Using Cvt2Mae to Convert Affymetrix Array Data for MAExplorer

http://www.lecb.ncifcrf.gov/Cvt2Mae

Peter F. Lemkin(1), Greg Thornwall (2), Bob Stephens(3)

(1) LECB/NCI/FCRDC, (2) SAIC/FCRDC, (3) ABCC/FCRDC

DRAFT - Revised: 01-28-2002
Cvt2Mae version 0.60

Accessing Arrays with MAExplorer

- MAExplorer works with any arrays using the schema (see Appendix C of MAExplorer Reference Manual for details)
- All data files are tab-delimited text files
- Databases could be constructed with tools like Excel for editing user data into the schema format
- The Cvt2Mae array data converter "Wizard" tool converts non-standard <User-defined> academic or commercial data to MAExplorer format
- **Affymetrix**, Incyte, GenePix, Scanalyze, and other array data formats may be converted using predefined "Array Layouts"

S.1 MAExplorer Data Schema

- MAExplorer works with any array data using our data schema

- The schema is described in detail in MAExplorer Reference Manual Appendix C.

- Data Schema: tab-delimited experiment data files:
  1. GIPO (Gene In Plate Order or "array print" file)
  2. List of hybridized samples in database
  3. Configuration data describing the array and conventions
  4. Separate spot quantification data files

- The Cvt2Mae "wizard" tool converts user array data to this schema

S.1.1 MAExplorer GIPO or Print File

- GIPO file maps a spot on the array to a particular gene

- Contains:
  1. location or grid-geometry
  2. one or more genomic identifiers (e.g., Clone ID, GenBank ID, LocusID, etc.)
  3. gene description as Gene Name (or other description)
  4. Optional: global spot quality (QualCheck)
  5. optional: plate coordinates for clones

S.1.2 MAExplorer Samples Database File

- List of hybridized samples file SamplesDB.txt file contains:

1. full sample description

2. base file name of quantification file (without .quant file extension)

3. optional sample ID number

4. other data you wish to carry with the samples (used in array reports)

S.1.3 MAExplorer Configuration Database File

- Configuration data file MaeConfig.txt describes particular type of array and hybridization labeling you are using. This includes:

- grid-geometry - # of replicate fields, grids, rows/grid, columns/grid

- spot hybridization labeling - intensity or ratio data, dye names

- various presentation options - use pseudo-array or actual (x,y) coordinates, etc.

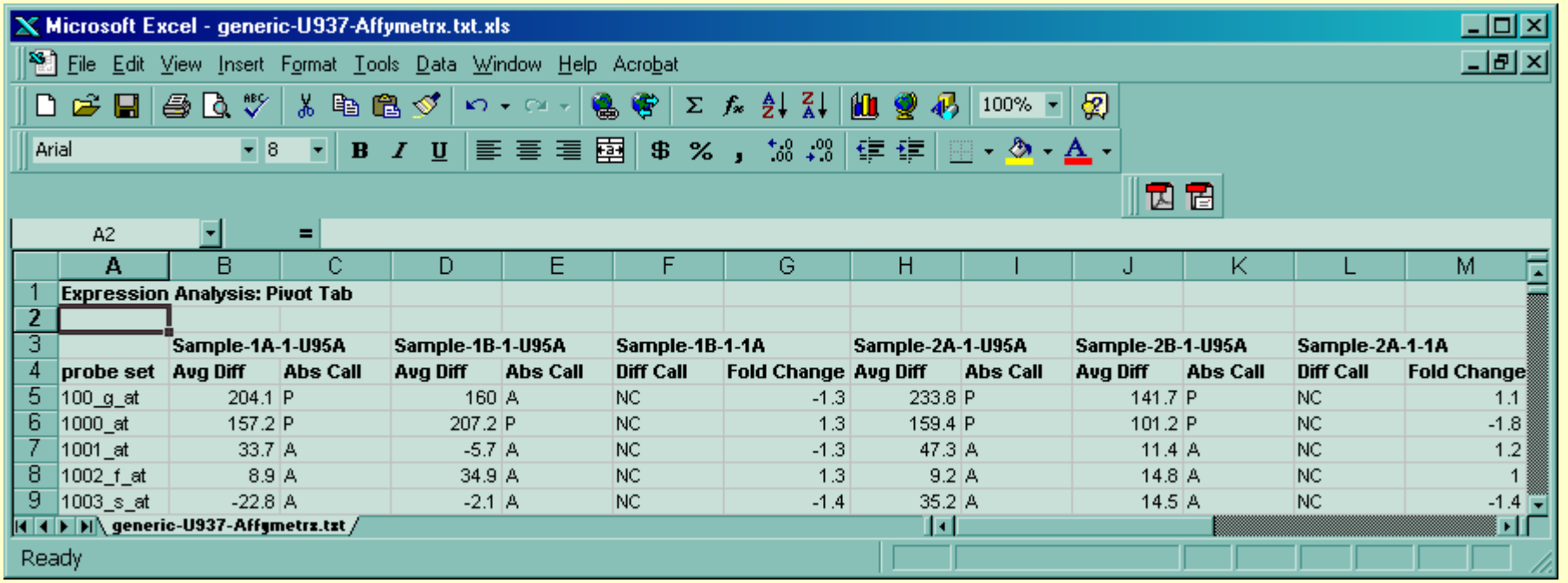
S.1.4 MAExplorer Spot Quantification Files

- Separate spot quantification data files (with .quant file extension) are used for each hybridized sample
- ^{33}P or biotin labeled samples are specified as one hybridization intensity information per file
- Fluorescent Cy3/Cy5-dye labeled samples are specified as two channels of hybridization intensity information per file
- Intensity background data is optional
- Spot quality (QualCheck) data is optional
- Grid-coordinates are specified the same as for GIPO file

S.2 Assumptions About User Data - Array Layout

- User data is tab-delimited ASCII text files (could generate with Excel)
- If the array geometry (#fields, grids, rows/grid, columns/grid) is known, that geometry may be used in MAExplorer
- Otherwise, a pseudo-array geometry is generated for visual use in MAExplorer from the total # of spots in the user data
- An Array Layout describes the user data. It may be edited and saved for subsequent use in converting other array data files of the same type
- The <User-defined> array layout gives users complete flexibility in describing the array

S.3 Example of tab-delimited Affymetrix Data



Microsoft Excel - generic-U937-Affymetrx.txt.xls

File Edit View Insert Format Tools Data Window Help Acrobat

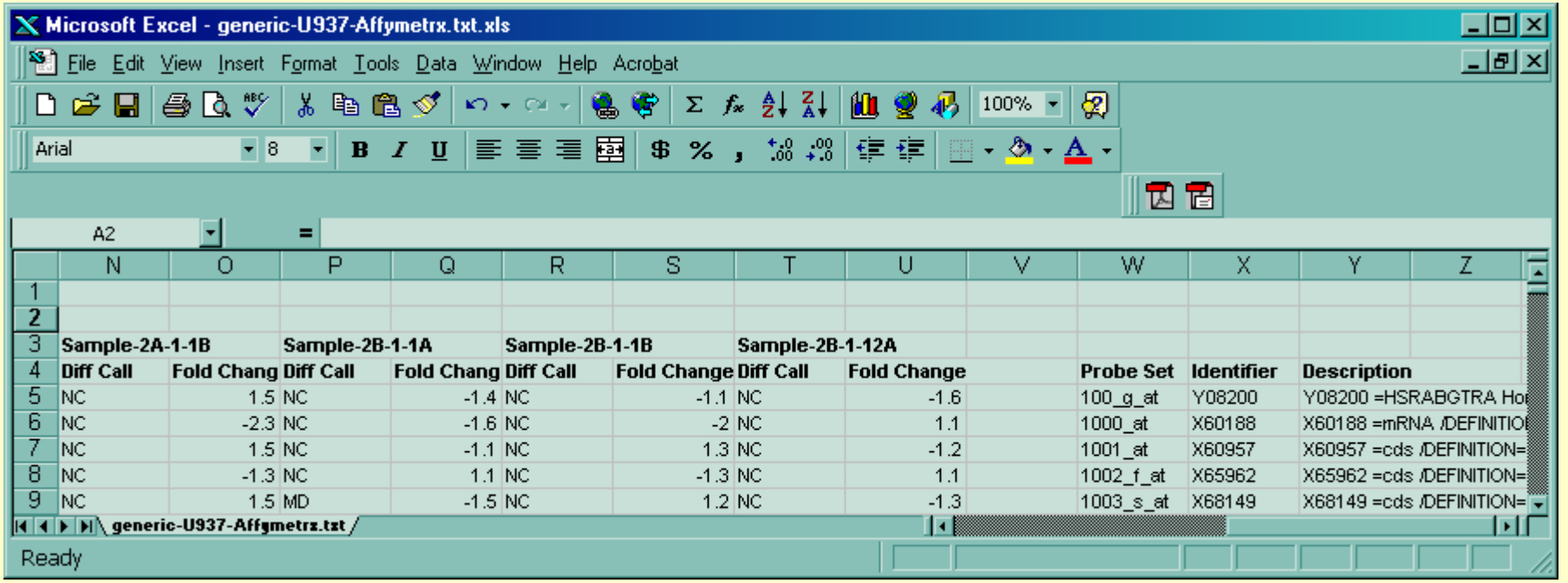
Arial 8 100%

A2 =

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Expression Analysis: Pivot Tab												
2													
3		Sample-1A-1-U95A		Sample-1B-1-U95A		Sample-1B-1-1A		Sample-2A-1-U95A		Sample-2B-1-U95A		Sample-2A-1-1A	
4	probe set	Avg Diff	Abs Call	Avg Diff	Abs Call	Diff Call	Fold Change	Avg Diff	Abs Call	Avg Diff	Abs Call	Diff Call	Fold Change
5	100_g_at	204.1	P	160	A	NC	-1.3	233.8	P	141.7	P	NC	1.1
6	1000_at	157.2	P	207.2	P	NC	1.3	159.4	P	101.2	P	NC	-1.8
7	1001_at	33.7	A	-5.7	A	NC	-1.3	47.3	A	11.4	A	NC	1.2
8	1002_f_at	8.9	A	34.9	A	NC	1.3	9.2	A	14.8	A	NC	1
9	1003_s_at	-22.8	A	-2.1	A	NC	-1.4	35.2	A	14.5	A	NC	-1.4

generic-U937-Affymetrx.txt

Ready



Microsoft Excel - generic-U937-Affymetrx.txt.xls

File Edit View Insert Format Tools Data Window Help Acrobat

Arial 8 100%

A2 =

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1													
2													
3	Sample-2A-1-1B		Sample-2B-1-1A		Sample-2B-1-1B		Sample-2B-1-12A						
4	Diff Call	Fold Chang	Diff Call	Fold Chang	Diff Call	Fold Change	Diff Call	Fold Change		Probe Set	Identifier	Description	
5	NC	1.5	NC	-1.4	NC	-1.1	NC	-1.6		100_g_at	Y08200	Y08200 =HSRABGTRA Ho	
6	NC	-2.3	NC	-1.6	NC	-2	NC	1.1		1000_at	X60188	X60188 =mRNA /DEFINITIO	
7	NC	1.5	NC	-1.1	NC	1.3	NC	-1.2		1001_at	X60957	X60957 =cds /DEFINITION=	
8	NC	-1.3	NC	1.1	NC	-1.3	NC	1.1		1002_f_at	X65962	X65962 =cds /DEFINITION=	
9	NC	1.5	MD	-1.5	NC	1.2	NC	-1.3		1003_s_at	X68149	X68149 =cds /DEFINITION=	

generic-U937-Affymetrx.txt

Ready

I. Procedure: Convert Data for Array Layouts

1. Select the Chip Set array layout (**Affymetrix - generic**) if in list, otherwise pick <User-defined>)

2. Select 1 or more input files using the "Browse input files" .

3. You may edit or change various array layout parameters at this time
   3.1 you may edit the array layout with "Edit Layout"
   3.2 you may "Assign GIPO fields" in user data file
   3.3 you may "Assign Quantification fields" in user data file
   3.4 if you changed any array layout parameters, you may save it with "Save Layout"

4. Select the project output directory (i.e., folder) to save generated files

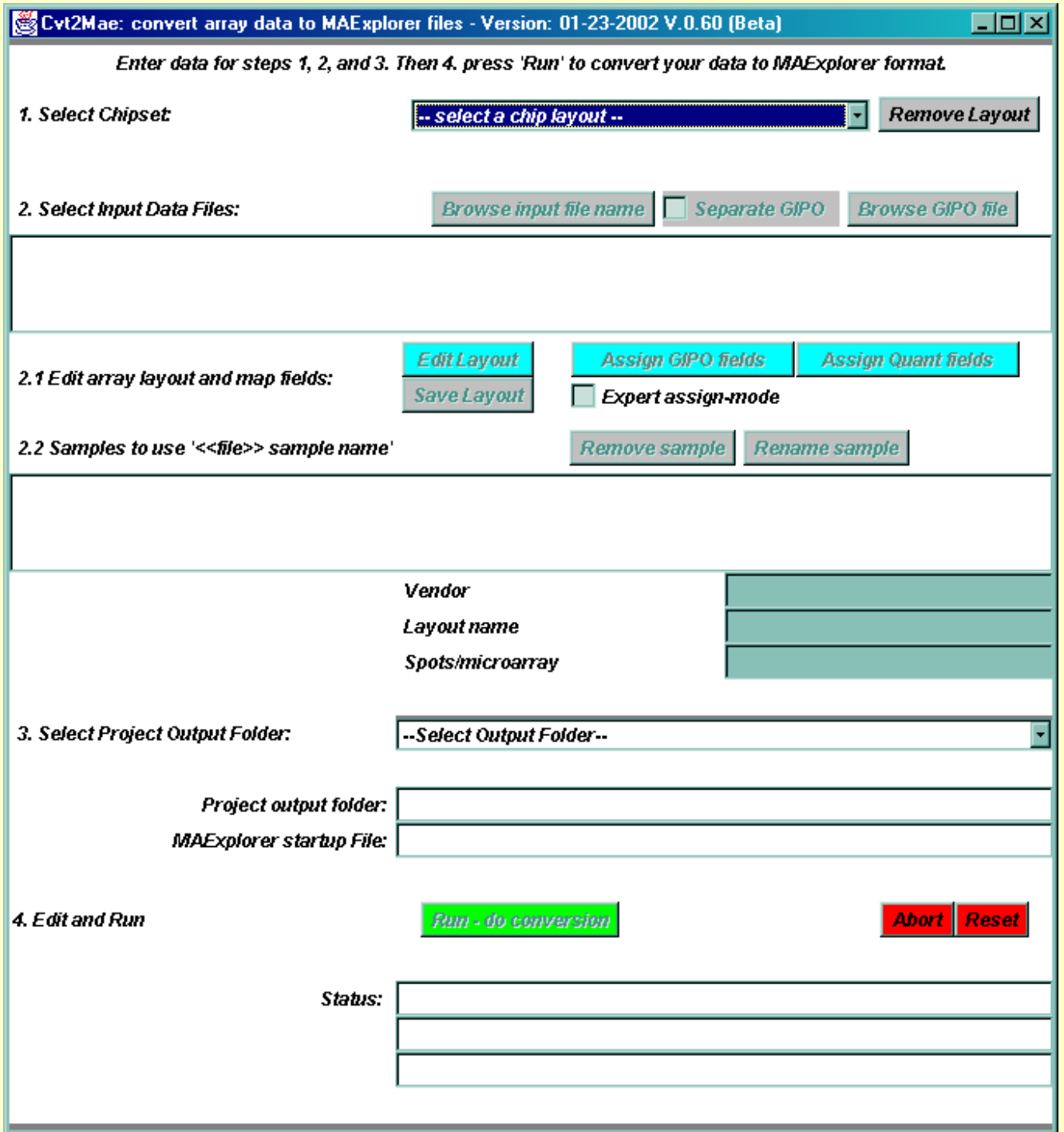
I. Procedure: continued...

5. Press “Run” to convert the data

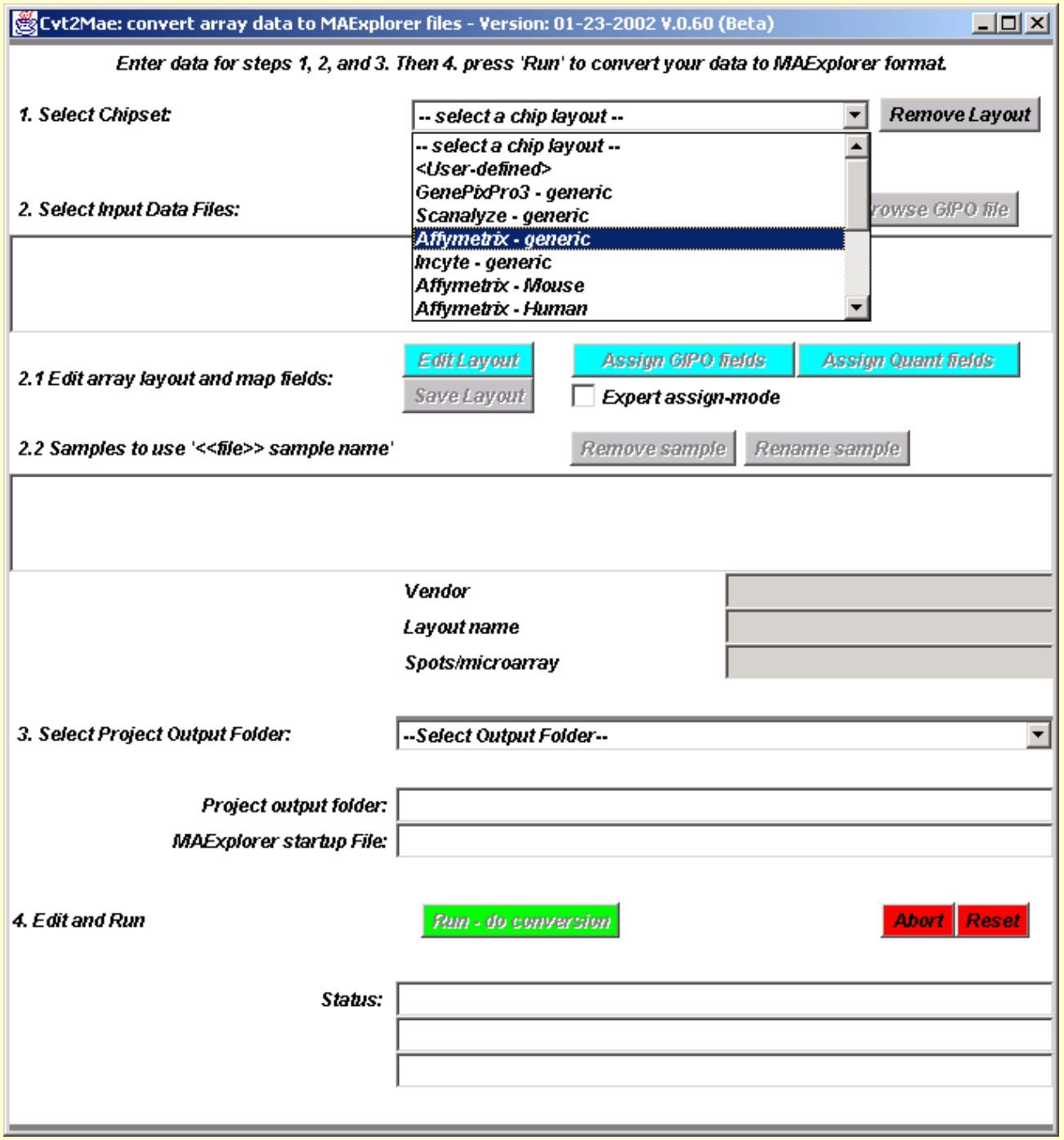
6. Press “Done” when it is finished.

7. Go to the project directory and then to the MAE sub-directory, click on the Start.mae file to start MAExplorer on the new data

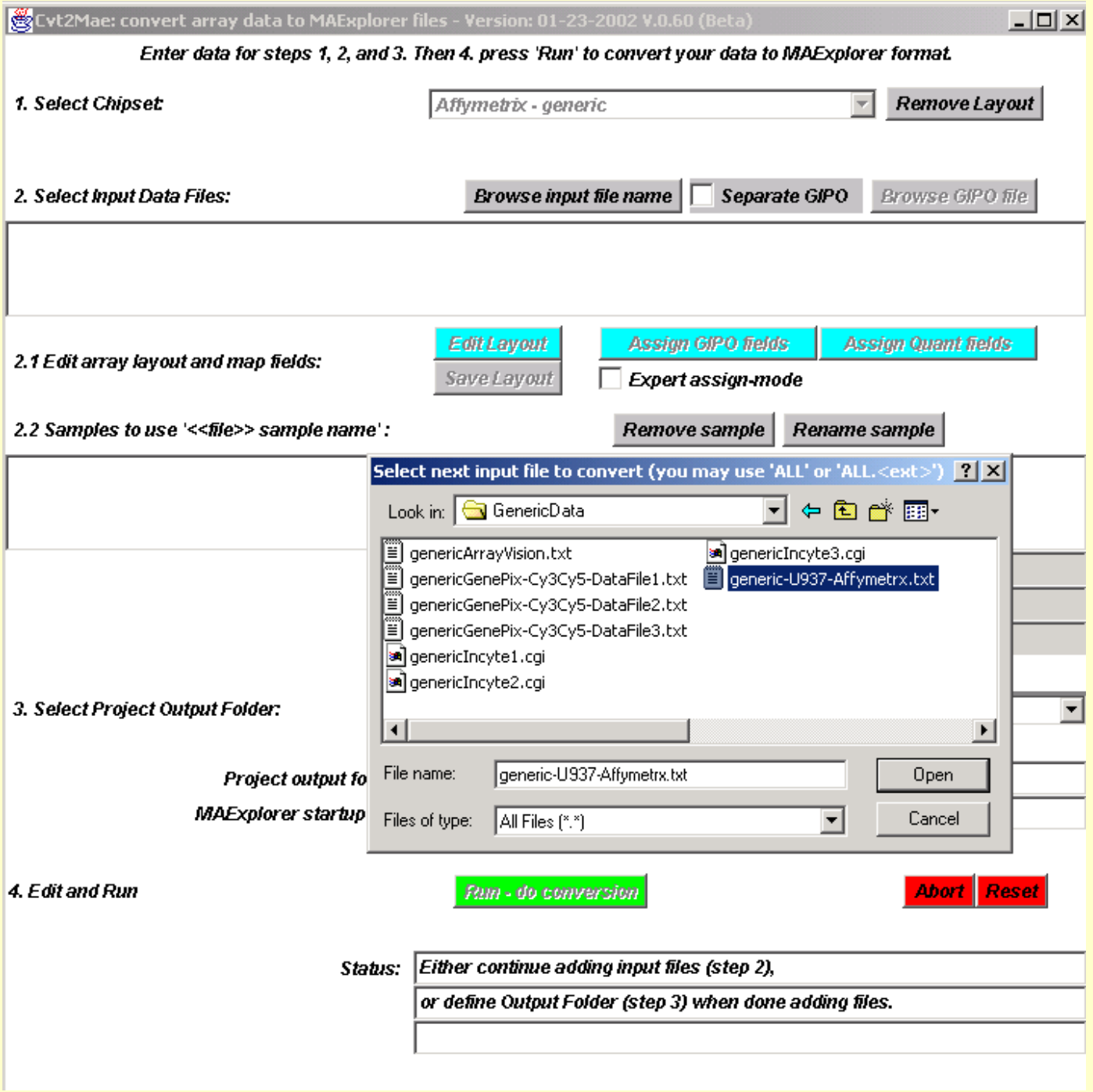
1. Initial State of Cvt2Mae Program



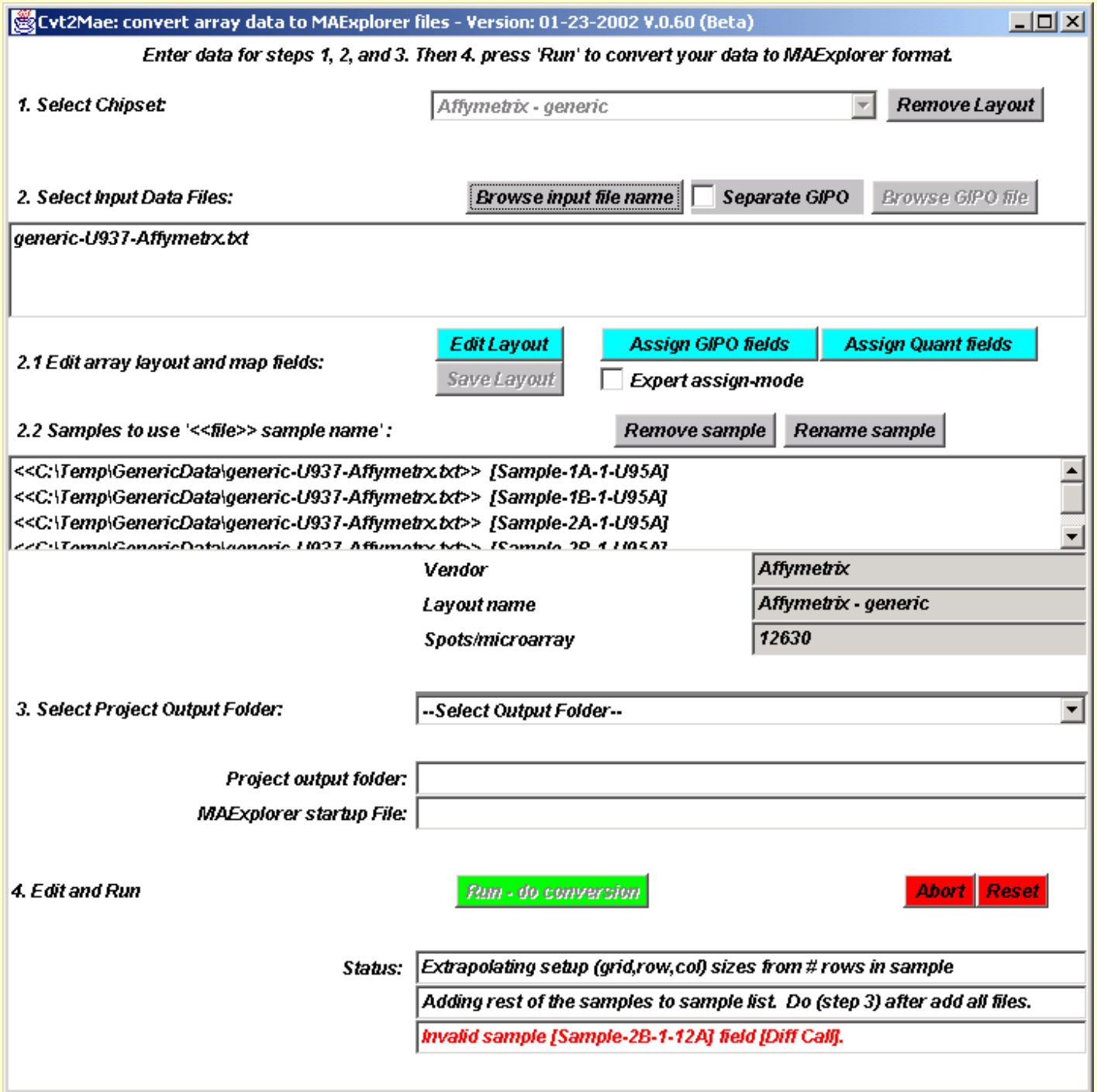
2. Selecting Affymetrix Chipset Array-Layout



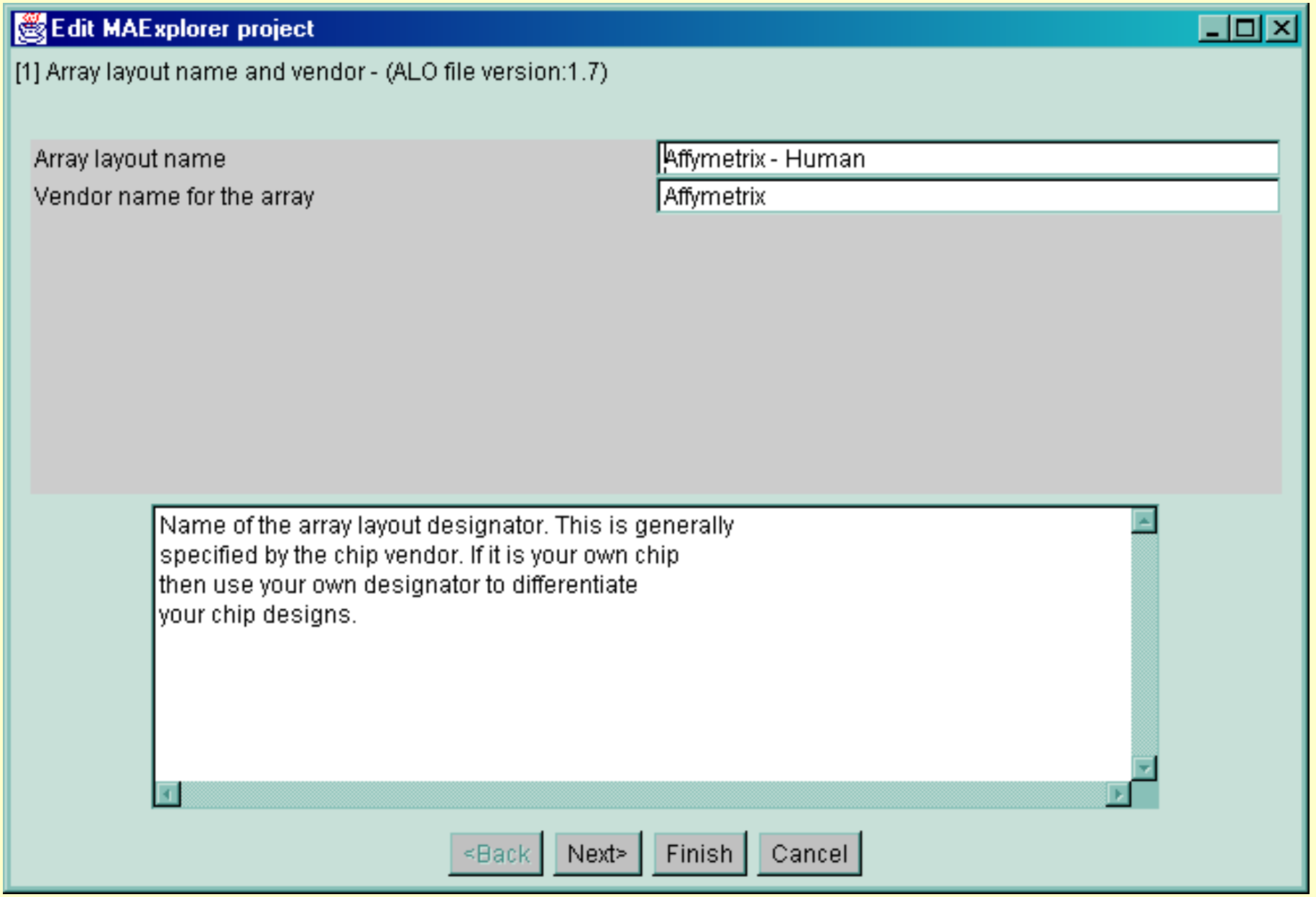
3. Select Files with "Browse input file" Name



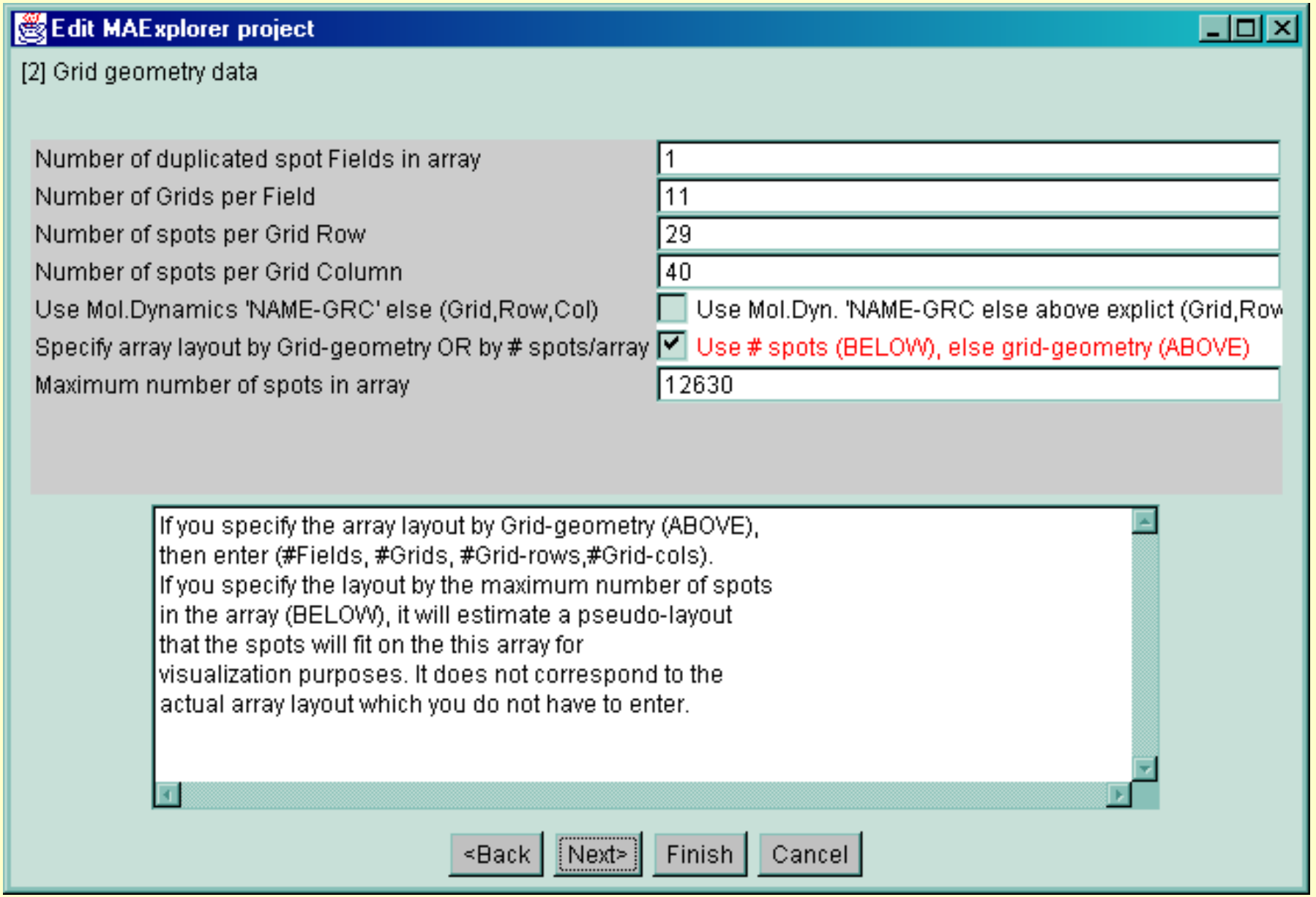
4. Input File(s) Analyzed for Multiple Samples



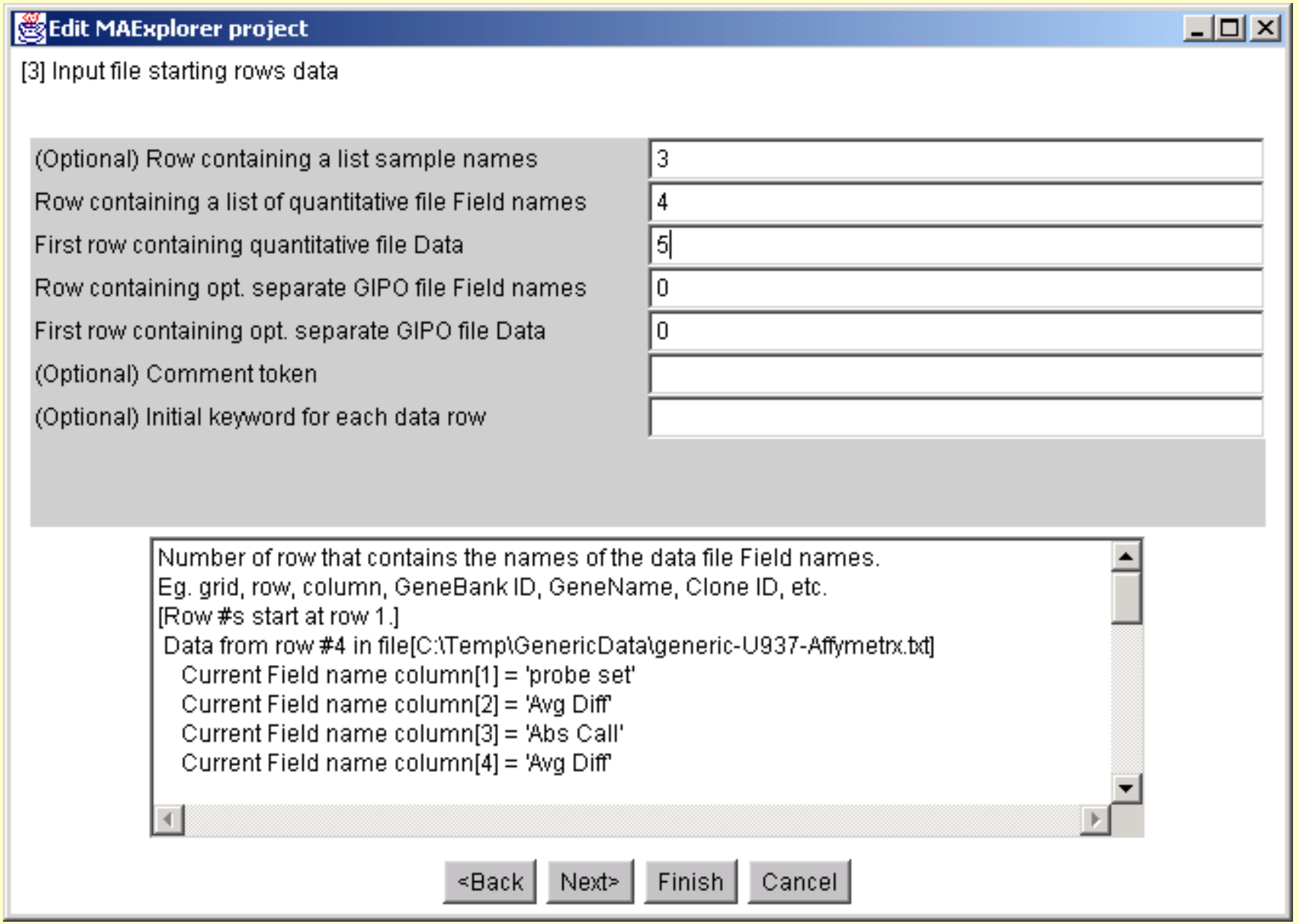
5.1 Edit Layout 'Wizard' Values for This Array



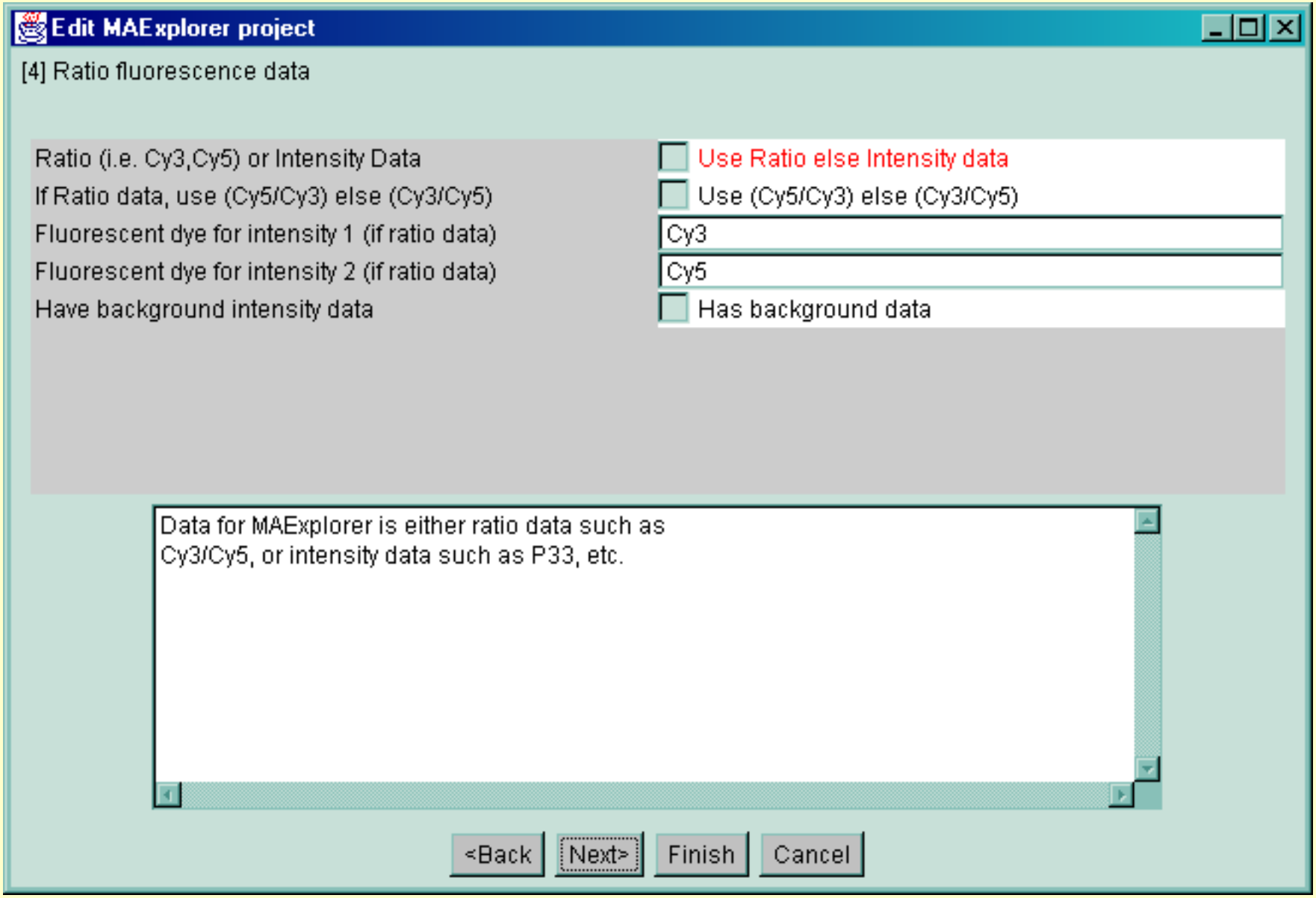
5.2 Edit Layout 'Wizard' Grid Geometry Values



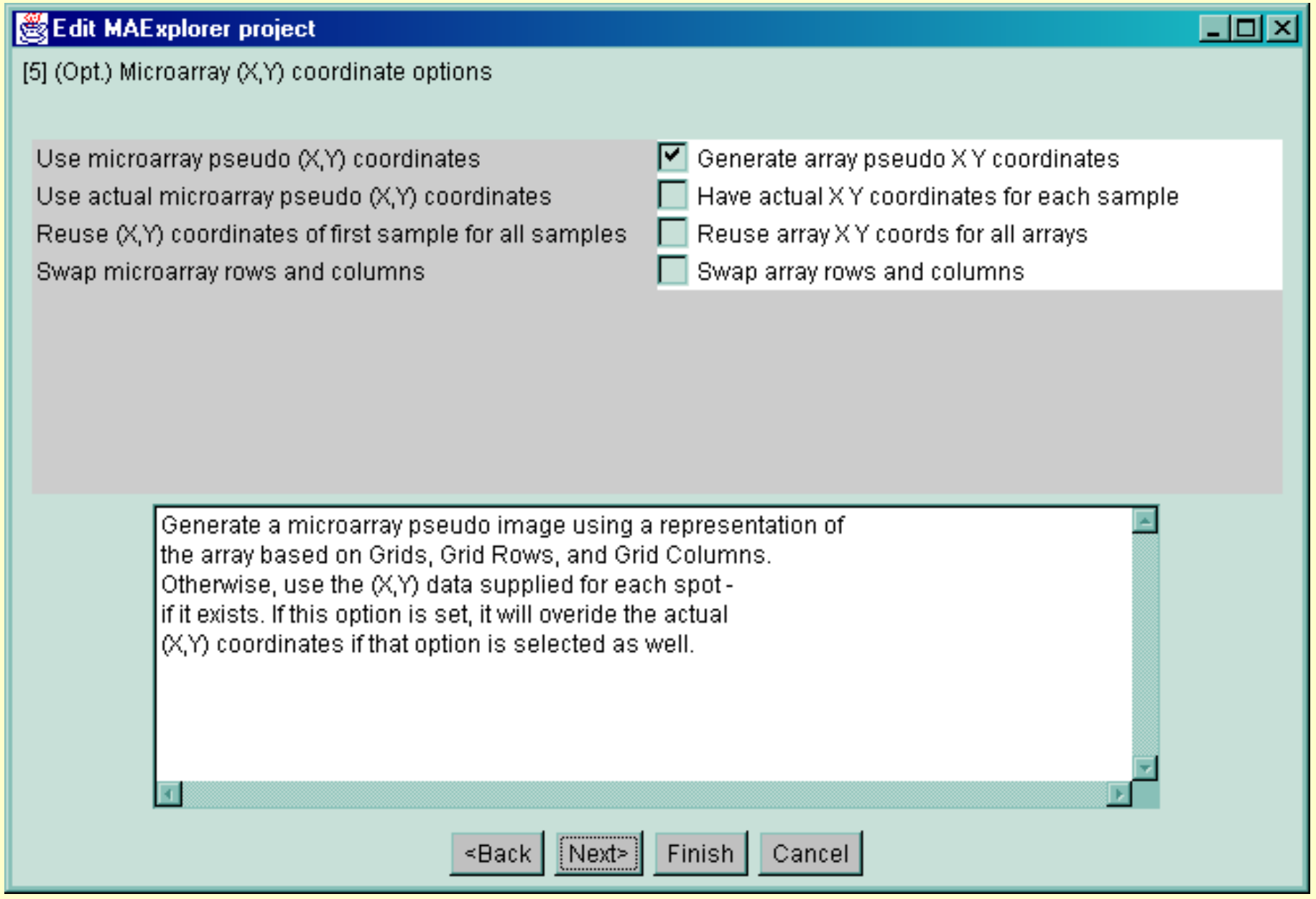
5.3 Edit Layout 'Wizard' Input File Row Values. Verify Rows for Sample & Field Names Defined



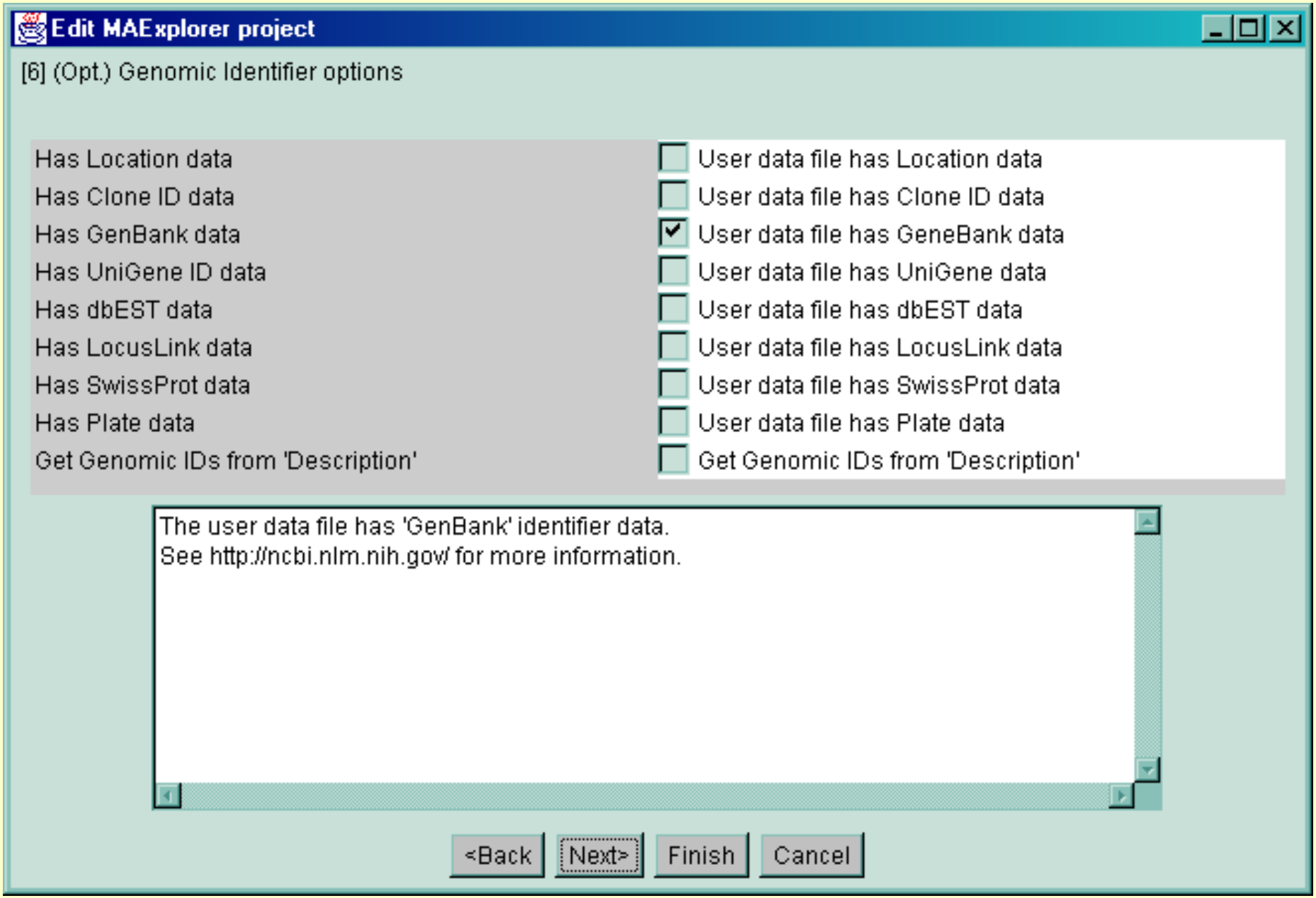
5.4 Edit Layout ‘Wizard’ Ratio or Intensity Values



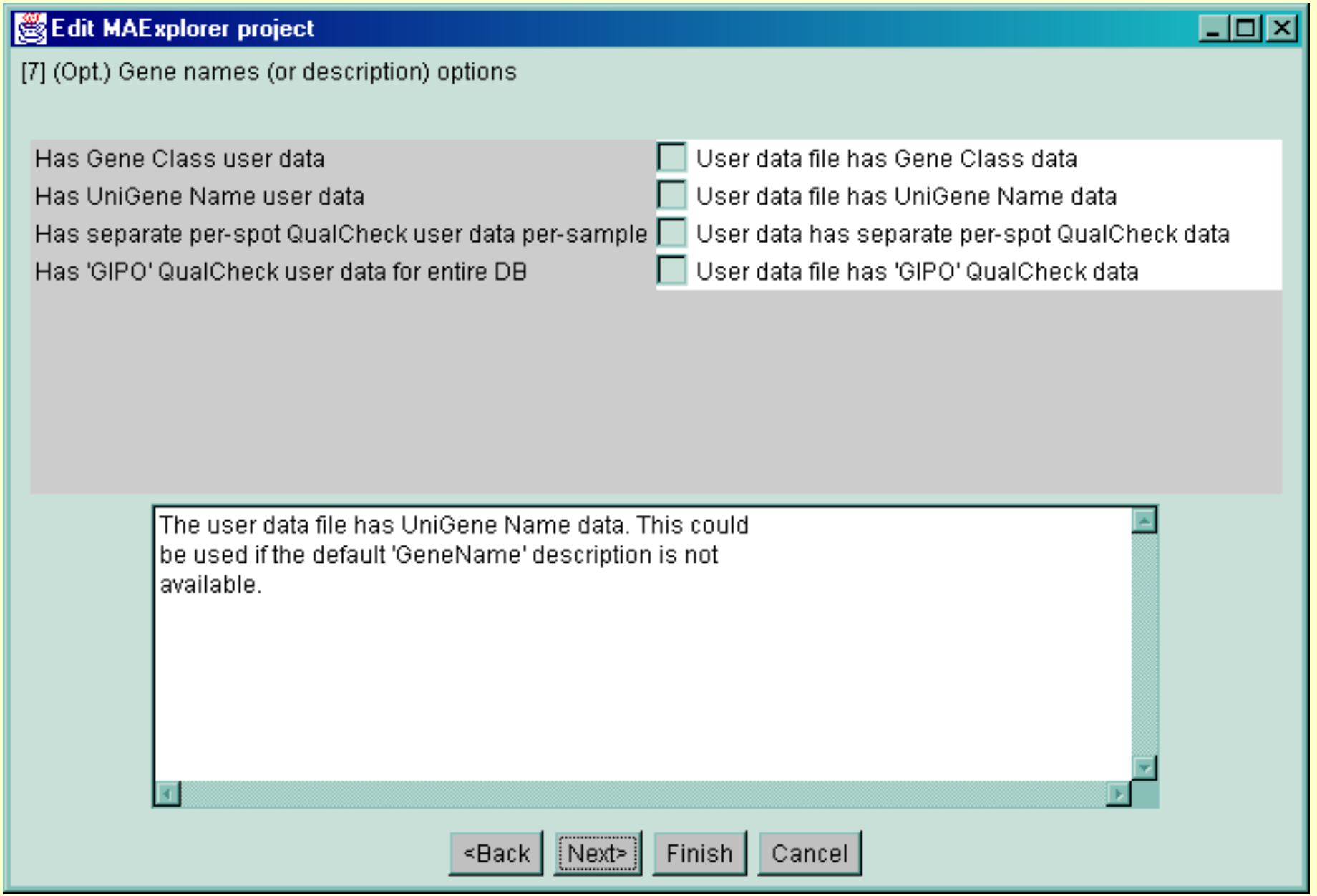
5.5 Edit Layout 'Wizard' optional (X,Y) Coordinate Values



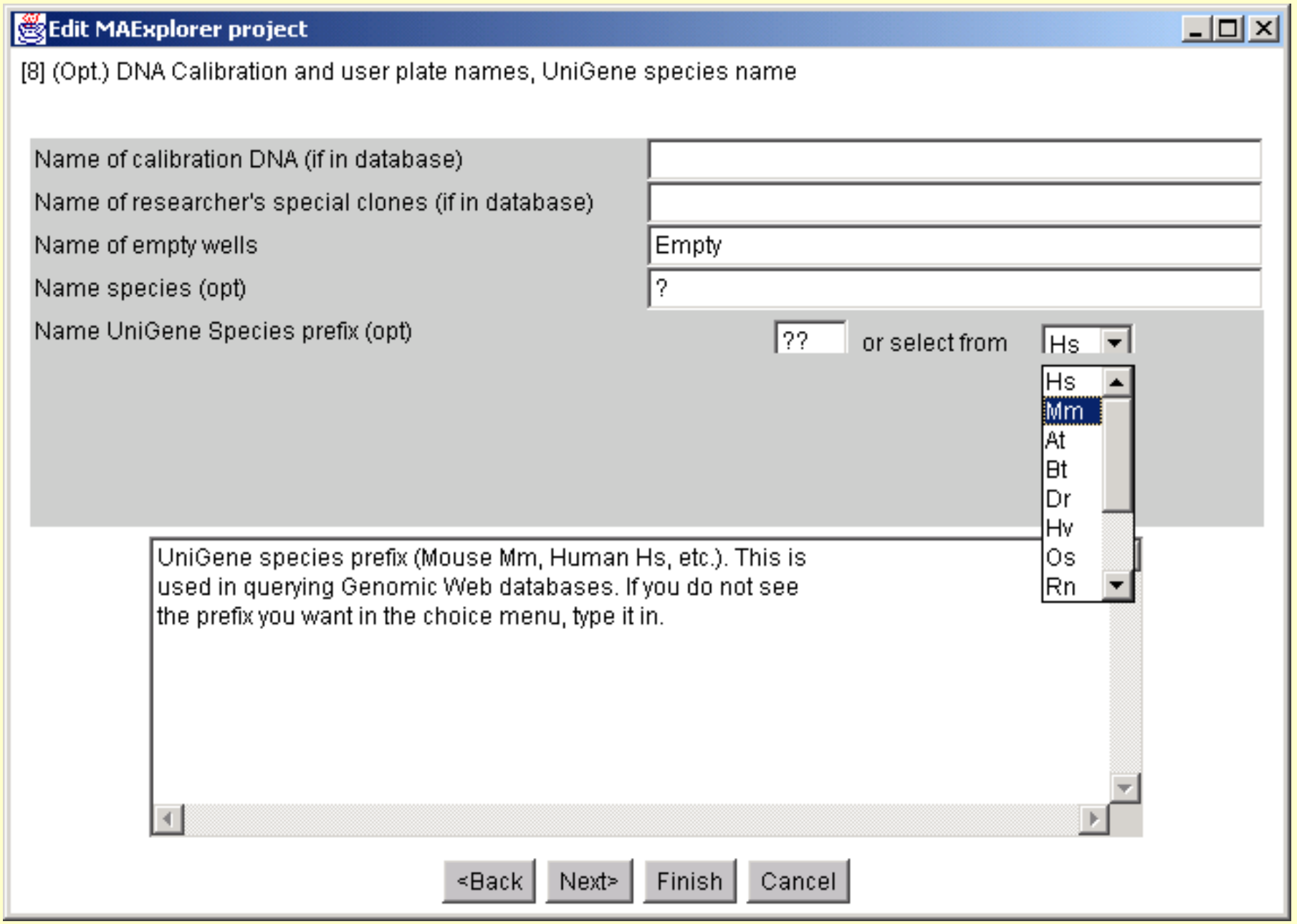
5.6 Edit Layout ‘Wizard’ Genomic ID Values



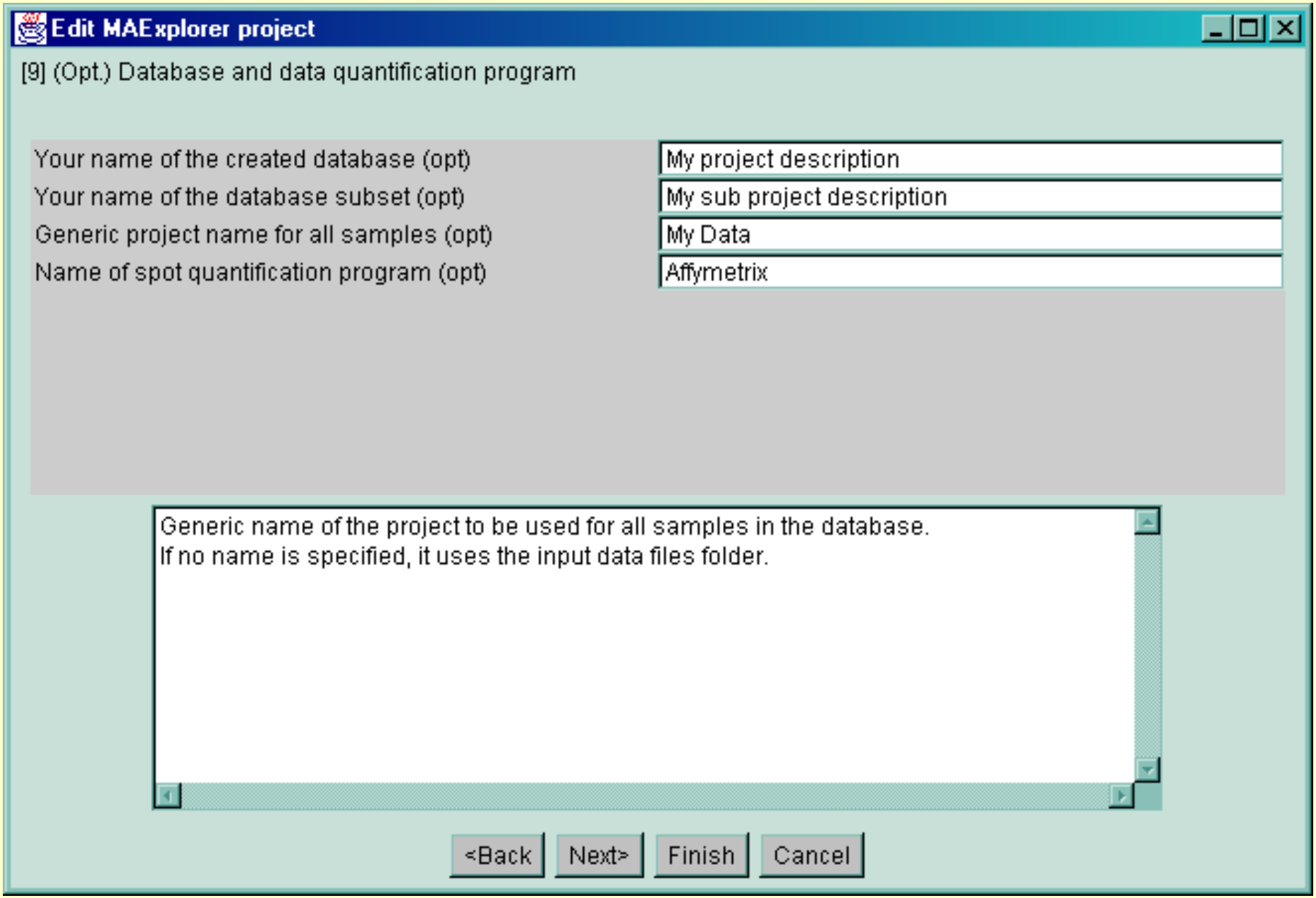
5.7 Edit Layout ‘Wizard’ Gene Names Description



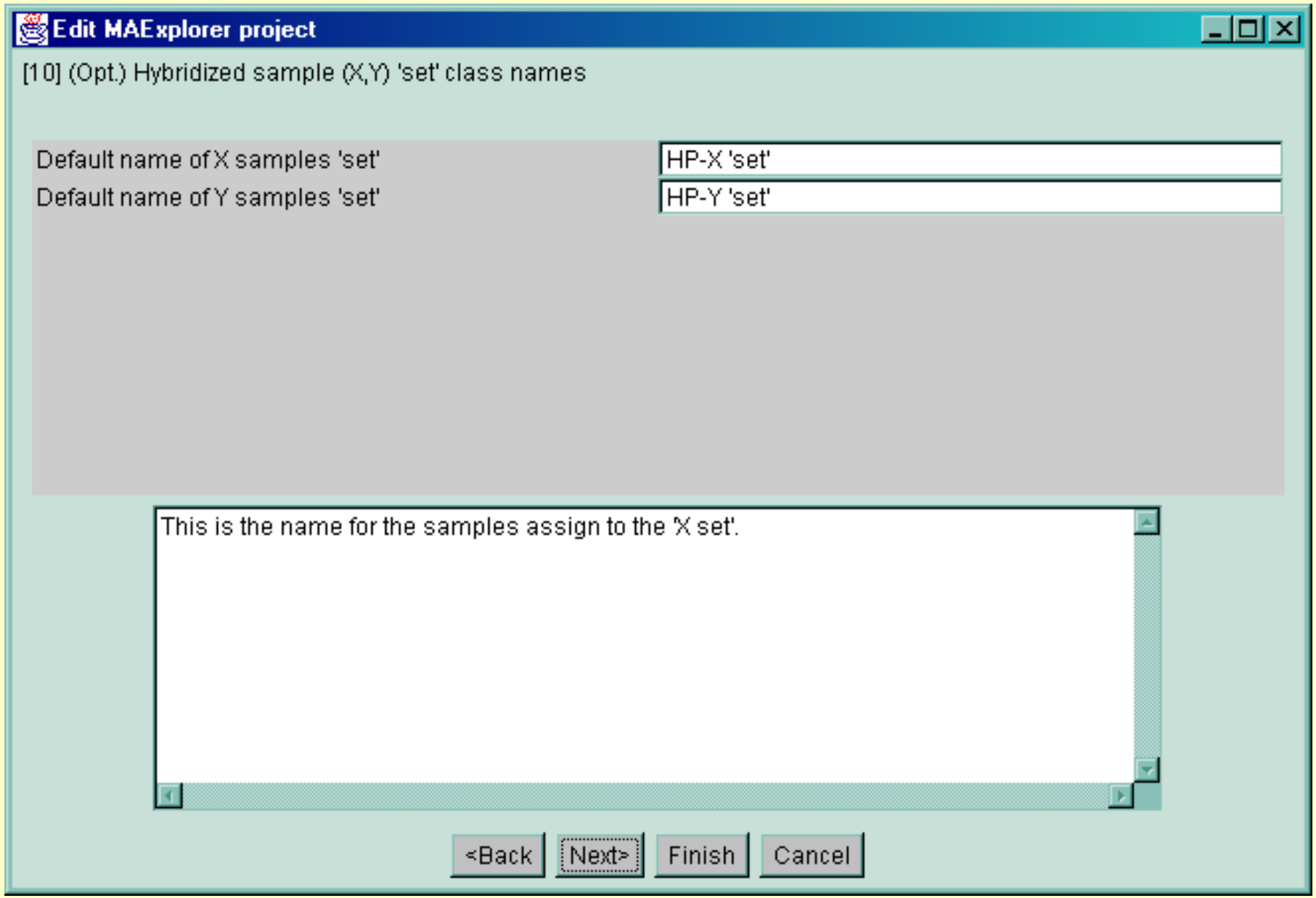
5.8 Edit Layout ‘Wizard’ Calibration Values



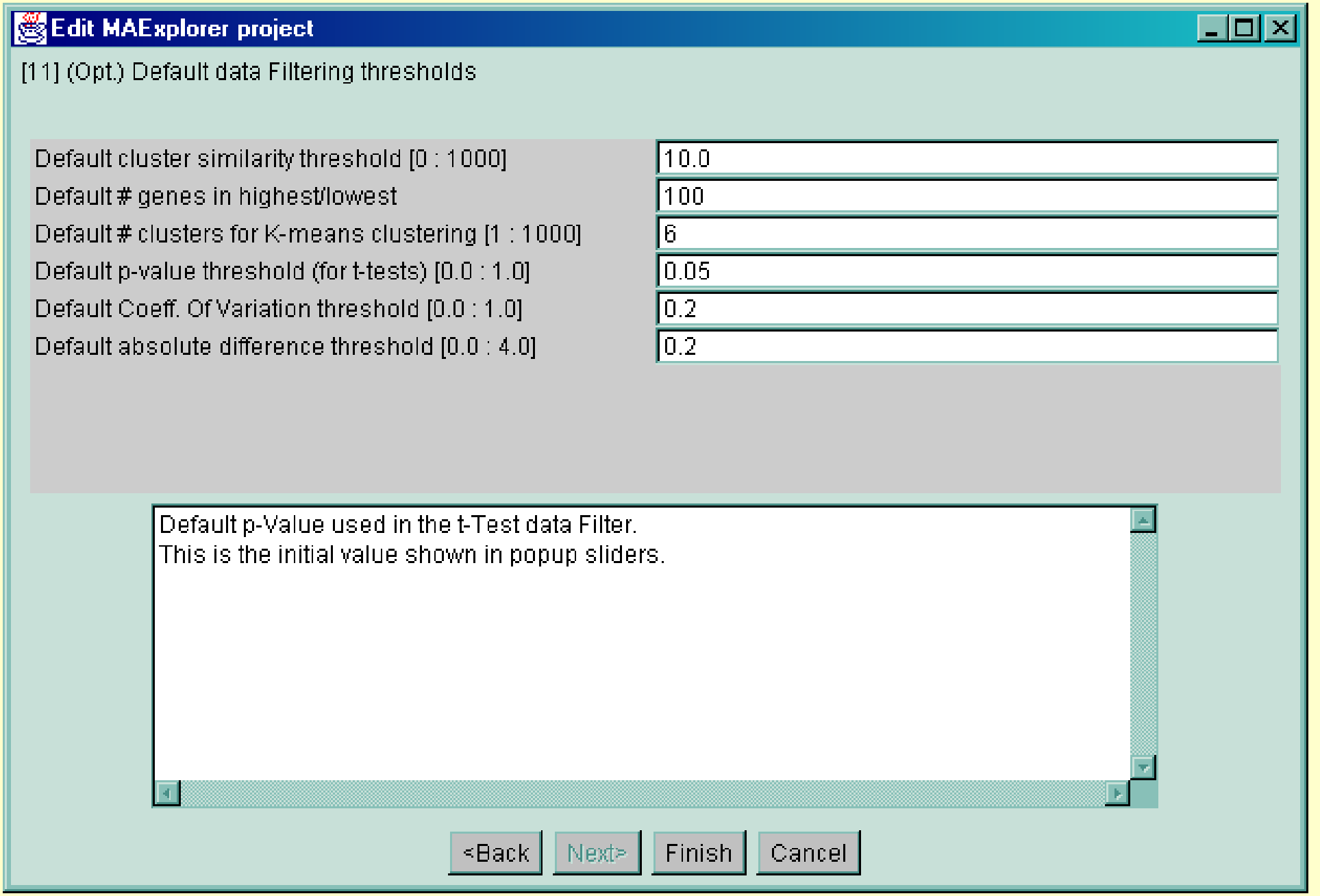
5.9 Edit Layout ‘Wizard’ Database Name Values



5.10 Edit Layout ‘Wizard’ HP-X,-Y Class Names



5.11 Edit Layout 'Wizard' Default Thresholds

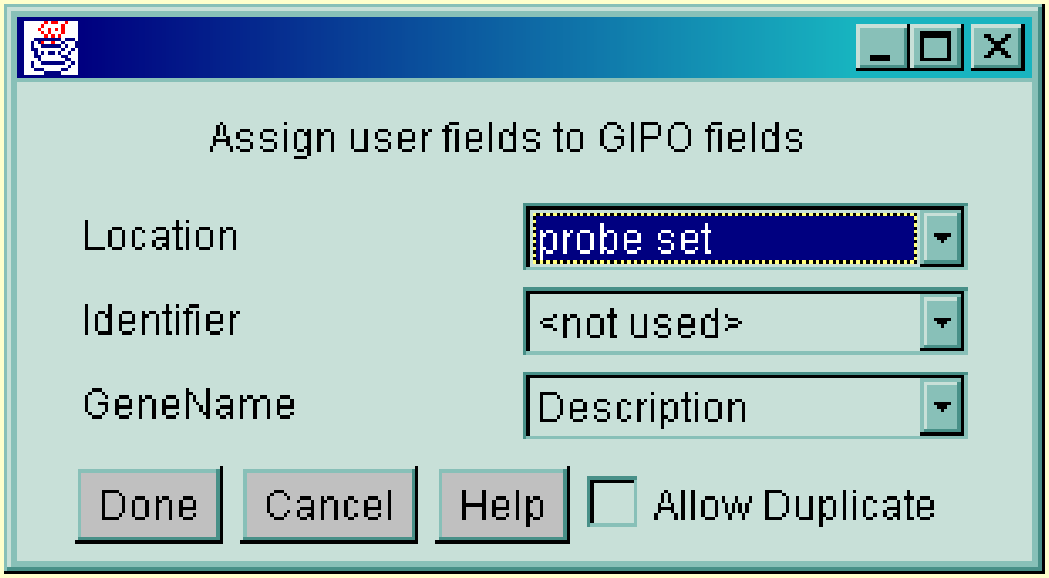


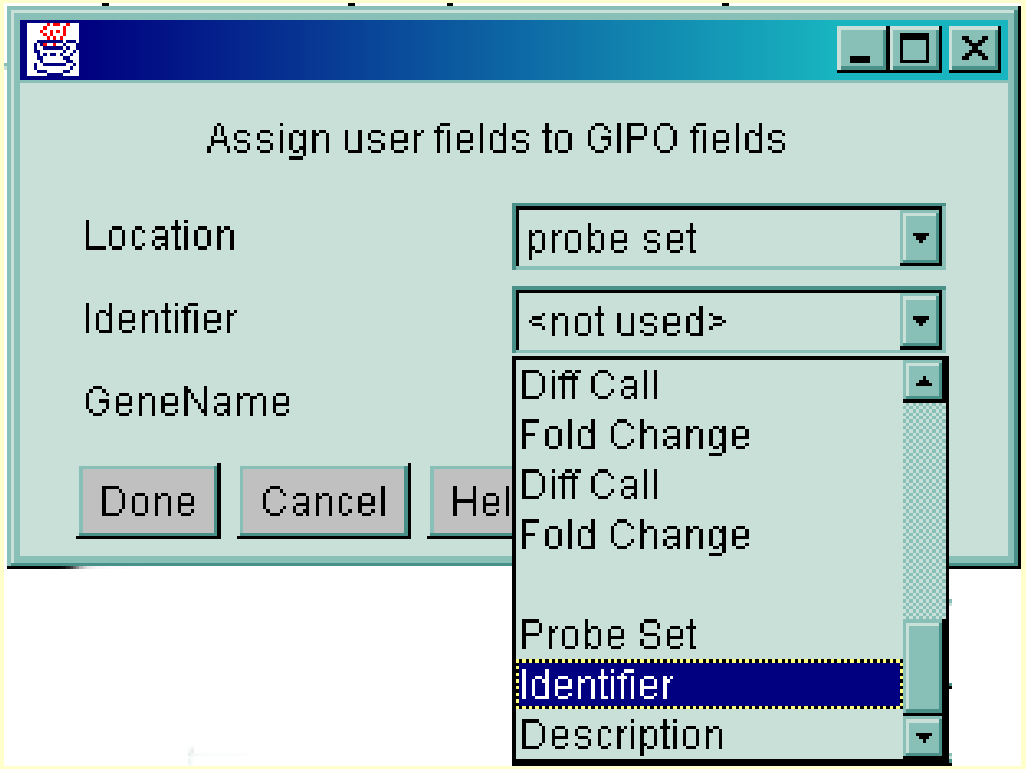
6. Other Options - Assigning User Data Fields to MAExplorer Fields

- GIPO (Gene In Plate Order or "array print" table) - assigns genes to positions on the array as well as GeneBank ID, Clone ID, LocusID (if available), Gene Name, etc.

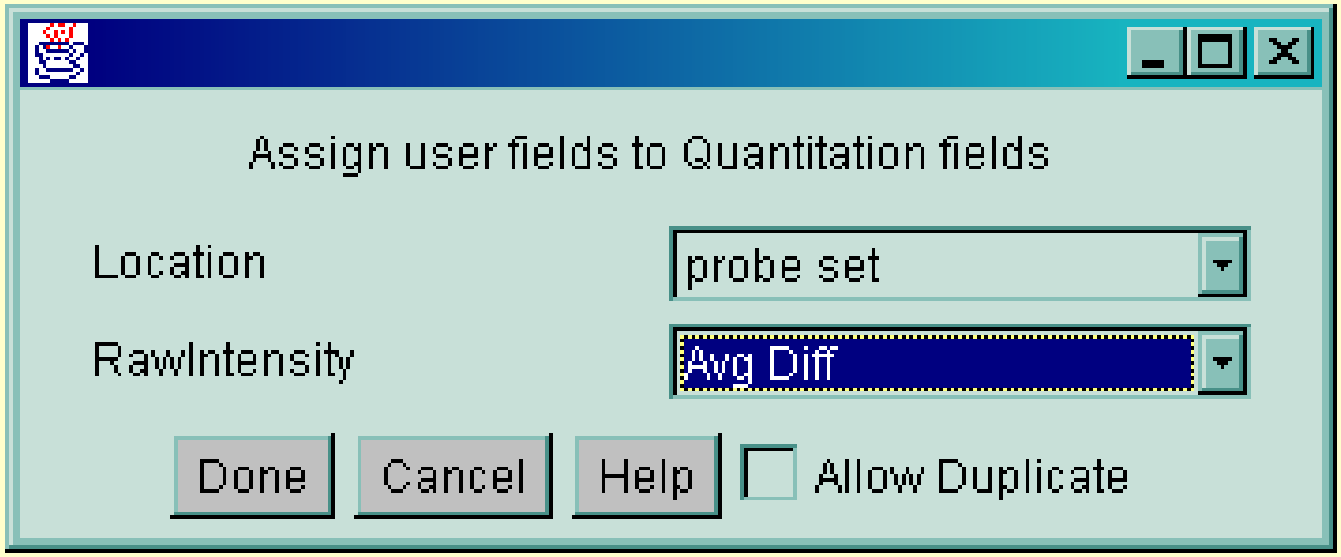
- Quant data - assigns names of quantified data in the user file to MAExplorer data (e.g. Cy3 intensity to RawIntensity1, Cy5 to RawIntensity2, etc).

6.1 “Assign user fields to GIPO fields”

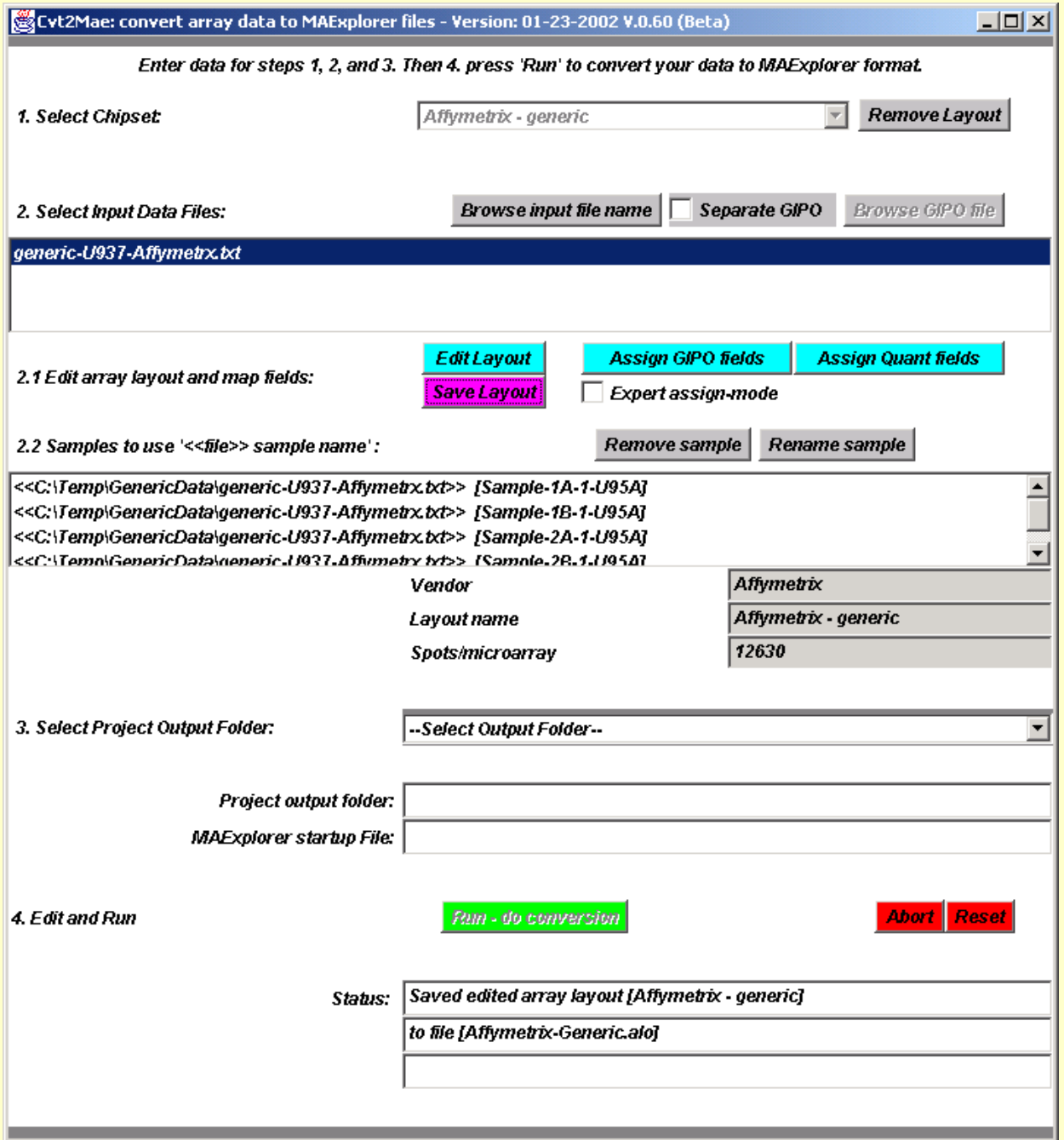




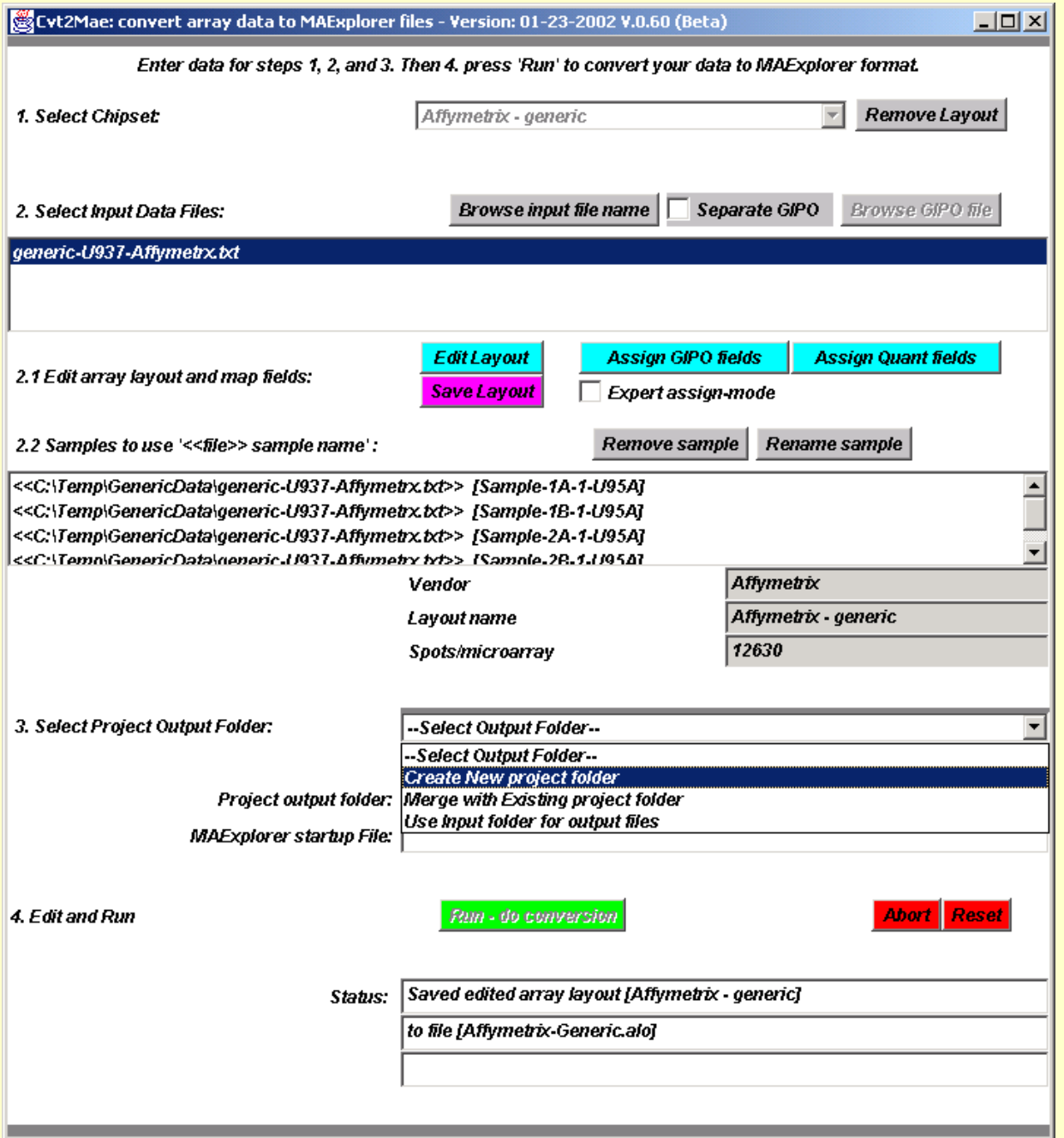
6.2 “Assign user fields to GIPO fields”



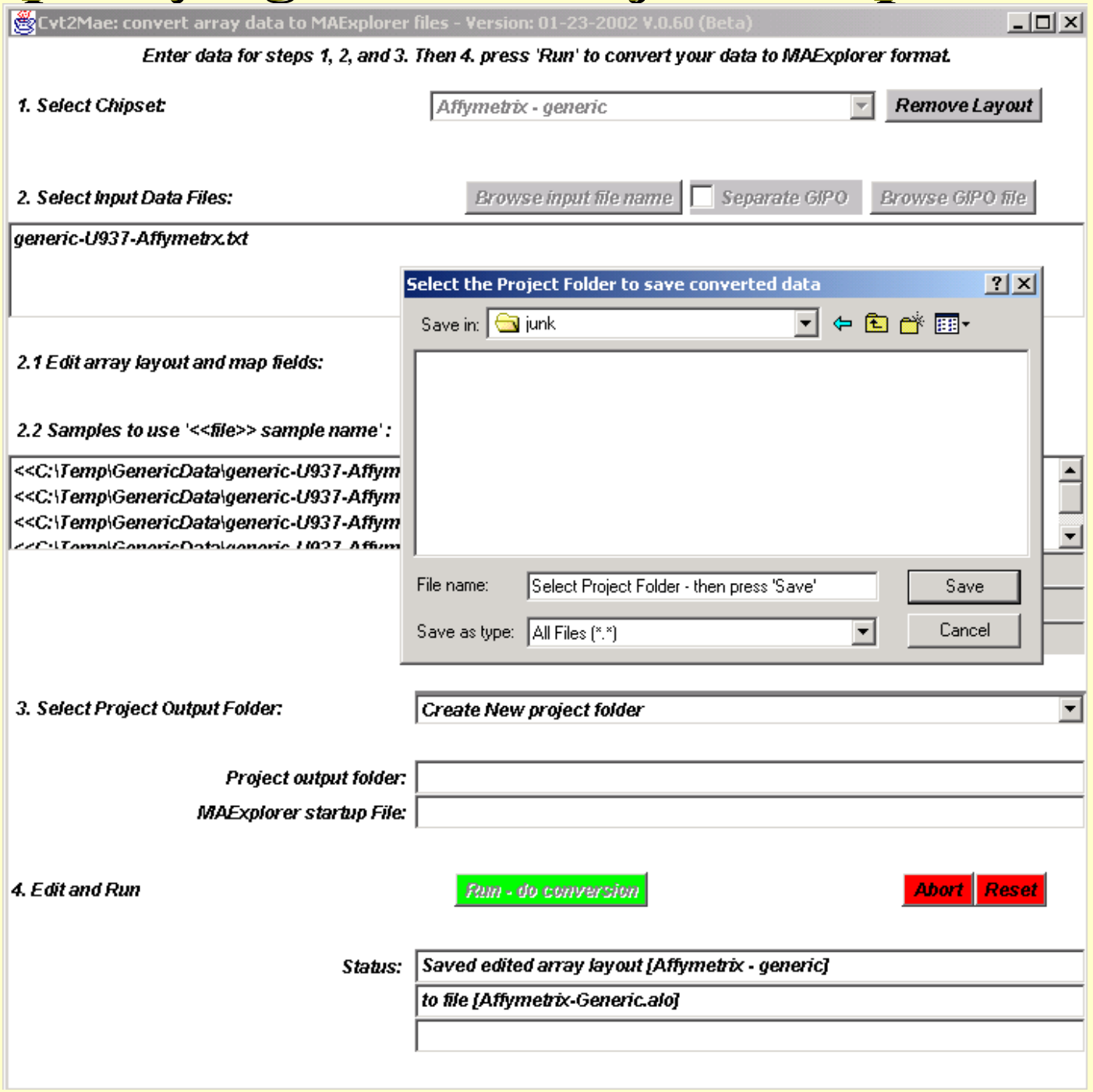
7. Optional "Save Layout" to Array Layout Database After Edit Layout and Assign fields



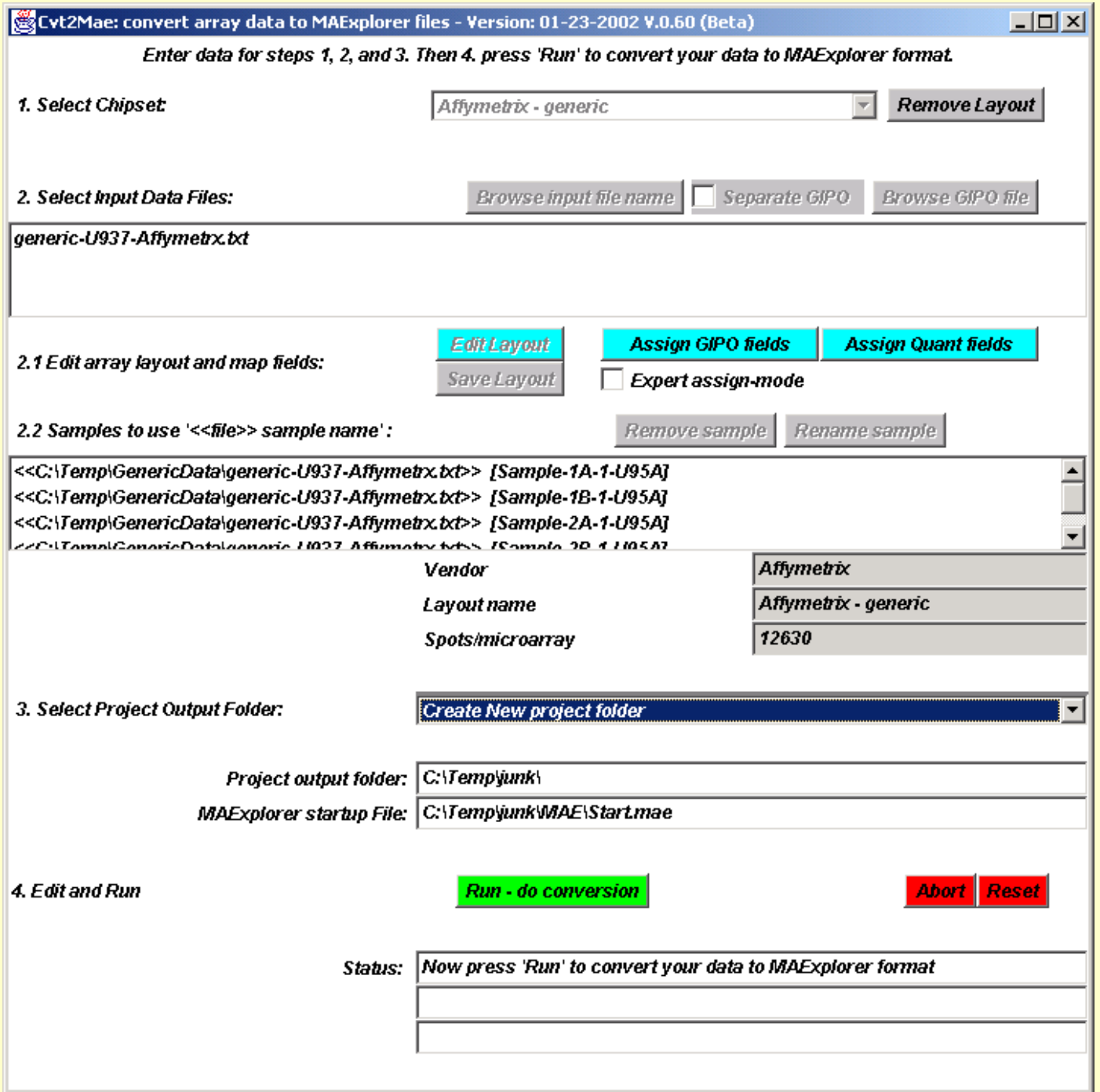
8. Specifying “Create new project folder” Option Where Generated Database Will Be Saved



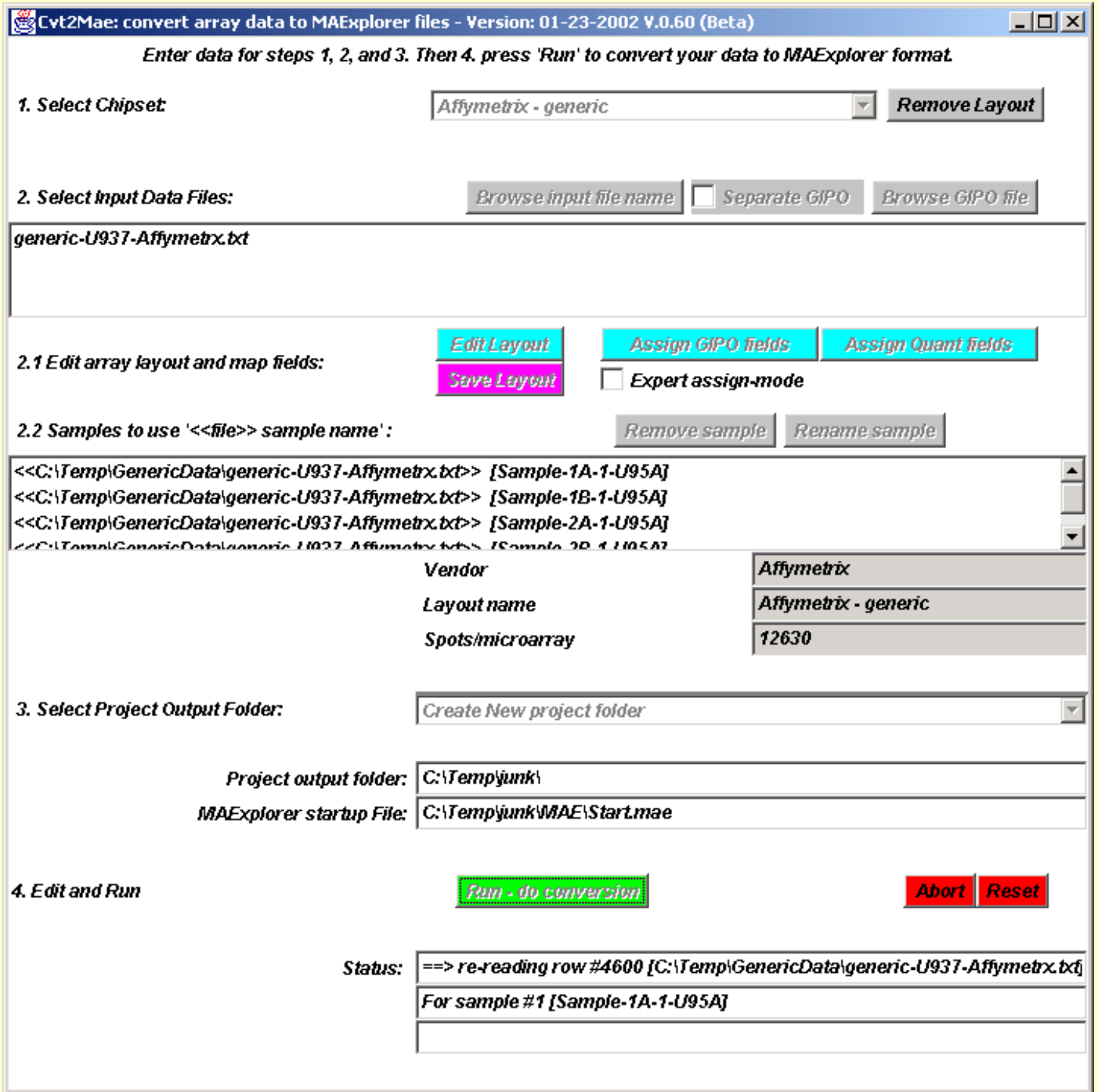
8.1 Specifying New "Project Output Folder"



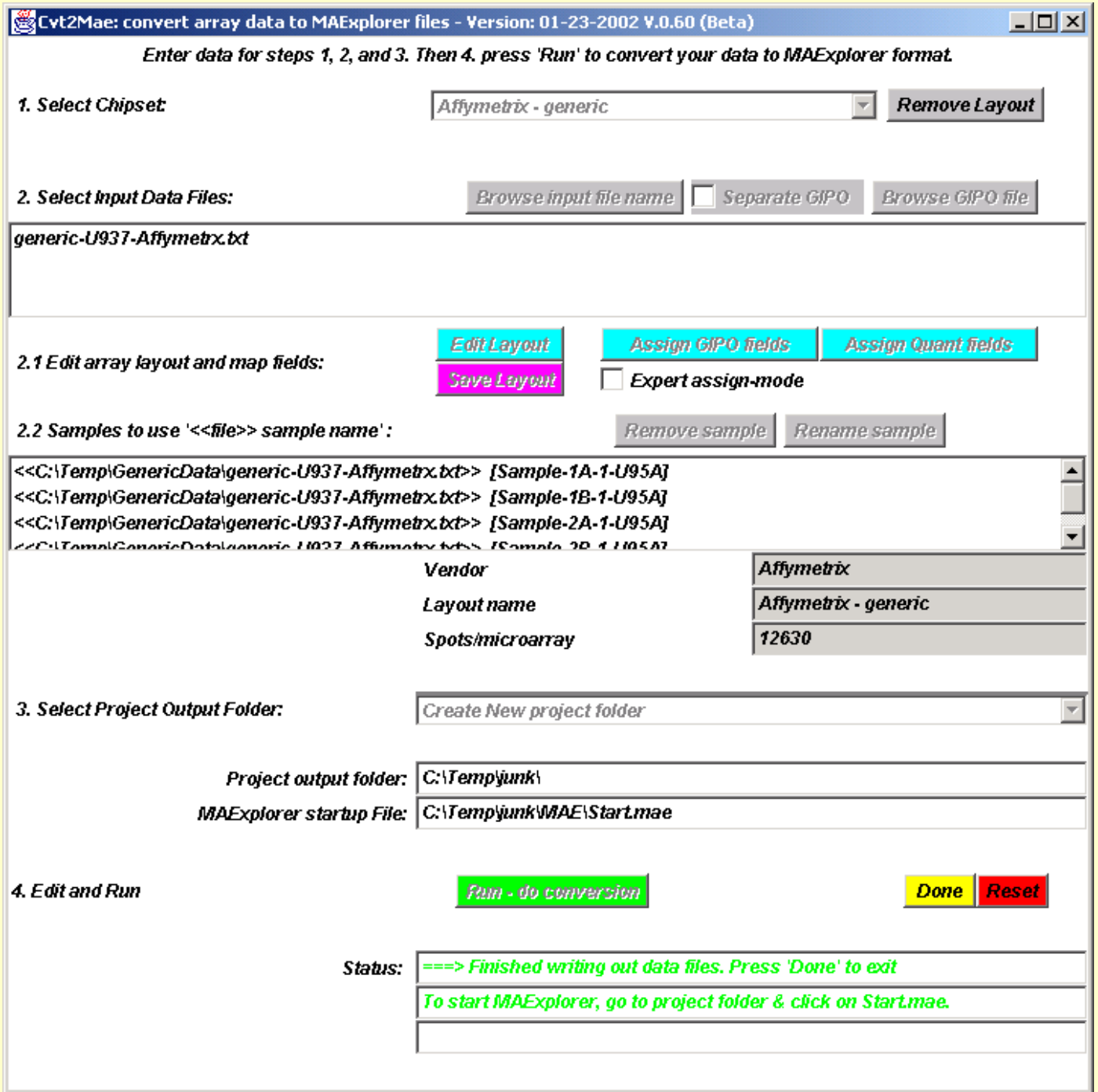
8.2 “Project Output Folder” & MAE startup file



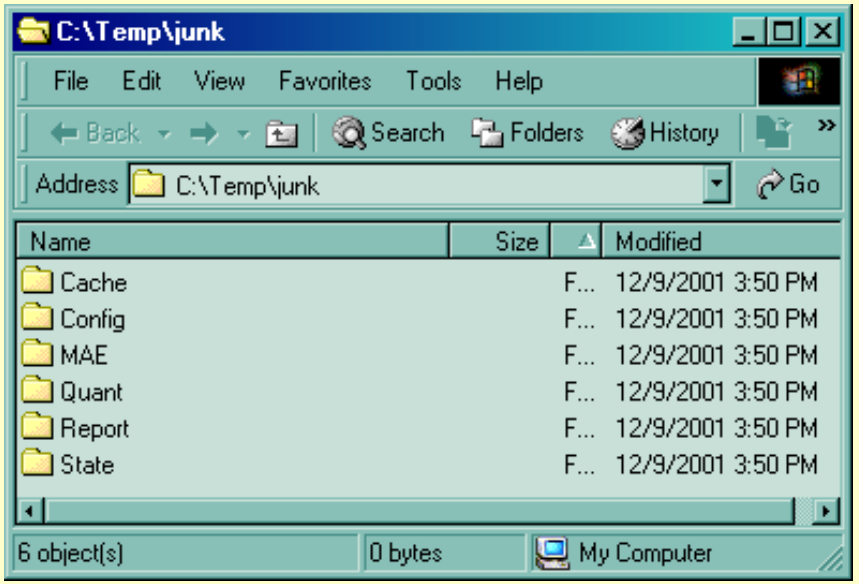
9. Conversion in Process After Pressing "RUN"

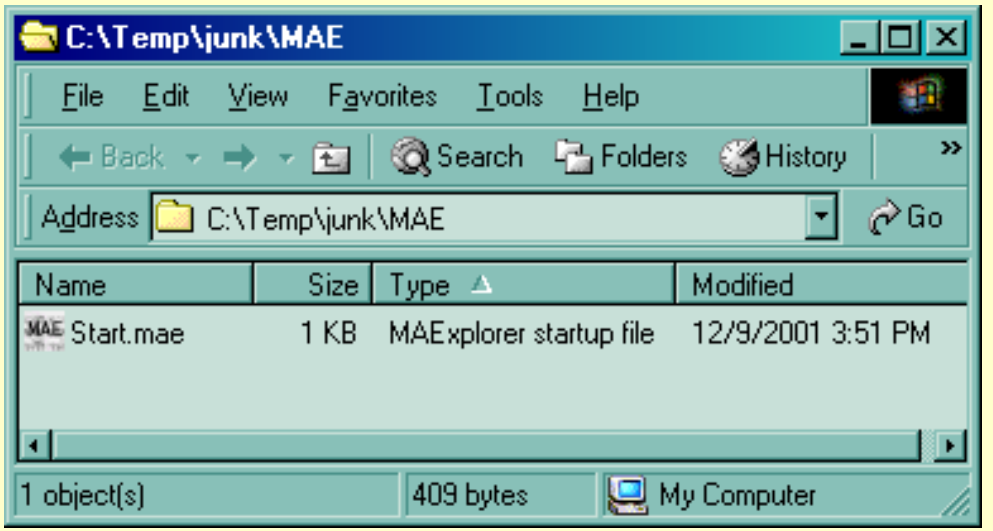


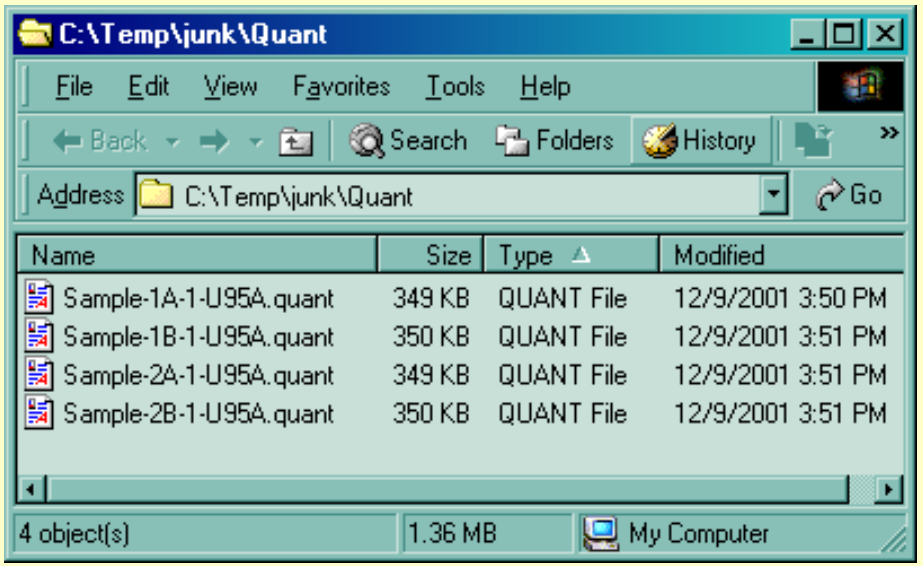
10. Notification that Conversion is Finished

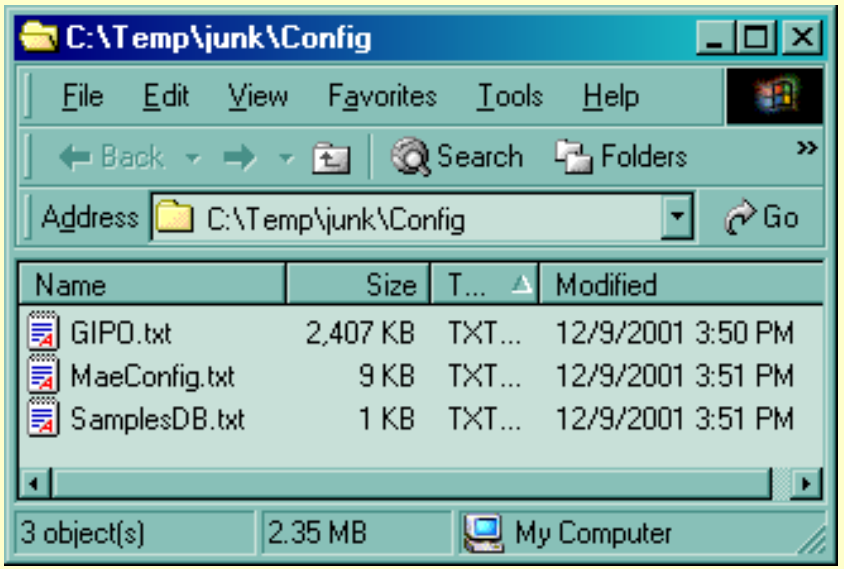


11. MAExplorer Data Created By Cvt2Mae









12. Running MAExplorer on the Converted Data