Machine Learning in Quality Assessment of Early Stage Next-Generation Sequencing Data

Steffen Albrecht[1,2,*], Miguel A. Andrade-Navarro[1], Jean-Fred Fontaine[1,*]

1 Johannes Gutenberg University Mainz, Faculty of Biology, Institute of Organismic and Molecular Evolution (iOME)
2 Institute of Molecular Biology (IMB Mainz)

*** {s.albrecht, fontaine}@uni-mainz.de**

Abstract

Next Generation Sequencing is a powerful technology highly relevant in biomedical research and pharmaceutical industry. Applied in combination with molecular assays it provides detailed insights in genomic properties such as chromatin accessibility or protein-DNA interactions. However, the biological relevance of results from these assays is extremely sensitive to the quality of the sequencing data. So far, quality control tools require extensive computational resources and manual inspection. This is critical considering the increasing amount of sequencing data due to decreasing costs. In this study, we investigated the possibility to automatically classify the quality of a large set of raw sequencing data in fastq format by using state-of-the-art machine learning algorithms and a comprehensive grid search to tune the parameters. The results showed high classification accuracy in discriminating between low and high quality files. Gradient Boosting machines were performing the best in most of the tested scenarios. Furthermore, some species-specific models were generalizable to classify data files from other species. The results suggest that the studied algorithms have high potential to be used routinely in biomedical and clinical applications to classify sequencing data according to quality. We provide a script that allows the application of the pre-trained classification models to new data: https://github.com/salbrec/seqQscorer.

Introduction

Next generation sequencing (NGS) assays are massively used in biomedical research and pharmaceutical industry. Among other applications, sequencing assays can be used to derive gene expression data (RNA sequencing assays; RNA-seq), protein-DNA interaction sites (chromatin immunoprecipitation DNA-sequencing assays; ChIP-seq), and open chromatin regions (DNase I hypersensitive sites sequencing assays; DNase-seq). These applications generate a large amount of data that require substantial computational resources for storage and processing and human resources for manual validation. Thus, the earlier the data quality can be determined, the earlier it is possible to exclude low quality samples to save resources that would have been unnecessarily spent otherwise. Importantly, using those low quality samples for downstream analyses, can lead to a crucial loss of reliability regarding biological results and findings. Taking into account that sequencing costs are continuously decreasing, it is expected that the amount of sequencing data will drastically increase in the future.

Consequently, quality control (QC) tools requiring a minimum amount of manual validation and producing highly reliable results will be of critical importance.

From the sequencing experiment the so-called fastq files are produced, which contain sequencing reads and can be seen as early stage sequencing data or raw data. While generic QC tools are applicable directly on the fastq files, other assay specific tools reveal their quality reports after certain additional processing steps. For example, there are several studies investigating comprehensive bioinformatic pipelines to assess the quality of ChIP-seq data [5, 12, 14] or RNA-seq data [11, 20]. Other QC tools are limited to specific sequencing platforms [16, 17]. Generic QC tools, for example FastQC, are applicable to data from any type of platform or assay and provide reports that comprehensively describe the quality of NGS data [6, 19]. AfterQC [3] and FaQCs [10] are tools that perform preprocessing such as read filtering and trimming as well as QC reporting. Because these quality reports are often extensive and complex, other tools such as MultiQC [7] or FQC Dashboard [1] aim to create one combined overview integrating reports from different tools and sources. Yet, manual inspection of such combined reports still requires substantial time and expertise.

To address this problem our study investigates a novel approach that uses existing public sequencing data in combination with state-of-the-art machine learning algorithms to predict the quality of fastq files automatically based on established QC feature sets. The major goal of this novel approach is to improve quality control decisions made for sequencing samples. This will ease the separation of low-quality files from those of higher quality and will have two major impacts, especially on clinical studies involving sequencing data from many patients. First, downstream analysis based on high-quality patient data will lead to more reliable results that are needed for instance in drug development. Second, in personalized medicine it is important to identify low-quality samples to avoid the misclassification of patients as this can lead to wrong decisions with respect to the treatment.

Materials and Methods

The underlying data for this study comes from the ENCODE project data portal, which provides fastq files together with the outcome of the project QC procedure determining each file status as released or revoked. We use this status to define the quality of a file as either low or high if its status is revoked or released, respectively. The explanations for the revocation of files in ENCODE are often described by so called audits. In ENCODE they describe files and whole experiment datasets with respect to QC features or (missing) annotations. However, it is not always possible to explain the status by these audits, since they are mainly considered as supporting information for specialists to make a decision about the status. Thus, human inspection and specialists expertise is needed to assess these difficult cases to guarantee a reliable annotation for this status. This reflects the work and knowledge behind the status that we use as a good quality indication. We were able to identify and download 1515 low-quality files plus the same amount of high-quality files in order to define balanced training and testing sets. For the selection of high-quality files we prioritized files that are associated to an experiment that contains also revoked files. The remaining high-quality files were chosen randomly. There were 2618 files for Homo Sapiens (human), 538 files for Mus musculus (mouse), and 54 files for Caenorhabditis elegans (worm). The most frequent sequencing assays were ChIP-seq with 2046 files, Dnase-seq with 640 files, RNA-seq with 240 files, enhanced cross-linking immunoprecipitation (eCLIP) with 68 files, and RNA immunoprecipitation sequencing (RIP-seq) with 36 files. The fastq files occurred either as a pair derived from paired-end sequencing or as a single file from single-end sequencing methods. The vast majority (88%) of revoked files from DNase-seq involved

paired-end files, while a large majority (72%) of revoked ChIP-seq files involved single-end files. For the study of different species-assay combinations with reasonable amount of data, we analyzed single-end files from ChIP-seq and paired-end files from Dnase-seq for human and mouse. Single-end files from RNA-seq were investigated only for human due to a very low number of files derived from mouse for this assay.

We derived four different quality feature sets for the set of downloaded fastq files. The first feature set *FQC* was defined by eleven features from the summary statistics of the FastQC tool (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). In the summary of a FastQC report each statistics is flagged as *Fail*, *Warning*, or *Pass*. We use these flags as values for the features. The second feature set *Bow* contained the mapping statistics after applying Bowtie2 [8] to map the sequencing reads of human and mouse against the genome assembly hg38 and mm10, respectively. The mapping statistics describe the percentage of reads being *unmapped*, *uniquely mapped*, or *multiple mapped* and their *overall mapping* rate. Accordingly, there are four features for single-end and eight features for paired-end as these statistics are done for both the concordantly and discordantly mapped reads. The third feature set was composed of nine features describing the distribution of reads mapped within the following types of genomic regions of interest: *promoter, first intron, other introns, 5'UTR, first exon, other exons, 3'UTR, distal intergenic*, and *downstream proximal to the 3'UTR*. We call this feature set *RA* (for read annotations) and the features were derived using the Bioconductor package ChIPseeker [21]. The fourth feature set is called *TSS*, it describes the distribution of reads around TSS (transcription start site) positions on the genome. The Bioconductor package ChIPpeakAnno [22, 23] was used to compute the amount of reads within the region 5kb up- and downstream the TSS divided into ten bins, resulting in ten features for this set. To reduce memory requirements during computation, the read annotations and *TSS* distribution were derived on one million mapped reads randomly sampled. For paired-end files, *FQC* features were derived independently for each of the two files, while *Bow*, *RA*, and *TSS* features were derived for the pair of files itself. In order to reduce redundancy in the dataset, we filtered out the *FQC* set for one member of each pair randomly.

We conducted classification experiments on either all samples of the dataset containing samples from different species and assays or on homogeneous subsets of the data containing samples for a certain combination out of species, assay, and run type. More details about the sizes of these subsets are provided together with the classification results. Given an underlying dataset for the classification experiment, different feature sets were provided to the classification algorithm composed out of features from only one feature set or the composition of more feature sets.

Based on the different species-assays and feature set combinations, machine learning algorithms were applied to train models classifying the ENCODE file status, that is the low or high quality annotation of an ENCODE data file. As the underlying subsets vary in terms of sample size and number of features, a comprehensive grid search was applied to be able to find the best fitting algorithm and parameter setting for each subset. The algorithm set for the grid search includes these ten algorithms: multilayer perceptron (MLP), random forest classifier (RFC), gradient boosting classifier (GBC), XG-Boost (XGB), support vector classifier (SVC), Gaussian Naïve Bayes (GNB), logistic regression (LRN), K-nearest neighbor classifier (KNN), extra-trees classifier (ETC), and Adaboost in combination with a decision tree learner (ADT). In combination with parameter settings specific to each algorithm a total of 19417 different models were trained and evaluated within the grid search for each classification case. Furthermore, for each parameter setting we applied a feature selection method prior to the classification. The feature selection is based on chi-squared statistics and selects the top k features. We analyzed four values for k: 100% (no feature selection), 75%, 50%, or 25% of all features

in the given set of quality features. The prior feature selection was added, because some classifiers are known to be sensitive to features that are intercorrelated.

The predictive performance was evaluated by the area under receiver operating characteristic curve (auROC). Within the grid search and feature selection, a ten-fold cross-validation was applied to evaluate the predictive performance. The entire grid search including the preprocessing methods as well as the chi-squared statistics and nine of the classification algorithms was implemented using the Python package scikit-learn [2, 18]. XGBoost was added from the project of Chen et al. [4].

Results

Potential of machine learning in quality assessment

We investigated the potential of machine learning algorithms to classify sequencing data according to quality by applying a grid search to the whole mixed dataset and five homogeneous subsets defined by different species-assay combinations. The features used by the classification models were the features from the different feature sets. In comparison we investigated the potential of single features to predict the quality. In order to evaluate the predictive performance of a single feature we used its normalized values for all samples as probability scores for the samples to be revoked. Based on these probabilities two auROC scores were calculated. One using the probabilities as explained and another auROC with the values of $1.0 - probability$, because features can be negatively correlated. For each feature the higher auROC was used and Figure 1 shows the best performing feature for each feature set while the blue bars show the performance of the best performing classification model identified by the grid search.

In most cases there is a drastic increase in predictive performance when a classification model is trained on the whole feature set. For DNase-seq data from mouse a high auROC can be achieved using only the *distal intergenic feature* from *RA* or the *-500 bin* from *TSS*. In theses two cases the performance increases just slightly with the machine learning. Concluding these first classification results we confirm that in all cases the predictive performance increases when using the machine learning approach instead of single features.

Before investigating and describing the results in more detail we analyzed whether there is a big impact in using homogeneous datasets for training the model, as we did so far, or whether similar performance scores can be achieved by a generalized model. This generalized model was trained on the mixed dataset using the best performing classification algorithm from the grid search within a ten-fold cross-validation. However, instead of calculating one auROC on the mixed data, for each sample that was correctly or incorrectly classified its conditions were derived to accumulate the true-positives and false-positives for each species-assay combination. Hence, during the validation of the model trained on the mixed dataset, the predicted probabilities are split up according to the experimental setup an individual sample belongs to.

As shown in Figure 2 the performance increases for a model that was trained only on samples from the same species-assay combination. This indicates that the best is still to use a model that was trained on the very same data conditions, even though this might reduce the sample size. However, the differences are not substantial which confirms that machine learning algorithms are capable to learn several patterns at the same time when these different patterns are present, represented by different subsets in the given dataset. The patterns within a bigger, mixed dataset might be more reliable due to a higher sample size.

In general, the auROC scores were remarkable on the homogeneous datasets, reaching the best possible value of 1.0 while the minimum was 0.701, and predictive

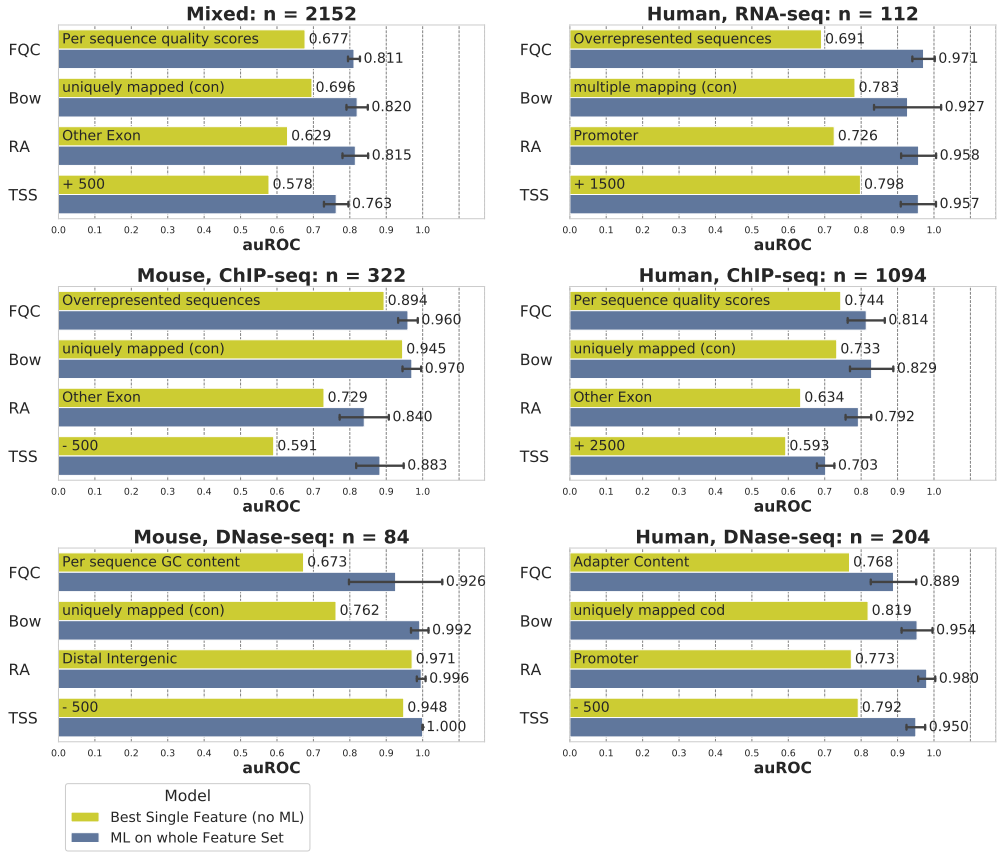


Figure 1. The yellow bars represent the best auROC of a single feature for each feature set. The name of the best feature is written in the bar. In comparison the auROC scores achieved by classification models trained with all features from a quality feature set, shown by the blue bars. The error bars describe the standard deviation of the auROC scores across the ten-fold cross-validation while the bar represents the mean auROC. Number of samples available for each subset are shown in the subheaders.

performance increased when all feature sets were included. The performance evaluated on the mixed dataset the auROC increased to 0.926 whit all feature sets. The values for auPRC were also high in the range between 0.782 and 1.0 as shown in Fig. S1b. In all cases, as shown in Fig. 2, the inspection of the auPRCs led to the same conclusions we derived from the auROCs.

Interestingly, the *RA* feature set was the best for DNase-seq and the worst for ChIP-seq in comparison to *FQC* and *Bow*. Moreover, combining *ALL* feature sets led to the best results in either assay. The underperformance of using the *RA* or *TSS* alone in ChIP-seq might reflect the higher heterogeneity of the ChIP-seq data for which our classification models did not know two important factors of variability: the biosample and the antibody. The complementarity of *RA* and *TSS* with other feature sets for ChIP-seq and the best results with DNase-seq might reflect the specificity of *RA* and *TSS* to model one factor of variability: the biosample.

The results as well that the predictive performance on mouse-ChIP-seq data was higher compared to human-ChIP-seq. This observation may be explained by an overrepresentation of immortalized cell lines in the human dataset. These cell-lines are often derived from cancer cells and are known for a higher number of genomic

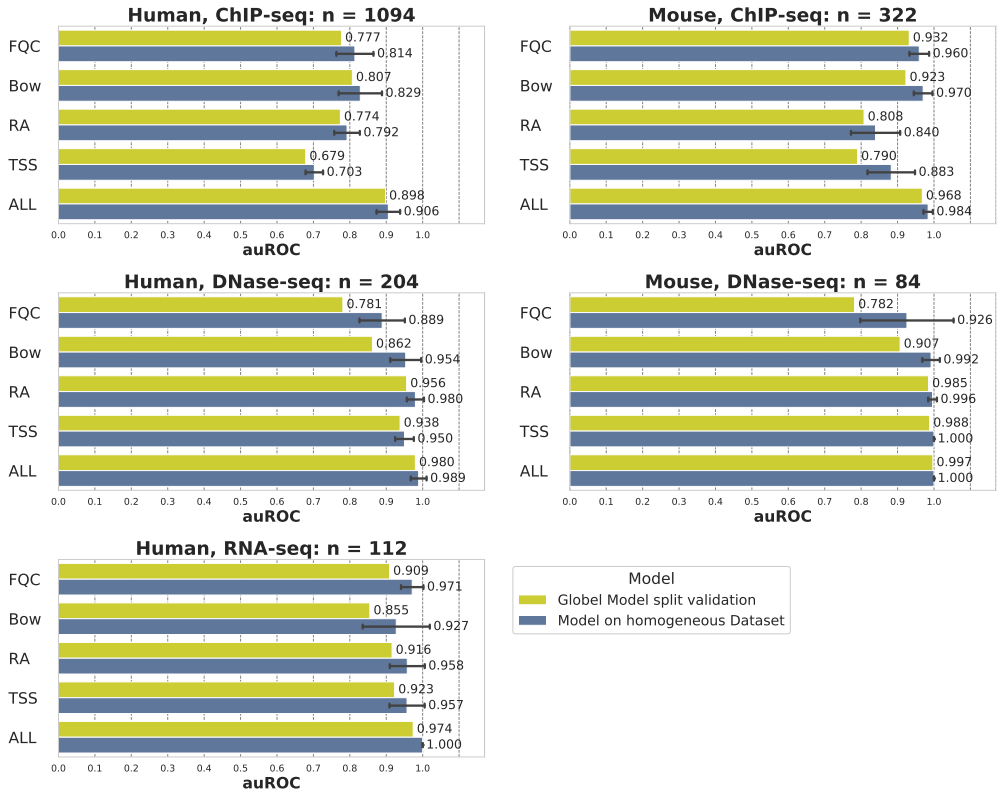


Figure 2. The performance of the global model validated by splitting up the samples in the testing sets, represented by the yellow bars. Compared to the performance of a model trained and tested on a homogeneous dataset, represented by the blue bars. The error bars describe the standard deviation across the ten-fold cross-validation applied within the grid search.

rearrangements and mutations, which might induce noise that has a negative impact to the training process of the classification models. In general there is a higher probability of the occurrence of bias in the evaluation of the classification models on the smaller subsets, especially for mouse-DNase-seq.

While Figure 2 shows only five cases defined by using either a single feature set or all feature sets, we ran the grid search for each possible combination that can be composed out of the four features sets. This resulted in 15 cases per species-assay combination, the values for auROC but also auPRC, accuracy, and F1 are shown in Figure S1.

Best algorithms

Out of the grid search the best classification algorithms were identified. Figure 3 shows the best performing algorithms across the 15 cases for each species-assay combination. RFC performed best on ChIP-seq data. This agrees with other studies, also based on sequencing data, in which a random forest based algorithm achieved the best performance [9, 15]. In case of ChIP-seq also MLP and XGB are performing well, while KNN and SVC have never been the best algorithm. In contrast, in all combinations without ChIP-seq SVC has the best performance in two cases each. We observe that SVC performs better on the smaller datsasets from DNase-seq and RNA-seq which

agrees with another study showing that support vector machines perform well on datasets with lower sample sizes [13]. Considering DNase-seq and RNA-seq data the best performing algorithm is definitely GBC that was the top algorithm in twelve out of 15 cases from the RNA-seq dataset. Moreover, RFC achieved the best auROC in six scenarios involving DNase-seq and MLP performed best in four cases considering the smallest subset from mouse, DNase-seq. Further investigations showed that: (i) there were in general small differences in performance between the algorithms RFC, XGB MLP, and GBC. (ii) performance was improved with prior feature selection more frequently for GBC. (iii) GBC was more stable with respect to changing numbers of features caused by the feature set combinations (results not shown). Note that more MLP parameter settings would have been evaluated using a GPU implementation, likely leading to better results.

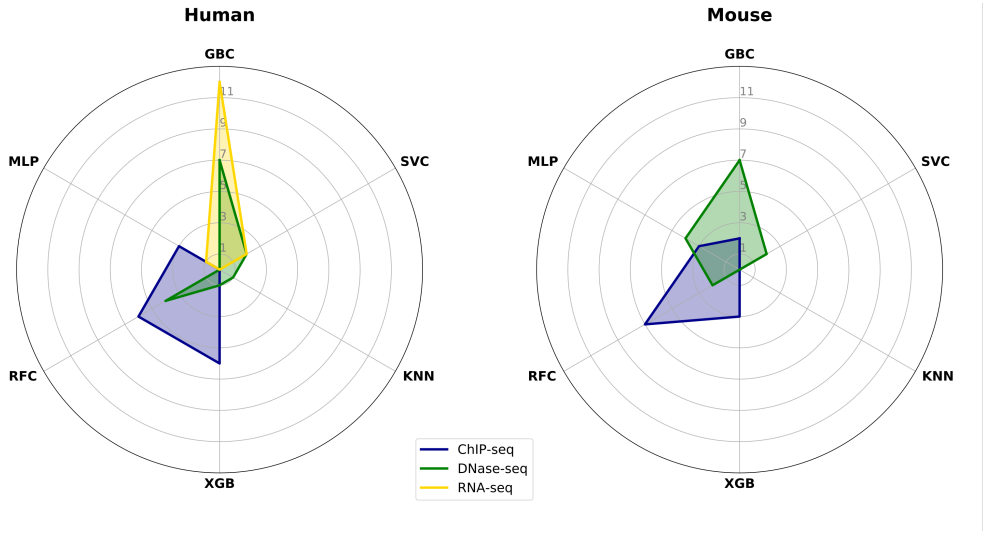


Figure 3. Radar-Plots showing the number of cases in which the different algorithms achieved the best auROC within the grid search. Subplots are separated by assay and the species are separated by color.

Generalization across species

For the different species-assay combinations we confirmed the possibility to predict the quality of sequencing data. However, we wondered whether this would still be possible when the data within the training and testing sets are from different species. This possibility would indicate that the classification models are generalizable across species and can be used also for species for which training data is not available.

In order to answer this question, we performed classification experiments using training and testing sets containing data from different species. Due to the data availability we investigated this question for ChIP-seq and DNase-seq data and applied the following procedure to perform a fair comparison regarding the different scenarios. Given one of the assays we created training and testing sets for each of the two species human respectively mouse. Using a five-fold cross-validation five training and testing sets were created for each species and the grid search was applied to the training sets. The best performing model for each training set was identified by the highest auROC achieved within a ten-fold cross-validation. Finally, each of these models was evaluated twice, firstly on the corresponding testing set from the five-fold cross-validation that contains data from the same species and secondly on the corresponding testing set

containing data from the differing species.

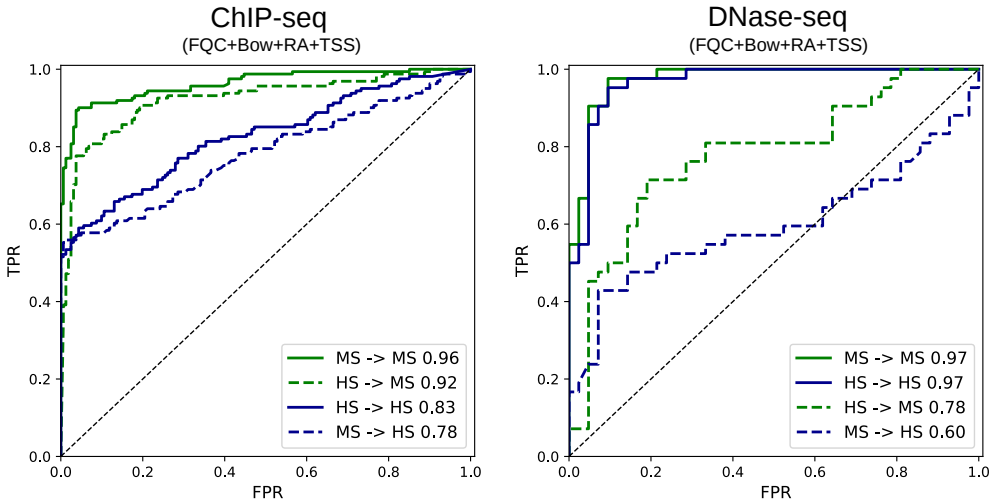


Figure 4. ROC-curves showing the classification performance for different species-assays combinations on all features *(FQC+Bow+RA+TSS)*. The solid lines represent cases in which data from the same species was used to define training and testing sets. The dashed lines show the performance on cases in which the species defining the training data differs from the species defining the testing data; e.g. "HS → MM" means the training set contains only data from human (Homo sapiens), while the testing data is only from mouse (Mus musculus)

The ROC-curves in Figure 4 represent the performance in the different scenarios using all feature sets. ROC-curves about the classification scenarios with a single feature set are shown in supplementary S2. For ChIP-seq, considering cases in which FQC or Bow are used or at least involved, the performance dropped slightly when the quality of mouse samples was classified and human data was used to train the model and vice versa. For the feature sets RA and TSS the drop is more substantial, while we observe a lower performance in general using these. However, these results indicate that predictive information exist, especially within the FQC and Bow features, that can be used across species to classify the quality of fastq files. Considering DNase-seq, the predictive performance was high when the species was the same for training and testing set especially when the features from RA and TSS were involved. However, across all cases, a large drop in predictive performance can be observed when a different species defined the data for training than for testing. Though, the auROC scores were still better than random expectations, the results indicate that within DNase-seq data models can not be generalized across different species.

Leave one experiment out - Application

In order to validate the concept of using classification models more focused on an application, we investigated its application to specific experiments in ENCODE. These experiments contain revoked files as well as released files. For the different species-assay combinations a different amount of such experiments could be queried from the data portal. However, for each of these experiments the best performing classification algorithm was identified according to the experiments setting and the algorithm was applied to train a classification model on all files from the same setting excluding the files that belong to the experiment. The classification model was then applied to the

files which were left out for training and their scores are plotted to see whether revoked files tend to have a higher score that describes their probability of being revoked.

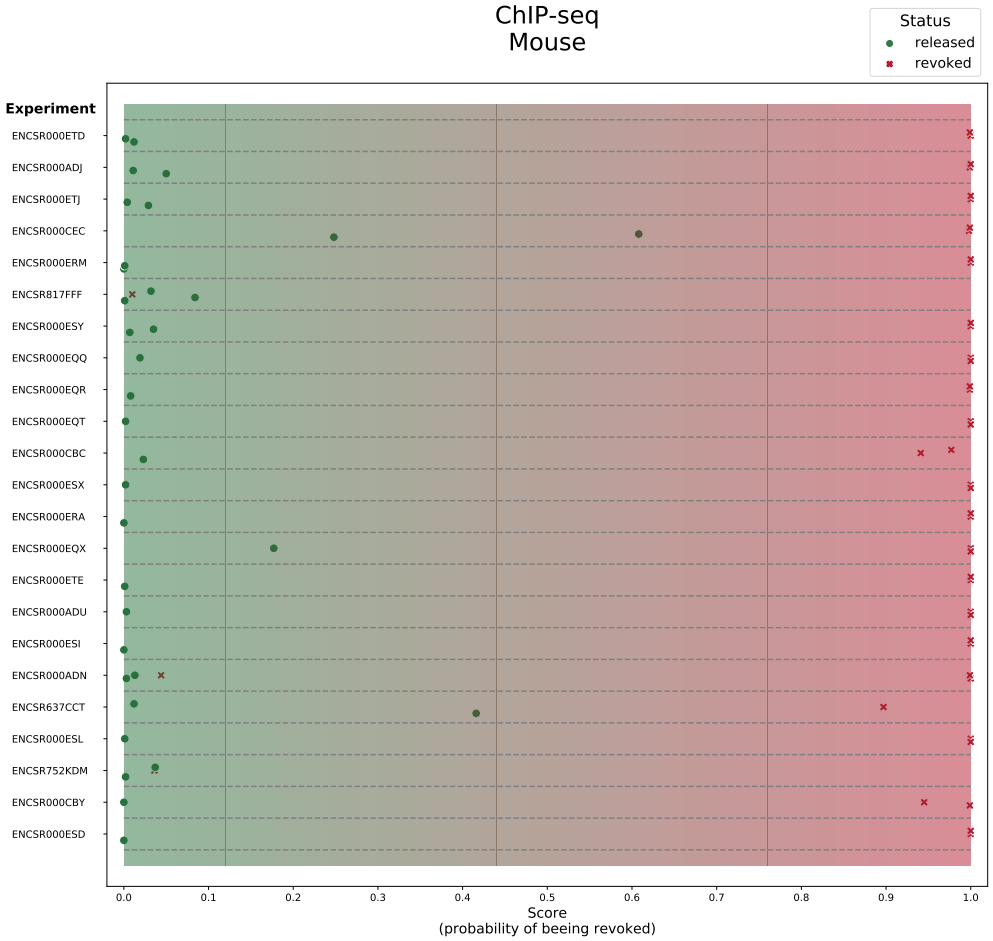


Figure 5. Experiment IDs are shown along the y-axis while the files that belong to one experiment are plotted within the area indicated by the dashed lines. The x-axis describes the score that is the predicted probability of a file to be revoked.

For most of the experiments the revoked files have a much higher score while the the scores of the released files are close to zero as shown in Figure 5. In some cases like ENCSR000CEC and ENCSR637CCT the difference is not as substantial, however the files are still ordered correctly. The same analysis was done for other species-assay combinations as well and similar plots are provided in supplementary S3. In the majority of all cases it was possible to see a clear separation of high and low quality files by the score especially for the cases from. Note that the real ENCODE status for most of these files was defined after uploading them to the ENCODE data portal, applying the preprocessing pipeline and confirming each individual case manually. The validation presented in this subsection shows that this procedure, that could be very time consuming, can be replaced by considering the predicted scores calculated by the pre-trained classification models computed for this study.

Novel quality assessment tool

In addition to the findings, we just described, we also provide a tool *seqQscorer* implemented in python. The code as well as all pre-trained models for the different species-assay and feature set combinations are available in a github repository: seqQscorer. A short description and an application example is also provided.

Conclusion

We have demonstrated that machine learning algorithms can perform automatic quality assessment of early stage sequencing data. Such a method will ease the quality assessment of NGS data and should have a beneficial impact on clinical studies involving a large number of patients. Moreover, the classification models are generalizable to some extent and might be highly relevant to classify data related to unseen cell types or diseases (e.g. rare diseases). Eventually, we demonstrated the applicability of trained models to single experiments in order to make a decision about the revocation of single files / replicates that have the potential of negatively influencing the results due to their low quality.

In future work we plan to improve the models by including more data sources, assays and species. Also, we will focus our investigations to large clinical datasets and the impact of our models to related applications such as diagnosis or prognosis. Finally, we would like to encourage the community to consider the collection of low quality NGS data. We were able to surely confirm the potential of our concept on the data that was available so far. However, as the ENCODE portal is not considered to provide a collection of low quality data, we were limited in creating larger datasets that would have been suitable for more powerful methods and classification algorithms from deep learning.

Supporting Information

S1 Figure

Predictive performance measures for all feature set combinations.

S2 Figure

Across species prediction using single feature sets.

S3 Figure

Remaining cases for leave-one-experiment-out analysis.

Acknowledgments

We thank the International PhD Programme (IPP) of the Institute of Molecular Biology, Mainz, for financial support. We thank Prof. Susanne Gerber and Prof. Leszek Wojnowski for meaningful discussions. We gratefully acknowledge the computing time granted on the supercomputer MOGON at Johannes Gutenberg University Mainz (hpc.uni-mainz.de).

References

1. J. Brown, M. Pirrung, and L. A. McCue. Fqc dashboard: integrates fastqc results into a web-based, interactive, and extensible fastq quality control tool. *Bioinformatics*, 33(19):3137–3139, 2017.

2. L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

3. S. Chen, T. Huang, Y. Zhou, Y. Han, M. Xu, and J. Gu. Afterqc: automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinformatics*, 18(3):80, 2017.

4. T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

5. M. D. Chikina and O. G. Troyanskaya. An effective statistical evaluation of chipseq dataset similarity. *Bioinformatics*, 28(5):607–613, 2012.

6. W. de Souza, B. de Sá Carvalho, and I. Lopes-Cendes. Rqc: A bioconductor package for quality control of high-throughput sequencing data. *Journal of Statistical Software*, 87(CN 2):1–14, 2018.

7. P. Ewels, M. Magnusson, S. Lundin, and M. Käller. Multiqc: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19):3047–3048, 2016.

8. B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357, 2012.

9. J. Li, B. Jew, L. Zhan, S. Hwang, G. Coppola, N. B. Freimer, and J. H. Sul. Forestqc: quality control on genetic variants from next-generation sequencing data using random forest. *bioRxiv*, page 444828, 2018.

10. C.-C. Lo and P. S. Chain. Rapid evaluation and quality control of next generation sequencing data with faqcs. *BMC Bioinformatics*, 15(1):366, 2014.

11. M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

12. G. K. Marinov, A. Kundaje, P. J. Park, and B. J. Wold. Large-scale quality analysis of published chip-seq data. *G3: Genes, Genomes, Genetics*, 4(2):209–223, 2014.

13. P. Matykiewicz and J. Pestian. Effect of small sample size on text categorization with support vector machines. In *Proceedings of the 2012 workshop on biomedical natural language processing*, pages 193–201. Association for Computational Linguistics, 2012.

14. M.-A. Mendoza-Parra, W. Van Gool, M. A. Mohamed Saleem, D. G. Ceschin, and H. Gronemeyer. A quality control system for profiles obtained by chip sequencing. *Nucleic Acids Research*, 41(21):e196–e196, 2013.

15. M. Oh and E. Cosgun. Exploring the consistency of the quality scores with machine learning for next-generation sequencing experiments. American Society of Human Genetics, October 2018. PgmNr 1480/T.

16. R. V. Pandey, S. Pabinger, A. Kriegner, and A. Weinhäusel. Clinqc: a tool for quality control and cleaning of sanger and ngs data in clinical research. *BMC Bioinformatics*, 17(1):56, 2016.

17. R. K. Patel and M. Jain. Ngs qc toolkit: a toolkit for quality control of next generation sequencing data. *PloS One*, 7(2):e30619, 2012.

18. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

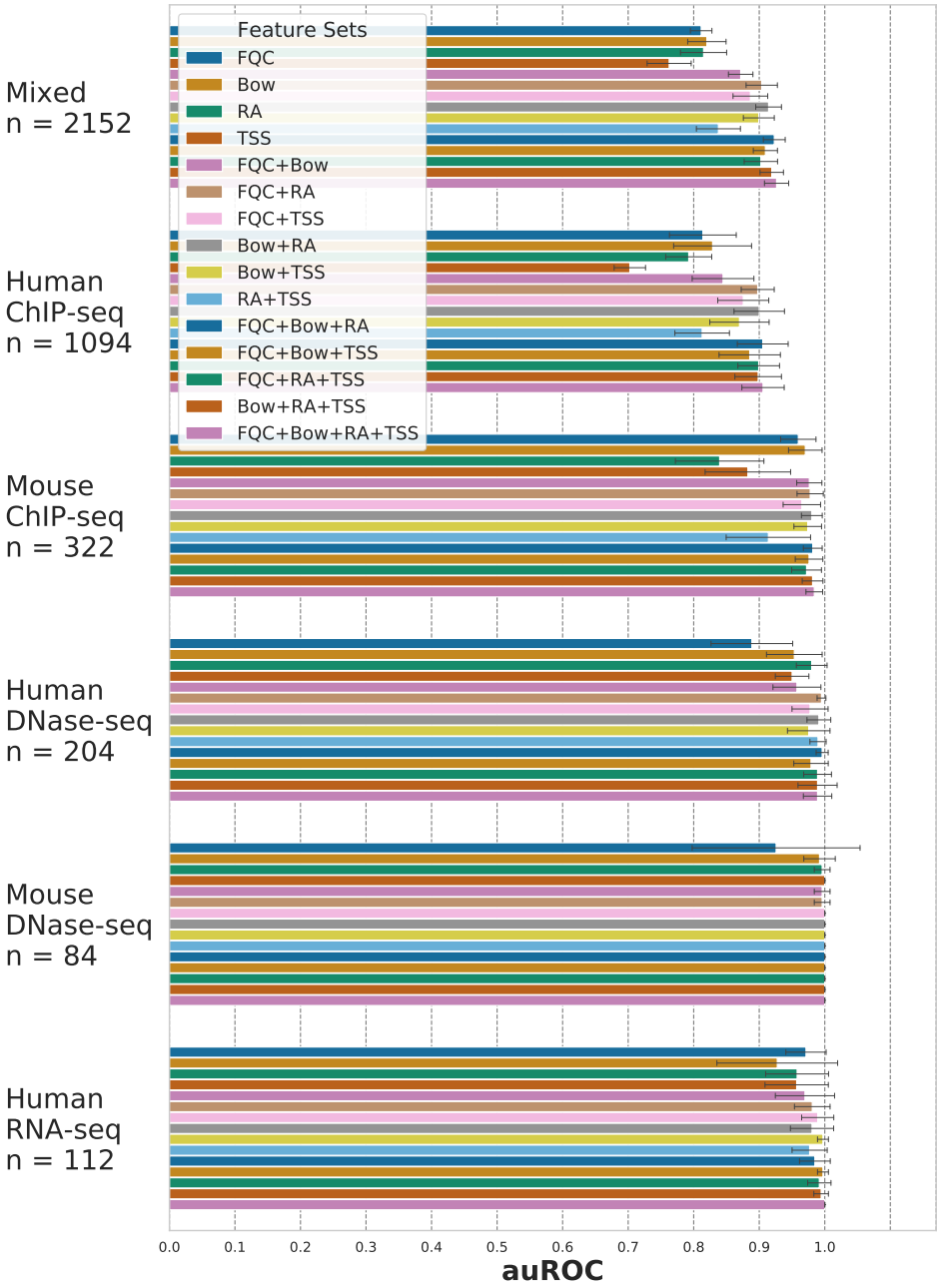
19. R. Schmieder and R. Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864, 2011.

20. C. Trapnell, L. Pachter, and S. L. Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, 03 2009.

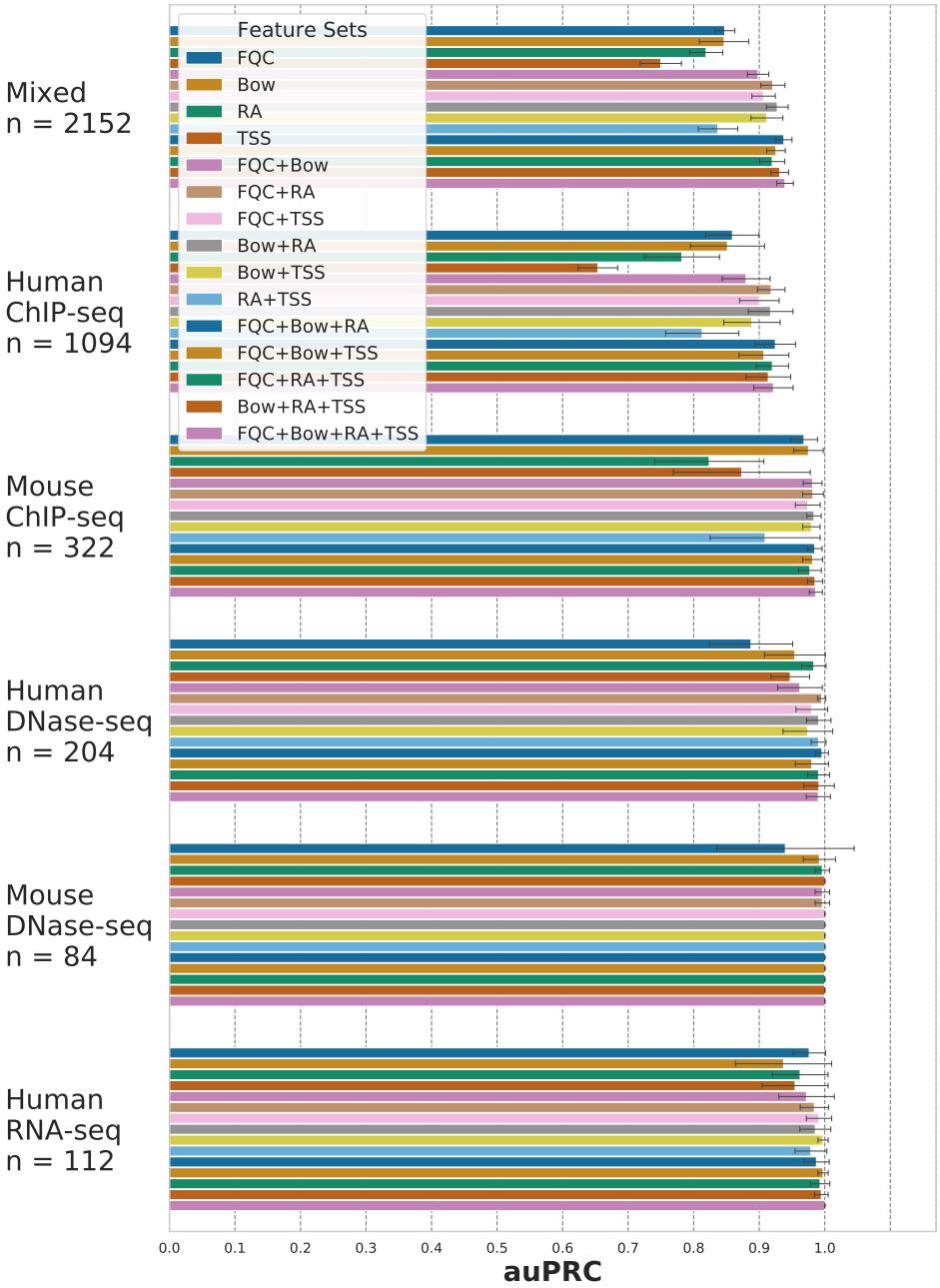
21. G. Yu, L.-G. Wang, and Q.-Y. He. Chipseeker: an r/bioconductor package for chip peak annotation, comparison and visualization. *BMC Bioinformatics*, 31(14):2382–2383, 2015.

22. L. J. Zhu. Integrative analysis of chip-chip and chip-seq dataset. In *Tiling Arrays*, pages 105–124. Springer, 2013.

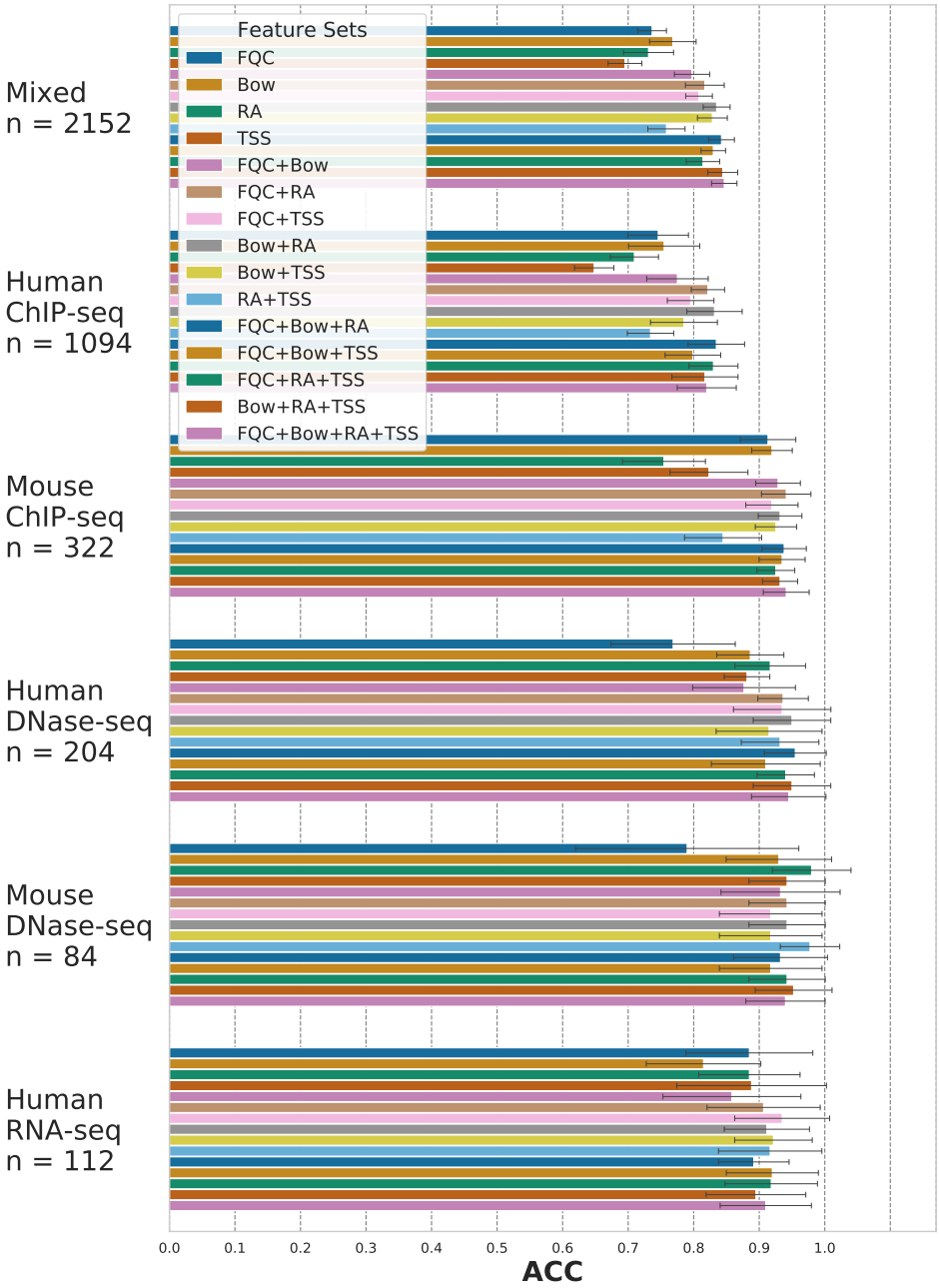
23. L. J. Zhu, C. Gazin, N. D. Lawson, H. Pagès, S. M. Lin, D. S. Lapointe, and M. R. Green. Chippeakanno: a bioconductor package to annotate chip-seq and chip-chip data. *BMC Bioinformatics*, 11(1):237, 2010.


Mixed
n = 2152
Human
ChIP-seq
n = 1094
Mouse
ChIP-seq
n = 322
Human
DNase-seq
n = 204
Mouse
DNase-seq
n = 84
Human
RNA-seq
n = 112
Feature Sets
FQC
Bow
RA
TSS
FQC+Bow
FQC+RA
FQC+TSS
Bow+RA
Bow+TSS
RA+TSS
FQC+Bow+RA
FQC+Bow+TSS
FQC+RA+TSS
Bow+RA+TSS
FQC+Bow+RA+TSS
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
auROC

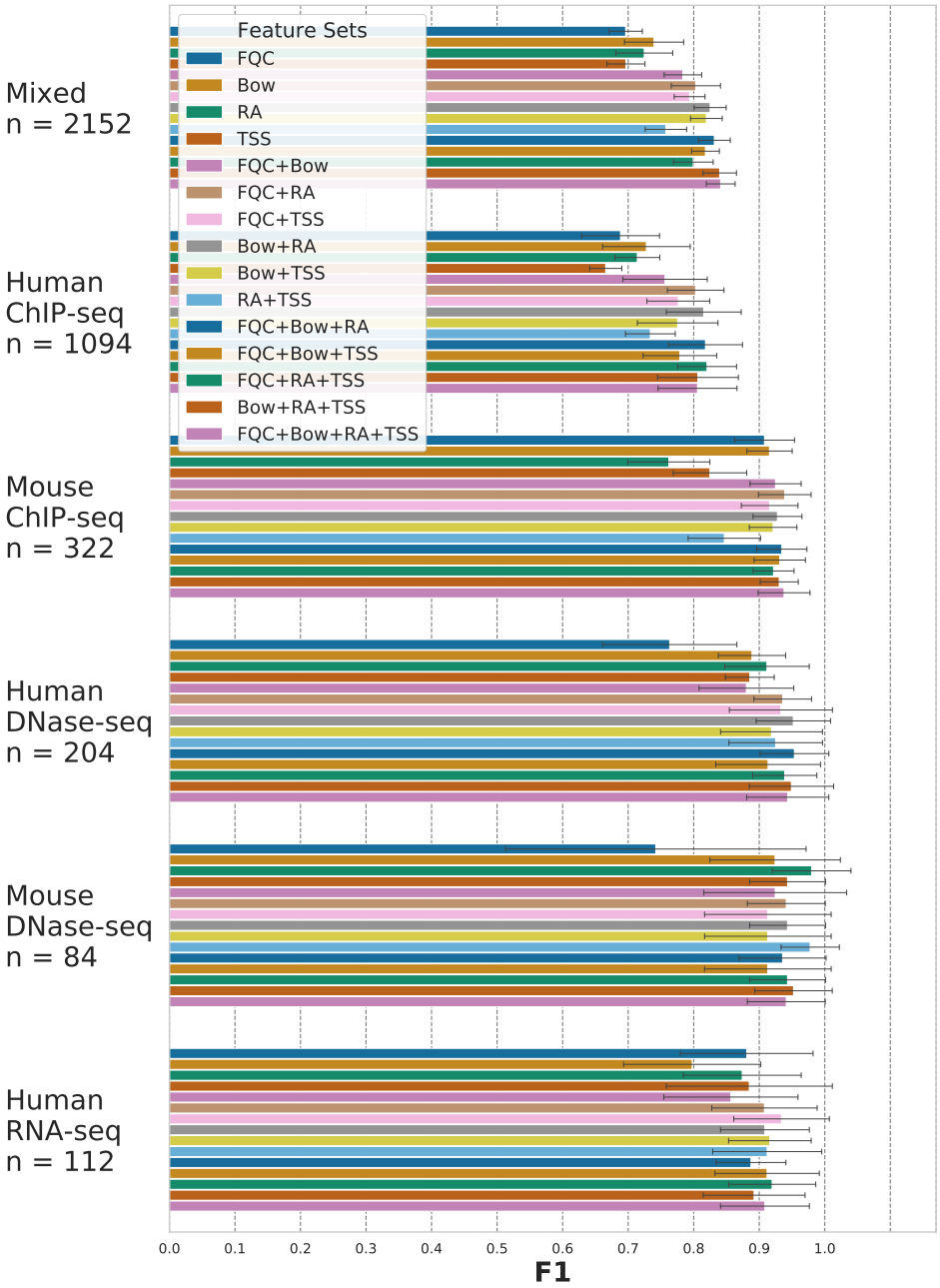
(a) auROC


Feature Sets
FQC
Bow
RA
TSS
FQC+Bow
FQC+RA
FQC+TSS
Bow+RA
Bow+TSS
RA+TSS
FQC+Bow+RA
FQC+Bow+TSS
FQC+RA+TSS
Bow+RA+TSS
FQC+Bow+RA+TSS
Mixed
n = 2152
Human
ChIP-seq
n = 1094
Mouse
ChIP-seq
n = 322
Human
DNase-seq
n = 204
Mouse
DNase-seq
n = 84
Human
RNA-seq
n = 112
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
auPRC

(b) auPRC


Feature Sets
FQC
Bow
RA
TSS
FQC+Bow
FQC+RA
FQC+TSS
Bow+RA
Bow+TSS
RA+TSS
FQC+Bow+RA
FQC+Bow+TSS
FQC+RA+TSS
Bow+RA+TSS
FQC+Bow+RA+TSS
Mixed
n = 2152
Human
ChIP-seq
n = 1094
Mouse
ChIP-seq
n = 322
Human
DNase-seq
n = 204
Mouse
DNase-seq
n = 84
Human
RNA-seq
n = 112
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
ACC

(c) Accuracy



(d) F1-Measure

Figure S1. Predictive performance measures for all feature set combinations. Bar-Plots representing the performance of classification models trained and evaluated on different subsets of the data. The colors describe the different feature set combinations. Several performance measures area under ROC-curve (auROC), area under precision-recall curve (auPRC), accuracy (ACC), F1 measure are shown in the subplots a), b), c), and d) respectively.

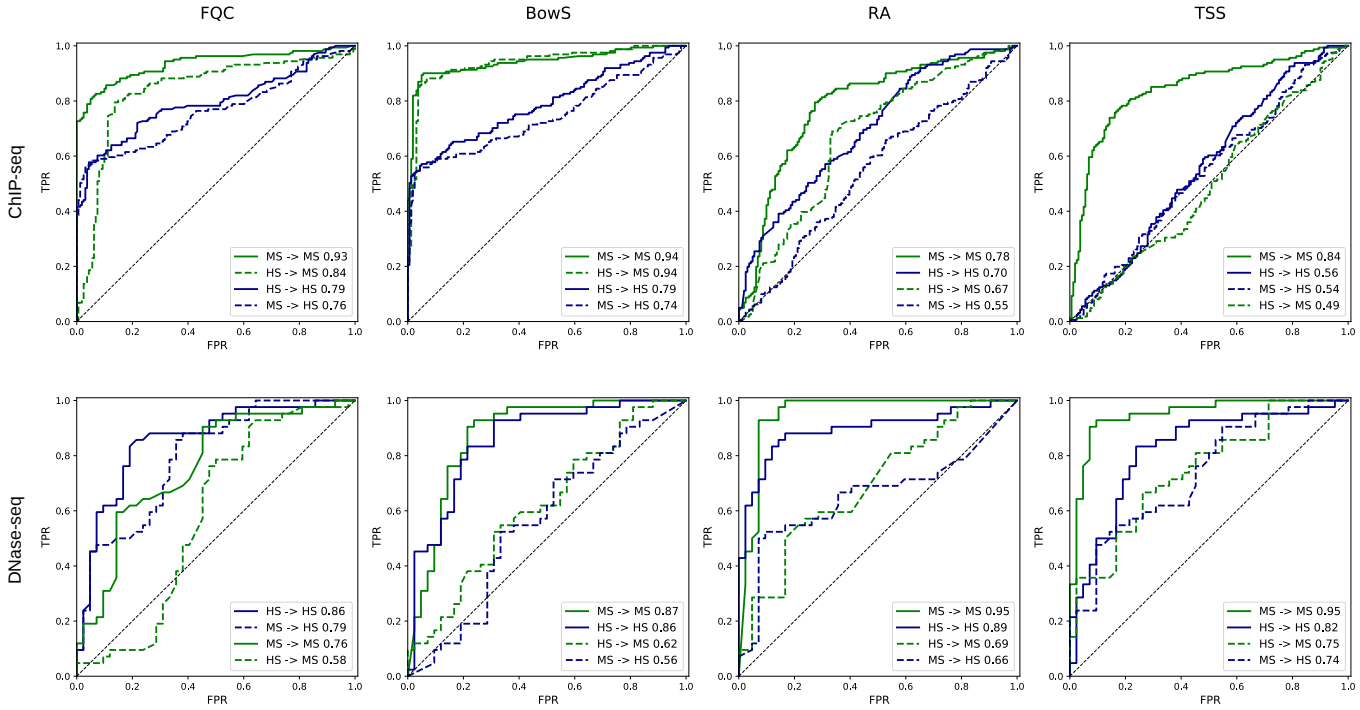
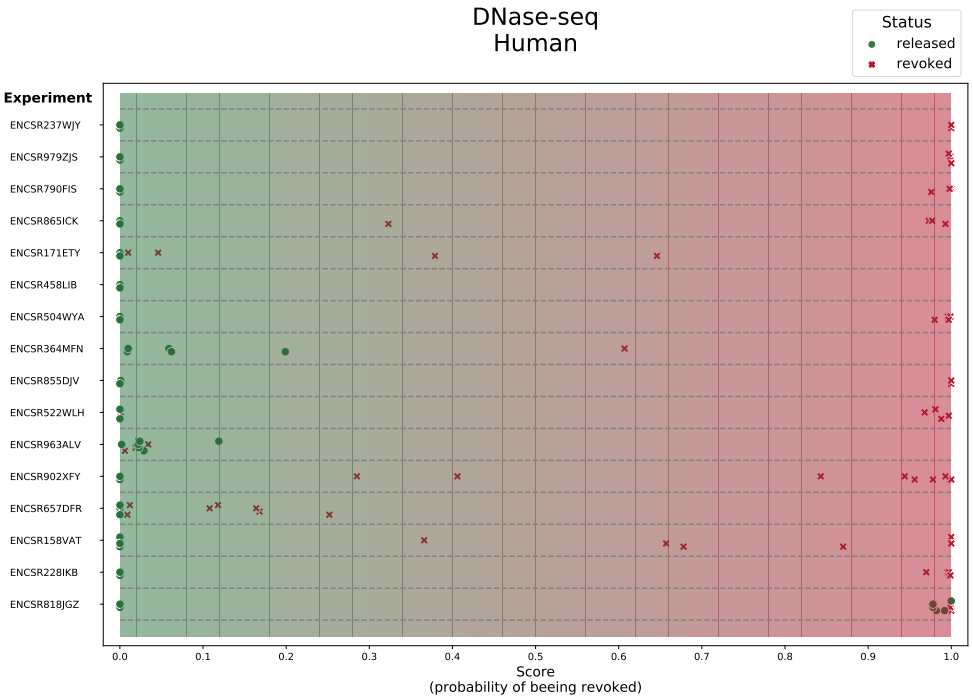
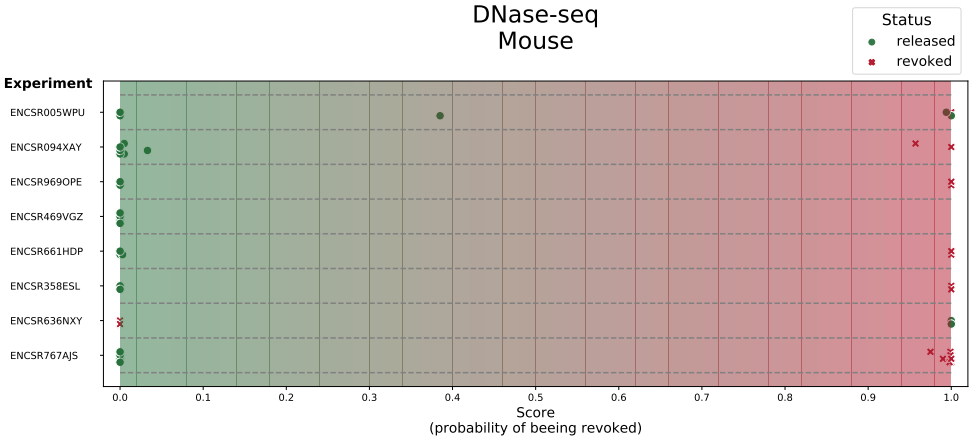


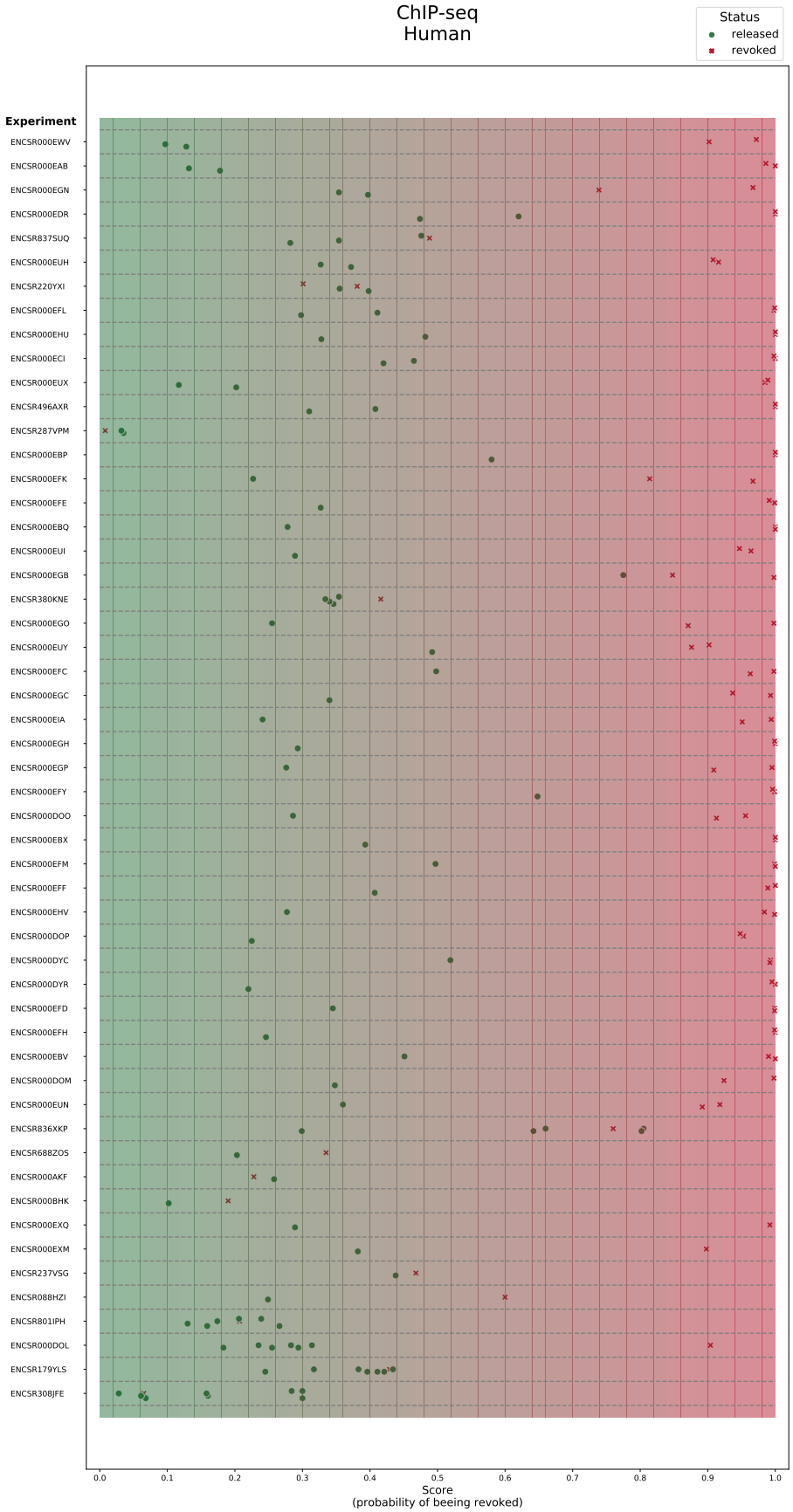
Figure S2. Across species prediction using single feature sets.
ROC-curves showing the classification performance for different species-assays combinations with different feature sets. The solid lines represent cases in which data from the same species was used to define training and testing sets. The dashed lines show the performance on cases in which the species defining the training data differs from the species defining the testing data; e.g. "HS → MM" means the training set contains only data from human (Homo sapiens), while the testing data is only from mouse (Mus musculus)



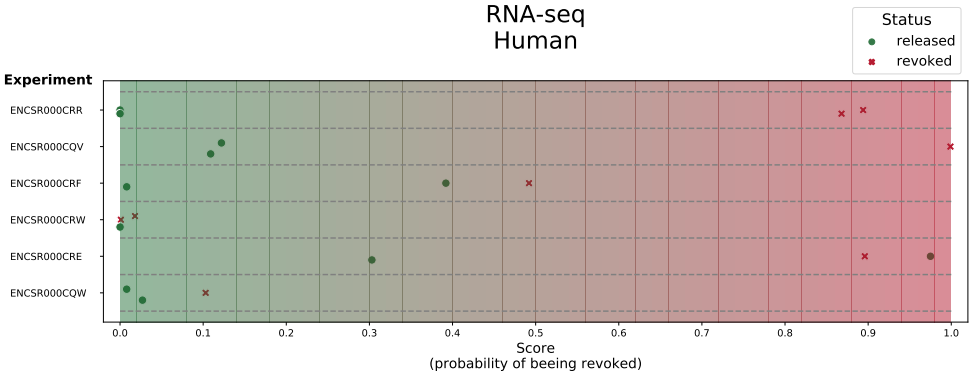
(a) DNase-seq data from human



(b) DNase-seq data from mouse



(c) ChIP-seq data from human



(d) RNA-seq data from human

Figure S3. Remaining cases for leave-one-experiment-out analysis. Experiment IDs are shown along the y-axis while the files that belong to one experiment are plotted within the area indicated by the dashed lines. The x-axis describes the score that is the predicted probability of a file to be revoked. Experiments were selected for different experimental settings as described in the subplot description of a), b), c), and d).