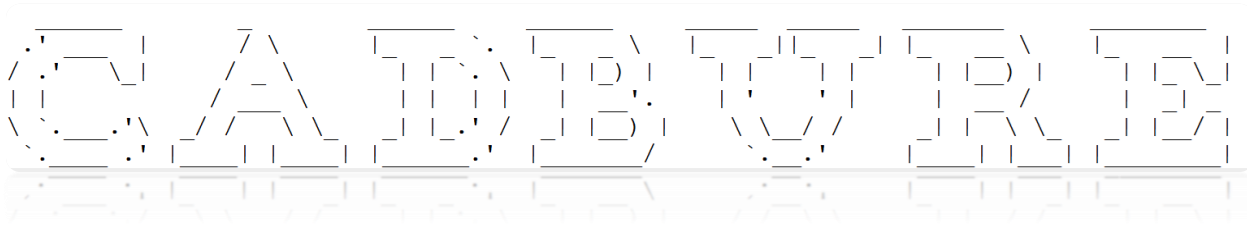Author: Praveen K. Raj Kumar

Contact: praveenk@pitt.edu

CADBURE

Introduction

CADBURE stands for comparing alignment results of user's RNA-Seq data based on the relative reliability of uniquely aligned reads. CADBURE is designed to easily evaluate different aligners, or different parameter sets using the same aligner, and selects the best alignment result for any RNA-Seq dataset. Strengths of CADBURE include the ability to compare alignment results without the need for synthetic data such as simulated genomes, alignment regeneration and randomly subsampled datasets. CADBURE works by subdividing aligned reads into eight distinct scenarios of potential alignment outcomes. These scenarios are binned into three categories: true positive, false positive and true negative. Finally, CADBURE calculates the specificity and accuracy for each alignment result using the true-positives, false-positives and true-negatives (see Article). Bootstrapping statistics ascertains the significance of differences between both the accuracy and specificity of alignment results. Higher accuracy and specificity means fewer false-positives. Non-significant difference in the accuracy and/or specificity measures of alignment results implies neither aligner holds a significant advantage over the other.

CADBURE is distributed under the GNU Public License (GPL). All its source codes are freely available to academic users. The latest version can be downloaded at the SourceForge (http://cadbure.sourceforge.net/).

Prerequisites

Hardware
CADBURE is a PERL script and should run on any computer (Linux/Unix, Windows or Mac) with PERL installed. However it is tested only on Linux.

Software

CADBURE can be obtained from here at http://cadbure.sourceforge.net/. Perl (http://www.perl.org/get.html) (version 5 and above) should be centrally installed or use the perl path to execute CADBURE like.

/wherever/perl CADBURE

Install the two Perl Modules

Bio::DB::Sam

HTML::Table.

Modules can be easily installed via CPAN as long as CPAN is installed. If CPAN is not installed see here [http://www.cpan.org/modules/INSTALL.html]

#/usr/bin/perl -MCPAN -e 'install Bio::DB::Sam'

Note: Installation of this module requires samtools. It is recommended to get samtools from Github. Instructions for installing samtools can be found here.

https://github.com/samtools/bcftools/wiki/HOWTOs

#/usr/bin/perl -MCPAN -e 'install HTML::Table'

Input files

Alignment results should be in the BAM format. For faster processing we recommend the input files to be sorted. See samtools for more information on how to convert the SAM format to sorted BAM format.

Available options

Mandatory flag	Description
-f	Name of the first aligner
-fb	BAM file for the first aligner
-s	Name of the second aligner
-sb	BAM file for the second

	aligner
-o	Name for the output summary file

Note: Uniquely and non-uniquely mapped reads were identified by value to the SAM flag **NH:i** (Number of Hits; value 1 for unique mappings and value 2 or more for ambiguous mappings). Hence it is mandatory that input BAM file alignment record contain NH flag. Also note only mapped read will have NH tag. Below is an example of SAM/BAM alignment record containing NH flag,

HWI-1KL140:55:C20PMACXX:7:2107:4150:3209   0   1   3009355 0   35M   *   0   0
TGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGT   IJIIHIIGGHFHHIHGHIGIHHDHHIIIHIHIIII
MD:Z:35 NH:i:1001   HI:i:1 NM:i:0 SM:i:0 XQ:i:40 X2:i:40

All RNA-Seq bam files used in our data analysis are available on
http://sourceforge.net/projects/cadbure/files/

Using the Sample files

Usage

    ./CADBURE -f aligner1 -fb Sample_aligner1.bam -s aligner2 -sb Sample_aligner1.bam -o sample-out

Make sure that you have proper permissions to create folder because CADBURE creates folders for each Scenarios.

CADBURE outputs summary file in the run directory as a html file with the name you provided to the -o flag. Open the html file using any web-browser. In addition it outputs the read name list associated with the scenarios in the respective folder it creates. For example, output files with the above sample usage will be

Scenario_1/Read-List_Scenario-1_TP_bothAligners.txt

Scenario_2/Read-List_Scenario-2_FP_aligner1.txt

Scenario_2/Read-List_Scenario-2_FP_aligner2.txt

Scenario_3/Read-List_Scenario-3_TP_aligner1.txt

Scenario_3/Read-List_Scenario-3_TP_aligner2.txt

Scenario_4/Read-List_Scenario-4_FP_bothAligners.txt

Scenario_5/Read-List_Scenario-5_FP_aligner1.TN_aligner1.txt

Scenario_5/Read-List_Scenario-5_FP_aligner2.TN_aligner2.txt

Scenario_6/Read-List_Scenario-6_TP_aligner1.txt

Scenario_6/Read-List_Scenario-6_TP_aligner2.txt

Scenario_7/Read-List_Scenario-7_TN_bothAligners.txt

Scenario_8/Read-List_Scenario-8_TN_aligner1.txt

Scenario_8/Read-List_Scenario-8_TN_aligner2.txt

TP ---> True Positive, FP ---> False Positive, TN ---> True Negative

Viewing the Read mappings of the scenarios

Open the aligner BAM file in any BAM viewer. We suggest TABLET (http://ics.hutton.ac.uk/tablet/) for its lightweight and high-performance graphical viewer.

We suggest you open two instances of TABLET for viewing the two alignment results. Thereby, you can see the difference in read mapping by the aligners. In the case of sample input files, open 'Sample_aligner1.bam' and 'Sample_aligner1.bam' in two different instances of TABLET. Now open any one of the Scenario folder described above and open the corresponding

aligner file to copy a read name. Use the read name to search for mapping in the two different instances of viewer to visualize the difference or similarity (Figure 1).
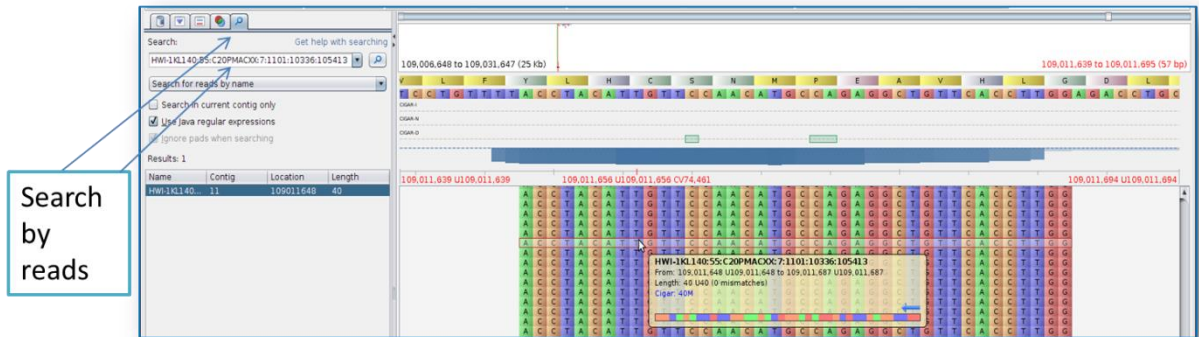

Figure 1: Screenshot of Mapping viewed in Tablet with the read searched being highlighted.