PEDCA Tutorial (Ploidy Estimation by Dynamic Coverage Analysis)

PEDCA is a ploidy estimation algorithm that infers copy number of the contigs submitted as input based on the read coverage that aligns to them. It only requires as an input an alignment file in .bam or .sam format of a library or set of libraries aligned to a reference file of the contigs that will be estimated.

Pre-processing the data (5 steps)

We need to align the reads against a reference.

Step 1. Index your reference.

Example using bwa (all command in one single line):

<*path_to_bwa_aligner*>/bwa index -a bwtsw <*path_to_reference_file/your_reference.fasta*>

Step 2. Align your reads to your reference

Example using bwa and paired end reads (all command in one single line):

<*bwa_aligner_path*>**/bwa mem** <*path_reference_file/your_reference.fasta*> <*readsPath/readsPairEnd1.fasta*> < *readsPath /readsPairEnd2.fasta*> **>** <*destination_folder /example.sam*>

Step 3. You might want to transform your .sam file into a .bam format

Example using samTools (all command in one single line):

<*samToolsPath*> **/samtools view -Sb** <*destination_folder /example.sam*> **>** <*destination_folder /example.bam*>

PEDCA just accepts one input file. If you have several libraries you can put all your bam files in a folder (or create a folder with symbolic links to all files you want to merge) and then:

<*samToolsPath*>**/samtools merge** <*bam_destination_folder*/finalBamFile.bam> ***.bam**

Step 4. Sort the .bam/.sam file

Example sorting a .bam file using samTools (all command in one single line):

<*samToolsPath*> **/samtools sort -o** <*destination_folder* /sorted*_example.bam*> **-O bam -T** <*temp_folderPath/tempName*> <*destination_folder /example.bam*>

Step 5. Index the sorted .bam/.sam file

Example indexing a sorted bam file using samTools (all command in one single line):

<*samToolsPath*> **/samtools index** <*destination_folder* /sorted*_example.bam*>

Using PEDCA

PEDCA has been designed to require minimal parameterization. It works by running a sliding window over the genome and measuring the average depth of coverage inside each bin. Most of its parameters are dependant of the window length and have default values that allows PEDCA to function on contigs > 500 bp and < 2.000 Kbp. Nevertheless, because each genome has its own particular characteristics it is possible to tune in the rest of the parameters. Here is a list of those options and how the influence the output.

At any moment you can obtain the following guide using: **`java –jar PEDCA.jar -help`**
(You can also use '**`–h`**' or '**`help`**')

```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++
PEDCA -help:

USAGE:    java –jar PEDCA.jar  -p <project name> -i < input sam/bam File> -o <output
Folder> <<OPTIONAL PARAMETERS>>

REQUIRED PARAMETERS:
-p (project name)          – (String)  Prefix used to generate the results file.
-i (input file)            - (Pathway to .bam/.sam file)   Pathway to the input file
containing the alignment file. Must be a .bam or .sam file
-o (output Folder)          - (String)   Pathway to the auto-generated output folder
that will contain the results

OPTIONAL PARAMETERS:
-m (multi Run)             - (no parameters) Runs a preselected set of default window
lengths {500,750,1000,2000,3000}
-w (windows length)        - (int)      Length of the sliding window, to measure the
coverage inside  contig. Default 500 bp
-c (coverage rate)          - (int)     Rate factor for the coverage sampling in the
Read count distribution. Default 100. The smaller it is, the less bins are sampled
-k (mode smoother window) - (int)       Number of points over which the ploidy
estimation is smoothed. The mode over k numbers of windows is used to average the
values of the bin. Default=49
-s (significant min)      - (double) Threshold to consider a cluster peak in the read
count to be significant. Default 0.1
-b (fitter bin factor)       - (double)  Affects the number of bins used to FIT the
read count distribution. Default 2.5; Recommended between min=2.0 and max=4.0
-v (allele frequencies)    - (Pathway to .vcf file)  Pathway to the file containing
the variant calling. Must be a .vcf file
-d (coverage data to use)    - (double) Fraction of coverage data that is used in the
read count distribution to infer the different ploidies and their ratio. Values
between 0 and 1. Default 0.97
++++++++++++++++++++++++++++++++++++++++++++++++++++++++
```

Downloading PEDCA https://github.com/AbeelLab/Pedca

REQUIRED PARAMETERS:

The first three arguments are required for PEDCA to function:

```
-p <project name> -i < input sam/bam File> -o <output Folder>
```

PEDCA creates a folder named by the concatenation of the project name and the size of the window length at the output pathway indicated by the user. The output has the following structure:

./<OutputFolderPath>
- ./BaseCall
  - . BaseCallHistogramCluster_1.jpg
  - . BaseCallHistogramCluster_2.jpg
  - . Matrix1stCluster.vcf
  - . Matrix2ndCluster.vcf
- ./<Project Name<*wl*>>
  - *./Ploidy_Estimation_Charts*
  - *.PEDCA*<Project Name<*wl*>>*PloidyEstimation.txt*
  - *.PEDCA*<Project Name<*wl*>>*PloidyEstimation_2nd_Round_.txt*
  - *. readsDistribution.jpg*
  - *.readsDistributionFittedFINALRESULT.jpg*

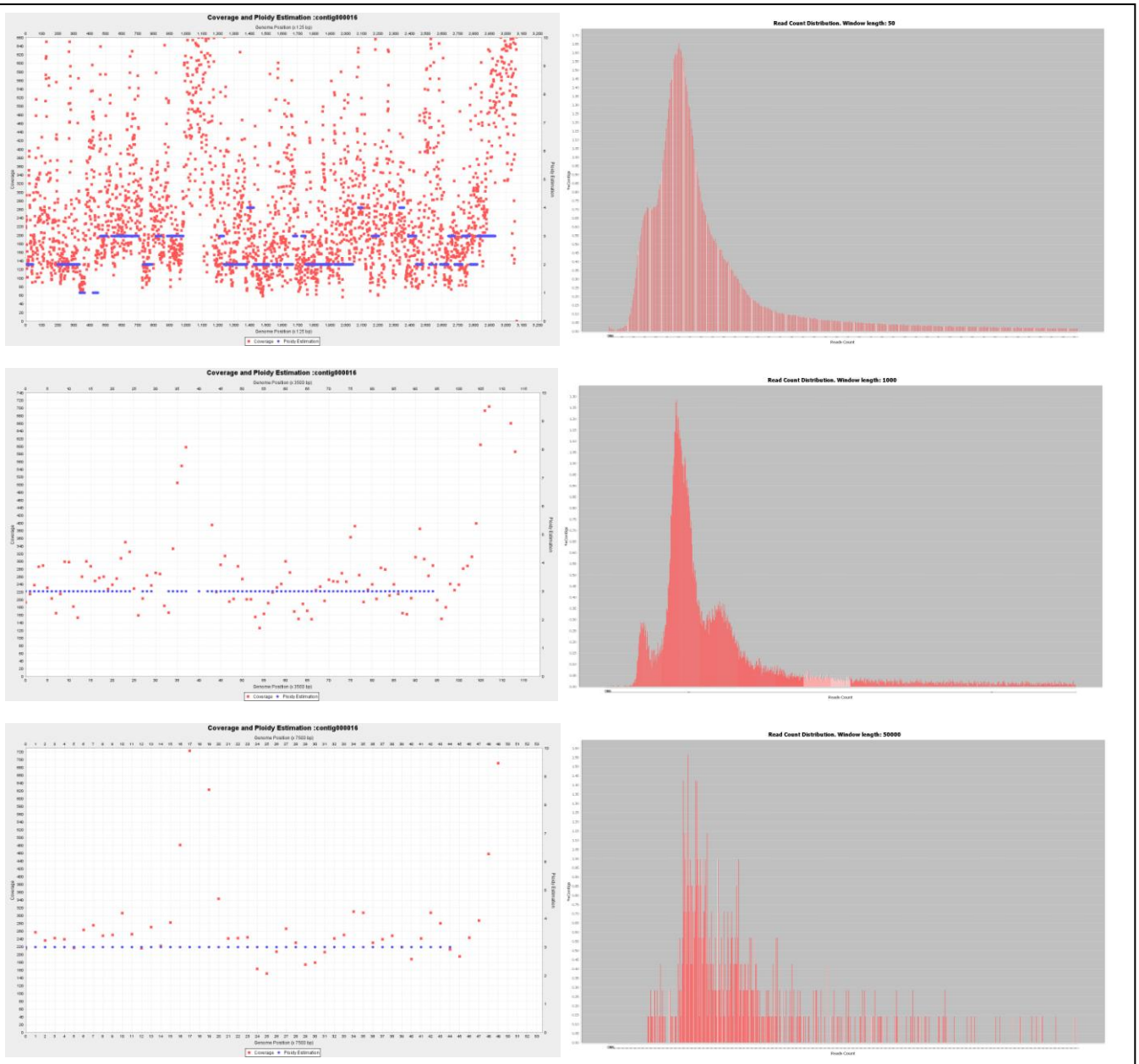
OPTIONAL PARAMETERS:

```
-w <window length>
```

Is worth noticing that the window length (*wl*), despite being the main parameter in PEDCA, is not a mandatory field. If no other preference is indicated, PEDCA runs with the default value of *wl*=500 bp. Even if the parameter is not required, the advantage of using PEDCA is to have a customizable window length so it is strongly recommended to use it with different values and compare the results.

A short *wl* provides more coverage data points to estimate the ploidy, you might want to shorten the *wl* if your ploidy estimation plot is too discontinuous or if it doesn't have much coverage information to support a reliable estimation **Tutorial Figure 1**. On the other hand, you might want a larger *wl* if your coverage/estimation plot looks overcrowded with coverage data with too much variation, which leads to a fragmented discontinuous copy number estimation **Tutorial Figure 1**. The minimum size of the window is 16 bp.

The *wl* also affects the sampling in the read count distribution, if it is too big, it will lead to a irregular sampling with unrecognizable clusters and false cluster ratios **Tutorial Figure 1**. If it is too small the read count distribution will have its clusters merged together with long and thick tails that might hide undetected peaks **Tutorial Figure 1**.



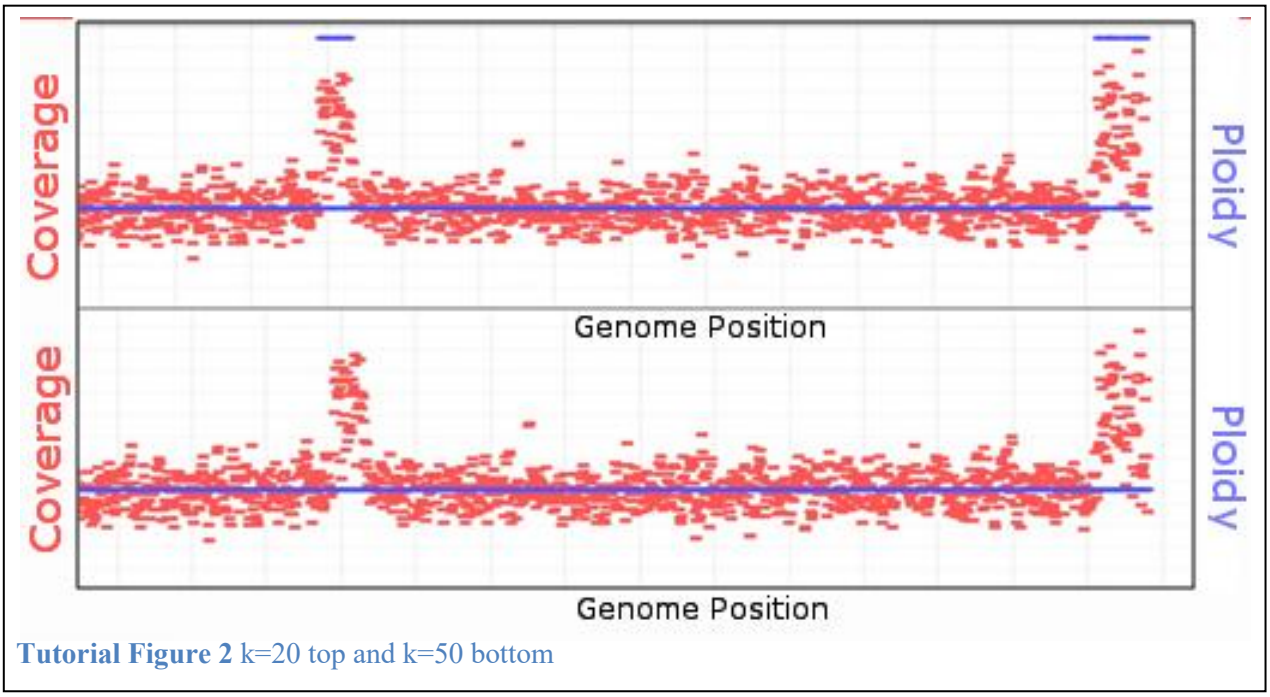
Tutorial Figure 1 Too short wl (50 bp; top figures), too long (30 Kbp bottom figure) and within optimal range (3 Kbp center figure)

`-v <Pathway to .vcf file>`

If this option is selected and a .vcf file submitted, a folder named BaseCall is also created at the same address., containing the allele frequencies plots and a matrix with the positions and frequencies o each base (order A,C,G,T).

`-k <mode smoother window>`

The coverage data has a certain degree of variation that we don't want to see reflected in the copy number estimation. In order to avoid undesired jumps in the ploidy plot, the points are averaged by the mode value over a bin of length *k*. If *k* is too small it might lead to fragmented ploidy estimation in regions with noisy coverage (**Tutorial Figure 2 top**). The continuity is smoothed with the default *k* value of 50 bp (**Tutorial Figure 2 bottom**). The correct length of *k* depends on the required precision, and can be parameterized. If *k* is too big, it might lead to the non detection of regions with different ploidies (i.e. large structural variations found in hybrid genomes)



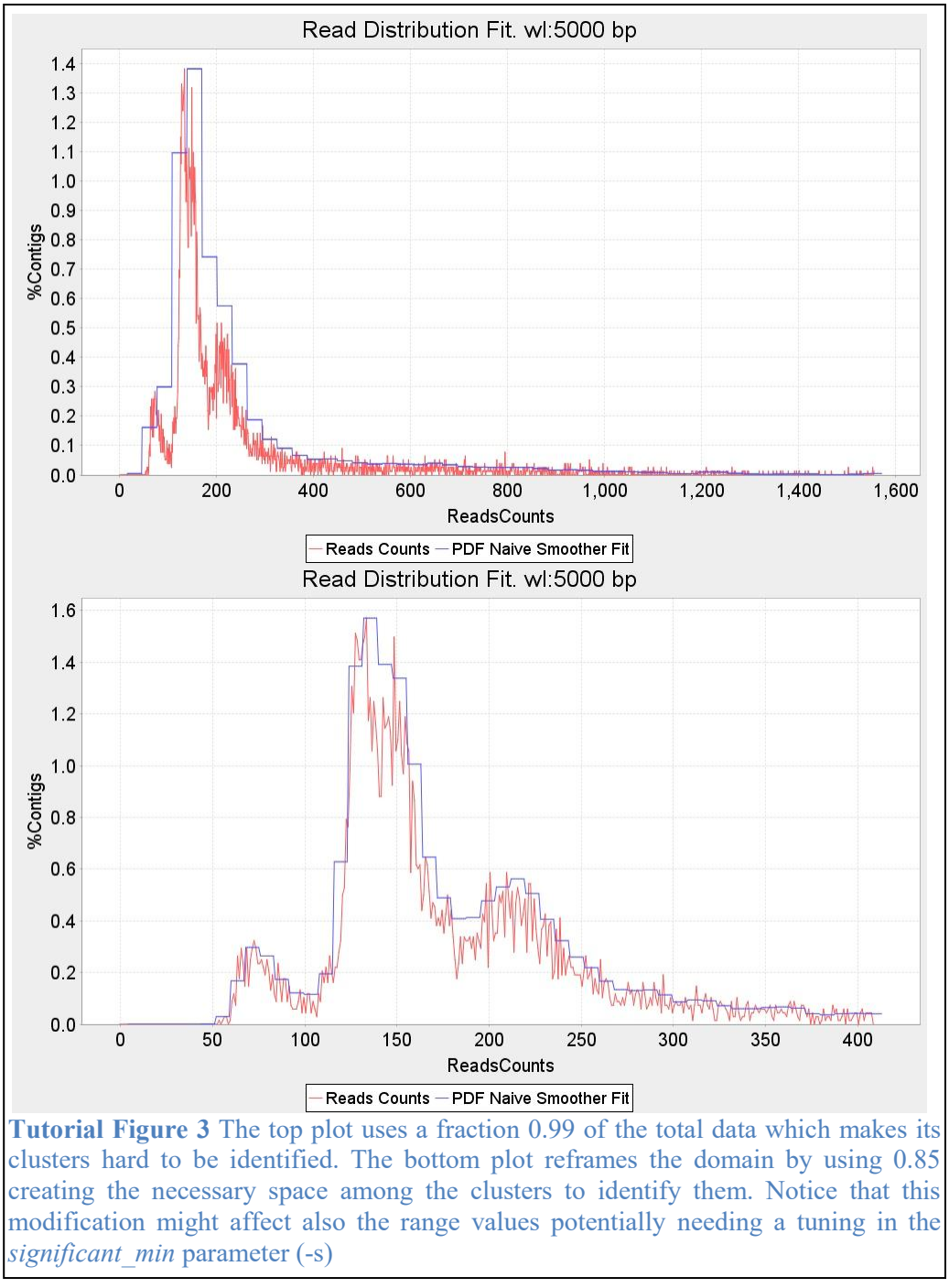
Tutorial Figure 2 k=20 top and k=50 bottom

`-m <multiple window lengths run>`

This parameter enables multirun mode. Instead of running PEDCA with one single *wl* value, it automatically runs it five times with the preset values {500, 750, 1000, 2000, 3000} and output the respective results to the output folder. These values work well for contigs larger than 500 bp and up to 2.000 Kbp.

```
-d <coverage portion of data to use (lower values)>
```

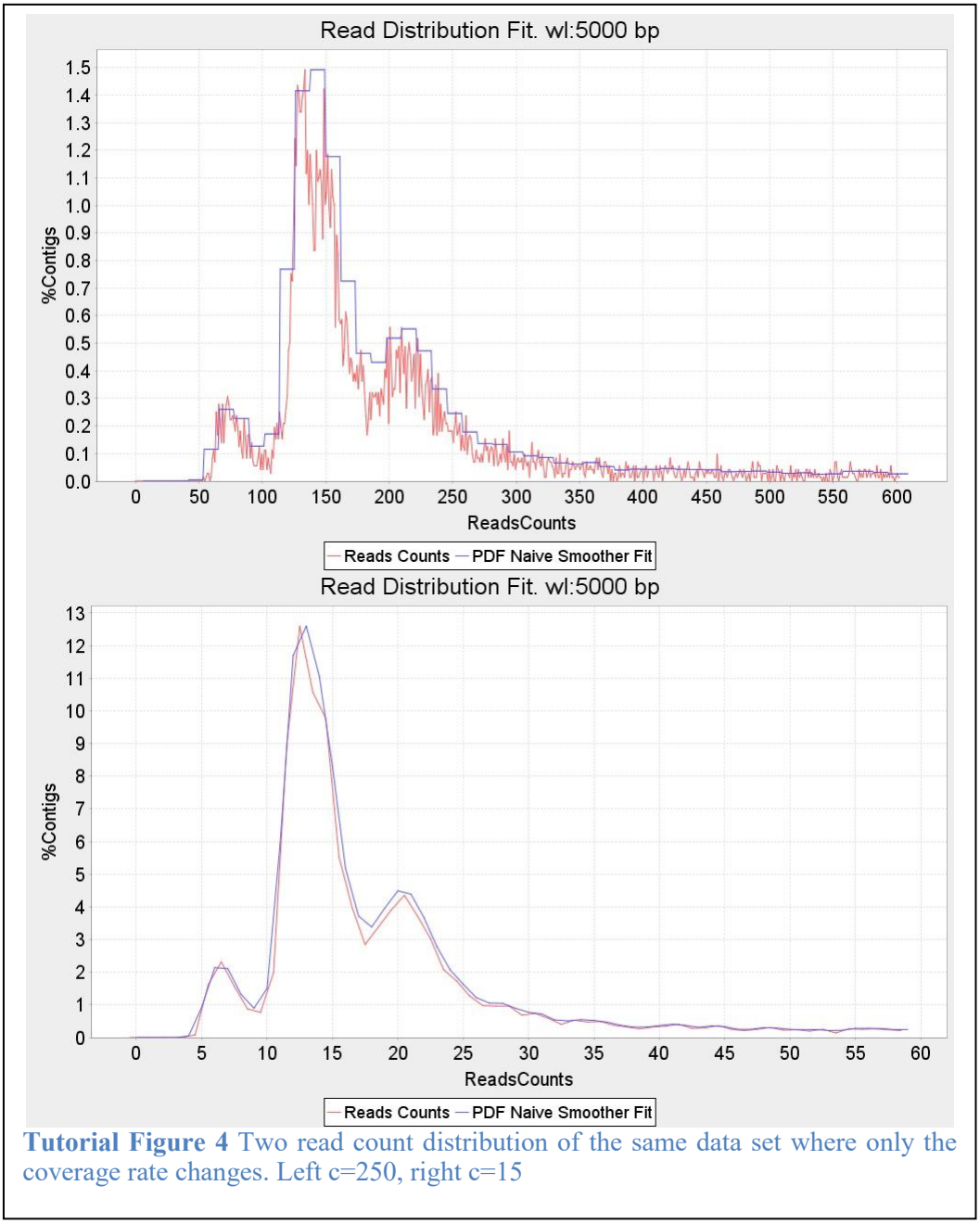
In order for PEDCA to correctly fit and identify the read count clusters it is important to select the correct domain of the plot. By default PEDCA keeps 0.97 of the data and rejects the top values, which doesn't affect the fitting of the read count PDF. Nevertheless, some highly noisy data sets might have a long tails on their last cluster with no significant ploidy information. This long tail might compress the significant clusters to a very small region of the domain, making it difficult to differentiate them and find their correct ratio. In those cases it can be very helpful to lower the fraction of data to use. Values can range from 0 to 1 **Tutorial Figure 3**



Tutorial Figure 3 The top plot uses a fraction 0.99 of the total data which makes its clusters hard to be identified. The bottom plot reframes the domain by using 0.85 creating the necessary space among the clusters to identify them. Notice that this modification might affect also the range values potentially needing a tuning in the *significant_min* parameter (-s)

`-c <coverage rate>`

The coverage rate is the definition with which the read count distribution is drawn. It affects the number of bins in the plot. The default value is 100. In some cases, when the plot is too irregular and sawed, it can be convenient to reduce this rate to have a fit that doesn't identify false peaks **Tutorial Figure 4** . If the sampling rate is too low, the bins might merge clusters, so it's recommended to use this option with caution.



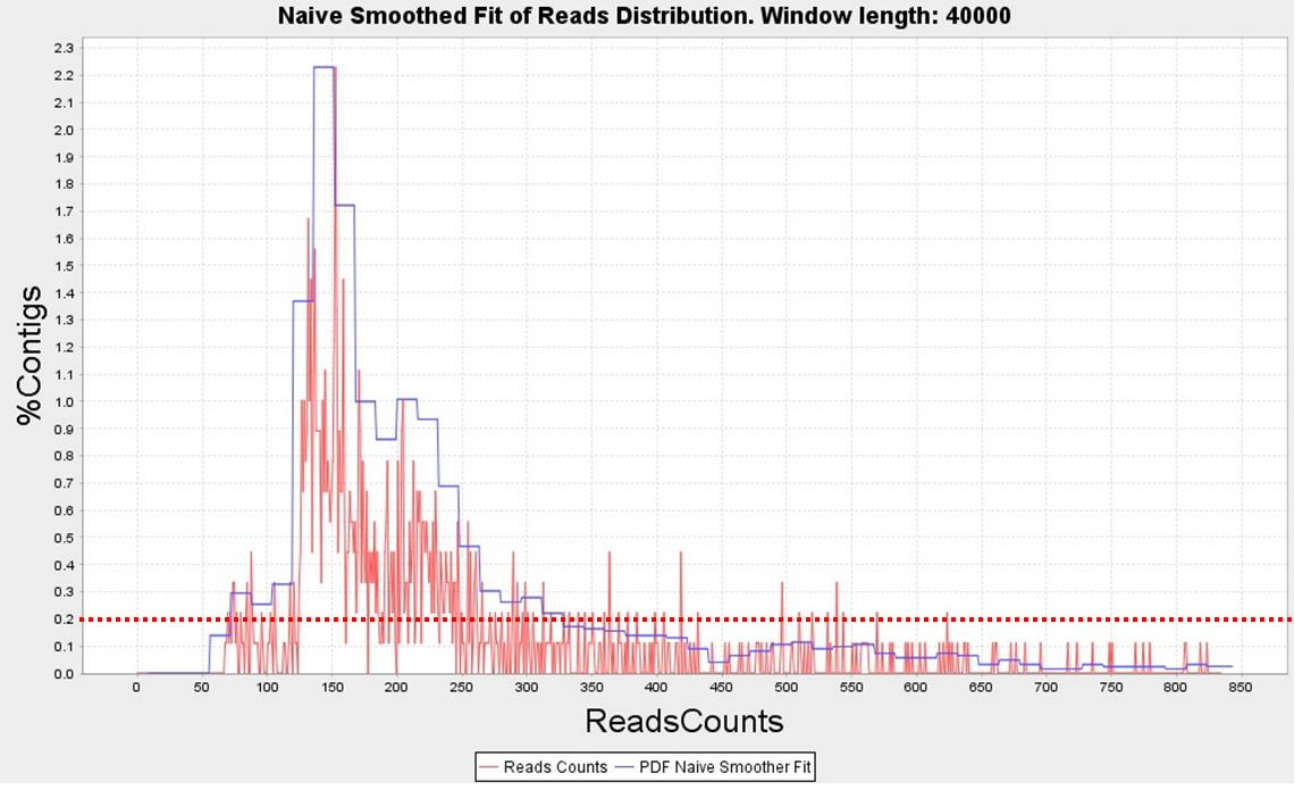
Tutorial Figure 4 Two read count distribution of the same data set where only the coverage rate changes. Left c=250, right c=15

```
-s <significant min>
```

When the read count is fitted, only peaks detected above a certain threshold are taken into account, otherwise insignificant oscillation peaks could be considered as clusters. The default values is preset to s= 0.1 % of the normalized number of windows for a given read count. For some genomes this value might be too big and real clusters could be missed with a potential misinterpretation of the correct cluster ratio. In the other hand, if the distribution has a long tail with isolated values that are not considered clusters, it is important to raise the threshold to ignore false peaks in that region that would also jeopardize finding the appropriate cluster ratio.

In the example in **Tutorial Figure 5** many micro peaks are detected in the long tail of the distribution. With a 0.2 significant minimum all peaks below the red line are discarded. If instead the default value was used, an error message would be displayed because the peaks' ratio would not make sense:

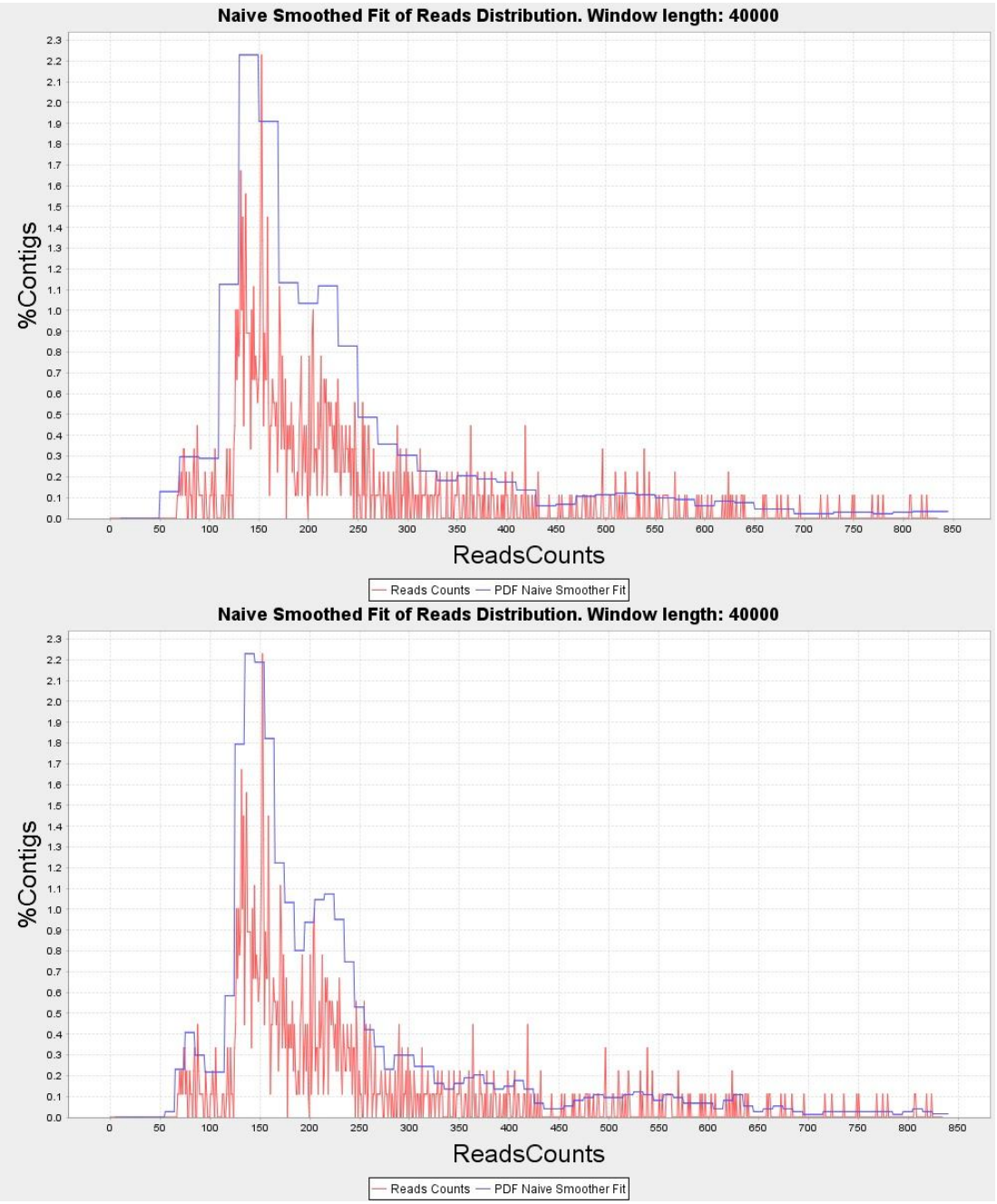
++++++++++ bestScore.candidateUnit: No CN mixture was able to satisfy the constraints. Result == null



Tutorial Figure 5 The red dots line indicates the significant min value (here 0.2%) below which peaks will not be taken into consideration

`-b <number of fit bins>`

With the default value (2.5) PEDCA fits the read count distribution with 25 bins, that is 2.5 x the maximum number of ploidies that PEDCA can detect. That number is adapted to detect a few clusters that are not very spread over the x axis of the read count distribution. If the clusters are far away from each other a higher number might better fit the distribution. It is recommended to remain between the values min=2.0 and max=4.0



Tutorial Figure 6 Two read count distribution fits where only the fit bin factor rate changes. Top b=2.0, bottom b=4.0