Chapter 13

NGS-QC Generator: A Quality Control System for ChIP-Seq and Related Deep Sequencing-Generated Datasets

Marco Antonio Mendoza-Parra, Mohamed-Ashick M. Saleem, Matthias Blum, Pierre-Etienne Cholley, and Hinrich Gronemeyer

Abstract

The combination of massive parallel sequencing with a variety of modern DNA/RNA enrichment technologies provides means for interrogating functional protein–genome interactions (ChIP-seq), genome-wide transcriptional activity (RNA-seq; GRO-seq), chromatin accessibility (DNase-seq, FAIRE-seq, MNase-seq), and more recently the three-dimensional organization of chromatin (Hi-C, ChIA-PET). In systems biology-based approaches several of these readouts are generally cumulated with the aim of describing living systems through a reconstitution of the genome-regulatory functions. However, an issue that is often underestimated is that conclusions drawn from such multidimensional analyses of NGS-derived datasets critically depend on the quality of the compared datasets. To address this problem, we have developed the NGS-QC Generator, a quality control system that infers quality descriptors for any kind of ChIP-sequencing and related datasets. In this chapter we provide a detailed protocol for (1) assessing quality descriptors with the NGS-QC Generator; (2) to interpret the generated reports; and (3) to explore the database of QC indicators (www.ngs-qc.org) for >21,000 publicly available datasets.

Key words Next-generation sequencing, Massive parallel sequencing, Quality control, ChIP-sequencing, Galaxy, Database

1 Introduction

The rapid development of next-generation sequencing (NGS) technologies poses multiple challenges to the bioinformatics-based analyses of the enormous amounts of data, which still increase exponentially. While in the past years computational efforts focused mostly on the description of local enrichment confidence in NGS-generated profiles, a number of other critical issues that limit the conclusions drawn from multi-profile comparisons have been neglected or only incompletely addressed. NGS technology can be

Ewy Mathé and Sean Davis (eds.), *Statistical Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1418, DOI 10.1007/978-1-4939-3578-9_13, © Springer Science+Business Media New York 2016

used to address a variety of molecular events. This includes the epigenetic histone modifications and transcription factor–chromatin association (ChIP-seq), DNA methylation profiling (MeDIP-seq and related technologies), analysis of transcriptional activity either by RNA-Polymerase II profiling or by evaluating the nascent transcriptomes (GRO-seq and similar technologies), or studying the association of RNAs with (particular) ribosomes by ribosome capture technologies followed by RNA-seq. It is important to emphasize that the integrative analysis of several different datasets requires particular attention to profile pattern variability, which is highly affected by the choice of antibodies, sequencing depth, cross-linking procedures, immunoprecipitation (IP) efficiency, and many other parameters.

For this reason, we have developed the NGS-QC Generator [1], a bioinformatics-based quality control (QC) system that uses the raw NGS data sets to (1) infer a set of global QC indicators that reveal the comparability of different NGS data sets; (2) provide local QC indicators to judge the robustness of cumulative read counts ("peaks or islands") in a particular region; (3) provide guidelines for the choice of the optimal sequencing depth for a given target; and finally (4) to have quantitative means of comparing different antibodies and antibody batches for ChIP-seq and related antibody-driven studies.

Briefly, this computational approach is based on the assumption that under optimal conditions NGS-generated profiles might conserve the relative enrichment patterns when a subset of the total sequenced/mapped reads are used for their reconstruction. In this context, it generates read count intensity (RCI) profiles from randomly selected subsets of the total originally mapped reads (TMRs) associated to the NGS-profile under study and evaluates the divergence from the theoretically expected read count intensities recovery after sampling relative to the original profile (described as dRCI or RCI dispersion). While most if not all NGS-generated datasets diverge systematically from the hypothesized "ideal behavior," the extent of such divergence represents a quantifiable descriptor of the quality of any NGS-generated profile.

In this chapter, we provide a step-by-step procedure for evaluating the quality of any ChIP-seq and related dataset by taking advantage of the multiple random samplings procedure embedded into the NGS-QC Generator. Furthermore, we provide to the reader a guide to interpret the quality reports generated by this tool. Finally we provide a global survey of the NGS-QC Generator database currently hosting quality grades for more than 21,000 publicly available datasets, such that the reader is able to compare publicly available datasets in the context of their quality grades, and to use this information to reveal the performance of "ChIP-seq grade" antibodies.

2 Materials

This chapter is dedicated to the use of the NGS-QC Generator tool and its associated database, both available through the Web access (www.ngs-qc.org). Users do not need to install any software; datasets to be processed are uploaded to a dedicated FTP server. For the purpose of this protocol, several datasets are available in the "Shared Data/Data Libraries".

2.1 Data Description: Preliminary Datasets Quality Evaluation (FastQC)

For the purpose of this protocol, we have selected the ChIP-seq datasets GSM733737 and GSM733720 (Table 1) from the publicly available ENCODE [2] histone modifications collection in the Gene Expression Omnibus (GEO) repository [3]. Both datasets were generated using antibodies targeting the active histone modification mark H3K4me3 on HepG2 and NHEK cell-lines respectively. A typical H3K4me3 profile is expected to display sharp and highly enriched occupancy of reads around Transcription Start Sites (TSS) of active genes while the rest of the genome including inactive genes could contain background noises. Each of the selected datasets has multiple replicates but for demonstration we have selected "Replicate3" and "Replicate1" from NHEK and HepG2, respectively (highlighted in Table 1).

While the NGS-QC Generator takes aligned datasets as input, we consider that the evaluation of the quality of the sequencing itself is also an essential part of the analysis. Thus, to illustrate the procedure to follow, we have downloaded raw sequence data SRR227526 and SRR227563 from the Short Read Archive (SRA) database [4] corresponding to the datasets GSM733737 and GSM733720 respectively. Throughout this chapter, use of terms NHEK and HepG2 data refers to the replicate 3 of NHEK and replicate 1 of HepG2 data respectively.

Table 1
Summary of the ChIP-seq datasets used in the illustrations

GEO ID	SRA ID	Replicate	Histone mark	Cell line	Raw reads	% Aligned reads	% Unique reads
GSM733720	*SRR227525*	Rep1	H3K4me3	NHEK	20855214	58.939908	93.03
GSM733720	*SRR227527*	Rep2	H3K4me3	NHEK	7121361	75.289583	95.92
GSM733720	*SRR227526*	*Rep3*	*H3K4me3*	*NHEK*	*6304701*	*69.269772*	*85.32*
GSM733737	*SRR227563*	*Rep1*	*H3K4me3*	*HepG2*	*10228152*	*87.698785*	*91.15*
GSM733737	*SRR227564*	Rep2	H3K4me3	HepG2	11600046	83.196291	95.92

2.2 Data Description: Datasets Format for NGS-QC Generator Analysis

NGS-QC Generator accepts the most commonly used formats of alignment, BAM and BED. In fact, most of the aligners reports the alignment output in a huge plain text SAM file format, which after compression and sorting generates BAM files. BED is a minimal format to represent the alignment information where only six columns (chromosome, start, end, score, id, strand) are used to represent each alignment. As this minimal information is enough for NGS-QC Generator, it is highly advised to convert other formats to BED format. Furthermore, providing sorted and compressed BED files to the NGS-QC Generator might be suitable since it saves time during the ftp upload and speed the analysis as well.

There are different tools and methods available to convert alignment file to an acceptable format. Samtools [5] can be used to convert SAM to sorted BAM as following:

```
samtools view -bS input.sam | samtools sort - output.srt
```

Use –q option to exclude alignments with low mapping quality in conversion, and with versions >=0.1.19 multithreading is available with -@ option. Use the input file name in the place of input.sam and output.srt and the output will be generated with automatic suffix .bam.

In a similar manner, Bedtools [6] can be used to convert BAM to BED:

```
bedtools bamtobed -i input.bam | sort -k 1,1 -k 2,2n -k 3,3n
-k 6,6 | gzip -f >output.srt.bed.gz
```

Finally, a simple awk script can be used to convert from bowtie default format to bed format:

```
awk -F "\t" 'BEGIN{OFS="\t";}{len=$4+length($5); print
$3,$4-1,len,$1,"0",$2}' input.bow bam | sort -k 1,1 -k
2,2n -k 3,3n -k 6,6 | gzip -f >output.srt.bed.gz
```

This script can be used to convert from any other alignment text files by changing the appropriate columns in the script. These are the corresponding fields from bowtie output:

$3 = Chromosome; $4 = Start position; $5 = Aligned sequence; $1 = Read ID; $2 = Strand.

The identity of the different columns can be retrieved by examining the first lines of the file to convert by using the "head" command and change the script and variable numbers accordingly (*see* **Note 1**).

3 Methods

3.1 Preliminary Sequencing Quality Assessment

Before stepping into the analysis of experiment quality based on enrichment with the NGS-QC Generator, it is important to check the quality of the DNA sequencing itself, as this factor can directly influence the results.

We currently use FastQC [7] to evaluate the DNA sequencing quality and potential contaminations. FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines prior to alignment. It can be used through simple GUI interface or through command-line for easy batch processing and the tool can handle different formats like FASTQ, SAM, and BAM file. FastQC generates a HTML output containing a variety of graphical displays illustrating the quality of the data under different aspects. Here we highlight the important modules that require close investigation from FastQC results to assess better the quality of the data.

3.1.1 Sequence Quality

"Per base sequence quality" and "Per sequence quality scores" are the two important plots that has to be paid attention to understand the sequencing quality of the data. The first plot is a boxplot illustrating the average, median, and distribution of quality per base from all the reads in the data (Fig. 1a). As Illumina's sequencing chemistry is sequencing per cycle approach, at each cycle one base of all reads are sequenced, thus the 36 cycle kit yields all reads with 36 bp length. While more recent versions of Illumina sequencing kits generate much longer reads (50 to even 150 bp length); a FastQC analysis will still be helpful to identify potential abnormalities during the sequencing process.

Hence it is expected to have dip or hike in quality at or from a particular base on in most of the sequences due to technical issues. This pattern can be easily observed in "Per base sequence quality" box plot with average and distribution of quality per base from all reads. It is important to remember that due to the reagents "burning" out over the period, quality of bases towards the end of the reads result in the gradual decrease of average quality in the plot.

While an average view of the read sequences provides a first way to evaluate sequence quality, the presence of reads with poor quality levels will affect the alignment accuracy to the reference genome (false or multiple alignment outcomes). One can address this problem by using the "Per sequence quality scores" plot which shows the distribution of each read average quality to get an idea about altogether bad reads.

Unlike the "per base distribution" approach, the analysis in a "per sequence quality" context allows to estimate the fraction of reads with low quality. It is important to note from Fig. 1a, that the NHEK dataset presents an overall poor sequencing quality, which is also reflected by the lower percentage of mapped reads in comparison to that observed for HepG2 datasets (Table 1).

Note that tools like FASTX-toolkit [http://hannonlab.cshl.edu/fastx_toolkit/], NGSQC-toolkit [8], or SeqQC [http://genotypic.co.in/Products/7/Seq-QC.aspx] are dedicated to trim and filter reads presenting low quality scores (*see* **Note 2**).

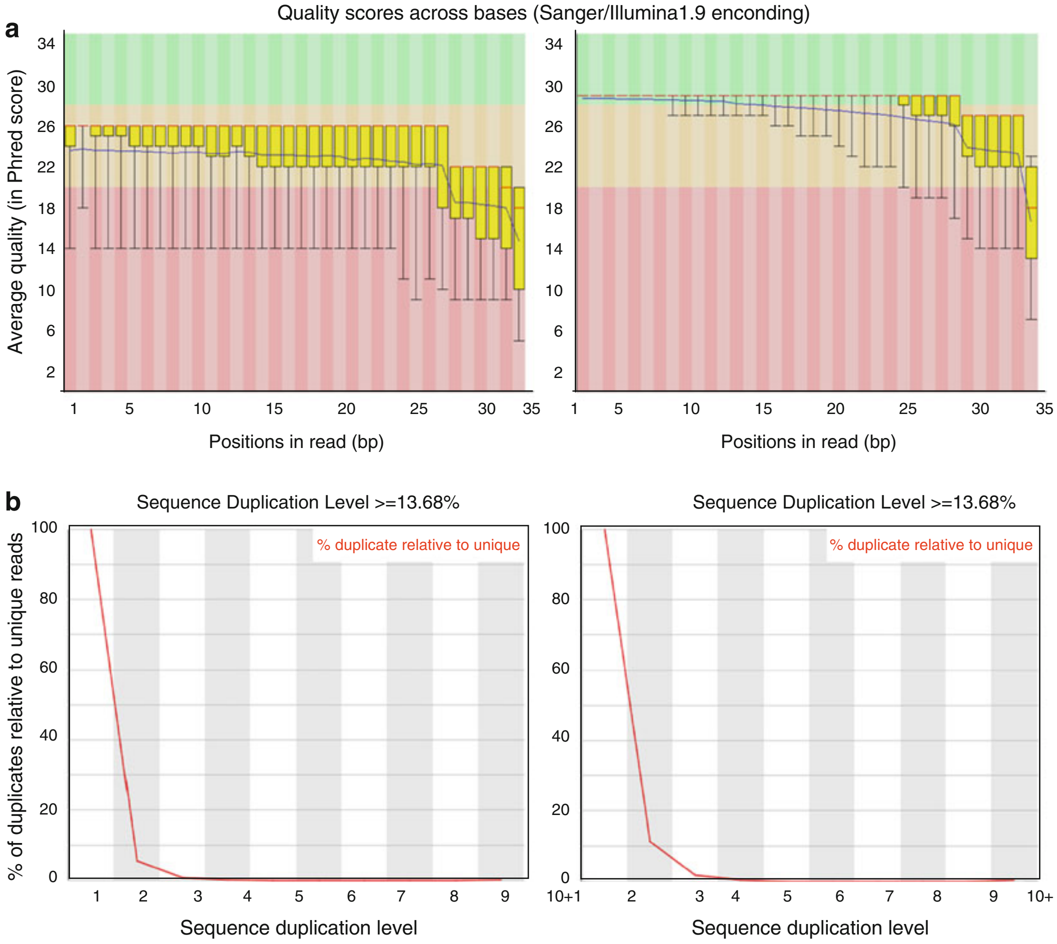


Fig. 1 (**a**) FastQC generated "Per base sequence quality" plot for GSM733720 NHEK (*left panel*) and GSM733737 HepG2 datasets (*right panel*). (**b**) FastQC generated "Sequence Duplication Level" for both datasets as in (**a**). Both the data show similar harmless level of clonal reads

3.1.2 Clonal Reads

"Sequence duplication level" is important aspect to understand the quality of data rather than the sequencing itself. When there is low amount of starting material to start with or by over-amplifying the fragments during library preparation could lead to high levels of sequence duplication. As clonal reads bias the analysis by over-representing the same fragment, they can generate false-positive enrichments. Hence, it is highly recommended to remove duplicate reads and keep only one copy.

Also it is important to point out that the currently existing methods to detect clonal reads from single-end sequencing assays may overestimate their presence. Clonal reads are defined as reads aligned with the same start and end position; indeed an analysis in which the two reads generated by paired-end sequencing from each DNA were treated separately as single reads demonstrated that reads considered as clonal in a single-end sequencing context can have different alignment coordinates in the corresponding pair-read analysis (data not shown).

It is highly recommended to remove clonal reads after alignment as sequencing errors could bias the similarity estimation among reads when it is carried out prior alignment. Hence, the FastQC sequence duplication level can be used for a rough assumption only, as it is done with raw data and only the first million reads are considered for the calculation. This deviation can be observed between FastQC sequences duplication level from Fig. 1b with the actual unique reads count from Table 1 (*see* **Note 3**).

3.1.3 Adapter Contamination

When fragmented DNA molecules are shorter than the applied sequencing length, DNA sequencing will continue over the adapter sequence at the end of the process. This is rare as most of the chromatin immunoprecipitated DNA is in average larger than several hundreds of nucleotides but when it happens (e.g., over-sonication; sequencing library primer-dimer contamination) it can be detected by the counts of "Overrepresented sequences." If there are bulk of reads with adapter contamination, tools like Cutadapt [9], FASTX-toolkit, NGSQC-toolkit, or SeqQC can be used to trim the adapter sequence but retain the actual DNA sequence.

3.2 Exploring the NGS-QC Generator Portal

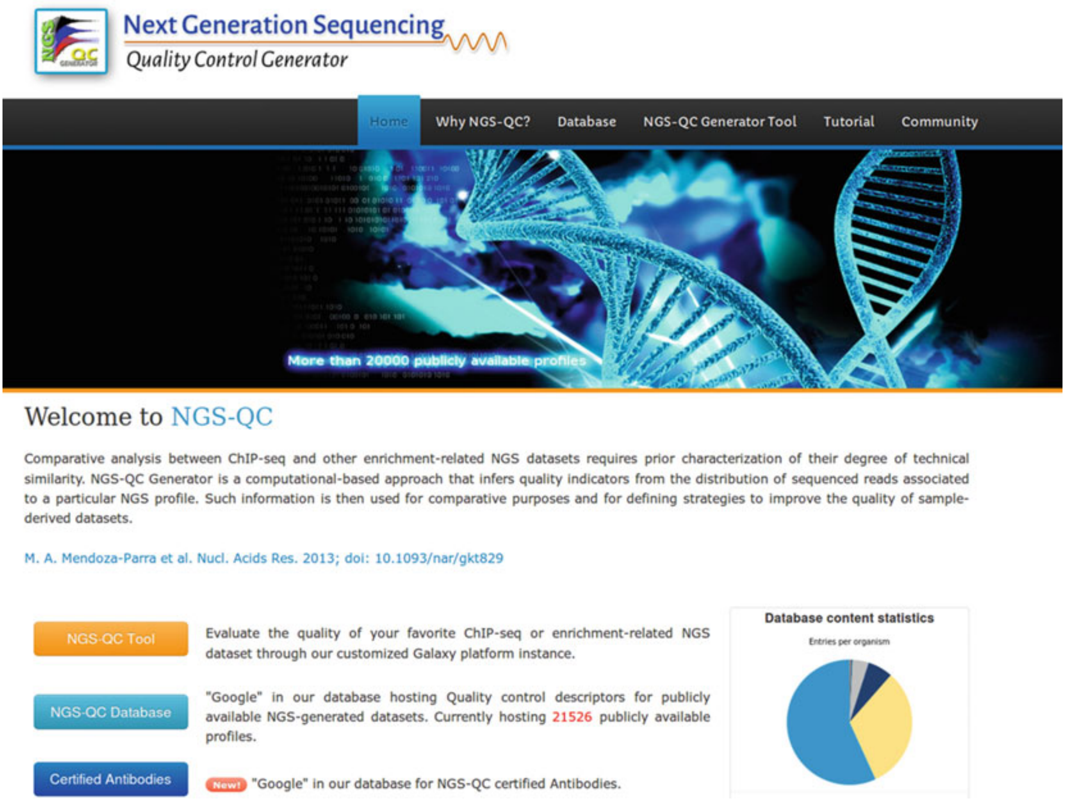
NGS-QC can be accessed through the portal www.ngs-qc.org, which gives access to different elements of NGS-QC method. Among the various components available in the main page we can cite the direct access to the customized Galaxy instance hosting the NGS-QC Generator tool, the access to the NGS-QC database as well as to a detailed tutorial describing the principles in use for assessing quality scores and their interpretation in the context of the enrichment behavior (Fig. 2).

3.3 Accessing to the NGS-QC Generator Tool via the Dedicated Galaxy Instance

3.3.1 An Overview of the Galaxy Interface

Galaxy [10–12] is an open, Web-based platform for biological data analysis. It enables to run bioinformatics tools in a user-friendly point-and-click environment. The NGS-QC Generator is available through a local Galaxy server reachable at http://galaxy.ngs-qc.org/. It is worth to mention that any user requires to create a login account such that a dedicated space is allocated to his jobs in the NGS-QC Generator associated server. Once logged in, users can access to the main Galaxy interface composed of three sections (as illustrated in Fig. 3):

1. A panel containing the list of available tools, grouped by field of analysis (left side).
2. A central panel where the interaction with the selected tools is enable as well as the display of the generated results.
3. A panel containing the history of data, requested jobs and results (right side). It is important to mention that this last panel



Welcome to NGS-QC

Comparative analysis between ChIP-seq and other enrichment-related NGS datasets requires prior characterization of their degree of technical similarity. NGS-QC Generator is a computational-based approach that infers quality indicators from the distribution of sequenced reads associated to a particular NGS profile. Such information is then used for comparative purposes and for defining strategies to improve the quality of sample-derived datasets.

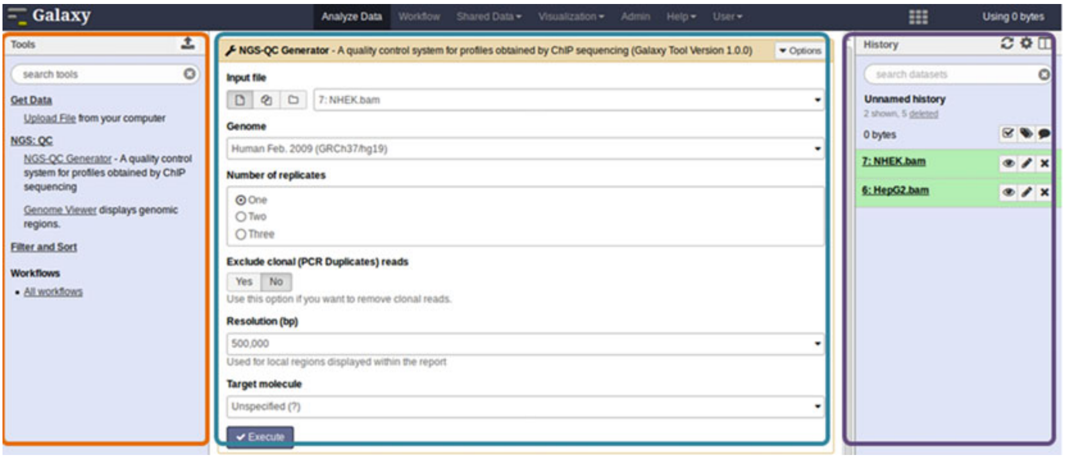
M. A. Mendoza-Parra et al. Nucl. Acids Res. 2013; doi: 10.1093/nar/gkt829

Evaluate the quality of your favorite ChIP-seq or enrichment-related NGS dataset through our customized Galaxy platform instance.

"Google" in our database hosting Quality control descriptors for publicly available NGS-generated datasets. Currently hosting 21526 publicly available profiles.

"Google" in our database for NGS-QC certified Antibodies.

Fig. 2 Home page of ngs-qc.org website. Notice that as part of the main display, access to the NGS-QC Generator tool and to the dedicated database is available. Furthermore, an access to the database content statistics is illustrated such that visitors could have a glance over the variety of the publicly available datasets currently hosted by this website



NGS-QC Generator - A quality control system for profiles obtained by ChIP sequencing (Galaxy Tool Version 1.0.0)

Input file: 7: NHEK.bam

Genome: Human Feb. 2009 (GRCh37/hg19)

Number of replicates: ● One ○ Two ○ Three

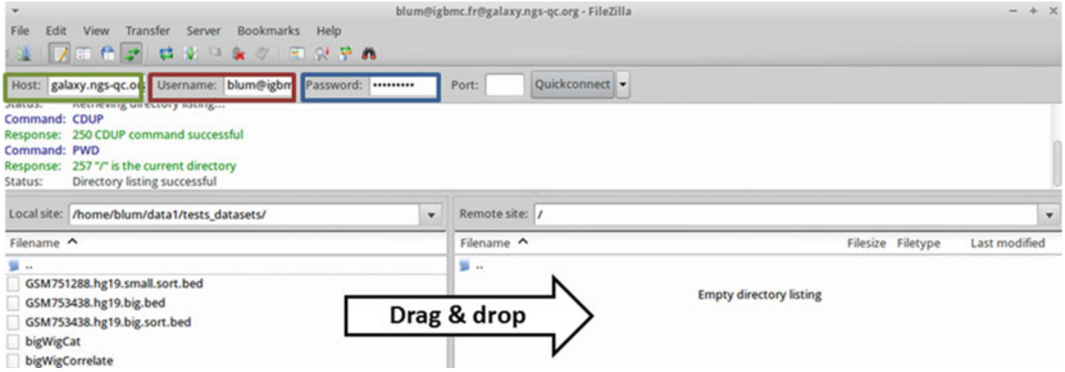
Exclude clonal (PCR Duplicates) reads: Yes [✓ No]
Use this option if you want to remove clonal reads.

Resolution (bp): 500,000
Used for local regions displayed within the report

Target molecule: Unspecified (?)

Unnamed history
2 shown, 5 deleted
0 bytes
7: NHEK.bam
6: HepG2.bam

Fig. 3 Illustration of a classical Galaxy interface composed by the "tools" panel (*left side*, demarcated in *red*), the central panel where tool parameters and results are displayed, and the history panel (*right side in purple*) containing the datasets available. In this example, the current history panel contains two datasets imported from the "Shared Libraries"


blum@igbmc.fr@galaxy.ngs-qc.org - FileZilla
File Edit View Transfer Server Bookmarks Help
Quickconnect
Drag & drop

Host: galaxy.ngs-qc.o
Username: blum@igbm
Password: ••••••••••
Port:

```
Command:  CDUP
Response: 250 CDUP command successful
Command:  PWD
Response: 257 "/" is the current directory
Status:   Directory listing successful
```

Local site: /home/blum/data1/tests_datasets/

Remote site: /

Filename
..
GSM751288.hg19.small.sort.bed
GSM753438.hg19.big.bed
GSM753438.hg19.big.sort.bed
bigWigCat
bigWigCorrelate

Filename	Filesize	Filetype	Last modified
..			

Empty directory listing

Fig. 4 Screenshot of the FileZilla interface. An FTP connection can be established by filling the "Host" field with "galaxy.ngs-qc.org" and the Galaxy user/password in the "Username"/"Password" fields, respectively, then clicking the "Quickconnect" button. Dragging files from the *left-side panel* and dropping them onto the *right-side panel* will start the upload

provides the history of every dataset/job associated to the user, thus a continuous cleaning of its content is a good practice to avoid overloading the allocated space in our servers (see below).

Running a tool on Galaxy requires importing the dataset of interest into their history. Locally stored files can be uploaded either through the user's Web browser, or by using our FTP server; nevertheless we might strongly recommend using the second option in the case of large size datasets (larger than 2Gbs). In the case of first option, locally stored files can be uploaded to Galaxy by clicking on "Get Data" in the tool panel, and select "Upload File". Either one can drag and drop files from their desktop file manager to the Galaxy interface, or select files from the local disk by clicking the "Choose local file" button. Once selected, click the "Start" button to start the upload.

To upload files by FTP, a connection to our FTP server with an FTP client is required (e.g., *FileZilla* [https://filezilla-project.org/]; an open source and cross-platform tool). In the FTP client, the host has to be given as "galaxy.ngs-qc.org" and users should use their Galaxy credentials for login (Fig. 4). Once the files are uploaded, they can be imported into the history by clicking on "Get Data", "Upload File", and "Choose FTP file" on the Galaxy left panel. Datasets successfully imported will be shown in green color box in the history panel (right side).

Though our Galaxy server is equipped with high computing capacity and storage space for handling multiple users, we cannot keep each and every dataset uploaded. For this reason, files, which are older than a month will be automatically removed from Galaxy. In case users are running out of space, they can reduce their disk usage by permanently deleting datasets they do not need anymore.

This is a two steps action: delete the unwanted dataset by clicking the "Delete" ("X") button in the history panel, then click on the history gear icon and click on "Purge deleted datasets".

The NGS-QC Galaxy instance is currently hosting a variety of tools generated by the Galaxy community (Text manipulation, Operate on Genomic Intervals) but also two dedicated tools for the assessment of Quality descriptors for ChIP-sequencing and enrichment related datasets, available under the "NGS: QC" group in the tool panel:

1. The *NGS-QC Generator* computes global and local quality descriptors by performing multiple random sub-samplings of the mapped reads retrieved in the provided dataset, to then use them for reconstructing and comparing their enrichment patterns with the original profile.
2. The *Local QCs viewer* allows users to visualize the read count intensity levels retrieved in their datasets in the context of the computed local quality descriptors. Regions to display can be chosen by the user either by (a) gene names, (b) genomic coordinates or (c) let the tool to select these regions based on local QC score ranking.

3.3.2 Performing a Global Quality Control Analysis

As mentioned earlier, we are going to run NGS-QC Generator on NHEK and HepG2 datasets, available in the "Shared libraries". To use these datasets from shared libraries, users have to import these datasets into their history panel as following:

1. Click "Shared Data" at the top of the page;
2. Navigate to "Data libraries", "Examples";
3. Select both "NHEK" and "HepG2" datasets;
4. Finally select "Import to current history" in the "For selected datasets" menu; click the "Go" button, and go back to the main page by clicking on "Analyze Data".

Now the datasets are into the history, we can run the NGS-QC Generator:

1. Click on "NGS-QC Generator" under "NGS: QC" in the tool panel
2. In the central panel, select one or more (batch mode) datasets, and the genome assembly; in this case hg19. The following options are also available:
   (a) Generate up to three technical replicates, to evaluate the results variability.
   (b) Remove the clonal reads (potential PCR duplicates; by default it is not removed).

(c) Exclude background noise (users can switch-off this option if interested in assessing its influence in QC scoring; but by default it is active).

(d) Select the resolution for the read-counts intensity representation of displayed genomic regions.

(e) Select the sample's antibody target. If the user provides the target identity, the output report will contain a scatter-plot showing a comparison of the evaluated dataset's quality with those retrieved in the NGS-QC collection.

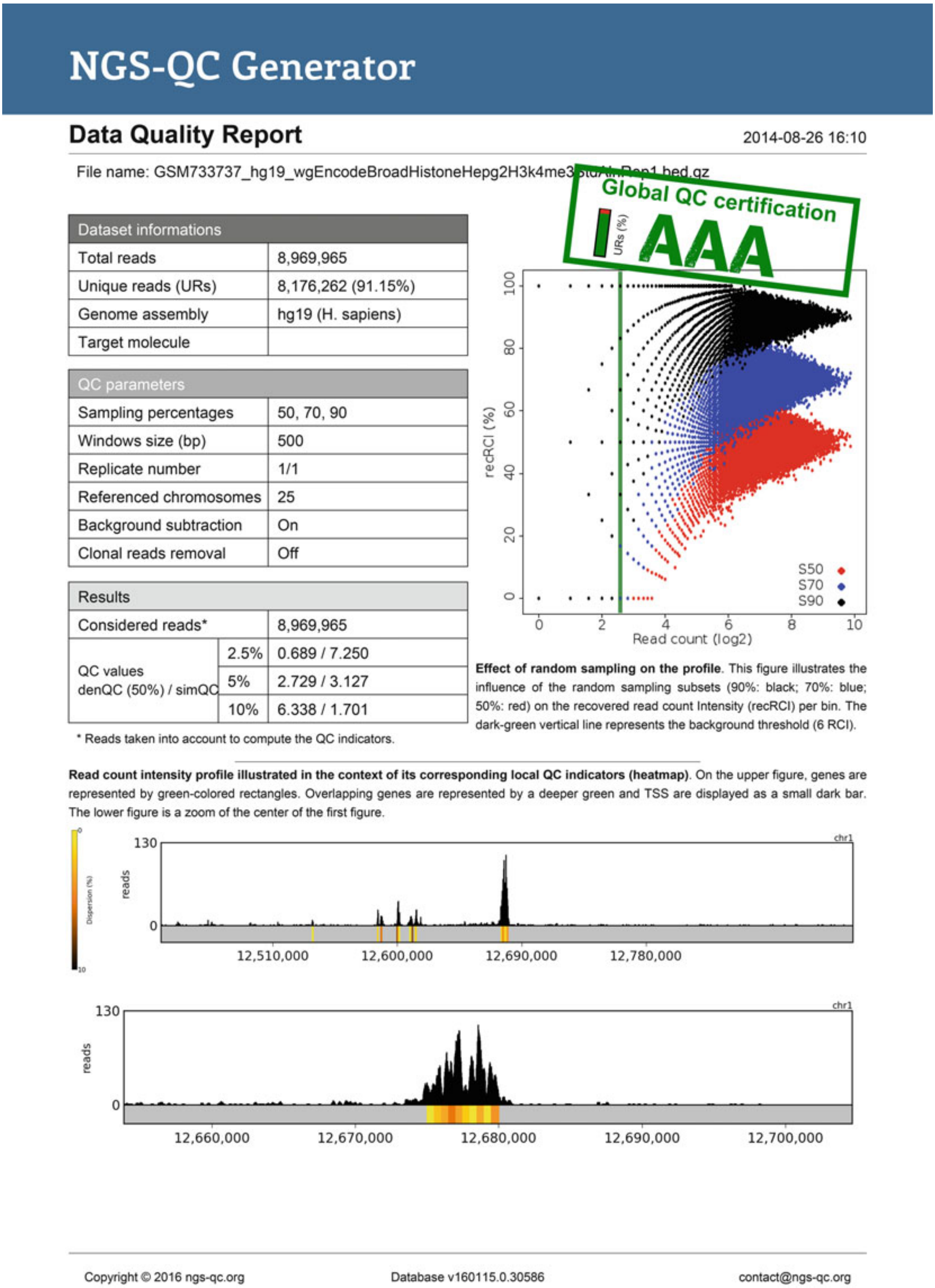
3. Click the "Execute" button to launch the job.

For each run the NGS-QC Generator produces two outputs which are made available in the history panel. The first one is an HTML page linking to the PDF report that summarizes the quality descriptors, and a ZIP archive containing several supplementary files.

A NGS-QC Generator report (Fig. 5) is composed by three major elements:

1. *Dataset information* where characteristics of the dataset under analysis like the total mapped reads, the fraction of unique mapped reads (i.e., without the clonal sequences), reads' length size, the genome assembly, and the target molecule (when available) are indicated.
2. *QC parameters* where a detailed overview of the various parameters used for generating the corresponding QC report are listed such that it could be reproduced if required.
3. *QC results* where a compilation of the global quality indicators assessed in the context of the considered mapped reads are displayed (either with or without clonal reads). The result panel is complemented with a scatter-plot displaying the read counts per genomic region (500 bp bin; *x*-axis) in comparison to the recovered counts after multiple reads' random subsampling (90 %, 70 %, and 50 % subsampled reads; *y*-axis). Furthermore a global QC certification score (from "AAA" to "DDD" for designating from high and low quality datasets) is stamped over the main page of the QC report such that the quality of the analyzed dataset is expressed in a rather intuitive manner without the need of getting deep into the assessed quality scores.

Each QC report is further complemented with series of genomic regions displays illustrating the enrichment patterns associated to the dataset under analysis complemented by the demarcation of genomic areas presenting high enrichment robustness (local QC indicators; heatmap display). In addition, when the target molecule identity is provided, a scatter-plot displaying the total mapped reads



NGS-QC Generator

Data Quality Report

2014-08-26 16:10

File name: GSM733737_hg19_wgEncodeBroadHistoneHepg2H3k4me3StdAlnRep1.bed.gz

Global QC certification

AAA

Dataset informations	
Total reads	8,969,965
Unique reads (URs)	8,176,262 (91.15%)
Genome assembly	hg19 (H. sapiens)
Target molecule	

QC parameters	
Sampling percentages	50, 70, 90
Windows size (bp)	500
Replicate number	1/1
Referenced chromosomes	25
Background subtraction	On
Clonal reads removal	Off

Results		
Considered reads*		8,969,965
QC values denQC (50%) / simQC	2.5%	0.689 / 7.250
	5%	2.729 / 3.127
	10%	6.338 / 1.701

* Reads taken into account to compute the QC indicators.

Effect of random sampling on the profile. This figure illustrates the influence of the random sampling subsets (90%: black; 70%: blue; 50%: red) on the recovered read count Intensity (recRCI) per bin. The dark-green vertical line represents the background threshold (6 RCI).

Read count intensity profile illustrated in the context of its corresponding local QC indicators (heatmap). On the upper figure, genes are represented by green-colored rectangles. Overlapping genes are represented by a deeper green and TSS are displayed as a small dark bar. The lower figure is a zoom of the center of the first figure.

Copyright © 2016 ngs-qc.org Database v160115.0.30586 contact@ngs-qc.org

Fig. 5 First page of the NGS-QC report for the HepG2 dataset (GSM733737; replicate 1). On the *upper right side* of the page, the global quality stamp is displayed along with a colored bar representing the proportion of unique reads(URs) in the dataset

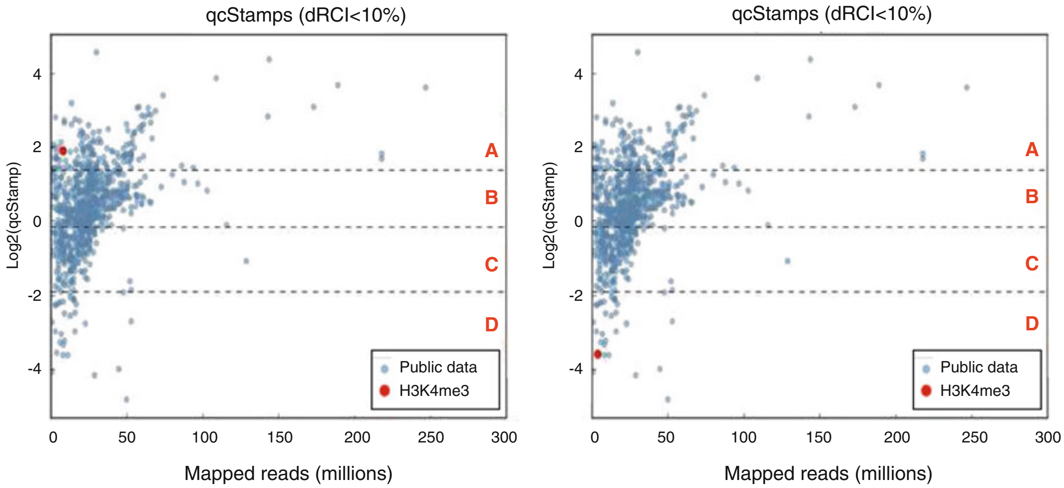


Fig. 6 Quality scores (*y*-axis) relative to their total mapped reads (*x*-axis) assessed for several public datasets (*blue*) in comparison to the dataset under evaluation (*red*). In the *left panel* this analysis has been performed for the HepG2 dataset; while in the *right side* it is illustrated for the NHEK dataset. This type of displays are included int he last page of the QC-report when the target molecule identity is provided

versus the quality descriptors assessed for entries available in the NGS-QC database in comparison with evaluated dataset is included at the end of the QC report (Fig. 6).

In addition to the QC report a set of supplementary files are generated (ZIP archive format). This supplementary folder contains further genomic regions display and local QC regions in either wiggle or BED format files such that they can be uploaded into a genome browsers like IGB [13] or on the UCSC Genome Browser [14]. Furthermore, a text file, referred here as "local Qci file", containing all genomic regions presenting enriched regions with high robustness (dRCI < 10 %) is also included in this supplementary file. This last file is also created in the current history, since it is required to generate multiple genomic regions displays (*see* Subheading 3.3.3).

3.3.3 Visualize Local Enrichment Patterns in the Context of Their Quality

Considering the interest of users to visualize genomic regions as a way to confirm the potential low or high quality of the dataset under study, we have developed a convenient way to display them together with a demarcation enriched regions based on their robustness or local QC indicators score. For it, our new tool called "*local QCs viewer*", takes alignment file (BED or BAM format) and the previously generated local QCi file as input to produce up to 25 genomic regions displays (PDF report format), with three selection strategies. By default, regions to be displayed are selected based on the local QC scores, the read count intensities and the presence of genes near a highly enriched region. Users have also the possibility to provide a list of genes or genomic positions as a way to customize these displays. As performed for the QC reports, genomic regions display enrichment patterns associated to the dataset under analysis complemented by the demarcation of genomic areas presenting high enrichment robustness (local QC indicators; heatmap display) (Fig. 7).

3.4 Exploring the NGS-QC Generator Database

3.4.1 A General Overview

In addition to the access to NGS-QC Generator tool, users can retrieve a large collection of quality indicators computed for a variety of publicly available datasets in the dedicated website. To access it, users have to select either "Database" on the top navigation bar or "NGS-QC Database" on the main page of the NGS-QC portal. There are two ways to browse our results; either by using the proposed query panel or through the interactive boxplots table located below. The *query panel* (Fig. 8a) allows specifying the request by multiple options like the model organism, target molecule, quality grades and also a public identifier from GEO database (GSM or GSE) or from ENCODE consortium (wgEncode). Importantly, each of these query options can be used in combinations. For example users can query for all the qualified datasets corresponding to mouse (*Mus musculus*) and human (*Homo sapiens*) model systems targeting the histone modification mark H3K4me3 and presenting quality grades between "A" and "B" (Fig. 8).

The *boxplot table* displays quality scores (QC-Stamp; dRCI < 10 %) distribution assessed over the whole database content (currently more than 21,000 datasets) as well as the discretized QC-STAMP intervals (from A to D). Furthermore, the quality score distribution per target molecule are displayed such that users might have a global overview of their associated quality scores. It is worth to mention that in addition to display each of these distributions under a boxplot format, a violin plot is also displayed such that a more detailed view of the distribution of population is available. Each target molecule in boxplot is complemented by a legend indicating its identity as well as the number of entries available (target molecules represented by less than 10 datasets are categorized as "Others"). By clicking on any of the target molecule boxplots, users are redirected to the results panel associated to that target molecule such that further refinement queries could be performed if required.

For instance by clicking on the boxplot associated to histone modification mark H3K4me3, users are redirected to the results page where the QC scores for the datasets associated to this target are displayed.

The results page (Fig. 9) is composed of five elements:

1. *Query panel* displaying the current request made.
2. *Boxplot table* displaying the QC scores distribution for each of the targets included in the request.

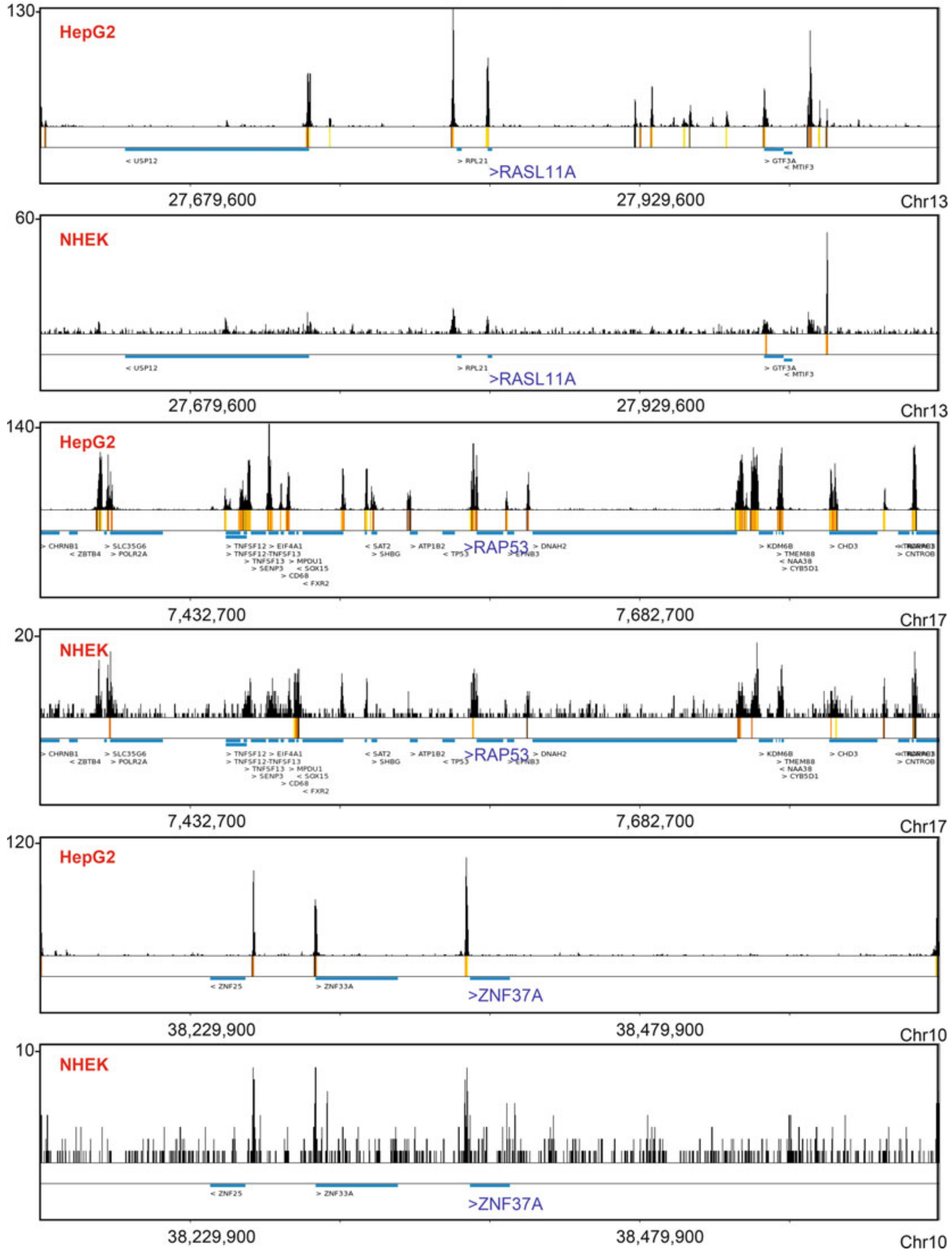
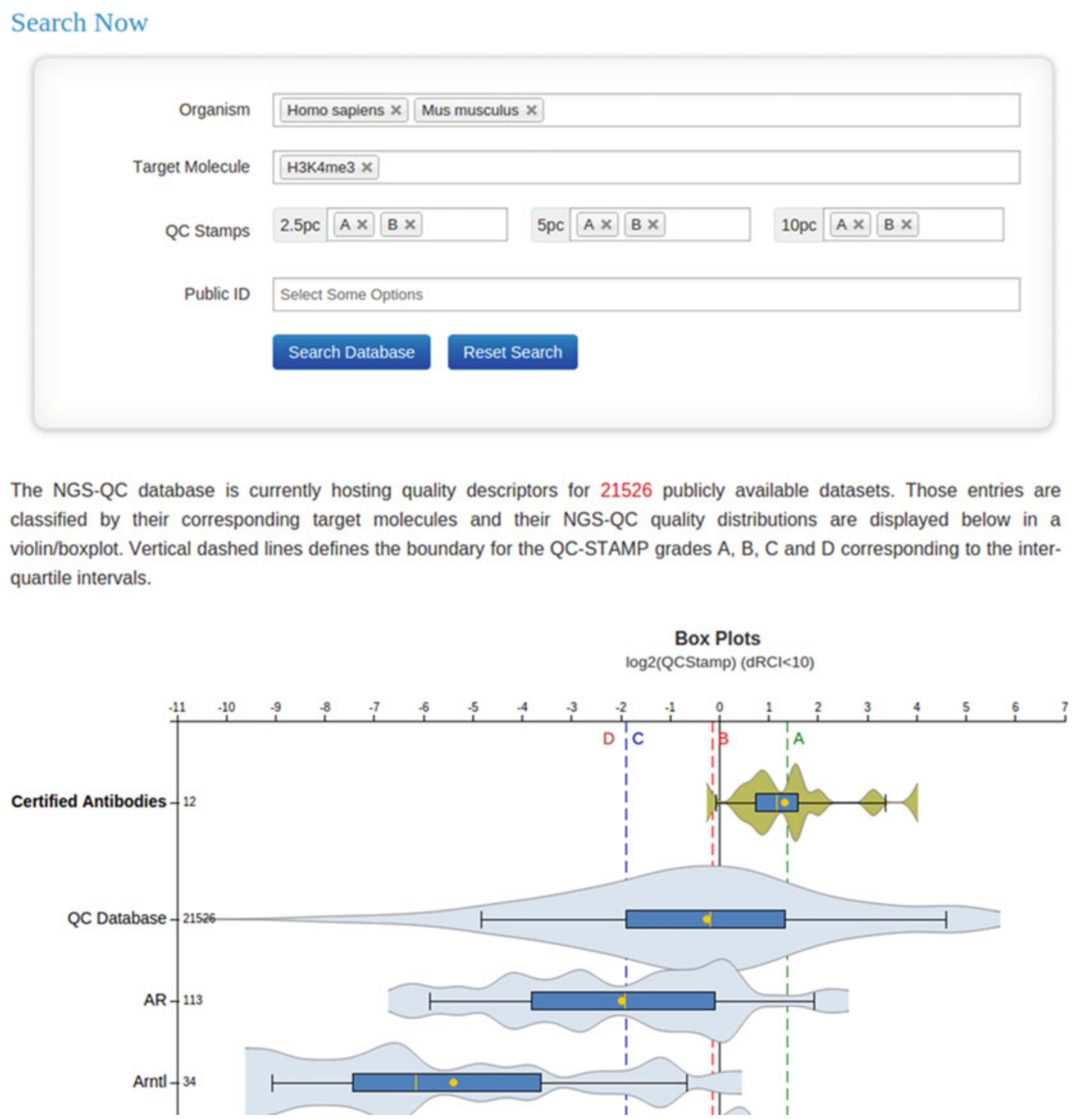


Fig. 7 Comparison of three genomic regions from HepG2 (*top*) and NHEK (*bottom*). Regions were selected using the RASL11A, WRAP53, and ZNF37A genes. Each plots contains three parts: (1) *top*, the representation of the read-counts intensity, (2) *center*, the position of the local regions with a dispersion between 0 % (*yellow*) and 10 % (*black*) and the genes' positions. The HepG2 plots present conserved local qc regions with highly peak intensity. On the contrary, NFEK plots are poor in enriched regions



Search Now

Organism: Homo sapiens × Mus musculus ×

Target Molecule: H3K4me3 ×

QC Stamps: 2.5pc A × B × 5pc A × B × 10pc A × B ×

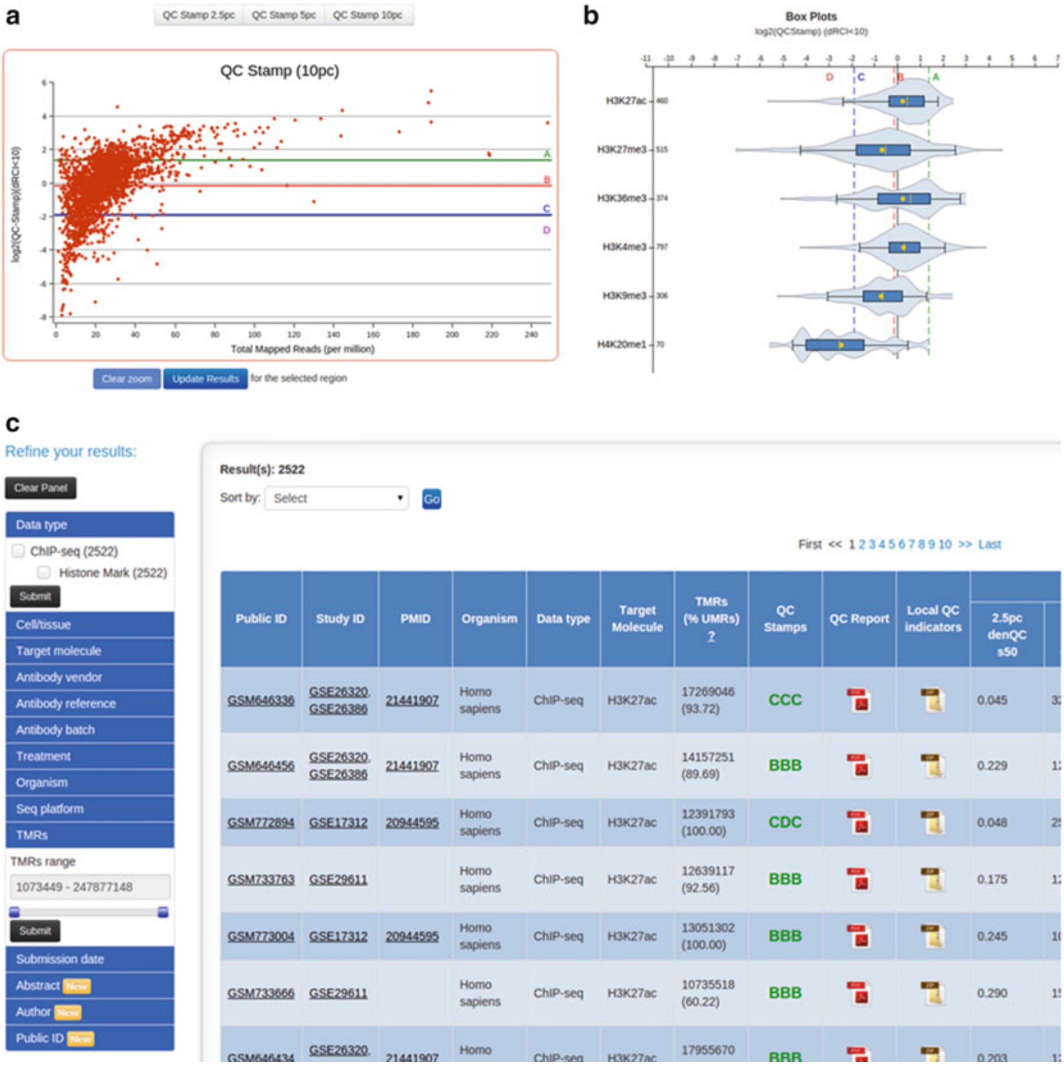
Public ID: Select Some Options

The NGS-QC database is currently hosting quality descriptors for 21526 publicly available datasets. Those entries are classified by their corresponding target molecules and their NGS-QC quality distributions are displayed below in a violin/boxplot. Vertical dashed lines defines the boundary for the QC-STAMP grades A, B, C and D corresponding to the inter-quartile intervals.

Fig. 8 Display illustrating the database page showing the search panel (*top*) and boxplots/violin plots table (*bottom*)

3. a *Scatter-plot* displaying the quality scores (QC-stamp) for each dataset in the context of their total mapped reads (TMRs).
4. *Results table* presenting an important number of information for each dataset retrieved.
5. *Refinement panel* providing further query options to be applied over the initial request to refine the results to specific interest.

All the described elements are meant to provide complementary information to the user, as well as means to narrow down the mining to users' specific interest. For instance, when multiple



□ ChIP-seq (2522)

□ Histone Mark (2522)

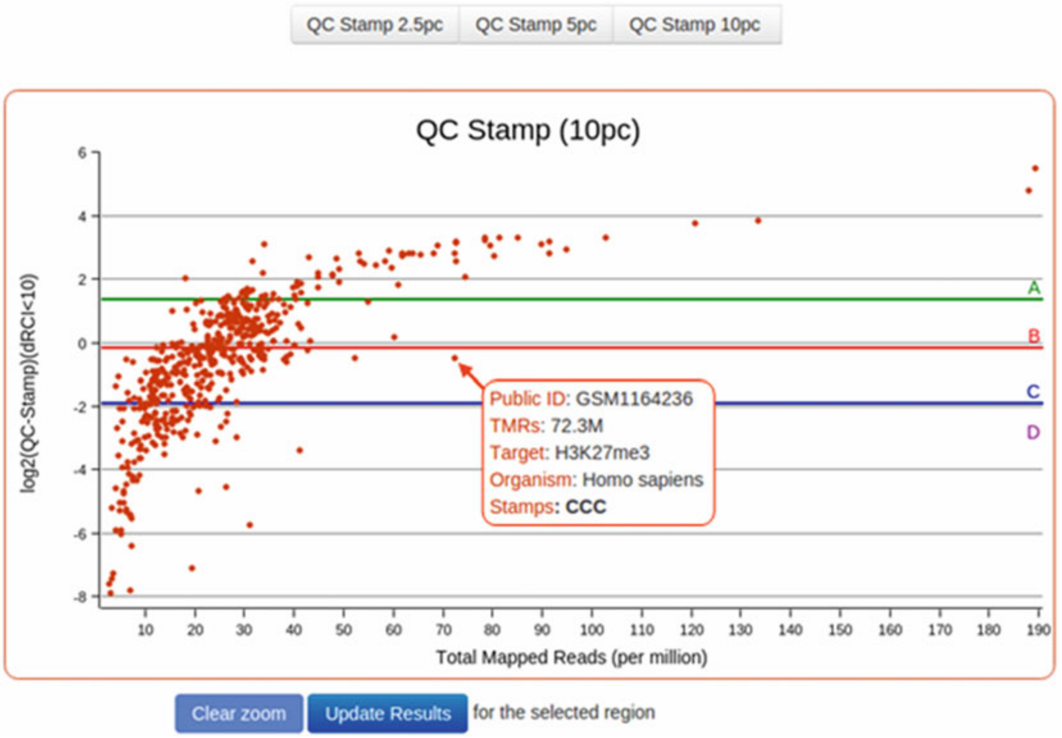
TMRs range: 1073449 - 247877148

Result(s): 2522

Public ID	Study ID	PMID	Organism	Data type	Target Molecule	TMRs (% UMRs) ?	QC Stamps	QC Report	Local QC indicators	2.5pc denQC s50
GSM646336	GSE26320, GSE26386	21441907	Homo sapiens	ChIP-seq	H3K27ac	17269046 (93.72)	CCC			0.045
GSM646456	GSE26320, GSE26386	21441907	Homo sapiens	ChIP-seq	H3K27ac	14157251 (89.69)	BBB			0.229
GSM772894	GSE17312	20944595	Homo sapiens	ChIP-seq	H3K27ac	12391793 (100.00)	CDC			0.048
GSM733763	GSE29611		Homo sapiens	ChIP-seq	H3K27ac	12639117 (92.56)	BBB			0.175
GSM773004	GSE17312	20944595	Homo sapiens	ChIP-seq	H3K27ac	13051302 (100.00)	BBB			0.245
GSM733666	GSE29611		Homo sapiens	ChIP-seq	H3K27ac	10735518 (60.22)	BBB			0.290
GSM646434	GSE26320,	21441907	Homo	ChIP-seq	H3K27ac	17955670	BBB			0.203

Fig. 9 Display illustrating the results obtained after performing a query in the NGS-QC database. (**a**) scatter-plot displaying the QC-indicators relative to the total mapped reads. (**b**) Boxplot/vioplots displaying the different target molecule retrieved on the query. (**c**) Results table including several additional information for each dataset (*right panel*) and the refinement panel (*left panel*)

targets are part of the initial query, the boxplot panel provides the possibility to select one of them (Fig. 9b). Furthermore, the *scatter-plot* display provides a powerful way to visualize the quality scores associated to each dataset in the context of their related TMRs. For instance, in the context of the query targeting the histone modification marks H3K27me3, the user can pass the pointer over the scatter-plot to have a pop-up displaying element like the dataset identifier, its associated TMRs, the target identity, the model organism and its computed quality grade (Fig. 10). Considering that the quality scores are computed at three different read count intensity dispersion conditions (dRCI), each of them are

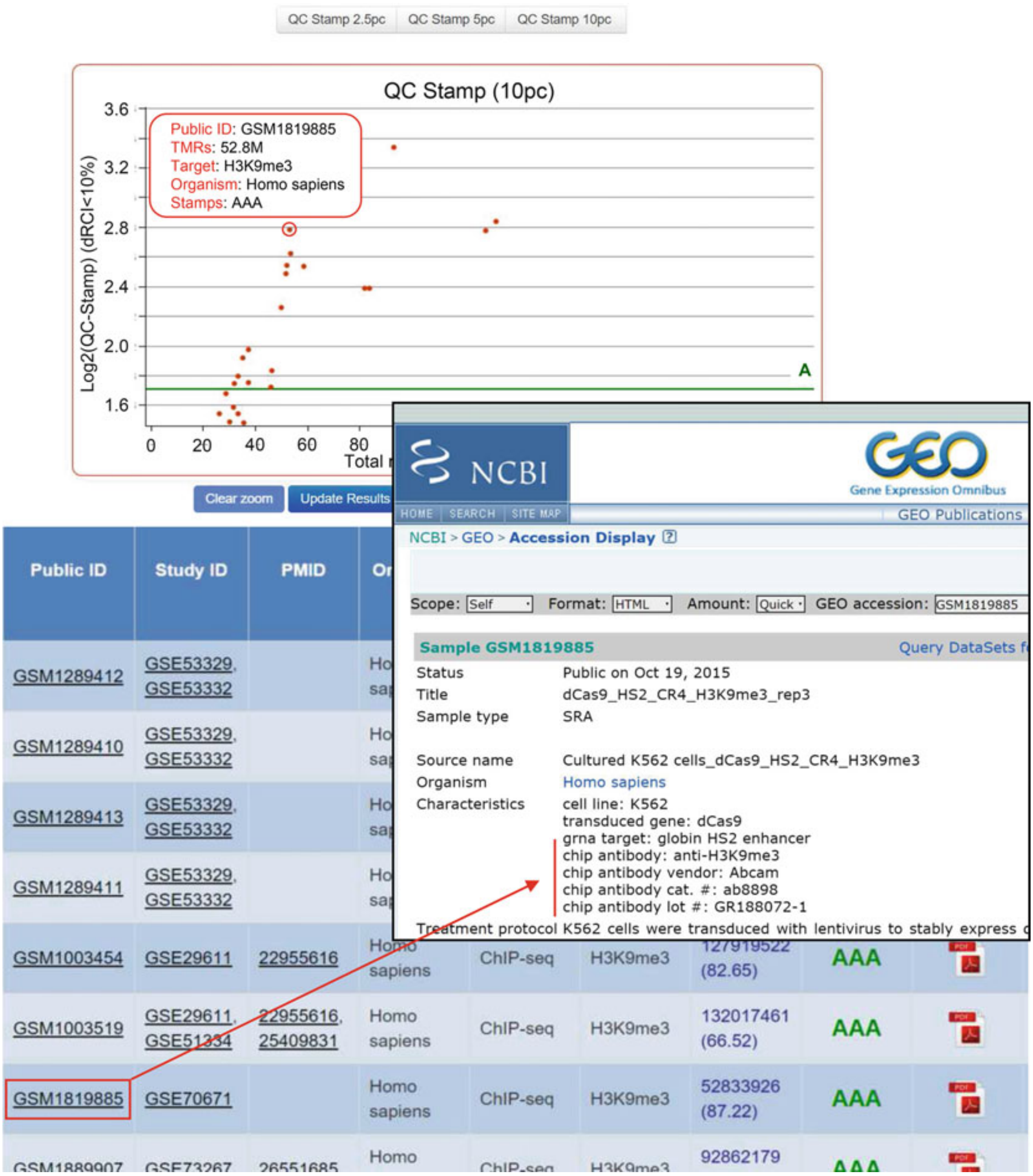


Public ID: GSM1164236
TMRs: 72.3M
Target: H3K27me3
Organism: Homo sapiens
Stamps: **CCC**

Fig. 10 Scatter-plot illustrating the quality scores computed for several H3K27me3 generated datasets in the context of their total mapped reads. The identity of one of these datasets presenting more than 70 million TMRs are highlighted as done in the NGS-QC Generator website

summarized in one of the letters corresponding to the QC-STAMP grade (e.g., A, A and A—represented in short AAA—for dRCI < 2.5 %, <5 % and <10 % respectively). In this context, users can switch the displays among all three corresponding scatter-plots using the buttons located on the top of the panel. Furthermore, users can zoom on each scatter-plot to perform for further refinement of the results for a desired group of datasets. To do so, users might select the "Zoom" icon located on the right top corner of the plot and then press and hold mouse button to select the region. Alternatively, zooming on the scatter-plot is possible by scrolling the wheel mouse button.

The *results table* (Fig. 9c), located under the scatter-plot provides further information for each of the datasets retrieved like public identifier (ID); study ID, associated publication ID (PMID), model organism, data type, target molecule, TMRs complemented by the fraction of unique mapped reads (i.e., by excluding the clonal reads), global quality grades (QC-Stamps) and a Global QC indicators report available in a pdf format. Further supplementary information is available under a ZIP file format available for downloading. Taken in consideration that users might systematically wish to associate a qualified dataset to an article describing it, users can access to the



Public ID	Study ID	PMID	Or					
GSM1289412	GSE53329, GSE53332		Ho sa					
GSM1289410	GSE53329, GSE53332		Ho sa					
GSM1289413	GSE53329, GSE53332		Ho sa					
GSM1289411	GSE53329, GSE53332		Ho sa					
GSM1003454	GSE29611	22955616	Homo sapiens	ChIP-seq	H3K9me3	127919522 (82.65)	AAA	PDF
GSM1003519	GSE29611, GSE51334	22955616, 25409831	Homo sapiens	ChIP-seq	H3K9me3	132017461 (66.52)	AAA	PDF
GSM1819885	GSE70671		Homo sapiens	ChIP-seq	H3K9me3	52833926 (87.22)	AAA	PDF
GSM1889907	GSE73267	26551685	Homo	ChIP-seq	H3K9me3	92862179	AAA	PDF

Scope: Self Format: HTML Amount: Quick GEO accession: GSM1819885

Sample GSM1819885

Status	Public on Oct 19, 2015
Title	dCas9_HS2_CR4_H3K9me3_rep3
Sample type	SRA
Source name	Cultured K562 cells_dCas9_HS2_CR4_H3K9me3
Organism	Homo sapiens
Characteristics	cell line: K562 transduced gene: dCas9 grna target: globin HS2 enhancer chip antibody: anti-H3K9me3 chip antibody vendor: Abcam chip antibody cat. #: ab8898 chip antibody lot #: GR188072-1
Treatment protocol	K562 cells were transduced with lentivirus to stably express

Fig. 11 Example illustrating the procedure by which the NGS-QC database can be used for retrieving the antibody in use for high quality grades datasets. A refined query has been performed for retrieving high quality datasets associated to the histone mark H3K9me3; followed by the identification of the best dataset ("AAA" QC grade with the least TMRs). Finally the corresponding GSM ID link has been used to explore over the original information retrieved in GEO

abstract of the related publication (when available) by placing the mouse pointer over the corresponding PMID.

Finally, some times the initial query might include too many outputs, such that users might wish to focus their view on a defined subset; thus a separate *refinement panel* is provided at the left of the results table (Fig. 9c). Importantly the refinement panel is composed by a list of 13 elements aiming at providing a large flexibility for this task such as Data type, target molecule, model organism, sequencing platform, TMRs, submission date, Abstract content, Author and Public ID. Other elements like the Cell line/tissue, the potential cell/tissue treatment or the antibody reference and batch available under a nonacademic access (*see* below).

3.4.2 The NGS-QC Database as a Source of Information for Providing Potential Hints to Interpret ChIP-Sequencing Assays with Low Quality Performances

An important application of the database is to provide potential hints to understand the reasons for the low quality grade associated to certain ChIP-seq or related datasets. To illustrate this aspect the user could query for the dataset "GSM733754" as described in the Subheading 3.4.1. In result, two entries associated to the histone modification mark "H3K27me3" are retrieved from the database, which correspond to two replicates retrieved under the same unique identifier (in general a unique ID is associated to a single dataset; but in some cases GEO entries include replicates under the same ID). Surprisingly, the computed quality grades are significantly different between these two entries, suggesting that though they are attributed as replicates, technical differences in preparation or sequencing could have led to these differences. In fact the dataset presenting "CBB" quality grade has 26 million TMRs, whereas "DDD" QC-stamp has only 9.2 million. This difference in the total mapped reads among datasets can at least partially explain their difference in their quality grades and strongly suggest that replicate datasets might also be sequenced at similar depths to avoid such potential sources of quality differences.

To further illustrate the use of the NGS-QC database, user could perform a query targeting all entries associated to the histone modification mark "H3K27me3" for the human model system. This time the scatter-plot is populated with more than 500 datasets in which a direct correlation between the quality grades and the TMRs per dataset can be observed. From this observed pattern, for instance it is possible to infer a minimal sequencing depth from which the majority of the retrieved datasets might present high quality grades (e.g., for TMRs higher than 30 million, most of the datasets are "CCC" or higher). It is important to mention that even for datasets generated with high TMRs levels, significant differences in quality scores can be observed. For instance, as illustrated in Fig. 10, datasets generated with more than 70 million reads could still present "CCC" quality grades.

3.4.3 Identifying ChIP-seq Grade Antibodies from the NGS-QC Database Content

While this in silico approach could provide guidelines concerning the sequencing-depth in use for a given factor, it is not guaranteed that all higher sequencing-depth experiment will be successful. In fact several other parameters like the antibody in use, the cell line/

tissue under analysis, as well as the performance of the experimenter are crucial factors in ChIP-seq pipeline that could have influence in the quality. To illustrate this important aspect, users might query for datasets generated with the histone modification mark H3K9me3 in the context of human studies (Fig. 11). Then, by taken advantage of the zoom option available in the results scatterplot we can refine the query to datasets presenting quality grades of "A". Finally, the refined output can be sorted based on TMRs using the results table such that high quality datasets (AAA) generated with the least number of TMRs might appear on the top of the result table panel. At this point, users might be interested on retrieving the Antibody source that gave rise to such optimal results, this is indeed possible by accessing the corresponding public repository page in GEO database by using the link embedded in the public ID displayed in the results table. Additionally, we are currently annotating the antibody source per QC-certified datasets in the NGS-QC database, such that users might directly access to this information through the previously described refinement panel.

4 Conclusions and Perspectives

As illustrated in the previous sections, the NGS-QC Generator and the QC database represent a powerful solution to evaluate the quality of any ChIP-seq and enrichment-related datasets. In april 2015, the qualities of more than 21,000 datasets available in the public domain have been assessed and the scientific community has full access to the corresponding quality reports (www.ngs-qc.org). This database is continuosly expanded. Indeed at the time of the final version of this document, the NGS-QC database bypassed the 30,000 processed datasets; covering more than 80 % of the publicly available ChIP-seq datasets in GEO (January 2016). We are currently developing additional modules, such as variable background detection and correction in cancer cells due to local DNA amplification or reduction, and expand its quality evaluation system to other types of datasets, like RNA-seq and all its related variants, as well as to chromatin conformation capture assays (e.g., Hi-C; ChIAPET).

While our methodology can currently detect poor quality datasets without necessarily providing an explanation for it, we are investing efforts in the evaluation of the antibodies in use. On one hand we have used text mining of all analyzed GEO entries to identify the cell line/tissue and antibody source and batch associated to each dataset (when available). We can thus provide a predictive analytics view of antibody performances based on several thousands of entries retrieved from the public domain. Considering that public datasets necessarily include variations originating from different experimenter skills and/or sample treatment conditions, we have recently set up an standardized automated pipeline for certifying antibodies dedicated to ChIP-seq applications. This last point is of great interest as there is currently no quantitative system apart from the NGS-QC Generator for certifying ChIP-seq quality grade antibodies. Indeed, the various commercial products labeled as "ChIP-seq grade" by their manufacturers receive their labels from a nonquantitative, subjectively chosen screenshot of a ChIP-seq profile, which has only very limited value to predict antibody performance along the entire genome.

We are convinced that a quantitative evaluation of the quality of enrichment-base datasets needs to be integrated in any dataset deposited in the public domain in the future. Similarly commercially available antibodies should have a quantified and independently certified quality performance associated to each antibody batch. Only this way the huge amount of information present in the public domain can be adequately extracted and waste of effort, time, and money can be saved when only high quality datasets are used for comparative or integrative analysis. Obviously, there is resistance towards the use of rigorous quality assessment tools by those performing these assays but at the end the science and the scientific community will benefit enormously from integrating such tools as a routine process as it has been done in the past for microarray analyses.

5 Notes

1. With coreutils (version>=8.6), –parallel option is available to speed up the sort command.
2. Preprocessing of raw data prior to alignment (i.e., trimming/filtering low quality or adapter-contaminated reads) is necessary if one is going to use end-to-end alignment approaches where there is a possibility of reads failing to align with long chunks of wrongly called bases (i.e., bad quality) or adapter contaminations at the end. Recent versions of alignment tools have "local alignment" approaches and low quality ends will be automatically clipped if not matching with the reference genome.
3. In experiments where enzymatic digestion is used for fragmenting DNA, the frequency of clonal reads might be enhanced due to a potential nuclease DNA-sequence specificity. In this scenario the removal of clonal reads is not advised.

Acknowledgements

This work was supported by funds from SATT/Conectus, the Fondation pour la Recherche Médicale (FRM), the Alliance Nationale pour les Sciences de la Vie et de la Santé–Institut Thématique Multi-organismes Cancer–Institut National du Cancer (INCa) grant "Epigenomics of breast cancer" and "EpiPCa," the Ligue National Contre le Cancer (to H.G.; Equipe Labellisée).

References

1. Mendoza-Parra MA, Van Gool W, Saleem MAM, Ceschin DG, Gronemeyer H (2013) A quality control system for profiles obtained by ChIP sequencing. Nucleic Acids Res 41, e196
2. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74. doi:10.1038/nature11247
3. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M et al (2013) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41: D991–D995. doi:10.1093/nar/gks1193
4. Kodama Y, Shumway M, Leinonen R (2012) The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res 40: D54–D56. doi:10.1093/nar/gkr854
5. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. doi:10.1093/bioinformatics/btp352
6. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. doi:10.1093/bioinformatics/btq033
7. Andrews S. FastQC: a quality control tool for high throughput sequence data [Internet]. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. citeulike-article-id:11583827
8. Patel RK, Jain M (2012) NGS QC toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7, e30619. doi:10.1371/journal.pone.0030619
9. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17(1). Next Gener Seq Data Anal. http://journal.embnet.org/index.php/embnetjournal/article/view/200/479
10. Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11:R86. doi:10.1186/gb-2010-11-8-r86
11. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P et al (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15:1451–1455. doi:10.1101/gr.4086505
12. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M et al (2001) Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol. doi:10.1002/0471142727.mb1910s89
13. Helt GA, Nicol JW, Erwin E, Blossom E, Blanchard SG, Chervitz SA et al (2009) Genoviz Software Development Kit: Java tool kit for building genomics visualization applications. BMC Bioinformatics 10:266. doi:10.1186/1471-2105-10-266
14. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM et al (2002) The Human Genome Browser at UCSC. Genome Res 12:996–1006. doi:10.1101/gr.229102