Contents

1 remoVecSec

remoVecSec is a library of modules that allows to remove contamination in assembled genomes prior to NCBI WGS submission.

1.1 Background

NCBI's Foreign Contamination Screens, November 2016

The purpose of the foreign contamination screens is to identify contaminating sequences that may be present for artificial reasons or for biological reasons. Artificial reasons include cloning artifacts (vector, linker/adaptor/primer, E. coli host DNA), contamination in the lab with human sequence, mixing of samples or sequencing runs with other organisms, and bacterial insertion sequences that have integrated into sequenced clones. Biological reasons include the presence of endosymbionts, infectious agents, or microbes residing on the surface of the organism or in the gut when the DNA prep was made.

Our suite of foreign contamination screens uses BLAST to screen the submitted sequences against:

1. a common contaminants database that contains vector sequences, bacterial insertion sequences, E. coli and phage genomes

2. a database of adaptors linkers and primers

3. a database of mitochondrial genomes
4. the chromosomes of unrelated organisms
5. a database of ribosomal RNA genes

Suspect spans are re-BLASTed against:

1. the chromosomes of unrelated organisms
2. the chromosomes of related organisms
3. the NCBI nt BLAST database of nucleotide sequence from all traditional divisions of GenBank, EMBL, and DDBJ
4. the NCBI htgs BLAST database of sequences from the HTG division of GenBank, EMBL, and DDBJ

Results similar to those obtained by NCBI could be generated by running the screens as described below.

1.1.1 Common contaminant screen

1. Databases

   (a) File to screen for the common contaminants in eukaryotic sequences:

   `ftp://ftp.ncbi.nlm.nih.gov/pub/kitts/contam_in_euks.fa.gz`

   Contains the cloning artifacts that are likely to show up as contaminants across all eukaryotic species: vector sequences, E.coli genome, phage genomes, bacterial Insertion Sequences and transposons.

   (a) File to screen for the common contaminants in prokaryotic sequences:

   `ftp://ftp.ncbi.nlm.nih.gov/pub/kitts/contam_in_prok.fa`

   Contains phiX174.

   These files need to be unzipped and the resulting FASTA sequence files formatted as BLAST databases using the makeblastdb program.

2. Programs

   blastn and makeblastdb are contained in the blast+ package which can be installed following the instruction in the BLAST help documents.

   "BLAST Command Line Applications User Manual":

   `https://www.ncbi.nlm.nih.gov/books/NBK279671/`

   "Standalone BLAST Setup for Windows PC":

   `https://www.ncbi.nlm.nih.gov/books/NBK52637/`

   "Standalone BLAST Setup for Unix":

   `https://www.ncbi.nlm.nih.gov/books/NBK52640/`

3. Execution

   A BLAST search is run against either the contam$\backslash_{\text{in}}\backslash_{\text{euks}}$ or contam$\backslash_{\text{in}}\backslash_{\text{prok}}$ database, depending on the origin of the input sequences. The common contaminant BLAST results are filtered for hits over various length and percent identity cut-offs.

   Command line:

   (a) for screening eukaryotic sequences:

   ```
   blastn -query _input_fasta_sequences_ -db contam_in_euks -task megablast -word
   ```

   OR with an intermediate file, these 2 commands:

   ```
   blastn -query _input_fasta_sequences_ -db contam_in_euks -task megablast -word_siz
   ```

   ```
   awk '($3>=98.0 && $4>=50)||($3>=94.0 && $4>=100)||($3>=90.0 && $4>=200)' _out_file
   ```

   (a) for screening prokaryotic sequences:

   ```
   blastn -query _input_fasta_sequences_ -db contam_in_prok -task megablast -word
   ```

   OR with an intermediate file, these 2 commands:

   ```
   blastn -query _input_fasta_sequences_ -db contam_in_prok -task megablast -word_siz
   ```

   ```
   awk '($3>=98.0 && $4>=50)||($3>=94.0 && $4>=100)||($3>=90.0 && $4>=200)' _out_file
   ```

1.1.2 Adaptor screen

VecScreen (`https://www.ncbi.nlm.nih.gov/tools/vecscreen/`) is run against either the adaptors\for\screening\euks.fa database or adaptors\for\screening\proks.fa database, depending on the origin of the input sequences. Hits are filtered to retain only those matches that VecScreen classifies as "Strong" or "Moderate" (see: `https://www.ncbi.nlm.nih.gov/tools/vecscreen/about/#Categories`).

1. Databases

   The adaptors\for\screening databases are available here:

   ```
   ftp://ftp.ncbi.nlm.nih.gov/pub/kitts/adaptors_for_screening_euks.fa
   ```

   ```
   ftp://ftp.ncbi.nlm.nih.gov/pub/kitts/adaptors_for_screening_proks.fa
   ```

   These FASTA sequence files need to be formatted as BLAST databases using the makeblastdb program.

2. Programs

   The VecScreen standalone program is available here:

   ```
   ftp://ftp.ncbi.nlm.nih.gov/blast/demo/vecscreen
   ```

   The script to filter the VecScreen results is here:

   ```
   ftp://ftp.ncbi.nlm.nih.gov/pub/kitts/VSlistTo1HitPerLine.awk
   ```

3. Execution

   Command line:

   (a) for screening eukaryotic sequences:

   ```
   vecscreen -d adaptors_for_screening_euks.fa -f3 -i _input_fasta_sequences_ -o
   ```

   (b) for screening prokaryotic sequences:

   ```
   vecscreen -d adaptors_for_screening_proks.fa -f3 -i _input_fasta_sequences_ -
   ```

   Filter out the "Weak" and "Suspect Origin" hits:

   ```
   VSlistTo1HitPerLine.awk suspect=0 weak=0 _vs_output_file_ > _filtered_vs_output_f
   ```

1.1.3 Mitochondrial genome screen

BLAST is used to screen the input sequences against a database of the mitochondrial genome sequences in the NCBI Reference Sequences (RefSeq) collection.

1. Database

   ```
   ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/mito.nt.gz
   ```

   This file needs to be unzipped and the resulting FASTA sequence file formatted as a BLAST database using the makeblastdb program.

2. Programs

   blastn and makeblastdb are contained in the blast+ package (see above).

3. Execution

   The BLAST hits to mitochondrial genomes are filtered for hits over 98.6% identity and at least 120 bases long.

   ```
   blastn -query _input_fasta_sequences -db mito.nt -out % -task megablast -word_size
   ```

1.1.4 Ribosomal RNA screen

Ribosomal RNA genes are the cause of many false positives because the include some segments that align to distantly related organisms. Segments that match rRNA genes are identified so that such segments are not reported as being foreign.

BLAST is used to screen the input sequences against a database of the rRNA gene sequences .

1. Database

   ```
   ftp://ftp.ncbi.nlm.nih.gov/pub/kitts/rrna.gz
   ```

   This file needs to be unzipped and the resulting FASTA sequence file formatted as a BLAST database using the makeblastdb program.

2. Programs

   blastn and makeblastdb are contained in the blast+ package (see above).

3. Execution

   The BLAST hits to rRNA genes are filtered for hits over 95% identity and at least 100 bases long.

   ```
   blastn -query _input_fasta_sequences_ -db rrna -task megablast -template_length 18
   ```

1.1.5 Foreign chromosome screen

Screens for matches to chromosome sequences from foreign organisms. Foreign organisms are those that belong to a different taxonomic group compared to the organism whose sequences are being screened. The taxonomic groups are:
arthropoda, chordata, other$\backslash_{\text{metazoa}}$,
viridiplantae, fungi, other$\backslash_{\text{eukaryota}}$,
bacteria, archaea, viruses$\backslash_{\text{and}}\backslash_{\text{viroids}}$

1. Databases

   Our databases to detect cross-contamination detection are limited to assemblies that have been publicly released in GenBank/ENA/DDBJ and subsequently picked up by RefSeq. Genome centers can do better by augmenting these databases with additional genomes that they have sequenced but which are not yet represented in the RefSeq collection.

   (a) archaea

   Query in Nucleotide :

   archaea[porgn] AND srcdb$\backslash_{\text{refseq}}$[prop] AND biomol$\backslash_{\text{genomic}}$[prop] AND complete[prop]

   (a) bacteria

   Query in Nucleotide :

   bacteria[porgn] AND srcdb$\backslash_{\text{refseq}}$[prop] AND biomol$\backslash_{\text{genomic}}$[prop] AND complete[prop]

   (a) fungi

   Query in Nucleotide :

   fungi[porgn] AND srcdb$\backslash_{\text{refseq}}$[prop] AND biomol$\backslash_{\text{genomic}}$[prop] AND (NC$\backslash_{000000}$:NC$\backslash_{999999}$[pacc] OR AC$\backslash_{000000}$:AC$\backslash_{999999}$[pacc] OR (NT$\backslash_{000001}$:NT$\backslash_{999999999}$[pacc]

AND ("chromosome 2L" OR "chromosome 2R" OR "chromosome 3L" OR "chromosome 3R")))

(a) arthropoda

Query in Nucleotide :

arthropoda[porgn] AND srcdb$\backslash_{\text{refseq}}$[prop] AND biomol$\backslash_{\text{genomic}}$[prop] AND (NC$\backslash_{000000}$:NC$\backslash_{999999}$[pacc] OR AC$\backslash_{000000}$:AC$\backslash_{999999}$[pacc] OR (NT$\backslash_{000001}$:NT$\backslash_{999999999}$[pacc] AND ("chromosome 2L" OR "chromosome 2R" OR "chromosome 3L" OR "chromosome 3R")))

(a) chordata

Query in Nucleotide :

chordata[porgn] AND srcdb$\backslash_{\text{refseq}}$[prop] AND biomol$\backslash_{\text{genomic}}$[prop] AND (NC$\backslash_{000000}$:NC$\backslash_{999999}$[pacc] OR AC$\backslash_{000000}$:AC$\backslash_{999999}$[pacc] OR (NT$\backslash_{000001}$:NT$\backslash_{999999999}$[pacc] AND ("chromosome 2L" OR "chromosome 2R" OR "chromosome 3L" OR "chromosome 3R")))

(a) other$\backslash_{\text{metazoa}}$

Query in Nucleotide :

metazoa[porgn] NOT (arthropoda[porgn] OR chordata[porgn]) AND srcdb$\backslash_{\text{refseq}}$[prop] AND biomol$\backslash_{\text{genomic}}$[prop] AND (NC$\backslash_{000000}$:NC$\backslash_{999999}$[pacc] OR AC$\backslash_{000000}$:AC$\backslash_{999999}$[pacc] OR (NT$\backslash_{000001}$:NT$\backslash_{999999999}$[pacc] AND ("chromosome 2L" OR "chromosome 2R" OR "chromosome 3L" OR "chromosome 3R")))

(a) viridiplantae

Query in Nucleotide :

viridiplantae[porgn] AND srcdb$\backslash_{\text{refseq}}$[prop] AND biomol$\backslash_{\text{genomic}}$[prop] AND (NC$\backslash_{000000}$:NC$\backslash_{999999}$[pacc] OR AC$\backslash_{000000}$:AC$\backslash_{999999}$[pacc] OR (NT$\backslash_{000001}$:NT$\backslash_{999999999}$[pacc] AND ("chromosome 2L" OR "chromosome 2R" OR "chromosome 3L" OR "chromosome 3R")))

(a) other$\backslash_{\text{eukaryota}}$

Query in Nucleotide :

eukaryota[porgn] NOT (metazoa[porgn] OR fungi[porgn] OR viridiplantae[porgn]) AND srcdb$\backslash_{\text{refseq}}$[prop] AND biomol$\backslash_{\text{genomic}}$[prop] AND (NC$\backslash_{000000}$:NC$\backslash_{999999}$[pacc] OR AC$\backslash_{000000}$:AC$\backslash_{999999}$[pacc] OR (NT$\backslash_{000001}$:NT$\backslash_{999999999}$[pacc] AND ("chromosome 2L" OR "chromosome 2R" OR "chromosome 3L" OR "chromosome 3R")))

(a) viruses$\backslash_{\text{and}}\backslash_{\text{viroids}}$

Query in Nucleotide :

(viruses[porgn] OR viroids[porgn]) AND srcdb$\backslash_{\text{refseq}}$[prop] AND biomol$\backslash_{\text{genomic}}$[prop] AND (NC$\backslash_{000000}$:NC$\backslash_{999999}$[pacc] OR AC$\backslash_{000000}$:AC$\backslash_{999999}$[pacc] OR (NT$\backslash_{000001}$:NT$\backslash_{999999999}$[pacc] AND ("chromosome 2L" OR "chromosome 2R" OR "chromosome 3L" OR "chromosome 3R")))

The FASTA sequence files resulting from these queries are formatted as nine BLAST databases using the makeblastdb program.

2. Execution

   Repeats in the input FASTA sequences are soft-masked to lowercase using WindowMasker. Then BLAST hits over 98% identity are generated to the databases for the 8 taxonomic groups to which the organism being screened does not belong.

```
blastn -query _input_fasta_sequences_ -db _distant_organism_dbs_ -task megablast -
```

1.1.6 First pass calls

The following heuristic rules help to get rid of most false matches.

1. Process contaminant matches from 1

   Contaminant matches from (1) are merged if they are from the same class of sequence (VECTOR, E.coli, IS, PHG) and they overlap or are separated by 50 bases or less.

   If the total coverage of contaminant matches from (1) is >75% of the sequence length then flag the sequence as a contaminant to be excluded.

   If the contaminant is classed as VECTOR, E.coli, IS:./, PERM:./ or PHG:* and the contaminant location is within 100 bases of the the start or end of the sequence (or gap is the sequence is not contiguous), or within 100 bases of another contaminant match that is at an end, flag the contaminant span for trimming.

If the contaminant is one of the above, and the match is longer than 700 bases flag the contaminant span for trimming.

Other matches may be false alarms. Treat them as suspect spans and reBLAST the hit span plus 10 Kbp of flanking sequence on each side against nr, HTGS, related and unrelated chromosomes (as described below).

2. Process contaminant matches from 2

   Flag all adaptor spans for trimming.

3. Process mitochondrion matches from 3

   If the total coverage of mitochondrial matches from (3) is >75% of the sequence length then flag the sequence as being mitochindrial sequence to be excluded.

4. Process unrelated chromosome matches from 5

   Ignore any matches to chromosomes from unrelated organisms that lie with a region identified as being rRNA genes from (4) (the spans matched in 4 plus 100 bases on both sides). These are likely to be false matches.

   Treat other matched spans as suspect and reBLAST the hit span plus 10 Kbp of flanking sequence on each side against nr, HTGS, related and unrelated chromosomes

5. ReBLAST against nr, HTGS, related and unrelated chromosomes

   Spans identified a contamination suspects in the first pass, plus 10 Kbp of flanking sequence on each side (up to the end of the contig), are BLASTed against nr, HTGS, related and unrelated chromosomes to generate additional data for calling contaminants to be excluded or trimmed.

6. Databases

   chromosome databases (a) to (i) from (5) above.

   `ftp://ftp.ncbi.nlm.nih.gov/blast/db/nt.*.tar.gz`

   `ftp://ftp.ncbi.nlm.nih.gov/blast/db/htgs.*.tar.gz`

7. Execution

   The suspect spans are BLASTed against each of the 10 databases.

   ```
   blastn -query _suspect_spans_plus_flanks_ -db _reblast_db_ -task megablast -dust y
   ```

8. Processing the reBLAST matches

   Automatically exclude sequence contigs that meet all the following criteria:

   ```
   >60% of length covered with foreign hits, or less than 200 bp that are NOT covered

   Each contributing hits must be 100 bp or longer with identity >= 98%
   ```

   The best match to chromosomes from unrelated organisms is longer than the best match to chromosomes from the related organism group

   Some of the other hits may be reviewed manually.

1.2 TODO Some vecscreen results have valid intervals at the begining and end of the sequences!

We solve that by

- keeping the intervals in three dictionaries for each contig:

  contig->begin [][][] for [].end < 100 contig->end [][][] for the rest contig->center [][][] for [].begin > 100

We remove the sequence by taking the max over all element (Beginning) or min (End). We do not touch the intervals that are located in the middle of the sequences.

1.3 TODO TODO: add more cases where vector are removed

Forking and feedback encouraged

1.4 More Resources

- vecscreen `https://www.ncbi.nlm.nih.gov/tools/vecscreen/`, `https://www.ncbi.nlm.nih.gov/tools/vecscreen/about/`

1.5 Usage

```
remoVecSec.py -g ‘pwd‘/test/MA6037.fasta -d UniVec > test/MA6037.fasta.cor
```

1.6 License

This is free and unencumbered software released into the public domain.

Anyone is free to copy, modify, publish, use, compile, sell, or distribute this software, either in source code form or as a compiled binary, for any purpose, commercial or non-commercial, and by any means.