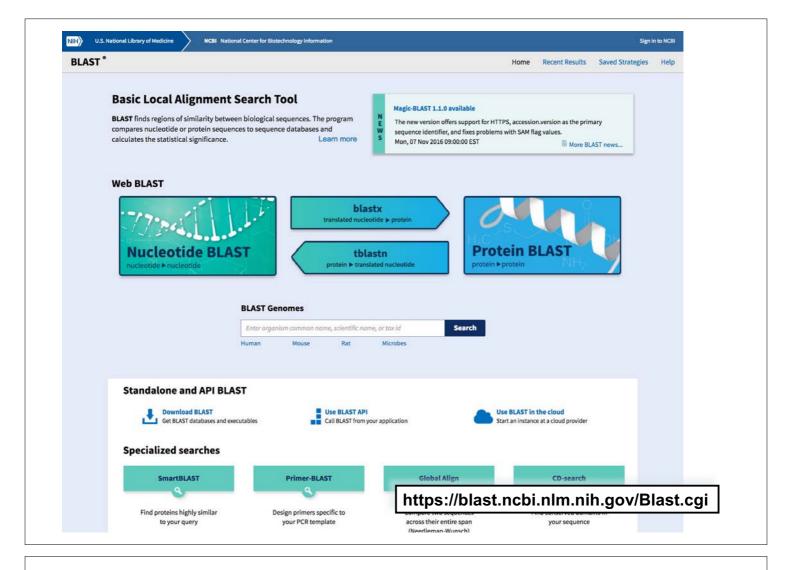# BLAST basics

Shuji Shigenobu /重信秀治

Aim
- BLASTの基本を学ぶ
- CUI環境でBLASTを実行する技術を習得する

What's **BLAST**

Basic Local Alignment Search Tool

▸ BLAST is the tool most frequently used for calculating sequence similarity

▸ Finds best local alignments

▸ Heuristic approach based on Smith Waterman algorithm

▸ Provides statistical significance

▸ BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families

http://blast.ncbi.nlm.nih.gov/Blast.cgi

https://blast.ncbi.nlm.nih.gov/Blast.cgi

BLAST search in GUI

▸ 宿題

BLASTに関する宿題

コースではコマンドライン上でCUIでBLAST解析を行いますが、予習として、NCBIのホームページ上でのGUIでのBLAST解析をやってみましょう。

accession no.: XM_002716705 はウサギのオートファジー遺伝子ATG3です。この遺伝子のマウスホモログ候補を探索する事を目的として、BLAST解析をやってみましょう。

- NCBI BLAST website

1) BLASTのトップヒットのマウスタンパク質は何ですか？

2) トップヒットのタンパク質とqueryとの、アミノ酸のペアワイズアライメントを示しなさい。

（コースでは同じ解析をCUIで行い、GUIでの結果、操作と比較します。）

Standard BLAST Programs

Query	DB	Program
DNA *vs*	DNA	**BLASTN**
Protein *vs*	Protein	**BLASTP**
DNA translation *vs*	Protein	**BLASTX**
Protein *vs*	DNA translation	**TBLASTN**
DNA translation *vs*	DNA translation	**TBLASTX**

Quick start: command line BLAST search

> ウサギのあるcDNA断片データを得た。これが何の遺伝子をコードしているのかを、マウスタンパク質データベースに対してsimilarity searchする事により明らかにする。

▸ Query: XM_002716705.3.fasta

▸ DB: mouse_proteins.pep.fasta

1. Format DB

```
$makeblastdb –in mouse_proteins.pep.fasta –dbtype prot –parse_seqids
```

2. Search

```
$blastx –db mouse_proteins.pep.fasta —query XM_002716705.3.fasta
```

Anatomy of BLAST output (format 0)

out.bl0.txt — Edited

```
BLASTX 2.5.0+

Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro A.
Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J.
Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of
protein database search programs", Nucleic Acids Res. 25:3389-3402.


Database: mouse_proteins.pep.fasta
           49,870 sequences; 21,467,191 total letters


Query= XM_002716705.3 PREDICTED: Oryctolagus cuniculus autophagy related 3
(ATG3), mRNA

Length=1516
                                                                  Score     E
Sequences producing significant alignments:                     (Bits)  Value
Q9CPX6  Ubiquitin-like-conjugating enzyme ATG3 OS=Mus musculus GN...  598    0.0
Q8R1P4  Ubiquitin-like-conjugating enzyme ATG10 OS=Mus musculus G...  35.0   0.13
A0A0R4J0Z9  Autophagy-related 10 (Yeast), isoform CRA_a OS=Mus mu...  34.7   0.14
A0A0X1KG62  Negative elongation factor B OS=Mus musculus GN=Nelfb...  30.4   5.0
Q0VF22  Coiled-coil domain-containing protein 138 OS=Mus musculus...  30.0   6.9
Q8C4Y3  Negative elongation factor B OS=Mus musculus GN=Nelfb PE=...  30.0   7.1
Q9JK38  Glucosamine 6-phosphate N-acetyltransferase OS=Mus muscul...  29.3   8.3
Q8C5K5  Uncharacterized protein CXorf38 homolog OS=Mus musculus P...  29.6   8.8


>Q9CPX6 Ubiquitin-like-conjugating enzyme ATG3 OS=Mus musculus GN=Atg3
PE=1 SV=1
Length=314

 Score = 598 bits (1542),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 309/314 (98%), Positives = 313/314 (99%), Gaps = 0/314 (0%)
 Frame = +1

Query  418  MQNVINTVKGKALEVAEYLTPVLKESKFKETGVITPEEFVAAGDHLVHHCPTWQWATGEE  597
            MQNVINTVKGKALEVAEYLTPVLKESKFKETGVITPEEFVAAGDHLVHHCPTWQWATGEE
Sbjct  1    MQNVINTVKGKALEVAEYLTPVLKESKFKETGVITPEEFVAAGDHLVHHCPTWQWATGEE  60

Query  598  LKVKAYLPTGKQFLVTKNVPCYKRCKQMEYSDELEAIIEEDDGDGGWVDTYHNTGITGIT  777
            LKVKAYLPT KQFLVTKNVPCYKRCKQMEYSDELEAIIEEDDGDGGWVDTYHNTGITGIT
Sbjct  61   LKVKAYLPTDKQFLVTKNVPCYKRCKQMEYSDELEAIIEEDDGDGGWVDTYHNTGITGIT  120

Query  778  EAVKEITLENKDSIKLQDCSALCeeeeeedegeaadmeeyeeSGLLETDEATLDTRKLVE  957
            EAVKEITLE+KDSIKLQDCSALC+EE+EEDEGEAADMEEYEESGLLETDEATLDTRK+VE
Sbjct  121  EAVKEITLESKDSIKLQDCSALCDEEDEEDEGEAADMEEYEESGLLETDEATLDTRKIVE  180

Query  958  ACKAKADAGGEDAILQTRTYDLYITYDKYYQTPRLWLFGYDEQRQPLTVEHMYEDISQDH  1137
            ACKAKADAGGEDAILQTRTYDLYITYDKYYQTPRLWLFGYDEQRQPLTVEHMYEDISQDH
Sbjct  181  ACKAKADAGGEDAILQTRTYDLYITYDKYYQTPRLWLFGYDEQRQPLTVEHMYEDISQDH  240

Query  1138 VKKTVTIENHPHLPPPPMCSVHPCRHAEVMKKIIETVAEGGGELGVHMYLLIFLKFVQAV  1317
            VKKTVTIENHPHLPPPPMCSVHPCRHAEVMKKIIETVAEGGGELGVHMYLLIFLKFVQAV
Sbjct  241  VKKTVTIENHPHLPPPPMCSVHPCRHAEVMKKIIETVAEGGGELGVHMYLLIFLKFVQAV  300

Query  1318 IPTIEYDYTRHFTM  1359
            IPTIEYDYTRHFTM
Sbjct  301  IPTIEYDYTRHFTM  314


>Q8R1P4 Ubiquitin-like-conjugating enzyme ATG10 OS=Mus musculus GN=Atg10
PE=1 SV=1
Length=215

 Score = 35.0 bits (79),  Expect = 0.13, Method: Compositional matrix adjust.
 Identities = 25/116 (22%), Positives = 52/116 (45%), Gaps = 11/116 (9%)
 Frame = +1

Query  994  AILQTRTYDLYITYDKYYQTPRLWLFGYDEQRQPLTVEHMYEDISQDHVKK-------TV  1152
            A++   ++ ++ Y   YQ P L+          +PL +E ++E + + +        T+
Sbjct  88   AVAEVIKHEYHVLYSCSYQVPVLYFRASFLDGRPLALEDIWEGVHECYKPRLLQGPWDTI  147
```

header

one-line summaries

alignments

Anatomy of BLAST output (format 0)

```
Sbjct  184  KKQGLL-PHEINQIYDELYHIHMMKLQYFTTAQKKFAEELQKREQFLAER--------EQL  234

Query  682  EYSDE-----LEAIIEEDDGDGGWVDTYHNTGITGITEAVKEITLENK  810
            +S E      ++ + EE       + H T I +TEA+KE  ENK
Sbjct  235  LFSHETALSKIKGVKEEVLTRFQILKEQHGTEIEHLTEALKEKNKENK  282


>Q8C4Y3 Negative elongation factor B OS=Mus musculus GN=Nelfb PE=1 SV=2
Length=580

 Score = 30.0 bits (66),  Expect = 7.1, Method: Compositional matrix adjust.
 Identities = 10/25 (40%), Positives = 16/25 (64%), Gaps = 1/25 (4%)
 Frame = -2

Query  549  ISSCHKLFWCYDTCFLKLRFLEDRR  475
            +  CHK  WC D C ++ RF++ +R
Sbjct  256  VDPCHKFTWCLDAC-IRERFVDSKR  279


>Q9JK38 Glucosamine 6-phosphate N-acetyltransferase OS=Mus musculus GN=Gnpnat1
PE=1 SV=1
Length=184

 Score = 29.3 bits (64),  Expect = 8.3, Method: Compositional matrix adjust.
 Identities = 20/73 (27%), Positives = 31/73 (42%), Gaps = 13/73 (18%)
 Frame = +1

Query  337  WGRPTAQLLPGLPSPRRRGAGRCSGPRMQNVINTVKGKALEVAEYLTPVLKESKFKETGV  516
            W + TA   P + SP  G G    P     K L           +    KETGV
Sbjct  19   WSQNTAIFSPAI-SPTHPGEGLVLRPLCTADLNKGFFKVL----------GQLTETGV  65

Query  517  ITPEEFVAAGDHL  555
            ++PE+F+ + +H+
Sbjct  66   VSPEQFMKSFEHM  78


>Q8C5K5 Uncharacterized protein CXorf38 homolog OS=Mus musculus PE=1
SV=1
Length=320

 Score = 29.6 bits (65),  Expect = 8.8, Method: Compositional matrix adjust.
 Identities = 22/83 (27%), Positives = 37/83 (45%), Gaps = 2/83 (2%)
 Frame = +1

Query  580  WATGEELKVKAYLPTGKQFLVTKNVPCYKRCKQMEYSDELEAIIEEDDGDGGWVDTYHNT  759
            W    ++K++ L  K   ++ V  Y R +Q+ SD     I EED+ DG     +
Sbjct  177  WLRDFQIKIQNFLNEFKN-IPEIVAVYSRIEQLLTSDWAVHIPEEDERDGCEFEIGSYL  234

Query  760  GITGITEAVKEITLENKDSIKLQ  828
            ++  I E   E+   + LQ
Sbjct  235  SVSQIHEIEIELLKEKLQEMYLQ  257


Lambda      K        H        a        alpha
   0.318    0.134    0.401    0.792     4.96

Gapped
Lambda      K        H        a        alpha    sigma
   0.267    0.0410   0.140    1.90      42.6     43.6

Effective search space used: 6420178198


 Database: mouse_proteins.pep.fasta
   Posted date:  Nov 18, 2016  4:07 PM
 Number of letters in database: 21,467,191
 Number of sequences in database:  49,870


Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Neighboring words threshold: 12
Window for multiple hits: 40
```

alignments

footer

Hierarchical structure of alignment part
- alignment #1
   - HSP #1
   - HSP #2
   - ...
- alignment #2
- alignment #3
- ...

Let's try

▸ help を表示させる

▸ -evalue オプションを使ってみる。evalue cutoff = 1.0e-8 を指定してみよう

▸ 他のフォーマットで出力してみよう

　　▸ format 6: table

　　▸ format 7: table with comment lines

　　▸ format 5: XML

Anatomy of BLAST output (format 7)

```
●●●                          📄 out.bl7.txt — Edited ⌄
# BLASTX 2.5.0+
# Query: XM_002716705.3 PREDICTED: Oryctolagus cuniculus autophagy related 3 (ATG3), mRNA
# Database: mouse_proteins.pep.fasta
# Fields: query acc., subject acc., % identity, alignment length, mismatches, gap opens, q. start,
q. end, s. start, s. end, evalue, bit score
# 8 hits found
XM_002716705.3      Q9CPX6 98.408 314     5      0      418     1359   1      314     0.0     598
XM_002716705.3      Q8R1P4 21.552 116     80     3      994     1320   88     199     0.13    35.0
XM_002716705.3      Q0VF22 27.778 108     64     3      502     810    184    282     6.9     30.0
XM_002716705.3      Q8C4Y3 40.000 25      14     1      549     475    256    279     7.1     30.0
XM_002716705.3      Q9JK38 27.397 73      40     2      337     555    19     78      8.3     29.3
XM_002716705.3      Q8C5K5 26.506 83      59     1      580     828    177    257     8.8     29.6
# BLAST processed 1 queries
```

Concise.
Useful for large scale analysis.
Tab-delimited text is easy to parse.
You can modified the format. (explain later)

Slide 1

Anatomy of BLAST output (format 6)

format 6/7 はcustomize可能

query	subject	%identity	align-len mismatch	gap_open	q_start q_end	s_start s_end	evalue	bit-score

query	subject	%identity	align-len	mismatch	gap_open	q_start	q_end	s_start	s_end	evalue	bit-score
spo:SPAC212.11	sce:YMR190C	27.297	370	238	12	1169	1525	653	1004	7.08e-23	105
spo:SPAC212.11	sce:YDR021W	38.053	113	64	3	1417	1529	267	373	1.42e-14	75.9
spo:SPAC212.11	sce:YNL112W	25.076	331	225	12	1204	1525	148	464	1.06e-12	70.9
spo:SPAC212.11	sce:YBR237W	24.403	377	237	17	1190	1532	276	638	1.68e-12	70.5
spo:SPAC212.11	sce:YOR204W	28.505	214	128	8	1352	1559	339	533	2.80e-12	69.7
spo:SPAC212.11	sce:YOR046C	28.205	234	140	8	1310	1530	227	445	1.65e-11	66.6
spo:SPAC212.11	sce:YDR243C	23.512	336	227	10	1208	1524	216	540	7.40e-11	65.1
spo:SPAC212.11	sce:YGL078C	22.689	357	239	11	1188	1527	130	466	7.62e-11	64.7
spo:SPAC212.11	sce:YPL119C	27.619	210	129	8	1352	1556	351	542	2.84e-10	63.2
spo:SPAC212.11	sce:YGL064C	22.195	410	216	16	1215	1541	166	555	8.25e-10	61.6
spo:SPAC212.11	sce:YDL084W	24.424	217	156	5	1308	1524	201	409	7.84e-07	51.6
spo:SPAC212.11	sce:YHR169W	27.273	143	93	4	1418	1555	257	393	1.09e-06	51.2
spo:SPAC212.11	sce:YLR276C	26.667	210	120	10	1189	1377	38	234	2.85e-06	50.1
spo:SPAC212.11	sce:YLR276C	34.146	41	27	0	1484	1524	386	426	0.12	35.0
spo:SPAC212.11	sce:YLL008W	21.965	346	233	13	1195	1524	258	582	4.71e-06	49.3
spo:SPAC212.11	sce:YPL082C	24.074	108	81	1	1417	1524	1648	1754	0.002	40.8
spo:SPAC212.11	sce:YGL070C	37.838	37	23	0	1583	1619	52	88	2.3	29.3
spo:SPBPB10D8.05C	sce:YPL092W	30.730	397	224	9	10	356	11	406	3.04e-47	165
spo:SPBPB10D8.05C	sce:YGL195W	31.034	58	37	2	140	195	2014	2070	0.37	30.8
spo:SPBPB10D8.05C	sce:YDR283C	51.613	31	14	1	221	250	589	619	0.66	30.0
spo:SPBPB10D8.05C	sce:YBR028C	35.135	37	24	0	229	265	127	163	0.84	29.6
spo:SPBPB10D8.05C	sce:YPL027W	27.957	93	54	3	138	221	53	141	3.1	27.3
spo:SPBPB10D8.05C	sce:YDR161W	28.571	49	35	0	318	366	102	150	3.2	27.7
spo:SPBPB10D8.05C	sce:YER166W	27.451	51	35	1	50	98	1273	1323	7.0	26.9
spo:SPBPB10D8.05C	sce:YGR040W	23.256	86	55	3	214	299	3	77	7.8	26.6
spo:SPBPB10D8.05C	sce:YAR019C	44.444	18	10	0	235	252	30	47	9.5	26.2
spo:SPBC359.02	sce:YOR368W	29.885	87	56	2	53	138	7	89	0.88	29.3
spo:SPCC330.07C	sce:YHR197W	26.761	71	39	3	65	130	287	349	2.0	28.9
spo:SPCC330.07C	sce:YGR224W	23.881	67	50	1	181	247	152	217	5.1	27.7

Slide 2

Multiple queries

ウサギのあるcDNA断片データを100個得た。これが何の遺伝子をコードしているのかを、マウスタンパク質データベースに対してsimilarity searchする事により明らかにする。多数の配列を解析したいときにこそCUIの本領発揮。

▸ Query: rabbit_100mRNAs.nuc.fastaa

▸ DB: mouse_proteins.pep.fasta

▸ options:

  ▸ evalue < 1.0e-5; format 7; 高速化のためにCPUを4コア使う。

1. Format DB

新たに作成する必要はない。

2. Search

自分で考えてみよう

Multiple queries

ウサギのあるcDNA断片データを100個得た。これが何の遺伝子をコードしているのかを、マウスタンパク質データベースに対してsimilarity searchする事により明らかにする。多数の配列を解析したいときにこそCUIの本領発揮。

▸ Query: rabbit_100mRNAs.nuc.fastaa

▸ DB: mouse_proteins.pep.fasta

1. Format DB

新たに作成する必要はない。

2. Search

```
blastx -db mouse_proteins.pep.fasta -query rabbit_100mRNAs.nuc.fasta
-evalue 1.0e-5 -num_threads 4 -outfmt 7 > practice.out.txt
```

Why and when command-line BLAST is useful?

▸ Automated

▸ Many queries

▸ Use local database (unpublished)

▸ Procedures can be recorded as a shell script

NCBI BLAST+ command line tools

- ▸ Provided as BLAST+ suite by NCBI
    - ▸ (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs &DOC_TYPE=Download)

- ▸ Functionality offered by BLAST+ suite
    - ▸ 1) Search tools
    - ▸ 2) BLAST database tools
    - ▸ 3) sequence filtering tools

Installation and configuration

- ▸ Installation
    - ▸ OSごとにコンパイル済みのバイナリが配布されている。ダウンロードして解凍するだけでよい。
    - ▸ Manuals
        - ▸ https://www.ncbi.nlm.nih.gov/books/NBK279671/
        - ▸ https://www.ncbi.nlm.nih.gov/books/NBK52640/

- ▸ Setup and configuration
    - ▸ blast programs のパスを通す
    - ▸ 環境変数 BLASTDB を設定する

```
export BLASTDB=$HOME/myblastdb
```

（コース用Macおよびbias4はすでに環境設定済み）

Exercise

- (詳細はwiki)
- ex 1-1: (blastn) mapping cDNA/EST to a genome
- ex 1-2: (blastp) homolog candidate search
- ex 1-3: (blastx) annotate RNA-seq contigs
- ex 1-4: (tblastn) mapping proteins to genome

Advanced usage

- Pairwise alignment by using blast commands
- Advanced parameters
- Customize output format
- blastdbcmdを使った配列データのmanipulation – Day 2
- advanced blast search tools – Day 2

Pairwise BLAST

BLASTプログラムを使って、2つの配列のローカルアライメントをとることも可能である。blastdbcmdでデータベースを作製する必要はない。
以下は、ウサギとマウスのAtg3タンパク質の配列である。blastpを使ってpairwiseアライメントをとってみよう。

▸ Query 1: OcAtg3.aa.fa
▸ Query 2: MmAtg3.aa.fa

Alignment

```
$blastp –query OcAtg3.aa.fa –subject MmAtg3.aa.fa
```

-dbのかわりに-subjectを使う以外は、通常のBLASTとほぼ同じ。

Exercise

▸ (詳細はwiki)
▸ ex 1-5: pairwise alignment by BLAST

BLASTN –task option

- **-task**
  - **megablast** (default): for very similar sequences
  - **dc-megablast**: typically used for inter-species comparisons
  - **blastn**: traditional program used for inter-species comparisons
  - **blastn-short**: optimized for sequences less than 30 nucleotides

---

BLASTN: -task

肺炎クラミジア Chlamydophila pneumoniae のCWL029株と、GPIC株のゲノム全長をBLASTNで比較しなさい。-task で megablast (default) , blastn, dc-megablast を指定したときで結果がどのように変わるだろうか。

- Seq1: CpneCWL029.NC_000922.uc.genome.fa
- Seq2: CpneGPIC.AE015925.uc.genome.fa

Alignment

```
$blastn –task megablast –subject seq1 –query seq2 –outfmt 6
```

-task を blastn, dc-megablast に変えてみよう

Customize output format

▸ **Format 6/7 can be arranged by users**

> OcAtg3.aa.fa と MmAtg3.aa.fa のpairwise BLASTをカスタム
> フォーマットで出力する例。

```
$blastp -query OcAtg3.aa.fa -subject MmAtg3.aa.fa \
 -outfmt "6 std qseq sseq"
```

標準の12カラム(qacc sacc pident length mismatch gapopen qstart qend sstart
send evalue　bitscore)に加えて、qseqとsseq を表示

```
Options 6, 7, 10 and 17 can be additionally configured to produce
a custom format specified by space delimited format specifiers.
The supported format specifiers for options 6, 7 and 10 are:
        qseqid means Query Seq-id
          qgi means Query GI
         qacc means Query accesion
      qaccver means Query accesion.version
         qlen means Query sequence length
       sseqid means Subject Seq-id
    sallseqid means All subject Seq-id(s), separated by a ';'
          sgi means Subject GI
       sallgi means All subject Gis
         sacc means Subject accession
      saccver means Subject accession.version
      sallacc means All subject accessions
         slen means Subject sequence length
       qstart means Start of alignment in query
         qend means End of alignment in query
       sstart means Start of alignment in subject
         send means End of alignment in subject
          qseq means Aligned part of query sequence
          sseq means Aligned part of subject sequence
       evalue means Expect value
     bitscore means Bit score
        score means Raw score
       length means Alignment length
        pident means Percentage of identical matches
        nident means Number of identical matches
     mismatch means Number of mismatches
     positive means Number of positive-scoring matches
      gapopen means Number of gap openings
```

```
      positive means Number of positive-scoring matches
       gapopen means Number of gap openings
          gaps means Total number of gaps
          ppos means Percentage of positive-scoring matches
        frames means Query and subject frames separated by a '/'
        qframe means Query frame
        sframe means Subject frame
          btop means Blast traceback operations (BTOP)
        staxid means Subject Taxonomy ID
      ssciname means Subject Scientific Name
      scomname means Subject Common Name
    sblastname means Subject Blast Name
     sskingdom means Subject Super Kingdom
        staxids means unique Subject Taxonomy ID(s), separated by a ';'
           (in numerical order)
     sscinames means unique Subject Scientific Name(s), separated by a ';'
     scomnames means unique Subject Common Name(s), separated by a ';'
   sblastnames means unique Subject Blast Name(s), separated by a ';'
           (in alphabetical order)
    sskingdoms means unique Subject Super Kingdom(s), separated by a ';'
           (in alphabetical order)
        stitle means Subject Title     salltitles means All Subject Title(s),
separated by a '<>'
       sstrand means Subject Strand
         qcovs means Query Coverage Per Subject
       qcovhsp means Query Coverage Per HSP
        qcovus means Query Coverage Per Unique Subject (blastn only)

When not provided, the default value is:   'qaccver saccver pident length
mismatch gapopen qstart qend sstart send   evalue bitscore', which is equivalent
to the keyword 'std'
```

More exercises

▸ (詳細はwiki)

▸ ex 1-7: blastn: CRISPR off-target check

▸ ex 1-8: compare two closely related genome

▸ ex 1-9: mapping a PacBio read to a genome