Genome Sequencing & Annotation

Yow-Ling Shiue 薛 佑 玲
Institute of Biomedical Science
National Sun Yat-sen University
Ylshiue@mail.nsysu.edu.tw

Bioinformatics Flow Chart

1a. Sequencing

↓

1b. Analysis of nucleic acid sequence

↓

2. Analysis of protein sequence

↓

3. Molecular structure prediction

↓

4. Molecular interaction

↓

5. Metabolic and regulatory networks

6. Gene & protein expression data

↓

7. Drug screening

↓

Ab initio drug design OR drug compound screening in database of molecules

↓

8. Genetic variability & drug effect

Automated DNA Sequencing

- The first objective of most genome projects
- **Automated** DNA **sequencing**
  - The Principle of Sanger Sequencing = Dideoxy (**Sanger**) sequencing

Dideoxy Sequencing of DNA (1)

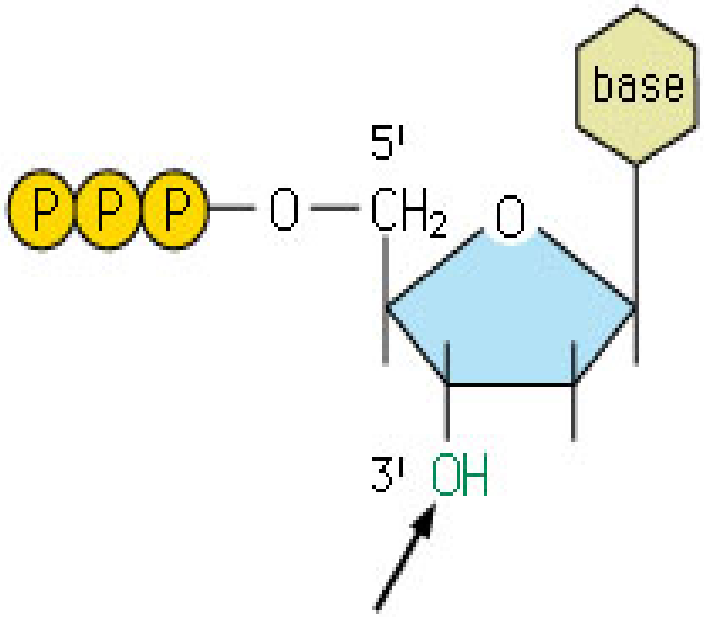
- The dideoxy or enzymatic method of DNA sequencing utilizes the principle that a **dideoxyribonucleotide triphosphate** can be incorporated into a growing DNA chain, but **can not continue synthesis**

- DNA synthesis is **terminated** and the type of dideoxyNTP (**ddNTP**) added reflects **the last nucleotide** incorporated

Dideoxy Sequencing of DNA (2)

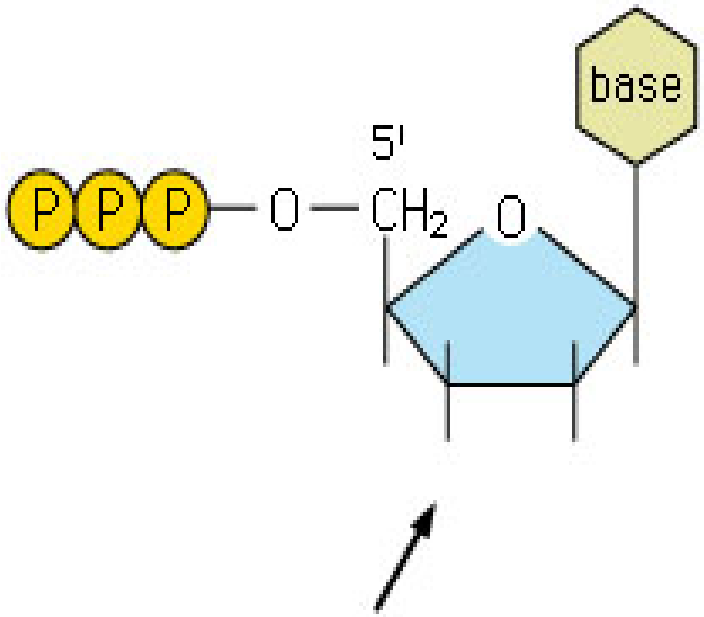
(A)

deoxyribonucleoside triphosphate



allows strand extension at 3' end

dideoxyribonucleoside triphosphate

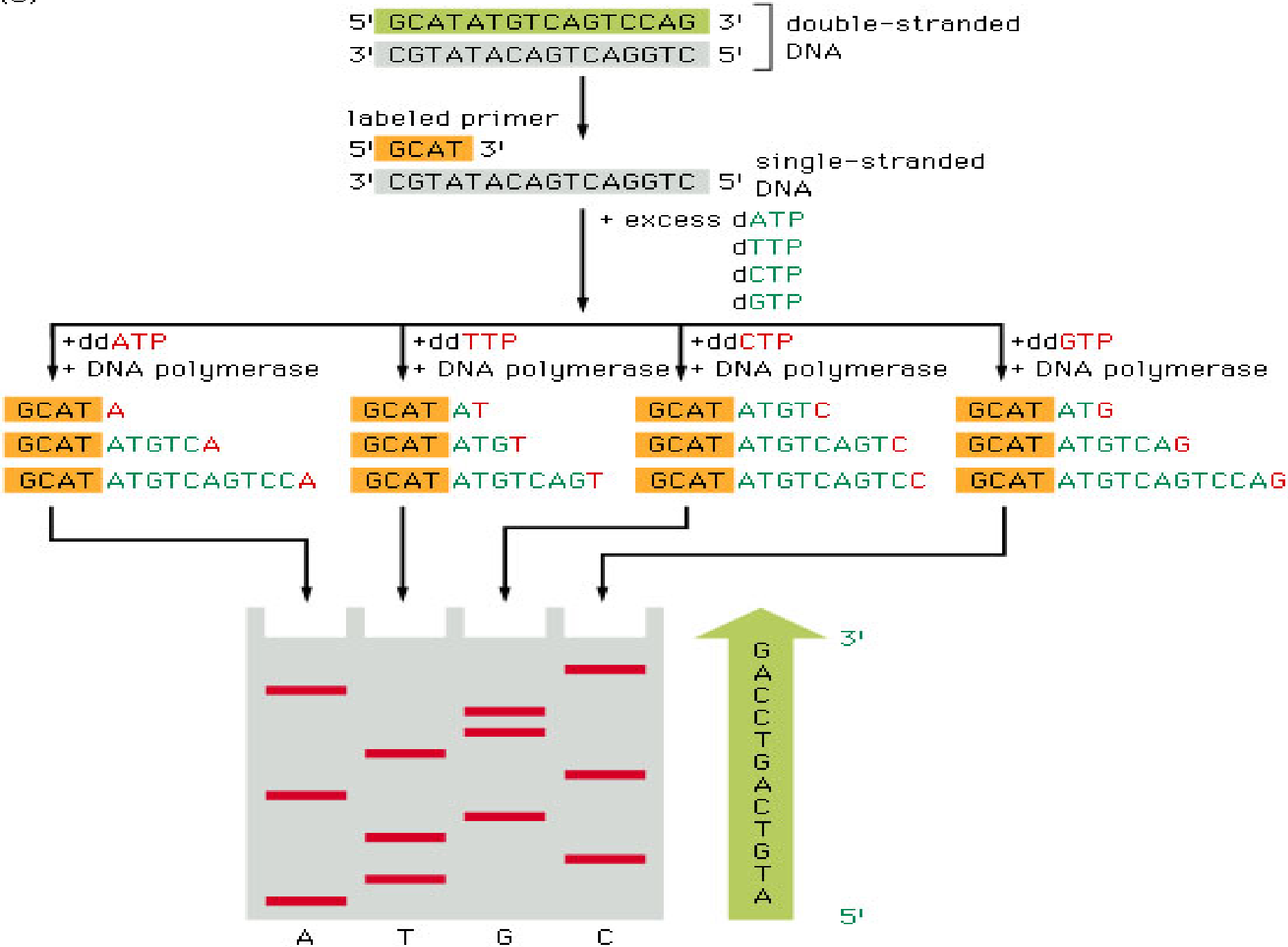


prevents strand extension at 3' end

Reactions Requirements for Dideoxy Sequencing of DNA

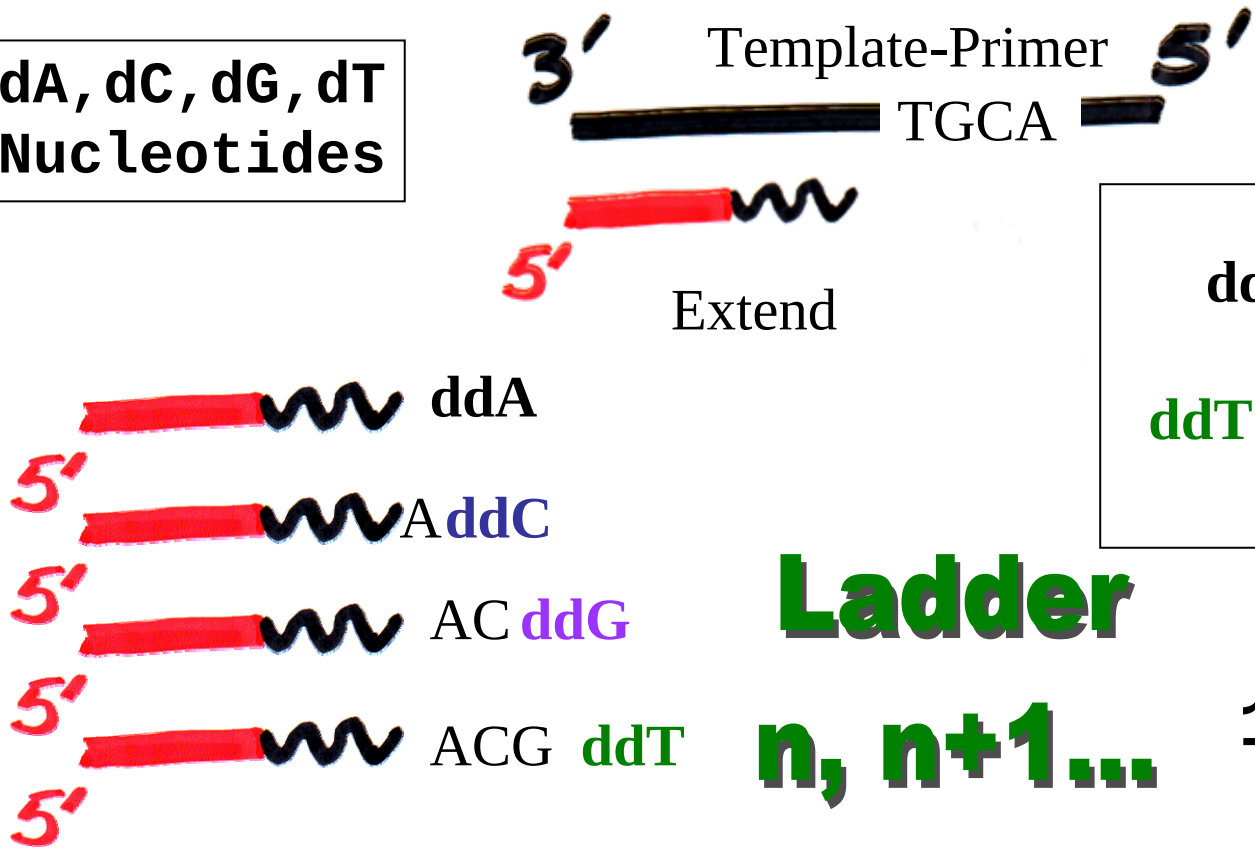
- **Single-stranded DNA** molecule (template) to be sequenced
- Oligonucleotide **primer** complementary to upstream region of template
- **DNA polymerase**
- All four dNTPs (dATP, dGTP, dCTP, dTTP)
- One of the four ddNTPs (ddATP, ddGTP, ddCTP, or ddTTP)

(C)



Automated Sequencing - Extension

dA,dC,dG,dT Nucleotides



ddA **ddC** **ddG** **ddT**

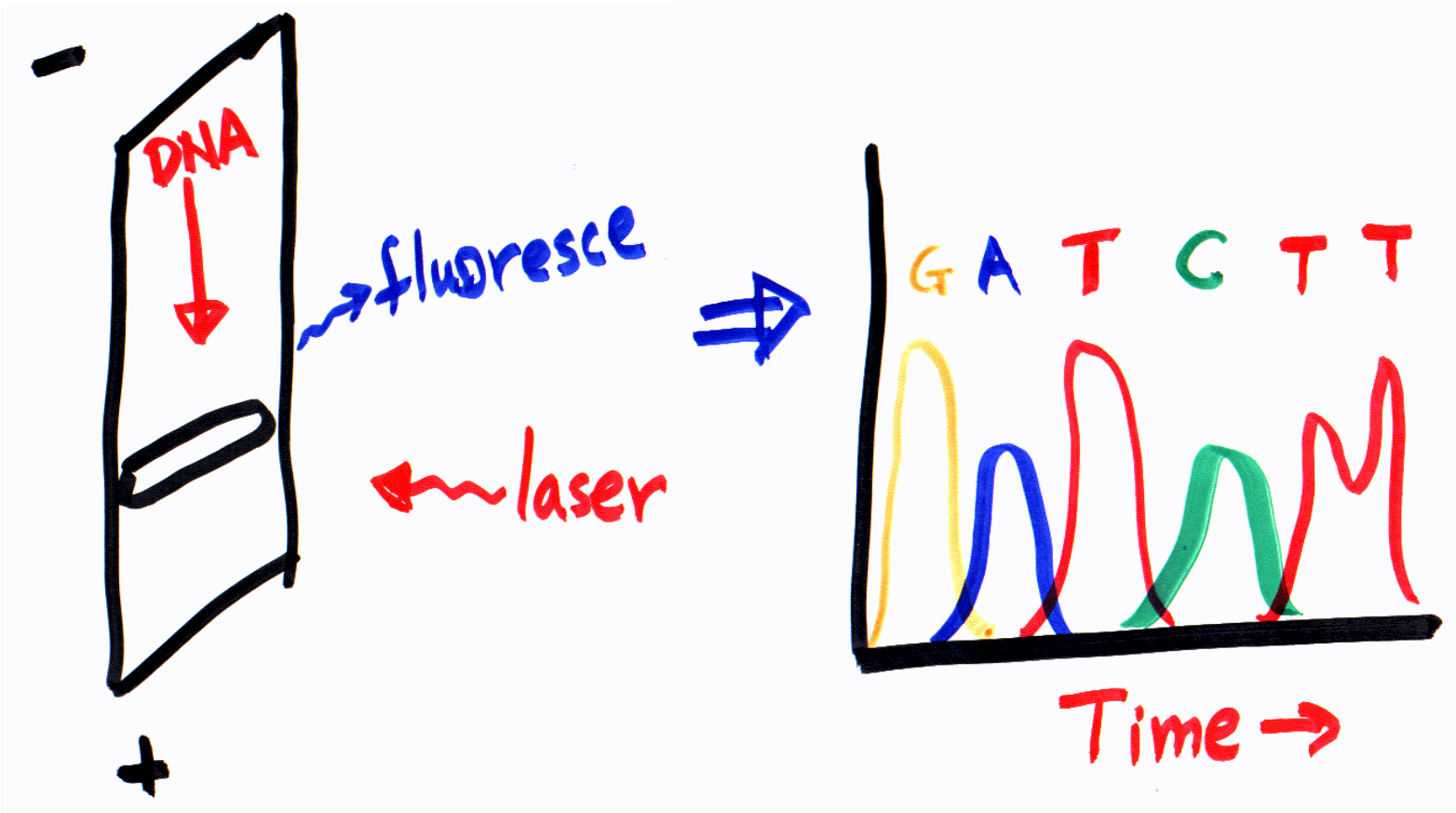
Terminators

Ladder

n, n+1…

dN	:	ddN
100	:	1

Electrophoresis



Gel or capillary tube

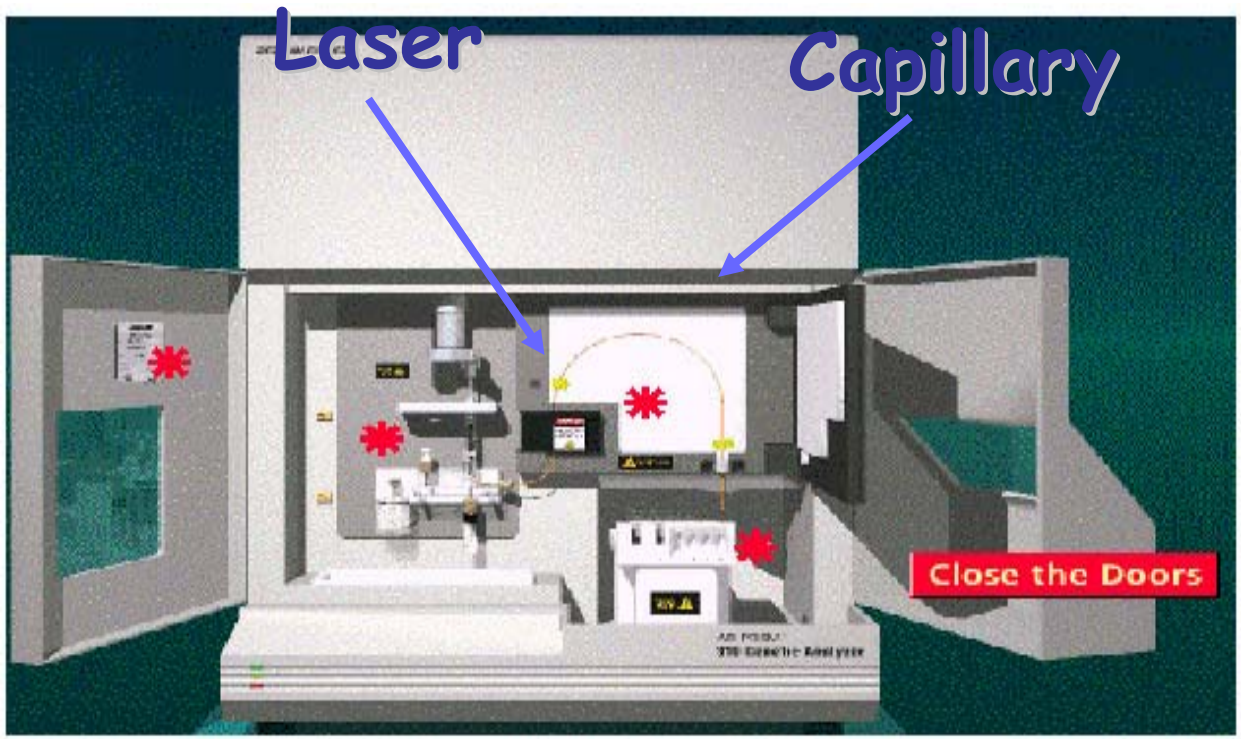
Chromatogram

Separation

- **Gel** electrophoresis
- **Capillary** electrophoresis
  - Suited to automation
    - Rapid (2 hrs vs 12 hrs)
    - Re-usable
    - Simple temperature control
    - 96 well format (= high throughput)

Automated Sequencing

- Now, we take the modified nucleotide idea further when we do "**automated sequencing**" (the most common method)
  - The ddNTPs are labeled with different color dyes, so we can mix all four in one tube
  - The move through the capillary past a **laser**, and then a photocell picks up the light emitted by the excited dye



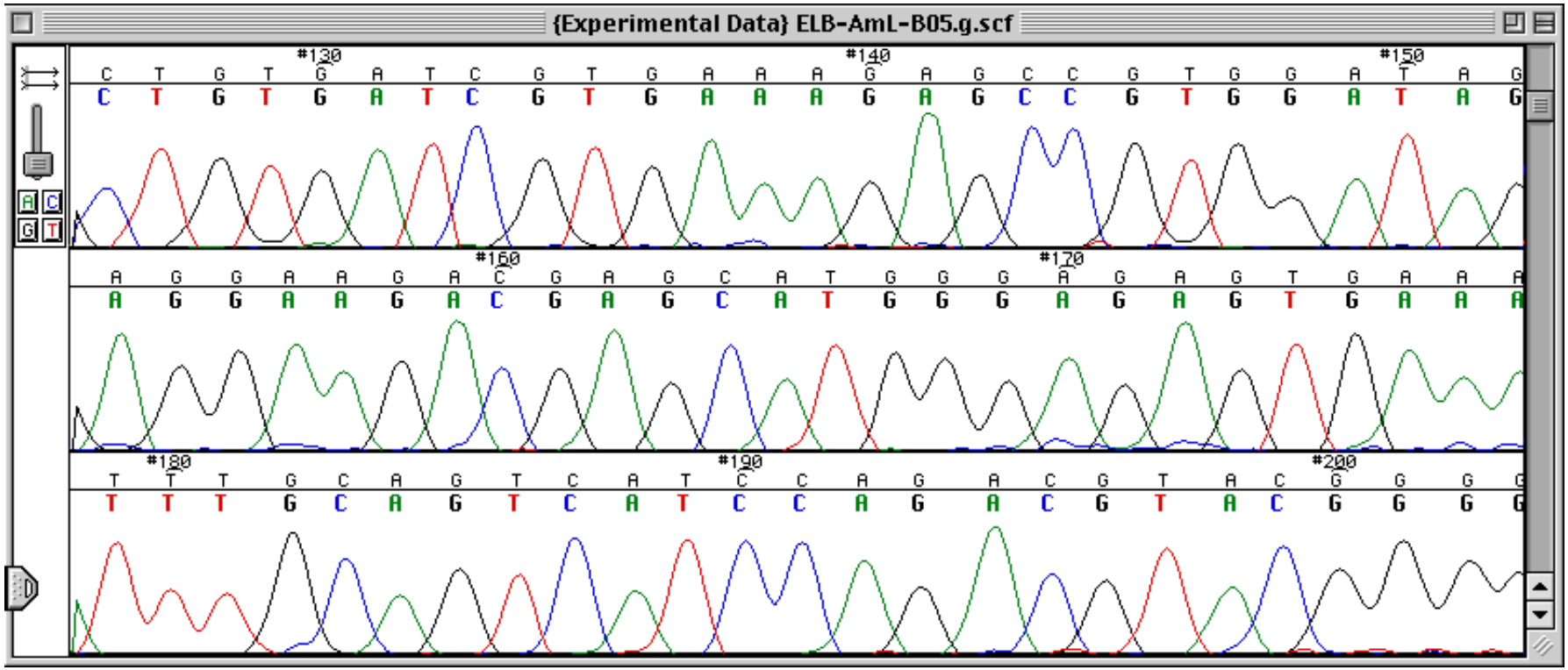
Sequencing Genomes

Many aspects of the sequencing process are truly **automated** in modern genome centers and **robotics** are used extensively

Automated Sequencing Statistics

- Electrophoresis can be done on a gel (older sequencers) or capillary
- **"All four reactions" done together**
- Can obtain **up to 1,000** nucleotides per run
- Entire process (post DNA prep.) take less than one day
- Makes graduate students happy!
- **Present day fact**: sequencing is cheap that it is often "sent out" at ~$12/reaction"
  - More cost effective and faster for labs to pay sequencing centers to perform this task than to do it themselves

- As the labeled fragments move past **the detector** the fluorescence of the dye will be detected
  - The results are shown as a chromatograph (below)
  - The dyes alter migration slightly and the signal intensity changes through the run, so some processing is necessary
  - Reading the **traces: base-calling**



Phred Base-calling Program

- University of Washington
  - Ewing *et al.* 1998; Ewing & Green 1998
- phred (free available)
  - Convert **computer-generated traces** into **base sequences** and access the probable accuracy of each base call (http://www.phrap.org)

The *Phred* Base-calling Algorithm

- Locate predicted peaks, using **Fourier methods** to fit best distribution
- Locate observed peaks, for which the area under the **concavity** exceeds 10% of the previous 10 peaks or 5% of the previous one
- Match observed and predicted peaks using a three-stage shifting algorithm
- Find missing peaks
- Assess error probabilities of each peak according to four-parameter model

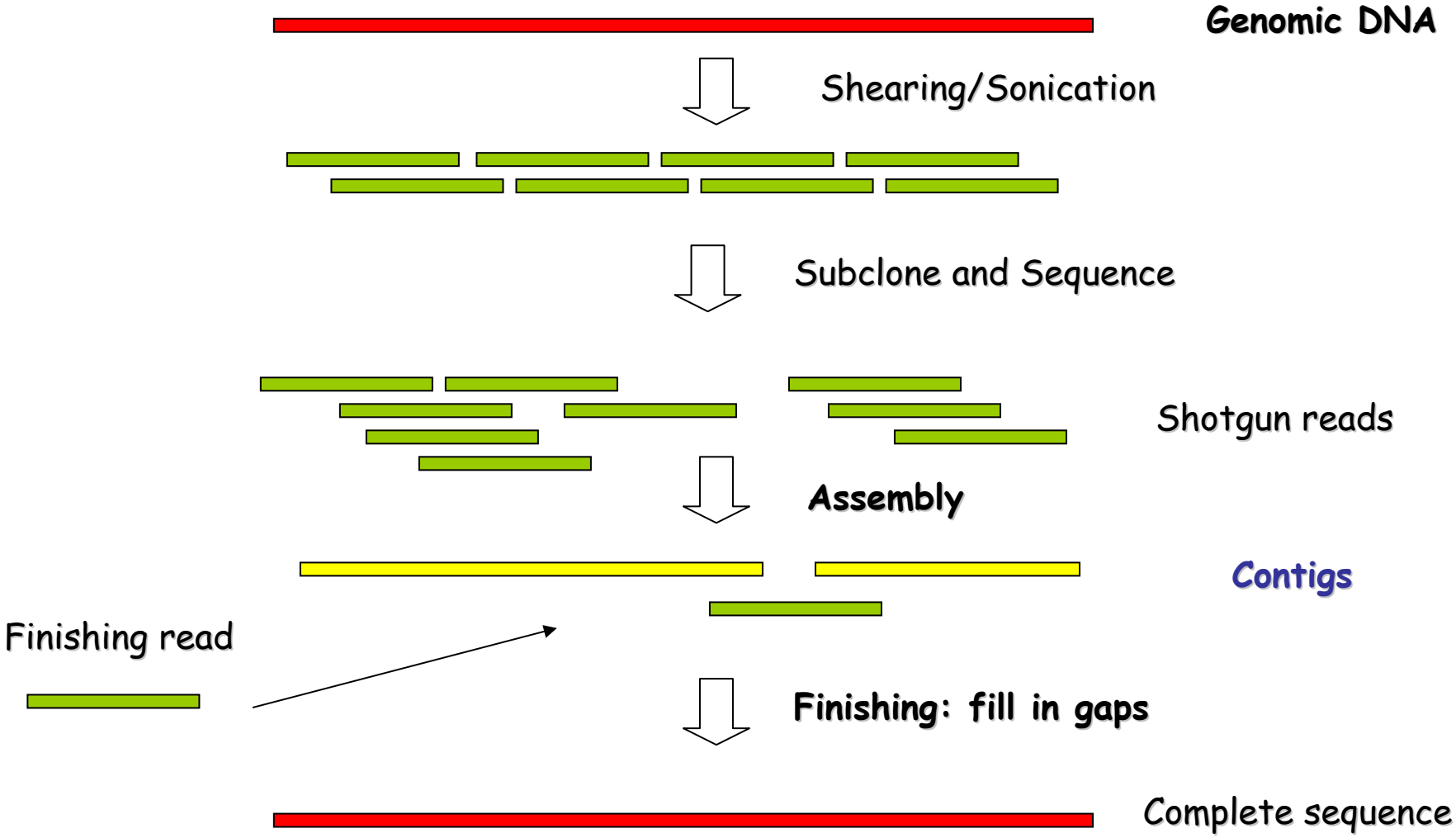
Sequencing Whole Genomes (1)

- **The reads are limited in length**
  - All sequencers ultimately produce reads in the range of **500 to 1,000 base pairs**
    - Some technologies that can produce reads a little longer, but **there is a limit**
    - So, longer sequences -- whether they are a specific gene, a chromosome region, a chromosome, or a genome -- must be **assembled from shorter sequences (contig assembly)**

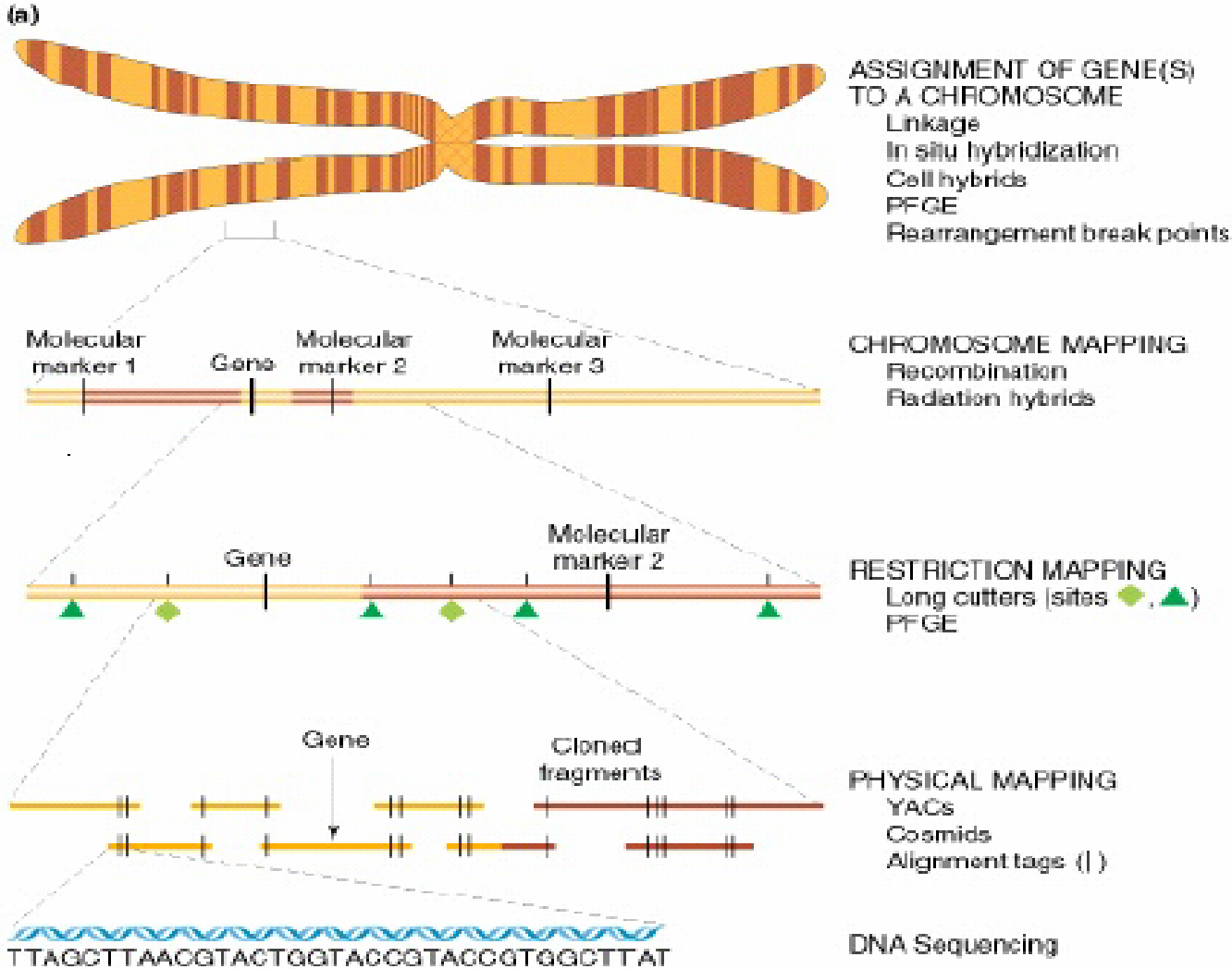
Sequencing Whole Genomes (2)

- For complete genomes, there are **two strategies**
  - **"Whole-genome shotgun"** -- a large number of reads are collected and then they are assembled computationally
    - Less expensive - popular for **microorganisms**
  - **Mapping followed by sequencing** -- large clones are mapped before they are sequenced, and then the complete genome is assembled using the map data **(map-based)**

Whole-genome Shotgun



Mapping Followed by Sequencing



Contig Assembly (1)

- Gordon *et al.* 1998
  - *Phrap* assembler
  - ***Consed*** graphic editor

- **Key features of successfully editors**
  - The use of **color** to illustrate key features such as the different bases
    - Different bases, **quality scores**, regions of **sequence conservation**, contrasts between automated and manual base calls

  - The ability to **view and navigate** along **the actual traces** of the sequences being compared, and to **tag** ambiguities and **features of interest** with notes

Contig Assembly (2)

- **Key features of successfully editors**
  - Easy display of **the complementary strand**
  - Tools for **manual sequence editing**, including inserting and deleting bases, without disrupting the original trace files yet propagating the edits throughout the assembly as appropriate, including **adjustments** to linked output files as requested by the editor
  - A **flexible alignment algorithm** that implements **user-defined alignment parameters**

Contig Assembly (3)

- **Key features of successfully editors**
  - Computation of **probability scores** associated with a calculated consensus sequence
  - The ability to identify **potentially polymorphic sites**
    - *E.g.*, SNPs
  - Provision of tools to guide **error correction**

Health or Disease?

DNA Sequence

Person 1 A A A T T T

Normal protein

Some DNA variations have no negative effects

Person 2 A A T T T T

Low or nonfunctioning protein

Person 3 A A C T T T

Other variations lead to disease (e.g., sickle cell) or increased susceptibility to disease (e.g., lung cancer)

YGG-00-0480

Trace Editing

- Other programs that allow you to view and edit chromatograms
  - Vector NTI
    - ContigExpress
  - Chromas
    - Windows/WinNT freeware
  - Consed
    - Unix-based

What Types of Sequences are Present in Genomes? (1)

- **Complete** (or nearly complete) sequences of genomes
  - The definition of many types of sequences and revealed the **complete set of genes** present in organisms

- The **types of sequences** present in genomes
  - **The relative copy number of different portions of the genome**
    - DNA **renaturation kinetics**
  - To examine the nature of the genomes
    - Density gradient centrifugation
  - Sequences were characterized based on **slight differences** in the **buoyant density** of **GC-rich and AT-rich** sequences

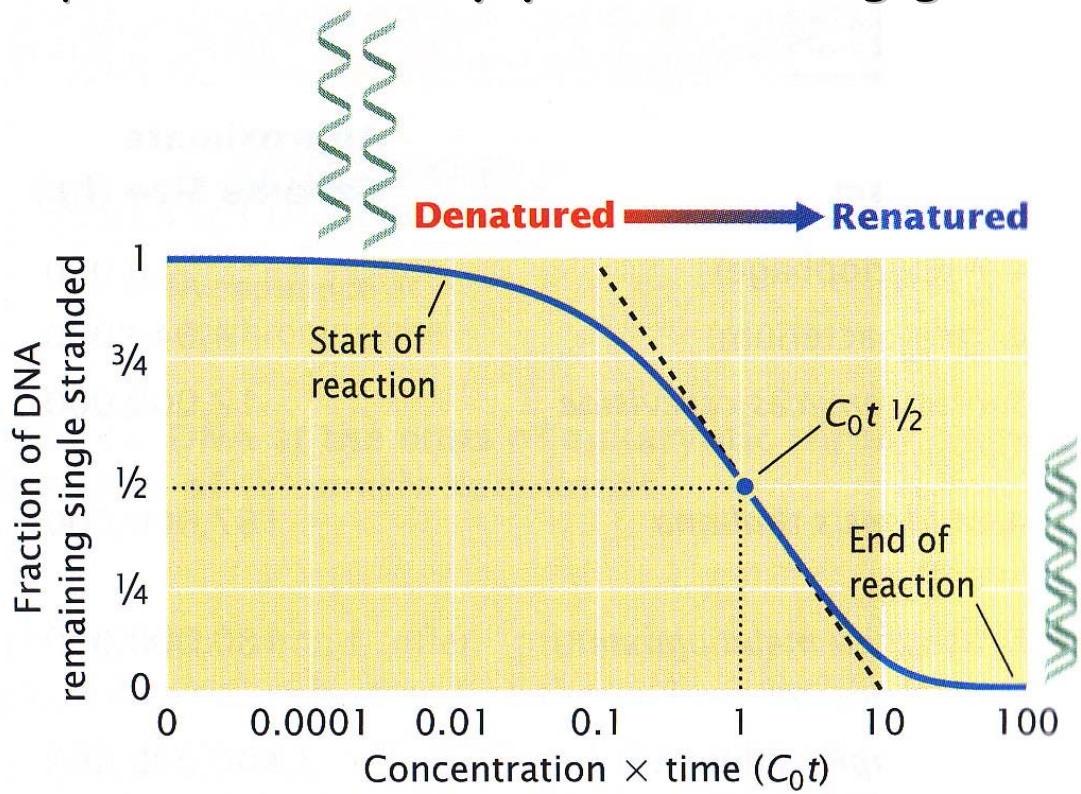
What Types of Sequences are Present in Genomes? (2)

- DNA **renaturation kinetics** involves "melting" the DNA and then monitoring the formation of double-stranded DNA
  - If we consider double stranded DNA, it is possible **to separate the strands** by increasing the temperature
    - Denaturing or "melting" the DNA
    - The temperature at which the DNA is separated is determined by **the GC content**, in large part
  - If we renature the DNA, the **concentration** (*C*) of double-stranded DNA **at time *t*** is given by:

$$\frac{C}{C_0} = \frac{1}{1 + kC_0 t}$$

A Single Class of Sequences in Bacteria

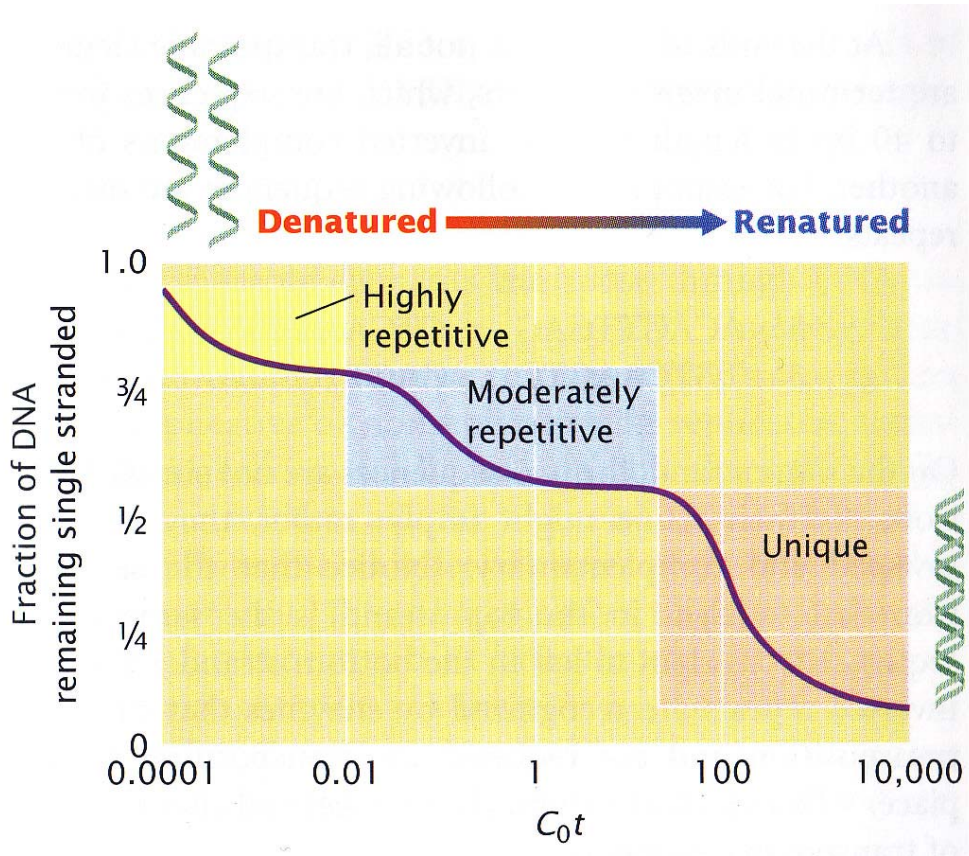
- Bacteria have **little repetitive DNA**, so the C_0t curve is fairly simple
  - There are **a few repetitive sequences** in bacteria, such as the **rRNA genes**, but bacterial genomes are characterized by the presence of **many protein-coding genes** and little else



…and Multiple Classes of Sequences in Complex Eukaryotes

- Complex eukaryotes like **plants** and **animals** tend to have **more complex C_0t curves**

- This reflects the fact that there are **many repetitive sequences** in eukaryotic genomes
- For example, **humans** have about **15% highly repetitive DNA** and 10% middle repetitive DNA
- Repeats with a **distinct** nucleotide composition appear as **SATELLITE DNA** in density gradient centrifugation



Annotation

Introduction

- "It is now apparent that the **bottleneck** in genomics is no longer in sequencing the genomes but lies in their **annotation**"
  - Thanaraj *et al.* - Paradigm shifts in the Approaches for Gene Annotation (2000)

Databases & Annotations

- **A database allows the**
  - Rapid and public scrutiny of data
  - **Analysis** of the data by others (*e.g.* for phylogeny and gene finding)
  - And provides a basis for **research** as a collective endeavor

- A well designed database that is **of optimal use to the researcher** should
  - Be well curated
  - Have a minimum level of redundancy
  - Have a high level of **integration** with other databases
  - Have a high level of **annotation**

What is Annotation? (1)

- Sequence database entries have **core data** that comprises of
  - **Sequence** data
  - Citation information
  - Taxonomic data

What is Annotation? (2)

- The **annotation** may consist
  - Identification of **gene structural elements**
  - **Functions** of the protein
  - Post-translational modifications
  - **Domains** and sites
  - Secondary structure
  - **Quaternary structure**: to describe the protein composed of multiple subunits (several monomers)
  - Similarities to other proteins
  - **Diseases**
  - Sequence conflicts
  - *Etc.*

EMBL/GenBank/DDBJ Annotations

Warning !

- DNA data base annotations are **full of errors**
  - In sequences, in annotations, in CDs attribution…
  - No consistency of annotations
  - Most annotations are done by the **submitters**
  - Heterogeneity of quality and updating

Some Interesting Sequence Annotation

```
FT   source        1..124
FT                 /db_xref="taxon:4097"
FT                 /organelle="plastid:chloroplast"
FT                 /organism="Nicotiana tabacum"
FT                 /isolate="Cuban cahibo cigar, gift from President Fidel
FT                 Castro"
```

Or:

```
FT   source        1..17084
FT                 /chromosome="complete mitochondrial genome"
FT                 /db_xref="taxon:9267"
FT                 /organelle="mitochondrion"
FT                 /organism="Didelphis virginiana"
FT                 /dev_stage="adult"
FT                 /isolate="fresh road killed individual"
FT                 /tissue_type="liver"
```

???

Taxonomy Browser @ EBI:

http://www.ebi.ac.uk/newt/

NEWT - Netscape 6

http://www.ebi.ac.uk/newt/index.html

Search For : delphis virginiana

To search the taxonomy database, please enter a search term in the box above then click on a button to search by keywords or taxonomic ID.

Archaea
Bacteria
Eukaryota
Viroids
Viruses

SWISS-PROT at ExPASy
SWISS-PROT at EBI
List of all species in SWISS-PROT
NCBI Taxonomy Database
Contact

Didelphis marsupialis virginiana (North American opossum)

Lineage	Tax ID	9267	External information
• Eukaryota • Metazoa • Chordata • Craniata • Vertebrata • Euteleostomi • Mammalia • Metatheria • Didelphimorphia • Didelphidae • Didelphis	OS code	DIDMA	
	Scientific name	Didelphis marsupialis virginiana	
	NCBI synonyms	Virginia opossum Didelphis virginiana North American opossum	
	Rank	species	
	Number of SWISS-PROT entries	35	
	Number of TrEMBL entries	21	

Taxonomy navigation	
Up taxonomy tree	Down taxonomy tree
Didelphis	*This is the last node of the tree*

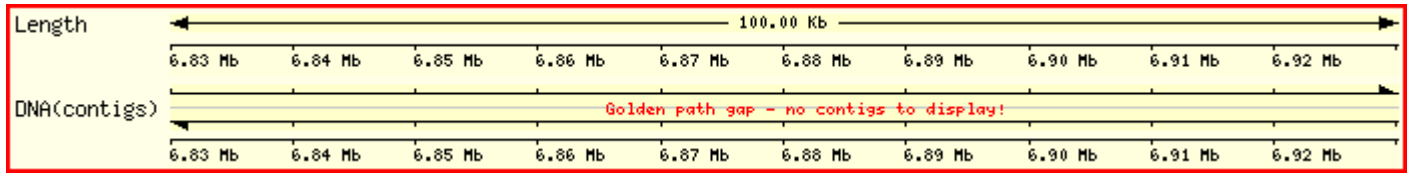
Source of data : Swiss-Prot

Document: Done (5.127 secs)

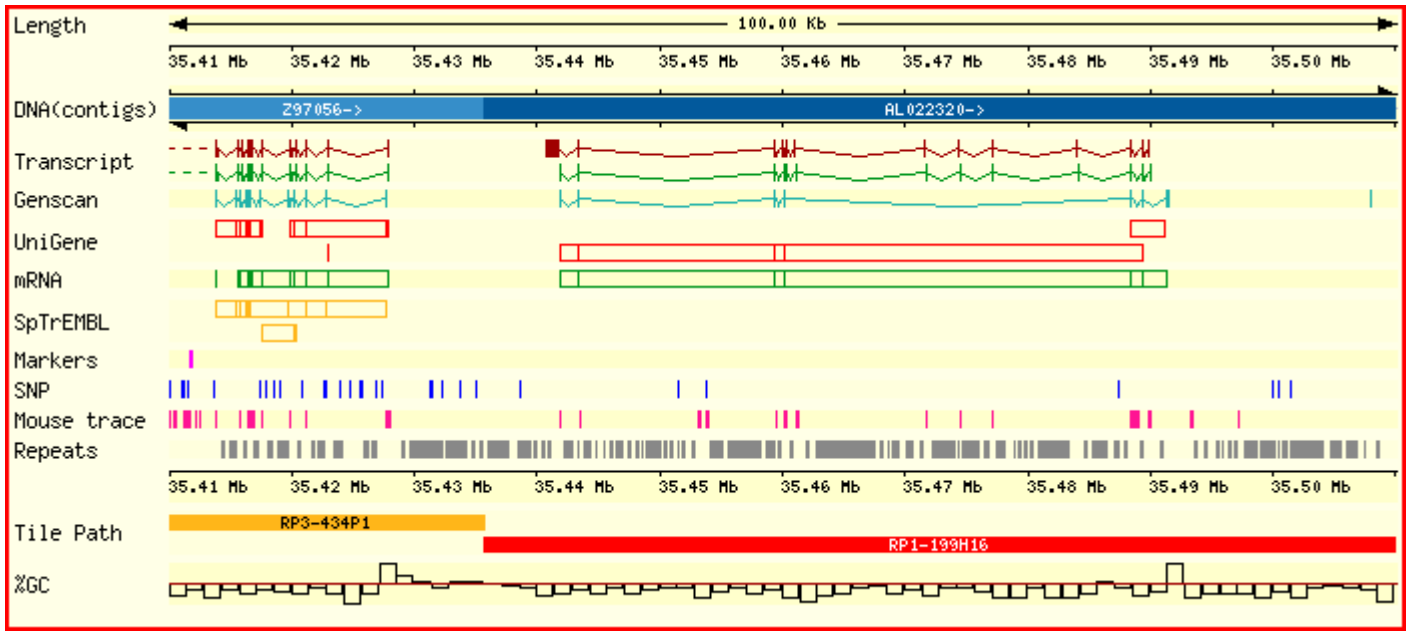
Related Projects

- **Whole genome sequencing**
  - Human & model organisms
- **EST sequencing**
  - CDs (coding DNAs)
  - Alternative splicing
- ***Ab initio* gene discovery**
  - *E.g.*, Hidden Markov Models (HMM)
- **Domain prediction**
  - Transmembrane
  - Signal peptied
- **Non-protein coding genes**
  - **Regulatory sequences**
- **Functional annotation and gene family clusters**, *e.g.*,
  - Homology search
  - Clustering of genes by **sequence similarity** (paralogues)
  - Clusters of **orthologous genes** (real homologues)
  - **Phylogenetic** classification of genes
  - **Gene ontology (GO)**

Genome Annotation – The Aim



Ensembl



UniGene

- A database created and maintained at NCBI as an experimental system for automatically partitioning **expressed nucleotide sequences** into a non-redundant set of gene-oriented clusters

- Each **UniGene cluster** contains sequences that represent **a unique gene**, as well as related information such as **the map location** and **tissue types** in which the gene has been expressed

- UniGene is particularly important for reducing the redundancy and complexity of **EST data** and is an important resource for gene discovery

Expressed Sequence Tag (EST)

- A short (300-1,000 nucleotide), single-pass, single-read DNA sequence derived from **a randomly picked cDNA clone**
- EST sequences comprise the largest GenBank division
- There are numerous **high-throughput sequencing projects** that continue to produce **large numbers** of EST sequences for important organisms
- Many ESTs are classified into **gene-specific clusters** in the UniGene data set

Genome Annotation – Context

- **Gene identification**
  - Known genes
    - Where? (location)
    - Genome structure/organization
    - Transcript(s)?
    - Protein(s)?
    - Attach useful links
  - Novel genes
    - How to predict?
      - Evidence required
    - Transcript(s)?
    - Protein(s)?
    - Attach useful links
  - Transcript= A sequence of RNA produced by transcription from a DNA template

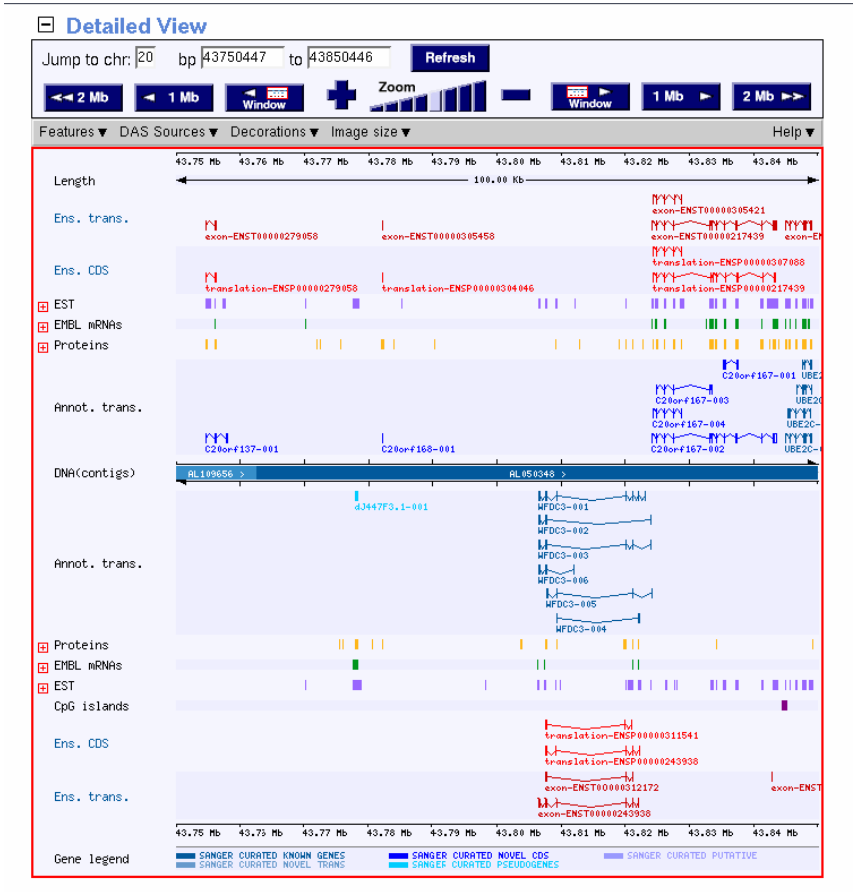
Genome Annotation – Approaches

- **Automatic annotation**
  - EnsEMBL, EBI
  - MapViewer, NCBI
  - Genome Browser, UCSC

- **Manual curation**
  - G16 Finishing Standards for the Human Genome Project
    - Standard Finishing Practices and Annotation of Problem Regions for the Human Genome Project
      - http://genomeold.wustl.edu/Overview/finrulesname.php?G16=1
      - General rules
      - Vocabulary
  - WormBase
  - FlyBase

Manual Curation – Identifying Genes

- Known
- Novel
- Novel transcript
- Putative
- Pseudogene



Vega Genome Browser - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

上一頁 搜尋 我的最愛

網址(D) http://vega.sanger.ac.uk/index.html 移至 連結

Google Vega AND annotation 搜尋 PageRank 17 已攔截 檢查 選項 Vega annotation

Vega

Search all Vega: Anything Go

Vega v16 - Dec 2005 Help

Use Vega to...

- BLAST
- Text search
- Export data

Information

- Site Map
- What's New
- Information

Other links

- Home
- Human
- Mouse
- Zebrafish
- Dog
- Ensembl

HAVANA

sanger

About Vega

The Vertebrate Genome Annotation (VEGA) database is a central repository for high quality, frequently updated, manual annotation of vertebrate finished genome sequence.

Details of the projects for each species are available through the homepages for human, mouse, zebrafish and dog.

The website is built upon code from the Ensembl project.

Browse a genome

Homo sapiens [12-12-2005]
browse | Ensembl | Glovar

Mus musculus [03-08-2005]
browse | Ensembl

Danio rerio [25-04-2005]
browse | Ensembl

Canis familiaris [14-02-2005]
browse | Ensembl

What's New

- New human database
- Website redesign
- Consolidation of gene track names
- Gene types

e!mpowered

We are keen to receive extra information on annotation, and to hear your comments, problems, and suggestions on Vega. Please send us feedback.

網際網路

NewsView - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

上一頁 搜尋 我的最愛

網址(D) http://vega.sanger.ac.uk/Multi/newsview?rel=16 移至 連結

Google Vega AND annotation 搜尋 PageRank 17 已攔截 檢查 選項 Vega annotation

Vega

Search all Vega: Anything Go

Vega v16 - Dec 2005 Help

News Archive

- Select news to view

Use Vega to...

- BLAST
- Text search
- Export data

Information

- Site Map
- What's New
- Information

Other links

- Home
- Human
- Mouse
- Zebrafish
- Dog
- Ensembl

Release 16 (12th December 2005)

Data updates

New Human Database (*Homo sapiens*)

This release of Vega contains a new human database with the full complement of chromosomes:

- Updated annotation for the 14 fully annotated chromosomes previously shown in Vega (1, 6, 7, 9, 10, 13, 14, 16, 18, 19, 20, 22, X and Y)
- New annotation of the 44 ENCODE regions spread across the whole genome.
- New annotation of ORFs and UTRs of selected loci as part of the CORF project (chr 2, 3, 4, 5, 8, 11, 12, 15, and 17).

Gene types in Vega (*Homo sapiens*, *Mus musculus*, *Danio rerio*, *Canis familiaris*)

Genes in Vega are now classified according to 'status' and 'biotype'. The former is a measure of the status of the annotation (NOVEL, KNOWN, etc) and the latter is indicative of biological function (Protein coding, Pseudogene, etc). Further details of the relationship between the HAWK and the new gene types can be found here

Improved links to external databases (*Homo sapiens*)

The method by which links from Vega genes to external databases such as HUGO and LocusLink are generated has been improved.

Web features

Website redesign

This release of Vega is the first using the new Ensembl webcode base. The aims of this redesign are to help navigation and to improve discoverability.

The large majority of pages are unchanged in content but the appearance of the site has altered considerably. Major new features are:

- Introduction of a menu bar on the left hand side of the page containing many of the links that were previously buried within the data itself (for example the link to ProteinView from GeneView is now shown as 'View Protein Information' in the menu bar). The sections displayed in the menu are themselves context sensitive.
- An 'Information' section, accessible from the menu bar, contains usefull information on the Vega project and the data shown on the web

網際網路

Five Agreed Groups of Genes (1)

- **Known**
  - Which have identities in databases like RefSeq (Entrez Gene) **(NCBI)** and are fully supported by **protein** and **DNA evidence**
- **Novel**
  - Which have evidence which either isn't complete or from a different organism *e.g* a human gene supported by **mouse ESTs** and **protein**

Five Agreed Groups of Genes (2)

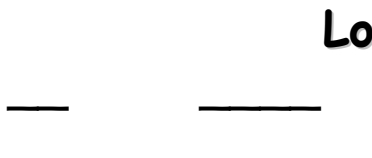
- **Novel transcript**
  - Have multiple potential ORFs and no protein evidence
- **Putative**
  - Only supported by ESTs and have no ORF covering on exon
- **Pseudogene**

Identification of Homologous Genes

- **Pairwise Sequence Alignments**
  - Needleman & Wunsch (1970)
    - **Global alignment** – similarity **across the full extent** of the sequence
      - End to end
  - Smith & Waterman (1981)
    - **Local alignment** – regions of similarity **in parts of** the sequences

Global

Local



Searching Sequence Database Using BLAST (Altschul *et al.* 1991)

- The single **most influential** bioinformatics tool
- Score (substitution) matrices
  - **PAMs**
    - Dayhoff *et al.* 1978
  - **BLOSUM** family
    - Henikoff & Henikoff (1992)
    - Practical experience suggests that the **BLOSUM62** matrix is quite useful for general use

RID=1143266588-22667-131914421252.BLASTQ1, gi|6633795|gb|AF123456.2|AF123456Gallus gallus doubl - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

上一頁 搜尋 我的最愛

網址(D) http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi 移至 連結

Google 搜尋 PageRank 17 已攔截 檢查 選項

Sequences producing significant alignments:

		Score (Bits)	E Value	
gi|6633795|gb|AF123456.2|AF123456	Gallus gallus doublesex and...	1362	0.0	U G
gi|6601569|gb|AF211349.1|AF211349	Gallus gallus doublesex and...	1237	0.0	U
gi|50762072|ref|XM_424927.1|	PREDICTED: Gallus gallus doubles...	1185	0.0	G
gi|40716455|gb|AY448019.1|	Gallus gallus DMRT1 isoform b mRNA...	1170	0.0	U G
gi|40716457|gb|AY448020.1|	Gallus gallus DMRT1 isoform c mRNA...	652	0.0	U
gi|32402377|gb|AY316537.1|	Trachemys scripta doublesex and ma...	371	5e-99	
gi|66841203|dbj|AB179697.1|	Pelodiscus sinensis DMRT1 mRNA fo...	371	5e-99	
gi|8572624|gb|AF201387.1|AF201387	Trachemys scripta doublesex...	355	3e-94	
gi|13384037|gb|AF335421.1|AF335421	Lepidochelys olivacea doub...	331	4e-87	
gi|40716465|gb|AY448024.1|	Gallus gallus DMRT1 isoform g mRNA...	321	4e-84	U
gi|40716463|gb|AY448023.1|	Gallus gallus DMRT1 isoform f mRNA...	321	4e-84	U
gi|40716459|gb|AY448021.1|	Gallus gallus DMRT1 isoform d mRNA...	305	2e-79	U
gi|40716461|gb|AY448022.1|	Gallus gallus DMRT1 isoform e mRNA...	297	5e-77	U G

網際網路

Automated Annotation (1)

- Currently databases cannot cope with manually annotating **the vast amount of sequence data**, they need **automatic methods** to **find the genes** and then **predict the structure** and **functionality** of the proteins

- The annotation level for **each species** crucially depends on **the existence of homologs** (vs. paralogs) of the proteins that are potentially encoded in the genome sequence
  - **Comparative genomics**

Automated Annotation (2)

- Regions **within a genome** can **differ in features** such as
  - **Gene density**
  - **GC content**

- Subtle details of **statistical properties** required for the *ab initio* gene prediction methods can differ for genome to genome

Automated Annotation (3)

- Limitations in the representatives of **protein** and **EST sequence** collections can limit the **reliability** of predictions

- Evolution of **function** and **sequence** **may not be as tightly linked** as is sometimes believed, making accurate predictions using homology inferences difficult

Automated Annotation (4)

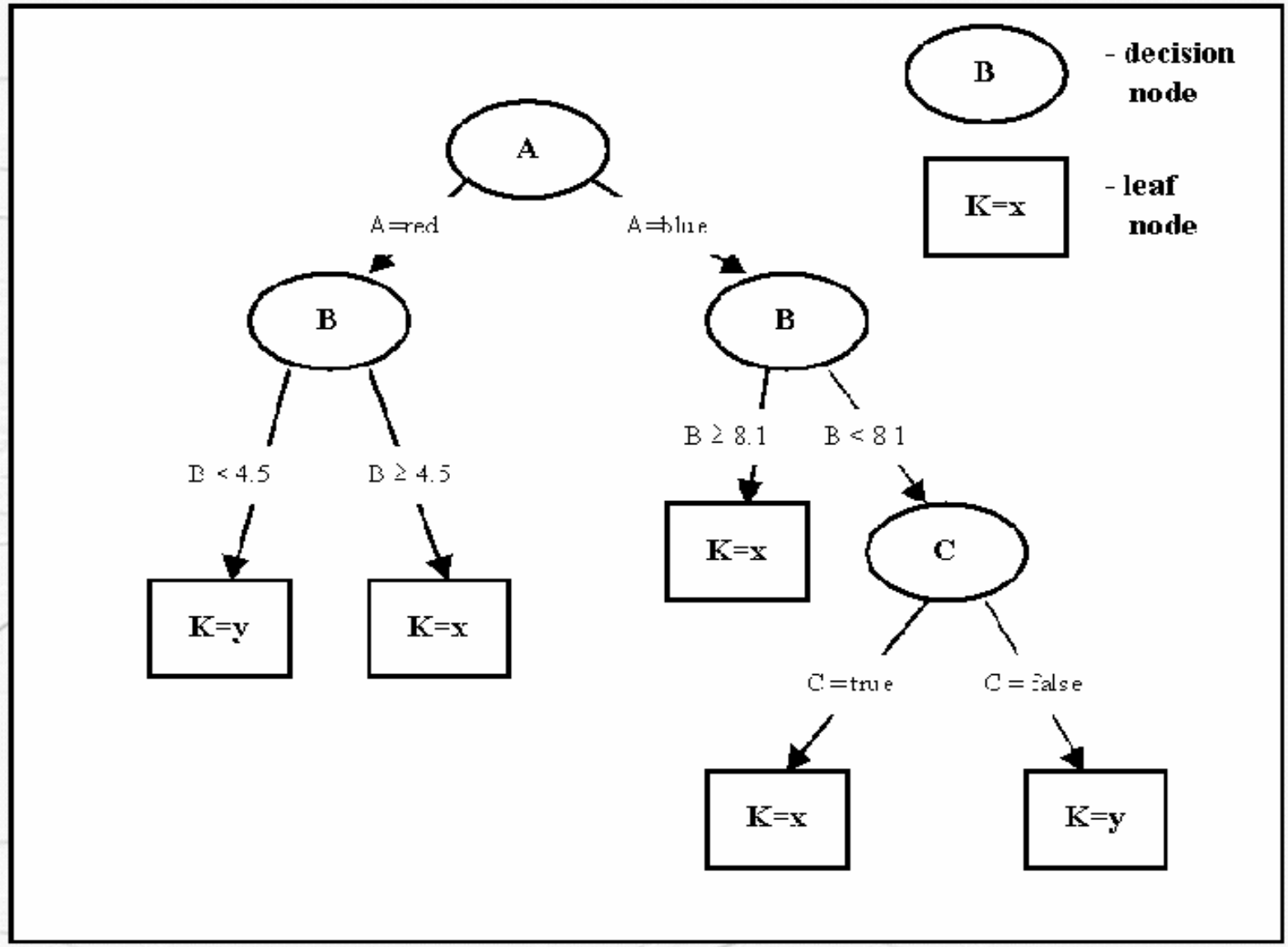
- A more general problem is that **the *ab initio* methods** for **gene prediction** are trained and optimized on **short to medium length DNA sequences** containing **single genes**, rather than the current and much longer genomic sequence, such as regions containing **long introns**, where the current programs are known to perform **poorly**

Gene Annotation Methods

- Basic methodologies for identifying gene structural elements
  - **Signal** based
  - **Content** based
  - **Similarity** based

- Statistical and mathematical techniques used include
  - **Decision tree approaches (powerful rules)**
    - The accuracy of a **classification** or **prediction** is the only thing that matters
  - **Discriminate analysis**
  - **Various statistical approaches, *etc*.**

Example of a Simple Decision Tree



Swiss-Prot and TrEMBL (1)

- Swiss-Prot
  - Curated protein sequence database
  - Provides a high level of annotation
  - Minimal redundancy and a high level of integration

Swiss-Prot and TrEMBL (2)

- **TrEMBL**
  - Contains the **translations** of all coding sequences (CDs) present in the EMBL nucleotide sequences
  - **SP-TrEMBL** CDs that should be given accession numbers and should be incorporated into SP
  - REM-TrEMBL entries that wont be included in SP
    - Contains the entries that we do not want to be included in SwissProt
      - REM-TrEMBL entries have **no accession numbers,** these including
        - **Immunoglobulins** and **T-cell receptors**, **synthetic sequences**, **patent application sequences**, small fragments, CDs not coding for real proteins
    - REMaining TrEMBL

Genequiz (1)

- http://www.cmbi.kun.nl/swift/genequiz/

- **Integrated system** for **large-scale biological sequence analysis** (highly **automated** analysis)
  - Protein sequence ⇨ **biochemical function**
    - Using a variety of **search** & **analysis methods**, and up-to-date **protein & DNA databases**
    - Applying **an "expert system" module** to the results of the different methods

- Create a compact summary of findings, focusing on
  - Deriving **a predicted protein function**, based on
    - The available evidence

Automated sequence analysis by GeneQuiz

On July 28, 1995, the complete DNA sequence of the genome of the bacterium *Haemophilus influenza* was released to the public and the results of TIGR's sequence analysis were published [Fleischmann et. al., 1995],The publication included functional assignments for many of the open reading frames (the predicted protein products).

However, the biological knowledge used for the analysis, most of it stored as annotation in the DNA and protein databases and in the primary literature, as well as the analysis methods are in fast and continuous evolution. As a result, any genome analysis can be easily out of date in a short time.

In these pages we offer an alternative: the analysis of the *H. influenza* open reading frames continuously updated with the latest version of the databases and with the application of an expanding variety of methods, including methods for homology, pattern and profile comparison and for the prediction of protein structure. The database update procedures and the search and analysis methods are embodied in the framework of **GeneQuiz**, a system for automated protein sequence analysis.

From these pages you can access in different ways the latest results obtained by **GeneQuiz** for the open reading frames of *Haemophilus influenza:* functional assignments, the reliability of the assignments, data on which the assignments are based, e.g., alignments, sequence patterns detected, sequence family information, 3D structure models etc.

Genequiz (2) – the Available Evidence

- The evaluation of **the similarity to** **the closest homologue** in the database
  - Identical
  - Clear
  - Tentative, or
  - Marginal

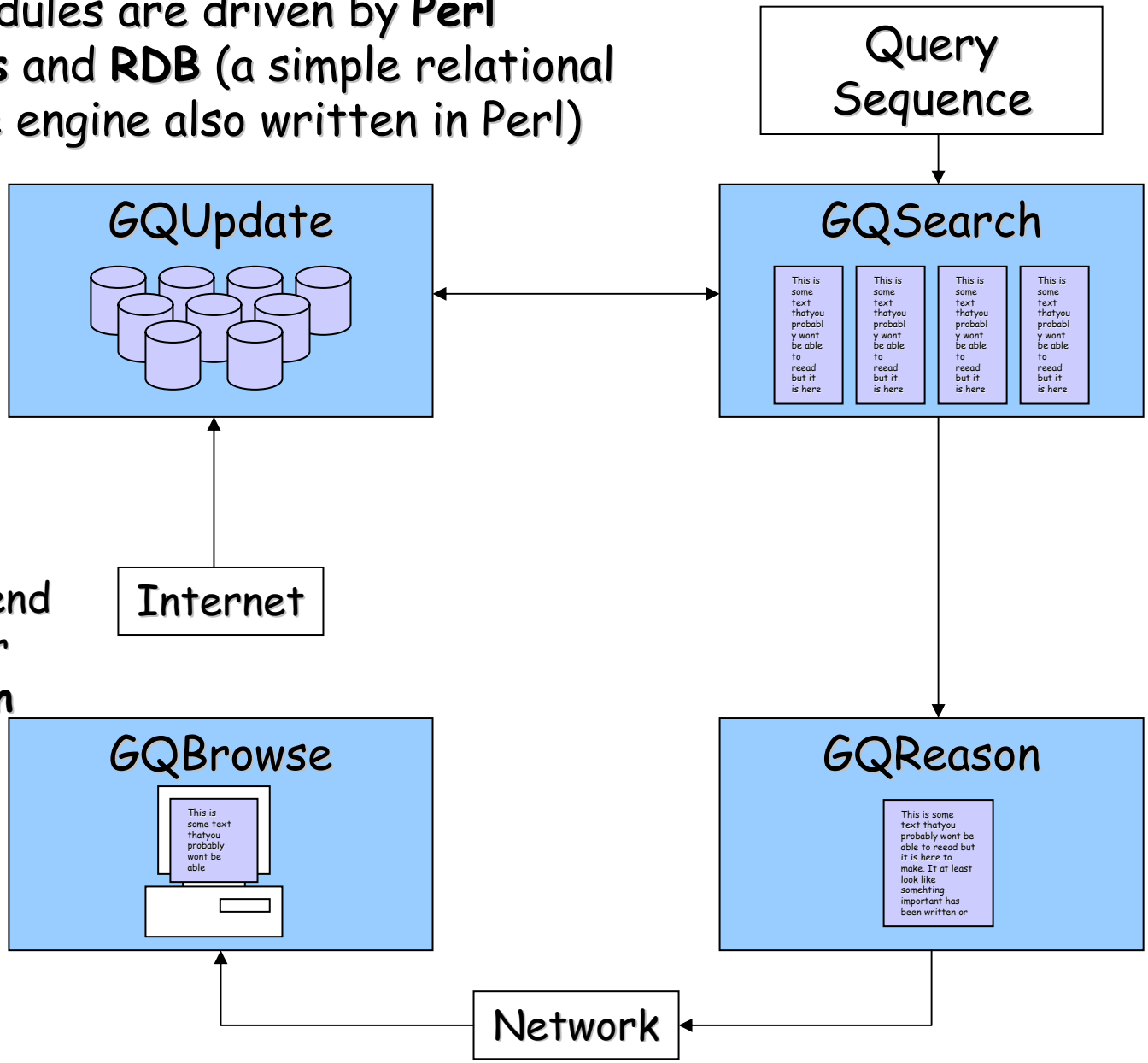
- The analysis yields **everything** that can possibly be extracted from **the current databases**, including **three-dimensional (3D) models by homology**, when the structure can be reliably calculated

Genequiz (3) – Four Modules

- GQupdate
  - The database update
- GQsearch
  - The search system
- GQreason
  - The interpretation module
- GQbrowse
  - The visualization & browsing system

- The principal design requirement - **the complete automation of all repetitive actions**
  - Database updates
  - Efficient **sequence similarity searches**
  - Sampling of results **in a uniform fashion**
  - The automated **evaluation** & **interpretation** of the results using expert knowledge coded in rules

- The modules are driven by **Perl programs** and **RDB** (a simple relational database engine also written in Perl)



The front-end program for **visualization** are WWW-browsers

Current Problems

- Re-annotation of the *Mycoplasma Pneumoniae* genome discovered **numerous errors**
- Finding **new drug targets** with **incomplete or incorrect** annotation **is very difficult**
- **Manual methods** are **time consuming** and **text abstraction** is extremely difficult in such **a varied vocabulary**
  - **Controlled vocabularies** are required
- Problems still exist when gene finding due to **errors in results** from software analysis

Future Developments

- Will fully sequenced genomes add value to function prediction?
- Knowledge of the mechanisms of **post-translational modifications** (PTM) need to be improved
- **More data** could be used to extend the function prediction **further than** the molecular level
- **Text analysis systems** to extract **keywords** and **other information** from abstracts and related articles

Ab initio Gene Discovery (1)

- **Protein coding sequences** within a whole genome sequence can be **identified** by
  - Software that recognizes **features** common to **protein coding reading frames**
    - *Esp.* **the codon bias** is typical of that observed for the **species** being studied
    - Proximity of
      - Transcriptional & translational initiation motifs,
      - 3'-polyadenylation sites,
      - Splicing consensus sequences at putative intron-exon boundaries

NCBI Taxonomy Homepage - Microsoft Internet Explorer

http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

NCBI Taxonomy Browser

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

Search for As complete name lock Go Clear

- Taxonomy browser
- Taxonomy common tree
- Taxonomy information
- Taxonomy resources
- Taxonomic advisors
- **Genetic codes**
  Translation tables
  1, 2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 14, 15, 16, 21, 22, 23
  Cited References
- Taxonomy Statistics
- Taxonomy Name/Id Status Report
- Taxonomy FTP site

The Genetic Codes

Compiled by Andrzej (Anjay) Elzanowski and Jim Ostell
National Center for Biotechnology Information (NCBI), Bethesda, Maryland, U.S.A.

Last update of the Genetic Codes: October 05, 2000

NCBI takes great care to ensure that the translation for each coding sequence (CDS) present in GenBank records is correct. Central to this effort is careful checking on the taxonomy of each record and assignment of the correct genetic code (shown as a /transl_table qualifier on the CDS in the flat files) for each organism and record. This page summarizes and references this work.

The synopsis presented below is based primarily on the reviews by Osawa et al. (1992) and Jukes and Osawa (1993). Listed in square brackets [ ] (under **Systematic Range**) are tentative assignments of a particular code based on sequence homology and/or phylogenetic relationships.

The print-form ASN.1 version of this document, which includes all the genetic codes outlined below, is also available here. Detailed information on codon usage can be found at the Codon Usage Database.

The following genetic codes are described here:

- The Standard Code
- The Vertebrate Mitochondrial Code
- The Yeast Mitochondrial Code
- The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code

Ab initio Gene Discovery (2)

- In **bacteria** & **eukaryotes** with small, compact genomes, programs such as **GeneFinder** & **Grail** have been found to predict **accurately** in excess 90% (>90%) of all true genes

- The genome of **higher eukaryotes,** tend to have **numerous introns** and extensive **non-coding intergenic regions**, **trans-splicing** and **genes located within the introns of other genes**
  - *Ab initio* gene discovery considerably **more difficult**
  - ***Programs***: *Genie*, *GenScan*, *HMMgene*, *FGENES* incorporate statistical approaches into **hidden Markov models (HMMs)**

Ab initio Gene Discovery (3)

- Irrespective of which discovery **algorithm** is used, all computationally identified putative genes **must be confirmed** by **a second line of evidence** before being elevated to gene status in the genome annotation

- Three biggest deficiencies of **computational gene discovery methods** are
  - **Imprecise** or **incomplete** characterization of **gene structure**
  - Characterization of **false positive genes**
  - Failure to identify true genes (**false negative**)

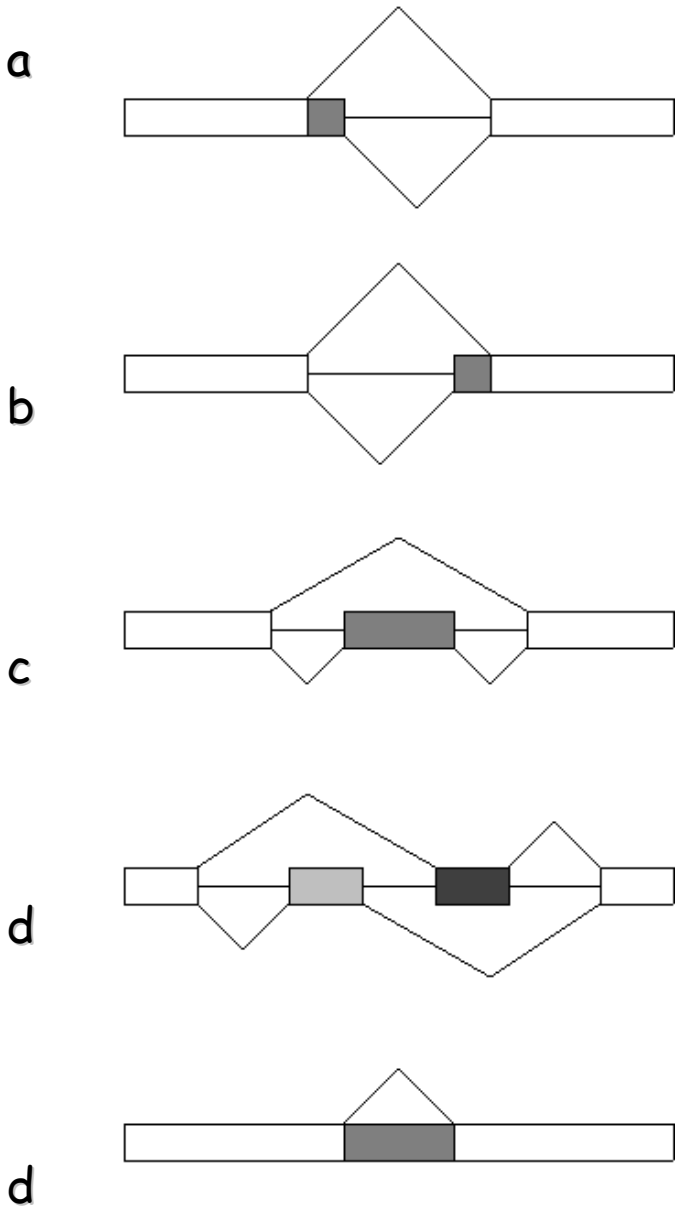
Ab initio Gene Discovery (4)

- **Improvements**
  - Sequencing of **complete cDNAs** ⇨ resolve the **complexity** of **alternative splicing in most of genes (MM, MapViewer, NCBI)**
  - Comparison of the genome sequences of **related species** ⇨ improve the annotation of **gene structure**
  - ~ 80-90% of **all true genes** are identified ⇨ < 10% false negative rate
  - **False negatives** can be detected using
    - **A different gene finding program**, or
    - By **adjusting the stringency** of the original program's search **parameters**, or even the properties of the database that is searched

Ab initio Gene Discovery (5)

- Confirmation that an annotated gene corresponds to a **true gene** ultimately depends on **functional data**
  - Functional genomics (all wet-based)
    - Gene knockout, **EST or cDNA expression,** **transcriptomics** by microarray, quantitative RT-PCR, RNAi, transgenics *etc.*

Alternative Splicing



- Every conceivable pattern of alternative splicing is found in nature
  - Exons have **multiple 5' or 3' splice sites** alternatively used (a, b).
  - Single cassette exons can reside between 2 constitutive exons such that **alternative exon** is either included or skipped (c)
  - Multiple cassette exons can reside between 2 constitutive exons such that the splicing machinery must choose between them (d)
  - Introns can be **retained in the mRNA** and become translated (d)

- Graveley, "Alternative splicing: increasing diversity in the proteomic world." *Trends in Genetics*, Feb., 2001.

The Five Standard Types Direct Evidence

- Identity to a previously **annotated reference sequence**
- A match to **one or more EST sequences** from **the same organism**
- Similarity of the **nucleotide** or **conceptually translated protein** sequence to such sequences **from other organisms** in GenBank or other databases
- **Protein structure prediction** that matches a **domain** in the Pfam database
- Association with **predicted promoter sequences**, including **a TATA box consensus sequence**, proximity to a **CpG island**

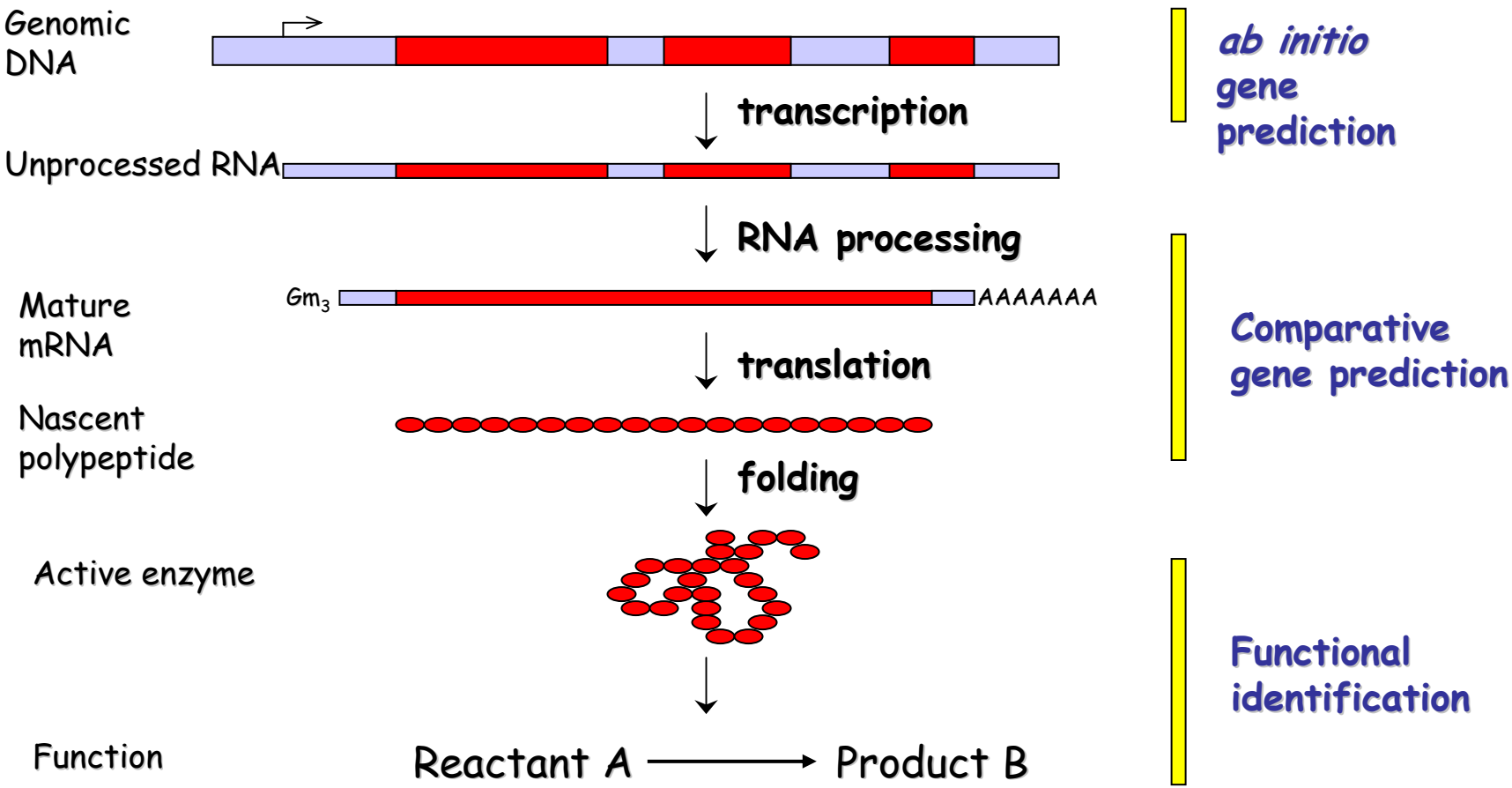
Non-Protein Coding Genes - Difficulties

- tRNA, rRNA, snoRNA, snRNA, miRNA (micro-)
- The transcripts are **not polyadenylated**, **not represented in standard cDNA libraries**
- The constraint on **sequence divergence** is at the level of **secondary structure**, rather than codon sequence ⇨ the **sequence divergence** between species is **often too great** to identify the genes purely by sequence similarity
- Relatively little is known about **the function** and **distributions** of non-protein coding RNAs (ncRNAs) **other than** those involved in transcriptional processing and translation
  - They are usually identified using **BLASTn**

Assigning Function to Genes

- Annotation process
  - Protein sequences
  - Features to look for
    - Similarity with other proteins
    - Signal peptide
    - Protein domains
    - Transmembrane domains
    - Low complexity regions
  - Other sources of data
  - Representation in database

Annotation of Eukaryotic Genomes



Levels of Annotation

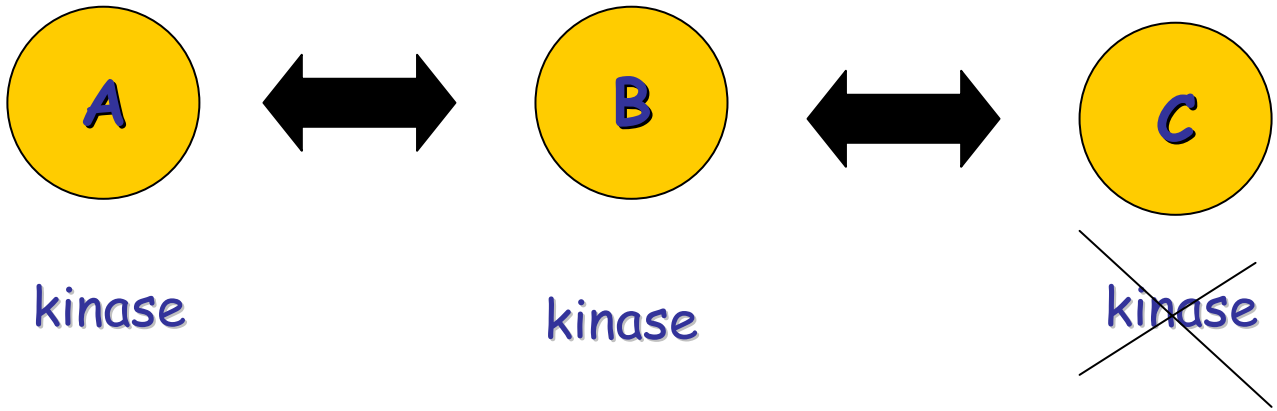
- Not: to give in depth annotation which may be highly species dependent
- To give generic information regarding the biochemical function of the protein

Similarity with "Known" Proteins

- Single most important procedure in assigning peptide function
- Orthologous proteins that share function are similar
- **BlastP** against a non-redundant protein database
  - Swiss-Prot: excellent annotation, limited coverage
  - TrEMBL: poor annotation, full coverage
- Multiple sequence alignment can be a trigger for re-prediction
- Remember that the protein sequences & annotation could be incorrect

The Danger of Inferring Function from Similarity Data

- Inferring the function of a peptide from similarity data is flawed
- Example: Protein B is similar to protein A, therefore the functional assignment of A is inherited to B
  - Protein C is similar to protein C, therefore the functional assignment of B is inherited to C
  - But protein C is not similar to protein A indicating that A & B are not orthologous & the functional assignments are invalid



Protein Features

- Peptide sequences have **features** which we can predict to aid **the functional assignment** of that peptide
- Features can be short (a few amino-acids) to very long (100's of amino acids)
- **Methods** of finding them range from the crude (regular expression), prosaic (**pairwise similarity**) to complex (**HMMs**, neural network & hierarchical knowledge based systems

Signal Peptides

- Information regarding whether a protein is to be **moved across membranes** in the cell
- Classic examples are **nuclear encoded peptides** which reside in the **mitochondria** or peptides which are to be presented on **the cell surface**
- SignalP 3.0 program

Transmembrane Domains

- Regions of the peptide which **span a membrane**
- Examples
  - The respiratory complex in mitochondria, transporters, ion-channels *etc.*
- A number of different methodologies to predict these
  - "sliding window" algorithms based on the amino-acid composition (**hydrophobic residues**)
    - TmPred
    - TMHMM (Hidden Markov Model)

Prediction Subcellular Localization

- PSORT
  - A knowledge-based system, which attempts to predict the subcellular localization for a given protein
  - Run SignalP & TMHMM

Low-complexity Regions

- **Peptide** regions which are repetitive or more formally have a low information content
  - Think of these regions as peptide repeats
  - Low-complexity regions can be indicative of some function, structural or molecular minicry/host evasion or incorrect gene prediction
  - SEG program (in BLASTs): a preliminary masking process

Secondary Protein Databases

- A number of secondary protein databases exist which are designed to collates similar (**orthologous**) proteins together
  - Pfam/SMART/PROSITE/InterPro
  - COGs (Clusters of Orthologous Genes)
- Searching against these databases give an excellent guide to assign function

Finding Protein Domains/Families (1)

- **Pfam & SMART** are collections of multiple sequence alignments for protein domains/families
  - Each family can be identified using an **HMM**
  - Each family has **limited annotation** & list of family members in the primary protein databases

- **PROSITE**
  - A collection of **regular expression** designed to identify important resides for protein (*e.g.*, active sites of enzymes)
  - Of limited use because of the searching methodology
  - Excellent **documentation**

Finding Protein Domains/Families (2)

- **InterPro**
  - The repository for all the annotation regarding the domain
  - Assigning an InterPro family to the peptide is an acceptable level of functional annotation

Clusters of Orthologous Genes (COGs)

- A database of orthologous genes from a range of completed genomes (currently this is weighted toward **microbial** genomes)
- Available from NCBI website
- Excellent coverage of the standard **metabolic enzyme families** make COGs ideal for cross-referencing between genomes

Genome Ontology (GO)

- An ontology is a restricted **vocabulary** used to describe/classify
- The GO consortium grew from efforts in the fly community to assign function/localization/process information to the biology of the fruitfly
- The GO consortium currently has representatives of **all the major model organisms** and provides a methodology of **comparing genomes** based on functional/biological terms
- There are several levels of assigning GO terms that **lowest** of which is by **similarity**
- To this end the consortium has prepared a list of InterPro to **GO mappings** from which GO terms can be added to the protein based on it's InterPro matches

Functional Assignments

- In addition to the searches mentioned thus far the annotator also read **the literature** regarding the gene under investigation
  - This comes from **cross-references** in the DNA/protein database entry
- Annotation should take the form of a one-line description of the proteins function (if assignable) with the ability for a user to search the database for the protein feature information
- More in-depth annotations should be stored in an atomic fashion (**single fact with evidence**) to aid searching/cross-referencing in a database

Utility of Functional Assignments

- Central to data mining & interpretation of **other experiments**
  - Example: a one line description for each gene in expression microarray work
- Generalized **multi-species attempts** to catalogue the proteomes of model organisms (*e.g.*, GO and COGs)

Functional Annotation & Gene Family Clusters (1)

- **First-pass classification**
  - **Protein-similarity** searches, BLASTp (or BLASTx) to screen for **amino acid sequence matches** in protein databases
    - 1/3~1/2 of all of **the predicted proteins** do not match **a protein** for which any functional data is available ⇨ **"unknown function"** or "**orphans**"
    - Others: **a finite number** of structural protein **domains** in the combined proteome of **all organisms**
      - To cluster all proteins into **gene families**

Functional Annotation & Gene Family Clusters (2) – Clustering of Gene by Sequence Similarity (1)

- **A common result**
  - Each query sequence matches **multiple proteins** from **one or more species**
  - **Reasons**
    - One **domain** in the query is present **in a family of proteins**
    - **Multiple domains** match different proteins

Functional Annotation & Gene Family Clusters (2) – Clustering of Gene by Sequence Similarity (2)

- When comparing **closely related species**, a query sequence containing **multiple domains** will typically identify **a protein** or **proteins** with the same domain structure

- Biological logic to **domain clustering**
  - **Multiple domain matches** tend to classify the gene product **in the same board category**
    - *E.g.*, transcriptional factors or receptors

NCBI CD-Search - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

網址(D) http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

NCBI Conserved Domain Search

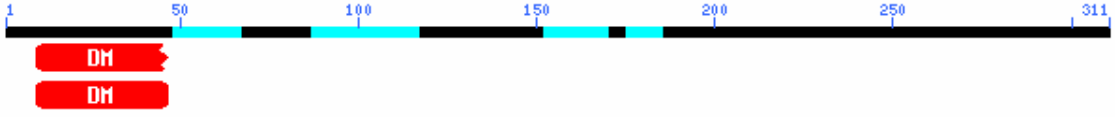
New Search | PubMed | Nucleotide | Protein | Structure | CDD | Taxonomy | Help?

RPS-BLAST 2.2.6 [Apr-09-2003]

Query= local sequence: doublesex and mab-3 related transcription factor 1 [Gallus gallus]
(311 letters)

Database: cdd.v2.00
11,382 PSSMs; 2,824,437 total columns

Click on boxes for multiple alignments



Show Domain Relatives

.. This CD alignment includes 3D structure. To display structure, download Cn3D!

PSSMs producing significant alignments:		Score (bits)	E value
gnl|CDD|24245	smart00301, DM, Doublesex DNA-binding motif;	63.1	4e-11
gnl|CDD|1307	pfam00751, DM, DM DNA binding domain. The DM domain is named a...	61.9	1e-10

完成 網際網路

InterPro Detailed matches for protein

Protein matches for protein STN1_HUMAN(P16949) from the the UniProt/Swiss-Prot database.

One line is shown per method for the protein. The vertical line are drawn at 10aa intervals.

- Go to the UniProt/Swiss-Prot entry for this protein.
- View the GOA annotation for this protein.
- View the matches in a table

Interpro Entry	Method accession	Graphical match [?]	Method name
IPR000956:	PF00836		Stathmin
IPR000956:	PR00345		STATHMIN
IPR000956:	PS00563		STATHMIN_1
IPR000956:	PS01041		STATHMIN_2

Key:		
Database	True match	False/Uncertain match

Functional Annotation & Gene Family Clusters (3) – Clustering of Gene by Sequence Similarity

- **Two different DNA-binding domains** may be combined on the same protein
  - *E.g.*, transmembrane region may be linked to a range of different intracellular and extracellular domains

- **Databases**: to classify **protein domains** according to criteria agreed upon by groups of experts
  - Enzyme Commission (EC) hierarchical classification of enzymes
  - Non-enzymatic proteins ⇨ not so obvious
  - Classification of **bacterial proteins** has attained wide acceptance

Functional Annotation & Gene Family Clusters (4) – Clustering of Gene by Sequence Similarity

- Structural biology
  - PFAM/the Sanger Centre
    - A large collection of **multiple sequence alignments** & hidden Markov models (HMMs) covering many **common protein domains** & **families**
  - InterPro
    - To classifying **individual protein domains**
    - Gene annotations now typically link directly to InterPro classifications
  - *Etc.*
- Classification of genes as **members** of the same family
  - Paralogs
  - Orthologs (homologues or homologs)

The "top ten" InterPro families in *H. sapiens*: numbers of genes per family in *H. sapiens* compared to other eukaryotes

(H=human, F=fly, W=worm, Y=yeast, At=*Arabidopsis*)

IHGSC (2001) *Nature* **409**: 860-921 Table 25 (modified)

	H	F	W	Y	At
1. Immunoglobulin domain	765	140	64	0	0
2. C2H2 zinc finger	706	357	151	48	115
3. Euk. Protein kinase	575	319	437	121	1049
4. Rhodopsin-like GPCR	569	97	358	0	16
5. P-loop motif	433	198	183	97	331
6. Reverse transcriptase	350	10	50	6	80
7. rrm domain	300	157	96	54	255
8. G-protein b WD-40	277	162	102	91	210
9. Ankryn repeats	276	105	107	19	120
10. Homeobox domain	267	148	109	9	118

Functional Annotation & Gene Family Clusters (5) – Clustering of Gene by Sequence Similarity

- **Protein function prediction**
  - Major classes of proteins
    - Enzyme
    - Signal transduction
      - Receptors and kinases
    - Nucleic acid binding
      - Transcriptional factors & nucleic acid enzymes
    - Structural
      - Cytoskeletal, extracellular matrix, motor protein
    - Channel
      - Voltage & chemically gated
    - **Others**
      - Immunoglobulins, calcium-binding proteins, transporters…

Functional Annotation & Gene Family Clusters (6) – Clusters of Orthologous Genes

- **PSI-BLAST (iterative)**
  - To align the sequences obtained in an initial protein database search and use this to construct **a profile**, which is then used to initiate a fresh search ⇨ until **no further matches are identified**
  - A **true family** of genes ought to be bounded by a significance cut-off
    - Gene family

- **COG = clusters of orthologous genes** (Tatusov *et al.* 2001)
  - Identification of **the best hit** for **each gene** in complete **pairwise comparisons** of a set of genomes

Functional Annotation & Gene Family Clusters (7) – Clusters of Orthologous Genes

- COG (http://www.ncbi.nlm.nih.gov/COG)
  - *E.g.*, comparison of the 45,350 proteins encoded by the genomes of **30 microbes** ⇨ 2,791 COGs
  - **Triad of proteins**
    - COGs are assembled by **merging triads of proteins** from **different species** that are **the best match to one another**
  - Both **orthologs** and **paralogs**
    - **Paralog**= a duplicate copy of a gene that arose subsequent to the split between two lineages that are being compared
    - **Ortholog**= a gene in another lineage that is derived from the same ancestral gene that was present prior to the lineage split

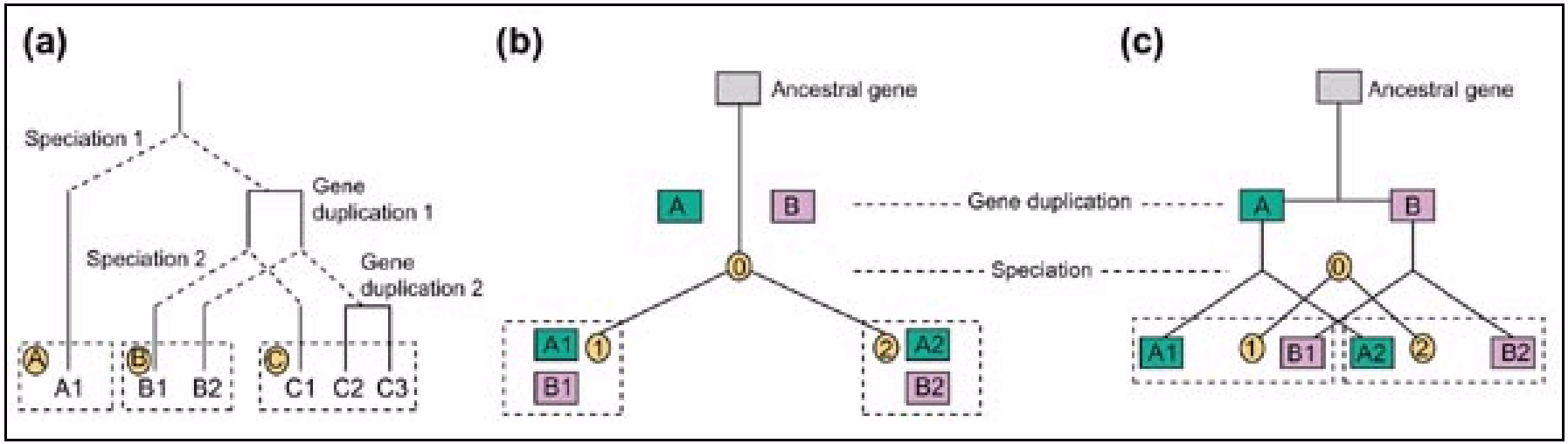
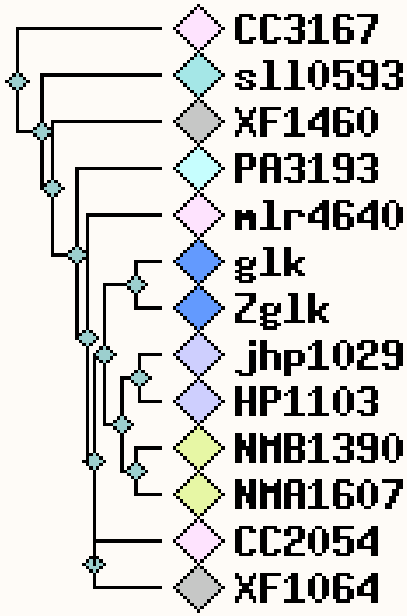
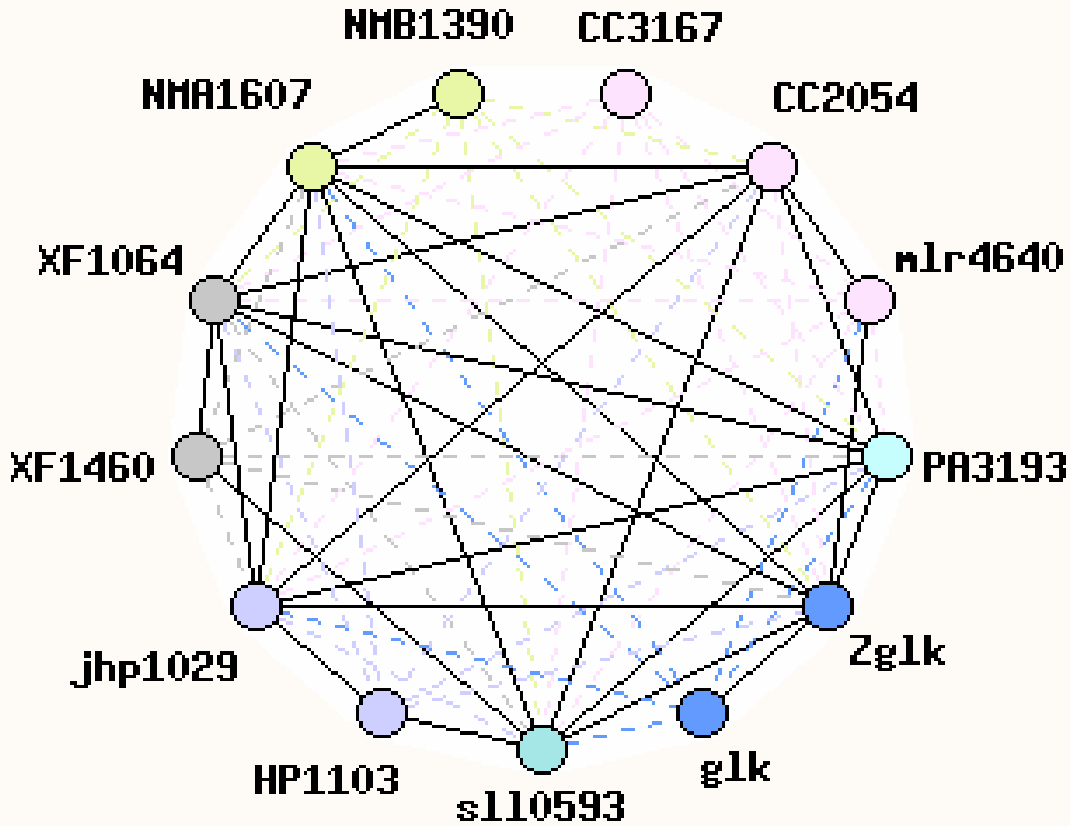


Figure 1.

(a) Simplified diagram of homology subtypes (showing orthologs and paralogs, but not xenologs); adapted from [4]. Speciation events produce the species A, B and C. The genes A1, B1, B2, C1, C2, and C3 have descended from the ancestral gene following evolutionary events of speciation and gene duplication. **(b,c)** Evolutionary descent of an ancestral gene to paralogs and orthologs following gene duplication in species 0, and then speciation to yield species 1 and 2. Diagram (b) shows the resulting relationship between paralogs and orthologs as illustrated by Koonin in his comment [1]. Diagram (c) is my version of Koonin's diagram using a Fitch diagram for visualization. Note that the two evolutionary events depicted are a subset of the four shown in (a) (gene duplication 1 and speciation 2), and that the use of capital letters for genes and numbers for species is the opposite of that used in (a).

COG0837: Glucokinase





- **Dashed lines**: the best match occurs only in **one direction**; solid lines: each gene is the match in the genome of the other
- The tree at the left: a more standard representation of **the relationship among the genes**

Gene Ontology (GO) (1)

- Annotation on the basis of **molecular function** alone is **insufficient** to describe or predict **biological function**

- Examples
  - The evolution of **a novel function** is the **reuse of enzymes** *e.g.*, Lactate dehydrogenase (LDH) as lens crystallins
    - Occurred multiple times in vertebrate lineages (an **enzyme** is converted into a **structural protein**)

  - The *Drosophila* ortholog of the **mammalian dioxin receptor** is encoded by the *spineless-aristapedia* gene (one of the key regulatory genes that control **antenna differentiation**)
    - Only clear **functional similarity** = the involvement of the protein product in **the olfactory system**
    - Knowledge from one species is not directly transferable to another

Gene Ontology (GO) (2)

- **Molecular function** has remained the same but the **physiological function** has evolved
  - *E.g*, HOX genes in patterning both the cranial（頭蓋）nerves of **vertebrates** & the appendages（附屬肢體） of **invertebrates** – a shared common function in patterning along the body axis
    - **One or more biological functions** are clearly derivative, the use of **Toll-dorsal pathway** in **the embryonic patterning** of flies as apposed to **innate immunity** in both **vertebrates** and **invertebrates**
- 1,500 genes have unambiguous orthologs among the fly, worm & human genomes
  - A **single** closest match in each genome

Gene Ontology (GO) (3)

- Gene Ontology (GO) consortium (Ashburner *et al.* 2000)
- To annotate all these features for the complete set of genes identified by genome projects
  - http://www.geneontology.org
- Three major categories
  - **Biological process**
    - The **nature of process** that is regulated or affected
    - Behavior, cell communication, cell growth & maintenance, death, developmental processes, perception of external stimulus, physiological processes, viral life cycle
      - *E.g.*, cell growth and division, respiration, signal transduction, cAMP biosynthesis

Gene Ontology (GO) (4)

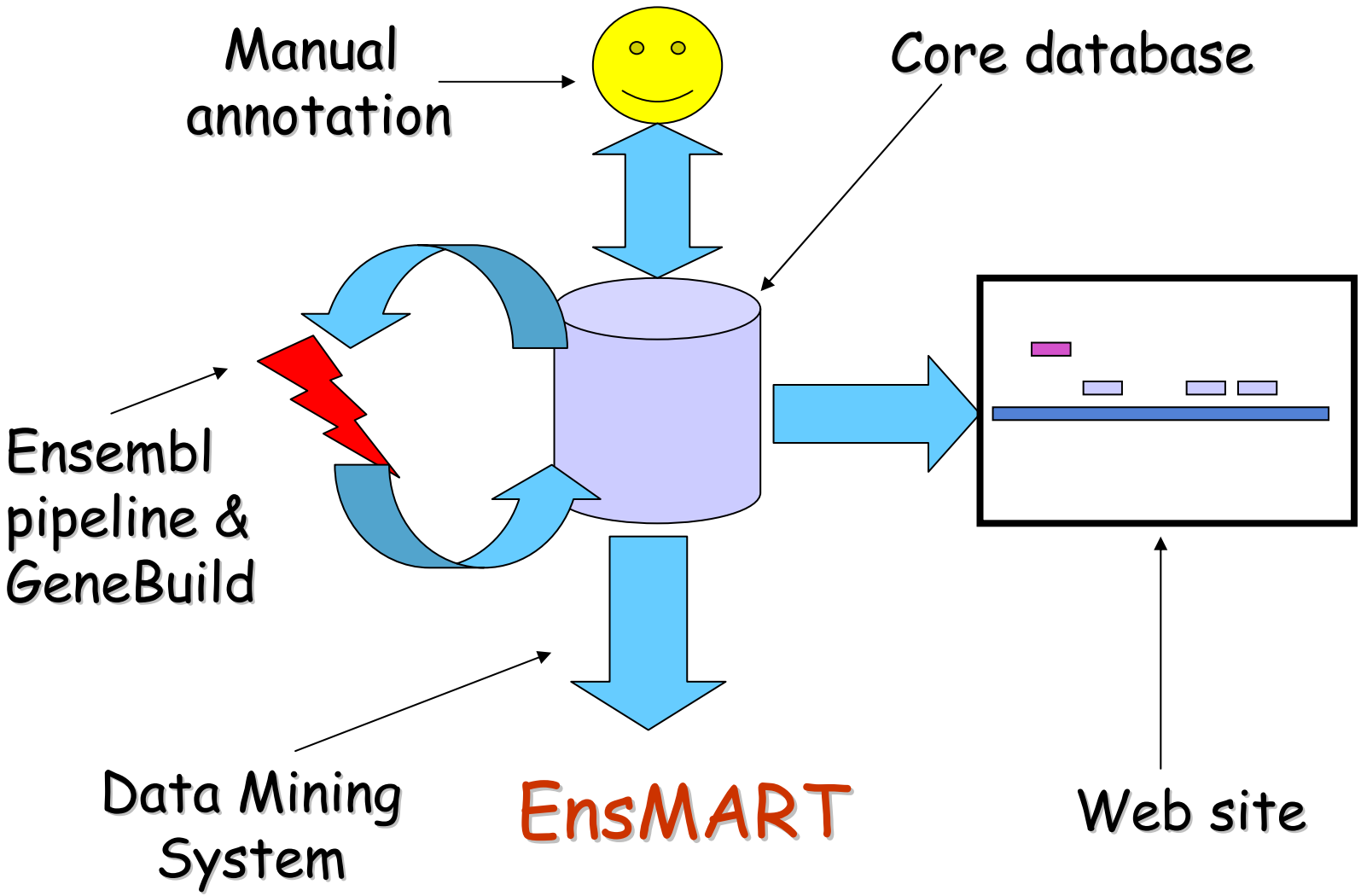
- **Molecular function**
  - The **biochemical activities**
  - Antitoxin, anticoagulant, antioxidant, apoptosis regulator, cell cycle regulatory, defense/immunity protein, vitamin transporter…
    - *E.g.*, enzymes, nucleic acid binding, DNA helicase, or tyrosine kinase

- **Cellular component**
  - The place in a cell where **the gene product is active**
  - Cell fraction, cell wall, extracellular, intracellular, membrane, unlocalized
    - *E.g.*, the cell surface, Golgi apparatus, or splicesome

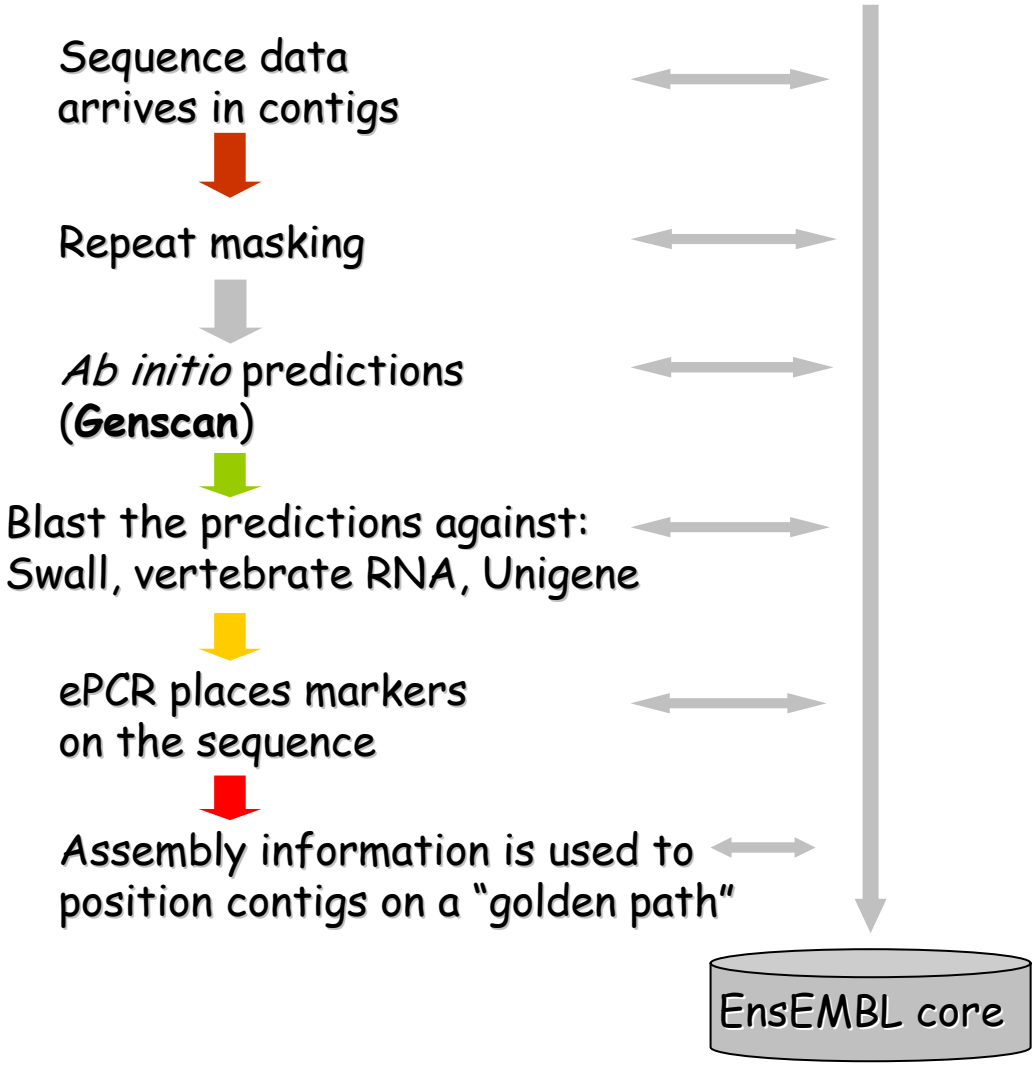
Gene Ontology (GO) (5)

- Each of these terms are arranged hierarchically, **from general to specific function**
  - Dynamic, continually updated
- Each organism portray different attributes in their annotation
  - Wormbase, Flybase…

Automatic Annotation - Ensemble



Automatic Annotation - Raw Computes



Database Search Service

http://www.sanger.ac.uk/DataSearch/databases.shtml

Ensembl Database Resources

Database Resources

Pfam:
Protein Families Database of Alignments and HMMs

Rfam:
RNA families database of alignments and CMs

Wormbase:
Caenorhabditis elegans genome database browser

GeneDB:
Sanger Institute Pathogen Sequencing Unit

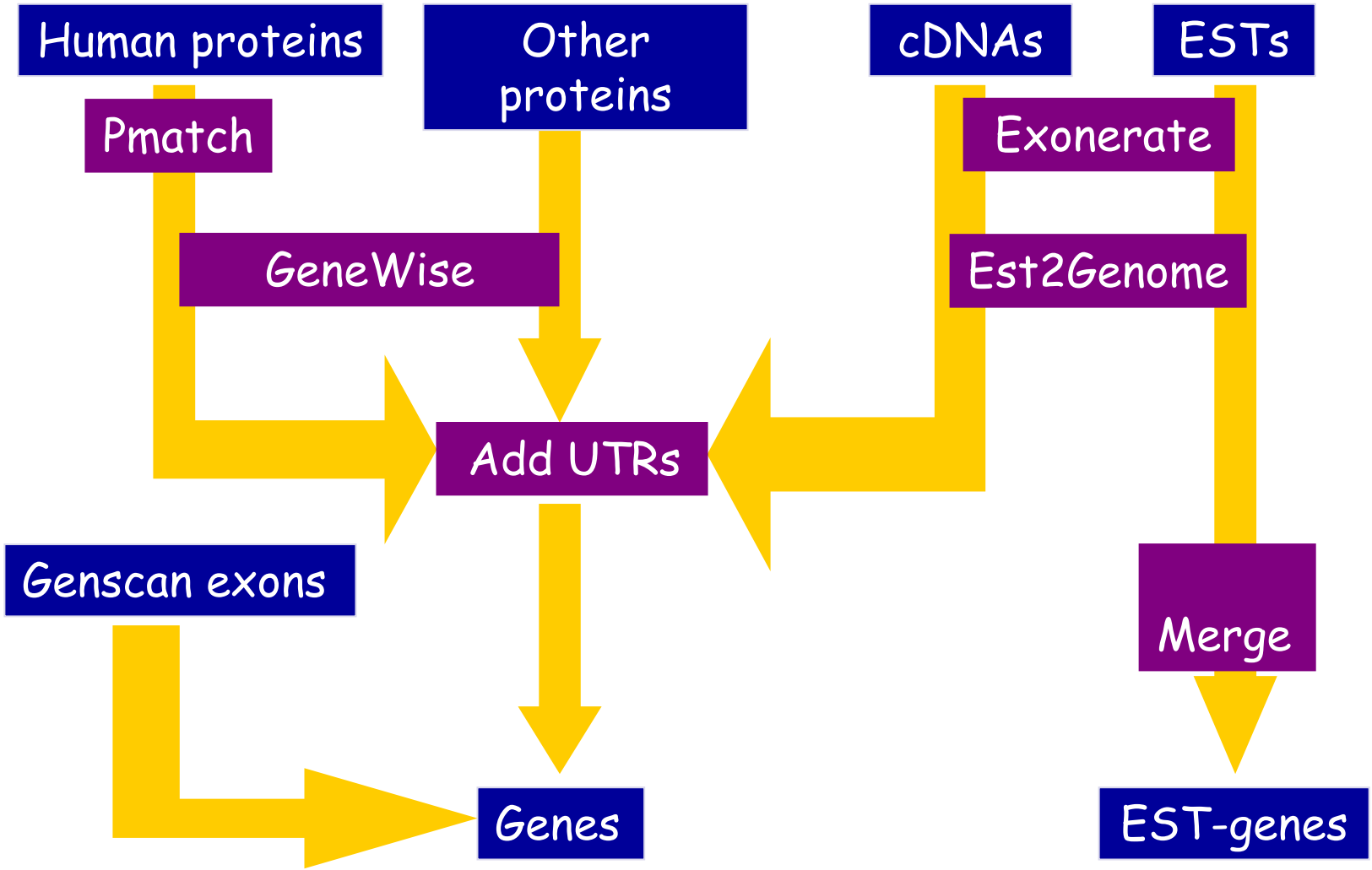
Vega:
Vertebrate Genome Annotation

SRS:
Sequence Retrieval System

AceDB:
A genome database designed specifically for handling bioinformatic data flexibly

MEROPS:
Provides a catalogue and structure-based classification of peptidases

Automatic Annotation - GeneBuild



GeneWise – by Ewan Birney

Protein Sequences

⇩

Aligned to the Genome

⇩

Blast ⇨ MiniSeqs

⇩

GeneWise

ESTs and cDNA

Map cDNAs and ESTs using Exonerate
(determine coverage, % identity and location in genome)

↓

Store hits and filter on percentage identity and length coverage

↓

Blast sequence and create a miniseq

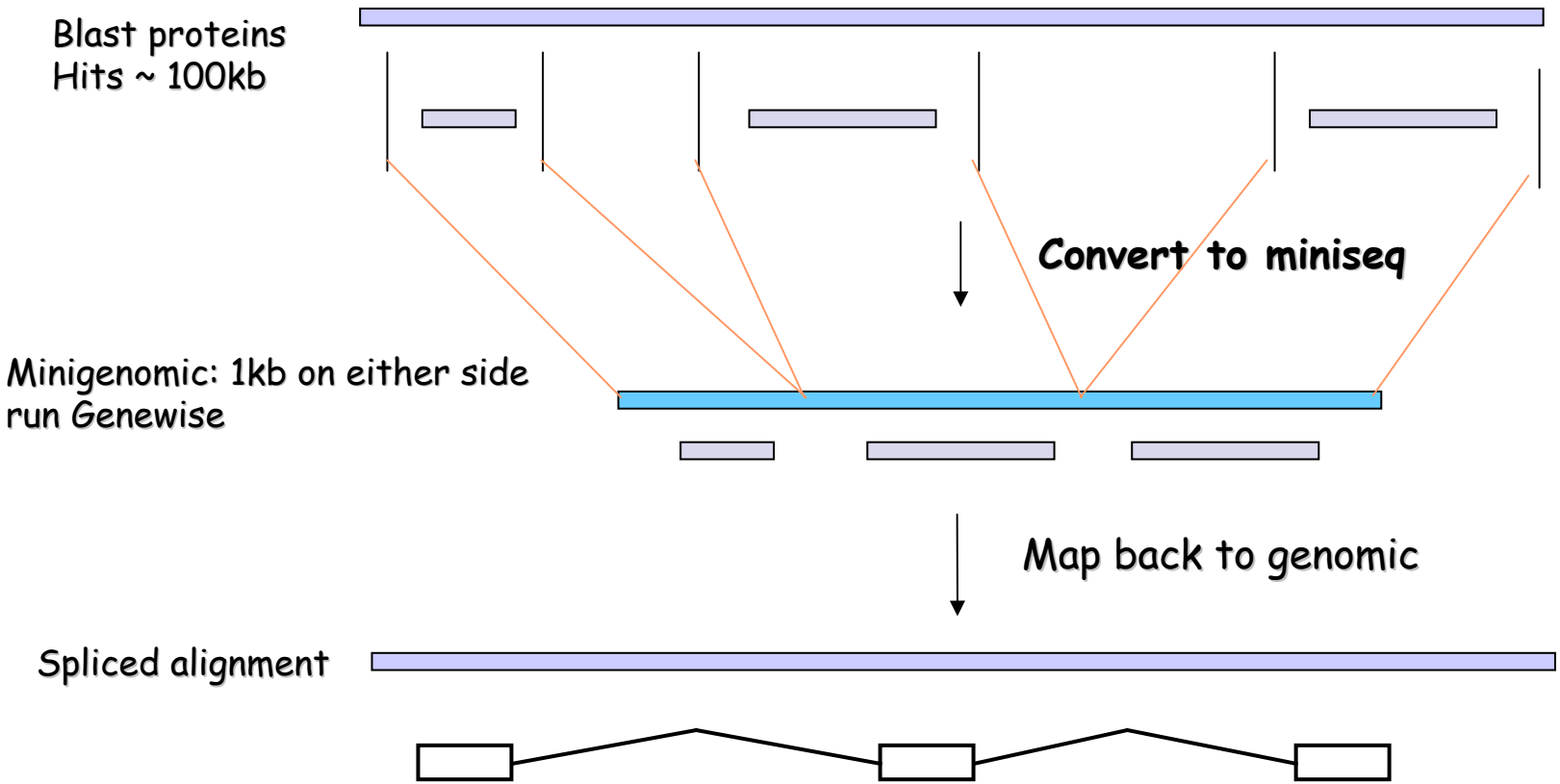
↓

Run est2genome on miniseq
(determine strand, splicing)

↓

Map transcripts back into genome-assembly

Miniseq - the Need for Speed



Resources

- 8x ES40 Alpha (667 MHz) with 2Tb fibre channel storage
- 6x ES45 Alpha (1GZ) with 4Tb fibre channel storage
- 360x DS10L (467 MHz) farm with 60Gb local disk storage
- 767xRLX800i with 80Gb of local disk storage
- Further 21Tb storage on farm
- Tru64 UNIX (avoids the 2Gb file limit)
- 7 MySQL (v. 3) instances
- Most binaries and all sequence databases stored locally (avoids using NFS)

Latest full Human Build - NCBI

- NCBI 34 build, version 1, released Dec. 2003
  - http://www.ncbi.nlm.nih.gov/mapview/stats/BuildStats.cgi?taxid=9606&build=34&ver=1#GBSourceSeq
  - HTG Phase 3
    - Sequence number used: 28,479
    - Length: 3.42288×10^9
    - Length of sequences assembled: 3,020,300,000

Statistics: Genome Annotation - Microsoft Internet Explorer

http://www.ncbi.nlm.nih.gov/mapview/stats/BuildStats.cgi?taxid=9606&build=34&ver=1#GBSourceSeq

Genes, Exons and Introns

Assembly: *reference*

Statistic	Evidence	count	min	max	avg
Exons	EST	6,378	2 bp	11,751 bp	290.65 bp
	Predicted	11,930	4 bp	8,415 bp	256.86 bp
	mRNA	698,192	1 bp	67,529 bp	225.07 bp
Introns	EST	4,267	2 bp	467,279 bp	5,236.63 bp
	Predicted	8,962	51 bp	497,817 bp	5,240.88 bp
	mRNA	355,342	2 bp	494,709 bp	5,320.31 bp
Transcripts	EST	1,169	347 bp	578,575 bp	24,512 bp
	Predicted	2,968	269 bp	531,390 bp	16,853.5 bp
	mRNA	41,254	104 bp	103,468,243 bp	75,670.89 bp
Transcripts per Gene	EST	-	1	2	1
	Predicted	-	1	2	1
	mRNA	-	1	23	1
One-exon Genes	EST	88	-	-	-
	Predicted	910	-	-	-
	mRNA	2,552	-	-	-
One-exon Transcripts	EST	88	-	-	-
	Predicted	912	-	-	-

http://www.ncbi.nlm.nih.gov/mapview/stats/BuildStats.cgi?taxid=9606&build=34&ver=1#GBSourceSeq

Human (Homo sapiens) - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

上一頁 搜尋 我的最愛

http://www.ensembl.org/Homo_sapiens/index.html

網址(D) http://www.ensembl.org/Homo_sapiens/index.html 移至 連結

Google ensembl 搜尋 PageRank 17 已攔截 檢查 選項 ensembl

- Homo sapiens
- Macaca mulatta
- Monodelphis domestica
- Mus musculus
- Pan troglodytes
- Rattus norvegicus

Other chordates

- Ciona intestinalis
- Danio rerio
- Takifugu rubripes
- Gallus gallus
- Tetraodon nigroviridis
- Xenopus tropicalis

Other eukaryotes

- Anopheles gambiae
- Apis mellifera
- Caenorhabditis elegans
- Drosophila melanogaster
- Saccharomyces cerevisiae

Other links

- Home
- Sitemap
- Vega
- Pre Ensembl
- View previous release of page in Archive!
- v36 Dec 2005
- v35 Nov 2005

Jump directly to sequence position

Chromosome: — or region

From (bp):

To (bp): Go

What's New in Ensembl 37

Homo sapiens News

- **cDNA updates**
  The Ensembl Human and Mouse cDNA database have received their regular updates.
- **Xrefs**
  New Xrefs have been run for Ensembl Human and Mouse.
- **Projected GO terms**
  GO terms have been projected across species as follows:
  - Human to Mouse, Rat and Dog
  - Mouse to Human, Rat and Dog
- **Projected gene names**
  Gene names have been projected across orthologues in the form.
  Read more...
- **Xref updates**
  Xrefs have been re-run for Human and Rat in this release. Mouse updates are scheduled for Release 38.

More news...

- More information about the CCDS project.

The ENCODE (ENCyclopedia Of DNA Elements) project aims to find functional elements in the human genome.

- More information about the ENCODE resources at Ensembl.

Statistics

Assembly:	NCBI 35, July 2004
Genebuild:	Ensembl, Nov 2004
Database version:	37.35j
Known genes:	20,134
Novel genes:	1,878
Pseudogenes:	1,947
RNA genes:	1,015
Genscan gene predictions:	68,101
Gene exons:	245,231
Gene transcripts:	35,845
Base Pairs:	3,272,187,692
Golden Path Length:	2,863,476,365
Most common InterPro domains:	Top 40 Top 500

*See the MapView help page for more information on gene counts

網際網路

Santa Cruz Genome Browser (1)

- http://www.genome.ucsc.edu

- Species: human, mouse, rat, *C. elegans*, *C. briggsae*, SARs
  - Most recent version of the mouse, rat and human genome data + several earlier assemblies (better annotations)

- **Search protocol**
  - Select the appropriate organism from the pull-down menu, *e.g.*, human chr22(:1-49396972)
    - Select the version of the human assembly to view
      - *E.g.*, May, 2004: based on an assembly of the human genome done by UCSC using sequence data available on that date

  - *E.g.*, position = *TBX1* ⇨ jump

Human TBX1 - UCSC Genome Browser v129 - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

上一頁 搜尋 我的最愛

網址(D) http://www.genome.ucsc.edu/cgi-bin/hgTracks?clade=vertebrate&org=Human&db=hg17&position=TBX1&pix=620&hgsid=68930088&hgt.dummyEnterButton.x=0&hgt.dum 移至 連結

Google 搜尋 PageRank 17 已攔截 檢查 選項

Known Genes

TBX1 (NM_080647) at chr22:18118779-18129409 - T-box 1 isoform C
TBX1 (NM_080646) at chr22:18118779-18141620 - T-box 1 isoform A
TBX1 (NM_005992) at chr22:18118779-18145664 - T-box 1 isoform B
TBX15 (NM_152380) at chr1:119137707-119244221 - T-box 15
TBX10 (NM_005995) at chr11:67155350-67163607 - T-box 10
TBX19 (NM_005149) at chr1:164981935-165015320 - T-box 19

RefSeq Genes

TBX1 at chr22:18118779-18145664 - (NM_005992) T-box 1 isoform B
TBX1 at chr22:18118779-18129409 - (NM_080647) T-box 1 isoform C
TBX1 at chr22:18118779-18141620 - (NM_080646) T-box 1 isoform A
TBX10 at chr11:67155350-67163607 - (NM_005995) T-box 10
TBX15 at chr1:119137707-119244221 - (NM_152380) T-box 15
TBX19 at chr1:164981935-165015320 - (NM_005149) T-box 19

Non-Human RefSeq Genes

Tbx1 at chr22:18123286-18127193 - (NM_001032450) T-box transcription factor
Tbx1 at chr11:67156991-67159146 - (NM_001032450) T-box transcription factor
tbx1 at chr22:18122920-18128803 - (NM_183339) T-box 1
tbx1 at chr22:18122920-18128803 - (NM_001035118) T-box transcription factor 1
Tbx1 at chr22:18122918-18129407 - (NM_011532) T-box 1
Tbx10 at chr11:67155750-67159313 - (NM_001001320) T-box 10
tbx15 at chr1:119139398-119242460 - (NM_153664) T-box 15
Tbx15 at chr1:119139071-119242444 - (NM_011534) T-box 15 isoform 2
Tbx15 at chr1:119137711-119242470 - (NM_009323) T-box 15 isoform 1
tbx18 at chr6:85503205-85530579 - (NM_153665) T-box 18
tbx18 at chr1:119139561-119242460 - (NM_153665) T-box 18
TBX18 at chr6:85503103-85530618 - (NM_204453) T-box 18
Tbx18 at chr6:85502304-85531030 - (NM_023814) T-box18
TBX19 at chr1:164982001-165009771 - (NM_204950) T-box 19

網際網路

Santa Cruz Genome Browser (2)

- RefSeq Genes
  - Known genes shows the mapping of the NCBI Reference mRNA sequences to the genome
  - NM_080647
- Human Aligned mRNA Search
  - The mRNA associated search results = the mapping of other GenBank mRNA sequences to the genome
- NonHuman Aligned mRNA Search

Human chr22:18,118,779-18,141,620 - UCSC Genome Browser v129 - Microsoft Internet Explorer

檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)

上一頁 搜尋 我的最愛

網址(D) http://www.genome.ucsc.edu/cgi-bin/hgTracks?position=chr22:18118779-18141620&hgsid=68930088&knownGene=pack&hgFind.matches=NM_080646,

移至 連結

Google 搜尋 PageRank 17 已攔截 檢查 選項

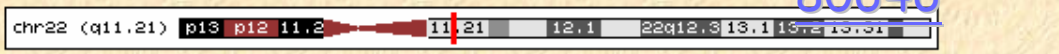
http://www.genome.ucsc.edu/cgi-bin/hgTracks?position=chr22:18118779-18141620&hgsid=68930088&knownGene=pack&hgFind.matches=NM_080646

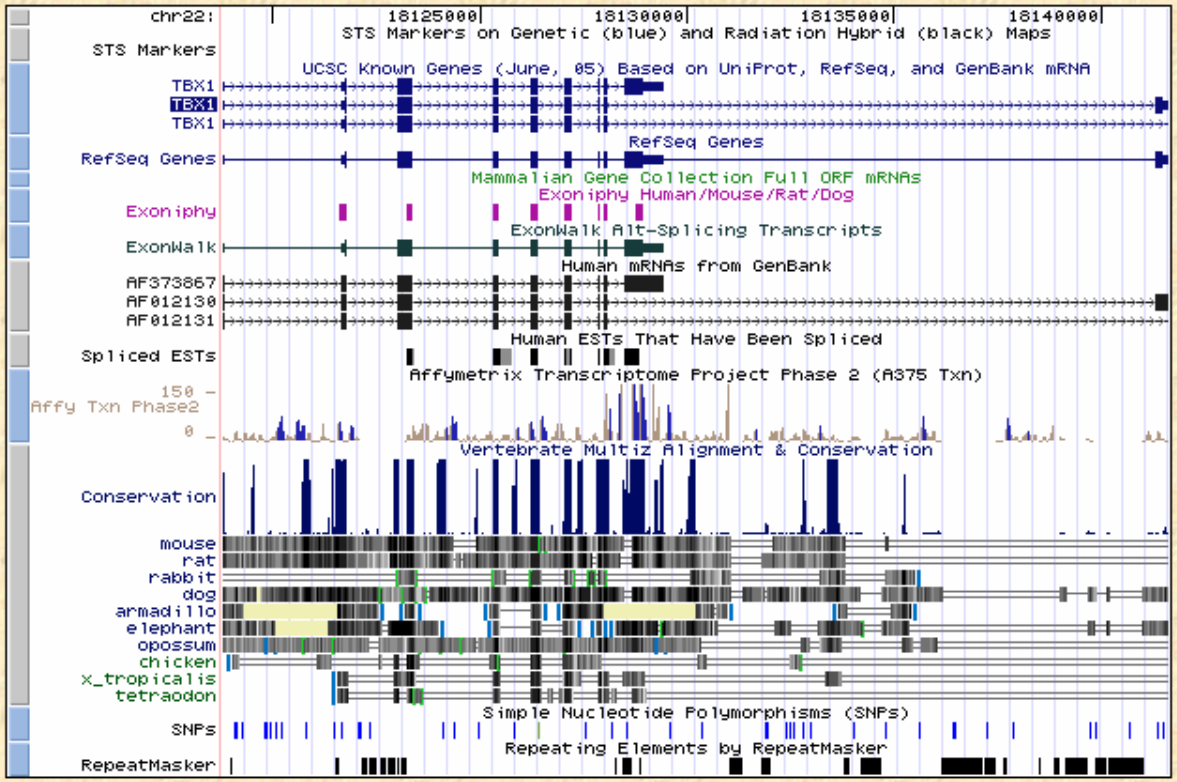
Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Help

UCSC Genome Browser on Human May 2004 Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chr22:18,118,779-18,141,620 jump clear size 22,842 bp. configure





move start Click on a feature for details. Click on base position to zoom in move end

網際網路

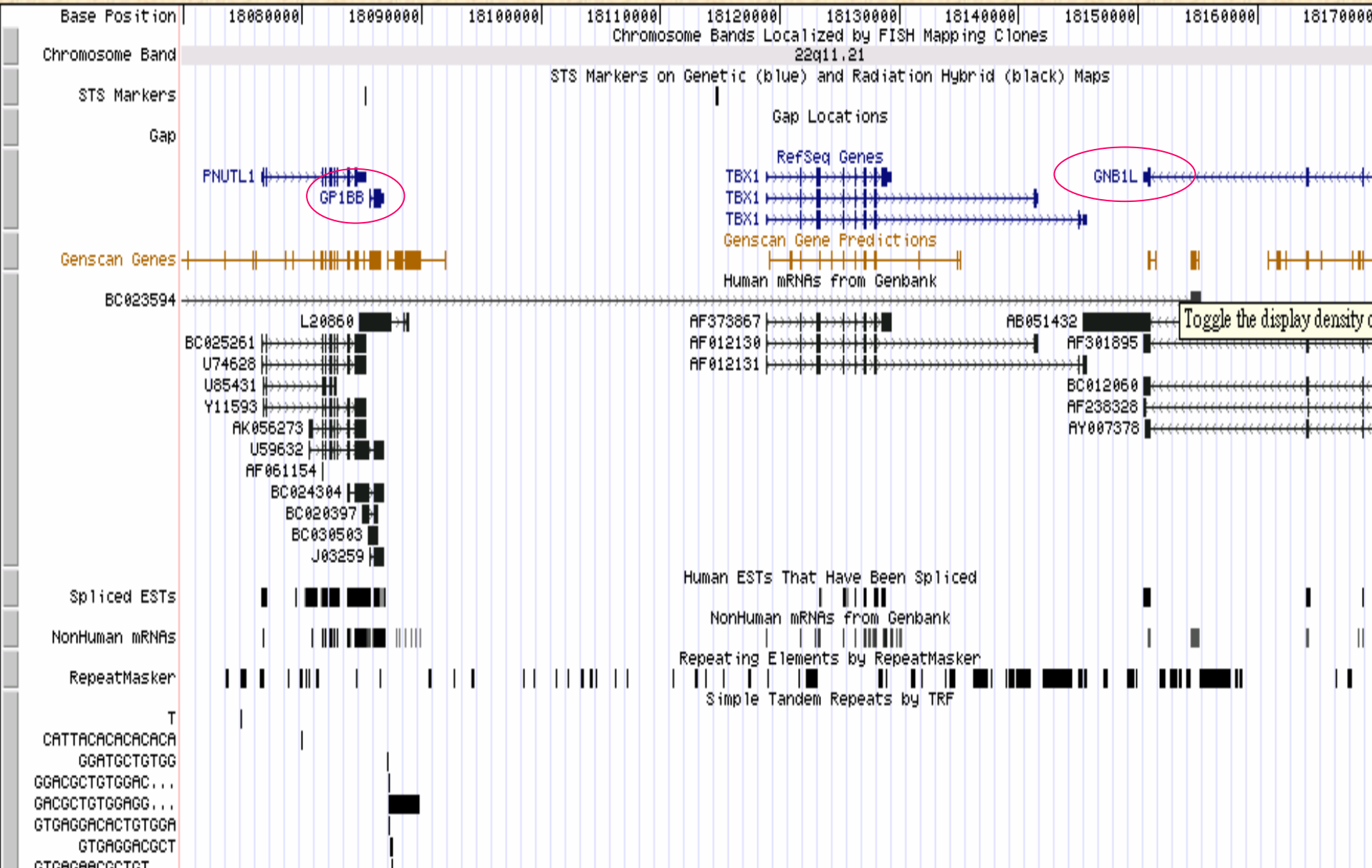
Santa Cruz Genome Browser (3)

- Blue track
  - RefSeq: intron-exon structure
    - Vertical boxes = exons
    - Horizontal lines = introns
    - Three isoforms
  - Direction of transcription
    - Arrowheads on the intron
- Genescan prediction
- Alignments for other database nucleotide sequences
  - Human mRNA from GenBank, spliced EST, UniGene & Nonhuman mRNA
- Tracks displaying single-nucleotide polymorphisms (SNPs), repetitive elements and microarray database are shown at the bottom
- To view the genomic context of TBX1, zoom out 10x by clicking on the zoom out 10x box
  - TBX1 is located between GP1BB and GNB1L

UCSC Genome Browser on Human July 2003 Freeze

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position chr22:18069850-18178579 size 108,730 image width 1000 jump



References

- Swiss-Prot and TrEMBL
  - http://www.expasy.ch/swissprot
- GeneQuiz
  - http://columba.ebi.ac.uk:8765/extgenequiz//genequiz.html
- PIPMaker Information
  - http://nog.cse.psu.edu/pipmaker/
- Gene Ontology Consortium
  - http://www.geneontology.org/