APPLICATION INFORMATION

Genetic Analysis: CEQ Series

STRATEGIES FOR AUTOMATING THE REVIEW OF DATA

Mark Dobbs
Beckman Coulter, Inc.

One of the most time-consuming tasks in projects that call upon large numbers of independent data sets is the segregation of data that is irrelevant or of poor quality from data that is useful. The task of selecting data that is relevant to a specific task in a study, but may be of marginal or no value in other tasks, is more difficult. For example, short electrophoretic separations with size fragments that are less than or equal to 250 nucleotides in length would be useful to discriminate DNA fingerprints where length polymorphisms occur in the vicinity of 150 nucleotides but not where the polymorphisms occur at 300 nucleotides. In most data management environments, the results would have to be screened visually to eliminate those that were run for a separation time that was too short. The CEQ™ software has a sophisticated set of tools for managing data that will be used for downstream fragment analysis.

The CEQ sorting and filtering tools operate on a set of over 70 independent properties of independent results. From the set of results that have been selected for further analysis, each recognized fragment has over 30 individual properties that can be used to discriminate it from other fragments. In this bulletin, we provide a variety of examples of how the sorting and filtering tools can be used to quickly reduce very large collections of DNA fragments to sets that are easily managed and suited to specific secondary analyses, such as allele binning and linkage analysis, peak ratio analysis, AFLP analysis, and LOH analysis. Using specifically customized **Filter sets,** a user can quickly and automatically remove from consideration all results or fragments that are not pertinent to the current experiment.

Definition of a "Study"

Within the context of the CEQ, a **Study** is defined as a collection of analyzed, non-sequencing results. This level of analysis, referred to as primary analysis to distinguish it from higher order levels of interpretation, consists of color separation, baseline normalization, dye mobility-shift correction and, in nearly all cases, fragment size determination and allele assignment. Each study has two unique parts: a single Results Set List (Figure 1) and a single Fragment List, which is the sum of all of the fragments from every result in the Results Set. All of the operations within the CEQ Fragment Analysis Software occur in the context of the single study that is open at any one time.

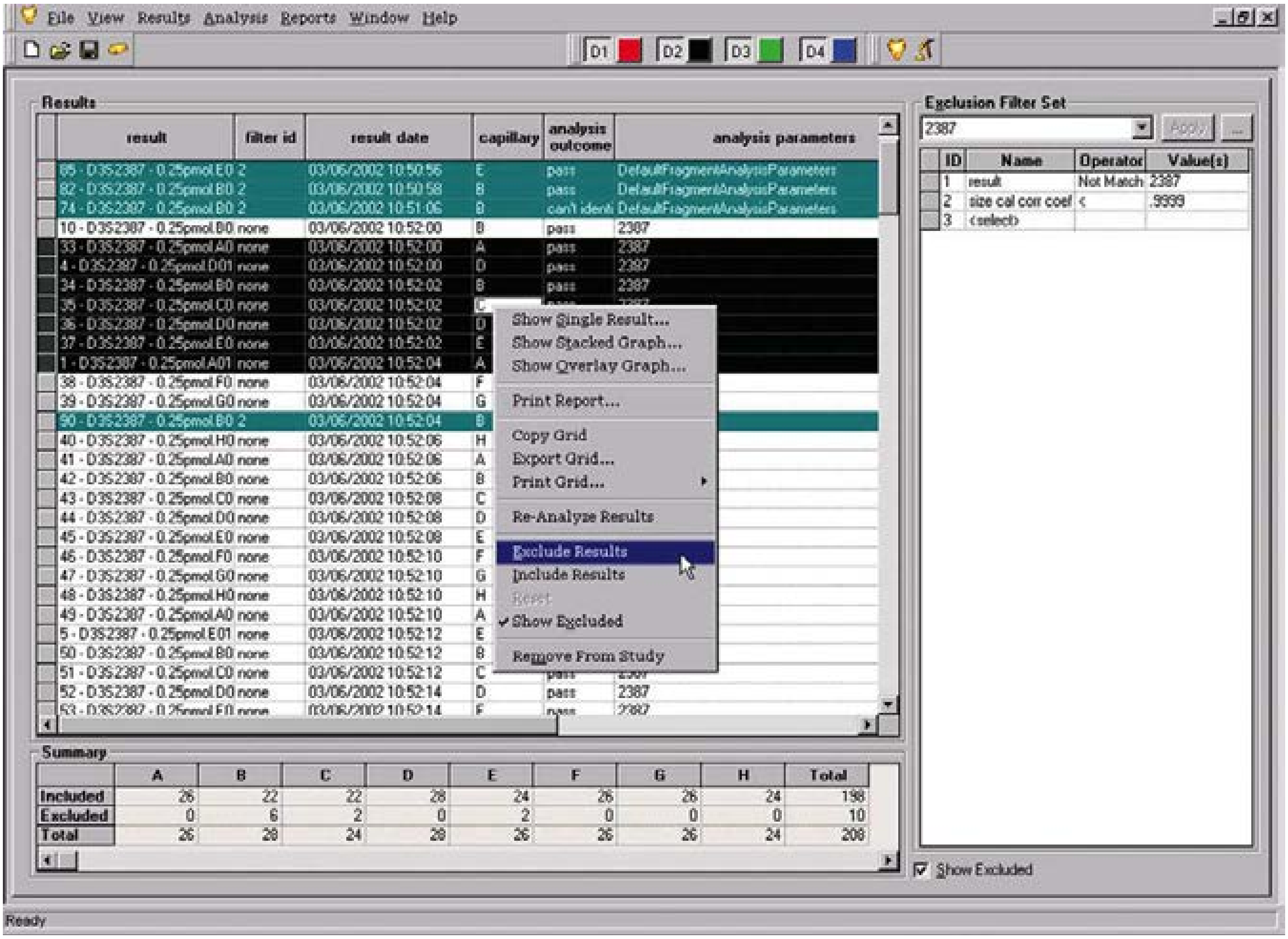
Column Selection

Each of the two lists or grids, the Results Set and the Fragment List, has a large number of data fields associated with it. Each data field can be selected for viewing using the Column Selector (Figures 2 and 3). The column selector has three functions:

1) it allows the selection of which columns to view
2) it determines the display order of these columns
3) it determines which, if any columns should be locked in place while the remaining columns scroll horizontally.

Table 1. Results Set Column Fields (Available Filters)

Abbreviated Description	***Full Description***
% peaks asymm	Percentage of size calibration peaks with asymmetry > 0.2
analysis outcome	Analysis Status
analysis param date	Date Parameters Modified
analysis parameters	Analysis Parameter Set
avg current (microA)	Average Separation Current
capillary	Capillary
channel 1 baseline	Average Baseline Level Ch1
channel 1 overrange	Overrange signal Channel 1
channel 1 rms noise	Raw Data RMS Noise Ch1
channel 2 baseline	Average Baseline Level Ch2
channel 2 overrange	Overrange signal Channel 2
channel 2 rms noise	Raw Data RMS Noise Ch2
channel 3 baseline	Average Baseline Level Ch3
channel 3 overrange	Overrange signal Channel 3
channel 3 rms noise	Raw Data RMS Noise Ch3
channel 4 baseline	Average Baseline Level Ch4
channel 4 overrange	Overrange signal Channel 4
channel 4 rms noise	Raw Data RMS Noise Ch4
current chng (%/min)	Relative Average Separation Current Slope
current noise (%)	Relative Standard Deviation of Separation Current
D1 fragments unsized	Fragments not sized D1
D1 broad peaks	Broad peaks >1 D1
D1 fragments <5% out	Fragments outside calibration range by < than 5% - D1
D1 fragments >5% out	Fragments outside calibration range by > than 5% - D1
D1 quant std missed	Expected Quantity standard not found D1
D1 spiky peaks	Spiky peaks >1 Ch1
D2 fragments unsized	Fragments not sized D2
D2 broad peaks	Broad peaks >1 D2
D2 fragments <5% out	Fragments outside calibration range by < than 5% - D2
D2 fragments >5% out	Fragments outside calibration range by > than 5% - D2
D2 quant std missed	Expected Quantity standard not found D2
D2 spiky peaks	Spiky peaks >1 D2
D3 fragments unsized	Fragments not sized D3
D3 broad peaks	Broad peaks >1 D3
D3 fragments <5% out	Fragments outside calibration range by < than 5% - D3
D3 fragments >5% out	Fragments outside calibration range by > than 5% - D3
D3 quant std missed	Expected Quantity standard not found D3
D3 spiky peaks	Spiky peaks >1 D3
D4 fragments unsized	Fragments not sized D4
D4 broad peaks	Broad peaks >1 D4
D4 fragments <5% out	Fragments outside calibration range by < than 5% - D4D4 quant std missed Expected Quantity standard not found D4
D4 spiky peaks	Spiky peaks >1 D4
denature time	Denature Duration
filter ID	Filter ID
low D1 SNR	>50% of peaks in D1 are <10 x rms baseline noise
low D2 SNR	>50% of peaks in D2 are <10 x rms baseline noise
low D3 SNR	>50% of peaks in D3 are <10 x rms baseline noise
low D4 SNR	>50% of peaks in D4 are <10 x rms baseline noise
minimum peak height	Relative Peak Height Threshold (Include Peaks)
no. cal stds missed	Number of calibration size standards not found
no. dyes in sample	Number of Dyes
no. siz stds missed	Number of size standards not found
number pks D1	Number of Peaks in D1
number pks D2	Number of Peaks in D2
number pks D3	Number of Peaks in D3
number pks D4	Number of Peaks in D4
peak slope threshold	Slope Threshold
result	Result Name
result date	Date Result Modified
sample	Sample Name
sample date	Date Sample Collected
separation method	Separation Method
separation temp	Capillary Temperature
separation time	Total Separation Duration
separation voltage	Separation Voltage
size cal corr coef	Size Calibration Correlation Coefficient
size cal fit std dev (nt)	Size Calibration Standard Deviation
size standard	Size Standard Name
used data spectra	Dye spectra estimated from data
used system spectra	Used system dye spectra
user properties	User Properties



Results

result	filter id	result date	capillary	analysis outcome	analysis parameters
65 - D3S2387 - 0.25pmol E0	2	03/06/2002 10:50:56	E	pass	DefaultFragmentAnalysisParameters
82 - D3S2387 - 0.25pmol B0	2	03/06/2002 10:50:58	B	pass	DefaultFragmentAnalysisParameters
74 - D3S2387 - 0.25pmol B0	2	03/06/2002 10:51:06	B	can't identi	DefaultFragmentAnalysisParameters
10 - D3S2387 - 0.25pmol B0	none	03/06/2002 10:52:00	B	pass	2387
33 - D3S2387 - 0.25pmol A0	none	03/06/2002 10:52:00	A	pass	2387
4 - D3S2387 - 0.25pmol D01	none	03/06/2002 10:52:00	D	pass	2387
34 - D3S2387 - 0.25pmol B0	none	03/06/2002 10:52:02	B	pass	2387
35 - D3S2387 - 0.25pmol C0	none	03/06/2002 10:52:02	C	pass	2387
36 - D3S2387 - 0.25pmol D0	none	03/06/2002 10:52:02	D		
37 - D3S2387 - 0.25pmol E0	none	03/06/2002 10:52:02	E		
1 - D3S2387 - 0.25pmol A01	none	03/06/2002 10:52:04	A		
38 - D3S2387 - 0.25pmol F0	none	03/06/2002 10:52:04	F		
39 - D3S2387 - 0.25pmol G0	none	03/06/2002 10:52:04	G		
90 - D3S2387 - 0.25pmol B0	2	03/06/2002 10:52:04	B		
40 - D3S2387 - 0.25pmol H0	none	03/06/2002 10:52:06	H		
41 - D3S2387 - 0.25pmol A0	none	03/06/2002 10:52:06	A		
42 - D3S2387 - 0.25pmol B0	none	03/06/2002 10:52:06	B		
43 - D3S2387 - 0.25pmol C0	none	03/06/2002 10:52:08	C		
44 - D3S2387 - 0.25pmol D0	none	03/06/2002 10:52:08	D		
45 - D3S2387 - 0.25pmol E0	none	03/06/2002 10:52:08	E		
46 - D3S2387 - 0.25pmol F0	none	03/06/2002 10:52:10	F		
47 - D3S2387 - 0.25pmol G0	none	03/06/2002 10:52:10	G		
48 - D3S2387 - 0.25pmol H0	none	03/06/2002 10:52:10	H		
49 - D3S2387 - 0.25pmol A0	none	03/06/2002 10:52:10	A		
5 - D3S2387 - 0.25pmol E01	none	03/06/2002 10:52:12	E		
50 - D3S2387 - 0.25pmol B0	none	03/06/2002 10:52:12	B		
51 - D3S2387 - 0.25pmol C0	none	03/06/2002 10:52:12	C	pass	2387
52 - D3S2387 - 0.25pmol D0	none	03/06/2002 10:52:14	D	pass	2387
53 - D3S2387 - 0.25pmol F0	none	03/06/2002 10:52:14	F	pass	2387

Summary

	A	B	C	D	E	F	G	H	Total
Included	26	22	22	28	24	26	26	24	198
Excluded	0	6	2	0	2	0	0	0	10
Total	26	28	24	28	26	26	26	24	208

Exclusion Filter Set

2387

ID	Name	Operator	Value(s)
1	result	Not Match	2387
2	size cal corr coef	<	.9999
3	<select>		

☑ Show Excluded

Figure 1. *Selected results can be launched using any of the top three right mouse button menu options.*

Beyond the viewing of data, the CEQ™ software allows direct copying, exporting, and printing of the grid as it is specified by the Column Selector. This allows a great deal of flexibility in preparing data for custom documents or for analysis by other software packages.

Sorting

For every column that is in the grid, the data can be sorted in ascending or descending order with a single left click of the mouse on the column header. The data in each row is linked together so that the associated data will follow the data of the primary sort. If secondary or tertiary sorts are desired, they may be selected using the **Sort…** right click menu option, which is accessible when the cursor is over the grid column headers.

Data may be sorted simply to get an idea of the range of values of a property of interest. If necessary, rows of data may be selected by manual highlighting, followed by exclusion using the right mouse click menu. When the **Show Excluded** checkbox is activated, manually excluded rows are colored purple. The individual traces corresponding to rows of either grid may be launched by highlighting the rows and using options from the right mouse button menus (one of the top 3 options in the right mouse button menu is shown in Figure 1).

Filtering the Results Set

The results set can be filtered upon any one of the Result Properties in Table 1. For each property, a pre-defined set of operators is available. The operators are of two types:

- Logical Operators
  =, Not Equal, >, >=, <, <=, Between
- Character or String-Based Operators
  =, Not Equal, Match, Not Match

Table 2. Fragment List Column Fields

Abbreviated Description	*Full Description*
abs frag amount	Absolute Fragment Quantity
allele ID	Allele ID
allele match qual	Allele ID Confidence
clstr area (rfu x mm)	Cluster Area
comment	User Comment
dye	Dye color
est frag size (nt)	Estimated Fragment Length
filter ID	Filter ID
frag size not est	No Fragment Size Estimated
locus name	Locus
mig time (min)	Average Migration Time
minus A peak	Minus A
mobility (cm^2/Vs)	Mobility
no alleles found	no alleles found
num pks in clstr	Cluster Size
pk area (rfu x mm)	Peak Area
pk assym	Peak Asymmetry
pk clstr ht order	Peak Cluster Height Rank
pk eff	Efficiency
pk height (rfu)	Peak Height
pk width (mm)	Peak Width
plus A pk	Plus A
quant std	Quantitation Standard
rel frag amount	Relative Fragment Quantity
res/base (1/nt)	Specific Resolution
result edited	Edited
RN	Result Name
sep method date/time	Date Method Modified
separation method	Separation Method
size highlim (nt)	Confidence Interval Upper
size limits not est	No confidence interval
size lowlim (nt)	Confidence Interval Lower
size std	Size Standard
spurious peak	Spurious Peak
std frag size (nt)	Standard Size
stutter peak	Stutter Peak
too many alleles	Too many alleles
unknown allele	Unknown allele

Below is a short list of possible criteria a user may want to use to organize and filter results:

- All results collected on a certain date
- All results analyzed on a certain date
- All results separated with a certain separation method
- All results analyzed with a certain analysis parameter set
- All results with only 2 of 4 possible dyes present
- All results with too much signal in dye 4.
- All results with ideal separation current values

The properties are accessed by clicking the <select> cell in the Filter set area (Figure 4). After selecting the correct operator and value, the software will *exclude* results that fit the filter criteria. Filter-excluded results, when displayed, are colored teal (Table 3). Each filter element in the collective filter set is executed sequentially, so that Filter sets can be tailored to the exact needs of the experiment. The same elements can be repeated with different operators and values to include rather than exclude results with values within a specified range. For example, to include all samples that were collected in February 2002, the following exclusion filter elements would be used:

sample date < 2/1/02
sample date > 2/28/02

Each element or property of the filter set is assigned a number, called the filter ID, which is a dynamic part of the filter set. The filter ID number is used to view how each result was excluded. The filter ID property is a selectable column of the result set, and can be sorted. The filter ID column is especially useful to tune filters so that valuable data is not excluded. The summary of included and excluded results is tabulated by capillary at the bottom of the Result Set Grid (Figures 1 and 4).

Table 3. Shading Color Codes for Data that Has Been Included or Excluded

Included samples have black text with light shading (none or gray) while excluded fragments have white text with dark shading (teal or purple).

State	*Shading Color*	*Options Available*
Normal	None	Filter exclusion, manual exclusion
Excluded By Filter	Teal	Remove filter, manual inclusion
Excluded by Manual Selection	Purple	Reset
Included by Manual Selection (Filter Override)	Gray	Reset

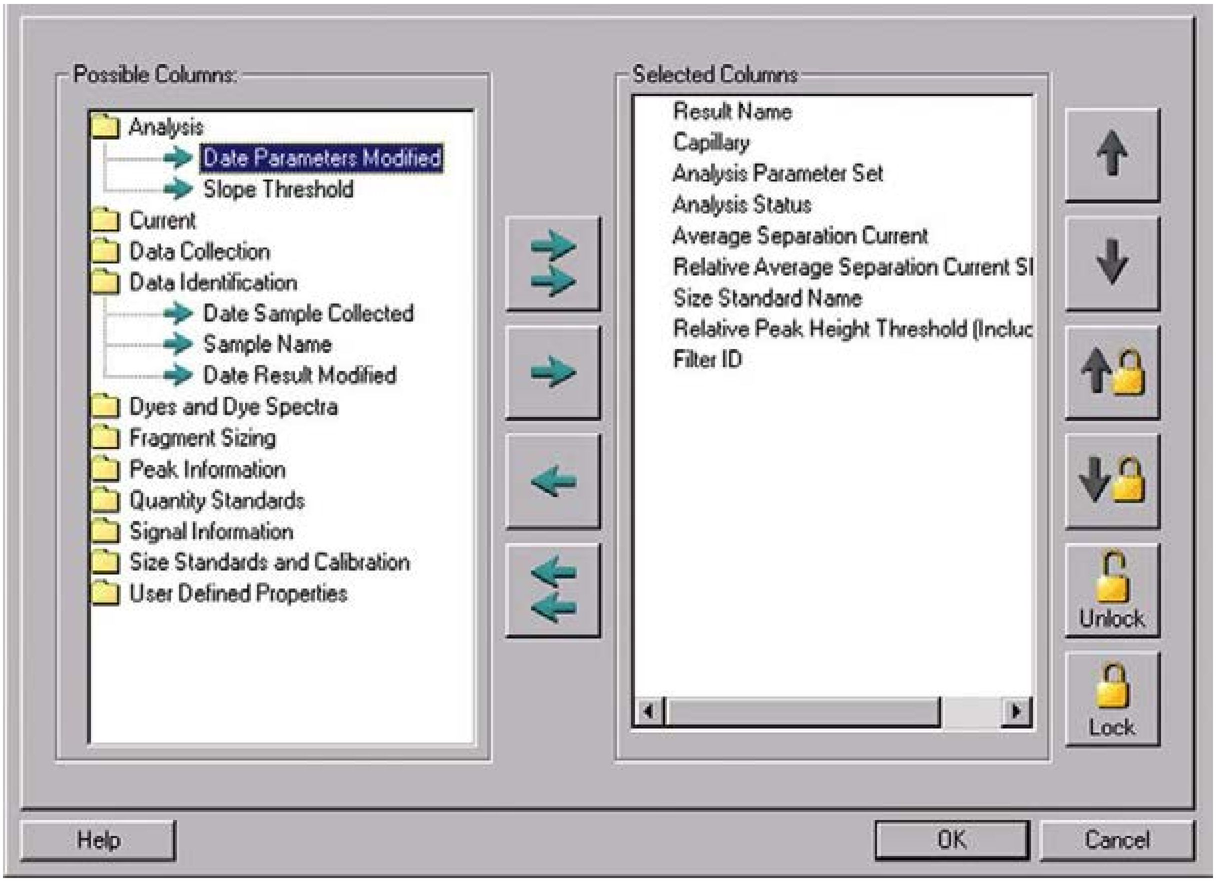


Figure 2. *Column Selector for the Results Set. Unselected properties are categorized in folders on the left.*

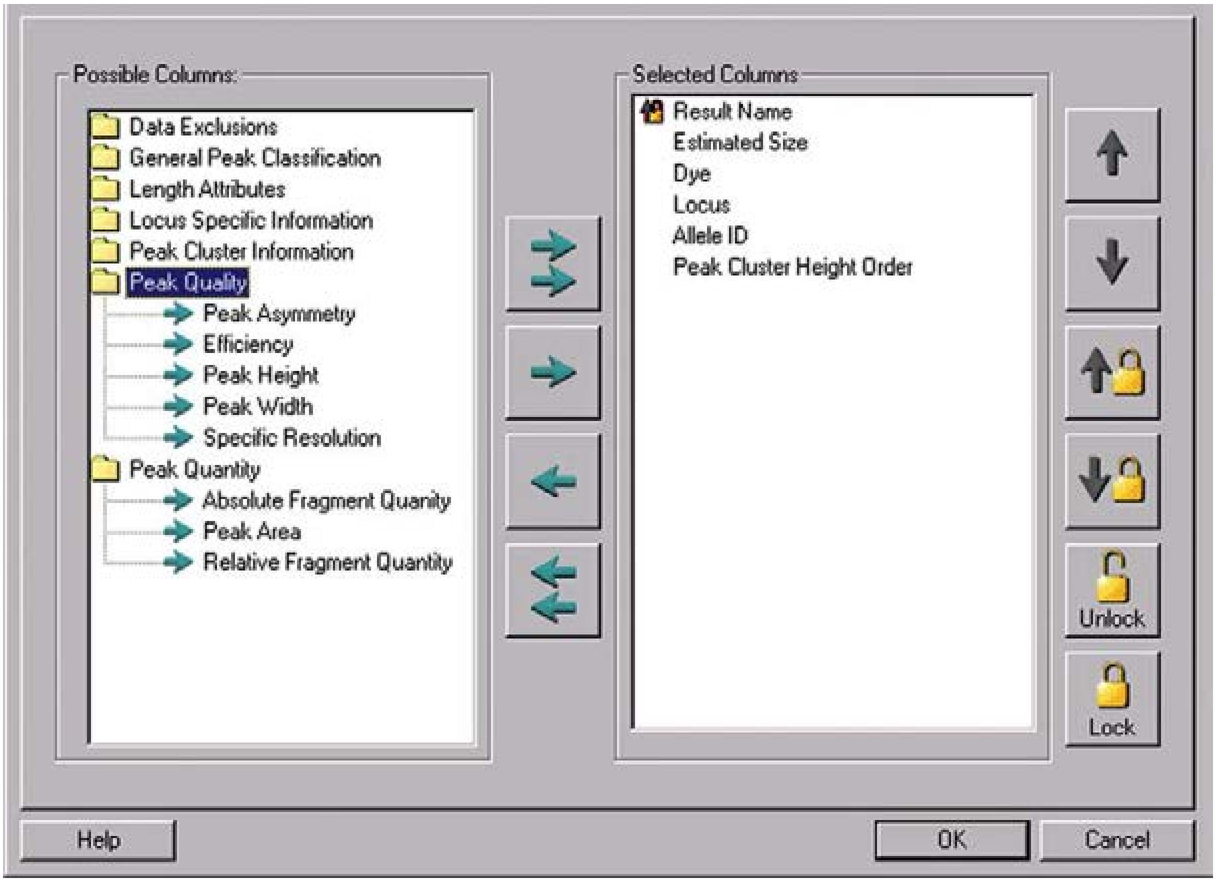
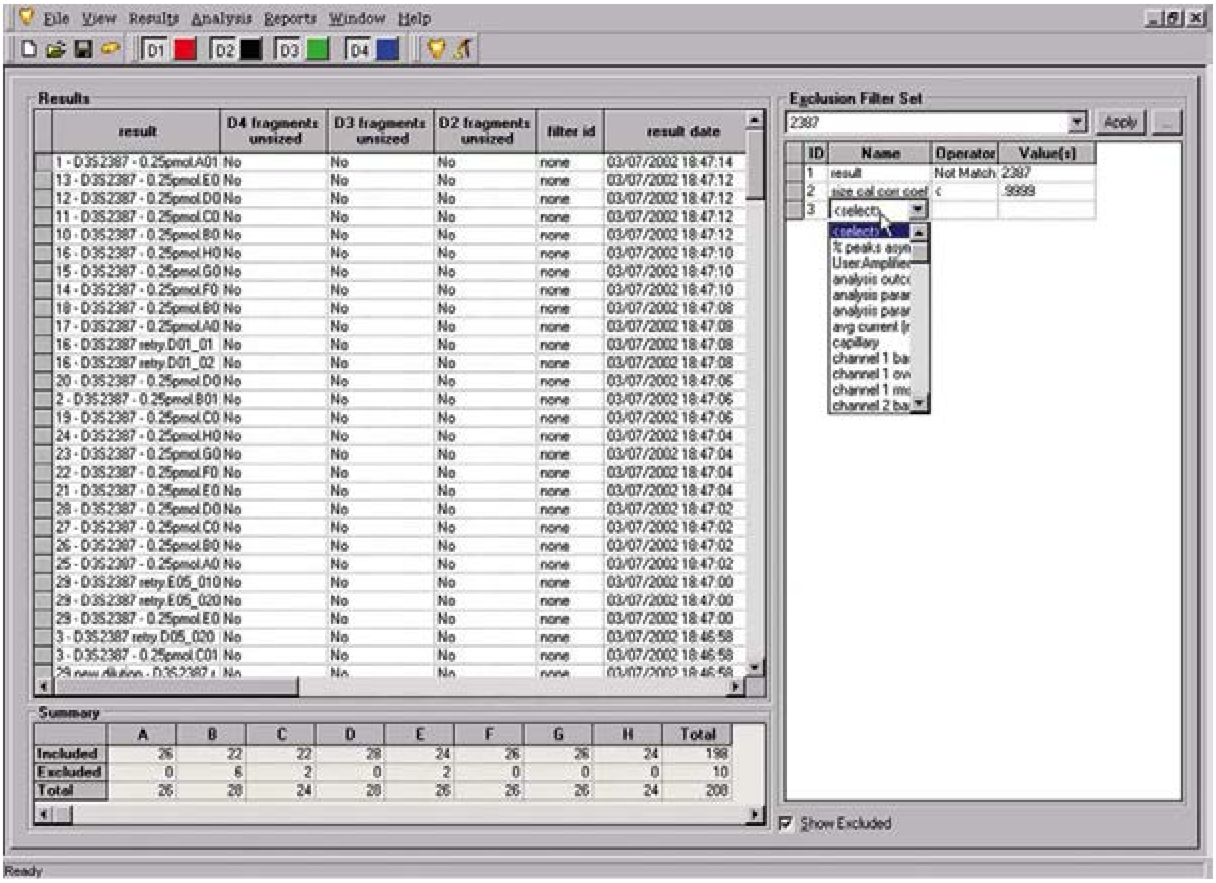


Figure 3. *Column Selector for the Fragment List. Unselected properties are categorized in folders on the left.*



Results

	result	D4 fragments unsized	D3 fragments unsized	D2 fragments unsized	filter id	result date
	1 - D3S2387 - 0.25pmol.A01	No	No	No	none	03/07/2002 18:47:14
	13 - D3S2387 - 0.25pmol.E0	No	No	No	none	03/07/2002 18:47:12
	12 - D3S2387 - 0.25pmol.D0	No	No	No	none	03/07/2002 18:47:12
	11 - D3S2387 - 0.25pmol.C0	No	No	No	none	03/07/2002 18:47:12
	10 - D3S2387 - 0.25pmol.B0	No	No	No	none	03/07/2002 18:47:12
	16 - D3S2387 - 0.25pmol.H0	No	No	No	none	03/07/2002 18:47:10
	15 - D3S2387 - 0.25pmol.G0	No	No	No	none	03/07/2002 18:47:10
	14 - D3S2387 - 0.25pmol.F0	No	No	No	none	03/07/2002 18:47:10
	18 - D3S2387 - 0.25pmol.B0	No	No	No	none	03/07/2002 18:47:08
	17 - D3S2387 - 0.25pmol.A0	No	No	No	none	03/07/2002 18:47:08
	16 - D3S2387 retry.D01_01	No	No	No	none	03/07/2002 18:47:08
	16 - D3S2387 retry.D01_02	No	No	No	none	03/07/2002 18:47:08
	20 - D3S2387 - 0.25pmol.D0	No	No	No	none	03/07/2002 18:47:06
	2 - D3S2387 - 0.25pmol.B01	No	No	No	none	03/07/2002 18:47:06
	19 - D3S2387 - 0.25pmol.C0	No	No	No	none	03/07/2002 18:47:06
	24 - D3S2387 - 0.25pmol.H0	No	No	No	none	03/07/2002 18:47:04
	23 - D3S2387 - 0.25pmol.G0	No	No	No	none	03/07/2002 18:47:04
	22 - D3S2387 - 0.25pmol.F0	No	No	No	none	03/07/2002 18:47:04
	21 - D3S2387 - 0.25pmol.E0	No	No	No	none	03/07/2002 18:47:04
	28 - D3S2387 - 0.25pmol.D0	No	No	No	none	03/07/2002 18:47:02
	27 - D3S2387 - 0.25pmol.C0	No	No	No	none	03/07/2002 18:47:02
	26 - D3S2387 - 0.25pmol.B0	No	No	No	none	03/07/2002 18:47:02
	25 - D3S2387 - 0.25pmol.A0	No	No	No	none	03/07/2002 18:47:02
	29 - D3S2387 retry.E05_010	No	No	No	none	03/07/2002 18:47:00
	29 - D3S2387 retry.E05_020	No	No	No	none	03/07/2002 18:47:00
	29 - D3S2387 - 0.25pmol.E0	No	No	No	none	03/07/2002 18:47:00
	3 - D3S2387 retry.D05_020	No	No	No	none	03/07/2002 18:46:58
	3 - D3S2387 - 0.25pmol.C01	No	No	No	none	03/07/2002 18:46:58
	29 new dilution - D3S2387 -	No	No	No	none	03/07/2002 18:46:58

Summary

	A	B	C	D	E	F	G	H	Total
Included	26	22	22	28	24	26	26	24	198
Excluded	0	6	2	0	2	0	0	0	10
Total	26	28	24	28	26	26	26	24	208

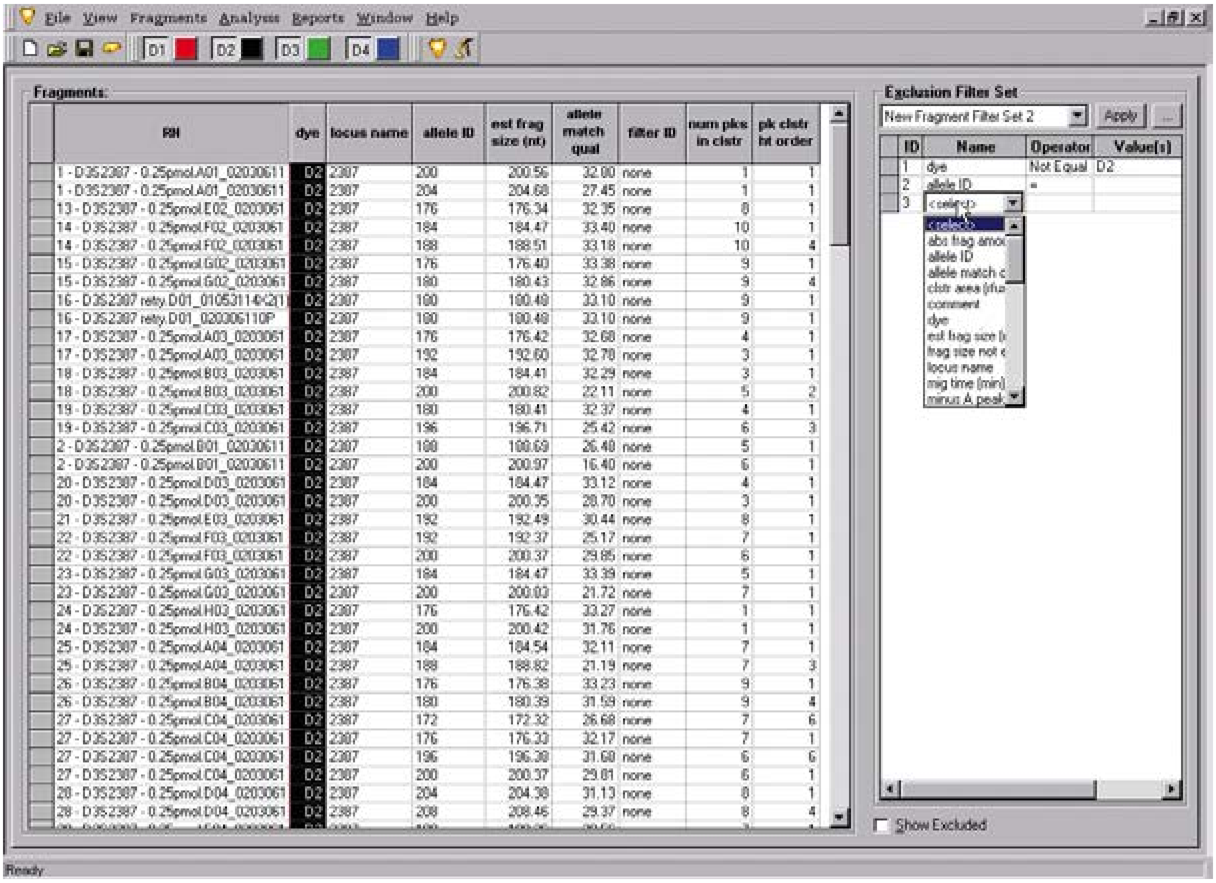
Exclusion Filter Set

2387

	ID	Name	Operator	Value(s)
	1	result	Not Match	2387
	2	size cal corr coef	<	.9999
	3	<select>		

☑ Show Excluded

Figure 4. *Building a Results Set Filter Set.*



Fragments:

	RH	dye	locus name	allele ID	est frag size (nt)	allele match qual	filter ID	num pks in clstr	pk clstr ht order
	1 - D3S2387 - 0.25pmol.A01_02030611	D2	2387	200	200.56	32.00	none	1	1
	1 - D3S2387 - 0.25pmol.A01_02030611	D2	2387	204	204.68	27.45	none	1	1
	13 - D3S2387 - 0.25pmol.E02_0203061	D2	2387	176	176.34	32.35	none	8	1
	14 - D3S2387 - 0.25pmol.F02_0203061	D2	2387	184	184.47	33.40	none	10	1
	14 - D3S2387 - 0.25pmol.F02_0203061	D2	2387	188	188.51	33.18	none	10	4
	15 - D3S2387 - 0.25pmol.G02_0203061	D2	2387	176	176.40	33.38	none	9	1
	15 - D3S2387 - 0.25pmol.G02_0203061	D2	2387	180	180.43	32.86	none	9	4
	16 - D3S2387 retry.D01_01053114K2(1	D2	2387	180	180.48	33.10	none	9	1
	16 - D3S2387 retry.D01_020306110P	D2	2387	180	180.48	33.10	none	9	1
	17 - D3S2387 - 0.25pmol.A03_0203061	D2	2387	176	176.42	32.68	none	4	1
	17 - D3S2387 - 0.25pmol.A03_0203061	D2	2387	192	192.60	32.78	none	3	1
	18 - D3S2387 - 0.25pmol.B03_0203061	D2	2387	184	184.41	32.29	none	3	1
	18 - D3S2387 - 0.25pmol.B03_0203061	D2	2387	200	200.82	22.11	none	5	2
	19 - D3S2387 - 0.25pmol.C03_0203061	D2	2387	180	180.41	32.37	none	4	1
	19 - D3S2387 - 0.25pmol.C03_0203061	D2	2387	196	196.71	25.42	none	6	3
	2 - D3S2387 - 0.25pmol.B01_02030611	D2	2387	188	188.69	26.40	none	5	1
	2 - D3S2387 - 0.25pmol.B01_02030611	D2	2387	200	200.97	16.40	none	6	1
	20 - D3S2387 - 0.25pmol.D03_0203061	D2	2387	184	184.47	33.12	none	4	1
	20 - D3S2387 - 0.25pmol.D03_0203061	D2	2387	200	200.35	28.70	none	3	1
	21 - D3S2387 - 0.25pmol.E03_0203061	D2	2387	192	192.49	30.44	none	8	1
	22 - D3S2387 - 0.25pmol.F03_0203061	D2	2387	192	192.37	25.17	none	7	1
	22 - D3S2387 - 0.25pmol.F03_0203061	D2	2387	200	200.37	29.85	none	6	1
	23 - D3S2387 - 0.25pmol.G03_0203061	D2	2387	184	184.47	33.39	none	5	1
	23 - D3S2387 - 0.25pmol.G03_0203061	D2	2387	200	200.03	21.72	none	7	1
	24 - D3S2387 - 0.25pmol.H03_0203061	D2	2387	176	176.42	33.27	none	1	1
	24 - D3S2387 - 0.25pmol.H03_0203061	D2	2387	200	200.42	31.76	none	1	1
	25 - D3S2387 - 0.25pmol.A04_0203061	D2	2387	184	184.54	32.11	none	7	1
	25 - D3S2387 - 0.25pmol.A04_0203061	D2	2387	188	188.82	21.19	none	7	3
	26 - D3S2387 - 0.25pmol.B04_0203061	D2	2387	176	176.38	33.23	none	9	1
	26 - D3S2387 - 0.25pmol.B04_0203061	D2	2387	180	180.39	31.59	none	9	4
	27 - D3S2387 - 0.25pmol.C04_0203061	D2	2387	172	172.32	26.68	none	7	6
	27 - D3S2387 - 0.25pmol.C04_0203061	D2	2387	176	176.33	32.17	none	7	1
	27 - D3S2387 - 0.25pmol.C04_0203061	D2	2387	196	196.38	31.68	none	6	6
	27 - D3S2387 - 0.25pmol.C04_0203061	D2	2387	200	200.37	29.81	none	6	1
	28 - D3S2387 - 0.25pmol.D04_0203061	D2	2387	204	204.38	31.13	none	8	1
	28 - D3S2387 - 0.25pmol.D04_0203061	D2	2387	208	208.46	29.37	none	8	4

Exclusion Filter Set

New Fragment Filter Set 2

	ID	Name	Operator	Value(s)
	1	dye	Not Equal	D2
	2	allele ID	=	
	3	<select>		

□ Show Excluded

Figure 5. *Building a Fragment List Filter Set.*

Filtering the Fragment List

The fragment list is generated from all results that are included in the study. The Fragment List set can be filtered upon any one of the fragment properties in Table 2. Whereas the Results Properties are characteristics of the complete results (groups of fragments), the properties of individual fragments are naturally distinct. The properties are accessed by pressing on the <select> button in the Filter set area (Figure 5).

The possible criteria a user may want to use to organize fragments may be:

- All fragments of a specific dye color
- All fragments of a specific size range
- All fragments above a specific height
- All fragments that have been identified to be part of a specific locus
- All fragments with identified alleles
- All fragments identified as unknown alleles
- All fragments that are one of too many alleles identified at a locus
- All fragments with peaks wider than a specified width

User-Defined Properties

There are some instances where important properties of results or fragments need to be assigned to data after it has been collected and passed through primary analysis. A good example is the assignment of pedigree data to results that are destined for linkage analysis. The information can be entered during sample plate setup, but it may also be entered into the Results Set Grid in the following way:

1. Create a new User-Defined Property Field (Figure 6)

    a. Launch a single sample view from any result in the study
    b. Go to the Property Set tab
    c. Right-click on the Property column header
    d. Select **Insert a Property** by left-clicking with the mouse
    e. Enter the Property Field Name
    f. Enter a value (optional)

2. Save the Study
3. Launch the Column Selector
4. Select the new Property from the User Defined Properties Folder, and position it for easy viewing
5. Sort the results in a manner so that the values are easily entered
6. Enter values

Note: Copying and pasting are allowed for one cell at a time.

User-Defined Properties can be sorted and filtered in the same way as all the other system-defined properties.

Removal of Filters

Individual filter elements can be removed by highlighting them in the Filter Set and selecting **Remove Selected Filter(s)** from the right-mouse-click menu. Alternatively, if the filtered column is displayed in the grid, the user may select **Remove Column Filter** from the right-mouse-click menu when the mouse is in the header region of that specific column. To remove all filters quickly, create or apply a blank Filter Set.

Managing Filter Sets

Filter Sets may be created and renamed using the Filter Set Manager which is accessed by pressing the [...] in the Exclusion Filter region of the Results Set (Figure 7). It is strongly recommended that the Filter Sets be given descriptive names shortly after they are created, and that a blank Filter Set be named to quickly undo the actions of custom Filter Sets. Filter Sets are applied automatically upon selection, but the [Apply] button must be used if edits are made to the current Filter Set.

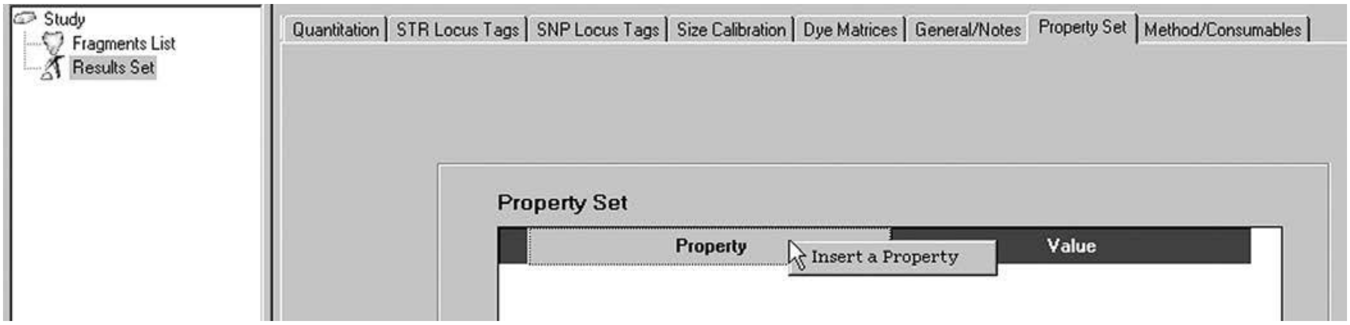


Figure 6. *Creating a new property for association with results. The Insert Property option appears by right clicking with the cursor over the Property Header. A specific value may be entered for this result. After creating the new user specified property, values can be entered for all results in the Result Set grid.*

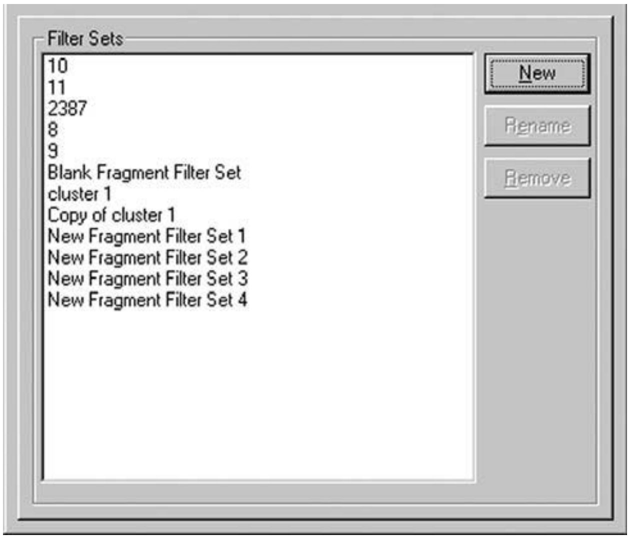

Filter Sets
10
11
2387
8
9
Blank Fragment Filter Set
cluster 1
Copy of cluster 1
New Fragment Filter Set 1
New Fragment Filter Set 2
New Fragment Filter Set 3
New Fragment Filter Set 4
New
Rename
Remove

Figure 7. *The Filter Set Manager. Named Filter Sets may be created, renamed, or removed.*

Persistence of Filter Sets and Study Management

Filter sets for both the Results Set and the Fragment List dynamically affect the data that is included in a study. In addition, all filter sets in the active database are available for use by all studies of the database. Therefore, to preserve the state of a Study's data content, it is useful to use the **Save Study As…** function to isolate a Study that has been completed. To limit the amount of unnecessary data in the new Study, excluded results can be highlighted and removed using the **Remove from Study** option from the File menu.

Practical Implementation of Filters for Minimum Acceptable Data Quality

With so many sample properties to review, is there a practical filter set which will recognize those samples that are most likely to be problematic? First, all samples must pass primary analysis, which is indicated in the analysis status column. Samples with serious current problems are the most likely to have missing data. This will be reflected in the electrical current values and the absence of an expected number of size standards. A simple, first-pass quality exclusion filter would look like this:

ID	Name	Operator	Value
1	analysis outcome	Not Equal	pass
2	current chng (%/min)	<	-0.2
3	current noise (%)	>	2

Beyond the initial filter, a user simply has to review the available filterable properties to see if one or more will be critical in the evaluation of the data.

Conclusion

The CEQ 8000 Genetic Analysis system software provides over 100 different filterable properties that can be mapped to the User's personal style to automatically review and organize his or her data. The software not only segregates data that does not meet the quality standards of the investigation but provides a flexible means of temporarily putting aside data that does not contribute to the current stage of the data analysis process.

Appendix A

Definitions for Selected Results Set Variables

General Result Information

Baseline

Each of the four channels will have a slightly different baseline level, with channels 1 and 3 (default trace colors blue and black) generally higher than channels 2 and 4. If baseline relative fluorescence units (RFU) are higher than 10,000 RFU in the raw data of any of the channels, sensitivity of the system to low abundance fragments will be reduced. The sensitivity is inversely proportional to the baseline noise.

Overrange

The optical detection system of the CEQ™ becomes non-linear above 100,000 raw data RFU and truncates raw data above 131,000 RFU due to limits of the optical detector. Signals above 100,000 may not be perfectly corrected for crosstalk, leaving signal behind that might be mistaken for real fragments.

RMS Noise

Noise expressed as the standard deviation of the observed baseline versus the normalized (flattened) baseline.

Relative Peak Height Threshold

The CEQ software identifies the two tallest peaks in each sample in each of the four dye colors. The user has the option during primary analysis to include all peaks above a percent height relative to the second tallest peak. This fraction is referred to as the relative peak height threshold.

Number of size standards not found/Number of calibration size standards not found

The number of size standards not found is the number of standards added to the sample minus the number of standards observed in the sample. Size standards may be missing if the separation time is too short for all of them to be observed. However, missing size standards could also be due to electrical current problems that retard the migration of all of the fragments of a run.

In setting up the analysis parameters, the user has the option of specifying subsets of available size standard fragments to be used in establishing the standard curve for each run. The subset becomes the new expected number of size standards, and any number smaller than the subset is recorded as the number of calibration size standards not found.

Fragment Quality

Broad peaks

Broad peaks are defined as peaks that are wider than 5/3 the half-height width of the nearest size standard in the same result. The presence of broad peaks generally indicates the presence of excess dye contaminants, some of which may be bound to sample components.

Spiky Peaks

Spiky peaks are defined as peaks that are narrower than 2/3 the half-height width of the nearest size standard in the same result. The presence of spiky peaks generally indicates the presence of sample contaminants.

Electrical Current Information

Relative Average Separation Current Slope

Ideally, after ramping up to full separation voltage, separation currents would remain constant (slope of 0). However, separation currents often decline slightly during CEQ™ separations. Decreases greater (more negative) than -0.2%/min may indicate a current abnormality in the run, shortening the effective separation of fragments.

Relative Standard Deviation of Separation Current

Also known as current noise, separation current tends to fluctuate slightly throughout the CEQ separation. Expected noise values are less than 2%. Fluctuations greater than 2% may indicate a current abnormality in the run, shortening the effective separation of fragments.

Sizing

Fragments Sizing Calibration Range

The CEQ will size all fragment between the smallest and largest size standards found after analysis of a separation. In addition, the software will extrapolate sizes 10% smaller than the 60-nt size standard and 10% larger than the largest size standard observed in the separation. The software will track the fragments outside the calibration range in two intervals (0-5% outside, and 5-10% outside). Fragments less that 5% outside the calibration range will have sizes and confidence intervals assigned, while fragments between 5 and 10% outside the range will have size assignments only.

Size Calibration Correlation Coefficient

The goodness of fit of the size standards to the user selected curve-fitting model is reflected in the Size Calibration Correlation Coefficient. Values <0.999 can be used to segregate rare separations where one or more of the size standards is shifted with respect to the others in the same run. The same shifting can be visualized by looking in the D1 channel in a binning analysis, or by viewing an Overlay Graph of the results.

Size Calibration Standard Deviation

The Size Calibration Standard Deviation is the average of all the deviations (in nucleotides) of the actual standard sizes from the sizes predicted by the fitted curve. Size Calibration Standard Deviations are generally <0.5 nt.

Color Correction

Dye spectra estimated from data/ Used system dye spectra

In every analysis, the software attempts to recalculate the dye spectra by examining the crosstalk between each pair of channels. In rare instances, the software is unable to calculate the spectra. In this case, the system will use the user-specified system dye spectra, if they are available. For each analysis, one of the two choices above will be selected. The user has the option of forcing the analysis software to use system spectra.

Appendix B

Definitions for Selected Fragment List Variables

Sizing

Confidence Interval Upper and Lower Bounds

Fragment size confidence intervals are statistical estimates derived from the user-specified confidence level and the size calibration standard deviation. At a confidence level of 95% (default), the upper bound and lower bound are the high and low size limits, respectively, that are likely to contain 95% of observed values of the same fragment separated many times.

Mobility

The mobility of a fragment is a measure of velocity of a charged species in an electric field divided by electrical field strength. While the migration time of a fragment will change when the run voltage is adjusted, the mobility estimate will remain constant.

Fragment Quality

Peak Width, and Efficiency

Peak Width and Efficiency and are interrelated values that reflect peak sharpness. Peak width is measured in mm, which reflects the baseline width of peaks as they pass the optical detection window. Under normal conditions, individual peaks range in width from ~0.5 mm to 2.0 mm. The efficiency, or theoretical plate number, is a measure of peak width taking into account distance the peak has traveled in the separation. Good efficiencies range from 1,000,000 to 5,000,000.

Specific Resolution

Specific Resolution is measured in 1/nt and combines the efficiency values with the selectivity (the ability to distinguish molecules that differ by 1 unit of size) of the separation system. Normal specific resolution values range from a low of 0.4 to a maximum of 1.5 (baseline resolved).

Peak Symmetry/Asymmetry

Fragments separated on the CEQ™ are usually symmetrical and readily fit to a Gaussian curve. Normal symmetry values range between -0.2 and 0.2, where a perfect symmetry value is 0. Lack of symmetry may indicate a capillary or gel problem.

Locus-Specific Information

Allele ID Confidence (match quality)

Alleles are identified based on the proximity of an unknown fragment of a specific color to a fragment in an allele list of a locus tag. Both the new fragment and the fragment in the allele list have confidence intervals associated with them. The greater the overlap between the two confidence intervals, the greater the match quality. Allele ID match qualities have a system maximum of 40. Quality values are calculated as -10 ($\log_{10}$ error probability):

Probability of Error	Quality Value
10%	10
1%	20
0.1%	30

Spurious Peak

Spurious peaks are peaks that are above the Relative Peak height threshold, but below the user specified Spurious Peak height threshold. Peaks above the spurious peak height threshold will be labeled as known or unknown alleles.

Stutter Peak

A stutter peak is defined as a peak that is an integral repeat unit away from an STR allele peak but is below a user defined Relative Stutter Peak height threshold.

Too Many Alleles

If the number of alleles (including unknown alleles) identified is more than the user-entered ploidy number, the software will register the state in the **Too many alleles** field.

Unknown Allele

As an unknown allele, a peak must satisfy the following conditions:

- Must be same color and in the size range of a locus tag that was used during analysis
- Must not be identified as one of the alleles at that locus
- Must be larger than the Spurious peak height threshold

No Alleles FounSd

If the no alleles are identified for a locus tag that is applied during primary data analysis, the software will register the state in the **No alleles found** field.

Fragment Quantity

Peak Cluster Size, Area, and Height Order

The software recognizes clusters as fragments of the same color in a single result that are not baseline resolved. The cluster size refers to the number of peaks that are in the same cluster as the peak in question, and the cluster area refers to the total area of the peak clusters in which the peak resides. The peak cluster height order identifies the relative peak height of the peak in its own cluster.

Relative Fragment Quantity

The relative fragment quantity is the fractional area that a peak constitutes of the total peak area of all sized peaks in the same color in a result.

* *All trademarks are the property of their respective owners.*

nsai
I.S. EN ISO 9001

For Research Use Only. Not for use in diagnostic procedures.

SCIEX
www.sciex.com/ce

View SCIEX products at www.sciex.com
Find your local office at www.sciex.com/offices

AB SCIEX Headquarters
500 Old Connecticut Path | Framingham, MA 01701 USA
Phone 508-383-7700
www.absciex.com