**KMerFineCompare ("KFC")**: a python analysis of DNA read data to look for differentially-expressed sequences as k-mers (avoiding any requirement for alignment of homology to a reference genome).

Minimal Command Line:

**python KmerFineCompare\_ar03\_101522.py up=PositiveFastQFile down=NegativeFastQFile**

**I. Essential File Inputs for KMerFineCompare:**

**up=<FileName>**. Illumina read file for the "positive" sample: fastq/fasta file that can be .gz compressed.

**down=<FileName>**. Illumina read file for the "negative" sample: fastq/fasta file that can be .gz compressed.

Bulk Compare Syntax: up=UpFile1 up=UpFile2 down=DownFile1 down=DownFile2

Yields k-mers where minimum fraction in all UpFiles is compared to maximum fraction in two DownFiles.

Individual Compare Syntax: up=UpFile1 down=DownFile1 up=UpFile2 down=DownFile2

Yields k-mers where each UpFile is compared to the neighboring DownFile in the list.

**II. Setting Additional Parameters (optional):**

Settings can be assigned in the command line in two ways (Parameter names are case-insensitive):

a. **Using equals signs**: e.g., Klen=25 FDR=2000

b. **Using *-parameter value* syntax:** e.g, *-Klen 25 -FDR 2000*

*Importantly note that file groups can be specified on the command line as a comma delimited list, with that list added to if needed through the command line argument MyFileList += onefile,anotherfile, etc.*

1. **Klen** (K-mer length). This is the word-length that the program uses to compare sequences. Generally a word length of 20 is sufficient to position a majority of unique sequences, but longer word lengths can be useful to position sequences with somewhat degererate character or in partly repetitive regions. Very long k-mer lengths will lose some assignments due to either errors or rearrangements. The default setting for Klen1 is 21 but it can be set as high as 32 with no reworking of the program.

2. **ReadsPerSample** How many reads to process for each sample? The default is to take all reads found in each file. This can also be set to a pre-assigned value (e.g. 1000000).

3. **FoldMinimum,Regularization** (Sets the minimum fold-change and other criteria to report a difference). The program uses a two-stage process to assess fold difference, with a workflow that first looks for a minimum raw-ratio difference and (for k-mers that meet that criteria) calculates a minimum fold-difference consistent with the numerical observations and assumptions of false discovery ratio. FoldMinimum is the underlying fold-difference that the program will lookfor. Setting the FoldMinimum at 2.0 requires a minimum 2-fold change for any given k-mer to report the difference. To ensure significance of diffferences in a "first pass", a fixed "regularization" is added to all k-mer counts. As an example, setting FoldMinimum to 2.0 and Regularization to 3.0 will insist that (UpCount+3.0)/(DownCount+3.0)>2.0 for considering a k-mer as potentially differentially expressed. Default values for FoldMinimum and Regularization respectively are 5.0 and 1.0.

4. **FDR** (False discovery rate). This controls the threshold set for a statistical test for differential expression candidates that is performed on candidates that emerge from from testing the raw (regularized) ratios. The statistical test here is handled by a separate program in the package (BayesRatioAH), which employs a maximum-likelihood Bayes approach to identifying differentially present k-mers. Documentation for that program is included. Note that a bonferroni-type correction is applied automatically to the FDR. The probability test applied by BayesRatio entails assuming a difference of exactly FoldMinimum and ensuring that if this was the case only a number of reads corresponding to FDR would be observed. The actual FDR for BayesRatio is number-of-kmers/klen.

5. **TetritisFiles,TetritisKmerLen** Tetritisfile is file (generally FastA format, can be gzippe) of linkers and other sequences that may be present in sequence datasets. Any k-mer longer then TetritisKmerLen is ignored or differential expression as these are generally artefactual differences representing differences in library production or differences in barcodes. Default is a file Tetritis.fa included with this distribution.

6. **JunkFiles,JunkKmerLen** Junk is file (generally FastA format, can be gzippe) of known contaminant sequences that may be present in sequence datasets. Any k-mer longer then JunkKmerLen is ignored or differential expression as these are likely due to differential contamination between samples. Default (for C. elegans) is the file OP50.fa which is an (unfortunately not completely accurate) assembly of the OP50 bacteria usually used as a C. elegans food source..

7. **RefFile** A reference file (FastA or gzipped FastA format) to align sequences (doesn't affect reporting or assembly but does report homologies. Two reference files are inlucded from C. elegans RDNA and a tRNA file and the default is to use these for annotation.

8. **ATrim**, **TTrim**. Set this to True to trim A (or T) residues from the 3' end of the read. Setting to an integer value trims at internal runs of As (so with ATrim=4, everything past the first run of 4 As will be skipped). This is for oligo-A tailed libraries and otherwise should be set to "False". Some sequences with runs of As will be lost with this setting (defaults: zero [no trimming]).

9. **FTrim, ETrim**. Set FTrim to any integer n to trim the first n bases from each read, set ETrim to any integer to trim the last n bases.

10.**JoinKLen**. Set this to the minimum overlap length in joining k-mers into longer segments (default is 11)

11. **OutputFileBase**. A name that will be used as the "Base name" for the output files. Default is for KFC to catenate input file names.

12. **ReportGranularity**. How often to report progress during main part of running program (assembling dictionary)

13. **Linker**. Look for a linker sequence at the 3' end of the read. Generally only the first few bases of the linker are provided in the command line and KFC will look for this sequence in each read starting with the 3' end. Sequences without linker are not analyzed in this case.

14. **RequireLinker**. Set to true insists that the linker tag above be present or the sequence is ignored. Default is False to catch a few reads with longer inserts of variant linkers.

**III. Output.**

**1. A file with the prefix "Concsnsi"**. This is a fastA file with the consensus sequences of short assemblies of any differentially expressed k-mers.

**2. A file with the prefix "CaughtReads"**. This shows the relevant reads supporting the differential expression of k-mers. The reads are A-trimmed is ATrim is set to True in the command line (default). Following each relevant read sequence are counts of positive reads in the down and up samples respectively

**3. A file with the prefix "Summary"**. Shows some features of the k-mer populations including total number in each sample and coincidence frequencies (probabilty of two arbitrary k-mers chosen from the same sequence matching each other).

**4. A cumulative KFCLog.txt file**. This shows descreibes each running of the program and saves both command line parameters and underlying source code

**IV. Packing List:**

A. The python code.

B. FastA files for OP50 (OP50.fa) and Some Illumina Linkers (IlluminaTetritis.fa). The program will complain and quit if these two files are not in the current directory.

C. Reference files for use in worm default alignments (ce10tRNAs.fa [Lowe ref] and cerDNA.fa [Ellis, Coulson, Sulston, 1986])

D. This documentation File (KmerFineCompareDocs##.docx).

E. The additional script file BayesRatioAH.py

**VII. Bug List and History:**

**KFC** Code written by A. Fire with inspiration from L. Hansen, M. Shoura, L. Wahba, S. Arur, and K. Artiles, Department of Pathology, Stanford University School of Medicine. Distributed with no guarantees of quality, efficacy, accuracy, or anything in life, let alone the functioning of software (**Copyright 2019-2022 Stanford**)

**10-15-22**

**Version AR** is a slightly refactored work-alike for originally distributed version AQ, with the adjustment of code to (i) run with Python3 (or pypy3), (ii) avoid the use of numpy (no longer required). This version seems substantially faster but may have slightly less efficient memory use than previous versions with Python2.