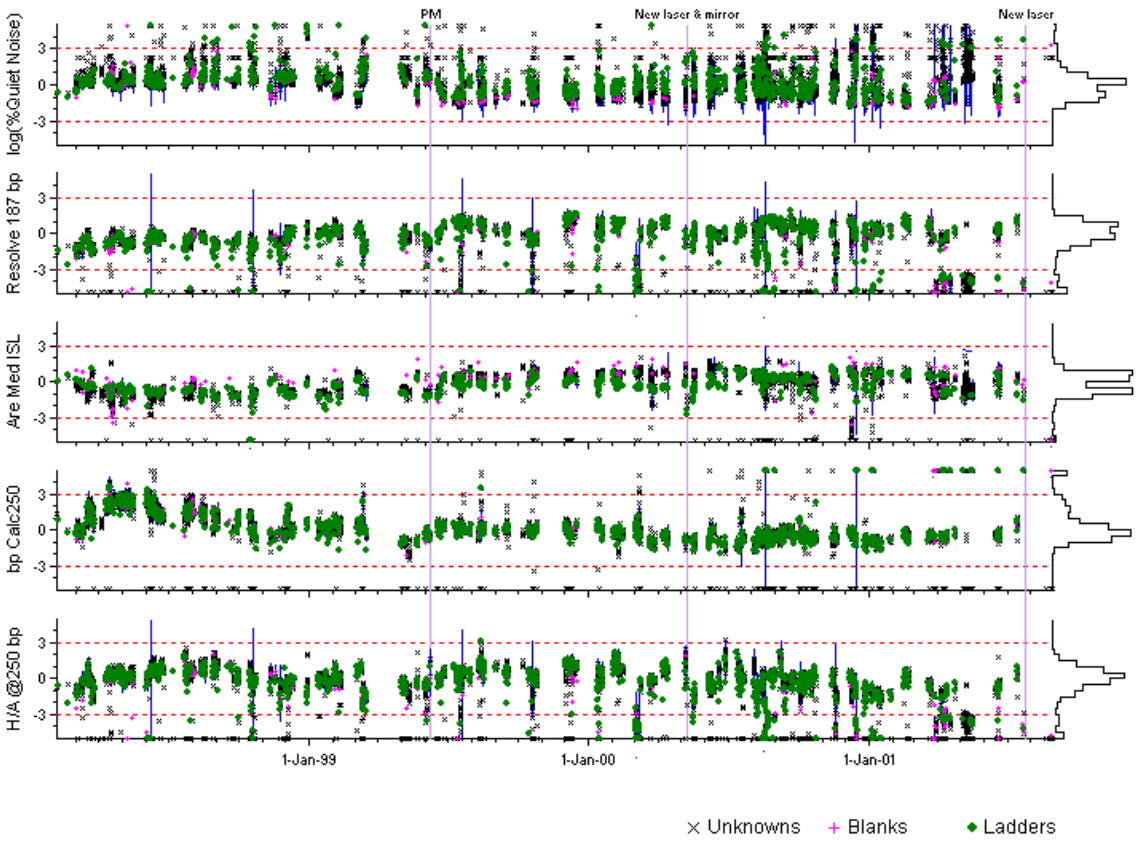
User's Manual
for
MULTIPLEX_QA
Version 10-Sep-2008

An Exploratory Quality Assessment Tool for STR Multiplex Assays

David Lee Duewer

Analytical Chemistry Division
Chemical Science and Technology Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899



DISCLAIMER

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

Points of view in this document are those of the author and do not necessarily represent the official position or policies of the U.S. Department of Justice or the U.S. Department of Commerce.

ABSTRACT

The Multiplex_QA system is a tool for visualizing short- and long-term changes in the quality of capillary electrophoretic analyses. Multiplex_QA does not itself read manufacturer-specific binary data files but rather uses text-formatted equivalents produced by the BatchExtract system developed by Dr. Stephen Sherry's Group at the National Center for Biotechnology Information of the National Library of Medicine. While the BatchExtract system is specifically designed to process the files generated by Applied Biosystem genotyping software, Multiplex_QA is capable of evaluating data from other capillary electrophoretic systems if text file of appropriate content and format become available.

Multiplex_QA uses the information in the text files to estimate quality metrics that capture changes in electropherographic resolution and efficiency. The metrics are mostly based on the behavior of the components of the internal size standard that is mixed with nearly every multiplex analysis. Several different graphical displays enable identifying unusual events over time scales ranging from a single analysis to all analyses performed on a given instrument.

While primarily intended to investigate the utility of the various quality metrics, the current Multiplex_QA system is sufficiently robust for use by forensic scientists with an interest in data analysis and access to a reasonably fast PC. This User's Manual is intended for such users. It focuses on how to operate Multiplex_QA rather than on how to interpret the graphical results that can be produced using Multiplex_QA.

KEYWORDS

Display, Document, Discover (D^3) plots

Electropherograms

Exploratory Data Analysis (EDA)

Internal Size Standard (ISS)

Quality Assessment (QA)

Short Tandem Repeat (STR)

(This page intentionally blank)

GLOSSARY

.fsa	The file extension for the binary data files produced by the various ABI electrophorectic instruments. The .fsa files store all the data associated with the sample, including sample identifiers, the raw electropherogram, and data analysis results. The BatchExtract system extracts selected chunks of these data into text files.
ABI	Applied Biosystems, Foster City, CA (USA)
Active window	The window unto your computer's soul that you use to interact with the Multiplex_QA system. This window should expand to cover your entire display screen shortly after you invoke Multiplex_QA, but you can then adjust it to whatever size you like. Multiplex_QA is designed to use only one Excel window, but you can have as many open at the same time as you like.
Active worksheet	The Multiplex_QA worksheet displayed in the active window. This worksheet will occupy nearly all of the screen area covered by the active window, excepting only the Excel Title and Menu bars at the very top and the Excel worksheet tabs and Status bar at the bottom.
Alert box	A standard Excel pop-up dialog box that provides you with either a warning or a request for action. Warning dialogs typically require that you click a **Yes** button before you can continue. Action dialogs typically require you to click a **Yes** button to continue with an action or a **Cancel** button to forgo the action. Clicking the "Close box" of the dialog window is equivalent to clicking **Cancel**.
Allele	The nucleotide sequence at a specified genetic locus. For STRs, an allele is first characterized by its apparent bp size via calibration to the ISS and then "typed" to a specified number of tandem repeats via calibration to the allelic ladder. The ISS is run as part of each sample; the allelic ladder is run as part of a given set of samples.
Allelic ladder	A set of all commonly encountered alleles for all genetic loci of interest, analyzed periodically within a given set of samples and used to convert allelic bp size to allelic type.
BatchExtract	A program that extracts most information embedded in the .fsa binaries into text files. BatchExtract system was developed in Dr. Stephen Sherry's Group at the National Center for Biotechnology Information of the National Library of Medicine.
bp	Basepair, a unit for expressing the length or "size" of a DNA fragment. Every pair of associated nucleotides in double-stranded DNA represents one bp. The rate of electrophorectic migration along a capillary column is roughly inversely proportional to the size of the fragment.

Cancel button	A command button labeled **Cancel** in many dialog boxes. Clicking **Cancel** will always abort the proposed action and return you from whence you came.
Checkbox	An Excel control device, visually represented as small square hole in the active worksheet, that controls whether a particular property of the system is selected or not selected. To select or deselect the property, click inside the checkbox. When selected, a "✓" fills the hole; when deselected, the hole is empty. When the property is not available, the checkbox is grayed-out (well, it's actually more a "checkered black" than gray.)
Close box	The square button marked with an "×" at the upper right hand corner of a window, used to close the window. It's better if you don't do this when working with Multiplex_QA, but (most of the time) clicking a Close box does the same thing as clicking **Cancel**.
Color	One of the unique dye labels used in an STR multiplex. Multiplexes typically use one color for the ISS and three or four other colors for various genetic loci. Within each color, the alleles are identified by their bp size as calibrated by the ISS peaks.
Command button	An Excel control device, visually represented as a raised rectangular tile on top of the active worksheet. Clicking on a command button invokes a particular Excel action – that is, it starts or stops a program.
Dataset	All of the sample identifier and quality metric data extracted from all .fsa files specified by the user. While the Dataset contains links to the BatchExtract files that were used to calculate the quality metrics, these text files are not themselves considered to be part of the Dataset.
Datetime	The calendar date plus the time of day. Datetimes are intended to be displayed in one of Excel date and/or time formats (e.g., 28-Jun-05 12:01). They are actually stored as the integer number of days since 1-Jan-1900 plus the decimal fraction of the day since midnight.
Dialog box	Any of the standard Excel pop-up windows used to request user interaction with Multiplex_QA or with Excel itself.
Excel	A "spreadsheet" data analysis and modeling product of Microsoft Corporation, Redman, WA. Multiplex_QA runs inside the Excel product. It has been tested on Excel 2000 on several OS Windows 2000 machines.
Folder-path	The *complete* textual description of where a particular file directory (folder) is located on your computer.
Formula bar	An Excel tool located along the top edge of the Excel window displaying the location and contents of the active cell. Multiplex_QA deactivates the display of this tool.
GS	One of ABI's GeneScan family of internal sizing standards.
ILS	One of Promega Corporation's Internal Lane Standard family of internal sizing standards.

Input box	A standard Excel pop-up dialog box that requests you to respond to a question more complex than something requiring just a "yes/no" decision. The input dialog will typically specify what is needed and the valid range of input values, as well as specifying a default value. If you do not want to proceed with the action that prompted the Input dialog, click the dialog box's **Cancel** button.
IQRe	The IQRe is a robust estimate of dispersion, It estimates the standard deviation of a group of data from the InterQuartile Range (IQR) and the known value of IQR for a standard normal distribution [1]. The IQR is the interval that contains the central 50% of a distribution of values.
ISS	Internal Size Standard, a mixture of DNA fragments of known bp size labeled with a distinctive dye that is mixed with every sample after amplification but before injection. The association of known peak elution time with bp size enables the typing of unknown alleles.
Kit	A specific STR multiplex.
Macro, VBA	An Excel function written in the Visual Basic language. Only the Auto_Open and Auto_Close macros used when you first open the Multiplex_QA system and when you finally leave it are written in Excel's VBA macro language.
Macro, XLM	An Excel function written in the Excel 95 macro language. Most of the Multiplex_QA functions are written in the Excel's XLM macro language.
Median	A robust estimate of location. The median is the value that is "in the middle of" a dataset. If there's an odd-number of values in the group, the median is the value in the middle of the rank-ordered list. If there's an even number, the median is the average of the middle two values.
Multiplex	A defined set of co-amplifiable PCR primers for two or more STR loci. The alleles for the various STR loci are identified on the basis of dye color and bp size.
Open box	A standard Excel pop-up dialog box that asks that you specify a file to be opened. There is no dialog *per se*, but a default directory, type of file, and all files in the default directory that match a particular wildcard search are specified in various input fields within the window. You can change the directory, the file type, and the wildcard search by using the standard navigation tools provided at the top and left side of the Open box. If you decide that you do not want to open any file, click the box's **Cancel** button.
Optionbuttons	Excel control devices, visually represented as small circular holes in the active worksheet, that control which single property of a list of properties of the system is active. To select a property, click inside the optionbutton associated with the desired option. When selected, the center of the optionbutton will be dark; when deselected, the center is empty. When one optionbutton of the series is selected, all the rest will be deselected.
PCR	Polymerase chain reaction.

Peak	The signal for an allele or ISS component, typically modeled as a symmetric Gaussian centered at the expected bp size of the allele with a dispersion that is a function of the bp size, various electrophorectic parameters, and the state of the column or gel.
Plate	A group of samples PCR-amplified at the same time in the same amplification plate. The current Multiplex_QA system only recognizes 96-well plates.
Promega	Promega Corporation, Madison, WI (USA)
R_{75}	The R_{75} is a robust estimate of bivariate correlation. It estimates the correlation of the entire dataset from the correlation between the central 75% of the samples.
Radio buttons	See "Optionbuttons".
rfu	"Relative fluorescence unit", the traditional intensity metric used for fluorometric signal detection systems.
Robust statistics	Summary estimates of various properties of a set of data that provide "reasonably good" estimates of those properties even when a small percentage of the data are "flawed" – representing glitches, incorrectly specified analysis parameters, or other artifacts rather than the desired information. Three robust statistical estimates are used in the Multiplex_QA system: the median as an estimate of central location, the IQRe as an estimate of dispersion about the central location, and the R_{75} as an estimate of bivariate correlation.
Sample	Any material loaded and analyzed with an STR multiplex kit as a separate entity. Multiplex_QA recognizes three sample types: blanks, ladders, and unknowns.
Save As box	A standard Excel pop-up dialog box that requests you to specify the name and directory for a file that is to hold about-to-be-stored information. There is no dialog *per se*, but a default directory, default file type, default file name, and all files in the default directory of the same type as the default type are specified in various input fields within the window. You can change the directory, the file type, and the file name using the standard navigation tools provided at the top and left side of the Open box. If you decide that you do not want to save the information, click the box's **Cancel** button.
Scroll bar	A standard Excel control device that allows you to specify a value using the mouse rather than typing characters. Scroll bars are visually represented as light grey rectangular strips with a solid triangle at either-end and a small dark gray box in between. The value specified by the scroll bar is proportional to the distance the central box is along the entire length of the scroll bar. The central box can be selected and moved along the scroll bar with the mouse. Clicking on the endpoint triangles moves the central box one unit in the direction in which the triangle is pointing;

	clicking on the space between a triangle and the central box moves the box a number of units in that direction.
Set	A group of samples run sequentially on a electropherographic instrument.
Sheets	See "Worksheet".
Status bar	An Excel tool located along the bottom edge of the Excel window displaying non-critical – but often very informative – messages from Multiplex_QA and/or Excel itself.
STR	Short Tandem Repeat.
Toolbars	Excel tools typically located along the top or bottom edge of the Excel window providing user access to system-level functions. Multiplex_QA deactivates the display of these tools.
Typing	Process for converting measured electrophorectic migration times into the conventional names for specific alleles at given genetic loci.
Workbook	In Microsoft Excel, a workbook is the file in which you work and store your data. Because each workbook can contain many sheets, you can organize various kinds of related information in a single file. The Multiplex_QA system is an Excel workbook.
Worksheet	A major structure of an Excel workbook that can store data, charts, and various control devices.
Worksheet tabs	The names of the worksheets appear on tabs at the bottom of the workbook window. To move from sheet to sheet, click the worksheet tabs.

User's Manual
for
MULTIPLEX_QA
An Exploratory Quality Assessment Tool for STR Multiplex Assays

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

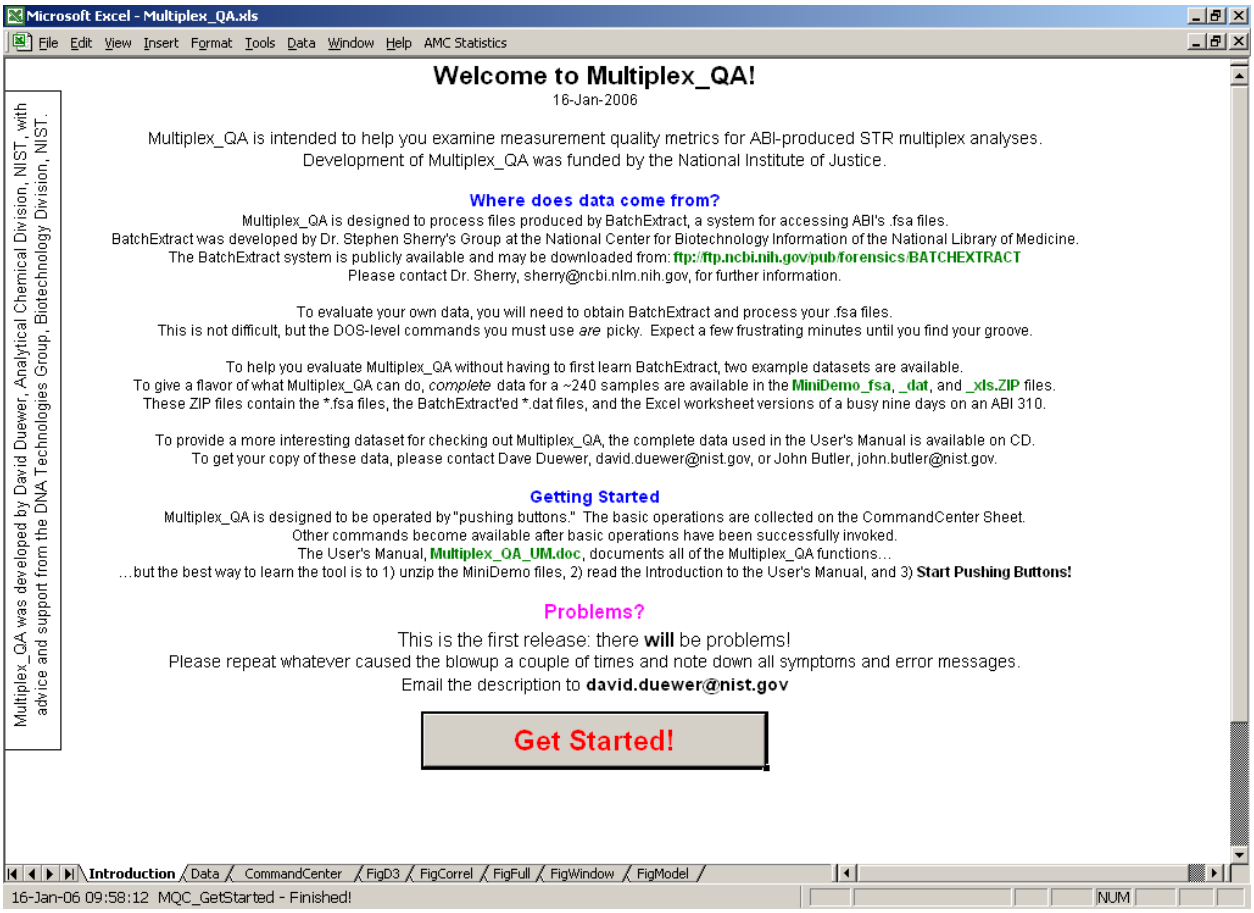
1 INTRODUCTION

Multiplex_QA is an Excel-based system for assessing capillary electrophoresis Short-Tandem Repeat (STR) measurement quality. Multiplex_QA enables the calculation and graphical evaluation of signal intensity, peak symmetry, electrophoretic resolution, and signal noise parameters. Multiplex_QA is operated by "clicking buttons." This User's manual documents the assumptions and calculations embedded in the Multiplex_QA system as well as describing the system's capabilities.

Multiplex_QA does *not* itself interpret binary data files that hold the original STR analysis data but rather processes text files produced by BatchExtract, a system developed by Dr. Stephen Sherry's Group at the National Center for Biotechnology Information of the National Library of Medicine, National Institutes of Health. While BatchExtract currently only interprets the .fsa binary files produced by Applied Biosystems (ABI) systems, Multiplex_QA can be used with non-ABI systems if the data from those systems can be extracted and stored in BatchExtract-style text files. Section 9 details how to use BatchExtract and the formats of the required files.

A demonstration dataset is provided along with the Multiplex_QA system. The best way of determining whether Multiplex_QA is of use to you is to… start clicking! The place to start is at the bottom of the Introduction worksheet.

Figure 1. Introduction Worksheet



1.1 Getting Started

Before you can start using Multiplex_QA, you should copy it from the CD or the www.cstl.nist.gov/biotech/strbase/software.htm website to your hard-disk.

1.1.1 Where to locate Multiplex_QA

Anyplace you want to. However, keeping track of things will be easier if you store both Multiplex_QA.xls (the tool) and Multiplex_QA_UM.doc (this manual) in a "Multiplex_QA" folder located in some easily remembered hard-drive location on your computer.

1.1.2 Checking and changing the "Read-only" file attribute

If the Multiplex_QA.xls "Read-only" attribute is enabled, you will not be able to conveniently save work you do with the Multiplex_QA tool. You can check what the Read-only setting is by: 1) viewing the contents of the directory where you have stored Multiplex_QA.xls, 2) activating (single-clicking) the Multiplex_QA.xls file, and 3) locating the "Read-only" checkbox at the bottom left of the "General" sheet of the "Properties" option of the Files" menu (in geek shorthand, "Files > Properties > General: Read-only"). If the checkbox is "checked", uncheck it. Click the **OK** button when finished.

Alternatively, rename your original copy of Multiplex_QA.xls to something like "Multiplex_QA_Original.xls," open it, and use the Excel **Save As** command (in the File menu) to save a working copy with the name "Multiplex_QA.xls." You will be graded on your spelling: the tool will not work under any other name.

1.1.3 Enabling macros

If you encounter the message "This workbook contains a type of macro (Microsoft Excel Version 4.0 macro) that cannot be disabled or signed. Therefore, this workbook cannot be opened under High Security Level." when you first try to open (double-clicking) the Multiplex_QA.xls workbook, you need to set the Security Level to "Medium." You do this on the "Security Level" sheet of the "Security" function of the "Macro" option of the "Tools" menu of the main Excel widow ("Tools > Macro > Security > Security Level: Medium"). After you reset the Security Level, exit from Excel and open Multiplex_QA.xls again. You should now see a message warning you that Multiplex_QA contains macros and giving you the option of either disabling them or enabling them. Click the **Enable Macros** button and you can begin!

1.1.4 Updates

Both Multiplex_QA and this User's Manual are occasionally updated, so you may want to occasionally visit the www.cstl.nist.gov/biotech/strbase/software.htm website and check if the Version Date has changed. It is strongly recommended that you save a copy of any "old" version of Multiplex_QA that you've used and **not** just over-write it with the new.

1.2 Excel's "Country Version" Parameters

The Excel system has a number of "Country Version"-specific words and formats that I haven't (yet) figured out how to automatically figure out. *If and only if* you are using a version of Excel that has not yet been parameterized (currently, only English and – coming soon – Danish), you will need to specify values for about a dozen parameters before you can proceed. A tool is provided to help you determine what the necessary values are. Please see Section 2.7.2 for a description of this tool.

1.3 The Introduction Worksheet

The Introduction worksheet should always become the active worksheet whenever Multiplex_QA is invoked. It is intended to remind the user of the system's origins. The first couple of times you use the tool, read the Introduction!

1.3.1 Turning off the automatic invocation of the Introduction worksheet

Startup invocation of the Introduction worksheet can be turned off by deleting, renaming, or editing the VBA macro "Auto_Open." Until you know enough about Excel to do this with confidence, you should leave well enough alone.

1.3.2 The "Get Started!" command button

Clicking **Get Started!** initializes a standard viewing environment for Multiplex_QA worksheets and activates the CommandCenter worksheet. All other Multiplex_QA functions are accessed directly or indirectly from the CommandCenter. There is no particular reason to revisit the Introduction worksheet during a given Multiplex_QA session.

1.3.2.1 Standard display

To display as much information as possible and to remove visual clutter not typically needed during operation, Multiplex_QA suppresses display of most Toolbars, the Formula Bar, and the row/column labels. Gridlines are turned off on the Introduction, CommandCenter, and the graphical display worksheets.

1.3.2.2 Window size and placement

While the size of the active worksheet is maximized within the Excel window, Multiplex_QA does not itself adjust the size and placement of the Excel window within the display screen. If you wish to change the location or size of the Excel window, you need to do so manually.

1.3.2.3 Zoom

The default magnification or "zoom" factor for all worksheets is 75%. However, the most appropriate magnification depends upon the screen resolution setting and the physical size of your display screen. With low-resolution displays, the default zoom may not allow you to see the **Get Started!** button without scrolling. You can manually reset both the default zoom factor

and the magnification for all Multiplex_QA worksheets using the **Zoom** button on the CommandCenter worksheet.

1.3.3 What if you dislike the standard display?

If you prefer to set your own viewing environment, *don't* click **Get Started!**. Just activate the CommandCenter worksheet by clicking its tab. The **Exit** button on the CommandCenter worksheet can also be used to restore your original environment.

1.3.4 History

A rough history of the Multiplex_QA system is recorded on the Introduction worksheet, just below the **Get Started!** Button.

1.4 Moving From Sheet to Sheet

The "natural" way to navigate among Excel worksheets is by clicking on the tabs at the bottom of the active window. However, particularly when you are unfamiliar with the Multiplex_QA system, it may be easier and more effective to use the **Back** and **Return to CommandCenter** buttons provided on many worksheets.

1.4.1 Back

Clicking a **Back** button activates the worksheet that was used to activate the current worksheet. While combining tab-clicking and button-clicking navigation modes does no harm, it may cause unexpected worksheets to become active when **Back** is clicked.

1.4.2 Return to CommandCenter

All worksheets that have a **Back** button also have a **Return to CommandCenter** button. Clicking **Return to CommandCenter** activates the CommandCenter worksheet.

1.5 Communication Tools

1.5.1 Messages in the Status bar

Multiplex_QA uses the Excel Status Bar to keep you informed of what, and sometimes what is not, happening. The message is updated at the start and completion of every major function. When involved in particularly time-consuming tasks, the message is generally updated when major subtasks are complete. When there has been a problem, the Status Bar will display a brief summary of what happened.

Excel's Status Bar is located at the very bottom of the Excel window. In Figure 1, the message is a date and the words "Welcome to Multiplex_QA!"

The Multiplex_QA Status bar messages are purely informative. None of them require any action from you.

1.5.2 Alerts

1.5.2.1 Notification alerts

When there is only one valid response to a situation requiring you to take action, Alert windows are issued that have only one possible response: **OK**. The essence of these Alerts is typically "You can't do that without doing this first." They are intended to grab your attention and force you to take some remedial action. You must click **OK** (or the Alert window's Close box) to regain control.

1.5.2.2 Choice-forcing alerts

When there are only two (sometimes three) valid actions that can be taken, Alert windows that have **Yes**, **No**, and/or **Cancel** buttons are used. A typical choice-forcing Alert is something like "Do you really want to do this?" When the choices aren't obvious, the Alert will briefly describe the actions taken on selection of each choice. Closing the Alert window using the window's Close box is equivalent to clicking **Cancel**.

1.5.3 Input prompts

When there are more than three possible responses to a particular question, an Input window is used. These windows ask a question, have in input-box for you to type a value, and present **Yes** and **Cancel** buttons. Clicking **Yes** accepts the typed value or, when nothing was typed into the input-box, some default value. Clicking **Cancel** will, depending on circumstances, either accept whatever the current value may be, put you into a "You must really respond to this in an appropriate manner" loop, or cause things to come to a screeching halt.

Responses that are of the wrong type (such as a word when a number is needed) just get an Excel-system level "Invalid Response" Alert that you must clear by clicking **OK** before you can proceed. Responses that are of the right type but outside an acceptable range just re-issue the Input Prompt. When in doubt, read the Input window question text carefully: it (should) specify the nature and range of the required responses.

1.6 Hardcopy

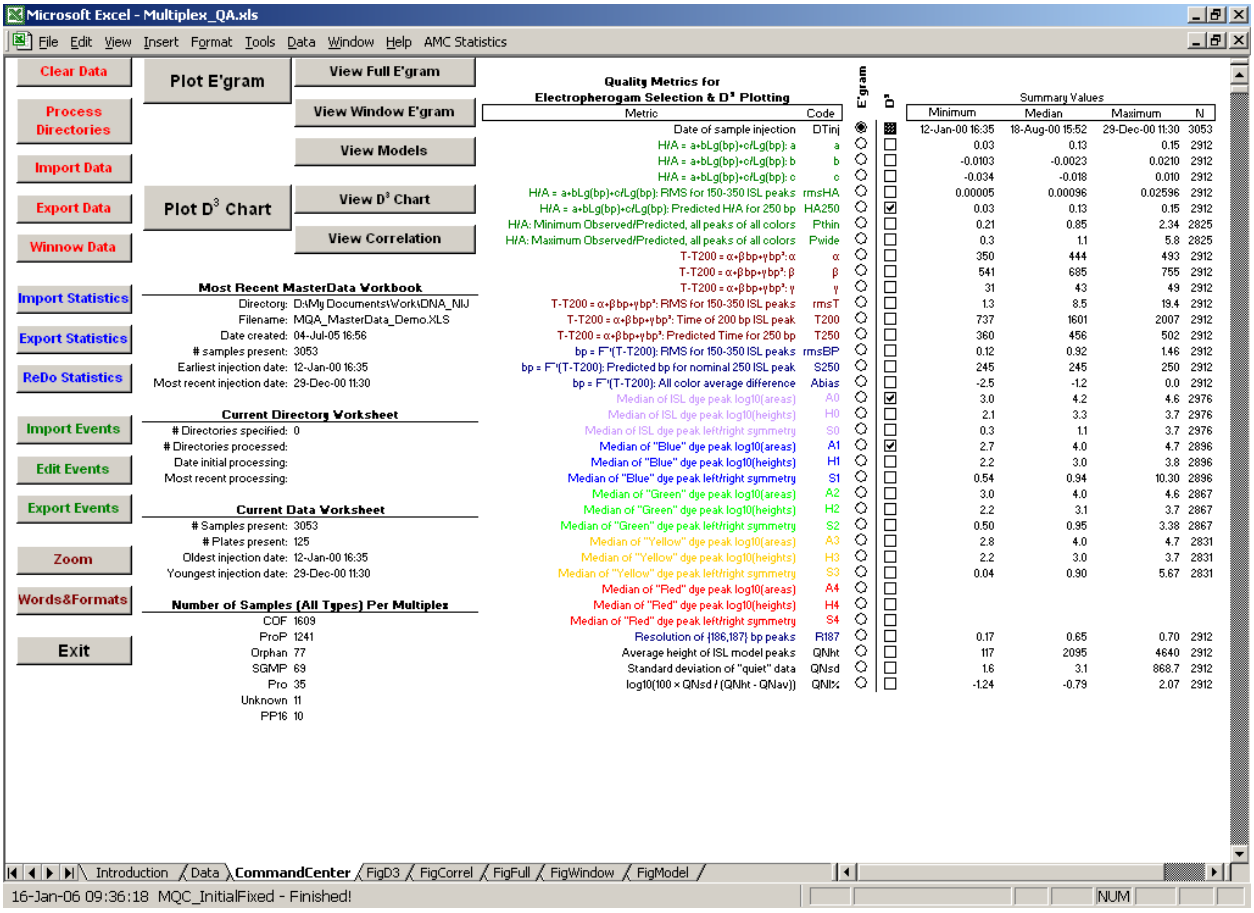
There are no special "print this page" functions in the Multiplex_QA system. All worksheets that store interesting information and/or graphical displays are designed to be printed using the standard Excel **Print** command, accessible either from the File menu or its keyboard shortcut (<Ctrl+P>, if you are using a Windows operating system).

The Figures in this document are screen captures, using the "Print Screen" key on a Windows 2000 machine.

2 THE COMMANDCENTER

The CommandCenter worksheet provides direct or indirect access to nearly all of Multiplex_QA's capabilities. The behavior of the **Plot E'gram** function depends on which of the **E'gram** optionbuttons on this worksheet is active. Likewise, the behavior of **Plot D³ Chart** depends on which of the **D³** checkboxes is/are active. In addition to these controls, the CommandCenter displays several blocks of summary information relevant to a current dataset.

Figure 2. CommandCenter Worksheet



2.1 The Command Buttons

All of the CommandCenter's functions are invoked (you guessed it) by clicking buttons. All of the buttons (\that are intended for your use are located to the left and upper left of the CommandCenter worksheet. Some of the Multiplex_QA functions are fairly simple and single-purpose while others are quite complex. Table 1 lists all of the (reasonably well-developed) Multiplex_QA function(s). All Multiplex_QA functions are described in detail in later sections of this manual.

Table 1, CommandCenter Command Buttons

Button	Function
Clear Old Data	Completely removes all currently defined data from the Multiplex_QA worksheets.
Process Directories	Processes all BatchExtract files contained in the directories specified on the worksheet Directory and adds them to the data stored on worksheet Data.
Import Data	Completely removes all currently defined data and imports a dataset previously defined using Process Directories and saved as a MasterData workbook.
Export Data	Saves the currently defined data as a user-named MasterData workbook in a user-specified directory.
Winnow Data	Allows selective deletion of data by date, average intensity of ISS components, type of sample, and the imputed type of multiplex kit.
Import Statistics	Replaces the current location and dispersion estimates with estimates from a previously saved MasterStat workbook. The location and dispersion estimates of all quality metrics are used to center and scale data displayed in the D^3 Chart functions.
Export Statistics	Saves the location and dispersion estimates calculated from the current dataset in a user-named file in a user-specified directory.
ReDo Statistics	Replaces the current location and dispersion estimates with values calculated from the current dataset.
Import Events	Replaces the current list of the dates on which potentially "interesting" events occurred with events from a previously saved MasterEvent workbook. "Interesting" is here defined as something that may have impact on data quality such as column or laser replacement.
Edit Events	Allows user to edit (add, delete, and/or modify) the current list of potentially "interesting" events.
Export Events	Saves the list of "interesting" events in a user-named MasterEvent workbook in a user-specified directory.
Zoom	Allows user to specify the "zoom" or effective display size of all Multiplex_QA worksheets.
Words&Formats	Allows user to customize date and time formats
Exit	Allows user to gracefully exit the Multiplex_QA system. Can also be used to re-establish the user's standard Excel display environment without exiting Multiplex_QA.

Button	Function
Plot E'gram	Allows user to display electropherograms. The order in which electropherograms are displayed is specified by the setting of the CommandCenter **E'gram** optionbuttons (Section 2.2.2). Once the first electropherogram has been displayed, a number of options become available. These options are described in Section 3.
View Full E'gram	If currently valid, activates the FigFull worksheet.
View Window E'gram	If currently valid, activates the FigWindow worksheet.
View Models	If currently valid, activates the FigModel worksheet.
Plot D^3 Chart	Allows user to display D^3 (Display, Document, Discover) or time series charts for one to five quality metrics. Which quality metrics are displayed is determined by the settings of the CommandCenter **D^3** checkboxes (Section 2.2.3). Once the D^3 chart has been generated, a number of options become available. These options are described in Section 5.
View D^3 Chart	If currently valid, activates the FigD3 worksheet.
View Correlation	If currently valid, activates the FigCorrel worksheet.

2.2 Quality Metrics, Controls, and Summary Values

All of the currently available "quality metrics," the control functions linked to them, and a sketchy statistical summary of their values in the current dataset are listed to the center right of the CommandCenter worksheet, as shown in Figure 3. A more complete discussion of the metrics currently available in Multiplex_QA is given in Section 8; the rest of this Section only presents how to make use of the critters.

A note of caution: These "quality metrics" are truly more *candidate* metrics for use in assessing aspects of either the quality of a given sample analysis or of the electrophorectic process when the sample was analyzed. Now that Multiplex_QA is more or less functional, we hope to be able to identify which of these metrics are most useful for specific tasks – and maybe even develop improved metrics that do the job(s) better.

2.2.1 Quality metrics

Each metric is listed in the CommandCenter worksheet with a more or less descriptive short phrase and a code name. For each metric, the optionbuttons, checkboxes, and summary statistics to the right of the code name "belong" to that particular metric.

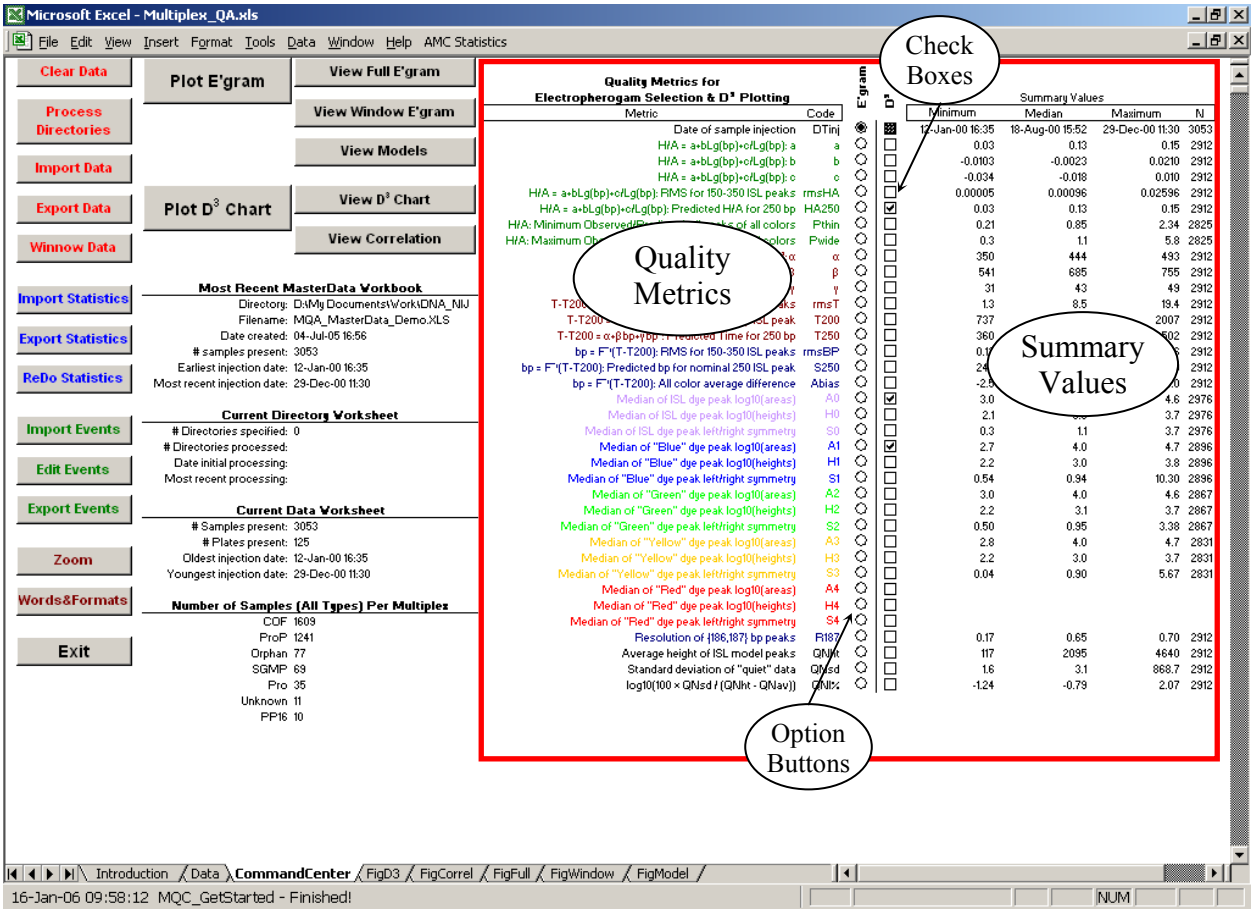
A note of caution: The code names are how Multiplex_QA links to the dataset stored on the Data worksheet. Changing a code name will break the link, probably just losing access to the metric… but if you really work at it, you can cause the system to erroneously link to another metric.

2.2.2 E'gram optionbuttons

The **E'gram** optionbuttons located immediately to the right of the code name communicate to the electropherogram plotting functions, invoked by clicking **Plot E'gram**. Only one optionbutton can be active at a given time. The electropherograms will be presented in order of increasing value of the specified metric.

A note of caution: Whenever you invoke **Plot E'gram**, the dataset is actually re-sorted in order of increasing value of the specified metric. Do not be concerned when the data order changes… but it is wise to **Export Data** before you start fiddlin' with a newly defined dataset.

Figure 3. Location of CommandCenter Quality Metrics, Controls, and Summary Values



2.2.3 D^3 checkboxes

The **D^3** checkboxes located immediately to the right of the **E'gram** optionbuttons communicate to the D^3 chart functions, invoked by clicking **Plot D^3 Chart**. At least one **D^3** checkbox must be active to be able to generate a D^3 chart. No more than five **D^3** checkboxes can be active at any one time.

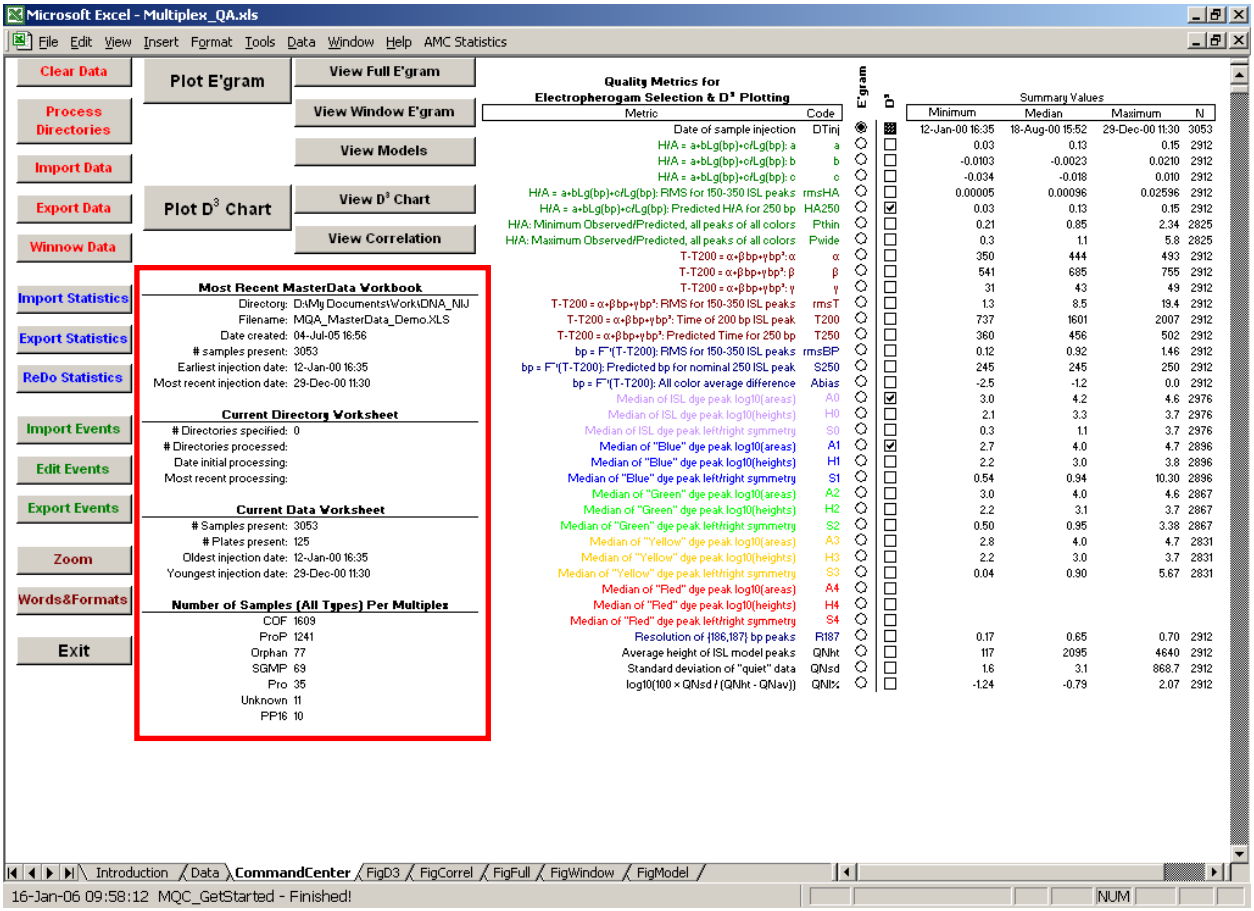
2.2.4 Summary values

The minimum, median, maximum, and number of valid values for each quality metric are listed to the right of the **D³** checkboxes. These summary values are displayed for your information only; they are not used elsewhere by the Multiplex_QA system.

2.3 Information Blocks

There are four blocks of information that help to identify any current dataset. These blocks are located to the center left of the CommandCenter worksheet, as shown in Figure 4. These information blocks are provided for your information only.

Figure 4. Location of CommandCenter Information Blocks



2.3.1 Most recent MasterData workbook

This block describes the most recently accessed MasterData workbook. This is both to remind you where the current dataset came from and to enable a rough check if any data have been deleted. Unless the dataset was built using the **Process Directories** (Section 2.4.2) command, the current dataset will be derived from this MasterData workbook. It may not be identical to the data saved in the MasterData workbook, however, as you may have deleted stuff – via **Winnow Data** (Section 2.4.5) or direct editing. The information displayed includes the

directory and filename of the MasterData workbook, the datetime it was created, number of samples described in the MasterData, and the earliest and most recent sample injection datetimes in the MasterData.

This information block is the only one of the four that is not cleared when you invoke the **Clear Old Data** (Section 2.4.1) command. It doesn't go away because it only describes the most recently accessed MasterData workbook, not the current dataset (if there is one). If you really want to clear the information in this block (like, for instance, when capturing a screen image), just select and delete.

2.3.2 Current Directory worksheet

This block describes the contents of the current Directory worksheet, if any, and is intended to help you keep track of progress while building a dataset from the .fsa files. The Directory worksheet is created by **Process Directories** (Section 2.4.2) and is removed from the Multiplex_QA system when you successfully tuck the competed dataset away for safekeeping with **Export Data** (Section 2.4.4).

2.3.3 Current Data worksheet

This block describes the contents of the current dataset, if any, and is intended to help you identify which dataset is actually present on the Data worksheet. The information displayed includes the number of samples, the number of identifiable plates of samples, the earliest sample injection datetime in the dataset, and the most recent injection datetime. The Data worksheet is created by the **Process Directories** (Section 2.4.2) or **Import Data** (Section 2.4.3) commands. It is removed from the Multiplex_QA system when you invoke the **Clear Old Data** (Section 2.4.1) or **Export Data** (Section 2.4.4) commands.

2.3.4 Number of samples (all types) per multiplex

This block describes the number of samples identified as having been analyzed using the multiplex kits currently recognized by the Multiplex_QA system. Like the Current Data Worksheet block, it is intended to help you identify which dataset is currently present.

2.4 Data Commands

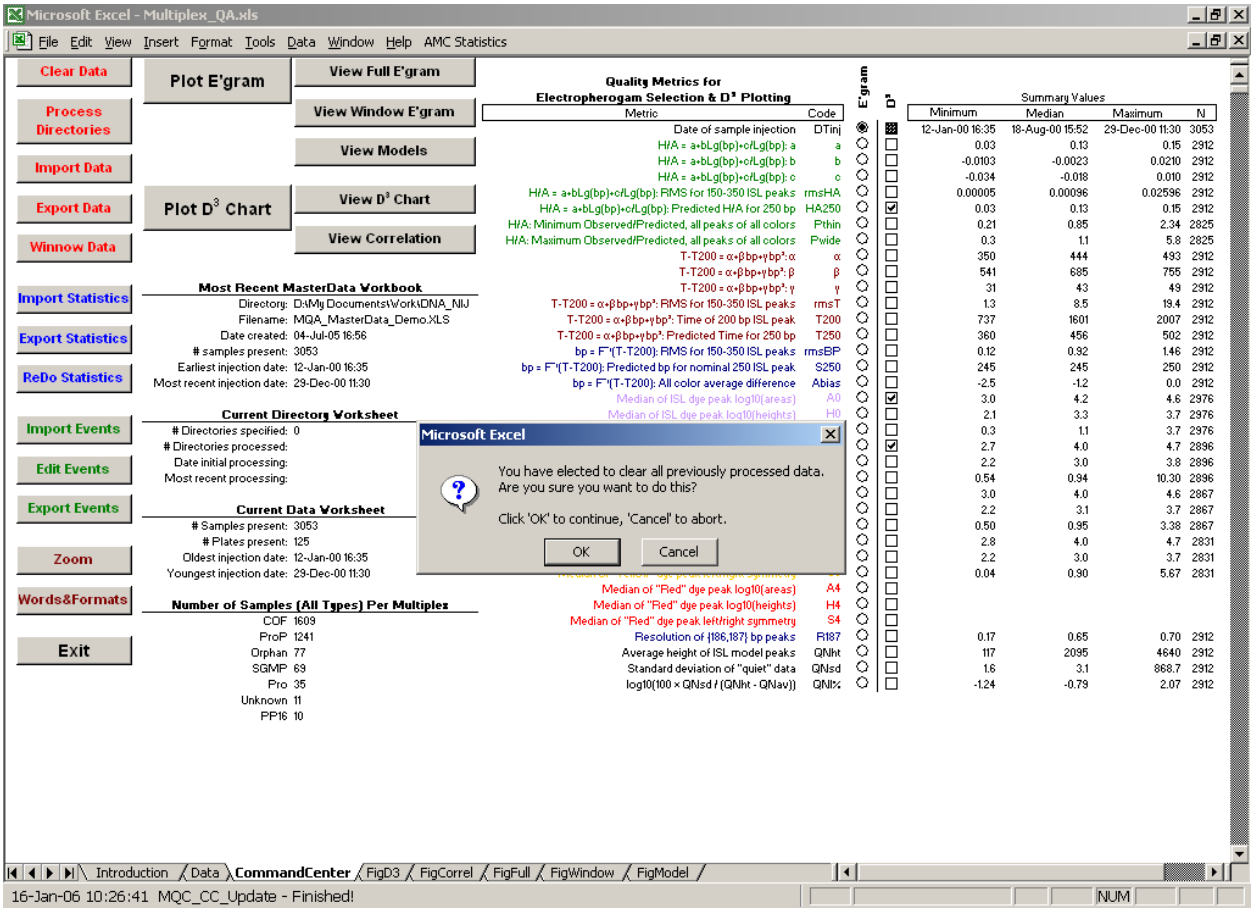
The following commands are the heart of the Multiplex_QA system: they allow you to specify and process new data, delete all or selected parts of the current dataset, and save and reuse datasets.

2.4.1 Clear Old Data

Clicking the **Clear Old Data** button allows you to clear all traces of the currently defined dataset from the Multiplex_QA system. This is most useful to do just before building new datasets. It is also useful when you wish to minimize the storage size of the Multiplex_QA workbook itself.

Figure 5 displays the dialog box that appears when **Clear Old Data** is clicked and there actually is a current dataset. If you wish to clear the decks, click **OK**; if you do not, click **Cancel**.

Figure 5. Clear Old Data



2.4.2 Process Directories

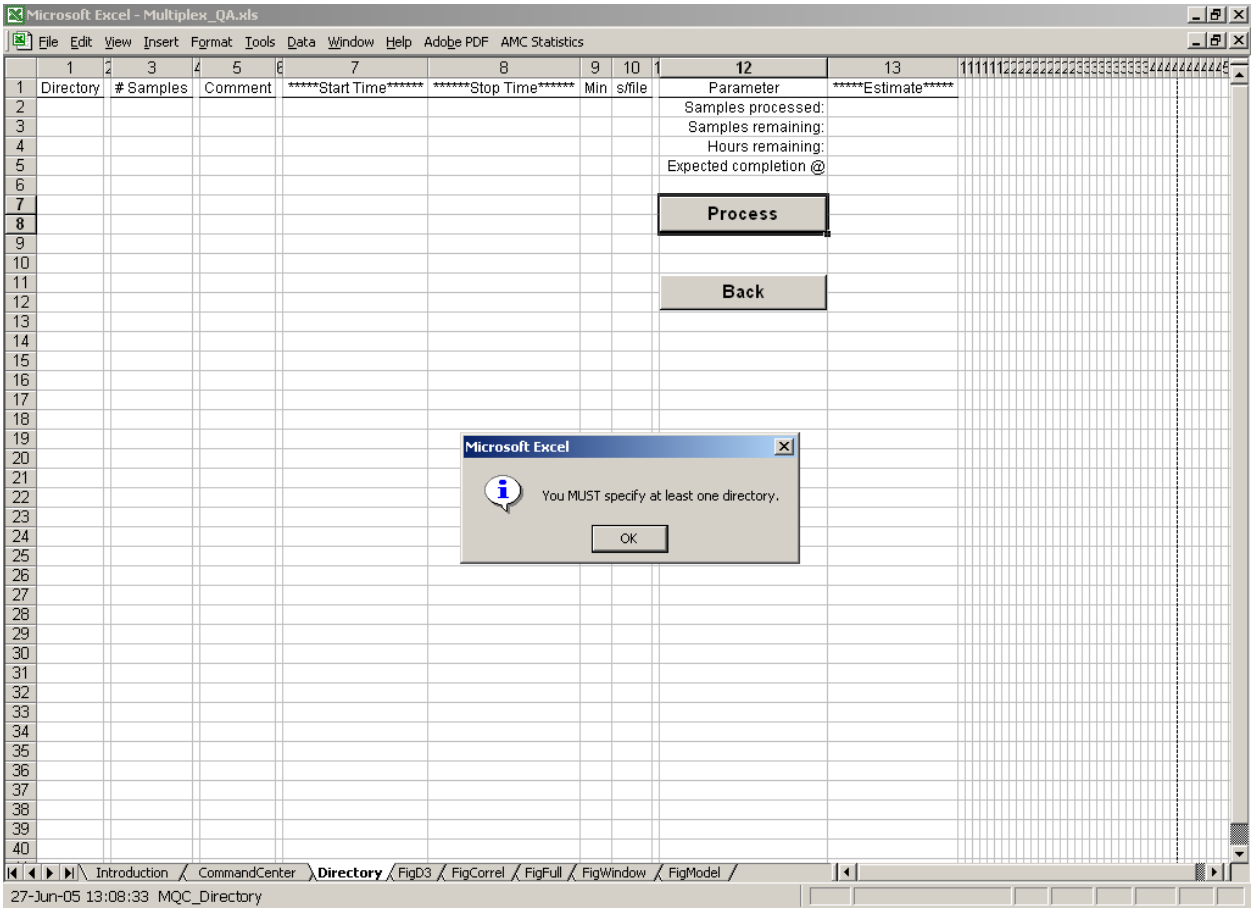
Clicking the **Process Directories** button is how datasets get built from the files extracted by the BatchExtract system (Section 9) from the ABI .fsa sample files. If you are intending to analyze your own data, it's critical that you master this somewhat tricky command. If you are just playing with test data, you can skip the rest of this section.

2.4.2.1 The Directory worksheet

Process Directories extracts Quality metric values from BatchExtract-generated files. While Multiplex_QA automatically identifies the needed files, it needs to be told where the files are located. In fact, since Excel can automatically detect a maximum of 256 files in a single directory ("folder"), the Multiplex_QA system actually needs to be told where these folders are located; *i.e.*, you need to specify the folder-paths to all the files you wish to process. Since processing files takes some time (on a fairly modern PC, a little less than a second a sample) and you may have many samples, progress needs to be recorded so that you can stop things mid-

stream without losing too much time and patience. The Directory worksheet, Figure 6, serves both these needs.

Figure 6. Process Directories, Directory Worksheet



2.4.2.2 Specifying folder-paths

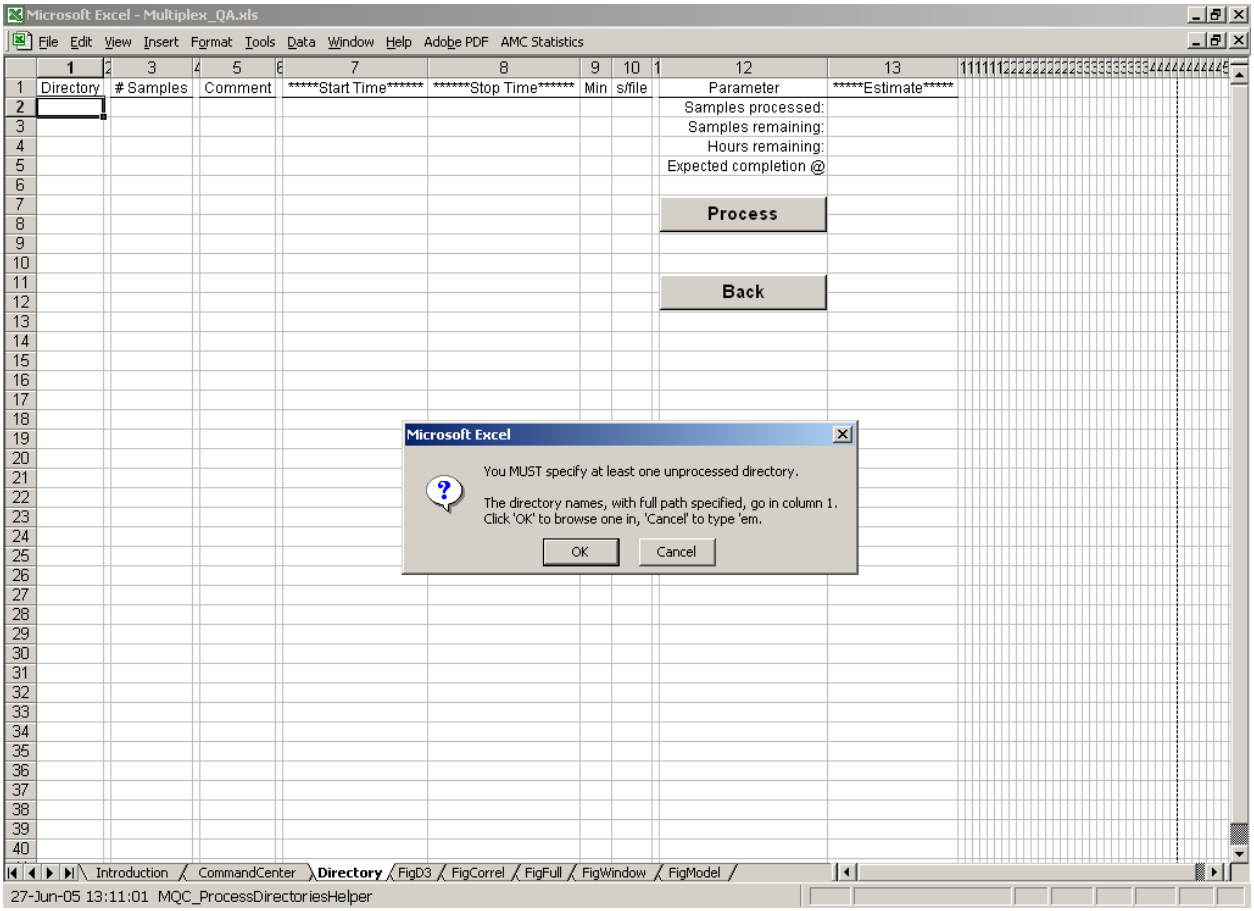
Figure 6 is representative of what you will see when you first invoke **Process Directories**. If there is no Directory worksheet in the Multiplex_QA workbook, one will be created when you invoke **Process Directories**. If the Directory worksheet is present, unless there are unprocessed folder-paths specified, the only difference in the display will be the presence of information about what's already been processed. In either case, click **OK** and the fun begins.

The "only" information you need to provide is the folder-path to one or more groups of files. All folder-paths need to be in the first column of the Directory worksheet. If you wish, type all of the folder-paths directly into separate rows of the first column of the Directory sheet. Don't worry about their order or about any blank rows, all will be sorted together and checked before the action starts. Once you've specified at least one folder-path, click the **Process** button.

Now, while directly typing folder-paths into the first column is conceptually easy, it is a bit of an error-prone pain… so is figuring out what to type. There is a better way. If you click **Process** without specifying a new folder-path, you should see something similar to Figure 7. If you click the Alert box's **Cancel**, Multiplex_QA just returns control to you and waits for you to

figure out what you wish to do next. However, clicking **OK** lets you use the "browse" function of the Excel Open dialog to specify a folder-path.

Figure 7. Process Directories, Processing an Empty List

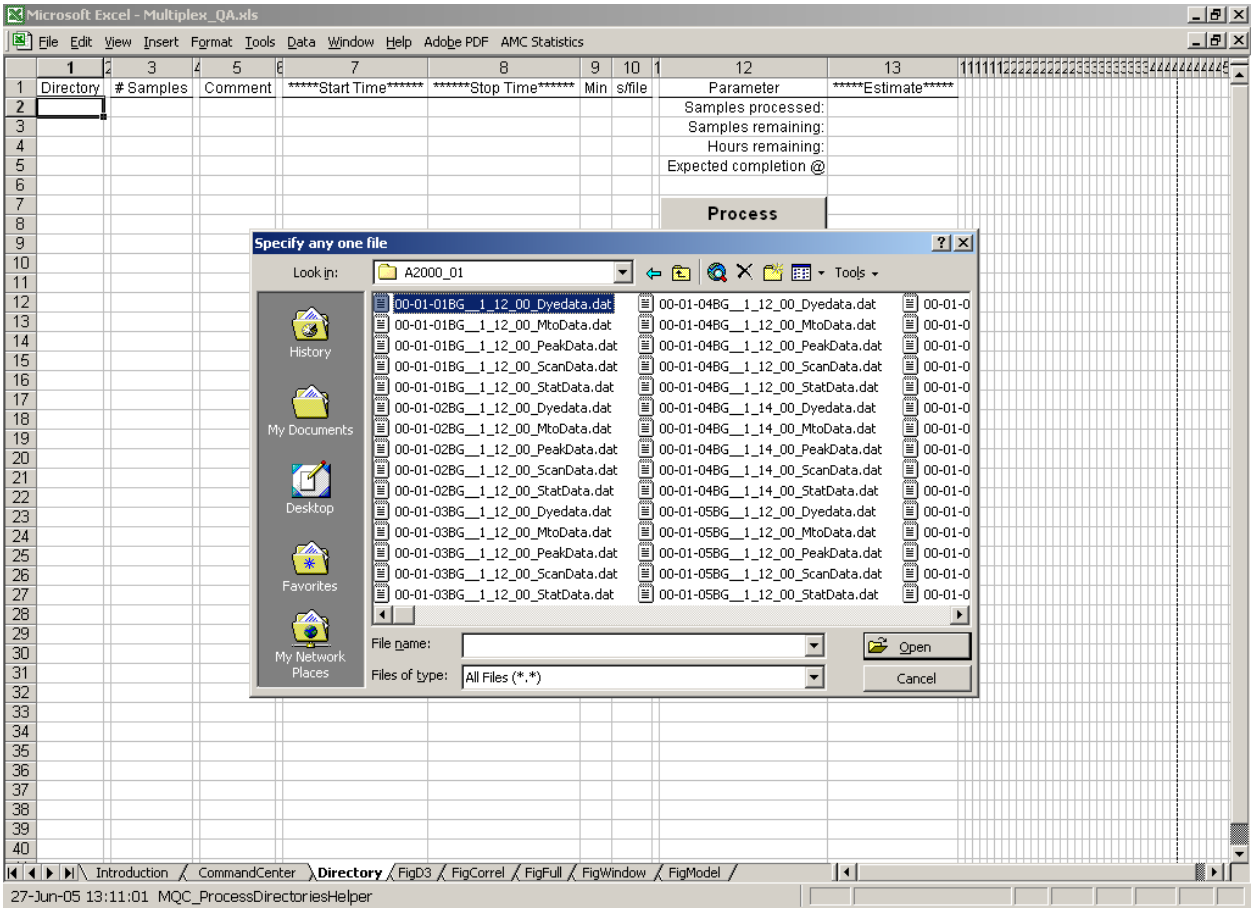


Once you've browsed into one of the folders you wish to process, you should see something similar to Figure 8. Select one of the listed files (it doesn't matter which one) and click **Open**. (Clicking **Cancel** just returns control to you; Multiplex_QA will then wait for you to figure out what you wish to do next). As soon as a valid file is specified, the folder-path will be logged into the Directory worksheet and processing will begin.

You can continue to browse folder-paths in one by one, but if you've located all of the folders in one central spot and if the folders are named in an easy to remember pattern, it may now be easier to copy the browsed-in folder-path, paste it into however many currently empty cells in the first column, then edit as needed.

If you are unafraid of using DOS-level commands, you can generate a text file listing of all folders in a directory and then copy and paste from this text list into the Directory worksheet (Section 9.2.6). However, you need to be modestly competent in DOS to make this worth your while. There's doubtless some Mac-ish way of doing this, too…

Figure 8. Process Directories, Specifying a Folder-Path by Browsing



2.4.2.3 Checking for common problems

Did I say "processing begins"? Well, a little error checking takes place first. Representative examples of the most common folder-path specification errors are shown in Figure 9. These problems are:

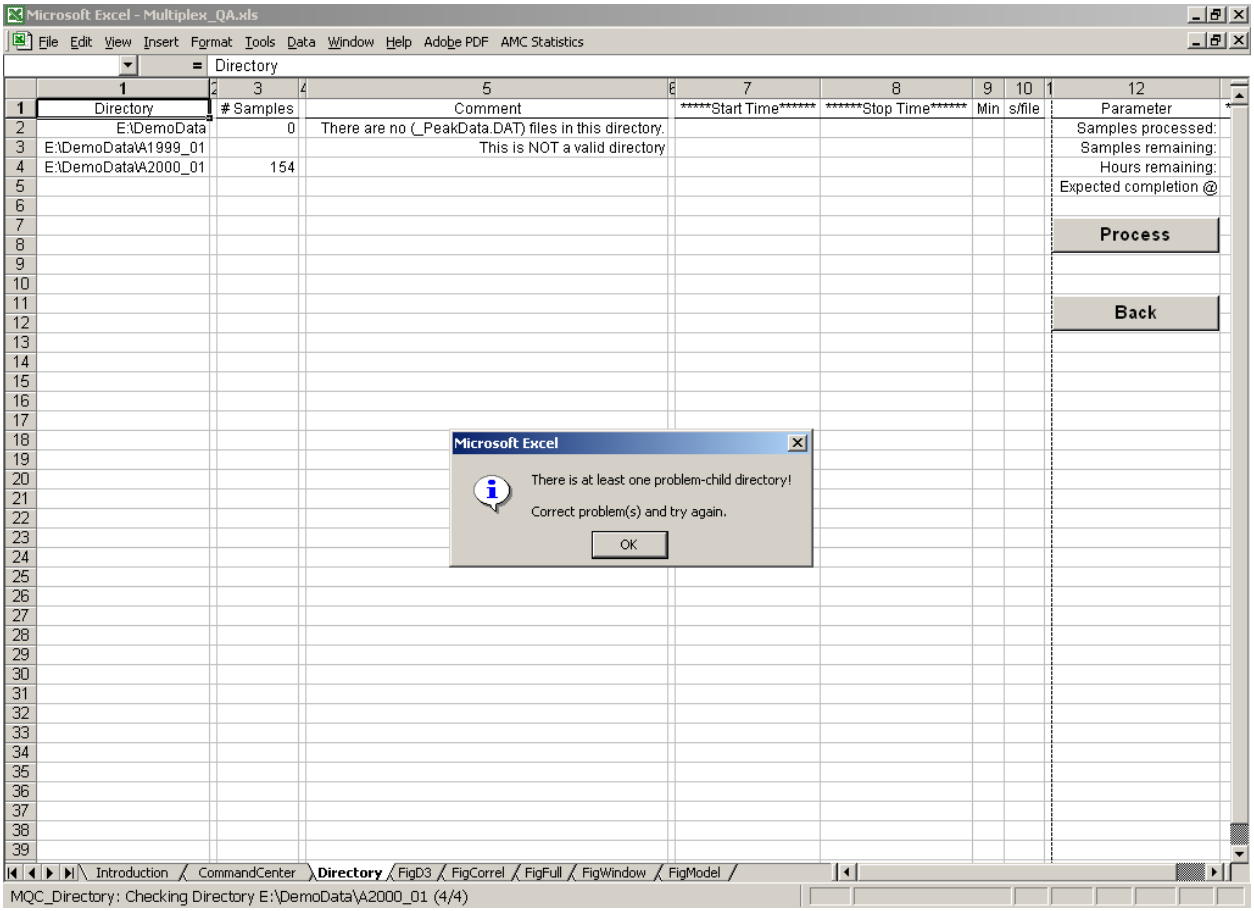
2.4.2.3.1 There is no such folder

You probably mistyped something. Check your spellling.

2.4.2.3.2 Empty folder

While there is a folder, there are no BatchExtract-extracted files in it. You probably incompletely specified the folder-path. However, it's possible that BatchExtract and Multiplex_QA have gotten "out of sync." Go check that the folder contains files that came from BatchExtract. If there are *_PeakData.dat files, there's probably some subtle error in the folder-path… try deleting it and "browsing" it back in (Section 2.4.2.2). If there are no *_PeakData.dat files, go read Section 10 …

Figure 9. Process Directories, Common Folder-Path Errors



2.4.2.3.3 More than 255 files

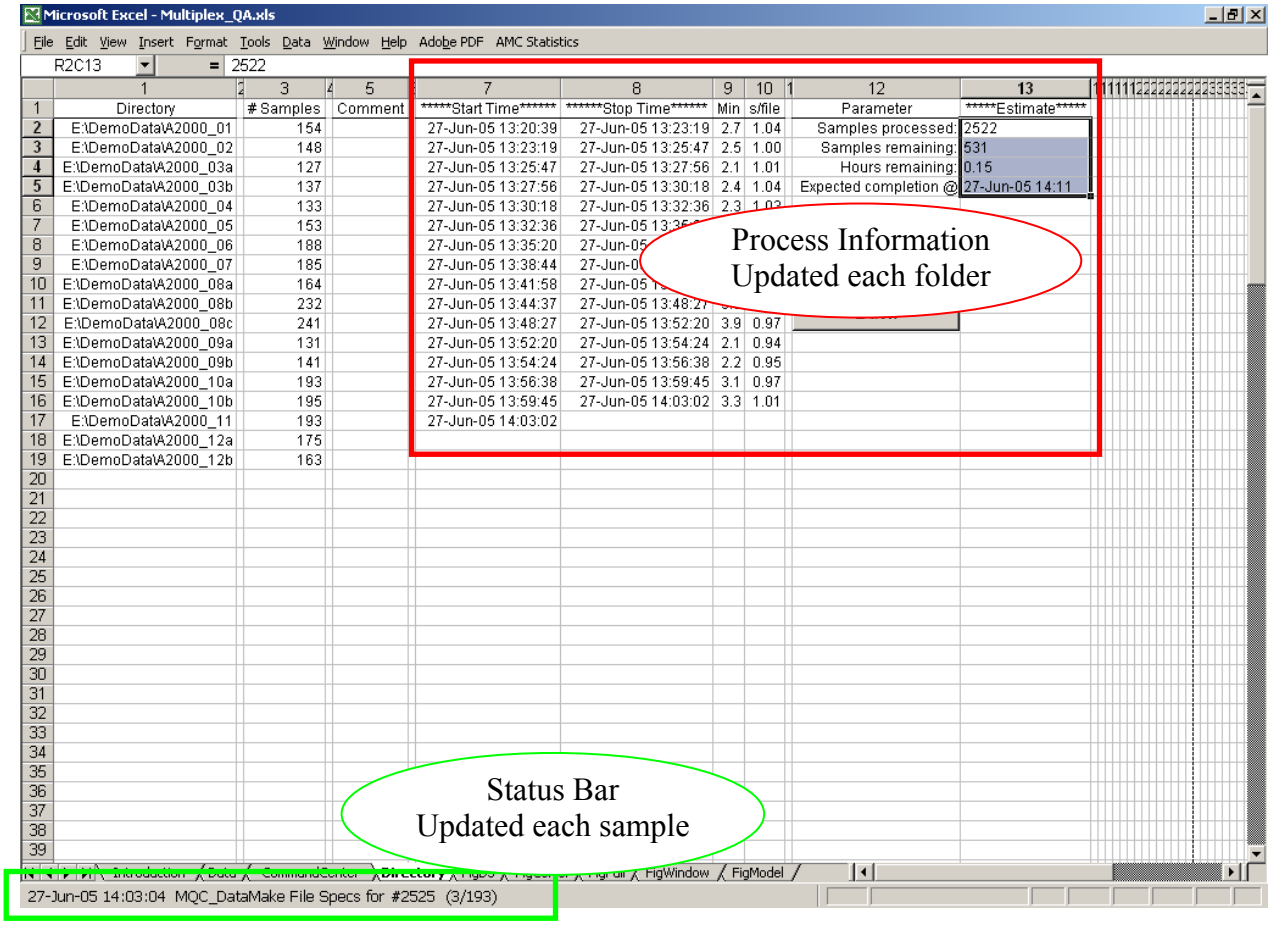
There are too many BatchExtract-ed files in a validly specified folder-path. Since Excel only recognizes a maximum of 256 files per folder, folders should contain at most 255 files to ensure that no sample is left behind. Split the sets of sample files into groups of no more than 255.

A note of caution: All duplicate folder-paths specified in the first column of the Directory worksheet are simply deleted without any warning. Before you start using a dataset just generated from Process Directories, double check that all the folder-paths you thought you specified are in fact still there.

2.4.2.4 Processing

Once a list of folders have been successfully specified and known problems corrected, the real processing begins. As shown in Figure 10, some information is provided during processing. The message in the Status bar at the bottom left of the active window is updated every time a new sample is processed. If an unexpected error occurs, you can identify the problem child from this info. The information block to the top right of the window is updated after each folder is completely processed. One of the more useful bits of information in this block is a guesstimate of when processing will finished, assuming the yet to be-processed files are similar to those that have already been processed.

Figure 10. Process Directories, Processing Status



	Directory	# Samples	Comment	*****Start Time******	******Stop Time******	Min	s/file	Parameter	*****Estimate*****
2	E:\DemoData\A2000_01	154		27-Jun-05 13:20:39	27-Jun-05 13:23:19	2.7	1.04	Samples processed:	2522
3	E:\DemoData\A2000_02	148		27-Jun-05 13:23:19	27-Jun-05 13:25:47	2.5	1.00	Samples remaining:	531
4	E:\DemoData\A2000_03a	127		27-Jun-05 13:25:47	27-Jun-05 13:27:56	2.1	1.01	Hours remaining:	0.15
5	E:\DemoData\A2000_03b	137		27-Jun-05 13:27:56	27-Jun-05 13:30:18	2.4	1.04	Expected completion @	27-Jun-05 14:11
6	E:\DemoData\A2000_04	133		27-Jun-05 13:30:18	27-Jun-05 13:32:36	2.3	1.03		
7	E:\DemoData\A2000_05	153		27-Jun-05 13:32:36	27-Jun-05 13:35				
8	E:\DemoData\A2000_06	188		27-Jun-05 13:35:20	27-Jun-05				
9	E:\DemoData\A2000_07	185		27-Jun-05 13:38:44	27-Jun-0				
10	E:\DemoData\A2000_08a	164		27-Jun-05 13:41:58	27-Jun-05 13				
11	E:\DemoData\A2000_08b	232		27-Jun-05 13:44:37	27-Jun-05 13:48:27				
12	E:\DemoData\A2000_08c	241		27-Jun-05 13:48:27	27-Jun-05 13:52:20	3.9	0.97		
13	E:\DemoData\A2000_09a	131		27-Jun-05 13:52:20	27-Jun-05 13:54:24	2.1	0.94		
14	E:\DemoData\A2000_09b	141		27-Jun-05 13:54:24	27-Jun-05 13:56:38	2.2	0.95		
15	E:\DemoData\A2000_10a	193		27-Jun-05 13:56:38	27-Jun-05 13:59:45	3.1	0.97		
16	E:\DemoData\A2000_10b	195		27-Jun-05 13:59:45	27-Jun-05 14:03:02	3.3	1.01		
17	E:\DemoData\A2000_11	193		27-Jun-05 14:03:02					
18	E:\DemoData\A2000_12a	175							
19	E:\DemoData\A2000_12b	163							

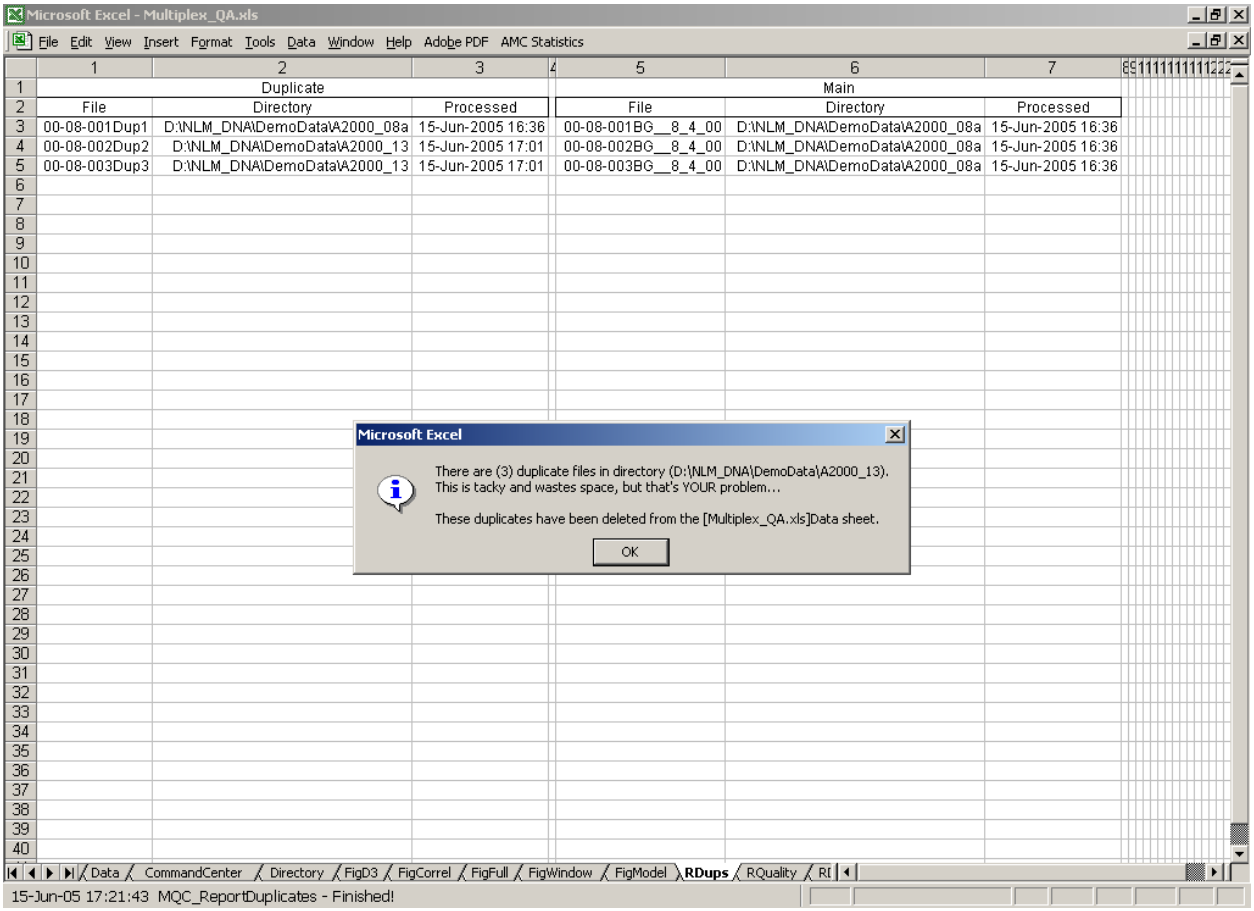
I generally monitor the processing of the first folder just to make sure that things are working, then – depending on when I started and the guesstimated time required to finish –take a break, switch computers, or leave the system to putter away by itself overnight. You can stop processing at any time (on PCs, hit the <Escape> key) without wasting more than one folder's worth of processing effort.

While in the expected course of events there's nothing in Multiplex_QA that you will need to interact with while processing proceeds, unanticipated fatal errors are possible. It's worth checking progress occasionally. If a particular sample is causing the problem, remove that set of sample files from its folder and retry **Process Directories**.

2.4.2.5 RDups worksheet: Duplicate File report

Immediately after sample file processing is complete, a check is made for the presence of duplicate samples. All simple duplications – the identical set of sample files in two or more folders – are corrected without comment. However, .fsa files are occasionally duplicated and renamed either by intent or misadventure when they are re-evaluated. These more complex duplications can be detected by examining the plate position and injection datetime characteristics. The Multiplex_QA system keeps the quality metrics only for the most recently re-evaluated of the files, but a record is made to allow you to investigate and correct your record-keeping.

Figure 11. Process Directories, RDups Worksheet



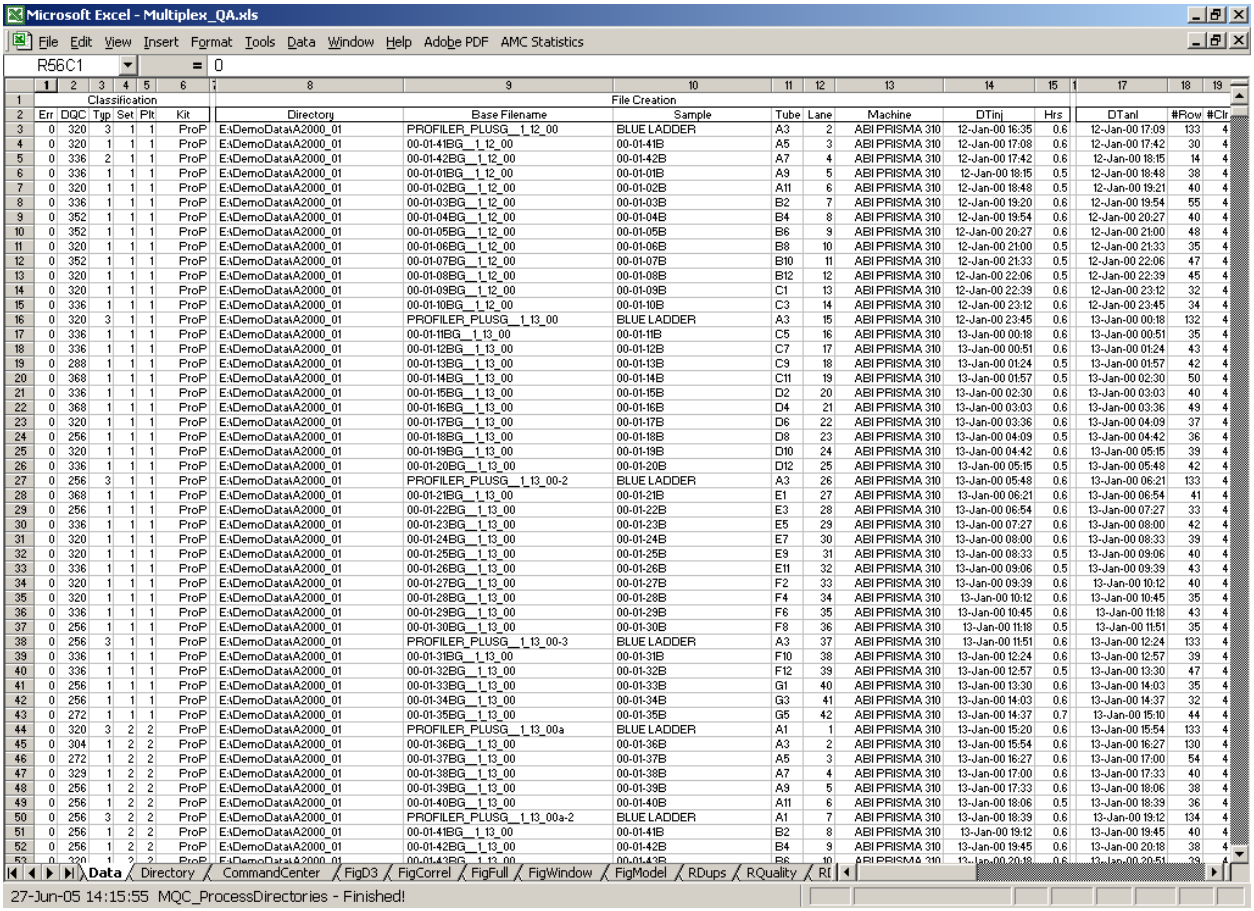
If one or more of these complex duplications is identified, the RDups worksheet is displayed along with a Warning similar to that in Figure 11. Clicking **OK** will let **Process Directories** proceed, clicking **Cancel** will return control to you should you wish to immediately investigate the duplication. In either case, the RDups worksheet remains with the Multiplex_QA system until additional folders are processed or the Data worksheet is saved with **Export Data** (Section 2.4.4).

2.4.2.6 Data worksheet

The primary product of **Process Directories** is the Data worksheet, similar to the example displayed in Figure 12. In addition to storing all of the quality metrics calculated for every sample (Section 8), the Data worksheet also contains a number of flags and identifiers that affect whether and how the quality metric data are used.

lists the column labels, a short description of what the column contains, and the source of the information. More complete descriptions of these data are provided in the following text.

Figure 12. Process Directories, Data Worksheet



2.4.2.6.1 Data integrity flags: Err and DQC

The Multiplex_QA system performs a number of checks on the integrity of the BatchExtracted data. The data gleaned for a particular sample is always stored in the Data worksheet, along with two sets of data integrity flags. These flags are summarized in Table 3. The values and descriptions of these flags are stored in the normally hidden PFlags worksheet.

The first column of the Data worksheet, labeled "Err," in general indicates whether or not the data are complete. There are currently eight different error or "E"-codes (flag-values of -99 to -92) that are issued when some expected data cannot be located or has an impossible value. These errors are detected during the time that the various BatchExtract files are open; when an E-code problem is recognized, the processing of that sample is stopped. Any value less than -1 in the Err column of the Data worksheet indicates that all to some of the data required for calculating the quality metrics are missing or invalid. A value of -1 indicates that the data for the sample are too suspect for further analysis.

Table 2, Data Worksheet: Flags and Identifiers

Label	What it is	Source
Err	Error flag: Data incomplete if < 0	Calculated
DQC	Information flag: Model quality concern if > 0	Calculated
Typ	Type of sample: Blank, Control, Ladder, Unknown	Calculated
Set	Index, sets of related samples	Calculated
Plt	Index, samples amplified on same 96-well plate	Calculated
Kit	STR multiplex kit type	Calculated
Directory	Location of BatchExtract-ed files	Directory
Base Filename	Base filename of BatchExtract-ed files	Directory
Sample	Sample name	Files
Tube	Sample well identifier (assumes 96-well plate)	Files
Lane	Electrophoresis lane identifier	Files
Machine	Electrophoresis instrument identifier	Files
DTinj	Datetime of sample injection	Files
Hrs	Elapsed hours from most recent previous sample injection	Calculated
DTanl	Datetime of sample data analysis	Files
#Row	Number of total peaks identified	Calculated
#Clr	Number of dye colors used	Files
Colors	Names of the dye colors used	Files
StdLdr	Name of the ISS: GS or ILS	Files

Table 3, Data Integrity and Quality Flags: Err and DQC

Code	Flag	Fatal?	Data Integrity Concern
E00	-99	Yes	Missing file or file structure error(s)
E01	-98	Yes	Analysis datetime earlier than Injection datetime
E02	-97	Yes	No "validly"-sized peaks of significant height or area
E03	-96	Yes	No ISS peaks
E04	-95	Yes	Unknown ISS label
E05	-94	Yes	Missing at least one required ISS peak(s)
E06	-93	Yes	The *PeakData.dat file has (at least) a bad header
E07	-92	Yes	The beginning, ending, or center index of one or more peaks is wrong
E08	-1	Yes	One or more fatal data quality concerns
I01	1	No	One or more low intensity or miss-sized peaks ignored
I02	2	No	At least one, but not all, sample color(s) without peaks
I03	4	No	Implausibly large number of peaks in at least one color
I04	8	No	ISS H/A = F(bp) model is suspiciously poor (RMSE high)
I05	16	No	One or more non-ISS H/A values is/are suspiciously large
I06	32	No	One or more non-ISS H/A values is suspiciously small
I07	64	No	ISS bp = F(Time) model is suspiciously poor (inverted RMSE high)
I08	128	No	ISS bp = F(Time) model suspiciously biased (calculated - stated)
I09	256	No	Nominal ISS 250 bp peak has distorted peak size assignment
I10	512	Yes	Missing "Tube" or "Lane" designation

The second column of the Data worksheets, labeled "DQC," in general indicates whether the data quality metrics were derived and, if derived, whether they "make sense." There are currently ten of these informational or "I"-codes. Since more than one of these I-code flags may be appropriate for a given record, for economy of storage the individual flag values are assigned as powers of 2, from 1 (2^0) to 512 (2^9). The value stored in the DOC column is the sum of the values of the individual flags.

If one or more "fatal" (a serious enough problem that the sample's data are so suspect that they are best ignored in further analysis) I-code event is encountered, a special E-code is used to prevent the sample from being used. There is currently only one I-code event regarded as this serious: one or more missing sample identifiers. These missing data are not in themselves critical, but may indicate more serious data file corruption.

2.4.2.6.2 Sample Type

All samples are assigned a sample type on the basis of the number of peaks in each of the colors reported for the given sample. Table 4 presents the current assignments.

A note of limitation: The process currently used to identify allelic ladders is, to say the least, imperfect, doubtless depends upon the datasets used to identify the expected number of peaks in the various allelic ladders, and should be replaced. While identification of allelic ladder files is not crucial to later analysis, it is used to visualize potential differences between ladder and "Unknown" samples.

Table 4, Sample Types

Code	Type	Meaning
0	Unclassifiable	No peak for any color (including the ISS) or way too many peaks for some color(s).
1	Unknown	Some peaks for some color as well as ISS peaks, but not as many as expected for an allelic ladder. Positive controls are considered to be "Unknowns."
2	Blank	No peaks other than for the ISS. These records may be for intentional non-template controls ("true blanks") or just samples that didn't amplify.
3	Known ladder	The number of peaks in each color is in the range expected for a known allelic ladder.
4	Unknown ladder	The number of peaks in each color is large enough to be an allelic ladder yet small enough that the data may be valid, but the numbers do not match any of the known ladders.

2.4.2.6.3 Sample Set and Plate

Samples are typically run as part of a set of materials injected either sequentially on a single column or simultaneously on different columns or in different lanes. After ordering by injection datetime, sets are identified by the amount of time between samples. To be assigned to the same Set, samples can be separated by no more than 2 hours. The sets in a given dataset are identified by a sequential integer index.

Samples are also typically amplified using 96-well plates. More than one Set of samples may have been amplified from the same Plate. To be assigned to the same Plate, no more than 16 hours can elapse between the injection datetime of the last sample of the earlier set and the initial injection datetime of the later set. Further, no sample in a candidate Set can have the same "Tube" identifier but a different "Sample" name as any of the samples already assigned to the Plate. The plates in a given dataset are identified by a sequential integer index.

2.4.2.6.4 Sample Kit

As discussed in Section 2.4.2.6.2, an attempt is made to identify allelic ladder samples. Identification is primarily based upon the number of peaks of the various colors relative to the expected number of peaks for confirmed allelic ladder samples. Table 5 presents the multiplexes and the associated minimum and maximum number of peaks for the allelic ladders thus far examined by the Multiplex_QA system.

All samples within a Plate are assumed to have been amplified with the same STR multiplex kit. When all samples that are assigned as allelic ladders in a Plate are for the same Kit, all of the samples are assigned to be of the same Kit. When no allelic ladder is present in a Plate or there are thought to be ladders of two or more different Kits in a plate, all samples between the differing ladder samples are assigned the Kit-type "Orphan."

Table 5, Recognized Multiplex Kits and Peak Numbers

		Minimum – Maximum number of Peaks			
Code	Name	Blue	Green	Yellow	Red
COF	ABI AmpFℓSTR COfiler	15 – 23	25 – 30	8 – 14	0
IDF	ABI AmpFℓSTR Identifiler	55 – 93	54 – 91	62 – 91	41 – 54
PP16	Promega PowerPlex 16	80 – 135	52 – 76	61 – 78	0
Pro	ABI AmpFℓSTR Profiler	19 – 38	25 – 63	12 – 31	0
ProP	ABI AmpFℓSTR Profiler Plus	30 – 36	48 – 60	25 – 30	0
SGMP	ABI AmpFℓSTR SGM Plus	39 – 49	54 – 73	52 – 82	0
Yfiler	ABI AmpFℓSTR Yfiler	31 – 52	34 – 78	41 – 60	20 – 34

The Kit parameters and the calculations used to assign the most probable Kit to a sample assigned as a probable allelic ladder is performed on the normally hidden PladType worksheet.

A note of limitation: The process currently used to identify Kit is, like the process used to assign the Type, "somewhat imperfect" and will eventually be replaced. While identification of Kit is not crucial to later analysis, it is used in various displays and can be used to winnow the dataset.

2.4.2.6.5 Directory and Base filename

The Directory is the last-known folder-path for the BatchExtract files from which the quality metrics were derived. The Base Filename is the name of the .fsa file from which the BatchExtract files were derived. These two identifiers are what the **E'gram** plotting commands use to link the samples with their electropherographic data.

2.4.2.6.6 Sample, Tube, Lane, and Machine

The Sample, Tube, Lane, and Machine identifiers are as extracted by the BatchExtract system. All four identifiers are used in other sections of the Multiplex_QA system. Sample is the name of the sample as the individual who originally set up the electropherographic analysis specified. Tube is the location of the sample in the 96-well plate. Lane is the gel lane or capillary designation for the electropherographic analysis. Machine is the name of the electropherographic instrumentation used, probably as set up by the first user of the equipment.

2.4.2.6.7 Injection Datetime, Hours, and Analysis Datetime

The DTinj datetime of sample injection is used extensively by other parts of the Multiplex_QA system; in particular, the D^3 plotting system displays quality metrics as a function of injection datetime. To ensure that all samples have a unique datetime, samples that are injected at the same time on multiple capillaries are assigned injection datetimes that differ from one another by three minutes.

The Hrs elapsed time is the time in decimal days between the injection datetime of the current sample and the sample immediately preceding it. While this parameter is used *en passant* to assigning Set and Plate for each sample, it is listed on the Data worksheet strictly as a convenience when manually reviewing the Set assignments.

The DTanl datetime of sample analysis is used to help identify multiple analyses of the same raw data. The analysis datetime occasionally is not properly BatchExtract-ed and so will be unassigned (*i.e.*, an empty cell) in the Data worksheet.

2.4.2.6.8 Number Peaks, Number Colors, and Color Names

The #Row, #Color, and Colors identifiers display the total number of peaks of all colors for the sample, the number of different dye colors present in the .fsa file, and the names of those dyes. These values are displayed for your convenience in reviewing the likely validity of the data. They are not otherwise used.

2.4.2.6.9 Internal Size Standard

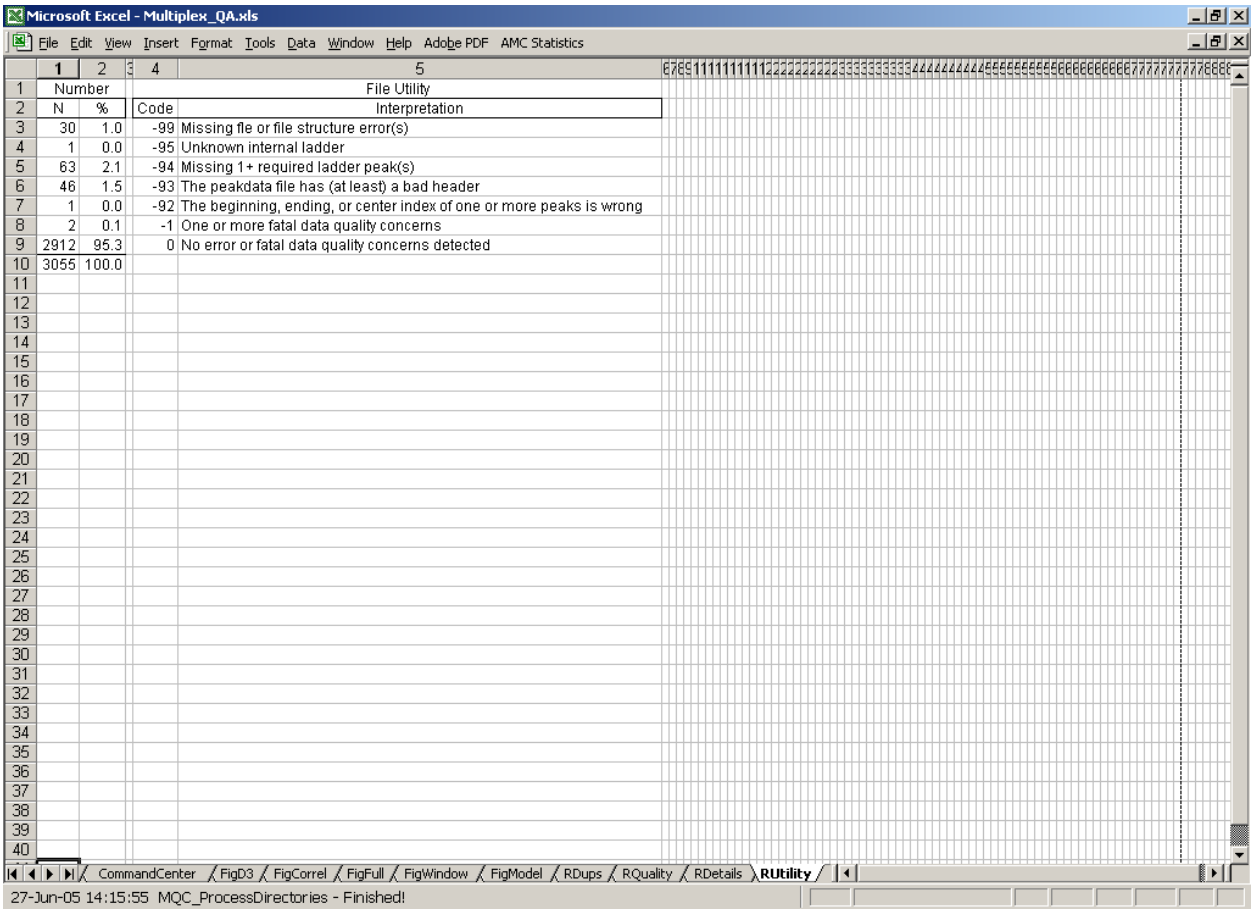
Three basic types of ISS are recognized: ABI's GeneScan GS350 and GS500 (GS350+), ABI's GS400, and Promega's Internal Lane Standard (ILS). The ISS types are assigned on the basis of the ISS name in the sample .fsa file: names that contain the string "IL" are assigned as ILS ladders, names that contain the string "GS4" are assigned as GS400 ladders, and names that contain either "GSx" where "x" is not "4" or "NONE" are assigned as GS350+ ladders. If the ISS type is not recognized, the type will be listed in the Data worksheet as "Unknown: xxx" where "xxx" is the full name from the .fsa file. Many of the Multiplex_QA quality metrics can only be calculated for samples having a known ISS type.

A note of warning: The process currently used to identify the ISS type is specific to the GS350+, GS400, and ILS ladders. When new ladders are introduced, the Multiplex_QA system will need to be modified.

2.4.2.7 RUtility worksheet: summary of fatal errors

Figure 13 displays an example of the "RUtility" worksheet report generated by **Process Directories**. This report summarizes the number of records afflicted with the various E-code fatal errors (Section 2.4.2.6.1).

Figure 13. Process Directories, RUtility Worksheet

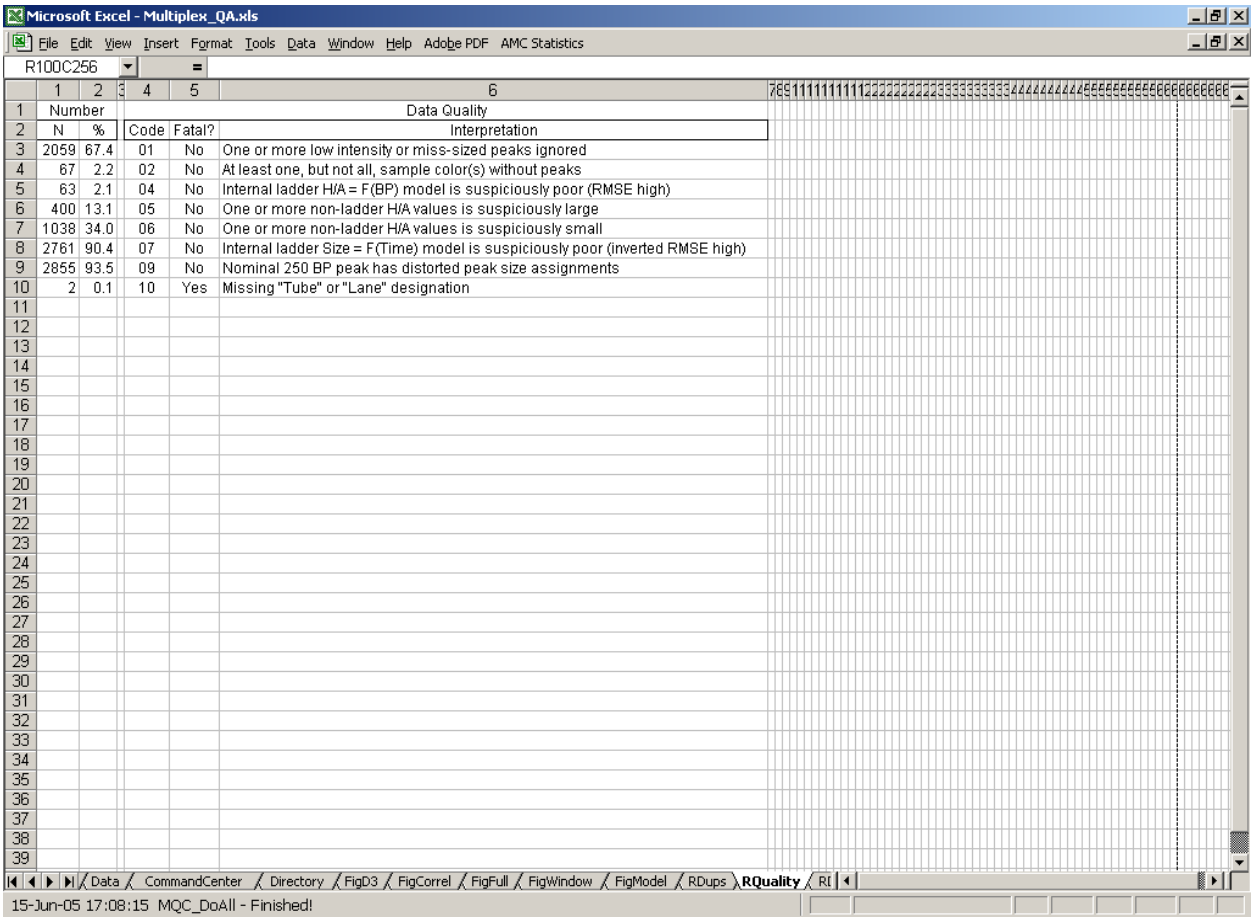


	Number		File Utility	
	N	%	Code	Interpretation
	30	1.0	-99	Missing fle or file structure error(s)
	1	0.0	-95	Unknown internal ladder
	63	2.1	-94	Missing 1+ required ladder peak(s)
	46	1.5	-93	The peakdata file has (at least) a bad header
	1	0.0	-92	The beginning, ending, or center index of one or more peaks is wrong
	2	0.1	-1	One or more fatal data quality concerns
	2912	95.3	0	No error or fatal data quality concerns detected
	3055	100.0		

2.4.2.8 RQuality worksheet

Figure 14 displays an example of the "RQuality" worksheet report generated by **Process Directories**. This report summarizes the number of records afflicted with the various I-code informational concerns (Section 2.4.2.6.1).

Figure 14. Process Directories, RQuality Worksheet



Microsoft Excel - Multiplex_QA.xls

File Edit View Insert Format Tools Data Window Help Adobe PDF AMC Statistics

R100C256 =

	1	2	4	5	6
1	Number				Data Quality
2	N	%	Code	Fatal?	Interpretation
3	2059	67.4	01	No	One or more low intensity or miss-sized peaks ignored
4	67	2.2	02	No	At least one, but not all, sample color(s) without peaks
5	63	2.1	04	No	Internal ladder H/A = F(BP) model is suspiciously poor (RMSE high)
6	400	13.1	05	No	One or more non-ladder H/A values is suspiciously large
7	1038	34.0	06	No	One or more non-ladder H/A values is suspiciously small
8	2761	90.4	07	No	Internal ladder Size = F(Time) model is suspiciously poor (inverted RMSE high)
9	2855	93.5	09	No	Nominal 250 BP peak has distorted peak size assignments
10	2	0.1	10	Yes	Missing "Tube" or "Lane" designation

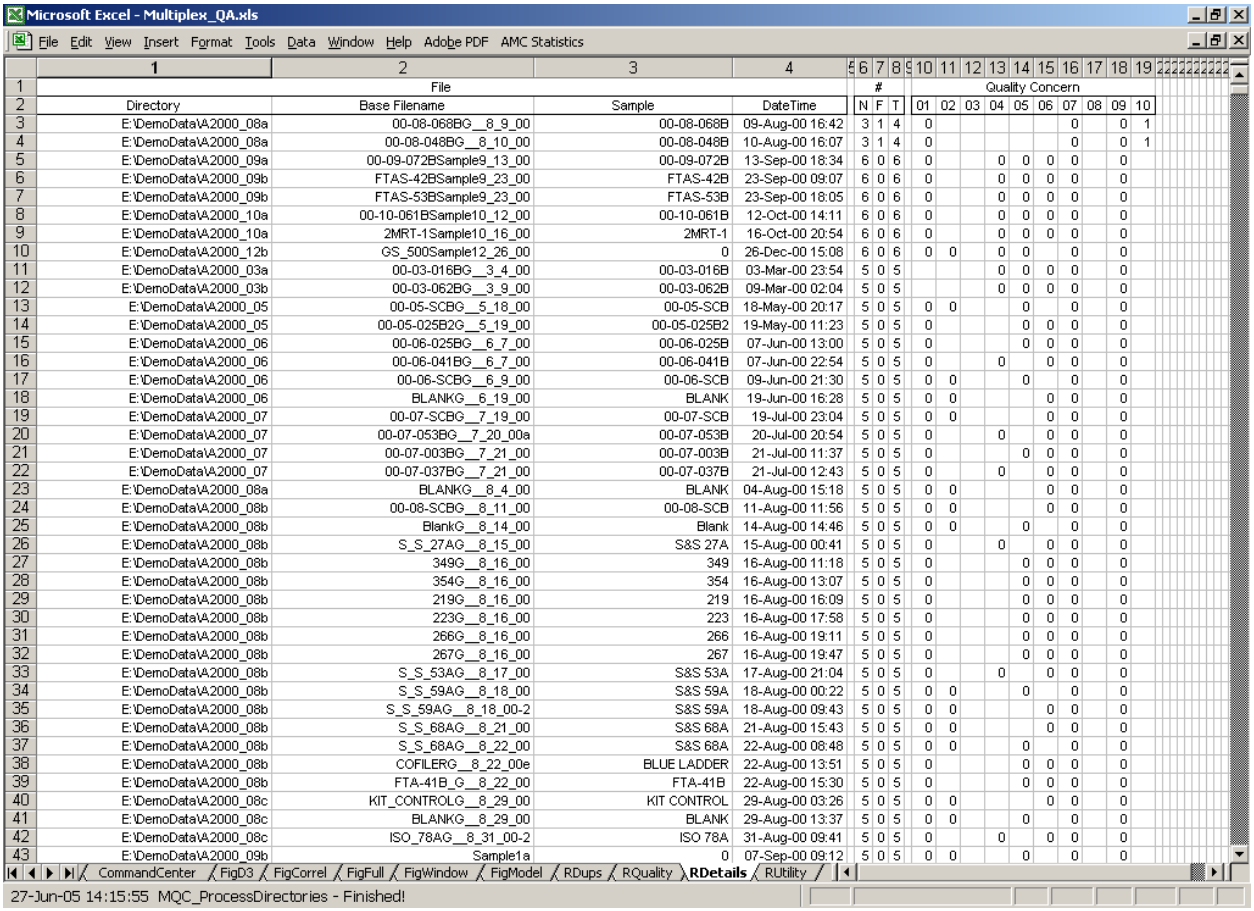
Data CommandCenter Directory FigD3 FigCorrel FigFull FigWindow FigModel RDups RQuality RI

15-Jun-05 17:08:15 MQC_DoAll - Finished!

2.4.2.9 RDetails worksheet

Figure 15 displays an example of the "RDetails" worksheet report generated by **Process Directories**. This report provides a detailed summary of all the I-code informational concerns for each sample afflicted with any concern (Section 2.4.2.6.1). In addition to sample identifiers and the breakout of the individual I-code flags for each sample, the number of nonfatal flags ("N"), the number of fatal flags ("F"), and the total number of flags ("T") are also listed. The samples are sorted by decreasing number of fatal and non-fatal flags.

Figure 15. Process Directories, RDetails Worksheet



Microsoft Excel - Multiplex_QA.xls

File Edit View Insert Format Tools Data Window Help Adobe PDF AMC Statistics

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	2222222222
1	File					#				Quality Concern										
2	Directory	Base Filename	Sample	DateTime		N	F	T		01	02	03	04	05	06	07	08	09	10	
3	E:\DemoData\A2000_08a	00-08-068BG__8_9_00	00-08-068B	09-Aug-00 16:42		3	1	4		0						0		0	1	
4	E:\DemoData\A2000_08a	00-08-048BG__8_10_00	00-08-048B	10-Aug-00 16:07		3	1	4		0						0		0	1	
5	E:\DemoData\A2000_09a	00-09-072BSample9_13_00	00-09-072B	13-Sep-00 18:34		6	0	6		0			0	0	0	0		0		
6	E:\DemoData\A2000_09b	FTAS-42BSample9_23_00	FTAS-42B	23-Sep-00 09:07		6	0	6		0			0	0	0	0		0		
7	E:\DemoData\A2000_09b	FTAS-53BSample9_23_00	FTAS-53B	23-Sep-00 18:05		6	0	6		0			0	0	0	0		0		
8	E:\DemoData\A2000_10a	00-10-061BSample10_12_00	00-10-061B	12-Oct-00 14:11		6	0	6		0			0	0	0	0		0		
9	E:\DemoData\A2000_10a	2MRT-1Sample10_16_00	2MRT-1	16-Oct-00 20:54		6	0	6		0			0	0	0	0		0		
10	E:\DemoData\A2000_12b	GS_500Sample12_26_00	0	26-Dec-00 15:08		6	0	6		0	0		0	0		0		0		
11	E:\DemoData\A2000_03a	00-03-016BG__3_4_00	00-03-016B	03-Mar-00 23:54		5	0	5					0	0	0	0		0		
12	E:\DemoData\A2000_03b	00-03-062BG__3_9_00	00-03-062B	09-Mar-00 02:04		5	0	5					0	0	0	0		0		
13	E:\DemoData\A2000_05	00-05-SCBG__5_18_00	00-05-SCB	18-May-00 20:17		5	0	5		0	0			0		0		0		
14	E:\DemoData\A2000_05	00-05-025B2G__5_19_00	00-05-025B2	19-May-00 11:23		5	0	5		0				0	0	0		0		
15	E:\DemoData\A2000_06	00-06-025BG__6_7_00	00-06-025B	07-Jun-00 13:00		5	0	5		0				0	0	0		0		
16	E:\DemoData\A2000_06	00-06-041BG__6_7_00	00-06-041B	07-Jun-00 22:54		5	0	5		0			0		0	0		0		
17	E:\DemoData\A2000_06	00-06-SCBG__6_9_00	00-06-SCB	09-Jun-00 21:30		5	0	5		0	0			0		0		0		
18	E:\DemoData\A2000_06	BLANKG__6_19_00	BLANK	19-Jun-00 16:28		5	0	5		0	0				0	0		0		
19	E:\DemoData\A2000_07	00-07-SCBG__7_19_00	00-07-SCB	19-Jul-00 23:04		5	0	5		0	0				0	0		0		
20	E:\DemoData\A2000_07	00-07-053BG__7_20_00a	00-07-053B	20-Jul-00 20:54		5	0	5		0			0		0	0		0		
21	E:\DemoData\A2000_07	00-07-003BG__7_21_00	00-07-003B	21-Jul-00 11:37		5	0	5		0				0	0	0		0		
22	E:\DemoData\A2000_07	00-07-037BG__7_21_00	00-07-037B	21-Jul-00 12:43		5	0	5		0			0		0	0		0		
23	E:\DemoData\A2000_08a	BLANKG__8_4_00	BLANK	04-Aug-00 15:18		5	0	5		0	0				0	0		0		
24	E:\DemoData\A2000_08b	00-08-SCBG__8_11_00	00-08-SCB	11-Aug-00 11:56		5	0	5		0	0				0	0		0		
25	E:\DemoData\A2000_08b	BlankG__8_14_00	Blank	14-Aug-00 14:46		5	0	5		0	0			0		0		0		
26	E:\DemoData\A2000_08b	S_S_27AG__8_15_00	S&S 27A	15-Aug-00 00:41		5	0	5		0			0		0	0		0		
27	E:\DemoData\A2000_08b	349G__8_16_00	349	16-Aug-00 11:18		5	0	5		0				0	0	0		0		
28	E:\DemoData\A2000_08b	354G__8_16_00	354	16-Aug-00 13:07		5	0	5		0				0	0	0		0		
29	E:\DemoData\A2000_08b	219G__8_16_00	219	16-Aug-00 16:09		5	0	5		0				0	0	0		0		
30	E:\DemoData\A2000_08b	223G__8_16_00	223	16-Aug-00 17:58		5	0	5		0				0	0	0		0		
31	E:\DemoData\A2000_08b	266G__8_16_00	266	16-Aug-00 19:11		5	0	5		0				0	0	0		0		
32	E:\DemoData\A2000_08b	267G__8_16_00	267	16-Aug-00 19:47		5	0	5		0				0	0	0		0		
33	E:\DemoData\A2000_08b	S_S_53AG__8_17_00	S&S 53A	17-Aug-00 21:04		5	0	5		0			0		0	0		0		
34	E:\DemoData\A2000_08b	S_S_59AG__8_18_00	S&S 59A	18-Aug-00 00:22		5	0	5		0	0			0		0		0		
35	E:\DemoData\A2000_08b	S_S_59AG__8_18_00-2	S&S 59A	18-Aug-00 09:43		5	0	5		0	0				0	0		0		
36	E:\DemoData\A2000_08b	S_S_68AG__8_21_00	S&S 68A	21-Aug-00 15:43		5	0	5		0	0				0	0		0		
37	E:\DemoData\A2000_08b	S_S_68AG__8_22_00	S&S 68A	22-Aug-00 08:48		5	0	5		0	0			0		0		0		
38	E:\DemoData\A2000_08b	COFILERG__8_22_00e	BLUE LADDER	22-Aug-00 13:51		5	0	5		0				0	0	0		0		
39	E:\DemoData\A2000_08b	FTA-41B_G__8_22_00	FTA-41B	22-Aug-00 15:30		5	0	5		0				0	0	0		0		
40	E:\DemoData\A2000_08c	KIT_CONTROLG__8_29_00	KIT CONTROL	29-Aug-00 03:26		5	0	5		0	0				0	0		0		
41	E:\DemoData\A2000_08c	BLANKG__8_29_00	BLANK	29-Aug-00 13:37		5	0	5		0	0			0		0		0		
42	E:\DemoData\A2000_08c	ISO_78AG__8_31_00-2	ISO 78A	31-Aug-00 09:41		5	0	5		0			0		0	0		0		
43	E:\DemoData\A2000_09b	Sample1a	0	07-Sep-00 09:12		5	0	5		0	0			0		0		0		

CommandCenter | FigD3 | FigCorrel | FigFull | FigWindow | FigModel | RDups | RQuality | RDetails | RUtility

27-Jun-05 14:15:55 MQC_ProcessDirectories - Finished!

2.4.2.10 Viewing the processing reports

The RDetails, RDups, RQuality, and RUtility worksheets can be viewed after the **Process Directories** process is complete by clicking on their tabs. These worksheets will be removed from the Multiplex_QA workbook when you Export Data (Section 2.4.4), so if you wish to review these reports you should do so sooner rather than later.

2.4.2.11 Manual cleanup

Many of the flag and identifier parameters described above can be "hand corrected" if the assigned values are incorrect. In particular, the Type, Set, Plate, and Kit assignments should be reviewed and corrected as necessary.

2.4.3 Import Data

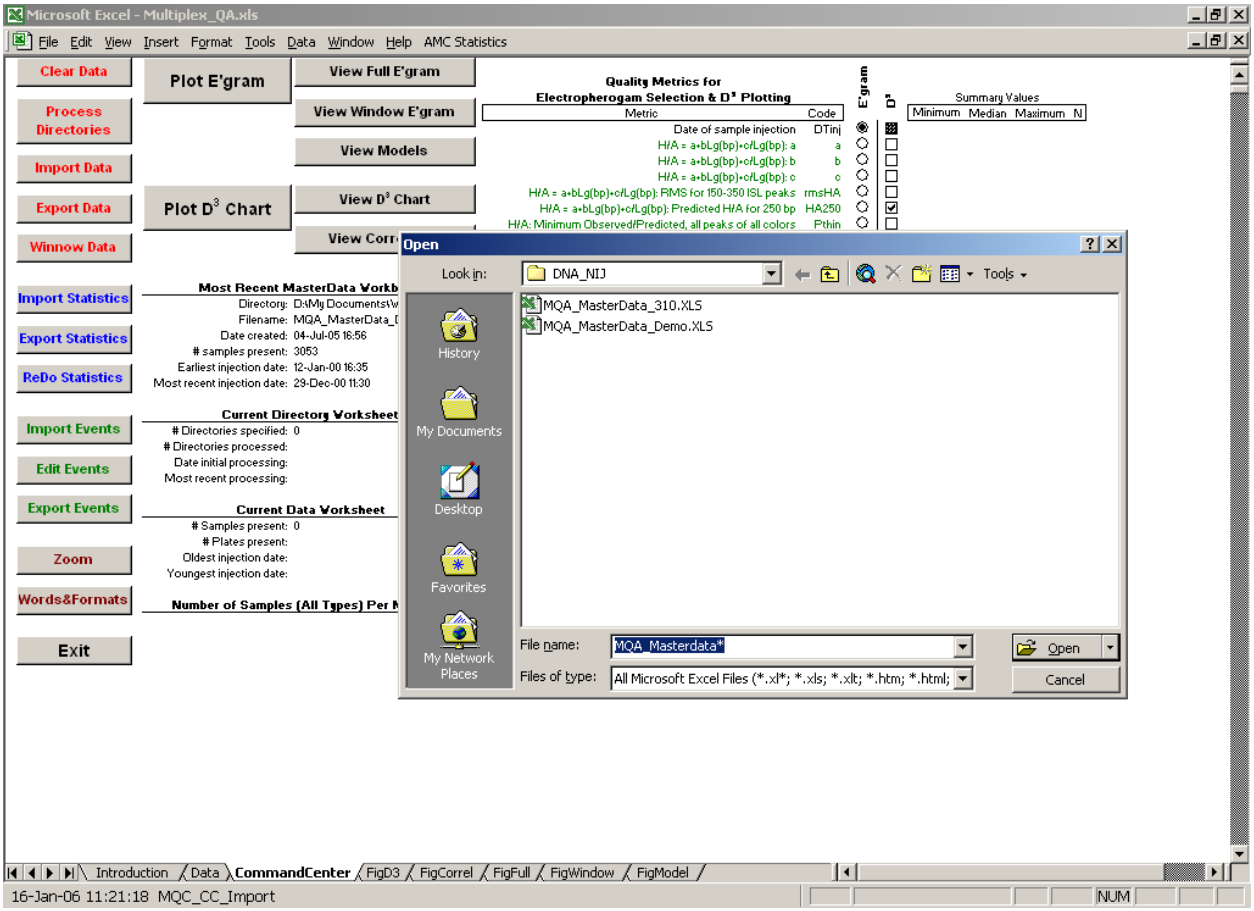
Clicking the **Import Data** button replaces any current QA dataset held on the (obscurely named) worksheet "Data" with that from a "MasterData" Excel workbook. Figure 5 displays the dialog box that appears when **Import Data** is invoked if a dataset is already defined. Note that summary information on the current dataset is displayed in the CommandCenter worksheet to the left of the dialog box. If you wish to replace the current QA dataset, click **OK**; if you do not, click **Cancel**.

Figure 16 displays the dialog box that appears if there is no current dataset or you elected to replace a current dataset, asking you to specify a MasterData workbook. For ease in identifying them, the MasterData workbooks are by default named "MQA_MasterData_*.xls," where the "*" represents the workbook's creation date. These files can be renamed as you like, but remembering which contains what is then your problem. No matter how the workbook is named, the worksheet containing the QA dataset must be named "Data."

If you confirm that you do indeed want to import a new QA dataset, you will be presented with a standard Excel file "Open" box similar to that shown in Figure 8. The Open box will list all of the files that match the "MQA_MasterData*.xls" wildcard search in the currently active directory (plus a list of all of the folders in that directory). You can select one of the MQA_MasterData*.xls files by clicking on the listed name and then clicking **Open**.

If the file you want is named something that is not matched by the wildcard search specified in the "File name" input field, you can type a specific filename or a new wildcard search into the input field followed by clicking Open or hitting <Enter>. If a specific file was named, it will be opened and the dataset transferred to the Data worksheet. If a new wildcard search was specified, everything that matches that search will be listed for your perusal. If the file you want isn't in the currently active directory, you can use the Open box's navigation functions to move to the desired directory.

Figure 16. Import Data, Specifying the MasterData File



A note of caution: The current dataset will be completely cleared from the Multiplex_QA system *as soon as you confirm that you do want to import a new dataset*; that is, the old dataset is cleared **before** the new set is specified.

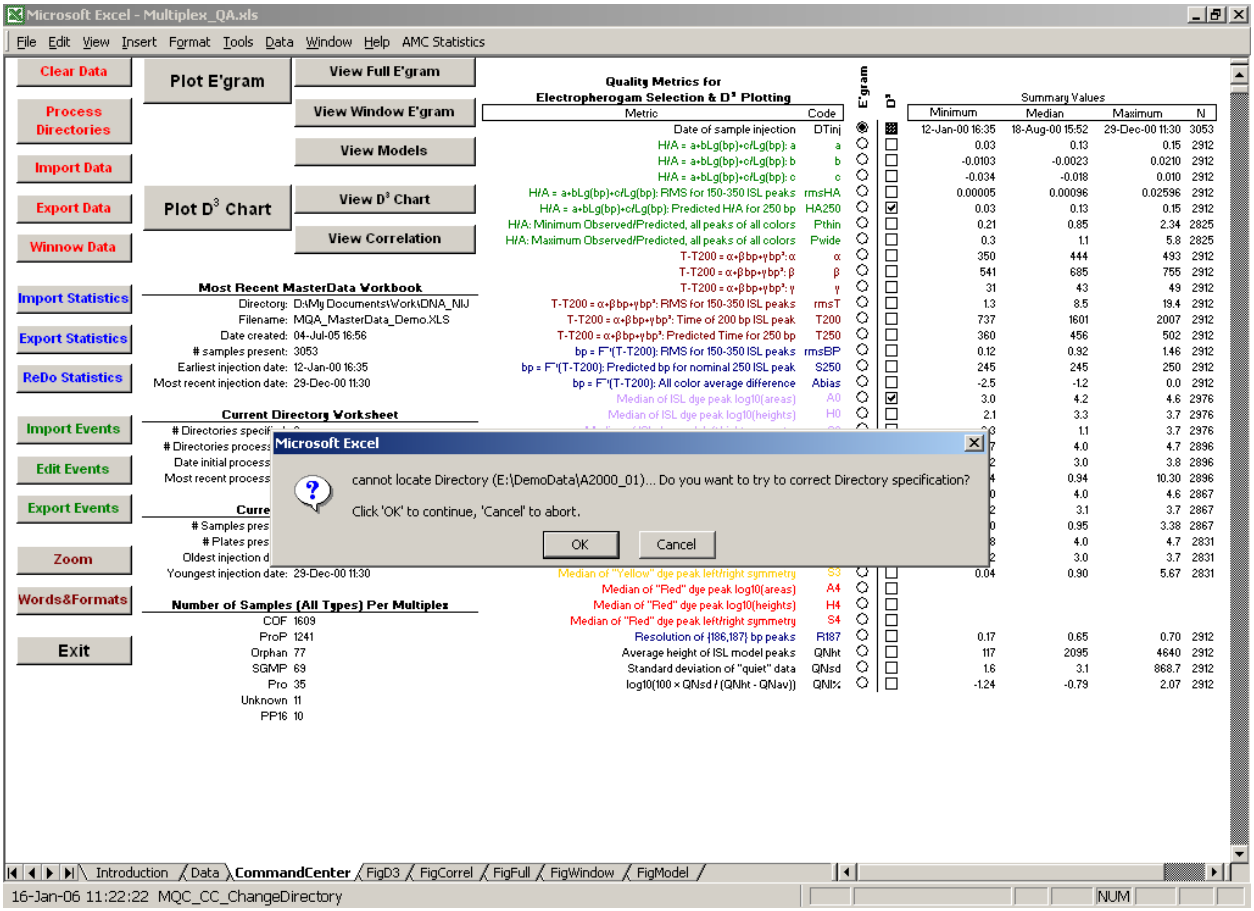
Clicking **Cancel** returns you to the CommandCenter. Remember, however, any dataset that had been specified when you first invoked **Import Data** is gone, gone, gone…

2.4.3.1 When the BatchExtract files aren't where they were

After a new dataset has been imported, Multiplex_QA attempts to confirm that the folder-paths specified in the "Directory" identifier of the Data worksheet are valid. The Directory identifiers will typically be as specified in the Directory worksheet used by **Process Directories** to define the dataset. However, when files are transported from one computer to another (such as the Test dataset accompanying this manual), the folder-paths may need to be updated.

If the linkage for the sample with the earliest injection datetime is invalid, a prompt dialog similar to that displayed in Figure 17 will be issued.

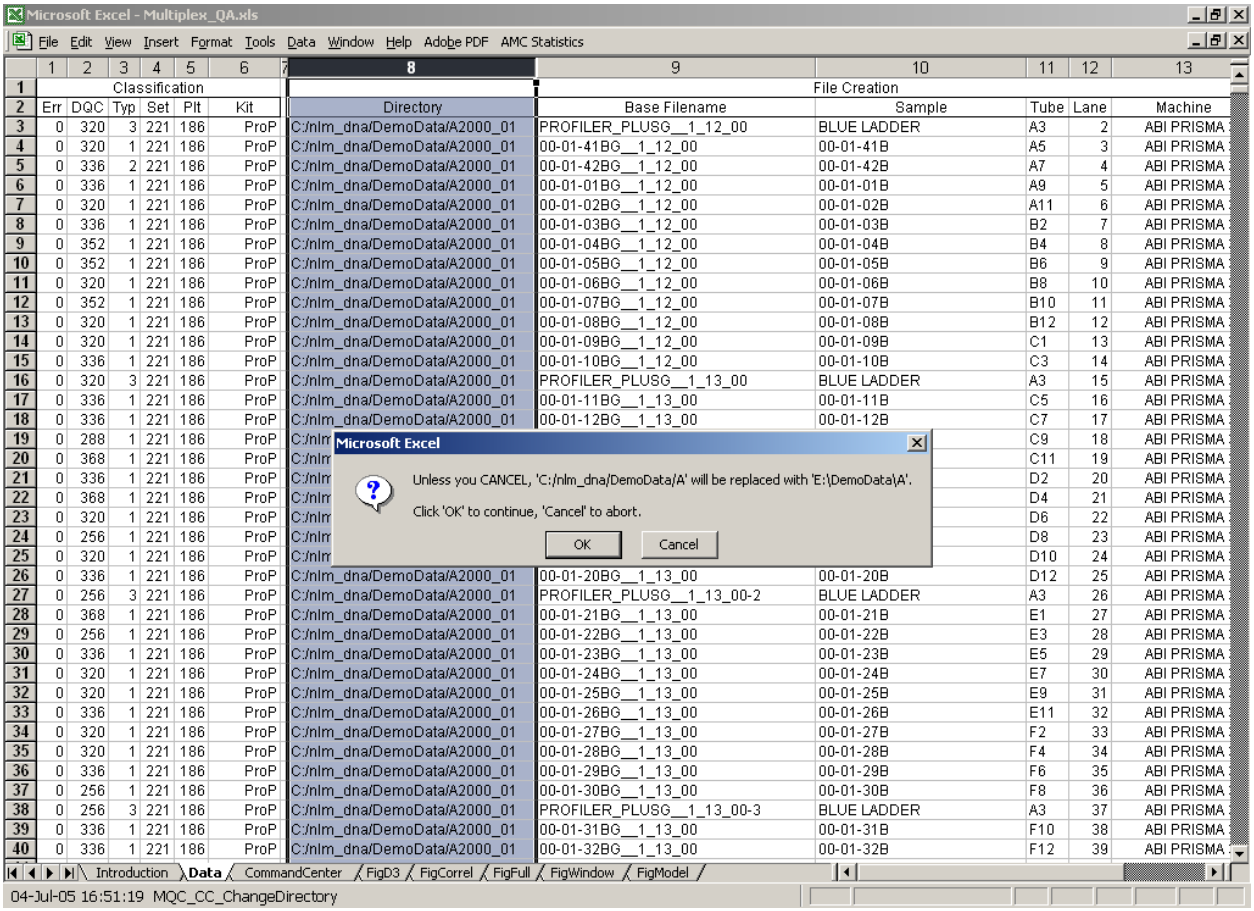
Figure 17. Import Data, Changing Folder-Paths.



If you do not wish to use the Plot E'gram functions, click **Cancel**. If you anticipate using the Plot E'gram functions and you know where the BatchExtract files are currently located, click **OK**.

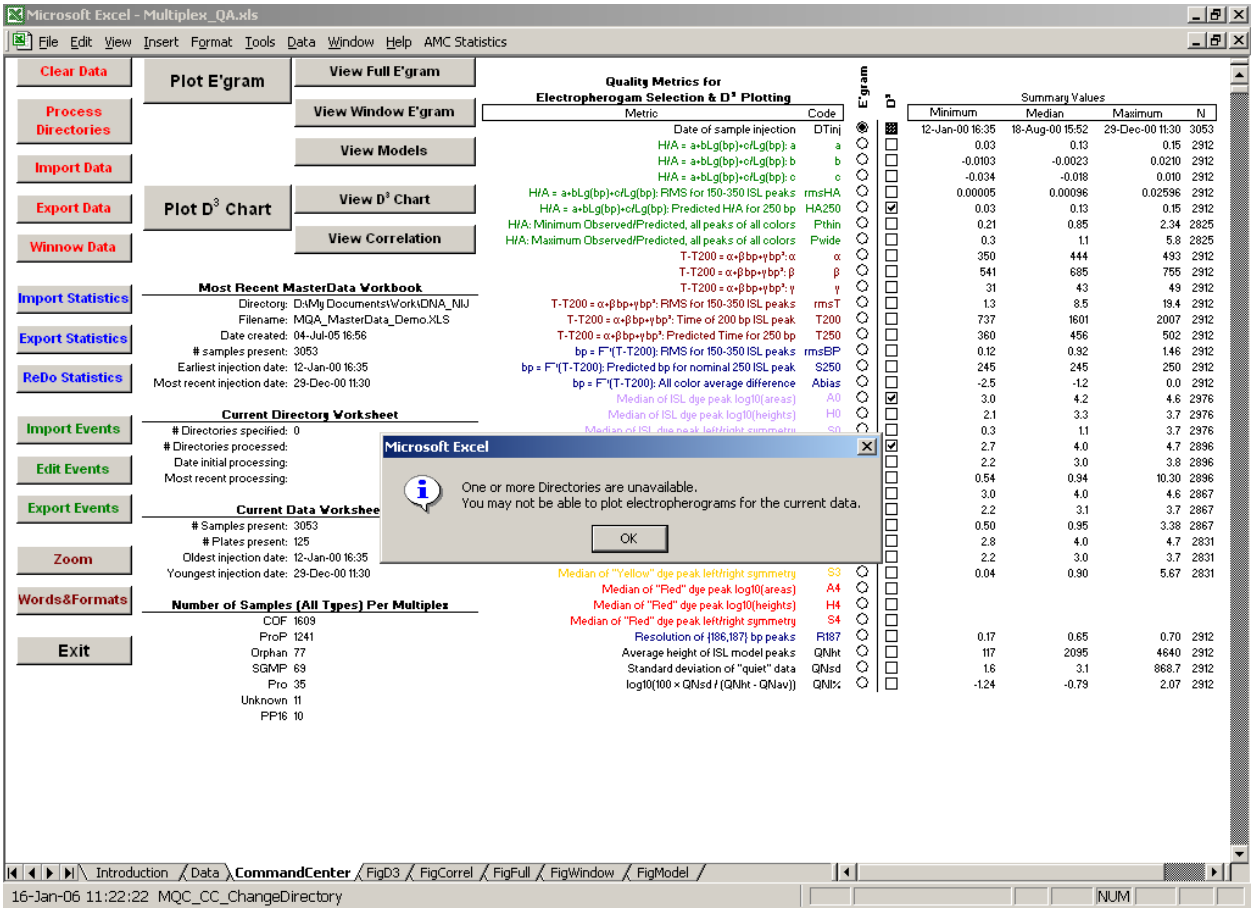
If you click **OK**, you will be requested to specify the current folder-path corresponding to that displayed in the dialog, in a process analogous to that described in Section 2.4.2.2. If you can locate the corresponding folder, you will be asked to confirm the change by an alert box similar to that in Figure 18. If you click **OK**, the folder-paths in the current dataset will be updated.

Figure 18. Import Data, Confirming a Folder-Path Change.



If you chose not to make the folder-path change or cannot locate the requested data, a warning alert similar to that in Figure 19 will be issued and you will be unable to use the E'gram functions.

Figure 19. Import Data, Unable to Locate .dat Files.



2.4.4 Export Data

Clicking the **Export Data** button actually moves the Data worksheet into a new workbook, along with all of the associated worksheets used to define the data or created while processing .fsa files (Directory, RDetails, RDups, RPlate, RQuality, and RUtility). Figure 20 displays the Alert box that appears when **Export Data** is invoked. If you wish to save the current data, click **OK**; if you do not, click **Cancel**. If there are no data, a Warning box will nag you about this deficiency and return you to the CommandCenter to regroup.

Figure 20. Export Data, Confirmation

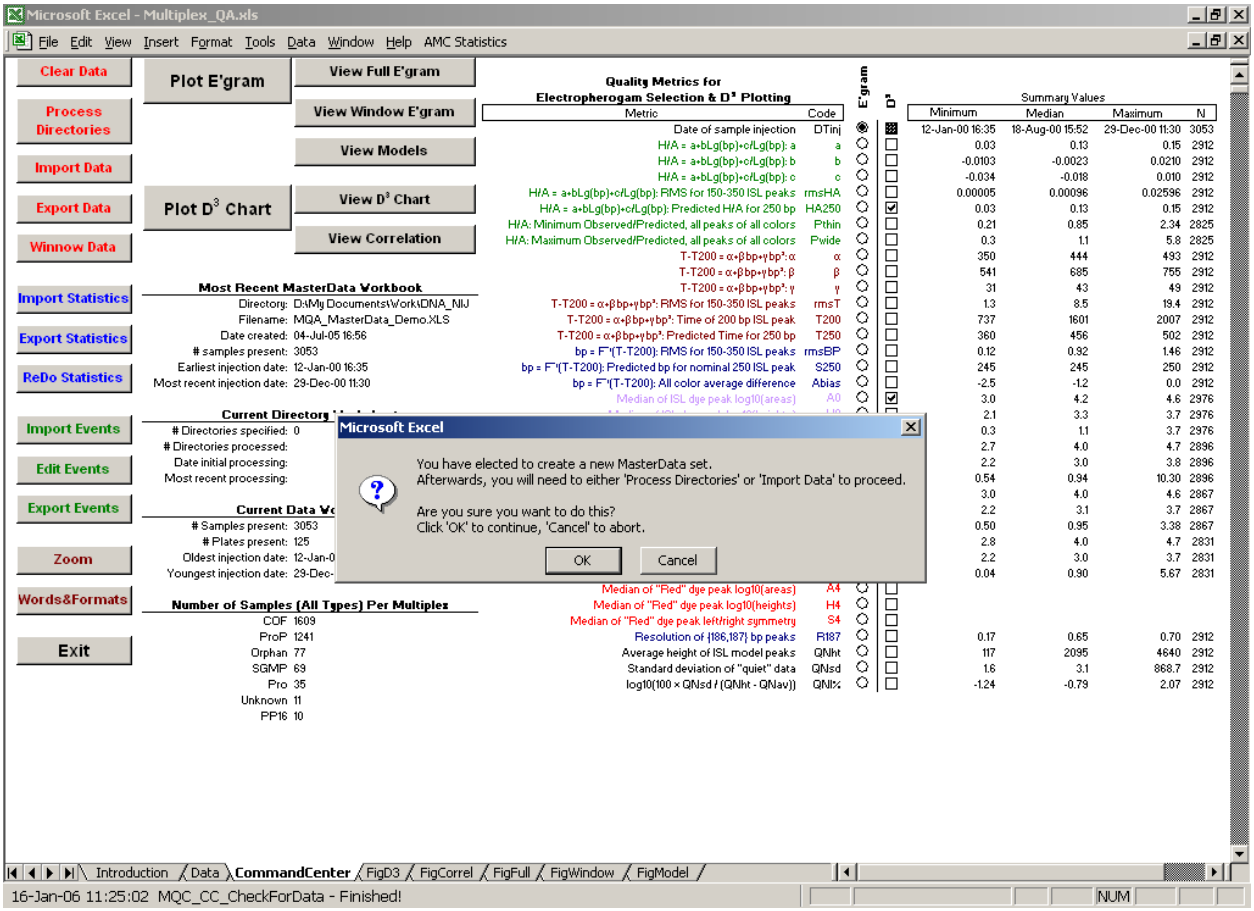
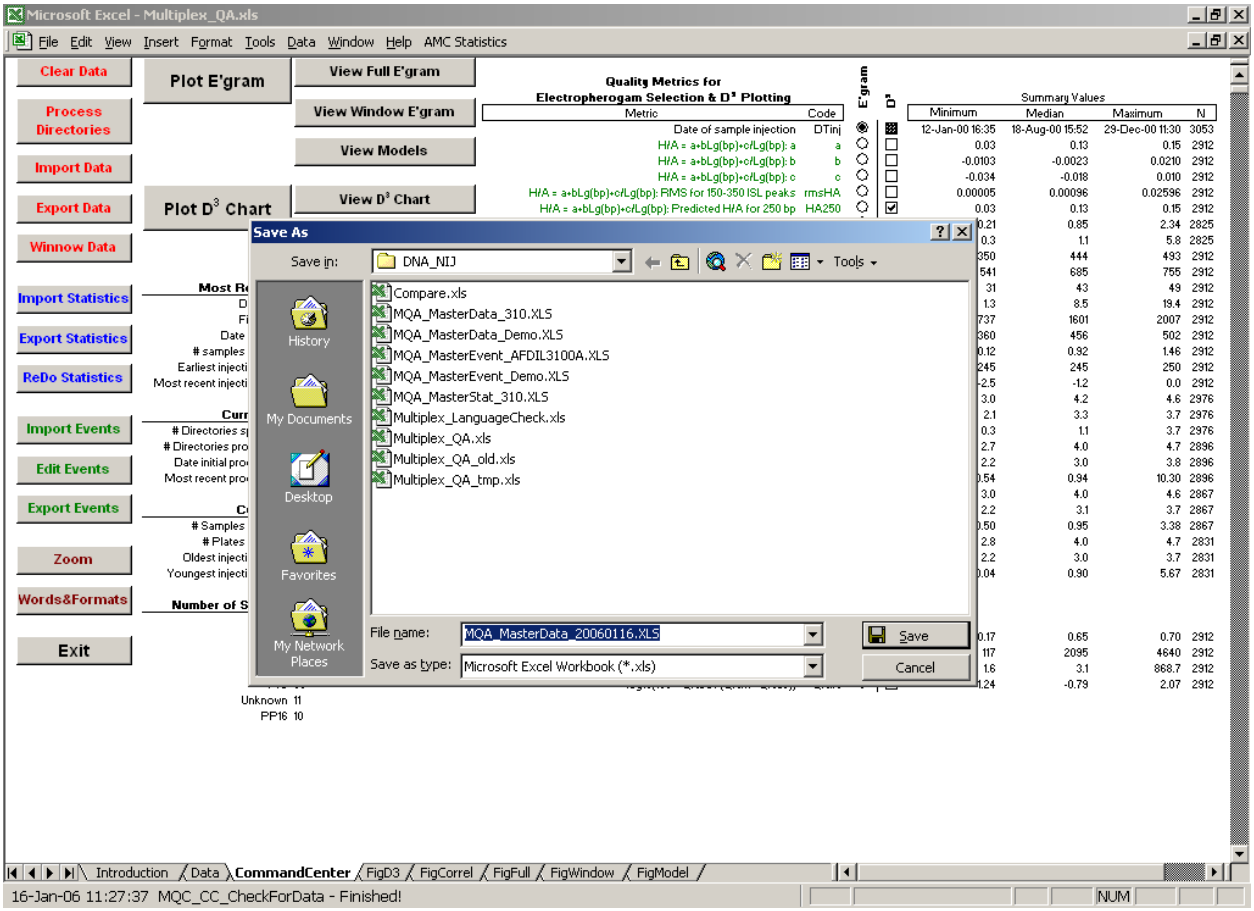


Figure 21 displays the "Save As" box used to specify a filename and directory for the to-be-stored Data and its associated worksheets. If you wish to use the suggested name in the currently active directory, click **OK**. If you wish to specify a different filename, type the name into the dialog box's "File name" input field and click **Save** or hit the <Enter> key. If you want to save the file in a directory other than the currently active one, use the Save As dialog box's navigation functions to move to the desired directory.

If there is already a file of the same name in the active directory, you will be asked if you wish to replace the old file with the new. If you wish to replace the file, click **Yes**; if you do not, click **No** and you will be returned to the Save As dialog. Since the Data and its associated worksheets have been removed from the Multiplex_QA workbook before you see this dialog, clicking **Cancel** returns to the Save As dialog: to continue, you *must* save the data under some name, in some directory.

Figure 21. Export Data, Specifying the Filename and Location



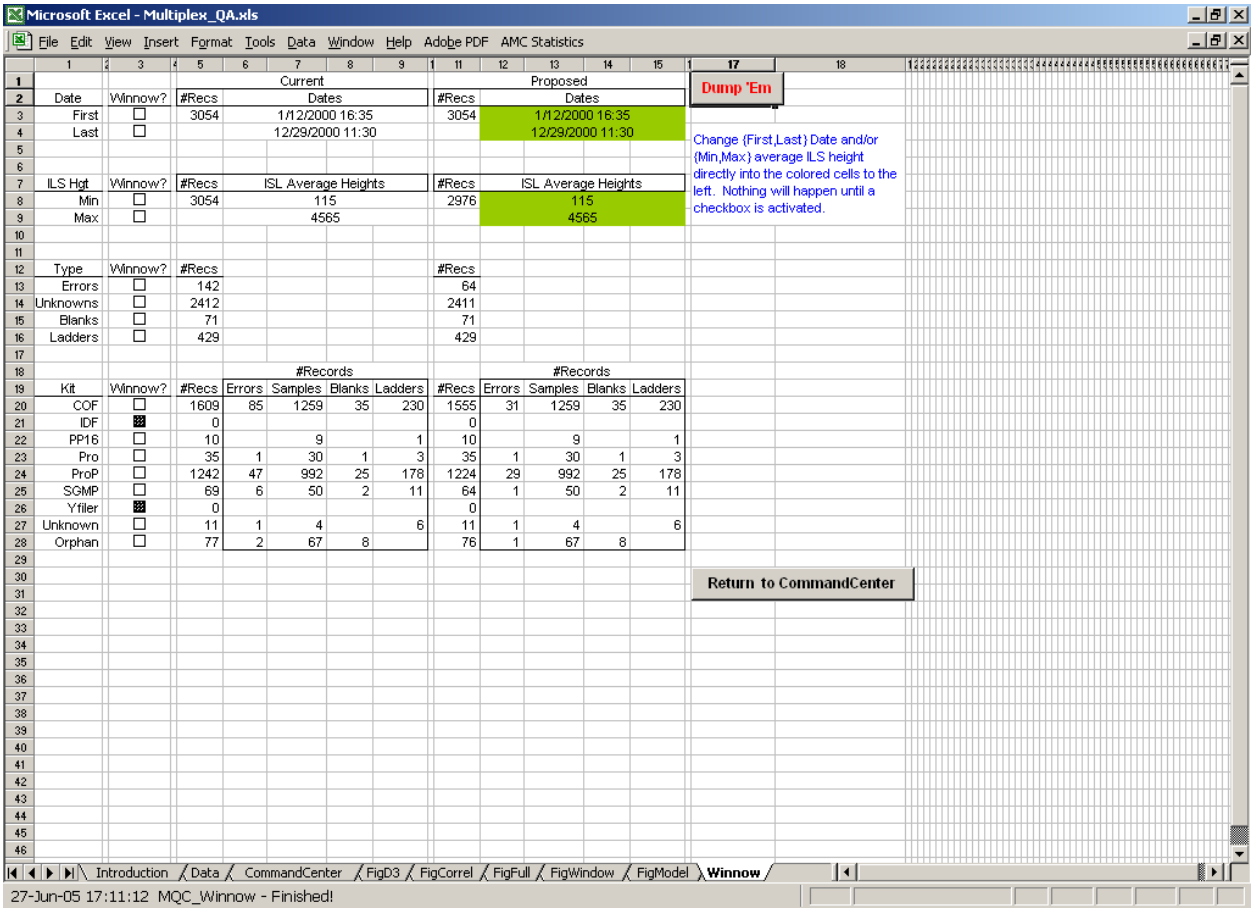
A note of caution: The Data and its associated worksheets are moved whole-cloth from the Multiplex_QA system to the new workbook. As the dialog says in Figure 20, you will need to **Process Directories** or **Import Data** to proceed. If you want to continue working with the same dataset as you just saved, read it back in with **Import Data**.

A note of explanation: The Data and associated worksheets are "moved" rather than "copied" to minimize frustration. The Data sheet in particular can be quite big, and the "Copy" part of Excel's Edit>Move or Copy Sheet function appears (on empirical evidence – I don't claim to understand why) much less efficient than the "Move" part for such large sheets.

2.4.5 Winnow Data

Clicking the **Winnow Data** button enables you to selectively eliminate data from the current dataset on the basis of four sample attributes: injection datetime, average height of the ISS peaks, sample error status and Type, and assigned Kit. Figure 22 displays an example of the normally hidden Winnow worksheet used to select the data to be removed. Any one to all of the four attributes can be used at the same time.

Figure 22. Winnow Data



Date	Winnow?	#Recs	Dates	#Recs	Dates
First	☐	3054	1/12/2000 16:35	3054	1/12/2000 16:35
Last	☐		12/29/2000 11:30		12/29/2000 11:30

ILS Hgt	Winnow?	#Recs	ISL Average Heights	#Recs	ISL Average Heights
Min	☐	3054	115	2976	115
Max	☐		4565		4565

Type	Winnow?	#Recs	#Recs
Errors	☐	142	64
Unknowns	☐	2412	2411
Blanks	☐	71	71
Ladders	☐	429	429

			#Records				#Records				
Kit	Winnow?	#Recs	Errors	Samples	Blanks	Ladders	#Recs	Errors	Samples	Blanks	Ladders
COF	☐	1609	85	1259	35	230	1555	31	1259	35	230
IDF	▩	0					0				
PP16	☐	10		9		1	10		9		1
Pro	☐	35	1	30	1	3	35	1	30	1	3
ProP	☐	1242	47	992	25	178	1224	29	992	25	178
SGMP	☐	69	6	50	2	11	64	1	50	2	11
Yfiler	▩	0					0				
Unknown	☐	11	1	4		6	11	1	4		6
Orphan	☐	77	2	67	8		76	1	67	8	

The total number of samples in the current dataset that could be winnowed by a given attribute is listed in the corresponding row of column 5. For the injection datetime and ISS peak height attributes, the number of samples that are not winnowed by the specified parameters are listed in column 11. For the sample Type and Kit attributes, the number of samples that can be winnowed after application of all current winnowing attributes is listed in column 11. No change is made to the current dataset until the **Dump 'Em** button is clicked.

2.4.5.1 Winnowing by Injection Datetime

The injection datetime winnowing can be accomplished using the entry fields (colored cells) and checkboxes at the top of the Winnow worksheet. To winnow samples that were injected before a given datetime, a datetime later than the earliest injection datetime in the current dataset should be entered into the cell at the intersection of the "First" row and the "Proposed" column (the first colored cell of column 12) of the Winnow worksheet. To winnow samples injected after a given datetime, similarly a datetime before the latest datetime in the current data set should be entered into the cell at the intersection of the "Last" row and the "Proposed" column (the second colored cell of column 12).

Nothing will happen until one of the checkboxes in column 3 is clicked. When a checkbox is clicked, the number of samples that were injected during the period between the specified First

and Last datetimes will be displayed. This number of remaining samples is unaffected by any of the other winnowing criteria.

2.4.5.2 Winnowing by Average ISS Peak Height

The average ISS peak height winnowing can be accomplished using the second set of entry fields (colored cells) and checkboxes, just below those for injection datetime. To winnow samples with an average ISS peak height less than a given value, a height larger than the smallest non-zero height in the current dataset should be entered into the cell at the intersection of the "Min" row and the "Proposed" column (the third colored cell of column 12) of the Winnow worksheet. To winnow samples with an average ISS height greater than a given value, similarly a height less than the maximum height in the current data set should be entered into the cell at the intersection of the "Max" row and the "Proposed" column (the fourth colored cell of column 12). Nothing will happen until one of the checkboxes in column 3 is clicked.

When a checkbox is clicked, the number of samples with average ISS heights in the interval between the specified Min and Max values will be displayed. This number of remaining samples is unaffected by any of the other winnowing criteria.

2.4.5.3 Winnowing by Sample Type

Sample Type winnowing is accomplished using the third set of checkboxes, just below those for average ISS peak height. Samples that are flagged as unusable (Err flag less than 0, Section 2.4.2.6.1) or are assigned sample Types (Section 2.4.2.6.2) of "Unknown" (Type = 1), "Blank" (Type = 2), or "Ladder" (Type = 3 or 4) can be winnowed by clicking the corresponding checkboxes.

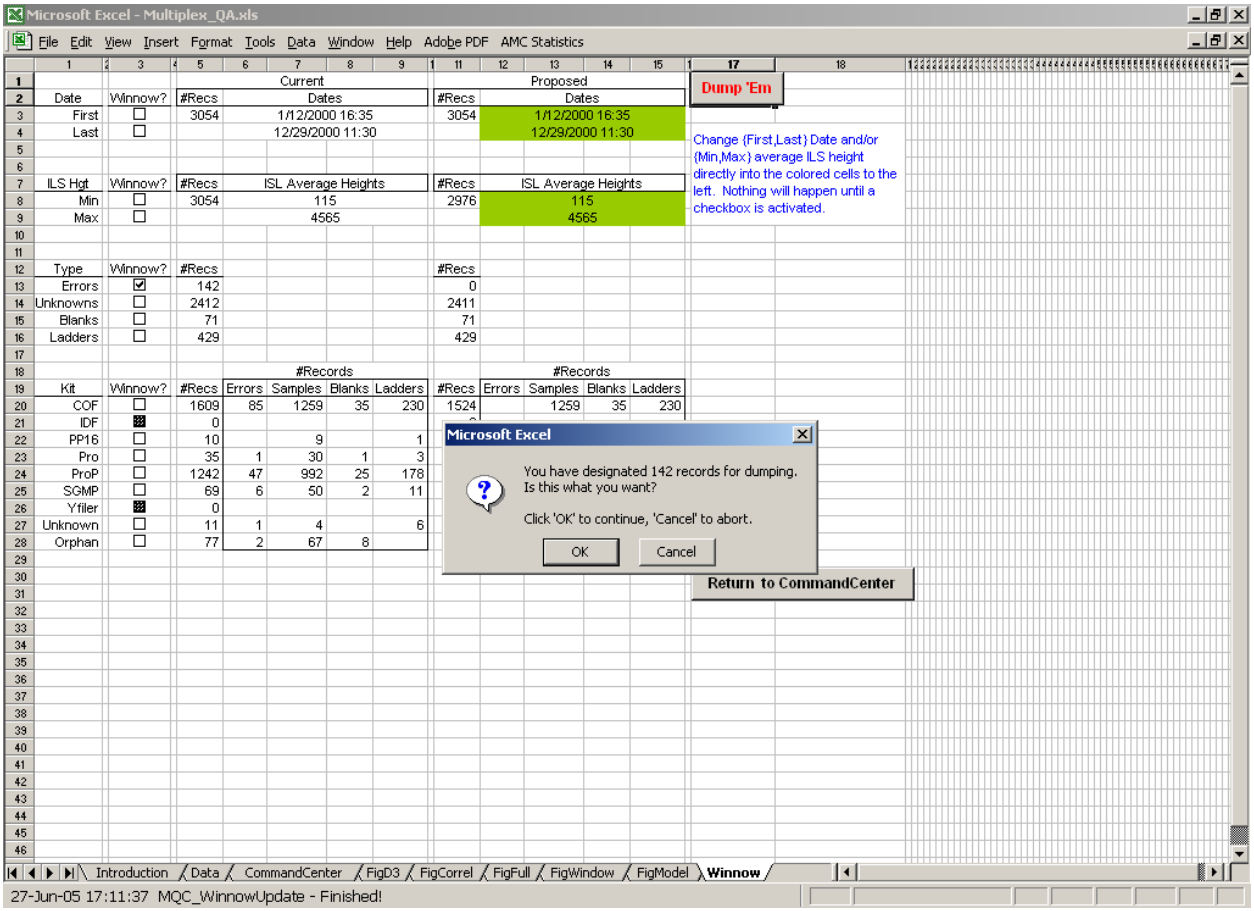
2.4.5.4 Winnowing by Kit

Sample Kit winnowing is accomplished using the last set of checkboxes. A separate checkbox is provided for the multiplex Kits listed in Table 5, along with those samples assigned to an "Unknown" Kit and those assigned as "Orphan" (Section 2.4.2.6.4).

2.4.5.5 Dump 'Em

While a summary of the number of samples that should remain after winnowing is issued whenever one of the checkboxes to the left of the worksheet is clicked, the current Data worksheet is not modified until **Dump 'Em** is clicked. When clicked, a confirmation dialog box similar to that displayed in Figure 23 will state the number of samples that will be winnowed from the current dataset and ask if you really want this to happen. Clicking **OK** will winnow these data; clicking **Cancel** will return you to the Winnow worksheet.

Figure 23. Winnow Data, Confirmation



2.4.5.6 Return to CommandCenter

Clicking **Return to CommandCenter** returns you to the CommandCenter worksheet and sets the "Hide" status of all Multiplex_QA worksheets to their default setting. See Section 11.1 for further information.

2.5 Statistics Commands

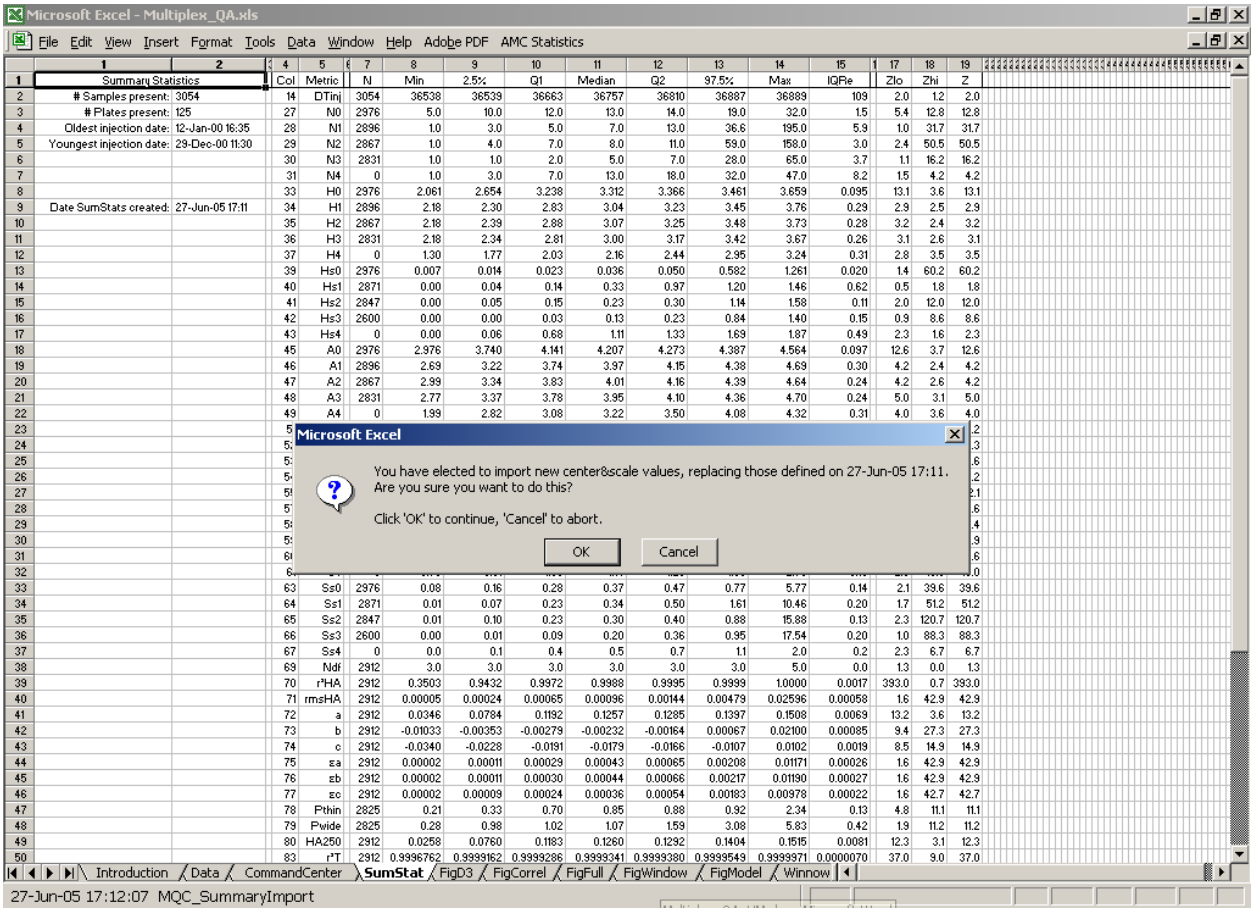
The QA metrics displayed in D^3 charts and correlation scattergrams (Sections 5 and 6) are standardized (centered and scaled) using the summary center and scale statistics stored in the normally hidden worksheet "SumStat." These center and scale parameters are automatically defined whenever a new data set is brought into Multiplex_QA. Most of the time, the D^3 and correlation displays are appropriately standardized using the dataset's own parameters. However, it is sometimes useful to display the data in a particular dataset using parameters from a different dataset, e.g., when comparing two or more charts or scattergrams for different time periods, instruments, or multiplex kits. The following three commands enable you to save and re-use particular sets of center and scale parameters.

2.5.1 Import Statistics

Clicking the **Import Statistics** button replaces the current center and scale parameters held on the normally hidden worksheet SumStat with those from a "MasterStat" Excel workbook. For ease in identifying them, the MasterStat workbooks are by default named "MQA_MasterStat_*.xls," where the "*" represents the workbook's creation date. These files can be renamed as you like, but remembering which contains what is then your problem. No matter how the workbook is named, the worksheet containing the summary statistics must be named "SumStat."

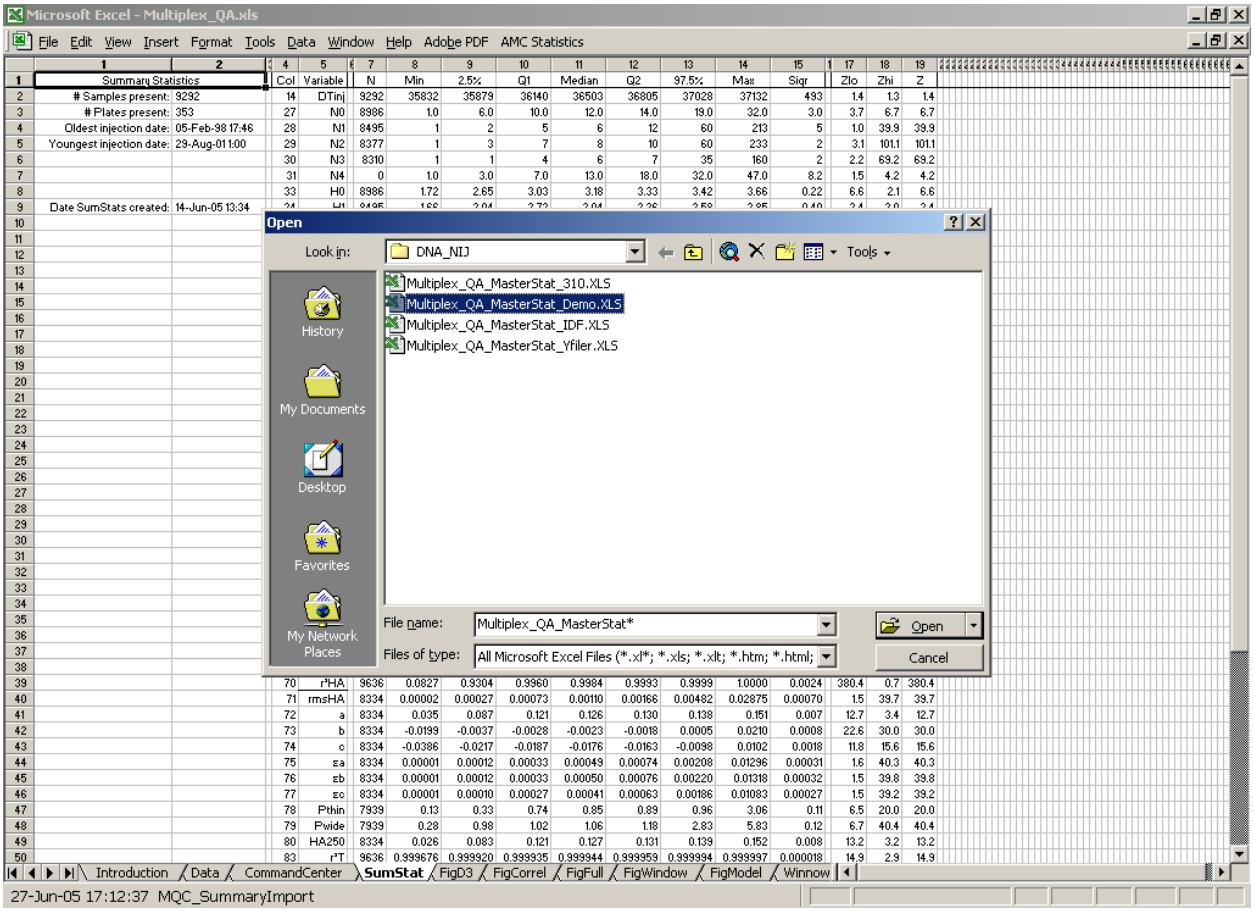
Figure 24 displays both the structure of a SumStat worksheet and the dialog box that appears when **Import Statistics** is clicked. If you wish to replace the current summary statistics, click **OK**; if you do not, click **Cancel**.

Figure 24. Import Statistics, Display of Statistics and Confirmation of Intent



If you confirm that you do indeed want to import a new set of center and scale parameters, you will be presented with a standard Excel file "Open" dialog box similar to that shown in Figure 25. The Open dialog box will list all of the files that match the "MQA_MasterStat*.xls" wildcard search in the currently active directory (plus a list of all of the folders in that directory). You can select one of the MQA_MasterStat*.xls files by clicking on the listed name and then clicking **Open**.

Figure 25. Import Statistics, Specifying the MasterStat File



If the file you want is named something that is not matched by the wildcard search specified in the "File name" input field, you can type a specific filename or a new wildcard search into the input field followed by clicking **Open** or hitting <Enter>. If a specific file was named, it will be opened and the center and scale parameters transferred to the SumStat worksheet. If a new wildcard search was specified, everything that matches that search will be listed for your perusal.

If the file you want isn't in the currently active directory, you can use the Open dialog box's navigation functions to move to the desired directory.

Clicking **Cancel** returns you to the CommandCenter, leaving the original center and scale parameters on the SumStat worksheet.

2.5.2 Export Statistics

Clicking the **Export Statistics** button saves the current summary center and scale statistics held on the normally hidden SumStat worksheet. Figure 26 displays both the structure of a SumStat worksheet and the Alert box that appears when **Export Statistics** is clicked. If you wish to save the current parameters, click **OK**; if you do not, click **Cancel**.

Figure 26. Export Statistics, Confirmation

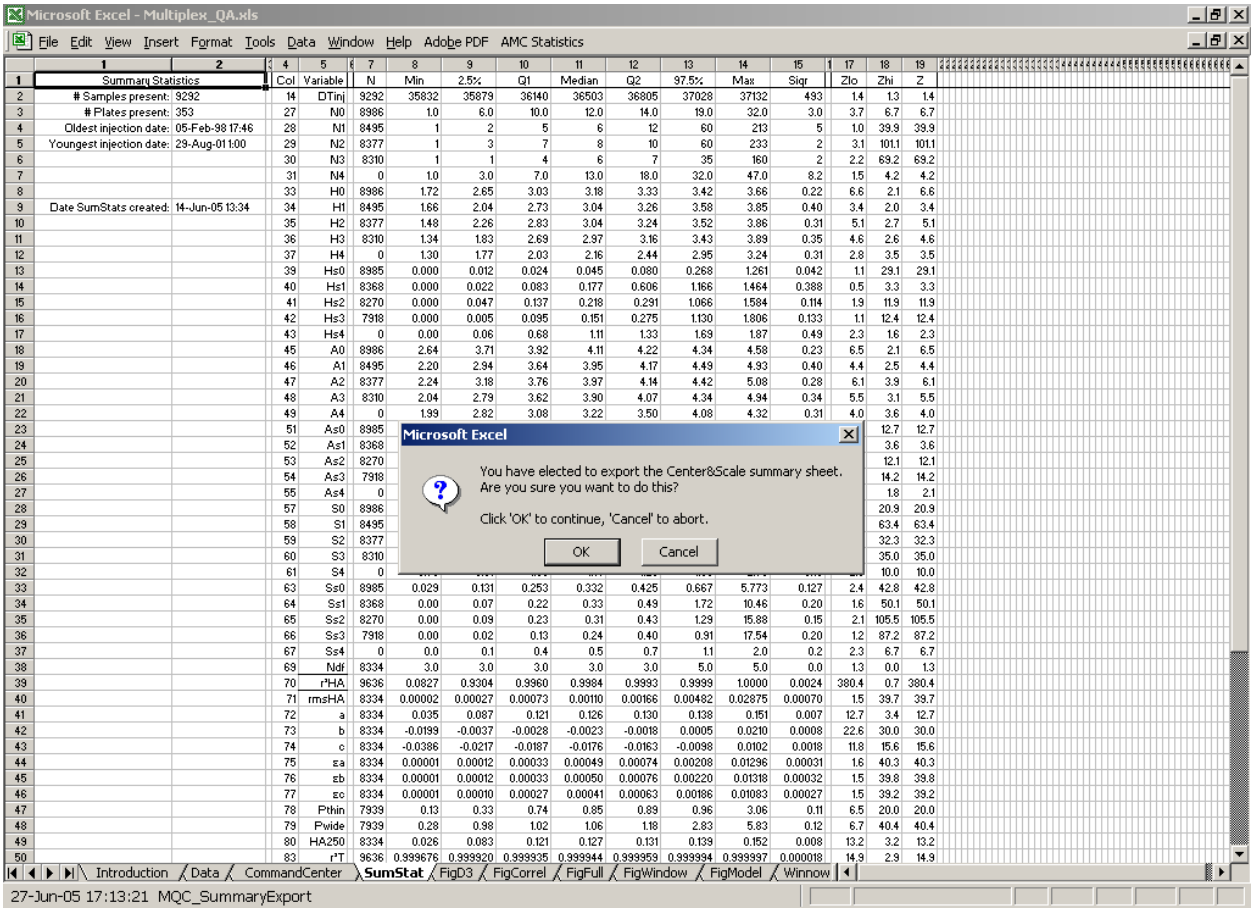


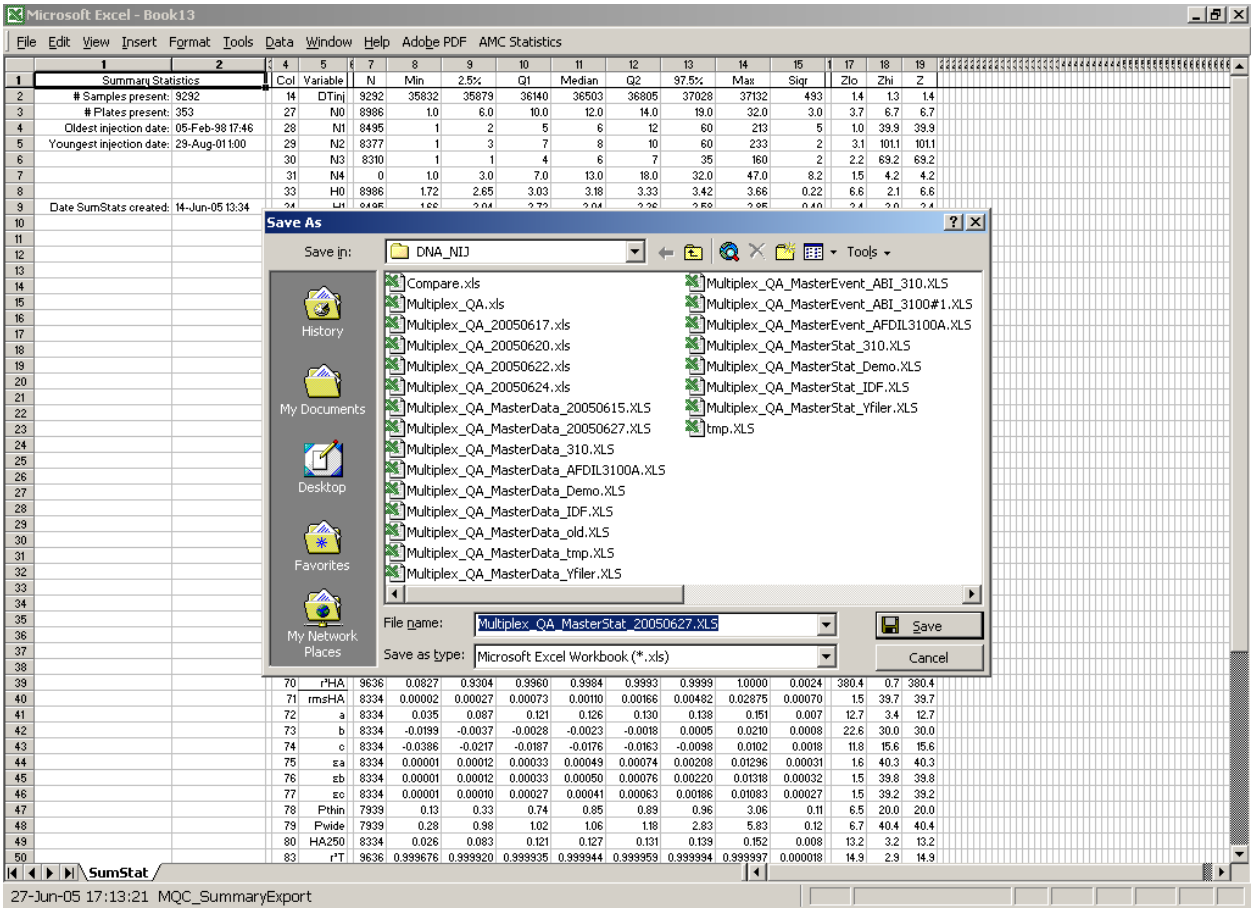
Figure 27 displays the "Save As" box used to specify a filename and directory for the to-be-stored center and scale parameters. If you wish to use the suggested name in the currently active directory, click the **OK**. If you wish to specify a different filename, type the name into the "File name" input field and click **Save** or hit the <Enter> key. If you want to save the file in a directory other than the currently active one, use the Save As dialog box's navigation functions to move to the desired directory.

If there is already a file of the same name in the active directory, you will be asked if you wish to replace the old file with the new. If you wish to replace the file, click **Yes**; if you do not, click **No** and you will be returned to the Save As dialog. Clicking **Cancel** returns you to the CommandCenter without creating a new MasterStat workbook.

2.5.3 ReDo Statistics

Clicking the **ReDo Statistics** button replaces the current center and scale parameters held on the normally hidden worksheet SumStat with values calculated directly from the current dataset. **ReDo Statistics** allows you make sure that the center and scale parameters are appropriate to the dataset after **Winnow Data** (Section 2.4.5) has been used. The dialog box that appears when you click **ReDo Statistics** is very similar to that displayed in Figure 24.

Figure 27. Export Statistics, Specifying the Filename and Location



2.6 Event Commands

The D3 chart (see Section 5) can mark the dates when "events of interest" took place; e.g., instrument repair, software upgrade, or column replacement. The list for a particular dataset is kept on the normally hidden worksheet "PEvents." The following commands enable you to save, modify, and re-use event files. It is your responsibility to identify the interesting events and to create and maintain the files for your instrument(s).

A note of caution: Multiplex_QA does **not** check that the events on the PEvents worksheet are associated with the dataset on the Data worksheet! This isn't so much an oversight as that I haven't figured out how to do it.

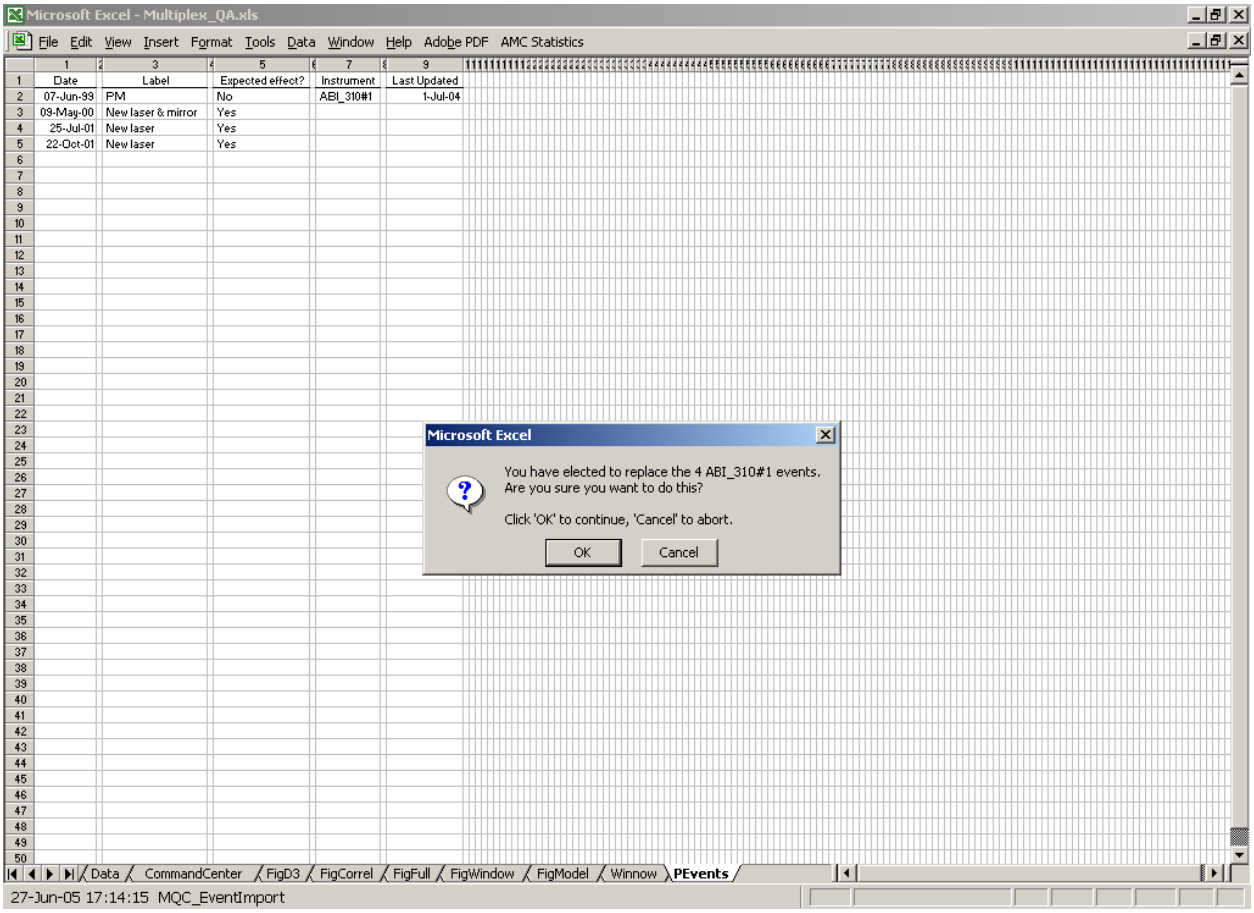
2.6.1 Import Events

Clicking the **Import Events** button replaces the current list of "events of interest" held on the normally hidden worksheet PEvents with those from a "MasterEvent" Excel workbook. For ease in identifying them, the MasterEvent workbooks are by default named "MQA_MasterEvent_*.xls" where the "*" represents the instrument name assigned by the user in Row 2 of column 7 of the PEvent worksheet. These files can be renamed as you like, but

remembering which contains what is then your problem. No matter how the workbook is named, the worksheet containing the events must be named "PEvents."

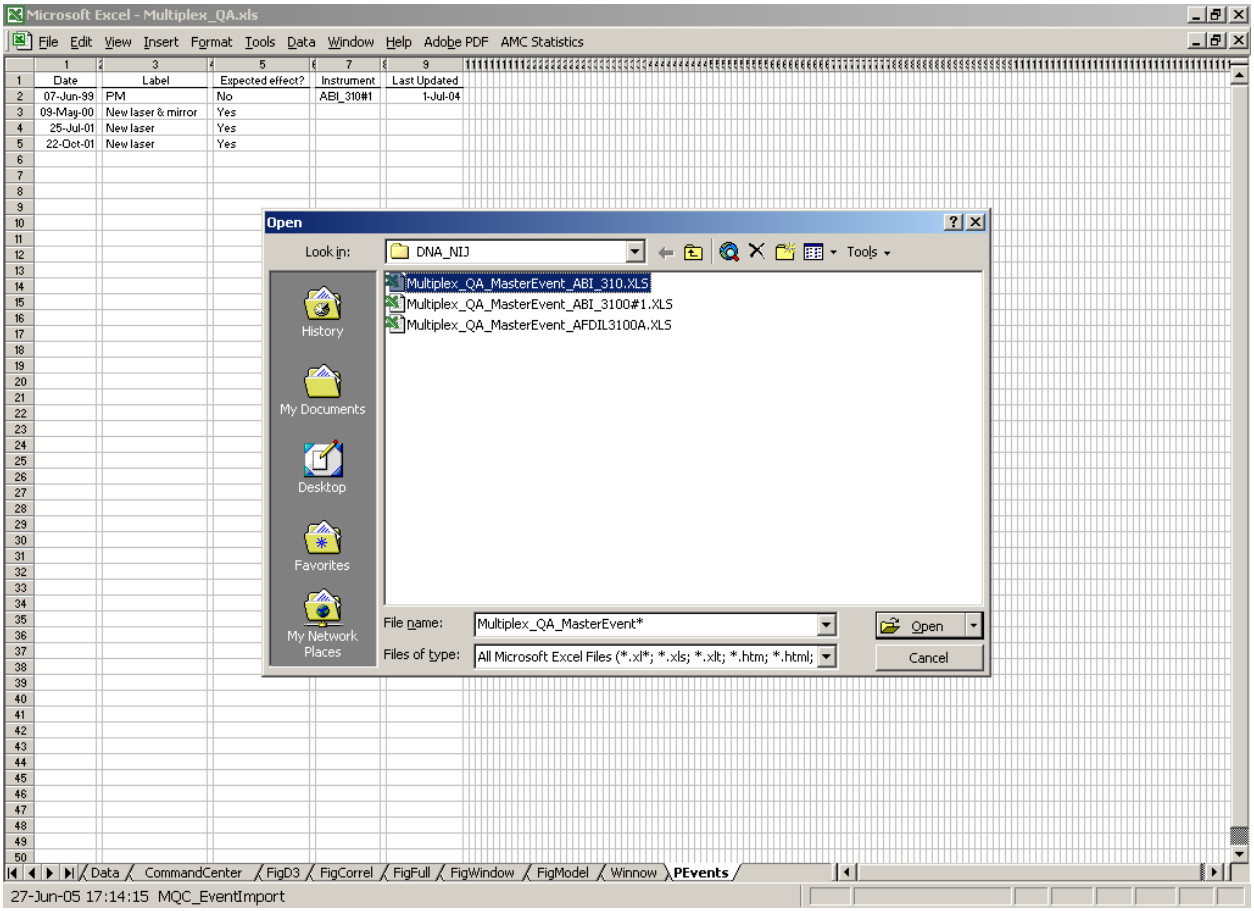
Figure 28 displays both the structure of a PEvents worksheet and the dialog box that appears when **Import Events** is clicked. If you wish to replace the current list of events, click **OK**; if you do not, click **Cancel**.

Figure 28. Import Events, Display of Events and Confirmation of Intent



If you confirm that you do indeed want to import a new list of events, you will be presented with a standard Excel file "Open" dialog box similar to that shown in Figure 29. The Open dialog box will list all of the files that match the "MQA_MasterEvent*.xls" wildcard search in the currently active directory (plus a list of all of the folders in that directory). You can select one of the MQA_MasterEvent*.xls files by clicking on the listed name and then clicking **Open**.

Figure 29. Import Events, Specifying the MasterEvent File



If the file you want is named something that is not matched by the wildcard search specified in the "File name" input field, you can type a specific filename or a new wildcard search into the input field followed by clicking **Open** or hitting <Enter>. If a specific file was named, it will be opened and the events transferred to the PEvents worksheet. If a new wildcard search was specified, everything that matches that search will be listed for your perusal.

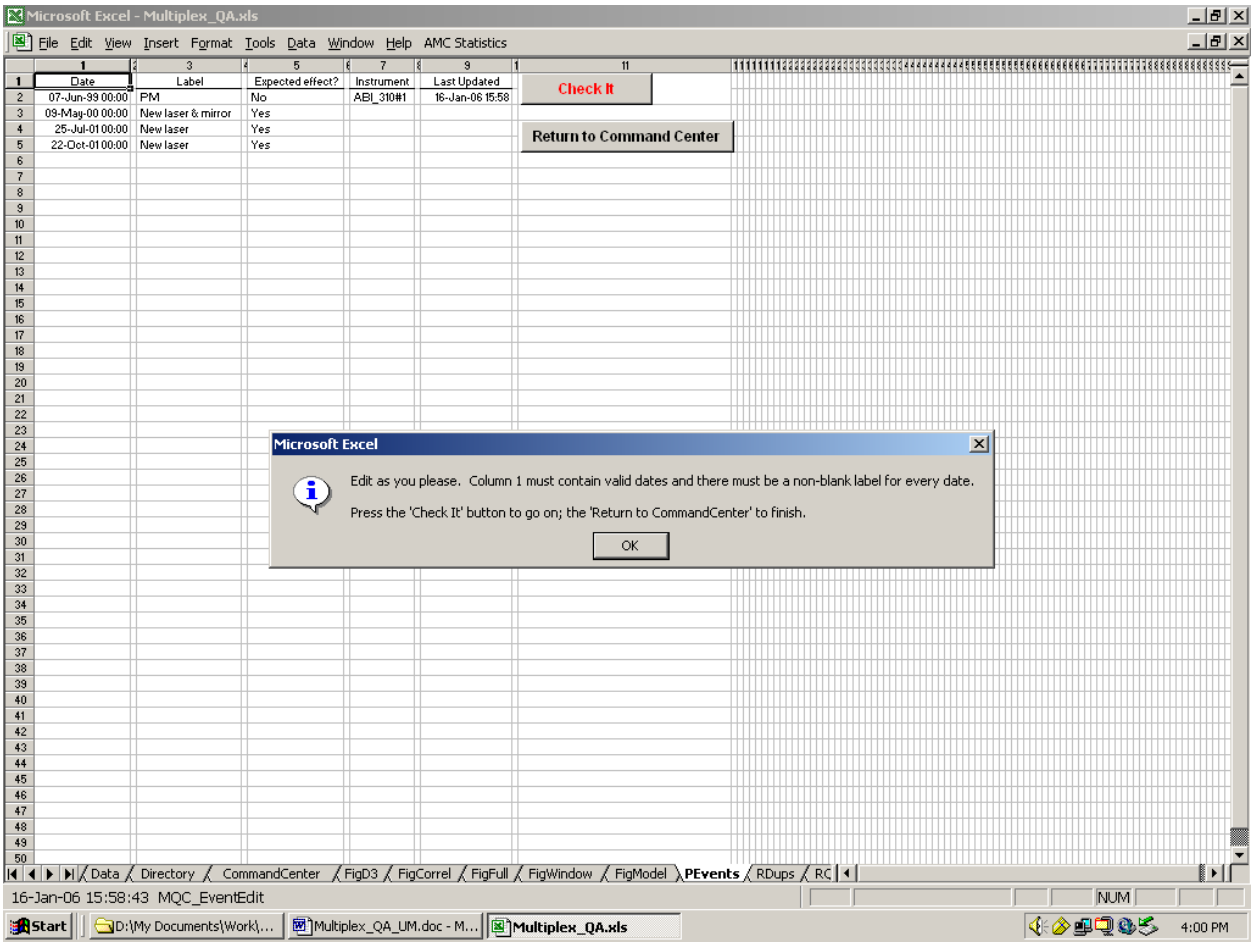
If the file you want isn't in the currently active directory, you can use the Open dialog box's navigation functions to move to the desired directory.

Clicking **Cancel** returns you to the CommandCenter, leaving the original list of events on the PEvents worksheet.

2.6.2 Edit Events

Clicking the **Edit Events** button "unhides" the PEvents worksheet and makes it available for you to edit as you please. Figure 30 is an example of the initial prompt message.

Figure 30. Edit Events



2.6.2.1 Parameters

The only event parameters in current use are a list of datetimes in column 1 and their associated labels in column 3. Datetimes may be specified as anything interpretable as a datetime in the range of 10 years before the current date to 1 year after the current date. Event labels should be meaningful to you while being as terse as practical; every label must consist of at least one non-blank alphanumeric character.

Column 5 allows you to "go on record" about whether the event had significant impact on any of the Multiplex_QA quality metrics, but it is not otherwise currently used. You may designate the particular instrument associated with this particular list of events in row 2 of column 7; however, this Instrument label is not currently used. (If you leave this cell blank, it will be assigned the value "Unspecified."). The current datetime is automatically stored in row 2 of column 9; this value is not currently used.

2.6.2.2 Check It

Clicking **Check It** assures that there is a non-blank instrument label, sorts the events by Date, evaluates if all the datetimes in column 1 are valid, and checks that there is an associated non-blank label for each specified datetime. If an invalid datetime is specified, you will be

prompted to either delete or correct the value. If a valid datetime does not have an associated non-blank event label, you will be prompted to provide a "useful" label.

2.6.2.3 Return to CommandCenter

Once valid datetimes and event labels have been specified, clicking **Return to CommandCenter** returns you to the CommandCenter worksheet. It also resets the "Hide" status of all Multiplex_QA worksheets to their default setting. See Section 11.1 for further information.

2.6.3 Export Events

Clicking the **Export Events** button saves the current list of "events of interest" held on the normally hidden PEvents worksheet. Figure 31 displays both the structure of a PEvents worksheet and the Alert box that appears when **Export Events** is clicked. If you wish to save the current events, click **OK**; if you do not, click **Cancel**.

Figure 31. Export Events, Confirmation of Intent

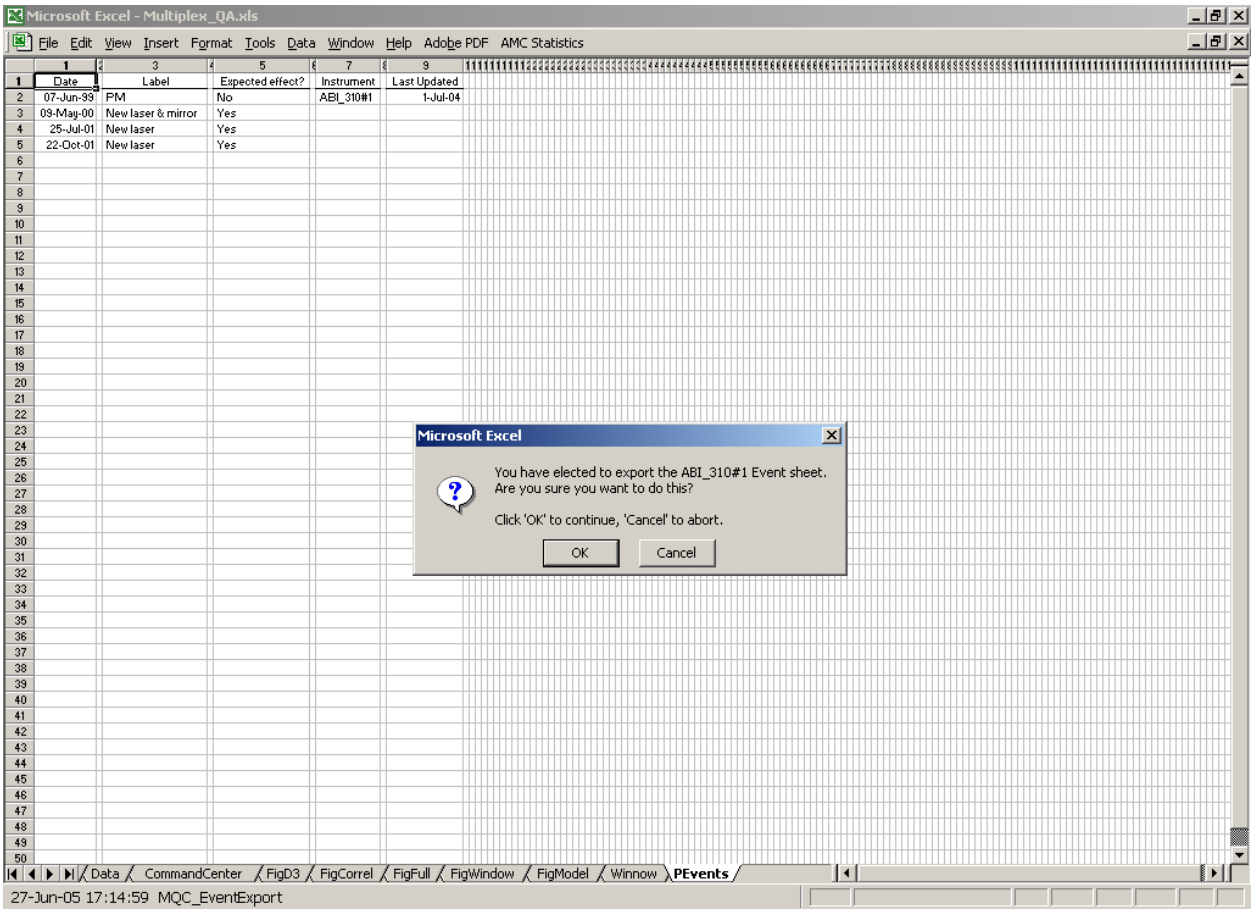
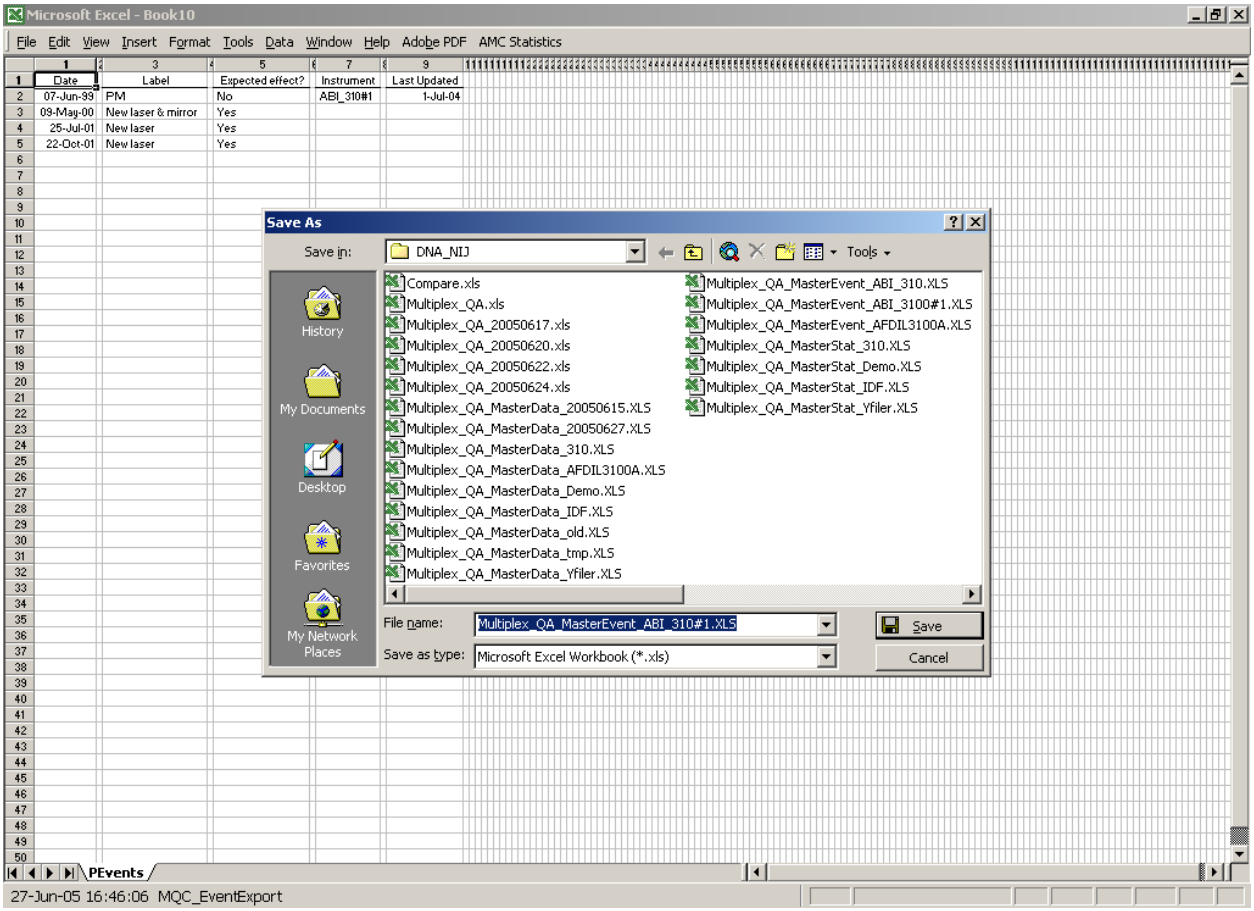


Figure 32 displays the "Save As" box used to specify a filename and directory for the to-be-stored list of events. If you wish to use the suggested name in the currently active directory, click **OK**. If you wish to specify a different filename, type the name into the dialog box's "File

name" input field and click **Save** or hit the <Enter> key. If you want to save the file in a directory other than the currently active one, use the Save As dialog box's navigation functions to move to the desired directory.

If there is already a file of the same name in the active directory, you will be asked if you wish to replace the old file with the new. If you wish to replace the file, click **Yes**; if you do not, click **No** and you will be returned to the Save As dialog. Clicking **Cancel** returns you to the CommandCenter without creating a new MasterEvent workbook.

Figure 32. Export Events, Specifying the Filename and Location



2.7 Workspace Commands

The following two commands provide some control of the Excel workspace.

2.7.1 Zoom

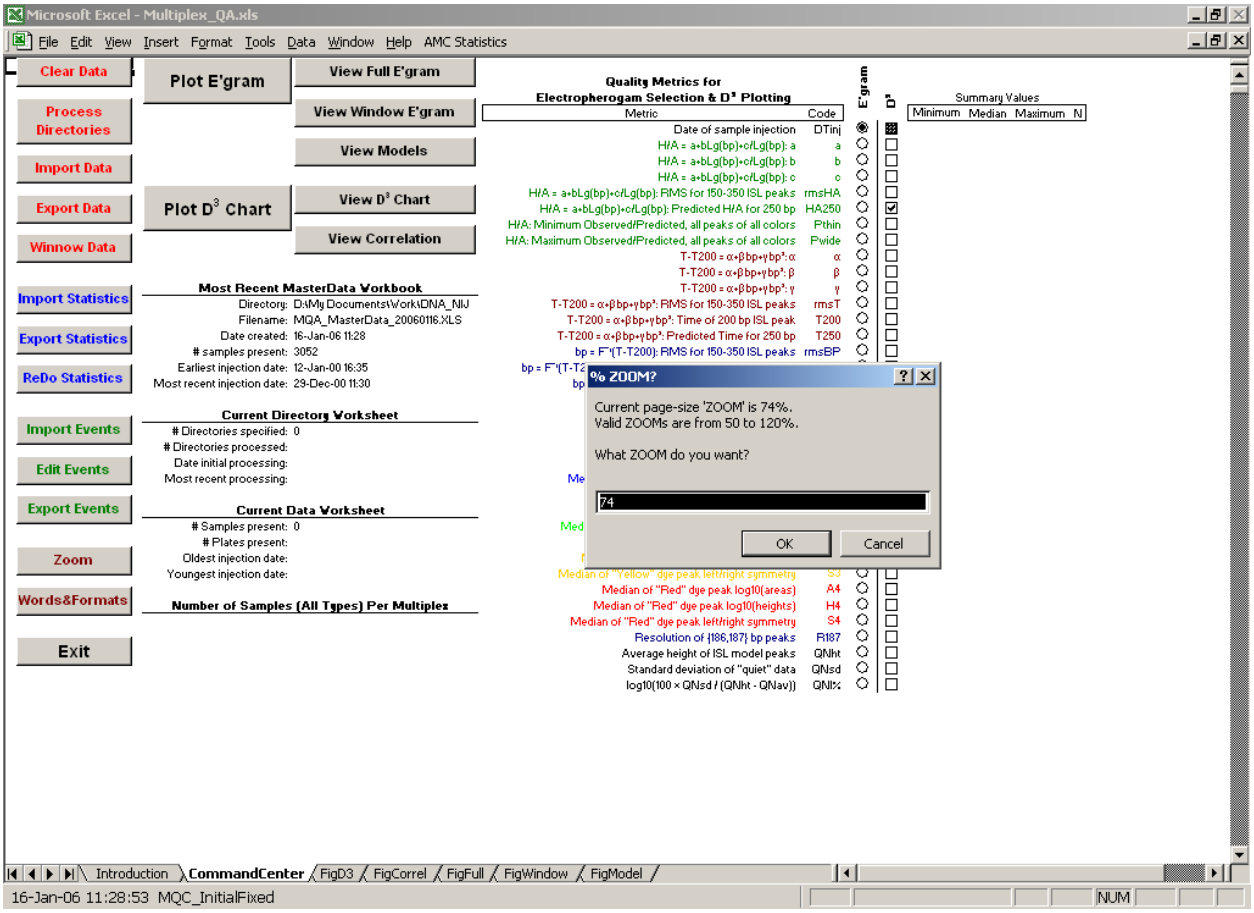
Because of the diversity in the physical size and resolution settings of computer displays, there is no one "best" setting for the "zoom" or display scale of the active worksheet in the fully expanded Multiplex_QA window. You can, of course, set the zoom using Excel's View>Zoom menu or toolbar commands (or, if using a suitably enabled wheel-mouse, by turning the wheel

with the <Ctrl> key held down). These changes, however, will probably need to be made for all the worksheets visited, every time the Multiplex_QA system is started.

Clicking the **Zoom** button allows you to set the zoom for all of Multiplex_QA worksheets at one time. All of the normally displayed Multiplex_QA worksheets are designed to occupy nearly the same visible space, so this "one-size-fits-all" zoom should be adequate. The initial zoom is 74%; the accepted values are currently 50% to 120%.

Figure 33 displays the dialog box used to establish the zoom. If you type in a new zoom value, the new zoom will be applied to the CommandCenter worksheet and the dialog box will reappear in the re-sized window. Since it often takes a few trials to find the best setting, the dialog will keep reappearing until you click **OK** without changing the zoom setting or you click **Cancel**. Clicking **OK** applies the new zoom value to all Multiplex_QA worksheets. Clicking **Cancel** applies the original zoom value to all worksheets.

Figure 33. Zoom

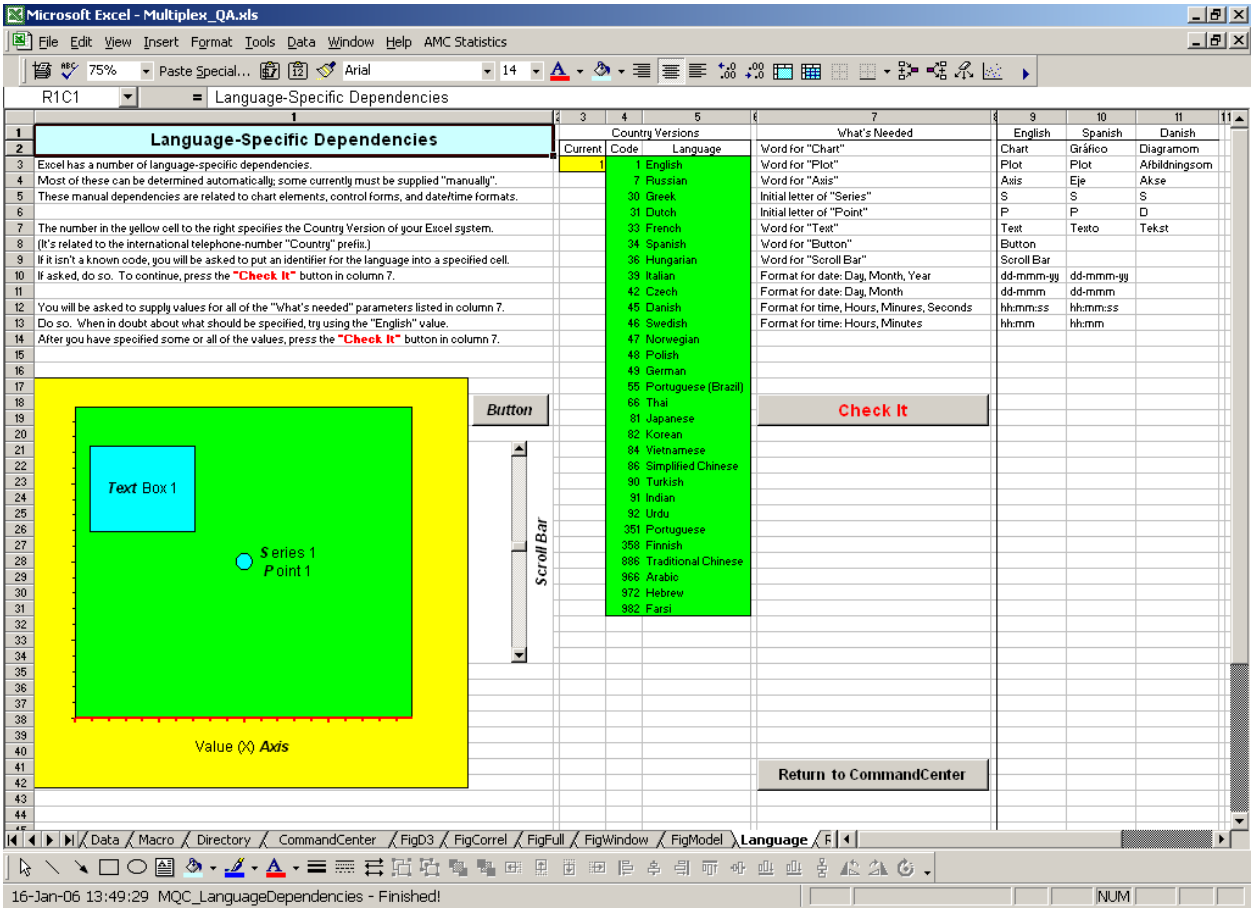


Once a new zoom is established, saving the Multiplex_QA system establishes the new zoom as the default setting. Clicking **Zoom** also resets the "Hide" state of all Multiplex_QA worksheets to their default setting. See Section 11.1 for further information.

2.7.2 Words&Format

There are a number of parameters used in the Multiplex_QA system that are specific to particular "Country Version's of Excel (*e.g.*, the words for various graphical components and control forms) or differ among cultures (*e.g.*, standard formats for date and time). Multiplex_QA stores these values on the Language worksheet and checks whether valid values are available at the start of each session. If you are using a non-English Country Version, you may be confronted with something similar to Figure 34 the first time you try to use Multiplex_QA. However, since this automatic process will occur at most once for any given user, the **Words&Formats** command is provided to allow you to access the worksheet and customize the date and time formats.

Figure 34. Language-Specific Dependencies



2.7.2.1 Country Version name

A number of Country Versions of the Excel system are extant. Excel references them internally with an integer code roughly corresponding to the international telephony prefix. The code for the version you are using is automatically determined and stored in row 2 column 3 of the Language worksheet (color-coded yellow). The codes and their corresponding (English-

language) references for a number of the extant Country Versions are stored in columns 4 and 5 (color coded green).

If the code for the Country Version you are using is not included in column 4, the code will be appended to the list and you will be asked to specify a name for the code in column 5. This enables referencing the parameters by a meaningful name rather than just the arbitrary code.

If requested, specify a suitable name (at least two alphabetic characters) in the indicated row of column 5 and click the **Check It** command (2.7.2.2).

2.7.2.2 Parameter specification

Once a valid name for the Country Version code is available, a column to hold the parameter values for the Country Version will be appended to the currently specified versions (currently, only the English version is fully specified). You will be asked to provide values for six components of a scattergram (in English: Chart, Plot, Axis, Series, Point, and Text), two control forms (in English: Button and Scroll bar), and four date or time formats ({day, month, year}, {day, month), {hour, minute, second}, and {hour, minute}). If you don't have a clue as to the appropriate value for a given parameter, use the "English" value. The validity of all of the values will be evaluated when you click the **Check It** command (2.7.2.2).

If any parameter is left unspecified or the assigned value is invalid, you will be prompted with advice on how to determine a valid value. Values for the needed words (or, in the case of "Series" and "Point", the initial letters of the words) can be found by clicking on the specified locations of the scattergram or control forms located in column 1. You may not get the "exact" value for the parameter by following the advice, but it should be close enough to figure out what the appropriate value should be. For example, the words "Chart Area" appear in the Formula Bar of the English Country Version when if you right-click in the yellow area of the scattergram as advised in the "Chart" prompt (Figure 35).

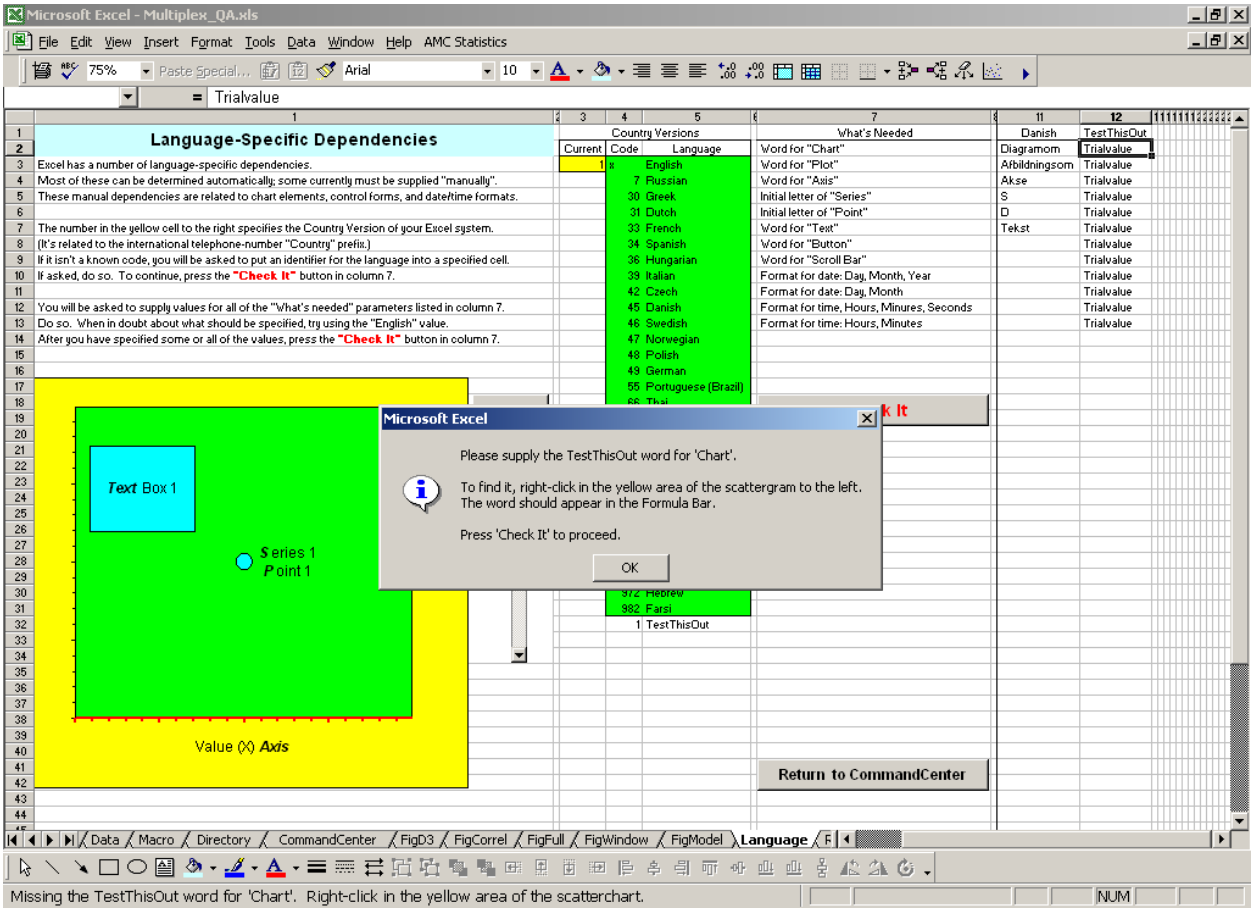
2.7.2.3 Check It

Clicking the **Check It** command evaluates whether valid values for all parameters (name for the Country Version and the 12 "What's Needed" words, letters, and formats). If the value for any parameter is left unspecified or is invalid, you will be prompted for what is required and where the value should be located.

2.7.2.4 Return to CommandCenter

Once valid values for all parameters have been specified, clicking **Return to CommandCenter** returns you to the CommandCenter worksheet. It also resets the "Hide" status of all Multiplex_QA worksheets to their default setting. See Section 11.1 for further information.

Figure 35. Example Prompt Message: Word for "Chart"

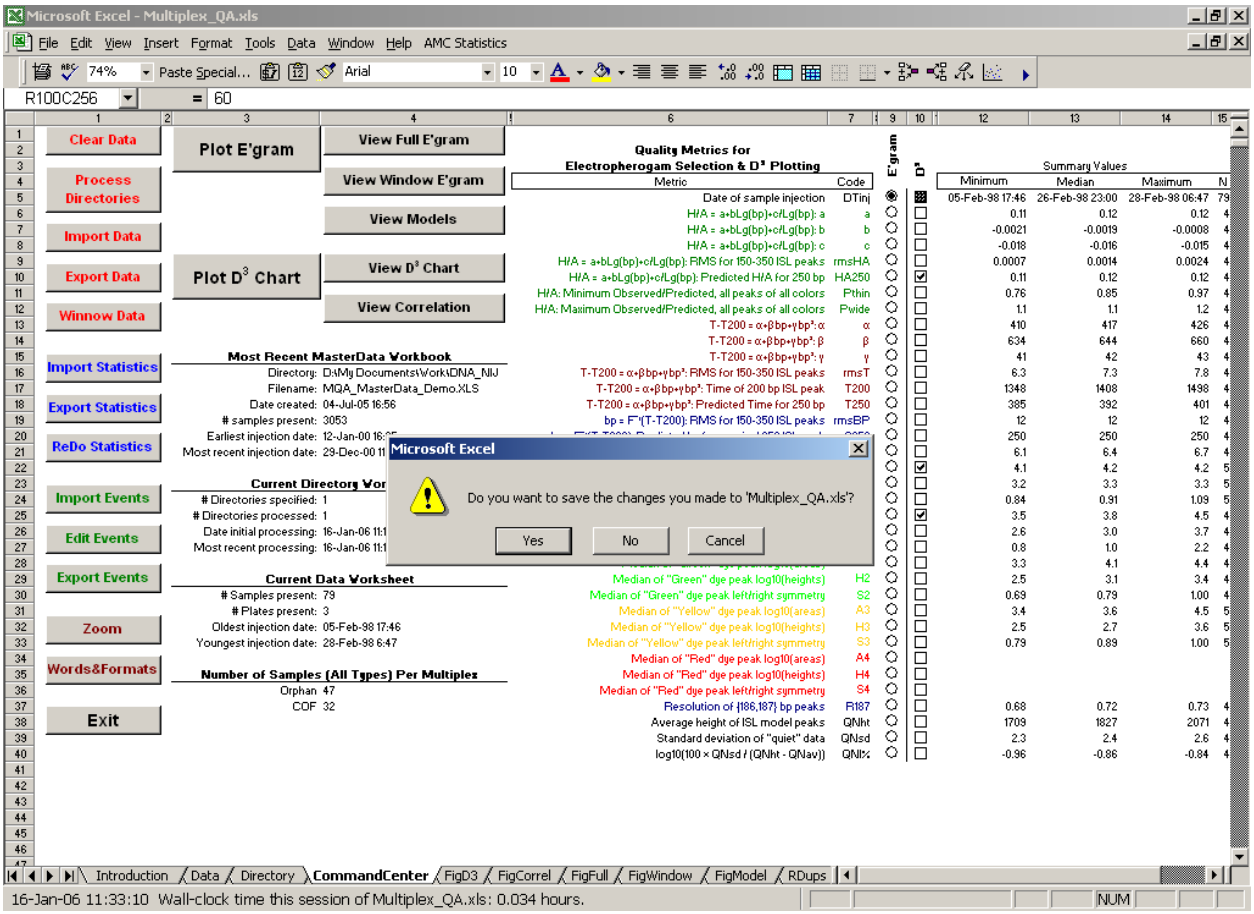


2.7.3 Exit

Clicking the **Exit** button restores Excel's calculation, address style, view, and toolbar options to the values in place at the time Multiplex_QA was opened. It also "cleans up" the temporary data used in the graphical displays, minimizing the disk space required if you chose to save the file.

Figure 36 displays the dialog box used to establish whether you wish the current Multiplex_QA workbook to replace the one you opened. Clicking **Yes** replaces the workbook and then causes Excel to exit. Clicking **No** causes Excel to quit without replacing the original workbook.

Figure 36. Exit, Saving Changes



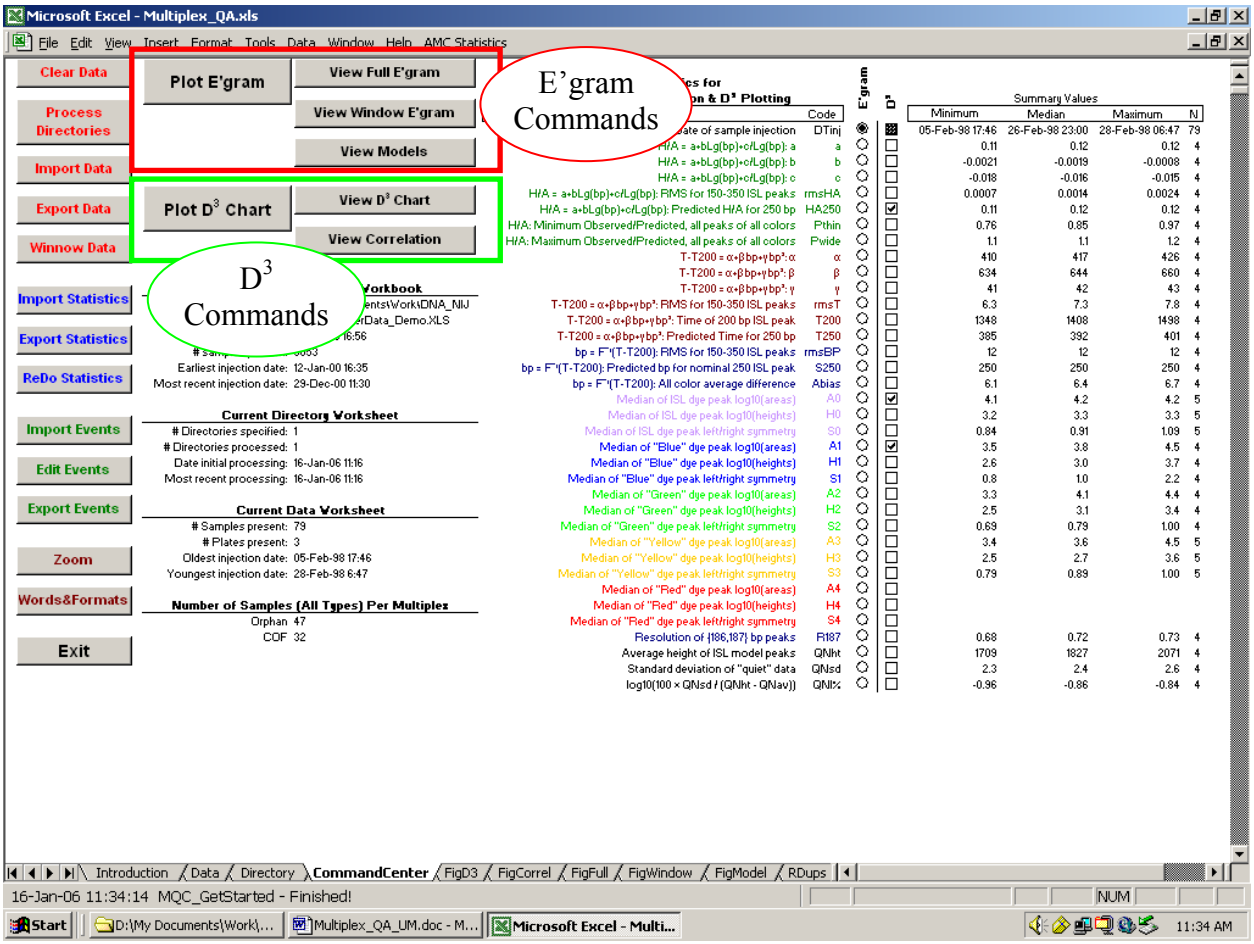
The **Cancel** option (equivalent to clicking the dialog box window's close box, see Section 1.5.2.2) re-establishes your Excel view and toolbar setting *without* closing the Multiplex_QA system. You will be "warned" that you have chosen not to exit. Your restored view and toolbar settings will then be used for the remainder of the Multiplex_QA session.

This sneakiness is primarily for my own convenience – it speeds retrieving all the tools I find convenient when debugging the system – but it may be useful if you find Multiplex_QA's choice of view to be inadequate. Remember (Section 1.3.3) that you can avoid Multiplex_QA changing your viewing options entirely by activating the CommandCenter worksheet yourself rather than clicking **Get Started!** on the Introduction worksheet.

2.8 Plot Commands

Two sets of buttons are located at the top center of the CommandCenter worksheet (Figure 37). The upper set controls access to the electropherogram (E'gram)-related functions. The lower set controls access to the D^3 Chart-related functions.

Figure 37. Location of CommandCenter Plot E'gram and D^3 Chart Commands



2.8.1 Plot E'gram

Clicking the **Plot E'gram** button initiates the electropherogram plotting functions described in Section 3. Electropherograms can only be plotted when the BatchExtract-ed files are available to the computer that is running the Multiplex_QA system. For the purposes of electropherogram plotting, the dataset ordering is controlled by the **E'gram** optionbuttons in the CommandCenter worksheet (Section 2.2.2).

2.8.1.1 View Full E'gram

If a complete electropherogram (a "Full E'gram")has been defined using the **Plot E'gram** button in the CommandCenter or the Plot Full commands in FigWindow, FigModel, or FigD3 worksheets, the **View Full E'gram** button will transfer control from the CommandCenter worksheet to the FigFull worksheet. If no Full E'gram has been defined, a warning dialog will appear. This dialog must be cleared before you can continue.

2.8.1.2 View Window E'gram

If a high-resolution chunk of an electropherogram (a "Window E'gram") has been defined using the **Plot Window** button on the FigFull, FigModel, or FigD3 worksheets, the **View Window E'gram** button will transfer control from the CommandCenter worksheet to the FigWindow worksheet. If no Window E'gram has been defined, a warning dialog will appear. This dialog must be cleared before you can continue.

2.8.1.3 View Models

If a Correlation Plot has been defined using the **Plot Models** button on the FigFull, FigWindow, or FigD3 worksheets, the **View Models** button will transfer control from the CommandCenter worksheet to the FigModel worksheet. If no Correlation plot has been defined, a warning dialog will appear. This dialog must be cleared before you can continue.

2.8.2 Plot D^3 Chart

Clicking the **Plot D^3 Chart** button initiates the D^3 (Display, Document, Discover) plotting functions described in Section 5. The quality metrics plotted in a D^3 chart are specified by the **D^3** checkboxes (Section 2.2.3).

2.8.2.1 View D^3 Chart

If a D^3 chart has been defined using **Plot D^3 Chart**, the **View D^3 Chart** button will transfer control from the CommandCenter worksheet to the FigD3 worksheet. If no D^3 chart has been defined, a warning dialog will appear. This dialog must be cleared before you can continue.

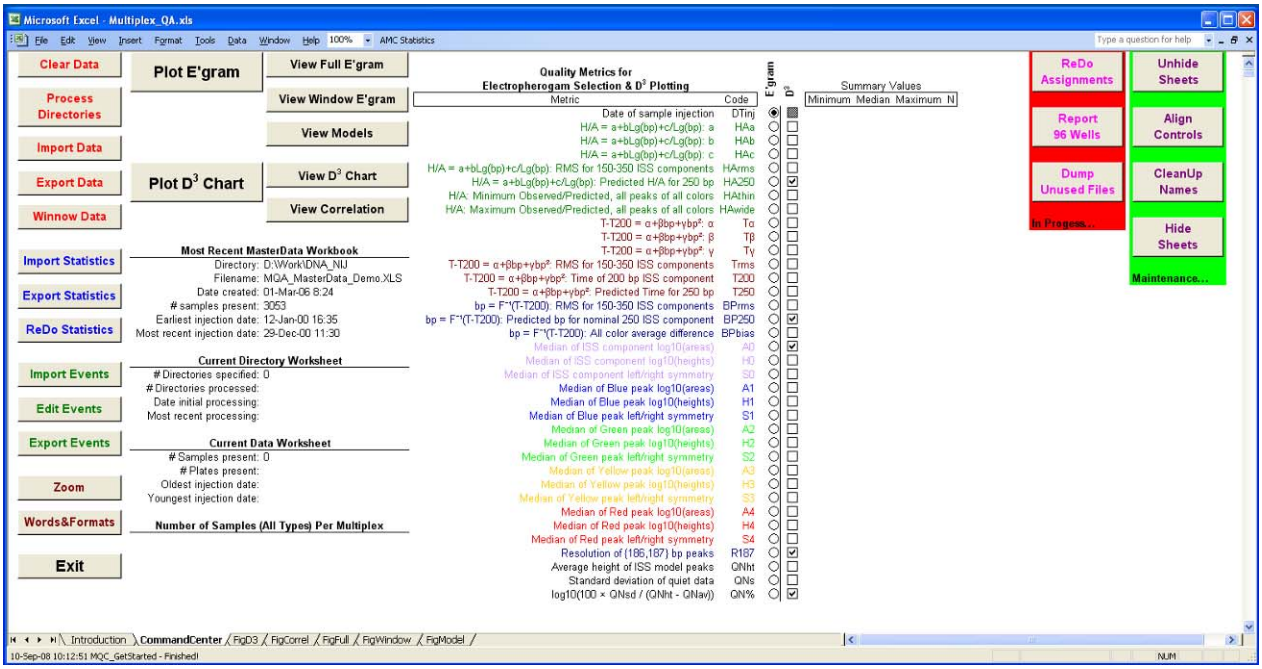
2.8.2.2 View Correlation

If a correlation scattergram has been defined using **Plot D^3 Chart**, the **View Correlation** button will transfer control from the CommandCenter worksheet to the FigCorrel worksheet. If no D^3 chart has been defined, a warning dialog will appear. This dialog must be cleared before you can continue.

2.9 Seven Somewhat Hidden Command Buttons

There are currently seven buttons to the far right hand edge of the CommandCenter worksheet as displayed in Figure 38. These buttons invoke functions that are not intended for routine use, either because they are archaic, still in development, or somewhat risky (the three one the red background) or are used to help maintain Multiplex_QA (the four on green).

Figure 38. Location of Somewhat Hidden Commands



2.9.1 ReDo Assignments

ReDo Assignments attempts to use the number of peaks of each color to identify samples that are allelic ladders and, if they are thought to be ladders, which multiplex kit they are for. These ladder assignments are, in turn, used to assign the multiplex kit of the remaining samples. This button is just here to make my life easier while fiddlin' with how the function works, since it currently doesn't work particularly well. If there is another release of the Multiplex_QA system, this button will disappear.

2.9.2 Report 96 Wells

Report 96 Wells was an early attempt to summarize the distribution of samples in a dataset over the 96 wells of a standard plate. Some aspects of the report may still be of interest, but except for the counts of sample Type (Blanks, Ladders, and Unknowns) per each of the 96 wells, it's probably of little use. If there is another release of the Multiplex_QA system, this button will likely disappear.

2.9.3 Delete Files

Clicking the **Delete Files** button can significantly reduce the storage required for the BatchExtract-processed .fsa files (see Section 9.3). However, since some of the files generated by the BatchExtract system are permanently deleted, using it is a tad risky… it's always possible that the (putative) next version of Multiplex_QA may require 'em. Then again, the BatchExtract system is now quite fast and fairly easy to use so the cost of re-processing has become relatively small. Anyway, the only reason to use **Delete Files** is if you are short of disk storage space. If you are short of space, consider using this function whenever you have successfully used **Process Directories** (Section 2.4.2).

2.9.4 Unhide Sheets

Unhide Sheets is a maintenance function that unhides ALL of Multiplex_QA's worksheets You're welcome to look at any-to-all; sheets are only hidden to reduce clutter. This function also restores the Excel workspace to its default conditions, but you are better off using the **Exit** command (Section 2.7.3) to accomplish this. Use the **Hide Sheets** command to return to the standard Multiplex_QA environment.

2.9.5 Align Controls

Align Controls is a maintenance function that repositions all of the Command Buttons on all of the Multiplex_QA worksheets to their intended locations. This may actually be of some use to you should you unintentionally move a button.

2.9.6 CleanUp Names

CleanUp Names is a maintenance function that deletes all of the temporary variables used in the Multiplex_QA macro code. It's pretty much a pure debugging aid

2.9.7 Hide Sheets

Hide Sheets is a maintenance function that undoes **Unhide Sheets**. It just returns Multiplex_QA to its standard environment.

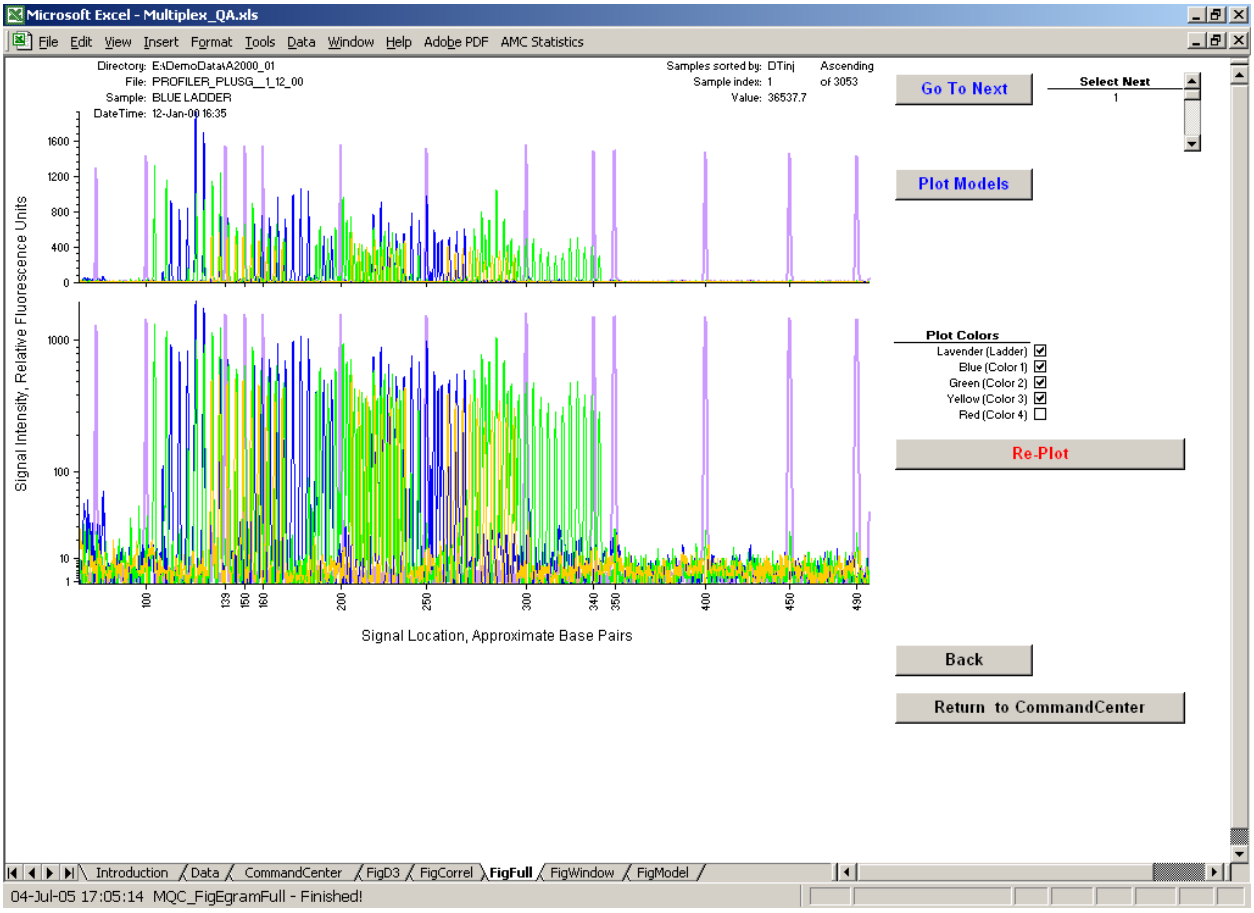
3 ELECTROPHEROGRAM PLOTS

The electropherogram ("E'gram") plots display fluorescence intensity in relative fluorescence units (RFUs) as a function of bp size for a specified sample. There are two E'gram formats: Full E'grams display all available electropherographic data while Window E'grams display selected intervals of the data.

3.1 Full E'gram and the FigFull Worksheet

The FigFull worksheet displays the full-scale ("Full") electropherogram plot. It also contains the controls that modify what data are displayed and the buttons that invoke other E'gram functions. The plot and the various controls are displayed in Figure 39.

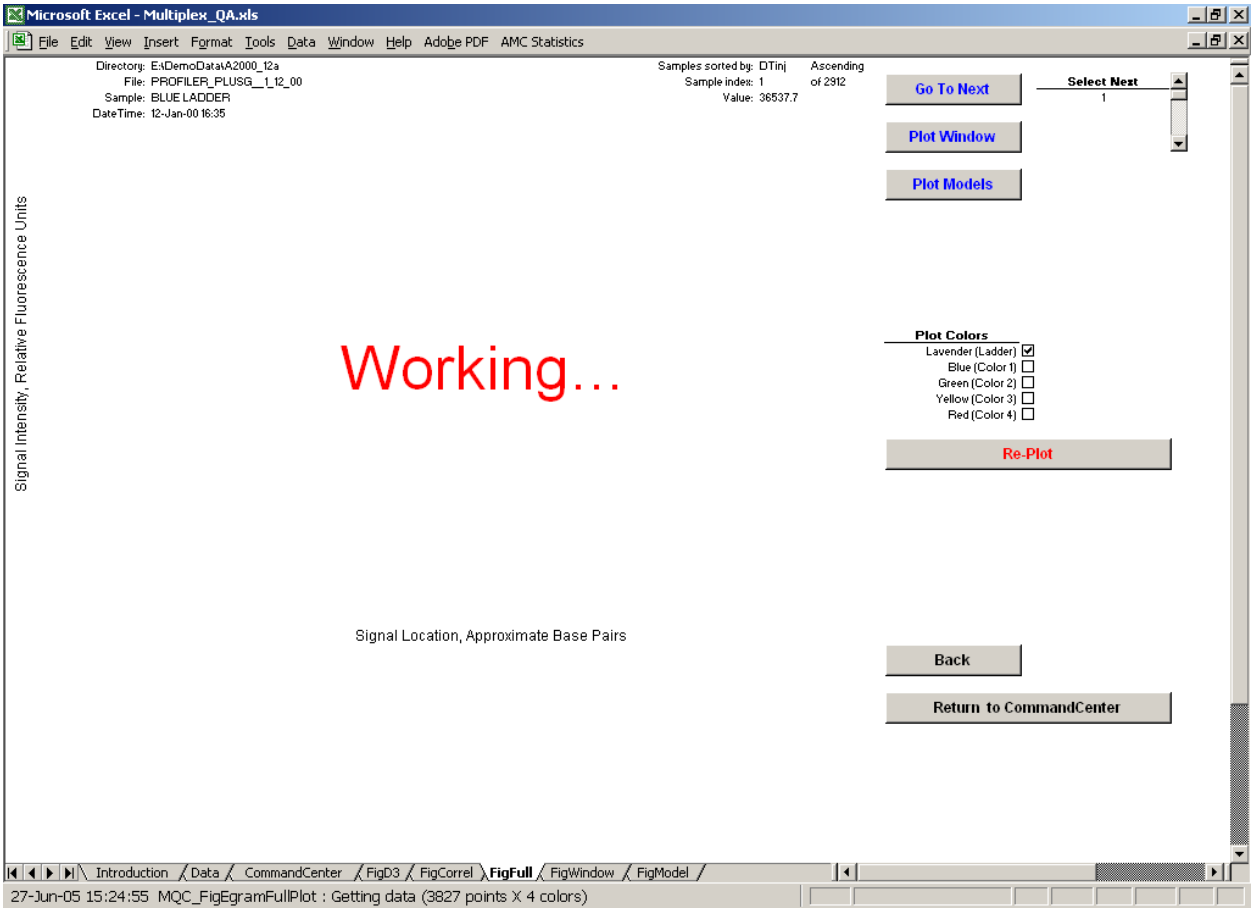
Figure 39. Full E'gram and FigFull Control Functions



The first Full E'gram must be invoked using the **Plot E'gram** button in the CommandCenter, giving an E'gram for the sample having the smallest value of the selected metric (Section 2.2.1). After this initialization, Full E'grams can be viewed by clicking the **Go To Next** button on the FigFull worksheet and the **Plot Full** button in the FigWindow, FigModel, and FigD3 worksheets. When invoked from FigFull, FigWindow or FigModel, the Full E'gram will be for the currently selected sample (3.1.3). When invoked from FigD3, the Full E'gram will be for the sample selected using the **Pick Sample** control on the FigD3 worksheet (Section 5.4).

A note of caution: Displaying the many thousands of data of an electropherogram stretches Excel pretty thin, particularly if you don't have a speedy processor. Figure 40 displays what you actually will see for a bit after you click **Plot E'gram**.

Figure 40. If You Don't Have Speed – Have Patience.



3.1.1 The Full E'gram

The full-scale electropherogram displays the entire electropherogram for the specified sample. The bp size range is from a little below the bp size of the smallest detected ISS peak to a little above the largest detected ISS peak, typically about 75 bp to 500 bp. The bp axis is labeled at the ISS peaks recognized by the ABI analysis software.

The RFUs are displayed on two scales in separate graphical segments: the upper segment is displayed on a linear scale from zero to the maximum RFU for any color. The lower segment is on a modified logarithmic scale that emphasizes the structure of the fluorescence signals near the baseline. The linear RFU-axis is labeled at uniform intervals from zero to the maximum; the modified logarithmic axis is labeled at each power of 10 from 1 RFU to the maximum.

3.1.2 Information block

The Directory, Base Filename, Sample, and Injection datetime sample identifiers (Section 2.4.2.6) are displayed to the upper left of the electropherogram. The metric used to determine

the sample order, whether the samples are being displayed in ascending or descending sequence, the index of the displayed sample in the dataset as it is currently sorted, the maximum possible sequence index, and the value of the ordering metric are displayed to the upper right of the electropherogram.

3.1.3 Choosing the next sample to display

The **Go To Next** button causes the electropherogram for the next available sample to be displayed. Unless the **Select Next** scroll bar to the right of the **Go To Next** button is used, the electropherograms will be presented sequentially in the order determined by the quality metric specified on the CommandCenter worksheet (Section 2.2.2). The ordering is initially ascending; it switches to descending after the last available sample has been selected.

The **Select Next** scroll bar allows you to select any sample to be next one displayed. If the selected sample is not displayable, the next sample for which the electropherogram can be displayed is selected. Whatever sample is specified for display on the FigFull worksheet is also specified on the FigWindow and FigModel worksheets.

A note of explanation: For most quality metrics, the "next available sample" will be the one in the next row of the sorted Data worksheet. However, there may well be gaps between the current and "next available" when the injection datetime is the ordering metric.

3.1.4 Re-Plot: Selecting which colors are displayed

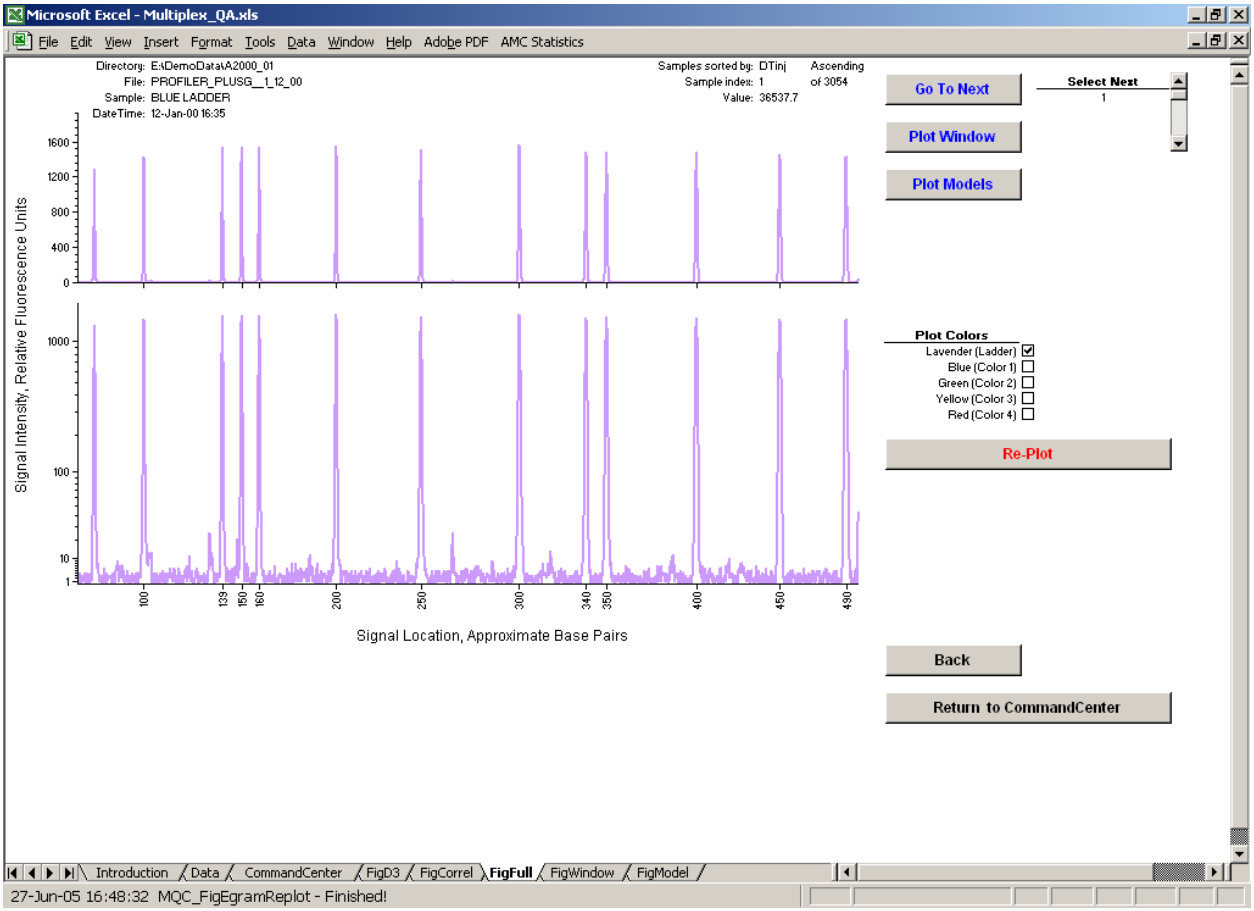
The **Re-Plot** button and the **Plot Color** checkboxes immediately above it control which fluorescence colors are displayed in the electropherogram. A checkbox can be activated only if data for that color are available. At least one color must be selected. The electropherogram is not re-plotted until **Re-Plot** is clicked. Figure 41 displays just the ISS data for the sample shown in Figure 39.

Color selections made on the FigFull worksheet are also applied to the FigWindow worksheet (Section 3.2.4.1).

3.1.5 Plot Window

Clicking the **Plot Window** button transfers control to the high-resolution ("Window") electropherogram plotting system on the FigWindow worksheet (Section 3.2). When first used, a section of the electropherogram for whichever sample is being displayed in the FigFull worksheet will be displayed in the FigWindow worksheet. This button is not active until the first Full E'gram is displayed.

Figure 41. Full E'gram, ISS Only



3.1.6 Plot Model

Clicking the **Plot Model** button transfers control to the FigModel worksheet (Section 4). This button is not active until the first Full E'gram has been displayed.

3.1.7 Back

Clicking the **Back** button transfers control to the worksheet that last transferred control to the FigFull worksheet.

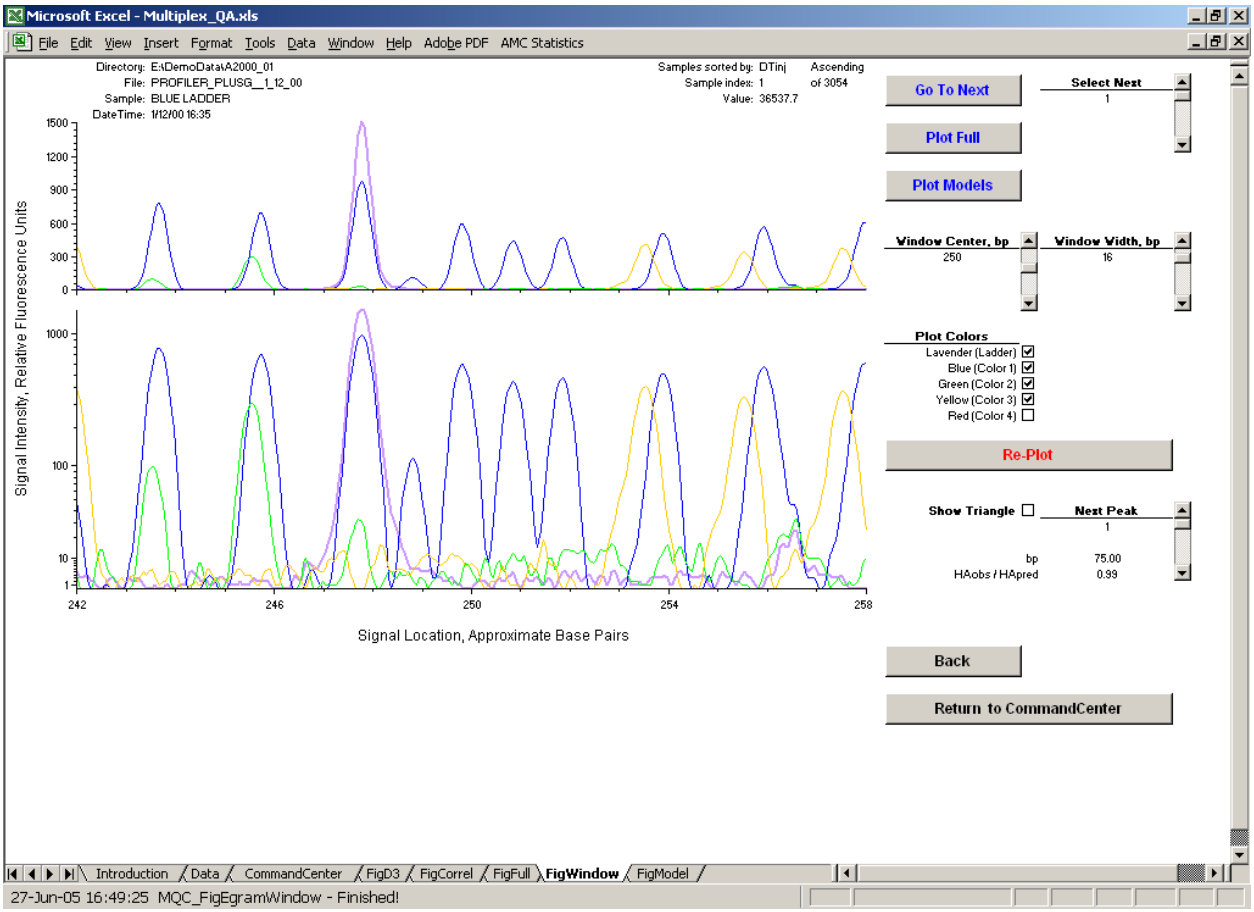
3.1.8 Return to CommandCenter

Clicking the **Return to CommandCenter** button transfers control to the CommandCenter worksheet.

3.2 Window E'gram and the FigWindow Worksheet

The Window E'gram displays a high-resolution slice ("Window") of an electropherogram. It is located in the FigWindow worksheet. The FigWindow worksheet contains the controls that modify what data are displayed and the buttons that invoke other E'gram functions. The plot and the various controls are displayed in Figure 42.

Figure 42. Window E'gram and FigWindow Control Functions



Window E'grams can only be viewed using the **Plot Window** button in the FigFull (Full E'gram), FigModel (Model plot) and FigD3 (D^3 Chart) worksheets. When invoked from FigFull or FigModel, the Window E'gram will be for the currently selected sample (3.1.3). When invoked from FigD3, the Window E'gram will be for the sample selected using the **Pick Sample** control in the FigD3 worksheet (Section 5.4).

A note of comfort: The Window E'grams typically display many fewer data than do the Full E'grams and therefore typically take less time (and patience) to display.

A note of explanation: The E'grams don't *really* plot RFU against bp, they plot RFU against an index related to the time the datum was obtained. Each sample's ISS is used to calibrate a quadratic model for the relationship between this index and bp size. The Full E'grams do not

use this model, as the ABI software provides a list of ISS peak locations and time indices that adequately define the whichness of what. However, this strategy won't work for Window E'grams since a given window may view few if any ISS peaks. Rather, the calibration function is used to estimate bp. Since the Multiplex_QA calibration model uses only a central subset of the potentially available ISS peaks (Section 8.4), the estimated bp values less than 150 bp and greater than 350 bp may be less accurate than those in between.

3.2.1 The Window E'gram

The Window E'gram is similar to the Full E'gram described in Section 3.1.1 except that 1) only a section of the data is displayed and 2) the bp axis is labeled at uniform intervals. The bp axis center and width are specified using the scroll bars located immediately below the **Plot Models** button (Section 3.2.6).

3.2.2 Information blocks

The displayed information is identical to that on the FigFull worksheet (Section 3.1.2).

3.2.3 Choosing the next sample to display

The **Select Next** scroll bar and **Go To Next** button work in the same manner as on the FigFull worksheet (Section 3.1.3). Whichever sample is specified on the FigWindow worksheet is also specified on the FigFull and FigModel worksheets.

3.2.4 Window E'gram options and the Re-Plot command

There are five sets of options for Window E'grams, rather than just the color selection as with Full E'gram. All options are "sticky" – once you set an option, it stays set (mostly: see Section 3.2.4.4). If you change an option, the Window E'gram will not be redrawn until **Re-Plot** is clicked.

3.2.4.1 Selecting which colors are displayed

The **Plot Color** checkboxes that select which colors will be displayed work in the same manner as in the FigFull worksheet (Section 3.1.4). Color selections made on the FigWindow worksheet are also applied to the FigFull worksheet.

3.2.4.2 Specifying the Window Center

The **Window Center, bp** scroll bar located immediately below the **Plot Models** button allows you to center the Window E'gram at a specific bp location. The minimum center value is 85 bp and the maximum is 500 bp. Clicking on this scroll bar temporarily inactivates the **Show Triangle** checkbox (Section 3.2.4.4).

3.2.4.3 Specifying the Window Width

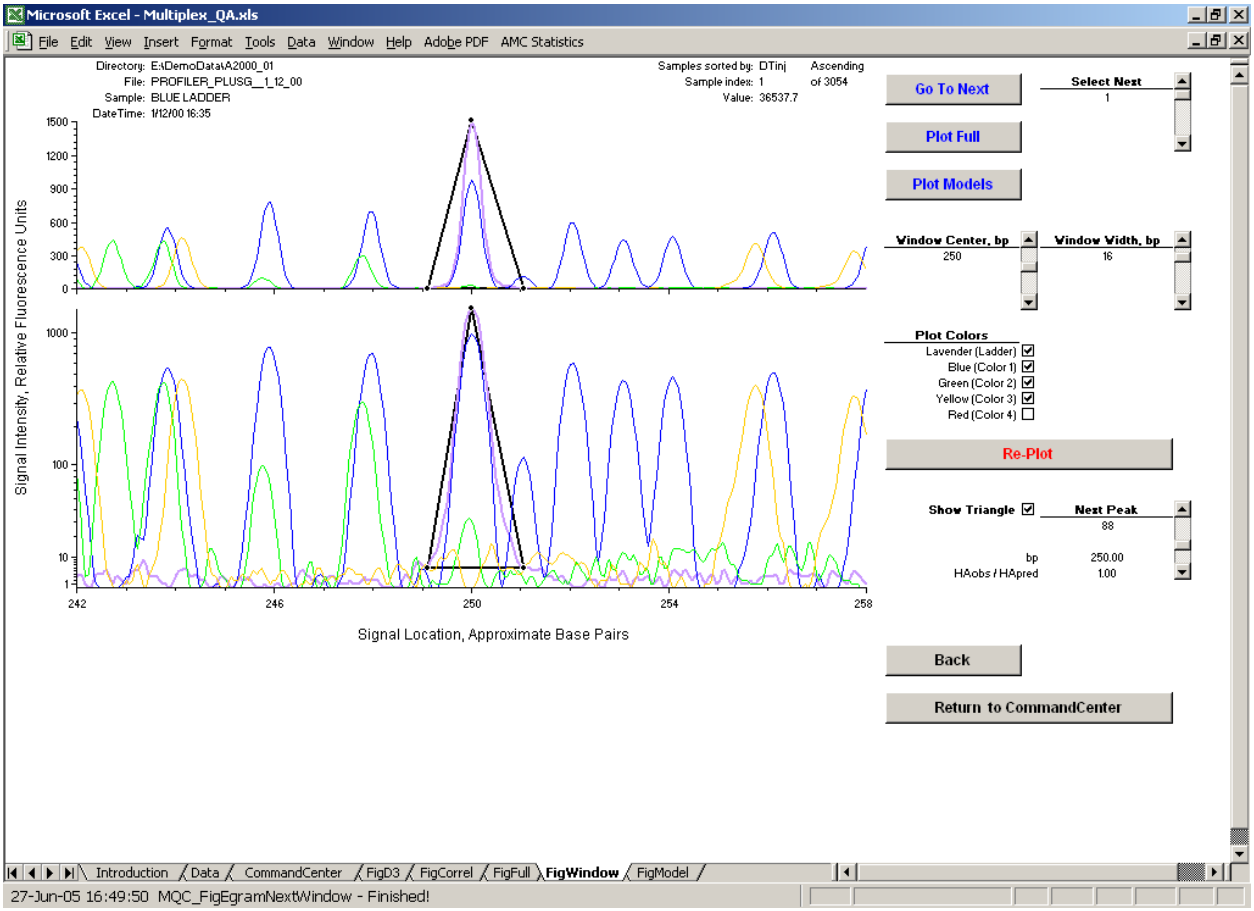
The **Window Width, bp** scroll bar located immediately to the right of the **Window Center, bp** scroll bar allows you to specify the bp width of the Window E'gram. The minimum width is

8 bp and the maximum is 200 bp. The width may be reduced if the center value minus one-half of the width is less than the bp value for the first available datum or if the center value plus one-half the width is greater than the bp value of the last available datum.

3.2.4.4 Displaying the Triangle Peak

The **Show Triangle** checkbox below the **Re-Plot** button turns on (when checked) and off (when empty) the display of the ABI-identified start-tip-stop points of a selected peak in the electropherogram. These peak recognition events are visualized as a black triangle in both the linear and semi-logarithmic plot segments, as shown in Figure 43. The bp location of the tip of the selected peak is displayed in the "bp" line below the checkbox; the **Window Center, bp** scroll bar (Section 3.2.4.2) is assigned this value to ensure that the Window E'gram is centered on the selected peak.

Figure 43. Window E'gram, Selected Peak



A note of explanation: Figure 42 and Figure 43 display the same window (centered at 250 bp, width of 16 bp), but the x-axes "Signal Location" labels are different by about 4 bp. Why? The nominal 250 peak in the GeneScan family of ISS does not play well with others [2]: on the ABI 310 that collected these demo data, its apparent electrophorectic size is consistently about 4 bp smaller than its nominal size. When **Show Triangle** is inactive, the x-axis is determined using just the peak retention model (Section 4.1.2). When **Show Triangle** is active, the window is

centered on the nominal size of the selected peak and only the Window's width is calculated from the retention model.

3.2.4.5 Selecting which Triangle Peak to display

When the **Show Triangle** checkbox is turned on, the **Next Peak** scroll bar to its right can be used to select which ABI-identified peak will be selected as the Triangle Peak. The bp location of the tip of the selected peak will be centered in the Window center. Clicking on the **Next Peak** scroll bar temporarily activates the checkbox.

3.2.4.5.1 bp size display

The bp location of the tip of the peak is displayed on the "bp" row below the **Show Triangle** checkbox.

3.2.4.5.2 Standardized Height/Area ratio display

When **Window E'gram** is invoked from the FigFull worksheet (Section 3.1.5), the row below the bp size displays the observed height/area ratio for the selected peak (H/A_{obs}) standardized to the height/area ratio expected for a peak at that bp location (H/A_{calc}): H/A_{obs} / H/A_{calc} All ABI-identified peaks can be selected, ordered (most of the time, see Section 3.2.4.5.3) from smallest to largest bp size.

When **Window E'gram** is invoked from the FigModel worksheet (Section 4.2.4), the next row displays the absolute value of the residual between the observed and calculated height/area ratio, $|H/A_{obs} - H/A_{calc}|$. Only the peaks used to define the height/area model can be selected, ordered from least well-modeled to best-modeled peak.

3.2.4.5.3 Peak selection order with Pthin or Pwide

When **Window E'gram** is invoked from the FigFull worksheet (Section 3.1.5) and Pthin (minimum height/area ratio, H/A_{obs} / H/A_{calc}) and Pwide (maximum H/A_{obs} / H/A_{calc}), the peaks are arranged in order of increasing H/A_{obs} / H/A_{calc}.

3.2.5 Plot Full

Clicking the **Plot Full** button transfers control to the complete ("Full") E'gram plotting system on the FigFull worksheet (Section 3.1). The Full E'gram for whichever sample is being displayed in the FigWindow worksheet will be displayed in the FigFull worksheet.

3.2.6 Plot Model

Clicking the **Plot Model** button transfers control to the FigModel worksheet (Section 4). The Model plot for whichever sample is being displayed in the FigWindow worksheet will be displayed in the FigModel worksheet.

3.2.7 Re-Plot

Clicking the **Re-Plot** button causes the Window E'gram to be redrawn using whatever plot options have been specified. None of the options take effect until **Re-Plot** is clicked.

3.2.8 Back

Clicking the **Back** button transfers control to the worksheet that last transferred control to the FigWindow worksheet.

3.2.9 Return to CommandCenter

Clicking the **Return to CommandCenter** button transfers control to the CommandCenter worksheet.

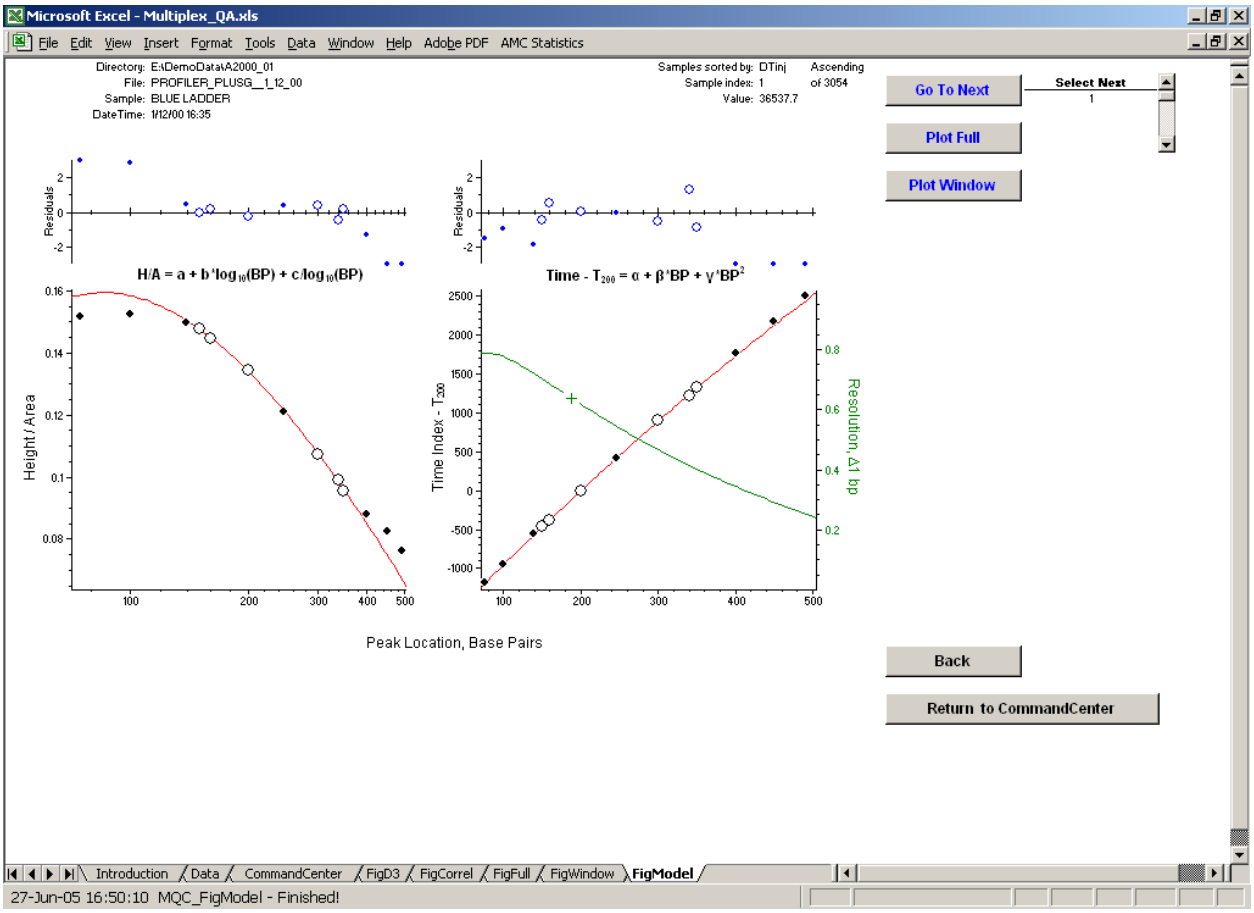
3.3 When the BatchExtract Files Aren't Where They Were

Sometime the folder-paths in the current dataset do not actually contain the BatchExtract-ed files that they are supposed to contain. This may result from premature ejection or deletion, but it more typically results from giving old folders new names or locations. Whenever the Full E'gram, Window E'gram, or Model functions encounter an invalid folder-path, you will be offered an opportunity to correct the folder-path. The procedure is the same as described in Section 2.4.3.1.

4 MODEL PLOTS

The Model plot visualizes the data that define the peak height/area ratio and retention time models for a particular sample. The Model plot is located in the FigModel worksheet. An example Model plot and the FigModel worksheet control functions are displayed in Figure 44.

Figure 44. Model Plot: Height/Area, Retention, and Resolution Models



Model plots are viewed by clicking the **Plot Model** button in the FigFull (Full E'gram), FigWindow (Window E'gram), and FigD3 (D^3 Chart) worksheets. When invoked from either FigFull or FigWindow, the Model plot will display the data for the currently selected sample (Section 3.1.3). When invoked from FigD3, the Model plot will display the data for the sample selected using the **Pick Sample** control in the FigD3 worksheet (Section 5.4).

A note of caution: The height/area and retention models are *empirical* – neither has any theoretical justification.

4.1 Graphical Elements

There are four graphical segments to the Model plot. The upper and lower segments on the left visualize the observed peak height/area ratios as a function of peak bp size. This ratio is the reciprocal of the expected peak width, assuming the shape of the peaks is approximately

triangular. The upper and lower segments to the right visualize the observed time index of the tip of the peak as a function of peak bp size. The time index for the peak tip is effectively the retention time, assuming the peak is "fairly symmetrical." The lower right segment also visualizes the effective resolution for peaks of 1 bp size difference, based on the height/area and retention models.

4.1.1 Height/Area model

The lower left graphical segment of the **Model plot** displays the observed height/area (H/A) data for the major ISS peaks and the empirical function used to relate these values (H/A_{obs}) to the nominal bp size of the ISS peaks:

$$H/A_{calc,i} = a + b \times \log_{10}(bp_i) + c / \log_{10}(bp_i) \ .$$

The data used to parameterize the model are shown as open circles, data for all other ISS peaks are shown as solid circles, and the parameterized function is denoted as a red line.

The upper left graphical segment displays the normalized residuals for the H/A data:

$$\text{Residual}_i = \frac{H/A_{obs,i} - H/A_{calc,i}}{\sum_{i=1}^{N}\left(H/A_{obs,i} - H/A_{calc,i}\right) \Big/ (N-3)}$$

where N is the number of ISS ladder peaks used to define the model. Any residual greater than 3 or less than –3 is displayed as if it were 3 or –3.

4.1.2 Retention model

The lower right graphical segment of the Model plot displays the observed time index (TI) data for the major ISS peaks and the empirical function used to relate these values (TI_{obs}) to the nominal bp size of the ISS peaks:

$$TI_{calc,i} - TI_{200} = \alpha + \beta \times bp_i + \gamma \times bp_i^2$$

where TI_{200} is the time index for the ISS peak of nominal size 200 bp. The TI_{200} offset is used to provide a "constant data recording start time" for each sample, which serves to stabilize the meaning of the constant (α) term of the equation but does not otherwise affect the regression. The data used to parameterize the model are shown as open circles, data for all other ISS peaks are shown as solid circles, and the parameterized function is denoted as a red line.

The upper right graphical segment displays the normalized residuals for the TI data:

$$\text{Residual}_i = \frac{TI_{obs,i} - TI_{calc,i}}{\sum_{i=1}^{N}\left(TI_{obs,i} - TI_{calc,i}\right) \Big/ (N-3)}$$

where N is the number of ISS ladder peaks used to define the model. Any residual greater than 3 or less than –3 is displayed as if it were 3 or –3.

A note of information: The time index (TI) values are expressed relative to the nominal 200 bp ISS peak (TI_{200}) because the nominal 200 bp is the smallest component shared by the two ISS materials (the GS and ILS families, see Section 2.4.2.6.9) in current use.

4.1.3 Resolution model

The green line in the lower right graphical segment displays the expected resolution of peaks 1 bp size different in size as a function of bp size. The calculated resolution for a peak of size 186.5 bp is indicated by the green "+"; the HUMTH01 {9.3, 10} alleles in the COfiler multiplex are of approximate size 186 and 187 bp.

The standard chromatographic definition of resolution for these peaks is [3]

$$R_{\Delta 1}(x) = 2\frac{(TI_{x+0.5} - TI_{x-0.5})}{W_{x-0.5} + W_{x+0.5}}$$

where TI is the Time Index and W is the base-width of a "peak" of size x-0.5 bp or x+0.5 bp. From the retention model:

$$\begin{aligned} TI_{x+0.5} - TI_{x-0.5} &= \left(TI_{200} + \alpha + \beta(x+0.5) + \gamma(x+0.5)^2\right) - \left(TI_{200} + \alpha + \beta(x-0.5) + \gamma(x-0.5)^2\right) \\ &= \beta(x+0.5-x+0.5) + \gamma\left((x+0.5)^2 - (x-0.5)^2\right) \\ &= \beta + \gamma 2x\,. \end{aligned}$$

Assuming that the peak shapes can be approximated as triangular:

$$\begin{aligned} W_{x-0.5} + W_{x+0.5} &= 2\frac{A_{x-0.5}}{H_{x-0.5}} + 2\frac{A_{x+0.5}}{H_{x+0.5}} \\ &= 2\left(\frac{A_{x-0.5}}{H_{x-0.5}} + \frac{A_{x+0.5}}{H_{x+0.5}}\right). \end{aligned}$$

From the height/area model:

$$\begin{aligned} \frac{A_{x-0.5}}{H_{x-0.5}} &= \frac{1}{a + b\log_{10}(x-0.5) + c/\log_{10}(x-0.5)} \\ \frac{A_{x+0.5}}{H_{x+0.5}} &= \frac{1}{a + b\log_{10}(x+0.5) + c/\log_{10}(x+0.5)}\,. \end{aligned}$$

Combining these equations and noting that the peak sizes of interest are all greater than 75 bp

gives:

$$R_{\Delta 1}(x) = (\beta + \gamma 2x) \Big/ \left(\frac{1}{a + b\log_{10}(x - 0.5) + c/\log_{10}(x - 0.5)} + \frac{1}{a + b\log_{10}(x + 0.5) + c/\log_{10}(x + 0.5)} \right).$$
$$\cong (\beta + \gamma 2x)(a + b\log_{10}(x) + c/\log_{10}(x))/2$$

A note of information: Peaks that have the same TI have resolution 0. Truly triangular peaks that are just baseline-resolved have resolution 1.0. For "real" peak shapes, the rule of thumb is that a resolution of 2.0 is needed to ensure baseline resolution.

4.1.4 Model parameterization

Both the height/area (H/A) and retention models are defined using data for the ISS peaks between 150 bp and 350 bp, excluding the 250 bp peak. While both the GS and ILS ladders can provide peaks from about 75 bp to about 500 bp, data are not always collected for the lowest or the highest peaks in the ISS. The GS family of ISS has six peaks in this region; the ILS family have eight. For both types of ISS, the 250 bp peak is "in the middle" of the peaks used to define the model. These defining data are denoted as open circles in all Model plot segments.

The nominal 250 bp peak is *not* used in the parameterization of either model. In the GS family of ISS, the retention time of this peak is "anomalous" – the peak typically migrates as if it were a few bp shorter that 250 bp and the "how much shorter?" is quite sensitive to the temperature of the column or gel [2]. While the H/A behavior of this peak is not much affected by its retention time anomalies, it is excluded from the H/A model to keep the two model-definition processes as similar as possible.

4.1.5 Model validation

The nominal 250 bp peak and all of the expected ISS peaks other than those used to define the models are used to help validate the models. These validation data are denoted as solid circles in all the Model plot segments.

The pattern of residuals for these peaks can help evaluate whether the models are predictive. Since it is "in the middle" of the models, the interpolated values for the nominal 250 bp peak are expected to be more reliable the extrapolated values of the peaks smaller than 150 bp and those larger than 350 bp.

4.2 Worksheet Elements

4.2.1 Information block

The information block displayed above the Model plot is identical to that of the FigFull worksheet (Section 3.1.2).

4.2.2 Choosing the next sample to display

The **Select Next** scroll bar and **Go To Next** button work in the same manner as on the FigFull worksheet (Section 3.1.3). Whichever sample is specified on the FigWindow worksheet is also specified on the FigFull and FigModel worksheets. If the specified sample is displayed on the D^3 chart, it will become the "picked" sample (Section 5.4).

4.2.3 Plot Full

Clicking the **Plot Full** button transfers control to the complete ("Full") E'gram plotting system on the FigFull worksheet (Section 3.1). The Full E'gram for whichever sample is being displayed in the FigModel worksheet will be displayed in the FigFull worksheet.

4.2.4 Plot Window

Clicking the **Plot Window** button transfers control to the high-resolution ("Window") E'gram plotting system on the FigWindow worksheet (Section 3.1.6). A Window E'gram for whichever sample is being displayed in the FigModel worksheet will be displayed in the FigWindow worksheet.

4.2.5 Back

Clicking the **Back** button transfers control to the worksheet that last transferred control to the FigModel worksheet.

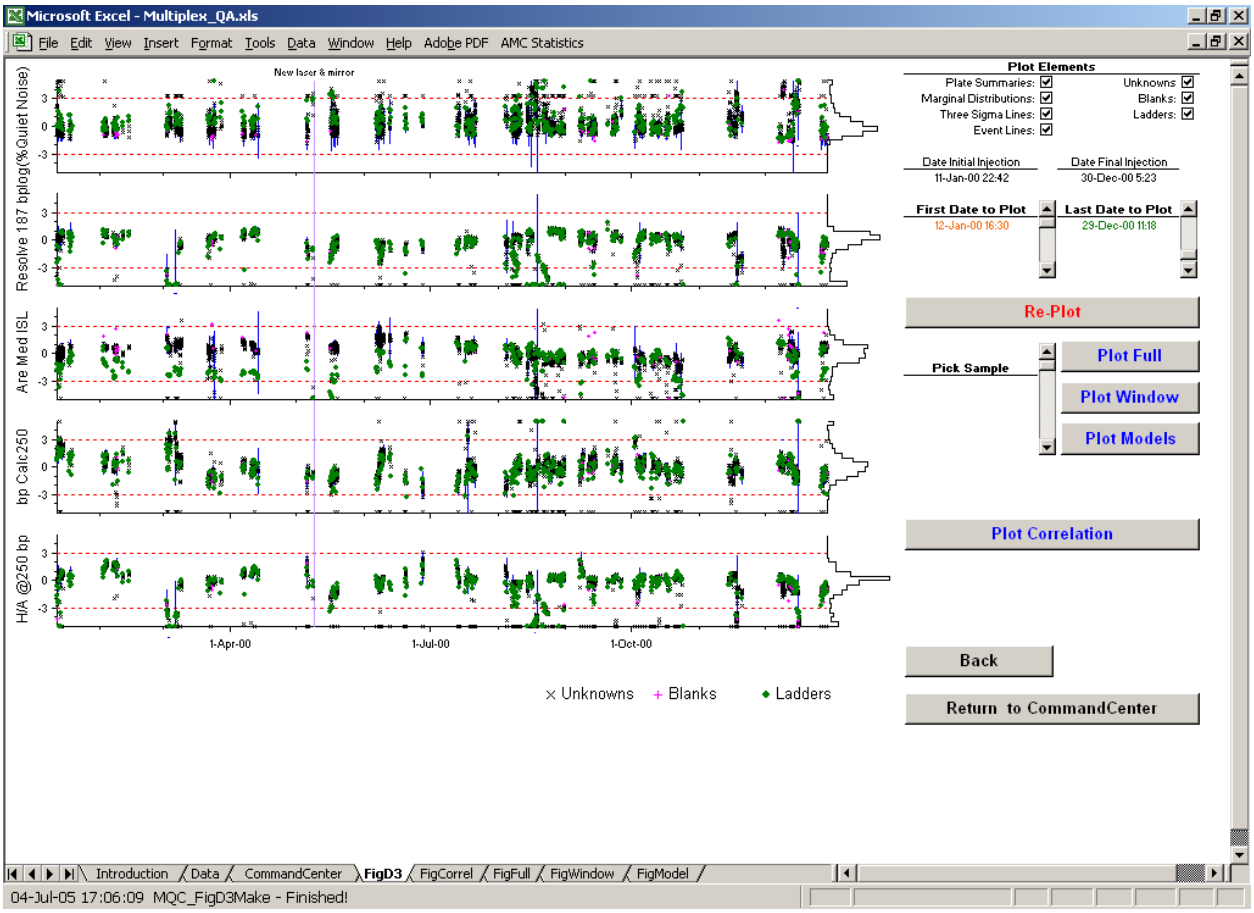
4.2.6 Return to CommandCenter

Clicking the **Return to CommandCenter** button transfers control to the CommandCenter worksheet.

5 D³ CHARTS

The "Display, Document, Discover" or "D³" chart presents the time-series behavior of selected quality metrics. The D³ chart is located in the FigD3 worksheet; it is accessed through the **Plot D³** button on the CommandCenter worksheet. A D³ chart for the Demo data and the worksheet FigD3 control functions are displayed in Figure 45.

Figure 45. Five-Metric D³ Chart and FigD3 Control Functions



The format of the D³ chart is controlled using the buttons, scroll bars, and checkboxes in the FigD3 worksheet. The FigD3 control functions become active only *after* clicking on **Plot D³**. If you click on any of the FigD3 buttons without having first clicked on **Plot D³**, a warning dialog will appear. This dialog must be cleared before you can continue.

The FigD3 worksheet also contains the **Plot Correlation** button. Clicking this button after the D³ chart is defined displays the pair-wise correlation among the quality metric values displayed in the D³ chart.

A note of explanation: A "D³ Chart" is a time series plot with bells and whistles. It is intended to let you *display* multiplex STR quality metric data so that you can *document* what's going on and hopefully help you *discover* why it's happening.

5.1 Selecting Quality Metrics for Display

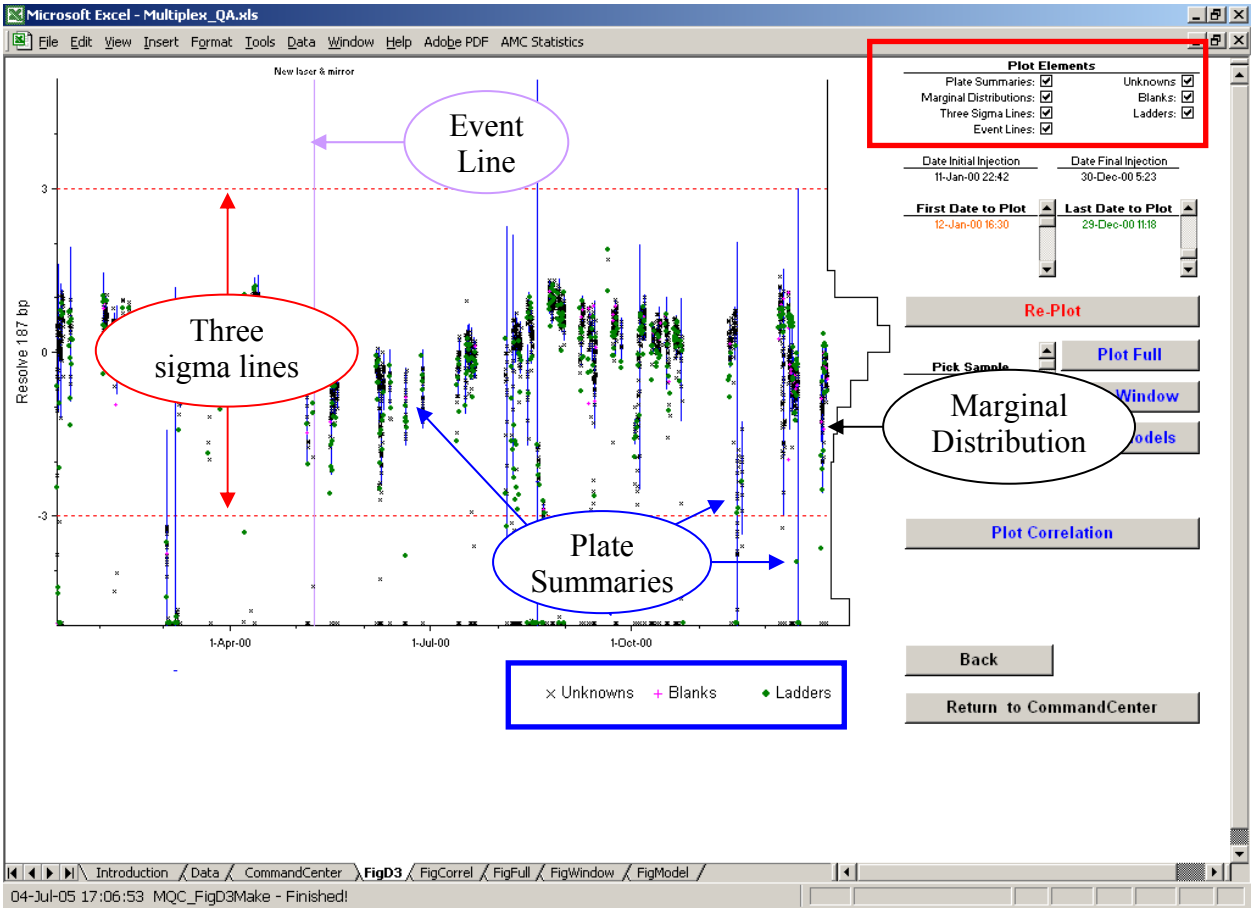
The quality metrics that are displayed in the D^3 chart are selected with the CommandCenter's **D^3** checkboxes (Section 2.2.3). At least one checkbox must be active for the CommandCenter's **Plot D^3 Chart** button to work. No more than five checkboxes can be active at any one time. Anytime that you change which checkboxes are active, you must click **Plot D^3 Chart** to generate a new D^3 chart.

A note of explanation: The requirement that at least one quality metric be selected is intrinsic – you can't plot nothing. However, the "no more than five" is just a practical limitation set by the dimensions of the plot. Five is about as many graphical segments as can be accommodated without changing the size of symbols and fonts.

5.2 Plot Elements

Figure 46 displays the data for just one of the plot metrics shown in Figure 45 to better reveal the various graphical elements of the D^3 chart. These elements are controlled using the **Plot Elements** checkboxes in the upper right corner of the FigD3 worksheet (the red rectangle in Figure 46). No change is made to the D^3 chart until **Re-Plot** is clicked.

Figure 46. One-Metric D^3 Chart and Plot Elements



5.2.1 Plate summaries

The blue cross-like "Plate summaries" display the range of a values for a given quality metric over all samples (unknowns, blanks, and ladders) assigned as members of the same plate (Section 2.4.2.6.3). The vertical line of each cross is a robust 95% confidence interval for the values, calculated from the median and IQRe estimates of the center and spread of the distribution. The horizontal line spans the entire range of injection times for the samples assigned to each plate. Each vertical line is centered at the mid-range of the injection time interval.

At least one of the four data-representation checkboxes (**Plate Summaries**, **Unknowns**, **Blanks**, or **Ladders**) must be active. If you attempt to deactivate all four checkboxes, all four will automatically become active.

5.2.2 Marginal distributions

The black histograms at the right end of each graphical segment display the probability density distribution of the values for all samples (unknowns, blanks, and ladders). Each bin of the histogram represents the relative number of samples that have values that fall within the bin – what you would see if you could tilt the graph up on its left edge and let the points slide straight down into clear glass jars set up on the right edge.

The Marginal distributions are an aid to evaluating the relative utility of the various quality metrics. The things to look for are: the number of modes (popular bins separated by less-popular ones indicate that more than one thing is going on), the location of the biggest mode (it's good if the most popular bin is close to the center), and the width of the most popular mode (not too narrow and not too wide).

The marginal distributions can be switched on or off without affecting any of the other graphical elements.

A note of explanation and caution: All histograms on a give D^3 chart can be directly compared as they use the same number of bins for the same normalized span of values and are displayed on the same scale. Each of the histograms has the same area (the same number of values), with the most popular bin among all of the histograms scaled to have maximum height. However, the height scale will in general *not* be the same across different D^3 charts.

5.2.3 Three-sigma lines

The horizontal red dashed lines in each segment represent the approximate 99% confidence interval on the values of the metric, under the assumption that the values are distributed approximately "normally" (unimodal, symmetric about the center, with the width at the base about two to three times the width at half-maximum). The median is used as a robust estimate of the center and the IQRe is used as a robust estimate of the spread.

The Three-sigma lines can be switched on or off without affecting any of the other graphical elements.

The Three-sigma lines are mostly a bow to the D^3 chart's "Control chart" cousins, but they may help you decide which of the quality metric values are worth examining in greater detail. However, the "99% confidence" bit applies only if the Marginal distributions look more or less "normal."

5.2.4 Event lines

The vertical lavender lines mark the datetime of any known "events" that may be of interest for the particular dataset (Section 2.6). The utility of these lines is completely dependent upon how diligently you document which things happened when. Unless you record in a MasterEvent file when something happened (e.g., column replaced, instrument PM –Preventive Maintenance as well as Post-Mortem repair, and environmental problems like "air conditioning failed"), it will be difficult to tie changes in the quality metrics to their root causes.

The Event lines can be switched on or off without affecting any of the other graphical elements.

5.2.5 Unknowns, blanks, and ladders

The individual values of a given quality metric can be displayed for unknowns (×), blanks (+), and ladders (●). A legend for the types of samples displayed is located at the lower right of the D^3 chart (the blue rectangle in Figure 46).

At least one of the four data-representation checkboxes (**Plate Summaries**, **Unknowns**, **Blanks**, or **Ladders**) must be active. If you attempt to deactivate all four checkboxes, all four will automatically become active.

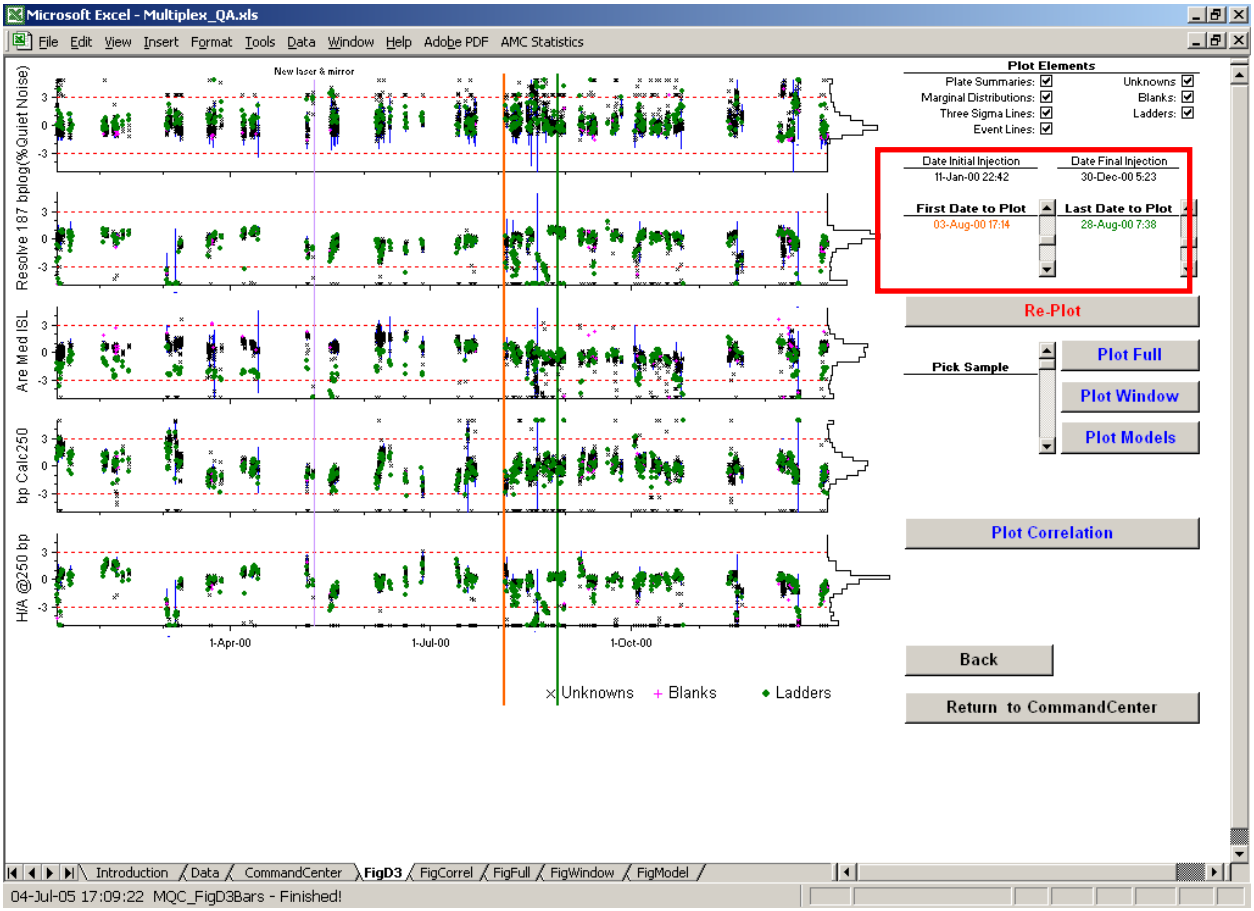
5.3 Specifying the Datetime Interval

It's often an excellent idea to start off seeing the big picture with as many data as are available. However, if you note any "funny stuff" worth exploring you will need to look at selected regions with much higher datetime resolution. The information block and scroll bars just below the Plot Elements allow you to specify the starting and stopping datetimes for the D^3 chart. Figure 47 is an example of what the D^3 chart should look like during the selection process.

5.3.1 Initial and final datetimes for the dataset

The earliest and most recent injection datetimes for any type of sample in the current dataset are displayed in the "Date Initial Injection" and "Date Final Injection" information lines immediately below the Plot Element controls. These values are provided just to remind you about the datetime limits of the current data.

Figure 47. Specifying the Datetime Interval

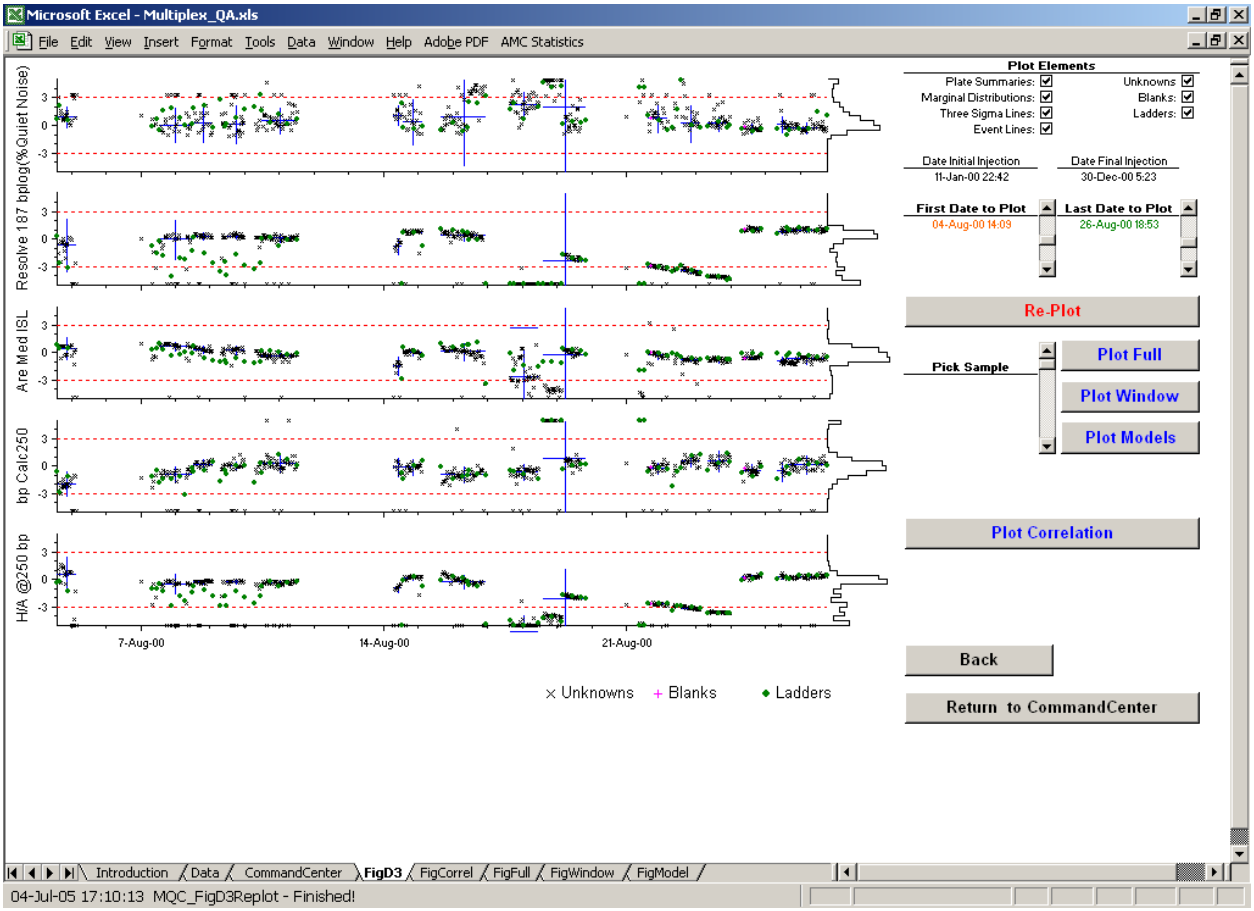


5.3.2 Initial and final datetimes for the D^3 chart

The earliest and latest injection datetimes that the D^3 chart will display are controlled by the **First Date to Plot** and **Last Date to Plot** scroll bars just above the **Re-Plot** button. You change the datetime by clicking on the scroll bars. When you click or drag on the **First Date to Plot** scroll bar, the orange datetime should change and a vertical orange line should appear on the currently displayed D^3 chart. When you click or drag on the **Last Date to Plot** scroll bar, the green datetime should change and a vertical green line should appear. Position the two lines so that the datetime interval you desire is bracketed by the two lines, then click **Re-Plot**. Figure 48 displays the resulting higher-resolution D^3 chart.

A note of explanation: While Excel supports a dynamic "drag the symbol to where you want it" facility that, in principle, could be used to specify the datetime interval without the hassle of the scroll bars, I have yet to make it work smoothly. At least with large datasets on the two systems used to develop Multiplex_QA, dynamic dragging is problematic… often requiring two tries before it works and – sometimes but not with any pattern yet recognized – inducing a fatal Excel error.

Figure 48. Expanded Datetime Interval



5.3.3 Using the scroll bars

Using the scroll bars is a bit tricky, particularly if there are many data displayed and you are running on a slow computer. Don't click too fast! It's best to wait until the lines move before clicking again. Remember: use the scroll bar's slider to make the big changes, click within the scroll bar for large step changes ($1/20^{th}$ of the displayed datetime interval) and click on the ends of a the scroll bar for small step changes.

The **First Date to Plot** must always be earlier than the **Last Date to Plot**. If you attempt to make the first last or vice versa, whichever limit is being changed will "bump" the other limit to ensure the correct ordering of things.

If you want to expand the datetime interval beyond what the current D^3 chart currently displays, use the scroll bar sliders. Since datetimes outside the current display can't be visualized, you get no feedback other than the position of the sliders and the value of the orange and green datetimes. While the coarse and fine controls of the two scroll bars are set by the datetime interval of the currently displayed D^3 chart, the scroll bar limits are always the first and last injection datetimes of the current dataset.

5.4 Selecting One Sample for Detailed Analysis

When you notice that some sample has unusual quality metric values, or that a sudden change in values occurred between one sample and the next, it is useful to visualize data for selected samples. The **Pick Sample** information and control connects the D^3 chart to the E'gram and Model plotting functions. Figure 49 is an example of what the D^3 chart should look like during the selection process.

Figure 49. Selecting a Sample

5.4.1 Specifying a sample

The **Pick Sample** scroll bar immediately below the **Re-Plot** button allows you to select any sample displayed in the D^3 chart for further analysis. The datetime of the selected sample is indicated by a red triangle along the datetime axis of the chart. The control metric values for the sample are centered in open red circle(s). If two or more control metrics are displayed in the chart, the red circles are connected with vertical red lines. The red triangle, circles, and lines appear when you first click the **Pick Sample** scroll bar.

5.4.2 Sample information

To identify what sample is currently selected, an information block immediately to the left of the **Pick Sample** scroll bar lists a number of identifiers: the injection datetime (Section 2.4.2.6.7), the multiplex kit (Section 2.4.2.6.4), the sample type (Section2.4.2.6.2), the set and plate number in the current dataset (Section 2.4.2.6.3), the sample name (Section 2.4.2.6.6), and the sample base filename and directory (Section 2.4.2.6.5). This information should change every time you manipulate the scroll bar.

5.5 Re-Plot

Clicking the **Re-Plot** button causes the D^3 chart to be redrawn using whatever plot options have been specified. None of the options take effect until **Re-Plot** is clicked.

5.6 Plot Full

Clicking the **Plot Full** button transfers control to the FigFull worksheet and generates the Full E'gram for the selected sample (Section 3.1).

5.7 Plot Window

Clicking the **Plot Window** button transfers control to the FigWindow worksheet and generates the Window E'gram for the selected sample (Section 3.2). Figure 50 displays the Window E'grams and Models for two selected samples. For these samples, the Window E'grams reveal slight "pull up" of the yellow signal into the ISS signal for both files, but also considerable peak tailing.

5.8 Plot Model

Clicking the **Plot Model** button transfers control to the FigModel worksheet and generates the Model plot for the selected sample (Section 4). Figure 50 displays the Window E'grams and Models for two selected samples. For these samples, the Model plots confirm that the peak height/area relationship for ISS peaks is significantly different for the two samples.

5.9 Plot Correlation

When two or more quality metrics are selected, clicking the **Plot Correlation** button transfers control to the FigCorr worksheet and generates a scattergram for each pair of metrics (Section 6).

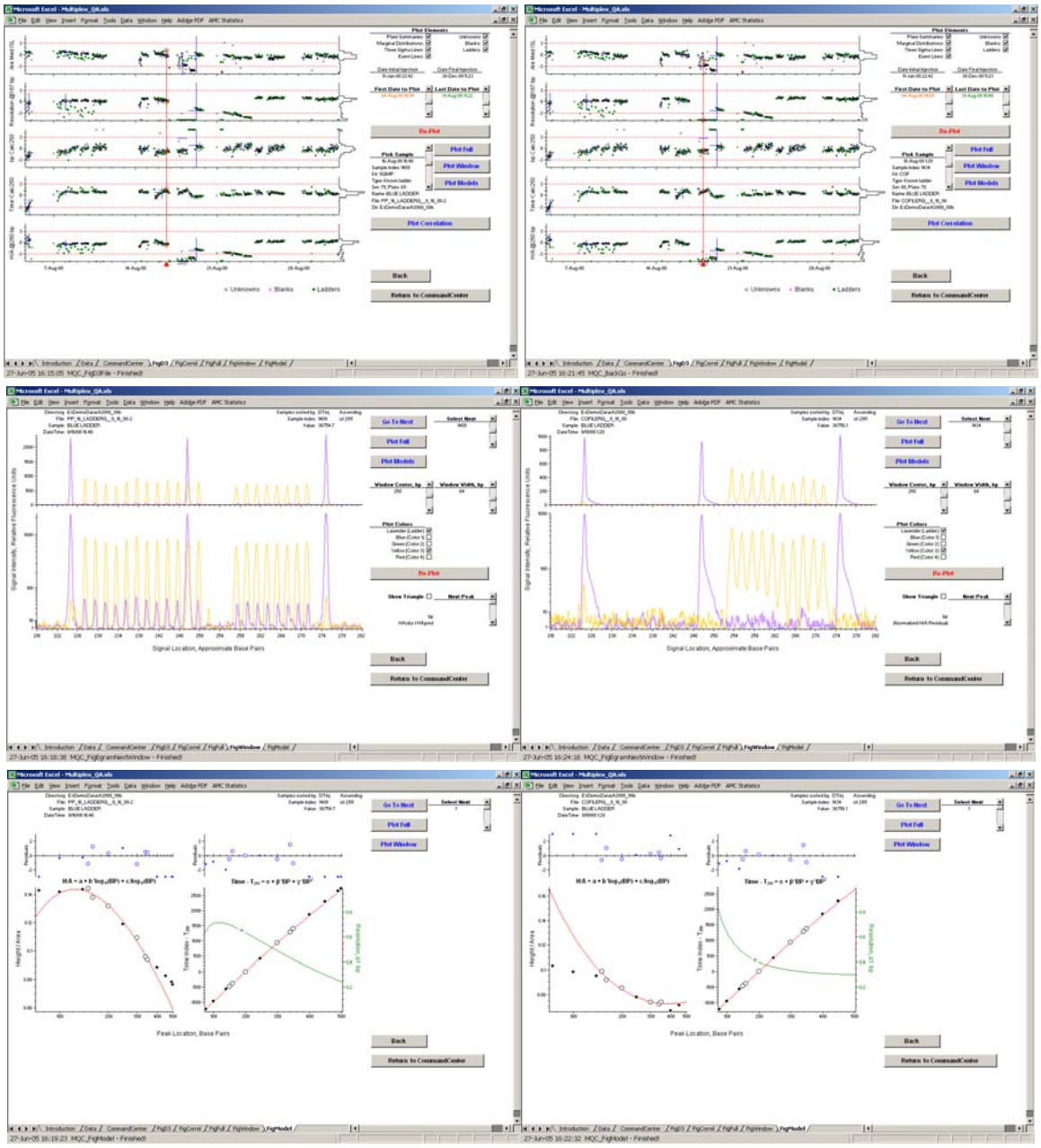
5.10 Back

Clicking the **Back** button transfers control to the worksheet that last transferred control to the FigD3 worksheet.

5.11 Return to CommandCenter

The **Return to CommandCenter** button transfers control to the CommandCenter worksheet.

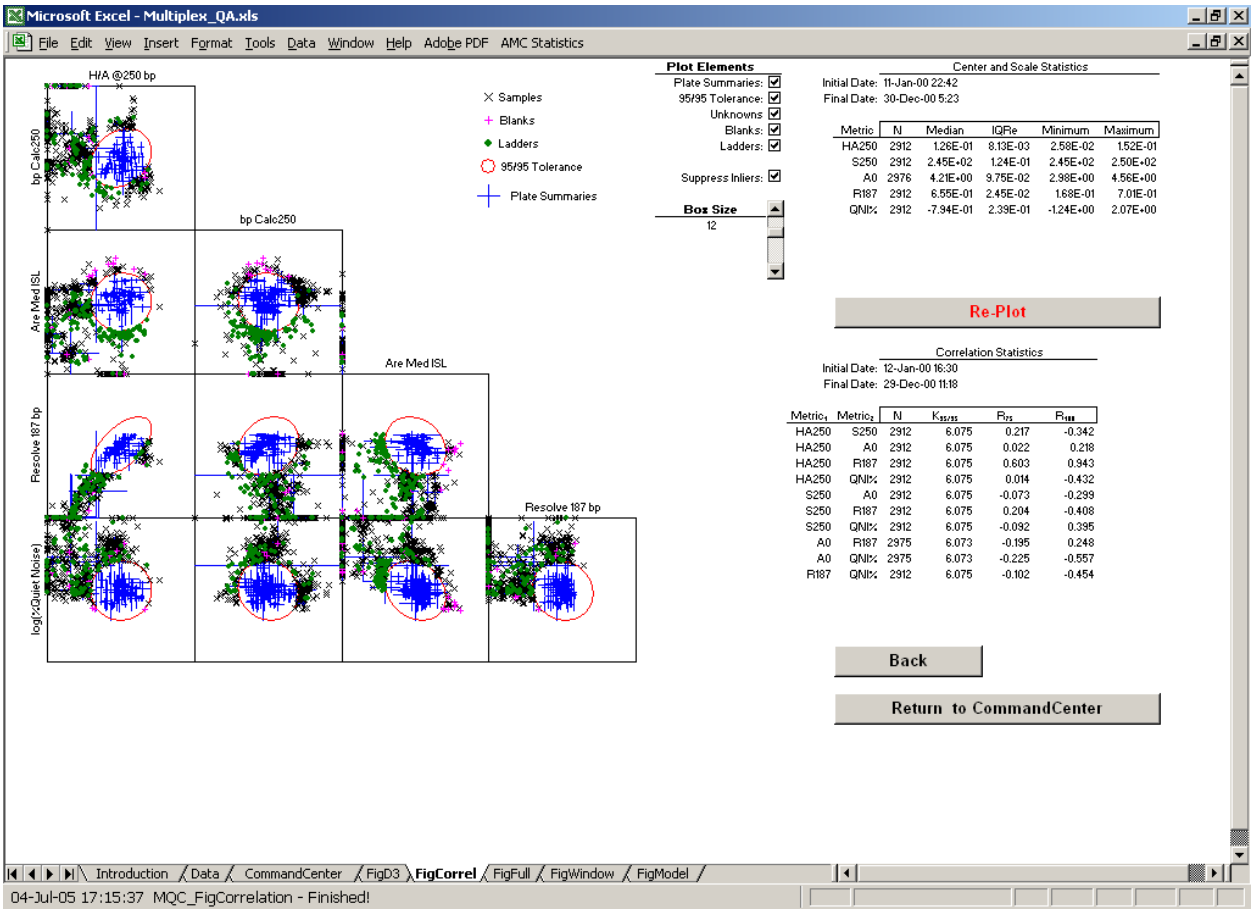
Figure 50. Window E'grams and Models for Two Selected Samples



6 CORRELATION PLOTS

The Correlation Plot displays the $n\times(n-1)/2$, where n is the number of active metrics, pair-wise scattergrams of the data displayed in the current D^3 time-series chart. The Correlation plot is located in the FigCorr worksheet; it is accessed with the **Plot Correlation** button in the FigD3 worksheet. Figure 51 is the Correlation plot for the data shown in the D^3 chart of Figure 45.

Figure 51. Five-Metric Correlation Plot



The format of the Correlation plot is controlled by the buttons, scroll bar, and checkboxes in the FigCorr worksheet. The FigCorr control functions become active only *after* clicking on the **Plot Correlation** button in the FigD3 worksheet. If you click on any of the FigCorr buttons without having first 1) defined a D^3 chart and 2) clicked on **Plot Correlation**, a warning dialog will appear. This dialog must be cleared before you can continue.

A note of warning: The Correlation plot can consume a fair bit of time when there are many data. A "Working…" notification, similar to that of Figure 40, will be displayed while the wheels are spinning.

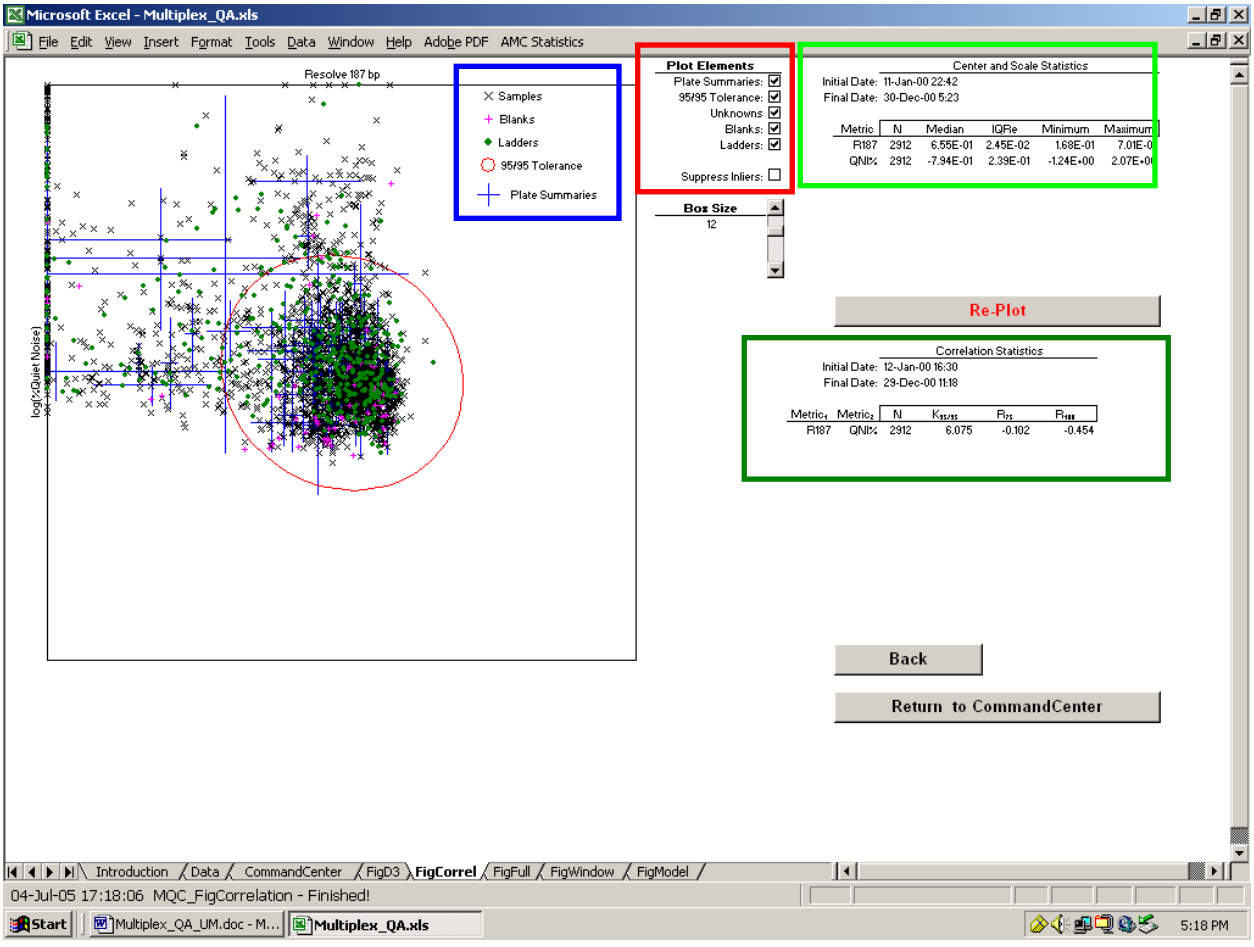
6.1 Selecting the Data for Scattergram Display

Each of the Correlation plot's scattergrams display data for all samples having injection datetimes within the interval specified in the FigD3 worksheet (Section 5.3). This datetime interval is stated in the top part of the "Correlation Statistics" information block below the **Re-Plot** button (the dark green rectangle in Figure 52).

6.2 Specifiable Plot Elements

Figure 52 displays the scattergram for just two of the plot metrics shown in Figure 45 to better reveal the various graphical elements of the Correlation plot. These elements are controlled using the **Plot Elements** checkboxes in the upper center of the FigCorr worksheet (the red rectangle in Figure 52). No change is made to the D^3 chart until **Re-Plot** is clicked.

Figure 52. Two-Metric Correlation Plot



There are checkboxes for five data-representation elements: **Plate Summaries**, **95% Tolerance**, **Unknowns**, **Blank**s, and **Ladders**. At least one of these five must be active. If you attempt to deactivate all five checkboxes, all five will automatically become active. The sixth checkbox, **Suppress Inliers**, controls how the data for the Unknown, Blank, and Ladder sample-types are displayed.

A note of encouragement: Once the basic Correlation plot is created, activating/deactivating these plot elements is very fast.

6.2.1 Plate summaries

For each pair of quality metrics, the blue cross-like "Plate summaries" display the expected range of values for all samples (unknowns, blanks, and ladders) assigned as members of the same plate (Section 2.4.2.6.3). Both the vertical and horizontal lines of each cross are robust 95% confidence intervals for the values of the respective quality metrics, calculated from the median and IQRe estimates of the center and spread of the distribution.

6.2.2 95/95 tolerance ellipse

For each pair of quality metrics, the red ellipse represents a 95/95 tolerance ellipse on their joint distribution. If the data for both metrics are representative of their populations and are more or less normally distributed (single, symmetrical mode near the middle of the pack), this ellipse should, with about 95% confidence, enclose about 95% of the population.

Calculation of the 95/95 tolerance ellipse is described in detail elsewhere [4]. Robust statistics are used for all three types of parameter required: the median for the center location of each metrics, the IQRe for the dispersions about the centers, and R_{75} for the bivariate correlation between the metrics.

6.2.3 Unknowns, blanks, and ladders

The individual values for each pair of quality metrics can be displayed for unknowns (×), blanks (+), and ladders (●). A legend for the types of samples displayed is located at the upper right corner (the blue rectangle in Figure 52) of the Correlation plot.

6.2.3.1 The number of samples that can be displayed

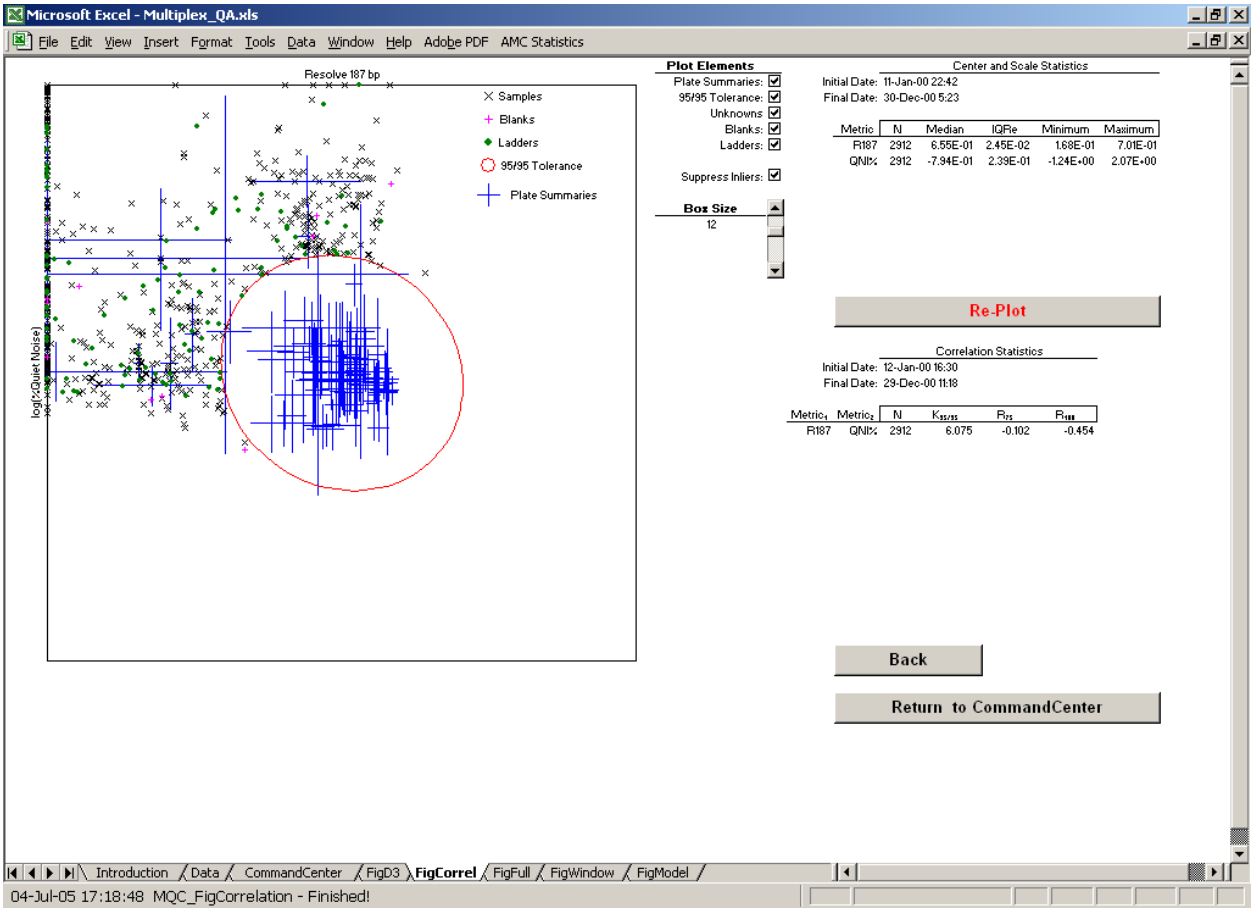
Excel can only display 30,000 data as a single line. This is sufficiently large that it seldom matters, but it impacts the Correlation plot when (the number of samples of a given type) × (the number of active metrics) × (the number of active metrics – 1) / 2 is greater than 30,000. With five active metrics, you can still have up to 3,000 Unknowns or Blanks or Ladders without worrying… but with more than 3,000 of some type, you might notice that the center of the scattergrams is suspiciously empty since the Correlation plot presents data in order of decreasing distance to the center of the scattergram.

If you should have too many data to be displayed, you will be notified of the truncation. If the small central holes truly bug you, activate the **Suppress Inliers** checkbox (Section 6.2.4) to make the central hole meaningful.

6.2.4 Suppress Inliers

When the **Suppress Inliers** checkbox is active, only those samples (Unknowns, Blanks, and Ladders) *outside* of the 95/95 tolerance ellipse will be displayed. Figure 53 is the same as Figure 52, but with inliers suppressed.

Figure 53. Two-Metric Correlation Plot with Inliers Suppressed



If the **Suppress Inliers** checkbox is clicked without at least one of these three elements being active, all three sample-types will be activated. If all three sample-types are deactivated, Suppress inliers will be deactivated.

6.3 Scattergram Construction

Each scattergram within a Correlation plot displays the values for a given pair of quality metrics. Samples can be displayed providing 1) their injection datetime is within the specified datetime interval (Section 5.3) and 2) they have valid numeric values for both metrics. The basic size of the distribution of the data within each of the scattergram boxes is set by the univariate center and scale parameters stored in the SumStat worksheet and the setting of the "Box Width in SDs" scroll bar. The 95/95 tolerance ellipse (Section 6.2.2) calculations require estimates of the bivariate correlation between each pair of metrics.

6.3.1 Univariate summary statistics

The "Center and Scale Statistic" information block at the upper right of the worksheet (the light green rectangle in Figure 52) lists the current center and scale parameters for the active metrics stored on the SumStat worksheet (Section 2.4.5.6). The upper part of the information block lists the earliest and most recent injection datetimes of the samples used. The lower part of

the block lists the code names for the active metrics, the number of valid data for each metric (N), the median, IQRe, minimum, and maximum values.

All of the data displayed in the Correlation plot are "standardized" or "z-scored": $z = (x - \text{Median})/\text{IQRe}$, where Median and IQRe are robust estimates of the location and dispersion of the data. This simple transformation ensures that each of the scattergrams is centered within its assigned box and that all of the distributions are about same size relative to the scattergram box.

A note of explanation: The SumStat-stored parameters are used to center and scale the scattergrams to facilitate comparisons among different datetime intervals.

6.3.2 Bivariate summary statistics

The "Correlation Statistics" information block just below the **Re-Plot** button (the dark green rectangle in Figure 52) lists correlation-related estimates for all samples with injection datetimes within the interval specified in the FigD3 worksheet. The upper part of the block lists the current datetime interval. The lower part of the block lists the code names for the pairs of metrics, the number of valid data pairs within the datetime interval (N), the 95/95 tolerance coverage factor for that number of data ($K_{95/95}$) [4], a robust estimate of correlation (R_{75}), and the usual estimate of correlation (R_{100}).

6.3.3 Box size

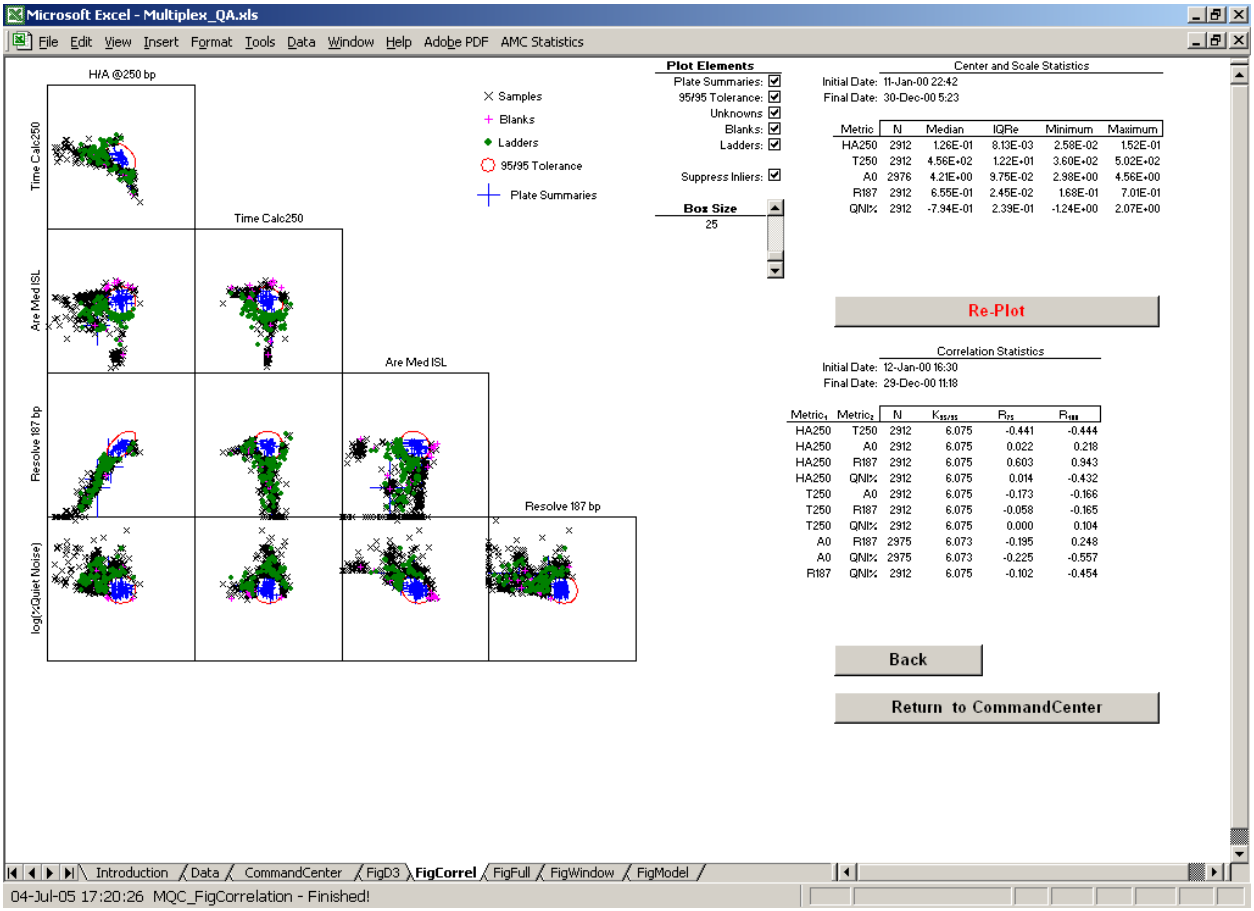
The scroll bar immediately below the **Suppress Inlier** checkbox controls the size of the scattergram distributions relative to the scattergram box. The box size, in units of IQRe-estimated standard deviation, can be from 6 standard deviations to 25 standard deviations.

If all the data belonged to well-behaved, normal distributions, then a scattergram box six standard deviations wide should contain about 99% of all values. In practice, the 12 standard deviations used in Figure 51 seems a more reasonable box size. When the distributions are multi-modal (more than one "most likely" value), even larger boxes can be interesting. Figure 54 re-displays the data shown in Figure 51 with box dimensions of 25 standard deviations.

6.3.4 Data outside the box

Any pair of values that cannot be displayed within its scattergram box is plotted at the edge of the box. Some data, such as the expected average height of the blue-dye peaks for blank samples, may legitimately be *very* far away from the center point of the distribution (Figure 54, bottom row of scattergrams).

Figure 54. Five-Metric Correlation Plot in Bigger Boxes



6.4 Re-Plot

Clicking the **Re-Plot** button causes the Correlation plot to be redrawn using whatever plot options have been specified. None of the options take effect until **Re-Plot** is clicked.

6.5 Back

Clicking the **Back** button transfers control to the FigD3 worksheet.

6.6 Return to CommandCenter

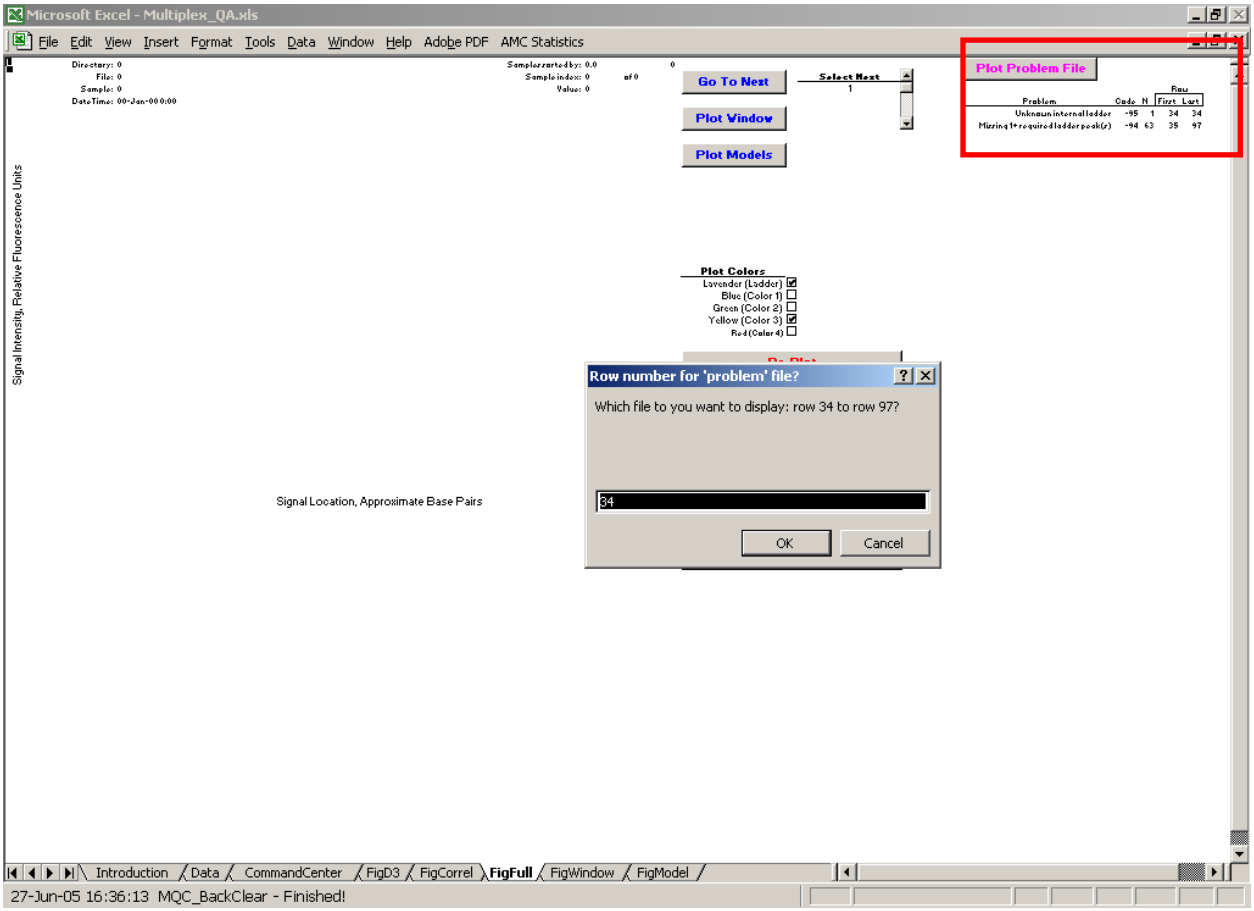
Clicking the **Return to CommandCenter** button transfers control to the CommandCenter worksheet.

7 EXPERIMENTAL FEATURES

7.1 Plotting Problem Files

While there's not much that can be done about samples with truly unreadable files, there are problems that prevent height/area or retention models being calculated for samples but that do *not* prevent reading-in and displaying at least the basic Full E'gram. Two such problems are addressed with the experimental **Plot Problem File** button semi-hidden on the FigFull worksheet. Figure 55 displays where the button lives and the input dialog that appears when you click it.

Figure 55. The Plotting Problem Files Command Button and Input Dialog



7.1.1 Plot Problem File

The **Plot Problem File** button can be used anytime that there is one or more samples with an "Err" code of –95, unrecognized ISS, or –94, known ISS missing one or more critical peaks. These file integrity errors are discussed in Section 2.4.2.6.1. Clicking on **Plot Problem File** sorts the dataset by "Err" code, updates the information block immediately below the button and, if there is at least one relevant sample, asksyou for the index of the sample you wish to visualize. If there is no such problem child, you will be so informed.

7.1.2 The Problem File information block

The information block immediately below the **Plot Problem File** button just lists the different **E'gram**-plottable "Err"s by name and code, the number of samples with each such problem, and the dataset row indices for the first and last sample so afflicted.

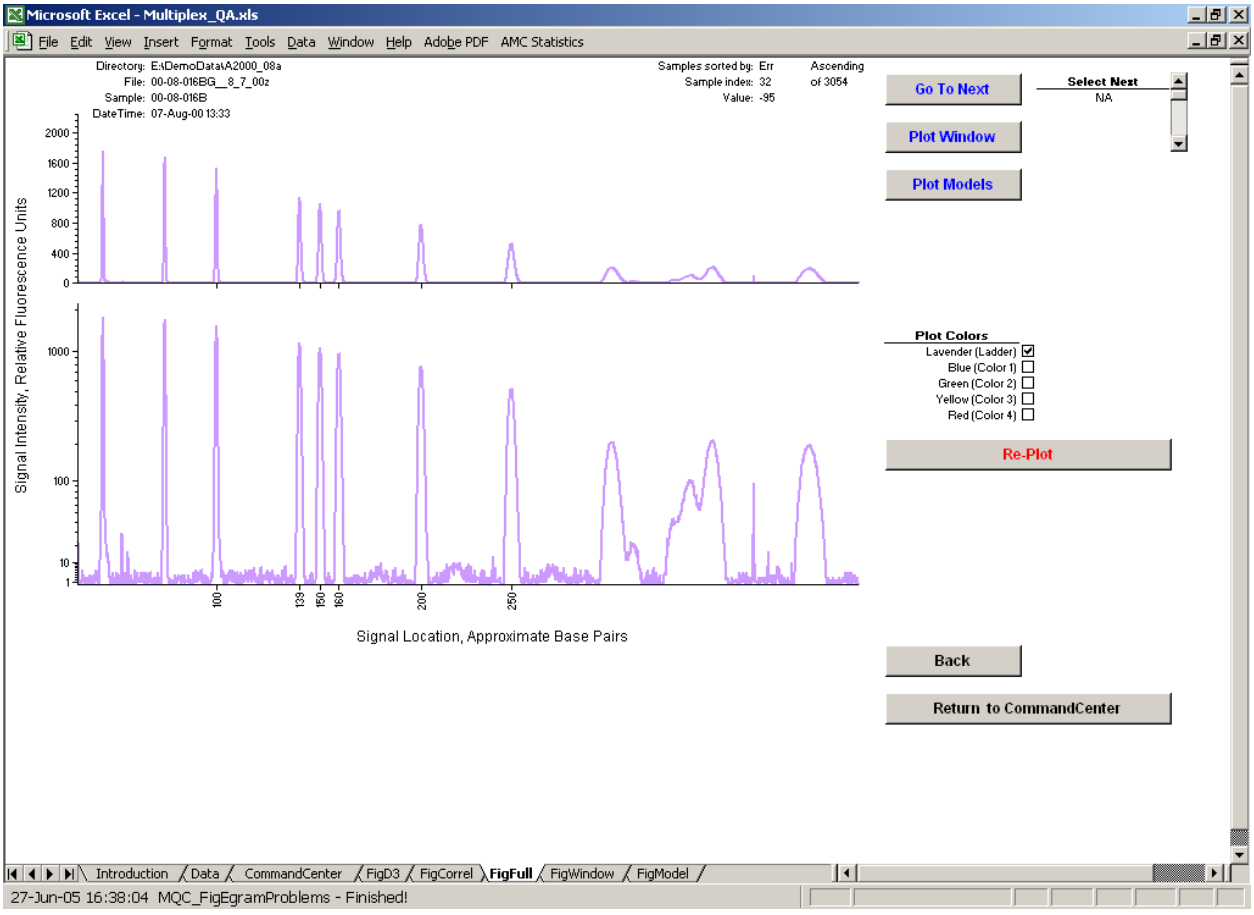
7.1.3 Specifying the Problem File

The input dialog that appears when you click the **Plot Problem File** button asks for the row index of the sample you wish to visualize. You are only allowed to input integer values in the range from the index of the first to the last "bad" file. Clicking the dialog box's **Cancel** clears the information block.

7.1.4 The (Partial) Full E'gram

Figure 56 displays the Full E'gram for one of the problem children in the Demo dataset. Note only is the ISS identifier in the .fsa file unrecognizable (Err code –95), the ladder data indicate that a "meltdown" occurred shortly after the 250 bp peak was detected. Since the injection datetime is mid-day in August in Gaithersburg, it is likely that there was a thunderstorm-related power outage during the run.

Figure 56. Example Problem E'gram



7.1.5 Why is this necessary?

The system that produces the Full E'gram for "good" samples does a fair bit of error checking and setup to ensure that Window E'grams and Model plots can be created. **Plot Problem Files** just tries to display as many of the raw data as can be read and doesn't worry about the details.

Since visualizing these problem samples is not integrated with the rest of the Multiplex_QA system, you can use the **Plot Problem File** button without first clicking on **Plot E'gram** in the CommandCenter.

Since the height/area and retention models cannot be defined for these problem children, neither the Window E'gram nor the Model plots can be invoked from the FigFull worksheet with a **Plot Problem Fil**e-created Full E'gram.

8 QUALITY METRICS

The following "Quality metrics" presented in the following Sections would more accurately be termed "column headings in the Data worksheet," as not all are intended for use in accessing STR multiplex quality.

8.1 Signal Intensity

The metrics listed in Table 6 are simple robust summaries of the height and area of the peaks as estimated by the genotyping software.

Table 6, Metrics Related to Peak Intensity

Code	Name	Description
N0	Pk# ISS	Number of ISS dye peaks
N1	Pk# Blu	Number of "Blue" dye peaks
N2	Pk# Grn	Number of "Green" dye peaks
N3	Pk# Ylw	Number of "Yellow" dye peaks
N4	Pk# Red	Number of "Red" dye peaks
A0	Are Med ISS	Median of ISS dye peak log10(areas)
A1	Are Med Blu	Median of "Blue" dye peak log10(areas)
A2	Are Med Grn	Median of "Green" dye peak log10(areas)
A3	Are Med Ylw	Median of "Yellow" dye peak log10(areas)
A4	Are Med Red	Median of "Red" dye peak log10(areas)
As0	Are IQR ISS	Inter-Quartile Range of ISS dye peak log10(areas)
As1	Are IQR Blu	Inter-Quartile Range of "Blue" dye peak log10(areas)
As2	Are IQR Grn	Inter-Quartile Range of "Green" dye peak log10(areas)
As3	Are IQR Ylw	Inter-Quartile Range of "Yellow" dye peak log10(areas)
As4	Are IQR Red	Inter-Quartile Range of "Red" dye peak log10(areas)
H0	Hgt Med ISS	Median of ISS dye peak log10(heights)
H1	Hgt Med Blu	Median of "Blue" dye peak log10(heights)
H2	Hgt Med Grn	Median of "Green" dye peak log10(heights)
H3	Hgt Med Ylw	Median of "Yellow" dye peak log10(heights)
H4	Hgt Med Red	Median of "Red" dye peak log10(heights)
Hs0	Hgt IQR ISS	Inter-Quartile Range of ISS dye peak log10(heights)
Hs1	Hgt IQR Blu	Inter-Quartile Range of "Blue" dye peak log10(heights)
Hs2	Hgt IQR Grn	Inter-Quartile Range of "Green" dye peak log10(heights)
Hs3	Hgt IQR Ylw	Inter-Quartile Range of "Yellow" dye peak log10(heights)
Hs4	Hgt IQR Red	Inter-Quartile Range of "Red" dye peak log10(heights)

8.1.1 Number

The number of peaks of a given color identified by the genotyping software.

8.1.2 Expected Peak Area and variability of Peak Areas

The median of the $\log_{10}$-transformed peak areas provides a robust description of the expected peak areas for each color of the identified peaks; the back-transformation of these values (10^{median}) is essentially the same as the median of the peak areas. The IQRe of the $\log_{10}$-transformed peak areas provides a robust description of the variability of the peak areas; the back-transformation of these values (10^{IQRe}) is a multiplicative factor standard deviation (*not* the usual ± additive factor) about the expected peak area [5].

8.1.3 Expected Peak Height and variability of Peak Heights

The median of the $\log_{10}$-transformed peak heights provides a robust description of the expected peak heights for each color of the identified peaks; the back-transformation of these values (10^{median}) is essentially the same as the median of the peak heights. The IQRe of the $\log_{10}$-transformed peak heights provides a robust description of the variability of the peak heights; the back-transformation of these values (10^{IQRe}) is a multiplicative factor standard deviation (*not* the usual ± additive factor) about the expected peak height [5].

8.2 Peak Symmetry

The metrics listed in Table 7 are simple robust summaries of a peak symmetry parameter,

$$\frac{T_{maximum} - T_{start}}{T_{end} - T_{maximum}}$$

where $T_{maximum}$ is the Time Index of the peak maximum assigned by the genotyping software, T_{start} is the Time Index of when the peak starts, and T_{end} is the Time Index of when the peak ends.

Table 7, Metrics Related to Peak Symmetry

Code	Name	Description
S0	Sym Med ISS	Median of ISS dye peak left/right symmetry
S1	Sym Med Blu	Median of "Blue" dye peak left/right symmetry
S2	Sym Med Grn	Median of "Green" dye peak left/right symmetry
S3	Sym Med Ylw	Median of "Yellow" dye peak left/right symmetry
S4	Sym Med Red	Median of "Red" dye peak left/right symmetry
Ss0	Sym IQR ISS	Inter-Quartile Range of ISS dye peak left/right symmetry
Ss1	Sym IQR Blu	Inter-Quartile Range of "Blue" dye peak left/right symmetry
Ss2	Sym IQR Grn	Inter-Quartile Range of "Green" dye peak left/right symmetry
Ss3	Sym IQR Ylw	Inter-Quartile Range of "Yellow" dye peak left/right symmetry
Ss4	Sym IQR Red	Inter-Quartile Range of "Red" dye peak left/right symmetry

The median of the ratios of provides a robust description of the expected peak left/right symmetry for each color of the identified peaks. The IQR of these ratios provides a robust description of the variability of the peak symmetries.

8.3 Height/Area Model

The metrics listed in Table 8 are estimated from the peak height, area, and bp size values for the ISS peaks between 150 bp and 350 bp, excluding the 250 bp peak. The metrics result from a least squares regression of the height/area (H/A) ratio onto the $\log_{10}(bp)$ and $1/\log_{10}(bp)$ of these peaks:

$$H/A_{calc,i} = a + b \times \log_{10}(bp_i) + c / \log_{10}(bp_i) \ .$$

Table 8, Metrics Related to Peak Resolution

Code	Name	Description
Ndf	Degrees of Freedom	H/A = a+bLg(bp)+c/Lg(bp): Number peaks – number parameters
r^2HA	H/A correlation	H/A = a+bLg(bp)+c/Lg(bp): Correlation^2 for 150-350 ISS peaks
rmsHA	H/A RMS error	H/A = a+bLg(bp)+c/Lg(bp): RMS for 150-350 ISS peaks
a	H/A intercept	H/A = a+bLg(bp)+c/Lg(bp): a
b	H/A 1/bp coef	H/A = a+bLg(bp)+c/Lg(bp): b
c	H/A bp coef	H/A = a+bLg(bp)+c/Lg(bp): c
εa	H/A intercept err	H/A = a+bLg(bp)+c/Lg(bp): Sigma(a)
εb	H/A 1/bp coef err	H/A = a+bLg(bp)+c/Lg(bp): Sigma(b)
εc	H/A bp coef err	H/A = a+bLg(bp)+c/Lg(bp): Sigma(c)
Pthin	H/A Min(Obs/Calc)	H/A: Minimum Observed/Predicted, all peaks of all colors
Pwide	H/A Max(Obs/Calc)	H/A: Maximum Observed/Predicted, all peaks of all colors
HA250	H/A @250 bp	H/A = a+bLg(bp)+c/Lg(bp): Predicted H/A for 250 bp

A note of explanation: The regression is *actually* performed on orthogonal variables that are linearly *related* to $\log_{10}(bp)$ and $1/\log_{10}(bp)$, Section 11.2. This doesn't impact the regression statistics nor their predictive utility in any way (other than making the calculations a bit messy), but it does make the values of the three regression parameters of comparable size and somewhat better fit for use as quality metrics.

8.3.1 Regression metrics

The initial nine metrics in Table 8 are provided by Excel's LINEST linear regression function: the number of degrees of freedom (Ndf), the square of the correlation between the observed and predicted values (r^2HA), the root-mean-square expected difference between the observed and predicted values (rmsHA), the regression parameters (a, b, and c), and the expected uncertainties of the regression parameters (εa, εb, and εc). These metrics are derived from the ISS peaks used in the regression.

8.3.2 Pthin and Pwide

The next two metrics in Table 8 provide a rough guide to samples with abnormally thin peaks (e.g., spikes) and wide peaks (e.g., dye blobs). They are both calculated observed height/area ratio divided by the predicted ratio for a peak of that bp size, using the regression model derived from the ISS peaks. Both the minimum (Pthin) and maximum (Pwide) ratio-of-ratios are stored.

Neither of these metrics is intended for use in D^3 charts; rather, they provide a mechanism for quickly screening electropherograms for "interesting" abnormal events using Full and/or Window E'gram (Section 3.2.4.5.3).

8.3.3 Predicting the 250 bp peak

The final "resolution model" metric in Table 8 (HA250) is (probably) the most useful: the height/area value predicted using the regression model for a peak of size 250 bp. Since the regression model uses peaks on either side of a nominal 250 bp ISS peak but not the peak itself (if present), this prediction is expected to be "the best" value that can be obtained from the model. That is, if the model can't predict at its center, it won't predict well at any bp.

8.4 Retention Model

The metrics listed in Table 9 are estimated from the bp size and the Time Index of the peak maximum, standardized so that the Time Index is zero for the nominal 200 bp peak, for the ISS peaks of size 150 bp to 250 bp, excluding the 250 bp peak. The metrics result from a quadratic least squares regression of the Time Index onto the bp size of these peaks:

$$TI_{calc,i} - TI_{200} = \alpha + \beta \times bp_i + \gamma \times bp_i^2 \ .$$

Table 9, Metrics Related to Peak Retention

Code	Name	Description
Ndf	Degrees of Freedom	T-T200 = $\alpha+\beta bp+\gamma bp^2$: Number of peaks – number of parameters
r^2T	Time correlation	T-T200 = $\alpha+\beta bp+\gamma bp^2$: Correlation^2 for 150-350 ISS peaks
rmsT	Time RMS error	T-T200 = $\alpha+\beta bp+\gamma bp^2$: RMS for 150-350 ISS peaks
α	Time bp intercept	T-T200 = $\alpha+\beta bp+\gamma bp^2$: α
β	Time bp coef	T-T200 = $\alpha+\beta bp+\gamma bp^2$: β
γ	Time bp^2 coef	T-T200 = $\alpha+\beta bp+\gamma bp^2$: γ
$\varepsilon\alpha$	Time bp intercept err	T-T200 = $\alpha+\beta bp+\gamma bp^2$: Sigma($\alpha$)
$\varepsilon\beta$	Time bp coef err	T-T200 = $\alpha+\beta bp+\gamma bp^2$: Sigma($\beta$)
$\varepsilon\gamma$	Time bp^2 coef err	T-T200 = $\alpha+\beta bp+\gamma bp^2$: Sigma($\gamma$)
T200	Time Obs200	T-T200 = $\alpha+\beta bp+\gamma bp^2$: Time of 200 bp ISS peak
T250	Time Calc250	T-T200 = $\alpha+\beta bp+\gamma bp^2$: Predicted Time for 250 bp
rmsbp	bp RMS error	bp = F^{-1}(T-T200): RMS for 15-350 ISS peaks
S250	bp Calc250	bp = F^{-1}(T-T200): Predicted bp for nominal 250 ISS peak
D250	bp S250-250	bp = F^{-1}(T-T200): S250 - 250
Abias	bp Average bias	bp = F^{-1}(T-T200): All color average difference
Sbias	bp Stdev bias	bp = F^{-1}(T-T200): All color standard deviation

A note of explanation: The regression is *actually* performed on orthogonal variables that are linearly *related* to bp and bp^2, Section 11.2. As with the resolution model, this doesn't impact the regression statistics nor their predictive utility in any way but it does make the values of the three regression parameters of comparable size and somewhat better fit for use as quality metrics. Likewise, the standardization of the Time Index to the nominal 200 bp peak is intended to numerically stabilize the constant term, α.

8.4.1 Regression metrics

The initial nine metrics in Table 9 are provided by Excel's LINEST linear regression function: the number of degrees of freedom (Ndf), the square of the correlation between the observed and predicted values (r^2T), the root-mean-square expected difference between the observed and predicted Time Index values (rmsT), the regression parameters (α, β, and γ), and the uncertainties on these parameters ($\varepsilon\alpha$, $\varepsilon\beta$, and $\varepsilon\gamma$). These metrics are derived from the ISS peaks used in the regression. The T200 metric is just the observed Time Index for the nominal 200 bp peak. The T250 metric is the predicted Time Index for a "real" peak of 250 bp size.

8.4.2 Inverse regression metrics

The next five metrics in Table 9 are derived from the inverse regression model (bp size as a function of Time Index): the root-mean-square expected difference between the observed and expected bp size values (rmsbp), the predicted bp size of the nominal 250 bp peak (S250), the bp bias of the nominal 250 bp peak (D250 = S250 – 250), the average bp bias over all non-ISS peaks between the peak bp sizes and those predicted by the inverse regression model (Abias), and the standard deviation of the bp bias of all peaks of all colors (Sbias). For numerical reasons, the inverse regression model is derived from the Time Index as a function of bp size model via your old friend, the "quadratic equation" of High School algebra.

8.5 Resolution Metric

The R187 metric is an estimate of the resolution of peaks at 186 bp and 187 bp (the bp size of the HUMTH01 9.3 and 10 alleles in the COfiler multiplex). This metric uses the height/area and retention model parameters to estimate the standard chromatographic definition of resolution for these two peaks [3]

$$R_{187} = 2\frac{(TI_{187} - TI_{186})}{W_{186} + W_{187}}$$

$$TI_x = \alpha + \beta x + \gamma x^2$$

$$W_x = 2A_x/H_x = 2/(a + b\log_{10}(x) + c/\log_{10}(x))$$

where TI_x is the Time Index for the 186 bp or 187 bp peaks and W_x is the width of the 186 or 187 peaks, assuming that the peaks can be validly approximated as triangles.

Table 10, Metrics Related to RFU Noise

Code	Name	Description
QNht	Mean ISS Height	Average height of ISS model peaks
QN#	# quiet ISS data	Number of "quiet" data used to calculate noise
QNav	Mean QN	Average of "quiet" data
QNsd	SD QN	Standard deviation of "quiet" data
QNl%	log(%QN)	$\log_{10}(100 \times QNsd / (QNht - QNav))$

8.6 Noise Metrics

The "noise" characteristic of a given analysis is estimated using the "quiet" intervals between ISS peaks. Table 10 lists the currently available noise-related metrics.

8.6.1 Estimating noise

The "noise" in an analytical signal is everything about the signal that you wish wasn't there. Since the signal for the ISS consists, in principle, of just a relatively few known peaks, the signal present in the "quiet" inter-ISS peak regions can be used to estimate one of the common types of analytical noise: the "blank" signal that's there when nothing interesting is going on.

The Time Index values of the peak maximum ($T_{maximum}$), peak start (T_{start}), and peak end (T_{end}) are used to identify the "quiet" regions between any pair of adjoining ISS peaks, T_1 and T_2. To minimize the contribution of the peaks to the noise, the quiet interval is defined as beginning at $T_{1,end}+(T_{1,end}-T_{1,maximum}) = 2\times T_{1,end}-T_{1,maximum}$ and ending at $T_{2,start}-(T_{2,maximum}-T_{2,start}) = 2\times T_{1,start}-T_{1,maximum}$. For the GS350+ family of ISS, the two intervals among the nominal 200 bp, 250 bp, and 300 bp peaks are used. For GS400, the four intervals among the nominal 200 bp, 220 bp, 240 bp, 260 bp, and 280 bp peaks are used. For ILS, the four intervals among the nominal 200 bp, 225 bp, 250 bp, 275 bp, and 300 bp peaks are used.

The metric "N#" is the number of data within all of the "quiet" intervals used in the noise estimate. The metric "Navg" is the mean of these data and "Nsd" is their standard deviation.

8.6.2 Quiet noise

The expected average quiet signal is one-half of the observed standard deviation of the signal when there truly is nothing "interesting" going on. If the average is much larger than this, there may well be an undesired component to the ISS signal (baseline shift?) but it's not the same sort of undesired signal as if the standard deviation grows along with the average ("pull up" or dye blob?). If the standard deviation of the quiet signal is proportional to the height of the ISS peaks, then the "noise" should not much effect the identification and characterization of the ISS peaks.

The effective level of quiet noise independent of baseline shifts and the strength of the signal for the ISS peaks can be estimated as the ratio of the quiet interval standard deviation to the difference between the average ISS peak height and the average baseline:

$$\%QN = \frac{100\times Nsd}{Npkh - Navg}$$

where Npht is the geometric mean of the peak heights of just the six (for the GS350+ family) or eight (for GS400 and ILS) peaks used to define the Height/Area and Retention models. The Npht metric is closely related to the signal intensity metric H0 (Section 8.1.3) but it is less sensitive to the presence of signal artifacts. Since %QN can be large if there are artifact peaks in the ISS signal, the metric is reported as in logarithmically transformed form:
$l\%QN = \log_{10}(\%QN)$.

9 WHEREFORE DATA?

The Multiplex_QA system is designed to use data contained in text files produced by the BatchExtract system developed in Dr. Stephen Sherry's Group at the National Center for Biotechnology Information of the National Library of Medicine. BatchExtract is designed to interpret the .fsa binary files generated by various flavors of Applied Biosystems (ABI) software. Therefore, to analyze data in Multiplex_QA you must first convert it to the format provided by BatchExtract.

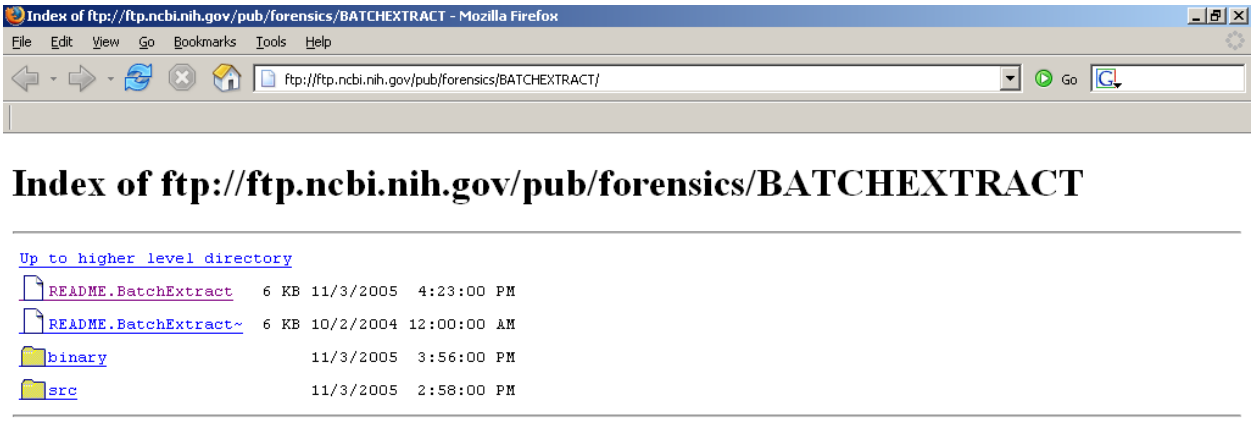
Section 9.1 describes how to obtain the BatchExtract system. Section 9.2 describes how to use BatchExtract to convert ABI .fsa binary files into analogous text files. Section 9.3 details the formats of the BatchExtract files used by Multiplex_QA, just in case some hearty (or desperate?) soul wants to use Multiplex_QA with data from a non-ABI system.

9.1 Getting BatchExtract

9.1.1 Where it lives

The BatchExtract system is publicly available and may be downloaded using your favorite web-browser from ftp.ncbi.nih.gov/pub/forensics/BATCHEXTRACT. You should see something similar to Figure 57 when you get to this site.

Figure 57. BatchExtract FTP site



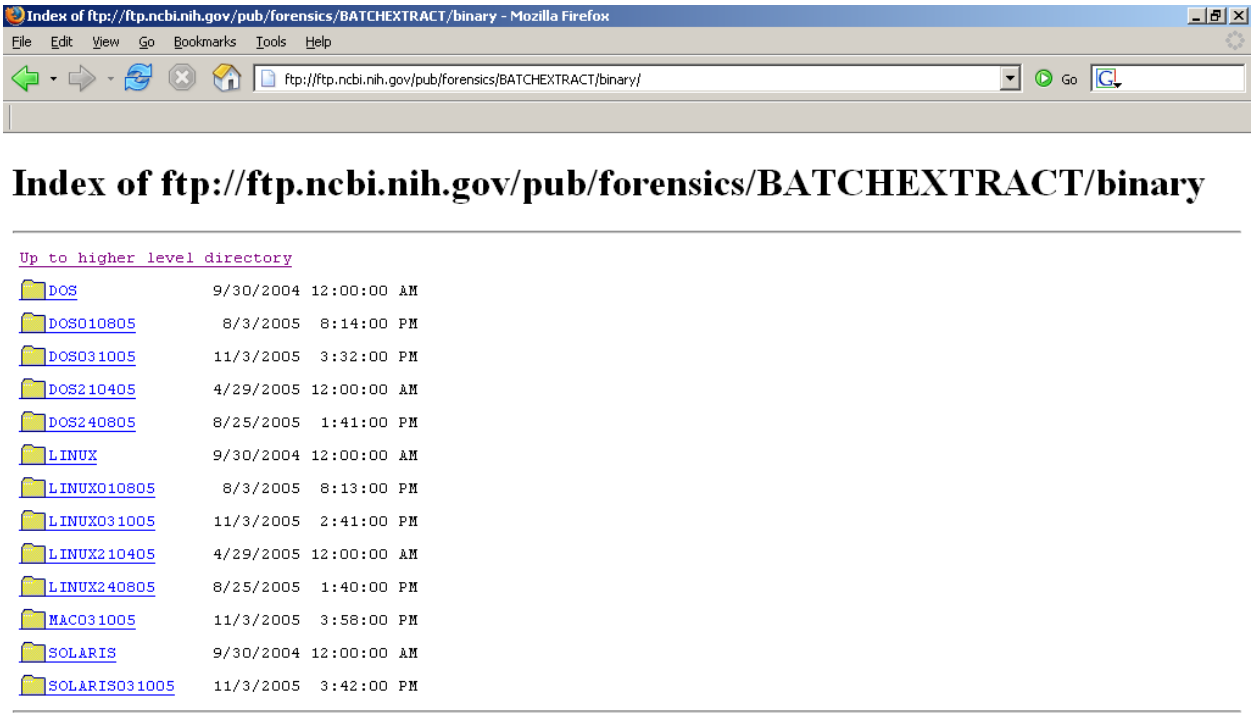
9.1.2 What to do with the README.BatchExtract text file

The README.BatchExtract file provides basic information about the current BatchExtract system. Click on the most recent version (currently, 3-Nov-05) of this text file, read it, and print it out for future reference. At this point, don't worry about *understanding* all of the technical stuff it contains. Unless you are an experienced programmer, much of it won't make sense until you've successfully used BatchExtract a few times.

9.1.3 How to fetch a BatchExtract executable system

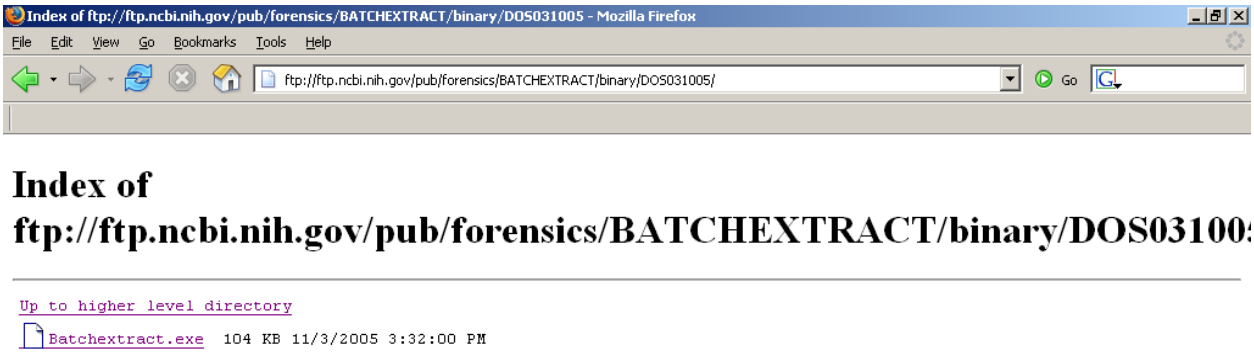
The "binary" folder contains executable systems for DOS, LINUX, Macintosh, and SOLARIS operating system. Click on the "Binary" folder and you should see something like Figure 58.

Figure 58. BatchExtract/binary



Locate the most recent folder suitable for your operating system. (Since you're still reading this, you probably are using an "IBM PC" computer running some version of the WINDOWS operating system. You need to use a "DOS" executable and so need to select one of the "DOSxxxxxx" folders. Currently, the most recent version of BatchExtract is in "DOS031005"). Click on the chosen folder and you should see something like Figure 59.

Figure 59. BatchExtract FTP Site /Binary/DOS031005



The sole content of the chosen folder is the needed executable binary: here, batchextract.exe. Click on this file; your web-browser will probably ask you whether to open the file or save it. Depending on your browser, it may also allow you to specify where the file should be stored. If you can't "browse" to your folder of choice, the file will be stored in whatever folder you have specified as the default storage area (often your active desktop). Select the "save" option and do whatever the browser's window requires you to do to go on. The file size is currently about 100 kB, so the download doesn't take long. If you weren't able to put the file directly where you want it to be stored, locate where the browser put it and drag it to where you want it.

If this web-browser approach does not work for you, ask your IT-guru how to use Telnet or some other FTP system.

9.1.4 Where it should store the executable files

To simplify your life, you should save the BatchExtract executable in a convenient root-level directory on a hard-drive with considerable free space. For WINDOWS operating systems, click on "My Computer," click on your (most appropriate) drive, and create a new folder named something simple like "DNA." If you do this on, say, drive "C:" then the full path to the created directory would be "C:\DNA."

9.2 Running BatchExtract

Earlier versions of BatchExtract were somewhat tedious to use, mostly due to Excel's inability to recognize more than 256 files in a single file directory. The current BatchExtract is much more convenient: you no longer need to futz around with moving the .fsa files into special folders or renaming .fsa files having the same filename. However, it is still a good idea to "start small". Until you've successfully worked through the following BatchExtract instructions and used **Process Directories** (Section 2.4.2) with a small set (say 10 to 20) of .fsa files, don't get too carried away.

BatchExtract does **not** yet run under WINDOWS, it runs under DOS. If you are using an "IBM PC", fear not: all WINDOWS systems provide a MS-DOS emulator ("shell"). If you don't already have a shortcut to such installed on your desktop, you should be able to find one by searching for "dos" or "command prompt" on the hard-drive your system is installed on. If you can't locate it, contact your local IT-guru. Note: "Command Prompt" is a newer and considerably advanced version of the traditional DOS shell, even if it is less obviously named. Use "Command Prompt" if you have it.

9.2.1 Get ready

Once you have located the DOS shell, click on it. A DOS window should open up with a copyright notice from the early 1990s and a ridiculously uninviting string of characters ending with a ">" and a blinking underscore.

9.2.1.1 Changing hard-drive designation

The initial DOS display is telling you what directory the DOS shell is located in and inviting you to make trouble. You need to tell it to go to the directory where the BatchExtract executable

is. You do this by first telling it which disk drive the desired file is located on, say "C." So type "C:" followed by <Enter> ("carriage return" to us fossils.)

9.2.1.2 Changing current directory

You then need to tell DOS which directory on this disk drive the BatchExtract.exe file is located. You do this by typing "cd *xxx* <Enter>" where "*xxx*" is the actual complete directory name for where you stuck BatchExtract. If you chose not to make life simple and have located the BatchExtract system in some remote corner of your file system, you will need to figure out where it is and how to describe that to DOS (i.e., cd *xxx**yyy**zzz*\…). When in doubt, cry on your IT-guru's shoulder – although he/she may be getting tired of talking with you by now.

9.2.2 Get set

Once this is accomplished, type "BatchExtract" to make sure everything is working. It should spit out at you a short list of the available options. The upper lines in Figure 60 show the process and the results on my system, where BatchExtract.exe is located in the NLM_DNA folder on hard-drive D.

9.2.3 Go

The keep-life-simple way of actually running BatchExtract is to locate a master directory that contains all of the .fsa files and/or all of the subdirectories that contain the .fsa files of interest, say "C:\DNA\myinput", and store the output files into a similarly named master folder, say "C:\DNA\myoutput." For these input and output directories, the command you need to type into the DOS window is:

BatchExtract –idpath .*myinput* –odpath .*myoutput* –mqa

The "." in the *myinput* and *myoutput* specifications is DOS-shorthand for "the current directory," but you could use "*C:DNA**myinput*" and "*C:DNA**myoutput*" if you want. The "–"character labels the characters immediately to the right as an instruction. The instructions don't have to be specified in any particular order. Things between instructions (like .*myinput*) are arguments to the immediately preceding instruction. Note that every separate item in every DOS command must be separated by one (and only one) "blank" – the character generated when you press the spacebar key.

The next-to-last lines in Figure 60 show an actual example on my system, where the desired input .fsa files are in directory D:\NLM_DNA\3100 and I want the extracted files to go to directory D:\NLM_DNA\3100o. It is easiest if you let BatchExtract create the directory that will hold the extracted files, but you can create it beforehand if you want. The "-mqa" at the end of the command tells BatchExtract to process all .fsa files within the 3100 input folder and store them in subfolders holding no more than 255 sets of processed files within an output folder named 3100o.

Other currently available options allow you to: specify the maximum number of files to be stored in a single subdirectory (-mqan *xxx*); process single files (-iABI), single directories (-batch), and all files within an entire "tree" of subdirectories (-r); control whether a trace file is

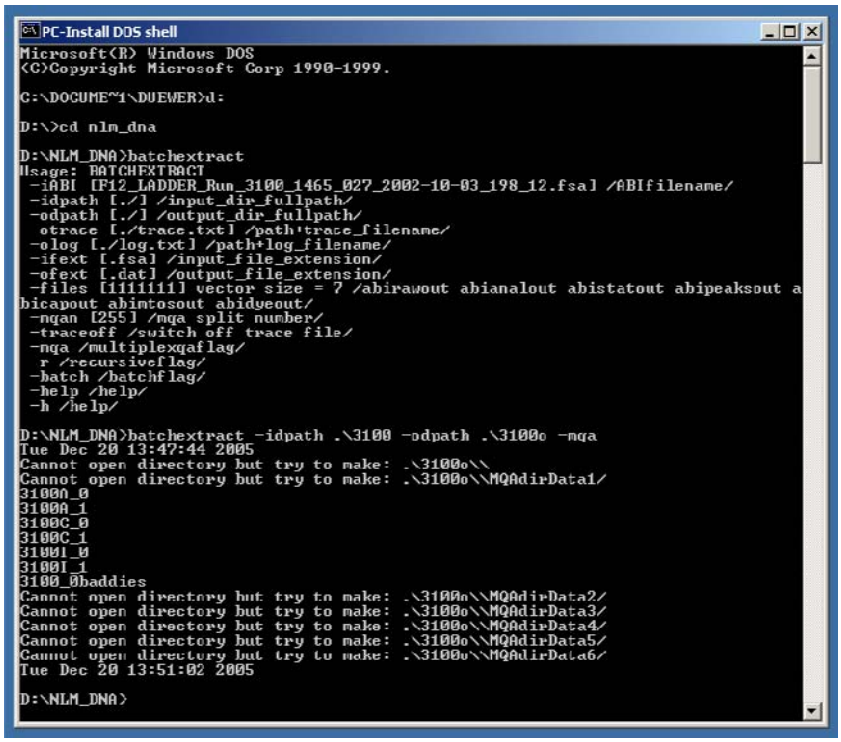
generated (-traceoff) and, if so, where it is stored (-otrace *xxx*); control where the log file is stored (-olog *xxx*), control what the input and output file extensions are (-ifext *xxx* and –ofext *xxx*). The "*xxx*" in the above represents the desired number, string, or directory appropriate to the instruction. The –files option is described in 9.3.

If a duplicate .fsa filename (such as "Ladder.fsa" or "Control.fsa") is encountered whilst stuffing things into the same output subdirectory, the basename of the duplicate will be augmented with an integer index before its BatchExtract files are stored.

9.2.4 Going…

If things go as they should, your DOS window will remain basically unchanged (other than noting when a new subdirectory is created) until BatchExtract has finished processing and issues another command-prompt (the last line in Figure 60).

Figure 60. Starting BatchExtract in DOS window



9.2.5 …or not

There are an infinite number of ways to miss-specify both DOS and BatchExtract commands. You have to get things *exactly* right or… well, mostly just nothing much happens other than having to stare at fairly uninformative error messages. And DOS shells don't make reusing/editing lines easy (use the arrow keys or the right mouse button), so it is often easier to retype the whole line from scratch. If you are prone to typos and/or want to use long, informative directory names… you'd better learn how to setup and run DOS "batch" files… …from someone else. However, if you "keep it simple" then you really don't need to muck

about in DOS very much at all. Note that one of the few good things about DOS is that it doesn't care about the "case" of the letters you type; upper or lower, doesn't matter.

9.2.6 Other DOS trivia

If you want to clear the DOS window, "CLS." When you are finished, "EXIT." If you want to list all the files, "DIR." If you need to be reminded about all the wondrous things you can do with DOS, "HELP"; if you need details on the "DIR" command, "HELP DIR." If you want to create a compact list of all the .fsa files in the current directory, "DIR .fsa /B"; if you want the list to go to a text file called "TEXT.TXT," "DIR .fsa /B > TEXT.TXT."

If you really want to harvest the power of this command-line operating system, you may want to obtain the sanity-saving *DOS for Dummies* [6].

9.2.7 BatchExtract text files

Unless told otherwise, BatchExtract produces the seven files listed in Table 11 for every .fsa file it successfully reads. The "Appended Name" text is appended to the files' basename (that is, the text to the left of the of ".fsa" extension.)

Table 11, BatchExtract Text Files

Appended Name	Stores Information On…	Required?
_CapData.dat	EPT voltage, power, current, and temperature	No
_DyeData.dat	Dye, sample	Yes
_MtoData.dat	Machine, model, and operating conditions	Yes
_PeakData.dat	Parameters for all "called peaks"	Yes
_RawData.dat	"Raw" intensity data for all colors	No
_ScanData.dat	Baseline-corrected intensity data for all colors	Yes, for E'grams
_StatData.dat	Analysis parameters	Yes

BatchExtract also creates two text files, log.txt and trace.txt, that keep track of what has and hasn't been processed. The log.txt files records all unexpected events. The trace.txt file records the processing of each .fsa file encountered. Both files are located in the directory where BatchExtract.exe lives. The information in these files is mostly of interest to the system's developers rather than to its users.

Note the *CapData.dat and *RawData.dat files are not used by Multiplex_QA. To generate just the files that Multiplex_QA requires, add the instruction " –files 0111011" to the end of the DOS command invoking BatchExtract. These two file types can also be easily, if not terrifically conveniently, done using the "DEL *capdata.dat" and DEL *rawdata.dat" commands from within a DOS shell. Clicking **Delete Files** (Section 2.9.3), semi-hidden on the CommandCenter worksheet, will more gracefully remove these files.

9.2.8 Getting ready for Multiplex_QA

The –mqa option of the current BatchExtract automatically groups input files into sets of no more than 255 files and stores them into subfolders inside the output folder you specified in the

BatchExtract command. The subfolders are named "MQAdirData#," where the # is an integer index. You will need these subfolder names to pull these data into Multiplex_QA with **Process Directories** (Section 2.4.2).

9.3 The BatchExtract Files

Multiplex_QA uses five of the text files BatchExtract produces from each .fsa file evaluated. The following Sections illustrate the formats of these files and describe the information that Multiplex_QA requires.

9.3.1 _DyeData.dat

The _DyeData.dat file contains two useful pieces of information: the names of the dyes used in the multiplex and the name of the sample analyzed. Table 12 lists the file for one particular sample. The bits that Multiplex_QA expects are in **bold** text. All other contents of this file are ignored.

Multiplex_QA expects "DYES" to be the first characters of row 18 and the names of the dyes, separated by tabs, to be on row 19. "SMPL" is expected to be the first characters of row 22 and the sample name to appear on row 23. The sample name is taken as all characters from the first character of row 23 up to the first tab.

Table 12, Format of the BatchExtract _DyeData.dat Files

Row	Content
1	AOFF:
2	0 0 0 0
3	CCDF:
4	0 0
5	0 0
6	0 0
7	0 0
8	PK_#:
9	0 0 0 0
10	SDOK:
11	0 0 0 0
12	SDNM:
13	0 0 0 0
14	SDSM:
15	Recp Recp Recp Recp
16	DYEN:
17	
18	**DYES:**
19	**Joe Fam Tamra Rox**
20	DYEZ:
21	
22	**SMPL:**
23	**00-08-002B** 00-08-002B 00-08-002B
24	SDIC:
25	…

9.3.2 _MtoData.dat

The _MtoData.dat file contains three useful (instrument name, sample tube, and lane number) and one critical (injection datetime) items of information. Table 13 lists the file for one

particular sample. The bits that Multiplex_QA expects are in **bold** text. All other contents of this file are ignored.

Multiplex_QA expects the character strings "MACHSPEC_MCHN:" and MACHSPEC_DYSN:" to be in row 3 of the file and the name of the instrument to be between the strings. The string "MACHSPEC_TUBE:" is expected to be in row 4 and the tube designation to immediately follow the ":". The strings "MACHSPEC_LANE:" and "MACHSPEC_LNTD:" are expected to be in row 5 and the lane designation to be between the strings. The strings "date:" and "time:" are expected to be in both rows 10 and 14, with the calendar date between the strings and the time of day immediately following "time:". The datetime specified in row 14, the time that data collection began, is evaluated to ensure that the "injection" datetime in row 10 is valid.

Table 13, Format of the BatchExtract _MtoData.dat Files

Row	Content
1	Machine Speciality:
2	MACHSPEC_InSc:5 MACHSPEC_InVt:15000 MACHSPEC_Tmpr:60
3	**MACHSPEC_MCHN:ABI PRISMA 310 MACHSPEC_DySN:** MACHSPEC_CMNT: MACHSPEC_StdA:
4	MACHSPEC_MODL:310 **MACHSPEC_TUBE:A11**
5	**MACHSPEC_LANE:7 MACHSPEC_LNTD:30** MACHSPEC_NAVG:0 MACHSPEC_OFFS:3200
6	MACHSPEC_TRKI:0 MACHSPEC_NLNE:0
7	----------
8	Electrophoresis tracking:
9	Start RunStart Run
10	**Electrophoresis Started date:2000/8/4 time:16:57:11:0**
11	Stop Run
12	Electrophoresis Stopped date:2000/8/4 time:17:30:5:0
13	Start Collection
14	**Data Collection Started date:2000/8/4 time:17:2:48:0**
15	Stop Collection
16	Data Collection Stopped date:2000/8/4 time:17:30:5:0
17	…

9.3.3 _StatData.dat

The _StatData.dat file contains three critical items of information: analysis datetime, ISS name, and the "official" number of dyes used in the multiplex. Table 14 lists the file for one particular sample. The bits that Multiplex_QA expects are in **bold** text. All other contents of this file are ignored.

Multiplex_QA expects the strings "date:" and "time:" to be in row 1, with the calendar date between the strings and the time of day immediately following "time:". The strings "stdName:" and "dyeStdNum:" are expected to be in row 4 with the name of the ISS between the strings. Oddly enough, the value to the right of the "dyeStdNum:" does not reliably specify the number of dyes used (neither is the number of dye names recovered from the _DyeData.dat file); this number is found by counting the number of rows that contain the string "hasAnalData:".

Table 14, Format of the BatchExtract _StatData.dat Files

Row	Content
1	**hasAnalData**:1 wasSizeCalled:1 **date:8/4/2000 time:17:30:9:0**
2	fullRange:0 analStartPt:3200 analStopPt:7296 baselined:1 multicomped:1 normalized:0
3	smmothValue:2 minPeakWith:3 allSizes:1 startSize:75 stopSize:500
4	paramName: <Analysis Parameters> **stdName::<GS500> dyeStdNum:4**
5	pkThresholds: 150 150 150 300 50 50 50 50 50 50 50 50 50 50 50
6	sizeMethod: RecpNNone pkCorrMethod: None
7	corrLimit:30 numDyeStdPks:14 numDetStdPks:15 numMatchedPks:14
8	**hasAnalData**:1 wasSizeCalled:1 date:8/4/2000 time:17:30:9:0
9	fullRange:0 analStartPt:3200 analStopPt:7296 baselined:1 multicomped:1 normalized:0
10	smmothValue:2 minPeakWith:3 allSizes:1 startSize:75 stopSize:500
11	paramName: <Analysis Parameters> stdName::<GS500> dyeStdNum:4
12	pkThresholds: 150 150 150 300 50 50 50 50 50 50 50 50 50 50 50
13	sizeMethod: RecpNNone pkCorrMethod: None
14	corrLimit:30 numDyeStdPks:14 numDetStdPks:15 numMatchedPks:14
15	**hasAnalData**:1 wasSizeCalled:1 date:8/4/2000 time:17:30:9:0
16	fullRange:0 analStartPt:3200 analStopPt:7296 baselined:1 multicomped:1 normalized:0
17	smmothValue:2 minPeakWith:3 allSizes:1 startSize:75 stopSize:500
18	paramName: <Analysis Parameters> stdName::<GS500> dyeStdNum:4
19	pkThresholds: 150 150 150 300 50 50 50 50 50 50 50 50 50 50 50
20	sizeMethod: RecpNNone pkCorrMethod: None
21	corrLimit:30 numDyeStdPks:14 numDetStdPks:15 numMatchedPks:14
22	**hasAnalData**:1 wasSizeCalled:1 date:8/4/2000 time:17:30:9:0
23	…

9.3.4 _PeakData.dat

Multiplex_QA expects "dyeCnt" to be the first characters in row 1 of the _PeakData.dat file. Each following row specifies parameters for one "called" (i.e., recognized and interpreted by the original software analysis system) peak. The parameters are separated by a single spaces (" "). Table 15 lists parts of the file for one particular sample, with the mandatory bit of row 1 in **bold** text.

Table 15, Format of the BatchExtract _PeakData.dat Files

Content
dyeCnt index position height beginPos endPos beginHI endHI area volume fragSize isEdited label
0 0 198 276 192 212 4 0 956 1397707348 95.174599 0
0 1 528 2262 515 539 7 4 14318 1129336396 132.269470 0
…
1 0 279 1605 273 294 148 0 9569 1397707348 104.219337 0
…
2 0 1676 805 1660 1691 12 2 6839 1397707348 275.797058 0
…
3 0 24 2152 9 48 6 0 13464 1397707348 75.000000 0
3 1 240 2342 226 257 3 2 14412 1129336396 100.000000 0
3 2 585 2535 572 607 9 5 15692 67108998 139.000000 0
3 3 669 2591 655 688 6 0 15983 261504 150.000000 0
…

Multiplex_QA extracts the first eleven parameters for each recognized peak. However, only the seven parameters described in Table 16 are currently used in any way.

Table 16, Peak Parameters in _PeakData.dat Used by Multiplex_QA

#	Name	Description
1	dyeCnt	Dye index, from 0 to number-of-dyes-1. The ISS is always the last dye.
3	position	Scan # of peak maximum.
4	height	Height of peak at peak maximum, in rfu.
5	beginPos	Initial scan # of peak (identifies when peak begins).
6	endPos	Final scan # of peak (identifies when peak ends).
9	area	Peak area, scan # × rfu.
11	fragSize	Size of DNA fragment associated with the peak, in bp. These are nominal sizes for ISS peaks and interpolated sizes for the others.

9.3.5 _ScanData.dat

If you wish to visualize electropherograms, Multiplex_QA requires access to the _ScanData.dat files holding the background-corrected rfu data for the sample. These files list the scan index and the rfu values for each dye color used by the multiplex. The rfu data for the ISS dye is always listed last. The scan index and the rfu values are separated by tabs. Table 17 lists the initial few rows of the file for one particular sample.

Table 17, Format of the BatchExtract _ScanData.dat Files

Content				
0	0	0	17	0
1	1	5	8	0
2	4	8	4	0
3	6	6	3	2
4	4	4	3	4
5	1	2	2	5
6	1	1	1	6
7	1	3	0	7
8	2	3	0	6
9	6	1	0	6
10	10	0	0	6
11	12	0	1	10
12	11	0	2	18
13	10	0	2	30
14	9	0	2	47
15	11	0	3	67
16	13	0	4	86
17	13	0	5	107
18	12	0	6	144
19	10	1	7	234
…				

10 WHAT TO DO WHEN THINGS BREAK

Multiplex_QA is a modestly large and complex set of interlocking systems. Particularly when first looking at new datasets and when new eyes examine even old datasets, the exploratory nature of the system tends to become embarrassingly apparent. In addition to occasional outright fatal errors, there may be incorrect calculations and logical inconsistencies. There certainly remain a number of interconnections between the functions that are either clumsily implemented or missing entirely, as well as many good ideas that the author has yet to stumble across.

"We apologize for any inconvenience." That said, it seems desirable to reduce the level of inconvenience and maybe even enhance the level of utility. To do so, users (i.e., you) need to identify what the error, clumsiness, missing capability, etc. actually is… and then report it to the author at david.duewer@nist.gov.

10.1 Fatal errors

Fatal errors are those that keep you from doing what you want to do when the data are actually available and in the right form for it to be accomplished. Three classes of fatal errors may be encountered: anticipated situations protected by "Alerts," unexpected problems resulting in "Macro Errors," and *really* unexpected errors causing Excel and/or your operating system to die without warning.

If either an uninformative Alert or a Macro Error is encountered, print the screen and save it to a text file along with notes on the functions you'd used immediately before using the function that just savaged you. Then and only then, try to recreate the error. (There's not much chance of identifying the cause of an error that can't be repeated, but it's at least worth documenting). You will (probably, unless it's the unmodified Demo dataset) also need to provide the dataset. If the error was in the **Process Data** or **E'gram** functions, the BatchExtract text files that are the proximate cause of the problem will also be needed.

10.1.1 Alerts: Anticipated error conditions

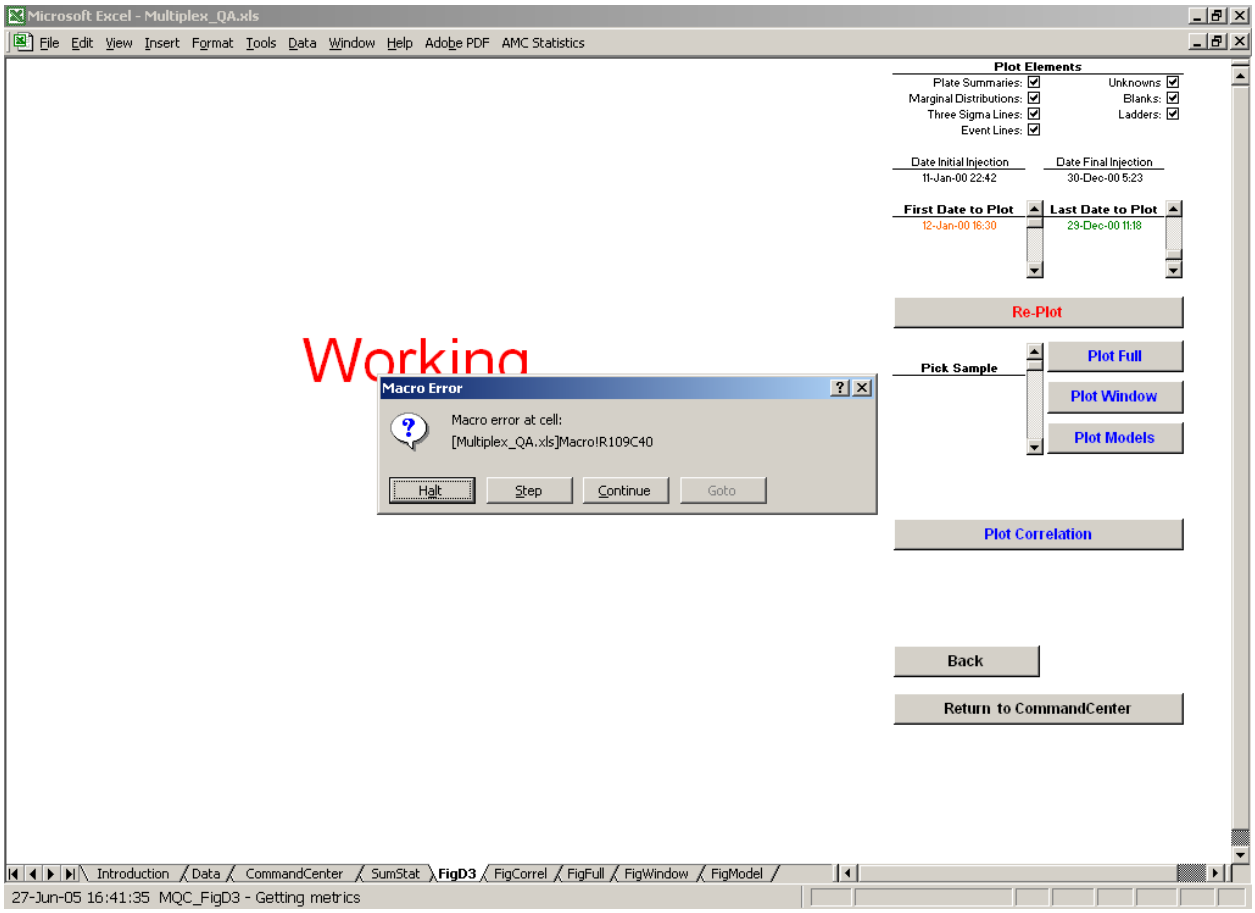
There are a large number of potential error conditions that Multiplex_QA monitors, most occur when a prerequisite task (e.g., loading a dataset or making the first **Full E'gram** plot or **D^3 chart**) hasn't been performed. When such an error occurs, an Alert window is generated that briefly describes the problem and suggests a remedy. These are *not* "fatal errors"; once you figure out how to do what Multiplex_QA wants, you'll probably be able to do what you want.

However, there are a number of "these should never happen" conditions that are nevertheless monitored. If such an error is encountered, an Alert window is generated that tersely identifies the problem and then states "Duewer's problem." Multiplex_QA will then halt no matter how you clear the Alert box.

10.1.2 Macro Errors: Unexpected error conditions

If the Multiplex_QA system encounters a fatal programmatic error, the Excel 95 macro language interpreter generates a Macro Error window (see Figure 61) that specifies where the error occurred and presents you with a series of options on how to proceed: Halt, Step, Continue, Goto, and close-the-window. You should "Halt" and document the error. The other options are debugging aids, but not viable for error recovery.

Figure 61. Macro Error



10.1.3 System collapse

I haven't encountered any truly disastrous errors that seem unique to Multiplex_QA, other than having two sessions open at the same time fighting each other for the same file. (I consider the wounds from that battle to have been pretty much self-inflicted). But it is *always* good practice to avoid having too many windows open at the same time, and to save all critical stuff early and often.

At any rate, if you do have the misfortune to discover a really nasty bug, record the circumstances as best you can remember and try to repeat the problem. If it's repeatable, there's a good chance Multiplex_QA can be fixed to at least avoid the hassle. If you can't get it to happen on demand, it was probably a one-off and there's not much to be done… beyond swearing a bit and re-booting.

10.2 Bad calculations or logical inconsistencies.

Make a hardcopy of the report and/or plot that's in error, along with some documentation of what the value should be.

10.3 Clumsy or missing connection between existing functions.

Describe briefly what's irritating and how you'd like to have it work.

10.4 Missing capability

Describe what you'd like to see analyzed or displayed in sufficient detail that one of your colleagues understands the description and honestly says something like "That'd be cool!"

11 BAGATELLES

The following details are provided for those obsessive folk who are *really* interested in how Multiplex_QA does things.

11.1 Worksheets

Table 18 lists all of the worksheets used by the Multiplex_QA system. You may add other worksheets as you see fit, but renaming any of the following will cause problems.

Table 18, Multiplex_QA Worksheets

Sheet	Visible?	Purpose
CommandCenter	Always	Provides access to most Multiplex_QA commands
Data	As needed	Holds the QA data for all processed files
Directory	As needed	Specifies the directories containing files to process
ErrorLog	Never	When & where info about errors and commands
FigCorrel	Always	Display correlation among 2 to 5 QA metrics
FigFull	Always	Display complete electropherogram of specified file
FigWindow	Always	Display a section of electropherogram of specified file
FigModel	Always	Display height/area and retention time models
FigD3	Always	Display 1 to 5 QA metrics as functions of date
Introduction	Always	Introductory comments
Language	Transient	Contains Excel's "Country Version" dependencies
Macro	Never	Contains the program that does all this stuff
PDetails	Never	Contains orthogonal-variable calculations and display formats
PEvents	Transient	Contains user-defined dates of events noted in FigD3 display
PFlags	Never	Details of error and file type codes used in file processing
PLadSize	Never	Details of ISS (GS350+, GS400, ILS)
PLadType	Never	Used in guessing multiplex kit type
RDetails	As needed	Details "quality" issues by sample
RDups	As needed	Lists any duplicate files found during processing
RPlate	As needed	Summarizes performance by sample location (96-well plate)
RQuality	As needed	Summarizes "quality" issues by error code
RUtility	As needed	Summarizes "quality" issues by utility class
TBatch	Never	List of samples probably amplified on same plate
TEgram	Never	Holds data displayed in FigFull, FigWindow, & FigModel
TFile	Never	Used when evaluating height/area and retention time models
TSet	Never	List of samples injected in same run
SumStat	Transient	Holds statistical summary info used in FigD3 & FigCorrel
TTime	Never	Holds the data displayed in FigD3 & FigCorrel
Winnow	Transient	Used when deleting subsets of data

11.1.1 Visibility Status of Multiplex_QA Worksheets

To focus your attention on the important information, a number of the Multiplex_QA worksheets are normally hidden from your view. The status of each Multiplex_QA worksheet is

specified in the Table 18 column under "Visible?" Those worksheets that are "Always" visible are always there and accessible, even when there are no active data being displayed. Those that are "Never" visible are either used for temporary calculations or hold data that are of no particular use except to a hard-core programmer. Those that are "Transient" are visible only when they are needed for a particular function. Those that are visible "As needed" are created when they are needed when specifying or processing .fsa files; they are deleted from Multiplex_QA when the data are exported (Section 2.4.4).

11.1.2 Revealing the invisible

All of the worksheets in the Multiplex_QA system can be accessed, if you really want to, using Excel's Format>Sheet>Unhide command.

11.2 Orthogonal Variable Regressions

Table 19, Orthogonal Variables

	Retention Model				Height/Area Model			
ISS	bp	bp^2	F_1	F_2	$\log_{10}$	$1/\log_{10}$	G_1	G_2
GS350+,GS400	150	22500	-1.52	-1.89	2.18	0.460	2.05	-1.89
GS350+,GS400,ILS	160	25600	-1.39	-1.33	2.20	0.454	1.13	-1.63
GS400, ILS	180	32400	-1.13	-0.39	2.26	0.443	-0.12	-1.16
GS400	190	36100	-0.99	0.00	2.30	0.435	-0.52	-0.95
GS350+,GS400,ILS	200	40000	-0.85	0.32	2.30	0.435	-0.80	-0.75
GS400	220	48400	-0.57	0.81	2.30	0.435	-1.09	-0.39
ILS	225	50625	-0.49	0.89	2.35	0.425	-1.12	-0.30
GS400	240	57600	-0.27	1.06	2.30	0.435	-1.09	-0.06
GS400	260	67600	0.05	1.08	2.30	0.435	-0.88	0.24
ILS	275	75625	0.29	0.95	2.44	0.410	-0.62	0.44
GS400	280	78400	0.37	0.87	2.30	0.435	-0.52	0.51
GS400	290	84100	0.54	0.68	2.30	0.435	-0.29	0.64
GS350+,GS400,ILS	300	90000	0.71	0.43	2.48	0.404	-0.04	0.76
GS400	320	102400	1.06	-0.23	2.30	0.435	0.54	0.99
ILS	325	105625	1.15	-0.44	2.51	0.398	0.69	1.04
GS350+,GS400	340	115600	1.43	-1.13	2.53	0.395	1.17	1.20
GS350+, ILS	350	122500	1.61	-1.67	2.54	0.393	1.51	1.30
Average	253	67946	0.00	0.00	2.40	0.419	0.00	0.00
Standard Deviation	64	32090	1.00	1.00	0.12	0.021	1.00	1.00
Correlation	0.994		0.000		-0.999		0.000	

To make the height/area and retention model coefficients more meaningful, the coefficients stored in the Data worksheet are from regression on orthogonal variables related to the nominal model variables. Table 19 lists the ISS peaks used in the regressions, the values for the nominal variables (bp and bp^2 for the retention model and $\log_{10}(bp)$ and $1/\log_{10}(bp)$ for height/area), and their respective orthogonal representations (F_1 and F_2 for the retention model and G_1 and G_2 for height/area. The means and standard deviations for all variables are also listed, as well as the correlation between each of the pairs.

The orthogonal variables are *linear rotations* of the nominal variables. This means that the predictive utility of the models isn't changed by using the orthogonalized version of the nominal variables. Table 20 lists the rotation coefficients that relate the orthogonal variables back to the nominal variables, using the equations

$$F_i = \Psi_0(F_i) + \Psi_1(F_i)bp + \Psi_2(F_i)bp^2$$
$$G_i = \Psi_0(G_i) + \Psi_1(G_i)\log_{10}bp + \Psi_2(G_i)/\log_{10}bp$$

where the index i is either 1 or 2 and the notation "$\Psi_0(F_1)$" translates to "the Ψ_0 term for the F_1 orthogonal variable."

Table 20, Rotation Coefficients

	Retention Model		Height/Area Model	
Term	F_1	F_2	G_1	G_2
Ψ_0	-3.051	-17.19	-950.4	-0.2202
Ψ_1	0.007862	0.1452	201.0	4.329
Ψ_2	0.00001560	-0.0002881	1121	-24.14

The following equations are the retention and height/area models for use with the nominal variables:

$$TI = (\alpha + \beta\Psi_0(F_1) + \gamma\Psi_0(F_2)) + (\beta\Psi_1(F_1) + \gamma\Psi_1(F_2))bp + (\beta\Psi_2(F_1) + \gamma\Psi_2(F_2))bp^2$$
$$H/A = (a + b\Psi_0(G_1) + c\Psi_0(G_2)) + (b\Psi_1(G_1) + c\Psi_1(G_2))\log_{10}bp + (b\Psi_2(G_1) + c\Psi_2(G_2))/\log_{10}bp$$

A note of explication: While variable orthogonalization does indeed serve to equalize and stabilize the magnitudes of the various the model coefficients, it sure makes using the coefficients less convenient! My enthusiasm for "robust" numerical methods was (probably) here misguided.

11.3 Where the Code Lives

The Multiplex_QA system is written almost entirely in the ancient Excel XLM macro language. No real reason other than I'm an ancient programmer. All of this code is located on the Macro worksheet (to access this, see Section 11.1.2). There are a number of default parameter settings in column 1 of this sheet that a *really* advanced user may plausibly want to modify.

The only bits of modern code are the VBA macros "Auto_Open" and "Auto_Close" that run when the Multiplex_QA system is opened or closed. They can be accessed through Excel's Tools>Macro>Macros>Edit facility.

Nothing in Multiplex_QA is protected – so if you go exploring, make sure that you have a backup copy!

12 ACKNOWLEDGMENTS

I thank Margaret C. Kline, Janette W. Redman, Peter M. Vallone, Michael Coble, John M. Butler, and Mary Satterfield of the Biochemical Science Division, Chemical Science and Technology laboratory, National Institute of Standards and Technology, for their support, advice, criticism, data, and – above all! – time and patience during the development and debugging of Multiplex_QA.

I thank Janos Murvai, Jon Baker, and Stephen Sherry of the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, for developing BatchExtract and making it publicly accessible. They have also been wonderfully responsive to questions and suggestions concerning its use and operation.

I thank Tim McMahon of the Armed Forces DNA Identification Laboratory for providing "before and after" data for several of their instruments. I thank Helle Smidt Mogensen of the Department of Forensic Genetics, Institute of Forensic Medicine at the University of Copenhagen, for "test driving" the BatchExtract/Multiplex_QA system with fresh eyes and enthusiasm. She managed to maintain good humor and positive suggestions through a variety of download, new ISS, non-English Excel version hassles, and just plain bugs. I thank Martin Mikkelsen, also at the University of Copenhagen's Department of Forensic Genetics, for his patient help in tracking down and finally defeating a subtle bug involved in reading BatchExtract files using Excel versions that don't recognize "." as a decimal point.

This project was funded in part by the National Institute of Justice through interagency agreement 2003-IJ-R-029 with the NIST Office of Law Enforcement Standards, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice.

13 REFERENCES

1 Stuart A, Ord JK, eds. Kendall's Advanced Theory of Statistics, 5th Edition, Volume 1. Oxford University Press, NY **1987**. Chapter 10, Section 11.

2 Rosenblum BB, Oaks F, Menchen S, Johnson B. Improved single-strand DNA sizing accuracy in capillary electrophoresis. Nucleic Acids Res 1997;25(19):3925-3929.

3 IUPAC, Compendium of analytical nomenclature, definitive rules 1997. ("Orange Book"), 3rd Edition. Blackwell, Oxford (1998). Section 9.2.3.10 of the Web edition: http://www.iupac.org/publications/analytical_compendium/

4 Duewer DL, Liu H-K, Reeder DJ. Graphical tools for RFLP DNA profiling. Single-locus Charts. J Forensic Sci 1999;44(5):969-977.

5 Duewer DL, Kline MC, Redman JW, Newall PJ, Reeder DJ. NIST Mixed Stain Studies #1 and #2: Interlaboratory Comparison of DNA Quantification Practice and Short Tandem Repeat Multiplex Performance with Multiple-Source Samples. J Forensic Sci 2001;46(5):1199-1210.

6 Dan Gookin. DOS for Dummies, 3rd Edition. Hungry Minds, Inc. 909 Third Avenue, New York, NY 10022 (1999). http://www.dummies.com