# GENOMICA

**Genómica**: Subdisciplina de la Genética que tiene que ver con la clonación y caracterización molecular de genomas completos.

- **Genómica estructural**: Es la caracterización de la naturaleza física de genomas completos (secuenciación, ensamblaje, anotación)

- **Genómica funcional:** Es la caracterización de transcriptomas y proteomas. Conjunto de genes de un genoma que determinan proteínas y de los patrones globales de expresión génica.

# Avances claves en genomica

| | | año |
|---|---|---|
| → | • Double helix structure of DNA - Watson and Crick | 1953 |
| → | • Sintesis in vitro de RNA – Severo Ochoa | 1959 |
| → | • Descubrimiento del codigo genetico (M.Nirenberg •y H. J. Matthaei) | 1968 |
| → | • DNA sequencing - Sanger method | 1973 |
| → | • Detection of specific DNA sequences - Southern | 1975 |
| → | • DNA amplification by PCR - Mullis | 1988 |
| → | • Array-based genetic determinations – P.Brown | 1995 |
| → | • Sequencing of the human genome- J.Ventner et al. | 2001 |

# Análisis Genómico

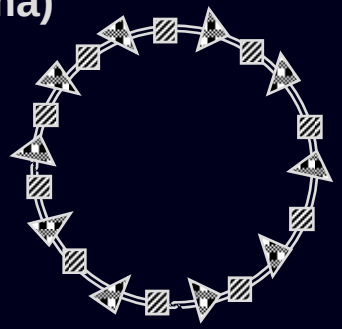
• Conocer los códigos contenidos en el genoma en un organismo y establecer su relación con otros genomas. (genómica comparada, filogenia, etc)

humano mosca plantas levadura

• Establecer la variabilidad de cada genoma (polimorfismo, SNPs)

caucásica asiatica negroide araucana

• Identificar regiones codificadoras y reguladoras (anotación del genoma)



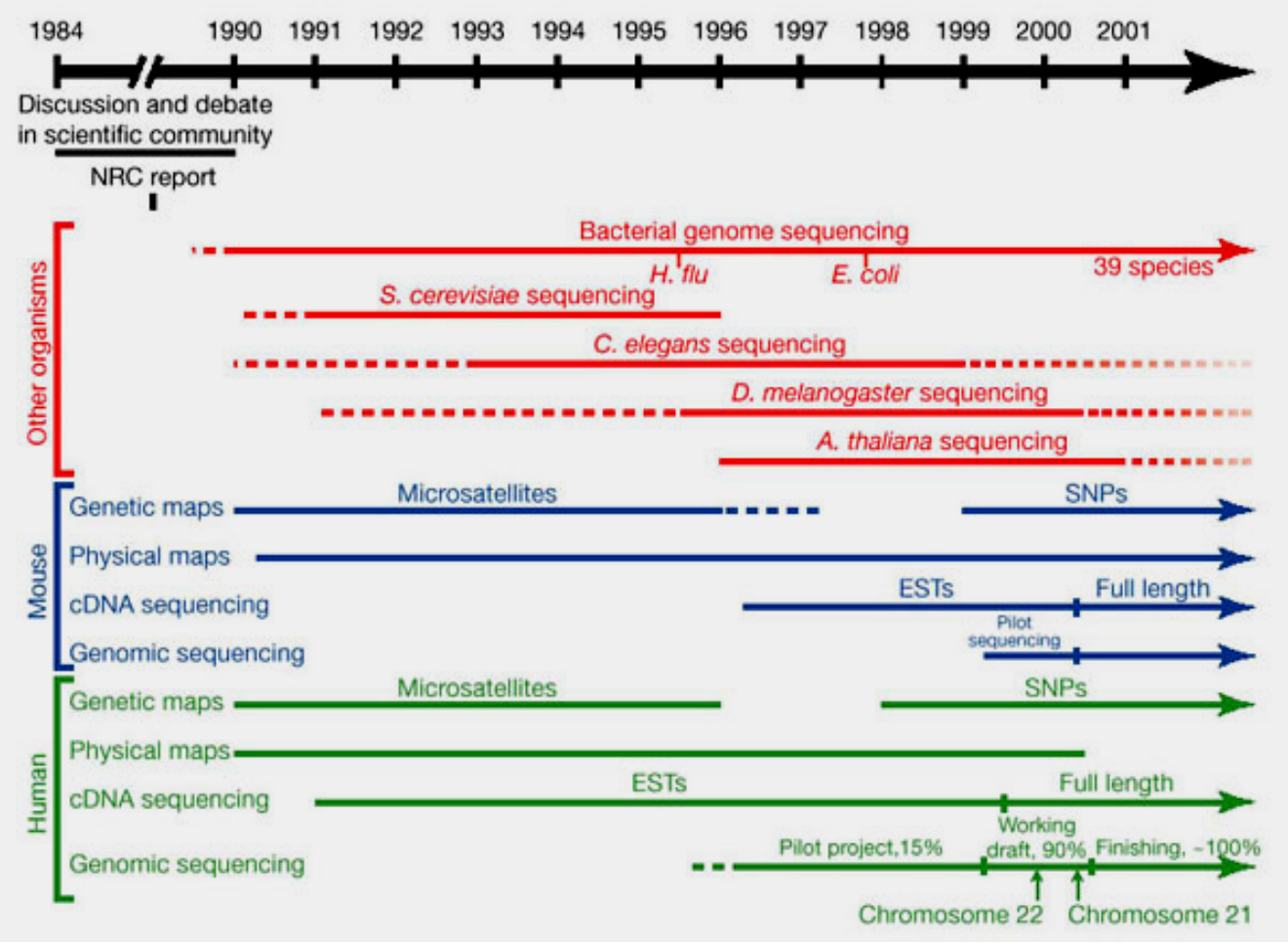
Codificadora, aquella que es transcrita en RNA

Reguladora, aquella que regula la transcripción

# QUE SE ESPERA DE LA GENOMICA FUNCIONAL

| Organism | # genes | % of genes with inferred function | Completion date of genome |
|---|---|---|---|
| *E. coli* | 4288 | 60 | 1997 |
| yeast | 6,600 | 40 | 1996 |
| *C. elegans* | 19,000 | 40 | 1998 |
| *Drosophila* | 12-14K | 25 | 1999 |
| *Arabadopsis* | 25,000 | 40 | 2000 |
| mouse | ~30,000? | 10-20 | 2003 |
| human | 30-40,000 | 10-20 | 2002 |

# HISTORIA DE LA GENOMICA



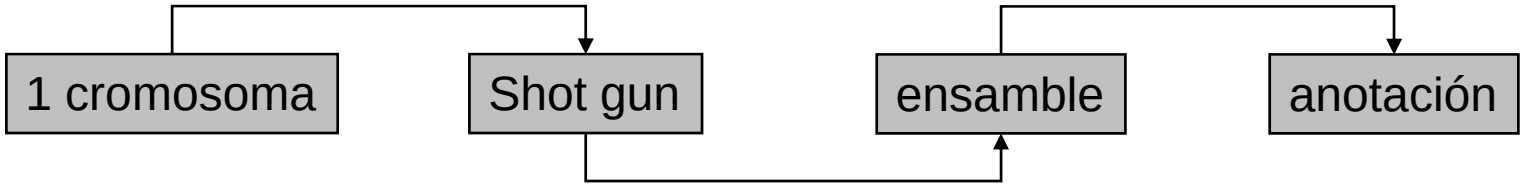

**27-10-2008**
**766-NCBI**

**Completo**
**July 2007 (mm9)**
**assembly $2{,}700 \times 10^6$**

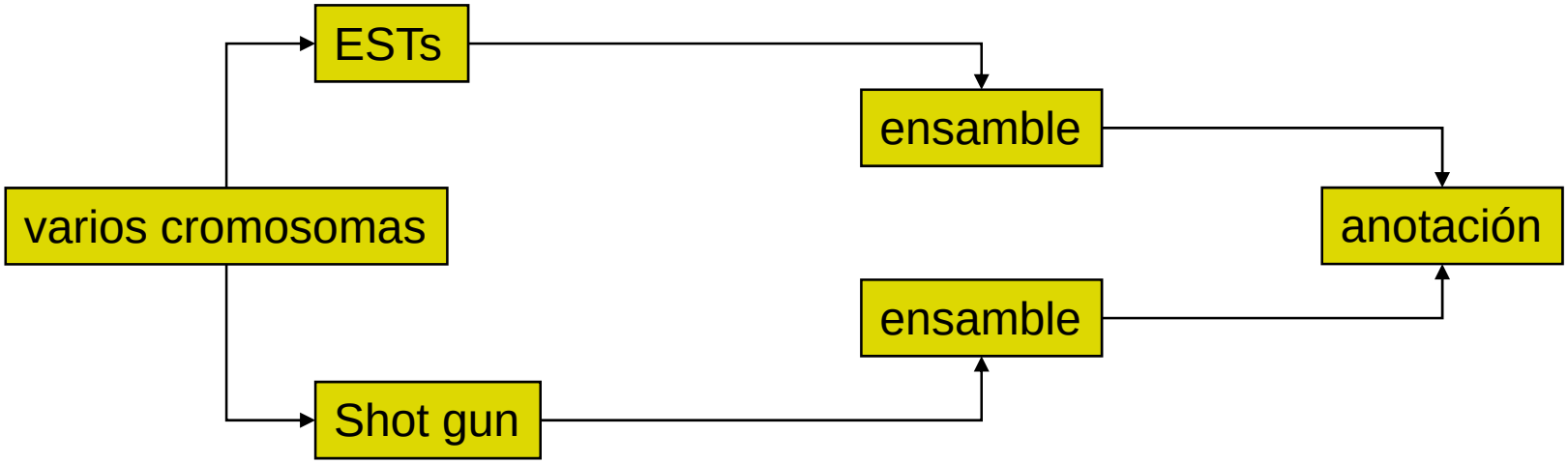
**Finalizado**
**NCBI Build 36.1, Mar. 2006 Assembly (hg18)**
**$3{,}080 \times 10 \times^6$**

**Otros 14 mamíferos; 10 vertebrados, 13 insectos; entre otros eucariontes han sido secuenciado total o parcialmente**

Bacterias




Eucariontes



Aislar el material genético de un organismo

1,5-4,5 x$10^6$ bp

Trozar el DNA (Shot-gun)

1-5 Kb y 40-60 Kb

Cloning

genoteca

1.000-100.000 clones

Secuenciación automática

10000 corridas (ambos extremos)

Lectura de secuencia nucleotídica

600-800 pb corridas

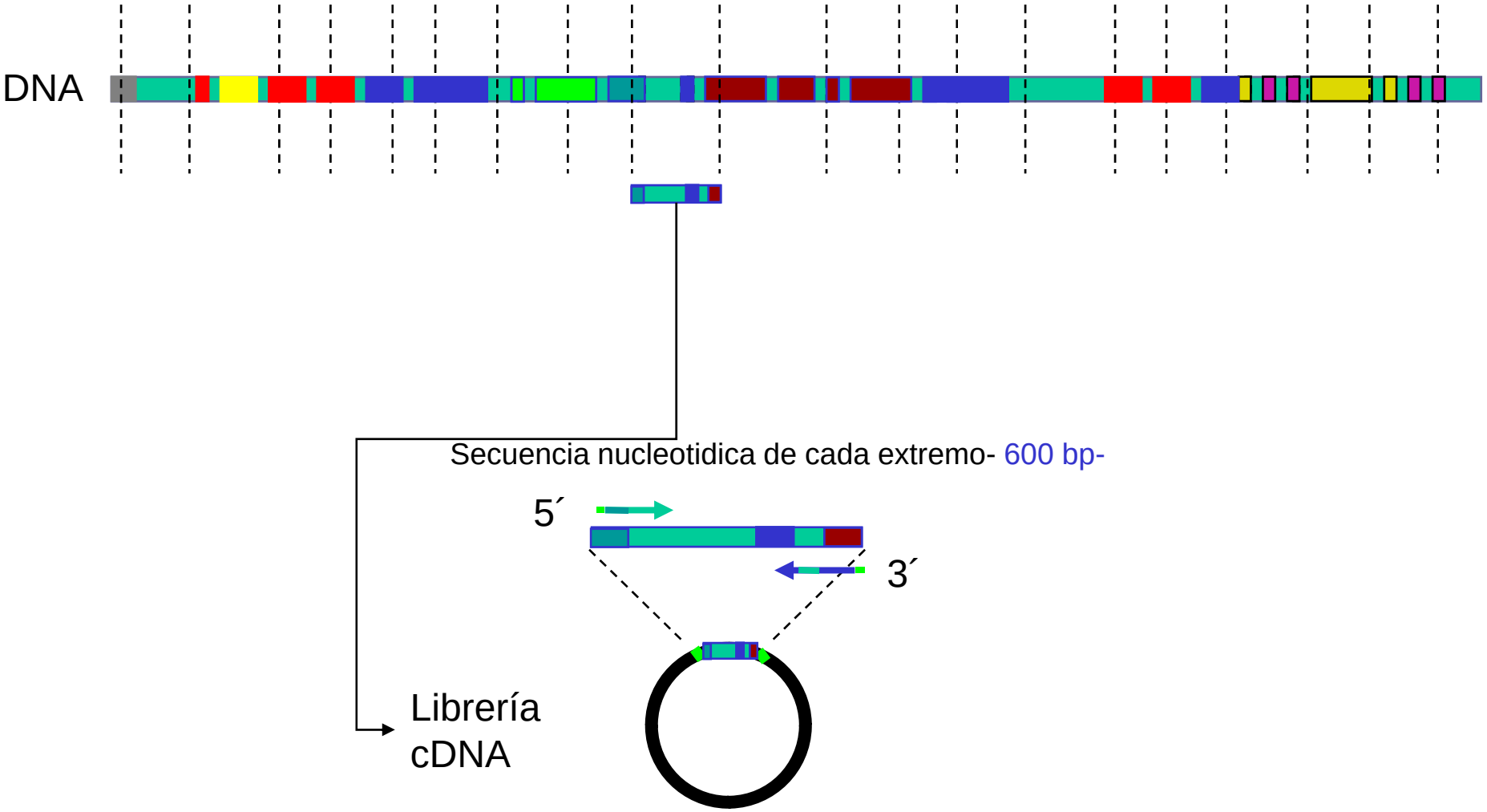
12 x$10^6$ bp

ensamblaje

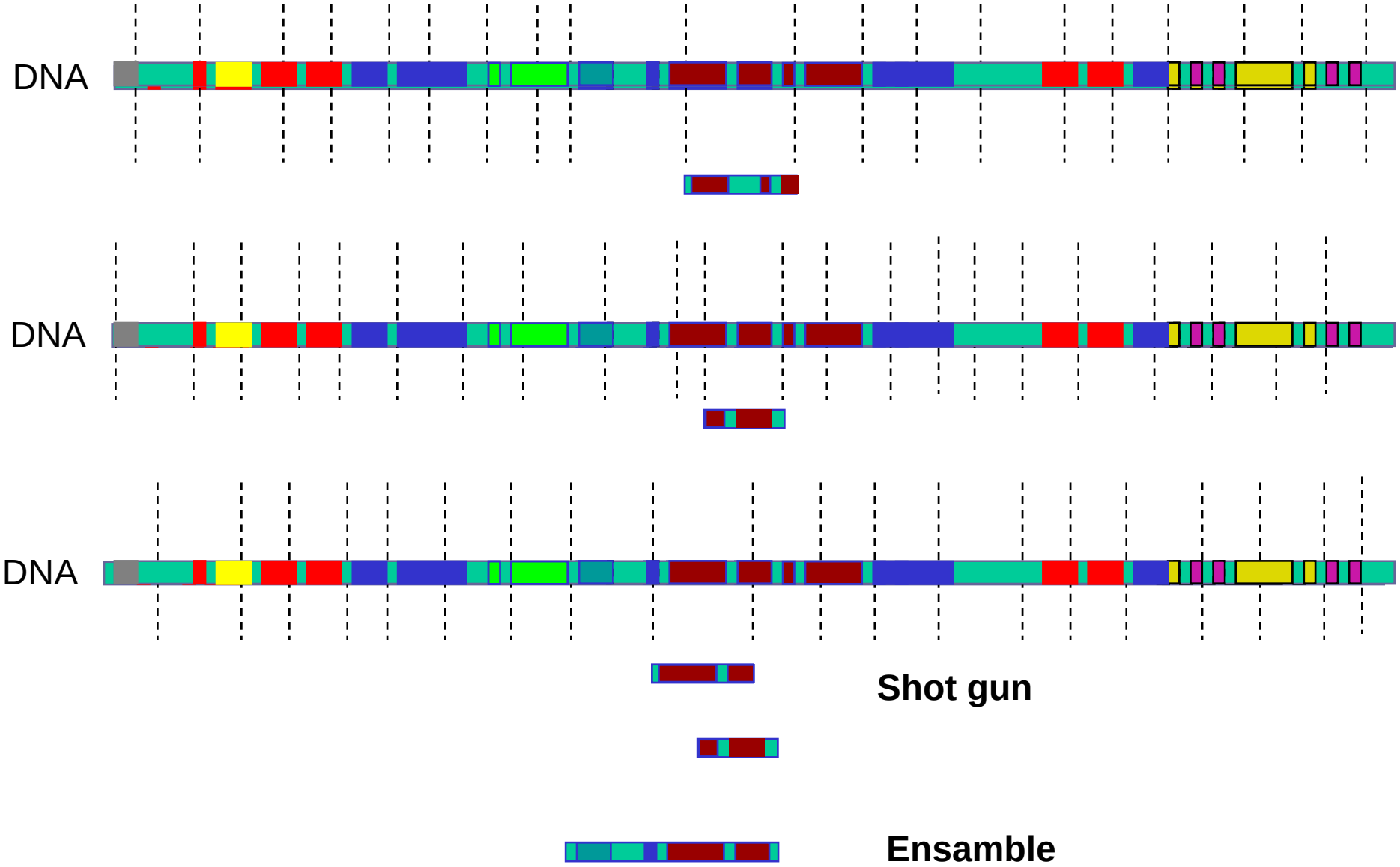
8-3 veces

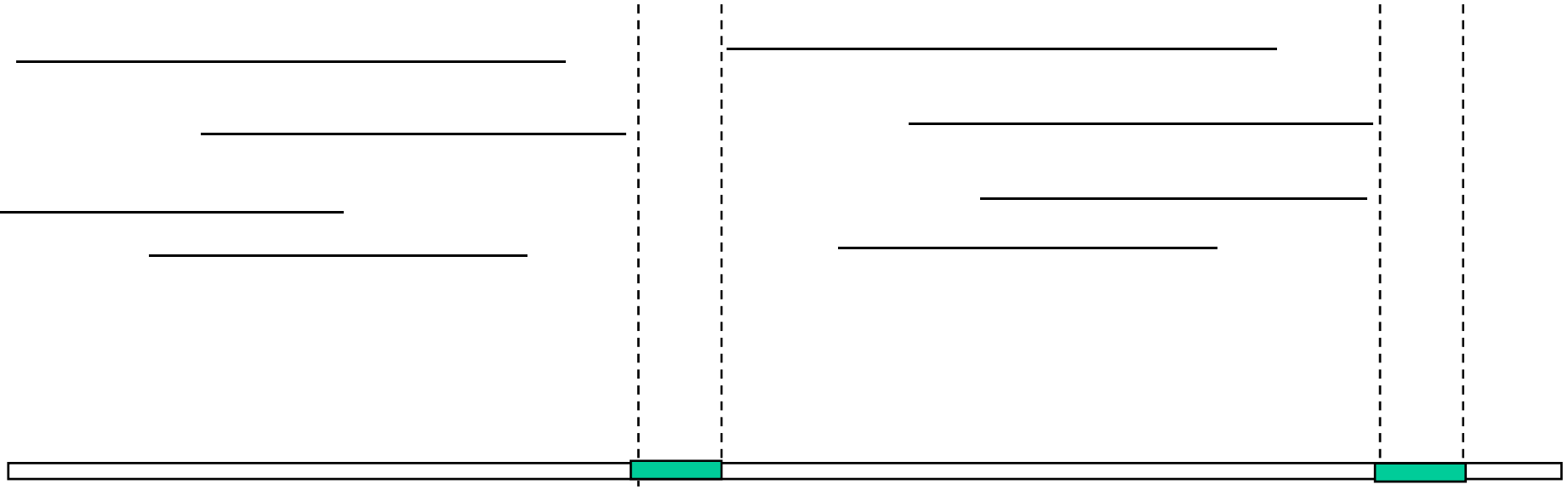
# Shot gun

**Cortes con enzimas de restricción**

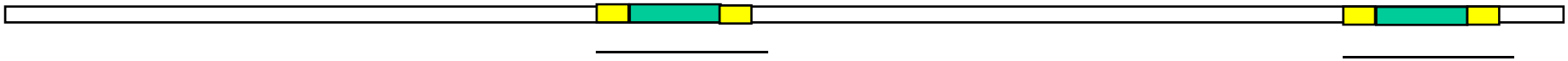


# Shot gun



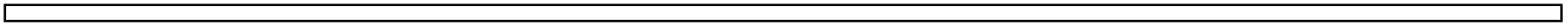


Recuperación de
fragmentos vía PCR



1

2

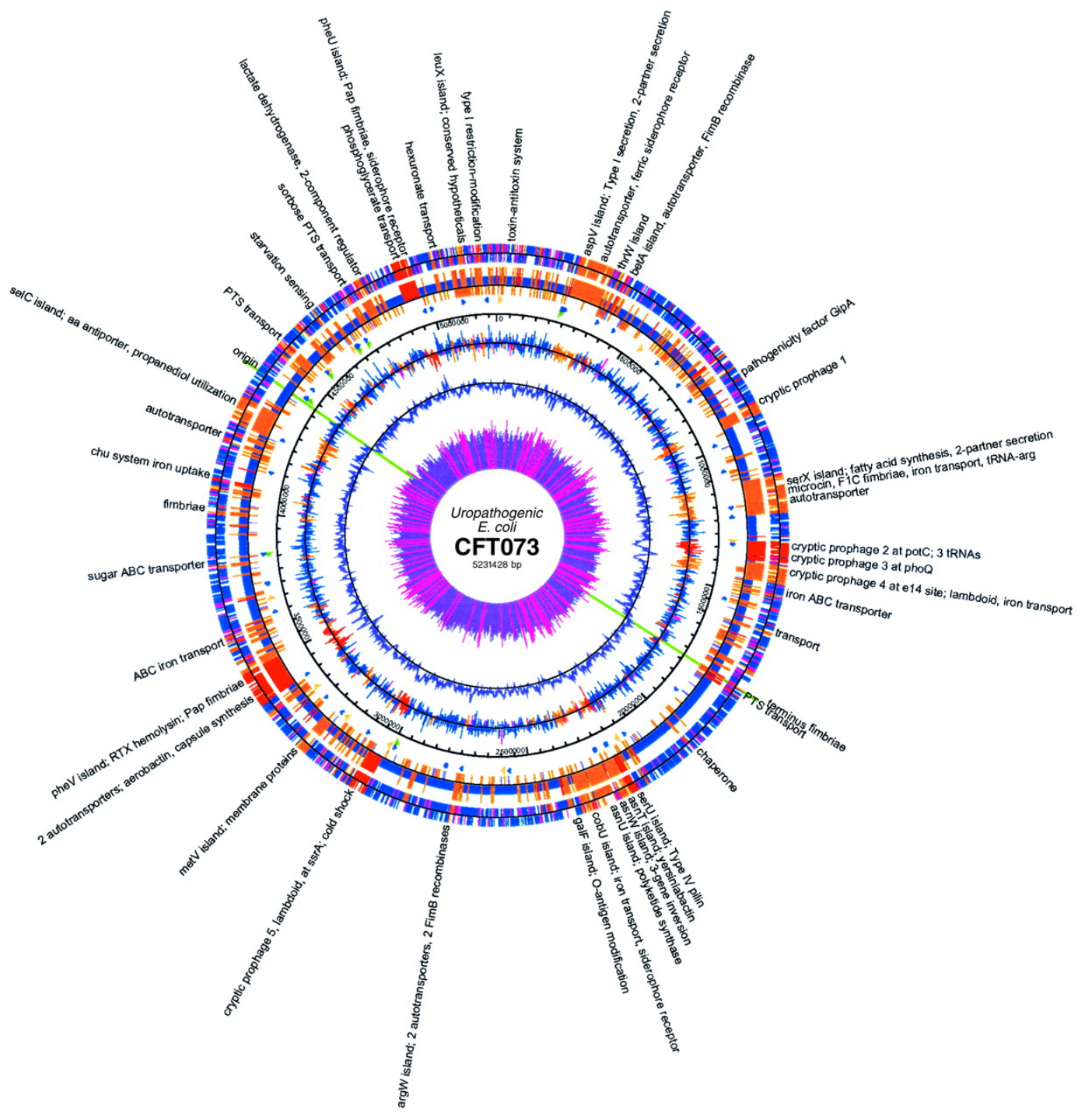


Repetición del protocolo de secuenciación con nuevos clones.

**Table 1. General Features of the *Thermoanaerobacter tengcongensis* Genome**

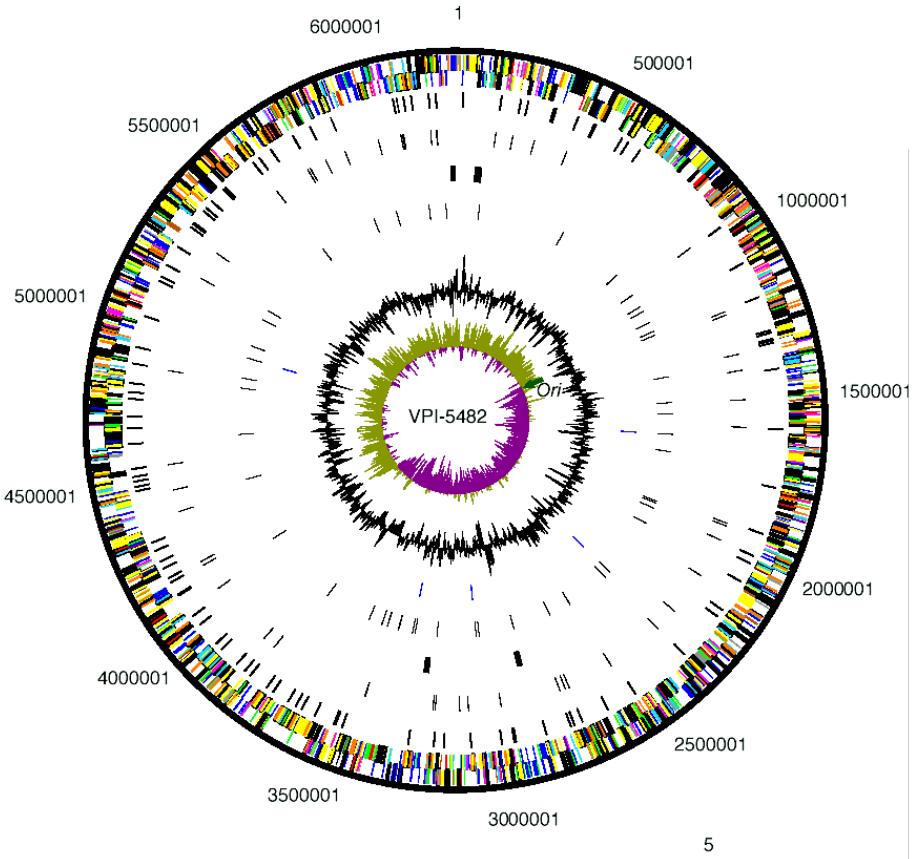
| | | |
|---|---|---|
| Genome size (bp) | | 2,689,445 |
| G+C content | | 37.6% |
| Protein coding | | 87.1% |
| Average CDS length (bp) | | 905 |
| Predicted CDS | | 2,588 |
| Homologous to | Known proteins | 1494 (57.8%) |
| | Protein domains/motifs | 270 (10.4%) |
| | Hypothetical proteins | 301 (11.6%) |
| No homology | | 523 (20.2%) |
| Stable RNAs | | 0.9% |
| rRNA operons | | 4 |
| rRNAs | | 55 |
| Major repetitive elements | | |
| Short noncoding repeats | | 0.5% |
| Long coding repeats | | 8.6% |

CDS, coding sequences.

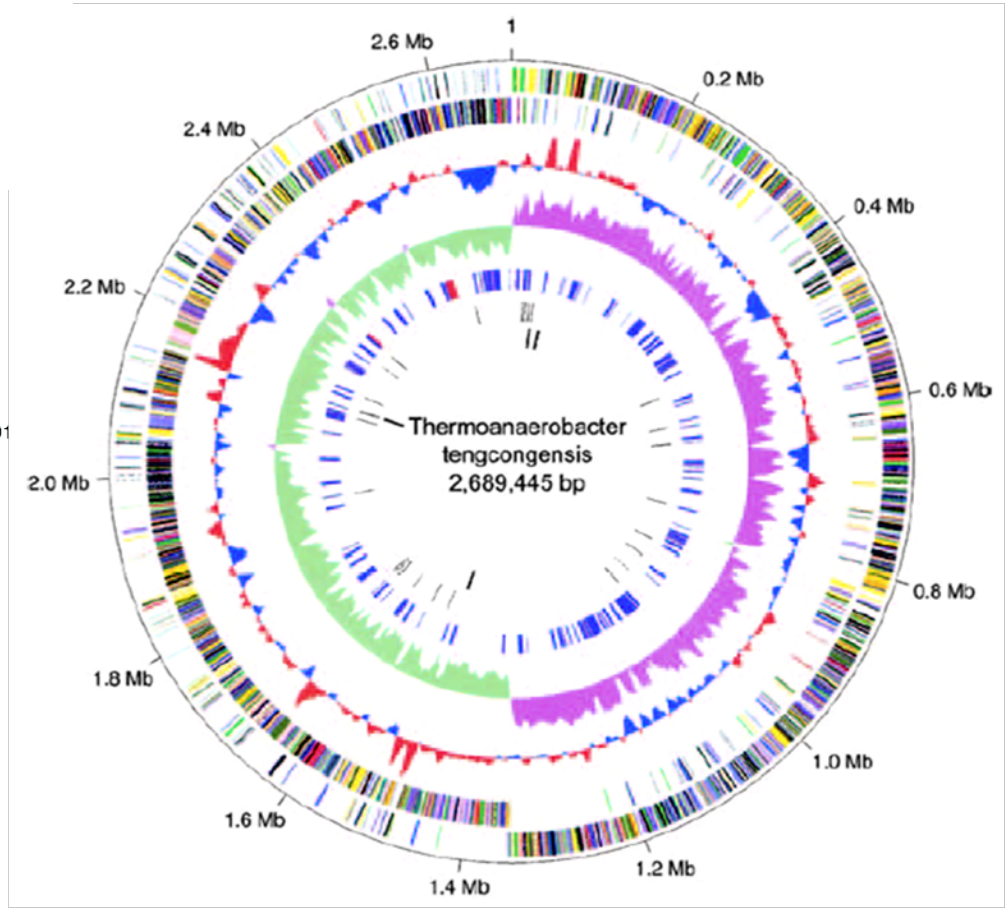



Map of the CFT073 genome and comparison with K-12 strain MG1655. The outer circle shows ORFs, colored according to the K-12 comparison in the second circle, where DNA regions are shown: blue, backbone, i.e., E. coli near match to MG1655; red, CFT073 islands (insertions); orange, islands (substitutions replacing K-12 segments); violet, K-12 islands. ORFs in the outer ring that span island-backbone junctions are pink. Third circle, RNAs: green, rRNA operons; blue, tRNAs; gold, miscellaneous RNAs. Fourth circle, scale in bp. Fifth circle, GC skew calculated for each ORF >100 aa, colored according to the same scheme of the ORF circle and plotted around the mean. Sixth circle, GC skew calculated over the whole sequence (window, 10 kb) plotted around the mean. Seventh circle, codon-adaptation index CAI (inverse, 1-CAI is plotted); pink rays indicate CAI values <0.2; purple rays, values >0.2. The pink rays can be seen to correspond with islands. A detailed linear map with ORF annotations is available at www.genome.wisc.edu. Maps were created by GENVISION from DNASTAR.

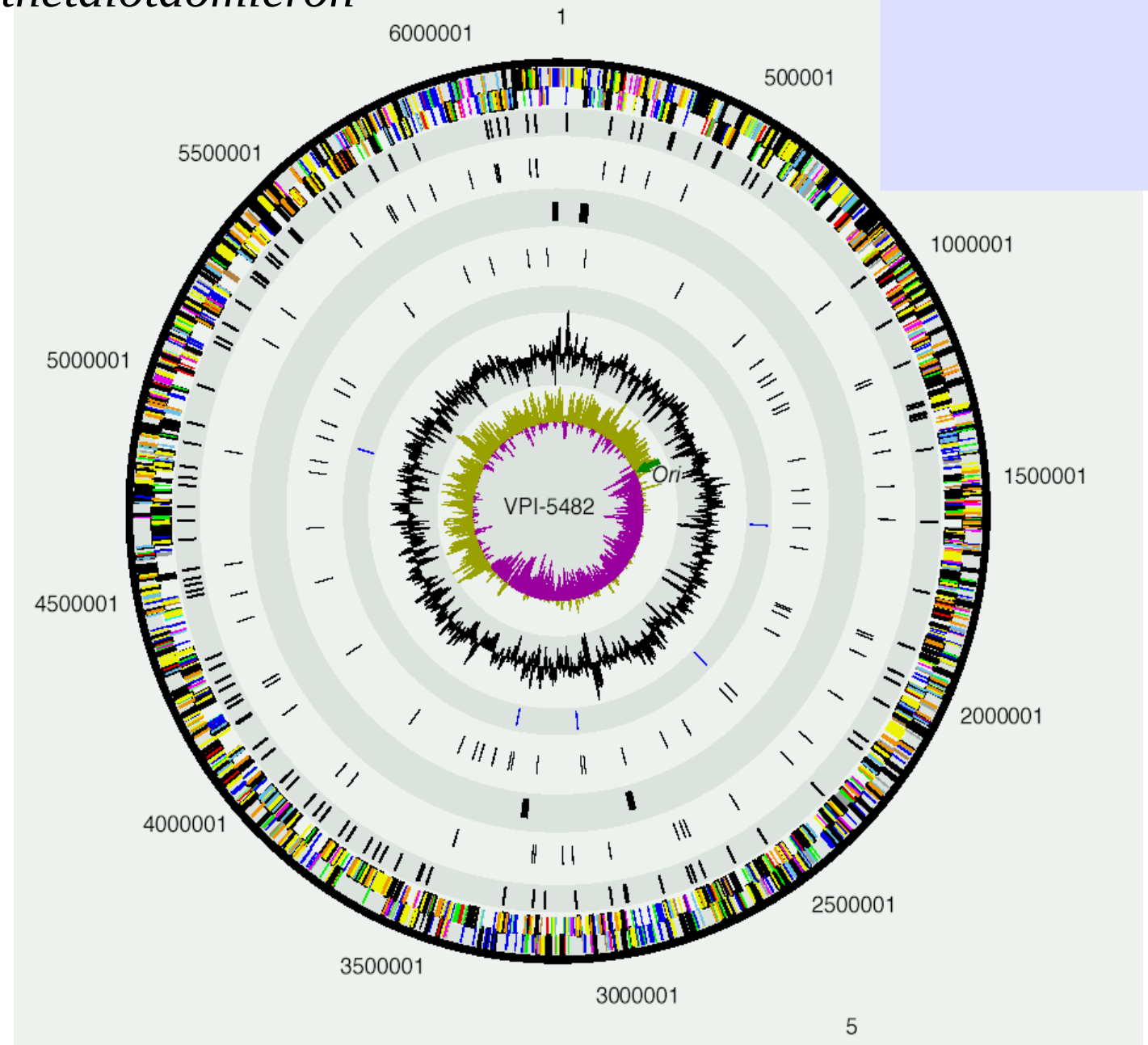
*B. thetaiotaomicron*




*T. tengcongensis*



*B. thetaiotaomicron*

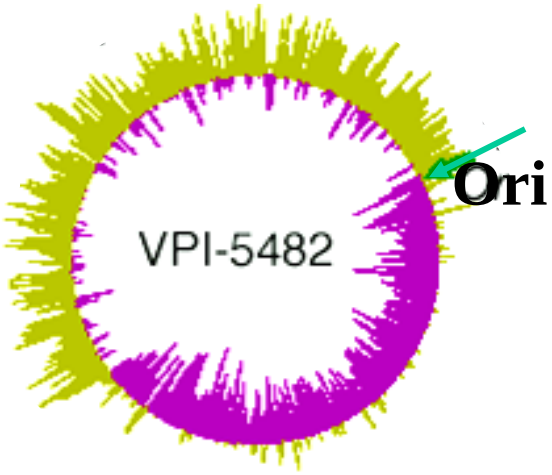


*B. thetaiotaomicron*

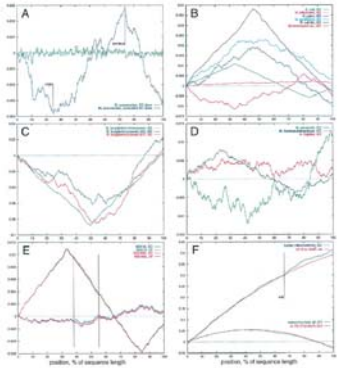
**Secuencia lineal de un cromosoma bacteriano**

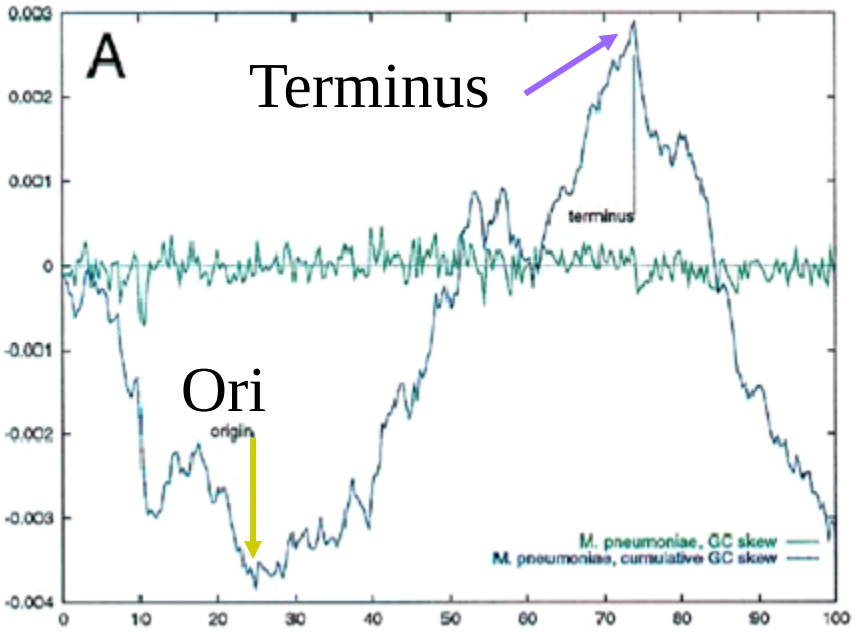
5` 3`

**Origin of DNA replication**:




**GC skew.**





1. GC nucleotide skew [(G-C)/(G+C)] analysis.

2 Identification by similarity, in this case we evaluated the co-localisation of several genes found around the origin of others bacterias. Typically bacteria contain the genes ***dnaA, dnaN, recF and gyrB****,* however the *Bacillus subtilis* chromosome include only the genes *dnaA and dnaN while* ***rpmH and recF*** are not include (Kuroda et al., 2001. The Lancet, 357: 1225), in *E. coli* dnaN and *gyrB* are present, dnaA is located 600bp away from dnaN – *gyrB*, and *recF* gene is missing. *(*Zawilaketal., 2001. Nucleid Acids Research **29**:2251***)***

3 In addition, we need evaluate if the replication origins region, contain short AT rich sequences and named **DnaA box** characteristic of bacterial (in *E. coli* and in *H. Pylori* there are 5 DnaA box).

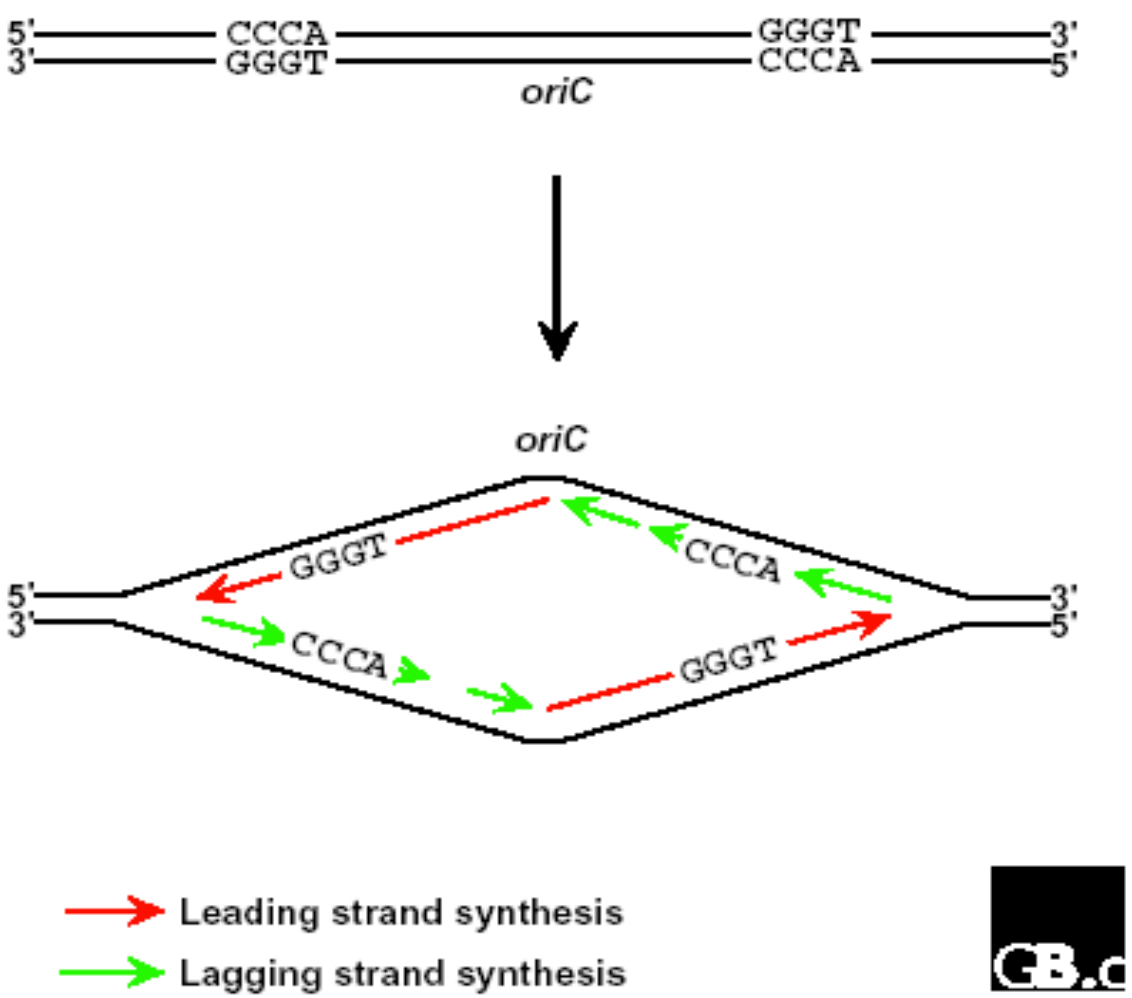
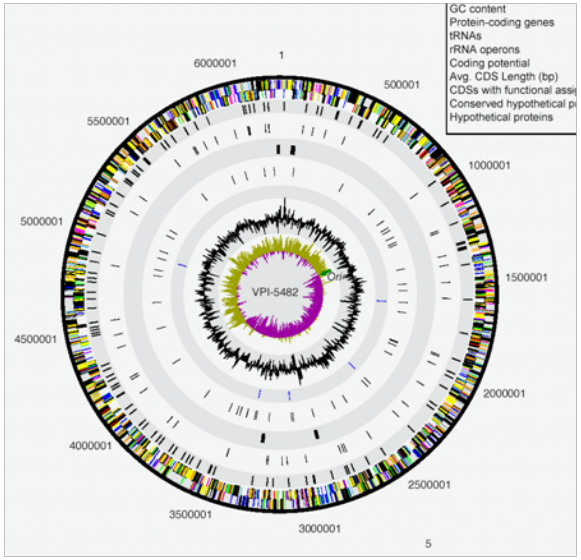
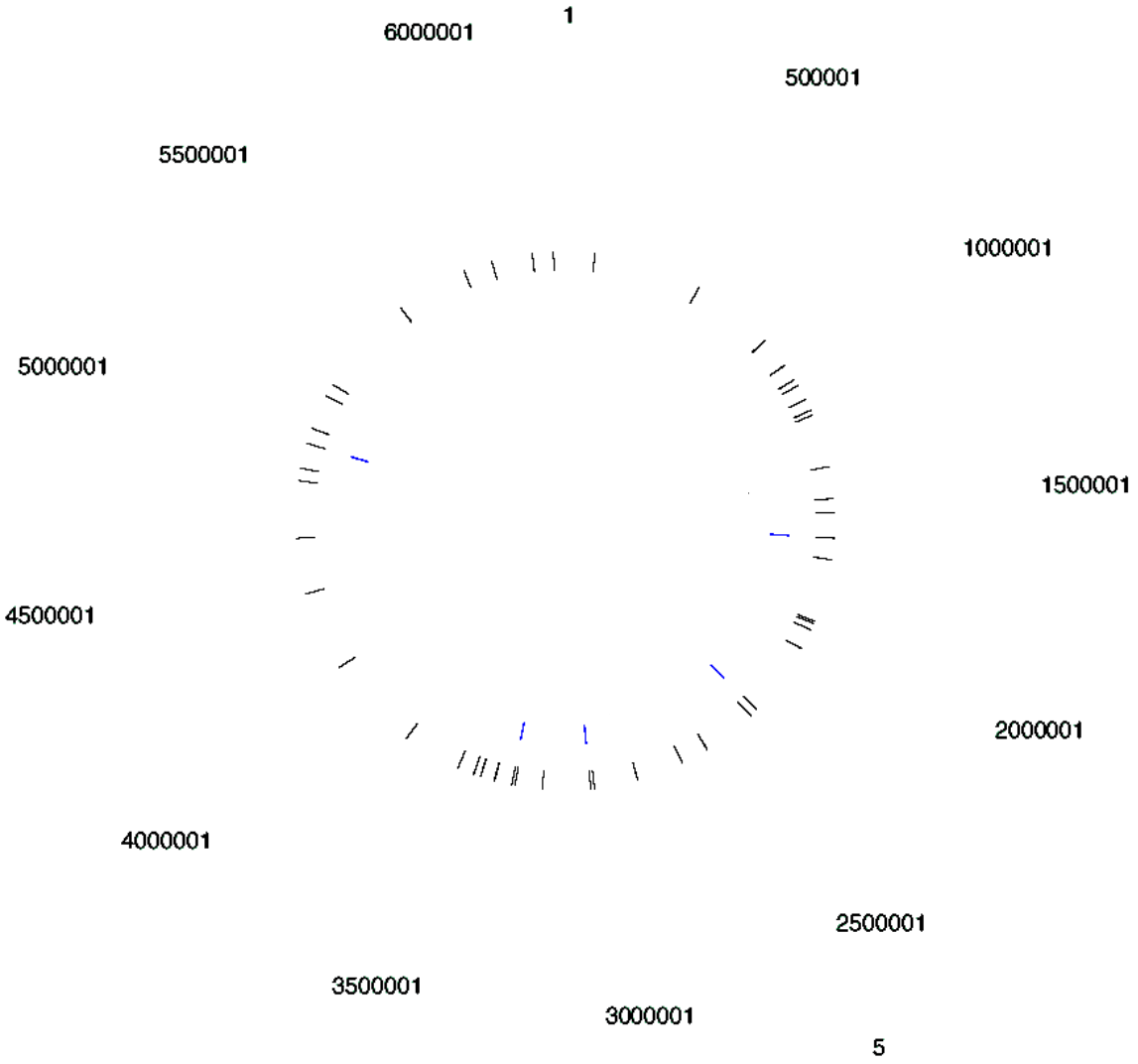


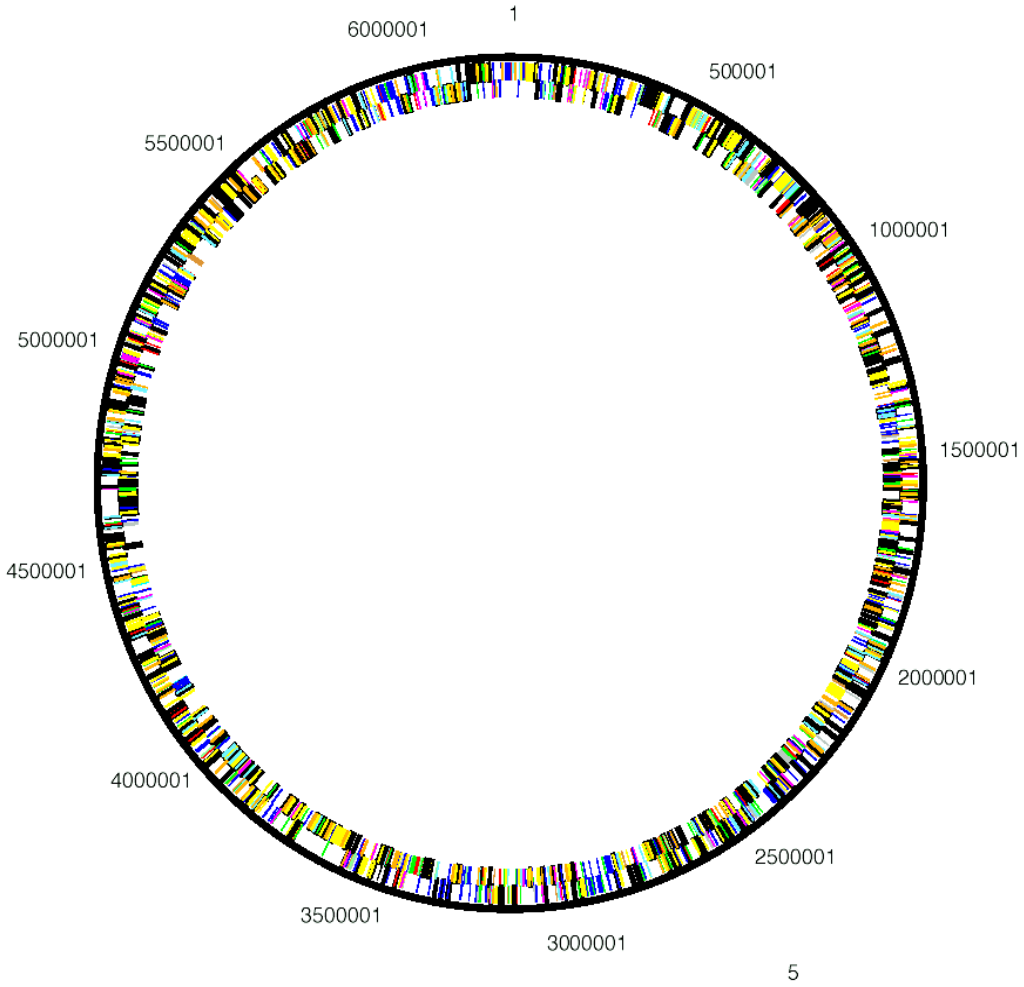

**Figure 2**
A bi-directional replication fork: DNA replicated by lagging-strand synthesis on one side of the origin will be replicated by leading-strand synthesis on the other side. In bacteria, there is a switch in the strand bias of guanine content at the origin. Myllykallio *et al*. [1] measured the strand bias resulting from GGGT as illustrated.

rRNA: BLASTN

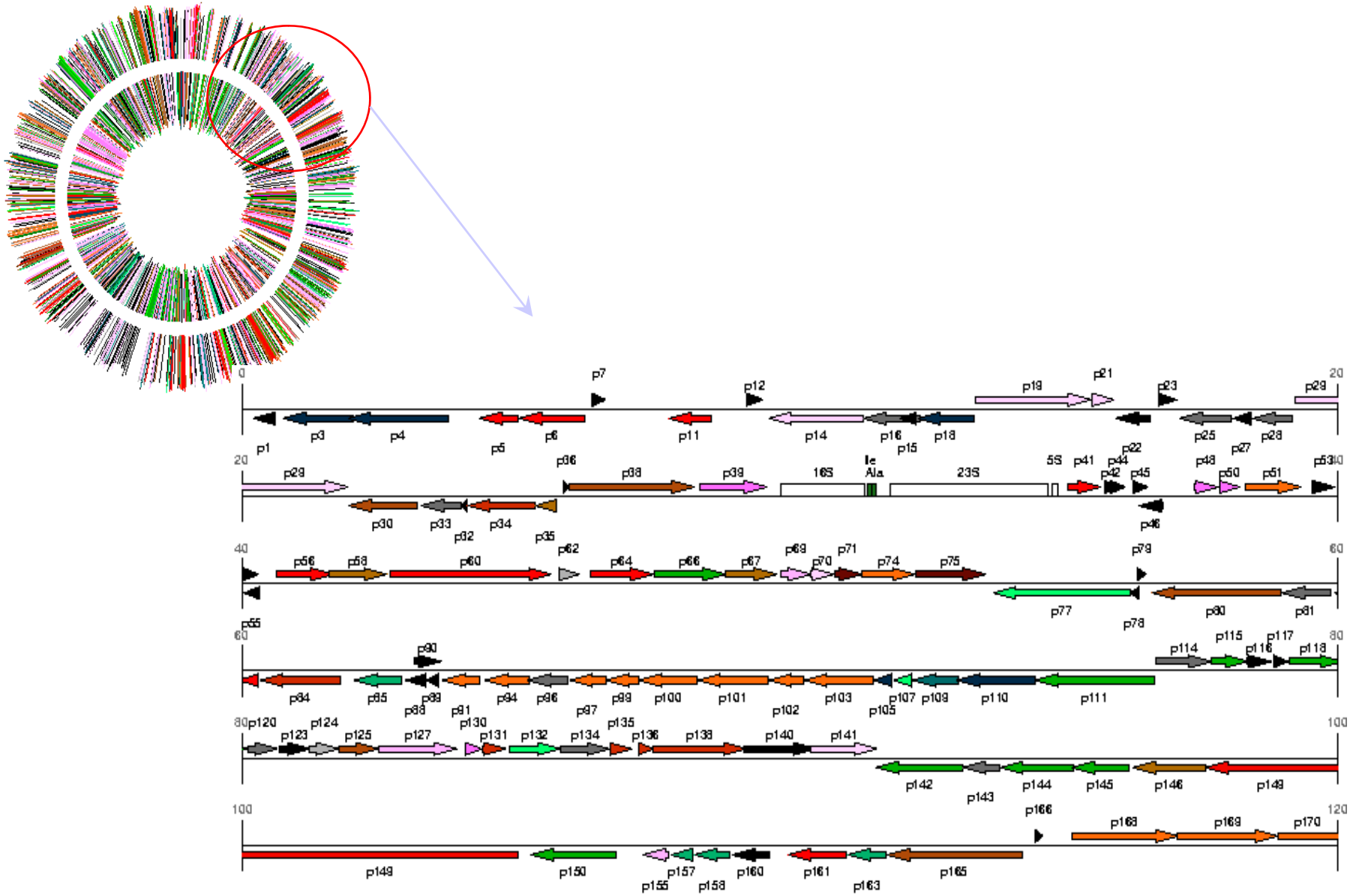
tRNA genes: tRNA-Scan



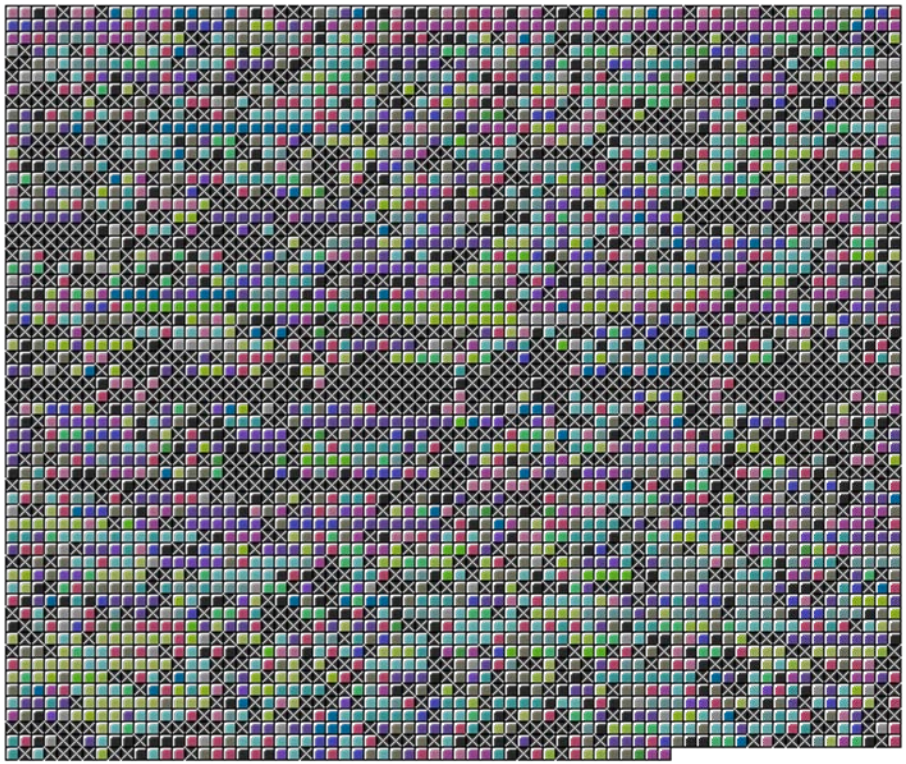



1. **Identificación de marcos de lectura por hebra. (utilizando GLIMMER, ORPHEUS and CRITICA .**
2. **Verificación por blast. % recuperado x homolg.**
3. **Revisión manual**
4. **Genes que codifican para proteinas con <60-100 codones que no presentan homologías con un cierto umbral de *E value* ($10^3$) deben ser eliminadas.**
5. **Clasificación de funciones (gen-ontology) por grandes grupos.**
6. **Mapa físico en megabases**

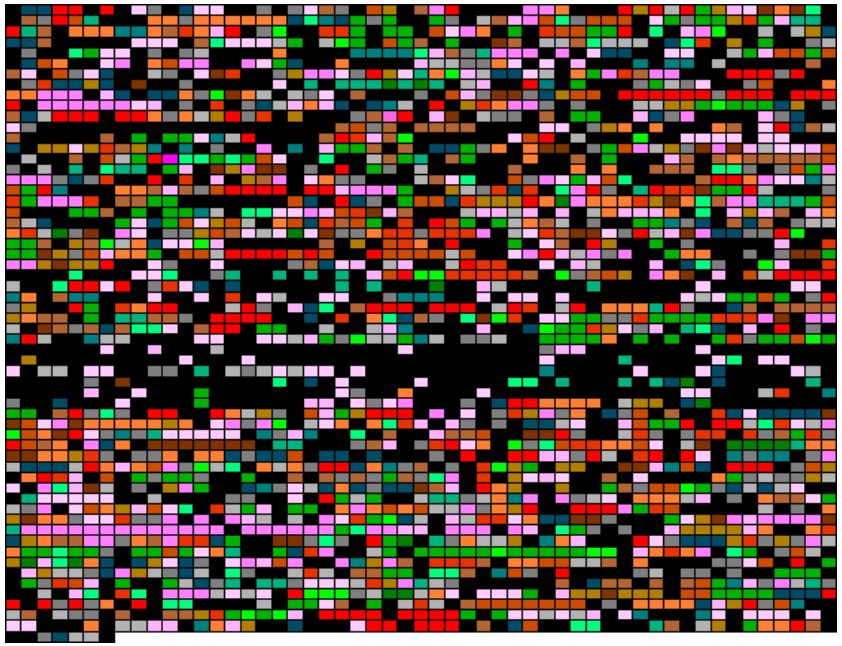

0
20
40
60
80
100
120
p1
p3
p4
p5
p6
p7
p11
p12
p14
p15
p16
p18
p19
p21
p23
p25
p27
p28
p29
p30
p32
p33
p34
p35
p36
p38
p39
16S
Ile
Ala
23S
5S
p41
p42
p44
p45
p46
p48
p50
p51
p53
p55
p56
p58
p60
p62
p64
p66
p67
p69
p70
p71
p74
p75
p77
p78
p79
p80
p81
p83
p84
p85
p88
p89
p90
p91
p94
p96
p97
p99
p100
p101
p102
p103
p105
p107
p109
p110
p111
p114
p115
p116
p117
p118
p120
p123
p124
p125
p127
p130
p131
p132
p134
p135
p136
p138
p140
p141
p142
p143
p144
p145
p146
p149
p150
p155
p157
p158
p160
p161
p163
p165
p166
p168
p169
p170

# Anotación funcional de genomas de bacterias

Anotación NCBI de *Bacillus subtilis*:



Anotación LBMG bacteria biominera

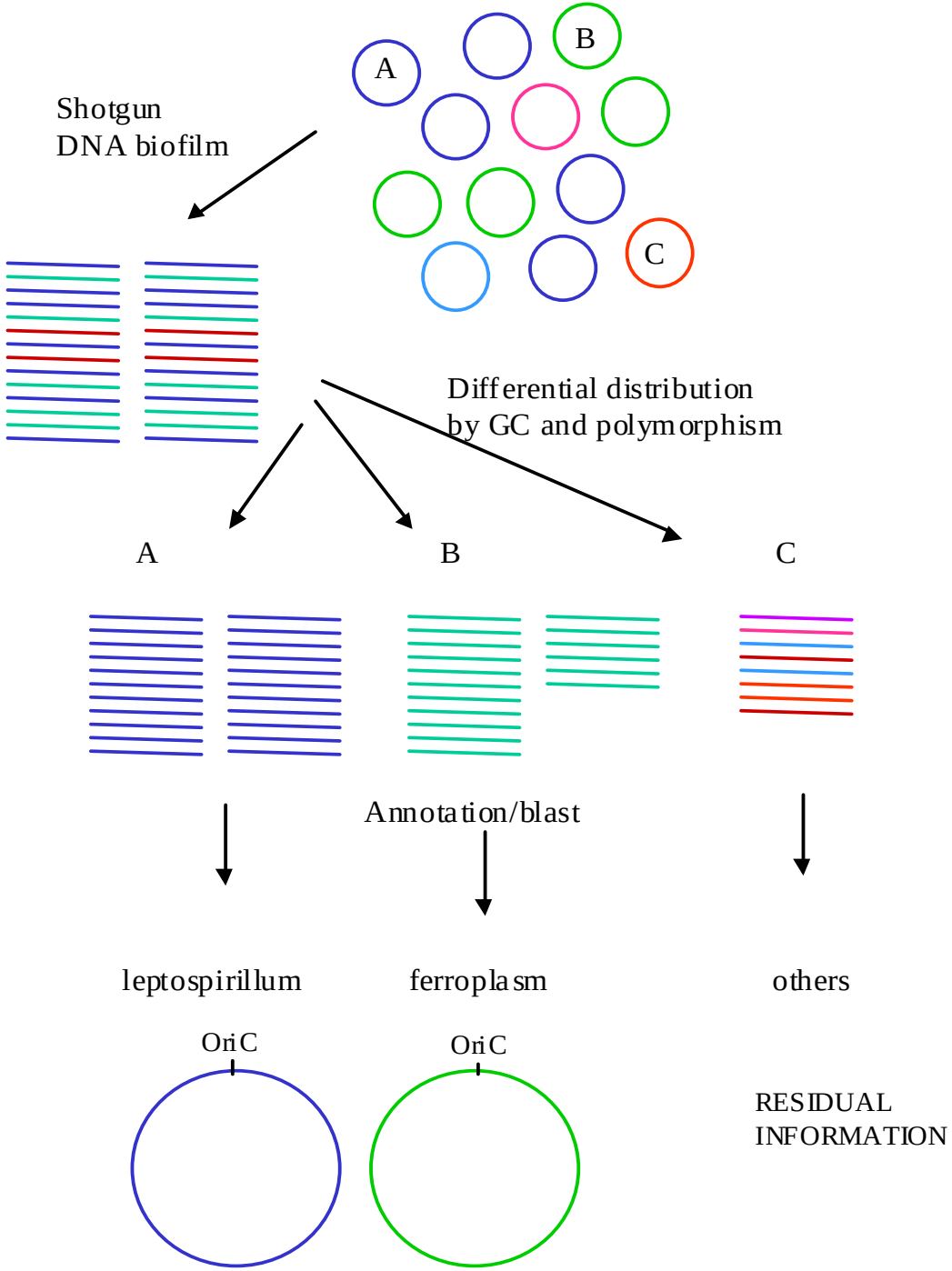


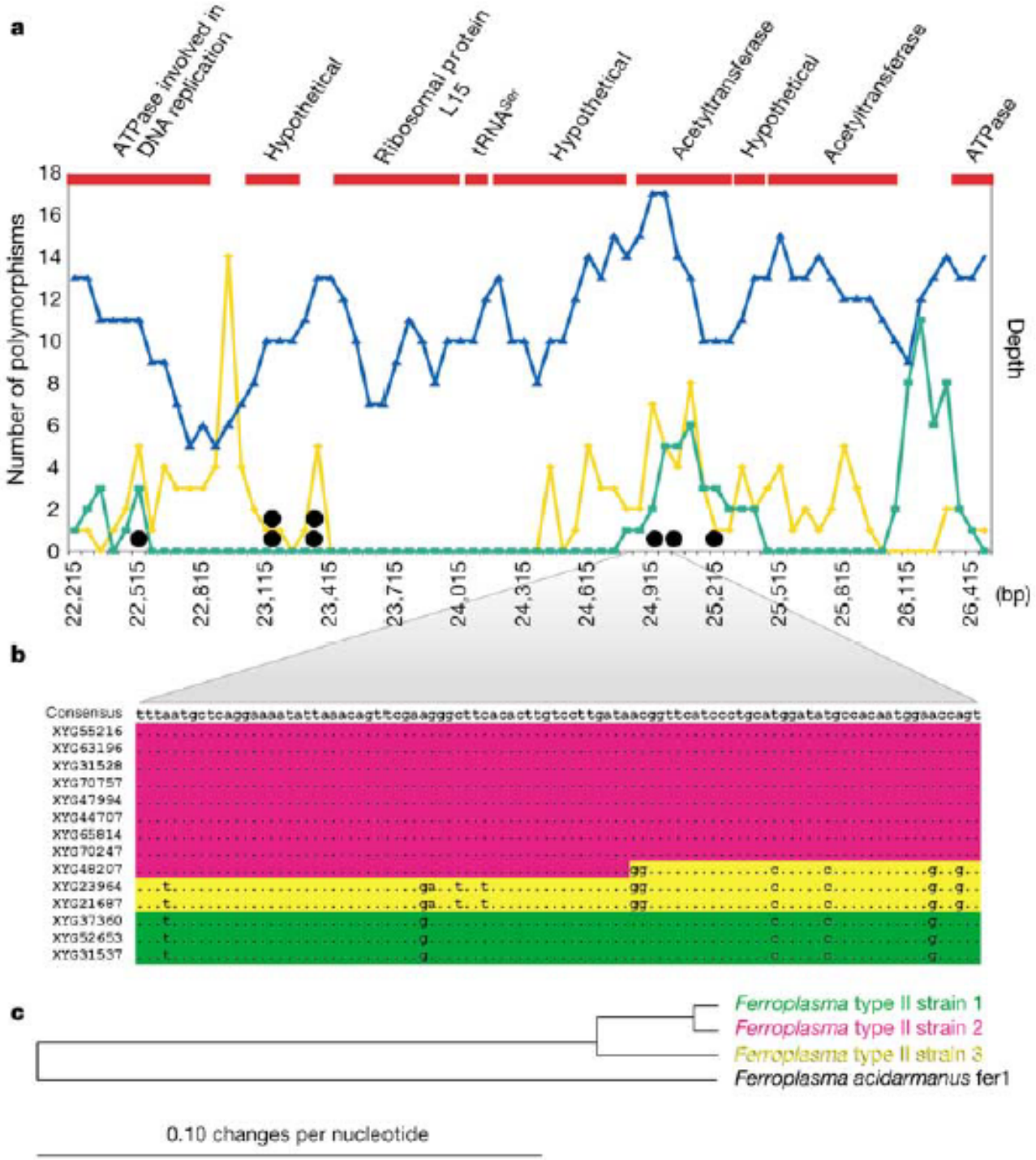
# Community structure and metabolism through reconstruction of microbial genomes from the environment

**Gene W. Tyson[1], Jarrod Chapman[3,4], Philip Hugenholtz[1], Eric E. Allen[1], Rachna J. Ram[1], Paul M. Richardson[4], Victor V. Solovyev[4], Edward M. Rubin[4], Daniel S. Rokhsar[3,4] & Jillian F. Banfield[1,2]**

[1]*Department of Environmental Science, Policy and Management,* [2]*Department of Earth and Planetary Sciences, and* [3]*Department of Physics, University of California, Berkeley, California 94720, USA*
[4]*Joint Genome Institute, Walnut Creek, California 94598, USA*

NATURE |VOL 428 | 4 MARCH 2004 | www.nature.com/nature


A
B
C
Shotgun
DNA biofilm
Differential distribution
by GC and polymorphism
A
B
C
Annotation/blast
leptospirillum
ferroplasm
others
OriC
OriC
RESIDUAL
INFORMATION


a
ATPase involved in DNA replication
Hypothetical
Ribosomal protein L15
tRNA^Ser
Hypothetical
Acetyltransferase
Hypothetical
Acetyltransferase
ATPase
Number of polymorphisms
Depth
(bp)
b
Consensus
c
Ferroplasma type II strain 1
Ferroplasma type II strain 2
Ferroplasma type II strain 3
Ferroplasma acidarmanus fer1
0.10 changes per nucleotide

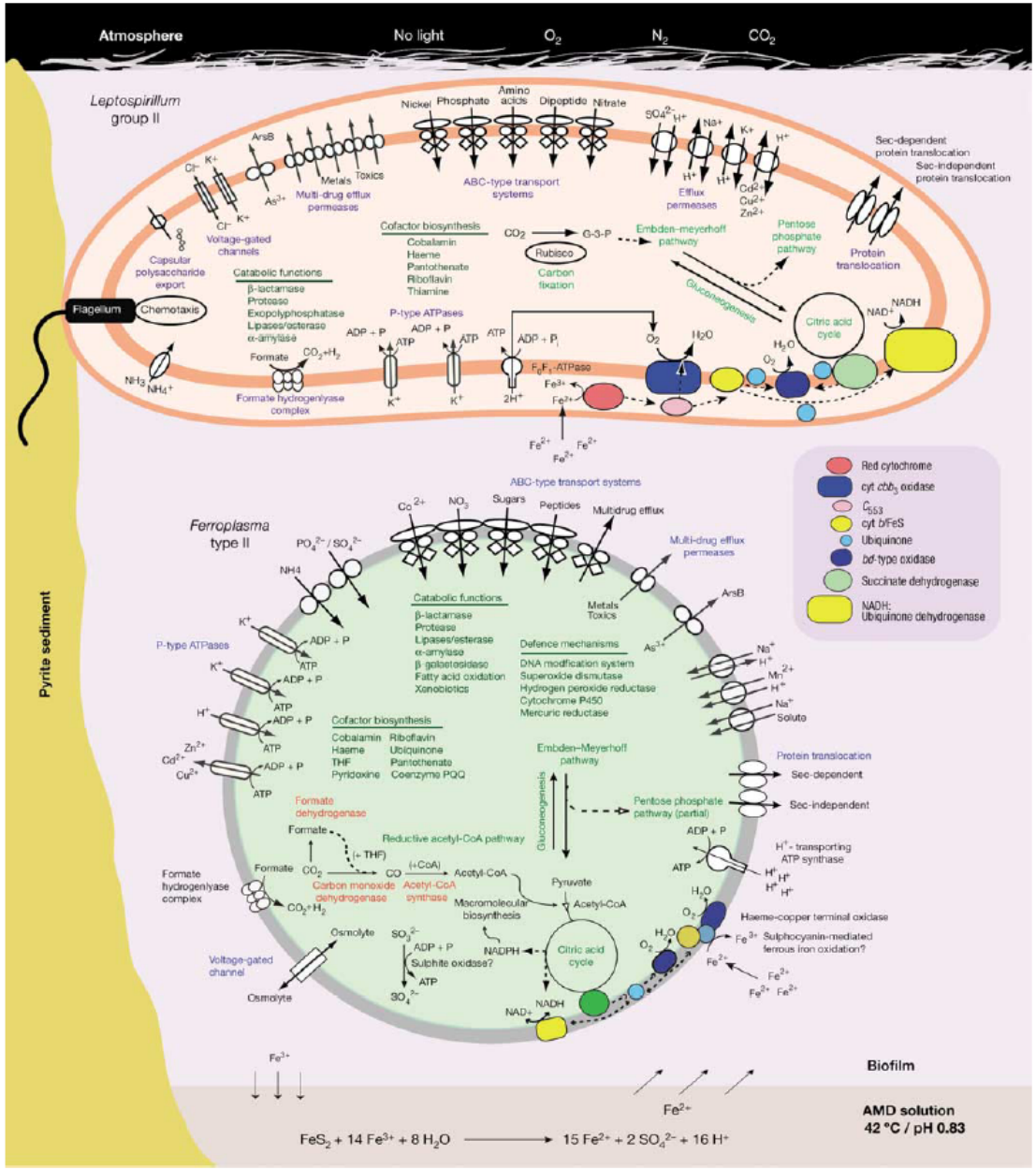
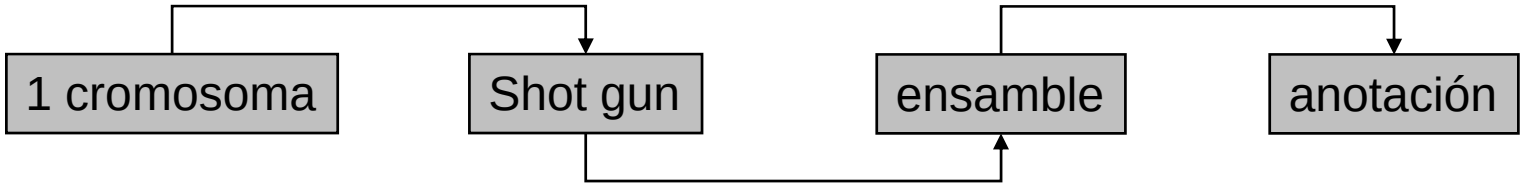
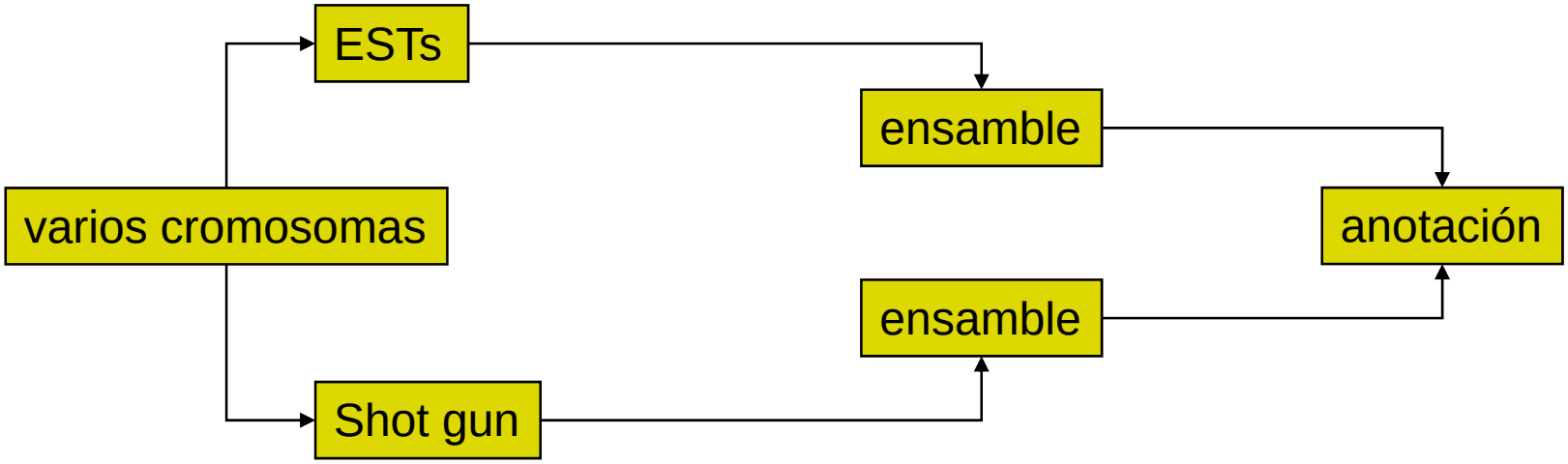



**Figure 4** Cell metabolic cartoons constructed from the annotation of 2,180 ORFs identified in the *Leptospirillum* group II genome (63% with putative assigned function) and 1,931 ORFs in the *Ferroplasma* type II genome (58% with assigned function). The cell cartoons are shown within a biofilm that is attached to the surface of an acid mine drainage stream (viewed in cross-section). Tight coupling between ferrous iron oxidation, pyrite dissolution and acid generation is indicated. Rubisco, ribulose 1,5-bisphosphate carboxylase–oxygenase. THF, tetrahydrofolate.

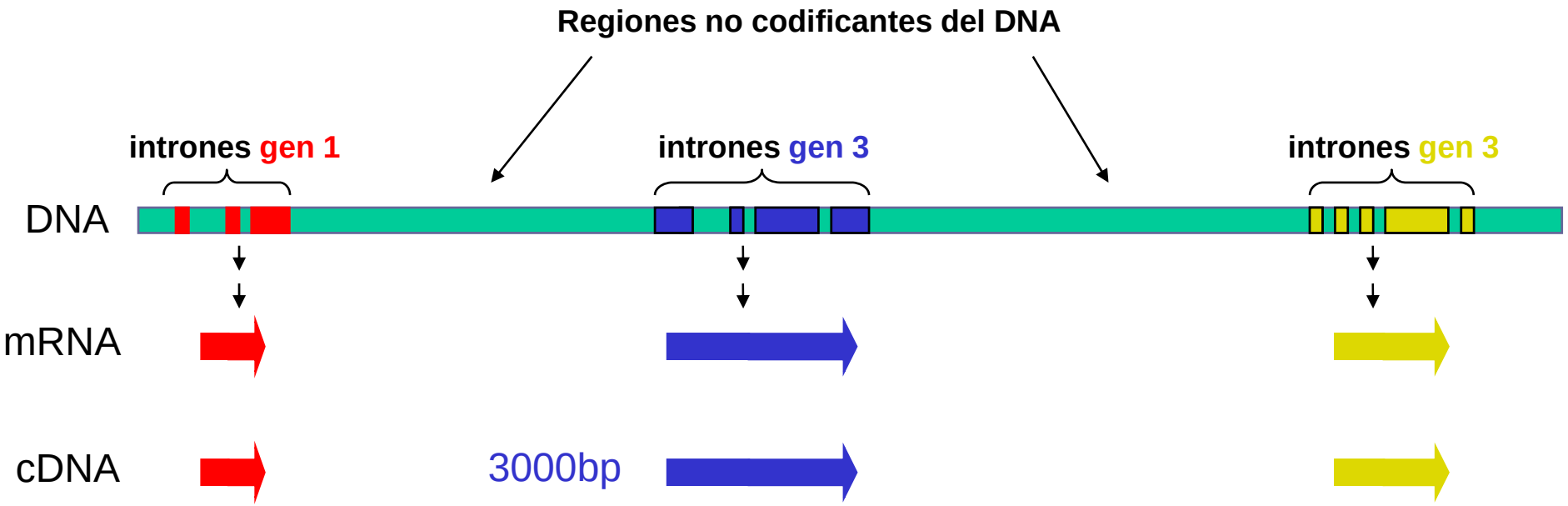
Bacterias



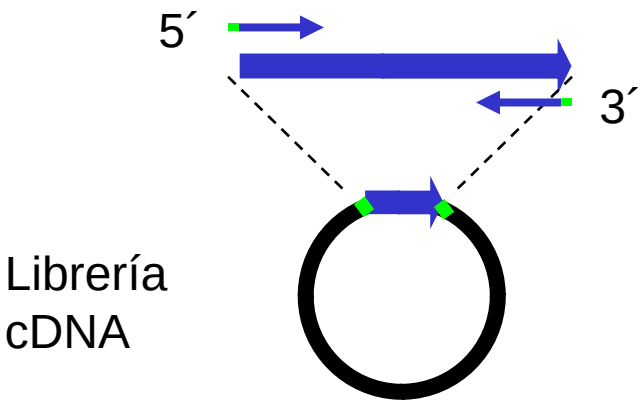

Eucariontes



# ESTs

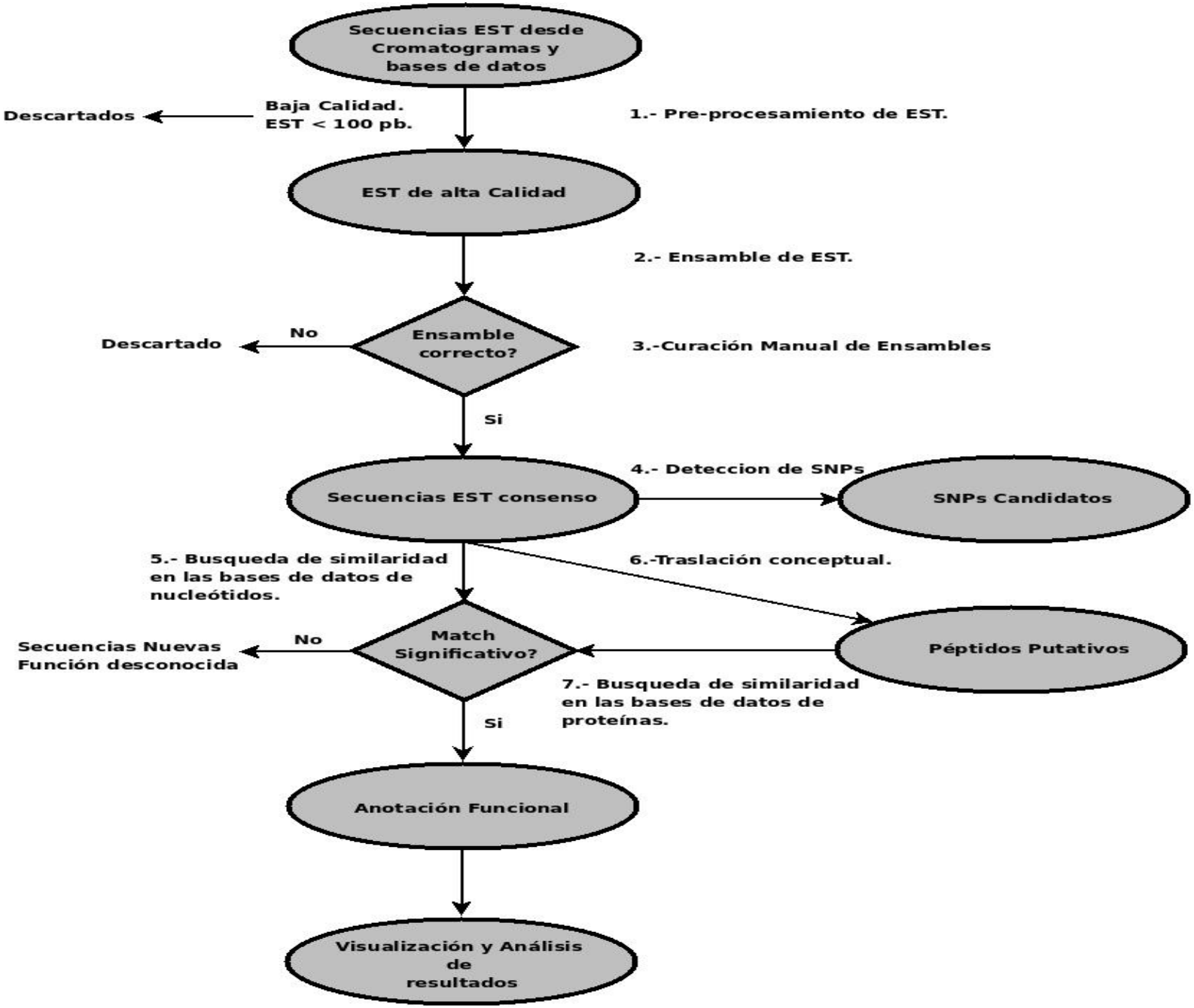



Secuencia nucleotidica de cada extremo- 600 bp-



# Avances proyecto
## genómica de salmones

Nombre : Alex Di Génova.


Secuencias EST desde Cromatogramas y bases de datos
Descartados
Baja Calidad. EST < 100 pb.
1.- Pre-procesamiento de EST.
EST de alta Calidad
2.- Ensamble de EST.
Descartado
No
Ensamble correcto?
3.-Curación Manual de Ensambles
Si
4.- Deteccion de SNPs
Secuencias EST consenso
SNPs Candidatos
5.- Busqueda de similaridad en las bases de datos de nucleótidos.
6.-Traslación conceptual.
Secuencias Nuevas Función desconocida
No
Match Significativo?
Péptidos Putativos
7.- Busqueda de similaridad en las bases de datos de proteínas.
Si
Anotación Funcional
Visualización y Análisis de resultados

# Bases de Datos

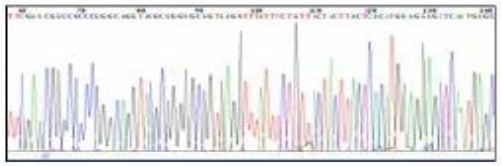
- cGRASP v/s NCBI:
  - Secuencias ESTs Salmo salar:

| | Total | Particular | Común |
|---|---|---|---|
| NCBI | 433337 | 139843 | 293494 |
| cGRASP | 487522 | 194028 | 293494 |
| | | | |
| TOTAL | 293494 | 333871 | 627365 |

# Ensamble

- Pre-Ensamble:
  - Determinación de Químicas de los cromatogramas.
  - Enmascarar Vector
    - Utilización de Univec (2862 sec).
    - PblueScriptIISK+, pCMVsport6 ect.
    - Crossmatch.
  - Enmascarar Poly-A
    - Seqclean y Masked.

# Ensamble



ACATGGAGGAGCTCAATGAGCTCATTTACGACAGAATCCACAACAAGA
TGATCTTCTTCTCCCTTTTCGTCCTACTTATCACCAGTAGAAACTGGA
TAGAAACATTTCCATTATATTACGTTATCAAGCTAGGCATTCTTCTCT
AAACTGAAGTTCTGGTTCAAGAGGAAACACCAAAAGCTCC

EST desde cromatogramas

EST desde base de datos

ACATGGAGGAGCTCAATGAGCTCATTTACGACAGAATCCACAACAAGA
TGATCTTCTTCTCCCTTTTCGTCCTACTTATCACCAGTAGAAACTGGA
TAGAAACATTTCCATTATATTACGTTATCAAGCTAGGCATTCTTCTCT
AAACTGAAGTTCTGGTTCAAGAGGAAACACCAAAAGCTCC

Identificar y remover
la secuencia del vector
(UNIVEC y EMVEC)

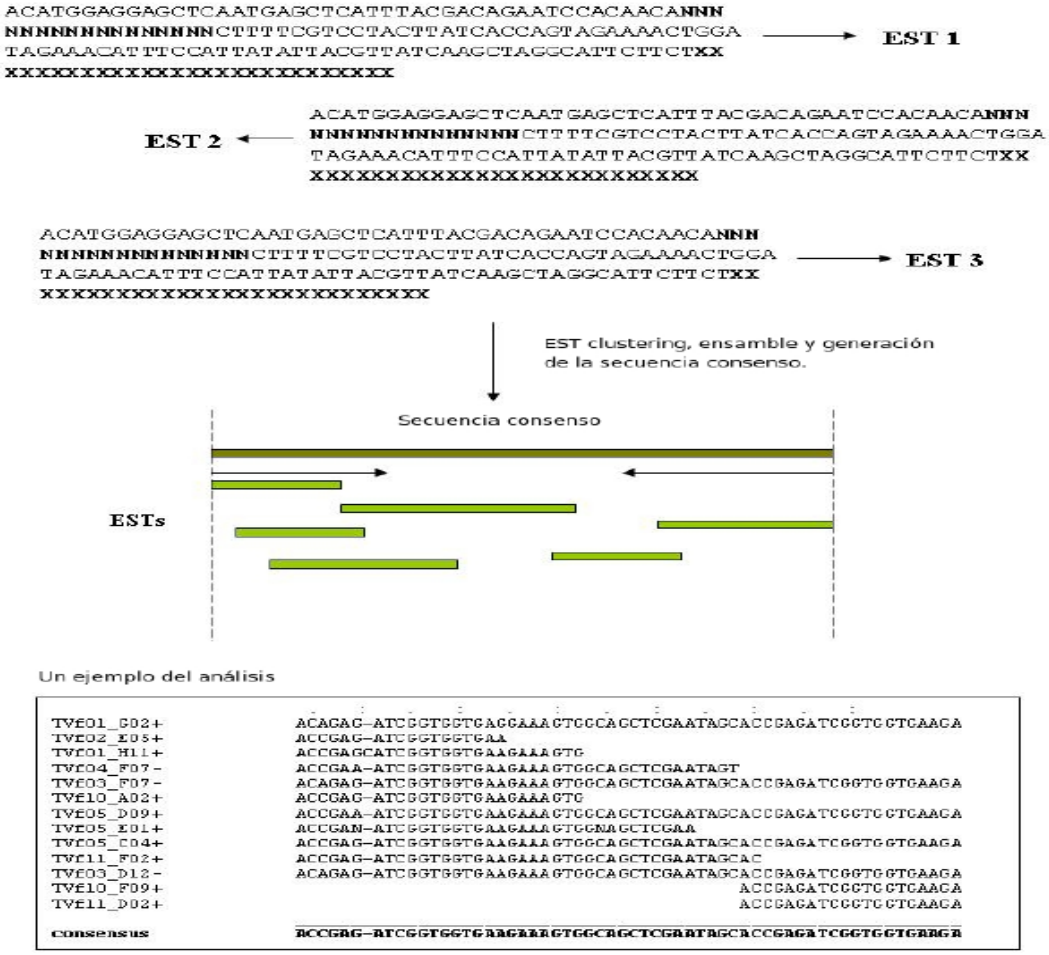
ACATGGAGGAGCTCAATGAGCTCATTTACGACAGAATCCACAACAAGA
TGATCTTCTTCTCCCTTTTCGTCCTACTTATCACCAGTAGAAACTGGA
TAGAAACATTTCCATTATATTACGTTATCAAGCTAGGCATTCTTCTXX
XXXXXXXXXXXXXXXXXXXXXXXXXX

Enmascaramiento de las
secuencias de baja
complejidad

ACATGGAGGAGCTCAATGAGCTCATTTACGACAGAATCCACAACANNN
NNNNNNNNNNNNNNNCTTTTCGTCCTACTTATCACCAGTAGAAAACTGGA
TAGAAACATTTCCATTATATTACGTTATCAAGCTAGGCATTCTTCTXX
XXXXXXXXXXXXXXXXXXXXXXXXXX

Secuecias de alta calidad para ensamblar

# Ensamble




Un ejemplo del análisis

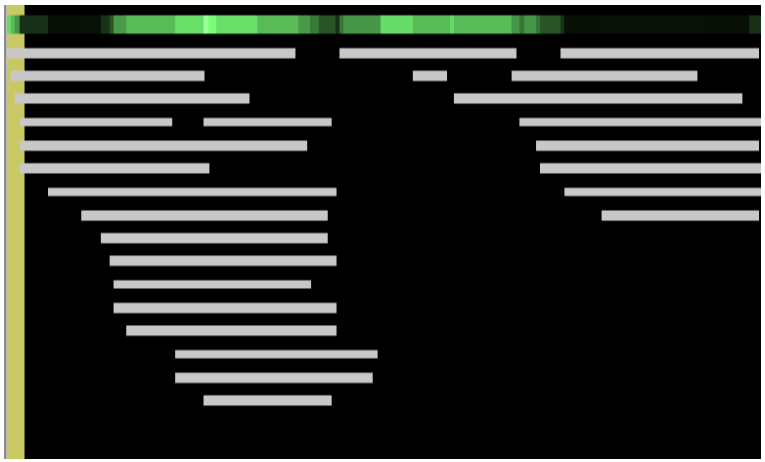
```
TVE01_G02+      ACAGAG-ATCGGTGGTGAGGAAAGTGGCAGCTCGAATAGCACCGAGATCGGTGGTGAAGA
TVE02_E05+      ACCGAG-ATCGGTGGTGAA
TVE01_H11+      ACCGAGCATCGGTGGTGAAGAAAGTG
TVE04_F07-      ACCGAA-ATCGGTGGTGAAGAAAGTGGCAGCTCGAATAGT
TVE03_F07-      ACAGAG-ATCGGTGGTGAAGAAAGTGGCAGCTCGAATAGCACCGAGATCGGTGGTGAAGA
TVE10_A02+      ACCGAG-ATCGGTGGTGAAGAAAGTG
TVE05_D09+      ACCGAA-ATCGGTGGTGAACAAAGTGGCAGCTCGAATAGCACCGAGATCGGTGGTGAACA
TVE05_E01+      ACCGAN-ATCGGTGGTGAAGAAAGTGGNAGCTCGAA
TVE05_C04+      ACCGAG-ATCGGTGGTGAAGAAAGTGGCAGCTCGAATAGCACCGAGATCGGTGGTGAAGA
TVE11_F02+      ACCGAG-ATCGGTGGTGAAGAAAGTGGCAGCTCGAATAGCAC
TVE03_D12-      ACAGAG-ATCGGTGGTGAAGAAAGTGGCAGCTCGAATAGCACCGAGATCGGTGGTGAAGA
TVE10_F09+                                                ACCGAGATCGGTGGTGAAGA
TVE11_D02+                                                ACCGAGATCGGTGGTGAAGA

consensus       ACCGAG-ATCGGTGGTGAAGAAAGTGGCAGCTCGAATAGCACCGAGATCGGTGGTGAAGA
```

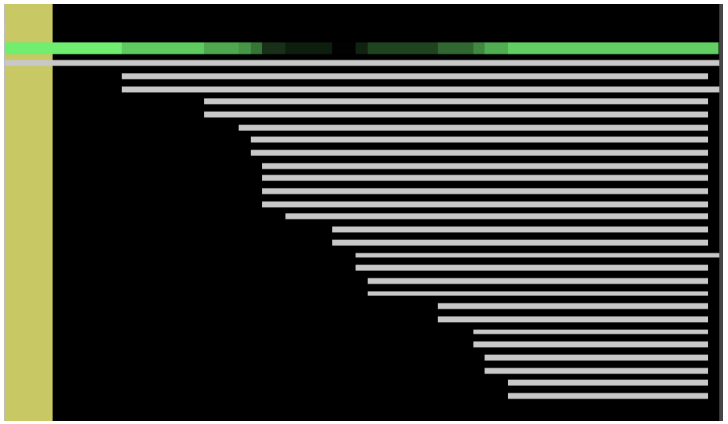
# Ensamble

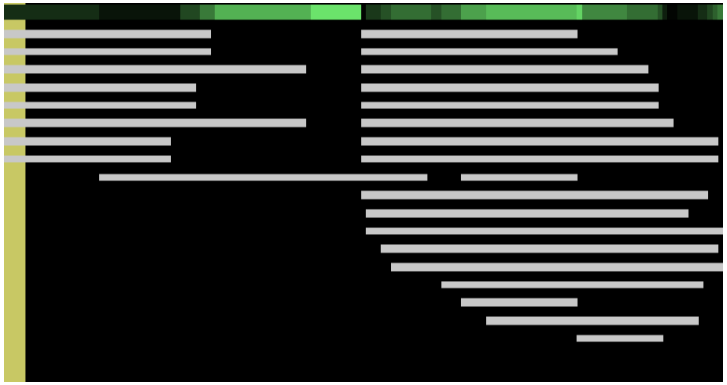
- Ensamble Salmo Salar:

| | |
|---|---|
| Numero de Lecturas | 485834 |
| Numero de Lecturas fwd | 291492 |
| Numero de Lecturas rev | 194342 |
| Mate pair | 189426 |
| Numero de Singlest | 111051 |
| Numero de Lecturas Ensamble | 374781 |
| Promedio Largo Lectura | 650 |
| Numero de Contig | 49297 |
| Promedio Largo Contig | 1218 |
| Desviación Estándar Largo Contig | 480 |
| Profundidad Promedio | 7 |
| Profundidad Máxima | 726 |

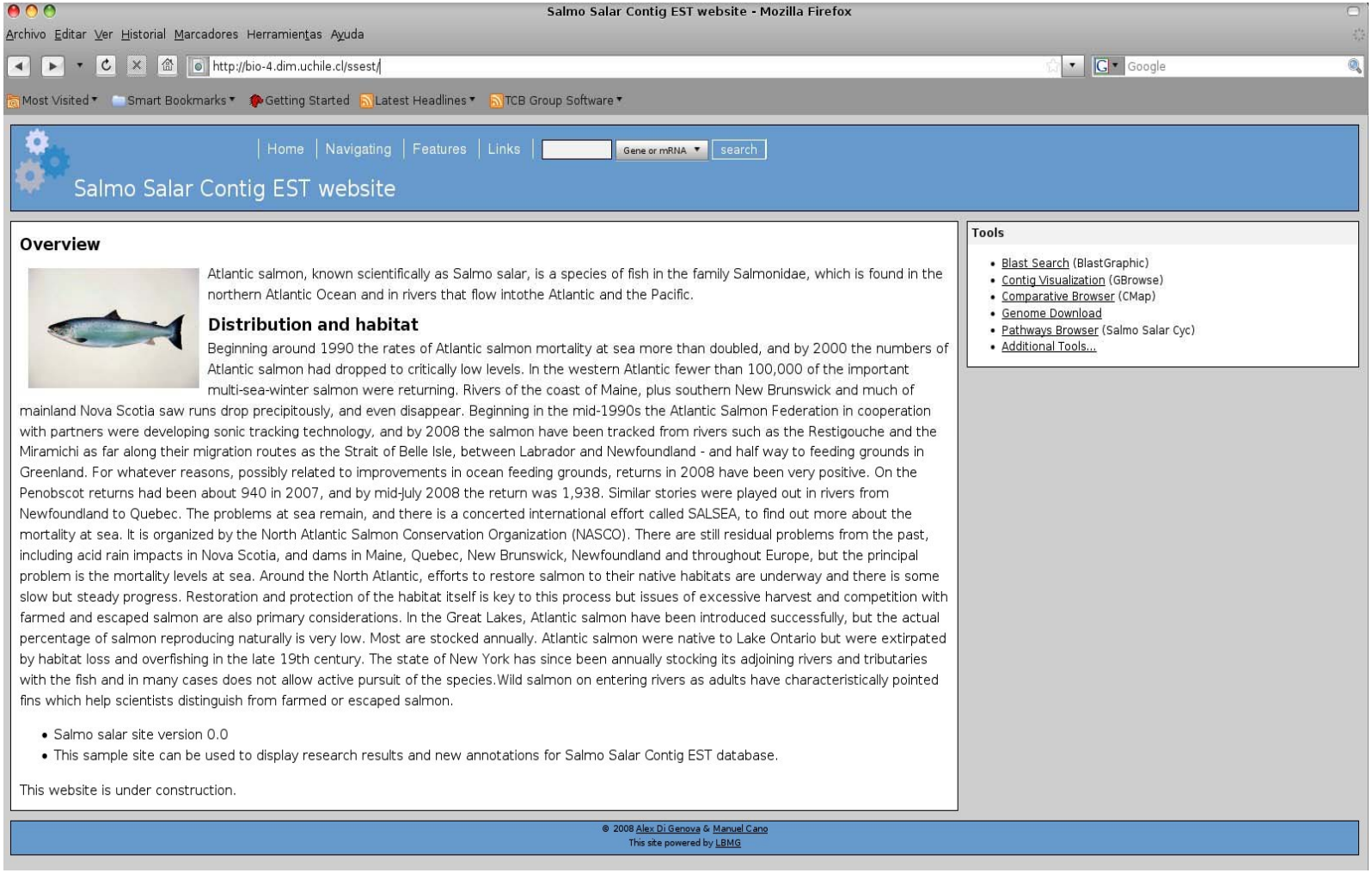
# Curación Ensambles







# Visualización Información



# Danio rerio

***Danio rerio***

version: 1.0

| Replicon | Total Genes | Protein Genes | RNA Genes | Pseudogenes | Size (bp) |
|---|---|---|---|---|---|
| Chromosome 1 (90 contigs) | 1211 | 1127 | 84 | 0 | |
| Chromosome 2 (125 contigs) | 1218 | 1105 | 113 | 0 | |
| Chromosome 3 (111 contigs) | 2276 | 1294 | 982 | 0 | |
| Chromosome 4 (89 contigs) | 2873 | 822 | 2051 | 0 | |
| Chromosome 5 (136 contigs) | 2087 | 1326 | 761 | 0 | |
| Chromosome 6 (117 contigs) | 1300 | 997 | 303 | 0 | |
| Chromosome 7 (103 contigs) | 1358 | 1239 | 119 | 0 | |
| Chromosome 8 (121 contigs) | 1780 | 1033 | 747 | 0 | |
| Chromosome 9 (87 contigs) | 979 | 843 | 136 | 0 | |
| Chromosome 10 (71 contigs) | 891 | 854 | 37 | 0 | |
| Chromosome 11 (83 contigs) | 861 | 807 | 54 | 0 | |
| Chromosome 12 (113 contigs) | 1176 | 835 | 341 | 0 | |
| Chromosome 13 (102 contigs) | 994 | 907 | 87 | 0 | |
| Chromosome 14 (111 contigs) | 2454 | 906 | 1548 | 0 | |
| Chromosome 15 (101 contigs) | 1251 | 934 | 317 | 0 | |
| Chromosome 16 (107 contigs) | 1483 | 1029 | 454 | 0 | |
| Chromosome 17 (101 contigs) | 1117 | 895 | 222 | 0 | |
| Chromosome 18 (93 contigs) | 893 | 795 | 98 | 0 | |
| Chromosome 19 (86 contigs) | 904 | 868 | 36 | 0 | |
| Chromosome 20 (83 contigs) | 1618 | 1076 | 542 | 0 | |
| Chromosome 21 (111 contigs) | 1936 | 900 | 1036 | 0 | |
| Chromosome 22 (79 contigs) | 1293 | 1037 | 256 | 0 | |
| Chromosome 23 (90 contigs) | 863 | 819 | 44 | 0 | |
| Chromosome 24 (90 contigs) | 983 | 648 | 335 | 0 | |
| Chromosome 25 (83 contigs) | 819 | 719 | 100 | 0 | |
| Chromosome MT (1 contigs) | 37 | 13 | 24 | 0 | |
| Total: | 34655 | 23828 | 10827 | 0 | |

| | |
|---|---|
| **Pathways:** | 123 |
| **Enzymatic Reactions:** | 837 |
| **Transport Reactions:** | 2 |
| **Polypeptides:** | 24229 |
| **Protein Complexes:** | 1 |
| **Enzymes:** | 913 |
| **Transporters:** | 6 |
| **Compounds:** | 702 |
| **Transcription Units:** | 0 |
| **tRNAs:** | 10211 |

Taxonomic lineage: **cellular organisms, Eukaryota, Fungi/Metazoa group, Metazoa, Eumetazoa, Bilateria, Coelomata, Deuterostomia, Chordata, Craniata, Vertebrata, Gnathostomata, Teleostomi, Euteleostomi, Actinopterygii, Actinopteri, Neopterygii, Teleostei, Elopocephala, Clupeocephala, Otocephala, Ostariophysi, Otophysi, Cypriniphysi, Cypriniformes, Cyprinoidea, Cyprinidae, Rasborinae, Danio, Danio rerio**

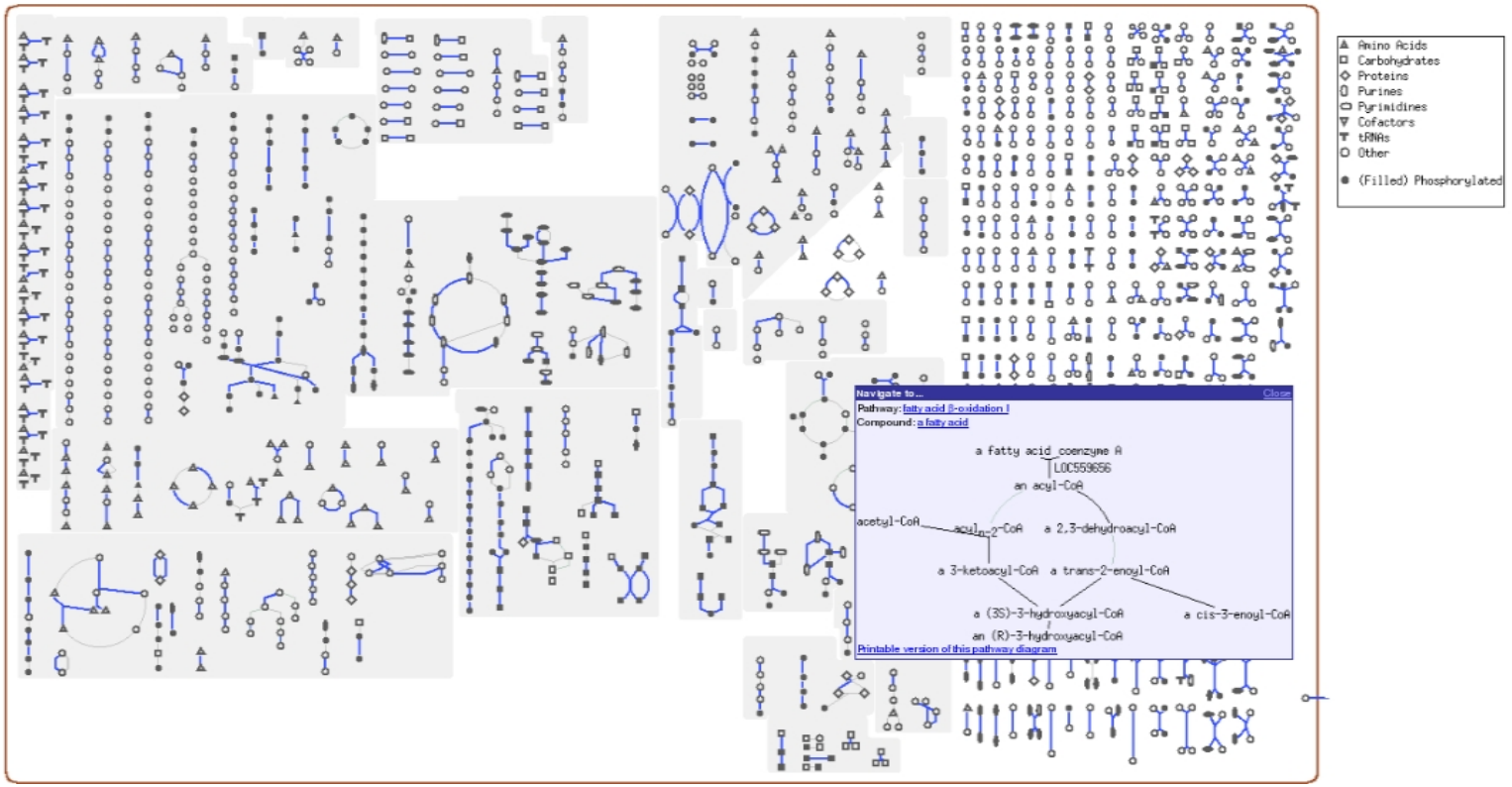
Unification Links: **NCBI-Taxonomy:7955**

# Danio rerio

**Overview of the *D. rerio* Metabolic Map**

This diagram provides a schematic of all pathways of *D. rerio* metabolism in the daniocyc database. Nodes represent metabolites, with shape indicating class of metabolite (see key to right). Lines represent reactions. Move the mouse over a metabolite icon to identify it. Click on a metabolite icon to navigate to the metabolite page or a related pathway page.

- Instructions
- Pathway Tools query page
- Omics Viewer: Paint omics data onto this diagram
- Species Comparison: Highlight reactions shared with other organisms



# Danio rerio

- Proteínas:

| | Number | Percentage |
|---|---|---|
| hypothetical protein | 7420 | 31,2 |
| know function | 16408 | 69,8 |
| total proteins | 23828 | 100 |