

apt-probeset-genotype Manual

Contents

- [Introduction.](#)
- [Quick Start - getting up and running.](#)
- [Beta Software - a word of caution.](#)
- [The Report File - explanation of contents.](#)
- [Program Options - command line options.](#)
- [FAQ - Frequently Asked Questions.](#)
- [Advanced Topics](#)
 - [BRLMM CHP file format - Same format, different values than DM genotyping CHPs.](#)
 - [Memory \(RAM\) issues - a common source of difficulty.](#)
 - [Custom Analyses - how to vary the way calls are made.](#)
 - [BRLMM transformations - different possible BRLMM clustering spaces.](#)

Introduction

apt-probeset-genotype is a program for making genotype calls from Affymetrix SNP microarrays using the model based algorithm BRLMM (pronounced "B-realm"). It uses a model based approach similar to the RLMM (pronounced "realm") model developed by Nusrat Rabbee and Terry Speed (<http://www.stat.berkeley.edu/users/nrabbee/RLMM/>) but with a Bayesian extension, hence the "B".

As BRLMM is a model based algorithm it needs to be run on multiple CEL files at once to estimate probe effect and SNP cluster parameters. For Mapping 500K data it is advisable to run on at least 50 distinct samples (excluding replicates) and ideally on about 100.

The beta version of the software has been developed and tested exclusively on Mapping 500K data. It is possible that it may also improve performance on Mapping 500K Early Access and Mapping 100K data but that has not yet been tested.

QuickStart

We illustrate the most basic way to run apt-probeset-genotype with an example. This example runs an analysis under the default parameter settings to generate a set of CHP files containing BRLMM genotype calls and confidence scores. The requirements are:

- A collection of CEL files to process. For this example the CEL files are all located in a directory called 'cel_sty'. Note that the program will run *much faster* if the CEL files are located on a local disk as opposed to being read across a network. For better overall performance it is advisable to filter

CEL files, running the BRLMM analysis only on those passing a certain specification based on the DM call rate. For Mapping 500K data the recommended per-chip spec is a call rate of 93% using DM at a threshold of 0.33.

- The CDF file corresponding to the array type of the CELs. For this example the CDF file is called 'Mapping250K_Sty.cdf'. It can be downloaded from <http://www.affymetrix.com/support/technical/libraryfilesmain.affx>
- A chrX file consisting of the SNP IDs for all non-pseudo-autosomal chrX SNPs on the array. For this example the chrX file is called 'Mapping250K_Sty.chrx'. It can be downloaded from the Affymetrix product page at <http://www.affymetrix.com>
- A directory where results will be stored. For this example it is called 'results_sty'

On unix systems a command using the default parameters looks like:

```
apt-probeset-genotype \  
-c Mapping250K_Sty.cdf \  
--chrX-snps Mapping250K_Sty.chrx \  
-o results_sty \  
cel_sty/*.CEL
```

The output will consist of a report file with some summary statistics about each chip analyzed and CHP format files written to a directory named chp in the specified output directory, in this example results_sty/chp. On Microsoft Windows the DOS prompt does not support wildcard expansion and the preferred method is to supply a text file with the path to the cel files via the '--cel-files' option (see below for details of file format). The Microsoft Windows DOS prompt does not allow a continuation of a command with the '\' character as does unix either.

On Microsoft Windows a command using the default parameters looks like:

```
apt-probeset-genotype -c Mapping250K_Sty.cdf --chrX-snps  
Mapping250K_Sty.chrx -o results_sty --cel-files  
cel_file_list.txt
```

apt-probeset-genotype runs 100 CELs in 1-2 hours on a 3GHz 2Mb RAM machine using local disk.

WARNING: apt-probeset-genotype will overwrite any existing output files it finds. If you wish to keep existing results make sure to specify a different output directory name.

Full details on the available options can be found [below](#); here we explain a few of the options used in the most common variants of the above workflow:

```
--txt-output      Output genotype calls and confidences in a
                  directory named 'txt' under the specified
                  output directory.  This option creates a
                  single txt file per CEL file analyzed.  The
                  txt file has 11 lines of header followed by
                  tab-delimited text with one line per SNP and
                  three columns:
                    Column 1: SNP_ID
                    Column 2: Genotype call
                              (-1=NoCall, 0=AA, 1=AB, 2=BB)
                    Column 3: Genotype call confidence
--table-output    Output genotype calls and confidences in a
                  pair of tab-delimited text matrices named
                  'brlmm.calls.txt' and
                  'brlmm.confidences.txt' in the specified
                  output directory.  The format of each file
                  is a few comment lines prefixed with '#',
                  a header line specifying the CEL file names
                  and then a line per SNP, the first field is
                  the SNP_ID and subsequent fields contain
                  genotype calls or confidences.  Calls are
                  encoded as -1=NoCall, 0=AA, 1=AB, 2=BB
--no-chp-output   Don't output CHP files.
-s, --probeset-ids  Allows for specification of a subset of SNPs
                   to which analysis will be restricted.  Run
                   time is directly proportional to the number
                   of SNPs so using this option can greatly
                   speed up the run time.  Note that it is not
                   possible to create CHP files with only a
                   subset of SNPs and so if this option is used
                   CHP output must be suppressed with the
                   --no-chp-output option.  The SNPs to analyze
                   are identified in a tab-delimited text file
                   which includes a column named 'probeset_id'
                   specifying the ids of the SNPs to be
                   analyzed.
```

Building upon the example above, here is an example in which only a subset of SNPs are analyzed and the results are written to a text table of genotype calls and a text table of call confidences. The subset of SNPs to be analyzed is specified in a tab-delimited text file called subset sty.txt, which must contain a column named 'probeset_id'.

```
apt-probeset-genotype \  
-s subset_sty.txt \  
--no-chp-output \  
--table-output \  
-c Mapping250K_Sty.cdf \  
--chrX-snps Mapping250K_Sty.chrx \  
-o results_sty \  
cel_sty/*.CEL
```

Important Note

The software is in a beta stage and is in the process of being finalized, so options, default parameters and even the genotype calls and confidences may change. If you encounter an issue please make sure to collect the following information and report the problem to devnet@affymetrix.com

- Contents of program log file (a file called apt-probeset-genotype.log is created in the output directory).
- Program output - cut and paste everything that the program reports to the screen.
- The specific command used.
- Type of machine (operating system, amount of memory).

The Report File

apt-probeset-genotype creates a summary report file in the output directory with file name extension '.report.txt'. The report file contains some summary information about each chip analyzed and is useful in getting a quick overview of the CELs analyzed. The format of the file is tab-delimited text with a header line followed by a line for each CEL file analyzed. The columns are all explained below, most users will be mainly interested in the first few entries. The additional entries are provided as potentially useful metrics to track and identify outlier chips and are expected to be mainly of interest to advanced users. The column entries are:

1. The CEL file name.
2. The estimated gender (based upon DM chrX calls at confidence of 0.33).
3. The BRLMM call rate at the default or user-specified threshold.
4. The percentage of SNPs called AB by BRLMM (i.e. the heterozygosity).
5. The percentage of SNPs called AA by BRLMM.
6. The percentage of SNPs called BB by BRLMM.
7. The average of the raw PM probe intensities.
8. The standard deviation of the raw PM probe intensities.
9. The average of the allele signal estimates (log2 scale).
10. The standard deviation of the allele signal estimates (log2 scale).
11. The average of the absolute difference between the log2 allele signal estimate and its median across all chips.

12. The standard deviation of the absolute difference between the log2 allele signal estimate and its median across all chips.
13. The average of the median absolute deviation (MAD) between observed probe intensities and probe intensities fitted by the model.
14. The standard deviation of the median absolute deviation (MAD) between observed probe intensities and probe intensities fitted by the model.
15. The average distance to the cluster center for the called genotype.
16. The standard deviation of the distance to the cluster center for the called genotype.

Options:

The explanation of options is divided into two - the first section deals with options which deal with input, output and other options which all still yield the default estimates of genotypes and confidences. The second deals with options which alter aspects of the analysis and will yield genotype calls and confidences different to the default.

The following options do not affect the analysis or genotype calls produced.

Input :	
-c, --cdf-file	File defining probe sets. Can be downloaded from www.affymetrix.com/support/technical/libraryfilesmain.affx
--cel-files	Text file specifying full or relative path of CEL files to process, one per line with the first line being 'cel_files'.
--chrX-snps	Text file specifying snps on chrX (non-pseudoautosomal region). Can be downloaded from the appropriate product page at www.affymetrix.com . Format consists of a single column of probeset names with column header name of 'all_chrx_no_par'. To help prevent accidentally using the wrong chrX file for a given chip type apt-probeset-genotype will exit with an error if there are any SNPs specified in the chrX file which are not found in the array. Use of this file is recommended for optimal performance on chrX but if you do not have or do not want to use it see the --chrX-force option.
-s, --probeset-ids	Allows for specification of a subset of SNPs to which analysis will be restricted. Run time is directly proportional to the number

of SNPs so using this option can greatly speed up the run time. Note that it is not possible to create CHP files with only a subset of SNPs and so if this option is used CHP output must be suppressed with the `--no-chp-output` option. The SNPs to analyze are identified in a tab-delimited text file which includes a column named 'probeset_id' specifying the ids of the SNPs to be analyzed.

Output:

<code>-o, --out-dir</code>	Directory into which result files and directories will be written. WARNING: any previously-generated results in the directory will be overwritten.
<code>--chp-output</code>	Create CHP format genotype calls and confidences in a directory called 'chp' under out-dir. Makes one CHP file per CEL file analyzed. [default: true]. A description of the CHP file format can be found at http://www.affymetrix.com/support/developer
<code>--no-chp-output</code>	Don't output CHP files.
<code>--table-output</code>	Output genotype calls and confidences in a pair of tab-delimited text matrices named 'brlmm.calls.txt' and 'brlmm.confidences.txt' in the specified output directory. The format of each file is a few comment lines prefixed with '#', a header line specifying the CEL file names and then a line per SNP, the first field is the SNP_ID and subsequent fields contain genotype calls or confidences. Calls are encoded as (-1==NoCall, 0==AA, 1==AB, 2==BB)
<code>--txt-output</code>	Output genotype calls and confidences in a directory named 'txt' under the specified output directory. This option creates a single txt file per CEL file analyzed. The txt file has 11 lines of header followed by tab-delimited text with one line per SNP and three columns: Column 1: SNP_ID Column 2: Genotype call

```
                (-1=NoCall, 0=AA, 1=AB, 2=BB)
                Column 1: Genotype call confidence
--summaries    Output the allele signal estimates for each
                allele in a file named
                brlmm.plier.summary.txt located in the
                specified output directory. The format of
                the file is tab-delimited text with one row
                for each allele of each SNP. The first
                column is the SNP_ID and the allele, the
                subsequent columns contain the estimated
                allele signal for each CEL file.
--feat-effects Output the empirically-determined feature
                (or probe) effects produced in the allele
                summarization step. Results are written to
                a file named
                brlmm.plier.feature-response.txt located in
                the specified output directory. This file
                can be used in future analyses with the
                --use-feat-eff option. The format of the
                file is tab-delimited text with one row per
                PM probe or PM,MM probe pair. The columns
                are:
                1 - probeset_id (i.e. SNP_id)
                2 - atom_id: identifier for PM probe or
                   PM,MM probe pair, unique within a SNP.
                3 - feature_id: identifier for PM probe or
                   PM,MM probe pair, unique within the
                   chip.
                4 - feature_response: the estimated
                   feature or probe effect.
--residuals    Output the residuals from the quantification
                method. Residuals are written as tab-
                delimited text and since there is one
                residual per PM probe per chip, analyzing
                all the probesets will generate huge output.
--write-sketch Write the estimated quantile (or sketch)
                normalization distribution to a file named
                quant-norm.normalization-target.txt, so that
                it can be re-used in future analyses with
                the --target-sketch option. Format of the
                file is text with the sorted probe
                intensities of the target distribution
                written one-per-line, with the first line
                being 'intensities'.
```

```
--write-prior Write the estimated generic prior
              information on SNP cluster centers and
              variances to a file named brlmm.prior.txt in
              the specified output directory. This file
              can be used in future analyses with the
              --prior-file option. The file format is
              tab-separated text with entries as follows:
              1 - 6 comma-separated values corresponding
                  to the cluster centers Values are
                  (center_x,center_y) for AA,AB,BB.
              2 - 9 comma-separated values corresponding
                  to the variances estimated for each
                  genotype. Values are (var_x,cov_xy,
                  ar_y) for AA,AB,BB
              3 - 36 comma-separated values
                  corresponding to the 6x6 variance
                  matrix estimated for the cluster
                  centers.

Other:
  --block-size How many probesets to process at once,
               useful when memory is limited. If set to 0
               program attempts to guess available RAM and
               set appropriately. For more info see the
               notes on memory in the advanced topics
               section below. [default: 0]
  -v, --verbose How verbose to be with status messages:
               0=quiet, 1=usual messages, 2=more messages.
  --version Output program version and quit.
  --explain Provide more detail on a particular analysis
            operation. Usage is '--explain <method>'
            where 'method' can be quant-norm, brlmm,
            plier or med-polish.
  -h, --help This message.
```

The following options will change the genotype calls produced.

```
-a, --analysis String specifying the kind of analysis to be
               applied. The default value is 'brlmm' which
               provides the default BRLMM implementation.
               Advanced users may wish to explore custom
               analyses with this option. For further
```

```

                                details see the 'custom analyses' section
                                below. Specifying 'brlmm' is an encoded
                                shortcut for the analysis string:
                                'quant-norm.sketch=50000,pm-
only,brlmm.transform=ccs.K=4'

                                Note that this -a option will determine the
                                base name for many of the output files
                                generated, such as those generated by the
                                options --table-output, --summaries,
                                --feat-effects and --write-prior.

--qmethod-spec Quantification Method to use for summarizing
alleles. Available methods are plier and
med-polish. Details on the options
available with each of these methods are
available by calling apt-probeset-genotype

with

                                the option '--explain <method>' where
                                <method> is 'plier' or 'med-polish'.
                                [default: plier.optmethod=1]

--max-score Maximum score (i.e. lowest confidence) at
which to make call, all calls of lower
confidence get no call. [default: .5]

-f, --force Ignore any disagreement between the array
types of the CEL files and the array type of
the CDF file.

--chrX-force Perform analysis even without using
--chrX-snps option. Note that this may
yield sub-optimal calls on chrX snps.

--prior-size How many probesets to use for determining
prior. [default: 10000]

--list-sample Only sample SNPs to be used in the prior
from list specified via --probeset-ids, not
entire chip. Useful in validation of
results from apt-probeset-genotype with

                                from external implementations.

--dm-thresh DM confidence threshold used for seeding
clusters. [default .17]

--dm-hetmult The DM algorithm's het dropout can be
ameliorated by adding a small constant to
the log likelihoods of all het calls, making
het calls more likely. This small constant
is called the 'DM het multiplier' and
setting it to something positive can help in
balancing het/hom performance for DM, which
```

```
is used to seed the clusters. [default = 0]
--use-feat-eff File defining a precomputed feature (or
                probe) effect to be used for each probe.
                Note that precomputed effects should only be
                used for an appropriately similar analysis
                (i.e. feature effects for a pm-only analysis
                may be different to feature effects for
                pm-mm). See --feat-effects option for
                details on file format.
--target-sketch File specifying a target distribution to
                use for quantile or sketch normalization.
                See --write-sketch option for details on
                file format.
--prior-file   File from which to load generic snp priors,
                see --write-prior option for details on file
                format.
```

Frequently Asked Questions

Q. What do I do when I don't have enough memory to process all the data?

A. You can manually set the `--block-size` command to specify how many probesets will be run at once. The program will then reduce memory by only loading those probesets into RAM. If the `block-size` option is unset the program will attempt to figure out how much available RAM you have and run in that memory. To fit in memory the program will often need to read the original CEL files multiple times. Also, if doing a quantile normalization try using a sketch (or subset) of the chip for the normalization. Sketch normalization is the default so this would only apply if you are using non-default options.

Q. How can I make apt-probeset-genotype run more probesets per iteration?

A. If you manually set the `--block-size` flag apt-probeset-genotype will not try to guess the amount of probesets to be run per iteration and will use the supplied value instead.

Q. The program died with an error message like "DmListener::getGenoCall() - Can't find genotypes for name: SNP_A-1780432". What does this mean?

A. This is symptomatic of having specified the wrong chrX file for the analysis. In order to reduce the likelihood of accidentally using the wrong chrX file apt-probeset-genotype checks to make sure that all the SNPs specified in the chrX file are present on the chips being analyzed. If it finds a SNP present in the chrX file that is not identified in the CDF file it will die with the above message. Note that if you want to bypass the requirement of a chrX file you can use the `--chrX-force` option.

Q. The program died and I got an error message saying "Killed". What does this mean and what can I do?

A. Linux has a "feature" that it will promise more memory than it actually has in the hope that many programs won't actually be using all their memory at once. However, if linux does run short of memory it will start killing programs arbitrarily. You can read more about linux's OOM (out of memory) killer at LWN.net.

Q. Why does apt-probeset-genotype require information regarding SNPs on chromosome X?

A. The SNPs on chromosome X are evaluated separately for XX (female) and XY (male) individuals as the intensity estimates for the males will generally be lower on X due to one missing chromosome. The prior is also adjusted to remove the het center as XY individuals should only have hom calls on the X chromosome. Using the tried and trusted gender estimation method employed in the GTYPE software, individuals are called male if less than 7.5% of the snps on X are called as hets by the initial DM calls using a .33 confidence threshold.

Advanced Topics

BRLMM CHP File format.

The CHP files produced by the BRLMM algorithm are different than those from DM. Historically the genotyping CHP file is closely tied to the DM model and while BRLMM uses the same format for backward compatibility it is important to note that the interpretation of many fields are different. Below are the names of the fields and corresponding BRLMM values that are stored in them.

- AlleleCall: Same as DM ALLELE_NO_CALL, ALLELE_A_CALL, ALLELE_AB_CALL, ALLELE_B_CALL.
- Confidence: Value between 0 and 1. The ratio of the distance to the closest cluster to the second closest cluster. Lower values are more confident.
- pvalue_AA: Mahalanobis distance to AA cluster.
- pvalue_AB: Mahalanobis distance to AB cluster.
- pvalue_BB: Mahalanobis distance to BB cluster.
- pvalue_NN: Not valid for BRLMM, value FLT_MAX inserted as placeholder.

The following parameters are saved in the CCHPFileHeader object:

- het-mult: Het multiplier.
- iterations: Number of BRLMM iterations.
- iter-thresh: Maximum score used during BRLMM iterations.
- K: Scaling parameter in transformations.
- transform: Which transformation was used for A & B allele.
- prior-weight: Number of pseudocounts used as weight for prior.

- prior-mincall: Minimum number of observations for seeing each AA,AB,BB cluster to be used in prior. Minimum is 2 as can't get variance for less than 2.
- lowprecision: Were summary values rounded off before transforming. This is only used for regression testing to be compatible with R prototype.

A word about memory (RAM)

The most common challenge people have running apt-probeset-genotype is with RAM. apt-probeset-genotype will attempt to split up jobs into the amount of RAM that appears free on your computer. The job is split by subdividing the analysis into blocks of probesets (or SNPs). Small block sizes will subdivide the job more and require less memory at the expense of having to read the CEL files more frequently. On the other hand, using a smaller number of large blocks will require more memory but will place minimal load on reading CEL files.

The default behavior is for apt-probeset-genotype to estimate the optimal block size based upon the amount memory that appears to be free at the time the job begins. This default behavior can be overridden by the user with the use of the --block-size option, which specifies the number of probesets to be processed at one time. For example, specifying --block-size=20000 will analyze the data in batches of 20,000 probesets at a time.

A problem that may be encountered (especially on a multi-user or multi-processor system) is running out of memory when a run of apt-probeset-genotype is initiated and then another big-memory process is started afterwards. In this circumstance the first instance of apt-probeset-genotype sees substantial free memory and chooses a large block-size, but then the second process grabs more of the memory and the first run of apt-probeset-genotype runs out of memory. This problem can be addressed by planning the work load on your machine and/or using an appropriately small block size with the --block-size option.

RAM usage (in bytes) for Mapping 500K data can be estimated by the following equation:

$$RAM \approx C \times [(B \times P \times (F + 1)) + (S \times F)] + (B \times N) + (B \times D) + K$$

- C is the number of CELs being analyzed
- B is the number of probesets being processed in each batch
- P is the average number of probes in a probeset (about 27 for Mapping 500K)
- F is the number of bytes in a float (4).
- S is the number of probes being used to approximate a full quantile normalization and is 50000 by default.
- N is the average size in bytes of SNP_IDs, for Mapping 500K it is about 25 bytes.
- D is the size in bytes for the probeset information (approximately 800 bytes, but dependent on pointer size).
- K is constant for hashes, book keeping etc. about 200MB on 32 bit system.

Below are some guidelines about how many probesets to run at once (i.e. the --block-size) in 1.9 Gig of RAM as a function of number of CEL files:

- 50 CELs -> 140,000 probesets (i.e. half the dataset for a Mapping 250K chip plus some for prior).
- 100 CELs -> 70,000 probesets
- 150 CELs -> 47,000 probesets
- 200 CELs -> 35,000 probesets

Note that the above recommendations won't use all of the 1.9 gigs of RAM. In addition to needing a relatively large amount of memory the program also needs relatively large blocks of contiguous memory and as RAM usage approaches the maximum available these get harder and harder to find. If you've got memory to spare the amount of RAM to run all the data at once as a function of Chips:

- 50 CELs -> 1.9 Gb
- 100 CELs -> 3.3 Gb
- 150 CELs -> 4.6 Gb
- 200 CELs -> 6.0 Gb
- 250 CELs -> 7.3 Gb
- 300 CELs -> 8.7 Gb

Note that on most 32 bit (i.e. Intel Pentium and Xeon) systems you can't use more than ~2 Gig of RAM with a single process, even if there is more available.

Custom Analyses:

While aliases for common analysis such as BRLMM with default parameters are provided it is possible to construct custom analyses on the command line. There are both program options and analysis parameters that can be set to affect the results. Most people are familiar with the standard method for setting program options, but the specification of the analysis method and its parameters in apt-probeset-genotype works a little differently. The method for setting custom parameters to the analysis involves supplying a text representation of the analysis and parameters desired. This enables flexibility as each piece of an analysis is self-contained and they can be (almost) arbitrarily combined. Note that when using a custom analysis rather than an alias it is necessary to specify the entire analysis and not acceptable to pass custom parameters to the alias. For example, if you wanted to change the number of iterations BRLMM performs you would have to specify 'quant-norm.sketch=50000,pm-only,brlmm.iterations=1' rather than just typing 'brlmm.iterations=1'

The current full default BRLMM analysis is: 'quant-norm.sketch=50000,pm-only,brlmm' where there can be multiple chipstream modules (in this case a single quant-norm) separated by commas and the last two entries are the pm adjuster (pm-only) and quantification method (brlmm). Parameters to a particular step in the analysis are supplied in key=value pairs and separated by periods. For example 'quant-norm.sketch=50000' indicates that the chips should be quantile normalized and that a sketch (subset of total data) of size 50000 should be used

to do the normalization. Using a sketch can significantly reduce the amount of memory needed with minimal impact on normalization values. To do quantile normalization with just the PM probes and resolve ties in the same manner as bioconductor's RMA version of quantile normalization you would specify 'quant-norm.sketch=50000.bioc=true.usepm=true'. All of the parameters possible can be seen by using the --explain option in conjunction with the name of the module (i.e. apt-probeset-genotype --explain quant-norm).

So a few examples custom analyses would be:

'pm-only,brlmm.transform=rvt' - No normalization, use rvt space for clustering in brlmm.

'med-norm,pm-mm,brlmm.het-mult=.9' - Do a median normalization, use a PM-MM adjustment for probes and a het multiplier of .9 to try and balance hom/het calls.

'rma-bg,quant-norm.sketch=50000.usepm=true.bioc=true,pm-only,brlmm.K=4.tranform=CCS' - Do an RMA style background subtraction followed by an RMA style quantile normalization using a subset of 50000 data points followed by BRLMM in CCS (contrast centers space) space with K = 4.

For BRLMM the parameters are:

```
apt-probeset-genotype --explain brlmm
brlmm: Do genotyping calls using the BRLMM (Bayesian RLMM)
algorithm.

Parameters:
  'K'                Scale parameter used in CCS and CES.
                    transformations [default = 4].
  'het-mult'         Number to balance het calls with to balance
                    performance on het/hom calls [default = 1
                    (no effect)]
  'iter-thresh'      Maximum confidence score to use when doing
                    iterations internally [0,1] [default = .3].
  'iterations'       Number of times to iterate BRLMM classifier,
                    feeding in new calls from previous iteration
                    [default=0]
  'prior-mincall'    Minimum number of genotypes per cluster for
                    inclusion in prior estimation, must be >= 2
                    [default = 2]
  'prior-weight'     What psuedocount weight should the prior have?
                    [default = 40]
  'transform'        What transformation of initial data are we
                    feeding into the classifier? {'CCS','CES',
```

```
'MvA', 'RvT' } [default = 'CCS']
```

BRLMM Transformations:

There are a number of different transformations that are implemented for different spaces which can be specified via the transform parameter to BRLMM and are detailed below. For all of these transformations **A** and **B** denote the intensity of the A and B alleles respectively as estimated by the quantification method (such as plier or RMA). **X** and **Y** denote the new coordinates that **A** and **B** will be transformed into.

- **CCS** = Contrast Centers Stretch:
$$X = a \sinh(K * (A - B) / (A + B)) / a \sinh(K)$$
$$Y = \log_2(A + B)$$
- **CES** = Contrast Extremes Stretch:
$$X = \sinh(K * (A - B) / (A + B)) / \sinh(K)$$
$$Y = \log_2(A + B)$$
- **MvA** = Minus Vs Average
$$X = (\log_2(A) + \log_2(B)) / 2$$
$$Y = \log_2(B) - \log_2(A)$$
- **RvT** = R vs Theta (polar coordinates)
$$X = \arctan(A/B)$$
$$Y = \ln(\sqrt{A^2 + B^2})$$

Copyright Notice

©2006 Affymetrix, Inc. All Rights Reserved.

Affymetrix®, GeneChip®, HuSNP®, GenFlex®, Flying Objective™, CustomExpress®, CustomSeq®, NetAffx™ Tools To Take You As Far As Your Vision®, The Way Ahead™ Powered by Affymetrix™ GeneChip-compatible™ and Command Console™ are trademarks of Affymetrix, Inc.

For Research Use Only, Not For Use in Diagnostic Procedures

