PeakFinder User's Guide

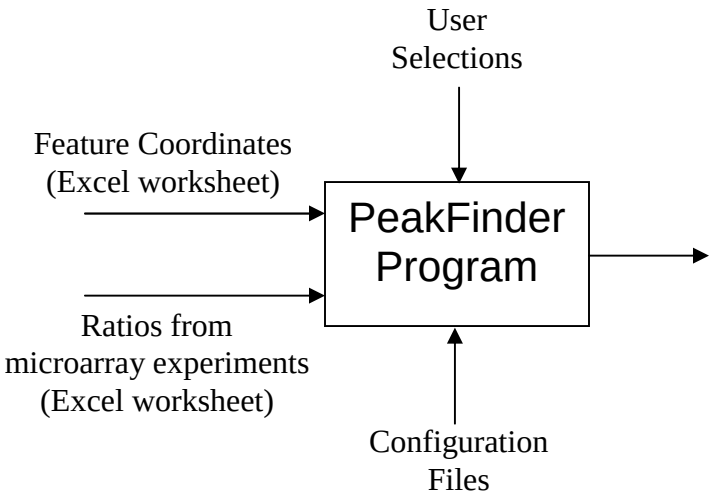
Supplementary material for
Genome-wide mapping of the cohesin complex in the yeast *Saccharomyces cerevisiae*
Glynn EF, Megee PC, Yu Hg, Mistrot C, Unal E, Koshland DE, DeRisi JL, and Gerton JL
http://research.stowers-institute.org/jeg/2004/cohesin

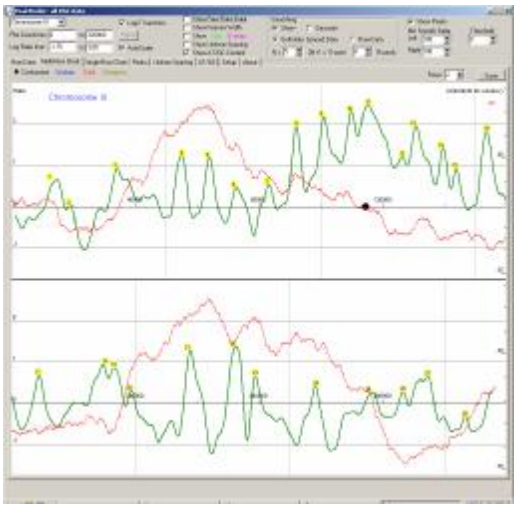
Software and User's Guide written by Earl F. Glynn

Please address *PeakFinder* questions to
e f g @ s t o w e r s – i n s t i t u t e . o r g

1 Overview

The PeakFinder program has been developed to find cohesin binding sites represented by the peaks in yeast chromatin immunoprecipitation (ChIP) microarray data, but can be applied to plot any measurement against a parameter such as genome coordinate, to interactively analyze the measurement plot, and to annotate the peaks on the basis of local properties of the curve.



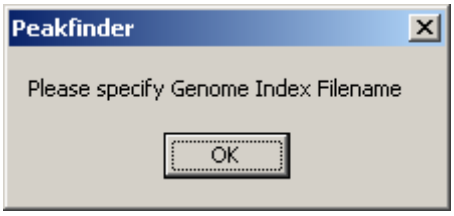


2 Files

The PeakFinder program needs a genome index file (ASCII text), two Excel worksheets, and optionally, a directory of sequence files. In addition, an initialization file is used to store certain selections between runs.

2.1 Genome Index File

If PeakFinder has never been used before, this message will be seen when the program first starts:



Press the button "1. Genome Index File" and specify this file.

The first row of this file has column headers and is skipped by the PeakFinder program. The columns expected in this file are as follows:

1. Number — Chromosome Number (e.g, 1 .. 16)
2. Name — Chromosome Name (usually Roman Numeral, e.g, I .. XVI).
3. First_bps — First bps in genome for this chromosome
4. Last_bps — Last bps in genome for this chromosome ("-" except for last)
5. PlotLen — Chromosome length used in charts
6. CenLow — Centromere "low" location
7. CenHigh — Centromere "high" location

Here's what a sample Genome Index File looks like:

YeastIndex.dat

#	Name	First_bps	Last_bps	PlotLen	CenLow	CenHigh
1	I	0	-	234000	151457	151595
2	II	230204	-	816000	238168	238284
3	III	1043346	-	320000	114379	114495
4	IV	1358685	-	1536000	449707	449818
5	V	2890659	-	582000	151960	152113
6	VI	3467529	-	276000	148503	148621
7	VII	3737677	-	1092000	496914	497032
8	VIII	4828613	-	564000	105579	105696
9	IX	5391251	-	444000	355626	355743
10	X	5831136	-	750000	436000	436116
11	XI	6576576	-	672000	439774	439888
12	XII	7243024	-	1080000	150827	150946
13	XIII	8321196	-	930000	268031	268141
14	XIV	9245626	-	786000	628757	628866
15	XV	10029956	-	1092000	326584	326702
16	XVI	11121239	12068940	954000	555952	556069

The first and second columns give the chromosome number and label. The third column, *First_bps*, gives the index of the zeroth nucleotide for the given chromosome. The first nucleotide for a given chromosome is one greater than the value shown. A dash ("-") can be used for the *Last_bps* value on all chromosomes except the last.

The First_bps index values were taken from a version of Joe DeRisi's (http://derisilab.ucsf.edu) Promoter V2.2 (See ScreenPromoter.cpp, lines 822-853).

The most up-to-date yeast chromosome lengths can be found in the Saccharomyces Genome Database (SGD):

ftp://genome-ftp.stanford.edu/pub/yeast/data_download/chromosomal_feature/chromosome_length.tab

1	NC_001133	230207
2	NC_001134	813138
3	NC_001135	316613
4	NC_001136	1531912
5	NC_001137	576869
6	NC_001138	270148
7	NC_001139	1090944
8	NC_001140	562639
9	NC_001141	439885
10	NC_001142	745445
11	NC_001143	666445
12	NC_001144	1078173
13	NC_001145	924430
14	NC_001146	784328
15	NC_001147	1091285
16	NC_001148	948061
17	NC_001224	85779

Here, chromosome 17 is the mitochondrial chromosome and can be ignored. The chromosome lengths here, and the differences in the *First_bps* values above should be the same, but are not (see reason below).

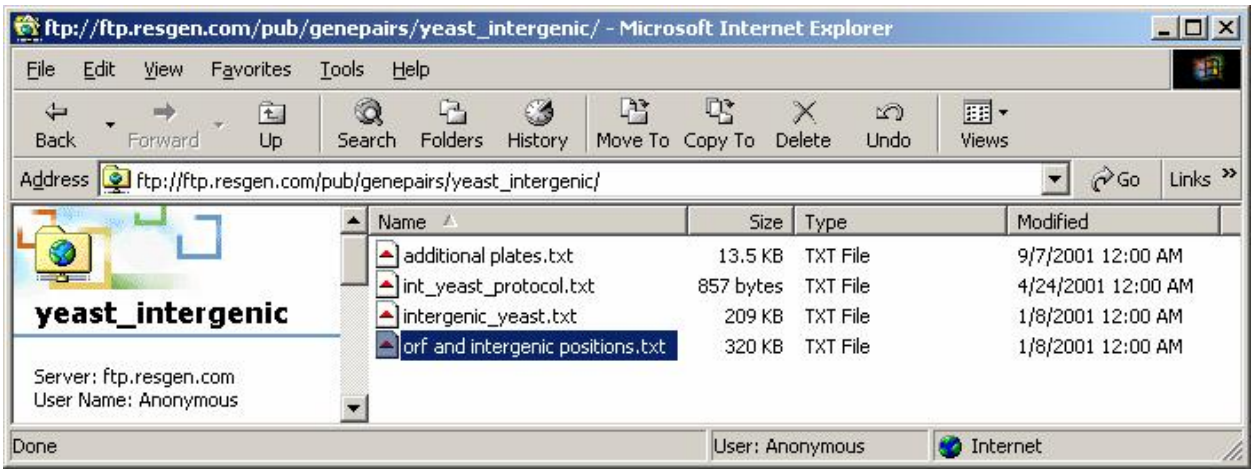
The *PlotLen* column of values gives the default maximum "X" value when the chromosome is plotted individually. This value is slightly larger than the actual chromosome length and is picked so that values along the X axis are somewhat rounded. For example, chromosome 3 is 1,358,685-1,043,346 = 315,339 bps long using the values in the First_bps column. This 315,339 bps chromosome is plotted with an "X" axis from 0 to 320,000 so intermediate X-axis markers are rounded numbers, e.g., 80,000, 160,000 and 240,000.

The CenLow and CenHigh columns give the coordinates of the location of the centromere for the given chromosome. The centromere range was from a 2 June 2003 E-mail from Jennifer Gerton.

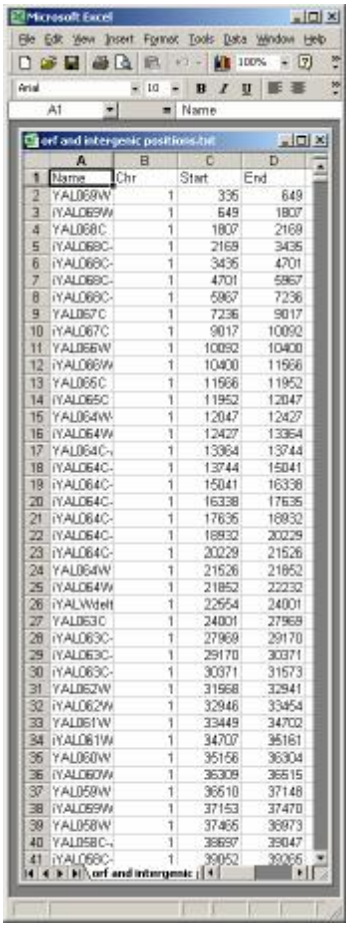
The most up-to-date centromere locations can be found by a "CEN" search on this page: http://db.yeastgenome.org/cgi-bin/SGD/search/featureSearch

Centromere Position Info
ChrI: coordinates 151457 to 151595
ChrII: coordinates 238168 to 238284
ChrIII: coordinates 114379 to 114495
ChrIV: coordinates 449707 to 449818
ChrV: coordinates 151960 to 152113
ChrVI: coordinates 148504 to 148622
ChrVII: coordinates 496921 to 497039
ChrVIII: coordinates 105579 to 105696
ChrIX: coordinates 355626 to 355743
ChrX: coordinates 436002 to 436118
ChrXI: coordinates 439774 to 439888
ChrXII: coordinates 150827 to 150946
ChrXIII: coordinates 268031 to 268141
ChrXIV: coordinates 628757 to 628866
ChrXV: coordinates 326584 to 326702
ChrXVI: coordinates 555952 to 556069

Instead of using the most recent data for SGD, the coordinates for the genes and intergenic regions are taken from ResGen data at http://www.resgen.com/products/YGP.php3.

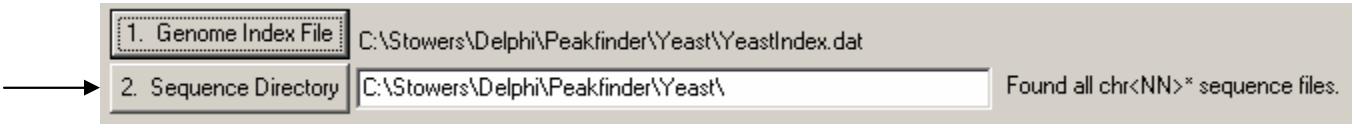


In particular, the file of interest is ftp://ftp.resgen.com/pub/genepairs/yeast_intergenic/org and intergenic positions.txt. The first few lines of this file are as follows:



2.2 Sequence Files

The directory containing the Genome Index File is assumed to have separate nucleotide sequence files for each chromosome in the genome, but this directory can be separately specified.

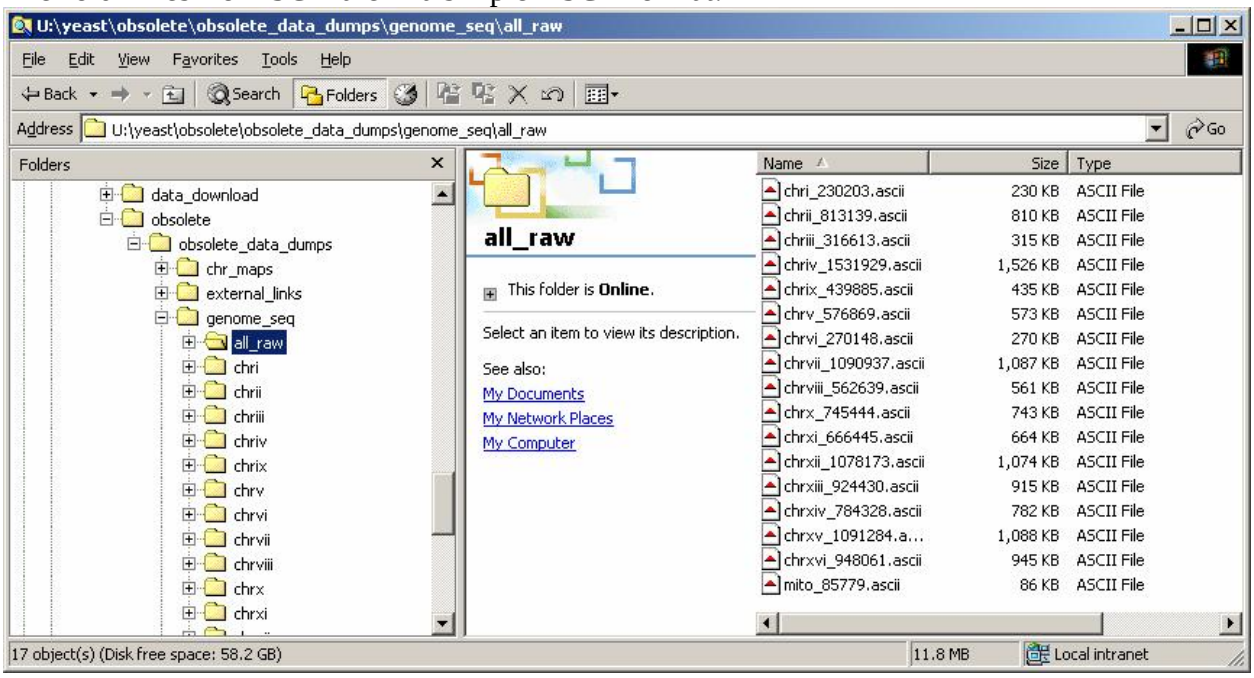


If desired, press button "2. Sequence Directory" and specify this directory. When this directory is specified, or assumed after pressing button 1, a search for the whole set of chromosome files is made. A message, such as the one shown above to the right, indicates if all the chromosome sequence files can be found.

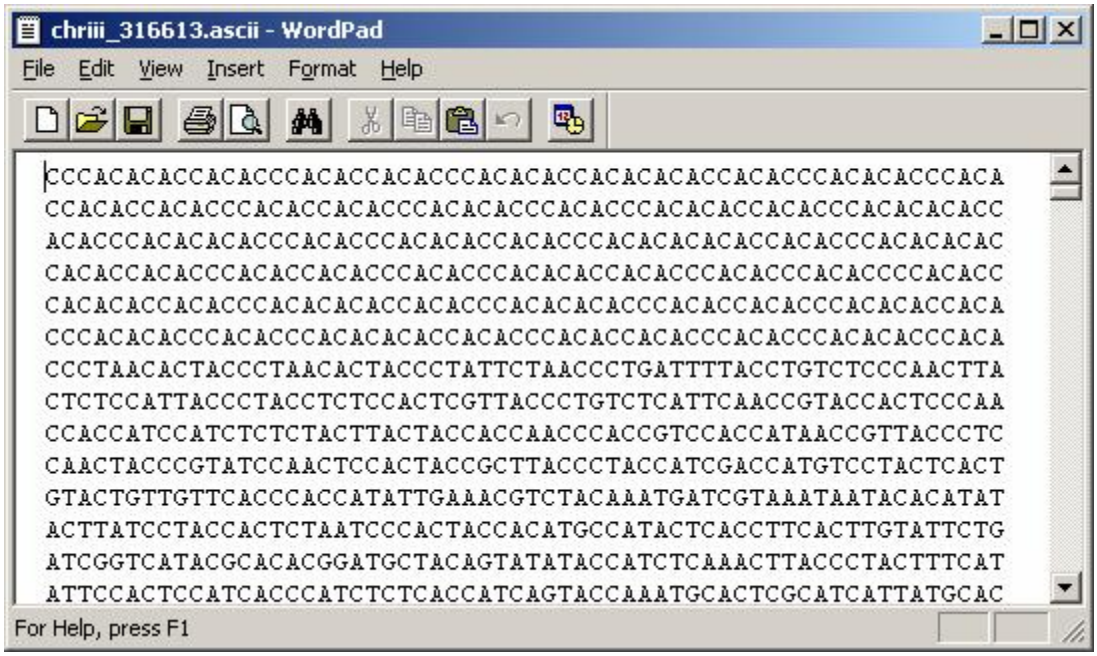
The chromosome sequence files are optional. When present, various nucleotide contents (e.g., AT content or GC content) can be displayed with the microarray ratio data.

There should be one file for each chromosome. The sequence data can be in the FASTA format, or simply an ASCII file. The names of the files must be chr<RomanNumeral>* or chrNN*. For example, chr05.fsa, or chrviii_562639.ascii. Either "old" or "new" file format from the *Saccharomyces* Genome Database (www.yeastgenome.org) is acceptable.

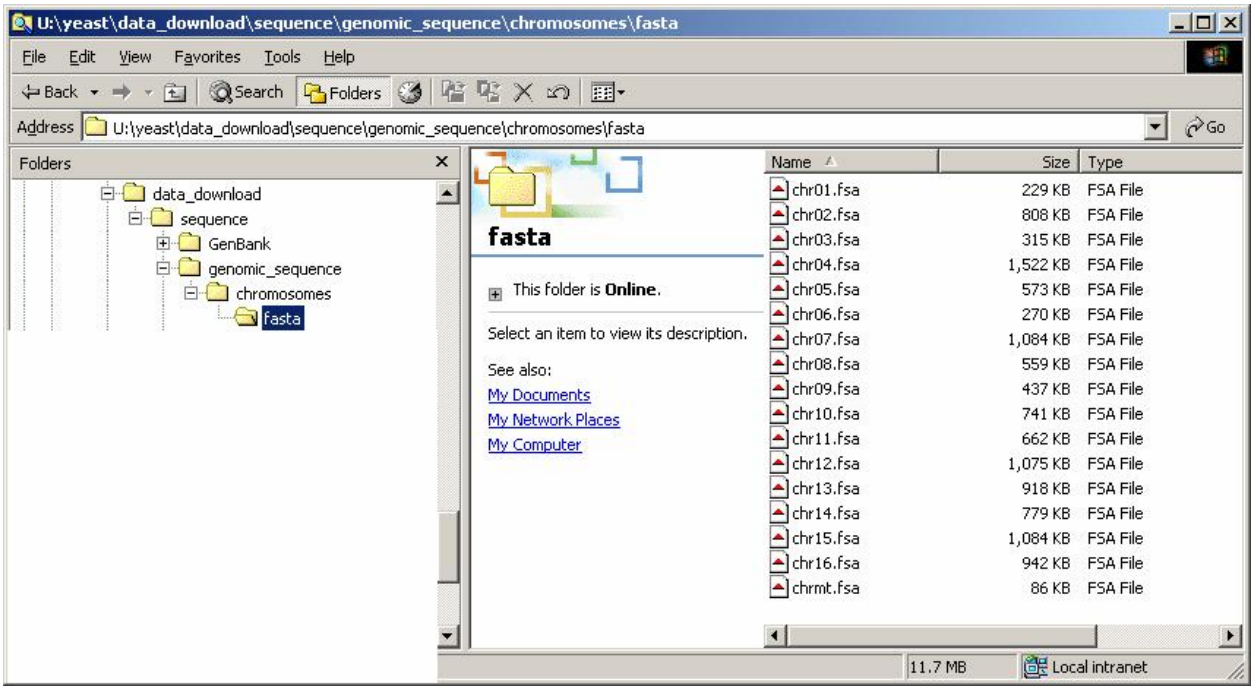
The "old" files from SGD are in a simple ASCII format:



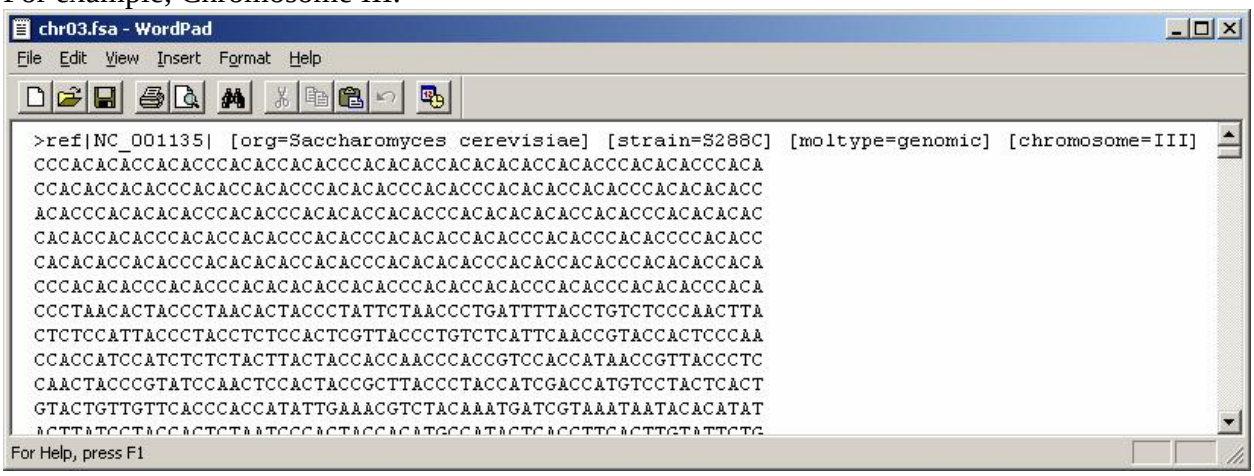
For example, chromosome III:



Newer files at SGD are in the FASTA format:

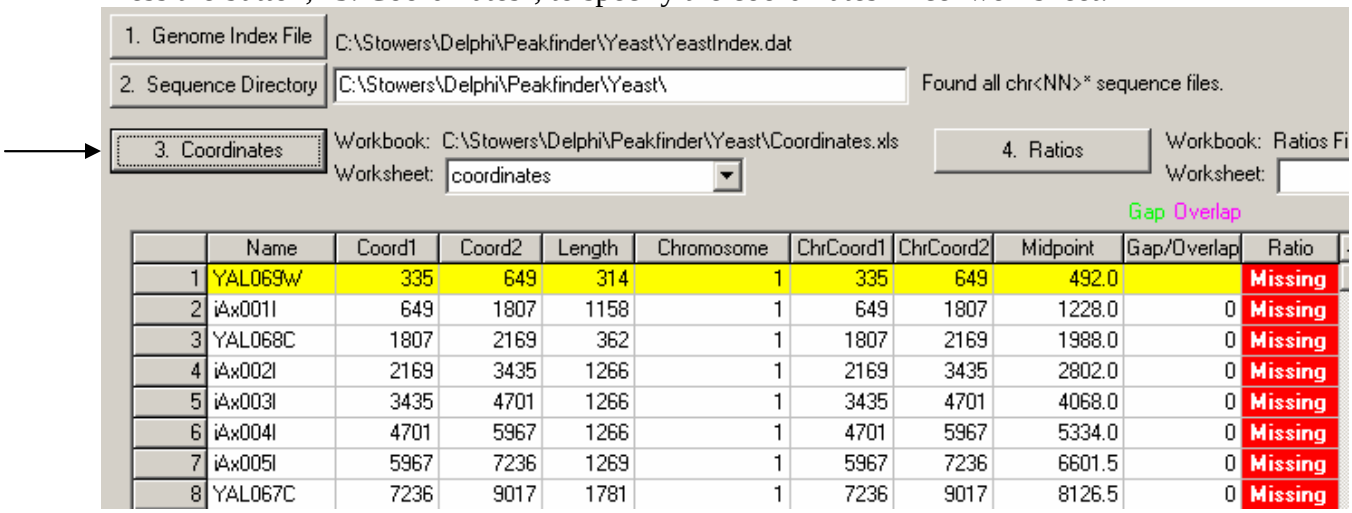


For example, Chromosome III:



2.3 Coordinates File

Press the button, "3. Coordinates", to specify the coordinates Excel worksheet:



	Name	Coord1	Coord2	Length	Chromosome	ChrCoord1	ChrCoord2	Midpoint	Gap/Overlap	Ratio
1	YAL069W	335	649	314	1	335	649	492.0		Missing
2	iAx001I	649	1807	1158	1	649	1807	1228.0	0	Missing
3	YAL068C	1807	2169	362	1	1807	2169	1988.0	0	Missing
4	iAx002I	2169	3435	1266	1	2169	3435	2802.0	0	Missing
5	iAx003I	3435	4701	1266	1	3435	4701	4068.0	0	Missing
6	iAx004I	4701	5967	1266	1	4701	5967	5334.0	0	Missing
7	iAx005I	5967	7236	1269	1	5967	7236	6601.5	0	Missing
8	YAL067C	7236	9017	1781	1	7236	9017	8126.5	0	Missing

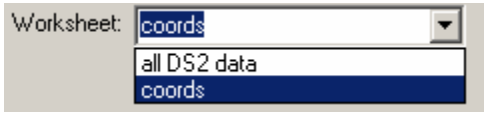
The coordinates file must be an Excel 2000 worksheet with at least these first three columns in the specified order (other columns are ignored) with these names as column headers in the first row:

1. Name
2. Coord1
3. Corod2

Coord1 and Coord2 are "genome coordinates" for each feature. The "chromosome coordinates" for the feature will be computed using index information from the Genome Index File.

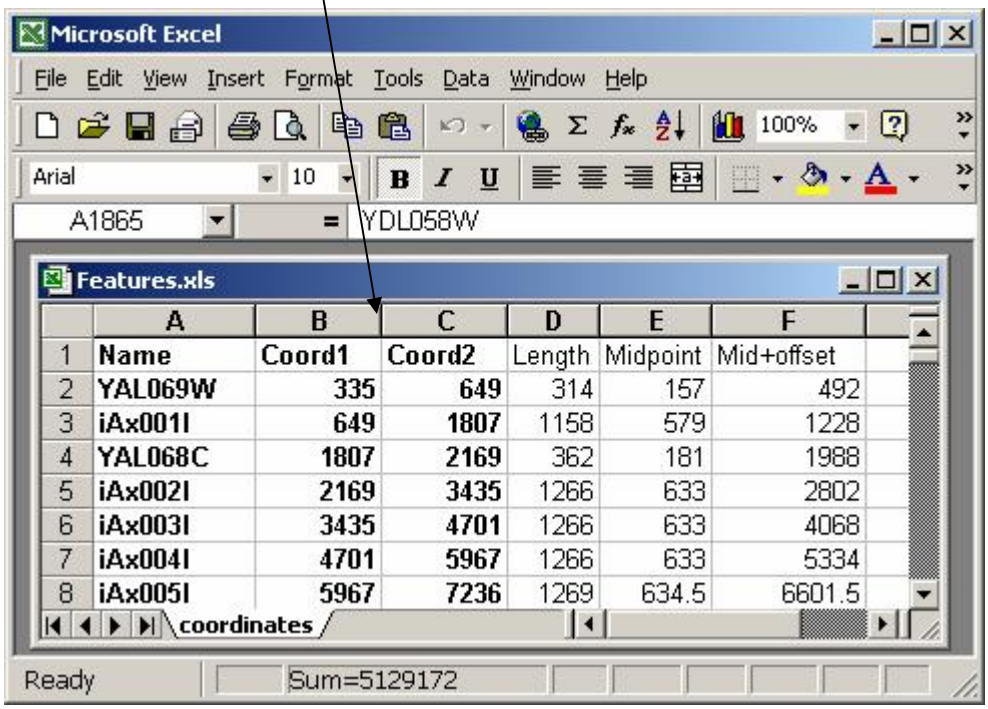
If the workbook has more than one worksheet, the ***first*** worksheet is automatically selected. If the headers for this worksheet are not correct (explained more below under PeakFinder.INI file), a message like this will be seen:

If a worksheet in the workbook other than the first one has the coordinate data (not recommended), select the other worksheet in the combobox:



In the following example, the first three columns of the worksheet are used, while the last three columns are ignored (but similar information is recomputed by the program):

Genome Coordinates



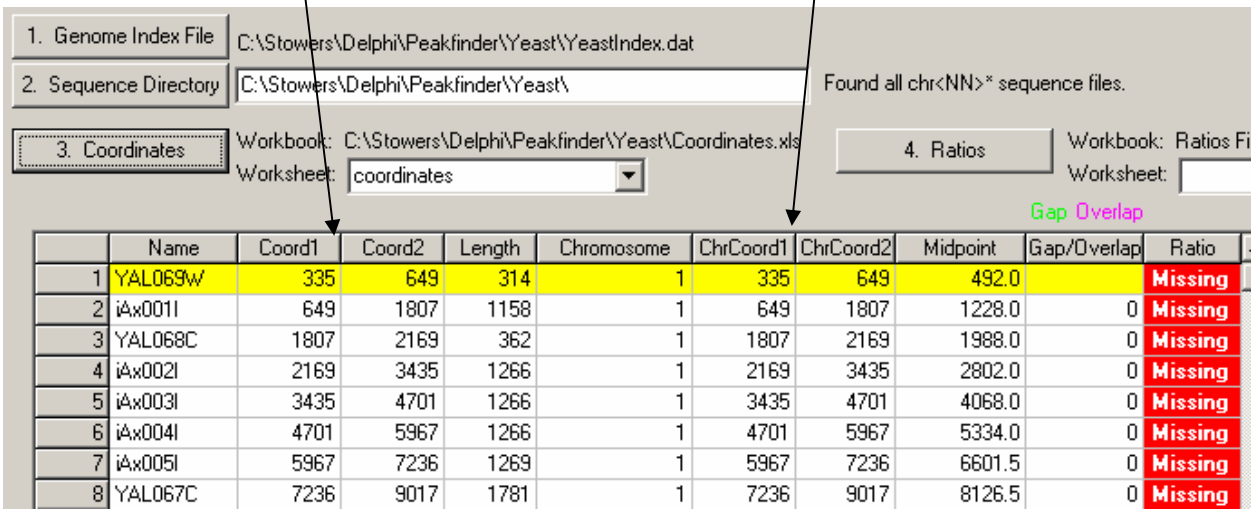
	A	B	C	D	E	F
1	Name	Coord1	Coord2	Length	Midpoint	Mid+offset
2	YAL069W	335	649	314	157	492
3	iAx001I	649	1807	1158	579	1228
4	YAL068C	1807	2169	362	181	1988
5	iAx002I	2169	3435	1266	633	2802
6	iAx003I	3435	4701	1266	633	4068
7	iAx004I	4701	5967	1266	633	5334
8	iAx005I	5967	7236	1269	634.5	6601.5

The first time the program is started (or when there is no PeakFinder.INI file – see below), or after the Rest Files button is pressed, the program will appear with a blank coordinate StringGrid.

When a coordinates worksheet is initially loaded, the "Ratio" column will be marked "Missing" as shown below, until it is loaded separately:

Genome coordinates

Computed Chromosome Coordinates



	Name	Coord1	Coord2	Length	Chromosome	ChrCoord1	ChrCoord2	Midpoint	Gap/Overlap	Ratio
1	YAL069W	335	649	314	1	335	649	492.0		Missing
2	iAx001I	649	1807	1158	1	649	1807	1228.0	0	Missing
3	YAL068C	1807	2169	362	1	1807	2169	1988.0	0	Missing
4	iAx002I	2169	3435	1266	1	2169	3435	2802.0	0	Missing
5	iAx003I	3435	4701	1266	1	3435	4701	4068.0	0	Missing
6	iAx004I	4701	5967	1266	1	4701	5967	5334.0	0	Missing
7	iAx005I	5967	7236	1269	1	5967	7236	6601.5	0	Missing
8	YAL067C	7236	9017	1781	1	7236	9017	8126.5	0	Missing

A sorted, in-memory index is made of the names in the coordinates files so the names can be located quickly when the ratio file is loaded. This index file does not allow duplicate entries. A message like the following will be seen if any duplicate names occur:

These duplicate entries should be investigated and manually deleted from the coordinates file. For example, the above message was caused by duplicate rows. The line 1864 refers to the StringGridCoord object in the program:

	Name	Coord1	Coord2	Length	Chromosome	ChrCoord1	ChrCoord2	Midpoint	Gap/Overlap
1862	iDx190I	1703636	1704348	712	4	344951	345663	345307.0	-4
1863	YDL058W	1704352	1709724	5372	4	345667	351039	348353.0	4
1864	YDL058W	1704352	1709724	5372	4	345667	351039	348353.0	-5372

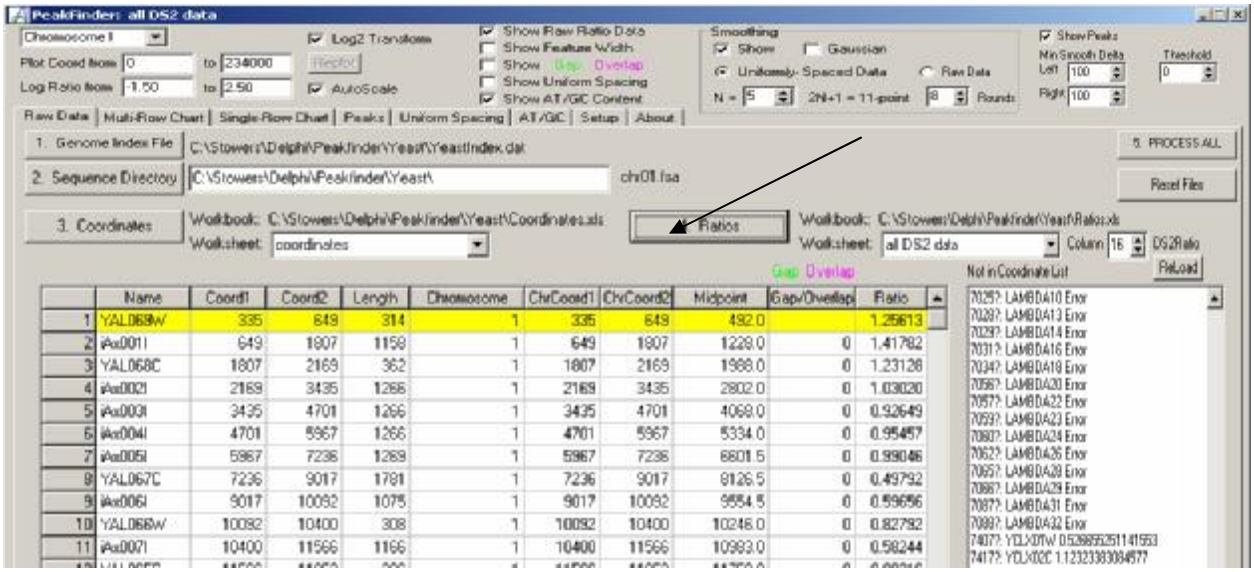
Unfortunately, line 1864 in this TStringGrid corresponds to line 1865 in the original spreadsheet:

	A	B	C	D	E	F
1863	**iDx190I**	**1703636**	**1704348**	712	356	1703992
1864	**YDL058W**	**1704352**	**1709724**	5372	2686	1707038
1865	**YDL058W**	**1704352**	**1709724**	5372	2686	1707038
1866	**iDx191I**	**1709720**	**1710117**	397	198.5	1709918.5

Every time a Coordinates File is loaded, a new file of ratios data must be loaded. This can be as simple as reselecting the worksheet name if a worksheet is already in memory.

2.4 Ratios File

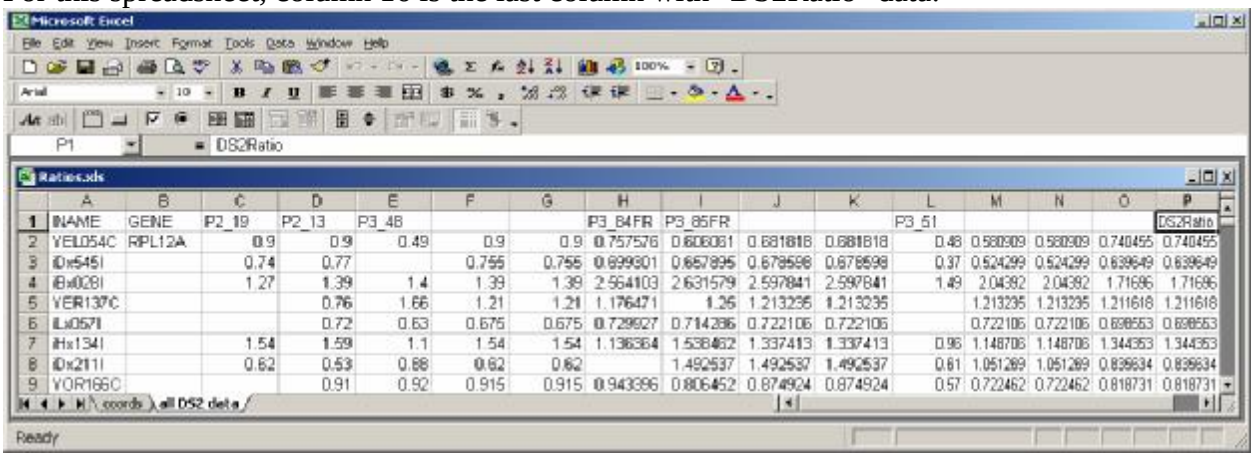
Press the "4. *Ratios"* button on the *Raw Data* tabsheet to load an Excel Workbook with ratio data.



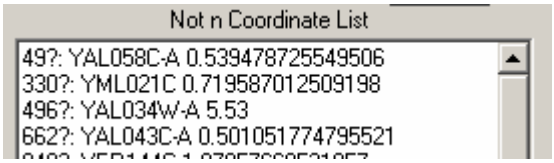
The ***last*** worksheet in the workbook is automatically selected. The means the Coordinates data are selected ***first***, and the Ratio data are selected ***last*** if both are present in the same Excel spreadsheet.

If the correct header row is present in column 1 ("IName" by default), the last column of that sheet is automatically loaded as the ratio data. Optionally, use the Spinbox to select another column by number, and then press the ReLoad button to load a different column than the last one. The column name from row 1 is shown (or "ColumnNN", if the original name is blank).

For this spreadsheet, column 16 is the last column with "DS2Ratio" data:



Features specified in the ratio file that cannot be found in the coordinates file are listed:

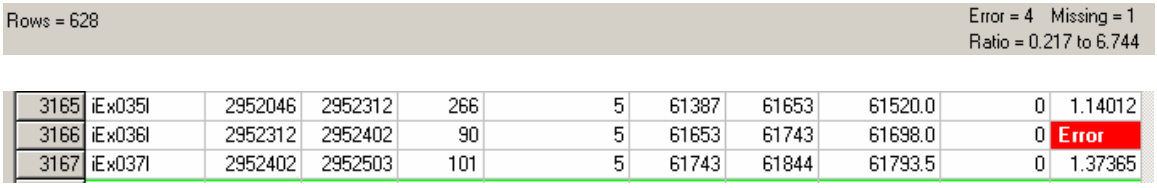


As discussed below, some features are not assigned a value and are "Missing", or some sort of conversion problem can exist while reading the spreadsheet and these ratios are tagged as an "Error".

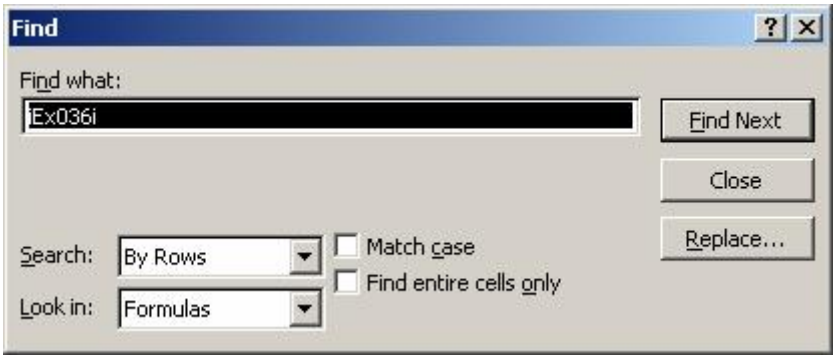
Chromosome I is automatically selected with ratio data is loaded, but the combobox at the upper left can be used to select either

or any single chromosome, e.g.,

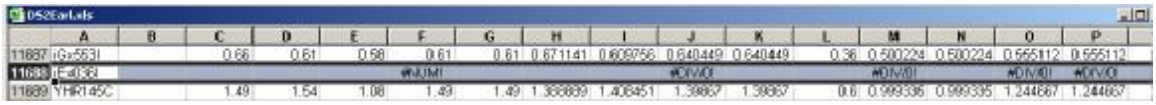
After a selection is made the StringGrid is adjusted to only show rows for the specified chromosome. Information at the bottom of the StringGrid identifies how many features are present, as well as how many conversion errors and missing data points were found.



This "Error" occurred with feature iEx036I from the coordinates file. To find the cause of this error, we select the INAME column | Edit | Find:



Find Next



So this error was caused by calculations on missing data in the original data file.

The ratios that are "Missing" for certain features should be reviewed:

3469	iEx207I	3212455	3213295	840	5	321796	322636	322216.0	-49	Missing

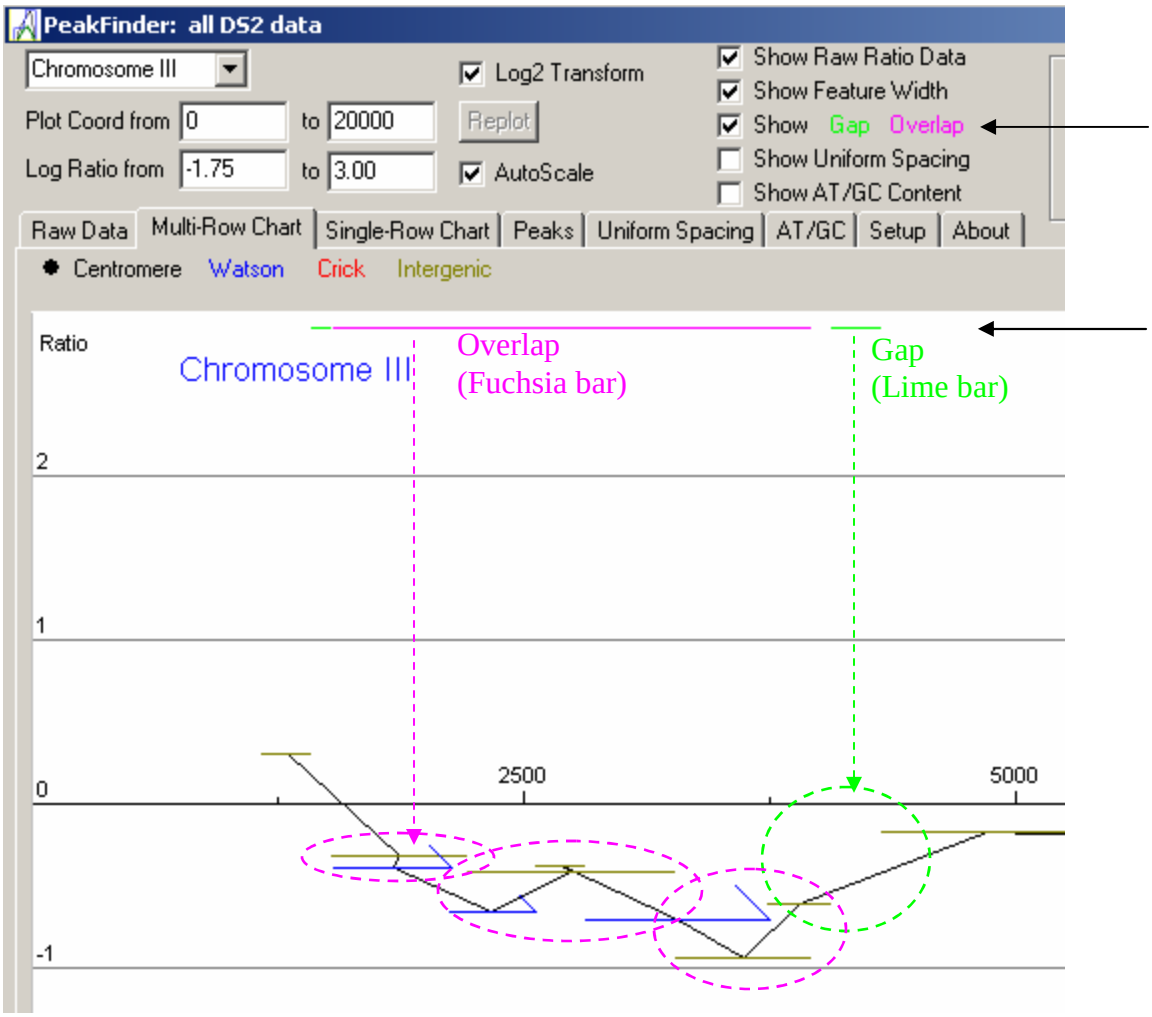
The "Missing" and "Error" points are ignored in any of the peak finding analysis.

Gaps / Overlaps. With adjacent features the *Coord2* of one feature is the same as *Coord1* of the next feature. For example, ORF YAL069W (below) has coordinates from 335 to 649, which is next to the intergenic feature iAx001I, which has coordinates from 649 to 1807. The first seven features of chromosome I are all adjacent.

	Name	Coord1	Coord2	Length	Chromosome	ChrCoord1	ChrCoord2	Midpoint	Gap/Overlap	Ratio
1	YAL069W	335	649	314	1	335	649	492.0		1.25613
2	iAx001I	649	1807	1158	1	649	1807	1228.0	0	1.41782
3	YAL068C	1807	2169	362	1	1807	2169	1988.0	0	1.23128
4	iAx002I	2169	3435	1266	1	2169	3435	2802.0	0	1.03020
5	iAx003I	3435	4701	1266	1	3435	4701	4068.0	0	0.92649
6	iAx004I	4701	5967	1266	1	4701	5967	5334.0	0	0.95457
7	iAx005I	5967	7236	1269	1	5967	7236	6601.5	0	0.99046

These adjacent features can be seen graphically using the Show Feature Width/Orientation option:

Some features overlap, and in other cases, there are gaps between features. The following shows both several overlaps and a gap in the coverage for Chromosome III:



Gaps and Overlaps are indicated by the colored lines in the StringGrid:

	Name	Coord1	Coord2	Length	Chromosome	ChrCoord1	ChrCoord2	Midpoint	Gap/Overlap	Ratio
1092	iCx001I	1044517	1044767	250	3	1171	1421	1296.0	2875	1.24187
1093	iCx002I	1044867	1045564	697	3	1521	2218	1869.5	100	0.80306
1094	YCL076W	1044884	1045474	590	3	1538	2128	1833.0	-680	0.76815
1095	YCL075W	1045465	1045905	440	3	2119	2559	2339.0	-9	0.63808
1096	iCx003I	1045564	1046611	1047	3	2218	3265	2741.5	-341	0.75558
1097	iCx004I	1045903	1046161	258	3	2557	2815	2686.0	-708	0.77239
1098	YCL074W	1046163	1047089	926	3	2817	3743	3280.0	2	0.61624
1099	iCx005I	1046611	1047309	698	3	3265	3963	3614.0	-478	0.52296
1100	iCx006I	1047087	1047410	323	3	3741	4064	3902.5	-222	0.65475
1101	iCx007I	1047659	1048737	1078	3	4313	5391	4852.0	249	0.88765
1102	iCx008I	1048737	1049815	1078	3	5391	6469	5930.0	0	0.86103

2.4.1 Gaps

1100	iCx006I	1047087	1047410	323	3	3741	4064	3902.5	-222	0.65475
1101	iCx007I	1047659	1048737	1078	3	4313	5391	4852.0	249	0.88765

The green line indicates a gap exists between the iCx006I and iCX007i features. The 249 in the gap column indicates the size of the feature.

Gap Math: 1047410 + 249 = 1047659

2.4.2 Overlaps

1093	iCx002I	1044867	1045564	697	3	1521	2218	1869.5	100
1094	YCL076W	1044884	1045474	590	3	1538	2128	1833.0	-680

The fuchsia line indicates the feature YCL076W overlaps with the previous feature.

Overlap Math: 1045564 – 680 = 1044884. Note the "680" is the difference between the end point and the start point. In this case the overlap itself is only 590 since that is the length of the shorter YCL076W feature.

The "Gap" between the first feature of one chromosome and the last feature of the previous chromosome is meaningless and should be ignored.

2.5 PeakFinder.INI File

An "INI" (pronounced "any") file contains configuration information that can be retained between runs.

The most recently used coordinate and ratio worksheets will be loaded automatically when the program is started. In a Windows NT or later Windows environment for user ID "efg", this file is store in this directory:

C:\Documents and Settings\efg\Local Settings\Application Data\StowersInstitute\PeakFinder

The last specified Coordinates and Ratios Workbooks are stored in the PeakFinder.INI file:

```
[Setup]
GenomeIndex=C:\Stowers\Delphi\Peakfinder\Yeast\YeastIndex.dat
CoordinatesWorkbook=C:\Stowers\Delphi\Peakfinder\Yeast\Coordinates.xls
RatioWorkbook=C:\Stowers\Delphi\Peakfinder\Yeast\Ratios.xls
SequenceDirectory=C:\Stowers\Delphi\Peakfinder\Yeast\
```

Unfortunately, the worksheet is not saved, so at present it's best to store coordinates data as the first worksheet of a workbook, or as a separate workbook with only a single worksheet.

Other column header names can be used for the coordinates data file if specified in a PeakFinder.INI file:

```
PeaksFinder.INI

[Setup]
CoordName=Name
Coord1=Coordx
Coord2=Coord2
RatioName=IName
```

Comparison of header names is NOT case sensitive.

If the names do not match, the first row will show what is expected while the second row shows what was found in the file.

Here the columns are OK in the Excel worksheet, but the names are wrong in the INI file:

Name	Coordx	Coord2
Name	Coord1	Coord2

Here there is an attempt to use the wrong worksheet for coordinates data:

Name	Coord1	Coord2
INAME	GENE	P2_19

When the column headers do no match what is expected, no coordinate data is loaded.

3 Data Samples

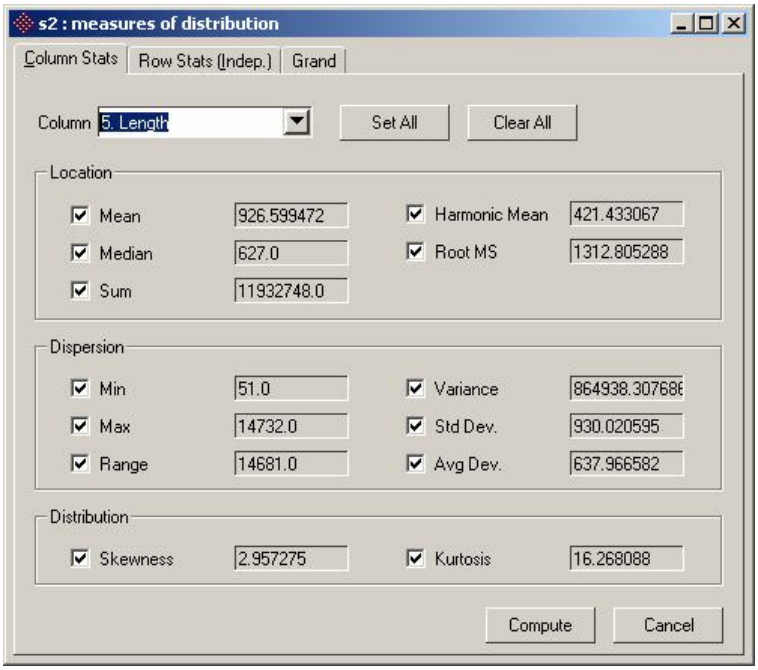
3.1 Raw Data

The "raw data" from the yeast cohesin microarray experiments can be displayed as a series of (X,Y) points, usually (Feature Midpoint, Log_2 Ratio) points, connecting the midpoints of the features.

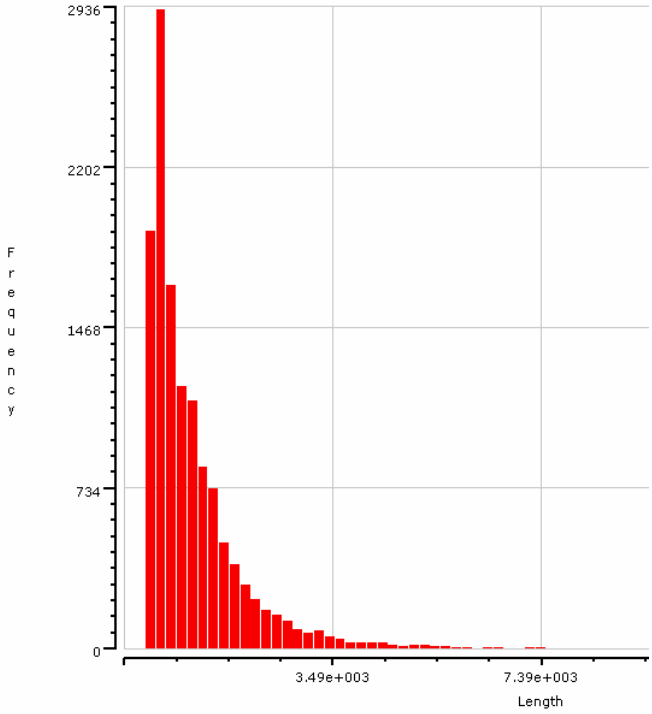
Feature	Midpoint (X)	Ratio	Log_2 Ratio (Y)
YAL069W	492.0	1.25613	0.32899
iAx001I	1228.0	1.41782	0.50367
YAL068C	1988.0	1.23128	0.30016
iAx002I	2802.0	1.03020	0.04292
iAx003I	4068.0	0.92649	-0.11015
iAx004I	5334.0	0.95457	-0.06708
iAx005I	6601.5	0.99046	-0.01383

3.2 Feature Length Statistics

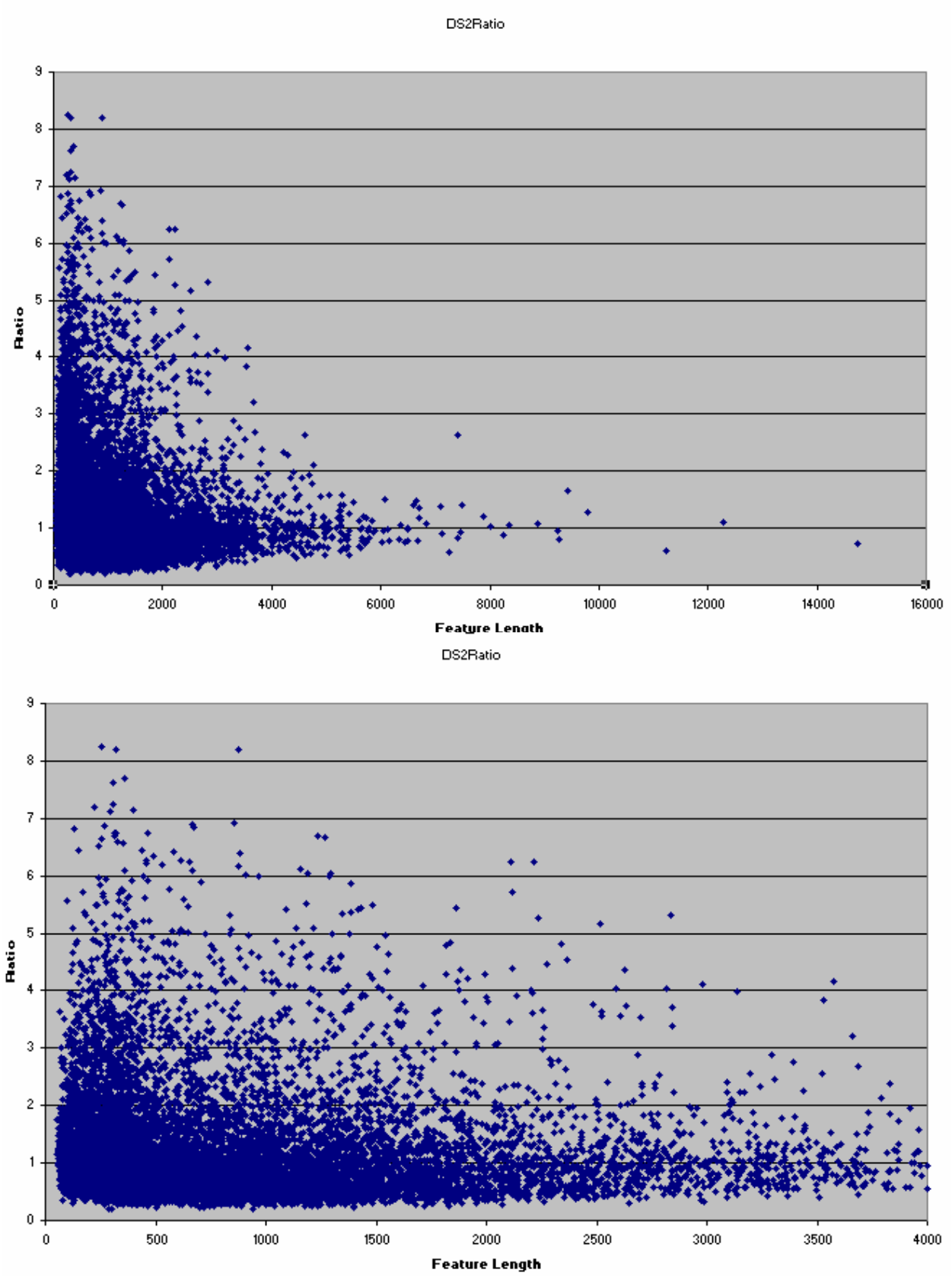
The statistics on the length features are useful to consider as part of the rationale in using uniformly spaced data. Consider these feature-length statistics created using Partek Pro:



Based on the 12878 features for all 16 yeast chromosomes:



Histogram of Feature Lengths
[Sorry for the strange scale shown by Partek Pro]



Scatterplot of Ratios by Feature Length (no clear pattern was expected / observed)

3.3 Uniformly Spaced Data

How to best treat data when features have a variable width, and gaps and overlaps can occur, is still a question open for additional discussion. Instead of the "raw" data points, uniformly spaced data is an alternative representation useful in smoothing the data and in other analyses, such as Fourier analysis. Tools and methods for analyzing unevenly spaced data are not common.

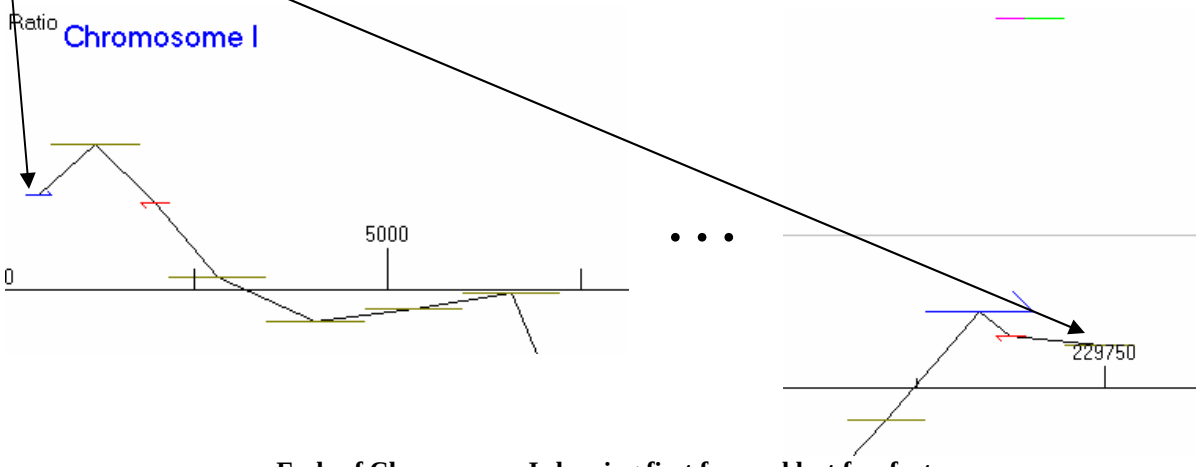
Some may object to creating "new" data points from the original data. The intent of creating uniformly spaced data is only to aid in the analysis of the original data. Once the analysis is complete, the results are applied to the original, non-uniformly spaced data. Final results are only reported in the context of the original data.

The choice of spacing size for uniformly spaced data is a balance between the length of the features in the experiment and the size that might have some biological significance. The default spacing now is 100 bps, but any other value can be selected. This spacing may miss a few of the very small features (the smallest feature is only 51 bps), but allows for 10 points per 1 kb for analysis purposes.

3.3.1 Features to Uniformly Space Data

To understand the mechanics of how uniformly spaced data is created from the original raw data points, let's consider the first and last features for Yeast Chromosome I from one of the datasets (the gap and overlap will be discussed later):

Feature	Coord1	Coord2	Length	Midpoint (X)	Ratio (Y)	Log_2(Ratio)
YAL069W	335	649	314	492	1.25613	0.32899
IAx134I	229310	230108	798	229709	1.21334	0.27898



Ends of Chromosome I showing first few and last few features
For convenience the line connecting the midpoints of features is plotted.

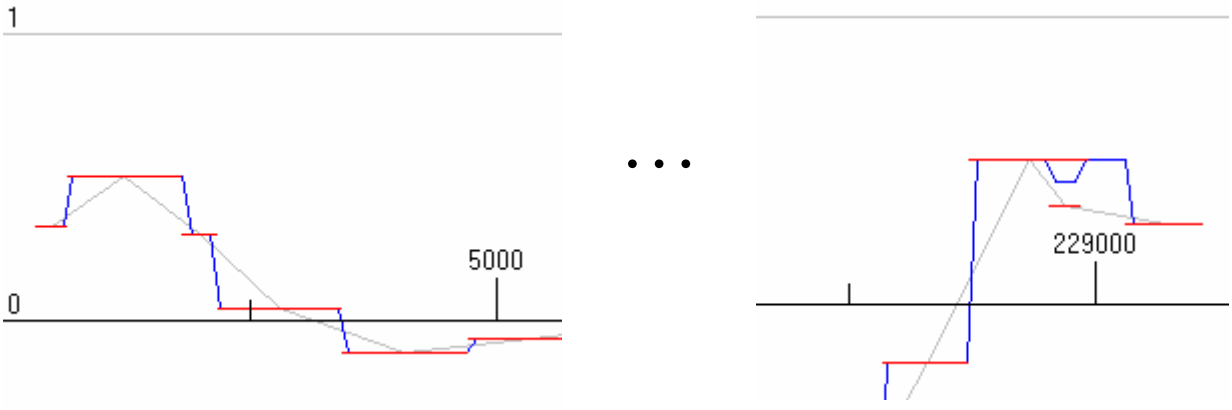
With 100 bps spacing, points may be created at bps index values of 0, 100, 200, 300, etc.

Since the first feature has coordinates 335 to 649, data points with a Ratio=1.25613 will be created at 400, 500 and 600 bps. Likewise, since the last feature has coordinates 229310 to 230108, data points with a Ratio=1.21334 will be created at 229400, 229500, 229600, 229700,

229800, 229900, 230000, and 230100. The following table summarizes the 100 bps uniformly spaced data:

Feature	X	Ratio (Y)	Log_2(Ratio)
YAL069W	400	1.25613	0.32899
	500	1.25613	0.32899
	600	1.25613	0.32899
IAx134I	229400	1.21334	0.27898
	229500	1.21334	0.27898
	229600	1.21334	0.27898
	229700	1.21334	0.27898
	229800	1.21334	0.27898
	229900	1.21334	0.27898
	230000	1.21334	0.27898
	231000	1.21334	0.27898

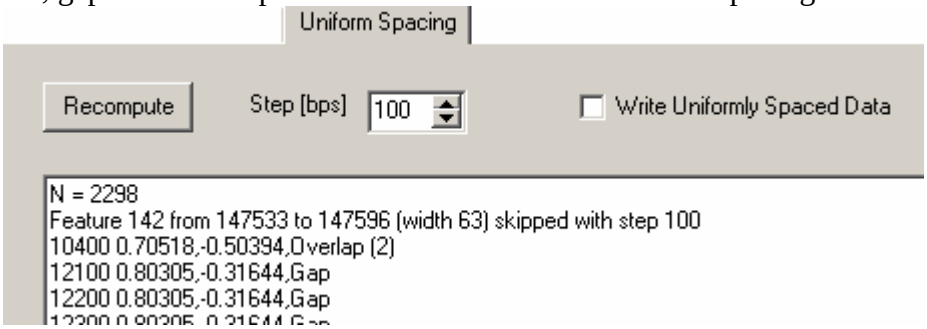
The plotted uniformly spaced data has a stair-step appearance (the red lines show the actual length of the features), which better reflects the true width of the various features instead of only plotting the midpoint of the feature:



Ends of Chromosome I showing first few and last few features
Represented by uniformly spaced data

3.3.2 Overlapping features and Gaps

Missed features, gaps and overlaps are summarized on the Uniform Spacing tabsheet:



Note the overlap in the 2nd and 3rd to last features (above and to the left of the number 229000 in the graph). For now, the average of all features is taken at each point. This is implemented in the PeakFinder program by counting the number of "hits" at each point and summing the values. Once all features are processed, those with more than one hit are averaged. Points without any "hits" are in gap areas. To prevent introducing any additional high-frequency noise when a gap occurs, a gap is assigned the last "Y" value, and in effect, extends any existing plateau area.

4 Smoothing

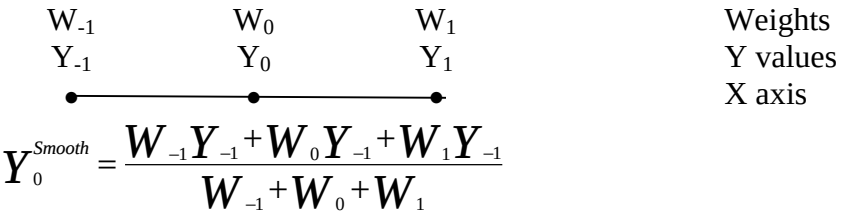
Finding peaks in the original raw data is a bit difficult because of the numerous spurious peaks caused by "noise". Smoothing algorithms applied to the original raw data do not properly account for the length of the various features. Smoothing of the uniformly spaced data seems to be a better approach.

Data smoothing helps eliminate small peaks caused by the noise. Our attempts to define a statistical basis for deciding "noise" from "signal" thus far did not give an adequate solution (a wavelet denoising technique was investigated and deserves further analysis). Lacking a rigorous statistical theory, we opted for the ability to visually and interactively decide how much to smooth the data.

4.1 Moving Average (Equal Weights)

A moving average is a simple "low pass" filter (lower-frequencies are "passed" but higher frequencies are not)

A weighted moving average of 2N+1 points can be used to smooth a data series. For example, for N=1, three points are used in smoothing:



$$Y_0^{Smooth} = \frac{W_{-1}Y_{-1} + W_0 Y_{-1} + W_1 Y_{-1}}{W_{-1} + W_0 + W_1}$$

In the simplest case, all the weights are 1.

$$Y_0^{Smooth} = \frac{Y_{-1} + Y_{-1} + Y_{-1}}{3}$$

A filter can be applied repeatedly for additional smoothing, which is equivalent to just a different weighted average scheme.

Consider what happens if the 3-point moving average is used with all weights = 1 and is applied three times:

X	1	2	3	4	5	6	7
Y	a	b	c	d	e	f	g
Moving Average Pass 1		$\frac{a+b+c}{3}$	$\frac{b+c+d}{3}$	$\frac{c+d+e}{3}$	$\frac{d+e+f}{3}$	$\frac{e+f+g}{3}$	
Moving Average Pass 2			$\frac{a+2b+3c+2d+e}{9}$	$\frac{b+2c+3d+2e+f}{9}$	$\frac{c+2d+3e+2f+g}{9}$		
Moving Average Pass 3	$\frac{a+3b+6c+7d+6e+3f+g}{27}$						

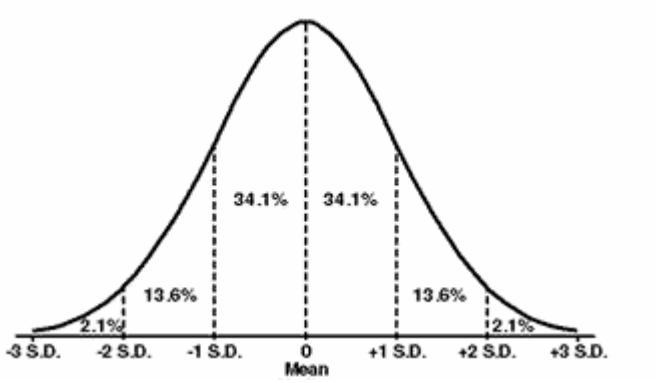
So applying a 3-point moving average (1 1 1)/3 three times is the same as applying a 7-point moving average (1 3 6 7 6 3 1)/27 a single time.

The boundary conditions were ignored above. One approach is to use as many points as possible in a "smaller" moving average. The treatment of boundary points is considered insignificant in finding peaks.

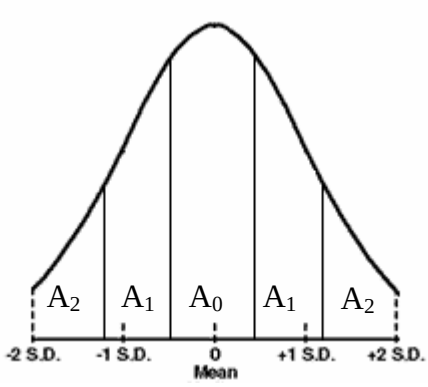
4.2 Moving Average (Gaussian Weights)

Instead of using weights of unity in a moving average, we can weight the nearer neighbors more and neighbors farther away less. This will retain the "peaks" a bit better/longer after smoothing.

Let's start with the normal (Gaussian) curve:



Let's approximate the whole area under the curve by the area from -2σ to +2σ (this ignores less than 5% of the area) and divide this area into 2N+1 intervals.



For N=2, this would be five intervals, as shown above. The areas under the curve can be computed[1]:

	Raw Weights
A_0	0.31084
A_1	0.22951
A_2	0.09232

The total area under this curve is $A = A_0 + 2A_1 + 2A_2$. Let's normalize the weights by dividing by this sum:

	Normalized Weights
A0	0.32566
A1	0.24045
A2	0.09672

The full set of weights used in computing the Gaussian moving average:

Gaussian Weights for Moving Average (N=2)
0.09672
0.24045
0.32566
0.24045
0.09672

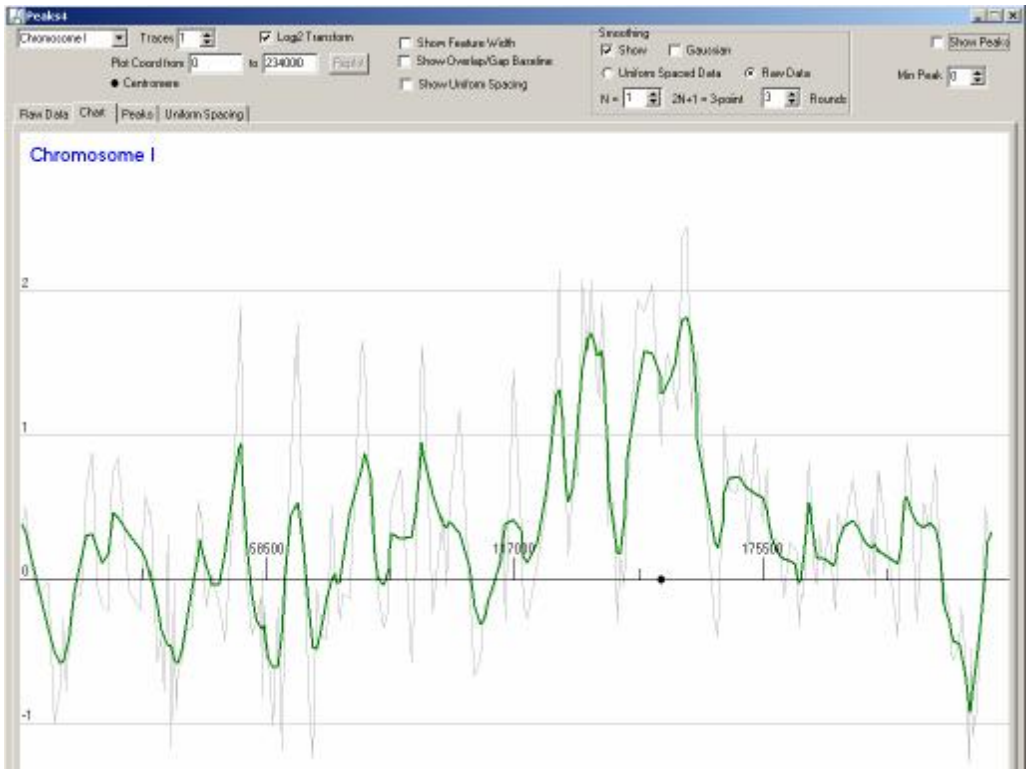
Let's compare the Gaussian Weights for N=3 with the result of three rounds of simple moving averages. Recall that applying a 3-point moving average (1 1 1)/3 three times is the same as applying a 7-point moving average (1 3 6 7 6 3 1)/27 a single time. Here are the Gaussian weights computed for N-3:

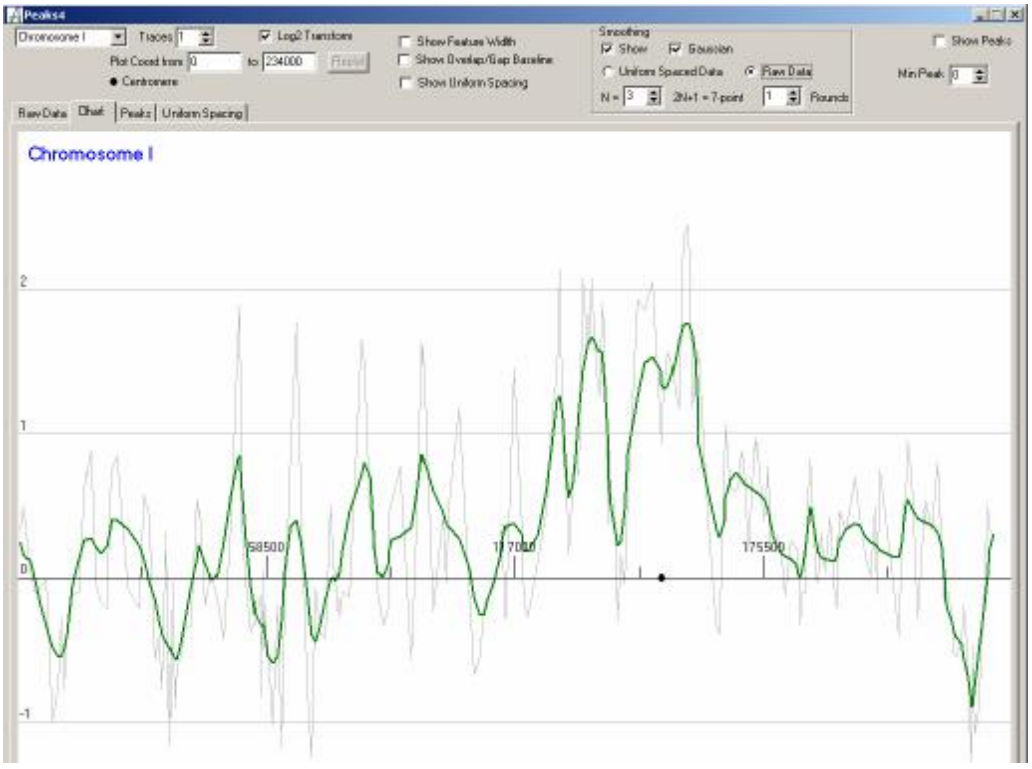
[1] PeakFinder uses the TPMath Library for this computation:

J. Debord (2003). TPMath, Mathematical library in Pascal. Available at http://www.unilim.fr/pages_perso/jean.debord/tpmath/tpmath.htm

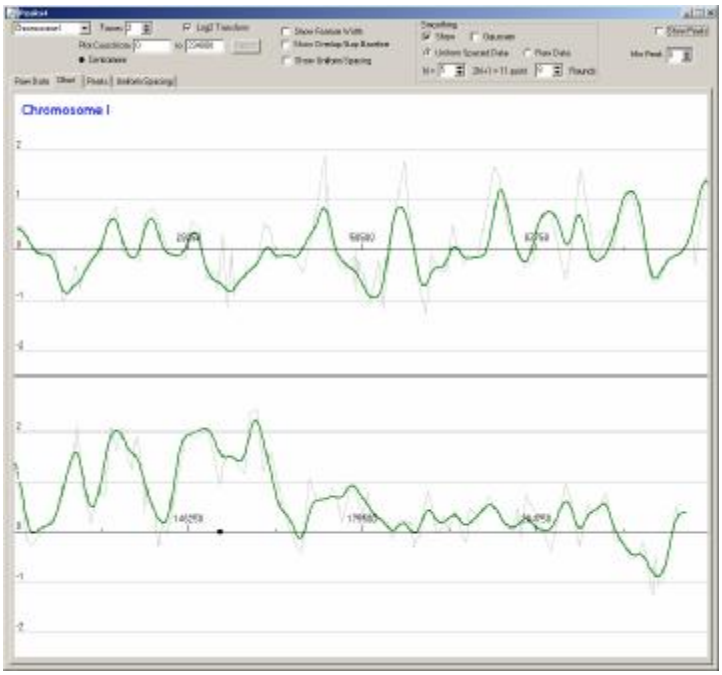
Index	Gaussian Weights for Moving Average (N=3)	Weights for three Rounds of 3-point moving averages
-3	0.05638	1/27 = 0.03704
-2	0.12480	3/27 = 0.11111
-1	0.20101	6/27 = 0.22222
0	0.23562	7/27 = 0.25926
1	0.20101	6/26 = 0.22222
2	0.12480	3/27 = 0.11111
3	0.05638	1/27 = 0.03704

The following shows that one round of 7-point Gaussian smoothing is nearly identical to three rounds of 3-point moving averages:

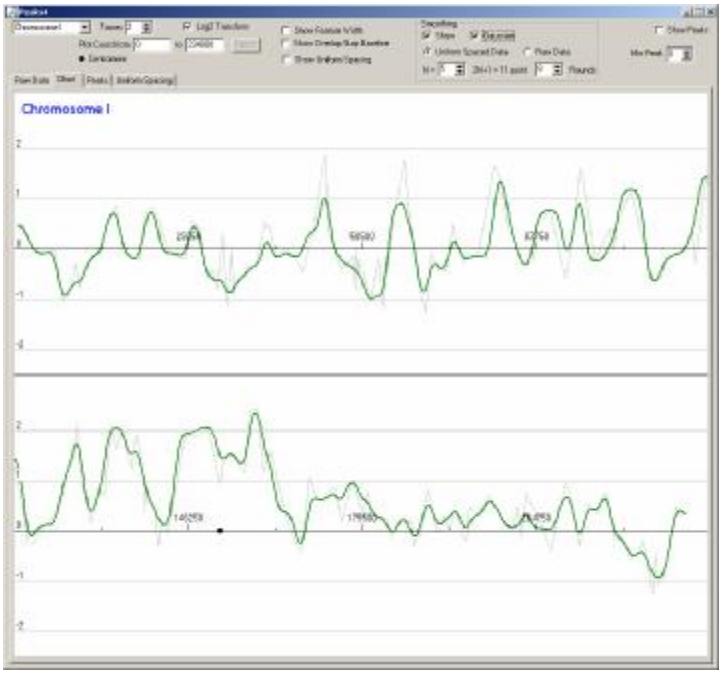

Peaks4
Chromosome I
Traces 1
Log2 Transform
Plot Coord from 0 to 234000
Replot
Centromere
Show Feature Width
Show Overlap/Gap Baseline
Show Uniform Spacing
Smoothing
Show
Gaussian
Uniform Spaced Data
Raw Data
N = 1
2N+1 = 3-point
3
Rounds
Show Peaks
Min Peak 0
Raw Data
Chart
Peaks
Uniform Spacing
Chromosome I
2
1
0
-1
58500
117000
175500


Peaks4
Chromosome I
Traces 1
Log2 Transform
Plot Coord from 0 to 234000
Replot
Centromere
Show Feature Width
Show Overlap/Gap Baseline
Show Uniform Spacing
Smoothing
Show
Gaussian
Uniform Spaced Data
Raw Data
N = 3
2N+1 = 7-point
1
Rounds
Show Peaks
Min Peak 0
Raw Data
Chart
Peaks
Uniform Spacing
Chromosome I
2
1
0
-1
58500
117000
175500

The uniformly spaced datasets have considerably more data points than the original raw data. The intent in the original paper was to smooth this data with a 1 KB moving average, which would be 11 points with 100 bps uniform sampling. Empirically, eight rounds of 11-point smoothing works well with the uniformly spaced datasets. Gaussian smoothing only helps a little.



"Regular" Smoothing



"Gaussian Smoothing"

5 Peak Finding

Once the data are sufficiently smoothed, peaks can be found when the first derivative (slope of the curve) is zero, i.e.,

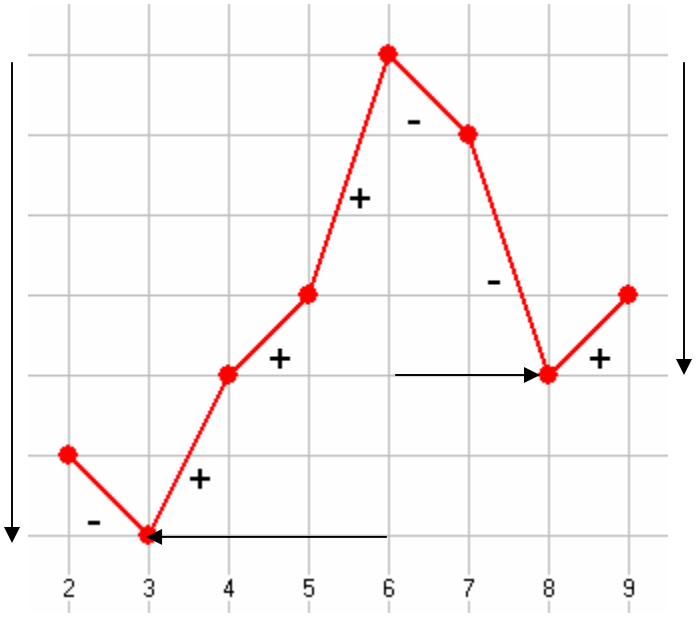
$$\frac{dy}{dx} = 0$$

The derivative can be approximated by computing the changes in x and y:

$$\frac{\Delta y}{\Delta x} = 0$$

With the uniformly spaced data, Δx is a constant, or even with the unevenly-spaced raw data, we can ignore Δx, and just consider the Δy value. We can find a zero crossing of the derivative by watching where its sign changes form "+" to "-".

In the diagram below, the "+" and "-" symbols show whether a segment has a Δy that is "increasing" or "decreasing".



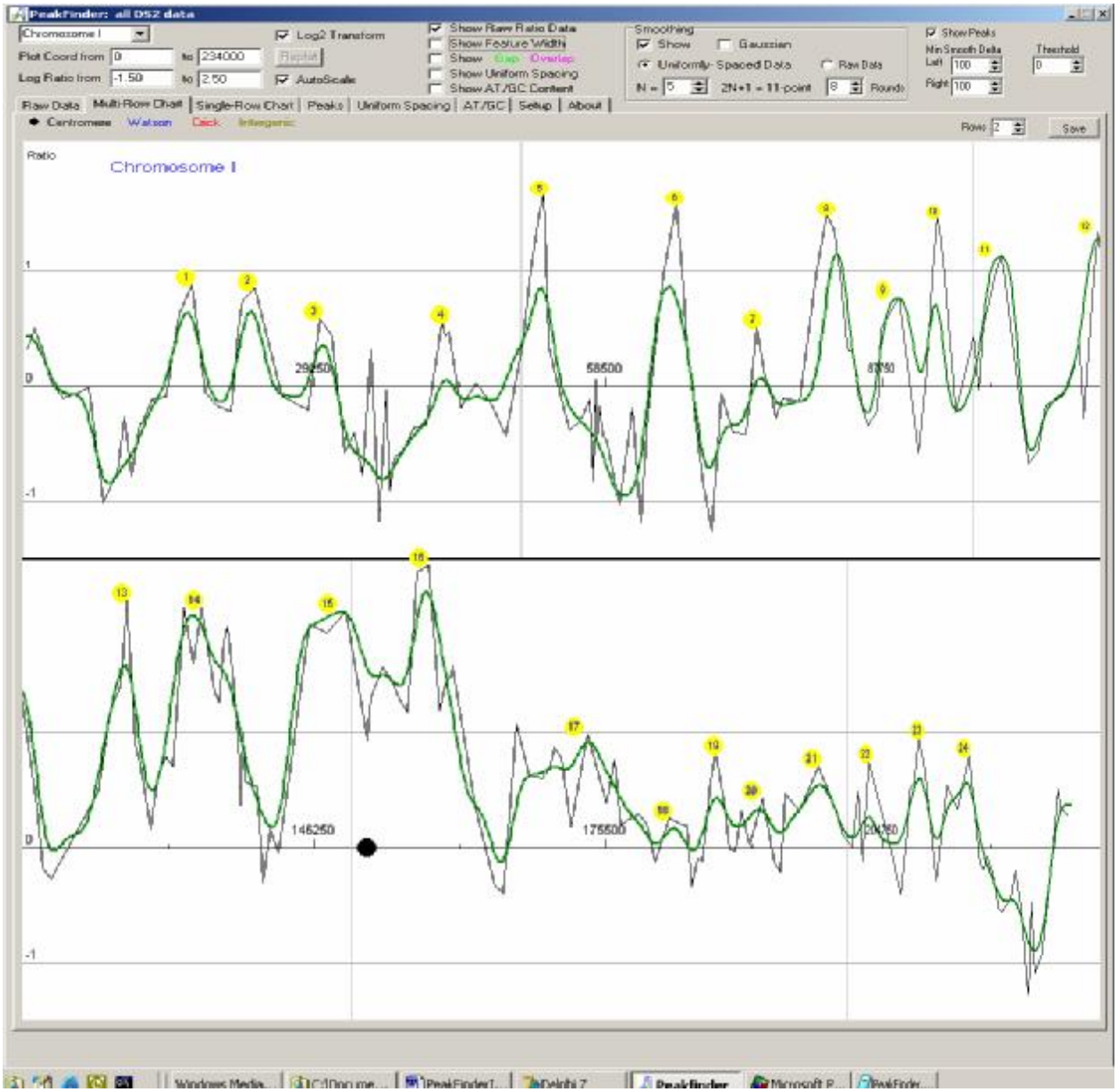
A "peak" can be found by identifying any point that has two (or more) "+" segments preceding it and is followed by two (or more) "-" segments – two consecutive ones are used instead of only one to avoid local maxima caused by noise. In the above diagram, there is a peak at X=6.

Once a peak is found, one can "slide" down both sides to find the valley and various statistics about the peak. For example, the peak had a left slide of 6 down over a distance of 3 units, and a right slide of 4 down over a distance of 2 units.

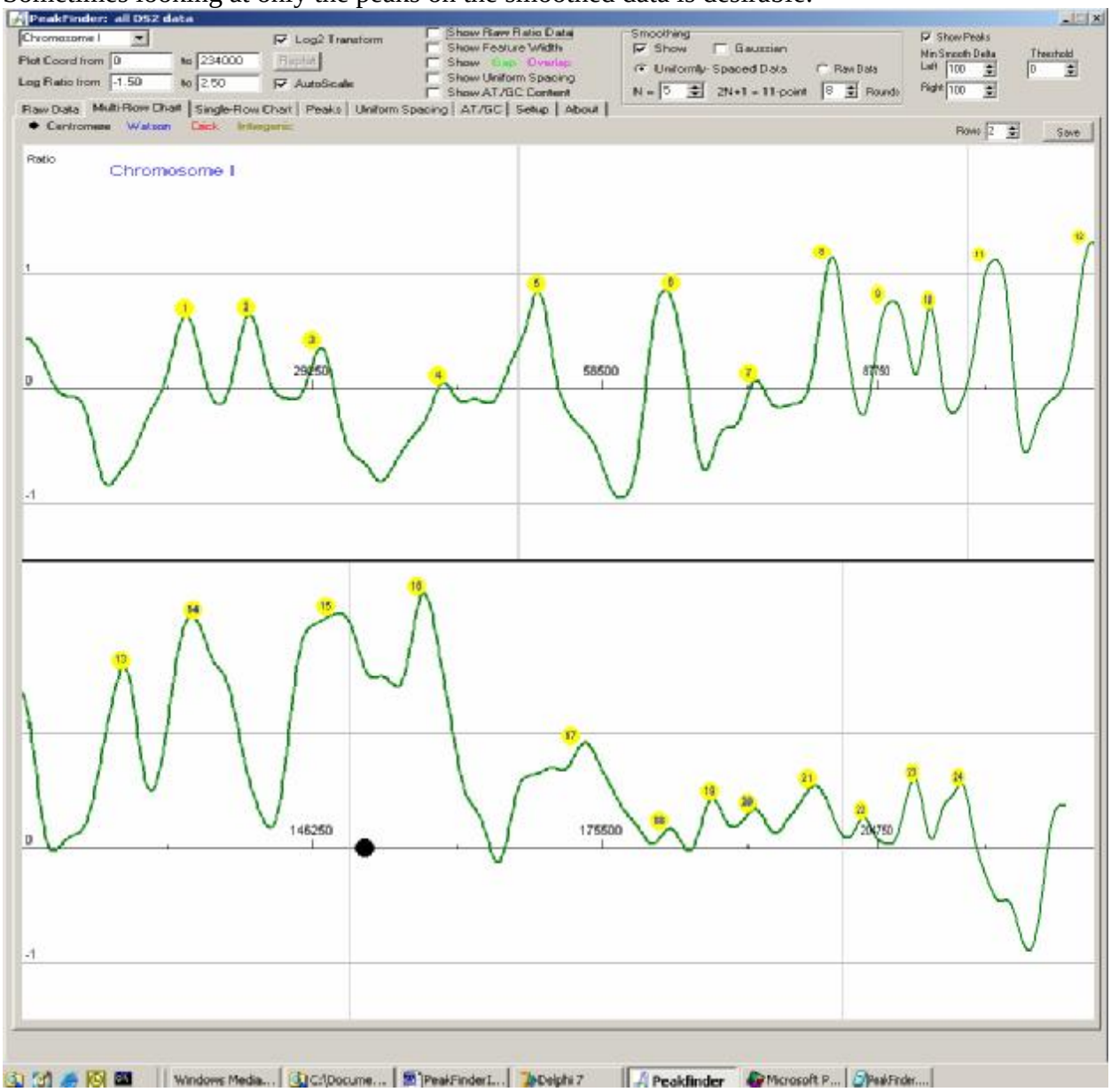
Because of the nature of the microarray data, peaks are ignored if they have a "Y" value < 1, when plotting raw ratio values, or $\log_2(Y) < 0$ when plotting log data.

The peak found from the smoothed data is not that interesting. Instead, the X range of a smoothed peak can be used to search for the real peak in the original, noisy data. Normally, the smoothed uniformly spaced data is used to find the location of a peak, but the original data is used to find the exact location of the peak. Peak statistics can also be computed using the "left slide" and "right slide" with the original data.

As shown below, the peaks found using the smoothed data are numbered consecutively within a chromosome:

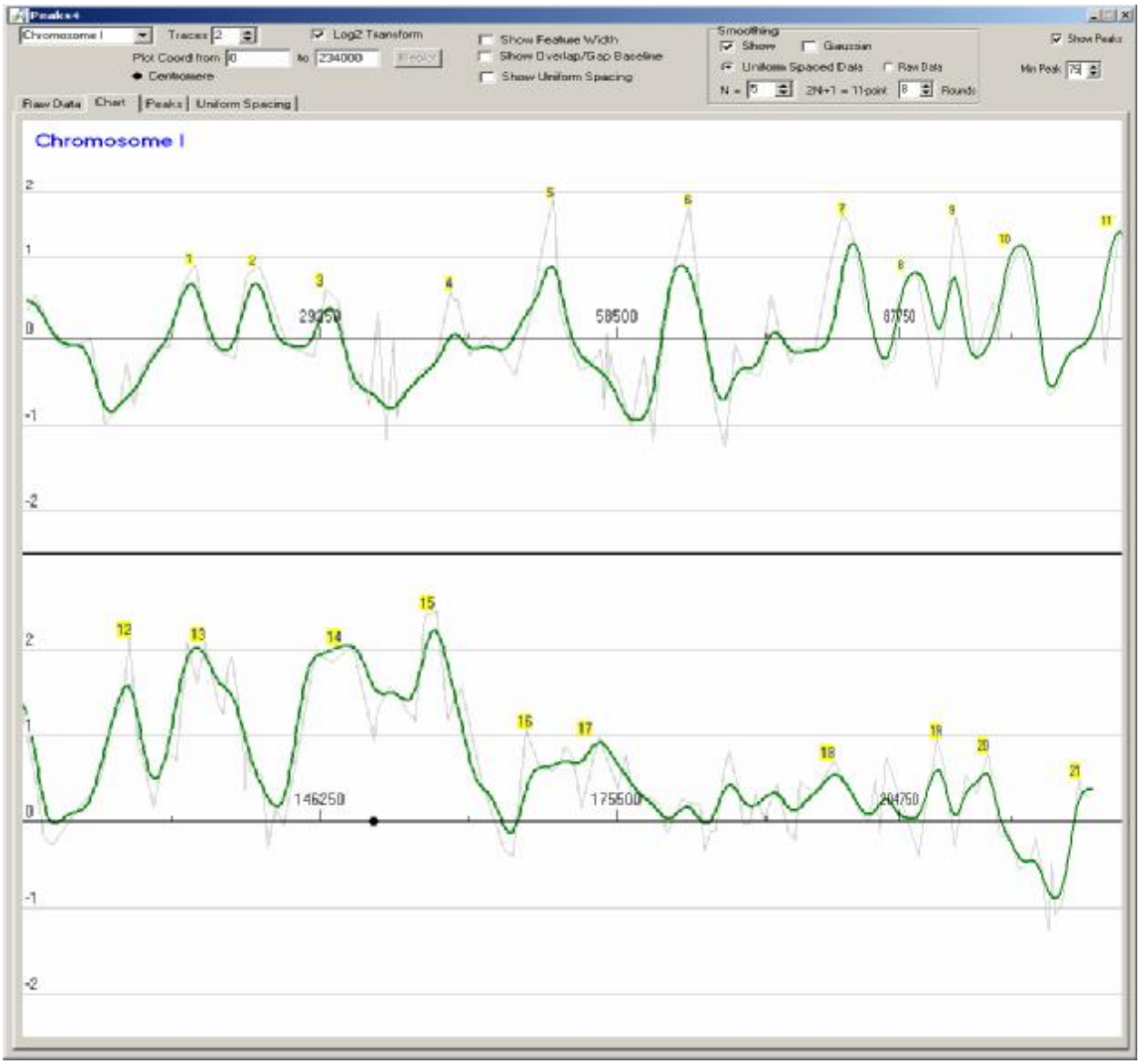


Sometimes looking at only the peaks on the smoothed data is desirable:



6 Avoiding Minimal Peaks

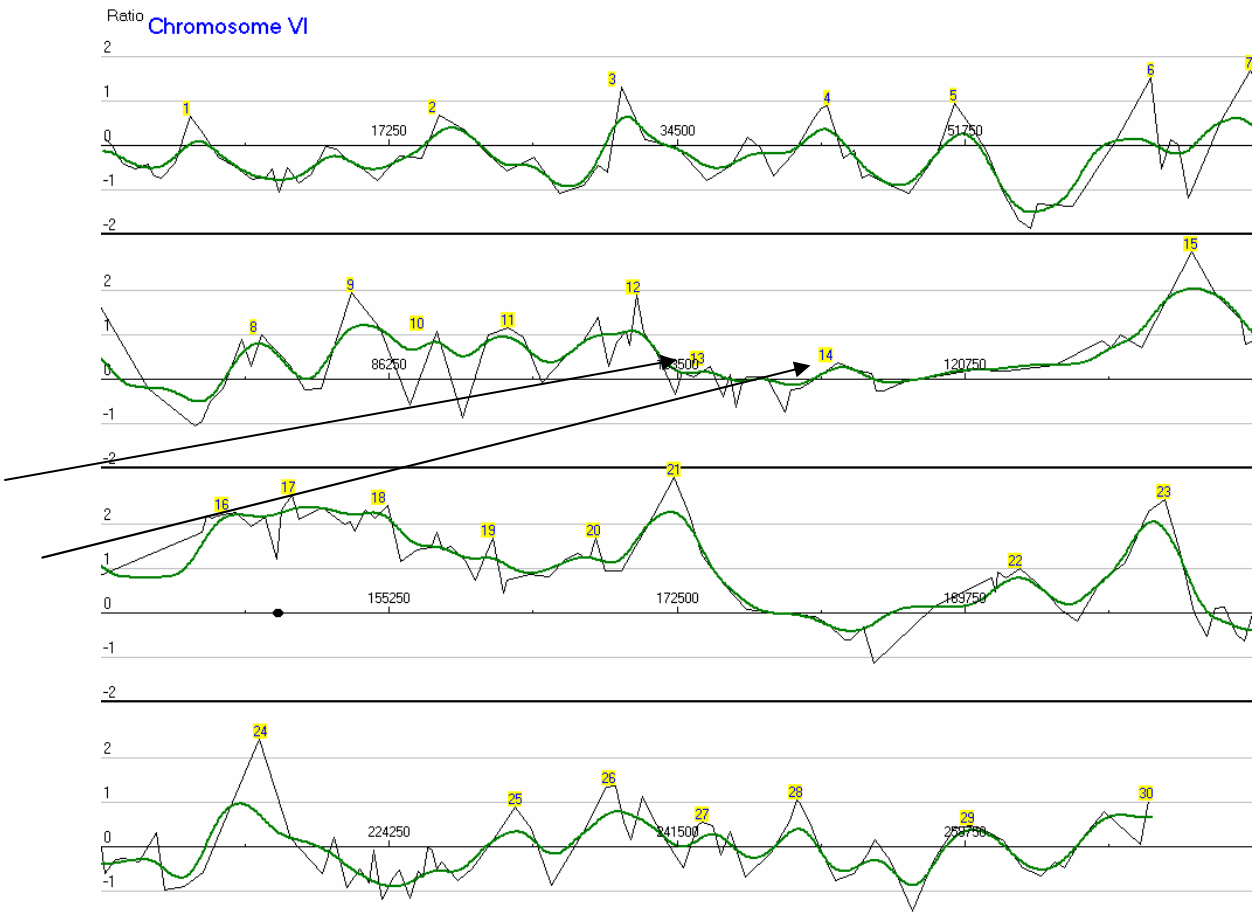
The Min Peak spin box can be used to eliminate peaks when the sum of the "left side" and "right slide" is too small. Above, 24 peaks were found with a value of 0 for Min Peak. Changing this value to 75 eliminates three peaks.

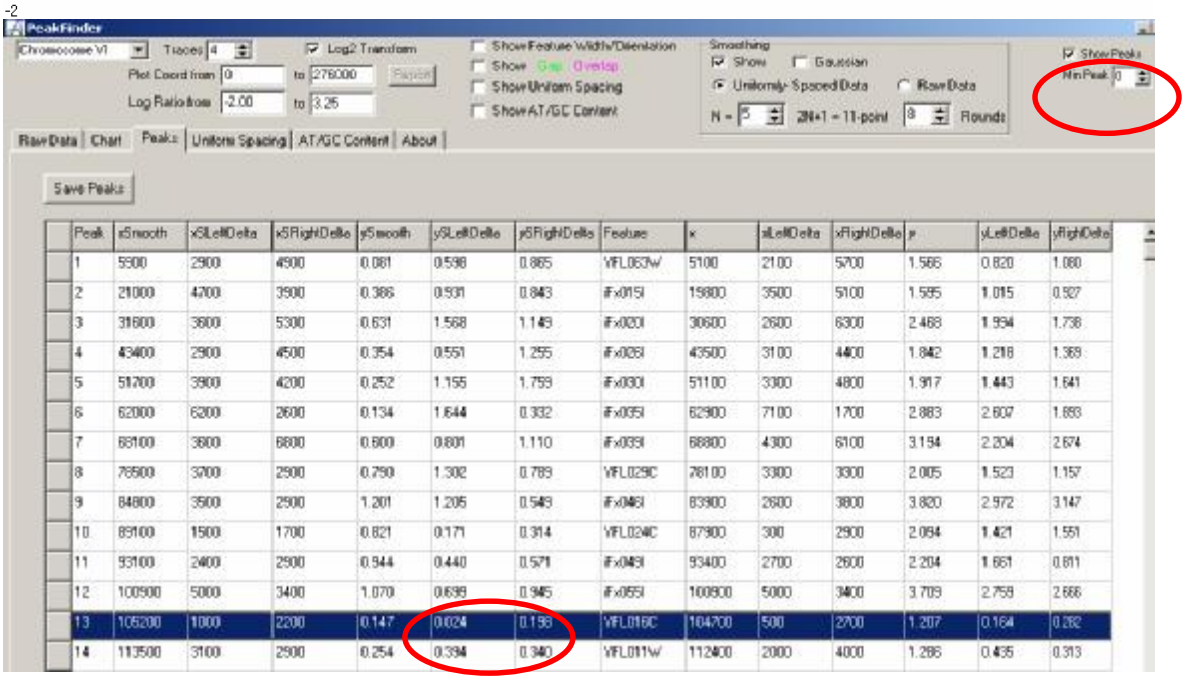


The 75 value means that in the peak statistics for the smoothed data, the

"left slide" + "right slide" < 0.01*75

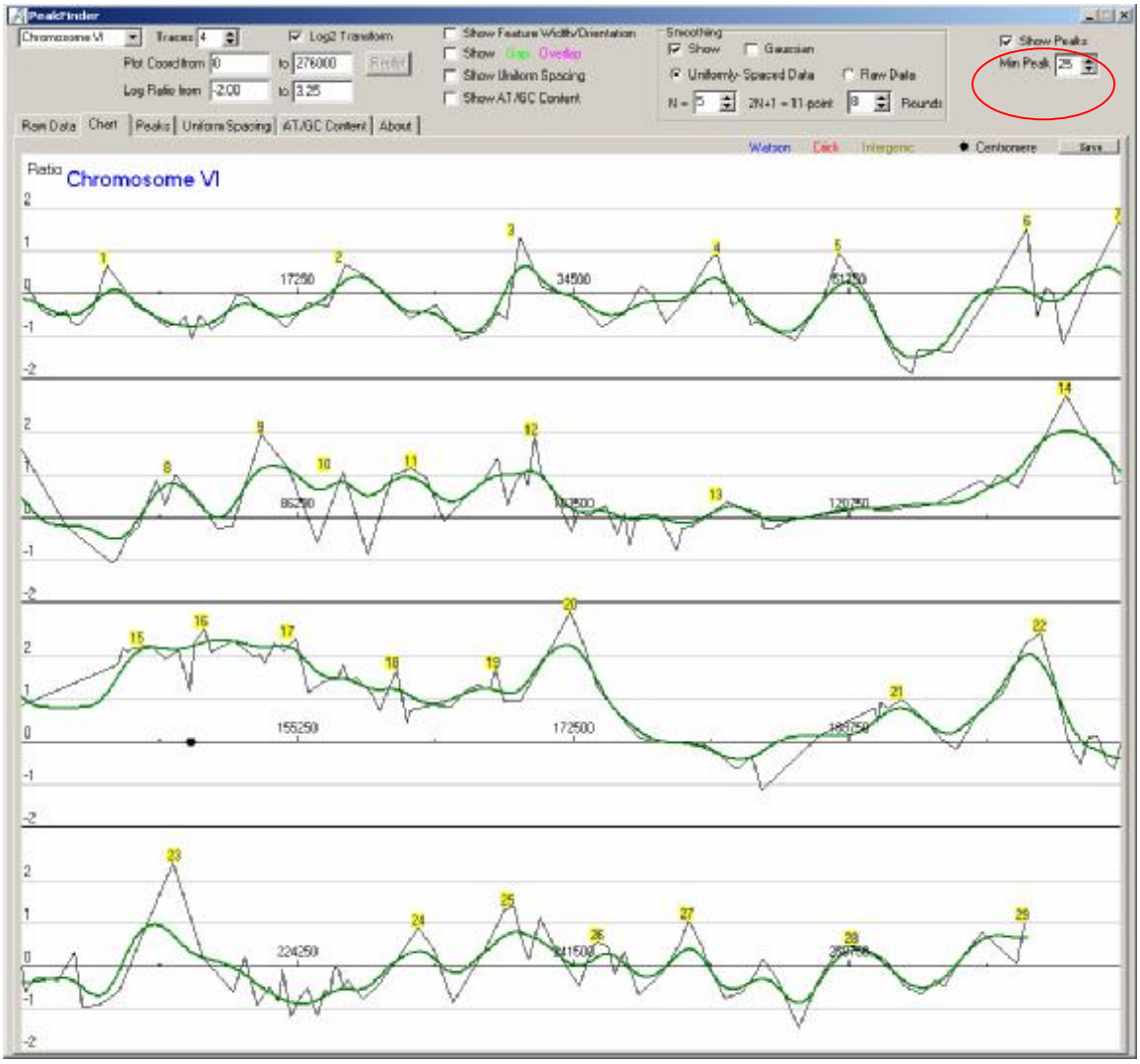
More examples:



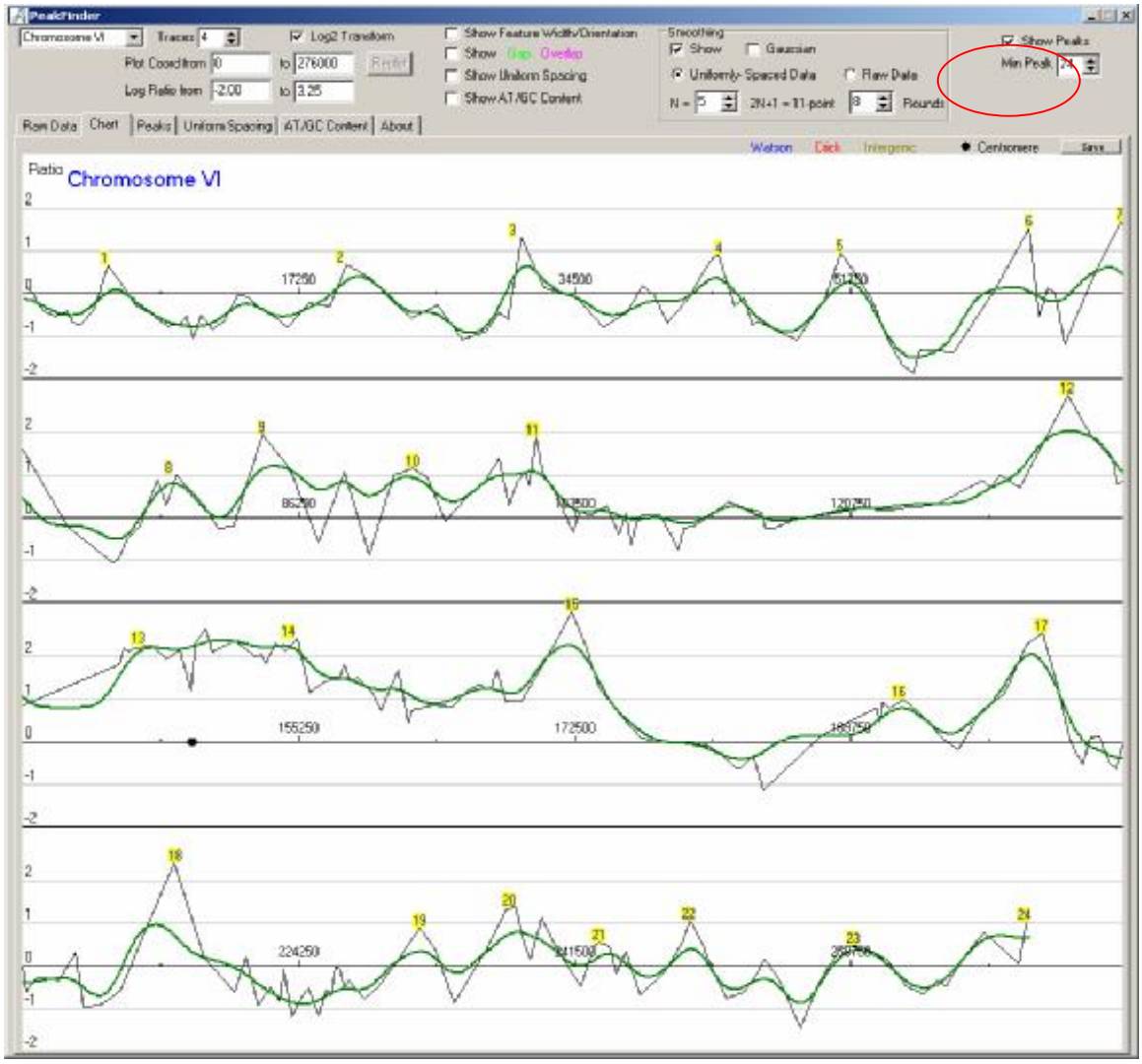


Peak	xSmooth	xSLeftDelta	xSRightDelta	ySmooth	ySLeftDelta	ySRightDelta	Feature	x	xLeftDelta	xRightDelta	y	yLeftDelta	yRightDelta
1	5500	2900	4900	0.081	0.598	0.885	VFL063W	5100	2100	5700	1.566	0.820	1.080
2	21000	4700	3900	0.386	0.931	0.843	#x015I	19800	3500	5100	1.595	1.015	0.927
3	31600	3800	5300	0.631	1.568	1.149	#x020I	30600	2600	6300	2.468	1.994	1.736
4	43400	2900	4500	0.354	0.551	1.255	#x026I	43500	3100	4400	1.842	1.218	1.369
5	51700	3900	4200	0.252	1.155	1.759	#x030I	51100	3300	4800	1.917	1.443	1.641
6	62000	6200	2600	0.134	1.644	0.332	#x035I	62900	7100	1700	2.883	2.607	1.883
7	68100	3800	6800	0.600	0.801	1.110	#x039I	68800	4300	6100	3.154	2.204	2.674
8	78500	3200	2500	0.750	1.302	0.789	VFL029C	78100	3300	3300	2.005	1.523	1.157
9	84800	3500	2500	1.201	1.205	0.549	#x046I	83900	2600	3800	3.820	2.972	3.147
10	89100	1500	1700	0.821	0.171	0.314	VFL024C	87900	300	2900	2.094	1.421	1.551
11	93100	2400	2500	0.944	0.440	0.571	#x048I	93400	2700	2600	2.204	1.661	0.611
12	100900	5000	3400	1.070	0.639	0.945	#x055I	100800	5000	3400	3.709	2.759	2.666
13	105200	1000	2200	0.147	0.024	0.193	VFL016C	104700	500	2700	1.207	0.164	0.282
14	113500	3100	2500	0.254	0.394	0.340	VFL011W	112400	2000	4000	1.286	0.435	0.313

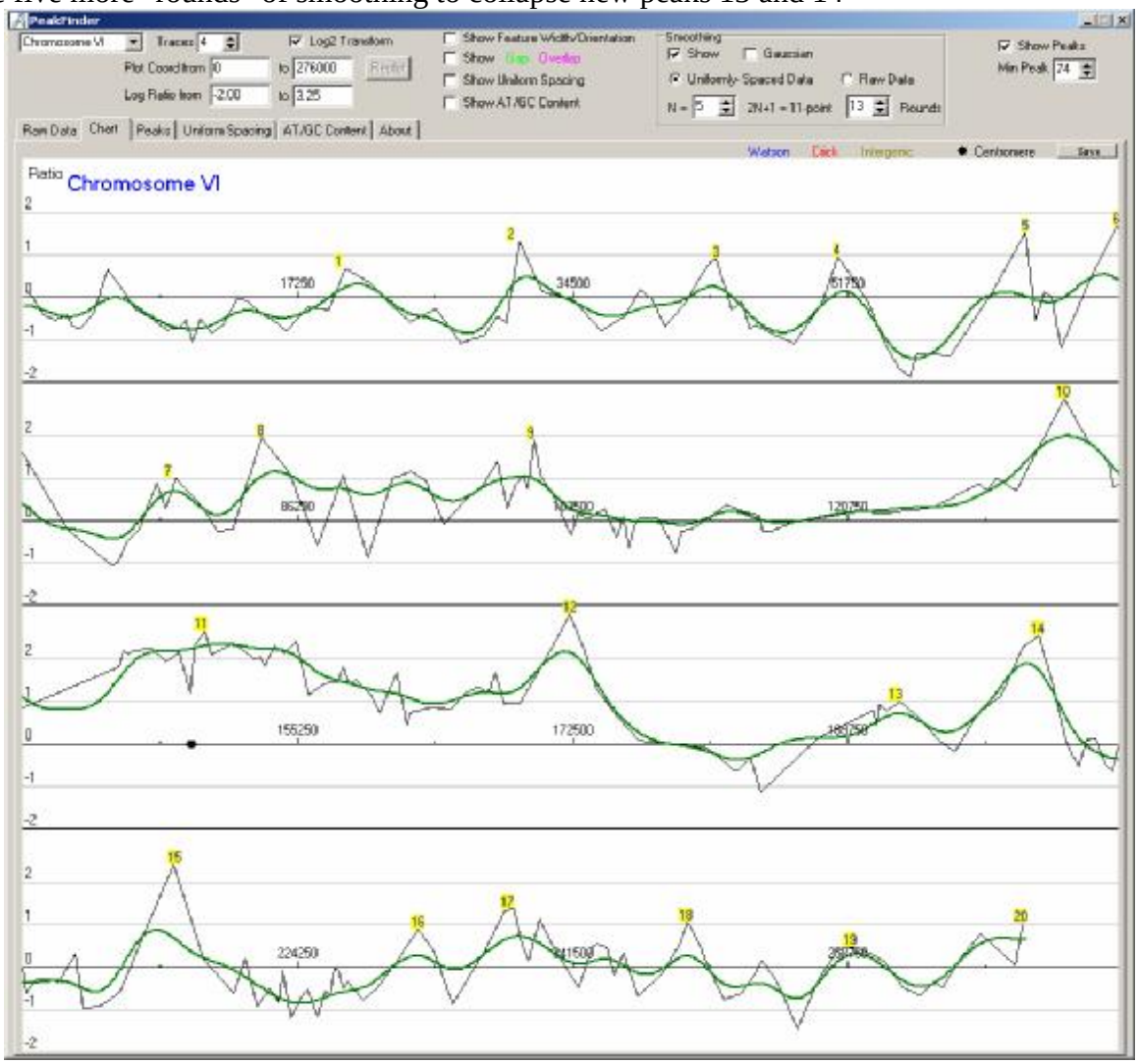
Get rid of original Peak 13:
Require sum of left and right "smoothed runs" to be greater than 0.25 (or any value over 0.024 + 0.198 = 0.222 ⇒ Value of at least 23 in the "Min Peak" spin box)



Get rid of original Peak 14:
This requires the sum of left and right "smoothed runs" to be greater than 0.74 (or any value over 0.394 + 0.340 = 0.734 ⇒ Value of 74 in the "Min Peak" spin box)



Use five more "rounds" of smoothing to collapse new peaks 13 and 14



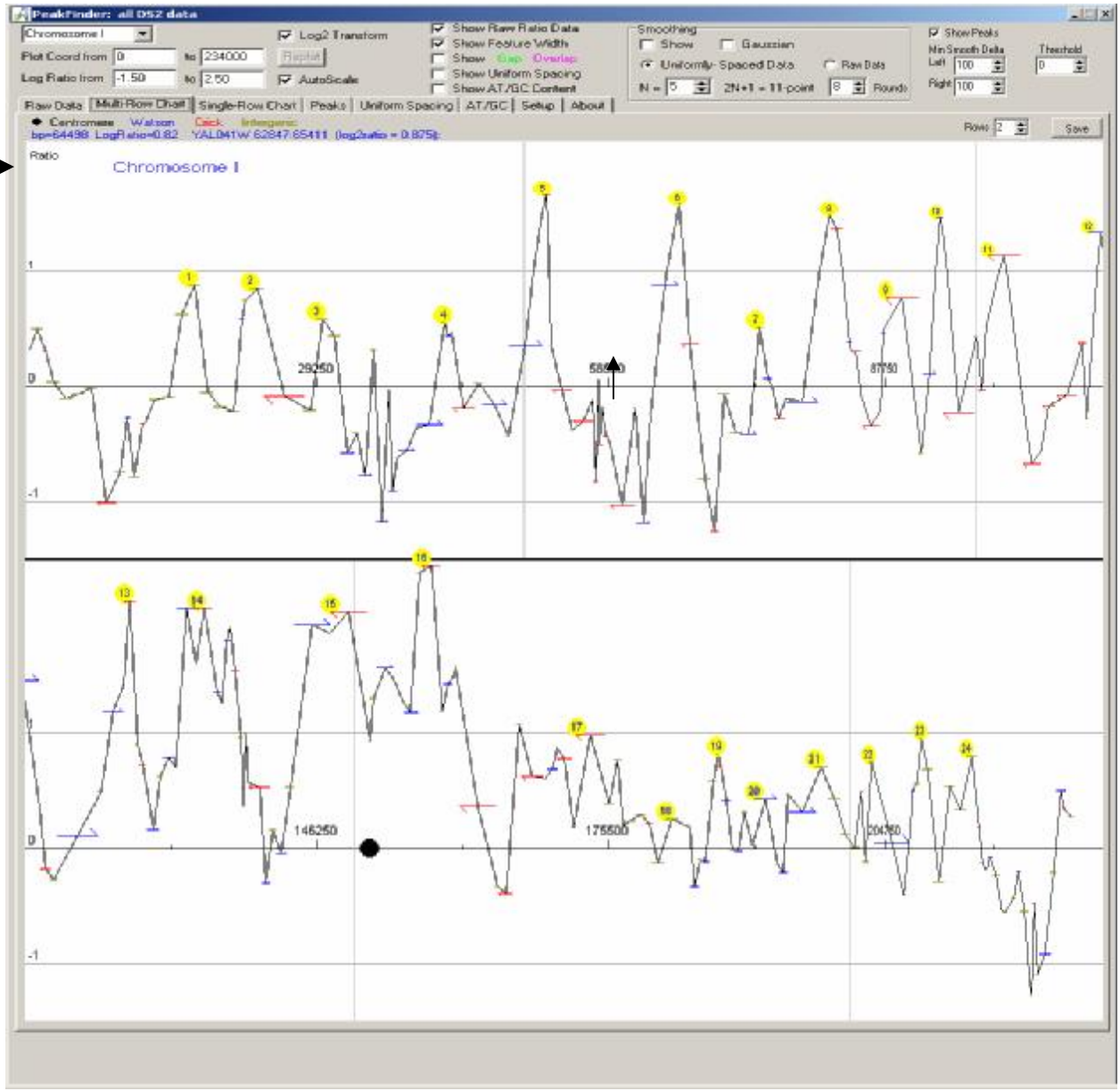
7 Display / Output Options

7.1 Multi-Row Chart

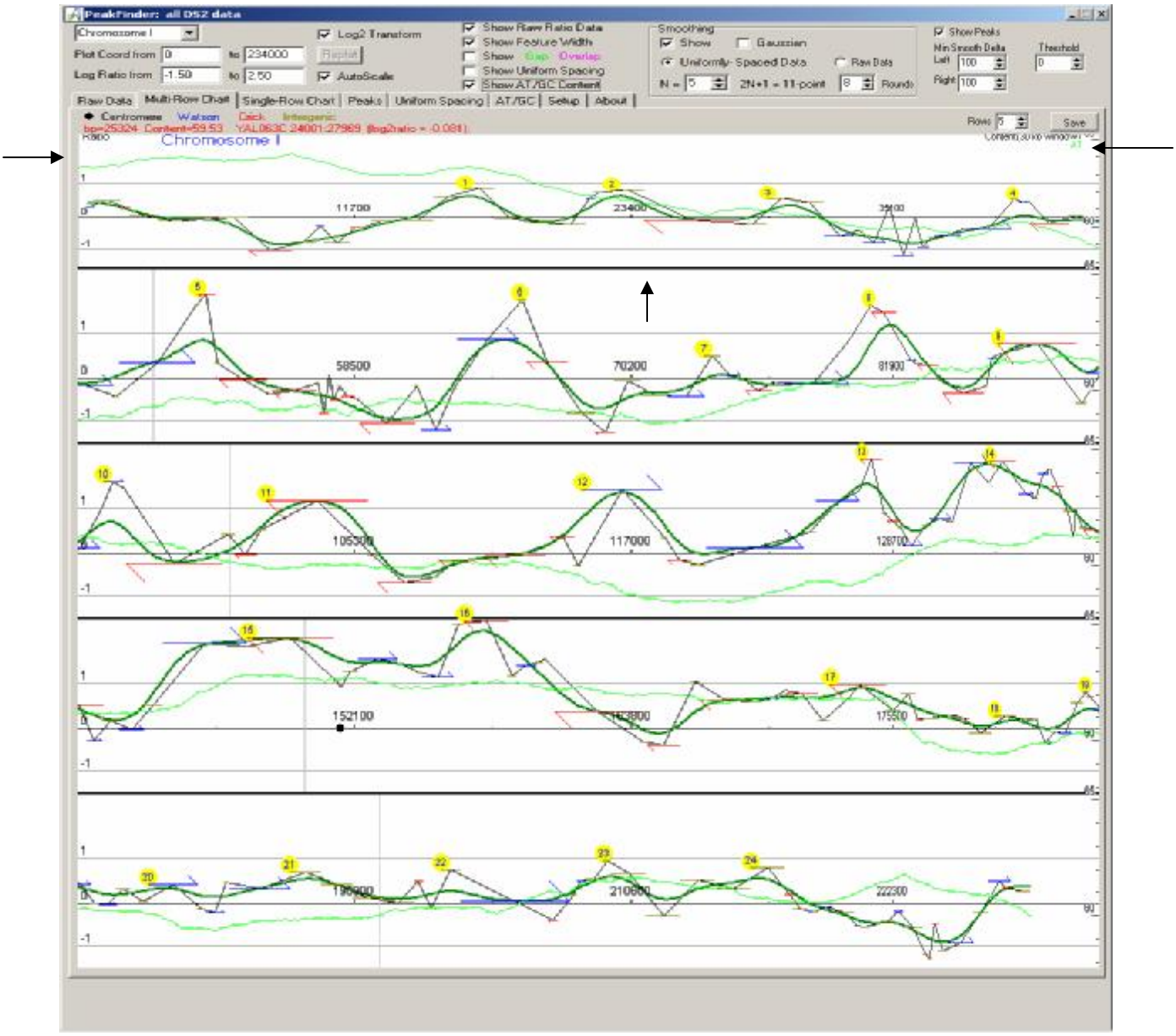
Most interactive selections are made on the Multi-Row Chart tabsheet. Many selections that alter the view of the chart automatically trigger recalculations and display. However, some changes do not trigger an automatic redisplay, but only enable the Replot button. This allows several changes to be made without an annoying slowdown for the redisplay for each change.

The "Rows" spinbox allows expansion/contraction of the X-axis to take a better look at the feature data.

Moving the mouse to the proximity of a feature (only the X coordinate matters), automatically triggers a color-coded display of information about the feature. This screen shows the X-axis divided into two rows and information about a Watson feature before peak 6.

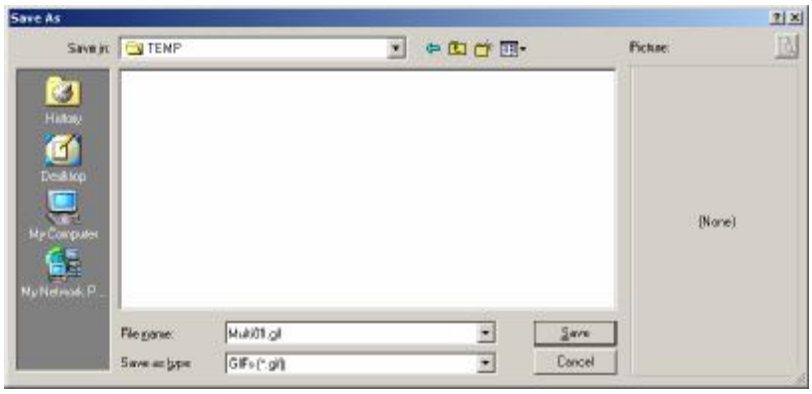

PeakFinder: all 052 data
Chromosome I
Log2 Transform
Plot Coord from 0 to 234000
Log Ratio from -1.50 to 2.50
AutoScale
Show Raw Ratio Data
Show Feature Width
Show Gaps Overlaps
Show Uniform Spacing
Show AT/GC Content
Smoothing
Show
Gaussian
Uniformly-Spaced Data
Raw Data
N = 5
2N+1 = 11-point
Rounds
Show Peaks
Min Smooth Delta
Left 100
Right 100
Threshold 0
Raw Data
Multi-Row Chart
Single-Row Chart
Peaks
Uniform Spacing
AT/GC
Setup
About
Rows 2
Save
Ratio
Chromosome I
1
0
-1
29250
87750
146250
175500
204750

This screen shows the X-axis divided into five rows and information about a Crick feature between peaks 2 and 3:

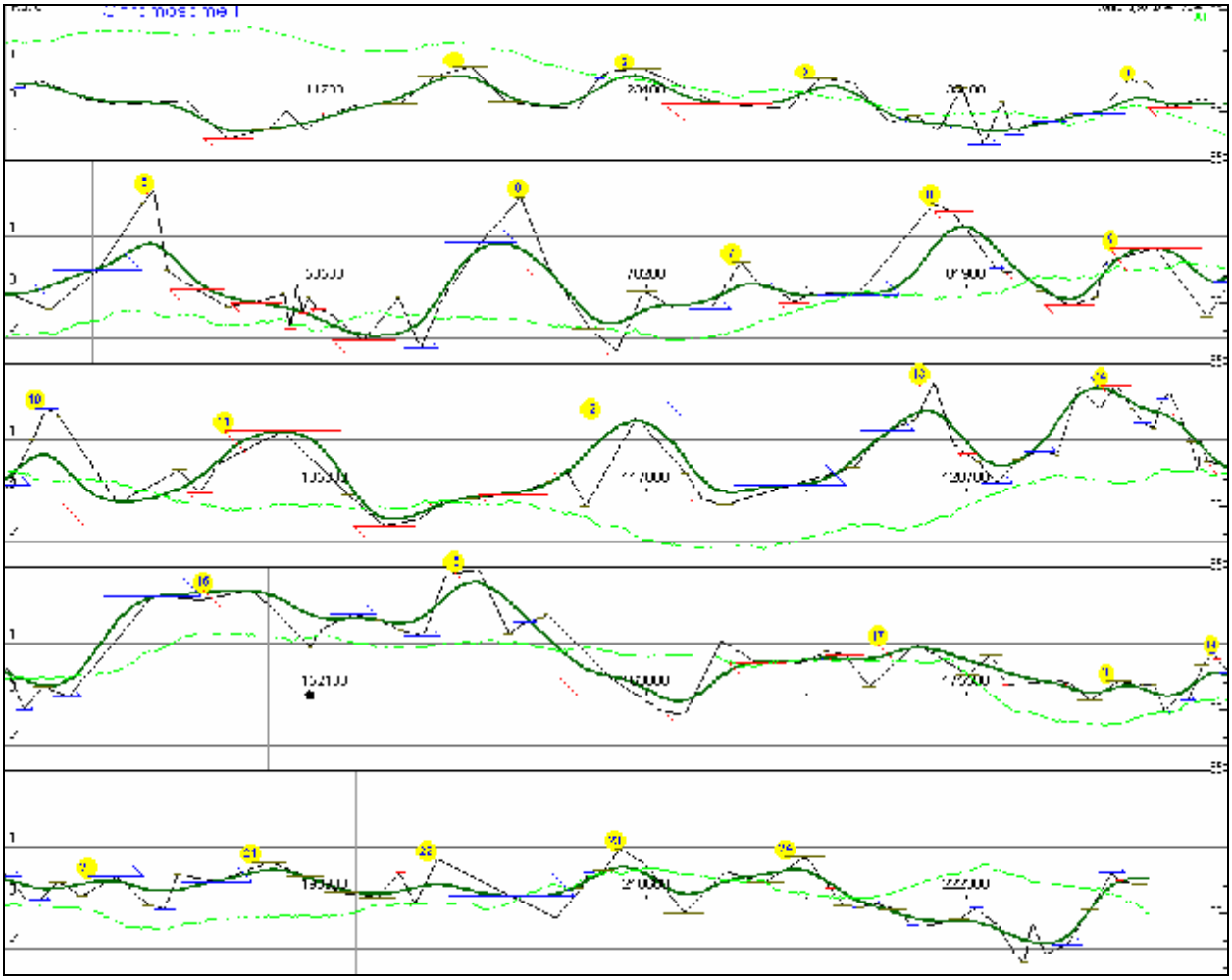


[Limitation: When the Postscript option was added, some of the labels drift off the top edge when more then 5 or 6 rows are used.]

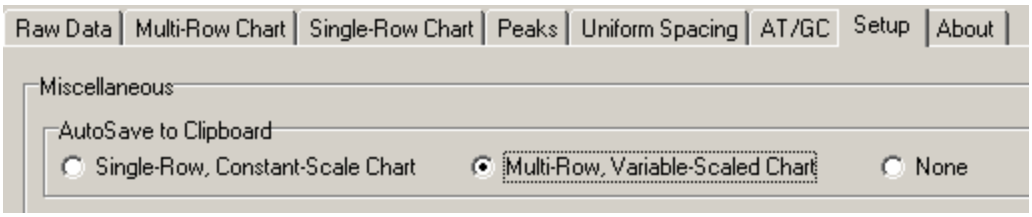
The save button (near arrow at the upper right above) can be used to save the graphic as a GIF file.



The default name for Chromosome I is Multi01.gif – but any name can be chosen. The GIF files are relatively small, but since they are bitmaps with lines and text, stretching them may not always result in a desirable look.



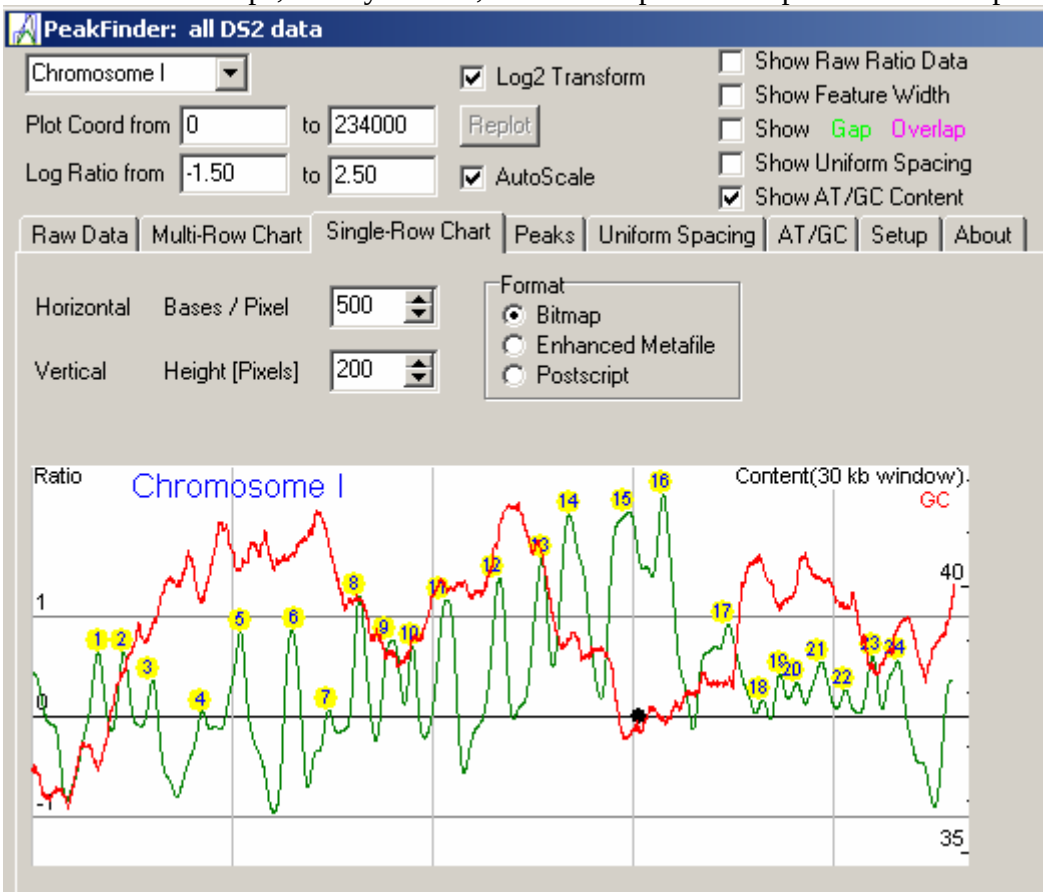
If the AutoSave to Clipboard option is set to the Multi-Row Chart (on the Setup tabsheet), this chart automatically is pasted to the Windows clipboard.



As soon as you see the chart on the screen, you can paste the graphic (Ctrl-V, or Edit | Paste in many programs) as a bitmap to Word, or any graphics program. Again, remember that stretching bitmaps with lines and text may cause aliasing artifacts that are not desirable.

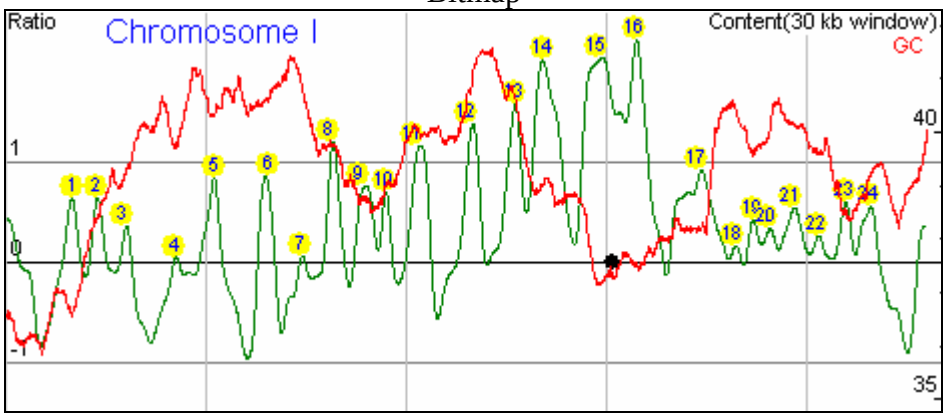
7.2 Single-Row Chart

This tabsheet gives several output options but with the chart as a single row. By default, the single-row charts are bitmaps, and by default, these bitmaps are also placed on the clipboard



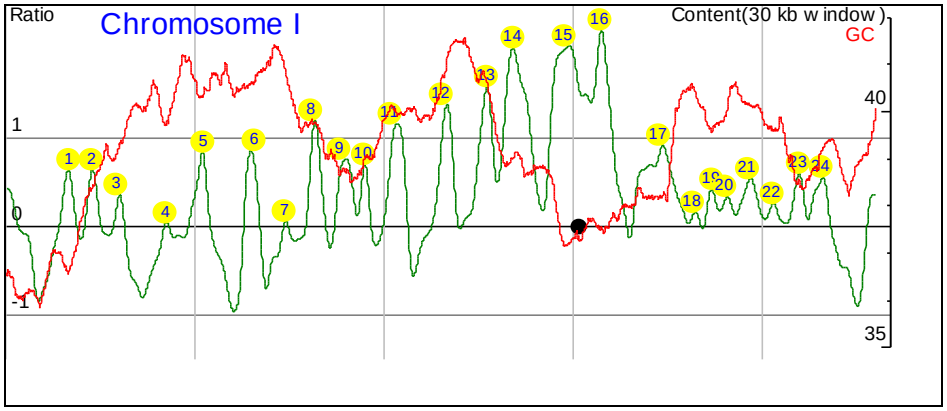
Pasting these graphics into other programs is quite convenient, such as this chart that was pasted into Word:

Bitmap



The Enhanced Metafile graphic and option was explored as a possible solution to eliminating the bitmap stretching artifacts. In Windows Enhanced Metafiles are not bitmaps but rather contain drawing instructions on how to redraw the graphic, which gives better stretching results.

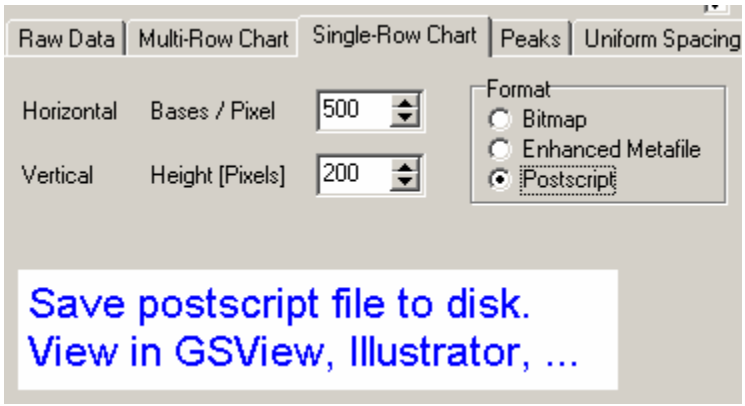
Enhanced Metafile



The Enhanced Metafiles worked fairly well for a single isolated graphic, but had two serious problems: (1) Enhanced Metafiles created on certain (not all) dual screen displays did not appear correctly when pasted or inserted in to Word files – the X dimension was only half of what was expected even though Word reported the correct size. (2) Alignment of several chromosomes within a genome does not work well at present with Metafiles. The unexplained gaps at the right and bottom of a metafile caused problems when aligning chromosomes of various sizes – at least in Word.

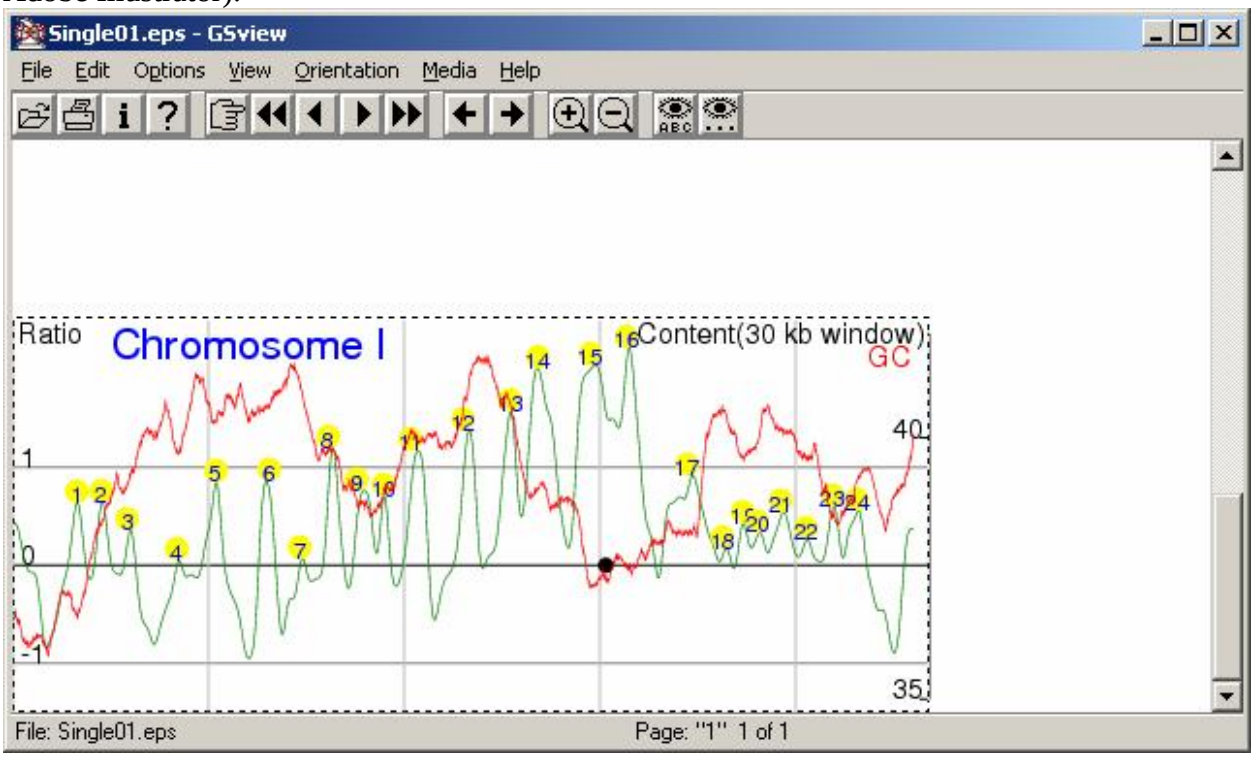
A Postscript alternative was explored [loosely based on the Postscript output created by Joe DeRisi's (http://derisilab.ucsf.edu) Promoter V2.2 program]. This option mixed the "device independent" nature of Postscript with several Windows device-dependent features. Having a common module that could draw using Postscript output, or draw directly on a Windows canvas was a bit of a challenge. Compromises were made to make both Postscript and a Windows display work from the same code, but the solution is not that general.

When the Postscript option is selected, no Windows graphic is drawn:



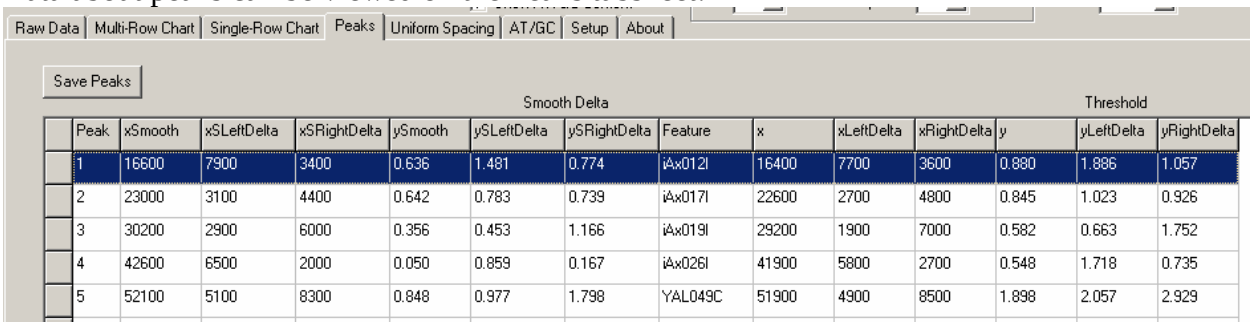
At this point, select the Save button and write the Postscript (.eps) file to disk.

The resulting .eps file can be viewed in Gsview or other programs that handle postscript. (See notes below about the "Process All Button" processing and alignment of Postscript files in Adobe Illustrator).



7.3 Peaks Table

Data about peaks can be viewed on the Peaks tabsheet:

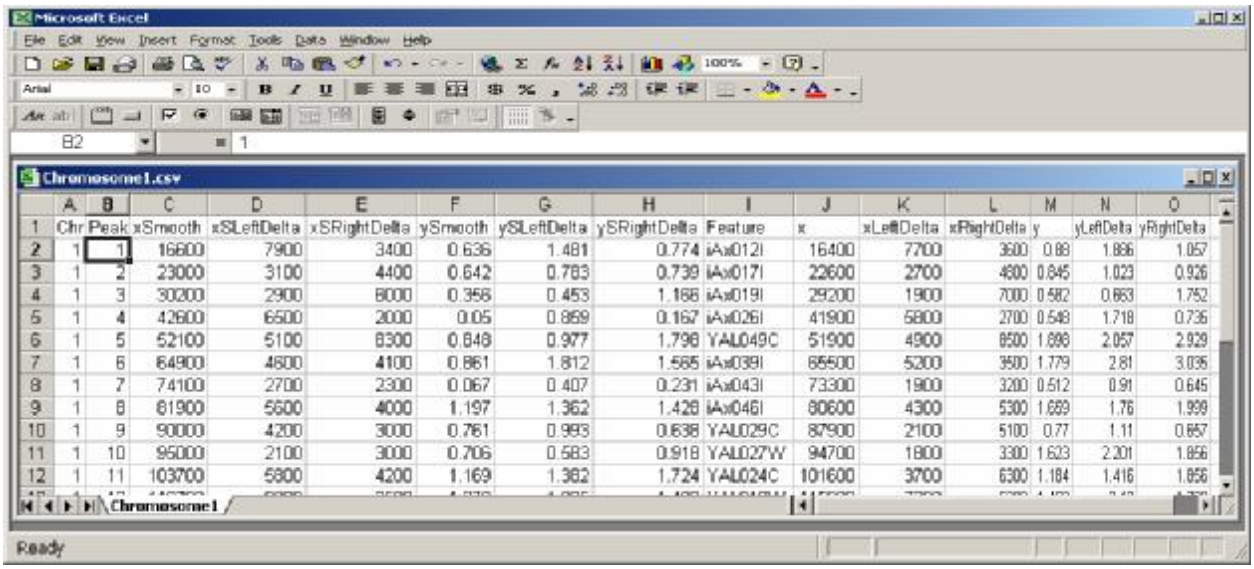


Raw Data | Multi-Row Chart | Single-Row Chart | Peaks | Uniform Spacing | AT/GC | Setup | About

Save Peaks

				Smooth Delta							Threshold		
	Peak	xSmooth	xSLeftDelta	xSRightDelta	ySmooth	ySLeftDelta	ySRightDelta	Feature	x	xLeftDelta	xRightDelta	y	yLeftDelta	yRightDelta
	1	16600	7900	3400	0.636	1.481	0.774	iAx012I	16400	7700	3600	0.880	1.886	1.057
	2	23000	3100	4400	0.642	0.783	0.739	iAx017I	22600	2700	4800	0.845	1.023	0.926
	3	30200	2900	6000	0.356	0.453	1.166	iAx019I	29200	1900	7000	0.582	0.663	1.752
	4	42600	6500	2000	0.050	0.859	0.167	iAx026I	41900	5800	2700	0.548	1.718	0.735
	5	52100	5100	8300	0.848	0.977	1.798	YAL049C	51900	4900	8500	1.898	2.057	2.929

Press the Save Peaks button to save this information into a .CSV file, which can be opened in Excel, Access, or other programs, for storage or additional filtering:



Microsoft Excel

File Edit View Insert Format Tools Data Window Help

B2 = 1

Chromosome1.csv

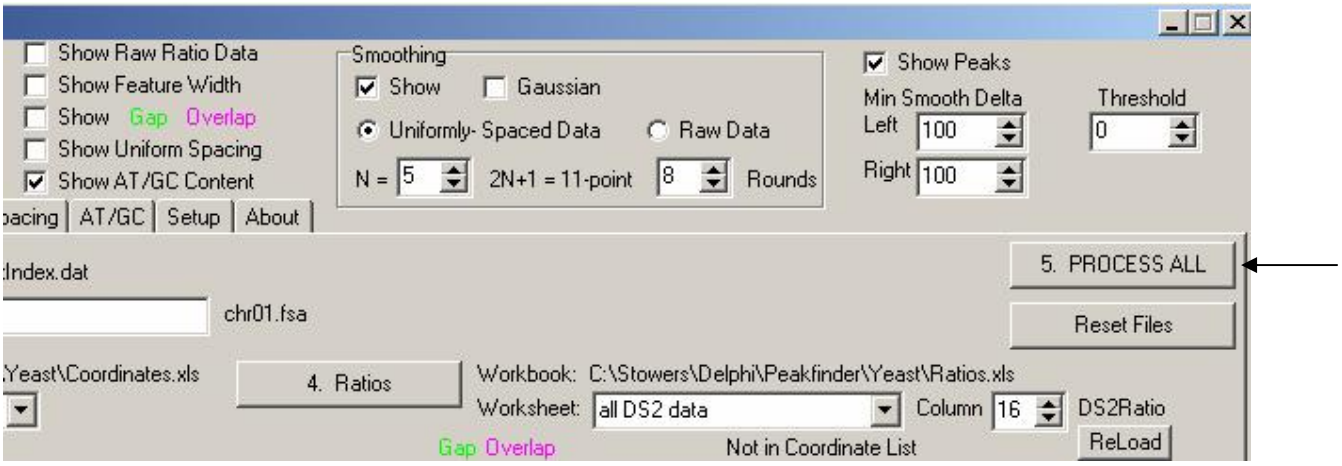
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Chr	Peak	xSmooth	xSLeftDelta	xSRightDelta	ySmooth	ySLeftDelta	ySRightDelta	Feature	x	xLeftDelta	xRightDelta	y	yLeftDelta	yRightDelta
2	1	1	16600	7900	3400	0.636	1.481	0.774	iAx012I	16400	7700	3600	0.88	1.886	1.057
3	1	2	23000	3100	4400	0.642	0.783	0.739	iAx017I	22600	2700	4800	0.845	1.023	0.926
4	1	3	30200	2900	6000	0.356	0.453	1.166	iAx019I	29200	1900	7000	0.582	0.663	1.752
5	1	4	42600	6500	2000	0.05	0.859	0.167	iAx026I	41900	5800	2700	0.548	1.718	0.735
6	1	5	52100	5100	8300	0.848	0.977	1.798	YAL049C	51900	4900	8500	1.898	2.057	2.929
7	1	6	64900	4600	4100	0.861	1.812	1.565	iAx039I	65500	5200	3500	1.779	2.81	3.036
8	1	7	74100	2700	2300	0.067	0.407	0.231	iAx043I	73300	1900	3200	0.512	0.91	0.645
9	1	8	81900	5600	4000	1.197	1.362	1.428	iAx046I	80600	4300	5300	1.689	1.76	1.999
10	1	9	90000	4200	3000	0.761	0.993	0.638	YAL029C	87900	2100	5100	0.77	1.11	0.657
11	1	10	95000	2100	3000	0.706	0.583	0.918	YAL027W	94700	1800	3300	1.623	2.201	1.856
12	1	11	103700	5800	4200	1.169	1.362	1.724	YAL024C	101600	3700	6300	1.184	1.416	1.856

Chromosome1

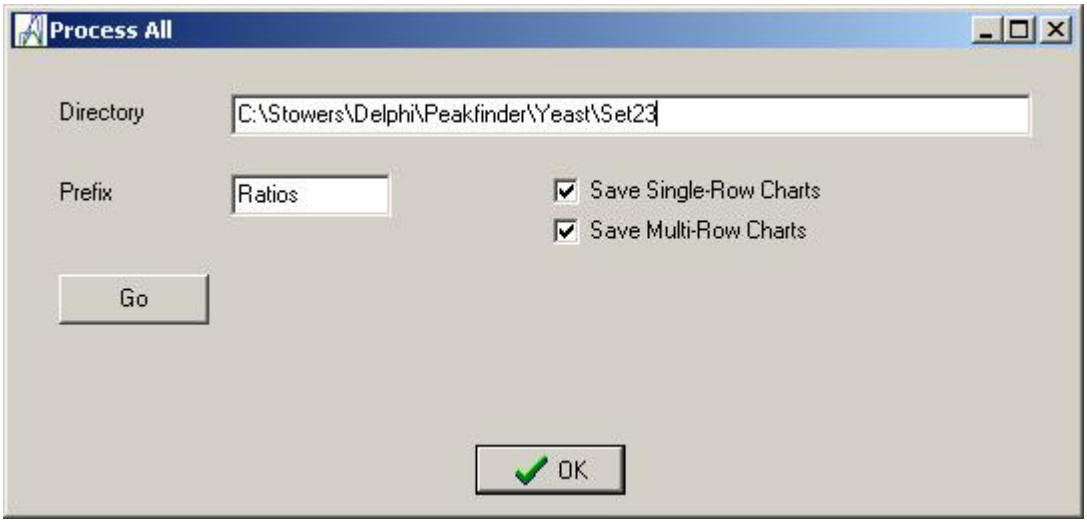
Ready

7.4 Process All Button

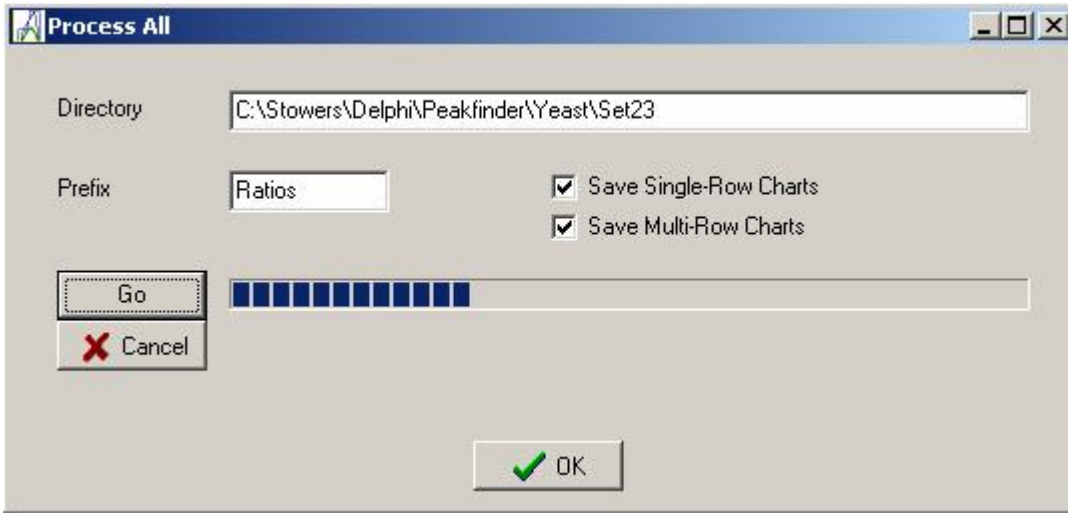
Once you have selected the desired smoothing and display options, you can process all the chromosomes in a genome by pressing the Process All button on the Raw Data tabsheet:



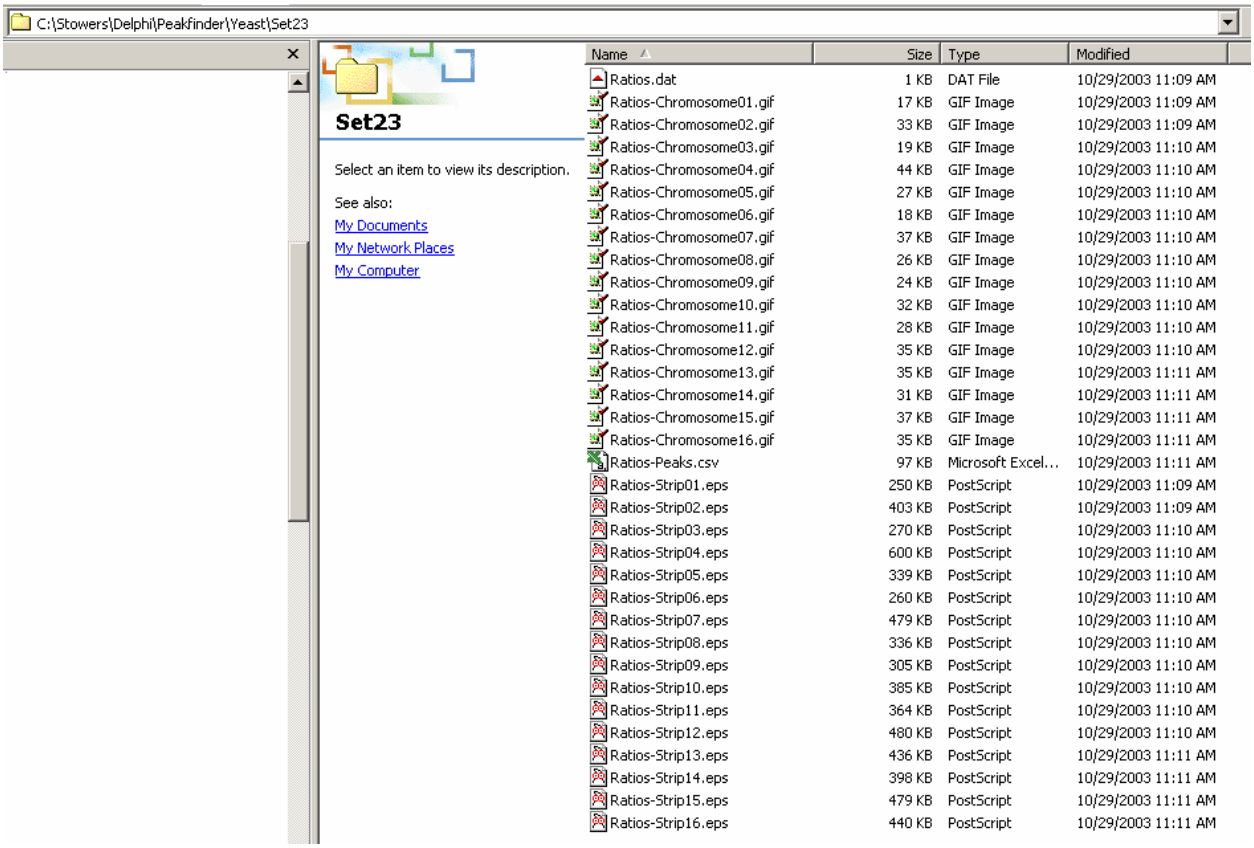
At this point you will be prompted to specify the directory to contain the output files. The default directory is the location of your Ratio data, so normally just add an additional level to store the results:



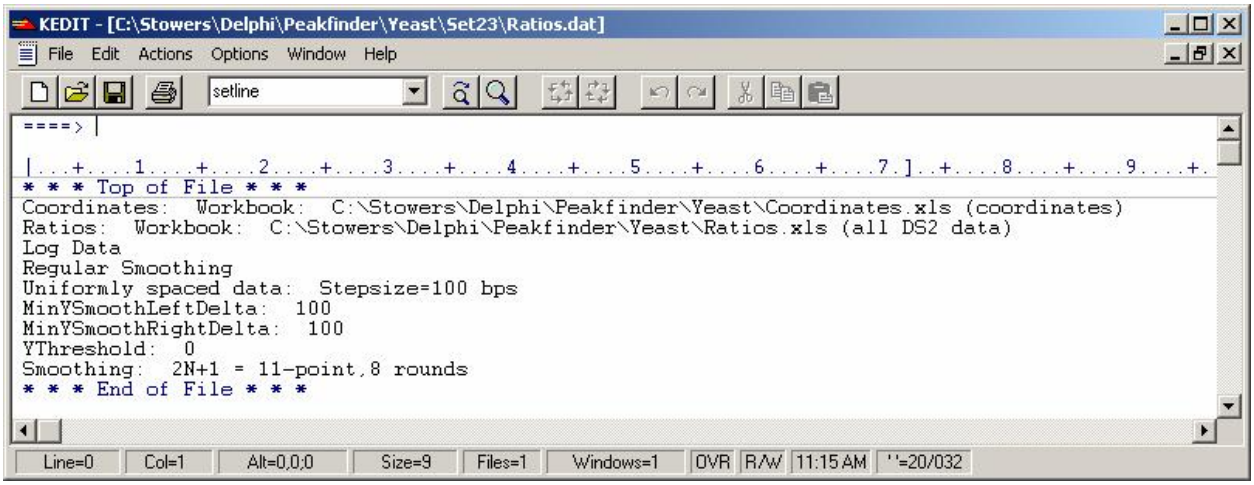
Press Go and watch the progress bar. When all process has completed, press the OK button.



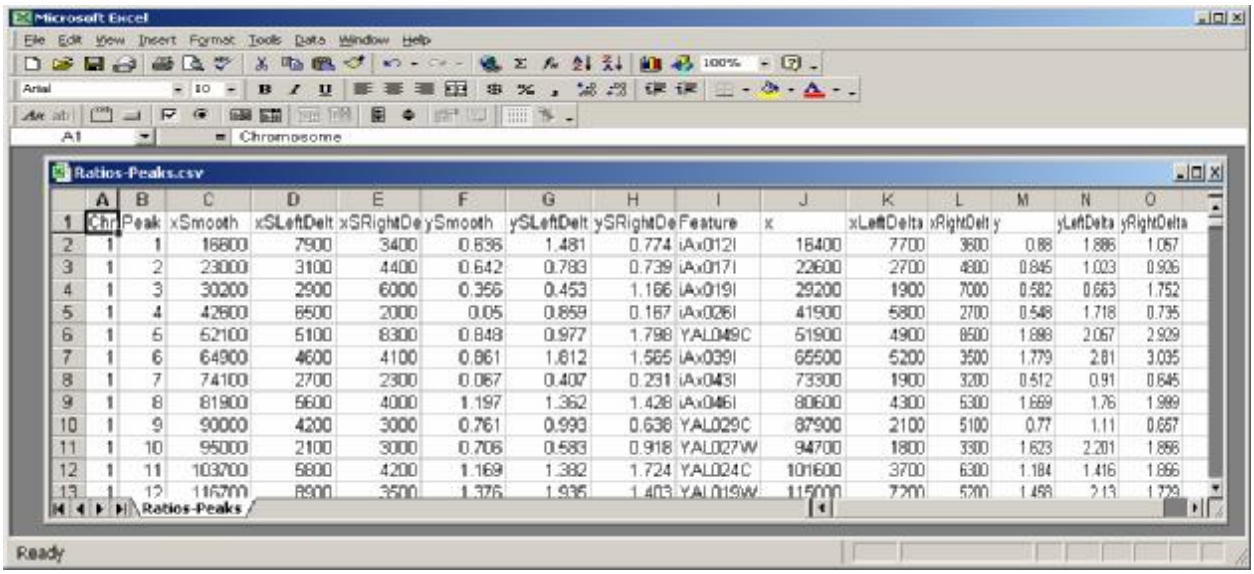
A number of files are created with the specified prefix:



The Ratios.dat file contains the PeakFinder settings so the analysis could be reproduced at a later date:



The Ratios-Peaks.csv file contains the peak information for all chromosomes:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Chr	Peak	xSmooth	xSLeftDelt	xSRightDe	ySmooth	ySLeftDelt	ySRightDe	Feature	x	xLeftDelta	xRightDelt	y	yLeftDelta	yRightDelta
2	1	1	16800	7900	3400	0.636	1.481	0.774	iAx012I	16400	7700	3600	0.88	1.886	1.057
3	1	2	23000	3100	4400	0.642	0.783	0.739	iAx017I	22600	2700	4800	0.845	1.023	0.906
4	1	3	30200	2900	6000	0.356	0.453	1.166	iAx019I	29200	1900	7000	0.582	0.663	1.752
5	1	4	42600	6500	2000	0.05	0.859	0.167	iAx026I	41900	5800	2700	0.548	1.718	0.735
6	1	5	52100	5100	8300	0.848	0.977	1.798	YAL049C	51900	4900	8500	1.888	2.067	2.929
7	1	6	64900	4600	4100	0.861	1.612	1.565	iAx039I	65500	5200	3500	1.779	2.81	3.035
8	1	7	74100	2700	2300	0.067	0.407	0.231	iAx043I	73900	1900	3200	0.512	0.91	0.645
9	1	8	81900	5600	4000	1.197	1.362	1.428	iAx046I	80600	4300	6300	1.669	1.76	1.989
10	1	9	90000	4200	3000	0.761	0.993	0.636	YAL029C	87900	2100	5100	0.77	1.11	0.657
11	1	10	95000	2100	3000	0.706	0.583	0.918	YAL027W	94700	1800	3300	1.623	2.201	1.866
12	1	11	103700	5800	4200	1.169	1.382	1.724	YAL024C	101600	3700	6300	1.184	1.416	1.866
13	1	12	116700	8900	3500	1.376	1.935	1.403	YAL015W	115000	7200	5200	1.458	2.13	1.729

The Multi-Row charts are saved as GIFs. Since Postscript was selected as the Single-Row Chart option, a series of encapsulated postscript (.eps) files were created. These files can be opened in Adobe Illustrator, scaled by ~20% (Transform | Scale | Uniform 20%), and aligned for a genomic display:

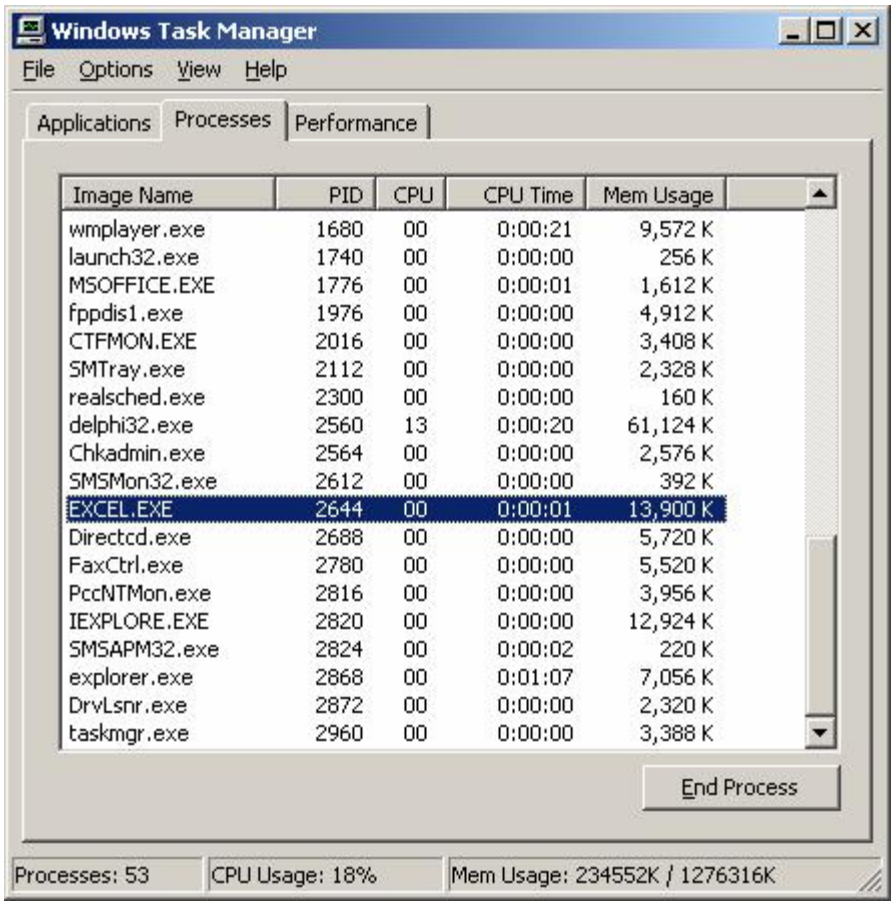
Aligning PeakFinder .eps Files in Adobe Illustrator

Alignment of the Postscript files worked much better than the Windows metafiles for publication diagrams.

8 Troubleshooting

If the program crashes…

If the Peaks program is interrupted, or if it crashes, it's possible that Excel is still loaded in memory and that instance of Excel can interfere in running the program again, or in running Excel directly. To fix such a problem, right click on the task bar and select "Task Manager". Select any instance of EXCEL.EXE (as shown below) and press the *End Process* button. Everything should work normally once all the extra instances of Excel are stopped.



If you encounter other problems, contact PeakFinder's author, Earl Glynn, at
e f g @ s t o w e r s – i n s t i t u t e . o r g