How To for Gene Level Summarization on Exon Arrays

Introduction

Affymetrix GeneChip® Human, Mouse, and Rat Exon 1.0 ST Arrays provide a powerful platform for understanding both the expression of particular exonic regions of a gene, and expression of the gene as a whole. By leveraging probes over the entire gene region, gene-level signal estimates from Exon Arrays have certain advantages over conventional 3' Expression Array designs, which have been the standard with almost all microarray vendors. These advantages with using Exon Arrays for gene-level analysis include:

- A single expression measurement for a gene versus several measures on the 3' ends for different transcript variants (including polyadenylation variants)
- A single expression measurement which takes into account the entire gene region, rather than assuming that the expression from one region is indicative of expression for the whole gene
- A more sensitive expression measurement because most genes are covered by many probes

Using gene-level results from Exon Arrays also affords the opportunity to drill down into exon-level expression results at the same time or at a latter date. Thus researchers gain insight into not just the gene-level expression quantation, but what is going on within the gene at the level of alternative splicing

Additional Resources

There are two technical White Papers which are considered must read for anyone interested in doing gene-level signal estimation work on Exon Arrays.

The first is the Exon Probeset Annotations and Transcript Cluster Groupings white paper (http://www.affymetrix.com/support/technical/whitepapers/exon_probeset_trans_clust_whitepaper.pdf). This white paper provides documentation on how gene-level groupings (called transcript clusters) are generated by Affymetrix for the various GeneChip Exon Arrays.

The second key resource is the Gene Level Signal Estimates from Exon Arrays white paper (http://www.affymetrix.com/support/technical/whitepapers/exon_gene_signal_estimate_whitepaper.pdf). This white paper provides detailed information about how gene-level estimates can be generated from Exon Arrays and the implications of including or excluding particular exon-level probesets from a transcript cluster grouping.

How to Generate Gene-Level Estimates – Using the ExACT GUI Workflow

1. Download and install ExACT (http://www.affymetrix.com/products/software/specific/exact.affx)

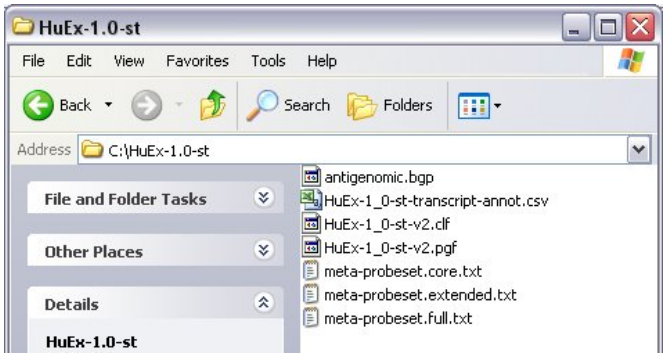
2. Download and unzip the necessary files for your particular Exon Array from the array support page, http://www.affymetrix.com/support/technical/byproduct.affx?cat=arrays
    a. For Human Exon 1.0 ST Array, http://www.affymetrix.com/support/technical/byproduct.affx?product=huexon-st
        i. ExACT Library Files: http://www.affymetrix.com/Auth/support/downloads/library_files/HuEx-1_0-st-v2_exactfiles.zip
        ii. One of the probeset level Design Time Annotation Files which include the meta probeset files. The meta probeset file is what groups multiple exon level probe sets into a single gene level probe set. There are different Design Time Annotation Files for different versions of the genome.
            1. NCBI Build 34: http://www.affymetrix.com/Auth/analysis/downloads/exon/HuEx-1_0-st-v2.design-annot-hg16.zip
            2. NCBI Build 35: http://www.affymetrix.com/Auth/analysis/downloads/exon/HuEx-1_0-st-v2.design-annot-hg17.zip
        iii. NetAffx™ transcript level annotation file which contains biological annotations for the transcript clusters defined in the meta probe set file: http://www.affymetrix.com/Auth/analysis/downloads/taf/HuEx-1_0-st-transcript-annot_csv.zip
    b. For Mouse Exon 1.0 ST Array, http://www.affymetrix.com/support/technical/byproduct.affx?product=moexon-st
        i. ExACT Library Files: http://www.affymetrix.com/Auth/support/downloads/library_files/MoEx-1_0-st-v1_exactfiles.zip
        ii. One of the probe set level Design Time Annotation Files which include the meta probe set files. The meta probe set file is what groups multiple exon level probe sets into a single gene level probe set. There are different Design Time Annotation Files for different versions of the genome.
            1. Mm5: http://www.affymetrix.com/Auth/analysis/downloads/exon/MoEx-1_0-st-v1.design-annot-mm5.zip
            2. Mm6: http://www.affymetrix.com/Auth/analysis/downloads/exon/MoEx-1_0-st-v1.design-annot-mm6.zip
            3. Mm7: http://www.affymetrix.com/Auth/analysis/downloads/exon/MoEx-1_0-st-v1.design-annot-mm7.zip

iii. NetAffx transcript level annotation file which contains biological annotations for the transcript clusters defined in the meta probeset file: http://www.affymetrix.com/Auth/analysis/downloads/taf/MoEx-1_0-st-transcript-annot_csv.zip

c. For Rat Exon 1.0 ST Array, http://www.affymetrix.com/support/technical/byproduct.affx?product=raexon-st
   i. ExACT Library Files: http://www.affymetrix.com/Auth/support/downloads/library_files/RaEx-1_0-st-v1_exactfiles.zip
   ii. One of the probe set level Design Time Annotation Files which include the meta probe set files. The meta probe set file is what groups multiple exon level probesets into a single gene level probe set. Typically there are different Design Time Annotation Files for different versions of the genome. Currently only one version of the rat genome is provided, Rn3: http://www.affymetrix.com/Auth/analysis/downloads/exon/RaEx-1_0-st-v1.design-annot-rn3.zip
   iii. NetAffx transcript level annotation file which contains biological annotations for the transcript clusters defined in the meta probeset file: http://www.affymetrix.com/Auth/analysis/downloads/taf/RaEx-1_0-st-transcript-annot_csv.zip

3. Unzip the downloaded files. You may find it easier to move the files you will need into a single folder. (Note, there are README files in most of the ZIP files which provide more information about the files contained in the ZIP.)

   For example, for the Human Exon 1.0 ST array and annotation files on build 34 you might want to place the following files into a single folder:



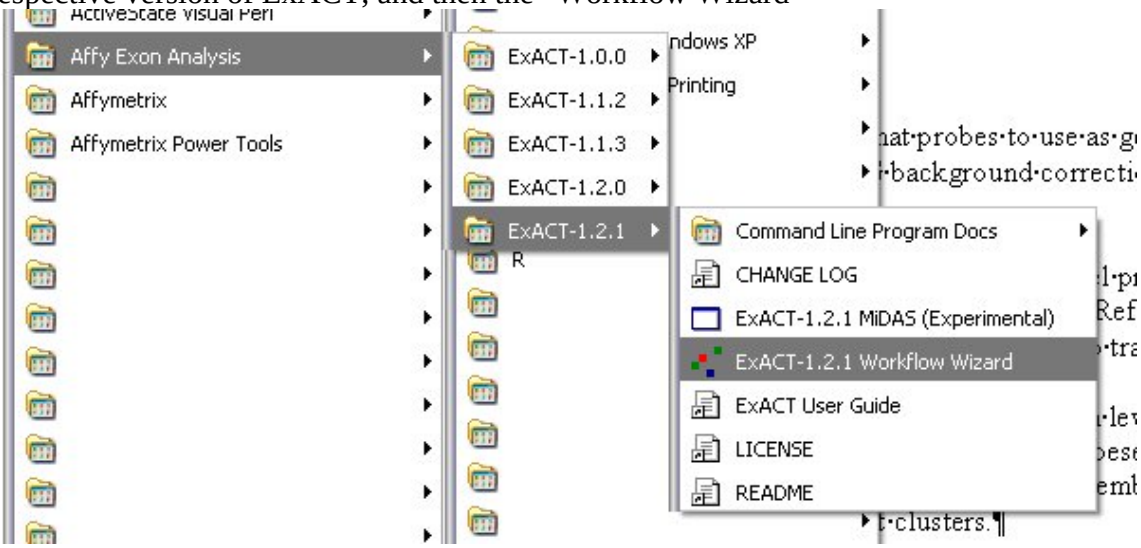
   a. From ExACT Library Zip File:
      i. HuEx-1_0-st-v2.pgf: This file indicates what probes belong to an exon level probe set. The content in this file together with the clf file (see below) replace what was previously provided in the CDF file for 3' Expression Arrays.

ii. HuEx-1_0-st-v2.clf: This file indicates what probe is where in the CEL file.
iii. antigenomic.bgp: This file indicates what probes to use as generic background probes (for the PM-GCBG background correction method).

b. From Design Time Annotation Zip File:
i. meta-probeset.core.txt: This file groups unique exon level probe sets with strong annotation support (IE the probe set hits RefSeq and other putative complete CDS mRNA sequences) into transcript clusters.
ii. meta-probeset.extended.txt: This file groups unique exon level probe sets with empirical annotation support (IE the probe set hits an mRNA or EST sequence or is part of the Vega or Ensembl transcript annotation set) into transcript clusters.
iii. meta-probeset.full.txt: This file groups all unique exon level probe sets into transcript clusters (including probe sets based only on *ab initio* transcript predictions).

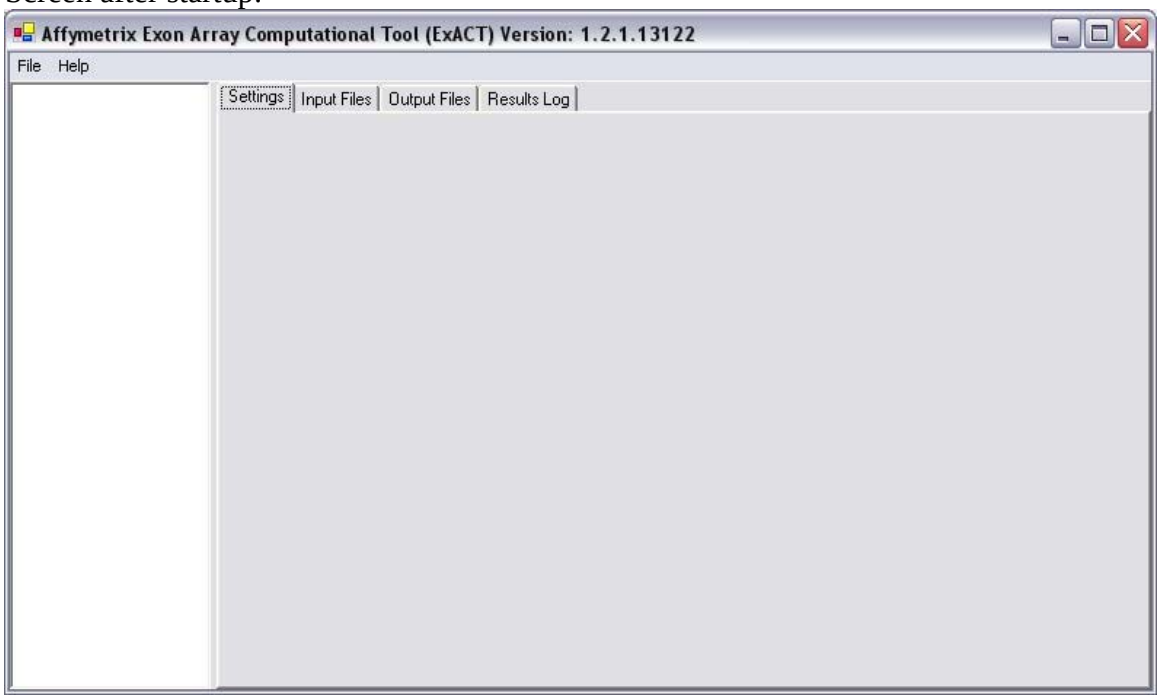
c. From NetAffx Transcript Annotation File:
i. HuEx-1_0-st-transcript-annot.csv: This file provides various biological annotations for each of the transcript clusters. See the associated README for more information on the various biological annotations provided.

4. Start ExACT Workflow GUI
a. From the Windows START menu select “Affy Exon Analysis”, the respective version of ExACT, and then the “Workflow Wizard”

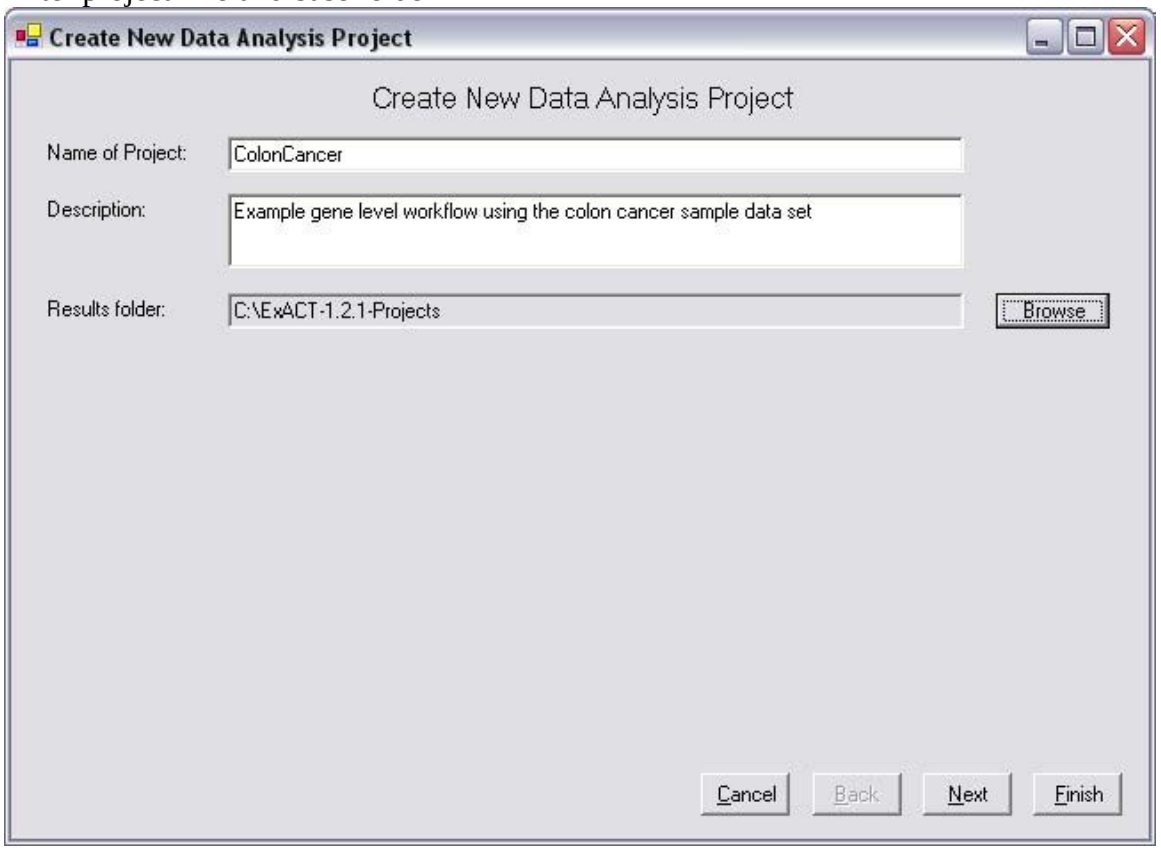


b. Take a look at the ExACT User Guide (also under the START menu) for more info on using the ExACT GUI

c. Screen after startup:

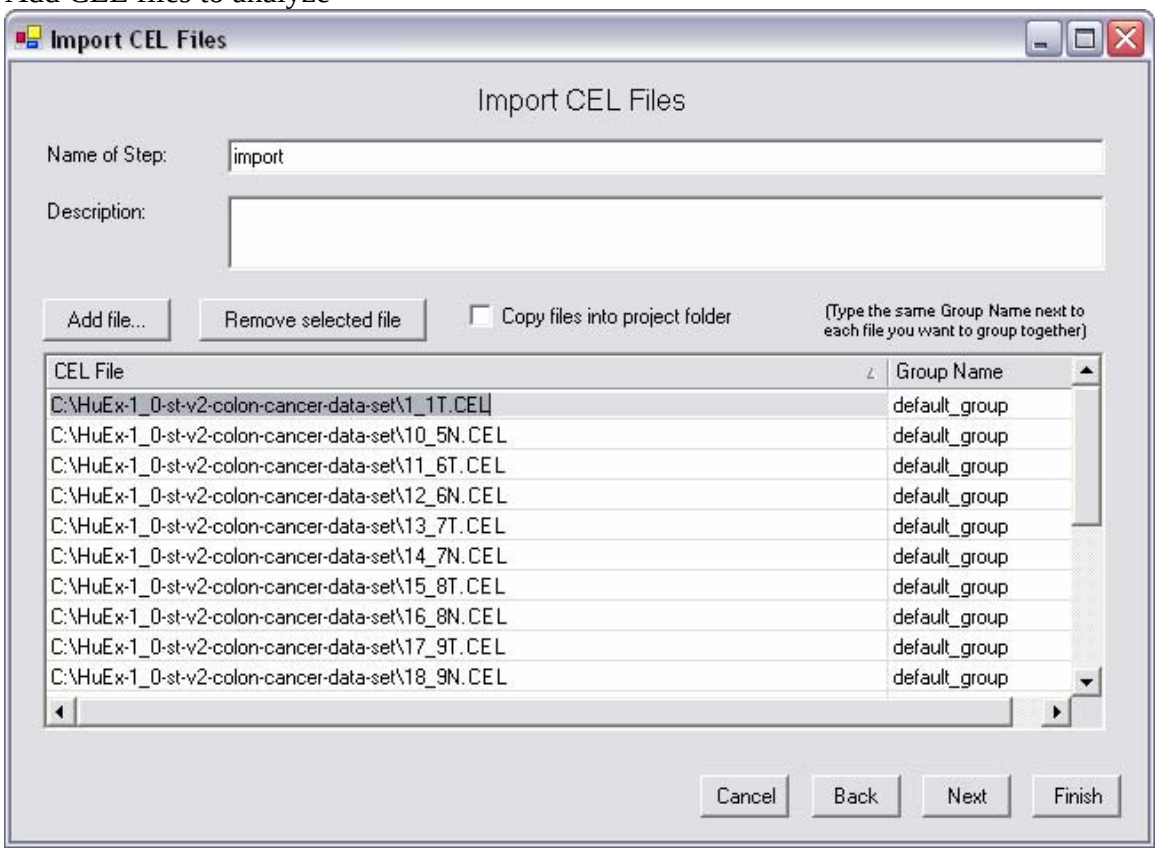


5. Create a new project
   a. File->New Project
   b. Enter project info and base folder



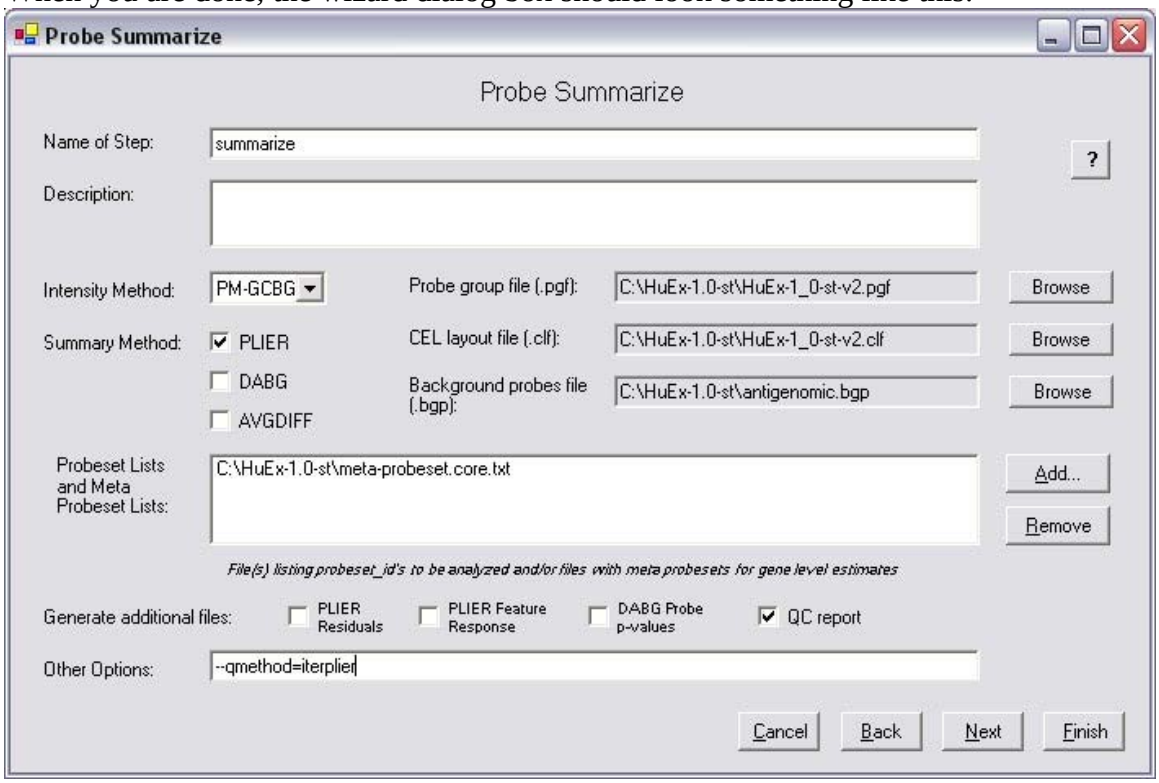
   c. Click next

d. Add CEL files to analyze



e. Click next 4 times in a row
    i. We are selecting the default normalization options by clicking through the wizard here.
    ii. This will bring you to a window titled “Probe Summarize”

f. Configure ExACT to compute gene-level estimates per the steps below. When you are done, the wizard dialog box should look something like this:



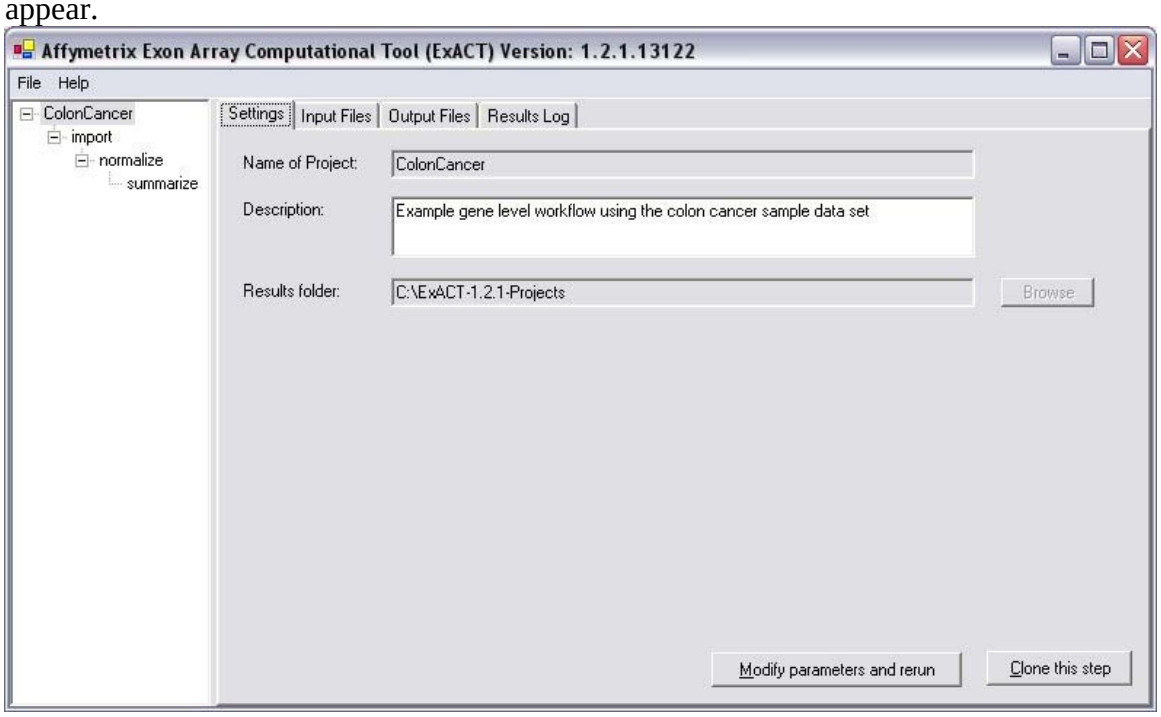
i. Populate the “Probe group file” field with the pgf file downloaded as part of the ExACT library files
ii. Populate the “CEL layout file” field with the clf file downloaded as part of the ExACT library files
iii. Populate the “Background probes file” field with the bgp file downloaded as part of the ExACT library files
iv. In this case we are using PM-GCBG background correction. PM (pm-only) is another common choice.
v. Deselected DABG. These values are useful when analyzing the data at the exon level, but they are difficult to interpret at the gene level.
vi. **The key part to generating a gene-level estimate is specifying a Meta Probeset List in the “Probeset Lists and Meta Probeset Lists” box.** In this example the “core” probesets (ie those exon level probesets with RefSeq and complete CDS mRNA support) will be used to generate gene level signal estimates.
vii. This example also uses both the PLIER method and the IterPLIER method are used to generate two different sets of gene- level estimates. See the white paper listed above for more information about these methods. In general, one will want to use IterPLIER for gene level results. **The key “trick” to getting IterPLIER results is to add “—qmethod=iterplier” to the “Other Options” field.** This will force the software to run IterPLIER and the results will be placed in the iterplier.summary.txt output file.

g. Click “Finish” to start the gene-level analysis

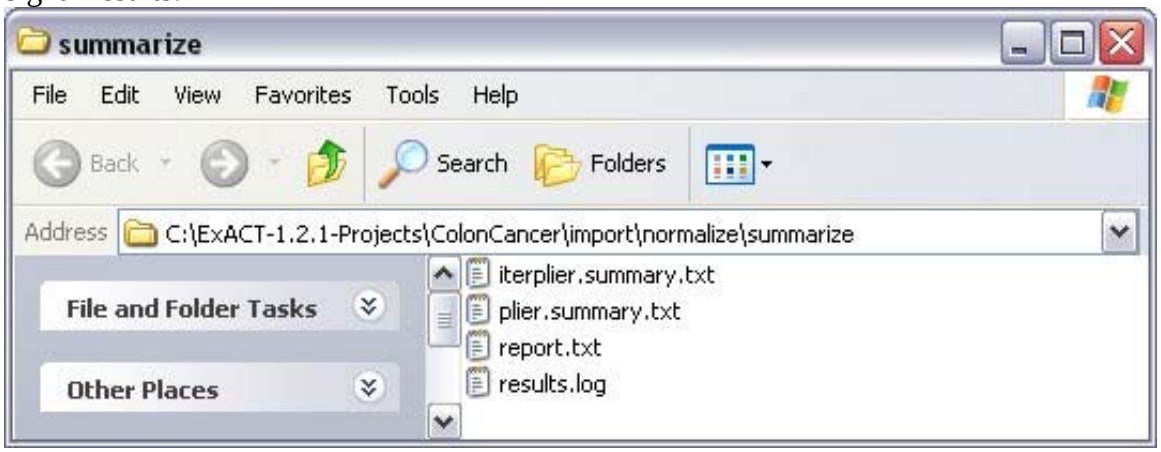
6. At this point a “Progress of Execution” window will appear along with a log window. First the CEL files will be normalized using Sketch normalization and then the gene-level signal estimates will be computed using PLIER and IterPLIER.

   On a Windows 1.8 GHz Pentium M with 2GB of memory, the normalization step takes about 10 minutes for 14 CEL files and the probe summarization step takes about 80 minutes. Using the “extended” or “full” meta probeset list will take longer.

   Once the analysis is complete you should see the ExACT Workflow main window appear.



7. Right click on the “summarize” node in the left hand panel then select “open folder” from the menu. This will open a folder which contains the gene-level signal results:



The iterplier.summary.txt file contains the gene-level summaries from the IterPLIER method. **For gene-level summaries the probeset_id field in the summary.txt file will contain an ID corresponding to the probeset_id defined in the meta probeset list used. For the meta probeset files provided by Affymetrix, this gene level probeset ID is a transcript cluster ID.**

How to Generate Gene-Level Estimates – Using the Affymetrix Power Tools (APT) apt-probeset-summarize

1. Reasons to use apt-probeset-summarize rather than the ExACT Workflow GUI:
    a. Users who are ok with minimal support through DevNet (see note below)
    b. Users who prefer a command line interface
    c. Users who want more control over the analysis
    d. Users who want a faster analysis (as of this writing, the apt-probeset-summarize command line program is significantly (ie 10x) faster than ExACT)
    e. Users want to run their analysis on non MS Windows computers
2. Download APT from http://www.affymetrix.com/support/developer/powertools/index.affx
    a. APT is a DevNet tool, and is only supported by DevNet (see support policy here http://www.affymetrix.com/support/technical/software_support_policy.affx)
    b. Windows, Mac OS-X, and Linux binaries are available, source code is also available
3. Install APT
    a. For MS Windows, simply run the installer. Under MS Windows, you can start up an APT shell for executing APT command line programs from the START menu. See START->AffymetrixPowerTools->…
    b. For other platforms, simply unzip the ZIP file in a convenient location (and optionally, add the binary location to your PATH.)
4. Review the apt-probeset-summarize manual (apt-probeset-summarize.html included in the download)
    a. For MS Windows, this is accessible from START->AffymetrixPowerTools.
    b. For non MS-Windows platforms, there is a docs folder with documentation that can be viewed using an html browser.
5. Download the necessary files per Items (2) and (3) under How to Generate Gene Level Estimates – Using the ExACT GUI Workflow above.
6. Run apt-probeset-summarize:

```
apt-probeset-summarize             \
    –p HuEx-1_0-st-v2.pgf          \
    -c HuEx-1_0-st-v2.clf          \
    -b antigenomic.bgp             \
    –o gene-level-analysis-output  \
    -s meta-probeset.core.txt      \
    -a 'quant-norm.sketch=65536,pm-gcbg,iter-plier'    \
```

```
*.CEL
```

The example above assumes that all the files are in the same folder. If not, you will need to specify paths with the filenames.

Note, on an Intel 2.4 GHz Pentium 4 running Linux it takes about 290 minutes to process 73 CEL files(~4 minutes per CEL file).

Also, note that the above command will do the normalization and signal summarization (no intermediate normalized CEL files).

For the above command,

a. –p is used to specify the pgf file
b. –c is used to specify the clf file
c. –b is used to specify the background probes
d. –o is used to specify the output directory
e. –s is used to specify a meta probeset file (or probeset list)
f. –a is used to specify the analysis string (note that you can have multiple –a options which will result in multiple analysis being run) In this case we are doing sketch normalization and using the BGP probes as surrogate MM values for the IterPLIER method.

Notes on Deviations from the Above Workflow

- A good starting point for gene-level signal estimates is to use the meta-probeset.extended.txt grouping file along with the IterPLIER summarization method.
- One may want to apply various variance stabilization techniques to the IterPLIER results in the summary files
    - Adding a small positive number to the summaries (ie adding 16 to the IterPLIER results)
    - Log transforming the data
    - ie log2(plier signal + 16)
- Users may want to create their own meta probeset files:
    - They may want to create curated groupings of exon level probe sets for particular genes of interest
    - They may want to apply different criteria for grouping probe sets into a transcript cluster (ie they may want to create transcript clusters based on exon level probe sets hitting the same Ensembl gene)
    - Users may want to create a smaller meta probe set file of only the genes their interested in. Such a smaller meta probe set file will speed up the analysis.

Now What

The gene-level summary files generated above are tab separated text files which can be pulled into a variety of downstream analysis applications. For example, one likely follow up step would be to pull the gene-level summary file into a statistical application to test

for significant differential gene expression. Once those significant hits are identified, the NetAffx transcript annotation file can be used to cross reference the significant transcript cluster IDs with biological information such as the gene symbol.

With the gene-level summary file and the NetAffx transcript annotation file, one can proceed with the same types of analysis that are applied to more traditional 3’ Expression Array data such as the Affymetrix GeneChip® HG-U133 Plus 2.0.