DO ORGANISMO …. À SEQUENCIA GENOMICA

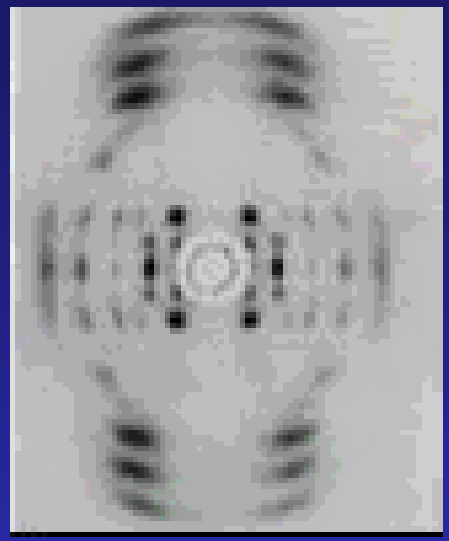
Slides por: Wim Degrave
Leila de Mendonça Lima
Antonio B. de Miranda

Departamente de Bioquímica e
Biologia Molecular
Instituto Oswaldo Cruz - Fiocruz
Rio de Janeiro, Brasil
wdegrave@fiocruz.br
http://www.dbbm.fiocruz.br

Galacidalacidesoxiribunucleicacid
Salvador Dali

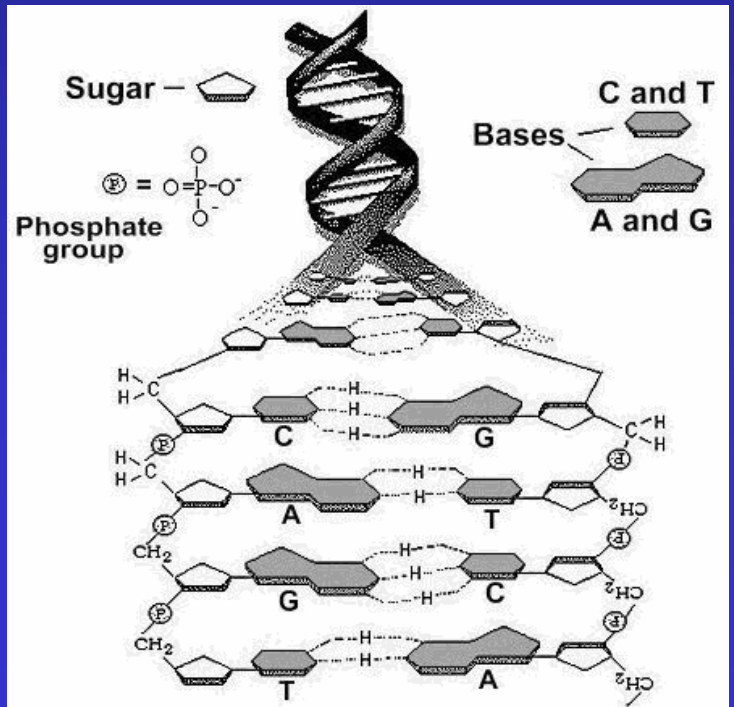
"O objetivo da Biologia Molecular é encontrar, na estrutura de macromoléculas, interpretações para os fundamentos da vida"

Jacques Monod




50th
Anniversary
DNA Double Helix

1953
2003


Sugar
C and T
Bases
A and G
Phosphate group
C
G
A
T
G
C
T
A

Depois da dupla hélice

BIOTECNOLOGIA

PROJETOS GENOMA

Sequenciamento

Clonagem

DNA

Genômica

RNA

TERAPIA GÊNICA

DESENVOLVIMENTO RACIONAL DE DROGAS

PCR

PROTEÍNA

Proteômica

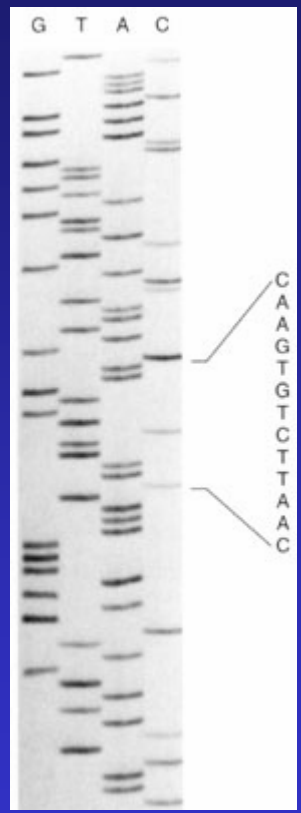
DIAGNÓSTICO MOLECULAR

TRANSFORMAÇÃO DA MEDICINA

O gene é uma seqüência definida de nucleotídeos

Sanger, F. and Coulson, A. R.. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94**: 444-448.
Método de sequenciamento de DNA baseado em terminação de cadeia por incorporação de ddNTPs

Maxam, A.M. and Gilbert, W. (1977) A new method of sequencing DNA. *Proc. Nat.Acad. Sci. USA* **74**: 560-564.
Método químico de sequenciamento de DNA.



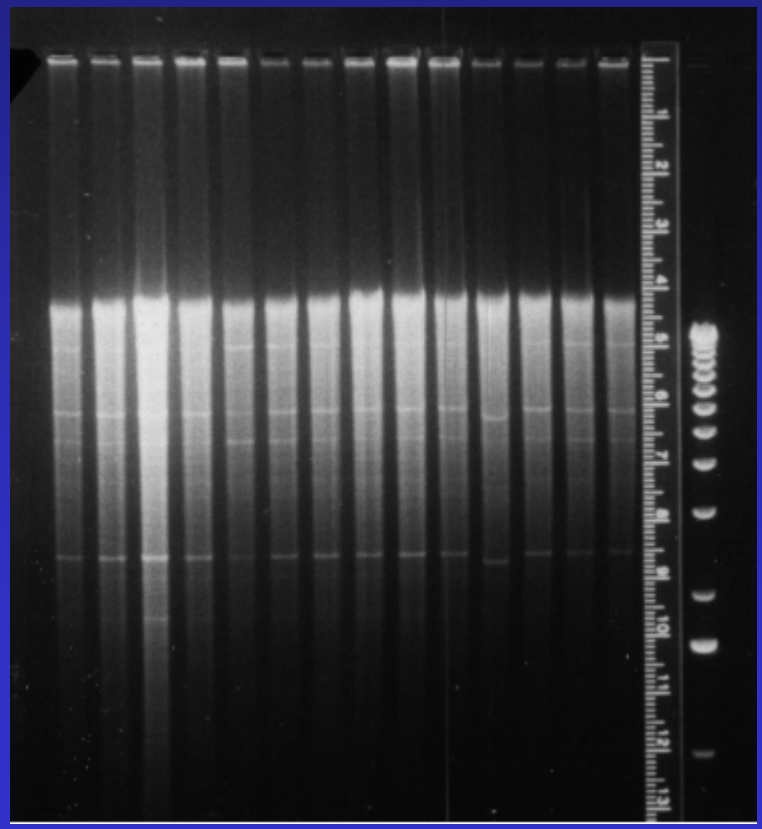
Fred Sanger

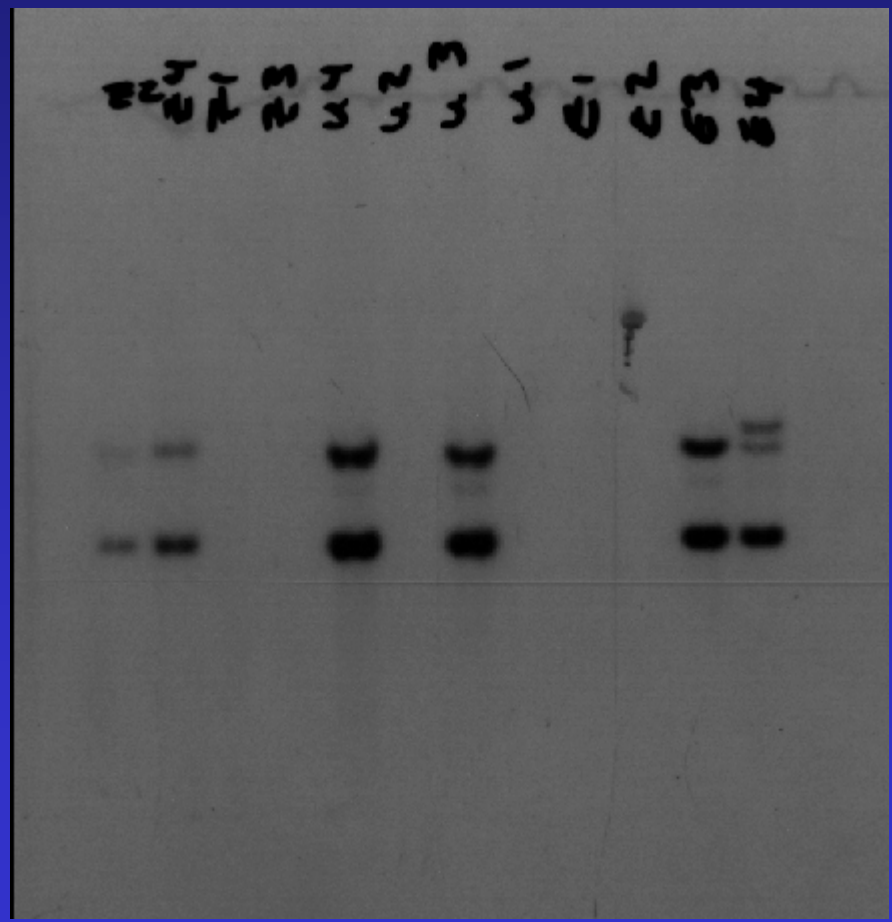
Courtesy of Dr. F. Sanger, MRC, Cambridge.
Noncommercial, educational use only.

Sanger recebeu o Nobel duas vezes : a primeira, nos anos 50, pelo seqüenciamento de proteínas (insulina) e a segunda pelo seqüenciamento de DNA, em 1980.

1975 - ‘Southern Blotting”

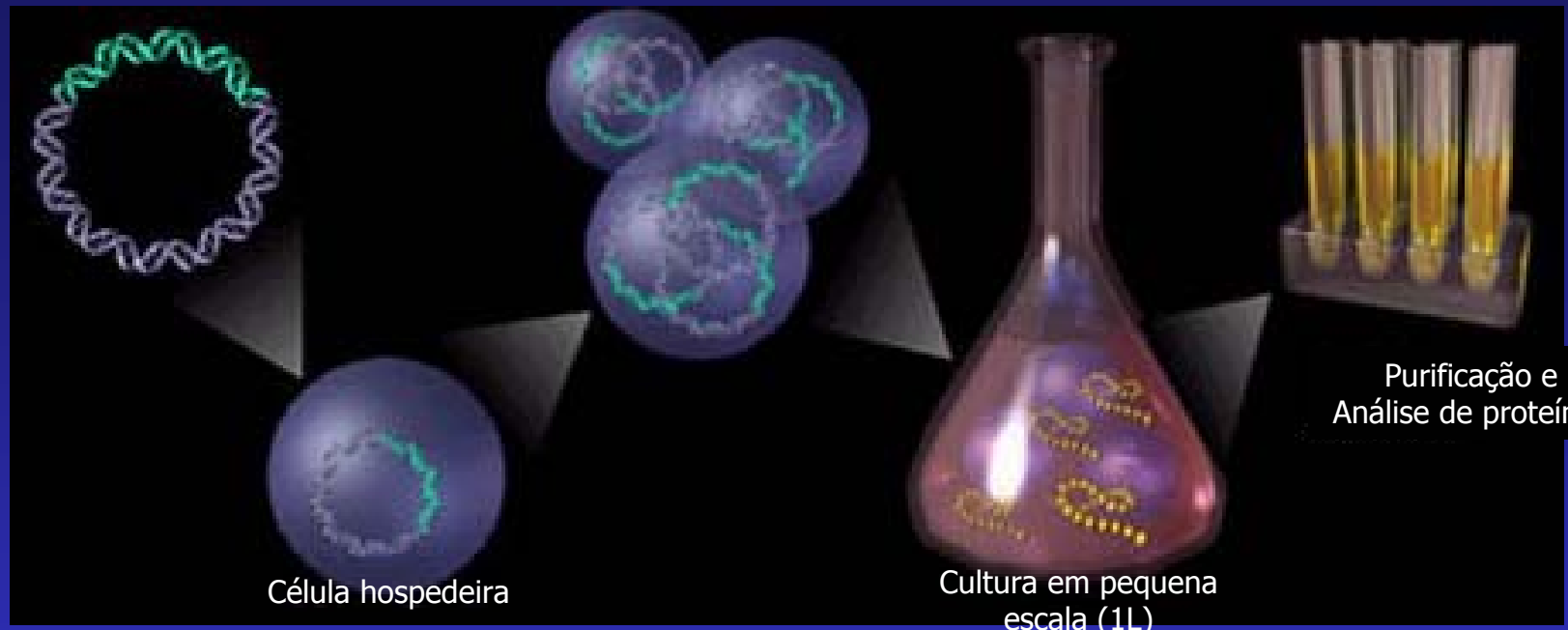
E.M. Southern. (1975) "Detection of specific sequences among DNA fragments separated by gel electrophoresis." J. Mol. Biol. 98: 503.





Genentech, Inc.
In Business for Life.

Célula recombinante



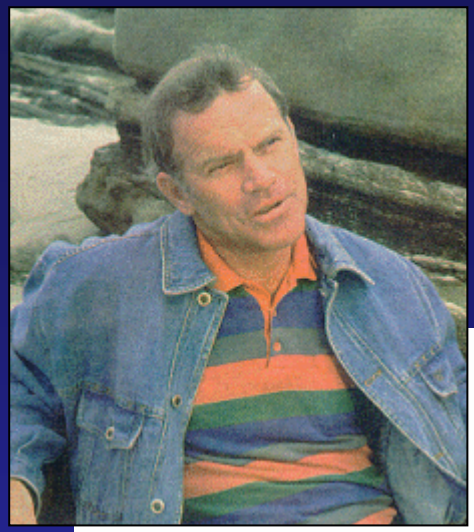
1976 Fundada por Robert Swanson and Herber Boyer

1977 Produção da primeira proteína humana (somatostatina) em um hospedeiro microbiano (*E.coli*)

1978 Clonagem da insulina humana

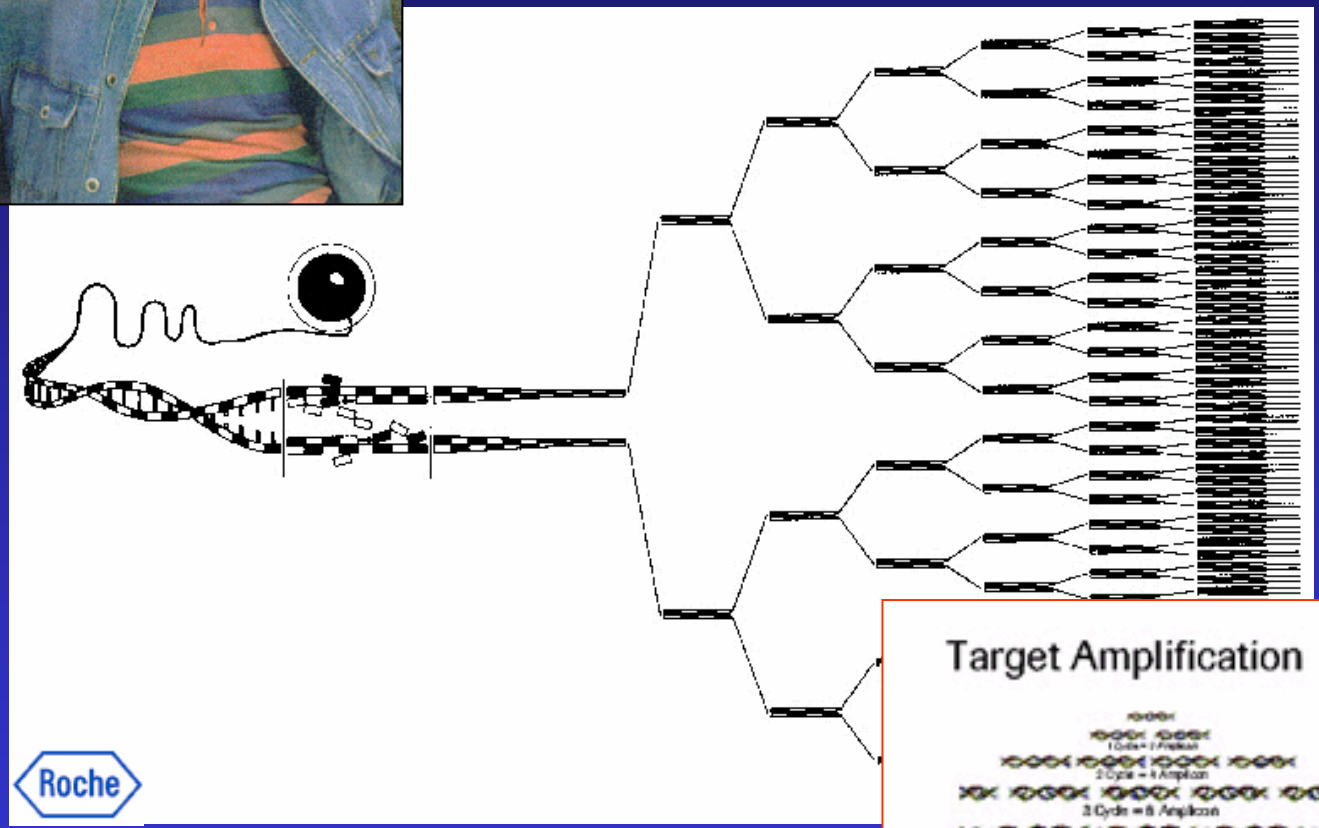
1979 Clonagem do hormônio de crescimento humano

1982 Primeira droga recombinante lançada no mercado (insulina humana)
Produção laboratorial do fator de coagulação VIII

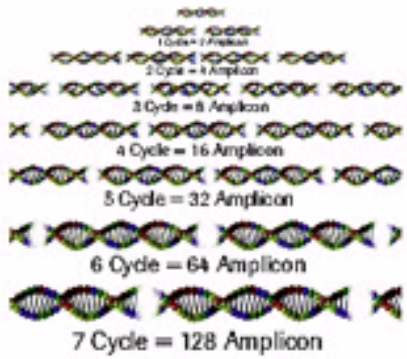


Kary Mullis (Cetus Corp.) – 1987

Permite obter, *in vitro*, grandes quantidades de uma sequência específica de DNA.



Target Amplification

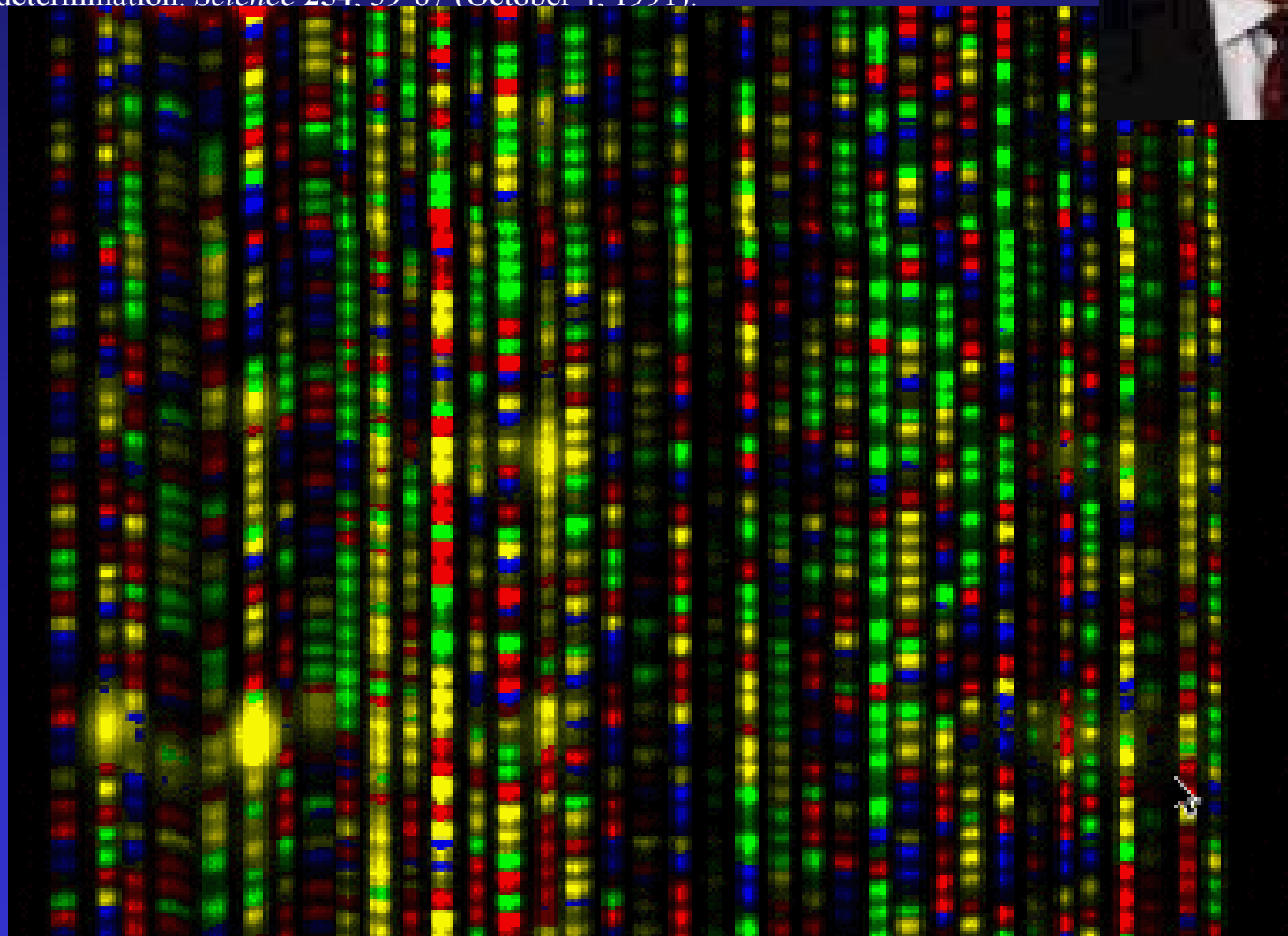


No. of Cycles	No. Amplicon Copies of Target
1	2
2	4
3	8
4	16
5	32
6	64
20	1.048,576
30	1.073,741,824

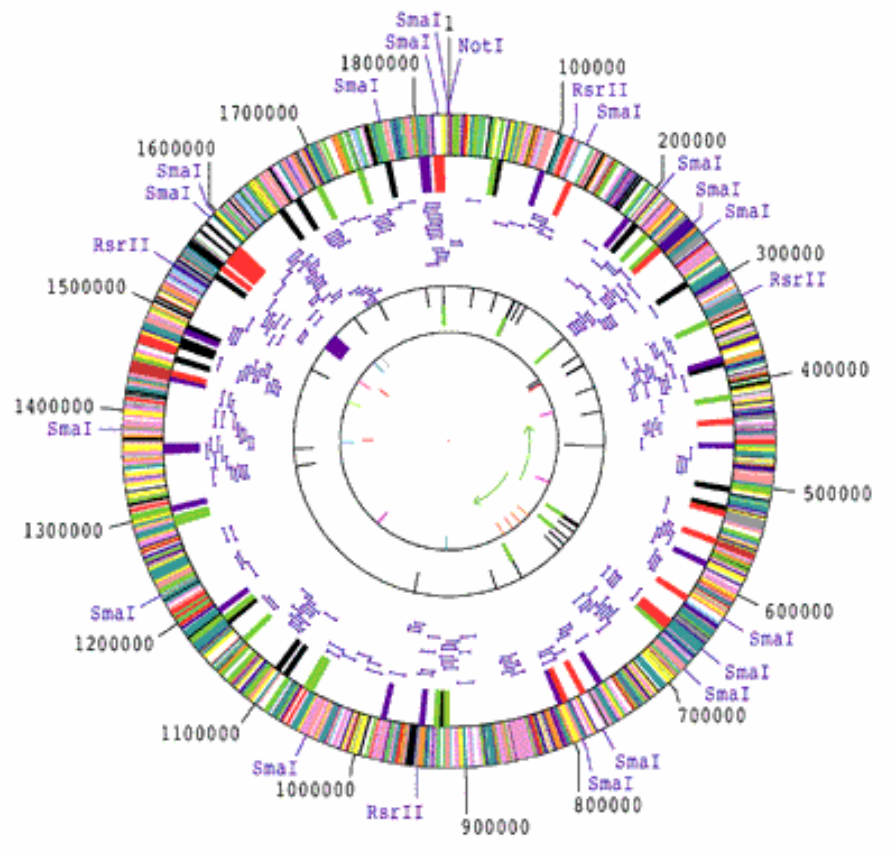
1986 - Sequenciador Automático de DNA (Leroy Hood)

Lloyd M. Smith et al. Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674-679 (June 12, 1986).

T. Hunkapiller, R. J. Kaiser, B. F. Koop, L. Hood. Large-Scale and automated DNA sequence determination. *Science* **254**. 59-67 (October 4, 1991).

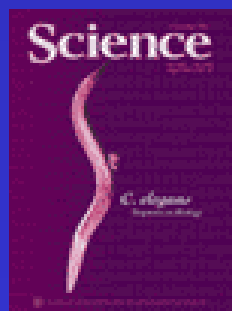


1995: Primeiro genoma sequenciado
Haemophilus influenzae



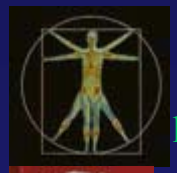
1998: *C. elegans* - primeiro genoma completo de um animal

A sequência genômica do nematódio de vida livre *Caenorhabditis elegans* está quase completa e representa o primeiro genoma de um organismo multi-celular a ser inteiramente sequenciado. O genoma tem aproximadamente 97 Mb de tamanho, e codifica cerca de 19.000 proteínas. O projeto de sequenciamento foi um esforço colaborativo entre o Genome Sequencing Center em St. Louis e o Sanger Center, em Hinxton, UK.



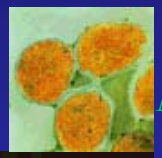
2000: Rascunho do genoma humano

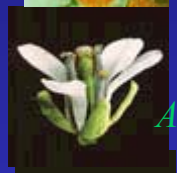
nature genomegateway


human


mouse


C. elegans

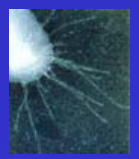

M. leprae


Arabidopsis


rat


Buchenerasp


Yersinia

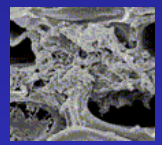

Ralstonia


Drosophila


Vibrio cholera


Plasmodium


Neisseria

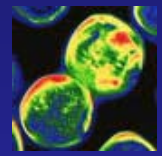

Xylella


E. coli


S. cerevisiae


S. pombe


Rickettsia


Archaeoglobus


M. tuberculosis


Helicobacter

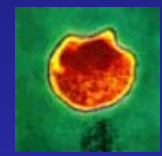

Borrelia

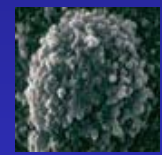

Bacillus

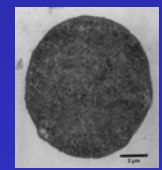

Salmonella


Campilobacter


Aquifex


Chlamidia


Ureaplasma


Thermoplasma

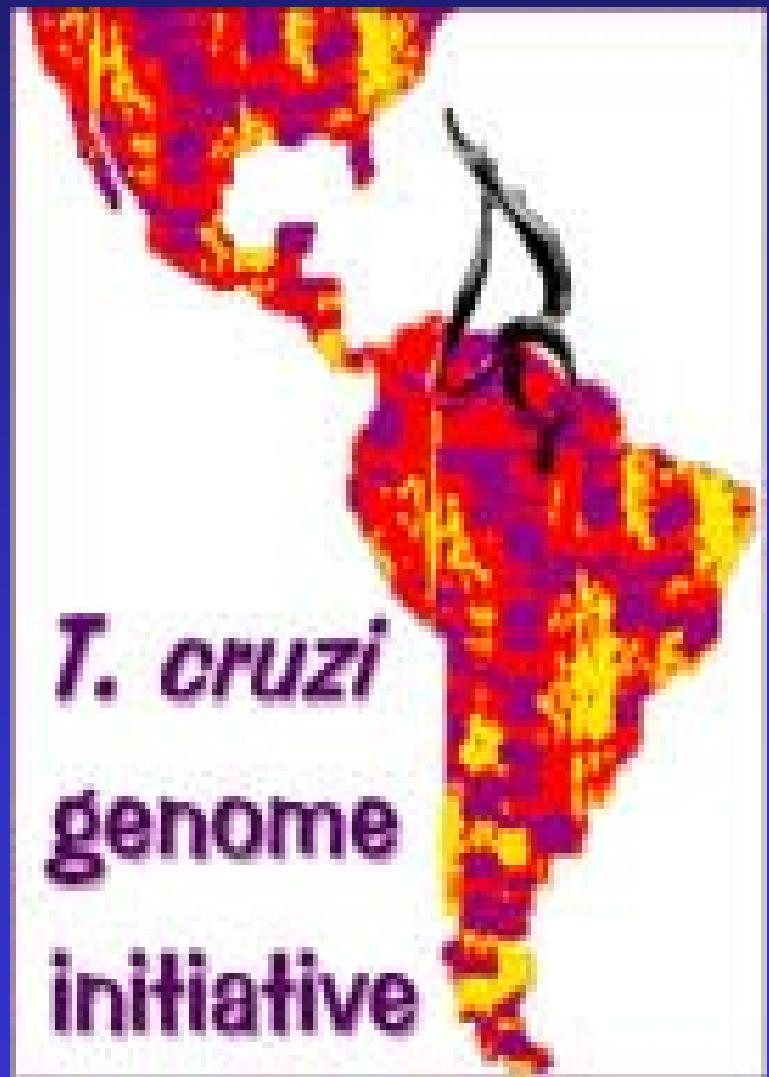

Thermotoga

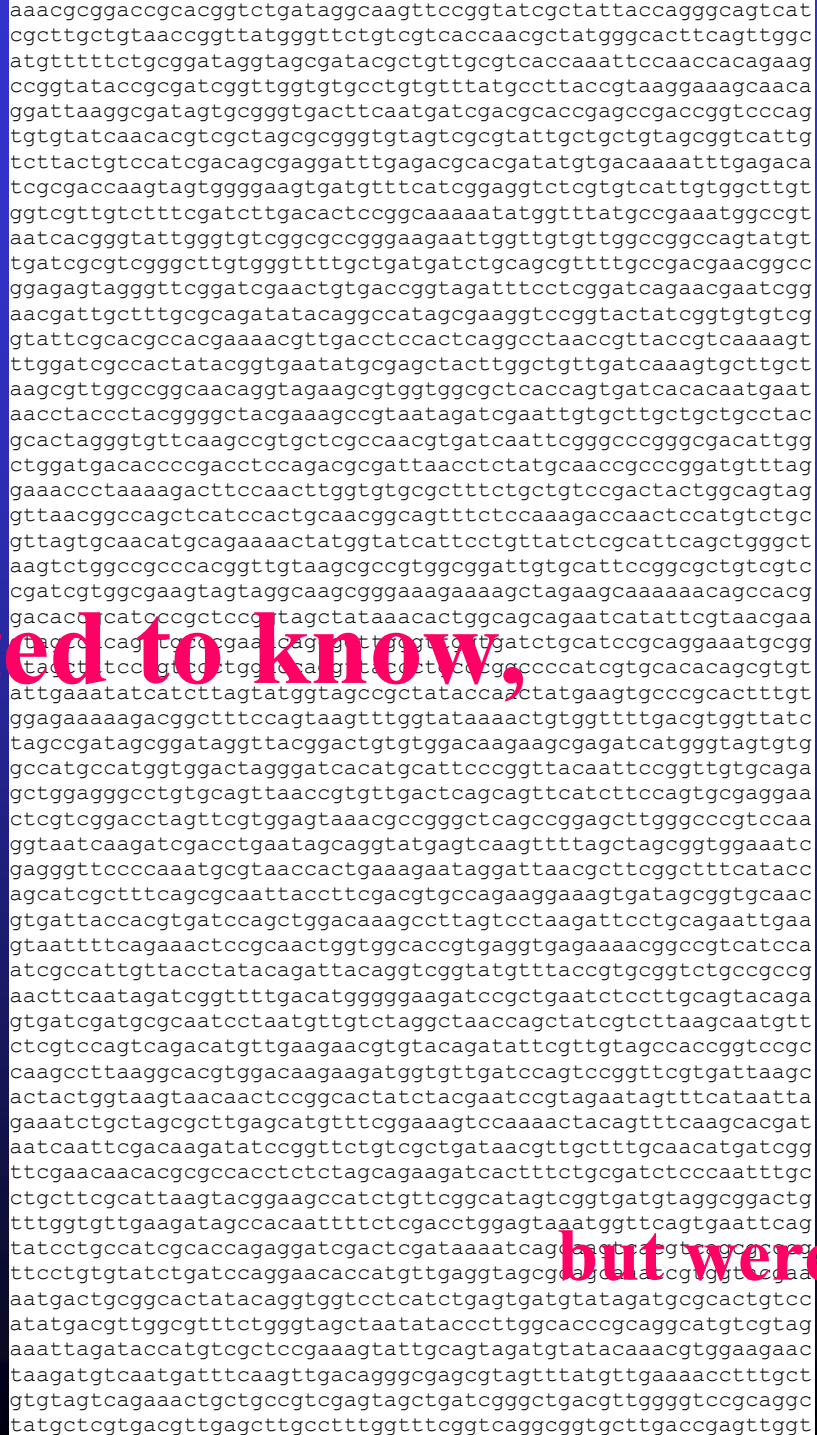

Pseudomonas

1997 - 2003

The *Trypanosoma cruzi* Genome Initiative

FOR MORE INFORMATION :
http://www.dbbm.fiocruz.br/



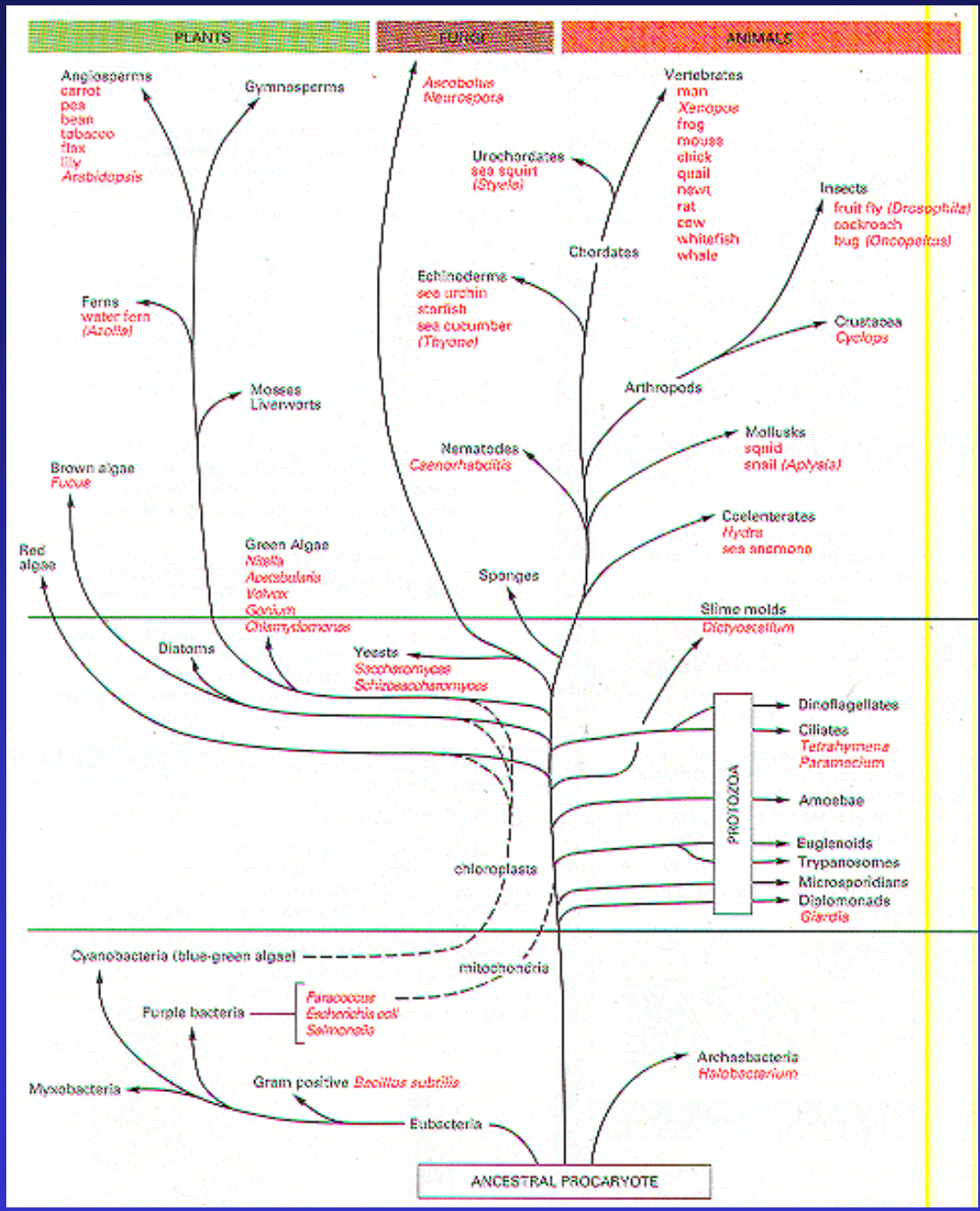

All you wanted to know,
but were afraid to ask...

Why Genome Projects : ??

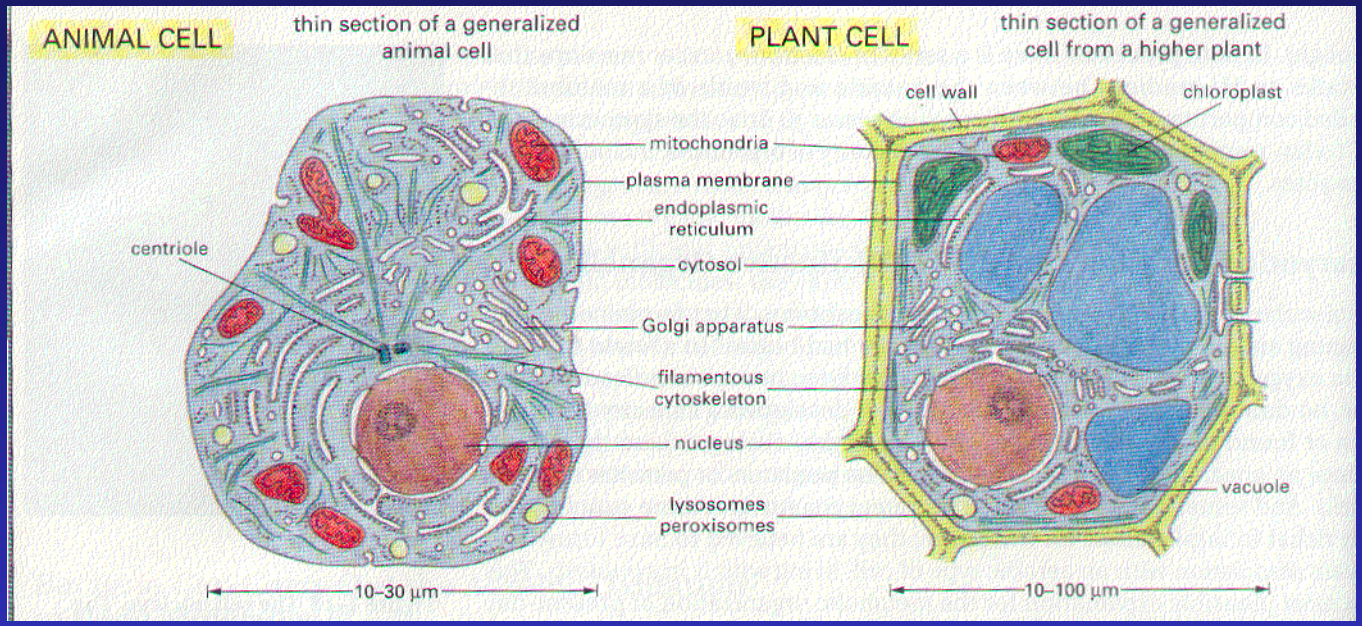
To :

- **Drastically increase knowledge on the (molecular) biology of the pathogen**
- **Develop new tools for vaccine development; find new drug targets, improve diagnostics**
- **Study the evolutionary relationship with orther oganisms, variability between strains, lineages, isolates**
- **Build expertise, discover keys of life, analyse structure-function relationships, learn about the interaction with the host etc.**

- **At least 60 microbial genome projects have been finished, and more than 150 are on-going.**

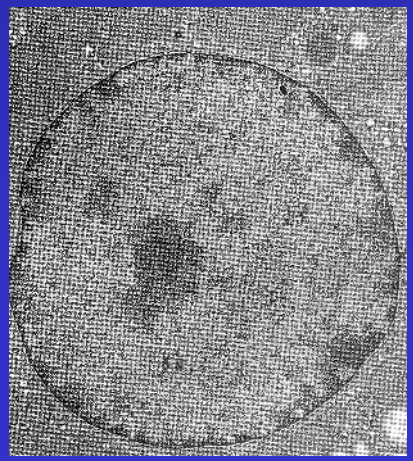

PLANTS
FUNGI
ANIMALS
Angiosperms
carrot
pea
bean
tobacco
flax
lily
Arabidopsis
Gymnosperms
Ferns
water fern
(Azolla)
Mosses
Liverworts
Brown algae
Fucus
Red algae
Green Algae
Nitella
Acetabularia
Volvox
Gonium
Chlamydomonas
Diatoms
Ascobolus
Neurospora
Yeasts
Saccharomyces
Schizosaccharomyces
Urochordates
sea squirt
(Styela)
Vertebrates
man
Xenopus
frog
mouse
chick
quail
newt
rat
cow
whitefish
whale
Chordates
Echinoderms
sea urchin
starfish
sea cucumber
(Thyone)
Insects
fruit fly (Drosophila)
cockroach
bug (Oncopeltus)
Crustacea
Cyclops
Arthropods
Nematodes
Caenorhabditis
Mollusks
squid
snail (Aplysia)
Coelenterates
Hydra
sea anemone
Sponges
Slime molds
Dictyostelium
Dinoflagellates
Ciliates
Tetrahymena
Paramecium
Amoebae
Euglenoids
Trypanosomes
Microsporidians
Diplomonads
Giardia
PROTOZOA
chloroplasts
Cyanobacteria (blue-green algae)
mitochondria
Purple bacteria
Paracoccus
Escherichia coli
Salmonella
Myxobacteria
Gram positive Bacillus subtilis
Eubacteria
Archaebacteria
Halobacterium
ANCESTRAL PROCARYOTE

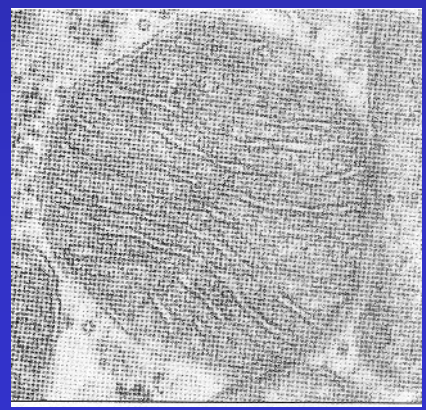
Animal and plant cells



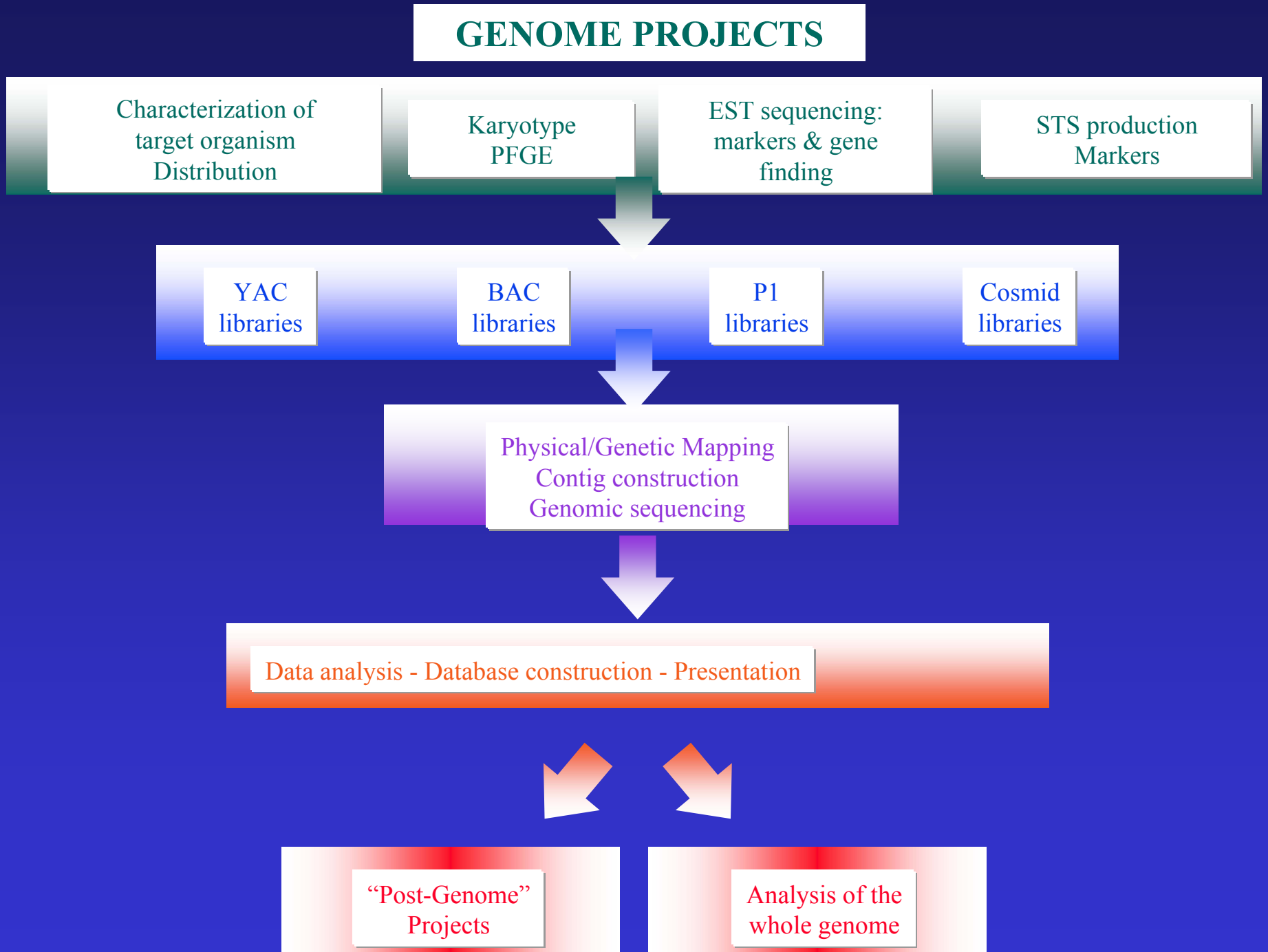
The mitochondrium

The nucleus

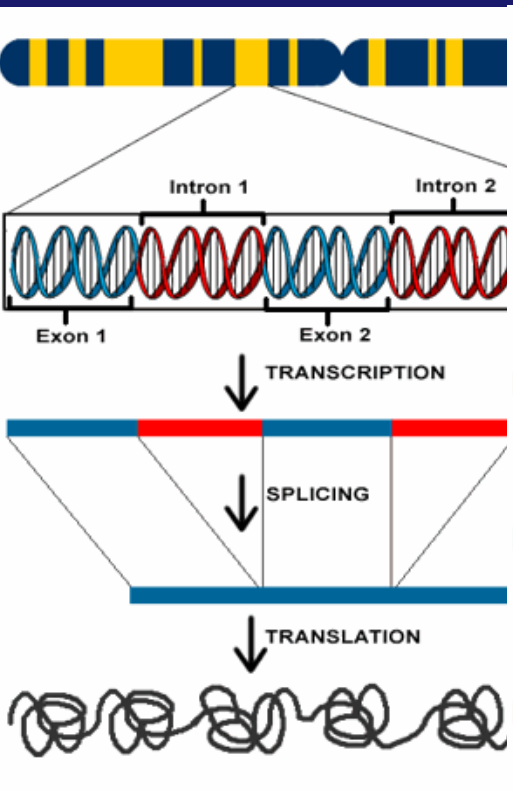


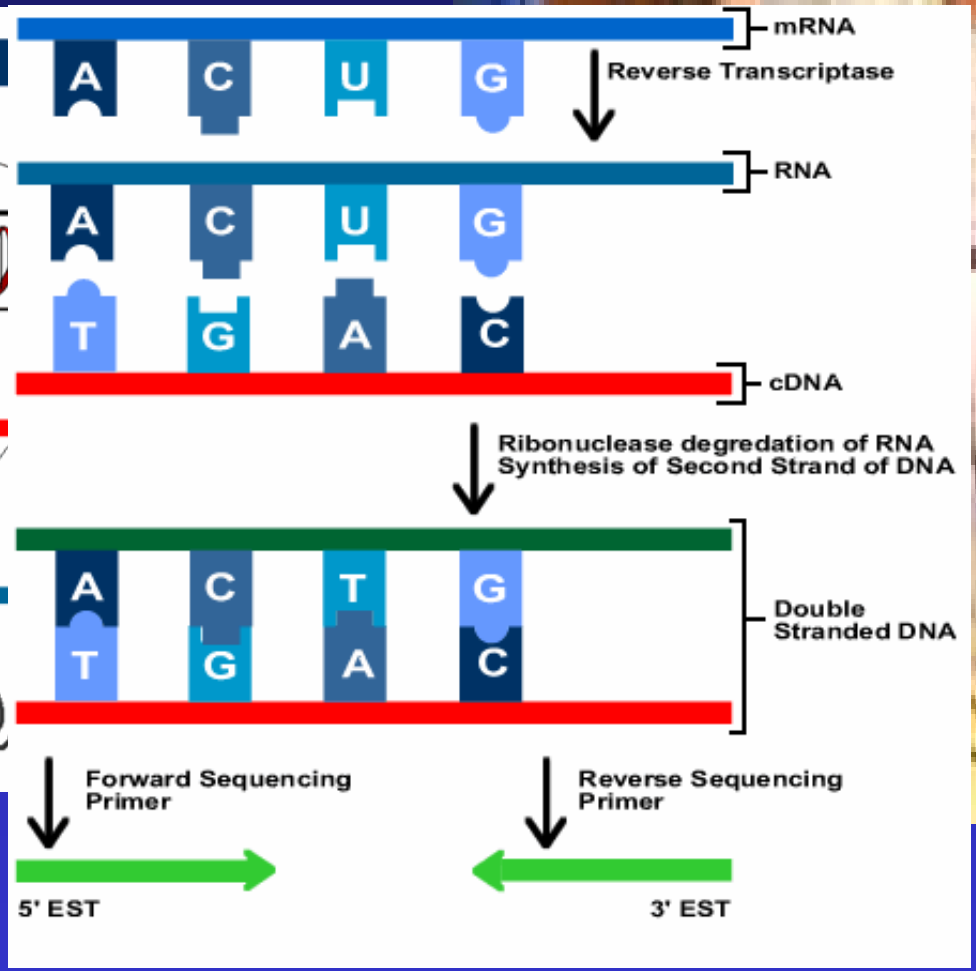


GENOME PROJECTS



1991 - J.Craig Venter
Expressed Sequence Tags





Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R., Kerlavage, A.R., McCombie, W.R., and Venter, J.C. Complementary DNA sequencing: "expressed sequence tags and the human genome project. *Science* **252**, 1651-1656 (1991).

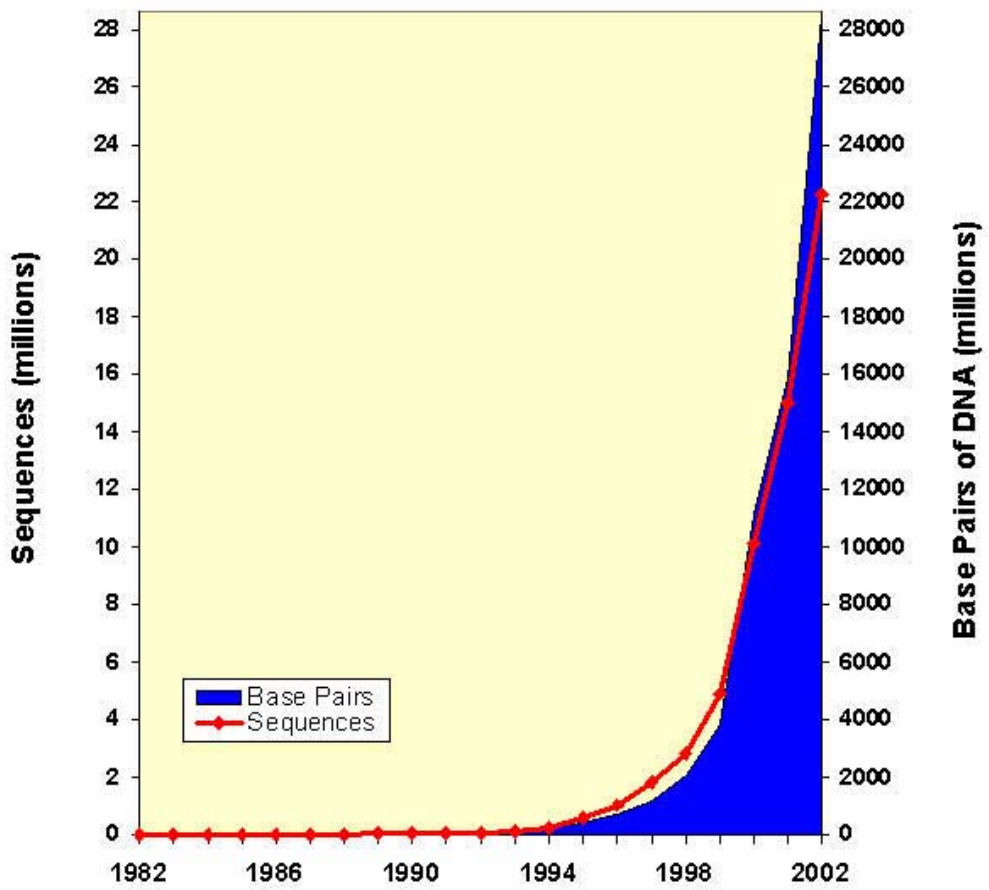
How to deal with the enormous amount of reads generated by the high throughput DNA sequencers?

Sanger Centre

NCBI

National Center for Biotechnology Information

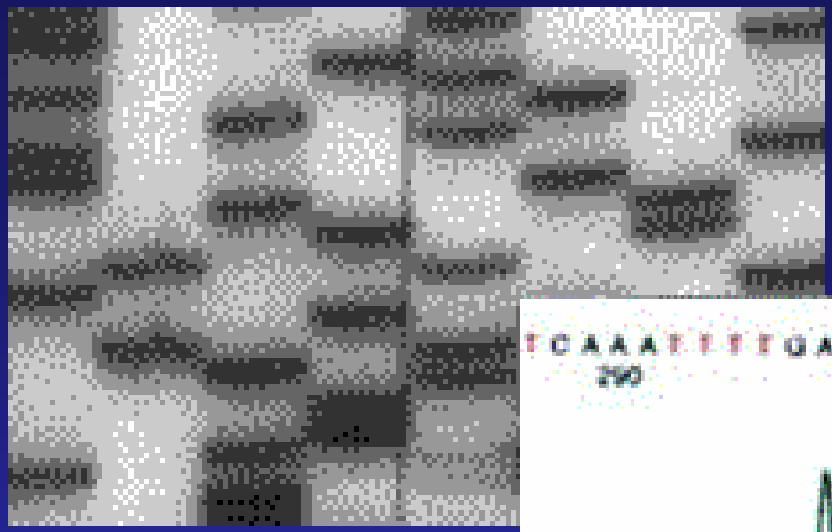
National Library of Medicine — National Institutes of Health



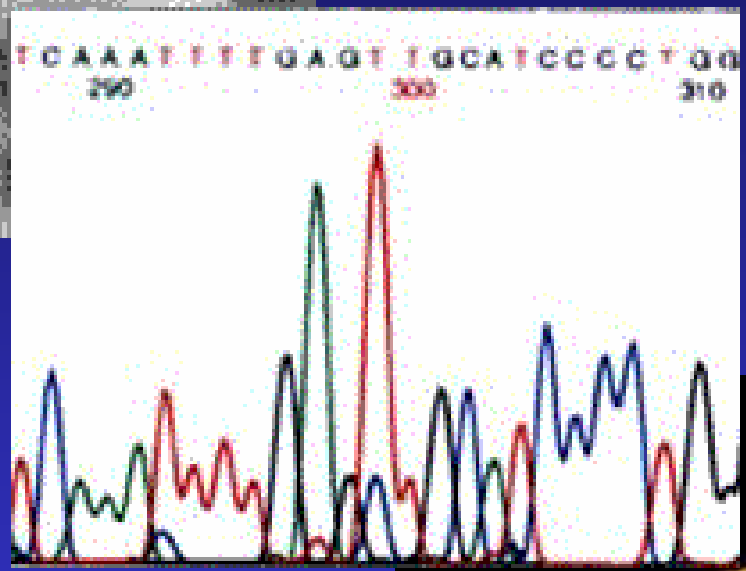
2002

28,507,990,166 bp

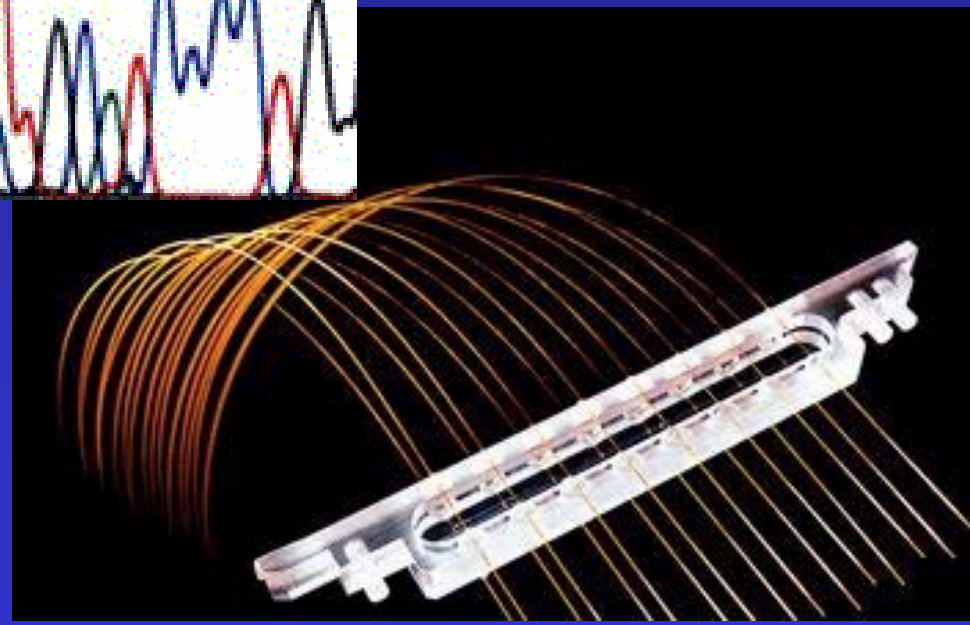
22,318,883 sequences



Sequenciamento manual



Sequenciamento automático



Sequenciamento capilar

Do gel a sequência: quatro etapas

1. lane tracking - identificação dos limites de cada lane.

2. lane profiling - cada um dos sinais é usado para a criação de um perfil (profile ou trace).

3. trace processing - são utilizados métodos matemáticos para o processamento das estimativas de sinal.

4. base-calling - os perfis são traduzidos numa sequência de bases.

Possíveis problemas:

- compressões
- variação nos sinais
    - eficiência no elongamento
    - eficiência da terminação
- quimeras
- reads de contaminantes
- mutações nos subclones
- sequências de vetores não removidas
- alinhamento errôneo

O que é Phred/Phrap/Consed?

R. É um pacote de programas para:

- Leitura de cromatogramas (trace files);
- Atribuição de qualidade a cada base individual;
- Identificação e mascaramento de sequências do vetor e repetições;
- Montagem de sequências;
- Visualização e edição da montagem;
- Acabamento automático.

Phred

Phred é um programa que realiza várias tarefas:

a. Lê cromatogramas – compatível com a maioria dos formatos de arquivo: SCF (standard chromatogram format), ABI (373/377/3700), ESD (MegaBACE) and LI-COR.

b. "Calls" bases – atribui uma base para cada pico identificado com uma taxa de erro menor do que os programas "standard".

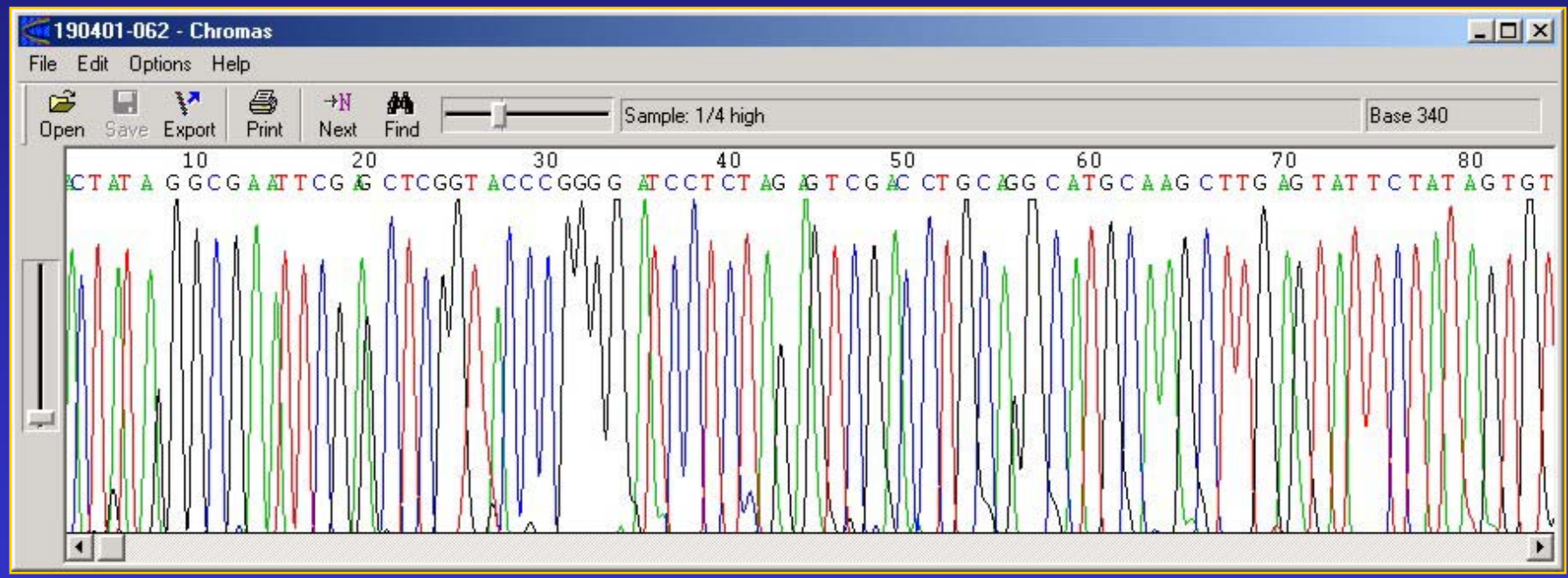
Phred

c. Atribui valores qualitativos para cada base (Phred value) – baseado em uma estimativa da taxa de erro calculada para cada base.

d. Cria arquivos de saída – as atribuições das bases e os valores de qualidade são escritos em arquivos de saída.

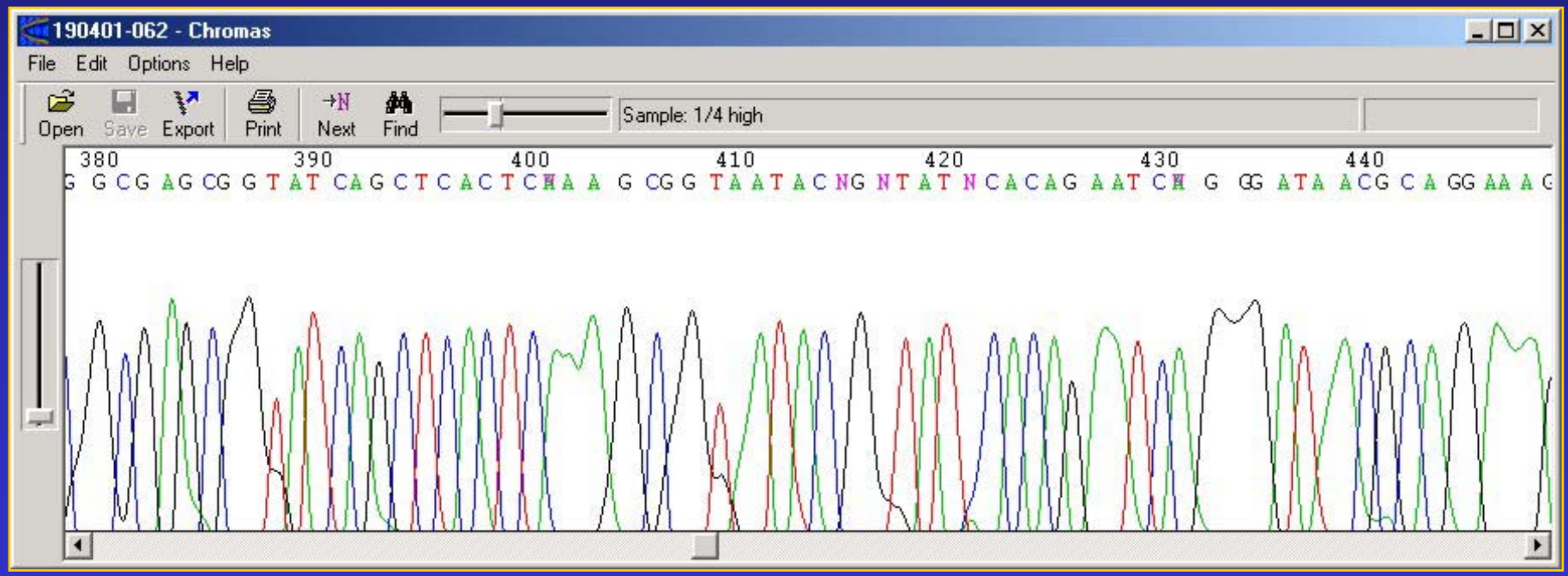
Trace File

Região de alta qualidade - nenhuma ambiguidade (N)



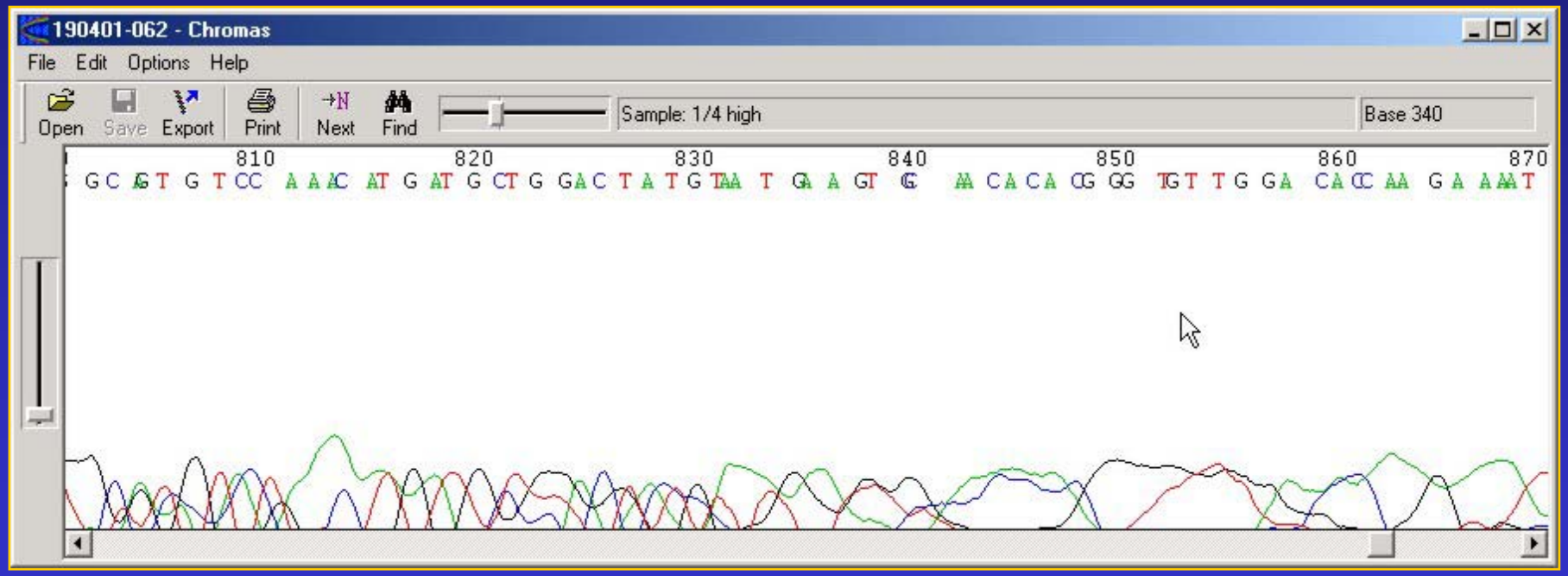
Trace File

Região de qualidade média - algumas ambiguidades (N)



Trace File

Região de baixa qualidade



Fórmula para calcular os valores Phred:

$q = -10 \times \log_{10}(p)$ onde:

q - valor da qualidade

p - probabilidade estimada de erro para uma atribuição de base

Exemplos:

q = 20 significa *p* = 10^{-2} (1 erro em 100 bases)

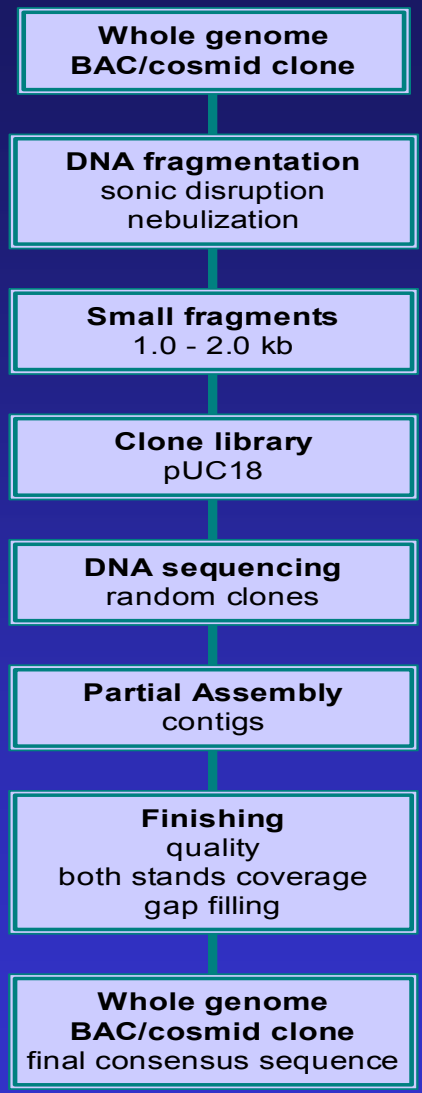
q = 40 significa *p* = 10^{-4} (1 erro em 10,000 bases)

A estrutura de um arquivo phd

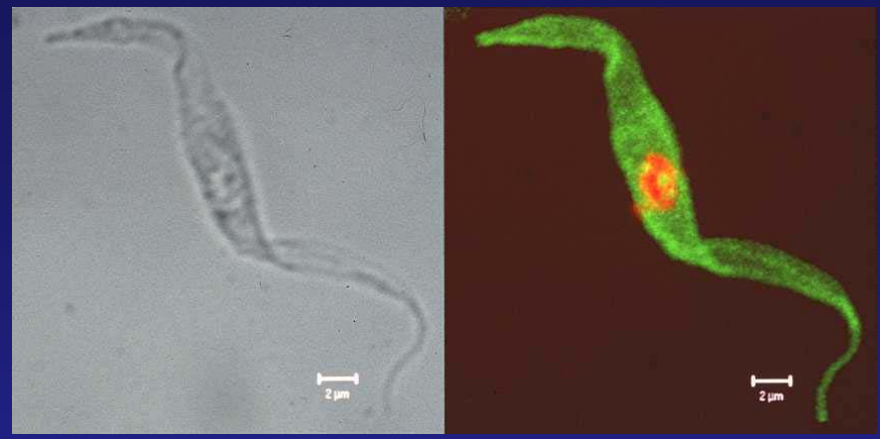
```
BEGIN_SEQUENCE 01EBV10201A02.g

BEGIN_COMMENT

CHROMAT_FILE: EBV10201A02.g
ABI_THUMBPRINT:
PHRED_VERSION: 0.990722.g
CALL_METHOD: phred
QUALITY_LEVELS:99
TIME: Thu May 24 00:18:58 2001
TRACE_ARRAY_MIN_INDEX: 0
TRACE_ARRAY_MAX_INDEX: 12153
TRIM:
CHEM: term
DYE: big

END_COMMENT

BEGIN_DNA
t 8 5
c 13 17
a 19 26
c 19 32
t 24 2221
a 24 2232
a 22 2245
a 27 2261
g 25 2272
c 19 2286
c 12 2302
t 19 2314
g 12 2324
g 15 2331
g 19 2346
g 23 2363
t 33 2378
g 36 2390
c 44 2404
c 44 2419
t 39 2433
a 39 2446
a 34 2460
t 35 2470
g 34 2482
t 16 8191
g 19 8200
t 13 8211
c 13 8229
g 4 8241
n 4 8253
c 4 8263
t 10 8276
t 9 8286
c 12 8301
t 16 8313
c 12 8329
c 12 8336
c 15 8343
t 19 8356
c 9 8371
g 13 8386
g 14 8397
a 7 8417
g 9 8427
g 4 8445
t 6 11908
a 6 11921
g 6 11927
t 6 11947
c 6 11953
a 6 11964
g 6 11981
c 4 11994
n 4 12015
c 4 12037
n 4 12044
n 4 12058
n 4 12071
n 4 12085
n 4 12098
n 4 12111
n 4 12124
c 4 12144
n 4 12151
END_DNA

END_SEQUENCE
```

Por que montar sequências?

- Os métodos atuais de sequenciamento geram sequências de 500-700 bp – limite de resolução da eletroforese.
- Genomas completos ou grandes clones precisam ser fragmentados - biblioteca de clones.
- Fragmentos pequenos são sequenciados aleatoriamente (shotgun) – as sequências são montadas para a obtenção do consenso final.

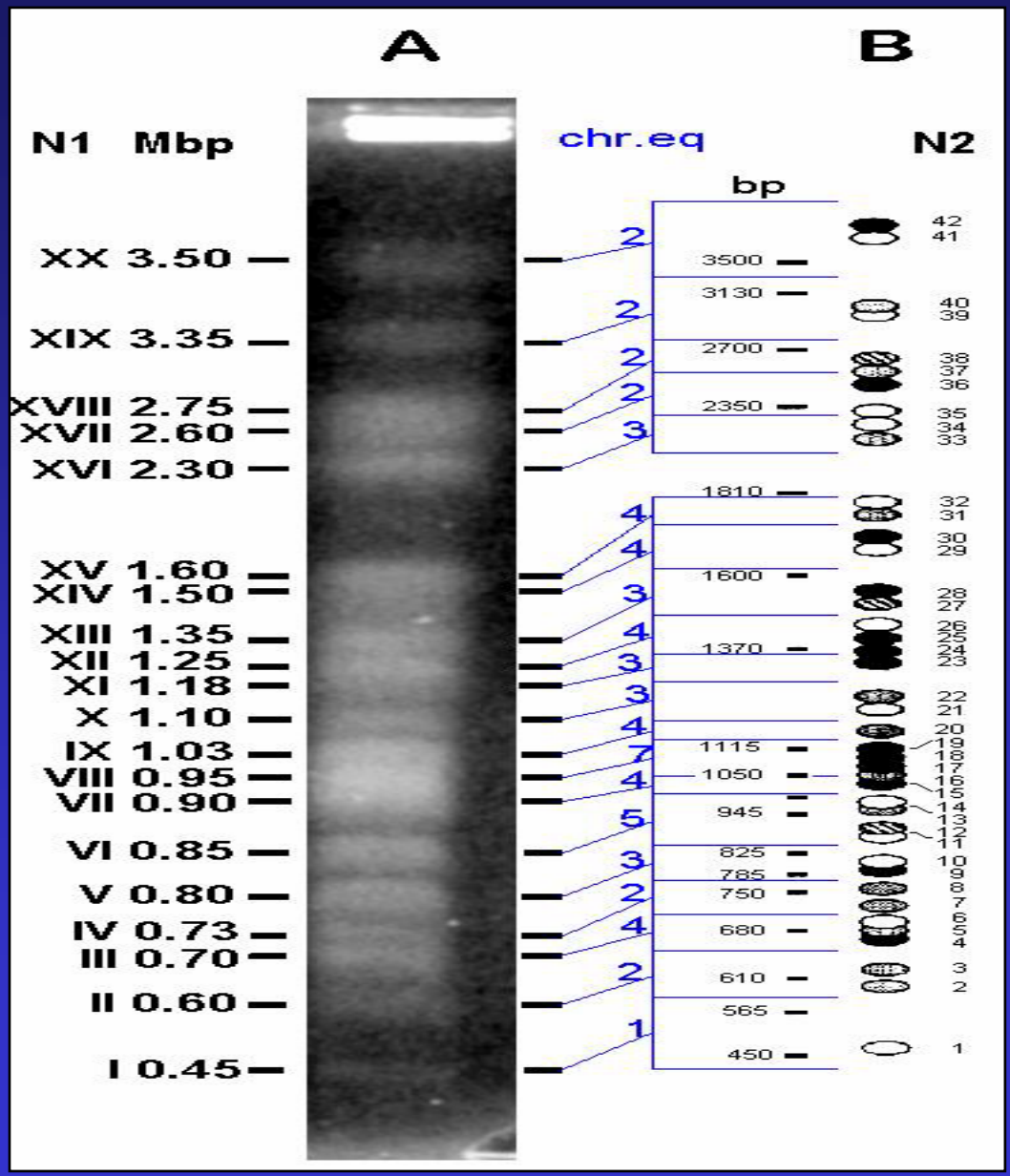


What do kinetoplastidae parasites have in common?

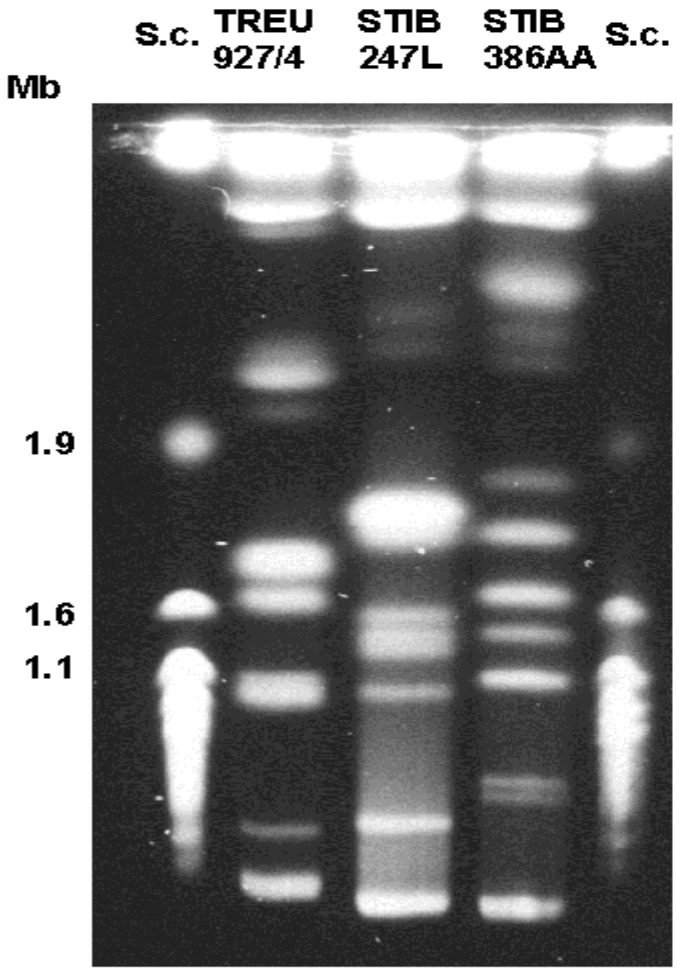


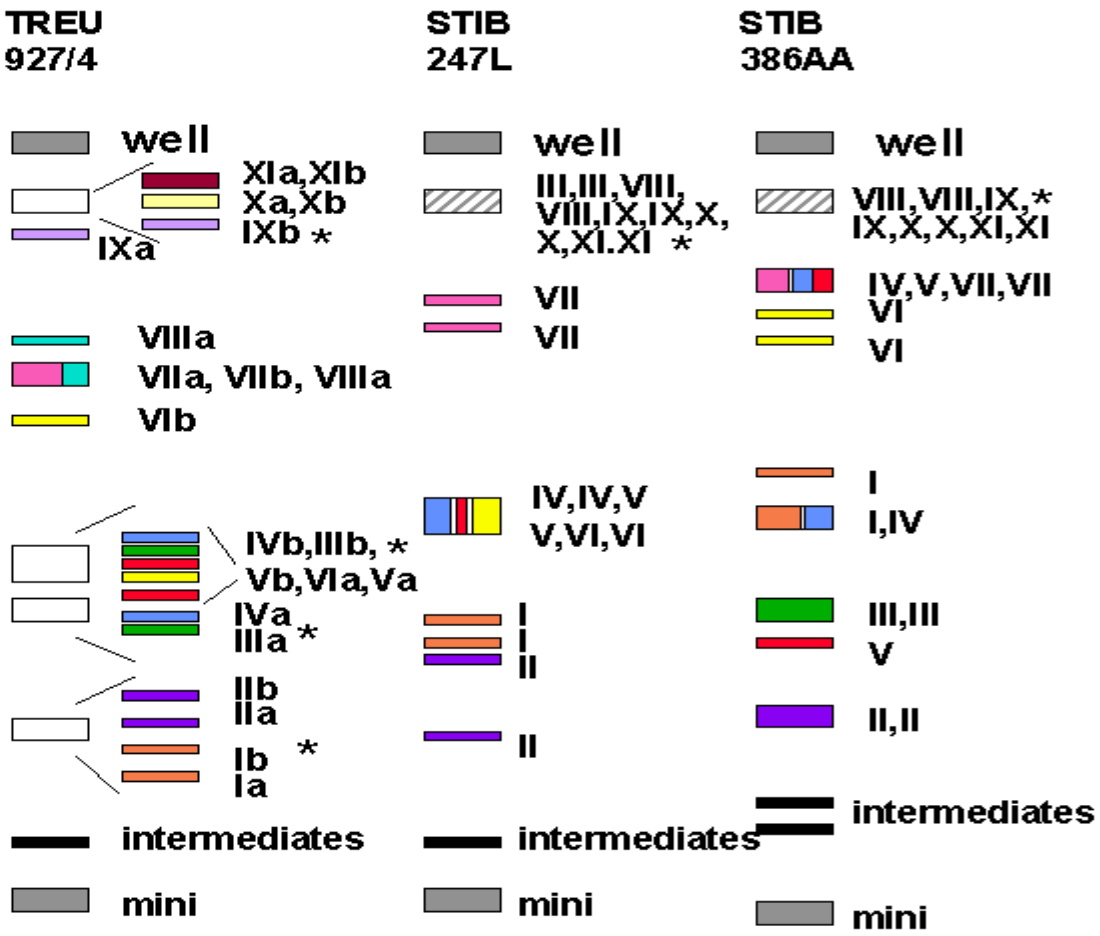
- Trans-splicing mechanism (also some cis-splicing?)
- RNA editing for mitochondrial genes (pan-editing for *T. cruzi* and *T. brucei*)
- Kinetoplast structure : maxi-circles and minicircles
- Life cycle and stages are similar, but have also important differences
- Chromosomes do not condense -> visualize through pulse-field gel electrophoresis
- Absent or very rare sexual recombination. T. cruzi has typical clonal structure
- Karyotype shows a lot of plasticity
- Important size differences exist between homologous chromosomes
- Telomeric structures are similar
- Chromosomes analysed thusfar have large polycistronic transcription
- No Pol II promoter structures, nor centromeres could be identified thusfar
- RNA processing occurs with concerted transsplicing and polyadenilation
- RNA stability is an important regulation factor
- Stage specific gene expression is often influenced by 3'-UTR sequences (?)

T.cruzi CL-Brener Karyotype



T. brucei karyotypes

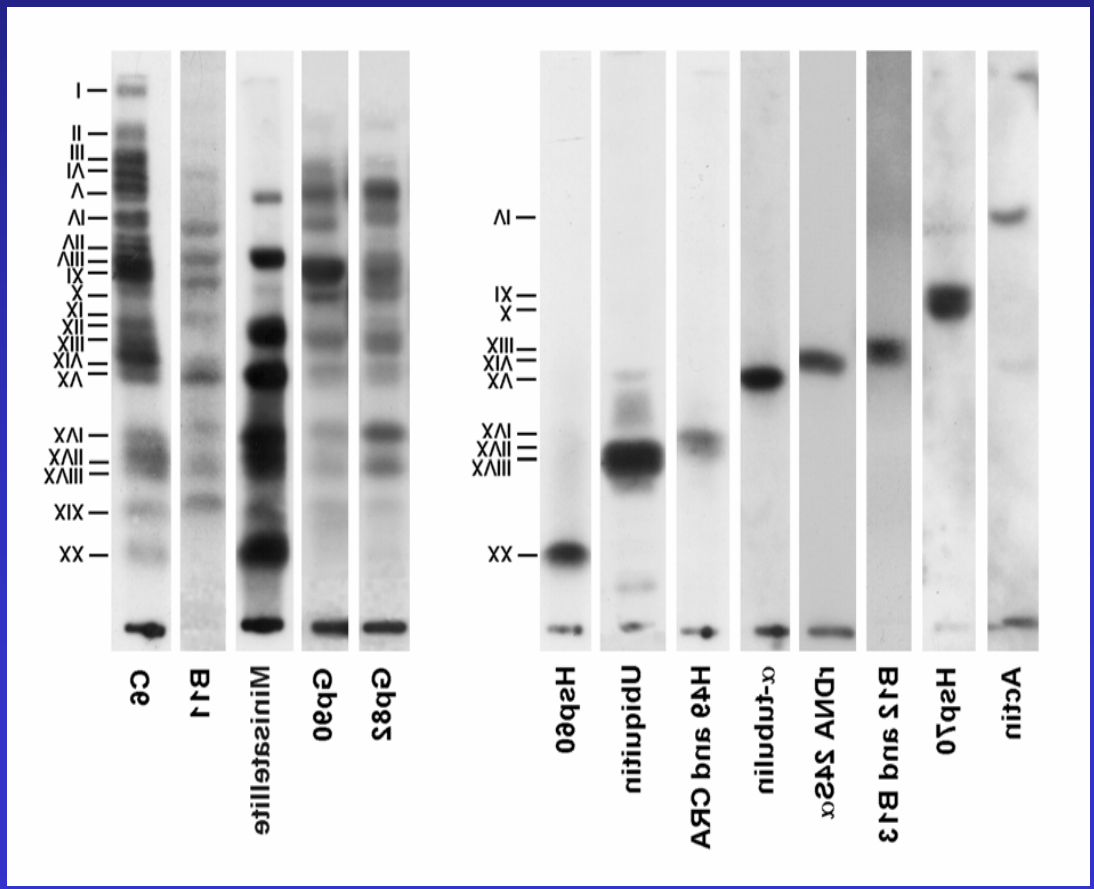


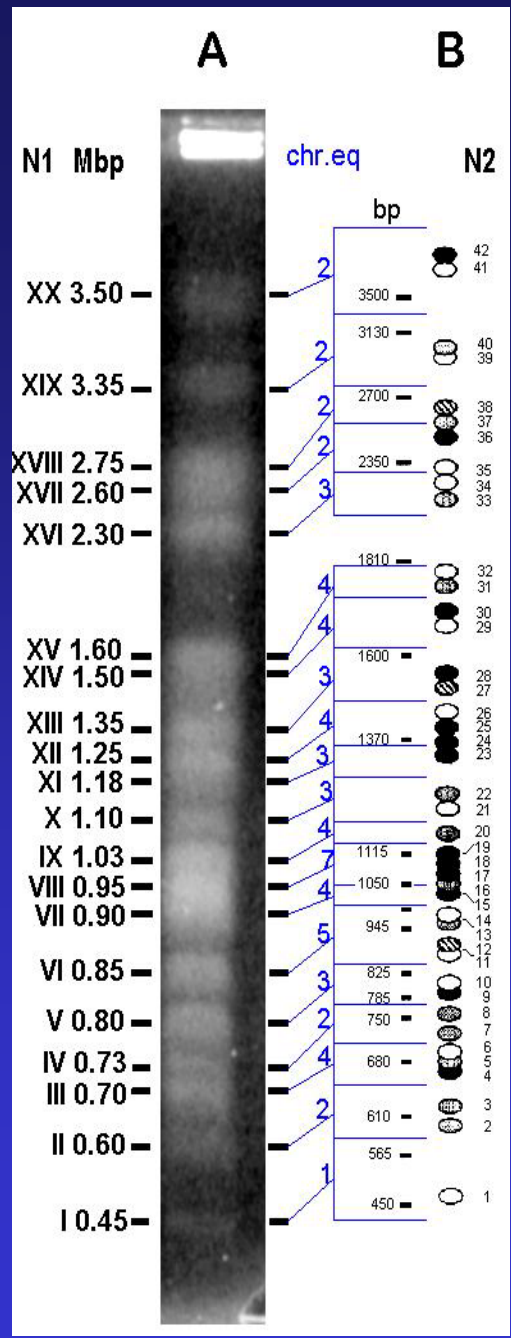


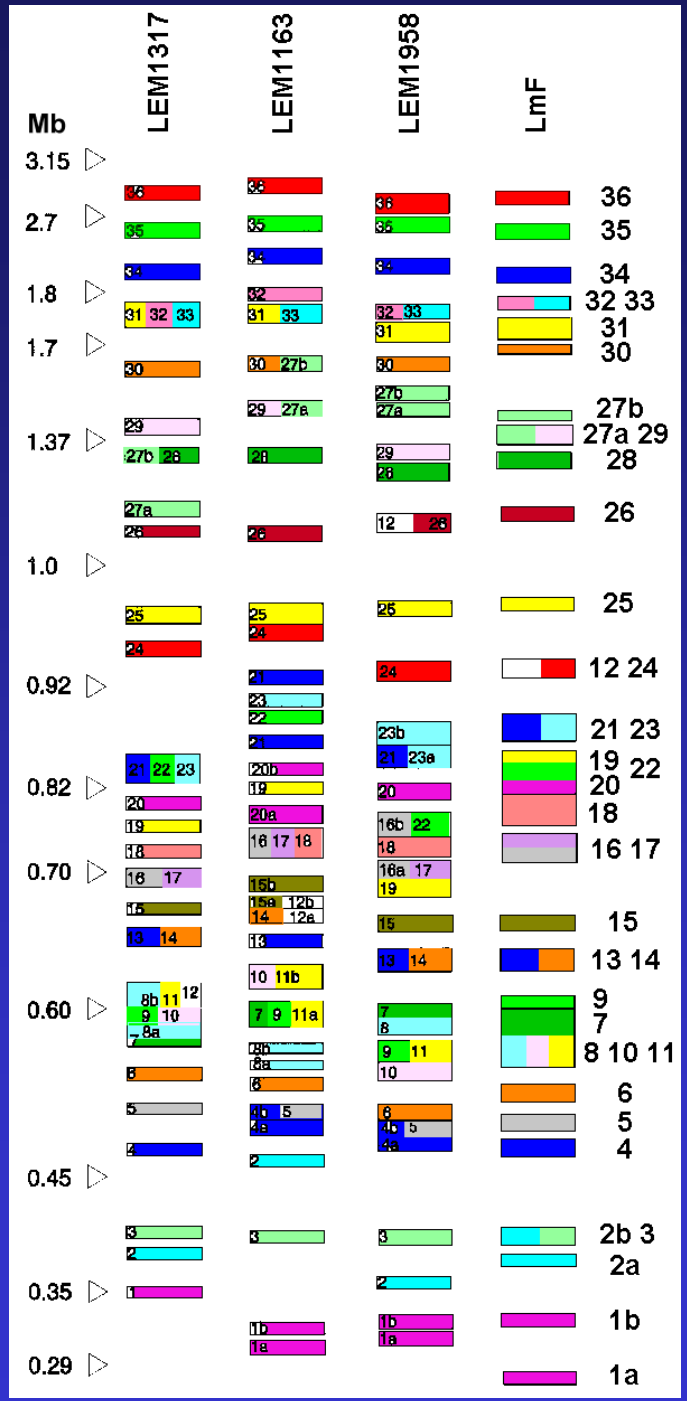
* size order determined on other gels
S.c.: *S. cerevisiae* strain YPH80

Karyotype and markers

T. cruzi CL-Brener :
64-70 chromosomal bands from 0.45 - 3.7 Mb
~ 47 Mb per haploid genome



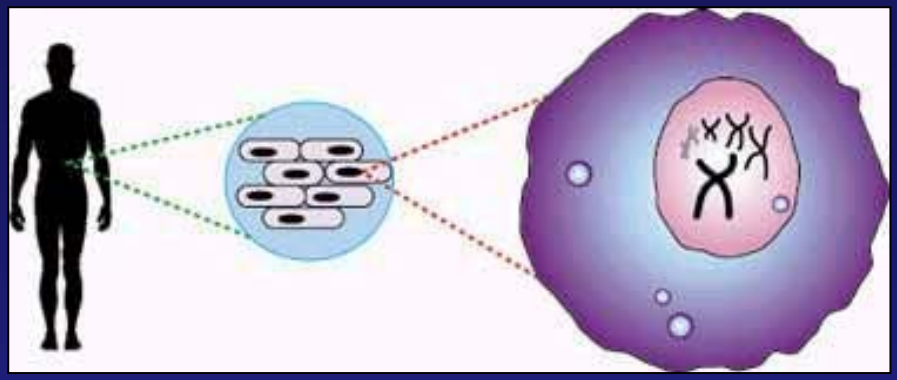


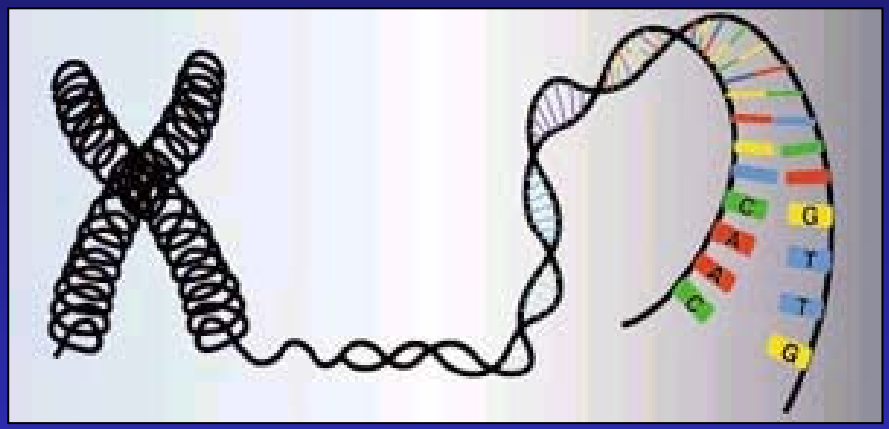


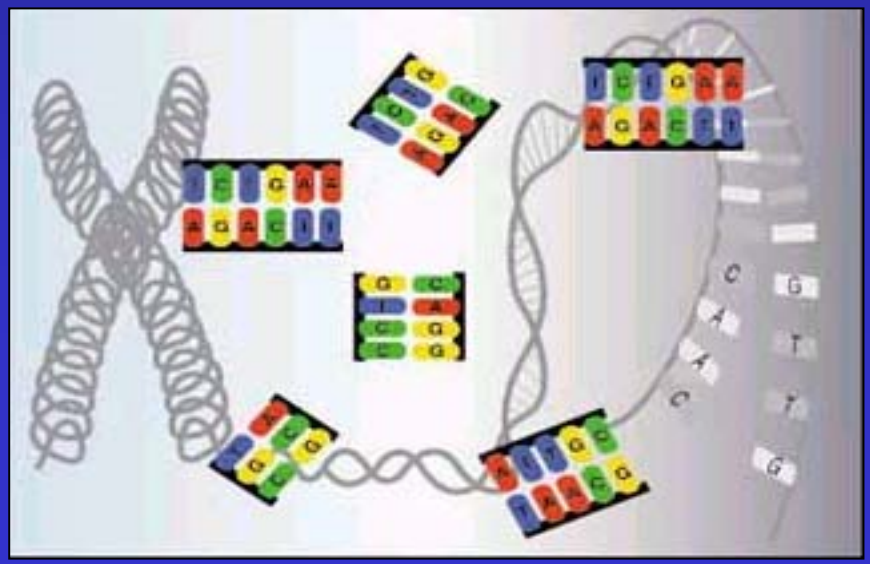
Leishmania karyotype ideograms

Leishmania infantum (LEM 1317, 1163)
Leishmania major (LEM 1958)
Leishmania major Friedlin

Adapted by Al Ivens from a drawing from
(Wincker et al. 1997 Gene 194:77-80





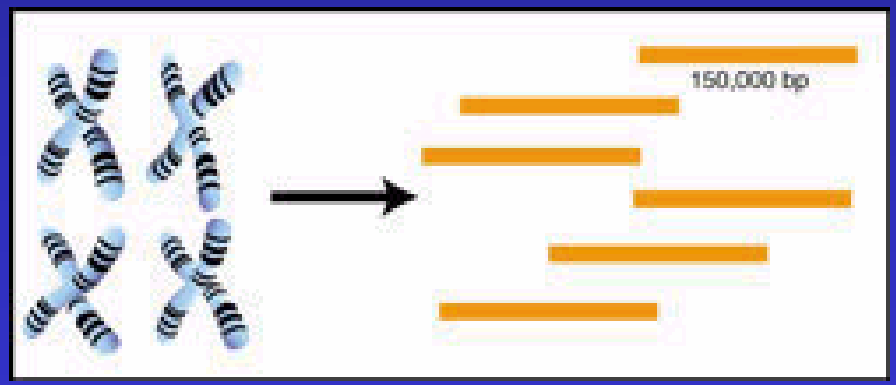



human DNA
CLEAVE WITH RESTRICTION NUCLEASE
millions of genomic DNA fragments
DNA FRAGMENTS INSERTED INTO PLASMIDS
recombinant DNA molecules
INTRODUCTION OF PLASMIDS INTO BACTERIA
genomic library

BAC to BAC Sequencing

The BAC to BAC approach first creates a crude physical map of the whole genome before sequencing the DNA. Constructing a map requires cutting the chromosomes into large pieces and figuring out the order of these big chunks of DNA before taking a closer look and sequencing all the fragments

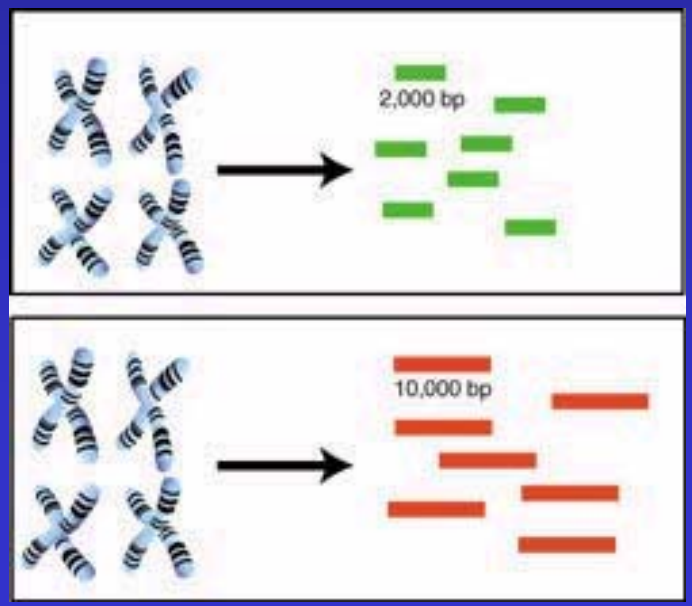
Several copies of the genome are randomly cut into pieces that are about 150,000 base pairs (bp) long.



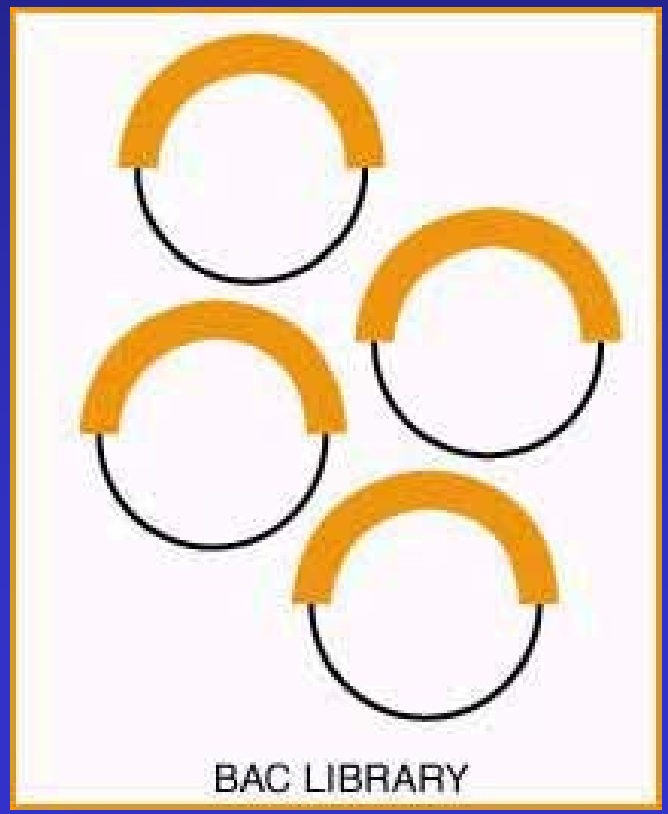
Whole Genome Shotgun Sequencing

The shotgun sequencing method goes straight to the job of decoding, bypassing the need for a physical map. Therefore, it is much faster.

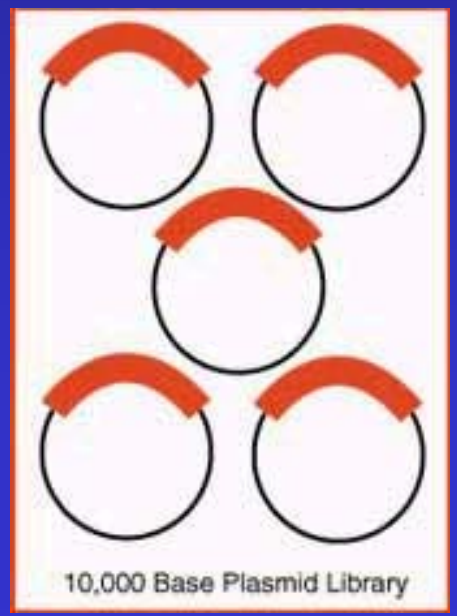
Multiple copies of the genome are randomly shredded into pieces that are 2,000 base pairs (bp) long by squeezing the DNA through a pressurized syringe. This is done a second time to generate pieces that are 10,000 bp long.



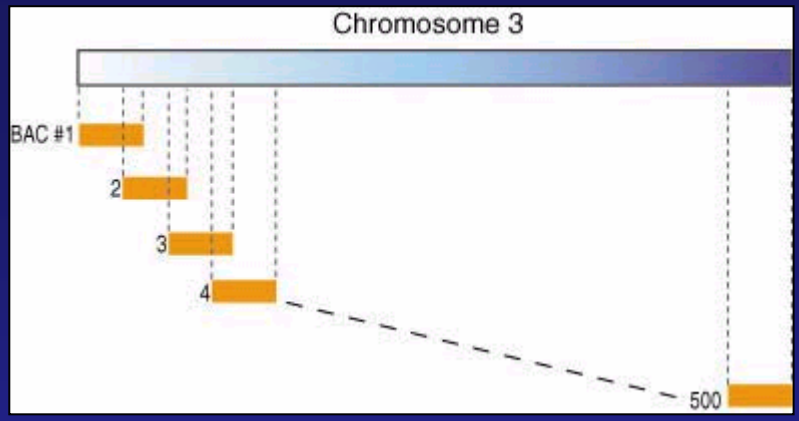
Each of these 150,000 bp fragments is inserted into a BAC-a bacterial artificial chromosome. A BAC is a man made piece of DNA that can replicate inside a bacterial cell. The whole collection of BACs containing the entire human genome is called a BAC library, because each BAC is like a book in a library that can be accessed and copied.



Each 2,000 and 10,000 bp fragment is inserted into a plasmid, which is a piece of DNA that can replicate in bacteria. The two collections of plasmids containing 2,000 and 10,000 bp chunks of human DNA are known as plasmid libraries.

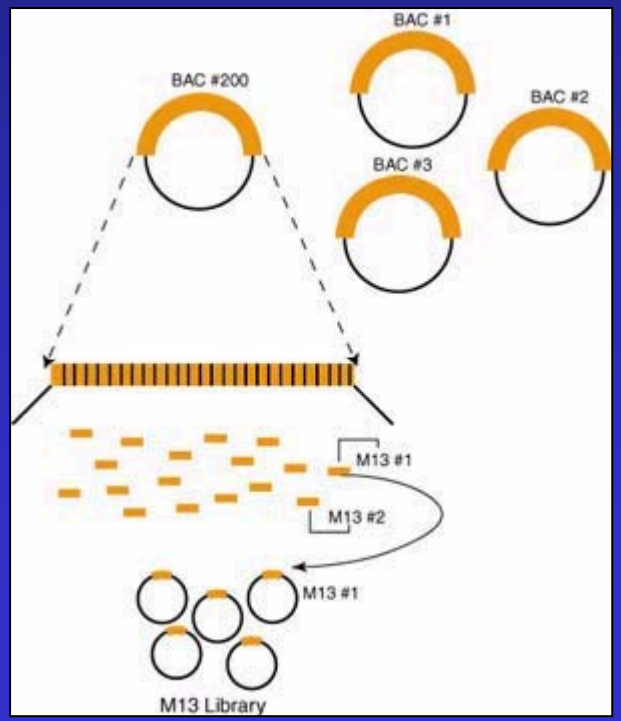


BAC sequencing



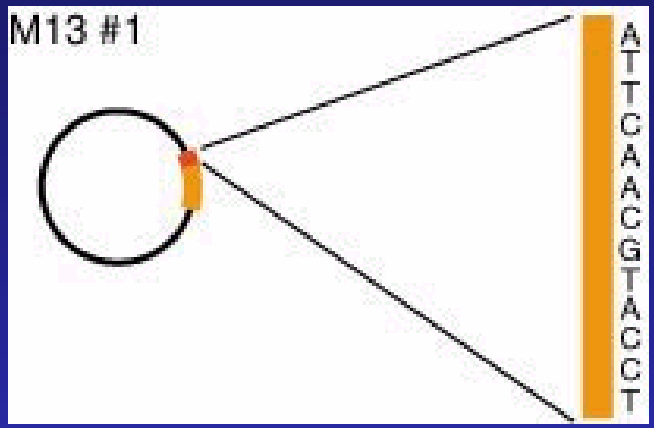
These pieces are fingerprinted to give each piece a unique identification tag that determines the order of the fragments. Fingerprinting involves cutting each BAC fragment with a single enzyme and finding common sequence landmarks in overlapping fragments that determine the location of each BAC along the chromosome. Then overlapping BACs with markers every 100,000 bp form a map of each chromosome.

OR : BAC end sequencing and map assembly



Each BAC is then broken randomly into 1,500 bp pieces and placed in another artificial piece of DNA called M13. This collection is known as an M13 library.

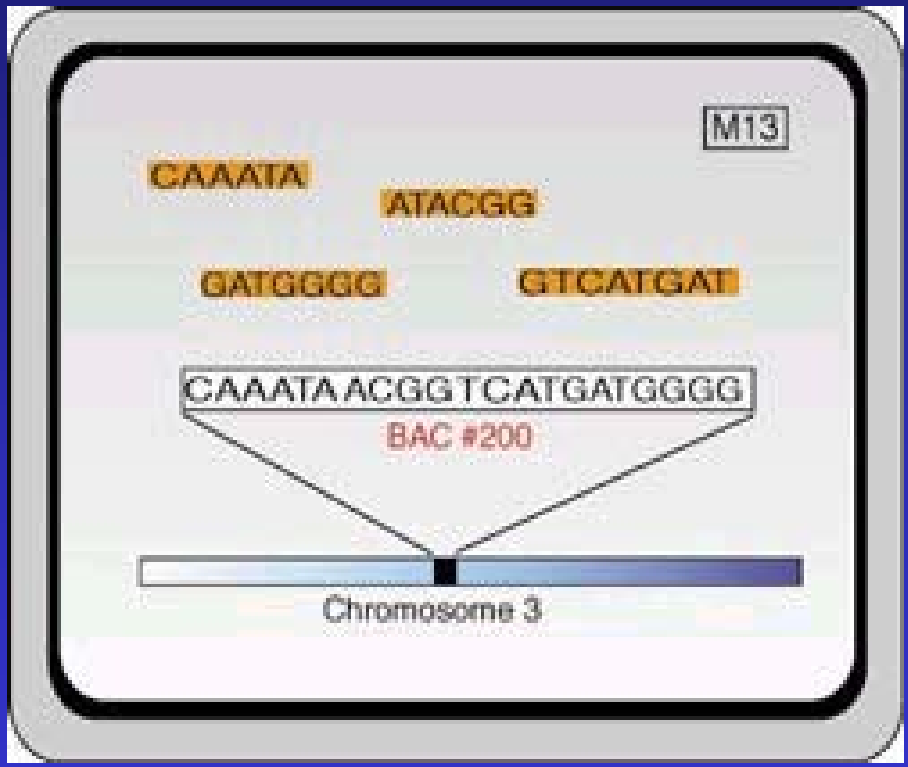
All the M13 libraries are sequenced. 500 bp from one end of the fragment are sequenced generating millions of sequences.

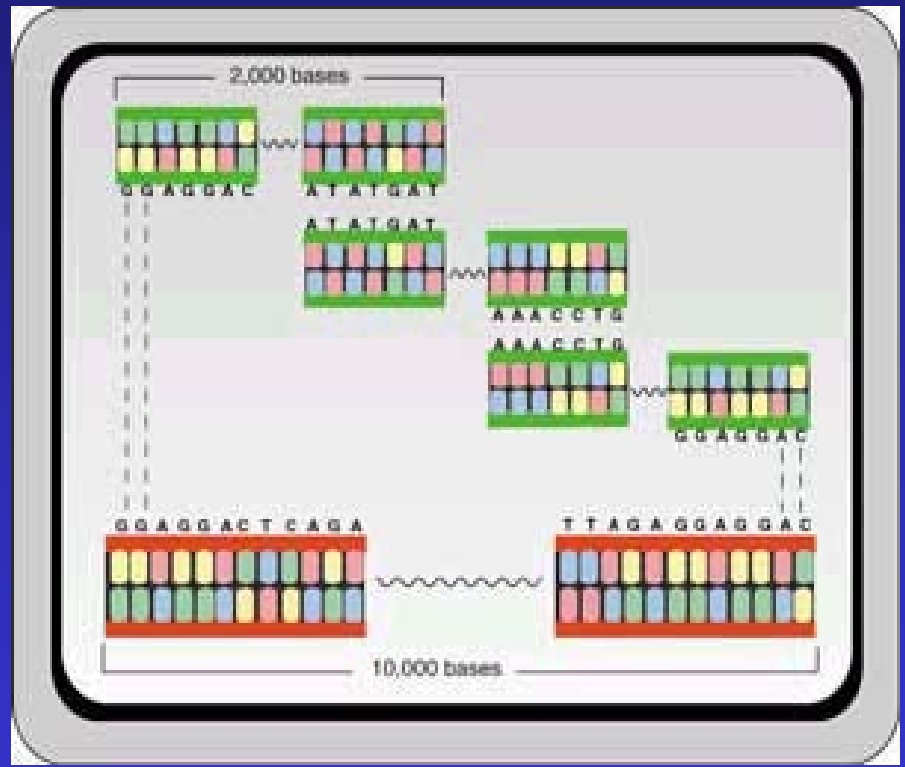


Both the 2,000 and the 10,000 bp plasmid libraries are sequenced. 500 bp from each end of each fragment are decoded generating millions of sequences. Sequencing both ends of each insert is critical for the assembling the entire chromosome.

These sequences are fed into a computer program called PHRAP that looks for common sequences that join two fragments together.

Computer algorithms assemble the millions of sequenced fragments into a continuous stretch resembling each chromosome.

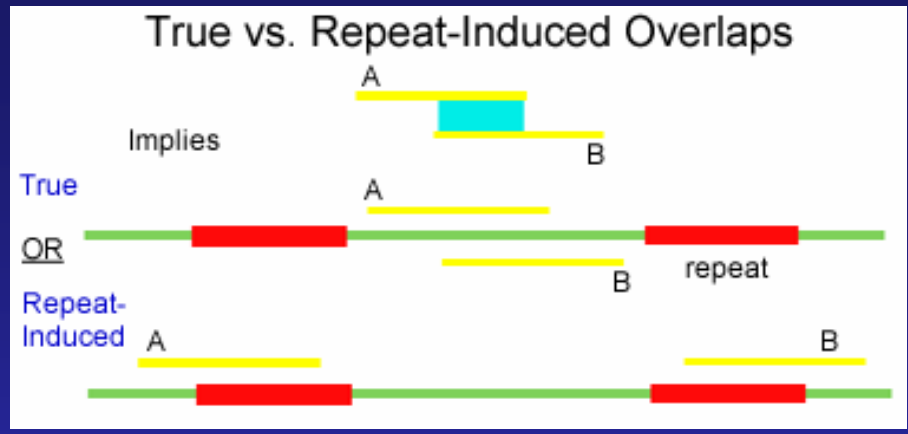


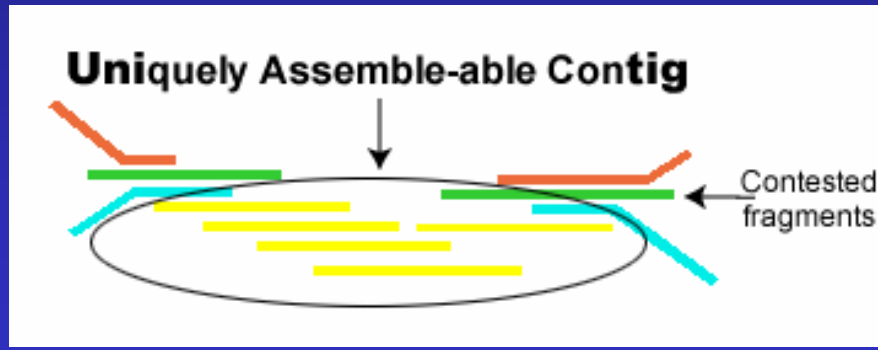


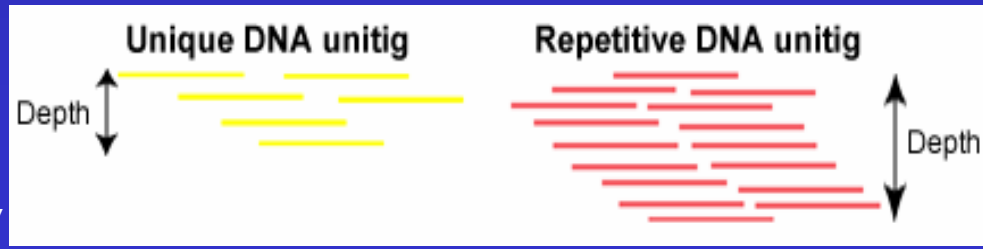
The GNN Assembler is actually a pipeline—a series of mathematical steps to sort, edit, and assemble fragments. The steps are stages in a layered strategy.

The first stage in assembly is the heavy-lifting: The assembler compares the millions of fragments against each other, finding all common segments between two fragments that are at least 40 letters long. These overlaps could not have occurred by chance, and they become the foundation of assembly.

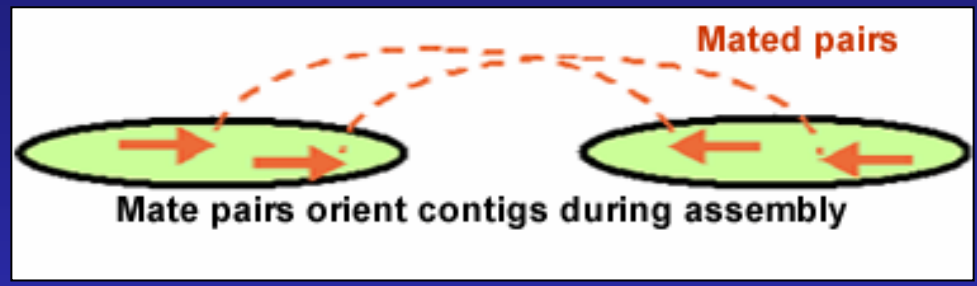
Of these overlaps, some are "true" and some are "repeat-induced." In true overlaps, the shared sequence involves fragments that come from overlapping sections of the genome. These fragments belong together. In repeat-induced overlaps, the shared sequence involves part of a repeat that occurs in several dispersed parts of the genome. These fragments do not belong together. If it were clear which overlaps were true, assembly would be a trivial matter.



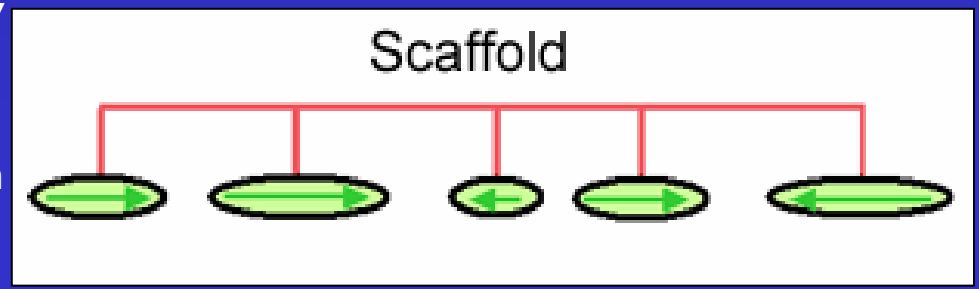




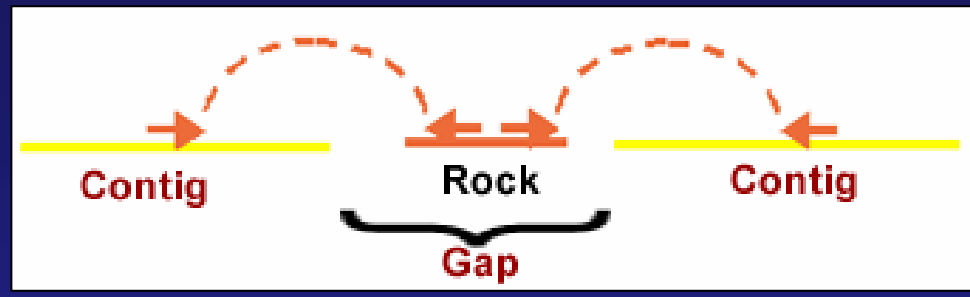
The U-unitigs are mini-phrases that are ready to be ordered in the genome. The scaffolding stage begins. Critical to this stage is the fact that most of the fragments were grabbed from the genome in pairs during sequencing. Known as mate pairs, these fragments are always separated by the same number of letters, either about 1,000 or about 9,000. Since most repeats are shorter than 7,000 letters, mates are a way to circumnavigate, or span, the repeats. However, about 1% of the time mate pairs are not actually paired at the given distance due to errors in the computer tracking of the fragments.



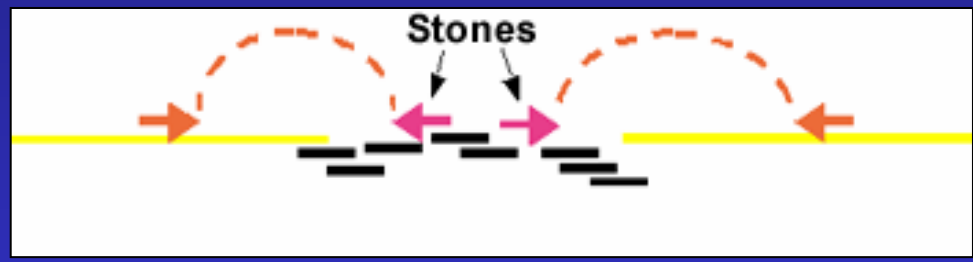
A contiguous sequence of ordered unitigs is a contig. During scaffolding, the assembler orients contigs using mates. Most mate pairs are reliable landmarks—they stick together and remain the same distance apart. If mates from the same pair lie on different contigs, for instance, the contigs are likely to be neighbors about 99% of the time. If two or more mate pairs enforce each other—that is, they indicate the same orientation—then the contigs involved are almost certain to be neighbors.



Some of these gaps are due to missing sequence; this is unavoidable. Other gaps contain repetitive sequence that can now be closed using the unitigs that were set aside earlier by the Discriminator. The same strategy — make progressively riskier moves — applies to closing gaps.

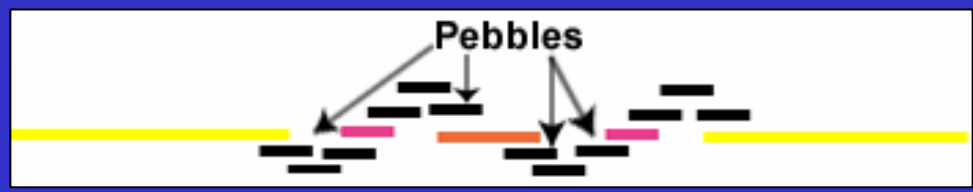


The assembler classifies repeat sequences by size and reliability, calling the largest and most reliable repeats "rocks." Rocks are tossed into the gaps first, to be followed by the lesser "stones," and finally the smallest and least reliable pieces, "pebbles." Rocks must be linked to the contigs on either side of a gap by two or more mates.

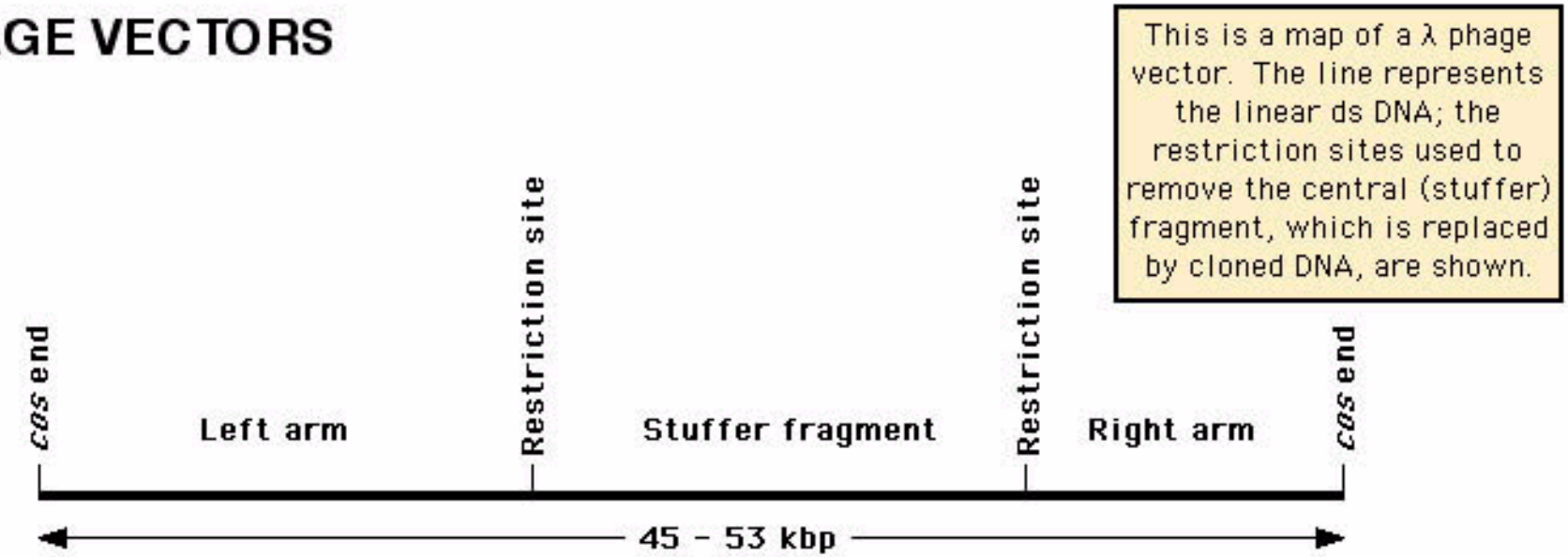


Stones are linked to the contigs by only one mate. Their position in a gap is confirmed by overlaps.

Pebbles are placed in a gap based on the quality of the overlaps between each other and the adjoining contigs



λ PHAGE VECTORS

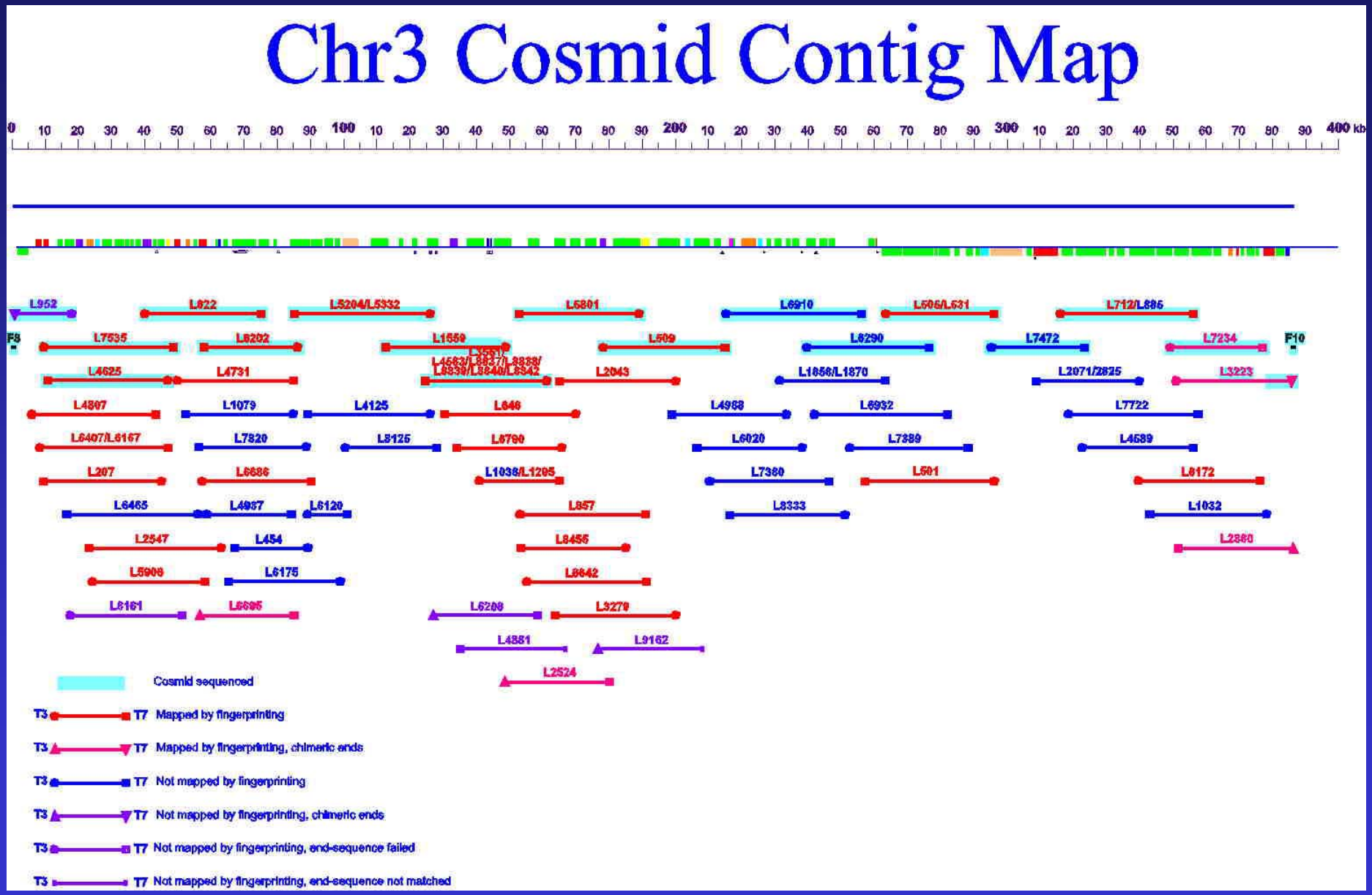


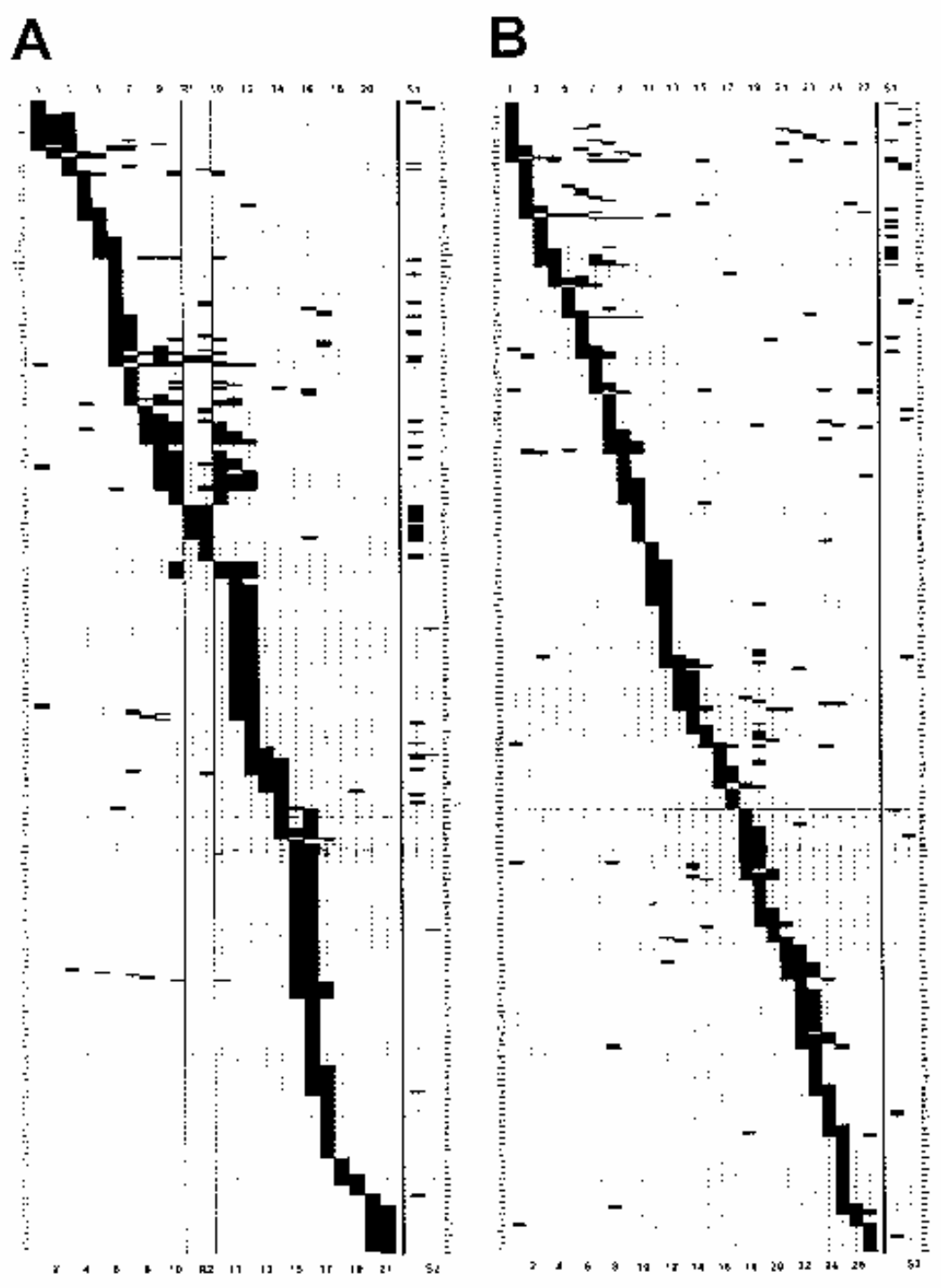
- λ phage vectors generally have a central stuffer fragment which is replaced by the DNA to be cloned.
- The central stuffer fragment contains genes used in the lysogenic phase of the life cycle – i.e. to cause the phage to integrate into the host genome, and to maintain the integrated state.
- λ phage cloning vectors thus generally only undergo the lytic part of the normal phage life cycle.
- When replacing the stuffer fragment, the DNA fragments to be cloned must allow functional phage to be produced; there is thus a restricted size range of fragments that can be accepted
- Vectors designed for cloning genomic DNA generally accept fragments in the range 12–20 kbp
- Vectors designed for cloning cDNA accept fragments in the range 0–10kbp
- Efficiency of host infection (transfection) is not dependent on insert size, provided that phage in the correct size range results

GENOMIC LIBRARIES AND cDNA LIBRARIES

GENOMIC LIBRARIES	cDNA LIBRARIES (ESTs)
Cloned DNA is directly from genome	Cloned DNA fragments are copies of mRNA sequences
Generally only library necessary for prokaryotes – expressed sequences and encoded proteins can be easily deduced	Both types of library necessary for eukaryotes – presence of large amounts of intergenic DNA and complex structure of genes hinders gene characterisation
Cloned fragments contain both transcribed and non-transcribed regions (i.e. include promoters, etc.)	Cloned fragments contain only transcribed regions
Cloned fragments contain introns and exons	Cloned fragments contain exons only
Cloned fragments have no poly(A) "tails"	Cloned sequences have poly(A) "tails" (and other mRNA modifications)
Library is the same when made from DNA of any tissue or developmental stage of organism	Libraries made from different tissues and/or developmental stages of organism differ because different genes are transcribed into mRNA
Need to contain more clones to represent entire genome ($>10^6$)	Need to contain fewer clones to represen all mRNA sequences present ($<10^6$)

Leishmania major Friedlin genome sequencing





Cosmid libraries provide insert fragments of 30-40 kb. They can be used for mapping (fingerprinting), for sequencing, and for transformation and transient expression or chromosome integration.

Minimal tiling pattern for T. cruzi CL-Brener chrom 3 and 4

	T. cruzi CL-Brener	L. major Friedlin	T. brucei TREU927/4	P. falcipa-rum (3D7)	T. gondii ME49/B7
Genome size (hapl)	45 Mb	33.6 Mb ~8000genes?	30 Mb	23 Mb 5300 genes	30 Mb
# chrom.	34?	36	11 (1-6M) ?(200-900K) 50-100 (25-100K)	14 (0.64-3.29 Mb)	9 (2-7 Mb)
# EST seq	10201		5133		
# GSS seq	26693 (~10 Mb)		96474 (46Mb)		
Full chrom	3 (~)	1,3,4,5,13, 14,19,21,23,35	I,II,IX,X, XI	80.6 % AT 54% genes with introns	
Complete	2005?	2003	2003	Sept.2002	2002/2003

Phrap - Phragment Assembly Program

Phrap é um programa para a montagem de sequências obtidas por shotgun.

Características principais:

a. Usa todo o "read" – não há necessidade de "trimming".

b. Utiliza dados fornecidos pelo usuário (i.e. Repbase) + dados computados internamente – maior precisão na montagem na presença de repetições.

c. A sequência do contig é constituída por um mosaico das regiões de alta qualidade dos reads – não é um consenso!!

Phrap - Phragment Assembly Program
ou... Phil's Revised Assembly Program!!

d. Proporciona extensa informação sobre a montagem – esta informação está contida nos arquivos phrap.out, *.ace and *.screen.contigs.qual.

e. Consegue gerenciar grandes conjuntos de dados – centenas de milhares de reads são facilmente manipulados.

f. Gera vários arquivos de saída – os quais contém dados importantes além de possibilitar a visualização da montagem por outros programas.

Arquivos de saída do Phrap

- *.contigs – arquivo no formato FASTA contendo os contigs
  - Contigs com mais de um read
  - Singletons (reads únicos com um match com outro contig mas que não puderam ser consistentemente fundidos ao mesmo).
- *.singlets – arquivos FASTA dos singlets
  - Reads com nenhum match a outro read
- *.ace – permite a visualização da montagem usando Consed
- *.view – necessário para a visualização da montagem usando Phrapview

Outros programas para a montagem:

GASP - Wendl MC, Dear S, Hodgson D and Hillier L. Automated sequence preprocessing in a large-scale sequencing environment. *Genome Research* 8:975-984 (1998).

CAP3 - Huang X & Madan A. CAP3: A DNA sequence assembly program. *Genome Research* 9:868-877 (1999).

GAP - Bonfield JK, Smith KF and Staden R. A new DNA sequence assembly program. *Nucleic Acids Research* 23(24):4992-4999 (1995).

Consed

Consed é um programa para a visualização e edição das montagens produzidas pelo Phrap.

Características principais:

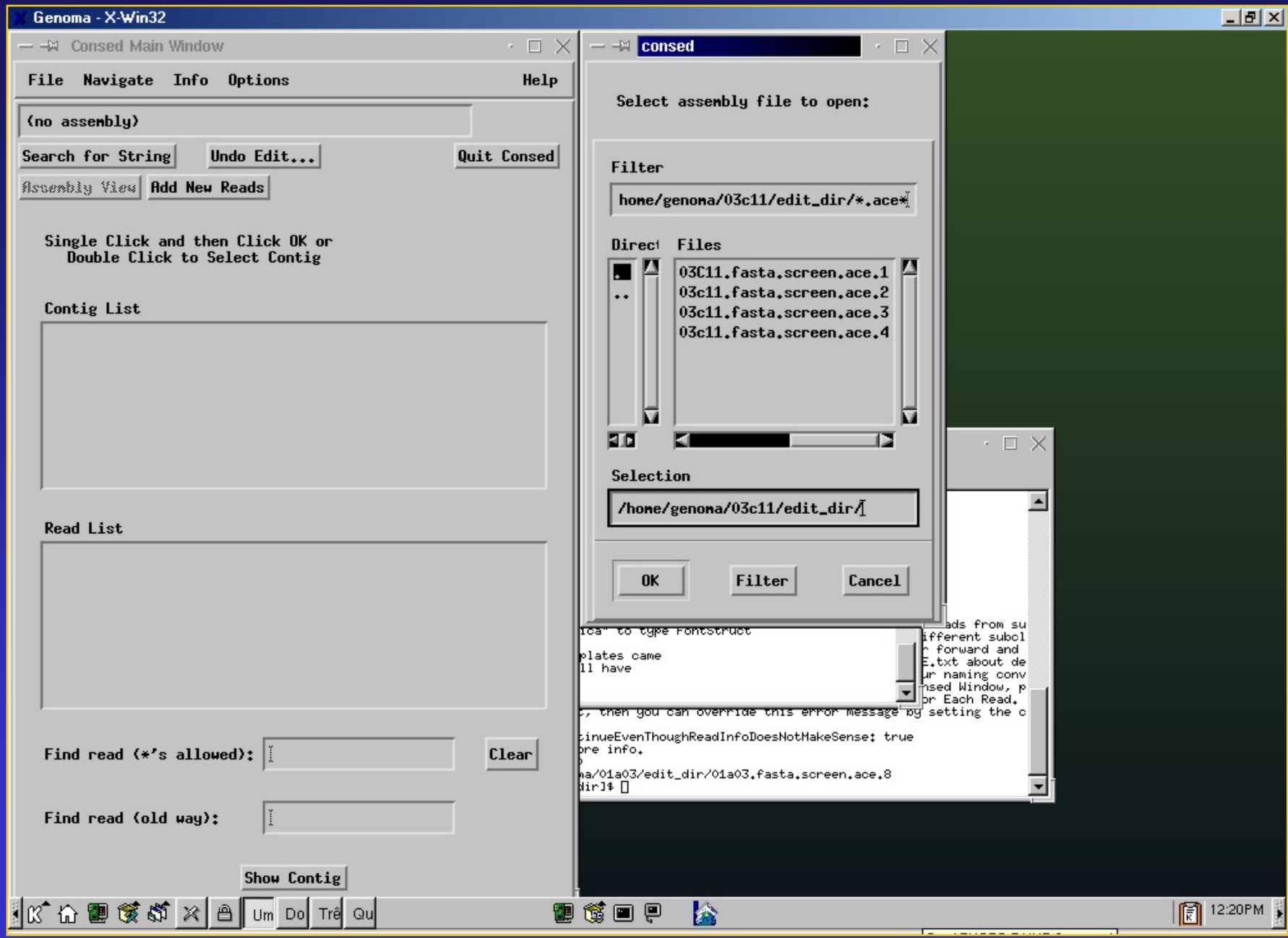
a. Visualizador da montagem - permite a visualização dos contigs, da montagem (reads alinhados), "quality values" dos reads e a sequência final.

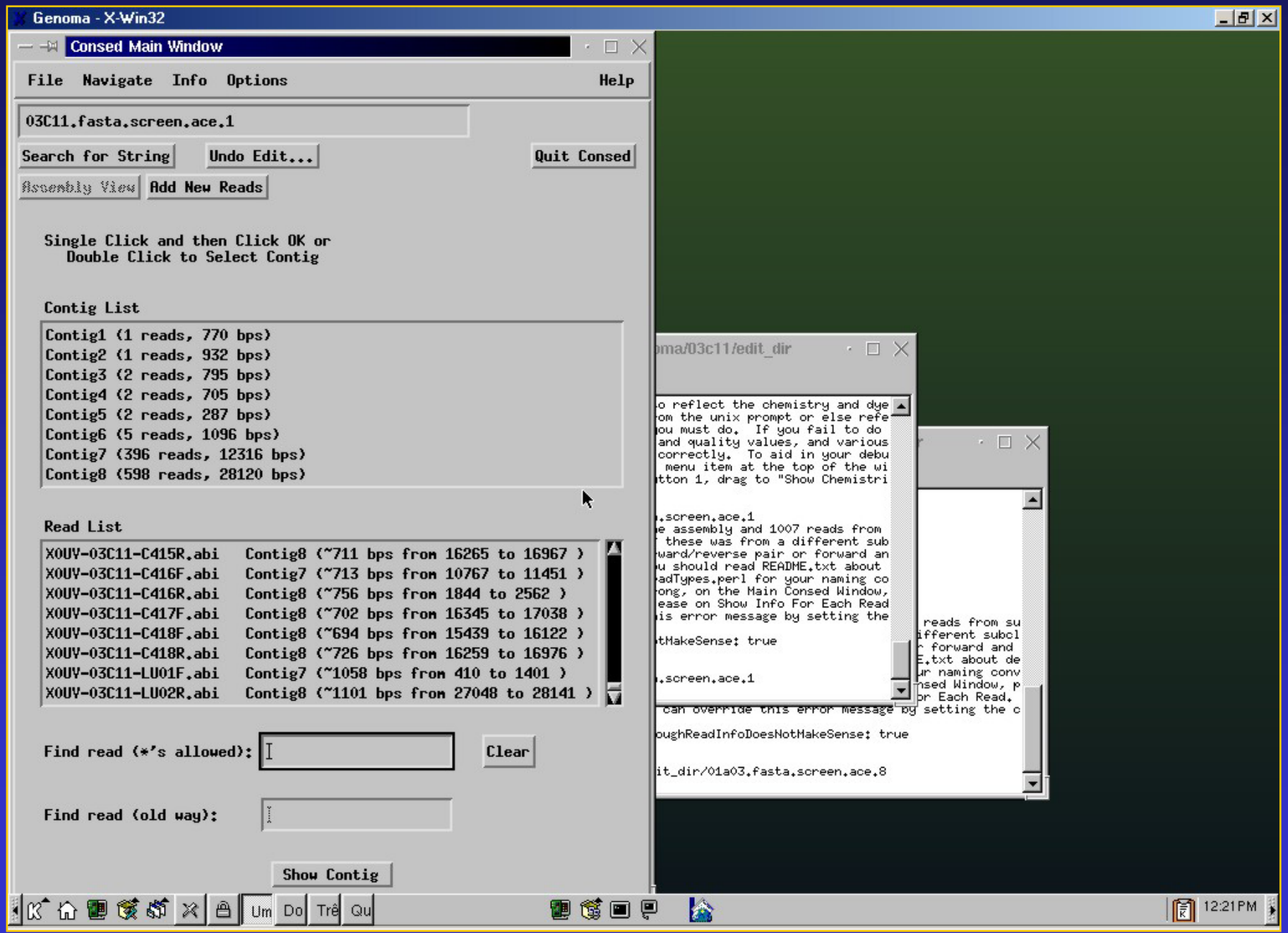
b. Visualizador dos trace files – múltiplos trace files podem ser visualizados permitindo a comparação de uma dada sequência em vários reads.

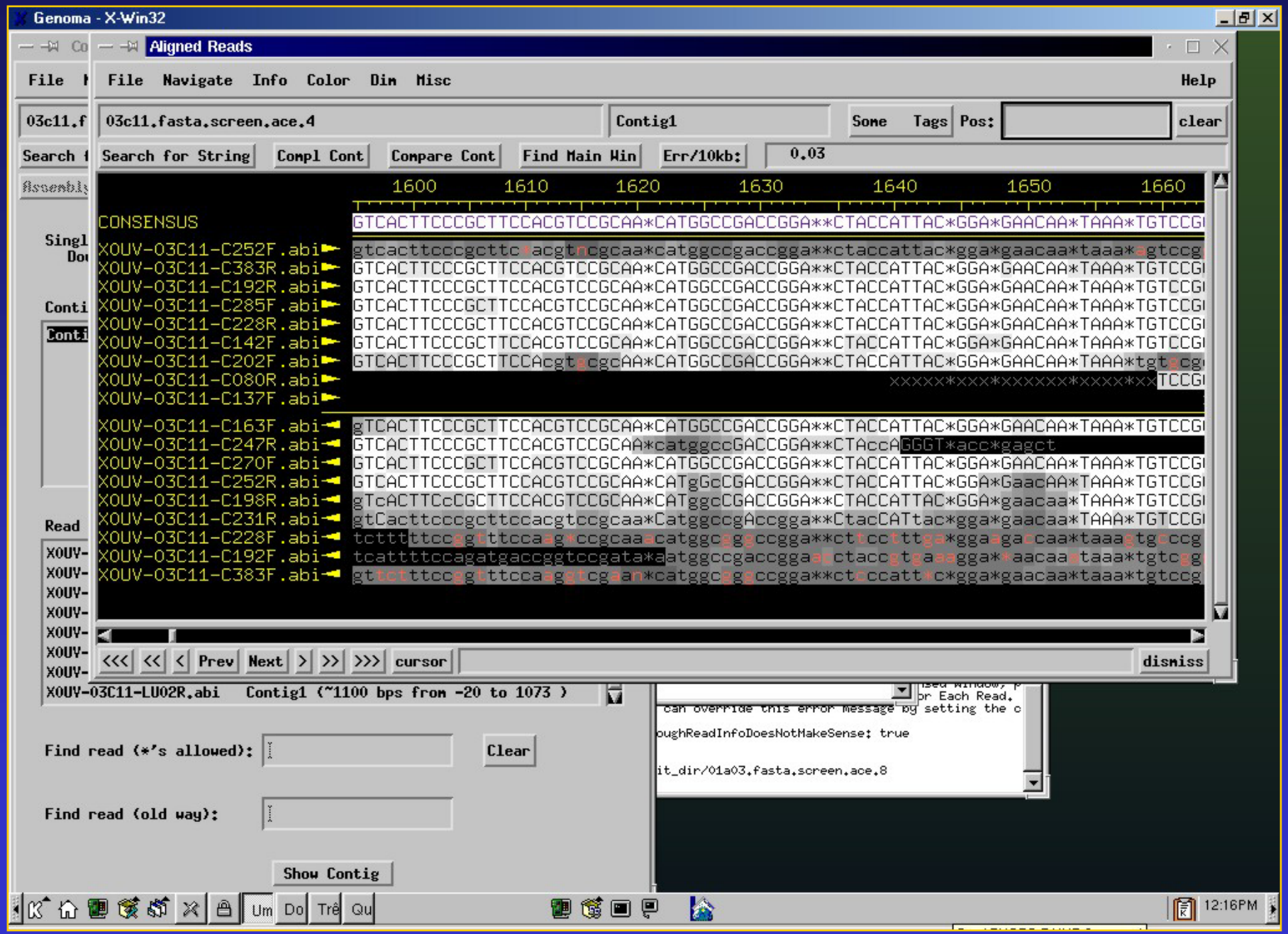
c. Navegação – identifica e lista regiões que estão abaixo de um determinado limite de qualidade, regiões que contém discrepâncias de alta qualidade, cobertura por uma única fita, etc.

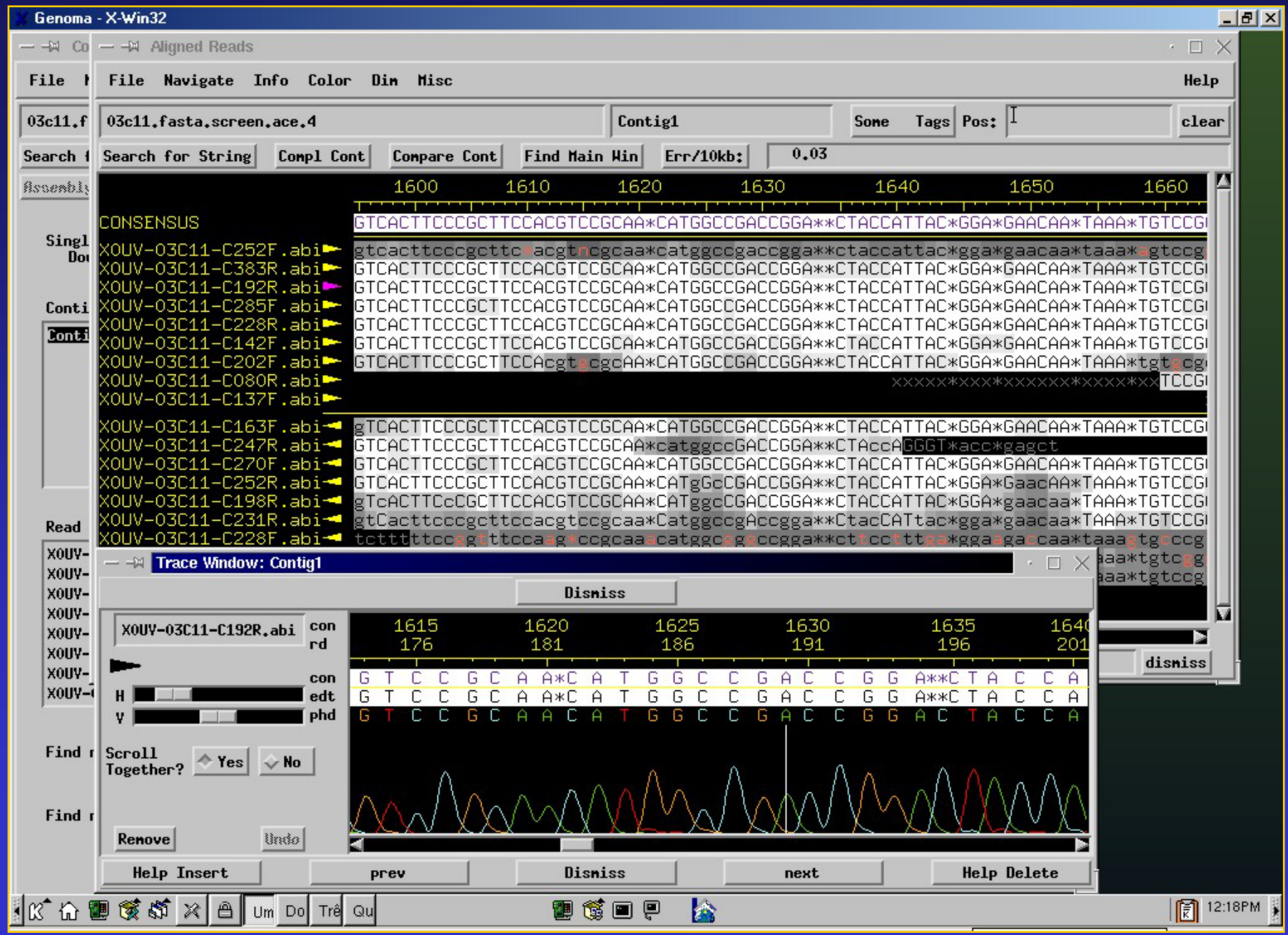
d. Autofinish – conjunto automático de funções para o fechamento de gaps, melhora na qualidade da sequência, determinação da orientação relativa dos contigs, identificação de regiões cobertas por um único read ou por uma única fita.

O programa seleciona automaticamente novos primers e seleciona os templates.


Genoma - X-Win32
Consed Main Window
File Navigate Info Options Help
(no assembly)
Search for String
Undo Edit...
Quit Consed
Assembly View
Add New Reads
Single Click and then Click OK or
Double Click to Select Contig
Contig List
Read List
Find read (*'s allowed):
Clear
Find read (old way):
Show Contig
consed
Select assembly file to open:
Filter
home/genoma/03c11/edit_dir/*.ace*
Direct
Files
03C11.fasta.screen.ace.1
03c11.fasta.screen.ace.2
03c11.fasta.screen.ace.3
03c11.fasta.screen.ace.4
Selection
/home/genoma/03c11/edit_dir/
OK
Filter
Cancel
12:20PM


Genoma - X-Win32
Consed Main Window
File Navigate Info Options Help
03C11.fasta.screen.ace.1
Search for String
Undo Edit...
Quit Consed
Assembly View
Add New Reads
Single Click and then Click OK or
Double Click to Select Contig
Contig List
Contig1 (1 reads, 770 bps)
Contig2 (1 reads, 932 bps)
Contig3 (2 reads, 795 bps)
Contig4 (2 reads, 705 bps)
Contig5 (2 reads, 287 bps)
Contig6 (5 reads, 1096 bps)
Contig7 (396 reads, 12316 bps)
Contig8 (598 reads, 28120 bps)
Read List
X0UV-03C11-C415R.abi Contig8 (~711 bps from 16265 to 16967)
X0UV-03C11-C416F.abi Contig7 (~713 bps from 10767 to 11451)
X0UV-03C11-C416R.abi Contig8 (~756 bps from 1844 to 2562)
X0UV-03C11-C417F.abi Contig8 (~702 bps from 16345 to 17038)
X0UV-03C11-C418F.abi Contig8 (~694 bps from 15439 to 16122)
X0UV-03C11-C418R.abi Contig8 (~726 bps from 16259 to 16976)
X0UV-03C11-LU01F.abi Contig7 (~1058 bps from 410 to 1401)
X0UV-03C11-LU02R.abi Contig8 (~1101 bps from 27048 to 28141)
Find read (*'s allowed):
Clear
Find read (old way):
Show Contig
oma/03c11/edit_dir
o reflect the chemistry and dye
om the unix prompt or else refe
ou must do. If you fail to do
and quality values, and various
correctly. To aid in your debu
menu item at the top of the wi
tton 1, drag to "Show Chemistri
.screen.ace.1
e assembly and 1007 reads from
these was from a different sub
ward/reverse pair or forward an
u should read README.txt about
adTypes.perl for your naming co
ong, on the Main Consed Window,
ease on Show Info For Each Read
is error message by setting the
tMakeSense: true
.screen.ace.1
oughReadInfoDoesNotMakeSense: true
it_dir/01a03.fasta.screen.ace.8
Um Do Tre Qu
12:21PM


Genoma - X-Win32
Aligned Reads
File Navigate Info Color Dim Misc Help
03c11.fasta.screen.ace.4
Contig1
Some Tags
Pos:
clear
Search for String
Compl Cont
Compare Cont
Find Main Win
Err/10kb:
0.03
1600 1610 1620 1630 1640 1650 1660
CONSENSUS
XOUV-03C11-C252F.abi
XOUV-03C11-C383R.abi
XOUV-03C11-C192R.abi
XOUV-03C11-C285F.abi
XOUV-03C11-C228R.abi
XOUV-03C11-C142F.abi
XOUV-03C11-C202F.abi
XOUV-03C11-C080R.abi
XOUV-03C11-C137F.abi
XOUV-03C11-C163F.abi
XOUV-03C11-C247R.abi
XOUV-03C11-C270F.abi
XOUV-03C11-C252R.abi
XOUV-03C11-C198R.abi
XOUV-03C11-C231R.abi
XOUV-03C11-C228F.abi
XOUV-03C11-C192F.abi
XOUV-03C11-C383F.abi
<<< << < Prev Next > >> >>> cursor
dismiss
XOUV-03C11-LU02R.abi Contig1 (~1100 bps from -20 to 1073)
Find read (*'s allowed):
Clear
Find read (old way):
Show Contig
oughReadInfoDoesNotMakeSense: true
it_dir/01a03.fasta.screen.ace.8
12:16PM


Genoma - X-Win32
Aligned Reads
File Navigate Info Color Dim Misc Help
03c11.fasta.screen.ace.4
Contig1
Some Tags
Pos:
clear
Search for String
Compl Cont
Compare Cont
Find Main Win
Err/10kb:
0.03
1600 1610 1620 1630 1640 1650 1660
CONSENSUS
XOUV-03C11-C252F.abi
XOUV-03C11-C383R.abi
XOUV-03C11-C192R.abi
XOUV-03C11-C285F.abi
XOUV-03C11-C228R.abi
XOUV-03C11-C142F.abi
XOUV-03C11-C202F.abi
XOUV-03C11-C080R.abi
XOUV-03C11-C137F.abi
XOUV-03C11-C163F.abi
XOUV-03C11-C247R.abi
XOUV-03C11-C270F.abi
XOUV-03C11-C252R.abi
XOUV-03C11-C198R.abi
XOUV-03C11-C231R.abi
XOUV-03C11-C228F.abi
Trace Window: Contig1
Dismiss
XOUV-03C11-C192R.abi
1615 1620 1625 1630 1635
176 181 186 191 196
Scroll Together?
Yes
No
Remove
Undo
Help Insert
prev
Dismiss
next
Help Delete
dismiss
12:18PM

Phred/Phrap/Consed Pipeline

Diretórios:

- Chromat_dir
- Phd_dir
- Edit_dir

Problemas na finalização (Finishing)

Problemas no sequenciamento:

a. Alto conteúdo GC – genomas apresentando um alto conteúdo GC podem gerar com maior frequência artefatos tais como compressões, quedas bruscas na qualidade, regiões com má qualidade. Tentar utilizar Dye Primer ao invés de Dye Terminator, modificar a química utilizada, adicionar DMSO, aumentar a temperatura de anelamento, usar deaza-dGTP ao invés de dGTP, etc.

b. Regiões palindrômicas – levam a formação de estruturas secundárias causando quedas súbitas. Usar deaza-dGTP ao invés de dGTP, amplificar a região problemática por PCR e sequenciar o produto.

c. Regiões homopoliméricas – podem reduzir a eficiência da síntese de DNA para algumas químicas. Utilizar Dye Primer ao invés de Dye Terminator, modificar a química (dRhodamine ao invés do BigDye).

Problemas na finalização (Finishing)

Problemas na montagem do DNA:

a. Alto conteúdo de repetições – elementos altamente repetitivos reduzem a precisão da montagem do DNA. Deve-se identificar a unidade de repetição, fazer o screening com o programa Cross_Match ou Repeat_Masker e mascará-la. Tentar montá-la novamente e adicionar a sequência repetitiva somente no final. Deve-se mapear a região repetitiva usando enzimas de restrição para estimar o seu tamanho e número de repetições.

b. Alto conteúdo AT no genoma – alguns genomas com grande desvio no seu conteúdo GC (i.e. *Plasmodium falciparum;* genomas de organelas) podem apresentar problemas para os programas de montagem. Muito difícil de se solucionar. Deve-se tentar a construção de um mapa de restrição e associá-lo com os dados do sequenciamento.

Métodos de predição:

Identificação de sinais

- ribosome binding sites
- start/stop codons
- RNA splice sites
- Polyadenylation signals

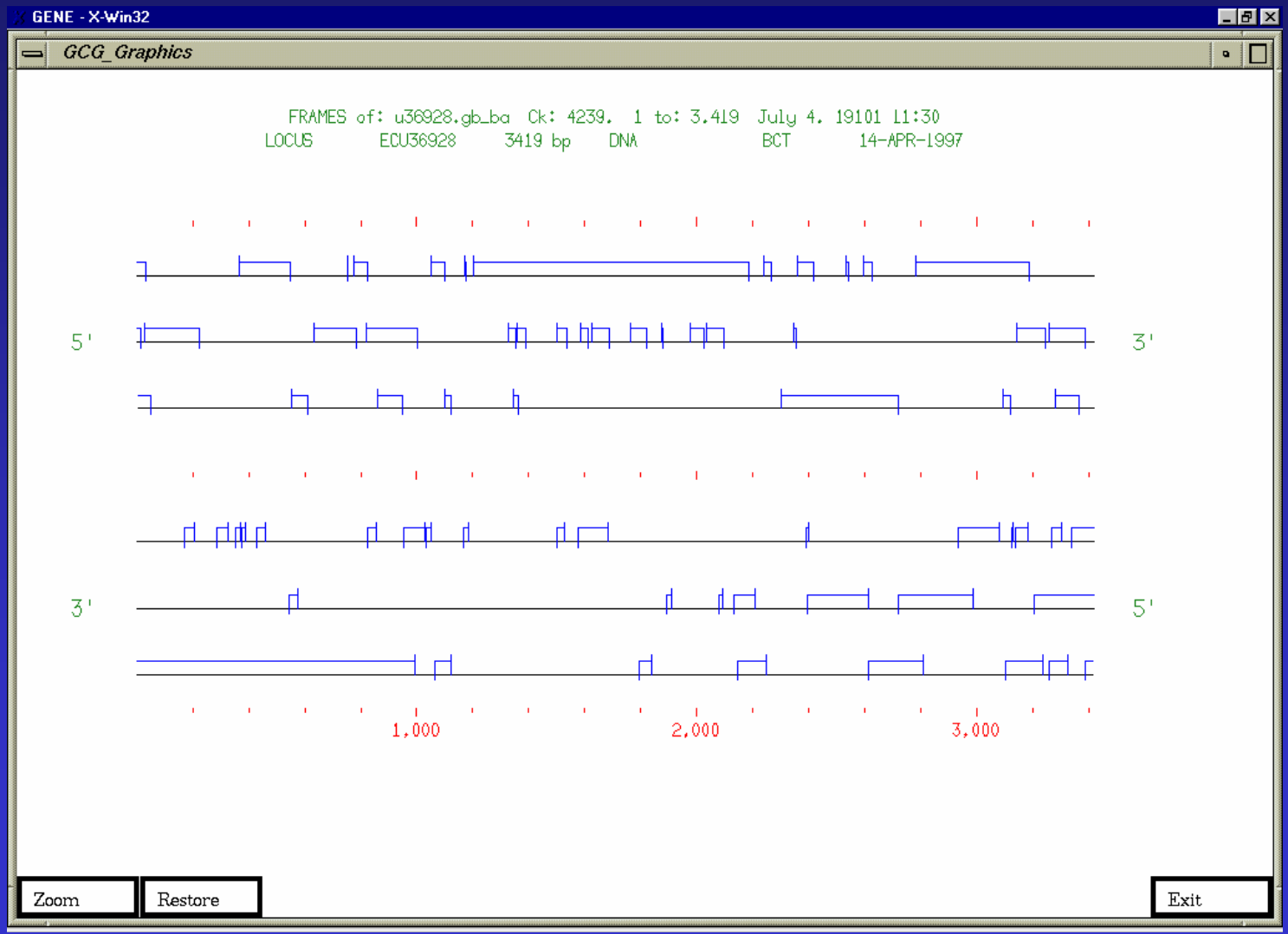
Desvios composicionais

- periodic base composition bias
- terceira posição do códon

Codon bias (codon preference)

Utilização de Markov Chains

Frames (GCG) - detecta open reading frames através da identificação de start/stop códons.



Considerações a respeito do método:

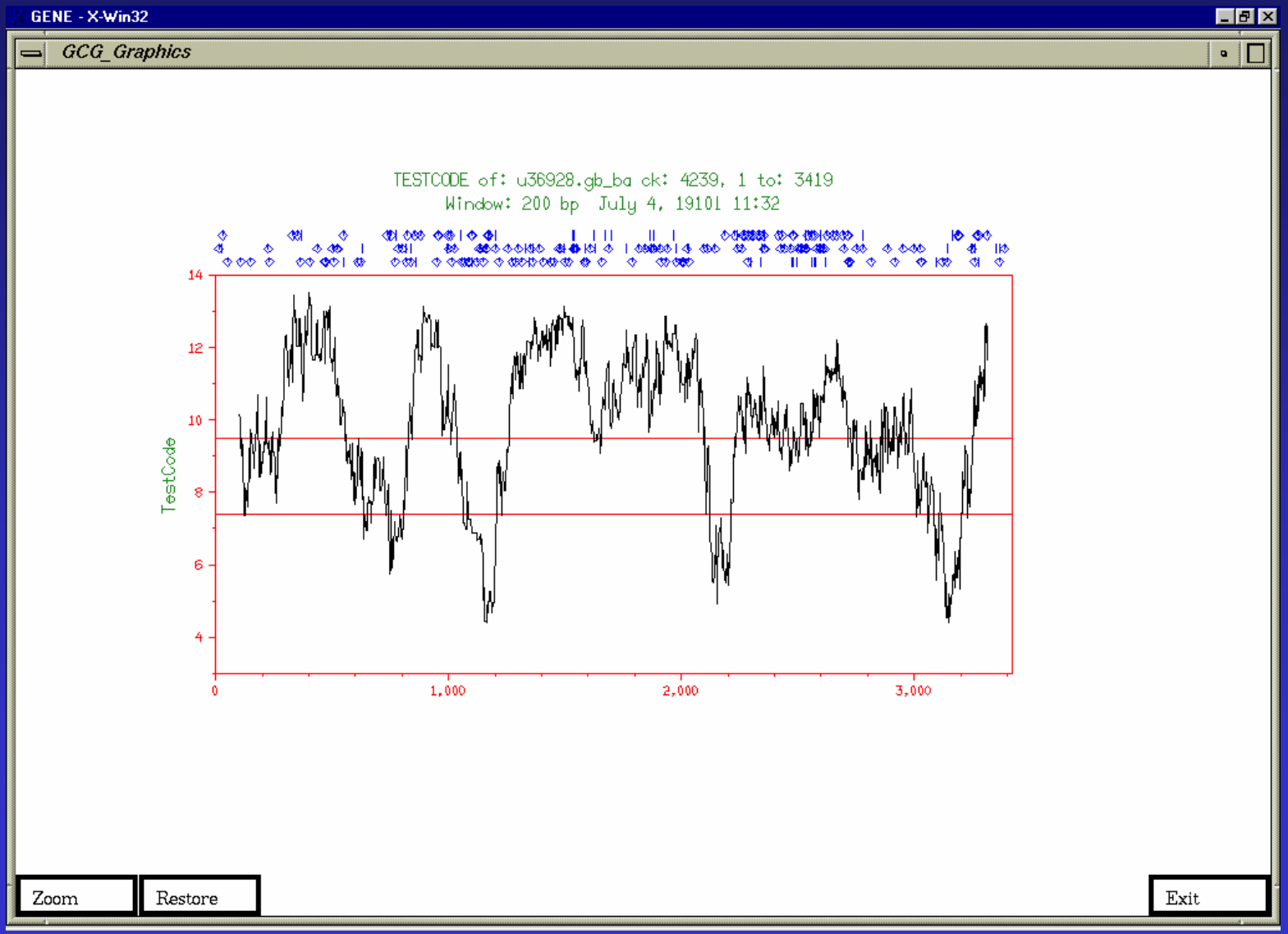
- difícil discriminação entre regiões codificantes e regiões não-codificantes.

- é necessária a identificação de sinais (RBS, início de transcrição/tradução, terminação, limites éxon/íntron), para assinalar a sequência como sendo codificante.

- existência de start códons alternativos.

- em sequências eucarióticas o método pode perder muito em eficiência (éxons/íntrons).

Testcode (GCG) - periodic base composition bias



Considerações a respeito do método:

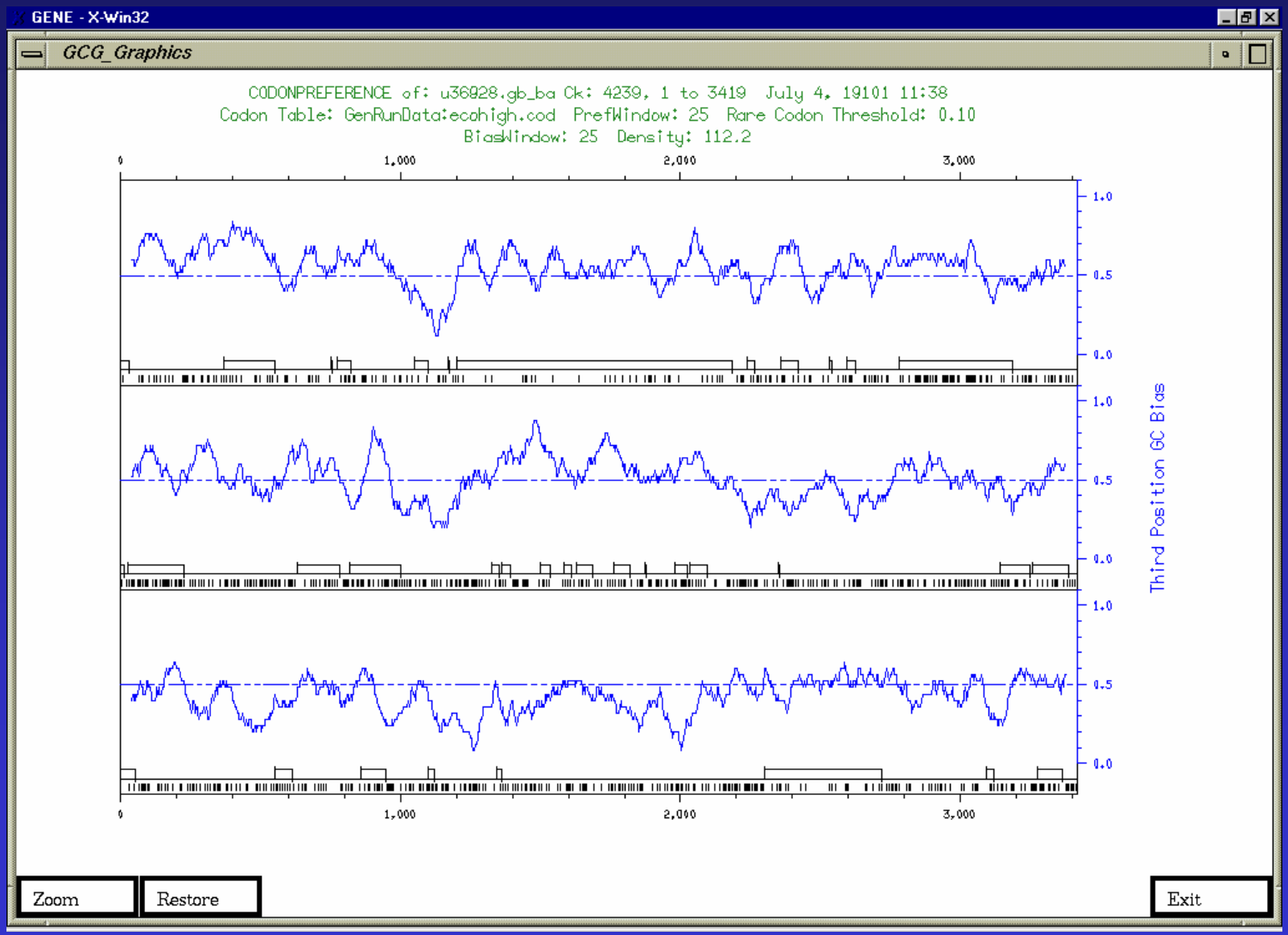
- um dos primeiros a possuir bases estatísticas.

- procura por "assimetrias" ao longo da molécula de DNA: 1º grupo: bases 1, 4, 7, ... 2º grupo: bases 2, 5, 8, ... 3º grupo: bases 3, 6, 9, ...

- não define a fase de leitura nem a fita.

- não determina de forma precisa a região codificante.

Third position GC bias



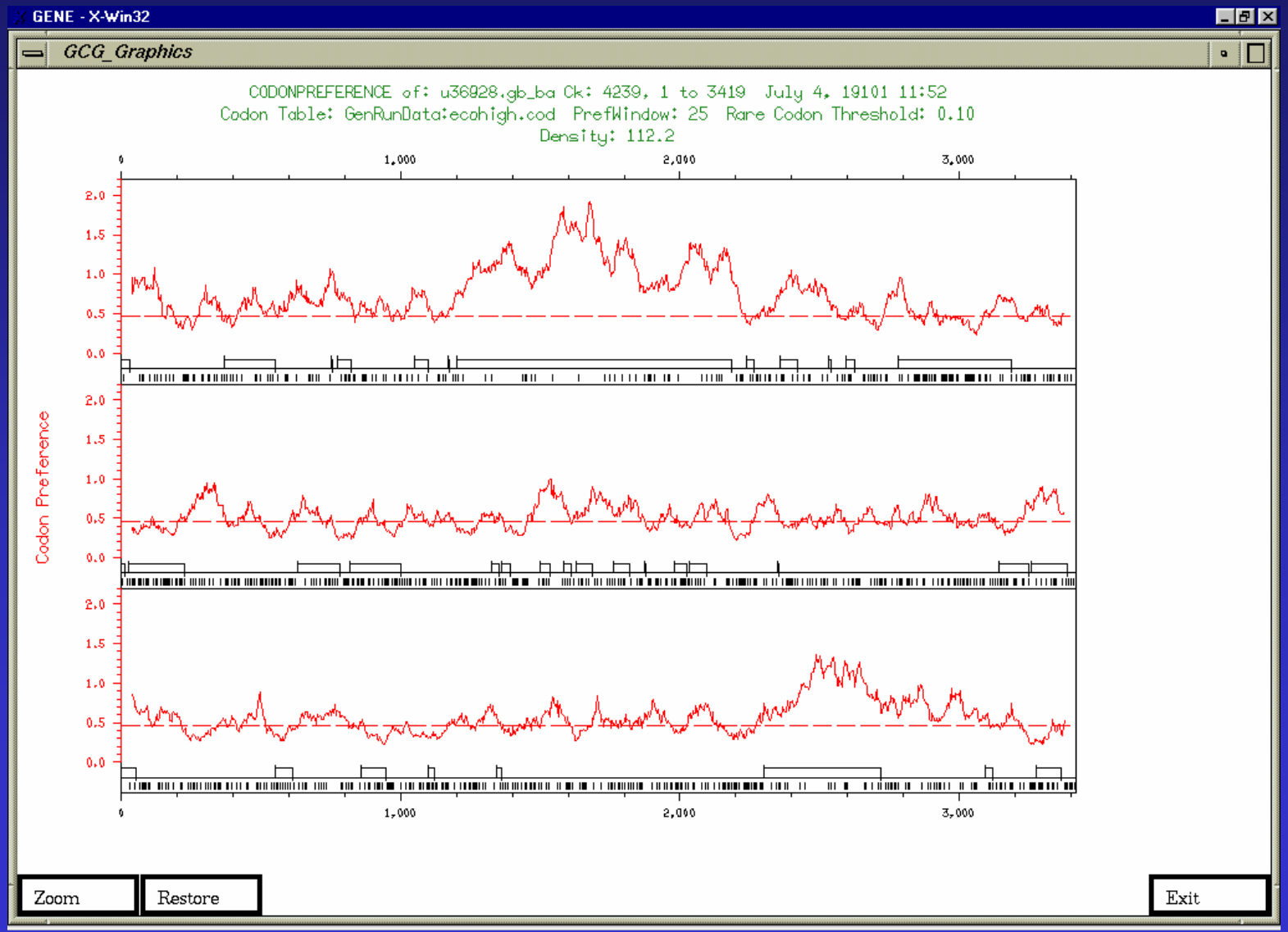
Considerações a respeito do método:

- funciona melhor em organismos com maior desvio composicional em seu genoma (alto/baixo) conteúdo GC.

- difícil discriminação de falsos positivos e/ou falsos positivos.

- proporciona melhores resultados quando usado em conjunto com outros métodos.

Codon preference - comparação com uma tabela de utilização de códons



Considerações a respeito do método:

- diferentes tecidos e diferentes estágios de desenvolvimento do organismo em questão podem apresentar diferentes "pools" de tRNA.

- detecta melhor genes com forte preferência por determinados códons (em geral, genes altamente expressos - seleção traducional).

- útil para a detecção de erros de sequenciamento causando frameshifts.

Glimmer

- utiliza um método estatístico chamado "Interpolated Markov Model" para reconhecer as regiões codificantes.

- Mas o que é isso???

- traduzindo: é a probabilidade de ocorrência de um determinado nucleotídeo, dado um determinado contexto (que é a sequência de bases imediatamente anterior a este nucleotídeo).

- ou ainda: qual a probabilidade da ocorrência de um G depois de um A? Ou depois de um AG?

Considerações sobre o método:

- é o método de escolha para a análise de grandes segmentos de DNA.

- processamento automático de grande eficiência - minimiza a interferência humana.

- alta taxa de acertos: prediz corretamente ~ 96% dos genes, com relativamente poucos falsos positivos.

- pode ser utilizado (com modificações) para a predição de sequências codificantes em genomas eucarióticos.

Blast - uma ferramenta para busca em bancos de dados por similaridade de sequências.

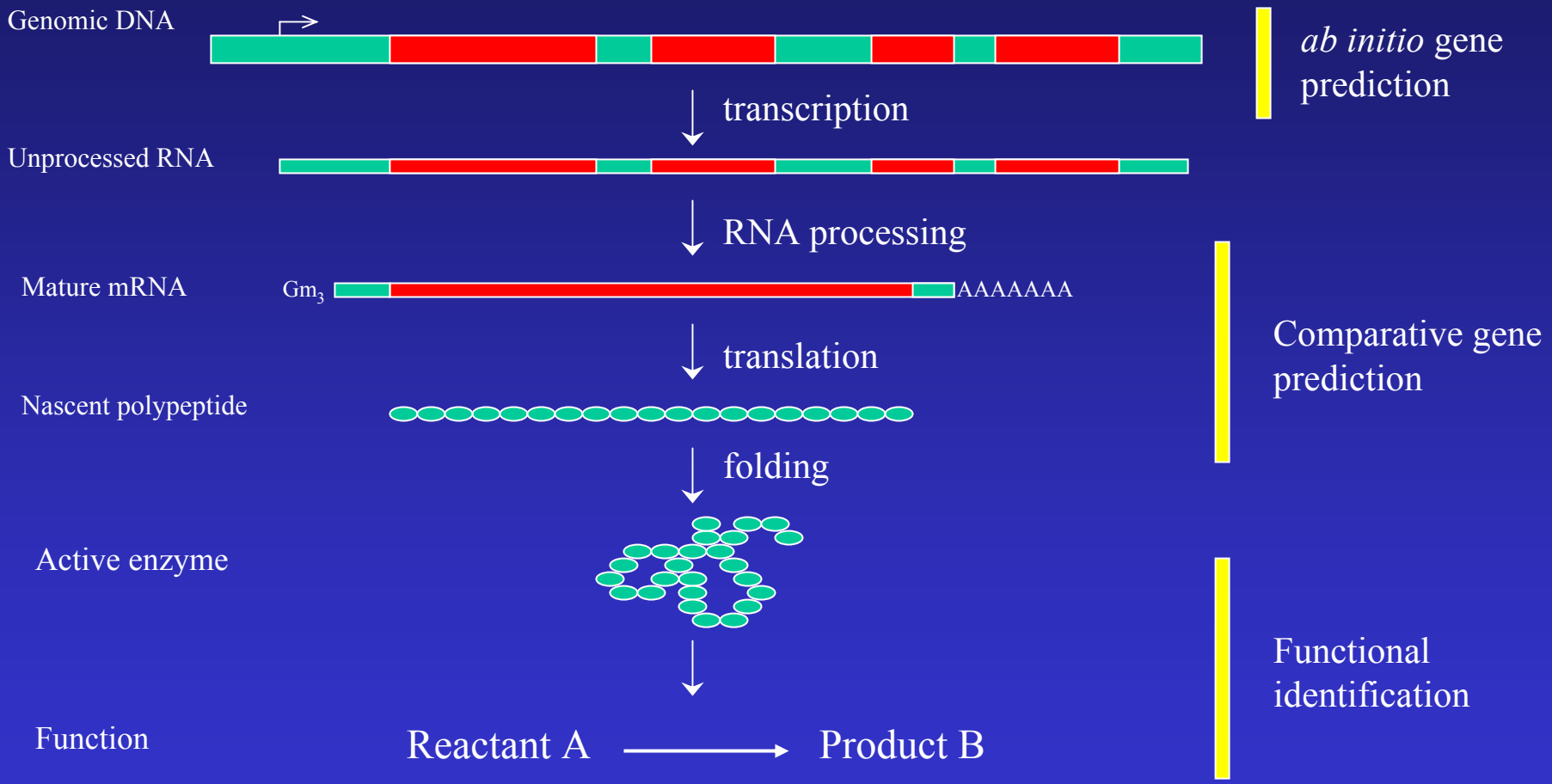
- métodos de busca em bancos de dados: keywords, identificadores, strings, por similaridade de sequência.

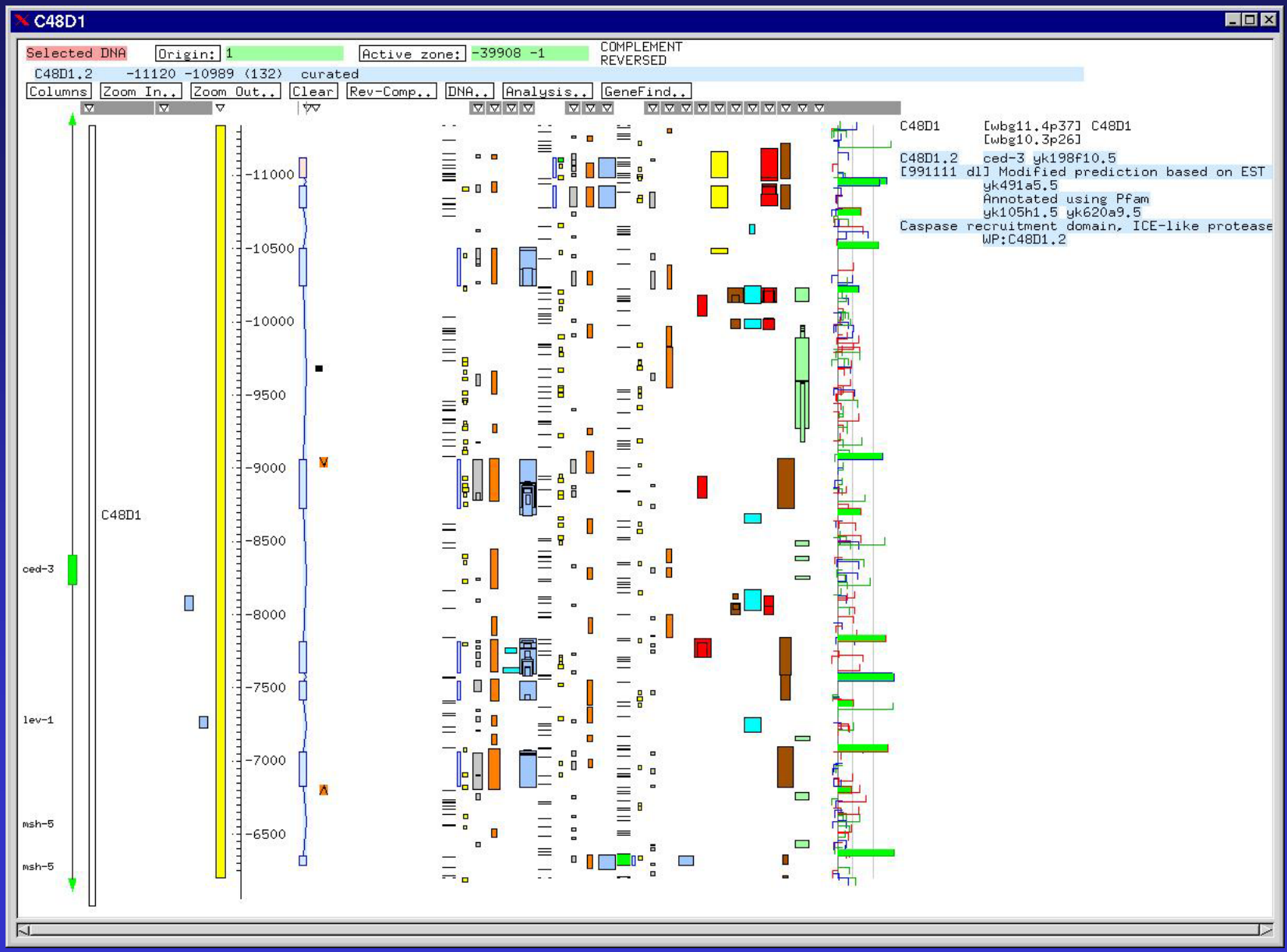
- outros métodos de busca por similaridade: FASTA.

- vantagens do BLAST: rápido e praticamente tão eficiente quanto o FASTA.

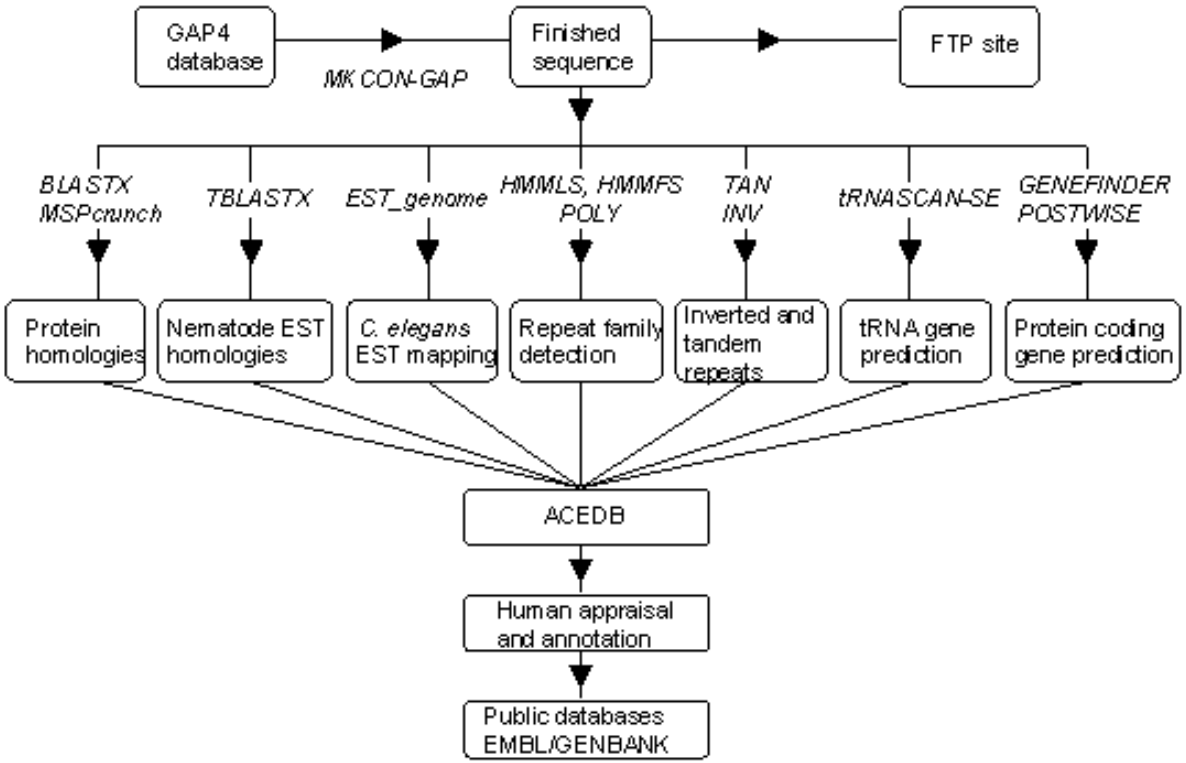
- pode ser realizado via Internet.

Annotation of eukaryotic genomes

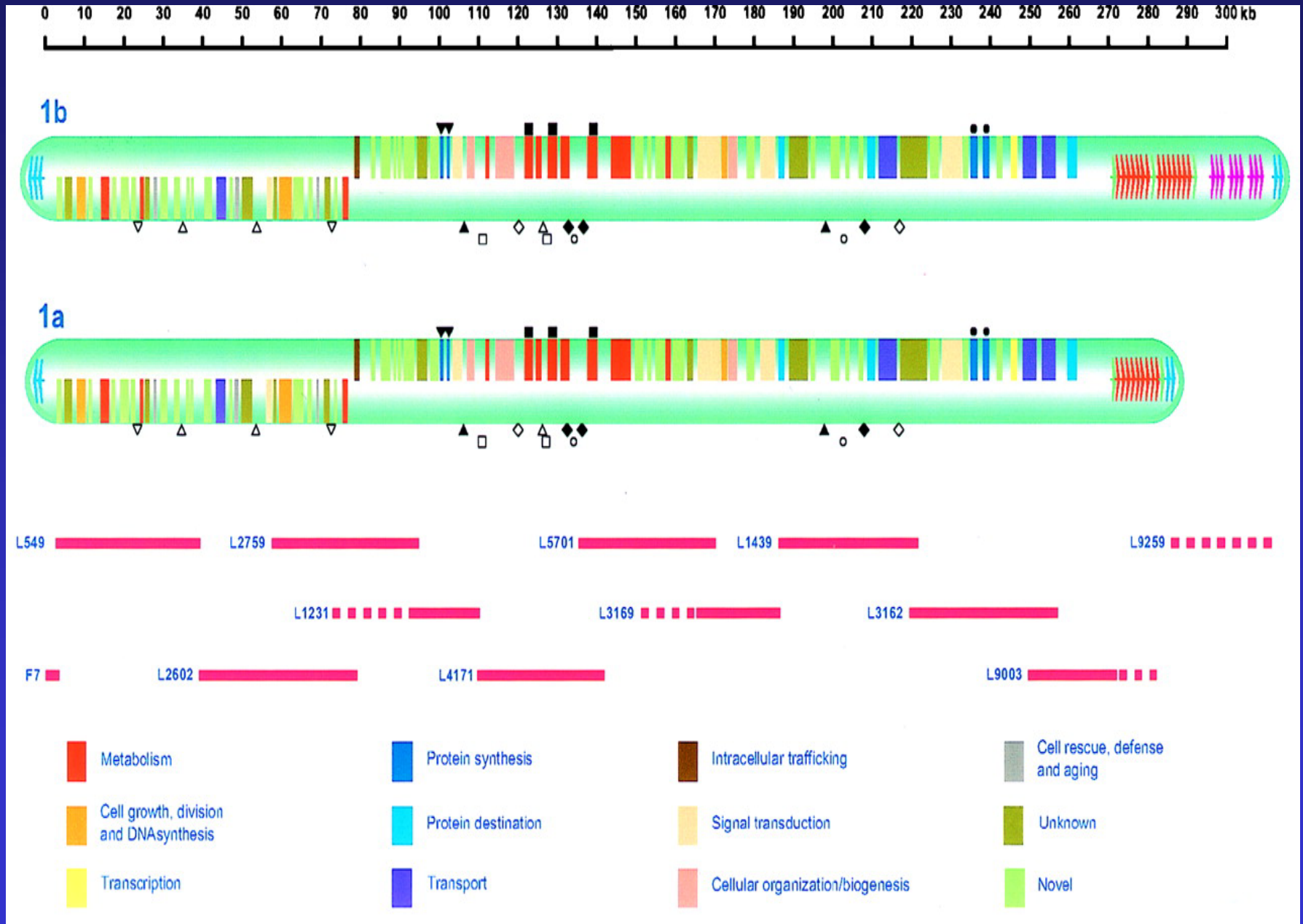



C48D1
Selected DNA
Origin: 1
Active zone: -39908 -1
COMPLEMENT
REVERSED
C48D1.2 -11120 -10989 (132) curated
Columns
Zoom In..
Zoom Out..
Clear
Rev-Comp..
DNA..
Analysis..
GeneFind..
C48D1 [wbg11.4p37] C48D1
[wbg10.3p26]
C48D1.2 ced-3 yk198f10.5
[991111 dl] Modified prediction based on EST
yk491a5.5
Annotated using Pfam
yk105h1.5 yk620a9.5
Caspase recruitment domain, ICE-like protease
WP:C48D1.2
-11000
-10500
-10000
-9500
-9000
-8500
-8000
-7500
-7000
-6500
C48D1
ced-3
lev-1
msh-5
msh-5

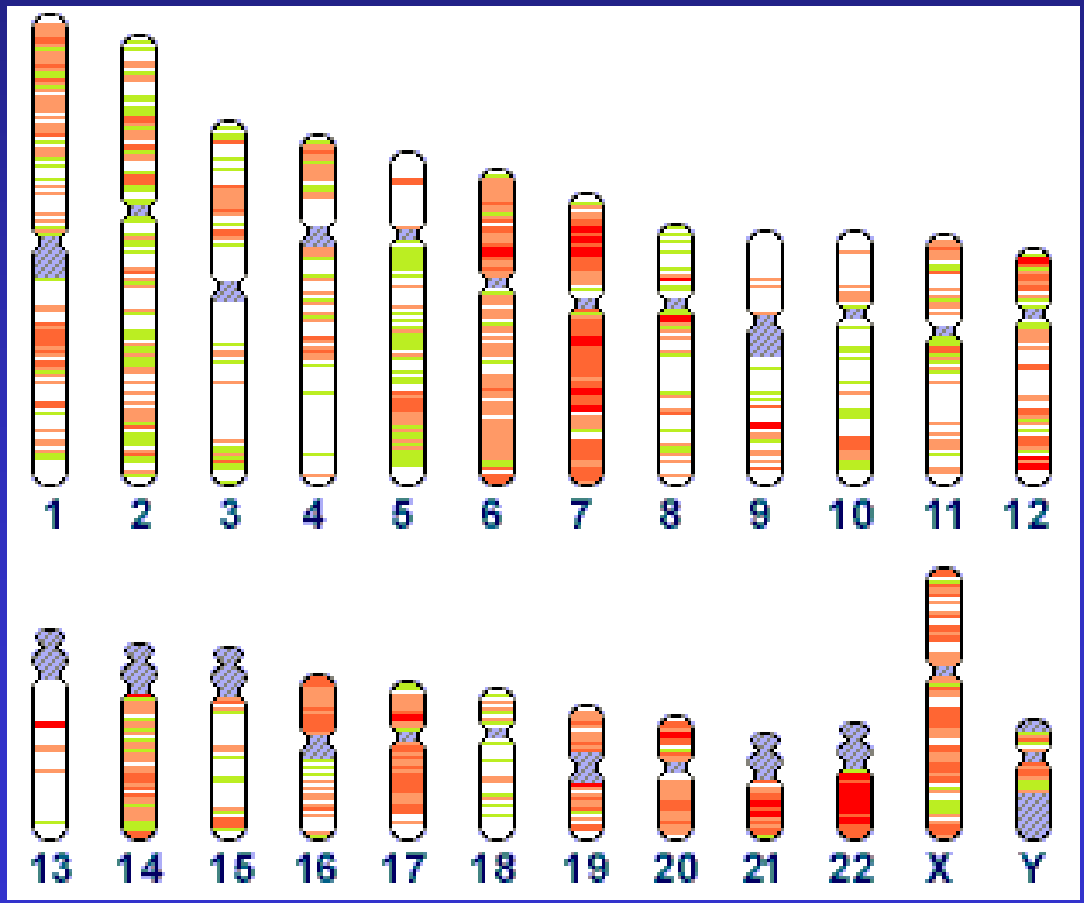
Genome analysis overview: *C.elegans*



L. major Friedlin chrom. 1 (Myler et al., PNAS 96, 2902-, 1999)

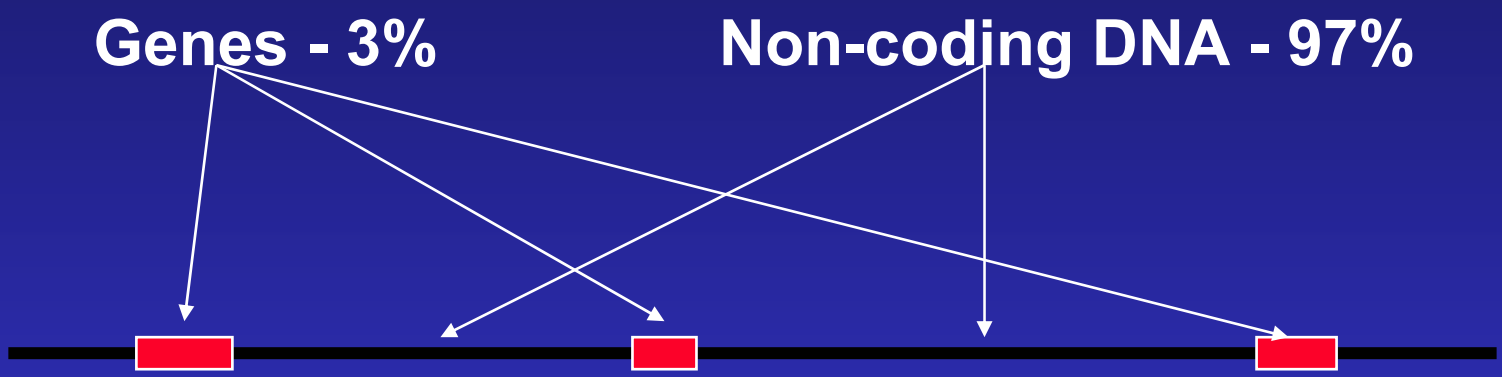


Human Genome Sequencing



The complete human genome sequence was announced in June 2000. Annotation of the draft sequences published February 2001

Human Genome Sequencing



HGP

Celera

CGCCAGCACTTCCGGGAAGTGACTTCGTCTCCGAAGCCGATTGGTTGTTGCTTTGCTCCCGCTCGCGTCGGTGGCGTTTTTCCTGCAGCGCGTGCGTGCTGCGCTACTGAGCAGCGCCATGGAGGACTCTG
AAGCACTGGGCTTCGAACACATGGGCCTCGATCCCCGGCTCCTTCAGGTACACGCGAGGGCTGGGGAGCCGGCTTACGGGCTCTGCGGGGCGCGCCATCGCTCTTCACGCCGCTTAAACCGCACTCCTGGT
CTCCTAGGCTGTCACCGATCTGGGCTGGTCGCGACCTACGCTGATCCAGGAGAAGGCCATCCCACTGGCCCTAGAAGGGAAGGACCTCCTGGCTCGGGCCCGCACGGGCTCCGGGAAGACGGCCGCTTATG
CTATTCCGATGCTGCAGCTGTTGCTCCATAGGAAGGCGGTGGGTAACGAGAGAGCTGAGGGGAGGAAGGAGGCAAGCTCCAAAAGCCTGGGAAGGGCGGTTCCCGTTTGTCTGAGGTTTTCTCTTGGCCCT
GTACCCGTGCAGGCCGGCCTGAGAACCTGGTGCTGTTGTGGCAAACACTCTGGGCTGGAGTTCAGGTTACCTGGATCCTTGTCCGGCCCTGCTACCACCAACCTTTGCGTAATCTTCGACAAAGCACTTTC
TTTTCTTTCTTACATAAAAAGGGAGCACATCTATCTTTTCTACTTACAGAATTATTGTGAGAATTTAGCTTCATAACTAGTATATTTAAAGTAGCTTCATAAACATCAGAGTACGTTATTCTTTTTGAGGG
TCAGTGCCTGGGGAAAGAACTCTCCACTCTGCATTCTGAGGCGGGCAGAGTGATAGATGATCAAAGTACTGCTAAGTAGTGTTGCAGCAGATGGGTCAGGTAGGCTGGAAGGGGTAGAGACACGTGGACAC
AGTGATGTGCACTGCTGGCTAAAGTCTTTAATTCATATTCTTACAGACAGGTCCGGTGGTAGAACAGGCAGTGAGAGGCCTTGTTCTTGTTCCTACCAAGGAGCTGGCACGGCAAGCACAGTCCATGATTC
AGCAGCTGGCTACCTACTGTGCTCGGGATGTCCGAGTGGCCAATGTCTCAGCTGCTGAAGACTCAGTCTCTCAGAGGTGGGTAAAAGCAGCAAAGCTGTACCTGAATGAAGCTACACAGTGTTGTGGGGTT
GGGTTTGTGTGTGGCAAAAAAGAGAGCAAATCCAGGGTGAGATCCCAGCTGCTACATTCTGCCTGATACTGATGTCTTGTCCACCTCCAGAGCTGTGCTGATGGAGAAGCCAGATGTGGTAGTAGGGACCC
CATCTCGCATATTAAGCCACTTGCAGCAAGACAGCCTGAAACTTCGTGACTCCCTGGAGCTTTTGGTGGTGGACGAAGCTGACCTTCTTTTTTCCTTTGGCTTTGAAGAAGAGCTCAAGAGTCTCCTCTGG
TAAGGCAGAGGTGGGTGTGATTCCTAGTGGAAACATCTGTGAGTAGGAGTTGGGACGAGAGCGGGGTGGCTGGAAGCCAGTTACTACAATTAGCGGCCCTTGGAGCTGGAATCTGATTGGATTCTTTCATT
TCAGTCACTTGCCCCGGATTTACCAGGCTTTTCTCATGTCAGCTACTTTTAACGAGGACGTACAAGCACTCAAGGAGCTGATATTACATAACCCGGTAAGAGGCACCATGGAAGTGTCTGGAGCTGCAGAC
ATGGGGGCACTCAAAGATCTTGATGCTCCTTCTTAGGGGATTCTTTGGTGTTTTGGGTGGGACAGTTGTCACTTAGTGTCTCATCCCTGGTCCTGAGGCACTAAAAGCCAGTGGTCTAAAATCACTATATA
TTTCCAAGTGTCCACAAGGGATGTCTCCCATTTCAGGCCATGCTTTGCCTAAAATCCTGAGCAAGGACCTCCCCTAAGGGGCAGCTTTGAGCAGCAGAGCCAAAATTCTAAGGCCAAGGTTCTCATCTTAA
GTAAACTTTACCTTTCAGAAGGCCTGTTGCTGTAGGCCTTCCCTTCTCAATGTAGTCCTTTATTGATGTGTTTCTCTTTGTTCTGTGCTTGGAAGTATTTTATATATGGTTTATATGGTATACTCTATATA
CCACAACAATAAGGGCATTTTGGGGTTTTAGGTTACAAAACTGGAGGAGAGTTAGGGTGCCAGGAATCCTTAAATGCATCTCTGCCCTGCACTAAAATGTTGATGCTTTGGTTGGTGAGTAAGTGGCCATA
CATCTCTGTGTTCTTTTCCTTTCTGACCACAGGCCTGTTTTCTCCCCCAGGTTACCCTTAAGTTACAGGAGTCCCAGCTGCCTGGGCCAGACCAGTTACAGCAGTTTCAGGTGGTCTGTGAGACTGAGGAA
GACAAATTCCTCCTGCTGTATGCCCTGCTCAAGCTGTCATTGATTCGGGGCAAGTCTCTGCTCTTTGTCAACACTCTAGAACGGAGTTACCGGCTACGCCTGTTCTTGGAACAGTTCAGCATCCCCACCTG
TGTGCTCAATGGAGAGCTTCCACTGCGCTCCAGGTCTGCCACAGCCAACATCTTGGTTGAAATAAGTTGAAGATAGAGATGGAAAGGGGACCCAGTTAATGTTCTGTTTCTTAAGCACTTAGTAGGGGCCA
GGTTCTAGATGTGACTGATACTGACTTCTCCCAACTCCAAAATACCTATCATGGCCGGGCACCATGGCTTATGCCTGCTGTAATCTCAGCACTTTGGGAGGCCGAGGTGGGCGGATCGCCTGAGGTCGGGA
GTTCAAGACCAGCCTGGCCAGCATGGTGAAACCCCGTCTCTACTAAAAATACAAAAATTAGCTGGACATGGTGGCAGGCACCTGTAATCCCAGCTACTCAGGAAGCTGAGATAGGAGAATTGCTTGAGCCC
GGGAGGTGGAGGTTGCAGTGAGCCAAGATCGTGCCATTGCACTCCAGCCTGGGCAACAGGAGTGAAACTCTGTCTCAAAAAAACAAAACCCTATAATTATTTCCAGCTGAGGAAACTGAGGCACAATGATT
AAGTAGGGAAAGAGATTAAGAAGAGGAAAAAGGAAAGGGTGATGGTTACTGTGATACTAGGGATGGCAGAGGGGCCTTGAGCTTGCTCTGCTGAGCTGATTCTCTGTCCGCTCTTGGCTGCAGGTGCCACA
TCATCTCACAGTTCAACCAAGGCTTCTACGACTGTGTCATAGCAACTGATGCTGAAGTCCTGGGGGCCCCAGTCAAGGGCAAGCGTCGGGGCCGAGGGCCAAAGGGGACAAGTGAGTCCATGCCTCTTTTT
CCATCCCTCCCCAGAAATGCCTGTGTTTTTAGCTTTTTGGAAGACTAAAACCAGAGTGCACAGAGCAGGGAGCCAAACCTTCCAGGCCTGGCTGGTAGTGTAGCCCAGAGAGCCCCACAGGTTCTTGCTCA
GCTGCCTGGATATAGAGAAGGGAGTGGATGGTGCACACTGCACATGCACCACGAAGGGCAAAACTGCCGGGGTTGTTGGCATGCAGAGCCCTGCAGGGGAGATGGCCCATCCTGCATTGGTGGTATGGCTG
TGACTTGCAGGGAGCATATTTCTGAAGGGAAAAGGAACCCCCCAACTCTCCAGTCTCTGTCCAGCTGAAGGCTTGACTAGCTCAGAGTTGGTTTTCAGATCACCATGTAGGGCAATGAGTTCTGCTGTTGT
CCCAGAACAGAGGTCAGGCCGAGATTTGGGTACATGTCAAAGCTCCAGGCTGCCCCAGGAAACCCTGACTCCTGGAACGGTTCCATTGTTGGAGAGTCCTCTGTATGTCAGGGTCTTATGATCTACAGGCA
TTTAGAGGAAGTTTTGCTGATTCAGCGTGTGAATACGTGCCCAGAGGAGAGGAAGGGTCCGGCTGACATTGAGTTATCTCTGCAGGGCCTCTGATCCGGAAGCAGGTGTGGCCCGGGGCATAGACTTCCAC
CATGTGTCTGCTGTGCTCAACTTTGATCTTCCCCCAACCCCTGAGGCCTACATCCATCGAGCTGGCAGGTAGTAGTGTGACGGCCCAGGCATCTGCATGGTAGGCACACTGAGGGACTTGGGGTGTGCTGG
ACAGAGCCTGCGGGTTGGAGATGCAAGCTGCACTGTCTTCCCTTGCAGGACAGCACGCGCTAACAACCCAGGCATAGTCTTAACCTTTGTGCTTCCCACGGAGCAGTTCCACTTAGGCAAGATTGAGGAGC
TTCTCAGTGGAGGTAAGAGCCTGGCTCTTGTGGTCCTGGGCCAGGGTCAGGCTTCTTCCACAATGCTTTAAAACTCCATGATAATGATGACAGAGGTCACAACATAGTGTGACAGGCCACTTCCACCATCC
ATCCTTGTTCTGCCCTGAGTGGCAGGCACTGTCCCCCTTGAGAGATAAACAAATTGAGGTAATTTGTCCAAAGTTGTGTTTACTGTCTGCCTCATGAGCGTTGAGTGACCTGACAGGCTGCTGTGACAGCT
CAGGACAGCACCTGACCCCAGGGTGCTGGGTGGTCCTGGACTGCTCTCTGTGGCCGTCGTCATGGGGGTACCTTGACTCCCAAGGAATACCATGGGGTACTCCTTGGGAGAGGAGAAGAGAGTGGGTGACG
GGTTCTTGGGCTTGGGGCCACACAGGCCACCCCCATCCACACACGGGGACAGATGGGTCATCACTGTAAGAGGCCCAGGTGCAGCTAACCTGCATGTTCGGCATCCCAGGAAGGCGGTGGGTCCCCTGCTG
CTTTCCCCCAAGGGGGAGGTGCAGGAGGCCTCCAATGAAGACCCTATCCTAAGGCCTCAGCCTGTGGGACCCTCGCTGCTTTCTTCTCCACAGAGAACAGGGGCCCCATTCTGCTCCCCTACCAGTTCCGG
ATGGAGGAGATCGAGGGCTTCCGCTATCGCTGCAGGGTGAGCTGCTGTGGTGGGGAGGGGAATGAGAGGGGAGGGGCTGTGGCCCAGGGATTGCACCGTCTTGCTGAGCATCCAGGTGTGAAGGGAGGATT
TGGGGCAGCCTCACTGTCTTGACCTTCAGTGTCCACCCCCAGGATGCCATGCGCTCAGTGACTAAGCAGGCCATTCGGGAGGCAAGATTGAAGGAGATCAAGGAAGAGCTTCTGCATTCTGAGAAGCTTAA
GGTGAGTGGATGGGAGGTGAGAAGGGGATAGATCTTAGACGGCTGCCCTTTTTGGAGACTGGCTGAGCTCCGAGTGGTGAGAAGCAGAGAACTGGGCAGTTTTCTGGCCTTTGGCACGGAAGGGGAGGAAA
TGGACCCAGAATCATGGAAGGAAGCCAGTCTGTTCTGCTTGGTGGTAAATTGGCACAACCTTATGGTGGACACTGTCCAGCAGAATTACGAGCTCATGTGTCCTTTCATCCGAAATTCCACTTCTGGAACT
TAATCCTGGTCACGCTTGTGAATGTGCACAGTCAAGCATGTGCCTGCATTCATCCATCCATGGCATTATCATGGAACCAAAAGATGGAAACAGCCTGGGGCCACCATAGGGGGCTTGCTAGGTAAACTCAG
GTGCATTCAGAGCCGAAGGTTACATGGGAAGGAATGAGGTTGGTTGCGTGTCCATATGGAACAGTCTGTAAGATGATGCCCAGCAAAAAGGGGTACAGGGTACTGCCATGTGTGTCATGGAGAAGGGAAAA
TGGAAACATCCACTCCCGGGAGGTTCTGAGAAATGCACAGAAGCAGCTGCCTCATGCCTTTTGAAACACATGAGTGTGTTATCCTTTGAAAAGCTAGGTCTGTGAAGTCACAGAAGAAAGATGCTCACTCT
GTGGCTCTCCCTCTTCCCCCGGCAGACATACTTTGAAGACAACCCTAGGGACCTCCAGCTGCTGCGGCATGACCTACCTTTGCACCCCGCAGTGGTGAAGCCCCACCTGGGCCATGTTCCTGACTACCTGG
GTGAGTGTGGCCTGACAGGGCAGGAGGCAGCAGGCTGGGGAAGTGGCATTAATTTCTCCACTGCTGGGTCAGCCCCTGTGCTTGGTGCTGGGGATGCTCAGGCAGAATAGAACCTGGAGACCCTGGCAGCA

http://www.dbbm.fiocruz.br
http://www.lacbionet.org
http://www.dbbm.fiocruz.br/helpdesk