CSBB - Computational Suite For Bioinformaticians and Biologists

-> -> -> version3.0

CSBB is learning and evolving. If you have questions or Bugs, please contact Praneet Chaturvedi on contact on https://github.com/csbbcompbio/CSBB-v3.0

\*\*\*\*\*\* Please report bugs and issues on https://github.com/csbbcompbio/CSBB-v3.0 \*\*\*\*\*\*\*\*\*\*\*

Author Info ::

Author/Developer : Praneet Chaturvedi

Designation : Senior Analyst Bioinformatics @ Cincinnati Children's Hospital and Medical Center

CSBB-v3.0 offers 18 statistical, visualization and bioinformatics pipelines/modules for several bioinformatics applications. Detailed instructions on how to execute each module is give below.

Also refer to white paper in the package for detailed explanation of each module!!!

What's New:

1) Added ChIP and ATAC seq pipelines for human, mouse and frog

2) Added ProcessPublicData Module for processing public data that is from downloading .sra files to end results of ChIP/ATAC and RNA-Seq pipelines [Sorted Bams, Indexed Bams, BigWigs, Peaks and Expression Results]

3) Updated install module for installing all dependencies automatically

4) CSBB now downloads required files and softwares from CSBB repository [https://sourceforge.net/projects/csbb-repository/]

5) FetchGEO and FetchSRA module removed

6) Various Bug Fixes

~~~Steps to make the application executable.

Step 1: Open the terminal/Command prompt in MAC OS, LINUX & Windows respectively.

Step 2: Browse to the directory where you have saved the package.

=================

cd —> change directory

ls —> list directory

=================

For example: let’s say you have downloaded the package in Downloads.

====================================================

MAC-OS :

cd /Users/xxx/Downloads/CSBB-v3.0

Linux:

cd /home/xxx/Downloads/CSBB-v3.0

====================================================

Requirements For CSBB-v3.0::

1) Perl -> -> -> [MACOS and Linux are preloaded with perl, but if perl is missing on your system then please intall perl using : https://www.perl.org/get.html]

2) Python2.7+ -> -> [MACOS and Linux are preloaded with python, but if python is missing on your system then please intall pyhton using : https://www.python.org/downloads/]

3) R -> -> -> -> -> [Please install R on your machine from : https://www.r-project.org/]

Note:

1) When you run CSBB-v3.0 first time, then please do run install module to automatically download files and install dependencies [tools/softwares]

2) CSBB-v3.0 can auto install samtools/1.8 but if auto install fails then please install samtools/1.8 using : http://www.htslib.org/doc/samtools.html

3) CSBB-v3.0 now downloads all the required files and indexes for running RNA-Seq and ChIP/ATAC-Seq pipelines automatically from https://sourceforge.net/projects/csbb-repository/files/?source=navbar.

If CSBB is taking a long time to download these files then please directly download files from : https://sourceforge.net/projects/csbb-repository/files/?source=navbar and unzip the files

and place all the files and folders inside the Modules folder

4) CSBB-v3.0 can auto install bowtie2 but if auto install fails then please install bowtie2 using : http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

5) CSBB-v3.0 now creates a log folder CSBB-v3.0\_Logs inside the package folder and write logs for all modules [Log, Error and Exit Status]

6) Current reference assemblies : Human : hg19 , Mouse : mm10 , Xenopus-Laevis : Laevis\_9.2 , Xenopus-Trop : Trop\_9.1 and Zebrafish : Zv10

7) If you wish to use your own reference indexes of Bowtie2 and RSEM then please match the current name and file format of the files in the index folders and place them in the Modules folder

8) Now CSBB supports processing of public data from SRA [https://www.ncbi.nlm.nih.gov/sra]

9) If running perl CSBB-v3.0\_MacOS.pl or perl CSBB-v3.0\_Linux.pl gives an error then follow steps mentioned below:

if MacOS then run commands mentioned below from command line:

\*) sudo perl -MCPAN -e 'install Bundle::CPAN'

\*) sudo perl -MCPAN -e 'install Capture::Tiny'

if Linux then run commands mentioned below from command line:

if Ubuntu

\*) sudo apt-get install build-essential

\*) sudo apt-get install libbundle-cpan-perl

\*) sudo apt-get install libcapture-tiny-perl

if Redhat

\*) sudo yum install build-essential

\*) sudo yum install perl-Bundle-CPAN

\*) sudo yum install perl-Capture-Tiny

10) CSBB-v3.0 requires complete path to file and folder in use: for example : /Users/xxxx/Desktop/Data.txt or /Users/xxxx/Desktop/CSBB-files

11) Please note that users can simply drag and drop the file/folder Finder Window (MAC) and LINUX respectively when path to a folder or a file is required

12) Please remove space and quotes when files and folders are dragged from finder

13) LINUX users please remove quotes on the file and folder path when using drag and drop

14) If you do not have administrative privileges on your machine where CSBB-v3.0 is installed then please run CSBB with sudo command: for example : sudo perl CSBB-v3.0\_MacOS.pl install or sudo perl CSBB-v3.0\_MacOS.pl Process-ChIP-ATAC\_SingleEnd

15) FetchGEO and FetchSRA modules have been removed from CSBB-v3.0 and three new modules have been added: ProcessPublicData, Process-ChIP-ATAC\_SingleEnd and Process-ChIP-ATAC\_PairedEnd

16) CSBB-v3.0 uses bash scripts to process RNA-Seq and ChIP/ATAC-Seq data inhouse or public [User can change parameters of tools that CSBB-v3.0 uses in bash scripts. By default, CSBB-v3.0 uses all default parameters.

If any change has been done to Bash scripts, then please on command line browse to Modules folder and do chmod +x BashScriptName.sh

How to Create SRA-DATA\_Table.txt ?

##### Please see all this information can be gathered from https://www.ncbi.nlm.nih.gov/sra

##### SRA PUBLIC DATA INFO FILE

##### SRA ID : For example - SRR2924628

##### SPECIES : human or mouse or xenopus-leavis or xenopus-trop or zebrafish

##### ASSAY TYPE : RNA-SEQ or ATAC or ChIP-TF or ChIP-Epigenetic

##### READ TYPE : SINGLE-END or PAIRED-END

##### SAMPLE NAME : WHOLE-Genome

SRRid Species Assay ReadType SampleName

SRR2924628 xenopus-leavis RNA-SEQ SINGLE-END Sample1

Please Find SRA\_DATA\_TABLE.txt in the package to test ProcessPublicData Module

~~~Running each module

All modules can be run in two ways:

1) Giving all arguments from command line in one line [Batch Mode Run if you have more than one file as input].

2) Providing arguments when prompted by the application.

Let’s see how to use —help option in the application.

====================================================

MAC USERS ::

perl CSBB-v3.0\_MacOS.pl —help

LINUX USERS ::

perl CSBB-v3.0\_Linux.pl —help

====================================================

Below is the sample output from —help option

Operating System is darwin

=============================================

CSBB - Computational Suite For Bioinformaticians and Biologists

=============================================

Version\_3.0

Developed and maintained by Praneet Chaturvedi!!

Please report issues and bugs on https://github.com/csbbcompbio/CSBB-v3.0

CSBB-v3.0 Last run : Wed May 9 17:35:16 2018

Current Run: Wed May 9 17:36:29 2018

/Users/pra7mx/Desktop/Development\_CSBB\_v3.0/CSBB-v3.0\_Logs directory is already present, CSBB will use the same to write module logs

Please See README and White-Paper for getting detailed instructions on running CSBB-v3.0

=============================================

Please See : Please run install module if using CSBB-v3.0 for first time [just one time process]

Please use options below to run CSBB - Computational Suite For Bioinformaticians and Biologists

Options::

Citation ---> for citing CSBB please use url : https://github.com/csbbcompbio/CSBB-v3.0

install ---> for installing all the required dependencies for Pipelines ::only one time process

UpperQuantile ---> for performing upper quantile normalization

BasicStats ---> for obtaining stats like mean, median, standard deviation, variance, Sum, min and max for each Gene Expression profile

ExpressionToZscore ---> for obtaining z-scores for Gene Expression in samples

ExtractGeneInfo ---> for obtaining info/expression of list genes from a huge matrix gene info/expression

ExpressionPlot ---> for generating line plot visualizing expression/information of genes/entities across samples/objects

InteractiveHeatmap ---> for generating interactive heatmaps for expression data. User has three options on clustering type and four choices on color theme. Please read README for descriptions and run command.

CorrelationProfiles ---> for obtaining genes correlation profile termed as positively and negatively correlated based on User threshold. One can obtain profile for all genes or just genes of interest (For genes of interest user needs to provide the path to gene list file). Heatmap will only be displayed for genes of interest

Biogrid-Gene-Protein-Search ---> for obtaining gene-protein interactions for Human and Mouse for genes of interest

DifferentialExpression ---> for obtaining DE genes in RNA-SEQ expriments. Uses RUVSeq package in R

PCA ---> for performing Principal Component Analysis

NMF ---> for performing Nonnegative Matrix Factorization on Samples in Expression dataset

ProcessPublicData ---> for processing public data [Includes downloading .sra files from SRA, mapping to reference genome and running processing pipelines based on Assay Type for human, mouse, frog and zebrafish]

InteractiveScatterPlot ---> for generating Interactive Scatter plot based on user preference. User needs to provide path to the file, Column number for x-axis values, Column number for y-axis values and Column number which user needs for color factorization (If users provides No color factorization will not be done) Please see README for extensive explanation

Process-RNASeq\_SingleEnd ---> for processing Single End RNASeq data using RSEM for human, mouse, frog and zebrafish

Process-RNASeq\_PairedEnd ---> for processing Paired End RNASeq data using RSEM for human, mouse, frog and zebrafish

Generate-TPM-Counts-Matrix ---> for generating TPM and Counts Matrix for Both Isoforms and Genes using RSEM result directory and species of interest

Process-ChIP-ATAC\_SingleEnd ---> for processing Single End ChIP and ATAC Seq data for human, mouse and frog

Process-ChIP-ATAC\_PairedEnd ---> for processing Paired End ChIP and ATAC Seq data for human, mouse and frog

I) Running install:

====================================================

perl CSBB-v3.0\_MacOS.pl install

perl CSBB-v3.0\_Linux.pl install

====================================================

II) Running UpperQuantile:

====================================================

perl CSBB-v3.0\_MacOS.pl UpperQuantile Path\_to\_file

====================================================

Example Mac Users: perl CSBB-v3.0\_MacOS.pl UpperQuantile /Users/xx/Desktop/SAMPLE.txt

Example Linux Users: perl CSBB-v3.0\_Linux.pl UpperQuantile /Users/xx/Desktop/SAMPLE.txt

Now if you did not provide the path to file as an argument

Example: perl CSBB-v3.0\_MacOS.pl UpperQuantile

— > Application will prompt to provide user input

III) Running BasicStats:

====================================================

perl CSBB-v3.0\_MacOS.pl BasicStats Path\_to\_file

perl CSBB-v3.0\_Linux.pl BasicStats Path\_to\_file

====================================================

Example Mac Users: perl CSBB-v3.0\_MacOS.pl BasicStats /Users/xx/Desktop/FIND.txt

Example Linux Users: perl CSBB-v3.0\_Linux.pl BasicStats /Users/xx/Desktop/FIND.txt

IV) Running ExpressionToZscore:

====================================================

perl CSBB-v3.0\_MacOS.pl ExpressionToZscore Path\_to\_file

perl CSBB-v3.0\_Linux.pl ExpressionToZscore Path\_to\_file

====================================================

Example Mac Users: perl CSBB-v3.0\_MacOS.pl ExpressionToZscore /Users/xx/Desktop/FIND.txt

Example Linux Users: perl CSBB-v3.0\_Linux.pl ExpressionToZscore /Users/xx/Desktop/FIND.txt

V) Running ExtractGeneInfo:

====================================================

perl CSBB-v3.0\_MacOS.pl ExtractGeneInfo Path\_to\_Expression/Info\_file Path\_to\_Gene\_List\_File

perl CSBB-v3.0\_Linux.pl ExtractGeneInfo Path\_to\_Expression/Info\_file Path\_to\_Gene\_List\_File

====================================================

VI) Running InteractiveHeatmap:

====================================================

perl CSBB-v3.0\_MacOS.pl InteractiveHeatmap Path\_to\_File Clustering\_Option [Row\_Clust, Col\_Clust or Row\_Col\_Clust] Color\_theme [YellowGreenOrange, BlueWhiteRed, YellowBlackBlue or GreenWhitePurple]

perl CSBB-v3.0\_Linux.pl InteractiveHeatmap Path\_to\_File Clustering\_Option [Row\_Clust, Col\_Clust or Row\_Col\_Clust] Color\_theme [YellowGreenOrange, BlueWhiteRed, YellowBlackBlue or GreenWhitePurple]

====================================================

Example : perl CSBB-v3.0\_MacOS.pl InteractiveHeatmap /Users/xxxx/Desktop/Heatmap.txt Row\_Col\_Clust BlueWhiteRed

In color theme : First color represents low expression and last color represents high expression

VII) Running CorrelationProfiles:

====================================================

perl CSBB-v3.0\_MacOS.pl CorrelationProfiles Path\_to\_File Correlation\_Threshold [-1 to 1] all/Path to Gene list Correlation\_Type[pearson or spearman or kendall]

perl CSBB-v3.0\_Linux.pl CorrelationProfiles Path\_to\_File Correlation\_Threshold [-1 to 1] all/Path to Gene list Correlation\_Type[pearson or spearman or kendall]

====================================================

Example : perl CSBB-v3.0\_MacOS.pl CorrelationProfiles /Users/xxxx/Desktop/File.txt 0.75 all pearson

Example : perl CSBB-v3.0\_MacOS.pl CorrelationProfiles /Users/xxxx/Desktop/File.txt 0.75 Users/xxx/Desktop/GeneList.txt pearson

\*\*use all to calculate correlation profiles all the genes/entities in the matrix.

\*\*using path to gene list file to calculate correlation profile for a specified set of genes.

\*\*Please specify which correlation method you want CSBB to use [use Pearson for linear dependency and Spearman when interested in ranked correlation]

VIII) Running Biogrid-Gene-Protein-Search:

====================================================

perl CSBB-v3.0\_MacOS.pl Biogrid-Gene-Protein-Search Human/Mouse Path\_to\_gene\_list

perl CSBB-v3.0\_Linux.pl Biogrid-Gene-Protein-Search Human/Mouse Path\_to\_gene\_list

====================================================

IX) Running DifferentialExpression

====================================================

perl CSBB-v3.0\_MacOS.pl DifferentialExpression Path\_to\_Counts\_File Number\_of\_Controls Number\_of\_Treatments Counts\_Threshold\_for\_filtering Number\_of\_Samples\_for\_Filtering\_per\_Gene Type\_of\_Normalization [UpperQuantile or UpperQuantile+Empirical]

perl CSBB-v3.0\_Linux.pl DifferentialExpression Path\_to\_Counts\_File Number\_of\_Controls Number\_of\_Treatments Counts\_Threshold\_for\_filtering Number\_of\_Samples\_for\_Filtering\_per\_Gene Type\_of\_Normalization [UpperQuantile or UpperQuantile+Empirical]

====================================================

Example: perl CSBB-v3.0\_MacOS.pl DifferentialExpression Path\_to\_Counts\_File 10 10 5 8 UpperQuantile

\*\*\* Please see that place your controls/control reps first and then treatments/treatment reps

\*\*\* Please read https://bioconductor.org/packages/release/bioc/vignettes/RUVSeq/inst/doc/RUVSeq.pdf for understanding types of Normalization

\*\*\* Generally if sequencing Quality is good then UpperQuantile normalization works best

\*\*\* CSBB advise using both Normalization Types separately and see which gives you robust/best results

X) Running PCA

====================================================

perl CSBB-v3.0\_MacOS.pl PCA Path\_to\_File

perl CSBB-v3.0\_Linux.pl PCA Path\_to\_File

====================================================

XI) Running NMF

====================================================

perl CSBB-v3.0\_MacOS.pl NMF Path\_to\_File

perl CSBB-v3.0\_Linux.pl NMF Path\_to\_File

====================================================

XII) Running ProcessPublicData

====================================================

perl CSBB-v3.0\_MacOS.pl ProcessPublicData Path\_to\_SRA-DATA\_Table Path\_to\_Folder\_to\_write\_results

perl CSBB-v3.0\_Linux.pl ProcessPublicData Path\_to\_SRA-DATA\_Table Path\_to\_Folder\_to\_write\_results

====================================================

\*\*\* Please see for SRA-DATA\_table example in the CSBB-v3.0 folder

\*\*\* All the information has been provided in the example table file

XIII) Running InteractiveScatterPlot

====================================================

perl CSBB-v3.0\_MacOS.pl InteractiveScatterPlot Path\_to\_File Column\_x\_axis\_values Column\_y\_axis\_values Column\_for\_Color\_Factorization/No

perl CSBB-v3.0\_Linux.pl InteractiveScatterPlot Path\_to\_File Column\_x\_axis\_values Column\_y\_axis\_values Column\_for\_Color\_Factorization/No

====================================================

Example : perl CSBB-v3.0\_MacOS.pl InteractiveScatterPlot /Users/xxx/Desktop/DE.txt 3 2 7 -> When you are using column 7 for color factorization

Example : perl CSBB-v3.0\_MacOS.pl InteractiveScatterPlot /Users/xxx/Desktop/DE.txt 3 2 No -> When no color factorization is needed

XIV) Running Process-RNASeq\_SingleEnd

====================================================

perl CSBB-v3.0\_MacOS.pl Process-RNASeq\_SingleEnd Path\_to\_Fastq\_File Species [human or mouse or xenopus-leavis or xenopus-trop or zebrafish ] Output\_Folder\_path Phred\_Quality\_encoding [phred33 or phred64 or solexa]

perl CSBB-v3.0\_Linux.pl Process-RNASeq\_SingleEnd Path\_to\_Fastq\_File Species [human or mouse or xenopus-leavis or xenopus-trop or zebrafish] Output\_Folder\_path Phred\_Quality\_encoding [phred33 or phred64 or solexa]

====================================================

Example :

perl CSBB-v3.0\_MacOS.pl Process-RNASeq\_SingleEnd /Users/xx/Desktop/my\_fastq.fastq human /Users/xx/Desktop phred33

perl CSBB-v3.0\_Linux.pl Process-RNASeq\_SingleEnd /Users/xx/Desktop/my\_fastq.fastq human /Users/xx/Desktop phred33

\*\*\* More information about phred quality encoding can be gathered from : https://en.wikipedia.org/wiki/FASTQ\_format - Encoding

\*\*\* From 2011 Illumina [1.8+] version of pipelines use phred33 encoding for reads

XV) Running Process-RNASeq\_PairedEnd

====================================================

perl CSBB-v3.0\_MacOS.pl Process-RNASeq\_PairedEnd Path\_to\_Fastq\_File\_pair1 Path\_to\_Fastq\_File\_pair2 Species [human or mouse or xenopus-leavis or xenopus-trop or zebrafish] Output\_Folder\_path Phred\_Quality\_encoding [phred33 or phred64 or solexa]

perl CSBB-v3.0\_Linux.pl Process-RNASeq\_PairedEnd Path\_to\_Fastq\_File\_pair1 Path\_to\_Fastq\_File\_pair2 Species [human or mouse or xenopus-leavis or xenopus-trop or zebrafish] Output\_Folder\_path Phred\_Quality\_encoding [phred33 or phred64 or solexa]

====================================================

Example :

perl CSBB-v3.0\_MacOS.pl Process-RNASeq\_PairedEnd /Users/xx/Desktop/my\_fastq\_pair1.fastq /Users/xx/Desktop/my\_fastq\_pair2.fastq human /Users/xx/Desktop phred33

perl CSBB-v3.0\_Linux.pl Process-RNASeq\_PairedEnd /Users/xx/Desktop/my\_fastq\_pair1.fastq /Users/xx/Desktop/my\_fastq\_pair2.fastq human /Users/xx/Desktop phred33

\*\*\* More information about phred quality encoding can be gathered from: https://en.wikipedia.org/wiki/FASTQ\_format - Encoding

\*\*\* From 2011 Illumina [1.8+] version of pipelines use phred33 encoding for reads

XVI) Running Generate-TPM-Counts-Matrix

====================================================

perl CSBB-v3.0\_MacOS.pl Generate-TPM-Counts-Matrix Path\_to\_RNA-SEQ\_Result\_Directory Species [human or mouse or xenopus-leavis or xenopus-trop or zebrafish] Path\_to\_Output\_Directory

perl CSBB-v3.0\_Linux.pl Generate-TPM-Counts-Matrix Path\_to\_RNA-SEQ\_Result\_Directory Species [human or mouse or xenopus-leavis or xenopus-trop or zebrafish] Path\_to\_Output\_Directory

====================================================

Example :

perl CSBB-v3.0\_MacOS.pl Generate-TPM-Counts-Matrix /Users/xx/Desktop/CSBB\_RNA-SEQ human /Users/xx/Desktop

perl CSBB-v3.0\_Linux.pl Generate-TPM-Counts-Matrix /Users/xx/Desktop/CSBB\_RNA-SEQ human /Users/xx/Desktop

XVII) Running ExpressionPlot

====================================================

perl CSBB-v3.0\_MacOS.pl ExpressionPlot Path\_to\_File

perl CSBB-v3.0\_Linux.pl ExpressionPlot Path\_to\_File

====================================================

XVII) Running Process-ChIP-ATAC\_SingleEnd

====================================================

perl CSBB-v3.0\_MacOS.pl Process-ChIP-ATAC\_SingleEnd Path\_to\_Fastq\_File Species [human or mouse or xenopus-leavis or xenopus-trop] Output\_Folder\_path Phred\_Quality\_encoding [phred33 or phred64 or solexa] ExperimentType [ATAC or ChIP-TF or ChIP-Epigenetic]

perl CSBB-v3.0\_Linux.pl Process-ChIP-ATAC\_SingleEnd Path\_to\_Fastq\_File Species [human or mouse or xenopus-leavis or xenopus-trop] Output\_Folder\_path Phred\_Quality\_encoding [phred33 or phred64 or solexa] ExperimentType [ATAC or ChIP-TF or ChIP-Epigenetic]

====================================================

Example :

perl CSBB-v3.0\_MacOS.pl Process-ChIP-ATAC\_SingleEnd /Users/xx/Desktop/my\_fastq\_pair1.fastq human /Users/xx/Desktop phred33 ChIP-TF

perl CSBB-v3.0\_Linux.pl Process-ChIP-ATAC\_SingleEnd /Users/xx/Desktop/my\_fastq\_pair1.fastq human /Users/xx/Desktop phred33 ChIP-TF

XVIII) Running Process-ChIP-ATAC\_PairedEnd

====================================================

perl CSBB-v3.0\_MacOS.pl Process-ChIP-ATAC\_PairedEnd Path\_to\_Fastq\_File\_pair1 Path\_to\_Fastq\_File\_pair2 Species [human or mouse or xenopus-leavis or xenopus-trop] Output\_Folder\_path Phred\_Quality\_encoding [phred33 or phred64 or solexa] ExperimentType [ATAC or ChIP-TF or ChIP-Epigenetic]

perl CSBB-v3.0\_Linux.pl Process-ChIP-ATAC\_PairedEnd Path\_to\_Fastq\_File\_pair1 Path\_to\_Fastq\_File\_pair2 Species [human or mouse or xenopus-leavis or xenopus-trop] Output\_Folder\_path Phred\_Quality\_encoding [phred33 or phred64 or solexa] ExperimentType [ATAC or ChIP-TF or ChIP-Epigenetic]

====================================================

Example :

perl CSBB-v3.0\_MacOS.pl Process-ChIP-ATAC\_PairedEnd /Users/xx/Desktop/my\_fastq\_pair1.fastq /Users/xx/Desktop/my\_fastq\_pair2.fastq human /Users/xx/Desktop phred33 ChIP-TF

perl CSBB-v3.0\_Linux.pl Process-ChIP-ATAC\_PairedEnd /Users/xx/Desktop/my\_fastq\_pair1.fastq /Users/xx/Desktop/my\_fastq\_pair2.fastq human /Users/xx/Desktop phred33 ChIP-TF