Title: Gene Expression Response to Genomic Inversions in *Escherichia coli*

Authors: Daniella F. Lato, Qing Zeng and G. Brian Golding

Journal: Genome

Corresponding Author Information:
G. Brian Golding
McMaster University
Department of Biology
1280 Main St. West
Hamilton, ON
Canada
L8S 4K1
Email: golding@mcmaster.ca

Supplementary Material

For the most up to date Supplementary Material, please visit GitHub at https://github.com/dlato/Genomic_Inversions_in_Ecoli_Alter_Gene_Expression/.

Further supplemental information and code are available on GitHub at https://github.com/dlato/Genomic_Inversions_in_Ecoli_Alter_Gene_Expression/.

Gene Expression Data

Strain	GEO Accession Number	Date Accessed	NCBI Accession Genome Used For Gene Position
E. coli K12 MG1655	GSE60522	December 20, 2017	U00096
	GSE114917	November 26, 2018	
	GSE54199	December 18, 2019	
	GSE40313	November 21, 2018	
E. coli K12 DH10B	GSE98890	March 13, 2018	NC_010473
E. coli BW25113	GSE73673	December 19, 2017	NZ_CP009273
	GSE85914	December 19, 2017	
E. coli ATCC 25922	GSE94978	November 23, 2018	NZ_CP009072 BA000007

Table S1: Strains and species used for each gene expression analysis. Gene Expression Omnibus accession numbers and date accessed are provided. NCBI genome accession numbers are listed for which genome was used to determine the gene position. Strains with multiple NCBI genome accession numbers had multiple genome versions/builds used to determine the genomic position.

Sequences

Strain	Accession Number	Date(s) Accessed
E. coli K-12 MG1655 *	U00096	September 26, 2016
E. coli K-12 DH10B	NC_010473	February 13, 2020
E. coli BW25113	NZ_CP009273	October 3, 2018
E. coli ATCC 25922	NZ_CP009072	December 18, 2018

Table S2: *E. coli* strains used for the analysis. Accession numbers and date accessed for each genome are provided. Multiple dates and accession numbers for one strain denote updated versions of the genome. An astrix (*) insicates the strain that was used as the representative strain.

Proteomes

Strain	UniProt Accession Number	NCBI Accession Number	Date(s) Accessed
E. coli K-12 MG1655	UP000000625	U00096	May 4, 2020
E. coli K-12 DH10B	UP000001689	NC_010473	May 4, 2020
E. coli BW25113	UP000029103	NZ_CP009273	May 4, 2020
E. coli ATCC 25922	UP000001410	NZ_CP009072	May 4, 2020

Table S3: Proteomes used for the *E. coli* analysis were downloaded from `UniProt`. Accession numbers for both `UniProt` and NCBI as well as date accessed are provided.

Correlation of Gene Expression Over Datasets

To assess uniform expression over *E. coli* strains with multiple data sets we looked at the mean normalized expression values. Multiple replicates from a data set were combined by finding the median normalized CPM expression value for each gene. This was done for any data sets that had multiple replicates. For each gene (x_i) the mean normalized expression value was calculated across all data sets ($\bar{x_{ij}}$). Then the normalized median expression value for each data set was subtracted from the mean across all expression values ($|x_{ij} - \bar{x_{ij}}|$). The distribution of these $|x_{ij} - \bar{x_{ij}}|$ across all genes are found in Figures S1. All data sets are well mixed, implying that the expression levels are consistent across all data sets. Only the *E. coli* K-12 MG1655 strain had multiple expression datasets available so this is the only one that were analyzed. *E. coli* ATCC 25922, *E. coli* BW25113, and *E. coli* K-12 DH10B had only one data set each and therefore were not analyzed.

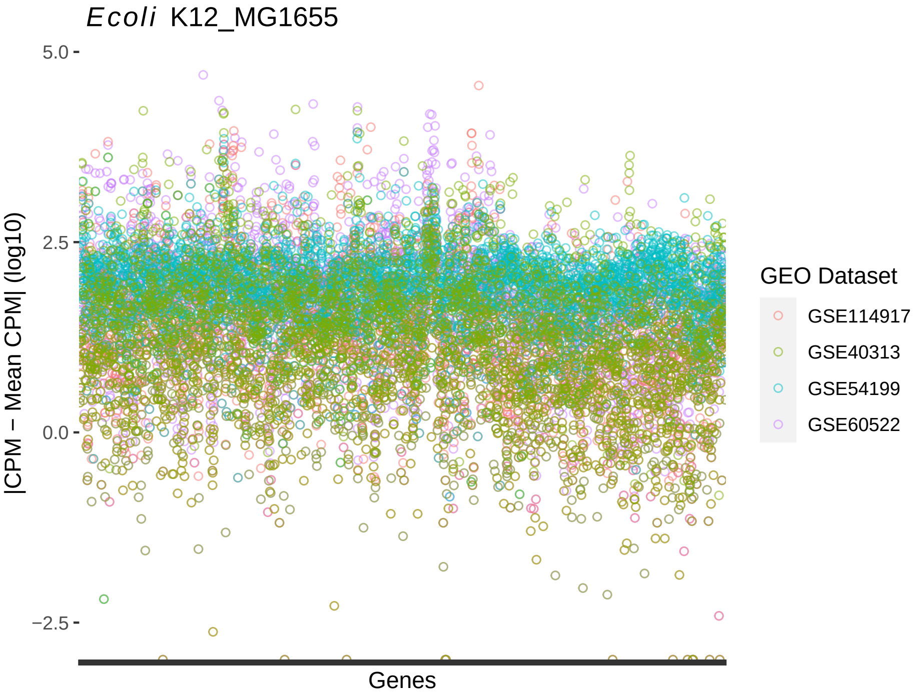

Figure S1: Dot plot distribution of the median expression value for each *E. coli* K-12 MG1655 data set minus the mean expression value for that gene across all data sets. Each gene is shown on the x-axis and the log base 10 values are on the y-axis. The values are coloured by GEO data set.

DIAMOND/BLAST Test Parameters

Command
`diamond blastp -query-cover 90 -evalue 1e6 -outfmt 6`
`diamond blastp -query-cover 95 -evalue 1e6 -outfmt 6`
`diamond blastp -sensitive -query-cover 95 -evalue 1e6 -outfmt "6"`
`diamond blastp -more-sensitive -query-cover 95 -evalue 1e6 -outfmt "6"`
`blastp -qcov_hsp_perc 90 -evalue 0.001 -outfmt "6" -use_sw_tback`
`blastp -qcov_hsp_perc 95 -evalue 0.001 -outfmt "6" -use_sw_tback`

Table S4: Commands used for testing appropriate `DIAMOND` and `BLAST` parameters. Only relevant parameters are shown. The command that yielded the best results and was used for the analysis is indicated in **bold** (`diamond blastp -more-sensitive`).

Length of Inverted Alignment Blocks

A Wilcoxon signed-rank test was used to determine if there was a difference in alignment block length between significant inverted alignment blocks and non-significant inverted alignment blocks. A significant correlation was determined (Wilcoxon signed-rank test: W=4293794.5, p-value < 0.001), indicating that there is a significant difference in the length of significant inverted alignment blocks and non-significant inverted alignment blocks. Significant inverted alignment blocks (mean = 12079bp, median = 10297bp) are on average longer than non-significant inverted alignment blocks (mean = 11310bp, median = 9662bp).

Higashi et al. (2016) H-NS Binding Criteria

The Higashi et al. (2016) data set had multiple criteria to define H-NS binding sites (see Table 3 in Main Paper). They are listed as follows: A: Genes whose coding regions overlap with the H-NS binding regions, B: Genes whose coding regions overlap with the H-NS binding regions and intergenic regions that were bound by H-NS, C: Genes whose coding regions overlap with the H-NS binding regions and intergenic regions that are "class I " (see Higashi et al. (2016)), D: Genes whose coding regions overlap with the H-NS binding regions and intergenic regions that contain known promoter sequences, E: Same as A, but genes on which H-NS binding is restricted to the 3' end and the length overlapping with H-NS-bound regions is $<$10% of the total gene length were excluded from H-NS-bound genes, F: When genes included in transcriptional units whose upstream regions or first coding regions overlapped with H-NS bound regions, all genes in the transcriptional units were judged as genes affected by H-NS binding.

Variation in Expression

	Test Statistic		Coefficient of Variation	
Group	Asymptotic	M-SLRT	Inversion	Non-Inversion
All Blocks	NS	NS	3.26	3.43
Only ATCC genes	NS	NS	3.24	3.78
Group	Asymptotic	M-SLRT	Significant Inversion	Non-Significant Inversion
Significant Inversions	8.738**	13.600***	4.39	3.08

Table S5: Tests for equality of coefficient of variances in gene expression. The “Asymptotic” test refers to the (Feltz and Miller 1996) asymptotic test. The “M-SLRT” test refers to the Modified Signed-Likelihood Ratio Test (M-SLRT) from (Krishnamoorthy and Lee 2014). “All Blocks” indicates all identified alignment blocks. “Only ATCC genes” indicates all ATCC genes that are both inverted and non-inverted. “Significant Inversions” indicates all inverted blocks that had a significant difference in gene expression between the inverted and non-inverted sequences. The coefficient variance in this group was calculated for the inversions that were significant inversions and non-significant inversions. All results are marked with significance codes as followed: < 0.001 = ‘***’, $0.001 < 0.01$ = ‘**’, $0.01 < 0.05$ = ‘*’, > 0.05 = ‘NS’.

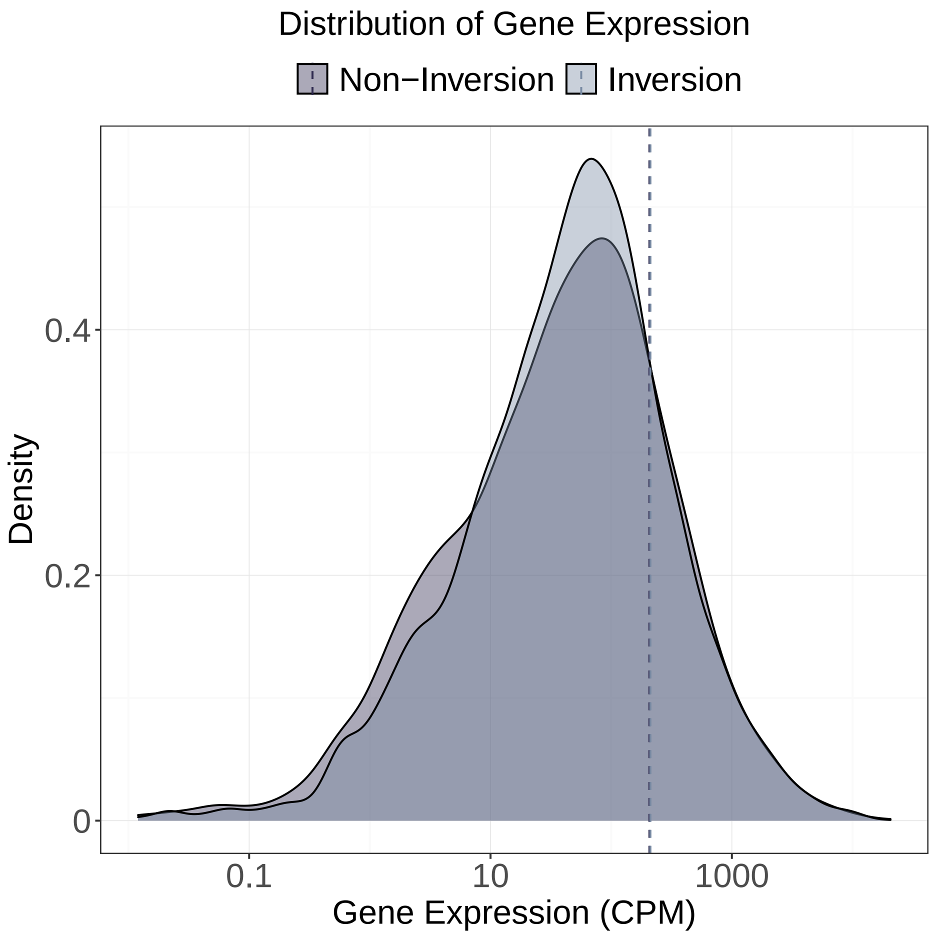

Figure S2: Distribution of gene expression values (CPM) for all genes in Inverted (light grey) and Non-inverted (dark purple) regions of the genomes of *E. coli* K-12 MG1655, *E. coli* K-12 DH10B, *E. coli* BW25113 and *E. coli* ATCC 25922. The expression value in CPM is on the x-axis on a $\log_{10}$ and the density of expression values is on the y-axis. The mean expression values for genes in the Inverted (light grey) and Non-inverted (dark purple) regions are denoted by vertical dashed lines. The means for the Inverted and Non-inverted groups are very similar, and nearly overlapping.

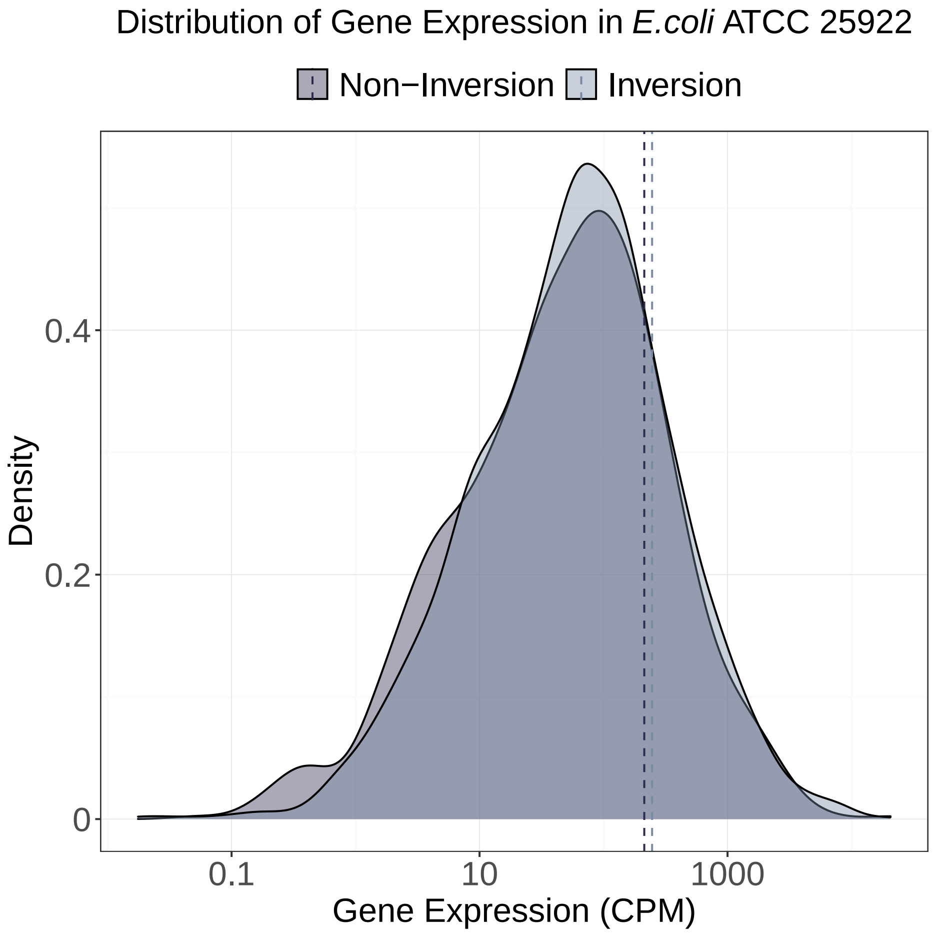

Figure S3: Distribution of gene expression values (CPM) for all genes in Inverted (light grey) and Non-inverted (dark purple) regions of the *E. coli* ATCC 25922 genome. The expression value in CPM is on the x-axis on a $\log_{10}$ and the density of expression values is on the y-axis. The mean expression values for genes in the Inverted (light grey) and Non-inverted (dark purple) regions are denoted by vertical dashed lines.

References

Feltz C J and Miller G E (1996). An asymptotic test for the equality of coefficients of variation from k populations. Stat Med 15, 646–658.

Higashi K, Tobe T, Kanai A, Uyar E, Ishikawa S, S uzuki Y, Ogasawara N, Kurokawa K, and Oshima T (2016). H-NS Facilitates Sequence Diversification of Horizontally Transferred DNA s during Their Integration in Host Chromosomes. PLoS Genet 12, e1005796.

Krishnamoorthy K and Lee M (2014). Improved tests for the equality of normal coefficients of variation. Computational Statistics 29(1-2), 215–232.