Title: Genomic inversions in *Escherichia coli* alter gene expression and are associated with nucleoid protein binding sites

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	Growth Information
E. coli K12 MG1655	GSE60522	December 20, 2017	U00096	LB media
	GSE114917	November 26, 2018		LB-Miller broth
	GSE54199	December 18, 2019		LB media
	GSE40313	November 21, 2018		LB media
E. coli K12 DH10B	GSE98890	March 13, 2018	NC_010473	MOPS EZ Rich Defined Medium
E. coli BW25113	GSE73673	December 19, 2017	NZ_CP009273	LB media
	GSE85914	December 19, 2017		M9 glucose (0.4% w/v)
E. coli ATCC 25922	GSE94978	November 23, 2018	NZ_CP009072	MH broth

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 (*) indicates 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.

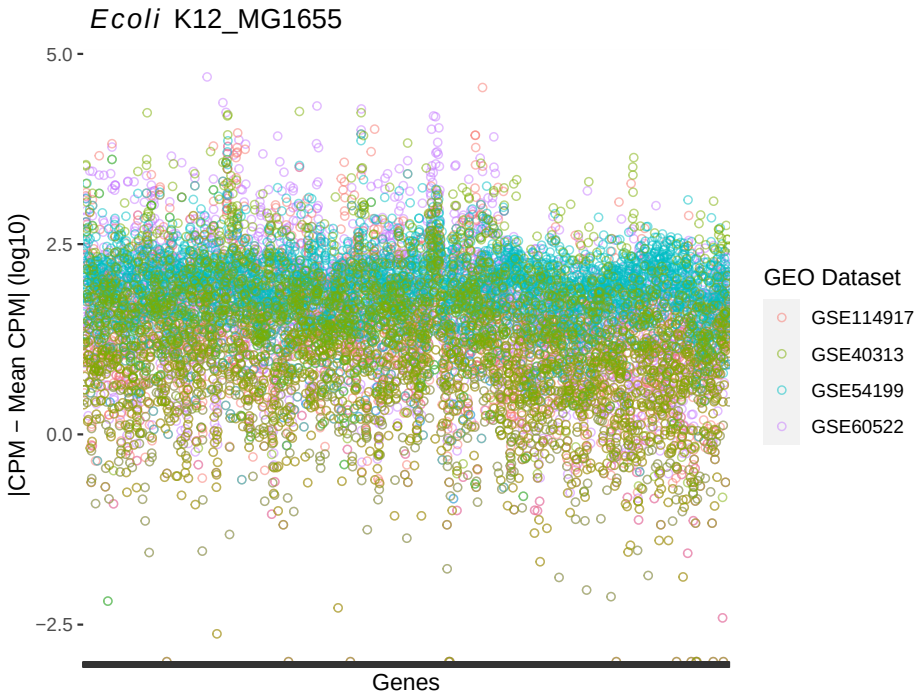

Figure S1: Dot plot distribution of the median expression value for each *E. coli* K-12 MG1655 data set minus the mean expression value for that gene across all data sets. Each gene is shown on the x-axis and the log base 10 values are on the y-axis. The values are coloured by GEO data set.

`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=485996, 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 = 6514bp, median = 7241bp) are on average shorter than non-significant inverted alignment blocks (mean = 11440bp, median = 9884bp).

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.

Lack of Patterns in Genomic Structure Within *E. coli* Phylotypes

Inspired by work from Denamur et al. (2021), we wanted to explore if there were any obvious patterns of inversions within *E. coli* phylotypes. A `Parsnp` (Treangen et al. 2014) core genome alignment was done using 38 *E. coli* genomes (see Table S5) from across 8 phylogroups. We identified a total of 1743 alignment blocks ranging from 22bp-14978bp long, with a mean of 1151bp and a median length of 887bp. A snapshot of the `Parsnp` alignment can be seen in Figure S4. *E. coli* from the same phylogroup do not appear to have similar genomic structure patterns which include genomic reorganization such as rearrangements, translocation, and inversions. Therefore, we did not detect strong inversion patterns within *E. coli* phylogroups. As an example, synteny (depicted in the alignment as a gradient of blue) is not consistent or conserved among the highlighted phylogroups A (purple) and E (red), indicating a high degree of variation in genomic structure between taxa within a phylogroup.

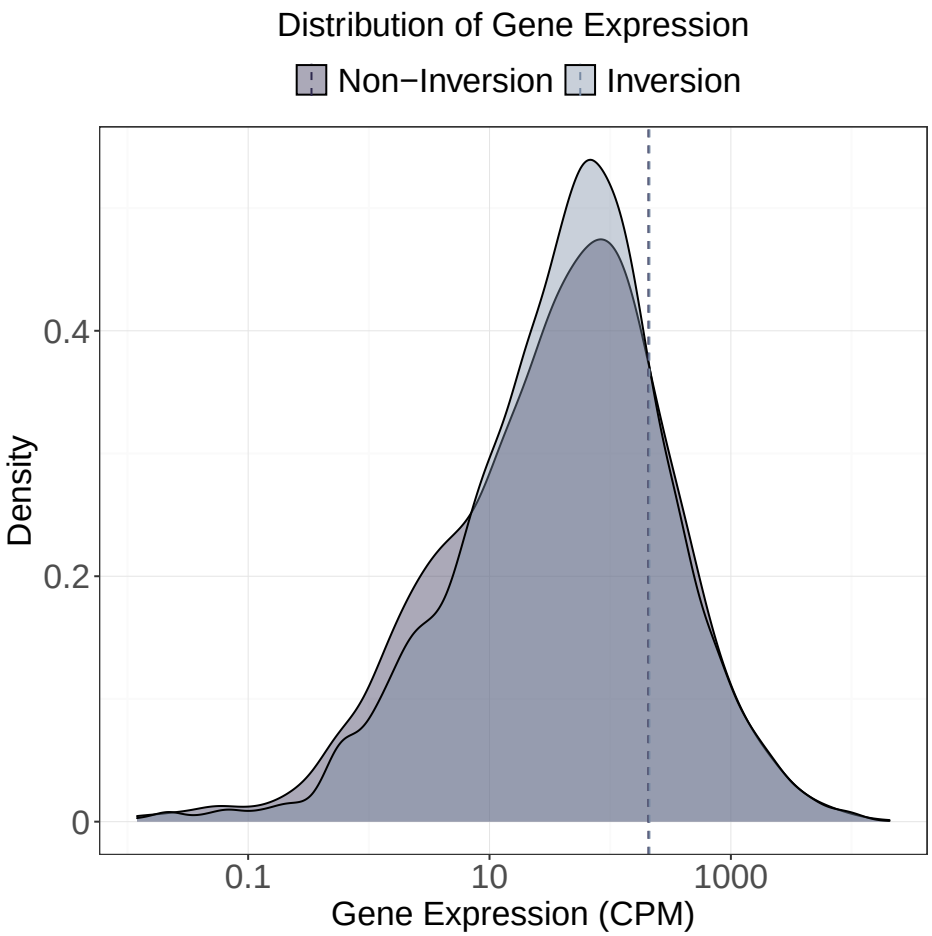

Figure S2: Distribution of gene expression values (CPM) for all genes in Inverted (light grey) and Non-inverted (dark purple) regions of the genomes of *E. coli* K-12 MG1655, *E. coli* K-12 DH10B, *E. coli* BW25113 and *E. coli* ATCC 25922. The expression value in CPM is on the x-axis on a $\log_{10}$ 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.

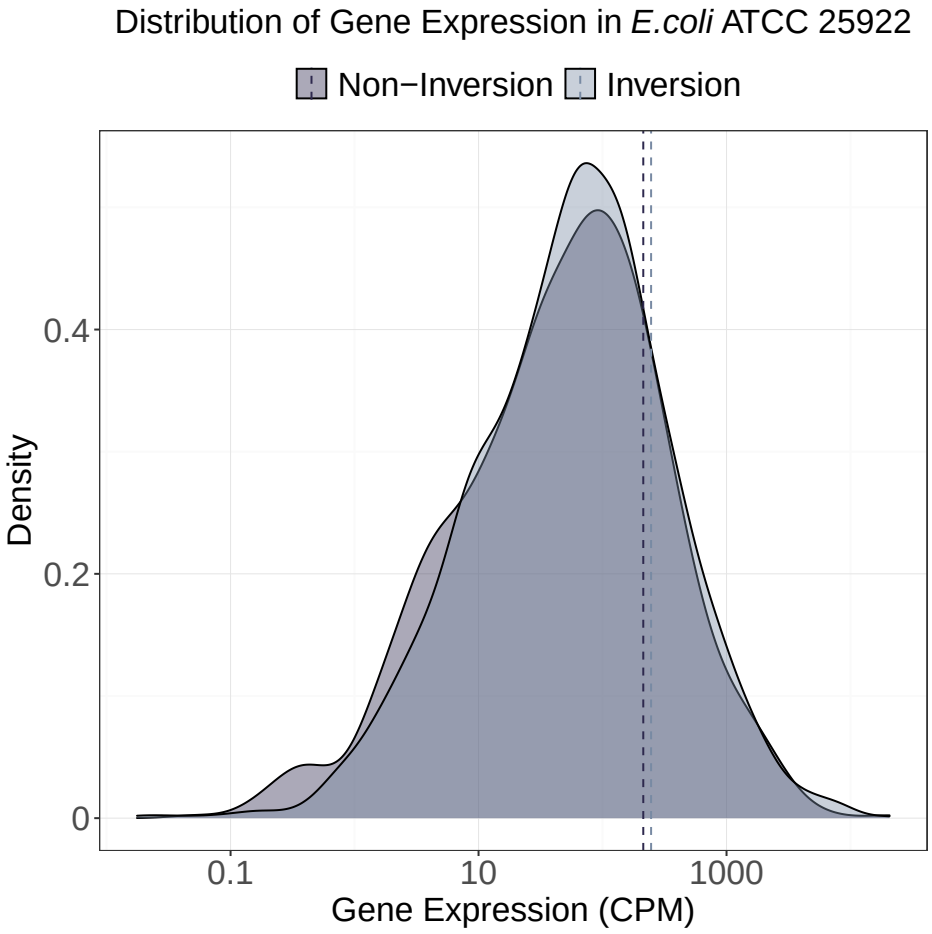

Figure S3: Distribution of gene expression values (CPM) for all genes in Inverted (light grey) and Non-inverted (dark purple) regions of the *E. coli* ATCC 25922 genome. The expression value in CPM is on the x-axis on a $\log_{10}$ and the density of expression values is on the y-axis. The mean expression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

Strain	Phylogroup	Accession Number
K-12 MG1655 *	A	U00096
K-12 DH10B	A	NC_010473
BW25113	A	NZ_CP009273
101-1	A	GCA_000168095.1
53638	A	GCA_000167915.2
ECOR24	A	GCA_002190595.1
HS	A	GCA_000017765.1
RDEx444	A	GCA_003123505.1
CIP61.11	A	GCA_900236115.1
ECOR01	A	GCA_002190105.1
H10407	A	GCA_000210475.1
11368	B1	GCA_000091005.1
12009	B1	GCA_000010745.1
55989	B1	GCA_000026245.1
E110019	B1	GCA_000167875.2
E24377A	B1	GCA_000017745.1
ATCC 25922	B2	NZ_CP009072
536	B2	GCA_000013305.1
CFT073	B2	GCA_000007445.1
E2348/69	B2	GCA_000026545.1
ECOR60	B2	GCA_002189835.1
ECOR63	B2	GCA_002189905.1
ECOR64	B2	GCA_002189945.1
ED1a	B2	GCA_000026305.1
FN-B26	B2	GCA_902505495.1
H223	B2	GCA_002110555.1
LF82	B2	GCA_000284495.1
NA114	B2	GCA_000214765.3
S88	B2	GCA_000026285.2
SE15	B2	GCA_000010485.1
APECO78	C	GCA_000332755.1
S286	C	GCA_013363015.1
ECOR49	D	GCA_002190975.1
042	D	GCA_000027125.1
UMN026	D	GCA_000026325.2
4608-58	E	GCA_000805835.1
ECOR31	E	GCA_001865905.1
ECOR42	E	GCA_002190935.1
SAKAI	E	GCA_003028755.1
IAI39	F	GCA_000026345.1
SMS-3-5	F	GCA_000019645.1
H299	G	GCA_000176695.2

Table S5: *E. coli* strains used for the supplementary `Parsnp` analysis. Phylogroup and accession numbers for each genome are provided. An astrix (*) indicates the strain that was used as the representative strain.

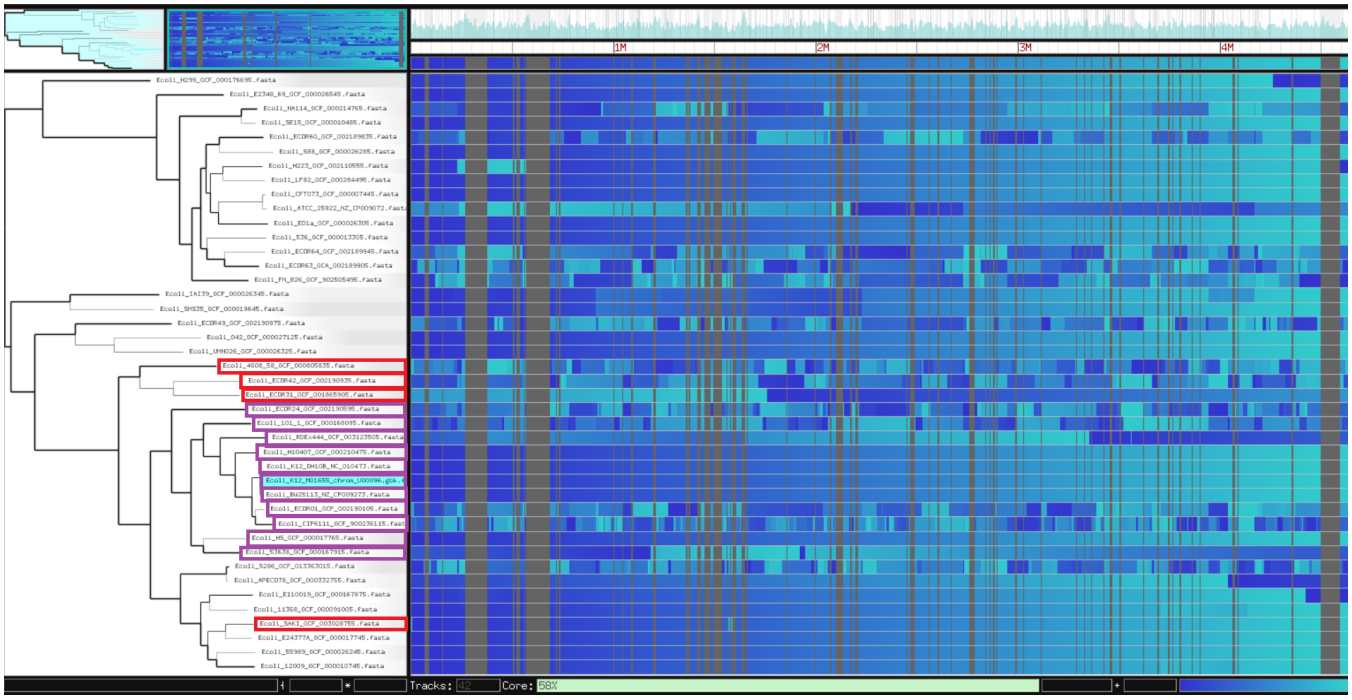

Figure S4: Visualization of `Parsnp` core genome alignment between 38 *E. coli* strains from 8 phylogroups (see Table S5). A phylogenetic tree produced by `Parsnp` is found on the left side of the diagram, which does not generally match the known phylogroup classification of the aligned *E. coli* strains. The core genome alignment is highlighted in blue on the right side of the diagram with a gradient from dark to light blue indicating synteny and genomic structure similarities to the *E. coli* K-12 MG1655 strain (highlighted in light blue in the phylogenetic tree). Dark grey regions of the alignment are not apart of the core genome identified by `Parsnp`. Taxa from the phylogroups A and E are highlighted by purple and red boxes respectively as an illustrative example of the variation in genomic structure between taxa within each group.

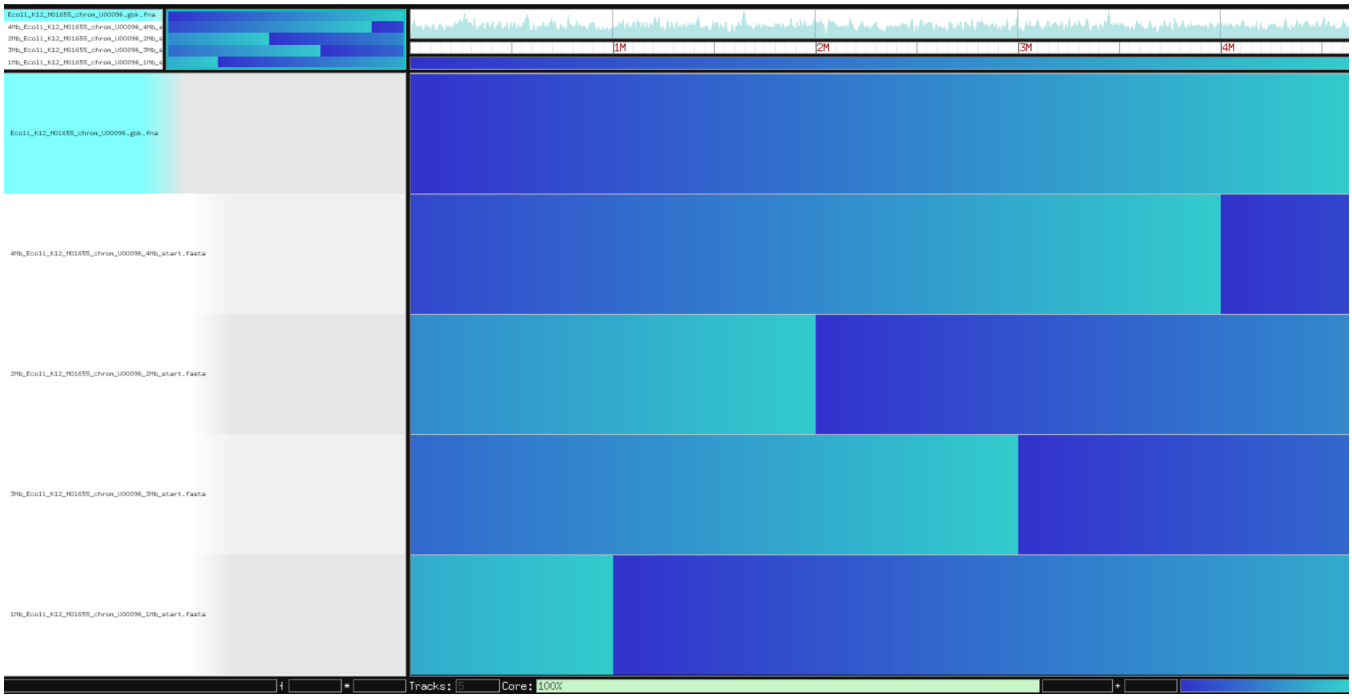

Figure S5: Visualization of `Parsnp` core genome alignment between five "versions" of the *E. coli* K-12 MG1655 genome. The gene order in four of the "versions" were altered to start at 1Mb, 2Mb, 3Mb, and 4Mb into the genome, thus "shifting" the annotation. The original genome of *E. coli* K-12 MG1655 (highlighted in light blue in the phylogenetic tree) is the first sequence listed vertically followed by the 4Mb, 2Mb, 3Mb, and 1Mb "shifted" genomes. A phylogenetic tree produced by `Parsnp` is found on the left side of the diagram. The core genome alignment is highlighted in blue on the right side of the diagram with a gradient from dark to light blue indicating synteny and genomic structure similarities to the original genome of *E. coli* K-12 MG1655.

`Parsnp` Illustrative Example

As mentioned in the manuscript, one of the benefits of using an alignment aware genome alignment program like `Parsnp` is that there is no need to ensure that all genomes begin with the same genes and share gene order throughout the genome. The alignment blocks specified by `Parsnp` are genomic rearrangement aware, meaning that an alignment block or set of genes, can be present in varying genomic locations across each taxa (rearrangements). Thus, even though in Figure 1 the *E. coli* ATCC 25922 has a shift in gene order, likely due to annotation differences, `Parsnp` can detect these changes and adjust the alignment accordingly. The identification of inversions by `Parsnp` is independent of the order of these genes. We have illustrated this point in Figure S5 where we took the genome of *E. coli* K-12 MG1655 and adjusted the sequence to start at 1Mb, 2Mb, 3Mb, and 4Mb into the genome, thus "shifting" the annotation in a similar fashion to what is observed in the annotation of *E. coli* ATCC 25922. An inversion in `Parsnp` is visualized in the alignment as the opposite colour gradient moving from light blue to dark blue. We aligned all of these sequences using `Parsnp` and as expected, there were no inversions detected (since none of the sequences were actually inverted), but the shift in gene order was detected. Thus, any annotation shift that is present in Figure 1 in the *E. coli* ATCC 25922 does not impact `Parsnp`'s ability to identify inverted regions of the genome.

Differential Expression Analysis on Various Growth Media

A differential gene expression analysis was performed to determine if there were any gene expression differences due to the various growth media used in this analysis. As an example case, we looked at the two samples from the

E. coli BW25113 strain which was grown in LB medium (GSE73673) and M9 medium (GSE85914) (see Table S1). For this analysis, the raw count data was used from the previously mentioned GEO datasets. Homologous genes were identified as mentioned previously using the `Parsnp` alignment. Any set of homologous genes that did not match the homology results from the `DIAMOND` protein alignment were excluded from this analysis. Only homologous genes that could be found in each GEO dataset were used for this analysis. The standard pipeline for performing differential gene expression analysis was followed using the `DESeq2` (Love et al. 2014) package in `R` (R Development Core Team 2014). The cutoff values we used for the differential gene expression analysis were a False Discovery Rate (FDR) of < 0.05 and a Log Fold Change (LFC) in gene expression of greater than 2 or less than -2. We detected 319 genes that were up-regulated (12%) and 197 genes that were down-regulated (7%) for a total of 516 (19%) differentially expressed genes between these two growth media. Although there is a non-zero difference in gene expression between media, we are confident that the robust permutation tests used in this analysis mitigate these gene expression changes.

References

Denamur E, Clermont O, Bonacorsi S, and Gordon D (2021). The population genetics of pathogenic *Escherichia coli*. Nat Rev Microbiol 19, 37–54.

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 DNAs during their integration in host chromosomes. PLoS Genet 12, e1005796.

Love M I, Huber W, and Anders S (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15(12), 550.

R Development Core Team (2014). *R: a language and environment for statistical computing*. Vienna, Austria.

Treangen T J, Ondov B D, Koren S, and Phillippy A M (2014). The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol 15, 524.