Homework #1
Plant Breeding 607
Fall '00

Practice safe computing
or
How to avoid STDs (sequence transmitted delusions)

Overview of the problem

In class we have been examining some of the pitfalls in the naïve application of sequence analysis tools. In this homework, you will be asked to explore the extent of one of the biggest problems in sequence analysis – the persistent contamination of sequence databases by cloning vectors. For general information on this problem see the VecScreen and web pages at NCBI. It is a good idea to start from the main VecScreen page, http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html, and follow all of the relevant links that discuss contamination in general, the implementation of the VecScreen system, and the contents of the UniVec database.

As a test, I used blastn with the standard settings to search the Drosophila genome for entries in the UniVec database. The vector sequences with the most significant hits (E values between 0.0 and 10^{-10}) listed in Appendix 1. This test was designed to determine if even state-of-the-art, highly curated sequence data obtained by reputable sequencing laboratories is contaminated.

Questions

A) Use the blastn program on the NCBI BLAST web server to search the Drosophila genome for 5-10 of the longest sequences in Appendix 1. What do you see in the alignments? What can you conclude from this experiment alone?

B) Use the blastn program on NCBI BLAST web server to search the EST database for some of the sequences in Appendix 1. What do you see? Could you be fooled into mis-assigning a function to a vector contamination based on EST annotations?

C) Is it possible that higher organisms have legitimate coding sequences that are similar to vector sequences. How can you use basic facts of molecular biology (e.g. the intro-exon structure of eukaryotic genes) to distinguish between vector contamination and real coding sequences?

D) To test the conjecture in previous question, try using the NCBI BLAST web site to search the SwissProt database for protein sequences that are similar to vector

sequences in Appendix 1. Which variant of BLAST should you use? Do you see any hits? What, if anything, do they mean?

Appendix 1: Possible contaminants of the Drosophila genomic sequence at NCBI from entries in the UniVec database

>gnl|uv|L05083.1:1-447 Cloning vector (pRL278) for sacB-mediated positive selection for double recombinants in gram-negative bacteria
GCTGGCTGTTTTACGCGTATGACAGGCTCCGGAAGACGGTTGTTGCGCACGTATTCGGTGAACGCACTAT
GGCGACGCTGGGGCGTCTTATGAGCCTGCTGTCACCCTTTGACGTGGTGATATGGATGACGGATGGCTGG
CCGCTGTATGAATCCCGCCTGAAGGGAAAGCTGCACGTAATCAGCAAGCGATATACGCAGCGAATTGAGC
GGCATAACCTGAATCTGAGGCAGCACCTGGCACGGCTGGGACGGAAGTCGCTGTCGTTCTCAAAATCGGT
GGAGCTGCATGACAAAGTCATCGGGCATTATCTGAACATAAAACACTATCAATAAGTTGGAGTCATTACC
AAAAGGTTAGGAATACGGTTAGCCATTTGCCTGCTTTTATATAGTTCATATGGGATTCACCTTTATGTTG
ATAAGAAATAAAAGAAAATGCCAATAG
>gnl|uv|L05083.1:5894-5941-49 Cloning vector (pRL278) for sacB-mediated positive selection for double recombinants in gram-negative bacteria
TGCATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGGCTGGCTGTTTTACGCGTATGA
CAGGCTCCGGAAGACGGTTGTTGCGCA
>gnl|uv|U87107.1:8348-8840-49 Cloning vector pAL-F insertion sequence IS1 galactokinase (galK), aminoglycoside 3'-phosphotransferase (kn), beta-galactosidase (lacZ), small ribosomal protein and beta-lactamase (Ap) genes, complete cds
TCGCCATAGTGCGTTCACCGAATACGTGCGCAACAACCGTCTTCCGGAGACTGTCATACGCGTAAAACAG
CCAGCGCTGGCGCGATTTAGCCCCGACATAGCCCCACTGTTCGTCCATTTCCGCGCAGACGATGACGTCA
CTGCCCGGCTGTATGCGCGAGGTTACCGACTGCGGCCTGAGTTTTTTAAGTGACGTAAAATCGTGTTGAG
GCCAACGCCCATAATGCGGGCTGTTGCCCGGCATCCAACGCCATTCATGGCCATATCAATGATTTTCTGG
TGCGTACCGGGTTGAGAAGCGGTGTAAGTGAACTGCAGTTGCCATGTTTTACGGCAGTGAGAGCAGAGAT
AGCGCTGATGTCCGGCGGTGCTTTTGCCGTTACGCACCACCCCGTCAGTAGCTGAACAGGAGGGACAGCT
GATAGAAACAGAAGCCACTGGAGCACCTCAAAAACACCATCATACACTAAATCAGTAAGTTGGCAGCATC
ACCCAATTACTGCAATGCCCTCGTAATTAAGTGAATTTACAATATCGTCCTG
>gnl|uv|U09365.1:1-1365 Broad host range plasmid Bin 19
CCGGGCTGGTTGCCCTCGCCGCTGGGCTGGCGGCCGTCTATGGCCCTGCAAACGCGCCAGAAACGCCGTC
GAAGCCGTGTGCGAGACACCGCGGCCGCCGGCGTTGTGGATACCTCGCGGAAAACTTGGCCCTCACTGAC
AGATGAGGGGCGGACGTTGACACTTGAGGGGCCGACTCACCCGGCGCGGCGTTGACAGATGAGGGGCAGG
CTCGATTTCGGCCGGCGACGTGGAGCTGGCCAGCCTCGCAAATCGGCGAAAACGCCTGATTTTACGCGAG
TTTCCCACAGATGATGTGGACAAGCCTGGGGATAAGTGCCCTGCGGTATTGACACTTGAGGGGCGCGACT
ACTGACAGATGAGGGGCGCGATCCTTGACACTTGAGGGGCAGAGTGCTGACAGATGAGGGGCGCACCTAT
TGACATTTGAGGGGCTGTCCACAGGCAGAAAATCCAGCATTTGCAAGGGTTTCCGCCCGTTTTTCGGCCA
CCGCTAACCTGTCTTTTAACCTGCTTTTAAACCAATATTTATAAACCTTGTTTTTAACCAGGGCTGCGCC
CTGTGCGCGTGACCGCGCACGCCGAAGGGGGGTGCCCCCCCTTCTCGAACCCTCCCGGCCCGCTAACGCG
GGCCTCCCATCCCCCCAGGGGCTGCGCCCCTCGGCCGCGAACGGCCTCACCCCAAAAATGGCAGCGCTGG
CAGTCCTTGCCATTGCCGGGATCGGGGCAGTAACGGGATGGGCGATCAGCCCGAGCGCGACGCCCGGAAG
CATTGACGTGCCGCAGGTGCTGGCATCGACATTCAGCGACCAGGTGCCGGGCAGTGAGGGCGGCGGCCTG
GGTGGCGGCCTGCCCTTCACTTCGGCCGTCGGGGCATTCACGGACTTCATGGCGGGGCCGGCAATTTTTA
CCTTGGGCATTCTTGGCATAGTGGTCGCGGGTGCCGTGCTCGTGTTCGGGGGTGCGATAAACCCAGCGAA
CCATTTGAGGTGATAGGTAAGATTATACCGAGGTATGAAAACGAGAATTGGACCTTTACAGAATTACTCT
ATGAAGCGCCATATTTAAAAAGCTACCAAGACGAAGAGGATGAAGAGGATGAGGAGGCAGATTGCCTTGA
ATATATTGACAATACTGATAAGATAATATATCTTTTATATAGAAGATATCGCCGTATGTAAGGATTTCAG
GGGGCAAGGCATAGGCAGCGCGCTTATCAATATATCTATAGAATGGGCAAAGCATAAAAACTTGCATGGA
CTAATGCTTGAAACCCAGGACAATAACCTTATAGCTTGTAAATTCTATCATAATTGGGTAATGACTCCAA
CTTATTGATAGTGTTTTATGTTCAGATAATGCCCG
>gnl|uv|U09365.1:1969-6239 Broad host range plasmid Bin 19

CGTTACGCACCACCCCGTCAGTAGCTGAACAGGAGGGACAGCTGATAGACACAGAAGCCACTGGAGCACC
TCAAAAACACCATCATACACTAAATCAGTAAGTTGGCAGCATCACCCATAATTGTGGTTTCAAAATCGGC
TCCGTCGATACTATGTTATACGCCAACTTTGAAAACAACTTTGAAAAAGCTGTTTTCTGGTATTTAAGGT
TTTAGAATGCAAGGAACAGTGAATTGGAGTTCGTCTTGTTATAATTAGCTTCTTGGGGTATCTTTAAATA
CTGTAGAAAAGAGGAAGGAAATAATAAATGGCTAAAATGAGAATATCACCGGAATTGAAAAAACTGATCG
AAAAATACCGCTGCGTAAAAGATACGGAAGGAATGTCTCCTGCTAAGGTATATAAGCTGGTGGGAGAAAA
TGAAAACCTATATTTAAAAATGACGGACAGCCGGTATAAAGGGACCACCTATGATGTGGAACGGGAAAAG
GACATGATGCTATGGCTGGAAGGAAAGCTGCCTGTTCCAAAGGTCCTGCACTTTGAACGGCATGATGGCT
GGAGCAATCTGCTCATGAGTGAGGCCGATGGCGTCCTTTGCTCGGAAGAGTATGAAGATGAACAAAGCCC
TGAAAAGATTATCGAGCTGTATGCGGAGTGCATCAGGCTCTTTCACTCCATCGACATATCGGATTGTCCC
TATACGAATAGCTTAGACAGCCGCTTAGCCGAATTGGATTACTTACTGAATAACGATCTGGCCGATGTGG
ATTGCGAAAACTGGGAAGAAGACACTCCATTTAAAGATCCGCGCGAGCTGTATGATTTTTTAAAGACGGA
AAAGCCCGAAGAGGAACTTGTCTTTTCCCACGGCGACCTGGGAGACAGCAACATCTTTGTGAAAGATGGC
AAAGTAAGTGGCTTTATTGATCTTGGGAGAAGCGGCAGGGCGGACAAGTGGTATGACATTGCCTTCTGCG
TCCGGTCGATCAGGGAGGATATCGGGGAAGAACAGTATGTCGAGCTATTTTTTGACTTACTGGGGATCAA
GCCTGATTGGGAGAAAATAAAATATTATATTTTACTGGATGAATTGTTTTAGTACCTAGATGTGGCGCAA
CGATGCCGGCGACAAGCAGGAGCGCACCGACTTCTTCCGCATCAAGTGTTTTGGCTCTCAGGCCGAGGCC
CACGGCAAGTATTTGGGCAAGGGGTCGCTGGTATTCGTGCAGGGCAAGATTCGGAATACCAAGTACGAGA
AGGACGGCCAGACGGTCTACGGGACCGACTTCATTGCCGATAAGGTGGATTATCTGGACACCAAGGCACC
AGGCGGGTCAAATCAGGAATAAGGGCACATTGCCCCGGCGTGAGTCGGGGCAATCCCGCAAGGAGGGTGA
ATGAATCGGACGTTTGACCGGAAGGCATACAGGCAAGAACTGATCGACGCGGGGTTTTCCGCCGAGGATG
CCGAAACCATCGCAAGCCGCACCGTCATGCGTGCGCCCCGCGAAACCTTCCAGTCCGTCGGCTCGATGGT
CCAGCAAGCTACGGCCAAGATCGAGCGCGACAGCGTGCAACTGGCTCCCCCTGCCCTGCCCGCGCCATCG
GCCGCCGTGGAGCGTTCGCGTCGTCTCGAACAGGAGGCGGCAGGTTTGGCGAAGTCGATGACCATCGACA
CGCGAGGAACTATGACGACCAAGAAGCGAAAAACCGCCGGCGAGGACCTGGCAAAACAGGTCAGCGAGGC
CAAGCAGGCCGCGTTGCTGAAACACACGAAGCAGCAGATCAAGGAAATGCAGCTTTCCTTGTTCGATATT
GCGCCGTGGCCGGACACGATGCGAGCGATGCCAAACGACACGGCCCGCTCTGCCCTGTTCACCACGCGCA
ACAAGAAAATCCCGCGCGAGGCGCTGCAAAACAAGGTCATTTTCCACGTCAACAAGGACGTGAAGATCAC
CTACACCGGCGTCGAGCTGCGGGCCGACGATGACGAACTGGTGTGGCAGCAGGTGTTGGAGTACGCGAAG
CGCACCCCTATCGGCGAGCCGATCACCTTCACGTTCTACGAGCTTTGCCAGGACCTGGGCTGGTCGATCA
ATGGCCGGTATTACACGAAGGCCGAGGAATGCCTGTCGCGCCTACAGGCGACGGCGATGGGCTTCACGTC
CGACCGCGTTGGGCACCTGGAATCGGTGTCGCTGCTGCACCGCTTCCGCGTCCTGGACCGTGGCAAGAAA
ACGTCCCGTTGCCAGGTCCTGATCGACGAGGAAATCGTCGTGCTGTTTGCTGGCGACCACTACACGAAAT
TCATATGGGAGAAGTACCGCAAGCTGTCGCCGACGGCCCGACGGATGTTCGACTATTTCAGCTCGCACCG
GGAGCCGTACCCGCTCAAGCTGGAAACCTTCCGCCTCATGTGCGGATCGGATTCCACCCGCGTGAAGAAG
TGGCGCGAGCAGGTCGGCGAAGCCTGCGAAGAGTTGCGAGGCAGCGGCCTGGTGGAACACGCCTGGGTCA
ATGATGACCTGGTGCATTGCAAACGCTAGGGCCTTGTGGGGTCAGTTCCGGCTGGGGGTTCAGCAGCCAG
CGCTTTACTGGCATTTCAGGAACAAGCGGGCACTGCTCGACGCACTTGCTTCGCTCAGTATCGCTCGGGA
CGCACGGCGCGCTCTACGAACTGCCGATAAACAGAGGATTAAAATTGACAATTGTGATTAAGGCTCAGAT
TCGACGGCTTGGAGCGGCCGACGTGCAGGATTTCCGCGAGATCCGATTGTCGGCCCTGAAGAAAGCTCCA
GAGATGTTCGGGTCCGTTTACGAGCACGAGGAGAAAAAGCCCATGGAGGCGTTCGCTGAACGGTTGCGAG
ATGCCGTGGCATTCGGCGCCTACATCGACGGCGAGATCATTGGGCTGTCGGTCTTCAAACAGGAGGACGG
CCCCAAGGACGCTCACAAGGCGCATCTGTCCGGCGTTTTCGTGGAGCCCGAACAGCGAGGCCGAGGGGTC
GCCGGTATGCTGCTGCGGGCGTTGCCGGCGGGTTTATTGCTCGTGATGATCGTCCGACAGATTCCAACGG
GAATCTGGTGGATGCGCATCTTCATCCTCGGCGCACTTAATATTTCGCTATTCTGGAGCTTGTTGTTTAT
TTCGGTCTACCGCCTGCCGGGCGGGGTCGCGGCGACGGTAGGCGCTGTGCAGCCGCTGATGGTCGTGTTC
ATCTCTGCCGCTCTGCTAGGTAGCCCGATACGATTGATGGCGGTCCTGGGGGCTATTTGCGGAACTGCGG
GCGTGGCGCTGTTGGTGTTGACACCAAACGCAGCGCTAGATCCTGTCGGCGTCGCAGCGGGCCTGGCGGG
GGCGGTTTCCATGGCGTTCGGAACCGTGCTGACCCGCAAGTGGCAACCTCCCGTGCCTCTGCTCACCTTT
ACCGCCTGGCAACTGGCGGCCGGAGGACTTCTGCTCGTTCCAGTAGCTTTAGTGTTTGATCCGCCAATCC
CGATGCCTACAGGAACCAATGTTCTCGGCCTGGCGTGGCTCGGCCTGATCGGAGCGGGTTTAACCTACTT
CCTTTGGTTCCGGGGGATCTCGCGACTCGAACCTACAGTTGTTTCCTTACTGGGCTTTCTCAGCCCCAGA
TCTGGGGTCGATCAGCCGGGGATGCATCAGGCCGACAGTCGGAACTTCGGGTCCCCGACCTGTACCATTC
GGTGAGCAATGGATAGGGGAGTTGATATCGTCAACGTTCACTTCTAAAGAAATAGCGCCACTCAGCTTCC
TCAGCGGCTTTATCCAGCGATTTCCTATTATGTCGGCATAGTTCTCAAGATCGACAGCCTGTCACGGTTA
AGCGAGAAATGAATAAGAAGGCTGATAATTCGGATCTCTGCGAGGGAGATGATATTTGATCACAGGCAGC
AACGCTCTGTCATCGTTACAATCAACATGCTACCCTCCGCGAGATCATCCGTGTTTCAAACCCGGCAGCT

TAGTTGCCGTTCTTCCGAATAGCATCGGTAACATGAGCAAAGTCTGCCGCCTTACAACGGCTCTCCCGCT
GACGCCGTCCCGGACTGATGGGCTGCCTGTATCGAGTGGTGATTTTGTGCCGAGCTGCCGGTCGGGGAGC
TGTTGGCTGGCTGGTGGCAGGATATATTGTGGTGTAAACAAATTGACGCTTAGACAACTTAATAACACAT
TGCGGACGTTTTTAATGTACTGGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCT
T
>gnl|uv|U75991.1:10205-10303 Cloning vector pBACe4
TTTTCCTCGCTTCCGGGCTGTCATCATTAAACTGTGCAATGGCGATAGCTTTCGTCATTTCATGACCAGC
GTTTATGCACTGGTTAAGTGTTTCCATGA
>gnl|uv|M64731.1:380-490 Cloning vector pSU2718 DNA
GGATTTATTTATTCTGCGAAGTGATCTTCCGTCACAGGTATTTATTCGGCGCAAAGTGCGTCGGGTGATG
CTGCCAACTTACTGATTTAGTGTATGATGGTGTTTTTGAGG
>gnl|uv|M64731.1:536-628 Cloning vector pSU2718 DNA
TACTGACGGGGTGGTGCGTAACGGCAAAAGCACCGCCGGACATCAGCGCCATTCGCCATTCAGGCTGCGC
AACTGTTGGGAAGGGCGATCGGT
>gnl|uv|X06403.1:538-626 Cloning vector pACYC184
CTGACGGGGTGGTGCGTAACGGCAAAAGCACCGCCGGACATCAGCGCTAGCGGAGTGTATACTGGCTTAC
TATGTTGGCACTGATGAGG
>gnl|uv|U13761.1:763-860 Cloning vector pDO192
CTCAACACGATTTTACGTCACTTAAAAAACTCAGGCCGCAGTCGGTAACATCGTCCATTCCGACAGCATC
GCCAGTCACTATGGCGTGCTGCTAGCTT
>gnl|uv|U13761.1:6428-6524 Cloning vector pDO192
TTCACCGAATACGTGCGCAACAACCGTCTTCCGGAGACTGTCATACGCGCATTCCCTGTTCCTGCCAGTT
CTGAATGGTACGGATACTCGCACCGAA
>gnl|uv|L08817.1:8943-9038 pAA3.7X cloning vector
CTACCGGAAGCAGTGTGACCGTGTGCTTCTCAAATGCCTGAGGCCAGTTGGTAATGACTCCAACTTATTG
ATAGTGTTTTATGTTCAGATAATGCC
>gnl|uv|L08817.1:9536-9583-49 pAA3.7X cloning vector
TGATTTTCTGGTGCGTACCGGGTTGAGAAGCGGTGTAAGTGAACTGCAGGCCTTTTGAAAAGCAAGCATA
AAAGATCAACAGTCGGTACGGCTGACC
>gnl|uv|U03463.1:764-939 Cloning vector pDO6
TCAACACGATTTTACGTCACTTAAAAAACTCAGGCCGCAGTCGGTAACCGGCCTCTTCATCGGGAATGCG
CGCGACCTTCAGCATCGCCGGCATGTCCCCCTGGCGGACGGGAAGTATCCAGCTCGATCAGGCAGCAGCT
GAACCAACTCGCGAGGGGATCGAGCCCGGGGTGGGC
>gnl|uv|U03463.1:8267-8360 Cloning vector pDO6
TTCGCTGCGTATATCGCTTGCTGATTACGTGCAGCTTTCCCTTCAGGCGTACGCATGACCGGCGGCGGAT
TTGGCGGCTGTATCGTCGCGCTGA
>gnl|uv|U03462.1:8530-8621 Cloning vector pDO2
TCACCGAATACGTGCGCAACAACCGTCTTCCGGAGACTGTCATACGCGTGCGTCATATACTGACTGAAAA
CGCCCGCACCGTTGAAGCTGCC
>gnl|uv|U03460.1:842-1010 Cloning vector pDO184
TCAGCATCGCCGGCATGTCCCCCTGGCGGACGGGAAGTATCCAGCTCGACATCGTCTGCGCGGAAATGGA
CGAACAGTGGGGCTATGTCGGGGCTAAATCGCGCCAGCGCTGGCTGTTTTACGCGTTTTGCCACGATCGG
GTCCAGTTCATGCGCAACAGCGTTGAACT
>gnl|uv|U03461.1:5370-5459 Cloning vector pDO19
CCGCAGCTTAGACCAAAACAGGAAGCTATGGGCCTGCTTAGGTGACGTCATCGTCTGCGCGGAAATGGAC
GAACAGTGGGGCTATGTCGG
>gnl|uv|X53938.1:1293-1389 Plasmid vector pSU 36
GCTTTTGCCGTTACGCACCACCCCGTCAGTAGCTGAACAGGAGGGACAGGGGGTGGGCGAAGAACTCCAG
CATGAGATCCCCGCGCTGGAGGATCAT
>gnl|uv|U55760.1:1-2749 Baculovirus transfer vector pBacPAK-His3, beta-
lactamase (bla) gene, complete cds
TCGACATGCTGAACAACAAGATTAATATGCCTCCGTGTATAAAAAAAATATTGAACGATTTGAAAGAAAA
CAATGTACCGCGCGGCGGTATGTACAGGAAGAGGTTTATACTAAACTGTTACATTGCAAACGTGGTTTCG
TGTGCCAAGTGTGAAAACCGATGTTTAATCAAGGCTCTGACGCATTTCTACAACCACGACTCCAAGTGTG
TCGGTGAAGTCATGCACCTTTTAATCAAATCCCAAGATGTGTATAAACCACCAAACTGCCAAAAAATGAA
AACTGTCGATAAGCTCTGTCCGTTTGCTGGCAACTGCAAGGGTCTCAATCCTATTTGTAATTATTGAATA
ATAAAACAATTATAAATGCTAAATTTGTTTTTTATTAACGATACAAACCAAACGCAACAAGAACATTTGT

AGTATTATCTATAATTGAAAAAGCGTAGTTATAATCGCTGAGGTAATATTTAAAATCATTTTCAAATGAT
TCACAGTTAATTTGCGACAATATAATTTTATTTTCACATAAACTAGACGCCTTGTCGTCTTCTTCTTCGT
ATTCCTTCTCTTTTTCATTTTTCTCCTCATAAAAATTAACATAGTTATTATCGTATCCATATATGTATCT
ATCGTATAGAGTAAATTTTTTGTTGTCATAAATATATATGTCTTTTTTAATGGGGTGTATAGTACCGCTG
CGCATAGTTTTTCTGTAATTTACAACAGTGCTATTTTCTGGTAGTTCTTCGGAGTGTGTTGCTTTAATTA
TTAAATTTATATAATCAATGAATTTGGGATCGTCGGTTTTGTACAATATGTTGCCGGCATAGTACGCAGC
TTCTTCTAGTTCAATTACACCATTTTTTAGCAGCACCGGATTAACATAACTTTCCAAAATGTTGTACGAA
CCGTTAAACAAAAACAGTTCACCTCCCTTTTCTATACTATTGTCTGCGAGCAGTTGTTTGTTGTTAAAAA
TAACAGCCATTGTAATGAGACGCACAAACTAATATCACAAACTGGAAATGTCTATCAATATATAGTTGCT
GATATCATGGAGATAATTAAAATGATAACCATCTCGCAAATAAATAAGTATTTTACTGTTTTCGTAACAG
TTTTGTAATAAAAAAACCTATAAATACCATGGGTCATCATCATCACCATCACGGGTCGACAAGCTTGGAT
CCCTGCAGGCCTCGAGTTCGAATCTAGAAGATCTGGTACCGAGCTCGAATTCCCGGGCGGCCGCTTAATT
AATTGATCCGGGTTATTAGTACATTTATTAAGCGCTAGATTCTGTGCGTTGTTGATTTACAGACAATTGT
TGTACGTATTTTAATAATTCATTAAATTTATAATCTTTAGGGTGGTATGTTAGAGCGAAAATCAAATGAT
TTTCAGCGTCTTTATATCTGAATTTAAATATTAAATCCTCAATAGATTTGTAAAATAGGTTTCGATTAGT
TTCAAACAAGGGTTGTTTTTCCGAACCGATGGCTGGACTATCTAATGGATTTTCGCTCAACGCCACAAAA
CTTGCCAAATCTTGTAGCAGCAATCTAGCTTTGTCGATATTCGTTTGTGTTTTGTTTTGTAATAAAGGTT
CGACGTCGTTCAAAATATTATGCGCTTTTGTATTTCTTTCATCACTGTCGTTAGTGTACAATTGACTCGA
CGTAAACACGTTAAATAACGCTTGGACATATTTAACATCGGGCGTGTTAGCTTTATTAGGCCGATTATCG
TCGTCGTCCCAACCCTCGTCGTTAGAAGTTGCTTCCGAAGACGATTTTGCCATAGCCACACGACGCCTAT
TAATTGTGTCGGCTAACACGTCCGCGATCAAATTTGTAGTTGAGCTTTTTGGAATTATTTCTGATTGCGG
GCGTTTTTGGGCGGGTTTCAATCTAACTGTGCCCGATTTTAATTCAGACAACACGTTAGAAAGCGATGGT
GCAGGCGGTGGTAACATTTCAGACGGCAAATCTACTAATGGCGGCGGTGGTGGAGCTGATGATAAATCTA
CCATCGGTGGAGGCGCAGGCGGGGCTGGCGGCGGAGGCGGAGGCGGAGGTGGTGGCGGTGATGCAGACGG
CGGTTTAGGCTCAAATGTCTCTTTAGGCAACACAGTCGGCACCTCAACTATTGTACTGGTTTCGGGCGCC
GTTTTTGGTTTGACCGGTCTGAGACGAGTGCGATTTTTTTCGTTTCTAATAGCTTCCAACAATTGTTGTC
TGTCGTCTAAAGGTGCAGCGGGTTGAGGTTCCGTCGGCATTGGTGGAGCGGGCGGCAATTCAGACATCGA
TGGTGGTGGTGGTGGTGGAGGCGCTGGAATGTTAGGCACGGGAGAAGGTGGTGGCGGCGGTGCCGCCGGT
ATAATTTGTTCTGGTTTAGTTTGTTCGCGCACGATTGTGGGCACCGGCGCAGGCGCCGCTGGCTGCACAA
CGGAAGGTCGTCTGCTTCGAGGCAGCGCTTGGGGTGGTGGCAATTCAATATTATAATTGGAATACAAATC
GTAAAAATCTGCTATAAGCATTGTAATTTCGCTATCGTTTACCGTGCCGATATTTAACAACCGCTCAATG
TAAGCAATTGTATTGTAAAGAGATTGTCTCAAGCTCGGATCGATCCCGCACGCCGATAACAAGCCTTTTC
ATTTTTACTACAGCATTGTAGTGGCGAGACACTTCGCTGTCGTCGCCTGATGCGGTATTTTCTCCTTACG
CATCTGTGCGGTATTTCAC
>gnl|uv|AF098285.1:118-350 Cloning vector pFR-Bgal
GTCCTCCGAGCGGAGACTCTAGAGGGTATATAATGGATCCCCGGGTACCAGCTTGGGGGATCGAAAGAGC
CTGCTAAAGCAAAAAAGAAGTCACCATGTCGTTTACTTTGACCAACAAGAACGTGATTTTCGTTGCCGGT
CTGGGAGGCATTGGTCTGGACACCAGCAAGGAGCTGCTCAAGCGCGATCCCGTCGTTTTACAACGTCGTG
ACTGGGAAAACCCTGGCGTTACC
>gnl|uv|U89935.1:560-659 Cloning vector pBI-GL
CCGCGGCCCCGAATTCGAGCTCGGTACCCGGGGATCCTCTAGAGGATCCCCGGGATCGAAAGAGCCTGCT
AAAGCAAAAAAGAAGTCACCATGTCGTTTA
>gnl|uv|U02433.1:1-258 Cloning vector pNASSbeta
GAATTCGAGCTCGGTACTCGAGGAACTGAAAAACCAGAAAGTTAACTGGTAAGTTTAGTCTTTTTGTCTT
TTATTTCAGGTCCCGGATCCGGTGGTGGTGCAAATCAAAGAACTGCTCCTCAGTGGATGTTGCCTTTACT
TCTAGGCCTGTACGGAAGTGTTACTTCTGCTCTAAAAGCTGCGGAATTGTACCCGCGGCCGCAATTCCCG
GGGATCGAAAGAGCCTGCTAAAGCAAAAAAGAAGTCACCATGTCGTTT
>gnl|uv|U13184.1:6-96 Cloning vector pbetagal-Basic
GAGGTACCGAGCTCTTACGCGTGCTAGCTCGAGATCTAAGTAAGCTTGGGGGATCGAAAGAGCCTGCTAA
AGCAAAAAAGAAGTCACCATG
>>gnl|uv|U13187.1:199-287 Cloning vector pbetagal-Promoter
GTAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAAAGCTTGGGGGATCGAAAGAGCCTGCTAAA
GCAAAAAAGAAGTCACCAT
>gnl|uv|M37847.1:848-1034 Bacterial cloning vector pMMB206
GGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTGATGTCCGGCGGTGCTTTTGC
CGTTACGCACCACCCCGTCAGTAGCTGAACAGGAGGGACAGCTGATAGAAACAGAAGCCACTGGAGCAAG
CACCTCAAAAACACCATCATACACTAAATCAGTAAGTTGGCAGCATC

>gnl|uv|U35235.1:534-629 Plasmid pBSL167 cloning vector
GCTACTGACGGGGTGGTGCGTAACGGCAAAAGCACCGCCGGACATCACGCGCGTAATACGACTCACTATA
GGGCGAATTGGGTACCGGGCCCCCCC
>gnl|uv|M37852.1:43-109 Cloning vector pMMB206 beta lactamase (bla,
truncated), 3' end
TGCCGTTACGCACCACCCCGTCAGTAGCTGAACAGGAGGGACAGCTGATGAAAACAGAAGCCACTGG
>gnl|uv|X06779.1:1-4734 Complete sequence of transformation vector
pPI25.1
GGATCCCAATCCGCATCGAATAAGACGCAAACTTTTGCATCGTATATTGAAAAGATGCGCTATCTCCATG
GCAAAGTTTTTGCGGTCGTCGCAGACAAAAGACGCGTGTTCCCTTTTAAAGTTTCATTAATCAATGGCTC
GGGTGGTGGATGTGGATGTGGCTGTAGCTATTGATGTAGCTGGAGGGGTTCTGCCACAGCTCCGGCTATA
GCTTTTATTATTTCAATTGCGGAACGGTGCCCGGGGCTCAAAAATATGAGGCATCATTTATCAGGCCGCC
ATTCCACCACTCCACTACGTCATACCATTCCGTTCTGGTGCGGATTGGCTGTCTGGGCTGTGGATTTCAA
ATTGGCAGCAACAATGCCGTGATTAATGGCCGTCAAAGCGAAAGAGTCAACAAATGTCAAGCGACAAACT
GCAATTGGCAATTGGAGGGAGGGGGGGTGGTGGTGGTGGTAAAGAGGACCAAATCCCAAAAAACCCCATA
CCATAACATATACCATATACCATACTATACCATGGTAGGAATCCCTTGGCCCGCGATGAAATGAATGCCA
ACACAAGTGCGGGTTAATTTTTTCAAATGTTTGTTAATTACATTTATTGTTTCCTTTTGTTTTTGCATTT
GCATTTGTTTTTATGTTTTTTTTTCTCTCGTTTCTCTTTAGTTTTGATTTCTTTCACTTGCAATCGTTCT
GGCTGCTTTCATTATTACTGATTGCCGGCCAAAACTCGGATTATGGTTTGTTTTTGTAATGTATATATAT
GCACGTATATATAAGTCTAAGCGGGTTTATCTTACACTTTAGATATTTTTGATTTAACTTTATGCCTGCT
CTCTAACAATACAATACAAAAAAAATATTGGATAATCGCTCTTACTTATTTCCGATTATTAGATATTATT
ACGATATTTAACATAGTTTCGGTAATTCGATAGATAGGTAGTATTGTTTATCGCTCGTAGATGATAGACT
GTTGCTTATACACACCATGATGAAATAACATAAGGTGGTCCCGTCGAAAGCCGAAGCTTACCGAAGTATA
CACTTAAATTCAGTGCACGTTTGCTTGTTGAGAGGAAAGGTTGTGTGCGGACGAATTTTTTTTTGAAAAC
ATTAACCCTTACGTGGAATAAAAAAAAATGAAATATTGCAAATTTTGCTGCAAAGCTGTGACTGGAGTAA
AATTAATTCACGTGCCGAAGTGTGCTATTAAGAGAAAATTGTGGGAGCAGAGCCTTGGGTGCAGCCTTGG
TGAAAACTCCCAAATTTGTGATACCCACTTTAATGATTCGCAGTGGAAGGCTGCACCTGCAAAAGGTCAG
ACATTTAAAAGGAGGCGACTCAACGCAGATGCCGTACCTAGTAAAGTGATAGAGCCTGAACCAGAAAAGA
TAAAAGAAGGCTATACCAGTGGGAGTACACAAACAGAGTAAGTTTGAATAGTAAAAAAAATCATTTATGT
AAACAATAACGTGACTGTGCGTTAGGTCCTGTTCATTGTTTAATGAAAATAAGAGCTTGAGGGAAAAAAT
TCGTACTTTGGAGTACGAAATGCGTCGTTTAGAGCAGCAGCTGAGGGAGTCTCAACAGTTGGAGGAGTCT
CTACGCAAAATCTTCACGGACACGCAGATACGGATACTGAAGAATGGTGGACAAAGAGCTACGTTCAATT
CCGACGACATTTCTACAGCTATTTGTCTCCACACCGCAGGCCCTCGAGCGTATAACCATCTGTACAAAAA
AGGATTTCCTTTGCCCAGTCGTACGACTTTGTACAGATGGTTATCAGATGTGGACATAAAAAGAGGATGT
TTGGATGTGGTCATAGACCTAATGGACAGTGATGGAGTTGATGACGCCGACAAGCTTTGCGTACTCGCTT
TCGACGAGATGAAGGTCGCTGCTGCCTTCGAGTATGACAGCTCTGCTGATATTGTTTACGAGCCAAGCGA
CTATGTCCAACTGGCTATTGTTCGTGGTCTAAAAAAATCGTGGAAGCAGCCAGTTTTTTTCGATTTTAAT
ACCCGAATGGACCCGGATACTCTTAACAATATATTAAGGAAACTGCATAGGAAAGGATATTTAGTAGTTG
CTATTGTATCCGATTTAGGTACCGGAAACCAAAAGCTATGGACAGAGCTCGGTATATCAGAATGTAAGTT
TCGTATATTACAAAAATCAGATAATCCTTGAAATTCCATTTTTTAGCAAAAACCTGGTTTAGCCATCCTG
CAGATGACCATTTAAAGATTTTCGTTTTTTCGGATACGCCACATTTAATTAAGTTAGTCCGTAACCACTA
TGTGGATTCCGGATTAACAATAAATGGGAAAAAATTAACAAAAAAAACAATTCAGGAGGCACTTCATCTT
TGCAACAAGTCCGATCTGTCTATCCTCTTTAAAATTAATGAAAATCACATTAATGTTCGATCGCTCGCAA
AACAGAAGGTTAAATTGGCTACCCAGCTGTTTTCGAATACCACCGCTAGCTCGATCAGACGCTGCTATTC
ATTGGGGTATGACATTGAAAATGCCACCGAAACTGCGGACTTCTTCAAATTGATGAATGATTGGTTCGAC
ATTTTTAATTCTAAATTGTCCACATCCAATTGCATTGAGTGCTCGCAACCTTATGGCAAGCAGTTGGATA
TACAGAATGATATTTTGAATCGAATGTCGGAAATTATGCGAACAGGAATTCTGGATAAACCCAAAAGGCT
CCCATTTCAAAAAGGTATCATTGTGAATAATGCTTCGCTTGATGGCTTGTATAAATATTTGCAAGAAAAC
TTCAGTATGCAATACATATTAACAAGCCGTCTCAACCAAGACATTGTGGAGCATTTTTTTGGCAGCATGC
GATCGAGAGGTGGACAATTCGACCATCCCACTCCACTGCAGTTTAAGTATAGGTTAAGAAAATATATAAT
AGGTATGACAAATTTAAAAGAATGCGTAAACAAAAATGTAATTCCATGATTTATAATTGTTTAATGTTTA
GCTATATGTTTCAGGAAAGTTTCAGTTGAGAATGTAGGTAGTTATGTGCTGTCTATTGTGTTTTGTCTTT
TATCTGTTTCTTTTCATTTTATTATTTAATCATTATCCTTTTGCTTATCCAGCCAGGAATACAGAAATGT
TAAGAAATTCGGGAAATATCGAAGAGGACAACTCTGAAAGCTGGCTTAATTTAGATTTCAGTTCTAAAGA
AAACGAAAATAAAAGTAAAGATGATGAGCCTGTCGATGATGAGCCTGTCGATGAGATGTTAAGCAATATA
GATTTCACCGAAATGGATGAGTTGACGGAGGATGCGATGGAATATATCGCGGGCTATGTCATTAAAAAAT
TGAGAATCAGTGACAAAGTAAAAGAAAATTTGACATTTACATACGTCGACGAGGTGTCTCACGGCGGACT

TATTAAGCCGTCCGAAAAATTTCAAGAGAAGTTAAAAGAGCTAGAATGTATTTTTTTGCATTATACAAAT
AATAATAATTTTGAAATTACAAATAATGTAAAGGAAAAATTAATATTAGCAGCGCGAAACGTCGATGTTG
ATAAACAAGTAAAATCTTTTTATTTTAAAATTAGAATATATTTTAGAATTAAGTACTTCAACAAAAAAAT
TGAAATTAAAAATCAAAAACAAAAGTTAATTGGAAACTCCAAATTATTAAAAATAAAACTTTAAAAATAA
TTTCGTCTAATTAATATTATGAGTTAATTCAAACCCCACGGACATGCTAAGGGTTAATCAACAATCATAT
CGCTGTCTCACTCAGACTCAATACGACACTCAGAATACTATTCCTTTCACTCGCACTTATTGCAAGCATA
CGTTAAGTGGATGTCTCTTGCCGACGGGACCACCTTATGTTATTTCATCATGATACACACACTTTGGTTT
CTCCCGTACTTAGTATATTCGCTAGCTCCTTTCGAACGCTAATTACGCATATTGATTAGTGTTACATTAG
ATTGCAAAAAAAAAAAAGGATTAATAAATACATCCATATTGTAACTAGTTCGCTTGGTTATAGATATACAC
ACGCACGCACACAAATCCTTAACTTTACATCCACCTATCTTAGGCGCTTTACATCTAACGCGCTGGATCA
TCGTGGATCTGGCCATCATACCTCATCTACAATCCATCTAGAAATCATCCAGCTCCAGCTACTTCTATAC
AACGTTCGCCGTGGAAAGAACATTCTTTAGCCTCGCACTAAACTAAACTTAAATAATAATCAACAACGTT
GCTCCAACAGACTTAAAATGGAATTAGAACAAAAATCATGTGTGCTTAAGTCGTGTGCAGAATGTCTCAA
TATTTACACATCAAAGTTCACTTTTGCCAATTGCCCTCTAGCTCCTCTAGCTCCTCATCATCTATCTCTC
TTCATCTGACTCTACTCTACGATCTGCGGCTCTTCGGATTCGAAGGACACATCGAAAATCGAACAGTGAT
CGGTGGGGATGTTTAGCTCTCTCGATTTCGCAGTCCACAAAATCCGATTCTCTCCACCACCAGCCGCCCT
TGTCATTCATTCCACATTCATCAGTTTAGTCCTTCGTTTATTTTTTTTTTTTTAACTTTTTGGCACTGCT
TTGAATTTCTTTTGGCACCGCACGCTTCGAGCGTGAAAATAAAATTATTTATGCATATACATAGATAAAT
ACGAATATATCTATATGGAGTATGTAAAAAAAAAAAAAGGATCC
>gnl|uv|M13369.1:1-85 Plasmid pNNL (cosmid) DNA with an inserted IS1
fragment
GAATTCTCATGTTTGACAGCTTATCATCGATAAGCTGGTGATGCTGCCAACTTACTGATTTAGTGTATGA
TGGTGTTTTTGAGGT
>gnl|uv|M13369.1:757-815 Plasmid pNNL (cosmid) DNA with an inserted IS1
fragment
GGGCATTATCTGAACATAAAACACTATCAATAAGTTGGAGTCATTACCACGATAAGCTT