The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200

Nascent vs. Steady-State Sequencing Reveals Extensive but Unstable Antisense Transcription in Yeast

Nishant Chaparala (author)[1]

[1]Aragon High School, San Mateo, California, United States of America

Abstract

While only around 1% of our entire human genome consists of protein-coding regions, nearly all of it is shown to be active through transcription. A lot of this transcription occurs on the opposite strand of the annotated gene (the so-called "antisense" strand). The exact biological function of this form of transcription has yet to be fully understood. To characterize the extent, distribution, and stability of antisense transcription, both nascent (using NET-seq) and steady-state (using RNA-seq) transcription data for 2649 actively transcribed yeast genes were analyzed. Antisense transcription has been detected in 98.8% of all actively transcribed genes by means of NET-seq, compared to 77% by RNA-seq steady-state measurement; however, the level of this transcription is generally lower than that of the corresponding sense-transcribed RNA. Additionally, the ratio of sense-to-antisense reads measured by RNA-seq is much higher than the corresponding ratio measured by NET-seq; this suggests that, following transcription, antisense transcripts are degraded at a faster rate than sense transcripts. This was further confirmed by means of stability studies; median RNA-seq/NET-seq ratios show that antisense transcripts have a much lower median ratio

than sense-transcribed RNA (-2.85 and -0.47 on a $\log_2$ scale, respectively), indicating a rapid rate of turnover for antisense RNA molecules. In summary, these findings provide evidence that not only is antisense transcription widely prevalent, but it is subject to strict regulation, primarily through the process of post-transcriptional degradation, which ultimately serves to limit accumulation and therefore expression of antisense transcripts.

Keywords: antisense transcription, non-coding RNA, RNA-seq, transcript stability, yeast

Introduction

The central dogma describes how genes are transferred from DNA to RNA to protein. When the central dogma was first proposed, transcription was believed to be a precise process in which a limited set of protein-coding genes produced mRNAs that were directly translated into proteins. However, recent advances in whole-genome sequencing have changed this view. It is now understood that almost all the DNA in eukaryotes (including yeast and humans) is transcribed at some level and gives rise to many different types of non-coding RNA, all of which researchers are now studying to discover their functions (Djebali et al., 2012). One interesting aspect of non-coding transcription is antisense transcription, which produces RNA from the complementary DNA strand (the strand that is opposite the strand that encodes the protein). Antisense RNA can be produced through: (a) divergent transcription from gene promoters; (b) read-through from convergent, upstream genes; or (c) independent promoters (Neil et al., 2009; Xu et al., 2009). In all cases, the antisense RNAs are non-coding and have typically been disregarded as transcriptional background noise. There is significant evidence that some antisense transcripts have a regulatory function. For example, antisense transcripts can regulate gene expression at the transcriptional level by

forming double-stranded RNA with their complementary sense transcript, inhibiting transcription factor binding, or recruiting chromatin-modifying complexes (Neil et al., 2009). Therefore, two important questions need to be addressed: What is the extent of antisense transcription? And what factors determine the relatively low levels of antisense compared to sense transcripts? Two complementary sequencing techniques that provide significant insights into these questions are NET-seq (Native Elongating Transcript sequencing), which profiles actively synthesized RNA by capturing RNA elongating from the engaged RNA polymerase, providing a snapshot of nascent transcription that is unaffected by any subsequent processing (Churchman & Weissman, 2011), and RNA-seq, which measures steady-state transcript levels that indicate the overall flux of transcription and degradation (Core et al., 2008). A direct comparison between these two datasets can provide insight into the fate of both sense and antisense transcripts after transcription.

In this investigation, NET-seq and RNA-seq datasets derived from 2,649 actively expressed yeast genes were evaluated. The aims addressed are: (i) understand the profile of antisense transcription throughout the entire gene body; (ii) evaluate genome wide distributions of sense/antisense transcription among the two collections of data; (iii) compare the level of concordance in antisense transcriptional activity as measured by the two sequencing methodologies; (iv) evaluate the sense to antisense transcription ratio for all 2,649 actively expressed yeast genes; (v) determine the relative stability of sense versus antisense transcripts. Together, these findings provide a detailed view of antisense transcriptional dynamics and offer clues to how cells manage to minimize the accumulation of antisense RNA.

The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200

Methods

Datasets

Data from both NET-sequencing (NET-seq) and RNA sequencing (RNA-seq) for Saccharomyces cerevisiae were obtained from existing publications (Churchman & Weissman, 2011; Xu et al., 2009). NET-seq measures newly synthesized (i.e., nascent) polymerase-transcribed RNA while RNA-seq measures steady-state (i.e., mature) transcript levels. Both datasets provide strand-specific read counts that are mapped to either the sense (+) or antisense (-) strands corresponding to annotated genes. From 5331 total annotated genes, 2649 were considered to be actively transcribed based upon the detection of a sense (+) strand signal at or above the background level across both NET-seq and RNA-seq datasets. The remaining 2682 genes were considered non-transcribed based upon the lack of evidence for active transcription above background levels.

Data Processing

To assess the sense and antisense transcription from the NET-seq and RNA-seq datasets, the average of per-gene read counts based on the respective genomic coordinates of each gene (Cherry et al., 2012) was calculated and then $\log_2$-transformed. Before applying the $\log_2$-transformation, all genes received an additional value of 0.001 (pseudo count) to offset any mathematical errors that would arise. For each gene, the Sense/Antisense (S/A) ratio was calculated by dividing the mean sense read count by the mean antisense read count for both the NET-seq and RNA-seq datasets. Transcript stability was defined as the $\log_2$ ratio of the signal from each gene in RNA-seq to NET-seq data; $\log_2$[(RNA-seq+0.001) / (NET-seq

+0.001)] (the greater the $\log_2$ value, the greater the accumulation of steady state transcripts compared to nascent).

Statistical Analysis

To determine the correlation between NET-seq and RNA-seq sense/antisense signals, as well as between sense expression levels and S/A ratios, Pearson correlation coefficients were calculated. All tests were two-tailed. Summary statistics (median, mean, standard deviation) are reported. Statistical analyses were performed in MATLAB software.

Results

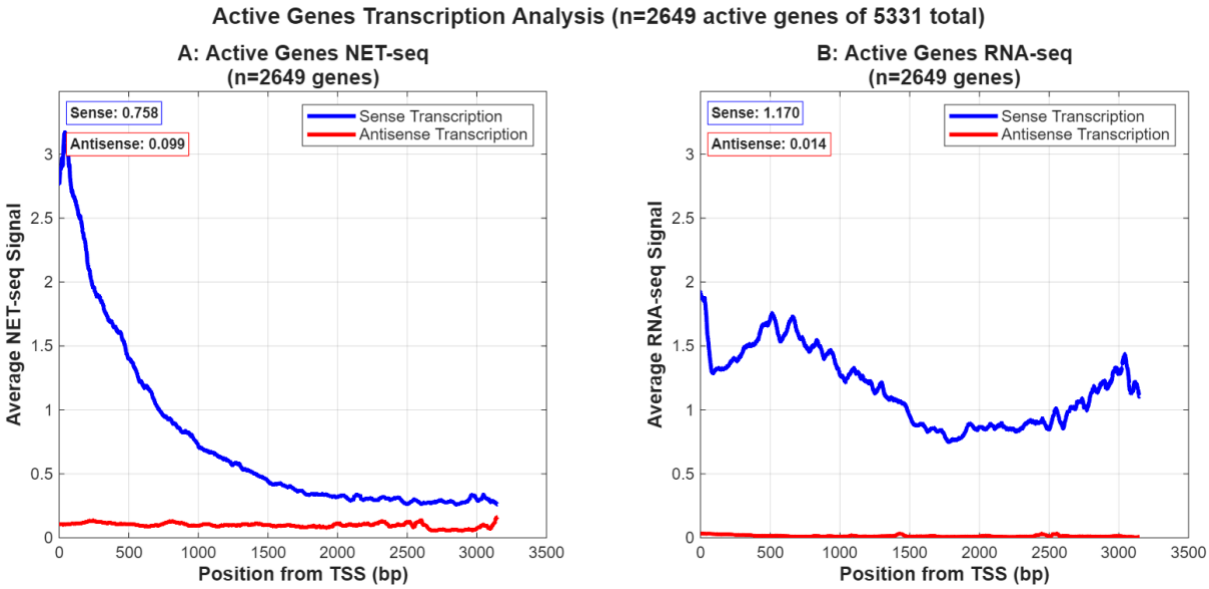
Antisense Transcription is a Near-Universal Feature of Active Genes

To characterize the overall pattern of transcription across active genes, the average sense and antisense signal was examined as a function of position from the transcription start site (TSS) in both RNA-seq and NET-seq datasets (Figure 1).

Figure 1

Active Genes Transcription Analysis (n = 2649 active genes of 5331 total)

The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200



Note: Average sense (blue) and antisense (red) transcription signal across 2,649 active genes as a function of position from the TSS. (A) NET-seq; (B) RNA-seq. Mean signal values are indicated in the inset boxes.

When analyzing the RNA-seq data, it was clear that the average sense signal (1.170) was grossly above the average antisense signal (0.014), with very little variation in the antisense signal across the length of the gene body. In contrast, the NET-seq data clearly demonstrated that the sense transcription is highest at the start of the transcription start site, where RNA polymerase is initially recruited to transcribe the gene, and steadily declines towards the end of the gene. This is consistent with known behavior whereby RNA polymerases pause before elongating transcription. Also, it is important to point out a large difference in means between the two datasets for the antisense signal; the antisense average in NET-seq (0.099) is higher than that in the RNA-seq antisense data (0.014) and is greater than zero at multiple points along the gene body (note that all comparisons use the same length of gene per data source). Further, this discrepancy is indicative of two key factors; 1) NET-seq

samples RNA being produced at a given point in time versus RNA-seq samples RNA indicates steady-state RNA within the cell; 2) antisense transcripts are being produced at an appreciable level as evidenced by their presence in NET-seq but do not have sufficient time to accumulate before they are degraded into the process of being formed into stable RNA within the cell.

Per-Gene Distribution of Sense and Antisense Transcription

To see how much variation there may be at the individual gene level beyond the average across all genes, the per-gene sense and antisense transcription in each of our datasets was calculated and their distributions were assessed (as shown in Figure 2).

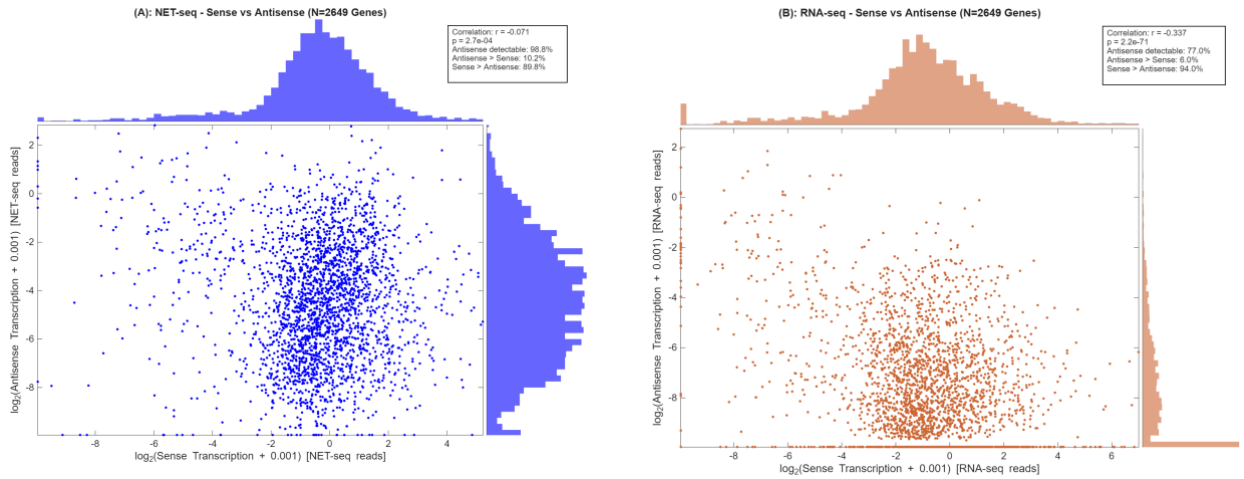
In the NET-seq dataset (Figure 2A), the antisense transcription had a wider distribution compared to the sense transcription, with many genes having expression closer to the detection limit of antisense expression. Antisense transcription was detectable in 98.8% of actively transcribed genes by NET-seq, confirming the broad prevalence of antisense transcription at the nascent level. For a small proportion of total genes (10.2%), the amount of antisense formed exceeded the amount of sense. Finally, the degree of correlation between net amounts of sense and antisense RNA was very low ($r = -0.071$; $p = 2.7e^{-4}$). This indicates that at the nascent RNA level, there is no trend of antisense transcription being proportionally related to sense transcription.

With regards to the RNA-seq dataset (Figure 2B), the antisense transcription had a highly skewed distribution compared to the sense transcription, with many genes having expression at the detection limit of antisense expression. Further, antisense transcription was detectable in only 77% of genes at steady state (compared to 98.8% seen in NET-seq),

reflecting post-transcriptional degradation of antisense transcripts. As was observed in NET-seq, a small proportion of total genes (6%) had higher antisense expression than sense. The correlation between sense and antisense transcription was negative ($r = -0.337$; $p = 2.2e^{-71}$), indicating that if a gene produced a large amount of sense RNA, it is not expected to produce a correspondingly large amount of antisense RNA. Thus, it is reasonable to conclude that the accumulation of antisense RNA is suppressed by the corresponding high transcriptional rate of sense RNA.

Figure 2

Sense vs Antisense Transcription (N = 2649 genes)



Note: Scatter plot of sense vs. antisense transcription for 2,649 active genes in (A) NET-seq and (B) RNA-seq. Marginal histograms show the distribution of each. Correlation statistics and percentage of genes where sense exceeds antisense are indicated.

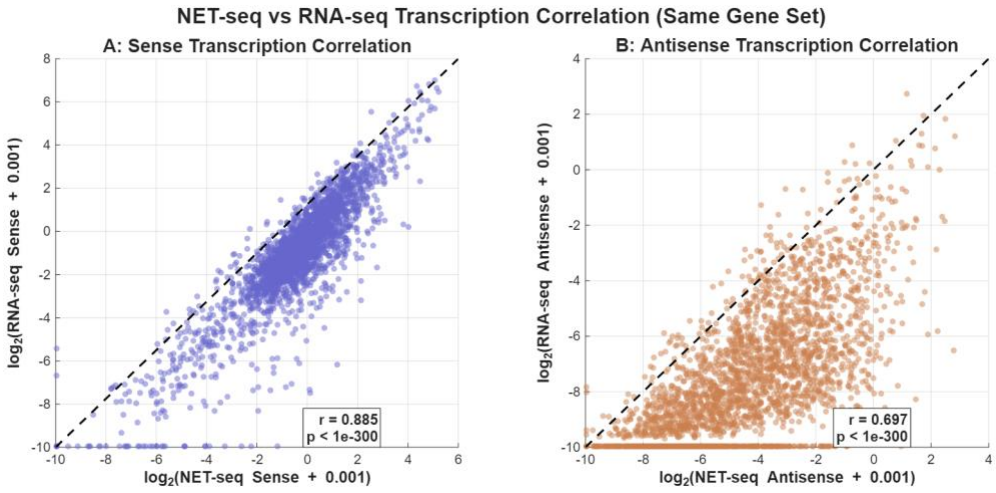
The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200

NET-seq and RNA-seq Are Concordant for Sense but Diverge for Antisense

Given that NET-seq and RNA-seq measure complementary aspects of gene expression, whether the two datasets agreed on the relative levels of sense and antisense transcription across genes was assessed (Figure 3).

Figure 3

Net-seq vs RNA-seq Transcription Correlation



Note: Correlation between NET-seq and RNA-seq for (A) sense transcription and (B) antisense transcription across 2,649 active genes. The dashed line indicates the identity (y = x) line. Pearson r and p-values are shown.

When looking for a correlation between NET-seq and RNA-seq for the nascent transcription of sense, the data showed a correlation coefficient (r = 0.885) at a highly significant level ($p < 1e^{-300}$). The high correlation indicated a strong linear relationship between sense nascent transcription and sense steady-state level of mRNA; genes with a high level of sense nascent transcription had a high sense steady-state level of mRNA. In contrast,

when considering the antisense transcription, although a statistically significant correlation was still evident ($r = 0.697$ and $p < 1e^{-300}$), the degree of correlation is much weaker than for sense. In addition, the data show that there was a consistent systematic offset below the identity line, such that RNA-seq antisense levels were consistently lower than NET-seq antisense levels. This indicates that although the antisense transcripts have a constant relative rate of production as measured by NET-seq across genes, they are subject to selective degradation that results in a substantially reduced steady-state level. The differences between the two datasets provide qualitative evidence that sense and antisense strands of the same gene are regulated differently after transcription and indicate that there are differences in the post-transcriptional regulatory methods that exist between the two strands. Therefore, the antisense transcription across the genome would be significantly underestimated if RNA-seq were used alone to estimate the level of antisense transcription because most of the signal of antisense transcription at the genome level is only found in the nascent state and cannot be detected in RNA-seq at steady state. Conversely, NET-seq alone is unable to determine which transcripts are accumulating at a steady-state level and which are degraded. The combination of both methods is therefore essential for a complete picture of antisense transcription.

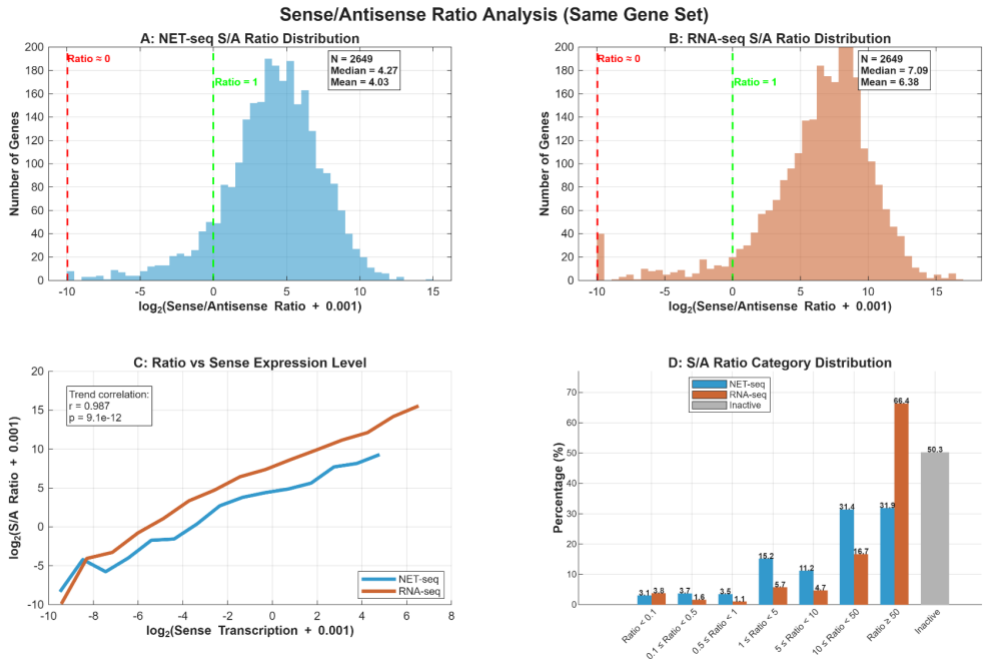
The Sense/Antisense Ratio is Higher in RNA-seq than NET-seq

To quantify the strand asymmetry more precisely, the sense/antisense (S/A) ratio for each gene in both datasets was computed and their distributions were compared (Figure 4).

The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200

Figure 4

Sense/Antisense Ratio Analysis



Note: Sense/antisense ratio analysis for the same set of 2,649 active genes. (A) NET-seq S/A ratio distribution; (B) RNA-seq S/A ratio distribution; (C) S/A ratio as a function of sense expression level; (D) categorical breakdown of S/A ratio ranges in NET-seq, RNA-seq, and inactive genes.

In both datasets, the S/A ratio distribution was positively and right-shifted (Figure 4A, 4B), indicating that the sense transcription of a gene predominates in the overall gene population; however, the median S/A ratio was significantly lower in NET-seq (median = 4.27, mean = 4.03) than that of RNA-seq (median = 7.09, mean = 6.38). This difference is indicative of an increased amount of antisense RNA being removed from the gene after transcription, as previously shown by Neil et al. (2009). The relative strength of sense versus

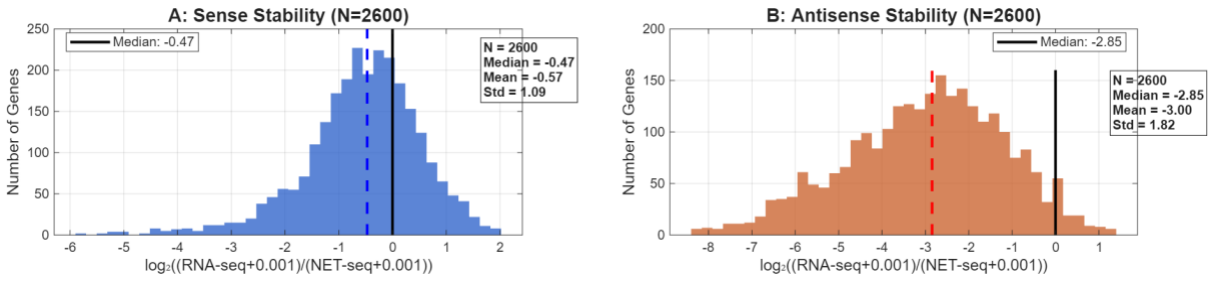
antisense transcription for all genes is shown in Figure 4D; for example, 66.4% of the genes had an S/A ratio in RNA-Seq ≥ 50, while only 31.9% of NET-seq genes had an S/A ratio that was ≥ 50. On the contrary, the percentage of genes with S/A ratios < 5 was greater in NET-seq than RNA-seq, reflecting a comparatively greater antisense signal compared to the number of sense signals at nascent levels. There is a strong positive correlation between the S/A ratio and the level of sense gene expression (Figure 4C; $r = 0.987$, $p = 9.1e^{-12}$), indicating that the genes showing the greatest amount of transcription also demonstrate the greatest sense bias, for both NET-seq and RNA-seq data.

Antisense Transcripts Are Substantially Less Stable than Sense Transcripts

The results above collectively suggest that antisense RNA is produced at the nascent level but then rapidly degraded. To test this directly, transcript stability as the $\log_2$(RNA-seq/NET-seq) ratio for each strand was estimated (Figure 5).

Figure 5

Sense and Antisense Stability



Note: Transcript stability was estimated by comparing steady-state RNA levels (RNA-seq) to actively transcribed RNA (NET-seq) using the formula $\log_2$ [(RNA-seq + 0.001)/ (NET-seq + 0.001)]. A positive value indicates a transcript is relatively stable and accumulates in the cell,

while a negative value suggests it is rapidly degraded before it can build up. This was calculated separately for (A) sense and (B) antisense transcripts across 2600 genes (49 genes outside the 0.5th-99.5th percentile range removed as outliers due to extreme stability values), with median values indicated by vertical dashed lines and summary statistics shown in inset boxes.

With respect to the sense transcripts (Figure 5A), the distribution of the stability values was approximately symmetric and slightly below zero (median = -0.47; mean = -0.57; SD = 1.09), which indicates that sense mRNA accumulates in a manner approximately proportional to their rates of synthesis with a slight negative net degradation. The distribution of the stability values for the antisense transcripts (Figure 5B) was sharply shifted towards a greater level of negative values than for the sense transcripts (median = -2.85; mean = -3.00; SD = 1.82). This very significant level of negative shift indicates that compared to their transcription rates, antisense transcripts accumulate to a much lower steady-state level, which is indicative of a very rapid degradation following transcription. The four-fold reduction in the median level (in $\log_2$ units) of the stability values for the antisense versus sense transcripts confirms that there are both active and specific mechanisms by which cells degrade antisense RNA, possibly via the exosome complex and/or other nuclear surveillance pathways.

Discussion

This study's results demonstrate that most active yeast genes (98.8% from NET-seq) undergo antisense transcription process. The antisense RNA, however, is rapidly degraded shortly after as evidenced by the markedly lower RNA-seq/NET-seq ratios observed for antisense transcripts compared to sense transcripts. There are several important implications

of this result. First, it appears that the antisense transcription process is a true biological phenomenon, not an accidental result of RNA polymerase activity. The fact that there is a dedicated machinery for degrading antisense RNA suggests that antisense transcription is expected and that the cell has developed means to deal with antisense transcripts on a frequent basis. The actions of the exosome and Rrp6 pathway for nuclear RNA surveillance and degradation of unstable transcripts (CUTs) in yeast is well-known; many of these transcripts come from genes that have no obvious, active transcription from the opposing DNA strand (Neil et al., 2009; Xu et al., 2009).

NET-seq and RNA-seq are highly correlated for sense transcription but have a lower correlation for antisense transcription; therefore, it is important to measure both nascent and steady-state transcription when analyzing the biology of non-coding RNA. Analyses using only RNA-seq will greatly underestimate the occurrence of antisense transcription by only analyzing those RNAs that do not degrade.

As sense transcription for genes increases, the corresponding ratio of sense to antisense also increases, which indicates that antisense transcription is not proportional to sense transcription. One explanation for this suppression of antisense transcription may be that when a gene is strongly transcribed in a sense direction, the act of strong sense transcription remodels the chromatin in such a way as to inhibit the initiation of antisense transcription via the directed modification of histones or repositioning of nucleosomes. Another possibility is that genes that have higher expression levels contain increased amounts of RNA polymerase (or other RNA surveillance factor) present, therefore making it easier for

the RNA polymerase (or other RNA surveillance factors) to clear any RNA transcripts produced in the antisense orientation.

One limitation of this research project is that it does not allow us to differentiate among the various ways that antisense RNAs can be produced and many of those types of producing mechanisms could be occurring from several locations within the genome. For example, antisense transcription may result from a process known as divergent transcription occurring at the 3' end of genes (where the transcriptional initiation for the sense strand stops and the transcription for the antisense strand begins), or from "read-through" transcription from upstream (convergent) genes, or from the presence of cryptic promoters that exist within the body of the gene itself. More work needs to be done to collect data regarding the binding of transcription factors (TBP, TFIIB) and the post-translational modification of chromatin (SWR1, Htz1) at antisense loci to be able to determine which of these mechanisms produce the majority of antisense RNAs and whether promoter-like elements exist at the 3' end of genes that are highly antisense-transcribed.

In conclusion, this study shows that antisense transcription is common in Saccharomyces cerevisiae, as well as active; however, much of that transcriptional output is heavily repressed shortly after its transcription has occurred. Thus, the cell produces antisense RNA while at the same time degrading the antisense RNA it produces — this balance could have been purposefully designed to regulate antisense RNAs, or could just be a byproduct of the fact that the DNA molecule is double-stranded and RNA polymerase cannot consistently transcribe from only one of the strands of the DNA molecule (Core et al. 2008).

The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200

Conclusion

Using matched NET-seq data and RNA-seq data from 2649 active (i.e., transcribed) yeast genes, we find that antisense transcription is a highly prevalent phenotype of gene expression. Despite that, antisense transcription represents a relatively low level of transcriptional activity with sense/antisense ratios consistently greater than one across both datasets. Sense transcription dominates over antisense transcription across both nascent and steady-state measurements, with significantly higher levels of sense transcription (particularly at the steady state). Antisense transcription could be detected in only 77% of the genes at steady-state, compared to 98.8% of the genes at the nascent level, meaning that antisense transcripts are lost more dramatically at steady-state than they are at the nascent level. Stability analyses directly corroborate this finding, where antisense RNAs displayed around 4 times less stability than sense transcripts based upon the median log2 ratios, suggesting that the rate and active degradation of antisense transcripts is the primary mechanism by which antisense transcript accumulation is limited. In conclusion, these studies expand upon previous observations regarding the prevalence of noncoding transcription; they reaffirm that noncoding transcription occurs with high frequency but is tightly regulated at the cellular level and suggest that the functional significance of noncoding transcription is dependent on the relative rates of transcription and degradation.

The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200

References

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Karra, K., Krieger, C. J., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., … Fisk, D. G. (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Research, 40(D1), D700–D705.

Churchman, L. S., & Weissman, J. S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature, 469(7330), 368–373.

Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science, 322(5909), 1845–1848.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., … Gingeras, T. R. (2012). Landscape of transcription in human cells. Nature, 489(7414), 101–108.

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

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

The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200

Acknowledgements

The author thanks Dr. Struan Murray for his guidance and for helping shape the research question addressed in this work.