**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 requires as an input an alignment file in .bam or .sam format of a library or set of libraries aligned to a reference file, containing all 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*>

**Downloading PEDCA**

<https://github.com/AbeelLab/Ploest>

**Using PEDCA**

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 (absolute if necessary) to the input file containing the alignment file. Must be a .bam or .sam file

-o (output Folder) - (String) Pathway (absolute if necessary) to the auto-generated output folder that will contain the results('./' if the folder points to the current directory)

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 (absolute if necessary) to the file containing the variant calling. Must be a .vcf file

+++++++++++++++++++++++++++++++++++++++++++++++++++++++

**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*

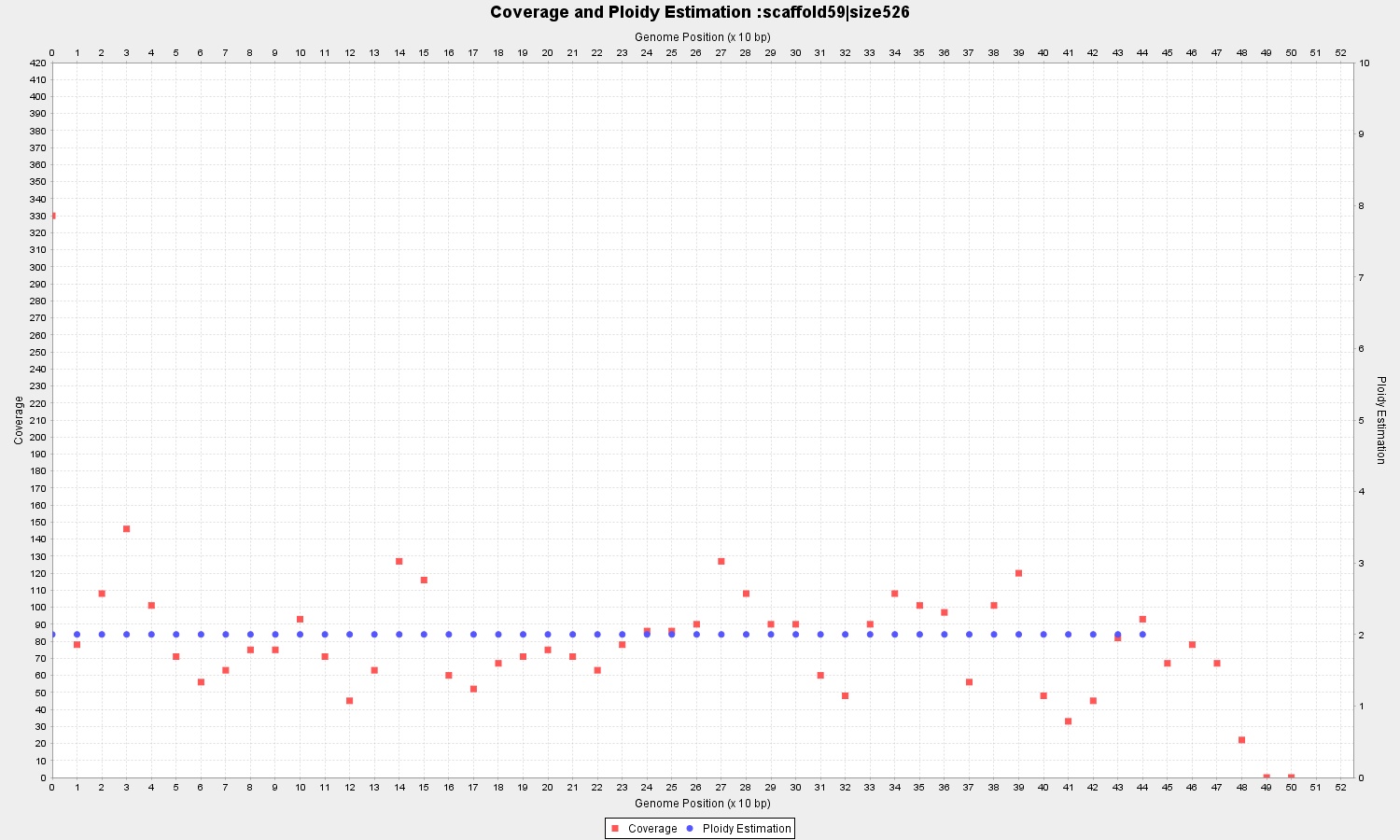
**OPTIONAL PARAMETERS:**

-v <Pathway to .vcf file>

If the 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)

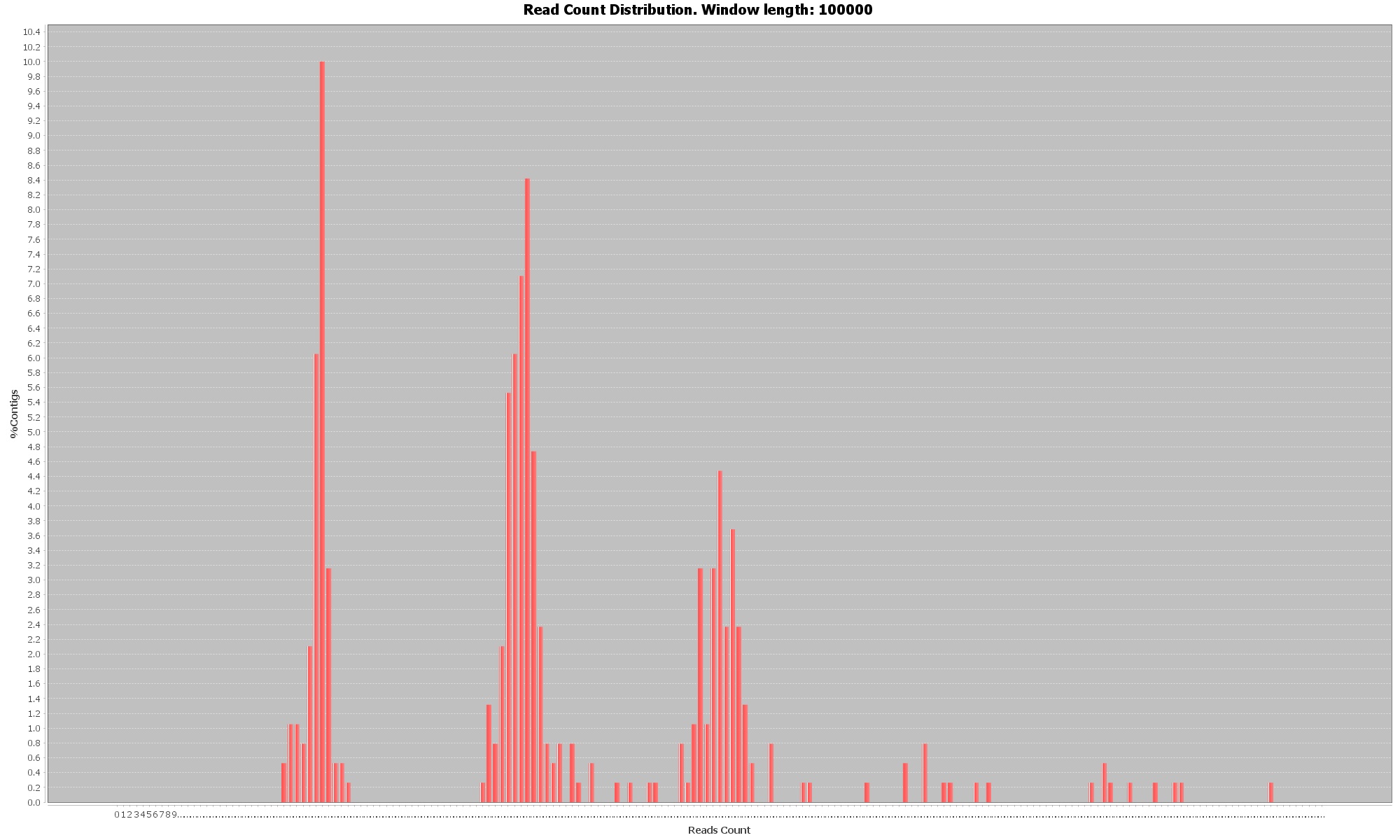
-w <window length>

Notice that the window length (*wl*), despite being the main parameter 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 advised to use it and compare among different results.



Tutorial Figure

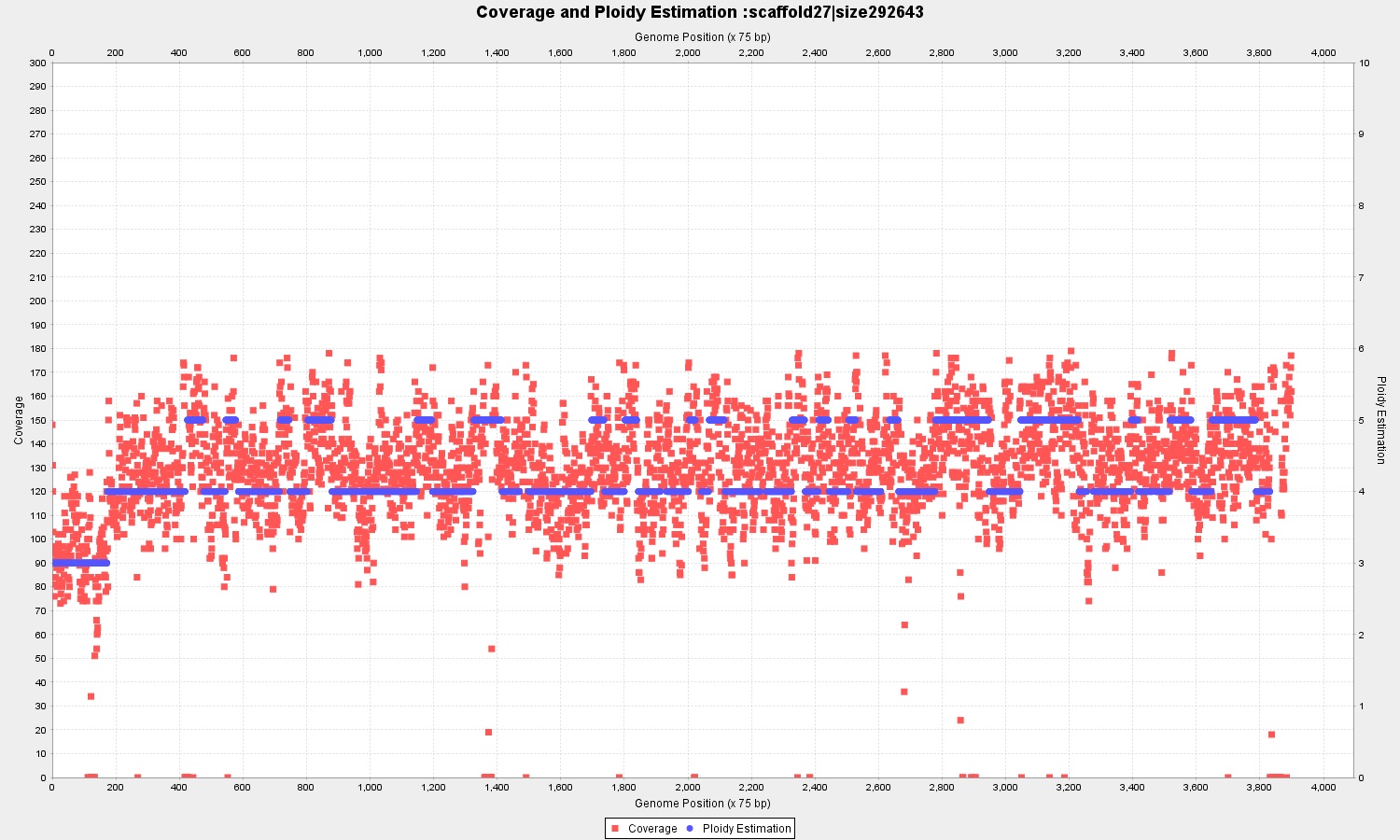
A short *wl* provides more coverage data points to estimate the ploidy, you might want to increase the *wl* if your ploidy estimation plot is too discontinuous or if it doesn’t have many coverage information to support the estimation. The minimum size of the window is 16 bp so with very short contigs (<500 bp) this is inevitable, which is one of the limits of the current implementation .



Tutorial Figure

The *wl* also affects the sampling in the read count distribution, if it is too big, some contigs might have irregular sampling which would lead to an erroneous fit, with unrecognizable clusters and false cluster ratios. You might want to shorten the bin if your read count distribution looks like .

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 . Same thing if the read count distribution has clusters that merge together and have too long and thick tails that might hide undetected peaks .



Tutorial Figure

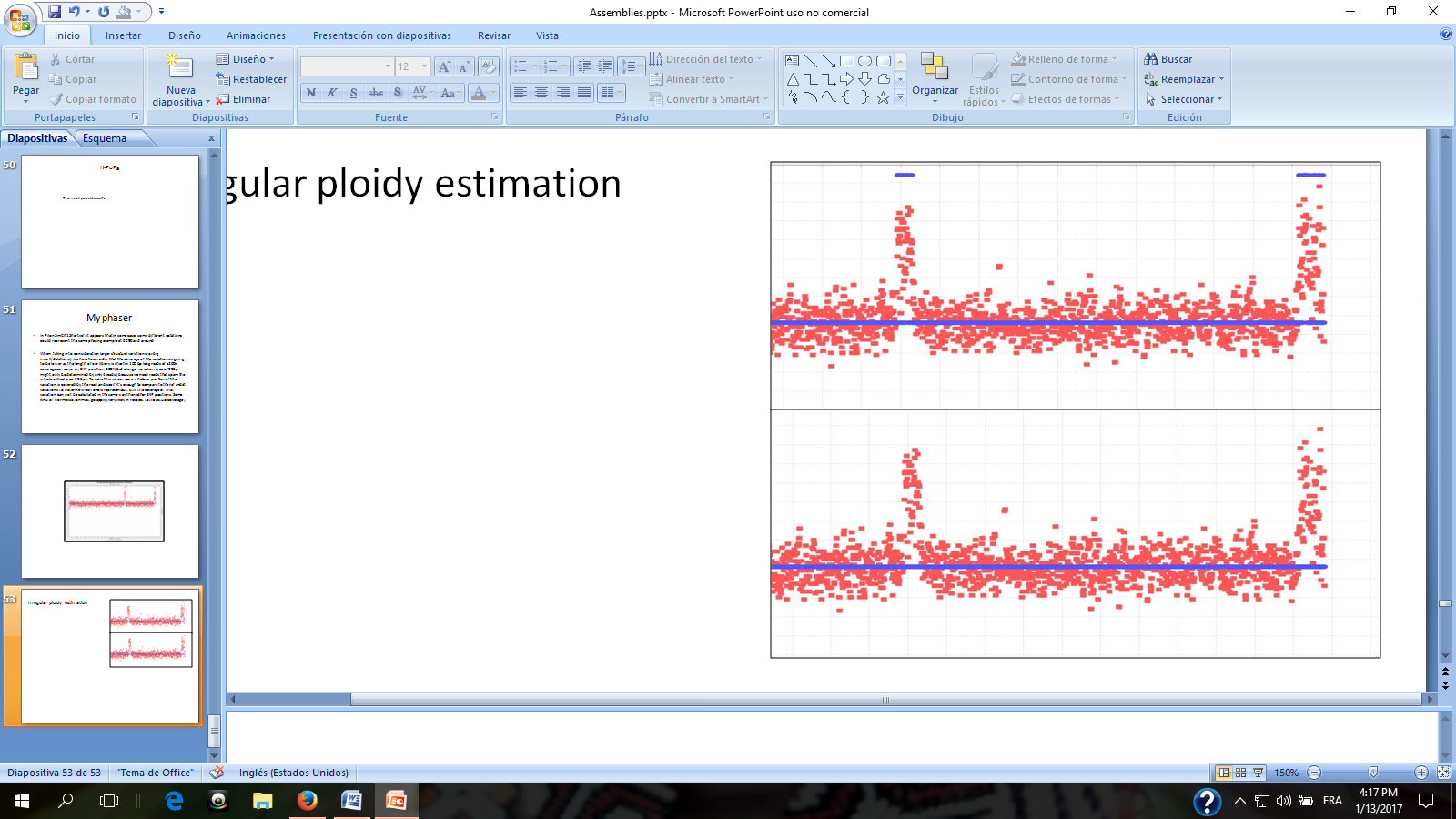


Tutorial Figure

-m

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.

-k <mode smoother window>

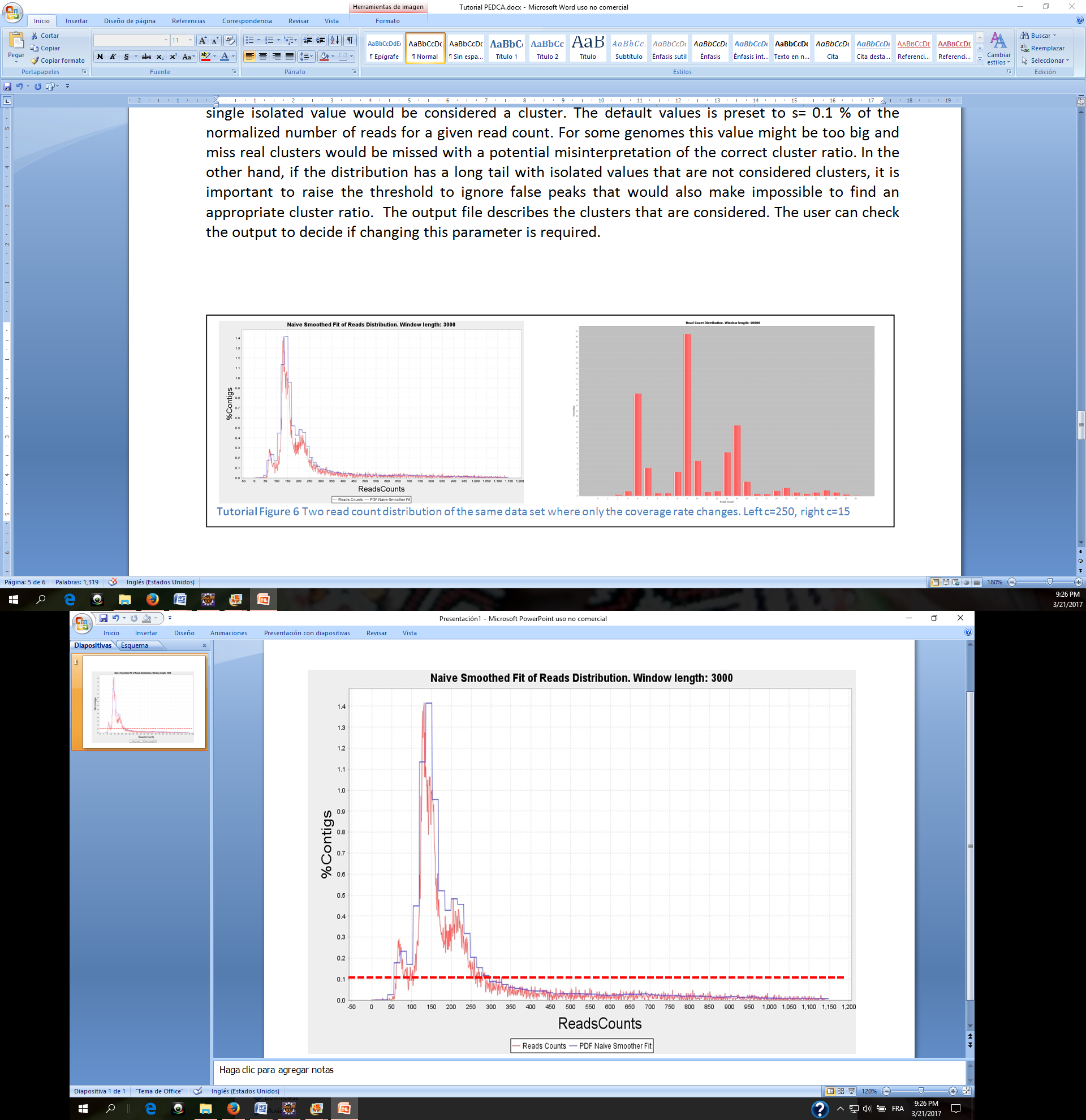


Tutorial Figure

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 their values 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 (top ) The continuity is smoothed with the default *k* value of 50 bp (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)

-s <significant min>

When the read count is fitted only peaks detected above a certain threshold are considered, otherwise a single isolated value would be considered a cluster. The default values is preset to s= 0.1 % of the normalized number of reads for a given read count. For some genomes this value might be too big and miss real clusters would 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 that would also make impossible to find an appropriate cluster ratio. In the example in many micro peaks are detected in the long tail of the distribution. With the default significant minimum all peaks below the red line are discarded. If instead the –s value was lowered to 0.01 an error message would be displayed:

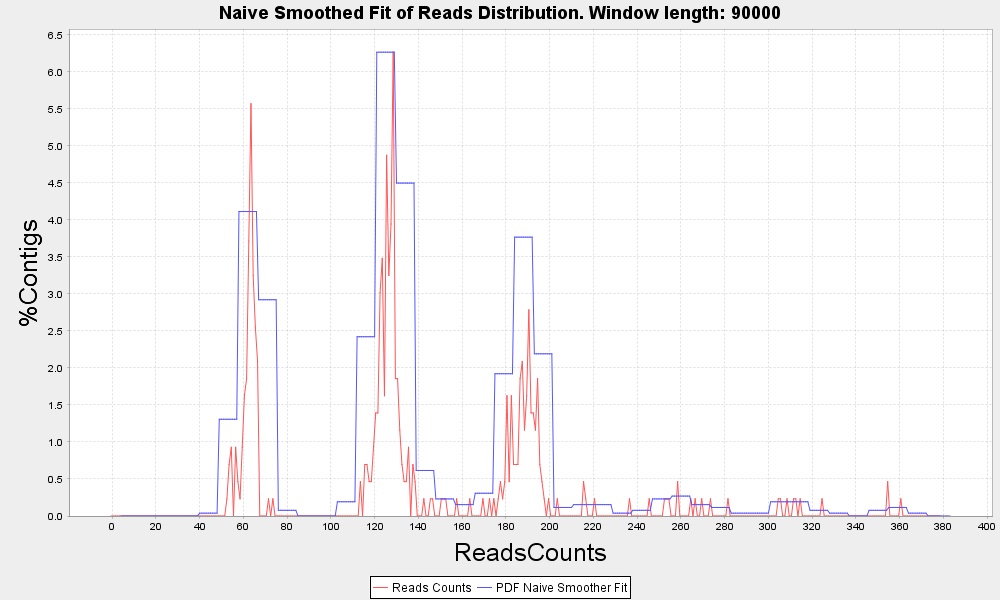
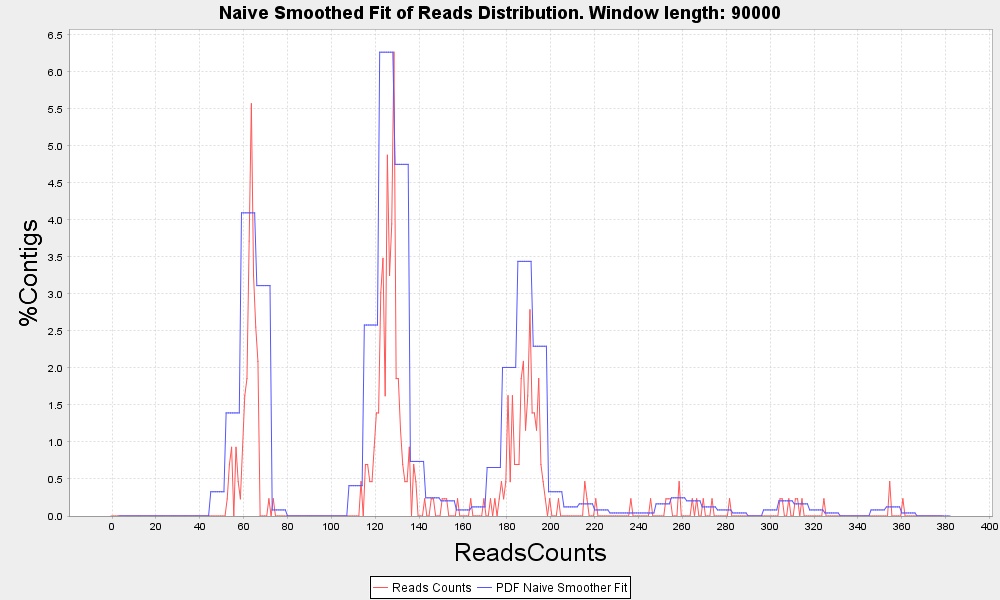


Tutorial Figure

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

-b <fitter bin factor>

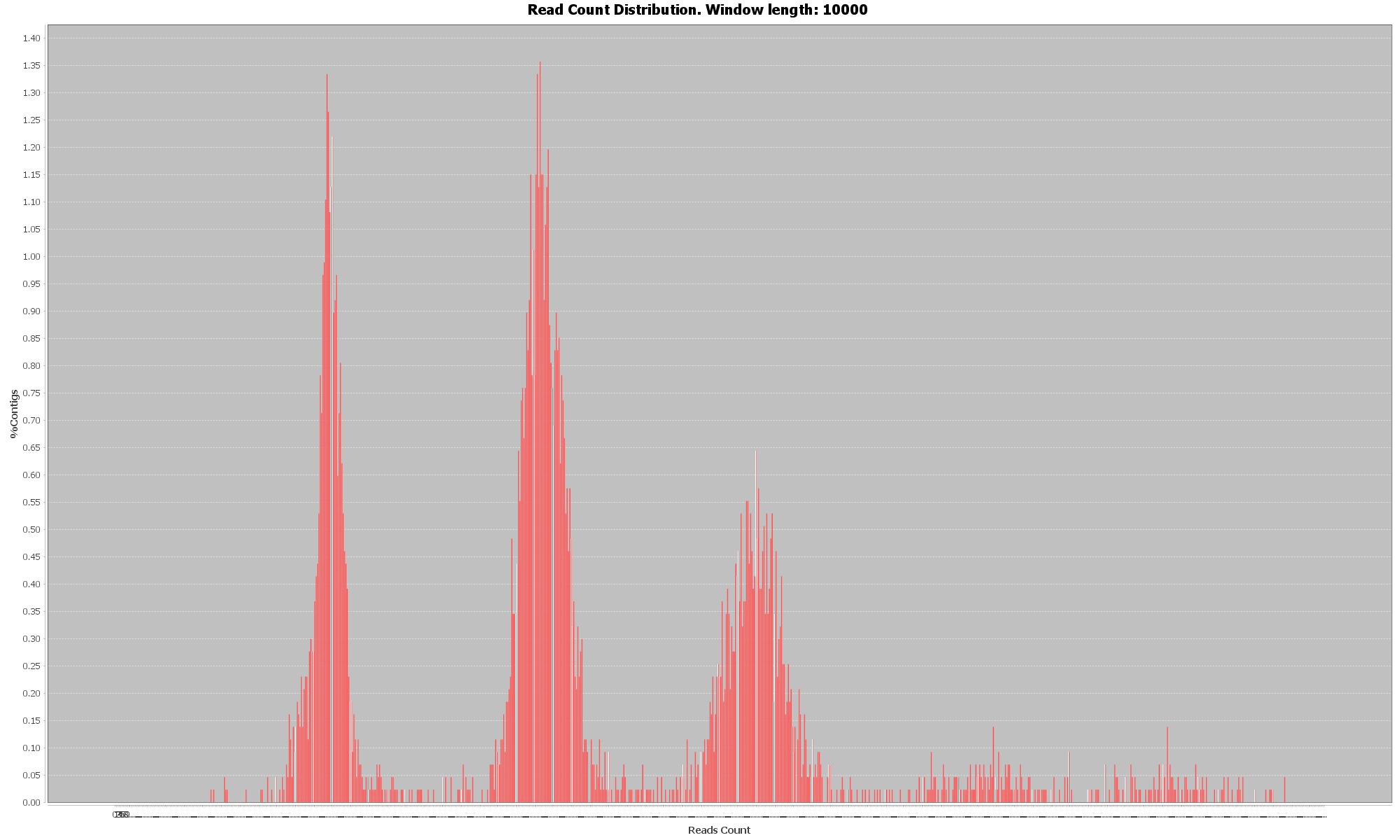
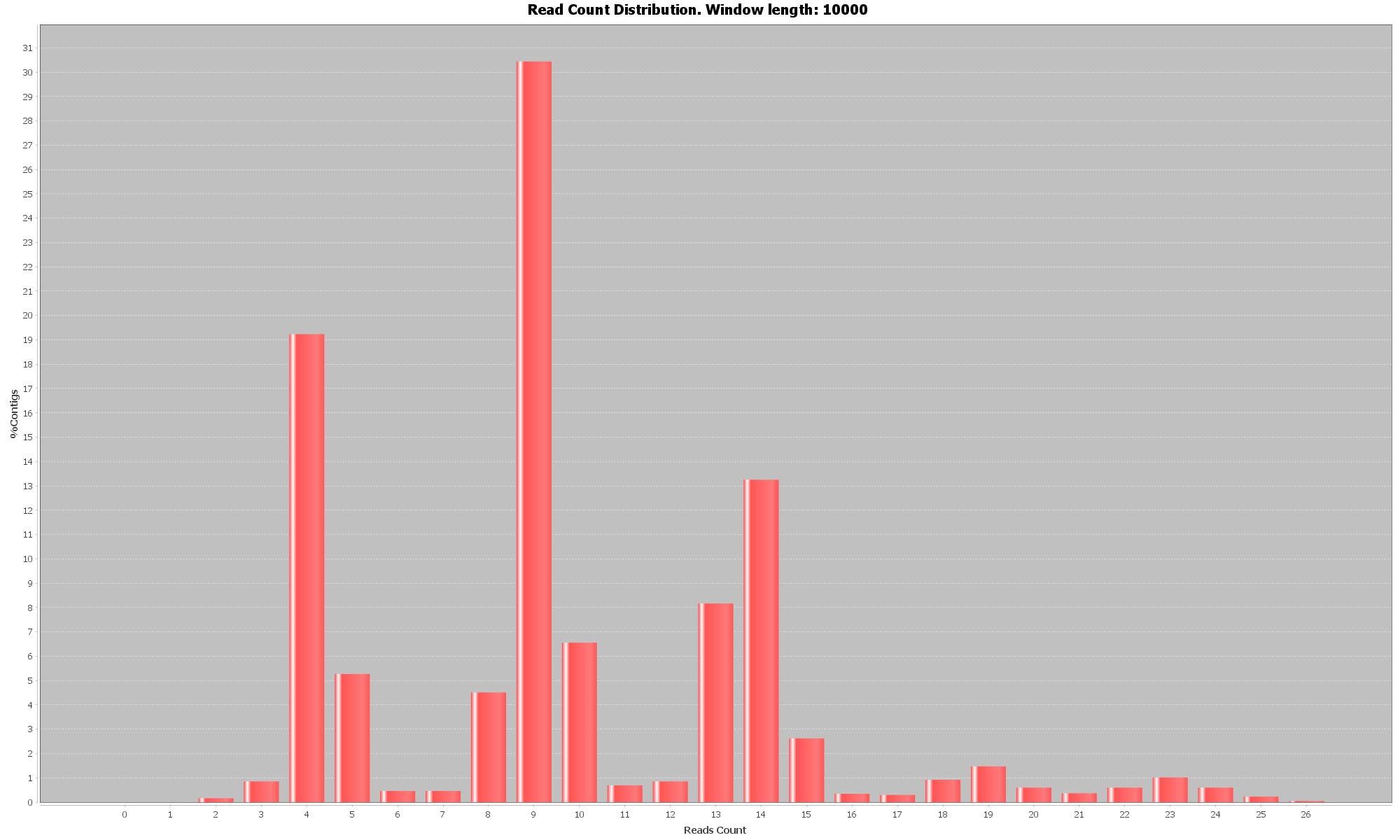
Default 2.5; We fit the read count distribution with 25 bins, that is 2.5 x the maximum number of ploidies that Pedca can detect. That number is adapted to detect a few clusters that are not very spread over the x axis of the read count distribution. If the clusters are far away from each other a higher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

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

-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 is convenient to reduce this rate to have a fit that doesn’t identify false peaks. If the sampling rate is too low, the bins might merge clusters, so it’s not recommended to go below a value of 50.

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