April, 2014

1. Introduction

Comparative analyses between Next generation sequencing (NGS) generated profiles, such as ChIP-seq, RNA-seq, Gro-seq, or MeDIP-seq require prior characterization of the degree of technical similarity of the various data sets, as individual profiles can vary significantly even between biological replicates, the use of different antibodies and batch-to-batch variations of the same antibody, sequencing depth and immunoprecipitation (IP) quality are only a few of the parameters that impact on the quality of a ChIP-seq profile. The present NGS-QC Generator infers global and local quality indicators based on a stand-alone approach, as it does not require additional wet-lab efforts. This computational approach generates read count intensity profiles from randomly selected subsets of the total originally mapped reads (TMRs) associated to the NGS-profile under study and defines the divergence from the theoretically expected read count intensities (RCIs) recovery after sampling relative to the original profile. For this, TMRs are first randomly sampled at three different densities (90%, 70% and 50%; referred to hereafter as s90, s70 and s50 subsets, respectively); then the genomic RCI profile is recorded for successive 500bp bins and compared to that of the original profile. This comparison is performed to evaluate the divergence from the ideal condition in which the RCI/bin for a s50 subset correspond to 50% of the original RCI/bin value. Importantly, NGS-sampled generated profiles diverge always to different degrees from the hypothesized "ideal behaviour", thereby generating a quantifiable denominator (referred to as profile "robustness"), which is linked to the quality of any NGS-generate profile (Mendoza-Parra et al.; manuscript in preparation).

Below we describe the different steps involved in the NGS-QC Generator's accessibility through the web-based platform GALAXY, the required input parameters and the information provided in the NGS-QC Generator Report. In addition, we provide an interpretation of the different quality control indicators, give examples and discuss additional applications of this methodology.

2. Running NGS-QC Generator

2.1 NGS-QC parameters: The NGS-QC Generator requires as input a single file containing the genome positions of the uniquely aligned reads (BAM or BED format). Depending on the user-defined analysis, the following additional parameters may be required:

- **Genome**: An important number of model organism genomes are supported, among them Homo sapiens (hg19, hg18); Mus musculus (mm9, mm8); Ratus norvegicus (rn4, rn3); Drosophila melanogaster (dm3, dm2); Caenorhabditis elegans (ce6, ce4); Dario rerio (dr6, dr4). If required we may add other reference genomes under request.
- **Windows size ('bin') for read counts enrichment assessment**: Currently the "default" parameter is set to 500 bp.

! *NOTE* While this parameter was available in previous versions of the NGS-QC Generator tool, we have decided to inactivate this option. In fact, for comparative analyses of several profiles, the QC indicators should be calculated using identical bin sizes, thus 500 bp windows should be used to compare QC indicators of a user profile with those displayed in the NGS-QC database available in our website: www.ngs-qc.org

We have studied the effects of bin size variation on the NGS-QC generator-calculated indicators and found the highest sensitivity (i.e., highest difference for the QC indicators assessed at different dispersion intervals) at a bin size of 500 bp.

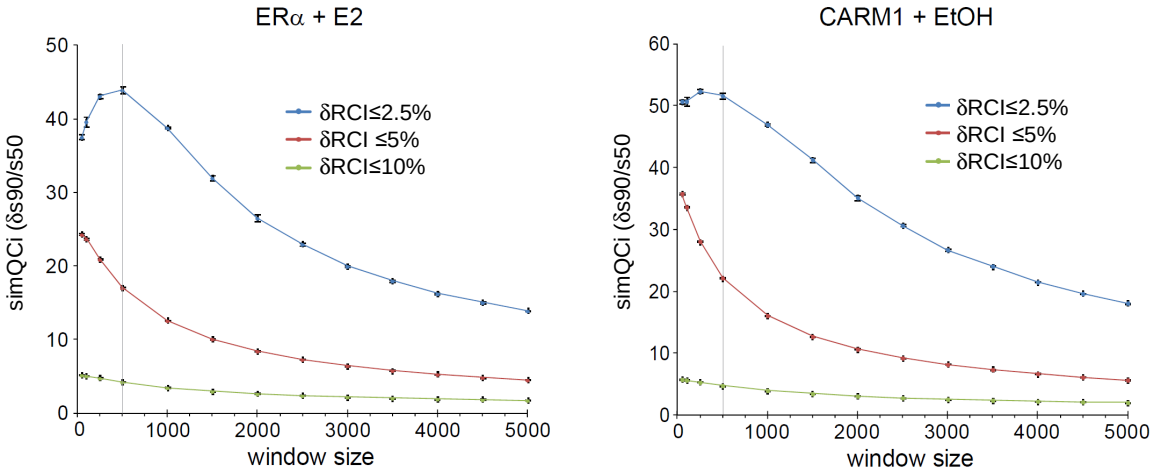


Figure I. Influence of the window size on the assessment of QC indicators for two different ChIP-seq profiles. Similarity QC indicator at different window sizes have been computed for Era and CARM1 ChIP-seq profiles. As highlighted by the vertical gray line, the highest difference for the simQC indicators assessed at three different dispersion intervals (2.5, 5, 10%) is retrieved for bins presenting a windows size between 250 and 500bp. Note that this value is in concordance with the expected chromatin fragmentation size.

- **Number of random sampling replicates**: We have obtained highly reproducible Global QC indicators (less than 2% coefficient of variation among five sampling replicates) when using several sampling replicates; however, users can choose the number of replicate samplings.

! ***NOTE*** This option is supported up to 3 replicates.

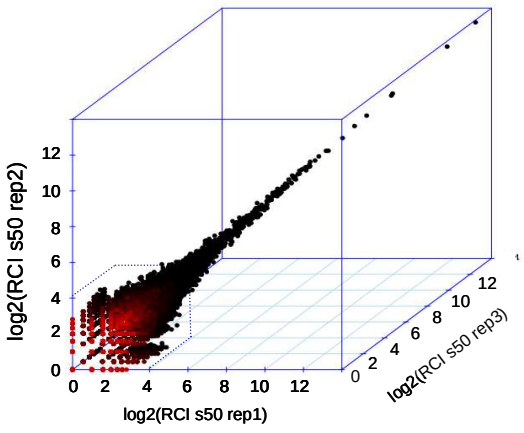
- **Bias corrections:** Two other optional parameters are available as part of the last version of the NGS-QC Generator:
  - ***Background noise bias subtraction:*** Background noise is unavoidable in ChIP-seq profiling but the ratio/intensity can vary for different datasets due to differences in antibody specificity, depth of sequencing and occasionally due to partial misalignment. Reads considered as background could bias the QC indicators calculation and QC-STAMP prediction. In order to exclude background noise while calculating the RCI/bin, we have incorporated a Poisson distribution-based model to estimate the background level threshold. From the Poisson distribution, a λ value (intensity cutoff) is estimated and genomic regions presenting RCIs under this threshold are removed from the analysis. By default, the option to remove background noise bias is activated; nevertheless users can uncheck this option if one wishes to experiment or for comparative purposes.
  - ***Clonal reads bias substraction:*** Clonal reads (PCR duplicates) are unavoidable in ChIP-Seq. In most cases, clonal reads are produced by PCR amplification during sequencing library preparation. Nevertheless, in

certain cases, clonal reads may originate from the procedure used to fragment the chromatin prior sequencing and cannot be addressed as PCR duplicates per se. Currently, the NGS-QC generator does not remove them systematically (whereas datasets with more than 50% of clonal reads out of total mapped reads are automatically subjected to a warning notice through the display of this information in red) , but users can have control over selecting the corresponding option for performing clonal read removal prior QC indicators assessment.

2.2 NGS-QC quality report: The NGS-QC Generator produces a certain number of output files which are summarized in a report available in PDF format. This report is subdivided in 3 sections as described below.

- ***Dataset informations:*** Information associated with the processed dataset. The File name should contain, in an ideal case, information specifying target and assay type (targeted factor, epitope, and antibody source and batch specification for ChIP-seq, origin and type of RNA in RNA-seq, GRO-seq, etc), treatment (if any) before ChIP, the model system used and any other information that is considered useful for future *meta* analyses. Furthermore, dataset informations like the Total mapped reads, the fraction of unique reads (excluding clonal events), the Genome assembly and when available the target molecule is displayed in this section.

a



b

simQCi (δs90/s50)

dataset 1	rep1	rep2	rep3	rep4	rep5	average	CV%
dRCI≤2.5	8.36	8.33	8.23	8.36	8.24	8.30	0.78
dRCI≤5	6.08	6.08	6.04	6.04	6.04	6.06	0.38
dRCI≤10	3.22	3.23	3.23	3.24	3.22	3.23	0.23

dataset 2	rep1	rep2	rep3	rep4	rep5	average	CV%
dRCI≤2.5	45.75	44.28	45.26	45.50	44.61	45.08	1.37
dRCI≤5	14.38	14.35	14.44	14.32	14.42	14.38	0.33
dRCI≤10	3.76	3.77	3.77	3.75	3.76	3.76	0.14

dataset 3	rep1	rep2	rep3	rep4	rep5	average	CV%
dRCI≤2.5	43.83	44.36	43.51	43.86	44.66	44.04	1.04
dRCI≤5	16.58	16.43	16.44	16.35	16.55	16.47	0.56
dRCI≤10	4.05	4.07	4.05	4.06	4.07	4.06	0.23

dataset 4	rep1	rep2	rep3	rep4	rep5	average	CV%
dRCI≤2.5	50.05	50.43	50.09	51.54	49.20	50.26	1.69
dRCI≤5	20.97	21.22	21.18	21.12	21.08	21.11	0.47
dRCI≤10	4.67	4.68	4.71	4.67	4.66	4.68	0.41

Figure II. QC indicators reproducibility over TMRs random sampling replicates. a) TMRs associated to given ChIP-seq profile have been randomly sampled three times to a 50% density (s50). The read count intensity (RCI) recorded per 500bps bins in all three replicated are compared. Note that for RCI higher than 16 (4 in log2) the RCI correlation is quite high. **b)** In order to quantify replicates sampling robustness, four different ChIP-seq samples were sampled 5 times. After that, the similarity QC indicator (δs90/s50) was compiled for three dispersion intervals (2.5, 5.0 and 10.0%). Chart tables show the different values obtained for each replicate. Average and coefficient of variation (CV%) from all replicates were also computed. Note that in all the cases the CV% is less than 2%.

- ***QC parameters*:** Specification of the different parameters for data processing, including the percent of sampled reads, the bin window size and the number of replicate samplings. In addition, the detected number of chromosomes as well as the use or not of the background subtraction model and the clonal reads removal option is also indicated.

- ***Results*:** This section presents a compendium of the computed indicators based on the number of considered sequenced reads. In fact, based on the use of the clonal reads removal option, the number of "considered reads" may differ from the total mapped reads described on the Dataset informations item. As is explained below we distinguish as global QC indicators two parameters, the "density QC (denQC)" and "similarity QC (simQC)", and offer the possibility to associate a "local QC" to a given profile. As part of the Results item, the global QC indicators; density or denQC and similarity or simQC are displayed for all three read count intensity dispersion intervals (2,5%; 5% and 10%). A detailed explanation of these QC indicators is available bellow.

  In addition to the global QC indicators, in the right side of the report a scatterplot illustrates the effect of random sampling on the read counts intensity of the evaluated profile. This scatter plot illustrates the **o**riginal **R**ead **C**ount **I**ntensity per bin (*oRCI*) in the studied profile (x-axis) relative to the **rec**overed **R**ead **C**ount **I**ntensity (*recRCI*) after sampling (y-axis). This relationship is displayed as following:

$$recRCI = (\frac{samRCI}{oRCI}) * 100$$

  Where *samRCI* corresponds to the RCI/bin retrieved after random sampling. As illustrated in **Figure III**, the theoretically expected *recRCI* is directly proportional to the random sampling density, i.e. 90% for s90, 70% for s70 and 50% for s50 respectively.
  The described RCI scaterplot provides intuitive information about the quality of the generated profile[1]. In fact, profiles of good quality show high number of bins that display a proportional decrease of RCI/bin in the sampled subsets compared to the original dataset. Thus, the less dispersed the recRCI pattern is, the better is the quality of the associated profile. Note that towards low signal intensities ($<2^4$ read counts/bin) the sampling process inevitably results in increased dispersion.

[1] « Quality » is defined here as the degree of dispersion from the theoretically expected recRCI scatter after sampling, which corresponds to a proportional decrease of all RCI/bin values relative to the sampling. With this definition a maximum of the quality indicator is reached when the recRCI/bin values are equal to the *oRCI* multiplied by the sampling percentage (i.e. 50% for s50). Any deviation – for whatever reason - from the expected RCI/bin scatter provides a quantitative indicator of the quality of a given NGS-profile.

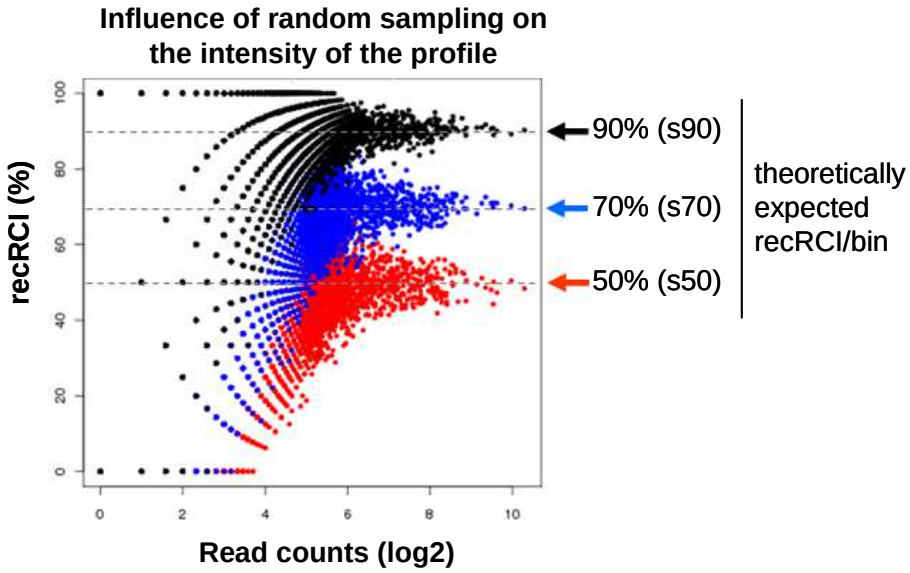


Figure III: Scatter plot illustrating the influence of random sampling on the read count intensity of a given profile. Each data point corresponds to the RCI within a 500 bp bin (x-axis) relative to the fraction of this intensity that is recovered after random sampling (y-axis). Note that for each of the three randomly sampled subsets, a different fraction of bins shows a proportional intensity recovery relative to the original read count measurements, and that with increasing number of reads the recRCI/bin increasingly approaches the theoretically expected value.

- ***Supplementary data:*** Further complementary information is also available as part of "supplementary data" that is generated together with the NGS-QC report (compressed zip file available through download). As part of such supplementary information, other plots describing the effect of random sampling on the read count intensities are also available; among them:

    - *Read count dispersion (dispersion_s50_s70_s90_replicate_1.png file).* To compare the dispersion effect relative to the expected proportional decrease in the RCI/bin values induced by random sampling, the scatter plot displayed in Figure 1 has been first centered by the following expression:

$$\partial RCI = samd - recRCI$$

      where "δRCI" corresponds to the ***read count intensity dispersion*** and "*samd*" to the random sampling density (i.e. 90%, 70% and 50% for s90, s70 and s50 respectively). This transformation facilitates to identify the subset of bins that display a δRCI *within in a given interval*, such as $\delta RCI \leq 5$ (illustrated in **Figure IV**). This information represents *per se* a quantifiable indicator for the quality of the studied profile. The current version of NGS-QCi Generator provides global quality indicators for dispersion intervals of 2.5%, 5% and 10%. In addition, the quality indicators for each 500bp bin are also generated in a wiggle format (described below as "local QC indicators").

★**Convention:** The measurement of the fraction of bins displaying a δRCI within in a given interval constitutes the global density QC indicator denQCi. The denQCi is described by the term "δs50/5" in which "s50" specifies the sampling in percentage and "5" the δRCI threshold.

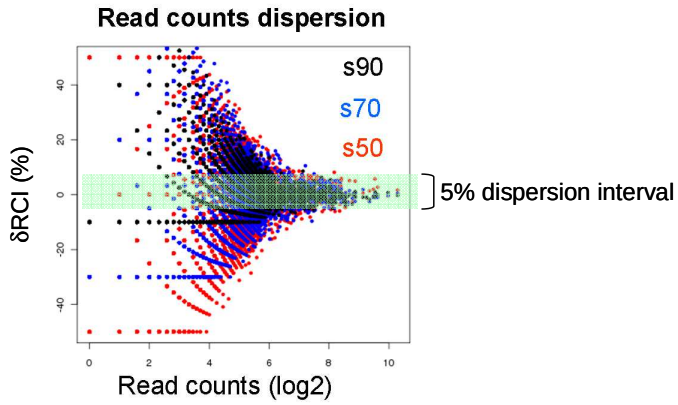


Figure IV: Scatter plot illustrating the recovered read count intensity dispersion of a given profile. This transformed scatter plot superimposes the three scatter plots obtained after sampling for the same original dataset. The scatter of bins after sampling at s50, s70 or s90 having RCI values that deviate ≤5% from the expected RCI/bin are highlighted.

- *δRCI at different Intensity thresholds (bins_dispersion s50 and s90.png files). A*s is apparent from **Figure IV** the dispersion of the read counts/bin after sampling and thus, the quality of the profile is inversely proportional to the RCI. In **Figure V** this dispersion (δRCI) is calculated for both the s90 and s50 randomly sampled and reconstructed profiles relative to the original s100 dataset. Note that in the illustrated example, bins with RCIs greater than 16 (4 in log2) present a median δRCI lower than 5% for both the re-sampled data sets. Importantly, for high quality profiles such a 5% threshold extends to lower RCI/bin values than for low quality profiles. Moreover, a similar dispersion pattern in s50 and s90 data sets is a sign for a high quality and "sampling robustness" of the evaluated profile; thus, the degree of similarity of the δRCIs of s90 and s50 data sets is a second quantifiable indicator that is evaluated by the NGS-QCi Generator.

★**Convention:** δRCI(s90/s50) constitutes the global similarity QC indicator simQCi. The simQCI is described as "δs90/s50/5", in which "δs90/s50" corresponds the ratio between the denQCi for s90 and the denQCi for s50 and "5" specifies the δRCI threshold.

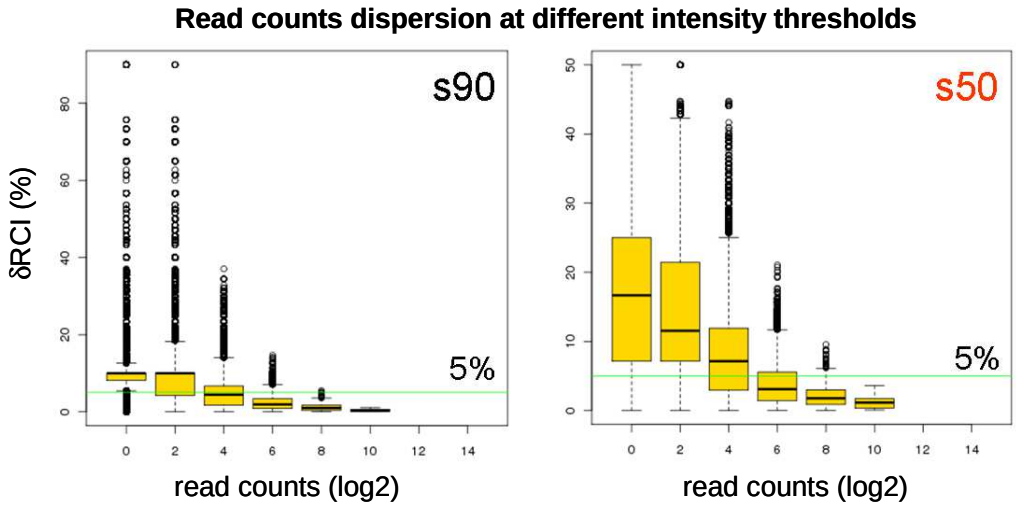


Figure V: δRCI evaluated at different intensity thresholds for both s90 and s50 random sampling subsets. The 5% δRCI threshold is indicated as a green line.

- *Number of bins at different δRCI intervals (bins_dispersion 2.5 5 and 10pc png files).* This analysis computes the fraction of bins in the sampled subset (i.e. s90 or s50) that exhibits a proportional decrease of their RCI for a given RCI threshold. A δRCI of 2.5% defines a very stringent condition, as nearly 50% of bins with RCI values above 256 (2^8) - corresponding to strong signals - are outside this interval (**Figure VI**, left panel). In contrast, more than 80% of such strong signals are within the interval defined by a δRCI threshold of 5% (middle panel); for more relaxed conditions, such as dRCI≤10%, all these signals in within the selected interval (right panel).

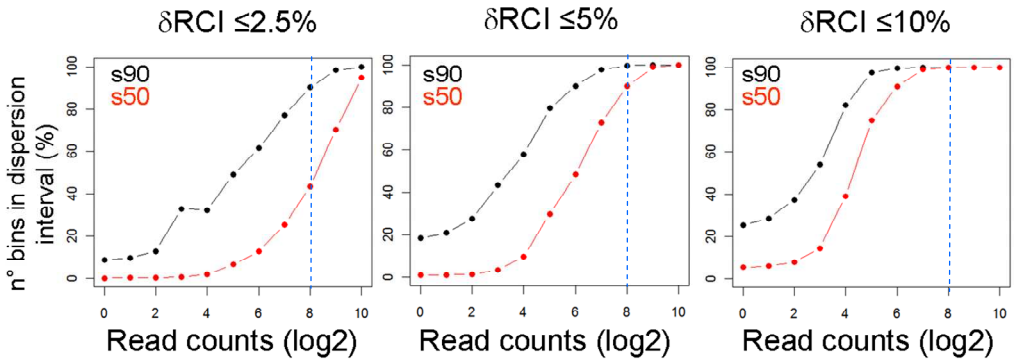


Figure Figure VI: Fraction of bins at different δRCI intervals. For a given RCI threshold the fraction of bins presenting a δRCI equal or lower than the indicated threshold (2.5%, 5% or 10%) is evaluated for s90 and s50 subsets.

3. Defining Global and local QC indicators

All previous analyses illustrate several characteristics associated to TMR distribution at different random sampling densities. Such characteristics represent a read-out for the quality of the evaluated profile, which is the consequence of several factors implicated in its genesis.

Below the global QC indicators, which represent "fingerprints" of an evaluated profile, and the corresponding acronyms are summarized:

- **Considered reads**: Mapped reads used for performing the random sampling analysis. Indeed, this parmetter could deviate from the total mapped reads (TMRs) if the "remove clonal reads" option is in use.

- **Density QC** (denQCi): The fraction of bins in the s90 or in s50 subsets with a δRCI lower than the default dispersion thresholds (2,5%; 5% and 10%). Note that the higher the density QCi is, the better is the quality of the associated profile. In the Results panel, only the denQCi for the s50 subset is displayed.

- **Similarity QC** [simQCi(s90/s50)]: Ratio between the density QCis for s90 and s50 subsets at the different dispersion thresholds. The simQCi reveals the similarity of the s90 and the s50 profiles. As a rule of thumb, the closer this value is to 1, the better is the quality of the studied profile.

Both the density and similarity QCis represent quantifiable NGS-profiles quality indicators, thus they can be used for comparative purposes as described below. Note that QC indicators associated to several publicly available NGS-generated profiles can be retrieved in our website: www.ngs-qc.org

Further supplementary information. Taken in consideration that the above analyses were computed for 500bp bins, the δRCI/bin data can be used to provide local QC indicators. Such information is provided by the NGS-QC Generator either in a wiggle or in a BED format and is available as part of the **supplementary data**. Note that by default bins with δRCI≤ 10%[2] are displayed in such local QC indicator files. **Figure VII** illustrates how the local QCis can be displayed in a heat-map format linked to the original read count intensity profile. This display option is useful to visualize the predicted δRCIs associated to a given chromatin region of interest. As part of the last version of the NGS-QC report, an example of genomic regions presenting the calculated local QCis is displayed at the bottom of the NGS-QC report. At this purpose, the user can define the genomic resolution at which the genomic display is generated. Furthermore, a second genomic display presenting a zoom-in view of the center of the first genomic display is also generated (50,000 nts resolution). In both cases the location of the RCI regions presenting dispersion lower than 10% is displayed in a heatmap context.

4. Interpretation of NGS-QC indicators

The quality indicators described by the NGS-QC Generator are derived from the question of how different a given NGS profile would be if only a subset of the total mapped reads were used? The underlying concept is that in the ideal case, the read counts intensities will decrease proportionally to the fraction of sampled reads. From this two quality indicators are derived.

[2] Local QC indicators in wiggle file format can be uploaded in the Integrated Genome Browser (IGB) and displayed in heat-map format; the corresponding BED file can be uploaded in the UCSC browser.

The ***density QC indicator (denQCi)*** makes reference to the fraction of the evaluated chromatin regions (sectioned into 500bp bins) that comply with this proportional within a defined dispersion margin, such as 5% at a sampling ration of 50% (i.e. δs50/5). The maximal theoretical value for denQCi is 100.

The ***similarity QC indicator (simQCi)*** refers to the fraction of chromatin regions which reveal a proportional decrease of RCIs in the subset sampled at 90% relative to that sampled at 50% and is given for a specified dRCI threshold (e.g., ds90/s50/5). The minimal theoretical value for simQCi is 1.

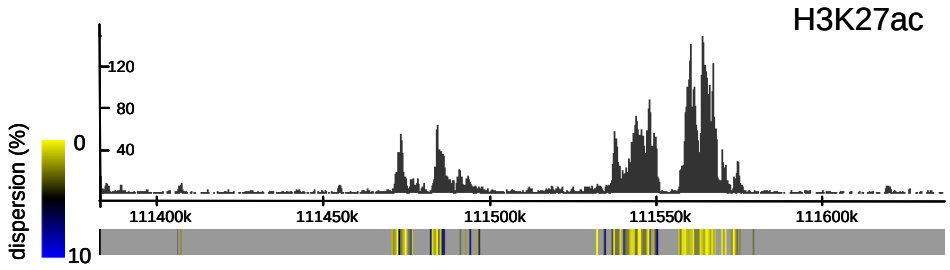


Figure VII: H3K27ac ChIP-seq profile displayed together with the corresponding local QC indicators. Bellow the ChIP-seq profile the corresponding δRCI for each 500bp bin are displayed for a 10% threshold using the heat map illustration indicated on the left. Only bins with δRCI ≤10% are shown.

4.1 denQCi and simQCi guide ChIP-seq experiments: **Figure VIII** illustrates that QC indicators can vary dramatically between experiments; indeed, publicly available ChIP-seq data provide useful information about the range of denQCi and simQCi that have been achieved in previous experiments for a given target and (batch of) antibody, such that a user can judge the QC performance of a ChIP-seq relative to past data sets. Moreover, the library of QC indicators (available at www.ngs-qc.org) provides a guide to users about the possible effect of the sequencing depth on ChIP-seq quality. Indeed, the comparison of several H4K20me1 profiles[3] demonstrates that at least 15 million total mapped reads are required to obtain QCis that differentiate between the ChIP-derived and the non-enriched ("input") datasets. In contrast, H3K4me3 ChIP-seq profiles present fairly good QCis even for TMRs lower than 15 million reads.

! ***NOTE*** Importantly, in both profiles individual ChIP-seq profiles can be observed which have been performed at similar sequencing depths but data analysis reveals nevertheless greatly varying global QCi indicators. This underscores the notion that in addition to the sequencing depth (multiple) other factors, whose effects cumulate along the experimental pathwtowards to final data set, influence the quality of the profile.

[3] The compared ChIP-seq profiles were taken from an study performed in nine human cell types following a production pipeline for chromatin immunoprecipitation (Ernst J. et al. 2011 Nature **473**; 43-49)

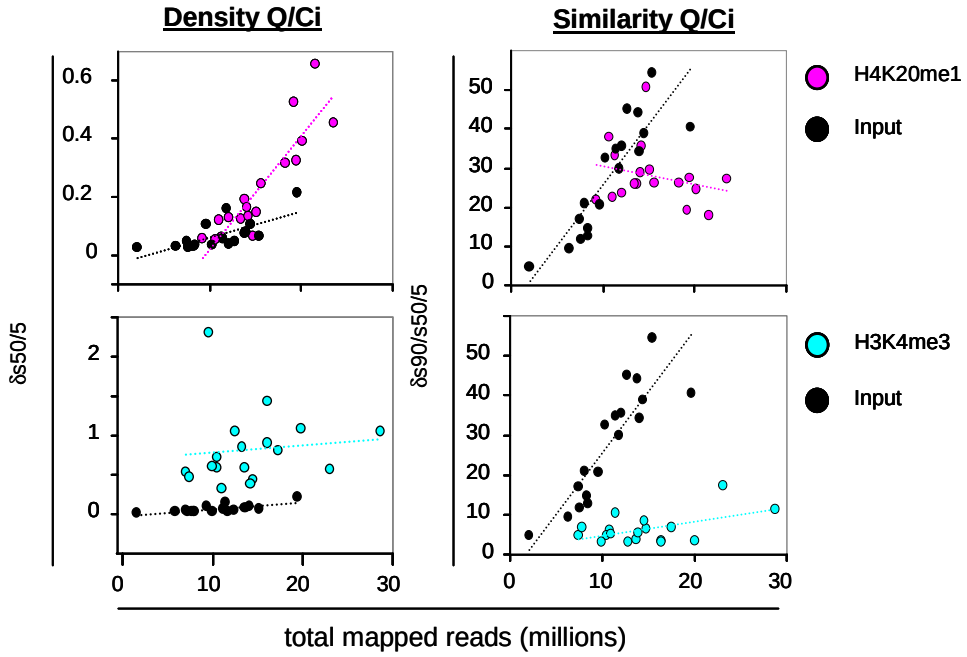


Figure VIII: Density and Similarity QCis for several ChIP-seq profiles in the context of their total mapped reads. Top and bottom panels illustrate QCis for H4K20me1 and H3K4me3 ChIP-seq profiles, respectively. In addition, QCis for the non-enriched input datasets are illustrated for comparative purpose. Notice that in contrast to the H3K4me3 datasets, H4K20me1 profiles reconstructed from up to 15 million reads present QCis similar to those observed for the input datasets. Importantly for such histone modification profiles, increase in the sequencing depth beyond this 15 million reads threshold allows to retrieve QC indicators diverging from the Input datasets behavior.

5. A dynamic publicly available database of global and local QC indicators

With the aim of establishing a dynamic guide for NGS users we have created a QC indicator database comprising a collection of global QCis for multiple NGS profiles. This database, which is available online to the scientific community through our website www.ngs-qc.org, will be expanded to include most, if not all, global and local QCis of the NGS profiles currently available from GEO. In addition, future profiles will be integrated and users may evaluate their NGS profiles and compare them with stored QCi.

To facilitate and simplify the recognition of QCi divergence between profiles we have defined QC-STAMP, a global descriptor that combines the information provided by denQCi and simQCi as following:

$$QC\_STAMP = \frac{denQCi}{simQCi}$$

In order to evaluate de divergence of this global descriptor over all enrichment-related NGS profiles currently compiled in the NGS-QC database, the QC-STAMP distributions assessed for three different RCI dispersion intervals was subdivided in four quantiles to which the following grades have been attributed: “D”, lower quartile (<25%); “C”, inter-quartile 25-50%: “B”, inter-quartile 50-75% and “A” upper quartile (>75%). The NGS-QC Generator database associates these grades for 2.5, 5 and 10% δRCI to each profile as a three letter symbol, such that, for example AAA (“triple A”) reveals an A grade for all three δRCIs. All available profiles are displayed as a dynamic QC-STAMP *vs.* TMR scatterplot, which allows judging of their QCi similarities in the context of the sequencing depth. Note that the global

QC-STAMP descriptor will be dynamically re-evaluated when novel entries are provided to the database.

Considering the inherent relationship between the current NGS repositories and our QC database, we aim to integrate in a long term a direct connection between the Galaxy version of the NGS-QCi Generator and the QCi database in order to simplify the repository of this information and to establish links with GEO in order to coordinate the generation of such indicators in a systematic manner[4].

6. Additional applications

The presented bioinformatics-based QC system uses the total mapped reads associated to any NGS data sets to infer a set of global QC indicators. In fact, profile's quality evaluation does not rely in a given Peak calling algorithm, thus it can be directly applied to any type of NGS-generated profile, including RNA-seq, GRO-seq, etc, in addition to the wide variety of ChIP-seq assays (transcription factors, insulators, histone modifications, RNA Polymerase II, etc). For the same reason, the inferred QC indicators are fully comparable, making of this approach a universal tool for multidimensional quality profiles comparison.
We believe that the global QC indicators will be useful for the development, characterization and comparison of antibodies directed towards a particular target. There are considerable variations between different antibodies and different batches of polyclonal antibodies. The certification of antibodies for ChIP-seq using the present QC systems should improve ChIP-seq reproducibility and comparability.

The quality of any NGS profile is the direct consequence of a complex number of factors, including aspects like crosslinking efficiency, chromatin shearing, antibody affinity and selectivity, as well as the variability between experiments and experimenters. While the QC indicators described here cannot *per se* identify the source for quality differences between profiles, they reveal the comparability and non-comparability of different NGS-generated profiles.

! ***NOTE*** The sequencing depth used to generating NGS-profiles can now be used as a tuneable parameter to identify profiles of similar quality. For this, correlative analyses between the inferred QC indicators and the performed sequencing depth will be very useful.

[4] The QCis generated in the current NGS-QCi Generator Galaxy version are not transferred into the QCi database, but in a further version we may establish such link; thus users will be invited to allow such a transfer. In addition, the identity of the sample will not be required, but certain information like the nature of the NGS profile, the antibody source, etc may be requested (without a mandatory condition) in order to associate a comprehensive description of the evaluated samples to the QCi data set.

7. Annexes

7.1 NGS-QC Generator availability: For providing a simple way to access to the community, NGS-QC Generator has been made available through a customized Galaxy cloud instance dedicate to this application (access provided in our website: www.ngs-qc.org).

Furthermore, an executable version of the NGS-QC Generator can be downloaded from our above indicated website. Importantly, such stand-alone version requires BEDtools to be installed on the hosting system. It can be retrieved at:

Stable releases: http://code.google.com/p/bedtools
Repository: https://github.com/arq5x/bedtools

A detailed description for the execution of such stand-alone version is available as part of the downloadable file.

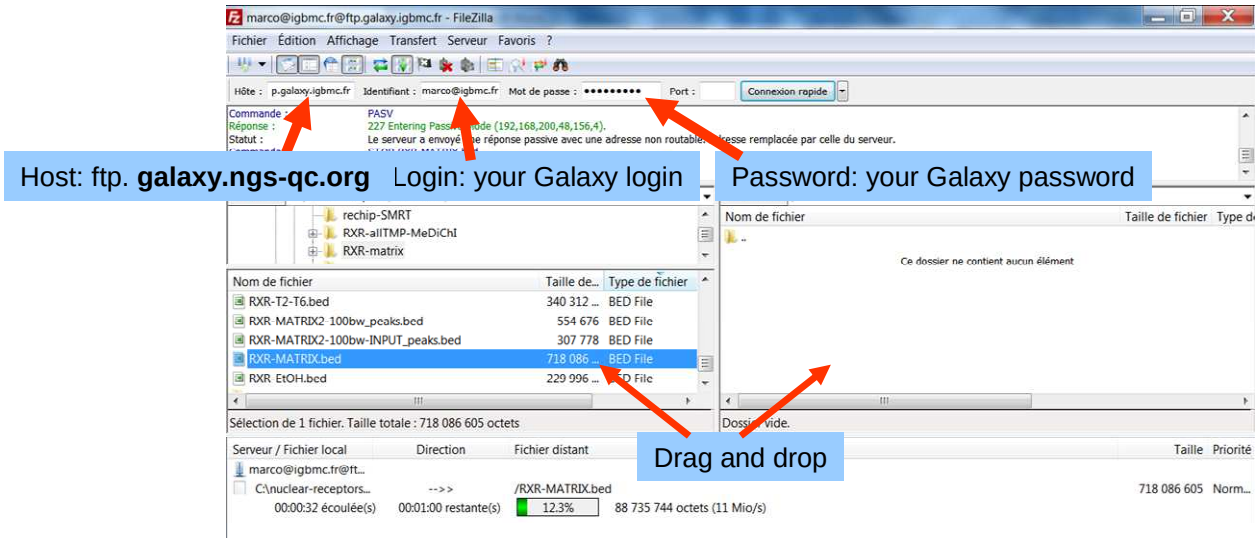
7.2 Practical aspects concerning the use of the NGS-QC Generator and database: The NGS-QC Generator and the corresponding QC indicator database are accessible from our website (www.ngs-qc.org). For assessing the quality of a dataset, users can access the NGS-QC Generator through our customised web-based GALAXY platform. For it users can register by providing an e-mail address as login and a password. This step is mainly required for the use of an FTP server to facilitate the uploading large size data files (see below).

Furthermore, due to storage space constraints, uploaded datasets into the Galaxy instance may not be available for more than 24hours, thus we strongly suggest users to download their processed files as early as possible.

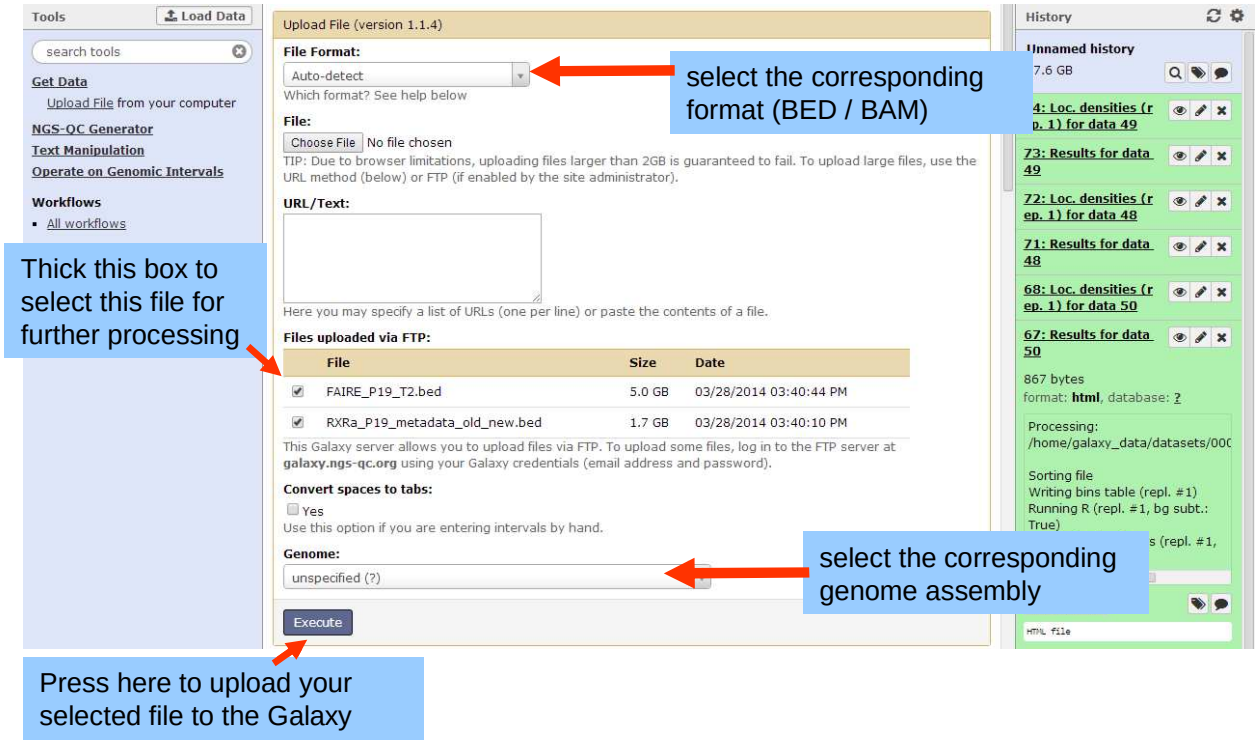
When required, some example datasets are available on the “shared_Data” access as part of the Data libraries, thus users may upload them for having a trial run on the NGS-QC generator tool.

Because the NGS-QC Generator tool requires mapped sequence files (in BED or BAM format which are quite large in size!!!) for processing, they should be uploaded by FTP. Furthermore, uploading compressed BED files is preferred to save time and space.

At this purpose, users may upload their datasets into the dedicated FTP server: **galaxy.ngs-qc.org** for instance by the cross-platform FTP software FileZilla (you can download FileZilla from here if required: https://filezilla-project.org/) as indicated below:



Once the transfer into the FTP server is done, you have to go back to the “Upload data” field in Galaxy and select your uploaded file for further processing.



At the end of this step the selected datasets will be displayed on the history panel (right side) in Galaxy ready to be used for running on the NGS-QC Generator tool.