slide01.jpg (504x378x24b jpeg)

Transcript analysis and reconstruction

slide02.jpg (504x378x24b jpeg)

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?

SANBI

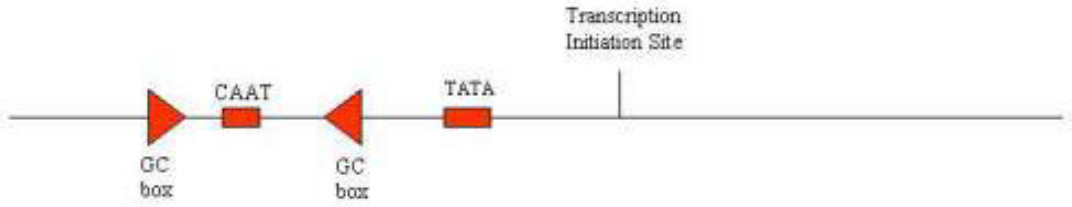
slide03.jpg (504x378x24b jpeg)

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

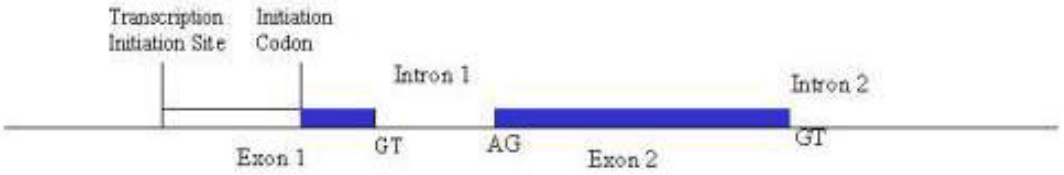
- 5' Flanking region with transcription regulation signals
- Transcription initiation start site (5')
- Initiation codon for protein coding sequence
- Exon-intron boundaries with splice site signals at the boundaries
- Termination codon for protein coding sequence
- 3' signals for regulation and polyadenylation

slide04.jpg (504x378x24b jpeg)



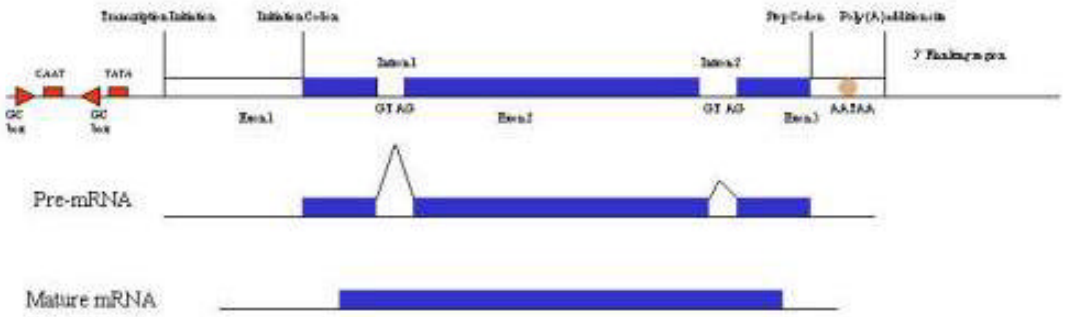
SANBI

slide05.jpg (504x378x24b jpeg)



SANBI

slide06.jpg (504x378x24b jpeg)



SANBI

slide07.jpg (504x378x24b jpeg)

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

SANBI

slide08.jpg (504x378x24b jpeg)

Examples of alternative splicing

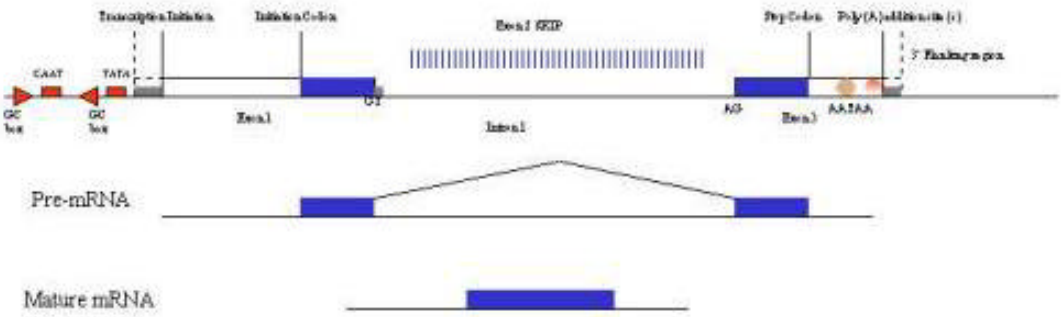
Alternative donor and acceptor splice sites

Alternative polyadenylation

Exon skipping

SANBI

slide09.jpg (504x378x24b jpeg)



SANBI

slide10.jpg (504x378x24b jpeg)

Capturing expressed transcripts

Databases - Sequences

dbEST

Several collapsed datasets

TIGR-THC	Allgenes
Unigene	BodyMap
STACK	Several more specialised

Genome Sequence as it appears:

SANBI

slide11.jpg (504x378x24b jpeg)

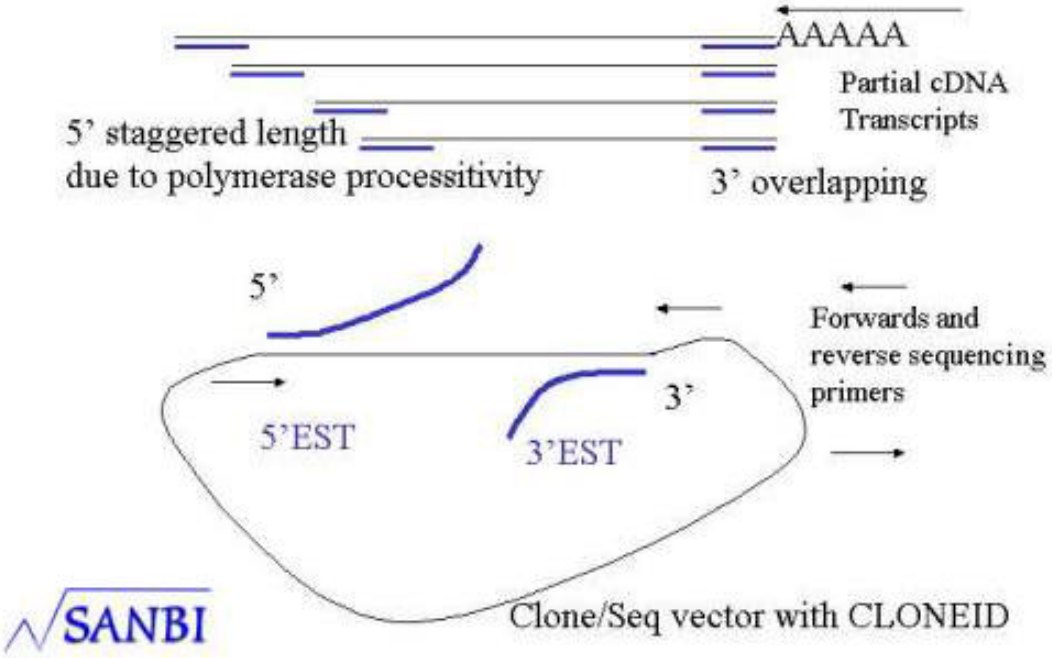
Expression Capture

- 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

slide12.jpg (504x378x24b jpeg)

What is an EST?



SANBI

slide13.jpg (504x378x24b jpeg)

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

SANBI

slide14.jpg (504x378x24b jpeg)

Use of Transcripts in Completed genomes

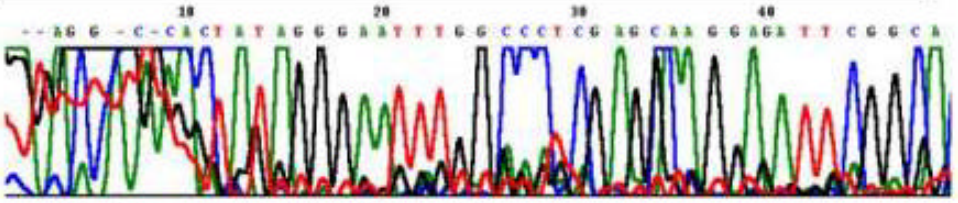
- Identification of genes
  - Exon boundaries
  - Alternate transcripts
- Genomic annotation
  - Expression sites of encoded genes
- Comparitive genomics

SANBI

slide15.jpg (504x378x24b jpeg)

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
AAGACCCCCCGTCTCTTTAAAAATATATATATTTTAAATATACTTAAATATATATTTCTAATATCTTTAAAT
ATATATATATATTTNAAAGACCAATTTATGGGAGANTTGCACACAGATGTGAAATGAATGTAATCTAATAG
ANGCCTAATCAGCCCACCATGTTCTCCACTGAAAAATCCTCTTTCTTTGGGGTTTTTCTTTCTTTCTTTTT
TGATTTTGCACTGGACGGTGACGTCAGCCATGTACAGGATCCACAGGGGTGGTGTCAAATGCTATTGAAAT
TNTGTTGAATTGTATACTTTTTCACTTTTTGATAATTAACCATGTAAAAAATG
```

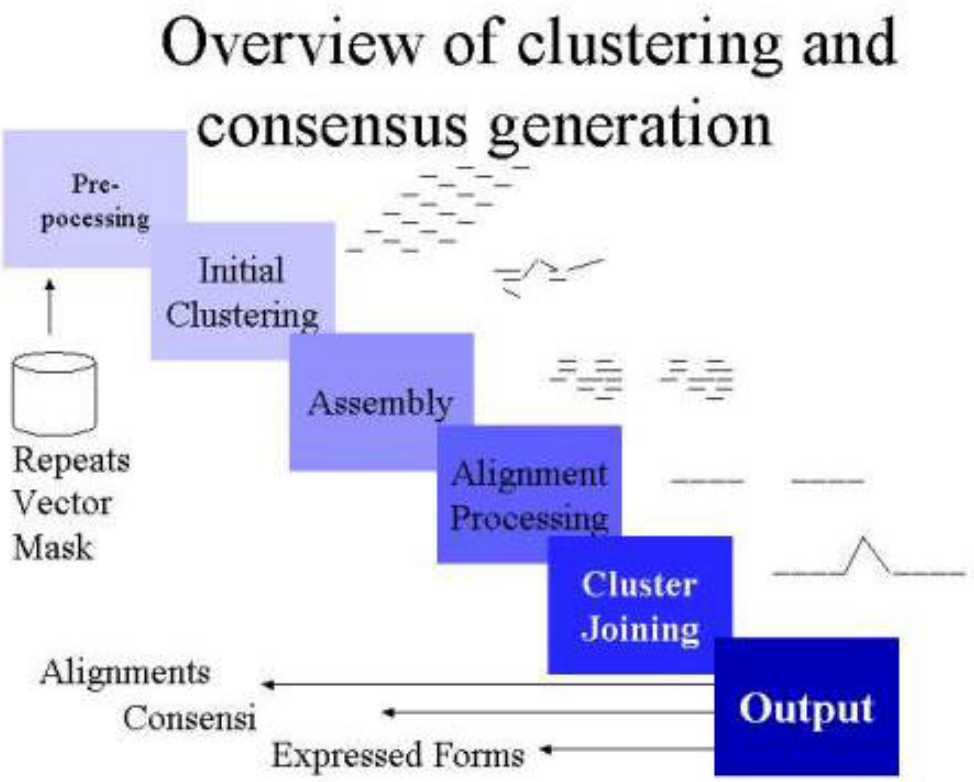


EST is Poor Quality data with contaminants

Vector Repeat MASK

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

slide16.jpg (504x378x24b jpeg)



slide17.jpg (504x378x24b jpeg)

Transcript reconstruction

slide18.jpg (504x378x24b jpeg)

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

slide19.jpg (504x378x24b jpeg)

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

SANBI

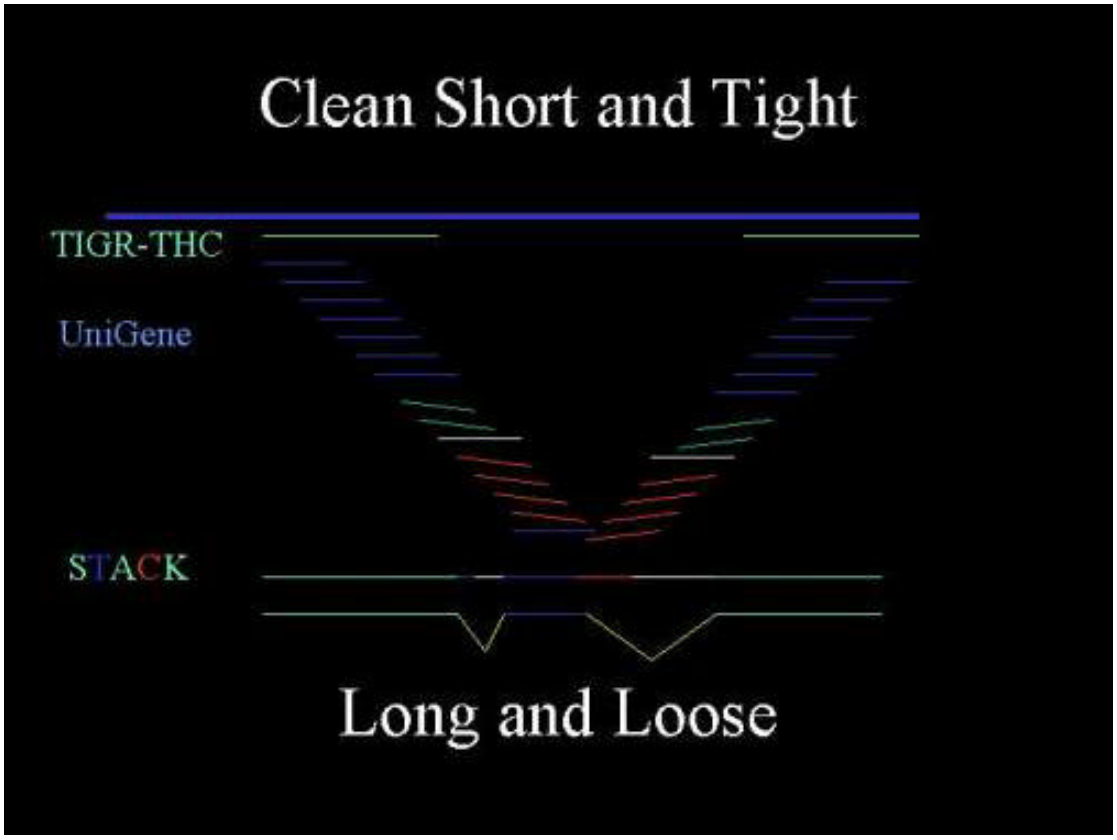
slide20.jpg (504x378x24b jpeg)

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

SANBI

slide21.jpg (504x378x24b jpeg)



slide22.jpg (504x378x24b jpeg)

Data apprehension and input format.

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

SANBI

slide23.jpg (504x378x24b jpeg)

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)

SANBI

slide24.jpg (504x378x24b jpeg)

Initial clustering

- Stepwise clustering 'Multistate'.
  - sequence identity
  - annotation
  - verification

SANBI

slide25.jpg (504x378x24b jpeg)

Assembly

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

SANBI

slide26.jpg (504x378x24b jpeg)

Alignment processing

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

slide27.jpg (504x378x24b jpeg)

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%

SANBI

slide28.jpg (504x378x24b jpeg)

Output

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

SANBI

slide29.jpg (504x378x24b jpeg)

Alignment scoring methods:

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

SANBI

slide30.jpg (504x378x24b jpeg)

EST clustering methods

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

SANBI

slide31.jpg (504x378x24b jpeg)

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.

slide32.jpg (504x378x24b jpeg)

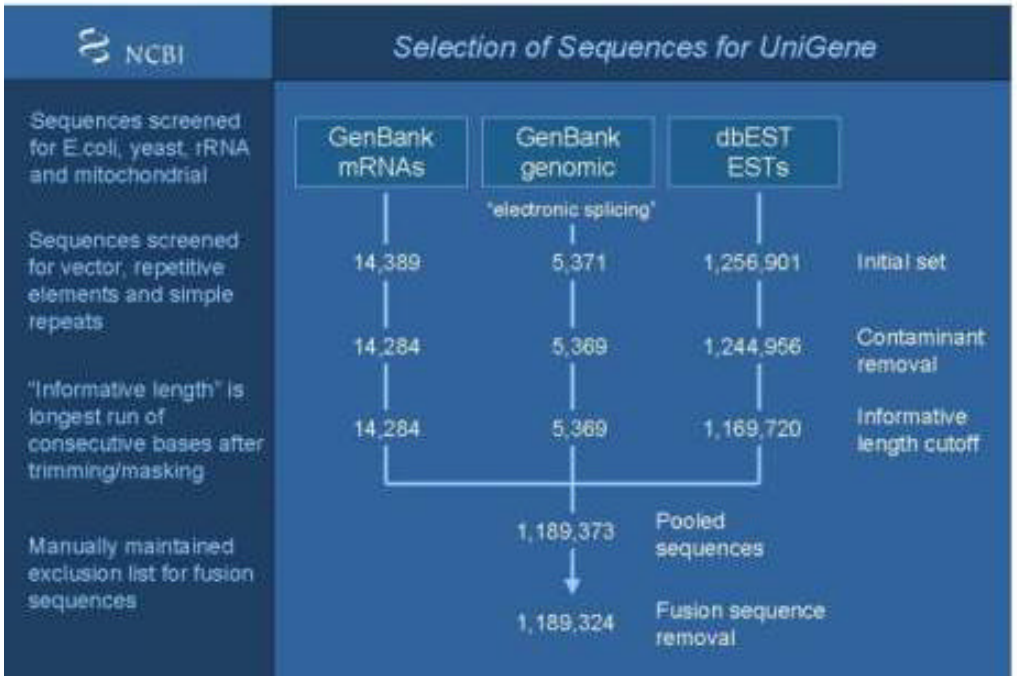
TIGR_ASSEMBLER

- 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

SANBI

slide33.jpg (504x378x24b jpeg)

UniGene



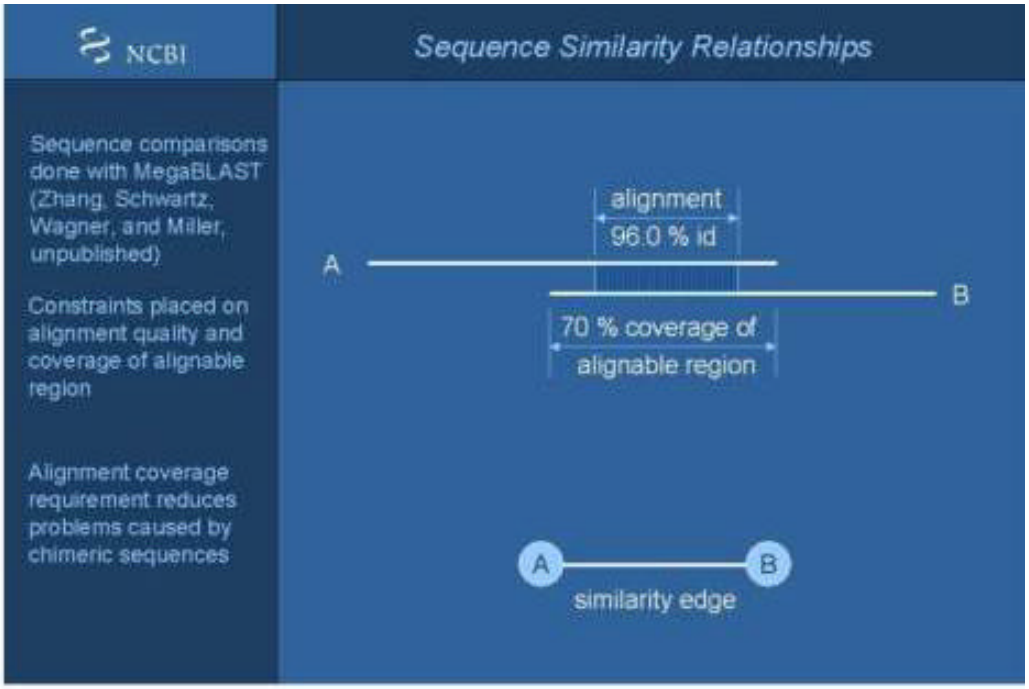
slide34.jpg (504x378x24b jpeg)

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

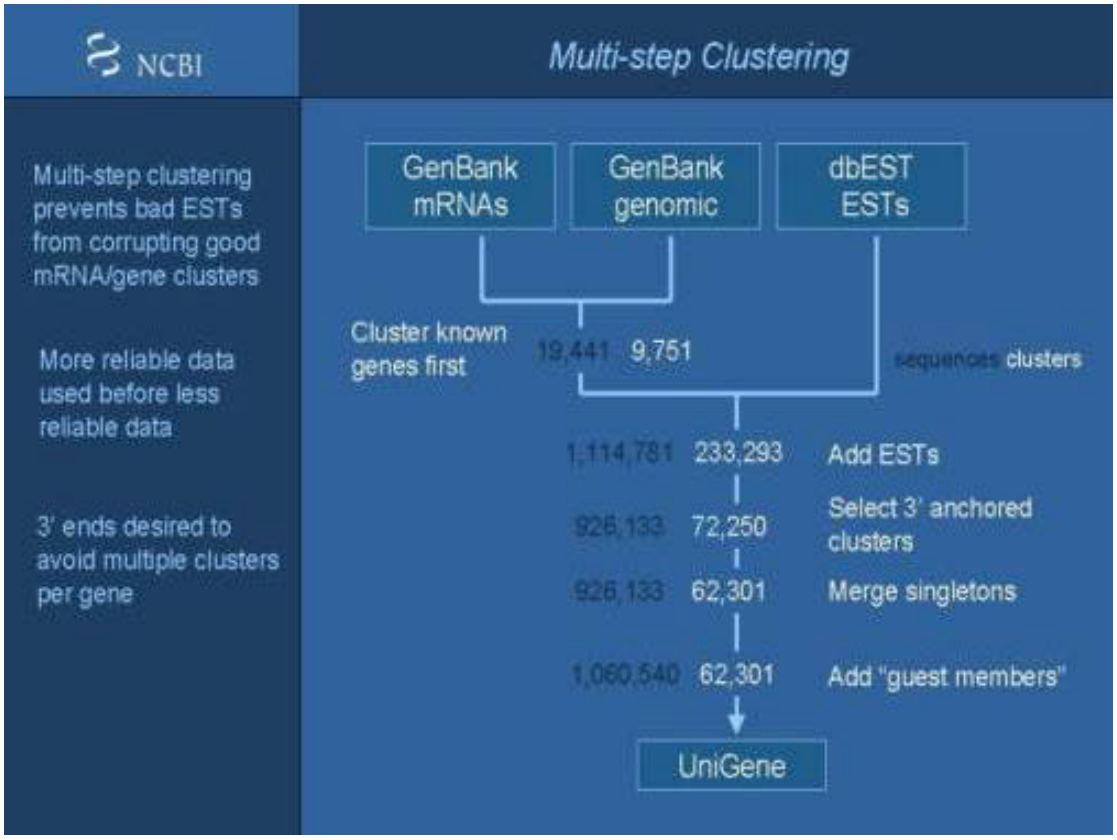
SANBI

slide35.jpg (504x378x24b jpeg)



SANBI Wagner et al. CSH 1999

slide36.jpg (504x378x24b jpeg)



slide37.jpg (504x378x24b jpeg)

Fragmentation Comparison

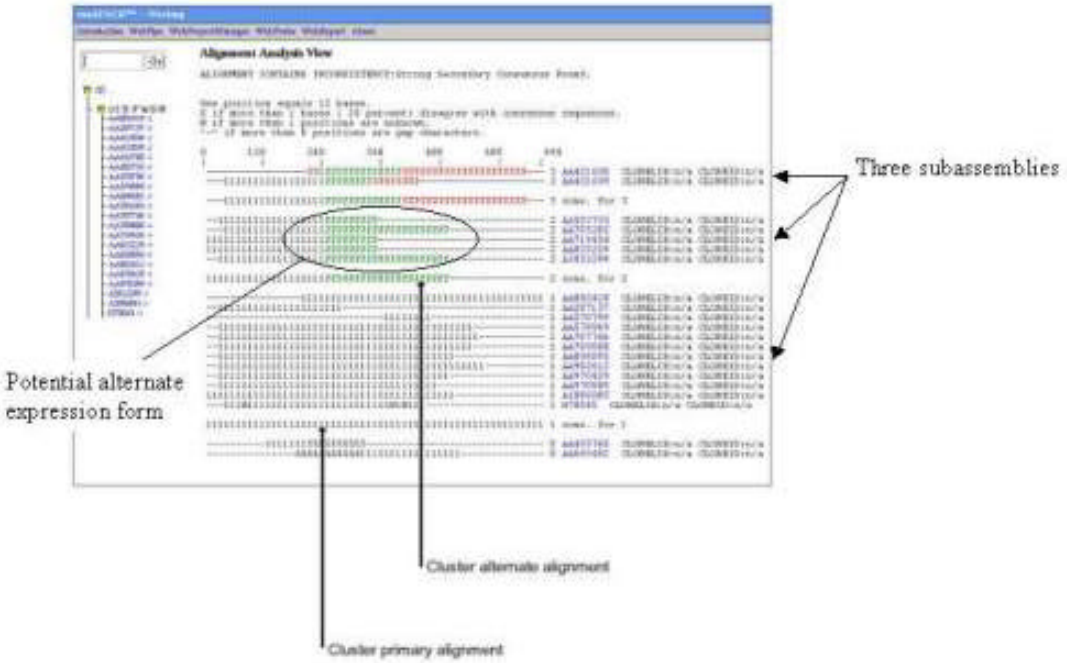
Methodology	Input Sequences	Singleton Groups	%Singleton Groups
TIGR Gene Index	626 163	135 140	21.83
STACK_PACK	415 833	58 070	13.96

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

SANBI

slide38.jpg (504x378x24b jpeg)

Alignment Analysis



slide39.jpg (504x378x24b jpeg)

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.

slide40.jpg (504x378x24b jpeg)

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

SANBI

slide41.jpg (504x378x24b jpeg)

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

SANBI

slide42.jpg (504x378x24b jpeg)