**PALADIN File and Test Organization**

**Data Sets**

**Root:** brain:~/local\_projects/paladin/data\_sets

**Description**

This directory contains:

1. References
   1. 6 MCBS913 data sets (fasta and GFF)
   2. The NT and AA translations of both versions of the UniProt DBs (full and filtered w/o the 6 datasets above)
2. Reads
   1. PE sets for each of the 6 MCBS913 data sets
   2. SE of the previous 6 concatenated, for simulated metagenomic reads (metareads.fq)
   3. PE of a real metagenomic set (Jun\_MW4)

All testing makes use of symbolic links to these files, with read mapping related files (PAC/BWT/SA) stored in the individual test's directory, and not with the dataset.

**Seed Testing**

**Root:** brain:~/local\_projects/paladin/test-seed\_length

**Description**

Testing the relationship between read mapped percentages and seed length

**Instructions**

1. genIndices.sh will index all references.
2. alignSeed.sh will run the testing for all single genome read sets (1-3 below)
3. alignMetagenome.sh will run the testing for metagenome read (4 below)

**Notes**

Each subdirectory under the root directory is a numeral identifying the read set being run against the reference. Each of 3 references is also stored within each subdirectory. Outputs will be in each directory in the form of samstat files which should be compiled with sam2csv script into a single CSV file. Values are as follows (Reads, References):

1. AcidovoraxAvenaeATCC19860
   1. Acidovorax\_citrulli\_AAC00\_1\_uid58429\_NC\_008752 (0.4%)
   2. Variovorax\_paradoxus\_EPS\_uid62107\_NC\_014931 (15.3%)
   3. Thiomonas\_intermedia\_K12\_uid48825\_NC\_014153 (31.1%)
2. EscherichiaColiStrK-12SubstrMG1655
   1. Escherichia\_coli\_042\_uid161985\_NC\_017626 (0.5%)
   2. Yersinia\_pestis\_A1122\_uid158119\_NC\_017168 (15.4%)
   3. Haemophilus\_parainfluenzae\_T3T1\_uid72801\_NC\_015964 (31%)
3. StaphylococcusEpidermidisATCC12228
   1. Staphylococcus\_pasteuri\_SP1\_NC\_022737 (3.8%)
   2. Macrococcus\_caseolyticus\_JCSC5402\_NC\_011995 (17%)
   3. Bacillus\_cellulosilyticus\_DSM2522\_NC\_014829 (N/A%)
4. Metagenome
   1. Iterates through directories/sets above

**ORF Length Testing**

**Root:** brain:~/local\_projects/paladin/test-orf\_length

**Description**

Testing the relationship between read mapped percentages and minimum ORF length filtering. **NOTE - this test is likely deprecated with new algorithm variants.**

**Instructions**

1. genIndices.sh will index all references.
2. alignOrfs.sh will run the testing for all single genome read sets (1-3 below)
3. alignMetagenome.sh will run the testing for metagenome read (4 below)

**Notes**

Each subdirectory under the root directory is a numeral identifying the read set being run against the reference. Each of 3 references is also stored within each subdirectory. Outputs will be in each directory in the form of samstat files which should be compiled with sam2csv script into a single CSV file. Values are as follows (Reads, References):

1. AcidovoraxAvenaeATCC19860
   1. Acidovorax\_citrulli\_AAC00\_1\_uid58429\_NC\_008752 (0.4%)
   2. Variovorax\_paradoxus\_EPS\_uid62107\_NC\_014931 (15.3%)
   3. Thiomonas\_intermedia\_K12\_uid48825\_NC\_014153 (31.1%)
2. EscherichiaColiStrK-12SubstrMG1655
   1. Escherichia\_coli\_042\_uid161985\_NC\_017626 (0.5%)
   2. Yersinia\_pestis\_A1122\_uid158119\_NC\_017168 (15.4%)
   3. Haemophilus\_parainfluenzae\_T3T1\_uid72801\_NC\_015964 (31%)
3. StaphylococcusEpidermidisATCC12228
   1. Staphylococcus\_pasteuri\_SP1\_NC\_022737 (3.8%)
   2. Macrococcus\_caseolyticus\_JCSC5402\_NC\_011995 (17%)
   3. Bacillus\_cellulosilyticus\_DSM2522\_NC\_014829 (N/A%)
4. Metagenome
   1. Iterates through directories/sets above

**No Hidden Stop Count per Frame Testing**

**Root:** brain:~/local\_projects/paladin/test-no\_hidden\_stop\_count

**Description**

Via PALADIN variant 1, index all 6 frames for the combined MCBS913 dataset, as well as the UniProt DB. The frame number is used as the first character in each sequence header of each AA sequence, with 0 being the correctly aligned read frame for the protein in question. Then the number of frames with no hidden stop codons are counted

**Instructions**

1. Run a PALADIN index using the all 6 frame index variant
2. Run ~/repos/paladin/Scripts/countNoHiddenStop.py file.pro startLength, endLength, stepLength
3. Redirect to CSV file, will contain column headings

**Notes**

The results of this test can be found in "No Hidden Stop Counts.xlsx"

**Order of Likelihood of Stop Codons by Frame**

**Root:** brain:~/local\_projects/paladin/test-stop

**Description**

Via PALADIN variant 1, index all 6 frames for the combined MCBS913 dataset, as well as the UniProt DB. The frame number is used as the first character in each sequence header of each AA sequence, with 0 being the correctly aligned read frame for the protein in question. Then the likelihood of stop codons per frame is reported in a matrix view

**Instructions**

1. Run a PALADIN index using the all 6 frame index variant
2. Run ~/repos/paladin/Scripts/countNoHiddenStop.py file.pro
3. Redirect to CSV file

**Notes**

The results of this test can be found in "Stop Stats.xlsx"

**ALL ALIGNMENT TESTS**

**Root:** brain:~/local\_projects/paladin/test-alignXXX

**Description**

All alignment tests are run with the same general steps

**Instructions**

1. Run a PALADIN index using the appropriate variant
2. Run alignment using the appropriate variant, redirected into SAM file
3. Convert SAM to BAM
4. Obtain flagstats
5. Run "~/repos/paladin/Scripts/listMappedCDS.py file.sam > file.cds" (This takes many hours, script is not optimized yet)
6. "cat file.cds | sort | uniq | wc -l" to see the number of CDS entries corresponding to reads that were successfully mapped
7. Repeat for 1-6 for additional seed sizes or different references

**Notes**

The results of these tests can be found in "PALADIN Test Stats.xlsx"

1. Align1 - PALADIN variant 1, MCBS913 metagenome reads, UniProt DB (full and filtered), seed length 9 and 11
2. BWA - BWA, MCBS913 metagenome reads, UniProt DB (full and filtered)