Fall 2005 Genomics Exam #1
Genomic Sequences

There is no time limit on this test, though I don't want you to spend too much time on it. I work hard to design challenging tests that continue your learning and hopefully will stimulate you too. You do not need to read any additional papers other than the ones I send to you. There are 4 pages, including this cover sheet, and 5 questions for this test. You are not allowed discuss the test with anyone until all exams are turned in at 11:30 am on Friday September 30. **EXAMS ARE DUE AT CLASS TIME ON FRIDAY SEPTEMBER 30**. You may use a calculator, a ruler, your notes, the book, and the internet. You may take it in as many blocks of time as you want. Submit your paper and electronic version before 11:30 am (eastern time zone:-).

The **answers to the questions must be typed in a Word file and emailed to me as an attachment**. Be sure to backup your test answers just in case (I suggest a thumb drive or other removable medium). You will need to capture screen images as a part of your answers which you may do without seeking permission since your test answers will not be in the public domain. Remember to explain your thoughts in your own words and use screen shots to support your answers. Screen shots without ***your*** words are worth very few points.

You may want to use some of the resources on this page <http://bioinformatics.org/sms/> but you may not need to. Just wanted to supply everyone with a common suite of tools.

DO NOT READ or DOWNLOAD ANY NEW PAPERS FOR THIS EXAM. RELY ONLY ON THE FIGURES PROVIDED IN THIS EXAM, YOUR EXPERIENCE, AND YOUR SKILLS.

-3 pts if you do not follow this direction.
Please do not write or type your name on any page other than this cover page.
Staple all your pages (INCLUDING THE TEST PAGES) together when finished with the exam.

Name (please print):

Write out the full pledge and sign (by typing a second time and signing paper version):

How long did this exam take you to complete (excluding typing)?

20 pts.

1) Access the GRAMENE Genome Browser at this web site (http://www.gramene.org/).

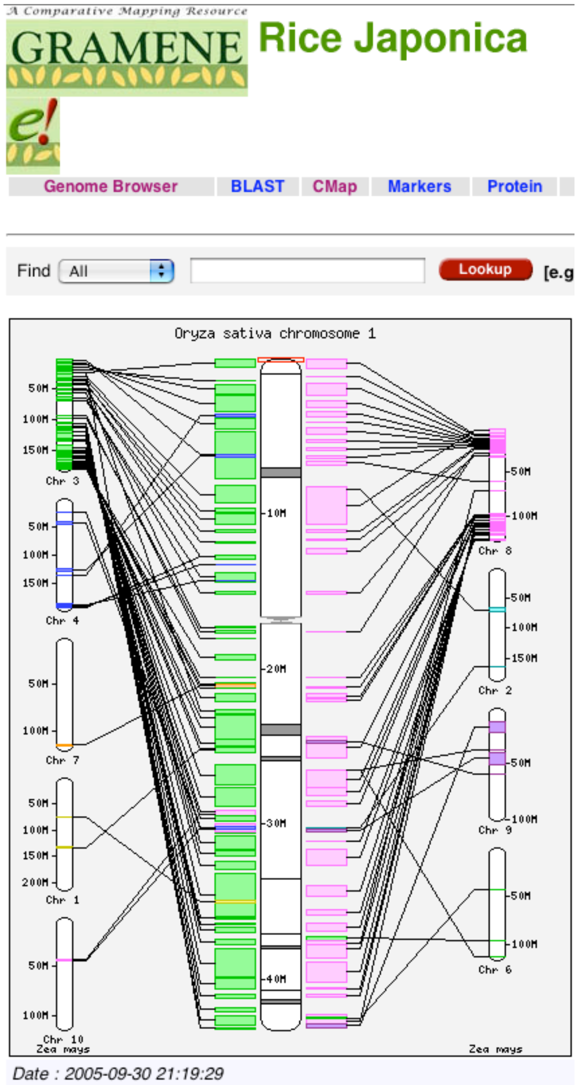
a. Consider the synteny of rice and maize and make an evaluation of the large-scale genomic changes that have occurred when comparing these two species. When you think you have a good sense of what has happened, use a screen shot of the most extreme example of what you observed to support your conclusions.

From this figure, you can see many maize chromosomes have DNA that map to a single rice chromosome. This is reminiscent of the puffer fish to human synteny comparison.

This is a good example to illustrate the probable duplication that took place in the maize genome.

Based on size alone, it is clear that maize must have undergone at least one duplication:

460 Mb genome = rice
3,000 Mb genome = maize

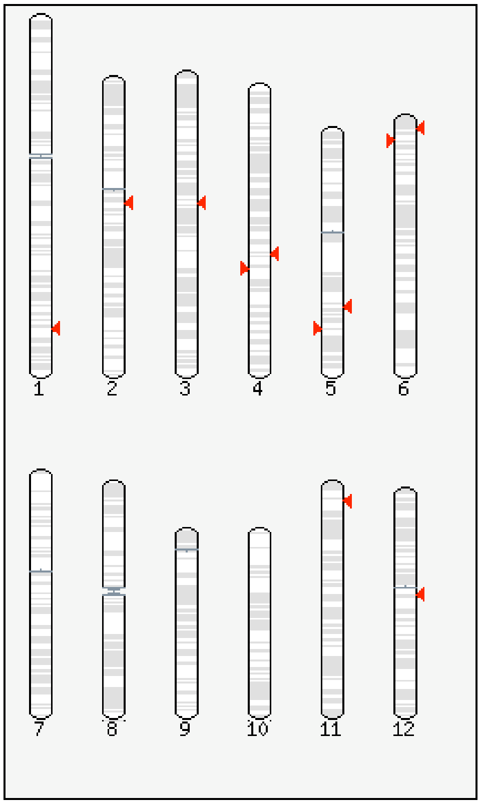


Date : 2005-09-30 21:19:29

b. Now search for Histone H3. Tell me where this gene is encoded in the Rice Japonica genome. Use a screen shot to support your findings.

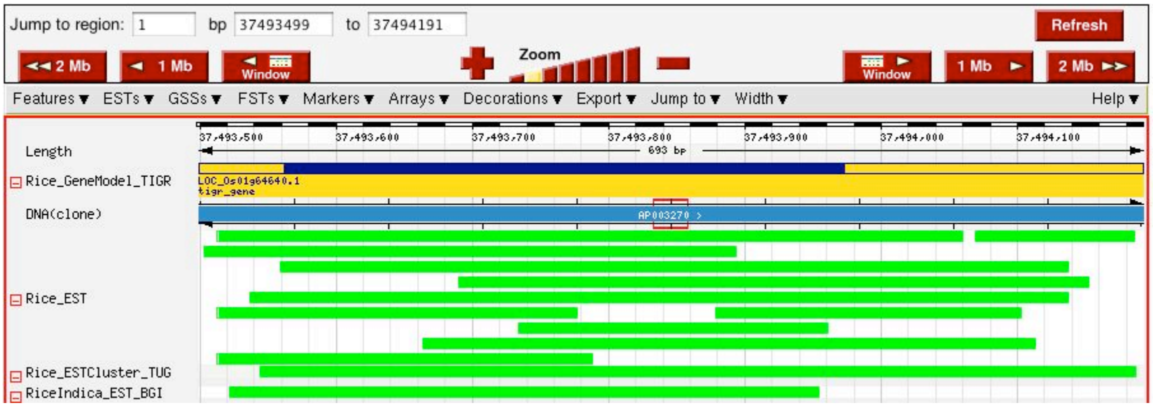
LOC_Os01g64640 37493499 - 37494192 bp (37.5 Mb) on chromosome 1
LOC_Os02g25940 15170251 - 15170925 bp (15.2 Mb) on chromosome 2
LOC_Os03g27310 15633410 - 15636041 bp (15.6 Mb) on chromosome 3

LOC_Os04g34240 20394571 - 20395338 bp (20.4 Mb) on chromosome 4
LOC_Os04g37780 22082063 - 22087295 bp (22.1 Mb) on chromosome 4
LOC_Os05g36280 21311103 - 21311835 bp (21.3 Mb) on chromosome 5
LOC_Os06g06460 3039364 - 3040035 bp (3.0 Mb) on chromosome 6
LOC_Os06g06500 3054213 - 3054623 bp (3.1 Mb) on chromosome 6
LOC_Os06g06510 3055172 - 3055582 bp (3.1 Mb) on chromosome 6
LOC_Os11g05730 2616346 - 2618061 bp (2.6 Mb) on chromosome 11
LOC_Os12g22650 12791053 - 12791463 bp (12.8 Mb) on chromosome 12
LOC_Os12g22680 12806609 - 12807019 bp (12.8 Mb) on chromosome 12



http://www.gramene.org/Oryza_sativa/domainview?domainentry=IPR000164

c. Use data to convince me whether histone h3 is highly transcribed or not.

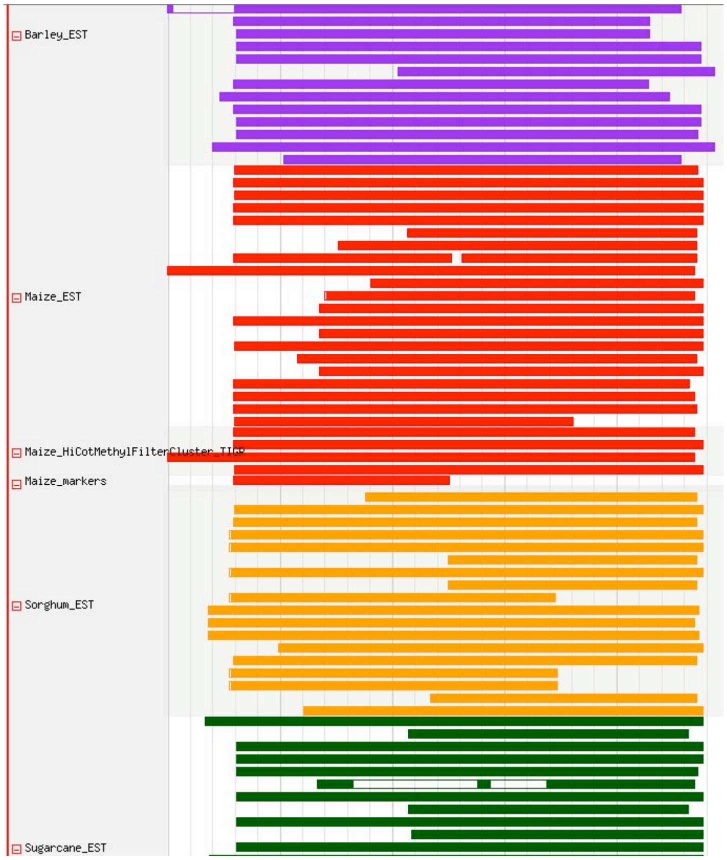


Jump to region: 1 bp 37493499 to 37494191

http://www.gramene.org/Oryza_sativa/contigview?highlight=LOC_Os01g64640&chr=1&vc_start=37493499&vc_end=37494191&x=54&y=14

Highly is a relative term, but based on the number of ESTs, it looks like Histone H3 is highly transcribed.

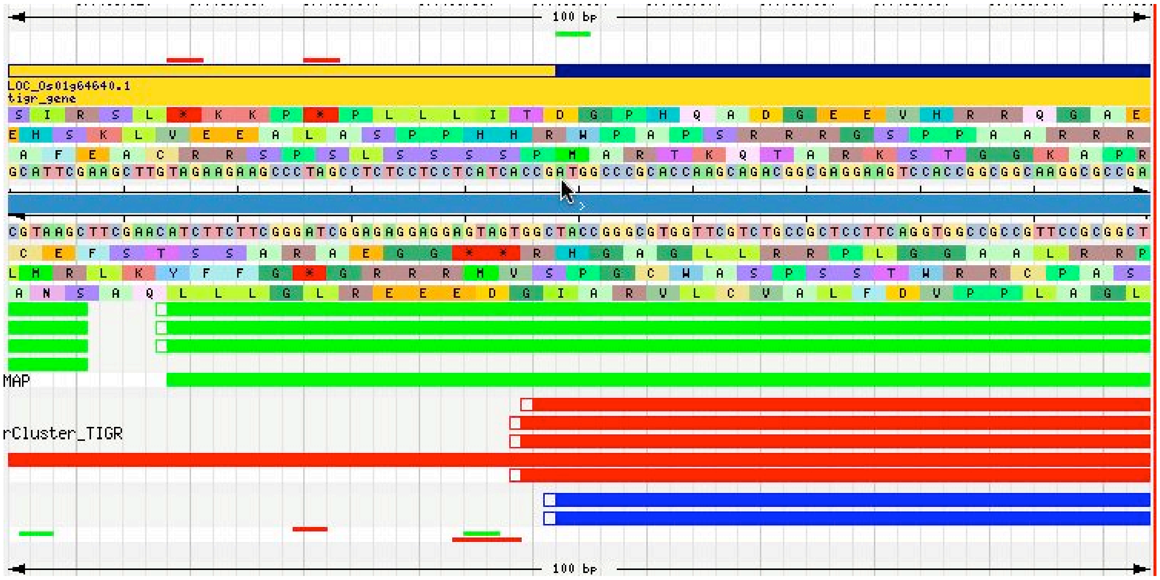
d. Is this gene similarly expressed in other grain plants? How do you know?
Yes, it appears fairly constant across the different species (see the number of ESTs).



e. What strand and which reading frame encodes the h3 protein you have selected? Explain how you found your answer and provide a screen shot of the DNA and protein sequences as shown in GRAMENE that validates your findings.

LOC_Os01g64640 37493499 - 37494192 bp (37.5 Mb) on chromosome 1

```
 CDS
>11667.m06477
ATGGCCCGCACCAAGCAGACGGCGAGGAAGTCCACCGGCGGCAAGGCGCCGAGGAAGCAG
CTGGCGACGAAGGCGGCGCGCAAGTCGGCCCCGGCCACCGGCGGCGTGAAGAAGCCCCAC
CGCTTCCGCCCCGGCACCGTCGCGCTCCGGGAGATCCGCAAGTACCAGAAGAGCACCGAG
CTGCTGATCCGCAAGCTGCCGTTCCAGCGCCTGGTGCGGGAGATCGCGCAGGACTTCAAG
ACCGACCTCCGCTTCCAGAGCTCCGCCGTCGCCGCGCTGCAGGAGGCGGCCGAGGCCTAC
CTCGTCGGGCTCTTCGAGGACACCAACCTCTGCGCCATCCACGCCAAGCGCGTCACCATC
ATGCCCAAGGACATCCAGCTCGCCCGCCGCATCCGTGGCGAGAGGGCCTAG

 Protein
>11667.m06477
MARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRFRPGTVALREIRKYQKSTE
LLIRKLPFQRLVREIAQDFKTDLRFQSSAVAALQEAAEAYLVGLFEDTNLCAIHAKRVTI

MPKDIQLARRIRGERA*
```


100 bp
LOC_Os01g64640.1
tigr_gene
S I R S L * K K P * P L L L I T D G P H Q A D G E E V H R R Q G A E
E H S K L V E E A L A S P P H H R W P A P S R R R G S P P A A R R R
A F E A C R R S P S L S S S S P M A R T K Q T A R K S T G G K A P R
GCATTCGAAGCTTGTAGAAGAAGCCCTAGCCTCTCCTCCTCATCACCGATGGCCCGCACCAAGCAGACGGCGAGGAAGTCCACCGGCGGCAAGGCGCCGA
CGTAAGCTTCGAACATCTTCTTCGGGATCGGAGAGGAGGAGTAGTGGCTACCGGGCGTGGTTCGTCTGCCGCTCCTTCAGGTGGCCGCCGTTCCGCGGCT
C E F S T S S A R A E G G * * R H G A G L L R R P L G G A A L R R P
L M R L K Y F F G * G R R R M V S P G C W A S P S S T W R R C P A S
A N S A Q L L L G L R E E E D G I A R V L C V A L F D V P P L A G L
MAP
rCluster_TIGR
100 bp

See the amino acid sequence MARTK…. just above the cursor arrow. The top strand has the ATG which means the coding strand is on the bottom strand. The reading frame closes to the DNA sequence is the appropriate reading from to start the CDS.

20 pts.

2) Start with this sequence and answer the following questions:

```
CCTAGTCTCCCTCCTCTTCGTCATGCTGCGCTACATGTACCGGCACAAGGGCACGTACCACACCAATGAGGCCAAGG
GCACGGAGTTTGCTGAGAGTGCAGATGCAGCCCTGCAGGGAGACCCTGCCCTCCAAGATGCTGGTGATAGCAGCAGA
AAGGAGTACTTTATTTGAGGGACAACAGACTTCACTTCCCTGAATGCCTCCCCCATCTCCATCAGGAAAAATACACC
CCATCGCCCAGCACCCCTGCTGATACCACCAGACAGAGAGAGAGAGCACTTGATTCTTCCCGAGATAGCCACCTGGA
AACACTAGGTGCCTGCCCAGGGAGGAACGGAGGAGGACTCGCGCTACAAGAG
```

a. What is this?

```
Homo sapiens glycophorin C (Gerbich blood group) (GYPC)
```

b. Provide me with the protein sequence.

```
MWSTRSPNSTAWPLSLEPDPGMSGWPDGRMETSTPTIMDIVVIAGVIAAVAIVLVSLLFVMLRYMYRHKGTYHTNEA
KGTEFAESADAALQGDPALQDAGDSSRKEYFI
```

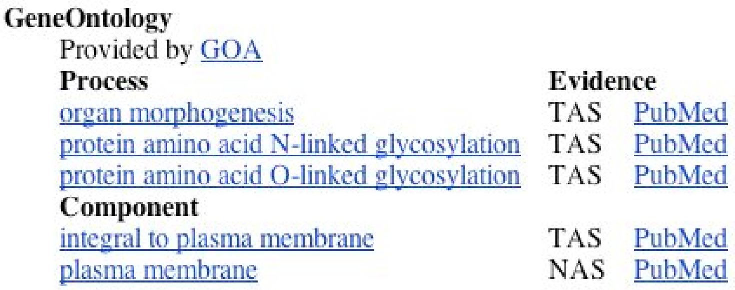
c. List the proteins functions and your source.

From Entrez Gene

GYPC glycophorin C (Gerbich blood group) [*Homo sapiens*]

GeneID: 2995 Locus tag: HGNC:4704; MIM: 110750

Glycophorin C (GYPC) is an integral membrane glycoprotein. It is a minor species carried by human erythrocytes, but plays an important role in regulating the mechanical stability of red cells.



GeneOntology

Provided by GOA

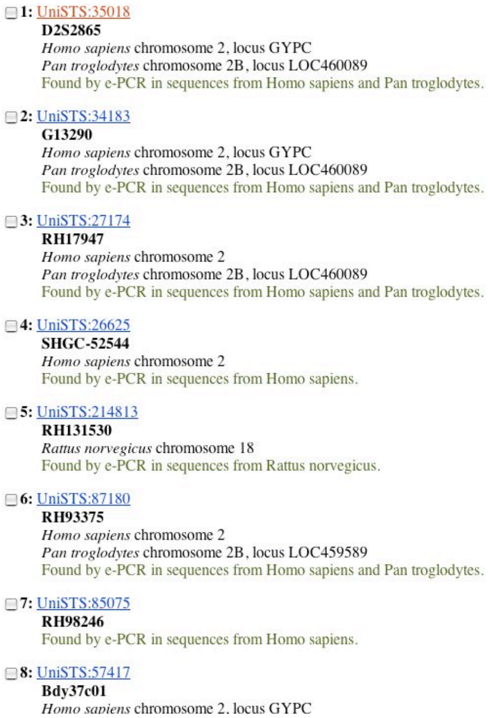
Process	**Evidence**	
organ morphogenesis	TAS	PubMed
protein amino acid N-linked glycosylation	TAS	PubMed
protein amino acid O-linked glycosylation	TAS	PubMed
Component		
integral to plasma membrane	TAS	PubMed
plasma membrane	NAS	PubMed

From OMIM:

It is a putative receptor for the merozoites of *Plasmodium falciparum* (Pasvol et al., 1984).

d. Are there any STS markers for this gene? Support your answer with data.

Yes, UniSTS found 7 for humans.



□ 1: UniSTS:35018
D2S2865
Homo sapiens chromosome 2, locus GYPC
Pan troglodytes chromosome 2B, locus LOC460089
Found by e-PCR in sequences from Homo sapiens and Pan troglodytes.

□ 2: UniSTS:34183
G13290
Homo sapiens chromosome 2, locus GYPC
Pan troglodytes chromosome 2B, locus LOC460089
Found by e-PCR in sequences from Homo sapiens and Pan troglodytes.

□ 3: UniSTS:27174
RH17947
Homo sapiens chromosome 2
Pan troglodytes chromosome 2B, locus LOC460089
Found by e-PCR in sequences from Homo sapiens and Pan troglodytes.

□ 4: UniSTS:26625
SHGC-52544
Homo sapiens chromosome 2
Found by e-PCR in sequences from Homo sapiens.

□ 5: UniSTS:214813
RH131530
Rattus norvegicus chromosome 18
Found by e-PCR in sequences from Rattus norvegicus.

□ 6: UniSTS:87180
RH93375
Homo sapiens chromosome 2
Pan troglodytes chromosome 2B, locus LOC459589
Found by e-PCR in sequences from Homo sapiens and Pan troglodytes.

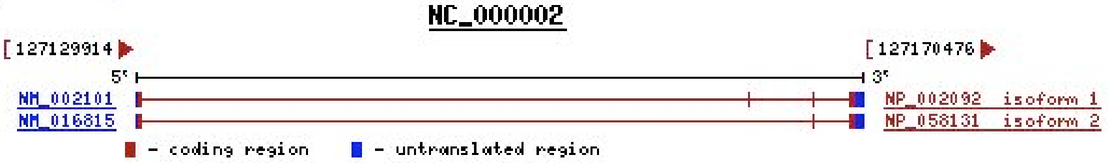
□ 7: UniSTS:85075
RH98246
Found by e-PCR in sequences from Homo sapiens.

□ 8: UniSTS:57417
Bdy37c01
Homo sapiens chromosome 2, locus GYPC

lodytes.

e. Are there any splice variants? Support your answer with data.

Yes, there are two.



f. Find the percent identity in the protein sequence of this gene with its chimp ortholog. Show data to support your findings.

Human NP_058131

MWSTRSPNSTAWPLSLEPDPGMASASTTMHTTTIAEPDPGMSGWPDGRMETSTPTIMDIV
VIAGVIAAVAIVLVSLLFVMLRYMYRHKGTYHTNEAKGTEFAESADAALQGDPALQDAGD
SSRKEYFI

MASASTTMHTTTIAEPDPGMSGWPDGRMETSTPTIMDIVVIAGVIAAVAIVLVSLLFVML
RYMYRHKGTYHTNEAKGTEFAESADAALQGDPALQDAGDSSRKEYFI

Pan ENSF00000008433

GLYCOPHORIN C PAS 2' GLYCOPROTEIN BETA GLPC GLYCOCONNECTIN SIALOGLYCOPROTEIN D GLYCOPHORIN D GPD

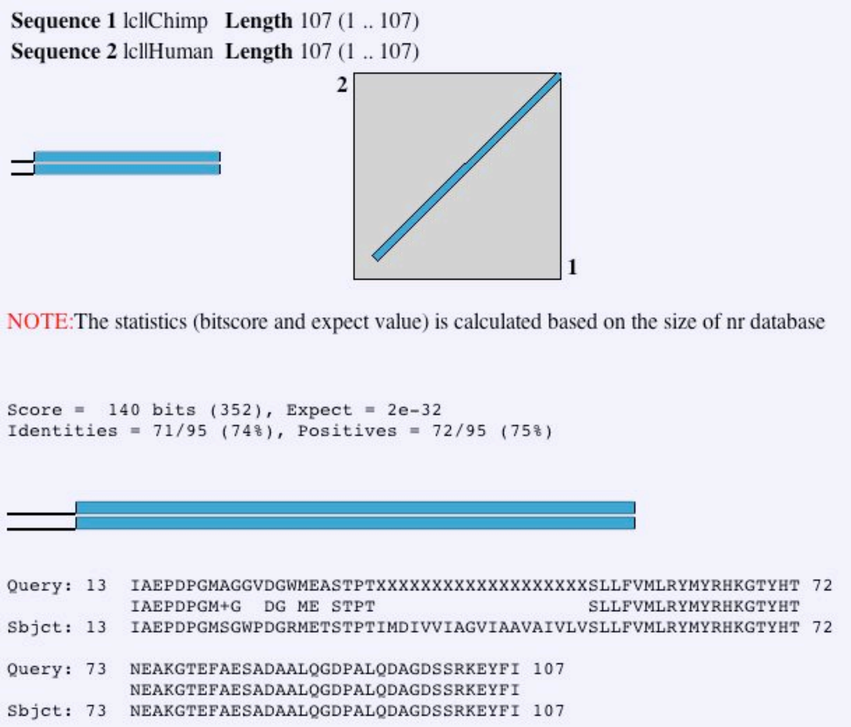
http://www.ensembl.org/Pan_troglodytes/geneview?gene=ENSPTRG00000012422

XPDPGMASASTTMHTTTIAEPDPGMAGGVDGWMEASTPTIIDIVIIAGVIAAVAIVLISL
LFVMLRYMYRHKGTYHTNEAKGTEFAESADAALQGDPALQDAGDSSRKEYFI

MASASTTMHTTTIAEPDPGMAGGVDGWMEASTPTIIDIVIIAGVIAAVAIVLISLLFVML
RYMYRHKGTYHTNEAKGTEFAESADAALQGDPALQDAGDSSRKEYFI

BLAST2

74% identical



Sequence 1 lcl|Chimp **Length** 107 (1 .. 107)
Sequence 2 lcl|Human **Length** 107 (1 .. 107)

NOTE:The statistics (bitscore and expect value) is calculated based on the size of nr database

```
Score =  140 bits (352), Expect = 2e-32
Identities = 71/95 (74%), Positives = 72/95 (75%)
```

```
Query: 13  IAEPDPGMAGGVDGWMEASTPTXXXXXXXXXXXXXXXXXXXSLLFVMLRYMYRHKGTYHT 72
            IAEPDPGM+G  DG ME STPT                   SLLFVMLRYMYRHKGTYHT
Sbjct: 13  IAEPDPGMSGWPDGRMETSTPTIMDIVVIAGVIAAVAIVLVSLLFVMLRYMYRHKGTYHT 72

Query: 73  NEAKGTEFAESADAALQGDPALQDAGDSSRKEYFI 107
            NEAKGTEFAESADAALQGDPALQDAGDSSRKEYFI
Sbjct: 73  NEAKGTEFAESADAALQGDPALQDAGDSSRKEYFI 107
```

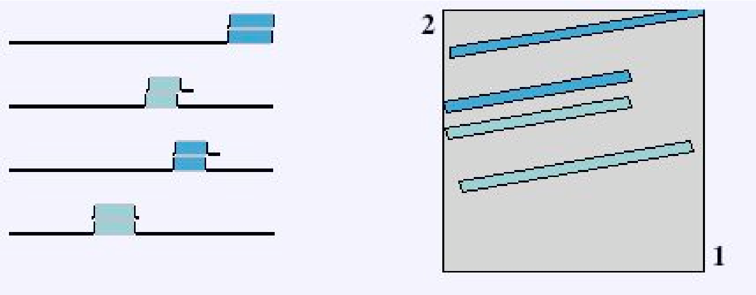
g. Tell me the percent identity for human and chimp proteins granulysin, protamine, and semenogelin. Use a table to show your results.

Using ENSEMBL databases

granulysin Identities = 117/126 (92%) (immunity)

protamine Identities = 27/47 (57%) (sperm protein)
semenogelin Identities = 54/97 (55%) (sperm protein)

BLAST2 alignment of human → and chimp semenogelin



h. Human and chimp have 29% of their orthologous proteins 100% conserved and the average number of amino acid changes is 2 per protein. Propose an hypothesis to explain these genome-wide numbers with the numbers you found in f. and g. above.

These 4 proteins appear to have evolved faster than the average protein. Two are related to sexual reproduction, one to immunity and the first one to malaria infection. There appears to a strong selection pressure for these to evolve faster than most proteins.

20 pts.

3) Open the attached file called worksheet.doc.

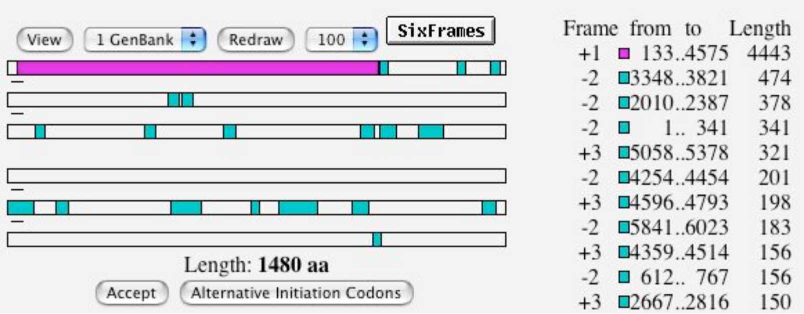
a. Find the largest open reading frame. In which frame is the largest ORF? How many amino acids? What is the predicted molecular weight? What conserved domains are there, if any? Show data for each of these answers.

First reading frame.

1480 amino acids

The protein weighs **168.14** kilodaltons

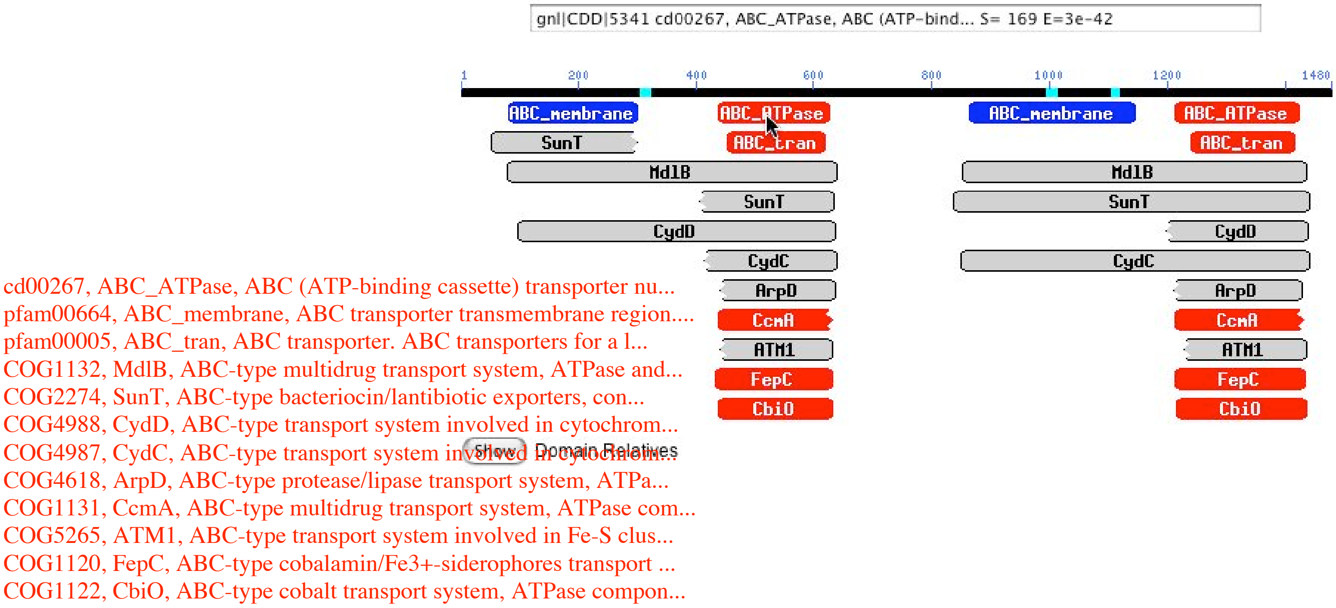
(The Sequence Manipulation Suite)



Length: **1480 aa**

Frame		from to	Length
+1	🟪	133..4575	4443
-2	🟦	3348..3821	474
-2	🟦	2010..2387	378
-2	🟦	1.. 341	341
+3	🟦	5058..5378	321
-2	🟦	4254..4454	201
+3	🟦	4596..4793	198
-2	🟦	5841..6023	183
+3	🟦	4359..4514	156
-2	🟦	612.. 767	156
+3	🟦	2667..2816	150

A list of conserved domains is shown below, in graphic and text formats.



cd00267, ABC_ATPase, ABC (ATP-binding cassette) transporter nu...
pfam00664, ABC_membrane, ABC transporter transmembrane region....
pfam00005, ABC_tran, ABC transporter. ABC transporters for a l...
COG1132, MdlB, ABC-type multidrug transport system, ATPase and...
COG2274, SunT, ABC-type bacteriocin/lantibiotic exporters, con...
COG4988, CydD, ABC-type transport system involved in cytochrom...
COG4987, CydC, ABC-type transport system involved in cytochrom...
COG4618, ArpD, ABC-type protease/lipase transport system, ATPa...
COG1131, CcmA, ABC-type multidrug transport system, ATPase com...
COG5265, ATM1, ABC-type transport system involved in Fe-S clus...
COG1120, FepC, ABC-type cobalamin/Fe3+-siderophores transport ...
COG1122, CbiO, ABC-type cobalt transport system, ATPase compon...

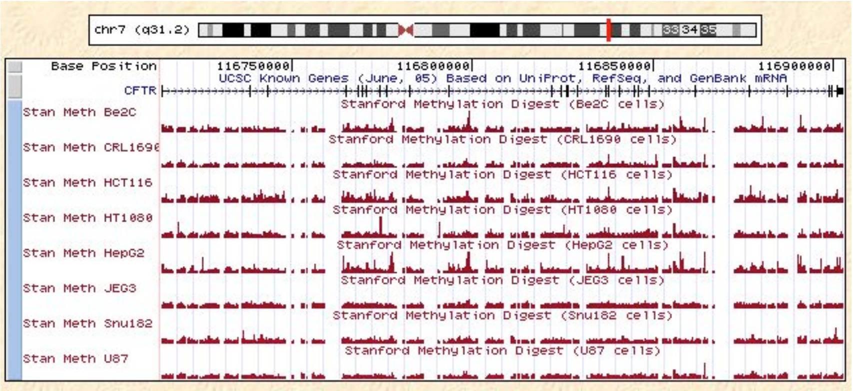
b. Is this the *wt* reference sequence or a mutant DNA sequence? Show data to support your answer.

When I performed a BLAST, I got this

`Identities = 6125/6129 (99%)`

Therefore, the query appears to be mutated.

c. Given this allele of the gene/cDNA, explain how it could lead to a disease in humans. Use a Genome Browser and consider the setting " Stanf Meth" to support your hypothesis with data.


chr7 (q31.2)
33 34 35
Base Position
116750000
116800000
116850000
116900000
UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA
CFTR
Stan Meth Be2C
Stanford Methylation Digest (Be2C cells)
Stan Meth CRL1690
Stanford Methylation Digest (CRL1690 cells)
Stan Meth HCT116
Stanford Methylation Digest (HCT116 cells)
Stan Meth HT1080
Stanford Methylation Digest (HT1080 cells)
Stan Meth HepG2
Stanford Methylation Digest (HepG2 cells)
Stan Meth JEG3
Stanford Methylation Digest (JEG3 cells)
Stan Meth Snu182
Stanford Methylation Digest (Snu182 cells)
Stan Meth U87
Stanford Methylation Digest (U87 cells)

Perhaps the 4 base changes alter the methylation status and thus its expression. Methylation tends to silence genes.

20 pts.

4) Find out some information about genome for the bacterium *Helicobacter pylori J99* to answer these questions:

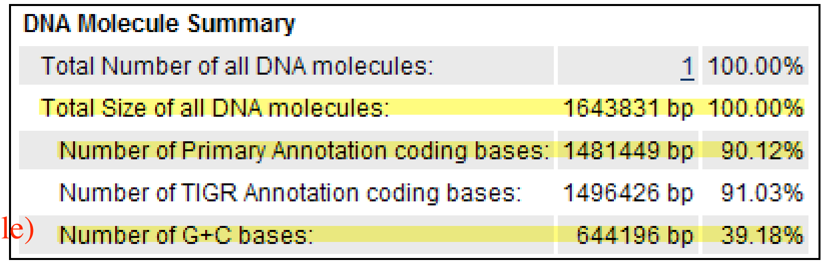
http://www.ncbi.nlm.nih.gov/genomes/framik.cgi?db=genome&gi=128

http://www.tigr.org/tigr-scripts/CMR2/GenomePage3.spl?database=ghp

a. How big is the genome?

1,643,831 or 1,667,866 base pairs accepted

b. What is the coding density?



DNA Molecule Summary

Total Number of all DNA molecules:	1	100.00%
Total Size of all DNA molecules:	1643831 bp	100.00%
Number of Primary Annotation coding bases:	1481449 bp	90.12%
Number of TIGR Annotation coding bases:	1496426 bp	91.03%
Number of G+C bases:	644196 bp	39.18%

91% (also found in paper freely available)

c. What is the overall %GC?

39% (not the same thing as frequency of G followed by C)

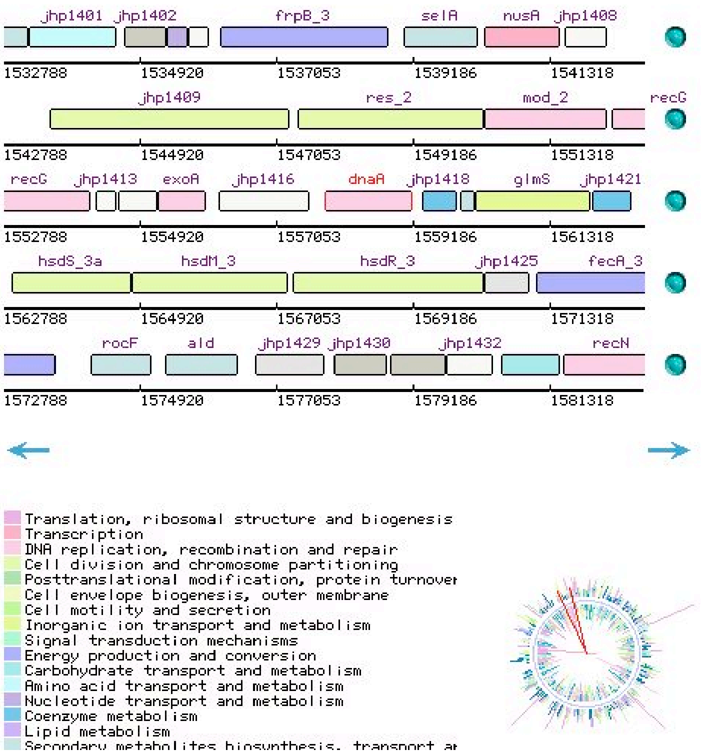
d. Find a gene with a significantly different %GC. Tell me the gene, its %GC, and provide me with the DNA sequence.

CGACCCTTGAAAGATTTGAACTTCCGTTTCCACCGTGAAAGGGTGGTATCCTTGGCCACTAGATGAAA
GGGTC tRNA-Glu from **Helicobacter pylori 26695**

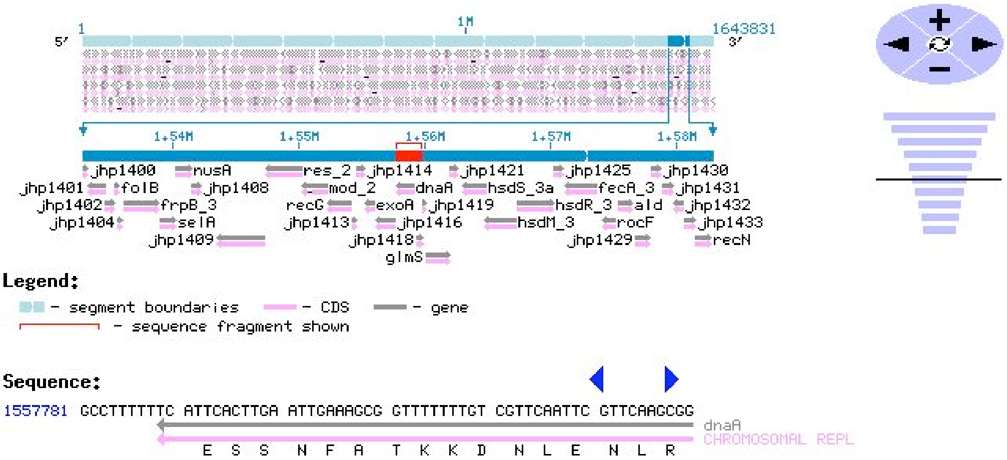
%GC = 49%

e. Identify the origin of replication and support your answer with two independent types of supporting data.

First, I searched for *DnaA* and found this image


jhp1401
jhp1402
frpB_3
selA
nusA
jhp1408
1532788
1534920
1537053
1539186
1541318
jhp1409
res_2
mod_2
recG
1542788
1544920
1547053
1549186
1551318
recG
jhp1413
exoA
jhp1416
dnaA
jhp1418
glmS
jhp1421
1552788
1554920
1557053
1559186
1561318
hsdS_3a
hsdM_3
hsdR_3
jhp1425
fecA_3
1562788
1564920
1567053
1569186
1571318
rocF
ald
jhp1429
jhp1430
jhp1432
recN
1572788
1574920
1577053
1579186
1581318
Translation, ribosomal structure and biogenesis
Transcription
DNA replication, recombination and repair
Cell division and chromosome partitioning
Posttranslational modification, protein turnover
Cell envelope biogenesis, outer membrane
Cell motility and secretion
Inorganic ion transport and metabolism
Signal transduction mechanisms
Energy production and conversion
Carbohydrate transport and metabolism
Amino acid transport and metabolism
Nucleotide transport and metabolism
Coenzyme metabolism
Lipid metabolism
Secondary metabolites biosynthesis, transport ar

chromosomal replication initiator protein (*dnaA*)



Legend:

▨ - segment boundaries ▬ - CDS ▬ - gene

┌──┐ - sequence fragment shown

Sequence:

```
1557781 GCCTTTTTTC ATTCACTTGA ATTGAAAGCG GTTTTTTTGT CGTTCAATTC GTTCAAGCGG
```

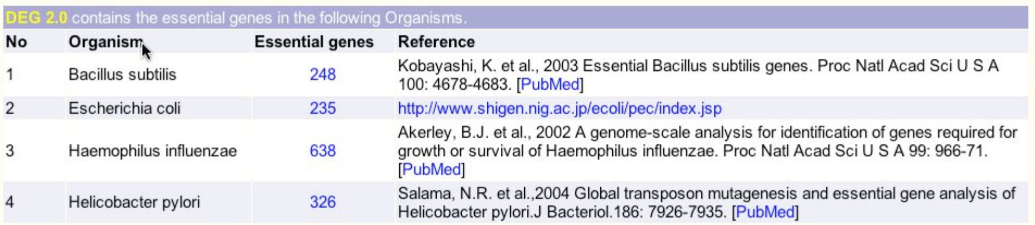
It is not universal, but many genes point their transcription away from DnaA.

f. What percentage of *H. pylori*'s genes are considered to be essential? Provide me with the numbers you used to calculate your answer and how you obtained these numbers.

DEG = 326

Total genes = 1587 (this number varied by source)

20% are essential.



DEG 2.0 contains the essential genes in the following Organisms.

No	Organism	Essential genes	Reference
1	Bacillus subtilis	248	Kobayashi, K. et al., 2003 Essential Bacillus subtilis genes. Proc Natl Acad Sci U S A 100: 4678-4683. [PubMed]
2	Escherichia coli	235	http://www.shigen.nig.ac.jp/ecoli/pec/index.jsp
3	Haemophilus influenzae	638	Akerley, B.J. et al., 2002 A genome-scale analysis for identification of genes required for growth or survival of Haemophilus influenzae. Proc Natl Acad Sci U S A 99: 966-71. [PubMed]
4	Helicobacter pylori	326	Salama, N.R. et al.,2004 Global transposon mutagenesis and essential gene analysis of Helicobacter pylori.J Bacteriol.186: 7926-7935. [PubMed]

20 pts.

5) I have just read an interesting paper about RNAi in mouse. See if you can make sense of these clues and data.

a. Find the DNA sequences for

NM_021476

Mus musculus cysteinyl leukotriene receptor 1 (Cysltr1), mRNA

```
>gi|31542448:394-1452 Mus musculus cysteinyl leukotriene receptor 1 (Cysltr1), mRNA
ATGTACCTCCAAGGCACCAAGCAGACATTCCTGGAGAACATGAATGGAACTGAAAATCTGACGACATCTC
TCATTAATAACACGTGTCATGACACAATTGATGAATTCCGAAATCAAGTATACTCCACTATGTATTCTGT
GATCTCTGTTGTGGGTTTCTTTGGCAATAGCTTTGTGCTCTATGTCCTCATAAAAACATACCATGAGAAA
TCAGCCTTCCAAGTATACATGATTAATCTAGCCATAGCAGATCTACTCTGTGTATGTACATTGCCTCTCC
GTGTGGTCTATTATGTTCACAAAGGCAAGTGGCTCTTTGGTGACTTTTTGTGCCGCCTCACCACCTATGC
CTTGTACGTTAACCTCTATTGTAGCATCTTCTTTATGACAGCCATGAGCTTTTTCCGGTGTGTTGCAATT
GTCTTTCCAGTCCAGAACATTAATTTGGTTACACAGAAAAAAGCCAGGTTCGTTTGCATTGGAATTTGGA
TTTTTGTGATTTTGACAAGTTCTCCCTTTTTAATGTACAAATCTTACCAAGATGAGAAAAACAATACTAA
ATGCTTTGAGCCTCCACAGAACAATCAAGCTAAAAAATACGTTTTGATCTTGCATTATGTGTCATTATTC
TTTGGTTTCATCATCCCTTTTGTCACCATAATTGTCTGTTACACAATGATCATTCTGACCTTACTAAAAA
ATACAATGAAGAAAAACATGCCAAGTCGTAGGAAGGCTATAGGGATGATCATAGTTGTGACAGCTGCCTT
TTTAGTGAGCTTCATGCCATATCATATTCAACGAACTATCCACCTTCACCTTTTACACAGTGAAACTAGA
```

```
CCCTGTGATTCTGTCCTTAGGATGCAGAAGTCAGTGGTCATAACCTTATCTCTAGCTGCATCAAATTGTT
GCTTTGATCCTCTGCTATATTTCTTTTCAGGTGGAAACTTTAGGAGAAGGCTATCTACATTTAGAAAGCA
TTCTTTGTCCAGTATGACTTATGTACCCAAGAAGAAAGCTTCCTTGCCAGAAAAAGGAGAAGAAATATGT
AACGAATAA
```

D530007L20

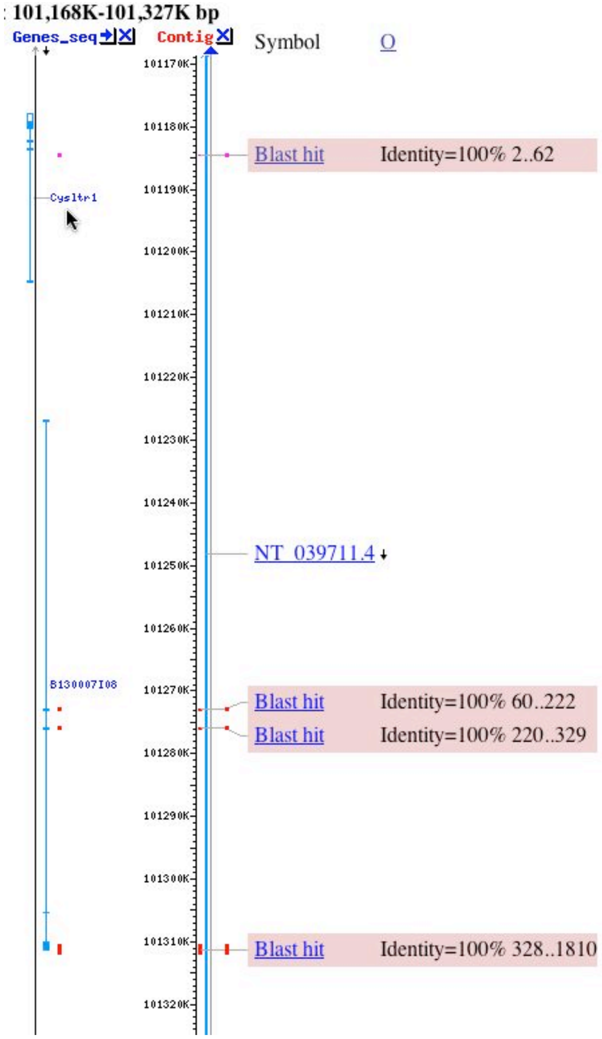
Mus musculus 13 days embryo stomach cDNA, RIKEN full-length enriched library, clone:D530007L20 product:hypothetical Rhodopsin-like GPCR superfamily/G-protein coupled receptors family. 1 profile containing protein, full insert sequence.

```
>gi|74210910:369-1007 Mus musculus 13 days embryo stomach cDNA, RIKEN full-length
enriched library, clone:D530007L20 product:hypothetical Rhodopsin-like GPCR
superfamily/G-protein coupled receptors family 1 profile containing protein, full
insert sequence
ATGAAAGACCTCCGCAGGTACCTGATGTTCAAATACATCCAGTATCTAGTCCTGCCTTCTCCGAATGTCA
TCATCATTCTGCTGGGATCTCTTGCCAGTCTCTATGTCATGCTGCTATTGACATTTCCTTCCGTTTCCAG
AAAATCCACTGCGGTGTTTATTGGTAGCATAGCTCAGGCAGACATCTTAGTTGTCTGCAGCTTGTTTTCT
GCTATCTCAGCATGCGTGATAAGGAGCGAGCCATCGTCAACTTCTTTTCAACTAGCCTTGAGGCAAAACT
TCCAAATTGCAAATATCCATGCCAGTTCCCTTTTACTCAGCTGTGTTACCCTTGAGGCTTTTCTGATTAC
TTTTCTTCCAGTAGAGACACGCCACATAAGGAATGTTAGATGTGCCAGAGTGGCTTCTAAAATCATCTGG
GCTGTTGTAATTACCGAGTGTTTCCTTTATCAGCTTGAATATGTCAAAGGCCTTAACATCTCCTATCTTG
GCATTCATAGGCAAATTCAGCTATTGATGAATTTCTTTTATGAGGCCACAGTGCTGCTGAAATTACTTAT
TTATCCCATTGGAGTTCTTCTAAGAATCTTCAACGTATATCTTTTTTATAAAATGTATTTCCGAGACCAT
TATAGTTAG
```

b. Find their chromosomal positions.

Used sequence accession number to
find map location
>gi|74210910|

What we see from this view is that NM_021476 and D530007L20 overlap on opposite strands of the same DNA. You can see that D530007L20 is cDNA due to the broken segments that match the BLAST results.



c. Look at Figure1.

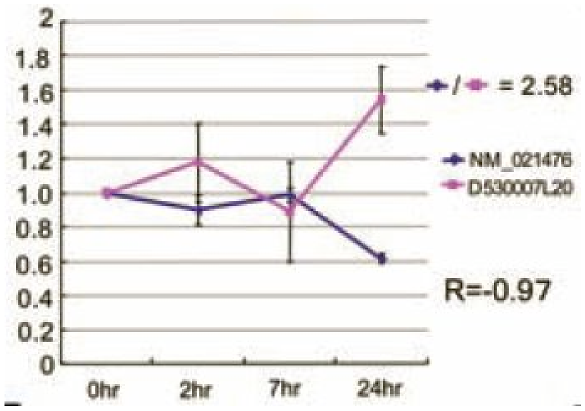
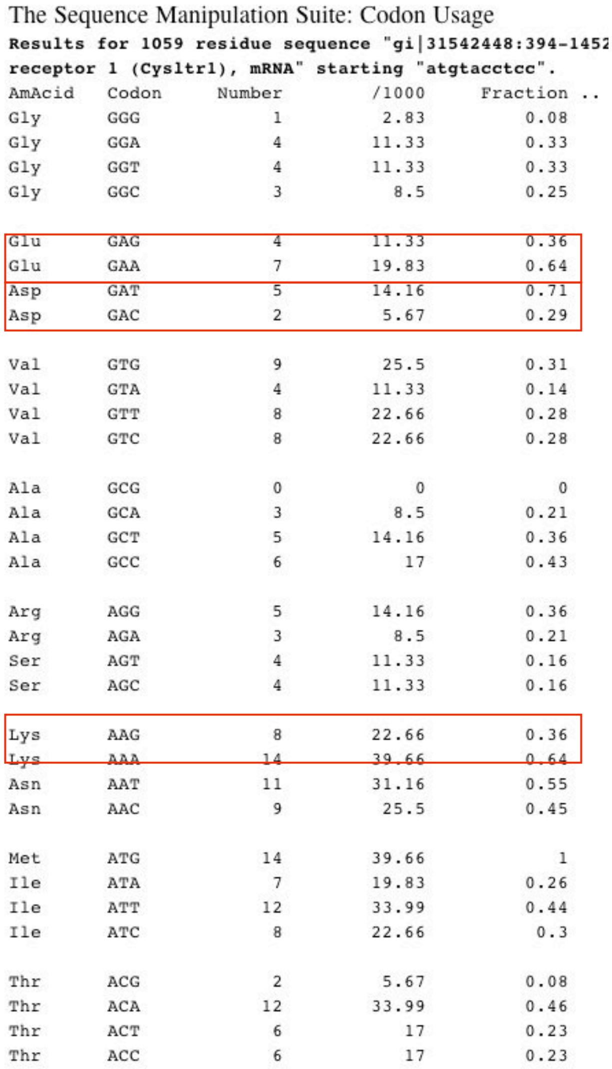


Figure 1. Time-course analysis of S/AS (sense/antisense) pairs. Expression of S/AS RNA pairs was verified by reverse transcription polymerase chain reaction over 24 hours after activation of macrophages with LPS. R, correlation coefficient. y axis, relative expression; blue/pink symbols ratio, actual expression levels at time 0 hours.

d. Use NCBI tools to determine if the sequences you retrieved in “a” are in fact sense antisense sequences. Show data to support your findings.

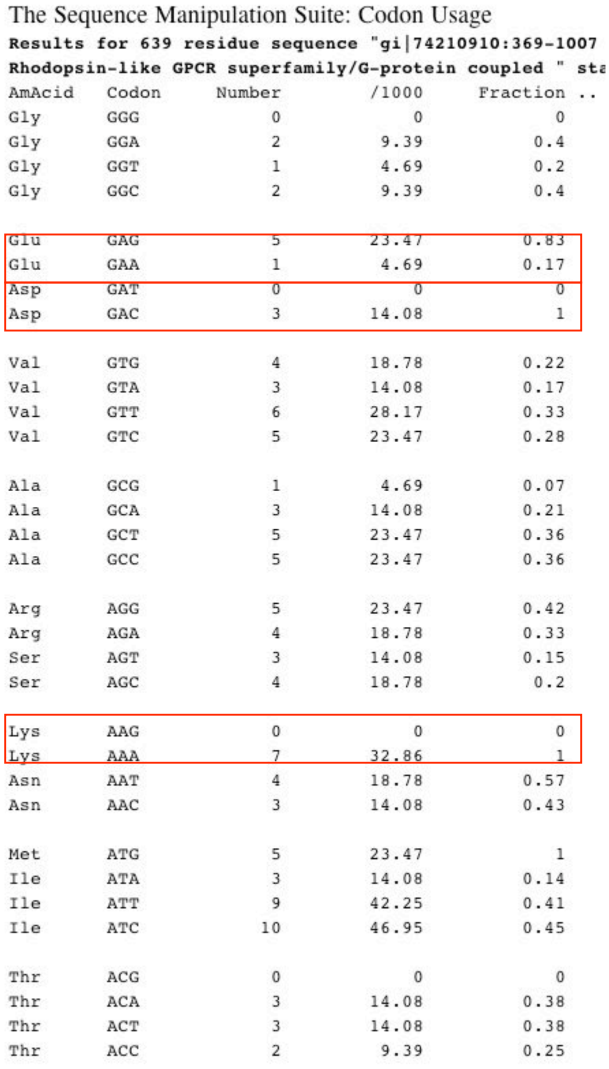
The figure above demonstrates they are antisense sequences in that the come from opposite strands, but the distributions do not over lap with the exons. So the mRNAs would not be sense and antisense mRNA. Pre-mRNA transcripts would be, however.

e. Do these two coding segments have the same codon bias? Support your answer with data and interpret the implication of your findings.



The Sequence Manipulation Suite: Codon Usage

```
Results for 1059 residue sequence "gi|31542448:394-1452
receptor 1 (Cysltr1), mRNA" starting "atgtacctcc".
AmAcid    Codon     Number        /1000     Fraction ..
Gly       GGG            1         2.83        0.08
Gly       GGA            4        11.33        0.33
Gly       GGT            4        11.33        0.33
Gly       GGC            3          8.5        0.25

Glu       GAG            4        11.33        0.36
Glu       GAA            7        19.83        0.64
Asp       GAT            5        14.16        0.71
Asp       GAC            2         5.67        0.29

Val       GTG            9         25.5        0.31
Val       GTA            4        11.33        0.14
Val       GTT            8        22.66        0.28
Val       GTC            8        22.66        0.28

Ala       GCG            0            0           0
Ala       GCA            3          8.5        0.21
Ala       GCT            5        14.16        0.36
Ala       GCC            6           17        0.43

Arg       AGG            5        14.16        0.36
Arg       AGA            3          8.5        0.21
Ser       AGT            4        11.33        0.16
Ser       AGC            4        11.33        0.16

Lys       AAG            8        22.66        0.36
Lys       AAA           14        39.66        0.64
Asn       AAT           11        31.16        0.55
Asn       AAC            9         25.5        0.45

Met       ATG           14        39.66           1
Ile       ATA            7        19.83        0.26
Ile       ATT           12        33.99        0.44
Ile       ATC            8        22.66         0.3

Thr       ACG            2         5.67        0.08
Thr       ACA           12        33.99        0.46
Thr       ACT            6           17        0.23
Thr       ACC            6           17        0.23
```

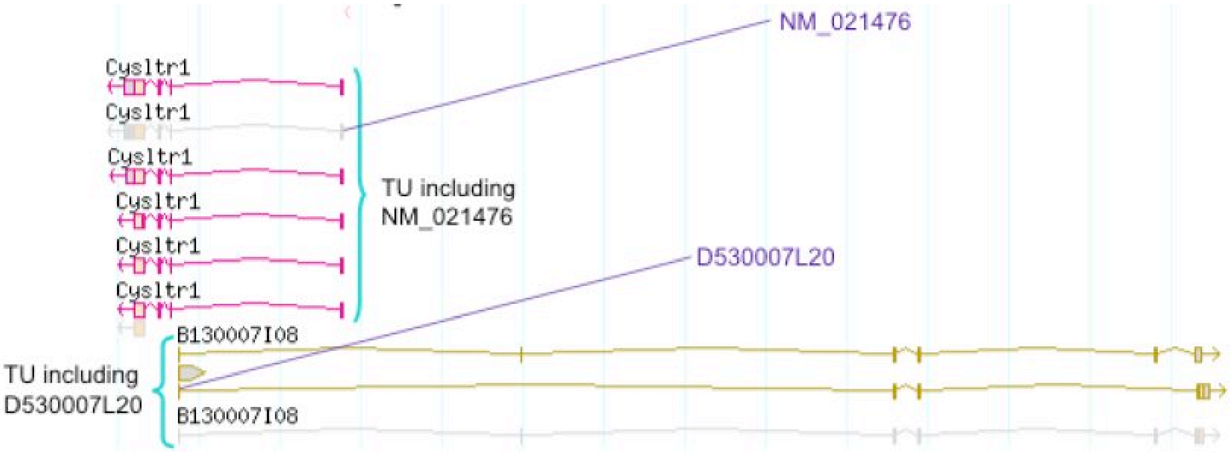


The Sequence Manipulation Suite: Codon Usage

```
Results for 639 residue sequence "gi|74210910:369-1007
Rhodopsin-like GPCR superfamily/G-protein coupled " sta
AmAcid    Codon     Number        /1000     Fraction ..
Gly       GGG            0            0           0
Gly       GGA            2         9.39         0.4
Gly       GGT            1         4.69         0.2
Gly       GGC            2         9.39         0.4

Glu       GAG            5        23.47        0.83
Glu       GAA            1         4.69        0.17
Asp       GAT            0            0           0
Asp       GAC            3        14.08           1

Val       GTG            4        18.78        0.22
Val       GTA            3        14.08        0.17
Val       GTT            6        28.17        0.33
Val       GTC            5        23.47        0.28

Ala       GCG            1         4.69        0.07
Ala       GCA            3        14.08        0.21
Ala       GCT            5        23.47        0.36
Ala       GCC            5        23.47        0.36

Arg       AGG            5        23.47        0.42
Arg       AGA            4        18.78        0.33
Ser       AGT            3        14.08        0.15
Ser       AGC            4        18.78         0.2

Lys       AAG            0            0           0
Lys       AAA            7        32.86           1
Asn       AAT            4        18.78        0.57
Asn       AAC            3        14.08        0.43

Met       ATG            5        23.47           1
Ile       ATA            3        14.08        0.14
Ile       ATT            9        42.25        0.41
Ile       ATC           10        46.95        0.45

Thr       ACG            0            0           0
Thr       ACA            3        14.08        0.38
Thr       ACT            3        14.08        0.38
Thr       ACC            2         9.39        0.25
```

Based on the 3 boxed areas, it appears these two genes do not share identical codon bias.

f. Now look at the file called “SOM Fig5.pdf” and try to make sense of all the data you have collected so far. Summarize your conclusions based on all the data you have, both from your own research and from this publication.



The key points are these:

1) When NM_021476 is induced, D530007L20 is repressed and vice versa.
2) These to genes overlap on opposite strands.
3) There exons do not appear to overlap at all.
4) They have different codon bias.
5) The question indicated it was related to RNAi.

The implication is that two genes may act as RNAi for each other. D530007L20 is only a hypothetical protein, but its mRNA is real. When NM_021476 is repressed, D530007L20 is induced. Perhaps D530007L20 acts as RNAi to completely silence NM_021476 residual mRNA. However, we have a problem that RNAi appears to work on mRNA but these two “genes” do not have codons in common. Therefore, it appears that RNAi may work inside the nucleus or perhaps non-spliced RNA is able to reach the cytoplasm and serve as RNAi raw material.