河豚鱼基因组序列片段和多药耐药基因序列分析

姓名 陈耿佳 学号 1301214752 组号 G01C

1. 研究背景和文献阅读

1) 认真阅读刘勇博士论文，简述该论文的研究目的、研究方法和主要研究结果。

研究目的：验证河豚鱼作为模式生物理论，验证其含有人同源的多药耐药基因。克隆河豚鱼多药耐药基因，分析河豚鱼多药耐药基因组全长序列和邻座基因。

研究方法：用人多药耐药基因全长 cDNA 为探针与河豚鱼 cosmid 文库杂交获得候选克隆，再通过反复杂交验证获得多药耐药基因的阳性克隆，通过 ExoIII 系统和 shotgun 法获得 cosmid 124A22 的全长序列；采用生物信息学的多种软件系统对全长序列进行分析、预测，主要分析其剪接结构。

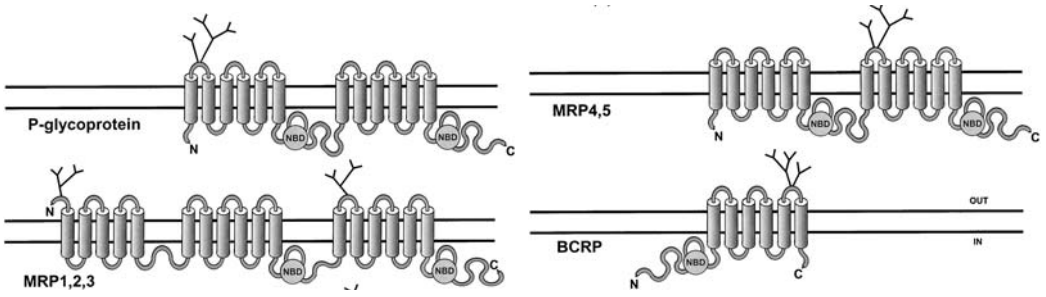
主要研究结果：河豚鱼基因组中至少含有三个多药耐药基因，头尾相连；在 cosmid 124A22 中含有两个读码框完整的多药耐药基因，其长度只有人同源基因的 1/10 左右，但 cDNA 长度与人同源基因一致，且在核苷酸水平和蛋白水平都具有较高同源性，结构也高度保守。

2) 阅读河豚鱼基因组测序计划相关论文，说明河豚鱼基因组的特点。

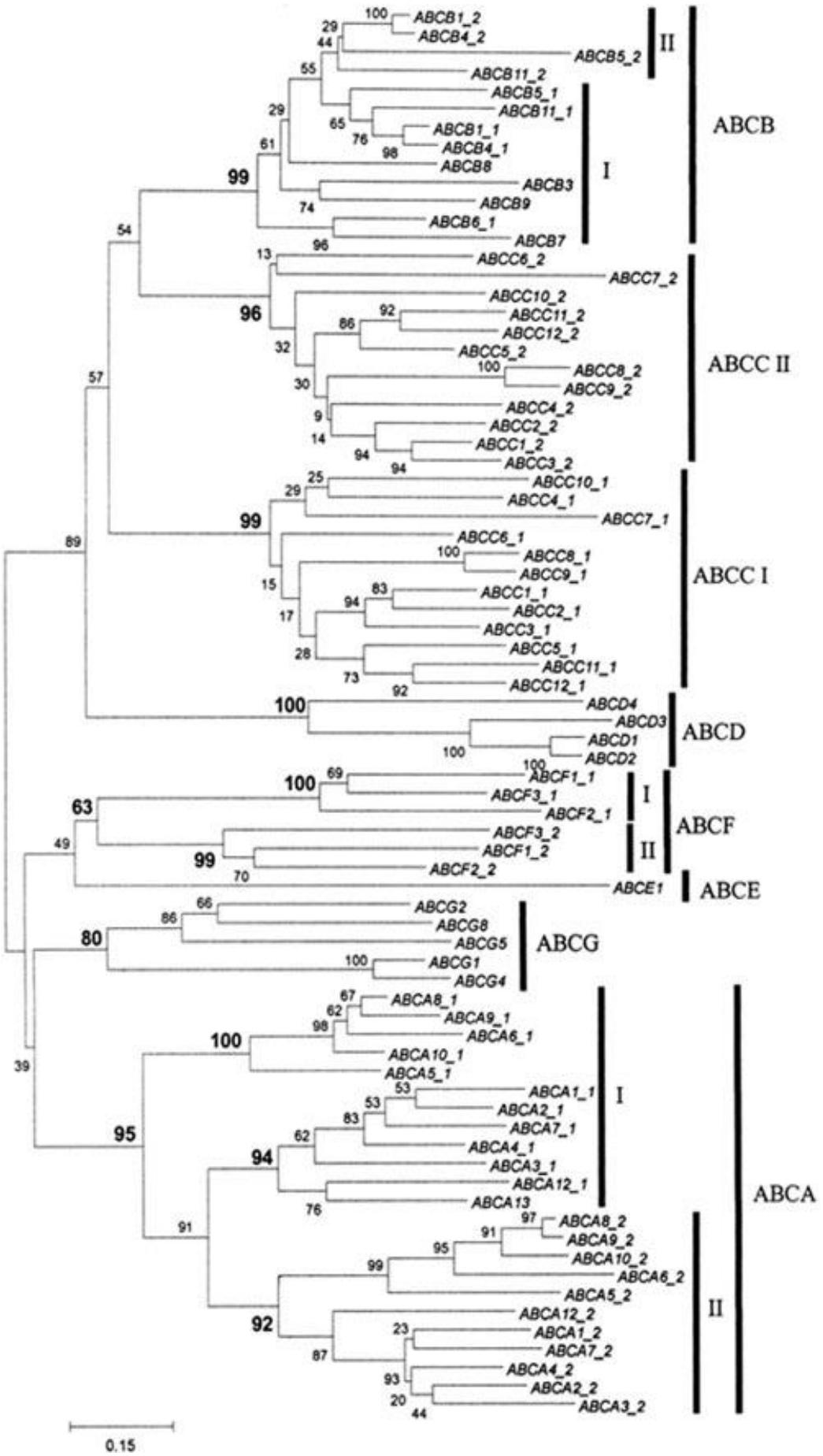
(1) 基因组序列密集，基因数与人相近，但其全长序列只有人类基因组的 1/7 – 1/8；
(2) 基因间序列少，重复序列少；
(3) 基因内含子小，但外显子数目与人基本一致；
(4) 基因分布均匀，没有明显的 G 带和 R 带；
(5) 基因之间呈一定的保守性，即同线保守。

3) 认真阅读 NCBI 免费书架提供的 ABC 转运蛋白超家族（The Human ATP-Binding Cassette (ABC) Transporter Superfamily）网络书籍及其它相关论文，简述 ABC 转运蛋白超家族分类和功能，说明人 ABCB1 家族（即多药耐药基因家，Mutidrug Resistance）的基因结构、染色体定位、序列特征和主要功能。

ABC 转运蛋白时一类具有 ATP 结合区域的单向底物转运泵，通过主动转运完成多种底物的跨膜转运。按结构可将 ABC 超家族分为全转运蛋白和半转运蛋白，前者含 2 个 ATP 结合位点（NBD）和 2 个疏水区，后者含 1 个 ATP 结合位点和 1 个疏水区，如下图所示：



目前基因序列确定的 ABC 转运蛋白超家族在人类基因组中共有 49 个成员，按序列同源性和功能可划分为 7 个亚家族：ABCA, ABCB, ABCC, ABCD, ABCE, ABCF, ABCG。人 ABC 转运蛋白超家族的系统发育树如下图所示：


ABCB1_2
ABCB4_2
ABCB5_2
ABCB11_2
ABCB5_1
ABCB11_1
ABCB1_1
ABCB4_1
ABCB8
ABCB3
ABCB9
ABCB6_1
ABCB7
ABCC6_2
ABCC7_2
ABCC10_2
ABCC11_2
ABCC12_2
ABCC5_2
ABCC8_2
ABCC9_2
ABCC4_2
ABCC2_2
ABCC1_2
ABCC3_2
ABCC10_1
ABCC4_1
ABCC7_1
ABCC6_1
ABCC8_1
ABCC9_1
ABCC1_1
ABCC2_1
ABCC3_1
ABCC5_1
ABCC11_1
ABCC12_1
ABCD4
ABCD3
ABCD1
ABCD2
ABCF1_1
ABCF3_1
ABCF2_1
ABCF3_2
ABCF1_2
ABCF2_2
ABCE1
ABCG2
ABCG8
ABCG5
ABCG1
ABCG4
ABCA8_1
ABCA9_1
ABCA6_1
ABCA10_1
ABCA5_1
ABCA1_1
ABCA2_1
ABCA7_1
ABCA4_1
ABCA3_1
ABCA12_1
ABCA13
ABCA8_2
ABCA9_2
ABCA10_2
ABCA6_2
ABCA5_2
ABCA12_2
ABCA1_2
ABCA7_2
ABCA4_2
ABCA2_2
ABCA3_2
ABCB
I
II
ABCC II
ABCC I
ABCD
ABCF
I
II
ABCE
ABCG
ABCA
I
II
0.15

它们广泛分布于机体各组织的细胞膜及亚细胞结构膜上，如肝细胞、肠上皮细胞、肾小管上皮细胞、内质网、线粒体和过氧化物酶体等。其转运底物包括糖、氨基酸、金属离子、多肽、胆固醇、胆汁酸盐、外源性疏水化合物、细胞代谢产物和药物等。此外，它们还是人体血脑屏障、血睾屏障以及血胎屏障的重要组成部分，可保护组织细胞免受毒性物质侵害。ABC 转运蛋白的编码基因及其在染色体上的定位和部分功能如下表所示：

Symbol	Alias	Location	Function
ABCA1	ABC1	9q31.1	Cholesterol efflux onto HDL
ABCA2	ABC2	9q34.3	Drug resistance
ABCA3	ABC3, ABCC	16p13.3	Surfactant secretion?
ABCA4	ABCR	1p21.3	*N*-Retinylidiene-PE efflux
ABCA5		17q24.3	
ABCA6		17q24.3	
ABCA7		19p13.3	
ABCA8		17q24.3	
ABCA9		17q24.3	
ABCA10		17q24.3	
ABCA12		2q34	
ABCA13		7p12.3	
ABCB1	PGY1, MDR1	7q21.12	Multidrug resistance
ABCB2	TAP1	6p21.3	Peptide transport
ABCB3	TAP2	6p21.3	Peptide transport
ABCB4	PGY3, MDR3	7q21.12	PC transport
ABCB5		7p21.1	
ABCB6	MTABC3	2q35	Iron transport
ABCB7	ABC7	Xq21-q22	Fe/S cluster transport
ABCB8	MABC1	7q36.1	
ABCB9		12q24.31	
ABCB10	MTABC2	1q42.13	
ABCB11	SPGP	2q24.3	Bile salt transport
ABCC1	MRP1	16p13.12	Drug resistance
ABCC2	MRP2	10q24.2	Organic anion efflux
ABCC3	MRP3	17q21.33	Drug resistance
ABCC4	MRP4	13q32.1	Nucleoside transport
ABCC5	MRP5	3q27.1	Nucleoside transport
ABCC6	MRP6	16p13.12	
CFTR	ABCC7	7q31.31	Chloride ion channel
ABCC8	SUR	11p15.1	Sulfonylurea receptor
ABCC9	SUR2	12p12.1	K(ATP) channel regulation
ABCC10	MRP7	6p21.1	
ABCC11		16q12.1	
ABCC12		16q12.1	
ABCD1	ALD	Xq28	VLCFA transport regulation
ABCD2	ALDL1, ALDR	12q11	
ABCD3	PXMP1,PMP70	1p22.1	
ABCD4	PMP69, P70R	14q24.3	
ABCE1	OABP, RNS4I	4q31.31	Oligoadenylate binding protein
ABCF1	ABC50	6p21.1	
ABCF2		7q36.1	
ABCF3		3q27.1	
ABCG1	ABC8, White	21q22.3	Cholesterol transport?
ABCG2	ABCP, MXR, BCRP	4q22	Toxin efflux, drug resistance
ABCG4	White2	11q23	
ABCG5	White3	2p21	Sterol transport
ABCG8		2p21	Sterol transport

人多药耐药基因家族有两个成员：MDR1（ABCB1）和 MDR3（ABCB4），均位于 7 号染色体 7q21.12，头尾相连，跨度为 230 kb，基因间序列 34 kb，基因长度分别为 120 kb 和 74 kb。MDR1 基因含 29 个外显子，其中 27 编码 P-glycoprotein；28 个内含子，其中 26 个位于基因编码区。全长 cDNA 约 4.5 kb。编码的蛋白 P-glycoprotein 有 1280 个氨基酸组成，N 端甲基化前后分子量分别为 140 kd 和 170 kd。P-glycoprotein 有 12 个跨膜

区和 2 个 ATP 结合位点。P-glycoprotein 可分为 N 端、C 端两个半侧，分别为[1-637]、[638-1280]位氨基酸，两个半侧各有 6 个跨膜区和 1 个 ATP 结合位点，两侧的氨基酸序列相似性达 78%，尤其是 ATP 结合位点序列高度保守。P-glycoprotein 广泛表达在与分泌和排泄有关的组织器官中，如肝脏、肾脏、小肠及血-组织屏障如血脑屏障及血胎盘屏障；但脑组织内神经元和胶质细胞不表达此类蛋白。生理状态下，P-glycoprotein 阻止外源性毒素入侵，同时排除内源性有毒物质，维持内环境的稳定，充当血脑屏障破坏之后的“第二屏障”作用，P-glycoprotein 可以限制药物或毒物进入脑内。P-glycoprotein 通过水解 ATP 提供能量，将药物从细胞内泵到胞外，也可将镶嵌在细胞膜脂质双分子层的药物泵到胞外，交叉抵抗若干结构和功能并不相关的亲脂类药物，使细胞内药物浓度下降，从而产生耐药性。

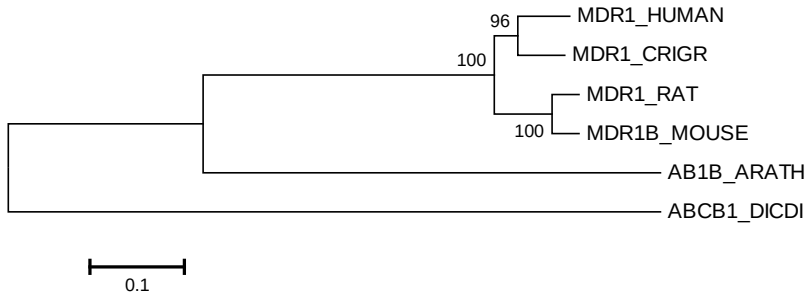
2. 数据库检索和序列分析

1) 检索 UniProtKB 序列数据库中 ABCB1 基因家族，列表说明已经过人工审阅的序列条目；对上述序列进行多序列比对，说明比对结果。

以“ABCB1”为关键词检索 UniProtKB，结果列表中前 6 条为人工审阅的 ABCB1 基因家族序列条目，如下表所示：

Entry	Entry name	Protein names	Organism	Length
Q9ZR72	AB1B_ARATH	ABC transporter B family member 1	*Arabidopsis thaliana*	1,286
Q54BU4	ABCB1_DICDI	ABC transporter B family member 1	*Dictyostelium discoideum*	909
P08183	MDR1_HUMAN	Multidrug resistance protein 1	*Homo sapiens*	1,280
P21448	MDR1_CRIGR	Multidrug resistance protein 1	*Cricetulus griseus*	1,276
P43245	MDR1_RAT	Multidrug resistance protein 1	*Rattus norvegicus*	1,277
P06795	MDR1B_MOUSE	Multidrug resistance protein 1B	*Mus musculus*	1,276

用 MEGA 6.0 软件进行多序列比对，发现 MDR1_HUMAN、MDR1_CRIGR、MDR1_RAT、MDR1B_MOUSE 的序列相似性较高，氨基酸序列高度保守，而与 AB1B_ARATH 及 ABCB1_DICDI 则差异较大。采用 Neighbor-Joining 算法，用 Bootstrap 法进行评估（500 次重复），构建系统发育树，如下图所示：



可见，在哺乳动物中 ABCB1 基因较为保守，人 ABCB1 基因与仓鼠、小鼠、大鼠同源基因亲缘关系较近，而与拟南芥和细菌的同源基因亲缘关系较远。

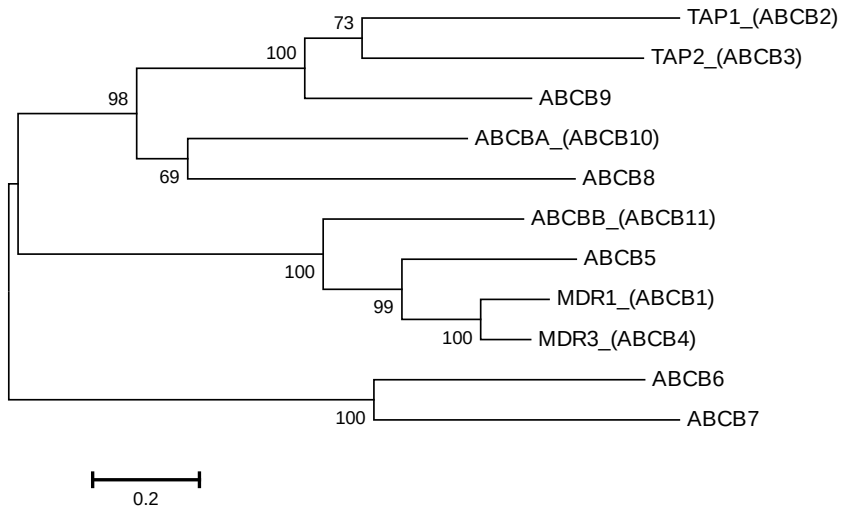
UniProt 收录的 ABCB1 基因家族成员较少，已通过人工审阅的仅 6 个物种的条目，下面对人 ABCB 基因家族进行检索。

以“family:ABCB and organism:human”进行检索，得 11 个已经过人工审阅的人 ABCB 基

因家族成员，如下表所示：

No	Entry	Name	Protein names	Length
1	Q2M3G0	ABCB5	ATP-binding cassette sub-family B member 5	812
2	Q9NP58	ABCB6	ATP-binding cassette sub-family B member 6	842
3	O75027	ABCB7	ATP-binding cassette sub-family B member 7	752
4	Q9NP78	ABCB9	ATP-binding cassette sub-family B member 9	766
5	O95342	ABCBB	Bile salt export pump	1,321
6	P08183	MDR1	Multidrug resistance protein 1	1,280
7	P21439	MDR3	Multidrug resistance protein 3	1,286
8	Q03518	TAP1	Antigen peptide transporter 1	808
9	Q03519	TAP2	Antigen peptide transporter 2	686
10	Q9NRK6	ABCBA	ATP-binding cassette sub-family B member 10	738
11	Q9NUT2	ABCB8	ATP-binding cassette sub-family B member 8	735

用 MEGA 6.0 软件进行多序列比对，发现 ABCB 基因家族各成员存在多个保守位点。采用 Maximum Likelihood 算法，构建系统发育树，如下图所示：



2) 浏览人多药耐药基因蛋白质注释信息、文献报道和数据库交叉链接，说明其功能、亚细胞定位、组织特异性、互作蛋白、结构域特征、剪接变体、序列特征、基因结构、基因组定位、表达特异性等。

功能：主动转运泵，在多药耐药细胞中将药物泵出细胞外，减少药物累积

亚细胞定位：细胞膜上

组织特异性：肝脏、肾脏、小肠、脑

互作蛋白：UBC、PIM1、MAPKAP1、CD4、BCCIP、UBA7、PSMB8、LNX1、DHX9、FBXO15、LAPTM4B、RNF2

结构域特征：包含 2 个 ABC transmembrane type-1 domain
2 个 ABC transporter domain

剪接变体：如下表所示

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype
ABCB1-001	ENST00000265724	4645	ENSP00000265724	1280	Protein coding
ABCB1-201	ENST00000543898	3651	ENSP00000444095	1216	Protein coding
ABCB1-003	ENST00000416177	461	ENSP00000399419	48	Protein coding
ABCB1-007	ENST00000488737	1864	No protein product	-	Processed transcript
ABCB1-006	ENST00000496821	913	No protein product	-	Processed transcript
ABCB1-009	ENST00000475929	787	No protein product	-	Processed transcript
ABCB1-004	ENST00000476862	582	No protein product	-	Processed transcript
ABCB1-008	ENST00000483831	642	No protein product	-	Retained intron
ABCB1-005	ENST00000482527	555	No protein product	-	Retained intron
ABCB1-010	ENST00000491360	539	No protein product	-	Retained intron

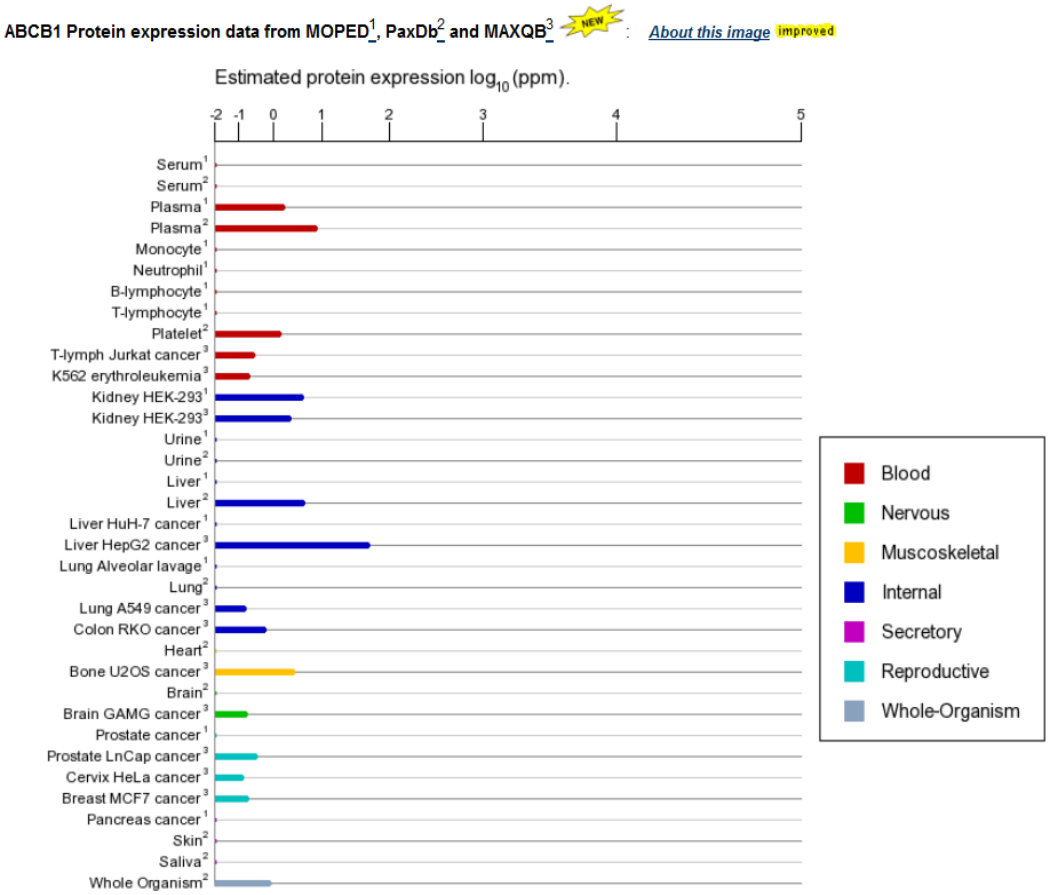
序列特征：12 个跨膜螺旋区与胞内/外连接部分交替相连

基因结构：基因长度 120 kb；含 29 个外显子，如下图所示，其中 27 编码 P-glycoprotein，长度为 1280 个氨基酸残基；28 个内含子，其中 26 个位于基因编码区；全长 cDNA 4645 bp；基因编码区各外显子和内含子长度和剪接位点如下表所示。

No	Chromosome	Start	Stop	Exon size (bp)	Intron size (bp)
1	7	87133559	87133765	206	1448
2	7	87135213	87135359	146	3232
3	7	87138591	87138797	206	5750
4	7	87144547	87144744	197	1081
5	7	87145825	87145981	156	2661
6	7	87148642	87148782	140	1310
7	7	87150092	87150192	100	10418
8	7	87160610	87160813	203	4961
9	7	87165774	87165857	83	2727
10	7	87168584	87168661	77	2012
11	7	87170673	87170780	107	2665
12	7	87173445	87173591	146	548
13	7	87174139	87174315	176	864
14	7	87175179	87175340	161	3324
15	7	87178664	87178834	170	333
16	7	87179167	87179370	203	117
17	7	87179487	87179612	125	172
18	7	87179784	87179894	110	147
19	7	87180041	87180154	113	2923
20	7	87183077	87183248	171	7331
21	7	87190579	87190703	124	4683
22	7	87195386	87195557	171	544
23	7	87196101	87196292	191	3196
24	7	87199488	87199539	51	15289
25	7	87214828	87214996	168	10086
26	7	87225082	87225130	48	4303
27	7	87229433	87229500	67	--

基因组定位：7 号染色体 NC_000007.14 (87503863..87713323, complement)

表达特异性：如下图所示：



3) 检索核酸序列数据库中人 ABCB1 基因的 mRNA 序列，提取其编码区核苷酸序列，并将其翻译成蛋白质序列。

在 RefSeq 中检索“ABCB1 human”，找到人 ABCB1 的 mRNA 序列：Homo sapiens ATP-binding cassette, sub-family B (MDR/TAP), member 1 (ABCB1), mRNA，登录号为：NM_000927.4，导出其 GenBank 格式文件，用 WebLab 中的 Coderet 软件进行分析，可获得其编码区核苷酸序列及蛋白质序列，分别如下所示：

```
>nm_000927_cds_1
atggatcttgaaggggaccgcaatggaggagcaaagaagaagaacttttttaaactgaac
aataaaagtgaaaaagataagaaggaaaagaaaccaactgtcagtgtattttcaatgttt
cgctattcaaattggcttgacaagttgtatatggtggtgggaactttggctgccatcatc
catggggctggacttcctctcatgatgctggtgtttggagaaatgacagatatctttgca
aatgcaggaaatttagaagatctgatgtcaaacatcactaatagaagtgatatcaatgat
acagggttcttcatgaatctggaggaagacatgaccaggtatgcctattattacagtgga
attggtgctggggtgctggttgctgcttacattcaggtttcattttggtgcctggcagct
ggaagacaaatacacaaaattagaaaacagtttttcatgctataatgcgacaggagata
ggctggtttgatgtgcacgatgttggggagcttaacacccgacttacagatgatgtctcc
aagattaatgaaggaattggtgacaaaattggaatgttctttcagtcaatggcaacattt
ttcactgggtttatagtaggatttacacgtggttggaagctaacccttgtgattttggcc
atcagtcctgttcttggactgtcagctgctgtctgggcaaagatactatcttcatttact
gataaagaactcttagcgtatgcaaaagctggagcagtagctgaagaggtcttggcagca
attagaactgtgattgcatttggaggacaaaagaaagaacttgaaaggtacaacaaaaat
ttagaagaagctaaaagaattgggataaagaaagctattacagccaatatttctataggt
gctgctttcctgctgatctatgcatcttatgctctggccttctggtatgggaccaccttg
```

gtcctctcaggggaatattctattggacaagtactcactgtattcttttctgtattaatt
ggggcttttagtgttggacaggcatctccaagcattgaagcatttgcaaatgcaagagga
gcagcttatgaaatcttcaagataattgataataagccaagtattgacagctattcgaag
agtgggcacaaaccagataatattaagggaaatttggaattcagaaatgttcacttcagt
tacccatctcgaaaagaagttaagatcttgaagggtctgaacctgaaggtgcagagtggg
cagacggtggccctggttggaaacagtggctgtgggaagagcacaacagtccagctgatg
cagaggctctatgaccccacagaggggatggtcagtgttgatggacaggatattaggacc
ataaatgtaaggtttctacgggaaatcattggtgtggtgagtcaggaacctgtattgttt
gccaccacgatagctgaaaacattcgctatggccgtgaaaatgtcaccatggatgagatt
gagaaagctgtcaaggaagccaatgcctatgactttatcatgaaactgcctcataaattt
gacaccctggttggagagagaggggcccagttgagtggtgggcagaagcagaggatcgcc
attgcacgtgccctggttcgcaaccccaagatcctcctgctggatgaggccacgtcagcc
ttggacacagaaagcgaagcagtggttcaggtggctctggataaggccagaaaaggtcgg
accaccattgtgatagctcatcgtttgtctacagttcgtaatgctgacgtcatcgctggt
ttcgatgatggagtcattgtggagaaaggaaatcatgatgaactcatgaaagagaaaggc
atttacttcaaacttgtcacaatgcagacagcaggaaatgaagttgaattagaaaatgca
gctgatgaatccaaaagtgaaattgatgccttggaaatgtcttcaaatgattcaagatcc
agtctaataagaaaaagatcaactcgtaggagtgtccgtggatcacaagcccaagacaga
aagcttagtaccaaagaggctctggatgaaagtatacctccagtttccttttggaggatt
atgaagctaaatttaactgaatggccttattttgttgttggtgtattttgtgccattata
aatggaggcctgcaaccagcatttgcaataatattttcaaagattataggggtttttaca
agaattgatgatcctgaaacaaaacgacagaatagtaacttgttttcactattgtttcta
gcccttggaattatttcttttattacatttttccttcagggtttcacatttggcaaagct
ggagagatcctcaccaagcggctccgatacatggttttccgatccatgctcagacaggat
gtgagttggtttgatgaccctaaaaacaccactggagcattgactaccaggctcgccaat
gatgctgctcaagttaaaggggctataggttccaggcttgctgtaattacccagaatata
gcaaatcttgggacaggaataattatatccttcatctatggttggcaactaacactgtta
ctcttagcaattgtacccatcattgcaatagcaggagttgttgaaatgaaaatgttgtct
ggacaagcactgaaagataagaaagaactagaaggttctgggaagatcgctactgaagca
atagaaaacttccgaaccgttgtttctttgactcaggagcagaagtttgaacatatgtat
gctcagagtttgcaggtaccatacagaaactctttgaggaaagcacacatctttggaatt
acattttccttcacccaggcaatgatgtatttttcctatgctggatgtttccggtttgga
gcctacttggtggcacataaactcatgagctttgaggatgttctgttagtattttcagct
gttgtctttggtgccatggccgtggggcaagtcagttcatttgctcctgactatgccaaa
gccaaaatatcagcagcccacatcatcatgatcattgaaaaaaccccttgattgacagc
tacagcacggaaggcctaatgccgaacacattggaaggaaatgtcacatttggtgaagtt
gtattcaactatcccacccgaccggacatcccagtgcttcagggactgagcctggaggtg
aagaagggccagacgctggctctggtgggcagcagtggctgtgggaagagcacagtggtc
cagctcctggagcggttctacgacccctggcagggaaagtgctgcttgatggcaaagaa
ataaagcgactgaatgttcagtggctccgagcacacctgggcatcgtgtcccaggagccc
atcctgtttgactgcagcattgctgagaacattgcctatggagacaacagccgggtggtg
tcacaggaagagattgtgagggcagcaaaggaggccaacatacatgccttcatcgagtca
ctgcctaataaatatagcactaaagtaggagacaaaggaactcagctctctggtggccag
aaacaacgcattgccatagctcgtgcccttgttagacagcctcatattttgcttttggat
gaagccacgtcagctctggatacagaaagtgaaaaggttgtccaagaagccctggacaaa
gccagagaaggccgcacctgcattgtgattgctcaccgcctgtccaccatccagaatgca
gacttaatagtggtgtttcagaatggcagagtcaaggagcatggcacgcatcagcagctg
ctggcacagaaaggcatctatttttcaatggtcagtgtccaggctggaacaaagcgccag
tga

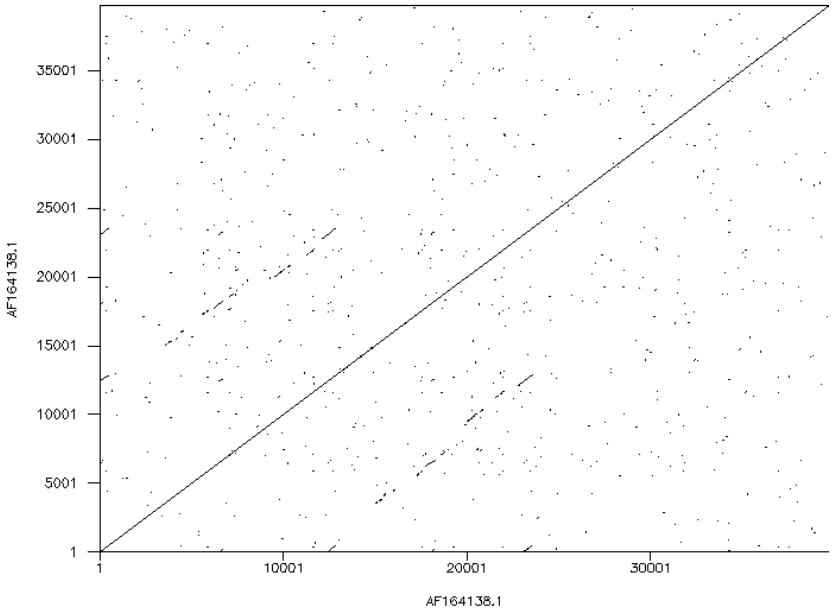
>nm_000927_pro_1
MDLEGDRNGGAKKKNFFKLNNKSEKDKKEKKPTVSVFSMFRYSNWLDKLYMVVGTLAAII
HGAGLPLMMLVFGEMTDIFANAGNLEDLMSNITNRSDINDTGFFMNLEEDMTRYAYYYSG
IGAGVLVAAYIQVSFWCLAAGRQIHKIRKQFFHAIMRQEIGWFDVHDVGELNTRLTDDVS
KINEGIGDKIGMFFQSMATFFTGFIVGFTRGWKLTLVILAISPVLGLSAAVWAKILSSFT
DKELLAYAKAGAVAEEVLAAIRTVIAFGGQKKELERYNKNLEEAKRIGIKKAITANISIG
AAFLLIYASYALAFWYGTTLVLSGEYSIGQVLTVFFSVLIGAFSVGQASPSIEAFANARG
AAYEIFKIIDNKPSIDSYSKSGHKPDNIKGNLEFRNVHFSYPSRKEVKILKGLNLKVQSG
QTVALVGNSGCGKSTTVQLMQRLYDPTEGMVSVDGQDIRTINVRFLREIIGVVSQEPVLF

ATTIAENIRYGRENVTMDEIEKAVKEANAYDFIMKLPHKFDTLVGERGAQLSGGQKQRIA
IARALVRNPKILLLDEATSALDTESEAVVQVALDKARKGRTTIVIAHRLSTVRNADVIAG
FDDGVIVEKGNHDELMKEKGIYFKLVTMQTAGNEVELENAADESKSEIDALEMSSNDSRS
SLIRKRSTRRSVRGSQAQDRKLSTKEALDESIPPVSFWRIMKLNLTEWPYFVVGVFCAII
NGGLQPAFAIIFSKIIGVFTRIDDPETKRQNSNLFSLLFLALGIISFITFFLQGFTFGKA
GEILTKRLRYMVFRSMLRQDVSWFDDPKNTTGALTTRLANDAAQVKGAIGSRLAVITQNI
ANLGTGIIISFIYGWQLTLLLLAIVPIIAIAGVVEMKMLSGQALKDKKELEGSGKIATEA
IENFRTVVSLTQEQKFEHMYAQSLQVPYRNSLRKAHIFGITFSFTQAMMYFSYAGCFRFG
AYLVAHKLMSFEDVLLVFSAVVFGAMAVGQVSSFAPDYAKAKISAAHIIMIIEKTPLIDS
YSTEGLMPNTLEGNVTFGEVVFNYPTRPDIPVLQGLSLEVKKGQTLALVGSSGCGKSTVV
QLLERFYDPLAGKVLLDGKEIKRLNVQWLRAHLGIVSQEPILFDCSIAENIAYGDNSRVV
SQEEIVRAAKEANIHAFIESLPNKYSTKVGDKGTQLSGGQKQRIAIARALVRQPHILLLD
EATSALDTESEKVVQEALDKAREGRTCIVIAHRLSTIQNADLIVVFQNGRVKEHGTHQQL
LAQKGIYFSMVSVQAGTKRQ

3. 河豚鱼柯氏质粒基因组序列片段分析

1) 提取 GenBank 中刘勇等提交的柯氏质粒河豚鱼基因序列片段全长序列，用点阵图方法确定其中是否包含重复片段。

用 WebLab 中 Dottup 程序分析 cosmid 124A22，两段输入序列均为 cosmid 124A22 序列，所得点阵图如下图所示：

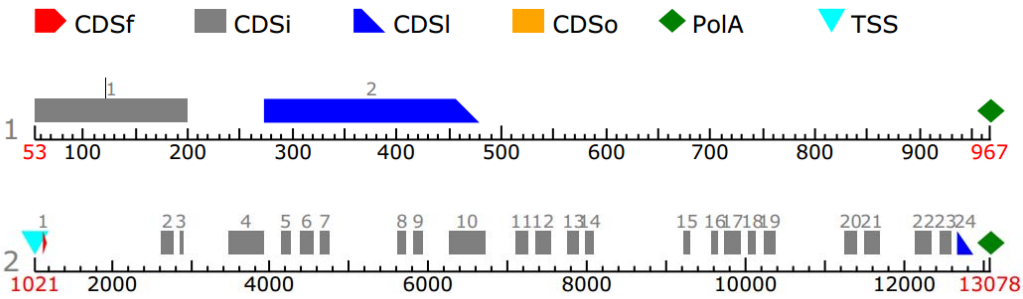


点阵图中主对角线外两段虚线表明该序列中存在重复片段，长度约为 12 kb，首尾相连，构成该序列前 24 kb 区域。

2) 用多种方法预测上述河豚鱼基因组重复片段基因结构，比较它与人 ABCB1 基因结构的异同。

提取 cosmid 124A22 序列的前 14 kb 片段，其中可能包含上述河豚鱼基因组第一个重复片段的完整基因序列。

用 Softberry 网站上的 FGENESH 工具分析上述 14 kb 片段，Organism 选择 *Takifugu rubripes* (pufferfish)，分析结果如下图所示及下表所示。预测到两个基因，都在正义链上，第一个基因不完整，不加以分析；第二个基因完整，包含 24 段 exon，基因编码区域为第 1122 – 12857 位核苷酸序列，CDS 长度为 3882 bp，编码 1293 aa 蛋白质。



G	Str	Feature	Start	End	Score	ORF	Len	IntronLen
1	+	1 CDSi	53	199	16.05	53 - 199	147	
1	+	2 CDSl	272	478	26.30	272 - 478	207	
1	+	PolA	967		1.60			
2	+	TSS	1021		- 3.84			
2	+	1 CDSf	1122	1125	2.06	1122 - 1124	3	1491
2	+	2 CDSi	2616	2768	15.55	2618 - 2767	150	77
2	+	3 CDSi	2845	2878	1.47	2847 - 2876	30	583
2	+	4 CDSi	3461	3911	31.94	3462 - 3911	450	209
2	+	5 CDSi	4120	4244	7.14	4120 - 4242	123	120
2	+	6 CDSi	4364	4535	27.01	4365 - 4535	171	86
2	+	7 CDSi	4621	4734	9.27	4621 - 4734	114	861
2	+	8 CDSi	5595	5705	7.80	5595 - 5705	111	87
2	+	9 CDSi	5792	5917	16.24	5792 - 5917	126	325
2	+	10 CDSi	6242	6700	50.29	6242 - 6700	459	378
2	+	11 CDSi	7078	7239	19.24	7078 - 7239	162	92
2	+	12 CDSi	7331	7525	23.03	7331 - 7525	195	215
2	+	13 CDSi	7740	7886	10.65	7740 - 7886	147	80
2	+	14 CDSi	7966	8070	6.51	7966 - 8070	105	1135
2	+	15 CDSi	9205	9282	7.02	9205 - 9282	78	274
2	+	16 CDSi	9556	9639	8.11	9556 - 9639	84	74
2	+	17 CDSi	9713	9916	36.69	9713 - 9916	204	102
2	+	18 CDSi	10018	10118	13.01	10018 - 10116	99	98
2	+	19 CDSi	10216	10356	13.94	10217 - 10354	138	874
2	+	20 CDSi	11230	11386	6.73	11231 - 11386	156	88
2	+	21 CDSi	11474	11671	26.16	11474 - 11671	198	452
2	+	22 CDSi	12123	12329	30.30	12123 - 12329	207	103
2	+	23 CDSi	12432	12578	15.80	12432 - 12578	147	73
2	+	24 CDSl	12651	12857	27.19	12651 - 12857	207	203
2	+	PolA	13078		1.60			

用 GENSCAN 进行分析，Organism 选择 Vertebrate，预测到 24 段外显子，基因编码区域为第 160 – 12946 位核苷酸序列，CDS 长度 4191 bp，编码蛋白长度为 1396 aa，各剪接位点如下表所示：

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Intr	+	160	306	147	0	0	22	100	231	0.960	18.13
1.02	Intr	+	379	568	190	0	1	78	-18	280	0.070	15.56
1.03	Intr	+	2723	2875	153	0	0	51	81	228	0.197	18.44
1.04	Intr	+	2952	2985	34	1	1	62	94	34	0.984	-1.32
1.05	Intr	+	3568	4018	451	1	1	85	94	476	0.913	41.20
1.06	Intr	+	4227	4351	125	2	2	45	113	140	0.999	11.58
1.07	Intr	+	4471	4642	172	1	1	101	74	244	0.999	24.25

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.08	Intr	+	4680	4841	162	2	0	-14	94	236	0.896	14.17
1.09	Intr	+	5702	5812	111	1	0	64	53	115	0.986	6.18
1.10	Intr	+	5899	6024	126	0	0	127	94	165	0.998	21.88
1.11	Intr	+	6349	6552	204	0	0	14	58	472	0.649	36.20
1.12	Intr	+	6637	6807	171	0	0	16	87	258	0.632	18.64
1.13	Intr	+	7185	7346	162	2	0	73	68	227	0.994	19.37
1.14	Intr	+	7438	7632	195	0	0	59	83	205	0.971	16.71
1.15	Intr	+	7847	7993	147	1	0	79	12	199	0.770	11.83
1.16	Intr	+	9297	9389	93	2	0	81	109	-12	0.493	0.36
1.17	Intr	+	9663	9746	84	2	0	47	83	109	0.948	6.32
1.18	Intr	+	9820	10023	204	0	0	66	2	461	0.990	34.70
1.19	Intr	+	10125	10225	101	2	2	100	21	115	0.999	4.91
1.20	Intr	+	10323	10463	141	0	0	102	76	126	0.997	12.27
1.21	Intr	+	11337	11493	157	0	1	61	59	159	0.996	10.31
1.22	Intr	+	11581	11778	198	0	0	107	69	278	0.993	27.35
1.23	Intr	+	12230	12685	456	1	0	50	100	667	0.640	57.52
1.24	Term	+	12758	12964	207	1	0	78	43	296	0.996	21.44
1.25	PlyA	+	13185	13190	6							1.05

用 Augustus 进行分析，Organism 选择与河豚鱼亲缘关系较近的象鲨 *Callorhinchus milii*，预测到 22 段外显子，基因编码区域为第 53 – 13544 位核苷酸序列，CDS 长度 4203 bp，编码蛋白长度为 1401 aa，各剪接位点如下表所示：

No	Start	End	Score	Str.	Fr.	Len	IntronLen
1	53	199	0.9	+	0	146	73
2	272	461	0.74	+	0	189	2155
3	2616	2768	0.85	+	2	152	623
4	3391	3911	0.52	+	2	520	209
5	4120	4244	1	+	0	124	120
6	4364	4535	1	+	1	171	1060
7	5595	5705	0.93	+	0	110	87
8	5792	5917	1	+	0	125	325
9	6242	6700	0.58	+	0	458	378
10	7078	7239	0.96	+	0	161	92
11	7331	7525	0.9	+	0	194	215
12	7740	7886	0.53	+	0	146	80
13	7966	8070	0.89	+	0	104	1182
14	9252	9282	0.62	+	0	30	438
15	9720	9916	0.61	+	2	196	102
16	10018	10118	1	+	0	100	98
17	10216	10356	1	+	1	140	874
18	11230	11386	0.99	+	1	156	88
19	11474	11671	0.97	+	0	197	452
20	12123	12578	0.74	+	0	455	73
21	12651	12824	0.48	+	0	173	564
22	13388	13544	0.27	+	0	156	--

用 HMMgene 进行分析，Organism 选择 *Homo sapiens*，正义链预测到 24 段外显子，基因编码区域为第 53 – 12857 位核苷酸序列，CDS 长度 4134 bp，编码蛋白长度为 1378 aa，各剪接位点如下表所示：

No	Start	End	Score	Str.	Fr.	Len	IntronLen
exon_1	53	199	0.472	+	0	146	73
exon_2	272	461	0.884	+	1	189	2155
exon_3	2616	2768	0.994	+	1	152	77
exon_4	2845	2878	0.922	+	2	33	583
exon_5	3461	3911	0.891	+	0	450	209
exon_6	4120	4244	1.000	+	2	124	120
exon_7	4364	4535	1.000	+	0	171	38
exon_8	4573	4734	0.942	+	0	161	861
exon_9	5595	5705	0.970	+	0	110	87
exon_10	5792	5917	0.963	+	0	125	325
exon_11	6242	6700	0.833	+	0	458	378
exon_12	7078	7239	0.985	+	0	161	92
exon_13	7331	7525	0.993	+	0	194	215
exon_14	7740	7886	0.762	+	0	146	1319
exon_15	9205	9282	0.887	+	0	77	274
exon_16	9556	9639	0.522	+	0	83	74
exon_17	9713	9916	0.965	+	0	203	102
exon_18	10018	10118	0.992	+	2	100	98
exon_19	10216	10356	0.990	+	2	140	874
exon_20	11230	11386	0.996	+	0	156	88
exon_21	11474	11671	0.984	+	0	197	452
exon_22	12123	12329	0.573	+	0	206	103
exon_23	12432	12578	0.756	+	0	146	73
lastex	12651	12857	0.512	+	0	206	--

反义链预测到 7 段外显子，如下表所示：

No	Start	End	Score	Str.	Fr.	Len	IntronLen
firstex	12102	12383	0.255	-	0	281	2204
exon_1	9758	9898	0.355	-	0	140	2564
exon_2	7085	7194	0.22	-	2	109	684
exon_3	6287	6401	0.301	-	0	114	2424
exon_4	3559	3863	0.247	-	2	304	3194
exon_5	216	365	0.387	-	2	149	31
lastex	92	185	0.796	-	0	93	--

用 GenID 进行分析，Organism 选择与河豚鱼亲缘关系较近的青斑河豚 *Tetraodon nigroviridis* (puffer fish)，预测到两个基因，各含 15 段和 5 段外显子，如下表所示：

Gene 1 (Forward). 15 exons. 915 aa. Score = 78.55

No	Start	End	Score	Str.	Fr.	Len	IntronLen
1	53	199	7.13	+	0	146	73
2	272	461	7.91	+	0	189	2155
3	2616	2702	1.23	+	2	86	143
4	2845	2878	0	+	2	33	583
5	3461	3911	10.84	+	1	450	209
6	4120	4244	-0.01	+	0	124	120
7	4364	4535	4.63	+	1	171	1257
8	5792	5917	0.25	+	0	125	325
9	6242	6700	13.34	+	0	458	378
10	7078	7239	6.53	+	0	161	92
11	7331	7525	8.53	+	0	194	215
12	7740	7886	0.63	+	0	146	1827
13	9713	9916	14.9	+	0	203	102
14	10018	10118	2.18	+	0	100	98
15	10216	10360	0.47	+	1	144	--

Gene 2 (Forward). 5 exons. 285 aa. Score = 31.01

No	Start	End	Score	Str.	Fr.	Len	IntronLen
1	11258	11386	3.46	+	0	128	88
2	11474	11671	4.08	+	0	197	452
3	12123	12329	9.71	+	0	206	103
4	12432	12578	6.43	+	0	146	73
5	12651	12824	7.33	+	0	173	-12824

总结：上述各种方法的预测结果有所差异，但总体都预测到了二十多段外显子，其总长度约 4 kb，编码蛋白长度 1300 aa，编码区在基因组中跨越约 13 kb。上述各种基因预测程序所用参考物种种类较少，仅 Softberry 网站上的 FGENESH 工具可选择河豚鱼 *Takifugu rubripes* (pufferfish)作为参考物种，预测结果可靠性较高，其他程序都只能找尽可能接近的物种，可靠性相对较差。

以 FGENESH 的预测结果为例，比较河豚鱼基因组重复片段基因结构与人 ABCB1 基因结构的异同，相关数据如下表所示：

物种	基因长度 (kb)	外显子数	Intron 总长度 (kb)	mRNA (bp)	编码蛋白 (aa)
人	120	27	96	3843	1280
河豚鱼	12	24	8	3882	1293

河豚鱼基因组重复片段所含基因与人 ABCB1 基因编码的蛋白长度非常接近，外显子数相近，外显子总长度相近，但人 ABCB1 基因中各内含子长度较大，内含子总长度约为河豚鱼相应内含子总长度的 10 倍，导致基因长度河豚鱼的约为人的 1/10。也就是说，河豚鱼该基因与人 ABCB1 基因的基因结构形式相似，都由 20 多段外显子构成，两者的有效编码信息长度非常接近，但河豚鱼该基因结构比较“紧凑”，总基因长度只有人 ABCB1 基因的 1/10。

3) 分析上述重复序列的全长基因 DNA 序列、编码区序列及其翻译得到的氨基酸序列的相似性，说明分析结果。

用 FGENESH 分别分析 cosmid 124A22 序列的 1 bp - 14 kb 片段和 12 kb – 26 kb 片段，即上述两个重复序列各自所在区域，取主要预测结果，相关信息如下表所示：

片段	外显子数	基因长度(bp)	mRNA (bp)	编码蛋白(aa)
Repeat_1	24	11735	3882 bp	1293
Repeat_2	26	9816	3873 bp	1290

用 WebLab 中 Water 程序分析上述重复序列的全长基因 DNA 序列（各 14 kb）、FGENESH 预测的编码区序列及其翻译得到的氨基酸序列，结果如下表所示：

	Length	Score	Identity	Similarity	Gaps
Genome	13773	11341	7043/13773 (51.1%)	7043/13773 (51.1%)	3510/13773 (25.5%)
mRNA	3891	13236	3165/3891 (81.3%)	3165/3891 (81.3%)	183/3891 (4.7%)
Protein	1295	5181	1015/1295 (78.4%)	1130/1295 (87.3%)	61/1295 (4.7%)

可看到，这对重复序列全长基因 DNA 序列相似性只有 51%，而编码区序列（仅外显子）相似性为 81%，编码蛋白产物氨基酸序列相似性为 87%。两基因产物相似性很高，外显子区保守程度较高，但由于内含子区差异较大，导致全长基因 DNA 序列相似性不高。

4) 用 WebLab 中 coderet 程序提取上述柯氏质粒河豚鱼基因组序列条目中所有基因的编码区序列和所编码的氨基酸序列。

下载 Takifugu rubripes cosmid 124A22 的 GenBank 格式文件[GenBank: AF164138.1]，用 WebLab 中 coderet 程序提取其所有基因的编码区序列和所编码的氨基酸序列，得 5 组序列。编码区序列如下：

>af164138_cds_1
atggccattgtgaacgggctggtgaatcctctgatgtgtatcgtgtttggtgagatgact
gacagcttcatccaggaagccaaactgtcccaaaaccacaacacaagcaaccccagagca
aacagcaccttagaagcagatatgcagagattctccatctactactccatcttggggttt
gctgtgctggtagtggcgtacctgcagatgtctctgtggaccctaacggccgcgcggcag
gccaaacgaattcgcgagttgtttttccacggcatcatgcagcaggacatcagctggtat
gacgtgactgagacaggagagctcaacacgcgtctcacagagtgggtgacgcacatcata
cacactccagttcctgtcacagctggcgtggtcgttatcatatgtggtgttcgattccct
ggtgcgcacgatgtctacaagatccaggagggcatcggtgacaaggcgggtctgctgatc
caggcggcctccacctttatcacttcctttgttattggttttgtacatggatggaagctc
accctggtcatcctggccatcagccctgtgttgggtctctcagctgccctttacagtaag
ttgctgacaagcttcaccagtaaagagcagacagcgtacgccaaagctggagctgtggca
gcggaggtgctatcctccatcaggactgtgtttgccttcagtggccaaagaaaagccatc
aaaagatatcataagaacctggaggatgcgagggacatgggaataaagaagggagttgct
gctaacacggccacaggcttctcctttctgatgatctacctgtcctatgctctggccttc
tggtacgggactactctggtcctcaacaaagagtacaccattggaaatttactgactaat
aagagcgttgctgcagaaacagtgaccacgtgtgtccaaatgaaggtgttcttcgtcgtc
ctctacggggcatacattattggacaggcctctcccaacgtccagtcctttgccagtgcc
agaggagcggcgtataaagtctacaacattatcgaccacaaaccgaatattgacagcttt
tcagaggacggatacaagcctgaatacatcaaaggtgacattgtattccagaacatccac
ttcagctacccttcgaggccagaaattaaaatcttaaacgacatgtcgtttcatgtgagg
aacggacagaccattgctttggtggggagcagcggctgtggtaaaagtaccaccatccag
ctgctgcagaggttctacgacccccagaaaggatccatatttatcgacggtcacgacatc
cgctccctcaacatccgctacctgagagaaatgatcggagtggtcagccaggagcccgtt
cttttcgccaccaccatcaccgagaacatcagatacggccgactggacgtgacgcaagag
gagatcgaacgagccactaaagagtccaacgcttatgacttcatcatgaaccttccagac
aagtttgagacgctggtgggagatcgagggactcagctgagcggaggacagaagcagagg
atcgccatcgctcgagctctggtccgcaaccctaaaatcctcctgctggacgaggccacg
tctgccctcgatgctgagagcgagaccatcgtacaggctgctctggacaaggtccgactg
ggtcgcaccaccatcgtcatcgctcaccgactctcgaccatcagaaacgccgacatcatt
gctggattcagtaatggtgaaatcgttgagcaggggactcacagccagctgatggagata
aagggagtctatcatggcctggtgaccatgcagagctttcagaagctggaggatctggaa
gactcagactacgagccctgggtcgctgagaagagccagctgatcgaatccttctcccag
tcctccctgcagaggaggaggtccactagaggctccttgcttgctgtctcagaaggaaca
aaagaggagaaagaaaaatttgagtgcgatcaggacaacatagaggaggatgagaacgtt
cctcccgtgtcgttctttaaagtgatgcgttacaacgtttctgagtggccgtatattttg
gtaggaaccatctgcgccatgatcaacggtgcgatgcagccagtgttcagcatcatcttc
accgagatcattatgttttggggtttccagggtttctgtttcagtaaatctggagaaatt
ctgaccctgaacctcagactcaaagccttcatatctatgatgagacaggacctcagctgg
tacgacaatcccaaaaacaccgttggcgctctcaccactaggctggctgccgacgccgcc
cacgtacaaggagctgcaggggtgcgcctggctgtaatgacgcagaacttcgccaacctg
ggcaccagcatcatcatcagcttcgtgtacggctgggagctgaccctgctcatcctggcc
gtggtgcccatcctggctgtggccggagccgctgaggtcaagctgctgacaggacacgcc
gccgaagacaagaaggagctggagatggccggaaagatcgccacagaggccatcgagaat
gtgagaactgtggtgtccctcaccagagaaccgacatttgtggctttatacgaggaaaat
ctaactgttccatacaagaactcccagaaaaaggccaaaatttatggcttaacctactcc
ttctcacaggccatgatcttctttgtttacgctgcctgtttccgctttggagcctggctg
atcgaagcaggacggatggatgtggagggagtgttccttgtggttatgacaatgctgtac
ggcgccatggctgtcggcgaggccaacacttatgctcccaacttcgccaaagccaaaatc
tcagcctcccacctgacgatgctaataaacagacagccggccatagataatctgtcagag
gaggaagcgagactggagaaatacgacggcaacgttcttttgaggacgtcaagtttaac

tacccgtcgcggcccgatgtgcctgtactacaagggctgaatctggaggtgcaaaaggga
gaaactctggccttggtgggcagcagcggttgtggaaagagcaccaccatccagctgctg
gagaggttttatgaccccagagaggggagagtgttgctggacggtgtcgatgtgaaacag
ctgaacgttcactggctgaggtctcagatcggcatcgtctcccaggagccggtgctgttc
gactgctccctggctgagaacatcgcctacggagacaacagtcgctccgtgtccatggat
gagatagtagctgctgctaaagcagccaacatccacagcttcatcgaagggctgcctcag
gtagcggctgtgaatcaggggaaatggttgattccacatttgatcgattcccatggagct
gcccatgaccatttacaccatatacaaactgtctctgagcagagatacgacactcaggct
ggtgataagggaacacagctgtcaggggccagaagcagcgtgtcgccatagcccgagcc
atcatccgcaaccccaaactgttgctcctggacgaggccacgtctgcgctcgacactgag
agtgagaaggtggtgcaggaggcgttggaccaggccaggaagggcaggacgtgcatcgtc
gtagcccaccgtctgtccaccatccagaacgccgactgcatcgctgtgttccagggagga
gtggtggtggaaaaggggacgcaccagcagctgatcgccaagaagggagtgtaccacatg
ctggtcaccaaacagatgggctatcacagtggatga

>af164138_cds_2

atggccttaaagatcgatacggccgaaacaaacggtgatctgagccatgattccaaggac
gatggtgccaagaatgaaaagaaaaagaagaataaaaaggaaaagccaccacaggagccc
atggtgggccccattactctgttccgatttgcagaccgctgggacgtcgtgctgctcatc
agcgggacagtgatggccatggtcaacggcaccgtgatgcccctcatgtgcattgtcttt
ggagaaatgacggacagttttatatacgctgacatggcccaacacaacgcaagtggctgg
aattctactactactattctgaacagcacgttacaggaggacatgcaaagattcgccatt
tattactccgtcttgggatttgttgtgctgctggccgcctacatgcaggtgtccttctgg
accataacagccgggcgccaggtgaaacgcatccgcagcttgtttttccactgcatcatg
cagcaggagatcagctggtttgacgtgaacgacacaggggagctcaacactcgactgacg
gaagagttcccagcttcagcgttcacgctctgtacggctacgctcggaggtgtagatgat
ctgatggacgtgcttctttttttccaatggcagcgacgtctacaagatccaggagggcatc
ggtgacaaggtggggctgctgatccaggcgtacaccaccttcatcacggccttcatcatc
ggcttcaccacgggctggaaactgacgctggtcatcctggccgtgagccccgcgctggcc
atttcggccgccttcttcagtaaagtgcttgcgtccttcaccagtaaggagcagacggcg
tacgccaaagccggagccgtggcggaggaagtgctgtccgccatcaggaccgtgttcgcc
ttcagtggtcagaccagagagattgagagataccacaagaacctgcgggacgcaaaggac
gtgggagtgaagaaggccatctcctccaacatcgccatgggcttcaccttcctgatgatc
tacctgtcctatgctctggccttctggtacgggagtacgctcatcctgaattttgagtac
accatcggcaatttactgactgtgttttttgtcgtgcttattggagcgttcagcgtcgga
cagacctctccgaacatccagaattttgccagcgcccgaggagccgcctataaagtgtac
agcatcatcgataacaagccaaacattgacagcttttcagaggacggtttcaagccggac
ttcatcaaaggtgacatcgagttcaagaacatccacttcaattacccttcgaggcctgaa
gtcaaaatcttgaacaacatgtctctgagcgtgaagagcggacagaccattgctttggtg
gggagcagcggctgtggcaaaagtaccaccatccagctgctgcagaggttctacgacccc
gaggaaggagctgtatttatcgacggtcacgacatccgctccctcaacatccgctacctg
agagagatgatcggagtggtcagccaggagcccgttcttttcgccaccaccatcaccgag
aacatcagatacggccgactggacgtgacgcaagaggagatcgaacgagccactaaagag
tccaacgcttatgacttcatcatgaaccttccagacaagtttgagacgctggtgggagat
cgagggactcagctgagcggaggacagaagcagaggatcgccatcgctcgagctctggtc
cgcaaccctaaaatcctcctgctggacgaggccacgtctgccctcgatgctgagagcgag
accatcgtacaggctgctctggacaaggtccgactgggtcgcaccaccatcgtggtcgct
caccgactctcgaccatcagaaacgccgacatcattgctggattcagtaatggcaaaatc
gtggagcaggggactcacagccagctgatggagataaagggagtctatcatggcctggtg
accatgcagacgttccacaatgtggaggaggaaaataccgccatgtcggagttatctgct
ggggagaagagccctgtggaaaagaccgtctcccagtcgtccatcatcaggaggaagtcc
accagagggtcctcgtttgccgcgtcagaaggaaccaaagaggaaaagacagaagaggat
gaagacgttcccgacgtgtcgttctttaaagtgctgcatctgaacatccccgagtggccc
tacatccttgtggggctcatctgcgctacgatcaatggagccatgcagccggtcttcgcc
atcctcttctccaagatcatcactgtgtttgcggatccagaccgtgattctgtcaggagg
aagagtgaattcatttctctgatgtttgtcgttattggctgtgtgtcatttgtcaccatg
tttttacagggttactgtttcggtaaatccggagagattctgacgctgaagctgagactc
cgggcgttcacggcgatgatgagacaggacctcagctggtacgacaatccccaaaacacc
gttggcgctctcaccactaggctggctgccgacgccgcccaagtacaaggagctgcaggg
gtgcgcctggcgacaataatgcagaacttcgccaacctgggcaccagcatcatcatcgcc

tttgtttacggctgggagctgaccctgctcatcctggccgtggtgcccctcatcgcggcc
gccggagccgctgagatcaagctgctcgcgggtcacgccgccaaagacaagaaggagctg
gagaaggccggaaagatcgccacagaggccatcgagaacgtcagaaccgtcgtgtccctc
agcagagaaccaaaatttgagtgtttatatgaggagaatctcagagtgccgtacaagaac
tcccagaaaaaggcccacgtgtacggcttaacctactccttctcccaggccatgatctac
tttgcttacgctgcctgtttccgcttcggagcctggctgattgaagcaggacggatggac
gtggagggagtgttcctggtggtttctgcggtgctgtacggcgccatggccgtgggggaa
gctaacacctttgctccgaactacgccaaggccaaaatggctgcttcctacctgatgatg
ctaataaacaagaagcccgccattgataacctctcagaggaggggacgtctccggaaaaa
tacgacggtaatgtgcatttcgagggtgttaaattcaactacccgtcgcggcccgatgtg
accatactccaggggctgaacctgaaggtgaaaaagggagaaactctggccttggtgggc
agcagcggttgtggaaagagcaccaccatccagctgctggagaggttttatgaccccaga
gaggggagagtgtcactggacggtgtcaacgtgaaacagctgaacattcactggctgagg
tctcagatcggcatcgtctcccaggagccggtgctgttcgactgctccctggctgagaac
atcgcctacggagacaacagtcgctccgtgtccatggatgagataagatacgacactcag
gctggtgataagggaacacagctgtcaggggccagaagcagcgtgtcgccatagcccga
gccatcatccgcaaccccaaactgttgctcctggacgaggccacgtctgcgctcgacact
gagagtgagaaggtggtgcaggaggcgttggaccaggccaggaagggcaggacgtgcatc
gtggtagcccaccgtctgtccaccatccagaacgccgactgcatcgctgtgttccaggga
ggagtggtggtggaaaaggggacgcaccagcagctgatcgccaagaagggagtgtaccac
atgctggtcaccaaacagatgggctatcacaacgactga
>af164138_cds_3
atgtccaagcaggcagaatttgagaagattgcagaggatgtgaagaaagtgaagacgagg
ccgacagaccaggagctgctggatctgtatggcctttacaaacaggcaattgttggagac
gtcaatacggacaggccaggacttctggatttaaagggaaaagctaagtgggatgcctgg
gaatccaggaaagttcgtccttttgcgtccgagaaggaattcaaggccacagaggacatt
gtgaggaacttccaacagggtgttggaaaagagctgcaccagaggctcctgcagagagct
gaaaccaggaggaactggatgtttaacactgtgctgtcctcccagctggagcaatggtgg
ttagacgctgcttatctggagggccgcagcccctctcagctgactgtgaacttcgcaggg
ccggcaccttatctggaacactgctggcctcctgccgagggaacagccctggagcgagcc
agtatttgctcgtggcatatgctgcagtactggaatctgatccgcacggagaggttggcg
ccgcagaaagctggcgaaacacctttagatatggaccagttcaggatgctgtactgcacc
tgcaaagtacccggggtgacgaaagacgctattcgtagctactttaaaacagagctcgag
gggaggtgcccttcccatttggtggttctttgtcgtggacgcatcttcacatttgatgcc
ctctgtgatggacaaatactgacgcccccagaactgttcaggcagctgagctacgtgaga
cagtgctgtgatgggaacccagaggggagggagtgagcgctctcactactgaagagagg
acgcgctgggcgaaggcccgagagtatctaataagtattgatccgcacaacgagaccatc
ctggagctcatccagagcagcctgttcaccatatgtctggatgagacgcagccttactcc
actccagagaactacaccaacctcacacgggagtctctcacgggtgatcccaccatccgc
tgggggacaaatcctacaattcagtcgtctattcagatggaacgtttggatccaactgt
gatcacgcgccgtacgacgccatggtgctggtgaccatgtgctggtacgtggaccagcga
attcaaagcaccggaggcaaatggaagggtgtggacacagtcagagtcctgccgcctccc
gaggagctggtgtttactgtggacgaaaaagtccgcagcgacatcggccgtgcgaaaaaa
caatactttgagtcggcgcaggacctgcaggttgtctgttacgccttcacggctttcgga
aaagccgccatcaagcagaaaaagctgcacccggacacgttcatccaactggcgatgcag
ctggcgtactttaaactgcaccagaggccagggtgttgctacgagacagccatgactcgc
aagttctaccacggcaggacggagaccatgaggccctgcaccgtggaggcggtgaaatgg
tgcacggccatgacggacccgtcgtgcgaggacaacgctaagaggaaagccatgcagctg
gcctttgagaaacacaacaacctgatggccgaggcccaggaaggacgaggcttcgacagg
caccttctcggcctgtatctcatcgccaaagaggagggacgtcctgttccggaactcttc
ttagatccgctctatgccaagagtggcggtggcggaaactttgtgctgtcgtccagcctg
gtgggctacaccacagttctgggcgcggtggcgccgatggttccccacggctacggcttc
ttctaccgtatccgagaggacaggattgtgatttccatatcggcctggaagtcctgccgc
cagaccgacgccgtgtccctgttcaacgtcttcagcagctgtctgcacgagatgctgcac
ctggcaacaacgtctcagctctga
>af164138_cds_4
atggctgctgcagcaaagtctgctaaaaaggaaagcaaaaggtatattcccacaaagacc
tgcttcacttcaccattcacgccaaaatggagcccgctcccgcaagaagacatgcatttc
atcctgaacaccttaaaggaaaactttgtttccatcggacttgtgaaaaaagagccaaag

gtgtttcgtccttggcgtaaaaagaaaaagcaggaagctgctcaatcacaagattcagac
ctccaggtgagccaggatgctgcgagtcaggaacctccaaaacgtggatggacagatgtg
gcagctagaagaaagctggccattggaatcaacgaggtcaccaaagctttggagaggaat
gagctcaaactgctgctagtgtgtaagtgtgtcaagccacaacacatgatggagcacctc
ataacgctgagcacaacgagagacgtccctgcctgccaggtgcctcggctcagccagagt
gtgtcggagcctctggggctaaaaagcgtcctcgccttaggattcagacaatgtctcccc
caagagagggatgtgttcagtaacgtggttgaagccattttacccaaagtgccaccactg
gatgttccctggctccaagatacaccggccagtataaaacctgacgaaaacagaggccag
aagaggaggctcgaaactgagtctgaggaagggacgcctgtctcctccacaactctacaa
cctctcaaagtgaagaaaatagttcccaactctgcgaggaaaggcaaaggaaagaaaaag
gtctaa

>af164138_cds_5

atggcggaggacagtgaatccgcagctagccagcagagcctggagctggacgaccaggac
acatgcgggatagacggggacaacgaagaggagaatgagcatctgcaagggagtccggga
ggggatttgggggctaaaaggaagaagaagaaacagaagaggaagaaagagaagccgagt
tcgggggggagccaagtccgactctgcctctgactcccaggagttcaagaaccctactttg
cccattcagaagctgcaggacattcaacgagccatggagttactctcctgtcagggtcct
gcaaagagcatcgacgaggcggccaagcacaagtaccagttctgggacacgcagcctgta
cccaagttaaacgaggtggtgacgagtcacgggccaatagaggccgacaaagaaaacatt
cgacaggagccatattctttacctcaaggttttatgtgggacacgctggatctgggcagc
gcagaagtgctgaaggagttgtacacgttactgaacgagaactacgtggaggacgacgac
aacatgttcagattcgactattcgccaaactttctcaaatgggctctgcgtccgccgggc
tggctcccccagtggcactgcggcgtgcgagtgtcgtcgaacaagaagctggtgggcttc
atcagcgccattcccgctgacatccgcatctacgacacagtgaagaggatggtggaaatc
aacttcctgtgtgtgcacaagaagctgcgttcgaagcgcgtcgccccggtgctcatcagg
gagatcacgcggagggtgaacctagagggcatatttcaggccgtttacacggcaggagtg
gttctgcccaaacccgtgtccacgtgcaggtactggcaccgttctctgaaccccaggaag
cttgtggaagtgaagttctcccacctgagcagaaacatgaccctgcaacggaccatgaag
ctctacagattaccagacagcacgaagactcccggtctgcggccaatggagaggcgcgac
atccgccaggtcacagagctgctacagaaattcctgaaacgcttccagctcgcaccttcc
atgacggaagaggaggtgtctcactggttcctgccgcaggacaacataattgacacttat
gtagtggaggtagccgggatcagcctgaaggactcagacccagagtgcctcggtgctggg
ggcgcgctgacagactttgctagtttctacactctgccctcgactgtgatgcaccaccct
ctccacaggagcctgaaggccgcctactctttttacaacgttcacacacaaaccccctctc
ctggatttgatgaacgacgcactgatcctggccaaactgaaagggttcgatgttttcaac
gccctggatctcatggagaataaagtgttcctggagaagctcaagtttggtataggagat
ggaaatctgcagtattacctctacaactggaaatgtccatctatggagcctgataagccg
tggttgcctttcaggtcggcctcgtccttcagtagcagggttcctcagggctgctacaca
aacactgggccaaaggtcaccacggaccgtggtagttgtcacctggacatgcagttcgct
gggaggggggggacaactcgtcaatcatcaacatcctgaatgtgaagtcatgctatgcc
tga

所编码的氨基酸序列如下：

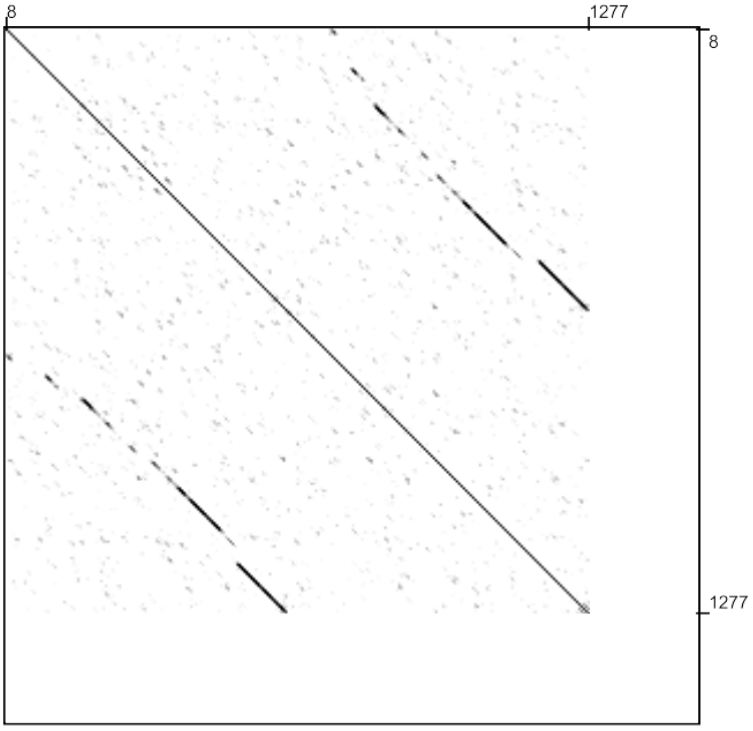
>af164138_pro_1

MAIVNGLVNPLMCIVFGEMTDSFIQEAKLSQNHNTSNPRANSTLEADMQRFSIYYSILGF
AVLVVAYLQMSLWTLTAARQAKRIRELFFHGIMQQDISWYDVTETGELNTRLTEWVTHII
HTPVPVTAGVVVIICGVRFPGAHDVYKIQEGIGDKAGLLIQAASTFITSFVIGFVHGWKL
TLVILAISPVLGLSAALYSKLLTSFTSKEQTAYAKAGAVAAEVLSSIRTVFAFSGQRKAI
KRYHKNLEDARDMGIKKGVAANTATGFSFLMIYLSYALAFWYGTTLVLNKEYTIGNLLTN
KSVAAETVTTCVQMKVFFVVLYGAYIIGQASPNVQSFASARGAAYKVYNIIDHKPNIDSF
SEDGYKPEYIKGDIVFQNIHFSYPSRPEIKILNDMSFHVRNGQTIALVGSSGCGKSTTIQ
LLQRFYDPQKGSIFIDGHDIRSLNIRYLREMIGVVSQEPVLFATTITENIRYGRLDVTQE
EIERATKESNAYDFIMNLPDKFETLVGDRGTQLSGGQKQRIAIARALVRNPKILLLDEAT
SALDAESETIVQAALDKVRLGRTTIVIAHRLSTIRNADIIAGFSNGEIVEQGTHSQLMEI
KGVYHGLVTMQSFQKLEDLEDSDYEPWVAEKSQLIESFSQSSLQRRRSTRGSLLAVSEGT
KEEKEKFECDQDNIEEDENVPPVSFFKVMRYNVSEWPYILVGTICAMINGAMQPVFSIIF
TEIIMFWGFQGFCFSKSGEILTLNLRLKAFISMMRQDLSWYDNPKNTVGALTTRLAADAA
HVQGAAGVRLAVMTQNFANLGTSIIISFVYGWELTLLILAVVPILAVAGAAEVKLLTGHA
AEDKKELEMAGKIATEAIENVRTVVSLTREPTFVALYEENLTVPYKNSQKKAKIYGLTYS

FSQAMIFFVYAACFRFGAWLIEAGRMDVEGVFLVVMTMLYGAMAVGEANTYAPNFAKAKI
SASHLTMLINRQPAIDNLSEEEARLEKYDGNVLFEDVKFNYPSRPDVPVLQGLNLEVQKG
ETLALVGSSGCGKSTTIQLLERFYDPREGRVLLDGVDVKQLNVHWLRSQIGIVSQEPVLF
DCSLAENIAYGDNSRSVSMDEIVAAAKAANIHSFIEGLPQVAAVNQGKWLIPHLIDSHGA
AHDHLHHIQTVSEQRYDTQAGDKGTQLSGGQKQRVAIARAIIRNPKLLLLDEATSALDTE
SEKVVQEALDQARKGRTCIVVAHRLSTIQNADCIAVFQGGVVVEKGTHQQLIAKKGVYHM
LVTKQMGYHSG
>af164138_pro_2
MALKIDTAETNGDLSHDSKDDGAKNEKKKKNKKEKPPQEPMVGPITLFRFADRWDVVLLI
SGTVMAMVNGTVMPLMCIVFGEMTDSFIYADMAQHNASGWNSTTTILNSTLQEDMQRFAI
YYSVLGFVVLLAAYMQVSFWTITAGRQVKRIRSLFFHCIMQQEISWFDVNDTGELNTRLT
EEFPASAFTLCTATLGGVDDLMDVLLFSNGSDVYKIQEGIGDKVGLLIQAYTTFITAFII
GFTTGWKLTLVILAVSPALAISAAFFSKVLASFTSKEQTAYAKAGAVAEEVLSAIRTVFA
FSGQTREIERYHKNLRDAKDVGVKKAISSNIAMGFTFLMIYLSYALAFWYGSTLILNFEY
TIGNLLTVFFVVLIGAFSVGQTSPNIQNFASARGAAYKVYSIIDNKPNIDSFSEDGFKPD
FIKGDIEFKNIHFNYPSRPEVKILNNMSLSVKSGQTIALVGSSGCGKSTTIQLLQRFYDP
EEGAVFIDGHDIRSLNIRYLREMIGVVSQEPVLFATTITENIRYGRLDVTQEEIERATKE
SNAYDFIMNLPDKFETLVGDRGTQLSGGQKQRIAIARALVRNPKILLLDEATSALDAESE
TIVQAALDKVRLGRTTIVVAHRLSTIRNADIIAGFSNGKIVEQGTHSQLMEIKGVYHGLV
TMQTFHNVEEENTAMSELSAGEKSPVEKTVSQSSIIRRKSTRGSSFAASEGTKEEKTEED
EDVPDVSFFKVLHLNIPEWPYILVGLICATINGAMQPVFAILFSKIITVFADPDRDSVRR
KSEFISLMFVVIGCVSFVTMFLQGYCFGKSGEILTLKLRLRAFTAMMRQDLSWYDNPQNT
VGALTTRLAADAAQVQGAAGVRLATIMQNFANLGTSIIIAFVYGWELTLLILAVVPLIAA
AGAAEIKLLAGHAAKDKKELEKAGKIATEAIENVRTVVSLSREPKFECLYEENLRVPYKN
SQKKAHVYGLTYSFSQAMIYFAYAACFRFGAWLIEAGRMDVEGVFLVVSAVLYGAMAVGE
ANTFAPNYAKAKMAASYLMMLINKKPAIDNLSEEGTSPEKYDGNVHFEGVKFNYPSRPDV
TILQGLNLKVKKGETLALVGSSGCGKSTTIQLLERFYDPREGRVSLDGVNVKQLNIHWLR
SQIGIVSQEPVLFDCSLAENIAYGDNSRSVSMDEIRYDTQAGDKGTQLSGGQKQRVAIAR
AIIRNPKLLLLDEATSALDTESEKVVQEALDQARKGRTCIVVAHRLSTIQNADCIAVFQG
GVVVEKGTHQQLIAKKGVYHMLVTKQMGYHND
>af164138_pro_3
MSKQAEFEKIAEDVKKVKTRPTDQELLDLYGLYKQAIVGDVNTDRPGLLDLKGKAKWDAW
ESRKVRPFASEKEFKATEDIVRNFQQGVGKELHQRLLQRAETRRNWMFNTVLSSQLEQWW
LDAAYLEGRSPSQLTVNFAGPAPYLEHCWPPAEGTALERASICSWHMLQYWNLIRTERLA
PQKAGETPLDMDQFRMLYCTCKVPGVTKDAIRSYFKTELEGRCPSHLVVLCRGRIFTFDA
LCDGQILTPPELFRQLSYVRQCCDGNPEGEGVSALTTEERTRWAKAREYLISIDPHNETI
LELIQSSLFTICLDETQPYSTPENYTNLTRESLTGDPTIRWGDKSYNSVVYSDGTFGSNC
DHAPYDAMVLVTMCWYVDQRIQSTGGKWKGVDTVRVLPPPEELVFTVDEKVRSDIGRAKK
QYFESAQDLQVVCYAFTAFGKAAIKQKKLHPDTFIQLAMQLAYFKLHQRPGCCYETAMTR
KFYHGRTETMRPCTVEAVKWCTAMTDPSCEDNAKRKAMQLAFEKHNNLMAEAQEGRGFDR
HLLGLYLIAKEEGRPVPELFLDPLYAKSGGGGNFVLSSSLVGYTTVLGAVAPMVPHGYGF
FYRIREDRIVISISAWKSCRQTDAVSLFNVFSSCLHEMLHLATTSQL
>af164138_pro_4
MAAAAKSAKKESKRYIPTKTCFTSPFTPKWSPLPQEDMHFILNTLKENFVSIGLVKKEPK
VFRPWRKKKKQEAAQSQDSDLQVSQDAASQEPPKRGWTDVAARRKLAIGINEVTKALERN
ELKLLLVCKCVKPQHMMEHLITLSTTRDVPACQVPRLSQSVSEPLGLKSVLALGFRQCLP
QERDVFSNVVEAILPKVPPLDVPWLQDTPASIKPDENRGQKRRLETESEEGTPVSSTTLQ
PLKVKKIVPNSARKGKGKKKV
>af164138_pro_5
MAEDSESAASQQSLELDDQDTCGIDGDNEEENEHLQGSPGGDLGAKRKKKKQKRKKEKPS
SGGAKSDSASDSQEFKNPTLPIQKLQDIQRAMELLSCQGPAKSIDEAAKHKYQFWDTQPV
PKLNEVVTSHGPIEADKENIRQEPYSLPQGFMWDTLDLGSAEVLKELYTLLNENYVEDDD
NMFRFDYSPNFLKWALRPPGWLPQWHCGVRVSSNKKLVGFISAIPADIRIYDTVKRMVEI
NFLCVHKKLRSKRVAPVLIREITRRVNLEGIFQAVYTAGVVLPKPVSTCRYWHRSLNPRK
LVEVKFSHLSRNMTLQRTMKLYRLPDSTKTPGLRPMERRDIRQVTELLQKFLKRFQLAPS
MTEEEVSHWFLPQDNIIDTYVVEVAGISLKDSDPECLGAGGALTDFASFYTLPSTVMHHP
LHRSLKAAYSFYNVHTQTPLLDLMNDALILAKLKGFDVFNALDLMENKVFLEKLKFGIGD
GNLQYYLYNWKCPSMEPDKPWLPFRSASSFSSRVPQGCYTNTGPKVTTDRGSCHLDMQFA
GRGGDNSSIINILNVKSCYA

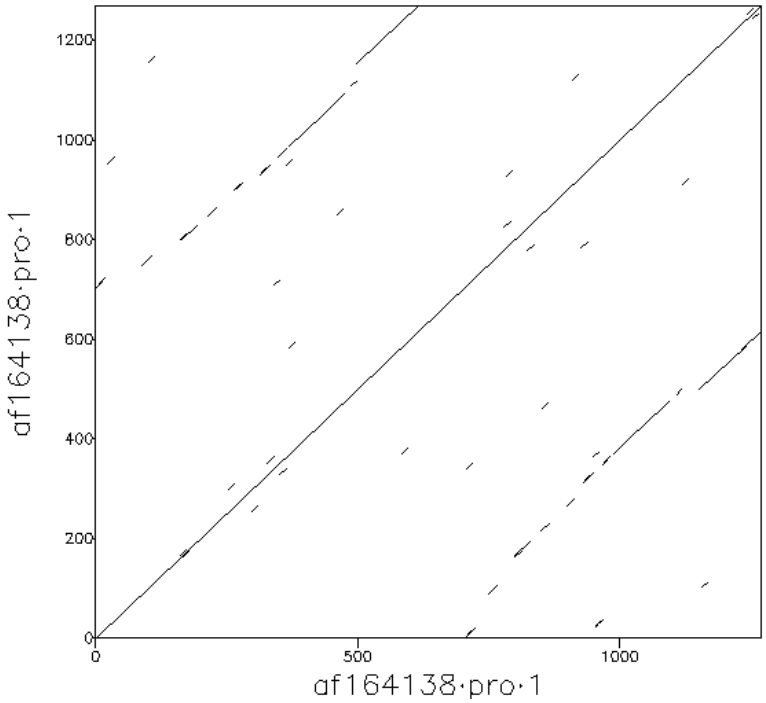
5) 用点阵图方法分析上述提取的序列中第一个基因（即多药耐药基因 MDR2）所编码的氨基酸序列，说明是否有重复结构域。

用 Dotlet 程序分析上述提取的序列中第一个基因所编码的氨基酸序列，所得点阵图如下所示：



Matrix : Blosum62
Sliding window : 15
Zoom : 5
Score range : -60 to 165
Gray scale : 49% - 31%

用 WebLab 中 DotMatcher 程序分析该序列，所得点阵图如下：

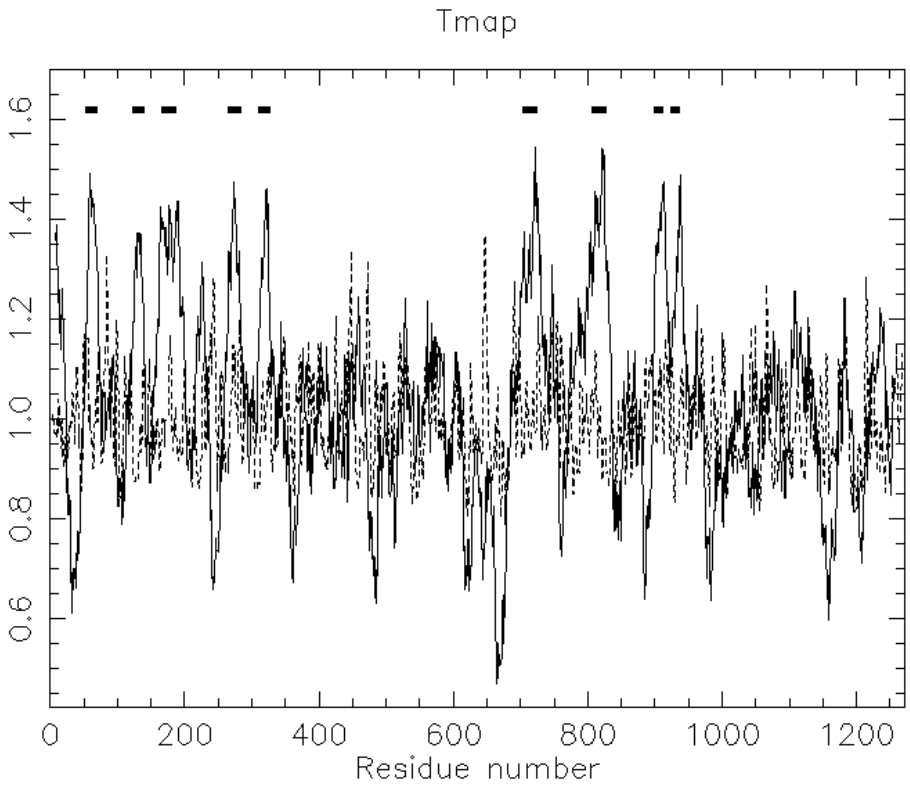


Windowsize = 10
Threshold = 23

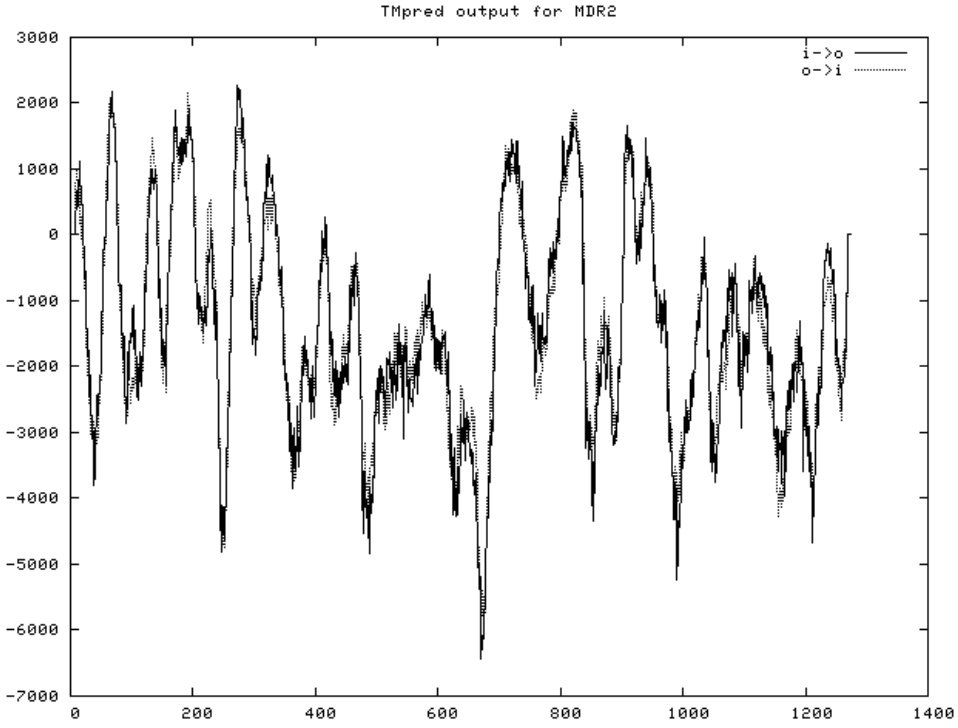
两个程序得到的点阵图大体一致，该蛋白质序列由两段重复序列首尾相连而成，该蛋白质可能存在重复结构。

6) 用多种跨膜螺旋预测程序预测上述 MDR2 蛋白质序列中可能的跨膜螺旋。

用 WebLab 中 Tmap 程序分析上述 MDR2 蛋白序列，预测到 9 段跨膜螺旋区，如下图所示：



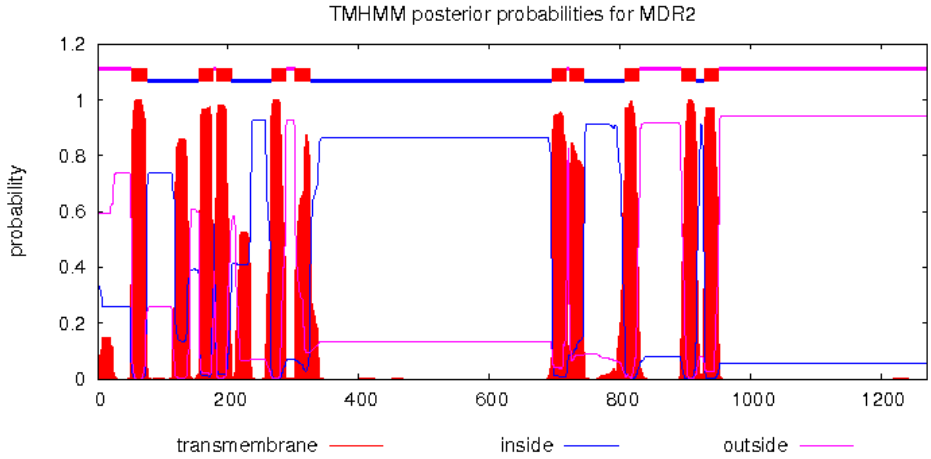
ExPASy 中的 TMpred 程序分析上述 MDR2 蛋白序列，结果如下图所示：



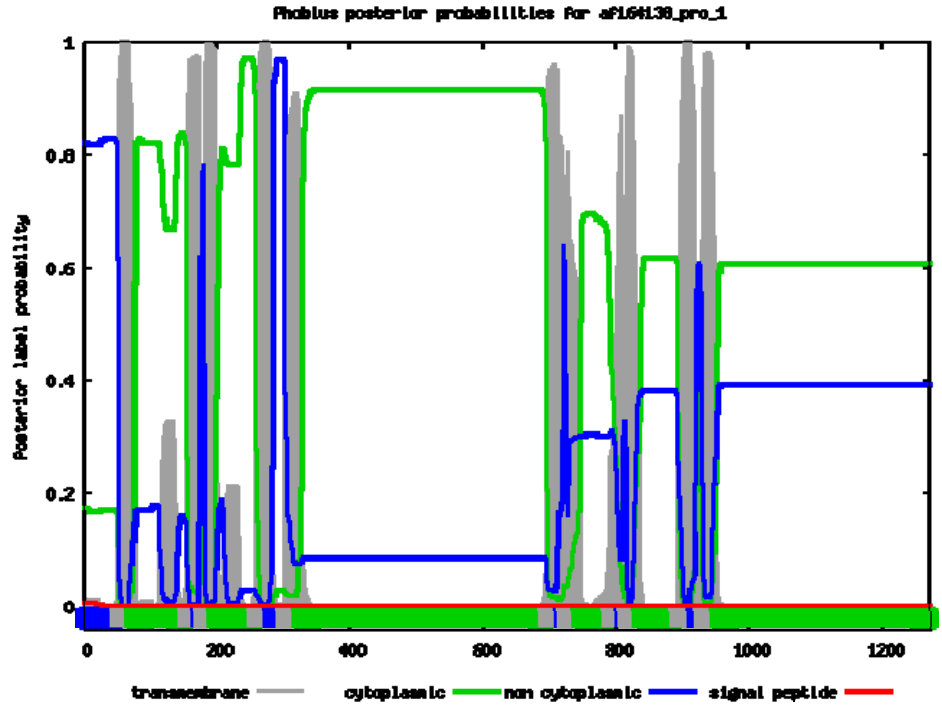
输出报告出两种跨膜螺旋模型，分别含 11 和 12 个跨膜螺旋，其中 12 个跨膜螺旋模型的详细信息如下表所示：

No	from	to	length	score	orientation
1	6	25	20	1107	i-o
2	56	73	18	1982	o-i
3	124	142	19	1007	i-o
4	161	183	23	1752	o-i
5	185	202	18	1908	i-o
6	221	238	18	528	o-i
7	263	287	25	2279	i-o
8	321	343	23	662	o-i
9	709	734	26	1441	i-o
1	6	25	20	1107	i-o
11	901	919	19	1649	i-o
12	936	953	18	1127	o-i

TMHMM 程序预测出 10 段跨膜螺旋：（53, 75）（156, 178）（183, 205）（266, 288）（303, 325）（696, 718）（723, 745）（808, 830）（895, 917）（930, 952），如下图所示：



Probius 程序预测了 11 段跨膜螺旋：（53, 75）（156, 174）（180, 198）（259, 283）（303, 327）（696, 720）（726, 745）（790, 810）（816, 836）（894, 918）（930, 952），如下图所示：



4. 数据库搜索

1) 以上述河豚鱼基因组多药耐药基因全长序列，用 BlastN 搜索人基因组核酸序列数据库，调节不同参数，比较搜索结果。

查询序列：cosmid 124A22 序列 1 bp - 14 kb 片段

Database：Refseq_rna Organism: human

Exclude Models (XM/XP) Exclude Uncultured/environmental sample sequences

Program Selection Optimize for Highly similar sequences (megablast)

Results: No significant similarity found

Program Selection Optimize for More dissimilar sequences (discontiguous megablast)

Results: 15 sequences，均为 ABCB 家族成员，如下表所示：

No	Accession	Description	Max score	Total score	Query cover	E value	Ident
1	NM_000927.4	ABCB1	149	1033	10%	2e-33	79%
2	NM_018849.2	ABCB4, transcript variant B	149	1143	10%	2e-33	79%
3	NM_000443.3	ABCB4, transcript variant A	149	1143	10%	2e-33	79%
4	NM_018850.2	ABCB4, transcript variant C	149	1143	10%	2e-33	79%
5	NM_001163941.1	ABCB5, transcript variant 1	96.9	365	4%	1e-17	71%
6	NM_178559.5	ABCB5, transcript variant 2	96.9	365	4%	1e-17	71%
7	NM_003742.2	ABCB11	96.9	559	6%	1e-17	73%
8	NM_001163942.1	ABCB5, transcript variant 3	89.7	150	2%	2e-15	72%
9	NM_001163993.2	ABCB5, transcript variant 4	87.8	148	2%	7e-15	71%
10	NM_001243014.1	ABCB9, transcript variant 5	69.8	139	1%	2e-09	75%
11	NM_001243013.1	ABCB9, transcript variant 6	69.8	139	1%	2e-09	75%
12	NM_203444.3	ABCB9, transcript variant 4	69.8	139	1%	2e-09	75%
13	NM_019624.3	ABCB9, transcript variant 2	69.8	139	1%	2e-09	75%
14	NM_019625.3	ABCB9, transcript variant 1	69.8	139	1%	2e-09	75%
15	NM_012089.2	ABCB10	41	41	0%	0.93	69%

Program Selection Optimize for Somewhat similar sequences (blastn)

Results: 77 sequences. 前 18 个高分匹配为 ABCB 家族成员，其余序列 Query cover 均在 1%以下。

2) 以上述河豚鱼基因组多药耐药基因编码区序列，用 BlastX 搜索人基因组蛋白质序列数据库，调节不同参数，比较搜索结果。

查询序列：MDR2 CDS 序列（上面 coderet 程序提取结果）

Database：Refseq_protein Organism: human

Exclude Models (XM/XP) Exclude Uncultured/environmental sample sequences

Expect threshold：0.001

Results: 84 sequences，大部分为 ABC 家族成员

Accession	Description	Max score	Total score	Query cover	E value	Ident
NP_000918.2	multidrug resistance protein 1	1439	1878	99%	0	58%
NP_000434.1	multidrug resistance protein 3 isoform A	1423	1423	99%	0e+00	57%
NP_061337.1	multidrug resistance protein 3 isoform B	1415	1415	99%	0e+00	57%
NP_061338.1	multidrug resistance protein 3 isoform C	1346	1346	99%	0e+00	56%
NP_001157413.1	ATP-binding cassette sub-family B member 5 isoform 1	1227	1639	99%	0e+00	50%
NP_003733.2	bile salt export pump	1076	1817	99%	0e+00	46%
NP_848654.3	ATP-binding cassette sub-family B member 5 isoform 2	856	1268	99%	0e+00	52%
NP_001269222.1	ATP-binding cassette sub-family B member 8,	327	585	91%	2e-99	34%

Accession	Description	Max score	Total score	Query cover	E value	Ident
	mitochondrial isoform d					
NP_009119.2	ATP-binding cassette sub-family B member 8, mitochondrial isoform b	323	580	86%	5e-97	36%
NP_036221.2	ATP-binding cassette sub-family B member 10, mitochondrial	323	610	84%	6e-97	37%
NP_001269220.1	ATP-binding cassette sub-family B member 8, mitochondrial isoform a	323	579	86%	8e-97	36%
NP_062571.1	ATP-binding cassette sub-family B member 9 isoform 1	297	551	88%	8e-88	32%
NP_001269221.1	ATP-binding cassette sub-family B member 8, mitochondrial isoform c	284	492	66%	6e-84	34%
NP_062570.1	ATP-binding cassette sub-family B member 9 isoform 2	256	496	64%	8e-74	45%
NP_001229943.1	ATP-binding cassette sub-family B member 9 isoform 5	249	445	77%	6e-72	31%
NP_982269.2	ATP-binding cassette sub-family B member 9 isoform 4	249	443	77%	1e-71	31%
NP_005680.1	ATP-binding cassette sub-family B member 6, mitochondrial	247	454	46%	5e-70	46%
NP_001229942.1	ATP-binding cassette sub-family B member 9 isoform 6	238	438	66%	9e-68	29%
NP_001278951.1	antigen peptide transporter 1 isoform 2	228	451	63%	9e-66	35%
NP_000584.2	antigen peptide transporter 1 isoform 1	228	449	63%	6e-64	35%
NP_000535.3	antigen peptide transporter 2 isoform 1	221	404	86%	5e-62	31%
NP_001276972.1	antigen peptide transporter 2 isoform 3	220	405	82%	6e-62	31%
NP_001157465.1	ATP-binding cassette sub-family B member 5 isoform 4	194	316	22%	1e-58	73%
NP_001157414.1	ATP-binding cassette sub-family B member 5 isoform 3	193	315	22%	2e-58	73%
NP_001258626.1	ATP-binding cassette sub-family B member 7, mitochondrial isoform 3	209	397	52%	3e-58	41%
NP_001258627.1	ATP-binding cassette sub-family B member 7, mitochondrial isoform 4	210	398	52%	3e-58	41%
NP_001258628.1	ATP-binding cassette sub-family B member 7, mitochondrial isoform 5	209	397	52%	3e-58	41%
NP_001258625.1	ATP-binding cassette sub-family B member 7, mitochondrial isoform 2	210	398	52%	3e-58	41%
NP_004290.2	ATP-binding cassette sub-family B member 7, mitochondrial isoform 1	210	398	52%	3e-58	41%
NP_061313.2	antigen peptide transporter 2 isoform 2	202	361	76%	3e-56	31%
NP_150229.2	multidrug resistance-associated protein 9	151	486	40%	9e-39	36%
NP_003777.2	canalicular multispecific organic anion transporter 2 isoform 1	144	458	41%	2e-36	32%
NP_149163.2	ATP-binding cassette sub-family C member 11 isoform a	144	459	41%	2e-36	36%
NP_000383.1	canalicular multispecific organic anion transporter 1	143	444	62%	2e-36	29%
NP_005679.2	multidrug resistance-associated protein 5 isoform 1	134	429	46%	2e-33	34%
NP_004987.2	multidrug resistance-associated protein 1	133	464	40%	3e-33	32%
NP_258261.2	multidrug resistance-associated protein 7 isoform MRP7A	126	401	38%	3e-31	35%
NP_001185863.1	multidrug resistance-associated protein 7 isoform MRP7	126	401	38%	4e-31	35%
NP_001162.4	multidrug resistance-associated protein 6 isoform 1	125	436	40%	7e-31	34%
NP_005836.2	multidrug resistance-associated protein 4 isoform 1	125	458	60%	9e-31	33%
NP_005682.2	ATP-binding cassette sub-family C member 9 isoform SUR2A	124	428	44%	2e-30	29%
NP_064693.2	ATP-binding cassette sub-family C member 9 isoform SUR2B	122	423	45%	6e-30	29%
NP_001098985.1	multidrug resistance-associated protein 4 isoform 2	116	213	59%	2e-28	26%
NP_000343.2	ATP-binding cassette sub-family C member 8	103	375	39%	5e-24	32%

Accession	Description	Max score	Total score	Query cover	E value	Ident
	isoform 2					
NP_001274103.1	ATP-binding cassette sub-family C member 8 isoform 1	103	374	39%	5e-24	32%
NP_660187.1	ATP-binding cassette sub-family C member 11 isoform b	103	420	41%	5e-24	31%
NP_005493.2	ATP-binding cassette sub-family A member 1	99.8	289	38%	6e-23	32%
NP_000483.3	cystic fibrosis transmembrane conductance regulator	92.8	333	39%	8e-21	31%
NP_061142.2	ATP-binding cassette sub-family A member 5	91.3	224	38%	3e-20	31%
NP_001597.2	ATP-binding cassette sub-family A member 2 isoform a	90.1	154	19%	5e-20	31%
NP_997698.1	ATP-binding cassette sub-family A member 2 isoform b	90.1	154	19%	6e-20	31%
NP_001080.2	ATP-binding cassette sub-family A member 3	88.6	232	41%	2e-19	28%
NP_061985.2	ATP-binding cassette sub-family A member 7	88.2	170	35%	2e-19	29%
NP_001275915.1	ATP-binding cassette sub-family A member 8 isoform 3	85.1	194	38%	2e-18	31%
NP_001275914.1	ATP-binding cassette sub-family A member 8 isoform 1	84.7	194	38%	2e-18	31%
NP_056472.2	ATP-binding cassette sub-family A member 12 isoform b	84.7	207	37%	2e-18	28%
NP_775099.2	ATP-binding cassette sub-family A member 12 isoform a	84.7	206	37%	3e-18	28%
NP_000341.2	retinal-specific ATP-binding cassette transporter	83.2	273	41%	8e-18	28%
NP_525022.2	ATP-binding cassette sub-family A member 9	81.6	136	19%	2e-17	28%
NP_525023.2	ATP-binding cassette sub-family A member 6	75.9	75.9	21%	1e-15	28%
NP_071882.1	ATP-binding cassette sub-family G member 8	73.9	124	23%	3e-15	27%
NP_689914.3	ATP-binding cassette sub-family A member 13	74.3	255	37%	4e-15	25%
NP_071881.1	ATP-binding cassette sub-family G member 5	72.8	72.8	16%	8e-15	26%
NP_058198.2	ATP-binding cassette sub-family G member 1 isoform 2	72.4	123	25%	8e-15	26%
NP_997510.1	ATP-binding cassette sub-family G member 1 isoform 5	72.4	123	25%	9e-15	26%
NP_997511.1	ATP-binding cassette sub-family G member 1 isoform 6	72.4	123	25%	9e-15	26%
NP_997512.1	ATP-binding cassette sub-family G member 1 isoform 7	72.4	123	25%	1e-14	26%
NP_004906.3	ATP-binding cassette sub-family G member 1 isoform 4	72.4	123	25%	1e-14	26%
NP_997057.1	ATP-binding cassette sub-family G member 1 isoform 3	72	122	25%	1e-14	26%
NP_525021.3	ATP-binding cassette sub-family A member 10	72	181	42%	2e-14	27%
NP_002849.1	ATP-binding cassette sub-family D member 3 isoform a	69.7	127	33%	6e-14	26%
NP_071452.2	ATP-binding cassette sub-family G member 4	67.8	121	25%	2e-13	26%
NP_004818.2	ATP-binding cassette sub-family G member 2 isoform 1	64.7	121	36%	2e-12	28%
NP_001244315.1	ATP-binding cassette sub-family G member 2 isoform 2	63.9	120	36%	3e-12	29%
NP_000024.2	ATP-binding cassette sub-family D member 1	61.6	121	34%	2e-11	29%
NP_005041.1	ATP-binding cassette sub-family D member 4	58.9	58.9	16%	1e-10	24%
NP_005155.1	ATP-binding cassette sub-family D member 2	57.8	114	34%	3e-10	26%
NP_009099.1	ATP-binding cassette sub-family A member 8 isoform 2	57.8	111	16%	4e-10	33%
NP_009120.1	ATP-binding cassette sub-family F member 2 isoform a	50.1	50.1	18%	6e-08	21%
NP_005683.2	ATP-binding cassette sub-family F member 2 isoform b	50.1	50.1	18%	6e-08	21%
NP_001081.1	ATP-binding cassette sub-family F member 1 isoform b	50.1	179	12%	8e-08	41%
NP_001020262.1	ATP-binding cassette sub-family F member 1 isoform a	50.1	179	12%	8e-08	41%
NP_002931.2	ATP-binding cassette sub-family E member 1	48.5	48.5	13%	2e-07	23%
NP_060828.2	ATP-binding cassette sub-family F member 3	45.4	45.4	5%	2e-06	37%

3) 以上述河豚鱼基因组多药耐药基因氨基酸序列，用 BlastP 搜索人基因组蛋白质序列数据库，调节不同参数，比较搜索结果。

查询序列：MDR2 protein 序列（上面 coderet 程序提取结果）
Database：Refseq_protein Organism: human
Exclude Models (XM/XP) Exclude Uncultured/environmental sample sequences
Expect threshold：0.001

Program Selection Algorithm blastp (protein-protein BLAST)
Results: 84 sequences，与上述 BlastX 结果完全一致。

Program Selection Algorithm PSI-BLAST (Position-Specific Iterated BLAST)
Iteration 1 Results: 84 sequences，与上述 BlastX 结果完全一致；
Iteration 2 Results: 88 sequences，在 iteration 1 基础上增加 4 个序列

Accession	Description	Max score	Total score	Query cover	E value	Ident
NP_001137542.1	canalicular multispecific organic anion transporter 2 isoform 2	68	68	14%	3e-11	17%
NP_005436.1	structural maintenance of chromosomes protein 3	50.6	99.7	11%	1e-05	29%
NP_005723.2	DNA repair protein RAD50	47.1	94.3	16%	1e-04	30%
NP_006435.2	structural maintenance of chromosomes protein 2	46.4	46.4	4%	2e-04	28%

Iteration 3 Results: 91 sequences，在 iteration 2 基础上增加 3 个序列

Accession	Description	Max score	Total score	Query cover	E value	Ident
NP_001275682.1	structural maintenance of chromosomes protein 4 isoform 2	51.1	100	18%	8e-06	15%
NP_001002800.1	structural maintenance of chromosomes protein 4 isoform 1	51.1	100	18%	8e-06	15%
NP_683515.4	structural maintenance of chromosomes protein 1B isoform 1	47.6	47.6	5%	8e-05	27%

Iteration 4 Results: 93 sequences，在 iteration 3 基础上增加 2 个序列

Accession	Description	Max score	Total score	Query cover	E value	Ident
NP_001268392.1	structural maintenance of chromosomes protein 1A isoform 2	59.5	107	11%	2e-08	24%
NP_006297.2	structural maintenance of chromosomes protein 1A isoform 1	59.5	107	11%	2e-08	24%

Iteration 5 Results: 没有增加新的序列

Program Selection Algorithm DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Results: 93 sequences，与 PSI-BLAST 经过 4 次 iteration 后所得序列一致。

4) 分析上述搜索策略，说明 Blast 数据库搜索工具的正确使用方法，举例说明不同参数对结果的影响。

以目标基因全长序列作为查询序列，由于可能存在较长非编码区域，非编码区保守性较低，因此难以准确搜索到匹配序列；用编码区序列或氨基酸序列保守性较高，更有利于搜索到同源序列。

用 BlastN 搜索时可选择不同的匹配程度，当查询序列保守性较差时可适当降低匹配的精

确程度，以尽可能搜索到同源序列。

用 BlastP 搜索时可根据搜索目的选择不同算法，blastp 根据氨基酸序列相似性搜索同源序列，容易漏掉序列差异较大的远缘同源序列；PSI-BLAST 根据已得结果构建位点特异性计分矩阵，通过跌代匹配可以搜索到序列差异较大但符具有特定保守位点的远缘同源序列，通常需进行多次迭代以尽可能搜索到所有同源序列；DELTA-BLAST 采用 domain 增强搜索方式，根据查询序列 domain 搜索具有相似 domain 的序列，可较高效地找到近缘、远缘同源序列。

5. 归纳总结

1) 通过以上河豚鱼柯氏质粒基因组序列片段分析过程，说明基因组序列分析的一般方法，所用软件种类和基本原理，以及不同软件的适用范围和优缺点。

A) 点阵图

通过目标基因组序列的点阵图（横纵坐标均为目标基因组序列）可以考察目标基因组序列是否含有重复片段，重复片段通常为具有重要功能的基因或者结构域所在区域，由此可粗略推测基因/结构域的大小、位置等信息以作进一步分析。如果已知目标基因组序列所含基因的功能或者有已知的可能的同源基因，可以通过目标基因组序列对已知的可能的同源基因序列作点阵图，可初步确认是否具有同源性，以便进行进一步分析。

绘制点阵图的工具很多，如 WebLab 中 Dottup、DotMatcher、DotPath 和 ExPASy 中的 Dotlet 等程序。Dottup 是精确匹配，两个序列比对中，word size 内精确匹配时以图上的点表示，从点阵图则可直观看到两个序列的相似程度及匹配范围，当两个序列匹配程度较高时适用，如果两个序列差异较大则不适用；会重复匹配，存在符合 word size 的重复序列时会予以显示，当定义的 word size 较小时往往会在对角线外出现较多短线。DotMatcher 是近似匹配，用给定的计分矩阵对 windowsize 内的序列进行打分，高于 threshold 的在图上以点显示，可用于匹配程度不高的两个序列；也存在重复匹配。DotPath 与 Dottup 相似，是 word size 内的精确匹配，但不进行重复匹配（non-overlapping wordmatch）。三种程序均有相应的参数设置，需根据具体情况及实验目的进行设置，用 RNA 序列进行匹配时，相对于蛋白序列而言 word size 或 windowsize 应适当提高。Dotlet 与 DotMatcher 较为相似，其优点在于参数调整较为方便直观，可以通过滑动条灵活调整显示的得分范围及显示的方式，并予以即时显示，可更高效地得到最优结果。

B) 基因预测

通过基因预测程序对目标基因组序列进行分析，可以得到目标基因组序列可能存在的基因及其基因结构、剪接方式、编码区序列、编码蛋白序列等信息，可用以进一步分析。基因预测程序所得结果不一定可靠，尤其是目标基因组序列所属物种在所用基因预测程序中没有相近的参考物种时预测结果可能偏差较大。而大部分基因预测程序所用的参考物种种类有限，局限性较大。基因预测结果

前面用到的基因预测的程序有 SoftBerry、AUGUSTUS、GenScan、GenID、HMMGene 等，上述程序的操作方式、参数设置、结果显示方式相差不多，对上述河豚鱼柯氏质粒基因组序列片段的分析结果大同小异。其中 Softberry 网站上的 FGENESH 工具所用参考物种种类较多，有利于得到较可靠的预测结果；预测结果既有列表也有图形标示，较为清晰直观。

C) 跨膜螺旋预测

通过预测目标基因蛋白序列是否含有跨膜螺旋，可以推测目标蛋白是否为膜蛋白以及其可能具有的功能等。预测结果不一定可靠，只能作为参考。

前面用到的跨膜螺旋预测程序包括 Tmap、ProtScale、TMpred、TMHMM、Phobius、DAS-TMfilter、HMMTOP 等。各跨膜螺旋预测程序所得结果大同小异，前面对 PPF1_PEA 及河豚鱼 MDR2 跨膜螺旋的预测过程发现 HMMTOP 程序与其他程序预测结果偏差较大。在各程序中 TMHMM 和 Phobius 的显示方式较为直观，DAS-TMfilter 给出了螺旋的具体位置和预测得分，比较清晰明了。

D) Blast

当已初步确定基因编码区域或已预测到可能的编码区序列或者蛋白序列，可通过 Blast 搜索目标基因的同源序列。以目标基因全长序列作为查询序列，由于可能存在较长非编码区域，非编码区保守性较低，因此难以准确搜索到匹配序列；用编码区序列或氨基酸序列保守性较高，更有利于搜索到同源序列。

课程学习中用到的是 NCBI 中的 Blast 程序，可选择多种搜索模式：BlastN 用核酸序列搜索核酸数据库；BlastP 用蛋白序列搜索蛋白数据库；BlastX 用核酸序列以所有六种读码框形式翻译成的蛋白序列搜索蛋白数据库，可用以搜索查询核酸序列所有可能基因产物的相似序列；tBlastN 用蛋白序列搜索将核酸数据库按六种读码框形式翻译成的蛋白序列库，当所用核酸数据库没有注释时可用；tBlastX 用核酸序列以所有六种读码框形式翻译成的蛋白序列搜索将核酸数据库按六种读码框形式翻译成的蛋白序列库，可最大限度避免注释错误造成的遗漏，也可用于发现新基因。

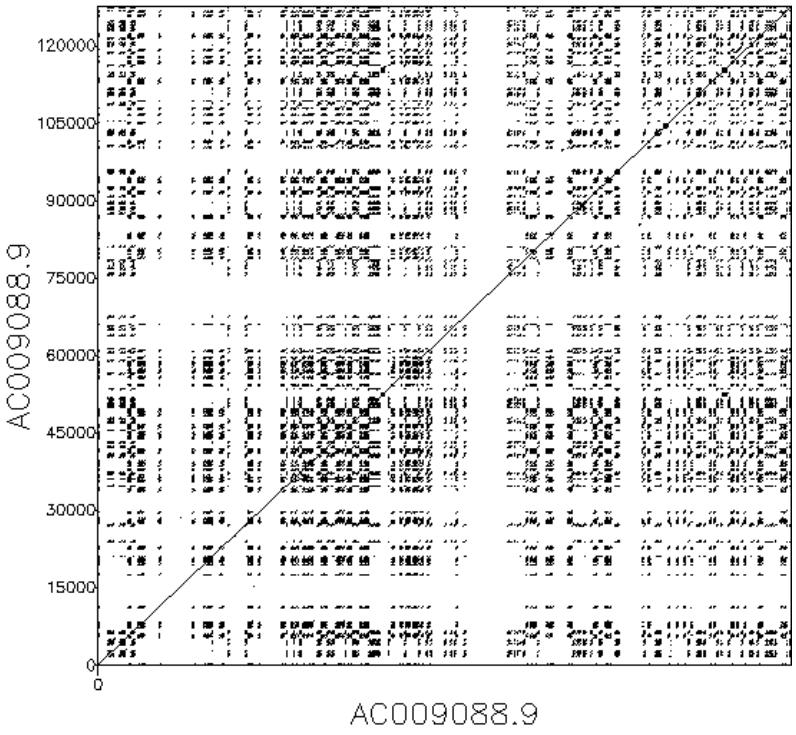
各种搜索模式下都有多种参数设置，可根据具体情况进行设置以高效准确地得到搜索结果。用 BlastN 搜索时可选择不同的匹配程度，当查询序列保守性较差时可适当降低匹配的精确程度，以尽可能搜索到同源序列。用 BlastP 搜索时可根据搜索目的选择不同算法，blastp 根据氨基酸序列相似性搜索同源序列，容易漏掉序列差异较大的远缘同源序列；PSI-BLAST 根据已得结果构建位点特异性计分矩阵，通过跌代匹配可以搜索到序列差异较大但符具有特定保守位点的远缘同源序列，通常需进行多次迭代以尽可能搜索到所有同源序列；DELTA-BLAST 采用 domain 增强搜索方式，根据查询序列 domain 搜索具有相似 domain 的序列，可较高效地找到近缘、远缘同源序列。

2) 将以上基因组序列片段分析方法用于与你研究课题相关基因组序列分析，说明分析结果。

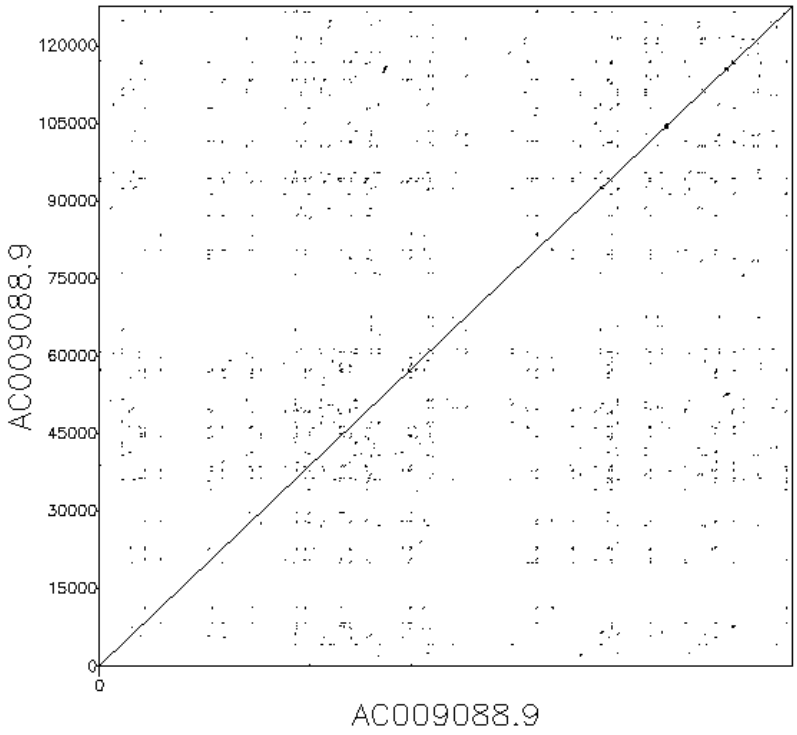
实验室主要关注的蛋白质是 MG53（Mitsugumin-53），也称为 TRIM72（Tripartite motif-containing protein 72）。已有完整的基因组注释，人 TRIM72 基因位于 16 号染色体 Chromosome 16: 31,225,342-31,236,510 forward strand，Ensembl 登录号为 ENSG00000177238。基因组序列长 11.17 kb，含 7 个 exon，其中 6 个含编码信息，转录本长度为 2156 bp，编码 477 aa 蛋白质。在 UniProt 中 TRI72_HUMAN 条目（Q6ZMU5）数据库交叉链接中可以获得其所在基因组片段序列：Homo sapiens chromosome 16 clone RP11-388M20, complete sequence（GenBank: AC009088.9），长度为 127769 bp（127.8 kb）。下面对该基因组片段进行分析：

(A) 点阵图

用 DotMatcher 对该基因组片段序列作自身的点阵图，如下图所示：



Matrix file: Ednafull
Windowsize = 60
Threshold = 150

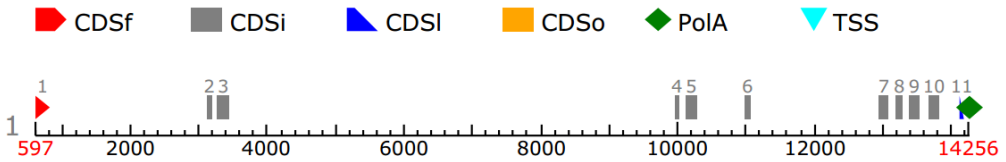


Matrix file: Ednafull
Windowsize = 65
Threshold = 280

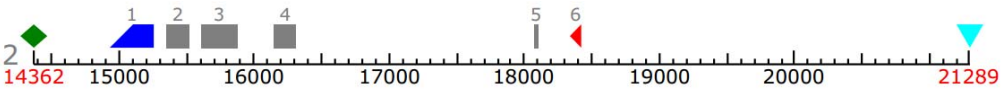
看不出明显的重复片段，一方面可能该片段中没有大范围的重复片段，另一方面由于序列长度过大，除非是较大范围的序列相似程度较高的重复片段才能够在点阵图中看出，绘制点阵图应选择合适长度的序列。另外，可以看到点阵图呈现有趣的 mosaic pattern，这可能是基因组中 non-coding 区域的重复所形成的。

(B) 基因预测

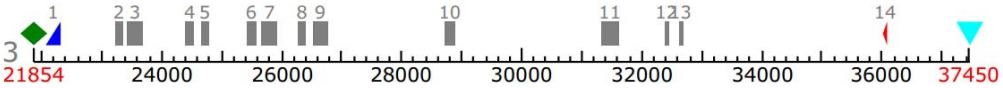
用 Softberry 网站上的 FGENESH 工具分析上述基因组片段，Organism 选择 Human，分析结果如下所示：



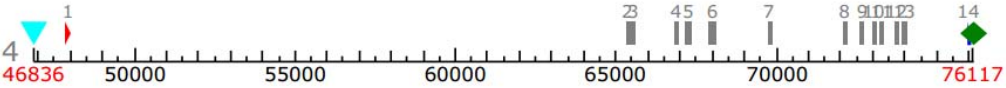
1	+	1	CDSf	597	-	807	11.12	597	-	806	210
1	+	2	CDSi	3101	-	3175	11.33	3103	-	3174	72
1	+	3	CDSi	3254	-	3429	19.86	3256	-	3429	174
1	+	4	CDSi	9950	-	10003	12.41	9950	-	10003	54
1	+	5	CDSi	10104	-	10268	21.25	10104	-	10268	165
1	+	6	CDSi	10965	-	11054	15.14	10965	-	11054	90
1	+	7	CDSi	12932	-	13072	11.66	12932	-	13072	141
1	+	8	CDSi	13182	-	13275	5.28	13182	-	13274	93
1	+	9	CDSi	13371	-	13521	19.12	13373	-	13519	147
1	+	10	CDSi	13661	-	13815	14.71	13662	-	13814	153
1	+	11	CDSl	14110	-	14174	2.73	14112	-	14174	63
1	+		PolA	14256			1.12				



2	-		PolA	14362			-1.07				
2	-	1	CDSl	14924	-	15250	8.02	14924	-	15250	327
2	-	2	CDSi	15344	-	15510	21.90	15344	-	15508	165
2	-	3	CDSi	15597	-	15868	19.03	15598	-	15867	270
2	-	4	CDSi	16141	-	16303	3.41	16143	-	16301	159
2	-	5	CDSi	18069	-	18086	6.43	18070	-	18084	15
2	-	6	CDSf	18329	-	18413	25.31	18330	-	18413	84
2	-		TSS	21289			-5.09				



3	-		PolA	21854			1.12				
3	-	1	CDSl	22053	-	22301	9.33	22053	-	22301	249
3	-	2	CDSi	23209	-	23330	6.27	23209	-	23328	120
3	-	3	CDSi	23407	-	23672	0.13	23408	-	23671	264
3	-	4	CDSi	24384	-	24525	7.25	24386	-	24523	138
3	-	5	CDSi	24638	-	24764	-0.52	24639	-	24764	126
3	-	6	CDSi	25396	-	25559	11.02	25396	-	25557	162
3	-	7	CDSi	25652	-	25908	11.73	25653	-	25907	255
3	-	8	CDSi	26257	-	26386	-7.94	26259	-	26384	126
3	-	9	CDSi	26503	-	26752	10.47	26504	-	26752	249
3	-	10	CDSi	28704	-	28870	9.59	28704	-	28868	165
3	-	11	CDSi	31310	-	31595	16.42	31311	-	31595	285
3	-	12	CDSi	32378	-	32413	0.26	32378	-	32413	36
3	-	13	CDSi	32611	-	32646	2.60	32611	-	32646	36
3	-	14	CDSf	36001	-	36072	6.88	36001	-	36072	72
3	-		TSS	37450			-4.89				



4	+		TSS	46836			-7.49				
4	+	1	CDSf	47805	-	47988	6.91	47805	-	47987	183
4	+	2	CDSi	65313	-	65337	5.47	65315	-	65335	21
4	+	3	CDSi	65428	-	65579	11.33	65429	-	65578	150
4	+	4	CDSi	66773	-	66917	13.02	66775	-	66915	141
4	+	5	CDSi	67124	-	67311	19.64	67125	-	67310	186
4	+	6	CDSi	67854	-	68094	21.92	67856	-	68092	237
4	+	7	CDSi	69717	-	69751	4.81	69718	-	69750	33
4	+	8	CDSi	72038	-	72141	2.61	72040	-	72141	102
4	+	9	CDSi	72580	-	72709	11.53	72580	-	72708	129
4	+	10	CDSi	72955	-	73056	11.84	72957	-	73055	99
4	+	11	CDSi	73190	-	73313	-1.17	73192	-	73311	120
4	+	12	CDSi	73657	-	73757	8.12	73658	-	73756	99
4	+	13	CDSi	73878	-	74025	20.34	73880	-	74023	144
4	+	14	CDSl	75922	-	75985	-2.54	75923	-	75985	63
4	+		PolA	76117			1.12				


```
5 -      PolA     83860                 -1.68
5 -    1 CDSl     84500 -     84756     20.94     84500 -     84754     255
5 -    2 CDSi     85062 -     85118      1.16     85063 -     85116      54
5 -    3 CDSf     85332 -     85605     38.34     85333 -     85605     273
5 -      TSS      86180                 -9.49

6 +      TSS      95728                -10.59
6 +    1 CDSo     97654 -     98118     35.92     97654 -     98118     465
6 +      PolA     98913                  1.12

7 -      PolA     99398                 -5.47
7 -    1 CDSo     99674 -     99943     42.72     99674 -     99943     270
7 -      TSS     100909                -10.09

8 +      TSS     101787                 -8.59
8 +    1 CDSf    101895 -    101954      0.16    101895 -    101954      60
8 +    2 CDSi    102010 -    102105      9.54    102010 -    102105      96
8 +    3 CDSi    102204 -    102434     23.90    102204 -    102434     231
8 +    4 CDSi    103811 -    103833      5.43    103811 -    103831      21
8 +    5 CDSi    105742 -    105860     11.78    105743 -    105859     117
8 +    6 CDSl    107096 -    107670     51.09    107098 -    107670     573
8 +      PolA    108106                -1.68
```

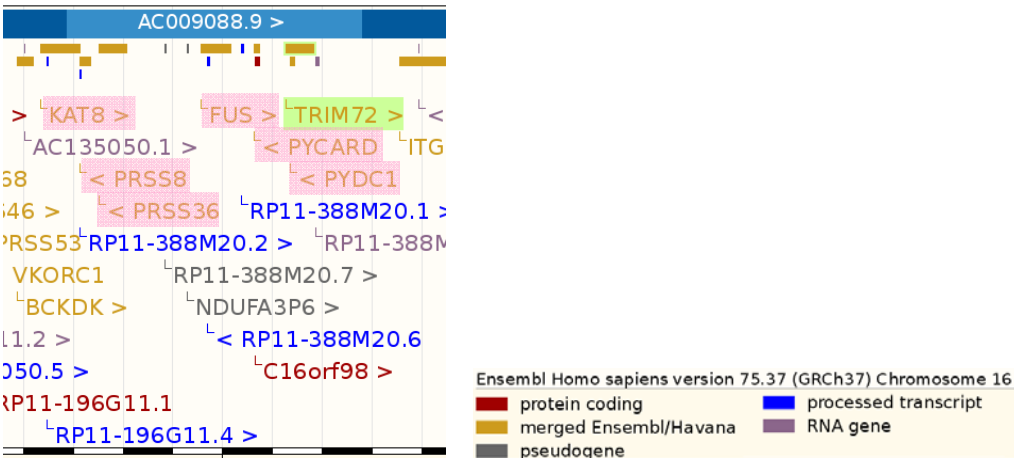
预测结果汇总如下：

No	Exon	Start	End	mRNA (bp)	Protein (aa)	Chain
FGENESH_1	11	597	14174	1377	458	+
FGENESH_2	6	14924	18413	1032	343	-
FGENESH_3	14	22053	36072	2304	767	-
FGENESH_4	14	47805	75985	1743	580	+
FGENESH_5	3	84500	85605	588	195	-
FGENESH_6	1	97654	98118	465	154	+
FGENESH_7	1	99674	99943	270	89	-
FGENESH_8	6	101895	107670	1104	367	+

对预测得到的蛋白质序列分别进行 blastp，Organism 选择 human，挑选最佳匹配结果，如下表所示：

No	Accession	Name	Total score	Query cover	E value	Ident	Length (aa)
FGENESH_1	NP_115564.2	KAT8	955	100%	0	100%	458
FGENESH_2	NP_002764.1	PRSS8	693	100%	0	100%	343
FGENESH_3	EAW52154.1	PRSS36	1407	93%	0	95%	855
FGENESH_4	NP_004951.1	FUS	907	86%	0	97%	526
FGENESH_5	NP_037390.2	PYCARD	398	100%	2e-141	100%	195
FGENESH_6	NP_001008275.2	TRIM72	266	87%	3e-86	97%	477
FGENESH_7	NP_690865.1	PYDC1	176	100%	5e-57	100%	89
FGENESH_8	NP_001008275.2	TRIM72	698	94%	0	99%	477

分析结果总没有 TRIM72 的完整序列，但第 6 和第 8 个预测序列包含 TRIM72 的部分序列，中间被第 7 个序列——反义链编码的 PYDC1 间隔开，这可能是该预测程序的局限性。根据 Ensembl 上人 TRIM72 基因邻近的基因组注释信息，如下图所示，可见 FGENESH 预测的结果总体而言较为准确。



(C) Blast

以上述基因组片段，用 BlastN 搜索人基因组核酸序列数据库：

查询序列：Homo sapiens chromosome 16 clone RP11-388M20, complete sequence（GenBank: AC009088.9），127769 bp

Database：Refseq_rna　　Organism: human

Exclude Models (XM/XP) Exclude Uncultured/environmental sample sequences

Program Selection Optimize for Highly similar sequences (megablast)

Results：100 sequences（Max target sequences），如下表所示：

No	Accession	Description	Max score	Total score	Query cover	E value	Ident
1	NM_001170634.1	FUS RNA binding protein (FUS), transcript variant 3, mRNA	6420	13798	7%	0	100%
2	NM_001170937.1	FUS RNA binding protein (FUS), transcript variant 4, mRNA	6420	13779	7%	0	100%
3	NR_028388.2	FUS RNA binding protein (FUS), transcript variant 2, long non-coding RNA	6420	13734	7%	0	100%
4	NM_004960.3	FUS RNA binding protein (FUS), transcript variant 1, mRNA	6420	13801	7%	0	100%
5	NM_001008274.3	tripartite motif containing 72, E3 ubiquitin protein ligase (TRIM72), mRNA	1869	8909	5%	0	100%
6	NR_102400.1	PYCARD opposite strand (PYCARDOS), long non-coding RNA	1836	2028	0%	0	100%
7	NM_002773.3	protease, serine, 8 (PRSS8), mRNA	1674	3586	1%	0	100%
8	NM_152384.2	Bardet-Biedl syndrome 5 (BBS5), mRNA	1500	7692	1%	0	90%
9	NM_001127395.1	methyltransferase like 21A (METTL21A), transcript variant 2, mRNA	1454	26340	13%	0	87%
10	NM_145280.4	methyltransferase like 21A (METTL21A), transcript variant 1, mRNA	1454	26340	13%	0	87%
11	NR_024425.1	PTGES2 antisense RNA 1 (head to head) (PTGES2-AS1), long non-coding RNA	1282	1282	0%	0	90%
12	NR_040023.1	endogenous retrovirus group K13, member 1 (ERVK13-1), long non-coding RNA	1225	6400	6%	0	89%
13	NM_182958.2	K(lysine) acetyltransferase 8 (KAT8), transcript variant 2, mRNA	1197	3399	1%	0	99%
14	NM_001007025.1	golgi SNAP receptor complex member 1 (GOSR1), transcript variant 2, mRNA	1155	5260	2%	0	90%

No	Accession	Description	Max score	Total score	Query cover	E value	Ident
15	NM_001007024.1	golgi SNAP receptor complex member 1 (GOSR1), transcript variant 3, mRNA	1155	5260	2%	0	90%
16	NM_004871.2	golgi SNAP receptor complex member 1 (GOSR1), transcript variant 1, mRNA	1155	5260	2%	0	90%
17	NM_001202519.1	CASP8 and FADD-like apoptosis regulator (CFLAR), transcript variant 8, mRNA	1064	28550	17%	0	88%
18	NM_001202518.1	CASP8 and FADD-like apoptosis regulator (CFLAR), transcript variant 7, mRNA	1064	28550	17%	0	88%
19	NM_001202517.1	CASP8 and FADD-like apoptosis regulator (CFLAR), transcript variant 6, mRNA	1064	28550	17%	0	88%
20	NM_001202516.1	CASP8 and FADD-like apoptosis regulator (CFLAR), transcript variant 5, mRNA	1064	28550	17%	0	88%
21	NM_001202515.1	CASP8 and FADD-like apoptosis regulator (CFLAR), transcript variant 4, mRNA	1064	28550	17%	0	88%
22	NM_001127183.2	CASP8 and FADD-like apoptosis regulator (CFLAR), transcript variant 2, mRNA	1064	28550	17%	0	88%
23	NM_003879.5	CASP8 and FADD-like apoptosis regulator (CFLAR), transcript variant 1, mRNA	1064	28550	17%	0	88%
24	NR_033959.1	SMG1 pseudogene 7 (SMG1P7), non-coding RNA	970	1766	0%	0	94%
25	NM_001258291.1	protease, serine, 36 (PRSS36), transcript variant 3, mRNA	913	4730	2%	0	100%
26	NM_001258290.1	protease, serine, 36 (PRSS36), transcript variant 2, mRNA	911	5277	2%	0	100%
27	NM_173502.4	protease, serine, 36 (PRSS36), transcript variant 1, mRNA	911	5303	2%	0	100%
28	NR_036504.1	uncharacterized LOC728752 (LOC728752), long non-coding RNA	835	1102	0%	0	93%
29	NR_027995.1	ankyrin repeat domain 20 family, member A9, pseudogene (ANKRD20A9P), non-coding RNA	732	38941	19%	0	82%
30	NM_145182.2	PYD and CARD domain containing (PYCARD), transcript variant 2, mRNA	667	1329	0%	0	100%
31	NM_013258.4	PYD and CARD domain containing (PYCARD), transcript variant 1, mRNA	667	1443	0%	0	100%
32	NR_109996.1	PRKAR2A antisense RNA 1 (PRKAR2A-AS1), transcript variant 1, long non-coding RNA	649	32544	18%	0	80%
33	NR_109997.1	PRKAR2A antisense RNA 1 (PRKAR2A-AS1), transcript variant 2, long non-coding RNA	649	32544	18%	0	80%
34	NM_152901.2	PYD (pyrin domain) containing 1 (PYDC1), mRNA	616	980	0%	6e-173	100%
35	NM_001168335.1	malic enzyme 2, NAD(+)-dependent, mitochondrial (ME2), transcript variant 2, mRNA	610	23226	14%	3e-171	86%
36	NM_002396.4	malic enzyme 2, NAD(+)-dependent, mitochondrial (ME2), transcript variant 1, mRNA	610	23226	14%	3e-171	86%
37	NM_000554.4	cone-rod homeobox (CRX), mRNA	590	30602	15%	4e-165	84%
38	NM_001105570.1	nudix (nucleoside diphosphate linked moiety X)-type motif 19 (NUDT19), mRNA	590	17289	12%	4e-165	85%
39	NM_033064.4	ataxia, cerebellar, Cayman type (ATCAY), mRNA	588	22766	15%	1e-164	84%
40	NM_207396.2	ring finger protein 207 (RNF207), mRNA	586	17788	13%	5e-164	84%
41	NM_004230.3	sphingosine-1-phosphate receptor 2 (S1PR2), mRNA	579	24996	16%	8e-162	84%
42	NM_052928.2	SET and MYND domain containing 4 (SMYD4), mRNA	573	24288	15%	4e-160	84%
43	NM_175066.3	DEAD (Asp-Glu-Ala-Asp) box polypeptide 51 (DDX51), mRNA	568	22355	16%	2e-158	85%
44	NM_000761.4	cytochrome P450, family 1, subfamily A, polypeptide 2 (CYP1A2), mRNA	564	17604	13%	2e-157	84%
45	NM_172209.2	TAP binding protein (tapasin) (TAPBP), transcript variant 3, mRNA	560	29403	15%	3e-156	84%
46	NM_003190.4	TAP binding protein (tapasin) (TAPBP), transcript variant 1, mRNA	560	29403	15%	3e-156	84%
47	NR_110974.1	zinc finger protein 154 (ZNF154), transcript variant 2, long non-coding RNA	556	28008	17%	4e-155	84%
48	NM_001085384.2	zinc finger protein 154 (ZNF154), transcript variant 1, mRNA	556	28008	17%	4e-155	84%
49	NM_001145078.1	zinc finger protein 805 (ZNF805), transcript variant 2, mRNA	547	24022	15%	2e-152	84%
50	NM_001023563.3	zinc finger protein 805 (ZNF805), transcript variant 1, mRNA	547	24022	15%	2e-152	84%
51	NM_017556.2	filamin binding LIM protein 1 (FBLIM1), transcript variant 1, mRNA	544	22997	15%	3e-151	84%
52	NM_001024216.1	filamin binding LIM protein 1 (FBLIM1), transcript variant 3, mRNA	544	22997	15%	3e-151	84%
53	NM_022347.3	torsin A interacting protein 2 (TOR1AIP2), transcript variant 1, mRNA	542	25478	16%	1e-150	83%
54	NM_001159508.1	isovaleryl-CoA dehydrogenase (IVD), transcript	536	27411	15%	5e-149	78%

No	Accession	Description	Max score	Total score	Query cover	E value	Ident
		variant 2, mRNA					
55	NM_002225.3	isovaleryl-CoA dehydrogenase (IVD), transcript variant 1, mRNA	536	27411	15%	5e-149	78%
56	NR_002612.1	deleted in lymphocytic leukemia 2 (non-protein coding) (DLEU2), long non-coding RNA	536	15170	10%	5e-149	85%
57	NM_001198536.1	Mediterranean fever (MEFV), transcript variant 2, mRNA	531	16296	13%	2e-147	83%
58	NM_000243.2	Mediterranean fever (MEFV), transcript variant 1, mRNA	531	16296	13%	2e-147	83%
59	NR_027412.1	long intergenic non-protein coding RNA 910 (LINC00910), transcript variant 2, long non-coding RNA	529	9206	9%	9e-147	83%
60	NR_109840.1	transmembrane protein 168 (TMEM168), transcript variant 3, long non-coding RNA	527	22705	15%	3e-146	83%
61	NM_022484.5	transmembrane protein 168 (TMEM168), transcript variant 2, mRNA	527	22705	15%	3e-146	83%
62	NM_001287497.1	transmembrane protein 168 (TMEM168), transcript variant 1, mRNA	527	22705	15%	3e-146	83%
63	NM_032779.3	coiled-coil domain containing 142 (CCDC142), mRNA	525	25193	14%	1e-145	83%
64	NM_001243650.1	glucose-fructose oxidoreductase domain containing 2 (GFOD2), transcript variant 3, mRNA	523	29033	16%	4e-145	82%
65	NM_001184819.1	guanine nucleotide binding protein-like 3 (nucleolar)-like (GNL3L), transcript variant 1, mRNA	520	30697	17%	5e-144	82%
66	NM_019067.5	guanine nucleotide binding protein-like 3 (nucleolar)-like (GNL3L), transcript variant 2, mRNA	520	30697	17%	5e-144	82%
67	NM_173079.2	RUN domain containing 1 (RUNDC1), mRNA	520	22311	14%	5e-144	81%
68	NM_001080435.2	WAS protein homolog associated with actin, golgi membranes and microtubules (WHAMM), mRNA	518	10869	9%	2e-143	84%
69	NR_046702.1	PRICKLE2 antisense RNA 3 (PRICKLE2-AS3), long non-coding RNA	518	20688	14%	2e-143	82%
70	NM_177538.2	cytochrome P450, family 20, subfamily A, polypeptide 1 (CYP20A1), mRNA	518	76373	24%	2e-143	82%
71	NM_018390.3	phosphatidylinositol-specific phospholipase C, X domain containing 1 (PLCXD1), transcript variant 1, mRNA	516	24479	15%	7e-143	83%
72	NM_001282670.1	chromosome 1 open reading frame 86 (C1orf86), transcript variant 6, mRNA	514	15523	9%	2e-142	82%
73	NM_001242826.1	ring finger protein 41, E3 ubiquitin protein ligase (RNF41), transcript variant 4, mRNA	514	20876	13%	2e-142	83%
74	NR_040053.1	ring finger protein 41, E3 ubiquitin protein ligase (RNF41), transcript variant 5, long non-coding RNA	514	20876	13%	2e-142	83%
75	NM_194359.2	ring finger protein 41, E3 ubiquitin protein ligase (RNF41), transcript variant 3, mRNA	514	20876	13%	2e-142	83%
76	NM_194358.2	ring finger protein 41, E3 ubiquitin protein ligase (RNF41), transcript variant 2, mRNA	514	20876	13%	2e-142	83%
77	NM_005785.3	ring finger protein 41, E3 ubiquitin protein ligase (RNF41), transcript variant 1, mRNA	514	20876	13%	2e-142	83%
78	NM_001205255.1	occludin (OCLN), transcript variant 3, mRNA	514	16335	12%	2e-142	82%
79	NM_001205254.1	occludin (OCLN), transcript variant 2, mRNA	514	16335	12%	2e-142	82%
80	NM_002538.3	occludin (OCLN), transcript variant 1, mRNA	514	16335	12%	2e-142	82%
81	NR_110695.1	uncharacterized LOC101928372 (LOC101928372), long non-coding RNA	510	2753	1%	3e-141	89%
82	NR_110005.1	low density lipoprotein receptor-related protein associated protein 1 (LRPAP1), transcript variant 2, long non-coding RNA	510	32709	17%	3e-141	82%
83	NM_002337.3	low density lipoprotein receptor-related protein associated protein 1 (LRPAP1), transcript variant 1, mRNA	510	32709	17%	3e-141	82%
84	NM_024345.4	DDB1 and CUL4 associated factor 10 (DCAF10), transcript variant 1, mRNA	508	21318	13%	1e-140	82%
85	NM_001286810.1	DDB1 and CUL4 associated factor 10 (DCAF10), transcript variant 2, mRNA	508	21318	13%	1e-140	82%
86	NR_121565.1	uncharacterized LOC101927954 (LOC101927954), long non-coding RNA	508	11839	9%	1e-140	83%
87	NM_015704.2	desumoylating isopeptidase 1 (DESI1), mRNA	507	26602	15%	4e-140	82%
88	NM_080706.3	transient receptor potential cation channel, subfamily V, member 1 (TRPV1), transcript variant 3, mRNA	505	16452	11%	1e-139	82%
89	NM_080705.3	transient receptor potential cation channel, subfamily V, member 1 (TRPV1), transcript variant 4, mRNA	505	16452	11%	1e-139	82%
90	NM_080704.3	transient receptor potential cation channel, subfamily V, member 1 (TRPV1), transcript variant 1, mRNA	505	16452	11%	1e-139	82%
91	NM_018727.5	transient receptor potential cation channel, subfamily	505	16452	11%	1e-139	82%

No	Accession	Description	Max score	Total score	Query cover	E value	Ident
		V, member 1 (TRPV1), transcript variant 2, mRNA					
92	NM_173827.2	COX18 cytochrome C oxidase assembly factor (COX18), mRNA	503	22502	14%	5e-139	83%
93	NM_147780.2	cathepsin B (CTSB), transcript variant 2, mRNA	496	23505	17%	9e-137	82%
94	NM_001908.3	cathepsin B (CTSB), transcript variant 1, mRNA	496	23505	17%	9e-137	82%
95	NM_145702.1	tigger transposable element derived 1 (TIGD1), mRNA	496	1148	0%	9e-137	81%
96	NM_147783.2	cathepsin B (CTSB), transcript variant 5, mRNA	496	23505	17%	9e-137	82%
97	NM_147782.2	cathepsin B (CTSB), transcript variant 4, mRNA	496	23505	17%	9e-137	82%
98	NM_147781.2	cathepsin B (CTSB), transcript variant 3, mRNA	496	23505	17%	9e-137	82%
99	NM_002886.3	RAP2B, member of RAS oncogene family (RAP2B), mRNA	494	9897	8%	3e-136	82%
100	NM_021167.4	GATA zinc finger domain containing 1 (GATAD1), transcript variant 1, mRNA	494	27870	15%	3e-136	83%

其中包含 KAT8、PRSS8、PRSS36、FUS、PYCARD、TRIM72、PYDC1 等 FGENESH 预测的结果，以及很多其他序列，大部分没有出现在 Ensembl 的注释信息中，可能为上述基因序列的同源序列。可见用基因组序列进行 blast，准确性较差，效率较低。

用 TRIM72 的基因编码区序列和氨基酸序列进行 Blast 在 homework4 中已经做过，此处不再重复。