

ChIPOTle v1.0: A Tool to Identify Genomic Regions Enriched in ChIP-chip Experiments

Michael Buck, UNC-CH Lieb lab, 2005

mjbuck@unc.email.edu

Description

ChIPOTle is a Microsoft Excel add-in Macro that analyzes yeast ChIP-chip data generated on whole-genome tiled arrays.

Introduction

In contrast to mRNA microarray experiments, in which each arrayed element usually measures the abundance of one mRNA species, in ChIP-chip experiments each element measures the abundance of a population of fragments of assorted lengths due to chromatin shearing. Therefore, arrayed elements representing genomic regions 1- to 2-kb downstream or upstream from the binding site will also detect enrichment. This effect produces a peak over several arrayed elements containing genomically adjacent DNA. This is non-random behavior that is not expected from spuriously high ratio measurements. ChIPOTle takes advantage of this fact and uses it as an independent confirmation of enrichment for a given genomic region.

ChIPOTle works by determining a sliding window average across each chromosome. A window of selected size (default 1 kb) is slid across a region or chromosome, and the average \log_2 ratio of any arrayed elements that fall within that window is determined. The window is moved downstream by the step size (default 0.25 kb), and then the calculation is repeated iteratively for the whole chromosome. This sliding average will identify binding sites as peaks. The height of peaks caused by spuriously high ratios will be reduced, since the probability of a neighboring genomic element also having a high ratio is low. ChIPOTle also defines a array density value for each peak based on the number of independent arrayed elements used to construct the peak.

The utility of this approach is that it does not depend on the absolute number of targets, but on the density of their distribution. It is appropriate for detecting any number of targets that are distributed with a frequency less than approximately three times the average sheared chromatin size. For example, if the average sheared chromatin size were 1 kb, this method would be useful for the detection of any protein predicted to be spaced at intervals of at least 3 kb. A drawback to this approach is that it requires high-resolution tiling arrays.

Setting up and running the Microsoft Excel CHIPOTLE Macro

1) Start excel by double clicking on the ChIPOTle.xla file. You may get a security warning, and if so click "enable macros". This will start Excel with the macro loaded. ChIPOTle will add a menu option to the excel Tools toolbar "CHIPOTLE".

If you are having problems, make sure Excel's security setting for macros is set to medium or low. Excel's security setting may be changed in the tools-macros -security menu option.

2) Open a spreadsheet containing your data in five columns (Spot name, \log_2 ratio, Chromosome or region ID, start coordinate, and stop coordinate).

3) To run ChIPOTle , go to the tools – CHIPOTLE menu option. You will be presented with a set of options.

Setting Parameters

1.) You will be prompted to select the cells containing the required input data. Select the cells containing the spot names (string), log ratios (real), chromosome number (string or number), start and stop coordinate (integer).

2) Selecting window size and step size: The program was designed to use a window size equal to the average shearing size of the DNA used in the ChIP. The step size should be set at $\sim\frac{1}{4}$ the shear size. Default settings – window size 1000 bases, step size 250 bases (Figure 1) .

3) Select significance criteria: (A) Peak height cutoff (\log_2 ratio value, default 1.0), to use as a cutoff for significant peaks. (B) Assume that the background distribution is Gaussian. (C) Estimate the background distribution using a permutation simulation. See “Picking a significance criterion” below for more details.

4) Permutation Parameters – If you selected permutation simulation, two additional parameters are required before the program will run. These are the number of permutations and the p-value. The number of permutations is the number of times all the data will be shuffled and the sliding window used to determine the negative peak distribution. The larger this number the longer it takes to run the program. The p-value is used in determining the cutoff via permutation simulations. In addition, the user should pay close attention to the number of significant negative regions (Significant Negative Regions). If there are many significant negative regions when compared to significant positive regions (Significant Regions), then the p-value cutoff should be decreased. A p-value cutoff that produces about 50 times more significant regions than false regions may be satisfactory.

Running the program

1) ChIPOTle retrieves the chromosome number, start, and end coordinates for each array element from the inputted data.

2) If selected, ChIPOTle estimates a cutoff for the selected p-value. The program updates its progress in the bottom left of the window.

3) The program calculates the sliding-window average for your data and outputs several data sheets.

Output

1) ChIPOTle will add the following sheets to the data workbook:

- SummarySheet - Contains all the data with the spot start and stop
- Significant Regions - Lists all regions above the positive cutoff
- Significant Negative Regions- Lists all regions below the negative cutoff
- Chromosomes aveP - Contains full output for each chromosome
- Peaks - Lists all the positive peaks above the positive cutoff
- Description - Lists the settings for CHIPOTLE run
- FDR - Lists all peaks identified by CHIPOTLE with the p-value and q-value for false discovery rate when using the permutation simulation approach.

2) Output Column Labels:

A) "Significant Regions" and "Significant Negative Regions"

Chromosome – Chromosome Number

Position – Start of window

Ave Log Ratio – Sliding window average for that region starting at position

of spots – Counts the number of independent spots used to get the average

Names – List the name of the spots for that region

B) Peaks above cutoff

Peak Number – Peak ID number by location

High Average – Highest window average for that peak

High Ratio – Highest log ratio for that peak

High Spot – The array element with the highest log ratio

Length – The length of the peak above the cutoff

Chromosome – Chromosome location or region

Peak Start – The first window average above the cutoff

Array density – A measure of the number of independent spots in the peak. A "1" means that only one spot was used to make that peak above the cutoff, therefore, this peak may not be reliably enriched.

P-value – Probability of enrichment via Gaussian or permutation

C) FDR

Peak Number – Peak ID number by height

High Average – Highest window average for that peak

High Ratio – Highest log ratio for that peak

High Spot – The array element with the highest log ratio

Chromosome – Chromosome location or region

Peak Start – The first window average above the cutoff

P-value – Probability of enrichment via permutation

Q-value – Q-value for determining FDR

Figure 1. Loading required input data and running ChIPOTle

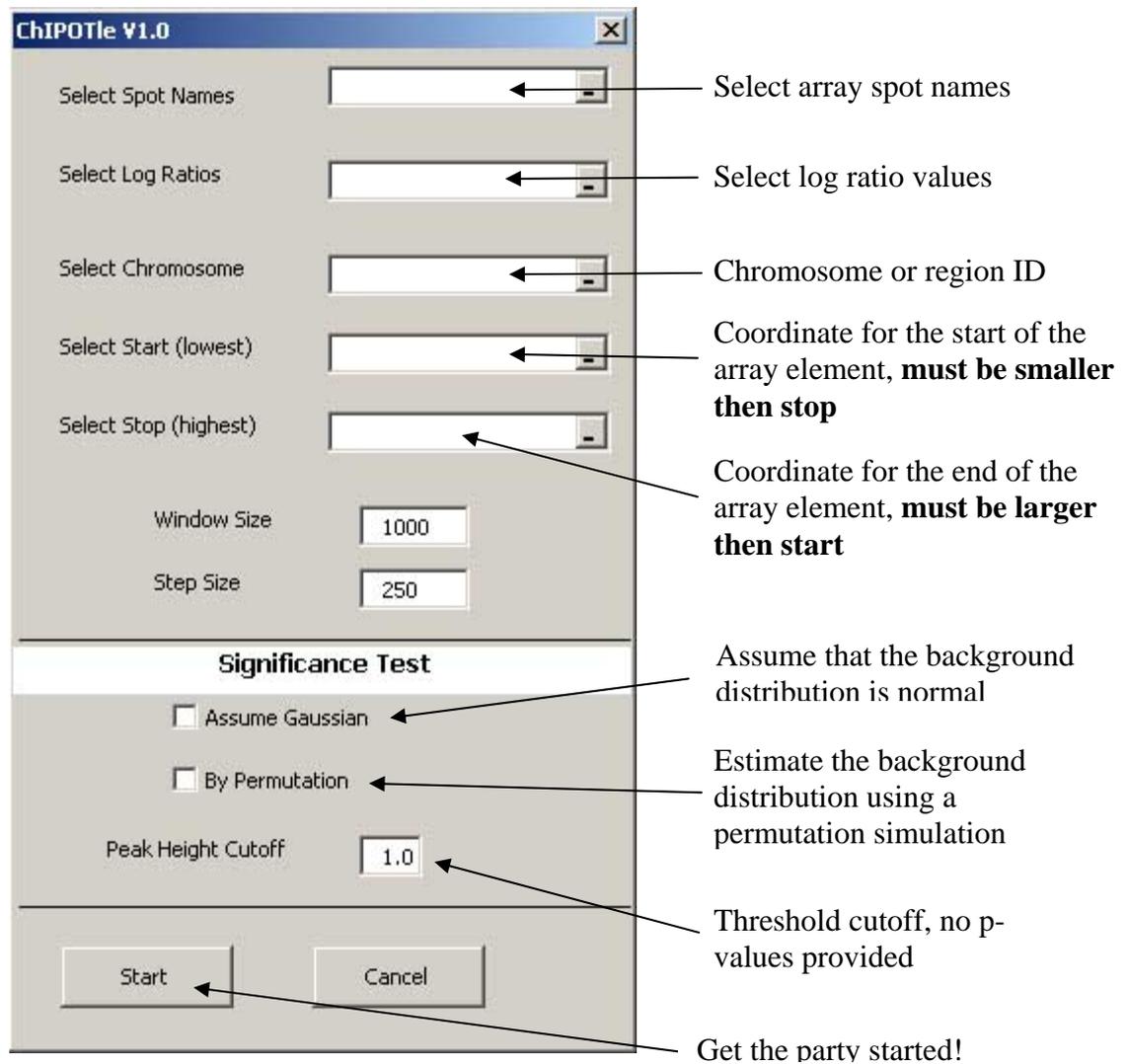


Figure 2. Looking at Results

SHEET – “Peaks above cutoff”

1	Peak Num	High Avera	High Ratio	High Spot	Length	Chromosome	Peak Start	Confidence	P-value
2	2	2.689125	2.77225	YAL069W	2000	1	600	2.25	2.291E-54
3	2	0.932625	1.27775	YAL063W	500	1	47750	2	0.00012694
4	3	2.744333	2.744333	YAL038W	2500	1	71500	3.19	2.0853E-29
5	4	1.686375	2.048	YAL034C	2000	1	81250	2.325	3.6466E-24
6	5	2.638	2.638	TEL1R	2250	1	225250	2.44444444	4.2737E-40
7	6	1.450667	1.450667	YBL113C	2750	2	600	1.63636364	5.4444E-10
8	7	3.00075	3.127	YBL109W-Q	5000	2	3750	2.6	3.589E-106
9	8	1.57725	1.800	YBL093C	2500	2	43750	3.1	0.7329E-24
10	9	2.22298	2.439667	YBL087C	3500	2	59750	2.57142857	2.2774E-52
11	10	1.197375	1.354	YBL072C	2000	2	88750	3.125	1.1231E-11
12	11	1.17125	1.73075	YBL061C	1500	2	107500	2.33333333	1.4672E-09
13	12	1.3328	1.69425	YBL026C	1500	2	167500	2.83333333	7.5048E-18
14	13	1.672003	2.12975	YBL021C	2000	2	180750	2.875	4.7151E-33
15	14	1.828	1.828	YBR003W	500	2	243250	1.5	1.5628E-11
16	15	1.897983	2.56525	YBR047W	2000	2	331600	3.25	2.5933E-43
17	16	1.34	1.72775	YBR040C-A	1750	2	414750	2.42667143	1.9315E-15
18	17	1.72725	1.72725	YBR117C	2250	2	476000	1.88888889	1.1795E-14
19	18	1.618125	1.91725	YBR181C	2250	2	592250	1.88888889	1.2305E-19
20	19	1.6795	1.83775	YBR190W	3000	2	603750	2.91888889	6.6165E-39
21	20	1.279292	1.706333	ILSR1	1250	2	681500	2.6	6.558E-12
22	21	1.985688	2.4155	YCL075W	4500	3	600	4.11111111	6.256E-91
23	22	2.208	2.38225	YCL066W	7250	3	9250	2.51724138	2.3172E-76
24	23	1.247125	1.77	YCL037C	1250	3	58500	2.6	1.4484E-13
25	24	1.031792	1.271333	YCL030C	250	3	88000	2	2.5816E-06
26	25	1.401167	2.58925	YCR025C	1500	3	163000	3.68666667	2.5547E-29
27	26	0.696938	1.293	YCR031C	750	3	177000	4	1.7471E-05
28	27	0.908983	1.64375	SNR189	500	3	178250	3.5	3.1704E-08
29	28	0.7245	1.64375	NR189	250	3	179250	3	0.00065676
30	29	2.008893	2.2389	YCR040W	2250	3	199250	3.33333333	1.833E-60
31	30	2.016	2.14275	YCR066C	4000	3	291500	3.9375	4.9003E-63
32	31	0.85625	1.45825	YCR103C	1000	3	307250	4	8.9027E-10
33	32	0.91525	2.077	YDL208W	500	4	87000	2	0.00023949
34	33	1.0775	2.077	YDL209W	500	4	88000	2	3.7929E-07
35	34	1.765	2.381	YDL192W	2250	4	116500	3.22222222	1.874E-05
36	35	1.583611	2.0925	YDL180C	2000	4	124500	2.875	2.2313E-29
37	36	1.616625	1.73825	YDL184C	2250	4	130000	3	7.8286E-27
38	37	0.963875	1.31225	YDL189C	1000	4	159250	2	3.8619E-05
39	38	1.5555	2.09775	YDL137W	2000	4	216250	2.625	2.983E-26
40	39	1.6405	2.30925	YDL133C-A	1500	4	221500	2.83333333	2.9374E-20
41	40	1.92	2.481	YDL130W-A	1750	4	228750	3.14285714	5.2701E-32
42	41	1.98675	2.66	YDL083C	1500	4	307500	2.33333333	3.3033E-31

SHEET – “Description”

1	Date	4/10/2005
2	Time	7:10:37 PM
3	Window size	1000
4	Step size	250
5	Significance technique	Gaussian
6	Gaussian Parameters	
8	Correct P-value	0.001
9	Number of unique windows	21208
10	Background Stdev	0.320573465
11	Results	
14	Number sites found	266

High window average for peak

Highest log ratio from original data for that peak

The array spot with the highest log ratio

Length of the peak above the cutoff

Chromosome ID

Start of peak

Average number of array elements in each window of the peak

P-value estimate by assuming a Gaussian background or by permutation

Date and time ChIPOTle was run

Window and step size

Significance technique

P-value threshold cutoff chosen

Number of unique windows used to correct for multiple comparisons

Standard deviation for background distribution

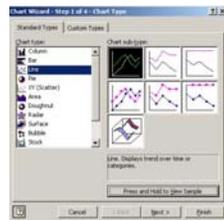
Number of binding sites found

Figure 3. Making Charts

Chromosomal Maps of sliding window average - Sheet "Chromosomes aveP" contains all the sliding window average data.

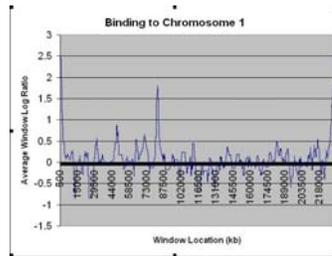
Step 1) Select the window average for the chromosome or region desired.

Step 2) Insert line chart



Step 3) Select chromosomal location for category X-axis

Step 4) Label chart as desired



Position Chr-1	Location Ave									
1	600	1 479106	YEL113C	600	NA	600	0.474819	YEL2P7C	600	
2	750	2 423181	FAL2B9V	750	2 102864	FCL2W6R	750	NA	750	
3	1000	2 423181	FAL2B9V	1000	2 102864	FCL2W6R	1000	NA	1000	
4	1250	2 423181	FAL2B9V	1250	2 294322	FCL2W6R	1250	NA	1250	
5	1500	1 462384	FAL2B9V	1500	NA	1500	2 294322	FCL2W6R	1500	
6	1750	1 184302	FAL2B9V	1750	NA	1750	2 170964	FCL2W6R	1750	
7	2000	1 184302	FAL2B9V	2000	NA	2000	2 107893	FCL2W7H	2000	
8	2250	1 184302	FAL2B9V	2250	0 681318	FBL112C	2250	NA	2250	
9	2500	0 44954	FAL2B9C	2500	0 54486	FBL112C	2500	0 81096	FCL2A8F	2500
10	2750	0 44954	FAL2B9C	2750	0 54486	FBL112C	2750	0 81096	FCL2A8F	2750
11	3000	0 44954	FAL2B9C	3000	0 54486	FBL112C	3000	0 81096	FCL2A8F	3000
12	3250	0 44954	FAL2B9C	3250	0 54486	FBL112C	3250	0 81096	FCL2A8F	3250
13	3500	0 44954	FAL2B9C	3500	1 108145	FBL111C	3500	1 791264	FCL2W7H	3500
14	3750	0 44954	FAL2B9C	3750	1 108145	FBL111C	3750	1 791264	FCL2W7H	3750
15	4000	0 26061	FAL2B9C	4000	1 386789	FBL111C	4000	0 21012	FCL2A8F	4000
16	4250	0 15307	FAL2B9C	4250	1 80864	FBL111C	4250	0 21012	FCL2A8F	4250
17	4500	0 15307	FAL2B9C	4500	1 80864	FBL111C	4500	0 21012	FCL2A8F	4500
18	4750	0 15307	FAL2B9C	4750	1 80864	FBL111C	4750	0 21012	FCL2A8F	4750
19	5000	0 59036	FAL2B9C	5000	2 148652	FBL110C	5000	0 226204	FCL2W6R	5000
20	5250	0 59036	FAL2B9C	5250	2 148652	FBL110C	5250	0 226204	FCL2W6R	5250
21	5500	0 59036	FAL2B9C	5500	2 148652	FBL110C	5500	0 226204	FCL2W6R	5500
22	5750	0 59036	FAL2B9C	5750	2 148652	FBL110C	5750	0 226204	FCL2W6R	5750
23	6000	0 59036	FAL2B9C	6000	2 148652	FBL110C	6000	0 226204	FCL2W6R	6000
24	6250	0 59036	FAL2B9C	6250	2 148652	FBL110C	6250	0 226204	FCL2W6R	6250
25	6500	0 213352	FAL2B9C	6500	2 148652	FBL110C	6500	0 226204	FCL2W6R	6500
26	6750	0 137488	FAL2B9C	6750	2 276372	FBL109F	6750	0 00574	FCL2A7F	6750
27	7000	0 137488	FAL2B9C	7000	2 276372	FBL109F	7000	0 00574	FCL2A7F	7000
28	7250	0 137488	FAL2B9C	7250	2 276372	FBL109F	7250	0 00574	FCL2A7F	7250
29	7500	0 137488	FAL2B9C	7500	2 276372	FBL109F	7500	0 00574	FCL2A7F	7500
30	7750	0 081639	FAL2B7C	7750	1 223789	FBL109F	7750	0 08113	FCL2A7F	7750
31	8000	0 081639	FAL2B7C	8000	1 223789	FBL109F	8000	0 08113	FCL2A7F	8000
32	8250	0 081639	FAL2B7C	8250	1 159522	FBL109F	8250	0 1342	FCL2A7F	8250
33	8500	0 081639	FAL2B7C	8500	0 081639	FBL109F	8500	0 1342	FCL2A7F	8500
34	8750	0 073898	FAL2B7C	8750	0 030717	FBL109A	8750	0 19843	FCL2A7F	8750
35	9000	0 073898	FAL2B7C	9000	0 301174	FBL109A	9000	0 19843	FCL2A7F	9000
36	9250	0 073898	FAL2B7C	9250	0 37205	FBL109A	9250	0 2395	FCL2A8C	9250
37	9500	0 073898	FAL2B7C	9500	0 220972	FBL109A	9500	0 26604	FCL2A8C	9500
38	9750	0 222884	FAL2B7C	9750	0 220972	FBL109A	9750	0 26604	FCL2A8C	9750
39	10000	0 222884	FAL2B7C	10000	0 33063	FBL109A	10000	0 26604	FCL2A8C	10000
40	10250	0 222884	FAL2B7C	10250	0 216194	FBL109A	10250	0 19819	FCL2A8C	10250
41	10500	0 222884	FAL2B7C	10500	0 264337	FBL107C	10500	1 90029	FCL2B9A	10500
42	10750	0 222884	FAL2B7C	10750	0 264337	FBL107C	10750	1 90029	FCL2B9A	10750

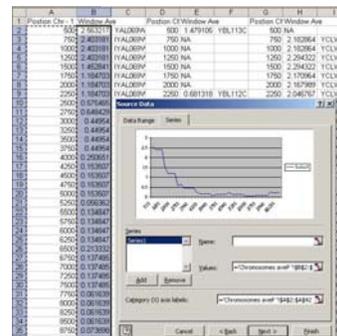


Figure 4. P-value vs Average Window log ratio chart (only for permutation simulations)– Sheet "P-value Histogram" contains the results from the permutation simulation.

Step 1) Select Average log ratio and p-value (cols A and B)

Step 2) Insert xy scatter chart

Step 3) Label chart as desired

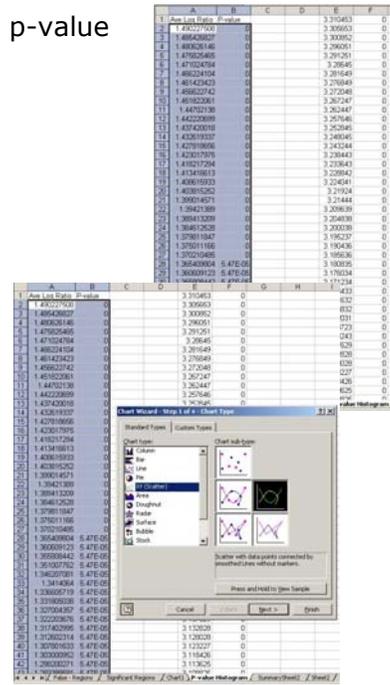
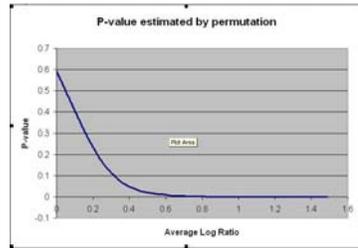


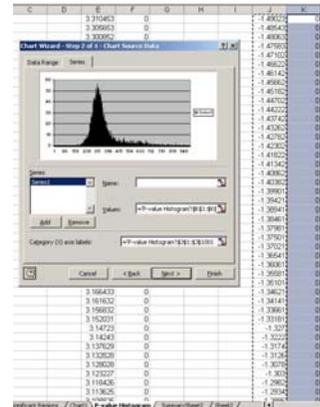
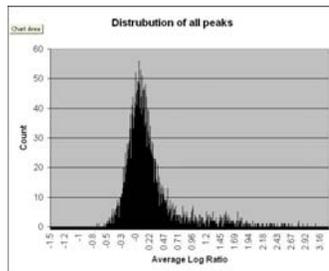
Figure 5. Distribution of all peaks – Sheet "P-value Histogram" contains the height of all peaks found in the experiment.

Step 1) Select column K

Step 2) Insert column chart

Step 3) Select column J for category X-axis

Step 4) Label chart as desired



Picking a significance criterion

ChIPOTle has three options for determining the significance of enrichment found in ChIP-chip experiment.

1. Peak height cutoff
Any peak with a height above the average \log_2 ratio inputted will be saved in the Significant Regions and the Peaks worksheets. This approach does not estimate the p-value for each window or peak.
2. Background Gaussian distribution
The background or non-enriched population is assumed to a symmetric Gaussian distribution about the mean of zero. For most ChIP-chip datasets this is the case but is not true for all experiments. See "Is my data Gaussian" below if your not sure if you data fits this assumption. Using the Gaussian distribution is the most powerful approach in ChIPOTle for estimating the p-value of enrichment. Under the null hypothesis, the distribution of the average \log_2 ratio within each window is again Gaussian, with mean zero and variance equal to the variance of a single log ratio divided by the number of elements in the window. Thus the nominal p-value for a window with average ratio w can be calculated using the standard error function (ERF) as follows:

$$(1) \quad P_{window} = 1 - ERF\left(\frac{\bar{w}}{\sigma/\sqrt{n}}\right)$$

where σ is the standard deviation for the background distribution, and n is the number of microarray elements used in the window. The p-values reported by ChIPOTle are corrected for multiple comparisons using the conservative Bonferroni correction.

3. Estimate background using permutation
The background or non-enriched population is assumed only to be symmetric about the mean of zero. This approach only looks for peaks in the sliding window averages and does not estimate a p-value for every window. In addition the p-values for peaks are not correct for multiple testing. Therefore, ChIPOTle includes an additional output sheet FDR which contains the false discovery rate statistics. The peaks are identified from the data as any window or group of windows with the same value having a preceding and following window of a lower value. Only these peaks will be tested for enrichment, reducing the total number of statistical test required. The significance of enrichment for a peak is estimated by comparing it's height to the height of peaks caused by chance (non-enriched). The height of peaks caused by chance is estimated by a permutation simulation of all non-enriched regions. Since, ChIP-chip experiments do not specifically deplete any genomic fragments, any array element or peak with negative log ratio can be assumed to belong to the non-enriched population. With the assumption of symmetry about the mean for the non-enriched population we can estimate the complete non-enriched population by reflecting the negative distribution onto the positive axis. For example, a negative peak of depth - 0.5, which should occur only by chance, will occur as often as a positive peak of height 0.5 by chance. From this distribution of the non-enriched positive

peaks, CHIPOTLE estimates the probability of enrichment for each peak found.

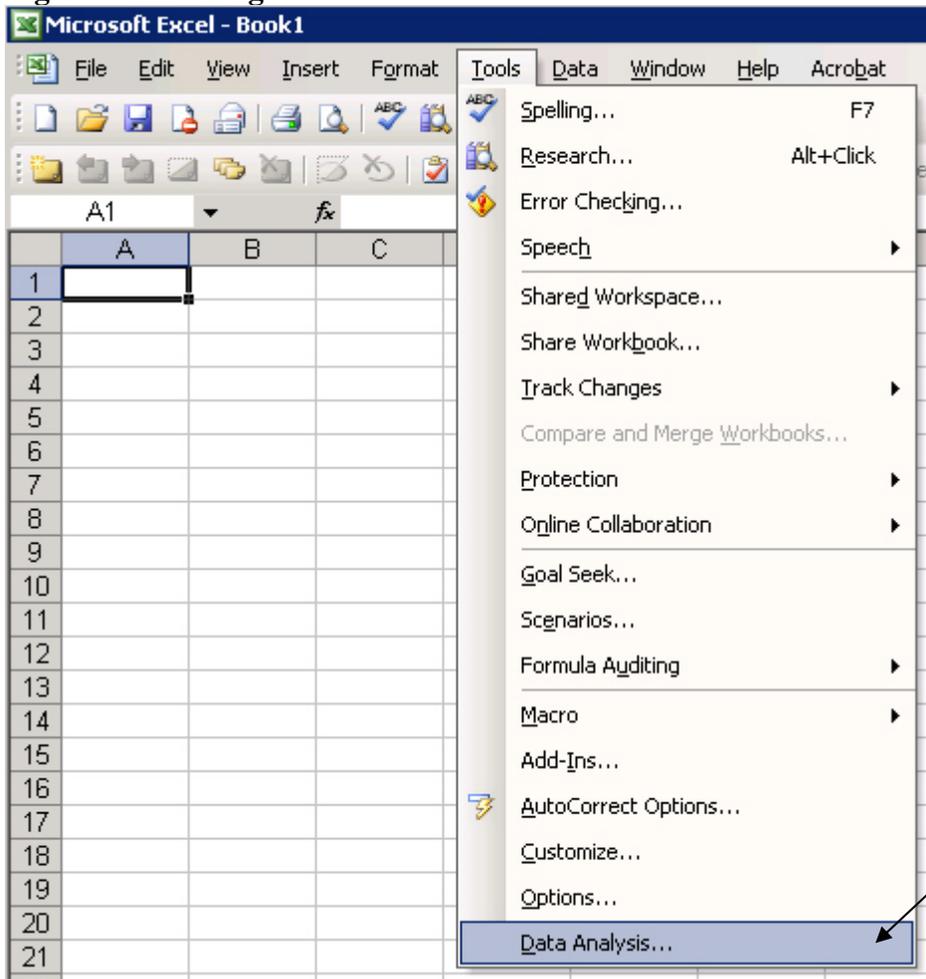
First the genomic order of the data is randomized and then a sliding window average is determined with the user specifications. Negative peaks are determined and their depth's counted. This is repeated a selected (default = 100) number of times and the distribution of the peaks is used to determine the p-value for enrichment.

Is my data Gaussian?

The quick and easy check using Microsoft excel data "Analysis ToolPak". The steps below demonstrate how to make a plot similar to a Q-Q plot in microsoft excel.

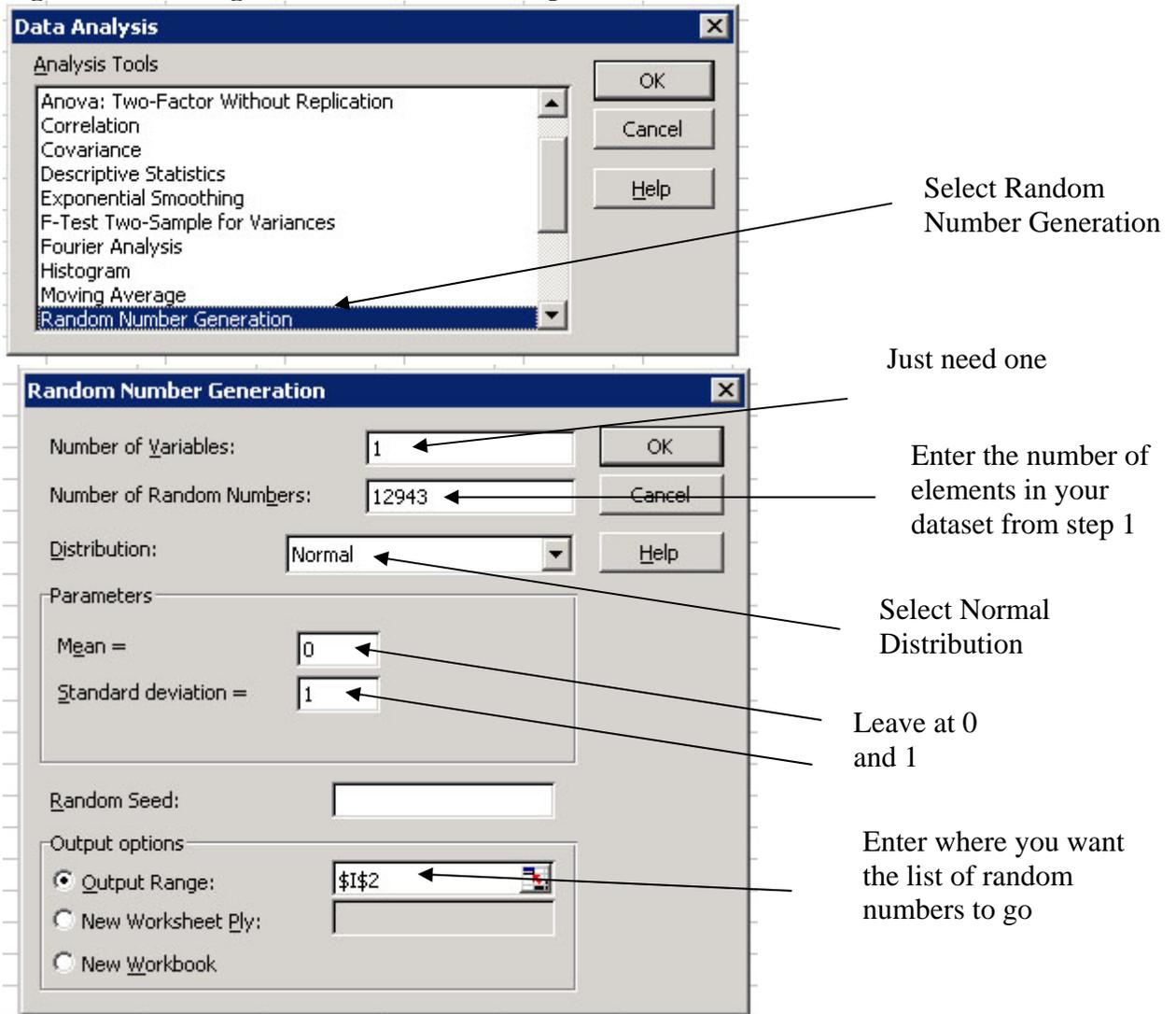
1. Count the number of elements in your dataset.
2. Create a list of random number from a Normal distribution using data analysis toolpak addin.

Figure 6. Creating random Gaussian data



Select data analysis, if you don't have this option make sure you have the ToolPak add-in installed.

Figure 7. Creating random Gaussian data part 2



3. Sort your dataset and the random number dataset individually in ascending order

Figure 8. Sorting a list

The screenshot shows the Microsoft Excel interface with the Data menu open. The 'Sort...' option is highlighted. Below the menu, the Sort dialog box is displayed. The 'Sort by' dropdown is set to 'real', and the 'Ascending' radio button is selected. The 'Then by' dropdowns are empty. The 'My data range has' section has 'Header row' selected. The dialog box has 'Options...', 'OK', and 'Cancel' buttons.

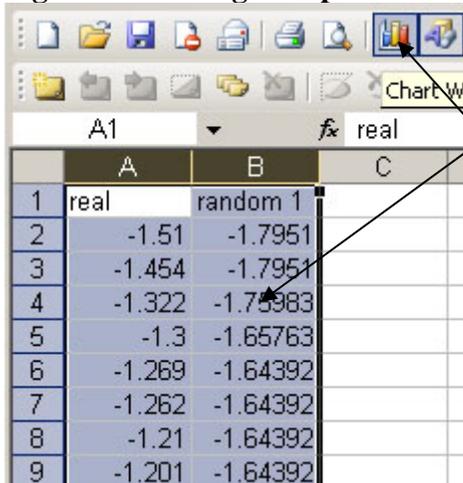
Select Sort from the Data menu

Make sure you do one column at a time

	A	B	C	D
1	real	random 1		
2	0.21	-0.13981		
3	-0.375	-0.60997		
4	0.151	0.122088		
5	0.195	0.618584		
6	0.733	0.581006		
7	-0.122	0.838237		
8	0.162	-1.04571		
9	1.253	-0.10804		
10	-0.18	0.531306		
11	0	-0.5181		
12	1.373	-0.32739		
13	-0.264	-0.8085		
14	0.305	-0.88376		
15	0.18	-0.46564		

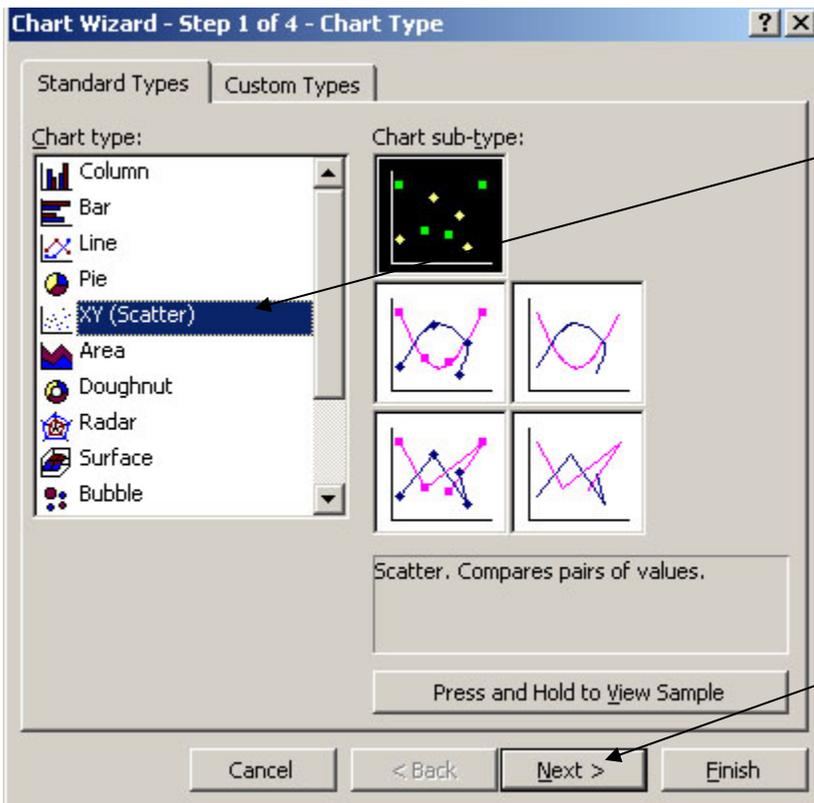
4. Make a x-y plot of the two data sets

Figure 9. Making X-Y plot



Select both columns

Then hit chart wizard



Select X-Y plot

Select next then fill in what ever you want

5. Interpret the plot! (Figure 10)

- a. Try drawing a line for the linear region of the plot? If there is not a linear region your data is not Gaussian. It may have a bimodal distribution depending on percentage of arrayed elements enriched in the IP. When you enrich greater than > 20 % of the arrayed elements the data distribution is more bimodal then normal.

- b. Is there a heavy skew to the left? Are there many spots above the line in the bottom left of the chart? If there is a heavy skew on the left side of the distribution then the Gaussian assumption may be too liberal. Depending on how heavy the tail is you may want to use the permutation simulation approach.
- c. Does the line intersect (0,0)? If not the data may need to be normalized or centered. A slight deviation < 0.05 from (0,0) is ok, but too much will invalidate the assumption of symmetry.

Figure 10. Q-Q plot in Excel

