

# UCSF

## Recent Work

### Title

Analysis of a Splice Array Experiment Elucidates Roles of Chromatin Elongation Factor Spt4-5 in Splicing

### Permalink

<https://escholarship.org/uc/item/3xc486h9>

### Authors

Xiao, Yuanyuan  
Yang, Yee Hwa  
Burckin, Todd A  
et al.

### Publication Date

2005-03-01

# **Analysis of a Splice Array Experiment Elucidates Roles of Chromatin Elongation Factor Spt4-5 in Splicing**

Yuanyuan Xiao<sup>\*</sup>, Yee H Yang<sup>\*</sup>, Todd A Burckin, Lily Shiue, Grant A Hartzog and Mark R Segal<sup>†</sup>

Addresses:

Yuanyuan Xiao      Department of Epidemiology and Biostatistics, Center for Bioinformatics and Molecular Biostatistics, University of California, San Francisco, CA 94143, USA. E-mail: [yxiao@itsa.ucsf.edu](mailto:yxiao@itsa.ucsf.edu)

Yee H Yang      Department of Medicine, Center for Bioinformatics and Molecular Biostatistics, University of California, San Francisco, CA 94143, USA. E-mail: [jean@biostat.ucsf.edu](mailto:jean@biostat.ucsf.edu)

Todd A Burckin      Department of Molecular, Cell and Developmental Biology, University of California, Santa Cruz, CA 95064, USA. E-mail: [burckin@biology.ucsc.edu](mailto:burckin@biology.ucsc.edu)

Lily Shiue      Center for Biomolecular Science and Engineering, Department of Molecular, Cell and Developmental Biology, University of California, Santa Cruz, CA 95064, USA. E-mail: [lshiue@soe.ucsc.edu](mailto:lshiue@soe.ucsc.edu)

Grant A Hartzog      Department of Molecular, Cell and Developmental Biology, University of California, Santa Cruz, CA 95064, USA. E-mail: [hartzog@biology.ucsc.edu](mailto:hartzog@biology.ucsc.edu)

Mark R Segal      Department of Epidemiology and Biostatistics, Center for Bioinformatics and Molecular Biostatistics, University of California, San Francisco, CA 94143, USA. Email: [mark@biostat.ucsf.edu](mailto:mark@biostat.ucsf.edu)

<sup>\*</sup> These authors contributed equally to this work

<sup>†</sup> Corresponding author

## **Abstract**

**Background:** Splicing is an important for regulation of gene expression in eukaryotes, and it has important functional links to other steps of gene expression. Two examples of these linkages include Ceg1, a component of the mRNA capping enzyme, and the chromatin elongation factors Spt4-5, both of which have recently been shown to play a role in the normal splicing of several genes in the yeast, *S. cerevisiae*.

**Principal Findings:** Using a genomic approach to characterize the roles of Spt4-5 in splicing, we extended our observations of splicing defects in *ceg1*, *spt4* and *spt5* mutants to the entire collection of intron-containing genes, employing splicing-sensitive DNA microarrays. In the context of the complex experiment design, highlighted by 22 dye-swap array hybridizations comprised of both biological and technical replications, we applied four ANOVA mixed models and a semiparametric hierarchical mixture model. To refine selection of differentially expressed genes whose normal splicing depends on Ceg1 or Spt4-5, we used a more robust model-synthesizing statistic, *Differential Expression via Distance Synthesis* (DEDS), to integrate all five models. We further analyzed the list of differentially expressed genes and found that highly transcribed genes with long introns were most sensitive to the *spt* mutations.

**Conclusions:** In this paper, we showcased splicing array technology and developed methodologies for their analysis in the context of a real, complex experimental design. Our result suggests that the Spt4-Spt5 complex may help coordinate splicing with transcription under conditions that present kinetic challenges to spliceosome assembly or function.

## **Introduction**

Eukaryotic genes are fragmented into exons by intervening sequences (introns). After a gene is transcribed into pre-mRNA, the introns are removed from the transcript and the exons are joined by the spliceosome. This reaction, splicing, can also be used to create multiple transcripts from a single gene. For example, a particular exon may be included in one version of an mRNA, and skipped in another. This process of alternative splicing is subject to regulation in response to tissue, developmental and environmental cues [1]. In humans, most genes are subject to splicing and more than half are likely subject to alternative splicing, which is credited as the most important source for the extraordinary enrichment in complexity of the human proteome relative to the genome [1]. Accurate splicing is crucial for normal protein function; aberrant transcripts

due to splicing mutations are known causes for 15% of genetic diseases [1]. Therefore, elucidation of splicing mechanisms will not only help us understand the operating mechanisms underneath the functional complexity and diversity of higher eukaryotes, but also aid in new therapeutic strategies for treatments in splicing-related genetic disorders.

Although the different steps of gene expression are typically studied in isolation, it is clear that there are important functional links between them [2, 3]. For example, the process of capping the 5' end of pre-mRNAs is thought to influence both transcription and splicing [4, 5]. Furthermore, the rate of transcription elongation appears to influence splicing and alternative splice site choice [6, 7]. In addition, a number of pre-mRNA processing factors are recruited to transcripts via interaction with RNA polymerase II [2, 3]. Thus, a comprehensive understanding of mRNA synthesis will require an understanding between these functional linkages of steps in gene expression.

We have recently found evidence that the chromatin elongation factors, Spt4 and Spt5, play a role in RNA processing in the yeast, *S. cerevisiae*. Spt4 and Spt5 form a complex that regulates transcription elongation by RNA polymerase II. This complex is conserved across eukaryotes and has been proposed to both facilitate transcription by removing a nucleosomal barrier to transcript elongation and also suppress inappropriate transcription by reassembling nucleosomes behind transcribing polymerase [8]. The recent finding that Spt5 interacts physically and genetically with pre-mRNA capping factors suggests a role for Spt4-Spt5 in capping [9-12]. Because pre-mRNA capping is thought to increase the efficiency of splicing, we further analyzed splicing in *spt4* and *spt5* mutants and found that several genes were not spliced with normal efficiency [9]. Thus, the function of Spt4-Spt5 is linked to the processing of at least several mRNAs.

Traditionally, splicing is studied on an individual gene basis by *ad hoc* experiments. With the advent of eukaryotic genomic sequences, a global genomic view of splicing is rendered achievable and will provide an unprecedented amount of information regarding the mechanisms and regulation of splicing [13]. *S. cerevisiae*, a simple yeast that has been used as a model to study eukaryotic gene expression, presents a convincing entry point to embark on this task. The yeast genome is completely sequenced and well annotated, and the splicing machinery of yeast is well conserved with that of humans. Among the more than ~5,800 genes in the yeast genome, only about 250 of them possess introns and only a handful have multiple introns or are alternatively spliced [14]. However, these 250 intronic genes give rise to 27% of the transcripts

synthesized by the cell, an indication of the importance of splicing in yeast [15, 16]. Clark *et al.* [17] have recently designed a DNA microarray for analysis of splicing in yeast. To discriminate between spliced and unspliced transcripts, oligonucleotide probes on these arrays were designed to detect the splice junctions, introns and second exons of intron containing genes (Figure 1A). Splice junctions are found only in spliced transcripts whereas introns exist only in unspliced transcripts and splicing intermediates. Second exons are present in both spliced and unspliced transcripts and are good indicators of total transcript level. To detect these different classes of transcripts, the arrays were competitively hybridized with probes derived from control and experimental yeast strains.

Here we have extended our observations of splicing defects in *spt4* and *spt5* mutants to the entire collection of intron-containing genes, using these splicing-sensitive DNA microarrays [17]. As the goal of this study is to identify genes whose normal splicing depends on Spt4 or Spt5, the primary statistical task for this study is to select a set of genes that have altered expression as reported by the splice junction and intron probes. There is no lack of methodologies to tackle this so-called differential expression (DE) problem and we next give a brief summary of such methods. Fold change has been applied extensively to yield lists of genes that have altered expression under a certain threshold. Despite its methodological simplicity and intuitive appeal, the fold change method lacks a statistical framework (there is no accommodation of expression variation) and is biased toward selecting genes at low expression levels. Another class of frequently used methods treats the task of comparing expression levels in different biological states as a univariate testing problem, employing various corrections for test multiplicity [18]. Kerr *et al.* [19] propose using traditional analysis of variance (ANOVA) since these readily handle known sources of variation due to, for example, dye labeling, and sample or array replicates. By removing these effects from the estimation of the random error term, we achieve a reduction in this term and correspondingly sharper inferences. Wolfinger *et al.* [20] extends the ANOVA framework by treating some factors, for example, dyes and arrays, as random representatives of a large population (i.e. as random effects) resulting in a mixed model. There are several Bayesian alternatives to the above approaches [21-24] as well as some intermediary approaches that yield regularized *t* statistics [25, 26].

Our study employs a complex experiment design, featuring 22 dye-swap array hybridizations comprised of both biological and technical replications; see details in the next section. As elaborated in the next section, we initially analyzed these data with four ANOVA mixed models

and the semiparametric hierarchical mixture model (SHMM) of Newton *et al.* [27]. Instead of arbitrating between these models and picking a single model on which to base differential expression declarations, we exploit the fact the all five models attempt to estimate differential expression and employ a novel synthesizing scheme [28], *Differential Expression via Distance Synthesis* (DEDS), to derive a list of differentially expressed genes in *spt* mutants. This method compares favorably with the best individual statistics, while enjoying robustness properties lacked by the individual statistics[28]. Further analysis of such genes, whose splicing is altered in *spt* mutants, reveals common biochemical characteristics and attributes, which may provide new insights into the mechanisms of RNA processing and its connections to transcription.

## **Results**

### **Experimental design and data preprocessing**

In yeast, *SPT4* is a nonessential gene encoding a 102 amino-acid protein and *spt4Δ* (null) mutants display mutant phenotypes and genetic interactions consistent with an elongation defect [8]. *SPT5* encodes a large protein, and *spt5* mutations typically display mutant phenotypes and genetic interactions similar to those observed for *spt4* mutations, although they are often phenotypically more severe, consistent with the fact that *SPT5* is essential for life [8]. In this work, we have analyzed an *spt4* null mutation, and three partial loss-of-function mutations in *SPT5*. Two of these, *spt5-4* and *spt5-194*, encode versions of Spt5 that are defective for binding Spt4 (G.A.H., Jena Yamada and Thea Egelhofer, unpublished). The third allele, *spt5-242*, causes a cold-sensitive growth defect [29], and displays splicing and other defects at all temperatures (G.A.H and T.B., unpublished; [9]). The Spt5-242 protein still binds Spt4, even at the non-permissive temperature (G.A.H., Jena Yamada and Thea Egelhofer, unpublished data). In addition, we include analysis of *ceg1-250*, a temperature-sensitive mutation that causes rapid inactivation of the capping enzyme at the non-permissive temperature [5].

Two independent mRNA samples were prepared from each mutant, fluorescently labeled and then hybridized to the splicing arrays competitively with a probe derived from wild-type cells. Experiments were performed using a replicated dye-swap study design (Figure 2a) [30]. Briefly, there were four arrays (A1-A4) for each mutant vs. wild-type experiment. The first mRNA sample was hybridized to arrays A1 and A2 (Figure 2b) and the second was hybridized to A3 and A4. In A1 and A3, the mutant mRNA sample was labeled with Cy5 dye, and the wild-type sample was labeled with Cy3. The dye assignment was reversed for arrays A2 and A4. In

addition to these twenty mutant arrays (4 arrays X 5 mutants), there were two separate wild-type self-hybridization experiments, in which the wild-type was labeled with both Cy5 and Cy3. These self-hybridizations serve as technical replicates, i.e. as controls for variation in labeling and hybridization.

To provide a global view of splicing defects in the *ceg* and *spt* mutants, we plotted unnormalized log intensity values for signals from the two channels, mutant against wild-type, in Figure 3. Points representing individual array features are color coded so that exon, splice junction, intron and intronless gene features can be visually differentiated. Genes lying on the diagonal have a ratio close to 1, and their expression in the mutants is therefore largely unaffected. For *ceg1-250*, shown in the lower right panel, introns (light blue points) deviate noticeably from the diagonal toward the *ceg1-250* axis. This is a clear indication of intron accumulation in the *ceg1* mutant. Splice junctions (dark blue points) in *ceg1-250*, on the other hand, largely display ratios  $< 1$ , indicating a decrease in splice junction formation. Taken together, an accumulation of introns and loss of splice junctions in *ceg1-250* is indicative of a splicing defect. Compared with *ceg1-250*, the four *spt* mutants exhibit fewer alterations in splicing, with *spt5-194* most severely affected, in agreement with its phenotypic characteristics. A control plot from the wild-type self-hybridization is depicted in the upper left panel. As expected, no separation is observed in introns and splice junctions, and all points conform closely to the diagonal.

Boxplots of normalized ratios of splicing related probes stratified by mutants are shown in Figure 4. The general trend of the splice junction probe ratios shows a shift from the horizontal zero line in the negative direction signaling a decreased expression of splice junction in the mutants. The *ceg1-250* mutant showed the largest decrease and *spt5-194* was the most severely affected of the *spt* mutants. The boxplots of the exon probe ratios display a similar pattern of change—the expression of exon probes was also decreased in the mutants. This is consistent with the idea that the majority of the exon 2 probe signal for a transcript is derived from mRNA, which is stable and long-lived in comparison to pre-mRNA. It is of interest to investigate if the decrease of the splice junction probe and exon probe ratios is correlated. Figure 5 displays the scatter plots between ratios of these probes. The upper panel shows evident correlation between splice junction and exon ratios. In contrast to the exons and splice junctions, ratios of the intron probes do not show any shift from the horizontal zero line but spread for the mutants is nonetheless increased. Furthermore, there is no obvious correlation between the intron and exon ratios. In both plots, however, the spread of the cloud of points is mutant dependent and related to the

severity of splicing defects. From Figure 4, it is clear that several of the mutants tested, *ceg1-250*, *spt5-194* and *spt5-242*, cause strong decreases in exon and splice junction signals and more idiosyncratic, gene specific changes in intron signals. Do these changes reflect altered transcription, splicing, RNA decay, or a mixture of potential defects? To focus on alterations of splicing efficiency independent of changes in transcription, we used the Intron Accumulation and Splice Junction indices of Clark *et al.* [17], which normalize ratios of intron and splice junction signals to the ratios measured for the second exon. The splice junction (SJ) index is the change of the splice junction probe signal normalized by the change of overall gene expression level as measured by the related exon probe signal:  $SJ = \log \frac{SpliceJunction_{mut} / SpliceJunction_{wt}}{Exon_{mut} / Exon_{wt}}$ . Similarly, the intron accumulation (IA) index is obtained as the normalized change of the intron probe signals:  $IA = \log \frac{Intron_{mut} / Intron_{wt}}{Exon_{mut} / Exon_{wt}}$ . Relating changes in the splice junction and intron signals to changes in the second exon takes into account changes in overall expression level that may occur as a result of alterations in other steps of gene expression.

## Differential Expression Models

The experimental design of the splice mutant study motivated the use of four different mixed ANOVA models in addition to the SHMM (Table 1). These were separately applied to the two splicing indices. We briefly discuss the models below and the reader is referred to Appendix I for details in model specifics.

### ANOVA mixed models

The experimental design for *ceg1-250* and the four *spt* mutants is identical; we illustrate the set-up using *spt4Δ* as an example. Two independent RNA samples were prepared from an *spt4Δ* and wild-type strain. The *spt4Δ* and wild-type RNA samples were fluorescently labeled and competitively hybridized to two arrays with reversed order of labeling for the second hybridization to avoid labeling bias. This yielded four indices (for each of SJ and IA—which we treat identically and separately) for the detection of splicing defects in *spt4Δ* for each gene. A test, such as the one-sample *t* test, can be used to examine if the mean of the indices is equal to zero. For each gene, acceptance of this hypothesis signals lack of evidence for DE, while rejection provides evidence for DE. Alternatively, if the two indices of wild-type vs. wild-type resulting from the two wild-type self hybridization slides are included, DE can be tested by



comparing the four *spt4Δ* indices to the two wild-type indices, using a two-sample test. We applied both the one-sample and two-sample approaches in the analysis of the splicing experiment.

In addition to the above two approaches, distinguished by including wild-type self-hybridizations or not, we also considered another two approaches distinguished by allowing gene-specific variance heterogeneity or not. This latter case imposes the assumption that all genes exhibit a similar degree of variability and so can be jointly analyzed using a common estimate of error variance. As illustrated subsequently (see Table 2), this pooling dramatically increases error degrees of freedom ( $df$ ). The former approach, on the other hand, does not impose the common variance assumption, allowing different variances for different genes. The resulting model is then fitted gene by gene.

The above approaches, four in all—see Table 1 -- can be fitted by appropriately specified ANOVAs. Due to the nature of the experimental design (Figure 2b) -- array effect  $A$  is nested in sample effect  $S$  ( $S/A$ ), and sample effect  $S$  is in turn nested in mutant effect  $V$  ( $V/S$ ), we consider model terms involving  $S$  and  $A$  to be random. The remaining effects, including gene ( $G$ ), mutant ( $V$ ), and gene-mutant interactions ( $GV$ ), are fixed effects. Therefore, the four models are mixed-effect ANOVA models; see Appendix I for model fitting details.

### **SHMM model**

To complement the ANOVA approaches described above we also employed the SHMM advanced by Newton *et al.* [27]. This methodology was selected for several reasons. Firstly, the SHMM is nonparametric where there is sufficient information (lots of genes) and parametric where there is limited information (observations per gene), and this synthesis makes for an appropriately balanced strategy. Secondly, as is standard, our ANOVA approaches treat gene ( $G$ ), mutant ( $V$ ), and gene-mutant interactions ( $GV$ ) as fixed effects. Thus, there is no information sharing between genes. The SHMM achieves such sharing and does so in a more principled and flexible manner than some of the *ad hoc* approaches proposed that yield regularized  $t$ -statistics [25, 31, 32]. Thirdly, the output posterior probabilities for (directional) DE have dual utilities: (i) ranking (genes), and (ii) calibration (providing false discovery rates (FDRs)). Our interest is primarily in the former since, in the next section, we describe a method for combining several measures of DE and computing associated FDRs. The SHMM also has limitations, the foremost of which perhaps is the adequacy of the parametric assumptions. The extent of such assumptions

has been appreciably relaxed compared to the preceding fully parametric treatment of Kendzioriski *et al.* [24]. Importantly, diagnostic tools are provided for assumption checking. Other limitations are implementation related. Estimation is very computationally intensive—while this was not an issue for the splice data due to the small number of (intron containing) genes, we have encountered lengthy run times for more typically dimensioned array studies[28]. Additionally, the present implementation (available from <ftp://ftp.biostat.wisc.edu/pub/newton/Arrays/tr1074/Rcode/> ) only supports two group comparisons. Thus, there is some potential efficiency loss for the nested design employed in the splice study (Figure 2b). Details on the estimation methodology as well as extensive illustration of calibration, diagnostic, and performance aspects are provided in Newton *et al.* [27].

### Comparisons of differential expression models

Models with heteroscedastic errors accommodate gene-specific variances but typically, as here, replication is very limited and so the precision of the estimates is compromised. Models imposing homoscedastic errors yield precise estimates of the common error variance, and tests based on many  $df$ , since they permit combination over the large number of genes. However, the homoscedasticity assumption is both strong and difficult to evaluate. Differences in error  $df$  for the different models are presented in Table 2. Note that there are more than 5000  $df$  for error for the homoscedastic models and only about 20  $df$  for the heteroscedastic models.

We used results pertaining to SJ indices of the *cegl* mutant to illustrate relationships between the five models. Figure 6 displays a scatter plot matrix of  $-\log_{10}(p)$ , where  $p$  either corresponds to the Model I through IV  $p$ -value for tests of DE or to the Model V posterior probability for non-DE. Note that by relating Model I through IV results to Model V results we may seemingly be perpetuating the “severe pedagogical problem of misinterpreting p-values as posterior probabilities” [33]. However, this is not the case. At no stage do we make probabilistic statements in terms of these quantities. Rather, they simply constitute a quantification of DE. The high correlations between Model I and Model III, and between Model II and Model IV, with correlation coefficients 0.97 and 0.95 respectively, are as anticipated. This attests to the fact that the gene expression log ratio measurements in the two self-hybridization experiments of wild-type are tightly centered around zero. The fact that Model V conforms more closely to the homoscedastic models (I and III) than to the heteroscedastic models (II and IV) is not surprising, since the SHMM utilizes information sharing between genes, which is absent for the gene specific heteroscedastic models.

## Model synthesis and selection of DE genes

*Differential Expression via Distance Synthesis* (DEDS) is a novel method combining statistics or summaries that measure the same phenomenon [28]. We applied it here so as to refine selection of DE genes as furnished by the above five individual models. The simple underlying principle of DEDS is that genes that are highly ranked (as being differentially expressed) by all five models are more likely to be truly differentially expressed than genes that are high only for a single model. Capturing this requires devising a ranking that reflects the joint (across model) distribution of the individual (within model) gene ranks. This is achieved as follows. The individual measures of DE are concatenated into gene specific vectors that in turn are represented by points in the correspondingly dimensioned space (here 5). Note that the DE summaries so combined need to be commensurate *e.g.*, all statistics or all  $p$  values. A fixed “extreme” point ( $E$ ), corresponding to the coordinate-wise maxima or minima (whichever indicates DE), is included. The distances (*e.g.*, Euclidean, Mahalanobis) of all genes to  $E$  are computed, and those genes for which this distance is “significantly small” (calibrated by an appropriate null referent distribution) are considered as DE. The null referents are obtained analogously to those used in calibrating gap statistics [34]. Further details concerning DEDS are provided in [28], while an algorithm outline is sketched in Appendix II.

We have applied five models for the analysis of the splicing arrays; there are no clear advantages of one model over the others. Therefore, rather than trying to arbitrate between models and pick a single model on which to base DE declarations, or informally distilling sets of genes that are DE under two or more models, we employed DEDS as a robust means for synthesizing results. A comparison of ranking of DE genes by DEDS and individual measures is provided in [28]. The numbers of genes identified as differentially expressed by DEDS under FDR 0.01 and 0.05 for SJ as well as IA indices are listed in Table 3. The observation of greater numbers of genes identified as DE based on the intron accumulation index data than on the splice junction data reinforced the finding in Clark *et al.* [17] that IA indices are a more sensitive indicator for splicing defects. The splicing defect in the yeast capping enzyme mutant, *ceg1-250*, is catastrophic, whereas in the *spt4* and *spt5* mutants, fewer genes exhibit a splicing defect. Overall, *spt5-194* is the most severe splicing mutant among all *spt* mutants, with *spt4Δ* being the least impaired. The complete list of DE genes is provided in Table S1.

## Validation of DE genes

The identification of genes affected by *spt4* and *spt5* mutations using statistically robust methodology offers insight into the function of the Spt4-Spt5 complex, as well as the opportunity to better equate changes in Intron Accumulation with *bona fide* splicing defects. To validate our findings, we have used quantitative RT-PCR (QPCR) analysis to quantitatively examine 5 intron-containing genes, as well as two 2 unspliced genes, in all 5 mutants. We previously performed a qualitative analysis of three of these genes, *U3B*, *RPS25A* and *RPL26A*, and found that they were inefficiently spliced in *spt4* and *spt5* mutants [9]. By choosing primers that flank the intron-exon2 junction, we can specifically detect unspliced pre-mRNA (Fig. 1B). We also picked primers to detect either the second exon, or spliced mRNA (Fig. 1B). As with the microarrays, we can normalize changes in pre-mRNA levels to changes in spliced mRNA or total mRNA (i.e. exon 2).

As shown in Table 7, the results of the RT-PCR analysis generally agreed with the microarray analysis. Strikingly, in the four *spt* mutants, genes identified by DEDS showed an absolute increase in pre-mRNA levels, while in the *ceg1* mutant none of the pre-mRNAs showed an absolute increase as compared to wildtype. After normalizing the pre-mRNA signals to the spliced mRNA or second exon signals to account for potential changes in transcription or transcript stability, *ceg1* also showed a splicing defect as predicted by DEDS. Furthermore, the performance of DEDS was superior to the four ANOVA models and equivalent to the SHMM in terms of numbers of false positives/negatives over all 5 mutants (data not shown).

## Description and analysis of DE genes

There are likely multiple molecular mechanisms by which different genes were differentially expressed in the mutants discussed here. To account for some of these mechanisms, we subdivided the lists of DE genes with a  $q \leq 0.05$  (controlling FDR) before further analysis. First, we reasoned that positive and negative changes in IA likely occurred via different molecular mechanisms. Therefore, for each of the 5 mutants examined, the DE genes were divided into lists of genes with either positive or negative fold change. Second, because ribosomal protein genes represent a large fraction of all spliced genes in yeast [35], and because they are subject to a common mode of regulation [36], we further subdivided our lists of DE genes into sublists of ribosomal (RP) and non-ribosomal (non-RP) genes (Table 4). Finally, we focused upon the intron accumulation index as it is more sensitive to alterations in splicing [17].

For the *spt5* and *ceg1* mutants, a large majority of the DE genes encoded ribosomal proteins, whereas only ~40% of all intron containing genes encode ribosomal proteins (Table 4 and [35]). Furthermore, a number of translation and rRNA processing factors are among the non-RP genes found in our analysis, and it is possible that these genes are regulated by the same strategies as the RPs. Interestingly, for those DE genes with a negative fold change, *i.e.* those that were apparently spliced more efficiently, we found no RP genes. This suggests that the genes with a negative or positive fold change in the intron accumulation index have distinct dependencies upon Spt4-Spt5 and Ceg1.

We next asked if the genes identified in this analysis shared any particular attributes. It has previously been noted that introns in yeast display a bimodal distribution of sizes and positions within genes [35]. Ribosomal protein genes have large introns that occur relatively early in a pre-mRNA, whereas non-RP genes typically have smaller introns that occur somewhat later in the mRNA. Furthermore, RP genes are highly transcribed whereas non-RP genes tend to be less highly transcribed [16]. We therefore compared the transcription rates and size and positions of introns within the DE genes that displayed a positive fold change (Table 5). In the *ceg1* mutant, the set of DE genes had no unusual properties other than the non-RP DE genes being transcribed somewhat more frequently than the average non-RP gene. In the *spt* mutants, intron position of the DE genes was not significantly different from the average for RP and non-RP genes (Table 5). In contrast, in the *spt5-4* and *spt5-194* mutants, the non-RP DE genes shared attributes of RP genes: they tended to have longer introns and be more highly expressed than the typical non-RP gene. The non-RP DE genes in the *spt4Δ* and *spt5-242* mutants represent an intermediate case; their introns are not significantly longer than those of the typical non-RP intron-containing genes, but they are more highly transcribed.

The DE genes with a negative fold change appear to represent a distinct class of genes. First, they only encoded non-RPs. Second, they resembled the typical non-RP intron containing genes in that they had short introns, however, they were expressed at even lower levels than the typical non-RPs (Table 6), contrary to the DE genes with positive fold changes. Again, this is consistent with the idea that these genes were DE for reasons distinct from those leading to DE of genes with a positive fold change.

## **Discussion**

In this paper, we showcased splicing array technology and developed methodologies for its analysis in the context of a real, complex experimental design. We applied four ANOVA mixed models and a semiparametric hierarchical mixture model and used DEDS [28] to derive a list of DE genes. The DEDS algorithm synthesizes statistics or methods that estimate the same quantity of interest. The underlying principle behind DEDS is that genes that are highly ranked by different methods are more likely to be truly differentially expressed than genes that rank highly on a single measure. In this and previous work, we have evaluated DEDS on diverse datasets, featuring both one-channel Affymetrix oligonucleotide arrays and two-channel spotted arrays[28]. Using a set of spike-in (Affymetrix) datasets, where differentially expressed genes are known, we demonstrated that DEDS compares favorably with the best individual statistics, while enjoying robustness properties lacked by the individual statistics [28].

Previous to this study only four genes had molecularly analyzed for splicing defects in *spt4* and *spt5* mutants. Recently, we have used splicing-sensitive DNA microarrays to compare patterns of splicing defects across a diverse set mutations affecting gene expression [37]; but this and the previous study lacked a statistical or quantitative framework for rigorous determination of specific genes that were differentially expressed. Here we have used splicing-sensitive DNA microarrays combined with DEDS to analyze all known intron-containing gene in the yeast genome and to specifically identify those genes whose proper splicing is dependent upon *SPT4*, *SPT5* or *CEG1*. Comparison of the lists of DE genes for the five mutants examined here revealed that most of the genes that were DE in the *spt* mutants were also DE in the *ceg1* mutant (see Figure 7 and Table S1). The *spt5-242* mutant differed from the other *spt5* mutants in that it did not preferentially affect the splicing of non-RP genes with long introns. We do not understand the mechanistic basis for this observation, although it is consistent with our previous observations that this *spt5* mutation is phenotypically distinct from other *spt5* alleles and therefore may cause a distinct biochemical defect [29, 38]. Our data further suggest that Spt4's contribution to splicing is modest, as only a handful of genes were DE in the *spt4* mutant. This is consistent with the observation that, in contrast to *SPT5*, *SPT4* is not essential for life. Since there is currently no evidence that Spt4 functions independently of Spt5 [39], this suggests that Spt4 assists in, but is not essential for the functions of Spt4-Spt5 in splicing.

The fewer number of DE genes in the *spt* mutants compared to *ceg1-250* may indicate a lesser effect on splicing rather than an effect on a distinct subset of intron-containing genes. It is

interesting to note however that highly transcribed genes with long introns, i.e. RP genes and a subset of non-RP genes with long introns, were most sensitive to the *spt* mutations. These data suggest that the Spt4-Spt5 complex may play a particular role in coordinating splicing with transcription under conditions that present kinetic challenges to the spliceosome or its assembly, i.e. when splice sites are widely separated, increasing the separation in time and space between the synthesis of the 5' and 3' splice sites, or when a gene is highly transcribed, creating the need for rapid and repeated assembly of spliceosomes over one site on a gene. In addition, these data are consistent with recent evidence demonstrating an effect of RNA polymerase II elongation rates on alternative splicing in higher eukaryotes [40]. In contrast, the non-RP genes spliced more efficiently in the *spt* mutants tend to be transcribed less frequently than the average non-RP gene (Table 6). Thus, as is the case for transcription, the Spt4-Spt5 complex may have both positive and negative effects on splicing [8]. Furthermore, this is consistent with previous observations that altered transcription elongation may lead to increased splicing, presumably due to increased opportunities for recognition of suboptimal splice sites [6, 7]. Whether the effects we have measured here are due to altered elongation rates or indicate a more direct role of Spt4-Spt5 in splicing is currently under investigation.

## **Materials and Methods**

### **Sample preparation and array hybridization**

All yeast strains (Table 7) used were isogenic to S288C and Gal<sup>+</sup> [41]. Yeast were grown overnight in rich medium (YPD) at 30°C to early log phase ( $>1 \times 10^7$  cells/ml), spun down and resuspended in pre-warmed 39°C media, and allowed to grow at 39°C for 45 minutes after shift to restrictive temperature. Cells were collected by centrifugation at room temperature for 4 minutes, washed once with sterile water, flash frozen in liquid nitrogen and stored at -80°C. Total RNA was isolated by a hot phenol method [42] and quantitated by UV absorbance. Fluorescently labeled probe preparation, hybridization and data acquisition were performed as previously described [17] using 15 ug of total RNA/sample. For each mutant, RNA was prepared from two independently grown cultures. Each RNA sample was used to probe two arrays, and was labeled with Cy3 for the first array and Cy5 for the second.

### **Data normalization and preprocessing**

To effectively and properly normalize the data, we used nonlinear *loess* normalization [43] based on the subset of intronless genes. After normalization, for each array, the four replicates of each

splice junction, intron and exon probes are summarized using averages. This is followed by the calculation of SJ and IA indices.

## ANOVA Mixed Models

### 1. Model specificities

#### Model I – one-sample / homoscedastic errors

Let  $Y_{gvs_a}$  be the splicing related index, SJ or IA, from gene  $g$  ( $g = 1, 2, \dots, 254$  for SJ and  $1, 2, \dots, 263$  for IA), mutant  $v$  ( $v = 1, 2, \dots, 5$ ), sample  $s$  ( $s = 1, 2$ ) and array  $a$  ( $a = 1, 2$ ; corresponding to the dye swap pair). The first model can be represented as

$$Y_{gvs_a} = \mu + G_g + V_v + (GV)_{gv} + (V/S)_{vs} + (V/S/A)_{vsa} + (GV/S)_{gvs} + \varepsilon_{gvs_a}$$

Effects  $(V/S)_s$ ,  $(V/S/A)_a$ ,  $(GV/S)_{gvs}$  and  $\varepsilon_{gvs_a}$  are assumed to be normally distributed normal variables with zero means and variance components  $\sigma_{V/S}^2$ ,  $\sigma_{V/S/A}^2$ ,  $\sigma_{GV/S}^2$  and  $\sigma^2$  respectively. The remaining effects in the model are fixed effects. The parameter of interest in this model is  $\mu_{gv} = \mu + G_g + V_v + GV_{gv}$ , which measures the mean of the SJ/IA indices of gene  $g$  in mutant  $v$ . The following null hypothesis therefore defines the absence of differential expression in mutant  $v$  and gene  $g$ :

$$H_0 : \mu_{gv} = 0$$

The variance of the treatment mean  $\hat{\mu}_{gv}$  can be computed by the following equation:

$$\hat{Var}(\hat{\mu}_{gv}) = \frac{1}{n_S} \hat{\sigma}_{V/S}^2 + \frac{1}{n_S} \hat{\sigma}_{GV/S}^2 + \frac{1}{n_A n_S} \hat{\sigma}_{V/S/A}^2 + \frac{1}{n_S n_A} \hat{\sigma}^2,$$

where  $n_S = 2$  and  $n_A = 2$ .

#### Models II – one-sample / heteroscedastic errors

Model II (one sample / homoscedastic errors) is different from Model I by assuming that each gene has its own error distribution, so the model is fitted gene by gene. It can be represented by the following equation:

$$Y_{gvs_a} = \mu_g + V_v + (V/S)_{vs} + \varepsilon_{gvs_a}$$

The parameter of interest in this model is  $\mu_{gv} = \mu_g + V_v$  which measures the mean of the SJ/IA indices of gene  $g$  in mutant  $v$ . The following null hypothesis defines the absence of differential expression in mutant  $v$  and gene  $g$ :



$$H_0 : \mu_{gv} = 0$$

The variance of the treatment mean  $\hat{\mu}_{gv}$  can be computed by the following equation:

$$\widehat{Var}(\hat{\mu}_{gv}) = \frac{1}{n_S} \hat{\sigma}_{V/S}^2 + \frac{1}{n_S n_A} \hat{\sigma}^2 = \frac{1}{2} \hat{\sigma}_{V/S}^2 + \frac{1}{4} \hat{\sigma}^2$$

### Models III – two-sample / homoscedastic errors

Models III differs from Model I by including the indices derived from the two wild-type self-hybridizations. Because of this inclusion, the study design is rendered unbalanced. To be more specific, the arrays in the two wild-type self-hybridizations came from the same sample, whereas the samples of four slides related to a mutant were from two distinct samples (see Figure 2b). The model can be represented by the following equation:

$$Y_{gvs_a} = \mu + G_g + V_v + (GV)_{gv} + (V/S)_{vs} + (V/S/A)_{vsa} + (GV/S)_{gvs} + \epsilon_{gvs_a}$$

The parameter of interest in this model is  $\mu_{gv} = \mu + G_g + V_v + GV_{gv}$ , which measures the mean of the SJ/IA indices of gene  $g$  in mutant  $v$ . The following null hypothesis defines the absence of differential expression in mutant  $v_m$  and gene  $g$  compared to the wild-type:

$$H_0 : D_{gv_m} = \mu_{gv_m} - \mu_{gv_w} = 0$$

The variance of the treatment mean  $\hat{\mu}_{gv}$  can be computed by the following equation:

$$\widehat{Var}(\hat{\mu}_{gv}) = \frac{1}{n_S} \hat{\sigma}_{V/S}^2 + \frac{1}{n_S} \hat{\sigma}_{GV/S}^2 + \frac{1}{n_A n_S} \hat{\sigma}_{V/S/A}^2 + \frac{1}{n_S n_A} \hat{\sigma}^2, \text{ where } n_S = 2 \text{ for mutants and } n_S = 1$$

for the wild-type.

### Models IV – two-sample / heteroscedastic errors

Models IV differs from Model II by including the indices derived from the two wild-type self-hybridizations. The model can be represented by the following equation:

$$Y_{gvs_a} = \mu_g + V_v + (V/S)_s + \epsilon_{gvs_a}$$

The parameter of interest in this model is  $\mu_{gv} = \mu_g + V_v$ , which measures the mean of the SJ/IA indices of gene  $g$  in mutant  $v$ . The following null hypothesis defines the absence of differential expression in mutant  $v_m$  and gene  $g$  compared to the wild-type:

$$H_0 : D_{g^{v_m}} = \mu_{g^{v_m}} - \mu_{g^{v_w}} = 0$$

The variance of the treatment mean  $\hat{\mu}_{g^v}$  can be computed by the following equation:

$$\widehat{Var}(\hat{\mu}_{g^v}) = \frac{1}{n_S} \hat{\sigma}_{V/S}^2 + \frac{1}{n_S n_A} \hat{\sigma}^2 = \frac{1}{2} \hat{\sigma}_{V/S}^2 + \frac{1}{4} \hat{\sigma}^2, \text{ where } n_S = 2 \text{ for mutants and } n_S = 1 \text{ for the}$$

wild-type.

## 2. Derivation of variance components

Component	Estimate	Results	
		SJ	IA
$\sigma^2$	$MS_E$	0.186	0.28
$\sigma_{GV/S}^2$	$(MS_{GV/S} - \hat{\sigma}^2)/n_A \quad (n_A = 2)$	0	0.036
$\sigma_{V/S/A}^2$	$(MS_{V/S/A} - \hat{\sigma}^2)/n_G \quad (n_G = 254 \text{ for SJ and } 263 \text{ for IA})$	0.056	0.054
$\sigma_{V/S}^2$	$(MS_{V/S} - \hat{\sigma}^2 - n_G \hat{\sigma}_{V/S/A}^2)/n_G n_A$	0.013	0.042

## 3. Application of DEDS

The procures of the application of DEDS are as follows:

1. Fit the five DE models, and assume the resulting  $p$  values for gene  $i$  and model  $j$  are  $p_{ij}$  ( $i = 1, 2, \dots, n, j = 1, 2, \dots, 5$ ) in data matrix  $P$ ;
2. Locate the most extreme point  $E$  as a vector of zeros of length five;
3. Calculate distance  $d_i$  of all genes to  $E$  and order  $d_{(1)} \leq d_{(2)} \leq \dots \leq d_{(n)}$ ;

$$d_i = \sqrt{(p_{i1} - E_1)^2 + (p_{i2} - E_2)^2 + \dots + (p_{i5} - E_5)^2}$$

4. Generate  $B$  sets of reference distribution by:
  - 1) Center the columns of  $P$  at mean 0;
  - 2) Compute the singular value decomposition  $P = UDV^T$ ;
  - 3) Calculate  $P^* = PV$ ;
  - 4) Create  $Z^*$  by drawing uniform distribution over the range of the columns of  $P^*$ ;
  - 5) Back transform  $Z^*$  by  $Z = Z^*V^T$  to obtain the reference data  $Z$ .
- 6) For each reference dataset  $b$ ,  $d_i$  values are calculated and ordered in the way of

$$d_{(1)}^{(b)} \leq d_{(2)}^{(b)} \leq \dots \leq d_{(n)}^{(b)};$$

5. For a typical gene  $i$ , compute the median number of falsely called genes by computing the median number of values among each of the  $B$  sets of  $d_{(i)}^{(b)}$  that are smaller than  $d_{(i)}$ ; and the  $q$ -value (controlling False Discovery Rate) of gene  $i$  is computed as the median of the number of falsely called genes divided by the number of genes called significant.

## Analysis of DE genes

Gene annotations were obtained from the Ares lab intron database ([http://www.cse.ucsc.edu/research/compbio/yeast\\_introns.html](http://www.cse.ucsc.edu/research/compbio/yeast_introns.html)), and transcription frequency data was obtained from the Young lab (<http://web.wi.mit.edu/young/expression/transcriptome.html>). The collection of all intron-containing genes was divided into sets of RP and non-RP genes and averages and standard deviations were calculated for their transcription frequencies, intron lengths and intron start sites. Several genes were omitted from these analyses because there is no good data concerning their transcription frequency or intron position or size. In addition, Mtr2, which has multiple, overlapping introns, was considered to have a single intron for this analysis (see Table S1). To determine if the properties of DE genes in a mutant were significantly different from those of all RP or non-RP intron containing genes, we used a nonparametric resampling method. Briefly, a referent null distribution was generated by first taking 10,000 random samples of size  $N$  from the sets of all intron containing RP or non-RP genes ( $N$  is the number of DE RP or non-RP genes for a particular mutant), and then calculating the averages of each sample. The  $p$  value was derived as the percentage within the referent distribution that is more extreme than the observed property.

## Quantitative PCR analysis

cDNA synthesis for quantitative PCR (QPCR) was performed as described for fluorescently labeled target synthesis except that equal concentrations of all four deoxyribonucleotides and no Cy dyes were used. Reactions lacking reverse transcriptase were performed to control for genomic DNA contamination. Amplifications were conducted in a Bio-Rad iCycler using iQ SYBR Green Supermix (Bio-Rad, Hercules CA) and 200uM primer according to the manufacturer's instructions, using the oligonucleotide primers found in Table S2. Representative transcripts were assayed in triplicate. To compare the QPCR with array values we normalized QPCR values to the *OSH3* mRNA. *OSH3* was chosen as a suitable reference gene, since the array data indicated that its expression was unchanged in the five mutants used in the comparison.

## Acknowledgements

This work was supported by grants to G.A.H. from the NIH (GM60479) and the University of California Cancer Research Coordinating Committee. L.S was supported by a grant from the Packard Foundation. We thank Manny Ares for many stimulating discussions related to this work.

## References

1. Maniatis T, Tasic B (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* 418:236-243.
2. Maniatis T, Reed R (2002) An extensive network of coupling among gene expression machines. *Nature* 416:499-506.
3. Proudfoot NJ, Furger A, Dye MJ (2002) Integrating mRNA processing with transcription. *Cell* 108:501-512.
4. Schwer B, Shuman S (1996) Conditional inactivation of mRNA capping enzyme affects yeast pre-mRNA splicing in vivo. *RNA* 2:574-583.
5. Fresco LD, Buratowski S (1996) Conditional mutants of the yeast mRNA capping enzyme show that the cap enhances, but is not required for, mRNA splicing. *RNA* 2:584-596.
6. Howe KJ, Kane CM, Ares M, Jr. (2003) Perturbation of transcription elongation influences the fidelity of internal exon inclusion in *Saccharomyces cerevisiae*. *RNA* 9:993-1006.
7. de la Mata M, Alonso CR, Kadener S, Fededa JP, Blaustein M, Pelisch F, Cramer P, Bentley D, Kornblihtt AR (2003) A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell* 12:525-532.
8. Hartzog GA, Speer JL, Lindstrom DL (2002) Transcript elongation on a nucleoprotein template. *Biochim Biophys Acta* 1577:276-286.
9. Lindstrom DL, Squazzo SL, Muster N, Burckin TA, Wachter KC, Emigh CA, McCleery JA, Yates JR, 3rd, Hartzog GA (2003) Dual roles for Spt5 in pre-mRNA processing and transcription elongation revealed by identification of Spt5-associated proteins. *Mol Cell Biol* 23:1368-1378.
10. Pei Y, Shuman S (2002) Interactions between fission yeast mRNA capping enzymes and elongation factor Spt5. *J Biol Chem* 277:19639-19648.
11. Wen Y, Shatkin AJ (1999) Transcription elongation factor hSPT5 stimulates mRNA capping. *Genes Dev* 13:1774-1779.
12. Mandal SS, Chu C, Wada T, Handa H, Shatkin AJ, Reinberg D (2004) Functional interactions of RNA-capping enzyme with factors that positively and negatively regulate promoter escape by RNA polymerase II. *Proc Natl Acad Sci U S A* 101:7572-7577.
13. Modrek B, Lee C (2002) A genomic view of alternative splicing. *Nat Genet* 30:13-19.
14. Barrass JD, Beggs JD (2003) Splicing goes global. *Trends Genet* 19:295-298.
15. Lopez PJ, Seraphin B (1999) Genomic-scale quantitative analysis of yeast pre-mRNA splicing: implications for splice-site recognition. *RNA* 5:1135-1137.
16. Ares M, Jr., Grate L, Pauling MH (1999) A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA* 5:1138-1139.

17. Clark TA, Sugnet CW, Ares M, Jr. (2002) Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* 296:907-910.
18. Dudoit S, Yang YH, Callow MJ, Speed TP (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 12:111-139.
19. Kerr MK, Martin M, Churchill GA (2000) Analysis of variance for gene expression microarray data. *J Comput Biol* 7:819-837.
20. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol* 8:625-637.
21. Efron E, Tibshirani R, Storey J, Tusher VG (2001) Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 96:1151-1160.
22. Lee ML, Lu W, Whitmore GA, Beier D (2002) Models for microarray gene expression data. *J Biopharm Stat* 12:1-19.
23. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comput Biol* 8:37-52.
24. Kendziorski CM, Newton MA, Lan H, Gould MN (2003) On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat Med* 22:3899-3914.
25. Lönnstedt I, Speed TP (2001) Replicated microarray data. *Statistica Sinica* 12:31-46.
26. Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3:Article 3.
27. Newton MA, Noueiry A, Sarkar D, Ahlquist P (2004) Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 5:155-176.
28. Yang YH, Xiao Y, Segal MR (2004) Identifying differentially expressed genes from microarray experiments via statistic synthesis. *Bioinformatics* <http://bioinformatics.oupjournals.org/cgi/reprint/bti108>.
29. Hartzog GA, Wada T, Handa H, Winston F (1998) Evidence that Spt4, Spt5, and Spt6 control transcription elongation by RNA polymerase II in *Saccharomyces cerevisiae*. *Genes Dev* 12:357-369.
30. Chen Y, Dougherty ER, Bittner ML (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* 2:364-374.
31. Baldi P, Long AD (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17:509-519.
32. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98:5116-5121.
33. Berger JO, Boukai B, Wang Y (1997) Unified frequentist Bayesian testing - Rejoinder. *Statistical Science* 12:156-160.
34. Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 63:411-423.
35. Spingola M, Grate L, Haussler D, Ares M, Jr. (1999) Genome-wide bioinformatic and molecular analysis of introns in *Saccharomyces cerevisiae*. *RNA* 5:221-234.
36. Wade JT, Hall DB, Struhl K (2004) The transcription factor Ifh1 is a key regulator of yeast ribosomal protein genes. *Nature* 432:1054-1058.
37. Burekin TA, Nagel R, Mandel-Gutfreund Y, Shiue L, Clark TA, Chong J-L, Chang T-H, Squazzo SL, Hartzog GA, Ares M, Jr. (2005) Exploring functional relationships between

- components of the transcription, splicing and mRNA export machinery by gene expression phenotype analysis. *Nat Struct Mol Biol* in press.
38. Lindstrom DL, Hartzog GA (2001) Genetic interactions of Spt4-Spt5 and TFIIS with the RNA polymerase II CTD and CTD modifying enzymes in *Saccharomyces cerevisiae*. *Genetics* 159:487-497.
  39. Kim DK, InuKai N, Yamada T, Furuya A, Sato H, Yamaguchi Y, Wada T, Handa H (2003) Structure-function analysis of human Spt4: evidence that hSpt4 and hSpt5 exert their roles in transcriptional elongation as parts of the DSIF complex. *Genes Cells* 8:371-378.
  40. Noguez G, Kadener S, Cramer P, de la Mata M, Fededa JP, Blaustein M, Srebrow A, Kornblihtt AR (2003) Control of alternative pre-mRNA splicing by RNA Pol II elongation: faster is not always better. *IUBMB Life* 55:235-241.
  41. Winston F, Dollard C, Ricupero-Hovasse SL (1995) Construction of a set of convenient *Saccharomyces cerevisiae* strains that are isogenic to S288C. *Yeast* 11:53-55.
  42. Zavanelli MI, Ares M, Jr. (1991) Efficient association of U2 snRNPs with pre-mRNA requires an essential U2 RNA structural element. *Genes Dev* 5:2521-2533.
  43. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 30:e15.

## Figure legends

### Figure 1: Splicing array probe and RT-PCR primer design.

a) Probe design of the splicing array. There are three oligonucleotide probes for each intron-containing gene: intron (red), splice-junction (blue) and exon (green). In addition, there are also about 800 probes for intronless genes (yellow). This figure is modified from Clark *et al.*[17]. b) Primer design of RT-PCR. Primers are chosen to flank the intron-exon2 junction and the second exon or spliced mRNA .

### Figure 2: Graphical representation of designs.

a) In this representation, *vertices* correspond to target mRNA samples and *edges* to hybridizations between two samples. By convention, we place the green-labeled sample at the tail and the red-labeled sample at the head of the arrow. b) Nested design of the experiment. The effect *A* is nested in *S* and *S* is in turn nested in *V*. Note that there are two samples (*S*) for each mutant, but only one sample for the wild-type.

### Figure 3: Scatter plots of the logarithm intensities of splicing-related probes.

Points are colored coded as indicated.

**Figure 4: Boxplots of normalized ratios of splicing related probes stratified by mutants.**

Splice-junction and exon probe ratios show a shift from the horizontal zero line in the negative direction, whereas intron probe ratios are centered at zero.

**Figure 5: Scatter plots of normalized ratios of splicing related probes.**

Points are colored coded by their mutant identity. Gray horizontal and vertical reference lines indicate zero expression ratios.

**Figure 6: Scatter plot matrix of DE models for SJ indices of the *ceg* mutant.**

Plotted are the  $-\log_{10}(p)$  of the corresponding models, where  $p$  either corresponds to the Model I through IV  $p$ -value for tests of DE or to the Model V posterior probability for non-DE.

Correlation coefficients between corresponding models are shown in the lower triangle of the matrix.

**Figure 7: Venn diagram of DE genes from different mutants**

Panel a) compares DE genes among the three *spt5* mutants (*spt5-194*, *spt5-4* and *spt5-242*).

Statistical test shows that the common 43 genes are highly significant with a  $p$  value  $< 0.001$ . In panel b), *spt5* refers to the 43 common genes among all *spt5* mutants. The overlaps between *spt5* and *ceg1-250* (40,  $p < 0.001$ ), *spt5* and *spt4* (8,  $p < 0.001$ ), *spt4*, *spt5* and *ceg1-250* (7,  $p < 0.001$ ) are all significant.

**Tables**

**Table 1: A summary of the five competing DE models.**

Model No	Model Description
I	Mixed ANOVA: one-sample / homoscedastic
II	Mixed ANOVA: one-sample / heteroscedastic
III	Mixed ANOVA: two-sample / homoscedastic
IV	Mixed ANOVA: two-sample / heteroscedastic
V	Semiparametric hierarchical mixture model

**Table 2: Degrees of freedom table for the ANOVA mixed models.**

Source	Models
--------	--------

	One-sample Homoscedastic	One-sample Heteroscedastic	two-sample Homoscedastic	two-sample Heteroscedastic
Intercept	1	1	1	1
G	253		253	
V	4	4	5	5
GV	1012		1265	
V/S	5	5	5	5
V/S/A	10		11	
GV/S	1265		1265	
Residuals	2530	10	2783	11
Total	5080	20	5588	22

**Table 3: Number of DE in SJ and IA indices.**

Mutant	SJ		IA	
	FDR 0.01	FDR0.05	FDR0.01	FDR0.05
<i>spt4Δ</i>	2	2	14	14
<i>spt5-242</i>	3	3	48	69
<i>spt5-4</i>	1	1	52	72
<i>spt5-194</i>	12	12	88	113
<i>cegl-250</i>	134	160	151	163

**Table 4: Distribution of DE genes.**

Mutant	Gene class	# DE genes with positive fold change in IA	# DE genes with negative fold change in IA
<i>spt4Δ</i>	RP	6	0
	non-RP	5	3
<i>spt5-242</i>	RP	44	0
	non-RP	17	5
<i>spt5-4</i>	RP	52	0
	non-RP	10	8
<i>spt5-194</i>	RP	72	0
	non-RP	13	24
<i>cegl-250</i>	RP	89	0
	non-RP	52	17



**Table 5: Properties of DE genes with a positive fold change (Average).**

Start is the nucleotide position in ORF where intron begins; mRNA/hr is the number of times a gene is transcribed per hour. Numbers in bold, italic text are significantly different from the corresponding value for all introns at the  $p < 0.05$  level.

mutant	RP			non-RP		
	intron length	start	mRNA/hr	intron length	start	mRNA/hr
All introns	405	48	94.52	156	128	8.27
<i>spt4Δ</i>	<b>342</b>	19	70.52	253	160	<b>48.20</b>
<i>spt5-242</i>	410	31	102.04	196	154	<b>22.92</b>
<i>spt5-4</i>	400	52	92.48	<b>396</b>	281	<b>42.51</b>
<i>spt5-194</i>	412	35	94.11	<b>324</b>	226	<b>30.68</b>
<i>ceg1-250</i>	408	51	96.24	164	134	<b>12.02</b>

**Table 6: Properties of DE genes with a negative fold change (Average).**

Numbers in bold, italic text are significantly different from the corresponding value for all nonRP introns at the  $p < 0.05$  level.

mutant	# DE genes	intron length	start	mRNA/hr
<i>spt4Δ</i>	3	105	<b>615</b>	<b>0.80</b>
<i>spt5-4</i>	8	107	44	<b>1.55</b>
<i>spt5-242</i>	5	106	327	<b>1.10</b>
<i>spt5-194</i>	24	133	170	<b>1.75</b>
<i>ceg1-250</i>	17	161	169	4.77

**Table 7: QPCR validation DE microarray data.**

GENE	QPCR target	Fold change				
		<i>spt4Δ</i>	<i>spt5-4</i>	<i>spt5-242</i>	<i>spt5-194</i>	<i>ceg1-250</i>
<i>YGR027C</i> ( <i>RPS25A</i> )	pre-mRNA	1.3	2.33	-0.77	2.17	-0.7
	spliced mRNA	-1.07	-1.07	-2.23	-1.17	-4.17
	pre-/spliced mRNA	<b>2.37*</b>	<b>3.40*</b>	1.47	<b>3.33*</b>	<b>3.47*</b>
<i>YLR344W</i> ( <i>RPL26A</i> )	pre-mRNA	-0.63	1.47	-1.07	2.57	-0.37
	spliced mRNA	-0.53	-0.63	-3.37	-3.60	-4.53
	pre-/spliced mRNA	<b>-0.10#</b>	<b>2.10*</b>	2.30	<b>6.17*</b>	<b>4.17*</b>
<i>YOL127W</i> ( <i>RPL25</i> )	pre-mRNA	-0.73	0.73	1.37	0.5	-2.47
	exon2	-0.53	-0.63	-2.1	-2	-4.93
	pre-mRNA/exon2	<b>-0.20#</b>	<b>1.37*</b>	<b>3.47*</b>	<b>2.50*</b>	<b>2.47*</b>
<i>YDR064W</i>	pre-mRNA	-2.13	-1.53	-0.93	-1.7	-1.43
	exon2	-0.97	-0.93	-2.3	-1.23	-3.83

<i>(RPS13)</i>	pre-mRNA/exon2	-1.17	<b>-0.60#</b>	1.37	<b>-0.47#</b>	<b>2.40*</b>
	pre-mRNA	-0.23	1.00	4.00	0.30	-0.23
<i>SNR17B</i> <i>(U3B)</i>	exon2	1.60	1.97	-0.03	1.93	0.83
	pre-mRNA/exon2	-1.83	<b>-0.97#</b>	<b>4.03*</b>	-1.63	-1.07

Numbers in bold text highlight concordance between the QPCR and Microarray (DEDS) analysis: 1) numbers with asterisks indicate genes identified as DE using DEDS; 2) numbers in bold with pound symbols indicate genes identified as non-DE using DEDS and whose QPCR fold changes are within the (-1, 1) thresholds.

**Table 8: Yeast Strains.**

Strain	Genotype	Source
FY120	Mat <b>a</b> <i>his4-912</i> $\delta$ <i>lys2-128</i> $\delta$ <i>leu2</i> $\Delta$ <i>ura3-52</i>	Fred Winston
GHY92	Mat $\alpha$ <i>his4-912</i> $\delta$ <i>lys2-128</i> $\delta$ <i>leu2</i> $\Delta$ <i>ura3-52 spt5-242</i>	Hartzog lab
GHY379	Mat $\alpha$ <i>his4-912</i> $\delta$ <i>lys2-128</i> $\delta$ <i>leu2</i> $\Delta$ <i>spt5-194</i>	Hartzog lab
GHY524	Mat <b>a</b> <i>his4-912</i> $\delta$ <i>lys2-128</i> $\delta$ <i>leu2</i> $\Delta$ <i>spt4</i> $\Delta$ 2:: <i>HIS3</i>	Hartzog lab
FY1668	Mat <b>a</b> <i>his4-912</i> $\delta$ <i>lys2-128</i> $\delta$ <i>spt5-4</i>	Fred Winston
OY163	Mat <b>a</b> <i>his3 lys2-128</i> $\delta$ <i>ura3 ceg1-250</i>	Hartzog lab

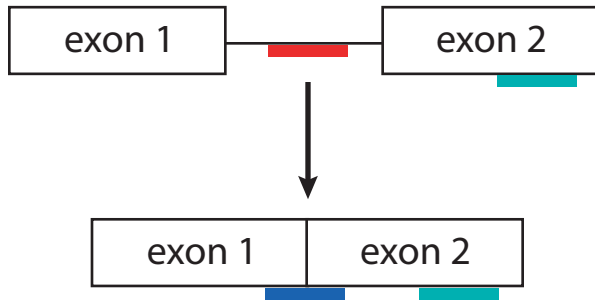
## **Supporting Information**

Table S1: A complete list of differentially\_expressed\_genes for the five mutants in both SJ and IA indices.

Table S2: Oligo\_sequences used in the QPCR validation of the microarray analysis.





## A) microarray probes

Intron-containing genes



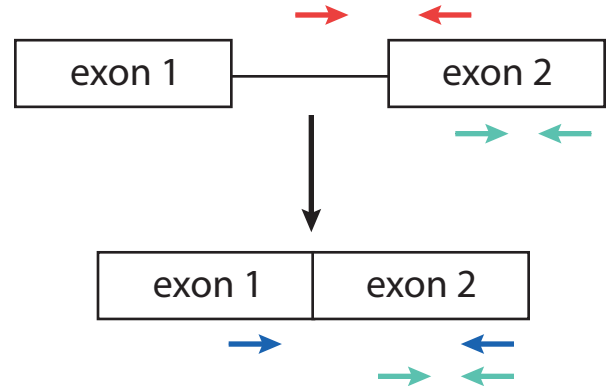
Intronless genes



-  Intron
-  Exon 2
-  Splice Junction
-  Intronless


## B) RT-PCR primers

Intron-containing genes

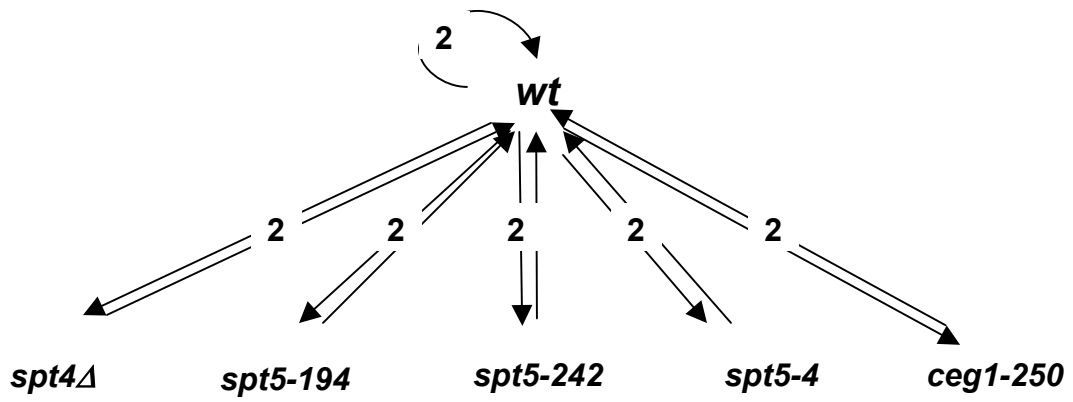


Intronless genes

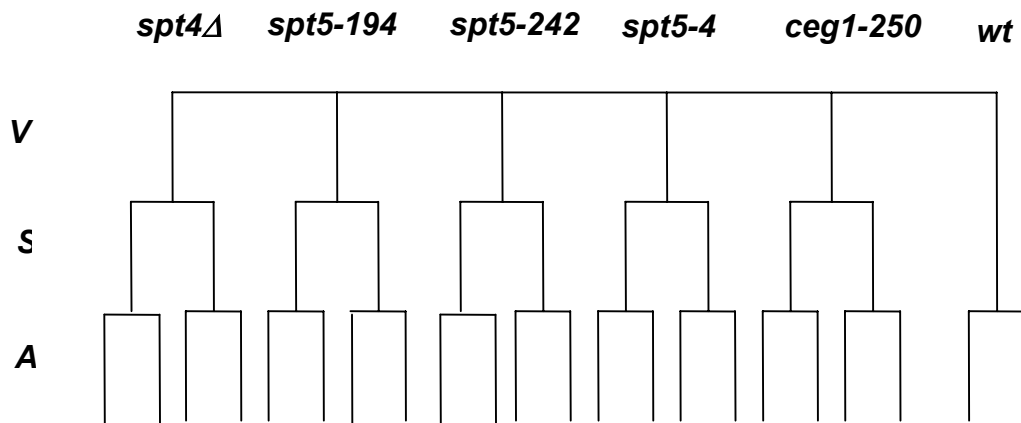


-  Unspliced
-  Exon 2
-  Spliced
-  Intronless

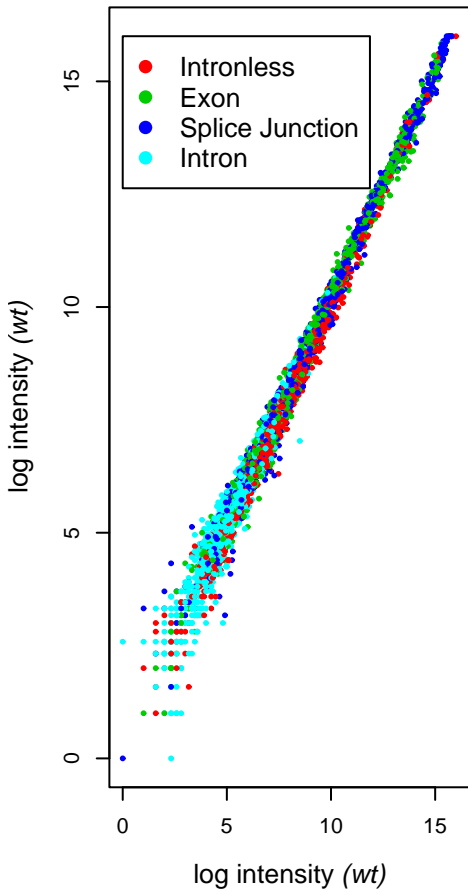
a)



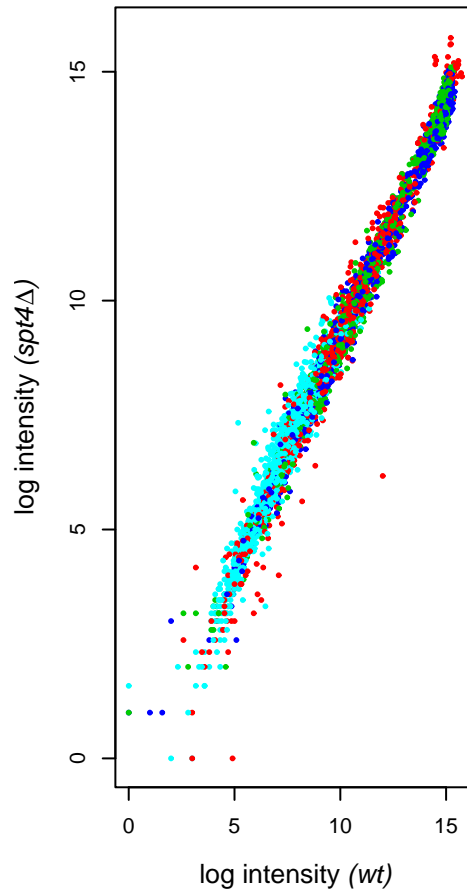
b)



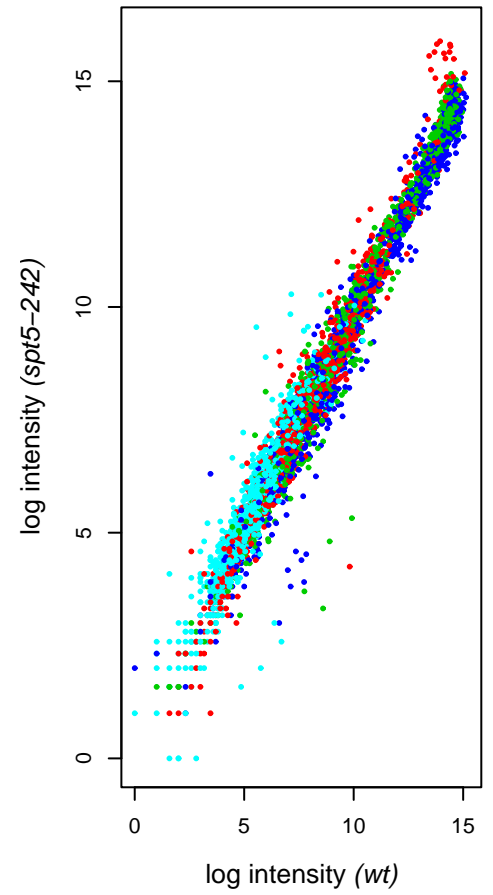
*wild-type*



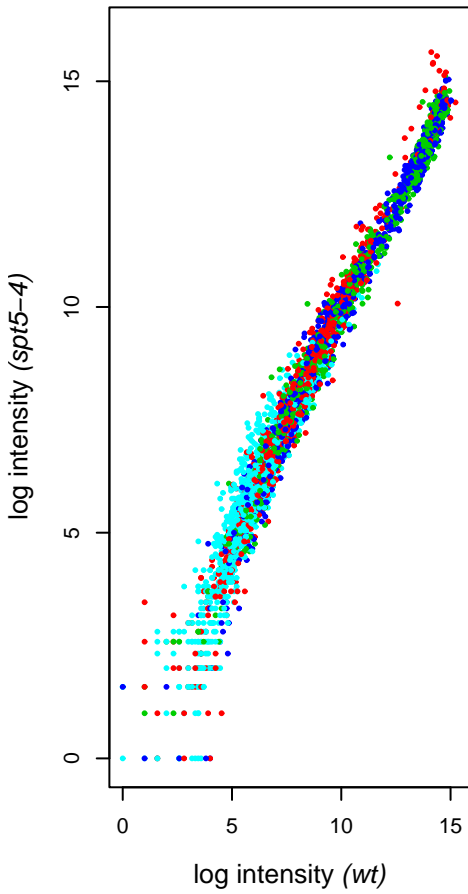
*spt4 $\Delta$*



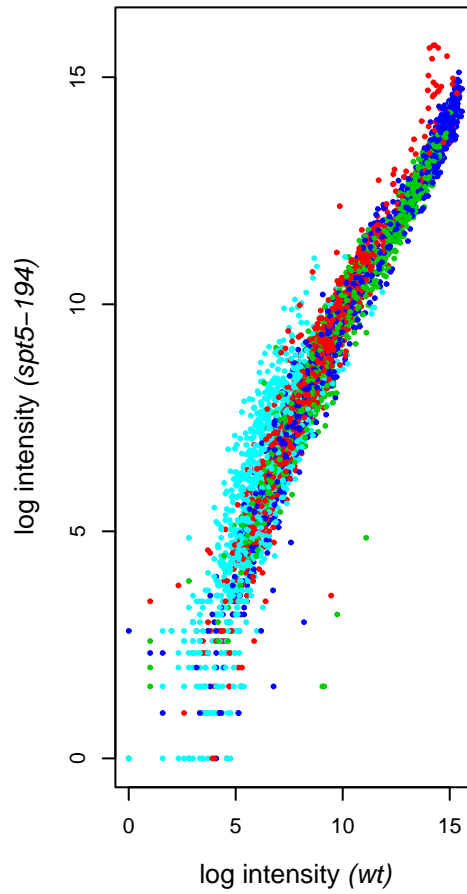
*spt5-242*



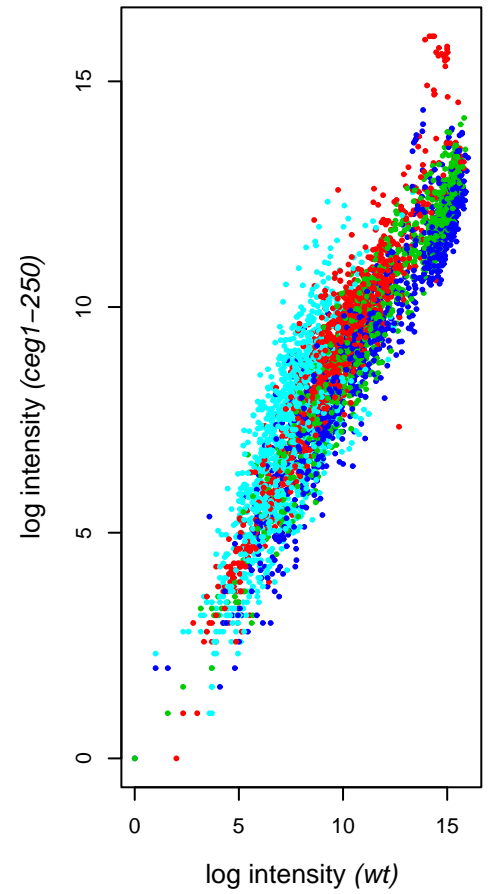
*spt5-4*

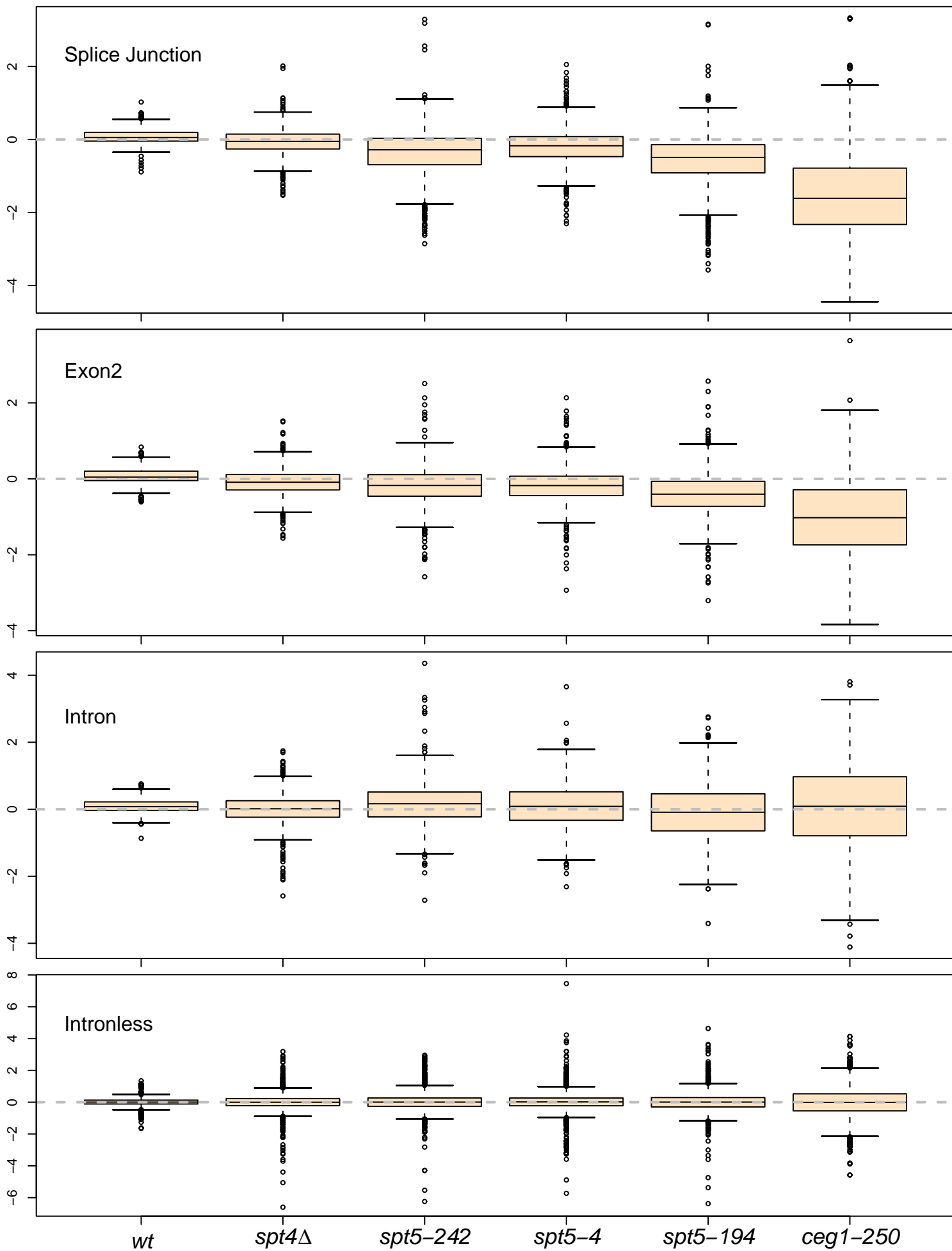


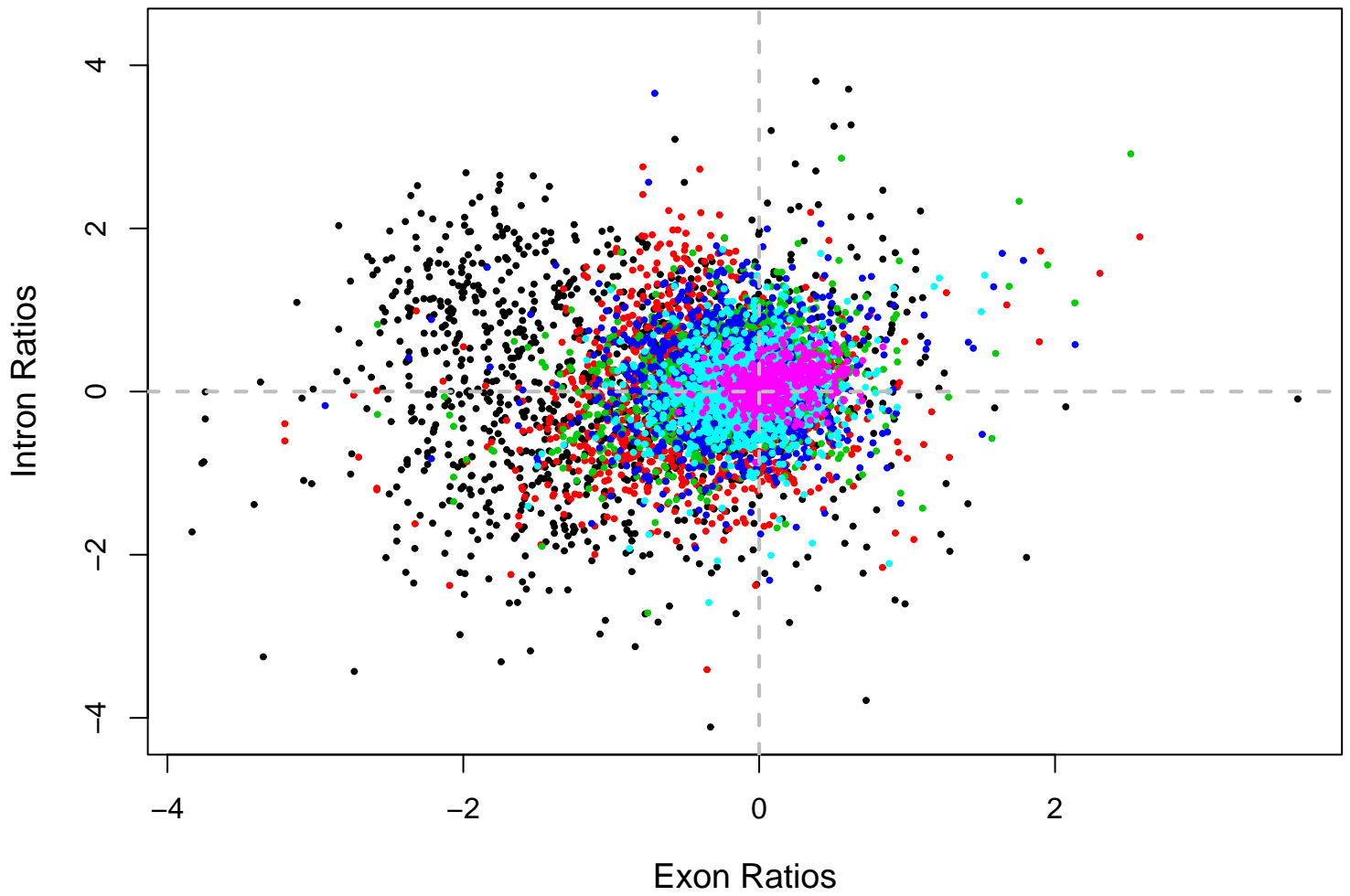
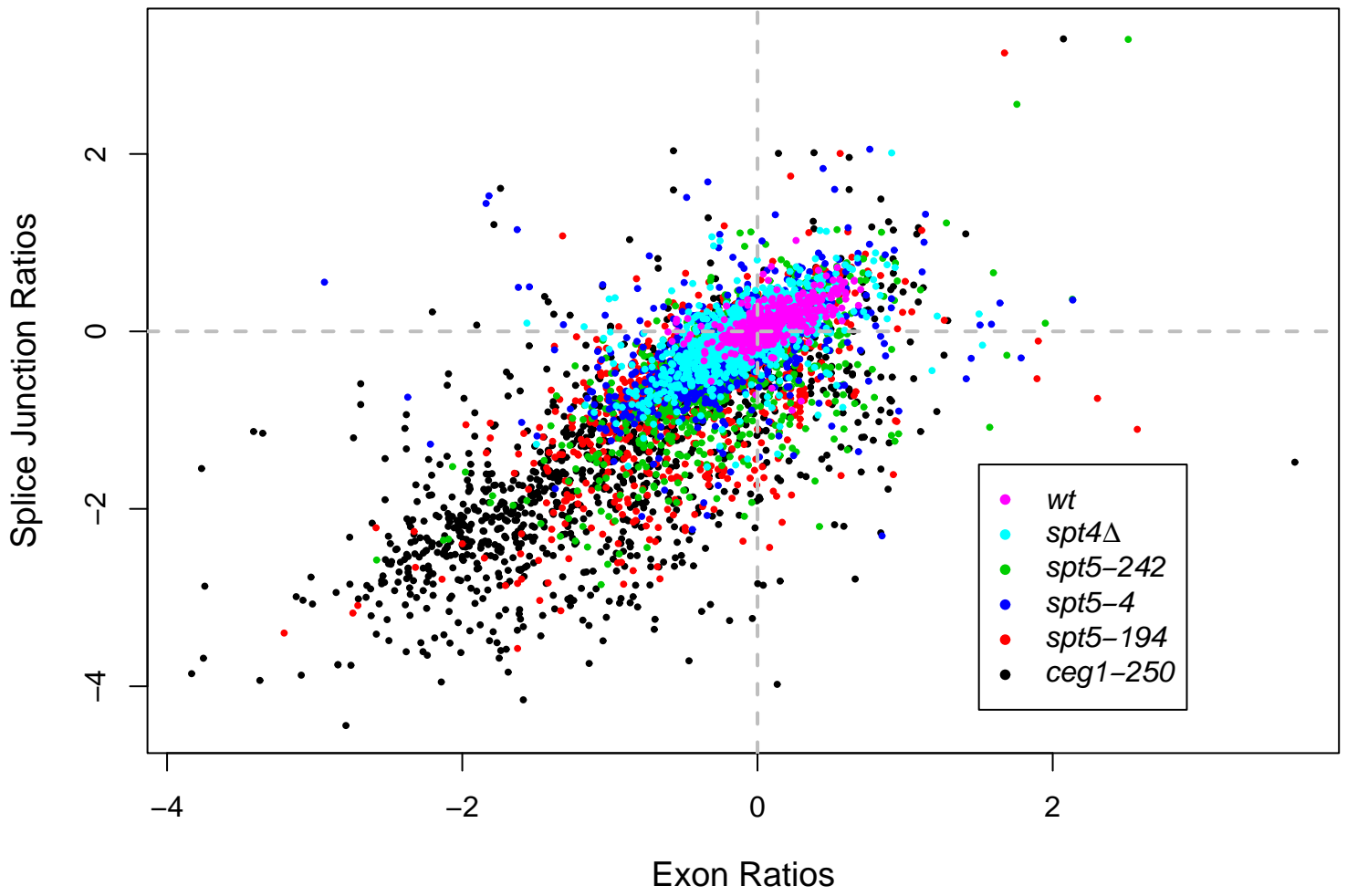
*spt5-194*



*ceg1-250*







### Splicing DE models for SJ indices / ceg1-250

