



**NAME**

**gpanalyze** – analyze GeneChip genotyping data from pooled genomic DNA

**SYNOPSIS**

```
gpanalyze
  [-F FILENAME] [--ExperimentFile=FILENAME]
  [-a FILENAME] [--AllRASOutFile=FILENAME]
  [-b INTEGER] [--BeadThreshold=INTEGER]
  [-c Weighted] [--Weighted=Weighted]
  [-d INTEGER] [--DistanceType=INTEGER]
  [-f INTEGER] [--RankThreshold=INTEGER]
  [-m INTEGER] [--DistMatrix=INTEGER]
  [-o FILENAME] [--OutFile=FILENAME]
  [-p Cluster] [--Cluster=Cluster]
  [-r INTEGER] [--RASMethod=INTEGER]
  [-s SNP_ID] [--SNPName=SNP_ID]
  [-u FILENAME] [--OutFileRASMean=FILENAME]
  [-K FILENAME] [--KCorrectFile=FILENAME]
  [-R FILENAME] [--MapFile=FILENAME]
  [-D] [--Defaults]
  [-h] [--Help]
  [-I] [--SysInfo]
  [-M] [--CheckFormat]
  [-T] [--Study]
  [-V] [--version]
  [-w INTEGER] [--MinWin=INTEGER]
  [-W INTEGER] [--MaxWin=INTEGER]
  [-A WinRankThld] [--WinRank=WinRankThld]
  [-t INTEGER] [--Stage=INTEGER]
  [--usage]
```

**DESCRIPTION**

**gpanalyze** is the second of three programs in the GenePool system. Together the programs allow for detection of shifts in relative allele frequency between two pools of genomic DNA hybridized on Affymetrix GeneChip of Illumina BeadChip SNP genotyping microarrays. The first GenePool program (**gpextract**) extracts intensity values from the Affymetrix CEL files or Illumina .txt files and aggregates them into a more compact customized binary file format that can be read by **gpanalyze**. **gpextract** relies on Affymetrix's FUSION library to read data from Affymetrix CEL files. **gpanalyze** takes the intensity values from the custom binary file and uses a variety of distance measures to assign a probability score to each SNP where the score indicates how unlikely is the observed difference in allele frequency between the hybridizations for the two DNA pools.

**Experiment File**

The Experiment file specified with the **-F** option is the key to the operation of **gpanalyze**. The Experiment file specifies the names of other auxilliary files that list the names of the case and control data files and the names and order in which SNPs occur in the data files. It also serves to associate files from a given chip platform. A brief description of the Experiment file appears below in the documentation for the **-F** option but this section contains the definitive documentation.

The Experiment file contains one or more tab-separated lines of the form:

```
CasesFile NumberOfCases ControlsFile NumberOfControls SNPNamesFile
```

If you are running **gpextract** directly then you will have to manually create an Experiment file and the first two of the three files listed within it (**CasesFile**, **ControlsFile**). However if you are using **gpcommand** the the Experiment file will be created for you based on information from the configuration file. Additionally, if the **-X** option is used, the graphical interface will create the Experiment file for you. The contents of an Experiment file for a 500K experiment might contain the following 2 lines:

```
AD_Cases500K_Nsp.txt 5 AD_Controls500K_Nsp.txt 5 500K_NspSnpNames.txt
AD_Cases500K_Sty.txt 5 AD_Controls500K_Sty.txt 5 500K_StySnpNames.txt
```

The 5 fields on each line are: (a) the name of a file containing the list of all data files from chips run with the pooled cases; (b) the number of lines (files) in a.; (c) the name of a file containing the list of all data files from chips run with the pooled controls; (d) the number of lines (files) in c.; (e) the name of a file that contains the IDs of the SNPs in exactly the order they occur in the files listed in a. and c.

There is one line for each chip type so if you only have data from one platform (for example Affymetrix 10K or Illumina HumanHap500) then the Experiment file will contain only one line. However, if your platform is the Affymetrix 100K or 500K then the Experiment file should contain 2 lines - one for each enzyme type.

Looking again at the example Experiment file shown above, the **AD\_Cases500K\_Nsp.txt** file that appears as the first parameter on the first line might contain text like:

```
AD_CA2_A_NSP.CEL.gpb
AD_CA3_A_NSP.CEL.gpb
AD_CA4_A_NSP.CEL.gpb
AD_CA5_A_NSP.CEL.gpb
AD_CA6_A_NSP.CEL.gpb
```

There are 5 file names listed, as specified by the second parameter of the first line of the Experiment file, and each file holds data for pooled cases on the Nsp enzyme chip. The **AD\_Controls500K\_Nsp.txt**, **AD\_Cases500K\_Sty.txt**, and **AD\_Controls500K\_Sty.txt** files would all have similar contents. These files that contain the names of the case and control data files for each platform are automatically created by **gpcommand** but must otherwise be created manually by the user. This is not true of the files that contain the SNP names (**500K\_NspSnpNames.txt** and **500K\_StySnpNames.txt** in our example above). The SNP names files are created directly by **gpextract** since they list the names of the SNPs in the order that the data is found in the **.gpb** files output by **gpextract**. For Illumina chips, the SNP names file generated by **gpextract** may not include all of the SNP names. Since the beads are randomly distributed on the chip, some SNPs may have no beads and thus no intensity values. Also, there are SNPs on the chip that are used by Illumina only for quality control so Illumina chips may have more SNPs than expected. Therefore it is important to use a precompiled SNP Names file when using Illumina chips. The contents of one of these SNP Names files would be approximately 250,000 lines of the form:

```
SNP_A-2128504
SNP_A-2183388
SNP_A-1999657
SNP_A-2298105
```

We are aware that the Experiment file is a cumbersome system and in the next version of GenePool the Experiment file will disappear entirely and **gpanalyze** will get all the information it needs directly from the Unified Experiment File (see the **TO DO** section of the **genepool** manpage for more details).

In the meantime, if at all possible, you should let **gpcommand** do the heavy lifting to create Experiment files for you.

## OPTIONS

Many of the options listed here show short and long formats, for example the **-p** option can also be specified using **--Cluster**. This behaviour relies on the **argp** library which does not appear to exist in the GNU Compiler Collection (**gcc**) on all platforms so if you have any problems trying to use the long format options, try swapping back to the short forms.

Mandatory or optional arguments to long options are also mandatory or optional for any corresponding short options.

**-a, --AllRASOutFile=FILENAME**

This option causes the RAS values to be printed out to a new file for all probe quartets and for all SNPs. The order of quartets is determined by the order they are stored in the **.gpb** file by the **gpextract** command which in turn stores the quartets in the order they are supplied by the code from the

Affymetrix Fusion library. All of the scores for the cases are printed then all of the controls. The order of cases and controls is according to the order in the **CasesFile** and **ControlsFile** that are named in the **ExperimentFile** (see documentation for the **-F** option). If single SNP mode (**-s** option) has been turned on then the RAS values will only be printed out for that one SNP, otherwise, the RAS values for all SNPs will be printed. The **-a** option takes precedence over the **-u** option.

- b, --BeadThreshold=INTEGER**  
This option can only be used with Illumina chip analysis. This option specifies a lower limit on the number of beads acceptable for either the cases or controls. For example, if we have **-b 5** then if the combined number of beads for the case chips is less than 5, we skip this SNP. Likewise for the control chips.
- c, --Weight=WEIGHT**  
The default value is **0** indicating that intensity values should not be weighted. A value of **1** indicates that the intensity values should be weighted. This option only has meaning when Consistency is used as the clustering method (**-p 1**) otherwise it is ignored.
- d, --DistanceType=INTEGER**  
There are currently 3 distance methods available. A value of **0** (the default) indicates that a basic Euclidean method should be used, a value of **1** indicates that a Manhattan method should be used, and a value of **2** indicates that a modified Manhattan method should be used. This option only has meaning when Silhouette is used as the clustering method (**-p 0**) otherwise it is ignored.
- f, --RankThreshold=INTEGER**  
This option specifies the number of SNPs to output. For example, if you only wish to see the top 5000 SNPs then use **-f5000**. This option is particularly useful in multistage analyses.
- m, --DistMatrix=MATRIX\_TYPE**  
There is currently only one distance matrix implemented (Pairwise) so this option is really just a placeholder for further expansion of **gpanalyze**.
- o, --OutFile=FILENAME**  
The name of the output file.
- p, --Cluster=CLUSTER\_METHOD**  
There are six clustering methods currently available. A value of **0** (the default) indicates that Silhouette should be used, a value of **1** indicates Consistency Unidirectional, **2** indicates Consistency Directional, **3** indicates Centroid Distance, **4** indicates Dunn Index, and **5** indicates T-test.
- r, --RASMethod=INTEGER**  
There are currently three methods implemented for calculating RAS scores. A value of **0** uses the formula  $A/(A+B)$ , where A is the intensity of the probe for allele A and B is the intensity of the probe for allele B. A value of **1** uses the formula  $A^k/(A^k+B)$ , and a value of **2** uses the formula  $\arctan(B/A)$ . All 3 values are valid for Affymetrix chips but only value 2 works for Illumina chips. A value of **1** for the RASMethod must be accompanied by the **-K** option which specifies a file containing predefined AA, AB, and BB average values allowing each quartet for each SNP to be k-corrected and normalized. See the **-K** option for more information on the format of the input file. In practice all methods in this option are dependent on the functionality of **gpextract** which discards all mismatch probe intensities unless the **-m** option is specified.
- s, --SNPName=SNP\_ID**  
If this option is used then only the results for the single SNP specified (Affy SNP ID) are output rather than the default behaviour which is to analyze all SNPs.
- t, --Stage=INTEGER**  
The Stage number for multi-stage pipelines where **gpanalyze** is invoked multiple times with each invocation taking as input the output from the previous stage.
- u, --OutFileRASMean=FILENAME**  
This option causes the mean RAS value for each SNP to be printed out to a new file for each sample. All of the scores for the cases are printed then all of the controls. The order of cases and controls is

according to the order in the **CasesFile** and **ControlsFile** that are named in the **ExperimentFile** (see documentation for the **-F** option). If single SNP mode (**-s** option) has been turned on then the RAS values will only be printed out for that one SNP, otherwise, the RAS values for all SNPs will be printed. The **-a** takes precedence over the **-u** option.

**-w, --MinWin=INTEGER**

**-W, --MaxWin=INTEGER**

**-D, --Defaults**

Display a brief help screen that lists the defaults for all parameters.

**-F, --ExperimentFile=FILENAME**

The Experiment file is created by **gpcommand** based on information from the configuration file and the format is detailed in the *Experiment File* section above. If you are running **gpextract** directly then you will have to manually create an Experiment file. **gpanalyze** cannot run without an Experiment file so if this option is not specified then a default filename is assumed: **Experiment.txt**. This is the name of the file created by **gpcommand** with the **--gpextract** option so if you check the log file written by **gpcommand** when it is called with the **--gpanalyze** option, you will not see an Experiment file specified.

**-K, --KCorrectFile=FILENAME**

This option is specific to Affymetrix chips. It specifies a k-correction input file to be used if the (**-r 1**) option is also specified. The k-correction file must be in the following format: (a) there are exactly the same number of rows in the file as the number of SNPs on the Affymetrix chip; (b) the first entry in each row is an Affymetrix SNP ID; (c) the values in the row are associated with the SNP ID and are organized into groups of three values; (d) the number of groups exactly corresponds to the number of quartets in the .gpb file for the given SNP; (e) the three values in a group are the AA average, AB average, and BB average in that order; (f) if a value is missing, then the string **NaN** must be inserted in the file in place of the missing value. The method will try to extrapolate values for a missing AA average, AB average, and BB average for a given group of three values in the same quartet. If more than one average is missing for a given group of three in a quartet, then the default formula of PA/(PA+PB) will be used. The order of the SNP IDs (and thus the order of the rows) in the k-correction file must correspond **EXACTLY** to the order of the rows in the **SNPNamesFile**. K-correction files are not currently distributed as part of GenePool although we do intend to add them in a future release.

**-R, --MapFile=FILENAME**

This option specifies the conversion file from SNP code to chromosome, position and rs code. For Affymetrix chips, each line in this file will have the Affymetrix ID, rs code, chromosome and position (tab delimited). For example,

```
SNP_A-1780270    rs987435        7           78244234
SNP_A-1780271    rs345783       15          31183071
SNP_A-1780272    rs955894        1           186539341
```

is three lines in a Affymetrix Map file. Additionally, the lines in the file must be sorted in alphanumeric ordering (shown in the example). For Illumina chips, each line in this file will have the Illumina code, RS code, chromosome and position (tab delimited).

For example,

```
10008           rs758676        7           12685347
10010           rs3916934       13          103143536
10014           rs2711935        4           38985023
```

is three lines in a Illumina Map file. Additionally, the lines in the file must be sorted in numeric ordering (as shown in the example). Note that currently the SNP codes for Illumina contain numbers and the SNP codes are treated as such.

**-M**

**--CheckFormat**

This option causes **gpanalyze** to check the format of all the binary **.gpb** data files for the current experiment.

**-T****--Study**

Display parameters without executing the program.

**KNOWN ISSUES**

Please see the **genepool(1)** manpage.

**AUTHORS**

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

**SEE ALSO**

**genepool(1)**, **gpextract(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](mailto: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/>*