

# Projetos Genoma

MO640A - Biologia Computacional

*Felipe Rodrigues da Silva*

*Embrapa Recursos Genéticos e Biotecnologia*

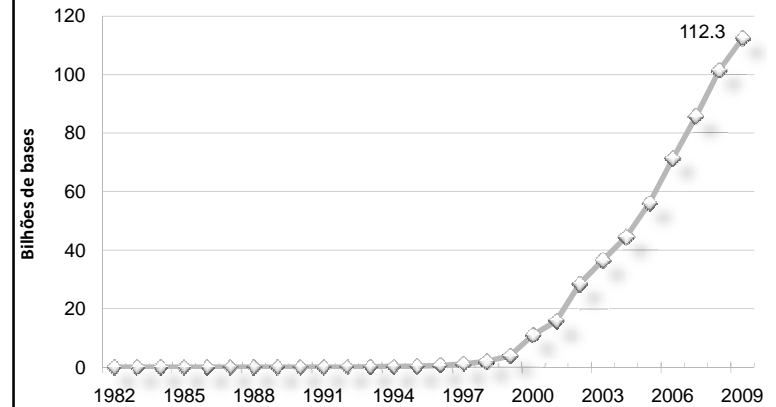


Recursos Genéticos e Biotecnologia

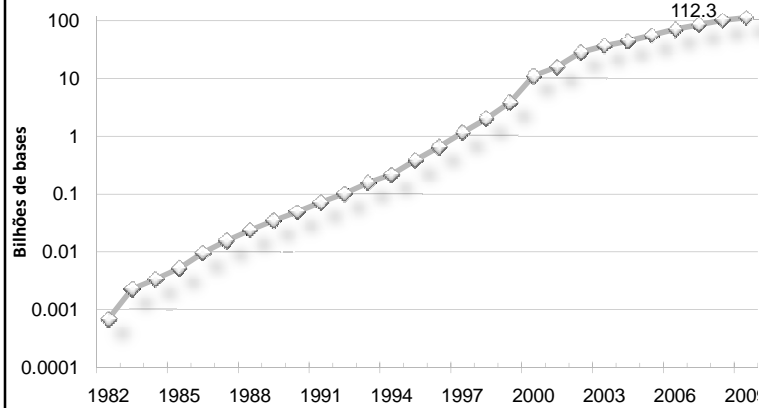


BRASIL GOV

GenBank 2010 (fevereiro)

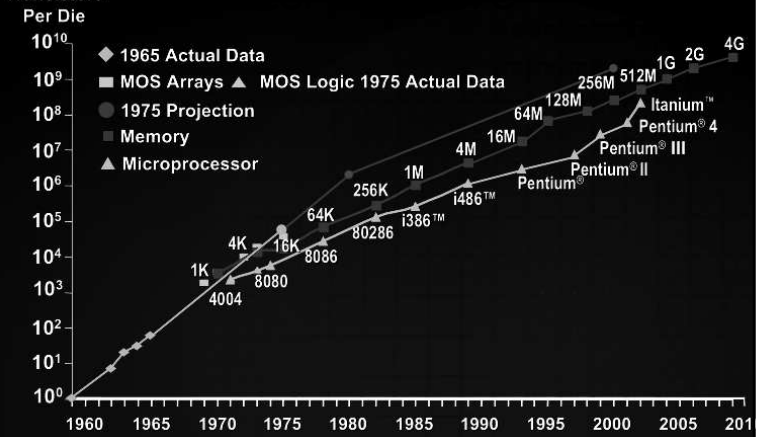


GenBank 2010 (fevereiro)

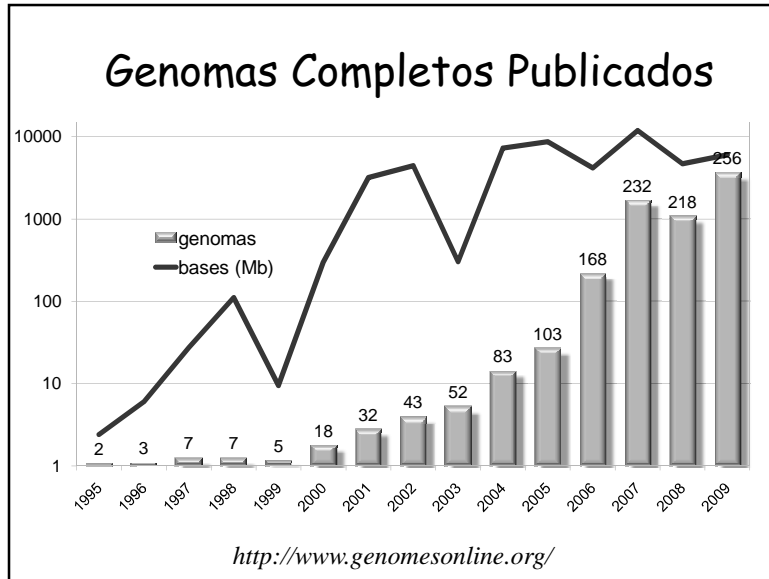


<http://www.ncbi.nlm.nih.gov/Genbank/>

Transistors



<http://www.intel.com/technology/mooreslaw/index.htm>



## Genomas Completos (fev 2010)

### 1.214

- Procariotos **1.010 + 80**
  - *Hodgkinia cicadicola* 143 Kb\*
  - *Mycoplasma genitalium* 580 Kb
  - *Sorangium cellulosum* 13.034 Kb
- Eucariotos **124**

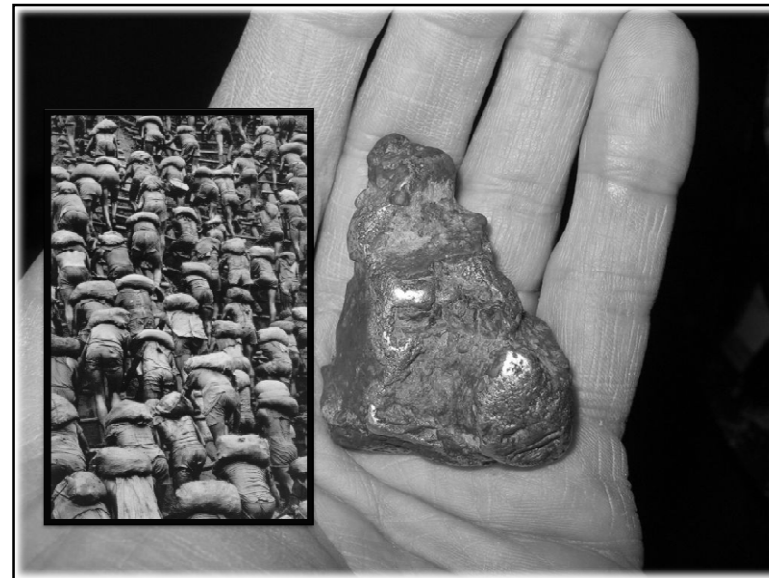
## Genomas em Andamento (fev 2010)

### 5.488

- Procariotos **3.963 + 202**
- Eucariotos **1.323**
  - Protozoa (*Trypanosoma cruzi*)
  - Algae (*Chlamydomonas*)
  - Fungy (*Candida albicans*)
  - Nematodes (*Ascaris suum*)
  - Plants (Maize, Wheat, Tomato, Cotto, Soy beam)
  - Insects (*Aedes egypti*, *Apis mellifera*)
  - Amphibious (*Xenopus*)
  - Birds (chicken)
  - Mammals (cow, dog, pig)

Mais 342 ESTs e 149 Metagenomas.....





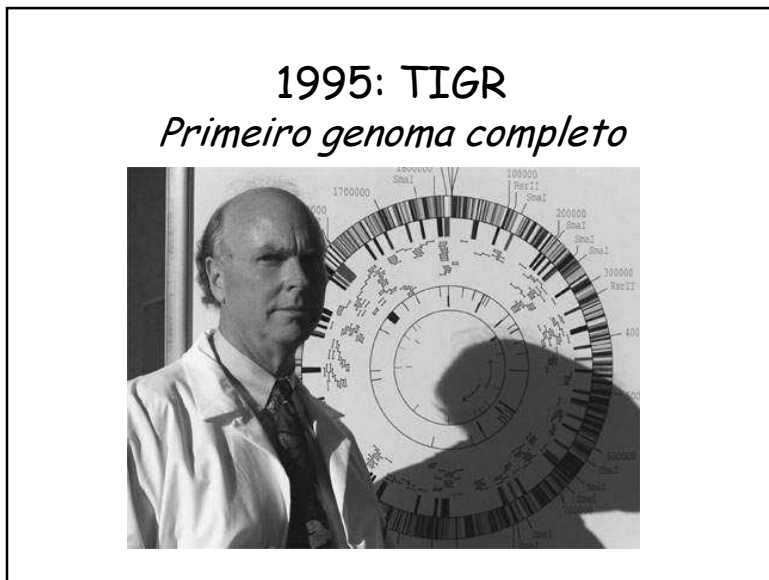
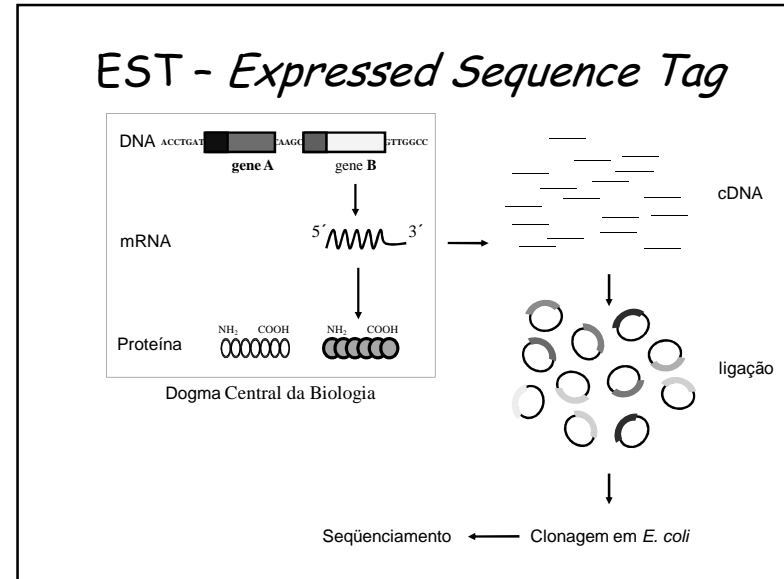
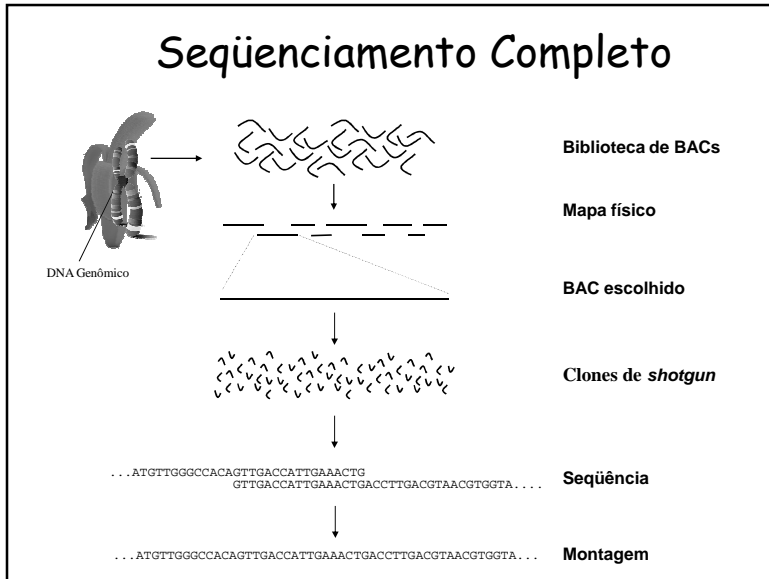
## Projeto Genoma

**Estrutural**  
Sequenciamento Completo do Genoma

– Região Gênica e Região Intergênica

**Funcional**  
EST – *Expressed Sequence Tag*

–Regiões que codificam proteínas (Genes)













## 1995: TIGR

### *Primeiro genoma completo*

**Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd**  
 Robert D. Fleischmann, Mark D. Adams, Owen White, Rebecca A. Clayton, Even F. Krieger, Anthony R. Kerlavage, Carl J. Sali, Jean-François Tomb, Brian A. Dougherty, Joseph M. Merrick, Keith McKenney, Granger Sutton, Will Fitzhugh, Chris Fields, Jeannine D. Gocayne, John Scott, Robert Shirley, Li-Ing Liu, Anna Glöckle, Jeremy M. Kelley, Janice E. Weidman, Cheryl A. Phillips, Tracy Spriggs, Eva Hedblom, Matthew D. Cotton, Teresa R. Utterback, Michael C. Harris, David T. Nguyen, Deborah M. Saudik, Rhonda C. Brandon, Leah D. Fine, Janice L. Fritchman, Joyce L. Fuhrmann, K. S. M. Geoghagen, Cheryl L. Gnehm, Lisa A. McDonald, Keith V. Small, Claire M. Fraser, Hamilton O. Smith, J. Craig Venter




**28 de Julho de 1995**  
*Fleischmann et al. (1995), Science 269:496*

**Bactéria**  
 1,6 Mb - 1.700 genes

<p><b>1997</b> Eukaryote 13 Mb - 6.000 genes [Nature 387:1]</p> 	<p><b>1998</b> Animal 100 Mb - 18.000 genes [Science 282:1945]</p> 	<p><b>2000</b> Insect 130 Mb - 13.000 genes [Science 287:2185]</p> 	<p><b>2000</b> Plant 120 Mb - 26.000 genes [Nature 408:791]</p> 	<p><b>2000</b> Bacteria - Plant pathogen 2,6 Mb - 3.000 genes [Nature 406:151]</p> 
<p><b>2001</b> Human 3,2 Gb - 40.000 genes [Science 291:1304]</p> 	<p><b>2001</b> Human 3,2 Gb - 40.000 genes [Nature 409:745]</p> 	<p><b>2002</b> Plant 420 Mb - 35.000 genes [Science 296:79]</p> 	<p><b>2002</b> Parasite - Host 23 Mb - 5.300 genes [Nature 419:498] 278Mb - 14.000 genes [Science 298:129]</p> 	<p><b>2002</b> Mouse 2,5 Gb - 30.000 genes [Nature 420:520]</p> 

## A Rede ONSA

- *Xylella fastidiosa*
- *Xanthomonas axonopodis* pv *citri*
- Sugar Cane EST
- Human Cancer EST
- *Xanthomonas campestris*
- AEG
  - *X. fastidiosa* / Pierce's Disease
  - *Leifsonia xyli* subsp. *Xyli*
  - eucalipto

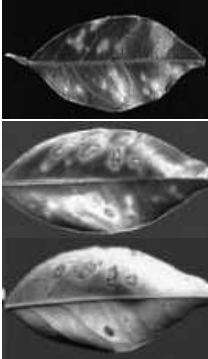

*The Virtual Genomics Institute*




# Genoma Xylella

## CVC

*clorose variegada dos citros*

## Alguns vetores de *X. fastidiosa*



*Dilobopterus costalimai*



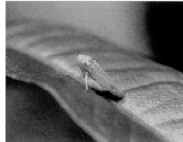
*Acrogonia terminalis*



*Oncometopia facialis*



*Homolodisca ignorata*

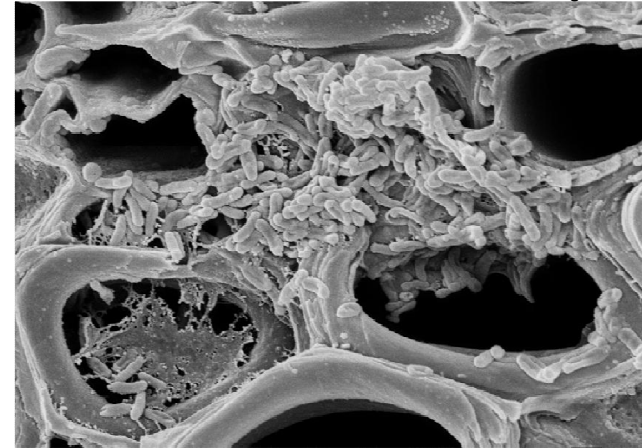


*Bucephalagonia xanthopis*



*Macugonalia sp*

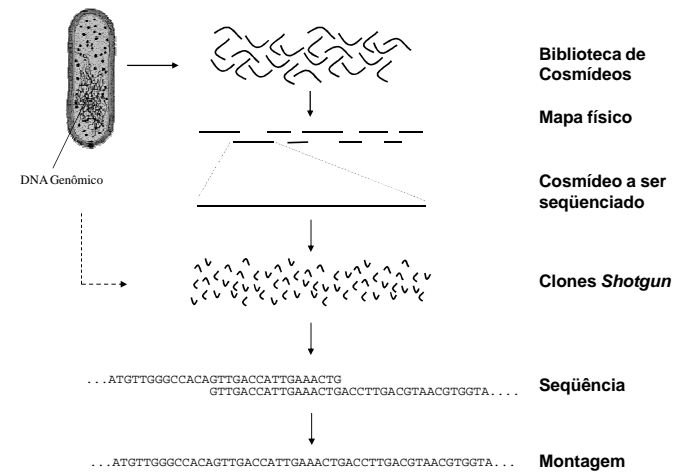
## *X. fastidiosa* em xilema de laranjeira



## Genomas

- Seqüenciamento
- Montagem
- Análise
- Anotação
- Pós-genoma

## Seqüenciamento Completo



## Estratégia de seqüenciamento

- Pequena escala
  - Leitura única
  - Deleções sucessivas
  - Subclonagem
  - *Primer walking*
  - *shotgun*

## Shotgun

- Amostrar fragmentos da seqüência-alvo da maneira mais aleatória possível.
- Determinar a maior porção possível das seqüências das extremidades destes fragmentos

Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. (1982) **Nucleotide sequence of bacteriophage lambda DNA.** *J Mol Biol* 162(4): 729-73.

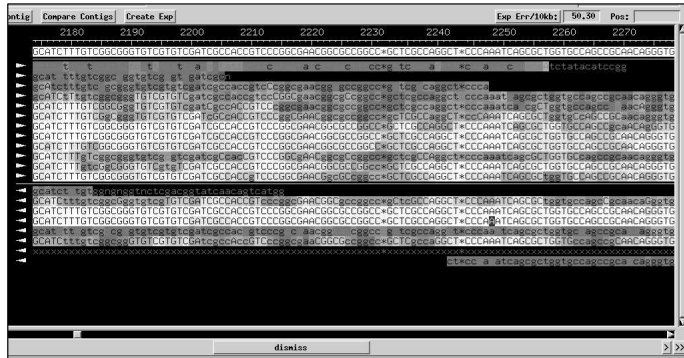
## Montagem shotgun



## Montagem

- *Trimming* (corte)
  - fixo
  - por ambigüidade
  - por qualidade
- *Consenso*
  - inclusivo (código de ambigüidades)
  - por freqüência
  - por qualidade

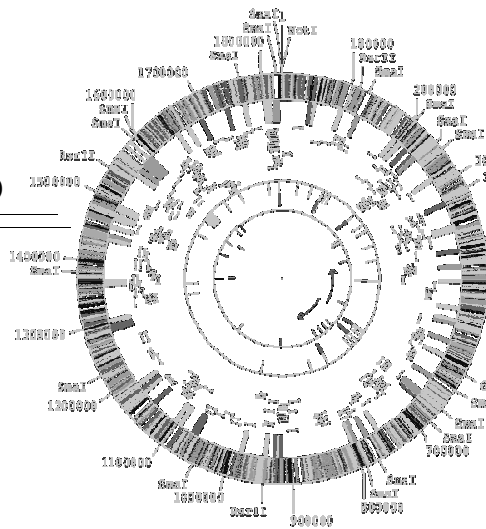
## Montagem com *phrap*



## Análise

- Reads
  - sem dados
  - ruins
  - errados
  - vetor
  - hospedeiro
- Montagem
  - gaps
  - conflitos
  - quimeras/deleções
- Checagem de co-linearidade

## Anotação



## Anotação

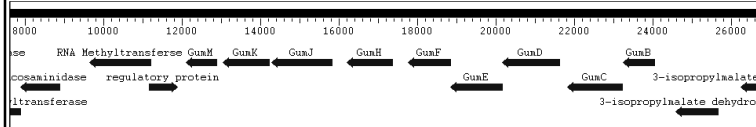
- ORFs
  - localização
  - identificação
- RNAs
- Seqüências regulatórias



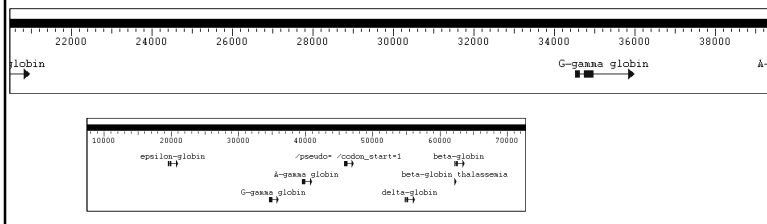


# Organização dos genomas

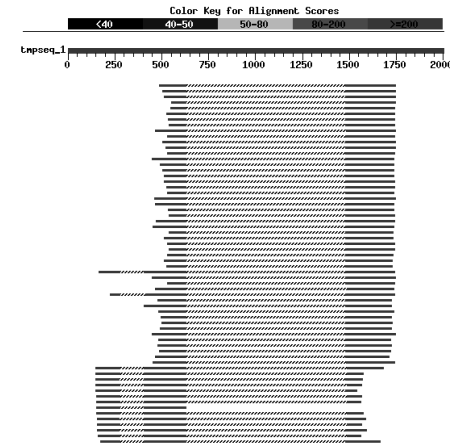
- Operon Gum de *Xylella*



- Região da beta globina humana no cromossomo 11



# Blastn β-globina (62000 to 64000)



# Blastn β-globina (62000 to 64000)

Sequences producing significant alignments:

	Score (bits)	E Value
emb F24593.1 F24593 HSPD11021 HM3 Homo sapiens cDNA clone s...	523	e-146
emb F26056.1 F26056 HSPD13352 HM3 Homo sapiens cDNA clone s...	523	e-146
gb T41162 T41162 ya31f08.s3 Homo sapiens cDNA clone 62247 3...	523	e-146
emb F27352.1 F27352 HSPD15247 HM3 Homo sapiens cDNA clone s...	523	e-146
emb F33173.1 F33173 HSPD26430 HM3 Homo sapiens cDNA clone s...	521	e-145
emb F21808.1 F21808 HSPD06401 HM3 Homo sapiens cDNA clone 0...	519	e-145
emb F21643.1 F21643 HSPD06227 HM3 Homo sapiens cDNA clone 0...	519	e-145
emb F20795.1 F20795 HSPD05201 HM3 Homo sapiens cDNA clone 0...	519	e-145
emb F31552.1 F31552 HSPD22843 HM3 Homo sapiens cDNA clone s...	517	e-144
gb R06757 R06757 yf11d02.s1 Homo sapiens cDNA clone 126531 ...	517	e-144
emb F26092.1 F26092 HSPD13402 HM3 Homo sapiens cDNA clone s...	515	e-143
emb F28594.1 F28594 HSPD17840 HM3 Homo sapiens cDNA clone s...	515	e-143
emb F24982.1 F24982 HSPD11692 HM3 Homo sapiens cDNA clone s...	515	e-143
emb F37427.1 F37427 HSPD36142 HM3 Homo sapiens cDNA clone s...	513	e-143
gb R91762 R91762 yp98b11.s1 Homo sapiens cDNA clone 195453 ...	513	e-143
gb AA296097 AA296097 EST10430 Adipose tissue, white I Homo ...	513	e-143
gb R93060 R93060 yq13c07.s1 Homo sapiens cDNA clone 196812 ...	513	e-143
gb R91514 R91514 yq09c11.s1 Homo sapiens cDNA clone 196436 ...	513	e-143
emb F31377.1 F31377 HSPD22507 HM3 Homo sapiens cDNA clone s...	513	e-143
emb F36284.1 F36284 HSPD33731 HM3 Homo sapiens cDNA clone s...	511	e-142
emb F24336.1 F24336 HSPD10580 HM3 Homo sapiens cDNA clone s...	509	e-142
emb F29589.1 F29589 HSPD19534 HM3 Homo sapiens cDNA clone s...	509	e-142
gb H95645 H95645 yv17b07.s1 Soares fetal liver spleen INFLS...	509	e-142

# Blastn β-globina (62000 to 64000)

emb|F26056.1|F26056 HSPD13352 HM3 Homo sapiens cDNA clone s4000047F08, mRNA sequence  
Length = 430

Score = 523 bits (264), Expect = e-146  
Identities = 264/264 (100%)  
Strand = Plus / Plus

```

Query: 1482 gtcctctgggcaacgtgctgtctgtgtgctggcccatcactttggcaagaattcaacccc 1541
          |||
Sbjct: 155 gtcctctgggcaacgtgctgtctgtgtgctggcccatcactttggcaagaattcaacccc 214

Query: 1542 accagtgcaggctgcctatcagaagaagggtggctggctgaatgcctggcccaaa 1601
          |||
Sbjct: 215 accagtgcaggctgcctatcagaagaagggtggctggctgaatgcctggcccaaa 274

Query: 1602 gtatcactaaagctcgtttcttctgtgtccaatttctattaagggttcccttgttccctaa 1661
          |||
Sbjct: 275 gtatcactaaagctcgtttcttctgtgtccaatttctattaagggttcccttgttccctaa 334

Query: 1662 gtccaactactaaactgggggatattatgaaggccttagcatctggattctgctcaat 1721
          |||
Sbjct: 335 gtccaactactaaactgggggatattatgaaggccttagcatctggattctgctcaat 394

Query: 1722 aaaaaacattttttcattgcaa 1745
          |||
Sbjct: 395 aaaaaacattttttcattgcaa 418
    
```

## Blastn $\beta$ -globina (62000 to 64000)

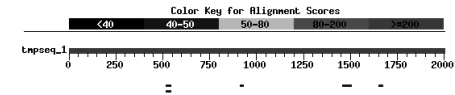
Score = 283 bits (143), Expect = 7e-74  
 Identities = 143/143 (100%)  
 Strand = Plus / Plus

Query: 490 accctaagggtgaaggctcatggcaagaagaagtgcggtgccttagtgatggcctggtc 549  
 |||||||  
 Sbjct: 13 accctaagggtgaaggctcatggcaagaagaagtgcggtgccttagtgatggcctggtc 72

Query: 550 acctggacaacctcaagggcacctttgccacactgagtgagctgcaactgtgacaagctgc 609  
 |||||||  
 Sbjct: 73 acctggacaacctcaagggcacctttgccacactgagtgagctgcaactgtgacaagctgc 132

Query: 610 acgtggatcctgagaacttcagg 632  
 |||||||  
 Sbjct: 133 acgtggatcctgagaacttcagg 155

## Blastn Operon Gum (21000 - 23000)



Sequences producing significant alignments:

	Score	E
	(bits)	Value

gb U22511 XCU22511 Xanthomonas campestris GumA, GumB, GumC...	46	0.034
emb Z68013 CEW02H3 Caenorhabditis elegans cosmid W02H3, com...	42	0.54
gb AF025669 AF025669 Tribolium castaneum GABA receptor subu...	40	2.1
emb X00274 HSHL07 Human gene for HLA-DR alpha heavy chain a...	40	2.1
gb AF172352.1 AF172352 Ceratitis capitata GABA receptor sub...	40	2.1

gb|U22511|XCU22511 Xanthomonas campestris GumA, GumB, GumC, GumD, GumE, GumF, GumG, GumH, GumI, GumJ, GumK, GumL, and GumM genes, complete cds.  
 Length = 16075

Score = 46.1 bits (23), Expect = 0.034  
 Identities = 41/47 (87%)  
 Strand = Plus / Minus

Query: 1458 ctgtgcgaggccagcaacgcattaagatcactcaggttctgtgcag 1504  
 |||||||  
 Sbjct: 2779 ctgtgcgaggccagcaacgcattgagatcggtcagattctgtgcag 2733

## Blastn Operon Gum (21000 - 23000)

emb|Z68013|CEW02H3 Caenorhabditis elegans cosmid W02H3, complete sequence  
 [Caenorhabditis elegans]  
 Length = 17716

Score = 42.1 bits (21), Expect = 0.54  
 Identities = 21/21 (100%)  
 Strand = Plus / Plus

Query: 1653 acgctgctgattattttcaat 1673  
 |||||||  
 Sbjct: 9659 acgctgctgattattttcaat 9679

gb|AF025669|AF025669 Tribolium castaneum GABA receptor subunit (Rd1) gene, partial cds  
 Length = 836

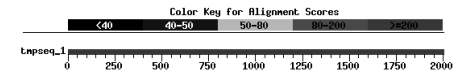
Score = 40.1 bits (20), Expect = 2.1  
 Identities = 23/24 (95%)  
 Strand = Plus / Plus

Query: 520 gatcaaaaaacgaagatttagat 543  
 |||||||  
 Sbjct: 565 gatcaataacacgaagatttagat 588

(...)

Database: Non-redundant GenBank+EMBL+DDBJ+PDB sequences  
 Posted date: Sep 22, 1999 2:18 PM  
 Number of letters in database: 1,306,157,801  
 Number of sequences in database: 477,403

## Blastx Operon Gum (21000 - 23000)



Sequences producing significant alignments:

	Score	E
	(bits)	Value

pir  S67819 GumC protein - Xanthomonas campestris >gi 11720...	419	e-116
pir  S67820 GumD protein - Xanthomonas campestris >gi 73314...	263	5e-69
pir  S31721 xps2A protein - Xanthomonas campestris >gi 4854...	258	1e-67
gb AAD43842.1 AF076290_12 (AF076290) Bme12 [Brucella melite...	80	8e-14
sp P33698 EXOP_RHIME SUCCINOGLYCAN BIOSYNTHESIS TRANSPORT P...	76	9e-13
pir  S37031 exoP protein - Rhizobium meliloti >gi 742331 pr...	68	2e-10
gi 3493609 (AF067140) exopolysaccharide polymerization prot...	68	3e-10
gi 2921558 (AF039306) integral membrane protein [Bradyrhizo...	67	4e-10
gb AAD35728.1 AE001738_8 (AE001738) hypothetical protein [T...	63	6e-09
emb CAA75431  (Y15162) ptk [Acinetobacter johnsonii]	54	4e-06
gb AAD34733.1 AF131869_2 (AF131869) putative undecaprenylph...	43	0.011
gi 3449364 (AF032862) intracellular hyaluronic acid binding...	41	0.024
pir  JC5016 hyaluronan receptor - human	41	0.041
sp Q46631 AMSA_ERWAM AMYLOVORAN BIOSYNTHESIS PUTATIVE MEMBR...	40	0.053
gb AAD32394.1 AAD32394 (AF065404) pXO1-90 [Bacillus anthracis]	40	0.070

# Blastx Operon Gum (21000 - 23000)

pir||867819 GumC protein - Xanthomonas campestris >gi|1172092 (U22511) GumC [Xanthomonas campestris]  
Length = 449

Score = 419 bits (1066), Expect = e-116  
Identities = 239/378 (63%), Positives = 310/378 (81%), Gaps = 6/378 (1%)  
Frame = -2

Query: 2000 IPVESPLDRDFQYQYQLQSSLSARAVIRKMNLDREPMLKPLVKVLSKVNQ-----1839  
+RER +E +L + +L L +EPLINSRIVV++DS DP L+K+AN Y K+P + Q  
+EVESE DRDFQYQYQLQSSLSARAVIR+ LD+EP K V++ L+K  
Sbjct: 58 MEVESPQDRDFQYQYQLQSSLSARAVIREAKLDQEPFAFKEQVEEALAKAEKNPEAK 117

Query: 1838 FVNTRNTAESALEMMLNSLKEIPILNSRLVYVHFDSDPTLSARVANAYAKMFINNQ 1659  
+RER +E +L + +L L +EPLINSRIVV++DS DP L+K+AN Y K+P + Q  
+EVESE DRDFQYQYQLQSSLSARAVIR+ LD+EP K V++ L+K  
Sbjct: 118 SLDSRQALVERSLTDTLLAGLVEPILNSRLVYVYDSDPFLAAKIANYPKVPVISTQ 177

Query: 1658 QRRSNAPSFAMKYLAGLEQLKLVKDESERNVAYSTDEKIVSVGDEKPSLAQNLSDIN 1479  
+RER +E +L + +L L +EPLINSRIVV++DS DP L+K+AN Y K+P + Q  
+EVESE DRDFQYQYQLQSSLSARAVIR+ LD+EP K V++ L+K  
Sbjct: 178 ERKMKASSFATPLAEKRLKQREKVESEKDLVSVSTEQIVSGDDKPSLAQNLSDIN 239

Query: 1478 ALLASAQNERIAREASMKQASIGDGLSIPQVLSNLLVQSLRTEQANNVYQKLSMFKP 1237  
+RER +E +L + +L L +EPLINSRIVV++DS DP L+K+AN Y K+P + Q  
+EVESE DRDFQYQYQLQSSLSARAVIR+ LD+EP K V++ L+K  
Sbjct: 238 ALLASAQARIKASAWRQASSGDGMSLPQVLSNLLVQSLRTEQANNVYQKLSMFKP 297

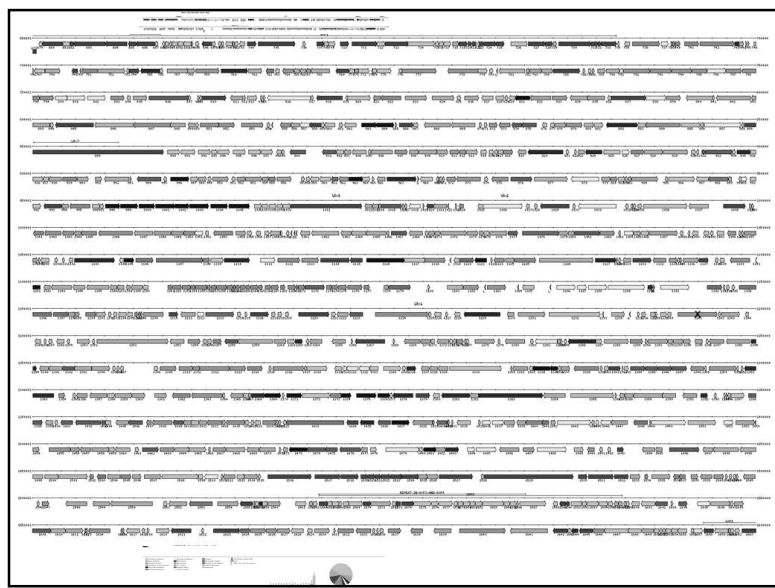
Query: 1298 EYFEMQRKARIKENTQINAEVINRQSLKSOVEATRQEMLINDRIAVLEKDELKLT 1119  
+RER +E +L + +L L +EPLINSRIVV++DS DP L+K+AN Y K+P + Q  
+EVESE DRDFQYQYQLQSSLSARAVIR+ LD+EP K V++ L+K  
Sbjct: 298 DYFEMQRKAKIERSRQINGEVINIRQSLKATDASVHQEGLINDRIAGLSNDELQ 357

Query: 1118 RLIRYNLREARETRQLYDALLQRYKEISLVGDVGSNNVTVDAAID PRSPISPNLLK 939  
+RER +E +L + +L L +EPLINSRIVV++DS DP L+K+AN Y K+P + Q  
+EVESE DRDFQYQYQLQSSLSARAVIR+ LD+EP K V++ L+K  
Sbjct: 358 RLIRYNLKRDRVDNRLQYDALLQRYKEISLVGSVGNANNVTVDVDTKPTSPKLIK 417

Query: 938 TILGGIPGVFLGLTVMVRYSMHG 867  
LG IPGVFLG+ VA+VRY + G  
Sbjct: 418 LALGLIPGVFLGVALVRYFLR 441

GumB (654 bp) (217 aa) | GumC (1404 bp) (467 aa) | GumD (1455 bp) (484 aa) | GumE (1275 bp) (424 aa) | GumF (1092 bp) (363 aa) | GumG (1143 bp) (380 aa) | GumH (1533 bp) (510 aa) | GumI (846 bp) (281 aa) | GumM (798 bp) (265 aa)

• xylella main chromosome from 2235870 to 2247690 (reverse complemented)



## O genoma publicado de Xylella fastidiosa

• Mar/98 – Jan/00  
• 2,7 Mbp  
• 2,904 ORFs

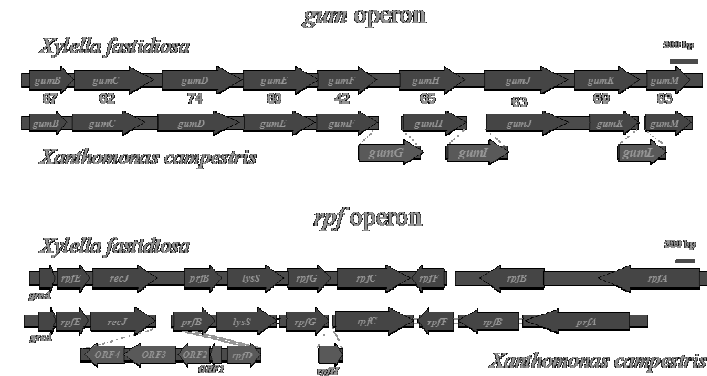
Nature, (July 13, 2000) 406:151-157.

## Grupo do genoma da *Xylella*

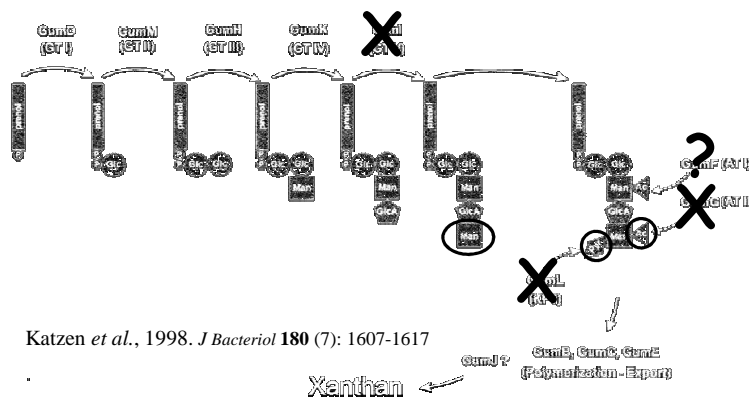


<http://ac>

## Mapa genético dos operons *gum* e *rpf*



## Via sintética da goma xantana

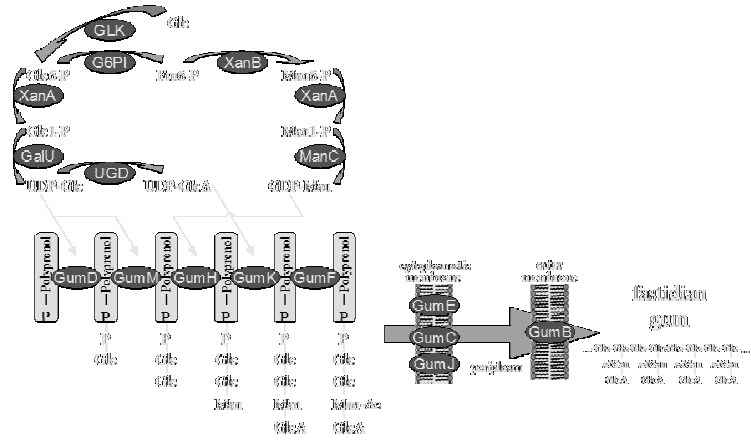


Katzen *et al.*, 1998. *J Bacteriol* **180** (7): 1607-1617

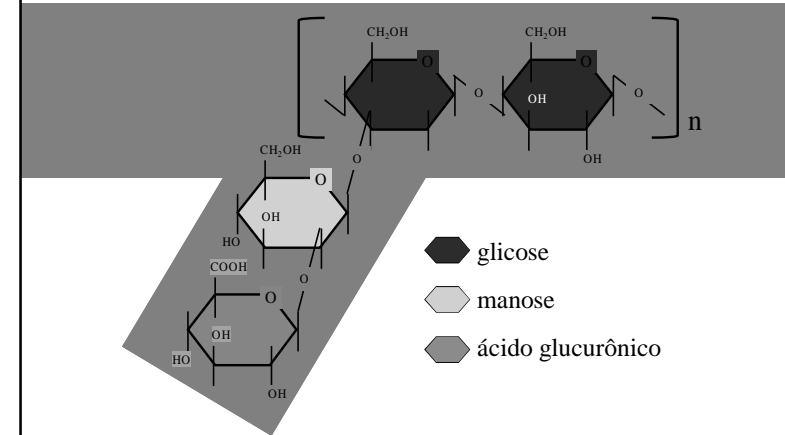
Xanthan

classe	gene ID	gene homologo	identidade	função
Regulador	XF0287	<i>rpfB</i>	72.3	regulatory protein (DSF)
Regulador	XF0290	<i>rpfA</i>	80.0	aconitase
Regulador	XF1109	<i>rpfE</i>	65.2	regulatory protein
Regulador	XF1113	<i>rpfG</i>	77.0	two-component system, regulatory protein
Regulador	XF1114	<i>rpfC</i>	60.0	fused two-component sensor-regulator protein
Regulador	XF1115	<i>rpfF</i>	65.7	regulatory protein (DSF)
Precursor	XF0232	<i>pgi</i>	79.1	glucose-6-phosphate isomerase
Precursor	XF0259	<i>xanB</i>	84.5	phosphomannose isomerase-GDP-mannose pyrophosphorylase
Precursor	XF0260	<i>xanA</i>	84.8	phosphoglucomutase /phosphomannomutase
Precursor	XF1064	<i>glk</i>	41.4	glucose kinase
Precursor	XF1460	<i>glk</i>	32.7	glucose kinase
Precursor	XF1606	<i>algD</i>	66.1	UDP-glucose dehydrogenase
Precursor	XF2432	<i>gtaB</i>	81.8	UTP-glucose-1-phosphate uridylyltransferase
EPS-sint	XF2360	<i>gumM</i>	63.1	GumM protein
EPS-sint	XF2361	<i>gumK</i>	68.7	GumK protein
EPS-sint	XF2364	<i>gumH</i>	64.7	GumH protein
EPS-sint	XF2365	<i>gumF</i>	41.9	GumF protein
EPS-sint	XF2367	<i>gumD</i>	73.6	GumD protein
EPS-exp	XF2362	<i>gumJ</i>	62.7	GumJ protein
EPS-exp	XF2366	<i>gumE</i>	59.9	GumE protein
EPS-exp	XF2369	<i>gumC</i>	61.2	GumC protein
EPS-exp	XF2370	<i>gumB</i>	67.1	GumB protein

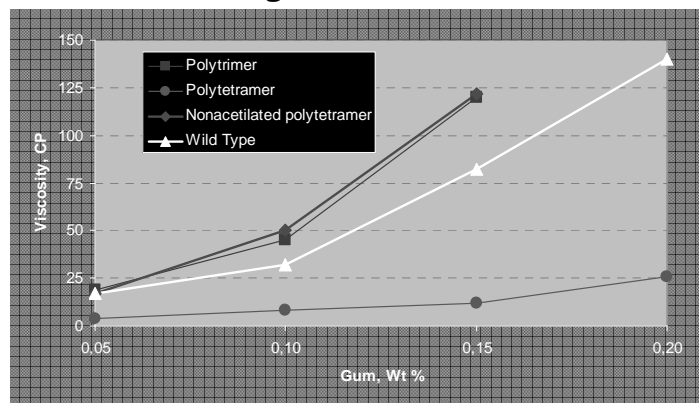
## Via sintética da goma fastidiana



## Possível estrutura da goma fastidiana



## Viscosidades de polímeros derivados da goma xantana



Hassler & Doherty, 1990. *Biotechnol. Prog.* **6** (3): 182-187

## Grupo Genoma - CBMEG





<http://www.cenargen.embrapa.br/~felipes/>

[felipes@cnptia.embrapa.br](mailto:felipes@cnptia.embrapa.br)