The Large Scale Blast Score Ratio (LS-BSR) pipeline

Updated 8/28/2017

Citation:

Jason W. Sahl, J. Gregory Caporaso, David A. Rasko, Paul S. Keim (2014). The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. PeerJ PrePrints 2:e220v1.

Jason W. Sahl, J. Gregory Caporaso, David A. Rasko, Paul S. Keim (2014). The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. PeerJ 2 (e332).

Contact:

Please address queries, concerns, improvements to jasonsahl at gmail dot com

What does it do?

The LS-BSR pipeline was designed as a way to quickly compare the genetic content between a large number of bacterial genomes. LS-BSR can calculate several pan-genome statistics in a population and the output can be easily visualized with a variety of third-party tools. Additionally, LS-BSR can be used to query a set of genes and intergenic regions against a large set of genomes to identify gene distribution and conservation. LS-BSR was developed to be easy to run and interpret.

Installation

-The code is kept here:
https://github.com/jasonsahl/LS-BSR.git

-You can clone the repository to your own system with git:
```
$git clone https://github.com/jasonsahl/LS-BSR.git
```

-Enter the directory, then:
```
$python setup.py install
```

-if your install directory is /Users/jsahl/LS-BSR, run:
```
$export PYTHONPATH=/Users/jsahl/LS-BSR:$PYTHONPATH
```

-You can add this to your .bashrc or .profile

-You can test your installation by running the tests:

```
$python /Users/jsahl/LS-BSR/tests/test_all_functions.py
```

-If your installation is correct, all tests should pass

Dependencies

1. Python >2.7 and <3.0
2. USEARCH (tested version is 6.0.307): must be in your $PATH as "usearch" – **At least one clustering method must be chosen if a set of genes is not supplied**. 32-bit version should be sufficient for most applications, including the analysis of 1000 *E. coli* genomes. Tested successfully with versions v7.0.959, 7.0.1090, 8.1.1861, and 9.0.2132. Can be freely obtained for academics/non-profits from: http://www.drive5.com/usearch/. Very different results have been observed with V8 compared to V7.
3. VSEARCH (tested version is 1.1.3, but works with 2.0.4): must be in your $PATH as "vsearch" – **At least one clustering method must be chosen if a set of genes is not supplied**. Can be freely obtained at: https://github.com/torognes/vsearch
4. CD-HIT (tested version is 4.6): must be in your path as "cd-hit-est" - **At least one clustering method must be chosen if a set of genes is not supplied.** Does not support clustering ID lower than 0.7. Can be freely obtained at: https://github.com/weizhongli/cdhit
5. BioPython, must be in $PYTHONPATH environmental variable. Can be freely obtained from: http://biopython.org/wiki/Main_Page
6. Blast+ (tested version is 2.2.28 up to 2.2.50), must be in path as "blastn, tblastn, blastp, makeblastdb' – **only required if you are using BLASTN or TBLASTN, and not BLAT**. Blast+ can be obtained from: ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.28/. Weird and incorrect bit scores have been observed with 2.2.31.
7. BLAT (tested version is v. 35x1), must be in path as 'blat' – **only required if you use choose blat for your alignment method**. Can be obtained from: http://hgdownload.cse.ucsc.edu/admin/exe/
8. Prodigal (tested version is 2.60), must be in path as 'prodigal' - **only required if a set of gene sequences is not supplied**. Can be obtained from: https://code.google.com/p/prodigal/

Command Line options

-d DIRECTORY: --directory=DIRECTORY : the directory to your fasta files, all must end in ".fasta". Can either be complete genomes or draft assemblies. Scaffolds are discouraged. Genbank files are supported and must end in “*.gbk” [REQUIRED]

-i ID: de-replication clustering value, defaults to 0.9 (range from 0.0-1.0). Low values (<0.8) are not supported if CD-HIT is chosen as the clustering method

-f FILTER: whether to use BLAST filtering, default is "F" or filter, turn off with "T". Turning this to “T” should speed up the analysis, but may throw out highly repetitive sequences.

-p PROCESSORS: number of processors to use, defaults to 2.

-g GENES: if you have a list of genes to screen, supply a nucleotide fasta file (.fasta) or a peptide file (.pep). Each gene sequence must be in frame, or questionable results will be obtained (only true for TBLASTN). If this flag is not invoked, then the *de novo* gene prediction method is invoked

-c CLUSTER_METHOD: determines which clustering method to choose. You can choose from “usearch”, “vsearch”, or “cd-hit”. These must be in your path as “usearch”, “vsearch”, or “cd-hit-est” to use.

-**b BLAST:** which alignment method to use. Default is 'tblastn', can be changed to 'blastn' or ‘blat’. Can be used with either a list of supplied genes or with the *de novo* method. Tblastn is not compatible with “-y T” flag set below.

-l LENGTH: minimum BSR value to be called a duplicate, defaults to 0.7. The BSR of the "duplicate" divided by the reference bit score must be greater than this value to be called a duplicate

-m MAX_PLOG: maximum value to be called a remote paralog, defaults to 0.85. If the BSR value is greater than this value, then it is considered to be a highly similar paralog

-n MIN_HLOG: minimum BLAST ID to be called a highly similar paralog, defaults to 75. If the BLAST ID is below this value, it is considered a remote paralog

-t F_PLOG: filter ORFs with a remote paralog from BSR matrix? Default is F (do not filter), values can be T (filter paralogs) or F

-k KEEP: keep or remove temp files, choose from T or F, defaults to False (F), choose from T or F

-s FILTER_PEPS: filter out short peps < 50AA during TBLASTN? Defaults to True T), choose from T or F

-e FILTER_SCAFFOLDS: filter any contig that contains an N? Defaults to F, choose from T or F

-x PREFIX: prefix name for output files, defaults to time/date. If prefix is given, the temporary directory will be named after the prefix

-a min_pep_length: after translating sequences, peptides of a length smaller than this value will be discarded, defaults to 33 (integer)

-y intergenics: Include intergenic regions greater than 50nts in the analysis? Regions at the end of contigs will not be included. Choose from T or F, defaults to (F).

Test data – give LS-BSR a whirl on small datasets

-Test data is present in the test_data directory. This data consists of:

- Genomes (4 *E.coli* genomes from 4 different pathogenic variants). Genomes are:

- H10407 - enterotoxigenic *E. coli* (ETEC)
- E2348/69 - enteropathogenic *E. coli* (EPEC)
- O157:H7 sakai - shiga toxin *E. coli* (STEC)
- SSON046 - Shigella sonnei

-Genes (5 different markers that delineate between the variants). These include:

- IpaH3 - Shigella invasion antigen. Mostly present in *Shigella* spp.
- LT - heat-labile toxin. Only present in ETEC (not all)
- ST2 - heat-stable toxin. Only present in ETEC (not all)
- bfpB - bundle forming pilus. Only present on plasmid in EPEC
- stx2a - shiga toxin. Present in STEC

-You can test out the LS-BSR functionality in 4 different ways:

1. Test the gene screen method with tblastn:

-enter test data directory, run LS-BSR

```
$python /Users/jsahl/LS-BSR/ls_bsr.py -d genomes -g genes/ecoli_markers.fasta –x test
```

-the output (test_bsr_matrix.txt) should show how each gene is only present in the correct pathovar

2. Test the gene screen method with blastn:

-enter test data directory, run LS-BSR

```
$python /Users/jsahl/LS-BSR/ls_bsr.py -d genomes -g genes/ecoli_markers.fasta -b blastn –x test
```

3. Test the de novo gene prediction method:

-enter test data directory, run LS-BSR, using USEARCH for clustering

```
$python /Users/jsahl/LS-BSR/ls_bsr.py -d genomes –c usearch –x test
```

-Sample output for each method is shown in the test data directory

Output files:

1. $prefix_bsr_matrix.txt: This is the 2x2 matrix of the BSR value for each CDS in each genome queried
2. $prefix_names.txt: The names of all of your genomes. This file can be helpful for running the compare_BSR script described below
3. $prefix_duplicate_ids.txt: A list of sequence IDs that are duplicated in at least one genome
4. $prefix_consensus.fasta: A multi-fasta of all unique CDS sequences in the pan-genome.
5. $prefix_consensus.pep (optional): A multi-fasta of protein sequences if TBLASTN is selected.
6. $prefix_dup_matrix.txt: A 2x2 matrix showing how many copies of a CDS are present in each genome, if conserved above a given threshold.
7. $prefix_run_parameters.txt: Documentation for the run that you performed.

Visualization of output:

1. The output of LS-BSR can be visualized in many different ways. One popular method to visualize the output as a heatmap is through integration with R. Many beginners find R to be intimidating, but this link by Kat Holt provides some excellent workflows on how to build heatmaps, and correlate the output with phylogenies:

http://bacpathgenomics.wordpress.com/2012/05/25/displaying-data-associated-with-phylogenetic-trees/

2. The Interactive Tree of Life (iTOL) project has an interface that is very user-friendly and can directly take LS-BSR output and a phylogeny and create publication ready figures. iTOL can be found here:

http://itol.embl.de/

3. Finally, MeV is designed as a way to visualize expression data, but can just as easily create heatmaps from LS-BSR output. MeV can also be used to cluster LS-BSR data. MeV is platform independent and can be found here:

http://www.tm4.org/

Post-matrix scripts

1. compare_BSR.py

-what does it do? Looks for CDS differences between two user-defined populations. Differences can be set by user-defined thresholds for presence and absence. The "names.txt" file contains the names as they should be listed in your separate groups file

-what do you need for the script to run? Requirements include:

- BSR matrix
- Two new-line delimited group files, taken from "names.txt"
- FASTA file of all CDS sequences

-what does output look like? If there are unique sequences to either group, they will be stored in the "groupX_unique_seqs.fasta" file. If there are no unique sequences, the "groups_combined_header.txt" can be analyzed to look at the variable distribution of regions between groups.

```
$python compare_BSR.py -1 group1.txt -2 group2.txt –f $prefix_consensus.fasta –b $prefix_bsr_matrix.txt
```

2. filter_BSR_variome.py

-what does it do? Filters out the conserved regions of the pan-genome, if you are only interested in looking at the "variome" or accessory genome

-what do you need for the script to run?

- BSR matrix
- Sometimes if a single genome is missing a value, it is still of interest. You can change the number of missing values that can still be considered as core, and therefore filtered with the "-t" flag

-what does output look like? A new BSR matrix (“variome_BSR_matrix”), with only variable positions included

```
$python filter_BSR_variome.py –b $prefix_bsr_matrix.txt
```

3. filter_column_BSR.py

-what does it do? Can remove a column from a BSR matrix, in the case where a genome doesn’t belong, or is of poor quality

-what do you need for the script to run?

- BSR matrix
- Prefix for output file
- New line delimited file of genome(s) to remove

-what does output look like? A new BSR matrix (“$prefix_genomes.matrix”), with genome columns removed

```
$python filter_column_BSR.py –b $prefix_bsr_matrix.txt –p pruned –g to_remove.txt
```

4. isolate_uniques_BSR.py

-what does it do? Isolates CDSs only present in a single genome, using a user defined threshold for the definition of absence

-what do you need for the script to run?

- BSR matrix
- Threshold for absence, defaults to 0.4

-what does the output look like? A new BSR matrix (“uniques_BSR_matrix”), with only CDSs present in a single genome

```
$python isolate_uniques_BSR.py –b $prefix_bsr_matrix.txt
```

5. pan_genome_stats.py

-what does it do? Calculates several popular pan-genome stats, based on the BSR matrix

-what do you need for the script to run?

- BSR matrix
- Upper and lower thresholds for BSR values

-what does output look like? Several stats are printed to screen. The script also creates two files for the IDs of core and unique CDSs. The frequency_data.txt file can be graphed in order to see the distribution of CDSs across the pan-genome

```
$python pan_genome_stats.py –b $prefix_bsr_matrix.txt
```

6. BSR_to_PANGP.py

-what does it do? Converts a BSR matrix to something that can be easily visualized with PanGP (http://PanGP.big.ac.cn). CDSs that are above a given threshold are converted to a "1" and below that threshold are converted to a "-".

-What do you need for the script to run?

- BSR matrix
- Threshold for presence/absence

-what does output look like? A new matrix ("panGP_matrix.txt") compatible with PanGP

```
$python BSR_to_PANGP.py –b $prefix_bsr_matrix.txt
```

7. BSR_to_gene_accumulation_scatter.py

-what does it do? For a given number of iterations, the script randomly samples genomes at various depths, and reports back the number of core and unique CDSs in the pan-genome. The script also determines the gene accumulation in the pan-genome. The output can be easily graphed in Excel.

-what do you need for script to run?

- BSR matrix
- Upper and lower bounds for presence/absence
- Output

-what does output look like? The mean for each sampling depth is printed to screen. The script can be run to output accumulation, core, uniques, or all (default).

```
$python BSR_to_gene_accumulation_scatter.py –b $prefix_bsr_matrix.txt –n 100 –p
out
```

8. quantify_BSR_uniques.py

-what does it do? Prints out the number of unique CDSs, sorted by a given tree. Nice way of annotating a tree with where unique CDSs are present

-what do you need for script to run?

- BSR matrix
- Newick tree

-what does output look like? The script prints to a file: “uniques_sorted_by_tree.txt”, which shows the taxa name and the number of unique CDSs

```
$python quantify_BSR_uniques.py –b bsr_matrix_values.txt –r test.tree
```

9. reorder_BSR_matrix_by_tree.py

-what does it do? Transposes a BSR matrix and re-orders the matrix based on the order of the taxa in a newick-formatted tree

-what do you need for script to run?

- BSR matrix
- Newick tree

-what does output look like? The script prints a reordered BSR matrix to file (“reordered_matrix.txt”). Note that since the matrix is transposed, the number of columns can be significant. Best for smaller analyses. Make sure that tree taxa in the tree do not start with numbers.

```
$python reorder_BSR_matrix_by_tree.py –b $prefix_bsr_matrix.txt –t test.tree
```

10. invert_select_group.py

-what does it do? Can be used in conjunction with the compare_BSR.py script. If you are comparing groups, you copy the IDs from group1 into a file, then uses the names.txt file to automatically create the group2 IDs.

-what do you need for script to run?

- New line delimited list of IDs in group1
- “names.txt” file which is standard output of LS-BSR

-what does output look like? The IDs from group2 are printed to screen, but can also be re-directed into an output file

```
$python invert_select_group.py group1.txt names.txt > group2.txt
```

11. select_seqs_by_IDs.py

-what does it do? Sub-selects a FASTA file based on a list of IDs. This can be used in conjunction with the pangenome_stats script to select the core or unique genes from the consensus.fasta file

-what do you need for script to run?

- New line delimited list of FASTA headers
- FASTA file

-what does output look like? A new FASTA file is generated that contains only the sub-selected sequences of interest

```
$python select_seqs_by_IDs.py -i in.fasta -d fasta_IDs.txt -o out.fasta
```

12. slice_ref_genome.py

-what does it do? This script splits a provided reference into fragments using a sliding window approach. This could be useful for the case where you are interested in intergenic regions, deletion junctions, etc.

-what do you need for script to run?

- Reference genome in FASTA format (currently must be finished assembly, only 1 chromosome)
- Fragment length (default is 1000bp)
- Step size (default is 100bp, set to 1000bp for non-overlapping windows)

-what does output look like? A FASTA file named by your input genome, containing all of the genomic fragments

```
$python slice_ref_genome.py -r reference.fasta -f 500 -s 200
```

-Let's say you want to run this script on a set of genomes. You could run a simple for loop. Enter the directory of your FASTA files, then:

```
$ for file in `find . -maxdepth 1 -name "*.fasta"`; do name=$(echo $file | cut -f 2 -d "/"); slice_ref_genome.py -r $file; done
```

-you could then cluster with USEARCH and use the resulting centroids in your LS-BSR analysis. Currently this script is limited to the case of only a single chromosome or sequence

13. transfer_annotation.py

-what does it do? If you have a file with annotated peptide sequences, it will transfer that annotation to your centroids, if there is a close match using a user-defined BSR threshold. In some cases, there might not be a close match, especially with highly plastic genomes.

-what do you need for script to run?

- Annotated peptides. If using genbank, this would be a ".faa" file.
- Query peptides. This could be the consensus.pep file produced by LS-BSR if the TBLASTN alignment option is used.

-What does output look like?

- A new fasta file ($prefix.annotated.fasta) is produced with comments like this in the header: `Y75_p3178::gi|47118301|dbj|BA000007.2|_4067 (original ID: annotation)`
- If no close match is found, the original nomenclature is retained

```
 python transfer_annotation.py -p NC_017633.pep -c $prefix_consensus.pep -o out
```

14. extract_locus_tags.py

-What does it do? Given a GenBank file, this script will generate a multi-FASTA nucleotide file with all locus tags. Helpful for transferring annotation in script #13.

-What do you need for the script to run?

- GenBank file
- BioPython

-What does the output look like?

- A new multi-fasta file (NC_007779.fasta in this case) is created with the locus tag header followed by the coding sequences.

```
$python extract_locus_tags.py test_data/NC_007779.gbk
```

15. extract_core_genome.py

-what does it do? This script extracts core genome regions from all genomes, aligns them, then concatenates them into a single multiple sequence alignment

-what do you need for script to run?

- Directory of genomes in FASTA format
- Input multi-FASTA
- MUSCLE aligner. Can be obtained from: http://www.drive5.com/muscle/
- BLAST+
- BioPython

```
$python extract_core_genome.py –d genomes_directory –g $prefix_consensus.fasta
```

16. annotate_matrix_by_locus_tags.py

-what does it do? Given a BSR matrix, a consensus file, and locus tags, such as those produced with script described above, a new BSR matrix will be generated and annotated by locus tags if the BSR values are >80% similar.

-What do you need for the script to run?

- BSR matrix
- Locus tags
- BioPython in your PYTHONPATH
- Multi-FASTA file (e.g. consensus.fasta)
- BLAT in your PATH

-What does the output look like?

- A new BSR matrix ("bsr_matrix_annotated.txt") ordered by the new annotation
- A new multi-FASTA ("$prefix.consensus_annotated.fasta") that matches the new BSR matrix

```
$python annotate_matrix_by_locus_tags.py –b $prefix_bsr_matrix.txt -c
$prefix_consensus.fasta –l test_data/NC_007779.fasta –p $prefix
```

17. extract_tree_names.py

-what does it do? Given a newick tree, this script extracts the tree IDs in a new line delimited text file. This can be useful for a number of applications, including for reordering a BSR matrix

-What do you need for the script to run?

- Newick formatted tree
- BioPython

-What does the output look like? A list of genomes, in phylogenetic order, is printed to screen

```
$python extract_tree_names.py -t tree.tree
```

18. reorder_matrix_by_list.py

-What does it do? Given a BSR matrix and a newline list of genomes, perhaps obtained from script 17, this script will reorder and transpose the BSR matrix

-What do you need for the script to run?

- BSR matrix
- Text files of new line delimited genomes in order (get this from Script 17)

-What does the output look like?

- Transposed BSR matrix ("reordered_matrix.txt")

```
$python reorder_matrix_by_list.py -b $prefix_bsr_matrix.txt -g tree_order.txt
```

19. BSR_to_scoary.py

-What does it do? Given a BSR matrix, this script generates output that is compatible with Scoary

-What do you need for the script to run?

- BSR matrix

-What does the output look like?

- New matrix (“Scoary_matrix.txt”)

```
$python ../tools/BSR_to_scoary.py -b test_bsr_matrix.txt
```

***Disclaimer**

TGen and ITS Affiliates, representatives and employees make no representations, warranties, guarantees, or claims, of any kind or nature, as to the accuracy of, reliability, utility, performance, effectiveness, or otherwise, of the Software or the results obtained therefrom, nor do they assume any responsibility for the results, quality of results, or lack of results. THE SOFTWARE IS PROVIDED "AS IS" AND WITHOUT WARRANTY OF ANY KIND, WRITTEN OR ORAL, EXPRESS OR IMPLIED, DIRECT, INDIRECT, BY ESTOPPEL, OR OTHERWISE, AND SPECIFICALLY EXCLUDING ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT OF ANY THIRD PARTY PATENT OR OTHER INTELLECTUAL PROPERTY RIGHTS, AND ANY OTHER WARRANTIES PROVIDED FOR BY THE UNIFORM COMMERCIAL CODE, USAGE IN THE INDUSTRY, OR COURSE OF DEALING BETWEEN THE PARTIES. TGEN AND IBBL DO NOT WARRANT THAT USE OF THE SOFTWARE WILL BE ERROR FREE, UNINTERRUPTED, SECURE, OR VIRUS-FREE. TO THE EXTENT TGEN AND IBBL MAY NOT AS A MATTER OF LAW DISCLAIM ANY WARRANTY, THE PARTIES AGREE THAT THE SCOPE AND DURATION OF ANY SUCH WARRANTY SHALL BE THE MINIMUM PERMITTED UNDER APPLICABLE LAW.

LIABILITY LIMITATION

TGEN SHALL NOT BE LIABLE FOR ANY DAMAGES WHATSOEVER, INCLUDING ACTUAL OR DIRECT DAMAGES, AND SHALL NOT IN ANY EVENT BE LIABLE FOR ANY INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF INAPPROPRIATE, SUBSTITUTE, OR ALTERNATIVE, SOFTWARE, LOSS OF USE, LOSS OF DATA, LOSS OF PROFITS, LOSS OF REVENUES, OR BUSINESS INTERRUPTION) HOWEVER CAUSED, RESULTING FROM, RELATED TO, OR ARISING OUT OF THE USE OF THE SOFTWARE, REGARDLESS OF FORESEEABILITY AND NOTWITHSTANDING WHETHER RECIPIENT HAS BEEN ADVISED OF THE POSSIBILITY OF ANY SUCH DAMAGES, AND REGARDLESS OF THE FORM OF THE CAUSE OF ACTION, WHETHER IN CONTRACT,

BREACH OF WARRANTY, OR TORT (INCLUDING NEGLIGENCE, STRICT LIABILITY, OR OTHERWISE), OR OTHERWISE.