

# EST clustering

Lorenzo Cerutti  
Swiss Institute of Bioinformatics

EMBNNet course, March 2003

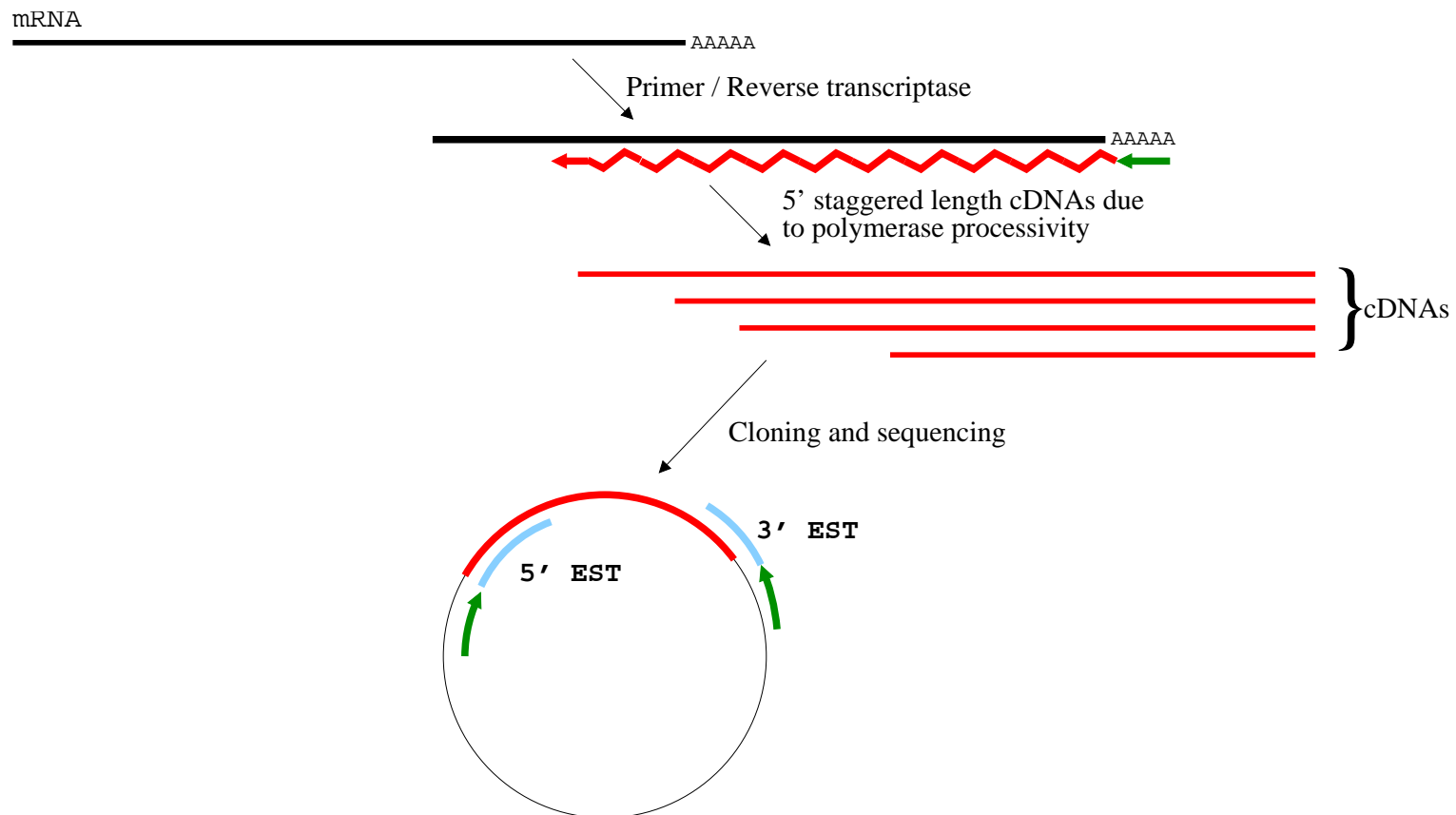
# Outline

- Introduction
  - What are ESTs
  - Gene Indexes
- Clustering and assembling
  - Pre-processing ESTs
  - Clustering
  - Assembling, processing, and cluster joining
- Gene Indexes databases
  - UniGene
  - TIGR Gene Indexes
  - STACK
  - trEST

Color code: *Keywords*, *Databases*, *Software*

# Expressed sequence tags (ESTs)

- ESTs represent partial sequences of cDNA clones (average  $\sim 360$  bp).
- Single-pass reads from the 5' and/or 3' ends of cDNA clones.



# Interest for ESTs

- ESTs represent the most extensive available survey of the transcribed portion of genomes.
- ESTs are indispensable for gene structure prediction, gene discovery and genomic mapping.
- Characterization of splice variants and alternative polyadenylation.
- *In silico* differential display and gene expression studies (specific tissue expression, normal/disease states).
- SNP data mining.
- High-volume and high-throughput data production at low cost.
- There are 15,795,385 of EST entries in GenBank (*dbEST*) (February 14, 2003) (<http://www.ncbi.nlm.nih.gov/dbEST/>):
  - 5,024,729 entries of human ESTs;
  - 3,691,862 entries of mouse ESTs;

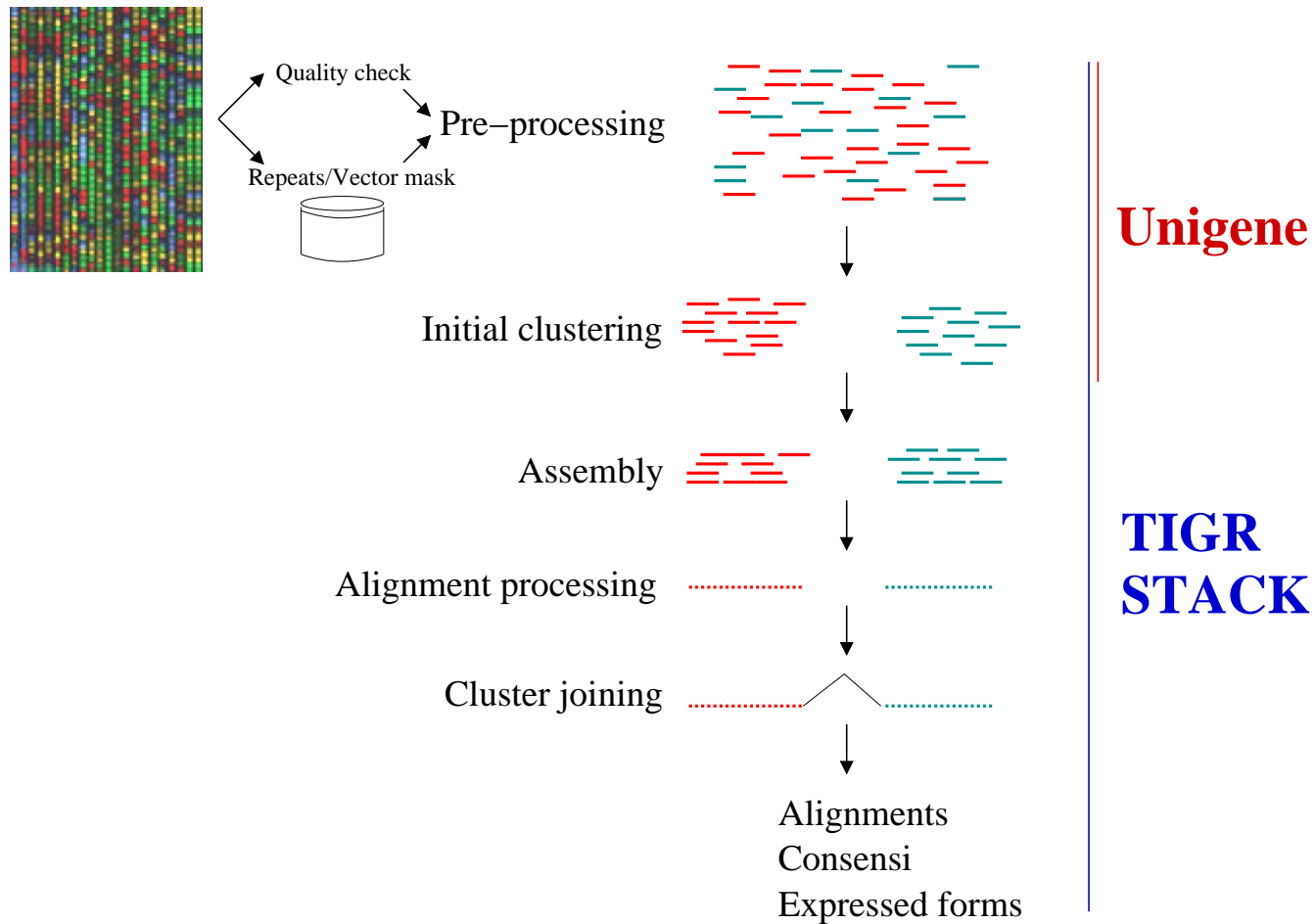
# Low data quality of ESTs

- High error rates ( $\sim 1/100$ ) because of the sequence reading single-pass.
- Sequence compression and frame-shift errors due to the sequence reading single-pass.
- A single EST represents only a partial gene sequence.
- Not a defined gene/protein product.
- Not curated in a highly annotated form.
- High redundancy in the data  $\Rightarrow$  huge number of sequences to analyze.

# Improving ESTs: Clustering, Assembling and Gene indices

- The value of ESTs is greatly enhanced by *clustering* and *assembling*.
  - solving redundancy can help to correct errors;
  - longer and better annotated sequences;
  - easier association to mRNAs and proteins;
  - detection of splice variants;
  - fewer sequences to analyze.
- *Gene indices*: All expressed sequences (as ESTs) concerning a single gene are grouped in a single index class, and each index class contains the information for only one gene.
- Different clustering/assembly procedures have been proposed with associated resulting databases (gene indices):
  - *UniGene* (<http://www.ncbi.nlm.nih.gov/UniGene>)
  - *TIGR Gene Indices* (<http://www.tigr.org/tdb/tgi>)
  - *STACK* (<http://www.sanbi.ac.za/Dbases.html>)

# EST clustering pipeline



# Pre-processing

# Data source

- The data sources for clustering can be in-house, proprietary, public database or a hybrid of this (chromatograms and/or sequence files).
- Each EST must have the following information:
  - A sequence ID (ex. sequence-run ID);
  - Location in respect of the poly A (3' or 5');
  - The CLONE ID from which the EST has been generated;
  - Organism;
  - Tissue and/or conditions;
  - The sequence.
- The EST can be stored in FASTA format:

```
>T27784 EST16067 Human Endothelial cells Homo sapiens cDNA 5'  
CCCCCGTCTCTTTAAAAATATATATATTTTAAATATACTTAAATATATATTTCTAATATC  
TTTAAATATATATATATATATTTNAAAGACCAATTTATGGGAGANTTGCACACAGATGTGAA  
ATGAATGTAATCTAATAGANGCCTAATCAGCCCACCATGTTCTCCACTGAAAAATCCTCT  
TTCTTTGGGGTTTTTTCTTTCTTTCTTTTTTTGATTTTGCCTGGACGGTGACGTCAGCCAT  
GTACAGGATCCACAGGGTGGTGTCAAATGCTATTGAAATTNTGTTGAATTGTATACTTT  
TTCCTTTTTGATAATTAACCATGTAAAAAATGAACGCTACTACTATAGTAGAATTGAT
```

# Pre-processing

- EST pre-processing consists in a number of essential steps to minimize the chance to cluster unrelated sequences.
  - Screening out low quality regions:
    - Low quality sequence readings are error prone.
    - Programs as *Phred* (*Ewig et al., 98*) read chromatograms and assesses a quality value to each nucleotide.
  - Screening out contaminations.
  - Screening out vector sequences (vector clipping).
  - Screening out repeat sequences (repeats masking).
  - Screening out low complexity sequences.
- Dedicated software are available for these tasks:
  - *RepeatMasker* (*Smit and Green*, <http://ftp.genome.washington.edu/RM/RepeatMasker.html>);
  - *VecScreen* (<http://www.ncbi.nlm.nih.gov/VecScreen>);
  - *Lucy* (*Chou and Holmes, 01*);
  - ...

# Vector-clipping and contaminations

- Vector sequences can skew clustering even if a small vector fragment remains in each read. Therefore vector sequences must be removed (*Vector-clipping*):
  - Delete 5' and 3' regions corresponding to the vector used for cloning.
  - Detection of vector sequences is not a trivial task, because they normally lies in the low quality region of the sequence.
  - *UniVec* is a non-redundant vector database available from NCBI:  
<http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html>
- Contaminations can also skew clustering and therefore must be removed:
  - Find and delete:
    - bacterial DNA, yeast DNA, and other contaminations;
- Standard pairwise alignment programs are used for the detection of vector sequences and other contaminants (for example *cross-match*, *BLASTN*, *FASTA*). They are reasonably fast and accurate.

# Repeats masking

- Some repetitive elements found in the human genome:

	Length	Copy number	Fraction of the genome
LINEs (long interspersed elements)	6-8 kb	850,000	21%
SINEs (short interspersed elements)	100-300 bp	1,500,000	13%
LTR (autonomous)	6-11 kb	} 450,000	8%
LTR (non-autonomous)	1.5-3 kb		
DNA transposons (autonomous)	2-3 kb	} 300,000	3%
DNA transposons (non-autonomous)	80-3000 bp		
SSRs (simple sequence repeats or microsatellite and minisatellites)			3%

# Repeats masking

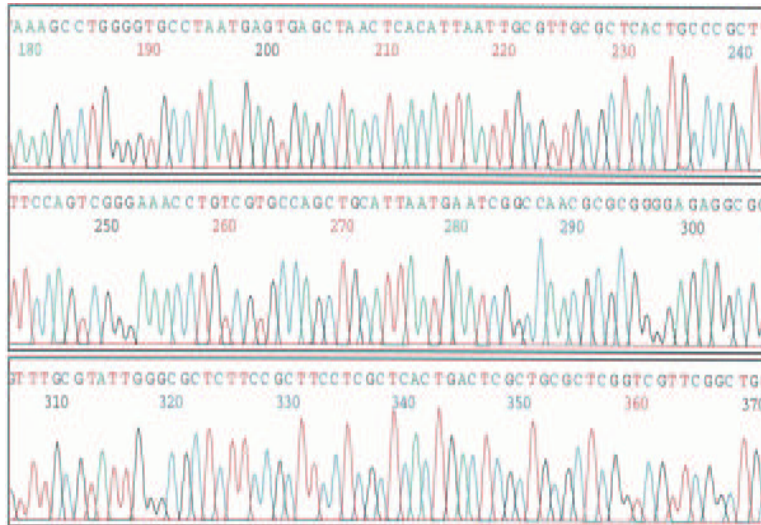
- Repeated elements:
  - They represent a big part of the mammalian genome.
  - They are found in a number of genomes (plants, ...)
  - They induce errors in clustering and assembling.
  - They should be masked, not deleted, to avoid false sequence assembling.
  - ... but also interesting elements for evolutionary studies.
  - SSRs important for mapping of diseases.
- Tools to find repeats:
  - *RepeatMasker* has been developed to find repetitive elements and low-complexity sequences. RepeatMasker uses the cross-match program for the pairwise alignments (<http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker>).
  - *MaskerAid* improves the speed of RepeatMasker by  $\sim 30$  folds using WU-BLAST instead of cross-match (<http://sapiens.wustl.edu/maskeraid>)
  - *RepBase* is a database of prototypic sequences representing repetitive DNA from different eukaryotic species.: [http://www.girinst.org/Repbase\\_Update.html](http://www.girinst.org/Repbase_Update.html).

# Low complexity masking

- Low complexity sequences contains an important bias in their nucleotide compositions (poly A tracts, AT repeats, etc.).
- Low complexity regions can provide an artifactual basis for cluster membership.
- Clustering strategies employing alignable similarity in their first pass are very sensitive to low complexity sequences.
- Some clustering strategies are insensitive to low complexity sequences, because they weight sequences in respect to their information content (ex. *d2-cluster*).
- Programs as *DUST* (NCBI) can be used to mask low complexity regions.

# Pre-processing

Base calling  
Select high quality reads



```
CCCCCGTCTCTTTAAAAATATATATATTTTTAAATATACTTAAATATATATTTCTAAATATC
TTTAAATATATATATATATATTTNAAAGACCAATTTATGGGAGANTTGCACACAGATGTGAA
ATGAATGTAATCTAATAGANGCCTAATCAGCCACCATGTTCTCCACTGAAAAATCCTCT
TTCTTTGGGGTTTTTCTTTCTTTCTTTTGGATTTTGCACCTGGACGGTGACGTCAGCCAT
GTACAGGATCCACAGGGGTGGTGTCAAATGCTATTGAAATNTGTTGAATTGTATACTTT
TTCACTTTTTGATAATTAACCATGTAAAAAATGAACGCTACTACTATAGTAGAATTGAT
```

↓ Vector clipping

```
CCCCCGTCTCTTTAAAAATATATATATTTTTAAATATACTTAAATATATATTTCTAAATATC
TTTAAATATATATATATATATTTNAAAGACCAATTTATGGGAGANTTGCACACAGATGTGAA
ATGAATGTAATCTAATAGANGCCTAATCAGCCACCATGTTCTCCACTGAAAAATCCTCT
TTCTTTGGGGTTTTTCTTTCTTTCTTTTGGATTTTGCACCTGGACGGTGACGTCAGCCAT
GTACAGGATCCACAGGGGTGGTGTCAAATGCTATTGAAATNTGTTGAATTGTATACTTT
TTCACTTTTTGATAATTAACCATGTAAAAAATGXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

↓ Repeat/Low complexity masking

```
CCCCCGTCTCTTTAAAAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
TTNAAAGACCAATTTATGGGAGANTTGCACACAGATGTGAA
ATGAATGTAATCTAATAGANGCCTAATCAGCCACCATGTTCTCCACTGAAAAATCCTCT
TTCTTTGGGGTTTTTCTTTCTTTCTTTTGGATTTTGCACCTGGACGGTGACGTCAGCCAT
GTACAGGATCCACAGGGGTGGTGTCAAATGCTATTGAAATNTGTTGAATTGTATACTTT
TTCACTTTTTGATAATTAACCATGTAAAAAATGXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

↓ Sequence ready for clustering

```
CCCCCGTCTCTTTAAAAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
TTNAAAGACCAATTTATGGGAGANTTGCACACAGATGTGAA
ATGAATGTAATCTAATAGANGCCTAATCAGCCACCATGTTCTCCACTGAAAAATCCTCT
TTCTTTGGGGTTTTTCTTTCTTTCTTTTGGATTTTGCACCTGGACGGTGACGTCAGCCAT
GTACAGGATCCACAGGGGTGGTGTCAAATGCTATTGAAATNTGTTGAATTGTATACTTT
TTCACTTTTTGATAATTAACCATGTAAAAAATG
```

# Clustering

# EST clustering

- The goal of the clustering process is to incorporate overlapping ESTs which tag the same transcript of the same gene in a single *cluster*.
- For clustering, we measure the *similarity* (distance) between any 2 sequences. The distance is then reduced to a simple binary value: accept or reject two sequences in the same cluster.
- Similarity can be measured using different algorithms:
  - Pairwise alignment algorithms:
    - Smith-Waterman is the most sensitive, but time consuming (ex. *cross-match*);
    - Heuristic algorithms, as *BLAST* and *FASTA*, trade some sensitivity for speed.
  - Non-alignment based scoring methods:
    - d2\_cluster algorithm: based on word comparison and composition (word identity and multiplicity) (*Burke et al.*, 99). No alignments are performed  $\Rightarrow$  fast.
  - Pre-indexing methods.
  - Purpose-built alignments based clustering methods.

# Loose and stringent clustering

- *Stringent clustering:*
  - Greater initial fidelity;
  - One pass;
  - Lower coverage of expressed gene data;
  - Lower cluster inclusion of expressed gene forms;
  - Shorter consensi.
- *Loose clustering:*
  - Lower initial fidelity;
  - Multi-pass;
  - Greater coverage of expressed gene data;
  - Greater cluster inclusion of alternate expressed forms.
  - Longer consensi;
  - Risk to include paralogs in the same gene index.

# Supervised and unsupervised EST clustering

- *Supervised clustering*
  - ESTs are classified with respect to known reference sequences or "seeds" (full length mRNAs, exon constructs from genomic sequences, previously assembled EST cluster consensus).
- *Unsupervised clustering*
  - ESTs are classified without any prior knowledge.
- The three major gene indices use different EST clustering methods:
  - *TIGR Gene Index* uses a stringent and supervised clustering method, which generate shorter consensus sequences and separate splice variants.
  - *STACK* uses a loose and unsupervised clustering method, producing longer consensus sequences and including splice variants in the same index.
  - A combination of supervised and unsupervised methods with variable levels of stringency are used in *UniGene*. No consensus sequences are produced.

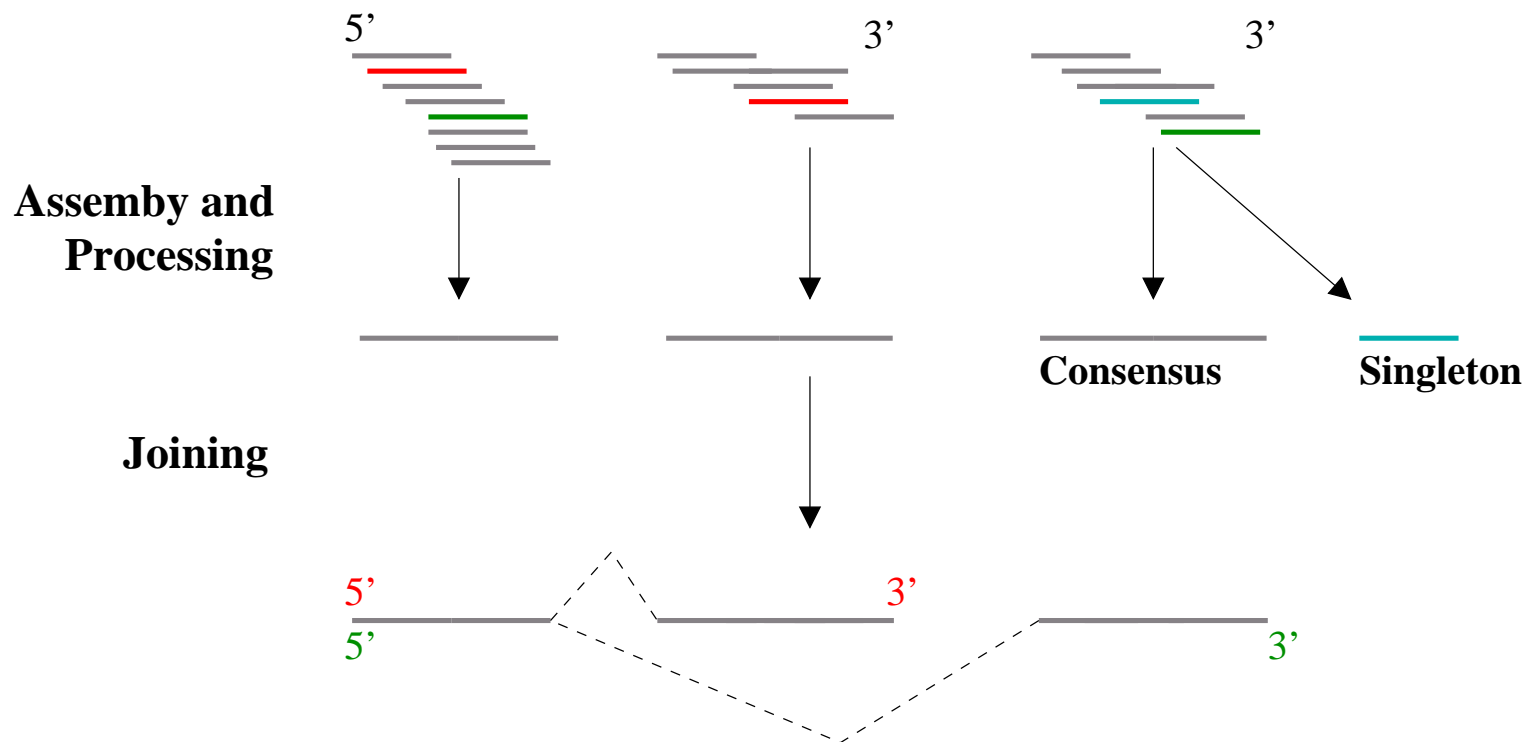
# **Assembling, processing, and cluster joining**

# Assembly and processing

- A *multiple alignment* for each cluster can be build (*assembly*) and consensus sequences generated (*processing*).
- A number of program are available for assembly and processing:
  - *PHRAP* (<http://www.genome.washington.edu/UWGC/analysistools/Phrap.cfm>);
  - *TIGR\_ASSEMBLER* (*Sutton et al.*, 95);
  - *CRAW* (*Burke et al.*, 98);
  - ...
- Assembly and processing result in the production of *consensus sequences* and *singletons*.
- Consensus sequences are useful:
  - helpful to visualize splice variants;
  - reduced size of data to analyze;
  - gene structure;
  - ...

# Cluster joining

- All ESTs generated from the same cDNA clone correspond to a single gene.
- Generally the original cDNA clone information is available ( $\sim 90\%$ ).
- Using the cDNA clone information and the 5' and 3' reads information, clusters can be joined.



# UniGene

- *UniGene Gene Indices* are available for a number of organisms.
- *UniGene* clusters are produced with a supervised procedure: ESTs are clustered using GenBank CDSs and mRNAs data as "seed" sequences.
- No attempts to produce contigs or consensus sequences.
- *UniGene* uses pairwise sequence comparison at various levels of stringency to group related sequences, placing closely related and alternatively spliced transcripts into one cluster.
- *UniGene* web site: <http://www.ncbi.nlm.nih.gov/UniGene>.

# UniGene procedure

- Screen for contaminants, repeats, and low-complexity regions in *GenBank*.
  - Low-complexity are detected using *Dust*.
  - Contaminants (vector, linker, bacterial, mitochondrial, ribosomal sequences) are detected using pairwise alignment programs.
  - Repeat masking of repeated regions (*RepeatMasker*).
  - Only sequences with at least 100 informative bases are accepted.
- Clustering procedure.
  - Build clusters of genes and mRNAs (*GenBank*).
  - Add ESTs to previous clusters (*megablast*).
  - ESTs that join two clusters of genes/mRNAs are discarded.
  - Any resulting cluster without a polyadenylation signal or at least two 3' ESTs is discarded.
  - The resulting clusters are called *anchored clusters* since their 3' end is supposed known.

# UniGene procedure

- Ensures that the 5' and 3' ESTs from the same cDNA clone belongs to the same cluster.
- ESTs that have not been clustered, are reprocessed with lower level of stringency. ESTs added during this step are called *guest members*.
- Clusters of size 1 (containing a single sequence) are compared against the rest of the clusters with a lower level of stringency and merged with the cluster containing the most similar sequence.
- For each build of the database, *clusters IDs change* if clusters are split or merged.

# TIGR Gene Indices

- *TIGR* produces Gene Indices for a number of organisms (<http://www.tigr.org/tdb/tgi>).
- *TIGR Gene Indices* are produced using strict supervised clustering methods.
- Clusters are assembled in consensus sequences, called *tentative consensus* (TC) sequences, that represent the underlying mRNA transcripts.
- The *TIGR Gene Indices* building method tightly groups highly related sequences and discard under-represented, divergent, or noisy sequences.
- *TIGR Gene Indices* characteristics:
  - separate closely related genes into distinct consensus sequences;
  - separate splice variants into separate clusters;
  - low level of contamination.
- TC sequences can be used for genome annotation, genome mapping, and identification of orthologs/paralogs genes.

# TIGR Gene Indices procedure

- EST sequences recovered from *dbEST* (<http://www.ncbi.nlm.nih.gov/dbEST>);
- Sequences are trimmed to remove:
  - vectors
  - polyA/T tails
  - adaptor sequences
  - bacterial sequences
- Get *expressed transcripts* (ETs) from *EGAD* (<http://www.tigr.org/tdb/egad/egad.shtml>):
  - *EGAD* (Expressed Gene Anatomy Database) is based on mRNA and CDS (coding sequences) from *GenBank*.
- Get TCs and singletons from previous database build.

# TIGR Gene Indices procedure

- Supervised and strict clustering:
  - Use ETs, TCs, and CDSs as template;
  - Compare cleaned ESTs to the template using *FLAST* (a rapid pairwise comparison program).
  - Sequences are grouped in the same cluster if both conditions are true:
    - they share  $\geq 95\%$  identity over 40 bases or longer regions
    - $< 20$  bases of mismatch at either end
- Each cluster is assembled using *CAP3* assembling program to produce *tentative consensus* (TC) sequences.
  - *CAP3* can generate multiple consensus sequences for each cluster
  - *CAP3* rejects chimeric, low-quality and non-overlapping sequences.
  - New TCs resulting from the joining or splitting of previous TCs, get a *new TC ID*.

# TIGR Gene Indices procedure

- Build TCs are loaded in the *TIGR Gene Indices* database and annotated using information from GenBank and/or protein homology.
- Track of the old TC IDs is maintained through a relational database.
- References:
  - Quackenbush *et al.* (2000) *Nucleic Acid Research*,**28**, 141-145.
  - Quackenbush *et al.* (2001) *Nucleic Acid Research*,**29**, 159-164.

# STACK

- *STACK* concentrates on human data.
- Based on "loose" unsupervised clustering, followed by strict assembly procedure and analysis to identify and characterize sequence divergence (alternative splicing, etc).
- The "loose" clustering approach, *d2\_cluster*, is not based on alignments, but performs comparisons via non-contextual assessment of the composition and multiplicity of words within each sequence.
- Because of the "loose" clustering, *STACK* produces longer consensus sequences than TIGR Gene Indices.
- *STACK* also integrates  $\sim 30\%$  more sequences than UniGene, due to the "loose" clustering approach

# STACK procedure

- Sub-partitioning.
  - Select human ESTs from *GenBank*;
  - Sequences are grouped in tissue-based categories ("bins"). This will allow further specific tissue transcription exploration.
  - A "bin" is also created for sequences derived from disease-related tissues.
- Masking.
  - Sequences are masked for repeats and contaminants using *cross-match*:
    - Human repeat sequences (*RepBase*);
    - Vector sequences;
    - Ribosomal and mitochondrial DNA, other contaminants.

# STACK procedure

- "Loose" clustering using *d2\_cluster*.
  - The algorithm looks for the co-occurrence of  $n$ -length words ( $n = 6$ ) in a window of size 150 bases having at least 96% identity.
  - Sequences shorter than 50 bases are excluded from the clustering process.
  - Clusters highly related sequences.
  - Clusters also sequences related by rearrangements or alternative splicing.
  - Because *d2\_cluster* weights sequences according to their information content, masking of low complexity regions is not required.
- Assembly.
  - The assembly step is performed using *Phrap*.
  - *STACK* don't use quality information available from chromatograms.
  - The lack of trace information is largely compensated by the redundancy of the ESTs data.
  - Sequences that cannot be aligned with Phrap are extracted from the clusters (singletons) and processed later.

# STACK procedure

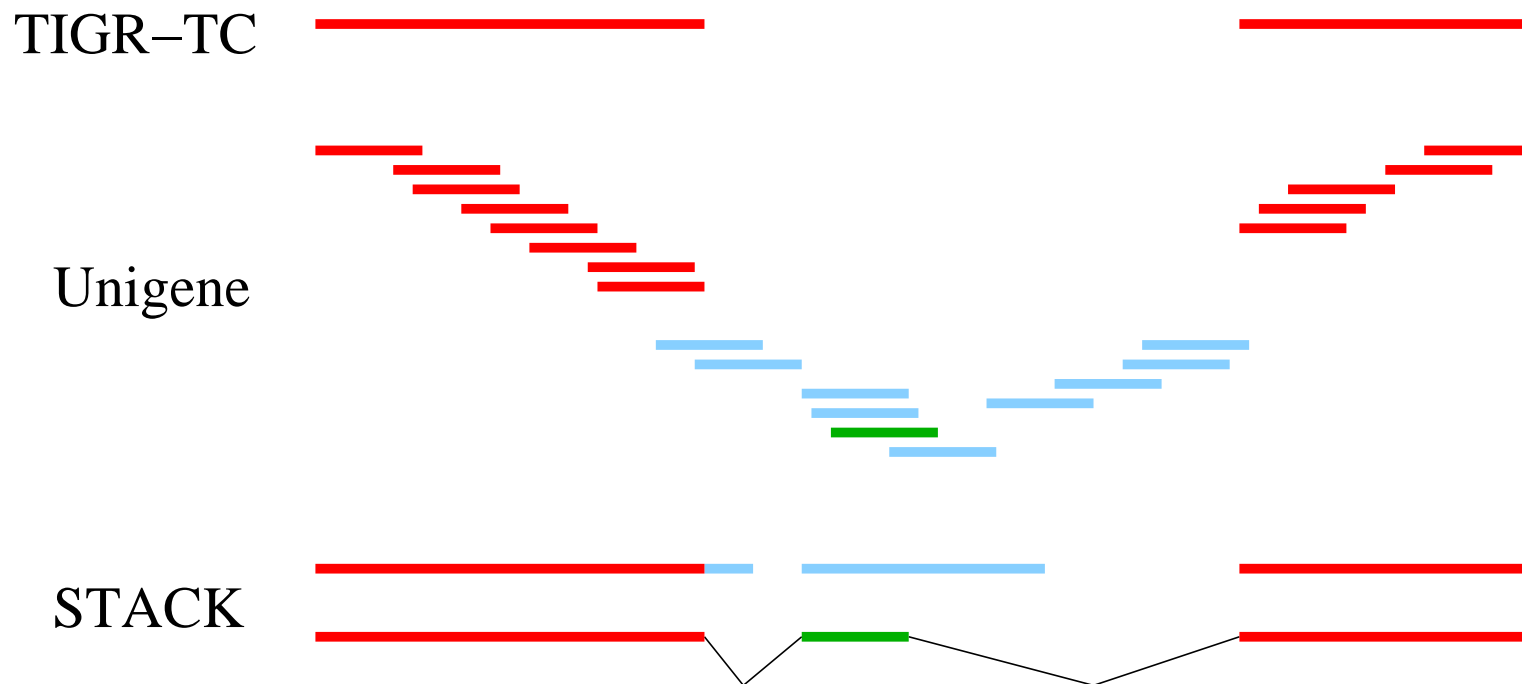
- Alignment analysis.
  - The *CRAW* program is used in the first part of the alignment analysis.
  - *CRAW* generates consensus sequence with maximized length.
  - *CRAW* partitions a cluster in sub-ensembles if  $\geq 50\%$  of a 100 bases window differ from the rest of the sequences of the cluster.
  - Rank the sub-ensembles according to the number of assigned sequences and number of called bases for each sub-ensemble (*CONTIGPROC*).
  - Annotate polymorphic regions and alternative splicing.
- Linking.
  - Joins clusters containing ESTs with shared clone ID.
  - Add singletons produced by Phrap in respect to their clone ID.

# STACK procedure

- STACK update.
  - New ESTs are searched against existing consensus and singletons using *cross-match*.
  - Matching sequences are added to extend existing clusters and consensus.
  - Non-matching sequences are processed using *d2-cluster* against the entire database and the new produces clusters are renamed  $\Rightarrow$  *Gene Index ID change*.
- STACK outputs.
  - Primary consensus for each cluster in FASTA format.
  - Alignments from Phrap in GDE (Genetic Data Environment) format.
  - Sequence variations and sub-consensus (from CRAW processing).
- References.
  - Miller *et al.* (1999) *Genome Research*,**9**, 1143-1155.
  - Christoffels *et al.* (2001) *Nucleic Acid Research*,**29**, 234-238.

# EST clustering procedures

Clean, short, and tight



Long and loose

# trEST

- *trEST* is an attempt to produce contigs from clusters of ESTs and to translate them into proteins.
- *trEST* uses UniGene clusters and clusters produced from in-house software.
- To assemble clusters *trEST* uses *Phrap* and *CAP3* algorithms.
- Contigs produced by the assembling step are translated into protein sequences using the *ESTscan* program, which corrects most of the frame-shift errors and predicts transcripts with a position error of few amino acids.
- You can access *trEST* via the *HITS* database (<http://hits.isb-sib.ch>).

# The end