Volumetric DNA microscopy
Please note: this documentation is a work-in-progress

GSE (image inference)

Input file:
link_assoc.txt (in separate directory)
Each row corresponds to a distinct UMI-UMI association
Has columns:

1. UEI "type" (ignored in current software version)
2. Source data "type I" UMI (can have any index enumeration, as long as unique among other type I UMIs)
3. Source data "type II" UMI (can have any index enumeration, as long as unique among other type II UMIs)
4. Number of UEIs for this particular association

Command line (example provided in numbers.sh):
python3 CODE_PATH/main.py gse -path DATA_PATH// -max_eig_cuts 5 -inference_dim 2 -inference_eignum 50 -final_eignum 100 -iterations 1 -ncpus 5

Arguments:

1. **max_eig_cuts: The number of distinct tessellations to be done on the data set**
2. **inference_dim: The number of dimensions being modeled/used in embedding**
3. **inference_eignum: The number of "raw" data eigenvectors to generate**
4. **final_eignum: The number of GSE eigenvectors to calculate for the final gradient descent**
5. **iterations: The number of GSE iterations**
6. **ncpus: The number of cpus to use (for parallelization)**

Optional arguments (all for data sets of size >>1e4; subset-GSE, whereby data subsets are analyzed, and merged through PCA and a final gradient descent):

7. **init_min_contig (used in manuscript: 10000)**
8. **num_subsets (used in manuscript: 25)**

Sequence analysis
The sequence analysis module of the VDNAmic pipeline borrows heavily from the original dnamic pipeline (https://github.com/jaweinst/dnamic)

The following settings pre-assume an amplicon format equivalent to those described in Fig S1 of the volumetric DNA microscopy manuscript.

Command line (example provided in seq-UEI.sh and seq-cDNA.sh):

```
gunzip RAWDATA_PATH/*.fastq.gz
python3 CODE_PATH/main.py lib UEI-directory/
python3 CODE_PATH/main.py lib cDNA-directory/
```

UEI-directory lib.settings:

```
-source_for ..//RAWDATA_PATH//i31sub_R1.fastq
-source_rev ..//RAWDATA_PATH//i31sub_R2.fastq
-seqform_rev U_GCTNWWNNNNWWSWNNNWSWNNNWSWWNNNNWWNTGA_2:39
-seqform_for
U_NWNNNNWWSWNNNWSWNNNWSWWNNNWNGCG_0:31|U_AGGNWWNNNNWWSWNNNWSWNNNWSWWN
NNNWWNAGC_39:76
-seqform_for
U_WNNNNWNSWNWNWNWSNNNSWNNWWWNNATG_0:31|U_AGGNWWNNNNWWSWNNNWSWNNNWSWWN
NNNWWNAGC_41:78
-min_mean_qual 30
-filter_umi0_amp_len 25
-inference_dim 3
-min_uei_per_umi 2
-min_reads_per_assoc 2
-min_assoc_per_umi 1
-max_eig_cuts 10
-inference_eignum 50
-u0 *,*,1
-u1 *,*,2
-u2 *,*,0
```

cDNA-directory lib.settings

```
-source_for ..//RAWDATA_PATH//i6sub_R1.fastq
-source_rev ..//RAWDATA_PATH//i6sub_R2.fastq
-seqform_for A_29:
-seqform_rev U_GCTNWWNNNNWWSWNNNWSWNNNWSWWNNNNWWNCCT_2:39
-seqform_rev U_GCTNWWNNNNWWSWNNNWSWNNNWSWWNNNNWWNTGA_2:39
-amplicon_terminate
GTTCAGACGTGT,ACACGTCTGAAC,CCGATCTTAGCT,AGCTAAGATCGG,TGACTCTCAGTG,CACT
GAGAGTCA,ACTATAGCAAAT,ATTTGCTATAGT,GATCTCTAGCTA,TAGCTAGAGATC,GCTCTTCC
```

```
GATC,GATCGGAAGAGC,CCTACCACTTAC,GTAAGTGGTAGG,GCAATACGACCA,TGGTCGTATTGC
,ACACTCTTTCCC,GGGAAAGAGTGT,TATAAGAGACAG,CTGTCTCTTATA,ATGTGTAAATCC,GGA
TTTACACAT,ACACTGAGAGTC,GACTCTCAGTGT,CGTAAGTGGTAG,CTACCACTTACG,TACCACT
TACGC,GCGTAAGTGGTA,CAGACGTGTGCT,AGCACACGTCTG,CCACTTACGCAT,ATGCGTAAGTG
G,GTATAAGAGACA,TGTCTCTTATAC,AGTGTGATGGCA,TGCCATCACACT,AGTGTCCAACCT,AG
GTTGGACACT
-min_mean_qual 30
-filter_umi0_amp_len 25
-filter_umi1_amp_len 25
-u0 *,0,0:revcomp
-u1 *,1,0
-a0 *,0,0
-a1 *,1,0
-STARindexdir dr_index // STAR-index directory (optional)
-gtffile  // GTF-file path (optional)
-uei_matchfilepath // UEI-match file (optional; will be low overlap
for under-sampled data, as is the case in the sample data set)
```