CNR - National Research Council
ITB - Institute of Biomedical Technologies

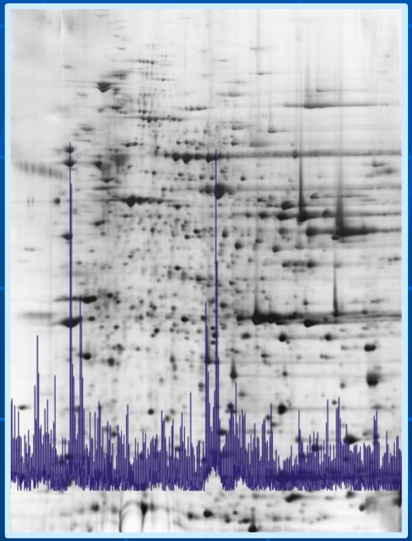
Workflows management: new abilities for the biological information overflow

Luciano Milanesi
Institute Biomedical Technologies CNR
luciano.milanesi@itb.cnr.it

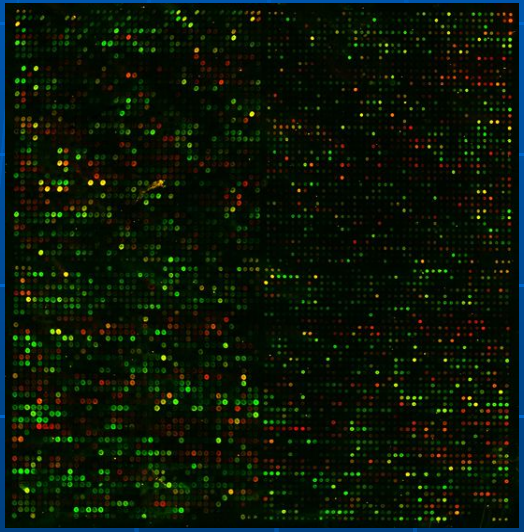
Workflows management: new abilities for the biological information overflow

Luciano Milanesi
Institute Biomedical Technologies CNR
luciano.milanesi@itb.cnr.it

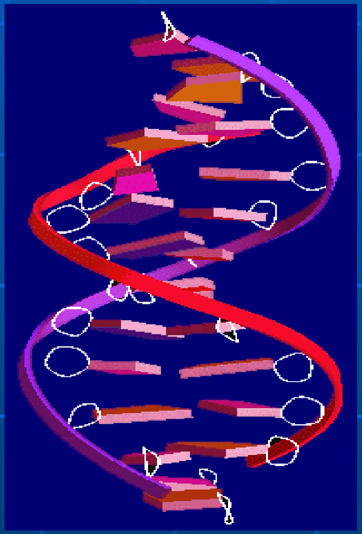
Complex Disease Mapping



Proteins
(Proteome)



Microarray
(Genome)



Gene & SNPs
(Genome)

HTS Data Project

Bioinformatics: Emerging Opportunities and Emerging Gaps

Paula E.Stephan and Grant Black

- A typical gene lab can produce **100 terabytes of information a year**, the equivalent of **1 million encyclopedias**.
- Few biologists have the computational skills needed to fully explore such an astonishing amount of data; nor do they have the skills to explore the exploding amount of data being generated from clinical trials.
- The immense amount of data that are available, and the knowledge that this is but the tip of the data iceberg.

HTS Data Project

EST

Microsatellite

MSMS

DNA High Throughput Sequencing

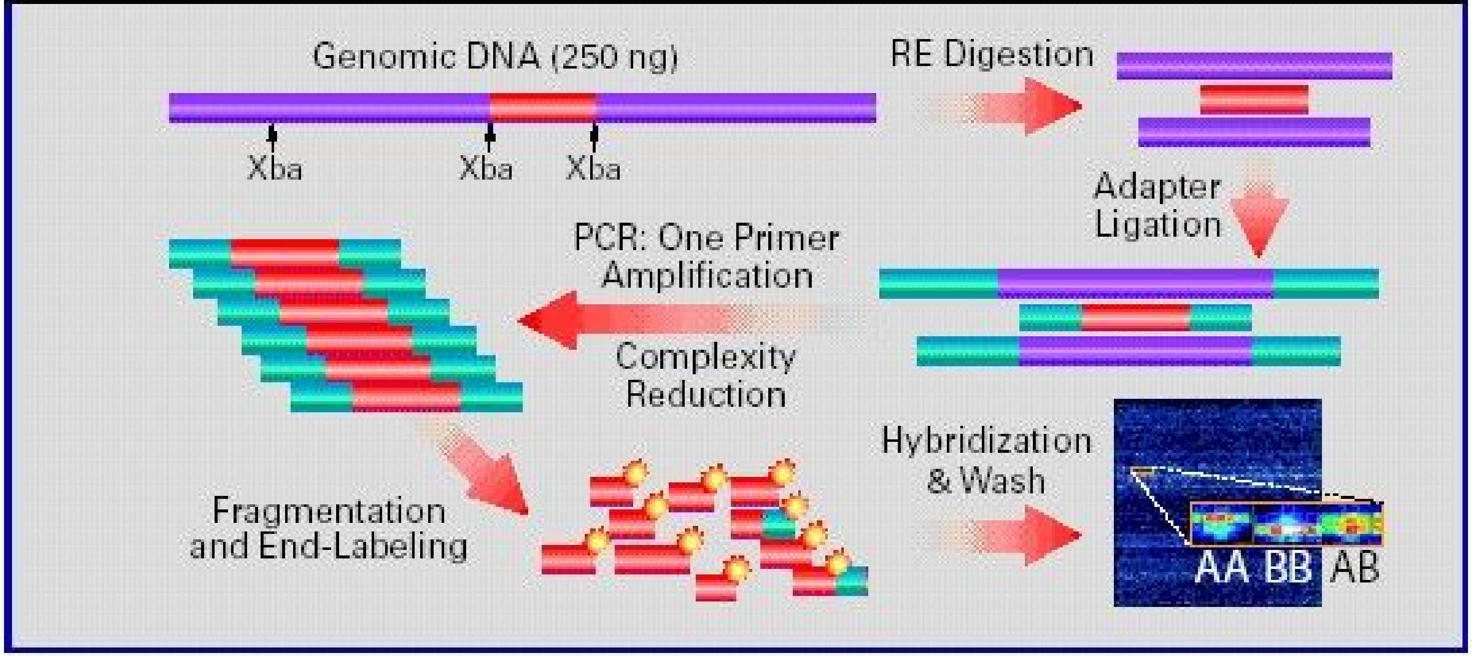
SNP's

Microarray

HTS Data Project

DNA High Throughput Sequencing

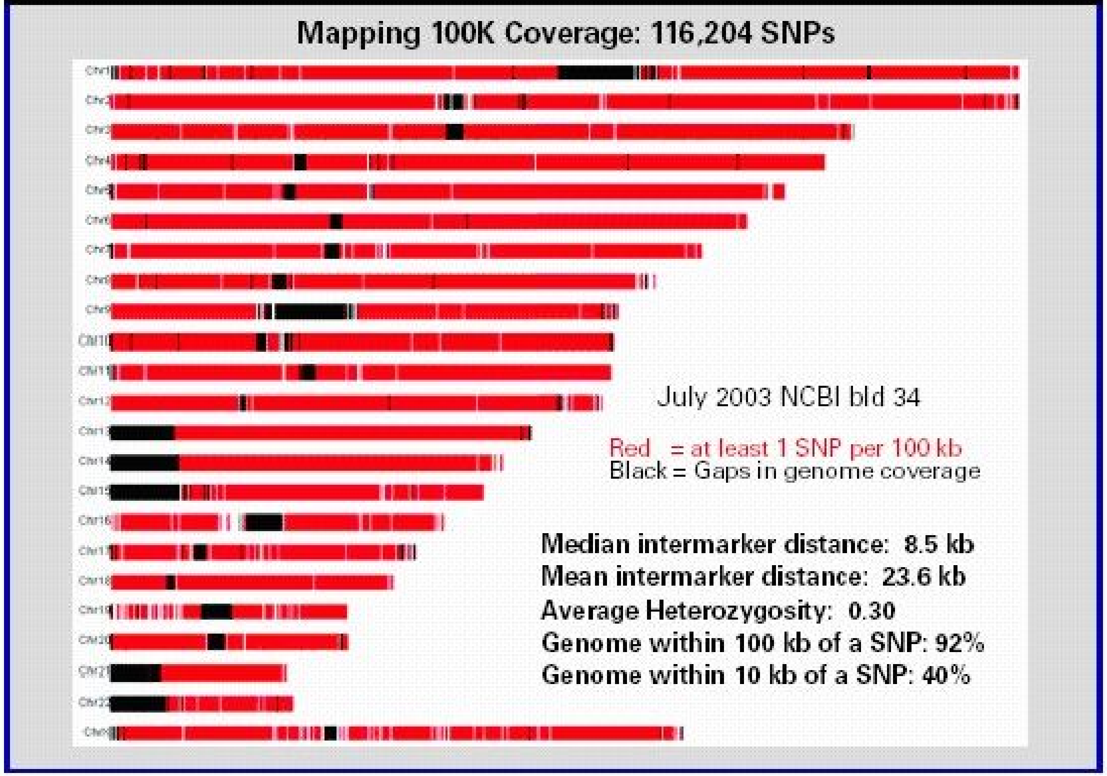
Figure 1: GeneChip® Mapping Assay Overview.



HTS Data Project

SNP's

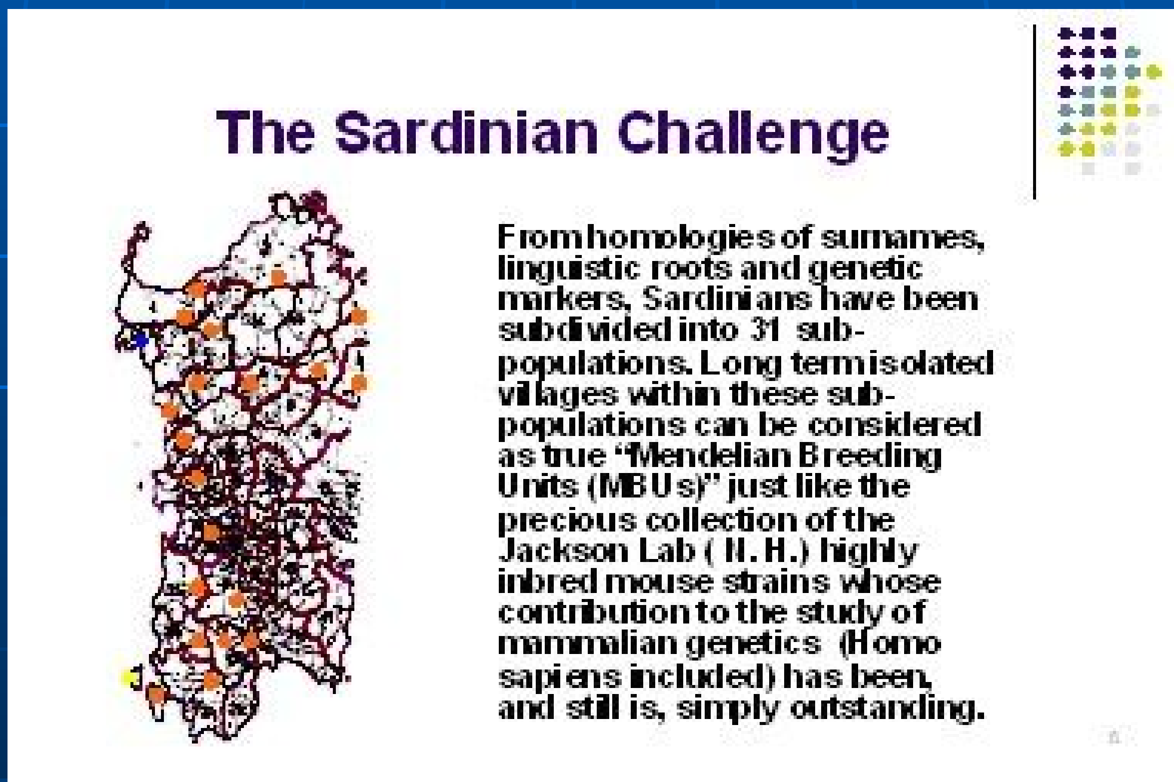
Figure 2: Genome Coverage of Mapping 100K SNPs by chromosome. Black areas represent gaps in the human genome sequence, primarily centromeres and telomeres.



HTS Data Project

SNP's

HTS

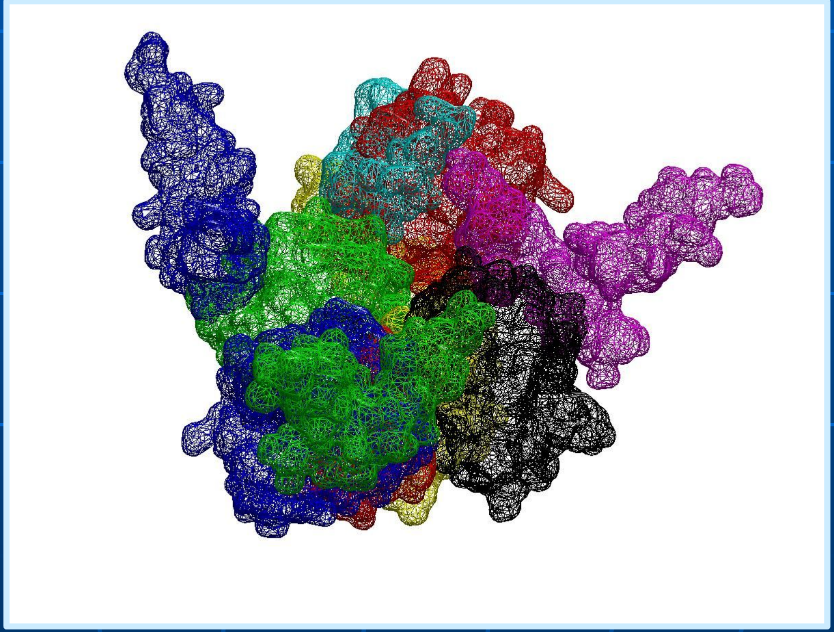


ANDTIASQDT PAKVVIKANK LKDLKDYVDD LKTYNNTYSN VVTVAGEDRI ETAIELSSKY YNSDDKNAIT
DKAVNDIVLV GSTSIVDGLV ASPLASEKTA PLLLTSKDKL DSSVKSEIKR VMNLKSDTGI NTSKKVYLAG GVNSISKDVE
NELKNMGLKV TRLSGEDRYE TSLAIADEIG LDNDKAFVVG GTGLADAMSI APVASQLKDG DATPIVVVDG
KAKEISDDAK SFLGTSDVDI IGGKNSVSKE IEESIDSATG KTPDRISGDD RQATNAEVLK EDDYFTDGEV
VNYFVAKDGS TKEDQLVDAL AAAPIAGRFK ESPAPIILAT DTLSSDQNVAVSKAVPKDGG TNLVQVGKGI
ASSVINKMKD LLDM

>monoisotopic mass = **39480**

position	sequence (**NCBI BLAST** link)
18- 22	ANKLK
245- 249	NSVSK
262- 265	TPDR
116- 119	SEIK
344- 347	AVPK
298- 302	DGSTK
266- 271	ISGDDR
163- 168	LSGEDR
116- 120	SEIKR
60- 66	YYNSDDK
224- 230	EISDDAK
155- 159	NMGLK
67- 72	NAITDK
126- 134	SDTGINTSK
160- 168	VTRLSGEDR
21- 25	LKDLK
108- 115	DKLDSSVK
135- 147	KVYLAGGVNSISK
14- 17	VVIK
1- 13	ANDTIASQDTPAK
126- 147	SDTGINTSKKVYLAGGVNSISK
14- 20	VVIKANK

……………………………………



Disease Gene

- Mutations in target sequences are usually revealed by either phenotypic selection in experimental test systems or, in case of disease-causing genes in humans, by clinical studies in which certain genes are sequenced in groups of patients and in control groups.
- Both the experimental test systems and the clinical studies rely on detectable (mutable) positions, which are sites where DNA sequence changes cause phenotypic changes.

Human genetic identity

- Genomic sequence **99.9%** identical
- **3,200,000** nucleotides different
- Single base differences in genomes between any two individuals: ca. **3 million**
- Amino acid differences in proteomes between any two individuals: ca. **100,000**

Variation types

- Macro:
  - Chromosome numbers
  - Segmental duplications, rearrangements, and deletions
- Medium:
  - Sequence Repeats
  - Transposable Elements
  - Short Deletions, Sequence and Tandem Repeats (including microsatellites)
- Micro:
  - **S**ingle **N**ucleotide **P**olymorphisms (SNPs)
  - Single Nucleotide Insertions and Deletions (Indels)

What are Single Nucleotide Polymorphisms (SNPs)?

```
ATGGTAAGCCTGAGCTGACTTAGCGT-AT
ATGGTAAACCTGAGTTGACTTAGCGTCAT
       ↑      ↑           ↑
      snp    snp        indel
```

SNPs result from replication errors and DNA damage

Are all SNPs really SNPs?

- A SNP is found by aligning overlapping DNA sequences and identifying variable positions
- Two types of errors in SNP finding
  - Inclusion of paralogs in sequence alignment
  - Sequencing errors

GCATGCAAGCAGATA
GCATGCACGCAGATA
GCATGCAAGCAGATA
GCATGCAAGCAGATA
GCATGCAAGCAGATA
GCATGCAAGCAGATA

Application of SNP data

- Study of evolution
  - Traces evolutionary history of different populations
- DNA fingerprinting
  - criminal or parental verification
- Markers for mapping of polygenic traits
- Genotype-specific medication
- Most genes contain SNPs
  - 93% genes have one or more SNPs
  - 39% have more than 10 SNPs)

Types of SNPs

Genic, coding SNPs

- non-synonymous
  - Maintaining vs. altering protein structure/function
- synonymous
  - Maintaining vs. altering splicing

Genic, non-coding SNPs

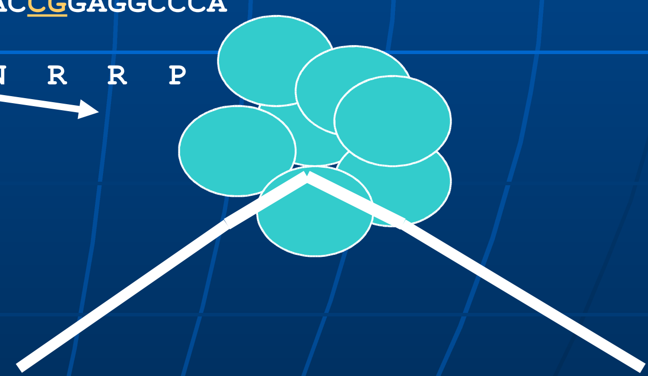
- Regulatory SNPs
  - Maintaining vs. altering gene expression
- Intronic SNPs
  - Maintaining vs. altering gene expression/splicing

Linked SNPs

- usually intergenic

Context-dependence of mutations

```
                                                         GA
                                                G        GA
                                                C        TA
                                                A        TA
                                                A        TA
                                                A        TA
                                        C       AA       TA
            T   G  T                AG  A       AA G     TA
   T
...ACCACCATCCACTACAACTACATGTGTAACAGTTCCTGCATGGGCGGCATGAACCGGAGGCCCA
   TC.
    T  T  I  H  Y  N  Y  M  C  N  S  S  C  M  G  G  M  N  R  R  P
   I
```



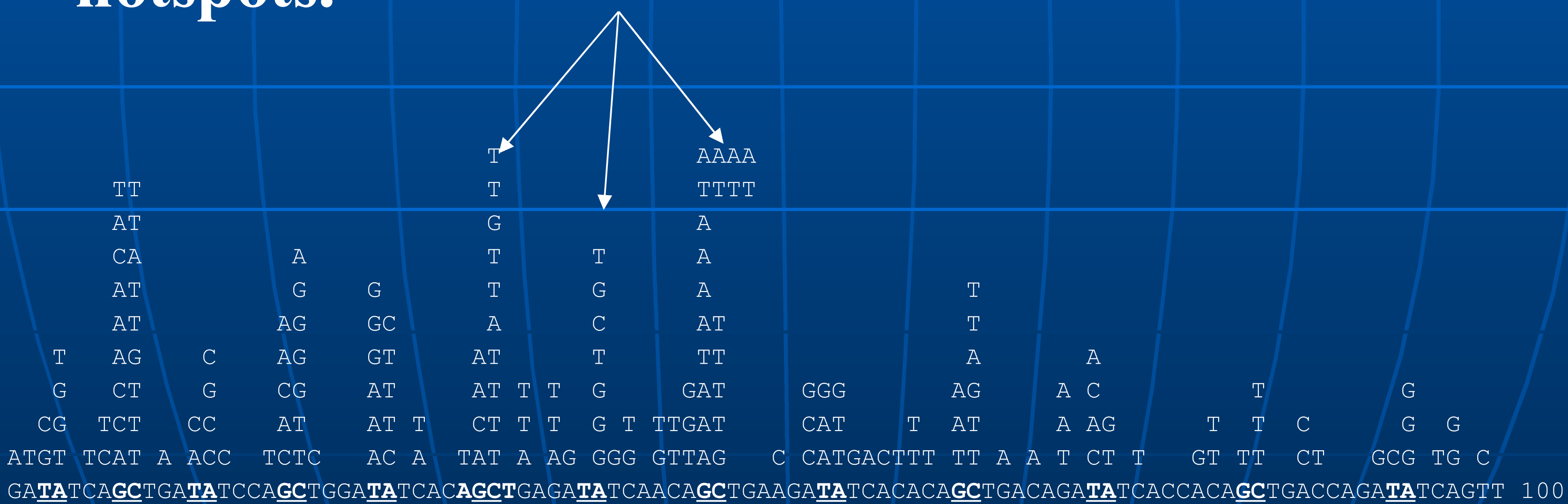
Mutational spectra

- **A mutation spectrum** is a distribution of mutation frequencies along nucleotide sequences and is compiled by the analysis of a large number of mutant target sequences.

```
                                T             AAAA
       TT                       T             TTTT
       AT                       G             A
       CA          A            T      T      A
       AT          G    G       T      G      A                 T
       AT         AG    GC      A      C      AT                T
   T   AG    C    AG    GT     AT      T      TT                A       A
   G   CT    G    CG    AT     AT T T  G     GAT     GGG       AG     A C          T         G
  CG  TCT   CC    AT    AT T   CT T T  G T TTGAT     CAT    T  AT     A AG      T  T  C      G  G
ATGT TCAT A ACC  TCTC   AC A  TAT A AG GGG GTTAG   C CATGACTTT TT A A T CT T   GT TT  CT   GCG TG C
GATATCAGCTGATATCCAGCTGGATATCACAGCTGAGATATCAACAGCTGAAGATATCACACAGCTGACAGATATCACCACAGCTGACCAGATATCAGTT 100
```

Mutational spectra

- Mutation frequencies vary significantly along nucleotide sequences such that mutations often concentrate at certain positions called **mutation hotspots.**



Hotspots in immunoglobulin V genes (Betz et al. 1993)

Mutational spectra

Mutation hotspots are frequently caused by mutable motifs .

R<u>G</u>YW/WR<u>C</u>Y (R = A/G, Y = T/C, W = A/T) **G/C mutator**

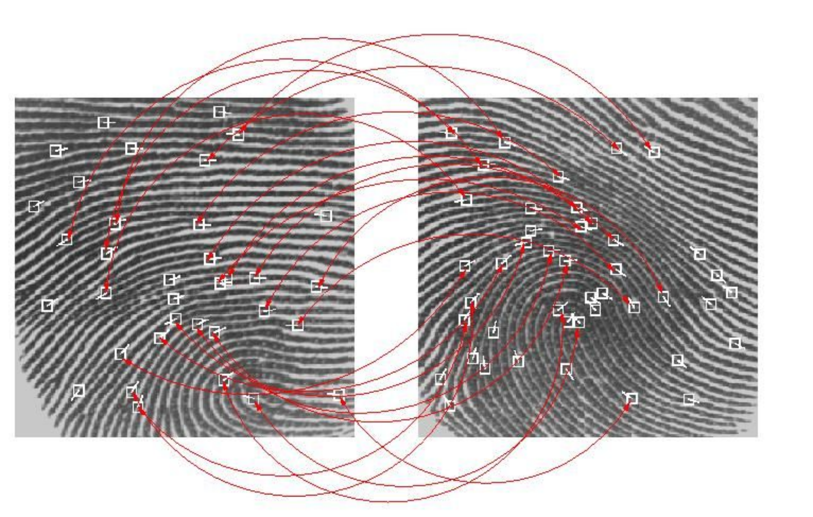
W<u>A</u>/<u>T</u>W (hotspot positions are underlined) **A/T mutator**

```
                                T             AAAA
       TT                       T             TTTT
       AT                       G             A
       CA          A            T      T      A
       AT          G    G       T      G      A                 T
       AT         AG    GC      A      C      AT                T
   T   AG    C    AG    GT     AT      T      TT                A       A
   G   CT    G    CG    AT     AT T T  G     GAT     GGG       AG      A C          T         G
  CG  TCT   CC    AT    AT T   CT T T  G T TTGAT     CAT    T  AT     A AG      T  T  C      G  G
ATGT TCAT A ACC  TCTC   AC A  TAT A AG GGG GTTAG   C CATGACTTT TT A A T CT T   GT TT  CT   GCG TG C
GATATCAGCTGATATCCAGCTGGATATCACAGCTGAGATATCAACAGCTGAAGATATCACACAGCTGACAGATATCACCACAGCTGACCAGATATCAGTT 100
  WA  RGYW  WA
  TW  WRCY  TW
```

Hotspot positions in **R<u>G</u>YW/WR<u>C</u>Y** and **W<u>A</u>/<u>T</u>W** mutable motifs

Mutational spectra

Mutation hotspots in DNA reflect intrinsic properties of the mutation process, such as sequence specificity, that manifests itself at the level of interaction between mutagens, DNA, and the action of the repair and replication machineries.



The nucleotide sequence context of mutation hotspots is a fingerprint of interactions between DNA and repair/replication/modification enzymes, and the analysis of hotspot context provides evidence of such interactions.

Mutational spectra

Context factors may influence mutation rates:

- Homonucleotide runs and microsatellites
- Direct and inverted repeats
- Local mononucleotide composition
- DNA conformation
- Oligonucleotide content
- higher-level features of gene
- chromatin structure

Methods

Various classification and statistical approaches are used for analysis of the nucleotide sequence context of mutations hotspots (Rogozin, Babenko, Milanesi, Pavlov 2003 *Brief. Bionform*. 4, 210-227).

DNA polymerase η mutation hotspots in *lacZ*

Sequence	Hotspot position	Type of changes	Number of mutations
CAATT	3	A$\rightarrow$G,T,C	15,1,1
TTATC	14	A$\rightarrow$G,C,T	14,1,1
GAAAT	21	A$\rightarrow$G,T	16,2
ATAGC	38	A$\rightarrow$G,T,C	9,2,1
CATAG	39	T$\rightarrow$G,A,C	9,9,2
TCATG	46	A$\rightarrow$G,T	13,1
GTAAT	50	A$\rightarrow$G,T	16,4
GAATT	56	A$\rightarrow$G	17
AAACG	70	A$\rightarrow$G,T	18,3
GTAAA	73	A$\rightarrow$G,T	14,1
CGACG	80	A$\rightarrow$G,T	11,2

WA	Consensus		

Hotspot positions are underlined.

Methods

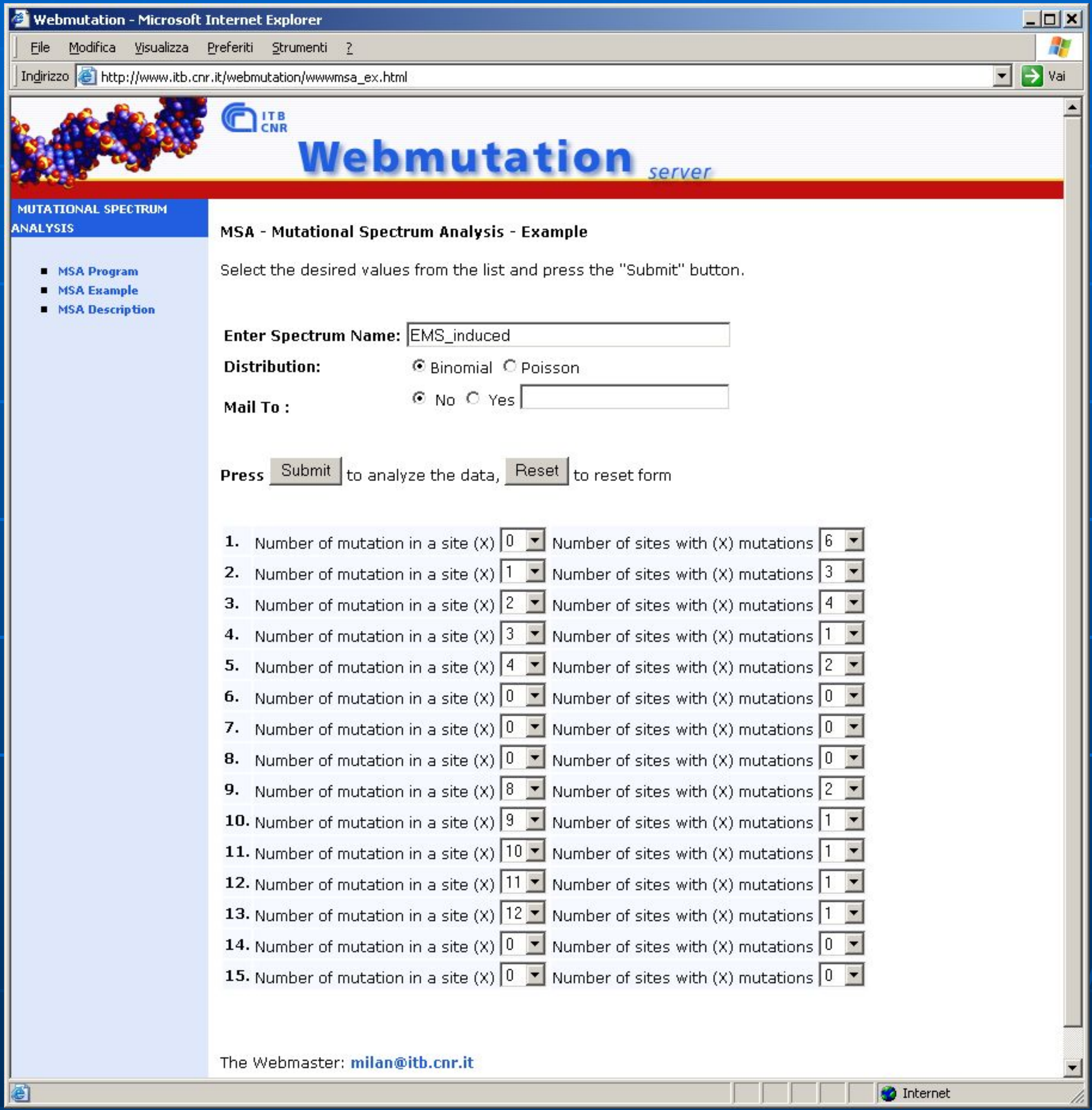
Correlation between nucleotide sequence features and mutation spectra. Nucleotide sequence features (mutable motifs) can be correlated with a mutation spectrum and the correlation can be tested for statistical significance.

```
        AT
   T    AG     C
   G    CT     G
  CG   TCT    CC
ATGT TCAT A ACC
GATATCAGCTGATATCCA
  WA   RGYW   WA
  TW   WRCY   TW
```

Motifs RGYW/WRCY+WA/TW

	+	-
Mutations	22	7
Positions	6	12

Fisher exact test P = 0.006 (non-random targeting of mutations to mutable motifs)


Webmutation - Microsoft Internet Explorer

File Modifica Visualizza Preferiti Strumenti ?

Indirizzo http://www.itb.cnr.it/webmutation/wwwmsa_ex.html Vai

ITB CNR

Webmutation server

MUTATIONAL SPECTRUM ANALYSIS

- MSA Program
- MSA Example
- MSA Description

MSA - Mutational Spectrum Analysis - Example

Select the desired values from the list and press the "Submit" button.

Enter Spectrum Name: EMS_induced

Distribution: (•) Binomial ( ) Poisson

Mail To : (•) No ( ) Yes

Press Submit to analyze the data, Reset to reset form

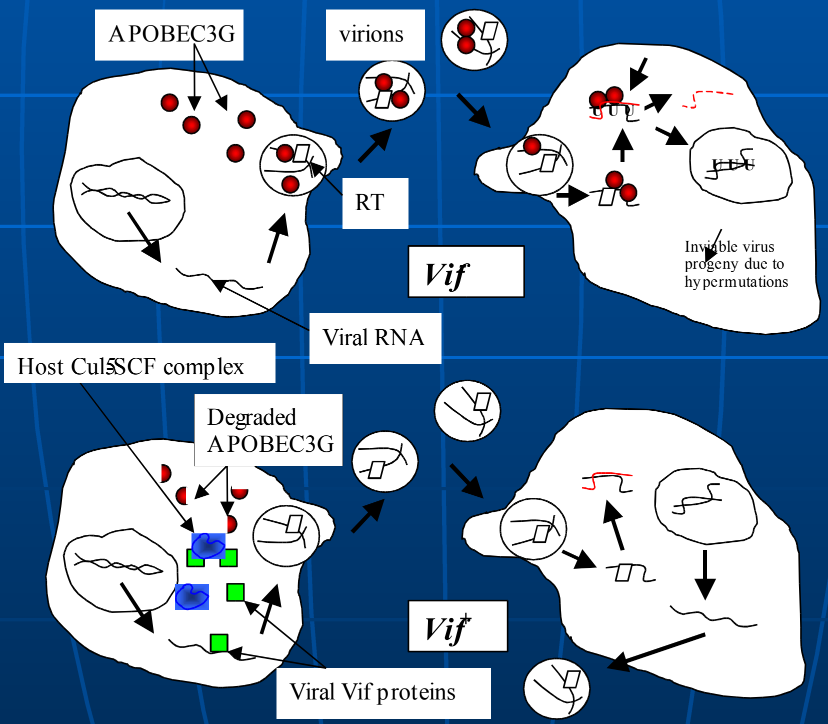
1.	Number of mutation in a site (X)	0	Number of sites with (X) mutations	6
2.	Number of mutation in a site (X)	1	Number of sites with (X) mutations	3
3.	Number of mutation in a site (X)	2	Number of sites with (X) mutations	4
4.	Number of mutation in a site (X)	3	Number of sites with (X) mutations	1
5.	Number of mutation in a site (X)	4	Number of sites with (X) mutations	2
6.	Number of mutation in a site (X)	0	Number of sites with (X) mutations	0
7.	Number of mutation in a site (X)	0	Number of sites with (X) mutations	0
8.	Number of mutation in a site (X)	0	Number of sites with (X) mutations	0
9.	Number of mutation in a site (X)	8	Number of sites with (X) mutations	2
10.	Number of mutation in a site (X)	9	Number of sites with (X) mutations	1
11.	Number of mutation in a site (X)	10	Number of sites with (X) mutations	1
12.	Number of mutation in a site (X)	11	Number of sites with (X) mutations	1
13.	Number of mutation in a site (X)	12	Number of sites with (X) mutations	1
14.	Number of mutation in a site (X)	0	Number of sites with (X) mutations	0
15.	Number of mutation in a site (X)	0	Number of sites with (X) mutations	0

The Webmaster: milan@itb.cnr.it

Internet

Hypermutation in HIV-1

The cytidine deaminase APOBEC3G confer resistance to HIV. Its antiviral action could be overcome by the presence of virion infectivity factor (Vif), encoded by the viral genome.



Hypermutation in HIV-1

Spectrum of mutations in the the *GFP* gene induced by the over-expression of APOBEC3G in *E.coli*

```
  AAA      AAAAAA              AAAAAAAAAAA          AAAAAAA
  A        AAAAAAA             A                    A                                             A
  A        A              AAAAAA   AAAAAAA          A                                             A
  A  A     A     A            AA   A              A A                                             A
  A  A     A     A            AAA  A           A  A A                                             A
  A  A     A A A A            AAAA A           A  A A       A           A              A        A A
ATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTGAACGGCCACAAGTTCAGCGTGTCCGGCGAGG 100
```

⇩

GG mutable motif

⇩

Spectrum of mutations in HIV-1 DNA in the absence of the Vif protein

```
                                                                             AAAAAA
                                           A                                  A
 A                            A            A                 A                AA
 A                            A            A  G     A  A G A AA  G         C  AA             G
TGGTAAATGTGACAGAAAATTTTAACATGTGGAAAAATGACATGGTAGAACAGATGCATGAGGATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGT 100
```

Hypermutation in HIV-1

Specificity of APOBEC3G for GG sequences, which is frequently a part of TGG tryptophan codons, results in a frequent generation of TAG nonsense codons which leads to a premature termination of protein synthesis.

The cytidine deaminase APOBEC3G may cause mutations in HIV-1

W

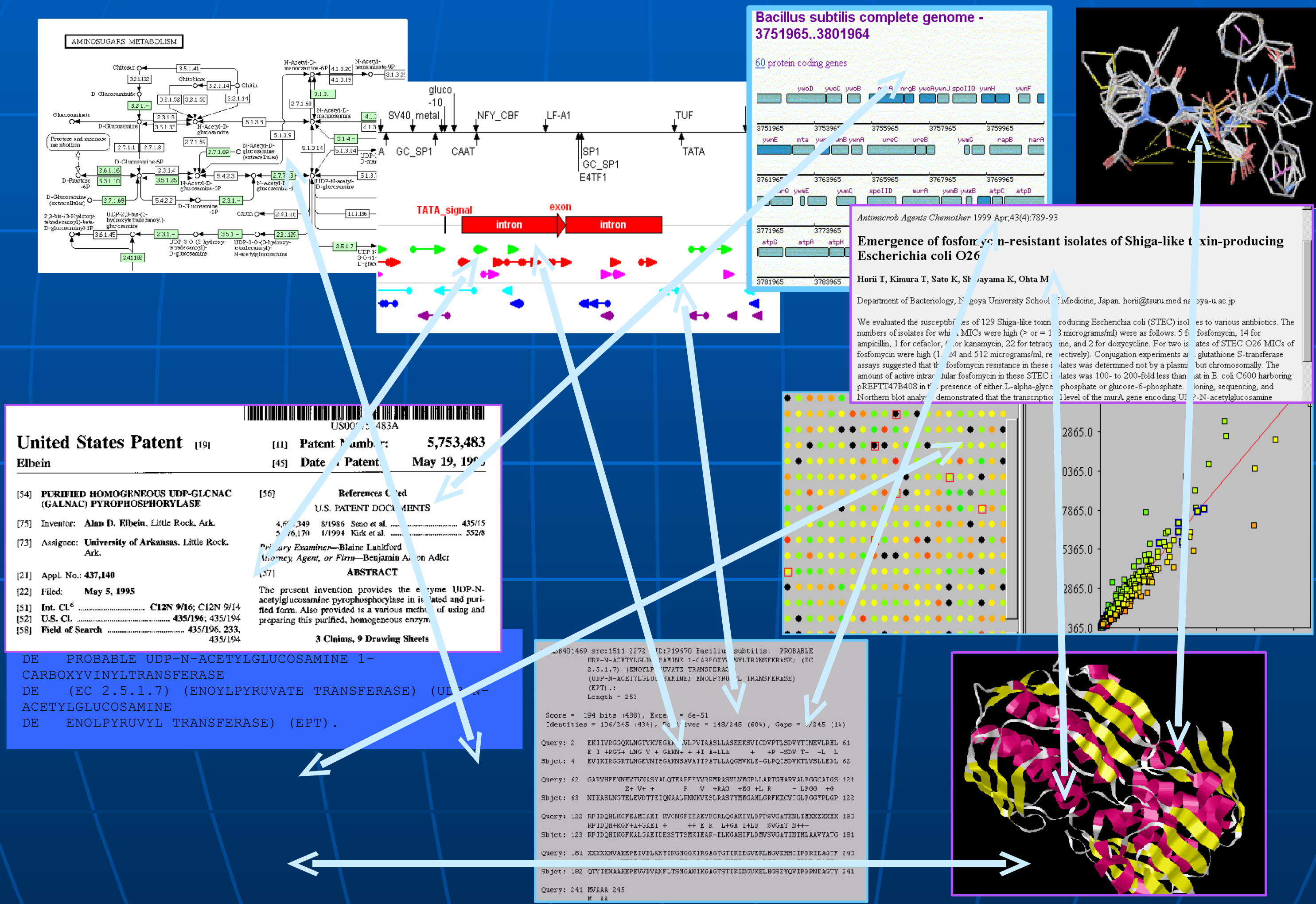
… TGG …

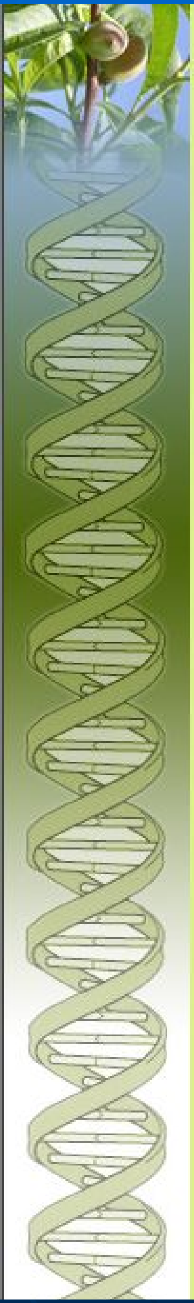
↓

*

… TAG …

HTS Data Project annotations





- Home Page
- Library details
- Processing, Assembly & Annotation Protocol
- Statistics on Sequences
- Sequence report
- Contig report
- SNPs
- Text Search
- Blast Search
- Links
- References
- Download
- DB contents and help

PROCESSING, ASSEMBLY AND ANNOTATION PROTOCOL

A fully automated pipeline has been prepared to process EST sequences using public software integrated by in-house developed Perl scripts.

Sequence files, produced by Dr. Pozzi's lab (libraries S3 and S4), together with the quality files produced by the base-calling program **phred** (Ewing et al, 1998) have been processed in order to identify vector contamination and low quality regions with the program **Lucy** (Chou et al, 2001) using default parameters.

Vector-free high quality sequences have been added to peach EST sequences produced by the groups of Padua (libraries Pp-S4 and S3II) and by the university of Clemson (library PP_Lea) and submitted to the program **CAP3** (Huan et al, 1999) to perform contig assembly. Stringency parameters have been modified (-p 95, -d 60) to identify paralogs.

All the input EST sequences and all the contig consensus sequences have been submitted to the **BLASTx** program for annotation (Altschul et al, 1990). BLASTx compares the six-frame conceptual translation products of a nucleotide query sequence against a protein sequence database. BLASTx is run locally against the Genbank nr protein database (nr contains all non-redundant GenBank CDS translations + RefSeq Proteins + PDB + SwissProt + PIR + PRF).
Blast output has been parsed and stored in a MySQL database together with other data produced in the intermediate steps of the pipeline.

The ESTree DB Unigene dataset is defined as the collection of all the non-redundant sequences present in the DB and is derived from CAP3 output marking as Unigene all the singleton sequences and the longest sequence of each contig.

SNP detection has been performed with the program AutoSNP version 9 (Baker et al, 2003). Tgicl parameters have been set to -p 95, -l 60, -v 20 for appropriate EST clustering.

A php-based web interface has been prepared to query the database. Users can view **sequence** data, **BLAST** outputs, **contig** alignments, and global **statistics** on sequences.

Single **sequence** and **contig** report pages are available, and fasta formatted sequences can be copied from here. Sequence quality files are available upon request.

A **text search** utility is available and queries can be performed on the **sequence** and **contig** report tables. BLAST **E-value** intervals can also be selected by users.

Local **BLASTn**, **BLASTp**, **BLASTx**, **tBLASTn** and **tBLASTx** programs are also available to perform BLAST and batch BLAST searches on the ESTree database.

In the web pages that present data in tabular form an EXCEL **spreadsheet** view is available, to retrieve and copy data in EXCEL format.

In the SNP report page SNP data can be accessed by sequence GI number, by sequence name or by SNP number.

Complete sequence and contig download is allowed from the Download page both in multifasta , CSV and NCBI format.

eSTree db

- Home Page
- Library details
- Processing, Assembly & Annotation Protocol
- Statistics on Sequences
- Sequence report
- Contig report
- SNPs
- Text Search
- Blast Search
- Links
- References
- Download
- DB contents and help

CONTIG

Contig name: Contig19
Contig length: 1311
Sequences in this contig: **Pp-S4_EST1698, PP_LEa0030J08f, PP_LEa0011B19f, PP_LEa0021I03f, S410N4, S319G11, PP_LEa0028N23f, PP_S3IIsel_A2_C10, S37D16, S411L16, S311K5, PP_S3IIall_A6_C05**
Blast Output
Best Blast Hit: **BAD33762.1**
E-value: 1e-102
Annotation: putative 6-phosphogluconolactonase [Oryza sativa (japonica cultivar-group)] ......
Source: Oryza sativa (japonica cultivar-group)...

View alignment details

Libraries and developmental stages:
- S3
- S4
- PP_Lea
- Pp_S4
- S3II


```
>Contig19 length=1311
ATCGGTCGCGCTACNTCTCGCGCGCCGTTGATCTCACTCCGAGTGTCACCAATATTCCAA
GCTCTCATCGATTTAAATCCGTTCGCACCAATTTCAAGAACCCAATTTCTGCATCGGCAT
CAGCATCAGCTAGGATGGCGGGTCAGAATAAGAAGATAGAGAAGTTTGAGACGGAGGAGG
AAGTGGCGGTGCGTTTGGCCAAGTACACCGCAGATCTGTCCGCTAAGTTNGTGAAAGAGA
```

eSTuberdb

- Home Page
- Library details
- Processing, Assembly & Annotation Protocol
- Statistics on Sequences
- Sequence report
- Contig report
- Text Search
- Blast Search
- Links
- Download
- DB contents and help

Welcome to ESTuber

A data bank of Est sequences developed at the Istituto Biologia e Biotecnologia Agraria in the frame of the CNR Strategic Project TUBER: biotecnologia della micorrizazione.

The ESTuber Consortium:

ESTuber is a Consortium of several research centers in Italy devoted to the implementation of genomics and functional genomics in truffle species.

The primary objectives of the Consortium are:

the development of an extensive EST database for Tuber species (ESTuber DB);
the analysis of several biochemical pathways based on oligonucleotide microarrays derived from ESTs collection;
the analysis of conserved genetic modules in truffle species and related filamentous fungi.

The current members of the Consortium are:

Istituto Biologia e Biotecnologia Agraria, IBBA-CNR, Milano
Istituto Genetica Vegetale, IGV-CNR, Section Perugia
Istituto per la Protezione delle Piante, IPP- CNR, Section Torino
Istituto di Biologia e Patologia Molecolari, IBPM-CNR, Roma
Dipartimento di Biochimica e Biologia Molecolare Università degli Studi di Parma
Istituto di Chimica Biologica, Centro di Biochimica delle Proteine, Università di Urbino

The ESTuber DB:

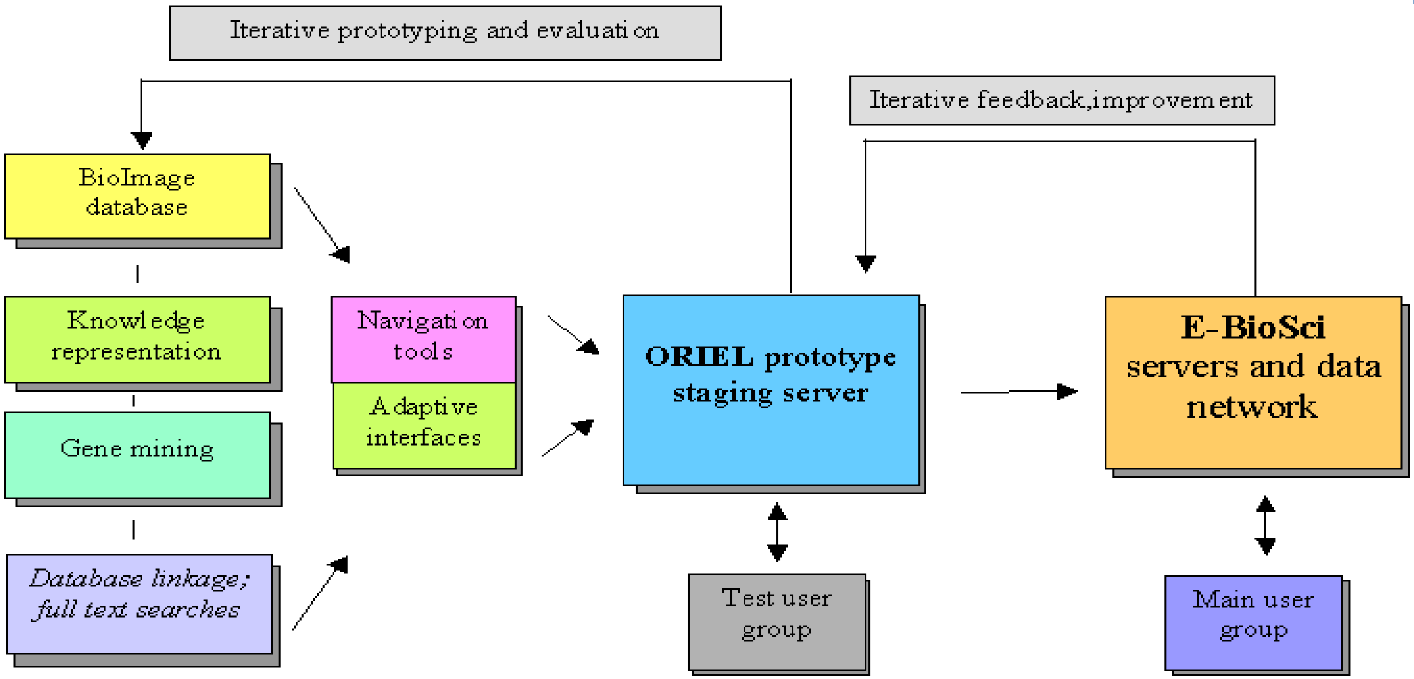
The ESTuber DB is a collection of *Tuber borchii* EST sequences.

The ESTuber DB has been developed by:

Dr. Barbara Lazzari and Dr. Andrea Caprera: DB designers and webmasters
Dr. Cristian Cosentino and Dr. Barbara Lazzari: cDNA library production and sequences managment
Dr. Luciano Milanesi: Bioinformatics
Dr. Angelo Viotti: ESTuber project manager

Italy

ORIEL & E-BIOSCI



GRID Networked data

Laboratory for Interdisciplinary Technologies in Bioinformatics

The GRID: networked data processing centres and "middleware" software as the "glue" of resources.

Researchers perform their activities regardless geographical location, interact with colleagues, share and access data

Scientific instruments and experiments provide huge amount of data

LITBIO Laboratory for Interdisciplinary Technologies in BIOinformatics

http://www.litbio.org

Molecular medicine has progressed to the point where the majority of human genes and proteins have been characterized and computational methods are necessary for further understanding of structure, mechanism and function.

Our aim is to develop a Laboratory for Interdisciplinary Technologies in Bioinformatics (LITBIO) applied to Genomics, Transcriptomics, Proteomics and Metabolomics.