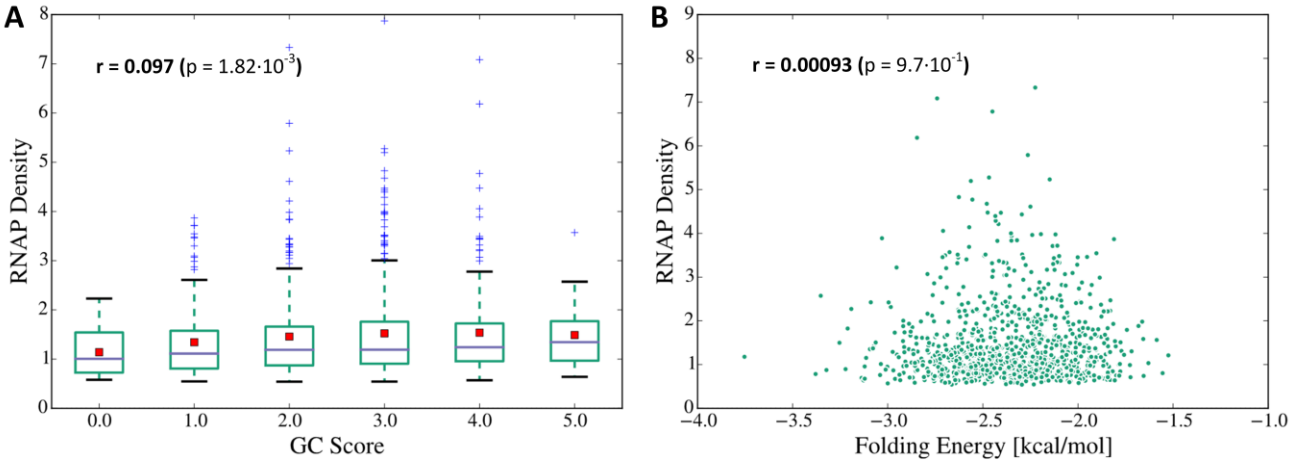
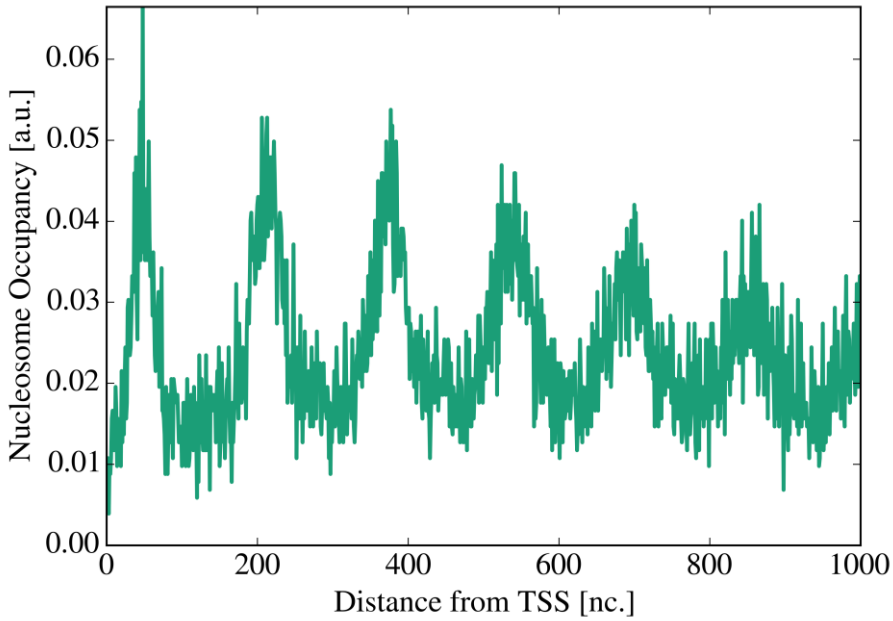
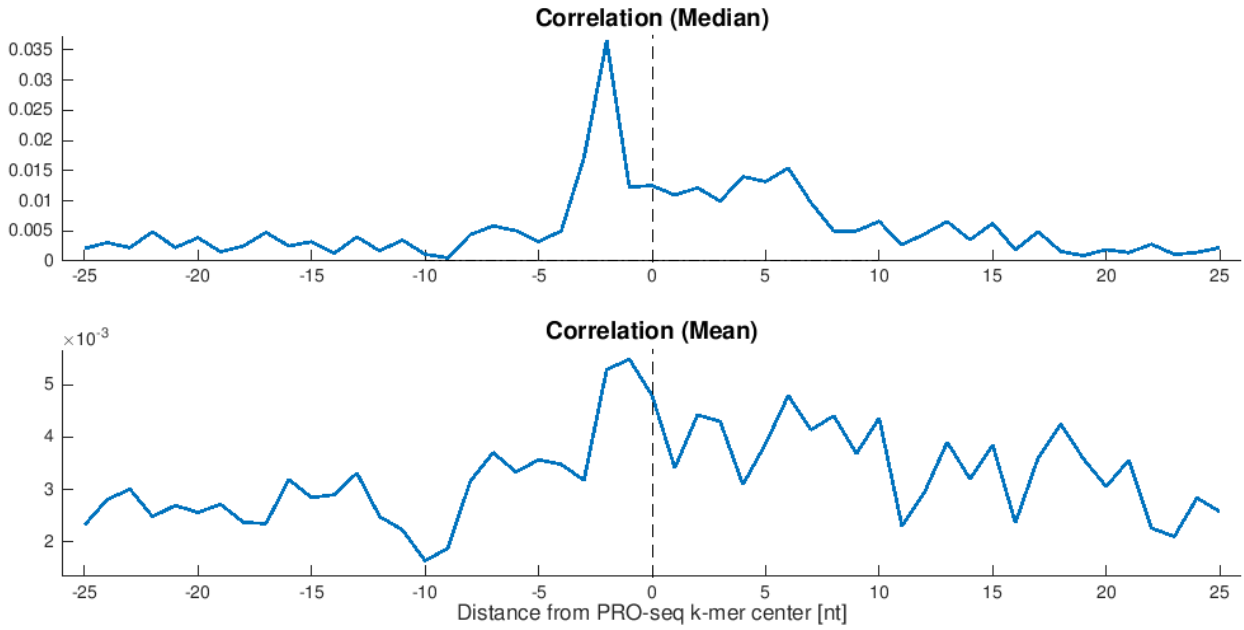
Supplementary Figures



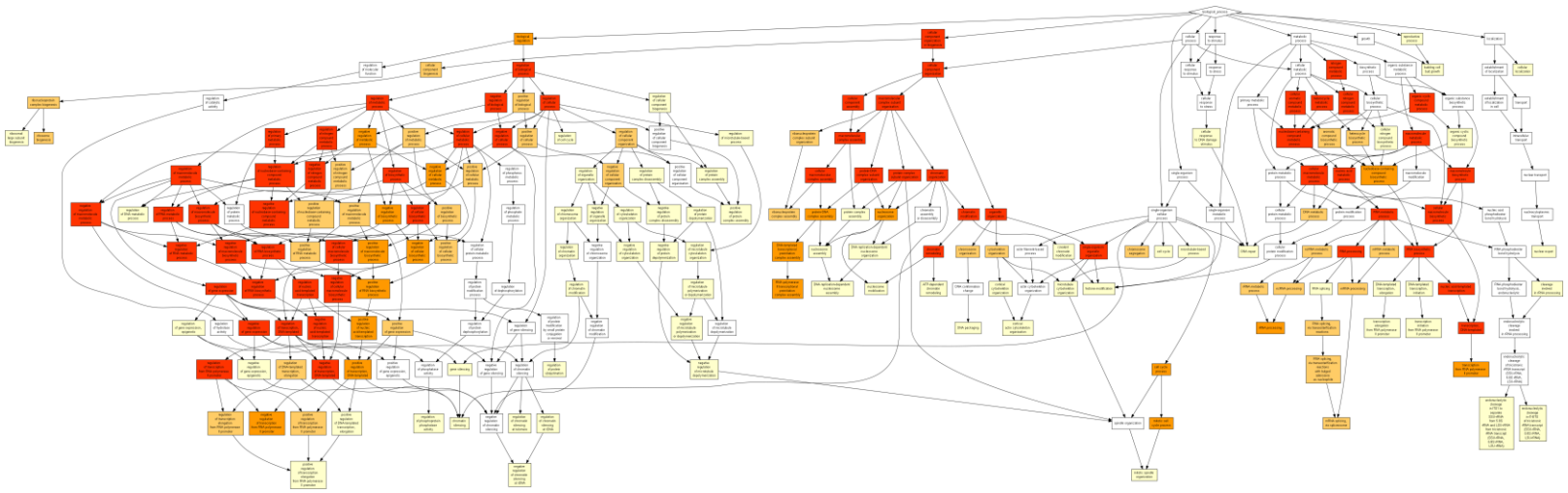
Supplementary Figure S1: Distribution of RNAP density score *vs.* GC content and local DNA folding energy. A) RNAP density score distribution (box-plot) for each of the possible 5-mer GC scores; the x-axis 0 through 5 is the number of G/C in the 5-mer. B) Scatter plot of RNAP density score *vs.* folding energy. For both panels, the Spearman correlation is placed at the top.



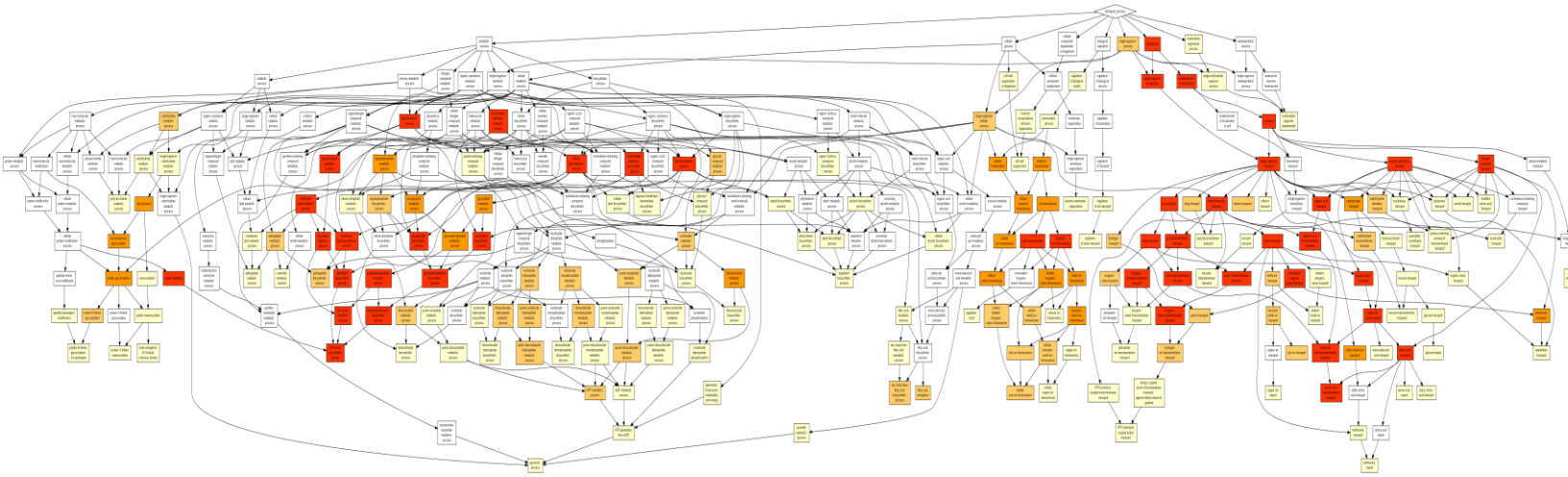
Supplementary Figure S2: Nucleosome occupancy profile. Nucleosome occupancy profile of 4232 genes aligned at the TSS.



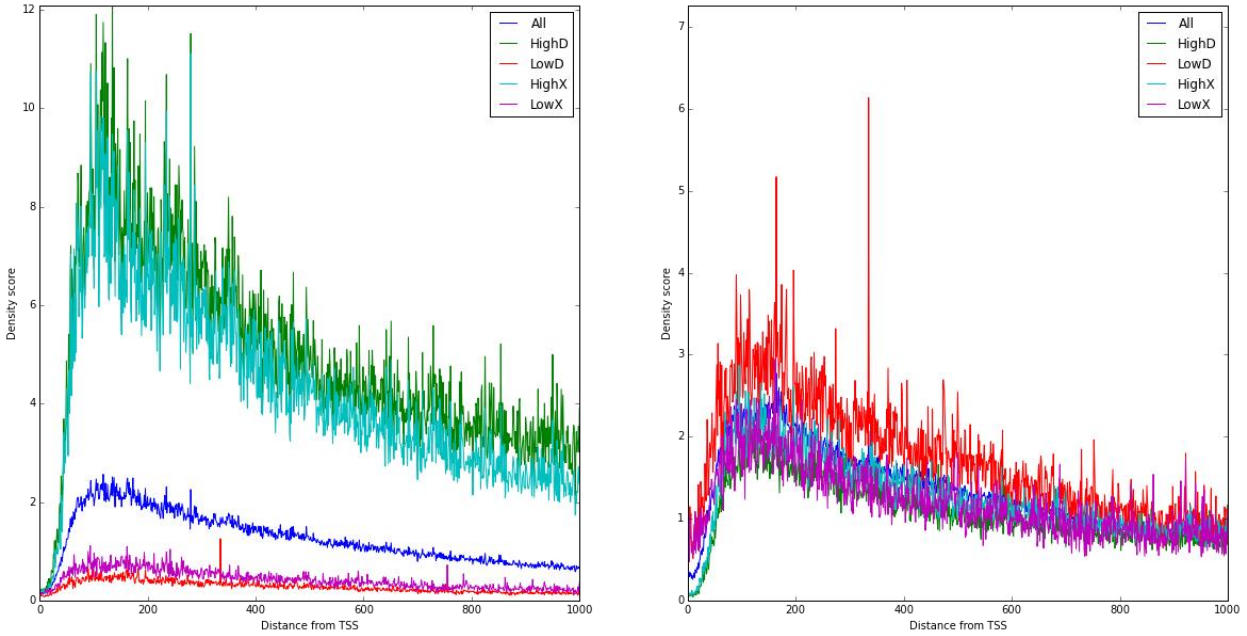
Supplementary Figure S3: PRO-seq *vs.* NET-seq Correlation. The RNAP read counts were compared between 1056 genes with the highest average RNAP in both datasets, and in a single nucleotide resolution. The median correlation was determined to be the highest in the position with lag of -2 ($r=0.0365$; $p=6.9\cdot10^{-8}$). Results for mean correlation also show relatively high value in this location.



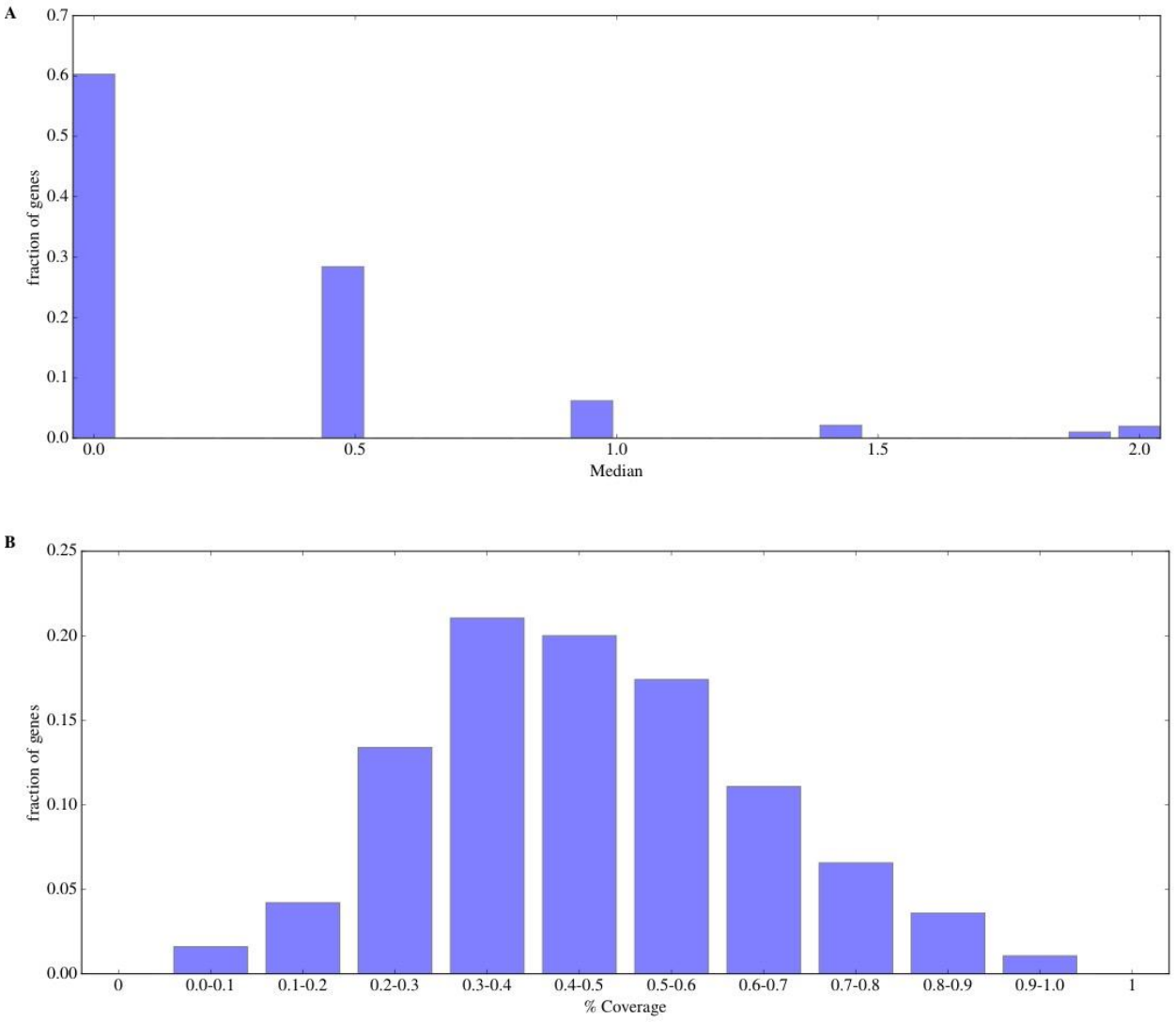
Supplementary Figure S4: GO Enrichment DAG for genes ordered by decreasing MTTR order.



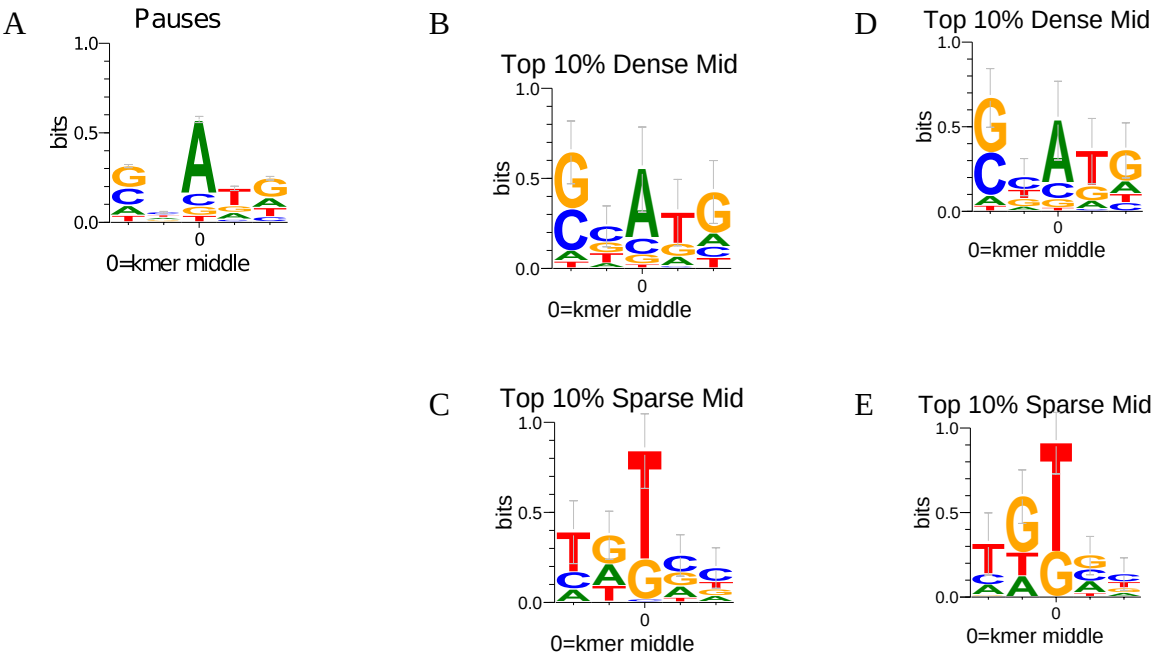
Supplementary Figure S5: GO Enrichment DAG for genes ordered by increasing MTTR order.



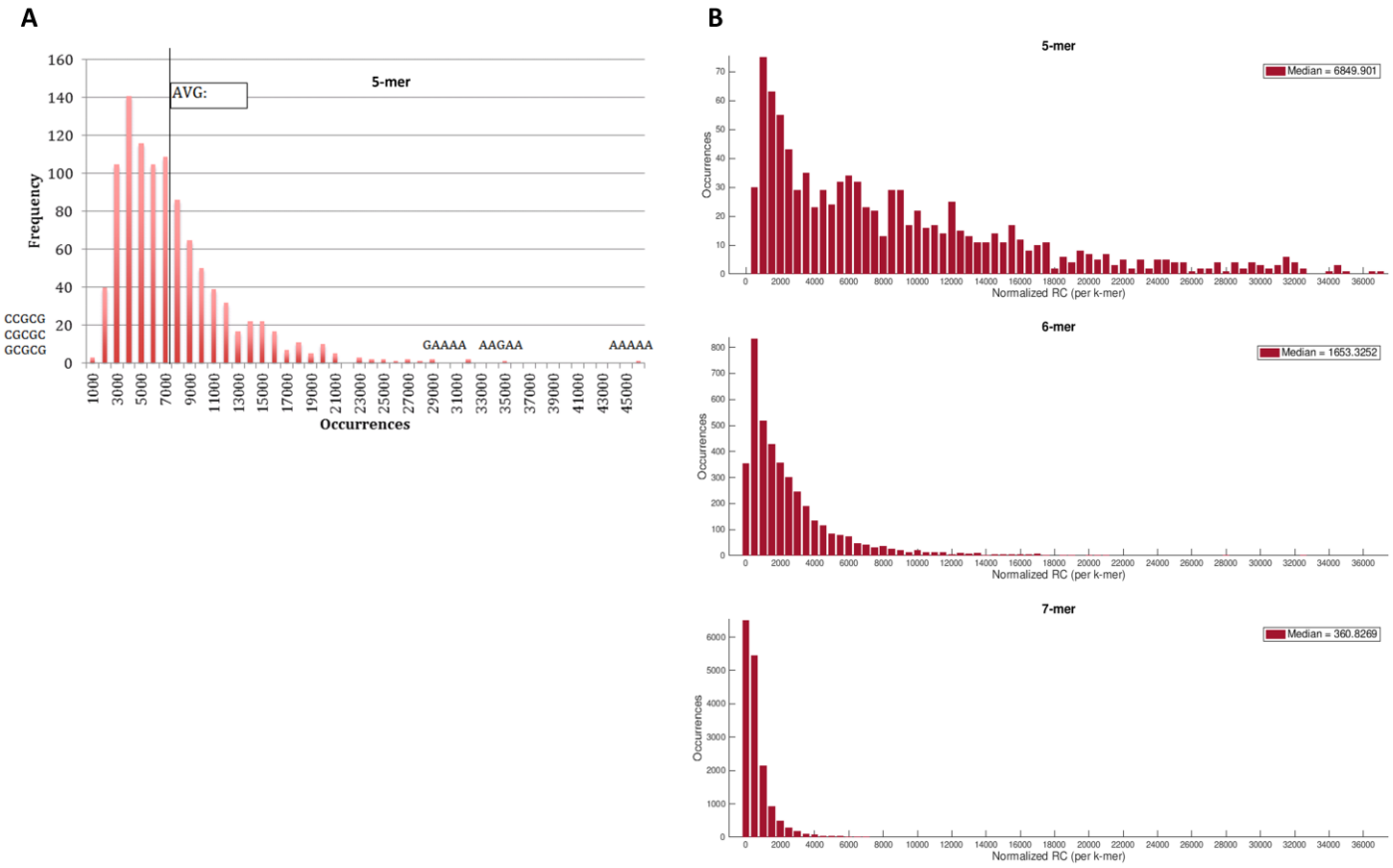
Supplementary Figure S6: Aggregate RNAP density profile for different gene groups. There are five profiles: all genes (All), high(High D)/low(Low D) density and high(High X)/low(Low X) mRNA level. The number of genes was used is 4232. 10% of the genes were used for each sub-group. On the right the same data, but normalized by the average density per gene.



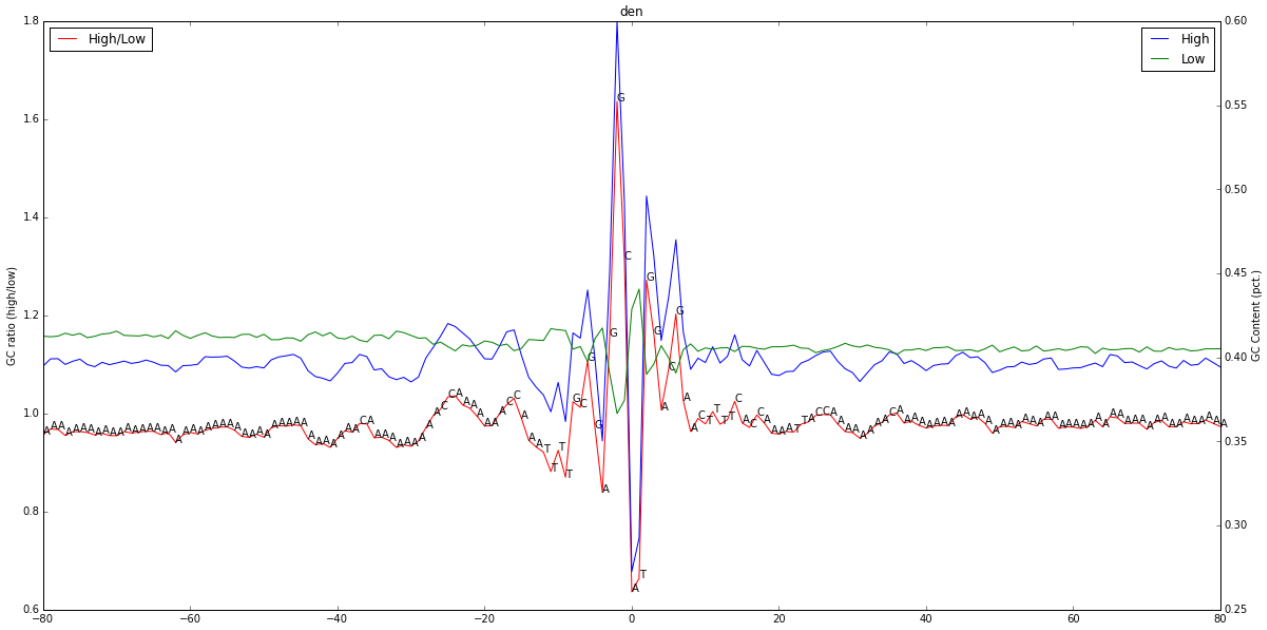
Supplementary Figure S7: RNAP reads per gene statistics. A) Histogram of the median read per gene. B) Histogram of the % coverage of the genes.



Supplementary Figure S8: PSSM of different groups of 5-mers. A) PSSM of pauses. B-C) Top 100 dense (sparse) 5-mers with pauses D-E) Top 100 dense (sparse) 5-mers without pauses.

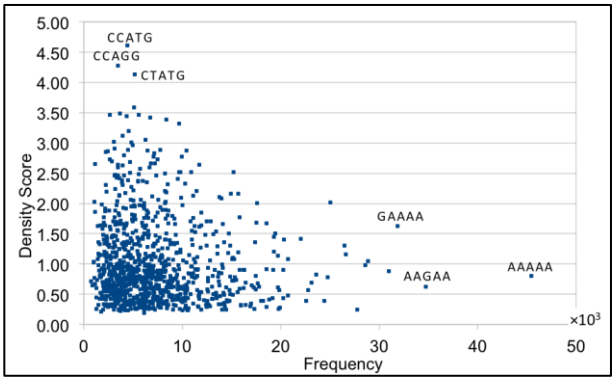


Supplementary Figure S9: A) Histograms of the 5-mers across all the genes. The X-axis shows the number of occurrences. The Y-axis is the number of k-mers having similar number of occurrences. Examples of the 3 most and least frequent k-mers can be viewed at their proper place in each panel. B) The Normalized Read Count (NRC) per k-mers covered in the NET-seq experiment (k = 5, 6, 7). There are significantly more NRCs related to the 5-mers in comparison to 6-mer and 7-mers. The 5-mer with lowest number of reads has 393.07 NRCs while the 6-mer with the lowest number of NRCs has 34.68 and six 7-mers having NRC value of zero. Results were similar in the PRO-seq experiment (not shown).

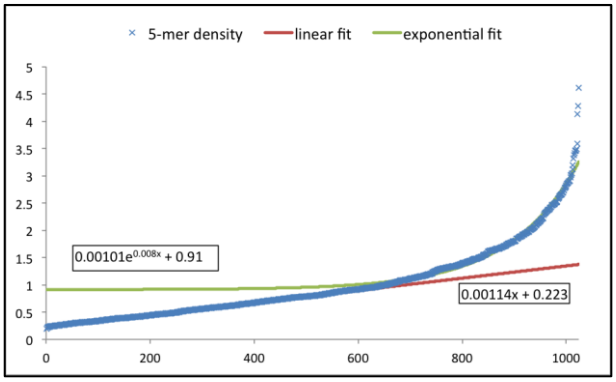


Supplementary Figure S10: GC landscape of relative high(blue)/low(green) density positions (ratio in red). All the genes TSSs were aligned. For each position we take the genes with the top 10% highest (lowest) density at that position and record the 80 nucleotides upstream and downstream. We end-up with two groups (H, L) of sequences (each is 161 nucleotides long). The GC content at each position in the H group (blue) and L group (green) is plotted as well as their ratio (red). The most common nucleotide at each position is indicated as well.

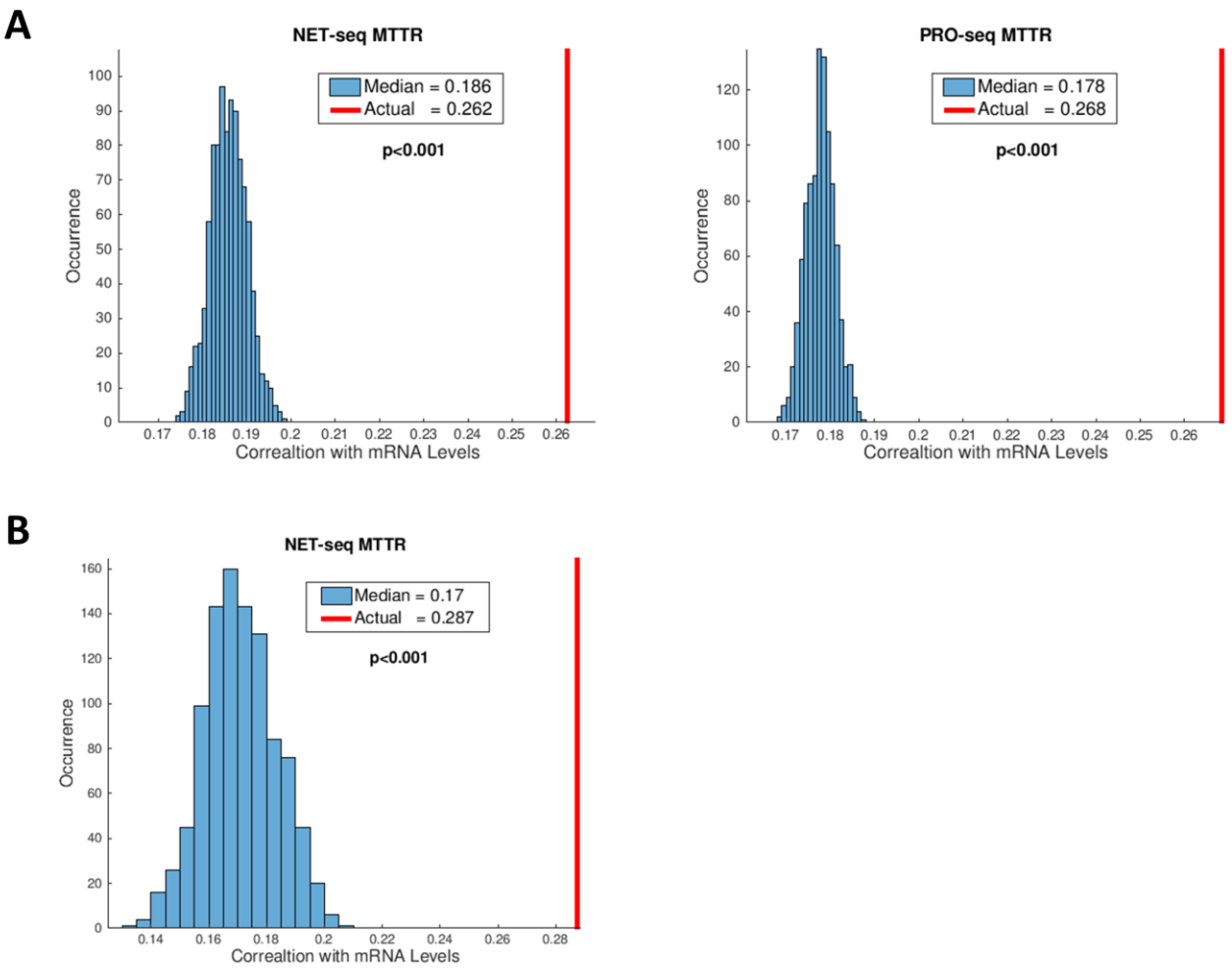
A



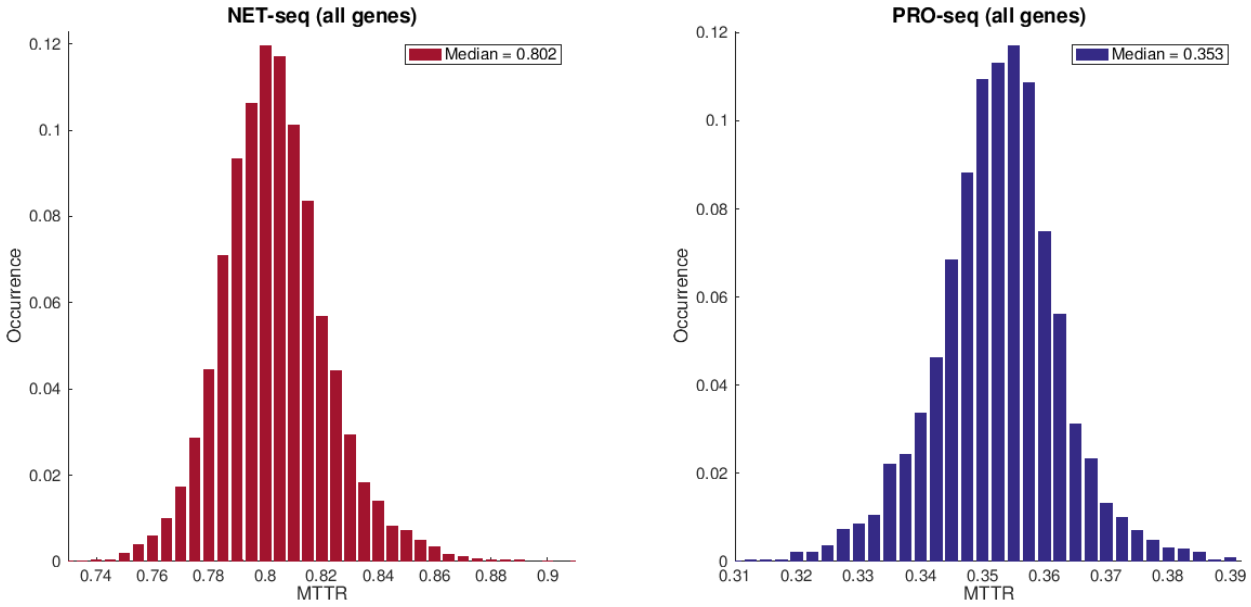
B



Supplementary Figure S11: 5-mers density distribution. A) 5-mer density scores as a function of 5-mer frequencies. B) The 5-mers were sorted in ascending density score order. The linear fit is based on the 562 lowest density 5-mers. The exponential fit is based on the rest of the 5-mers (462).



Supplementary Figure S12: Correlation distribution of the randomized models for MTTR with the mRNA levels for NET-seq and PRO-seq. A) Based on the 3rd random model, the correlation is lower than the actual value: Left) NET-seq: actual=0.262, random median=0.186. Right) PRO-seq: actual=0.268, random median=0.178; empirical $p<1\cdot10^{-3}$. B) Based on the 5th random model, the correlation is lower than the actual value. NET-seq: actual=0.262, random median=0.186; results remain consistent based on the 4th random model that randomize transcripts while preserving the ORF's CUB (not shown). MTTR/mRNA values are for all genes.



Supplementary Figure S13: Histograms of the MTTR (Mean of Typical Transcription-elongation Rate) values for all *S. cerevisiae* genes; NET-seq (Left) and PRO-seq (Right).

Supplementary Notes

Supplementary Note 1 - On the Choosing of 5 Nucleotide k-mer

In general, if enough statistics exist, longer k-mers potentially contain more encoded information then shorter ones and are therefore more preferable. On the other hand, models based on too long k-mers can be problematic, noisy, and non-statistically robust - since some of them do not appear in the genome or repeat a very few times in the genome.

We decided to work with 5-mers due to the following reasons:

1) There are significantly more normalized read counts (NRCs) and thus robust statistics related to the 5-mers in comparison to 6-mer and 7-mers. For example the median NRCs per 5-mer for NET-seq is 6849.9 while it is 1653.33 and 360.83 for 6-mers and 7-mers, respectively. The 5-mer with lowest number of reads has 393.07 NRCs while the 6-mer with the lowest number of NRCs has 34.68 and six 7-mers having NRC value of zero; see Supplementary **Figure S9B**.

2) The results obtained for 6-mers and 7-mers are similar to the ones obtained for 5-mers; thus, with agreement with the Occam's razor principle we decided to work with the simplest model. For example, we calculated the k-mer density scores for both 6-mers and 7-mers. Following we calculated their respective MTTR scores and found the correlation between their MTTR scores and the 5-mer MTTR to be 0.968/0.947 when analyzing the NET-seq experiment and 0.99/0.947 when analyzing the PRO-seq experiment (respectively).

3) The k-mers potentially capture various types of phenomena and information such as interaction between RNAP and the DNA, interaction between pre-mRNA and DNA, interaction between the RNAP and nucleosomes, folding of RNA and DNA molecules, interaction with other transcription factors, etc. Thus, the DNA-RNAP surface area is probably not the only relevant parameter. In addition, to the best of our knowledge, the effective RNA-DNA surface interaction length (*i.e.* the length that is relevant to its movement) is unknown and non-trivial to measure.

Supplementary Note 2 - Analysis of Highly an Lowly Expressed Genes

We examined the k-mer density scores while separately learning from highly and lowly expressed genes (based on their corresponding mRNA levels), and ranked them based only on the data of the highly/lowly expressed genes, respectively. Specifically, we looked at genes with sufficient read count coverage and generated equally sized groups of highly and lowly expressed ones. We found that the ranking between the two group's k-mer scores was very high for both NET-seq and PRO-seq ($r>0.94$, $p<1\cdot10^{-324}$; Spearman correlation). Furthermore, when learning from different parts of the genes (in a similar manner to the 2nd random model shown in **Figure 2C-D**), we found the Spearman correlation between highly and lowly to be higher than 0.91 ($p<1\cdot10^{-324}$). Finally, the correlation between the corresponding MTTR scores over all the genome, calculated by these sets was higher than 0.98/0.95, respectively ($p<1\cdot10^{-324}$).

Thus, these results suggest that the typical elongation rate is similar in highly and lowly expressed genes and cannot be trivially explained by biases related to the higher number of reads in highly expressed genes.

Supplementary Note 3 - On the Coefficient of Variation Analysis

The CV analysis demonstrates that k-mer variance is generally very low; this suggests that the differences among the k-mer scores are not due to variability or noise in the data. Indeed there are 26 5-mers that are very noisy (*i.e.* the only ones with significant high variation in comparison to a random model and thus no signal at all). One explanation for this group may be related to the fact that they have less reads: the average read count is 787 in comparison to 1,041 for the significant ones; thus, they are actually the ones that are less biologically interesting. The other 998 k-mers (out of 1,024) show a consistent typical density/speed of the RNAP; for instance, 741 out of them have variance significantly lower than the randomized model.

Supplementary Note 4 - Nucleotide Composition Related to 5-mer with High and Low Density

The top 3 most frequent 5-mers (between 33 to 46 thousands occurrences) and top 3 least frequent 5-mers (between 0.7 to 1 thousands occurrences) can be found in Supplementary **Figure S9A**. Note that frequent 5-mers tend to be AT-rich and least frequent ones are more GC-rich; in line with the relatively low GC content 38% in *S. cerevisiae*.

The most frequent 5-mers have relatively low density scores and the densest 5-mers are relatively rare (see Supplementary **Figure S11A**). So, there is relatively small number of elongation rate-limiting factors. The density distribution of 5-mers shows that the ordering of the lower density 5-mers has a linear ascent compared to the ordering of the higher density 5-mers that appears to fit an exponential function (see Supplementary **Figure S11B**).

Supplementary **Figure S8** includes the nucleotide composition near transcription pauses (see definition in the Method section). We build a Position-Specific Scoring Matrix (PSSM) of the top 100 most dense 5-mers and top 100 least dense 5-mers with (Supplementary **Figure S8B-C**) and *without pauses* (Supplementary **Figure S8D-E**). The resulting nucleotide signature of the densest 5-mers resembles the PSSM of the pauses. We get similar signatures for all the combinations of parameters used in this study as well as for 3-mers and 4-mers. These results support the conjecture that the effect of nucleotide composition on the transcription elongation speed is continuous; this is different than the model that was suggested before: extreme pauses in specific sequences and constant speed in the rest of the sequences.

Supplementary Note 5 - Typical Features of the Global Density Profile

There is a strong pattern of the RNAP aggregated density profile shared by different groups of genes (see the Methods section and Supplementary **Figure S6**). The groups differ by their average RNAP density (high and low) and by their expression level (high and low). Note that the density profile of genes with high (low) density is very similar to the density profile of gene with high (low) expression level. The profile is characterized by a narrow deep (50 nucleotides) near the transcription start site (TSS) followed by a peak stretched for ~150 nucleotides, which is followed by a slow decent to a plateau. These results are similar to those reported earlier by Churchman and Weissman (2011) for a group of 471 well-expressed genes. As these characteristics might be a result of an experimental bias (*e.g.*, nascent transcripts are too short to be uniquely aligned to the reference genome for the first ~20 nucleotides) or due to specific mechanisms of transcription *initiation* that are relevant only for the transcript 5'end, we explore the data with and without the first 200 nucleotides.

Hence, we aligned all the genes to the TSSs. For each position starting at the TSS+80 we selected those genes with the highest density and those with the lowest density and recorded their genomic landscape (80 nucleotides upstream and downstream) at that position. Finally, we have two groups (landscapes of high density and low density positions) of 161 nucleotide long sequences. For each group and position we compute the frequency of each nucleotide (see Supplementary **Figure S10**). It can be seen that the nucleotide signature of the high-density group is *GCATG*.