Fall 2003 Genomics Exam #1 Answer Key
Genomic Medicine and Sequencing Tools

There is no time limit on this test, though I have tried to design one that you should be able to complete within 4 hours, except for typing and web searches. There are three pages for this test, including this cover sheet. You are not allowed discuss the test with anyone until all exams are turned in at 11:30 am on Wednesday October 1. **EXAMS ARE DUE AT CLASS TIME ON WEDNESDAY OCTOBER 1**. You may use a calculator, a ruler, your notes, the book and the internet. I have to say, this is a challenging test, so do NOT put it off too long. You may take it in as many blocks of time as you need to.

The **answers to the questions must be typed on a separate sheet of paper** unless the question specifically says to write the answer in the space provided. If you do not write your answers in the appropriate location, I may not find them. 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. If you are asked to print out any pages, you do not have to print in color, though it is permitted.

-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:
On my honor I have neither given nor received unauthorized information regarding this work, I have followed and will continue to observe all regulations regarding it, and I am unaware of any violation of the Honor Code by others.

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

1) Start with this partial sequence:

```
atggcaaagg tcacaaactg ggataagttt gtgctgcttt tgtggaagaa
ctggaccctc caatggaacc acaagtggca gatggttata gagctggtgc
tgccagcgat attctccctg ctcctcgttc tagtccgcac cttggtggat
acggagcaaa aaggagtccg gtattataat gagcagaact taacagacct
caatctgctg cagaagaatg gcggcttttc aaaatttgag ttcatcctgt
gctactcgcc cgtcaatccc gtgttgaaga aactggtaga agaggcgtgg
cagagccttg gtaagaacaa aatctgtgag tcggaaaatg ccacccaact
ggagttggat
```

a. What is the species and gene name for the full sequence?

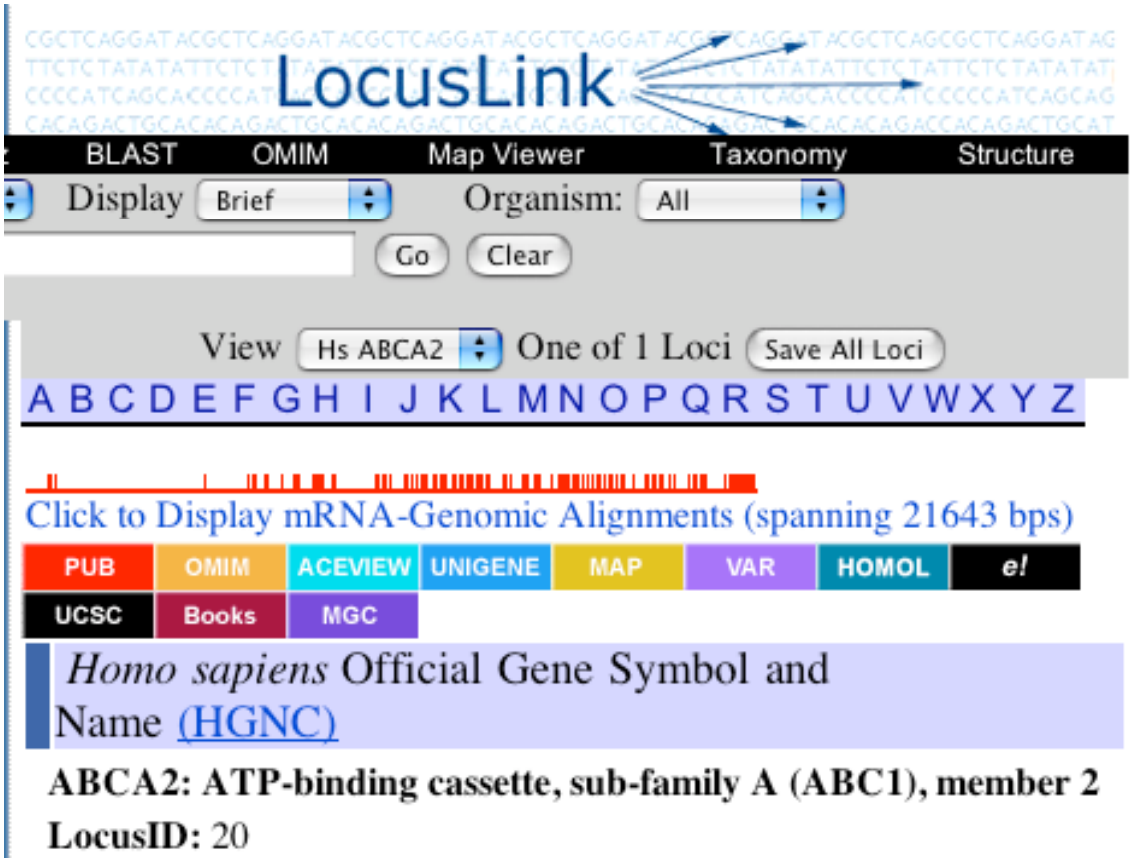
Drosophila melanogaster CG1718.

Sequences producing significant alignments:		Score (bits)	E Value
gi|24643673|ref|NM_134601.1|	Drosophila melanogaster CG1718...	714	0.0
gi|22833171|gb|AE003568.3|	Drosophila melanogaster chromoso...	422	e-115
gi|14578130|gb|AC010848.13|AC010848	Drosophila melanogaster...	422	e-115

b. Find the human ortholog. (no answer needed for letter b)

c. What is the name of the encoded human protein?

ABC1



d. Is the human gene/protein associated with any human diseases?
Yes, Tangier disease.

ALLELIC VARIANTS
(selected examples)

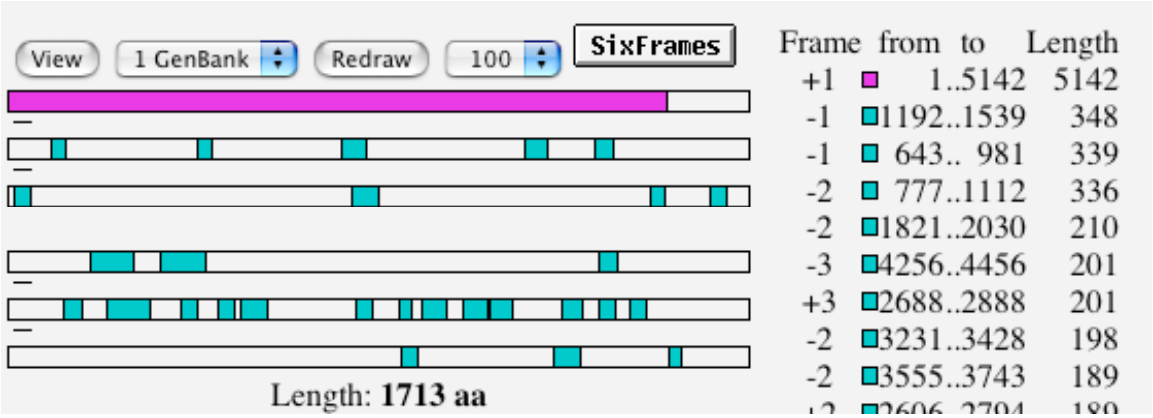
.0001 TANGIER DISEASE [ABCA1, CYS1417ARG]

In the proband with Tangier disease (205400) in a Dutch family, Broc
mutation on 1 chromosome in the proband was a T-to-C transition pre
mutation was a G-to-C transversion in the splice donor site of exon 2

e. How long is the complete mRNA (document your answer with an accession number).

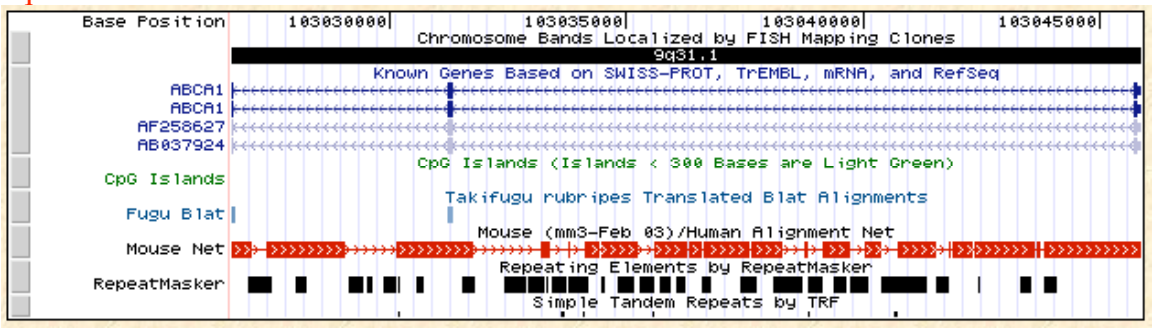
```
LOCUS       NM_134601                5142 bp    mRNA
DEFINITION  Drosophila melanogaster CG1718-PA [Drosophila
melanogaster](CG1718) mRNA, complete cds.
```

f. How many amino acids are in the largest ORF of the human sequence. Document your answer with data.
1713 amino acids.



g. Where is the human gene located?

9q31.1



h. Are there any STS markers for this gene? Document your answer.

yes.

RH81050

UniSTS:87958

Primer Information

Forward primer:	**CTGTCACAGCTTTATTTTGTGACTC**
Reverse primer:	**CTTTGTTCATCATTGGCCCT**
PCR product size:	245 (bp), Homo sapiens

Homo sapiens

Name:	RH81050
Also known as:	sts-N63586

Cross References

LocusLink	LocusID:	19
	Symbol:	ABCA1
	Description:	ATP-binding cassette, sub-family A (ABC1), member 1
	Position:	9q31.1
UniGene	Hs.147259	ATP-binding cassette, sub-family A (ABC1), member 1
SNP	rs1137170	Summary
RH details	RH81050	Genebridge4

i. Is the human protein post-translationally modified? Explain your answer and support it

with data.
It is heavily glycosylated. See this site:
http://www.mips.biochem.mpg.de/cgi-bin/proj/human/pedant/wwwdbp.pl?Db=human&Name=pros&String=48923&Title=human+PROSITE+patterns&ContigId=0&Cmd=Fetch

or this site
http://www.ncbi.nlm.nih.gov:80/entrez/viewer.fcgi?cmd=Retrieve&db=protein&list_uids=13123945&dopt=GenPept&term=&qty=1

j. Describe the human protein's cellular roles in Gene Ontology terminology.
Biological Process = cholesterol metabolism
Molecular Funciton = cholesterol transport
Cellular component = membrane protein (PM?)

Gene Product: Abca1

Full Name: ATP-binding cassette, sub-family A (ABC1), member 1
Synonyms: None
Data Source: MGI

Associated To Terms:

Term Name		Association Evidence
GO:0008203 : cholesterol metabolism	Tree View	Inferred from Direct Assay
GO:0030301 : cholesterol transport	Tree View	Inferred from Direct Assay
GO:0016021 : integral to membrane	Tree View	Traceable Author Statement

k. Is the human protein an integral membrane protein? Support your answer with evidence showing how many times it spans the membrane or that it does not span the membrane.
Yes, see this site and GO annotations. Spans about 13 times.
http://www.ncbi.nlm.nih.gov:80/entrez/viewer.fcgi?cmd=Retrieve&db=protein&list_uids=13123945&dopt=GenPept&term=&qty=1
You could also do a Kyte Doolittle, but this is only a predictor.

l. Document one SNP that produces a disease-causing allele.
In the proband with Tangier disease (205400) in a Dutch family, Brooks-Wilson et al. (1999) found compound heterozygosity for mutations in the ABC1 gene. The mutation on 1 chromosome in the proband was a T-to-C transition predicted to result in a cys1417-to-arg substitution. The mutation was located in exon 30. The other mutation was a G-to-C transversion in the splice donor site of exon 24, predicted to cause alternative splicing, deleting a significant part of the transcript.

m. Are there any non-coding mutations that can cause this disease, or are all documented cases caused by changes in the amino acid sequence? Document your answer.

Yes:

Zwarts et al. (2002) identified several SNPs in noncoding regions of ABCA1 that may be important for the appropriate regulation of ABCA1 expression (i.e., in the promoter, intron 1, and the 5-prime untranslated region), and examined the phenotypic effects of these SNPs in 804 Dutch men with proven coronary artery disease. They presented data suggesting that common variation in noncoding regions of ABCA1 may significantly alter the severity of atherosclerosis, without necessarily influencing plasma lipid levels.

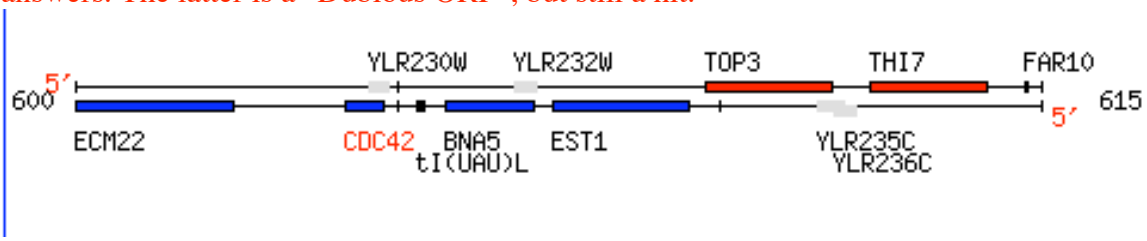
2. Here is a tricky sequence for you to identify.

```
ATGACAAGGGTCTCAATCGATCGTAATCTTCTTGACCGGCCGTATCAAACAAACCTAAC
GTATATGGTTCATCACCAATCATCACAGTCACCGCATAGTTATCGAACACTGTTGGAAC
ATAGTCGGCTGGAAATTGATTCGTTGTATAGGAGATTAGAAGGCACGTTTTCCCAACAG
CACCATCACCGACAACAACACACTTTAGCGTTTGCATTTTGTGGAAGAGCTAATACGTT
TATTTCTTGTTTTATAAGTTTTGCTTCTCTAATTCGTTTGTTAACCTATCTCCTCAGGA
AAATAGAATAG
```

Your task is to tell me why this is a tricky sequence. Use any online tools you can to document why you should not get a single answer. What I want you to do is tell me the two correct answers. Document how you came to your conclusions.

Two acceptable answers. One was that there are two accession numbers that contain this sequence; the first is a cosmid that contains this as one of its genes. The second hit is just this coding sequence.

However, the really tricky part is that there are two ORFs at this locus, one on each strand. CDC42 and YLR230W are both at this site and both are correct but different answers. The latter is a "Dubious ORF", but still a hit.



3) Find the structure file for "Rho transcription terminator"

a. Give me the most recent publication citation for this structure.

Cell. 2003 Jul 11;114(1):135-46.

Structure of the Rho transcription terminator: mechanism of mRNA recognition and helicase loading.

Skordalakes E, Berger JM.

b. What analogy do the authors use to explain this protein's mode of action?

Lock washer

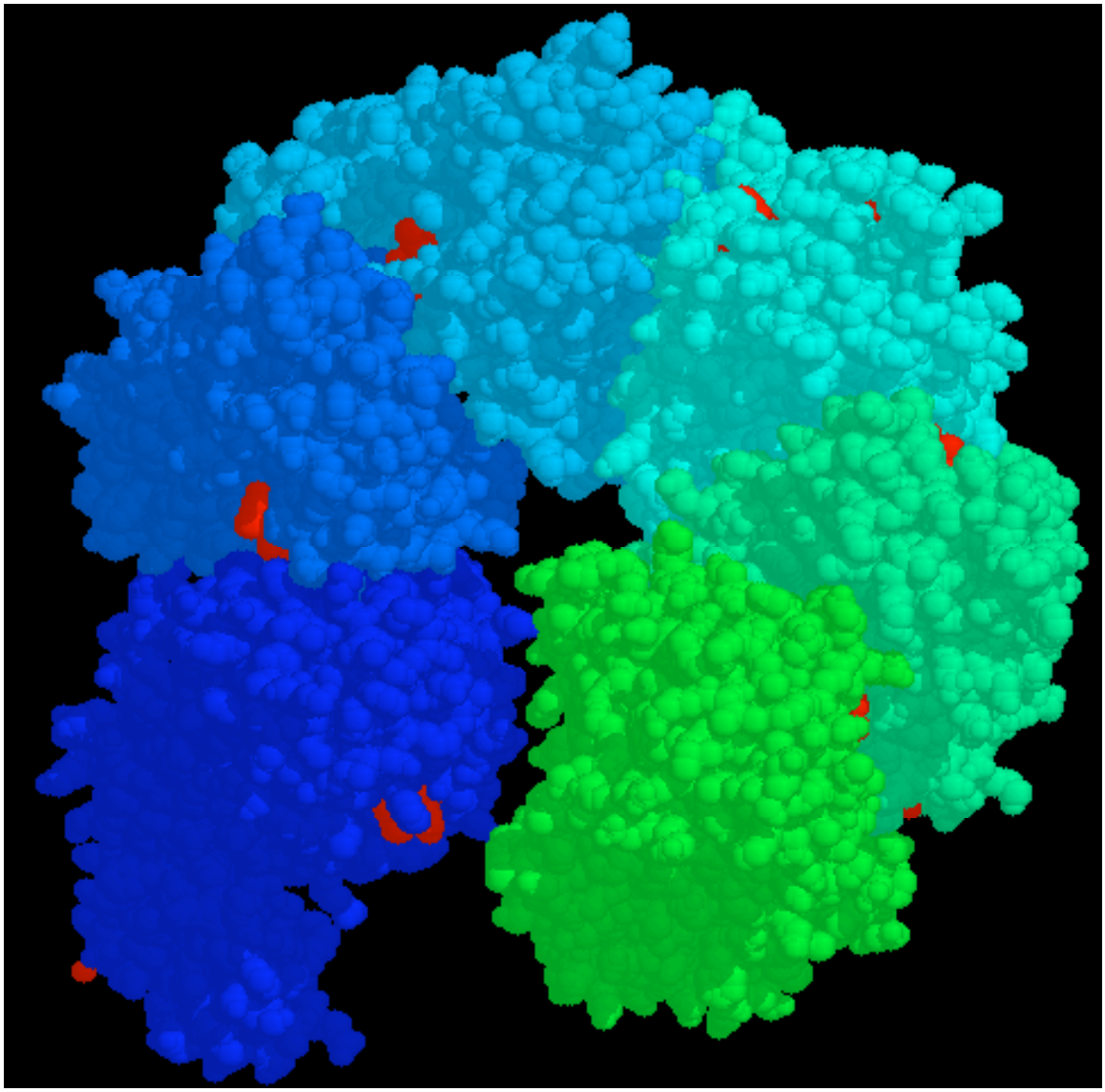
http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=12859904&dopt=Abstract

c. What is the PDB ID number?

PDB: 1PVO or 1PV4

d. How many protein subunits are present? Tell me both the number of subunits and how many different protein sequences are present in your structure.
6 protein subunits but it is a homohexamer so there is only one protein sequence.

e. Show me an image of this structure that matches the analogy and shows the number of subunits.



4) Go to the Human Mouse Homology map and focus on Human Chromosome 7.
a. Find a very small region on the human chromosome that has its ortholog on a small piece of a mouse chromosome that is not syntenic with other mouse genes shown on this image. Show me an image of which gene you found.
As with many of these questions, there are many correct answers. Here are some:
CRIP1

7q11.23	FGL2	5	b	Fgl2
7q11.23	PTPN12	5	b	Ptpn12
7q21	KIAA0705	5	b	Acvrip1-pending
7q11.23	*CRIP1* *	12	b	Crip
7q11.23	*CLDN3* *	5	b	Cldn3
7q11.23	*YWHAG* *	5	b	Ywhag
7q21	GNAI1	5	b	Gnai1
7q11.2	CD36	5	b	Cd36

b. Find a gene in the human map that appears to be mislocated. Name that gene and show a picture of the gene you have chosen. Explain why you think this human gene is mislocated in this map. Verify your choice by further evidence using another source of information.

MAP2K2 is mislocated in this list of genes and this is validated by many other databases.

	7p22.2	RBAK	5	b	Rbak-pending
	7q32	MAP2K2	10	b	Map2k2
sts	7p22	*NUDT1* *		b	Mth1
sts	7p22	*MAD1L1* *	5	b	Mad1l1

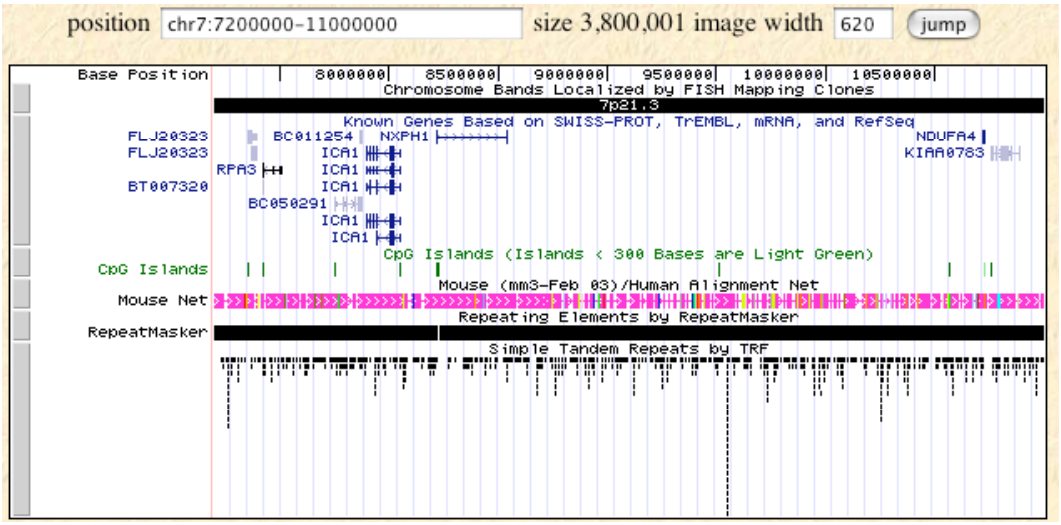
c. Zoom in on small region of the human chromosome 7 and focus on the human gene ICA1.

What is the location of the mouse orthologs for ICA1's nearest neighbors? What is the problem depicted in this image?

sts	7p22	RPA3	5	b	C330026P08Rik	sts
sts	7p22	ICA1	6	b	Ica1	2.9
sts	7p21.3	NDUFA4		b	Ndufa4	

Some of the genes have no location and yet if you click on them, there locations are known.

d. Use a database to find the chromosomal location of ICA1, the GC islands in this region and the two flanking neighbors of ICA1. Zoom in as tightly as you can and take a screen shot.



e. Which human gene neighbor is closer to ICA1?
It depends on the database you used. But the closest is BC050291, then NXPH1. If you used the homology map, then it would be PRA3.

f. Describe the location of the GC islands relative to the genes you tracked down?
There are GC islands on both ends but none in the middle. This is consistent with GC islands in promoter regions, but a bit of a surprise given genes tend to have more GC content (might have expected more islands).

g. Does the human mouse homology map show all the genes in this area of the human genome?
No. See e. above.

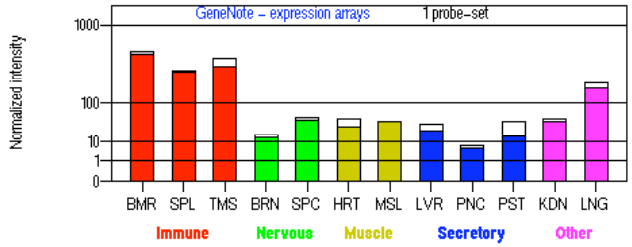
h. Given what the authors Eichler and Sankoff (the PDF file I emailed to you a couple weeks ago) said about chromosomal break points, what is surprising about this region of the human genome?
We might have expected more breaks given the repeats in this area. However, it is not easy to see why there would be a break between PRA3 and ICA1. Maybe this explains the partial nature of BC050291.
This question was too obscure so many answers were acceptable.

5) a. Find the tissues that express ARHGDIB and DAT1 (DAT1 is also called SLC6A3). Make sure you are working with the two genes located on chromosome 12.

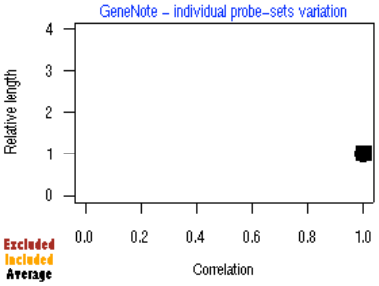
b. Which gene appears to be a housekeeping gene based on expression patterns?
ARHGDIB

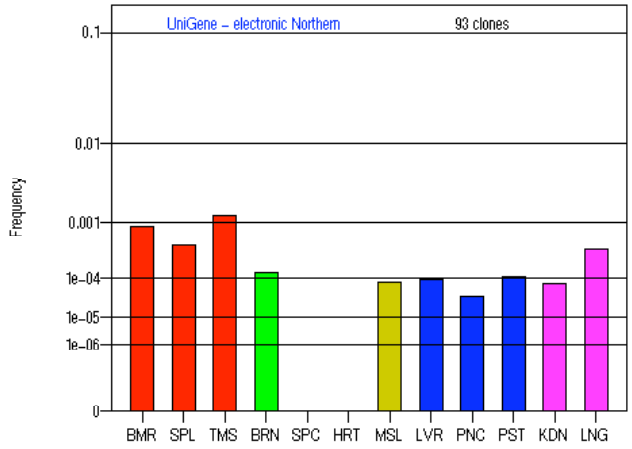
ARHGDIB expression in normal human tissues according to GeneNote/GeneAnnot

Affymetrix probe-set	array	sensitivity	specificity	# of associated genes
1984_s_at	A	1	1	1



White rectangles above bars show min-max range for duplicates

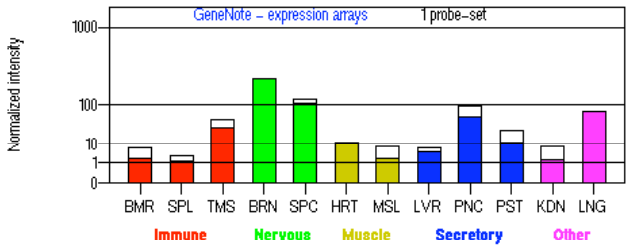




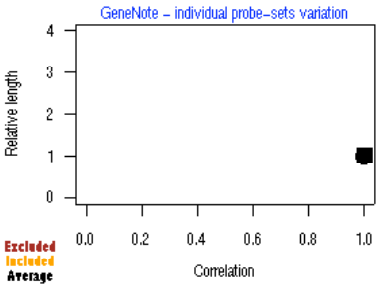
Tissue		Clones per gene	Total clones
BMR	Bone marrow	17	19,606
SPL	Spleen	4	9,440
TMS	Thymus	5	3,941
BRN	Brain	20	160,338
SPC	Spinal cord	0	472
HRT	Heart	0	3,287
MSL	Skeletal muscle	2	25,046
LVR	Liver	7	78,110
PNC	Pancreas	2	55,159
PST	Prostate	6	58,485
KDN	Kidney	5	67,619
LNG	Lung	25	72,024

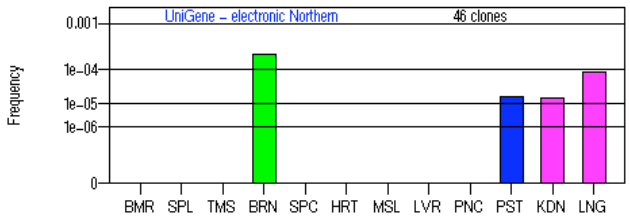
DAT1 expression in normal human tissues according to GeneNote/GeneAnnot

Affymetrix probe-set	array	sensitivity	specificity	# of associated genes
38028_at	A	1	1	1



White rectangles above bars show min-max range for duplicates





Tissue		Clones per gene	Total clones
BMR	Bone marrow	0	19,606
SPL	Spleen	0	9,440
TMS	Thymus	0	3,941
BRN	Brain	38	160,338
SPC	Spinal cord	0	472
HRT	Heart	0	3,287
MSL	Skeletal muscle	0	25,046
LVR	Liver	0	78,110
PNC	Pancreas	0	55,159
PST	Prostate	1	58,485
KDN	Kidney	1	67,619
LNG	Lung	6	72,024

SOURCE GeneReport for Unigene cluster Hs.301914

c. What structural feature about housekeeping genes can you detect when comparing these two genes when looking at overall gene structure?
Housekeeping genes are more compact than typical genes.
ARHGDIB gene is 19.6 kb and DAT1 is 59.7 kb.
Coding lengths are also different, though proteins are not so different.