

# WORKSHOP06 – Processing, Querying and Visualizing Gene Expression Data CANB 7640

Aik Choon Tan, Ph.D.

Associate Professor of Bioinformatics

Division of Medical Oncology

Department of Medicine

[aikchoon.tan@ucdenver.edu](mailto:aikchoon.tan@ucdenver.edu)

10/16/2018

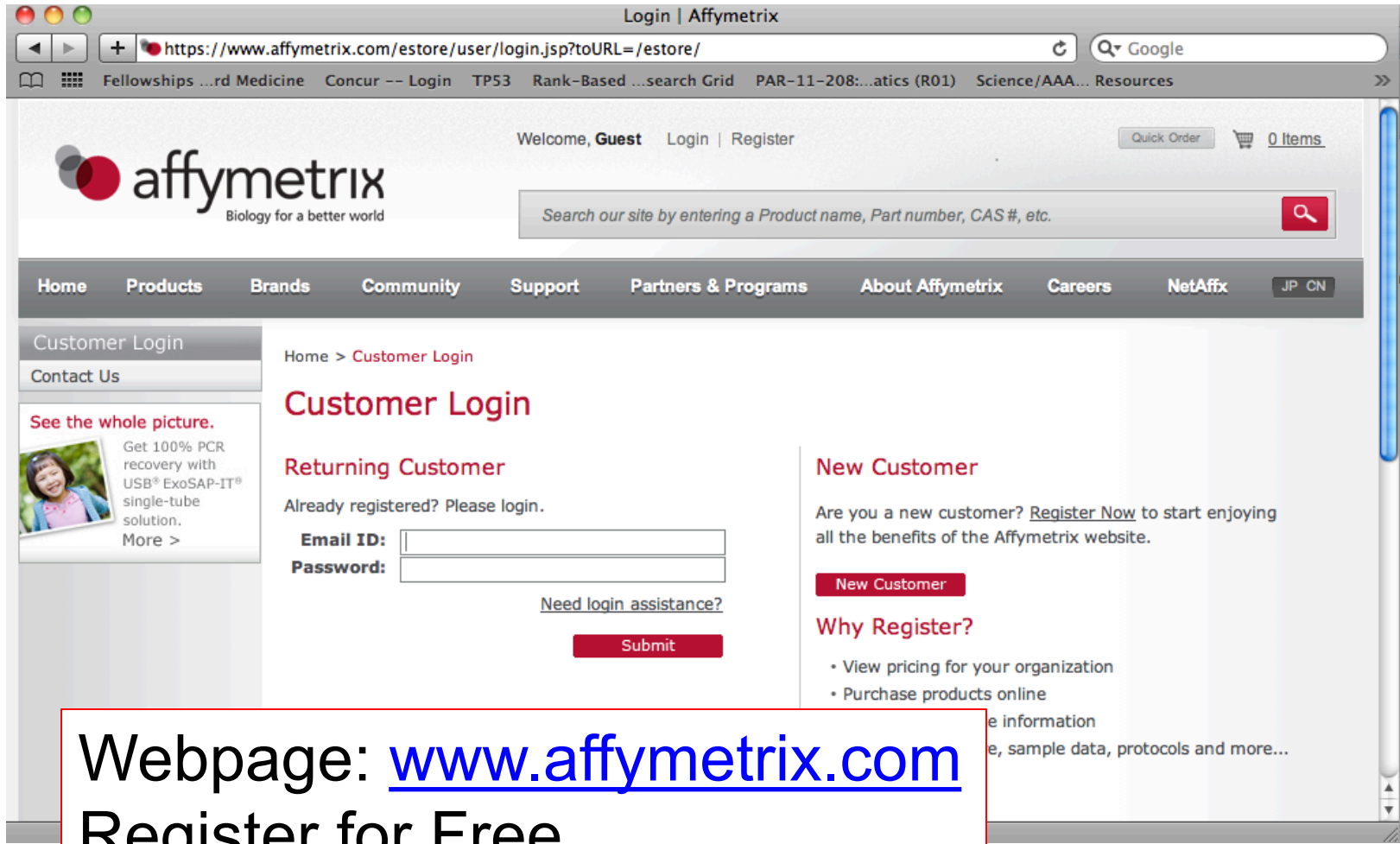
<http://tanlab.ucdenver.edu/labHomePage/teaching/CANB7640/>

# Affymetrix Power Tools

---

- Affymetrix Power Tools (APT) are a set of ***cross-platform command line programs*** that implement ***algorithms for analyzing and working*** with ***Affymetrix GeneChip® arrays***.
- APT is an ***open-source*** project licensed under the GNU General Public License (GPL). (Developers who need a non-GPL license may purchase a commercial license from Affymetrix.)
- APT programs are intended for "***power users***" who prefer programs that can be utilized in scripting environments and are sophisticated enough to handle the complexity of extra features and functionality.
- The vision is that APT provides a platform for developing and deploying new algorithms without waiting for the GUI implementations.

# How to get Affymetrix Power Tools?



The screenshot shows the Affymetrix website's Customer Login page. The browser address bar displays the URL <https://www.affymetrix.com/estore/user/login.jsp?toURL=/estore/>. The page features the Affymetrix logo and a navigation menu with links to Home, Products, Brands, Community, Support, Partners & Programs, About Affymetrix, Careers, and NetAffx. A search bar is located at the top right. The main content area is titled "Customer Login" and includes a "Returning Customer" section with fields for "Email ID:" and "Password:", a "Submit" button, and a link for "Need login assistance?". A "New Customer" section prompts users to "Register Now" and lists benefits such as viewing pricing and purchasing products online. A sidebar on the left contains a "See the whole picture." section with a photo of a woman and text about PCR recovery with USB® ExoSAP-IT®.

Webpage: [www.affymetrix.com](https://www.affymetrix.com)

Register for Free

Login

# Search “APT” in the searchbox

<https://www.thermofisher.com/us/en/home/life-science/microarray-analysis/microarray-analysis-partners-programs/affymetrix-developers-network/affymetrix-power-tools.html>

## Affymetrix Power Tools

• [Affymetrix Developers' Network \(ADN\)](#)

Affymetrix DevNet Tools

### Affymetrix Power Tools

Fusion Software Developer's Kit (SDK)

MAS5 Statistical Software Developer's Kit (SDK)

Affymetrix Power Tools (APT) is a set of cross-platform command line programs that implement algorithms for analyzing and working with GeneChip™ arrays. APT programs are intended for "power users" who prefer programs that can be utilized in scripting environments and are sophisticated enough to handle the complexity of extra features and functionality.

Two of the most popular programs and their features are:

**apt-probeset-summarize: an application for analyzing expression (i.e., U133 and Exon Arrays)**

#### Features include:

- Multiple summarization methods like PLIER, RMA, MAS5, DABG, and IterPLIER
- Sketch quantile normalization (saves memory)
- Lots of parameters available to power users
- Save and reuse feature effects and target normalizations
- Use meta-probeset files to group probe sets into larger probe sets (i.e., combine exons into genes)
- Can run thousands of chips in 1-2 GB of RAM by dividing job into smaller pieces
- Jobs can be split for running on cluster
- Run multiple analysis at once (i.e., Plier and RMA)
- Relatively fast and robust: process 848 Human Exon Arrays with RMA-sketch in about two days on Windows 2 GB RAM machine

# Click on APT Packages

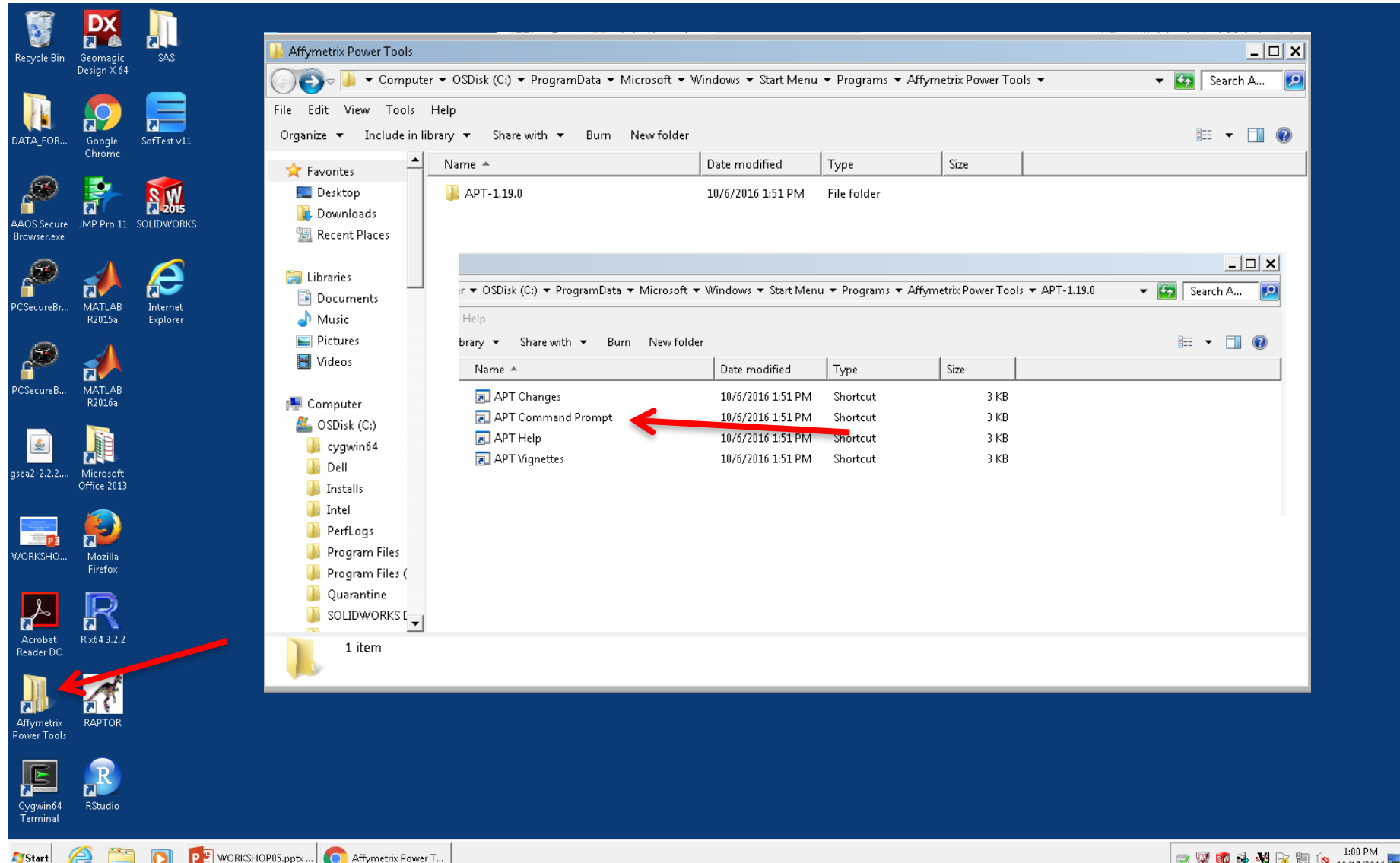
---

## APT packages

APT is available as a Windows installer package, pre-built binaries for Linux and Mac OS-X. APT versions available for download are listed below with the most recent release at the top. Note that multiple versions of APT can be installed simultaneously on the same computer.

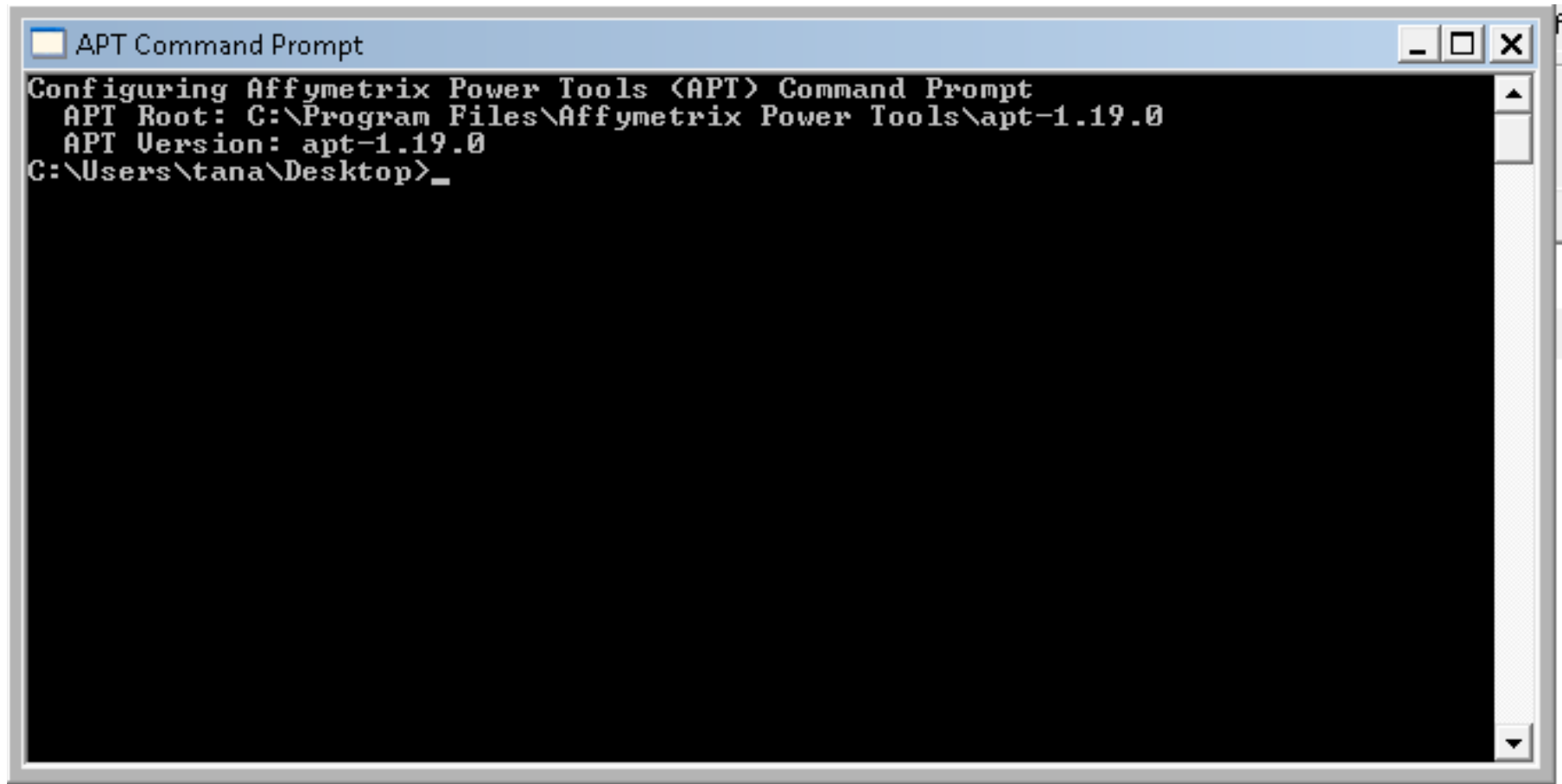
File	Size	md5 Checksum
<a href="#">APT 1.20.5 Windows Installer - win64</a>	94 MB	a1b6adbc8969a8dd99351bc89a5561dc
<a href="#">APT 1.20.5 Apple Yosemite 64 bit x86</a>	119 MB	bbc161972feb0daa8b0d0cb8801a7f06
<a href="#">APT 1.20.5 Linux 64 bit x86 binaries</a>	134 MB	72025032786999e1e7e91283334d41a6

# APT will be installed in C/Program Files



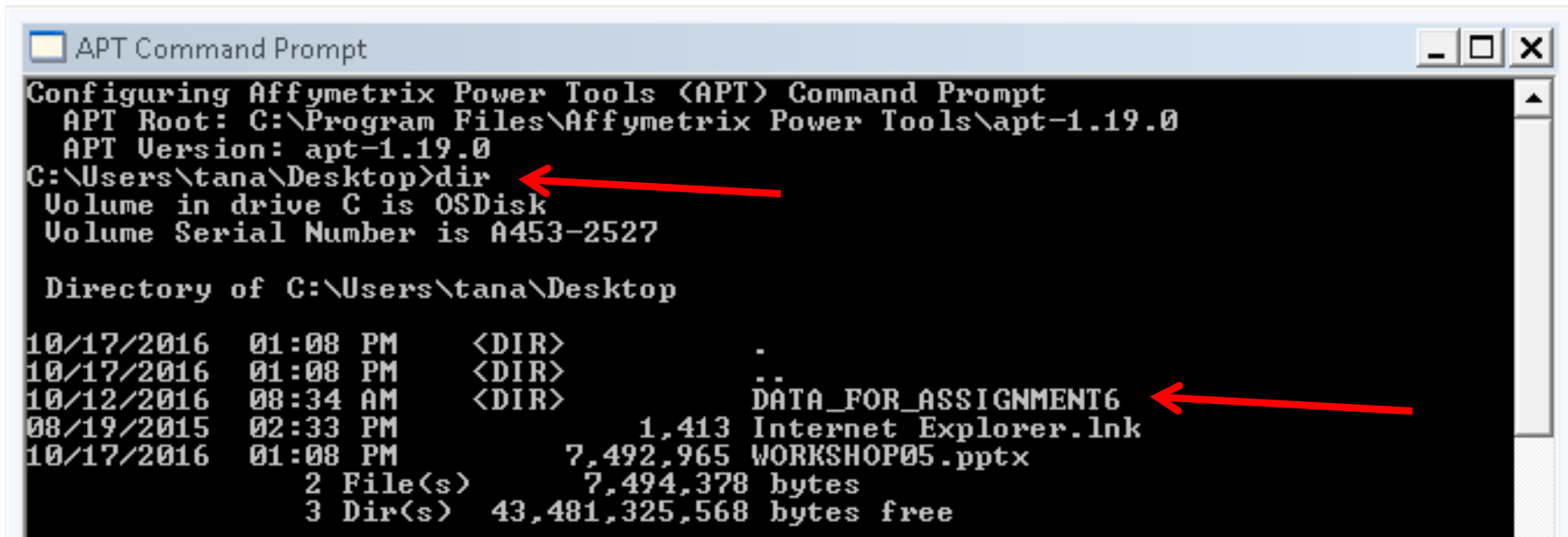
# Go to APT-14.2/Bin/ using Cygwin

---



```
APT Command Prompt
Configuring Affymetrix Power Tools (APT) Command Prompt
APT Root: C:\Program Files\Affymetrix Power Tools\apt-1.19.0
APT Version: apt-1.19.0
C:\Users\tana\Desktop>
```

# List out all Programs in APT-14.2/Bin



```
APT Command Prompt
Configuring Affymetrix Power Tools (APT) Command Prompt
APT Root: C:\Program Files\Affymetrix Power Tools\apt-1.19.0
APT Version: apt-1.19.0
C:\Users\tana\Desktop>dir
Volume in drive C is OSDisk
Volume Serial Number is A453-2527

Directory of C:\Users\tana\Desktop

10/17/2016  01:08 PM    <DIR>          .
10/17/2016  01:08 PM    <DIR>          ..
10/12/2016  08:34 AM    <DIR>          DATA_FOR_ASSIGNMENT6
08/19/2015  02:33 PM             1,413 Internet Explorer.lnk
10/17/2016  01:08 PM      7,492,965 WORKSHOP05.pptx
           2 File(s)      7,494,378 bytes
           3 Dir(s)  43,481,325,568 bytes free
```

The screenshot shows a command prompt window titled "APT Command Prompt". It displays the configuration for Affymetrix Power Tools (APT) and the results of a directory listing command. Two red arrows are present: one pointing to the command "C:\Users\tana\Desktop>dir" and another pointing to the file "DATA\_FOR\_ASSIGNMENT6" in the directory listing.



# List out all Programs in APT-14.2/Bin

```
C:\Users\tana\Desktop>cd DATA_FOR_ASSIGNMENT6
C:\Users\tana\Desktop\DATA_FOR_ASSIGNMENT6>dir
Volume in drive C is OSDisk
Volume Serial Number is A453-2527

Directory of C:\Users\tana\Desktop\DATA_FOR_ASSIGNMENT6

10/12/2016  08:34 AM    <DIR>          .
10/12/2016  08:34 AM    <DIR>          ..
10/08/2014  11:06 AM                325  _KRAS_DEP_LUNG.png
10/08/2014  10:47 AM                229  _NCIH2009.CEL.gz
10/08/2014  10:47 AM                229  _NCIH23.CEL.gz
10/08/2014  10:47 AM                229  _NCIH358.CEL.gz
10/08/2014  10:47 AM                229  _NCIH441.CEL.gz
10/08/2014  10:47 AM                229  _NCIH460.CEL.gz
10/08/2014  10:47 AM                229  _NCIH727.CEL.gz
10/08/2014  11:08 AM                325  _SAMPL0T_FDR5.png
10/08/2014  10:47 AM                229  _SKLU1.CEL.gz
10/08/2014  10:47 AM                229  _SW1573.CEL.gz
10/27/2005  01:19 PM                229  _U133_X3P.cdf.gz
10/08/2014  10:55 AM                 49  KRASDEP.txt
10/08/2014  10:55 AM                 44  KRASIND.txt
10/08/2014  11:06 AM            55,604  KRAS_DEP_LUNG.png
10/08/2014  10:47 AM        13,551,663  NCIH2009.CEL
10/08/2014  10:47 AM        13,553,150  NCIH23.CEL
10/08/2014  10:47 AM        13,553,051  NCIH358.CEL
10/08/2014  10:47 AM        13,552,955  NCIH441.CEL
10/08/2014  10:47 AM        13,553,091  NCIH460.CEL
10/08/2014  10:47 AM        13,552,430  NCIH727.CEL
10/08/2014  11:08 AM            47,548  SAMPL0T_FDR5.png
10/08/2014  10:47 AM        13,552,471  SKLU1.CEL
10/08/2014  10:47 AM        13,551,248  SW1573.CEL
10/12/2016  08:34 AM    <DIR>          TEST
10/27/2005  01:19 PM        123,036,175  U133_X3P.cdf
                24 File(s)      231,562,190 bytes
                3 Dir(s)      43,481,325,568 bytes free

C:\Users\tana\Desktop\DATA_FOR_ASSIGNMENT6>
```

# apt-probeset-summarize

```
APT Command Prompt
C:\Users\tana\Desktop\DATA_FOR_ASSIGNMENT6>apt-probeset-summarize ! more
apt-probeset-summarize - A program for summarizing expression probe
data from cel files. Can use either a cdf file or pgf/clf files for defining
probesets. Use the '--explain' flag for further documentation on a
particular data transformation or summary value.

usage:
  apt-probeset-summarize -a rma-sketch -a plier-mm-sketch \
    -p chip.pgf -c chip.clf -o output-dir *.cel

options:
Common Options (not used by all programs)
-h, --help                Display program options and extra
                           documentation about possible analyses. See
                           --explain for information about a specific
                           operation. [default 'false']
-v, --verbose             How verbose to be with status messages 0 -
                           quiet, 1 - usual messages, 2 - more
                           messages. [default '1']
--console-off             Turn off the default messages to the
                           console but not logging or sockets.
                           [default 'false']
--use-socket              Host and port to print messages over in
                           localhost:port format [default '']
--version                 Display version information. [default
                           'false']
-f, --force               Disable various checks including chip
                           types. Consider using --chip-type option
                           rather than --force. [default 'false']
--throw-exception         Throw an exception rather than calling
                           exit() on error. Useful for debugging. This
                           option is intended for command line use
                           only. If you are wrapping an Engine and
                           want exceptions thrown, then you should
                           call Err::setThrowStatus(true) to ensure
                           that all Err::errAbort() calls result in an
                           exception. [default 'false']
--analysis-files-path     Search path for analysis library files.
                           Will override APPX_ANALYSIS_FILES_PATH
                           environment variable. [default '']
--xml-file                Input parameters in XML format (Will
                           override command line settings). [default
                           '']
--temp-dir                Directory for temporary files when working
                           off disk. Using network mounted drives is
                           not advised. When not set, the output
                           folder will be used. The default is
                           typically the output directory or the
                           current working directory. [default '']
-o, --out-dir             Directory for output files. Defaults to
                           current working directory. [default '.']
--log-file                The name of the log file. Generally
                           defaults to the program name in the out-dir
                           folder. [default '']
```

# User Manual (Help Page)

<http://media.affymetrix.com/support/developer/powertools/changelog/index.html>

## Affymetrix Power Tools (APT) -- Release 1.20.5

The Affymetrix Power Tools (APT) is a collection of command line programs for analyzing and working with Affymetrix microarray data. These programs are generally focused on CEL file level analysis. APT also refers to the underlying C++ source code. Binaries and source code are available from the main APT website, <http://www.affymetrix.com/support/developer/powertools/index.affx>.

### Power User Manuals

Power user manuals are available for specific command line applications:

- Main Applications:
  - `apt-probeset-summarize`: A program for analyzing expression arrays including 3' IVT and exon arrays. Supports background correction (MAS5,RMA), normalization (linear scaling, quantile, sketch), and summarization (PLIER, RMA, MAS5) methods.
  - `apt-probeset-genotype`: A program for analyzing mapping arrays. Supports BRLMM-P, Birdseed, and BRLMM methods for genotype calling.
  - `apt-genotype-axiom`: A program for performing recommended genotype calling analysis on Axiom arrays.
  - `apt-genotype-eureka`: A program for performing recommended genotype calling analysis on Eureka binning files.
  - `nibls`: A program for converting sequence data from the Eureka platform into binning files to be used in genotyping or visualization.
  - `apt-geno-qc`: A program for doing single chip QC of WGS genotyping arrays.
  - `apt-copynumber-axiom-ssa`: A program that reports copy number states in pre-defined regions for Axiom library packages that support it.
  - `apt-copynumber-workflow`: A program to run the copy number analysis workflow on SNP6 arrays.
  - `apt-copynumber-cyto-ssa`: A program to run single-sample copynumber and LOH analysis on CytoScan family of arrays.
  - `apt-copynumber-cyto-ref`: A program to generate reference model files for copynumber and LOH analysis of CytoScan family of arrays.
  - `apt-copynumber-wave`: A program to add additional waves to copynumber reference file. Most users should use default wave corrections provided by Affymetrix.
  - `apt-canary`: A program to compute copy number variation calls given a known set of CNV regions.
  - `apt-dmet-genotype`: A program to compute genotypes and copy number variation from DMET Plus CEL files. DMET CHP files are generated.
  - `apt-dmet-translation`: A program to compute star allele translation reports from DMET Plus CHP files.
  - `apt-copynumber-onco-ref`: A program to generate copynumber reference model files for OncoScan arrays.
  - `apt-copynumber-onco-ssa`: A program to perform copynumber analysis on OncoScan arrays and matched normal/tumor pairs.
  - `apt-copynumber-onco-som-ref`: A program to generate somatic mutation reference model files for OncoScan arrays.
  - `apt-copynumber-onco-som-ssa`: A program to implement the somatic mutation analysis pipeline for OncoScan arrays.
  - `ps-metrics`: A program to generate various QC metrics for SNPs for Axiom arrays.
  - `ps-classification`: A program which reads a metrics table generated by `ps-metrics` and classifies SNPs based on a number of customizable criteria.
  - `otv-caller`: A program for identifying *off-target variants*.
- Utility Programs:
  - `apt-cel-transformer`: A program for applying arbitrary chipstream methods (ie quantile normalization, RMA background correction) to a set of cel files, resulting in a new set of cel files.
  - `apt-cel-extract`: A program for extracting feature level intensities from CEL files.
  - `apt-cel-convert`: A program for converting CEL files to different formats.
  - `apt-chp-to-txt`: A program to dump AGCC and XDA chp files as text.
  - `apt-file5-util`: A program to convert between a5 and text formats.
  - `apt-engine-wrapper`: A program to directly call analysis engines. The main use is to run it with the help option in order to find out what options various sub-engines will except.
  - `apt-summary-genotype`: A program to run BRLMM-P family of algorithms on allele summaries.
  - `apt-annotation-converter`: A program to create custom SQLite format annotation files from csv files.
  - `apt2-dset-util`: A program for converting between the file formats supported by the APT2 framework, including OSCHP and text files.
  - `apt-param-convert`: A program for converting XML parameter files used in legacy applications to those used in newer APT2 applications (e.g. from `apt-probeset-genotype` to `apt-genotype-axiom`).
  - `apt-package-util`: A program for creating .suitecase files used by Axiom Analysis Suite from TXT or CHP output produced by APT or GTC.
  - `apt-format-result`: A program for creating VCF or PLINK file formats from Axiom Analysis Suite .suitecase files or TXT files.
  - `apt-suitecase-extract`: A program for converting .suitecase files generated by Axiom Analysis Suite version 1 to folder format required by version 1.1.
- Legacy Programs (likely to be removed in later APT releases):
  - `apt-midas`: A program to compute MiDAS (alternative splice detection) scores from exon array results.

# Quick Start

## Quick Start

Most users will just want to generate summaries using RMA and/or Plier for each probeset on the microarray. We provide both 'rma' and 'rma-sketch' where 'rma-sketch' will closely approximate a full quantile normalization using a much smaller amount of memory.

On unix systems a command to do both rma-sketch and plier-sketch analysis at the same time with the default parameters looks like:

```
apt-probeset-summarize -a rma-sketch -a plier-mm-sketch -d chip.cdf -o output-dir *.cel
```

when using a CDF file or alternatively a PGF and CLF files can be specified:

```
apt-probeset-summarize -a rma-sketch -a plier-mm-sketch -p chip.pgf -c chip.clf -o output-dir *.cel
```

As the windows command prompt does not natively support wild card expansion the preferred method is to supply a text file list via the --cel-files option (see below for details of file format). A windows a command using the default parameters looks like:

```
apt-probeset-summarize -a rma-sketch -a plier-mm-sketch -d chip.cdf -o output-dir --cel-files cel_list.txt
```

Where -a specifies an analysis to do and -o specifies a directory to put the output files in. You can specify the probesets on a chip with either a CDF file via a -d or using a PGF/CLF file pair via the -p and -c flags.

If the microarray does not have mismatch probes you can specify use a surrogate mismatch based on probes with similar GC content by using the plier-gcbg analysis and specifying the background probes using the --bgp-file flag.

**WARNING:** apt-probeset-summarize will overwrite any existing output files it finds. If you wish to keep existing results make sure to specify a different output directory name. It is also important to note that consistent with the Bioconductor implementation all RMA output has been log2 transformed.

# Usage

```
hslib@HSL-TL-04 /cygdrive/c/Program Files/Affymetrix Power Tools/APT-14.2/Bin
$ ./apt-probeset-summarize.exe -a ANALYSIS_TYPE -d CDF -o OUTPUT *.CEL
```

Call the  
program

Type of  
analysis

Chip annotations:  
-d chip.CDF  
or  
-p chip.pgf  
-c chip.clf

Output  
Folder

Input  
CEL  
files

```
usage:
apt-probeset-summarize -a rma-sketch -p plier-mm-sketch \
-p chip.pgf -c chip.clf -o output-dir *.cel
```

# Input Options (Required)

---

## **CEL FILES (INPUTS)**

`--cel-files` Text file specifying cel files to process,  
one per line with the first line being  
'cel\_files'. [default '']

OR

`*.CEL` specify CEL file names (or `*` represents all CEL files in  
current working directory)

## **CHIP TYPES:**

`-d, --cdf-file` File defining probe sets. Use either  
`--cdf-file`, `--spf-file`, or `--pgf-file` and  
`--clf-file`. Automatically sets `--names`.  
[default '']

OR

`-p, --pgf-file` File defining probe sets. [default '']  
`-c, --clf-file` File defining x,y <-> probe id conversion.  
Required when using PGF file. [default '']

# Analysis Options (Required)

---

`-a, --analysis String` representing analysis pathway desired.

For example:

```
'plier-gcbg-sketch',  
'plier-gcbg',  
'plier-mm-sketch',  
'plier-mm',  
'rma-sketch',  
'rma'
```

Multiple analysis allowed at same time. When using quantile normalization, you may need to use the sketch option to avoid running out of memory. [default '']

# Output Options (Required)

---

`-o, --out-dir` Directory for output files. Defaults to  
current working directory[default '.']



# Where to find Chip Annotations?

## From Affymetrix Website

<https://www.thermofisher.com/search/browse/results?customGroup=Human+Expression+Profiling+Arrays+%26+Assays&persona=Catalog&resultPage=1&resultsPerPage=60>

[Home](#) › [Shop All Products](#) › [Human Expression Profiling Arrays & Assays](#) › [DNA & RNA Microarray Analysis](#)

### Human Expression Profiling Arrays & Assays

#### Product Category

☒ [DNA & RNA Microarray Analysis](#)

#### Product Category

☐ [Transcriptome Profiling Arrays & Assays \[27\]](#)

#### Application

☐ [DNA & RNA Microarray Analysis \[27\]](#)

#### Brand

☐ [Applied Biosystems™ \[27\]](#)

[Catalog \[27\]](#)

[Learn More \[0\]](#)

[Documents & Support \[0\]](#)



#### [GeneChip™ Human Gene 1.1 ST Array Plate](#) (Applied Biosystems™)

The Human Gene 1.1 ST 24-Array Plate and Trays provide the most accurate, sensitive, and comprehensive measurement of protein coding and long intergenic non-coding RNA transcripts.



#### [PrimeView™ Human Genome U219 Array Strip](#) (Applied Biosystems™)

##### Proven performance from the industry standard

The Affymetrix™ Human Genome U219 Array Strip enables expression profiling of four samples at a time using probe sets with an emphasis on established, well-annotated content. Sequences used in



#### [GeneChip™ HT HG-U133+PM Array Plate](#) (Applied Biosystems™)

##### Description: Proven performance from the industry standard

The GeneChip™ HT HG-U133+ PM 16-Array Plate enables high-throughput expression profiling of multiple samples at a time using the same content as the industry-standard GeneChip Human

# Annotations for HG-U133 Plus 2.0 Array

<https://www.thermofisher.com/order/catalog/product/900466?SID=srch-srp-900466>



## GeneChip™ Human Genome U133A 2.0 Array (Applied Biosystems™)

The Human Genome U133A 2.0 Array is a single array representing 14,500 well-characterized human genes that can be used to explore human biology and disease processes. Newer design and reduced feature size mean that you can use smaller sample volumes than the previous HG-U133 Array without



## Clariom™ D Pico Assay, human (Applied Biosystems™)

Accelerate your biomarker discovery from deep within the transcriptome with Clariom D Pico Assays for human, the next generation of transcriptome-level expression profiling tools. Human Clariom D Pico Assays provide a highly detailed view of the transcriptome and offer the fastest path to



## GeneChip™ Human Transcriptome Assay 2.0 (Applied Biosystems™)

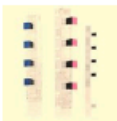
Designed to empower next-generation expression profiling studies, the GeneChip™ Human Transcriptome Assay 2.0 provides the ability to go beyond gene-level expression profiling by providing the coverage and accuracy required to accurately detect all known transcript isoforms produced by a



## GeneChip™ Human Genome U133 Plus 2.0 Array (Applied Biosystems™)

### GeneChip™ Human Genome U133 Plus 2.0 Array

**Benefits of the first and most comprehensive whole human genome expression array**



## GeneChip™ Human Gene 2.1 ST Array Strip (Applied Biosystems™)




### Comprehensive design

Keeping pace with the research community's understanding of the transcriptome, we have designed whole-transcript arrays that include probes to measure both messenger (mRNA) and long intergenic













# Annotations for HG-U133 Plus 2.0 Array

## Documents

### Manuals & protocols

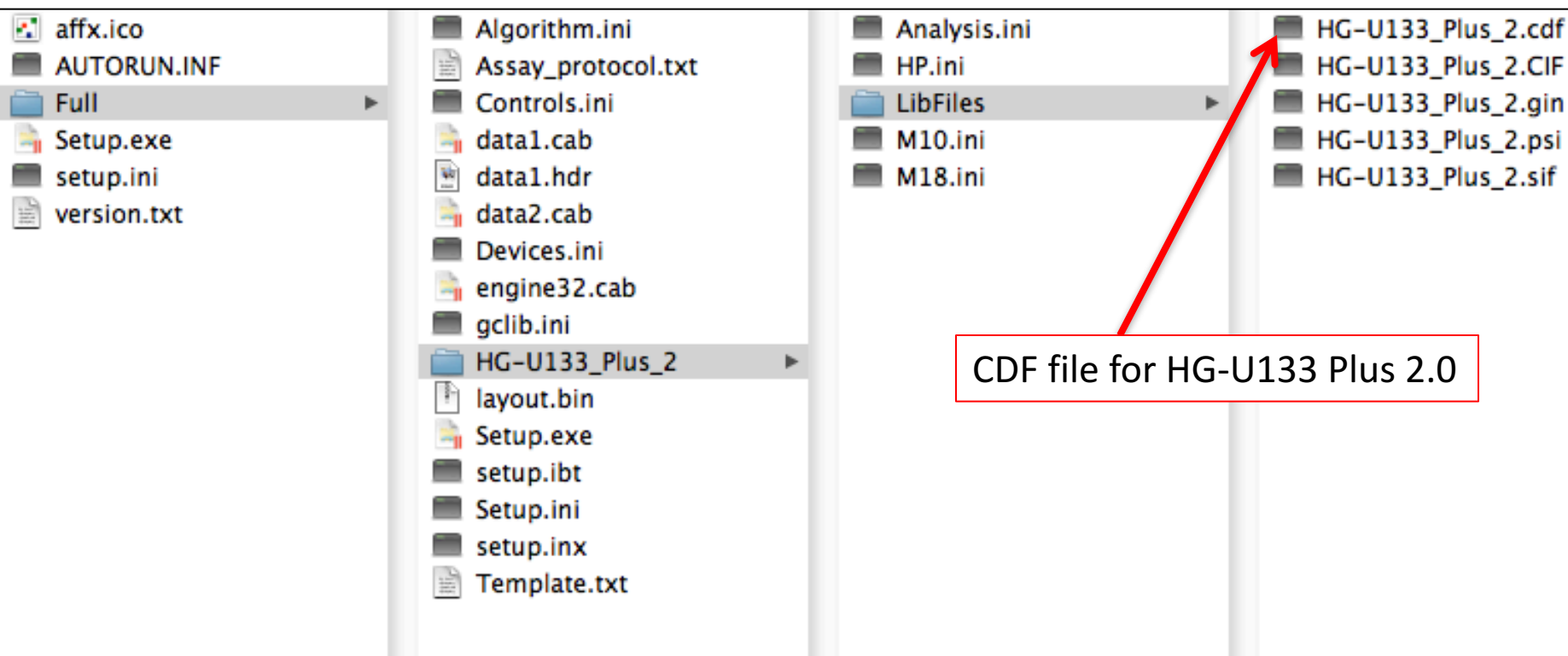
-  [Manuals: 3' IVT PLUS Reagent Kit User Manual](#)
-  [Manuals: Expression Analysis Technical Manual, with Specific Protocols for Use with the Hybridization, Wash, and Stain Kit](#)
-  [Package Inserts: Human Genome U133 Plus 2.0 Array](#)

### Support files

-  [Array Comparisons: Bovine to Human Genome U133 Plus 2.0 Array: Best Match](#)  
Version: Jan. 2017, File size: 99.6 kB
-  [Array Comparisons: Bovine to Human Genome U133 Plus 2.0 Array: Complex](#)  
Version: Jan. 2017, File size: 507.8 kB
-  [Library Files: Human Genome U133 Plus 2.0 Array](#)  
Version: Jan. 2017, File size: 30.2 MB
-  [Mask Files: Human Genome U133 Plus 2.0 Array Normalization Controls](#)  
Version: Jan. 2017, File size: 402 B
-  [NetAffx Alignment Files: HG-U133\\_Plus\\_2 Alignments to Genome, PSL](#)  
Version: 9/28/2006, File size: 7.4 MB
-  [NetAffx Alignment Files: HG-U133\\_Plus\\_2 BED File](#)  
Version: 8/22/2016, File size: 9.9 MB
-  [Sequence Files: HG-U133\\_Plus\\_2 Consensus Sequences, FASTA](#)  
Version: 8/20/2008, File size: 22.8 MB
-  [Sequence Files: HG-U133\\_Plus\\_2 Control Sequences, FASTA](#)  
Version: 10/17/2003, File size: 20 kB
-  [Sequence Files: HG-U133\\_Plus\\_2 Exemplar Sequences, FASTA](#)  
Version: 10/17/2003, File size: 12.2 MB
-  [Sequence Files: HG-U133\\_Plus\\_2 Probe Sequences, FASTA](#)  
Version: 8/20/2008, File size: 11 MB
-  [Sequence Files: HG-U133\\_Plus\\_2 Probe Sequences, Tabular](#)  
Version: 8/20/2008, File size: 9.9 MB
-  [Sequence Files: HG-U133\\_Plus\\_2 Target Sequences, FASTA](#)  
Version: 8/20/2008, File size: 11 MB

Library Files contain:  
CDF or PLG and CLF

# Library Files Folder



# Querying NCBI GEO

## Gene Expression Omnibus

### Query by GSE ID



GEO is a public functional genomics data repository supporting Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.



### Getting Started

[Overview](#)[FAQ](#)[About GEO DataSets](#)[About GEO Profiles](#)[About GEO2R Analysis](#)[How to Construct a Query](#)[How to Download Data](#)

### Tools

[Search for Studies at GEO DataSets](#)[Search for Gene Expression at GEO Profiles](#)[Search GEO Documentation](#)[Analyze a Study with GEO2R](#)[GEO BLAST](#)[Programmatic Access](#)[FTP Site](#)

### Browse Content

[Repository Browser](#)

DataSets: 3413

Series:  41905

Platforms: 12087

Samples: 1007387

# Types of Data Available for Download

Platforms (1) [GPL570](#) [HG-U133\_Plus\_2] Affymetrix Human Genome U133 Plus 2.0 Array

Samples (53) [GSM452148](#) C1  
[More...](#) [GSM452149](#) C2  
[GSM452150](#) C3

**Relations**  
BioProject [PRJNA119367](#)

**Analyze with GEO2R**

**Download family**

	Format
<a href="#">SOFT formatted family file(s)</a>	SOFT ?
<a href="#">MINiML formatted family file(s)</a>	MINiML ?
<a href="#">Series Matrix File(s)</a>	TXT ?

Supplementary file	Size	Download	File type/resource
<a href="#">GSE18088_RAW.tar</a>	243.6 Mb	<a href="#">(http)</a> <a href="#">(custom)</a>	TAR (of CEL)

*Raw data provided as supplementary file*  
*Processed data included within Sample table*

**SOFT data**

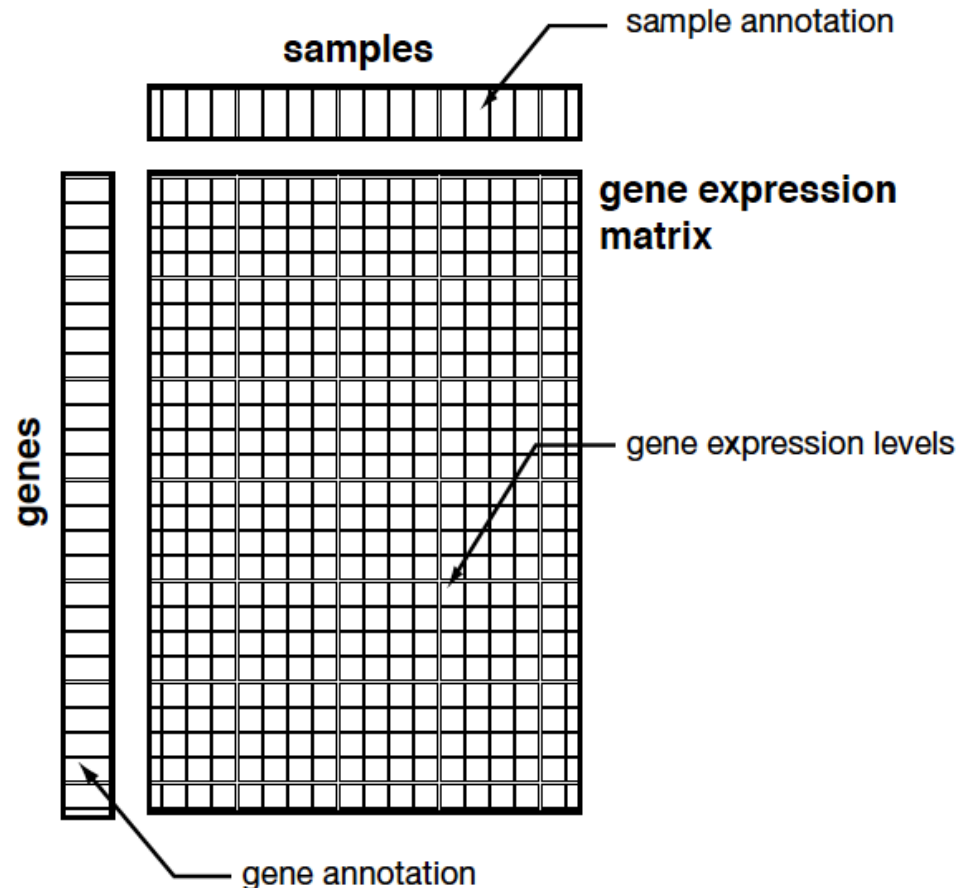
**MINiML data**

**SERIES data**

**RAW data**

# Series Matrix Format

SeriesMatrix/: This directory contains tab-delimited value-matrices generated from the VALUE column of the Sample tables of each Series entry. Files also include Series and Sample metadata and are ideal for opening in spreadsheet applications such as MicrosoftExcel. Most users find SeriesMatrix files the most convenient format for handling data that have not been assembled into a DataSet



# Download Series Matrix File(s)

1

## Analyze with GEO2R

### Download family

SOFT formatted family file(s)  
MINiML formatted family file(s)  
Series Matrix File(s)

### Format

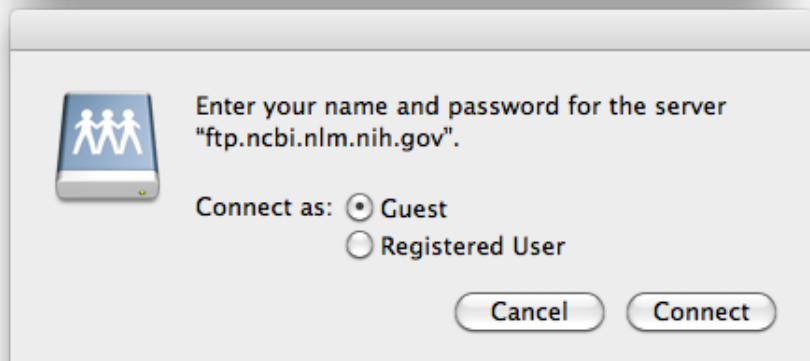
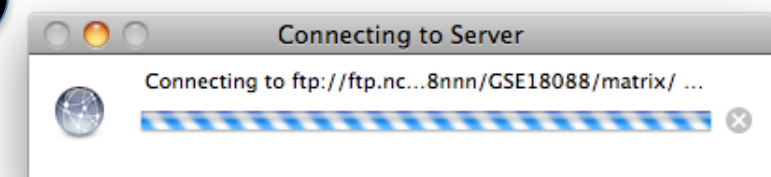
SOFT [?](#)  
MINiML [?](#)  
TXT [?](#)

Supplementary file	Size	Download	File type/resource
GSE18088_RAW.tar	243.6 Mb	( <a href="#">http</a> )( <a href="#">custom</a> )	TAR (of CEL)

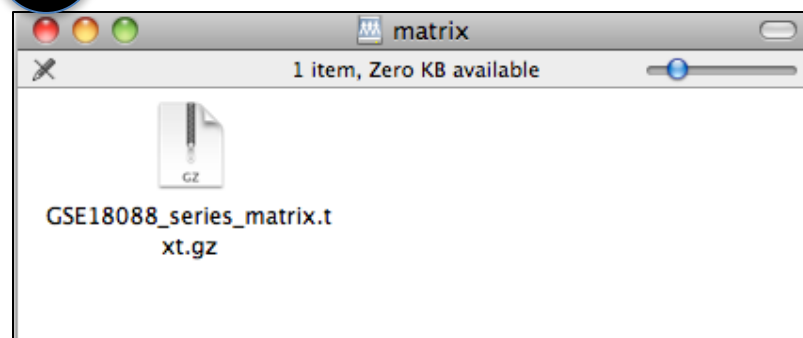
*Raw data provided as supplementary file*

*Processed data included within Sample table*

2



3





# Sample Descriptions

39	!Sample_title	C1	C2	C3
40	!Sample_geo_accession	GSM452148	GSM452149	GSM452150
41	!Sample_status	Public on Apr 10 2011	Public on Apr 10 2011	Public on Apr 10 2011
42	!Sample_submission_date	Sep 11 2009	Sep 11 2009	Sep 11 2009
43	!Sample_last_update_date	Apr 10 2011	Apr 10 2011	Apr 10 2011
44	!Sample_type	RNA	RNA	RNA
45	!Sample_channel_count	1	1	1
46	!Sample_source_name_ch1	colon	colon	colon
47	!Sample_organism_ch1	Homo sapiens	Homo sapiens	Homo sapiens
48	!Sample_characteristics_ch1	localization: proximal	localization: distal	localization: distal
49	!Sample_characteristics_ch1	gender: male	gender: female	gender: female
50	!Sample_characteristics_ch1	relapse: no	relapse: no	relapse: no
51	!Sample_characteristics_ch1	microsatellite status: MSS	microsatellite status: MSS	microsatellite status: MSS
52	!Sample_characteristics_ch1	age at diagnosis, years: 65	age at diagnosis, years: 65	age at diagnosis, years: 65
53	!Sample_characteristics_ch1	grading: G2	grading: G3	grading: G3
54	!Sample_characteristics_ch1	pt: 3	pt: 3	pt: 3
55	!Sample_molecule_ch1	total RNA	total RNA	total RNA
56	!Sample_extract_protocol_ch1	For microarray analyses,	For microarray analyses, snap fr	For microarray analyse
57	!Sample_label_ch1	biotin	biotin	biotin
58	!Sample_label_protocol_ch1	Biotinylated cRNA were	Biotinylated cRNA were prepare	Biotinylated cRNA wer
59	!Sample_taxid_ch1	9606	9606	9606
60	!Sample_hyb_protocol	Following fragmentation	Following fragmentation, 10 ug	Following fragmentati
61	!Sample_scan_protocol	GeneChips were scanned	GeneChips were scanned using 1	GeneChips were scann
62	!Sample_description	none	none	none
63	!Sample_data_processing	Data were normalized w	Data were normalized with VSN	Data were normalized
64	!Sample_platform_id	GPL570	GPL570	GPL570
65	!Sample_contact_name	Dido,,Lenze	Dido,,Lenze	Dido,,Lenze
66	!Sample_contact_email	dido.lenze@charite.de	dido.lenze@charite.de	dido.lenze@charite.de
67	!Sample_contact_department	Pathologie, Campus Ben	Pathologie, Campus Benjamin F	Pathologie, Campus Be
68	!Sample_contact_institute	CharitÄ©-UniversitÄtss	CharitÄ©-UniversitÄttsmedizin	CharitÄ©-UniversitÄt
69	!Sample_contact_address	Hindenburgdamm 30	Hindenburgdamm 30	Hindenburgdamm 30
70	!Sample_contact_city	Berlin	Berlin	Berlin
71	!Sample_contact_zip/postal_code	12200	12200	12200
72	!Sample_contact_country	Germany	Germany	Germany
73	!Sample_supplementary_file	ftp://ftp.ncbi.nlm.nih.gov	ftp://ftp.ncbi.nlm.nih.gov/pub/g	ftp://ftp.ncbi.nlm.nih.g
74	!Sample_data_row_count	54675	54675	54675

Class Label for  
comparison in  
SAM



# Using matrix2png

<http://www.chibi.ubc.ca/matrix2png/>

## matrix2png interface

Fill in the following form to generate images from your own data files. Information on the required data file format is available [here](#), including a sample data file. Detailed documentation is available [here](#). The matrix2png home page is [here](#). While that documentation is for the command line interface, most of it applies here.

### Quick tips to avoid common file format problems:

- There is a [sample file](#) available
- The input files are *tab* delimited. Comma or space-delimited files will not work.
- Missing values are OK, but rows that have missing values at the end must contain extra "tabs": "ragged" rows are not permitted.
- Notice the 'upper left hand corner' string in the [example file](#) - all columns including the example names have a heading. It does not matter what you put in the corner. The parser uses the header to figure out how many features you have, so if you skip the corner string it will appear that you have extra data, resulting in an error message.
- You can only have one column of descriptors, all other data must be your numeric feature data. In other words, don't include extra columns in your file that are not part of the data or the example labels. Extra columns will either result in an error (most likely) or invalid results (if your extra columns look like data). If you have multiple columns of non-numeric data in your file, you should combine them into one column. Use spaces or other symbols to separate your "text fields".

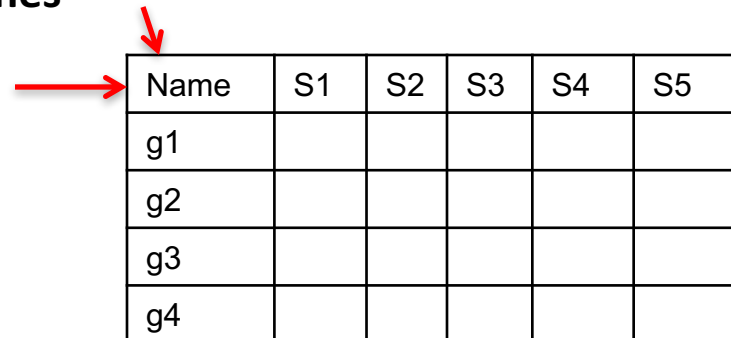
Use test file: ☐

Enter the path to your data file here:  no file selected

First Row :  
Genes

INPUT FORMAT

First Row :  
Samples



Name	S1	S2	S3	S4	S5
g1					
g2					
g3					
g4					

# Options in matrix2png

**Set any of the options below.**

<http://www.chibi.ubc.ca/matrix2png/>

The most commonly changed options are shown in **color**. For some options, default values are already set.

## Set the options for the basic appearance of the image:

Size in pixels for each element: width  height  (Will be adjusted to at least 8 if you want row or column labels)

☐ Normalize the rows of the matrix to have mean zero, variance one

Range of values to display as different colors: minimum  maximum

Trim values by percent:  (To limit effect of extreme outlier values; not used if you set range above)

☐ Use ellipses instead of rectangles

## Set color options:

Background color

Missing values are shown as:

Adjust contrast: by  (Ignored if you set range above)

## Decide how you want colors assigned to values:

<b>Set the color bounds by hand:</b> <input checked="" type="radio"/>	<b>OR use a preset map:</b> <input type="radio"/>	<b>OR, Use a discrete mapping of values to colors:</b> <input type="radio"/>
Min color <input type="text" value="green"/>	Map: <input type="text" value="1 (black body)"/>	Optional discrete map file: <input type="button" value="Choose File"/> no file selected
Middle color <input type="text" value="auto (blend of min and max colors)"/>	<input type="checkbox"/> Reverse the map	
Max color <input type="text" value="red"/>	Map choices are shown <a href="#">here</a>	

Number of colors to use:  Default: 64. (Ignored if using discrete mapping)

# Options in matrix2png

<http://www.chibi.ubc.ca/matrix2png/>

---

## Add widgets:

- ☐ Draw a scale bar
- ☐ Display the row labels
- ☐ Display the row labels on the left side of the image (default=right side)
- ☐ Display the column labels
- ☐ Display the column labels on the bottom of the image (default=top)
- ☐ Draw dividers between cells

## Limit the data used for making the picture:

By default, all the data is used and all values are used to calculate the data range.

Start from row  Number of rows to process:

Start from column  Number of columns to process:

# Assignment #6

---

1. Download GSE18088 Raw data from NCBI GEO.
2. Extract the data using APT (Get CDF file from Affymetrix website).
3. Perform Significance Analysis of Microarray on the data to find out the genes that were differentially expressed between “Disease Free” (DF) group and “Relapse” (R) group at FDR 5%.
4. Plot the differentially expressed genes as heat map using matrix2png.

Send me the list of differentially expressed genes and heat map.



# Read and Locate Data from NCBI GEO

Int J Colorectal Dis (2011) 26:847–858  
DOI 10.1007/s00384-011-1176-x

## ORIGINAL ARTICLE

### Molecular profiles and clinical outcome of stage UICC II colon cancer patients

Jörn Gröne · Dido Lenz · Vindi Jurinovic · Manuela Hummel · Henrik Seidel · Gabriele Leder · Georg Beckmann · Anette Sommer · Robert Grützmann · Christian Pilarsky · Ulrich Mansmann · Heinz-Johannes Buhr · Harald Stein · Michael Hummel

Accepted: 3 March 2011 / Published online: 5 April 2011  
© Springer-Verlag 2011

#### Abstract

**Purpose** Published multigene classifiers suggesting outcome prediction for patients with stage UICC II colon cancer have not been translated into a clinical application so far. Therefore, we aimed at validating own and published gene expression signatures employing methods

which enable their reconstruction in routine diagnostic specimens.

**Methods** Immunohistochemistry was applied to 68 stage UICC II colon cancers to determine the protein expression of previously published prognostic classifier genes (CDH17, LAT, CA2, EMR3, and TNFRSF11A). RNA from macrodissected tumor samples from 53 of these 68 patients was profiled on Affymetrix GeneChips (HG-U133 Plus 2.0). Prognostic signatures were generated by “nearest shrunken centroids” with cross-validation. Previously published gene signatures were applied to our data set using “global tests” and leave-one-out cross-validation.

**Results** Correlation of protein expression with clinical outcome failed to separate patients with disease-free follow-up (group DF) and relapse (group R). Although gene expression profiling allowed the identification of differentially expressed genes (“DF” vs. “R”), a stable classification/prognosis signature was not discernable. Furthermore, the application of previously published gene signatures to our data was unable to predict clinical outcome (prediction rate 75.5% and 64.2%; n.s.). T-stage was the only independent prognostic factor for relapse with established clinical and pathological parameters including microsatellite status (multivariate analysis).

**Conclusions** Our protein and gene expression analyses do not support application of molecular classifiers for prediction of clinical outcome in current routine diagnostic as a basis for patient-orientated therapy in stage UICC II colon cancer. Further studies are needed to develop prognosis signatures applicable in patient care.

**Keywords** Colon cancer · Immunohistochemistry · Gene expression signature · Prognosis

parameters (age, gender, tumor localization, grading, T-stage, microsatellite status), available scores were then tested in multivariate Cox regression analysis. Correlation of expression of selected proteins (CDH17 and EMR3, one probe set each; TNFRSF11A and LAT, two probe sets each) and corresponding RNA expression data was demonstrated by scatter plots.

#### Microarray analyses

**Tumor sample preparation and array hybridization** For microarray analyses, snap frozen tissue specimens were cut into 7- $\mu$ m-thick sections that were stained with H&E. Stained sections were reviewed by a pathologist to identify areas of vital tumor cells and to ensure a tumor content of 80–90%. Corresponding tumor areas were macrodissected by vertical 3-mm incision into the frozen tissue with a sterile blade. Incision was followed by a series of ten 20- $\mu$ m frozen sections. Separated tumor areas were harvested by sterile micropipette tip and collected in buffer (RLT buffer, RNeasy Mini Kit; Qiagen, Hilden, Germany). Each series of ten sections was followed by a 7- $\mu$ m H&E-stained section to control tissue composition. The number of tissue sections used to extract RNA was dependent on the expanse of the area of individual tumor tissue.

Total RNA was isolated using the RNeasy Mini Kit (Qiagen) according to the manufacturer’s instructions and quantified using the Nanodrop ND-1000 UV–vis spectrophotometer (Nanodrop Technologies, USA). The quality of the RNA was controlled using the BioAnalyzer (Agilent Technologies, USA), and exclusively high quality RNA (RIN $\geq$ 7.6) was used for further analysis. For Affymetrix GeneChip analysis, 3  $\mu$ g total RNA of each sample was converted to biotin-labeled cRNA and hybridized on HG-U133 Plus 2.0 arrays (Affymetrix, USA), following the manufacturer’s recommendations.

**Microarray data analysis** The quality of all microarrays was reviewed by inspection of scatter plots (MvA plots)

[25]. Variation of non-biological origin between the arrays were reduced by normalization (variance stabilization) using the *vs*n package in R (language and environment for statistical computing and graphics). “*vs*n” is a robust method for normalization of large-scale gene expression data. When running experiments that involve multiple high-density oligonucleotide arrays, it is important to remove sources of variation between arrays of non-biological origin. Normalization is a process for reducing this variation that works also on values that are negative after background subtraction [10]. For construction of a classifier for relapse (yes/no), the method of “nearest shrunken centroids” was applied [26] based on all stage UICC II patients and on the subgroup of microsatellite stable (MSS) patients. To avoid overfitting, a repeated double cross-validation procedure was used [27]. The data have been deposited in NCBI’s Gene Expression Omnibus and are accessible through GEO Series accession number GSE18088 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=hncvtygayqgmghg&acc=GSE18088>).

Data of previously published prognostic gene expression signatures involving patients with stage UICC II colon cancer were analyzed by testing their power to separate between patients with relapse or disease-free patients in our data set using “global test.” This test can determine whether the global expression pattern of a group of genes is significantly related to clinical variable [28] (Table 2). The two data sets of Lin et al. [22] were validated as published by the authors (New Zealand data: support vector machine; German data set: three nearest neighbor classifier, leave-one-out cross-validation, permutation approach).

#### Results

Our study comprised paraffin-embedded and formalin-fixed tissues from 68 patients all of which have been employed for immunohistochemistry (IHC) detection of protein expression (“protein collection”). In addition, frozen tissue specimens were available for 53 of these 68 patients (78%).

J. Gröne (✉) · H.-J. Buhr  
Department of General, Vascular and Thoracic Surgery, Charité University Medicine Berlin,  
Campus Benjamin Franklin, Hindenburgdamm 30,  
12200 Berlin, Germany  
e-mail: joern.groene@charite.de

D. Lenz · H. Stein · M. Hummel  
Institute of Pathology, Charité University Medicine Berlin,  
Campus Benjamin Franklin, Hindenburgdamm 30,  
12200 Berlin, Germany

V. Jurinovic · U. Mansmann  
Institut für Medizinische Informatik Biometrie  
Epidemiologie (IBE),  
Munich, Germany

M. Hummel  
Core Facilities-Microarray Unit, Centre for Genomic Regulation,  
C/Dr. Aiguader 88,  
08003 Barcelona, Spain

H. Seidel · G. Leder · G. Beckmann · A. Sommer  
Target Discovery, Bayer Schering Pharma AG,  
Müllerstr. 178,  
13353 Berlin, Germany

R. Grützmann · C. Pilarsky  
Department of Visceral, Thoracic and Vascular Surgery,  
University Hospital Dresden,  
Fetscherstr. 74,  
01307 Dresden, Germany