Discovering Regulatory Overlapping RNA Transcripts

Citation: Danford, Timothy et al. "Discovering Regulatory Overlapping RNA Transcripts." Research in Computational Molecular Biology. Ed. Bonnie Berger. LNCS Vol. 6044. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. 110-122.

As Published: http://dx.doi.org/ 10.1007/978-3-642-12683-3_8

Publisher: Springer Berlin/Heidelberg

Persistent URL: http://hdl.handle.net/1721.1/73688

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Discovering Regulatory Overlapping RNA Transcripts

Timothy Danford[1], Robin Dowell[1]*, Sudeep Agarwala[2],
Paula Grisafi[2], Gerald Fink[2], David Gifford[1]

[1] Massachusetts Institute of Technology
[2] Whitehead Institute

Abstract. STEREO is a novel algorithm that discovers cis-regulatory RNA interactions by assembling complete and potentially overlapping same-strand RNA transcripts from tiling expression data. STEREO first identifies coherent segments of transcription and then discovers individual transcripts that are consistent with the observed segments given intensity and shape constraints. We used STEREO to identify 1446 regions of overlapping transcription in two strains of yeast, including transcripts that comprise a new form of molecular toggle switch that controls gene variegation.

* Currently: University of Colorado

1 Introduction

Evidence has recently emerged from high-throughput expression datasets that overlapping RNA transcripts can play an important role in gene regulation. For example, an antisense transcript can be used to regulate its corresponding sense gene [8,3]. In budding yeast, a sense/antisense toggle has been shown to regulate the mating type of the cell [6]. The interference of a transcript on the same strand as a coding transcript is also sufficient to play a repressive role in the regulation of downstream genes [10,11].

Discovering RNA transcript based cis-regulation requires the precise spatial localization of transcripts and the identification of their overlap with other, nearby transcripts. Contemporary algorithms for analyzing tiling microarray identify non-overlapping segments of coherent transcription [15], but they do not attempt to identify the transcripts that generated and potentially span the observed segments. A genomic locus which is multiply transcribed may produce a region of complex segmentation, but no additional resolution of such regions into separate, overlapping transcripts can be provided by dynamic programming or the probabilistic models used by segmentation algorithms.

We present a new algorithm, `STEREO`, for the computational analysis of overlapping transcription. `STEREO` is organized into two phases. The first phase implements segmentation and discovers genomic intervals which are transcribed in one of the input experiments. The identified genomic intervals are classified into transcript or background classes using both observed probe intensities and the $3'$ to $5'$ transcript intensity fall-off caused by reverse transcriptase processivity. The second phase, transcript reconstruction, resolves this labeled segmentation into consistent arrangements of explanatory RNA transcripts. `STEREO` performs a combinatorial search of all possible transcripts given the constraints of transcript additivity and differential expression to yield a segmentation.

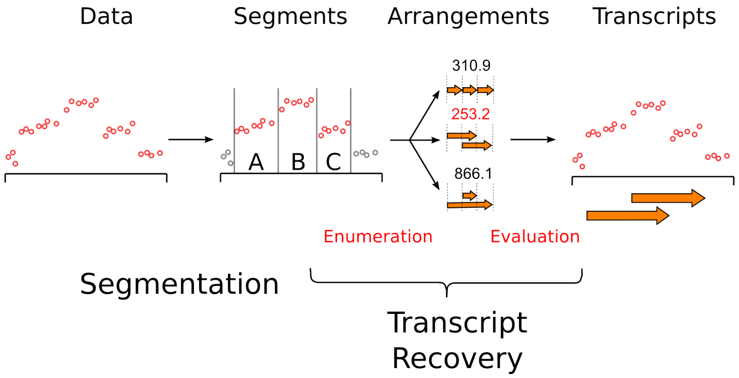

Fig. 1. Workflow description for segmentation and transcript recovery phases of the `STEREO` algorithm. The phases operate on sequence in a genomic region tiled by a microarray. The first phase partitions the genome into intervals and assigns each interval a local transcription label. The second phase identifies clusters of transcription. For each cluster, transcript arrangements are enumerated and evaluated, and the optimal arrangement is chosen as the explanation for that cluster.

STEREO is the first algorithm to computationally discover regions of complex transcription from segmentation, and to resolve those regions into overlapping transcripts. Overlapping transcripts fall into two mutually-exclusive categories, opposite and same-strand overlap, both of which may be expected to exhibit mutual interference or other regulatory properties. Opposite-strand overlap involves two transcripts are transcribed from complementary DNA strands and whose spatial extents overlap despite their different directions. Sense/antisense pairs of transcripts over the same gene are an example of opposite-strand overlap. Same-strand overlap occurs when two or more transcripts transcribed from overlapping portions of a DNA strand.

We tested STEREO on tiling expression data from two strains of yeast and discovered 1,446 instances of transcriptional overlap. Of these, 564 (39.0%) were overlapping in the same strand, a percentage consistent with previous estimates of alternate promoter usage in known yeast coding regions [12]. Northern blot analysis confirmed a same-strand interaction predicted by STEREO, and STEREO also identified opposite-strand transcripts that are organized into a novel form of molecular toggle switch [2] that controls the state of gene variegaton.

The remainder of our paper is organized into sections that describe notation, previous work, and experimental design (Section 2), expression segmentation and results (Section 3), transcript discovery and results (Section 4), and a discussion (Section 5).

2 Preliminaries

2.1 Notation

We begin by outlining some basic notation for arrays and transcripts, shown in Table 1. For geometric descriptions of locations along the genome, we will use two terms: points and intervals. A point will be a location of a single nucleotide in a genome assembly. Probes map to a genome assembly as point locations, based on the center of the interval to which their sequence is uniquely

i, j, s, t	probes i, experiments j, segments s, transcripts t
x_i	genomic location of probe i
y_{ij}	intensity of probe i in experiment j
$\mathbf{5'}_{[s,t]}, \mathbf{3'}_{[s,t]}$	$5'$ and $3'$ ends of segment s or transcript t
$\lvert x - x' \rvert$	linear distance along the genome, in bp
$\mathbf{t}_s$	type of segment s
θ_s	parameters of segment s
δ_{it}	the distance $\lvert x_i - \mathbf{3'}_t \rvert$ from probe i to the $3'$ end of transcript t
T_i	set of transcripts that overlap probe i
γ_{tj}	intensity of transcript t in experiment j
λ_t	3' log-linear slope of transcript t

Table 1. Array, segmentation, and transcript notation summary.

mapped. Intervals are convex subsets of the genome, single coherent loci specified completely by start and end positions.

A breakpoint set $B = (b_1, \ldots, b_N)$ is an ordered list of genomic locations which partition the genome into a set of non-overlapping intervals called segments. A segmentation is a set of segments which partition a complete genome. For a given set of breakpoints B, we use $\mathbb{S}_B$ to indicate the segmentation defined by those breakpoints. If $s \in \mathbb{S}_B$, then s is a genomic interval whose endpoints ($\mathbf{5'}_s$ and $\mathbf{3'}_s$) are consecutive elements of the list B. A segmentation algorithm assigns each segment a type $\mathbf{t}_s$ and a set of parameters θ_s which provide a local description of the probe values within that segment.

A transcript is a genomic interval, characterized by its start and end points $\mathbf{5'}_t$ and $\mathbf{3'}_t$. It is a single, coherent message transcribed from the genome in one or more cells. It may be edited or it may be present in an unedited form, in which case it will appear as an interval when matched to the genome which produced it. Overlapping transcripts will produce complex regions of transcription. Transcribed regions are sections of the genome which may form part or all of a single mapped transcript or multiple adjacent and overlapping transcripts.

2.2 Prior Work

Analysis of tiling microarray data by segmentation was originally used for the analysis of comparative genomic hybridization [15, 19]. Picard et al. described the first dynamic-programming based segmentation algorithms for discovering regions of copy number variation in array-CGH experiments [15]. They later extended their algorithm to provide automatic labeling of segments using a hybrid dynamic programming/expectation maximization approach [16]. These methods derive their computational efficiency from the fundamental assumption that the segments they identify form a non-overlapping partition of the genome into spatially coherent intervals, an assumption which allows the use of dynamic programming approaches to discover optimal segmentations.

Tiling microarrays are also used to measure the transcription of genomic regions, and segmentation algorithms were similarly adapted to uncover consistently transcribed regions in those datasets [18]. Huber et al. adapted the segmentation algorithm of Picard to identify transcribed regions from tiling arrays [7]. This method was then used in David et al., which published the first tiling microarray study of transcription in yeast [4]. One additional feature of tiling microarrays was their ability to discover strand-specific transcription through the use of strand-specific probes and experimental protocols which preserved the strand-specificity of the sample. The array results of David et al. were strand-specific, and so were able to identify regions of opposite-strand overlapping transcription.

Microarrays are not the only method for analyzing transcription on a genome-wide scale and in an unbiased manner; sequencing of cDNA has been a standard way to identifying unknown transcripts. Miura et al. sequenced expressed cDNA tags to produce a catalog of $5'$ and $3'$ transcript end-points throughput the yeast genome [12]. Sequencing measures single transcripts (and not transcribed regions) directly, and therefore can give information about the structure of transcript overlaps, starting, and ending points assuming that the read length is long enough relative to the transcript lengths.

The use of new, high-throughput short read sequencing machines to investigate transcription has led to the recent adoption of RNA-seq as a measurement of genome wide transcription [9, 21]. RNA-seq experiments sequence fragments of transcripts which are randomly selected from the sample. Nagalakshmi et al. provided the first strand-insensitive view of transcription through RNA-seq in budding yeast [13]. These results have been extended in a strand-specific manner in related strains of yeast by Wilhelm et al. [20]. Unlike traditional sequencing, which produces longer reads, these unpaired-end short read sequencing techniques are unable to give us a full picture of the transcripts from which they were taken and suffer from the same problem of transcript mixture as microarrays. Some sequencing protocols produce reads which are insensitive to the strand of the underlying transcript, requiring that downstream computational analyses include strand-differentiation as one of their goals [14].

2.3 RNA *cis*-regulation in S288C and Σ1278b

Using an array designed to probe the S288C genome at approximately 50 base-pair resolution, we designed a set of experiments intended to reveal differences in transcription regulation between two closely related strains of *Saccharomyces cerevisiae*: S288C and Σ1278b [5]. Each array had two channels, Cy3 and Cy5, which were used to simultaneously measure the expression in the two strains. In addition to the haploid (mat-α) dataset of [5], we generated diploid expression in rich media with a technical replicate of each experiment. Treating each channel of each array as a separate logical experiment, this design provided us with eight total experiments on which to perform our segmentation and analysis. Data was normalized across experiments using quantile normalization [1].

3 Segmenting expression using multiple constraints

The segmentation phase of our algorithm partitions the genome into a complete set of non-overlapping regions. Each block, or segment, in the partition is labeled either TRANSCRIBED or BACKGROUND and assigned a set of local parameters that model the microarray probe observations within the segment. The segmentation considers multiple microarray experiments as input and learns a single segmentation that jointly explains all the input experiments. Segments may be assigned local parameters on an experiment specific basis, but the locations of the segments and the breakpoints that divide them are common across all experiments. Each label (TRANSCRIBED and BACKGROUND) corresponds to a model class, each with different complexities (requiring a penalty for the choice of a more complex class). The algorithm chooses from two classes, a flat model class that fits a mean and a variance to a given segment and represents the BACKGROUND segment label, and a linear model class that fits a line to the log intensities of the probes in a segment and is used to model the TRANSCRIBED label. The linear model class captures the $3'$ falloff effect created by the reverse transcriptase step of our experimental protocol. Both model classes can be represented by their log likelihood functions:

$$\mathcal{L}_j^{(1)}(x_1, x_2, \mu_j, \sigma) = \frac{1}{2}\log(\sigma) \sum_{i:x_1 \le x_i \le x_2} \frac{(y_{ij} - \mu_j)^2}{2\sigma^2} \tag{1}$$

$$\mathcal{L}_j^{(2)}(x_1, x_2, \mu_j, \lambda, \sigma) = \mathit{\Pi} + \frac{1}{2}\log(\sigma) \sum_{i:x_1 \le x_i \le x_2} \frac{(y_{ij} - \log(\mu_j e^{\delta_i \lambda}))^2}{2\sigma^2} \tag{2}$$

Here the x_1 and x_2 parameters are the bounds of the segment, while x_i and y_{ij} are the location of probe i and the value of probe i in experiment j, respectively. $\mathit{\Pi}$ is a penalty term which corrects for the choice of the more complex (linear) model class, and is set through training against synthetically generated data. For a fixed pair of segment bounds, the choice of parameters for either model class are obtained by maximizing the corresponding log likelihood functions $\mathcal{L}^{(1)}$ or $\mathcal{L}^{(2)}$. For either likelihood function, we will write $\theta^* \equiv \arg\max_\theta \sum_j \mathcal{L}_j(x_1, x_2, \theta)$ to indicate the maximum likelihood values of the parameters given the segment boundaries x_1 and x_2, and $\mathbb{L}(x_1, x_2) \equiv \mathcal{L}(x_1, x_2, \theta^*)$ for the log-likelihood as a function of just the segment endpoints.

3.1 Segmentation phase uses dynamic programming to find an optimal segmentation

The algorithm finds an optimal set of segmentation boundaries such that the total log likelihood of all probe observations from all experiments is maximized. Since a segmentation is a partition that separates the genome into non-overlapping regions, this can be accomplished through dynamic programming on the recursive formulation for $\mathbb{L}$.

$$\mathbb{L}(x_1, x_2) = \begin{cases} \mathbb{L}^{(1)}(x_1, x_2) \\ \mathbb{L}^{(2)}(x_1, x_2) \\ \max_{b \in [x_1, x_2]} \mathbb{L}(x_1, b) + \mathbb{L}(b, x_2) \end{cases} \tag{3}$$

The identity of any segment can be tracked by remembering which choice is maximizing. Those segments $[x_1, x_2]$ for which the $\mathbb{L}^{(1)}(x_1, x_2)$ is optimal are given the BACKGROUND label, while the TRANSCRIBED label is assigned to those for which $\mathbb{L}^{(2)}(x_1, x_2)$ was optimal.

3.2 Segmentation phase discovers novel regulatory transcription

We used the segmentation and labeling phase of our algorithm to analyze the tiling microarray experiments described in Section 2.3. We ran our segmentation algorithm on each strand of the tiling microarray dataset separately, and we identified $14,076$ segments on the Watson strand and $13,792$ on the Crick strand. From the segmentation on the Watson strand, we identified 37.0% of the tiled genome as transcribed, and from the Crick strand we identified 40.3%; taken together, accounting for overlap, we identified 65.2% of the complete genome sequence as transcribed.

The segmentation recovered two noncoding transcripts whose regulatory function is related to their spatial overlap and interference with the production of a downstream coding transcript. In

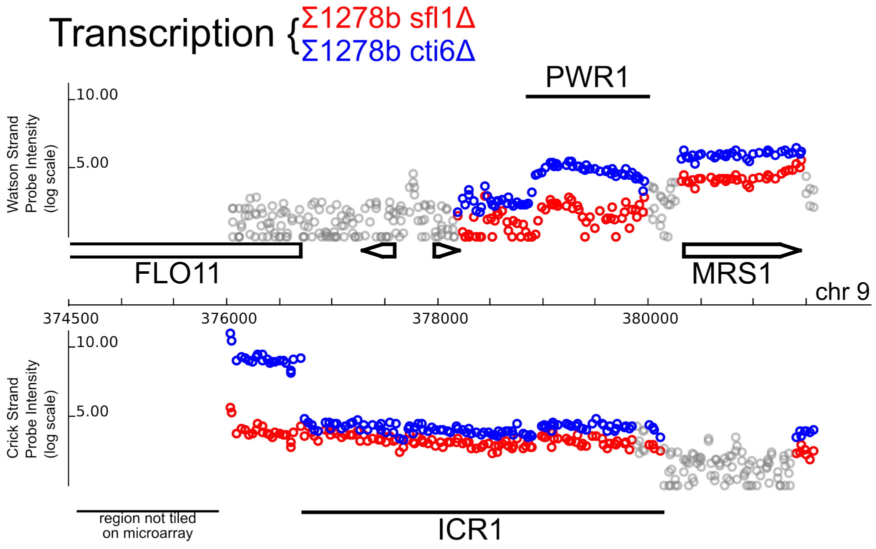

Fig. 2. We are able to discover noncoding transcription which is known to play a role in the regulation both of a downstream coding transcript (FLO11) and of each other. ICR1 and PRW1 are noncoding RNAs, reported in [2], whose regulatory function is related to their spatial overlap. The segmentation phase of our `STEREO` algorithm is able to find the complete PRW1 transcript and the 3′ end of the ICR1 transcript in the Bumgarner dataset. Regions identified as `BACKGROUND` are shown in grey, `TRANSCRIBED` regions are shown in color. Genes are identified as arrowed boxes. The x-axis is genomic coordinates and the y-axis is log intensity.

Figure 2, we show the locations of two noncoding transcripts, PWR1 and ICR1, that we showed implement a new type of RNA molecular toggle [2]. We also ran an implementation of the Picard segmentation algorithm on the S288C and Σ1278b dataset. Although this method can be easily adapted to handle multiple experiments simultaneously, it lacks the ability to identify regions with a shape other than a flat regions of transcription; instead, it separates the sloped regions of transcription into "steps" of multiple flat segments. Therefore, the Picard algorithm is unable to handle a key feature of our experimental protocol (the 3′ falloff) and unnecessarily single units of transcription into artificially complex sets of segments.

4 `STEREO` assembles transcripts from expressed segments

4.1 An additive model for overlapping transcripts

Our model for overlapping transcripts employs two key constraints. First, we constrain same-strand overlapping transcripts that are co-expressed to display additive expression in their region of overlap. Second, we constrain transcripts to display 3′ to 5′ fall off in intensity corresponding to the processivity of reverse transcriptase in our experimental method. Our additivity constraint is reflected in the summation in Equation 4, and the slope constraint is reflected in the parameter λ

that uniformly applies to all modeled transcripts. Equation 4 models observed intensities y_{ij} as the sum of transcript levels γ_{tj} associated with a particular transcript t in experiment j. Equation 4 makes the assumption that the noise of the array is log-normal, but that the transcripts themselves are additive in the non-logarithmic-scale of the array.

$$y_{ij} = \log(\sum_{t \in T_i} \gamma_{tj} e^{\lambda \delta_{it}}) + e_{ij} \tag{4}$$

If we give the unit level error term a probability distribution, $\epsilon_{ij} \sim \mathcal{N}(\cdot; 0, \sigma_y)$, we turn Equation 4 into a probabilistic model with log-likelihood function:

$$\mathcal{L}(\Gamma, \sigma, \lambda) = -N\sigma - \sum_i \sum_j \frac{(y_{ij} - \log(\sum_{t \in T_i} \gamma_{tj} e^{\lambda \delta_{it}}))^2}{2\sigma^2} \tag{5}$$

The vector of transcript intensities $\Gamma = \{\gamma_{tj}\}$, along with the transcript slope λ and probe level variance σ, are chosen to maximize the log likelihood function in Equation 5. Since this equation is a non-linear function of a sum, there is not a simple closed-form solution for the maximizing parameters. Instead, we compute the derivatives of the log-likelihood function and maximize numerically using gradient ascent.

4.2 Enumerating and evaluating overlapping transcripts

A maximum likelihood solution to Equation 5 provides a method for finding local parameters for a set T of overlapping transcripts. However, it does not answer the question of how we determine T. A poor choice of T will lead to estimates of transcript intensities that do not correspond to biological reality.

STEREO uses an enumeration-based search method to choose the transcript arrangement T which best explains the transcribed regions that are provided as input by the segmentation and labeling phase. We assume that the segmentation provided to the transcript discovery phase has correctly identified the starts and ends of transcripts as breakpoints in the segmentation, and has identified each segment as transcribed or noise. Furthermore, we assume that every transcribed segment will be explained by at least one transcript, while noise segments will not be explained by any transcript.

We break the problem of transcript calling into independent sub-problems, called clusters. Each cluster is a spatially-consecutive sequence of transcribed segments, separated from every other cluster either by one or more noise segments or a chromosome boundary.

The STEREO algorithm first identifies the clusters corresponding to the input segmentation of tiling microarray data. Then for each cluster, it enumerates all possible transcript arrangements. Each cluster will have a finite number of arrangements, since there are a finite number of breakpoints in the cluster and we assume that the total number of transcripts does not exceed the number of segments in the cluster. For each enumerated transcript arrangement T, we find an optimal set of parameters $\Theta_T \equiv \langle \Gamma_T, \lambda_T, \sigma_T \rangle = \arg\max \mathcal{L}(\Gamma, \lambda, \sigma)$ by maximizing the log-likelihood equation of the probes within the cluster. The penalized log-likelihood $\mathcal{L}(\Gamma_T, \lambda_T, \sigma_T) - C(T)$ then provides a score by which to evaluate the fit of the transcript arrangement T to the cluster. The complexity penalty $C(T) = \alpha|T| + \beta c_{\text{over}}(T)$ assesses a constant penalty for the total number of transcripts in T and for the number of overlaps $c_{\text{over}}(T)$ in the arrangement. The penalties (α and β) are chosen to optimize transcript discovery against synthetically-generated data.

4.3 STEREO Transcript discovery recovers appropriate SER3/SRG1 transcripts

An example of overlapping transcripts with regulatory interactions in yeast is the SER3 gene and its upstream intergenic transcript, SRG1. The SER3 gene is involved in serine biosynthesis and under repressing conditions its promoter is bound by significant levels of both TATA binding protein (TBP) and RNA polymerase II (Pol II). The expression of a short transcript that runs through the SER3-proximal TATA element is associated with decreased expression of the SER3 transcript itself [10]. Furthermore, a nearly 2 kb read-through transcript starting from the SRG1 TATA element and extending through the entire SER3 gene itself was observed by northern analysis in the same study.

The SER3 and SRG1 genes, and their observed architecture of overlapping transcription, provide a convenient test of our ability to estimate relative intensities of overlapping transcripts. In Figure 3, we show that our tiling array data in S288C (red) and Σ1278b (blue) around the SER3 and SRG1 locus. The figure depicts the locations of three overlapping transcripts, shown as orange

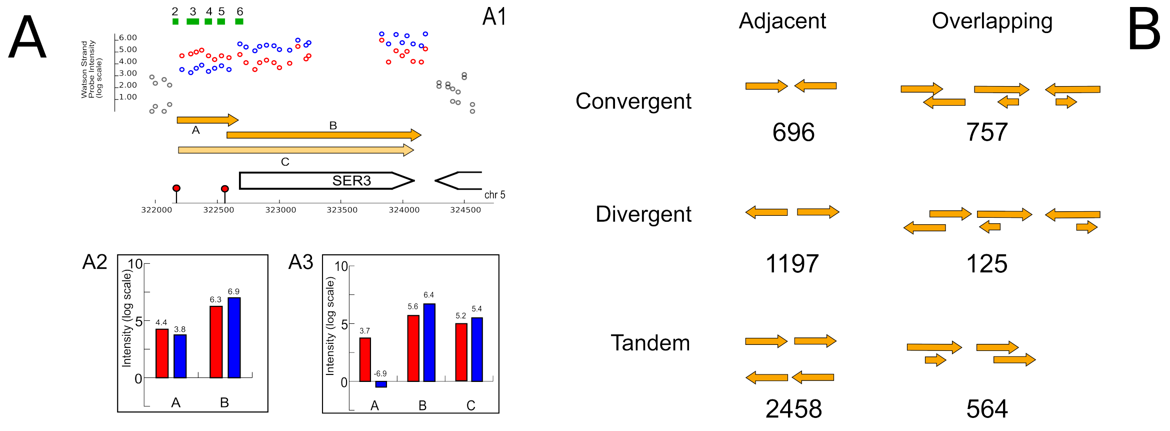

Fig. 3. **A** Re-construction of transcript intensities at the SER3/SRG1 locus. **A1**. Probes which are included in either the SER3 or SRG1 region and are included in this analysis are displayed in either red (S288C) or blue (Σ1278b). Original probes from Martens et al. enriched for the SRG1 transcript are green. Putative transcripts A, B, and C are shown in orange and TATA elements with red dots. Transcript A corresponds to Martens SRG1 transcript while Transcript B corresponds to SER3 transcript. Transcript C is the "readthrough" transcript Martens detected, extending exactly 2 kb. Transcript intensity analyses were carried out for two arrangements, (**A2**) just the A and B transcripts and (**A3**) all three transcripts. Each transcript has reconstructed intensities for both S288C (red) and Σ1278b (blue) experimental data. **B** Schematics for each category along with the total number of STEREO transcripts identified within each category are shown.

arrows: one from the upstream SRG1 TATA element extending to the annotated start of the SER3 gene, the second from the SER3 TATA element extending to the end of the SER3 gene, and one 2 kb-long transcript starting from the SRG1 TATA element and extending through the SER3 gene itself.

Using our transcript intensity estimation method we reconstructed relative log-intensities of 4.4 and 6.3 for the A and B transcripts respectively; these values are consistent with previously reported concentrations for SRG1 and SER3 respectively [10]. Moreover, the fitted intensities are anti-correlated across cell types, between the two measured strains of yeast. When the SRG1 transcript drops the SER3 transcript rises, consistent with the claim that SRG1's transcription represses that of SER3.

4.4 Identification of 1446 overlapping transcripts

STEREO resolved the collected S288C and Σ1278b expression datasets into 6609 transcripts. Most transcripts (5233) inferred by our method were strand singletons, covering a single region without a second overlapping transcript. However, our algorithm identified 1446 regions of overlapping transcription, of which 564 were same-strand overlapping transcripts. Figure 3 Part B shows a classification of transcript pairs into six categories depending on their relative orientation and overlap, and gives the number of transcript pairs that fell into each category from our dataset.

The segmentation and labeling phase has also been able to uncover overlapping transcript pair predictions which show differential expression between different cell types and strains, and whose variation is consistent with potential repressive regulatory interactions between the overlapping transcripts.

4.5 Northern analysis of overlapping predictions

In order to confirm one of our predictions we chose three of the predictions made by our algorithm to test with northern blot analysis. To facilitate northern blot analysis we chose examples to test that had a larger outer transcript with a smaller inner contained transcript that would readily be immediately apparent in the experimental result. In one of the three locations tested northen blot analysis showed same-strand overlapping transcription with transcript lengths matching those produced by STEREO. This validated locus, YCR082W, provides a new example of tandem overlapping transcripts previously unknown in the literature. Instead of reporting overlapping transcripts in this location, an alternate explanation would have been two tandem transcripts aligned head-to-tail; in this case, the transcript discovery algorithm reconstructs the more complex overlap based on our prior distribution over transcript intensities and our belief that higher-intensity transcripts are less likely than lower ones.

Zheng et al. have previously attempted to quantify the intensities of multiple overlapping transcripts using a hierarchical Bayesian model [22]. Their approach is limited, however, to the quantification of transcript intensities whose locations have already been specificed from gene annotations or an external datasource. Rochette et al. have reported a set of overlapping transcripts at a

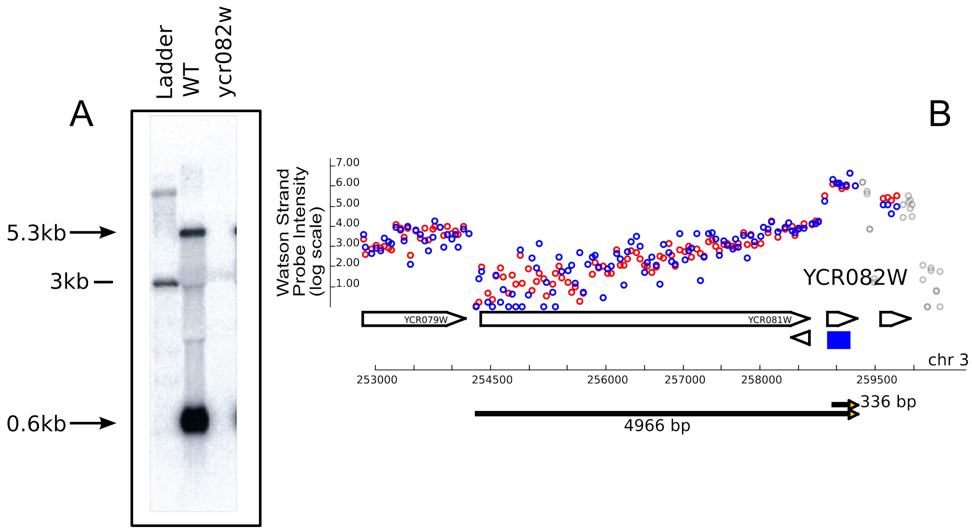

Fig. 4. Northern analysis was performed at YCR082W to test for the presence of multiple overlapping transcripts. Probes were chosen to cover the first 400 bp of the gene, shown as a blue square. The blot (**A**) shows two transcripts with lengths approximately 5 kb and 600 bp. These transcripts correspond (**B**) to two overlapping transcripts called by the `STEREO` algorithm with lengths of approximately 5 kb and 300 bp. For clarity, only the Watson strand is shown. Probes in `TRANSCRIBED` regions are shown in color for S288C (red) and Σ1278b (blue) data.

genome-wide level in the parasite *Leishmania* [17]. These transcripts were identified by experimental means (5′-RACE) in a genome significantly smaller than yeast, however, and do not represent a comprehensive computational approach to transcript discovery.

5 Discussion

We introduced several unique features of our `STEREO` algorithm. In the segmentation phase, we simultaneously incorporated multiple experiments and utilized the slope of the transcription data to identify transcribed segments. In the transcript discovery phase we employed both additive intensity and differential expression to evaluate likely configurations of transcripts.

`STEREO` also has certain limitations. While a 3′ to the 5′ intensity fall off provides a useful constraint, it also makes it more difficult to accurately locate the 5′ ends of long, low-abundance transcripts. In addition, `STEREO` is sometimes unable to separate same-strand overlapping transcripts without differential expression between conditions or strains. In these cases, overlapping transcript calling depends on our prior distributions on transcript intensities. A better understanding of the distribution of transcript abundances will improve the accuracy of our transcript reassembly algorithm. The combinatorial architecture of gene regulation is in part implemented by RNA based cis-regulation. We are making our set of 1446 candidate interactions available for other investigators.

References

1. B. Bolstad. Probe Level Quantile Normalization of High Density Oligonucleotide Array Data. Technical report, Division of Biostatistics, University of California, Berkeley, December 2001.
2. S. Bumgarner, R. Dowell, P. Grisafi, D. Gifford, and G. Fink. A Toggle Involving *Cis*-Interfering Noncoding RNAs Controls Variegated Gene Expression in Yeast. *PNAS*, 2009.
3. J. Camblong, N. Iglesias, C. Fickentscher, G. Dieppois, and F. Stutz. Antisense RNA Stabilization Induces Transcriptional Gene Silencing via Histone Deacetylation in *S. cerevisiae*. *Cell*, 131:706–717, November 2007.
4. L. David, W. Huber, M. Granovskaia, J. Toedling, C. J. Palm, L. Bofkin, T. Jones, R. W. Davis, and L. M. Steinmetz. A high-resolution map of transcription in the yeast genome. *PNAS*, 103(14):5320–5325, 4 April 2006.
5. R. D. Dowell, O. Ryan, A. Jansen, D. Cheung, S. Agarwala, T. W. Danford, D. Bernstein, P. A. Rolfe, G. R. Fink, D. K. Gifford, and C. Boone. Genotype to Phenotype: A Comparison of Two Interbreeding Yeast Strains Reveals Complex Genetics of Conditional Essential Genes. in submission.
6. C. Hongay, P. Grisafi, T. Galitski, and G. Fink. Antisense transcription controls cell fate in Saccharomyces cerevisiae. *Cell*, 127(4):735–745, 2006.
7. W. Huber, J. Toedling, and L. Steinmetz. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, 22(16):1963–1970, 2006.
8. T. A. Hughes. Regulation of gene expression by alternative untranslated regions. *Trends in Genetics*, 22(3):119–122, March 2006.
9. J. Marioni, C. Mason, S. Mane, M. Stephens, and Y. Gilad. RNA-Seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18:1509–1517, June 2008.
10. J. A. Martens, L. Laprade, and F. Winston. Intergenic transcription is required to repress the *S*accharomyces cerevisiae SER3 gene. *Nature*, 429:571–574, May 2004.
11. I. Martianov, A. Ramadass, A. S. Barros, N. Chow, and A. Akoulitchev. Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature*, 445:666–670, February 2007.
12. F. Miura, N. Kawaguchi, J. Sese, A. Toyoda, M. Hattori, S. Morishita, and T. Ito. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *PNAS*, 103(47):17486–17851, 21 November 2006.
13. U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320:1158441–1349, May 2008. 10.1126/science.1158441.
14. D. Parkhomchuk, T. Borodina, V. Amstislavskiy, M. Banaru, L. Hallen, S. Krobitsch, H. Lehrach, and A. Soldatov. Transcriptome analysis by strand-specific sequencing of complementary dna. *Nucleic Acids Research*, July 2009.
15. F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J.-J. Daudin. A Statistical Approach for Array CGH Data Analysis. *BMC Bioinformatics*, 6(27), February 2005.
16. F. Picard, S. Robin, E. Lebarbier, and J.-J. Daudin. A segmentation/clustering model for the analysis of array CGH data. *Biometrics*, 63:758–766, 2007.
17. A. Rochette, F. Raymond, J.-M. Ubeda, M. Smith, N. Messier, S. Boisvert, P. Rigault, J. Corbeil, M. Ouellette, and B. Papadopoulou. Genome-wide gene expression profiling analysis of *leishmania major* and *leishmania infantum* developmental stages reveals substantial differences between the two species. *BMC Genomics*, 9(255), May 2008.
18. T. Royce, J. Rozowsky, P. Bertone, M. Samanta, V. Stolc, S. Weissman, M. Snyder, and M. Gerstein. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends in Genetics*, 21(8):466–475, August 2005.
19. A. E. Urban, J. O. Korbel, R. Selzer, T. Richmond, A. Hacker, G. Popescu, J. F. Cubells, R. Green, B. S. Emanuel, M. B. Gerstein, S. M. Weissman, and M. Snyder. High-resolution mapping of DNA copy alterations in human chromosome 22 using high-density tiling oligonucleotide arrays. *PNAS*, 103(12):4534–4539, March 2006.
20. B. Wilhelm and J.-R. Landry. Rna-seq: quantitative measurement of expression through massively parallel rna-sequencing. *Methods*, 48(3):249–257, July 2009.
21. B. Wilhelm, S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C. Penkett, J. Rogers, and J. Bahler. Dynamic repertoire of a eukaryotic transcriptome surveyed at a single-nucleotide resolution. *Nature*, 453:1239–1243, June 2008.

22. S. Zheng and L. Chen. A hierarchical Bayesian model for comparing transcriptomes at the individual transcript isoform level. *Nucleic Acids Research*, pages 1–16, May 2009.