A Novel Approach Towards a Comprehensive Consensus Representation of the Expressed Human Genome

Winston Hide winhide@sanbi.ac.za
John Burke info@sanbi.ac.za

Alan Christoffels **Robert Miller**

South African National Bioinformatics Institute, The University of the Western Cape
Private Bag X17, Cape Town, South Africa

Abstract

In order to provided a novel maximised approach to the generation of accurate, comprehensive, consensus sequences of the expressed human genome, we have developed and produced a system for a novel-representation, broad gene coverage, consensus database of expressed human gene fragments (ESTs). To perform clustering of ESTs, we have developed and employed D2-cluster, an algorithm based on the d2-search algorithm (Hide et al. 1994) specifically for EST clustering. D2-cluster does not require alignment in order to perform clustering (Burke, Davison and Hide, in prep). We have incorporated d2-cluster into a portable and novel system to perform clustering, alignment and automated error analysis of publicly available expressed sequence tags (STACK_PACK). The system includes a statistically robust algorithm that can detect and compensate for error within an aligned cluster of ESTs. We have manufactured a database of partial human consensus sequences from 552 013 ESTs from dbEST 040896 and TIGR. The database is termed Sequence Tag Alignment and Consensus Knowledgebase (STACK). STACK 1.0 contains 18 divisions based on tissue annotation identifying 204 431 unique sequences and generating 76 131 consensi which represent 321 134 ESTs. The consensus sequences have an average length of 497 bases, a 39% increase over the 357 base average length of the input data set. Clone Ids are used to join 92 759 unique sequences and 48 858 consensi into 61 632 linked sequences, averaging 900 bases each. The distribution of clusters compares favourably with UniGene, reflecting the difference in methodology of clustering and the higher input number of sequences into STACK. SANIGENE high accuracy database is also generated, consisting of sequences which agree in at least two ESTs. STACK is a distributable, core information resource upon which a comprehensive knowledgebase can be built.

1 Introduction

Expressed Sequence Tags (ESTs) represent a major gene expression and functional discovery resource. Due to the high volume and high throughput automated mode of manufacture however, ESTs present a major processing problem to DNA sequence based analytical systems. Undesirable characteristics of ESTs include: annotation errors, sequencing errors, short length, errors in reading frame, rearrangements, artifacts of generation, contaminants, and alternate representations of the same gene (Aaronson et al. 1996)

These characteristics can result in a significant error rate being present in public EST databases, such as UniGene and dbEST. The rate of error varies according to the method of database generation, and the source of EST. A non-trivial amount of processing of the sequence data is necessary to discard error-filled sequence regions, and to provide novel, low-error, non redundant human gene consensi and expression analysis candidates.

Current EST clustering and processing projects reduce error output in several pre-processing steps, that include masking of repeat sequences, pruning poor quality sequence and masking low information sequence (Adams et.al 1995, Okubo et. al 1992, Houlgatte et. al 1995). The subsequent strategies of EST clustering projects are quality-based, building a cluster based on strict close identity overlap criteria. There is a resultant sacrifice of longer EST sequence consensus for increased accuracy of shorter, but better quality consensus sequence (Sutton et al., 1995). Alignment-based clustering requires that matching sequences be highly identical, and that surrounding regions with low-fidelity sequence do not interfere with the assignment of an EST to a cluster.

2 Aim

We have set out to generate a database of EST alignments and consensus sequences that reflect a maximum possible useful EST consensus by utilising both poor quality and good quality ESTs to contribute to the composite consensus sequence.

3 Implementation

In order to utilise exhaustive comparison techniques, we have implemented STACK_PACK on a multiprocessor MasPar 2216 16 000 processor system, and a Silicon Graphics Origin 2000 multiprocessor system. Subsequent alignment of the clusters has been performed using the simulated annealing approach of TIGR_MSA-contig, a sensitive code developed at TIGR for EST alignment. We have processed the resulting aligned ESTs using a combination of two error analysis systems, CRAW and CONTIGPROC.

The resultant consensi have been collected into a qualitated-error Sequence Tag Alignment and Consensus Knowledgebase (STACK) made up of all publicly available expressed human genes.

For production of extended consensi, sequences are put into loose groups by similarity threshold and then further segmented into sub-clusters. Alternate splice forms and alignment errors are isolated but can be viewed in the context of the entire sampled gene. We decouple the representative sequence generation and error-checking from the actual sequence clustering. The decoupling allows the introduction of higher error sequence into the consensus construction resulting in broader gene sequence sampling.

4 Methods

Clustering of ESTs

EST data does not share the characteristics of most DNA sequences found in full length entries in GenBank. Clustering of ESTs requires that a clustering method be highly tolerant of error, inconsistencies and re-arrangements. The system must be able to assign ESTs to clusters based on a statistic that reflects the properties of the data, and be able to align large numbers of highly identical stretches of DNA bounded by very low quality sequence. The resulting consensus has to be generated according to a set of rules that reflects the highly variable nature of the data.

D2-cluster

Use of a high performance method termed D2-cluster, which does not use alignment in order to make clusters (Burke, Davison, Hide in prep), allows successful incorporation of ESTs into a cluster, even if they do not have a long region of overlap. The algorithm relies on the presence of multiple identical words within each EST, and if the identical words produce a similarity above a statistically defined threshold, the algorithm assigns an EST to a parent cluster. Every sequence begins in its own cluster and the final clustering is made through a series of mergers. D2-cluster is an agglomerative clustering method appropriate to single read sequence data.

Word-based methods such as D2-cluster can be used to identify regions that can align well by use of a transfer function between word-similarity and alignable similarity.

Alignment and EST assembly

Sequences have been aligned for STACK using TIGR_MSA-contig, a high performance simulated annealing application written by Granger Sutton (Institute of Genome Research) and Tim Bussey (formerly MasPar Computer Corporation), that is tolerant of error and can align ESTs. We have found that extant algorithms that have an assembly approach such as TIGR_ASSEMBLER (Sutton et al. 1995) or the PHRAP package written by Philip Green at the University of Washington, tend to produce larger numbers of smaller clusters (less ESTs) because they are stringent and require similarity between sequences at their ends (overlap) These approaches are not as effective for production of extended consensi with error prone data. Our strategy has been to develop a non-alignment based engine that is devoted to EST clustering and to complement the core engine with available assembly and alignment systems.

Consensus Sequence Representation

The database system that results differs markedly from indices such as TIGR Gene Index (`http://www.tigr.org/tdb/hgi/`), and also databases of clusters of ESTs such as UniGene (`http://www.ncbi.nlm.nih.gov/UniGene/index.html` and Bogusky et al 1995) because of its organisation. Records are designed to be useful to the gene discovery researcher. Each record contains a header that explains the source of the consensus, the degree of matching of the ESTs to the consensus, and describes the coverage of the consensus (Figure 2, 3).

Processing of the 040896 GenBank format release of dbEST using STACK_PACK

(1) The first processing step is the conversion of the GenBank sequence files into FASTA format and the division of the complete database into organism directories and tissue/library files for each organism. All sequences with the same tissue or library name are grouped into files based on that name.

(2) Each organism subdirectory contains a myriad of files named exclusively by tissue type or clone library as specified in the original GenBank source files. The filenames are essentially random and must be grouped by hand into related tissue subdirectories. We have defined the hierarchy shown in Table 1 based on tissue relations and limits on the number of sequences which can reasonably be clustered in a single run with current resources.

(3) Files in the hand-organized subdirectories from step (2) are then concatenated into single tissue files.

(4) Sequence files from step (3) are then masked against vector and human repeat sequences (VecBase and RepBase accessed from NCBI November 1996).

(5) Files of masked EST sequences are transferred to a high performance architecture such as a MASPAR or SGI ORIGIN 2000 for clustering where they are processed by MPD2_CLUSTER or D2-CLUSTER and BUILD_CLUSTERS. About 45-55% of the single EST sequences subsequently form clusters containing two or more EST sequences (Table 2).

(6) For the STACK version 1.0 project, each individual cluster is further processed by TIGR_MSA_CONTIG on a MASPAR to generate alignment and assembly information in GDE format. We have used PHRAP in some cases at this step. Alignment can be highly problematical, as sequence quality varies greatly. Some clusters cannot be processed because of limitations on performance

tissue name	5’	3’	total	contents include
adipose	123	79	672	brown, white
connective	3524	3416	7631	bone, fibroblast, skin
digestive	422	522	1686	colon, gall bladder
disease-duplicates	10714	11142	23070	copies of all disease related
genomic	777	3403	7767	chromosome, clone sequences
glands	17370	12602	31640	breast, endocrine
muscle	0	0	7122	leg, skeletal, pectoral
nervous/brain	48194	41473	117132	fetal, infant, adult
nervous/eye	5389	4559	15036	retina
nervous/cochlea	1219	3158	4377	fetal cochlea
nervous/olfactory	951	1649	2600	olfactory epithelium
nervous/synovial	134	0	134	synovial membrane
other	11115	11786	22957	melanocyte, monocyte
reproductive	29430	21602	52150	genital, embryo, placenta
resp-circ/heart	18648	9255	27903	aorta, fetal heart
resp-circ/hemato-lymph	57702	55549	113721	blood, liver, kidney, lymph
resp-circ/lung	12532	10857	23391	fetal, adult,
totals:	222351	191257	470280	

Table 1: Description of tissue-types and constituent sequence numbers for version 1.0 of STACK. Tissue cluster types were named according to groupings that commonly represent classes of tissue. Eg; Digestive tissues are grouped to include colon and gall bladder.

Tissue	bases	sequences	clusters
adipose	238069	672	640
connective	3089144	7631	5004
digestive	351508	1686	1601
disease	8652083	23070	15468
genomic	1850653	7767	7012
glands	11153719	31640	18395
muscle	1922538	7122	3204
brain	41876662	117132	46825
eye	6921349	15036	10426
olf.epithelium	898739	2600	1740
fet.cochlea	1476881	4377	2730
synovial membrane	40007	134	123
other	9295815	22957	12406
reproductive	18671171	52150	24415
heart	13238079	39194	19518
hemato-lymphatic	41442595	113721	52454
lung	8304608	23391	14010
totals:	169423620	470280	235971

Table 2: Numbers of clusters produced using d2-cluster on EST sequences divided into tissue types by clone annotation.

tissue	problem clusters	lost sequences
glands	1	174
muscle	1	608
brain	5	2262
eye	1	143
reproductive	6	2143
heart	2	1634
hemato	12	11386
totals: 28	18350	

Table 3: Numbers of problematical clusters produced by D2-cluster on dbEST 040896.

tissue	1-sequence clusters	multi-seq clusters	sequences in multi-seq clusters
adipose	626	14	30
connective	3849	1155	3686
digestive	1556	45	98
disease	11920	3548	10855
genomic	6619	393	1081
glands	13769	4623	17277
muscle	2624	579	3617
brain	27679	19141	85622
eye	8605	1818	5938
olfactory	1465	275	896
cochlea	1987	743	2302
synovial	114	8	19
other	8383	4021	14316
reproductive	17125	7282	31985
heart	14237	5277	22753
hemato-lymph	22457	11129	50142
lung	10977	3032	11880

Table 4: Cluster formation in dbEST using D2-cluster.

of TIGR_MSA_CONTIG. We have subsequently processed large and problematical clusters by hand alignment (Table 3).

(7) Successful clusters in step (6) are concatenated together and the resulting file is processed by CONTIGPROC.PL, which invokes CRAW (John Burke, University of Houston) on each cluster to evaluate it for alignment quality and presence of subclusters. CONTIGPROC.PL generates consensus sequence and assembly information in GIO format (used by Genome Sequence Database at National Centre for Genome Resources), consensus and optional high-quality consensus information in FASTA format, and/or assembly information in GDE format for each cluster.

(8) The original sequence files from step (4) are processed by CLONELIST to extract clone IDs and sequence accession numbers, while the cluster files from step (6) are processed by CONTIGLIST to extract cluster IDs and clustered sequence accession numbers. The results of these two programs are processed by XCLUST2.PL to generate a list of clone-ID-linked clusters.

(9) JOIN.PL combines individual clusters from step (7) according to the result list from step (8) to generate a set of clone-linked clusters in GIO and/or FASTA format output files.

The sequence alignments are processed to generate consensi, and error checking and compensation is performed at this stage using CRAW. CRAW takes an alignment as input and characterises variation within each cluster. If there is significant variation of sequences, it divides the cluster into alignable sub-clusters and outputs maximum agreed subconsensi groupings. These are then processed for similarity and characterised. The most frequent class of output sub-consensi result from mis-alignments of the clustered ESTs.

Good consensi are identified using CONTIGPROC which sorts the best output consensus according to:

(1) number of ESTs assigned to consensus

(2) number of ATCG bases in consensus

(3) (lesser first) number of VHDB (iupac not-T, not-G, not-C, not-A) bases in consensus

The next step is clone linking which generates sequences linked by 20-N stretches. The ordering for the sequences is (1) 5'/other/3' assignment, (2) order of the cluster-ids.

5 Results

Clustering of dbEST produces a set of "single sequence clusters" which have no matches with other clusters in the database, "multiple sequence clusters", which are assigned to share a cluster based on high sequence similarity, "single consensus clusters" which contain one clear consensus when aligned, and "multiple consensus clusters" which reflect the low quality information in the sequence and generate more than one consensus (Table 5). The resulting consensus alignments form the basis for records in STACK. Comparisons of cluster distribution with UniGene (Bogusky and Schuler, 1995) demonstrate that STACK demonstrates a similar distribution of total numbers of multiple sequence clusters, in some cases, such as connective tissue and heart, far exceeding those for UniGene (Figure 1). The discrepancy is a result of the clustering method. UniGene originally relies on 3' EST clustering only, followed by 5' clone linking. D2-cluster performs 3' and 5' clustering, followed by clone linking and does not have alignment dependency. STACK has a significantly larger number of input sequences into the clustering process (Table 6), which can positively impact the resulting cluster consensus length and quality.

Gene Representation

STACK has been generated in order to provide viewable alignments of EST clusters and an assembly of ESTs that provide extended consensi. The expressed gene sequence data that results is collated into "gene-sets". Each STACK entry contains all available expressed sequence data from a particular gene (Figure 2)

The sequence representation method allows accurate representation of consensi where normally it would be necessary to discard the consensus due to error. As a result, the average length of all records in the dataset exceeds 510 bases. Clustering of non-redundant records from STACK with the latest dbEST release, and with data released from other projects will allow STACK to make an even more representative contribution to the genome projects.

STACK and STACK_PACK represents a unique, multi-platform EST clustering system that has broad application for laboratories requiring clustering of EST data for combination into the public data. The resulting composite linked clusters provide a powerful discovery resource. STACK is curated by the South African National Bioinformatics Institute, to which inquiries should be addressed for errors, additions and distribution requests.

tissue	single consensus clusters	% of total clusters	mult consensus clusters
adipose	13	93	0
brain	16,195	85	526
cochlea	669	90	5
connective	987	85	20
digestive	43	96	0
disease	3,084	87	42
eye	1,392	77	43
genomic	363	92	1
gland	3,914	85	96
heart	4,448	84	150
hemato	11,895	79	484
lung	2,497	82	94
muscle	545	94	10
olf eptihelium	248	90	0
other	3,382	84	77
reproductive	5,697	78	264
synovial mem.	8	100	0
tigr	8,506	93	132
totals:	63,886	84	1,944

Table 5: Clusters which have more than one consensus sequence.

Tissue	SANBI sequences	UniGene sequences	SANBI MS clusters	UniGene clusters
adipose	672	202	14	77
brain	117,132	73,167	19,141	15,492
fet.cochlea	4,377	2,144	743	870
connective	7,631	139	1,155	4
digestive	1,686	351	45	65
eye	15,036	6,346	1,818	1,883
glands	31,640	19,190	4,623	4,370
heart	39,194	3,528	5,277	104
hemato-lymph	113,721	86,444	15,071	15,649
lung	23,391	9,504	3,032	1,948
olf.epithelium	2,600	2,521	275	775
reproductive	52,150	48,866	7,282	9,924
totals:	409,230	252,402	58,476	51,161

Table 6: Distribution of sequences input into STACK and Unigene present in clusters per tissue type.

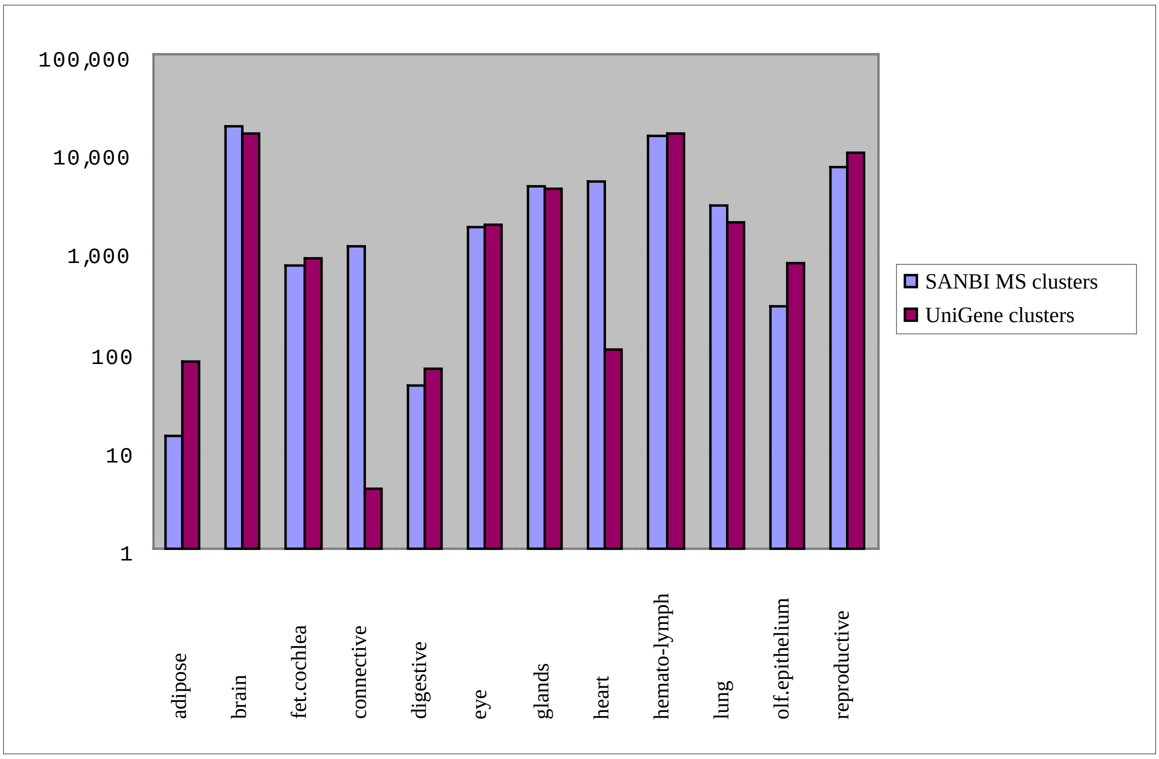

Figure 1: Log comparison of tissue distribution of numbers of multiple sequence clusters in STACK (MS clusters) and UniGene.

STACK distribution

The files are available at `http://ziggy.sanbi.ac.za/stack/stackrequest.html` and have been submitted to NCGR for inclusion in GSDB.

References

[1] Aaronson, J.S., Eckman, B., Blevins, R.A., Borkowski, J.A., Myerson, J., Imran, S., Elliston, K.O., Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data, Genome Res. 6(9), 829–845 (1996).

[2] Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D., White, O., Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence, Nature 377 (6547 Suppl), 3–174 (1995).

[3] Houlgatte, R., Mariage-Samson, R., Duprat, S., Tessier, A., Bentolila, S., Lamy, B., Auffray, C., The Genexpress Index: a resource for gene discovery and the genic map of the human genome, Genome Res. 5(3), 272–304 (1995).

[4] Torney, D.C., Davison, D., Burkes, C., Computation of d2: a measure of sequence dissimilarity, Computers and DNA, SFI Studies in the Sciences of Complexity, Vol. VII, Eds. G. Bell and T. Marr, Addison-Wesley (1990).

```
> 9856-0-eye-001-1997-0.1
COVERAGE: 0.9247 OTHER_CONSENSI: 3 ASSIGNED: W28267 W2
2290 W28033 W22856 W22548 W22633 W27813 W21991 W22259 W27212
THYKKKGNWHNMHVWSMHWDKHNNSSHSVBSRBBBTBBKSSTHWTCMSHSVDWMRDGBSCY
SSYYCRRGTYWYYKCCYCCCTGRGKAMGGSBBWCSBVVVVADSMMYWMCYMCMMYTRSGKG
GASGSYKKCWHSRYSGKGVCAGACCATGTTCTCCCTSYTGGTGACGGGAAAGCTGAAGCSC
TACTTCACGGRCCTAGAGGCCTTGGCCATGGYCMCTSCTGCTTTCTBCCATGACATTGACC
ACAAGAGGCACCAATADCCTCTACCAGATGAAATCCCAGAACCCACTGGSCAAGCTCCATG
GGTCCTCTATCTKGGWAAGACACCACTTGGAGTTTGSCMAARCACTGCTCAGAGACGAGAG
CCTGAATATCTTTCMAAACCTCAATCGTCGACAGYATGAGCATKCCATCCACATGATGGRC
ATTGCAATCATTKBCACAGACCTCGCCYTGTRTTTCAAGAAGAGGACGATGTYCCMAWAGW
TCSBGGRTCARTCTWAGACATWTKAGAGTGAACAGGRGTRRASAMMRTRMWKKWKGMKGRA
GMMRASRMGGRRRGWMRKYKTTWKGSCMWKRATGRWKRMCSCYYKTKHKCTCTYWKCMAKC
AMCAAMCMMWSSSWKGKGSRGRGSSARGKRSCWSKGTTKTGRYTKYSMKKYWRKTSSCWWR
SRGGKKSKGGKVHMHSYNGKTYCYRGKKKKKWYYAAAGTCAANGAGGGTTGKKKNTTATNT
NAAGNCCAGGTTYCMGGACCCAGTTCAACCTNGGTTCCCAYYYCCCCNTTTCCAAAGAAAA
GGGNTTCATTTTCGGNTTTNTCNAGNC
```

Figure 2: Representation of a cluster of ESTs in consensus format. COVERAGE: 0.9247 is the average of the consensus for a cluster for each called base in sequence. OTHER_CONSENSI: 3 CRAW generated three other possible consensus, based on the degree of error in the sequence and possible alternate splicing. The top consensus has been selected for representation in the record. ASSIGNED: describes the ESTs assigned which have provided good consensus data. These ESTs match where the called bases are shown.

[5] Sutton, G., White, O., Damas, M., Kerlavage, A., TIGR assembler: A new tool for assembling large shotgun sequencing projects, Genome Science and Technology, 1995.

[6] Boguski, M.S.and Schuler G., ESTablishing a Human Transcript Map, Nat. Genet. 10(4) 369–371 (1995).

[7] Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., Matsubara, K., Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression, Nat. Genet. 2(3), 173–179 (1992).

[8] Hide, W., Burke, J., and Davison, D.B., Biological evaluation of d2, an algorithm for high-performance sequence comparison, J. Comput. Biol. 1(3), 199–215 (1994).

```
> eye2 TOTAL_ESTS: 29 COVERAGE: 0.9532
CLONE_LINK_OF: 1002 989 4171 6496 6692 7627 8425 14982
WTCKGCACAGGNATCTGACTTTAAAAATTATTCTAGAATTTCTGTGCTTCAATATTAATGC
CAGAAGACTTGGAATTGTTTATTTGTAGGTAACTGCCTTTAAGGAAACTTGACCAAATATT
AACTAAGTTATGTATTTCCTTTTGGCAACAGTTGTGACTTCTCACCAGGAGAKTTGGTTTG
GGCCMMRRWGGRGGGTTACCCCTGGKKGCCTTGTCTGGTTTACAACCACCCCTTTGATGGA
ACATTCATCCGCGRKAAAGGGAAATCAGTCCGTGTTCATGTACAGTTTTTTGATGACAGCC
CAACAAGGGGCTGGGTTATYAAAAGGCTTTTAAAGCCATATACAGGTTCAACTCCCTTCCC
CCTTCCSCCACCAAAAAAATAAARSMAGGGCACGSCGKKKYTTTACCTGTWAAWTCCTAGS
TTACCTAAGGAGGSTTGACACGAAGAGGTCTKTYCNYGGGGTWACMGAGGCMAGRCACTGT
YTWRWWMRMWAAWTYYTKTKYKMKATATTAAAGACTGAAGAAAGGCCAGGCGCAATGGGTC
ATGCCNNNNNNNNNNNNNNNNNNNNATCGAACAAAAnnnnnnnnnnTACAGGTAAGCACCG
GCGTGCCCTGCnnnnnnnnnnAAGGGAGTTAACCTGTATATGGCTTAAAAGCCTTTTNAAA
CCCAGCCCCTTTTTGGNTGTCACAAAAAACTGACATGACACGNNNNNNNNNNNNNNNNNNN
NAAAAAATAAAGCRGGGCACGCCGGTGCTTACCTGTAAACCCTAGCTACCTAAGAGGCTGA
CACGAGAGGnnnnnnnnnnCCYTTTKKTNNNNNNNNNNNNNNNNNNNNCAAAAnnnnnnnn
nnCTCGTGTTCAGCCNCTTAGGGNAGGCNAGGGATTTACAGGNAAGCACCNGCGTGCCCTN
TTTTNNNNNNNNNNNNNNNNNNNNNCTCTCGTGTCAGCCTCTTAGGNAGCTAGGATTTACA
GGNAAGCACCGGCGNGCCCNGCTTTATTNNNNNNNNNNNNNNNNNNNNTACAGGNAAGCAC
CGGCGTGCCCTGCTTTATTTNTTTGGTGNTGGNANGGGGGAANGGAAGTTGAAACCTGTAN
ATGGGCTTNAAAAAGCCCTTTTGATAACCCCAGCCCCCTTGTTGGGGCTGGTCATCAAAAA
ACTGGACATGAACACGGACTGAATTCCCCTTCTCGCGGANGAATGNTCCNTCAAAAGGNNN
NNNNNNNNNNNNNNNNNCAAAAGGnnnnnnnnnnTACAGGTAAGCACCGGCGTGCCCTGGC
TTTAATTTTTTGGGGGGGNAAAANGGGGAAAGGAAGTTGAANCCGGTAAAGGGCCTTAAAA
ANNNNNNNNNNNNNNNNNNNNCTTTCTTCAGTCTTTAATAATCGAACAAAAnnnnnnnnnn
TTACAGGTAAG
```

Figure 3: Representation of a SANBI linked cluster. The cluster record is a consensus of several EST sequences. Each sequence that comprises the cluster can be found in a separate alignment file, and the identities of the sequences can be found in a separate table. The record comprises of joined consensi from 8 clusters generated by D2_CLUSTER. The average of the coverage scores for the clusters, 1002 and 4171, is 0.95. The 8 clusters joined are given following the identifier: CLONE_LINK_OF:. The 8 consensus sequences follow the FASTA header line, each separated from the previous by a sequence of 20 'N's. Within an individual cluster consensus, long regions of N's or X's are replaced by a single sequence of 10 'n's; this is shown in the second incorporated sequence, (singleton) cluster 989.