Supplemental course assignment #1

Individual projects on comparative genomics of three *E. coli* O157:H7 strains.

Focus will be on the genomes of three *E. coli* O157:H7 strains
#1) 933EDL version2 (Ground beef outbreak 1982)
#2) Sakai version1 (radish sprout outbreak 1996)
#3) EC 4042 version 2 (Spinach outbreak 2006)

Topics to address:
#1) Run a blast analysis (BLASTN) with your virulence gene from strain EDL933 against the other two strains and provide the results of the % identity in a list or table. Is the gene, and the corresponding open reading frame for the protein sequence (ORF), conserved in all three genomes and are they all the same length? Is there more than one copy of the gene in any of the genomes you queried? Are they present in the Mauve genome alignment of the three genomes? From the Mauve alignment, provide the coordinate positions or create an image to include in your report.

#2) Describe how the protein produced from this gene may be involved in the microorganism's ability to cause human disease. Briefly summarize the supporting evidence by clicking the link from the ASAP database for the for the subsystem annotation: virulence or putative virulence factor (refer to student instructions if you are unclear how to find the subsystem annotations for a gene in ASAP).

#3) Is this gene or a homolog found in other *Enterobacteria*? (hint run a blast in the ASAP database against all other organisms) Is this gene or a homolog found in other microorganisms? (hint run a blast search at NCBI against all bacteria and archaea. Briefly provide the five best blast "hits" with % identity).

#4) Using the Mauve alignment of the three genomes, identify a unique island in the genome for one strain and briefly summarize the predicted products from the genes located in the genomic island (provide an image generated in Mauve showing the genomic region(s) you worked with). Next, identify a region that is unique to the genomes of two strains and briefly summarize the predicted products (provide coordinates for the genomic island or an image). Overall, based on your analysis of your two identified genomic regions, do you think that the genes located in either genomic island play a role in virulence and/or the evolution of *E. coli* O157:H7 genomes? How important do you think phages are in variation of the genomes? (OR, an alternative question: From your analyses, what is one mechanism that has caused divergence of these genomes? Feel free to compare notes with those around you in answering this question.)

Supplemental course assignment #2

Individual projects for *Yersinia pestis* 5 genome alignment module

There are currently five finished genomes of *Y. pestis* strains that you will be working with:
#1)KIM
#2)CO92
#3)91001
#4)Antiqua
#5)Nepal516

Traditional classification of *Y. pestis* biovars was based on the following three phenotypes.
-Antiqua (glycerol positive, arabinose positive, and nitrate positive)
-Mediaevalis (glycerol positive, arabinose positive, and nitrate negative)
-Orientalis (glycerol negative, arabinose positive, and nitrate positive)

These biovar phenotypes can be traced to the following three genes:
glpD (glycerol)
napA (nitrate)
araC (arabinose)

#1) By examining the three genes (*glpD*, *napA*, and *araC*) in each genome for each of the five strains, determine the biovar classification and determine what type of mutational event (ie, deletion of a region of DNA within the ORF, SNPs, or insertion of additional DNA into the ORF) may have occurred on the genetic level for dysfunctional genes (these dysfunctional genes are referred to as pseudogenes in ASAP):

Historically, this is what has been proposed for the three biovars:
-Biovar Antiqua, from East Africa, may have descended from *Y. pestis* strains that caused the first pandemic
- Medievalis, from central Asia, may have descended from *Y. pestis* strains that caused the second pandemic
-*Y. pestis* strains linked to the third pandemic are all of the Orientalis biovar

#2) Corpses were unearthed from the periods of the first and second pandemics (based on carbon dating) and the DNA for *Y. pestis* from the dental pulp was sequenced (Drancourt *et al*. 2004). Using the available corpse *Y. pestis* DNA (available in the Supplemental course assignment 2 Ypestis corpse and CA88-4125YPE gene sequences) determine which biovar is most similar to the one that caused the first pandemic, and which biovar is most similar to the one that caused the second.

#4) *Y. pestis* strains are believed to have arrived in North America via shipping routes. Now, given the *glpD*, *napA*, and *araC* gene sequences from a strain recently isolated

from North America (genome CA88-4125 YPE) compare the gene sequences from this strain to the five used in the Mauve alignment and based on sequence similarity decide which strain(s) were most likely isolated from North America. Based on your analysis, has the North American lineage (strain CA88-4125 YPE) arrived via the Pacific (biovar Medievalis or Orientalis) or Atlantic trade routes (biovar Antiqua)?

#5) In the five genome Mauve alignment, one strain (91001) has lost the ability to cause disease in higher mammals (humans). Identify a backbone region that is absent in the 91001 genome yet present in the remaining four genomes that are still capable of causing human disease. Analyze the genes in this genomic island and try to identify genes that may play a role in virulence.

#6) Next, provided with one or more gene(s) that have been identified as known or putative virulence factors, determine if the full ORF is conserved in all five of the *Y.pestis* strains used in the Mauve analysis. Are there any that may not functional in the 91001 genome. What would you propose to do to follow up on your findings, to see if your gene(s) are important virulence factors in humans?

Supplemental course assignment #2 (cont.)

Genetic sequences for BlastN analysis against entire genomes of *Y. pestis* in ASAP

Ancient DNA specimen #1 This DNA sequence was amplified from from the dental pulp of a corpse in a grave located in Sens, France carbon dated to 500-600 A.D and thought to represent the 1st pandemic of the black plague.

1 ggtgaggtac agctaaacgg tgatgtcata acgtttctcc caagtttaat ggctacagcg

61 agtcgacaat gacgagcgcc aacaataacg agcgctaaca attaccagtt tcaacgatta

121 ccagctccaa cgattaccag ctccaacaat taccagctcc aacaattacc agctccaaca

181 attatcagtt tcaacaatta cagacgtcga taaagtgaca aataacctac ggcggcaagt

241 tgccaaccaa agtcgagcct catcgcatgg cgctggacgt gacgcca

Ancient DNA specimen #2 This DNA sequence was amplified from the dental pulp of a corpse in a grave located in Dreux, France carbon dated from the 12th to the 14th century and thought to represent the 2nd pandemic of the black plague.

1 acgggtgagg tatagctaaa cggtgatgtc ataacgtttt tcccaagttt aatggctaca

61 acgagtcgac aatgacgagc gccaacaata acgagcgcta acaattacca gtttcaacga

121 ttaccagctc caacgattac cagctccaac aattaccagc tccaacaatt accagctcca

181 acaattatca gtttcaacaa ttacagacgt cgataaagtg acaaataatc tacggcggca

241 agttgccaac caaagtcgag cctcatcgca tggcgctgga cgtgacgcca atgcttcggc

301 ctgctccaca gtacaaaggc acgg

Ancient specimen #3 This DNA sequence is from the dental pulp of a corpse in a grave located in Montpellier, France carbon dated from the 13th and 14th centuries and thought to represent the 2nd pandemic of the black plague.

1 ggtgaggtat agctaaacgg tgatgtcata acgtttttcc caagtttaat ggctacaacg

61 agtcgacaat gacgagcgcc aacaataacg agcgctaaca attaccagtt tcaacgatta

121 ccagctccaa cgattaccag ctccaacaat taccagctcc aacaattacc agctccaaca

181 attatcagtt tcaacaatta cagacgtcga taaagtgaca aataatctac ggcggcaagt

241 tgccaaccaa agtcgagcct catcgcatgg cgctggacgt gacgcca

Genes for *glpD*, *napA*, and *araC* from a known North American strain, *Y. pestis* CA88-4125 (aka YPE) for BlastN analysis against entire genomes of *Y. pestis* in ASAP

glpD

1 ttaagaaacc agcggcagcg cctgttgttt ttcagtgtgc gcatcagcca gccactgggc

61 tacccgctgt ttctcttcat cgctgagggc catgcctaat ttggtacgac gccagatagc

121 atcatccagc tcaacgaccc actcgttctc aaccaaatag cgcaattcag cctcatacaa
181 gccgtgacca aagtgctcac ctaggctttc aagacgggtc gcgttggcta aaattagctc
241 gctgtggcta ccataggtac gggtatagcg gcgggctaac ccttccggca accagttata
301 gcggtggcgt aattgcacgg tatagctatc acgactaccg ccgatatccc caccgggcaa
361 ggcaccggtt ttagtccacg ctgggccgac attcgggtag tacgctgaca gtttttccag
421 cgcatgttct gccaatttac ggtacgtggt gagcttaccg ccgaagaccg acagcagtgg
481 tgcctgaccc gcttcatcgg ccacatctag cgtgtaatcg cgggtaacgg cttgcggtga
541 atctgattca tcgtcgcata gcgggcgcac accagagtag gtccagacga tatcgtcacg
601 acccaactgt tttttaaagt ggtcgttata gactttcagc agataagtca tttcctgatc
661 gtcaattttc acctcttttg gatcgccgtg gtattccacg tcggtagtac cgatgatgga
721 atagtcatct aaccaaggga taacgaaaac gatacggtga tcttcatttt gcagaatata
781 cgcctgcggt tggttatgaa cccgaggcac cacaatgtgg ctgcctttaa ttaggcggat
841 gccataaggt gatttgagct ttaggccatc gtcgaagaac tgtttaaccc aagggccagt
901 ggcattcact aagcctttag cccgccaggt gaaggttttg ccggtattga catcaagggc
961 ttcaaccatc cataggcctt gttcacgcca tgtgcgggtc actttggtac gggttcggac
1021 ttcaccgccg tgtttaacca cttcctgcac attcagcacc accaggcggg catcatccac
1081 ccaacagtca gaatattcga aaccgcgcac caactcgggc tttaacacag attccggccc
1141 aaaacgcagc cctttactgg caggcaggct ggtacgtttg cccaaatggt catacaagaa
1201 caagccggtg cggatcatcc atgccgggcg tagatgaggc tggtggggta ggcggaagcg
1261 catagggaat gcgatatgcg gtgccagttt cagtaacact tcacgctcgg ccaaggcttc
1321 acttaccaag cggaattcat aatgttccag atagcgtaag ccaccatgga taagtttgga
1381 actggcggaa gacgtagcac aggccaagtc tttggtttcc at

napA

1 atgaaactca gtcgccggga ctttatgaaa gcgaatgcgg ccgttgctgc ggctgccgcc
61 gccggaatga ccatacctac tgtcgctaaa gcggtgggtg agacaaccaa tgcgatcaag
121 tgggataaag caccttgccg attctgcggc accggctgcg gtgtactggt aggaacgcaa
181 aatggccgta tcgtggcctc acaaggtgac ccggactcac cggttaaccg tgggctgaac
241 tgcatcaaag gctatttcct gccaaaaatc atgtacggca aagaccggct gacacagcca
301 ctgctgcgta tgaaagacgg tcaatacgat aaagaaggcg atttcacccc aataagctgg
361 gagaaagcct ttgatatcat ggaactgaaa ttcaaaaatg cgctaaaaga gaaaggcccg
421 accgcggtcg gtatgttcgg ctccgggcaa tggaccgtgt gggaaggcta cgcggcgttg
481 aagttgctga aaggggggtt ccgctcaaat aaccttgatc ctaatgcccg ccattgtatg
541 gcgtcctcgg ttgttggatt catgcgtacc ttcggtatgg atgagccgat ggggtgctac
601 gatgatattg aagaagccga tgccttcgtg ctctggggct ccaatatggc ggaaatgcac
661 ccggtattat ggtcgcgtat gaccagccgc cgcctgacca atgcgcatgt cagaattgcc
721 gtcctctcca cttacgaaca ccgcagtttt gaattggccg acaacccgat cgtctttacc

781 ccacaaaccg atctggtcat catgaattac atcgccaatt acatcattca aaataatgcc
841 gttgataaag acttcctggc tcaacacgtg aatttccgcc gcggcgcgac cgatatcggc
901 tatggcttac ggccaaccca tccgttggaa aaagcggcga agaatcccgg cagcgatgct
961 tctgaaccga tgagttttga ggatttcaaa acctttgtcg ctgaatacac gttagaaaaa
1021 gccgccaaaa tgagcggtgt accagaagat cagcttgagt cgttggccca gttgtatgct
1081 gatccaaagg tgaaattggt ctcttactgg accatgggct ttaaccagca tacccgcggc
1141 gtgtgggcca acaacatgtg ctacaacctg cacctgttaa ccggcaagat ctccacgccg
1201 ggatcggggc ctttctccct gacggggcag ccttccgcgt gtggcaccgc ccgcgaagtg
1261 gggacattct cccatcgtct gcctgcagat atggtggtca cgaatgaaaa acatcggcag
1321 attgctgaaa ccacatggca gttaccggcg gggactatcc cggaaaaagt gggtttacat
1381 gcggtagcac aagatcgggc gctgaaagac ggcaccctca acgcctactg ggtgatgtgc
1441 aacaacaaca tgcaggccgg accgaatatt aatgaagagc gtatgccggg ctggcgtgat
1501 ccgcgcaact ttattgtggt ctccgatccc tatcccacca tcagtgcgct gtctgctgac
1561 ttgattttac cgacctcaat gtgggtcgag aaagagggcg catacggcaa tgctgaacgc
1621 cgtactcaat tctggcgtca gcaagtcccc tcaccggggg aggctaaatc ggatttatgg
1681 caaatcgtcg agttcgccaa acgctttaac gtcgaagaag tctggcccgc tgagttggtg
1741 aatcaaaaac ctgagtatcg cggtaaaaat ttatatgagg tgctgtttgc caacgatgta
1801 gtcagtaaat acccactgag cgagatccct gacgatcaat tgaacgacga agcgcgcgat
1861 tttggtttct acatacagaa aggattattt gaagagtacg ccagctttgg gcgtgggcac
1921 gctcatgatc tggctccttt cgatgtatat catcaggtac gcggcctgcg ctggccggtg
1981 gttgacggta aggaaacact ctggcgctac cgtgaaggtt ttgatccctt cgtaccgaaa
2041 ggcgaagagg tgcgcttcta tggcaaacca gacggtaagg cggtcatttt tgccctgcct
2101 tatgaaccag cagcagaaag cccggaccaa gaatatgacc tctggctctc taccggccgg
2161 gtgctggaac attggcacac gggttcaatg acccgccggg tacctgaatt gcaccgtgct
2221 ttcccagagg ccgtgttatt cattcatcca ttggatgcca aagcgcgtgg tttacaccgt
2281 ggtgacaaag tgaaagtgat ttcacgccga ggtgaggtta tttctctggt tgaaacccgt
2341 ggccgtaacc gcccaccgcg agggctggtg tacatgccgt tcttcgatgc cgcacagttg
2401 gtcaataacc tgaccttaga cgcgaccgat ccgctctcga aagaaactga ctttaagaaa
2461 tgcgcagtga aactggaacg ggtagtggcc tga

araC

1 ttaaacccgc accaatgccc cctgattatc ctgccagtcc gctggacgaa aagtacgttg
61 tggatagtta gtttcgatac tgcgccgcct gaaatcgctg gggctgaccc ccactcgttt
121 acggaaaaca cgggaaaaat agagttggtc atcatagcct accacccggc caatggtcgc
181 aattggctct tgggtcgttt gcagtaataa tttcgcccgg atcacccgct gatcctcacg
241 ccaacgtaat atattaatgc caacctgttc acggaataaa tgcgctaagc gtgatggtga
301 taggcaaaca tggcgggcaa cttcgtcaat acgtaattcc cctgccagat ttccggtaat

361 aaattgacag gcctcaataa tacgtgggtc cataatacgt tgtggactaa gtgggtcttc
421 ttccattgct cgtagcagta accgctcgag caaattcata cccagttctt caccgaagcg
481 ccgccccgat ttctgtgtct gctcaatatt ggcaaataag cggtcaaact ccagcattaa
541 gttattgtta ggtaaagata aacgccctac ctcatgggtt ttactgtgcc attccaacca
601 atcggcccaa taagcccgtg gtcggaaata gacccagcgg tgataccaac aatcactatt
661 cggtgaacga ccataatggt gaggcgactt aggtgagaac aacagtaaat caccaggatt
721 actgtagatg gtattttcac catcgaaaat cttcccctgc cccttaatgg tcaaattcag
781 aatatagccc ttcatgccgc caggccgatc aatgaagaaa tcgagtgggc cgtcagccag
841 aatcggggtt aatcctgcga ccagataagc attgaacgtg tagcctggca gcaaaggatt
901 gggttgcggt tcctgaacca tgcgttgata cat

Supplemental course assignment #3

Exploration of LEE gene product function: formulation of concept maps

I would like you to work in small topic-focused groups to construct mini concept maps on a particular aspect of LEE function in EPEC/EHEC pathogenesis.

Learning goal: To gain practice in synthesizing pieces of information from different papers, and in building regulatory networks from individual pieces of data

As is true in any research area, no one person or lab will have all the information to build a complete story. Each student or pair of students has one abstract relevant to the gene you initially explored in our MAUVE lab. You have been asked to download the paper in question and to take a look at the introduction and discussion. You are the "expert" on that piece of data. Design a concept map that depicts the relationships among the proteins in your cluster, addressing the question posed to your group below.

Group 1: How do interactions between host cell proteins and delivered effectors alter the host cell cytoskeleton to bring about pedestal formation and other intracellular changes in the host?
espF, espZ/sepZ, espH, espG

Group 2: How is traffic through the T3SS regulated?
sepL, escN, cesT, cesD2

Group 3: How do interactions among transcriptional regulator proteins and interactions between these proteins and DNA regulate expression of the LEE operons?
grlA, grlR, ler

Group 4: How do the proteins of the T3SS itself work together to assemble a functional injectisome, and what approaches are used to gain insight into such a structure?
escJ, rORF1, espB, espA, espD

Once you have done this, we will reconvene and work as a group to synthesize all the information you have unearthed into one large concept map that captures the entire LEE function.