ANALYSIS OF A FIVE-WAY GENOME COMPARISON

We applied GRIL to five enterobacteria: *Escherichia coli* K-12 MG1655 (Blattner, Plunkett et al. 1997), *E. coli* O157:H7 EDL933 (Perna, Plunkett et al. 2001), *E. coli* O157:H7 VT-2 Sakai (Hayashi, Makino et al. 2001), *Salmonella enterica* serovar Typhi CT18 (Parkhill, Dougan et al. 2001), and *S. enterica* serovar Typhimurium LT2 (McClelland, Sanderson et al. 2001). *E. coli* MG1655 is a common laboratory strain derived from the original K-12 isolate, which was isolated in 1922 from an anonymous hospital patient with an unrelated pathology. *E. coli* O157:H7 strains cause a bloody diarrhea that can lead to fatal hemolytic uremic syndrome. The two sequenced strains EDL933 and VT-2 Sakai are associated with outbreaks of food-borne illness in the U.S. and Japan, respectively. *Salmonella enterica* serovar Typhimurium causes a relatively mild form of diarrhea, while serovar Typhi is the responsible agent for typhoid fever, a systemic infection with a 10% mortality rate. These *Salmonella* serovars were previously referred to as *S. Typhi* and *S. Typhimurium.* These genomes have been the subject of a number of comparative analyses aimed toward revealing the genetic basis for the diversity of disease phenotypes. Fasta files are available http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html. All genomes are uniformly linearized at a common marker upstream of a single copy thrA gene.

Using a seed match mer size of 23 base pairs yields 8434 MUMs, identifying 290 kb of exactly matching sequence. During the first execution, GRIL constructs a separate sorted mer list for each genome. The minimum seed match size *m* must be also be specified to search for MUMs. The command line is thus:

```
> gril.exe –m 23 –o "5wayMUMs.txt"  f1.fas f1.sml f2.fas f2.sml f3.fas f3.sml
f4.fas f4.sml f5.fas f5.sml
```

Where f1 through f5 are the file names of the genome sequences. This step saves all MUMs in the output file "5wayMUMs.txt".

Using the MUMs located during the first execution, putative LCBs can be found during a second execution of GRIL. To find LCBs, GRIL can be executed with the following additional parameters: -s, minimal filtered MUM size *s* ($s \geq m$), -f, a generalized offset threshold *f*, -d, a minimal LCB density (percent identity) *d*, and –r, a minimal LCB range *r*. Additionally, -l should be specified to indicate that GRIL should report putative LCBs. We used $f = 100{,}000$ and $r = 10{,}000$ (and default settings for the rest) to generate Figure 1 on the web page:

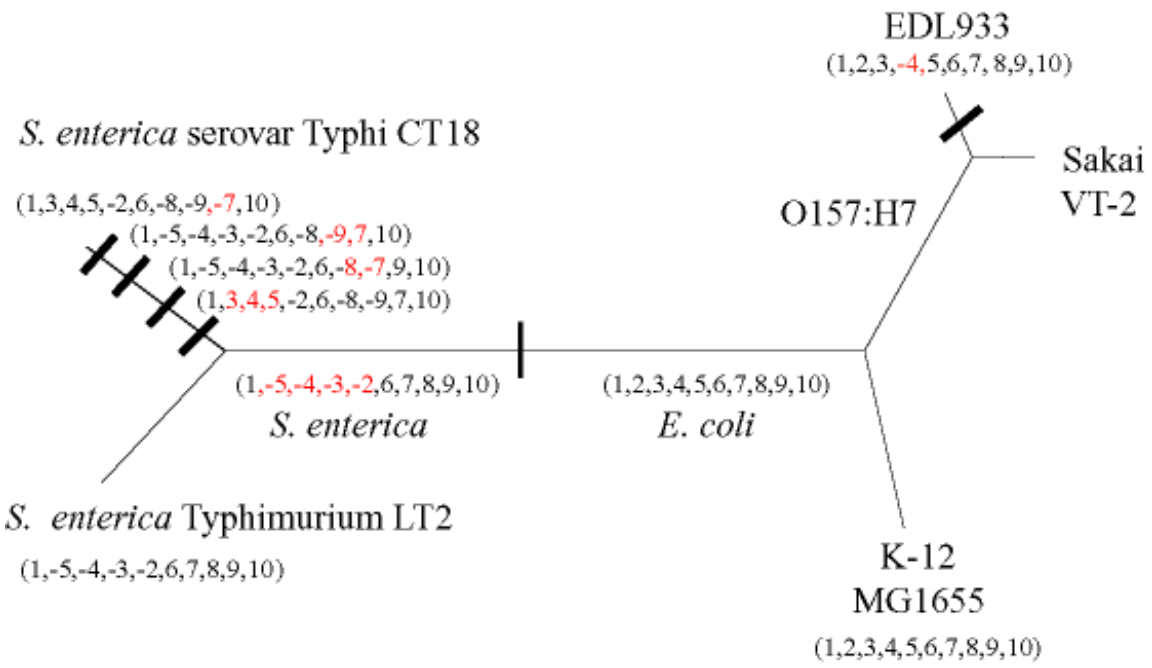
```
> gril.exe –l –i "5wayMUMs.txt"  -f  100000 –r 10000  –o  "5wayLCBs.txt"
```

The above command line takes the MUMs in "5wayMUMs.txt" as input and generates the output file "5wayLCBs.txt" containing putative LCBs found using the given parameters.

An unrooted phylogenetic tree was constructed by neighbor-joining (Saitou and Nei 1987) as implemented in MEGA2 (Kumar, Tamura et al. 2001). The number of pair-wise unique matching-mers, denoted as $S^{\#}(G_i, G_j)$, is a similarity measure that is converted to a distance metric by the formula:

$$d(G_i, G_j) = |G_i| + |G_j| - 2S^{\#}(G_i, G_j)$$

From Figure 5, it is obvious that the number of inversions on each branch is uncorrelated with the length of the branch.



Permutations of sections indices from Figure 4 are attached to labeled leaf nodes. K-12 MG1655 and O157:H7 Sakai VT-2 are both represented by the identity permutation (1,2,3,4,5,6,7,8,9,10). Inversions are drawn as thick bars on the tree branches they transform. Putative ancestral states are indicated by permutations inserted between contiguous bars and/or leaf nodes. Red integers denote the range of the inversion along the path in a direction emanating from the reference genome. No bar bars the path between MG1655 and Sakai VT-2, reflecting their complete collinearity.

Recombinational events among rRNAs D, E, C and B provide a unified explanation for inversions about the origin of replication. Although confined to *S. Typhi* in this particular strain comparison, other examples of rRNA-mediated rearrangement abound in the literature. (e.g.(Schmid and Roth 1983; Liu and Sanderson 1995; Liu, Rahn et al. 2002).The interested reader can identify rRNAs with inversion endpoints by comparing permutations on the *S. Typhi* branch with corresponding labels in Figure 4 via Figure 5.

Each inversion about the terminus is different. The best understood example is the 440 kb inversion between the O157:H7 isolates. EDL933 is slightly larger than the VT2-Sakai, due to a duplicate copy of the urease-tellurite island (87.5 kb) located nearby the original in the first replichore. Replichore sizes vary dramatically between isolates: in VT2-Sakai, the first replichore is 290 kb larger compared to EDL933 where the second replichore is 70 kb larger. The size differential is mitigated by the asymmetric inversion

about the terminus. In Figure 4, the left hand boundary of the inversion appears to be tightly bracketed between segments in MG1655 coordinates, while the location of the right hand boundary occurs in an area without any MUMs. Earlier pairwise analyses have placed both boundaries inside cryptic prophages (B. Mau, unpublished) present in O157:H7 but not K-12. The absolute collinearity between MG1655 and O157:H7 VT-2 Sakai supports the hypothesis that this inversion is a recent evolutionary event in EDL933, possibly the product of biological pressures favoring replichores of equal size (Schmid and Roth 1983) in response to the uneven distribution of large lateral transfers near the terminus in replichore one.

Next, a large inversion phylogenetically separates *Salmonella* from *Escherichia* (Sanderson and Roth 1988) in Figure 5. A comparison of gene order in MG1655 versus *S. Typhimurium* places the locus of recombination within a 302 basepair stretch between *minE* and *rnd* orthologs in both *Salmonella.* The location of the other frontier is unclear due to differential horizontal transfer at a site immediately to the right of orthologs of the *icdA* gene: 28 kb in MG1655, 51 kb in EDL933, 48 kb in Sakai, 36 kb in *S. Typhimurium*, but only a mere 750 basepairs in *S. Typhi.* Hence, the size of the inversion varies between 552 to 592 kb in *S Typhimurium* and 591 to 604 kb in *S. Typhi*, depending on how one allocates acquired sequence. The large substitutions can be characterized as a pair of small phage-related sequences in K-12, cryptic lambdoid prophages in both O157:H7 strains, a 31-gene pathogenicity island in *S. Typhimurium* (beginning at STM1239 in (McClelland, Sanderson et al. 2001)), and a tandem repeat of IS200 elements in *S. Typhi*. This inversion is the most ancient large-scale rearrangement in our study, perhaps explaining our inability to find the recombination endpoints.

Finally, a second inversion (McClelland, Sanderson et al. 2001; Parkhill, Dougan et al. 2001), situated inside the larger one, is sandwiched between two copies of the insertion sequence IS 200 in *S. Typhi*. IS 200 encodes a transposase tnpA occurring 38 times in *S. Typhi*, but only six times in *S. Typhimurium* and not at all the *E. coli* strains. Since neither bounding IS element has an ortholog in *S. Typhimurium*, the inversion is placed in the *S.Typhi* lineage, along with the rRNA mediated inversions about the origin.

When the range *r* is dropped to 1,000 bp, the number of LCBs doubles to 21. A compensatory increase the minimum MUM size *s* to 27 recovers the large.ten LCBs. But what of these smaller intervals that are not explicable as inversions? Analysis reveals that three regions of modest range comprise three gene clusters, consisting of 9, 6, and 3 genes, organized into operons. These include an operon pair (hyaABCDEF and appCBA), forming a hydrogenase/cytochrome oxidase complex, a tor operon, (torSTRCAD) acting on trimethylamine N-oxide, and finally a cad operon (cadABC) involving cadeverine and lysine. Their placement and rearrangement are overlaid onto Figure 2 on the GRIL web site. A fourth, weak match occurred among ambiguously annotated regions in the five genomes.

REFERENCES

Blattner, F. R., G. Plunkett, 3rd, et al. (1997). "The complete genome sequence of Escherichia coli K-12." Science **277**(5331): 1453-74.

Hayashi, T., K. Makino, et al. (2001). "Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12." DNA Res **8**(1): 11-22.

Kumar, S., K. Tamura, et al. (2001). "MEGA2: molecular evolutionary genetics analysis software." Bioinformatics **17**(12): 1244-5.
Liu, G. R., A. Rahn, et al. (2002). "The evolving genome of Salmonella enterica serovar Pullorum." J Bacteriol **184**(10): 2626-33.
Liu, S. L. and K. E. Sanderson (1995). "Rearrangements in the genome of the bacterium Salmonella typhi." Proc Natl Acad Sci U S A **92**(4): 1018-22.
McClelland, M., K. E. Sanderson, et al. (2001). "Complete genome sequence of Salmonella enterica serovar Typhimurium LT2." Nature **413**(6858): 852-6.
Parkhill, J., G. Dougan, et al. (2001). "Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18." Nature **413**(6858): 848-52.
Perna, N. T., G. Plunkett, 3rd, et al. (2001). "Genome sequence of enterohaemorrhagic Escherichia coli O157:H7." Nature **409**(6819): 529-33.
Saitou, N. and M. Nei (1987). "The neighbor-joining method: a new method for reconstructing phylogenetic trees." Mol Biol Evol **4**(4): 406-25.
Sanderson, K. E. and J. R. Roth (1988). "Linkage map of Salmonella typhimurium, edition VII." Microbiol Rev **52**(4): 485-532.
Schmid, M. B. and J. R. Roth (1983). "Selection and endpoint distribution of bacterial inversion mutations." Genetics **105**(3): 539-57.