HIDE-seq Data Analysis Read Me

Release 1.0 rev2012-07-05

Table of Contents

Section 1. HIDE-seq Data Analysis

Step 1. Sorting and trimming raw data

Note: HIDE-seq was performed using Roche's 454 Titanium chemistry. Other sequencing platforms may be used but the downstream analysis will need to be modified accordingly (e.g. different primers & barcodes). Off the shelf software packages will perform many of the steps described herein, but the HIDE-seq tool requires generation of particular file types for its usage.

Your data will likely be saved in a compressed format that needs to first be unzipped.

Sequence files are in FASTA format. The .sff files are not required.

Reads with correct 454 primers, barcodes and subtraction primers (e.g. 1F and 2R; see Extended Experimental Procedures in the online supplement of Berg, M. et al 2012 for sequence information) must be identified accordingly for each experiment and filtered from those without them. Reads should then be sorted by barcodes (e.g. subtraction direction) and saved into separate files.

The fusion primer sequences must then be removed from reads (personalized for each experiment, which takes into account the barcodes used) to create a new files of trimmed sequences in FASTA formats. A mapping of read identifier to experiment type/identifier must also be created and stored in two columns in a map file (by convention we use the ".map" suffix to create such a file). The columns are separated by a tab, with the first column storing the read identifier, and the second column storing the corresponding experiment type/identifier.

Map files must combine both subtractions (e.g. control-variable and variable-control) into a single experimental file and named appropriately.

Step 2. Alignment of reads to respective genomes

Align the trimmed FASTA files to the required genome. The HIDE-seq tool accepts alignments only in .psl format (produced by aligners such as GMAP).

.psl files created containing the alignments can be viewed in the genome browser and should be copied back to your local directory as they are required for the remaining steps.

Step 3. Creation of HIDE-seq bedfiles and annotations

To start a HIDE-seq project from the command line: java –jar HIDE-seq.jar

A new window will be generated to allow users to load individual files.

To create the necessary files for the first time, click on the "PSL" tab at the top window.

The first entry line is for selecting .map files generated after sequence trimming. Click on the tab on the right to search your directory and upload the correct file.

The second entry line is for the .psl file generated after alignment. Upload as above.

The third entry line is for creation of output files. Copy the file name from the second line, remove the ".psl" extension, and replace it with ".bed".

In the fourth entry line, again copy the file name and replace ".psl" with ".hns".

The sequence coverage and identity are set at a default of 90% but each parameter can be adjusted by moving the blue arrows.

The overlap window is a measure of the maximum number of bases between reads to merge them into a single read and is typically set at 1 bp.

A title for the analysis may be entered at this time. If, for example, multiple experiments (e.g. numerous time-points compared to the same control) are being compared simultaneously, this can be indicated here. In this case, a combined .psl file is required which must be created during the alignment step.

Next, click on the bottom left tab, "edit treatment colors". One box will be checked for each subtraction direction. Colors can be changed by clicking on the default black box which brings up a new window of swatches from which to select. Click on a given color (e.g. red for variable-control) and press "OK". Repeat for the reverse subtraction (e.g. green for control-variable). If you want to see which reads were

found in both subtractions (e.g. background), click “add”, check both subtraction boxes, and select a third color.

Press “Start” to run the program which generates a “.hns” file and takes users to the next screen in which the “HNS” tab at the top is highlighted in blue. The “.hns” file name should already be found in the first entry line.

Note: For repeated use, the saved “.hns” files may be uploaded on this page directly and file creation performed on the previous page can be skipped.

In the second entry line for “Gene annotation file”, a text file of RefSeq (or similar database) genes for a given organism and genome version is uploaded. Currently, the software accepts database table dumps of hg18, hg19, mm9, etc. from the UCSC genome browser for the corresponding refgene tables.

The “Output Bedfile” line may be changed or left empty. Bedfiles created at this stage can be uploaded into the UCSC genome browser, for example, for viewing of data.

Adjacent reads are assembled together into “blocks.”

The “minimum block length” is a measure of the minimum block to consider (blocks smaller than this size are discarded) and is typically 10 bp, which again can be adjusted to one’s preference.

Press “Analyze” and a new window entitled “Edit control-treatment values” opens. Using your mouse, drag the “control-variable” title into the “control” box beneath it on the right and drag the “variable-control” title into the “treatment” box below on the left.

Press OK and a spreadsheet of all the unique reads mapped is generated. Users can directly sort the data within this window or press “save” to generate a text file.

Examples:

- To organize the data based on a particular parameter (e.g. pattern of reads), simply click on the top of the column and it will sort the entire list on this.

- If you only want to look at reads in the “control-variable” subtraction, right click on this designation within the spreadsheet, mouse across the “filter” option that appears, and highlight the “=” from the drop down menu.

- Similarly, if one wants to identify genes in the “variable-control” subtraction with reads in introns in the first 30% of the transcript, first sort on the “variable-control” reads (e.g. filter, “=”), then again on “intron” from the exon/intron start type column, then on 0.3 within the exon/intron rank

column, but select the "<" sign when filtering. The spreadsheet now will be reduced to reads fitting these criteria. To go back or undo, simply press "unfilter".

Plotting Data:

- Users may use the histogram tab to plot the filtered/unfiltered reads.

- The "column" tab allows users to designate which parameter to plot (e.g. exon/intron start rank). The "type" tab allows data to be plotted by frequency, relative frequency, or density. The "bins" tab allows one to designate the number of bins to create (e.g. the number of bars shown in the histogram) and "color" for the appearance of the graph.

- Press "draw" to update and "save" to create picture files. The "info" tab reports the number of reads that went into the annotations.

Section 2. Installation Notes

To run the project from the command line, go to the HIDE-seqDist_1.0 folder and type the following: java -jar "HIDE-seq.jar"

To distribute this project, zip up the HIDE-seqDist folder (including the lib folder) and distribute the ZIP file. Please note that this software requires JFreeChart (included), which is licensed under the terms of the GNU Lesser General Public License. Please see: www.jfree.org/jfreechart/.

If you redistribute or make changes to the software or this document, please include this original document.

Section 3. License

- Neither the name of the Dreyfuss Laboratory nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL DREYFUSS LABORATORY BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

Section 4. Known Bugs

Although we have attempted to verify the accuracy of this software to the best of our ability, there may still be some problems. For instance, even though the software is designed to allow different sets of controls and experiments, there may be some issues with the color schemes when using more than one pair of control and experiment. Use this feature and the software at your own risk.