Xpression Graphical Interface User Guide
Harwood Lab UW
March 2012

Introduction
Requirements
Example dataset
Quick Start
Workflow
Xpression options

Introduction

Xpression is a RNA-seq analysis pipeline that uses free and open technologies to create gene expression profiles, mapping statistics and visualization files from next-gen sequencing data in a fast, flexible manner. Xpression is designed to be an easy-to-use, integrated solution for processing next-gen sequencing data derived from various types of samples, such as samples mapped to draft genomes and multiplexed samples. Automated steps include quality filtering for sequencing and mapping, and gene expression profiling and normalization. It is especially well suited for analysis of bacterial sequence data.

Xpression can be run from command-line or from a graphic user interface (GUI), which is the focus of this guide. The GUI is relatively intuitive and simple, allowing the running and queuing of multiple samples quickly and easily. This GUI runs in a Linux environment, but can be made to run on other Unix-like environments. A simple installation script allows all supporting software to be installed automatically in Linux using common system tools. If the user's system is not Linux, we have provided a complete graphical Linux environment to run Xpression from Windows or MacOS without any further installation or modification required.

This document will guide the user through the process of getting Xpression up and running, including an example analysis run. The later sections deal with some specifics of each option field found in the Xpression interface. Further documentation is available on the Harwood lab website https://depts.washington.edu/cshlab/html/rnaseq.html.
A guide to installing the graphical virtual Linux system, Xpression VE, is also available.

Things you will need on your computer

- A graphical Linux system or the provided graphical virtual Linux system, Xpression VE
- Supporting software installed (please see installation guide)
- Xpression_GUI.jar file
- a fastq sequencing file
- a fasta genome reference file
- a genbank reference file containing complete CDS / gene information

These reference files may be found from sources such as NCBI http://www.ncbi.nlm.nih.gov/nuccore or JGI http://img.jgi.doe.gov/cgi-bin/w/main.cgi.

Example Dataset

We have provided an example dataset which includes everything needed to run Xpression. A fastq file is from the GAII platform in Illumina 1.3+ format. It contains 8 multiplexed samples by means of 8 unique 4-mer barcodes used in the cDNA library construction. The barcodes

used were ACCC, CGTA, GAGT, TTAG, AGGG, CCAT, GTCA, and TATC. The fastq contains a total of 2,979,809 reads, and has a compressed size of 84 MB. A fasta file and a genbank file are included, and are necessary to run the pipeline for this example.

Quick start guide

To get started, we will now process a sample sequencing file using Xpression. Each step is explained in further detail in later sections.

Step 1 – **Obtain Xpression**: Download the Xpression_GUI.jar from the Harwood lab website at http://depts.washington.edu/cshlab/html/rnaseq.html to your harddrive.
Step 2 – **Download dataset**: Download the example_dataset.tar.bz2 file and unpack it into a convenient directory. This includes a small fastq file, fasta and genbank reference files.
Step 3 – **Start Xpression**: Start the Xpression application by double-clicking the Xpresson_GUI.jar from a file browser or your desktop.
Step 4 – **Enter sample-specific information**: In the Sample Details area, enter the barcode sequence of ACCC into the 'barcode' field, and the name 'Example1' into the Sample ID field.

In the Sample References area, there are fields for each of the three required reference files. For each of these three reference files, do the following:
Step 5 – **Define reference files**: Click the left-hand button to open a file selection window. Locate and select the appropriate file from the example dataset package where it was decompressed on your computer.

Step 6 – **Begin analysis**: Click the Start Run button in the upper-right corner. This button will now change to Stop Run, to allow processing to be cancelled. Two windows will become visible, a run list and an Xpression output log window.
Step 7 – **Explore results**: Once Xpression analysis is finished, the upper-right button will change to Start Run and the status in the run list will change from running to complete. The processed files are located at the output directory in the folder labelled by the sample ID. In this case this will be your home folder > Xpression_results > example > Example1, since output directory is your_home_directory/Xpression_results, the sequence filename is example.fastq, and the sample ID is Example1.

Pipeline workflow overview

Xpression analyzes FASTQ sequencing files, the output of “next-generation” sequencing from platforms including the Illumina GAII and HiSeq. Xpression uses reference files specific to the organims or strain under investigation. Sequence reads are mapped to a genome fasta file and a genome loci genbank file provides loci boundaries and gene information. The user must supply sample-specific and sequencing-file-specific information such as sample name, sample barcode, sequencing file format, and so on. Once the pipeline is started, these options cannot be changed during the run. If an error occurs the pipeline may be stopped by clicking Stop Run. Once the error has been resolved by providing proper information, the sample can be requeued by click Add to Queue.

Xpression also supports the use of non-finished genomes that may be in multiple contigs. For these files, a single fasta and/or a single genbank should be provided which include all genome contigs as separate entries.

Xpression will process one sample at a time, completing each step of the pipeline sequentially for each sample.

- **Step 1 - Quality filtering and barcode extraction**
  The sequencing fastq is filtered based on sequencing quality. If the sample is barcoded for multiplexing, Xpression can separate and extract each barcoded sample by providing barcode sequences as a semicolon-separated list in the 'multiplexed barcodes' field in the Options menu.

  Note: In the case of samples multiplexed within the same sequencing file, specifying all the barcode sequences used in the 'Multiplexed barcodes' field will reduce the number of times the lengthy sequencing file is processed, shortening overall running time significantly. Supplying these additional barcodes will result in these barcoded reads being extracted as well as the current sample. These reads will be separated and saved for later runs of the pipeline on the samples corresponding to those barcodes. For instance, entering the the first barcode, ACCC, in the 'sample barcode' field and the remaining seven barcodes in 'Multiplexed barcodes' field will extract all eight barcodes from the example dataset.

- ***Step 2 – Map reads to genome reference***
  The extracted reads for the sample (determined by barcode if present) are mapped to the genome reference as given in the 'Genome reference' field.

  File produced → mapping statistics table
  The mapping statistics file is a table of total reads considered, 'good' reads based on sequencing quality, 'mapped' reads that mapped to the sample's reference fasta file, and 'unique' reads that are the 'good' and 'mapped' reads mapped to a single postion on the reference. Percentage of 'unique' reads typically indicate the level of rRNA depletion.

- **Step 3 - Generate expression profile**
  Generate an expression profile by counting reads within gene boundaries using the high-quality, sample-specific, mapped reads,

  File produced → expression profile table
  The expression profile can be opened in MS Excel or another spreadsheet program, and includes raw count data (reads), read normalised to total reads (pM), and reads normalised to both total reads and loci length (pKM).

- **Step 4 – Generate visualisation file**
  Create a visual representation of the expression profile for this sample.

  File produced → visualisation file
  This file is a visualisation of the expression profile of the sample. It can be viewed in programs like Sanger Institute's ARTEMIS or the Broad institute's Integrated Genome Viewer.

Obtaining the Xpression GUI

If the Xpression_GUI.jar file is not already on your computer, or if you are not using the virtual system, you need to download it from the Harwood lab website.

Direct your web browser to the RNA-seq section of the Harwood lab website found at:
http://depts.washington.edu/cshlab/html/rnaseq.html
Then find the link for Xpression_GUI.jar file and right-click to save the file on your computer's hard drive.

Starting Xpression

First find the Xpression_GUI.jar file that you saved to the computer. Double-click to open this application. If double-clicking is unsuccessful, try right-clicking and choosing the “open with java” option that may be presented.

Entering sample details

Upon opening the Xpression application, you should be presented with a window like *Figure 1*. A number of options and fields are available for modification. Some of these fields can be left as defaults, but some inputs are needed before Xpression can start analyzing your data.

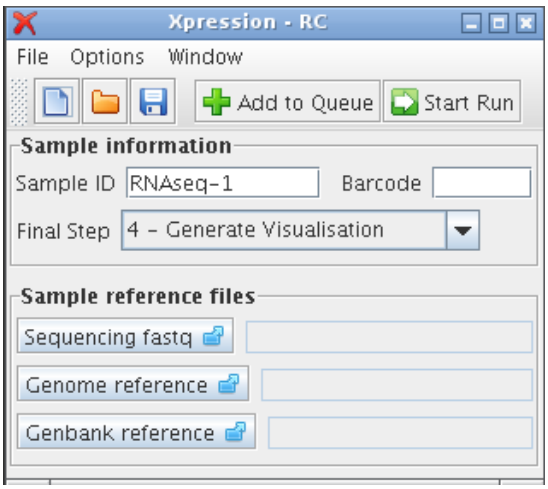

Figure 1: Xpression main window

Sample Information

The 'Sample information' fields change the sample's identification for output files, a nucleotide barcode if sample have been multiplexed in the same sequencing run, and the final step to be run in the pipeline.

Sample ID This can be anything to help describe your sample, but should be unique from other samples you have run previously to avoid overwriting data.

Barcode This is unique to a multiplexed sample, and tends to be a short (3-6 nt) sequence ligated before the biologically relevant sequence. If only one sample is present in the sequencing file, this field should be left empty. Only the letters A, G, T, C, and N are allowed.

Final Step The pipeline will start at step 1 and finish at the step indicated here. These steps are more fully described in the reference material. The default setting is to run the pipeline from step 1 to step 4. However, if only mapping statistics are required for instance, the pipeline only needs to be run to step 2, and the genbank file can be disregarded. To change this option, click on the arrow to see the list of steps.

Sample reference files

These files are specific to the sample you want to analyze. They must be located physically on your hard drive or attached external media. To choose the files, you simply browse for its location, letting Xpression know where they are. The sequencing file may contain multiplexed, barcoded samples, so the same sequencing file should be chosen for all samples within that file. The genome reference fasta and Genbank reference files are specific to each sample as well and must correspond to the organism or strain of the sample.

Sequencing fastq This is the sequencing file obtained from your sequencing facility. To select the file, click on the left-hand 'Sequencing fastq' button to bring up a file selection window. Navigate to the location of this file, which could be a place like the Desktop, your personal home folder or an external source like a DVD or removable harddrive. This sequence file format should be FASTQ, but may end with file extensions such as 'fastq', 'txt', or 'seq'. If it is compressed (e.g. with zip or gunzip) it may be read without being decompressed first, to save time and space.

Note: If you do not see the file you know to exist, try changing the 'files of type' menu at the bottom of the window from 'fastq and txt files' to 'All files' to see all available files in the current directory.

Genome reference This file is the next field and is required for mapping the sequence reads to a reference. If you want to run step 2 or higher of the pipeline, a valid genome file in the standard fasta format must be supplied.

Genbank reference The last file is the Genbank reference file which contains loci information such as CDS boundaries and annotations. On NCBI's database this file is known as genbank (full) or genbank with parts.

Running Xpression

Other options exist that will be discussed next, but once the fields just mentioned are filled, Xpression is now ready to run with these settings. The run can begin immediately by clicking ‘Start Run’, or can be queued to be run later after a currently running sample by clicking 'Add to Queue’. Either of these buttons will load this sample and all of its parameters into the pipeline.

Note: If a field is invalid or is required but empty, the field will be highlighted, and will need to be fixed before the sample can be run.

To see the currently queued samples, select Sample List from the Window menu.

Xpression output window

This window displays information about pipeline output. Potentially useful information will be displayed as the sample runs, including which step it is processing, files it is using or generating and errors if they occur.

Run list window

The Sample List shows each sample that has been queued. The columns show the sample’s ID, barcode used, if any, and status. Status can be queued, running, stopped, or error. If the sample is stopped or if an error occurs, the sample will have to be requeued and run again after potential issue has been resolved. Clicking a sample then the 'Details’ button in the same window will display all the parameters used for that sample. In this ‘details’ window, the settings for that run can be saved if it was not already.

Options window

The options window contains parameters that do not generally require user input and are set to defaults. These defaults can be modified to suit the sample or library preparation method used. They can be reverted to defaults by selecting ‘restore defaults’ from the options menu in the main window.

Sample Options

Strand Specific If the library preparation method has maintained strand specificity, select this option. The reads will map to a specific strand on the reference genome, but if the library preparation method did not allow directionality to be maintained, this ‘strandedness’ mapping information will not be accurate. Expression data will be divided into 'genic’ and 'intergene’ for non-strand-specific and ‘sense', ‘antisense', 'igpos’, 'igneg’ for strand-specific samples.

Native Direction If the library preparation method has maintained strand specificity and the resulting sequencing data is oriented forward relative to the native strand. If this is un-checked, the resulting read data is a reverse-complement to the native strand. Library preparation methods like the standard Illumina protocol produce native-strand reads, while

spRNA-seq produces reads that are a reverse-complement to the native strand. This option is disregarded if Strand Specific is not also selected.

Read start position This is the start of valid biological sequence. If a barcode was used, this field is automatically set to the position following the barcode unless you choose for this check not to occur. For example, with a barcode of ACGT, the start position will be 5.

Multiplex barcodes In addition to the current sample being extracted from the sequencing file, you may indicate which other barcodes to extract as well in this field. Sequence reads with these barcodes will be extracted and saved for later use by additional runs of the pipeline for samples corresponding to them. This is an optional field, but can shorten overall running time by processing the sequencing file once instead of each time. To indicate which barcodes to extract, use upper-case barcodes separated by a semi-colon, such as AAAA;CCCC;GGGG. The current sample's barcode will be extracted even if this field is left empty, and only corresponds to the other samples in the sequencing file.

Input format The sequencing file output assigns a quality score to each base position in a read. 'fastq' refers to the Sanger/Illumina 1.8+ and fastq-illumina refers to Solexa/Illumina 1.3+/1.5+. The sequencing file you receive may not likely give any indication of this format, the file being labelled with a general '.txt', '.seq', or '.fastq' file extension. For further explanation, please see the other reference material or visit http://www.wikipedia.org/wiki/FASTQ_format for further clairification. The pipeline will raise an error while processing if this format appears to be incorrect. The default is the most recent type, 'fastq'.

General Options
These option may vary less from sample to sample, and are maintained between sessions of Xpression. The '.xpression_settings' option file is located in the user's home folder.

Allowed mismatches Number of allowed mismatches between the read to be mapped and the fasta reference genome in step 2. The default is 2 mismatches.

CPU processes Number of processes to be run simultaneously (in parallel). This is usually set to 2 processes per CPU core, so quad-core computers would have 8 processes and dual-core would be half of that, 4. This value can be set higher or lower depending on the capacity of the CPU, but higher values may make the system unresponsive or unstable.

Output location Directory where sample runs are saved to. Each sequencing file will have a folder in this location and each sample a subfolder within this folder. This is the folder to look in to find the results when a sample has been processed.

Package results Compress mapping statistics, expression_profile, and/or visualization files after analysis for conveniece.

Saving Xpression parameter file

The set of options used for a sample can be saved as a parameter file for later use. For each sample, an option file is automatically generated and saved in the sample's output folder. To save the current sample settings, select Save or Save As from the File menu.

If the sample was queued and the details are now changed in the main window, that sample can still be saved by clicking the 'Save' button from the Details window accessed via the Sample List window.

Opening an Xpression parameter file

An Xpression parameter file can be opened to use a previously saved set of options. Xpression will look for the reference files defined in the parameter file, so if these have been moved or deleted a pop-up window will notify you of anything that needs to be fixed. Missing or invalid entry fields will be highlighted. To open a file, select ‘Open’ from the File menu and browse to the file location.

Example dataset walkthrough

A walkthrough video is embedded on the Xpression website. This written guide will follow that video which shows a sample of the example dataset as it is run in Xpression.

First you need to make sure all required software is installed properly and that you have a copy of Xpression.jar. This walkthrough will use the example data set that can be downloaded from the Harwood lab website. The dataset includes a fastq file, a genome reference file, and a genbank annotation reference file.

Start Xpression by right-clicking the Xpression_GUI.jar icon and choosing Java as the program to run this file with. If there is no application option, look for a “custom command" field and enter in ‘java -jar’ and press enter. Alternatively, you can type ‘java -jar ./path/to/Xpression_GUI.jar’ into a terminal emulator.

Now the Xpression window is open.

First, lets change the Sample ID to Example1. Next, the barcode needs to be defined to allow reads with a certain barcode to be extracted. For our multiplexed example, the four-nucleotide barcode is ‘ACCC’, so we type ‘ACCC’ into the barcode field. Now all reads starting with this barcode will be included in the sample output.

Xpression will complete all the steps of analysis by default, but you can click on the drop-down box ‘Final Step’ to explore which steps will occur.

Now the sequencing file itself must be located on the filesystem. This file is in fastq format, with the 'fastq' quality encoding which is default. This file can remain compressed to save time.

Next, the location of the sample organism’s genome reference in fasta format is entered. This file is used as a reference to map the reads.

Finally, since we want to generate expression data for our sample, a genbank reference file corresponding to the organism of the sample needs to be located. If the pipeline is only run to step 2, this is not needed.

Now click the Start Run button located in the upper right hand corner of the main window. This will add the sample to the list and start the pipeline for this sample.

Various output shows the steps and status of the pipeline as it is running.

Once Xpression has completed, find the output files located in directory chosed by the Output location field, by default Xpression_results in your home directory.