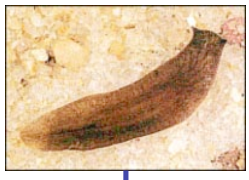
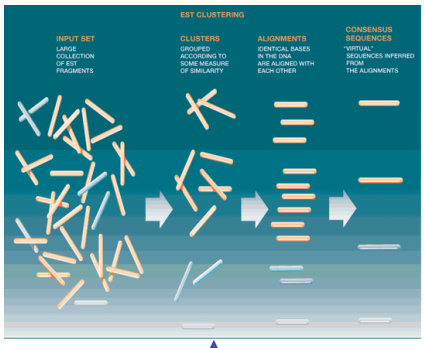
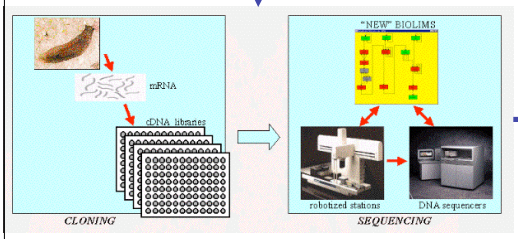
Applied Bioinformatics
WS2007/2008

Ingo Ebersberger
Arndt von Haeseler
Anne Kupczok
Heiko Schmidt

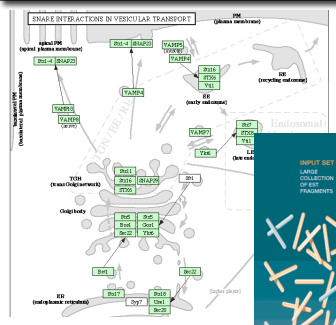
Day 1: Data generation and processing

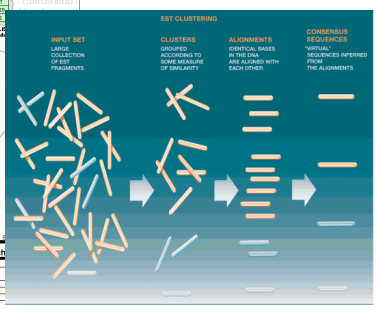


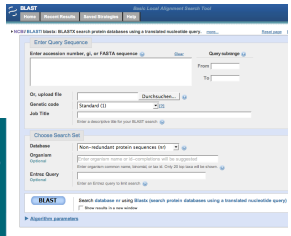


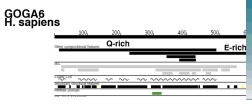


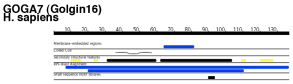
Day 2: Annotating EST sequences

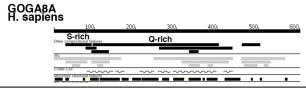


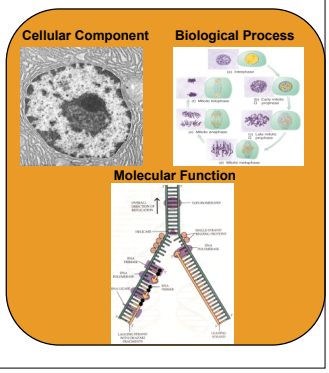




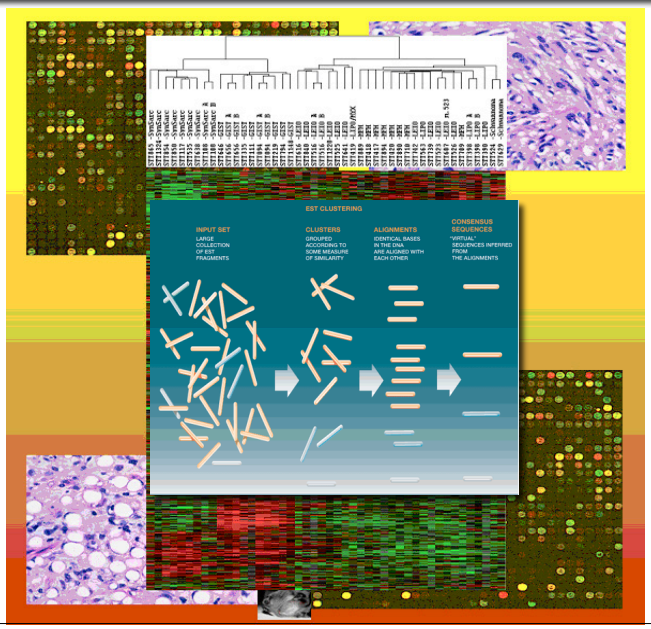






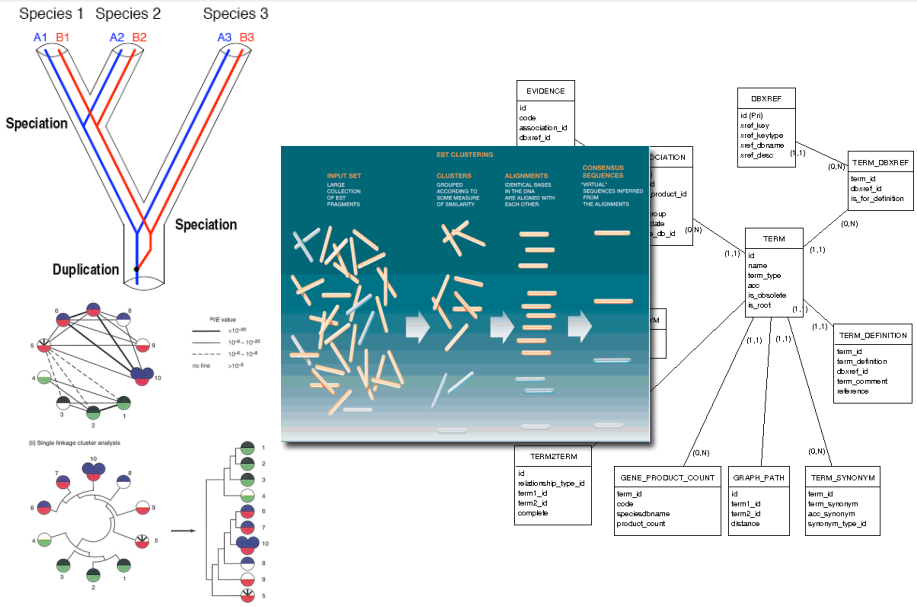


Day 3: Analysis of Gene expression



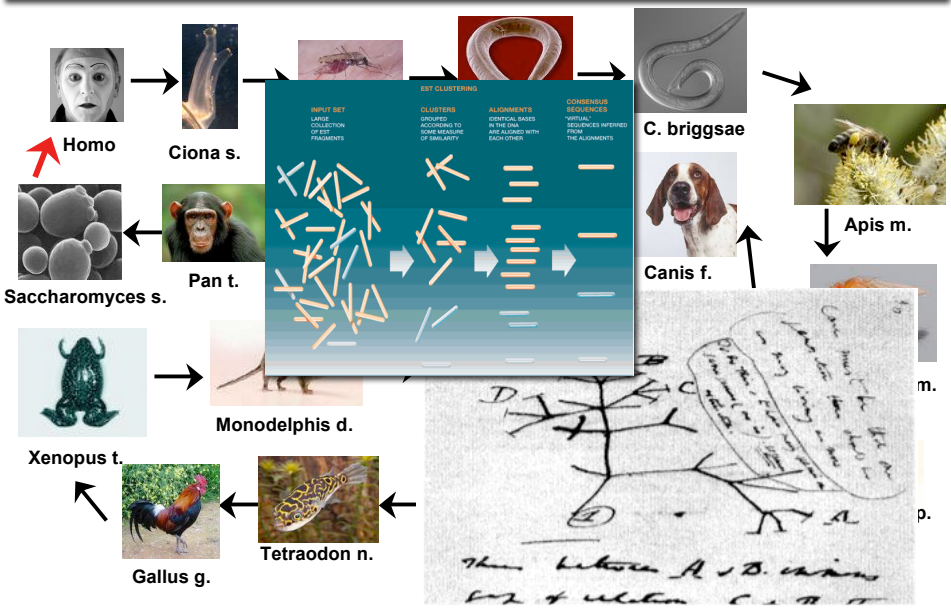
Day 4: Evolutionary Relationships of genes and Data Management

CIBIV MFPL



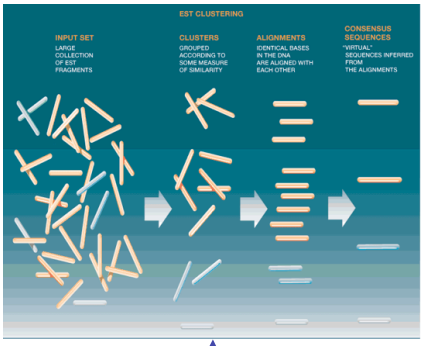
Day 5: Phylogeny Reconstruction

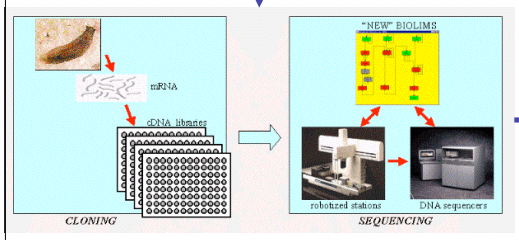
CIBIV MFPL



Day 1: Data generation and processing

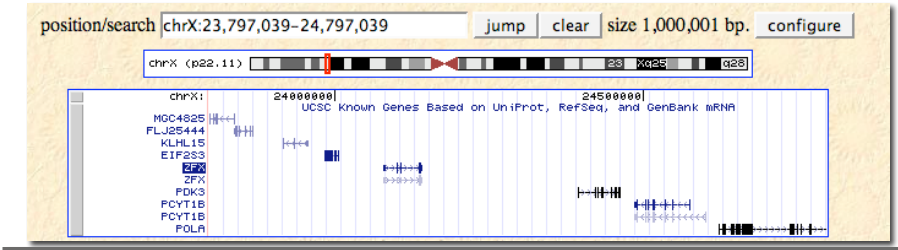
CIBIV MFPL

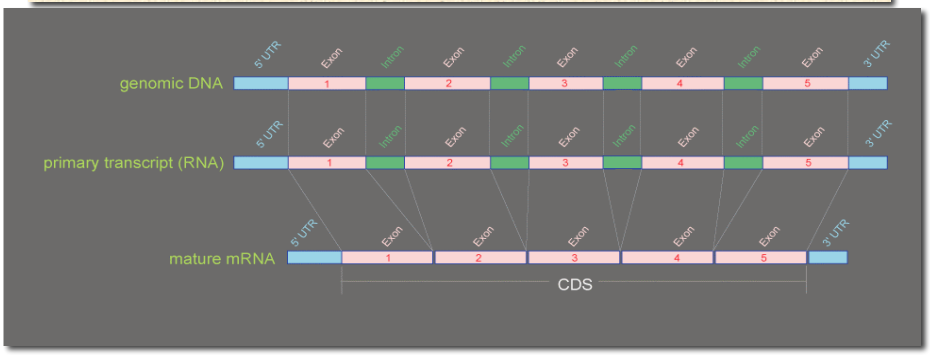




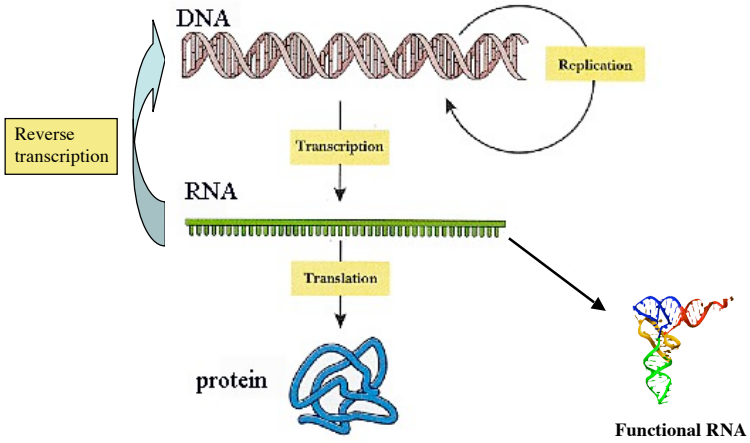
Identifying protein Coding Information in a Genome

CIBIV MFPL



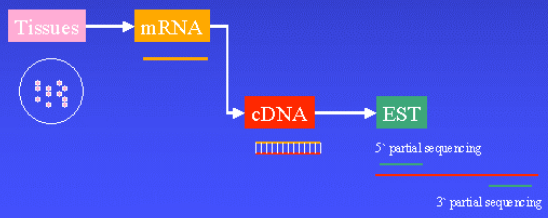


From DNA to Gene products

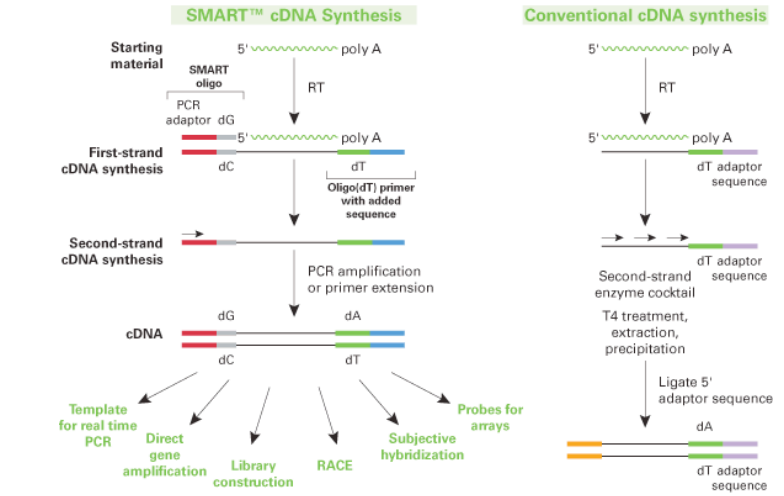


Expressed Sequence Tag (EST)

Partial cDNA sequences of genes expressed in different tissues

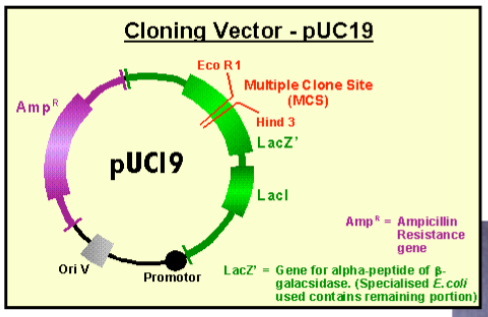


cDNA Generation: The SMART approach

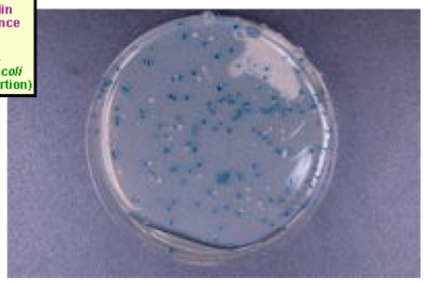


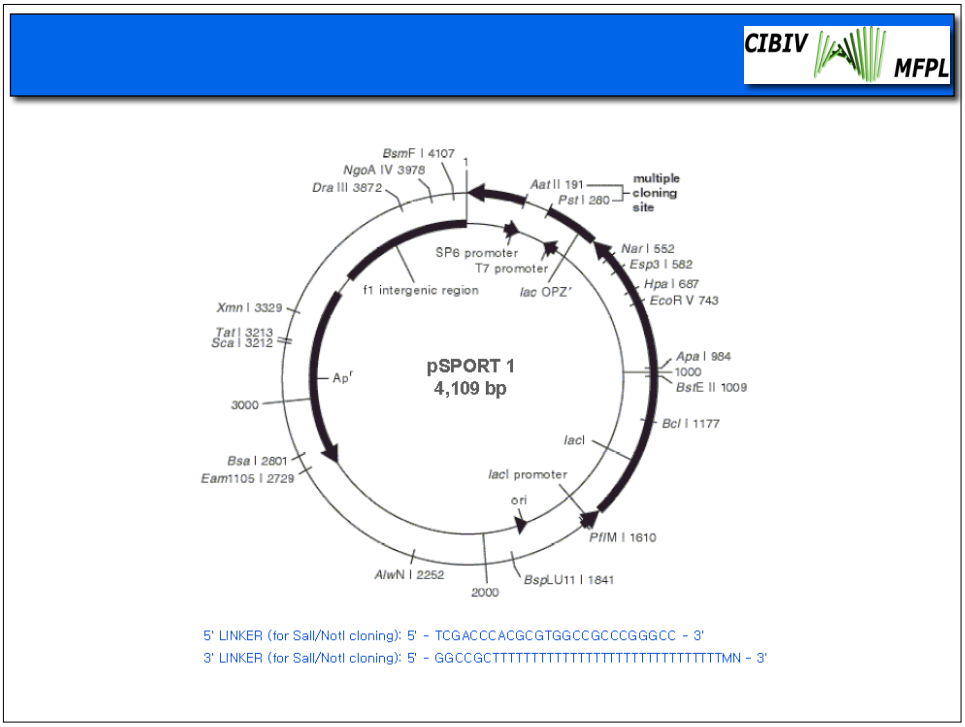
SMART™ (Switching Mechanism at 5' End of RNA Template)

Cloning of cDNA

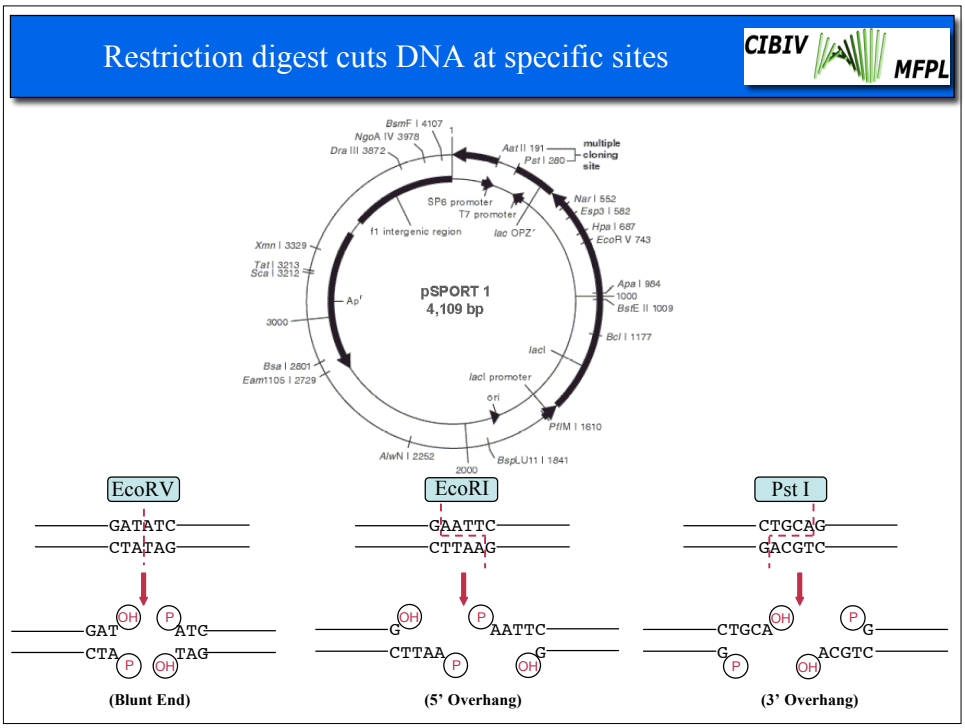


- Origin of Replication (copy number)
- Selection Marker
- Regulatory Elements for Gene Expression
- Cloning site (MCS)

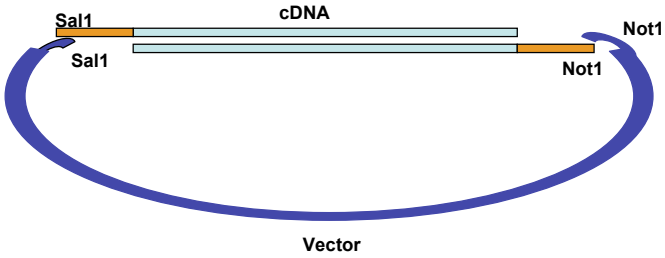




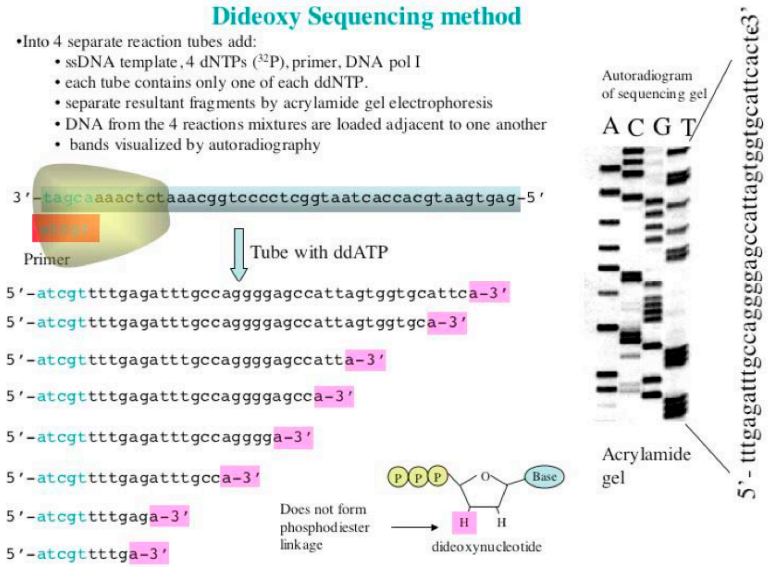
Restriction digest cuts DNA at specific sites



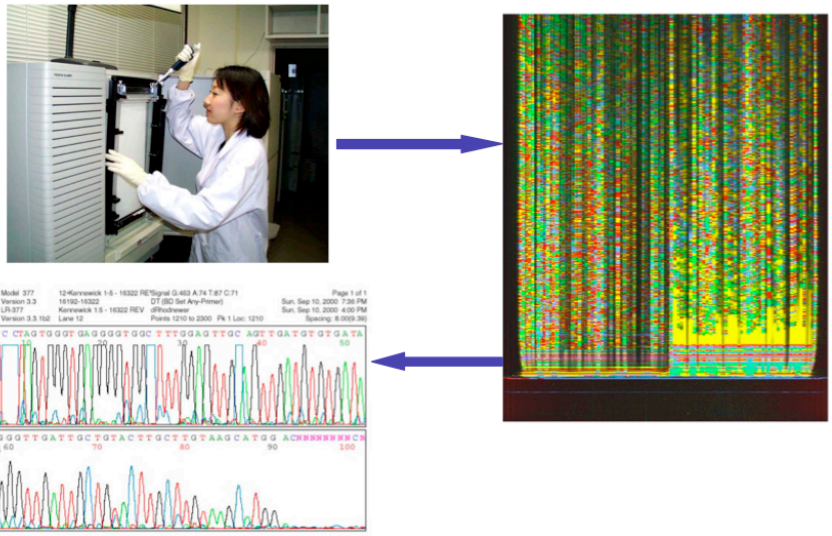
Linker are required for directed cloning of cDNAs



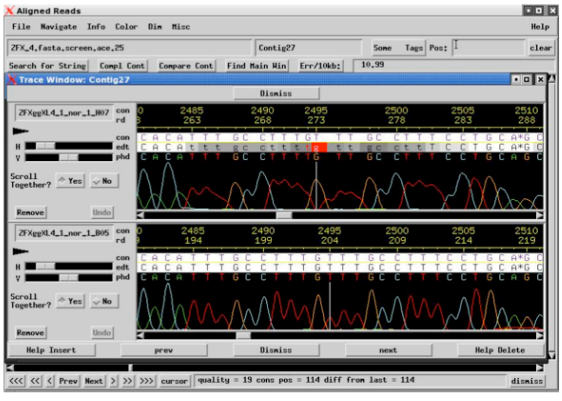
Sanger Sequencing



Sanger Sequencing: The Trace Data



The Base calling procedure



Quality Parameters

- Peak Spacing (7)
- Uncalled/Called Ratio (7)
- Uncalled/Called Ratio (3)
- Peak Resolution

Calibration Dataset

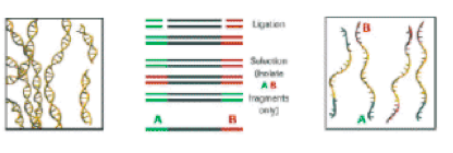
$$Q = -10 \log10(P_e)$$

Ewing B, Green P: Basecalling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8:186-194 (1998).

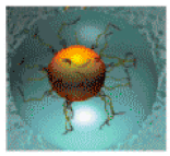
Sequencing Methods: Real time sequencing

CIBIV MFPL

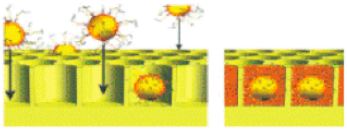
Figure 1. Overview of the 454 sequencing system



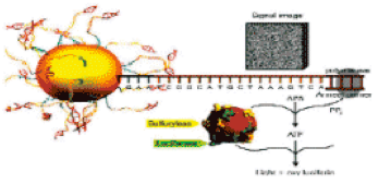
1) Prepare Adapter Ligated ssDNA Library (A-[insert]-B)



2) EmPCR: Clonal Amplification on 28 µ beads followed by enrichment



3) Load beads and enzymes in PicoTiter Plate™



4) Perform sequencing-by-synthesis on the GS 20 Sequencer

www.roche-applied-science.com 454 LIFE SCIENCES

Sequencing Methods: Real time sequencing

CIBIV MFPL

Step 1 - Polymerase

One of the four nucleotides dNTP (dATP, dCTP, dGTP, dTTP) is added to the reaction mixture. If the added nucleotide is complementary to the base in the DNA strand, it is incorporated and inorganic pyrophosphate (PP_i) is released.

Step 2 - ATP sulfurylase

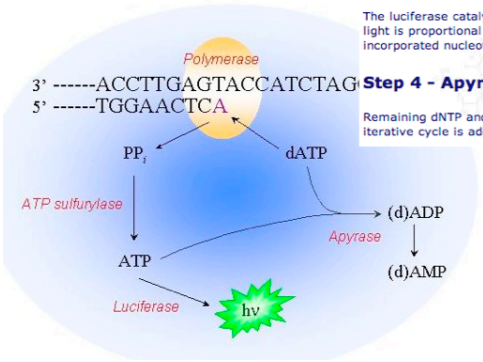
The PP_i is converted into ATP by the enzyme ATP sulfurylase.

Step 3 - Luciferase

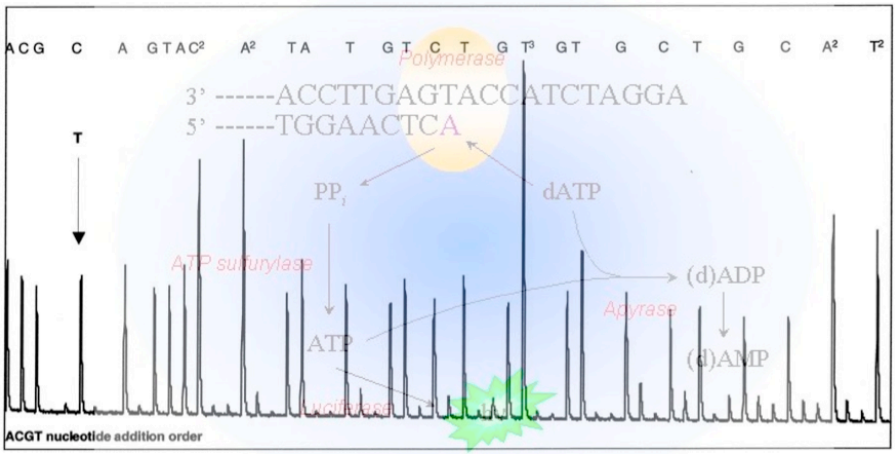
The luciferase catalyzes a reaction where ATP is used to generate light. The amount of light is proportional to the amount of ATP, and hence also proportional to the amount of incorporated nucleotides via the PP_i. The light is then detected by a CCD camera.

Step 4 - Apyrase

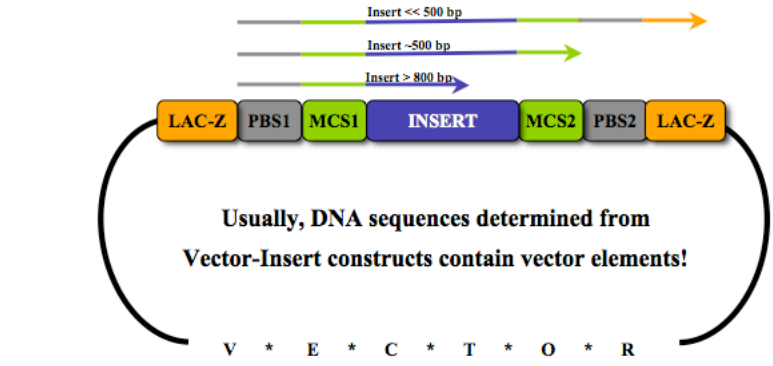
Remaining dNTP and ATP are degraded by the apyrase before the next nucleotide in the iterative cycle is added to the reaction mixture.



Light intensity reflects the number of inserted bases



Sequences from cloned inserts usually contain vector elements



LAC-Z — LAC-Z gene for blue-white selection

MCS1 — Multiple Cloning site

PBS1 — Primer Binding site

EST information helps sometimes

1: EC906561. Reports XBT_L-1_E18.t Xen...[gi:116673034]

IDENTIFIERS

dbEST Id:	**40460848**
EST name:	XBT_L-1_E18.t
GenBank Acc:	EC906561
GenBank gi:	116673034

CLONE INFO

DNA type:	cDNA

PRIMERS

PolyA Tail:	Unknown

SEQUENCE

GATTACTGGGACCCAGCGATAGACGTACAAAAACTGGTCCCATGGAAGACACTTCACGGA
GAACGATACACTAATAGTATAAAATATCAAAGAAAAATGTTTGGAAGATTTGGCAAATCT
TCGGGTGTAGATCCCTGTCTACTCTGGCCGTCGAGGAGCGAATTGAAGGAATTGATGTCC
CAGGCCTCTTCAGAGCGACCCGTGGACGAAGTGAAGGCTGAGATAGATGAACGGAAAATA
ATCGCGAATGAGGAGAAAGCAAAGAGGGAGAGAATTATAGCTGAGAATATGGCAAAGATG
CCAGCCCTGATAGCACGCTGGGAGAAGAAAGTCGACCAAGATAAAGCCAAGCAGTTGGAA
GCAAAGGAAACGCGTGAAAAAAATGCTGAAAGAAGCAAGAAGGACTTATGGACAGAGATT
GAGAGAGGACAACCCGCAATTCAAAGAACTTTTAGCCGAACTTCAAAAAGAAGAAGAGAA
AAAGGACAAGGAACTTAGAAAGAGAGGAATAATATTACCAGATAAATGATTTGTGATGAT
AATGTTGTGTAAATGAATGATACGTTAATAGAATATTACGAATAACACAAAAAAAAGTAC
CGACTGCG

Entry Created:	Oct 26 2006
Last Updated:	Oct 26 2006

LIBRARY

Lib Name:	Xenoturbella bocki, whole animal expression library (XBT_L)
Organism:	Xenoturbella bocki
Tissue type:	whole animal
Vector:	pGemT
Description:	Amplified cDNA library. Library split by size >1.5 KB (XBT_L) and size <1.5KB (XBT_S)

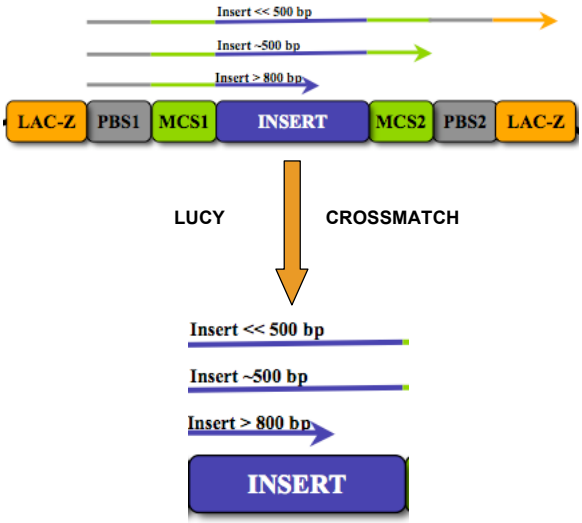
SUBMITTER

Name:	Leonid L. Moroz and Andrea B. Kohn
Lab:	Whitney Laboratory, Dept of Neuroscience
Institution:	University of Florida
Address:	9505 Ocean Shore Blvd, St. Augustine, FL 32080, USA
Tel:	(904) 461-4020
Fax:	(904) 461-4052
E-mail:	abk@whitney.ufl.edu, moroz@whitney.ufl.edu

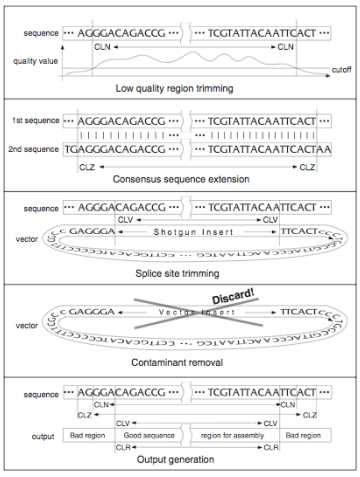
Trace Information also provides relevant information


```
<?xml version="1.0"?>
    <trace_volume>
        <trace>
            <trace_name>Xb_MM1_01A01</trace_name>
            <center_name>KML-UH</center_name>
            <submission_type>NEW</submission_type>
            <species_code>XENOTURBELLA BOCKI</species_code>
            <strategy>EST</strategy>
            <trace_type_code>EST</trace_type_code>
            <source_type>NON GENOMIC</source_type>
            <center_project>ANIMAL PHYLOGENOMICS</center_project>
            <trace_file>XENOTURBELLA_BOCKI/KML-UH/traces/Xb_MM1_01A01.scf</trace_file>
            <chemistry>BIGDYEV3.1</chemistry>
            <chemistry_type>TERMINATOR</chemistry_type>
            <clone_id>Xb_MM1_01A01</clone_id>
            <insert_flank_left>CGACTGGAGCACGAGGACACTGACATGGACTGAAGGAGTAGAAA</insert_flank_left>
            <insert_flank_right>AAAAAAAAAAAAAAACACTGTCATGCCGTTACGTAGCGTATCGTTGACAGC</insert_flank_right>
            <library_id>XB_MM1</library_id>
            <plate_id>01</plate_id>
            <program_id>KB BASECALLER VERSION=1.2</program_id>
            <run_machine_type>ABI3730XL</run_machine_type>
            <template_id>Xb_MM1_01A01</template_id>
            <trace_end>FORWARD</trace_end>
            <trace_format>SCF</trace_format>
            <ncbi_trace_archive>
                <ti>1926486127</ti>
                <taxid>242395</taxid>
                <basecall_length>1174</basecall_length>
                <load_date>Oct  2 2007  3:45AM</load_date>
                <state>active</state>
            </ncbi_trace_archive>
        </trace>
```

Clipping of non-Insert DNA is essential



LUCY: Identification of Vector contamination and low quality regions



1. Determine longest continuous high-qual region (overall quality must exceed a user defined value)

Fig. 2. Illustrations of LUCY's major processing steps. See the main text for explanations.

From: Chou H.-H. and Holmes M. H. (2001) Bioinformatics 17:1093-104

LUCY: Identification of Vector contamination and low quality regions

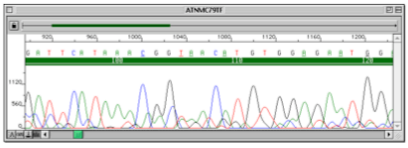


Fig. 3. The raw data in a chromatogram file can be viewed as four sets of overlapping peaks, one each for the A, C, G and T sequencing reactions.

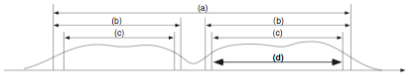


Fig. 4. LUCY's quality trimming steps. (a) Low quality areas are trimmed from each end, then (b) regions of poor quality within the sequence are identified and removed from the clean ranges. (c) The resulting candidate clean ranges are further trimmed to satisfy the overall average probability of error criterion and the criterion of the probability of error at terminal bases. (d) The largest remaining candidate is chosen as the final clean range.

1. Determine longest continuous high-qual region (overall quality must exceed a user defined value)

a) **End Trimming (bracket): Find the first and the last window meeting a given quality limit, e.g. 10 consecutive base pairs with max_avg_error of 0.02**

From: Chou H.-H. and Holmes M. H. (2001) Bioinformatics 17:1093-104

LUCY: Identification of Vector contamination and low quality regions

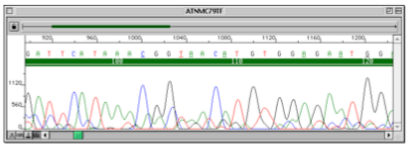


Fig. 3. The raw data in a chromatogram file can be viewed as four sets of overlapping peaks, one each for the A, C, G and T sequencing reactions.

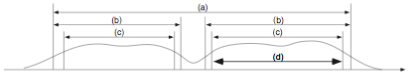


Fig. 4. LUCY's quality trimming steps. (a) Low quality areas are trimmed from each end, then (b) regions of poor quality within the sequence are identified and removed from the clean ranges. (c) The resulting candidate clean ranges are further trimmed to satisfy the overall average probability of error criterion and the criterion of the probability of error at terminal bases. (d) The largest remaining candidate is chosen as the final clean range.

1. Determine longest continuous high-qual region (overall quality must exceed a user defined value)

a) End Trimming (bracket): Find the first and the last window meeting a given quality limit, e.g. 10 consecutive base pairs with max_avg_error of 0.02

b) **Identification of regions with high error probability (window). Find the longest subsequence that meets the quality criteria specified by window_size and max_avg_error (multiple window sizes are allowed)**

From: Chou H.-H. and Holmes M. H. (2001) Bioinformatics 17:1093-104

LUCY: Identification of Vector contamination and low quality regions

CIBIV MFPL

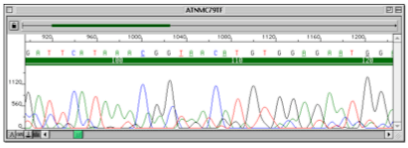


Fig. 3. The raw data in a chromatogram file can be viewed as four sets of overlapping peaks, one each for the A, C, G and T sequencing reactions.

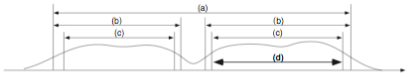


Fig. 4. LUCY's quality trimming steps. (a) Low quality areas are trimmed from each end, then (b) regions of poor quality within the sequence are identified and removed from the clean ranges. (c) The resulting candidate clean ranges are further trimmed to satisfy the overall average probability of error criterion and the criterion of the probability of error at terminal bases. (d) The largest remaining candidate is chosen as the final clean range.

1. Determine longest continuous high-qual region (overall quality must exceed a user defined value)

a) End Trimming (bracket): Find the first and the last window meeting a given quality limit, e.g. 10 consecutive base pairs with max_avg_error of 0.02

b) Identification of regions with high error probability (window). Find the longest subsequence that meets the quality criteria specified by window_size and max_avg_error (multiple window sizes are allowed)

c) Identify the longest subsequence fulfilling the overall sequence quality criteria (error). Options to specify are max_avg_error and max_error_at_ends (maximum error probability to the two bases on each end of the subsequence).

From: Chou H.-H. and Holmes M. H. (2001) Bioinformatics 17:1093-104

LUCY: Identification of Vector contamination and low quality regions

CIBIV MFPL

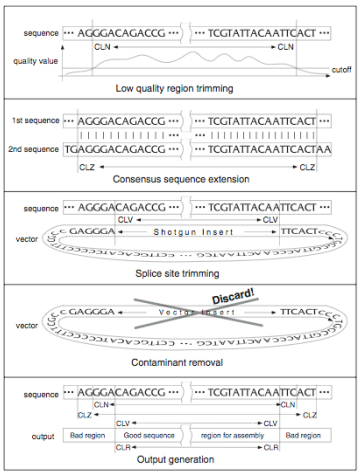


Fig. 2. Illustrations of LUCY's major processing steps. See the main text for explanations.

1. Determine longest continuous high-qual region (overall quality must exceed a user defined value)

2. (Optional): Extend the high quality region by using a 2nd, independent base call from the same chromatogram.

From: Chou H.-H. and Holmes M. H. (2001) Bioinformatics 17:1093-104

LUCY: Identification of Vector contamination and low quality regions

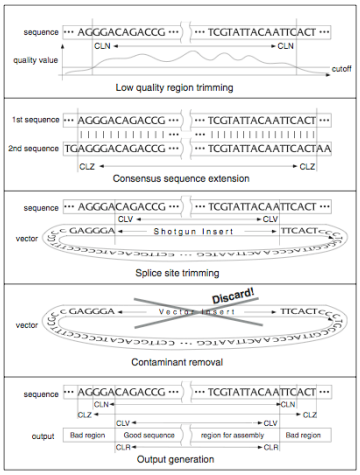
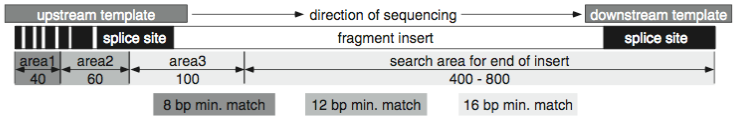


Fig. 2. Illustrations of LUCY's major processing steps. See the main text for explanations.

1. **Determine longest continuous high-qual region (overall quality must exceed a user defined value)**
2. **(Optional): Extend the high quality region by using a 2nd, independent base call from the same chromatogram.**
3. **Identification of 'Vector Splice Sites'. Different search stringencies are used to account of variable base quality over the sequence**

From: Chou H.-H. and Holmes M. H. (2001) Bioinformatics 17:1093-104

LUCY: Identification of 'Vector Splice Sites'.

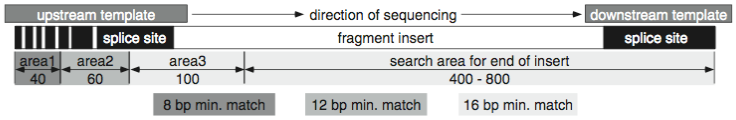


a) **splice_site_file: Used for vector clipping and contains the flanking sequences of the cloning site, plus any adaptor sequences used during cloning. Option: bidirectional trimming.**
b) **vector_file: Used for contaminant identification. Contains the entire sequence of the cloning vector.**

1) **Search within the first 200 bp for the 5' splice site. In the first 40 bp, an 8 bp long optimal local alignment of splice site and sequence is sufficient. For the next 60 bp and 100 bp, the minimum alignment length is set to 12 and 16, respectively.**

From: Chou H.-H. and Holmes M. H. (2001) Bioinformatics 17:1093-104

LUCY: Identification of 'Vector Splice Sites'.



a) **splice_site_file: Used for vector clipping and contains the flanking sequences of the cloning site, plus any adaptor sequences used during cloning. Option: bidirectional trimming.**
b) **vector_file: Used for contaminant identification. Contains the entire sequence of the cloning vector.**

1) Search within the first 200 bp for the 5' splice site. In the first 40 bp, an 8 bp long optimal local alignment of splice site and sequence is sufficient. For the next 60 bp and 100 bp, the minimum alignment length is set to 12 and 16, respectively.
2) **Search in the remainder of the sequence for the 3' splice site. Note, search is done only with the highest stringency.**

From: Chou H.-H. and Holmes M. H. (2001) Bioinformatics 17:1093-104

LUCY: Identification of polyA tails

1) **Search for the first *min_span* oligo-dT in the *initial_search_range* bp of the insert sequence,**
2) **Extend from this poly-T seed towards the center of the sequence**
3) **Allow for *max_error* mismatches between the oligo-dT of *min_span* length and the sequence.**
4) **Repeat the procedure with oligo-dA from the end of the sequence.**

From: Chou H.-H. and Holmes M. H. (2001) Bioinformatics 17:1093-104

LUCY: Identification of Vector contamination and low quality regions

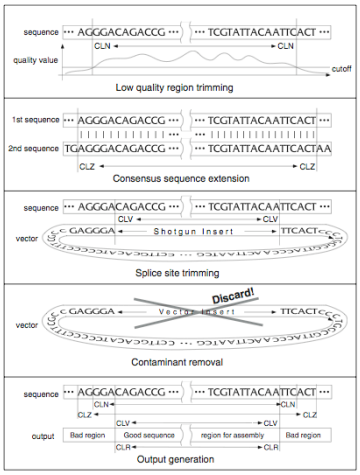


Fig. 2. Illustrations of LUCY's major processing steps. See the main text for explanations.

1. **Determine longest continuous high-qual region (overall quality must exceed a user defined value)**
2. **(Optional): Extend the high quality region by using a 2nd, independent base call from the same chromatogram.**
3. **Identification of 'Vector Splice Sites'. Different search stringencies are used to account of variable base quality over the sequence**
4. **Identification and removal of contaminating sequences, e.g. Vector inserts (cut the vector into a tag-library and search for hits.**

From: Chou H.-H. and Holmes M. H. (2001) Bioinformatics 17:1093-104

LUCY: Identification of Vector contamination and low quality regions

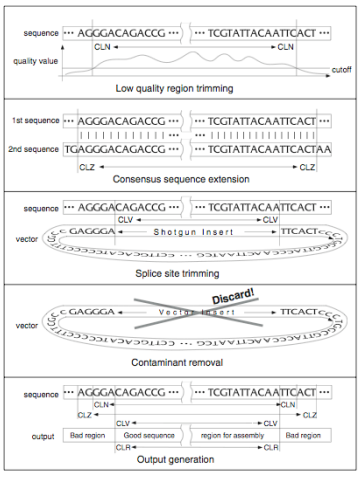


Fig. 2. Illustrations of LUCY's major processing steps. See the main text for explanations.

1. **Determine longest continuous high-qual region (overall quality must exceed a user defined value)**
2. **(Optional): Extend the high quality region by using a 2nd, independent base call from the same chromatogram.**
3. **Identification of 'Vector Splice Sites'. Different search stringencies are used to account of variable base quality over the sequence**
4. **Identification and removal of contaminating sequences, e.g. Vector inserts (cut the vector into a tag-library and search for hits.**

5. **Summary of the results and output generation**

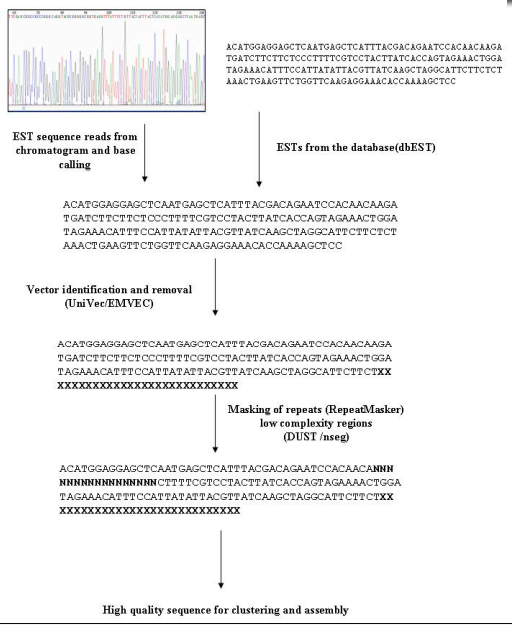
From: Chou H.-H. and Holmes M. H. (2001) Bioinformatics 17:1093-104

LUCY: Parameter and their default values

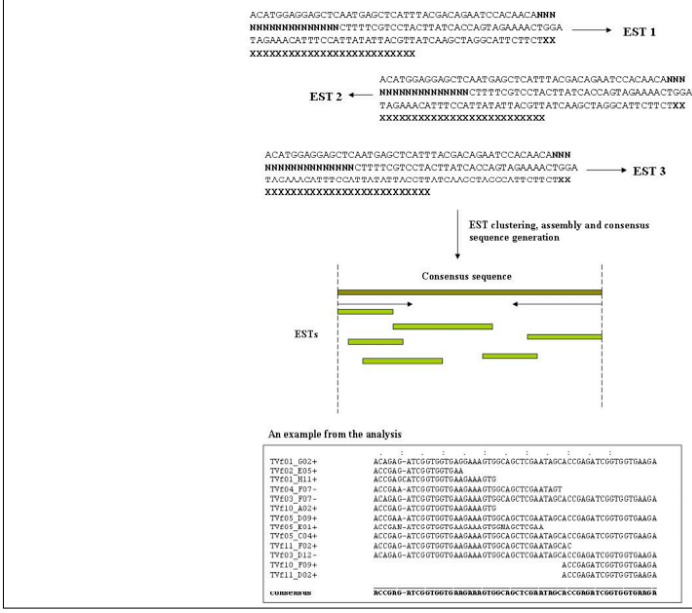
Parameters	Default values	Related operation steps
pass_along min_value max_value med_value	0 0 0	pass to assembly program
error max_avg_error max_error_at_ends	0.025 0.02	quality area determination
window window_size max_avg_error ...	50 0.08 10 0.3	quality area determination
bracket window_size max_avg_error	10 0.02	quality area determination
range area1 area2 area3	40 60 100	vector splice site trimming
alignment area1 area2 area3	8 12 16	vector splice site trimming
vector vector_sequence_file splice_site_file	none	vector splice site trimming
cdna [min_span max_error initial_search_range]	none or 10 3 50	poly-A/T trimming
keep	none	poly-A/T trimming
size vector_tag_size	10	contaminant removal
threshold vector_cutoff	20	contaminant removal
minimum good_sequence_length	100	overall quality control
xtra cpu_threads	1	overall program control
output, quiet, inform_me, debug	none	overall program control

From: Chou H.-H. and Holmes M. H. (2001) Bioinformatics 17:1093-104

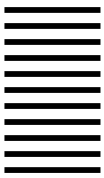
Additional processing: Id of repeats



EST-Clustering

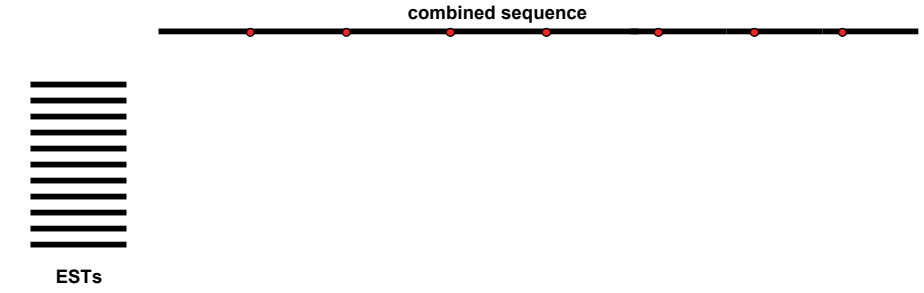


CAP3: Overlap identification



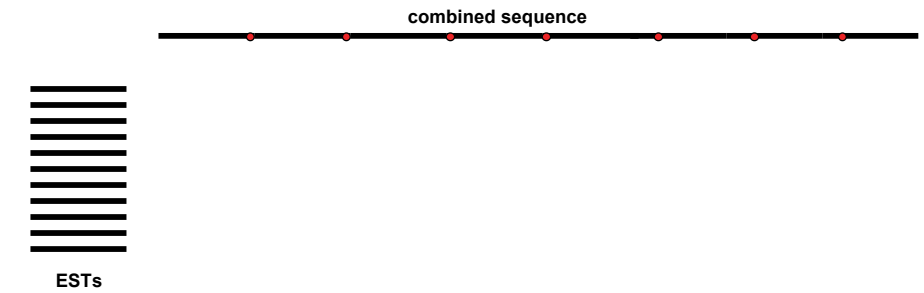
ESTs

CAP3: Overlap identification



1. Concatenate sequences into a combined sequence. Reads are separated by a separation character

CAP3: Overlap identification



1. Concatenate sequences into a combined sequence. Reads are separated by a separation character.
2. Compute high scoring chains of segments between each EST and the combined sequence using the Blast heuristic. Identify candidate pairs. Every pair counted only once.

CAP3: Overlap identification

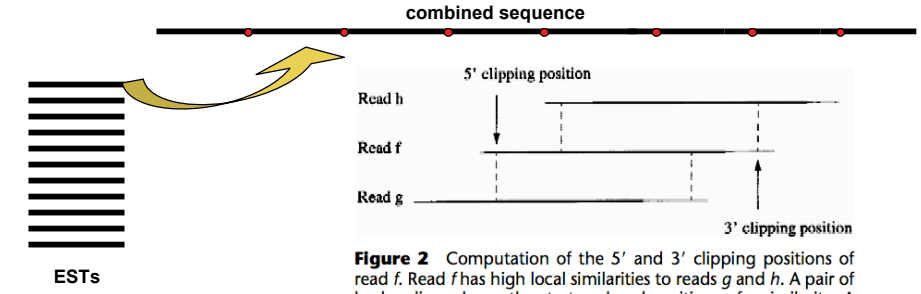
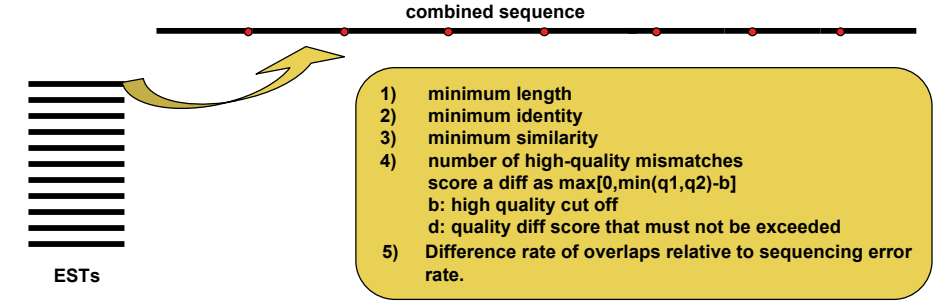


Figure 2 Computation of the 5′ and 3′ clipping positions of read *f*. Read *f* has high local similarities to reads *g* and *h*. A pair of broken lines shows the start and end positions of a similarity. A thick line indicates the high-quality region of a read.

1. **Concatenate sequences into a combined sequence. Reads are separated by a separation character**
2. **Compute high scoring chains of segments between each EST and the combined sequence using the Blast heuristic. Identify candidate pairs. Every pair counted only once.**
3. **Remove poor quality sequence ends**

What is good?
a) any sufficiently long region of high quality that is highly similar to a region of another read
b) any sufficiently long region that is similar to a high quality region of another read.
Method: Smith-Waterman alignment with quality-weighted scored: Match = m * min(q1,q2); Mismatch = n * min(q1,q2), Gap-extension = -g * min(q1,q2).

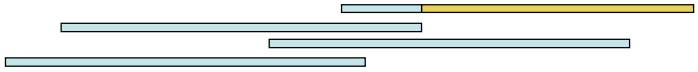
CAP3: Overlap identification



1. **Concatenate sequences into a combined sequence. Reads are separated by a separation character**
2. **Compute high scoring chains of segments between each EST and the combined sequence using the Blast heuristic.**
3. **Remove poor quality sequence ends**
4. **Compute global alignment for the high quality sequence pairs to verify overlaps. Evaluate according to the following criteria:**

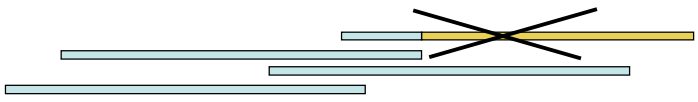
CAP3: Contig Building

1) Generate a general layout using the overlapping reads from the pair-wise analysis (Greedy algorithm in decreasing order of overlap scores).
2) In a simple view: Check the layout for incompatibilities, remove incompatible reads and align.



CAP3: Contig Building

1) Generate a general layout using the overlapping reads from the pair-wise analysis (Greedy algorithm in decreasing order of overlap scores).
2) In a simple view: Check the layout for incompatibilities, remove incompatible reads and align.



CAP3: Contig Building

1) Generate a general layout using the overlapping reads from the pair-wise analysis (Greedy algorithm in decreasing order of overlap scores).
2) In a simple view: Check the layout for incompatibilities, remove incompatible reads and align. (For the whole story refer to Huang and Madan (1999) Genome Res. 9:868-877)