

Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution

Brian T. Wilhelm^{1*†}, Samuel Marguerat^{1*†}, Stephen Watt^{1†}, Falk Schubert^{1†}, Valerie Wood¹, Ian Goodhead^{1†}, Christopher J. Penkett^{1†}, Jane Rogers¹ & Jürg Bähler^{1†}

Recent data from several organisms indicate that the transcribed portions of genomes are larger and more complex than expected, and that many functional properties of transcripts are based not on coding sequences but on regulatory sequences in untranslated regions or non-coding RNAs^{1–9}. Alternative start and polyadenylation sites and regulation of intron splicing add additional dimensions to the rich transcriptional output^{10,11}. This transcriptional complexity has been sampled mainly using hybridization-based methods under one or few experimental conditions. Here we applied direct high-throughput sequencing of complementary DNAs (RNA-Seq), supplemented with data from high-density tiling arrays, to globally sample transcripts of the fission yeast *Schizosaccharomyces pombe*, independently from available gene annotations. We interrogated transcriptomes under multiple conditions, including rapid proliferation, meiotic differentiation and environmental stress, as well as in RNA processing mutants to reveal the dynamic plasticity of the transcriptional landscape as a function of environmental, developmental and genetic factors. High-throughput sequencing proved to be a powerful and quantitative method to sample transcriptomes deeply at maximal resolution. In contrast to hybridization, sequencing showed little, if any, background noise and was sensitive enough to detect widespread transcription in >90% of the genome, including traces of RNAs that were not robustly transcribed or rapidly degraded. The combined sequencing and strand-specific array data provide rich condition-specific information on novel, mostly non-coding transcripts, untranslated regions and gene structures, thus improving the existing genome annotation. Sequence reads spanning exon–exon or exon–intron junctions give unique insight into a surprising variability in splicing efficiency across introns, genes and conditions. Splicing efficiency was largely coordinated with transcript levels, and increased transcription led to increased splicing in test genes. Hundreds of introns showed such regulated splicing during cellular proliferation or differentiation.

To analyse the *S. pombe* transcriptome at the best possible resolution, we used Illumina 1G to sequence directly cDNA synthesized from poly(A)-enriched RNA. This approach kept the proportion of sequence reads from ribosomal RNA low (<10%) without biasing against messenger RNAs with short poly(A) tails¹². We obtained >23 million reads of an average length of 39.1 base pairs (bp), representing ~60 genome lengths, from cells proliferating exponentially in rich medium. In addition, we acquired >99 million reads of transcriptomes from five stages of meiotic differentiation, representing an additional ~190 genomes (Supplementary Table 1). Sequence reads were mapped back to both the spliced and the unspliced reference genome¹³ to determine the numbers of reads hitting each

genomic base-pair position. Approximately 60% of all reads specifically mapped to one genomic region over 100% of their sequence, whereas >85% of the reads uniquely mapped over 90% of their

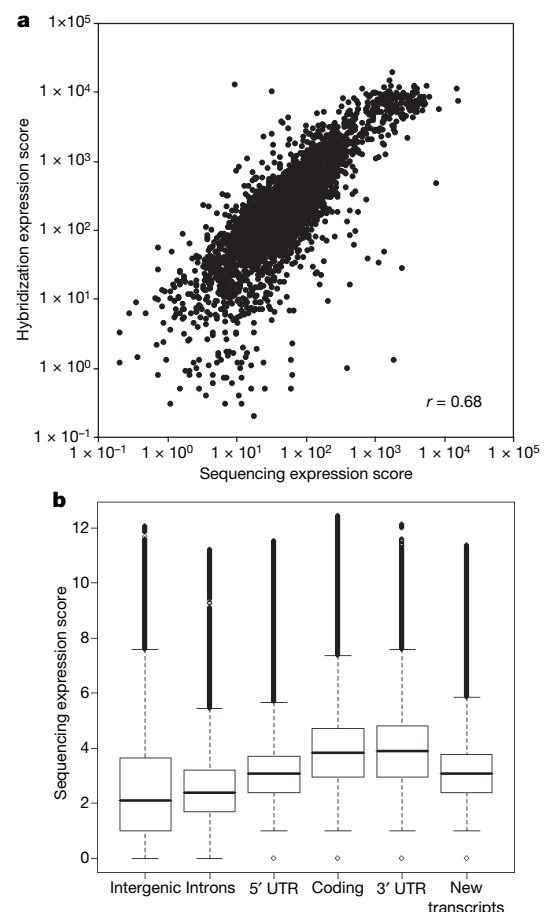


Figure 1 | Quantitation of sequence coverage. **a**, Scatterplot comparing gene-expression scores based on Affymetrix expression-chip hybridization signals (y axis) with gene-expression scores based on high-throughput sequencing (x axis). The dynamic range of hybridization signals is limited by the scanner. The corresponding Pearson correlation is shown at the bottom right. **b**, Box-and-whisker plots (in which the whiskers denote the 5th and 95th quantiles) of \log_2 -transformed numbers of sequence reads per nucleotide for the following genomic regions: all intergenic sequences, introns, coding sequences, 5' and 3' UTRs (based on sequencing), and newly identified transcripts (based on sequencing and tiling chips). Diamonds represent data outside of the quantiles.

¹Cancer Research UK Fission Yeast Functional Genomics Group, Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1HH, UK. [†]Present addresses: Institut de Recherche en Immunologie et en Cancérologie (IRIC), Montreal, H3C 3J7, Canada (B.T.W.); Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK (S.M., S.W., F.S. and J.B.); School of Biological Sciences, University of Liverpool, L69 7ZB, UK (I.G.); EMBL-European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK (C.J.P.). *These authors contributed equally to this work.

sequence. The remaining reads either mapped to repeated sequences or were of poor quality. RNA expression levels determined from sequence-read numbers strongly correlated with those determined from hybridization signals, indicating that sequencing provides quantitative data on transcript levels (Fig. 1a).

The 5% of transcripts present at the lowest steady-state levels in rapidly proliferating cells¹² accumulated ~777 sequence-read hits and 94.9% coverage on average, indicating that the transcriptome was sampled deeply enough to detect even genes with low expression levels. We modelled sequencing depth for rapidly proliferating cells: given the expression scores for all annotated genes, the model predicts that 99% of these genes have >50% sequence-read coverage (Supplementary Fig. 1). In agreement with this prediction, we obtained >50% sequence-read coverage for 99.3% of all annotated genes. The 41 genes with <50% coverage included 20 transposon-related long terminal repeats and 13 dubious genes or pseudogenes (Supplementary Table 2). Using cDNA microarrays, only 80–90% of genes yield measurable signals in proliferating cells¹⁴, whereas the remaining genes are only highly expressed under specific conditions such as meiosis or stress^{15,16}. These data suggest that the sequencing approach is sensitive enough to detect basal 'transcriptional noise' from genes that are not actively expressed.

As expected, intergenic regions were hit by fewer sequence reads than coding regions (Figs 1b and 2a). However, we obtained sequence data from ~94% and >99% of the nuclear and mitochondrial genomes, respectively, suggesting that almost the entire genome is

transcribed to some degree, consistent with the considerable overlap and complexity among different transcripts reported for other eukaryotes⁹. Reverse transcription followed by polymerase chain reaction (RT-PCR) controls verified that even intergenic regions with poor sequence-read coverage reflect expressed RNAs rather than technical noise from spurious sequences (Supplementary Fig. 2). Thus, our sequence data provide direct evidence for widespread transcription; it has been suggested that as much as 90% of all RNA polymerase II (Pol II) initiation events represent transcriptional noise¹⁷. Taken together, unlike for hybridization-based approaches, sequencing appears to produce little or no background noise, and the dynamic range of detected transcripts is only limited by sequencing depth.

To verify and compare the sequence data with an established platform, we used Affymetrix chips containing 25-mer probes tiled at ~20-nucleotide intervals across both strands of the *S. pombe* genome. We interrogated transcriptomes under a wide range of conditions (Supplementary Table 1), thus independently sampling gene expression at lower resolution but with strand-specific information (Fig. 2a).

The combined sequence and hybridization data revealed hundreds of novel transcribed regions. To distinguish between separate transcripts and extensions to known gene structures, we analysed tiling-chip data from a *prp2* splicing-factor mutant¹⁸ along with sequence 'trans-reads' spanning unannotated splice junctions (Figs 2a and 3d). Combined with manual curation, these analyses helped to refine annotated gene structures, including 75 revisions of protein-coding

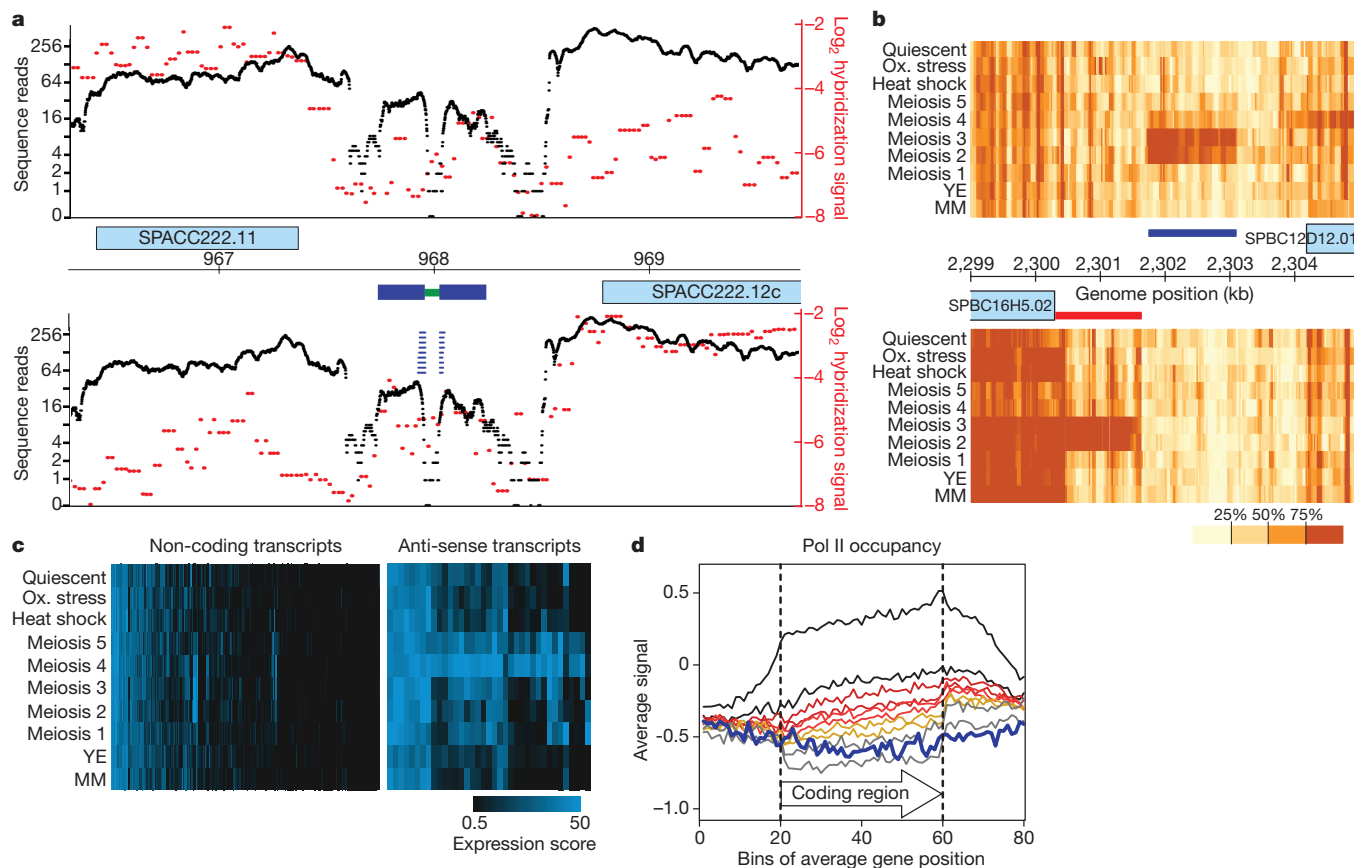


Figure 2 | Analyses of novel transcripts. **a**, Plot depicting numbers of sequence reads (log-scale, black) and tiling-chip hybridization signals (log₂, red) across the genomic region indicated in the centre (coordinates in kb) for forward (top) and reverse (bottom) strands. The sequence data are not strand-specific. A novel non-coding transcript (dark blue) containing an intron (green) is indicated in the middle. The nine trans-reads across the exon–exon junction are indicated as broken blue bars. **b**, Tiling-chip hybridization signals (in which the strength of colour reflects the signal-distribution quartile) across the genomic region shown in the centre for

forward (top) and reverse (bottom) strands, with rows reflecting ten experimental conditions (Supplementary Table 1). Rapid proliferation was sampled in rich (YE) and minimal (MM) media. Blue bar, novel meiosis-specific transcript; red bar, alternate meiosis-specific 5' UTR. **c**, Hierarchical clustering of non-coding and anti-sense transcripts by their tiling-chip expression scores across multiple conditions as in **b**. **d**, Average Pol II occupancy across coding regions for genes with lowest to highest mRNA levels (grey to black via red to yellow shades). The average profile of the novel transcripts is shown as a thick blue line.

regions and identification of ~20 new introns in known genes. Conservative data analysis also revealed 453 novel transcripts, only 26 of which seemed to be coding for small proteins (<150 amino acids); 37 of the apparently non-coding transcripts overlapped known genes in the anti-sense direction (Supplementary Table 3). The 427 non-coding RNAs showed an average length of ~825 nucleotides and a GC content that was similar to the 135 annotated non-coding RNAs but higher than for intergenic regions overall (33.0% versus 30.6%; $P < 2 \times 10^{-16}$, Wilcoxon test). The non-coding RNAs included the elusive, recently discovered Ter1 telomerase RNA^{19,20}, which was induced during meiosis (SPNCRNA.214; Supplementary Table 3). Expression of 14 non-coding RNAs was independently confirmed by RT-PCR (Supplementary Fig. 3). This analysis revealed bi-directional transcription across all tested regions, including the well-characterized *nmt1* gene, although most regions showed more transcripts from one strand. Given the ubiquitous transcription throughout the genome, the novel transcripts described here probably only hint at the true level of transcriptional complexity.

Sequence-read numbers across the newly identified transcribed regions were lower than numbers across annotated coding regions (Fig. 1b). Only 13 of the novel transcripts were evident from the tiling-chip data in proliferating cells, whereas another 79 were only substantially expressed under specific conditions, most notably during meiosis or quiescence (Fig. 2b, c and Supplementary Table 3). The antisense RNAs were particularly enriched for highly regulated transcripts, many of which peaked during the meiotic divisions (Fig. 2c).

To test whether some of the newly identified regions reflect cryptic transcripts that are degraded in the nucleus, we analysed RNA isolated from an *rrp6* mutant defective in nuclear exosome function^{21,22}; 36 of the novel transcripts were more highly expressed in this mutant such that they became evident also on tiling chips (Supplementary Table 3). These data raised the possibility that many newly identified regions are strongly transcribed but rapidly degraded by different surveillance systems²¹. To test this hypothesis, we globally measured Pol II occupancy (reflecting transcriptional activity¹²). Overall, Pol II occupancy across the new regions was comparable to the location of 10–20% of genes with the lowest levels of transcription (Fig. 2d). We conclude that most newly identified regions were not robustly expressed in proliferating cells, but that the sequencing approach was sufficiently sensitive to detect transcriptional traces below the detection limit of hybridization-based approaches.

The combined sequence and hybridization data provided a rich source to analyse transcript structures at maximal resolution. High densities of overlapping transcripts can confound the sequence data, and decreasing read-numbers towards the 5' ends, reflecting oligo(dT) priming (Figs 1b and 3a), render it difficult to determine accurately transcript lengths of long genes. The hybridization data are less affected by these issues because they distinguish transcriptional direction and do not show any 5' bias (Fig. 3a and Supplementary Fig. 4). Together, the two approaches provided complementary data on untranslated regions (UTRs) for most *S. pombe* genes (Supplementary Table 4). For many other genes, which were mostly expressed at low levels and did not pass our confidence cutoffs, the

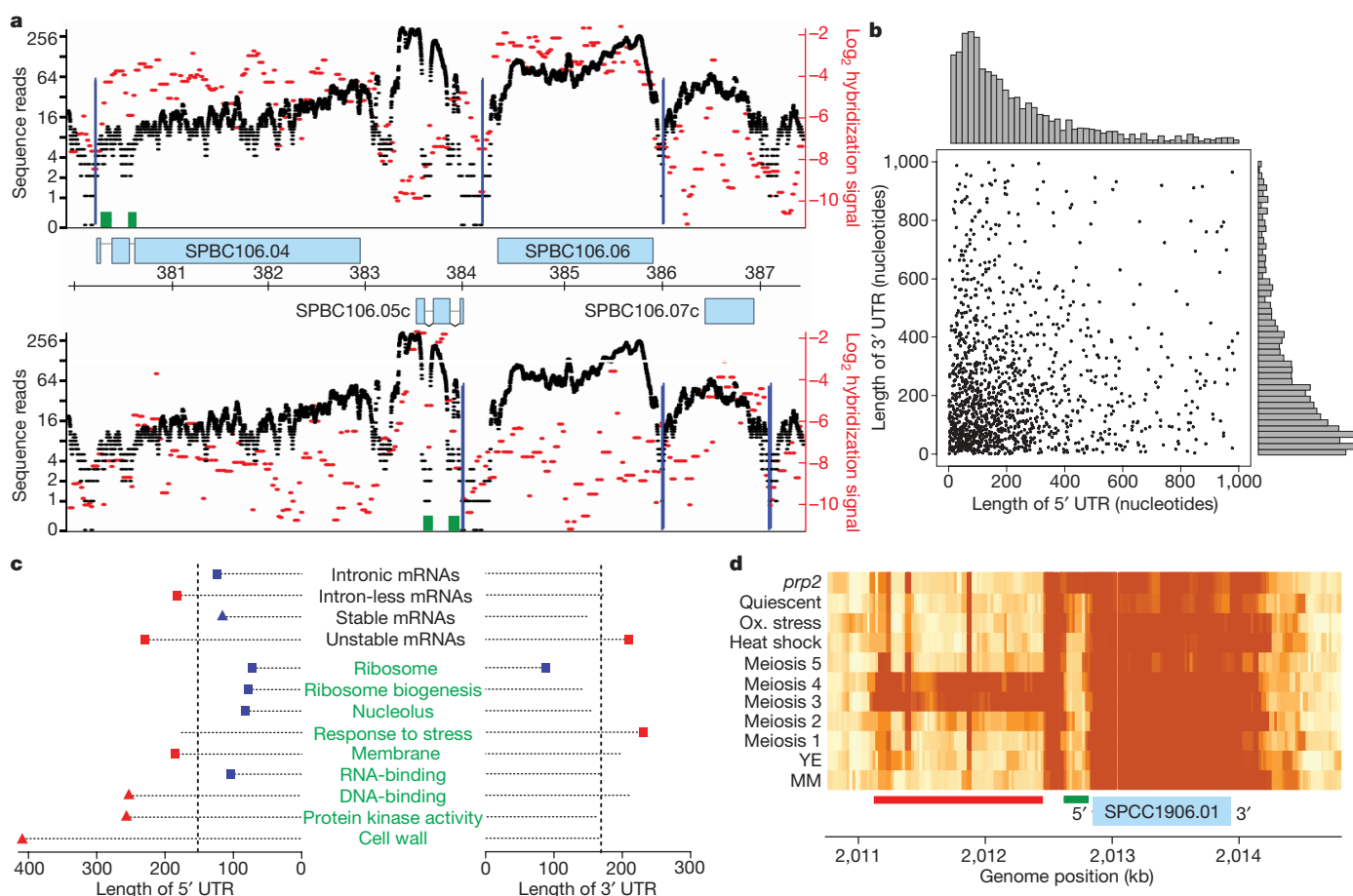


Figure 3 | Analyses of transcript structures. **a**, Plot depicting sequence reads and tiling-chip signals as in Fig. 2a. Vertical dark blue lines, transcription start and end sites determined by sequencing; green boxes, introns. **b**, Scatterplot and histograms showing length distributions of 3' and 5' UTRs based on tiling-chip data. **c**, Transcripts with significantly larger (red) or smaller (blue) UTRs for selected mRNA properties (top) or GO

categories (bottom, green), based on tiling-chip data (triangles) or on tiling-chip and sequence data (squares) (Wilcoxon test, $P < 0.05$, Hochberg-adjusted for multiple tests). Vertical dashed lines: median UTR lengths. **d**, Tiling-chip hybridization signals as in Fig. 2b, showing a novel intron that is not spliced in the *prp2* mutant (green bar) and an alternate 5' UTR (red bar).

UTRs could be mapped by visual inspection. UTRs determined by hybridization or sequencing showed good agreement with each other and also with the previously known UTRs (Supplementary Table 5). The median 5'- and 3'-UTR lengths determined by hybridization were 152 and 169 nucleotides, respectively, with a mean combined length of 465 nucleotides (Fig. 3b). Thus, the UTRs of fission yeast are substantially larger than those of budding yeast, which show a mean combined length of 211 nucleotides⁵.

We compared UTR-length distributions for different functional categories (Fig. 3c). The most stable transcripts¹² had short 5' UTRs, whereas the least stable transcripts had long 5' and 3' UTRs, which may contain regulatory signals for RNA turnover. An analysis of Gene Ontology (GO) categories with significantly longer or shorter UTRs (Fig. 3c) uncovered similarities to budding yeast⁵. For example, transcripts encoding protein kinases and membrane proteins had long 5' UTRs, whereas ribosome-biogenesis genes had short 5' UTRs in both yeasts, indicating that UTR-length distributions show some conservation in these distantly related yeasts.

Sampling UTR lengths under different conditions allowed detection of transcript-size regulation (Supplementary Table 4). Our data confirmed the known transcripts with alternate start sites or polyadenylation sites produced from *cig2* and *wos2*, respectively^{23,24}. Using a conservative approach, we identified 27 additional transcripts with alternate start sites during meiosis or stress (Figs 2b and 3d, and Supplementary Table 6). Alternate polyadenylation sites were more abundant, affecting ~187 transcripts (Supplementary Table 6). Transcription-termination sites were generally less well defined than start sites and also varied across different conditions (Fig. 3d and Supplementary Fig. 5).

The resolution of the tiling chips was limiting to analyse splicing owing to the small size of most introns (<100 nucleotides). The sequence data, however, provided unprecedented insights into splicing of the 45.4% intronic genes of *S. pombe*¹³. Both unspliced and spliced transcripts were present in the total RNA preparations;

accordingly, we also obtained reads covering introns, albeit at lower numbers than for exons (Figs 1b and 3a). Importantly, sequencing provided direct evidence for splicing owing to 'trans-reads' spanning exon-exon junctions, thus confirming ~93% of predicted introns and hugely reducing unsupported gene structures. We found no evidence for the existence of alternate splicing in *S. pombe*.

To estimate splicing efficiencies, we determined normalized numbers of sequence reads spanning exon-exon and corresponding exon-intron junctions for all introns (Supplementary Table 7). This calculation of splicing efficiency exploits relative read numbers and is therefore internally normalized for expression levels and sequencing depth. Median numbers of spliced transcripts were only ~2-fold higher than numbers of corresponding unspliced transcripts, suggesting a surprisingly large cellular portion of unprocessed mRNAs (Supplementary Table 7). Average splicing efficiency was similar for different intron positions within genes (Supplementary Fig. 6). Splicing efficiency strongly varied, however, among different genes and conditions. A conservative analysis uncovered 254 genes (314 introns) that were more efficiently spliced during meiotic differentiation than in proliferating cells (Supplementary Table 8). These genes included 9 of 12 known meiotically spliced genes²⁵, whereas the 3 remaining genes showed increased meiotic splicing below our cutoff. Such 'regulated' splicing was evident in all five differentiation stages tested, but was most prevalent during meiotic prophase and nuclear divisions (Fig. 4a). In some genes all introns showed regulated splicing, whereas in others only selected introns were regulated (Supplementary Table 8)—a finding that was robust to lowering the cutoff. The median proportion of introns per gene showing regulated splicing was 50%, and regulated splicing showed no preference for specific intron positions.

The surprisingly large, yet conservative, list of genes with increased meiotic splicing was highly enriched for genes showing increased transcript levels during meiosis²⁶ ($P \sim 2 \times 10^{-20}$, hypergeometric test). Coordinated increases of meiotic gene expression and splicing

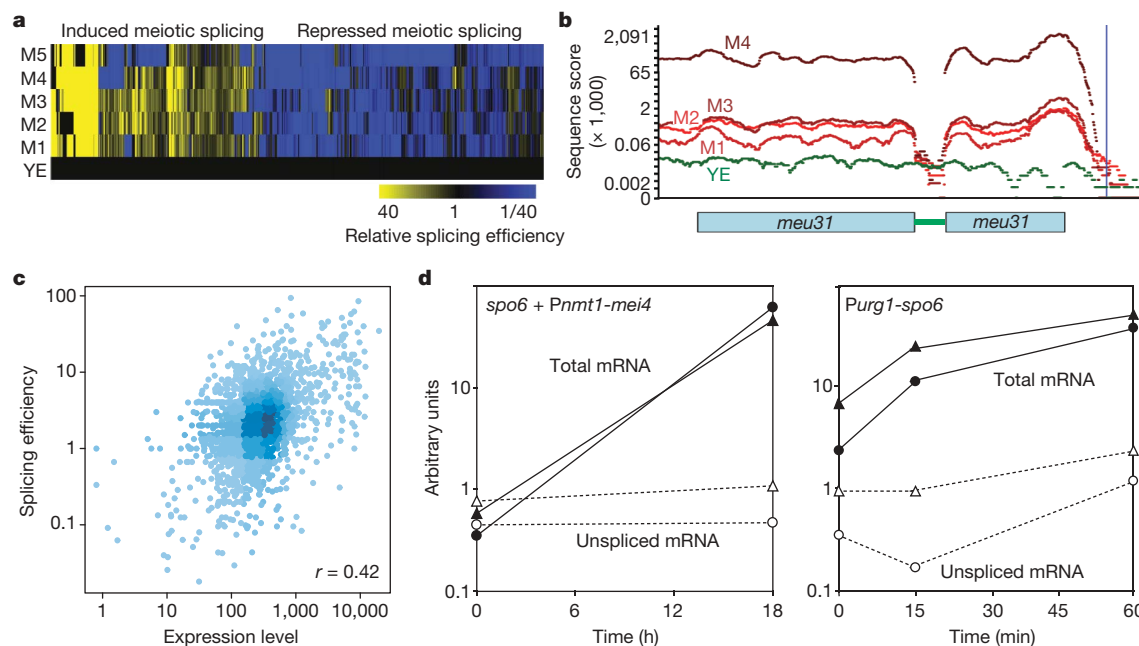


Figure 4 | Dynamics of splicing efficiency reflects transcription.

a, Hierarchical clustering of introns by their splicing efficiency in five stages of meiotic differentiation (M1 to M5) relative to their splicing efficiency during rapid proliferation (YE). **b**, Plot depicting log-scale numbers of sequence reads normalized for sequencing depth across *meu31* (intron depicted as green bar), colour-coded by experimental condition. Numbers of trans-reads across the exon-exon junction range from 0 (YE) to 490 (M3). **c**, Scatterplot comparing median splicing efficiency for intron-containing genes with mRNA levels based on expression-chip hybridization signals.

Shades of blue reflect the gene density, and Pearson correlation is shown at the bottom right. **d**, RT-PCR data to quantify splicing of *spo6* transcript as a function of transcription. Left: RNA levels before and 18 h after overexpression of Mei4 using the *nmt1* promoter (*Pnmt1*)²⁷; right: RNA levels before and up to 1 h after direct overexpression of *spo6* using the *urg1* promoter (*Purg1*)³⁰. Data from primers within exons (solid) or across exon-intron junctions (dashed) are shown for two different exons or junctions, respectively.

were also directly evident from the sequence data (Fig. 4b). Moreover, meiotic transcripts showed similar profiles for gene expression and splicing efficiency during meiosis (Supplementary Fig. 7). A reciprocal analysis uncovered 478 genes (559 introns) that were more efficiently spliced in proliferating cells than during meiosis (Fig. 4a and Supplementary Table 8). This list was enriched for genes highly expressed in proliferating cells¹⁶, including ribosomal-protein genes ($P < 2 \times 10^{-7}$, hypergeometric test). These data suggest that increased transcription can promote splicing. Indeed, splicing efficiency was significantly correlated with mRNA levels (Fig. 4c). Moreover, a functional analysis revealed widespread relationships between expression levels and splicing efficiency in proliferating cells (Supplementary Table 9). For example, highly expressed genes, such as those repressed during stress¹⁵, or conserved genes¹⁶ were more efficiently spliced than genes induced during stress or than *S. pombe*-specific genes.

To test directly whether increased transcription can lead to increased splicing, we activated transcription of the meiotically spliced *spo6* and *spn7* genes, either by placing them under the control of an ectopic regulatable promoter or by overexpressing the transcription factor Mei4, which activates *spo6* and *spn7* (ref. 27) and has been implicated in the regulation of meiotic splicing²⁸. The proportion of spliced transcripts increased after activating transcription, using either the ectopic or the native transcription factor (Fig. 4d; Supplementary Fig. 7). We conclude that activation of transcription itself is sufficient to promote splicing during meiosis, without the specific need for the meiotic factor Mei4. This finding raises the possibility that transcriptional and splicing efficiencies are mechanistically linked. Taken together, our results reveal a surprising genome-wide regulation of splicing, largely reflecting transcript levels during proliferation or differentiation. These data point to a global and condition-specific coupling between splicing efficiency and transcription, which may help to optimize and streamline gene expression programmes.

METHODS SUMMARY

Strains and experimental conditions are listed in Supplementary Table 1. cDNA for sequencing and array hybridization was prepared using oligo(dT) or random primers, respectively. For sequencing, fragment sizes of 120–170 bp were attached to the FlowCell at an average concentration of 3 pM, amplified isothermally, and sequenced using Solexa reversible-terminator chemistry on the Illumina Genome Analyser. Sequence reads were mapped to the reference genome using BLAT. Analyses of tiling-chip data were based on the Bioconductor package ‘tilingArray’²⁹.

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

Received 17 March; accepted 15 April 2008.

Published online 18 May 2008.

1. Yamada, K. *et al.* Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302**, 842–846 (2003).
2. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242–2246 (2004).
3. Stolc, V. *et al.* A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306**, 655–660 (2004).
4. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
5. David, L. *et al.* A high-resolution map of transcription in the yeast genome. *Proc. Natl Acad. Sci. USA* **103**, 5320–5325 (2006).
6. Li, L. *et al.* Genome-wide transcription analyses in rice using tiling microarrays. *Nature Genet.* **38**, 124–129 (2006).
7. The Encode Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).

8. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007).
9. Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. *Nature Rev. Genet.* **8**, 413–423 (2007).
10. Blencowe, B. J. Alternative splicing: new insights from global analyses. *Cell* **126**, 37–47 (2006).
11. Hughes, T. A. Regulation of gene expression by alternative untranslated regions. *Trends Genet.* **22**, 119–122 (2006).
12. Lackner, D. H. *et al.* A network of multiple regulatory layers shapes gene expression in fission yeast. *Mol. Cell* **26**, 145–155 (2007).
13. Wood, V. *et al.* The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**, 871–880 (2002).
14. Lyne, R. *et al.* Whole-genome microarrays of fission yeast: characteristics, accuracy, reproducibility, and processing of array data. *BMC Genomics* **4**, 27 (2003).
15. Chen, D. *et al.* Global transcriptional responses of fission yeast to environmental stress. *Mol. Biol. Cell* **14**, 214–229 (2003).
16. Mata, J. & Bähler, J. Correlations between gene expression and gene conservation in fission yeast. *Genome Res.* **13**, 2686–2690 (2003).
17. Struhl, K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Struct. Mol. Biol.* **14**, 103–105 (2007).
18. Potashkin, J., Li, R. & Frendewey, D. Pre-mRNA splicing mutants of *Schizosaccharomyces pombe*. *EMBO J.* **8**, 551–559 (1989).
19. Leonardi, J., Box, J. A., Bunch, J. T. & Baumann, P. TER1, the RNA subunit of fission yeast telomerase. *Nature Struct. Mol. Biol.* **15**, 26–33 (2008).
20. Webb, C. J. & Zakian, V. A. Identification and characterization of the *Schizosaccharomyces pombe* TER1 telomerase RNA. *Nat. Struct. Mol. Biol.* **15**, 34–42 (2008).
21. Bickel, K. S. & Morris, D. R. Silencing the transcriptome's dark matter: mechanisms for suppressing translation of intergenic transcripts. *Mol. Cell* **22**, 309–316 (2006).
22. Harigaya, Y. *et al.* Selective elimination of messenger RNA prevents an incidence of untimely meiosis. *Nature* **442**, 45–50 (2006).
23. Borgne, A., Murakami, H., Ayté, J. & Nurse, P. The G1/S cyclin Cig2p during meiosis in fission yeast. *Mol. Biol. Cell* **13**, 2080–2090 (2002).
24. Munoz, M. J., Daga, R. R., Garzon, A., Thode, G. & Jimenez, J. Poly(A) site choice during mRNA 3'-end formation in the *Schizosaccharomyces pombe* *wos2* gene. *Mol. Genet. Genomics* **267**, 792–796 (2002).
25. Averbeck, N., Sunder, S., Sample, N., Wise, J. A. & Leatherwood, J. Negative control contributes to an extensive program of meiotic splicing in fission yeast. *Mol. Cell* **18**, 491–498 (2005).
26. Mata, J., Lyne, R., Burns, G. & Bähler, J. The transcriptional program of meiosis and sporulation in fission yeast. *Nature Genet.* **32**, 143–147 (2002).
27. Mata, J., Wilbrey, A. & Bähler, J. Transcriptional regulatory network for sexual differentiation in fission yeast. *Genome Biol.* **8**, R217 (2007).
28. Malapeira, J. *et al.* A meiosis-specific cyclin regulated by splicing is required for proper progression through meiosis. *Mol. Cell. Biol.* **25**, 6330–6337 (2005).
29. Huber, W., Toedling, J. & Steinmetz, L. M. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* **22**, 1963–1970 (2006).
30. Watt, S. *et al.* *urg1*: a uracil-regulatable promoter system for fission yeast with short induction and repression times. *PLoS ONE* **3**, e1428 (2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Gould and M. Yamamoto for strains, W. Huber and R. Durbin for advice, and J. Mata, W. Huber, V. Pancaldi, D. Stemple, J.-R. Landry and D. Lackner for comments on the manuscript. B.T.W. was supported by Sanger Postdoctoral and Canadian NSERC fellowships, and S.M. by a fellowship for Advanced Researchers from the Swiss National Science Foundation. This research was funded by Cancer Research UK grant number C9546/A6517, by the Wellcome Trust, and by DIAMONDS, an EC FP6 Lifescihealth STREP (LSHB-CT-2004-512143).

Author Contributions B.T.W., S.M. and J.B. designed and supervised the research and discussed the results; S.W. performed most experiments with help of B.T.W. and S.M.; B.T.W. and S.M. analysed the data with help of F.S., V.W., C.J.P. and J.B.; I.G. and J.R. helped with sequencing; and J.B. drafted the manuscript.

Author Information Raw data are available from ArrayExpress (<http://www.ebi.ac.uk/arrayexpress>) under accession numbers E-MTAB-5 (sequence data) and E-MTAB-18 (array data). Transcript data-plots are available from our TranscriptomeViewer at http://www.sanger.ac.uk/PostGenomics/S_pombe/. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.B. (jurg@sanger.ac.uk).

METHODS

cDNA preparation for high-throughput sequencing. All cDNA samples for Illumina were prepared by first treating ~1 mg of total RNA for 30 min with amplification-grade RNase-free DNase (Invitrogen), according to the manufacturer's protocols. Poly(A)-enriched RNA was then prepared using an oligo(dT) selection kit (Oligotex Direct mRNA miniKit, Qiagen). The resulting RNA was converted to double-stranded cDNA using a cDNA synthesis kit (Superscript choice system for cDNA synthesis, Invitrogen), primed by an oligo(dT) primer. RNA samples from the pooled meiotic time points were subjected to amplification by *in vitro* transcription (IVT) after a poly(A)-enrichment step as described above.

DNA libraries were prepared following the manufacturer's instructions (Illumina). DNA was sheared by nebulization, followed by simultaneous end-repair and phosphorylation using T4 DNA polymerase, Klenow fragment of DNA polymerase I and T4 PNK. DNA recovery was performed after each stage using QIAquick PCR purification columns (Qiagen). These repaired fragments were 3'-adenylated using Klenow exonuclease-minus (Illumina) and were purified using a MinElute PCR purification column (Qiagen). Illumina adaptors were ligated to the adenylated ends of the fragments and gel-purified on a 2% TAE (Tris-acetate-EDTA)-agarose gel (Certified Low-Range Ultra Agarose, Biorad), stained using ethidium bromide and visualized on a Dark Reader (Clare Chemical). A range of fragment sizes (120–170 bp) was excised from the gel and extracted using a QIAquick gel extraction kit. Seventeen rounds of PCR amplification were performed using primers complementary to the previously ligated adaptors and compatible to oligonucleotides attached to the FlowCell. DNA was recovered using a QIAquick PCR purification column. DNA was subsequently diluted to a working concentration of 10 nM in TE (Tris-EDTA) after quantification on a Nanodrop-1000 spectrophotometer.

Sequencing data processing and expression scores. FASTQ files of sequence reads were converted into FASTA files, and were filtered to remove sequences <15 bp after trimming the sequence from the position of the first N. All remaining FASTA sequences were matched back to the *S. pombe* genome using BLAT (tilesize 8, oneoff 1) in parallel on the Sanger Institute computer farm. All FASTA reads were also matched back as above to a spliced genome with all known or predicted intron sequences removed. The result files of matches to the spliced and unspliced genomes were compiled into a complete and non-redundant set used for subsequent analysis.

For Fig. 1a, expression scores for every genomic base pair position were assigned on the basis of how many sequence reads covered each position. The log₂ of the score for each base pair position was then plotted using R/Bioconductor. The numbers of sequence reads drop towards the 5'-end of long genes. To ensure that expression scores are not biased against long genes, scores were determined first by taking the sum of the sequencing expression scores for only 300 bp at the 3'-end of each coding region, or for the entire length if the coding region was <300 bp, and then dividing by the corresponding length used.

Expression-chip hybridization and processing. Total RNA was isolated as described¹⁴, and 0.3 µg RNA were labelled using the standard Affymetrix Genechip eukaryotic hybridization protocols. Hybridizations were performed on Affymetrix Yeast 2.0 Genechip arrays. Scanning was performed on a Genechip Scanner 3000, and data extraction was carried out using Affymetrix GCOS 1.4 (Figs 1a and 4c).

Tiling-chip labelling, hybridization and normalization. Total RNA was isolated as described¹⁴. Labelling and hybridization to the Affymetrix GeneChip *S. pombe* Tiling 1.0FR arrays were performed as described⁵. Affymetrix CEL files were normalized using the 'normalizeByReference' function from the Bioconductor package 'tilingArray' (<http://www.bioconductor.org>)²⁹. In this procedure, the individual hybridization behaviour of every probe was corrected using the signal of three genomic DNA hybridizations. Genomic DNA was extracted, labelled and hybridized to the Affymetrix GeneChip *S. pombe* Tiling 1.0FR arrays as described⁵. A second normalization step was applied using the signals of intergenic probes as a reference. Finally, between-array normalization and variance-stabilizing transformation were applied using the Bioconductor package 'vsn'.

Pol II ChIP-chip analysis. Chromatin immunoprecipitation (ChIP) was performed as described¹² using an antibody specific for the Pol II C-terminal domain (4H8, Abcam). The immunoprecipitated material and input control were amplified in two steps as described³¹. During the second step, dUTPs were added to the PCR mix for subsequent fragmentation of the products. Fragmentation and labelling of the amplified products were performed using the GeneChip WT double-stranded DNA terminal labelling kit (Affymetrix). The duplicated immunoprecipitated samples and corresponding input material were hybridized on four separate Affymetrix GeneChip *S. pombe* Tiling 1.0FR

arrays. The log₂ signals of the probes on the input arrays were subtracted from the log₂ signals of the Pol II arrays. The two normalized Pol II data sets were averaged and smoothed using a five-probe moving average. Average gene profiles were created using R and Bioconductor.

Data visualization along genomic coordinates. The tiling-chip data were visualized using the 'plotAlongChrom' function⁵ (Figs 2b and 3d). The sequence data were visualized using an in-house R script (Figs 2a and 3a). Normalized sequence scores were generated by dividing the sequencing expression score for a given base pair position by the sum of the expression scores for this base pair position in each condition sequenced.

Novel transcript analysis using tiling-chip data. The normalized data were smoothed using a five-probe moving average. Signal breakpoints in the probe signals along genomic coordinates were then determined using a dynamic programming algorithm for finding a globally optimal fit of a piecewise constant expression profile along genomic coordinates²⁹. Segments ≥100 bp and a median probe signal higher than the 75th percentile of the chip and outside of any annotation were selected for visual analysis. To screen for anti-sense transcripts, similar criteria were applied except that the segments had to overlap annotated genes on the opposite strand (Supplementary Table 3).

Novel transcript analysis using sequence data. Stretches of contiguous expression in intergenic regions were identified after removing all UTRs (see below) from the intergenic search space. Novel transcribed regions were required to have a length of ≥70 bp and an average sequence-expression score of ≥5 reads per bp. All predicted novel transcripts were then visually validated to remove inaccurate UTRs before a final manual curation (Supplementary Table 3).

Expression profiling analysis of the novel transcripts. Expression profiles of the novel transcribed regions determined by sequencing and tiling chips were visually inspected from their expression across the 12 biological conditions tested (Supplementary Table 3). For the clustering analysis of Fig. 2c, a Wilcoxon rank sum test was used to determine if the probe signals in each new transcribed region were significantly greater than the signals of a reference set containing probes located outside of any annotated regions in any condition. An expression score was defined as $-\log_2$ of the *P*-value of this test.

UTR determination using tiling-chip data. CEL files were processed as for novel transcripts. The UTR boundaries were the closest breakpoint to the start of an annotated gene, where the median of the four probes immediately upstream of the breakpoint was lower than the one of the four probes downstream of the breakpoint. If no breakpoint could be defined that way and a breakpoint was present <50 bases inside the coding region, the UTR was set to 1. UTRs called inside neighbouring genes or sharing UTR boundaries with neighbouring genes were discarded. UTRs >1,000 nucleotides were discarded, because they were highly enriched in wrong calls based on visual inspection of the data (Supplementary Table 4).

UTR determination using sequence data. UTR lengths were determined by screening for a break in the transcribed region around genes, denoted by positions with sequence scores of 0 or 1, starting from either end of every gene. If a score of 0 was not found in the section between the start and/or end of the neighbouring regions, 1 was used as a cutoff. If no break was found using either cutoff, the UTR was denoted as undetermined (Supplementary Table 4).

Alternate 5'- and 3'-end analysis using tiling-chip data. Genes with UTRs containing several breakpoints caused by 'steps' in the decreasing probe signals moving away from the gene boundaries were automatically selected from 12 biological conditions. A Wilcoxon rank sum test was then used to determine if the probe signals in each region were significantly greater than the signals of a reference set containing probes located outside of any annotated regions in any condition. A score was defined as $-\log_2$ of the *P*-value of this test. Candidate regions with scores >10 in ≥12 conditions were selected for visual inspection (Supplementary Table 6).

Splicing analysis using sequence data. The initial BLAT results generated a set of sequence reads with gaps in the reference sequence (that is, representing potential spliced reads). Spurious matches within this data set caused by poly(A/T) tracts splitting reads between two distant regions in the genome were filtered out using a limit of ≤1 kb for the maximum sequence spanned by trans-reads. The remaining trans-reads were compared to all known and predicted introns for intron validation. Trans-reads that did not span any known introns were clustered on the basis of their splice junctions, where putative junctions had to overlap ±1 bp to belong to the same cluster. Clusters were ranked by the number of novel trans-reads in each cluster and a conservative set of 33,466 reads with ≥6 reads per cluster (defining 485 potential splice sites) were manually curated. 'False-positive' trans-read clusters that did not seem to reflect splicing were mostly within complex repeated regions, and some may reflect errors in the original genome sequence.

Regulated splicing was determined by calculating a ratio of reads that span exon-exon junctions (EE) to those that span the two corresponding exon-intron

junctions (2EI) (Supplementary Table 7). The latter were divided by two to normalize for relative frequency. To obtain a conservative estimate of regulated splicing, the EE:EI ratio for one condition had to be ≥ 5 -times greater than the EE:EI ratio of another stage. Junctions covered by < 2 sequence reads in any condition were not considered. Genes that were ≥ 5 -times higher spliced in any meiotic-differentiation stage (M1 to M5) compared to rapidly proliferating cells as well as those that were ≥ 5 -times higher spliced in rapidly proliferating cells compared to ≥ 1 meiotic-differentiation stage were determined (Fig. 4a). Additional analysis was also performed using absolute read numbers, in cases where ratios could not be calculated because of 0 values. In these cases, to obtain a conservative estimate of regulated splicing, where EE = 0 in rapidly proliferating cells, the EE in ≥ 1 meiotic-differentiation stage was required to be > 6 . With EE = 1 or 2 but EI = 0 in rapidly proliferating cells, the EE in ≥ 1 meiotic-differentiation stage was required to be 8 or 9, respectively. With EE ≥ 3 in rapidly proliferating cells, the EE in ≥ 1 meiotic-differentiation stage was required to be ≥ 5 -times higher than in rapidly proliferating cells, or ≥ 20 -times higher when identifying introns spliced more efficiently in rapidly proliferating cells to account for the greater sequence depth in this condition (Supplementary Table 8).

Measurement of splicing efficiency by quantitative RT-PCR. To test the relationship between transcription rate and splicing efficiency (Fig. 4d and Supplementary Fig. 7), the uracil-inducible *urg1* promoter was integrated upstream of *spo6* and *spn7* (ref. 30). Cells were grown in exponential phase for 16 h in minimal medium (MM) in the absence of uracil. A cell sample was then harvested, and uracil was added to the remaining culture at a final concentration of 2 mg ml^{-1} . Further cell samples were harvested 15 min and 60 min after uracil addition.

spo6 and *spn7* are putative targets of Mei4 and were induced in a strain over-expressing Mei4 under the control of the *nmt1* promoter²⁷. Such a strain (Supplementary Table 1) was grown in the presence of thiamine to early exponential phase. A cell sample was then harvested before the cells were diluted and was grown for 18 h in the absence of thiamine.

Primers were designed inside the exons 1 and 2 of *spo6* and across the exon 1/intron 1 and exon 2/intron 2 junctions. Similarly, primers were designed inside exons 1 and 4 of *spn7* and across the exon1/intron1 and intron 3/exon 4 junctions. RNA was extracted and qRT-PCR performed as described³⁰. The data were normalized to the signal of the *fbal* control gene. No signals above background levels were detected in control runs in the absence of reverse transcriptase.

Curator methods. Novel transcribed regions were converted to gff3 format and visualized in the context of the existing annotation using Artemis software and methods described previously¹³. The corresponding sequence plots were examined and discrete features designated 'non-coding RNAs'. Manual inspection of the strand-specific tiling-chip data identified several 'antisense' transcripts. 'Non-coding RNAs' were inspected for the presence of methionine-containing ORFs > 60 amino acids, identifying three protein-coding genes. Less discrete features, which may correspond to transcriptional noise, occurring mainly in low-complexity regions were designated 'miscellaneous features'. Some transcribed features were clearly related to their proximal genes and curated as 5' and 3' UTRs (occasionally intron-containing).

Sequence trans-reads obtained from proliferating cells validated 3,796 of the 4,811 known and predicted introns, and trans-reads only obtained from meiotic cells validated an additional 666 introns (Supplementary Table 7). The remaining 349 introns either were in poorly expressed genes with insufficient sequence reads, or were not spliced under any of the conditions tested. Among the latter, manual inspection coupled with homology searches and intron branch, acceptor and donor consensus-sequence data allowed refinement of 25 protein-coding gene structures, and deletion of 6 unsupported intron-containing genes. A number of the introns confirmed by trans-read sequences were not previously annotated in the database. These 'false negative' introns were mapped onto the genomic sequence and used to identify 22 new genes and revise a further ~ 60 gene structures.

All these alterations have been incorporated in *S. pombe* gene database (<http://www.genedb.org/genedb/pombe/>). The new transcribed regions are listed in Supplementary Table 3, and the corrected gene structures are listed at <http://www.genedb.org/genedb/pombe/coordChanges.jsp>.

31. Bernstein, B. E., Humphrey, E. L., Liu, C. L. & Schreiber, S. L. The use of chromatin immunoprecipitation assays in genome-wide analyses of histone modifications. *Methods Enzymol.* **376**, 349–360 (2004).