Algorithmic Based Automatic Contiguation of Assembled Shotgun sequences (ABACAS)

Overview:
ABACAS is intended to rapidly contiguate (align, order, orientate) , visualize and design primers to close gaps on shotgun assembled contigs based on a reference sequence. It used MUMmer to find alignment positions and identify syntenies of assembly contigs against the reference. The output is then processed to generate a pseudomolecule taking overlaping contigs and gaps in to account. MUMmer's alignment generating programs, Nucmer and Promer are used followed by the 'delta-filter' utility function. Users could also run tblastx on contigs that are not used to generate the pseudomolecule. If the blast search results in mapping of extra contigs, finishers can use the visualization tool to easily modify ordering of contigs on the pseudomolecule. Gaps in the pseudololecule are represented by "N"s. ABACAS could automatically extract gaps on the pseudomolecule and generate primer oligos for gap closure using Primer3. Uniqueness of primer sets is checked by running a sensitive NUCmer alignment. If a quality file (contig_name.qual) exists in the working directory, users will be asked for a minimum quality cutoff while picking primers.

Requirement:
ABACAS requires MUMmer to be installed in the working path for ordering and orienting of contigs. The Arthemis Comparision Tool (ACT) should be downloaded for visualizing scaffolding of contigs. Primer design part of the programme requires Primer3. Optionally, BLASTALL is required in order to run tblastx on the contigs that are not mapped using Nucmer or Promer.

USAGE

```
abacas.pl  -r <reference file: single fasta> -q <query sequence file: fasta>
           -p <nucmer/promer> [Options]


    -r         reference sequence in a single fasta file
    -q         contigs in multi-fasta format
    -p         MUMmer program to use: 'nucmer' or 'promer'

    [OPTIONS]
    -h              print help information
    -d         0/1  use default nucmer/promer parameters [default 0]
    -s         int  minimum length of exact matching word (nucmer default)
```

```
    -m         0/1  print ordered contigs to file in multifasta
format [default 0]
    -b         0/1   print contigs in bin to file [default 0]
    -N         0/1   print a pseudomolecule without "N"s [default 0]
    -i         int   mimimum percent identity [default 40]
    -v    int   mimimum contig coverage [default 40]
    -V         int      minimum contig coverage difference [default
1]
    -l         int      minimum contig length [default 1]
    -t         0/1  run tblastx on contigs that are not mapped
[default 0]
    -g         string (file name)     print gaps on reference to file
name
    -a         0/1       append contigs in bin to the pseudomolecule
    -o         prefix    output files will have this prefix
    -P         0/1            pick primer sets to close gaps
    -f         int            number of flanking bases on either side
of a gap for primer design (default 350)
```

Input:

Two fasta files containing the reference and query (contigs) sequences are required. The reference file should be in a single fasta format for speedy contig ordering and orientation.

Outout:

Running the script with default options will generate the following files:

1. ordered and orientated sequence file (reference_query.fasta)
2. a feature file (reference_query.tab)
3. a bin file thatcontains contigs that are not used (reference_query.bin)
4. a comparision file (reference_query.crunch)
5. a file with gap information (reference_query.gaps)
6. a file that contains information on contigs that have a mapping information but could not be used in the ordering (unused_contigs.out)
7. a feature file to view contigs with ambiguous mapping (reference.notMapped.contigs.tab). This file should be uploaded on the reference side of ACT view.
8. a file that shows how repetitive the reference genome is (reference.Repeats.plot).

   Files 7 & 8 should be uploaded on the reference side of ACT view.

Please note that contigs in the '.fasta' file will be reverse complemented if they are found to map on the reverse strand. However, the ACT view shows the initial orientation of these contigs i.e. will be shown on the reverse strand. If you write a fasta file of the pseudomolecule from ACT, the resulting sequence will be a set of ordered contigs (the orientation will not change). It is therefore recommended to use the '.fasta' pseudomolecule file automatically generated for further investigation.

Optional Output files:

It is also possible to generate other files including:

1. A list of ordered and orientated contigs in a multi-fasta format (-m 1)
2. A pseudomolecule with all unmapepd contigs appended to the end for reordering (-a 1)
3. A pseudomolecule where the gaps are not padded with N (-N 1)
4. A multi-fasta file of all unmapped contigs (-b 1)
5. A multi-fasta file of regions on the reference that correspond to gaps on pseudomolecule (-g file_name)
6. A list of sense and antisense primer sets in separate files
7. A list of locations where sense and antisense primers are found in two separate files
8. A standard primer3 output summary file with a detailed information on oligos.

Options:

```
-d     default 0 i.e. increase mapping sensitivity by using 'all anchor matches
regardless of their uniqueness during mapping' ( i.e. --maxmatch)
       -d  1 runs default NUCmer or PROmer parameters

-m     default -m 0
       *this option is helpful if the user wants to further investigate the
ordering using other alignment algorithms such as blast.
       -m 1 to print ordered and orientated contigs to file

-b    default -b 0
       *contigs that are not used in generating the pseudomolecule will be placed
        in a '.bin' file. This file only contains contig names.
      -b 1 will print those contigs in the bin to multi-fasta file. This option
       must be used with -t 1 in order to run 'tblastx' on contigs in bin

-N    default -N 0
      *ABACAS produces a pseudomolecule ('.fasta' file) and fills gaps with 'N's.
       This option will remove "N"s.
       -N 1 will generate another pseudomolecule without 'N's

-i    default 40
       *minimum percent identity could vary from 0 to 100 depending on the
      closeness of the two genomes. Choosing a smaller value will pull in more
```

```
      contigs and vice versa

-v    default 40
      *minimum contig coverage: set a value between 0 and 100

-V    default 1
      *minimum contig coverage difference. Use -V 0 to place contigs randomly to
       one of the positions (in cases where a contig maps to multiple places)

-l    default 100
       * contigs below this cutoff will not be used

-t    default 0
      *runs tblastx on contigs that are not used to generate the pseudomolecule.
      -t 1 will run blastall on contigs in the .bin file

-g     file_name
       * will print sequences of the reference that correspond with gaps on the
       pseudomolecule in a multi-fasta format

-a    default -a 0
       * this option will append contigs that re not used to the end of the
      pseudomolecule. Contigs will then be easily manipulated and re-ordered using
      ACT's graphical interface

-o    prefix (string)
       *output files will have this prefix

-P    default 0
      * -P 1 will pick primer oligos to close gaps

-f    default 350
      *number of flanking bases on either side of a gap for primer design (default
350)
```

Colour code:

The feature (file 2 from default output section) file has the following colour codes:

```
Dark blue (4):     contigs with forward orientation
Dark green (3):    contigs with reverse orientations
Sky blue (5):      contigs that overlap with the next contig
Yellow (7):        contigs that have no hit (only added to the pseudomolecule if
'-a 1' is used)
```