FlexiBin

A Program to Automate the Binning of Microsatellite Alleles

by Bill Amos

Department of Zoology, Downing Street, Cambridge, CB2 3EJ, UK

FlexiBin is a program written in Microsoft Visual Basic for Excel and runs as a Macro on either PC or Macintosh computers. It was written because of the inadequacy of extremely expensive programs that were available at the time as the primary software for ABI sequencers. The main problem I address is that the effective allele size is seldom exactly 2 bases (for a dinucleotide repeat), but instead is often somewhat longer or shorter. Consequently, forcing alleles into an exact 2 base periodicity will result in miscalling.

The program is emailed with representative data in worksheet "Input", comprising data from a human genome screen with markers from chromosome 13. Input data are in columns 1 to 3 and the output is to the right and on sheet "OutPut". Input data are in the form of a single column of genotypes with no missing data and arranged Locus name, allele 1, allele 2. After binning, the program generates an ordered array of alleles in columns S (= locus), T (= allele number), U (= raw length), V (= binned allele number), W (= sample number) and X (=allele number). Columns W and X are included to facilitate manual editing if the binning is thought to be incorrect: simply edit the offending alleles and sort first by allele number then by sample number. Summary statistics are output to a table (columns M to Q) and a graph depicting a summary of how the binning process went lies between the input data and the output table.

The Binning Graphs

These are the standard way many people bin alleles manually. The graph represents a cumulative length distribution. Alternate bin classes are coloured red and blue alternately. This way binning problems are easily spotted. The first locus is deliberately chosen to be the worst case (out of 300 loci!), with obvious miscalling in bins 4 – 7. The likely reason is a single base deletion / insertion in the shorter alleles, though this needs to be confirmed. In all other cases the binning appears to be as good as you might achieve by eye, though there are a few instances where two classes meet and may even overlap.

The Binning Table

As a second, numeric guide to the binning process, data are also presented as a table. This give summary statistics for the binning process: specifically the effective length of a repeat unit and the adjust value (the remainder left after an exact number of repeat units has been subtracted). In addition, the bin classes are listed, with for each an expected length, a mean length of alleles placed in that bin, a count of how many alleles were found and, most importantly, a standard deviation of binned alleles. When the binning works well, the standard deviation seldom rises above about 0.35. Higher values are indicative of problems and may encourage closer scrutiny of the relevant graph.

The program assumes that most binning will proceed quite happily, so automatically creates an output genotype list in sheet "OutPut" in which all the input alleles have been replaced with their inferred bin numbers. Bin numbers are estimated *relative* repeat numbers. Absolute repeat numbers can be obtained by reference to a sequenced allele.

Running the program

To run the program, paste you data into the first three columns. The current version will accept 4 alleles in each genotype as a result of a historical request. However, there must be **no gaps**. Multiple loci can be analysed (as in the example data) by stacking each locus one above the other with no gaps, a change being inferred at every place the locus name changes. Be careful, the program is currently case sensitive, so if you sometimes type locus 'abc' and sometime 'ABC', you will end up analysing lots of very small datasets!

After pasting your data into the sheet to replace (i.e. delete) the existing data, choose <Tools>, <Macro>, <Macros . >, select 'Binning' and <Run>. Other ways involve entering the Visual Basic editor, placing the cursor in the first subroutine ('Binning') and pressing key F5. Current settings favour resolution over speed. I am pretty sure that you will get essentially identical results in a tenth the time. If this is important, let me know and I'll make the changes.

Development

The current version is set for dinucleotide loci where problems are most intense. It will probably work well with tetranucloetides, though the need here is likely to be much less. Future versions aim to allow the user to input the repeat size for each locus independently. I also intend to add a program 'ReBin' automatically to rectify problem loci and hence to save the need for manual changes to the genotypes.

Any and all comments welcome, as well as requests for new features. Needless to say, this program is distributed in good faith as a freeware tool and, while I have tried my best to eliminate bugs, I cannot accept responsibility for downstream problems should any bugs be present. I hope very much that it saves you time and improves accuracy!

Bill Amos
December 2005