CLEANIT:

An automated procedure for filtering pull-up, complex pull-up, and minus A from an electropherogram

Developed and Written with Contributions By:
Lauren E. Alfonse
Catherine M. Grgicak

Boston University School of Medicine
Program in Biomedical Forensic Sciences
72 East Concord Street, Boston, MA 02118

BOSTON UNIVERSITY

Contents

1. INTRODUCTION.....3

2. FILTERING DATA WITH CLEANIT.....4

2.1 PULL-UP FILTER.....4

2.2 COMPLEX PULL-UP FILTER.....5

2.3 MINUS A FILTER.....7

3. REQUIREMENTS AND PROCEDURE 8

1. Introduction

CleanIt is an automated procedure for filtering pull-up, complex pull-up, and minus A from an electropherogram. It is written in Microsoft® Visual Basic for Applications and should be used in conjunction with data exported from a genotyping software. The user inputs the data to be filtered and selects which artifacts to filter; with minimal hands-on time, common artifacts are then removed according to user-defined parameters. CleanIt has been tested on data generated with commercially available 16- and 24-locus short tandem repeat (STR) amplification kits which amplify tri-, tetra-, and pentanucleotide STR loci. Up to 180 electropherograms have been tested and filtered with CleanIt in one run.

Use of this tool assumes prior training and education in analyzing and exporting STR data. We have provided this User's Guide in order to provide fundamental information and instructions regarding the use of CleanIt. This program is distributed without any warranty. While we have attempted to ensure that the information provided in this document is correct and that CleanIt works properly in most scenarios, CleanIt is offered "as is." Users of information and/or downloadable material implicitly agree to use the information and material at their own risk and responsibility. We recommend that users validate this tool using their own data and check results against other methods.

2. Filtering Data with CleanIt

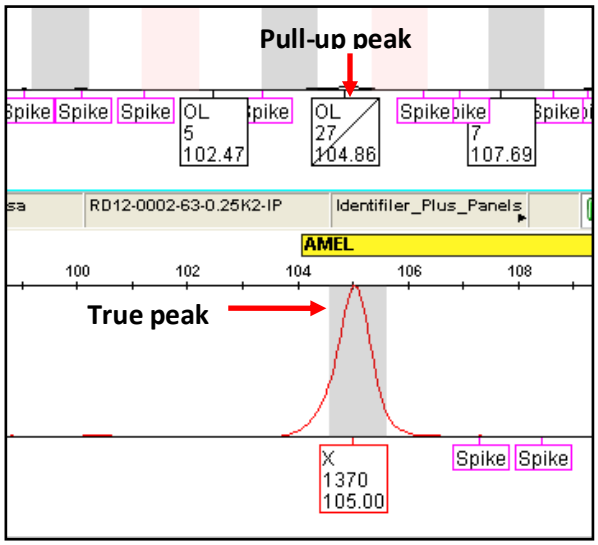
CleanIt seeks to automate the removal of pull-up, complex pull-up, and minus A by applying a distinct set of rules and conditions for each artifact as outlined and defined below.

2.1 Pull-up Filter

In some protocols, removal of pull-up may be performed manually by applying three rules. The first consists of a dye color restriction that only considers peaks that are different colors (i.e. the pull-up peak color cannot be the same as the true peak color). The second applies a size restriction by determining whether the size of the pull-up peak falls within +/- *x* base pairs (bp) of the size of the true peak, where *x* is a decimal value that accounts for small variations in peak sizing. The third applies a height restriction by asking if the height of the pull-up peak divided by the height of the true peak is less than or equal to *y*, where *y* is a ratio determined by the laboratory. If all three criteria are fulfilled, the peak is classified as pull-up and removed from the data.

In CleanIt, these three criteria are transposed to three if/then statements to automate filtering of pull-up. Note that the size restriction of +/- *x* bp corresponds to the Pull-up Size Range, and the height restriction of less than/equal to *y* is the Pull-up Height Ratio; both are variables which the user must specify in the CleanIt interface.

An example of a pull-up peak is pictured in a segment of a DNA profile featured at right. If the pull-up size range is set to 0.3 bp and the pull-up height ratio is set to 0.06, the process automated by CleanIt would proceed as follows:



1. Color of pull-up peak=yellow (shown as black in image) , Color of true peak=red

   Because the color of the pull-up peak is not the same as the color of the true peak, the dye color criteria is fulfilled, and the next if/then statement is considered.

2. Size of pull-up peak=104.86, Size of true peak=105.00

   Because 105.00-104.86 = 0.14, the size requirement of +/-0.6 is fulfilled, and the next if/then statement is considered. Note that the pull-up size range is automatically doubled to account for the scenario in which one peak has positive sizing error (i.e. +0.3), and the other peak has negative sizing error (i.e. -0.3).

3. Height of pull-up peak=27, Height of true peak=1370

   Because 27/1370 = 0.0197, the height requirement of less than/equal to 0.06 is fulfilled, and the pull-up peak is removed.

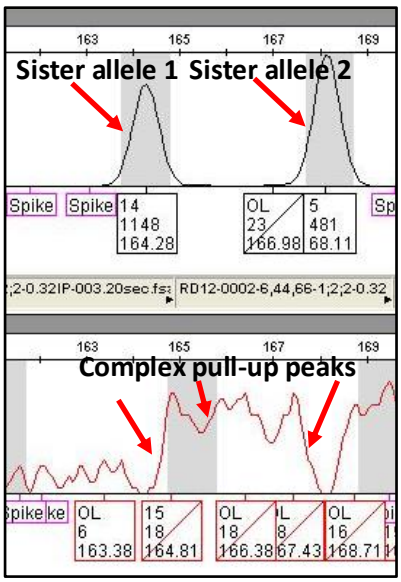
2.2 Complex Pull-up Filter

The manual filtering of complex pull-up is often subjective and based upon the observance of a plateau-shaped peak(s) appearing between two sister alleles in a different dye channel.

CleanIt was designed to filter complex pull-up that fulfills the following rules sometimes employed during manual filtering: (1) The observance of two sister alleles at the same locus within one repeat unit apart. (2) The two sister alleles must be greater than or equal to *x* of each other in height, where *x* is a ratio determined by the laboratory. (3) The complex pull-up peak(s) must appear in a dye channel different from that of the sister alleles. (4) The size of the complex pull-up peak (in bp) must fall between the sizes of the two sister alleles, +/- *y* bp where *y* is a decimal value that accounts for small variations in peak sizing. (5) The height of the complex pull-up peak(s) divided by the height of the shorter sister allele must be less than or equal to *z*, where *z* is a threshold set by the laboratory. If all five criteria are fulfilled, the peak may be classified as complex pull-up and removed from the data.

In CleanIt, these criteria are transposed to multiple if/then statements to automate filtering of complex pull-up. Note the following pertaining to the CleanIt interface: the user must specify the repeat unit length of the sister alleles in the field Max STR Repeat Unit Length, the sister allele height restriction of greater than or equal to *x* corresponds to the Sister Allele Peak Height Ratio, the +/- *y* bp size restriction of the complex pull-up peak(s) corresponds to the Complex Pull-up Size Range, and the less than/equal to *z* height ratio corresponds to the Complex Pull-up Height Ratio.

An example of a complex pull-up peak(s) is featured in the segment of a DNA profile shown below. Given that the maximum STR repeat unit length is set to 4 bp, the sister allele peak height ratio is set to 0.5, the complex pull-up size range is set to 0.3 bp, and the complex pull-up height ratio is set to 0.06, the process automated by CleanIt would proceed as follows:



1. Size of sister allele 1=164.28
   Size of sister allele 2=168.11
   Because 168.11–164.28 = 3.83, the size requirement of 4 +/- 0.6 is fulfilled (the alleles are within one repeat unit apart), and the next if/then statement is considered. Note that the decimal portion of the size range is automatically doubled to account for the scenario in which one peak has positive sizing error (i.e. +0.3), and the other peak has negative sizing error (i.e. -0.3).

2. Height of sister allele 1=1148
   Height of sister allele 2=1481
   Because 1148/1481=0.78, the sister allele peak height ratio of greater than/equal to 0.5 is fulfilled, and the next if/then statement is considered.

3. Color of sister alleles=yellow (shown in image as black)
   Color of complex pull-up peak(s)=red
   Because the color of the complex pull-up peak(s) is not the same as the color of the sister allele peaks, the dye channel criteria is fulfilled, and the next if/then statement is considered.

4. Size of complex pull-up peak=164.81
   Range of sister alleles sizes=164.28 (-0.3) to 168.11 (+0.3)
   Because the size of the complex pull-up peak falls within the range of the sizes of the two sister alleles, the size criteria is fulfilled, and the next if/then statement is considered. Note that three total complex pull-up peaks are shown, and the sizes of the remaining two complex pull-up peaks also fall within the size range.

5. Height of complex pull-up peak=18
   Height of shorter sister allele=1148
   Because 18/1148=0.016, the complex pull-up height ratio of less than/equal to 0.06 is fulfilled, and the peak would be removed as complex-pull up. Note that three total complex pull-up peaks are shown, and the height ratios of the remaining two complex pull-up peaks are also less than/equal to 0.06. Given the criteria articulated, all three complex pull-up peaks indicated on the image above would be filtered and removed as complex pull-up.

2.3 Minus A Filter

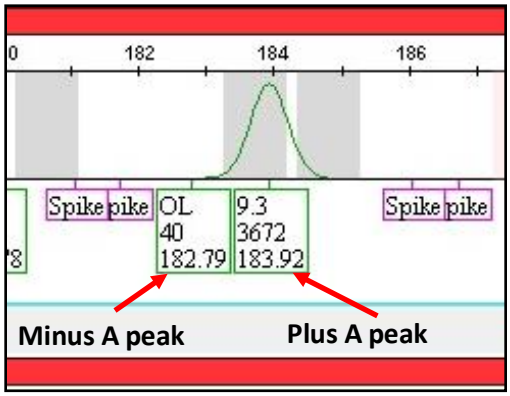
Removal of minus A may be performed manually by applying two "rules." The first rule applies a size restriction by asking if the size of the minus A peak falls within -1 +/- *x* bp of the size of the plus A peak (note that -1 is a constant, while the term following +/- can vary). The second applies a height restriction by asking if the height of the minus A peak divided by the height of the plus A peak is less than or equal to *y*, where *y* is a ratio determined by the laboratory. If these two criteria are fulfilled, the peak is classified as minus A and removed from the data.

In CleanIt, these two criteria are transposed to two if/then statements to automate removal of minus A. Note that the size restriction of +/- *x* bp corresponds to the Minus A Size Range, and the height restriction of less than/equal to *y* is the Pull-up Height Ratio; both are variables which the user must specify in the CleanIt interface.

An example of a minus A peak is pictured in a segment of a DNA profile shown below. If the minus A size range is set to +/- 0.3 bp and the minus A height ratio is set to 0.06, the process automated by CleanIt would proceed as follows:

1. Size of minus A peak=182.79
   Size of plus A peak=183.92

   Because 183.92 - 182.79 =1.13, the size requirement of -1 +/-0.6 is fulfilled, and the next if/then statement is considered. Note that the pull-up size range is automatically doubled to account for the scenario in which one peak has positive sizing error (i.e. +0.3), and the other peak has negative sizing error (i.e. -0.3).



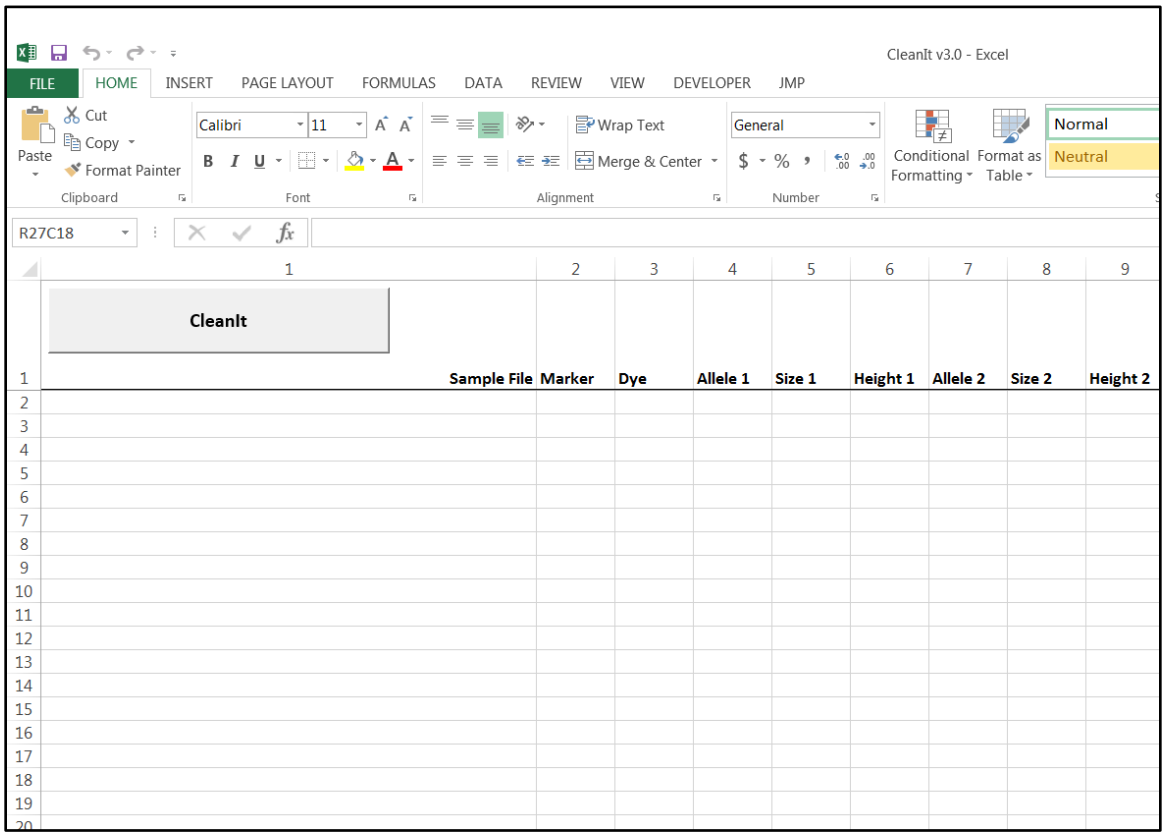
2. Height of minus A peak=40
   Height of plus A peak=3672

   Because 40/3672 = 0.011, the height requirement of less than/equal to 0.06 is fulfilled, and the minus A peak is removed.

3. Requirements and Procedure

1. CleanIt is written in Microsoft® Visual Basic for Applications. Users must have Microsoft® Excel installed with Macros enabled. Instructions to enable Macros may be found here.

2. Download CleanIt v3.1 from Boston University's Laboratory for Forensic Research and Technology Development website at the following link: http://sites.bu.edu/grgicak/software/. Open CleanIt by double-clicking on the Microsoft® Excel icon. The window that opens should appear as follows:



3. Export (or obtain) data to be filtered from the genotyping software of choice, and open the exported file. The data must follow the format of the row one header shown below, with general information about the sample appearing the first column, followed by the marker/locus and corresponding dye, and the allele call, size, and height for all peaks observed at each locus. For a given dataset to be filtered using CleanIt, **all samples must contain the same number of loci**. Data should be grouped first by sample, then sorted by dye within a given sample. As shown in the example below, Sample 1 and Sample 2 both contain 16 loci, and within each sample, the loci labeled with a blue ("B") dye are grouped together, followed by green ("G"), yellow ("Y"), and red ("R") loci.

	1	2	3	4	5	6	7	8	9	10	11	12
1	**Sample File**	**Marker**	**Dye**	**Allele 1**	**Size 1**	**Height 1**	**Allele 2**	**Size 2**	**Height 2**	**Allele 3**	**Size 3**	**Height 3**
2	Sample 1	D8S1179	B	OL	113.92	3	OL	115.38	3	OL	116.84	3
3	Sample 1	D21S11	B	24	182.41	30	25.2	188.03	5	26	190.62	7
4	Sample 1	D7S820	B	OL	249.26	2	7	258.22	4	OL	258.93	4
5	Sample 1	CSF1PO	B	5	298.81	3	OL	305.63	3	7	306.73	3
6	Sample 1	D3S1358	G	OL	94.97	7	OL	96.04	11	OL	98.68	2
7	Sample 1	TH01	G	OL	159.65	4	OL	170.28	10	7	172.9	7
8	Sample 1	D13S317	G	OL	207.01	4	OL	217.54	6	OL	218.63	3
9	Sample 1	D16S539	G	OL	254.12	2	7	259.38	43	8	263.3	1411
10	Sample 1	D2S1338	G	14	300.27	5	15	304.89	78	16	308.85	1616
11	Sample 1	D19S433	Y	OL	98.93	1	OL	104.77	13	12	112.09	9
12	Sample 1	vWA	Y	OL	147.28	12	OL	148.59	22	OL	161.65	3
13	Sample 1	TPOX	Y	OL	212.94	17	OL	215.18	9	6	220.9	5
14	Sample 1	D18S51	Y	OL	262.1	5	OL	263.3	55	OL	267.43	6
15	Sample 1	AMEL	R	X	105.02	1426	OL	107.6	11	Y	110.93	1121
16	Sample 1	D5S818	R	OL	127.31	8	9	140.6	19	10	144.85	146
17	Sample 1	FGA	R	OL	205.26	6	18	216.72	7	20	225.14	72
18	Sample 2	D8S1179	B	OL	113.92	3	OL	114.88	2	OL	120.64	7
19	Sample 2	D21S11	B	OL	181.67	30	24	182.64	33	OL	185.81	4
20	Sample 2	D7S820	B	6	254.17	7	OL	260.09	4	8	262.07	6
21	Sample 2	CSF1PO	B	5	298.57	3	OL	307.68	2	9	314.48	73
22	Sample 2	D3S1358	G	OL	94.88	12	OL	105.04	7	OL	107.02	4
23	Sample 2	TH01	G	OL	155.74	11	OL	157.21	7	OL	163.08	4
24	Sample 2	D13S317	G	8	215.25	8	9	219.37	94	10	223.36	1649
25	Sample 2	D16S539	G	OL	257.44	4	8	263.33	84	9	267.4	2470
26	Sample 2	D2S1338	G	15	304.89	63	16	308.77	1424	18	316.89	116
27	Sample 2	D19S433	Y	OL	95.37	16	12	112.28	141	13	116.32	2284
28	Sample 2	vWA	Y	OL	148.39	5	OL	152.79	9	OL	154.96	8
29	Sample 2	TPOX	Y	OL	214.09	7	7	224.78	36	8	228.78	1503
30	Sample 2	D18S51	Y	OL	267.4	93	10	273.46	8	OL	274.46	5
31	Sample 2	AMEL	R	X	105.04	3066						
32	Sample 2	D5S818	R	OL	129.73	7	11	148.91	16	OL	149.77	9
33	Sample 2	FGA	R	17	212.35	7	18	217.4	5	19	220.88	10

4. Select and copy all rows of exported data **except the header row** (row one). Return to the CleanIt window, and paste the data into row two, as shown below. **DO NOT MODIFY THE FIRST ROW IN CLEANIT.**

CleanIt

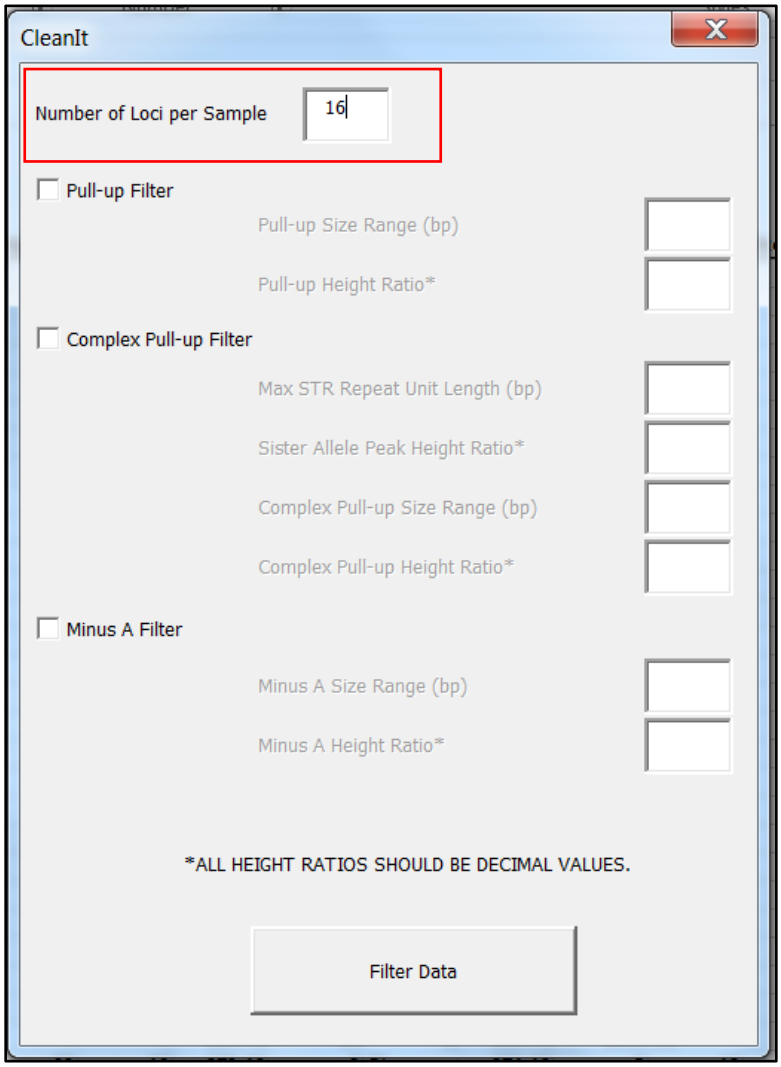
	Sample File	Marker	Dye	Allele 1	Size 1	Height 1	Allele 2	Size 2	Height 2	Allele 3	Size 3	Height 3
2	Sample 1	D8S1179	B	OL	113.92	3	OL	115.38	3	OL	116.84	3
3	Sample 1	D21S11	B	24	182.41	30	25.2	188.03	5	26	190.62	7
4	Sample 1	D7S820	B	OL	249.26	2	7	258.22	4	OL	258.93	4
5	Sample 1	CSF1PO	B	5	298.81	3	OL	305.63	3	7	306.73	3
6	Sample 1	D3S1358	G	OL	94.97	7	OL	96.04	11	OL	98.68	2
7	Sample 1	TH01	G	OL	159.65	4	OL	170.28	10	7	172.9	7
8	Sample 1	D13S317	G	OL	207.01	4	OL	217.54	6	OL	218.63	3
9	Sample 1	D16S539	G	OL	254.12	2	7	259.38	43	8	263.3	1411
10	Sample 1	D2S1338	G	14	300.27	5	15	304.89	78	16	308.85	1616
11	Sample 1	D19S433	Y	OL	98.93	1	OL	104.77	13	12	112.09	9
12	Sample 1	vWA	Y	OL	147.28	12	OL	148.59	22	OL	161.65	3
13	Sample 1	TPOX	Y	OL	212.94	17	OL	215.18	9	6	220.9	5
14	Sample 1	D18S51	Y	OL	262.1	5	OL	263.3	55	OL	267.43	6
15	Sample 1	AMEL	R	X	105.02	1426	OL	107.6	11	Y	110.93	1121
16	Sample 1	D5S818	R	OL	127.31	8	9	140.6	19	10	144.85	146
17	Sample 1	FGA	R	OL	205.26	6	18	216.72	7	20	225.14	72
18	Sample 2	D8S1179	B	OL	113.92	3	OL	114.88	2	OL	120.64	7
19	Sample 2	D21S11	B	OL	181.67	30	24	182.64	33	OL	185.81	4
20	Sample 2	D7S820	B	6	254.17	7	OL	260.09	4	8	262.07	6
21	Sample 2	CSF1PO	B	5	298.57	3	OL	307.68	2	9	314.48	73
22	Sample 2	D3S1358	G	OL	94.88	12	OL	105.04	7	OL	107.02	4
23	Sample 2	TH01	G	OL	155.74	11	OL	157.21	7	OL	163.08	4
24	Sample 2	D13S317	G	8	215.25	8	9	219.37	94	10	223.36	1649
25	Sample 2	D16S539	G	OL	257.44	4	8	263.33	84	9	267.4	2470
26	Sample 2	D2S1338	G	15	304.89	63	16	308.77	1424	18	316.89	116
27	Sample 2	D19S433	Y	OL	95.37	16	12	112.28	141	13	116.32	2284
28	Sample 2	vWA	Y	OL	148.39	5	OL	152.79	9	OL	154.96	8
29	Sample 2	TPOX	Y	OL	214.09	7	7	224.78	36	8	228.78	1503
30	Sample 2	D18S51	Y	OL	267.4	93	10	273.46	8	OL	274.46	5
31	Sample 2	AMEL	R	X	105.04	3066						
32	Sample 2	D5S818	R	OL	129.73	7	11	148.91	16	OL	149.77	9
33	Sample 2	FGA	R	17	212.35	7	18	217.4	5	19	220.88	10

5. Click on the button labeled “CleanIt” in the top left-hand corner of the window.

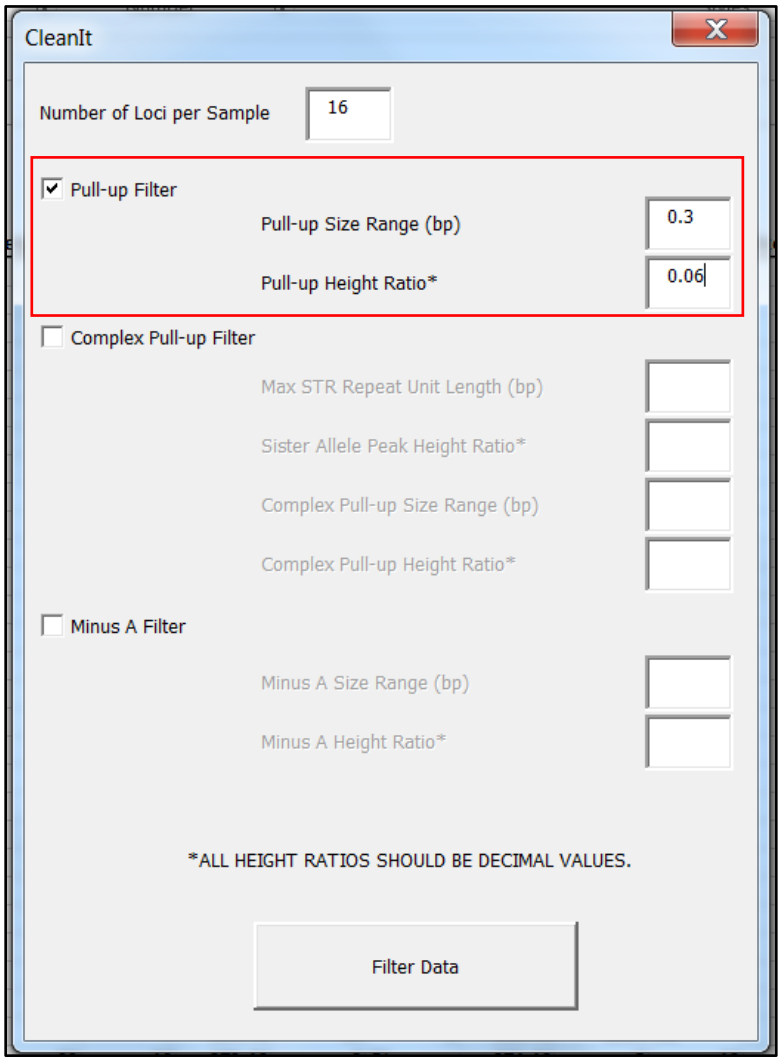
CleanIt

	Sample File	Marker	Dye	Allele 1	Size 1	Height 1
2	Sample 1	D8S1179	B	OL	113.92	3
3	Sample 1	D21S11	B	24	182.41	30
4	Sample 1	D7S820	B	OL	249.26	2
5	Sample 1	CSF1PO	B	5	298.81	3
6	Sample 1	D3S1358	G	OL	94.97	7
7	Sample 1	TH01	G	OL	159.65	4

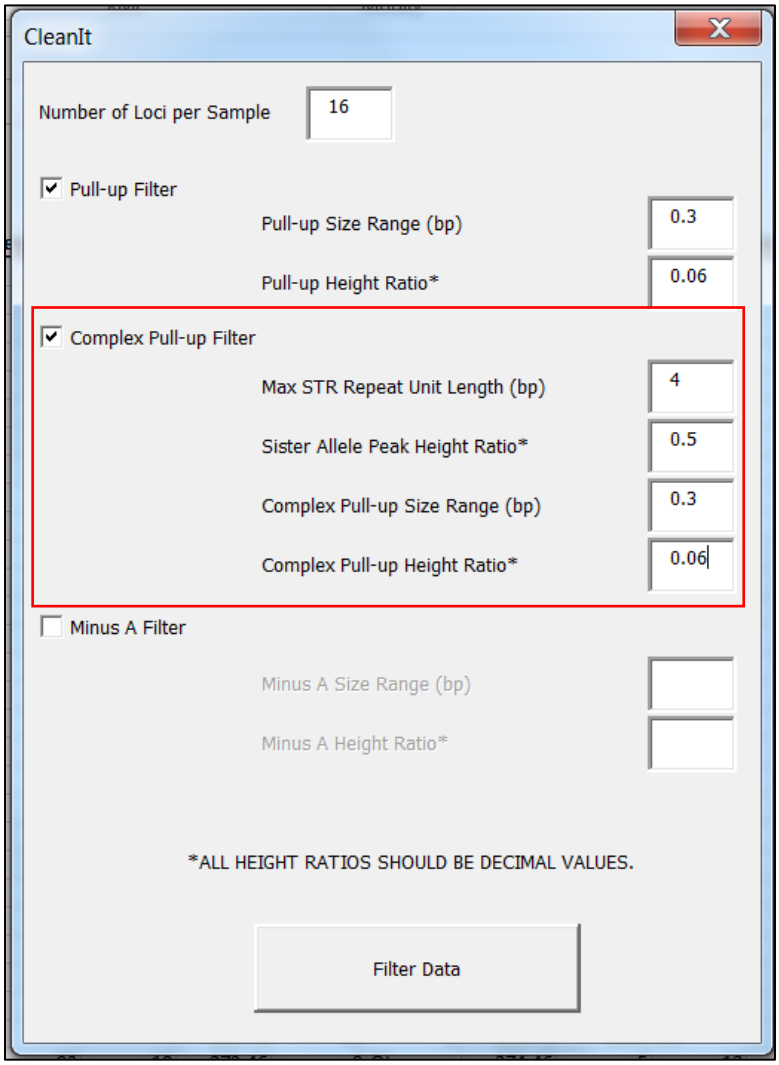
6. In the interface that appears, first enter the number of loci in each sample by clicking on the input box next to "Number of Loci per Sample," and typing the desired value, shown as "16" in the example below. **Note that this field cannot be left blank,** the number of loci should be a positive, whole number, and all samples being filtered should contain the same number of loci.



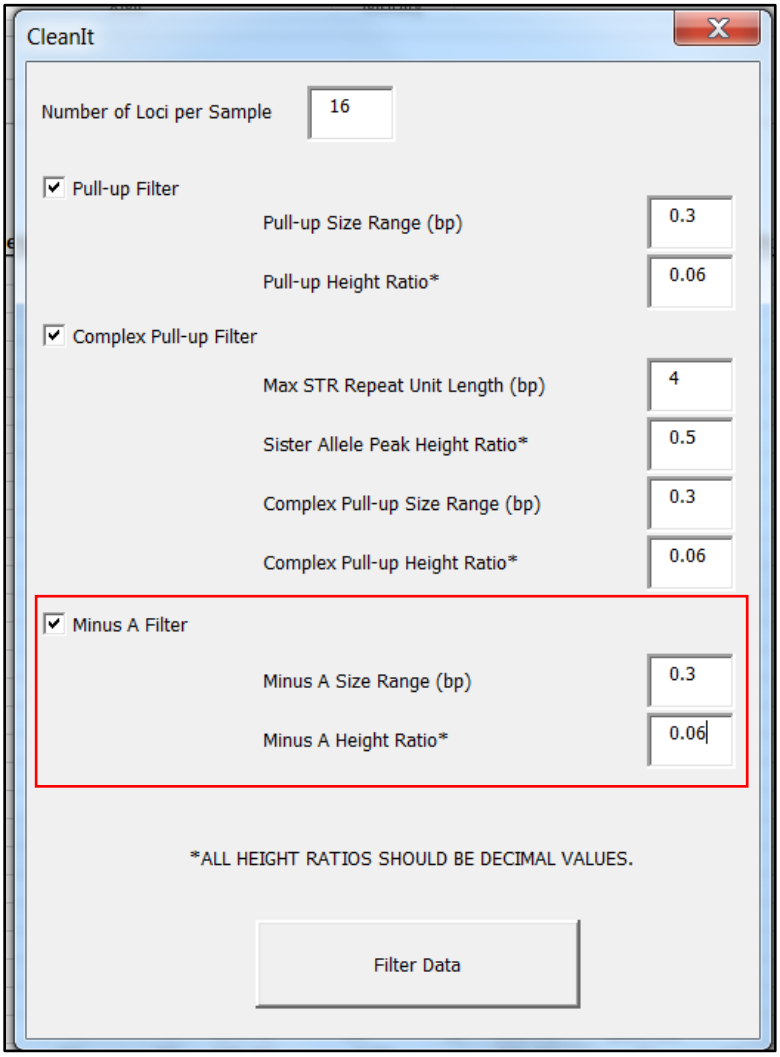
7. In the same interface, the user must then decide whether or not to run the pull-up filter. If the user does not want to filter pull-up, the check-box to the left of "Pull-up Filter" should be un-checked. If the user would like to filter pull-up, check the box to the left of "Pull-up Filter." The fields "Pull-up Size Range (bp)" and "Pull-up Height Ratio" should now be enabled, and **the user must provide values for both.** The pull-up size range, shown in the example below as "0.3," should be a positive, decimal value. Similarly, the pull-up height ratio, shown in the example below as "0.06," should also be a positive, decimal value. **Do not enter the pull-up height ratio as a percentage.**



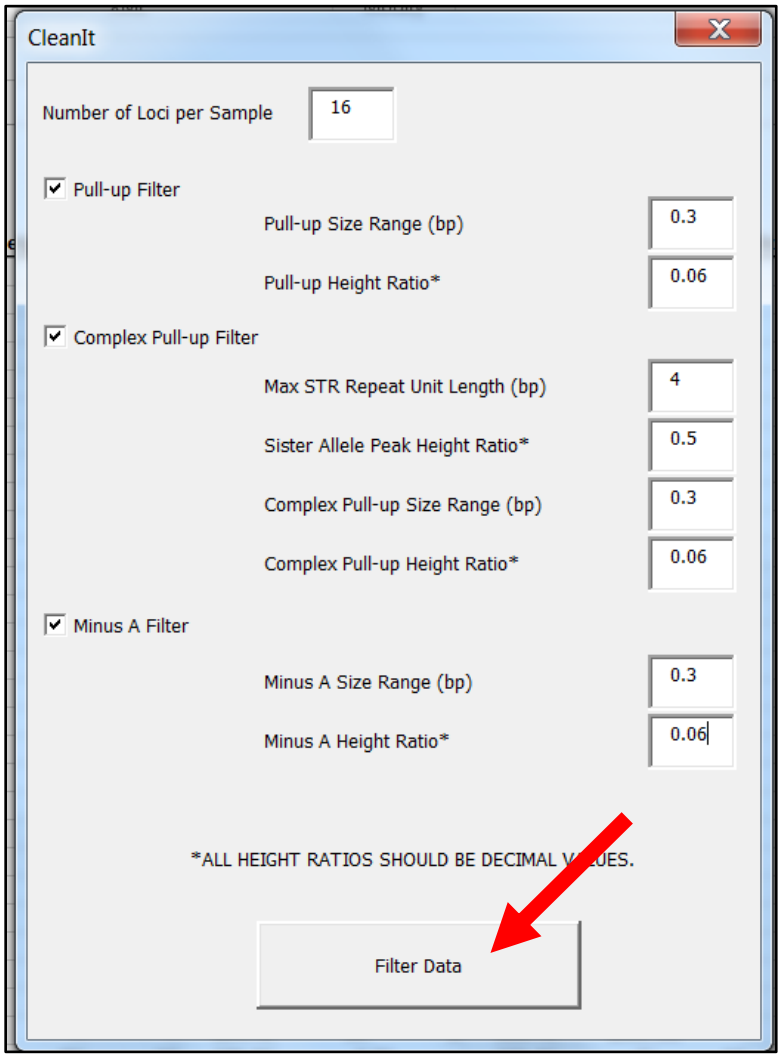
8. In the same interface, next decide whether or not to run the complex pull-up filter. If the user does not want to filter complex pull-up, the check-box to the left of "Complex Pull-up Filter" should be un-checked. If the user would like to filter complex pull-up, check the box to the left of "Complex Pull-up Filter." The fields "Max STR Repeat Unit Length (bp)," "Sister Allele Peak Height Ratio," "Complex Pull-up Size Range (bp)," and "Complex Pull-up Height Ratio" should now be enabled, and **the user must provide values for all.** The maximum STR repeat unit length, shown below as "4," should be a positive, whole number. The sister allele peak height ratio ("0.5"), complex pull-up size range ("0.3"), and complex pull-up height ratio ("0.06"), should all be positive, decimal values. **Height ratios must not be entered as percentages.**



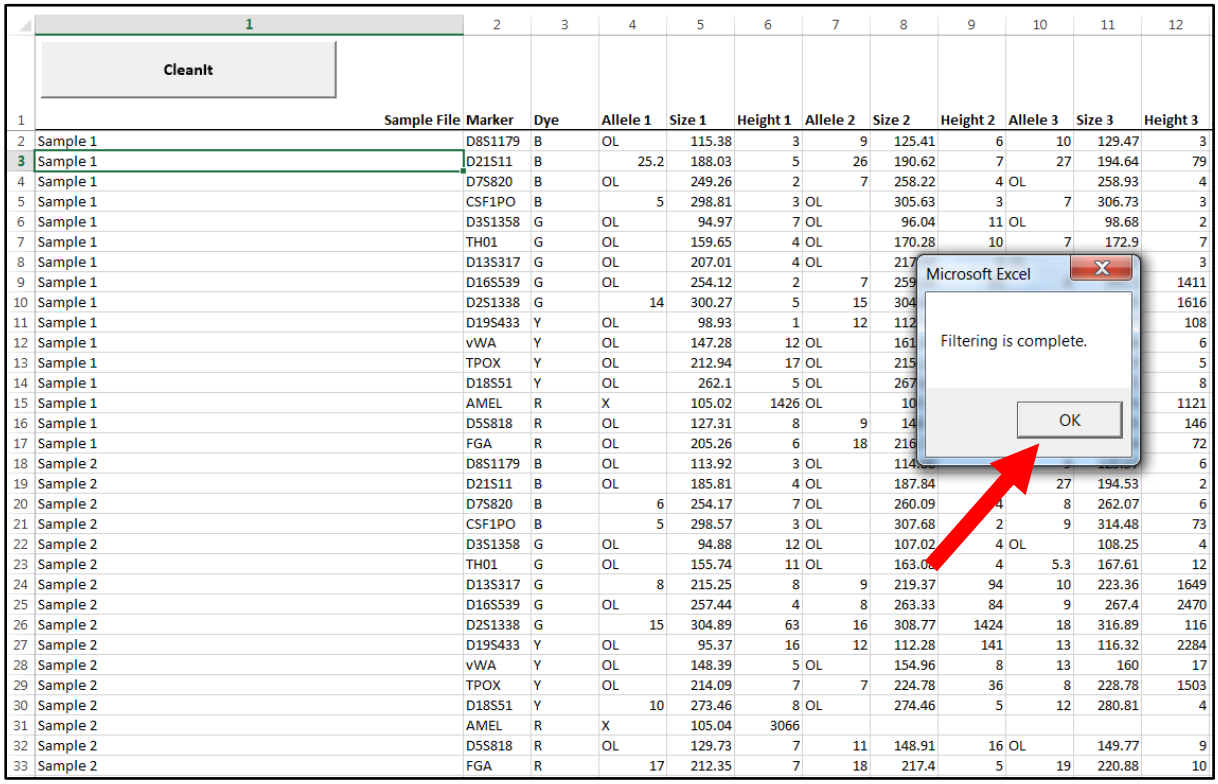
9. In the same interface, decide whether or not to run the minus A filter. If the user does not want to filter minus A, the check-box to the left of “Minus A Filter” should be un-checked. If the user would like to filter minus A, check the box to the left of “Minus A Filter.” The fields “Minus A Size range (bp)” and “Minus A Height Ratio” should now be enabled, and the user must specify values for both. The minus A size range (“0.3”) and the minus A height ratio (“0.06”) should both be positive, decimal values. **The height ratio must not be entered as a percentage.**



10. To run the selected filters, press button labeled “Filter Data.” The interface should disappear, and the filtering code will begin to run in the background. The Microsoft® Excel window may appear frozen and/or unresponsive during this time. This is routine, and Microsoft® Excel will become responsive once filtering is complete. In addition, it is important to note that CleanIt run-time increases with increasing sample number. The developers recommend testing CleanIt on one or a few samples at first use to be able to gauge run-times, as the run-time may vary from system to system. For large datasets, it may be beneficial to run CleanIt overnight on a computer that will not be used.



11. The message box shown below will appear when filtering is complete. Press “OK.” **The filtered data should now be copied to a new spreadsheet and saved.** The user can now exit CleanIt or clear the existing data, input new data, and repeat the procedure as desired.



	1	2	3	4	5	6	7	8	9	10	11	12
1	Sample File	Marker	Dye	Allele 1	Size 1	Height 1	Allele 2	Size 2	Height 2	Allele 3	Size 3	Height 3
2	Sample 1	D8S1179	B	OL	115.38	3	9	125.41	6	10	129.47	3
3	Sample 1	D21S11	B	25.2	188.03	5	26	190.62	7	27	194.64	79
4	Sample 1	D7S820	B	OL	249.26	2	7	258.22	4	OL	258.93	4
5	Sample 1	CSF1PO	B	5	298.81	3	OL	305.63	3	7	306.73	3
6	Sample 1	D3S1358	G	OL	94.97	7	OL	96.04	11	OL	98.68	2
7	Sample 1	TH01	G	OL	159.65	4	OL	170.28	10	7	172.9	7
8	Sample 1	D13S317	G	OL	207.01	4	OL	217				3
9	Sample 1	D16S539	G	OL	254.12	2	7	259				1411
10	Sample 1	D2S1338	G	14	300.27	5	15	304				1616
11	Sample 1	D19S433	Y	OL	98.93	1	12	112				108
12	Sample 1	vWA	Y	OL	147.28	12	OL	161				6
13	Sample 1	TPOX	Y	OL	212.94	17	OL	215				5
14	Sample 1	D18S51	Y	OL	262.1	5	OL	267				8
15	Sample 1	AMEL	R	X	105.02	1426	OL	10				1121
16	Sample 1	D5S818	R	OL	127.31	8	9	14				146
17	Sample 1	FGA	R	OL	205.26	6	18	216				72
18	Sample 2	D8S1179	B	OL	113.92	3	OL	114				6
19	Sample 2	D21S11	B	OL	185.81	4	OL	187.84		27	194.53	2
20	Sample 2	D7S820	B	6	254.17	7	OL	260.09	4	8	262.07	6
21	Sample 2	CSF1PO	B	5	298.57	3	OL	307.68	2	9	314.48	73
22	Sample 2	D3S1358	G	OL	94.88	12	OL	107.02	4	OL	108.25	4
23	Sample 2	TH01	G	OL	155.74	11	OL	163.0	4	5.3	167.61	12
24	Sample 2	D13S317	G	8	215.25	8	9	219.37	94	10	223.36	1649
25	Sample 2	D16S539	G	OL	257.44	4	8	263.33	84	9	267.4	2470
26	Sample 2	D2S1338	G	15	304.89	63	16	308.77	1424	18	316.89	116
27	Sample 2	D19S433	Y	OL	95.37	16	12	112.28	141	13	116.32	2284
28	Sample 2	vWA	Y	OL	148.39	5	OL	154.96	8	13	160	17
29	Sample 2	TPOX	Y	OL	214.09	7	7	224.78	36	8	228.78	1503
30	Sample 2	D18S51	Y	10	273.46	8	OL	274.46	5	12	280.81	4
31	Sample 2	AMEL	R	X	105.04	3066						
32	Sample 2	D5S818	R	OL	129.73	7	11	148.91	16	OL	149.77	9
33	Sample 2	FGA	R	17	212.35	7	18	217.4	5	19	220.88	10