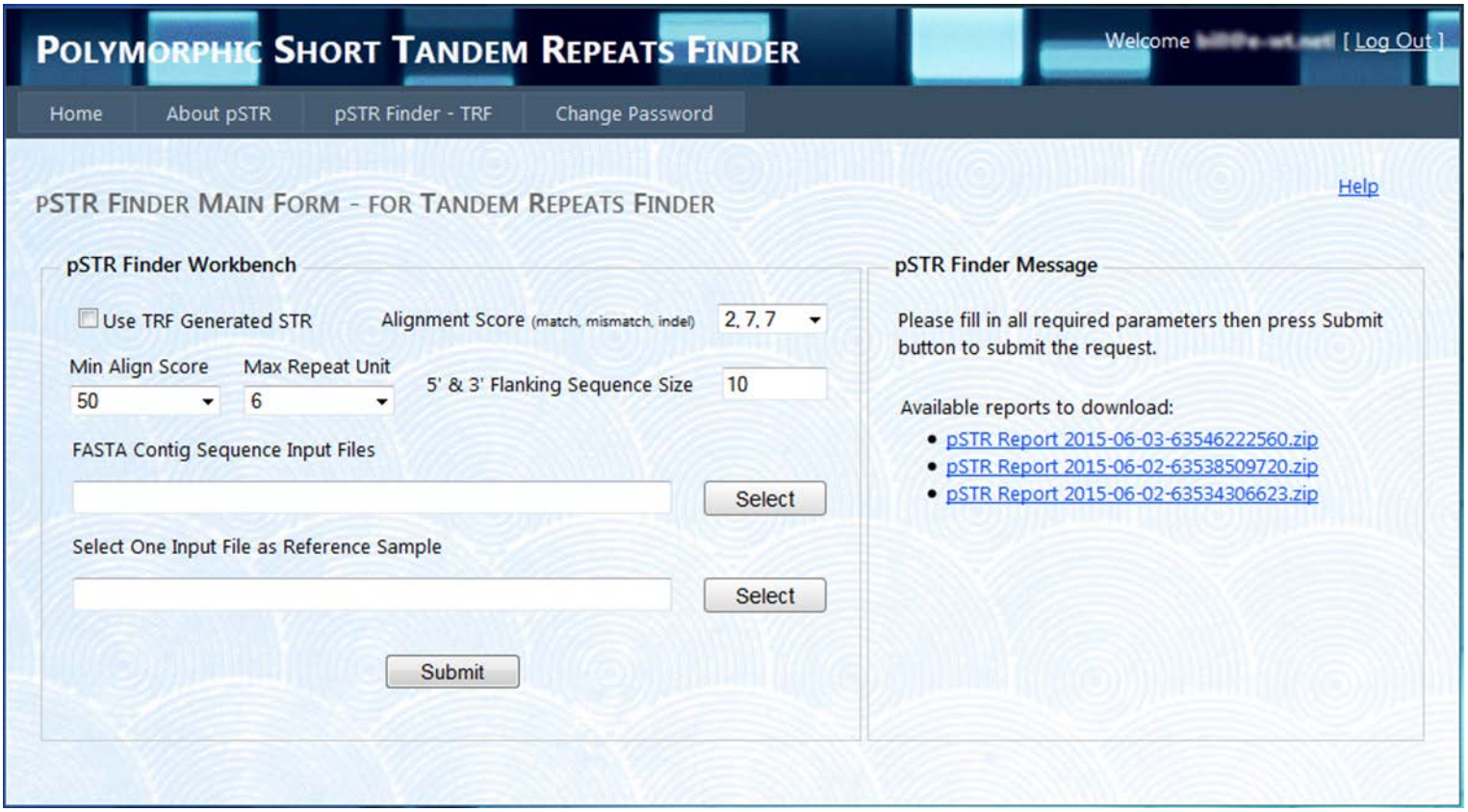
How to cite Tandem Repeats Finder (TRF)

Author: G. Benson, Tandem repeats finder: a program to analyze DNA sequences, Nucleic Acids Research (1999), Vol. 27, No. 2, pp. 573-580.
Site: http://tandem.bu.edu/trf/trf.html

TRF 4.09 has been tested and integrated with pSTR.

pSTR Default Screen



The pre-determined default options should be sufficient for most users. Below explains how parameters can be specified. Most parameters are adopted from TRF as all sample genome data files will use TRF to discover and generate STR sequences for each genome data file.

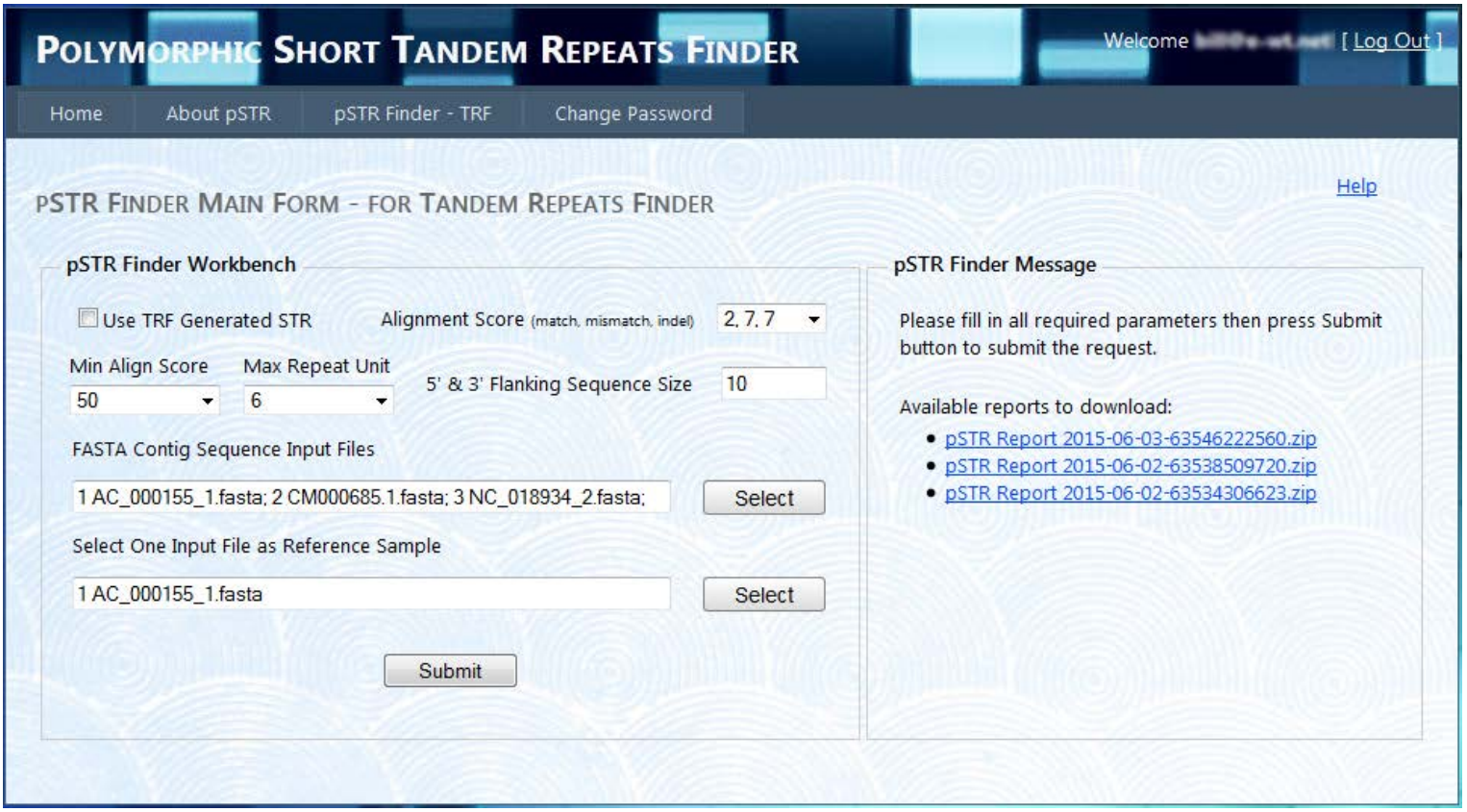
1. Alignment Score – Weights for match, mismatch and indels (insert/delete). Lower weights allow alignments with more mismatches and indels. Match weight is +2 in all options here. Mismatch and indel weights (interpreted as negative numbers) are either 3, 5, or 7. A 3 is more permissive and a 7 less permissive of these types of alignments choices.
2. Minimum Alignment Score – The alignment score must meet or exceed this value for the repeat to be reported.
3. Maximum Repeat Unit – The repeat unit (period size) must not be larger than this value for the repeat to be reported. Repeat unit is TRF's best guess at the pattern size of the tandem repeat. TRF will find all repeats with repeat unit between 1 and 2000.
4. 5' & 3' Flanking Sequence Size – The default size is 10 bp. pSTR will designate 10 bp from both ends of the sequences as respective 5' and 3' flanking sequences. Recall all sample genome data files will use TRF to generate STR sequences for each genome first. For each pair of STR of two comparing samples, pSTR compares the 5' and 3' of both STR. When a match is found and if both comparing STR have the same copy number then both STR are identical, otherwise polymorphic. User may change the size to see fit.
5. Use TRF Generated STR – If it is checked, pSTR will expect user has managed to use TRF standalone application to generate STR sequences for all the sample genome data files. Please see below for more details. The default option is not checked. This means that pSTR will invoke TRF to generate the STR sequences in the background.

FASTA input file by default will have file extension .fasta. FASTA file requires the first line to be the header line starting with a ">" sign then followed by a sequence identification code. The sequence data will follow starting the second line.

Please note that we require one sample FASTA file containing only one header line and one sequence stream. In addition, user must select at least 2 FASTA files as the input and designate one of the selected input files as the reference sample.

Submit a Request

The screen below demonstrates 3 human Chromosome X samples "1 AC_000155_1.fasta", "2 CM000685.1.fasta" and "3 NC_018934_2.fasta" selected and the designated reference sample is "1 AC_000155_1.fasta". All default options are used.



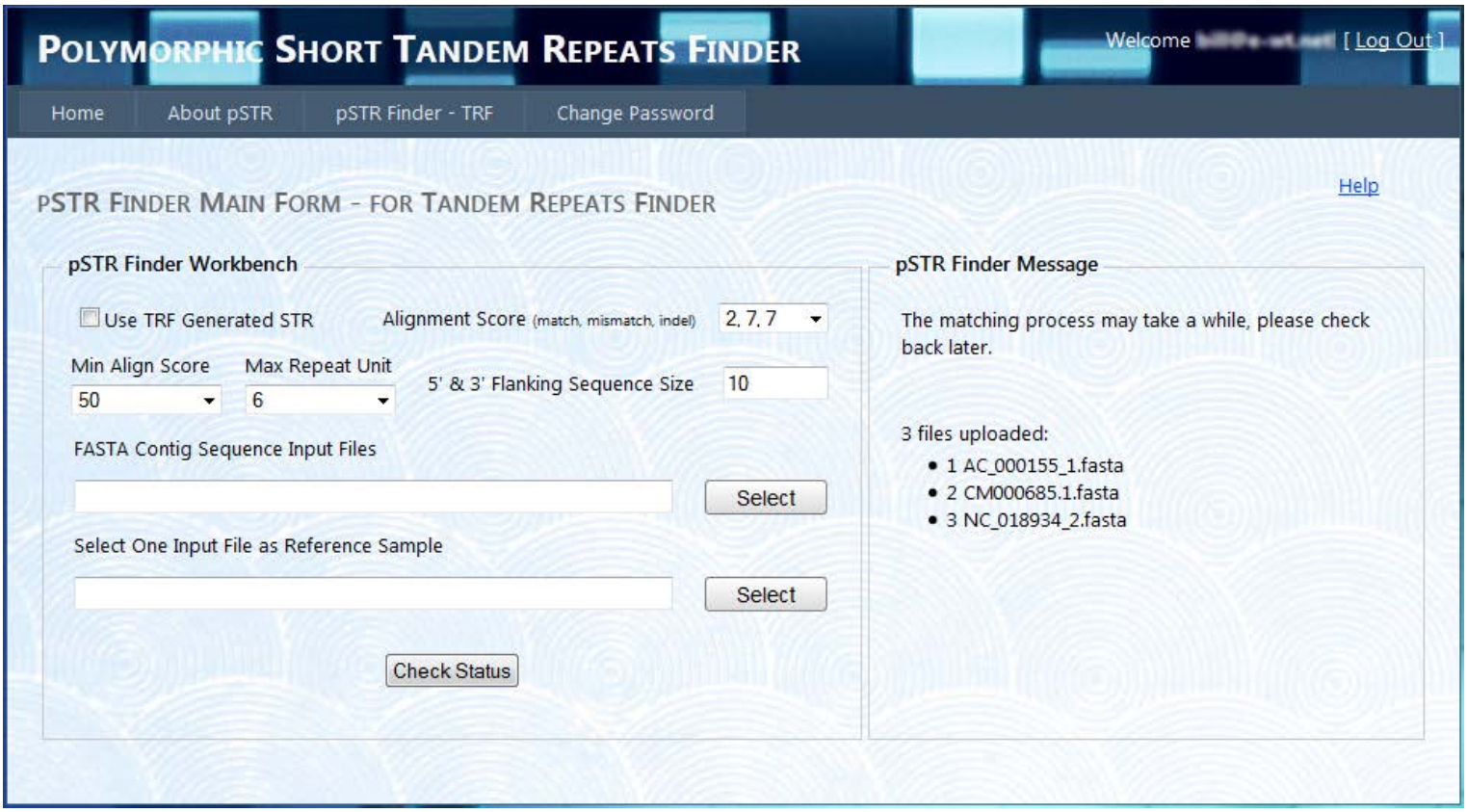
User is now ready to submit the request. Just press the Submit button and the sample files will be uploaded to the server. The upload time varies based on the data size and internet speed.

Notice from below screen after the upload completed the Submit button changed to Check Status and the upload status appears in the pSTR Finder Message panel.

Applications handling large amount of data like DNS sequences may be very time and resource consuming. In order to better protect our limited system resource, we only allow a user to submit one request at a time. User needs to check the status to ensure the current request has completed then user may submit the next request.

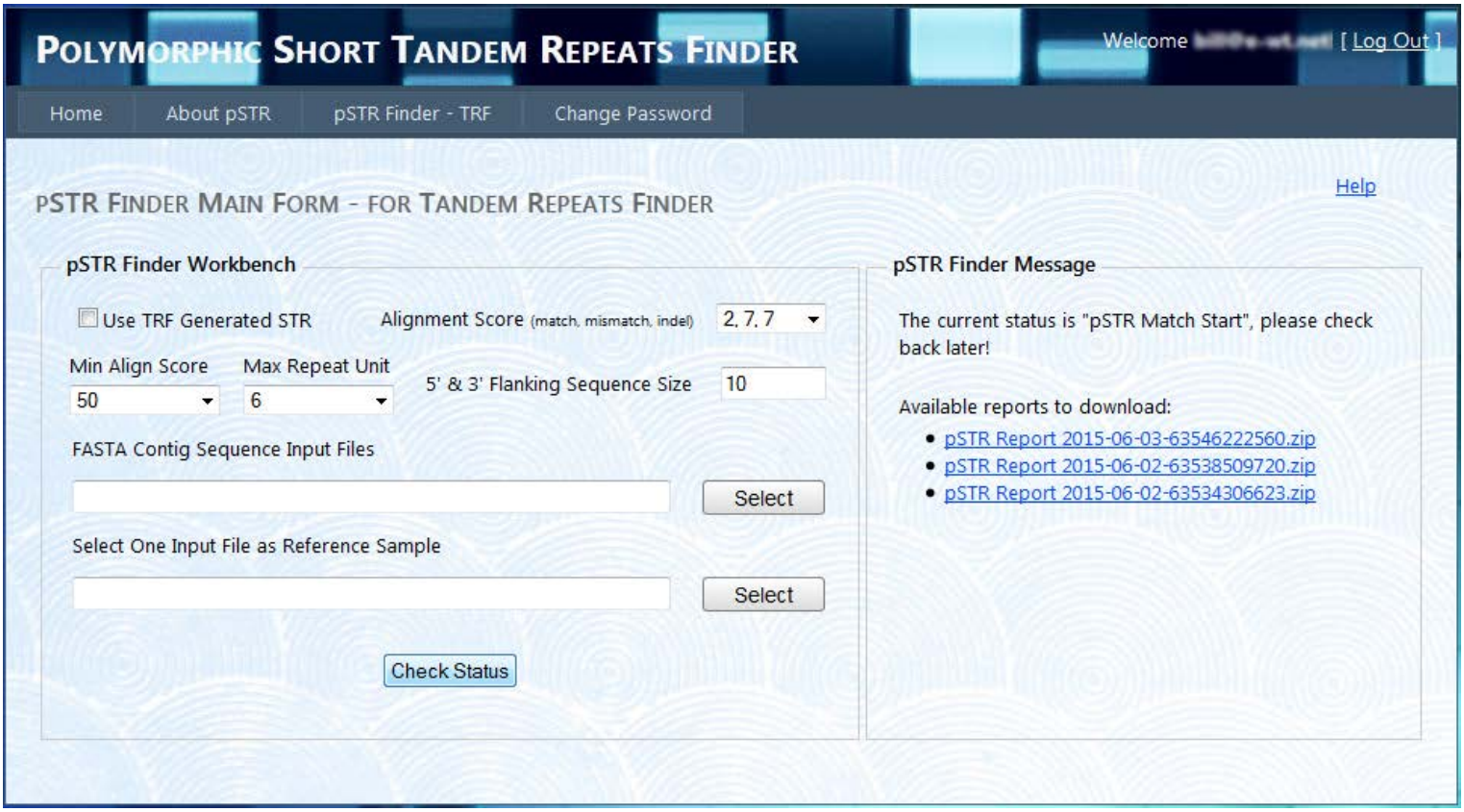
Sample Genome Data Files Uploaded

Below screen demonstrates that user has selected and successfully uploaded 3 sample data files to the server.



Check Status

Press Check Status to retrieve the pSTR status.

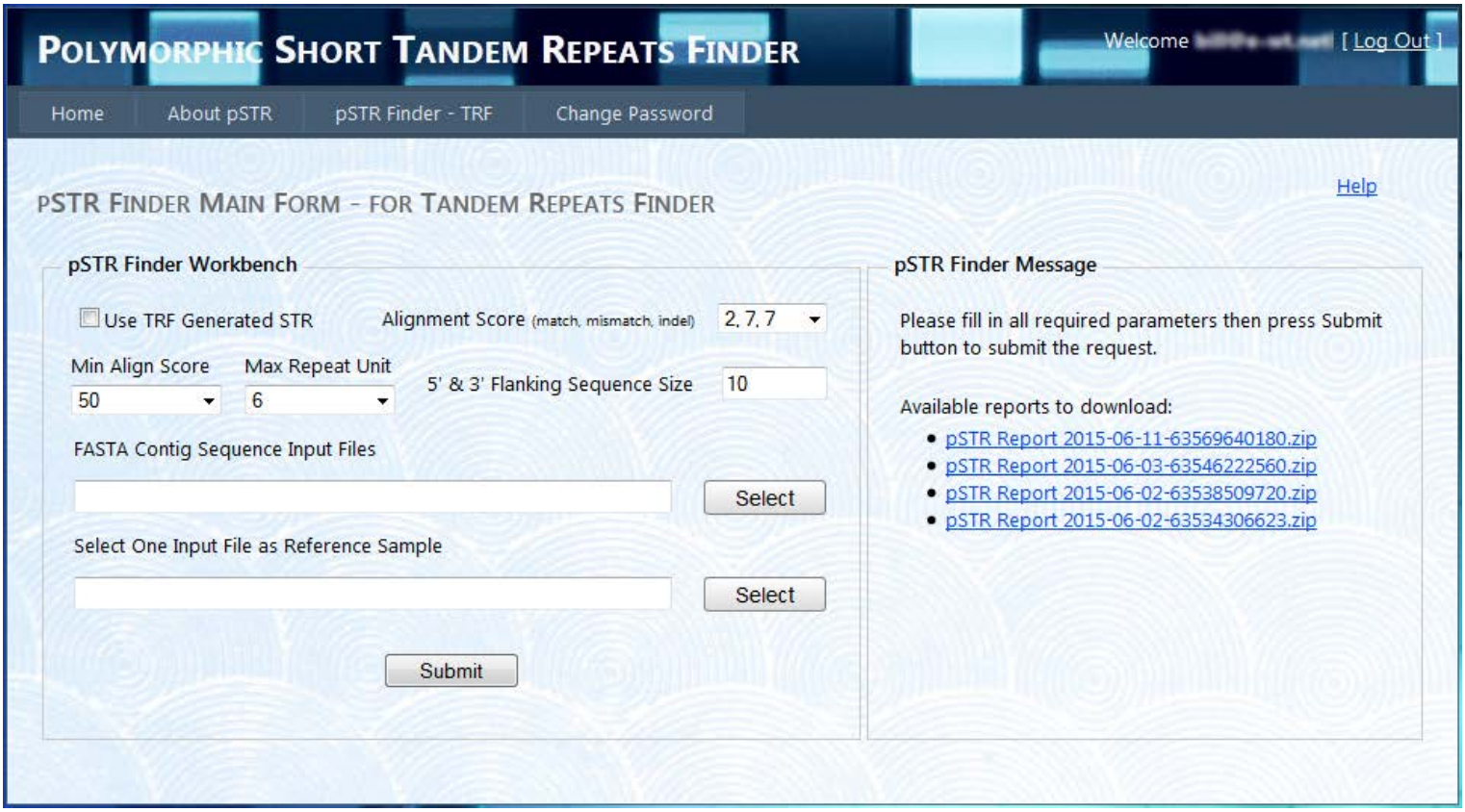


Below is the list of pSTR status designated a specific process being handled by pSTR:

1. pSTR Match Start – pSTR will start comparing all sample STR sequences once all sample data files uploaded to server.
2. STR Gen Complete – pSTR invokes TRF to generate STR sequences for each sample data file. When user checks to see this status, pSTR has completed generating STR sequences.
3. Load Sequence Complete – pSTR will start loading TRF generated STR sequences for all sample data. When user checks to see this status, pSTR has completed loading the STR sequences. Normally this process is very fast.
4. pSTR Match Complete – pSTR has completed matching all sample STR sequences.
5. pSTR Report Ready – pSTR has completed generating reports. At this point the Check Status button should change to Submit. This means user may proceed with the next request.

Reports Ready

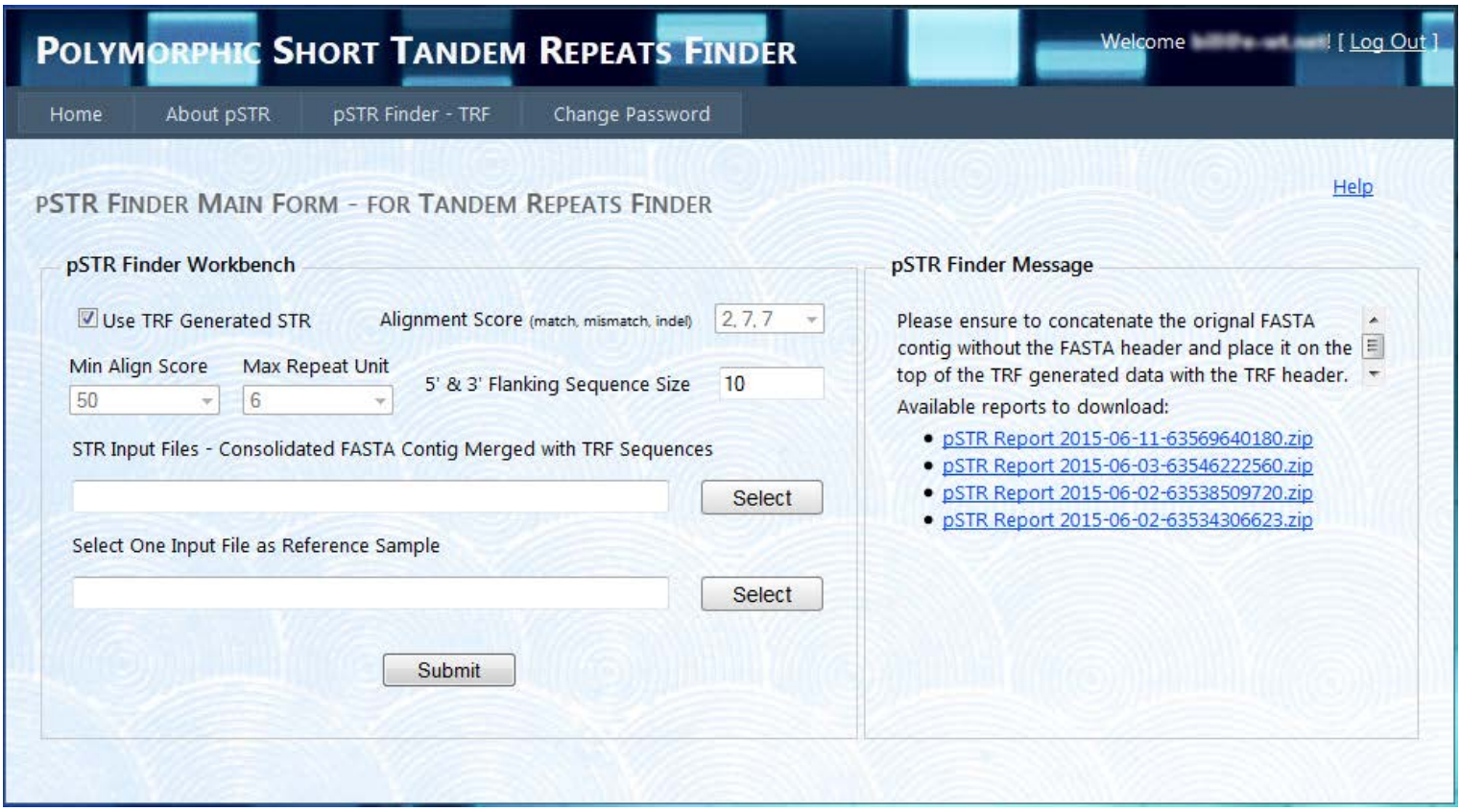
Eventually the pSTR process will complete and the result can be downloaded from the pSTR Finder Message panel and at this point user may submit another request.



Use TRF Generated STR Sequences Data Files

Check "Use TRF Generated STR" and notice most options are disabled except for the 5' & 3' Flanking Sequence Size. TRF standalone application will provide all the options for the user.

Notice in the pSTR Finder Message panel a brief instruction appears to remind user how to prepare the input sample data files.



Please note that most of the UI options are primarily inherited from TRF web site at http://tandem.bu.edu/trf/trf.advanced.submit.html. User will notice the standalone command line TRF application offers more options. The general idea is that if user uses TRF to produce more STR sequences than necessary like by lowering the minimal alignment score for instance, the pSTR matching will take longer time to process and it may produce more results than necessary.

Below is the command line used by the standalone TRF command line application:

trf407b.dos.exe "1 AC_000155_1.fasta" 2 7 7 80 10 50 6 -d -h

The flanking sequence size option is needed by pSTR.

Please make sure to manually merge the FASTA contig sequences without the FASTA header to the TRF generated data file. If the original FASTA contig sequences were broken into multiple lines, please concatenate them first then place the consolidated contig sequences on the top of the TRF generated file. Then you may use it as the input file.

For instance, below is a trimmed down sample of a tea DNA. Notice the original FASTA contig is on the top of the TRF generated STR. The original FASTA contig sequence is needed in order for pSTR to extract the flanking sequences for each specific STR. Please note that the FASTA contig sequence was wrapped to the second line due to the width limitation of this document. In fact the entire FASTA contig sequences is in the same line and this is a requirement if "Use TRF Generated STR" is checked.

GGGCGAACGACGGGAATTGAACCCGCGCATGGTGGATTCACAATCCACTGCCTTAATCCACTTGGCTACATCCGCCCCCTTACTCCACTATTT
AAATTATAAAATAAAGAATTAAAATTCAACCATTCATGATTATTTCTTTCTTCT...
Tandem Repeats Finder Program written by:

Gary Benson
Program in Bioinformatics
Boston University
Version 4.09

Sequence: 00_Chinshin.v4.5 |Camellia sinensis chloroplast, complete genome

Parameters: 2 7 7 80 10 20 6

136 150 4 3.5 4 90 9 21 0 26 0 73 0.84 TTTC TTTCTTTCTTCTCTT
356 368 1 13.0 1 100 0 26 100 0 0 0 0.00 A AAAAAAAAAAAAA
613 624 6 2.0 6 100 0 24 50 0 16 33 1.46 TTGAAA TTGAAATTGAAA
1485 1498 6 2.3 6 100 0 28 0 35 21 42 1.53 GCTTTC GCTTTCGCTTTCGC
1646 1655 4 2.5 4 100 0 20 30 20 0 50 1.49 TATC TATCTATCTA
2995 3008 6 2.3 6 100 0 28 35 28 0 35 1.58 ATTCCA ATTCCAATTCCAAT
3090 3099 4 2.5 4 100 0 20 30 0 50 20 1.49 GATG GATGGATGGA
3750 3762 5 2.6 5 100 0 26 15 0 0 84 0.62 TTTAT TTTATTTTATTTT
3778 3789 6 2.0 6 100 0 24 16 0 16 66 1.25 ATTTTG ATTTTGATTTTG
3791 3800 1 10.0 1 100 0 20 100 0 0 0 0.00 A AAAAAAAAAA
...

pSTR Reports

At the end of the process multiple reports will be generated and saved in CSV (comma separated values) format.

1. A Summary Report shows the number of identical STRs, the number of polymorphic STRs, and the number of different STRs between two matching samples. The Summary Report also details the total number of identical STR loci, and polymorphic STR loci, and unique STR loci for all samples analyzed, based on the specific reference sample.
2. Detail Report: this report relates to the specified reference sample. The following are recorded in the CSV file: the 5’ and 3’ flanking sequences; the number of bases constituting the repeat unit; the sequence of the STR motif; the position of the STR starting from the first base in the repeat; the variation in number of repeat units based on the samples included and with the same 5' & 3' flanking sequences; the number of repeats for the reference sample; and the number of repeats for every other sample included. Excluded in this Detail Report are STR sequence records that are at the same putative STR locus but are recorded as having a shorter repeat motif, and STR sequence records having the same 5' and 3' flanking sequences and smaller repeat number. These two types of STR sequence are excluded from the Detail Report are saved for each sample in the Duplicated STR Report.
3. Duplicated STR Report: a separate report for each sample captures the "duplicated" STR records mentioned above.
4. Identical STR Report: this report captures all "identical" STR sequence records from the comparison of two samples. An "identical" STR means that both the 5' & 3' flanking sequences and the repeat number are the same between both matching STR loci within the two input samples.
5. Polymorphic STR Report: this report captures all STR sequence records having the same 5' & 3' flanking sequences but a different repeat number between two input samples.
6. Different STR Report: this report, like 4 & 5 above, requires two input samples and captures all STR sequence records that exist in the source sample only and not in the other sequence file.
7. Different STR Report after switching samples: this report also captures all "different" STR sequence records but the "source" and "target" sample are switched and then re-matched.

We will be briefly discussing the summary report and detail report. All other reports are adequately self-explanatory.

Summary Report:

pSTR Finder: A rapid method to discover polymorphic STR markers from genome-wide sequences

James Chun-I Lee; Bill Tseng; Bing-Ching Ho; Adrian Linacre
Department of Forensic Medicine; College of Medicine; National Taiwan University

Matching Options: .2.7.7.80.10.50.6.10.10

Results of STR matching among 4 samples

	Reference Sample - gi|157734237|ref|AC_000155.1| Homo sapiens chromosome X; alternate assembly HuRef; whole genome shotgun sequence	Sample 2 - gi|224384746|gb|CM000685.1| Homo sapiens chromosome X; GRCh37 primary reference assembly	Sample 3 - gi|528476524|ref|NC_018934.2| Homo sapiens chromosome X; alternate assembly CHM1_1.1; whole genome shotgun sequence	Sample 4 - gi|74273659|gb|CM000274.1| Homo sapiens chromosome X; whole genome shotgun sequence
Reference Sample		7034 (4654)	6716 (4096)	8592 (2906)
Sample 2	4935 (2197)		10720 (2790)	10930 (3017)
Sample 3	4807 (3073)	3113 (2109)		8655 (3546)
Sample 4	5033 (2387)	2676 (2584)	4330 (3418)	

Note:
1. Above diagonal: No. of identical STR (No. of polymorphic STR).
   Below diagonal: No. of different STR (No. of different STR after switching samples).
2. Results of matching 4 samples using "gi|157734237|ref|AC_000155.1| Homo sapiens chromosome X; alternate assembly HuRef; whole genome shotgun sequence" as reference:
   No. of Identical STR: 5443
   No. of Polymorphic STR: 4305
   No. of Unique STR: 0

Notice "*gi|157734237|ref|AC_000155.1| Homo sapiens chromosome X; alternate assembly HuRef; whole genome shotgun sequence*" is the designated reference sample.

The matching options used for this request are:

1. Alignment Score: 2.7.7
2. Detection Parameters: Matching probability Pm and indel probability Pi. Pm = .80 and Pi = .10 by default and cannot be modified in this version of the TRF. Both Pm and Pi are hidden parameters.
3. Minimum Alignment Score: 50
4. Maximum Repeat Unit: 6
5. 5' & 3' Flanking Sequence Size: 10

Detail Report (partial):

No	5' Flanking Region of Reference Sample	3' Flanking Region of Reference Sample	Repeat Unit	STR Sequence of Reference Sample	STR Position of Reference Sample	Difference of Repeat Number	Repeat Number of Reference Sample - gi|157734237|ref|AC_000155.1| Homo sapiens chromosome X; alternate assembly HuRef; whole genome shotgun	Repeat Number of Sample 2 - gi|224384746|gb|CM000685.1| Homo sapiens chromosome X; GRCh37 primary reference assembly	Repeat Number of Sample 3 - gi|528476524|ref|NC_018934.2| Homo sapiens chromosome X; alternate assembly CHM1_1.1; whole genome shotgun sequence	Repeat Number of Sample 4 - gi|74273659|gb|CM000274.1| Homo sapiens chromosome X; whole genome shotgun sequence
1	TATTATCTCA	ATCACATGAG	4	AGATAGATAGAAAGATAGATGATAGATAGAAGATAGATAGATAGATAGATAGATAATAGATA	2926	0	18.3			18.3
2	ATATATATAT	TACTATATAT	2	ACACACACACACACACACACATATATACACACACA	27417	0	19.5			19.5
3	GTCAGAAGCT	TCCTGTATAG	2	ACACACACACACACACACACACACACACACACACACACACACACA	42757	0	24.5			24.5
4	TCAATCAGAA	GAGATGGAGT	1	TTTTTTTTTTTTTTTTTTTTTTTTTT	60771	0	26			26
5	CCTGGCTAAC	AGATGGAGTA	5	TTTTGTTTTGTTTTGTTTTGTTTTG	62491	0	5			5
6	TATATAGGAT	TATATTTTTT	4	ATAGATAGATAGATAGATAGATAGAGATAGATA	65897	1	8.8		9.8	8.8
7	ATTATTGTTA	GAGATGGATT	4	TTATTTATTTATTTATTTATTTATT	69390	0	6.3	6.3		
8	AAGCAGGAAC	TGGTTGACAA	2	TGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGT	71144	10	16.5	23.5	17.5	26.5
9	ACATGACTAC	AATGAATGAA	4	ATGGATGAATGGATGGATGGATGAATGGATGGATGGATGGATGGACAGATGCATGGATGGATGGATAGATG	75097	5	18.8	18.8	13.8	18.8
10	AGACCCTGTT	GGAGGGAGGA	4	GAAAGAAAAAAAAAATGGAAGAAAGAAAAAAAGGAAAGAAGAAACAGAGAAAGAAGAGAGAAGGAGAGAGAGAAAGAGAGAGAGAAAGAAAGAAAGGAAGGAAGGAAGGAAAGAGAGAGAGAGAGAAAG	76713	0	32.3			32.3
11	GAAAATAAAG	TGGAGAAAGA	4	AGAATGAAAGAAGGAAGAAAGGAGAGAAAGAAAGAAAGGAAGAAAGCAAAAGAAAAAGAAAAGAAAGAAAAAGAAAGGAGAAAGAAAGA	77520	0	22.8	22.8		22.8

Notice in record no 8, the STR TGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGT at the position 7114 of the reference sample compares with the other three samples results in a match on 5' and 3' flanking sequences but different repeat numbers. This means that all four comparing STR sequences are polymorphic. The difference of repeat number is 10 which is calculated as the largest gap of repeat number among all four comparing STRs.

Below demonstrates the “identical STR” scenario in record no 46 and 47. Notice all four comparing STRs have the same repeat number.

No	5' Flanking Region of Reference Sample	3' Flanking Region of Reference Sample	Repeat Unit	STR Sequence of Reference Sample	STR Position of Reference Sample	Difference of Repeat Number	Repeat Number of Reference Sample - gi|157734237|ref|AC_000155.1| Homo sapiens chromosome X; alternate assembly HuRef; whole genome shotgun	Repeat Number of Sample 2 - gi|224384746|gb|CM000685.1| Homo sapiens chromosome X; GRCh37 primary reference assembly	Repeat Number of Sample 3 - gi|528476524|ref|NC_018934.2| Homo sapiens chromosome X; alternate assembly CHM1_1.1; whole genome shotgun sequence	Repeat Number of Sample 4 - gi|74273659|gb|CM000274.1| Homo sapiens chromosome X; whole genome shotgun sequence
45	ATATTCTATC	CTGGGCTGAA	3	ATTATTATCATTATTACTATTATTATTATTATTA	355616	0	11.3	11.3	11.3	11.3
46	TATTTGTTAT	AGACAGAGTT	5	TTTTGTTTTGTTTTGTTTTGTTTTG	366083	0	5	5	5	5