerase complex (Zappulla and Cech, 2006). Others, like the miRNA class, form a RNA-RNA double strand by hybridizing to their target mRNA (He and Hannon, 2004). This is one way to post-transcriptionally regulate the activity of a gene, as this can either (a) result in the degradation of these RNA molecules; or (b) prevent the mRNA from being translated.

Antisense transcripts are a specific type of non-coding RNAs. As explained above, the DNA has two strands, and each gene is located on a specific strand of the sequence. In some cases, we can find a non-coding gene that overlaps a protein-coding gene, only on the opposite strand. In this context, these are called the *antisense* and *sense* transcript, respectively. When the antisense gene is transcribed, it can down-regulate the expression of the sense gene. There are a few mechanisms suggested for this down-regulation (Faghihi and Wahlestedt, 2009): (1) through the formation of a RNA-RNA double strand, as explained above (**Fig. 1.4**); (2) the machinery that transcribes the antisense genes physically interferes with the sense transcription machinery and prevents it from transcribing the sense gene, hence less mRNA of the sense gene is available for translation; and (3) transcription of the antisense gene leaves histone marks on the chromatin that repress the transcription of the sense gene.

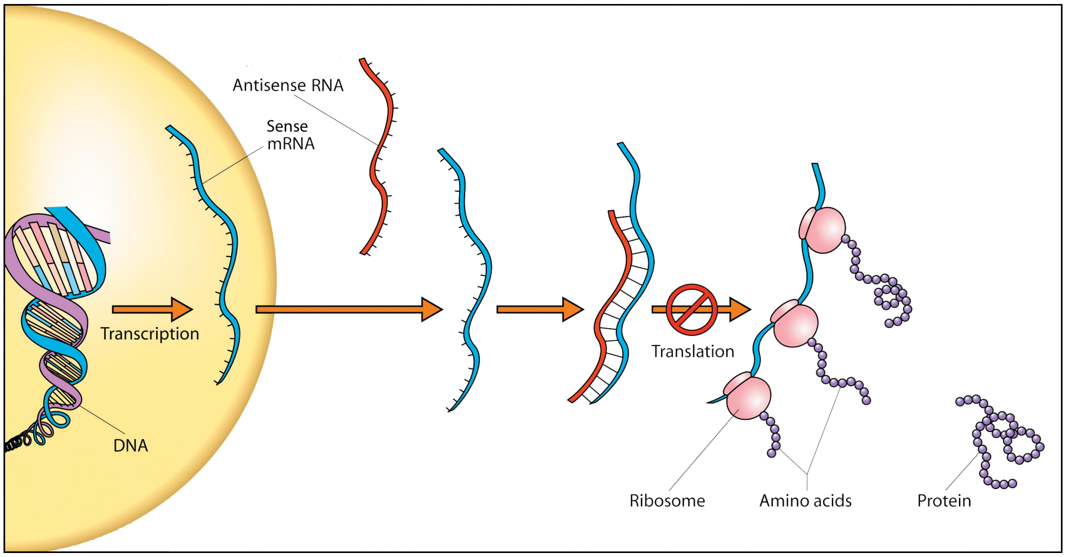


Figure 1.4: **Antisense transcription.** Adapted from Robinson (2004). The Central Dogma of Biology is presented for the sense gene as it is transcribed to mRNA (blue) and then translated to protein (purple). The antisense RNA (red) can form a RNA-RNA double strand, and prevent the sense gene from being translated. Other forms of down-regulation include increased mRNA degradation, and transcription interference of the sense gene.

1.4 Gene Discovery

One of the first steps in understanding a newly sequenced organism is to annotate its genes, which will enable us to predict its repertoire of proteins. By comparing the predicted proteins to all known proteins, we can better understand how this organism lives, regulates its behavior, and copes with different stress conditions. This comparison can also teach us which components are unique to this organism, and can shed some light on different mechanisms that have not been modeled before.

Ultimately we would like to annotate the genes, find their genomic position, and understand when, why and how they are turned on and off. The simplest task is to first identify their genomic position. In some simple eukaryote organisms (like the budding yeast), the genome is very dense with genes, and the vast majority of them are not spliced. In mammals, however, the genes comprise an extremely small part of the genomic sequence. For example, in humans only 2% of the genomic sequence is protein coding, making the task of finding the genes in the sea of the genomic sequence far from trivial. Thus, sequencing the genome is only the first step in our journey, and additional steps are required for better understanding an organism.

Although genomes and the genes they contain differ greatly between different organisms, some common rules are universal (except for few special organisms). All open reading frames share some characteristics that can help us locate them. There are special codons to specify the beginning and end of the proteins, these are called the *start-* and *stop- codons*, respectively. There are computational approaches that use this knowledge of the special codons, and other properties of codons (like conservation) to discover the genes, given the genomic sequences (Stanke and Waack, 2003; Majoros et al., 2004).

Even if we have a good computational tool for identifying the ORFs, there are still caveats to this approach: (1) low specificity, since it finds many sequences that have the start and stop codons but are not transcribed (*spurious ORFs*), at least in the examined conditions; (2) the sensitivity of finding single exon genes is fair, but it decreases dramatically at the multi exon genes; (3) this approach will fail to identify non-coding genes, as it searches for ORFs; (4) it finds only the ORF and not the entire gene sequence that includes the UTRs; and (5) we need to have the reference genome sequence of our model organism to perform this search.

A different approach is to examine the collection of mRNA molecules in the cell (also known as the cell's *transcriptome*). Most commonly this is done by either tiling microarrays or sequencing the transcriptome, and in both cases the RNA sequences are first transformed to their *complementary DNA* sequence (cDNA).

Microarrays

Building on existing reference sequence and properties of hybridization (the process by which one-stranded DNA molecules bind to the complementary sequence) allowed the design of DNA microarrays to detect the presence and quantity of specific mRNA sequences (Bertone et al., 2004; David et al., 2006). Each sequence on the array is called a *probe*, and we can design the array to hold overlapping probes for the genomic region of interest at a given resolution. The cDNA sequences are then hybridized with the array. Probes of the array that complement the sequences in the cDNA will by hybridized and identified. The major caveat of this technology is that we still rely on having a sequenced reference genome. We need to know the sequences of the regions we are interested in for designing the array.

Transcriptome Sequencing

In the sequencing approach we harness existing DNA sequencing technologies to sequence the cDNA library. In a pioneering work in the early 90s, Craig Venter and colleagues describe how they sequence parts of a human cDNA library (Adams et al., 1991). These sequence parts are called *Expressed Sequence Tags* (ESTs).

Although only parts of the cDNA are sequenced, with sufficient amount of ESTs and a reference sequence of minimal quality, we can annotate the coding regions of a genome. However, sequencing ESTs is a very long and laborious process. Each cDNA part is inserted into a bacterial clone, and each clone is grown to a colony, so we have an amplified and homogenous population. Finally, each colony is sequenced using primers for the known flanking sequences of the bacterial clone, generating sequences in the range of a few hundred basepairs.

1.5 Next Generation Sequencing

In recent years new sequencing technologies (*"next generation sequencing"* or *"high-throughput sequencing"*) have emerged. These technologies allow us to take a single sample (small amount of liquid with DNA fragments) and sequence tens of millions of short *reads*, at unprecedented high speed and low cost. Depending on the particular technology, these reads represent 30-300bp off the end of the fragment or off the two ends of the fragment (paired-end sequencing). The main breakthrough in this technology is that the amplification and sequencing is done in parallel for all fragments. The double stranded DNA sample is first fragmented into ∼300bp long pieces, and then these pieces are spread on a glass surface. The amplification is performed while the fragments are connected to the glass, and generates many copies of the same fragment, in close proximity to the location of the original fragment (these are referred to as *clusters*). The clusters are homogenous, as they contain many copies of the same fragment, and if they are distinct enough, the sequencing machine can sequence all the clusters simultaneously.

In this dissertation I focus on the Illumina (Solexa) platform that currently provides sequenced reads of length 32-150bp. The most common application of Illumina sequencing is to derive a host of sequenced reads from a DNA sample of interest, identify them by mapping to a finished reference genome, and deriving biological insight relevant to the measured sample. For example, by re-sequencing the DNA of specific individuals we can find the differences from the reference genome (Nielsen et al., 2011). Moreover, they open up intriguing possibilities in studies of other aspects of the genome. For example, as we are interested in characterizing the transcriptome we can sequence the cDNA library, which represents all the mRNA molecules present at our sample. This assay is called *RNA-Seq*, and can be used not only to characterize the transcripts, but also to quantify the expression levels of all genes and their different splicing isoforms (**Fig. 1.2**).

There are two major approaches in analyzing high throughput sequencing data: (1) mapping-first and (2) assembly-first (**Fig. 1.5**).

Mapping-first approach

In the *mapping-first* approaches, we rely on having a sequenced reference genome at some minimal quality. The first step is to find the genomic location that each read originated from. This is done by mapping the read sequences to the given genomic reference, and can be performed by a variety of aligner methods (Kent, 2002; Langmead et al., 2009; Li and Durbin, 2009). The second step is to generate coverage plots, that represent how many reads originated from each position in the genome. Finally, we can examine these coverage plots to identify patterns associated with transcribed regions (**Fig. 1.5**).

Assembly-first approach

Assembly-first approaches do not rely on a sequenced reference genome. Instead, they start by assembling all sequenced reads. After we have assembled the reads, if we do have a sequenced reference genome, we can map the assembled sequences to it. We can then examine the sequences assembled by predicting their protein sequences, and comparing to known genes.

DNA Assembly

There is a long history of DNA assembly methods, as part of the effort to sequence many genomes (Zerbino and Birney, 2008; Gnerre et al., 2010; Li et al., 2010). All these methods build on the initial concept of using de-Bruijn graphs for DNA assembly originally suggested by Pavel Pevzner (Pevzner, 1989). There are two main challenges in this field: (1) how to generate the enormous de-Bruijn graphs

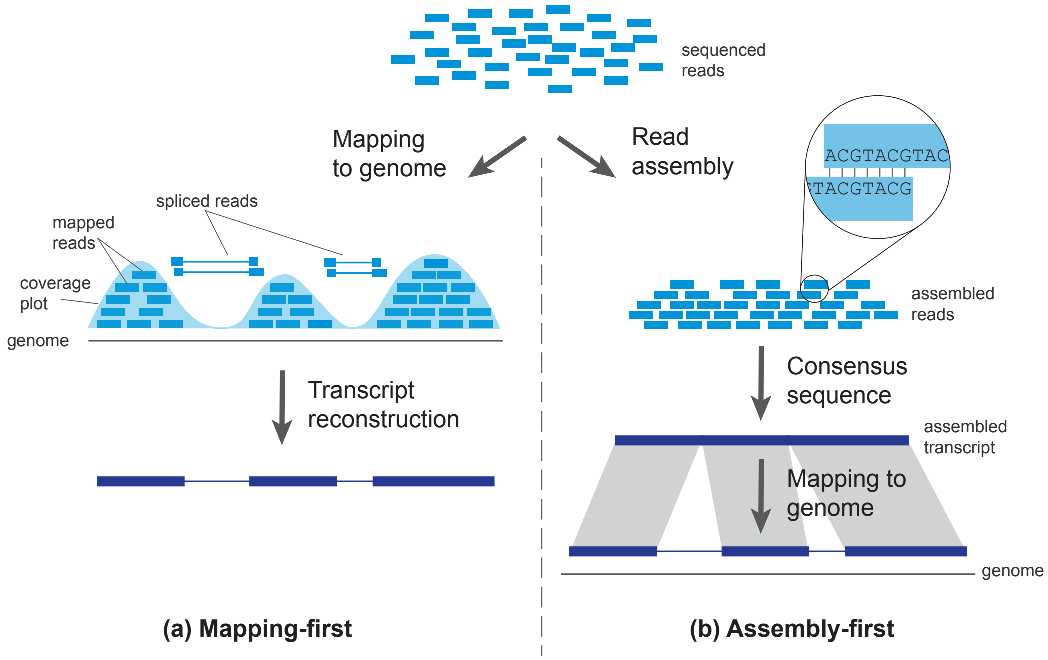


Figure 1.5: **Transcriptome assembly approaches.** **(a)** In the mapping-first approach the reads (blue rectangles) are first mapped to the reference genome, and together with the mapping of the spliced reads, coverage plots are calculated. These functions hold the number of reads mapped to it for each position in the genome. From the plots we can reconstruct the transcript and its structure. **(b)** In the assembly-first approach, the reads are first assembled by finding reads that share overlapping sequences (see insert). The assembled transcript is then extracted and in the presence of a sequenced reference genome, it can be mapped to reconstruct its structure.

that represent our data; and (2) once the graphs are built, how to extract a single coherent long DNA sequence. The first challenge is usually solved by using extremely high memory computers (*e.g.* with 512GB memory), which allows only established institutes designated for this task to perform such assemblies. To address the second challenge, there are many methods that traverse the graphs, and output their most probable sequence (Zerbino and Birney, 2008; Gnerre et al., 2010; Li et al., 2010).

RNA assembly

The task of assembling RNA-Seq reads shares these challenges, but adds on new challenges. As we have tens of millions of reads, the challenge of generating the graphs also exists in RNA-Seq assembly methods, and can be relaxed by using heuristic approaches. The underlying assumption in RNA-Seq assembly is that ideally, each gene or gene family should assemble separately. This implies there is no single enormous graph, and the challenge is to accomplish this, even if we do not know a priori which read originated from which transcript. One way to do this is to first use a greedy method for the crude assembly of the reads to linear sequences, and then combine these sequences based on their similarity (Grabherr et al., 2011). Another approach, in the case we have a sequenced reference, is to first map the reads to the genome and then assemble them based on this mapping (Trapnell et al., 2010; Guttman et al., 2010).

Finally, assembling RNA-Seq reads is different from assembling DNA-Seq reads, for additional two reasons: (1) We do not expect uniform coverage of the genome, as we have about four orders of magnitude difference in the expression levels of genes; and (2) In DNA assembly we expect to have a linear assembly graph (except for repetitions). In RNA assembly, the different isoforms will generate non-linearity in the assembly graph, and unlike in the DNA case, there is no single right answer, as we would like to capture all possible transcripts.

There is a tradeoff in choosing between the mapping-first and the assembly-first approaches, as each approach has its pros and cons (Haas and Zody, 2010). First and foremost, mapping-first methods require a sequenced reference genome. In addition, mapping-first approaches heavily rely on the software for mapping the short reads to the genome. There are a few challenges in aligning the reads to the genome, including handling tens of millions of reads, taking into account that some reads might originate from a few genomic loci, and doing all this in an efficient manner. In addition, there can be discrepancies between the read-

and the reference sequence. On the one hand, sequencing is known to introduce errors (roughly, at a ∼0.1%-1% rate depending on the technology), thus the sequence outputted by the sequencing machine might not be accurate. On the other hand, there is no true reference, as there is some level of diversity in our sample and not all cells in our sample have the exact same reference sequence. For example, there could be a *single nucleotide polymorphism* (SNP) in one of the genes, and while the majority of our sample, and the reference sequence, have a "T" in that position, a non-trivial portion of our sample has "C" there. When we try to map a read with a "C" to this position, we would count this position as a mismatch, while it could just belong to the second, less frequent variant. Finally, mapping the reads that originate from the splicing junction (also referred to as *spliced reads*) is difficult, as we do not know the position within the read where this splicing occurs. The way current aligners handle this problem is by examining all reads that have not been mapped in full, and mapping pieces of them in the hope to find that different pieces would map to the different exons. The problem of mapping spliced reads has become more and more relevant, as the reads are getting longer we observe reads that have up to 4 splice junctions within their sequence. However, if we rely on currently known annotation, we can align the RNA reads to all possible transcripts, thus overcome this problem but introduce a new challenge of assigning each read to a single transcript. To conclude, the mapping first approach relies on a high quality sequenced reference genome, and performs well if we rely on known annotations or study a relatively simple transcriptome (not too spliced).

On the other hand, although the assembly-first approach does not rely on a sequenced reference genome, it has other caveats. The assembly process itself is challenging, as we do not know a priori which reads originate from which transcripts. Ideally, we would divide our read set into groups, each group corresponds to a transcript, and assemble all reads within a group. Instead we have to assemble all the tens of millions of reads as a single group, and in theory each gene (or gene family) could assemble separately. In addition, taking into account the orders of magnitude difference in transcript abundance and the sequencing errors discussed above, we are faced with a new challenge. The erroneous versions of some highly expressed genes are more abundant than the lowly expressed genes in our sample. The assembly method should be able to tell these two cases apart, by filtering all the sequencing errors, considering the abundance of their alternative variants. An additional advantage of the assembly-first method, is that if we do have a reference genome sequenced, mapping the longer sequences is much easier, and can overcome many splicing junctions. To conclude, the assembly-first approach is appropriate if the reference genome is fragmented, or not sequenced at all. In addition, it has a great added value in studying complex, highly spliced, transcriptomes, and in cases where we do not rely on current annotations. Such *ab-initio* work is of great importance to the annotation of genomes for species on which we know relatively little or where the genome has massive genomic aberrations and rearrangements, as occurs in many tumor tissues.

1.6 Research Goals

The advent of high-throughput sequencing opens new opportunities to study transcriptome profiling and gene expression in a genome-wide and unbiased manner. Harnessing the full power of these technologies poses significant analytical challenges both in processing the raw data and its biological interpretation. My goal is to develop methodologies to address this challenge and apply them to several central problems in molecular biology. I have the following specific objectives:

Computational platform

Develop a computational framework for processing high-throughput sequencing in both the mapping- and assembly-first approaches.

Library construction methods

Understand the differences in the various RNA-Seq library construction methods, and find the ideal protocol for each task (*e.g.,* characterization vs. expression measurements).

Transcriptome characterization

Develop methodology for processing sequencing results from RNA-Seq assays to define the repertoire of transcripts, including their exact boundaries, strand specificity, splicing isoforms, and abundance. My emphasis will be on *ab initio* methodology that assumes we do not know the genes structure in advance.

Chapter 2

Paper: *Ab initio* Construction of a Eukaryotic Transcriptome by Massively Parallel mRNA Sequencing

Moran Yassour*, Tommy Kaplan*, Hunter B. Fraser, Joshua Z. Levin, Jenna Pfiffner, Xian Adiconis, Gary Schroth, Shujun Luo, Irina Khrebtukova, Andreas Gnirke, Chad Nusbaum, Dawn-Anne Thompson, Nir Friedman and Aviv Regev

In *Proc Natl Acad Sci U S A.*, 2009

*These authors contributed equally.

Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing

Moran Yassour[a,b,1], Tommy Kaplan[a,c,1], Hunter B. Fraser[b], Joshua Z. Levin[b], Jenna Pfiffner[b], Xian Adiconis[b], Gary Schroth[d], Shujun Luo[d], Irina Khrebtukova[d], Andreas Gnirke[b], Chad Nusbaum[b], Dawn-Anne Thompson[b], Nir Friedman[a,2], and Aviv Regev[b,e,2]

[a]School of Computer Science and Engineering, The Hebrew University, Jerusalem, 91904, Israel; [b]Broad Institute of Massachusetts Institute of Technology and Harvard, 7 Cambridge Center, Cambridge, MA 02142; [c]Department of Molecular Genetics and Biotechnology, Faculty of Medicine, The Hebrew University, Jerusalem 91120, Israel; [d]Illumina, Inc., 25861 Industrial Boulevard, Hayward, CA 94545; and [e]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142

Communicated by Eric S. Lander, The Broad Institute, Cambridge, MA, December 18, 2008 (received for review October 14, 2008)

Defining the transcriptome, the repertoire of transcribed regions encoded in the genome, is a challenging experimental task. Current approaches, relying on sequencing of ESTs or cDNA libraries, are expensive and labor-intensive. Here, we present a general approach for ab initio discovery of the complete transcriptome of the budding yeast, based only on the unannotated genome sequence and millions of short reads from a single massively parallel sequencing run. Using novel algorithms, we automatically construct a highly accurate transcript catalog. Our approach automatically and fully defines 86% of the genes expressed under the given conditions, and discovers 160 previously undescribed transcription units of 250 bp or longer. It correctly demarcates the 5′ and 3′ UTR boundaries of 86 and 77% of expressed genes, respectively. The method further identifies 83% of known splice junctions in expressed genes, and discovers 25 previously uncharacterized introns, including 2 cases of condition-dependent intron retention. Our framework is applicable to poorly understood organisms, and can lead to greater understanding of the transcribed elements in an explored genome.

computational biology | RNAseq | next generation sequencing | transcriptome profiling | *Saccharomyces cerevisiae*

Experimentally defining the complete transcriptome of eukaryotic organisms has traditionally been a challenging task, involving large, costly, and slow experimental efforts for sequencing of ESTs and full-length cDNA libraries. Unlike the genome, RNA transcripts are not present at equimolar concentrations, and are typically expressed in a context-specific manner. Thus, despite the fact that the genomes of >1,000 species have been sequenced, only few transcriptomes have been extensively characterized.

Recent advances in massively parallel sequencing technology (1, 2) offer new and powerful approaches to the study of transcriptomes. Recent studies (3–7) have shown that, by sequencing the mRNA content of cells, one can quantify the expression levels of known genes (by counting how often sequences from a given gene are observed) and refine their boundaries. For example, Nagalakshmi *et al.* (3) studied the *Saccharomyces cerevisiae* transcriptome by mapping reads to the location of known genes to quantify expression, and to known splice sites to measure their occurrence. Similarly, Mortazavi *et al.* (5) studied the mouse transcriptome by mapping reads to known exons and known splice junctions, as well as to "putative" junctions between known exons. Thus, in both cases (and in additional studies, see refs. 4–7) the analysis critically depended on existing annotation.

A more challenging problem is to define a transcriptome ab initio, based only on the unannotated genome sequence and millions of short reads from cDNA samples. Rapid and efficient methods to do so would transform our ability to define transcripts and study transcription in any genome. This ability would be particularly important in a new genome project involving phylogenetically isolated species and in cancer genome projects, where the genome annotation may fail to reflect pathological aberrations. The full goal would include: (*i*) identification of all regions encoding transcripts (coding and noncoding RNAs) in a given condition or cell type; (*ii*) demarcation of the 5′- and 3′- ends of transcripts; (*iii*) determination of splice junctions and identification of different splice variants; and (*iv*) identification of posttranscriptional transcript editing.

Here, we present a general approach to accomplish all of these goals, based solely on an unannotated genome sequence and data from a single sequencing run on an Illumina sequencer (2). To test our approach, we apply it to the budding yeast *S. cerevisiae*, and compare our ab initio results to the known transcript annotation (8). Our approach automatically and fully defines 86% of the genes expressed under the given conditions, and discovers 160 previously undescribed transcription units of 250 bp or longer. The approach correctly demarcates the correct 5′ and 3′ UTR boundaries of 86 and 77% of expressed genes, respectively. The method identifies 83% of known splice junctions in expressed genes, and discovers 25 previously uncharacterized introns, including evidence for 2 rare cases of condition-dependent "alternative splicing." Last, we use the data to quantify absolute and relative expression levels of each transcript, showing remarkable agreement with well-established microarray technologies.

Our results demonstrate that massive, cost-efficient, and fast sequencing can be used to accurately define and quantify a transcriptome ab initio. To evaluate the strength of our approach, we have refrained from using other sets of data and gene predictions methods. However, in many practical cases, these methods can be incorporated into a single bioinformatics pipeline for a more powerful outcome. This framework can be readily applied to study poorly understood organisms, for which only the genomic sequence is known.

Results

Sequencing the Budding Yeast Transcriptome. To define the budding yeast transcriptome ab initio, we generated cDNA libraries from poly(A)$^+$ mRNA from the budding yeast *S. cerevisiae* under 2 growth conditions: in rich medium (YPD) and after heat shock (HS). We used a cDNA preparation procedure that combines a random priming step with a shearing step (see *Materials and*

Author contributions: M.Y., T.K., H.B.F., J.Z.L., C.N., D.-A.T., N.F., and A.R. designed research; M.Y., T.K., H.B.F., J.Z.L., J.P., X.A., D.-A.T., N.F., and A.R. performed research; M.Y., T.K., J.Z.L., J.P., X.A., G.S., S.L., I.K., A.G., C.N., D.-A.T., and N.F. contributed new reagents/analytic tools; M.Y., T.K., and N.F. analyzed data; and M.Y., T.K., N.F., and A.R. wrote the paper.

Conflict of interest statement: G.S., S.L., and I.K. are from Illumina Corporation, which developed the sequencing technology we used, and thus have competing financial interests.

Freely available online through the PNAS open access option.

[1]M.Y. and T.K. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: nir@cs.huji.ac.il or aregev@broad.mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0812841106/DCSupplemental.

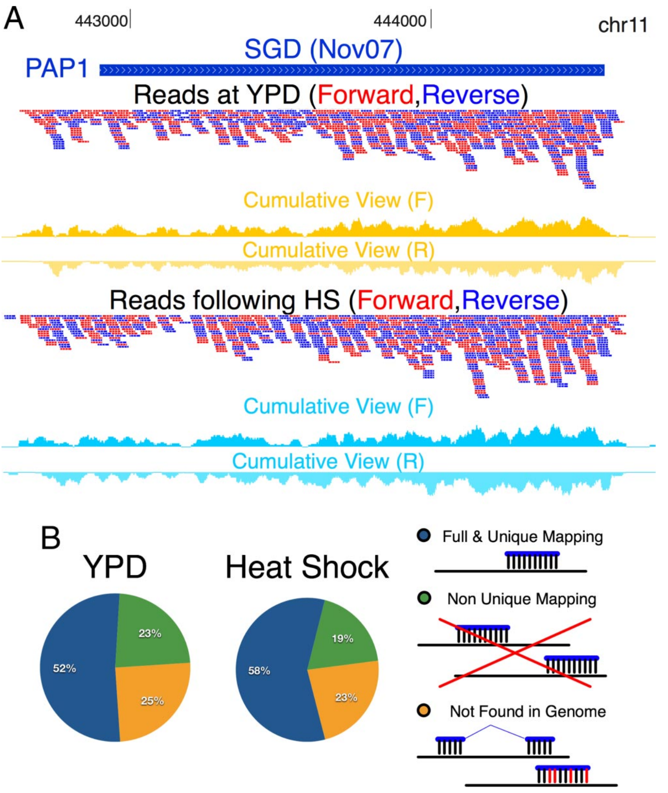


Fig. 1. Unbiased sequencing of the yeast transcriptome. (*A*) Distribution of reads mapped to the PAP1 locus. Shown are SGD annotations (downloaded at November 2007) (8), and mapped reads (red, W strand; blue, C strand). Additional tracks plot the cumulative number of reads covering each base position (yellow, YPD; light blue, HS). Full data can be accessed at http://compbio.cs.huji.ac.il/RNASeq, and is visualized using the University of California, Santa Cruz, genome browser (22). (*B*) Distribution of reads matched to the genome. Of the 26,050,414 reads sequenced in YPD (*Left*), 13,424,957 (52%, blue) were uniquely mapped to a single genomic locus, 6,144,595 (23%, green) were mapped to several locations, and 6,480,862 (25%, yellow) could not have been aligned, and were later used to detect splice junctions. Similar numbers were found after a HS (*Right*).

Methods). This approach has 2 benefits, which are essential for ab initio predictions. First, unlike other methods that provide a signal only in the 5′ or the 3′ end of transcripts, our method results in signal that covers the whole transcript (Fig. 1*A*). Second, for sequencing with short reads, random priming alone results in extensive nonuniformity in the start sites (9), whereas we obtain better uniformity.

We sequenced each library using an Illumina 1G Analyzer to generate 36-bp long reads. We obtained 25,043,976 reads from the YPD sample (2 biological replicates) and 11,776,251 reads after HS (see *Materials and Methods*). The entire experiment (RNA extraction, library preparation, and sequencing) required <14 workdays.

Then, we developed an accurate method to map reads to their genomic locations. The sequence matching approach used in previous studies (3, 4) may fail due to errors in the sequencing process or repetitive genomic regions (as a result of low-complexity or homology). Therefore, we developed a detailed probabilistic error model that scores the genomic matches of reads according to the position-specific probability of sequencing errors [see *Materials and Methods*; also, supporting information (SI) Fig. S1 and Dataset S1]. To minimize mapping errors, a read should match a specific genomic sequence at a strict threshold and should not match any other genomic location, even at a more relaxed threshold (see *Materials and Methods*). Applying this strategy to our data, we uniquely mapped 52% of the reads in YPD. We discarded an additional 23% of the reads that mapped to >1 genomic locus; this proportion is consistent with expectations due to genomic repeats (25.5% for 36-bp reads based on simulation). The remaining 25% reads did not map to any genomic locus at the required stringency (Fig. 1*B*). A minority is due to posttranscriptional modifications, such as splicing (see below). We obtained similar results with the reads in the HS experiment (Fig. 1*B*).

Ab Initio Construction of a Transcript Catalog for *S. cerevisiae*. We next developed a procedure to ab initio define all of the transcriptional units expressed under the 2 conditions, using only the mapped cDNA reads and the (unannotated) genome sequence of *S. cerevisiae* (Fig. 2*A*). Based on the current annotation of the yeast genome and microarray-based expression studies (8, 10), we expect 4,630 known genes to be expressed in YPD (at >0.2 transcripts per cell; see *Materials and Methods*). We started by identifying contiguous regions with a density of cDNA reads above a given threshold. Because genes are densely packed in the *S. cerevisiae* genome, such regions can span several genes. Thus, we developed a procedure that breaks these regions into segments of consistent read density, reflecting the expectation that transcript levels should be much more consistent within genes, than between genes (see *Materials and Methods* and Fig. 2*B*; also, Fig. S2). Last, we predicted transcription orientation based on different read densities between ends of genes (even in our relatively uniform libraries, there is a higher read density toward the 3′ end, which may be due to the library preparation protocol; see *Materials and Methods* and Fig. 2*B*). In total, we identified 6,248 segments, demarcating putative transcribed regions.

Before assembling a gene catalog, we next searched for splicing events. We analyzed the 25% of reads (9,212,859) that did not match the genome to identify those that may originate from splicing events. In such events, sequences from 2 exons that are separated in the genomic sequence are adjacent in the mature mRNA, yielding reads with a "gapped alignment" (Fig. 2*C*).

We developed an automatic method to systematically discover splice junctions. First, we identified reads with a gapped alignment, involving 2 sites of at least 10 bp each separated by at most 2 Kb (and together adding up to 36 bp). We required the same noise thresholds as before to filter out mismatches and nonunique matches (see *Materials and Methods*). Because we allow only a single gap, the probability of finding a spurious match is extremely low, although the precise gap location might be ambiguous by 1 or 2 base pairs, depending on the exact sequence at the gap boundaries. To eliminate spurious events, we required splice junctions to be supported by multiple observations. Specifically, we included all putative junctions that were either (*i*) supported by at least 5 independent reads (possibly starting at different locations; 243 junctions); (*ii*) supported by at least 3 independent reads and contain donor (5′) and acceptor (3′) splice site motifs (263 junctions; see http://compbio.cs.huji.ac.il/RNASeq); or (*iii*) supported by 2 independent reads and contain very strong splice motifs (13 junctions). This scoring allows us to resolve ambiguities, increase confidence in gapped reads, and assign an orientation to the junction (see *Materials and Methods* and Fig. 2*C*). The remaining putative junctions had little support and were discarded. In particular, shorter junctions are likely due to short deletions in the genomic DNA of the particular strain, consistent with Illumina sequencing of the DNA of this specific strain (data not shown). The resulting set had 285 junctions of 40 bp or longer. Notably, the majority of these junctions (243/285) were identified by the first criterion (5 strong junctions lack canonical splice site signals altogether; see http://compbio.cs.huji.ac.il/RNASeq), demonstrating the power of ab initio detection.

Joining the putative transcribed units based on the splice junctions, we built a final catalog of the yeast transcriptome in the 2 measured conditions (Fig. 2*A*; Dataset S2). This catalog includes 6,160 transcripts, 264 of them with at least 1 splicing junction.

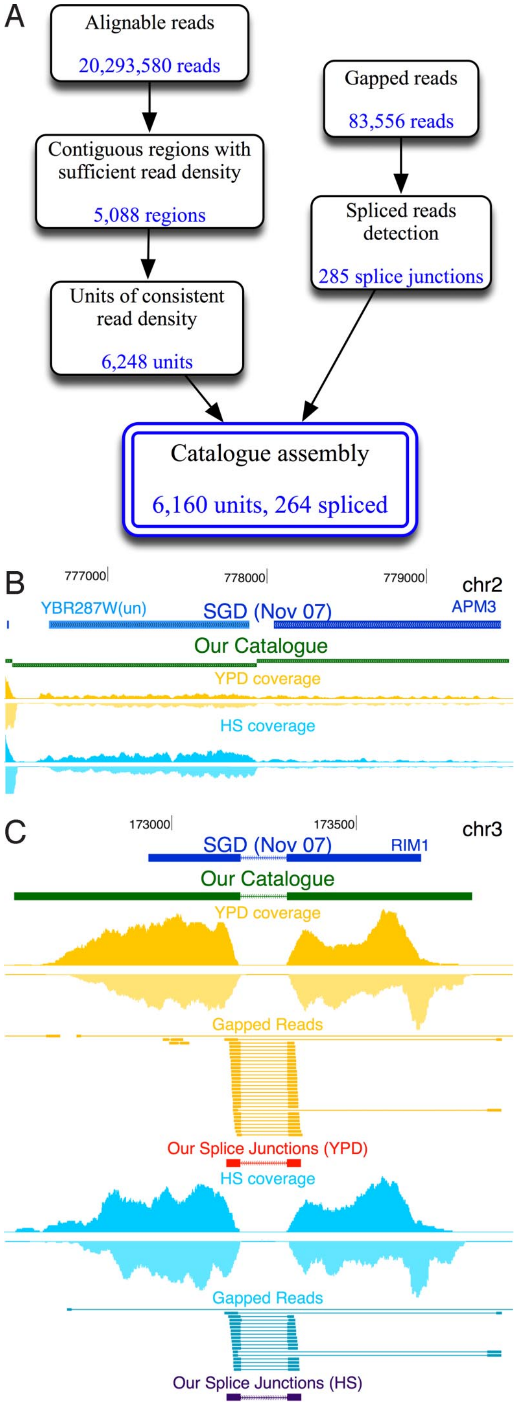


Fig. 2. Ab initio assembly of a transcript catalog. (*A*) Outline of steps in the catalog construction pipeline. (*B*) Segmentation of a contiguously transcribed region into 2 regions of distinct expression levels corresponding to the genes YBR287W and APM3. When using YPD reads alone, both genes exhibit similar coverage and thus cannot be segmented. However, in HS, they are differentially expressed, and hence by combining observations from both conditions the automatic segmentation procedure (see *Materials and Methods*) correctly separates them to 2 units. Tracks from top to bottom: SGD annotations (blue), our catalog (green), read coverage at YPD (yellow), and read coverage at HS (blue). (*C*) Detection of splice junctions. Full and gapped reads mapped to the RIM1 genomic locus. Tracks are as in *B*, together with gapped reads (connected segments), our putative splice junctions (in red and blue), including the junction orientations as estimated by donor and acceptor sequence motifs (arrows). As shown, our procedure identifies the exact coordinates and orientation of the known splice site.

Assessment of Transcription Units of the Catalog. We compared our ab initio catalog of transcripts with the current annotated transcriptional catalog of the yeast genome. Approaches based on sequencing of mRNAs cannot discover genes that are not expressed. Also, because we rely on short reads, we are limited to identifying transcripts in alignable (nonrepetitive) genomic regions. By using conservative thresholds, there are 5,437 (94%) known genes (classified as "verified" or "uncharacterized" ORF genes; see ref. 8) in the yeast genome that are "alignable" (at 50% coverage or more) with 36-bp reads, of which 4,784 are expressed in YPD (see *Materials and Methods*).

Overall, the ab initio transcriptional units in our catalog cover 99% of these expressed genes over >80% of the length of genes (Fig. 3*A*). For 86% of the genes, the transcriptional units fully cover the known genes (4096/4784; see Fig. 3*A*). For the remaining 13% of genes, the genes are largely covered, but correspond to multiple transcriptional units that have not been confidently connected (due to gaps or unevenness in coverage, particularly for highly expressed genes); this problem should be largely eliminated by connecting transcribing units through the use of "paired-end" reads, which are now becoming routinely available on the Illumina platform (11). Last, we correctly assigned orientation to 3,432 genes (84%), based solely on the pattern of increasing read density from 5′ to 3′-end. Overall, these results demonstrate that we can reconstruct the compendium of transcripts with great sensitivity and specificity.

Notably, our analysis indicates transcription from some "dubious ORFs" loci (62 of 206 expressed alignable dubious ORFs that do not overlap any other gene). In comparison, only 1% of nontranscribed loci based on ultradense tiling arrays (12) are covered by transcription units in YPD. This observation suggests that these are less likely to be spurious transcription events, and that some of these loci encode for functional transcripts (possibly noncoding RNAs).

The transcripts in our catalog assign the correct gene structure in terms of boundaries (and splicing; see below). Notably, because RNA-sequencing only samples short reads from transcripts, it has limited ability to accurately determine transcript boundaries in a highly compact genome (as compared with 5′ sequencing methods). Nevertheless, our transcript boundaries reasonably match several previous annotations of transcript boundaries in *S. cerevisiae*. These include the known annotations (SGD) as well as start site definitions based on previous full-length cDNA sequencing (13) and ultradense tiling arrays (12). In particular, our 5′ UTR positions match 80% of previous definitions within 50 bp, but have limited agreement in higher resolution [47% with Miura *et al.* (13); 22% with David *et al.* (12) in 10-bp resolution]. This latter result may be because our protocol likely misses 8–21 nt at the 5′ end of the transcript (14). Notably, we correctly predict the 3′ boundaries of 307 of 501 (60%) pairs of converging genes, and miss the boundary by at most 50 bp for an additional 58 cases (11%). Differential expression is a major contributor to correct detection. For correctly predicted pairs, the mean differential expression ratio is 8.5, whereas for those pairs that we cannot correctly differentiate, the mean differential expression ratio is 2.9. By considering the predicted ORFs within our transcripts, we estimate the typical lengths of 5′ and 3′ UTRs as 153 bp (SD of 145 bp), and 169 bp (SD of 142 bp), respectively (see http://compbio.cs.huji.ac.il/RNASeq; also, Dataset S3).

To our surprise, although 93% of our catalog corresponds to known genes (Fig. 3*B*; Dataset S2), we also discovered 160 transcription units of length ≥250 bp that did not overlap any previously annotated transcripts (Dataset S2; see ref. 8). Many of these units are clearly transcribed, for example, a ≈3,694-bp region at Chromosome 1, coordinates 196277–199970, that we also validated experimentally (see below). Many of these transcripts have supporting evidence in the raw data from hybridization to tiling arrays (129 units overlap; see ref. 12) and cDNA sequencing (92 units overlap; see ref. 13); although these previous studies did not report them as transcriptional units per se. Some of the units are differentially expressed between YPD and HS (Dataset S2). Most

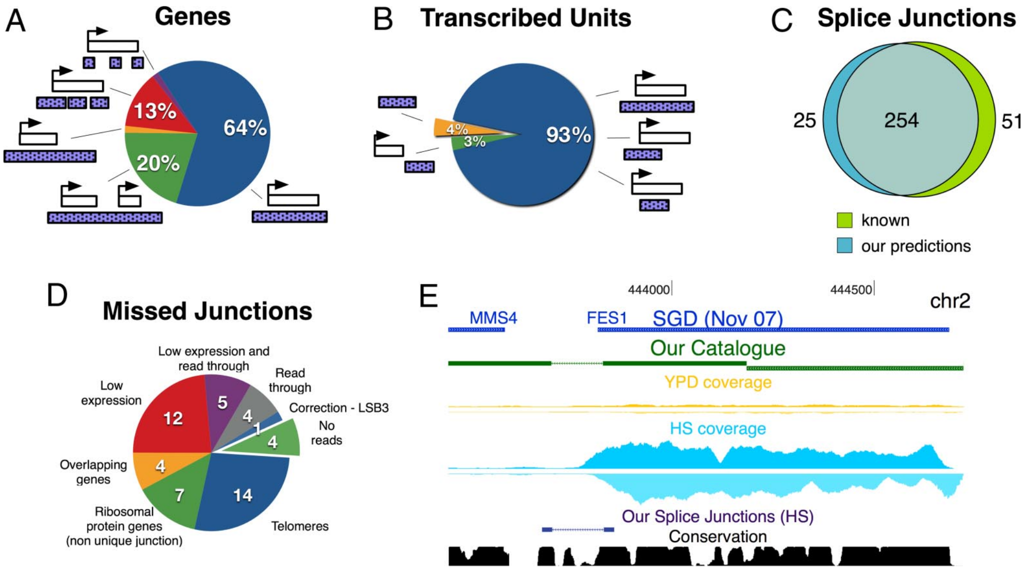


Fig. 3. Validation of the transcript catalog. (*A*) Coverage of the top 86% expressed genes by our predicted transcribed units, based on different patterns of coverage. (*B*) Relationship between found transcribed units and annotated transcribed features from SGD. In both *A* and *B*, white boxes denote genes, and purple boxes denote transcribed units. (*C*) Comparison of our putative splice junctions (blue) to known ones (green). (*D*) The 51 known introns missed by our predictions are partitioned into 8 categories. (*E*) Validation of splicing read-through in the gene FES1. Tracks are as in Fig. 2*C*, including the evolutionary conservation of each position across 7 yeast species (15).

notably, 12/160 novel units have are induced 10-fold or higher in HS vs. YPD, and 2 of those are not detected at all in YPD.

Overall, the previously undescribed units are mostly short (mean length of 713 bp, SD of 431 bp), and many are likely not coding for a protein. Several lines of evidence support this conclusion. First, the predicted ORFs are usually short (mean predicted ORF of 51 aa, SD of 19 aa, 20 units $>$80 aa; see Dataset S2), and do not match predicted or known proteins in other fungal species. Second, when sampling regions of the same length at random from intergenic regions, the median length of predicted ORFs is 146 aa, in contrast to the much shorter median length of predicted ORFs in these transcription units (48 aa). Last, relatively few of the units are evolutionary conserved (28/160 units $>$50% conservation; see ref. 15), which is not significant when compared with random ($P = 0.059$).

We experimentally tested and verified 4 of these novel transcripts by RT-PCR followed by sequencing. These included: (*i*) the novel ≈3,694-bp transcript discussed above (Chromosome 1, 196277–199970; see Fig. S3*A*); (*ii*) a transcribed pseudogene at Chromosome 15, coordinates 36742–38650 (Fig. S3*B*); (*iii*) a novel transcription unit at the YMR194C locus that spans both a dubious ORF (YMR194C-B) and the gene YMR194C-A (Fig. S3*C*); and (*iv*) a predicted 900-bp 3′ UTR for the FEN2 gene. In the latter 2 cases, the novel transcriptional units overlap, expand, or modify dubious ORFs or pseudogenes. For example, the novel transcription unit at the YMR194C locus also includes a 200-bp 3′ UTR past the predicted stop codon of YMR194C-A, suggesting a recent pseudogene.

Validation of Splice Junctions. Our splice site predictions are also highly accurate and sensitive, as compared with the known annotated junctions. The 285 ab initio detected splice junctions include most of the annotated junctions in the yeast genome (Fig. 3*C*; Dataset S2). We predict 254 (83%) out of 305 known junctions within 5-bp resolution. Of the 51 missed junctions, 21 are in non unique "unalignable" regions (telomeres and ribosomal protein genes), and 21 have very low read coverage (Fig. 3*D*). From the remaining 9 cases, we see read-through transcription in 4 undetected junctions, whose introns are matched by a significant number of reads (see http://compbio.cs.huji.ac.il/RNASeq), and determine a corrected location for 1 junction (LSB3 gene; see below). Thus, in only 4 of the 51 cases, we do not detect spliced reads for unknown reasons.

We also discovered 25 previously uncharacterized splice junctions that are not close to any annotated ones (one is an "artifact" caused by the HIS3 deletion in this strain). To study the implications of these splice junctions, we examined their effect on transcript structure. We found that 11 of the putative junctions are within annotated coding regions and affect the encoded protein, either by modifying existing introns, or by introducing additional ones (Dataset S3). For example, in the LSB3 gene, our putative intron is 24-bp shorter than the known one, adding 8 aa to the translated protein. When compared with other yeast species, the 8-aa stretch shows clear evolutionary conservation in the orthologous proteins (Fig. S4; see ref. 16); thus, it appears to be a conserved part of the protein.

In 6 of these junctions, we see evidence for alternative splicing (intron retention), because 3 junctions appear only in YPD and 3 only in HS (while taking into consideration the number of full reads aligned in both conditions; see http://compbio.cs.huji.ac.il/RNASeq). For example, in the MRM2 gene, the discovered intron is spliced out only in YPD; thus, creating a shorter protein, which perfectly aligns with orthologs of this gene in *Kluyveromyces lactis*, *Candida lusitaniae*, *Debopriya hansenii*, *Candida guilleromondi*, *Candida tropicalis*, and *Candida albicans*. In *C. albicans*, for example, the intronic sequence is completely missing from the genome, strongly supporting the functionality of this spliced form. Similarly, in the APE2 gene, the HS intron is slightly shorter, which creates a protein that is 6-aa shorter than the regular one. This modified protein has a domain that fits orthologs of this gene in *Saccharomyces paradoxus*, *Saccharomyces mikatae*, and *Saccharomyces bayanus*.

We experimentally tested 6 predicted splicing events and validated 4 of them (in the genes FES1, YMR148W, RPS22B, and AGA2) using RT-PCR and sequencing (Fig. 3*E*; Fig. S5). For example, in the FES1 gene, our catalog identified a previously uncharacterized intron with full reads through the splice junction and inside the intron, suggesting alternative splicing (Fig. 3*E*). In the spliced variant, the annotated stop codon is abolished and a later stop codon is introduced, resulting in a 10-aa extension. Validation by RT-PCR shows bands consistent with both the spliced and unspliced variants (sequencing of these bands confirmed the splice site). Another example of alternative splicing is the SUS1 gene, where, in addition to the 2 known introns, we also observe clear read-through at both junctions (Fig. S5*A*). Experimental validation confirms our predictions by revealing 3 bands, 2 bands consistent with just 1 intron spliced, and a stronger band consistent with both introns spliced out. A third example is an intron from the end of the snoRNA, SNR44, to the acceptor site of its hosting intron, inside RPS22B (Fig. S5*B*). All experimental validations were performed by RT-PCR followed by sequencing of the bands to verify the exact splice site. The predicted splice junctions that we could not validate may be in low-abundance or represent partial splicing.

Inferring Expression from Massively Parallel Sequencing. Having defined a gene catalog, we then examined the ability to infer quantitative expression levels from sequence abundance. We estimated the mRNA abundance of known annotated ORFs by calculating the average density of reads along each ORF and compared the results with expression data from microarrays. We converted the read densities per gene to rough assessments of absolute mRNA copy numbers per cell, using a conservative estimation of 15,000 transcripts per yeast cell (17). This analysis reveals at least 4 orders of magnitude differences in mRNA copy number among genes. For example, we find an average of 26 mRNA copies per cell for the top 5% of expressed genes, in contrast to an average of 0.0026 copies per cell for the bottom 5% (Fig. S6*A*). The top 5% of expressed genes in YPD account for 58% of the transcriptome, mostly comprised of transcripts encoding protein biosynthesis proteins and central carbon metabolism enzymes. Our mRNA copy number estimates are consistent with previous estimates using DNA microarrays (Pearson correlations of 0.67, $P < 10^{-300}$; 0.72, $P < 10^{-300}$; and 0.83, $P < 10^{-300}$, respectively; see Dataset S3 and Fig. S7) (3, 10, 18).

To calculate the relative expression level of each gene in HS vs. YPD, we compared the read densities in the 2 conditions. We compared the result with relative expression levels for the same mRNA samples inferred by commercial 2-dye microarrays (see *Materials and Methods*). Indeed, these ratios show strong agreement (Pearson correlation coefficient of 0.87, $P < 10^{-300}$; see Fig. S6*B*). These results were reproducible across sequencing and microarray replicates (Dataset S3; http://compbio.cs.huji.ac.il/RNASeq), consistent with recent studies (5).

Discussion

We set out to test whether it is possible to define a complete yeast transcriptome ab initio using only the (unannotated) genome sequence and massively parallel sequencing of cDNA from 1 or more experimental conditions. Our approach independently identifies the vast majority of known genes transcribed under the tested conditions, correctly infers splicing events, and detects the correct gene structure. Also, it corrects a number of current annotations and identifies previously undescribed transcriptional units and splice junctions, several of which we validated experimentally. Last, the method can also accurately quantify the expression levels of transcripts.

There are several crucial steps in the strategy. First, the creation of the cDNA fragments determines the transcript coverage. The laboratory protocol that we used here is only mildly biased toward the 3′ end of the transcript and thus provides efficient coverage throughout the transcript, allowing us to effectively assemble transcripts from short reads. Second, to accurately map reads to the reference genome, we created a sequencing noise model to limit the errors in mapping. Because the yeast genome has large unique regions, we can estimate the error model from the data without requiring calibration runs. Using this model, we correct for varying quality among batches. Unlike previous read mapping approaches (19), our method estimates the noise model separately for each batch; thus, it is more specific and, depending on the model, may allow for more mismatches if their probability is higher. Third, using the error model and sequence similarity tests, we reliably identify reads that are split between 2 genomic positions. This step is crucial for identifying splice junctions ab initio and defining correct gene structures, and is distinct from previous read mapping approaches (19).

Our approach has several limitations. First, we are unable to predict transcriptional units for low-copy transcripts and nonunique regions (e.g., at the telomeres). Although we can estimate relative expression of some low-copy transcripts, we cannot reliably determine splicing events or boundaries in such genes. We partially address this issue by creating libraries from YPD and HS. Deeper sequencing and libraries from additional conditions can further improve the completeness of the catalog. Second, we miss splicing events due to local nonuniqueness at the splice junction. We can alleviate this problem by sequencing either longer reads or paired-end fragments, both of which are becoming available (11). Last, our approach is limited in detecting and distinguishing antisense transcripts and differentiating between close divergent transcription units due to the lack of strand specificity. Although in most cases we can recover transcript orientation, we can further improve the predictions by constructing strand-specific cDNA libraries.

Unlike recent studies (3, 5), we demonstrate the use of massively parallel sequencing for complete, ab initio construction of a eukaryotic transcriptome, independent of any existing genome annotation. For example, Mortazavi *et al.* (5), and several similar approaches (3–7), use a step-wise mapping approach that relies on mapping reads to known gene models, exons and splice junctions. De novo discovery in these schemes is also limited, and is based on mapping reads to all possible combinations of known exons. Such approaches cannot detect splice junctions between unannotated exons. Also, they are not applicable to a genome for which there are poor (or no) gene predictions. In contrast, our approach searches for all the locations where a spliced version of an unaligned read can be mapped in the genome. Thus, our approach will be useful for both smaller more compact genomes, such as those of fungi or protists that often involve phylogenetically isolated groups for which there are poor gene predictions (20), as well as for aberrant cancer genomes.

Our work powerfully demonstrates the feasibility of constructing a transcriptome of an organism in a comprehensive, fast, and cheap way. To estimate the power of this approach, we conducted our analysis in isolation from any other source of data or gene prediction methods. Nevertheless, we anticipate that in many practical setups it can be powerfully combined with other gene prediction approaches. Applying our approach to explore the transcriptomes of less characterized organisms in an ab initio fashion, can have a significant impact on genomics studies.

Materials and Methods

Yeast Strains and Growth Conditions. ***HS experiment.*** The strain used was a derivative of the *S. cerevisiae* strain S288c (BY4741; see ref. 21). We grew 1-L cultures overnight in YPD medium (1% yeast extract, 2% peptone, 2% dextrose) to an OD_{600} of ≈1.0. The cultures were split and 1 flask was submerged in a 37 °C water bath and the other in a 22 °C water bath; 50-mL samples were harvested after 0 and 15 min.

RNA extraction and library preparation. Total RNA and polyA$^+$ RNA were isolated by using the RNeasy Midi Kit (Qiagen) and Poly(A) Purist kit (Ambion), respectively. Samples were quality controlled with the RNA 6000 Nano II kit of the Bioanalyzer 2100 (Agilent). Sheared cDNA libraries were created for 6 samples (22 °C, 0 min; 22 °C, 15 min; 37 °C, 15 min; 2 replicates per condition; 150 ng of polyA$^+$ RNA per sample). The cDNA was synthesized by using the SuperScript Double-Stranded cDNA Synthesis kit (Invitrogen) with SuperScript III (Invitrogen), 15-ng random hexamers (Invitrogen), and 20 units SUPERase·In (Ambion). Primer annealing was done at room temperature for 10 min followed by 1 h at 55 °C for first strand synthesis and 2 h at 16 °C for second strand synthesis; cDNA was sheared by sonication with 12 alternating cycles between "high intensity" (30 s; duty cycle, 20%; intensity, 10%; cycles per burst, 200) and "low intensity" (4 s; duty cycle, 5%; intensity, 10%; cycles per burst, 200) in the Frequency Sweeping mode (Covaris S2 machine). Adapters for Illumina sequencing were added following the instructions provided, except that 5 times less adapter mix was ligated to the cDNAs and PCR primers were removed by digestion with RecJ (New England Biolabs). Each library had an insert size of 60 to 110 bp. One lane of sequence (5.4 to 7.0 M reads) was generated for each sample on an Illumina 1G sequencer.

Genomic Mapping of Reads. ***Error model.*** We developed a detailed probabilistic model for scoring the quality of matching reads to the genome. Our score depends on the specific type of sequencing error made (e.g., genomic A sequenced as C) and its position within the sequenced read. Formally, the score of obtaining a read R originating from a genomic sequence G equals $\sum_{i=1}^{36}\log_2(\Pr(R_i|G_i,i))$, where i is the position within the read, R_i is the sequenced nucleotide at position i, and G_i is the nucleotide at the corresponding genomic position. To estimate the error parameters, we identified reads with up to 4 mismatches to highly unique regions of the genome. Then, we estimated the fraction of errors at each position for each genomic nucleotide (Fig. S1).

Mapping method. To map the sequenced reads to the genome with minimal errors, we devised the following strategy. Each read was compared with every possible 36-bp window in the genome and scored according to the error model above. We developed a procedure that uses suffix trees to efficiently finds all of the matches above a predefined threshold. To filter the matches, we require that the read matches the assigned genomic sequence with a threshold (−8.3) that assures correct mapping of 95% of reads (based on simulations). Also, to ensure uniqueness, we require the match to remain unique even when allowing a more relaxed threshold (−11.5).

Detection of Transcriptional Units. ***Segmentation of transcriptional units.*** To identify transcriptional units, we first artificially extended mapped reads to partially reconstruct the dsDNA segments they originated from. Because the segment size in our library varies between 60 and 110 bp, we chose a conservative approach and extended each read by an additional 40 bp (each read is now 76 bp). We then identified contiguously covered genomic regions. In many cases, these regions contained >1 gene, due to overlapping neighboring transcripts in the dense yeast genome. To refine these regions into single transcribed units, we developed an automated segmentation algorithm to fit the genomic patterns of mapped reads using piecewise linear regression (Fig. S2). Neighboring genes often exhibit different expression levels allowing an accurate partition. To achieve a coherent segmentation, we applied our algorithm to YPD and HS data simultaneously. This strategy also allows us to use the transcriptional differences of genes between the 2 conditions. For example, the 2 neighboring genes YBR287W and APM3 (Fig. 2*B*; Fig. S2) have similar expression levels at YPD; hence, preventing a proper segmentation to 2 transcription units. However, at HS, YBR287W is expressed in much higher levels than APM3, allowing us to position the boundary between the 2 genes.

Definition of nontranscribed loci. For a negative control, we applied a sliding window of 75 bp over the data of David *et al.* (12), and identified 892 loci that presented the lowest mRNA to genome signal in YPD.

Automated determination of orientation. As demonstrated in Fig. 2*B*, the typical density of reads is not completely uniform along the transcript with higher density toward the 3′ end. We use this pattern to estimate the orientation of each transcription unit. We use the slope of our piecewise linear fit to determine the orientation of each transcription unit. Specifically, we estimate a 95% confidence interval of the regressed slope parameters, and assign a forward or reverse orientation to the transcription unit if the entire interval is orientation-consistent (above or below zero, respectively).

Detection of splice junctions. First, we map gapped reads by searching for coordinated partial matches to 2 genomic loci within 2 Kb, each one of at least 10 bp (and together adding up to 36 bp). We require the same noise thresholds to filter out mismatches and nonunique matches. Specifically, we score each putative match with the score described above, allowing a single gap in the genomic sequence. Second, we calculate the position-specific scoring matrix (PSSM) score for each gapped read, according to the splice motifs we learned from the known introns (see http://compbio.cs.huji.ac.il/RNASeq). Third, we cluster gapped reads by the genomic location of their gaps. Each cluster defines a putative junction in the transcriptome, and is characterized by the number of supporting reads and the PSSM score of the junction. We assign orientation to each putative junction using these asymmetric PSSM motifs. We define a threshold over the PSSM log-odd scores (2.78), such that 95% of the known splice junctions (based on SGD annotations, October 2007) are identified in the correct orientation.

Definition of Alignable Expressed Genes. A genomic location is ''nonalignable'' if reads originating from that location will be mapped by our method to at least one other location in the genome; otherwise, we say that the location is alignable. We define a gene to be alignable if at least 50% of locations within its coding region are alignable. We define genes, known from previous studies to have at least 0.2 mRNA copies per cell (on average) (10), as ''expressed,'' reflecting 85% of the transcriptome at YPD condition.

Estimation of Gene Expression Levels. Using annotations from SGD (October 2007), we calculate the number of reads mapped to each coding region. We approximate the expression level of each gene by the average density of reads along the unique (alignable) part of the coding region. This measure is expressed in arbitrary units of number of reads per lane per 1-K base pairs, and is assumed to be proportional to the actual number of mRNA molecules per cell. Assuming a conservative estimation of 15,000 transcripts per cell (17), we can assess the expected number of copies for each gene. Relative expression levels (HS vs. YPD) are calculated by comparing the average density of each gene at the 2 conditions.

Relative Gene Expression Using Commercial Arrays. PolyA$^+$ RNA samples from one replicate each of the 37 °C, 15 min (HS) and 22 °C, 15 min (YPD reference) were labeled with either Cy3 or Cy5 by using a modification of the protocol developed by De Risi (University of California, San Francisco) and Rosetta Inpharmatics that can be obtained at http://www.microarrays.org. For the detailed modified protocol see http://compbio.cs.huji.ac.il/RNASeq. Four technical replicates of the HS samples were hybridized against the reference on commercial *S. cerevisiae* (S288C strain) 2-color 60-mer oligo Agilent arrays in the 4 × 44 K format (Agilent). After hybridization and washing per Agilent instructions, arrays were scanned by using a scanner (Agilent) and analyzed with a feature extraction software (Agilent).

Validation of Novel Transcription Units and Splice Sites. RNA from the HS and YPD reference samples was treated with TURBO DNA-free Kit (Ambion) to remove trace amounts of genomic DNA; cDNA was synthesized from this RNA by using a SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen). Assays were designed to detect predicted RNA species by the PCR. Reactions were performed under the conditions specified in the Amplitag gold polymerase product manual (Applied Biosystems) by using 10 ng of cDNA as template in a volume of 50 μL. For primer sequences, see http://compbio.cs.huji.ac.il/RNASeq. Products were amplified by using the following Thermocycler program: *i*, 95 °C for 5 min; *ii*, 95 °C for 30 s; *iii*, 56 °C for 30 s; *iv*, 70 °C for 45 s; go to step 2 for 40 cycles; *v*, 70 °C for 7 min; *vi*, 4 °C forever. PCR products were separated by using 3% Metaphor agarose (Cambrex) gels. The DNA fragments were isolated from the gel by using a QIAEX II Gel extraction kit (Qiagen). These fragments were cloned by using a TOPO TA Cloning Kit for Sequencing (with pCR4-TOPO) with One Shot TOP10 Chemically Competent *Escherichia coli* and PureLink Quick Plasmid Miniprep Kit (Invitrogen). Insert containing constructs were sequenced at the Massachusetts Institute of Technology core facility. Sequences were verified by using the BLAST function at the *Saccharomyces* genome database (www.yeastgenome.org/).

Supplementary Web Site. Raw data and additional notes and figures can be found at our supplementary web site (http://compbio.cs.huji.ac.il/RNASeq).

ACKNOWLEDGMENTS. We thank Ariel Jaimovich, Ilan Wapinski, Daphne Koller, Hanah Margalit, Jill Mesirov, Erin O'Shea, Oliver Rando, and especially Eric Lander, for comments and discussions; and Jason Funt, Toshiro Osumi, and the Broad Institute's Sequencing Platform for help with processing of Illumina sequence data. This work was supported by the Leibniz Research Center at the Hebrew University (T.K.), National Institutes of Health and Israel Science Foundation grants (to M.Y. and N.F.), a Human Frontiers Science Program grant (to D.-A.T.), funding from the Broad Institute (J.P.), and a Scientific Interface Career Award from the Burroughs Wellcome Fund (to A.R.).

1. Margulies M, *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380.
2. Bentley DR (2006) Whole-genome re-sequencing. *Curr Opin Genet Dev* 16:545–552.
3. Nagalakshmi U, *et al.* (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320:1344–1349.
4. Wilhelm BT, *et al.* (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453:1239–1243.
5. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621–628.
6. Cloonan N, *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods* 5:613–619.
7. Salehi-Ashtiani K, *et al.* (2008) Isoform discovery by targeted cloning, 'deep-well' pooling and parallel sequencing. *Nat Methods* 5:597–600.
8. Cherry JM, *et al.* (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res* 26:73–79.
9. Wang ET, *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456:470–476.
10. Holstege FC, *et al.* (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95:717–728.
11. Hillier LW, *et al.* (2008) Whole-genome sequencing and variant discovery in C. elegans. *Nat Methods* 5:183–188.
12. David L, *et al.* (2006) A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci USA* 103:5320–5325.
13. Miura F, *et al.* (2006) A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc Natl Acad Sci USA* 103:17846–17851.
14. D'Alessio JM, Gerard GF (1988) Second-strand cDNA synthesis with E. coli DNA polymerase I and RNase H: The fate of information at the mRNA 5′ terminus and the effect of E. coli DNA ligase. *Nucleic Acids Res* 16:1999–2014.
15. Siepel A, *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15:1034–1050.
16. Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449:54–61.
17. Hereford LM, Rosbash M (1977) Number and distribution of polyadenylated RNA sequences in yeast. *Cell* 10:453–462.
18. Liu CL, *et al.* (2005) Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLoS Biol* 3:e328.
19. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18:1851–1858.
20. Gardner MJ, *et al.* (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* 419:498–511.
21. Brachmann CB, *et al.* (1998) Designer deletion strains derived from Saccharomyces cerevisiae S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* 14:115–132.
22. Kent WJ, *et al.* (2002) The human genome browser at UCSC. *Genome Res* 12:996–1006.

Chapter 3

Paper: Comprehensive comparative analysis of strand-specific RNA sequencing methods

Joshua Z. Levin*, Moran Yassour*, Xian Adiconis, Chad Nusbaum, Dawn Anne Thompson, Nir Friedman, Andreas Gnirke, and Aviv Regev

In *Nature Methods*, 2010

*These authors contributed equally.

Comprehensive comparative analysis of strand-specific RNA sequencing methods

Joshua Z Levin[1,6], Moran Yassour[1–3,6], Xian Adiconis[1], Chad Nusbaum[1], Dawn Anne Thompson[1], Nir Friedman[3,4], Andreas Gnirke[1] & Aviv Regev[1,2,5]

Strand-specific, massively parallel cDNA sequencing (RNA-seq) is a powerful tool for transcript discovery, genome annotation and expression profiling. There are multiple published methods for strand-specific RNA-seq, but no consensus exists as to how to choose between them. Here we developed a comprehensive computational pipeline to compare library quality metrics from any RNA-seq method. Using the well-annotated *Saccharomyces cerevisiae* transcriptome as a benchmark, we compared seven library-construction protocols, including both published and our own methods. We found marked differences in strand specificity, library complexity, evenness and continuity of coverage, agreement with known annotations and accuracy for expression profiling. Weighing each method's performance and ease, we identified the dUTP second-strand marking and the Illumina RNA ligation methods as the leading protocols, with the former benefitting from the current availability of paired-end sequencing. Our analysis provides a comprehensive benchmark, and our computational pipeline is applicable for assessment of future protocols in other organisms.

Recent advances in massively parallel cDNA sequencing (RNA-seq) have opened the way for comprehensive analysis of any transcriptome[1]. In principle, RNA-seq allows analysis of all expressed transcripts, with three key goals: (i) annotating the structures of all transcribed genes including their 5′ and 3′ ends and all splice junctions[2–4], (ii) quantifying expression of each transcript[5,6] and (iii) measuring the extent of alternative splicing[7–11].

Standard libraries for RNA-seq do not preserve information about which strand was originally transcribed. Synthesis of randomly primed double-stranded cDNA followed by addition of adaptors for next-generation sequencing leads to the loss of information about which strand was present in the original mRNA template. In some cases, strand information can be inferred by subsequent computational analyses using, for example, open reading frame (ORF) information in protein-coding genes, biases in coverage between 5′ and 3′ ends[4] or splice-site orientation in eukaryotic genomes[4,10,11].

Nevertheless, direct information on the originating strand can substantially enhance the value of an RNA-seq experiment. For example, such information would help to accurately identify antisense transcripts, with potential regulatory roles[12], determine the transcribed strand of other noncoding RNAs, demarcate the exact boundaries of adjacent genes transcribed on opposite strands and resolve the correct expression levels of coding or noncoding overlapping transcripts. These tasks are particularly challenging in small microbial genomes, prokaryotic and eukaryotic, in which genes are densely coded, with overlapping untranslated regions (UTRs) or ORFs and in which splice-site information is limited or nonexistent.

Many methods have been recently developed for strand-specific RNA-seq, and they fall into two main classes. One class relies on attaching different adaptors in a known orientation relative to the 5′ and 3′ ends of the RNA transcript (**Fig. 1a**). These protocols generate a cDNA library flanked by two distinct adaptor sequences, marking the 5′ end and the 3′ end of the original mRNA. A second class of methods relies on marking one strand by chemical modification, either on the RNA itself by bisulfite treatment or during second-strand cDNA synthesis followed by degradation of the unmarked strand (**Fig. 1b**). Both modification methods essentially follow the standard protocol for RNA-seq with the exception of these marking steps.

Although standard RNA-seq largely relies on one protocol, the great diversity of published protocols for strand-specific RNA-seq poses several challenges. First, when conducting an experiment, researchers are challenged to identify a suitable protocol. Furthermore, if protocols vary considerably in their performance, the chosen method can dramatically affect the conclusions drawn from an experiment, confounding interpretation and comparison across studies. There is therefore a substantial need for a systematic evaluation of the performance of different protocols for strand-specific RNA-seq.

Here we present a comprehensive comparison of seven protocols for strand-specific RNA-seq. Using *Saccharomyces cerevisiae* poly(A)$^+$ RNA, we built a compendium of libraries using these

[1]Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, USA. [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]School of Engineering and Computer Science, Hebrew University, Jerusalem, Israel. [4]Alexander Silberman Institute of Life Sciences, Hebrew University, Jerusalem, Israel. [5]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [6]These authors contributed equally to this work. Correspondence should be addressed to J.Z.L. (jlevin@broadinstitute.org) or A.R. (aregev@broad.mit.edu).

RECEIVED 26 MARCH; ACCEPTED 20 JULY; PUBLISHED ONLINE 15 AUGUST 2010; DOI:10.1038/NMETH.1491

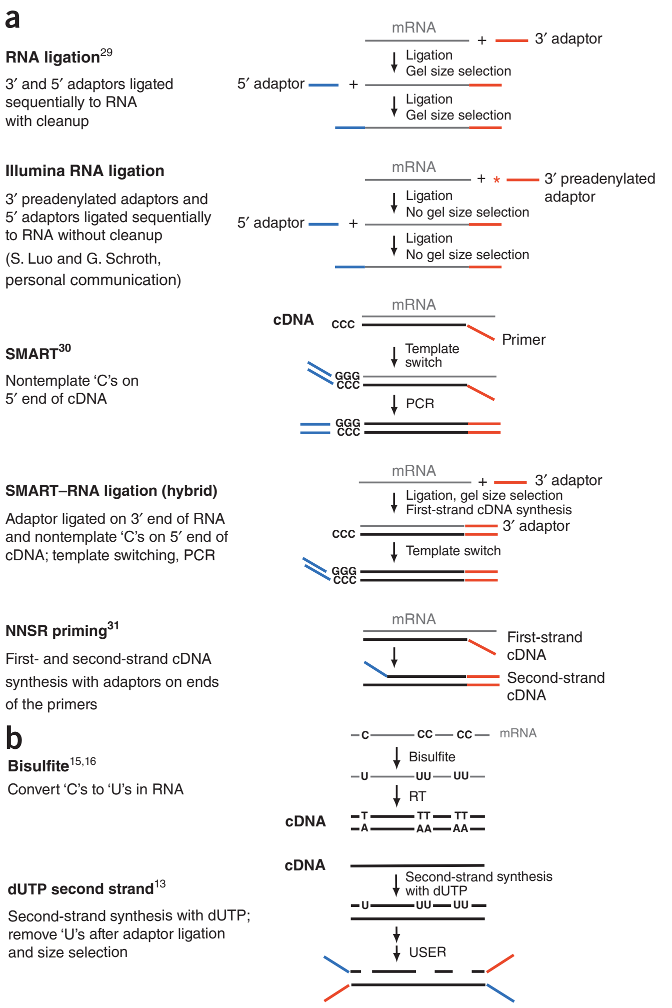

a
RNA ligation[29]
3′ and 5′ adaptors ligated sequentially to RNA with cleanup
mRNA
+
3′ adaptor
Ligation
Gel size selection
5′ adaptor
+
Ligation
Gel size selection
Illumina RNA ligation
3′ preadenylated adaptors and 5′ adaptors ligated sequentially to RNA without cleanup
(S. Luo and G. Schroth, personal communication)
mRNA
+ *
3′ preadenylated adaptor
Ligation
No gel size selection
5′ adaptor
+
Ligation
No gel size selection
SMART[30]
Nontemplate 'C's on 5′ end of cDNA
cDNA
CCC
mRNA
Primer
Template switch
GGG
CCC
PCR
GGG
CCC
SMART–RNA ligation (hybrid)
Adaptor ligated on 3′ end of RNA and nontemplate 'C's on 5′ end of cDNA; template switching, PCR
mRNA
+
3′ adaptor
Ligation, gel size selection
First-strand cDNA synthesis
CCC
3′ adaptor
Template switch
GGG
CCC
NNSR priming[31]
First- and second-strand cDNA synthesis with adaptors on ends of the primers
mRNA
First-strand cDNA
Second-strand cDNA
b
Bisulfite[15,16]
Convert 'C's to 'U's in RNA
C
CC
CC
mRNA
Bisulfite
U
UU
UU
RT
cDNA
T
TT
TT
A
AA
AA
cDNA
dUTP second strand[13]
Second-strand synthesis with dUTP; remove 'U's after adaptor ligation and size selection
Second-strand synthesis with dUTP
U
UU
UU
USER

Figure 1 | Methods for strand-specific RNA-seq. (**a**,**b**) Salient details for differential adaptor methods including RNA ligation[29], SMART[30] and NNSR priming[31] (**a**) and differential marking methods (**b**). USER, uracil-specific excision reagent. mRNA is shown in gray and cDNA in black. For differential adaptor methods, 5′ adaptors are shown in blue, and 3′ adaptors are shown in red.

protocols and sequenced each of them on an Illumina Genome Analyzer instrument to deep coverage. We developed a computational pipeline to assess each library's quality according to library complexity, strand specificity, evenness and continuity of coverage, agreement with known genome annotation and quantitative accuracy for expression profiling, in addition to considering the ease of laboratory and computational manipulations. We identified the dUTP and Illumina RNA ligation methods as the leading protocols, with the dUTP library providing the added benefit of the ability to conduct paired-end sequencing.

RESULTS

A comparison of strand-specific RNA-seq

We evaluated 13 stand-specific libraries. We constructed 11 libraries based on seven strand-specific RNA-seq methods (**Fig. 1**), including two variations for four of the methods. We also compiled comparable data for two published libraries: a dUTP library[13] and a library based on another (eighth) method from the differential adaptor class[14] (3′ split adaptor; **Supplementary Fig. 1**). Finally, we prepared a standard, non–strand-specific cDNA library to use as a control in these comparisons.

We explored two different variations for four of the seven methods to improve our libraries (Online Methods). These variations were the addition of actinomycin D to the 'not not so random' (NNSR) library protocol, two published variations of the bisulfite library protocol ('H' and 'S'; Online Methods[15,16]), different size-selection methods for the Illumina RNA ligation libraries and different reverse transcription primers for the dUTP libraries. We present results only for the 'S' bisulfite library because we found no substantial differences between the two libraries in our analyses.

We used each method to prepare a cDNA library for Illumina sequencing from *S. cerevisiae* poly(A)$^+$ RNA. We chose *S. cerevisiae* because this eukaryotic model organism has an exceptionally well-annotated genome, facilitating quality evaluations. We used paired-end Illumina sequencing for each library (Online Methods), except for the RNA ligation and Illumina RNA ligation libraries, which we sequenced only from the 3′ end of each cDNA because of the RNA adaptors used in these protocols. These approaches could be modified in the future to accommodate paired-end sequencing by changing the RNA adaptor and PCR primer sequences.

An analysis framework for assessing RNA-seq libraries

To compare the quality of the different libraries, we defined six assessment criteria (**Fig. 2**) implemented in a computational pipeline (Online Methods). These criteria were library complexity, defined as the number of unique reads (**Fig. 2a**); strand specificity, defined as the number of reads mapping to known transcribed regions at the expected strand (**Fig. 2b**); evenness and continuity of coverage at annotated transcripts (**Fig. 2c**,**d**); performance at 5′ and 3′ ends, defined as agreement with known end annotation (**Fig. 2d**); and performance in expression profiling, defined by sensitivity, linearity and dynamic range. With the exception of strand specificity, we compared each criterion to that for the control library. We focused on only one variation per method unless there were substantial differences in performance between variations. We provide the full evaluation results in **Supplementary Tables 1–2** and **Supplementary Figures 2–4**.

Equal sampling of reads enables direct library comparisons

We mapped each library's reads to the *S. cerevisiae* genome using Arachne[17]. For paired-end libraries, we mapped unique pairs with opposite orientations and an appropriate separation; for single-end libraries, we identified unique mappings for individual reads[17] (Online Methods).

The libraries had a broad range of yields, measured by the total number of reads and by the number of reads or paired reads mapping to a unique location (**Supplementary Table 1**). In this initial comparison, the dUTP library had the highest percentage of paired-end mapped reads (**Supplementary Table 1**). The Illumina RNA ligation–solid-phase reversible immobilization (SPRI) library, which we prepared using SPRI-based size selection, had a smaller percentage of unique reads than the Illumina RNA ligation library, which we prepared using gel-based size selection (35% versus 59%; **Supplementary Table 1**). This was likely due to the difficulty in physically removing cDNAs shorter than 76 base pairs with the SPRI method, resulting in the ends

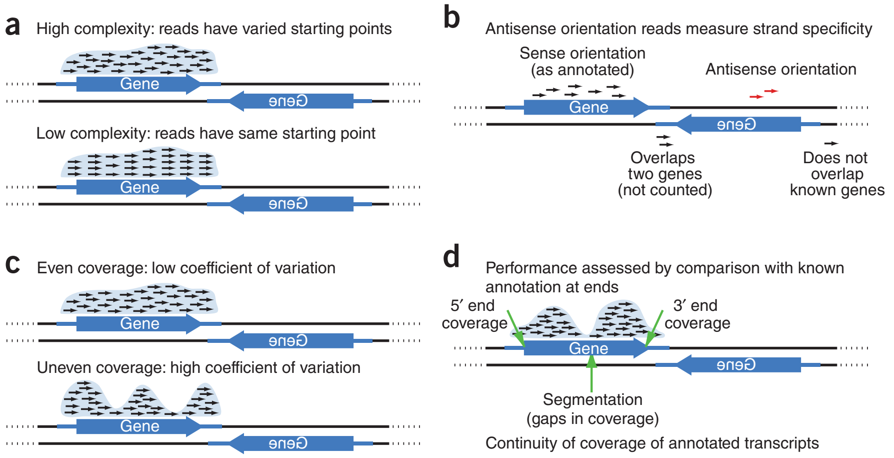

a
High complexity: reads have varied starting points
Gene
Gene
Low complexity: reads have same starting point
Gene
Gene
b
Antisense orientation reads measure strand specificity
Sense orientation (as annotated)
Antisense orientation
Gene
Gene
Overlaps two genes (not counted)
Does not overlap known genes
c
Even coverage: low coefficient of variation
Gene
Gene
Uneven coverage: high coefficient of variation
Gene
Gene
d
Performance assessed by comparison with known annotation at ends
5′ end coverage
3′ end coverage
Gene
Gene
Segmentation (gaps in coverage)
Continuity of coverage of annotated transcripts

Figure 2 | Key criteria for evaluation of strand-specific RNA-seq libraries. (**a**–**d**) Categories of quality assessment were complexity (**a**), strand specificity (**b**), evenness of coverage (**c**) and comparison to known transcript structure (**d**). Double-stranded genome with gene ORF orientation (blue arrows) and UTRs (blue lines) are shown along with mapped reads (black and red arrows, reads mapped to sense and antisense strands, respectively).

of sequencing reads containing an Illumina adaptor sequence that could not be aligned to the yeast genome. Indeed, when we trimmed these reads to 51 bases, the percentage of aligned reads improved dramatically (data not shown). Below, we report results only for the Illumina RNA ligation library, which we prepared using gel-based size selection.

Some of this variation in performance may reflect variation in sequencing yields between sequencing runs and lanes (**Supplementary Table 1**), unrelated to the library protocol. As many of our measures were sensitive to read quantity and length, we used sampling to obtain the same number of reads from each library (Online Methods). Unless specifically noted, we conducted all subsequent comparisons with 2.5 million sampled reads from each library. The 'switching mechanism at 5′ end of RNA template' (SMART) library had only 930,686 reads because of repeated poor yields, but with the exception of complexity, we obtained overall similar results when using the SMART reads 'as is' (without any compensatory calculations for there being fewer than 2.5 million reads) or when randomly resampling the same reads more than once to reach 2.5 million (data not shown). To compare libraries with different read lengths (51 or 76 bases in our libraries and 36 bases in published data), we sampled the first 36 bases of every read.

Complexity of single- and paired-end libraries

We next assessed the complexity of each library, defined as the number of distinct (unique) read start positions (**Fig. 2a**). A high complexity library, with many different start positions, is preferable as it does not suffer from 'jackpot' effects in fragment amplification or a strong bias in selection of fragment ends. Using single-end mapping (**Fig. 3a** and **Supplementary Table 2**), we observed the best complexity for the control library (42% unique) followed closely by the 3′ split adaptor method (42% unique), SMART (41% unique) and the published dUTP method (40% unique).

Single-read complexity calculations may overestimate the number of redundant cDNAs in a library. For paired-end libraries, we also estimated complexity as unique pairs of start and end positions (**Fig. 3b**), because cDNAs that have the same start site for one read can be distinguished based on a different start site for the other read in the pair. Comparing paired-end libraries by this measure, we found that the control and dUTP libraries performed best, with 88% and 84% unique paired reads, respectively. This demonstrates that paired-end sequencing substantially improves estimates of library complexity relative to estimates using only single reads.

Strand specificity across libraries

We measured the strand specificity of each library by comparing the mapped reads to the expected transcribed strand based on the known *S. cerevisiae* annotation (Online Methods). Based on recent studies[18], we conservatively assumed that most of the *S. cerevisiae* genes are not transcribed from the antisense strand and used the fraction of reads mapped to the opposite (antisense) strand of known transcripts as a measure of strand specificity (**Fig. 2b**, **Supplementary Table 2** and Online Methods).

Four of the protocols, RNA ligation, Illumina RNA ligation, dUTP and NNSR (with actinomycin D), performed best, whereas the SMART approach was the least strand-specific method, by a wide margin (**Fig. 4** and **Supplementary Fig. 5**). Only 0.47–0.63% of the reads mapped to the antisense strand for the four best performing methods. Notably, addition of actinomycin D dramatically improved the strand specificity of the NNSR method (**Supplementary Table 2**). Actinomycin D treatment cannot be used to improve the strand specificity of SMART because it inhibits both second-strand synthesis and template switching[19] (X.A. and J.Z.L.; data not shown).

Evenness and continuity of annotated transcript coverage

Using RNA-seq for effective transcriptome annotation, which includes transcript assembly[3,4], separating neighboring genes correctly and identifying full-length transcripts with correct 5′ and 3′ ends requires even, continuous and complete coverage along each transcript's length.

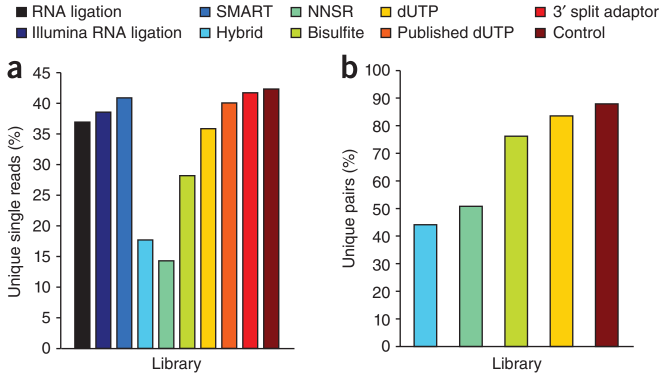


Figure 3 | Complexity of single- and paired-end libraries. (**a**,**b**) Percentage of unique reads mapping out of the total number of mapped reads, when considering only single-mapped reads (**a**; all libraries) or uniquely mapped pairs (**b**; only paired-end libraries).

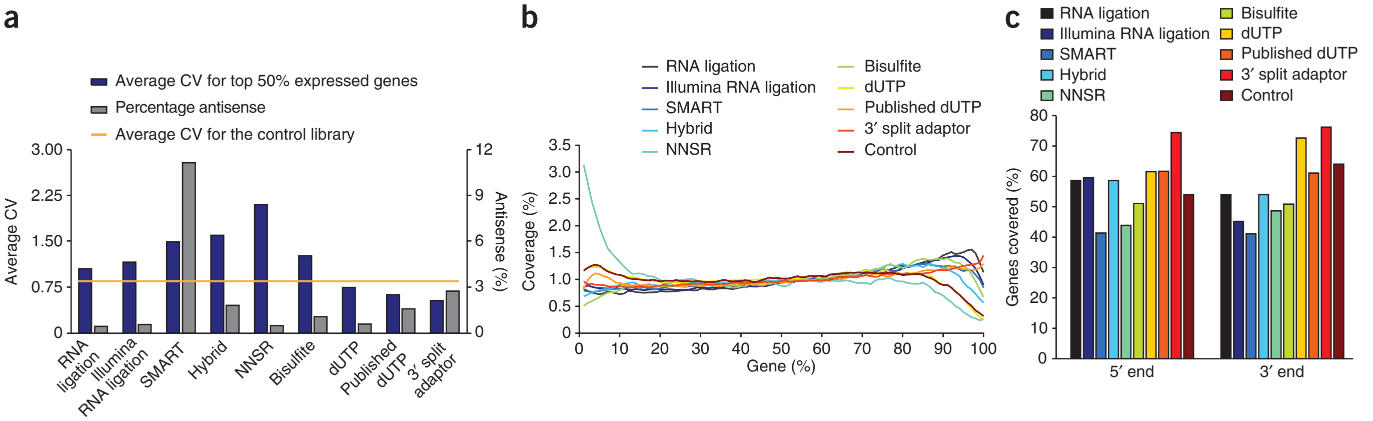


Figure 4 | Strand specificity and evenness of transcript coverage. (**a**) Strand specificity (percentage antisense) and evenness of coverage (average coefficient of variation (CV)) for all libraries. (**b**) Relative gene coverage at each percentile of a gene's length, averaged across all genes in each library. The 5′ end is on the left. (**c**) Percentage of genes with 5′-end and 3′-end coverage in each library.

To measure evenness of coverage for each library, we calculated the average of the coefficient of variation of gene coverage for the top 50% expressed genes (**Figs. 2c** and **4a**, **Supplementary Fig. 5** and **Supplementary Table 2**). We found the most even coverage for the 3′ split adaptor method[14] (average coefficient of variation, 0.54), closely followed by that for the dUTP approach (average coefficient of variation of 0.64 in the original dataset[13] and 0.76 in our hands).

We defined two measures of continuity of coverage. First, we counted the number of segments into which each known transcript was broken, where we defined a break as a stretch of at least five bases without read coverage (**Figs. 2d** and **5a** and **Supplementary Table 2**). We then averaged this measure across all genes, weighting by the relative expression of each gene (we expected low-expressed genes to be less covered and more segmented). The best performing methods by this measure were the 3′ split adaptor method[14] (2.29 segments per gene), the dUTP libraries (2.41 and 2.48 segments per gene with published data[13] and in our hands, respectively) and the Illumina RNA ligation libraries (2.61 segments per gene).

Second, we calculated the fraction of bases without coverage in each transcript (**Figs. 2d** and **5b–e** and **Supplementary Fig. 2**) and examined the distribution of this fraction at different expression levels, as defined by pooling data across libraries (Online Methods). As expected, in all libraries, the fraction of uncovered bases decreased as expression increased (**Fig. 5b–e** and **Supplementary Fig. 2**). However, both the rate of decrease and the coverage per transcript at higher expression levels were variable between better performing libraries (**Fig. 5c**,**d**) and poorly performing ones (**Fig. 5e**). To systematically assess this difference, we compared the Lowess fits of each of the distributions (**Fig. 5b** and **Supplementary Fig. 2**). We found that the dUTP (both in our hands (**Fig. 5c**) and in published data[13]) and 3′ split adaptor (**Fig. 5d**) methods performed best.

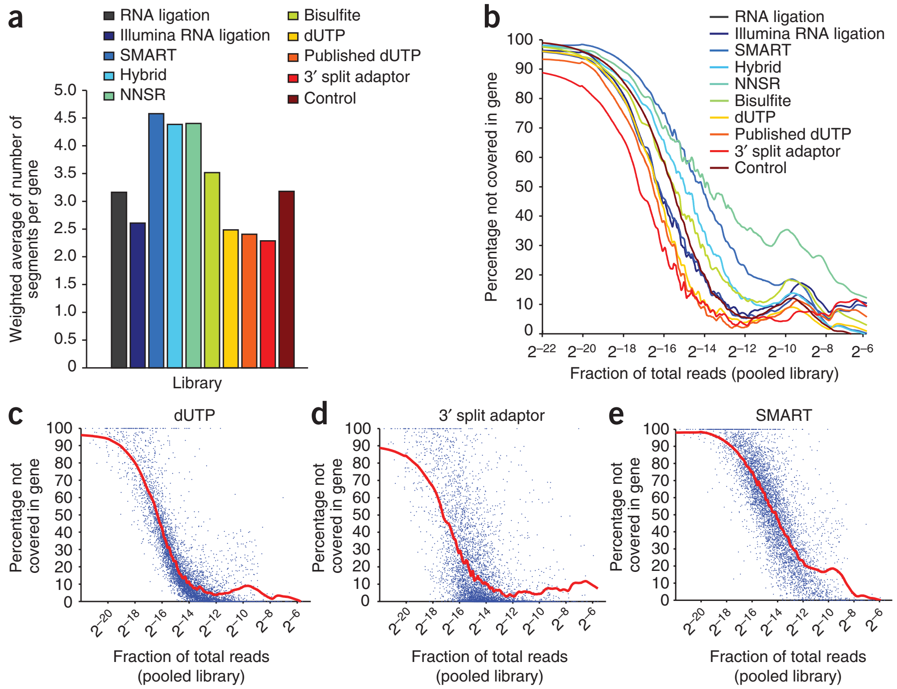


Figure 5 | Continuity of transcript coverage. (**a**) Average number of segments (separated by at least five bases of zero coverage) weighted by the average expression of each gene, in each library. (**b**) Lowess fit for each library. (**c**–**e**) Plots for the dUTP method (**c**), the 3′ split adaptor method (**d**) and the SMART method (**e**). In **c**–**e**, a Lowess fit is shown as a red curve, and each gene is represented by a blue dot.

Coverage at 5′ and 3′ ends

Coverage at 5′ and 3′ ends is crucial for correctly identifying full-length transcripts. To estimate this, we computed for each library the average coverage at each percentile of length from the annotated 5′ end to the annotated 3′ end of known transcripts[18] (**Figs. 2d** and **4b**), as well as the number of genes with complete coverage of their 5′ and 3′ ends (**Fig. 4c**). For paired-end libraries, we computed 5′ and 3′ end coverage based on both read pairs, thus estimating coverage of each end based on the relevant read.

We found substantial variation in the average coverage along a gene's length, with

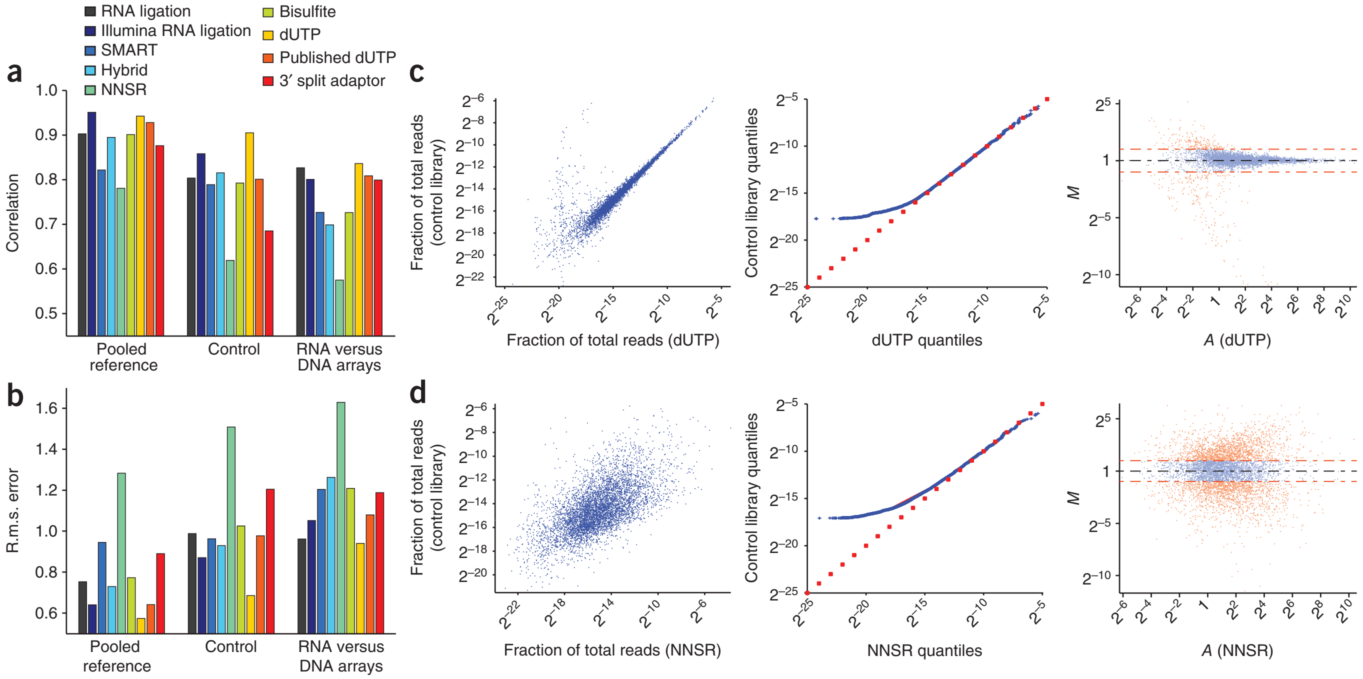


Figure 6 | Digital expression profiling using strand-specific RNA-seq. (**a**,**b**) Pearson correlation coefficient (**a**) and r.m.s. error (**b**) for each library when compared to a pooled reference, the control library and Agilent microarrays (right). (**c**,**d**) Scatter (left), Q-Q (middle) and MA (right) plots for the best performing (dUTP; **c**) and worst performing (NNSR; **d**) libraries, in comparison to the control library. The scatter plots show the fraction of total reads for each gene (blue dot) in the control library against a strand-specific library. The Q-Q plot shows the level at each quantile (rank) of expression in the control library against the strand-specific library. A slope = 1 line is shown for reference (red). The MA plot shows for each gene (dot) the difference in expression levels between the control and strand-specific libraries (*M*; *y* axis) compared to their mean expression level (*A*; *x* axis). Red and blue dashed lines indicate twofold and onefold difference in expression, respectively.

specific biases in 5′ and 3′ coverage (**Fig. 4b**,**c**, **Supplementary Fig. 3** and **Supplementary Table 2**). The NNSR library data had more coverage at the 5′ ends of transcripts, whereas the remaining libraries had modestly increased coverage of the 3′ ends (**Fig. 4b** and **Supplementary Fig. 3**). Consistent with its evenness and continuity, the 3′ split adaptor method had the best coverage of both 5′ and 3′ ends (75% and 77% of genes covered completely at each end, respectively), followed by the dUTP method (62% and 73%) (**Fig. 4c** and **Supplementary Table 2**). The addition of oligo(dT) primers for reverse transcription for the dUTP method, both in our results and in the published data[13], did not increase the coverage at the 3′ ends (**Supplementary Table 2**), although more lenient read mapping may assist with this task in reads that contain portions of the poly(A) tail.

Performance for digital expression profiling

We compared the performance of each library in digital expression profiling relative to reference expression measurements estimated from three ‘standard’ sources: the control (non–strand-specific) library; a pooled estimate generated from the sampled reads of nine of the strand-specific libraries (Online Methods); and expression profiles measured by competitive hybridization of a mid-log phase RNA sample versus genomic DNA using Agilent arrays (Online Methods). We calculated the expression of each gene as its length-normalized read coverage and normalized all values for the total number of reads.

We used several standard quality measures[20] to estimate each library’s performance. These included the Pearson correlation coefficient of expression levels across all genes (**Fig. 6a** and **Supplementary Table 2**); the root mean squared (r.m.s.) error of the expression measurements in each library using the reference measurement as the expected level (**Fig. 6b** and **Supplementary Table 2**); and scatter, quantile-quantile (Q-Q) and MA[21] plots—the last of which compare for each gene the difference in expression between two libraries to the mean expression of that gene in the two libraries (Online Methods, **Fig. 6c**,**d** and **Supplementary Fig. 4**) that help compare differences in expression levels across the dynamic range.

We found that the dUTP library had the best correlation and lowest r.m.s. error relative to all three references (**Fig. 6b** and **Supplementary Table 2**). The only exception was that the Illumina RNA ligation method had a slightly better (0.95 versus 0.94) correlation to the pooled library (**Supplementary Table 2**). Furthermore, visual inspection of the scatter, Q-Q and MA plots showed an excellent linear relation between the dUTP library and the control library across a broad range of values, with weaker performance only for genes with very low expression (**Fig. 6c**). The Illumina RNA ligation protocol also performed reasonably well based on the correlation and r.m.s. error measures but with noticeably broader scatter across the expression range (**Supplementary Fig. 4**). The worst performing methods were the SMART, NNSR and 3′ split adaptor libraries (**Fig. 6d** and **Supplementary Fig. 4**), by all measures.

DISCUSSION

The evaluated RNA-seq protocols broadly represent existing approaches (for a summary of their relative merits, see **Supplementary Table 3**), and we excluded some protocols because of well-known technical limitations, incomplete method development or high similarity to tested methods. These excluded protocols comprise single-stranded cDNA library synthesis[22] (owing to chimeric cDNAs created); deep sequencing of ribosome-protected mRNA fragments[14] (because cDNA lengths are too short, and the original method involves a complex procedure for RNA preparation; we included published data from the nonprotected library designated as the 3′ split adaptor method; **Supplementary Fig. 1**); Helicos single-molecule digital gene expression[23] and direct RNA sequencing[24] (coverage heavily biased to the 5′ or 3′ ends of transcripts, respectively; the latter is currently still under development); and ligation of adaptor to 5′ end and C-tailing at 3′ end of RNA[25] and the double-random priming method[26] (similar to NNSR). We did not include FRT-seq[27] and SOLiD Whole Transcriptome Analysis kit (Applied Biosystems)[28] because they are similar to the two RNA ligation methods we tested, and it would be difficult to distinguish differences owing to library construction methods from those because of the different sequencing methods.

In addition to the formal criteria we evaluated, there is substantial variation in the experimental complexity of different protocols (**Supplementary Table 4**). The original RNA ligation method is the most labor intensive and requires substantial amounts of starting material. The NNSR protocol is the simplest. It is unclear how well the original RNA ligation method can be adapted to larger fragments (greater than 152 base pairs) needed for paired-end sequencing with 76-base reads as it requires the adaptor-ligated RNA to be separated on a gel from unligated RNA, an increasing challenge as the length of the RNA increases.

The libraries also vary in the facility of computational analysis, in particular at early processing steps. The bisulfite method is the most computationally challenging, as reads must be aligned to two reference 'genomes' that have all the cytosine bases converted to thymine bases on one of the two strands. This alignment is complicated both by the imperfect efficiency of the bisulfite treatment and by inherent sequencing errors.

Our analysis allowed us to assess the tradeoff between different protocol modifications. For example, we found that actinomycin D improved the strand specificity of the NNSR protocol (**Supplementary Table 2**) but had the opposite effect on the coefficient of variation, 5′ and 3′ end coverage and correlation of expression levels (**Supplementary Table 2**). For the Illumina RNA ligation libraries, it is preferable to use gel size selection rather than SPRI because removing the shorter cDNAs increased the fraction of reads aligning to the yeast genome. If read length is reduced below 76 bases, this may be less of an issue, but such a choice would also impact other sequencing outputs. Notably, SPRI is amenable to liquid handling automation and can increase the throughput and convenience of any of the other methods, except for RNA ligation. Although these modifications impacted library quality for the NNSR and Illumina RNA ligation methods, most of the variations tested did not alter the performance characteristics of the libraries (**Supplementary Table 2** and **Supplementary Figs. 2–4**), an indication of the reproducibility of the methods. We did not directly evaluate the experimental features, such as PCR conditions or adaptor sequences, that contributed to each method's success (or lack thereof) because these may be complex. We note, however, that the amount of starting material did not correlate with library complexity (**Supplementary Tables 2** and **4**).

The dUTP protocol provided the most compelling overall balance across criteria, followed closely by the Illumina RNA ligation protocol (**Supplementary Note 1**). Currently, the dUTP protocol is compatible with paired-end sequencing, whereas the present Illumina RNA ligation protocol is not. Paired-end sequencing increases the number of mappable reads (unique as pairs), and in higher eukaryotes provides substantial power in transcriptome reconstruction[10,11]. The 3′ split adaptor method[14] excelled in measures critical for genome annotation, but was less well suited for expression profiling. Finally, our compendium and analysis pipeline, which is available online (http://www.broadinstitute.org/regev/rnaseqmethods/) and will be provided as a GenePattern module (http://www.broadinstitute.org/cancer/software/genepattern/), are important resources and include a general benchmarking dataset and tools for testing the quality of future libraries.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Accession code. Gene Expression Omnibus: GSE21739 (sequence and microarray data).

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank members of the Broad Genome Sequencing Platform for sequencing work, J. Meldrim for advice on monotemplate sequencing issues, T. Fennell for help with read processing, S. Luo and G. Schroth (Illumina) for sharing their Illumina RNA ligation protocol, L. Gaffney for assistance with figure graphics, J. Weissman for discussions and T. Liefeld and M. Reich for assistance with the GenePattern module. This work was supported by a US National Institutes of Health Director's Pioneer award, a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, the Human Frontiers Science Program, a Sloan Fellowship, the Merkin Foundation for Stem Cell Research at the Broad Institute, and Howard Hughes Medical Institute (A.R.), by the US-Israel Binational Science Foundation (N.F. and A.R.), by the Canadian friends of the Hebrew University (M.Y.) and by US National Human Genome Research Institute grant HG03067-05 (C.N.).

AUTHOR CONTRIBUTIONS

J.Z.L., M.Y., X.A., D.A.T., N.F. and A.R. wrote the paper. J.Z.L., M.Y., X.A., C.N., D.A.T., N.F., A.G. and A.R. assisted in editing the paper. D.A.T. prepared the poly(A)$^+$ RNA. J.Z.L. and X.A. prepared the cDNA libraries. M.Y., N.F. and A.R. developed and performed the computational analysis. J.Z.L., X.A., M.Y., N.F. and A.R. conceived the research.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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

1. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
2. Wilhelm, B.T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239–1243 (2008).
3. Denoeud, F. *et al.* Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* **9**, R175 (2008).
4. 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).
5. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517 (2008).
6. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
7. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).

8. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
9. Sultan, M. *et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 956–960 (2008).
10. Guttman, M. *et al. Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28**, 503–510 (2010).
11. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
12. Core, L.J., Waterfall, J.J. & Lis, J.T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
13. Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37**, e123 (2009).
14. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R. & Weissman, J.S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
15. He, Y., Vogelstein, B., Velculescu, V.E., Papadopoulos, N. & Kinzler, K.W. The antisense transcriptomes of human cells. *Science* **322**, 1855–1857 (2008).
16. Schaefer, M., Pollex, T., Hanna, K. & Lyko, F. RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Res.* **37**, e12 (2009).
17. Jaffe, D.B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–96 (2003).
18. Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033–1037 (2009).
19. Guo, J., Wu, T., Bess, J., Henderson, L.E. & Levin, J.G. Actinomycin D inhibits human immunodeficiency virus type 1 minus-strand transfer in *in vitro* and endogenous reverse transcriptase assays. *J. Virol.* **72**, 6716–6724 (1998).
20. Gentleman, R., Carey, V., Huber, W., Irizarry, R. & Dudoit, S. (eds.). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, 473 (Springer, Secaucus, NJ, 2005).
21. Yang, Y.H. *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15 (2002).
22. Croucher, N.J. *et al.* A simple method for directional transcriptome sequencing using Illumina technology. *Nucleic Acids Res.* **37**, e148 (2009).
23. Lipson, D. *et al.* Quantification of the yeast transcriptome by single-molecule sequencing. *Nat. Biotechnol.* **27**, 652–658 (2009).
24. Ozsolak, F. *et al.* Direct RNA sequencing. *Nature* **461**, 814–818 (2009).
25. Affymetrix / Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009).
26. 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).
27. Mamanova, L. *et al.* FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat. Methods* **7**, 130–132 (2010).
28. Linsen, S.E. *et al.* Limitations and possibilities of small RNA digital gene expression profiling. *Nat. Methods* **6**, 474–476 (2009).
29. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
30. Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R. & Siebert, P.D. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* **30**, 892–897 (2001).
31. Armour, C.D. *et al.* Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat. Methods* **6**, 647–649 (2009).

ONLINE METHODS

Yeast RNA preparation. We grew *S. cerevisiae* strains Bb32 and BY4741 to mid-log phase. We used mid-log phase RNA from Bb32 for the original RNA ligation and SMART libraries; other libraries were made from a single sample of BY4741 RNA (the two strains are closely related and interchangeable for this study). We made one library (hybrid) from post–diauxic shift BY4741 RNA (slightly impacting its performance in expression profiling and not otherwise). We isolated total and poly(A)$^+$ RNA and treated it with Turbo DNA-free (Ambion) as described[4].

RNA ligation library. We created the library using a previously described method[29] starting from 1.2 μg of poly(A)$^+$ RNA with these modifications. We fragmented RNA by incubation at 70 °C for 8 min in 1× fragmentation buffer (Ambion) and isolated 65–80-nucleotide RNA fragments from a gel. We reverse-transcribed RNA with SuperScript III (Invitrogen) at 55 °C and amplified the cDNA with Herculase (Stratagene) in the presence of 5% DMSO for 16 cycles of PCR followed by a cleanup with 1.8 volumes of AMPure beads (Agencourt) rather than gel purification.

Illumina RNA ligation library. The Illumina method used a preadenylated 3′ adaptor, which enables the subsequent ligation of the 5′ adaptor without an intermediate purification step. Our method has been modified from the version provided by Illumina. We created our libraries starting from 100 ng of poly(A)$^+$ RNA as follows. We decapped RNA by adding 10 U of tobacco acid pyrophosphatase (Epicentre), 1 μl of 10× buffer, 40 U of RNaseOut (Invitrogen) and water in a 10-μl reaction, and incubated it at 37 °C for 90 min, followed by extraction with 25:24:1 phenol: chloroform:isoamyl alcohol (PCIA; Invitrogen), ethanol precipitation and resuspension in 16 μl of H_2O. We fragmented decapped RNA by heating at 94 °C for 6 min in 1× fragmentation buffer (Affymetrix), followed by ethanol precipitation and resuspension in 16 μl of H_2O. We 3′ dephosphorylated fragmented RNA by adding 2 μl of 10× phosphatase buffer, 5 U of Antarctic phosphatase (New England Biolabs (NEB)) and 40 U of RNaseOut and incubating at 37 °C for 30 min followed by 5 min at 65 °C before chilling on ice. We 5′ phosphorylated the RNA by adding 5 μl of 10× PNK buffer, 20 U of T4 polynucleotide kinase (NEB), 5 μl of 10 mM ATP (Epicentre), 40 U of RNaseOut, 17 μl of water and incubating at 37 °C for 60 min. We adjusted the reaction volume to 100 μl with water and cleaned up with the RNeasy MinElute kit (Qiagen) following the instructions of the manufacturer except 400 μl of 100% ethanol were used in step two. We concentrated RNA to 6 μl by Vacufuge (Eppendorf), followed by mixing with 1 μl 1× v1.5 sRNA 3′ adaptor (Illumina), incubating at 70 °C for 2 min and chilling on ice for 2 min. We prepared the 3′ ligation with this RNA adaptor mix, 1 μl 10× T4 RNA ligase 2 truncated reaction buffer, 0.8 μl of 100 mM $MgCl_2$ (Sigma), 20 U of RNaseOut, 300 U of T4 RNA ligase 2, truncated (NEB) and incubated at 22 °C for 1 h. We denatured 1 μl of SRA 5′ adaptor (Illumina) at 70 °C for 2 min and chilled it on ice before combining it with the 3′ adaptor–ligated RNA, 1 μl of 10 mM ATP and 1 μl of T4 RNA ligase (Illumina) and incubating at 20 °C for 1 h. We combined 12 μl of this doubly adaptor-ligated RNA with 3 μl of 0.2× SRA reverse transcription (RT) primer (Illumina), followed by incubation at 70 °C for 2 min, and chilling on ice. We synthesized single-stranded cDNA with this RNA primer mix by adding 6 μl 5× first-strand buffer, 6 μl 100 mM DTT, 1.5 μl 12.5 mM dNTPs, 600 U SuperScript III and 30 U SUPERase-In (Ambion) and incubating for 1 h at 55 °C. We divided the cDNA into two aliquots that we processed with different size selection methods yielding libraries with differing insert lengths. In the first method, we mixed two-thirds of the cDNA with 5 U RNase H (NEB), incubated at 37 °C for 1 h and 75 °C for 15 min, PCIA extracted, ethanol precipitated and resuspended in 10 μl H_2O. We selected single-stranded cDNA ranging in size from 175 to 225 nt on a Criterion 10% TBE-urea gel (Bio-Rad). We crushed the gel slice and eluted with 250 μl 0.3 M NaCl by rotating at room temperature (20–23 °C) for over 4 h. We filtered the crushed gel slice and buffer mixture through a Spin-X cellulose acetate filter (Corning) by centrifugation at 16,000*g* for 3 min. We ethanol-precipitated the eluate and resuspended it in 10 μl RNase-free water. We prepared a 50 μl PCR with 5 μl water, 25 μl 2× Phusion High-Fidelity Master Mix with GC buffer (NEB), 13 μl 5 M betaine (Sigma), 1 μl each primer GX1.0 and 2.0 (Illumina) and 5 μl size-selected cDNA. Thermocycling conditions were: 30 s at 98 °C, 14 cycles of 98 °C for 10 s, 60 °C for 30 s, and 72 °C for 15 s, followed by 10 min at 72 °C. We removed PCR primers using 1.8 volumes of AMPure beads. This generated a cDNA library ranging in size from 180 to 240 base pairs (bp) (insert size of 110–170 bp). In the second method (SPRI), we used one-sixth of the cDNA without size selection in a 50 μl PCR prepared as in the first method. We purified the PCR product twice with 1.3 volumes of AMPure beads to generate a library ranging in size from 120 to 250 bp (insert size of 50–180 bp).

SMART library. We adapted the SMART method[30] developed for SOLiD[32] to Illumina Genome Analzyer sequencing. In our method, reverse transcriptase–primed cDNA synthesis with an oligonucleotide comprised of an Illumina adaptor sequence 5′ of a random hexamer, added three nontemplate cytosine nucleotides at the 3′ end of the cDNA, followed by template switching to a second oligonucleotide containing a second Illumina adaptor sequence 5′ of three guanine ribonucleotides. Specifically, we created the SMART library starting from 100 ng of poly(A)$^+$ RNA as follows. We fragmented RNA by heating at 98 °C for 40 min in 0.2 mM sodium citrate, pH 6.4 (Ambion), followed by concentrating it to 3.5 μl, mixing with 1 μl 2 μM SMART tagged random primer, incubating at 70 °C for 10 min and chilling on ice for 2 min. (Sequences of all custom primers used in this study are listed in **Supplementary Table 5**.) We synthesized first-strand cDNA from this RNA primer mix by adding 2 μl 5× buffer, 1 μl 20 mM DTT, 0.5 μl 10 mM dNTPs, 50 U SMARTScribe reverse transcriptase (Clontech), and 10 U SUPERase-In and incubating at room temperature for 10 min followed by 45 min at 42 °C. We denatured 1 μl 10 μM 5′ SMART oligo at 70 °C for 5 min and added it to the cDNA synthesis reaction, which we then incubated at 42 °C for another 15 min and chilled on ice. We cleaned up the cDNA using 1× volume of AMPure beads and eluted with 20 μl of elution buffer (Qiagen). We prepared a 160 μl PCR with 96 μl water, 16 μl 10× HF 2 PCR buffer, 16 μl 10× HF 2 dNTP mix, 6.4 μl 25 μM primer PE 1.0 (Illumina), 6.4 μl 5μM SMART reverse primer, 3.2 μl 50× Advantage-HF 2 polymerase mix (Clontech) and 16 μl cDNA. Thermocycling conditions were: 5 min at 94 °C, 19 cycles of 94 °C for 15 s and 68 °C for 30 s. We PCIA extracted, ethanol precipitated and resuspended the PCR products in 10 μl

doi:10.1038/nmeth.1491

H_2O. We selected PCR products ranging in size from 220 to 420 bp on a 4% NuSieve 3:1 agarose (Lonza) TAE gel and purified them with the MinElute Gel Extraction kit (Qiagen).

SMART-RNA ligation 'hybrid' library. The SMART–RNA ligation ('hybrid') library combined ligation of an RNA adaptor to the 3′ end of fragmented RNA with SMART's template switching to attach a second adaptor at the 3′ end of the cDNA. We created the library starting from 500 ng poly(A)$^+$ RNA as follows. We fragmented RNA as described for the SMART library and dephosphorylated it with 1.5 µl 10× buffer 3 (NEB), 15 U calf intestinal alkaline phosphatase (NEB), 40 U RNaseOut and water in a final volume of 15 µl for 1 h at 37 °C and then chilled it on ice. We PCIA extracted, ethanol precipitated and resuspended this RNA in 5 µl H_2O. We denatured this RNA and 1 µl 4 µM 3′ RNA adaptor oligo at 70 °C for 2 min, chilled them on ice, combined them with 40 U RNaseOut, 1 µl 100% DMSO (NEB), 10 U T4 RNA ligase (Promega), and 1 µl 10× T4 RNA ligase buffer, and incubated for 6 h at 20 °C and then 4 h at 4 °C. We cleaned up adaptor-ligated RNA using 1.8 volumes of RNAClean beads (Agencourt) and eluted with 10 µl water. We repeated this process to minimize the amount of unincorporated RNA adaptor oligos. We used half of this RNA for cDNA synthesis as described for the SMART library, except we used 1 µl 10 µM Hybrid reverse transcription primer in the reverse transcription reaction for 45 min at 42 °C before adding the 5′ Hybrid oligo. We degraded RNA by adding 2.5 U RNase H, 1.5 µl 10× RNase H buffer, 3 µl water and incubating at 37 °C for 1 h. We PCIA extracted, ethanol precipitated and resuspended the cDNA in 6 µl H_2O. We selected single stranded cDNA ranging in size from 300 to 500 nt on a Criterion 5% TBE-Urea gel and eluted it as described for the Illumina RNA ligation library. We prepared a 125 µl PCR with 2.5 µl water, 62.5 µl 2× Phusion High-Fidelity Master Mix with GC buffer, 50 µl 5 M betaine, 2.5 µl each 25 µM Hybrid forward and Hybrid reverse primers and 5 µl size-selected cDNA. Thermo-cycling conditions were: 30 s at 98 °C, 5 cycles of 98 °C for 10 s, 50 °C for 30 s and 72 °C for 30 s, 13 cycles of 98 °C for 10 s, 65 °C for 30 s and 72 °C for 30 s, followed by 5 min at 72 °C. We removed PCR primers using 1.8 volumes of AMPure beads.

NNSR library. We modified the original NSR method[31], which creates a strand-specific library, by replacing the 'not so random' primers for cDNA synthesis with random (or 'not not so random') primers. The NNSR method used two different primers, each comprised of a different adaptor sequence and random hexamers, for first- and second-strand cDNA synthesis. We created the NNSR library starting from 250 ng of poly(A)$^+$ RNA. We concentrated RNA to 5 µl, mixed it with 2 µl of 100 µM tagged first-strand NNSR primers, incubated them at 65 °C for 5 min and placed them on ice. We synthesized first-strand cDNA with this RNA primer mix by adding 4 µl of 5× first-strand buffer, 2 µl of 100 mM DTT, 1 µl of 10 mM dNTPs, 4 µg actinomycin D (USB), 200 U SuperScript III and 20 U SUPERase-In and incubating at 45 °C for 30 min followed by 15 min at 70 °C. We PCIA extracted twice, ethanol precipitated and resuspended first-strand cDNA in 10 µl H_2O. We treated it with RNase H in 1× RNase H buffer at 37 °C for 20 min followed by 15 min at 75 °C, clean up using 1.8 volumes of RNAClean beads and elution with 30 µl water. We synthesized second-strand cDNA in a 100 µl reaction by adding 10 µl 10× buffer 2 (NEB), 5 µl 10 mM dNTPs, 20 U Klenow Fragment (3′ to 5′ exo$^-$; NEB), 10 µl of 100 µM tagged second-strand NNSR primers and water and incubating at 37 °C for 30 min. We purified the cDNA with 1.8 volumes of AMPure beads. We prepared a 50 µl PCR with 9.5 µl water, 10 µl of 5× reaction buffer 2, 2.5 µl of 10 mM dNTP mix, 5 µl of 25 mM $MgCl_2$, 5 µl of each 10 µM NNSR forward and NNSR reverse primers, 0.5 µl of Expand^{PLUS} enzyme (Roche) and 12.5 µl cleaned up cDNA. Thermo-cycling conditions were: 2 min at 94 °C, two cycles of 94 °C for 10 s, 40 °C for 2 min and 72 °C for 1 min; eight cycles of 94 °C for 10 s, 60 °C for 30 s and 72 °C for 1 min; four cycles of 94 °C for 15 s, 60 °C for 30 s and 72 °C for 1 min with an additional 10 s added at each cycle; 72 °C for 5 min. We purified PCR products using 1.8 volumes of AMPure beads. We selected PCR products ranging in size from 325 to 525 bp on a Criterion 10% TBE gel and eluted them as described for the Illumina RNA ligation library.

We made a second NNSR library in parallel without actinomycin D.

Bisulfite libraries. We created the 'H' and 'S' bisulfite libraries using two previously described methods[15,16], respectively, starting from 1 µg of poly(A)$^+$ RNA with the following modifications. The S library bisulfite reaction followed the 6× cycles for human 28S RNA treatment[16] and was ethanol precipitated before and after desulfonation. We cleaned up the H library bisulfite reaction with an Amicon Ultra-15 3k MWCO filter (Millipore) centrifuged at 4,000*g* at 25 °C for 50 min. In subsequent steps we followed a previously published procedure[15], except as noted. We synthesized first-strand cDNAs from 100 ng of bisulfite-treated poly(A)$^+$ RNA with 1.5 µg 'random octamer' mixture, prepared as described[15], in a 40 µl reaction for 10 min at 25 °C followed by 60 min at 55 °C. We synthesized second-strand cDNA with 5× second-strand buffer (Invitrogen) in a 300 µl reaction. Because bisulfite treatment fragmented the RNA (data not shown), it was not necessary to fragment the cDNA. We prepared a paired-end library for Illumina sequencing as for the dUTP library, except that we gel-purified the final PCR products with an insert size of 160–300 bp.

dUTP library. We created the dUTP second strand library starting from 200 ng of poly(A)$^+$ RNA using a previously described method[13] with the following modifications. All reagents were from Invitrogen except as noted. We fragmented RNA as described for the SMART library, concentrated it to 5 µl, mixing with 3 µg random hexamers, followed by incubation at 70 °C for 10 min and chilling on ice. We synthesized first-strand cDNA with this RNA primer mix by adding 4 µl 5× first-strand buffer, 2 µl 100 mM DTT, 1 µl 10 mM dNTPs, 4 µg of actinomycin D, 200 U SuperScript III and 20 U SUPERase-In, incubating at room temperature for 10 min followed by 1 h at 55 °C. We cleaned up first-strand cDNA by PCIA extraction twice, ethanol precipitation with 0.1 volumes 5 M ammonium acetate to remove dNTPs and resuspension in 104 µl H_2O. We synthesized second-strand cDNA by adding 4 µl of 5× first-strand buffer, 2 µl of 100 mM DTT, 4 µl of 10 mM dNTPs with dTTP replaced by dUTP (Sigma), 30 µl of 5× second-strand buffer, 40 U of *Escherichia coli* DNA polymerase, 10 U of *E. coli* DNA ligase and 2 U of *E. coli* RNase H, and incubating at 16 °C for 2 h. We prepared a paired-end library for Illumina sequencing according to the instructions provided, with the following modifications. First, we ligated five times less adaptor mix to the cDNAs. Second, we incubated 1 U USER (NEB) with

180 to 480 bp size-selected, adaptor-ligated cDNA at 37 °C for 15 min followed by 5 min at 95 °C before PCR. Third, we performed PCR with Phusion High-Fidelity DNA polymerase with GC buffer and 2 M betaine. Fourth, we removed PCR primers using 1.8 volumes of AMPure beads.

In addition, we made a second cDNA library in parallel with 2.7 µg random hexamers plus 1.1 µg anchored oligo$(dT)_{20}$ (Invitrogen) in the first-strand synthesis.

'Control' (non–strand-specific) library. We prepared a control library that used dTTP instead of dUTP for second-strand cDNA synthesis at the same time as the dUTP library. In addition, we made a second control cDNA library in parallel with 2.7 µg of random hexamers plus 1.1 µg of anchored oligo$(dT)_{20}$ in the first-strand synthesis.

Illumina sequencing. We sequenced each of the cDNA libraries with an Illumina Genome Analyzer II (one or two lanes of 76 base reads) using the standard SBS3 and SBS8 sequencing primers (Illumina), except as noted below. We sequenced the SMART library with the standard SBS3 primer for the first read and the custom SBS11 primer for the second read; both reads were 51 bases. We sequenced the RNA ligation and Illumina RNA ligation libraries with the small RNA sequencing primer (Illumina). The NNSR, SMART and Hybrid libraries have a short, identical sequence at the start of every read that leads to 'monotemplate' issues during cluster image processing (**Supplementary Note 2**).

Library read mapping. For SMART, Hybrid and NNSR libraries, we trimmed reads before mapping, to remove specific adaptor-derived bases expected at the start of the read. We mapped reads using Arachne[17]. We mapped reads in single end libraries uniquely, allowing up to four mismatches. We first mapped reads in paired-end libraries non-uniquely allowing up to four mismatches and then searched for unique pairing of the non-unique read mappings (a single pair of mappings on the same chromosome, up to 500 bp apart, with reads on opposite strands). For the bisulfite libraries, we first converted each 'C' in the genome to 'T', resulting in two pseudo-genomes (one per strand), to which the reads were mapped (a unique read mapped to a single location in exactly one of those pseudo genomes).

Read sampling and trimming. We sampled 2.5 million mapped read 'starts' from the aligned reads of each library, with the exception of the SMART and Bisulfite 'H' libraries where we used all reads (~0.9 million and 2.1 million reads, respectively), owing to their repeated low yields. (Resampling these libraries to 2.5 million did not change the results substantially, data not shown.) As the libraries have various read lengths, we used only the first 36 bases of each mapped read (the shortest fragment length in our compendium). We used the sampled 36 base extended coverage for all subsequent method comparison.

Library complexity. We calculated the fraction of reads starting at a distinct (unique) genomic location. In paired libraries we measured the fraction of pairs whose combination of start and end locations was unique, as a proxy for the number of unique cDNAs loaded on the sequencer.

Strand specificity. We used the known annotation from (*Saccharomyces* Genome Database (SGD), http://www.yeastgenome.org/; downloaded in November 2007), and published estimates of UTR lengths[18], or when absent an estimation of 100 bp for each of the UTRs. We considered only high-quality annotations ('verified' or 'uncharacterized'; SGD) and excluded all regions with annotated overlapping transcripts (at UTRs or ORFs) and all genes designated as 'dubious'. We calculated the number of reads that map to the sense and opposite strand of known transcripts.

Evenness of coverage. We used the known annotation from SGD, divided the length of each gene into 100 bins of equal length and calculated the relative coverage in each bin compared to the entire gene. We averaged across all 'verified' and 'uncharacterized' annotated genes.

Continuity of coverage. We measured for each gene the fraction of the gene's total length that had no read coverage. We plotted these values against the relative expression of the gene based on a 'pooled' library (below) and calculated in each plot the Lowess fit of these data (Matlab version 2009b; MathWorks). For each gene, we also counted the number of segments of length 5 bp or longer that had no read coverage. We averaged these measurements across all genes, weighting by the relative expression of each gene.

Comparison to *S. cerevisiae* annotation of 5′ and 3′ ends. Conservatively, we used known annotation of verified and uncharacterized genes (SGD). For each end, we measured the number of genes where a window of ten bp around the translation start and end sites was fully covered by aligned reads.

Expression. We used three standards: microarray data, the 'control' library and a 'pooled' library with 2.5 million sampled mapped reads from each of nine strand-specific libraries (RNA ligation, Illumina RNA ligation, SMART, Hybrid, NNSR, bisulfite, our dUTP, published dUTP and 3′ split adaptor). For each library, we calculated the relative expression level of known genes (SGD) by calculating the mean coverage over the coding region length, and normalizing it to a distribution over all genes[4]. We compared each library to each reference using the Pearson correlation coefficient and the r.m.s. error measures. We also generated scatter, Q-Q and MA plots for each library-reference pair.

MA and Q-Q plots. Both plots compare two sets of data (D_1, D_2). An MA plot displays the $\log_2(D_1) + \log_2(D_2)$ versus $\log_2(D_1) - \log_2(D_2)$. If the samples are very similar, they should be close to the $y = 0$ axis regardless of the x-axis position. A Q-Q plot displays a quantile-quantile plot of D_1 (x axis) and D_2 (y axis). If the samples were drawn from the same distribution, the plot should be a straight line.

Microarray data. Microarray data preparation methods are described in **Supplementary Note 3**.

32. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008).

doi:10.1038/nmeth.1491

Chapter 4

Paper: Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species

Moran Yassour, Jenna Pfiffner, Joshua Z. Levin, Xian Adiconis, Andreas Gnirke,
Chad Nusbaum, Dawn Anne Thompson, Nir Friedman, and Aviv Regev
In *Genome Biology*, 2010.

Yassour *et al. Genome Biology* 2010, **11**:R87
http://genomebiology.com/2010/11/8/R87

Genome **Biology**

RESEARCH **Open Access**

Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species

Moran Yassour[1,2,3], Jenna Pfiffner[1†], Joshua Z Levin[1†], Xian Adiconis[1], Andreas Gnirke[1], Chad Nusbaum[1], Dawn-Anne Thompson[1*], Nir Friedman[3,4*], Aviv Regev[1,2*]

Abstract

Background: Recent studies in budding yeast have shown that antisense transcription occurs at many loci. However, the functional role of antisense transcripts has been demonstrated only in a few cases and it has been suggested that most antisense transcripts may result from promiscuous bi-directional transcription in a dense genome.

Results: Here, we use strand-specific RNA sequencing to study anti-sense transcription in *Saccharomyces cerevisiae*. We detect 1,103 putative antisense transcripts expressed in mid-log phase growth, ranging from 39 short transcripts covering only the 3′ UTR of sense genes to 145 long transcripts covering the entire sense open reading frame. Many of these antisense transcripts overlap sense genes that are repressed in mid-log phase and are important in stationary phase, stress response, or meiosis. We validate the differential regulation of 67 antisense transcripts and their sense targets in relevant conditions, including nutrient limitation and environmental stresses. Moreover, we show that several antisense transcripts and, in some cases, their differential expression have been conserved across five species of yeast spanning 150 million years of evolution. Divergence in the regulation of antisense transcripts to two respiratory genes coincides with the evolution of respiro-fermentation.

Conclusions: Our work provides support for a global and conserved role for antisense transcription in yeast gene regulation.

Background

Antisense transcription plays an important role in gene regulation from bacteria to humans. While the role of antisense transcripts is increasingly studied in metazoans [1], less is known about its relevance for gene regulation in the yeast *Saccharomyces cerevisiae*, a key model for eukaryotic gene regulation. Recent genomic studies using tiling microarrays showed evidence of stable antisense transcription in *S. cerevisiae* [2,3] and *Schizosaccharomyces pombe* [4,5].

It is unclear how broad the role of antisense transcription is and what key functional processes in yeast it affects. A few functional antisense transcripts have been implicated in the control of several key genes, including the meiosis regulator gene *IME4* [6], the phosphate metabolism gene *PHO84* [7], the galactose metabolism gene *GAL10* [8], and the inositol phosphate biosynthetic gene *KCS1* [9]. In contrast, genome-scale analysis in yeast suggested that antisense transcripts largely arise from bi-directional, possibly promiscuous, transcription from nucleosome free regions in promoters or 3′ UTRs of upstream protein coding genes [2,3]. The ability to massively sequence cDNA libraries (RNA-seq) can facilitate the discovery of novel transcripts [10-12], but most studies have not distinguished the transcribed strand.

Here, we used massively parallel sequencing to sequence a strand-specific cDNA library from RNA isolated from *S. cerevisiae* cells at mid-log phase. We

* Correspondence: dawnt@broadinstitute.org; nir@cs.huji.ac.il; aregev@broadinstitute.org
† Contributed equally
[1] Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA
[3] School of Engineering and Computer Science, Hebrew University, Ross Building, Givat Ram Campus, Jerusalem, 91904, Israel
Full list of author information is available at the end of the article

found 1,103 putative antisense transcripts in those cells, ranging from short ones that cover only the 3′ UTR of sense genes to over a hundred long ones that cover the entire sense ORF. Many of the putative sense targets encode proteins with important roles in stationary phase, stress responses, or meiosis. We validated the differential regulation of 67 antisense transcripts and their sense targets in conditions ranging from nutrient limitation to stress, and show that the exosome component *Rrp6* affects their levels, but that the histone deacetylase *Hda2* does not. Furthermore, for a few examples we show that antisense transcripts and their differential regulation are conserved over 150 million years across five yeast species. Our results support a potential conserved role for antisense transcription in yeast gene regulation.

Results

Strand-specific RNA-seq of *S. cerevisiae* cells

To identify antisense transcripts in yeast, we used massively parallel sequencing (Illumina) to sequence a strand-specific cDNA library from *S. cerevisiae* during mid-log growth in rich media. The approach we used [13] relies on the incorporation of deoxy-UTP during the second strand synthesis, allowing subsequent selective destruction of this strand (Materials and methods). Our sequencing yielded 9.22 million 76-nucleotide paired-end reads that map to unique positions in the genome.

Of the reads that map to regions with a known annotation for uni-directional transcription (from the *Saccharomyces* Genome Database (SGD) [14]), only 0.62% were mapped to the opposite (antisense) strand, demonstrating the strand-specificity of our library [15] (Materials and methods). We next combined these reads to define consecutive regions of strand-specific transcription (Materials and methods), and found 8,778 units, covering 4,944 of the 5,501 (90%) genes expressed in this condition (top 85% [12]) at the correct orientation, for at least 80% of the length of each gene (Materials and methods; Additional files 1 and 2).

Identification of 1,103 antisense transcripts that vary in sense coverage from the 3′ UTR to the entire ORF

We found 1,103 putative units that have an antisense orientation relative to annotated transcripts and cover at least 25% of a known transcript on the opposite strand, using published UTR estimates [2] (Materials and methods; Additional file 1). While antisense reads are only a small minority (0.62%) of the total reads, they aggregate in a relatively small number of loci, with 62% of the antisense reads concentrated in the 1,103 units we defined. The remaining 38% are mostly isolated reads scattered across the genome (Figure S1 in Additional file 3).

We observe a range of antisense unit lengths (Figure S2 in Additional file 3). At one extreme are 39 units that cover at least 25% of the transcript but none of the ORF, most commonly at the 3′ UTR (for example, *Unit3689*, a putative antisense transcript to *NOP10*; Figure 1a). Other units cover a substantial portion of the sense ORF. For example, 438 units overlap with at least 50% of the sense ORF, and 145 units cover the entire sense ORF (for example, *Unit4966*, a putative antisense to the *MBR1* gene; Figure 1b). In some cases a single sense gene may be covered by more than one antisense unit, most likely due to low antisense expression levels that result in gaps in coverage (for example, *Unit8753*, *Unit8754*, *Unit8756* and *Unit8758* all opposite to the *OPT2* gene; Figure S3 in Additional file 3). To avoid spurious or 'gapped' calls by our automatic method, we manually inspected each of the units, and focused on the 402 units that passed manual inspection and overlap at least 75% of a sense ORF (Materials and methods).

The 402 antisense units are supported by several lines of evidence. First, comparing the units to published data from strand-specific tiling arrays [2], we find that 143 of our 402 units (36%) are at least 80% covered by stable antisense units as previously defined [2], while 224 units were not detected at all on tiling arrays (Additional file 1; Materials and methods). Finally, 336 of the 402 units are supported by an independent RNA-seq experiment we conducted using an RNA ligation protocol [16] for strand-specific library preparation (Materials and methods) [15]. The lower number of units detected using the independent library reflects the less continuous nature of the data collected by the alternative protocol [15].

Antisense units are unlikely to result solely from leaky transcription

We next assessed the previously suggested possibility [2,17] that antisense transcription is a consequence of leaky transcriptional regulation, through either unterminated transcription, bi-directional transcription initiation from promoters, or transcription from potential nucleosome-free regions (NFRs) in 3′ UTRs. We found that 48 and 27 units might reside within a long 3′ or 5′ UTR, respectively. Of the remaining 333 antisense units, 149 appear to share the (divergent) promoter of a known neighbor transcript, consistent with previous reports [2,3]. An additional 43 units may be transcribed from potential NFRs in the 3′ UTR of an adjacent transcript [18]. The remaining 141 units (35%) cannot be accounted for by transcription from a known promoter or 3′ UTR (when considering 400-bp margins; Figure S4 in Additional file 3).

We compared the change in expression of antisense units and such neighboring genes between cells grown in rich media containing glucose (yeast peptone dextrose

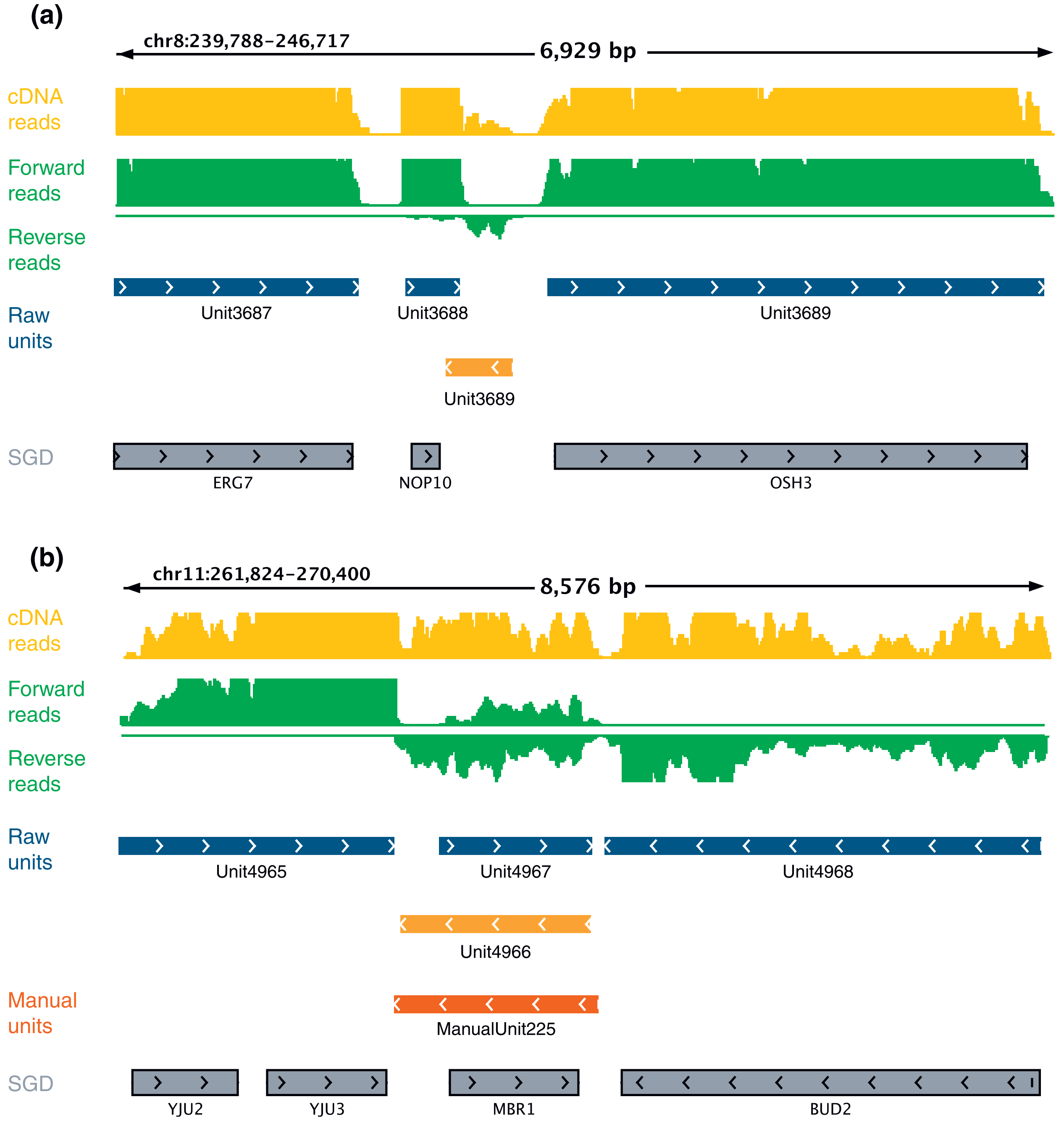


Figure 1 Strand-specific RNA-seq identifies 1,103 antisense units associated with stationary phase, stress, and meiosis genes in *S. cerevisiae*. **(a)** Typical short antisense (*Unit3689*, antisense to *NOP10*). Shown are reads mapped from a standard cDNA sequencing library [15] (yellow), and from the strand-specific library prepared and run side-by-side on the same flow cell (green: forward reads above, reverse reads below). All coverage tracks were normalized to the total number of reads mapped, and are shown up to a threshold of 3 × 10^{-8} of total mapped reads (genome-wide). Units were called from the strand-specific library (blue units, known genes; orange, putative antisense), and are shown along with the manually curated units (red) and the known gene annotations from the SGD (gray). **(b)** Typical long antisense (*ManualUnit225*, antisense to *MBR1*). Tracks are as in (a). The figures are shown using the Integrative Genome Viewer [36].

(YPD)) and ethanol (yeast peptone ethanol (YPE)) as the main carbon source [2]. We reasoned that 'leaky transcription' would result in strong positive correlation in expression between the antisense transcript and the neighboring gene. However, we found a very low correlation (R^2 = 0.07; Figure S5 in Additional file 3), suggesting only weak co-regulation through leaky transcription, from divergent promoters or 3′ NFRs, if at all. Thus, even among the units that could hypothetically arise from leaky transcription, there is little if any evidence of such events.

We also examined the hypothesis that antisense is transcribed to prevent the neighboring gene from runthrough transcription. Of the 402 units, 72 (18%) end relatively close (< 200 bp) to the 3' ends of known genes (for example, *Unit3689* ends close to the *NOP10* gene shown in Figure 1a). On average, the 3′ UTRs of these 72 genes are shorter than those of other genes ($P <$ 0.0058, Wilcoxon test; Figure S6 in Additional file 3). This minority of units could thus potentially play a role in curbing runthrough transcription.

Stress, meiosis and nutrient limitation genes are associated with antisense transcripts at mid-log phase

To explore the potential function of the antisense units, we examined the known function and expression pattern of their associated sense transcripts. We found that the set of ORFs with 75% or more antisense coverage is enriched for genes induced after the diauxic shift ($P < 6 \times 10^{-14}$) or in stationary phase ($P < 2 \times 10^{-10}$), during stress ($P < 2 \times 10^{-27}$), and in some meiosis and sporulation experiments (for example, 85 of 805 genes induced at 8 h in a sporulation time course, $P < 3 \times 10^{-6}$), and include multiple central genes in these processes. For example, the genes encoding the key meiosis proteins *IME4*, *NDT80*, *REC102*, *GAS2*, *SPS19*, *SLZ1*, *RIM9*, and *SMK1* are all associated with long antisense transcription. This is consistent with previous studies in *S. pombe* [4] showing a preponderance of antisense transcription in genes induced during meiosis. Long antisense is also found in many key respiration and mitochondrial genes, including *HAP3*, *COX8*, *CYB2*, *CYC3*, *COX5B*, *MMF1*, *NCA3*, *CYC1*, *MBR1*, *PET10*, *COX12*, and *ATP14*. Genes from other processes repressed during mid-log phase are also associated with long antisense transcripts. Notably, these include at least five members of the PHO regulon (*VTC1*, *PHO5*, *PHM8*, *ICS2*, *PHO3*) and three genes from the GAL regulon (*GAL4*, *GAL10*, *GAL2*). This suggests that antisense regulation may be prevalent across these regulons rather than at single target genes (as found in [6-8]). Furthermore, the expression of 149 of the antisense transcripts is inversely related to that of their sense targets, as measured on tiling arrays [2] in several conditions (glucose versus ethanol, versus galactose, and in Δ*rrp6*; Figure S7 in Additional file 3). Certain key genes that are highly expressed in mid-log phase are also associated with detectable transcription of long antisense units. These include some of the ribosomal protein genes (for example, *RPS26A*, *RPS20*), glycolytic enzymes (for example, *CDC19*, *PGK1*), and cell cycle regulators (for example, *PCL2*, *APC11*, *ASK1*). Nevertheless, these observations suggest that antisense transcription may be regulated in a condition-specific manner in *S. cerevisiae* and may be involved in the repression of stress, stationary phase and meiosis genes in rich growth conditions.

Differential regulation of antisense-sense pairs in nutrient limitation and stress

To test this hypothesis, we first experimentally measured the existence and differential expression of nine pairs of sense and antisense transcripts in *S. cerevisiae*, where the sense gene was known to be induced and important in stress or stationary phase states. We used strand-specific RT-PCR (Materials and methods) followed by sequencing to check for the presence of each sense and antisense transcript in mid-log (rich media), and found that all of the nine tested antisense units were present as expected (Additional file 4). Next, we used strand-specific quantitative real-time PCR (qRT-PCR; Materials and methods) to quantify the differential expression of six sense and antisense transcript pairs between mid-log and early stationary phase. We found that all six of the pairs were differentially expressed, with induction of the sense accompanied by repression of the antisense (Figure 2a; Additional file 5). Third, we devised a novel assay based on the nCounter technology for sensitive multiplex measurement of mRNAs [19,20] (Materials and methods) to measure the absolute level of expression of the nine pairs across five conditions, including mid-log, early stationary phase, stationary phase, high salt and heat shock. We found that the gene pairs exhibited inverse transcription patterns across all the tested conditions (Figure 2b). The differential expression we observed is consistent with antisense interference with sense expression (Figure 2b; Additional file 6), and with the known function and regulation of the sense genes. These included proteins with roles in respiration and mitochondria (*PET10* and *MBR1* [21,22]), repression of ribosomal protein gene expression in stress and poor nutrients (*CRF1* [23]), and the response to caloric restriction (*CTA1* [24]). Thus, differentially regulated antisense transcription may play a role in the distinction between mid-log non-stress growth and stationary phase and stress conditions in *S. cerevisiae*.

Finally, to test the generality of these suggestive patterns, we expanded the nCounter assay to measure the

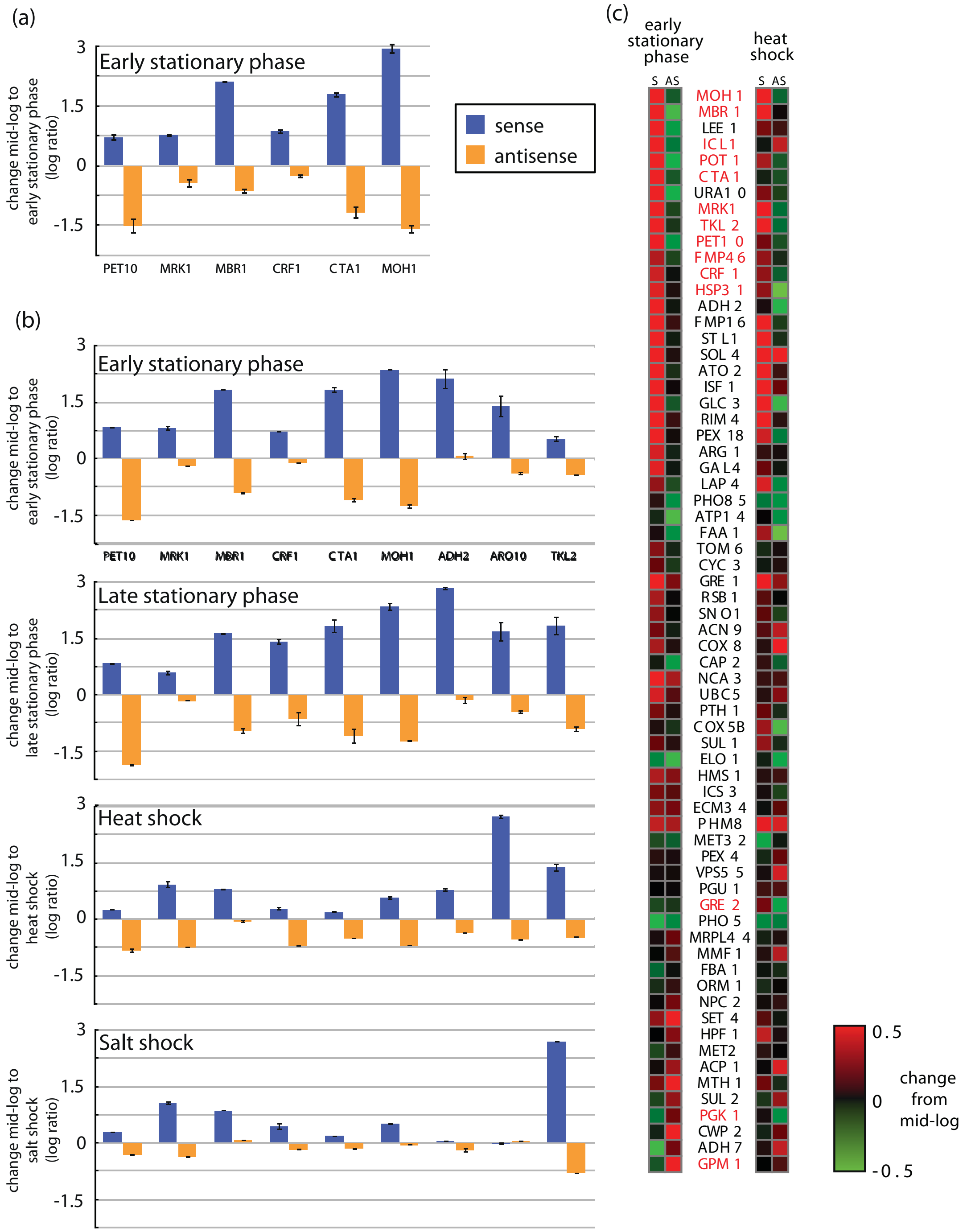


Figure 2 Quantitative expression measurements of putative antisense units and the corresponding sense genes in *S. cerevisiae*. **(a)** Strand-specific qRT-PCR measurements of six pairs of known sense genes and their putative antisense units in comparing mid-log and early stationary phase (the y-axis shows the $\log_2$ ratio of expression in early stationary phase versus mid-log). Error bars indicate the standard deviation between biological replicates and different primers. **(b)** nCounter [20] measurements of nine representative putative antisense units, comparing mid-log to early stationary phase, stationary phase, heat shock and salt stress (the y-axis is as in (a) for the examined condition). Error bars indicate the standard deviation between biological replicates. (**c**) nCounter measurement for 67 tested sense-antisense pairs in early stationary phase (left) and heat shock (right), each relative to a mid-log (no stress) control. The columns marked 'S' and 'A' represent the sense and antisense change, respectively. Red, induced; green, repressed; black, no change. The names of genes highlighted in the main text are shown in red.

expression of 67 sense-antisense pairs in log-phase, early stationary phase, and after 15 minutes under heat shock conditions (Figure 2c; Additional file 6). We found 25 pairs where the sense was induced while the antisense was repressed in either early stationary phase or heat shock (12 in early stationary phase, 21 in heat shock, 8 in both), and 12 pairs where the sense was repressed while the antisense was induced (6 in early stationary phase, 8 in heat shock, 2 in both). Notably, 17 of the 25 pairs with induced sense and repressed antisense in early stationary phase (relative to mid-log) involved sense genes important in respiration, mitochondrial function, alternative carbon source metabolism and starvation response (for example, *PET10*, *MBR1*, *FMP46*, *POT1*, *MOH1*, *TKL2*, *ICL1*, *CTA1*). Conversely, four of the six pairs with the opposite pattern involved sense genes with key roles in glycolysis and fermentation (for example, *GPM1*, *PGK1*). Many of the pairs with induced sense and repressed antisense following heat shock overlapped with those responsive to early stationary phase (consistent with known metabolic changes under stress [25]). Furthermore, they also included four genes known to be important under environmental stresses (the regulators *CRF1* and *MRK1*, and the effectors *HSP31* and *GRE2*). Thus, antisense regulation may play a regulatory role at coordinating the major metabolic changes in the diauxic shift and early stationary phase, and some of the changes in the environmental stress response [21-24].

The exosome component Rrp6 affects antisense levels, but the histone deacetylase Hda2 does not

To explore the mechanistic regulation of antisense transcription, we measured the expression of the 67 pairs of sense and antisense units using the nCounter assay in strains deleted for the exosome component RRP6 (Δ*rrp6*), the histone deacetylase HDA2 (Δ*hda2*), or both (Δ*rrp6*Δ*hda2*). Previous studies [2,7] have suggested that Δ*rrp6* increases the levels of antisense transcription in the *PHO84* locus, and that Hda2 is required for mediating the effect of antisense transcription on the sense transcripts in this locus. If these findings apply more broadly, we expect higher levels of antisense transcripts in Δ*rrp6*, and a change in the relative levels of sense to antisense in either the Δ*hda2* or Δ*rrp6*Δ*hda2* strains.

We found increased transcription of the antisense units in the Δ*rrp6* mutant, with a mild reduction of the sense transcripts (R = -0.36; Figure 3a,c; Figure S8a in Additional file 3). This is consistent with regulation of antisense transcript levels by the exosome, and with a possible, albeit mild, effect of this increase in antisense on reduction in the level of sense transcripts. We found only a very mild, if any, effect on either sense or antisense transcripts levels in Δ*hda2* (Figure 3b; Figure S8b in Additional file 3), suggesting that Hda2 plays at most a very minor independent role in the regulation of our transcripts. We also found no evidence for a synergistic effect between the mechanisms, since transcript levels in the double mutant were very close to those in Δ*rrp6* (Figure S8c in Additional file 3). Finally, the differential expression of the sense genes between conditions was not substantially affected in any of these mutants (for example, R > 0.93 in all conditions; Figure 3d; Figure S9 in Additional file 3), suggesting that relative regulation itself was not compromised in any of these mutants. This may be due to a comparable effect of the deletion in all conditions. Thus, the mechanistic basis of sense-antisense regulation involved Rrp6, but may be more complex than that in the simple model suggested for *PHO84* [7].

Evolutionary conservation of six antisense transcripts and their regulation in five species of yeast

Finally, we tested whether the presence and regulation of antisense transcripts is conserved in five other species of yeast. We reasoned that while the biochemical function and mechanistic basis of each antisense unit may be distinct or complex, their conservation would provide additional support for their functional and ancestral role in gene regulation. We chose five species with diverse lifestyles and a broad phylogenetic range spanning approximately 150 million years (Figure 4). These include three *sensu stricto Saccharomyces* species (*S. paradoxus*, *S. mikatae*, *S. bayanus*), a more distant species that diverged after the whole genome duplication (WGD; *S. castellii*), and one species that diverged pre-WGD (*Kluyveromyces lactis*). Importantly, post-WGD species are known to follow a respiro-fermentative lifestyle, repressing the expression of respiration genes (for example, *PET10*) in mid-log phase, whereas pre-WGD species follow a respirative lifestyle without such repression. We used conserved synteny and gene orthology of *S. cerevisiae* loci [26,27] to identify orthologous regions for candidate antisense transcription in the five species. We focused on six of the units validated in *S. cerevisiae* (*PET10*, *MRK1*, *MBR1*, *CRF1*, *CTA1*, *MOH1*), used strand-specific RT-PCR and sequencing to validate the presence of the orthologous sense and antisense transcripts in each species in mid-log and early stationary phase, and used strand-specific quantitative real-time PCR to quantify transcript levels (Additional file 5).

We found that the tested antisense units are largely conserved in the *sensu stricto* species, and less so at increasing evolutionary distances. All six units were detected in at least one species besides *S. cerevisiae*. Five of the six units are present in *sensu stricto Saccharomyces*, and four are still observed in *S. castellii* and *K. lactis*. The absence in *K. lactis* of an antisense transcript to the *PET10* gene, important for respiratory growth, is

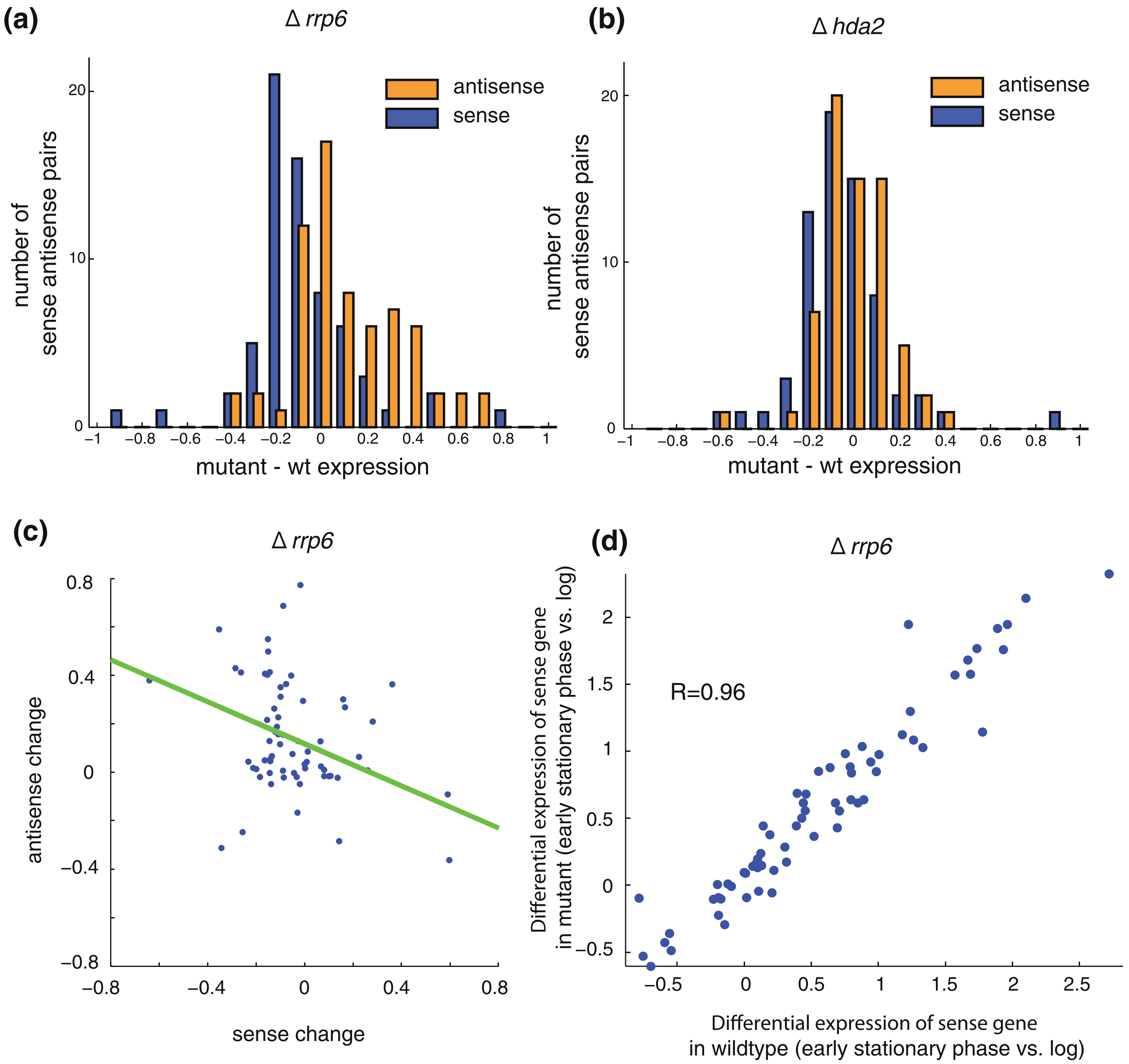


Figure 3 Effect of Rrp6 and Hda2 on antisense transcript levels and sense-antisense regulation. **(a,b)** The distribution of changes in expression levels (x-axis) for sense (blue) and antisense (orange) transcripts in the Δ*rrp6* (a) and Δ*hda2* (b) mutants compared to the wild type (wt). In the Δ*rrp6* mutant (a) there is a mild increase in antisense levels and decrease in sense levels. No such changes are observed in the Δhda2 mutant (b). **(c)** Negative correlation between change in antisense transcript (y-axis) and in sense transcript (x-axis) in the Δ*rrp6* mutant relative to the wild-type strain. **(d)** Similarity in differential sense gene expression from mid-log to early stationary phase between the wild type (x-axis) and the Δ*rrp6* mutant (y-axis).

consistent with its respiratory lifestyle, and suggests that antisense transcription in this gene may have appeared after the whole genome duplication. We cannot rule out the possibility, however, that other antisense units are present in the *K. lactis* genome, or that the missing antisense units are expressed under different conditions.

The anti-correlation between sense and antisense units observed in *S. cerevisiae* is conserved in most post-WGD species, but not in the pre-WGD *K. lactis*. The differential expression of five sense-antisense pairs (*PET10*, *MRK1*, *MBR1*, *CRF1*, *CTA1*) is conserved in at least two out of three other *sensu stricto* species. The more distant *S. castellii* shows less conservation of transcriptional regulation, most prominently in the *PET10* gene. In contrast, although we could detect four of the antisense units in *K. lactis*, their differential expression was not conserved. This is consistent with the lack of repression of the corresponding sense gene in mid-log

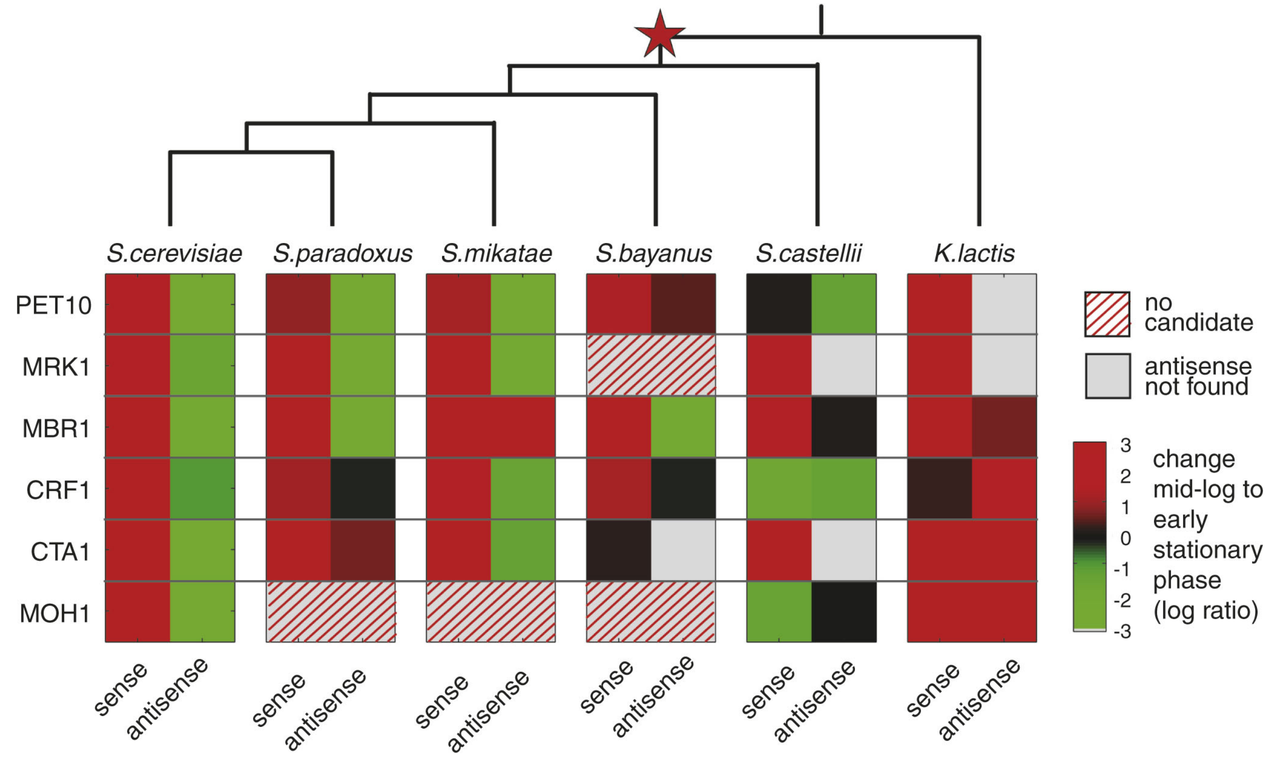


Figure 4 Conservation of the presence and regulation of antisense units in *Hemiascomycota*. Shown are the differential expression values of antisense and sense units comparing mid-log and early stationary phase across *S. cerevisiae* and the five other species (red, higher in early stationary phase; green, lower in early stationary phase; black, no change; hatched, no candidate orthologous contig; grey, no antisense transcription detected in species). A phylogenetic tree of the species included in this study [27] is shown above (the star indicates the WGD).

K. lactis cultures. The absence of antisense (for two genes) and the observed correlated (rather than anti-correlated) regulation (for three others) in *K. lactis* may reflect either the increased phylogenetic distance or may be more directly related to the shift to a respiro-fermentative lifestyle. In the latter case, either antisense transcription or its regulatory pattern in those genes may have evolved concomitantly with the emergence of fermentative growth, and the repression of respiratory genes, such as *PET10* and *MBR1*. Further experiments are needed to elucidate this relationship.

Discussion

In this study, we used strand-specific mRNA sequencing to explore the extent of antisense transcription in yeast, and found 1,103 putative antisense transcripts expressed in mid-log phase in *S. cerevisiae*, ranging from 39 short ones covering only the 3′ UTR of sense genes to 145 long ones covering the entire sense ORF. We focus on 402 long antisense units (each spanning over 75% of a coding unit). In this category, our sequencing based methodology allowed us to identify 224 new antisense transcripts that, in previous studies based on tiling microarrays [2], were either undetected or annotated as long UTRs of neighboring genes.

What could be the role of such prevalent antisense transcription? To date, functional studies have identified a regulatory role for a few antisense transcripts [6-8], whereas genome-wide analyses have suggested that antisense transcripts may represent promiscuous leaky transcription from NFRs at the promoter of a neighboring gene or the 3′ UTR of the sense gene [2,3,28]. The diversity of lengths in our 1,103 antisense units - ranging from long antisense units covering entire ORFs to shorter ones mostly at the 3′ UTR - suggests that there may be more than a single underlying mechanism for their formation and function.

Our results do not support promiscuous or aberrant transcription as the primary cause of the observed antisense transcripts. We find antisense transcription at only 18% of the genes. Moreover, many of the units are long and show robust sequence coverage, in contrast to what we might expect in a noisy process. Finally, antisense genes are only very weakly correlated to their neighbors, inconsistent with leaky transcription from divergent promoters or 3′ NFRs.

Characterizing the functional effect of each unit requires delicate assays to disable the antisense unit, without harming the sense gene, which have been successfully performed only in a few examples [6-8]. We therefore instead examined whether the changes in expression of sense and antisense are consistent with a regulatory function. We chose to focus on the long antisense units because they exhibit strong signal in our data, are less well-studied, are less likely to reflect noise, and can be verified more rigorously.

We found that the sense transcripts corresponding to longer antisense units are significantly enriched for key processes in *S. cerevisiae*, including stress response, the differential regulation of growth and stationary phase, and possibly meiosis and sporulation. The high level of antisense expression is consistent with the repression of these processes in fast growing yeast, and suggests a potential global function. Indeed, when we examined the relative change in expression in sense and antisense units across multiple conditions using three technologies (tiling arrays [2], strand-specific qPCR, and nCounter measurements), we found a strong and consistent anti-correlation between sense genes and the corresponding antisense units. While these results are consistent with regulatory function of antisense units (for example, reduction of antisense transcription leads to increased sense transcription), we cannot rule out the possibility that anti-correlation can occur without active regulation of the antisense transcript. For example, it is possible that when a sense gene is repressed, there is a relieved hindrance of antisense-transcription. Notably, we found support for the role of Rrp6 in the regulation of antisense levels, resulting in an increase in antisense levels in the Δ*rrp6* mutant, and a concomitant, albeit very mild, decrease in sense levels. We could not demonstrate a general effect of Hda2 on the levels of sense or antisense transcripts (either alone or together with Rrp6), and - in all mutants - the differential expression of sense and antisense remained highly correlated to the wild-type regulation. This suggests that it may be challenging to generalize the mechanisms shown for specific transcripts (*PHO84*) to all antisense transcripts.

Independent support for a potential function is the conservation of expression and regulation of six antisense units tested across five species that have diverged more than 150 million years ago, suggesting purifying selection. Notably, previous studies in mammals have shown that certain non-coding RNAs (that are not antisense) can be conserved at the sequence level [17,29], but the applicability of such analyses to antisense transcripts that cover ORFs is limited, and hence experimental data are needed to show conservation. We find that both the presence and the regulation of antisense transcripts are most diverged in the distant, pre-WGD species *K. lactis*. This may reflect either the increased phylogenetic distance *per se*, or an evolved role in regulating respiration genes in post-WGD species. Another possibility for the lack of conservation in expression or absence of antisense in *S. castellii* and *K. lactis* may be the presence of RNA interference in these species [30]. Further experiments will be needed to elucidate these possibilities and characterize the full functional scope of antisense transcription in yeasts.

Conclusions

Our results expand and strengthen the existing body of evidence that antisense transcription is a substantial phenomenon in yeast, and not solely a noisy by product of imprecise transcription regulation. While the mechanism and function of antisense transcription is still elusive, our results indicate that antisense transcription is often conserved and plays a regulatory role in the yeast transcriptional response.

Materials and methods

Supplementary website

All tables, figures, raw sequenced reads, and a link to a browser with the mapped reads appear on our supplementary website [31].

Strains and growth conditions

Strains are listed in Table 1. Cultures were grown in the following rich medium: yeast extract (1.5%), peptone (1%), dextrose (2%), SC Amino Acid mix (Sunrise Science - San Diego, CA, USA) 2 g/l, adenine 100 mg/l, tryptophan 100 mg/l, uracil 100 mg/l, at 200 RPM in a New Brunswick Scientific (Edison, NJ, USA) air-shaker. The medium was chosen to minimize cross-species variation in growth. Following the experimental treatments described below, stressed and mock cultures were transferred to shaking water baths.

To generate strain RGV 69(*rrp6Δ::KANMX6*, *hda2Δ::NatMX4*), strain RGV 71(*rrp6Δ::KANMX6*) was transformed with a PCR product constructed by using the pAG25 containing the NatMX4 cassette using the following primers: GTAAAAGTATTTGGCTTCATTAG TGTGTGAAAAATAAAGAAAATAGATACAATACTATCGACGGTCGACGGATCCCCGGGTT and AAGA AAGTATATAAAATCTCTCTATATTATACAGGCTACTTCTTTTAGGAAACGTCACATCGATGAATTCGAGCTCGTT [32]. Correct integration of this construct was confirmed with the following: (5′ left) left TGGCGTATATGGTTCATTGC; (5′ right) GTATGGG CTAAATGTACGGG; (3′ left) left TGGCGTATATGGT TCATTGC; (3′ right) GGTTGGAGAGGCAAATTGAG.

Heat shock

Overnight cultures of *S. cerevisiae* were grown in 650 ml of media at 22°C to between 3 × 10^7 and 1 × 10^8 cell/

Table 1 Strains and growth conditions

Strain number	Species	Background	Genotype	Source
BB32	*Saccharomyces cerevisiae*			Gift from Leonid Kruglyak's lab
BY4741	*Saccharomyces cerevisiae*	S288c	MATa, his3Δ1, leu2Δ0, met15Δ0, ura3Δ0	Gift from Andrew Murray's lab
	Saccharomyces cerevisiae	BY4741	Same as above with rrp6Δ::KANMX6	ATCC
	Saccharomyces cerevisiae	BY4741	Same as above with hda2Δ::URA3	Gift from Oliver Rando's lab
	Saccharomyces cerevisiae	BY4741	Same as above with rrp6Δ::KANMX6, hda2Δ::NatMX4	This study
NCYC2600	*Saccharomyces paradoxus*			NCYC Stock Center
IFO 1815	*Saccharomyces mikatae*			ATCC
CLIB 592	*Saccharomyces castellii*			CLIB Stock Center
CLIB 209	*Kluyveromyces lactis*			CLIB Stock Center

ATCC, American Type Culture Collection.

ml, OD_{600} = 1.0. The overnight culture was split into two 300 ml cultures and cells from each were collected by removing the media via vacuum filtration (Millipore - Billerica, MA, USA). The cell-containing filters were re-suspended in pre-warmed media to either control (22°C) or heat-shock temperatures (37°C). Density measurements were taken approximately 1 minute after cells were re-suspended to ensure that concentrations did not change during the transfer from overnight media. We harvested 12 ml of culture at 15 minutes and quenched by adding to 30 ml liquid methanol at -40°C, which was later removed by centrifugation at -9°C, and stored these overnight at -80°C. Cell density measurements were repeatedly taken every 5 to 15 minutes for the first 2 hours after treatment. Harvested cells were later washed in RNase-free water and archived in RNAlater (Ambion - Austin, TX, USA) for future preparations. Cells were also harvested from cultures just before treatment for use as controls.

Salt stress

Overnight cultures of *S. cerevisiae* (*BB32*) were grown in 600 ml of media at 30°C until reaching a final concentration of 3×10^7 and 1×10^8 cell/ml. The culture was split into two parallel cultures of 250 ml and sodium chloride was added to one culture for a final concentration of 0.3 M NaCl. Cells were harvested by vacuum filtration at 15 minutes after the addition of sodium chloride and from cultures immediately before the addition of sodium chloride for use as controls (t = 0 minutes). Filters were placed in liquid nitrogen and stored at -80°C and were later archived in RNAlater for future use.

Diauxic shift

Overnight cultures for each species were grown to saturation in 3 ml rich medium. From the 3 ml overnight cultures, 300 ml of rich media was inoculated at the OD_{600} corresponding to 1×10^6 cell/ml: *S. cerevisiae* 0.016, *S. paradoxus* 0.016, *S. mikatae* 0.023, *S. bayanus* 0.016, *S. castellii* 0.020, and *K. lactis* 0.024. The density measurements were taken approximately 1 minute after cells were re-suspended to ensure that concentrations did not change during the transfer from overnight media. Cells were harvested and quenched at a final concentration of 60% methanol at the mid-log and early stationary phase time points. Mid-log was taken at the following OD_{600} values: *S. cerevisiae*, 0.35; *S. paradoxus*, 0.40; *S. mikatae*, 0.40; *S. bayanus*, 0.30; *S. castellii*, 0.35; and *K. lactis*, 0.30. The early stationary phase time points were taken 2 hours after the glucose levels reached zero. Glucose levels were monitored hourly using the YSI 2700 Select Bioanalyzer (YSI Life Sciences - Yellow Springs, OH, USA). OD_{600} values for early stationary phase time points were: *S. cerevisiae*, 4.6; *S. paradoxus*, 3.9; *S. mikatae*, 4.3; *S. bayanus*, 2.8; *S. castellii*, 3.2; and *K. lactis*, 5.0. Harvested cells were later washed in RNase-free water, archived in RNAlater (Ambion) for future preparations, and frozen at -80°C.

Stationary phase

Stationary phase was done for *S. cerevisiae* (*BB32*) only. This experiment was set up identically to the diauxic shift, but samples were taken at mid-log, and 5-day time points. The 5-day samples were taken at the same time of day as the mid-log samples.

Strand-specific cDNA library

The library was created by modifying the previously described dUTP second strand method [13]. All reagents were from Invitrogen (Carlsbad, CA, USA) except as noted. We fragmented 200 ng of *S. cerevisiae* polyA$^+$ RNA by heating at 98°C for 40 minutes in 0.2 mM sodium citrate, pH 6.4 (Ambion). Fragmented RNA was concentrated to 5 μl, mixed with 3 μg random hexamers, incubated at 70°C for 10 minutes, and placed on ice. First-strand cDNA was synthesized with this RNA primer mix by adding 4 μl of 5× first-strand buffer, 2 μl of 100 mM DTT, 1 μl of 10 mM dNTPs, 4 μg of actinomycin D (USB), 200 U SuperScript III, and 20 U

SUPERase-In (Ambion) and incubating at room temperature for 10 minutes followed by 1 hour at 55°C. First-strand cDNA was cleaned up by extraction twice with phenol:chloroform:isoamyl alcohol (25:24:1), followed by ethanol precipitation with 0.1 volumes 5 M ammonia acetate to remove dNTPs and re-suspension in 104 μl H_2O. Second-strand cDNA was synthesized by adding 4 μl 5× first-strand buffer, 2 μl 100 mM DTT, 4 μl 10 mM dNTPs with dTTP replaced by dUTP (Sigma - Aldrich, St Louis, MO, USA), 30 μl 5× second strand buffer, 40 U *Escherichia coli* DNA polymerase, 10 U *E. coli* DNA ligase, 2 U *E. coli* RNase H and incubating at 16°C for 2 hours. A paired-end library for Illumina sequencing was prepared according to the instructions provided with the following modifications. First, five times less adapter mix was ligated to the cDNAs. Second, 1 U USER (New England Biolabs - Ipswich, MA, USA) was incubated with 180- to 480-bp size-selected, adapter-ligated cDNA at 37°C for 15 minutes followed by 5 minutes at 95°C before PCR. Third, PCR was performed with Phusion High-Fidelity DNA Polymerase with GC buffer (New England Biolabs) and 2 M betaine (Sigma). Fourth, PCR primers were removed using 1.8× volume of AMPure PCR Purification kit (Beckman Coulter Genomics - Danvers, MA, USA).

Strand-specific library based on the RNA ligation method

The RNA ligation library was created using a previously described method [16] starting from 1.2 μg of polyA$^+$ RNA with the following modifications. RNA was fragmented by incubation at 70°C for 8 minutes in 1× fragmentation buffer (Ambion) and 65- to 80-nucleotide RNA fragments were isolated from a gel. RNA was reverse transcribed with SuperScript III (Invitrogen) at 55°C and cDNA was amplified with Herculase (Agilent - Santa Clara, CA, USA) in the presence of 5% DMSO for 16 cycles of PCR followed by a clean up with 1.8× volumes of AMPure beads (Beckman Coulter Genomics - Danvers, MA, USA) rather than gel purification.

Illumina sequencing

Both cDNA libraries were sequenced with an Illumina Genome Analyzer II (San Diego, CA, USA). The dUTP library was sequenced using 1 lane of 76-nucleotide paired reads, and the RNA ligation library was sequenced using 2 lanes of 51-nucleotide reads. All RNA-seq data are available in the Gene Expression Omnibus [GEO:GSE21739].

Data pre-processing

We used the Arachne mapper [33] to map the reads to the genome. We next identified consecutive regions of transcription by segmenting the centers of the paired-end segments with coverage >1 and maximum signal gaps of size 20 nucleotides.

Assessment of the strand specificity of the library

To evaluate the strand specificity of our library, we used the known annotation from SGD [14], and published estimates of UTR lengths [2], or when absent an estimation of 100 bp. According to these annotations we found that only 53,803 reads (0.62%) mapped to the opposite strand of known transcripts.

Identification of sense and antisense transcriptional units

We assigned a putative unit to a known gene if it is in the same orientation as the unit and it overlaps the known transcript boundaries, including published estimates of UTR length [2], or when absent an estimation of 100 bp was used. When comparing our transcription units to known annotations in the SGD [14], we examined the top 85% of expressed genes, as previously described [12].

Manual annotation of 402 antisense units

We have manually annotated the boundaries of antisense units covering 75% or more of an opposite ORF, resulting in 402 antisense units covering 75% or more of 412 ORFs.

Comparing the antisense units to published data from strand-specific tiling arrays

We compared our units to the published catalog of [2] using the following criteria. For each of our units, we searched for units in the catalog of [2] that are on the same strand and overlap it. We chose the unit with the highest overlap, and required a minimal threshold of 50% overlap.

Functional analysis of sense units

We constructed a gene set from the 377 sense genes, for which at least 75% of the ORF is covered by an antisense unit, and tested it for functional enrichment using a collection of functional categories as previously described [27]. We also tested the genes for enriched induction or repression in a compendium of 1,400 annotated arrays, as previously described [27].

Identification of candidate regions in other species

We searched for orthologs of the sense gene in other species, using our published orthogroup catalog [27], and used the relative coordinates of the antisense transcripts in *S. cerevisiae* relative to the sense gene to predict their locations in other species. In cases where there were no clear candidates for orthologs, or the synteny block was broken [26], we did not define a candidate.

Strand-specific RT-PCR

Strand-specific RT-PCR followed an adaptation of a published protocol [34]. Total RNA was isolated from strain Bb32(3) at late log time point for two biological replicates. RNA was Turbo DNase treated (Ambion) following the manufacturer's stringent protocol followed by phenol chloroform extraction. For each assay, a gene-specific, strand-specific reverse transcription (RT) was performed. The four reactions for each sample were: +RT L-primer (sense), +RT R-Primer (antisense), +RT no primer, -RT both primers. First strand cDNA synthesis started RNA denaturation and the hybridization of the 2 pmol of gene specific primer. Total RNA with primer (10 ng) was heated to 70°C for 10 minutes and incubated on ice for at least 1 minute. A primer targeting ACT1 mRNA was always included as an internal control for strand specificity. This was followed by adding a Master mix containing 200 U SuperScript III (Invitrogen), 40 U RNaseOut (Invitrogen) and 10 mM dNTP mix for at 55°C for 15 minutes. The enzyme was heat-inactivated at 70°C for 15 minutes. RNA complementary to the cDNA was removed by *E. coli* RNase H (10 U; Ambion) and remaining RNAs were digested with 20 U of RNase Cocktail (Ambion) by incubating at 37°C for 20 minutes. PCR was performed for the sense and antisense transcripts independently. We added 5 μl of RT to each reaction as template with two gene-specific primers each at 250 nM final concentration (the same primers that were used for the sense and antisense RT; Additional file 7), 300 μM dNTP and 1 U of Ampli Taq Gold (Applied Biosystems - Carlsbad, CA, USA), in a 50 μl reaction. RNA contaminated with genomic DNA was used as a positive control. The touch down amplification program used was as follows: incubation of 95°C for 5 minutes followed by 10 cycles of 95°C for 30 s, 60°C for 30 s -1 degree per cycle, 70°C for 45 s, then followed by 17 to 20 cycles of 95°C for 30 s, 50°C for 30 s, 70°C 45 s, 72°C for 10 minutes (a step required for future Topo TA cloning (Invitrogen)).

Strand-specific RT-PCR across species

Strand-specific RT-PCR across species used an adaptation of a published protocol [35]. Total RNA was isolated from each species at both the mid-log and early stationary phase time points. Genomic DNA contamination was removed with Turbo DNase (Ambion) using the stringent protocol, and phenol:chloroform to extract the RNA and to inactivate the DNase. For each of the species two biological replicates of the mid-log and early stationary phase time points were tested. Four reactions were performed for each sample: +RT L-primer (sense), +RT R-primer (antisense), +RT no primer, -RT. The sense, antisense, and -RT reactions were done with 2 pmol of primer (Additional file 7; only the primers with A1 in the title were used for the initial RT-PCR, and all primers used were designed for the target species). RT was done with first strand synthesis only in 20-μl reactions, using 4 units of Omniscript reverse transcriptase (Qiagen - Valencia, CA, USA) and 500 ng of total RNA. Each reaction was carried out at 50°C for 20 minutes, and heat inactivated at 70°C for 15 minutes. PCR was conducted as for the *S. cerevisiae* RT-PCR described above.

Strand-specific qRT-PCR across species

The same RT protocol was followed for the qRT-PCR across species as for the RT PCR above. For each sense-antisense pair validated, two sets of primers were tested, and primers for two internal control genes (*ACT1* and *PDA1*) were included in each reaction. Control primers ('right primer', Additional file 7) were added at a concentration of 2 pmol to each of the RT reactions. qPCR was done using the Roche Light Cycler 480 in 12-μl reactions in a 384 well plate (Roche - Indianapolis, IN, USA). qPCR was done independently for sense, antisense, and control genes. RT samples were diluted 1:40 in water then 1:2 in Light Cycler 480 SYBR Green I Master with gene specific primer pair (each primer at 200 nM final concentration). The program protocol used was as follows: activation, 95°C for 5 minutes; cycling, 95°C for 15 s and 60°C for 45 s; melt, 95°C continuous.

Analysis of strand-specific qRT-PCR data

The ratios reported in Additional file 5 and Figure 2a are $\log_2$ ratios of early stationary phase and mid-log qRT-PCR reads (after normalization by the control gene *PDA1*), averaged over the two sets of primers and the two biological repeats.

nCounter measurements

The following experiments were done in biological duplicates: heat shock - 0 and 15 minutes; salt stress - 0 and 15 minutes; diauxic shift - log and early stationary phase; and stationary phase - log and 5 days. Details on the nCounter system are presented in full in [20]. In a nutshell, the nCounter system uses pre-defined probes labeled with molecular barcodes ('code sets') and single molecule imaging to detect and directly count millions of unique transcripts (from up to hundreds of genes) in a single reaction. The assay is performed in cell lysates, involves no enzymatic steps prior to detection, and is highly accurate. Code sets were constructed to detect putative antisense units and sense genes and additional controls (Additional file 8). We lysed 7×10^7 (or 2×10^7, depending on the code set) cells according to the RNeasy (Qiagen) yeast mechanical lysis protocol. The protocol was stopped after spinning the lysate to remove debris, and 3 μl of the lysate was hybridized for 16 hours followed by processing in the nCounter Prep Station and quantification by the nCounter Digital Analyzer. We normalized the nCounter data in two steps as previously described [19]. In the first step, we controlled for small variations in the efficiency of the automated sample processing. To this end, we followed the manufacturer's instructions, and normalized measurements from all samples analyzed on a given run to the levels of a chosen sample (in all cases we used the first sample in the set). This was done using the positive spiked-in controls provided by the nCounter instrument. In the second step, we used the control genes for which we designed probes to normalize for sample variation.

Additional material

Additional file 1: Table S1. Strand-specific (sense and antisense) transcribed units in mid-log *S. cerevisiae*.

Additional file 2: Table S2. Sense and antisense coverage of SGD annotated genes.

Additional file 3: Figure S1 to S9. Figure S1: read coverage at antisense units. **(a,b)** The distribution (a) and cumulative distribution **(**CDF) (b) of read coverage at antisense units 'called' by our method (gray) and at all other loci in the genome with at least one antisense read (orange). The called units have substantially deeper coverage, whereas 80% of sporadic loci are covered by a single read. (**c**) Sense coverage (x-axis) versus antisense coverage (y-axis) of all verified genes. Genes that we have detected antisense units opposite them are shown in orange. Figure S2: statistics for transcription units. (**a**) Distribution of antisense unit length, colored by the percentage of overlap with the opposite ORF. Dark blue, units with at least 25% overlap with the opposite transcript; light blue, units with at least 50% overlap with the opposite ORF; green, units with at least 75% overlap with the opposite ORF; orange, units with 100% overlap with the opposite ORF. (**b**) Cumulative distribution function of the units length. Blue, antisense units; red, other units. Figure S3: an example of an over-segmented antisense unit. Shown is the genomic region of *OPT2*; tracks and colors are as in Figure 1, with the addition of the brown tracks showing the centers of the paired end segments (forward and reverse), which were used for the segmentation (Materials and methods). All coverage tracks are normalized and shown up to a threshold of 3×10^{-8} of the total (genome-wide) number of mapped reads. Due to low read coverage, both the sense (blue) and the antisense units (yellow) are over-segmented. After the manual curation of the antisense units, we defined one long antisense unit (*ManualUnit402*) that covers the entire ORF of the gene *OPT2*. The figure is shown using the Integrative Genome Viewer [36]. Figure S4: promoter types associated with antisense units. Shown are two examples of promoter types of antisense units; tracks and colors as in Figure 1. *ManualUnit69* included the *BTT1* gene, and a very long 3′ UTR, as an antisense to the gene *MET32*. *ManualUnit70* is a long antisense to the gene *CTA1*, and is transcribed from the divergent promoter of *RMD5*. The figures are shown using the Integrative Genome Viewer [36]. Figure S5: correlation between differential expression of antisense units and their neighboring (non-overlapping) genes. Expression of antisense units versus neighboring genes, which could be co-regulated (using published tiling array data [2]). Shown is the log ratio of change from glucose (YPD) to ethanol (YPE). Blue, antisense units with shared promoter (as in Figure S3 in Additional file 3); red, antisense units with a nearby 3′ UTR; green, linear fit. Figure S6: differences in UTR length between genes with nearby antisense units, compared to all genes. Cumulative distribution of the UTR lengths of all genes (blue) and those with antisense units ending close to the 3′ UTR end. Figure S7: differential expression of antisense units and their target sense transcripts. (**a**) Expression of sense versus antisense units (using published tiling array data [2]). Shown is the log ratio of change in sense gene expression from YPD to YPE (x-axis) plotted versus the same for the antisense strand (y-axis). Red, differentially expressed genes; green, linear fit. (**b,c**) The same as (a), only comparing YPD to galactose growth and to an rrp6 deletion mutant, respectively. Figure S8: mutant effect on transcription. **(a-c)** Expression changes of the sense genes (x-axis) versus expression changes of the antisense units (y-axis) in the Δ*rrp6* mutant (a), the Δ*hda2* mutant (b), and the Δ*rrp6*Δ*hda2* mutant (c). Figure S9: mutant effect on differential expression. **(a-c)** Differential expression of the sense genes from mid-log to early stationary phase in the wild type (x-axis) versus the Δ*rrp6* mutant (a), the Δ*hda2* mutant (b), and the Δ*rrp6*Δ*hda2* mutant (c).

Additional file 4: Table S3. Antisense units validated in RT experiments in *S. cerevisiae*.

Additional file 5: Table S4. qRT-PCR results in each gene and species.

Additional file 6: Table S5. Nanostring results in *S. cerevisiae*.

Additional file 7: Table S6. RT and qRT-PCR primers in each gene and species.

Additional file 8: Table S7. Control genes used for the Nanostring nCounter assays.

Abbreviations

bp: base pair; DTT: dithiothreitol; NFR: nucleosome-free region; ORF: open reading frame; qRT-PCR: quantitative reverse transcriptase PCR; RNA-seq: RNA sequencing; SGD: *Saccharomyces* Genome Database; UTR: untranslated region; WGD: whole genome duplication; YPD: yeast peptone dextrose; YPE: yeast peptone ethanol.

Acknowledgements

MY was supported by the Canadian Friends of the Hebrew University. The work was supported by the Howard Hughes Medical Institute, the Human Frontiers Science Program, a Career Award at the Scientific Interface from the Burroughs Welcome Fund, an NIH PIONEER award, the Broad Institute, and a Sloan Fellowship (AR), the National Human Genome Research Institute (NHGRI) (CN), and by a US-Israel Bi-national Science Foundation award (NF and MY).

Author details

[1]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA 02142, USA. [2]Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, 31 Ames Street, 68-132, Cambridge, MA 02139, USA. [3]School of Engineering and Computer Science, Hebrew University, Ross Building, Givat Ram Campus, Jerusalem, 91904, Israel. [4]Alexander Silberman Institute of Life Sciences, Hebrew University, Edmond J Safra Campus, Givat Ram, Jerusalem, 91904, Israel.

Authors' contributions

MY, JP, JZL, AG, CN, D-AT, NF, and AR designed the research; MY, JP, JZL, XA, D-AT, NF, and AR performed research; MY, NF, and AR analyzed data; JZL and XA contributed text to the methods section; and MY, NF, and AR wrote the paper with editorial input from all authors. All authors read and approved the final manuscript.

Received: 15 February 2010 Revised: 26 July 2010
Accepted: 26 August 2010 Published: 26 August 2010

References

1. Faghihi MA, Wahlestedt C: **Regulatory roles of natural antisense transcripts.** *Nat Rev Mol Cell Biol* 2009, **10**:637-643.
2. Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM: **Bidirectional promoters generate pervasive transcription in yeast.** *Nature* 2009, **457**:1033-1037.
3. Neil H, Malabat C, d'Aubenton-Carafa Y, Xu Z, Steinmetz LM, Jacquier A: **Widespread bidirectional promoters are the major source of cryptic transcripts in yeast.** *Nature* 2009, **457**:1038-1042.
4. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J: **Dynamic repertoire of a eukaryotic**

transcriptome surveyed at single-nucleotide resolution. *Nature* 2008, **453**:1239-1243.

5. Dutrow N, Nix DA, Holt D, Milash B, Dalley B, Westbroek E, Parnell TJ, Cairns BR: **Dynamic transcriptome of *Schizosaccharomyces pombe* shown by RNA-DNA hybrid mapping.** *Nat Genet* 2008, **40**:977-986.
6. Hongay , Grisafi , Galitski , Fink : **Antisense transcription controls cell fate in *Saccharomyces cerevisiae*.** *Cell* 2006, **127**:735-745.
7. Camblong J, Iglesias , Fickentscher , Dieppois , Stutz : **Antisense RNA stabilization induces transcriptional gene silencing via histone deacetylation in *S. cerevisiae*.** *Cell* 2007, **131**:706-717.
8. Houseley , Rubbi , Grunstein , Tollervey , Vogelauer : **A ncRNA modulates histone modification and mRNA induction in the yeast GAL gene cluster.** *Mol Cell* 2008, **32**:685-695.
9. Nishizawa , Komai , Katou , Shirahige , Ito , Toh-E : **Nutrient-regulated antisense and intragenic RNAs modulate a signal transduction pathway in yeast.** *PLoS Biol* 2008, **6**:2817-2830.
10. 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**:1344-1349.
11. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**:621-628.
12. Yassour M, Kaplan T, Fraser HB, Levin JZ, Pfiffner J, Adiconis X, Schroth G, Luo S, Khrebtukova I, Gnirke A, Nusbaum C, Thompson DA, Friedman N, Regev A: ***Ab initio* construction of a eukaryotic transcriptome by massively parallel mRNA sequencing.** *Proc Natl Acad Sci USA* 2009, **106**:3264-3269.
13. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A: **Transcriptome analysis by strand-specific sequencing of complementary DNA.** *Nucleic Acids Res* 2009, **37**: e123.
14. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M, Weng S, Botstein D: **SGD: *Saccharomyces* Genome Database.** *Nucleic Acids Res* 1998, **26**:73-79.
15. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A: **Comprehensive comparative analysis of strand-specific RNA sequencing methods.** *Nat Methods* 2010.
16. 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**:523-536.
17. Berretta J, Morillon A: **Pervasive transcription constitutes a new level of eukaryotic genome regulation.** *EMBO Rep* 2009, **10**:973-982.
18. Tsankov A, Thompson DA, Socha A, Regev A, Rando OJ: **The role of nucleosome positioning in the evolution of gene regulation.** *PLoS Biol* 2010, **8**:e1000414.
19. Amit I, Garber M, Chevrier N, Leite AP, Donner Y, Eisenhaure T, Guttman M, Grenier JK, Li W, Zuk O, Schubert LA, Birditt B, Shay T, Goren A, Zhang X, Smith Z, Deering R, McDonald RC, Cabili M, Bernstein BE, Rinn JL, Meissner A, Root DE, Hacohen N, Regev A: **Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses.** *Science* 2009, **326**:257-263.
20. Geiss GK, Bumgarner RE, Birditt B, Dahl T, Dowidar N, Dunaway DL, Fell HP, Ferree S, George RD, Grogan T, James JJ, Maysuria M, Mitton JD, Oliveri P, Osborn JL, Peng T, Ratcliffe AL, Webster PJ, Davidson EH, Hood L, Dimitrov K: **Direct multiplexed measurement of gene expression with color-coded probe pairs.** *Nat Biotechnol* 2008, **26**:317-325.
21. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *Proc Natl Acad Sci USA* 2000, **97**:262-267.
22. Reisdorf P, Boy-Marcotte E, Bolotin-Fukuhara M: **The MBR1 gene from *Saccharomyces cerevisiae* is activated by and required for growth under sub-optimal conditions.** *Mol Gen Genet* 1997, **255**:400-409.
23. Martin DE, Soulard A, Hall MN: **TOR regulates ribosomal protein gene expression via PKA and the Forkhead transcription factor FHL1.** *Cell* 2004, **119**:969-979.
24. Agarwal S, Sharma S, Agrawal V, Roy N: **Caloric restriction augments ROS defense in *S. cerevisiae*, by a Sir2p independent mechanism.** *Free Radic Res* 2005, **39**:55-62.
25. Chechik G, Oh E, Rando O, Weissman J, Regev A, Koller D: **Activity motifs reveal principles of timing in transcriptional control of the yeast metabolic network.** *Nat Biotechnol* 2008, **26**:1251-1259.
26. Byrne KP, Wolfe KH: **Visualizing syntenic relationships among the hemiascomycetes with the Yeast Gene Order Browser.** *Nucleic Acids Res* 2006, **34**:D452-455.
27. Wapinski I, Pfeffer A, Friedman N, Regev A: **Natural history and evolutionary principles of gene duplication in fungi.** *Nature* 2007, **449**:54-61.
28. He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW: **The antisense transcriptomes of human cells.** *Science* 2008, **322**:1855-1857.
29. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES: **Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.** *Nature* 2009, **458**:223-227.
30. Drinnenberg IA, Weinberg DE, Xie KT, Mower JP, Wolfe KH, Fink GR, Bartel DP: **RNAi in budding yeast.** *Science* 2009, **326**:544-550.
31. **Supplementary Website.** [http://compbio.cs.huji.ac.il/YeastAntisense].
32. Goldstein AL, McCusker JH: **Three new dominant drug resistance cassettes for gene disruption in *Saccharomyces cerevisiae*.** *Yeast* 1999, **15**:1541-1553.
33. Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES: **ARACHNE: a whole-genome shotgun assembler.** *Genome Res* 2002, **12**:177-189.
34. Perocchi F, Xu Z, Clauder-Munster S, Steinmetz LM: **Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D.** *Nucleic Acids Res* 2007, **35**:e128.
35. Haddad F, Qin AX, Giger JM, Guo H, Baldwin KM: **Potential pitfalls in the accuracy of analysis of natural sense-antisense RNA pairs by reverse transcription-PCR.** *BMC Biotechnol* 2007, **7**:21.
36. **Integrative Genomics Viewer.** [http://www.broadinstitute.org/igv].

doi:10.1186/gb-2010-11-8-r87
Cite this article as: Yassour *et al.*: **Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species.** *Genome Biology* 2010 **11**:R87.

Chapter 5

Paper: Full-length transcriptome assembly from RNA-Seq data without a reference genome

Manfred G Grabherr*, Brian J Haas*, Moran Yassour*, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman and Aviv Regev

In *Nature Biotechnology*, 2011

*These authors contributed equally.

nature biotechnology

Full-length transcriptome assembly from RNA-Seq data without a reference genome

Manfred G Grabherr[1,8], Brian J Haas[1,8], Moran Yassour[1–3,8], Joshua Z Levin[1], Dawn A Thompson[1], Ido Amit[1], Xian Adiconis[1], Lin Fan[1], Raktima Raychowdhury[1], Qiandong Zeng[1], Zehua Chen[1], Evan Mauceli[1], Nir Hacohen[1], Andreas Gnirke[1], Nicholas Rhind[4], Federica di Palma[1], Bruce W Birren[1], Chad Nusbaum[1], Kerstin Lindblad-Toh[1,5], Nir Friedman[2,6] & Aviv Regev[1,3,7]

Massively parallel sequencing of cDNA has enabled deep and efficient probing of transcriptomes. Current approaches for transcript reconstruction from such data often rely on aligning reads to a reference genome, and are thus unsuitable for samples with a partial or missing reference genome. Here we present the Trinity method for *de novo* assembly of full-length transcripts and evaluate it on samples from fission yeast, mouse and whitefly, whose reference genome is not yet available. By efficiently constructing and analyzing sets of de Bruijn graphs, Trinity fully reconstructs a large fraction of transcripts, including alternatively spliced isoforms and transcripts from recently duplicated genes. Compared with other *de novo* transcriptome assemblers, Trinity recovers more full-length transcripts across a broad range of expression levels, with a sensitivity similar to methods that rely on genome alignments. Our approach provides a unified solution for transcriptome reconstruction in any sample, especially in the absence of a reference genome.

Recent advances in massively parallel cDNA sequencing (RNA-Seq) provide a cost-effective way to obtain large amounts of transcriptome data from many organisms and tissue types[1,2]. In principle, such data can allow us to identify all expressed transcripts[3], as complete and contiguous mRNA sequence from the transcription start site to the transcription end, for multiple alternatively spliced isoforms. However, reconstruction of all full-length transcripts from short reads with considerable sequencing error rates poses substantial computational challenges[4]: (i) some transcripts have low coverage, whereas others are highly expressed; (ii) read coverage may be uneven across the transcript's length, owing to sequencing biases; (iii) reads with sequencing errors derived from a highly expressed transcript may be more abundant than correct reads from a transcript that is not highly expressed; (iv) transcripts encoded by adjacent loci can overlap and thus can be erroneously fused to form a chimeric transcript; (v) data structures need to accommodate multiple transcripts per locus, owing to alternative splicing; and (vi) sequences that are repeated in different genes introduce ambiguity. A successful method should address each challenge, be applicable to both complex mammalian genomes and gene-dense microbial genomes, and be able to reconstruct transcripts of variable sizes, expression levels and protein-coding capacity.

There are two alternative computational strategies for transcriptome reconstruction[4]. Mapping-first approaches[5], such as Scripture[3] and Cufflinks[2], first align all the reads to a reference (unannotated) genome and then merge sequences with overlapping alignment, spanning splice junctions with reads and paired-ends. Assembly-first (*de novo*) methods, such as ABySS[1], SOAPdenovo[6] or Oases (E. Birney, European Bioinformatics Institute, personal communication), use the reads to assemble transcripts directly, which can be mapped subsequently to a reference genome, if available. Mapping-first approaches promise, in principle, maximum sensitivity, but depend on correct read-to-reference alignment, a task that is complicated by splicing, sequencing errors and the lack or incompleteness of many reference genomes. Conversely, assembly-first approaches do not require any read-reference alignments, important when the genomic sequence is not available, is gapped, highly fragmented or substantially altered, as in cancer cells.

Successful mapping-first methods were developed in the past year[4], but substantially less progress was made to date in developing effective assembly-first approaches. As the number of reads grows, it is increasingly difficult to determine which reads should be joined into contiguous sequence contigs. An elegant computational solution is provided by the de Bruijn graph[7,8], the basis for several whole-genome assembly programs[9–11]. In this graph, a node is defined by a sequence of a fixed length of k nucleotides ('k-mer', with k considerably shorter than the read length), and nodes are connected by edges, if they perfectly overlap by $k - 1$ nucleotides, and the sequence data support this connection. This compact representation allows for enumerating all possible solutions by which linear sequences can be reconstructed given overlaps of $k - 1$.

[1]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts, USA. [2]School of Computer Science, Hebrew University, Jerusalem, Israel. [3]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, Massachusetts, USA. [5]Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden. [6]Alexander Silberman Institute of Life Sciences, Hebrew University, Jerusalem, Israel. [7]Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [8]These authors contributed equally to this work. Correspondence should be addressed to N.F. (nir@cs.huji.ac.il) or A.R. (aregev@broad.mit.edu).

Received 3 December 2010; accepted 28 April 2011; published online 15 May 2011; doi:10.1038/nbt.1883

For transcriptome assembly, each path in the graph represents a possible transcript. A scoring scheme applied to the graph structure can rely on the original read sequences and mate-pair information to discard nonsensical solutions (transcripts) and compute all plausible ones.

Applying the scheme of de Bruijn graphs to *de novo* assembly of RNA-Seq data represents three critical challenges: (i) efficiently constructing this graph from large amounts (billions of base pairs) of raw data; (ii) defining a suitable scoring and enumeration algorithm to recover all plausible splice forms and paralogous transcripts; and (iii) providing robustness to the noise stemming from sequencing errors and other artifacts in the data. In particular, sequencing errors would introduce a large number of false nodes, resulting in a massive graph with millions of possible (albeit mostly implausible) paths.

Here, we present Trinity, a method for the efficient and robust *de novo* reconstruction of transcriptomes, consisting of three software modules: Inchworm, Chrysalis and Butterfly, applied sequentially to process large volumes of RNA-Seq reads. We evaluated Trinity on data from two well-annotated species—one microorganism (fission yeast) and one mammal (mouse)—as well as an insect (the whitefly *Bemisia tabaci*), whose genome has not yet been sequenced. In each case, Trinity recovers most of the reference (annotated) expressed transcripts as full-length sequences, and resolves alternative isoforms and duplicated genes, performing better than other available transcriptome *de novo* assembly tools, and similarly to methods relying on genome alignments.

RESULTS

Trinity: a method for *de novo* transcriptome assembly

In contrast to *de novo* assembly of a genome, where few large connected sequence graphs can represent connectivities among reads across entire chromosomes, in assembling transcriptome data we expect to encounter numerous individual disconnected graphs, each representing the transcriptional complexity at nonoverlapping loci. Accordingly, Trinity partitions the sequence data into these many individual graphs, and then processes each graph independently to extract full-length isoforms and tease apart transcripts derived from paralogous genes.

In the first step in Trinity, Inchworm assembles reads into the unique sequences of transcripts. Inchworm (**Fig. 1a**) uses a greedy *k*-mer–based approach for fast and efficient transcript assembly, recovering only a single (best) representative for a set of alternative variants that share *k*-mers (owing to alternative splicing, gene duplication or allelic variation). Next, Chrysalis (**Fig. 1b**) clusters related contigs that correspond to portions of alternatively spliced transcripts or otherwise unique portions of paralogous genes. Chrysalis then constructs a de Bruijn graph for each cluster of related contigs, each graph reflecting the complexity of overlaps between variants. Finally, Butterfly (**Fig. 1c**) analyzes the paths taken by reads and read pairings in the context of the corresponding de Bruijn graph and reports all plausible transcript sequences, resolving alternatively spliced isoforms and transcripts derived from paralogous genes. Below, we describe each of Trinity's modules.

Inchworm assembles contigs greedily and efficiently

Inchworm efficiently reconstructs linear transcript contigs in six steps (**Fig. 1a**). Inchworm (i) constructs a *k*-mer dictionary from all sequence reads (in practice, *k* = 25); (ii) removes likely error-containing *k*-mers from the *k*-mer dictionary; (iii) selects the most frequent *k*-mer in the dictionary to seed a contig assembly, excluding both low-complexity

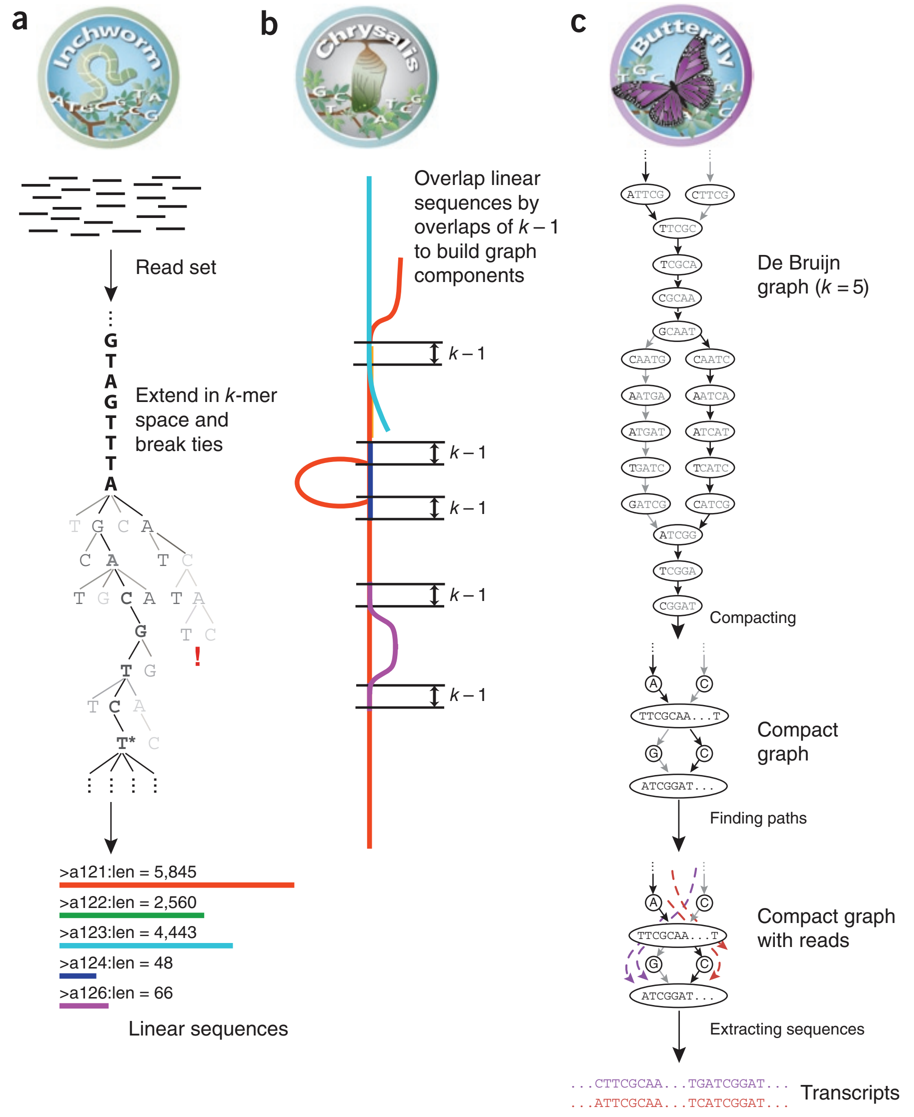
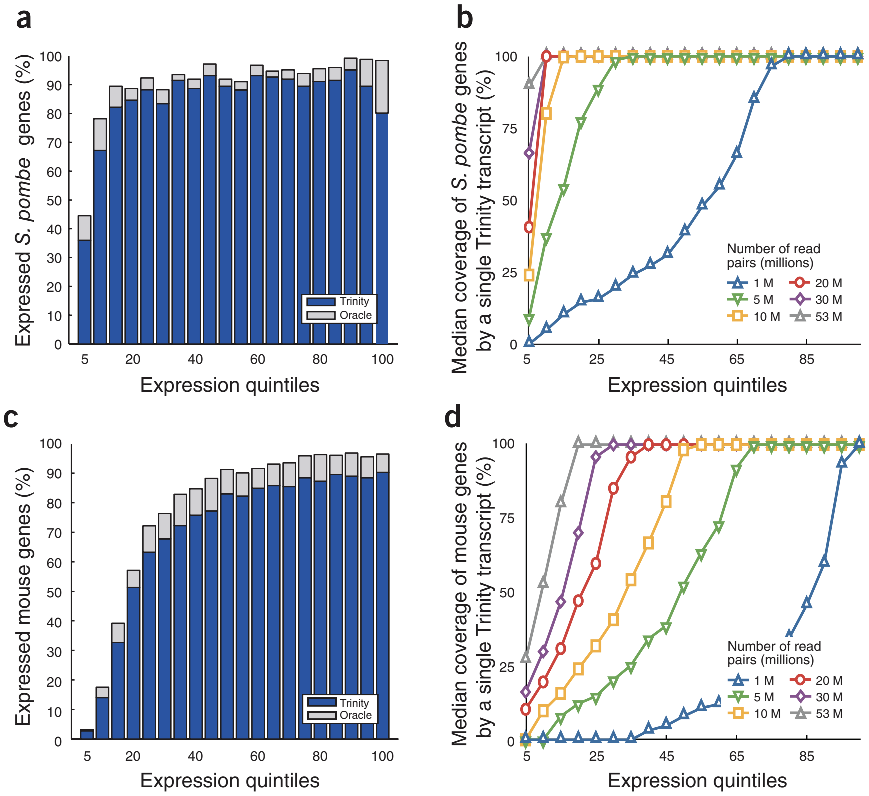


Figure 1 Overview of Trinity. (**a**) Inchworm assembles the read data set (short black lines, top) by greedily searching for paths in a *k*-mer graph (middle), resulting in a collection of linear contigs (color lines, bottom), with each *k*-mer present only once in the contigs. (**b**) Chrysalis pools contigs (colored lines) if they share at least one *k* – 1-mer and if reads span the junction between contigs, and then it builds individual de Bruijn graphs from each pool. (**c**) Butterfly takes each de Bruijn graph from Chrysalis (top), and trims spurious edges and compacts linear paths (middle). It then reconciles the graph with reads (dashed colored arrows, bottom) and pairs (not shown), and outputs one linear sequence for each splice form and/or paralogous transcript represented in the graph (bottom, colored sequences).

Figure 2 Trinity correctly reconstructs the majority of full-length transcripts in fission yeast and mouse. (**a**,**c**) The fraction of genes that are fully reconstructed and in the Oracle Set in different expression quintiles (5% increments) in fission yeast (50 M pairs assembly) (**a**) and the fraction of genes that have at least one fully reconstructed transcript and are in the Oracle Set in different expression quintiles in mouse (53 M pairs assembly) (**c**). Each bar represents a 5% quintile of read coverage for genes expressed. Gray bars show the remaining fraction of transcripts that are in the Oracle Set but not fully reconstructed. For example, ~36% of the *S. pombe* transcripts at the bottom 5% of expression levels are fully reconstructed by Trinity; ~45% of the transcripts in this quintile are in the Oracle Set. (**b**,**d**) Curves show the median values for coverage (as fraction of length of reference transcripts) by the longest corresponding Trinity-assembled transcript, according to expression quintiles in yeast (**b**) and mouse (**d**), depending on the number of read pairs that went into each assembly.



and singleton *k*-mers (appearing only once); (iv) extends the seed in each direction by finding the highest occurring *k*-mer with a *k* – 1 overlap with the current contig terminus and concatenating its terminal base to the growing contig sequence (once a *k*-mer has been used for extension, it is removed from the dictionary); (v) extends the sequence in either direction until it cannot be extended further, then reports the linear contig; (vi) repeats steps iii–v, starting with the next most abundant *k*-mer, until the entire *k*-mer dictionary has been exhausted.

The contigs reported by Inchworm alone do not capture the full complexity of the transcriptome; for example, only one alternatively spliced variant can be reported at full length per locus, with partial sequences reported for unique regions of any alternatively spliced transcripts. However, its contigs do maintain the information required by subsequent Trinity components to reconstruct and search the entire graph containing all possible sequences. Indeed, except for low-complexity and singleton k-mers excluded from seeds or discarded in contigs shorter than the minimum length required, Inchworm's contigs provide a complete representation of the sequence overlap–based de Bruijn graph, with each *k*-mer being unique in the set, and the *k* – 1 subsequences implicitly defining the edges in the graph. This approach is much more efficient than computing a full graph from all reads at once, and it quickly provides a meaningful intermediate output of the contigs strongly supported by many *k*-mers in the reads. By eliminating singleton *k*-mers as initial seeds for contig extensions, Inchworm further reduces the inclusion in assemblies of *k*-mers likely resulting from sequencing errors.

Chrysalis builds de Bruijn transcript graphs

Chrysalis clusters minimally overlapping Inchworm contigs into sets of connected components, and constructs complete de Bruijn graphs for each component (**Fig. 1b**). Each component defines a collection of Inchworm contigs that are likely to be derived from alternative splice forms or closely related paralogs. Chrysalis works in three phases. (i) It recursively groups Inchworm contigs into connected components. Contigs are grouped if there is a perfect overlap of *k* – 1 bases between them and if there is a minimal number of reads that span the junction across both contigs with a (*k* – 1)/2 base match on each side of the (*k* – 1)-mer junction. (ii) It builds a de Bruijn graph for each component using a word size of *k* – 1 to represent nodes, and *k* to define the edges connecting the nodes. It weights each edge of the de Bruijn graph with the number of *k*-mers in the original read set that support it. (iii) It assigns each read to the component with which it shares the largest number of *k*-mers, and determines the regions within each read that contribute *k*-mers to the component.

Butterfly resolves alternatively spliced and paralogous transcripts

Butterfly reconstructs plausible, full-length, linear transcripts by reconciling the individual de Bruijn graphs generated by Chrysalis with the original reads and paired ends. It reconstructs distinct transcripts for splice isoforms and paralogous genes, and resolves ambiguities stemming from errors or from sequences >*k* bases long that are shared between transcripts.

Butterfly consists of two parts (**Fig. 1c**). During the first part, called graph simplification, Butterfly iterates between (i) merging consecutive nodes in linear paths in the de Bruijn graph to form nodes that represent longer sequences and (ii) pruning edges that represent minor deviations (supported by comparatively few reads), which likely correspond to sequencing errors. Diploid polymorphisms are expected to be more frequent than sequencing errors and will likely be maintained. In the second part, called plausible path scoring, Butterfly identifies those paths that are supported by actual reads and read pairs, using a dynamic programming procedure that traverses potential paths in the graph while maintaining the reads (and pairs) that support them. Because reads and sequence fragments (paired reads) are typically much longer than *k*, they can resolve ambiguities and reduce the combinatorial number of paths to a much smaller number of actual transcripts, enumerated as linear sequences.

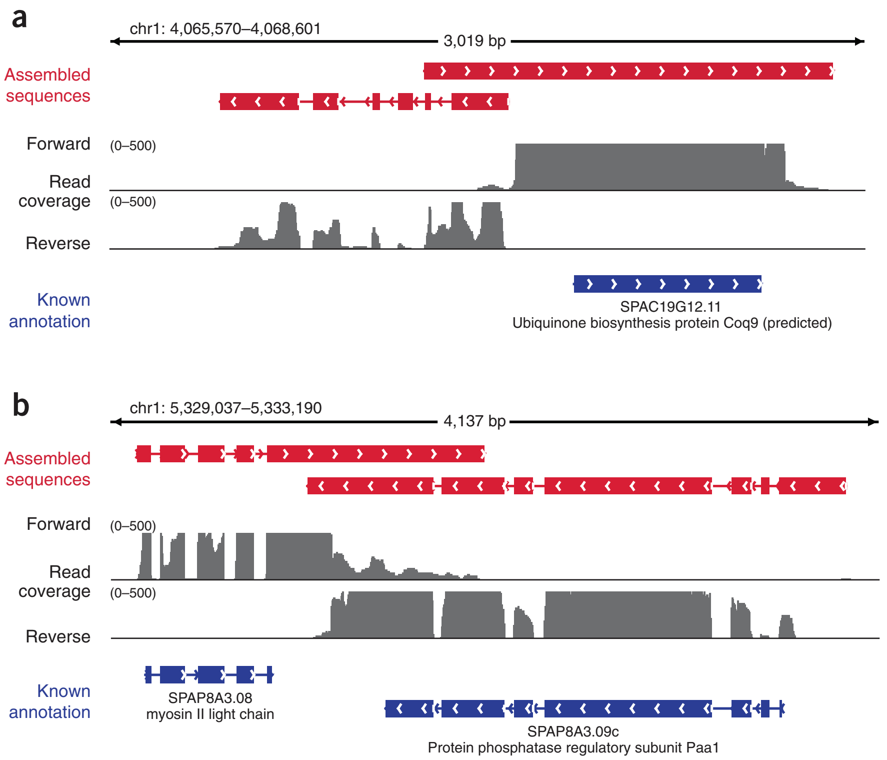


Figure 3 Trinity improves the yeast annotation. Shown are examples of Trinity assemblies (red) along with the corresponding annotated transcripts (blue) and underlying reads (gray) all aligned to the *S. pombe* genome (read alignment is shown for graphical clarity; no alignments were used to generate the assemblies). (**a**) Trinity identifies a new multi-exonic transcript (left) and extends the 5′ and 3′ UTRs of the *coq*9 gene (right). (**b**) Trinity extends the UTRs of two convergently transcribed and overlapping genes.

RNA-Seq of *Schizosaccharomyces pombe*

We first generated RNA-Seq data from the fission yeast *S. pombe*. The *S. pombe* transcriptome[12] has relatively substantial splicing for a eukaryotic microorganism, with short introns (mean intron length = 80.6 bp) and dense transcripts (mean intergenic region = 938 bp based on coding genes only). To maximize transcript coverage, we pooled ~154 million pairs of strand-specific[13,14], 76-base Illumina read sequences from four biological conditions: mid-log growth, growth after all glucose has been consumed, late stationary phase and heat shock[15].

Sensitivity limit for full-length reconstruction

We next estimated the upper sensitivity limit for which annotated transcripts can possibly be perfectly reconstructed given a particular data set of sequences. Any assembly approach based on a particular *k*-length oligomer is limited to those sequences that are represented by the exact *k*-mer composition of the RNA-Seq read set. To determine this empirical upper sensitivity limit, we built a *k*-mer dictionary from all the reads and identified all known reference protein-coding sequences that are reconstructable to full length given the read set, as those sequences that can be populated by adjacent and overlapping *k*-mers across their entire length. We call this set of sequences the 'Oracle Set'. Because this set also contains transcript sequences that are covered by *k*-mers, but not entire reads, some transcripts will appear reconstructable but are not. Conversely, the Oracle Set reflects only annotated known genes and known isoforms, which are likely an underestimate, especially in mammals[16]. Nevertheless, the Oracle Set provides a useful sensitivity benchmark.

In the *S. pombe* data set, nearly all (91%, 4,600/5,064) reference protein-coding sequences exist in the Oracle Set (25-mer dictionary, 154 M paired-reads), as almost all encoded transcripts (98%) are expressed in the measured conditions (≥ 0.5 fragments per transcript kilobase per million fragments mapped (FPKM)), consistent with previous studies in yeasts[5,17,18]. When reducing the coverage by random sub-sampling, the size of the Oracle Set is saturated at 50 M paired reads (4,494/5,064, **Supplementary Fig. 1**), which we chose as our subsequent benchmarking set.

Trinity recovered most *S. pombe* transcripts

From the 50 M pairs of reads, Trinity fully reconstructed 86% of annotated transcripts (4,338/5,064, **Supplementary Table 1**) at full length, including 94% of the stringently defined oracle transcripts (4,218/4,494). Of the 276 oracle transcripts not fully reconstructed, 90 (33%) are reconstructed over at least 90% of their length, and 177 (64%) are reconstructed over at least 50% of their length.

Overall, Trinity generated 27,841 linear contigs longer than 100 bases, grouped into 23,232 components (**Supplementary Note**). Only 2,454 of the 27,841 Trinity contigs did not align to the genome using GMAP[19]. Of those, 30% match a Uniref90 (ref. 20) protein (BLASTX $E \leq 10^{-10}$), almost invariably (90%) a *Schizosaccharomyces* protein, and likely reflect assemblies with error-rich reads.

Trinity reconstructs full-length transcripts across a broad range of expression levels and sequencing depths (**Fig. 2**). For example, it accurately captured the full-length transcript of 71% of genes from the second quintile (5–10%), and had full-length coverage of 81–95% of annotated transcripts in the remaining quintiles (**Fig. 2a**). Considering both full-length and partial reconstructions, Trinity reconstructed a large fraction of the bases in each transcript (**Fig. 2b**).

In many cases, Trinity accurately resolved the sequences of closely related paralogous transcripts. Out of 77 gene families containing 185 paralogs[21], Trinity recovered at full length all members of 33 families (68 genes), at least one member from an additional 33 families (46 genes found, 45 genes missing), and missed all 26 genes in the remaining 11 families, often involving genes not highly expressed. Some of the most highly expressed transcripts in *S. pombe* are derived from paralogous genes with very similar sequences (e.g., those encoding ribosomal proteins[21]), yet were resolved by Trinity.

Extended UTRs and long anti-sense transcripts in *S. pombe*

Compared to the existing annotation, Trinity extended the 5′ untranslated region (UTR) of 312 transcripts (median extension, 80 bp; average, 176 bp), and the 3′ UTR of 543 transcripts (median, 72 bp; average, 172 bp) (**Supplementary Fig. 2a**,**b**). It also found 3,726 previously unannotated 5′ UTRs (median length, 183 bp; average length, 288 bp), and 3,416 3′ UTRs (median length, 272 bp; average length, 397 bp).

Trinity identified 2,319 transcripts at 1,235 intergenic loci as novel transcribed sequences (**Fig. 3a**) and 612 long antisense transcripts that covered >75% of the length of the corresponding sense

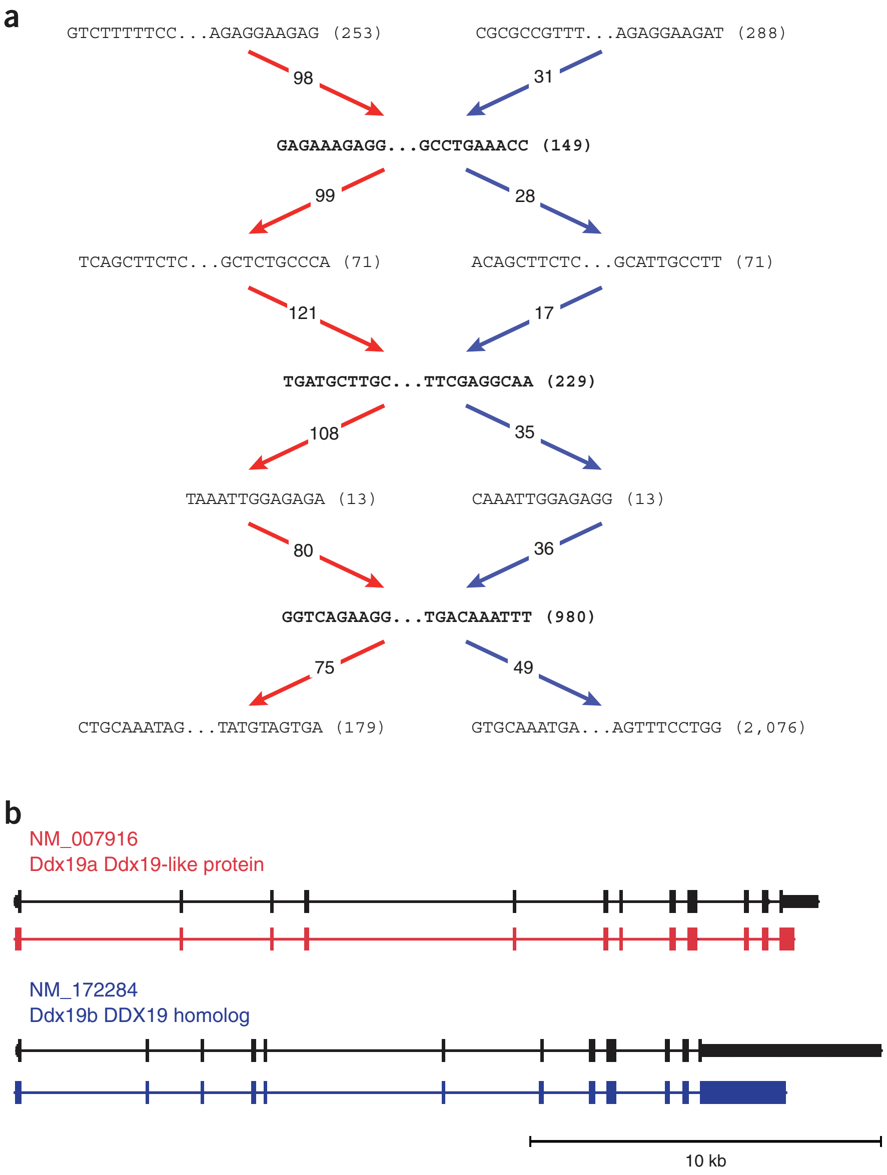

a
GTCTTTTTCC...AGAGGAAGAG (253)
CGCGCCGTTT...AGAGGAAGAT (288)
98
31
GAGAAAGAGG...GCCTGAAACC (149)
99
28
TCAGCTTCTC...GCTCTGCCCA (71)
ACAGCTTCTC...GCATTGCCTT (71)
121
17
TGATGCTTGC...TTCGAGGCAA (229)
108
35
TAAATTGGAGAGA (13)
CAAATTGGAGAGG (13)
80
36
GGTCAGAAGG...TGACAAATTT (980)
75
49
CTGCAAATAG...TATGTAGTGA (179)
GTGCAAATGA...AGTTTCCTGG (2,076)
b
NM_007916
Ddx19a Ddx19-like protein
NM_172284
Ddx19b DDX19 homolog
10 kb

Figure 4 Trinity resolves closely paralogous genes. (**a**) The compacted component graph for two paralogous mouse genes, *Ddx19a* and *Ddx19b* (93% identity). Red and blue arrows highlight the two paths chosen by Trinity out of the 64 possible paths in this portion of the graph alone. Numbers on the edges indicate the number of supporting reads; numbers in parentheses represent the sequence length at each node. (**b**) Alignments between the transcripts represented by the red and blue paths in **a** and the paralogous genes *Ddx19a* and *Ddx19b* relative to the mouse reference genome (genome alignment shown for graphical clarity only; no alignments were used to generate the assemblies).

transcript (**Fig. 3b**), and were not likely to be derived from extended transcription of a neighboring gene. One hundred thirteen of the intergenic transcripts and 612 long antisense transcripts were multiexonic. Although both were expressed at lower levels on average than annotated protein-coding genes (**Supplementary Fig. 3**), 49 long antisense transcripts (at 35 loci) were at least fivefold more highly expressed than the corresponding sense coding transcript (e.g., an antisense transcript to the meiotic gene *mug*27/*slk*1 (SPCC417.06c) was >100-fold more highly expressed, **Supplementary Fig. 4**). This supports a role for antisense transcriptional regulation in meiosis for *S. pombe*[15,22–24], and is consistent with previous findings in *S. cerevisiae*[25].

Trinity recovered most expressed annotated mouse transcripts

Compared to yeasts, mammalian transcriptomes exhibit substantially more complex patterns of alternative splicing[26]. To test Trinity's ability to identify different isoforms, we sequenced ~52.6 million 76-base read pairs from C567BL/6 mouse primary immune dendritic cells. Unlike in *S. pombe*, only 54% of known mouse genes (10,724) were identified as expressed (≥0.5 FPKM), and of those, the Oracle Set determined 8,358 to be full-length reconstructable (727 loci have two or more isoforms variable in the protein-coding sequences, totaling 9,258 transcripts).

Trinity reported 48,497 contigs longer than 350 bp, capturing 8,185 transcripts to full-length (**Supplementary Table 2** and **Supplementary Note**), corresponding to 7,749 loci (including 7,947 (86%) transcripts at 7,573 (91%) loci in the mouse Oracle Set). The percentage of transcripts recovered to full-length and the fraction of length captured were high across a broad range of expression levels (**Fig. 2c**,**d**).

Trinity resolved splice isoforms and gene paralogs in a manner consistent with the mouse Oracle Set. Trinity found 872 full-length, alternatively spliced, isoforms from 385 loci (53% of the loci with alternatively spliced variants in the Oracle Set), and matched the full-length transcripts for 463 (61.6%) of 752 paralogous transcripts in the Oracle Set (>70% identity between paralogs, **Fig. 4**).

Trinity extended the annotated 5′ UTR for 5,265 transcripts (5,036 loci, median length, 43; average length, 91, **Supplementary Fig. 2c**), and included one or more additional 5′ UTR exons in 305 cases (**Supplementary Fig. 5**). It extended the 3′ UTR in 2,918 transcripts (2,819 loci, median length, 20; average length, 248; **Supplementary Fig. 2d**), adding 3′ UTR exons in 62 cases (**Supplementary Fig. 2b**). Differences in UTR length were often due to alternative splicing events restricted to the UTR.

High sequence fidelity of reconstructed transcripts

We measured the assembled transcript base error rate by aligning the full-length transcripts to the corresponding reference genome (using BLAT), and capturing mismatches, insertions and deletions from the highest scoring alignment (**Supplementary Table 3**). In fission yeast, rates of mismatches, insertions and deletions are each <1 in 10,000. In mouse, rates were approximately twice as high, reflecting the lower transcript fold-coverage. As the raw read error rate is ~1%, Trinity thus resolved ~99% of sequencing errors.

Comparing Trinity's performance to other methods

We compared Trinity's performance to that of other assemblers by several measures. First, we examined the number of reference transcripts reconstructed to full-length by each method ('sensitivity'). In *S. pombe*, Trinity outperformed the *de novo* sequence assemblers, ABySS[1], Trans-ABySS[27] and SOAPdenovo[6], as well as the mapping-first programs Scripture[3] and Cufflinks[2] (**Fig. 5a**). Trinity performed well across a range of 10 M to the full 150 M input sequence reads, whereas the alternative methods tended to peak at ~50 M pairs or smaller inputs (**Supplementary Fig. 6a**). In mouse (Ref-Seq annotation set, **Fig. 5b**), Trinity (8,185 transcripts; 7,749 genes) outperformed the other *de novo* assembly methods ABySS (5,561; 5,500), Trans-ABySS (7,025; 6,598) and SOAPdenovo (761; 760), with the mapping-first programs Cufflinks (9,010; 8,536) and Scripture (9,086; 8,293) exhibiting better sensitivity. Furthermore, Trinity and Cufflinks appear best-tuned in their sensitivity across the broadest range of expression levels (**Supplementary Fig. 7**). Unlike Trinity, several of the *de novo* methods did not perform well in fully reconstructing transcripts within the highest expression quintiles (**Supplementary Fig. 7**).

Second, we assessed the accuracy of splice pattern detection. We mapped all the reconstructed transcripts (annotated or not) back to the reference genome and considered each individual intron or the combinations of introns (splicing patterns) defined by this mapping (**Fig. 5c–f**). We compared the number of annotated reference introns (or splicing patterns) captured by each method (**Fig. 5c–f**, *y* axis), and the number of previously unannotated introns (or extended splicing patterns) defined by each method's transcripts (**Fig. 5c–f**, *x* axis). Unannotated introns or splice patterns captured by more than one method are less likely to be false positives. In *S. pombe*, Trinity identified the largest number of reference introns (4,543) (**Fig. 5c**) and 1,582 unannotated introns, most of which are in putative, unannotated UTRs. Of these, 1,174 (74%) are also identified by at least one other method and thus are more likely genuine. Trinity also identifies the largest number of annotated splicing patterns in *S. pombe* (**Fig. 5e**). The alternative methods also report large numbers of falsely fused *S. pombe* transcripts, which are distinct transcripts encoded by adjacent genes that are reported as a single merged transcript by the assemblers. These contribute to the lack of sensitivity of the alternative methods.

In mouse, most methods had similarly high sensitivity for detecting individual annotated introns (**Fig. 5d**), but varied in detecting complete splicing patterns (**Fig. 5f**). Scripture identifies the most annotated splicing patterns (7,274), closely followed by Trinity (7,127). However, Scripture reports >110,000 unique splicing patterns, about tenfold more than Trinity and all other methods (each less than 10,000 unique patterns), suggesting many false positives in Scripture, and excellent precision in Trinity. Overall, relatively few of the nonannotated splicing patterns predicted by each method are supported by at least one other method (18–25%). (The notable exceptions were the particularly low fraction for Scripture (2%) and high fraction for ABySS (66%)).

Finally, we examined the number of distinct contigs that mapped to each reference genomic locus, as well as the coverage (tiers) of reconstructed transcripts per locus. This accounts for multiple reported transcripts that represent the same region of a locus owing, for example, either to alternative splicing, captured allelic variation or enumerating transcripts with otherwise undetected sequencing errors. In *S. pombe*, Trinity reports 7,057 transcripts that map to 4,874 genes with an average coverage of 1.37 tiers per gene, similar to all the alternative

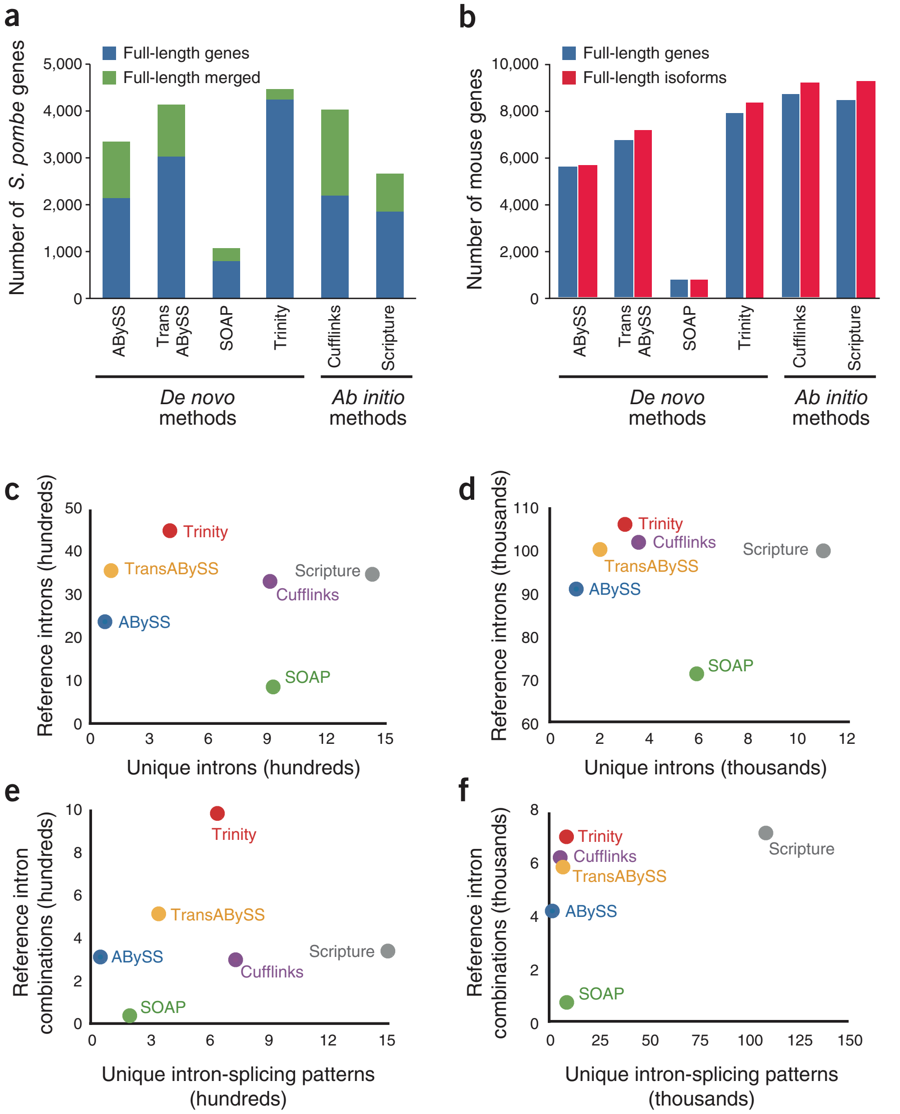
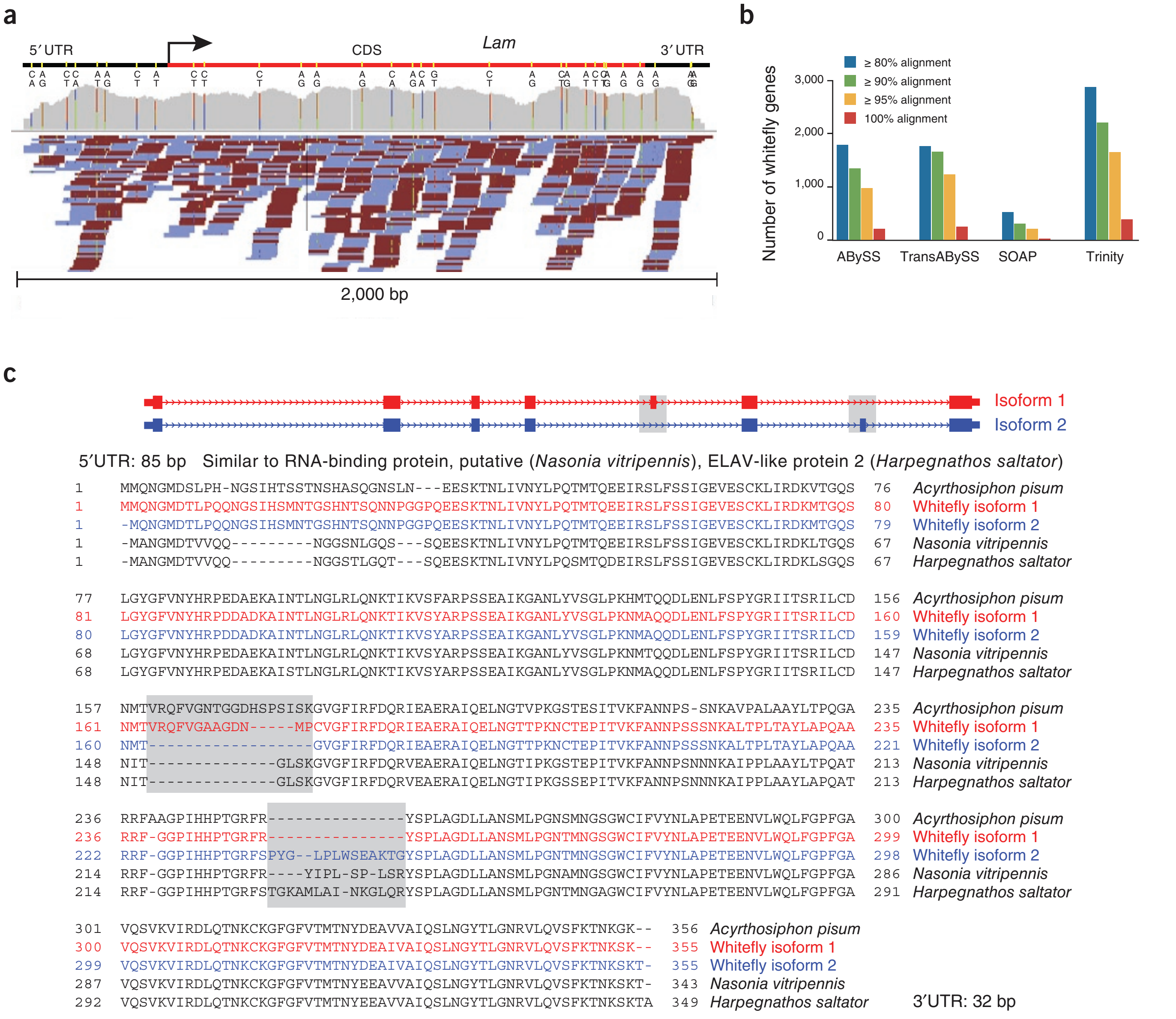


Figure 5 Comparison of Trinity to other mapping-first and assembly-first methods. (**a**,**b**) Evaluation based on number of full-length annotated transcripts reconstructed by each method in *S. pombe* (50 M read pair assemblies) (**a**) and mouse (53 M read pair assemblies) (**b**). Number of genes reconstructed in full length (blue) or as fusions of two full-length genes (green, yeast only) and the number of full-length reconstructed transcript isoforms (red, mouse only) in each of four assembly-first (*de novo*) and two mapping-first approaches. (**c**,**d**) Evaluation based on the number of introns defined by the transcripts from each method for *S. pombe* (**c**) and mouse (**d**). Shown is the number of distinct introns consistent with the reference annotation (*y* axis) versus the number of uniquely predicted introns (*x* axis), based on mapping to the genome of the transcripts reconstructed by the different methods. (**e**,**f**) Evaluation based on the number of splicing patterns (complete sets of introns in multi-intronic transcripts) defined by the transcripts from each method for *S. pombe* (**e**) and mouse (**f**). Shown are the numbers of distinct splicing patterns (*y* axis) consistent with the reference annotation versus the number of unique splicing patterns (*x* axis), for each method.



5′UTR: 85 bp Similar to RNA-binding protein, putative (*Nasonia vitripennis*), ELAV-like protein 2 (*Harpegnathos saltator*)

```
1     MMQNGMDSLPH-NGSIHTSSTNSHASQGNSLN---EESKTNLIVNYLPQTMTQEEIRSLFSSIGEVESCKLIRDKVTGQS   76  Acyrthosiphon pisum
1     MMQNGMDTLPQQNGSIHSMNTGSHNTSQNNPGGPQEESKTNLIVNYLPQTMTQEEIRSLFSSIGEVESCKLIRDKMTGQS   80  Whitefly isoform 1
1     -MQNGMDTLPQQNGSIHSMNTGSHNTSQNNPGGPQEESKTNLIVNYLPQTMTQEEIRSLFSSIGEVESCKLIRDKMTGQS   79  Whitefly isoform 2
1     -MANGMDTVVQQ---------NGGSNLGQS---SQEESKTNLIVNYLPQTMTQEEIRSLFSSIGEVESCKLIRDKLTGQS   67  Nasonia vitripennis
1     -MANGMDTVVQQ---------NGGSTLGQT---SQEESKTNLIVNYLPQSMTQDEIRSLFSSIGEVESCKLIRDKLSGQS   67  Harpegnathos saltator

77    LGYGFVNYHRPEDAEKAINTLNGLRLQNKTIKVSFARPSSEAIKGANLYVSGLPKHMTQQDLENLFSPYGRIITSRILCD  156  Acyrthosiphon pisum
81    LGYGFVNYHRPDDADKAINTLNGLRLQNKTIKVSYARPSSEAIKGANLYVSGLPKNMAQQDLENLFSPYGRIITSRILCD  160  Whitefly isoform 1
80    LGYGFVNYHRPDDADKAINTLNGLRLQNKTIKVSYARPSSEAIKGANLYVSGLPKNMAQQDLENLFSPYGRIITSRILCD  159  Whitefly isoform 2
68    LGYGFVNYHRPEDAEKAINTLNGLRLQNKTIKVSYARPSSEAIKGANLYVSGLPKNMTQQDLENLFSPYGRIITSRILCD  147  Nasonia vitripennis
68    LGYGFVNYHRPEDAEKAISTLNGLRLQNKTIKVSYARPSSEAIKGANLYVSGLPKNMAQQDLENLFSPYGRIITSRILCD  147  Harpegnathos saltator

157   NMTVRQFVGNTGGDHSPSISKGVGFIRFDQRIEAERAIQELNGTVPKGSTESITVKFANNPS-SNKAVPALAAYLTPQGA  235  Acyrthosiphon pisum
161   NMTVRQFVGAAGDN-----MPCVGFIRFDQRIEAERAIQELNGTTPKNCTEPITVKFANNPSSSNKALTPLTAYLAPQAA  235  Whitefly isoform 1
160   NMT------------------GVGFIRFDQRIEAERAIQELNGTTPKNCTEPITVKFANNPSSSNKALTPLTAYLAPQAA  221  Whitefly isoform 2
148   NIT--------------GLSKGVGFIRFDQRVEAERAIQELNGTIPKGSTEPITVKFANNPSNNNKAIPPLAAYLTPQAT  213  Nasonia vitripennis
148   NIT--------------GLSKGVGFIRFDQRVEAERAIQELNGTIPKGSSEPITVKFANNPSNNNKAIPPLAAYLAPQAT  213  Harpegnathos saltator

236   RRFAAGPIHHPTGRFR---------------YSPLAGDLLANSMLPGNSMNGSGWCIFVYNLAPETEENVLWQLFGPFGA  300  Acyrthosiphon pisum
236   RRF-GGPIHHPTGRFR---------------YSPLAGDLLANSMLPGNTMNGSGWCIFVYNLAPETEENVLWQLFGPFGA  299  Whitefly isoform 1
222   RRF-GGPIHHPTGRFSPYG--LPLWSEAKTGYSPLAGDLLANSMLPGNTMNGSGWCIFVYNLAPETEENVLWQLFGPFGA  298  Whitefly isoform 2
214   RRF-GGPIHHPTGRFR----YIPL-SP-LSRYSPLAGDLLANSMLPGNAMNGSGWCIFVYNLAPETEENVLWQLFGPFGA  286  Nasonia vitripennis
214   RRF-GGPIHHPTGRFSTGKAMLAI-NKGLQRYSPLAGDLLANSMLPGNTMNGAGWCIFVYNLAPETEENVLWQLFGPFGA  291  Harpegnathos saltator

301   VQSVKVIRDLQTNKCKGFGFVTMTNYDEAVVAIQSLNGYTLGNRVLQVSFKTNKGK--  356  Acyrthosiphon pisum
300   VQSVKVIRDLQTNKCKGFGFVTMTNYDEAIVAIQSLNGYTLGNRVLQVSFKTNKSK--  355  Whitefly isoform 1
299   VQSVKVIRDLQTNKCKGFGFVTMTNYDEAIVAIQSLNGYTLGNRVLQVSFKTNKSKT-  355  Whitefly isoform 2
287   VQSVKVIRDLQTNKCKGFGFVTMTNYEEAVVAIQSLNGYTLGNRVLQVSFKTNKSKT-  343  Nasonia vitripennis
292   VQSVKVIRDLQTNKCKGFGFVTMTNYEEAVVAIQSLNGYTLGNRVLQVSFKTNKSKTA  349  Harpegnathos saltator        3′UTR: 32 bp
```

Figure 6 Trinity reconstructs polymorphic transcripts in whitefly. (**a**) Allelic variation evident from mapping RNA-Seq reads to a full-length whitefly transcript reconstructed by Trinity. At the top is a schematic of a single transcript orthologous to the *Drosophila melanogaster* Lamin gene *Lam*, identified by grouping reconstructed transcripts having allelic variants (colored yellow). Gray coverage plot shows cumulative read coverage along the transcripts. SNPs are marked with colored bars and scaled based on the relative proportions of each variant (blue: C, red: T, orange: G, green: A). Individual reads are shown below coverage plot (forward reads, blue; reverse, red). (**b**) Comparison of performance for *de novo* assembly of the whitefly transcriptome. The *y* axis is a count of the unique top-matching (BLASTX) uniref90 (ref. 20) protein sequences aligned Trinity transcripts across a minimal percent of their length. (**c**) Example of two alternatively spliced transcripts resolved even in the absence of a reference genome. Shown are two isoforms of an ELAV-like gene reconstructed by Trinity (gray boxes indicate alternative exons). Exon structure is determined for visualization by the *D. melanogaster* ortholog. The protein sequence alignment shows the similarity between the two whitefly isoforms and orthologous proteins from other insects, and it confirms the splice variants (gray boxes).

methods except Scripture (4.37 tiers per gene) and trans-ABySS (5.08 tiers per gene). In mouse, the performance of Trinity (31,706 contigs map to 11,334 genes, 2.05 tiers per gene on average) is similar to that of all other methods except trans-ABySS (111,000 contigs, 10,685 genes, 5.93 tiers). The large numbers of Trans-ABySS transcripts covering similar regions of loci is not reflected in the number of distinct splicing patterns, indicating that multiple similar transcript sequences are being generated at individual loci, rather than many different splice isoforms. ABySS alone, although lacking the higher sensitivity of Trans-ABySS, reports a smaller number of contigs (~1 transcript tier per locus).

De novo assembly of the whitefly transcriptome

In the absence of a sequenced genome, *de novo* assembly of RNA-Seq is the only viable option to study the transcriptomes of most organisms to date. For example, although the highly diverse class Insecta contains several key model organisms, it is not densely covered by high-quality draft genome sequences. In addition, insect transcriptomes exhibit complex alternative splicing patterns[28]. The whitefly *B. tabaci* is one such example; the genome was not sequenced, and the RNA-Seq samples are genetically polymorphic, as they are derived from a mixture of individuals from an outbred population[28].

We applied Trinity to a published RNA-Seq data set from whitefly, consisting of ~21.9 million pairs of 76-base Illumina reads, sequenced using conventional non-strand-specific methods[29]. Trinity produced 196,000 transcripts, 14,522 >1,000 base pairs, capturing allelic variants (**Fig. 6a**). Of those, 4,323 had top BLASTX matches ($E \leq 10^{-10}$) to 2,880 unique Uniref90 (ref. 20) protein sequences, along at least 80% of the corresponding homologous protein sequence. This number of approximately full-length Trinity-assembled transcripts is substantially higher than achieved by other *de novo* assemblers (**Fig. 6b**).

To assess the extent to which alternative splice forms are captured by the Trinity assembly, we aligned all pairs of contigs derived from individual graph components, and searched for evidence of at least one alternative internal exon of minimum length 21 bp and a multiple of 3. By this definition, 325 components contain at least two different isoforms. One such example (**Fig. 6c**) is a highly conserved ortholog to an ELAV-like protein in the ant *Harpegnathos saltator*, which is present as two different isoforms involving inclusion of two different, alternatively spliced exons.

DISCUSSION

We presented Trinity, a method for *de novo* reconstruction of the majority of full-length transcripts in a sample from RNA-Seq reads directly, across a broad range of expression levels. Trinity resolved ~99% of the initial sequencing errors, determined splice isoforms, distinguished transcripts from recently duplicated and identified allelic variants. Unlike existing short-read assembly tools initially developed for genome assembly, Trinity was designed specifically for transcriptome assembly. To this end, Trinity leverages several properties of transcriptomes in its assembly procedure: it uses transcript expression to guide the initial Inchworm transcript assembly procedure in a strand-specific manner, it partitions RNA-Seq reads into sets of disjoint transcriptional loci, and it traverses each of the transcript graphs systematically to explore the sets of transcript sequences that best represent variants resulting from alternative splicing or gene duplication by exploiting pairs of RNA-Seq reads.

Trinity's transcripts substantially enhance our annotation of the mouse and fission yeast transcriptomes. In yeast, we identified a large number of UTR extensions, antisense transcripts and novel intergenic transcripts. In mouse, we identified many novel transcripts and novel exons for reference transcripts. Trinity reconstructed many full-length transcripts from the whitefly transcriptome in the presence of substantial polymorphisms, as well as alternatively spliced variants.

Paired-reads are important to increase the distance at which Trinity can resolve ambiguities. For example, a component representing two paralogous genes (e.g., **Fig. 4**) or alternative isoforms can have an enormous number of possible paths, but often only very few of them represent real transcripts. Read pairs, representing longer fragments allow us to resolve differences (e.g., two pairs of single nucleotide polymorphisms (SNPs), or two different exons) that occur at that distance or below. At longer distances, there is no physical unit to support alternative paths, although similarity in expression levels could be used in the future, as well as longer reads and fragments from improved high-throughput sequencing technologies.

Evaluating the performance of transcript assemblers introduces several challenges, primarily because many transcripts, especially alternative isoforms, are not thoroughly defined as part of existing genome annotations. To address these challenges we used several complementary benchmarks. Our Oracle Set allowed us to assess sensitivity, by defining a 'gold standard' of expressed annotated transcripts present at full length. To assess our ability to reconstruct other reference transcripts, we considered the number of reference loci to which reconstructed transcripts map, and the coverage (tiers) of reconstructed transcripts per locus. Finally, we assessed precision by considering all the reconstructed transcripts and the number of 'correct' intron boundaries and splice patterns. Each measure represents a useful benchmark, and showed that Trinity performs better than other *de novo* methods and on par with mapping-first methods depending on the organism.

Trinity is important for both genome annotation and the study of non-model organisms. For example, all but two vertebrate genomes are available only as unfinished drafts, containing sequence gaps, scaffolds that cannot be anchored to chromosomes and assembly errors[30]. Each of these limitations hinders genome annotation and read mapping. We expect that new genomes, assembled from next-generation, high-throughput sequencing data, will be even more fragmented. Thus, high-quality *de novo* transcriptome reconstruction, as implemented in Trinity, featuring low base-error rates and the ability to capture multiple isoforms, will prove crucial to maintain acceptable levels of accuracy when characterizing genes. Finally, genomic sequences are available for only a tiny fraction of the enormous variety of organisms. Trinity provides an effective starting point to examine the transcriptomes of such species.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/nbt/index.html.

Accession Code. GEO (mouse data): GSE29209; SRA (fission yeast data): SRP005611. Trinity and its open source code are publicly available at http://TrinityRNASeq.sourceforge.net

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank L. Gaffney for help with figure preparation, J. Bochicchio for project management, the Broad Sequencing Platform for all sequencing work, A. Papanicolaou and M. Ott for Inchworm software testing and code enhancements, and F. Ribeiro for helpful discussions regarding error pruning. The work was supported in part by a grant from the National Human Genome Research Institute (NIH 1 U54 HG03067, Lander), the Howard Hughes Medical Institute, a National Institutes of Health PIONEER award, a Burroughs Wellcome Fund–Career Award at the Scientific Interface (A.R.), the US-Israel Binational Science Foundation (N.F. and A.R.), and funds from the National Institute of Allergy and Infectious Diseases under contract no. HHSN27220090018C. M.Y. was supported by a Clore Fellowship. K.L.-T. is a recipient of the European Young Investigator Award (EYRYI) funded by the European Science Foundation. A.R. is a researcher of the Merkin Foundation for Stem Cell Research at the Broad Institute.

AUTHOR CONTRIBUTIONS

M.G.G., M.Y., B.J.H., K.L.-T., N.F. and A.R. conceived and designed the study. B.J.H., M.G.G. and M.Y. developed the Inchworm, Chrysalis and Butterfly components, respectively. N.R., F.D.P., B.W.B., C.N., K.L.-T. contributed to the study's conception and execution. J.Z.L., D.A.T., X.A., L.F., R.R., I.A., N.H., A.R. and A.G. designed and performed all experiments. Q.Z., Z.C. and E.M. contributed computational analyses. M.G.G., B.J.H. and M.Y. designed, implemented and evaluated all methods. A.R., N.F., M.G.G., B.J.H. and M.Y. wrote the manuscript, with input from all authors. A.R. and N.F. contributed equally to this paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/nbt/index.html.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Birol, I. *et al.* De novo transcriptome assembly with ABySS. *Bioinformatics* **25**, 2872–2877 (2009).
2. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
3. Guttman, M. *et al.* Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **28**, 503–510 (2010).

4. Haas, B.J. & Zody, M.C. Advancing RNA-Seq analysis. *Nat. Biotechnol.* **28**, 421–423 (2010).
5. 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).
6. Li, R. *et al.* SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
7. De Bruijn, N.G. A combinatorical problem. *Koninklijke Nederlandse Akademie v. Wetenschappen* **46**, 758–764 (1946).
8. Good, I.J. Normal recurring decimals. *J. Lond. Math. Soc.* **21**, 167–169 (1946).
9. Pevzner, P.A., Tang, H. & Waterman, M.S. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. USA* **98**, 9748–9753 (2001).
10. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
11. Butler, J. *et al.* ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res.* **18**, 810–820 (2008).
12. Hertz-Fowler, C. *et al.* GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res.* **32**, D339–D343 (2004).
13. Levin, J.Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7**, 709–715 (2010).
14. Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37**, e123 (2009).
15. Rhind, N. *et al.* Comparative functional genomics of the fission yeasts. *Science* published online, doi:10.1126/science.1203357 (21 April 2011).
16. Wang, E.T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
17. Wilhelm, B.T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239–1243 (2008).
18. Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033–1037 (2009).
19. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859–1875 (2005).
20. Wu, C.H. *et al.* The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.* **34**, D187–D191 (2006).
21. Wapinski, I., Pfeffer, A., Friedman, N. & Regev, A. Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449**, 54–61 (2007).
22. Molnar, M. *et al.* Characterization of rec7, an early meiotic recombination gene in *Schizosaccharomyces pombe. Genetics* **157**, 519–532 (2001).
23. Nakamura, T., Kishida, M. & Shimoda, C. The *Schizosaccharomyces pombe* spo6+ gene encoding a nuclear protein with sequence similarity to budding yeast Dbf4 is required for meiotic second division and sporulation. *Genes Cells* **5**, 463–479 (2000).
24. Watanabe, T. *et al.* Comprehensive isolation of meiosis-specific genes identifies novel proteins and unusual non-coding transcripts in *Schizosaccharomyces pombe. Nucleic Acids Res.* **29**, 2327–2337 (2001).
25. Yassour, M. *et al.* Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species. *Genome Biol.* **11**, R87 (2010).
26. Matlin, A.J., Clark, F. & Smith, C.W.J. Understanding alternative splicing: towards a cellular code. *Nat. Rev. Mol. Cell Biol.* **6**, 386–398 (2005).
27. Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nat. Methods* **7**, 909–912 (2010).
28. Graveley, B.R. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* **17**, 100–107 (2001).
29. Wang, X.-W. *et al.* De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics* **11**, 400 (2010).
30. Salzberg, S.L. & Yorke, J.A. Beware of mis-assembled genomes. *Bioinformatics* **21**, 4320–4321 (2005).

ONLINE METHODS

Inchworm. Inchworm decomposes each sequence read into overlapping *k*-mers (default $k = 25$). Each *k*-mer is stored in a hash table as a key-value pair, where the key is the *k*-mer sequence and the value is the abundance of that *k*-mer in the input data set. The *k*-mer key is stored as a 64-bit unsigned integer with 2-bit nucleotide encoding. Likely sequencing error-containing *k*-mers are identified by examining *k*-mers that have identical $k - 1$ prefixes, differing only at their terminal nucleotide, and removing those *k*-mers that are <5% abundant as compared to the most highly abundant *k*-mer of the group. After processing the entire read set into a set of *k*-mers and pruning the likely error *k*-mers, the most frequently occurring *k*-mer is identified as a seed *k*-mer for reconstruction of draft transcript contigs. The information content of the seed *k*-mer is computed as Shannon's Entropy[31], and only *k*-mers having entropy $H \geq 1.5$, occurring at least twice in the complete set of input reads, and not palindromic, are allowed as seed *k*-mers. The seed *k*-mer is extended at both ends in a coverage-guided manner, first from 5′ to 3′, followed by extension from 3′ to 5′. Seed selection by Inchworm was largely inspired by similar methods implemented in the RepeatScout algorithm[32]. Extension from the seed is performed greedily based on the frequencies of candidate overlapping *k*-mers, with the single most abundant *k*-mer with $(k - 1)$ overlap chosen to provide a single-base extension. In the case of tied extensions, paths are recursively explored to identify the extension yielding the cumulatively maximal coverage. Extension continues until no *k*-mer exists in the data set to provide an extension. The sequence yielded from the bidirectional seed *k*-mer extension is reported as a draft transcript contig, and the set of overlapping *k*-mers comprising the contig are removed from the hash table. The entire cycle of seed selection and bidirectional *k*-mer extension continues until all *k*-mers in the hash table have been exhausted.

In strand-specific mode (default), *k*-mers are derived from only the sense strand of the RNA-Seq read. Double-stranded mode, used with non-strand-specific RNA-Seq data involves several modifications: both the sense and the reverse-complemented read sequence are parsed into overlapping *k*-mers; during Inchworm contig extension, a *k*-mer chosen to extend a given path has the reverse-complemented *k*-mer sequence disabled for further *k*-mer extensions; and when an Inchworm contig is reported at the end of one iteration of contig assembly, both the sense and reverse-complemented *k*-mers are removed from the *k*-mer dictionary.

Only Inchworm contigs with an average *k*-mer coverage of 2 and length at least 48 ($2^*(k - 1)$, $k = 25$), the minimal contig length required to capture variation anchored by $(k - 1)$ at each terminus, are used by Chrysalis, as described below.

Chrysalis. To convert the linear contigs into a proper de Bruijn graph, Chrysalis first builds a $k - 1$-mer lookup table and recursively pools contigs that share sequences (excluding low-complexity sequence, as above in Inchworm) into components, given that there are reads that span across a potential junction (the 'welds') and extend perfect matches by $(k - 1)/2$ bases on each side. The number of welds must exceed 0.04 times the average $k - 1$-mer coverage of each contig (twice the sequencing error rate in a read, the upper bound of which we estimate at ~2%), as computed by Inchworm. In addition, the $k - 1$-mer coverage of one contig cannot exceed the coverage of the other by a factor of 100 (empirically determined). Next, Chrysalis processes each component individually and computes a de Bruijn graph from the linear inchworm contigs. The reads are then mapped to components by selecting the component that shares the most $k - 1$-mers with the read, with a single $k - 1$-mer being sufficient for assignment. Chrysalis also counts all *k*-mers and stores them as 'edge weight' to indicate their support in the read set. Components with less than a minimum number of nodes are discarded (a configurable parameter that defaults to an empirically determined value of $300 - (k - 1) = 276$).

Butterfly. The input to butterfly is a de Bruijn graph component as built by Chrysalis. First, Butterfly trims edges in the de Bruijn graph. It uses two criteria. (1) We reasoned that if there is a node with several outgoing edges, such that one of them has a much smaller read support than the total outgoing reads (less than 5%), then it probably represents a sequencing error or a variant with very low expression (**Supplementary Fig. 8a**). (2) If the outgoing edge has less than 2% support from the total incoming reads, then it is more likely a spurious transcript extension (**Supplementary Fig. 8b**). Outgoing or incoming edges that fail according to one of these criteria are removed (both these numbers are parameters to the program, and can be changed for specific requirements).

Second, Butterfly transforms the modified graph into a weighted sequence graph, where each node is a sequence, rather than an individual *k*-mer providing a single-base path extension as in the de Bruijn graph. In this step, Butterfly generates a compact graph—the set of paths in the compacted graph is identical to that of the original de Bruijn graph. As a result, linear paths will be compacted into a single node, and polymorphisms will be minimized. The weight on each edge of the modified graph corresponds to the number of reads supporting the edge in the original de Bruijn graph. For each compound node, we compute the average coverage, which corresponds to the weights of the original edges that made up the sequence divided by the length of the node.

We then repeat the trimming step, except that when examining compound nodes of length >1, we also use the node coverage as a measure of opposite flow in the second criterion. These two steps (trimming and graph compaction) are reiterated until convergence. The resulting graph represents possible transcripts as paths through the graph.

Finally, Butterfly uses read sequences, read-pairings and Chrysalis' read mappings to the graph to select the paths that are best supported by read sequences. The goal is to look for paths with physical evidence for contiguity, by either reads or read pairings. To do so, we first represent all the reads that contributed to the de Bruijn graph by the list of the nodes that they traverse. We then use a dynamic programming algorithm for finding supported path prefixes. The procedure is initialized with source nodes in the graphs (one without incoming edges), and at each step one path prefix is extended by an additional node.

When extending a path prefix that ends at node *n,* we consider all outgoing edges from *n*, and evaluate the support for the extension. By construction, each edge in the graph is supported by reads. We however, further require that the last *L* nucleotides of the path be supported by reads. We define a path as *L*-supported at coverage *c* if at each extension of this path, we have at least c reads supporting the *L* nucleotide suffix of this path (**Supplementary Fig. 8c**). A read supports a path fragment either if it contains that fragment as a subsequence, or in the case of paired-reads, if the fragment lies on all paths from nodes that correspond to the first sequence mate to the second sequence mate. In addition, to avoid combinatorial explosion because of small variations (most likely caused by sequence errors), once we extend a path prefix, we examine other paths ending at the same node, and merge the new path with previous path prefix ending at the same node if the two are >95% identical.

In the results here we used $L = 250$ and $c = 2$. The requirement for 250-supported paths emerges from the expected insert size of our library, as we do not expect to have support for a longer suffix if our read pairs (derived from a single fragment) do not span that far. We note that the resolution of ambiguities, which includes alternative splicing and allelic variation, is limited to the insert size of the read pairs, or the read lengths for unpaired data. Although this program can be in theory exponential in size, in practice its cost is defined by the number of supported paths.

Yeast and mouse cell growth conditions. We used the *S. pombe* strain SPY73 975h+ and dendritic cells isolated from C57BL/6J mice. Details of cell isolation and growth conditions are in the **Supplementary Methods**.

RNA isolation for yeast samples. Total yeast RNA was isolated using Qiagen RNeasy kit following manufacturers' protocol for mechanical lysis using 0.5 mm zirconia/silica beads (Biospec). PolyA$^+$ RNA was isolated from total RNA using Poly(A) purist kit (Ambion) or Dynabeads mRNA purification kit (Invitrogen). Total RNA and polyA$^+$ RNA were treated with Turbo DNA-*free* (Ambion), as described. The integrity of the RNA was confirmed using the Agilent 2100 Bioanalyzer and quantified using RNA Quant-It assay for the Qubit Fluorometer (Invitrogen).

RNA preparation for mouse RNA. Dendritic cells were lysed using QIAzol reagent and total RNA was extracted the miRNeasy kit's procedure (Qiagen), sample quality was controlled on a 2100 Bioanalyzer (Agilent).

RNA-Seq library preparation. For the mouse dendritic cell sample, we created a dUTP second strand library starting from 200 ng of Turbo DNase treated and poly(A)$^+$ RNA using a previously described method[14] except that we fragmented RNA in 1× fragmentation buffer (Affymetrix) at 80 °C for 4 min, purified and concentrated it to 6 µl after ethanol precipitation. For the *S. pombe* samples, we prepared dUTP second-strand libraries similarly, with the following additional modifications. We added an index (8-base barcode) to each library to enable pooling of these libraries (S. Fisher, Broad Institute, personal communication). In addition, the adaptor ligation step was done with 1.2 µl of index adaptor mix and 4,000 cohesive end units of T4 DNA Ligase (New England Biolabs) overnight at 16 °C in a final volume of 20 µl. Finally, we generated libraries with an insert size ranging from 225 to 425 bp.

RNA-Seq library sequencing. We sequenced all the cDNA libraries with an Illumina Genome Analyzer IIx. We pooled the four *S. pombe* libraries together with four other indexed libraries and sequenced them using eight lanes of 76-base paired reads. We sequenced the mouse library using two lanes of 76-base paired reads.

Defining empirical limits of full-length transcript reconstruction. Inchworm was used to construct a *k*-mer dictionary based on the input reads as described above. Reference protein-coding sequences were examined by searching for each overlapping *k*-mer sequence in the dictionary. Reference protein-coding sequences lacking at least one *k*-mer in the Inchworm *k*-mer graph were classified as inaccessible for full-length reconstruction by means of the *k*-mer graph method. Those reference sequences fully represented within the *k*-mer dictionary were included in the Oracle Set.

Finding paralogous genes in mouse. To determine paralogous transcripts, we aligned all isoforms of all genes present in the Oracle Set against each other, using the alignment program *Satsuma*[33] We required alignments to be longer than half of the shorter of both sequences and at sequence identity of 70% and up. If at least one pair of transcripts from two genes met the criteria, we called both genes paralogous.

Short-read spliced alignments and transcript reconstructions using Cufflinks and Scripture. The *S. pombe* genome was obtained from the Sanger Institute (http://www.sanger.ac.uk/Projects/S_pombe/download.shtml). The mouse genome version 9 was obtained from the UCSC mouse genome browser gateway (http://genome.ucsc.edu/cgi-bin/hgGateway?db=mm9). Left and right fragment reads were separately aligned to the genomes using TopHat (version 1.1.4)[34] with mouse RNA-Seq reads, and BLAT with *S. pombe* RNA-Seq reads; the BLAT short-read alignment pipeline is provided at http://inchworm.sourceforge.net/blat_short_read_alignment.html . We found BLAT to provide more accurate short-read alignment with *S. pombe*, with TopHat lacking sensitive detection of the very short introns in *S. pombe*. In addition, both Scripture and Cufflinks demonstrated better performance using the BLAT alignments for *S. pombe* as compared to the TopHat alignments (**Supplementary Fig. 9a**). Conversely, performance of Scripture and Cufflinks using TopHat alignments in mouse exceeded that using BLAT alignments (**Supplementary Fig. 9b**). Hence, for evaluation purposes, we leveraged BLAT short-read spliced alignments in *S. pombe* and TopHat alignments in mouse.

BLAT alignments of short reads to the *S. pombe* genome were performed using the pipeline described above with the following settings: maximum intron length set to 500 bases, maximum distance between read pairs of 500, and only the single best alignment was reported per read. TopHat alignments to the mouse genome were performed using the following parameters: minimum intron length of 50 bases, maximum intron of 100 kb and mate inner distance set to 300 bases. Transcribed strand information was assigned to the individual reads based on knowledge of the fragment type (left or right) and the aligned strand of the genome. Both Cufflinks (version 0.9.3)[2] and Scripture[3] (version VPaperR3, obtained from Scripture author Manuel Garber) were executed on these alignments.

Evaluation of published *de novo* methods. Illumina reads were *de novo* assembled using ABySS[1] (version 1.2.1), SOAPdenovo[6] (version 1.04) or Trans-ABySS[27]. Command-line parameters used with ABySS were "abyss-pe k=25 E=0 n=10 in='left.fa right.fa' ", using a *k*-mer length of 25. Likewise, a 25-mer length was used with SOAPdenovo along with other default parameters. Trans-ABySS[27] was run on mouse and *S. pombe* using a set of *k*-mers including 26, 31, 36, 41 and 46 followed by merging the results by running the first stage of the trans-ABySS analysis pipeline. In the case of whitefly, all *k*-mers from 26 through 46 were used so as to maximize sensitivity given the smaller input number of reads.

Comparisons to reference transcripts. Current gene annotations for *S. pombe* were downloaded as file 'pombe_290110.gff' from GeneDB (http://old.genedb.org/genedb/pombe/). Ref-Seq transcript gene annotations were downloaded for mouse at the UCSC mouse genome browser gateway (http://genome.ucsc.edu/cgi-bin/hgGateway?db=mm9) in BED format. Protein coding nucleotide sequences were extracted from the genome sequences based on the gene annotations using custom PERL scripts. The mouse reference coding sequences were further distilled to remove entirely identical sequences corresponding to isoforms encoding identical proteins and paralogous sequences: the original 19,947 genes encoding 23,881 transcripts were reduced to 19,857 genes encoding 22,717 non-identical coding transcripts.

Reconstructed transcript sequences (by *de novo* assembly, Scripture or Cufflinks) were mapped to the reference coding sequences using BLAT[35]. Full-length reference annotation mappings were defined as having at least 95% sequence identity covering the entire reference coding sequence and containing at most 5% insertions or deletions (cumulative gap content). In evaluating methods that leverage the strand-specific data (Trinity and Cufflinks), proper sense-strand mapping of sequences was required. Transcripts reconstructed by the alternative methods (Scripture, ABySS and SOAPdenovo) were allowed to map to either strand. Fusion transcripts were identified as individual reconstructed transcripts that mapped as full-length to multiple reference coding sequences and lacked overlap among the matching regions within the reconstructed transcript. One-to-one mappings were required between reconstructed transcripts and reference transcripts, including alternatively spliced isoforms, with the exception of fusion transcripts.

Analysis of alignment-inferred introns and splicing patterns from reconstructed transcripts. Reconstructed transcripts were mapped to genome sequences using GMAP, reporting only the single top-scoring alignment per sequence. Individual introns and complete splicing patterns were extracted from each of the alignments and compared to reference annotations using custom PERL scripts. Unique introns (missing from the reference annotations) were required to contain consensus dinucleotide splice sites (GT or GC donors and AG acceptors).

Locus coverage (tiering) by reconstructed transcripts. The BLAT alignments between reference coding sequences (loci) and reconstructed transcripts described above were organized into locus-level coverage tiers as follows. Given a set of different reconstructed transcripts that have a best match to a reference sequence, the first match is selected and applied to that reference contig at the first coverage tier. The remaining matches are then examined for placement in the first tier. If a subsequent reference-matching region in common between two matches exceeds 30% of the shorter match length, then this subsequent match is propagated to the next highest tier lacking such restrictive match overlap. Tier placement continues until all matches are placed. The maximal tier level defines the locus-level coverage for that reference sequence and can be at most equal to the number of reconstructed transcripts mapped to that locus. Strand-specific transcript reconstructions were tiered in a strand-specific manner (as in the case of Trinity and Cufflinks). In the case of a highly fragmented transcriptome assembly, it is possible for many reconstructed transcripts to populate the first tier yielding a coverage of 1. In the case of alternatively spliced isoforms or redundant transcript generation at a given locus, the coverage value will exceed 1.

Running Trinity on data sets of varying read depth. We randomly subsampled pairs in the mouse data set to generate such subsets. Inchworm and Chrysalis were run on a server with 256 GB of RAM, Butterfly on a server (*load sharing facility* (LSF)) farm in parallel. Wall-clock run times are: ~17 h (10 M pair set), ~36 h (30 M pair set), and ~60 h (full 50 M pair set). All experiments were performed with Trinity using parameters: minimum contig length of 100 bases and average fragment length of 300 bases.

Computing gene expression values from aligned RNA-Seq reads. The aligned reads (by TopHat in the case of mouse leveraging the full 52.6M read pairs, and by BLAT in the case of *S. pombe* leveraging the 50 M read pairs) were used for computing gene (and other feature) expression values. The number of fragments mapped to segments (exons) of a genome-mapped feature were tallied based on overlap of the segment's coordinates by either read from a sequenced fragment, counting fragments as opposed to counting individual reads. Expression was computed as the normalized value of fragments per kilobase of feature sequence per million fragments mapped, or FPKM[2]. Calculations were performed using custom PERL scripts. Genes were defined as 'expressed' if observed to have expression values of at least 0.5 FPKM, and these genes were divided into expression quintiles at 5% intervals for purposes of analysis.

31. Shannon, C.E. Prediction and entropy of printed English. *Bell Syst. Tech. J.* **30**, 50–64 (1951).
32. Price, A.L., Jones, N.C. & Pevzner, P.A. De novo identification of repeat families in large genomes. *Bioinformatics* **21** Suppl 1, i351–i358 (2005).
33. Grabherr, M.G. *et al.* Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* **26**, 1145–1151 (2010).
34. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
35. Kent, W.J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).

doi:10.1038/nbt.1883

Chapter 6

Discussion

In my dissertation I set out to develop tools for transcriptome characterization using RNA-Seq data without relying on pre-existing annotation. I have applied these tools to study a range of organisms: from the dense transcriptomes of the budding- and fission-yeast to the highly spliced and complex mouse transcriptome.

6.1 Characterizing the budding yeast transcriptome using the mapping-first approach

In the paper described in **Chapter 2** (Yassour et al., 2009) we aimed to test whether it is possible to *ab-initio* define a complete yeast transcriptome using only the (unannotated) genome sequence and massively parallel cDNA sequencing. Our approach identified 85% of expressed genes and correctly inferred 254 of the 305 known splicing events. This is impressive as not all splice junctions are used in our samples. Also, it corrected a number of current annotations and identified previously undescribed transcriptional units and splice junctions, several of which we validated experimentally. Last, the method can also accurately quantify the expression levels of transcripts.

This mapping-first approach had several limitations. First, as in all RNA-Seq based studies, we are limited to the expressed portion of the transcriptome of our sample. We partly addressed this issue by creating libraries from two physiological conditions. Second, we missed splicing events due to local non-uniqueness at the splice junction. This was at the early days of RNA-Seq and we had only single-end 32bp long reads. With current read length and paired reads, this problem is less severe. Finally, due to the lack of strand specificity our approach was limited in detecting and distinguishing antisense transcripts and differentiating between close divergent transcription units. In most cases we could recover transcript orientation from biases in read coverage along a gene, but we can further enhance the predictions by constructing strand-specific cDNA libraries, that were not available then, but are currently the standard.

Unlike previous RNA-Seq studies (Nagalakshmi et al., 2008; Mortazavi et al., 2008), we demonstrated the use of RNA-Seq for complete, *ab-initio* construction of a eukaryotic transcriptome, independent of any existing genome annotation. For example, Mortazavi et al. (2008) use a mapping approach that relies on mapping reads to known gene models, exons and splice junctions. Such approaches cannot detect splice junctions between unannotated exons.

Our work powerfully demonstrates the feasibility of constructing a transcriptome of an organism in a comprehensive, fast, and cheap way. Applying our approach to explore the transcriptomes of less characterized organisms in an *ab-initio* fashion can have a significant impact on genomics studies.

6.2 Comparing strand specific library construction methods

One of the major caveats of the work of **Chapter 2** (Yassour et al., 2009), as mentioned above, is the lack of strand specificity in the RNA-Seq data. To address this issue we evaluated existing strand specific library construction protocols (**Chapter 3**, Levin et al. (2010)). I have developed a computational framework to estimate the performance of each protocol. It is unclear how to measure the success of such protocols, as they differ greatly in the experimental work and output, and depending on our task one can be better than the other. To address this, the framework is comprised of a few metrics that address several aspects of the data: (1) the complexity of the library, specifically, how many unique reads we have, which indicates how many artifacts were introduced in the amplification step; (2) the strand specificity of the reads which was calculated by the percentage of the reads mapped to the expected strand; (3) even-ness of coverage along genes; (4) the coverage of the 5' and 3' ends of genes; and (5) correlation in expression level estimations with the microarray technology. In addition to these formal criteria, we found a substantial variation in the experimental complexity of different protocols.

We concluded that the dUTP protocol provided the most compelling overall balance across criteria, followed closely by the Illumina RNA ligation protocol. Our compendium and analysis pipeline, which are available online, are important resources, include a general benchmarking dataset and tools for testing the quality of future libraries, and have been used thus far by various labs around the world (Tariq et al., 2011; Wang et al., 2011).

6.3 Annotating antisense transcripts in the budding yeast transcriptome

Once we have identified the best protocol for strand specific RNA-Seq, I went back to explore the extent of antisense transcription in yeast (**Chapter 4**, Yassour et al. (2009)). Towards this end, I have used the strand specific RNA-Seq data from the dUTP library, generated from *Saccharomyces cerevisiae* cells grown to mid-log phase. I found 1,103 putative antisense transcripts expressed in this condition, ranging from 39 short ones covering only the 3' UTR of sense genes to 145 long ones covering the entire sense ORF. I focused on 402 long antisense units (each spanning over 75% of a coding unit). In this category, I identified 224 new antisense transcripts that in previous microarray studies (Xu et al., 2009) were either undetected or annotated as long UTRs of neighboring genes. Using the paired reads in our data, we can distinguish between UTR extensions and independent transcriptional units.

We are still unsure why so many genes have antisense transcripts. Could it be that they are all side effect of the sense transcription? The cell is investing a great deal of energy and materials into transcribing these antisense unit, thus the question of their functionality is even more interesting. To date, functional studies have identified a regulatory role for only a few antisense transcripts (Hongay et al., 2006; Camblong et al., 2007; Houseley et al., 2008). The diversity of lengths in our antisense units suggests there may be more than a single underlying mechanism for their formation and function.

Genome-wide analyses have suggested that antisense transcripts are the results of promiscuous transcription (He et al., 2008; Xu et al., 2009; Neil et al., 2009). Our results do not support promiscuous or aberrant transcription as the primary cause of the observed antisense transcripts. We find antisense transcription at only 18% of the genes. Moreover, many of the antisense units are long and show robust sequence coverage, in contrast to what we might expect in a noisy process. Finally, antisense transcripts are only very weakly correlated to their neighbors, inconsistent with the leaky transcription theory.

We found that the sense transcripts corresponding to longer antisense units are significantly enriched for key processes in *S. cerevisiae*, including stress response, the differential regulation of growth and stationary phase, and possibly meiosis and sporulation. The high level of antisense expression is consistent with the repression of these processes in fast growing yeast cells. Indeed, when we examined the relative change in expression in sense and antisense units across multiple conditions, we found a strong and consistent anti-correlation between sense genes and their corresponding antisense units.

In search for a mechanistic understanding of this potential regulation, we measured the expression levels of 67 sense and antisense pairs in the ΔRrp6 and ΔHda2 strains, as these genes were suggested to play a mechanistic role by Camblong et al. (2007). Notably, we found support for the role of Rrp6 in the regulation of antisense levels, resulting in an increase in antisense levels in the Δrrp6 mutant, and a mild decrease in sense levels. We could not demonstrate a general effect of Hda2 on the levels of sense or antisense transcripts. This suggests that it may be challenging to generalize the mechanisms shown for specific transcripts (PHO84, Camblong et al. (2007)) to all antisense transcripts.

Independent support for a potential function is the conservation of expression and regulation of six antisense units tested across five species that have diverged more than 150 million years ago, suggesting purifying selection.

Lastly, following our identification of several antisense units in meiosis related genes, I was involved in a study on the transcription and translation regulation during meiosis in yeast (Brar et al., 2012). In this work, we measured RNA-Seq and protein production through the yeast meiotic sporulation program. We found strong, stage-specific expression for most genes, achieved through control of both mRNA levels and translational efficiency. Meiotic translation is also shifted toward non-canonical sites, including short ORFs on unannnotated transcripts and upstream regions of known transcripts (*upstream ORFs*, or uORFs). This work reveals pervasive translational control in meiosis and helps to illuminate the molecular basis of the broad restructuring of meiotic cells.

Since our publication there has been growing evidence of antisense transcription in fungi (Donaldson and Saville, 2012), especially in genes related to stress and meiosis (Chen and Neiman, 2011), compatible with our findings. As more and more studies of antisense transcription are preformed, the debate regarding their functionality settles and makes way to the more interesting discussion regarding their regulation and mechanism of inhibition. A recent study by Murray et al. (2012) finds that antisense transcripts and their neighboring genes are independent in their regulation, inconsistent with Xu et al. (2009) but consistent with our conclusion. Regarding their functionality, a new study from the Steinmetz lab (Xu et al., 2011) finds that antisense transcripts assist in a complete "shut-off" of the sense genes, and that this type of inhibition specifically affects low levels of sense gene expression. Furthermore, they argue that antisense transcripts initiating from bi-directional promoters assist in spreading the repression signal to adjacent genes (Xu et al., 2011). Regarding the inhibition mechanism, several studies have found evidence that chromatin take part in this process, although much remains to be discovered. Recently, Magistri et al. (2012) find that antisense transcripts regulated their sense genes by recruiting epigenetic effectors (*e*.g., via H3K27me3 and H3K9me3), and van Dijk et al. (2011) show how H3K4me3 plays an important role in controlling the antisense repressive activity.

To conclude, it is now clear that antisense transcripts provide an additional layer of regulation, spanning from fungi to mammals, but the exact inhibition mechanisms are still unclear and remain to be fully characterized.

6.4 The development and application of an assembly-first method to characterize complex transcriptomes

In **Chapters 1-3** I have discussed only mapping-first approaches, which have some caveats, mainly the requirement of a high quality reference genome, and the difficult task of mapping spliced reads. In our recent work studying the genome and highly spliced transcriptome of the fission yeasts (Rhind et al., 2011), these caveats became major obstacles. To address these challenges we turned to the assembly-first strategy, which as explained above in details, first assembles all the RNA-Seq reads, and then maps the longer sequences to a reference genome, if such has been sequenced.

In the paper described in **Chapter 5** (Grabherr et al., 2011) we presented Trinity, a method for *de-novo* reconstruction of full-length transcripts using RNA-

Seq assembly. Unlike existing short-read assembly tools initially developed for genome assembly, Trinity was designed specifically for transcriptome assembly. To this end, Trinity leverages several properties of transcriptomes in its assembly procedure: it uses transcript expression to guide the initial Inchworm transcript assembly procedure in a strand-specific manner, it partitions RNA-Seq reads into sets of disjoint transcriptional loci, and it traverses each of the transcript graphs systematically to explore the sets of transcript sequences that best represent variants resulting from alternative splicing or gene duplication by exploiting pairs of RNA-Seq reads.

We applied Trinity to annotate the dense transcriptome of the fission yeast and the spliced and complex transcriptome of mouse. Trinity resolved ∼99% of the initial sequencing errors, determined splice isoforms, distinguished transcripts from recently duplicated and identified allelic variants. In yeast, we identified a large number of UTR extensions, antisense transcripts and novel intergenic transcripts. In mouse, we identified many novel transcripts and novel exons for reference transcripts. In addition, when applying Trinity to RNA-Seq data from whitefly, an organism with no sequenced reference genome, we reconstructed many full-length transcripts, including alternatively spliced variants, even in the presence of substantial polymorphisms.

Paired-reads are important to increase the distance at which Trinity can resolve ambiguities. Read pairs, representing longer fragments allow us to resolve differences (*e.g.*, two pairs of SNPs, or inclusion of two distant exons) that occur at that distance or below. At longer distances, there is no physical unit to support alternative paths, but future RNA-Seq libraries with longer fragment size can improve our performance greatly.

Trinity is important for both genome annotation and the study of non-model organisms. High-quality *de-novo* transcriptome reconstruction, as implemented in Trinity, featuring low base-error rates and the ability to capture multiple isoforms, will prove crucial to maintain acceptable levels of accuracy when characterizing genes. Furthermore, genomic sequences are available for only a tiny fraction of the enormous variety of organisms. Thus, Trinity provides an effective starting point to examine the transcriptomes of such species as well as aberrant cancer genomes. In the year since its publication Trinity has been used in many studies of characterizing transcriptomes, with or without a sequenced reference genome (van Bakel et al., 2011; Zhang et al., 2012; Lulin et al., 2012; Wang et al., 2012).

Bibliography

M. D. Adams, J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno, A. R. Kerlavage, W. R. McCombie, and J. C. Venter. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science (New York, N.Y.)*, 252 (5013):1651–1656, June 1991.

J. D. Anderson and J. Widom. Poly(dA-dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites. *Molecular and Cellular Biology*, 21(11):3830–3839, June 2001.

P. Bertone, V. Stolc, T. E. Royce, J. S. Rozowsky, A. E. Urban, X. Zhu, J. L. Rinn, W. Tongprasit, M. Samanta, S. Weissman, M. Gerstein, and M. Snyder. Global identification of human transcribed sequences with genome tiling arrays. *Science (New York, N.Y.)*, 306(5705):2242–2246, Dec. 2004.

G. A. Brar, M. Yassour, N. Friedman, A. Regev, N. T. Ingolia, and J. S. Weissman. High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science (New York, N.Y.)*, 335(6068):552–557, Feb. 2012.

J. Camblong, N. Iglesias, C. Fickentscher, G. Dieppois, and F. Stutz. Antisense RNA Stabilization Induces Transcriptional Gene Silencing via Histone Deacetylation in S. cerevisiae. *Cell*, 131(4):706–717, Nov. 2007.

P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V. B. Bajic, S. E. Brenner, S. Batalov, A. R. R. Forrest, M. Zavolan, M. J. Davis, L. G. Wilming, V. Aidinis, J. E. Allen, A. Ambesi-Impiombato, R. Apweiler, R. N. Aturaliya, T. L. Bailey, M. Bansal, L. Baxter, K. W. Beisel, T. Bersano, H. Bono, A. M. Chalk, K. P. Chiu, V. Choudhary, A. Christoffels, D. R. Clutterbuck, M. L. Crowe, E. Dalla, B. P. Dalrymple, B. de Bono, G. Della Gatta,

D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C. F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T. R. Gingeras, T. Gojobori, R. E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T. K. Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S. P. T. Krishnan, A. Kruger, S. K. Kummerfeld, I. V. Kurochkin, L. F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K. C. Pang, W. J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J. F. Reid, B. Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S. L. Salzberg, A. Sandelin, C. Schneider, C. Schönbach, K. Sekiguchi, C. A. M. Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S. L. Tan, S. Tang, M. S. Taylor, J. Tegner, S. A. Teichmann, H. R. Ueda, E. van Nimwegen, R. Verardo, C. L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S. M. Grimmond, R. D. Teasdale, E. T. Liu, V. Brusic, J. Quackenbush, C. Wahlestedt, J. S. Mattick, D. A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai, Y. Hayashizaki, FANTOM Consortium, and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group). The transcriptional landscape of the mammalian genome. *Science (New York, N.Y.)*, 309(5740):1559–1563, Sept. 2005.

H. M. Chen and A. M. Neiman. A conserved regulatory role for antisense RNA in meiotic gene expression in yeast. *Current Opinion in Microbiology*, 14(6): 655–659, Dec. 2011.

L. David, W. Huber, M. Granosvskaia, J. Toedling, C. J. Palm, L. Bofkin, T. Jones, R. W. Davis, and L. M. Steinmetz. A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences of the United States of America*, pages 1–6, Mar. 2006.

M. E. Donaldson and B. J. Saville. Natural antisense transcripts in fungi. *Molecular microbiology*, 85(3):405–417, Aug. 2012.

A. E. Ehrenhofer-Murray. Chromatin dynamics at DNA replication, transcription and repair. *European journal of biochemistry / FEBS*, 271(12):2335–2349, June 2004.

M. A. Faghihi and C. Wahlestedt. Regulatory roles of natural antisense transcripts. *Nature reviews. Molecular cell biology*, 10(9):637–643, Sept. 2009.

Y. Field, N. Kaplan, Y. Fondufe-Mittendorf, I. K. Moore, E. Sharon, Y. Lubling, J. Widom, and E. Segal. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Computational Biology*, 4(11): e1000216, Nov. 2008.

S. Gnerre, I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, and D. B. Jaffe. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, Jan. 2010.

M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29(7):644–652, May 2011.

M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn, and E. S. Lander. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 457(7235):223–227, Mar. 2009.

M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander, and

A. Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature biotechnology*, 28(5):503–510, May 2010.

B. J. Haas and M. C. Zody. Advancing RNA-Seq analysis. *Nature biotechnology*, 28(5):421–423, May 2010.

L. He and G. J. Hannon. MicroRNAs: small RNAs with a big role in gene regulation. *Nature reviews. Genetics*, 5(7):522–531, July 2004.

Y. He, B. Vogelstein, V. E. Velculescu, N. Papadopoulos, and K. W. Winzler. The Antisense Transcriptomesof Human Cells. *Science (New York, N.Y.)*, pages 1–3, Dec. 2008.

C. F. Hongay, P. L. Grisafi, T. Galitski, and G. R. Fink. Antisense Transcription Controls Cell Fate in Saccharomyces cerevisiae. *Cell*, 127(4):735–745, Nov. 2006.

J. Houseley, L. Rubbi, M. Grunstein, D. Tollervey, and M. Vogelauer. A ncRNA Modulates Histone Modification and mRNA Induction in the Yeast GAL Gene Cluster. *Molecular Cell*, 32(5):685–695, Dec. 2008.

W. J. Kent. BLAT–the BLAST-like alignment tool. *Genome research*, 12(4): 656–664, Apr. 2002.

B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25, 2009.

J. Z. Levin, M. Yassour, X. Adiconis, C. Nusbaum, D. A. Thompson, N. Friedman, A. Gnirke, and A. Regev. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature methods*, 7(9):709–715, Aug. 2010.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–1760, July 2009.

R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, and J. Wang. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research*, 20 (2):265–272, Feb. 2010.

P. T. Lowary and J. Widom. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *Journal of molecular biology*, 276(1):19–42, Feb. 1998.

K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, 389(6648):251–260, Sept. 1997.

H. Lulin, Y. Xiao, S. Pei, T. Wen, and H. Shangqin. The First Illumina-Based ¡italic¿De Novo¡/italic¿ Transcriptome Sequencing and Analysis of Safflower Flowers. *PLoS ONE*, 7(6):e38653 EP –, Jan. 2012.

M. Magistri, M. A. Faghihi, G. St Laurent, and C. Wahlestedt. Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts. *Trends in genetics : TIG*, 28(8):389–396, Aug. 2012.

W. H. Majoros, M. Pertea, and S. L. Salzberg. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics (Oxford, England)*, 20(16):2878–2879, Nov. 2004.

T. N. Mavrich, I. P. Ioshikhes, B. J. Venters, C. Jiang, L. P. Tomsho, J. Qi, S. C. Schuster, I. Albert, and B. F. Pugh. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome research*, 18 (7):1073–1083, July 2008.

A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5 (7):621–628, May 2008.

S. C. Murray, A. Serra Barros, D. A. Brown, P. Dudek, J. Ayling, and J. Mellor. A pre-initiation complex at the 3'-end of genes drives antisense transcription independent of divergent sense transcription. *Nucleic Acids Research*, 40(6): 2432–2444, Mar. 2012.

U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science (New York, N.Y.)*, 320(5881):1344–1349, June 2008.

H. Neil, C. Malabat, Y. d'Aubenton Carafa, Z. Xu, L. M. Steinmetz, and A. Jacquier. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature*, 457(7232):1038–1042, Feb. 2009.

R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song. Genotype and SNP calling from next-generation sequencing data. *Nature Publishing Group*, 12(6): 443–451, June 2011.

P. A. Pevzner. 1-Tuple DNA sequencing: computer analysis. *Journal of biomolecular structure & dynamics*, 7(1):63–73, Aug. 1989.

O. J. Rando and K. Ahmad. Rules and regulation in the primary structure of chromatin. *Current Opinion in Cell Biology*, 19(3):250–256, June 2007.

N. Rhind, Z. Chen, M. Yassour, D. A. Thompson, B. J. Haas, N. Habib, I. Wapinski, S. Roy, M. F. Lin, D. I. Heiman, S. K. Young, K. Furuya, Y. Guo, A. Pidoux, H. M. Chen, B. Robbertse, J. M. Goldberg, K. Aoki, E. H. Bayne, A. M. Berlin, C. A. Desjardins, E. Dobbs, L. Dukaj, L. Fan, M. G. FitzGerald, C. French, S. Gujja, K. Hansen, D. Keifenheim, J. Z. Levin, R. A. Mosher, C. A. Müller, J. Pfiffner, M. Priest, C. Russ, A. Smialowska, P. Swoboda, S. M. Sykes, M. Vaughn, S. Vengrova, R. Yoder, Q. Zeng, R. Allshire, D. Baulcombe, B. W. Birren, W. Brown, K. Ekwall, M. Kellis, J. Leatherwood, H. Levin, H. Margalit, R. Martienssen, C. A. Nieduszynski, J. W. Spatafora, N. Friedman, J. Z. Dalgaard, P. Baumann, H. Niki, A. Regev, and C. Nusbaum. Comparative functional genomics of the fission yeasts. *Science (New York, N.Y.)*, 332(6032): 930–936, May 2011.

J. L. Rinn, M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, J. A. Helms, P. J. Farnham, E. Segal, and H. Y. Chang. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129(7):1311–1323, June 2007.

R. Robinson. RNAi therapeutics: how likely, how soon? *PLoS Biology*, 2(1):E28, Jan. 2004.

E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, July 2006.

M. Stanke and S. Waack. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics (Oxford, England)*, 19(suppl 2):ii215–ii225, Jan. 2003.

M. A. Tariq, H. J. Kim, O. Jejelowo, and N. Pourmand. Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Research*, Jan. 2011.

C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):516–520, May 2010.

H. van Bakel, J. Stout, A. Cote, C. Tallon, A. Sharpe, T. Hughes, and J. Page. The draft genome and transcriptome of Cannabis sativa. *Genome biology*, 12 (10):R102, 2011.

E. L. van Dijk, C. L. Chen, Y. d'Aubenton Carafa, S. Gourvennec, M. Kwapisz, V. Roche, C. Bertrand, M. Silvain, P. Legoix-Né, S. Loeillet, A. Nicolas, C. Thermes, and A. Morillon. XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature*, 475(7354):114–117, July 2011.

L. Wang, Y. Si, L. K. Dedow, Y. Shao, P. Liu, and T. P. Brutnell. A Low-Cost Library Construction Protocol and Data Analysis Pipeline for Illumina-Based Strand-Specific Multiplex RNA-Seq. *PLoS ONE*, 6(10):e26426 EP –, Oct. 2011.

S. Wang, X. Wang, Q. He, X. Liu, W. Xu, L. Li, J. Gao, and F. Wang. Transcriptome analysis of the roots at early and late seedling stages using Illumina paired-end sequencing and development of EST-SSR markers in radish. *Plant cell reports*, Apr. 2012.

I. Whitehouse, O. J. Rando, J. Delrow, and T. Tsukiyama. Chromatin remodelling at promoters suppresses antisense transcription. *Nature*, 450(7172):1031–1035, Dec. 2007.

Z. Xu, W. Wei, J. Gagneur, F. Perocchi, S. Clauder-Münster, J. Camblong, E. Guffanti, F. Stutz, W. Huber, and L. M. Steinmetz. Bidirectional promoters generate pervasive transcription in yeast. *Nature*, 457(7232):1033–1037, Feb. 2009.

Z. Xu, W. Wei, J. Gagneur, S. Clauder-Münster, M. Smolik, W. Huber, and L. M. Steinmetz. Antisense expression increases gene expression variability and locus interdependency. *Molecular systems biology*, 7:468, Feb. 2011.

M. Yassour, T. Kaplan, H. B. Fraser, J. Z. Levin, J. Pfiffner, X. Adiconis, G. Schroth, S. Luo, I. Khrebtukova, A. Gnirke, C. Nusbaum, D. A. Thompson, N. Friedman, and A. Regev. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(9):3264–3269, Mar. 2009.

D. C. Zappulla and T. R. Cech. RNA as a flexible scaffold for proteins: yeast telomerase and beyond. *Cold Spring Harbor symposia on quantitative biology*, 71:217–224, 2006.

D. R. Zerbino and E. Birney. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5):821–829, Feb. 2008.

G. Zhang, X. Liu, Z. Quan, S. Cheng, X. Xu, S. Pan, M. Xie, P. Zeng, Z. Yue, W. Wang, Y. Tao, C. Bian, C. Han, Q. Xia, X. Peng, R. Cao, X. Yang, D. Zhan, J. Hu, Y. Zhang, H. Li, H. Li, N. Li, J. Wang, C. Wang, R. Wang, T. Guo, Y. Cai, C. Liu, H. Xiang, Q. Shi, P. Huang, Q. Chen, Y. Li, J. Wang, Z. Zhao, and J. Wang. Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. *Nature biotechnology*, 30(6): 549–554, May 2012.

Appendices

Supplementary Information: *Ab initio* construction of a eukaryotic transcriptome by massively parallel mRNA sequencing

Supporting Information

Yassour *et al*. 10.1073/pnas.0812841106

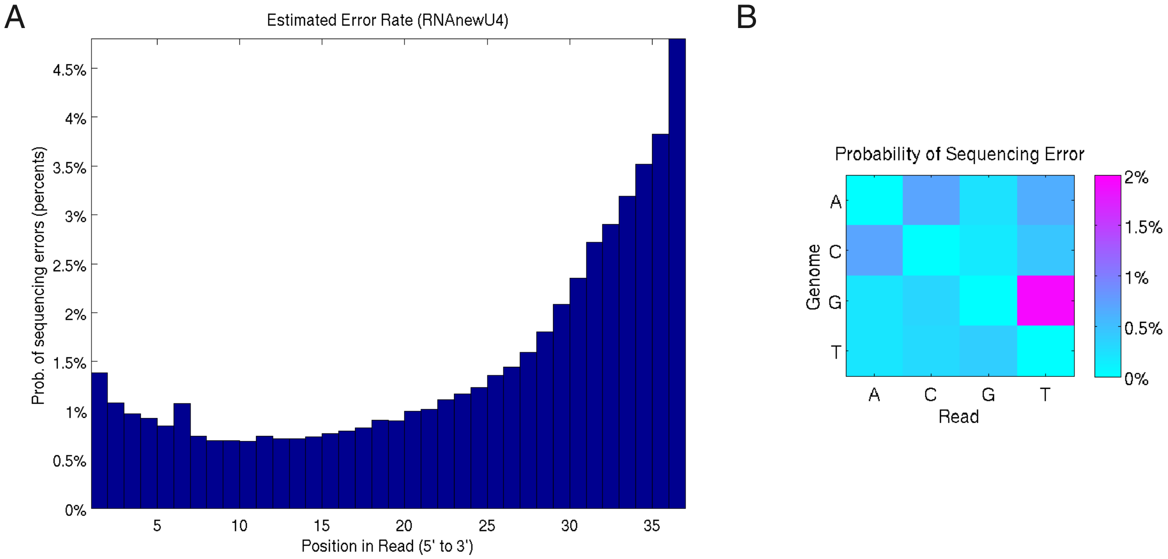


Fig. S1. Error model. (*A*) Estimated error rate for each position in the read. (*B*) The error rate of each specific error, averaged over all positions.

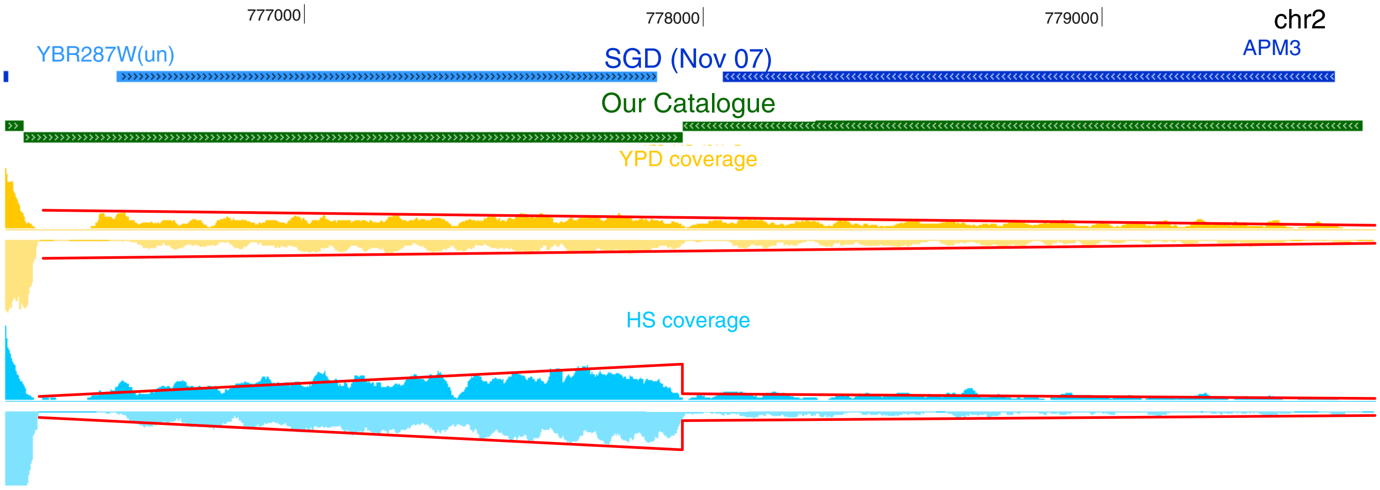


Fig. S2. Segmentation example. A visualization of the segmentation method applied on the locus chr2:776000–780000. In this example, the segmentation is almost impossible based on the YPD data alone, but when considering the HS data, it is very clear.

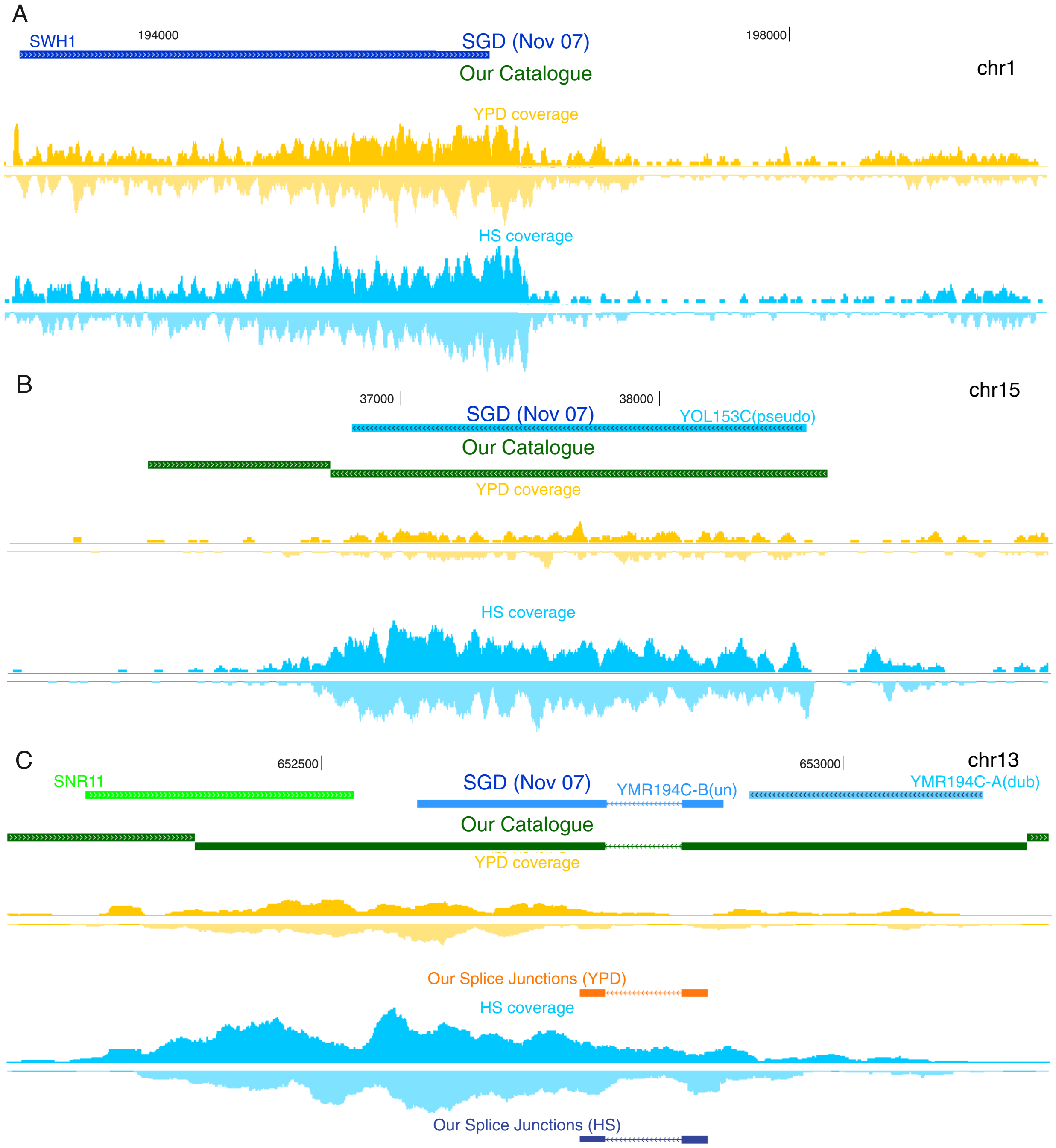
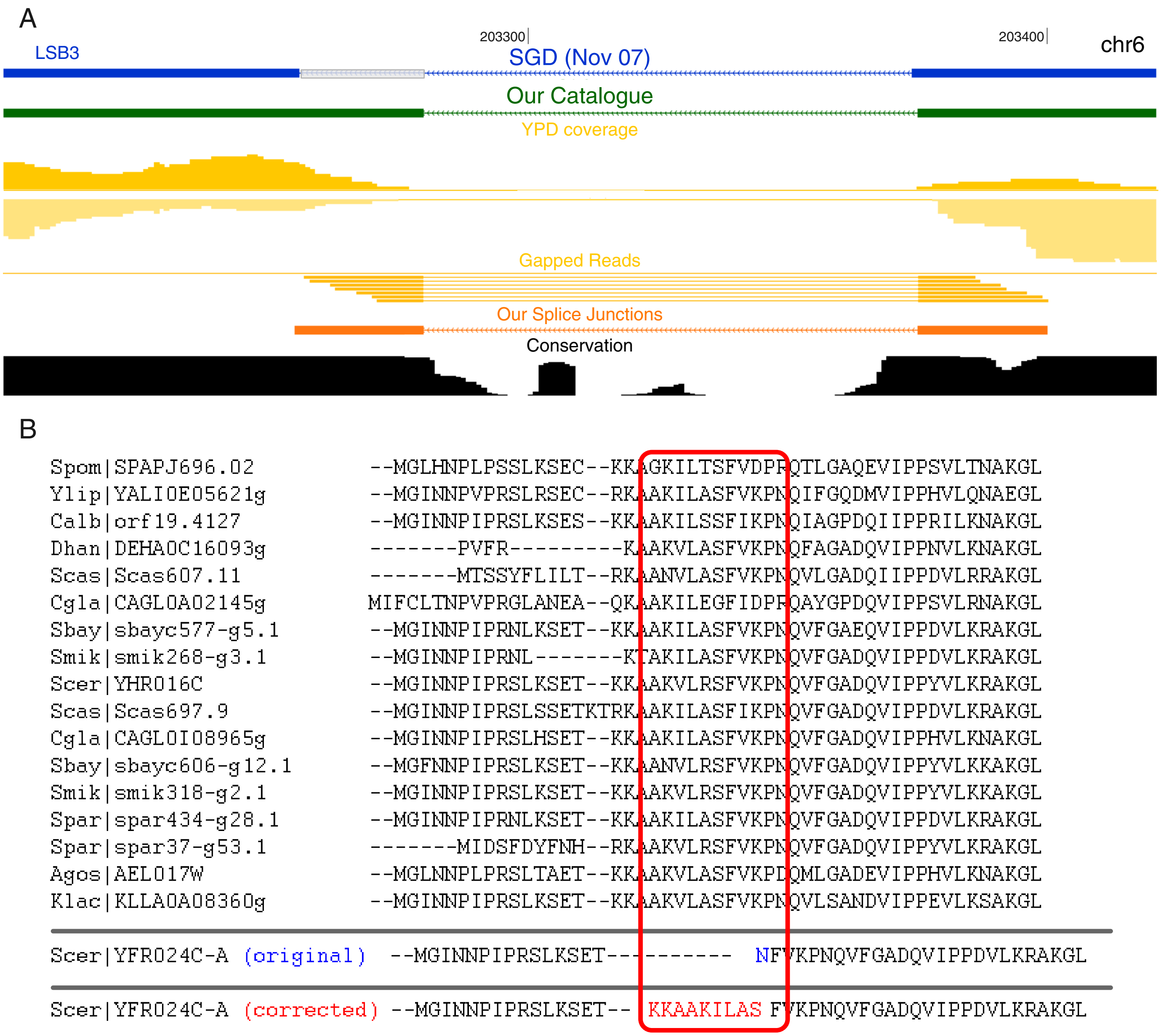


Fig. S3. Transcription validation. (*A*) A new transcribed element at chr1:196277–199970. (*B*) A transcribed pseudogene at chr15:36742–38650. (*C*) A novel transcription unit at the YMR194C locus that spans both a dubious ORF (YMR194C-B) and the gene YMR194C-A.



```
Spom|SPAPJ696.02         --MGLHNPLPSSLKSEC--KKAGKILTSFVDPRQTLGAQEVIPPSVLTNAKGL
Ylip|YALI0E05621g        --MGINNPVPRSLRSEC--RKAAKILASFVKPNQIFGQDMVIPPHVLQNAEGL
Calb|orf19.4127          --MGINNPIPRSLKSES--KKAAKILSSFIKPNQIAGPDQIIPPRILKNAKGL
Dhan|DEHA0C16093g        -------PVFR---------KAAKVLASFVKPNQFAGADQVIPPNVLKNAKGL
Scas|Scas607.11          -------MTSSYFLILT--RKAANVLASFVKPNQVLGADQIIPPDVLRRAKGL
Cgla|CAGL0A02145g        MIFCLTNPVPRGLANEA--QKAAKILEGFIDPRQAYGPDQVIPPSVLRNAKGL
Sbay|sbayc577-g5.1       --MGINNPIPRNLKSET--KKAAKILASFVKPNQVFGAEQVIPPDVLKRAKGL
Smik|smik268-g3.1        --MGINNPIPRNL-------KTAKILASFVKPNQVFGADQVIPPDVLKRAKGL
Scer|YHR016C             --MGINNPIPRSLKSET--KKAAKVLRSFVKPNQVFGADQVIPPYVLKRAKGL
Scas|Scas697.9           --MGINNPIPRSLSSETKTRKAAKILASFIKPNQVFGADQVIPPDVLKRAKGL
Cgla|CAGL0I08965g        --MGINNPIPRSLHSET--KKAAKILASFVKPNQVFGADQVIPPHVLKNAKGL
Sbay|sbayc606-g12.1      --MGFNNPIPRSLKSET--KKAANVLRSFVKPNQVFGADQVIPPYVLKKAKGL
Smik|smik318-g2.1        --MGINNPIPRSLKSET--KKAAKVLRSFVKPNQVFGADQVIPPYVLKKAKGL
Spar|spar434-g28.1       --MGINNPIPRNLKSET--KKAAKILASFVKPNQVFGADQVIPPDVLKRAKGL
Spar|spar37-g53.1        -------MIDSFDYFNH--RKAAKVLRSFVKPNQVFGADQVIPPYVLKRAKGL
Agos|AEL017W             --MGLNNPLPRSLTAET--KKAAKVLASFVKPDQMLGADEVIPPHVLKNAKGL
Klac|KLLA0A08360g        --MGINNPIPRSLKSET--KKAAKVLASFVKPNQVLSANDVIPPEVLKSAKGL

Scer|YFR024C-A (original)  --MGINNPIPRSLKSET----------  NFVKPNQVFGADQVIPPDVLKRAKGL

Scer|YFR024C-A (corrected) --MGINNPIPRSLKSET-- KKAAKILAS FVKPNQVFGADQVIPPDVLKRAKGL
```

Fig. S4. Splicing correction example. (*A*) In the gene LSB3, we find an intron that is shorter than reported by SGD [Cherry JM, *et al.* (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res* 26:73–79]. The gray box represents the addition to the exon, according to our results. (*B*) The multiple sequence alignment of this region with the original and corrected annotation of the gene LSB3 [Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449:54–61]. It is clear that the added segment is highly conserved in other yeast species

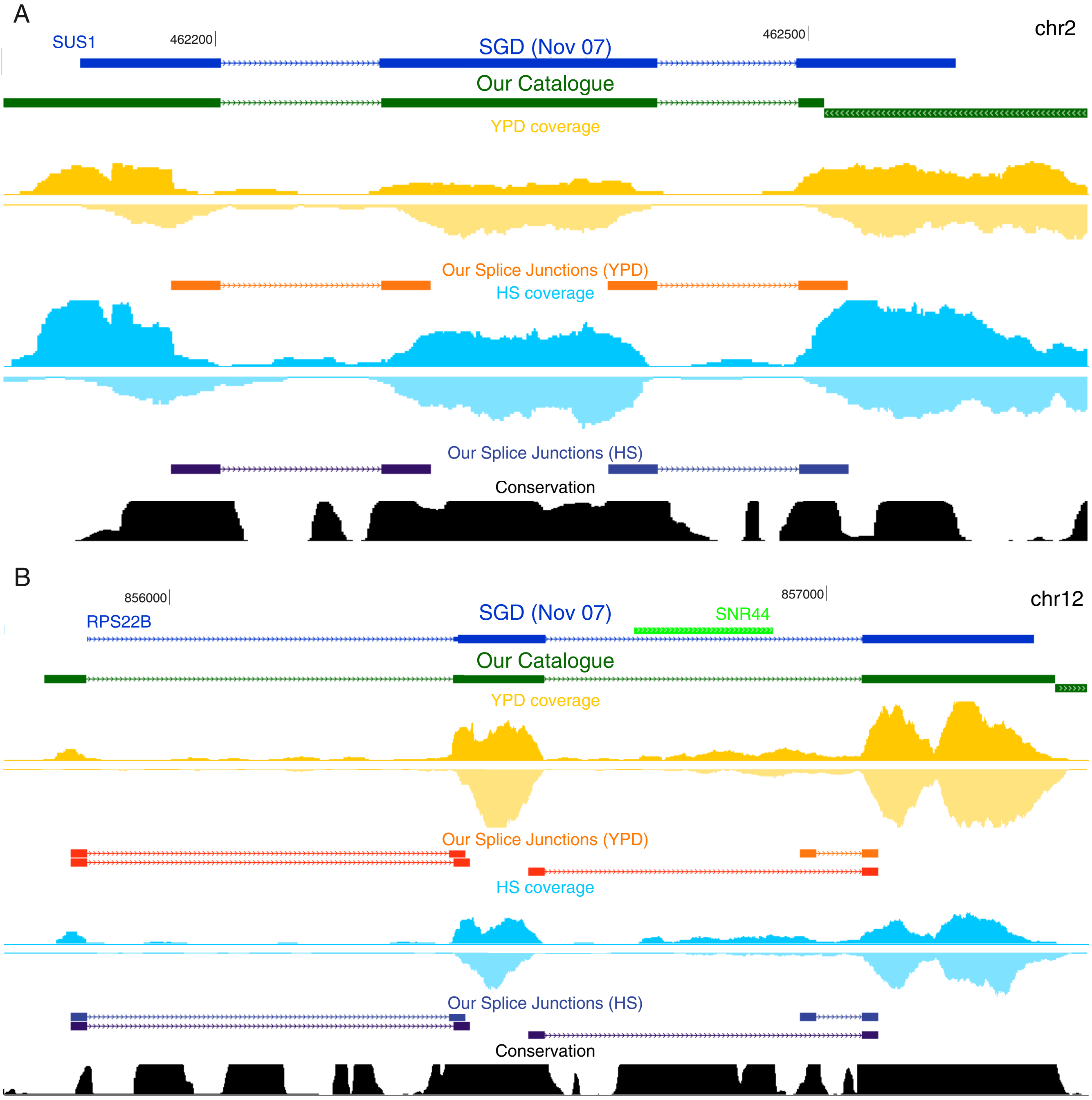


Fig. S5. Splicing validation. (*A*) Alternative splicing in the SUS1 gene, where, in addition to the 2 known introns, we also observe clear read-through at both junctions. Experimental validation confirms our predictions by revealing 3 bands, 2 bands consistent with just 1 intron spliced, and a stronger band consistent with both introns spliced out. (*B*) A previously uncharacterized intron from the end of the snoRNA, SNR44, to the acceptor site of its hosting intron, inside RPS22B.

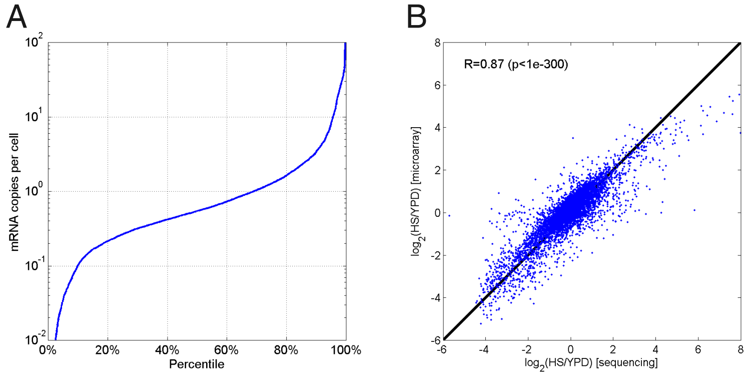


Fig. S6. Quantifying expression using sequencing. (*A*) Distribution of estimated mRNA copies per cell in YPD. Quantitative mRNA expression levels were estimated based on the density of reads along ORFs, with an estimate of 15,000 mRNA molecules per cell. (*B*) For each ORF, we computed the log2 ratio of HS and YPD (*x* axis), and compare it to its log2 ratio as measured by commercial 2-dye DNA microarrays (*y* axis).

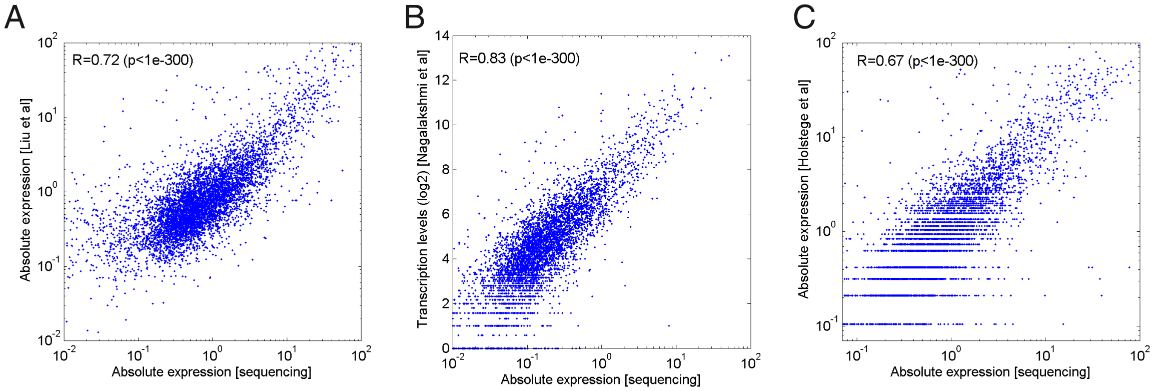


Fig. S7. Absolute expression comparison to previous studies [Nagalakshmi U, *et al.* (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320:1344–1349; Holstege FC, *et al.* (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95:717–728; Liu CL, *et al.* (2005) Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLoS Biol* 3:e328.

Other Supporting Information Files

Dataset S1 (XLS)
Dataset S2 (XLS)
Dataset S3 (XLS)

Supplementary Information: Comprehensive comparative analysis of strand-specific RNA sequencing methods

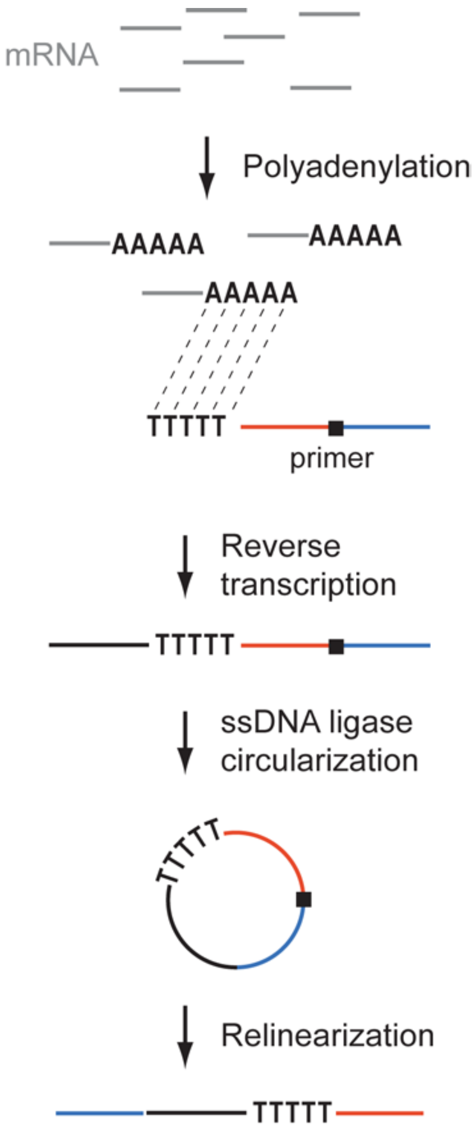
nature | methods

Comprehensive comparative analysis of strand-specific RNA sequencing methods

Joshua Z Levin, Moran Yassour, Xian Adiconis, Chad Nusbaum, Dawn Anne Thompson, Nir Friedman, Andreas Gnirke & Aviv Regev

Supplementary figures and text:

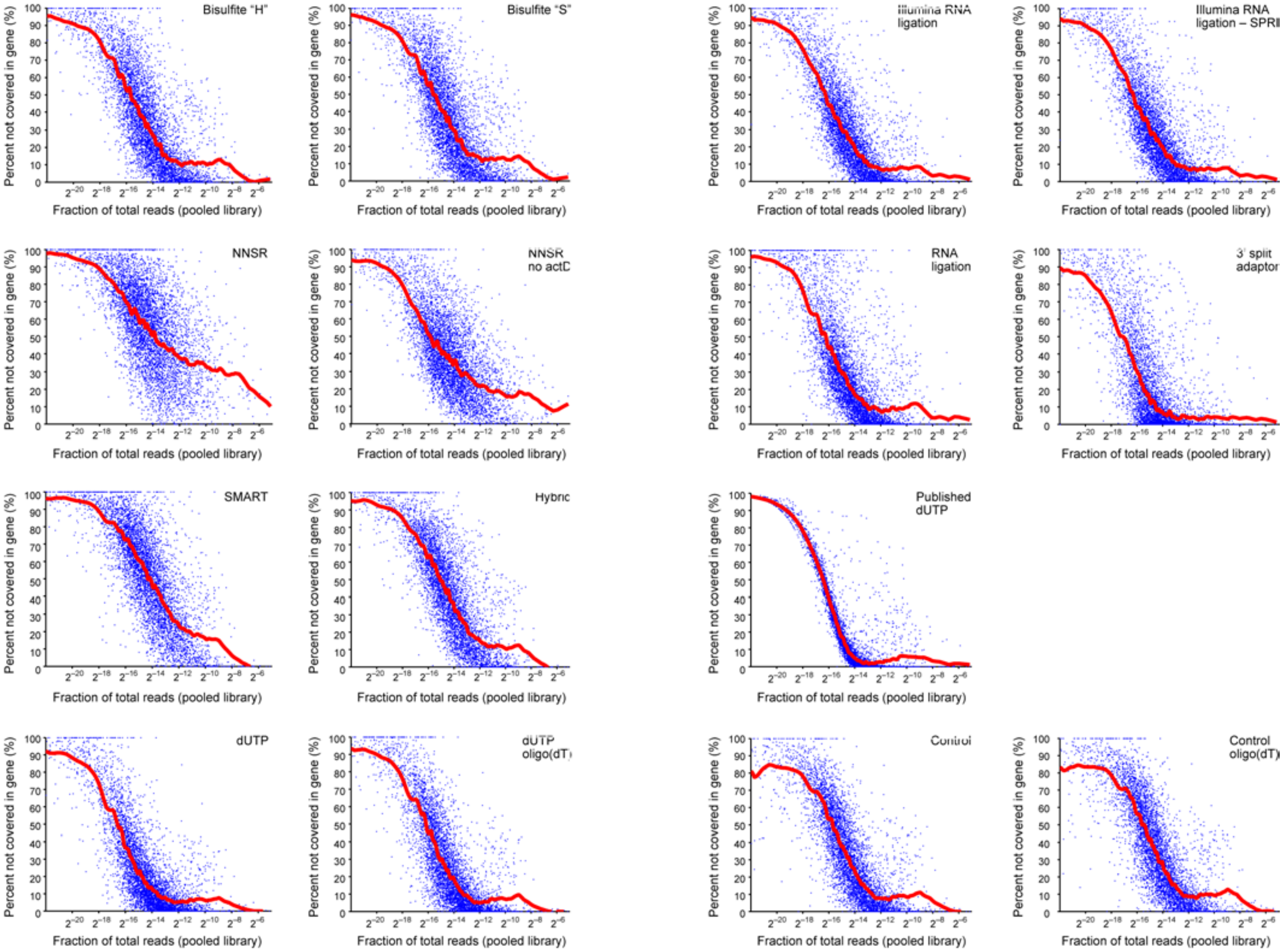
Supplementary Figure 1



Supplementary Figure 1. The 3’ split adaptor method.
Shown are the salient details for the 3’ split adaptor method[14].

Nature Methods: doi: 10.1038/nmeth.1491

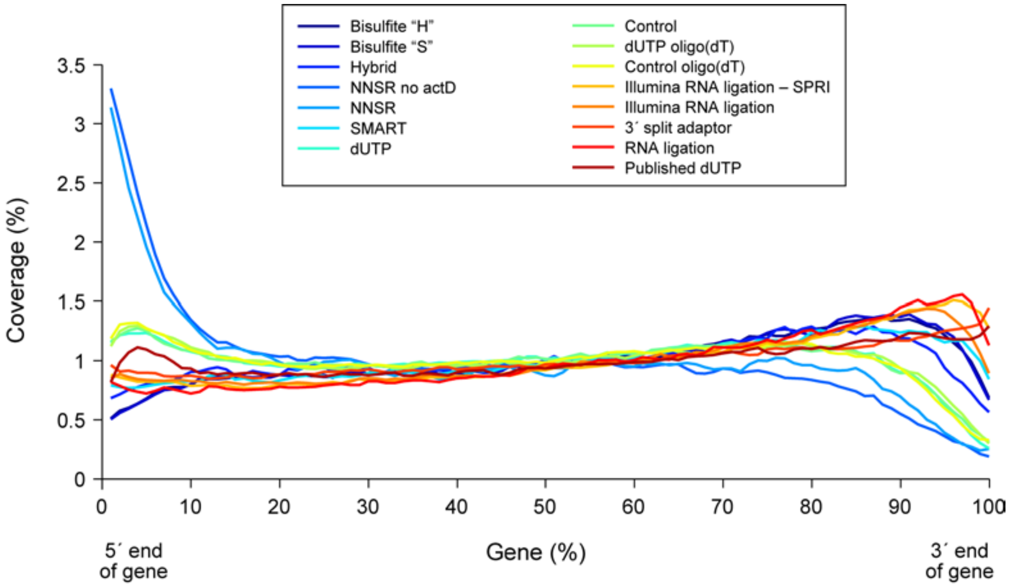
Supplementary Figure 2



Supplementary Figure 2. Fraction of transcript coverage.

Shown is the percentage of bases with zero coverage (Y axis) for each gene (blue dot) in the genome, vs. the fraction of total reads for that gene in the pooled library. Plots are shown for each library in the compendium, as noted. In each case, a Lowess fit is shown as a red curve.

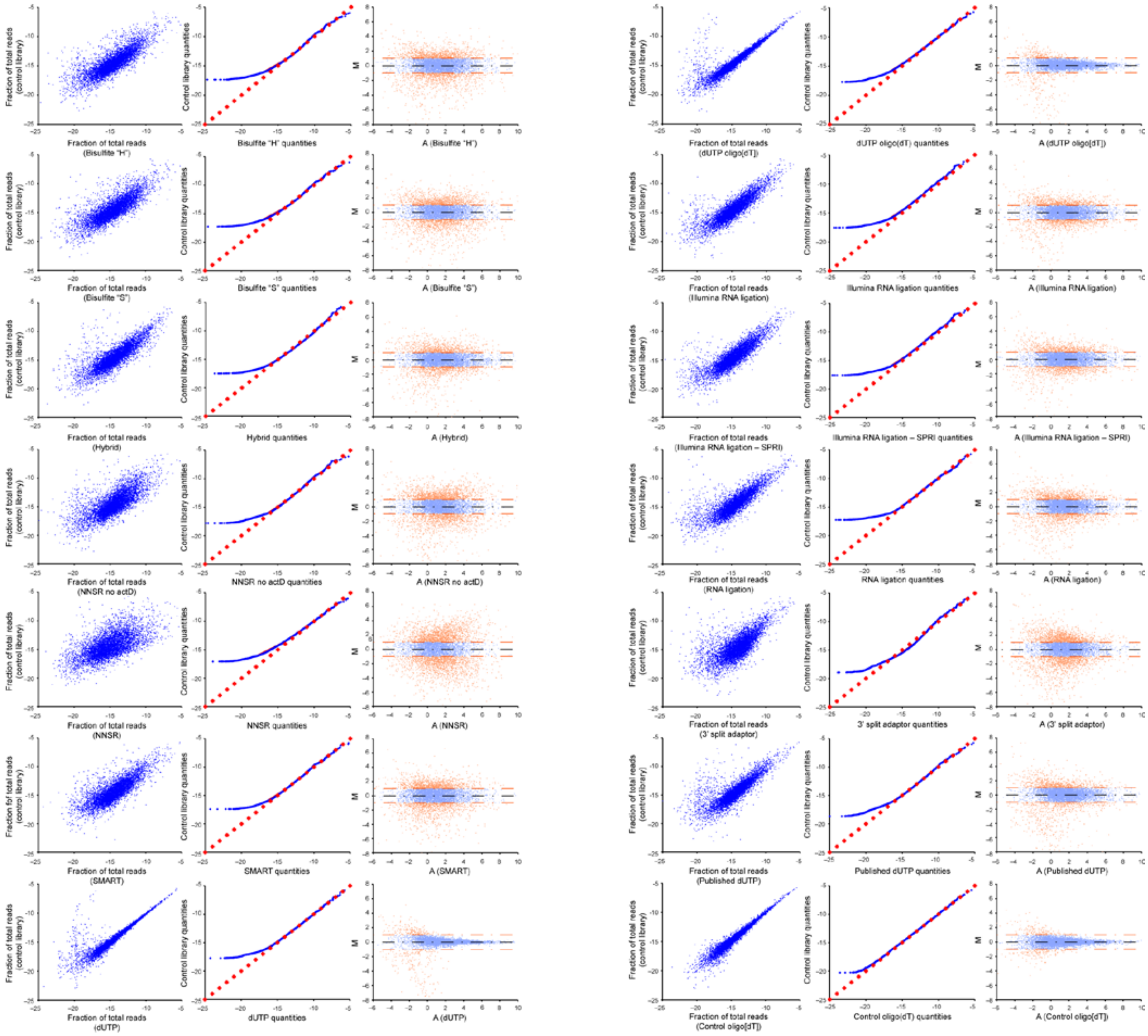
Supplementary Figure 3



Supplementary Figure 3. Average gene coverage.

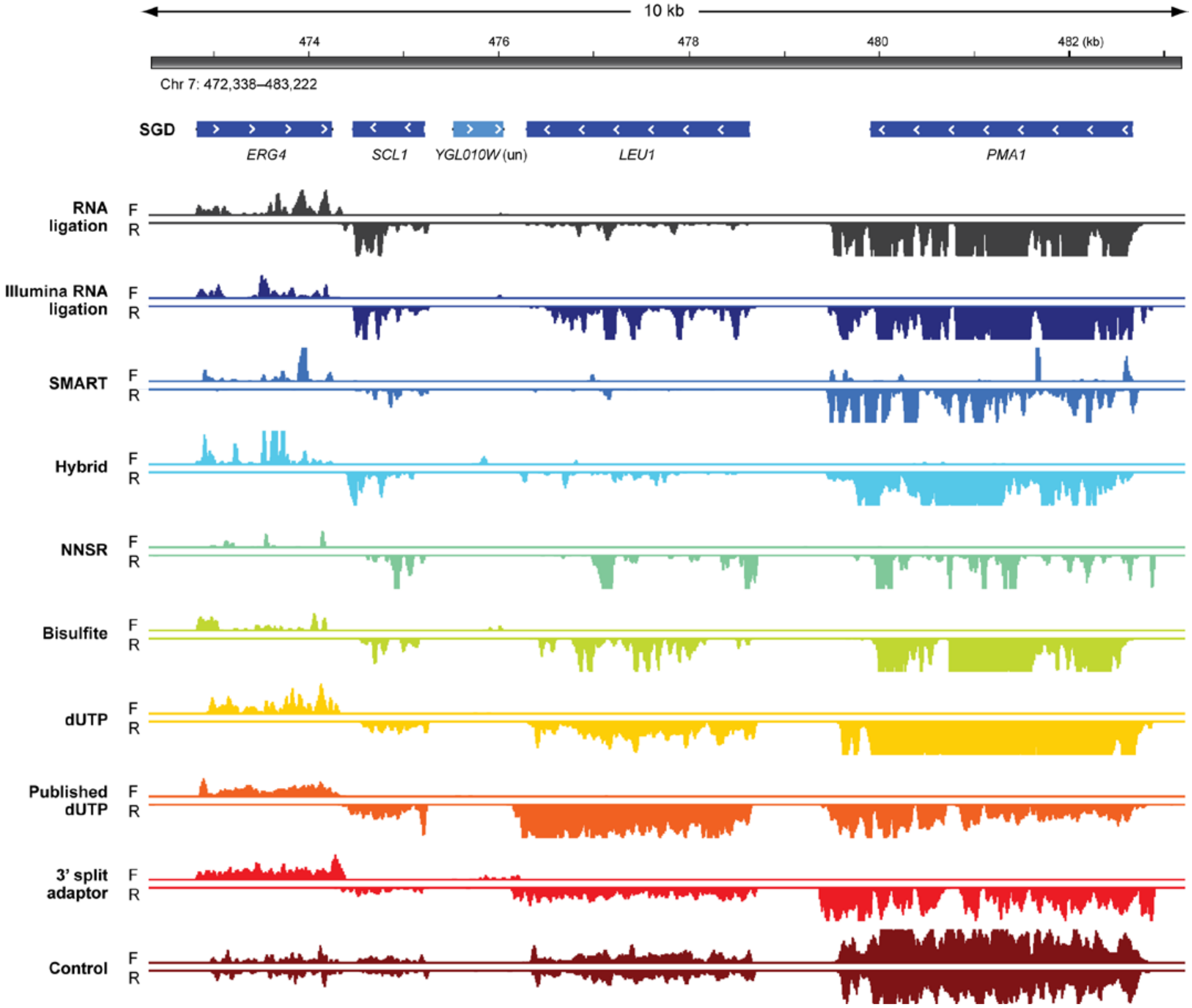
Shown is the average gene coverage at each percentile of a gene’s length, for all genes in each library. Libraries are color coded as specified in the legend.

Supplementary Figure 4



Supplementary Figure 4. Scatter, Q-Q, and MA plots. Shown are the scatter (left panel), Q-Q (middle panel) and MA (right panel) plots for each library, in comparison to the control library. The scatter plot shows the fraction of total reads for each gene (blue dot) in the reference library (Y axis) *vs.* a strand specific library (X axis). The Q-Q plot shows the level at each quantile (rank) of expression in the reference library (Y axis) *vs.* the strand-specific library (X axis). A slope = 1 line is shown for comparison (red crosses). The MA plot shows for each gene (dot) the difference in expression levels between the reference and strand-specific libraries (Y axis) *vs*. their mean expression level (X axis). Red dashed lines — two-fold difference in expression.

Supplementary Figure 5



Supplementary Figure 5. Coverage at example genomic locus.

Shown are the genome annotations from SGD (top track, boxes with arrow heads), followed by the aligned read coverage in each library on each strand (maximum scale is 100 reads), for the Chromosome 7: 472,338-483,222 locus. Coverage is calculated only with reads from the 2.5 million sampled reads per library.

Supplementary Table 1: Alignment of all reads for each library in our compendium.

Single-end Libraries

Library	# Lanes	Total # reads	# Mapped reads	Mapped / lane	% Mapped	Mapped uniquely	Mapped uniquely / lane	% Mapped uniquely
RNA Ligation	1	24,504,932	19,188,938	19,188,938	78	15,249,242	15,249,242	62
Illumina RNA Ligation	2	48,120,669	33,843,481	16,921,741	70	28,519,438	14,259,719	59
Illumina RNA Ligation - SPRI	2	51,475,621	21,444,010	10,722,005	42	18,074,114	9,037,057	35
3' Split Adaptor	1	9,612,690	9,231,502	9,231,502	96	3,695,252	3,695,252	38
Published dUTP	1	12,216,063	7,652,683	7,652,683	63	5,140,634	5,140,634	4

Paired-end Libraries

Library	# Lanes	Total # reads	Read 1 mapped non-uniquely	% Read 1 mapped non-uniquely	Read 2 mapped non-uniquely	% Read 2 mapped non-uniquely	Paired matches	Paired / lane	% Paired-end mapped reads	% Unique reads
SMART	2	5,076,555	2,868,582	57	2,543,430	50	930,686	465,343	18	81
Hybrid	2	14,788,936	5,752,937	39	5,664,015	38	2,900,346	1,450,173	20	81
NNSR	1	6,873,972	4,636,153	67	3,628,894	53	2,683,010	2,683,010	39	81
NNSR no actD	2	16,399,019	8,328,130	51	7,975,082	49	5,291,376	2,645,688	32	82
BiSulfite "S"	1	10,168,083	7,235,219	71	7,564,178	74	4,570,831	4,570,831	45	63
BiSulfite "H"	1	6,896,242	3,708,647	54	3,992,780	58	2,022,728	2,022,728	29	78
dUTP	1	13,614,820	11,895,357	87	11,689,118	86	9,222,678	9,222,678	68	58
dUTP oligo(dT)	1	9,899,691	8,512,926	86	8,590,913	87	6,580,247	6,580,247	67	48
Control	1	14,596,122	12,565,360	86	12,654,534	87	9,872,609	9,872,609	68	54
Control oligo(dT)	1	13,843,046	11,712,442	85	11,857,471	86	9,059,171	9,059,171	65	53

Supplementary Table 2: Basic statistics and comprehensive performance measures for each library in our compendium.

Library	total number of reads (sampled all regions)	unique read starts	% unique read starts	% unique pairs	number of reads on expected strand	number of reads on opposite strand	number of reads outside known annotations	total number of reads in single feature regions	% antisense (opposite strand)	average coefficient of variation (CV) for top 50% expressed genes)	genes with 5' end covered	genes with 3' end covered	weighted average of number of segments per gene	correlation to control	RMSE to control	correlation to pooled	RMSE to pooled	correlation to microarrays	RMSE to microarrays
RNA Ligation	2,500,018	922,327	37%		2,290,117	10,947	13,668	2,314,732	0.47%	1.06	59%	54%	3.16	0.80	0.99	0.90	0.75	0.83	0.96
Illumina RNA Ligation	2,500,018	962,917	39%		2,293,081	13,837	21,374	2,328,292	0.59%	1.17	60%	45%	2.61	0.86	0.87	0.95	0.64	0.80	1.05
Illumina RNA Ligation - SPRI	2,500,016	979,204	39%		2,287,502	15,207	24,742	2,327,451	0.65%	1.16	62%	49%	2.69	0.85	0.87	0.95	0.62	0.81	1.03
SMART	930,686	380,169	41%		756,529	96,080	4,750	857,359	11.20%	1.50	41%	41%	4.59	0.79	0.96	0.82	0.95	0.73	1.21
Hybrid	2,500,017	442,037	18%	44%	2,289,299	43,642	23,849	2,356,790	1.85%	1.61	59%	54%	4.39	0.81	0.93	0.89	0.73	0.70	1.27
NNSR	2,500,020	356,534	14%	51%	2,270,422	12,268	62,007	2,344,697	0.52%	2.11	44%	49%	4.40	0.62	1.51	0.78	1.29	0.57	1.63
NNSR no actD	2,500,019	591,443	24%	64%	2,273,087	51,888	62,623	2,387,598	2.17%	1.75	58%	62%	4.43	0.72	1.15	0.87	0.82	0.73	1.23
BiSulfite "S"	2,500,017	704,275	28%	78%	2,263,206	25,518	11,504	2,300,228	1.11%	1.28	51%	51%	3.52	0.79	1.03	0.90	0.77	0.73	1.21
BiSulfite "H"	2,019,595	738,479	37%	76%	1,828,045	23,254	10,544	1,861,843	1.25%	1.25	51%	50%	3.29	0.81	0.99	0.90	0.75	0.73	1.20
dUTP	2,500,019	895,698	36%	84%	2,319,635	14,609	14,320	2,348,564	0.62%	0.76	62%	73%	2.48	0.90	0.69	0.94	0.57	0.84	0.94
dUTP oligo(dT)	2,500,018	794,635	32%	81%	2,303,415	15,554	13,632	2,332,601	0.67%	0.86	58%	72%	2.54	0.89	0.78	0.92	0.74	0.82	1.03
3' Split Adaptor	2,500,016	1,042,152	42%		2,139,848	63,904	91,288	2,295,040	2.78%	0.54	75%	77%	2.29	0.68	1.21	0.88	0.89	0.80	1.19
Published dUTP	2,500,019	1,000,797	40%		2,192,118	37,033	57,559	2,286,710	1.62%	0.64	62%	61%	2.41	0.80	0.98	0.93	0.64	0.81	1.08
Control	2,500,017	1,057,315	42%	88%	1,148,156	1,167,471	17,423	2,333,050	50.04%	0.85	54%	64%	3.18	1.00	0.00	0.75	1.22	0.67	1.46
Control oligo(dT)	2,500,016	996,806	40%	87%	1,157,204	1,150,325	17,463	2,324,992	49.48%	0.90	51%	63%	3.20	0.97	0.36	0.74	1.23	0.65	1.48

Supplementary Table 3: Summary of advantages and disadvantages of library construction methods.

Library	Advantages	Disadvantages
RNA Ligation	High complexity; High strand specificity	Lengthy method with multiple size selection steps requiring large amounts of RNA; Uneven coverage; Single end sequencing[a]
Illumina RNA Ligation	Overall high quality	Single end sequencing[a]; Uneven coverage; Low coverage of 3' ends
Illumina RNA Ligation - SPRI	Overall high quality	Shorter cDNAs not removed from library; Single end sequencing[a]; Uneven coverage; Low coverage of 3' ends
SMART		Inefficient process -- few reads; Overall low quality
Hybrid	Better than SMART	Overall low quality
NNSR	High strand specificity; Simple library construction	Overall low quality
NNSR no actD	Simple library construction	Overall low quality

BiSulfite "S"	Similar to standard library construction	Sequence alignment issues;
		Low strand specificity; Uneven coverage;
		Low coverage of 5' ends

BiSulfite "H"	Similar to standard library construction	Sequence alignment issues;
		Low strand specificity; Uneven coverage;
		Low coverage of 5' ends

dUTP	Overall high quality; Similar to standard library construction	
dUTP oligo(dT)	Overall high quality; Similar to standard library construction	

a: Not an intrinsic limitation of the protocol; with appropriate modification of the protocol, paired-end sequencing can presumably be performed

Supplementary Table 4: Comparison of technical details of library construction methods.

Library	Time Required (days)	Total number of steps	Approx. Reagent cost ($)	Starting material used (RNA, ng)	Applicability to small RNA	Kits available?
RNA Ligation	8	19	250	1200	Yes	No
Illumina RNA Ligation	5	16	240	100[b]	Yes	Partially (Small RNA Library Construction v1.5)
Illumina RNA Ligation - SPRI	4	12	220	100[b]	Yes	Partially (Small RNA Library Construction v1.5)
SMART	5	8	80	100	Yes	No
Hybrid	5	13	90	500	Yes	No
NNSR	4	9	90	250	Unclear	No
NNSR no actD	4	9	90	250	Unclear	No
BiSulfite "S"	6	19	540[a]	1000[c]	No	Mostly (Bisulfite & Standard Library Construction)
BiSulfite "H"	6	19	540[a]	1000[d]	No	Mostly (Bisulfite & Standard Library Construction)
dUTP	5	17	430[a]	200	No	Mostly (Standard Library Construction)
dUTP oligo(dT)	5	17	440[a]	200	No	Mostly (Standard Library Construction)
Control	5	15	430[a]	200	No	Standard Library Construction
Control oligo(dT)	5	15	430[a]	200	No	Standard Library Construction

a: Cost is lower if individual reagents are used instead of Illumina standard library construction kit

b: Starting material was 100 ng -- cDNA was split later for the two variants of this method (see Methods for details).

c: Starting material was 1000 ng -- only 96 ng of 212 ng was used for reverse transcription

d: Starting material was 1000 ng -- only 40 ng of 152 ng was used for reverse transcription

Supplementary Table 5: Primer sequences.

Primer Name	Primer Sequence
SMART tagged random primer	5’-CATTGAGCTGAACCGAGTCCAGCAGNNNNNN
5’ SMART oligo	5’-TTTCCCTACACGACGCTCTTCCGATCTrGrGrG
SMART reverse primer	5’- CAAGCAGAAGACGGCATACGACGATCTCGACATTGAGCTGAACCGAGTCCAGCAG
3’ RNA adaptor oligo	5’- AGAUCGGAAGAGCGGUUCAGCAGInvdT
Hybrid reverse transcription primer	5’- GGCATTCCTGCTGAACCGCTCTTCCGATCT
5’ Hybrid oligo	5’- CTCTTTCCCTACACGACGCTCTTCCGATCTrGrGrG
Hybrid forward	5’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCT
Hybrid reverse	5’- CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGC
1st strand NNSR primers	5’-TCCGATCTCTNNNNNNNN
2nd strand NNSR primers	5’-TCCGATCTGANNNNNNNN
NNSR forward	5’- AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCT
NNSR reverse	5’- CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTGA
SBS11	5’-CGATCTCGACATTGAGCTGAACCGAGTCCAGCAG

Supplementary Note 1

Comparison of dUTP and Illumina RNA ligation methods

Overall, the dUTP and Illumina RNA ligation protocols performed best across the broadest range of evaluation criteria, including strand specificity, measures critical for genome annotation (evenness and continuity of coverage), and measures critical for expression profiling. The dUTP approach performed significantly better in fraction of mapped reads and evenness of coverage (important for genome annotation), and slightly better at expression profiling (especially based on RMSE measures) and 3' end coverage. The Illumina RNA ligation methods performed somewhat better for strand specificity and single-end complexity, but paired-end dUTP reads had excellent complexity. The two methods were comparable for continuity and 5' end detection, involve comparably simple protocols (with dUTP being slightly simpler), and require no specialized computational processing.

Supplementary Note 2

Monotemplate sequencing issue

Because each of the NNSR, SMART, and Hybrid libraries has a short, identical sequence at the start of every read and must be sequenced at a lower cluster density, these libraries generate less usable sequence per lane than standard libraries with current Illumina sequencing protocols.

For the NNSR libraries, we used a lower cluster density to resolve issues resulting from the first two bases being identical in each read, creating a "monotemplate." This monotemplate issue is a problem for the Illumina Genome Analyzer software (v.1.5) because it uses the first two cycles to determine where clusters reside in an image (template generation) and this results in some images being "denser" than they would be given a random base distribution, i.e. 100% of the clusters lighting up in the "A" image compared to 25% lighting up in the "A" image. As a result of this higher image density, the software is unable to find cluster locations using cross-correlation of the pixel intensities. Lowering the cluster density alleviates this problem, but results in less sequence being generated than for a library without monotemplate issues loaded at standard cluster density. This was also a potential problem that may have reduced the fraction of Passing Filter bases for the SMART and Hybrid libraries, but without any special handling their cluster densities turned out to be somewhat lower relative to other contemporary sequencing runs.

Supplementary Note 3

Microarray data

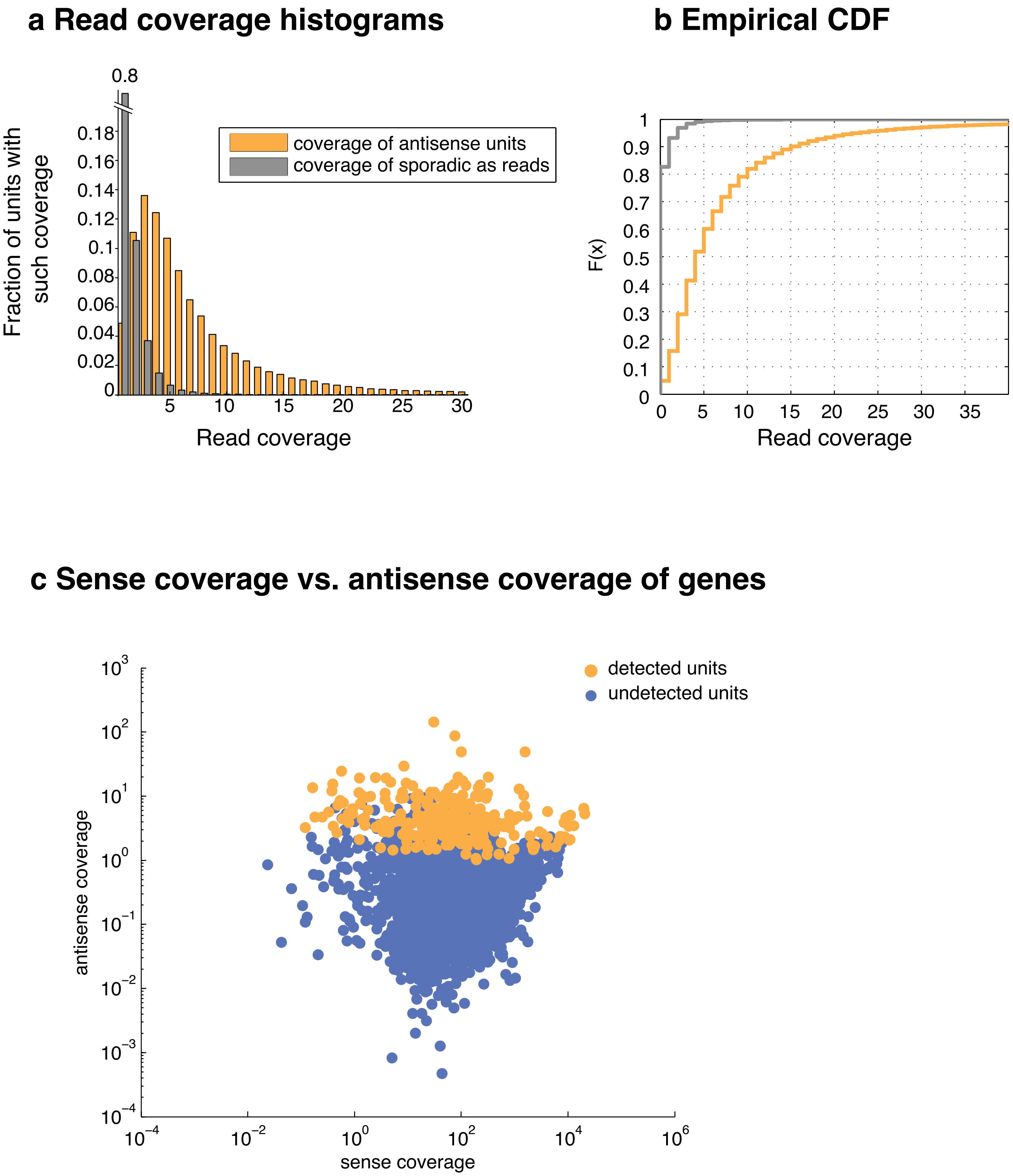
Saccharomyces cerevisiae strain BY4741 was grown to mid-log and cells were harvested by freezing in liquid nitrogen. Total RNA was isolated using the RNeasy Midi or mini Kits (Qiagen) according to the provided instructions for mechanical lysis. Samples were quality controlled with the RNA 6000 Nano (series II) kit for the Bioanalyzer 2100 (Agilent). Genomic DNA from *Saccharomyces cerevisiae* strain BY4741 was isolated using Genomic-tip 500/G (Qiagen) using the provided protocol for yeast. DNA samples were sheared using Covaris sonicator to 500-1000 bp fragments, as verified using DNA 7500 and DNA 12000 kit for the Bioanalyzer 2100 (Agilent). Independently sheared samples labeled with Cy3 and Cy5 were highly correlated (R> .97 in each of four independent hybridizations), indicating that the shearing procedure is reproducible and unbiased. Total RNA samples were labeled with Cy3 (cyanine fluorescent dyes) and genomic DNA samples were labeled with Cy5 using a modification of the protocol developed by Joe DeRisi (UCSF) and Rosetta Inpharmatics (Kirkland, WA) that can be obtained at www.microarrays.org and as described[33]. Two biological replicates of Cy3 labeled RNA samples were mixed with a reference Cy5 labeled genomic DNA sample and hybridized on a two-color Agilent 4x44K *S. cerevisiae* array (commercial Agilent array; four to five probes per target gene). After hybridization and washing per Agilent instructions, arrays were scanned using an Agilent scanner and analyzed with Agilent's feature extraction software version 10.5.1.1. For each probe, the median signal intensities were background subtracted for both channels and combined by taking the log2 of their ratio. To estimate the absolute expression values for each gene, we took the median of the log2 ratios across all probes. The experiments were highly reproducible; most biological replicates correlated at R = .99 and replicates with R < .95 were removed. Different biological replicates were combined using Quantile normalization to estimate the absolute expression level per gene.

33. Wapinski, I. et al. Gene duplication and the evolution of ribosomal protein gene regulation in yeast. *Proc. Natl. Acad. Sci. USA* 107, 5505–5510 (2010).

Supplementary Figures[*]: Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species

[*] Figure captions appear in **Chapter 4**

Supplementary Figure 1 - Antisense reads coverage: units vs. sporadic



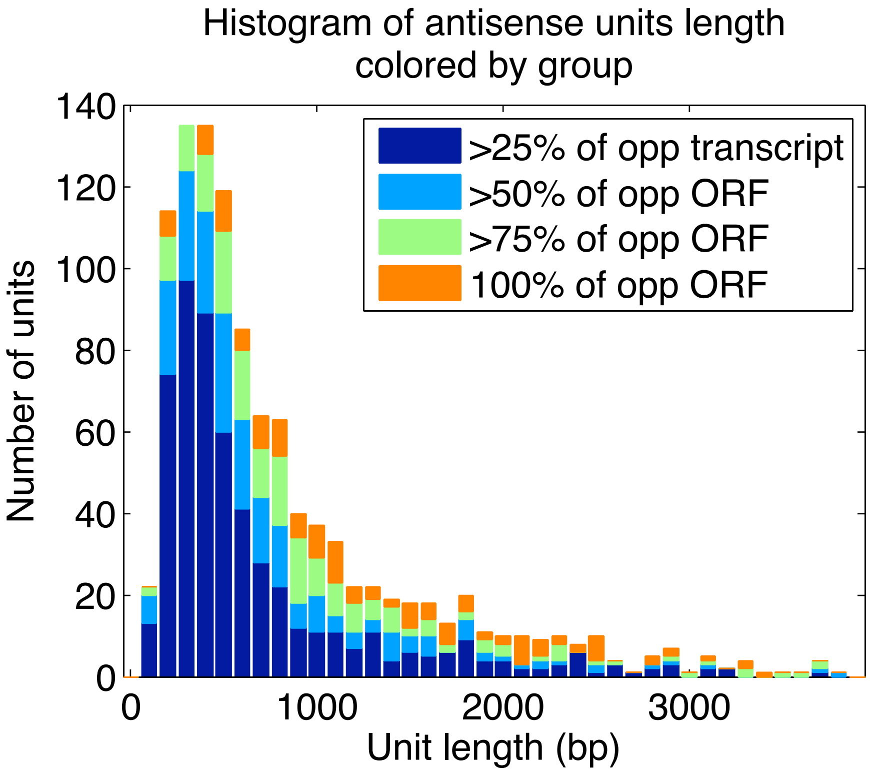
a Read coverage histograms

b Empirical CDF

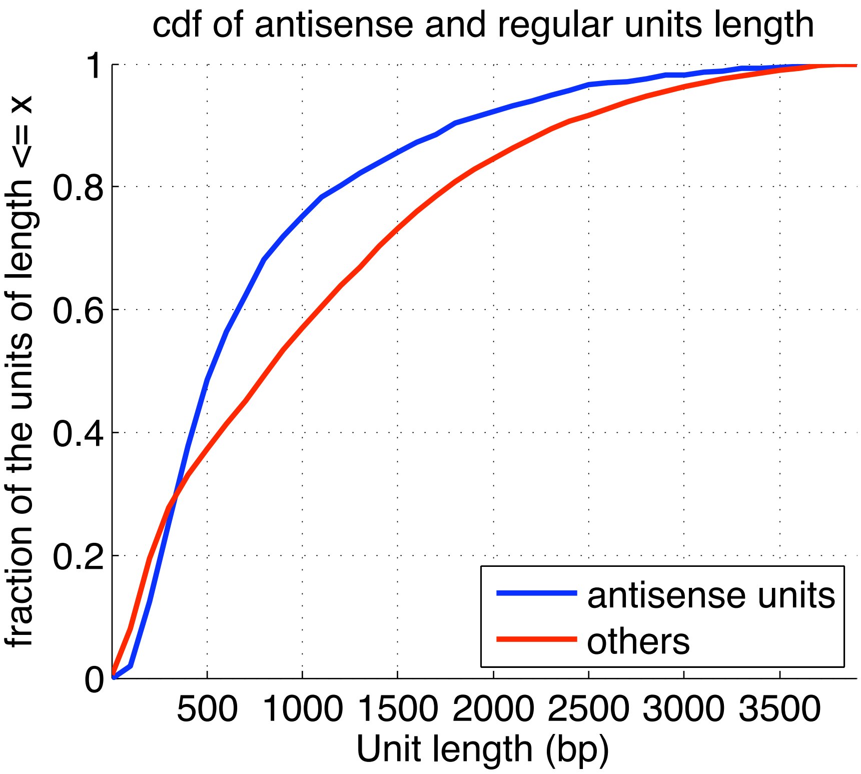
c Sense coverage vs. antisense coverage of genes

Supplementary Figure 2 - Units statistics

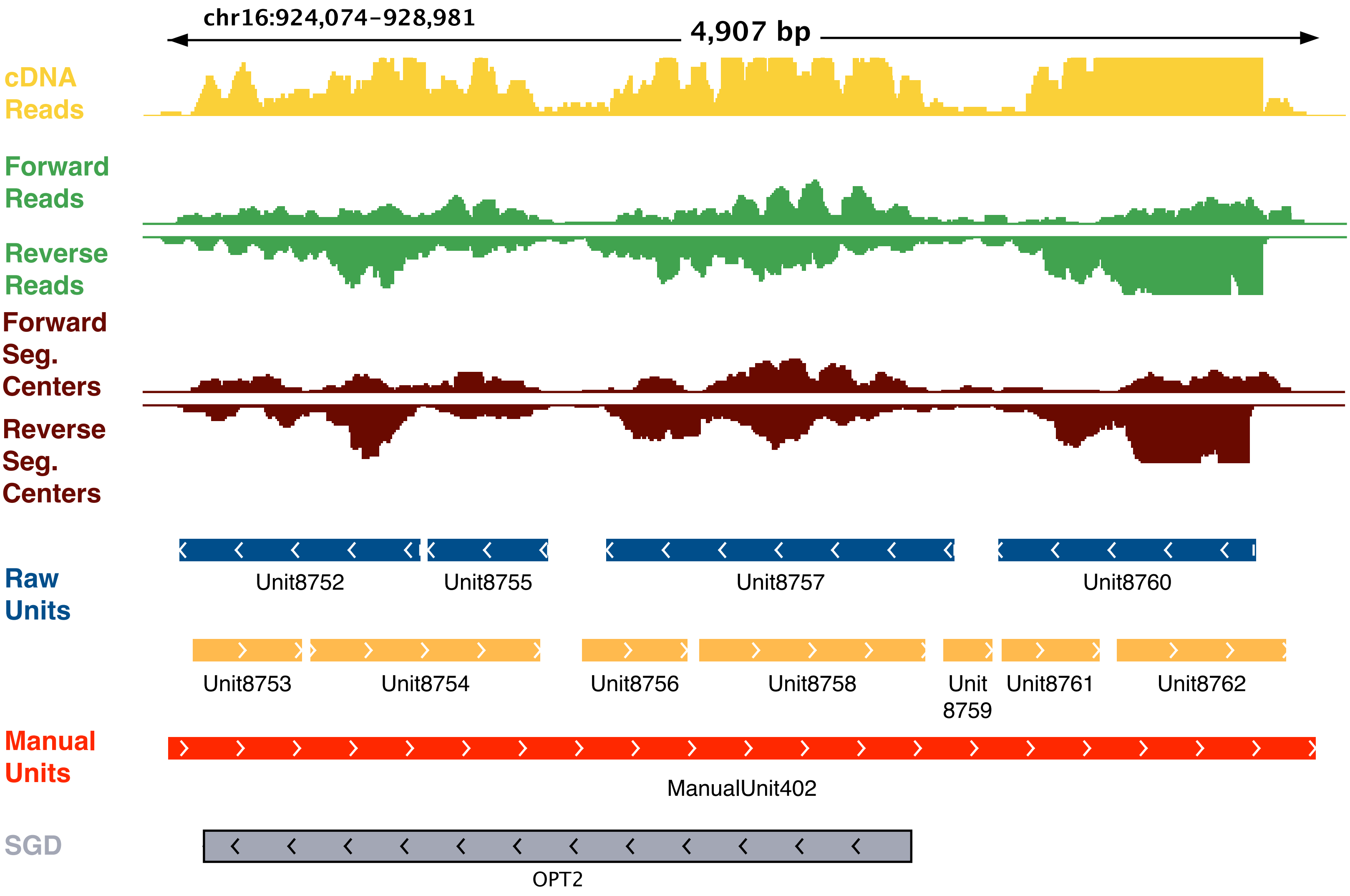
a Antisense unit length histogram



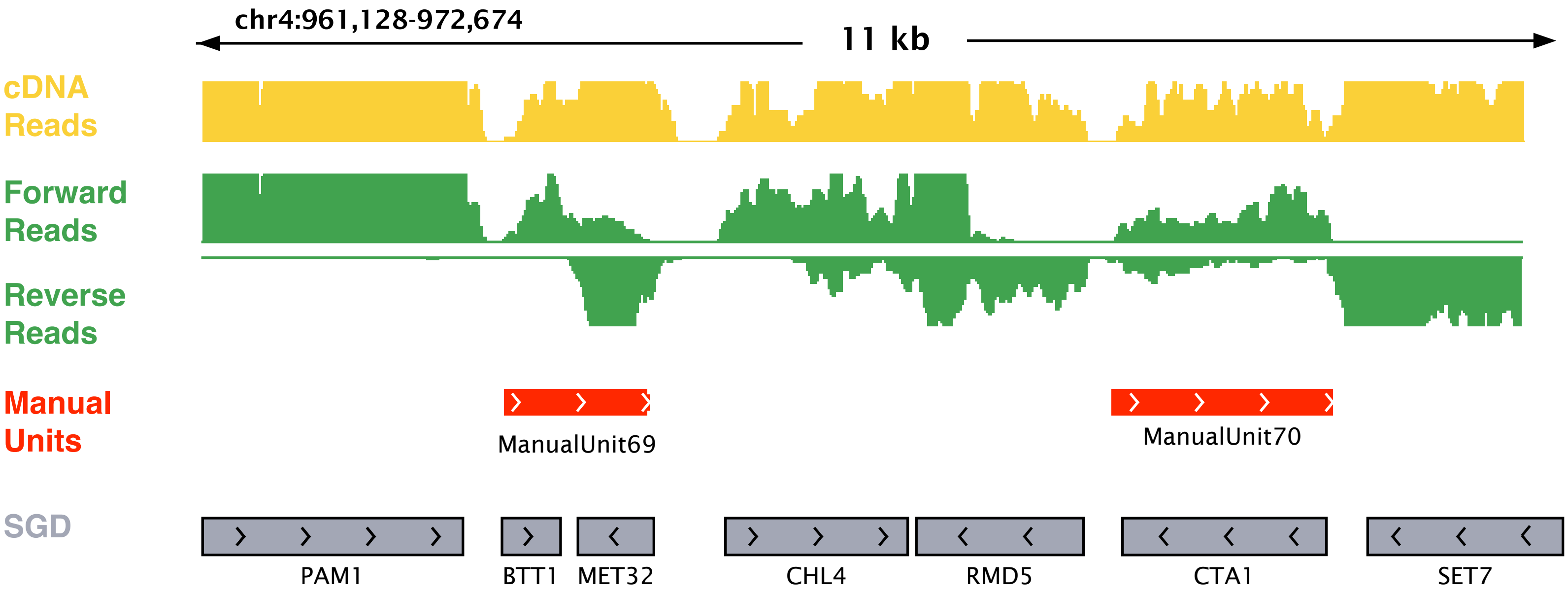
b Cumulative distribution function (cdf) of antisense units vs. other units



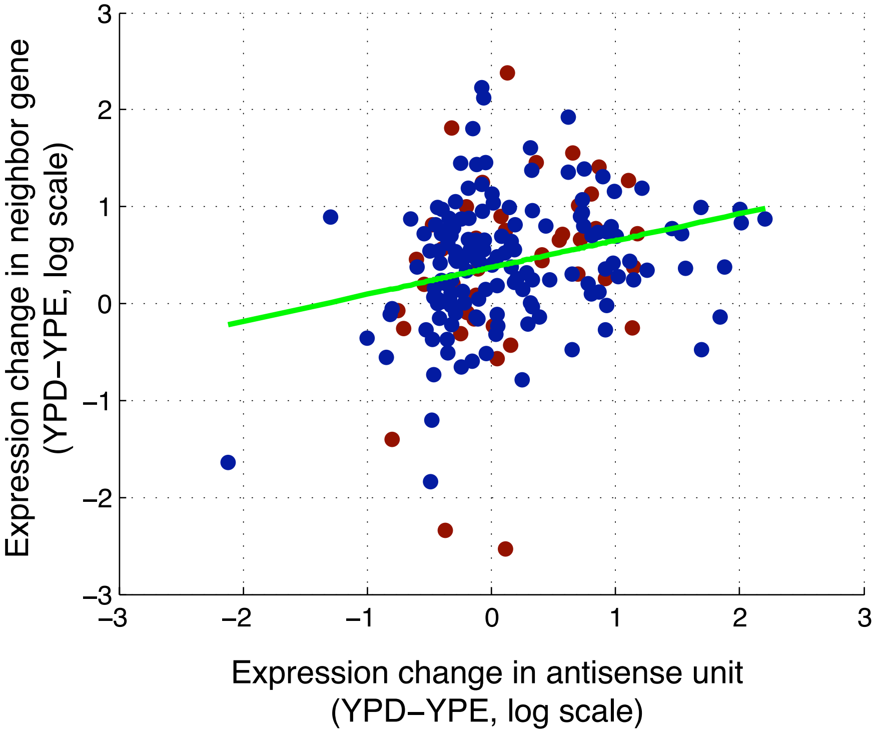
Supplementary Figure 3 - Manual Curation Example



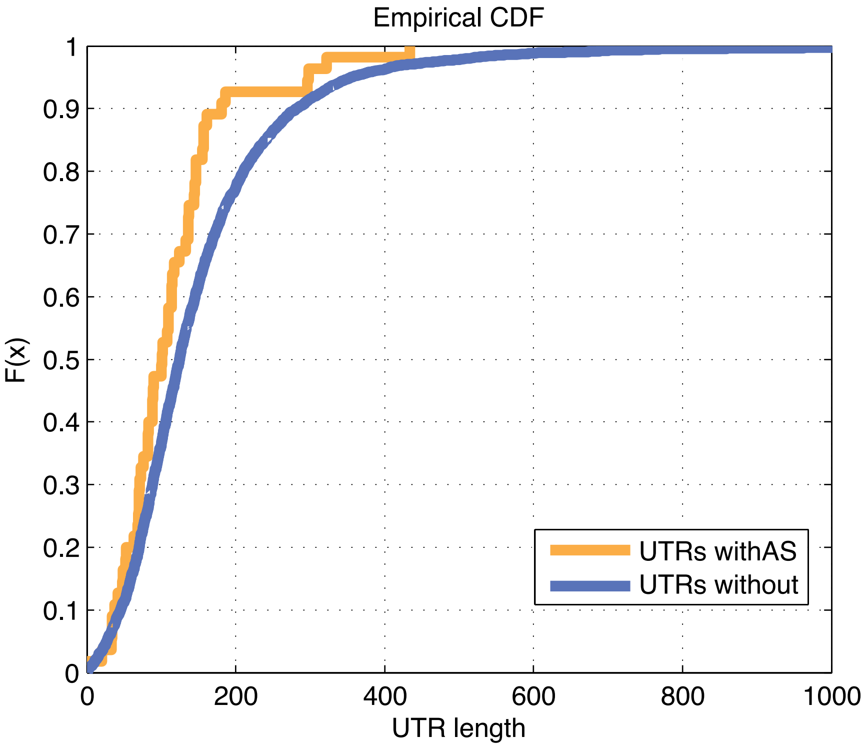
Supplementary Figure 4 - Antisense Units' Promoter Types



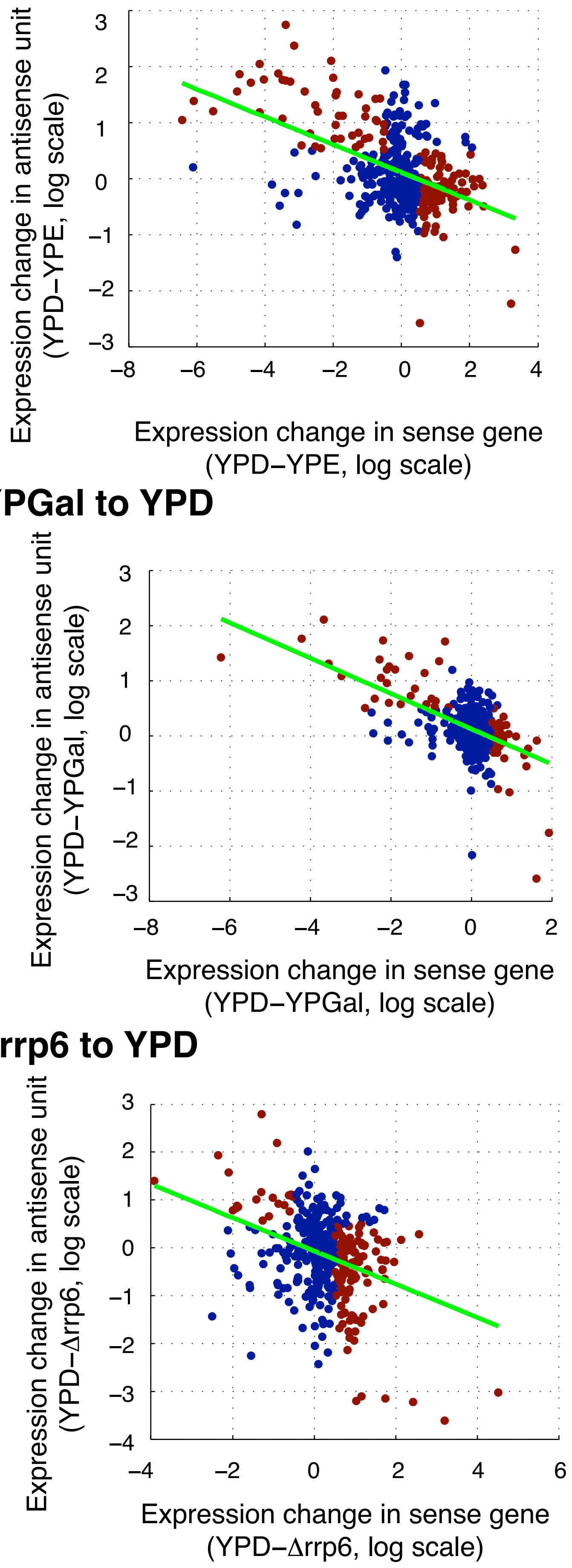
Supplementary Figure 5 - Expression patterns of antisense units and their neighboring genes



Supplementary Figure 6 - UTR length of genes with antisense ending close by



Supplementary Figure 7 - Expression Measurements

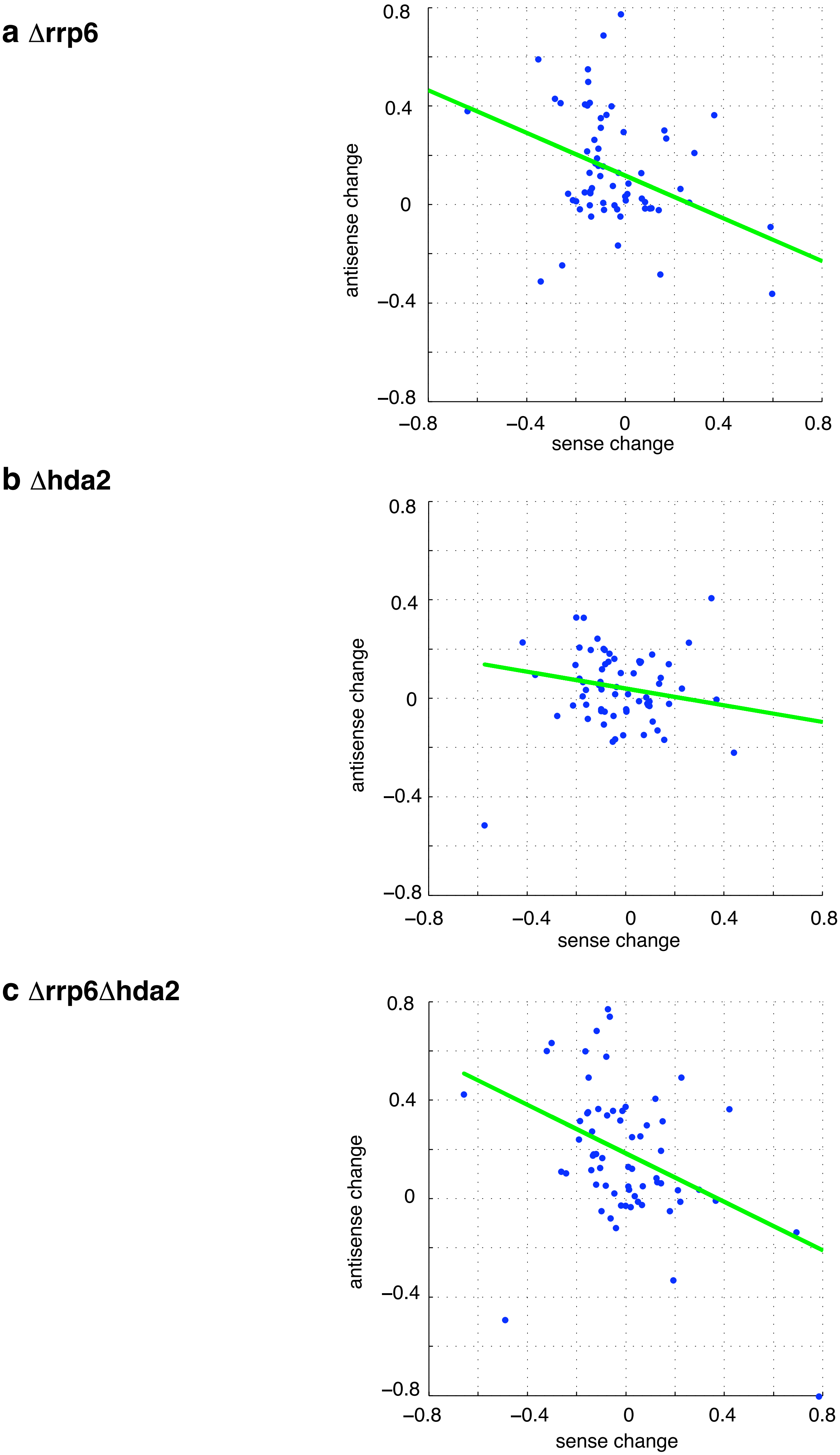


a Comparing YPE to YPD

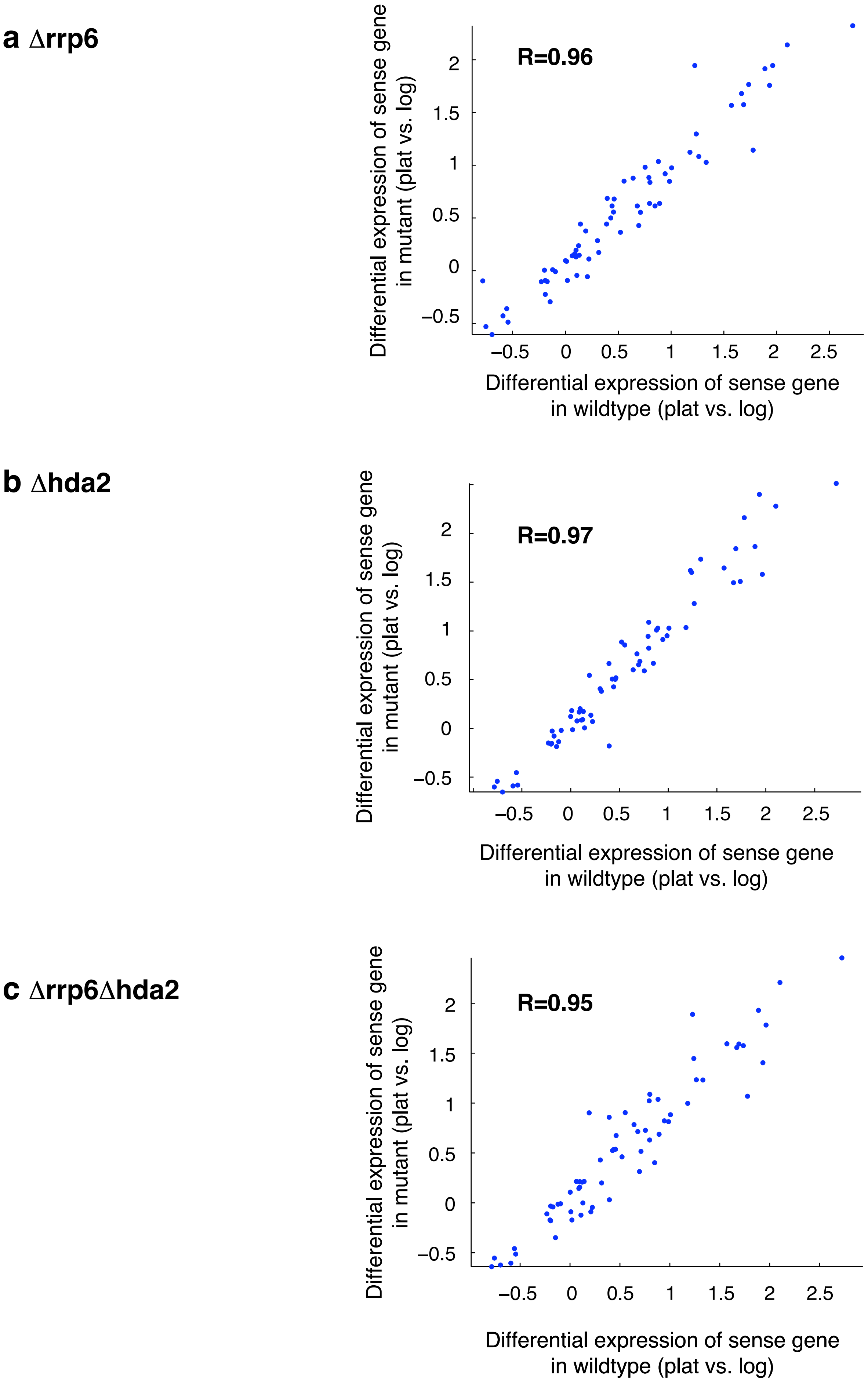
b Comparing YPGal to YPD

c Comparing Δrrp6 to YPD

Supplementary Figure 8 - Mutant Effect on Transcription



Supplementary Figure 9 - Mutant effect on sense differential expression



a Δrrp6

b Δhda2

c Δrrp6Δhda2

Supplementary Information: Full-length transcriptome assembly from RNA-Seq data without a reference genome

Supplementary Information for Grabherr et al., 2011

Nature Biotechnology: doi:10.1038/nbt.1883

SUPPLEMENTARY NOTE

Assembly of the fission yeast transcriptome

Inchworm assembled the yeast data set into 811,364 contigs with length at least 48 bases (2*(k-1), k=25, see above). Only 8,234 of the contigs are at least 350 bases long (approximately the mean insert size in our RNA-Seq library) and those comprise 13.4 Mb of total sequence. At this stage, 15% (660 of 4265) of the Inchworm-reconstructed, Oracle-matching, transcripts were recovered as falsely fused into single contigs. These mostly correspond to adjacent genes that overlap in their untranslated regions (UTRs), a common phenomenon in yeasts[1, 2]. By examining the clustering of read mate-pairings, 375 of the 660 falsely-fused transcripts were automatically teased apart into individual full-length transcripts (see above). Chrysalis grouped all contigs into 23,607 components and built a set of de Bruijn graphs, with a total of 24M unique k-mer nodes. After filtering and analyzing the graph, Butterfly outputs 27,841 linear contigs longer than 100 bases, grouped into a final set of 23,232 components.

Assembly of the mouse transcriptome

First, Inchworm assembled the reads into ~1.9M contigs (43 Mb resides in 32,466 sequences >= 350 bp), containing 7,346 annotated full-length transcripts. **Second**, Chrysalis pooled the contigs into 156,211 components. **Finally**, Butterfly reported 179,340 contigs (48,497 of length greater than 350bp), residing in 151,115 remaining components, fully capturing the 8,185 transcripts at 7,749 loci at full length.

SUPPLEMENTARY METHODS

Yeast strains and growth conditions.

Cultures were grown in the following rich medium: Yeast extract (1.5%), Peptone (1%), Dextrose (2%), SC Amino Acid mix (Sunrise Science) 2 grams per liter, Adenine 100 mg/L, Tryptophan 100 mg/L, Uracil 100 mg/L, at 200 RPM in an New Brunswick Scientific air-shaker.

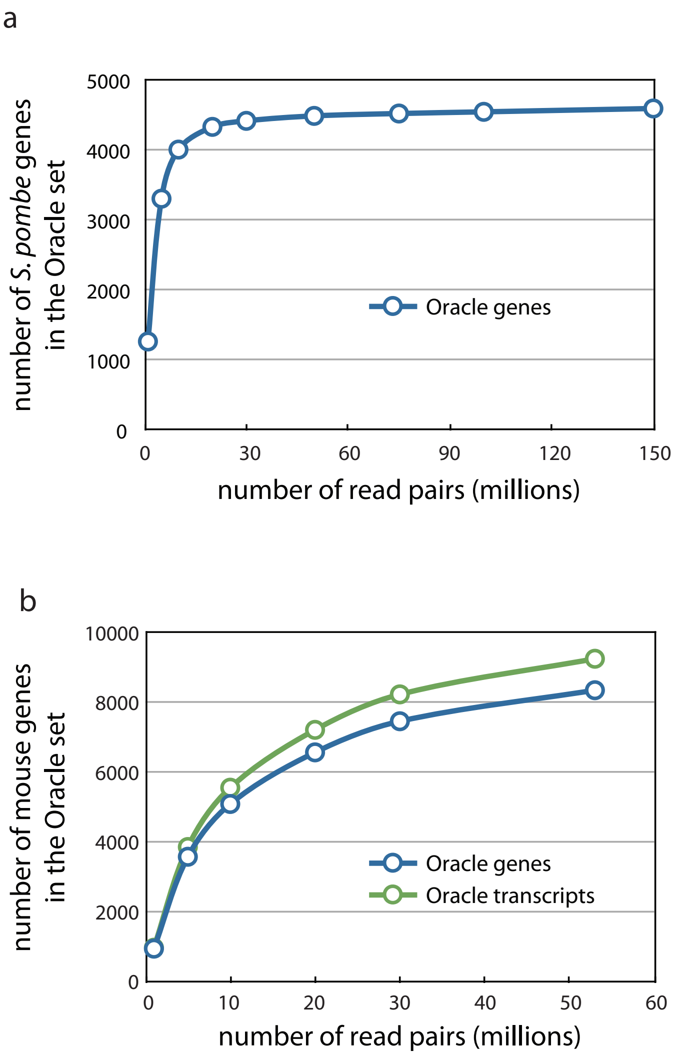
For glucose depletion (mid-log, diauxic shift, and stationary phase samples), overnight cultures were grown to saturation in 3 ml rich medium. From the 3 ml overnight cultures, 300 ml of rich media was inoculated at the OD_{600} corresponding to $1x10^6$ cell/ml and grown in New Brunswick Scientific shaking water baths. Culture density was monitored by OD_{600}. Glucose levels were monitored using the YSI 2700 Select Bioanalyzer. Cells were harvested at mid-log, diauxic shift (defined as the timepoint when glucose is depleted from the medium), and when growth plateaus by quenching them in 60% liquid methanol at -40°C that was later removed by centrifugation at -9°C and stored overnight at -80°C. Harvested cells were later washed in RNAse-free water and archived in RNAlater (Ambion) for future preparations.

For heat shock, overnight cultures were grown in 650ml of media at 22°C to between $3x10^7$ and $1x10^8$ cell/ml $OD_{600} = 1.0$. The overnight culture was split into two 300ml cultures and cells from each were collected by removing the media via vacuum filtration (Millipore). The cell-containing filters were re-suspended in pre-warmed media to either control (22°C) or heat-shock temperatures (37°C). Density measurements were taken approximately one minute after cells were re-suspended to ensure that concentrations did not change during the transfer from overnight media. 60ml of culture were harvested at 15 minutes after re-suspension by quenching them in 60% liquid methanol at -40°C that was later removed by centrifugation at -9°C and stored overnight at -80°C. Harvested cells were later washed in RNAse-free water and archived in RNAlater (Ambion) for future preparations. Cells were also harvested from cultures just before treatment for use as controls.

Mouse dendritic cell isolation and tissue culture

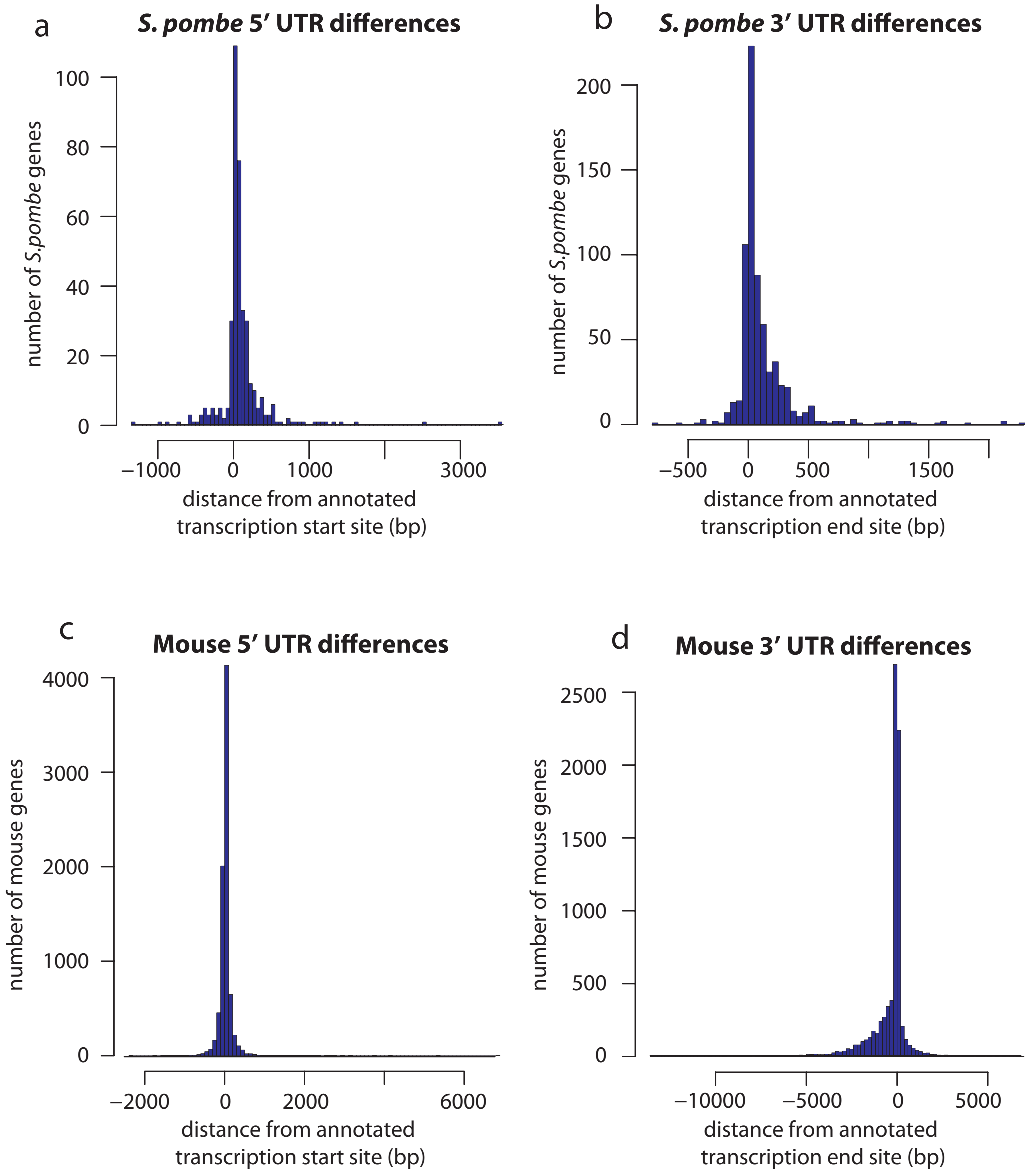
6-8 weeks female C57BL/6J mice were obtained from the Jackson Laboratories. Bone Marrow DCs were collected from femora and tibiae and plated on non-tissue culture treated plastic dishes in RPMI medium (Gibco Invitrogen) supplemented with 10% FBS, L-glutamin, penicillin/streptomycin, MEM non-essential amino acids, HEPES, sodium pyruvate, β-mercaptoethanol, and GM-CSF (15 ng/mL; Peprotech). At day 5, floating CD11c+ cells were collected and sorted on MACS columns using the CD11c (N418) MicroBeads kit (Myltenyi Biotec). CD11c+ cells where replated at a concentration of 10^6 cells/ml and collected 12 hours post sorting.

SUPPLEMENTARY FIGURES AND LEGENDS



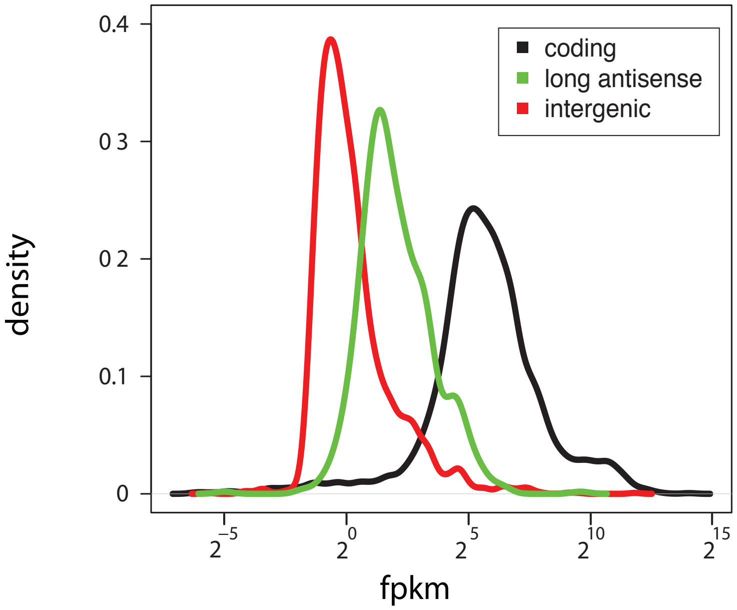
Supplementary Figure 1. Impact of the number of reads on the oracle set.

Shown are the numbers of *S. pombe* genes (**a,** blue) or mouse genes (**b**, blue) or transcripts (**b**, green) that are captured by the Oracle set at different numbers of input read pairs (x axis). The oracle set begins to saturate at 25M read pairs (or 50M reads) for the *S. pombe* RNA-Seq data (**a**), but is likely not saturated with the entire set of 53M read pairs on the mouse data set (**b**).



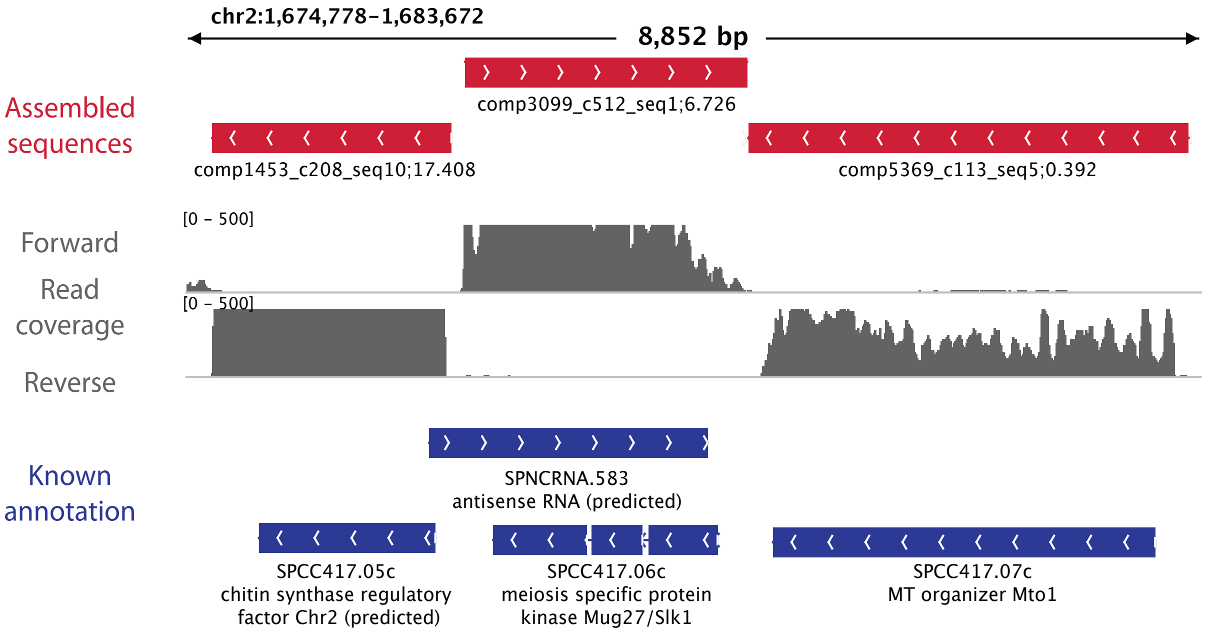
Supplementary Figure 2. UTR differences between Trinity transcripts and the annotated reference.

Shown are the distributions of changes in UTR length between Trinity transcripts and the annotated reference at the 5'UTR (a,c) and 3'UTR (b,d) of *S. pombe* (**a,b**) and mouse (**c,d**).



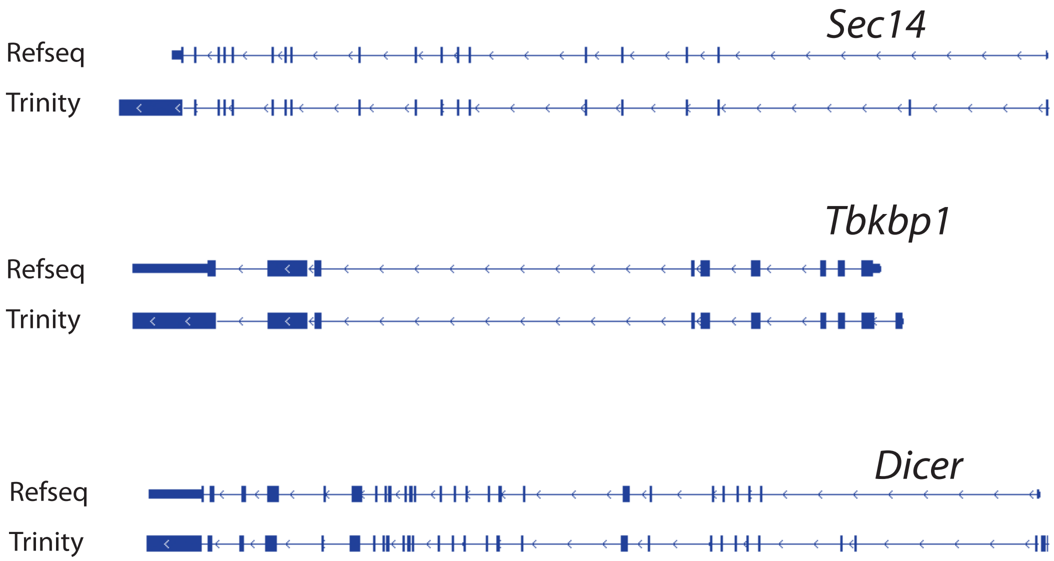
Supplementary Figure 3. Distribution of expression levels for protein-coding and antisense transcripts.

Shown are the distributions of expression levels (FPKM) for coding (blue), long antisense (green), and intergenic (red) Trinity-assembled transcripts in *S. pombe*.



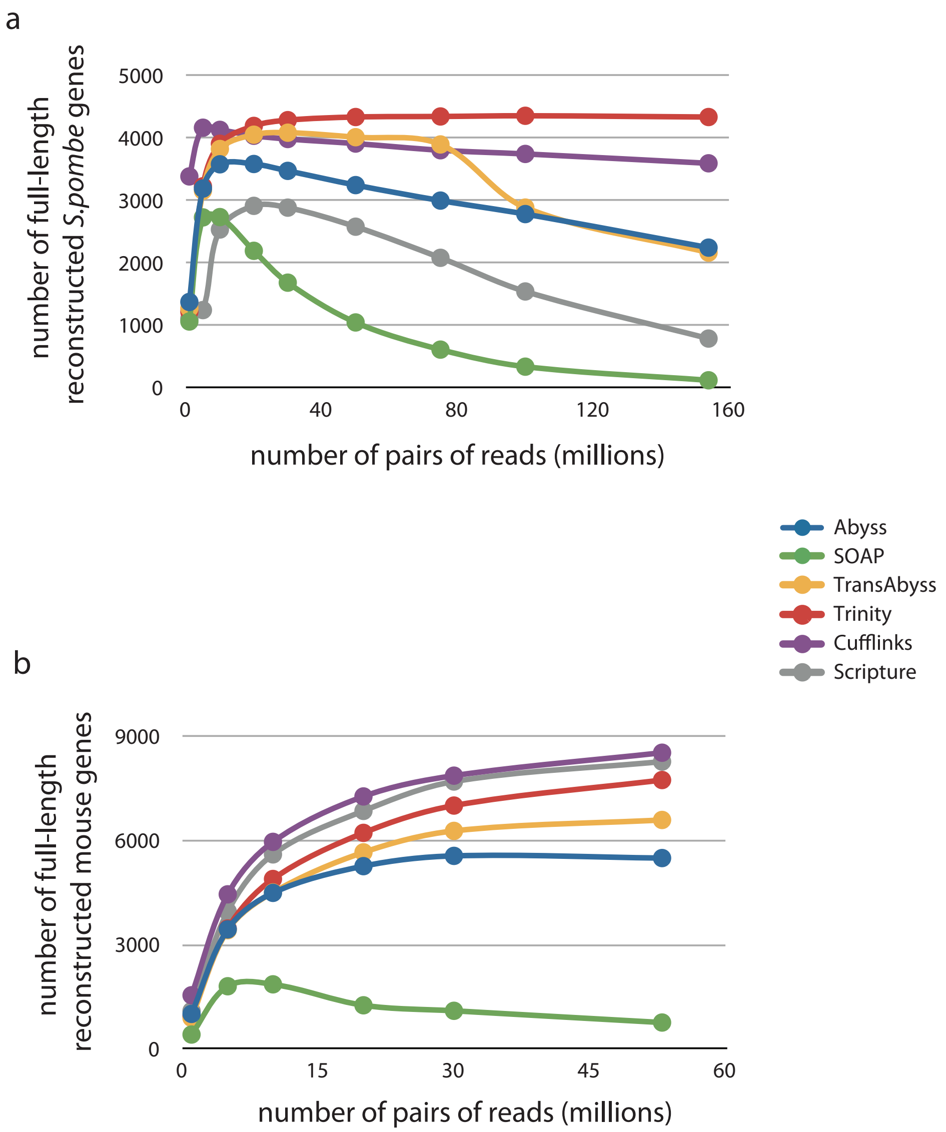
Supplementary Figure 4. Trinity identifies antisense transcription in yeast.

Shown are examples of Trinity assemblies (red) along with the corresponding annotated transcripts (blue) and coverage of underlying reads (green) all aligned to the *S. pombe* genome (for graphical clarity; no alignments were used to generate the assemblies). Trinity's assembly of comp3099 corresponds to the predicted antisense transcript SPNCRNA.583.



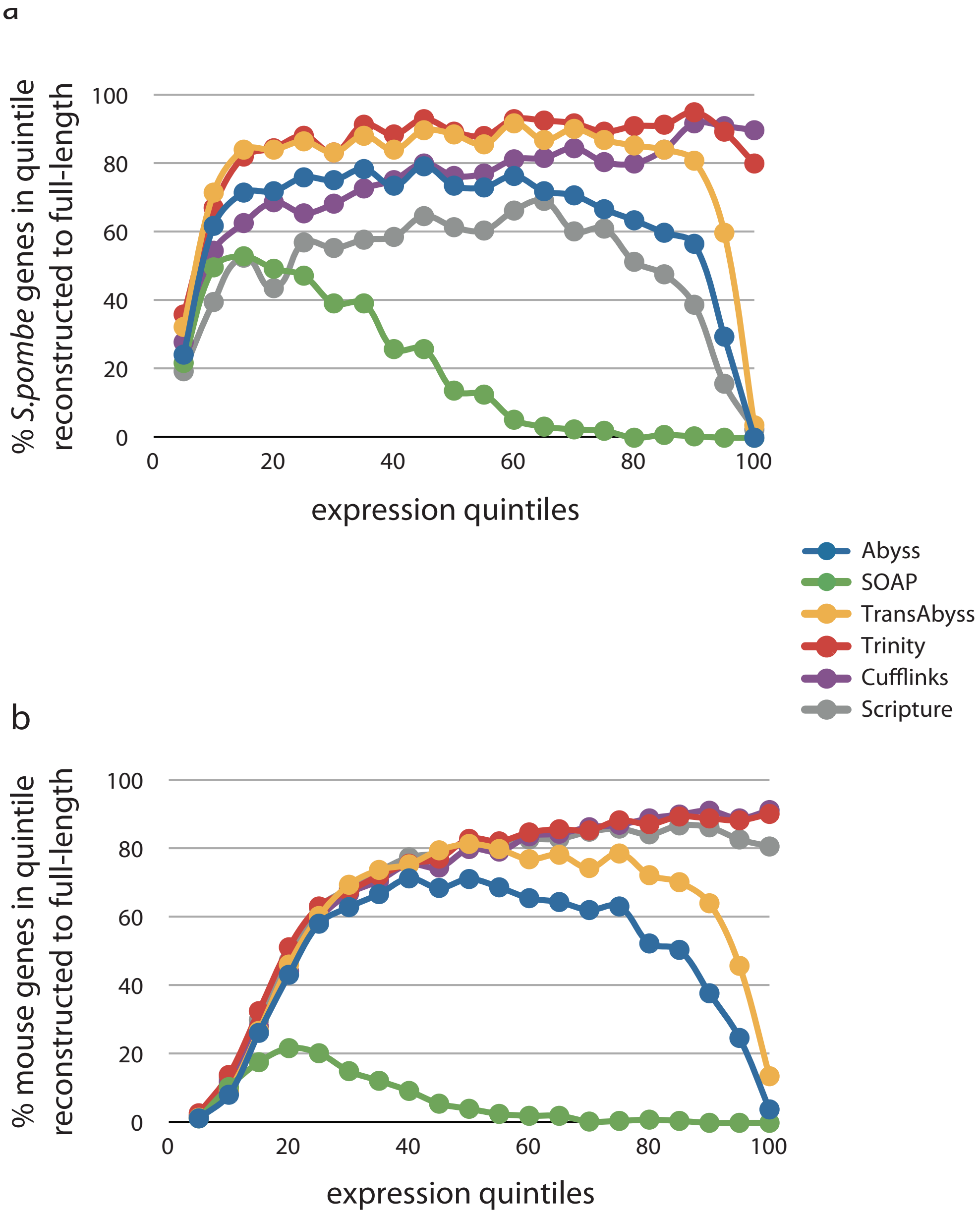
Supplementary Figure 5. Examples for UTR exon additions in mouse.

Shown are examples of Trinity assemblies (bottom) and the corresponding reference annotation (top) for (**a**) *Sec14* (one extra internal UTR exon), (**b**) *Tbkbp1* (one extra UTR exon at the 5' end), and (**c**) *Dicer* (multiple internal and 5' end UTR exons).



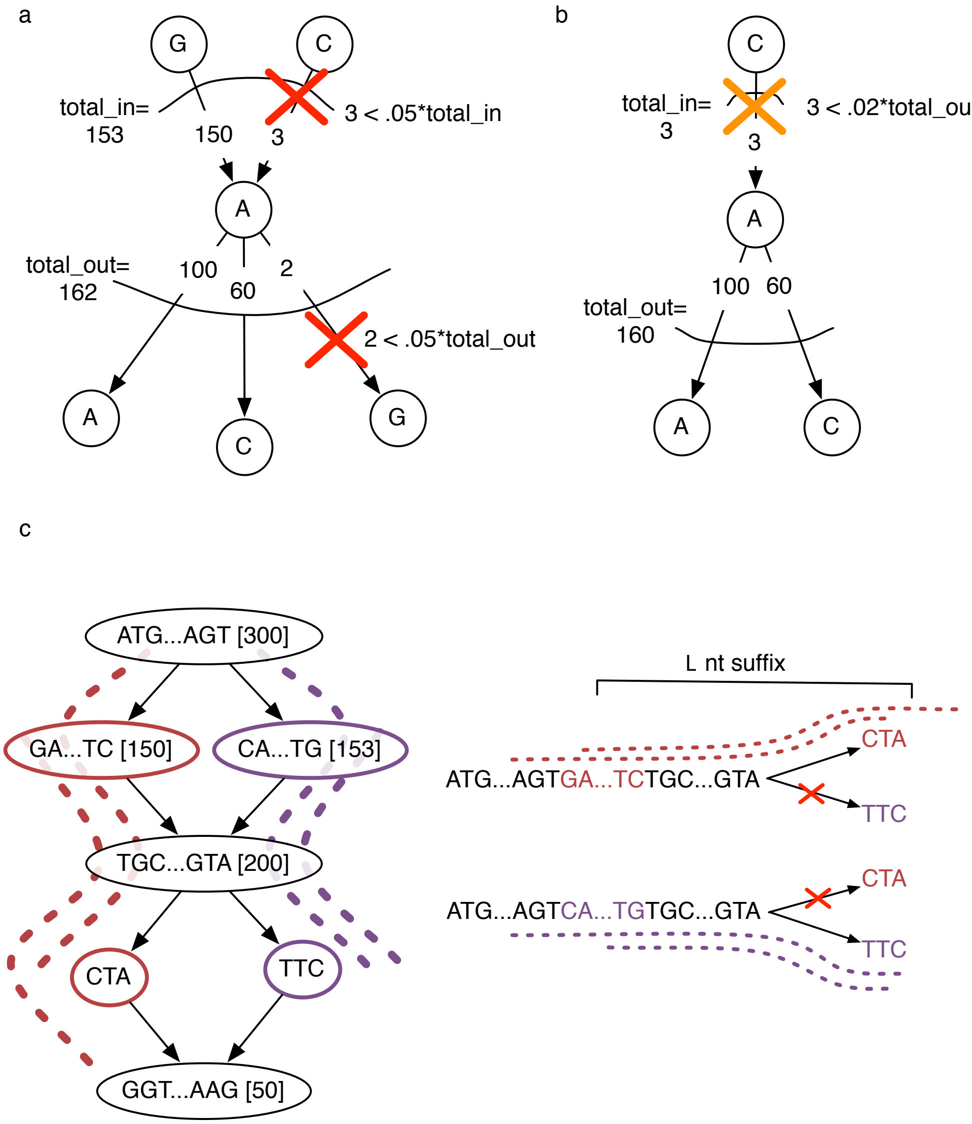
Supplementary Figure 6. The number of full-length transcripts reconstructed by each method at different numbers of input reads.

Shown are the number of annotated full-length transcripts (Y axis) reconstructed at different input read numbers (X axis) for each of Trinity (red), TransAbyss (yellow), Abyss (blue), SOAPdenovo (green), Scripture (purple) and Cufflinks (grey) in yeast (**a**) and mouse (**b**).



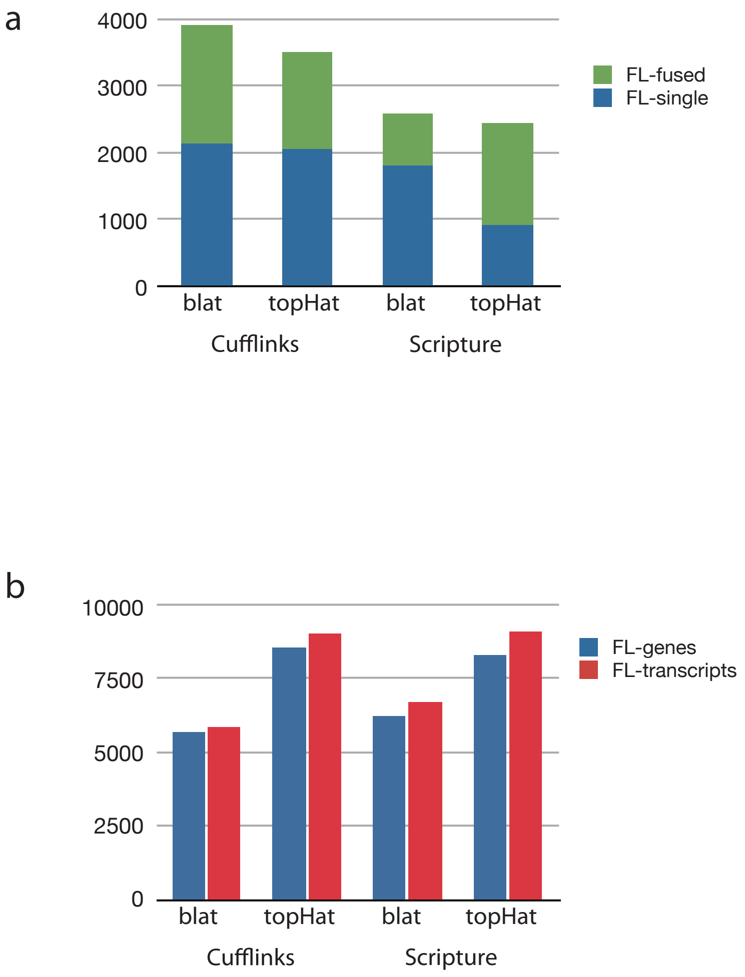
Supplementary Figure 7. The number of full-length transcripts reconstructed by each method at different expression levels.

Shown are the numbers of full-length Oracle transcripts (Y axis) reconstructed at different expression quintiles (X axis) by each of Trinity (red), TransAbyss (yellow), Abyss (blue), SOAPdenovo (green), Scripture (purple) and Cufflinks (grey) in yeast (**a**) and mouse (**b**).



Supplementary Figure 8. Butterfly edge pruning and path finding.

(a,b) Shown are the two cases where we would remove an edge (respectively, see **Methods**). **(c)** Illustrates the progress of the path finding process. On the left we see the compacted graph, each node shows the beginning and end of its sequence, and its length in square brackets. On the right we examine 2 possible extensions for the two paths that reached node (TGC…GTA), and show that we have L-suffix support only for the red and purple paths, and not their chimera paths.



Supplementary Figure 9. The effect of the choice of alignment program on mapping-first transcriptome reconstruction.

Shown are the numbers of annotated full-length transcripts (blue) and full-length fused transcripts (green) reconstructed by Cufflinks and Scripture, based on Blat and the latest version of Tophat in yeast (**a**) and mouse (**b**).

SUPPLEMENTARY TABLES

Supplementary Table 1. Comparison of sensitivity of different methods on the S. pombe transcriptome.

Listed are the number of full-length (FL) genes, the percentage of false fusions, the total number of contigs, the number of contigs that could be mapped to the genome, the number of genes hat overlap with mapped contigs, and the average number of contigs per gene.

	Scripture (blat)	Cufflinks (blat)	ABySS	Trans-ABySS	SOAP-denovo	Trinity
FL genes	2585	3913	3248	4015	1049	4338
% falsely fused genes	30	45	36	27	26	5
Total contigs	14909	4605	6343	39178	12392	27841
Contigs mapped	11714	3258	4601	31974	5456	7057
Genes captured	3838	4182	4533	4871	3400	4874
Average contig coverage/ gene	4.37	1.07	1.06	5.08	1.01	1.37

Nature Biotechnology: doi:10.1038/nbt.1883

Supplementary Table 2. Comparison of sensitivity of different methods on the mouse transcriptome.

Listed are the number of full-length (FL) genes, the percentage of false fusions, the total number of contigs, the number of contigs that could be mapped to the genome, the number of genes hat overlap with mapped contigs, and the average number of contigs per gene.

	Scripture (tophat)	Cufflinks (tophat)	ABySS	Trans-ABySS	SOAP-denovo	Trinity
FL transcripts	9086	9010	5561	7025	761	8185
FL genes	8293	8536	5500	6598	760	7749
Total contigs	300148	31121	46783	203085	145518	179340
Contigs mapped	119515	19342	17427	111309	34816	31706
Genes captured	10432	10806	9879	10685	10035	11334
Average contig coverage / gene	12.0	1.65	1.25	5.93	1.12	2.05

Supplementary Table 3. Base error stats for Trinity transcripts.

Listed are the number of aligned bases, matches, mismatches, insertions and deletions.

	S. pombe	Mouse
# Full-length Trinity Transcripts	4230	8178
# aligned bases	8942895	21400061
# matching bases	8942241	21397375
# mismatches	654	2686
Mismatch rate	7.31e-05	1.26e-04
# genome inserted bases	299	1551
Genome inserted base rate	3.34e-05	7.25e-05
# transcript inserted bases	528	2875
Transcript inserted base rate	5.90e-05	1.34e-04

References

1. Wilhelm, B.T. et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239-1243 (2008).
2. 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).

אפיון הטרנסקריפטום תוך שימוש
בטכנולוגיות ריצוף מתקדמות

חיבור לשם קבלת תואר דוקטור לפילוסופיה

מאת

מורן יסעור

הוגש לסנאט האוניברסיטה העברית, בירושלים

יולי 2012

קרובים, מה שמחזק את דעתנו שהם פוקציונלים ולא רק תופעת לוואי של שעתוק הגן המקודד.

לבסוף, אני מציגה שיטה לאפיון טרנסקריפטומים מורכבים ומשוחברים, אפילו בהעדר רצף הגנום. שיטה זו נקראת Trinity, והיא פותחה יחד עם בריאן האס ומנפרד גרבהר (Brian Haas, Manfred Grabherr) ממכון הברוד. בשיטה זו, אנו מרכיבים מכל המילים מעין פאזל במטרה ליצר רצפים ארוכים, שמכוסים כולם במילים מרוצפות. כדי לעשות זאת אנו משתמשים בעקרונות הרכבה של דנ"א (DNA assembly), ומתאימים אותם להרכבת רנ"א. גישה זו נקראת "הרכבה תחילה", כיוון שראשית אנו מרכיבים את המילים בינן לבין עצמן, ולבסוף ממפים את הרצפים הארוכים לגנום, במידה והוא רוצף. בשיטה זו אנו מתגברים בנקל על בעיית מיפוי המילים שמקורן בצמתי השחבור וכן מצליחים לאפיין איזופורמים שונים במלואם. בעזרת שיטה זו, אפיינו טרנסקריפטומים שונים, החל בטרנסקריפטום הפשוט והצפוף של שמר האופים וכלה בטרנסקריפטום המורכב והמשוחבר רבות של העכבר. בנוסף, אפיינו את הטרנסקריפטום של עש הטבק (*Bemisia tabaci*), יצור שאין לו כיום גנום מרוצף, והשוונו את החלבונים שקיבלנו למאגר החלבונים הידועים כיום.

לסיכום, בעבודה זו אני מציגה כלים חישוביים לאפיון הטרנסקריפטום מנתוני RNA-Seq, שניתן להפעילם על מגוון אורגניזמים, בין אם רצף הגנום שלהם ידוע ובין אם לאו. בנוסף, השתמשתי בכלים אלו כדי ללמוד על בקרת השעתוק בשמרים, תוך דגש על זיהוי גני אנטיסנס אשר מבקרים גנים מקודדי חלבון. שיטות חישוביות לאפיון טרנסקריפטום, ובמיוחד אלו שאינן מסתמכות על ריצוף גנום קיים פותחות צוהר למחקר של יצורים חדשים, וכן של תאים בהם הרצף הקיים שונה מזה שרוצף בעבר, כמו למשל במקרים רבים של רקמות סרטניות.

טכנולוגיות אלו פתחו צוהר למחקרים גנומיים חדשים, כמו למשל ריצוף של הטרנסקריפטום השלם (RNA-Seq). מרבית המחקרים השתמשו ב RNA-Seq בעיקר לשם מדידה מדוייקת יותר של רמות ביטוי הגנים, זיהוי של איזופורמים חלופיים (splice isoforms) ושיפור הבנתנו של גבולות הגנים. אבל רבים ממחקרים אלו מסתמכים על ידע מוקדם של רצף הגנים, מיקומם על פני הגנום או על ריצוף הגנום כולו, וכך מגבילים את יכולתנו לזהות גנים חדשים ולחקור יצורים שהגנום שלהם לא רוצף עדין.

במהלך עבודה זו אני מציגה אסופת מחקרים בהם פיתחתי טכנולוגיות וכלים לאנליזה של ניסויי RNA-Seq, ואפליקציה של כלים אלה באורגניזמים שונים, החל משמר האופים ועד עכבר.

ראשית, אני מציגה שיטה חדשה לאפיון טרנסקריפטומים פשוטים יחסית, כמו למשל זה של שמר האופים, המסתמכת על ריצוף קיים של הגנום (פרק 2). בגישה זו הנקראת "מיפוי תחילה", אנו ממפים את המילים הקצרות לגנום, ולאחר מכן מאפיינים אזורים בגנום המכוסים על ידי מילים אלה, ולכן שועתקו לרנ"א. בעבודה זו גם פיתחנו אלגוריתם חדש למיפוי המילים לגנום, כדי להתגבר על בעית מיפוי מילים שמחולקות על פני כמה אזורים בגנום, כמו למשל במקרה של מילה הנמצאת בצומת של גן משוחבר (splice junction). כך הצלחנו לשחזר במלואם 85% מהגנים המבוטאים, וכן לאפיין 254 מתוך 305 צמתי השחבור. אחד האתגרים הגדולים היה להצליח להפריד בין גנים שכנים, אשר חופפים מעט על פני הגנום, אך שוכנים על גדילים שונים. כדי להתגבר על אתגר זה, פנינו לחפש שיטה שבה רק הגדיל המשועתק יהיה זה שירוצף בניסוי.

בעבודה המוצגת בפרק 3, שיתפתי פעולה עם ג'ושוע לוין ממכון הברוד (Joshua Levin, Broad Institute) כדי למצוא את הפרוטוקול האידאלי לריצוף רנ"א, תוך שמירה על הגדיל המשועתק. אני פיתחתי תשתית חישובית להשוואת כל הספריות שקיבלנו על ידי הגדרת קריטריונים המודדים איכות של כל ספריה, לשימושים השונים. כאשר שיקללנו גם את כמות העבודה הניסויית הנדרשת ליצור כל ספריה, מצאנו את הפרוטוקול המייטבי לספריות RNA-Seq, והוא הפרוטוקול הסטנדרטי מאז בכל הניסויים במכון זה.

לאחר מציאת הפרוטוקול האידאילי, חזרתי למשימה המקורית של אפיון הטרנסקריפטום של שמר האופים. כעת כשהיו בידי נתונים רק על הגדיל המשועתק, יכולתי להבדיל בקלות בין גנים שכנים, וגם לזהות מקרים רבים של גני אנטיסנס, המשועתקים באזור חופף לגנים מקודדים לחלבון, רק בגדיל השני (פרק 4). גיליתי כ 225 גנים מקודדי חלבון עם שעתוק בגדיל השני, וכשבחנתי גנים אלו מצאתי שרבים מהם קשורים להתמודדות של התאים עם תנאי עקה (חום, חוסר במזון וכן הלאה). בניתוח של ביטוי גנים אלו בתנאים שונים, גילינו שיש התאמה הפוכה בין ביטוי הגן המקודד לבין ביטוי האנטיסנס, כך שיתכן מאד שהאחד מבקר את השני. בנוסף, מצאנו כי חלק מגני האנטיסנס הללו שמורים בחמישה מיני שמרים

תקציר

המידע התורשתי שלנו אגור ברצף הדנ"א (הגנום) ומכיל את המתכונים ליצירת כל הרכיבים של התא, כך שכל מתכון כזה נמצא באזור מסויים בגנום ונקרא גן. לפי התפיסה המרכזית של הביולוגיה, הדנ"א של הגן משועתק ל רנ"א שליח (messenger RNA), שלאחר מכן מתורגם לחלבון. החלבונים הם אבני הבניין של התא, ומבצעים את מרבית הפעולות בו.

כל התאים ביצור החי מכילים את אותו הגנום בדיוק, ולכן מכילים את אותם הגנים, אך עדין ישנם הבדלים משמעותיים הן בצורה והן בתפקיד של תאים ברקמות שונות, או כתגובה לסביבה שונה. מרבית מהבדלים אלו מבוקרים על ידי בחירת הגנים שיופעלו בכל רגע נתון ובקרה זו נקראת בקרת השעתוק.

באופן היסטורי, אזורים בדנ"א נקראו גנים רק במידה והם תורגמו לחלבון, אבל היום אנחנו מכירים סוג נוסף של גנים שמתפקדים ברמת הרנ"א ואינם מקודדים לחלבון כלל. דוגמא לגנים מסוג זה הם גני ה אנטיסנס (antisense), אשר נמצאים בחפיפה לגנים מקודדי חלבון, רק בגדיל השני של הדנ"א. כאשר גן האנטיסנס משועתק לרנ"א, הוא יכול לגרום לירידה ברמת השעתוק של הגן המקודד.

בבואנו לחקור אורגניזם שזה עתה רוצף הגנום שלו, אחד הצעדים הראשונים הוא לאפיין את כל הגנים שיש לו (מקודדים ולא מקודדים כאחד), כך נוכל גם לחזות את אוסף החלבונים הפוטנציאלים ביצור זה. באופן אידיאלי, נרצה לדעת את המיקום המדויק של הגנים, וגם להבין מתי, מדוע וכיצד הם מבוטאים, אך המשימה הפשוטה ביותר היא ראשית לאפיין את מיקומם בגנום. ביצורים אאוקריוטיים פשוטים (כמו שמר האופים *Saccharomyces cerevisiae*), הגנים ממוקמים בצפיפות רבה על פני הגנום, ומרביתם אינם משוחברים (spliced). ביונקים, לעומת זאת, הגנים מהווים חלק קטן מאד מהגנום, כך שבאדם למשל, רק שני אחוז מרצף הגנום מכיל גנים. עובדה זו מקשה מאד מציאת מיקום הגנים רק מתוך הרצף, ולכן ריצוף הגנום הוא רק צעד ראשון בדרך לחקר אורגניזם חדש, אך נדרשים צעדים משמעותיים נוספים.

דרך נוספת לאפיון כל הגנים המשועתקים, היא בחינת אוסף מולקולות הרנ"א שיש בתא (אוסף זה נקרא גם הטרנסקריפטום של התא). אפיון הטרנסקריפטום של יצור אאוקריוטי היא משימה מאתגרת מאד, ובעבר נעשה שימוש בשבבי דנ"א (tiling microarrays) או בריצוף קטעים קצרים מתוך הרנ"א (Expressed Sequenced Tags) לשם מטרה זו. גישות אלו מסתמכות על רצף גנום ידוע או דורשות השקעת משאבים רבה עבור כל גן וגן. בשנים האחרונות פותחו טכנולוגיות ריצוף מתקדמות הנקראות ”next generation sequencing“ או ”high-throughput sequencing“. טכנולוגיות אלו מאפשרות לנו לרצף עשרות מליוני מילים קצרות (reads) מדוגמת דנ"א בודדת, בעלות נמוכה ובמהירות חסרת תקדים.

עבודה זו נעשתה בהדרכתם של:

פרופ' ניר פרידמן ופרופ' אביב רגב

אפיון הטרנסקריפטום תוך שימוש בטכנולוגיות ריצוף מתקדמות

חיבור לשם קבלת תואר דוקטור לפילוסופיה

מאת

מורן יסעור

הוגש לסנאט האוניברסיטה העברית, בירושלים

יולי 2012