Spring 5-2010

A Parameterization Study of Short Read Assembly Using the Velvet Assembler

Alex Christopher Elliot
University of Southern Mississippi

The University of Southern Mississippi

A PARAMETERIZATION STUDY OF SHORT READ ASSEMBLY USING THE VELVET ASSEMBLER

by

Alex Christopher Elliot

A Thesis
Submitted to the Graduate School
of The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Master of Science

Approved:

Dean of the Graduate School

May 2010

ABSTRACT

A PARAMETERIZATION STUDY OF SHORT READ ASSEMBLY

USING THE VELVET ASSEMBLER

by Alex Christopher Elliot

May 2010

In this study, we examine approaches to the problem of assembling large, contiguous sections of genetic code from short reads generated from laboratory techniques. We explore the Eulerian Path approach in detail, utilizing a de Bruijn Graph, and demonstrate current software technologies and algorithms using a sample genome. We investigate the input parameters of Velvet and discuss choice implications in the context of the E. coli putA/b1014 gene.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figure

LIST OF TABLES

CHAPTER I

BACKGROUND INFORMATION

In order to fully understand the challenges and scope of read assembly, we must first discuss the biological and chemical background to the problem. All known living organisms and some viruses encode genetic information in the form of Deoxyribonucleic Acid (DNA) (Darnell, Lodish and Baltimore 66-74). DNA is a double helix shaped polymer consisting of two long, spiraling chains of alternating sugars and phosphate groups bonded to a series of nucleotides. These nucleotides, or bases, can be one of either adenine (A) thymine (T) guanine (G) or cytosine (C). Each nucleotide forms a bond with exactly one other base, known as its compliment, and the order of these bases houses the genetic information.

Through the processes of transcription and translation, an organism uses its genetic blueprint to first create ribonucleic acid (RNA). RNA is similar to DNA except that it contains uracil (U) in the place of thymine (T). Once a segment of DNA is translated to RNA, ordered amino acid chains are generated based on the RNA's nucleotide sequence. These amino acids are then assembled into proteins, which can range in size from tens to tens of thousands of amino acids each.

Each amino acid is encoded by three bases, and there are twenty possible amino acids in use by known organisms (Darnell, Lodish and Baltimore 88). Some areas of DNA also contain information to control gene regulation and expression as well as sections of "filler" data known as introns. With four

possible bases, there are 4^3, or 64 base combinations. Some amino acids have multiple coding triplets, and the information can be encoded in the forward or reverse direction. Furthermore, varying the start base can shift the read frame and thus the resulting codons, amino acids and proteins.

Eukaryotic cells such as those found in mammals and plants, encapsulate their DNA within a cell nucleus, often in chromosomes or tightly wound structures of DNA. Prokaryotic cells such as bacteria and archea without a true, membrane bound nucleus, store their genetic material in cytosol or intracellular fluid. The genetic material found in prokaryotes is normally arranged into smaller circular chromosomes. The collection of all genetic material contained in an organism is known as that organism's genome.

Since the discovery of the DNA double helix in 1953 (Watson and Crick, A Structure for Deoxyribose Nucleic Acid), science has sought to fully understand the information contained within it (Watson and Crick, Genetical Implications of the structure of Deoxyribonucleic Acid). In a macro view, understanding an organism's genome can help reveal its phylogeny and origins, while the micro view can uncover information about disease susceptibility and cure. Small sections of an individual organism's genetic fingerprint that indicate the presence or absence of a particular trait are called genetic biomarkers. These biomarkers can be used to, for example, determine relation between organisms, gauge exposure to a particular genetic toxicant, predict inherited disease, or determine an optimal treatment approach.

In order to understand genetic information, one must find a way to read

the information contained within DNA or RNA. Genetic sequencing techniques were first developed in the early 1970's (Gilbert and Maxam). These complex methods, including the wandering-spot technique were very labor intensive. Fredrick Sanger (Sanger, Nicklen and Coulson) and Gilbert (Maxam and Gilbert) independently published research in 1977 that greatly simplified the sequencing process. Both were awarded the Nobel Prize in 1980 for their work.

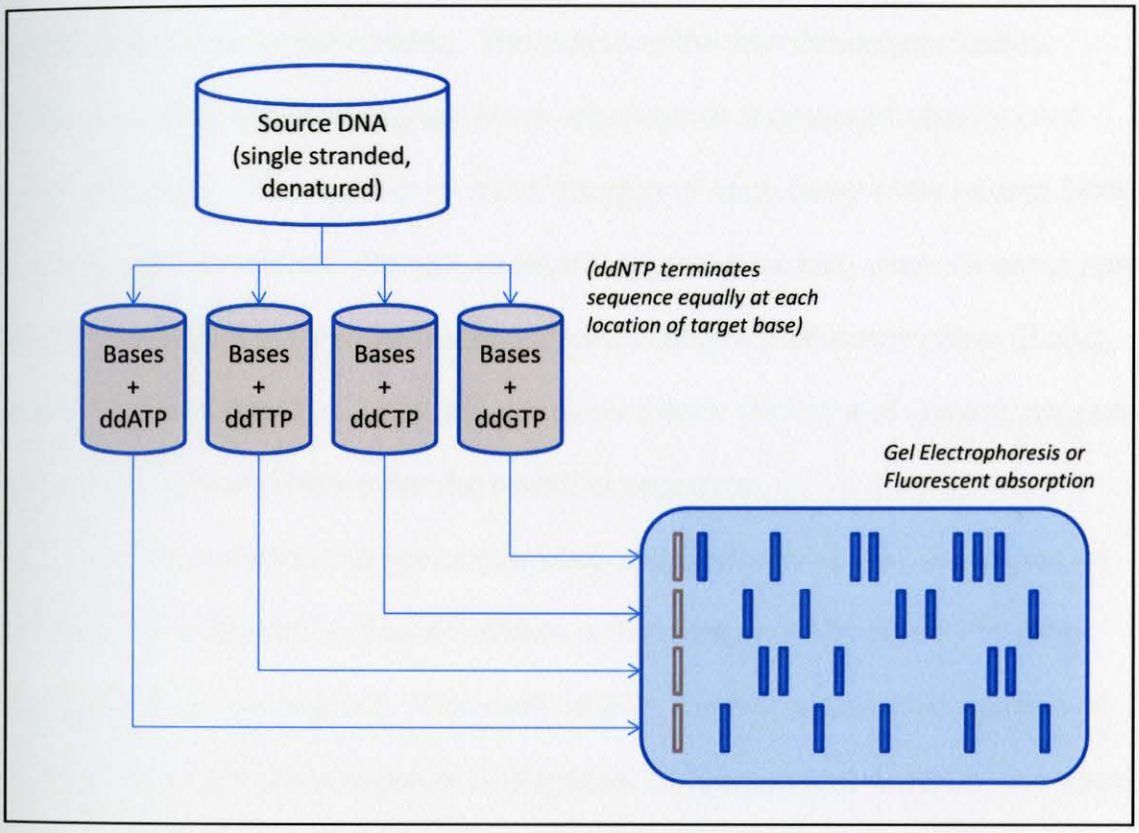

Figure 1: Sanger method overview. The Sanger method uses ddNTPs to terminate cloned genetic source material at each base location. The resultant strands are then separated by length via Gel Electrophoresis or Fluorescent Absorption and recorded. The Sanger method usually results in few, long (~1kbp) reads (Applied Biosystems).

The Sanger method (Fig. 1) is a chain terminating technique that uses of dideoxynucleotide triphosphates (ddNTPs) to selectively terminate long strands of genetic material (Tamarin 334-6). In this method, single stranded, denatured DNA source material is cloned and separated into four separate solutions containing one of ddATP, ddTTP, ddCTP, or ddGTP each. The dideoxynucleotides terminate the multiple copies of the DNA strand at each location of the target base, resulting in strands that begin at the origin and have length of the base location index. The output of the four dideoxynucleotide solutions is then separated by gel electrophoresis or fluorescent absorption if dyes were used. The result is an index location of each base in the source DNA to a one-base resolution. Sanger sequencing generates long reads of about one thousand bases, but requires weeks to months of costly laboratory time (Ewing, et al.). This technique is susceptible to cloning error (Ewing and Green), as parts of the cloning vector may enter the resulting sequence.

An alternative to Sanger sequencing, pyrosequencing was developed by Nyrén and Ronaghi at the Royal Institute of Technology in Stockholm in 1998 (Ronaghi, Uhlén and Nyrén). This method (Fig. 2) involves iterative addition of bases in an enzymatic solution of Sulfurylase, Luciferase and Apryase. As each base bonds to the source material, a measurable amount of light is released per base. Repeat bases yield proportionately more light. After each base is introduced and bound, an enzyme is added to remove all unused bases before the next base is added.

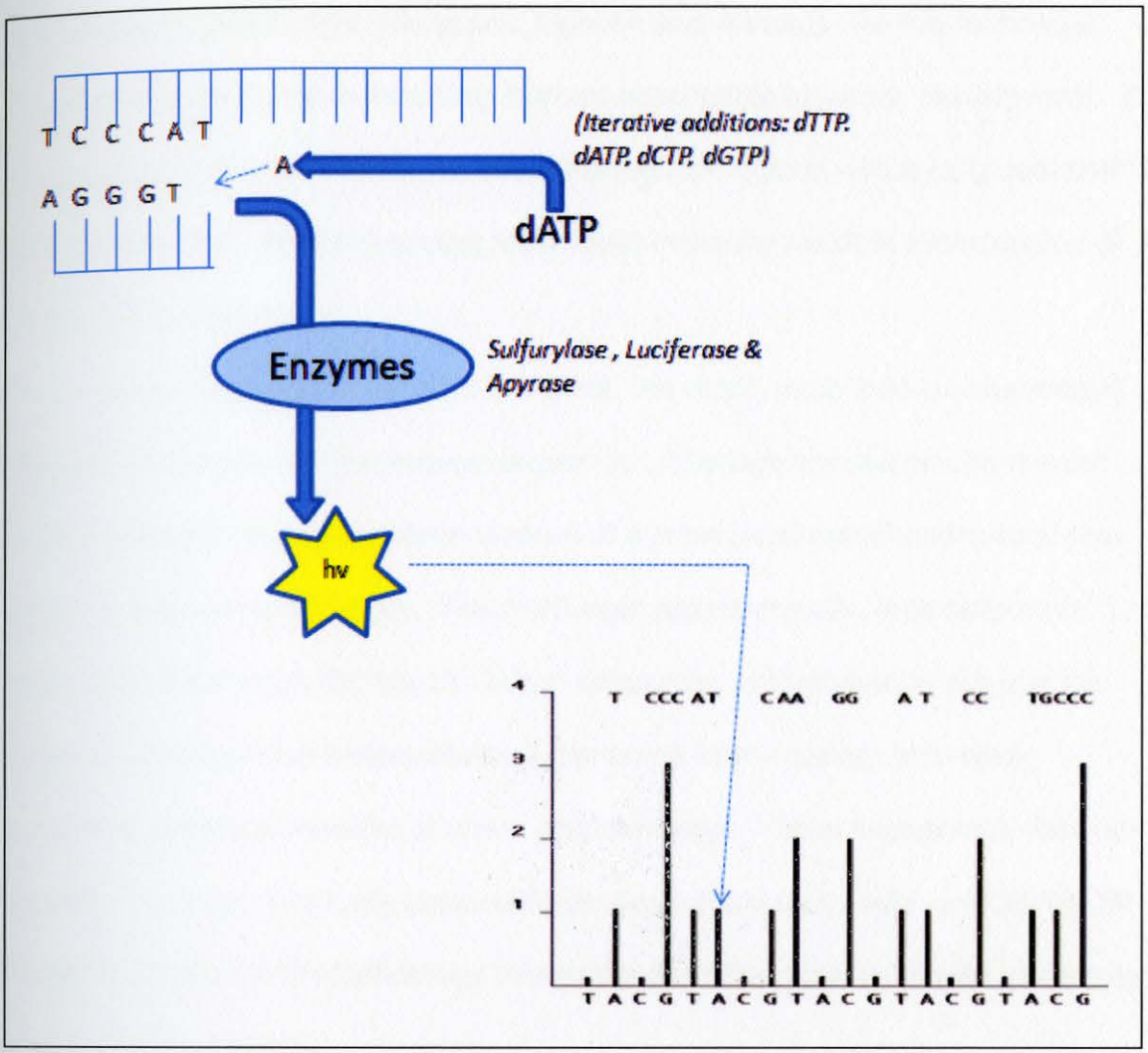

Figure 2: Pyrosequencing overview. Pyrosequencing introduces bases to source material iteratively. If a base binds to the source material, a measurable unit of light (photon) is released. The intensity of this light is quantifiable and relates to number of sequential bases encountered. Pyrosequencing quickly (~25Mbp/4hr) generates many; short (100-400bp) reads (Marguiles, Egholm and Altman).

Pyrosequencing results in short length reads with an upper limit of approximately 500 bases; however commercial implementations are constantly increasing the maximum read length. Pyrosequencing is also less expensive to perform than traditional techniques, with companies such as 454 Life Sciences producing all-in-one units (Roche Diagnostics Co.). This technique can produce

approximately 25Mbp/4hr (Marguiles, Egholm and Altman). As this technique does not require traditional cloning, it is not susceptible to vector cloning error. It is, however, potentially less accurate in homopolar regions with a long series of repeating bases. Pyrosequencing techniques normally result in many copies of overlapping short reads.

After the laboratory work is complete, the reads must then be assembled into a representation of the source sequence. Although various solutions exist for this problem, all require some amount of *a priori* assumption and reliance on yet to be fully verified metrics. The challenge, algorithmically, is to determine how each of the reads fits into the larger sequence. Information to support the selection amongst candidate solutions can come from existing, reference genomes, statistical models, or sheer read coverage. Once sequenced, data can be added to large, publically accessible genome databases such as NCBI (NCBI [National Center for Biotechnology Information]) or GenomeNet (Kyoto University Bioinformatics Center).

The NCBI Basic Local Alignment Search Tool (BLAST) can be used to find regions of local similarity between sequences. The program (Altschul, Madden and Schaffer) compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

CHAPTER II

THE EULERIAN APPROACH

In this section, we describe the application of the Eulerian path to short read assembly and its differences as compared to earlier methods. We will discuss one available implementation – Velvet, and provide insight into its algorithm.

Older approaches to the problem of read assembly were designed around the assembly of few, long reads. Many available programs utilized the "overlap-layout-consensus" paradigm which tests each possible read pair combination to determine the best matches. Each read is represented as a node, and each detected overlap is drawn as an arc between the overlapping nodes. Once matches are scored, the assembly is generated based on overlap scoring. Unfortunately, determining the layout leads to the NP-complete Hamiltonian Path Problem (Cormen, Leiserson and Rivest). The difficulty of the Hamiltonian Path Problem is exacerbated when attempting to operate on an increased number of reads.

Pevzner proposed an alternative solution to the read assembly problem for sequencing by hybridization (P. A. Pevzner). By making use of the de Bruijn Graph, he reduced read assembly to a solvable Eulerian Path Problem. Further work by Idury and Waterman (Idury and Waterman) applied the Eulerian path to short fragment assembly by treating short fragment assembly as a Sequencing by Hybridization problem. Pevzner, Tang and Waterman refined their Eulerian graph techniques in 2001 to include methods of error correction and repeat

handling in data (Pevzner, Tang and Waterman).

An n-dimensional de Bruin graph contains vertices representing all sequence permutations of length n over a given alphabet with repeats (de Bruijn). For gene sequencing, our alphabet consists of a relatively few symbols - A,T/U, G, C. Nodes are adjacent whenever they represent a one character shift left with any of the alphabet symbols shifted into the last location. Such a graph covers all possibilities. In read assembly algorithms, the de Bruin graph is built from the bottom up, adding vertices only when that permutation is found within the genetic information itself. This restricts the size of the graph, and also allows it to represent all possible overlap combinations for the underlying data.

Zerbino and Birney released a set of algorithms called "Velvet" (Zerbino and Birney) to manipulate de Bruijn graphs for genomic sequence assembly. In their implementation of the graph (Fig. 3), a *k*-mer is defined as a substring of length *k*, extracted from a read. Each node contains a series of overlapping k-mers, with each overlap having length *k*-1 bases. Each node is attached to another, "mirror" node which contains the reverse series of *k*-mers. These mirror nodes take into account the complementary nature of genetic material. Nodes whose last *k*-mer overlaps with the first *k*-mer of another node are connected by a directed arc. The assembled contiguous sequence or "contig" is represented by a traversal from the first *k*-mer of the first node through connected arcs to each other node.

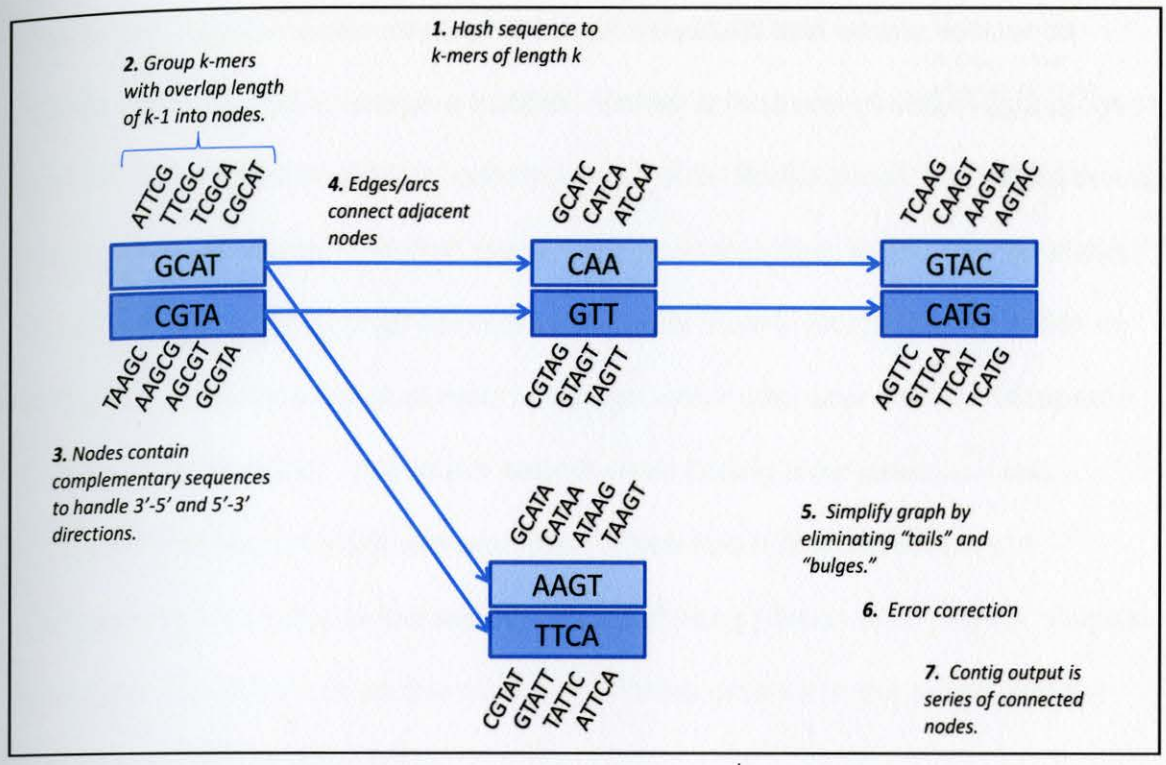

Figure 3: Steps in the Eulerian Approach. 1. Hash source sequence into k-mers of length k. 2. Group k-mers with overlap length of k-1 into nodes. 3. Nodes contain complimentary/mirror sequences. 4. Directed edges/arcs are formed to connect adjacent nodes where k-1 overlap is found. 5. Graph is simplified by condensing adjacent, single pathway nodes. "Tips" and "Bulges" are removed. 6. Error Correction. 7. Contig output sequence is path through series of connected nodes.

Once the input reads have been hashed into *k*-mers and assembled into nodes and arcs, the resulting graph must be simplified and cleared of errors. Velvet simplifies the graph by combining adjacent nodes with only one incoming and outgoing arc. This reduces the node count to only nodes with multiple arcs. Error correction is performed by eliminating "tips" and "bulges." A "tip" is defined as a chain of nodes connected at only one end, and Velvet removes tips that do not meet minimum length and coverage requirements. A "bulge" is a redundant path that starts and ends at the same nodes as other paths with similar

sequences. Velvet again employs a length threshold and simple sequence identity to condense or merge a bubble. Velvet is thus composed of four stages: hashing the reads into *k*-mers, constructing the de Bruijn graph, correcting errors, and resolving repeats. The first stage, graph construction, is memory intensive. The time complexity of error correction depends mainly on number of nodes in the graph, which is a result of read coverage, error rate, and number of repeats in the source material. The graph search used during error detection and correction employs the Dijkstra algorithm which has a time complexity of O(NlogN) when implemented with a Fibonacci heap (Gross and Yellen). Repeat resolution also depends on the number of nodes present in the graph and the average length of those nodes.

CHAPTER III

METHODS

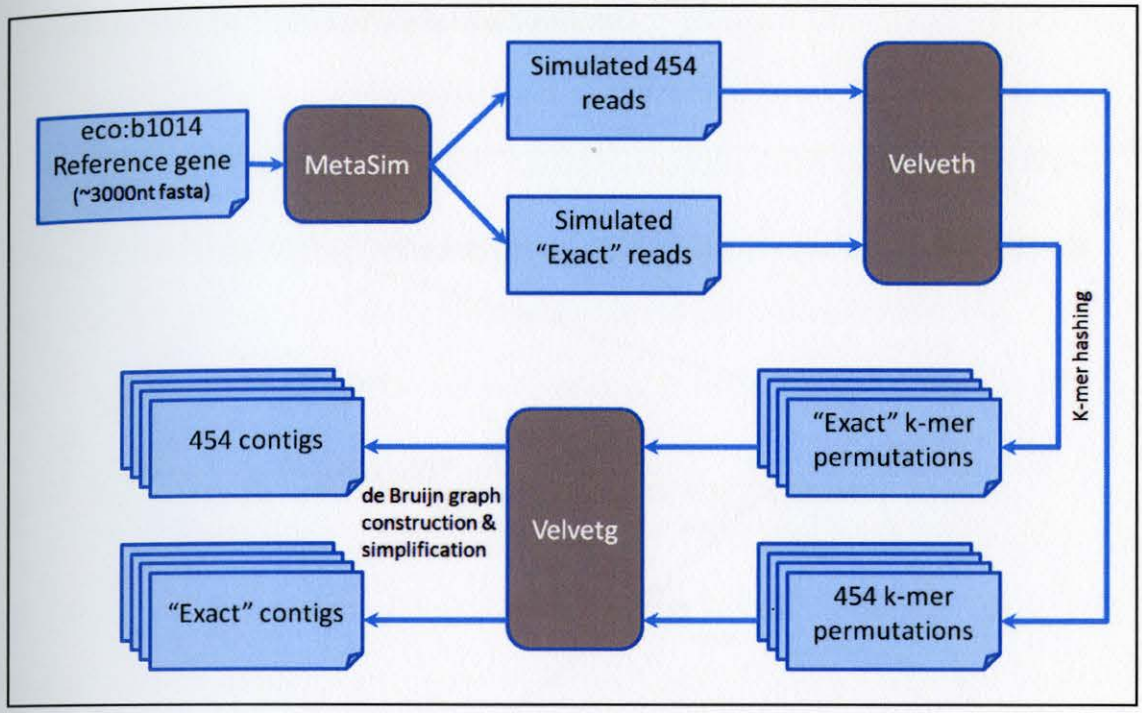

Figure 4: Assembly Programmatic Flow. Metasim generated simulated reads from source gene. These simulated read sets were hashed using velveth with varying k-mer lengths then assembled into de Bruijn graphs using velvetg.

To illustrate the operation of Velvet in conjunction with a series of other utilities (Fig.4), we chose a specific, active coding gene of Escherichia coli str. K-12 substr. MG1655 (Fig. 5). This gene, NP_415534, codes for the enzyme proline dehydrogenase/pyrroline-5-carboxylate dehydrogenase which functions as a fused DNA-binding transcriptional regulator (Kyoto University Bioinformatics Center). The E. coli genome has been extensively studied and fully sequenced (Blattner, Plunkett and Bloch) allowing for comparison of our assembly results with established sequence data. The NP_415534 gene sequence was obtained

from GenomeNet (Kyoto University Bioinformatics Center) in its full form as an ASCII formatted fasta file (NCBI [Fasta]) (Appendix A). This reference gene contains a total of 3963 ordered nucleotides.

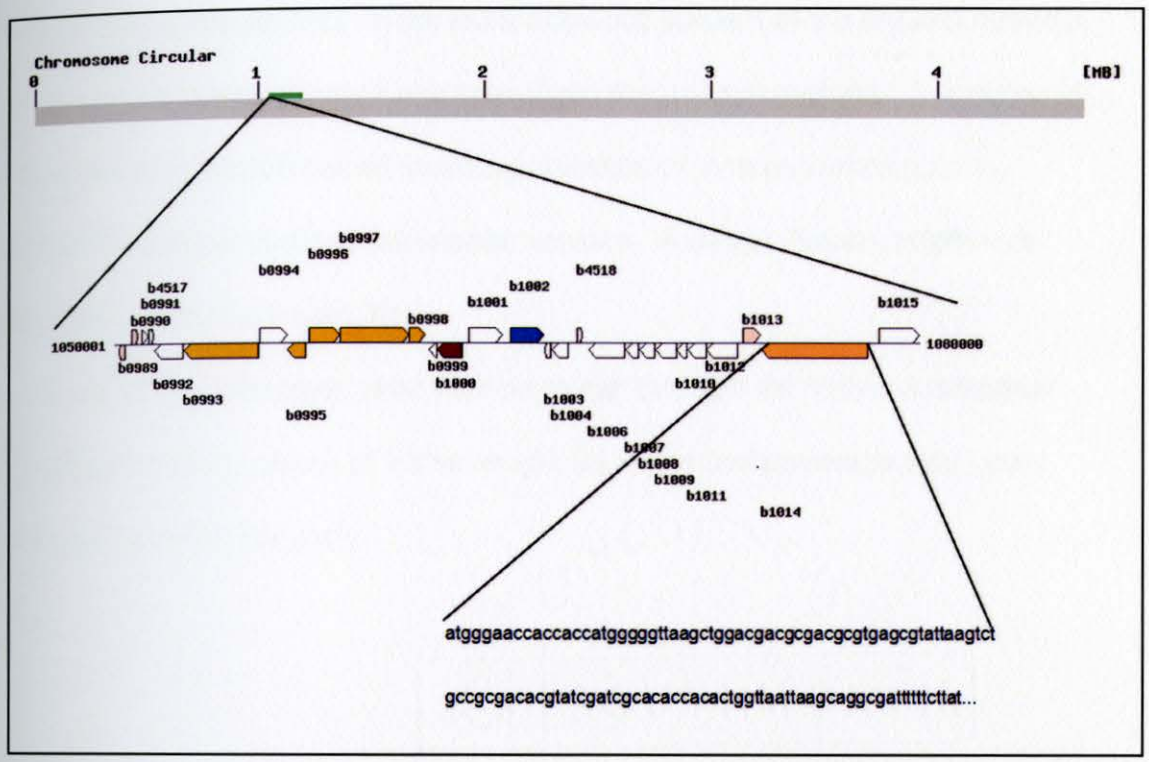

Figure 5: Gene NP_415534 Physical Map. Gene NP_415534, putA was selected from the circular E. Coli chromosome to illustrate the assembly algorithms. This figure shows its physical location with reference to neighboring genes. (Kyoto University Bioinformatics Center)

From the reference file, we used the read simulation function of MetaSim (Richter, et al.) to output two sets of simulated reads (Fig.4). The first set represents an "exact" or reference set (Appendix E) in which, 5000 reads were taken directly from the source gene without introduced error. The output reads have a normal distribution across the source gene and an average read length of

997.87 base pairs (Appendix E).

To illustrate real world data, we also generated a set of reads modeling the read output of the LifeSciences 454 sequencer (Roche Diagnostics Co.). These 5000 simulated reads contained 29890 insertions and 7321 deletions. Each insertion is the addition of an extra base not present in the original material. Each deletion is the removal of one base from the original material. Locations of these induced errors are based on characteristics of pyrosequencing such difficulties accurately reading homopolar regions. Average Read Length was 258.21 base pairs (Appendix F).

Each of the simulated read sets were run through the Velvet Assembler (Fig.4) using varying values of k-mer length (k), expected coverage (exp_cov) and coverage cutoff (cv_cut).

Table 1: Assembly Parameter Permutations.

kmer	cvCut	exp	ctgs	asmLg	N50	mean	1k	max	tiles	rdPc
21	2	4	4253	114471	26	26	0	56	1394	27.88
21	3	6	2603	70464	26	27	0	44	2236	44.72
21	4	8	1484	40732	27	27	0	70	2638	52.76
21	6	12	107	3896	36	36	0	144	3375	67.50
21	10	20	60	4007	91	66	0	381	4997	99.94
21	12	24	56	4366	110	77	0	411	5000	100.00
23	2	4	4337	128149	29	29	0	60	1378	27.56
23	3	6	2638	78246	28	29	0	66	2914	58.28
23	4	8	692	20845	29	30	0	57	1438	28.76
23	6	12	212	7328	35	34	0	119	3065	61.30
23	10	20	65	4453	99	68	0	311	4999	99.98
23	12	24	46	4099	133	89	0	344	4998	99.96
27	2	4	3989	137723	33	34	0	82	1655	33.10
27	3	6	2418	83130	32	34	0	62	1588	31.76
27	4	8	1067	37296	34	34	0	81	2490	49.80
27	6	12	233	9410	41	40	0	130	3637	72.74
27	10	20	47	4250	136	90	0	559	4999	99.98
27	12	24	44	4331	145	98	0	594	5000	100.00
31	2	4	3804	150919	38	39	0	81	1118	22.36
31	3	6	1251	49497	37	39	0	73	596	11.92
31	4	8	383	15473	39	40	0	88	1053	21.06
31	6	12	134	7062	54	52	0	195	4228	84.56
31	10	20	22	3605	266	163	0	497	4975	99.50
31	12	24	22	3790	273	172	0	491	5000	100.00

This table shows the parameter permutations used and their results for the simulated 454 reads. "kmer" is the selection of k or kmer length. "cvCut" is coverage cutoff, a threshold used to determine if a node in the constructed de Bruijn graph should be included as part of the final assembly. "exp," expected coverage, is the expected frequency of repeats of each source base. "ctgs" is the number of contigs. "asmLg," "mean," and "max" refer to the total length, mean, and maximum length of all assembled contigs respectively. "N50" refers to the length of the shortest contig in an assembly such that the sum of contigs of equal length or longer is at least 50% of the total length of all contigs. "1k" is the number of contigs over 1000 bases long. "tiles" is the number of reads that are used in an assembly. "rdPc" is percentage of input reads used in the assembly.

Automation of parameter variation (Table 1) and report generation (Fig. 6-9) was assisted by the standardized velvet assembly report script project (leipzig). Expected coverage is the expected frequency of repeats of each source base. This is a function of the source material and the depth at which the sequencing was performed. Coverage cutoff is the value used to determine if a

node in the constructed de Bruijn graph should be removed from the final assembly. Lower frequency nodes with coverage below the coverage cutoff value are suspected to be erroneous and are subsequently removed during graph error correction, especially during tip and bulge removal. This threshold specifies how many read k-mers must overlap for each contig k-mer. The number of k-mers per read is a function of read length L and k-mer length K (L-K+1) (leipzig).

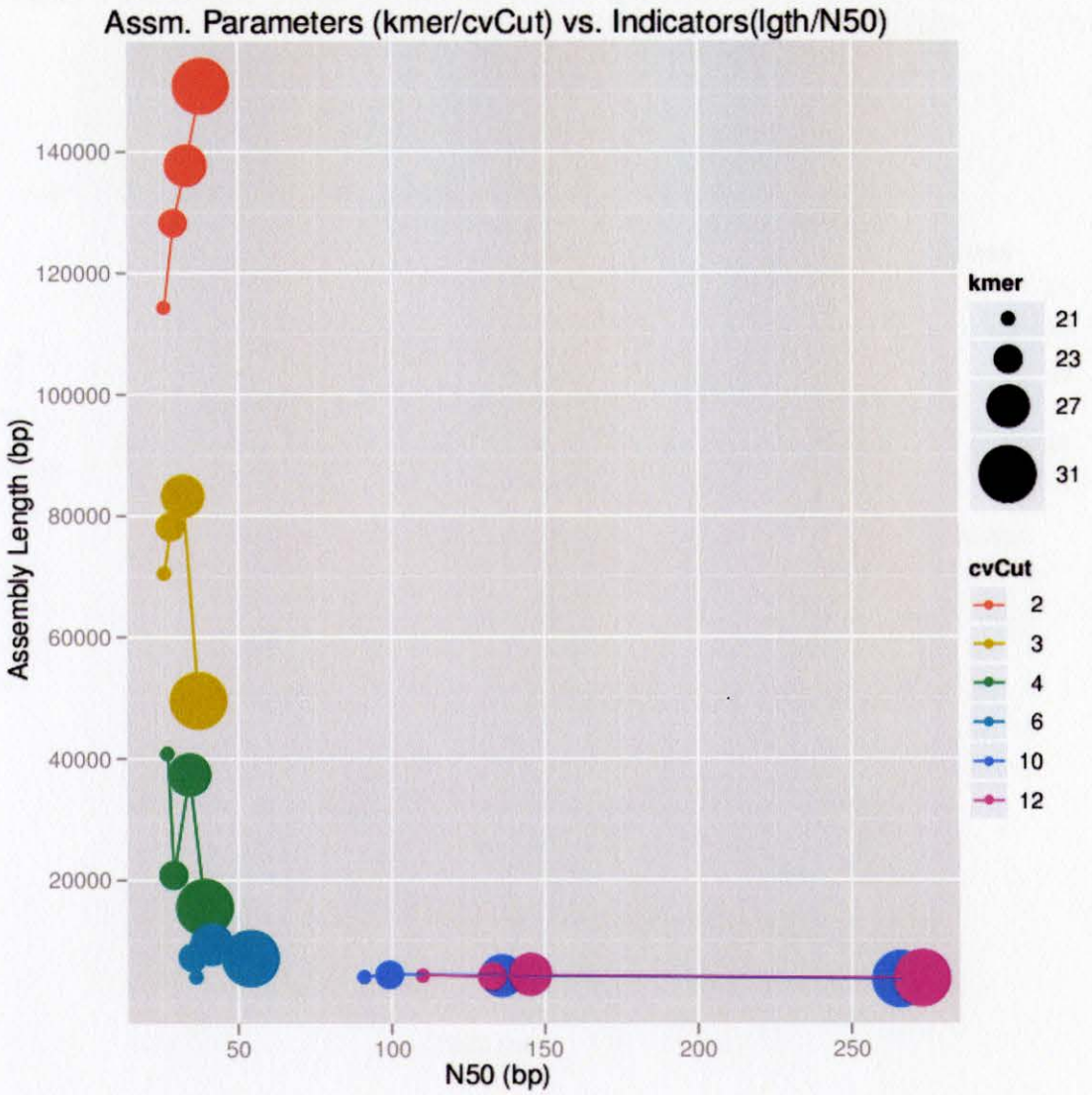

Figure 6: Assembly Parameters (kmer/cvCut) vs. Indicators (lgth/N50). This scatter plot illustrates the effect of kmer length and coverage cutoff on N50 and assembly length. N50 refers to the length of the shortest contig in an assembly such that the sum of contigs of equal length or longer is at least 50% of the total length of all contigs. Here we see a logarithmic distribution where higher cvCut values generate larger contigs.

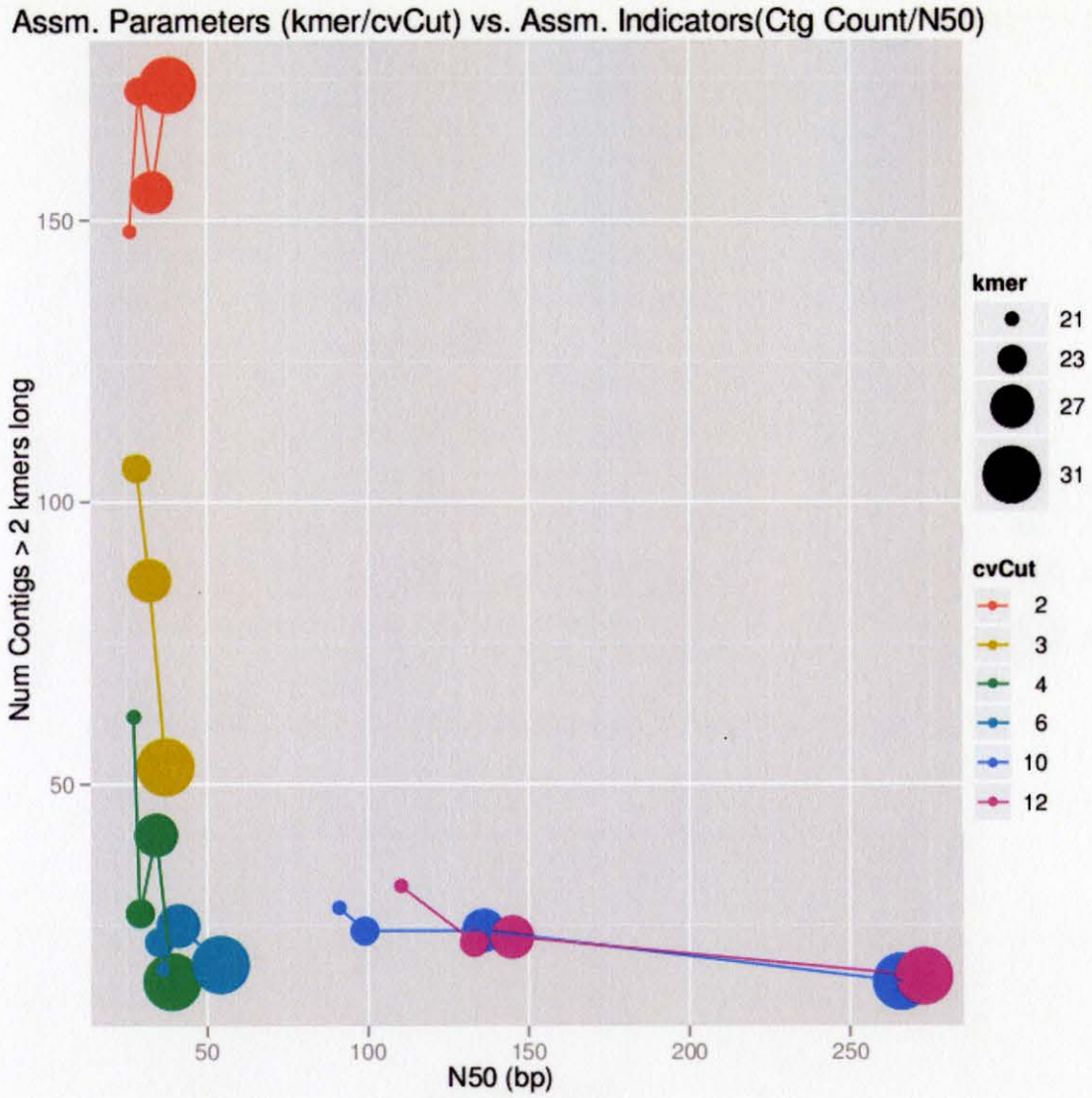

Figure 7: Assembly Parameters (kmer/cvCut) vs. Assembly Indicators (Ctg Count/N50). This plot shows the influence of kmer length and cvCut on the number of contigs produced with greater than 2*k length. Again, we see a somewhat logarithmic function with higher cvCut and higher kmers producer longer and fewer isolated contigs.

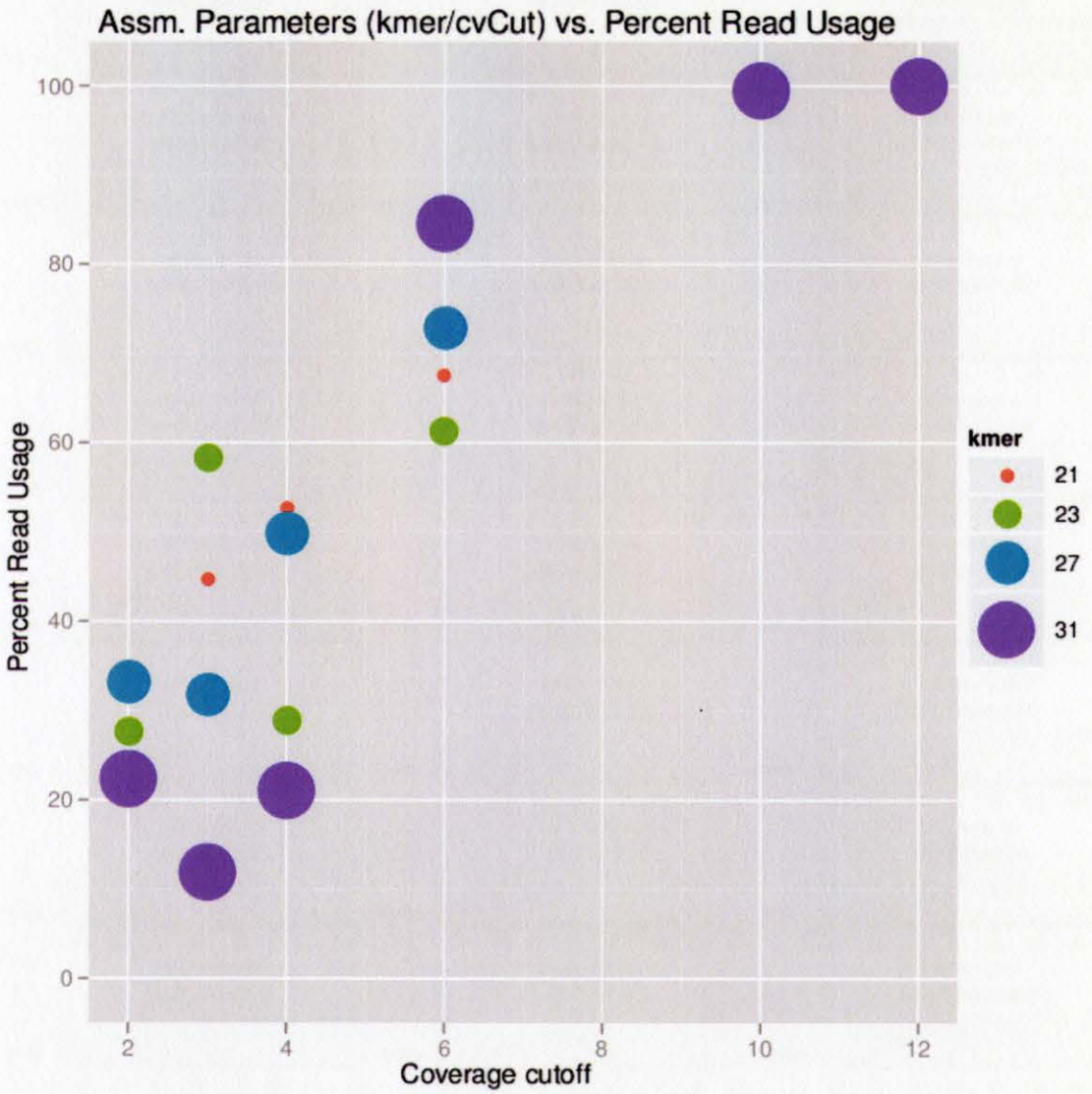

Figure 8: Assembly Parameters (k-mer/cvCut) vs. Percent Read Usage. This figure compares k-mer and cvCut to the percent read usage. With sufficiently high k, read utilization increases with coverage cutoff, due to the relaxation of selectivity of node removal.

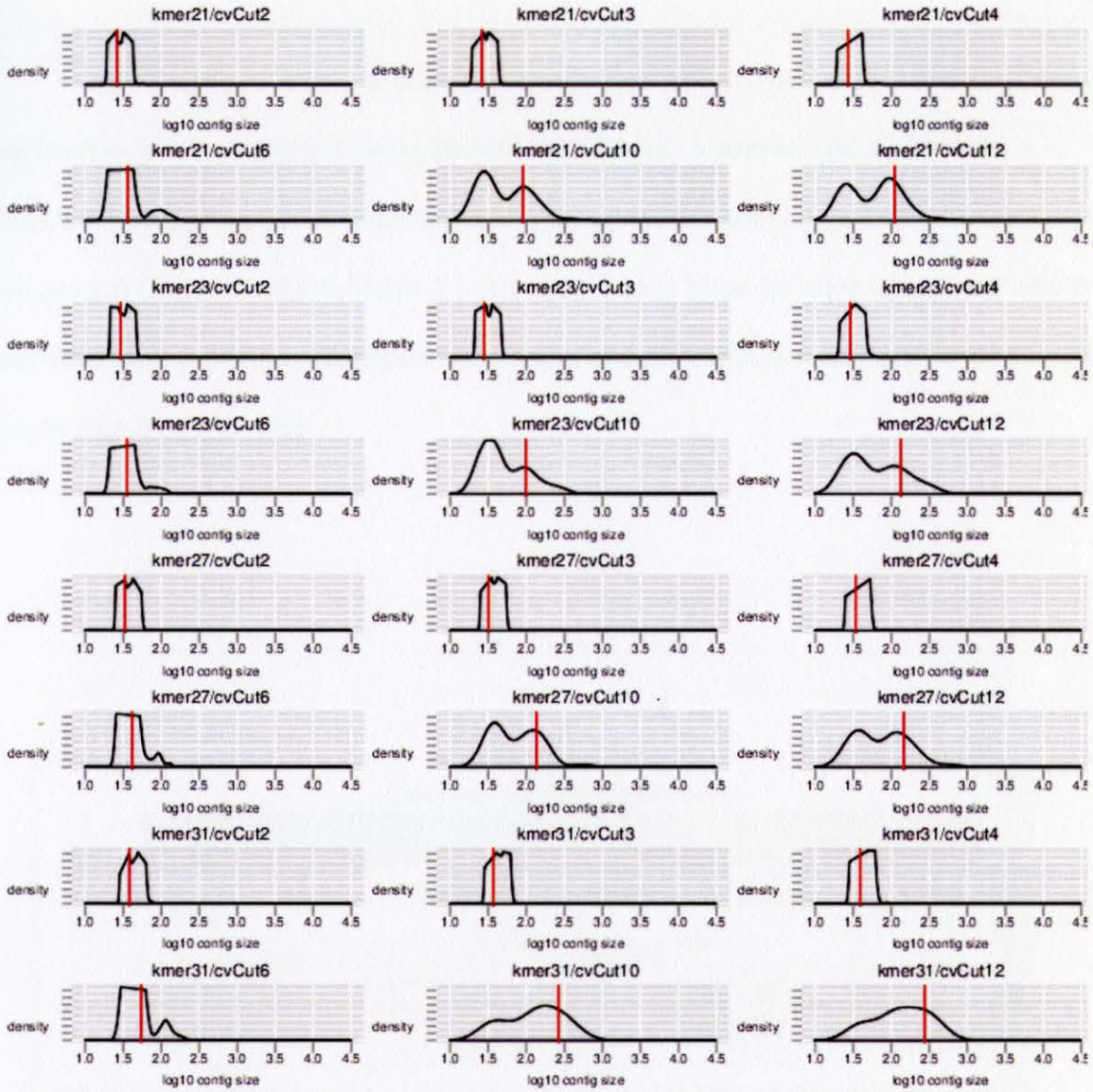

Figure 9: Kernel density plot of log contig size distribution by controlling for overall contig count. It is interesting to note the bimodal nature of this data. Future research into the cause of this may uncover underlying information about the source data or function of the assembler. The highest kmer and cvCut runs appear to be more normal in distribution.

The results of these assemblies were then compared back to the original, reference gene sequence using BLASTN (Altschul, Madden and Schaffer). BLASTN outputs a percent identity which shows the similarity to the reference sequence (Fig. 10 – 12 & Table 2 – 3) as well as a base by base alignment which shows direct matches, deletions, insertions and substitutions for each assembled node (Appendix B – E).

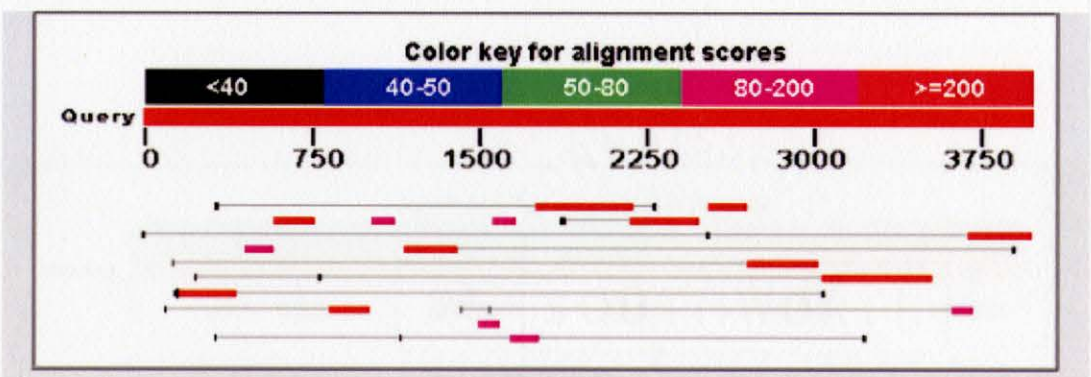

Figure 10: BLAST Map for Simulated 454 Reads (k = 31, coverage cutoff = 12, expected coverage = 24). This figure, from BLAST, maps the 16 resultant contigs of the 454 simulated reads at k-mer length 31, expected coverage 24, and coverage cutoff 12. The level of similarity to the reference gene is shown by the color, with red being the best quality. Contigs appear to high quality and well distributed about the gene.

Table 2: BLAST Contig Scoring for Simulated 454 Reads (k = 31, coverage cutoff = 12, expected coverage = 24)

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
63088	NODE_1_length_393_cov_219.491089	623	713	11%	0.0	100%
63089	NODE_2_length_282_cov_195.992905	484	511	8%	2e-140	96%
63090	NODE_3_length_267_cov_206.041199	428	479	7%	1e-123	100%
63091	NODE_4_length_206_cov_205.684464	340	361	6%	3e-97	100%
63092	NODE_5_length_158_cov_201.348099	293	293	4%	3e-83	95%
63093	NODE_6_length_291_cov_199.041245	524	570	8%	2e-152	100%
63094	NODE_7_length_461_cov_209.563995	690	732	13%	0.0	100%
63095	NODE_8_length_243_cov_213.312759	416	458	7%	6e-120	100%
63096	NODE_9_length_98_cov_212.948975	197	197	3%	1e-54	96%
63097	NODE_10_length_148_cov_229.418915	300	321	4%	2e-85	100%
63098	NODE_11_length_76_cov_221.750000	156	156	2%	3e-42	95%
63099	NODE_12_length_62_cov_208.145157	129	171	2%	4e-34	100%
63100	NODE_13_length_73_cov_221.095886	187	187	2%	2e-51	100%
63101	NODE_15_length_142_cov_198.697189	239	239	4%	6e-67	92%
63102	NODE_16_length_65_cov_216.415390	114	114	2%	8e-30	88%
63103	NODE_17_length_101_cov_213.732666	185	252	4%	8e-51	100%

This table, also from BLAST, shows the 16 resultant contigs of the 454 simulated reads at k-mer length 31, coverage cutoff 12 and expected coverage 24. The contigs averaged 97.63% accuracy.

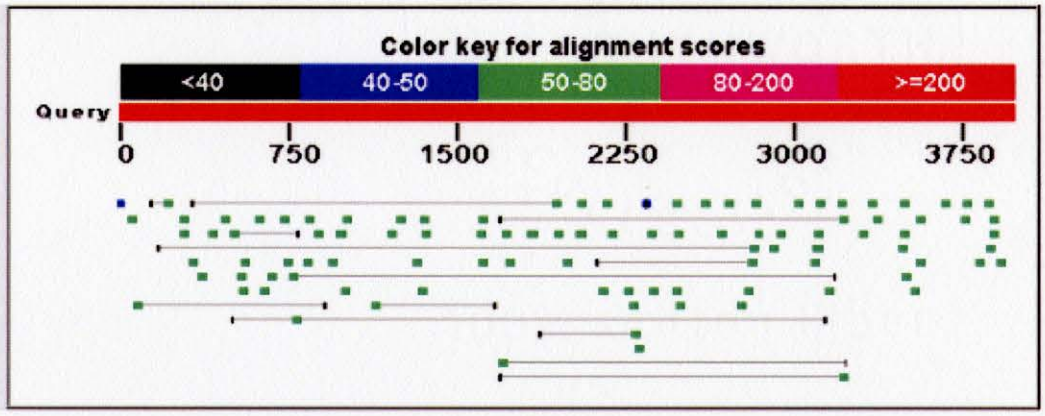

Figure 11: BLAST Contig Scoring for Simulated 454 Reads (k = 21, coverage cutoff = 2, expected coverage = 4). This assembly resulted in 4253 nodes and n50 of 16. BLAST, maps the resultant contigs above a scoring threshold. The level of similarity to the reference gene is shown by the color, with red being the best quality. Contigs appear to be very small with medium to poor quality, yet well distributed about the gene.

Table 3: BLAST Contig Scoring for Simulated 454 Reads (k = 21, coverage cutoff = 2, expected coverage = 4)

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max Ident	Links
52974	NODE_11_length_21_cov_7.142857	66.2	87.3	1%	1e-15	100%	
52975	NODE_159_length_21_cov_2.000000	59.0	59.0	0%	2e-13	95%	
52976	NODE_273_length_21_cov_3.952381	66.2	66.2	1%	1e-15	97%	
52977	NODE_893_length_21_cov_5.047619	66.2	92.7	1%	1e-15	97%	
52978	NODE_1046_length_21_cov_6.666667	66.2	66.2	1%	1e-15	97%	
52979	NODE_1086_length_23_cov_93.173912	53.6	53.6	1%	9e-12	93%	
52980	NODE_1139_length_21_cov_27.714285	46.4	46.4	0%	1e-09	96%	
52981	NODE_1229_length_21_cov_3.571429	66.2	66.2	1%	1e-15	97%	
52982	NODE_1278_length_21_cov_4.047619	66.2	66.2	1%	1e-15	97%	
52983	NODE_1407_length_21_cov_4.904762	66.2	66.2	1%	1e-15	97%	
52984	NODE_1540_length_36_cov_27.888889	42.8	83.7	0%	2e-08	100%	
52985	NODE_1819_length_25_cov_2.360000	59.0	81.9	1%	2e-13	93%	
52986	NODE_1834_length_21_cov_5.523809	66.2	66.2	1%	1e-15	97%	
52987	NODE_1931_length_21_cov_4.904762	66.2	92.7	1%	1e-15	100%	
52988	NODE_1961_length_21_cov_2.428571	66.2	66.2	1%	1e-15	97%	
52989	NODE_1994_length_21_cov_5.000000	66.2	87.3	1%	1e-15	100%	
52990	NODE_2005_length_21_cov_4.714286	66.2	66.2	1%	1e-15	97%	
52991	NODE_2082_length_21_cov_3.238095	66.2	66.2	1%	1e-15	97%	
52992	NODE_2191_length_21_cov_5.095238	66.2	66.2	1%	1e-15	97%	
52993	NODE_2199_length_21_cov_7.380952	66.2	66.2	1%	1e-15	97%	
52994	NODE_2212_length_21_cov_4.238095	66.2	66.2	1%	1e-15	97%	
52995	NODE_2217_length_21_cov_5.904762	66.2	66.2	1%	1e-15	97%	
52996	NODE_2236_length_21_cov_6.190476	66.2	66.2	1%	1e-15	97%	
52997	NODE_2238_length_21_cov_2.857143	66.2	66.2	1%	1e-15	97%	
52998	NODE_2257_length_21_cov_4.714286	66.2	66.2	1%	1e-15	97%	
52999	NODE_2258_length_21_cov_4.523809	66.2	66.2	1%	1e-15	97%	
53000	NODE_2275_length_21_cov_3.095238	66.2	66.2	1%	1e-15	97%	
53001	NODE_2294_length_21_cov_3.428571	66.2	66.2	1%	1e-15	97%	
53002	NODE_2316_length_21_cov_4.714286	66.2	66.2	1%	1e-15	97%	
53003	NODE_2332_length_21_cov_2.619048	66.2	66.2	1%	1e-15	97%	
53004	NODE_2338_length_21_cov_4.857143	66.2	66.2	1%	1e-15	97%	
53005	NODE_2351_length_21_cov_4.761905	66.2	66.2	1%	1e-15	97%	
53006	NODE_2355_length_21_cov_7.142857	66.2	66.2	1%	1e-15	97%	
53007	NODE_2369_length_21_cov_3.428571	66.2	66.2	1%	1e-15	97%	
53008	NODE_2370_length_21_cov_7.190476	66.2	66.2	1%	1e-15	97%	
53009	NODE_2371_length_21_cov_4.190476	66.2	66.2	1%	1e-15	97%	

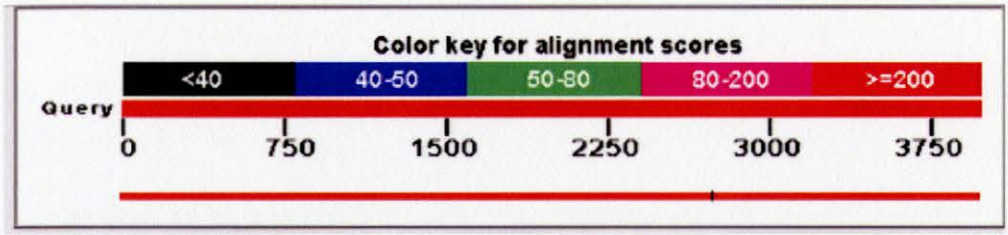

Figure 12: BLAST Contig Scoring for Simulated "Exact" Reads (k = 21, coverage cutoff = 2, expected coverage = 4). This assembly maps the single resultant contig of the simulated "exact" reads at k-mer length 21, expected coverage 4, and coverage cutoff 2. This assembly resulted in a complete gene sequence regardless of input parameters, indicating that the importance of coverage cutoff and expected coverage lies mainly in error handling and correction. When the input was of complete consensus with the reference sequence, the output remained error free. The assembled contig achieved 100% reference gene coverage at 100% identity with a length of 3963 bases.

AMOS files of selected final assemblies were generated with velvet and opened for analysis with Hawkeye (Fig. 13 – 15) (Schatz, Phillippy and Shneiderman).

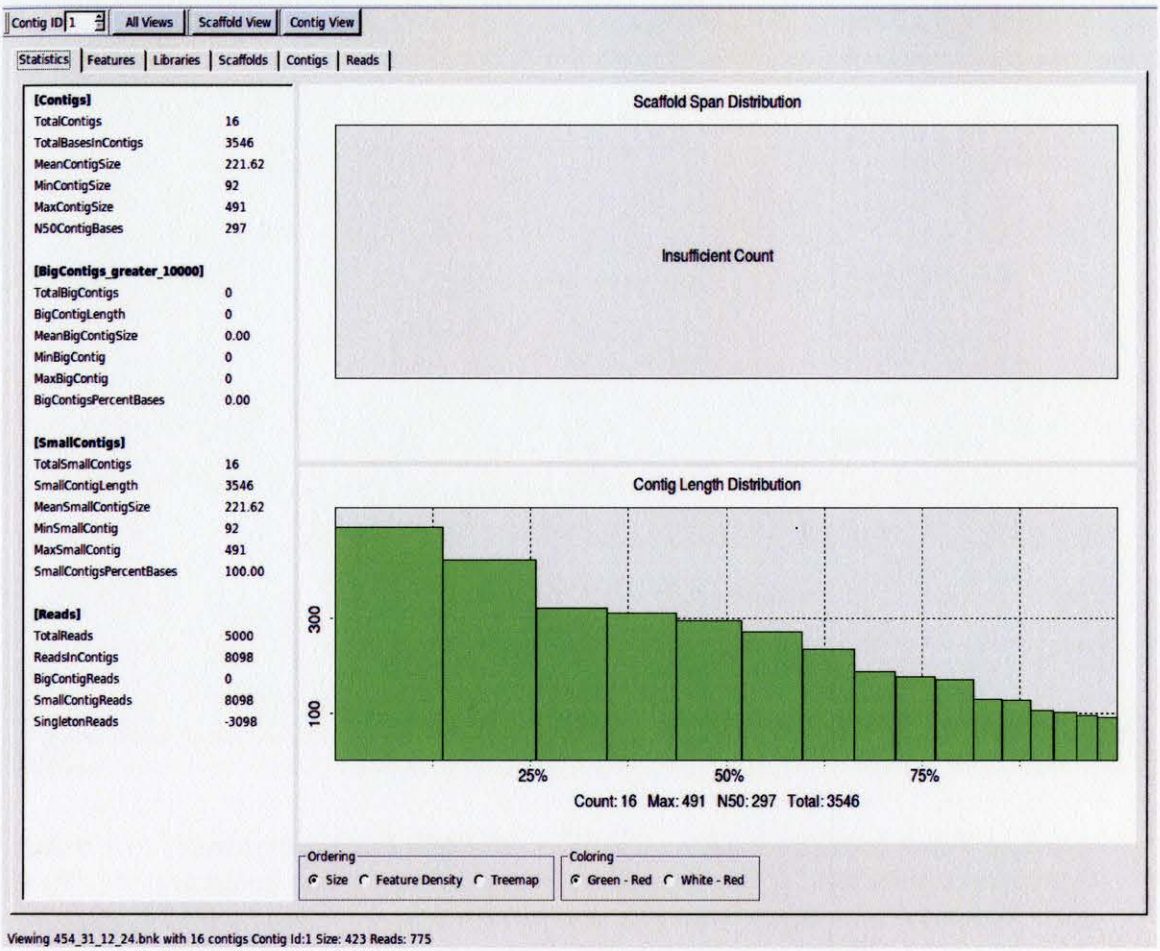

Figure 13: Hawkeye overview display. Simulated 454 simulated reads at k-mer length 31, expected coverage 24, and coverage cutoff 12. Various statistics are shown, including a graph of the contig length distribution.

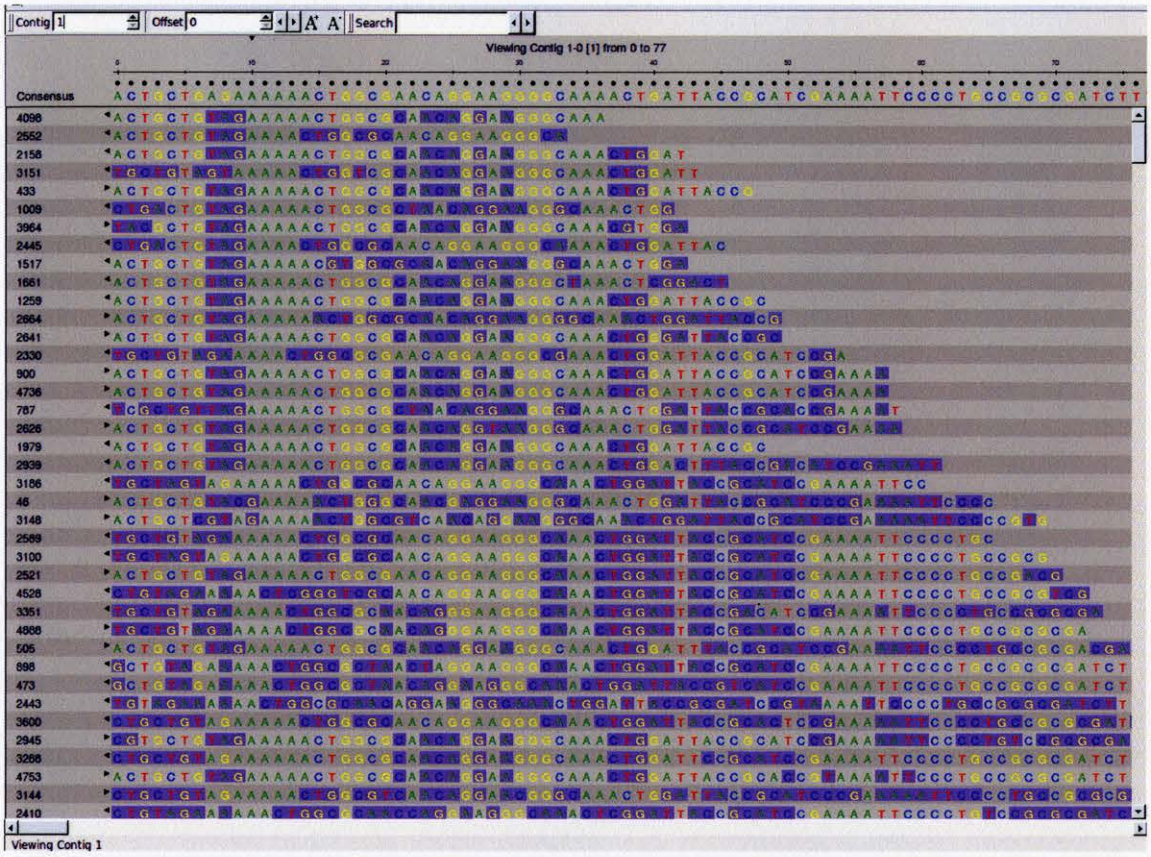

Figure 14: Hawkeye Contig Alignment Display. 454 simulated reads at k-mer length 31, expected coverage 24, and coverage cutoff 12 showing consensus alignments of the reads used to create one of the 16 generated contigs. This figure also shows the introduced errors in the simulated reads.

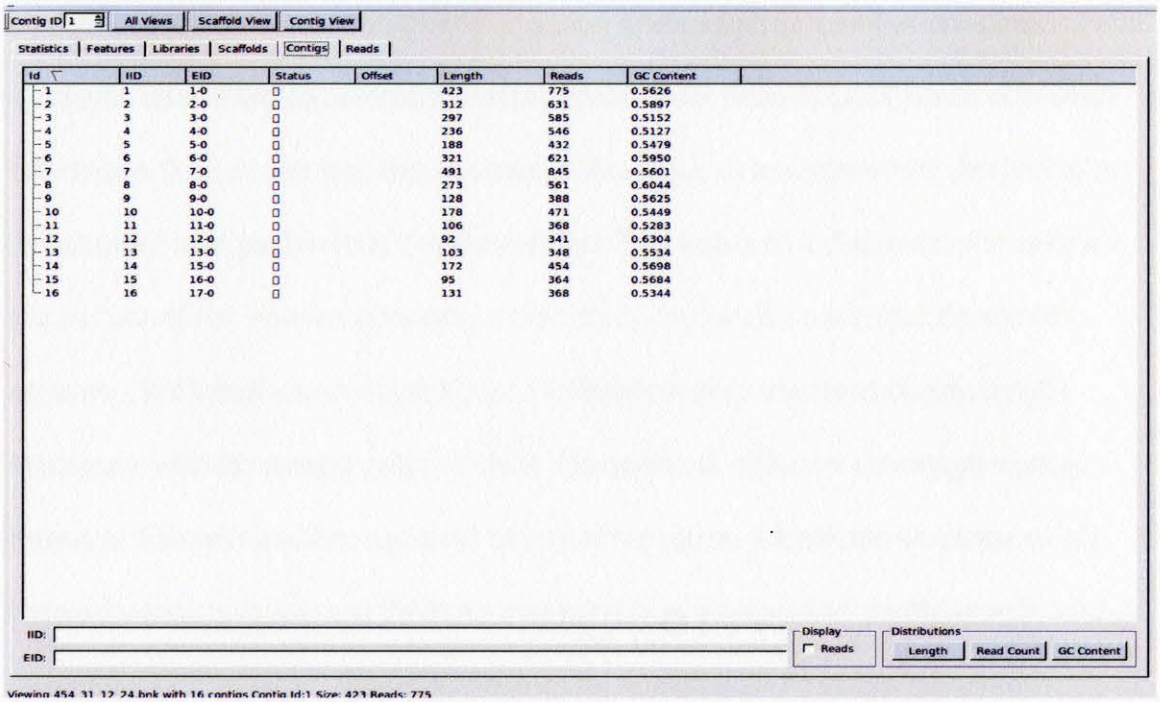

Figure 15: Hawkeye Contig Detail Display. 454 simulated reads at k-mer length 31, expected coverage 24, and coverage cutoff 12 showing information about each of the 16 generated contigs including length and number of reads used per contig.

CHAPTER IV

CONCLUSIONS

Read assembly remains an inexact science, relying heavily on statistical modeling and inference for error correction and graph simplification. Our synthetic assembly experiment demonstrates how heavily parameter selection influences final assembly, thus consideration must be made when designing an experiment and performing the assembly. The value of k depends primarily on the nature of the source genome, particularly the length and abundance of repeats. With sufficiently high k, read utilization and resultant contig length increases with coverage cutoff, due to the removal of lower coverage nodes, however this elimination can lead to misassemblies. A delicate balance exists between easing coverage limits to increase final assembled contig length and a reduction in accuracy. Some experiments, such as preliminary genome sequencing may seek wider coverage and fewer but longer nodes at the expense of 100% accuracy of individual bases, whereas small target sequencing of short gene segments may obtain the higher accuracy required by increasing read coverage.

Future Work

As the algorithms continue to mature, research into the automated choice of parameters will assist scientists when faced with the challenge of read assembly. Obtaining and integrating the various scripts and applications was a chore, as each had its own set of dependencies and special setup instructions.

Velvet assembly and the associated tools would benefit from a cloud implementation, similar to that of NCBI's BLAST to provide a full suite of assembly tools with minimal or no configuration.

Further efforts to understand the parameterization of short read assembly using Velvet should expand both the source data and selected parameter value set, possibly to include eukaryotic data. Application of a similar study to eukaryotic source data could aid in understanding how the various stages of De Novo assembly are influenced by the fundamental differences in complexity and arrangement of more complex genomes. Understanding the effects that source data repeat rates and complexity have on assembly contig outputs may lead to a better understanding or expansion of the use boundaries of this algorithm.

A more detailed study of k-mer length selection could also include recursive scanning of a reference genome for maximum repeat length and *a priori* comparison to the genomes of similar organisms.

Continued effort to understand and evaluate the decisions used when simplifying or error correcting the de Bruijn graph will lead to higher quality assemblies, eventually reducing our dependence on De Novo techniques. Improved and standardized methods can server to unify the field in the areas of statistical decision making as well as reference to biological markers and archived genomic data.

APPENDIX A

FASTA FILE FOR REFERENCE GENE PUTA/B1014

>eco:b1014 putA, ECK1005, JW0999, poaA, putC; fused DNA-binding transcriptional regulator/proline dehydrogenase/pyrroline-5-carboxylate dehydrogenase (EC:1.5.99.8 1.5.1.12); K00294 1-pyrroline-5-carboxylate dehydrogenase [EC:1.5.1.12] K00318 proline dehydrogenase [EC:1.5.99.8] (N)

atgggaaccaccaccatgggggttaagctggacgacgcgacgcgtgagcgtattaagtct
gccgcgacacgtatcgatcgcacaccacactggttaattaagcaggcgatttttcttat
ctcgaacaactggaaaacagcgatactctgccggagctacctgcgctgctttctggcgcg
gccaatgagagcgatgaagcaccgactccggcagaggaaccacaccagccattcctcgac
tttgccgagcaaatattgccccagtcggtttcccgcgccgcgatcaccgcggcctatcgc
cgcccggaaaccgaagcggtttctatgctgctggaacaagcccgcctgccgcagccagtt
gctgaacaggcgcacaaactggcgtatcagctggccgataaactgcgtaatcaaaaaaat
gccagtggtcgcgcaggtatggtccaggggttattgcaggagttttcgctgtcatcgcag
gaaggcgtggcgctgatgtgtctggcggaagcgttgttgcgtattcccgacaaagccacc
cgcgacgcgttaattcgcgacaaaatcagcaacggtaactggcagtcacacattggtcgt
agcccgtcactgtttgttaatgccgccacctgggggctgctgtttactggcaaactggtt
tccacccataacgaagccagcctctcccgctcgctgaaccgcattatcggtaaaagcggt
gaaccgctgatccgcaaaggtgtggatatggcgatgcgcctgatgggtgagcagttcgtc
actggcgaaaccatcgcggaagcgttagccaatgcccgcaagctggaagagaaaggtttc
cgttactcttacgatatgctgggcgaagccgcgctgaccgccgcagatgcacaggcgtat
atggtttcctatcagcaggcgattcacgccatcggtaaagcgtctaacggtcgtggcatc
tatgaagggccgggcatttcaatcaaactgtcggcgctgcatccgcgttatagccgcgcc
cagtatgaccgggtaatggaagagctttacccgcgtctgaaatcactcaccctgctggcg
cgtcagtacgatattggtatcaacattgacgccgaagagtccgatcgcctggagatctcc
ctcgatctgctggaaaaactctgtttcgagccggaactggcaggctggaacggcatcggt
tttgttattcaggcttatcaaaaacgctgcccgttggtgatcgattacctgattgatctc
gccacccgcagccgtcgccgtctgatgattcgcctggtgaaaggcgcgtactgggatagt
gaaattaagcgtgcgcagatggacggccttgaaggttatccggtttatacccgcaaggtg
tataccgacgtttcttatctcgcctgtgcgaaaaagctgctggcggtgccgaatctaatc
tacccgcagttcgcgacgcacaacgcccatacgctggcggcgatttatcaactggcgggg
cagaactactacccgggtcagtacgagttccagtgcctgcatggtatgggcgagccactg
tatgagcaggtcaccgggaaagttgccgacggcaaacttaaccgtccgtgtcgtatttat
gctccggttggcacacatgaaacgctgttggcgtatctggtgcgtcgcctgctggaaaac
ggtgctaacacctcgtttgttaaccgtattgccgacacctctttgccactggatgaactg
gtcgccgatccggtcactgctgtagaaaaactggcgcaacaggaagggcaaactggatta
ccgcatccgaaaattcccctgccgcgcgatctttacggtcacgggcgcgacaactcggca
gggctggatctcgctaacgaacaccgcctggcctcgctctcctctgccctgctcaatagt
gcactgcaaaaatggcaggccttgccaatgctggaacaaccggtagcggcaggtgagatg
tcgcccgttattaaccctgcggaaccgaaagatattgtgggctatgtgcgtgaagccacg
ccgcgtgaagtagaacaggcgctggaaagtgcggttaataacgcgccaatctggtttgcc
acgcctccggctgaacgcgcagcgattttgcaccgcgctgccgtgctgatggaaagccag
atgcagcaactgattggtattctggtgcgtgaggccggaaaaaccttcagtaacgccatt
gccgaagtgcgcgaagcggtcgattttctccactactacgccggacaggtgcgggatgat

ttcgctaacgaaacccaccgtccattagggcctgtggtgtgtatcagtccgtggaacttc
ccgctggctattttcaccgggcagatcgccgccgcactggcggcaggtaacagcgtgctg
gcaaaaccggcagaacaaacgccgctgattgccgcgcaagggatcgccattttgctggaa
gcgggtgtaccgccaggcgtggtgcaattgctgccaggtcggggtgaaaccgtgggcgcg
caactgacgggtgatgatcgcgtgcgcggggtgatgtttaccggttcaaccgaagtcgct
acgttactgcagcgcaatatcgccagccgcctggacgctcagggtcgccctattccgctc
atcgctgaaaccggcggcatgaacgcgatgattgtcgattcttcagcactgaccgaacag
gtcgtcgtggatgtactggcctcggcgttcgacagtgcgggtcagcgttgttcggcgctg
cgcgtgctgtgcctgcaagatgagattgccgaccacacgttgaaaatgctgcgcggcgca
atggccgaatgccggatgggtaatccgggtcgcctgaccaccgatatcggtccagtgatt
gatagcgaagcgaaagccaatattgagcgccatattcagaccatgcgtagcaaaggccgt
ccggtgttccaggcggtgcgggaaaacagcgaagatgcccgtgaatggcaaagcggcacc
tttgtcgccccgacgctgatcgaactggatgactttgccgaattgcaaaaagaggtcttt
ggtccggtgctgcatgtggtgcgttacaaccgtaaccagctaccagagctgatcgagcag
attaacgcttccggttatggtctgacgcttggcgtccatacgcgcattgatgaaaccatc
gcccaggtcactggctcggcccatgttggtaacctgtatgttaaccgtaatatggtgggc
gcagtggttggtgtgcagccgttcggcggcgaagggttgtccggtaccgggccgaaagca
ggcggtccgctctatctctaccgtctgctggcgaatcgcccggaaagtgcgctggcagtg
acgctcgcgcgtcaggatgcaaagtatccggtcgatgcgcagttgaaagccgcattgact
cagccgctaaatgcactgcgggaatgggcagcaaatcgtccagaattgcaggcgttatgt
acgcaatatggcgagctggcgcaggcaggaacacaacgattgctgccggggccgacgggt
gaacgcaacacctggacgctgctgccgcgtgagcgcgtgttgtgtattgccgatgatgag
caggatgcgctgactcagctcgccgccgtgctggcggtgggcagccaggtactgtggccg
gatgacgcgctgcatcgtcagttagtgaaggcattgccatcggcagtcagcgaacgtatt
caactggcgaaagcggaaaatataaccgctcaaccgtttgatgcggtgatcttccacggt
gattcggatcagcttcgcgcattgtgtgaagcagttgccgcgcgggatggcacaattgtt
tcggtgcagggttttgcccgtggcgaaagcaatatccttctggaacggctgtatatcgag
cgttcgctgagtgtgaataccgctgccgctggcggtaacgccagcttaatgactataggt
taa

APPENDIX B

BLAST REPORT "EXACT" ASSEMBLED k = 21 EXPECTED COVERAGE = 4

COVERAGE CUTOFF = 2 VS. REFERENCE

```
 velveth out_NP_415534-Exact_21 21 -fasta -shortPaired NP_415534-Exact.fasta
 velvetg out_NP_415534-Exact_21_2_4_dir -exp_cov 4 -cov_cutoff 2 -read_trkg yes -
amos_file yes -unused_reads yes
Final graph has 1 nodes and n50 of 3963, max 3963, total 3963, using 5000/5000 reads

BLASTN 2.2.23+
Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro
A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and
David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new
generation of protein database search programs", Nucleic
Acids Res. 25:3389-3402.


RID: U317VMZJ112


Query=  eco:b1014 putA, ECK1005, JW0999, poaA, putC; fused DNA-binding
transcriptional regulator/proline
dehydrogenase/pyrroline-5-carboxylate dehydrogenase (EC:1.5.99.8
1.5.1.12); K00294 1-pyrroline-5-carboxylate dehydrogenase
[EC:1.5.1.12] K00318 proline dehydrogenase [EC:1.5.99.8] (N)
Length=3963


                                                                 Score     E
Sequences producing significant alignments:                      (Bits)  Value

lcl|29253  NODE_1_length_3963_cov_1233.750977                     4911    0.0

ALIGNMENTS
>lcl|29253 NODE_1_length_3963_cov_1233.750977
Length=3983

 Score = 4911 bits (5446),  Expect = 0.0
 Identities = 2723/2723 (100%), Gaps = 0/2723 (0%)
 Strand=Plus/Plus

Query  1     ATGGGAACCACCACCATGGGGGTTAAGCTGGACGACGCGACGCGTGAGCGTATTAAGTCT  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1261  ATGGGAACCACCACCATGGGGGTTAAGCTGGACGACGCGACGCGTGAGCGTATTAAGTCT  1320

Query  61    GCCGCGACACGTATCGATCGCACACCACACTGGTTAATTAAGCAGGCGATTTTTTCTTAT  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1321  GCCGCGACACGTATCGATCGCACACCACACTGGTTAATTAAGCAGGCGATTTTTTCTTAT  1380

Query  121   CTCGAACAACTGGAAAACAGCGATACTCTGCCGGAGCTACCTGCGCTGCTTTCTGGCGCG  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1381  CTCGAACAACTGGAAAACAGCGATACTCTGCCGGAGCTACCTGCGCTGCTTTCTGGCGCG  1440

Query  181   GCCAATGAGAGCGATGAAGCACCGACTCCGGCAGAGGAACCACACCAGCCATTCCTCGAC  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1441  GCCAATGAGAGCGATGAAGCACCGACTCCGGCAGAGGAACCACACCAGCCATTCCTCGAC  1500

Query  241   TTTGCCGAGCAAATATTGCCCCAGTCGGTTTCCCGCGCCGCGATCACCGCGGCCTATCGC  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1501  TTTGCCGAGCAAATATTGCCCCAGTCGGTTTCCCGCGCCGCGATCACCGCGGCCTATCGC  1560


Query  301   CGCCCGGAAACCGAAGCGGTTTCTATGCTGCTGGAACAAGCCCGCCTGCCGCAGCCAGTT  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1561  CGCCCGGAAACCGAAGCGGTTTCTATGCTGCTGGAACAAGCCCGCCTGCCGCAGCCAGTT  1620

Query  361   GCTGAACAGGCGCACAAACTGGCGTATCAGCTGGCCGATAAACTGCGTAATCnnnnnnnT  420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1621  GCTGAACAGGCGCACAAACTGGCGTATCAGCTGGCCGATAAACTGCGTAATCAAAAAAAT  1680

Query  421   GCCAGTGGTCGCGCAGGTATGGTCCAGGGGTTATTGCAGGAGTTTTCGCTGTCATCGCAG  480
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1681  GCCAGTGGTCGCGCAGGTATGGTCCAGGGGTTATTGCAGGAGTTTTCGCTGTCATCGCAG  1740
```

```
Query  481   GAAGGCGTGGCGCTGATGTGTCTGGCGGAAGCGTTGTTGCGTATTCCCGACAAAGCCACC  540
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1741  GAAGGCGTGGCGCTGATGTGTCTGGCGGAAGCGTTGTTGCGTATTCCCGACAAAGCCACC  1800

Query  541   CGCGACGCGTTAATTCGCGACAAAATCAGCAACGGTAACTGGCAGTCACACATTGGTCGT  600
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1801  CGCGACGCGTTAATTCGCGACAAAATCAGCAACGGTAACTGGCAGTCACACATTGGTCGT  1860

Query  601   AGCCCGTCACTGTTTGTTAATGCCGCCACCTGGGGGCTGCTGTTTACTGGCAAACTGGTT  660
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1861  AGCCCGTCACTGTTTGTTAATGCCGCCACCTGGGGGCTGCTGTTTACTGGCAAACTGGTT  1920

Query  661   TCCACCCATAACGAAGCCAGCCTCTCCCGCTCGCTGAACCGCATTATCGGTAAAAGCGGT  720
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1921  TCCACCCATAACGAAGCCAGCCTCTCCCGCTCGCTGAACCGCATTATCGGTAAAAGCGGT  1980

Query  721   GAACCGCTGATCCGCAAAGGTGTGGATATGGCGATGCGCCTGATGGGTGAGCAGTTCGTC  780
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1981  GAACCGCTGATCCGCAAAGGTGTGGATATGGCGATGCGCCTGATGGGTGAGCAGTTCGTC  2040

Query  781   ACTGGCGAAACCATCGCGGAAGCGTTAGCCAATGCCCGCAAGCTGGAAGAGAAAGGTTTC  840
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2041  ACTGGCGAAACCATCGCGGAAGCGTTAGCCAATGCCCGCAAGCTGGAAGAGAAAGGTTTC  2100

Query  841   CGTTACTCTTACGATATGCTGGGCGAAGCCGCGCTGACCGCCGCAGATGCACAGGCGTAT  900
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2101  CGTTACTCTTACGATATGCTGGGCGAAGCCGCGCTGACCGCCGCAGATGCACAGGCGTAT  2160

Query  901   ATGGTTTCCTATCAGCAGGCGATTCACGCCATCGGTAAAGCGTCTAACGGTCGTGGCATC  960
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2161  ATGGTTTCCTATCAGCAGGCGATTCACGCCATCGGTAAAGCGTCTAACGGTCGTGGCATC  2220

Query  961   TATGAAGGGCCGGGCATTTCAATCAAACTGTCGGCGCTGCATCCGCGTTATAGCCGCGCC  1020
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2221  TATGAAGGGCCGGGCATTTCAATCAAACTGTCGGCGCTGCATCCGCGTTATAGCCGCGCC  2280

Query  1021  CAGTATGACCGGGTAATGGAAGAGCTTTACCCGCGTCTGAAATCACTCACCCTGCTGGCG  1080
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2281  CAGTATGACCGGGTAATGGAAGAGCTTTACCCGCGTCTGAAATCACTCACCCTGCTGGCG  2340

Query  1081  CGTCAGTACGATATTGGTATCAACATTGACGCCGAAGAGTCCGATCGCCTGGAGATCTCC  1140
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2341  CGTCAGTACGATATTGGTATCAACATTGACGCCGAAGAGTCCGATCGCCTGGAGATCTCC  2400

Query  1141  CTCGATCTGCTGGAAAAACTCTGTTTCGAGCCGGAACTGGCAGGCTGGAACGGCATCGGT  1200
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2401  CTCGATCTGCTGGAAAAACTCTGTTTCGAGCCGGAACTGGCAGGCTGGAACGGCATCGGT  2460

Query  1201  TTTGTTATTCAGGCTTATCAAAAACGCTGCCCGTTGGTGATCGATTACCTGATTGATCTC  1260
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2461  TTTGTTATTCAGGCTTATCAAAAACGCTGCCCGTTGGTGATCGATTACCTGATTGATCTC  2520

Query  1261  GCCACCCGCAGCCGTCGCCGTCTGATGATTCGCCTGGTGAAAGGCGCGTACTGGGATAGT  1320
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2521  GCCACCCGCAGCCGTCGCCGTCTGATGATTCGCCTGGTGAAAGGCGCGTACTGGGATAGT  2580

Query  1321  GAAATTAAGCGTGCGCAGATGGACGGCCTTGAAGGTTATCCGGTTTATACCCGCAAGGTG  1380
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2581  GAAATTAAGCGTGCGCAGATGGACGGCCTTGAAGGTTATCCGGTTTATACCCGCAAGGTG  2640

Query  1381  TATACCGACGTTTCTTATCTCGCCTGTGCGAAAAAGCTGCTGGCGGTGCCGAATCTAATC  1440
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2641  TATACCGACGTTTCTTATCTCGCCTGTGCGAAAAAGCTGCTGGCGGTGCCGAATCTAATC  2700

Query  1441  TACCCGCAGTTCGCGACGCACAACGCCCATACGCTGGCGGCGATTTATCAACTGGCGGGG  1500
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2701  TACCCGCAGTTCGCGACGCACAACGCCCATACGCTGGCGGCGATTTATCAACTGGCGGGG  2760

Query  1501  CAGAACTACTACCCGGGTCAGTACGAGTTCCAGTGCCTGCATGGTATGGGCGAGCCACTG  1560
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2761  CAGAACTACTACCCGGGTCAGTACGAGTTCCAGTGCCTGCATGGTATGGGCGAGCCACTG  2820

Query  1561  TATGAGCAGGTCACCGGGAAAGTTGCCGACGGCAAACTTAACCGTCCGTGTCGTATTTAT  1620
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2821  TATGAGCAGGTCACCGGGAAAGTTGCCGACGGCAAACTTAACCGTCCGTGTCGTATTTAT  2880

Query  1621  GCTCCGGTTGGCACACATGAAACGCTGTTGGCGTATCTGGTGCGTCGCCTGCTGGAAAAC  1680
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2881  GCTCCGGTTGGCACACATGAAACGCTGTTGGCGTATCTGGTGCGTCGCCTGCTGGAAAAC  2940
```

```
Query  1681  GGTGCTAACACCTCGTTTGTTAACCGTATTGCCGACACCTCTTTGCCACTGGATGAACTG  1740
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2941  GGTGCTAACACCTCGTTTGTTAACCGTATTGCCGACACCTCTTTGCCACTGGATGAACTG  3000

Query  1741  GTCGCCGATCCGGTCACTGCTGTAGAAAAACTGGCGCAACAGGAAGGGCAAACTGGATTA  1800
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3001  GTCGCCGATCCGGTCACTGCTGTAGAAAAACTGGCGCAACAGGAAGGGCAAACTGGATTA  3060

Query  1801  CCGCATCCGAAAATTCCCCTGCCGCGCGATCTTTACGGTCACGGGCGCGACAACTCGGCA  1860
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3061  CCGCATCCGAAAATTCCCCTGCCGCGCGATCTTTACGGTCACGGGCGCGACAACTCGGCA  3120

Query  1861  GGGCTGGATCTCGCTAACGAACACCGCCTGGCCTCGCTCTCCTCTGCCCTGCTCAATAGT  1920
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3121  GGGCTGGATCTCGCTAACGAACACCGCCTGGCCTCGCTCTCCTCTGCCCTGCTCAATAGT  3180

Query  1921  GCACTGCAAAAATGGCAGGCCTTGCCAATGCTGGAACAACCGGTAGCGGCAGGTGAGATG  1980
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3181  GCACTGCAAAAATGGCAGGCCTTGCCAATGCTGGAACAACCGGTAGCGGCAGGTGAGATG  3240

Query  1981  TCGCCCGTTATTAACCCTGCGGAACCGAAAGATATTGTGGGCTATGTGCGTGAAGCCACG  2040
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3241  TCGCCCGTTATTAACCCTGCGGAACCGAAAGATATTGTGGGCTATGTGCGTGAAGCCACG  3300

Query  2041  CCGCGTGAAGTAGAACAGGCGCTGGAAAGTGCGGTTAATAACGCGCCAATCTGGTTTGCC  2100
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3301  CCGCGTGAAGTAGAACAGGCGCTGGAAAGTGCGGTTAATAACGCGCCAATCTGGTTTGCC  3360

Query  2101  ACGCCTCCGGCTGAACGCGCAGCGATTTTGCACCGCGCTGCCGTGCTGATGGAAAGCCAG  2160
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3361  ACGCCTCCGGCTGAACGCGCAGCGATTTTGCACCGCGCTGCCGTGCTGATGGAAAGCCAG  3420

Query  2161  ATGCAGCAACTGATTGGTATTCTGGTGCGTGAGGCCGGAAAAACCTTCAGTAACGCCATT  2220
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3421  ATGCAGCAACTGATTGGTATTCTGGTGCGTGAGGCCGGAAAAACCTTCAGTAACGCCATT  3480

Query  2221  GCCGAAGTGCGCGAAGCGGTCGATTTTCTCCACTACTACGCCGGACAGGTGCGGGATGAT  2280
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3481  GCCGAAGTGCGCGAAGCGGTCGATTTTCTCCACTACTACGCCGGACAGGTGCGGGATGAT  3540

Query  2281  TTCGCTAACGAAACCCACCGTCCATTAGGGCCTGTGGTGTGTATCAGTCCGTGGAACTTC  2340
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3541  TTCGCTAACGAAACCCACCGTCCATTAGGGCCTGTGGTGTGTATCAGTCCGTGGAACTTC  3600

Query  2341  CCGCTGGCTATTTTCACCGGGCAGATCGCCGCCGCACTGGCGGCAGGTAACAGCGTGCTG  2400
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3601  CCGCTGGCTATTTTCACCGGGCAGATCGCCGCCGCACTGGCGGCAGGTAACAGCGTGCTG  3660

Query  2401  GCAAAACCGGCAGAACAAACGCCGCTGATTGCCGCGCAAGGGATCGCCATTTTGCTGGAA  2460
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3661  GCAAAACCGGCAGAACAAACGCCGCTGATTGCCGCGCAAGGGATCGCCATTTTGCTGGAA  3720

Query  2461  GCGGGTGTACCGCCAGGCGTGGTGCAATTGCTGCCAGGTCGGGGTGAAACCGTGGGCGCG  2520
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3721  GCGGGTGTACCGCCAGGCGTGGTGCAATTGCTGCCAGGTCGGGGTGAAACCGTGGGCGCG  3780

Query  2521  CAACTGACGGGTGATGATCGCGTGCGCGGGGTGATGTTTACCGGTTCAACCGAAGTCGCT  2580
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3781  CAACTGACGGGTGATGATCGCGTGCGCGGGGTGATGTTTACCGGTTCAACCGAAGTCGCT  3840

Query  2581  ACGTTACTGCAGCGCAATATCGCCAGCCGCCTGGACGCTCAGGGTCGCCCTATTCCGCTC  2640
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3841  ACGTTACTGCAGCGCAATATCGCCAGCCGCCTGGACGCTCAGGGTCGCCCTATTCCGCTC  3900

Query  2641  ATCGCTGAAACCGGCGGCATGAACGCGATGATTGTCGATTCTTCAGCACTGACCGAACAG  2700
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3901  ATCGCTGAAACCGGCGGCATGAACGCGATGATTGTCGATTCTTCAGCACTGACCGAACAG  3960

Query  2701  GTCGTCGTGGATGTACTGGCCTC  2723
             |||||||||||||||||||||||
Sbjct  3961  GTCGTCGTGGATGTACTGGCCTC  3983


 Score = 2273 bits (2520),  Expect = 0.0
 Identities = 1260/1260 (100%), Gaps = 0/1260 (0%)
 Strand=Plus/Plus

Query  2704  GTCGTGGATGTACTGGCCTCGGCGTTCGACAGTGCGGGTCAGCGTTGTTCGGCGCTGCGC  2763
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     GTCGTGGATGTACTGGCCTCGGCGTTCGACAGTGCGGGTCAGCGTTGTTCGGCGCTGCGC  60
```

```
Query  2764  GTGCTGTGCCTGCAAGATGAGATTGCCGACCACACGTTGAAAATGCTGCGCGGCGCAATG  2823
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    GTGCTGTGCCTGCAAGATGAGATTGCCGACCACACGTTGAAAATGCTGCGCGGCGCAATG  120

Query  2824  GCCGAATGCCGGATGGGTAATCCGGGTCGCCTGACCACCGATATCGGTCCAGTGATTGAT  2883
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   GCCGAATGCCGGATGGGTAATCCGGGTCGCCTGACCACCGATATCGGTCCAGTGATTGAT  180

Query  2884  AGCGAAGCGAAAGCCAATATTGAGCGCCATATTCAGACCATGCGTAGCAAAGGCCGTCCG  2943
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181   AGCGAAGCGAAAGCCAATATTGAGCGCCATATTCAGACCATGCGTAGCAAAGGCCGTCCG  240

Query  2944  GTGTTCCAGGCGGTGCGGGAAAACAGCGAAGATGCCCGTGAATGGCAAAGCGGCACCTTT  3003
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241   GTGTTCCAGGCGGTGCGGGAAAACAGCGAAGATGCCCGTGAATGGCAAAGCGGCACCTTT  300

Query  3004  GTCGCCCCGACGCTGATCGAACTGGATGACTTTGCCGAATTGCAAAAAGAGGTCTTTGGT  3063
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301   GTCGCCCCGACGCTGATCGAACTGGATGACTTTGCCGAATTGCAAAAAGAGGTCTTTGGT  360

Query  3064  CCGGTGCTGCATGTGGTGCGTTACAACCGTAACCAGCTACCAGAGCTGATCGAGCAGATT  3123
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  361   CCGGTGCTGCATGTGGTGCGTTACAACCGTAACCAGCTACCAGAGCTGATCGAGCAGATT  420


Query  3124  AACGCTTCCGGTTATGGTCTGACGCTTGGCGTCCATACGCGCATTGATGAAACCATCGCC  3183
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  421   AACGCTTCCGGTTATGGTCTGACGCTTGGCGTCCATACGCGCATTGATGAAACCATCGCC  480

Query  3184  CAGGTCACTGGCTCGGCCCATGTTGGTAACCTGTATGTTAACCGTAATATGGTGGGCGCA  3243
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  481   CAGGTCACTGGCTCGGCCCATGTTGGTAACCTGTATGTTAACCGTAATATGGTGGGCGCA  540

Query  3244  GTGGTTGGTGTGCAGCCGTTCGGCGGCGAAGGGTTGTCCGGTACCGGGCCGAAAGCAGGC  3303
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  541   GTGGTTGGTGTGCAGCCGTTCGGCGGCGAAGGGTTGTCCGGTACCGGGCCGAAAGCAGGC  600

Query  3304  GGTCCGCTCTATCTCTACCGTCTGCTGGCGAATCGCCCGGAAAGTGCGCTGGCAGTGACG  3363
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  601   GGTCCGCTCTATCTCTACCGTCTGCTGGCGAATCGCCCGGAAAGTGCGCTGGCAGTGACG  660

Query  3364  CTCGCGCGTCAGGATGCAAAGTATCCGGTCGATGCGCAGTTGAAAGCCGCATTGACTCAG  3423
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  661   CTCGCGCGTCAGGATGCAAAGTATCCGGTCGATGCGCAGTTGAAAGCCGCATTGACTCAG  720

Query  3424  CCGCTAAATGCACTGCGGGAATGGGCAGCAAATCGTCCAGAATTGCAGGCGTTATGTACG  3483
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  721   CCGCTAAATGCACTGCGGGAATGGGCAGCAAATCGTCCAGAATTGCAGGCGTTATGTACG  780

Query  3484  CAATATGGCGAGCTGGCGCAGGCAGGAACACAACGATTGCTGCCGGGGCCGACGGGTGAA  3543
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  781   CAATATGGCGAGCTGGCGCAGGCAGGAACACAACGATTGCTGCCGGGGCCGACGGGTGAA  840

Query  3544  CGCAACACCTGGACGCTGCTGCCGCGTGAGCGCGTGTTGTGTATTGCCGATGATGAGCAG  3603
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  841   CGCAACACCTGGACGCTGCTGCCGCGTGAGCGCGTGTTGTGTATTGCCGATGATGAGCAG  900

Query  3604  GATGCGCTGACTCAGCTCGCCGCCGTGCTGGCGGTGGGCAGCCAGGTACTGTGGCCGGAT  3663
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  901   GATGCGCTGACTCAGCTCGCCGCCGTGCTGGCGGTGGGCAGCCAGGTACTGTGGCCGGAT  960

Query  3664  GACGCGCTGCATCGTCAGTTAGTGAAGGCATTGCCATCGGCAGTCAGCGAACGTATTCAA  3723
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  961   GACGCGCTGCATCGTCAGTTAGTGAAGGCATTGCCATCGGCAGTCAGCGAACGTATTCAA  1020

Query  3724  CTGGCGAAAGCGGAAAATATAACCGCTCAACCGTTTGATGCGGTGATCTTCCACGGTGAT  3783
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1021  CTGGCGAAAGCGGAAAATATAACCGCTCAACCGTTTGATGCGGTGATCTTCCACGGTGAT  1080

Query  3784  TCGGATCAGCTTCGCGCATTGTGTGAAGCAGTTGCCGCGCGGGGATGGCACAATTGTTTCG  3843
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1081  TCGGATCAGCTTCGCGCATTGTGTGAAGCAGTTGCCGCGCGGGGATGGCACAATTGTTTCG  1140

Query  3844  GTGCAGGGTTTTGCCCGTGGCGAAAGCAATATCCTTCTGGAACGGCTGTATATCGAGCGT  3903
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1141  GTGCAGGGTTTTGCCCGTGGCGAAAGCAATATCCTTCTGGAACGGCTGTATATCGAGCGT  1200

Query  3904  TCGCTGAGTGTGAATACCGCTGCCGCTGGCGGTAACGCCAGCTTAATGACTATAGGTTAA  3963
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1201  TCGCTGAGTGTGAATACCGCTGCCGCTGGCGGTAACGCCAGCTTAATGACTATAGGTTAA  1260
```

```
 Score = 22.9 bits (24),  Expect = 2.0
 Identities = 12/12 (100%), Gaps = 0/12 (0%)
 Strand=Plus/Minus

Query  2860  ACCGATATCGGT  2871
             ||||||||||||
Sbjct  168   ACCGATATCGGT  157


 Score = 22.9 bits (24),  Expect = 2.0
 Identities = 12/12 (100%), Gaps = 0/12 (0%)
 Strand=Plus/Plus

Query  132  GGAAAACAGCGA  143
             ||||||||||||
Sbjct  258  GGAAAACAGCGA  269


 Score = 22.9 bits (24),  Expect = 2.0
 Identities = 17/19 (89%), Gaps = 1/19 (5%)
 Strand=Plus/Plus

Query  1690  ACCTCGTTTGTTAACCGTA  1708
              |||| || ||||||||||
Sbjct  509   ACCT-GTATGTTAACCGTA  526


 Score = 22.9 bits (24),  Expect = 2.0
 Identities = 12/12 (100%), Gaps = 0/12 (0%)
 Strand=Plus/Plus

Query  2961  GGAAAACAGCGA  2972
              ||||||||||||
Sbjct  1392  GGAAAACAGCGA  1403


 Score = 22.9 bits (24),  Expect = 2.0
 Identities = 17/19 (89%), Gaps = 1/19 (5%)
 Strand=Plus/Plus

Query  3212  ACCT-GTATGTTAACCGTA  3229
              |||| || ||||||||||
Sbjct  2950  ACCTCGTTTGTTAACCGTA  2968


 Score = 21.1 bits (22),  Expect = 6.8
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Plus

Query  238  GACTTTGCCGA  248
             ||||||||||
Sbjct  328  GACTTTGCCGA  338


 Score = 21.1 bits (22),  Expect = 6.8
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Minus

Query  797  CGGAAGCGTTA  807
             ||||||||||
Sbjct  430  CGGAAGCGTTA  420


 Score = 21.1 bits (22),  Expect = 6.8
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Plus

Query  787  GAAACCATCGC  797
             ||||||||||
Sbjct  469  GAAACCATCGC  479


 Score = 21.1 bits (22),  Expect = 6.8
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Plus
```

```
Query  1418  TGCTGGCGGTG  1428
             ||||||||||
Sbjct  926   TGCTGGCGGTG  936


 Score = 21.1 bits (22),  Expect = 6.8
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Minus

Query  206   CTCCGGCAGAG  216
             ||||||||||
Sbjct  1416  CTCCGGCAGAG  1406


 Score = 21.1 bits (22),  Expect = 6.8
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Minus

Query  146   CTCTGCCGGAG  156
             ||||||||||
Sbjct  1476  CTCTGCCGGAG  1466


 Score = 21.1 bits (22),  Expect = 6.8
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Plus

Query  3031  GACTTTGCCGA  3041
             ||||||||||
Sbjct  1498  GACTTTGCCGA  1508


 Score = 21.1 bits (22),  Expect = 6.8
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Plus

Query  3172  GAAACCATCGC  3182
             ||||||||||
Sbjct  2047  GAAACCATCGC  2057


 Score = 21.1 bits (22),  Expect = 6.8
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Minus

Query  3123  TAACGCTTCCG  3133
             ||||||||||
Sbjct  2067  TAACGCTTCCG  2057


 Score = 21.1 bits (22),  Expect = 6.8
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Plus

Query  3629  TGCTGGCGGTG  3639
             ||||||||||
Sbjct  2678  TGCTGGCGGTG  2688


 Score = 21.1 bits (22),  Expect = 6.8
 Identities = 13/14 (92%), Gaps = 0/14 (0%)
 Strand=Plus/Minus

Query  2073  GGTTAATAACGCGC  2086
             |||||||||| ||
Sbjct  3256  GGTTAATAACGGGC  3243


 Score = 21.1 bits (22),  Expect = 6.8
 Identities = 13/14 (92%), Gaps = 0/14 (0%)
 Strand=Plus/Minus

Query  1983  GCCCGTTATTAACC  1996
             || ||||||||||
Sbjct  3346  GCGCGTTATTAACC  3333
```

APPENDIX C

BLAST REPORT 454 ASSEMBLED k = 21 EXPECTED COVERAGE = 4

COVERAGE CUTOFF = 2 VS. REFERENCE

```
 velveth out_NP_415534-454_21 21 -fasta -shortPaired NP_415534-454.fasta
 velvetg out_NP_415534-454_21_2_4_dir -exp_cov 4 -cov_cutoff 2 -read_trkg yes -amos_file
yes -unused_reads yes
Final graph has 4253 nodes and n50 of 16, max 36, total 29411, using 1394/5000 reads

BLASTN 2.2.23+
Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro
A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and
David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new
generation of protein database search programs", Nucleic
Acids Res. 25:3389-3402.


RID: U30ZCU4Y114


Query=  eco:b1014 putA, ECK1005, JW0999, poaA, putC; fused DNA-binding
transcriptional regulator/proline
dehydrogenase/pyrroline-5-carboxylate dehydrogenase (EC:1.5.99.8
1.5.1.12); K00294 1-pyrroline-5-carboxylate dehydrogenase
[EC:1.5.1.12] K00318 proline dehydrogenase [EC:1.5.99.8] (N)
Length=3963
```

Sequences producing significant alignments:		Score (Bits)	E Value
lcl|52974	NODE_11_length_21_cov_7.142857	66.2	1e-15
lcl|52975	NODE_159_length_21_cov_2.000000	59.0	2e-13
lcl|52976	NODE_273_length_21_cov_3.952381	66.2	1e-15
lcl|52977	NODE_893_length_21_cov_5.047619	66.2	1e-15
lcl|52978	NODE_1046_length_21_cov_6.666667	66.2	1e-15
lcl|52979	NODE_1086_length_23_cov_93.173912	53.6	9e-12
lcl|52980	NODE_1139_length_21_cov_27.714285	46.4	1e-09
lcl|52981	NODE_1229_length_21_cov_3.571429	66.2	1e-15
lcl|52982	NODE_1278_length_21_cov_4.047619	66.2	1e-15
lcl|52983	NODE_1407_length_21_cov_4.904762	66.2	1e-15
lcl|52984	NODE_1540_length_36_cov_27.888889	42.8	2e-08
lcl|52985	NODE_1819_length_25_cov_2.360000	59.0	2e-13
lcl|52986	NODE_1834_length_21_cov_5.523809	66.2	1e-15
lcl|52987	NODE_1931_length_21_cov_4.904762	66.2	1e-15
lcl|52988	NODE_1961_length_21_cov_2.428571	66.2	1e-15
lcl|52989	NODE_1994_length_21_cov_5.000000	66.2	1e-15
lcl|52990	NODE_2005_length_21_cov_4.714286	66.2	1e-15
lcl|52991	NODE_2082_length_21_cov_3.238095	66.2	1e-15
lcl|52992	NODE_2191_length_21_cov_5.095238	66.2	1e-15
lcl|52993	NODE_2199_length_21_cov_7.380952	66.2	1e-15
lcl|52994	NODE_2212_length_21_cov_4.238095	66.2	1e-15
lcl|52995	NODE_2217_length_21_cov_5.904762	66.2	1e-15
lcl|52996	NODE_2236_length_21_cov_6.190476	66.2	1e-15
lcl|52997	NODE_2238_length_21_cov_2.857143	66.2	1e-15
lcl|52998	NODE_2257_length_21_cov_4.714286	66.2	1e-15
lcl|52999	NODE_2258_length_21_cov_4.523809	66.2	1e-15
lcl|53000	NODE_2275_length_21_cov_3.095238	66.2	1e-15
lcl|53001	NODE_2294_length_21_cov_3.428571	66.2	1e-15
lcl|53002	NODE_2316_length_21_cov_4.714286	66.2	1e-15
lcl|53003	NODE_2332_length_21_cov_2.619048	66.2	1e-15
lcl|53004	NODE_2338_length_21_cov_4.857143	66.2	1e-15
lcl|53005	NODE_2351_length_21_cov_4.761905	66.2	1e-15
lcl|53006	NODE_2355_length_21_cov_7.142857	66.2	1e-15
lcl|53007	NODE_2369_length_21_cov_3.428571	66.2	1e-15
lcl|53008	NODE_2370_length_21_cov_7.190476	66.2	1e-15
lcl|53009	NODE_2371_length_21_cov_4.190476	66.2	1e-15
lcl|53010	NODE_2377_length_21_cov_2.190476	66.2	1e-15
lcl|53011	NODE_2396_length_21_cov_2.714286	66.2	1e-15
lcl|53012	NODE_2409_length_21_cov_5.142857	66.2	1e-15
lcl|53013	NODE_2417_length_21_cov_2.238095	66.2	1e-15
lcl|53014	NODE_2418_length_21_cov_6.190476	66.2	1e-15
lcl|53015	NODE_2453_length_21_cov_6.571429	66.2	1e-15
lcl|53016	NODE_2516_length_21_cov_3.047619	66.2	1e-15

lcl|53017	NODE_2544_length_21_cov_3.000000		66.2	1e-15
lcl|53018	NODE_2551_length_21_cov_4.333333		66.2	1e-15
lcl|53019	NODE_2559_length_21_cov_5.190476		66.2	1e-15
lcl|53020	NODE_2566_length_21_cov_6.000000		66.2	1e-15
lcl|53021	NODE_2633_length_21_cov_3.571429		66.2	1e-15
lcl|53022	NODE_2655_length_21_cov_2.666667		59.0	2e-13
lcl|53023	NODE_2660_length_21_cov_6.238095		66.2	1e-15
lcl|53024	NODE_2695_length_21_cov_3.619048		66.2	1e-15
lcl|53025	NODE_2717_length_21_cov_3.238095		66.2	1e-15
lcl|53026	NODE_2718_length_21_cov_2.857143		66.2	1e-15
lcl|53027	NODE_2773_length_21_cov_8.380953		66.2	1e-15
lcl|53028	NODE_2791_length_21_cov_3.523809		66.2	1e-15
lcl|53029	NODE_2808_length_21_cov_4.047619		66.2	1e-15
lcl|53030	NODE_2822_length_21_cov_8.142858		66.2	1e-15
lcl|53031	NODE_2832_length_21_cov_8.380953		66.2	1e-15
lcl|53032	NODE_2898_length_21_cov_3.190476		66.2	1e-15
lcl|53033	NODE_2899_length_21_cov_2.428571		59.0	2e-13
lcl|53034	NODE_2906_length_21_cov_4.714286		66.2	1e-15
lcl|53035	NODE_2922_length_21_cov_4.476191		66.2	1e-15
lcl|53036	NODE_2940_length_21_cov_3.619048		66.2	1e-15
lcl|53037	NODE_2949_length_21_cov_8.047619		66.2	1e-15
lcl|53038	NODE_2981_length_21_cov_3.428571		66.2	1e-15
lcl|53039	NODE_3005_length_21_cov_2.428571		66.2	1e-15
lcl|53040	NODE_3009_length_21_cov_4.476191		66.2	1e-15
lcl|53041	NODE_3023_length_21_cov_4.142857		66.2	1e-15
lcl|53042	NODE_3058_length_21_cov_5.952381		66.2	1e-15
lcl|53043	NODE_3067_length_21_cov_3.809524		66.2	1e-15
lcl|53044	NODE_3101_length_21_cov_5.047619		66.2	1e-15
lcl|53045	NODE_3119_length_21_cov_6.142857		66.2	1e-15
lcl|53046	NODE_3122_length_36_cov_2.083333		77.0	1e-18
lcl|53047	NODE_3153_length_21_cov_6.190476		66.2	1e-15
lcl|53048	NODE_3162_length_21_cov_3.285714		59.0	2e-13
lcl|53049	NODE_3168_length_21_cov_4.285714		59.0	2e-13
lcl|53050	NODE_3170_length_21_cov_2.380952		66.2	1e-15
lcl|53051	NODE_3181_length_21_cov_2.619048		66.2	1e-15
lcl|53052	NODE_3182_length_21_cov_3.952381		66.2	1e-15
lcl|53053	NODE_3234_length_21_cov_2.476191		66.2	1e-15
lcl|53054	NODE_3245_length_21_cov_5.238095		66.2	1e-15
lcl|53055	NODE_3262_length_21_cov_4.714286		66.2	1e-15
lcl|53056	NODE_3264_length_21_cov_6.428571		66.2	1e-15
lcl|53057	NODE_3273_length_21_cov_3.190476		66.2	1e-15
lcl|53058	NODE_3276_length_21_cov_5.428571		66.2	1e-15
lcl|53059	NODE_3279_length_21_cov_3.571429		66.2	1e-15
lcl|53060	NODE_3294_length_21_cov_4.761905		66.2	1e-15
lcl|53061	NODE_3301_length_21_cov_3.619048		66.2	1e-15
lcl|53062	NODE_3302_length_21_cov_3.761905		66.2	1e-15
lcl|53063	NODE_3312_length_21_cov_4.000000		66.2	1e-15
lcl|53064	NODE_3314_length_21_cov_2.000000		66.2	1e-15
lcl|53065	NODE_3317_length_21_cov_2.809524		66.2	1e-15
lcl|53066	NODE_3330_length_21_cov_3.238095		66.2	1e-15
lcl|53067	NODE_3347_length_21_cov_3.333333		66.2	1e-15
lcl|53068	NODE_3370_length_21_cov_3.142857		66.2	1e-15
lcl|53069	NODE_3387_length_21_cov_2.714286		66.2	1e-15
lcl|53070	NODE_3404_length_21_cov_5.476191		66.2	1e-15
lcl|53071	NODE_3428_length_21_cov_5.523809		66.2	1e-15
lcl|53072	NODE_3470_length_21_cov_2.571429		59.0	2e-13
lcl|53073	NODE_3488_length_21_cov_2.904762		66.2	1e-15

ALIGNMENTS
>lcl|52974 NODE_11_length_21_cov_7.142857
Length=41

Score = 66.2 bits (72), Expect = 1e-15
Identities = 40/41 (97%), Gaps = 1/41 (2%)
Strand=Plus/Plus

```
Query  1925  TGCAAAAATGGCAGGCCTTG-CCAATGCTGGAACAACCGGT  1964
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     TGCAAAAATGGCAGGCCTTGTCCAATGCTGGAACAACCGGT  41
```

Score = 21.1 bits (22), Expect = 0.052
Identities = 11/11 (100%), Gaps = 0/11 (0%)
Strand=Plus/Plus

```
Query  329  TGCTGGAACAA  339
            ||||||||||| 
Sbjct  26   TGCTGGAACAA  36
```

>lcl|52975 NODE_159_length_21_cov_2.000000

```
Length=41

 Score = 59.0 bits (64),  Expect = 2e-13
 Identities = 39/41 (95%), Gaps = 2/41 (4%)
 Strand=Plus/Plus

Query  3837  TGTTTCGGTGCAGGGTTTTG-CCCGT-GGCGAAAGCAATAT  3875
              |||||||||||||||||||| ||||| |||||||||||||
Sbjct  1     TGTTTCGGTGCAGGGTTTTGACCCGTCGGCGAAAGCAATAT  41


>lcl|52976 NODE_273_length_21_cov_3.952381
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  2040  GCCGCGTGAAGTAGAACAGG-CGCTGGAAAGTGCGGTTAAT  2079
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    GCCGCGTGAAGTAGAACAGGACGCTGGAAAGTGCGGTTAAT  1


>lcl|52977 NODE_893_length_21_cov_5.047619
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  3196  TCGGCCCATGTTGGTAACCT-GTATGTTAACCGTAATATGG  3235
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    TCGGCCCATGTTGGTAACCTAGTATGTTAACCGTAATATGG  1


 Score = 26.5 bits (28),  Expect = 0.001
 Identities = 17/19 (89%), Gaps = 0/19 (0%)
 Strand=Plus/Minus



Query  1690  ACCTCGTTTGTTAACCGTA  1708
              |||| || ||||||||||
Sbjct  25    ACCTAGTATGTTAACCGTA  7


>lcl|52978 NODE_1046_length_21_cov_6.666667
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  3185  AGGTCACTGGCTCGGCCCAT-GTTGGTAACCTGTATGTTAA  3224
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    AGGTCACTGGCTCGGCCCATAGTTGGTAACCTGTATGTTAA  1


>lcl|52979 NODE_1086_length_23_cov_93.173912
Length=43

 Score = 53.6 bits (58),  Expect = 9e-12
 Identities = 40/43 (93%), Gaps = 3/43 (6%)
 Strand=Plus/Minus

Query  2583  GTTACTGCAGCGCAA-TATCG-CCAGCCGCCTGGACGCTCAGG  2623
              ||||||||||||||| ||||| |||||||||||| |||||||
Sbjct  42    GTTACTGCAGCGCAACTATCGTCCAGCCGCCTGGA-GCTCAGG  1


>lcl|52980 NODE_1139_length_21_cov_27.714285
Length=41

 Score = 46.4 bits (50),  Expect = 1e-09
 Identities = 29/30 (96%), Gaps = 1/30 (3%)
 Strand=Plus/Minus

Query  1   ATGGGAACC-ACCACCATGGGGGTTAAGCT  29
           ||||||||| ||||||||||||||||||||
Sbjct  30  ATGGGAACCTACCACCATGGGGGTTAAGCT  1
```

```
>lcl|52981 NODE_1229_length_21_cov_3.571429
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  47  AGCGTATTAAGTCTGCCGCG-ACACGTATCGATCGCACACC  86
           |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1   AGCGTATTAAGTCTGCCGCGTACACGTATCGATCGCACACC  41


>lcl|52982 NODE_1278_length_21_cov_4.047619
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  2034  AGCCACGCCGCGTGAAGTAG-AACAGGCGCTGGAAAGTGCG  2073
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1      AGCCACGCCGCGTGAAGTAGTAACAGGCGCTGGAAAGTGCG  41


>lcl|52983 NODE_1407_length_21_cov_4.904762
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  464  TTTCGCTGTCATCGCAGGAA-GGCGTGGCGCTGATGTGTCT  503
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41   TTTCGCTGTCATCGCAGGAACGGCGTGGCGCTGATGTGTCT  1


>lcl|52984 NODE_1540_length_36_cov_27.888889
Length=56

 Score = 42.8 bits (46),  Expect = 2e-08
 Identities = 31/35 (88%), Gaps = 1/35 (2%)
 Strand=Plus/Minus

Query  2325  CAGTCCGTGGAACTTCCCGC-TGGCTATTTTCACC  2358
              ||||||||||||| |   || ||||||||||||||
Sbjct  35     CAGTCCGTGGAACTCCGGGCCTGGCTATTTTCACC  1


 Score = 41.0 bits (44),  Expect = 8e-08
 Identities = 22/22 (100%), Gaps = 0/22 (0%)
 Strand=Plus/Minus

Query  2321  GTATCAGTCCGTGGAACTTCCC  2342
              ||||||||||||||||||||||
Sbjct  56     GTATCAGTCCGTGGAACTTCCC  35


>lcl|52985 NODE_1819_length_25_cov_2.360000
Length=45

 Score = 59.0 bits (64),  Expect = 2e-13
 Identities = 43/46 (93%), Gaps = 3/46 (6%)
 Strand=Plus/Minus

Query  2795  ACACGTTGAAAA-TGCTG-CGCGGCGCAATGGCCGAATGCCGGATG  2838
              |||||||||||| ||||| |||||| ||||||||||||||||||||
Sbjct  45     ACACGTTGAAAAATGCTGACGCGGC-CAATGGCCGAATGCCGGATG  1


 Score = 22.9 bits (24),  Expect = 0.017
 Identities = 14/15 (93%), Gaps = 0/15 (0%)
 Strand=Plus/Minus

Query  173  CTGGCGCGGCCAATG  187
             ||| ||||||||||
Sbjct  30   CTGACGCGGCCAATG  16


>lcl|52986 NODE_1834_length_21_cov_5.523809
Length=41
```

```
 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  1706  GTATTGCCGACACCTCTTTG-CCACTGGATGAACTGGTCGC  1745
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     GTATTGCCGACACCTCTTTGTCCACTGGATGAACTGGTCGC  41


>lcl|52987 NODE_1931_length_21_cov_4.904762
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  2788  GCCGACCACACGTTGAAAAT-GCTGCGCGGCGCAATGGCCG  2827
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    GCCGACCACACGTTGAAAATCGCTGCGCGGCGCAATGGCCG  1


 Score = 26.5 bits (28),  Expect = 0.001
 Identities = 14/14 (100%), Gaps = 0/14 (0%)
 Strand=Plus/Plus

Query  2116  CGCGCAGCGATTTT  2129
             ||||||||||||||
Sbjct  13    CGCGCAGCGATTTT  26


>lcl|52988 NODE_1961_length_21_cov_2.428571
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  3464  AATTGCAGGCGTTATGTACG-CAATATGGCGAGCTGGCGCA  3503
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     AATTGCAGGCGTTATGTACGTCAATATGGCGAGCTGGCGCA  41


>lcl|52989 NODE_1994_length_21_cov_5.000000
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  500  GTCTGGCGGAAGCGTTGTTG-CGTATTCCCGACAAAGCCAC  539
            |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41   GTCTGGCGGAAGCGTTGTTGACGTATTCCCGACAAAGCCAC  1


 Score = 21.1 bits (22),  Expect = 0.052
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Minus

Query  796  GCGGAAGCGTT  806
            ||||||||||
Sbjct  36   GCGGAAGCGTT  26


>lcl|52990 NODE_2005_length_21_cov_4.714286
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  2686  GCACTGACCGAACAGGTCGT-CGTGGATGTACTGGCCTCGG  2725
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    GCACTGACCGAACAGGTCGTACGTGGATGTACTGGCCTCGG  1


>lcl|52991 NODE_2082_length_21_cov_3.238095
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
```

```
 Strand=Plus/Plus

Query  277  GCCGCGATCACCGCGGCCTA-TCGCCGCCCGGAAACCGAAG  316
            |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1    GCCGCGATCACCGCGGCCTACTCGCCGCCCGGAAACCGAAG  41


>lcl|52992 NODE_2191_length_21_cov_5.095238
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  550  TTAATTCGCGACAAAATCAG-CAACGGTAACTGGCAGTCAC  589
            |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1    TTAATTCGCGACAAAATCAGTCAACGGTAACTGGCAGTCAC  41


>lcl|52993 NODE_2199_length_21_cov_7.380952
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  2145  GCTGATGGAAAGCCAGATGC-AGCAACTGATTGGTATTCTG  2184
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    GCTGATGGAAAGCCAGATGCTAGCAACTGATTGGTATTCTG  1


>lcl|52994 NODE_2212_length_21_cov_4.238095
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  835  GGTTTCCGTTACTCTTACGA-TATGCTGGGCGAAGCCGCGC  874
            |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1    GGTTTCCGTTACTCTTACGAGTATGCTGGGCGAAGCCGCGC  41


>lcl|52995 NODE_2217_length_21_cov_5.904762
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  2461  GCGGGTGTACCGCCAGGCGT-GGTGCAATTGCTGCCAGGTC  2500
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    GCGGGTGTACCGCCAGGCGTAGGTGCAATTGCTGCCAGGTC  1


>lcl|52996 NODE_2236_length_21_cov_6.190476
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  1933  TGGCAGGCCTTGCCAATGCT-GGAACAACCGGTAGCGGCAG  1972
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    TGGCAGGCCTTGCCAATGCTCGGAACAACCGGTAGCGGCAG  1


>lcl|52997 NODE_2238_length_21_cov_2.857143
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  277  GCCGCGATCACCGCGGCCTA-TCGCCGCCCGGAAACCGAAG  316
            |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1    GCCGCGATCACCGCGGCCTAGTCGCCGCCCGGAAACCGAAG  41


>lcl|52998 NODE_2257_length_21_cov_4.714286
Length=41
```

```
 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  3536  CGGGTGAACGCAACACCTGG-ACGCTGCTGCCGCGTGAGCG  3575
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    CGGGTGAACGCAACACCTGGTACGCTGCTGCCGCGTGAGCG  1


>lcl|52999 NODE_2258_length_21_cov_4.523809
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  3467  TGCAGGCGTTATGTACGCAA-TATGGCGAGCTGGCGCAGGC  3506
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    TGCAGGCGTTATGTACGCAACTATGGCGAGCTGGCGCAGGC  1


>lcl|53000 NODE_2275_length_21_cov_3.095238
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  3862  GGCGAAAGCAATATCCTTCT-GGAACGGCTGTATATCGAGC  3901
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     GGCGAAAGCAATATCCTTCTAGGAACGGCTGTATATCGAGC  41


>lcl|53001 NODE_2294_length_21_cov_3.428571
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  873  GCTGACCGCCGCAGATGCAC-AGGCGTATATGGTTTCCTAT  912
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41   GCTGACCGCCGCAGATGCACGAGGCGTATATGGTTTCCTAT  1


>lcl|53002 NODE_2316_length_21_cov_4.714286
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  3458  GTCCAGAATTGCAGGCGTTA-TGTACGCAATATGGCGAGCT  3497
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    GTCCAGAATTGCAGGCGTTAGTGTACGCAATATGGCGAGCT  1


>lcl|53003 NODE_2332_length_21_cov_2.619048
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  3534  GACGGGTGAACGCAACACCT-GGACGCTGCTGCCGCGTGAG  3573
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     GACGGGTGAACGCAACACCTCGGACGCTGCTGCCGCGTGAG  41


>lcl|53004 NODE_2338_length_21_cov_4.857143
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  2805  AATGCTGCGCGGCGCAATGG-CCGAATGCCGGATGGGTAAT  2844
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    AATGCTGCGCGGCGCAATGGACCGAATGCCGGATGGGTAAT  1
```

```
>lcl|53005 NODE_2351_length_21_cov_4.761905
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  1820  TGCCGCGCGATCTTTACGGT-CACGGGCGCGACAACTCGGC  1859
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    TGCCGCGCGATCTTTACGGTACACGGGCGCGACAACTCGGC  1


>lcl|53006 NODE_2355_length_21_cov_7.142857
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  3743  TAACCGCTCAACCGTTTGAT-GCGGTGATCTTCCACGGTGA  3782
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     TAACCGCTCAACCGTTTGATAGCGGTGATCTTCCACGGTGA  41


>lcl|53007 NODE_2369_length_21_cov_3.428571
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  828  AGAGAAAGGTTTCCGTTACT-CTTACGATATGCTGGGCGAA  867
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1    AGAGAAAGGTTTCCGTTACTACTTACGATATGCTGGGCGAA  41


>lcl|53008 NODE_2370_length_21_cov_7.190476
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  3319  TACCGTCTGCTGGCGAATCG-CCCGGAAAGTGCGCTGGCAG  3358
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     TACCGTCTGCTGGCGAATCGACCCGGAAAGTGCGCTGGCAG  41


>lcl|53009 NODE_2371_length_21_cov_4.190476
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  3476  TATGTACGCAATATGGCGAG-CTGGCGCAGGCAGGAACACA  3515
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     TATGTACGCAATATGGCGAGTCTGGCGCAGGCAGGAACACA  41


>lcl|53010 NODE_2377_length_21_cov_2.190476
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  534  AGCCACCCGCGACGCGTTAA-TTCGCGACAAAATCAGCAAC  573
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1    AGCCACCCGCGACGCGTTAACTTCGCGACAAAATCAGCAAC  41


>lcl|53011 NODE_2396_length_21_cov_2.714286
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus
```

```
Query  2998  ACCTTTGTCGCCCCGACGCT-GATCGAACTGGATGACTTTG  3037
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    ACCTTTGTCGCCCCGACGCTCGATCGAACTGGATGACTTTG  1


>lcl|53012 NODE_2409_length_21_cov_5.142857
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  3081  GCGTTACAACCGTAACCAGC-TACCAGAGCTGATCGAGCAG  3120
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     GCGTTACAACCGTAACCAGCGTACCAGAGCTGATCGAGCAG  41


>lcl|53013 NODE_2417_length_21_cov_2.238095
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus


Query  1971  AGGTGAGATGTCGCCCGTTA-TTAACCCTGCGGAACCGAAA  2010
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     AGGTGAGATGTCGCCCGTTAGTTAACCCTGCGGAACCGAAA  41


>lcl|53014 NODE_2418_length_21_cov_6.190476
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  3080  TGCGTTACAACCGTAACCAG-CTACCAGAGCTGATCGAGCA  3119
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     TGCGTTACAACCGTAACCAGTCTACCAGAGCTGATCGAGCA  41


>lcl|53015 NODE_2453_length_21_cov_6.571429
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  764   TGGGTGAGCAGTTCGTCACT-GGCGAAACCATCGCGGAAGC  803
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    TGGGTGAGCAGTTCGTCACTAGGCGAAACCATCGCGGAAGC  1


 Score = 21.1 bits (22),  Expect = 0.052
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Minus

Query  3172  GAAACCATCGC  3182
              ||||||||||| 
Sbjct  17    GAAACCATCGC  7


>lcl|53016 NODE_2516_length_21_cov_3.047619
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  1600  AACCGTCCGTGTCGTATTTA-TGCTCCGGTTGGCACACATG  1639
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     AACCGTCCGTGTCGTATTTACTGCTCCGGTTGGCACACATG  41


>lcl|53017 NODE_2544_length_21_cov_3.000000
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus
```

```
Query  2349  TATTTTCACCGGGCAGATCG-CCGCCGCACTGGCGGCAGGT  2388
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     TATTTTCACCGGGCAGATCGTCCGCCGCACTGGCGGCAGGT  41


>lcl|53018 NODE_2551_length_21_cov_4.333333
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  724  CCGCTGATCCGCAAAGGTGT-GGATATGGCGATGCGCCTGA  763
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1    CCGCTGATCCGCAAAGGTGTCGGATATGGCGATGCGCCTGA  41


>lcl|53019 NODE_2559_length_21_cov_5.190476
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  1239  GATCGATTACCTGATTGATC-TCGCCACCCGCAGCCGTCGC  1278
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     GATCGATTACCTGATTGATCGTCGCCACCCGCAGCCGTCGC  41


>lcl|53020 NODE_2566_length_21_cov_6.000000
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  318  GGTTTCTATGCTGCTGGAAC-AAGCCCGCCTGCCGCAGCCA  357
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41   GGTTTCTATGCTGCTGGAACTAAGCCCGCCTGCCGCAGCCA  1


>lcl|53021 NODE_2633_length_21_cov_3.571429
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  1195  ATCGGTTTTGTTATTCAGGC-TTATCAAAAACGCTGCCCGT  1234
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     ATCGGTTTTGTTATTCAGGCGTTATCAAAAACGCTGCCCGT  41


>lcl|53022 NODE_2655_length_21_cov_2.666667
Length=41

 Score = 59.0 bits (64),  Expect = 2e-13
 Identities = 36/37 (97%), Gaps = 1/37 (2%)
 Strand=Plus/Plus

Query  2820  AATGGCCGAATGCCGG-ATGGGTAATCCGGGTCGCCT  2855
              ||||||||||||||| ||||||||||||||||||||
Sbjct  5     AATGGCCGAATGCCGGTATGGGTAATCCGGGTCGCCT  41


>lcl|53023 NODE_2660_length_21_cov_6.238095
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  2466  TGTACCGCCAGGCGTGGTGC-AATTGCTGCCAGGTCGGGGT  2505
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    TGTACCGCCAGGCGTGGTGCGAATTGCTGCCAGGTCGGGGT  1


>lcl|53024 NODE_2695_length_21_cov_3.619048
Length=41
```

```
 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  2458  GAAGCGGGTGTACCGCCAGG-CGTGGTGCAATTGCTGCCAG  2497
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     GAAGCGGGTGTACCGCCAGGTCGTGGTGCAATTGCTGCCAG  41


>lcl|53025 NODE_2717_length_21_cov_3.238095
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  999   GCATCCGCGTTATAGCCGCG-CCCAGTATGACCGGGTAATG  1038
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     GCATCCGCGTTATAGCCGCGACCCAGTATGACCGGGTAATG  41


>lcl|53026 NODE_2718_length_21_cov_2.857143
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  204  GACTCCGGCAGAGGAACCAC-ACCAGCCATTCCTCGACTTT  243
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1    GACTCCGGCAGAGGAACCACGACCAGCCATTCCTCGACTTT  41


 Score = 21.1 bits (22),  Expect = 0.052
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Minus

Query  146  CTCTGCCGGAG  156
             ||||||||||| 
Sbjct  13   CTCTGCCGGAG  3


>lcl|53027 NODE_2773_length_21_cov_8.380953
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  1339  ATGGACGGCCTTGAAGGTTA-TCCGGTTTATACCCGCAAGG  1378
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    ATGGACGGCCTTGAAGGTTAGTCCGGTTTATACCCGCAAGG  1


>lcl|53028 NODE_2791_length_21_cov_3.523809
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  540  CCGCGACGCGTTAATTCGCG-ACAAAATCAGCAACGGTAAC  579
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1    CCGCGACGCGTTAATTCGCGTACAAAATCAGCAACGGTAAC  41


>lcl|53029 NODE_2808_length_21_cov_4.047619
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  1723  TTGCCACTGGATGAACTGGT-CGCCGATCCGGTCACTGCTG  1762
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     TTGCCACTGGATGAACTGGTACGCCGATCCGGTCACTGCTG  41


>lcl|53030 NODE_2822_length_21_cov_8.142858
Length=41
```

```
 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  1596  ACTTAACCGTCCGTGTCGTA-TTTATGCTCCGGTTGGCACA  1635
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1      ACTTAACCGTCCGTGTCGTAGTTTATGCTCCGGTTGGCACA  41


>lcl|53031 NODE_2832_length_21_cov_8.380953
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  935  GTAAAGCGTCTAACGGTCGT-GGCATCTATGAAGGGCCGGG  974
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1    GTAAAGCGTCTAACGGTCGTCGGCATCTATGAAGGGCCGGG  41


>lcl|53032 NODE_2898_length_21_cov_3.190476
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  2175  TGGTATTCTGGTGCGTGAGG-CCGGAAAAACCTTCAGTAAC  2214
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    TGGTATTCTGGTGCGTGAGGTCCGGAAAAACCTTCAGTAAC  1


>lcl|53033 NODE_2899_length_21_cov_2.428571
Length=41

 Score = 59.0 bits (64),  Expect = 2e-13
 Identities = 39/41 (95%), Gaps = 2/41 (4%)
 Strand=Plus/Minus

Query  2131  CACCGCGCTGCCGTG-CTGA-TGGAAAGCCAGATGCAGCAA  2169
              ||||||||||||||| |||| ||||||||||||||||||||
Sbjct  41    CACCGCGCTGCCGTGACTGACTGGAAAGCCAGATGCAGCAA  1


>lcl|53034 NODE_2906_length_21_cov_4.714286
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  72   TATCGATCGCACACCACACT-GGTTAATTAAGCAGGCGATT  111
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41   TATCGATCGCACACCACACTAGGTTAATTAAGCAGGCGATT  1


 Score = 21.1 bits (22),  Expect = 0.052
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Minus

Query  914  AGCAGGCGATT  924
             |||||||||||
Sbjct  11   AGCAGGCGATT  1


>lcl|53035 NODE_2922_length_21_cov_4.476191
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  1348  CTTGAAGGTTATCCGGTTTA-TACCCGCAAGGTGTATACCG  1387
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    CTTGAAGGTTATCCGGTTTACTACCCGCAAGGTGTATACCG  1


>lcl|53036 NODE_2940_length_21_cov_3.619048
Length=41
```

```
 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  1310  ACTGGGATAGTGAAATTAAG-CGTGCGCAGATGGACGGCCT  1349
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     ACTGGGATAGTGAAATTAAGACGTGCGCAGATGGACGGCCT  41


>lcl|53037 NODE_2949_length_21_cov_8.047619
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  1332  TGCGCAGATGGACGGCCTTG-AAGGTTATCCGGTTTATACC  1371
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    TGCGCAGATGGACGGCCTTGTAAGGTTATCCGGTTTATACC  1


>lcl|53038 NODE_2981_length_21_cov_3.428571
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  2888  AAGCGAAAGCCAATATTGAG-CGCCATATTCAGACCATGCG  2927
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     AAGCGAAAGCCAATATTGAGTCGCCATATTCAGACCATGCG  41


>lcl|53039 NODE_3005_length_21_cov_2.428571
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  3512  CACAACGATTGCTGCCGGGG-CCGACGGGTGAACGCAACAC  3551
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     CACAACGATTGCTGCCGGGGTCCGACGGGTGAACGCAACAC  41


>lcl|53040 NODE_3009_length_21_cov_4.476191
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  2471  CGCCAGGCGTGGTGCAATTG-CTGCCAGGTCGGGGTGAAAC  2510
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     CGCCAGGCGTGGTGCAATTGACTGCCAGGTCGGGGTGAAAC  41


>lcl|53041 NODE_3023_length_21_cov_4.142857
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  2241  CGATTTTCTCCACTACTACG-CCGGACAGGTGCGGGATGAT  2280
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     CGATTTTCTCCACTACTACGACCGGACAGGTGCGGGATGAT  41


>lcl|53042 NODE_3058_length_21_cov_5.952381
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  2915  TTCAGACCATGCGTAGCAAA-GGCCGTCCGGTGTTCCAGGC  2954
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    TTCAGACCATGCGTAGCAAACGGCCGTCCGGTGTTCCAGGC  1
```

```
>lcl|53043 NODE_3067_length_21_cov_3.809524
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  775  TTCGTCACTGGCGAAACCAT-CGCGGAAGCGTTAGCCAATG  814
            |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41   TTCGTCACTGGCGAAACCATACGCGGAAGCGTTAGCCAATG  1


 Score = 26.5 bits (28),  Expect = 0.001
 Identities = 17/19 (89%), Gaps = 0/19 (0%)
 Strand=Plus/Plus

Query  3123  TAACGCTTCCGGTTATGGT  3141
              |||||||||||  ||||||
Sbjct  8     TAACGCTTCCGCGTATGGT  26


 Score = 21.1 bits (22),  Expect = 0.052
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Minus

Query  505  GCGGAAGCGTT  515
            ||||||||||
Sbjct  19   GCGGAAGCGTT  9


>lcl|53044 NODE_3101_length_21_cov_5.047619
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  3862  GGCGAAAGCAATATCCTTCT-GGAACGGCTGTATATCGAGC  3901
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    GGCGAAAGCAATATCCTTCTCGGAACGGCTGTATATCGAGC  1


>lcl|53045 NODE_3119_length_21_cov_6.142857
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  356  CAGTTGCTGAACAGGCGCAC-AAACTGGCGTATCAGCTGGC  395
            |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41   CAGTTGCTGAACAGGCGCACGAAACTGGCGTATCAGCTGGC  1


>lcl|53046 NODE_3122_length_36_cov_2.083333
Length=56

 Score = 77.0 bits (84),  Expect = 1e-18
 Identities = 53/56 (94%), Gaps = 3/56 (5%)
 Strand=Plus/Plus

Query  966   AGGGCCGGGCATTTCAA-TC-AAACTGTCGGCGCT-GCATCCGCGTTATAGCCGCG  1018
              ||||||||||||||||| || ||||||||||||| ||||||||||||||||||||
Sbjct  1     AGGGCCGGGCATTTCAAATCGAAACTGTCGGCGCTCGCATCCGCGTTATAGCCGCG  56


>lcl|53047 NODE_3153_length_21_cov_6.190476
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  2352  TTTCACCGGGCAGATCGCCG-CCGCACTGGCGGCAGGTAAC  2391
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    TTTCACCGGGCAGATCGCCGTCCGCACTGGCGGCAGGTAAC  1


>lcl|53048 NODE_3162_length_21_cov_3.285714
Length=41
```

```
 Score = 59.0 bits (64),  Expect = 2e-13
 Identities = 39/41 (95%), Gaps = 2/41 (4%)
 Strand=Plus/Minus

Query  2266  CAGGTGCGGGATGATTT-CG-CTAACGAAACCCACCGTCCA  2304
             |||||||||||||||| || ||||||||||||||||||||
Sbjct  41    CAGGTGCGGGATGATTTTCGACTAACGAAACCCACCGTCCA  1


>lcl|53049 NODE_3168_length_21_cov_4.285714
Length=41

 Score = 59.0 bits (64),  Expect = 2e-13
 Identities = 39/41 (95%), Gaps = 2/41 (4%)
 Strand=Plus/Plus

Query  609  ACTGTTTGTTAATG-CCGCC-ACCTGGGGGCTGCTGTTTAC  647
             ||||||||||||| ||||| ||||||||||||||||||||
Sbjct  1    ACTGTTTGTTAATGACCGCCGACCTGGGGGCTGCTGTTTAC  41


>lcl|53050 NODE_3170_length_21_cov_2.380952
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  739  GGTGTGGATATGGCGATGCG-CCTGATGGGTGAGCAGTTCG  778
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1    GGTGTGGATATGGCGATGCGACCTGATGGGTGAGCAGTTCG  41


>lcl|53051 NODE_3181_length_21_cov_2.619048
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  3732  AGCGGAAAATATAACCGCTC-AACCGTTTGATGCGGTGATC  3771
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    AGCGGAAAATATAACCGCTCTAACCGTTTGATGCGGTGATC  1


>lcl|53052 NODE_3182_length_21_cov_3.952381
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  3093  TAACCAGCTACCAGAGCTGA-TCGAGCAGATTAACGCTTCC  3132
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    TAACCAGCTACCAGAGCTGAGTCGAGCAGATTAACGCTTCC  1


>lcl|53053 NODE_3234_length_21_cov_2.476191
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  2777  AAGATGAGATTGCCGACCAC-ACGTTGAAAATGCTGCGCGG  2816
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    AAGATGAGATTGCCGACCACTACGTTGAAAATGCTGCGCGG  1


>lcl|53054 NODE_3245_length_21_cov_5.238095
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  1602  CCGTCCGTGTCGTATTTATG-CTCCGGTTGGCACACATGAA  1641
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     CCGTCCGTGTCGTATTTATGACTCCGGTTGGCACACATGAA  41
```

```
>lcl|53055 NODE_3262_length_21_cov_4.714286
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  1127  GCCTGGAGATCTCCCTCGAT-CTGCTGGAAAAACTCTGTTT  1166
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1      GCCTGGAGATCTCCCTCGATACTGCTGGAAAAACTCTGTTT  41


 Score = 21.1 bits (22),  Expect = 0.052
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Plus

Query  1669  CTGCTGGAAAA  1679
              ||||||||||| 
Sbjct  22     CTGCTGGAAAA  32


>lcl|53056 NODE_3264_length_21_cov_6.428571
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  664  ACCCATAACGAAGCCAGCCT-CTCCCGCTCGCTGAACCGCA  703
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41   ACCCATAACGAAGCCAGCCTACTCCCGCTCGCTGAACCGCA  1


>lcl|53057 NODE_3273_length_21_cov_3.190476
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  3844  GTGCAGGGTTTTGCCCGTGG-CGAAAGCAATATCCTTCTGG  3883
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    GTGCAGGGTTTTGCCCGTGGACGAAAGCAATATCCTTCTGG  1


>lcl|53058 NODE_3276_length_21_cov_5.428571
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  3068  TGCTGCATGTGGTGCGTTAC-AACCGTAACCAGCTACCAGA  3107
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1      TGCTGCATGTGGTGCGTTACGAACCGTAACCAGCTACCAGA  41


>lcl|53059 NODE_3279_length_21_cov_3.571429
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  3279  GTCCGGTACCGGGCCGAAAG-CAGGCGGTCCGCTCTATCTC  3318
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1      GTCCGGTACCGGGCCGAAAGTCAGGCGGTCCGCTCTATCTC  41


>lcl|53060 NODE_3294_length_21_cov_4.761905
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  3348  TGCGCTGGCAGTGACGCTCG-CGCGTCAGGATGCAAAGTAT  3387
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1      TGCGCTGGCAGTGACGCTCGTCGCGTCAGGATGCAAAGTAT  41
```

```
>lcl|53061 NODE_3301_length_21_cov_3.619048
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  2659  ATGAACGCGATGATTGTCGA-TTCTTCAGCACTGACCGAAC  2698
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1      ATGAACGCGATGATTGTCGACTTCTTCAGCACTGACCGAAC  41


>lcl|53062 NODE_3302_length_21_cov_3.761905
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  2741  GTCAGCGTTGTTCGGCGCTG-CGCGTGCTGTGCCTGCAAGA  2780
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1      GTCAGCGTTGTTCGGCGCTGTCGCGTGCTGTGCCTGCAAGA  41


>lcl|53063 NODE_3312_length_21_cov_4.000000
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  632  GGGGGCTGCTGTTTACTGGC-AAACTGGTTTCCACCCATAA  671
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1    GGGGGCTGCTGTTTACTGGCTAAACTGGTTTCCACCCATAA  41


>lcl|53064 NODE_3314_length_21_cov_2.000000
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  3893  ATATCGAGCGTTCGCTGAGT-GTGAATACCGCTGCCGCTGG  3932
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1      ATATCGAGCGTTCGCTGAGTCGTGAATACCGCTGCCGCTGG  41


>lcl|53065 NODE_3317_length_21_cov_2.809524
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  3133  GGTTATGGTCTGACGCTTGG-CGTCCATACGCGCATTGATG  3172
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1      GGTTATGGTCTGACGCTTGGACGTCCATACGCGCATTGATG  41


>lcl|53066 NODE_3330_length_21_cov_3.238095
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  2277  TGATTTCGCTAACGAAACCC-ACCGTCCATTAGGGCCTGTG  2316
              |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1      TGATTTCGCTAACGAAACCCGACCGTCCATTAGGGCCTGTG  41


 Score = 22.9 bits (24),  Expect = 0.015
 Identities = 14/15 (93%), Gaps = 0/15 (0%)
 Strand=Plus/Plus

Query  1867  GATCTCGCTAACGAA  1881
              ||| ||||||||||
Sbjct  2      GATTTCGCTAACGAA  16
```

```
>lcl|53067 NODE_3347_length_21_cov_3.333333
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  2294  CCCACCGTCCATTAGGGCCT-GTGGTGTGTATCAGTCCGTG  2333
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     CCCACCGTCCATTAGGGCCTAGTGGTGTGTATCAGTCCGTG  41


>lcl|53068 NODE_3370_length_21_cov_3.142857
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  3798  CGCATTGTGTGAAGCAGTTG-CCGCGCGGGATGGCACAATT  3837
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     CGCATTGTGTGAAGCAGTTGACCGCGCGGGATGGCACAATT  41


>lcl|53069 NODE_3387_length_21_cov_2.714286
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus


Query  991   TCGGCGCTGCATCCGCGTTA-TAGCCGCGCCCAGTATGACC  1030
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    TCGGCGCTGCATCCGCGTTAGTAGCCGCGCCCAGTATGACC  1


>lcl|53070 NODE_3404_length_21_cov_5.476191
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  1692  CTCGTTTGTTAACCGTATTG-CCGACACCTCTTTGCCACTG  1731
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  1     CTCGTTTGTTAACCGTATTGTCCGACACCTCTTTGCCACTG  41


 Score = 21.1 bits (22),  Expect = 0.052
 Identities = 13/14 (92%), Gaps = 0/14 (0%)
 Strand=Plus/Plus

Query  3216  GTATGTTAACCGTA  3229
             || ||||||||||
Sbjct  4     GTTTGTTAACCGTA  17

>lcl|53071 NODE_3428_length_21_cov_5.523809
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Minus

Query  3196  TCGGCCCATGTTGGTAACCT-GTATGTTAACCGTAATATGG  3235
             |||||||||||||||||||| ||||||||||||||||||||
Sbjct  41    TCGGCCCATGTTGGTAACCTCGTATGTTAACCGTAATATGG  1

 Score = 30.1 bits (32),  Expect = 1e-04
 Identities = 18/19 (94%), Gaps = 0/19 (0%)
 Strand=Plus/Minus

Query  1690  ACCTCGTTTGTTAACCGTA  1708
             ||||||| ||||||||||
Sbjct  25    ACCTCGTATGTTAACCGTA  7


>lcl|53072 NODE_3470_length_21_cov_2.571429
Length=41
```

```
 Score = 59.0 bits (64),  Expect = 2e-13
 Identities = 39/41 (95%), Gaps = 2/41 (4%)
 Strand=Plus/Minus

Query  3649  GTAC-TGTGGCCGGATGACG-CGCTGCATCGTCAGTTAGTG  3687
              |||| ||||||||||||||| ||||||||||||||||||
Sbjct  41    GTACGTGTGGCCGGATGACGACGCTGCATCGTCAGTTAGTG  1

>lcl|53073 NODE_3488_length_21_cov_2.904762
Length=41

 Score = 66.2 bits (72),  Expect = 1e-15
 Identities = 40/41 (97%), Gaps = 1/41 (2%)
 Strand=Plus/Plus

Query  407  GTAATCnnnnnnnTGCCAGT-GGTCGCGCAGGTATGGTCCA  446
             ||||||||||||||||||| ||||||||||||||||||||
Sbjct  1    GTAATCAAAAAAATGCCAGTAGGTCGCGCAGGTATGGTCCA  41
```

APPENDIX D

BLAST REPORT 454 ASSEMBLED k = 31 EXPECTED COVERAGE = 24

COVERAGE CUTOFF = 12 VS. REFERENCE

```
 velveth out_NP_415534-454_31 31 -fasta -shortPaired NP_415534-454.fasta
 velvetg out_NP_415534-454_31_12_24_dir -exp_cov 24 -cov_cutoff 12 -read_trkg yes -
amos_file yes -unused_reads yes
Final graph has 22 nodes and n50 of 267, max 461, total 3130, using 5000/5000 reads

BLASTN 2.2.23+
Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro
A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and
David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new
generation of protein database search programs", Nucleic
Acids Res. 25:3389-3402.


RID: U2ZPZ325112


Query=  eco:b1014 putA, ECK1005, JW0999, poaA, putC; fused DNA-binding
transcriptional regulator/proline
dehydrogenase/pyrroline-5-carboxylate dehydrogenase (EC:1.5.99.8
1.5.1.12); K00294 1-pyrroline-5-carboxylate dehydrogenase
[EC:1.5.1.12] K00318 proline dehydrogenase [EC:1.5.99.8] (N)
Length=3963


                                                                      Score     E
Sequences producing significant alignments:                           (Bits)  Value

lcl|63088  NODE_1_length_393_cov_219.491089                            623    0.0
lcl|63089  NODE_2_length_282_cov_195.992905                            484    2e-140
lcl|63090  NODE_3_length_267_cov_206.041199                            428    1e-123
lcl|63091  NODE_4_length_206_cov_205.684464                            340    3e-97
lcl|63092  NODE_5_length_158_cov_201.348099                            293    3e-83
lcl|63093  NODE_6_length_291_cov_199.041245                            524    2e-152
lcl|63094  NODE_7_length_461_cov_209.563995                            690    0.0
lcl|63095  NODE_8_length_243_cov_213.312759                            416    6e-120
lcl|63096  NODE_9_length_98_cov_212.948975                             197    1e-54
lcl|63097  NODE_10_length_148_cov_229.418915                           300    2e-85
lcl|63098  NODE_11_length_76_cov_221.750000                            156    3e-42
lcl|63099  NODE_12_length_62_cov_208.145157                            129    4e-34
lcl|63100  NODE_13_length_73_cov_221.095886                            187    2e-51
lcl|63101  NODE_15_length_142_cov_198.697189                           239    6e-67
lcl|63102  NODE_16_length_65_cov_216.415390                            114    8e-30
lcl|63103  NODE_17_length_101_cov_213.732666                           185    8e-51

ALIGNMENTS
>lcl|63088 NODE_1_length_393_cov_219.491089
Length=423

 Score =  623 bits (690),  Expect = 0.0
 Identities = 410/434 (94%), Gaps = 18/434 (4%)
 Strand=Plus/Plus

Query  1756  ACTGCTGTAGAAAAACTGGCGCAACAGGAAGGG-CAAACTGGATTACCGCATCCGAAAAT  1814
              |||||||   |||||||||||| |||||||||||  ||||   |||||||||||| ||||||
Sbjct  1      ACTGCTGAGAAAAAACTGGCG-AACAGGAAGGGGCAAAACTGATTACCGCATC-GAAAAT  58


Query  1815  TCCCCTGCCGCGCGATCTTT-ACGGTCACGGGCGCGACAACTCGGCAGGGCTGGATCTCG  1873
              ||||||||||||||||||| |||||||||| ||||||||||||||||||||||||||||
Sbjct  59    TCCCCTGCCGCGCGATCTTTTACGGTCACGG-CGCGACAACTCGGCAGGGCTGGATCTCG  117

Query  1874  CTAACGAACACCGCCTGGCCTCGCTCTCCTCTGCCCTGCTCAATAGTGCACTGCAAAAAT  1933
              |||||||||||||||||||||||||||||||||||  |||||||||||||||||||||||
Sbjct  118   CTAACGAACACCGCCTGGCCTCGCTCTCCTCTGCCC--CTCAATAGTGCACTGCAAAAAT  175

Query  1934  GG-CAGGCCTTGCCAATGCTGGAACAACCGGTAGCGGCAGGTGAGATGTCGCCCGTTATT  1992
              || |||||||||||||||||||||||||| ||||||||| ||||||||||  ||||||
Sbjct  176   GGGCAGGCCTTGCCAATGCTGGAACAACCG-TAGCGGCAG-TGAGATGTCGC--GTTATT  231
```

```
Query  1993  AACCCTGCGGAACCG-AAAGATATTGTGGGCTATGTGCGTGAAGCCACGCCGCGTGAAGT  2051
             ||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||
Sbjct  232   AACCCTGCGGAACCGGAAAGATATTGTGGGCTATGTGCGTGAAGCCACGCCGCGTGAAGT  291

Query  2052  AGAACAGGCGCTGGAAAGTGCGGTTAATAACGCGCCAATCTGGTTTGCCACGCCTCCGGC  2111
             ||||||||||||||||||||||||||||||||||| ||||||||||||||||||||||||
Sbjct  292   AGAACAGGCGCTGGAAAGTGCGGTTAATAACGCG-CAATCTGGTTTGCCACGCCTCCGGC  350

Query  2112  TGAACGCGCAGCGATTTTGCACCGCGCTGCCGTGCTGATGGAAAGCCAGATGCAGCAACT  2171
             ||||||||| ||||||||||||||||||| ||||||||||||||||||||||||||||||
Sbjct  351   TGAACGCG-AGCGATTTTGCACCGCGCT-CCGTGCTGATGGAAAGCCAGATGCAGCAACT  408

Query  2172  GATTGGTATTCTGG  2185
             |||||  |||||||
Sbjct  409   GATTG--ATTCTGG  420


 Score = 24.7 bits (26),  Expect = 0.058
 Identities = 15/16 (93%), Gaps = 0/16 (0%)
 Strand=Plus/Minus

Query  2073  GGTTAATAACGCGCCA  2088
             |||||||||||||  ||
Sbjct  235   GGTTAATAACGCGACA  220


 Score = 22.9 bits (24),  Expect = 0.20
 Identities = 14/15 (93%), Gaps = 0/15 (0%)
 Strand=Plus/Plus

Query  2278  GATTTCGCTAACGAA  2292
             ||| ||||||||||||
Sbjct  111   GATCTCGCTAACGAA  125


 Score = 21.1 bits (22),  Expect = 0.71
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Plus

Query  329  TGCTGGAACAA  339
             ||||||||||||
Sbjct  192  TGCTGGAACAA  202


 Score = 21.1 bits (22),  Expect = 0.71
 Identities = 13/14 (92%), Gaps = 0/14 (0%)
 Strand=Plus/Minus

Query  1983  GCCCGTTATTAACC  1996
             || ||||||||||||
Sbjct  326   GCGCGTTATTAACC  313


>lcl|63089 NODE_2_length_282_cov_195.992905
Length=312

 Score =  484 bits (536),  Expect = 2e-140
 Identities = 305/317 (96%), Gaps = 12/317 (3%)
 Strand=Plus/Plus

Query  2174  TTGGTATTCTGGTGCGTGAGGCC-GGAAAAACCTTCAGTAACGCCATTGCCGAAGTG-CG  2231
             |||||||||||||||||||||||| |||||||||||||||||||||||||||||||| ||
Sbjct  1     TTGGTATTCTGGTGCGTGAGGCCCGGAAAAACCTTCAGTAACGCCATTGCCGAAGTGGCG  60

Query  2232  CGAAGCGGTCGATTTTCTCCACTACTACGCCGGACAGGTGCGGGATGATTTCGCTAACGA  2291
             ||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    CGAAGCG-TCGATTTTCTCCACTACTACGCCGGACAGGTGCGGGATGATTTCGCTAACGA  119

Query  2292  AACCCACCGTCCATTAGGG-CCTGTGGTGTGTATCAGTCCGTGGAACTTCCCGCTGGCTA  2350
             |||||||| ||||||||| |||||| ||||||||||||||||||||||||||||||||||
Sbjct  120   AACCCACCG-CCATTAGGGGCCTGTG-TGTGTATCAGTCCGTGGAACTTCCCGCTGGCTA  177

Query  2351  TTTTCACCGGGCAGATCGCCGCCGCACTGGCGGCAGGTAACA-GCGTGCTGGCAAAACCG  2409
             |||||||||||||||||||||||||||  |||||||||| ||||||||||||||||||
Sbjct  178   TTTTCACCGGGCAGATCGCCGCCGCACTG--GGCAGGTAACAGGCGTGCTGGCAAAACCG  235

Query  2410  GCAGAACAAACGCCGCTGATTGCCGCGCAA-GGGATCGCCATTTTGCTGGAAGCGGGTGT  2468
             |||||||||||||||||||||||||||||| |||||||||||||||||||||||||||||
Sbjct  236   GCAGAACAAACGCCGCTGATTGCCGCGCAAGGGGATCGCCATTTTGCTGGAAGCGGGTGT  295
```

```
Query  2469  ACCGCC--AGGCGTGGT  2483
             ||||||  |||||||||
Sbjct  296   ACCGCCAGAGGCGTGGT  312


 Score = 26.5 bits (28),  Expect = 0.012
 Identities = 21/25 (84%), Gaps = 3/25 (12%)
 Strand=Plus/Plus

Query  1867  GATCTCGCTAACGAA---CACCGCC  1888
             ||| ||||||||||   |||||||
Sbjct  106   GATTTCGCTAACGAAACCCACCGCC  130


>lcl|63090 NODE_3_length_267_cov_206.041199
Length=297

 Score =  428 bits (474),  Expect = 1e-123
 Identities = 274/288 (95%), Gaps = 13/288 (4%)
 Strand=Plus/Plus

Query  3681  GTTAGTGAAGGCATTGCCATCGGCAGTCAGCGAACGTATTCAACTGGCGAAAGCGGAAAA  3740
             |||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct  2     GTTAGTGAAGGCATTTCCATCGGCAGTCAGCGAACGTATTCAACTGGCGAAAGCGGAAAA  61

Query  3741  T-ATAACCGCTCAACCG-TTTGATGCGGTGATCTTCCACGGTGATTCGGATCAGCTTCGC  3798
             | |||||||||||||| |||||||||||||||||||||||||||||| ||||||||
Sbjct  62    TAATAACCGCTCAACCGTTTTGATGCGGTGATCTTCCACGGTGATTCGGAT-AGCTTCGC  120

Query  3799  GCATTGTGTGAAGCAGTTGCCGCGCGGGATGGCACAATTGTTTCGGTGCAGGGTTTTGCC  3858
             ||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||
Sbjct  121   GCATTGTGTGAAGCAGTTG-CGCGCGGGATGGCACAATTGTTTCGGTGCAGGGTTTTGCC  179

Query  3859  CGTGGCGAAAGCAATATCCTTCTGGA--ACGGCTGTATATCG-AGCGTTCGCTGAGTGTG  3915
             |||||||  |||||||| |||||||  |||||||||||||| ||||||||||||||||||
Sbjct  180   CGTGGCG--AGCAATAT--TTCTGGAACACGGCTGTATATCGAAGCGTTCGCTGAGTGTG  235

Query  3916  AATACCGCTGCCGCTGGCGGTAACGCCAGCTTAATGACTATAGGTTAA  3963
             ||||||||||||||||||  ||||||||||||||||||||||||||
Sbjct  236   AATACCGCTGCCGCTGGCG--AACGCCAGCTTAATGACTATAGGTTAA  281


 Score = 30.1 bits (32),  Expect = 0.001
 Identities = 16/16 (100%), Gaps = 0/16 (0%)
 Strand=Plus/Plus

Query  1    ATGGGAACCACCACCA  16
             ||||||||||||||||
Sbjct  282  ATGGGAACCACCACCA  297


 Score = 21.1 bits (22),  Expect = 0.49
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Minus

Query  2516  GCGCGCAACTG  2526
             ||||||||||| 
Sbjct  144   GCGCGCAACTG  134


>lcl|63091 NODE_4_length_206_cov_205.684464
Length=236

 Score =  340 bits (376),  Expect = 3e-97
 Identities = 225/239 (94%), Gaps = 13/239 (5%)
 Strand=Plus/Plus

Query  1170  GCCGGAACTGG-CAGG-CTGGAACGGCATCGGTTTTGTTATT--CAGGCTTATCAAAAA-  1224
             |||||||||| |||| |||||||||||   ||||||||||  |||||||||||||||| 
Sbjct  2     GCCGGAACTGGGCAGGTCTGGAACGGCATA--TTTTGTTATTTTCAGGCTTATCAAAAAA  59

Query  1225  -CGCTGCCCGTTGGTGATCGATTACCTGATTGA-TCT-CGCCACCCGCAGCCGTCGCCG-  1280
              |||||||||||||||||||||||||||||||| ||| |||||||||||||||||||| 
Sbjct  60    ACGCTGCCCGTTGGTGATCGATTACCTGATTGAATCTTCGCCACCCGCAGCCGTCGCCGG  119

Query  1281  TCTGATGATTCGCCTGGTGAAAGGCGCGTACTGGGATAGTGAAATTAAGCGTGCGCAGAT  1340
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  120   TCTGATGATTCGCCTGGTGAAAGGCGCGTACTGGGATAGTGAAATTAAGCGTGCGCAGAT  179
```

```
Query  1341  GGACGGCCTTGAAGGTTATCCGGTTTATACCCGCAAGGTGTATACCGACGTTTCTTATC  1399
             ||||||||||||||||||||||||||||||||||||||||||||  ||||||||||||
Sbjct  180   GGACGGCCTTGAAGGTTATCCGGTTTATACCCGCAAGGTGTATAC--ACGTTTCTTATC  236


 Score = 21.1 bits (22),  Expect = 0.38
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Plus

Query  3879  TCTGGAACGGC  3889
             ||||||||||| 
Sbjct  18    TCTGGAACGGC  28


>lcl|63092 NODE_5_length_158_cov_201.348099
Length=188

 Score =  293 bits (324),  Expect = 3e-83
 Identities = 182/190 (95%), Gaps = 4/190 (2%)
 Strand=Plus/Minus

Query  584  AGTCACACATTGGTCGTAGCCCGTCACTGTTTGTTAATGCCGCCACCTGGGGGCTGCTGT  643
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||  ||||
Sbjct  188  AGTCACACATTGGTCGTAGCCCGTCACTGTTTGTTAATGCCGCCACCTGGGGGCGTCTGT  129

Query  644  TTACTGGCAAACTGGTTTCCACCCATAACGAAGCCAGCCTCTCCCGCTCGCTGAACCGCA  703
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  128  TTACTGGCAAACTGGTTTCCACCCATAACGAAGCCAGCCTCTCCCGCTCGCTGAACCGCA  69

Query  704  TTATCGGTAAAAGCGGTGAACCGCTGATCCG-CAAAGGTGTGGATATGGCGATGCGCC-T  761
             |||||||||||||  |||||||||||||| |||||||||||| |||||||||||||  |
Sbjct  68   TTATCGGTAAAAG-GTAGAACCGCTGATCCGTCAAAGGTGTGGA-ATGGCGATGCGCCTT  11

Query  762  GATGGGTGAG  771
             ||||||||||
Sbjct  10   GATGGGTGAG  1


>lcl|63093 NODE_6_length_291_cov_199.041245
Length=321

 Score =  524 bits (580),  Expect = 2e-152
 Identities = 315/324 (97%), Gaps = 7/324 (2%)
 Strand=Plus/Plus

Query  2692  ACCGAACAGGTCGTCGTGGATGTACTGGCCTCGGCGTTCGACAGTGCGGGTCAGC-GTTG  2750
             |||||||||||||||||||||||||||| ||||||||||||||||||||||||| ||||
Sbjct  1     ACCGAACAGGTCGTCGTGGATGTACTGGCCGCGGCGTTCGACAGTGCGGGTCAGCCGTTG  60

Query  2751  TTCGGCGCTGCGCGTGCTGTGCCTGCAAGATGAGA-TTGCCGACCACACGTTGAAAATGC  2809
             |||||||||||||||||||| |||||||||| |||||||||||||||||||||||||||
Sbjct  61    TTCGGCGCTGCGCGTGCTGTGCC-GCAAGATGAGAATTGCCGACCACACGTTGAAAATGC  119

Query  2810  TGCGCGGCGCAATGGCCGAATGCCGGATGGGTAATCCGGGTCGCCTGACCACCGATATCG  2869
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  120   TGCGCGGCGCAATGGCCGAATGCCGGATGGGTAATCCGGGTCGCCTGACCACCGATATCG  179

Query  2870  GTCCAGTGATTGATAGCGAAGCGAAAGCCAATATTGAGCGCCATATTCAGACCATGCGTA  2929
             |||||||||||||||| |||||||||||||||||||| ||||||||||||||||||||
Sbjct  180   GTCCAGTGATTGATAGC-AAGCGAAAGCCAATATTGAGCGCCCTATTCAGACCATGCGTA  238

Query  2930  GCAAAGGCCGTCCGGTGTTCCAGGCGGTGCGGGAAAACAGCGAAGATGCCCGTGAATGGC  2989
             |||||||||||||||||||| ||||||||||||||||||||||||||||||||||||||
Sbjct  239   GCAAAGGCCGTCCGGTGTTCCA-GCGGTGCGGGAAAACAGCGAAGATGCCCGTGAATGGC  297

Query  2990  AA--AGCGGCACCTTTGTCGCCCC  3011
             ||  ||||||||||||||||||||
Sbjct  298   AAAGAGCGGCACCTTTGTCGCCCC  321


 Score = 22.9 bits (24),  Expect = 0.15
 Identities = 12/12 (100%), Gaps = 0/12 (0%)
 Strand=Plus/Minus

Query  2860  ACCGATATCGGT  2871
             ||||||||||||
Sbjct  181   ACCGATATCGGT  170

 Score = 22.9 bits (24),  Expect = 0.15
```

```
 Identities = 12/12 (100%), Gaps = 0/12 (0%)
 Strand=Plus/Plus

Query  132  GGAAAACAGCGA  143
            ||||||||||||
Sbjct  269  GGAAAACAGCGA  280


>lcl|63094 NODE_7_length_461_cov_209.563995
Length=491

 Score =  690 bits (764),  Expect = 0.0
 Identities = 471/506 (93%), Gaps = 28/506 (5%)
 Strand=Plus/Plus

Query  3027  GGATGACTTTGCCGAATTGCAAAAAGAGGTCTTTGGTCCGGTGCTGCAT-GTGGTGCGTT  3085
             ||||||||||||||||||  |||||||||||||| ||||||||||||   ||||||||||
Sbjct  1      GGATGACTTTGCCGAATTCGAAAAAGAGGTCTT-GGTCCGGTGCTGCGCAGTGGTGCGTT  59

Query  3086  ACAACCGTAACCAGCTACCAGAGCTGATCGAGCAGATTAACGCTTCCGGTTA-TGGTCTG  3144
             |||||||||||||||||||||||||||||||||||||||||||  | ||||||  |||||||
Sbjct  60     ACAACCGTAACCAGCTACCAGAGCTGATCGAGCAGATTAACGGCTTCGGTTAATGGTCTG  119

Query  3145  ACGCTTGGCGTCCATACGCGCATTGA-TGAAACCATCGCCCAGGTC-ACTGGCTCGGCCC  3202
             ||  ||||||||||||||||  |||| ||||||||||||| ||||| ||||||||||||
Sbjct  120   AC--TTGGCGTCCATACGCG--TTGAATGAAACCATCGCC-AGGTCCACTGGCTCGGC--  172

Query  3203  ATGTTGGTAACCTGTATGTTAA--CCGTAATATGGTGGGCGCAGTGGTTGGTGTGCAGCC  3260
             ||||||||||||||||||||||  |||||||||| ||||||| ||||| ||||||||||||
Sbjct  173   ATGTTGGTAACCTGTATGTTAAAACCGTAATATG-TGGGCGC-GTGGT-GGTGTGCAGCC  229

Query  3261  GTTCGGCGGCGAAGGGTTGTCCGGTACCGGGCCGAAAGCA-GGCGGTCCGCTCTATCTCT  3319
             |||||||||||||||||||||||||||||||||||||||| ||||||||||||||||||||
Sbjct  230   GTTCGGCGGCGAAGGGTTGTCCGGTACCGGGCCGAAAGCAAGGCGGTCCGCTCTATCTCT  289

Query  3320  ACCGTCTGCTGGCGAATCGCCCGGAAAGTGCGCTGGCAGTGACGCTCGCGCGTCAGGATG  3379
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  290   ACCGTCTGCTGGCGAATCGCCCGGAAAGTGCGCTGGCAGTGACGCTCGCGCGTCAGGATG  349

Query  3380  CAAAGTATCCGGTCGATGCGCAGTTGAAAGCCGCATTGACTCAGCCGCTAAATGCACTGC  3439
             |||   ||||||||||||||||||||||||||||||||||  |||||||||||||
Sbjct  350   CAA---ATCCGGTCGATGCGCAGTTGAAAGCCGCATTGACTCAGC--CTAAATGCACTGC  404

Query  3440  GGGAAT-GGGCAGCAAATCGTCCAGAATTGCAGGCGTTATGTACGCAATATGGCGAGCT-  3497
             ||||||  ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  405   GGGAATGGGGCAGCAAATCGTCCAGAATTGCAGGCGTTATGTACGCAATATGGC-AGCTG  463

Query  3498  GGCGCAGGCAGGA--ACACAACGATT  3521
             |||||||||||||  ||||||||||
Sbjct  464   GGCGCAGGCAGGAACACACAACGATT  489


 Score = 21.1 bits (22),  Expect = 0.82
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Plus

Query  238  GACTTTGCCGA  248
            ||||||||||
Sbjct  5    GACTTTGCCGA  15


 Score = 21.1 bits (22),  Expect = 0.82
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Plus

Query  787  GAAACCATCGC  797
            ||||||||||
Sbjct  144  GAAACCATCGC  154


>lcl|63095 NODE_8_length_243_cov_213.312759
Length=273

 Score =  416 bits (460),  Expect = 6e-120
 Identities = 264/277 (95%), Gaps = 11/277 (3%)
 Strand=Plus/Plus

Query  150  GCCGGAGCTACCTGCGCTGCTTTCTGGCGCGGCCAATGAGAGCGATGAAGCACCGACTCC  209
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    GCCGGAGCTACCTGCGCTGCTTTCTGGCGCGGCCAATGAGAGCGATGAAGCACCGACTCC  60
```

```
Query  210  GGCAGAGGAACCACACCA--GCCATTCCTCGACTTTGCCGAGCAAATATTGCCCCAGTCG  267
            ||||||||||||||||||  ||||||||||||||||||||||||||||||||||||||||
Sbjct  61   GGCAGAGGAACCACACCAGCGCCATTCCTCGACTTTGCCGAGCAAATATTGCCCCAGTCG  120

Query  268  GTTTCCCGCGCCGCGATCACCGCGGCCTATCGCCGCCCGGAAACCGAAGCGGTTTCTATG  327
            ||||||||||||||||||||||||||||||||||    ||||||   ||||||||||||
Sbjct  121  GTTTCCCGCGCCGCGATCACCGCGGCCTATCGCCGC---GAAACC-CGGCGGTTTCTATG  176

Query  328  CTGCT-GGAACAAGCCCGCCTGCCGCAGCCAGTTG-CTGAACAGGCGC-ACAAACTGGCG  384
            ||||| ||||||||||||||||||||||||||||| |||||||||||| ||||||||||
Sbjct  177  CTGCTGGGAACAAGCCCGCCTGCCGCAGCCAGTTGCCTGAACAGGCGCAACAAACTGGCG  236

Query  385  TATCAGCTGGCCGATAAA-CTGCGT-AATCnnnnnnn  419
            |||||||||||||||||| |||||| ||||||||||
Sbjct  237  TATCAGCTGGCCGATAAACCTGCGTAAATCAAAAAAA  273


 Score = 21.1 bits (22),  Expect = 0.45
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Minus

Query  146  CTCTGCCGGAG  156
            ||||||||||
Sbjct  67   CTCTGCCGGAG  57


 Score = 21.1 bits (22),  Expect = 0.45
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Plus

Query  3031  GACTTTGCCGA  3041
             ||||||||||
Sbjct  91    GACTTTGCCGA  101


>lcl|63096 NODE_9_length_98_cov_212.948975
Length=128

 Score =  197 bits (218),  Expect = 1e-54
 Identities = 121/125 (96%), Gaps = 3/125 (2%)
 Strand=Plus/Plus

Query  462  GTTTTCGCTGTCATCGCAGGAAGGCGTGGCGCTGATGTGTCTGGC-GGAAGCGTTGTTGC  520
            ||||||||||||||||||||||||||||| ||||||||||||||| |||||||||||||
Sbjct  2    GTTTTCGCTGTCATCGCAGGAAGGCGTGGAGCTGATGTGTCTGGCGGGAAGCGTTGTTGC  61

Query  521  GTATTCCCGACAAAGCCA-CCCGCGACGCG-TTAATTCGCGACAAAATCAGCAACGGTAA  578
            |||||||||||||||||| |||||||||| ||||||||||||||||||||||||||||||
Sbjct  62   GTATTCCCGACAAAGCCACCCCGCGACGCGTTTAATTCGCGACAAAATCAGCAACGGTAA  121

Query  579  CTGGC  583
            |||||
Sbjct  122  CTGGC  126


>lcl|63097 NODE_10_length_148_cov_229.418915
Length=178

 Score =  300 bits (332),  Expect = 2e-85
 Identities = 178/182 (97%), Gaps = 4/182 (2%)
 Strand=Plus/Plus

Query  831   GAAAGGTTTCCGTTACTCTTACGATATGCTGGGCGAAGCCGCGCTGACCGCCGCAGATGC  890
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     GAAAGGTTTCCGTTACTCTTACGATATGCTGGGCGAAGCCGCGCTGACCGCCGCAGATGC  60

Query  891   ACAGGCGTATATGGTTTCCTATCAGCAGGCGATTCACGCCATCGGTAAAGCGTCTAACGG  950
            |||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    ACAG-CGTATATGGTTTCCTATCAGCAGGCGATTCACGCCATCGGTAAAGCGTCTAACGG  119

Query  951   TCGTGGCATCTATGAAGGGCCGGGCATTTCAATCAAACTGTCGGCGCTGCATCCGCGTTA  1010
            |||||||||| |||||||||||||||||||||||||||  ||||||||||||||||||||
Sbjct  120   TCGTGGCATC-ATGAAGGGCCGGGCATTTCAATCAAACTG--GGCGCTGCATCCGCGTTA  176

Query  1011  TA  1012
            ||
Sbjct  177   TA  178


 Score = 21.1 bits (22),  Expect = 0.29
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
```

```
 Strand=Plus/Plus

Query  101  AGCAGGCGATT  111
            ||||||||||
Sbjct  83   AGCAGGCGATT  93


>lcl|63098 NODE_11_length_76_cov_221.750000
Length=106

 Score =  156 bits (172),  Expect = 3e-42
 Identities = 101/106 (95%), Gaps = 5/106 (4%)
 Strand=Plus/Plus

Query  1026  TGACCGGGTAATGGAAGAGCTTTACCCGCGTCTGAAATCACTCACCCTGCTGGCGCGTCA  1085
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4     TGACCGGGTAATGGAAGAGCTTTACCCGCGTCTGAAATCACTCACCCTGCTGGCGCGTCA  63

Query  1086  GTACGATATTGGTATCAAC-ATTGACGCCGAAGAGTCCGA-TCGCC  1129
            |||||||||  ||||||| |||||||||||||| ||||| |||||
Sbjct  64    GTACGATAT--GTATCAACGATTGACGCCGAAGA-TCCGATTCGCC  106


>lcl|63099 NODE_12_length_62_cov_208.145157
Length=92

 Score =  129 bits (142),  Expect = 4e-34
 Identities = 89/94 (94%), Gaps = 5/94 (5%)
 Strand=Plus/Minus

Query  3611  TGACTCAG-CTCGCCGCCGTGCTGGCGGTGGGC-AGCCAGGTACTGTGGCCGGATGACGC  3668
            |||||||| |||||||||||||||||||||||| |||||||||||||||| ||||||||
Sbjct  92    TGACTCAGTCTCGCCGCCGTGCTGGCGGTGGGCGAGCCAGGTACTGTGGCC-GATGACGC  34

Query  3669  G-CTGCATCGTCAGTTAGTGAAGGCATTGCCATC  3701
            | |||||||||||||| ||||||||||||||||
Sbjct  33    GCCTGCATCGTCAGTTAG-GAAGGCATTGCCATC  1


 Score = 21.1 bits (22),  Expect = 0.14
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Minus

Query  1547  TGGGCGAGCCA  1557
            ||||||||||||
Sbjct  64    TGGGCGAGCCA  54


 Score = 21.1 bits (22),  Expect = 0.14
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Minus

Query  1418  TGCTGGCGGTG  1428
            ||||||||||||
Sbjct  73    TGCTGGCGGTG  63


>lcl|63100 NODE_13_length_73_cov_221.095886
Length=103

 Score =  187 bits (206),  Expect = 2e-51
 Identities = 103/103 (100%), Gaps = 0/103 (0%)
 Strand=Plus/Plus

Query  1565  AGCAGGTCACCGGGAAAGTTGCCGACGGCAAACTTAACCGTCCGTGTCGTATTTATGCTC  1624
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1     AGCAGGTCACCGGGAAAGTTGCCGACGGCAAACTTAACCGTCCGTGTCGTATTTATGCTC  60

Query  1625  CGGTTGGCACACATGAAACGCTGTTGGCGTATCTGGTGCGTCG  1667
            |||||||||||||||||||||||||||||||||||||||||||
Sbjct  61    CGGTTGGCACACATGAAACGCTGTTGGCGTATCTGGTGCGTCG  103


>lcl|63101 NODE_15_length_142_cov_198.697189
Length=172

 Score =  239 bits (264),  Expect = 6e-67
 Identities = 164/177 (92%), Gaps = 11/177 (6%)
 Strand=Plus/Minus
```

```
Query  2522  AACTGACGGG-TGATGATCGCGTG--CGCGGGGTGATGTTTACCGGTTCAACCGAAGTCG  2578
             |||||||||| |||||||||||||  |||||||||||||||||||||||||||||||||
Sbjct  172   AACTGACGGGGTGATGATCGCGTGTGCGCGGGGTGATGTTTACCGGTTCAACCGAAGTCG  113

Query  2579  CTACGTTACTGCAGCG-CAATATCGCCAGCCGCCTGGACGCTCAGGGTCGCCCTATTCCG  2637
             | ||||||||||||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct  112   C-ACGTTACTGCAGCGACAATATCGCCAGCCGCCTGGACGCTCAGGGTCGCCCTATTCGA  54

Query  2638  CTCATCGCTGAAACCGGCGGCATG--AACGCGATGATTGTCGATTCTTCAGCACTGA  2692
             ||||||||||||||  |||||||  |||||||||||||||||||  ||||||||
Sbjct  53    CTCATCGCTGAAACC--CGGCATGTGAACGCGATGATTGTCGATTCT--AGCACTGA  1


>lcl|63102 NODE_16_length_65_cov_216.415390
Length=95

 Score =  114 bits (126),  Expect = 8e-30
 Identities = 86/97 (88%), Gaps = 7/97 (7%)
 Strand=Plus/Plus

Query  1502  AGAACTACTACCCGGGTCAGTACG--AGTTCCAGTGCCTGCATGGTATGGGC--GAGCCA  1557
             |||||||||||||||||||||||    ||||  ||||||||||||||||||  ||||||
Sbjct  1     AGAACTACTACCCGGGTCAGTACGTACGTTC--GTGCCTGCATGGTATGGGCGAGAGCCA  58

Query  1558  CTGTATG-AGCAGGTCACCGGGAAAGTTGCCGACGGC  1593
             ||||||| |||||||| | || ||||||||||||||||
Sbjct  59    CTGTATGAAGCAGGTCCCGGGAAAAGTTGCCGACGGC  95


>lcl|63103 NODE_17_length_101_cov_213.732666
Length=131

 Score =  185 bits (204),  Expect = 8e-51
 Identities = 126/136 (92%), Gaps = 8/136 (5%)
 Strand=Plus/Plus

Query  1638  TGAAACGCTGTTGGCGTATCTGGTGCGTCGC--CTGCTGGAAAACGGTGCTAACACCTCG  1695
             ||||||||||||||||||  |||||||  |  ||||||||||||||||  |||||||
Sbjct  1     TGAAACGCTGTTGGCGTAT--GGTGCGTGCCTGCTGCTGGAAAACGGTGC--ACACCTC-  55

Query  1696  TTTGTTAACCGTATTGCCGACACCTCTTTGCCACTGGATGAACTGGTC-GCCGATCCGGT  1754
             ||||||||||||||||||||||||||||||||||||||||||||||| ||||||||||
Sbjct  56    TTTGTTAACCGTATTGCCGACACCTCTTTGCCACTGGATGAACTGGTCGGCCGATCCGGT  115

Query  1755  CACTGCTGTAGAAAAA  1770
             ||||||||||||||||
Sbjct  116   CACTGCTGTAGAAAAA  131


 Score = 24.7 bits (26),  Expect = 0.017
 Identities = 16/18 (88%), Gaps = 0/18 (0%)
 Strand=Plus/Plus

Query  3212  ACCTGTATGTTAACCGTA  3229
             |||| | ||||||||||
Sbjct  51    ACCTCTTTGTTAACCGTA  68


 Score = 21.1 bits (22),  Expect = 0.20
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Plus

Query  326  TGCTGCTGGAA  336
            ||||||||||
Sbjct  30   TGCTGCTGGAA  40


 Score = 21.1 bits (22),  Expect = 0.20
 Identities = 11/11 (100%), Gaps = 0/11 (0%)
 Strand=Plus/Plus

Query  1147  CTGCTGGAAAA  1157
             ||||||||||
Sbjct  32    CTGCTGGAAAA  42
```

APPENDIX E

METASIM "EXACT" EXCERPT

>r1.1 |SOURCES={KEY=0f5d3d4b...,fw,2703-3330}|ERRORS={}|SOURCE_1="eco:b1014 putA, ECK1005, JW0999, poaA, putC; fused DNA-binding transcriptional regulator/proline dehydrogenase/pyrroline-5-carboxylate dehydrogenase (EC:1.5.99.8 1.5.1.12); K00294 1-pyrroline-5-carboxylate dehydrogenase [EC:1.5.1.12] K00318 proline dehydrogenase [EC:1.5.99.8] (N)" (0f5d3d4b481b65cb949b9c3998839ebd73b6a641)

gtcgtggatgtactggcctcggcgttcgacagtgcgggtcagcgttgttcggcgctgcgcgtgct
gtgcctgcaagatgagattgccgaccacacgttgaaaatgctgcgcggcgcaatggccgaatgcc
ggatgggtaatccgggtcgcctgaccaccgatatcggtccagtgattgatagcgaagcgaaagcc
aatattgagcgccatattcagaccatgcgtagcaaaggccgtccggtgttccaggcggtgcggga
aaacagcgaagatgcccgtgaatggcaaagcggcacctttgtcgccccgacgctgatcgaactgg
atgactttgccgaattgcaaaaagaggtctttggtccggtgctgcatgtggtgcgttacaaccgt
aaccagctaccagagctgatcgagcagattaacgcttccggttatggtctgacgcttggcgtcca
tacgcgcattgatgaaaccatcgcccaggtcactggctcggcccatgttggtaacctgtatgtta
accgtaatatggtgggcgcagtggttggtgtgcagccgttcggcggcgaagggttgtccggtacc
gggccgaaagcaggcggtccgctctatctctaccgtctgctg

>r2.1 |SOURCES={KEY=0f5d3d4b...,fw,2534-3180}|ERRORS={}|SOURCE_1="eco:b1014 putA, ECK1005, JW0999, poaA, putC; fused DNA-binding transcriptional regulator/proline dehydrogenase/pyrroline-5-carboxylate dehydrogenase (EC:1.5.99.8 1.5.1.12); K00294 1-pyrroline-5-carboxylate dehydrogenase [EC:1.5.1.12] K00318 proline dehydrogenase [EC:1.5.99.8] (N)" (0f5d3d4b481b65cb949b9c3998839ebd73b6a641)

tgatcgcgtgcgcggggtgatgtttaccggttcaaccgaagtcgctacgttactgcagcgcaata
tcgccagccgcctggacgctcagggtcgccctattccgctcatcgctgaaaccggcggcatgaac
gcgatgattgtcgattcttcagcactgaccgaacaggtcgtcgtggatgtactggcctcggcgtt
cgacagtgcgggtcagcgttgttcggcgctgcgcgtgctgtgcctgcaagatgagattgccgacc
acacgttgaaaatgctgcgcggcgcaatggccgaatgccggatgggtaatccgggtcgcctgacc
accgatatcggtccagtgattgatagcgaagcgaaagccaatattgagcgccatattcagaccat
gcgtagcaaaggccgtccggtgttccaggcggtgcgggaaaacagcgaagatgcccgtgaatggc
aaagcggcacctttgtcgccccgacgctgatcgaactggatgactttgccgaattgcaaaaagag
gtctttggtccggtgctgcatgtggtgcgttacaaccgtaaccagctaccagagctgatcgagca
gattaacgcttccggttatggtctgacgcttggcgtccatacgcgcattgatgaaaccatc

>r3.1 |SOURCES={KEY=0f5d3d4b...,fw,163-820}|ERRORS={}|SOURCE_1="eco:b1014 putA, ECK1005, JW0999, poaA, putC; fused DNA-binding transcriptional regulator/proline dehydrogenase/pyrroline-5-carboxylate dehydrogenase (EC:1.5.99.8 1.5.1.12); K00294 1-pyrroline-5-carboxylate dehydrogenase [EC:1.5.1.12] K00318 proline dehydrogenase [EC:1.5.99.8] (N)" (0f5d3d4b481b65cb949b9c3998839ebd73b6a641)

cgctgctttctggcgcggccaatgagagcgatgaagcaccgactccggcagaggaaccacaccag
ccattcctcgactttgccgagcaaatattgccccagtcggtttcccgcgccgcgatcaccgcggc

ctatcgccgcccggaaaccgaagcggtttctatgctgctggaacaagcccgcctgccgcagccag

ttgctgaacaggcgcacaaactggcgtatcagctggccgataaactgcgtaatcaaaaaaatgcc

agtggtcgcgcaggtatggtccaggggttattgcaggagttttcgctgtcatcgcaggaaggcgt

ggcgctgatgtgtctggcggaagcgttgttgcgtattcccgacaaagccacccgcgacgcgttaa

ttcgcgacaaaatcagcaacggtaactggcagtcacacattggtcgtagcccgtcactgtttgtt

aatgccgccacctgggggctgctgtttactggcaaactggtttccacccataacgaagccagcct

ctcccgctcgctgaaccgcattatcggtaaaagcggtgaaccgctgatccgcaaaggtgtggata

tggcgatgcgcctgatgggtgagcagttcgtcactggcgaaaccatcgcggaagcgttagccaat

gcccgca

>r4.1 |SOURCES={KEY=0f5d3d4b...,fw,3344-3963;KEY=0f5d3d4b...,fw,0-

40}|ERRORS={}|SOURCE_1="eco:b1014 putA, ECK1005, JW0999, poaA, putC; fused DNA-binding

transcriptional regulator/proline dehydrogenase/pyrroline-5-carboxylate dehydrogenase

(EC:1.5.99.8 1.5.1.12); K00294 1-pyrroline-5-carboxylate dehydrogenase [EC:1.5.1.12]

K00318 proline dehydrogenase [EC:1.5.99.8] (N)"

(0f5d3d4b481b65cb949b9c3998839ebd73b6a641)

aagtgcgctggcagtgacgctcgcgcgtcaggatgcaaagtatccggtcgatgcgcagttgaaag

ccgcattgactcagccgctaaatgcactgcgggaatgggcagcaaatcgtccagaattgcaggcg

ttatgtacgcaatatggcgagctggcgcaggcaggaacacaacgattgctgccggggccgacggg

tgaacgcaacacctggacgctgctgccgcgtgagcgcgtgttgtgtattgccgatgatgagcagg

atgcgctgactcagctcgccgccgtgctggcggtgggcagccaggtactgtggccggatgacgcg

ctgcatcgtcagttagtgaaggcattgccatcggcagtcagcgaacgtattcaactggcgaaagc

ggaaaatataaccgctcaaccgtttgatgcggtgatcttccacggtgattcggatcagcttcgcg

cattgtgtgaagcagttgccgcgcgggatggcacaattgtttcggtgcagggttttgcccgtggc

gaaagcaatatccttctggaacggctgtatatcgagcgttcgctgagtgtgaataccgctgccgc

tggcggtaacgccagcttaatgactataggttaaatgggaaccaccaccatgggggttaagctgg

acgacgcga

>r5.1 |SOURCES={KEY=0f5d3d4b...,fw,3400-3963;KEY=0f5d3d4b...,fw,0-

118}|ERRORS={}|SOURCE_1="eco:b1014 putA, ECK1005, JW0999, poaA, putC; fused DNA-binding

transcriptional regulator/proline dehydrogenase/pyrroline-5-carboxylate dehydrogenase

(EC:1.5.99.8 1.5.1.12); K00294 1-pyrroline-5-carboxylate dehydrogenase [EC:1.5.1.12]

K00318 proline dehydrogenase [EC:1.5.99.8] (N)"

(0f5d3d4b481b65cb949b9c3998839ebd73b6a641)

agttgaaagccgcattgactcagccgctaaatgcactgcgggaatgggcagcaaatcgtccagaa

ttgcaggcgttatgtacgcaatatggcgagctggcgcaggcaggaacacaacgattgctgccggg

gccgacgggtgaacgcaacacctggacgctgctgccgcgtgagcgcgtgttgtgtattgccgatg

atgagcaggatgcgctgactcagctcgccgccgtgctggcggtgggcagccaggtactgtggccg

gatgacgcgctgcatcgtcagttagtgaaggcattgccatcggcagtcagcgaacgtattcaact

ggcgaaagcggaaaatataaccgctcaaccgtttgatgcggtgatcttccacggtgattcggatc

agcttcgcgcattgtgtgaagcagttgccgcgcgggatggcacaattgtttcggtgcagggtttt

gcccgtggcgaaagcaatatccttctggaacggctgtatatcgagcgttcgctgagtgtgaatac

[...]

APPENDIX F

METASIM 454 EXCERPT

>r1.1 |SOURCES={KEY=0f5d3d4b...,fw,1824-
2078}|ERRORS={22_1:T,38_1:T,57_1:G,72_1:G,89_1:T,119_1:T,153_1:G,206_1:A}|SOURCE_1="eco:b
1014 putA, ECK1005, JW0999, poaA, putC; fused DNA-binding transcriptional
regulator/proline dehydrogenase/pyrroline-5-carboxylate dehydrogenase (EC:1.5.99.8
1.5.1.12); K00294 1-pyrroline-5-carboxylate dehydrogenase [EC:1.5.1.12] K00318 proline
dehydrogenase [EC:1.5.99.8] (N)" (0f5d3d4b481b65cb949b9c3998839ebd73b6a641)
cgcgatctttacggtcacgggcgTcgacaactcggcagggTctggatctcgctaacgaacGaccg
cctggcctcgcGtctcctctgccctgctcTaatagtgcactgcaaaaatggcaggccttgTccaa
tgctggaacaaccggtagcggcaggtgagaGtgtcgcccgttattaaccctgcggaaccgaaaga
tattgtgggctatgtgcgtAgaagccacgccgcgtgaagtagaacaggcgctggaaagtgcggtt
aa
>r2.1 |SOURCES={KEY=0f5d3d4b...,fw,2006-2253}|ERRORS={18_1:G,59:-,77_1:G,151_1:G,154:-
}|SOURCE_1="eco:b1014 putA, ECK1005, JW0999, poaA, putC; fused DNA-binding
transcriptional regulator/proline dehydrogenase/pyrroline-5-carboxylate dehydrogenase
(EC:1.5.99.8 1.5.1.12); K00294 1-pyrroline-5-carboxylate dehydrogenase [EC:1.5.1.12]
K00318 proline dehydrogenase [EC:1.5.99.8] (N)"
(0f5d3d4b481b65cb949b9c3998839ebd73b6a641)
gaaagatattgtgggctatGgtgcgtgaagccacgccgcgtgaagtagaacaggcgctggaagtg
cggttaataacgcGgccaatctggtttgccacgcctccggctgaacgcgcagcgattttgcaccg
cgctgccgtgctgatggaaagccGagtgcagcaactgattggtattctggtgcgtgaggccggaa
aaaccttcagtaacgccattgccgaagtgcgcgaagcggtcgattttctccac
>r3.1 |SOURCES={KEY=0f5d3d4b...,fw,3721-3963;KEY=0f5d3d4b...,fw,0-
16}|ERRORS={18_1:C,26_1:A,95_1:C,134_1:A,139_1:C,168_1:A,190:-
,200_1:T,228_1:C,243_1:C}|SOURCE_1="eco:b1014 putA, ECK1005, JW0999, poaA, putC; fused
DNA-binding transcriptional regulator/proline dehydrogenase/pyrroline-5-carboxylate
dehydrogenase (EC:1.5.99.8 1.5.1.12); K00294 1-pyrroline-5-carboxylate dehydrogenase
[EC:1.5.1.12] K00318 proline dehydrogenase [EC:1.5.99.8] (N)"
(0f5d3d4b481b65cb949b9c3998839ebd73b6a641)
aactggcgaaagcggaaaaCtataaccgActcaaccgtttgatgcggtgatcttccacggtgatt
cggatcagcttcgcgcattgtgtgaagcagttgCccgcgcgggatggcacaattgtttcggtgca
gggttttgAcccgtCggcgaaagcaatatccttctggaacggctAgtatatcgagcgttcgctga
ggtgaataccgTctgccgctggcggtaacgccagcttaatCgactataggttaaatCgggaacca
ccacca
>r4.1 |SOURCES={KEY=0f5d3d4b...,fw,1209-1469}|ERRORS={20:-,40_1:C,44_1:C,175:-,198:-
,209_1:A,237_1:T}|SOURCE_1="eco:b1014 putA, ECK1005, JW0999, poaA, putC; fused DNA-
binding transcriptional regulator/proline dehydrogenase/pyrroline-5-carboxylate
dehydrogenase (EC:1.5.99.8 1.5.1.12); K00294 1-pyrroline-5-carboxylate dehydrogenase

[EC:1.5.1.12] K00318 proline dehydrogenase [EC:1.5.99.8] (N)"
(0f5d3d4b481b65cb949b9c3998839ebd73b6a641)
caggcttatcaaaaacgctgccgttggtgatcgattacctCgattCgatctcgccacccgcagcc
gtcgccgtctgatgattcgcctggtgaaaggcgcgtactgggatagtgaaattaagcgtgcgcag
atggacggccttgaaggttatccggtttatacccgcaaggtgtatacgacgtttcttatctcgcc
tgtcgaaaaagctgActggcggtgccgaatctaatctacccgcTagttcgcgacgcacaacgccc
a
>r5.1 |SOURCES={KEY=0f5d3d4b...,fw,1985-2239}|ERRORS={13_1:A,26:-,49_1:A,139_1:T,251:-
}|SOURCE_1="eco:b1014 putA, ECK1005, JW0999, poaA, putC; fused DNA-binding
transcriptional regulator/proline dehydrogenase/pyrroline-5-carboxylate dehydrogenase
(EC:1.5.99.8 1.5.1.12); K00294 1-pyrroline-5-carboxylate dehydrogenase [EC:1.5.1.12]
K00318 proline dehydrogenase [EC:1.5.99.8] (N)"
(0f5d3d4b481b65cb949b9c3998839ebd73b6a641)
cgttattaaccctgAcggaaccgaaagtattgtgggctatgtgcgtgaagAccacgccgcgtgaa
gtagaacaggcgctggaaagtgcggttaataacgcgccaatctggtttgccacgcctccggctga
acgcgcagcgaTttttgcaccgcgctgccgtgctgatggaaagccagatgcagcaactgattggt
attctggtgcgtgaggccggaaaaaccttcagtaacgccattgccgaagtgcgcgaaggg
>r6.1 |SOURCES={KEY=0f5d3d4b...,bw,1905-
2155}|ERRORS={2_1:A,24_1:T,44_1:T,89_1:A,95_1:T,203_1:C,205_1:T}|SOURCE_1="eco:b1014
putA, ECK1005, JW0999, poaA, putC; fused DNA-binding transcriptional regulator/proline
dehydrogenase/pyrroline-5-carboxylate dehydrogenase (EC:1.5.99.8 1.5.1.12); K00294 1-
pyrroline-5-carboxylate dehydrogenase [EC:1.5.1.12] K00318 proline dehydrogenase
[EC:1.5.99.8] (N)" (0f5d3d4b481b65cb949b9c3998839ebd73b6a641)
TTTACCATCAGCACGGCAGCGCGGTGTCAAAATCGCTGCGCGTTCAGTCCGGAGGCGTGGCAAAC
CAGATTGGCGCGTTATTAACCGCACTTTACCAGCGTCCTGTTCTACTTCACGCGGCGTGGCTTCA
CGCACATAGCCCACAATATCTTTCGGTTCCGCAGGGTTAATAACGGGCGACATCTCACCTGCCGC
TACCGGTTGTTCCACGCTATTGGCAAGGCCTGCCATTTTTGCAGTGCACTATTGAGCAGGGC
>r7.1 |SOURCES={KEY=0f5d3d4b...,bw,566-
822}|ERRORS={22_1:A,52_1:A,86_1:C,90_1:T,176_1:A,216_1:G,230_1:C}|SOURCE_1="eco:b1014
putA, ECK1005, JW0999, poaA, putC; fused DNA-binding transcriptional regulator/proline
dehydrogenase/pyrroline-5-carboxylate dehydrogenase (EC:1.5.99.8 1.5.1.12); K00294 1-
pyrroline-5-carboxylate dehydrogenase [EC:1.5.1.12] K00318 proline dehydrogenase
[EC:1.5.99.8] (N)" (0f5d3d4b481b65cb949b9c3998839ebd73b6a641)
CTTGCGGGCATTGGCTAACGCTTACCGCGATGGTTTCGCCAGTGACGAACTGCTACACCCATCAG
GCGCATCGCCATATCCACACCTTTCGCGGTATCAGCGGTTCACCGCTTTTACCGATAATGCGGTT
CAGCGAGCGGGAGAGGCTGGCTTCGTTATGGGTGGAAACCAGTTTGCCAGTAAAACAGCAGCCCC
CAGGTGGCGGCATTAACAAACAGTGACGGGGCTACGACCAATCGTGTGACTGCCAGTTACCGTTG
CTG
[...]

REFERENCES

Altschul, Stephen F., et al. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acid Res.* 25 (1997): 3389-3402.

Applied Biosystems. *Applied Biosystems Product Catalog.* 22 March 2010. Web. 22 March 2010 <https://products.appliedbiosystems.com/ab/en/US/adirect/ab>.

Blattner, F. R., et al. "The Complete Genome Sequence of Escherichia coli K-12." *Science* (1997): 1453-1462.

Chaisson, Mark, Pavel Pevzner and Haixu Tang. "Fragment Assembly with Short Reads." *Bioinfomatics* 20.13 (2004): 2067-2074.

Cormen, Thomas H., et al. *Introduction to Algorithms.* 2nd ed. New York: McGraw Hill, 2000.

Darnell, James, Harvey Lodish and David Baltimore. *Molecular Cell Biology.* 2nd ed. New York: Scientific American Books, 1990.

de Bruijn, N. G. "A Combinatorial Problem." *Koninklijke Nederlandse Akademie v. Wetenschappen* 49 (1946): 758-764.

DiGuistini, Scott, et al. "De Novo Genome Sequence Assembly of a Filamentous Fungus Using Sanger, 454 and Illumina Sequence Data." *Genome Biology* 10 (2009): R94.

Ewing, Brent and Phil Green. "Base-Calling of Automated Sequencer Traces UsingPhred. II. Error Probabilities." *Genome Res.* (1998): 186-194.

Ewing, Brent, et al. "Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment." *Genome Res.* 8 (1998): 175-185.

Gilbert, W. and A. Maxam. "The Nucleotide Sequence of the Lac Operator." *Proc Natl Acad Sci U.S.A.* 12.70 (1973): 3581-3584.

Gross, J. L. and J. Yellen. *Handbook of Graph Theory.* Boca Raton: CRC Press LLC, 2004.

Idury, Ramana M. and Michael S. Waterman. "A New Algorithm for DNA Sequence Assembly." *Journal of Computational Biology* 2.2 (1995): 291-306.

Kyoto University Bioinformatics Center. *GenomeNet.* 22 March 2010. Web. 22 March 2010 <http://www.genome.jp/>.

Lehninger, Albert L., David L. Nelson and Michael M. Cox. *Principles of Biochemistry.* 2nd ed. New York: Worth Publishers, 1993.

leipzig. *Standardized-Velvet-Assembly-Report - Project Hosting on Google Code.* 22 March 2010. Web. 22 March 2010 <http://code.google.com/p/standardized-velvet-assembly-report/>.

Marguiles, Marcel, et al. "Genome Sequencing in Open Microfabricated High Density Picoliter Reactors." *Nature* 437 (2005): 376-380.

Maxam, A. M. and W. Gilbert. "A New Method for Sequencing DNA." *Proc Natl Acad Sci U S A.* 2.74 (1977): 560-564.

NCBI. *FASTA Format Description.* 22 March 2010. Web. 22 March 2010 <http://www.ncbi.nlm.nih.gov/blast/fasta.shtml>.

—. *National Center for Biotechnology Information.* 22 March 2010. Web. 22 March 2010 <http://www.ncbi.nlm.nih.gov/>.

Pevzner, P. A. "1-Tuple DNA Sequencing: Computer Analysis." *J Biomol Struct Dyn* 7 (1989): 63-73.

Pevzner, Pavel A., Haixu Tang and Michael S. Waterman. "An Eulerian Path Approach to DNA Fragment Assembly." *PNAS* 98.17 (2001): 9748-9753.

Richter, Daniel C., et al. "MetaSim—A Sequencing Simulator for Genomics and Metagenomics." *PLoS ONE* 3.10 (2008): e3373.

Roche Diagnostics Co. *Products & Solutions - Syetem Benefits : 454 Life Sciences, a Roche Company.* 22 March 2010. Web. 22 March 2010 <http://454.com/products-solutions/system-benefits.asp>.

Ronaghi, M., M. Uhlén and P. Nyrén. "A Sequencing Method Based on Real-Time Pyrophosphate." *Science* (1998): 363, 365.

Sanger, Fredrick, S. Nicklen and A. R. Coulson. "DNA Sequencing with Chain-Terminating Inhibitors." *Proc Natl Acad Sci U.S.A.* 12.74 (1977): 5463-5467.

Schatz, Michael C., et al. "Hawkeye: An Interactive Visual Analytics Tool for Genome Assemblies." *Genome Biology* 8 (2007): R34.

Tamarin, Robert H. *Principles of Genetics.* 4th ed. Dubuque: Wm. C. Brown Publishers, 1993.

Watson, James D. and Francis H. Crick. "Genetical Implications of the structure of Deoxyribonucleic Acid." *Nature* 171 (1953): 964-967.

Watson, James D. and Franics H. Crick. "A Structure for Deoxyribose Nucleic Acid." *Nature* 171 (1953): 737-738.

Zerbino, Daniel R. and Ewan Birney. "Velvet: Algorithms for De Novo Short Read Assembly Using de Bruijn Graphs." *Genome Research* 18 (2008): 821-829.