Integrating the ENCODE blocklist* for machine learning quality control of ChIP-seq with seqQscorer

Steffen Albrecht[1,7], Clarissa Krämer[2], Philipp Röchner[2], Johannes U Mayer[3,4,5], Franz Rothlauf[2], Miguel A Andrade-Navarro[1,5], Maximilian Sprang[3,5]

[1] Institute of Organismic and Molecular Evolution, Faculty of Biology, Johannes Gutenberg University Mainz, 55128 Mainz, Germany
[2] Information Systems and Business Administration, Johannes Gutenberg University Mainz, 55128 Mainz, Germany
[3] Department of Dermatology, University Medical Center of the Johannes Gutenberg University, 55131 Mainz, Germany
[4] Research Center for Immunotherapy (FZI), University Medical Center of the Johannes Gutenberg University Mainz, 55131 Mainz, Germany
[5] Institute of Quantitative & Computational Biosciences, Johannes Gutenberg University Mainz, 55128 Mainz, Germany.
[7] Current Address: Department of General Practice and Primary Health Care, Faculty of Medical and Health Sciences (FMHS), The University of Auckland, Auckland 1023, New Zealand

Abstract

Motivation: Quality assessment of next-generation sequencing (NGS) data is a complex but important task to ensure correct conclusions from experiments in molecular biology, biomedicine, and biotechnology. While there are multiple software tools to support this process, it is still mainly driven by domain experts manually curating quality reports. This time-consuming effort is expected to increase with the future capacity to generate more sequencing samples, at lower costs, using advanced technologies. By leveraging quality-labelled data from large databases and automatically created quality-related features derived from raw sequencing samples, supervised machine learning shows great potential for supporting quality assessment, as shown by the seqQscorer tool. This tool shows outstanding results for automated quality assessment and has been used to investigate several quality-related aspects, integrating large-scale sequencing datasets. To improve seqQscorer in terms of accuracy and processing time, we explored the potential of creating features that are more informative and faster in generation than those conventionally used by seqQscorer's models. For this purpose, we used the ENCODE blocklist*, a set of problematic genomic regions known to be related to the quality of NGS data.

Results: Our results show that we can improve the quality assessment for ChIP-seq samples derived from human tissues and cell lines when using features based on the ENCODE blocklist by up to 4.4%, when using a generic dataset containing single-end and paired-end sequencing samples. For other assays, these new blocklist features allow for highly accurate quality assessment, in general, but are less accurate in comparison to using the conventional seqQscorer features.

Outlook: These results strongly encourage an extension of the seqQscorer tool that automatically integrates the new blocklist features. Users investigating human protein-DNA interactions would benefit from this extension, as it not only improves model performance during quality assessment but also enhances usability by simplifying the installation procedure and reducing the computational resources required for feature generation.

---

* Following inclusive language recommendations, we use the term ENCODE blocklist, previously introduced as the ENCODE blacklist (1)

Introduction

In this article, we propose an extension of seqQscorer, a machine learning-driven software for automated quality assessment of next-generation sequencing (NGS) samples. Next-generation sequencing, also called high-throughput or massive parallel sequencing, has revolutionized the past decades of molecular biology, biotechnology, and biomedical research. Besides whole-genome sequencing, NGS has been coupled with molecular assays to study important drivers of gene regulation and better understand cellular function and cell differentiation. Widely used assays are RNA-seq (RNA sequencing) to quantify gene expression and DNase-seq (DNase I hypersensitive sites sequencing) to detect open chromatin sites (2,3). Open chromatin is a prerequisite for transcription factor binding and is strongly associated with histone modification. As these protein-DNA interactions strongly influence gene expression, the ChIP-seq (chromatin immunoprecipitation followed by high-throughput sequencing) assay has been developed to reveal such interactions (4). These assays are essential for both fundamental research and biomedical studies investigating cancer and other life-threatening diseases (5–8).

Technological advancements for NGS allowed scientists to create an enormous amount of data, filling publicly available databases (9–11). These advancements also reduce costs, resulting in future research projects with more samples (12). A key challenge, considering large datasets of NGS samples, is quality assessment. Low-quality samples must be accurately identified to be refined or even replaced to ensure reliable downstream analyses. This is even more important when sequencing results are involved in decision-making processes for precision medicine (13).

While quality assessment is an important task, it is also challenging and requires the experience of domain experts. Several tools exist to support this process, providing quality reports incorporating different statistics derived from NGS samples. Some of these are generic to different assays (14,15), while others are tailored to the specificity of the different molecular assays (16,17). These reports can be comprehensive, providing an overview of the quality of NGS samples from multiple perspectives (14,18). However, the inspection of all this information can be time-consuming and requires domain expertise (19).

To reduce manual work, we previously introduced *seqQscorer*, which applies machine learning (ML) models trained on quality-related reports and statistics to automate the quality assessment of NGS samples from RNA-seq, ChIP-seq, DNase-seq, and ATAC-seq (another open chromatin assay). The models are trained by supervised ML algorithms and return a single quality score that is used to rank potentially large sets of NGS samples by their quality. The quality labels used for model training are defined by the manually curated ENCODE status, considering *revoked* fastq files as low-quality and *released* fastq files as high-quality raw NGS samples.

In the original and following work, we showed that seqQscorer captures quality across assays and species (19,20). Additionally, its quality score relates to batch effects (21), and it enabled us to introduce *Quality Imbalances*, a new type of technical bias that arises in experiments dealing with multiple samples (e.g. controls and conditions, with several replicates) when the mean quality scores of the groups of interest, for example control and treatment strongly differ (22). While seqQscorer achieves outstanding results, using commonly used quality-related features derived

from raw sequencing samples, we continued to further explore possibilities to improve seqQscorer in terms of accuracy, usability, and processing time.

This led us to investigate the ENCODE blocklist, a set of problematic genomic regions defined on data from the ENCODE portal (1). These regions can, for instance, contain collapsed repeats in the genome assembly that lead to a high number of mapped reads. It has been shown that removing these regions before downstream analysis is important to achieve reliable results, and bioinformatic analysis tools incorporate automated filtering of these regions, thereby removing artifacts before downstream analysis (23). According to the original work, the blocklisted regions can also be used to compute a single-value quality metric describing the fraction of reads that overlap with any of these regions.

Given the strong relationship between these blocklisted regions and data quality, we explore the potential of creating ML features based on these regions to train quality assessment models as introduced by seqQscorer. In comparison to a single value for the fraction of reads overlapping with any blocklist region, the *blocklist features* we propose are described by multiple values describing the number of overlapping reads for each single blocklisted region. We show that the blocklist features allow seqQscorer to perform very well on data across assays and organisms and to even improve model performance for ChIP-seq data derived from human tissues and cell lines.

Datasets and Methods

The datasets used for investigating the blocklist features are the same as those used in the seqQscorer study, composed of samples derived from tissues and cell lines, from now on called *biosamples*, of two organisms (human, mouse), using three different assays (ChIP-seq, DNase-seq, RNA-seq), further specified by the two sequencing modes (single-end, paired-end). Importantly, the samples have been labelled using the ENCODE status *revoked* as the low-quality label. For the assays used, all revoked files have been downloaded. The same number of files with the *released* status have been downloaded, representing the high-quality data. Different bioinformatics tools have been applied to derive comprehensive quality-related information from the raw sequencing samples to assemble the conventional feature sets, see seqQscorer (19) for more details.

To create the blocklist features explored in this study, we first mapped the sequencing reads for each raw sequencing sample, provided in fastq format. As done for seqQscorer, we use *Bowtie2* with default settings consistently across samples to align sequencing reads against the genome assemblies hg38 and mm10 for human and mouse, respectively. Then, for each blocklisted region, the number of overlapping reads was counted to receive a feature value for each region. According to the size of the blocklist for the different genomes, this results in 636 features for human and 3435 for mouse samples.

The ML experiments performed for this article aim to compare the quality classification performance of machine learning models trained with different sets of features using (i) the conventional seqQscorer features (seqQ), (ii) the new blocklist features (BL) as described in the previous paragraph, and a concatenation of both (seqQ+BL). Models were evaluated by the area under the ROC curve (auROC) within a ten-fold cross-validation. To obtain a statistically robust average performance measure, we repeated the cross-validation five times to receive 50 training-testing splits and visualize the average auROC with the 95% confidence intervals.

As the number of BL features is much higher compared to the seqQ features, we face the small-data learning challenge occurring for datasets with a high number of ML features but a comparatively low number of samples (24). Therefore, we integrate dimensionality reduction using the Principal Component Analysis (PCA) (25). Fitted on the training set, the PCA is used to transform training and testing sets to reduce the number of features by keeping only the principal components explaining most of the data variance. The number of relevant components is automatically determined in such a way that 99.99% of the variance is explained. As a result, the number of relevant components may vary between different training-testing splits. Note that this has not been done for seqQscorer before and could potentially improve the performance of seqQscorer using its conventional features.

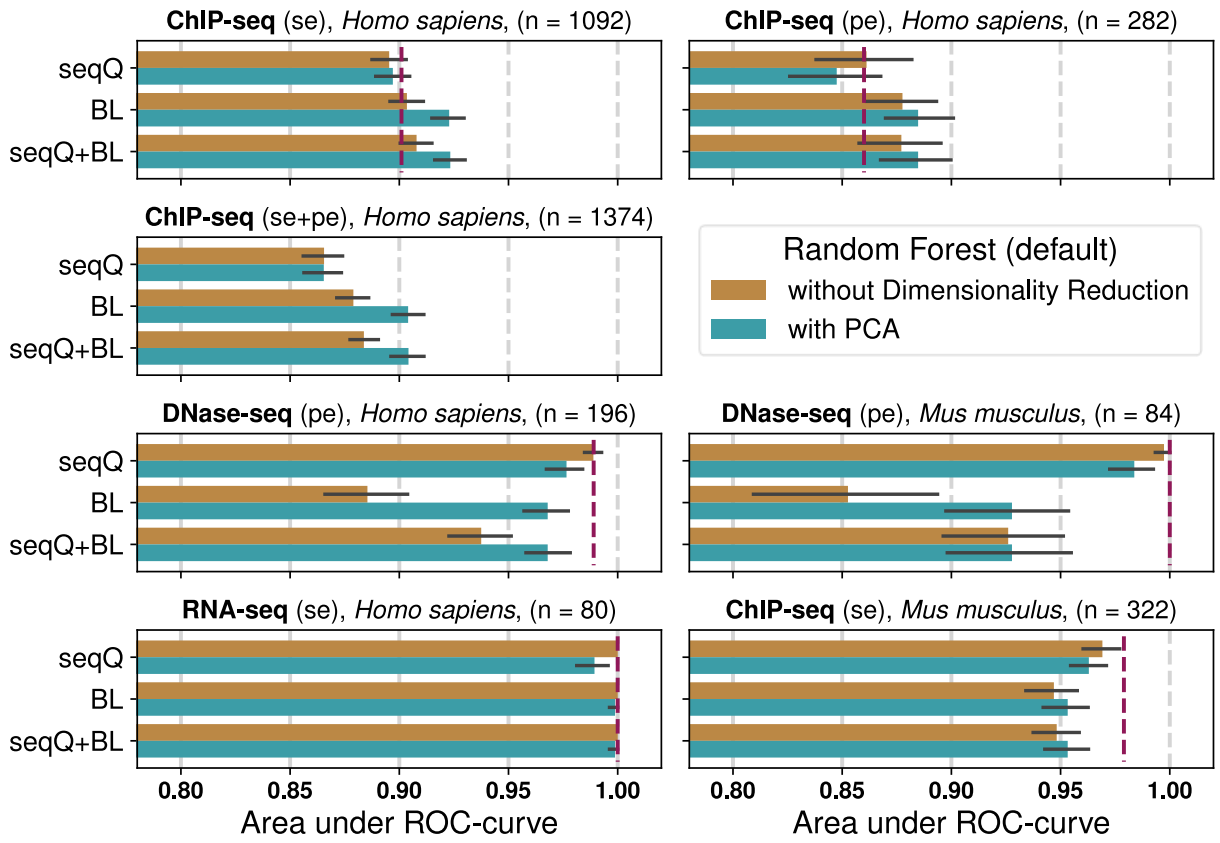

Figure 1 – Performance of models using different machine learning features

Here, we explore the performance of models trained by the Random Forest algorithm with default settings (see scikit-learn package (26)). The key question is how the model performance changes when using the different feature sets composed of the conventional seqQscorer features (seqQ), backlist features (BL), and concatenation of these (seqQ+BL). The bars represent the mean area under the ROC curve evaluated on 50 training-testing splits. The error bars show the 95% confidence intervals of the performance measures obtained. Subtitles summarize the characteristics of the underlying datasets with assay name, organism, and run type, with (se) and (pe) abbreviating single-end and paired-end, respectively. The number "n" refers to the number of samples in the subsets. The dashed red line shows the mean performance of models that were

fine-tuned on the specific subsets during a more comprehensive ML analysis performed for the original seqQscorer study, solely based on the seqQ features. Note that these results were derived from only one ten-fold cross-validation, without repetition. Additionally, the (se+pe) ChIP-seq setting has not been explored in the original seqQscorer study. Note that a few samples had been removed from the ENCODE portal in the meantime, resulting in slightly lower sample sizes for some subsets compared to the subsets originally used.

Results

The main question to be answered by the ML experiments was how the models perform in classifying NGS samples by quality using the BL features in comparison to the seqQ features conventionally used by seqQscorer. Using the BL features can result in improved model performance, but this is dependent on the assay and model organism (**Fig. 1**). For ChIP-seq data derived from human biosamples, we observed an improvement of 2.7% to 4.4% when using the BL features, regardless of the run type. Using the BL features on DNase-seq data resulted in a performance drop for data from both organisms, with a greater difference observed for data from mouse biosamples. For RNA-seq data, for which the performance was very high in general, no difference could be observed between using seqQ and BL features. For ChIP-seq data from mouse biosamples, the performance dropped when using the BL features compared to the conventional features of seqQscorer. In most cases, applying the PCA dimensionality reduction led to higher performance when using the BL features. Importantly, incorporating PCA is key for achieving performance boosts of human ChIP-seq subsets. Depending on the data subsets and feature sets used, the number of principal components used differs (**see Table 1 in the Appendix**). In general, using a comprehensive feature set composed of seqQ and BL did not enable seqQscorer to further improve the quality assessment.

Discussion and Outlook

Our results clearly advise users of seqQscorer to make use of the ENCODE blocklist features when using ChIP-seq in experiments investigating human biosamples. Regarding model robustness, the corresponding ChIP-seq subset, especially for single-end sequencing, is the biggest available for seqQscorer.

Small sample size might be problematic for the ChIP-seq subset from mouse biosamples, for which the models trained with the blocklist features perform worse in comparison to the conventional models. The high dimensionality with 3,435 blocklist features and the low sample size of this subset define a small data learning challenge known to result in decreased model performance in general (24,27).

Even though the integration of the ENCODE blocklist can result in lower model performance, the auROC is high in these cases. Hence, depending on the specific assay and organism, users might decide to accept slightly lower model accuracy using blocklist features that are easier to generate in comparison to the conventional seqQscorer features. In contrast to the conventional features, a mapping against the full genome assembly is not required for the blocklist features. With the read mapping being computationally the most expensive feature-generating step, the runtime for this procedure can be reduced significantly when using only the blocklisted genomic regions.

Additionally, the blocklist feature generation does not require any additional bioinformatic packages, simplifying the software setup and installation.

The ML models trained and evaluated in this study are simple in comparison to those explored in the original seqQscorer study, in which different algorithms were used with hyperparameter tuning. Interestingly, considering the conventional features, the default Random Forest models perform almost as accurately as the fine-tuned seqQscorer models for most of the subsets investigated. Nevertheless, other, possibly fine-tuned, ML algorithms and alternative dimensionality reduction methods might enable the improvement of the model performance using the new blocklist features.

ENCODE uses the blocklist to create a single-value quality control metric described by *the fraction of reads in blocklisted regions* (1). Hence, this fraction has been considered when samples got revoked in the past, and a relationship could potentially exist between the ENCODE status we use to label our data and the blocklist features we propose here as a novel feature set for a seqQscorer extension. However, we assume that this fraction, as a single-value quality control metric, does not capture the level of detail encoded in the blocklist features we propose, describing one value for each region. Furthermore, it is unlikely that this fraction is always the sole reason for sample revocation, and the proposed blocklist features are expected to relate to other quality-related problems. Indeed, the fraction of reads in the ENCODE blocklist is poorly related to the quality labels, while the seqQscorer quality score allows for a much better differentiation between low- and high-quality samples (**see Fig. A1 in the Appendix**). Nevertheless, we are aiming to integrate more data from the CistromeDB to perform an external evaluation of the proposed blocklist-based approach before we release this extension.

The data used by seqQscorer comes from bulk sequencing only, and the ML models are not necessarily applicable to samples from single-cell sequencing, which is a limitation considering the importance of sequencing assays at single-cell resolution (28). However, data from single-cell sequencing has increasingly been uploaded to the ENCODE database during the last five years. Given that this data also underwent the comprehensive quality annotation procedures and manual curation protocols of the ENCODE project, these data provide a great foundation to explore, in future work, how the seqQscorer approach works for automated quality assessment of single-cell sequencing data.

Appendix

		dim	# PCs min	# PCs median	# PCs max
ChIP-seq (se) human, n = 1092	seqQ	34	17	17	17
	BL	637	73	74	78
	seqQ+BL	671	73	74	78
ChIP-seq (pe) human, n = 282	seqQ	34	20	21	21
	BL	637	54	56	58
	seqQ+BL	671	54	56	58
ChIP-seq (se+pe) human, n = 1374	seqQ	34	18	18	18
	BL	637	77	79	81
	seqQ+BL	671	77	79	81
DNase-seq (pe) human, n = 196	seqQ	34	14	15	15
	BL	637	22	23	25
	seqQ+BL	671	22	23	25
DNase-seq (pe) mouse, n = 84	seqQ	34	12	13	14
	BL	3436	9	11	12
	seqQ+BL	3470	9	11	12
RNA-seq (se) human, n = 80	seqQ	34	22	23	23
	BL	637	3	3	3
	seqQ+BL	3470	3	3	3
ChIP-seq (se) mouse, n = 322	seqQ	34	18	18	18
	BL	3436	12	12	14
	seqQ+BL	3470	12	12	14

Table 1 – Data dimensionality before and after dimensionality reduction using the PCA

The PCA is fitted based on the training set for each training-testing split. There are 50 splits, and as the PCA is fitted on the training data, the number of principal components (PCs) varies between the training-testing splits. This table provides an overview of the dimensionality of the data without reduction (dim) and information about the number of PCs (# PCs) used by the classifiers after dimensionality reduction. Across the 50 splits, we here report the minimum (min), maximum (max), and median number of PCs required to explain 99.99% of the variance of the training data.

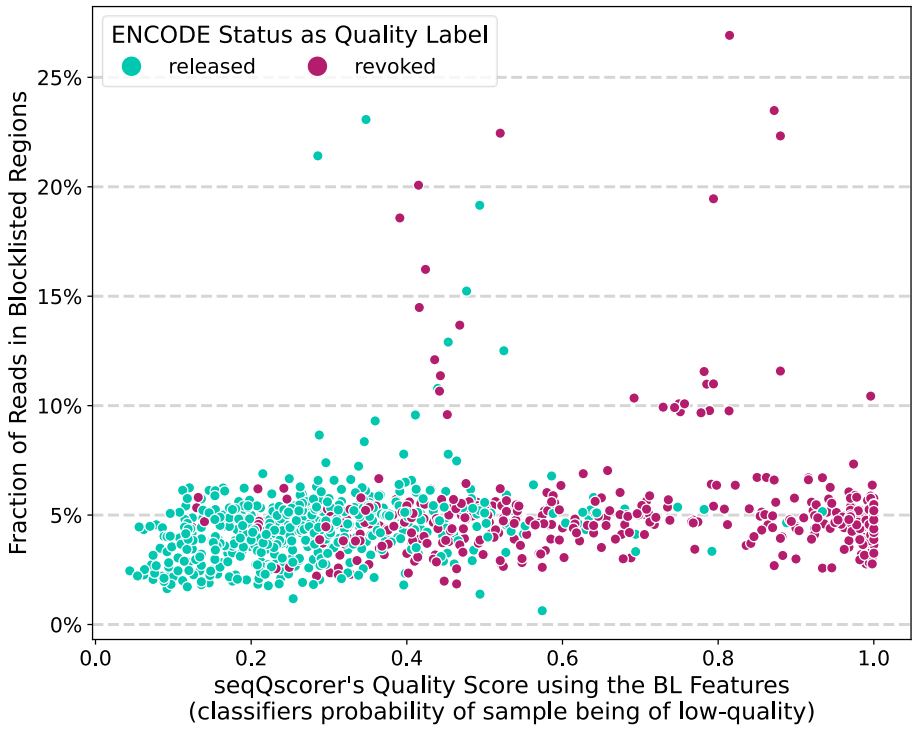

Figure A1 – Comparing a single-value blocklist quality metric with the quality score from seqQscorer in relation to the ENCODE quality labels

The dots in this scatter plot represent the 1092 samples from the single-end ChIP-seq subset for human biosamples. The y-axis shows the fraction of reads overlapping with regions from the ENCODE blocklist for these samples. The x-axis shows the mean quality score for these samples using the seqQscorer models trained on the blocklist features. The higher this score, the lower the expected quality, as these scores are essentially the probability of a sample being of low quality. The color of the dots describes the quality label derived from the ENCODE status, using revoked samples as low-quality and released samples as high-quality data. Low-quality samples show a clear trend to be associated with higher quality scores from our tool, as expected from the high model performance (see ***Fig. 1****, upper left panel). However, the overall fraction of reads in blocklisted regions does not show such a strong relationship, which strongly indicates that the seqQscorer models are able to extract more information from the blocklist by incorporating a feature for each region instead of considering the blocklist as a whole.*

References

1. Amemiya HM, Kundaje A, Boyle AP. The ENCODE blacklist: identification of problematic regions of the genome. Scientific reports. 2019;9(1):9354.

2. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nature reviews genetics. 2011;12(2):87–98.

3. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harbor Protocols. 2010;2010(2):pdb-prot5384.

4. Mardis ER. ChIP-seq: welcome to the new frontier. Nature methods. 2007;4(8):613–4.

5. Wang C, Han B. Twenty years of rice genomics research: From sequencing and functional genomics to quantitative genomics. Molecular Plant. 2022;15(4):593–619.

6. Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, et al. Next-generation sequencing technology: current trends and advancements. Biology. 2023;12(7):997.

7. Prasher D, Greenway SC, Singh RB. The impact of epigenetics on cardiovascular disease. Biochem Cell Biol. 2020 Feb;98(1):12–22.

8. Esteller M. Epigenetics in Cancer. N Engl J Med. 2008 Mar 13;358(11):1148–59.

9. de Souza N. The ENCODE project. Nature methods. 2012;9(11):1046–1046.

10. Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, et al. Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic acids research. 2019;47(D1):D729–35.

11. Clough E, Barrett T. The gene expression omnibus database. In: Statistical genomics. Springer; 2016. p. 93–110.

12. Kumar KR, Cowley MJ, Davis RL. Next-Generation Sequencing and Emerging Technologies*. Semin Thromb Hemost. 2024 Oct;50(07):1026–38.

13. Mosele MF, Westphalen CB, Stenzinger A, Barlesi F, Bayle A, Bièche I, et al. Recommendations for the use of next-generation sequencing (NGS) for patients with advanced cancer in 2024: a report from the ESMO Precision Medicine Working Group. Annals of Oncology. 2024;35(7):588–606.

14. Brown J, Pirrung M, McCue LA. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics. 2017;33(19):3137–9.

15. Meyer CA, Liu XS. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nature Reviews Genetics. 2014;15(11):709–21.

16. D. Chikina M, G. Troyanskaya O. An effective statistical evaluation of ChIPseq dataset similarity. Bioinformatics. 2012;28(5):607–13.

17. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105–11.

18. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047–8.

19. Albrecht S, Sprang M, Andrade-Navarro MA, Fontaine JF. seqQscorer: automated quality control of next-generation sequencing data using machine learning. Genome biology. 2021;22(1):1–20.

20. Sprang M, Krüger M, Andrade-Navarro MA, Fontaine JF. Statistical guidelines for quality control of next-generation sequencing techniques. Life Science Alliance [Internet]. 2021 [cited 2025 May 10];4(11). Available from: https://www.life-science-alliance.org/content/4/11/e202101113.abstract

21. Sprang M, Andrade-Navarro MA, Fontaine JF. Batch effect detection and correction in RNA-seq data using machine-learning-based automated assessment of quality. BMC bioinformatics. 2022;23(6):1–15.

22. Sprang M, Möllmann J, Andrade-Navarro MA, Fontaine JF. Overlooked poor-quality patient samples in sequencing data impair reproducibility of published clinically relevant datasets. Genome Biol. 2024 Aug 16;25(1):222.

23. Stark R, Brown G. DiffBind: differential binding analysis of ChIP-Seq peak data. R package version. 2011;100(4.3).

24. Vecchi E, Pospíšil L, Albrecht S, O'Kane TJ, Horenko I. eSPA+: Scalable entropy-optimal machine learning classification for small data problems. Neural Computation. 2022;34(5):1220–55.

25. Kherif F, Latypova A. Principal component analysis. In: Machine learning [Internet]. Elsevier; 2020 [cited 2025 May 10]. p. 209–25. Available from: https://www.sciencedirect.com/science/article/pii/B9780128157398000122

26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825–30.

27. Vecchi E, Berra G, Albrecht S, Gagliardini P, Horenko I. Entropic approximate learning for financial decision-making in the small data regime. Research in International Business and Finance. 2023;65:101958.

28. Jovic D, Liang X, Zeng H, Lin L, Xu F, Luo Y. Single-cell RNA sequencing technologies and applications: A brief overview. Clinical & Translational Med. 2022 Mar;12(3):e694.