

NAME

gpextract – assemble data for GenePool analysis

SYNOPSIS

```
gpextract
  -c ChipType
  -f CDFfilename
  -s SampleType
  -i IlluminaFiles
  -l CELfilename
  -e EnzymeString
  [-d ExpDir]
  [-h]
  [-m]
  [-n ChannelNormalizeMethod]
  [-v]
  [-F]
  [-L FilterLimit]
  [-S FilterStdDev]
  [-v]
```

DESCRIPTION

gpextract is the first of 3 programs in the GenePool system. Together the three programs allow for detection of shifts in relative allele frequency between two pools of genomic DNA hybridized on either Affymetrix GeneChip or Illumina BeadChip SNP genotyping microarrays. For Affymetrix analysis, **gpextract** uses Affymetrix's Fusion library to extract intensity values from the Affymetrix CEL files and write them out to a more compact customized binary file format that will be read by **gpanalyze**. For Illumina analysis, **gpextract** used the raw **.txt** Illumina files to extract intensity values and writes them to a more compact customized binary file format that will be read by **gpanalyze**. Note that for Illumina chips **gpextract** may report extracting intensity values for more SNPs than is stated for a given chip since there are SNPs on the chip that are used only for quality control and testing purposes. Additionally, each Illumina chip may have a different number of SNPs and intensity values since the distribution of beads on the chip is random.

The basic sequence of events within **gpextract** is that the Affymetrix Fusion library is used to read CEL and CDF files and assemble feature intensities by SNP and then the intensities are written out in a custom format to a new file with a **.gpb** (GenePool Binary) extension. In most cases it is simpler to then use **gpcommand** which is a wrapper for **gpextract** rather than using **gpextract** directly. **gpcommand** automatically creates a number of auxiliary files required by **gpanalyze**. If you use **gpextract** directly, you will have to create and correctly format these files by hand. These files are dealt with in more detail in the man page for **gpanalyze**.

Negative Illumina bead values

It appears that bead-level values from Illumina chips are not completely raw intensities as some beads have been observed to have negative values. The current behaviour of **gpextract** on encountering a bead with a negative value is to drop the bead entirely.

OPTIONS**-c ChipType**

This option is required for all chip types. Specify **-c 0** for an Affymetrix chip type. Specify **-c 1** for an Illumina chip type.

-d ExpDir

During the extraction process each SNP name encountered is written to an output file that has "SnpNames.txt" appended to the file name. The directory specified with this option determines where that SNP names file is to be created. The default if no directory is specified is to place the SNP names file in the current working directory. For Affymetrix chips the SNP names file should be the same for chips within the same chipset (for example 500,000 SNP chips). For a given Illumina chip, a given

SNP may have zero beads and so may be absent from the data file. Thus the SNP names file may be different between Illumina chips within the same chipset. Therefore it may be necessary to store the SNP names for each chip extracted separately.

-e EnzymeType

This option is necessary for Affymetrix chips only. This is a string that is incorporated into the name of the output files so it is not critical to the functioning of the analysis however if you do not adopt a standard way of naming the enzymes, later analysis stages will become unnecessarily complicated. We suggest using the following strings: *Xba*, *Hind*, *Nsp*, and *Sty*.

-f CDFFileName

This options is necessary for Affymetrix chips only. CDF files are created by Affymetrix and are effectively the "decoder rings" for the corresponding CEL files. The CDF file stores the identity and x-y location of every probe in the matrix of features on the chip. This is critical because a CEL file just contains a matrix of intensity values so without the CDF file, the CEL file is just a big string of meaningless numbers. Effectively CDF files are tied to the manufacturing process - if the layout of a chip changes then a new CDF file must be created to document the new layout. The **gpextract** user must supply the correct CDF file but because we use the Affymetrix Fusion library to read CEL and CDF files, the library should spot cases where a CDF file does not match the supplied CEL file.

-h Print out a help screen shwoing the short-form descriptions of each command line option.

-i IlluminaFiles

This option is required for Illumina chips only. This is the filename that stores the number and names of each file that contains raw data for a given strip. The data files for the Illumina chips are **.txt** files typically numbered according to their associated strip. The specified file must on the first line have the number of strips (i.e the number of files for the given Illumina chip). Then on each subsequent line there is the name of the strip file. The strip files must be ordered so that the SNP codes within the **.txt** files are ordered in increasing order. Typically this achieved by naming the first strip first, the second strip second and so on.

-l CELFileName

This option is necessary for Affymetrix chips only. This is the name of the CEL genotyping file to be processed for Affymetrix chips. CEL file contains intensities for features ("spots") on the chip. CEL files are derived from DAT files which are effectively big scanned images of the chip and so contains intensities for pixels. CEL files tend to be about one order of magnitude smaller than the corresponding DAT files. You need the CEL files for GenePool, DATs are no use.

-m For Affymetrix chips, the extraction routines normally only keep the perfect match probe intensity values however if the **-m** option is used, the mismatch probe intensities are also transferred into the **.gpb** files. **gpanalyze** can automatically distinguish between **.gpb** files that contain only perfect match intensities and those that contain both perfect and mismatch intensities.

-n ChannelNormalizeMethod

This option can only be used with Illumina chips. A value of **1** indicates that for each bead and channel, the signal intensity is to be divided by the mean of the respective channel. The default behaviour is to do no normalization.

-s SampleType

This string must have one of the values *Case* or *Control*. Like *EnzymeType*, this string is used to name the output files.

-v Print the version string for **gpextract**. It will typically be something like: *TGen GenePool gpextract version 0.0.2*.

-F The extraction routines are run but no output file are created.

-L FilterLimit

This option only works with Illumina chips (with option **-c 1**). The specified value is the absolute lowest signal intensity for both the channels. For a given bead, if either of the channels signal intensity is below the given value, the bead is skipped.

-S FilterStdDev

This option only works with Illumina chips (with option **-c 1**). The specified value is an intensity threshold expressed in Standard Deviations and for a given bead, if either of the channel signal intensities is more than the specified number of standard deviations above or below the mean channel intensity then the bead is skipped.

- V** Verbose mode. By default the program does not display any progress messages. This option can be specified multiple times to enable higher levels of verbosity, for example specifying **-V -V** on the command line would enable verbose level 2. At level 3 the individual RAS scores are dumped to the screen so level 2 is probably as high as you'll usually want to set the verbose level.

EXAMPLES

Example 1. Extract information from Affymetrix (**-c 0**) CEL file **Case_1.CEL (-l)** which is a 10K chip run using Xba (**-e**) enzyme with a case sample (**-s**) interpreted using the Affymetrix CDF file **Mapping10K_Xba142.CDF (-f)**. This example shows the minimal set of options (**-c -l -f -e -s**) required in order to extract information from an Affymetrix GeneChip.

```
gpextract -c 0 -l Case_1.CEL -f Mapping10K_Xba142.CDF -e Xba -s Case
```

KNOWN ISSUES

Please see the `genepool(1)` manpage.

AUTHORS

Waibhav Tembe <wtembe@tgen.org>
John Pearson <jpearson@tgen.org>
David Craig <dcraig@tgen.org>
Matt Huentelman <mhuentelman@tgen.org>
Nils Homer <nhomer@tgen.org>

SEE ALSO

`genepool(1)`, `gpanalyze(1)`, `gpcommand(1)`.

COPYRIGHT

GenePool is copyright 2006 by The Translational Genomics Research Institute. All rights reserved. This License is limited to, and you may use the Software solely for, your own internal and non-commercial use for academic and research purposes. Without limiting the foregoing, you may not use the Software as part of, or in any way in connection with the production, marketing, sale or support of any commercial product or service. For commercial use, please contact licensing@tgen.org. By installing this Software you are agreeing to the terms of the LICENSE file distributed with this software.

In any work or product derived from the use of this Software, proper attribution of the authors as the source of the software or data must be made. The following URL should be cited:

<http://bioinformatics.tgen.org/software/genepool/>