Transcript analysis and reconstruction

Brazil 2001

SANBI

Genes

Why are there only a few tens of thousands of genes in the human genome?

How do genes express themselves to manufacture the proteome?

How can available sequence information be processed in order to deliver understanding of gene expression?

Genomic expression

Within eukaryotes, genes have shared basic characteristics. They have single or multiple exons and introns distributed along the gene in coding and non-coding regions with

5’ Flanking region with transcription regulation signals

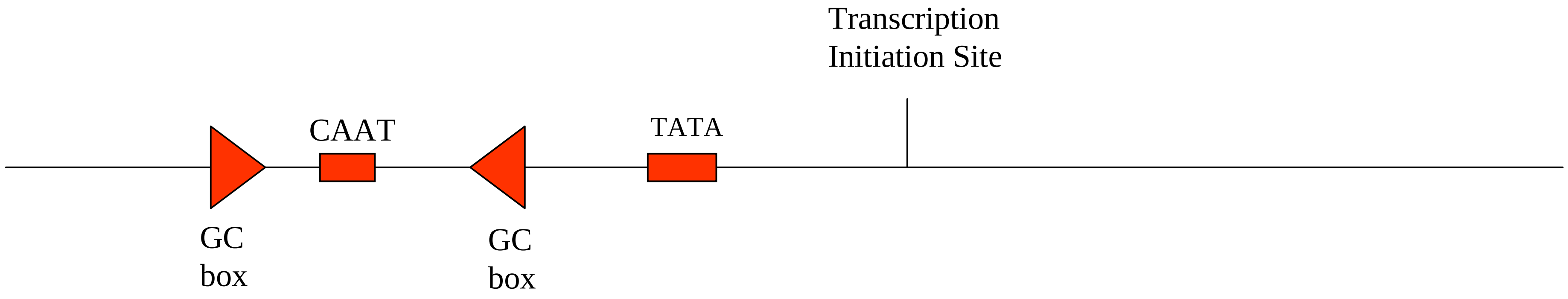
Transcription initiation start site (5’)

Initiation codon for protein coding sequence

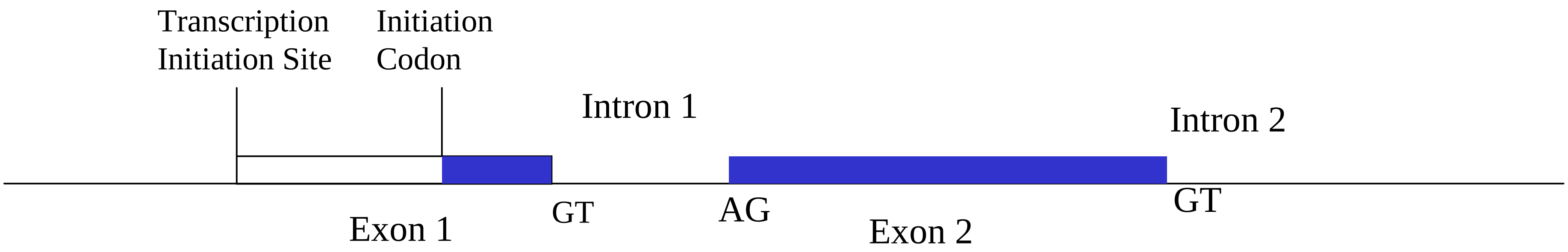
Exon-intron boundaries with splice site signals at the boundaries

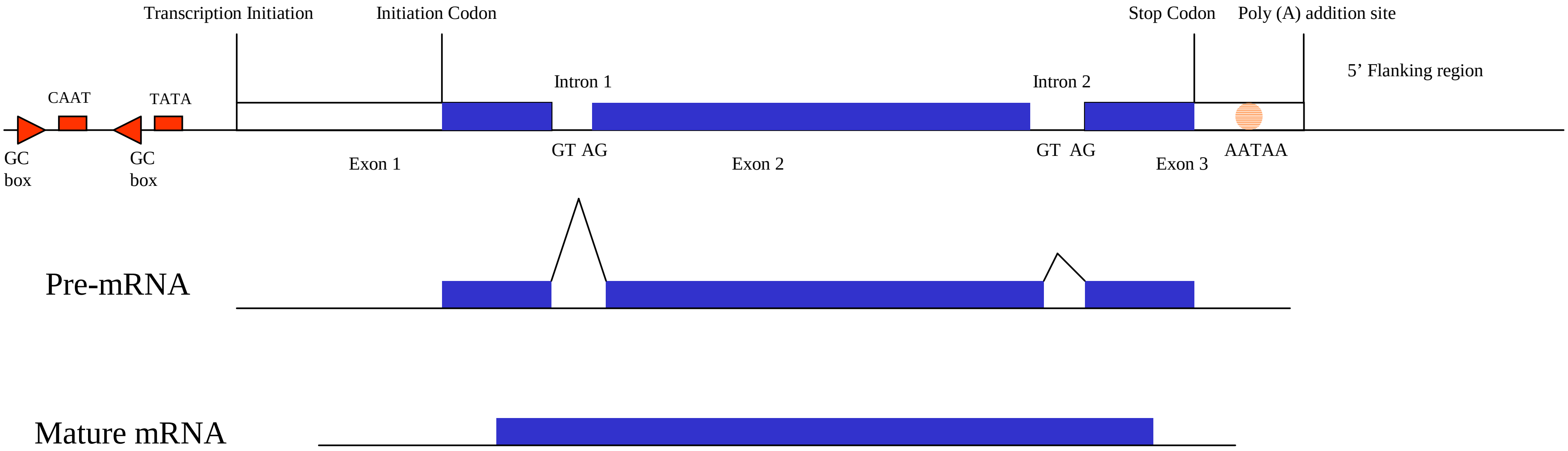
Termination codon for protein coding sequence

3’ signals for regulation and polyadenylation



SANBI


Transcription
Initiation Site
Initiation
Codon
Intron 1
Intron 2
GT
AG
GT
Exon 1
Exon 2


Transcription Initiation
Initiation Codon
Stop Codon
Poly (A) addition site
5’ Flanking region
Intron 1
Intron 2
CAAT
TATA
GC box
GC box
Exon 1
GT AG
Exon 2
GT AG
Exon 3
AATAA
Pre-mRNA
Mature mRNA

SANBI

Gene Expression

Transcription products can vary.

Transcription initiation at the start site (TSS)

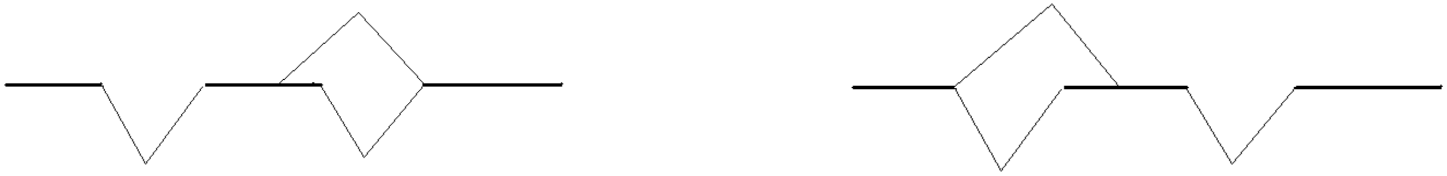
Exon length

Exon prescence/absence in the mature transcript

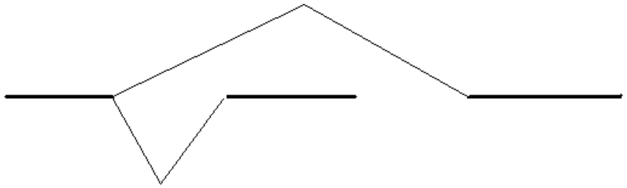
Alternate transcription termination and polyadenylation

Examples of alternative splicing

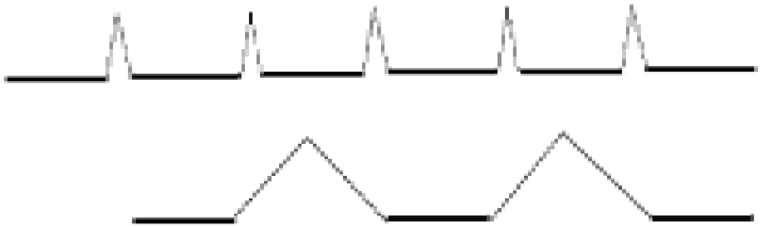
Alternative donor and acceptor splice sites

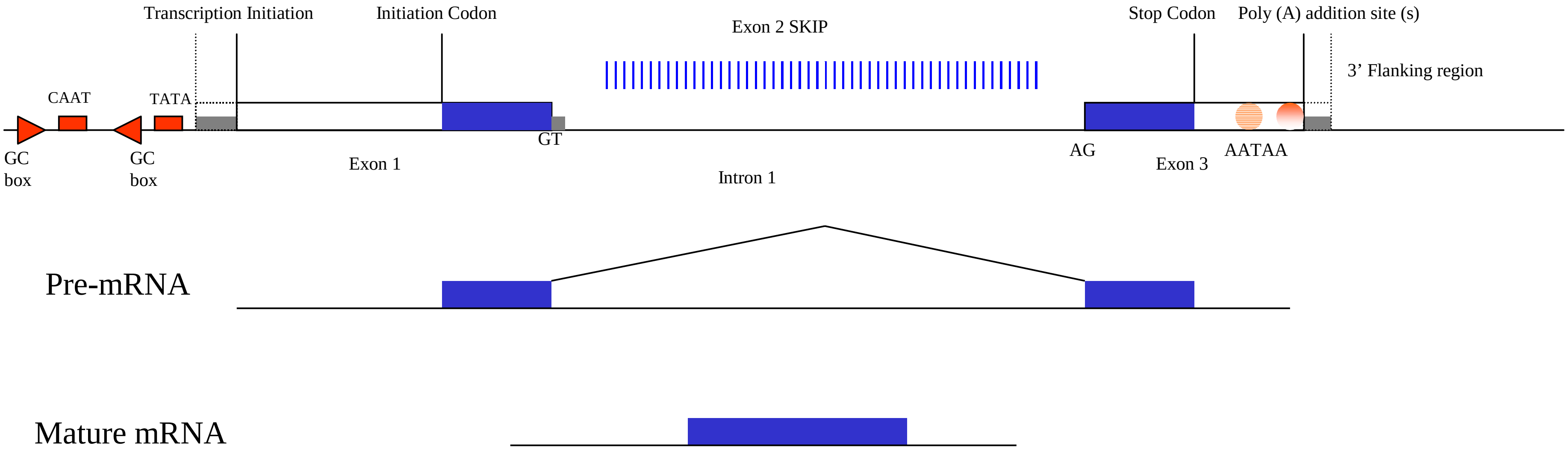


Alternative polyadenylation



Exon skipping




Transcription Initiation
Initiation Codon
Exon 2 SKIP
Stop Codon
Poly (A) addition site (s)
3’ Flanking region
CAAT
TATA
GC box
GC box
Exon 1
GT
Intron 1
AG
Exon 3
AATAA
Pre-mRNA
Mature mRNA

SANBI

Capturing expressed transcripts

Databases - Sequences

dbEST

Several collapsed datasets

TIGR-THC

Unigene

STACK

Allgenes

BodyMap

Several more specialised

Genome Sequence as it appears:

Expression Capture

- Serial Analysis of Gene Expression
  - – DNA fragments that act as unique markers of gene transcripts.

  - – Assay of numbers of each marker in a set of sequence yields a measure of gene expression
- Array
  - – Laydown of sequence clones to provide an organised series for hybridisation

SANBI

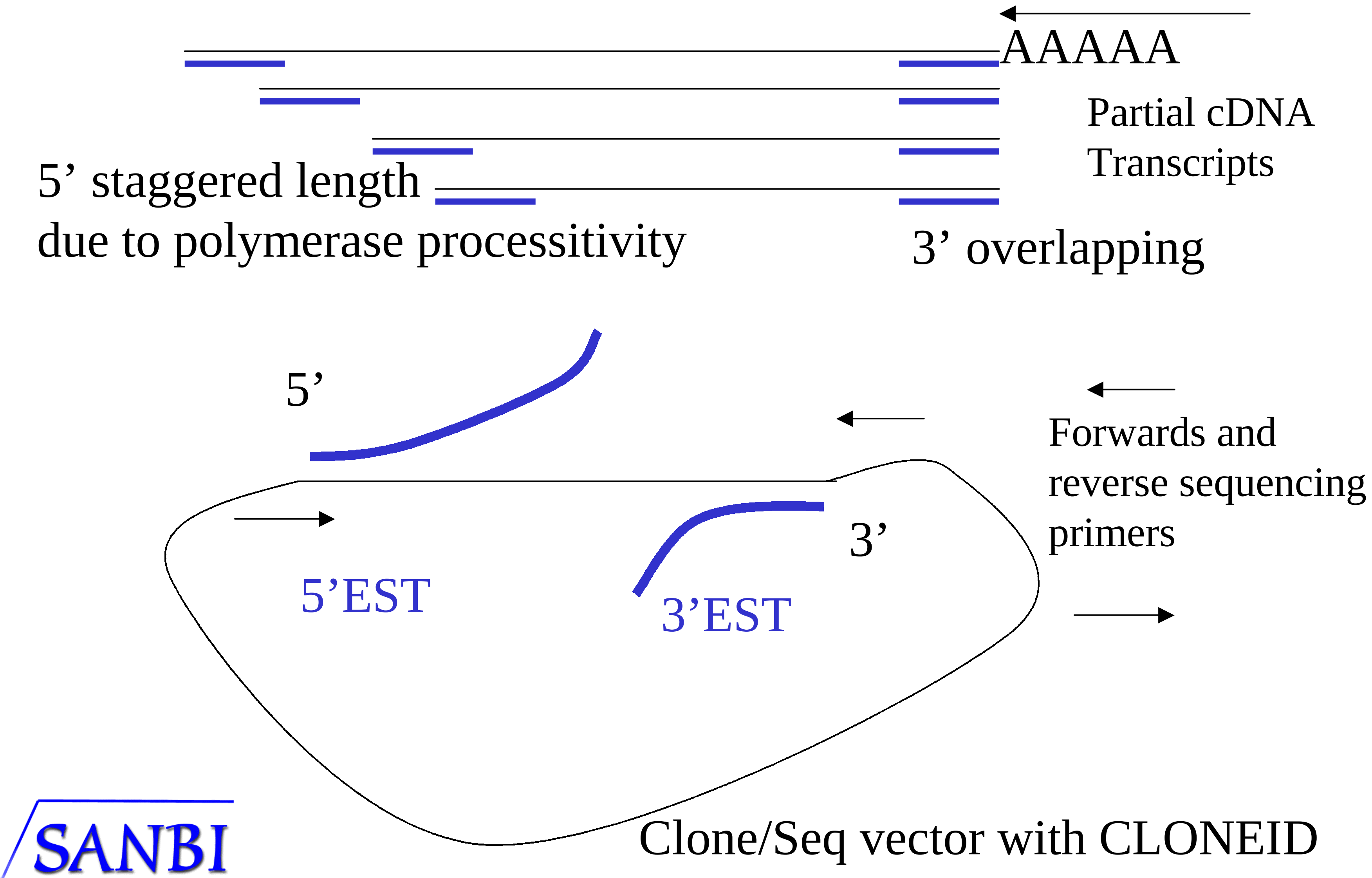
Resolution of Captured Expression

ESTS Low resolution, broad capture, provides template for SAGE and Array

SAGE Medium resolution, need template, noise can be an issue, stoichiometry is revealed but standardisation a problem

ARRAY High resolution, need template, noise, stoichiometric resolution highest, standardisation a problem.

What is an EST?



AAAAA

Partial cDNA
Transcripts

5’ staggered length
due to polymerase processitivity

3’ overlapping

5’

3’

5’EST

3’EST

Forwards and
reverse sequencing
primers

Clone/Seq vector with CLONEID

SANBI

What potential do ESTs hold?

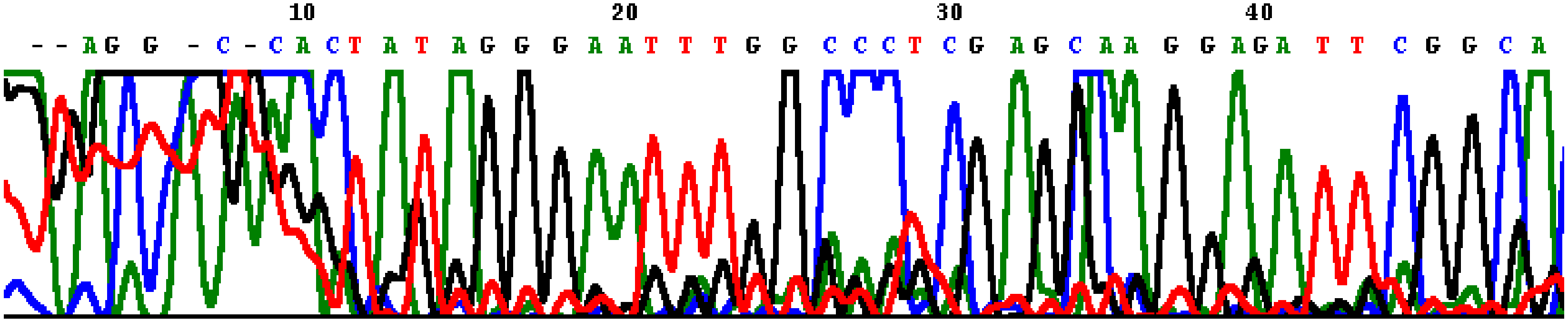
- Expression counts
- Consensus sequences
- Alternate expression-form characterisation
- Identification of genes expressed in a pilot gene discovery project
- Identification of genes specifically expressed in a chosen library or tissue

Use of Transcripts in Completed genomes

- Identification of genes
  - Exon boundaries
  - Alternate transcripts
- Genomic annotation
  - Expression sites of encoded genes
- Comparitive genomics

EST data quality

```
>T27784   g609882 | T27784 CLONE_LIB: Human Endothelial cells. LEN: 337
b.p. FILE gbest3.seq 5-PRIME DEFN: EST16067 Homo sapiens cDNA 5' end
AAGACCCCCGTCTCTTTAAAAATATATATATTTTAAATATACTTAAATATATATTTCTAATATCTTTAAAT
ATATATATATATTTNAAAGACCAATTTATGGGAGANTTGCACACAGATGTGAAATGAATGTAATCTAATAG
ANGCCTAATCAGCCCACCATGTTCTCCACTGAAAAATCCTCTTTCTTTGGGGTTTTTCTTTCTTTCTTTTT
TGATTTTGCACTGGACGGTGACGTCAGCCATGTACAGGATCCACAGGGGTGGTGTCAAATGCTATTGAAAT
TNTGTTGAATTGTATACTTTTTCACTTTTTGATAATTAACCATGTAAAAAATG
```



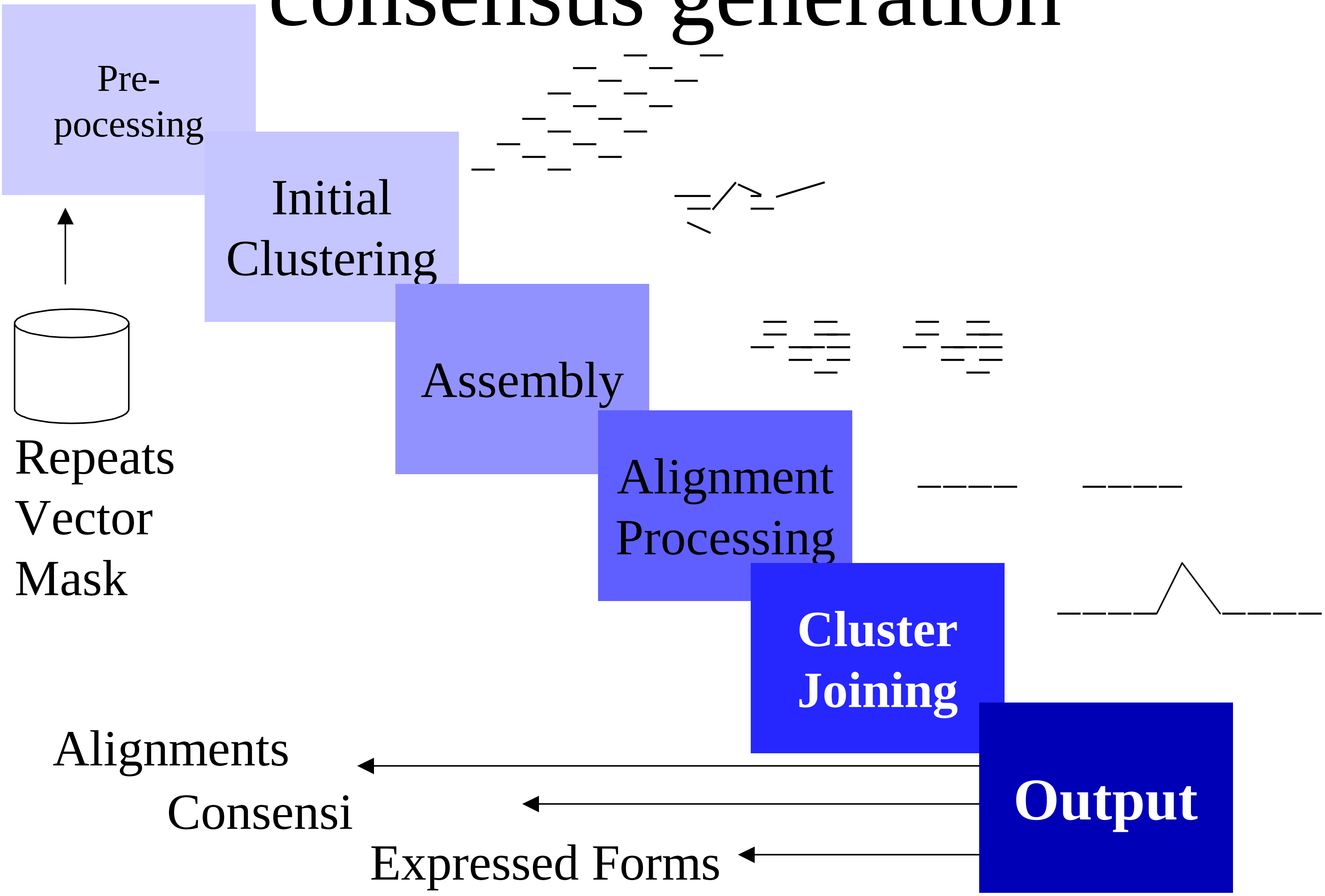
EST is Poor Quality data with contaminants

Vector

Repeat MASK

Individual items are prone to error but an entire collection contains valuable genetic information

Overview of clustering and consensus generation



Transcript reconstruction

What is an EST cluster?

A cluster is fragmented, EST data and (if known) composite exon transcript sequence data, consolidated, placed in correct context and indexed by gene such that all expressed data concerning a single gene is in a single index class, and each index class contains the information for only one gene.

(Burke, Davison, Hide, Genome Research 1999).

SANBI

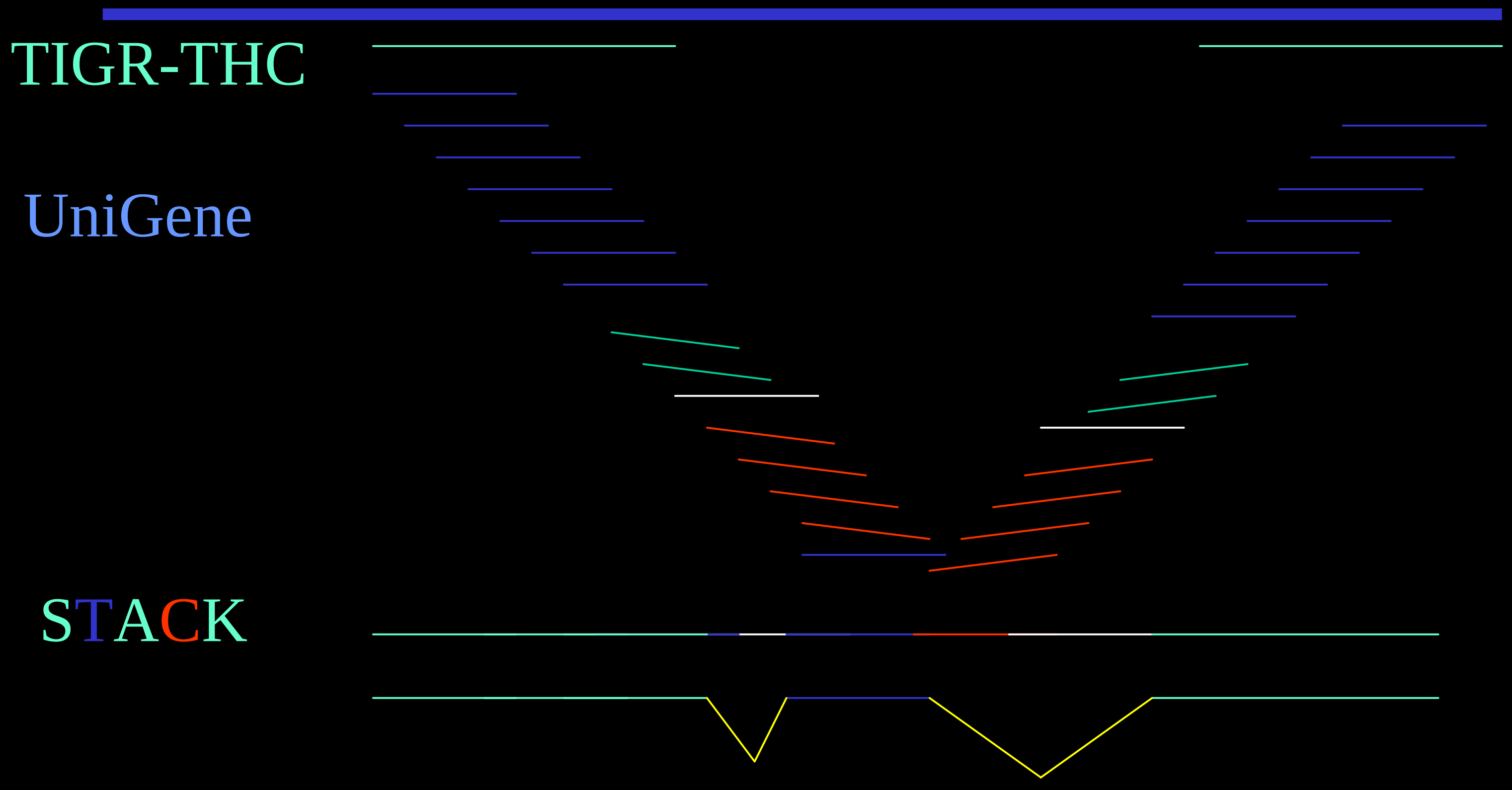
Loose and stringent clustering

- Stringent - greater fidelity, lower coverage
  - One pass
  - Shorter consensi
  - Lower inclusion rate of expression-forms
- Loose - lower fidelity, higher coverage
  - Multi-pass
  - Longer consensus sequences but paralogs need attention
  - Comprehensive inclusion of expression-forms

Supervised clustering

- ‘Template for hybridisation’ is a transcript composite derived from:
  - A captured ‘full length’ mRNA
  - A composite exon construct from a genomic sequence
  - An assembled EST cluster consensus

Clean Short and Tight



Long and Loose

Data apprehension and input format.

- Sources: In-House, Public, Proprietary
- ‘Accession’ / Sequence-run ID
- Location/orientation
- Source Clone
- Source library and conditions

Pre-processing

- Minimum informative length
- Low complexity regions
- Removal of common contaminants
  - Vector, Repeats, Mitochondrial, Xenocontaminants
  - XBLAST,
  - Repeatmasker, VecBase and others
  - BLIND masking
- Pre-clustering *vs* known transcripts (data reduction)

Initial clustering

- Stepwise clustering ‘Multistate’.
  - sequence identity
  - annotation
  - verification

Assembly

- Including chromatograms - SNPs and Paralogs
- PHRAP and CAP series
- Multiple assemblies can fragment from one input cluster
  - fidelity
  - alt. forms
  - error

Alignment processing

- Consensus generation
- Alternate forms
- Errors
- Choosing the ‘correct consensus’

SANBI

Cluster joining

- Clone joining
  - Choosing to accept a clone annotation
    - 1 clone ID
    - 2 clone ID’s
- Available parents
  - mRNA (incomplete/alternate)
  - Composite(constructed from Genomic)
    - intronic sequence ~ 2%

Output

- Alignment
  - alternate expression-forms
  - polymorphisms
  - error assessment
- Cluster
  - raw cluster membership
  - contextual links
- Formats: FASTA, GenBank, EMBL

Alignment scoring methods:

- Correct position of sequence elements against each other maximizes some score
- BLAST and FASTA
  - Heuristic
  - cutoff and identity
  - pairwise alignment
  - ~fast

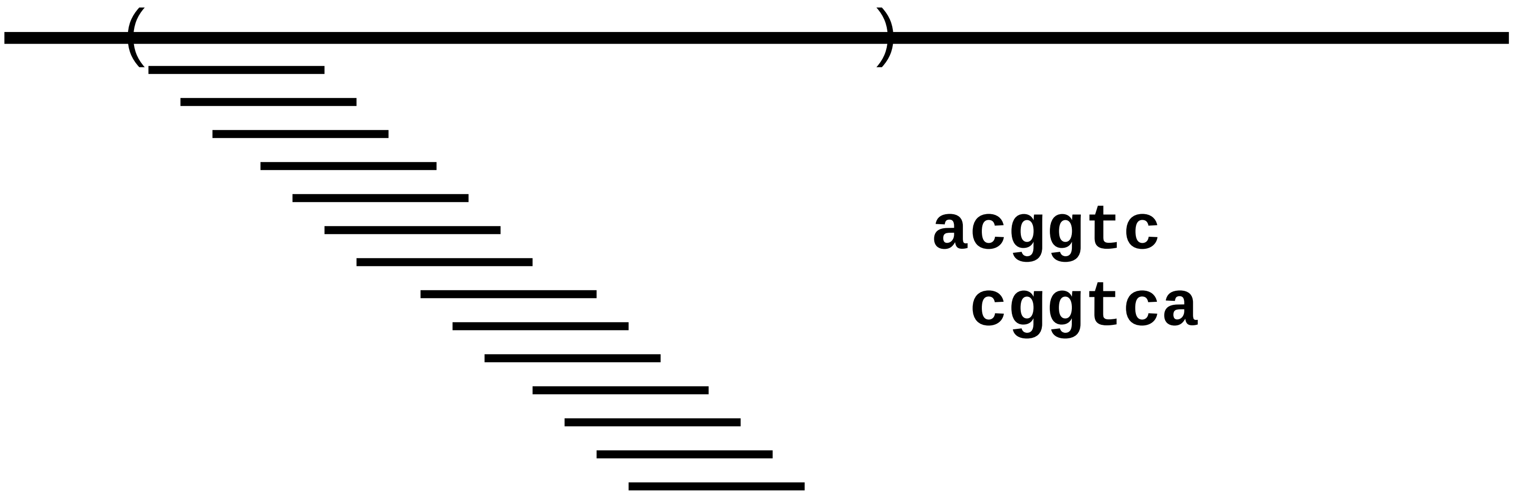
EST clustering methods

- Est sequence is littered with errors, stutters, in-dels and re-arrangements
- alignment approach is sensitive to these
- 3’ only comparison

Non-alignment based scoring methods: D2-cluster

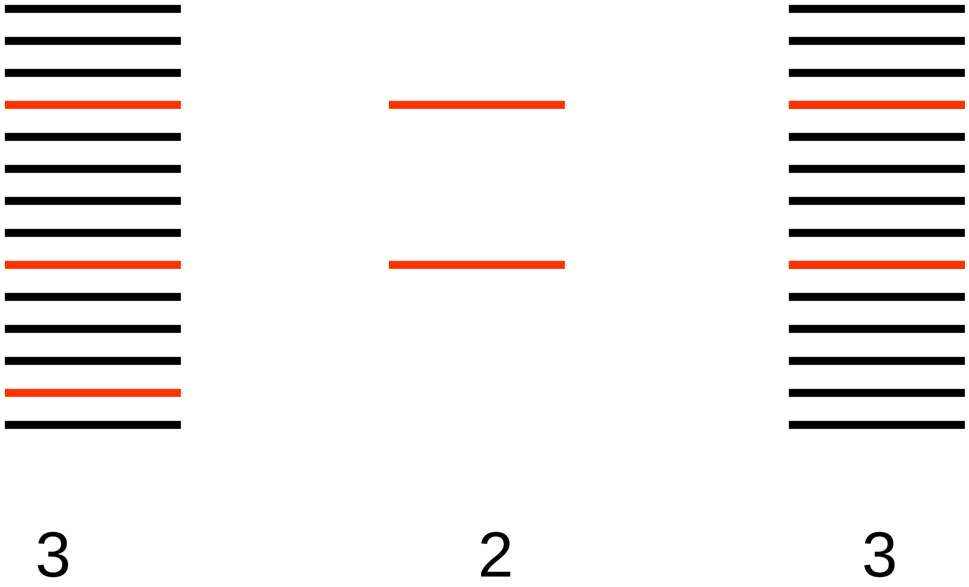
- No alignment so a speedup
- Sensitivity improved by multiplicity measure
- low weight to low complexity
- very error tolerant
- transitive closure
- 96% ID over 100 or 150 bases.

Word table



```
acggtc
 cggtca
```

Multiplicity comparison



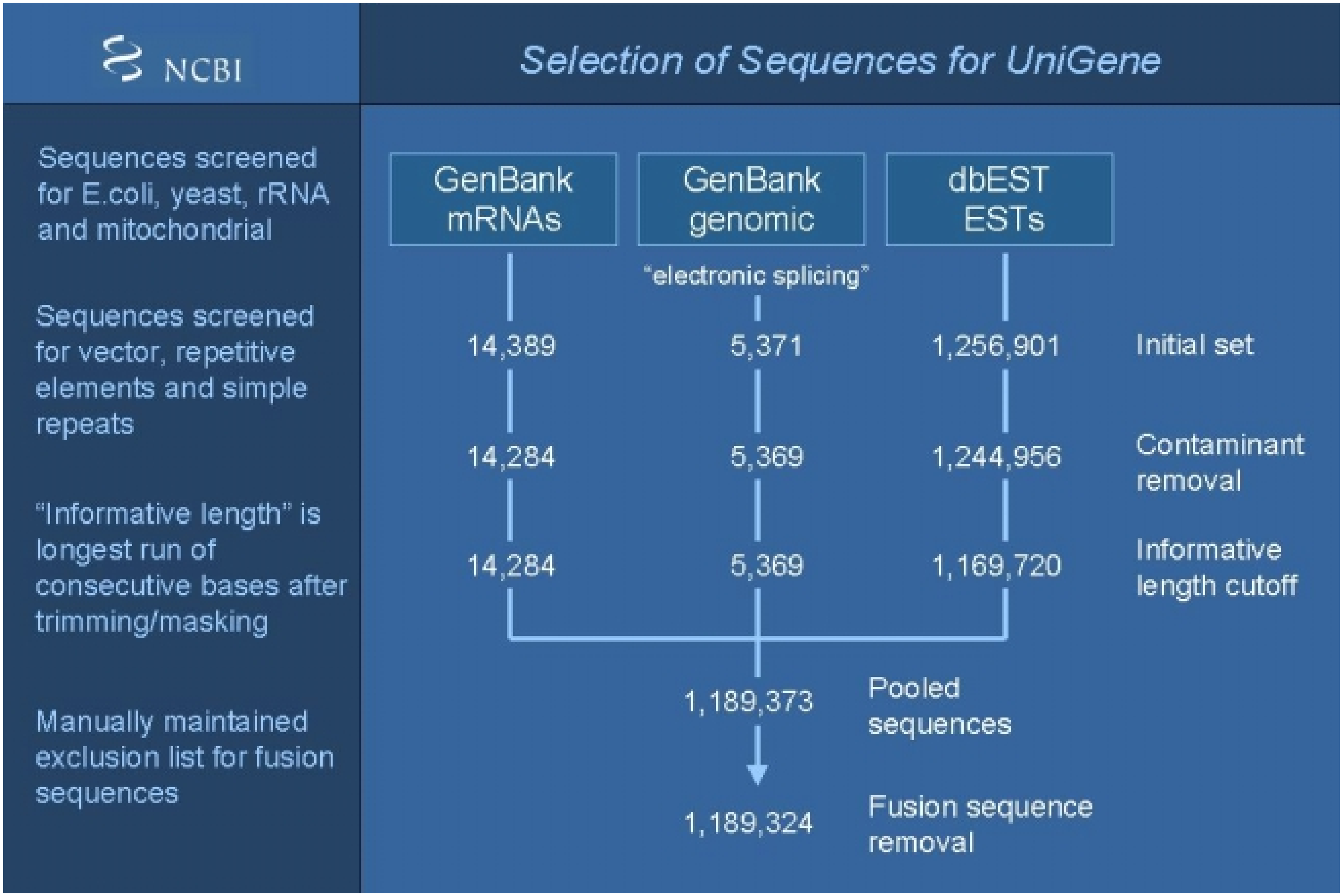
(d)^2= 4

SANBI

TIGR_ASSEMBLER

- THC_BUILD: BLAST-FASTA id all overlaps and are stored.
- Tigr-assembler then uses rapid oligo nucleotide comparison and assembles non-repeat overlaps. (95% ID over 40bp)
- matching constraints on sequence ends
- minimum sequence id within a sequence group - more fragmented as a result
- Other TIGR approaches are similar

UniGene



NCBI

Selection of Sequences for UniGene

Sequences screened for E.coli, yeast, rRNA and mitochondrial

Sequences screened for vector, repetitive elements and simple repeats

"Informative length" is longest run of consecutive bases after trimming/masking

Manually maintained exclusion list for fusion sequences

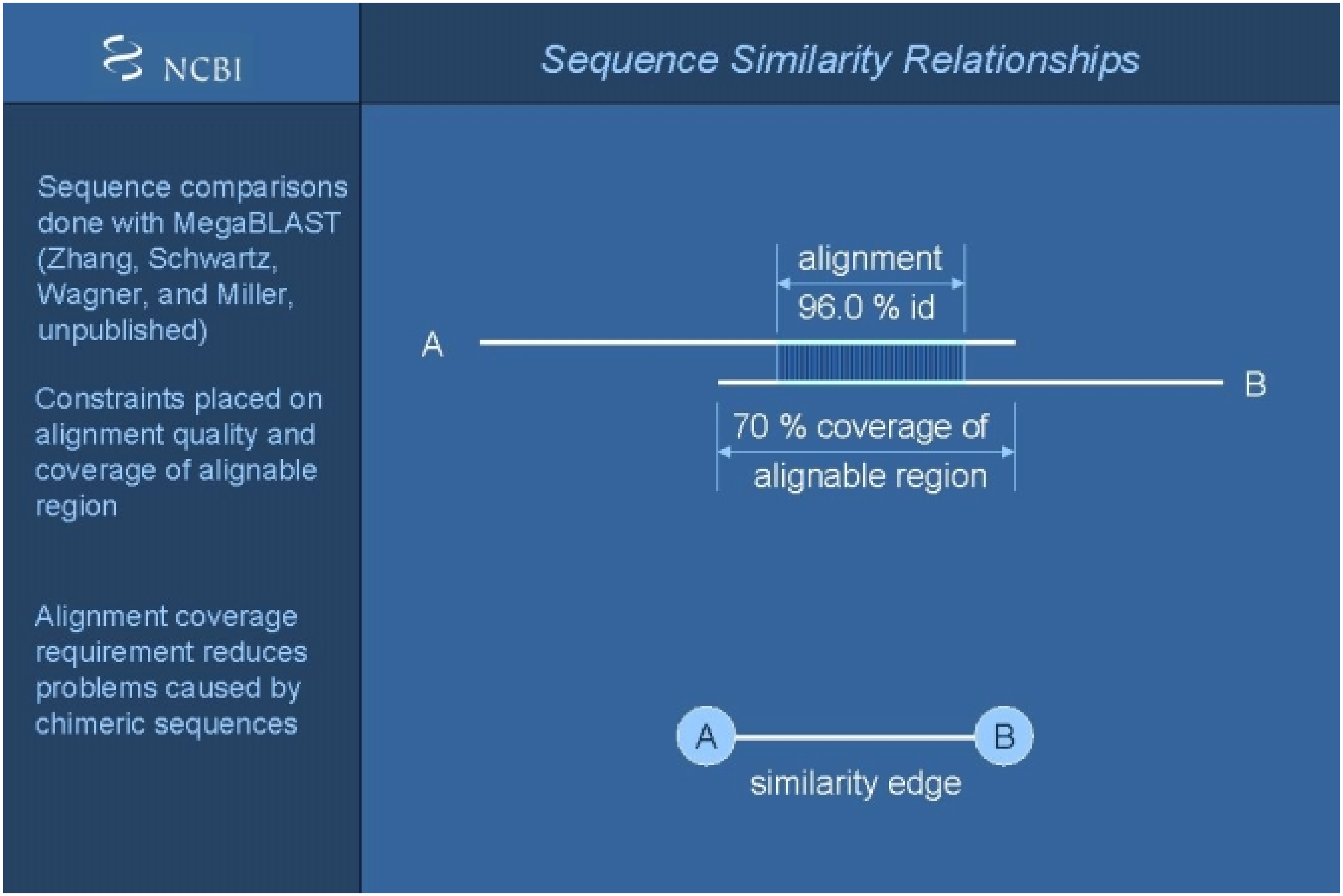
GenBank mRNAs	GenBank genomic	dbEST ESTs	
	"electronic splicing"		
14,389	5,371	1,256,901	Initial set
14,284	5,369	1,244,956	Contaminant removal
14,284	5,369	1,169,720	Informative length cutoff

1,189,373 Pooled sequences

1,189,324 Fusion sequence removal

Unigene approach

- Originally 3’ only + mRNA common words of length 13 separated by no more than 2 bases.
- ID>Annotation>Shared clone ID
- Genbank, genomic ad dbEST > DUST > 100bp min >MEGABLAST



NCBI

Sequence Similarity Relationships

Sequence comparisons done with MegaBLAST (Zhang, Schwartz, Wagner, and Miller, unpublished)

Constraints placed on alignment quality and coverage of alignable region

Alignment coverage requirement reduces problems caused by chimeric sequences

SANBI

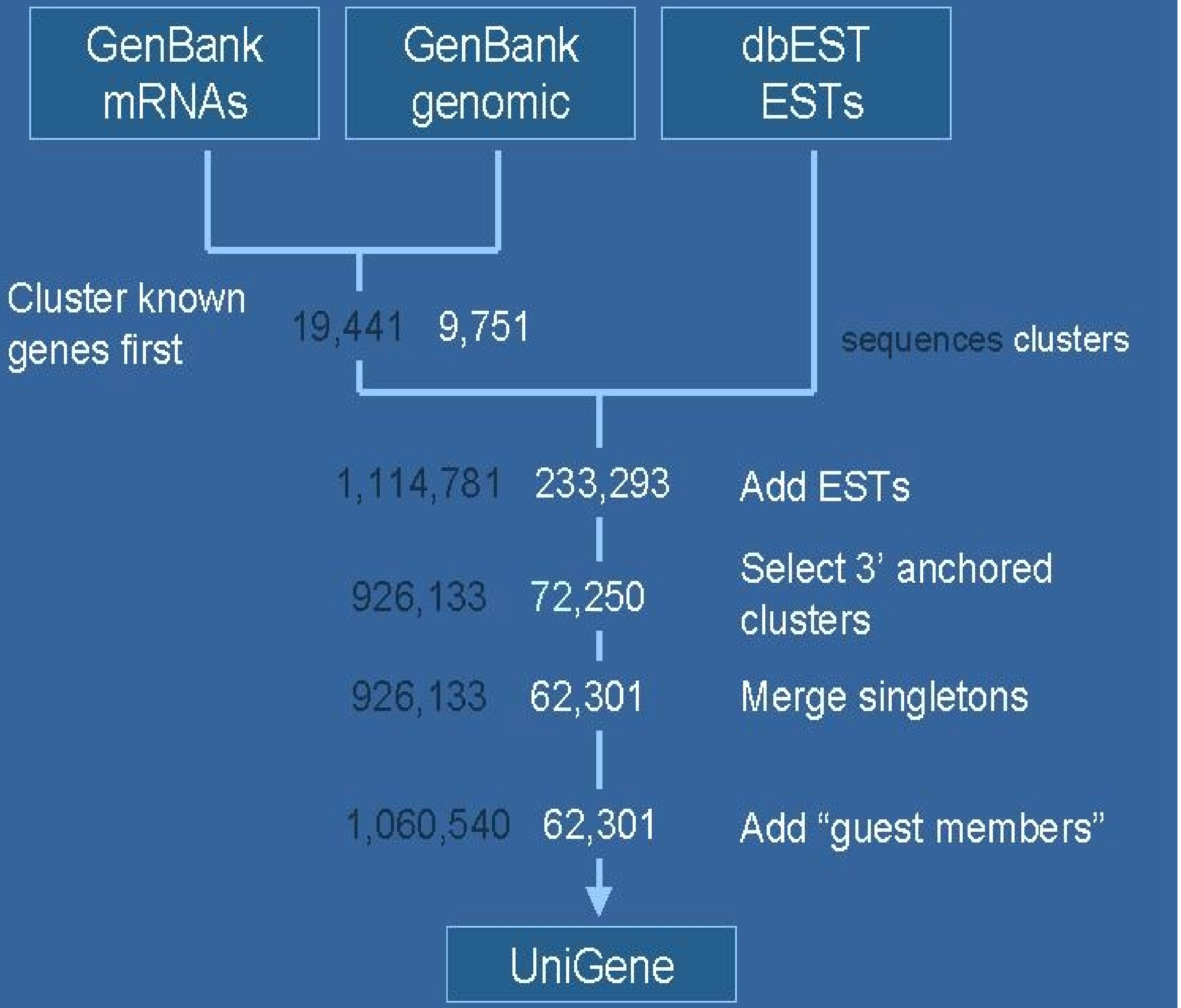
Wagner et al. CSH 1999

Multi-step Clustering

Multi-step clustering prevents bad ESTs from corrupting good mRNA/gene clusters

More reliable data used before less reliable data

3’ ends desired to avoid multiple clusters per gene



GenBank mRNAs

GenBank genomic

dbEST ESTs

Cluster known genes first

sequences clusters

19,441 9,751

1,114,781 233,293 Add ESTs

926,133 72,250 Select 3’ anchored clusters

926,133 62,301 Merge singletons

1,060,540 62,301 Add “guest members”

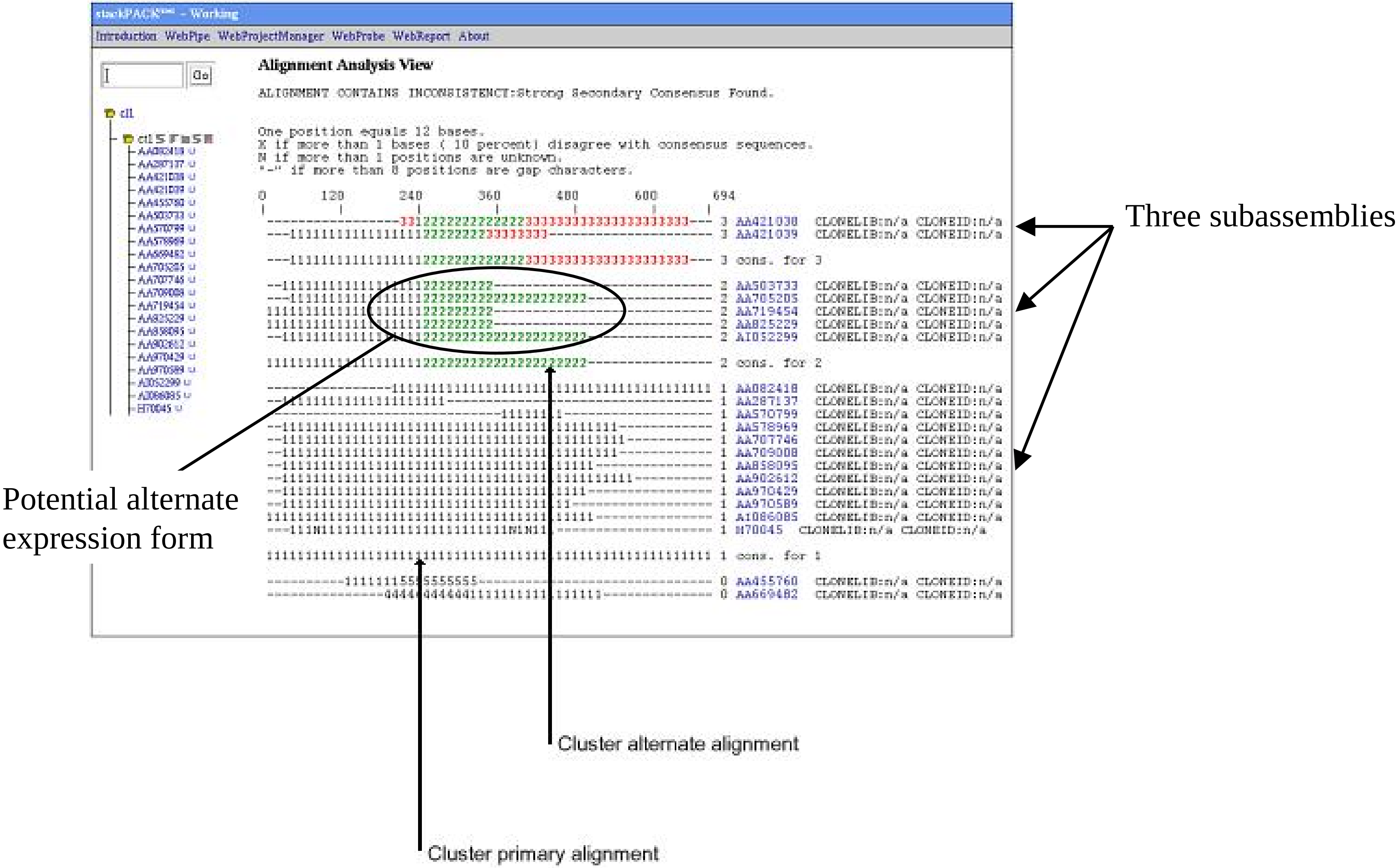
UniGene

Fragmentation Comparison

Methodology	**Input Sequences**	**Singleton Groups**	**%Singleton Groups**
TIGR Gene Index	626 163	135 140	21.83
STACK_PACK	415 833	58 070	13.96

STACK_PACK analysis of UniGene clusters resulted in a fragmentation rate just over half of the TIGR index.

Alignment Analysis


stackPACK™ - Working
Introduction WebPipe WebProjectManager WebProbe WebReport About
Go
cl1
cl1

- AA082418
- AA287137
- AA421038
- AA421039
- AA455760
- AA503733
- AA570799
- AA578969
- AA669482
- AA705205
- AA707746
- AA709008
- AA719454
- AA825229
- AA858095
- AA902612
- AA970429
- AA970589
- AI052299
- AI086085
- H70045

Alignment Analysis View

```
ALIGNMENT CONTAINS INCONSISTENCY:Strong Secondary Consensus Found.


One position equals 12 bases.
X if more than 1 bases ( 10 percent) disagree with consensus sequences.
N if more than 1 positions are unknown.
"-" if more than 8 positions are gap characters.

0         120       240       360       400       600       694
|         |         |         |         |         |         |
 -----------------331222222222222233333333333333333333333--- 3 AA421038  CLONELIB:n/a CLONEID:n/a
 ---11111111111111111222222223333333333--------------------- 3 AA421039  CLONELIB:n/a CLONEID:n/a

 ---111111111111111112222222222222233333333333333333333333--- 3 cons. for 3

 --111111111111111111222222222---------------------------- 2 AA503733  CLONELIB:n/a CLONEID:n/a
 ---11111111111111111222222222222222222222---------------- 2 AA705205  CLONELIB:n/a CLONEID:n/a
 11111111111111111111222222222---------------------------- 2 AA719454  CLONELIB:n/a CLONEID:n/a
 11111111111111111111222222222---------------------------- 2 AA825229  CLONELIB:n/a CLONEID:n/a
 --111111111111111111222222222222222222222---------------- 2 AI052299  CLONELIB:n/a CLONEID:n/a

 11111111111111111111222222222222222222222---------------- 2 cons. for 2

 ----------------11111111111111111111111111111111111111111 1 AA082418  CLONELIB:n/a CLONEID:n/a
 --111111111111111111111---------------------------------- 1 AA287137  CLONELIB:n/a CLONEID:n/a
 ------------------------------11111111------------------- 1 AA570799  CLONELIB:n/a CLONEID:n/a
 --1111111111111111111111111111111111111111111------------ 1 AA578969  CLONELIB:n/a CLONEID:n/a
 --11111111111111111111111111111111111111111111----------- 1 AA707746  CLONELIB:n/a CLONEID:n/a
 --1111111111111111111111111111111111111111111------------ 1 AA709008  CLONELIB:n/a CLONEID:n/a
 --1111111111111111111111111111111111111111--------------- 1 AA858095  CLONELIB:n/a CLONEID:n/a
 --111111111111111111111111111111111111111111111---------- 1 AA902612  CLONELIB:n/a CLONEID:n/a
 --111111111111111111111111111111111111111---------------- 1 AA970429  CLONELIB:n/a CLONEID:n/a
 --1111111111111111111111111111111111111------------------ 1 AA970589  CLONELIB:n/a CLONEID:n/a
 11111111111111111111111111111111111111111---------------- 1 AI086085  CLONELIB:n/a CLONEID:n/a
 ---111N11111111111111111111111111N1N11------------------- 1 H70045  CLONELIB:n/a CLONEID:n/a

 111111111111111111111111111111111111111111111111111111111 1 cons. for 1

 ----------1111111555555555555---------------------------- 0 AA455760  CLONELIB:n/a CLONEID:n/a
 ---------------44444444441111111111111111111-------------- 0 AA669482  CLONELIB:n/a CLONEID:n/a
```

Three subassemblies

Potential alternate expression form

Cluster alternate alignment

Cluster primary alignment

Orthologs and Paralogs

- Orthologs
  - Genes that share the same ancestral gene that perform the same biological function in different species but have diverged in sequence makeup due to selective evolution
- Paralogs
  - Genes within the same genome that share an ancestral gene that perform diverse biological functions.

Needs

- Functional assignments
- Expression states of alternate forms and their sites of expression
- Exon level resolution of expression
- Representative forms for application to arrays
- Physical gene locations
- Relationship to disease

Exploration

- Availability of genomic sequence and partial transcription products means characterisation of alternate transcription can begin in earnest.
- Contribution to variation of expressed products and effects on biology are likely to be significant

How to trap useful genome sequence to manufacture a genome virtually?

Gene level approach

Trap Expressed Sequence Tags

~1.8 M tags, ~35-100K genes

Combine to form virtual genes

Annotate and analyse these genes

Correlate with phenotype(s) = disease

Understand the expression basis of disease

Reconstruction of transcripts

- Derive understanding of expressed gene products
  - Use of expressed sequence data requires complex processing
  - Processed datasets are badly needed

- Capture a first glimpse of a genome's activites
  - Genomic level sequence is the final state, but its products can provide powerful information very early.

- Characterize underlying gene structure
  - Exon boundaries are difficult to define accurately and consistently

- Assess effect of an intervention on gene expression products
  - A rough EST profile is a quick identifier of key expression products

- Associate isoforms with expression states
  - Expression forms vary, how and when?
  - What does a full length cDNA really mean?

Why is transcript data a problem?

Transcript Data

Full length cDNA

- GenBank has many entries that confuse ‘full length’ with ‘complete Coding Sequence’

Partial cDNA

- Redundant partial cDNA sequences

Exon Composite

- All confirmed exons combined to form a ‘complete transcript’

Expressed Sequence Tag

- Single pass sequence

Genome Survey Sequence

- Single pass sequence
  - Small genomes contain more coding sequences in GSS than larger genomes

Genome Sequence:

Characterizing underlying gene structure

Fanfare fragment

- First Pass Annotated

Exon boundaries

- Predicted
- Cross species conservation
- Transcript confirmation

Composite exon transcript

- How do you define a transcript?

STACKing approach

Distill quality from quantity

Accurate consensus sequence representation

Identify expression variation, both spatial and developmental

Facilitate better understanding of gene expression

Exon-level gene expression profile

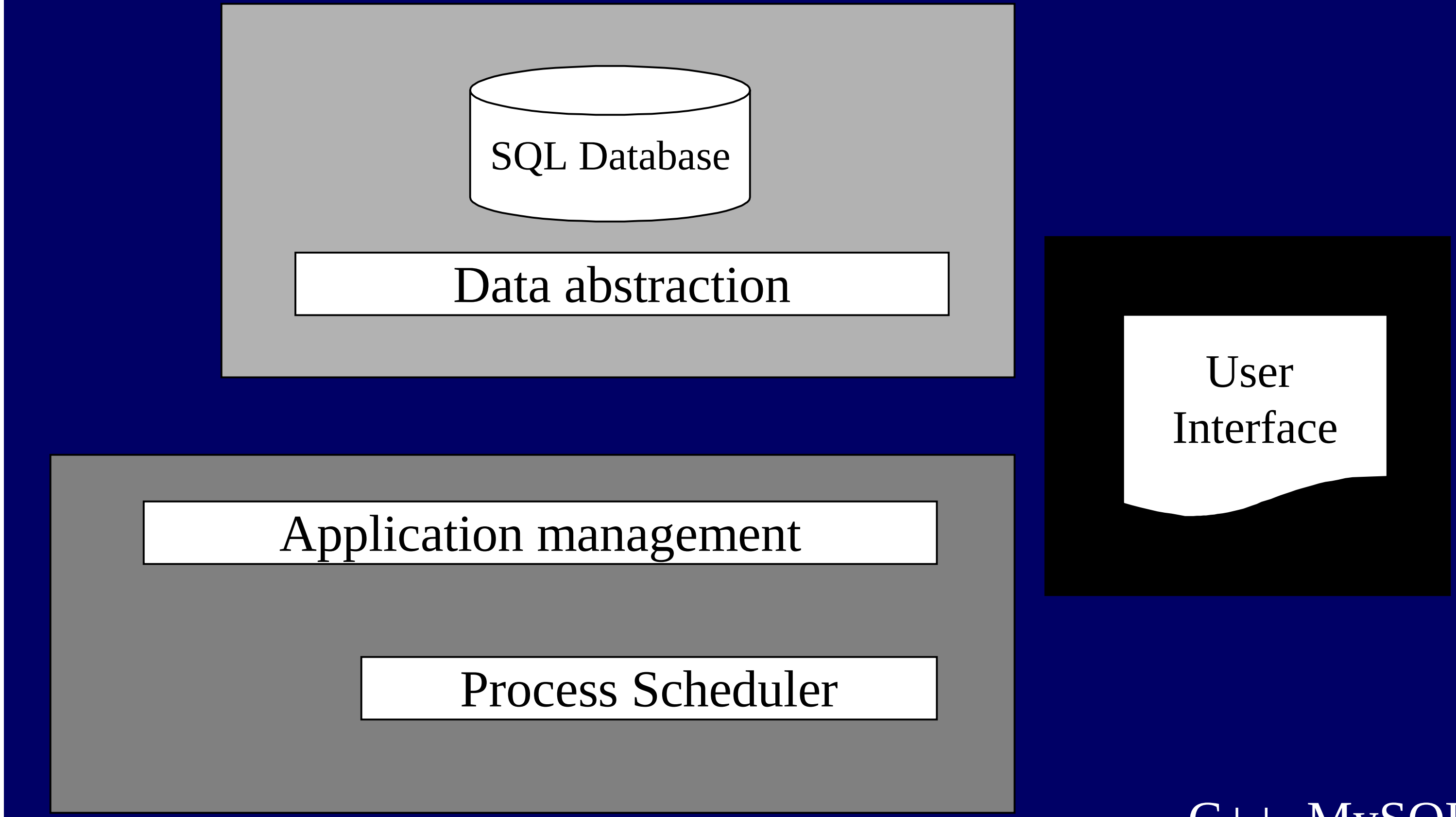
Integration of expression with genome sequence

Confirm and discover expressed exons

Provide gene candidacy delivery

Integrate with phenotype

STACKPACK



SQL Database

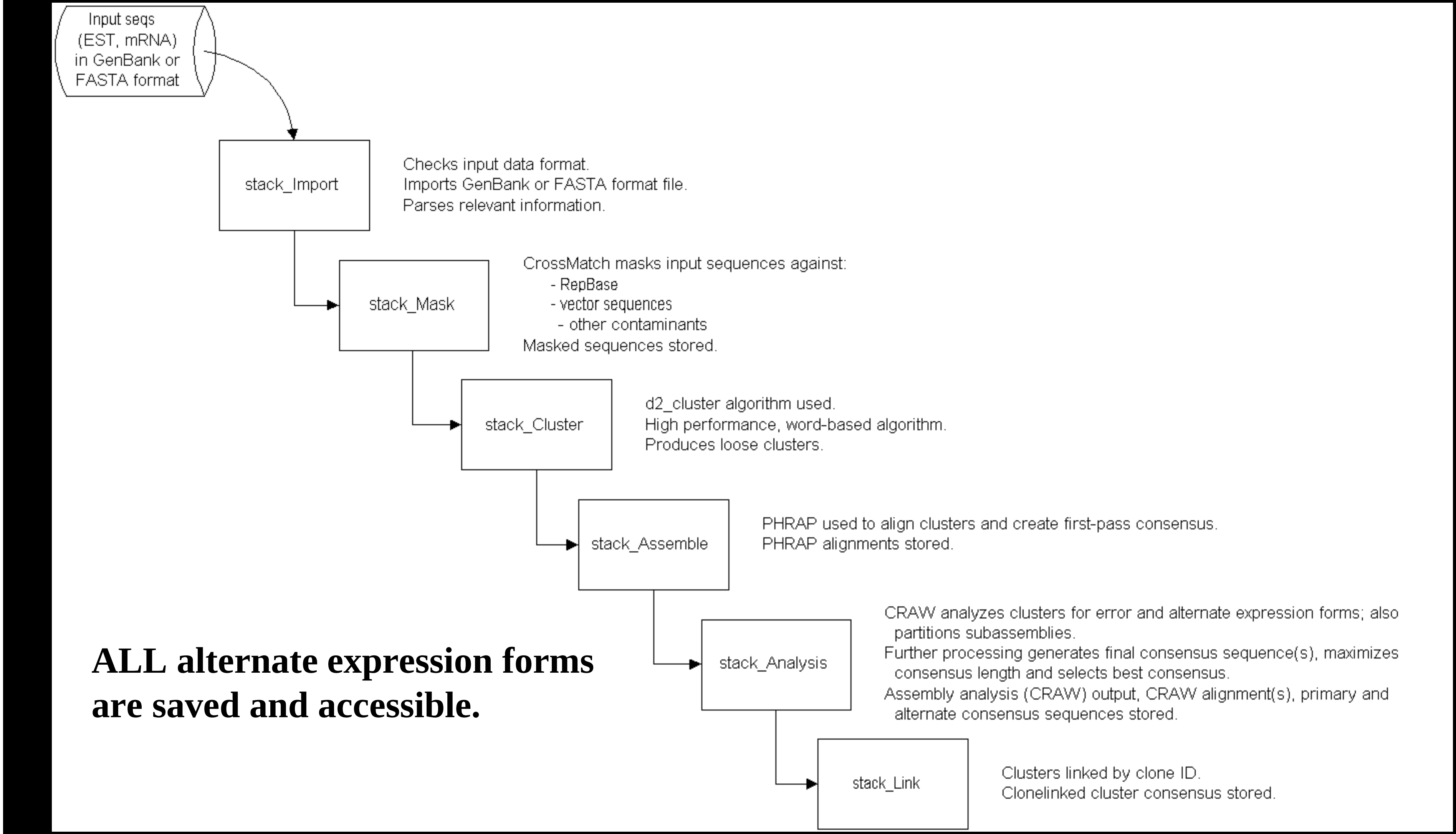
Data abstraction

User Interface

Application management

Process Scheduler

- C++, MySQL, HTML, Java



Checks input data format.
Imports GenBank or FASTA format file.
Parses relevant information.

CrossMatch masks input sequences against:
- RepBase
- vector sequences
- other contaminants

Masked sequences stored.

d2_cluster algorithm used.
High performance, word-based algorithm.
Produces loose clusters.

PHRAP used to align clusters and create first-pass consensus.
PHRAP alignments stored.

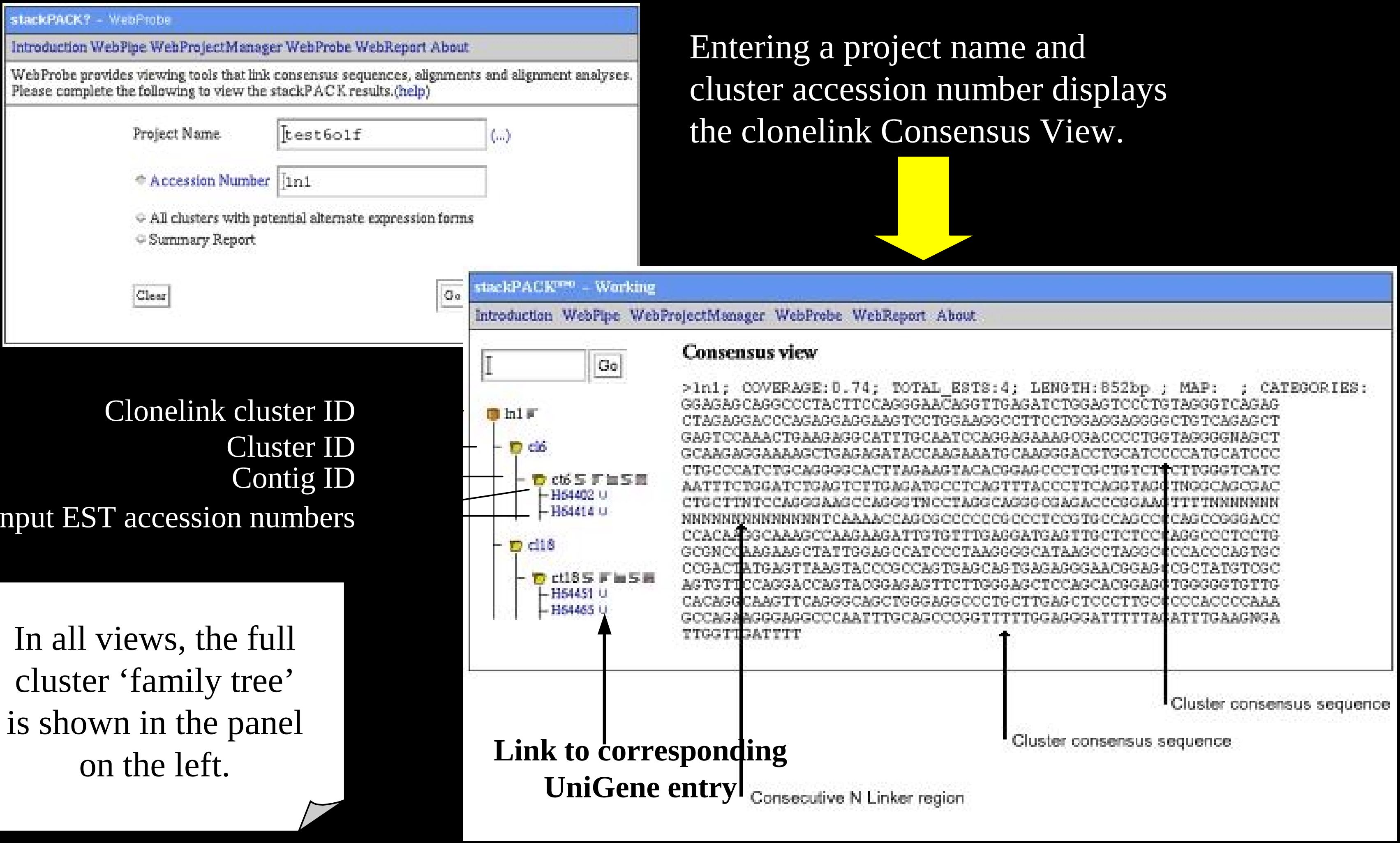
CRAW analyzes clusters for error and alternate expression forms; also partitions subassemblies.
Further processing generates final consensus sequence(s), maximizes consensus length and selects best consensus.
Assembly analysis (CRAW) output, CRAW alignment(s), primary and alternate consensus sequences stored.

Clusters linked by clone ID.
Clonelinked cluster consensus stored.

ALL alternate expression forms are saved and accessible.

WebProbe - View by clonelink accession

Entering a project name and cluster accession number displays the clonelink Consensus View.



stackPACK? - WebProbe

Introduction WebPipe WebProjectManager WebProbe WebReport About

WebProbe provides viewing tools that link consensus sequences, alignments and alignment analyses. Please complete the following to view the stackPACK results.(help)

Project Name: test6o1f (...)

● Accession Number: ln1

○ All clusters with potential alternate expression forms

○ Summary Report

Clear | Go

stackPACK™ - Working

Introduction WebPipe WebProjectManager WebProbe WebReport About

Consensus view

```
>ln1; COVERAGE:0.74; TOTAL_ESTS:4; LENGTH:852bp ; MAP:  ; CATEGORIES:
GGAGAGCAGGCCCTACTTCCAGGGAACAGGTTGAGATCTGGAGTCCCTGTAGGGTCAGAG
CTAGAGGACCCAGAGGAGGAAGTCCTGGAAGGCCTTCCTGGAGGAGGGGCTGTCAGAGCT
GAGTCCAAACTGAAGAGGCATTTGCAATCCAGGAGAAAGCGACCCCTGGTAGGGGNAGCT
GCAAGAGGAAAAGCTGAGAGATACCAAGAAATGCAAGGGACCTGCATCCCCATGCATCCC
CTGCCCATCTGCAGGGGCACTTAGAAGTACACGGAGCCCTCGCTGTCTTCTTGGGTCATC
AATTTCTGGATCTGAGTCTTGAGATGCCTCAGTTTACCCTTCAGGTAGGTNGGCAGCGAC
CTGCTTNTCCAGGGAAGCCAGGGTNCCTAGGCAGGGCGAGACCCGGAAGTTTTNNNNNNN
NNNNNNNNNNNNNNNTCAAAACCAGCGCCCCCCGCCCTCCGTGCCAGCCCCAGCCGGGACC
CCACAAGGCAAAGCCAAGAAGATTGTGTTTGAGGATGAGTTGCTCTCCCAGGCCCTCCTG
GCGNCCAAGAAGCTATTGGAGCCATCCCTAAGGGGCATAAGCCTAGGCCCCACCCAGTGC
CCGACTATGAGTTAAGTACCCGCCAGTGAGCAGTGAGAGGGAACGGAGCCGCTATGTCGC
AGTGTTCCAGGACCAGTACGGAGAGTTCTTGGGAGCTCCAGCACGGAGGTGGGGGTGTTG
CACAGGCAAGTTCAGGGCAGCTGGGAGGCCCTGCTTGAGCTCCCTTGCCCCCACCCCAAA
GCCAGAAGGGAGGCCCAATTTGCAGCCCGGTTTTTGGAGGGATTTTTAGATTTGAAGNGA
TTGGTTGATTTT
```

Go
ln1
cl6
ct6
H64402 U
H64414 U
cl18
ct18
H64451 U
H64465 U
Cluster consensus sequence
Cluster consensus sequence
Consecutive N Linker region

Clonelink cluster ID

Cluster ID

Contig ID

Input EST accession numbers

Link to corresponding UniGene entry

In all views, the full cluster 'family tree' is shown in the panel on the left.

Alignment and Analysis

- PHRAP Alignment
  - first alignment created
  - all ESTs in one alignment
- Alignment Analysis
  - CRAW used to look for subassemblies
  - Identifies potential alternate expression forms
- CRAW Alignment
  - Final alignment for each subassembly
- Consensus Analysis
  - Statistics used to select best consensus
  - Notes degree of matching between EST & consensus

The Value of Cluster Data

Microarray Studies

Clusters represent unique forms associated with a specific state

Gene Discovery

Unique transcripts revealed in association with expression libraries – especially in little studied organisms

Functional Annotation

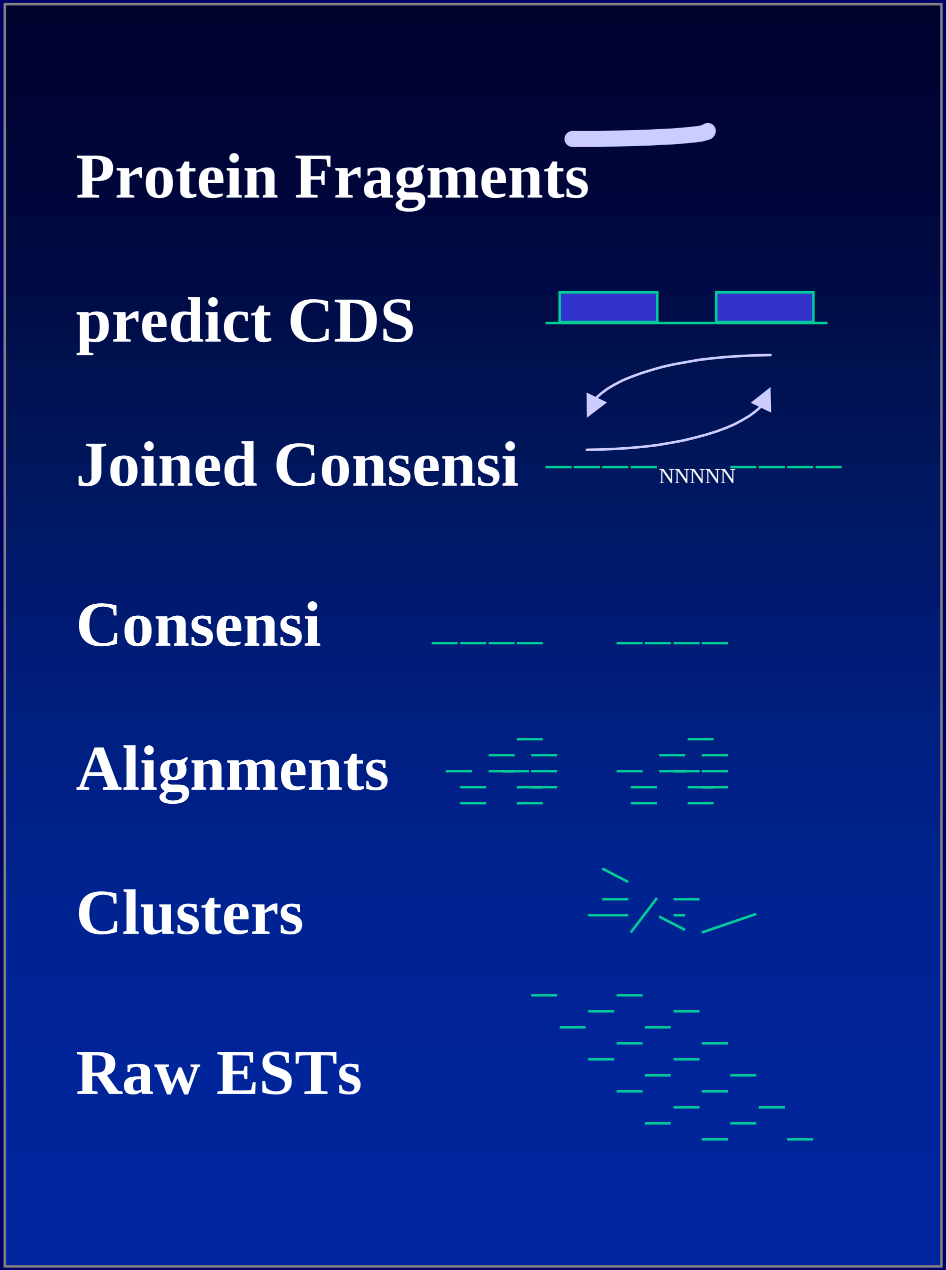
Virtual genes can be searched against the database to provide functional annotation of the products of a genome

Expressed Gene Structure

Exons boundaries are revealed by transcript confirmation

How to trap useful genome sequence to manufacture a genome virtually?

- Gene level approach
- Trap Expressed Sequence Tags
- Combine to reconstruct virtual genes
- Maufacture a substrate for microarray studies
- Annotate and analyse these genes
- Compare between species
  - Species-specific characteristics
  - Reveal genes under selection



Protein Fragments

predict CDS

Joined Consensi

NNNNN

Consensi

Alignments

Clusters

Raw ESTs

Virtual Protein
Sequence and transcript
reconstruction

Detection of virulence genes in malarial pathogens

Rahlston Muller

Reconstruction of transcripts from gene expression projects in the USA

Collaboration with Jane Carlton at NCBI

Delivery of over several previously unknown genes in *Plasmodium* spp.

Discovery of 76 genes that may be involved in virulence and pathogenicity

Vaccine and drug candidates

Sequence re-construction and assembly

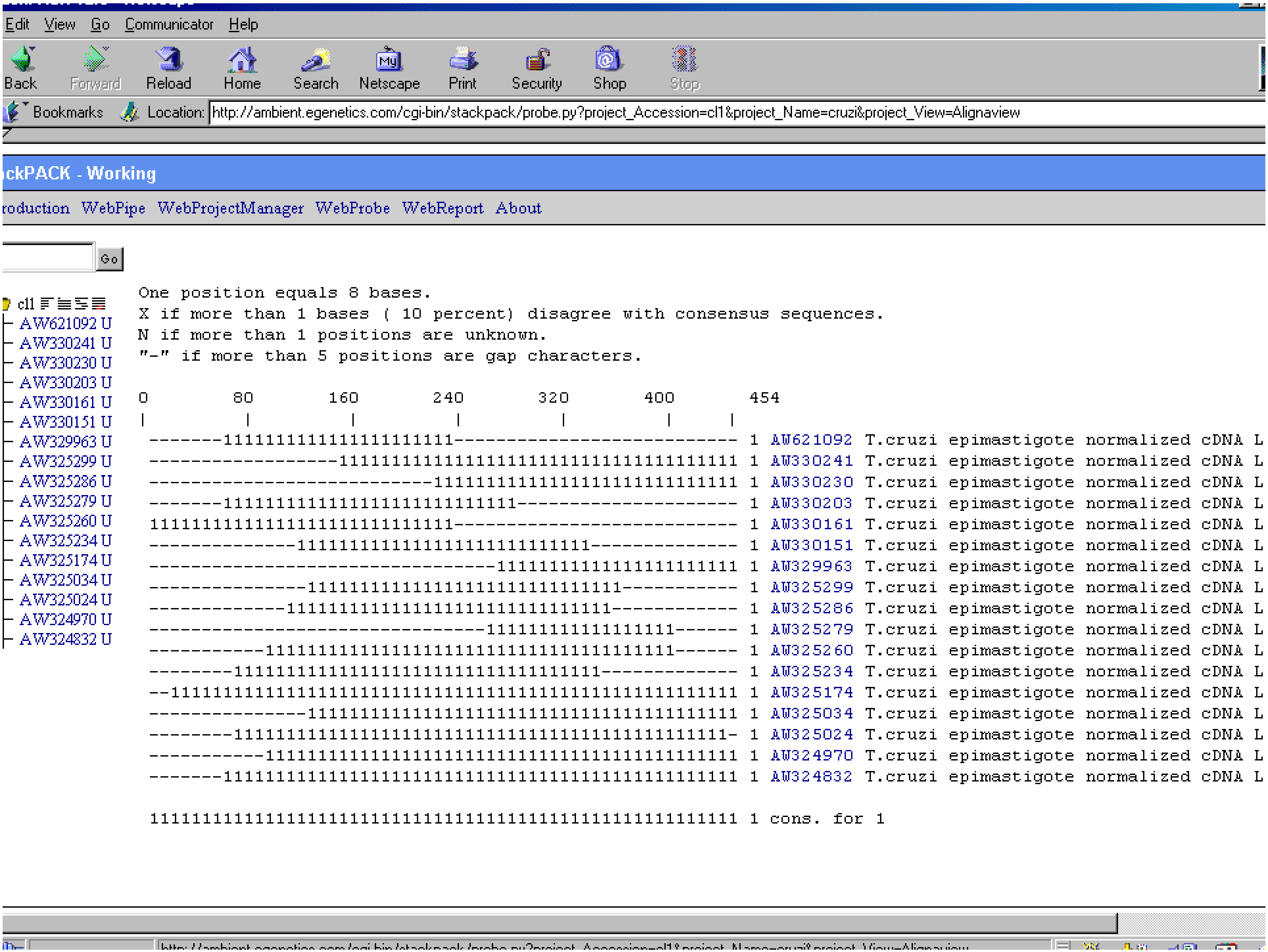
- ESTs re-constructed using stackPack
    - 6,697 submitted
    - 860 Multiple Sequence clusters, and
    - 2,786 singletons

- GSSs assembly using PHRAP
    - Clones may contain a higher proportion of CDS
    - 18,082 submitted
    - 2,784 contigs
    - 10,979 singletons

- All together now : 17,409 consensus sequences
- Subsequent analysis

Redundancy determination

- PF
  - ESTs 15%
  - GSSs 14%
- PB
  - ESTs 50%, not normalized
  - GSSs 24%
- PV
  - Sal I 26%
  - Belem 25%



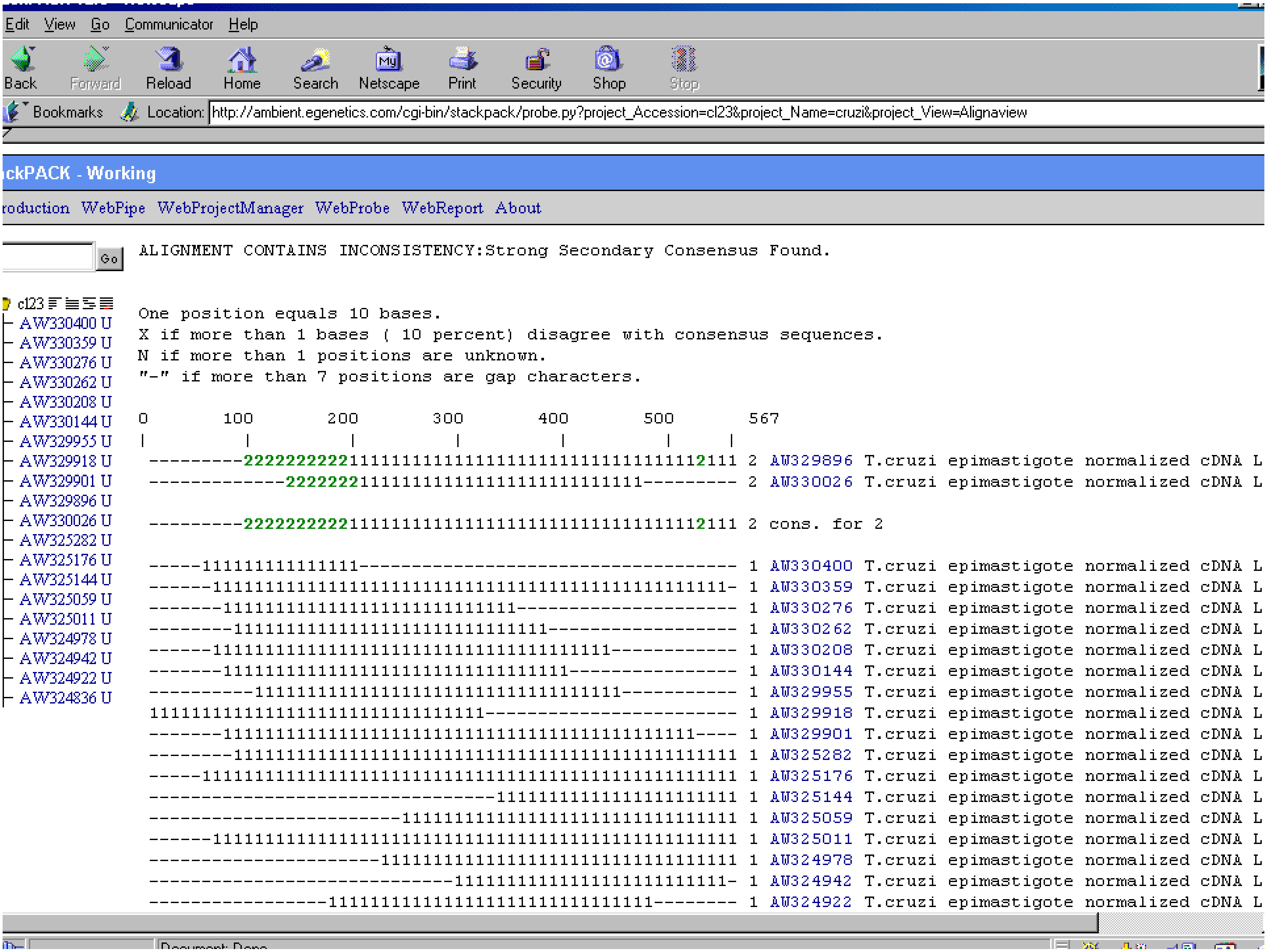
http://ambient.egenetics.com/cgi-bin/stackpack/probe.py?project_Accession=cl1&project_Name=cruzi&project_View=Alignaview

cl1
- AW621092 U
- AW330241 U
- AW330230 U
- AW330203 U
- AW330161 U
- AW330151 U
- AW329963 U
- AW325299 U
- AW325286 U
- AW325279 U
- AW325260 U
- AW325234 U
- AW325174 U
- AW325034 U
- AW325024 U
- AW324970 U
- AW324832 U

```
One position equals 8 bases.
X if more than 1 bases ( 10 percent) disagree with consensus sequences.
N if more than 1 positions are unknown.
"-" if more than 5 positions are gap characters.

0         80        160       240       320       400       454
|          |          |          |          |          |      |
 -------1111111111111111111111--------------------------- 1 AW621092 T.cruzi epimastigote normalized cDNA L
 ------------------11111111111111111111111111111111111111 1 AW330241 T.cruzi epimastigote normalized cDNA L
 ---------------------------11111111111111111111111111111 1 AW330230 T.cruzi epimastigote normalized cDNA L
 -------1111111111111111111111111111--------------------- 1 AW330203 T.cruzi epimastigote normalized cDNA L
 11111111111111111111111111111--------------------------- 1 AW330161 T.cruzi epimastigote normalized cDNA L
 --------------1111111111111111111111111111-------------- 1 AW330151 T.cruzi epimastigote normalized cDNA L
 ---------------------------------11111111111111111111111 1 AW329963 T.cruzi epimastigote normalized cDNA L
 ---------------111111111111111111111111111111----------- 1 AW325299 T.cruzi epimastigote normalized cDNA L
 -------------1111111111111111111111111111111------------ 1 AW325286 T.cruzi epimastigote normalized cDNA L
 --------------------------------111111111111111111------ 1 AW325279 T.cruzi epimastigote normalized cDNA L
 -----------111111111111111111111111111111111111111------ 1 AW325260 T.cruzi epimastigote normalized cDNA L
 --------11111111111111111111111111111111111------------- 1 AW325234 T.cruzi epimastigote normalized cDNA L
 --111111111111111111111111111111111111111111111111111111 1 AW325174 T.cruzi epimastigote normalized cDNA L
 ---------------11111111111111111111111111111111111111111 1 AW325034 T.cruzi epimastigote normalized cDNA L
 --------11111111111111111111111111111111111111111111111- 1 AW325024 T.cruzi epimastigote normalized cDNA L
 -----------111111111111111111111111111111111111111111111 1 AW324970 T.cruzi epimastigote normalized cDNA L
 -------1111111111111111111111111111111111111111111111111 1 AW324832 T.cruzi epimastigote normalized cDNA L

 11111111111111111111111111111111111111111111111111111111 1 cons. for 1
```



Location: http://ambient.egenetics.com/cgi-bin/stackpack/probe.py?project_Accession=cl23&project_Name=cruzi&project_View=Alignaview

cl23
- AW330400 U
- AW330359 U
- AW330276 U
- AW330262 U
- AW330208 U
- AW330144 U
- AW329955 U
- AW329918 U
- AW329901 U
- AW329896 U
- AW330026 U
- AW325282 U
- AW325176 U
- AW325144 U
- AW325059 U
- AW325011 U
- AW324978 U
- AW324942 U
- AW324922 U
- AW324836 U

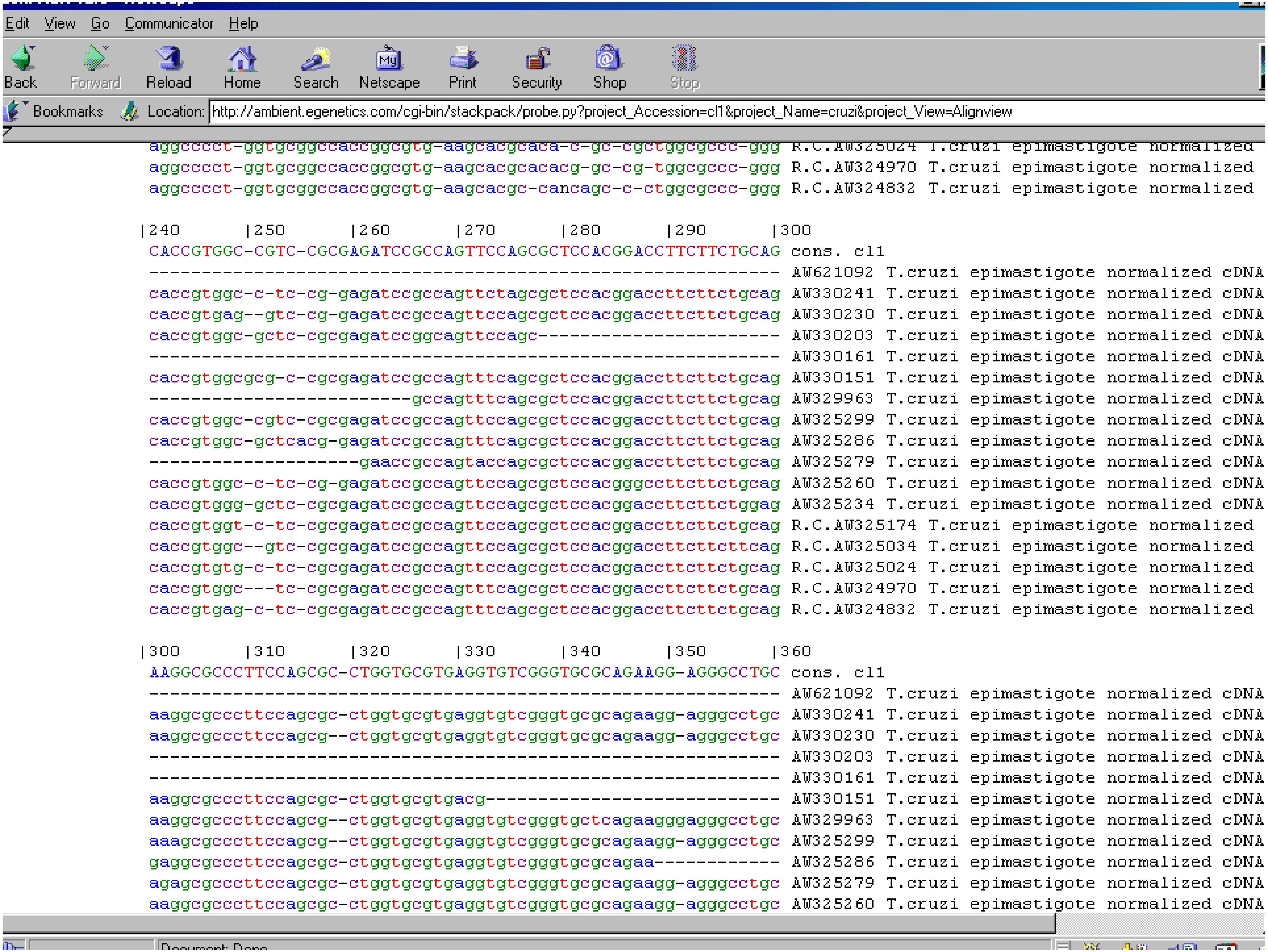
```
ALIGNMENT CONTAINS INCONSISTENCY:Strong Secondary Consensus Found.


One position equals 10 bases.
X if more than 1 bases ( 10 percent) disagree with consensus sequences.
N if more than 1 positions are unknown.
"-" if more than 7 positions are gap characters.

0       100       200       300       400       500       567
|         |         |         |         |         |     |
 ---------2222222222111111111111111111111111111111111121112 AW329896 T.cruzi epimastigote normalized cDNA L
 -------------22222221111111111111111111111111111--------- 2 AW330026 T.cruzi epimastigote normalized cDNA L

 ---------22222222221111111111111111111111111111111111112111 2 cons. for 2

 -----111111111111111------------------------------------- 1 AW330400 T.cruzi epimastigote normalized cDNA L
 ------11111111111111111111111111111111111111111111111111- 1 AW330359 T.cruzi epimastigote normalized cDNA L
 -------1111111111111111111111111111---------------------- 1 AW330276 T.cruzi epimastigote normalized cDNA L
 --------111111111111111111111111111111------------------- 1 AW330262 T.cruzi epimastigote normalized cDNA L
 ------11111111111111111111111111111111111111------------- 1 AW330208 T.cruzi epimastigote normalized cDNA L
 -------111111111111111111111111111111111----------------- 1 AW330144 T.cruzi epimastigote normalized cDNA L
 ----------11111111111111111111111111111111111------------ 1 AW329955 T.cruzi epimastigote normalized cDNA L
 11111111111111111111111111111111------------------------- 1 AW329918 T.cruzi epimastigote normalized cDNA L
 -------1111111111111111111111111111111111111111111111---- 1 AW329901 T.cruzi epimastigote normalized cDNA L
 --------1111111111111111111111111111111111111111111111111 1 AW325282 T.cruzi epimastigote normalized cDNA L
 -----1111111111111111111111111111111111111111111111111111 1 AW325176 T.cruzi epimastigote normalized cDNA L
 ---------------------------------111111111111111111111111 1 AW325144 T.cruzi epimastigote normalized cDNA L
 ------------------------111111111111111111111111111111111 1 AW325059 T.cruzi epimastigote normalized cDNA L
 ------111111111111111111111111111111111111111111111111111 1 AW325011 T.cruzi epimastigote normalized cDNA L
 ----------------------11111111111111111111111111111111111 1 AW324978 T.cruzi epimastigote normalized cDNA L
 -----------------------------111111111111111111111111111- 1 AW324942 T.cruzi epimastigote normalized cDNA L
 -----------------11111111111111111111111111111111-------- 1 AW324922 T.cruzi epimastigote normalized cDNA L
```



http://ambient.egenetics.com/cgi-bin/stackpack/probe.py?project_Accession=cl1&project_Name=cruzi&project_View=Alignview

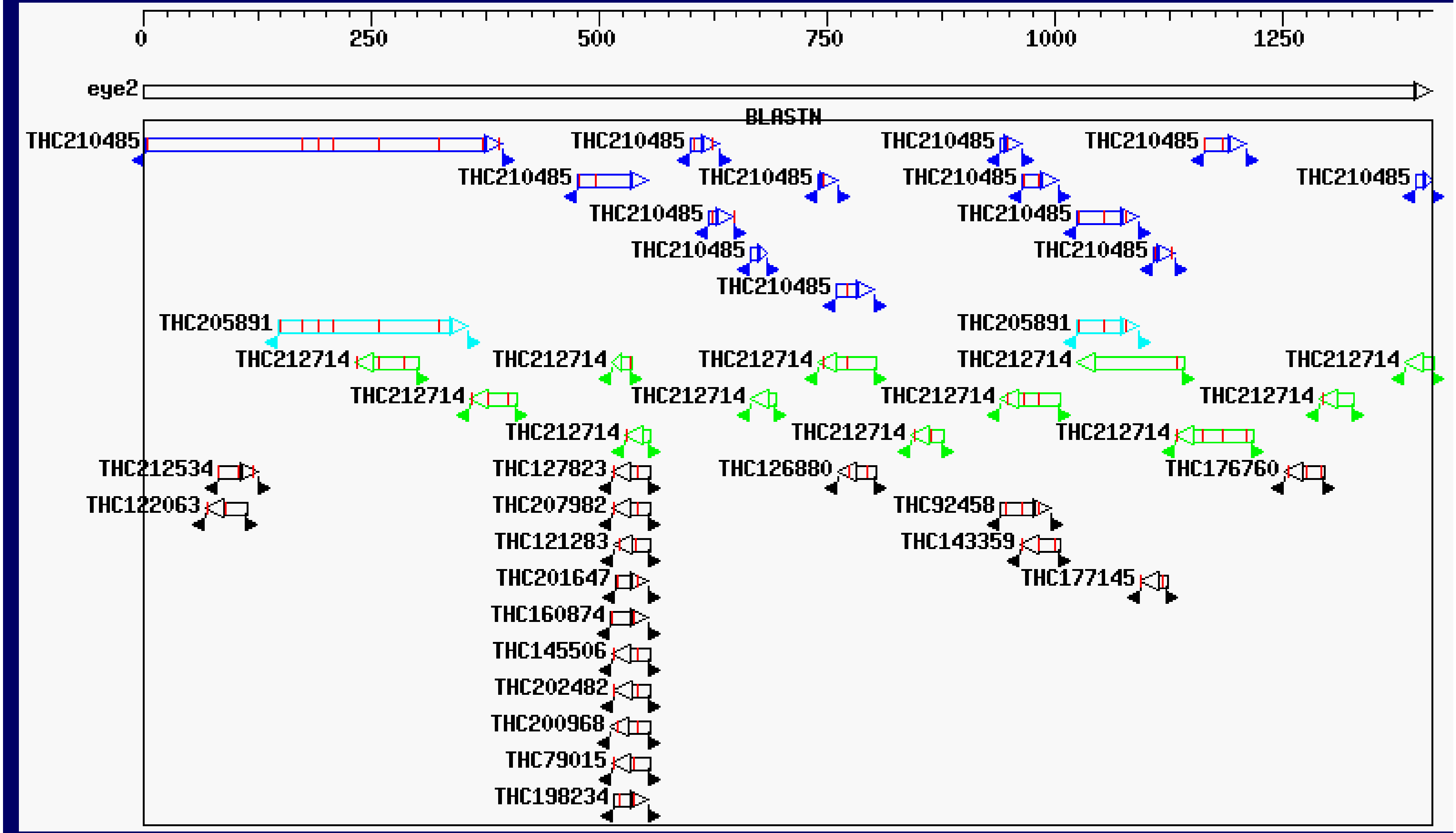
```
aggcccct-ggtgcggccaccggcgtg-aagcacgcaca-c-gc-cgctggcgccc-ggg R.C.AW325024 T.cruzi epimastigote normalized
aggcccct-ggtgcggccaccggcgtg-aagcacgcacacg-gc-cg-tggcgccc-ggg R.C.AW324970 T.cruzi epimastigote normalized
aggcccct-ggtgcggccaccggcgtg-aagcacgc-cancagc-c-ctggcgccc-ggg R.C.AW324832 T.cruzi epimastigote normalized

|240      |250      |260      |270      |280      |290      |300
CACCGTGGC-CGTC-CGCGAGATCCGCCAGTTCCAGCGCTCCACGGACCTTCTTCTGCAG cons. cl1
------------------------------------------------------------ AW621092 T.cruzi epimastigote normalized cDNA
caccgtggc-c-tc-cg-gagatccgccagttctagcgctccacggaccttcttctgcag AW330241 T.cruzi epimastigote normalized cDNA
caccgtgag--gtc-cg-gagatccgccagttccagcgctccacggaccttcttctgcag AW330230 T.cruzi epimastigote normalized cDNA
caccgtggc-gctc-cgcgagatccggcagttccagc----------------------- AW330203 T.cruzi epimastigote normalized cDNA
------------------------------------------------------------ AW330161 T.cruzi epimastigote normalized cDNA
caccgtggcgcg-c-cgcgagatccgccagtttcagcgctccacggaccttcttctgcag AW330151 T.cruzi epimastigote normalized cDNA
-------------------------gccagtttcagcgctccacggaccttcttctgcag AW329963 T.cruzi epimastigote normalized cDNA
caccgtggc-cgtc-cgcgagatccgccagttccagcgctccacggaccttcttctgcag AW325299 T.cruzi epimastigote normalized cDNA
caccgtggc-gctcacg-gagatccgccagtttcagcgctccacggaccttcttctgcag AW325286 T.cruzi epimastigote normalized cDNA
--------------------gaaccgccagtaccagcgctccacggaccttcttctgcag AW325279 T.cruzi epimastigote normalized cDNA
caccgtggc-c-tc-cg-gagatccgccagttccagcgctccacgggccttcttctgcag AW325260 T.cruzi epimastigote normalized cDNA
caccgtggg-gctc-cgcgagatccgccagttccagcgctccacggaccttcttctggag AW325234 T.cruzi epimastigote normalized cDNA
caccgtggt-c-tc-cgcgagatccgccagttccagcgctccacggaccttcttctgcag R.C.AW325174 T.cruzi epimastigote normalized
caccgtggc--gtc-cgcgagatccgccagttccagcgctccacggaccttcttcttcag R.C.AW325034 T.cruzi epimastigote normalized
caccgtgtg-c-tc-cgcgagatccgccagttccagcgctccacggaccttcttctgcag R.C.AW325024 T.cruzi epimastigote normalized
caccgtggc---tc-cgcgagatccgccagtttcagcgctccacggaccttcttctgcag R.C.AW324970 T.cruzi epimastigote normalized
caccgtgag-c-tc-cgcgagatccgccagtttcagcgctccacggaccttcttctgcag R.C.AW324832 T.cruzi epimastigote normalized

|300      |310      |320      |330      |340      |350      |360
AAGGCGCCCTTCCAGCGC-CTGGTGCGTGAGGTGTCGGGTGCGCAGAAGG-AGGGCCTGC cons. cl1
------------------------------------------------------------ AW621092 T.cruzi epimastigote normalized cDNA
aaggcgcccttccagcgc-ctggtgcgtgaggtgtcgggtgcgcagaagg-agggcctgc AW330241 T.cruzi epimastigote normalized cDNA
aaggcgcccttccagcg--ctggtgcgtgaggtgtcgggtgcgcagaagg-agggcctgc AW330230 T.cruzi epimastigote normalized cDNA
------------------------------------------------------------ AW330203 T.cruzi epimastigote normalized cDNA
------------------------------------------------------------ AW330161 T.cruzi epimastigote normalized cDNA
aaggcgcccttccagcgc-ctggtgcgtgacg---------------------------- AW330151 T.cruzi epimastigote normalized cDNA
aaggcgcccttccagcg--ctggtgcgtgaggtgtcgggtgctcagaagggagggcctgc AW329963 T.cruzi epimastigote normalized cDNA
aaagcgcccttccagcg--ctggtgcgtgaggtgtcgggtgcgcagaagg-agggcctgc AW325299 T.cruzi epimastigote normalized cDNA
gaggcgcccttccagcgc-ctggtgcgtgaggtgtcgggtgcgcagaa------------ AW325286 T.cruzi epimastigote normalized cDNA
agagcgcccttccagcgc-ctggtgcgtgaggtgtcgggtgcgcagaagg-agggcctgc AW325279 T.cruzi epimastigote normalized cDNA
aaggcgcccttccagcgc-ctggtgcgtgaggtgtcgggtgcgcagaagg-agggcctgc AW325260 T.cruzi epimastigote normalized cDNA
```

Document: Done

Sample Graphical Output of a STACK Eye sequence eye2 BLASTN search Vs TIGR Tentative Human Consensus Sequences.



Outputs

Raw State Expression

- Representative unique forms associated with a specific state

Gene Discovery

- Unique transcripts revealed in association with expression libs

Isoform coupled expression

Gene Structure

- Exons boundaries are revealed by transcript confirmation

Protein prediction, using PHAT

- Putative open reading identified, using criteria other than db searches
- HMM gene finder for *Plasmodium*
  - *P.falciparum* 56% predicted
  - *P.berghei* 60% predicted
  - *P.vivax* 84% predicted
- 72% (12,530/17,408) predicted proteins