1 Quality assessment of strand-specific RNA-seq

To test whether the amplification step preserves gene expression levels, we compared the expression levels of all protein coding genes in the unsorted whole root library without amplification to the libraries with amplification. Although aligned reads tend to be enriched towards the $3'$ ends of transcripts after amplification, we found high linear correlation of the expression levels between the amplified libraries unWR2/unWR3 and the unamplified library unWR1 (average $R^2 = 0.945$). This result suggests that the linear amplification protocol used to generate unWR2 and unWR3 faithfully retains relative gene expression levels at the genome scale.

We further compared the gene expression levels of the sorted libraries (WR1, WR2 and WR3) to the unsorted libraries (unWR2 and unWR3); we found a mere 6.5% of all protein coding genes in the Arabidopsis genome to be upregulated by the sorting process (> 2 fold change, FDR adjusted $p-value < 0.01$ using DESeq [1]). This is consistent with the previous observation that most genes are not affected by the cell-sorting process [2]. The high quality of our RNA-seq data is also supported by the strong pairwise correlation between biological replicates (average $R^2 > 0.97$ for all 9 pairwise comparisons). We further compared gene expression levels from our RNA-seq data to published cell type-specific expression levels using Affymetrix microarrays [3]. For each protein coding gene that is on the ATH1 22K microarray, we calculated the average expression level of that gene from the replicates of the microarray and from the replicates of the RNA-seq data. For both CORT and ENDO libraries, we found high reproducibility of the gene expression levels across platforms (linear regression, $R^2 = 0.79$ and 0.71 respectively).

2 Statistical model of antisense transcription.

2.1 Derivation of the sense only model M_0

Because our method is evaluated on one locus at a time, we drop the index i for genes and j for replicates in this section to simplify the model notation and the explanation. We denote read count from the forward strand as N^+ and read count from the reverse strand as N^-. Given the observed number of reads from forward strand and reverse strand (N^+ and N^-, $N = N^+ + N^-$), the goal of our inference is to compare two possible models that can explain the observed read counts. If we assume only one gene is transcribed in the locus of interest on the forward strand, the probability of the observed read counts can be calculated by the sense only model (M_0):

$$\begin{aligned}\Pr\left(N^+, N^-\right) &= \Pr\left(N^-, N\right) \\ &= \mathcal{B}(N^-|N, pe) \times \mathcal{NB}(N|\mu, \sigma) \qquad (1)\end{aligned}$$

Because the strand-specific protocol generates a certain percentage of reads from the unexpected strand of a gene, N^- is not always equal to zero. We suppose that the N^- is distributed according to a binomial distribution ($\mathcal{B}(N^-|N, pe)$), where *pe* represents the probability of observing reads from the unexpected strand (called protocol error rate or PE in the main text), and is estimated as the average PE across all genes. The *pe* equals 0.5 if the reads generated from the sense strand gene have equal probability to be mapped to the sense strand or antisense strand. For a strand-specific protocol, *pe* should always be smaller than 0.5. The parameters for the negative binomial part of the equation is estimated in the same way as in DESeq package [1]. In DESeq, the expected count (μ) was estimated by averaging the normalized read counts. The variance (σ) was estimated by fitting a generalized linear model.

2.2 Derivation of the antisense model M_1

When a pair of antisense genes are transcribed in one locus, all the reads that mapped to the sense strand are generated from two sources. Some of the sense reads are from the sense strand gene; others are from the antisense strand gene but mapped to the sense strand due to the non-perfect strand specificity. We introduce N_s and N_a to denote the number of reads originated from sense transcription and antisense transcription respectively. We also use $N_s^+, N_s^-, N_a^+ and N_a^-$ to denote the read counts from forward and reverse strands for sense and antisense transcriptions respectively. These variables are not observed but are related to the observed read count N^+ and N^- by the following equations:

$$
\begin{aligned}
N_s &= N_s^+ + N_s^- && (2)\\
N_a &= N_a^+ + N_a^- && (3)\\
N^+ &= N_s^+ + N_a^+ && (4)\\
N^- &= N_s^- + N_a^- && (5)
\end{aligned}
$$

The probability of the observed data ($N^+ and N^-$) given the antisense model (M_1) is expressed as (M_1 is dropped for clarity):

$$
\begin{aligned}
\Pr(N^+, N^-) &= \sum \Pr(N_s^+, N_s^-, N_a^+, N_a^-)\ \textit{over all possible}\ N_s^+, N_s^-, N_a^+, N_a^- && (6)\\
&= \sum \Pr(N_s^-, N_s, N_a^+, N_a)\ \textit{over all possible}\ N_s^-, N_s, N_a^+, N_a && (7)\\
&= \sum_{N_s^-=0}^{N^-} \sum_{N_a^+=0}^{N^+} \Pr(N_s^-, N_a^+, N_s, N_a) && (8)
\end{aligned}
$$

The probability of the observed data ($N^+\, and\, N^-$) equals the sum of probabilities of missing data over all the possible values of the missing data (equation 6). Because of the linear constraints (equations 2 and 3), we can write the the probability in equation 6 as the probablity in equation 7. In fact, only two non-complementary missing data N_s^- and N_a^+ are needed to calculate the values of the other two missing data (N_s and N_a). One can choose other pairs of missing data, for example, N_s^+ and N_a^-, for the optimization process. However, because *pe* is always smaller than 0.5, the expected values of N_s^- and N_a^+ are smaller as compared to other possible pairs. Therefore, we decide to optimize the model using N_s^- and N_a^+. These steps are similar to the derivation of the sum of Poisson random variable with the total count as a constant: one first decides the joint distribution of all random variables and then substitutes some of the random varibles using the known constraints.

Following the model assumption in 1, we can write the joint distribution of the missing data ($N_s^-, N_a^+, N_s \, and \, N_a$) under an antisense model as:

$$\begin{aligned} \Pr(N_s^-, N_a^+, N_s, N_a) &= \mathcal{B}(N_s^-|N_s, pe) \times \mathcal{NB}(N_s|\mu_1, \sigma_1) \\ &\times \mathcal{B}(N_a^+|N_a, pe) \times \mathcal{NB}(N_a|\mu_2, \sigma_2) \end{aligned} \quad (9)$$

We call this model as the antisense model (M_1). To determine the significance of antisense transcription, we compare the sense only model (M_0) to the antisense model (M_1). The parameters for the Binomial and Negative Binomial distributions are estimated as described above and treated as known in the model comparison. In an ideal situation, one would calculate the sum in (8), and compare the two models using Bayes Factor, which is the ratio of probabilities of the data given two models ($\Pr(N^+, N^-|M_1)/\Pr(N^+, N^-|M_0)$). When working on actual RNA-seq data sets, however, thousands of reads could map to one gene; the summation of many small probabilities is slow to calculate and may cause numerical error. Instead of calculating the summation, we approximate the probability of the observed data using maximum likelihood estimation of N_s^- and N_a^+. This is essentially using Bayesian information criterion to approximate the Bayes Factor by treating the two missing data N_s^- and N_a^+ as parameters. We call this summary statistics the NASTI score.

$$\begin{aligned} NASTIscore &= log(\Pr(D|\hat{N_s^-}, \hat{N_a^+}, M_1)) - log(\Pr(D|M_0)) \\ &- \frac{1}{2} \times (d_1 - d_0) \times log(n) \end{aligned} \quad (10)$$

We performed simulation studies to compare the difference between log(BF) and NASTI score using model M_1. In the simulations, we first set $N^+ = 20, 50, 100, 200$ and $pe = 0.01, 0.02, 0.05, 0.10, 0.20$. We then choose N^- such that the ratio $N^-/N^+ = 1\%, 2\%, 5\%, 10\%, 20\%, 50\%, 100\%, \, and \, 200\%$. We found that both log BF and NASTI score increase very fast as the number of reads on the reverse strand increases (Figure 1). The value of log BF and NASTI score are similar; the absolute value of the difference is very small as compared to the magnitude of log BF or NASTI score (Figure 2).

The simulations were performed over a wide range of parameters, and we found that the differences are small in all simulations (Figure 3).

2.3 Model formulation with multiple genes and replicates

We denote the number of mapped reads in replicates i, gene j and strand k as N_{ijk}, where $i = 1, \cdots, I, j = 1, \cdots, J$ and $k = 0 \, or \, 1$. We use $k = 0$ to indicate that the reads are mapped to the expected strand, whereas $k = 1$ indicates the reads are mapped to the unexpected strand. Because in the commonly used read mapping data format (bam files [4]), the direction of a read is always represented according to the strand of the genome. When counting reads using bam files, the value of k is determined by 1) which strand of the genome the reads mapped to and 2) the strand of the gene of interest. For one example, given a gene on the forward strand, $N_{i,j,k=0}$ is equvilent to N^+ and $N_{i,j,k=0}$ is

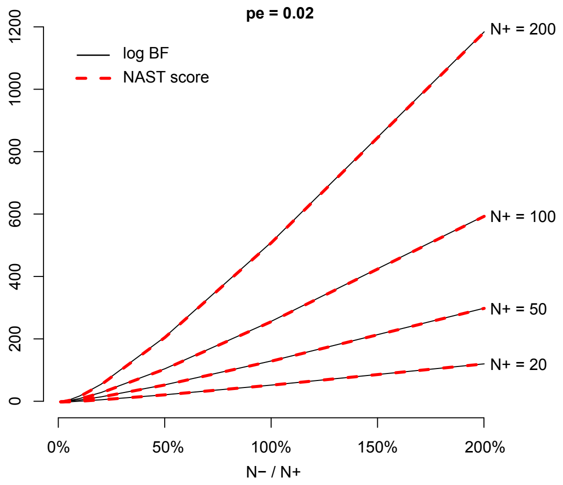

Figure 1: Comparison of the full Bayes Factor and the NASTI score

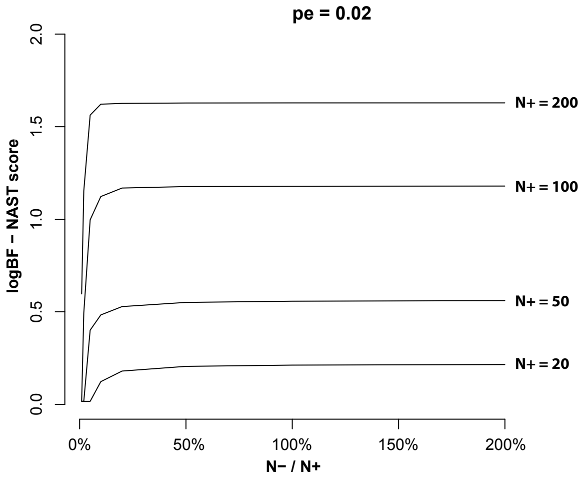

Figure 2: Difference between the full Bayes Factor and the NASTI score

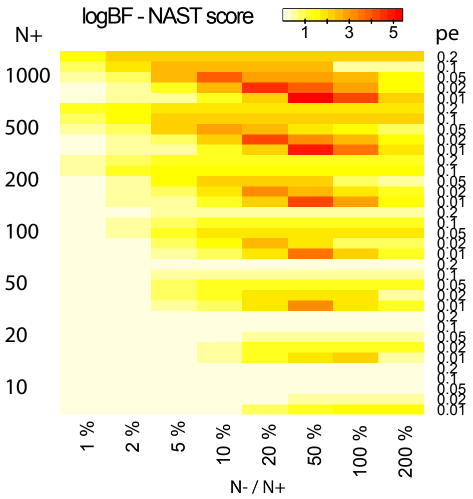

Figure 3: Difference between the full Bayes Factor and NASTI score over a range of parameters. The simulations were performed by varying N^{+} and pe. N^{-}s were choosen according to the ratio N^{-}/N^{+}

equvilent to N^{-} in the simplified model in section 2.1. For another example, if a gene "g" is on the reverse strand of the genomic DNA, and we found 10 reads mapped to the forward strand of the genomic DNA of gene "g" in replicate "r". In this situation, we assign $N_{i=r,j=g,k=1} = 10$, because for an antisense gene, the unexpected strand is the forward strand of the genomic DNA. We denote $N_{ij} = \sum_k N_{ijk}$ as the total number of reads mapped to gene j under condition i, which is equivalent to N in the simple model 2.1. We model this number using a negative binomial distribution ($N_{ij} \sim NB(\mu_{ij}, \sigma_{ij})$) [1]. We also model $N_{i,j,k=1}$ according to a binomial distribution ($N_{ij1} \sim B\left(N_{ij}, pe_{ij}\right)$), where pe_{ij} represents the protocol error rate. The pe_{ij} for each gene is calculated as the fraction of unexpected reads from that gene, and the average PE across all genes under one condition is used as the estimated pe_i for each gene ($pe_{ij} = pe_i$). For any strand specific protocol, pe_i would always be smaller than 0.5 by definition.

For the sense only model (M0), the likelihood of the observed read count is the following.

$$\Pr(N_{ijk}) = \mathcal{B}(N_{ij1}|N_{ij}, pe_i) \times \mathcal{NB}(N_{ij}|\mu_{ij}, \sigma_{ij}) \tag{11}$$

The binomial term in this equation can also be written as $\mathcal{B}(N_{ij0}|N_{ij.}, 1-pe_i)$ because of complementarity. For the antisense model (M1), we introduce N_{ijks} to denote the unobserved read counts, with s as an indicator variable for the strand of the underlying gene that generates the read counts. If the reads are from the sense strand gene, then s equals 0, otherwise, s equals 1. For example, N_s^{-} in the 2.1 is equivalent to $N_{i,j,k=1,s=0}$, where $s = 0$ indicates the gene is from sense strand and $k = 1$ indicates the reads are from the unexpected strand (reverse strand). The likelihood of the unobserved data is then written as:

$$\Pr(N_{ijks}) = \prod_{s=0,1} \mathcal{B}(N_{ij1s}|N_{ij.s}, pe_i) \times \mathcal{NB}(N_{ij.s}|\mu_{ij.s}, \sigma_{ij.s}) \tag{12}$$

The likelihood in (12) is maximized by iteratively estimating the unobserved data N_{ijks} and the parameters for the negative binomial distribution. The expected number of antisense reads coming from the sense gene was initialized as $\hat{N_{ij10}} = N_{ij.0} \times pe_i$. $\hat{N_{ij11}}$ is initialized in a similar fashion. At each iteration, new $\hat{N_{ij1s}}(s = 0, 1)$ were found as one of the closest integers to the $\hat{N_{ij1s}}$ in the previous step such that the likelihood increases. The algorithm converges very fast (usually less than 10 steps, when the improvement of the likelihood is smaller than a given threshold ($\sigma = 1e - 20$)). After the algorithm converges, the NASTI score is calculated based on (10).

To demonstrate the effectiveness of our method in identifying cis-NAT pairs, we made use of an existing annotation of natural antisense pairs [5], in which more than 2000 pairs of anti-sense genes were identified by analyzing EST databases. Among these pairs, we found 874 cis-NAT pairs that are still supported by the latest Arabidopsis annotation (TAIR10) as overlapping and antisense transcripts. These 874 cis-NAT pairs were used as the positive training set. Newly annotated natural antisense pairs in TAIR10 were kept as an independent validation set for the downstream analyses. To generate a negative training set, we scanned the TAIR10 annotation for pairs of genes that are unlikely to form natural antisense pairs. We first identified all groups of four genes such that the four genes are located next to each other along the genome. For each group of four genes, we require that either end of the second and third genes are at least 500 base pairs away from each other and are at least 500 base pairs away from the ends of the first gene and the fourth gene. We found 3972 groups that satisfied our criteria, and we used each second gene and third gene as negative training pairs. Both positive and negative training pairs were required to have at least 5 reads from each gene. The criteria for selecting training sets are also listed in Supplemental Table 3.

The model was first trained on data from all loci to calculate the NASTI score for each locus, and then the maximum NASTI score for each pair of locus was used as a score to classify positive training pairs against the negative training pairs. Ten-fold cross validation was carried out and the model was evaluated by receiver operation characteristic curve (ROC curve, Figure 1B) and the area under the ROC curve (auROC, Figure 1C) in the main text.

3 Logistic regression

A logistic regression model was fitted to the read count data to predict antisense transcription, using the following equation: $\Pr(Y_i = 1|X_i) = logit^{-1}(X_i)$

For each sense strand gene i, the numbers of the sense strand reads and the antisense strand reads from each replicate were used as predictor variables (represented by the vector X_i). The indicator variable, Y_i, equals one if the gene is in the positive training set, and equals zero if the gene is in the negative training set. Model parameters were estimated using R (www.cran.org). In the classification step, for each pair of genes in the training set, the final score is the maximum score ($logP(Y_i = 1|X_i)$) of the two genes.

References

[1] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010.

[2] K. Birnbaum, D. E. Shasha, J. Y. Wang, J. W. Jung, G. M. Lambert, D. W. Galbraith, and P. N. Benfey. A gene expression map of the arabidopsis root. *Science*, 302(5652):1956–60, 2003.

[3] S. M. Brady, D. A. Orlando, J. Y. Lee, J. Y. Wang, J. Koch, J. R. Dinneny, D. Mace, U. Ohler, and P. N. Benfey. A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science*, 318(5851):801–6, 2007.

[4] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and Subgroup Genome Project Data Processing. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9, 2009.

[5] X. J. Wang, T. Gaasterland, and N. H. Chua. Genome-wide prediction and identification of cis-natural antisense transcripts in arabidopsis thaliana. *Genome biology*, 6(4):R30, 2005.