Dataset S3 from ***Whole genome sequences from wild-type and laboratory evolved strains define the alleleome and establish its hallmarks***. Edward Alexander Catoiu, Patrick Phaneuf, Jonathan Monk, Bernhard O Palsson*. *Correspondence to Bernard Palsson, palsson@ucsd.edu

DatasetS3 contains the nucleic acid and amino acid sequences for 4,349 genes across 2,661 WT *E. coli* strains. This allele-IDs (nucleic acid allele ID) match the values in DatasetS2. The data is present as 4 text files: which contain the allele sequences for genes 1-1000, 1001-2000, 2001-3000, 3001-4349. The first line of each text file are the column headers: 'gene', 'na_allele_id', 'na_allele_count', 'na_seq_length', 'aa_allele_id', 'aa_seq_length', 'aa_seq_length_fraction', 'aa_seq_length_zscore', 'passed_allele_qcqa', 'na_seq', 'aa_seq_pre_indel', separated by semi-colons (';'). Each subsequent line is a unique nucleic acid allele in the WT alleleome with the following information:

1. **'gene'** – Blattner Number derived from RAST.
2. **'na_allele_id'** – a unique nucleic acid allele ID "{gene}-{allele #}".
3. **'na_allele_count'** – the number of strains with the specific nucleic acid sequence allele.
4. **'na_seq_length'** – length of nucleic acid sequence.
5. **'aa_allele_id'** – amino acid allele ID "{gene}-{allele #}" (these can be repeated across multiple lines) and do not correspond with the nucleic acid allele ID.
6. **'aa_seq_length'** – length of amino acid sequence (until first STOP codon).
7. **'aa_seq_length_fraction'** – fraction of nucleic acid sequence translated (pre-indel).
8. **'aa_seq_length_zscore'** – QCQA metric, determined from the distribution of amino acid allele sequence lengths for a particular gene.
9. **'passed_allele_qcqa'** – QCQA determination. Failed alleles do not enter the sequence-alignment pipeline and do not contribute to the WT alleleome consensus sequence.
10. **'na_seq'** – Nucleic acid sequence from genome.
11. **'aa_seq_pre_indel'** – Amino acid sequence translated until first STOP codon.

A sample line is shown below. NOTE: This allele does NOT pass QCQA.

```
'b1068;b1068-228;1;923;b1068-143;8;0.02600216684723727;-7.823686986541865;
nan;GTGAAAAATTACGTATCGGCGTAGTGGGATTAGGTGGCATTGCGCAAAAAGCGTGGTTACCGGTGCTGGC
GGCAGCGTCTGACTGGACGTTACAAGGAGCCTGGTCGCCTACGCGCGAAAGCCCTGCCAATTTGTGAAAGCT
GGCGCATTCCTTATGCCGATTCGTTATCCAGCCTTGCCGCCAGTTGCGATGCGGTTTTTGTGCATTCCAGCACC
GCCAGCCACTTTGACGTGGTCAGTACGTTACTCAATGCGGGTGTACATGTCTGTGTCGATAAACCGCTGGCAGA
AAATCTGCGCGATGCTGAACGGCTGGTGGAACTGGCGGCGCGTAAAAAACTGACGTTGATGGTCGGTTTTAACC
GTCGTTTCGCACCACTCTACGGTGAGTTAAAAACGCAACTCGCTACCGCAGCTTCGCTAAGAATGGATAAACAT
CGTAGCAATAGTGTCGGGCCACACGATCTTTATTTCACGTTGCTGGATGATTATCTGCATGTGGTGGATACCGC
GCTGTGGTTGTCGGGCGGCAAAGCCTCTCTGGATGGCGGTACGCTACTGACTAACGACGCTGGCGAAATGCTGT
TTGCCGAGCACCATTTTTCGGCCGGTCCTTTGCAGATCACCACCTGTATGCATCGCCGTGCCGGAAGTCAGCGT
GAAACCGTGCAGGCCGTGACTGACGGTGCGCTCATCGACATTACGGATATGCGCGAATGGCGTGAGGAGCGCGG
GCAGGGCGTAGTGCATAAACCGATTCCTGGTTGGCAGAGCACTCTTGAACAACGTGGATTTGTCGGCTGTGCAC
GGCACTTCATTGAATGTGTGCAAAATCAGACAGTTCCGCAAACCGCCGGCGAACAGGCCGTGCTGGCGCAACGT
ATCGTTGACAAGATCTGGCGCGATGCGATGAGTGAATAA;VKNYVSA*\n'
```

Usage: This data was used to generate Fig. S2. There are 729,212 unique codon alleles that 1) pass QCQA (value in dataset) and 2) code for genes present in more than 133 (5%) of WT strains. The latter needs to be calculated before recreating Fig. S2b-c. For each gene, sequence-alignment of all its alleles yields the WT consensus sequence and variants that define the WT alleleome (Dataset S4).