Continuing lecture 6

• The BLAST+ binaries have been downloaded and linked into the bin directory

• The sequence has been extracted from the GFF file

• We know how to create a blast database
makeblastdb -parse_seqids option

• Use the \texttt{-parse_seqids} flag when invoking makeblastdb \texttt{\textasciitilde} allows the retrieval of sequences based upon sequence identifiers.

• In that case, each sequence must have a \texttt{unique identifier}, and that identifier must have a specific format

see also section \textit{5.14 Limiting a Search with a List of Identifiers} in the BLAST+ handbook

• This is an advanced option may be also used to restrict the search to a subset of sequences
```
# look at what files are created when you run
# each of the following

~/bin/makeblastdb -in sc.fa -dbtype nucl

~/bin/makeblastdb -in sc.fa -dbtype nucl -parse_seqids
```

More files are created →
# FASTA sequence ID format values

## Table 5. FASTA sequence ID format values

<table>
<thead>
<tr>
<th>Type</th>
<th>Format(s)</th>
<th>Example(s)</th>
</tr>
</thead>
</table>
| local                    | lcl|integer  
                         | lcl|string          | lcl|123  
                         | lcl|hmm271       |
| GenInfo backbone seqid   | bbs|integer  | bbs|123                      |
| GenInfo backbone moltype | bbm|integer  | bbm|123                      |
| GenInfo import ID        | gim|integer  | gim|123                      |
| GenBank                  | gb|accession|locus               | gb|M73307|AGMA13GT |
| EMBL                     | emb|accession|locus          | emb|CAM43271.1    |
| PIR                      | pir|accession|name     | pir|G36364            |
| SWISS-PROT               | sp|accession|name     | sp|P01013|OVAX_CHICK |
| patent                   | pat|country|patent|sequence | pat|US|RE33188|1   |
| pre-grant patent         | pgp|country|application-number|seq-number | pgp|EP|0238993|7 |
| RefSeq ²                 | ref|accession|name     | ref|NM_010450.1     |
| general database reference | gn1|database|integer | gn1|taxon|9606 |
|                          | gn1|database|string  | gn1|PID|e1632      |
| GenInfo integrated database | gi|integer      | gi|21434723            |
| DDBJ                     | dbj|accession|locus          | dbj|BAC85684.1|
| PRF                      | prf|accession|name     | prf|0806162C      |
blastdbcmd

• A useful tool with an unfortunate name

• and unfortunate parameters

• and unfortunate documentation
One of the most common questions

How to extract a small sub-sequence from a genome?

<table>
<thead>
<tr>
<th>Id</th>
<th>Type</th>
<th>Title</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>11947</td>
<td>Question</td>
<td>Extraction of part of different trp of nucleotide sequence in 10000 of sequences.</td>
<td>sequence in 10000 of sequences. Hi I would like... I am having 10000 sequences as a contig and... ranges and I want to extract in all 10000 sequences. I also want to</td>
</tr>
<tr>
<td>1195</td>
<td>Question</td>
<td>extracting sequence from a 3GB fasta file</td>
<td>extracting sequence from a 3GB fasta... Hi, How to extract fasta sequence from an huge 3gb...fasta file by giving sequence id as input using</td>
</tr>
<tr>
<td>10220</td>
<td>Question</td>
<td>Extract domain sequences from multiple sequences</td>
<td>Extract domain sequences from multiple...sequences Hi, I have 100 protein sequences with some conserved...domains. I want to extract the domain sequences in a go. Is it possible</td>
</tr>
<tr>
<td>15538</td>
<td>Question</td>
<td>how to extract patterns from protein sequence</td>
<td>sequence I just want to know when a protein sequence is given with a... such as biopm will extract data so that I can... algorithms do to extract data. whether they</td>
</tr>
<tr>
<td>12392</td>
<td>Question</td>
<td>Extracting protein sequences</td>
<td>Extracting protein sequences Hi, I am trying... extract all the protein sequences of Rho GEF in humans... redundancy and also extract all the sequences. How</td>
</tr>
</tbody>
</table>

There are a number of answers – blastdbcmd may not the simplest but is a good start and usually you already have the commands installed.
Displaying the content of a blast database

```bash
ialbert@porthos
$ ~/bin/blastdbcmd -db ~/refs/yeast/sc_full -entry 'all' | head -3
>lcl|chrI
CCACACACACACACACACACACACACACACACACACACACACACACACACACACATCTAATACACTTACCTAAAC
ACAGCCCTAATCTAACCCCTGCGCAACCTGTCTCTCAACCTTTACCTCACCTGCTCCACCTCGTTACCCTGTCCCAT

ialbert@porthos
$ ~/bin/blastdbcmd -db ~/refs/yeast/sc_full -entry 'all' -outfmt "%a" | head -3
chrI
chrII
chrIII

ialbert@porthos
$ ~/bin/blastdbcmd -db ~/refs/yeast/sc_full -entry 'all' -outfmt "%l" | head -3
230218
813184
316620

ialbert@porthos
$ 
```
Default = `-'
-outfmt <String>

Output format, where the available format specifiers are:

%f means sequence in FASTA format
%s means sequence data (without defline)
%a means accession
%g means gi
%o means ordinal id (OID)
%i means sequence id
%t means sequence title
%l means sequence length
%h means sequence hash value
%T means taxid
%e means membership integer
%L means common taxonomic name
%S means scientific name
%P means PIG
%m means sequence masking data.

Masking data will be displayed as a series of 'N-M' values
separated by ';' or the word 'none' if none are available.
If '%f' is specified, all other format specifiers are ignored.
For every format except '%f', each line of output will correspond to
a sequence.
Extract regions of your sequences

```
ialbert@porthos ~/work/lec6
$ ~/bin/blastdbcmd -db refs/sc -entry 'all' -range '1-10' | head -6
>lcl|chrI:1-10
CCACACCACA
>lcl|chrII:1-10
AAATAGCCCT
>lcl|chrIII:1-10
CCACACACCC

ialbert@porthos ~/work/lec6
$ ~/bin/blastdbcmd -db refs/sc -entry 'all' -range '1-10'
```
Fetch a subsequence by accession

```
ialbert@porthand ~/work/lec6
$ ~/bin/blastdbcmd -db refs/sc -entry 'chrV' -range '1000-1020'
>lcl|chrV:1000-1020
CTTTTTACGCTAAAATATTTTC

ialbert@porthand ~/work/lec6
$ 
```

```
~/bin/blastdbcmd -db refs/sc -entry 'chrV' -range '1000-1020'
```
Batch entries

```
ialbert@porthos
$ cat ids.txt
chrI 100-110 plus
chrI 100-110 minus
ialbert@porthos
$ ~/bin/blastdbcmd -db ~/refs/yeast/sc_full -entry_batch ids.txt
>lcl|chrI:100-110
GGCCAACCTGT
>lcl|chrI:c110-100
ACAGGTGGCC
ialbert@porthos
$ 
```
Beware of tacit behavior

```
ialbert@porthos
$ cat sc.fa | head -2 > mini.fa

ialbert@porthos
$ cat mini.fa
>chrI
CCACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACACAC

ialbert@porthos
$ ~/bin/blastn -db ~/refs/yeast/sc_full -query mini.fa -outfmt 7
# BLASTN 2.2.28+
# Query: chrI
# Database: /Users/ialbert/refs/yeast/sc_full
# 0 hits found
# BLAST processed 1 queries

ialbert@porthos
$ ...
```
Low complexity filter is ON by default

IALBERT@PORTHOS

$ ~/bin/blastn -db ~/refs/yeast/sc_full -query mini.fa -dust no -outfmt 7
# BLASTN 2.2.28+
# Database: /Users/ialbert/refs/yeast/sc_full
# Fields: query id, subject id, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, eval, bit score
# 7 hits found
chrI chrI 100.00 80 0 0 1 80 1 80 1e-36 148
chrI chrVIII 96.77 62 0 2 2 61 562608 562548 9e-23 102
chrI chrXII 92.06 63 3 2 2 62 9 71 3e-18 87.9
chrI chrXII 91.80 61 1 4 2 61 9 58 1e-16 82.4
chrI chrXII 87.18 78 2 7 1 78 1065079 1065010 1e-16 82.4
chrI chrXIII 91.07 56 3 2 6 59 924429 924374 2e-14 75.0
chrI chrIV 97.14 35 0 1 1 34 1525405 1525371 2e-09 58.4
# BLAST processed 1 queries
IALBERT@PORTHOS
$
Options 6, 7, and 10 can be additionally configured to produce a custom format specified by space delimited format specifiers. The supported format specifiers are:

- `qseqid` means Query Seq-id
- `qgi` means Query GI
- `qacc` means Query accession
- `qaccver` means Query accession.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

-`outfmt '7 qseqid sseqid mismatch'`
Homework 7

• Using `blastdbcmd` create a list of the chromosomes in the yeast genome sorted by their sizes (show the sizes as well)

• Show the 25 base pair long region that lies directly ahead of gene `YBR299W`

Hint: find your gene in `features.gff`. 