

Description of data files

ALEXA – A microarray design platform for alternative expression analysis.

Malachi Griffith¹, Michelle J. Tang¹, Obi L. Griffith¹, Ryan D. Morin¹, Susanna Y. Chan¹, Jennifer K. Asano¹, Thomas Zeng¹, Stephane Flibotte¹, Adrian Ally¹, Agnes Baross², Martin Hirst¹, Steven J.M. Jones¹, Gregg B. Morin¹, Isabella T. Tai¹ and Marco A. Marra¹.

¹Canada's Michael Smith Genome Sciences Centre. British Columbia Cancer Agency. Vancouver, BC. Canada.

²Genome British Columbia, Vancouver, BC. Canada

- 1.) Array design files
- 2.) Array design annotation files
- 3.) Raw and processed array data
- 4.) Result tables
- 5.) Control gene list

In addition to the figures and tables provided in the supplementary materials, additional data files for validation experiments of the ALEXA platform are provided as described below. These include the following: (1) The array design file submitted to NimbleGen which was used for the layout and synthesis of our validation arrays. (2) A complete annotation of the probes on this array including information on the probe sequences themselves as well as their genomic coordinates. (3) RAW probe-level data from both the ALEXA experiments as well as the Affymetrix experiments (only for those genes included in the ALEXA design). Files containing background corrected and normalized data are also provided. (4) Excel results tables. These contain gene-level summaries for all genes profiled by both the ALEXA and Affymetrix platforms. Complete lists of differential expression results for individual exons, introns or junctions are also provided. Extended versions of the 'top' DE gene and isoform lists presented in the supplementary tables are also provided. (5) A list of control genes targeted by both the ALEXA and Affymetrix designs. These genes were used for some of the analyses described in this manuscript.

The files described below can be downloaded from:

http://www.bcgsc.ca/people/malachig/htdocs/alex_a_platform/data/Supplementary-Data-1.zip

1.) Array Design Files ('ArrayDesignFiles' directory)

The file, '2006-09-29_MG_Human.ndf' is the design file used by NimbleGen to describe the layout and synthesis of our custom array. This file describes the sequence and position of all 385,000 probes for the ALEXA array design used in our validation

experiments. It also includes some additional probes which NimbleGen uses for internal control purposes when synthesizing the array.

2.) Array Design Annotation Files ('ArrayDesignAnnotation' directory)

For convenience, additional information for each probe can be found in the file, '[MIP101_vs_5FUR_CompleteDesignAnnotation.txt](#)'. For example, this file contains chromosome coordinates, gene IDs, probe sequence, tm, etc. The coordinates are provided for each probe such that where appropriate the position of each half of a junction probe is provided. For example, an exon-exon probe will have coordinates defining the chromosome position of the probe in each of the two exons it spans. These coordinates are referred to as ('Unit1_start_chr' to 'Unit1_end_chr') and ('Unit2_start_chr' to 'Unit2_end_chr'). In the case of Exon and Intron probes only a single set of coordinates are provided as there is no junction involved. Note that this file does not include annotations for the 4392 random 'Control-Negative' probes included on the array. Since these are random probe sequences they do not have chromosome coordinates or associated genes. Finally, remember that considerable additional annotation information for these probes can be retrieved from the database for ALEXA_hs_35_35h. Refer to the ALEXA platform website for more details (www.AlexaPlatform.org).

3.) Raw and Processed Array Data ('ArrayData' directory)

A.) Raw Data

The '[Raw_Data](#)' directory contains intensities as they were reported in Affymetrix or NimbleGen (for ALEXA data) image files.

B.) Background corrected and normalized data

The '[BGC_Normalized_Data](#)' directory contains fully processed data. Briefly, intensities were background corrected by plotting the Tm and observed intensity of a pool of random probe sequences. A loess model was fit to this plot and used to interpolate the background hybridization for every probe on the array according to its known Tm. Negative values that resulted from this procedure were set to 0. Arrays were normalized with the quantiles normalization procedure of GCRMA in Bioconductor. Finally a value of 16 was added to all values to stabilize variance.

Each of these archives contains a number of files:

I.) [ALEXA_data_###.txt](#)

These files contain probe-level intensities for all 385,000 probes in the ALEXA_hs_35_35h validation design. Each probe is one of the following types: Exon, Exon Junction (Exon-Exon), Exon Boundary (Exon-Intron or Intron-Exon), Intron, or Control-Negative. Probe and ProbeSet IDs correspond to those in the ALEXA_35_35h database and can be used to gather additional information on each probe. Additional information for each probe is provided.

II.) [Affymetrix_data_###.txt](#)

Affymetrix data for all ~2500 genes targeted by the ALEXA_hs_35_35h validation design. Since Affymetrix's Exon Array does not include junction or boundary probes, all probes are of the type: Exon, Intron or Control-Negative. A probe was only considered 'Exonic' if it mapped within an EnsEMBL exon. Probe and probeset IDs correspond to the original Affymetrix probe IDs and can be used to get additional information from Affy's annotation files.

III.) [AA_affy_data_###.txt](#) AND [AA_alex_data_###.txt](#)

To allow head-to-head comparisons of expression estimates between the ALEXA and Affymetrix designs a set of common probesets was defined. These were called 'AA' or 'Affy-ALEXA' probesets. Each 'AA' probeset corresponds to a single EnsEMBL exon (or portion thereof) which was targeted by at least one probe in BOTH platforms. Intensities from each of these two files with the same AA_probeset_ID can thus be compared directly, as they are both measuring the same target exons.

4.) Result tables

Six excel spreadsheets are provided as detailed summaries of the analysis. Complete summaries of expression at the gene level are provided for both the Affymetrix and ALEXA experiments ('[Affymetrix Data - Final Summary.xls](#)' and '[ALEXA Data - Final Summary.xls](#)'). **NOTE:** These files contain web links to custom UCSC tracks. In order to display these tracks as they appear in the manuscript, turn off the option to 'Display description above each track', in the 'configure' area of the UCSC browser.

A complete list of the significant differential expression events, each corresponding to a single probeset ID (individual exons, exon junctions and exon boundaries) are provided ('[Significant Differential Expression Events](#)'). This is basically the complete set of data which Table 2 in the manuscript summarizes. A complete list of those significant differential expression events which are most likely to correspond to an actual alternative transcription event (not just DE of the entire gene) is also provided ('[Significant Differential AT Events.xls](#)').

A detailed summary of all significant genes (MTP corrected p-value < 0.05) which also showed a 4-fold or greater change are provided. ('[Top Differential Gene Expression Events - 4Fold.xls](#)'). An abbreviated version of this table is presented as **Supplementary Table 2** in the manuscript. Finally a detailed summary of 22 manually examined differential alternative transcription events is provided ('[Top Differential AT Events.xls](#)'). An abbreviated version of this table is presented as **Supplementary Table 3** in the manuscript.

5.) Control Gene List – Human Housekeeping Genes

The file '[Human_Housekeeping_GeneList.txt](#)' contains a list of 96 housekeeping genes. These genes were used for various comparisons of the ALEXA and Affymetrix Exon Array platforms. These genes were selected by Affymetrix and have been used on several of their expression array platforms. Because of their apparently fundamental functions, they are assumed to be expressed in most tissues. In the Affymetrix exon array platform, these genes are comprehensively targeted with probes in their intronic and exonic regions. In the ALEXA_hs_35_35h validation design, these same genes were also

targeted with both exonic and intron probes. Since these genes are expected to be expressed, the exonic probes act as positive controls. Since introns are removed in mature transcripts, the intronic probes act as negative controls. Together these two categories of probes for ~100 housekeeping genes, act as an overall measure of the signal-to-noise ratio for each platform.