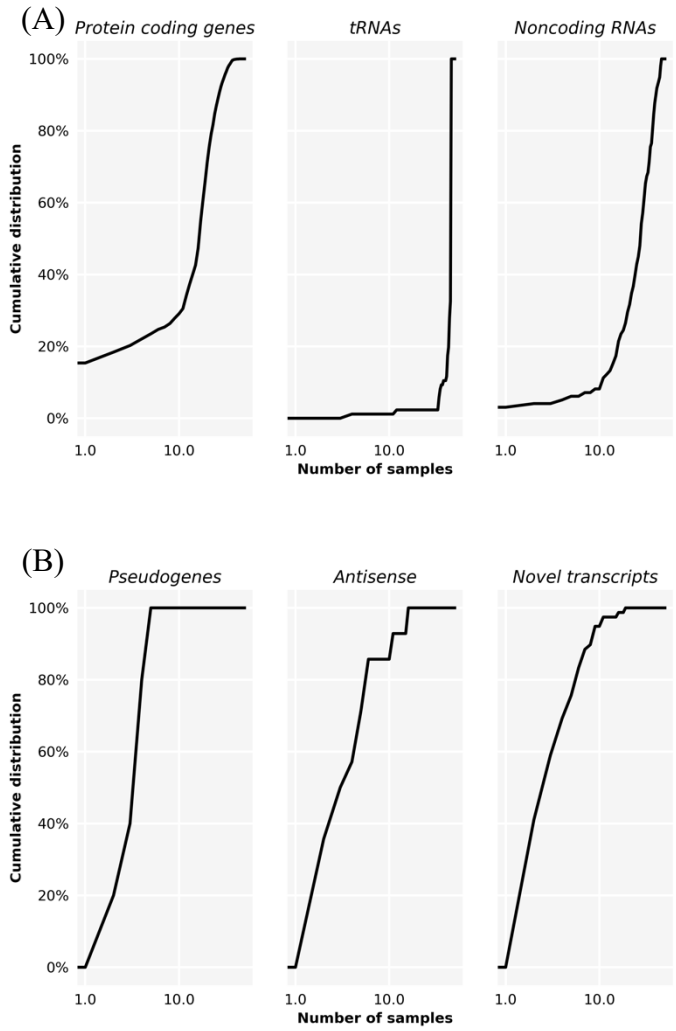
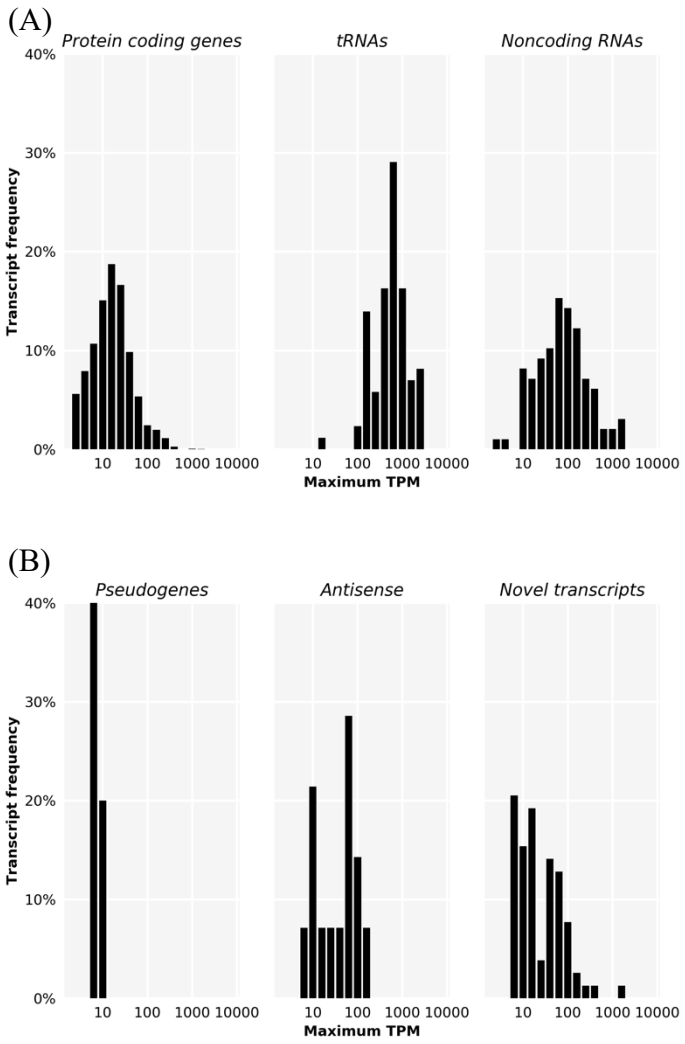
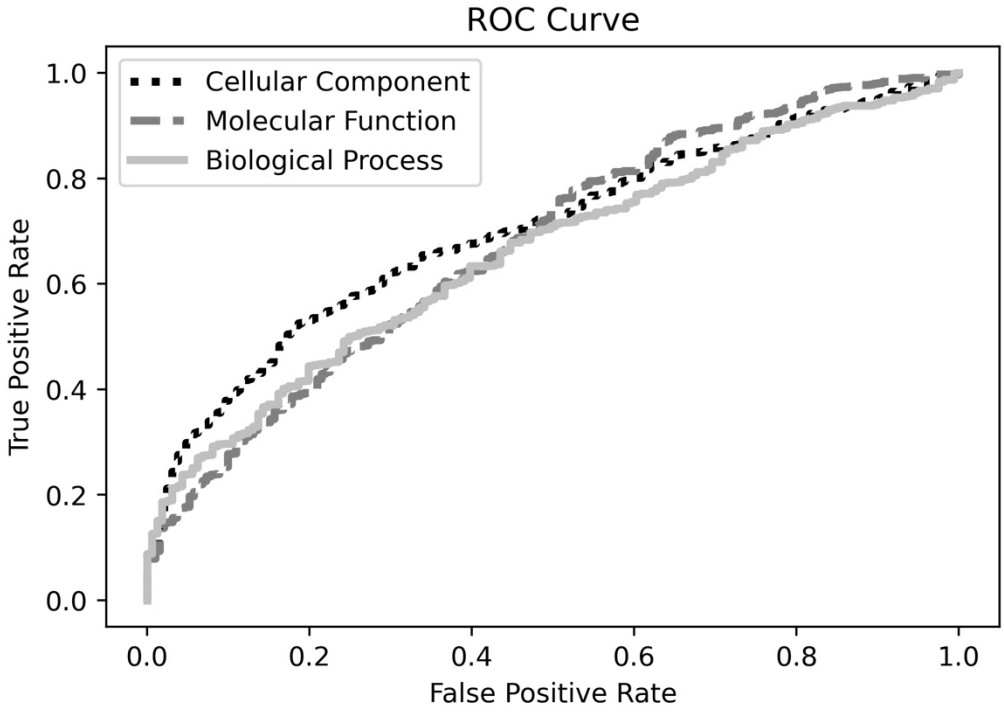


Supplementary Figure 1. Cumulative distribution of the percentage of transcripts expressed in a given number of size-selected samples on a log scale. (A) Cumulative distributions for annotated *E. coli* genes corresponding to protein coding genes, tRNAs, and noncoding RNAs. (B) Cumulative distributions for transcripts corresponding to pseudogenes, antisense transcripts, and other novel transcripts that are not part of annotated *E. coli* genes.



Supplementary Figure 2. Frequency of transcripts expressed at different levels, as measured by the maximum TPM of a transcript across size-selected samples on a log scale. (A) Frequency for annotated *E. coli* genes corresponding to protein coding genes, tRNAs, and noncoding RNAs. (B) Frequency for transcripts corresponding to pseudogenes, antisense transcripts, and other novel transcripts that are not part of annotated *E. coli* genes.



Supplementary Figure 3. Receiver operating characteristic (ROC) curves indicating the trade-off between true positive rate and false positive rate as different thresholds are used for identifing GO roles for genes based on their expression profiles. There are three ROC curves, one for each of the GO roles: cellular component, molecular function, biological process. Thresholds correspond to different probability scores output by a logistic regression classifier representing how likely it is that a gene participates in the given GO role.

Supplementary Table 1. Information about the 3,376 samples from the Sequence Read Archive. The table includes data, for each sample, on the release data, sequencing reads, library selection method, model of sequencing platform used, corresponding BioProject and BioSample, and hyperlink to the sample source.

Available online at: https://dataverse.harvard.edu/api/access/datafile/6562493

Supplementary Table 2. Information about the transcript expression of 4,510 annotated *E. coli* genes as determined by the pipeline. The table includes data, for each gene, on the type of gene, its location within the genome, the gene product, and transcript expression as measured by TPM in each sample.

Available online at: https://dataverse.harvard.edu/api/access/datafile/6562494

Supplementary Table 3. Information about the transcript expression of 5,071 transcripts not corresponding to annotated *E. coli* genes as determined by the pipeline. The table includes data, for each transcript, on the corresponding expressed region of the genome, whether the transcript is antisense to any annotated gene, and expression of the transcript as measured by TPM in each sample.

Available online at: https://dataverse.harvard.edu/api/access/datafile/6562495

Supplementary Table 4. Information about the transcript expression of 4,510 annotated *E. coli* genes as determined by the tool Rockhopper. The table includes data, for each gene, on the type of gene, its location within the genome, the gene product, and transcript expression as measured by TPM in each sample.

Available online at: https://dataverse.harvard.edu/api/access/datafile/6767650

Supplementary Table 5. Co-transcription of pairs of genes. Each row of the table corresponds to a pair of consecutive genes on the same strand in the genome. For each pair, the table indicates its likelihood of being co-transcribed as part of the same transcription unit, if the pair is reported in RegulonDB to be part of the same operon, and information on each gene in the pair.

Available online at: https://dataverse.harvard.edu/api/access/datafile/6562496

Supplementary Table 6. Multi-gene transcription units. Each row of the table corresponds to a set of multiple genes for which there is evidence that the genes are all transcribed as part of the same transcription unit.

Available online at: https://dataverse.harvard.edu/api/access/datafile/6562497