APPLICATION INFORMATION

Genetic Analysis: CEQ Series

AUTOMATED BINNING PROCESS FOR THE GENERATION OF LOCUS TAGS

Heather Gull, Dana Campbell, and Mark Dobbs
Beckman Coulter, Inc.

Short tandem repeats (STRs) have become one of the most widely used genomic markers for identity testing and gene mapping due to their high degree of heterozygosity. Accurate genotyping of STRs by fragment sizing relies on precise relative migration of identical alleles and prior knowledge of the spectrum of most or all possible apparent sizes that are derived from those relative migrations. The list of all possible alleles for a genetic locus is called an allele list. The allele list serves as a lookup table that will label fragments according to a naming scheme chosen by the user. In most cases, the apparent sizes will require sampling of a population of results that are separated under the same conditions that will be used in the larger experiment.

The most commonly used STR loci have alleles that are expected to increase in size by a fixed interval, generally between 2 and 6 nucleotides. However, exceptions to this rule are often observed. Some alleles do not adhere strictly to the common inter-allelic spacing at their loci. The alleles that fall between the more regularly-spaced alleles are sometimes referred to as non-integer repeats. In addition, alleles at some loci display spacing patterns that are marginally shorter or longer than expected. The trend is not identical for all STR loci and thus requires inspection on a locus-by-locus basis.

In the CEQ™ 8000 Fragment Analysis software, the process of binning compiles all of the pertinent fragment lengths from real data, estimates their most likely apparent sizes, and, taking size drift into account, assigns integer lengths (nominal sizes) to them. The product of the binning process is an allele list that then can be used to identify alleles whenever amplification products of the same genetic locus are separated under the same conditions.

Here we describe the process of automated allele binning using simple tools for recognizing the migration trends of STR alleles, identifying both integer repeat alleles and those fragments that may represent non-integer repeats.

Software Wizards

The CEQ 8000 Fragment Analysis software uses a number of software interface wizards—tools that aid the user in progressing through the necessary steps of selecting data and entering required values. You may not proceed to the next step of a wizard if critical information is incorrect or missing. The software usually will indicate the problem area using yellow highlights. The Binning wizard is one of the more sophisticated wizards, consisting of four interdependent screens.

Automated Generation of a Locus Tag Using the Automatic Binning Wizard

When data are plentiful, binning is the best way of creating an STR allele list. The observed sizes of fragments are rarely, if ever, integers but highly reproducible non-integer lengths (*A-1876A:* "Highly Precise DNA Sizing on the CEQ™ Genetic Analysis System"). The first step of the automated binning process is cluster analysis—the organization of observed fragment lengths into groups. Where the clusters are tight, the mean observed

sizes within each cluster represent the most likely length that will be observed for an unknown fragment with the same number of nucleotides.

To start a new Binning Analysis, select this option from the analysis menu or by right-clicking on the Binning folder in the **Analyses** tab of the Study Explorer. Each peak included in the study is a candidate for the binning process. On Screen 1, the user simply views the data (Figure 1) in each of the dye colors, selects the dye, size range, maximum bin width, minimum number of data points per bin, repeat unit length, and an allele naming convention before proceeding to the next step of the process. The wizard steps are fast and reversible so not all selections need to be correct at the outset.

Screen 2 of the binning wizard normalizes the peak heights for the fragment length range selected and displays a prediction of the positions of the regularly spaced alleles. Several options are now available to validate new allele lists (Figure 2).

Trace Inspection

View the trace of any point in the **Bin View** scatter plot by left-clicking on it. To hide a trace that has been launched, select this option from the right mouse button menu. The bins overlaying the traces can be turned off by unchecking the **Show Bins** checkbox in the **Trace Views** area.

Phase Shift

Shift the phase of the bins in the plus or minus direction. The effect will be to shift all the bins in single nucleotide increments in the event that the software did not automatically select the desired peaks to build the allele list. Both the Nominal and Apparent sizes will be shifted by those same increments until a shift equivalent to a full repeat unit is reached, at which point the effect of the shift will be nullified.

Minimum Relative Peak Height

Change the minimum relative peak height to exclude additional peaks that are used in the cluster analysis. The feature is useful for excluding peaks from binning that are clearly not critical for building the allele list but were not previously excluded by other means.

Show Phantom Bins

Phantom bins are the bins that fall in between perfect repeat allele positions. Phantom bin positions are easily calculated by the software based on the repeat unit length and the register, or spacing, of the perfect repeats and the peaks associated with them (e.g., +A). When peaks do not fall into perfect repeat bins, two observations lend support to the idea that they are non-integer repeats:

1) The peak cluster pattern, or signature, of the amplification product is similar to the signatures of the perfect repeats; and
2) The candidate allele peak falls in a phantom bin, suggesting that its secondary structure is consistent with other products of the locus (see Discussion).

Phantom bins may be converted into real bins by left-clicking on them and then selecting **Create Allele** from the right-click menu. Alternatively, one can add an allele above or below any row of the allele list using the right-click menu option when the cursor is over the allele list grid. Conversely, alleles may be removed from the allele list using the opposite commands.

Regression Plot

The regression plot displays the linear relationship between nominal sizes and apparent sizes (Figure 3). The regression plot can be accessed in the right-

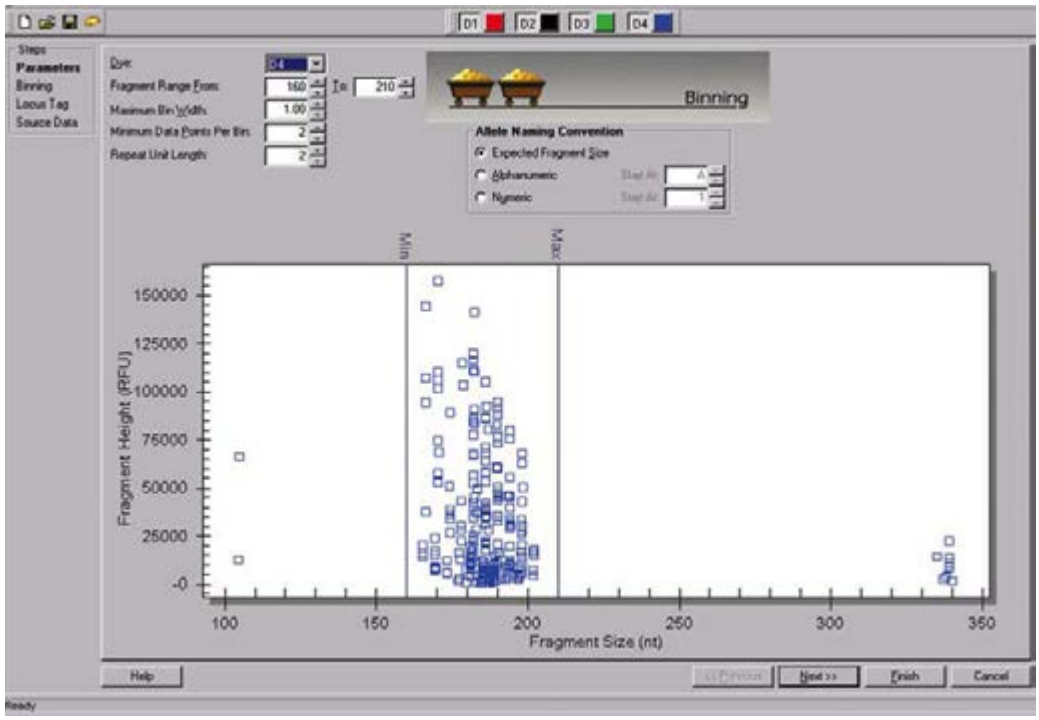

Figure 1. *First screen of the binning wizard.*

click menu when the cursor is on the Bin View. The software selects nominal sizes to obtain the best linear fit with the available data from the cluster analysis. However, because the nominal sizes and apparent sizes are user-editable, it is possible for the user to inadvertently reduce the goodness of fit of the regression line. Editing a nominal size so that it is one integer nucleotide too high or too low would

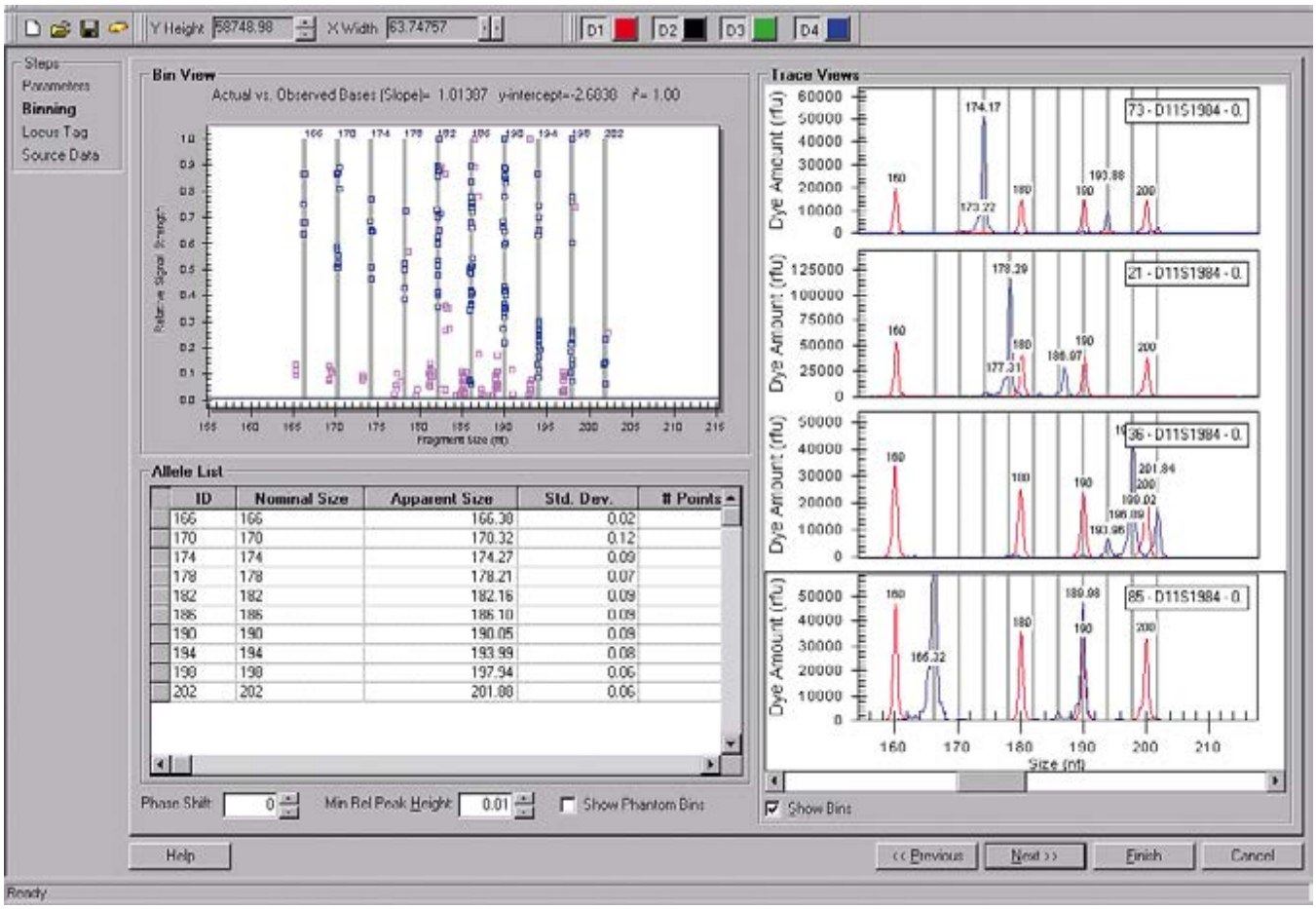

Figure 2. *Second screen of the binning wizard.*

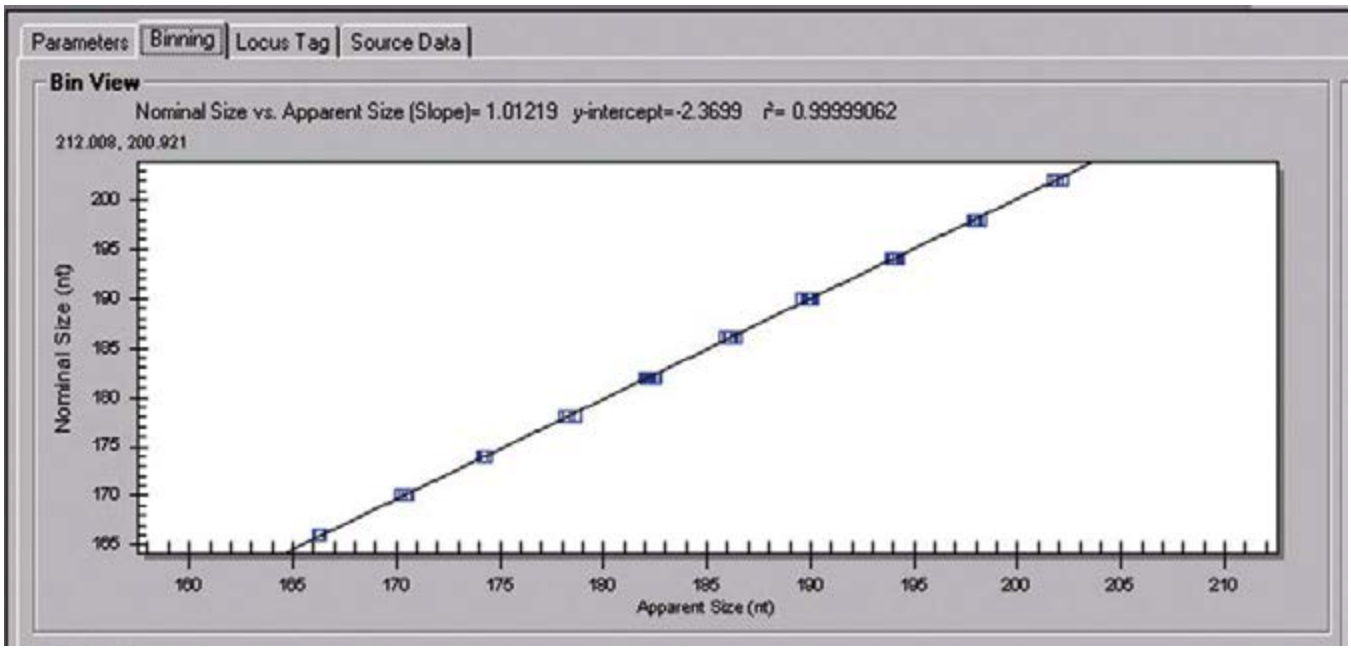

Figure 3. *Regression Plot. The regression plot displays the nominal size versus the apparent sizes for fragment length clusters from the results in a study.*

create a disjointed set of points at the position of the error. The regression plot will take user edits into account and reflect the reliability of the nominal size estimates. The fit of the line is described by three values: the nominal versus apparent size slope, the y-intercept, and the correlation coefficient (r^2). Based on our observations, the slope should always be between 0.95 and 1.05, and r^2 should be >0.9999. The most significant impact of a bad edit would be a reduction in r^2. The values are displayed above both the regression plot and the Bin View scatter plot.

Screen 3 of the binning wizard prompts the user to specify a unique **Locus Tag,** which is the name referenced by the software, and a **Locus Label,** which is the name that the software applies to the alleles when they are identified. The two names can be the same. The allele list is carried forward from Screen 2.The **Locus Tag** tab (Figure 4) contains locus-specific information, some of which has already been entered in the binning process (*e.g.*, repeat unit length) and some of which is for documentation purposes only (*e.g.*, primer set names and sequences). The **Allele ID Criteria** tab (Figures 7 and 10) provides options for interpreting the +A artifact and discriminating against stutter. The proper use of these options is described in the *CEQ™ 8000 Genetic Analysis System User's Guide* (Beckman Coulter PN 608315).

The final Screen 4 of the wizard reviews the source data that was used to initiate the binning analysis. If no changes to the result set have been made, the source data list will be identical to the results set list. However, the source data list will remain with the binning analysis that it gave rise to even as the results set is modified or manipulated for other purposes.

Using the Allele List to Identify Alleles

It is important to keep two points in mind during the creation of an allele list using the binning process:

1) Not all true allele peaks are required to build the list; and
2) Peaks that are used to build the list are not automatically assigned allele IDs.

The first point is important because it enables the user to sample the data without including weaker peaks that may blend with noise peaks from other results. In the second analysis of the data that is required to perform the actual allele assignments with the new Locus Tag, the sensitivity of peak detection can be raised to include all alleles, strong or weak. While this two-step approach is valid for simple and clean alleles, it is not foolproof. Complex allele patterns may require the use of more advanced tools of the CEQ 8000 software. Both simple and complex allele patterns are considered in the examples following.

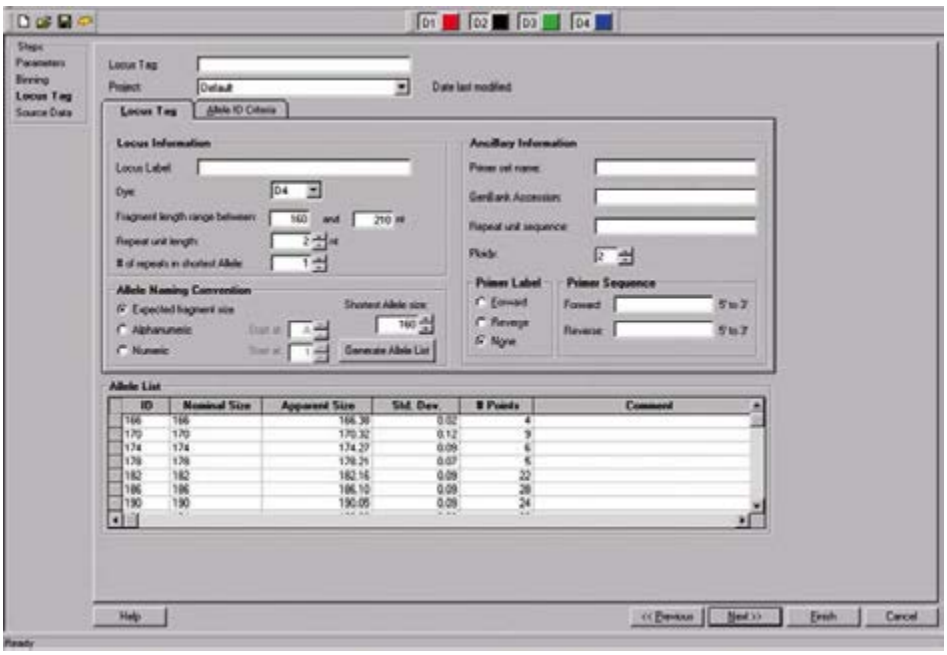

Figure 4. *Third screen of the binning wizard, displaying the Locus Tag tab.*

Case 1—One peak per allele, no stutter

Sample Locus	D22S683
Repeat Unit Length	4 (GATA)
Peaks per Allele	1
Conventional Stutter	None

Special Binning Procedures

There were so many non-integer repeats at this locus that it was appropriate to set up a binning analysis with a repeat unit length of 2 instead of 4.

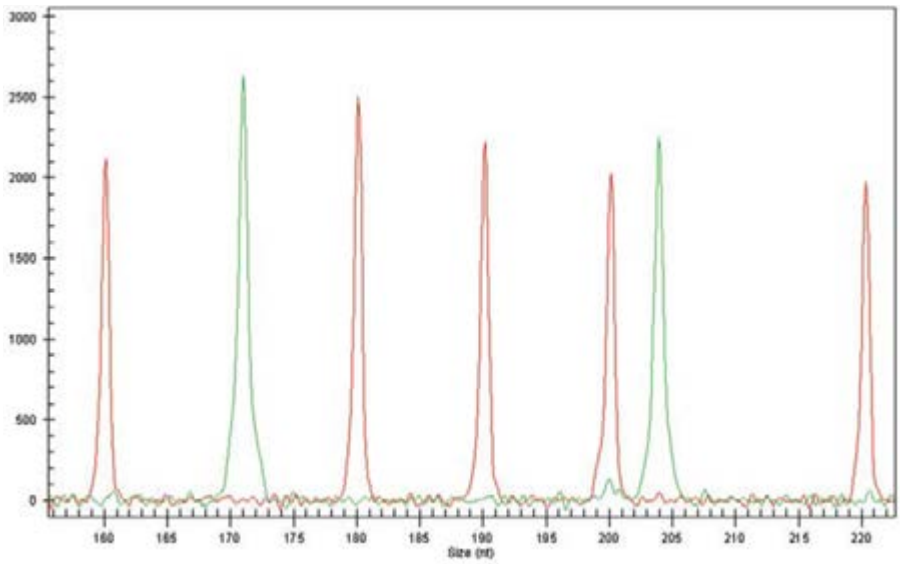

Figure 5. *D22S683 Allele Signature (heterozygote).*

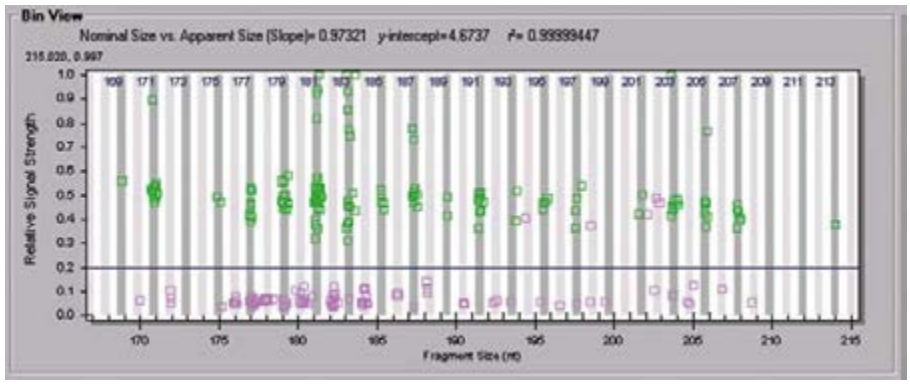

Figure 6. *D22S683 Initial Bin View.*

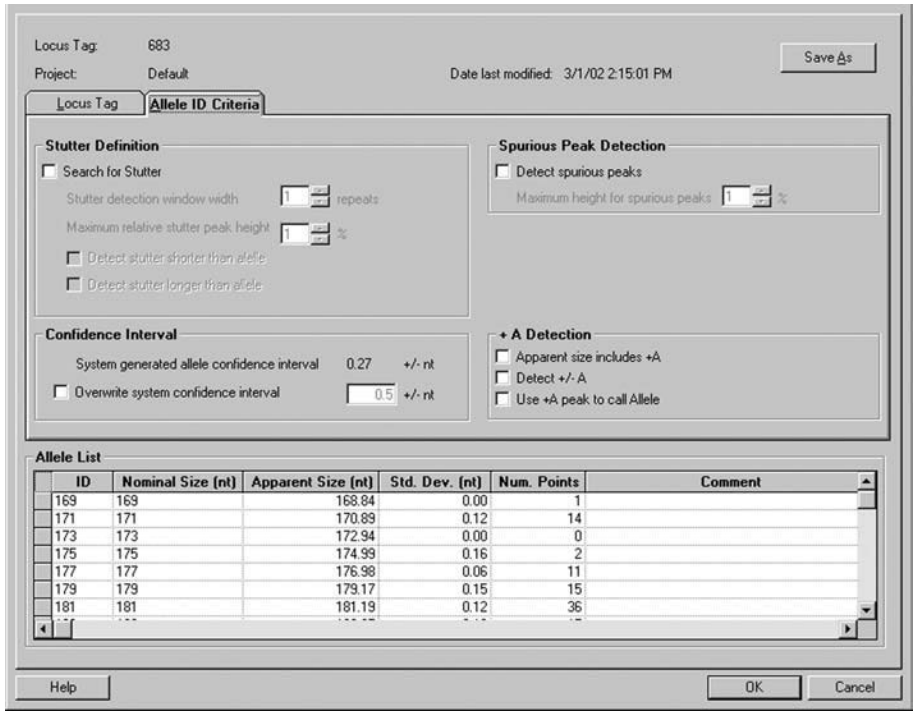

Figure 7. *D22S683Allele ID Settings.*

Case 2—Strong +A conversion, weak stutter

Sample Locus	D11S1984
Repeat Unit Length	4 (GGAA)
Peaks per Allele	2
Conventional Stutter	Weak

Special Binning Procedures

None.

Allele ID Notes

Conventional stutter peaks though smaller than true alleles will often be registered as identified alleles because they are full repeat units away from the true alleles.

Both -A and +A forms of an allele, if they are large enough, will be identified as alleles unless Detect +/-A is selected.

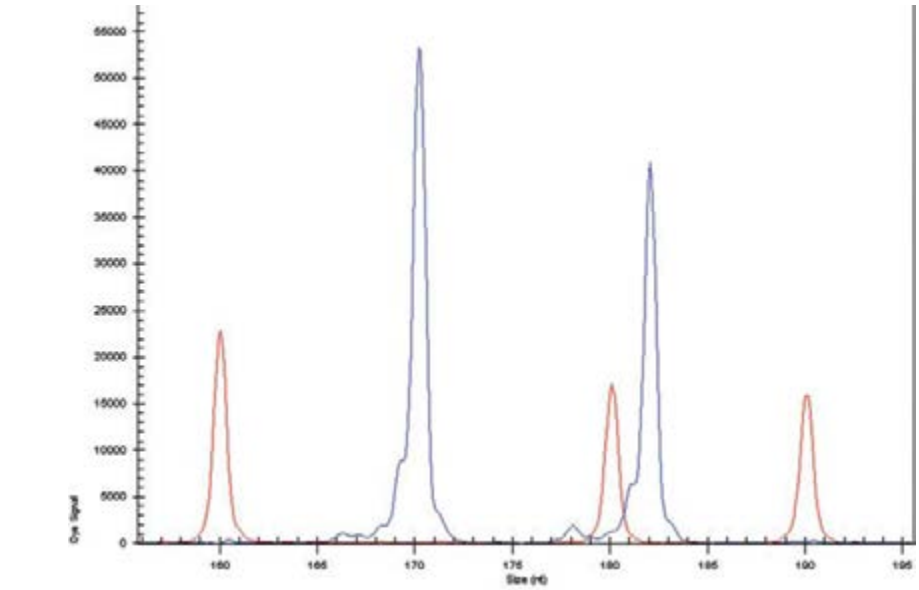

Figure 8. *D11S1984 Allele Signature (Heterozygote).*

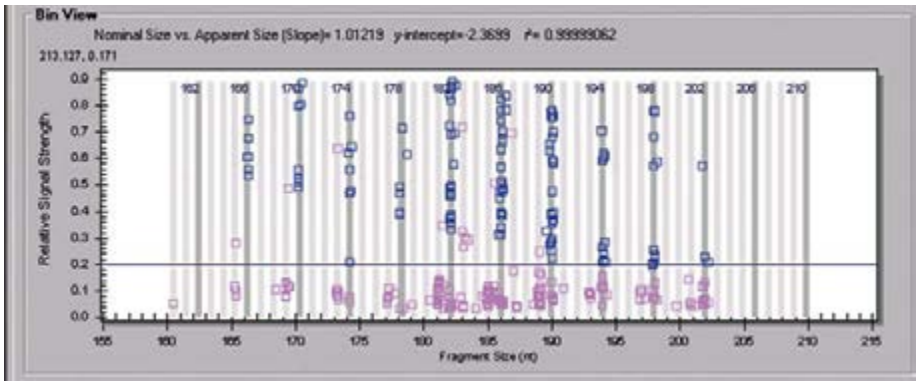

Figure 9. *D11S1984 Initial Bin View.*

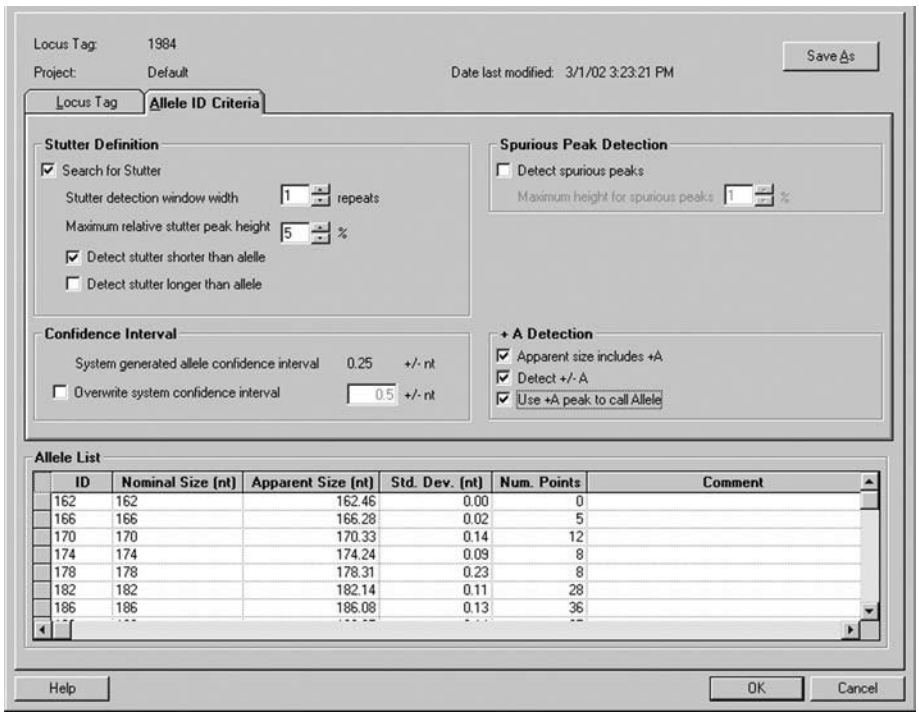

Figure 10. *D11S1984 Allele ID Settings.*

Case 3—Two peaks of similar height per allele, leftmost selected as allele

Sample Locus	GATA193A07
Repeat Unit Length	4 (GATA)
Peaks per Allele	4
Conventional Stutter	Weak

Special Binning Procedures

Check to ensure that the proper phase has been selected by the automatic binning software (the leftmost peak of each doublet). Launch some trace views by clicking on points from within bins. If incorrect member of doublet is selected, use Phase shift tool to correct.

Allele ID Notes

In the case of greater than two peaks per allele, it is not clear which peak is the -A form and which is the +A form. However, since two peaks per allele are taller than the others (the others can be excluded during analysis, or after analysis using filtering), we can use the +A tools to prevent the second peak of the tall doublet from being called an unknown allele as indicated in the allele ID settings.

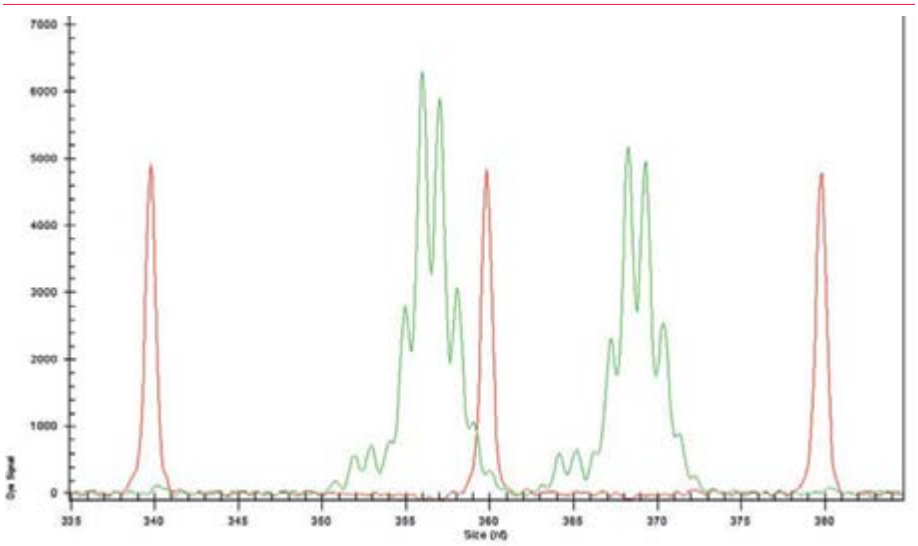

Figure 11. *GATA193A07 Allele Signature (Heterozygote).*

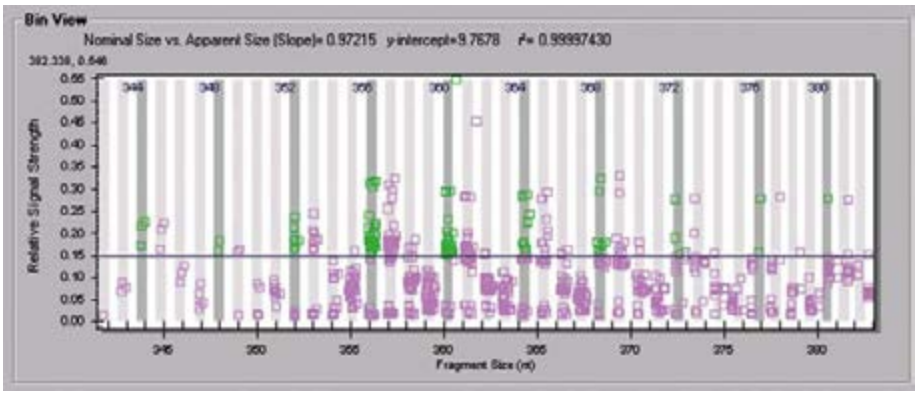

Figure 12. *GATA193A07 Initial Bin View.*

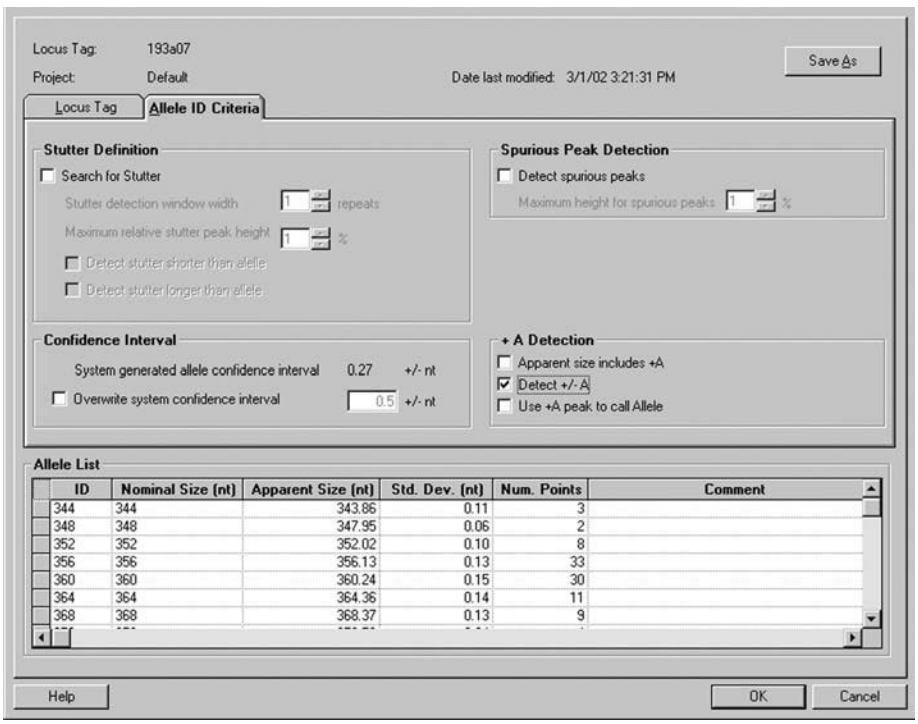

Figure 13. *GATA193A07 Allele ID Settings.*

Case 4—Large number of peaks per allele

Sample Locus	D3S2387
Repeat Unit Length	4 (GATA)
Peaks per Allele	7-9
Conventional Stutter	Unknown

Of the four cases presented, this is the most difficult because the allele clusters are composed of several peaks of similar height. This situation is problematic for both the binning process and for the identification of allele peaks after the allele lists have been generated. Let's examine the two problems one at a time.

1. Binning Procedures

The biggest challenge in the construction of the allele list from complex alleles using the binning algorithm is noise. In our example, the compiled collection of fragments spans the entire locus range with some peaks at every possible position. To address the large number of peaks per allele, we consider four different strategies for automating the allele list construction.

Strategy 1: During primary data analysis, only those fragments that are ≥ 99% the height of the second tallest fragment are included (relative peak height threshold = 99%).

Advantages: works when applied to both simple and complex alleles, because the two tallest peaks in a trace are always included.

Disadvantages: independent analysis method just for binning; when multiplexing products of the same color in different size ranges, weaker loci will be excluded.

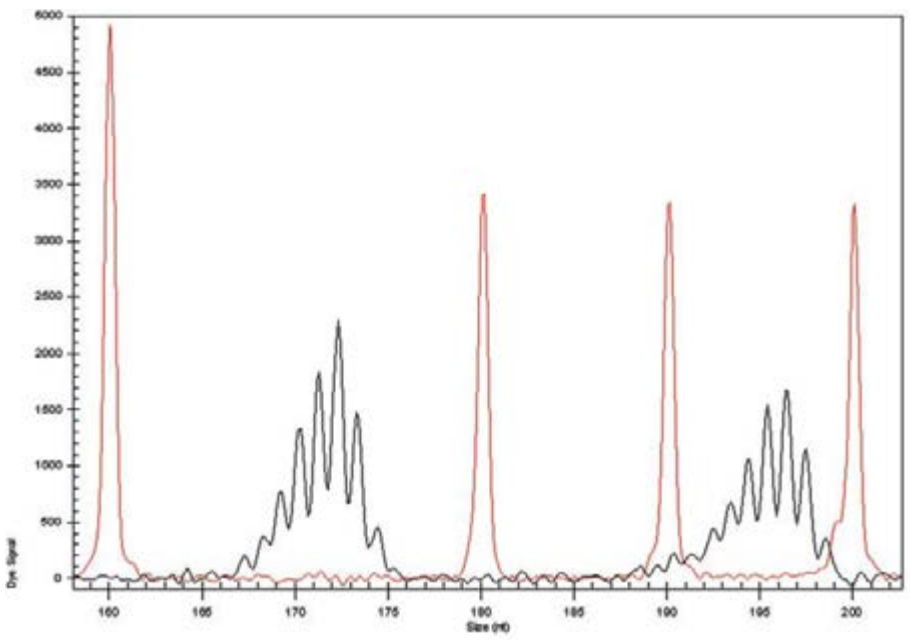

Figure 14. *D3S2387 Allele Signature (Heterozygote).*

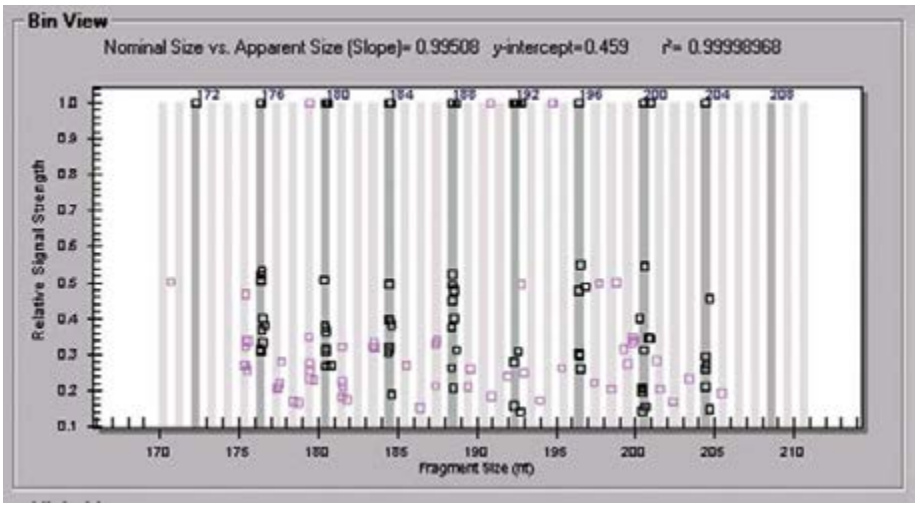

Figure 15a: *D3S2387 Bin View scatter plot using fragments that are >99% the height of the second tallest fragment are included (relative peak height threshold set to 99% during primary data analysis). Note that the Y-axis starts at 0.1 because there are no peaks with a Relative Signal Strength below this number.*

Strategy 2: Exclude peaks of below a fixed relative peak amount by applying a filter to the Fragment List before the binning process is initiated (Figure 15b).

Advantages: offers more flexibility than **Strategy 1** enabling the exclusion of any desired relative quantity of fragment.

Disadvantages: when multiplexing products of the same color in different size ranges, weaker loci will be excluded.

Strategy 3: Exclude data by raising the minimum relative peak height during the binning process (Figure 15c).

Advantages: includes weaker peaks in the general analysis; excludes them only in the phase of analysis where they are not needed; relative peak heights in binning are calculated for the locus range only, to take into account loci of different intensities within the same samples (multiplexed samples).

Disadvantages: sometimes difficult to decide where to set the minimum relative peak height.

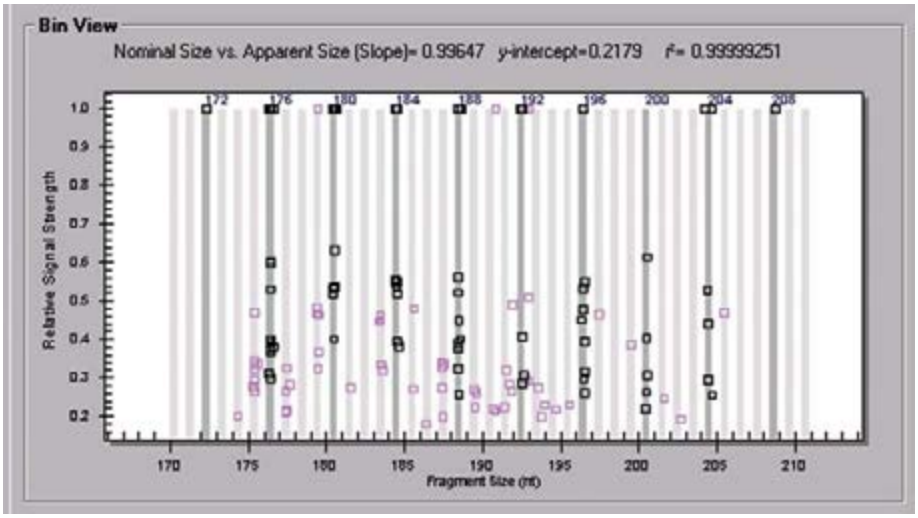

Figure 15b. *D3S2387 Bin View scatter plot from fragments that are above a* ***Relative Signal Strength*** *of 0.15 by applying a filter to the Fragment List before binning.*

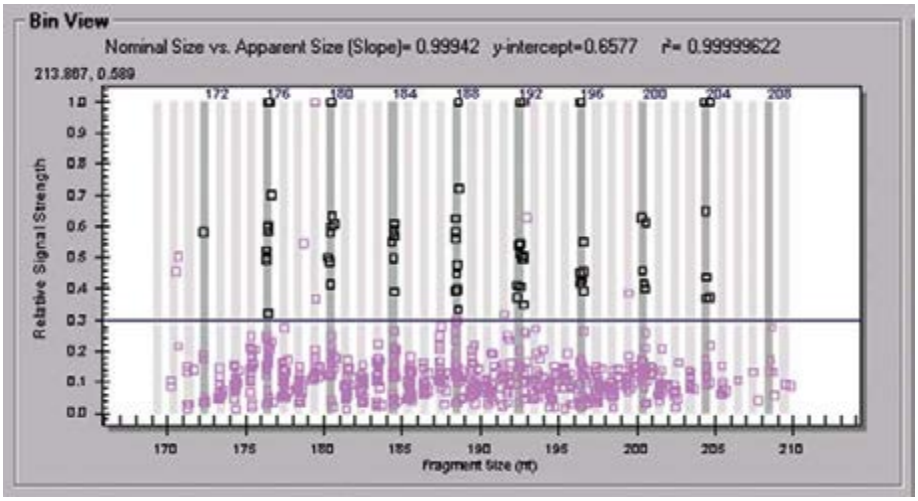

Figure 15c. *D3S2387 Bin View scatter plot from fragments where the minimum relative peak height was raised to 0.3 during the binning process.*

Strategy 4: Filter on the fragment list to exclude all fragments that were not the tallest peaks within their own peak clusters.

Advantages: preserves the effective recognition of smaller peaks during primary analysis, while recognizing the tallest members within each fragment cluster.

Disadvantages: when peak clusters overlap, the tallest peaks within lower height sub-clusters will not be recognized.

The decision regarding which noise reduction strategy to use will depend on the complexity of the locus, the differences in peak heights between shorter and longer alleles at the same locus, the potential for overlap of peak clusters, and the peak height differences between peaks chosen as alleles and the remaining peaks. For D3S2387, Strategy 4 is perhaps the best because it selects nearly all of the peaks that we would select by manual inspection of the traces. For the purposes of allele list construction, the losses due to overlapping peak clusters are not significant.

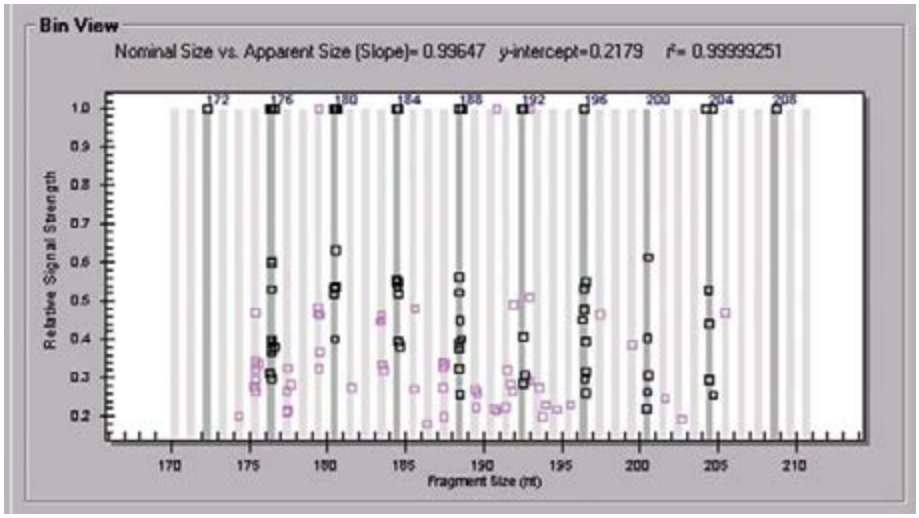

Figure 15d. *D3S2387 Bin View scatter plot using fragments that were the tallest peaks within their own peak clusters.*

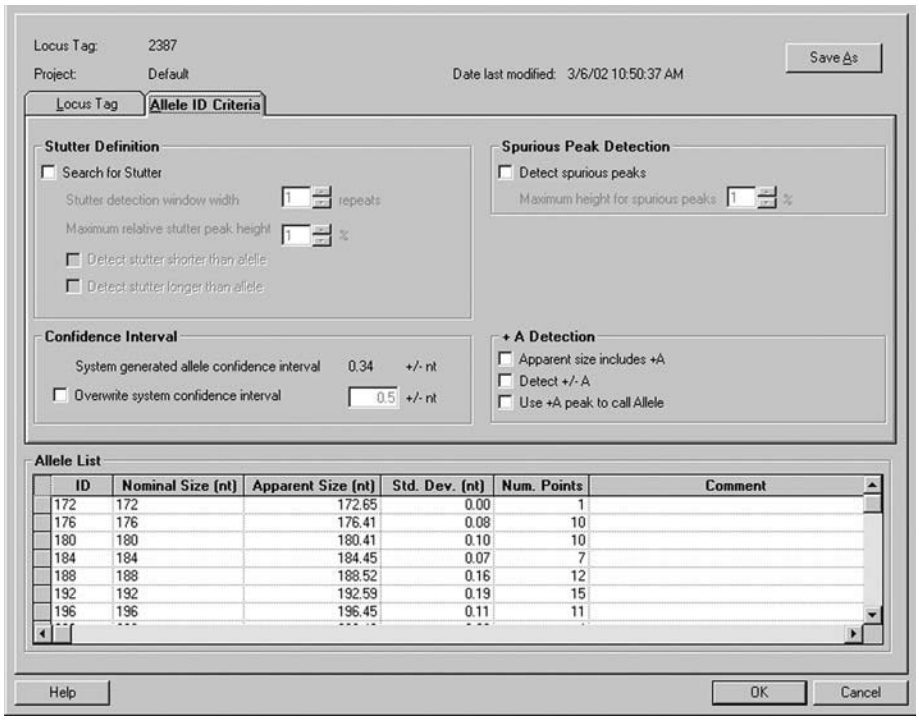

Figure 16. *D3S2387 Allele ID settings.*

2. The identification of allele peaks after the allele lists have been generated

Allele ID Notes

After building the allele list, one now has to deal with the problem of preventing non-allele peaks of being identified as unknown alleles. If there were no overlapping peak clusters, one could filter out all but the primary peaks, and be left with only the identified alleles. In this approach, alleles that were part of overlapping peak clusters would be lost. The only remaining approach is to apply the Locus Tag to the data using reasonable sensitive peak detection parameters (the default slope threshold = 10, relative peak height threshold =10% is sufficient) and none of the +A checkboxes selected. This process will identify multiple alleles per locus, only some of which are real. The list must be reduced by selecting the allele peaks manually.

Left click on the fragment list and choose **Manually Select Peaks**. When given the option to clear the fragment list, select **Yes**. A stacked graph of all the results in the study will be presented on the right. Point the cursor at the apex of the tallest peak of each cluster and left click to select it. Then, right click, and select **Include**. Each selected peak, including its allele ID, will be added to the list.

Finally, if the tallest peaks in the allele clusters do not happen to coincide with the perfect repeats in the allele list, their allele IDs will be blank in the fragment list. Sort the fragment list based on esti-

mated fragment size (nt). It should be relatively easy to interpret the allele IDs of the imperfect repeats based on the surrounding alleles. Enter these IDs into the fragment list.

Application and Management of Locus Tags

When a Binning Analysis is saved, the allele list is frozen as part of the specified Locus Tag. The Locus Tag name is not editable except through the Data Manager. However, the saved binning result keeps the original Locus Tag name, even if it is renamed or deleted.

There are two ways to edit the allele list:

1. The binning analysis can be re-opened, edits made, and the study saved. In this instance, the original Locus Tag name is used, regardless of whether the Locus Tag was renamed or deleted from the database.
2. The Locus Tag can be accessed through the **STR Locus Tags** tab of the Analysis Parameters menu option by selecting Edit when any Analysis Parameter set is selected. First the Locus Tag is selected and then the **Edit Locus** button is pressed. The list of available Locus Tags reflects the current elements of the database.

Using the **Re-analyze results** command to apply Locus Tags replaces the current study results with new results, and displaces, but does not discard, the old results. The consequence will be to remove from the study the first results that comprised the source data list. It is also important to remember that the data points that are viewed in the binning analyses are dynamic - they reflect the current state of the results set and the fragment list. Thus the application of new filters or the manual exclusion or inclusion of results or fragments will be reflected in the Bin Views, despite the static bin positions and statistics. To force the binning analysis to conform to the new data, **Re-bin** must be executed from the Binning menu option (the Binning menu is available only when the binning tab is viewed). Note that re-binning will undo most manual edits to the allele list. Re-binning is useful when results with contributing alleles for the locus have been added to the study.

The above points are all important to consider when deciding how to segregate results that have been used for binning and in the construction of allele lists from those results that have identified alleles. In most cases the same data is used for both. To preserve the integrity of the binning analysis, the best choice is to create new study from raw data, applying the new Locus Tags to the data. However, if you believe that the allele list will not require any future edits, you may re-analyze the results using Analysis Parameters that include the new Locus Tags. If re-construction of the original binning analysis is then required, the list of source data can be used to re-build the study from those results.

Discussion

The most reliable method of determining longer DNA fragment lengths is by DNA sequencing. However, DNA sequencing cannot be multiplexed, and provides more information at each locus than is normally required for identity testing or linkage studies. Fragment sizing of end-labeled polymorphic short tandem repeat amplicons by high resolution gel electrophoresis has become a rapid alternative tool for genetic analysis in mammals. Precise sizing of DNA fragments is essential in dealing with both common alleles and rare variants that may differ in length by as little as 1 nucleotide, a feat that is well within the capabilities of the CEQ™ 8000 (*Application Information Bulletin A-1876A*). In addition, many short tandem repeat sequences have been well characterized, their estimated sizes need to be re-evaluated whenever new separation systems or new separation conditions are used.

We noted that when all peaks from a locus are taken into account, the differences in apparent nucleotide length are not always identical to the differences in true size, *i.e.*, a spacing of 1.00 real nucleotides per 1.00 observed nucleotides is not always observed. The CEQ 8000 binning software quantitates the relationship in a term called the **nominal versus apparent size slope**. In two of the loci above (case 1 and case 3), we demonstrate a slope of ~0.97, indicating that, for every fragment length increase of 1 nucleotide, we observe an apparent increase of 0.97 nucleotides. The most likely explanation for this phenomenon is that the DNA fragments from different loci have slightly different mobilities due to secondary structure that are not corrected for by the internal size standards. If rounding or truncation of the apparent sizes were used to assign nominal sizes, the subtle drift would eventually lead to errors in predicting allele lengths. The CEQ 8000 binning software uses the nominal versus apparent size slope to take the drift into account. For all loci that we have examined, the observed base change per actual base difference is linear and well behaved.

The process of automated allele binning greatly facilitates the development of an allele list that pre-

dicts the observed lengths of all alleles at a genetic locus. In the case of simple allele peak signatures, allele IDs may be specified for both common and rare alleles. As illustrated in the cases with complex alleles, some degree of population sampling may be required to establish the integer repeat peak distribution pattern. For most allele peak patterns, the ID names may be applied directly during a second analysis step. For some complex allele patterns, the user has the freedom to use the electropherogram traces to manually select those peaks that represent true alleles.

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