GigaScience, 10, 2021, 1–13

https://doi.org/10.1093/gigascience/giab064
Technical Note

TECHNICAL NOTE

Bias-invariant RNA-sequencing metadata annotation

Hannes Wartmann, Sven Heins, Karin Kloiber and Stefan Bonn*

Institute of Medical Systems Biology, Center for Biomedical AI, University Medical Center Hamburg-Eppendorf, 20251 Hamburg, Germany

*Correspondence address. Stefan Bonn, Institute of Medical Systems Biology, Center for Molecular Neurobiology, Center for Biomedical AI, University Medical Center Hamburg-Eppendorf, 20251 Hamburg, Germany. E-mail: sbonn@uke.de http://orcid.org/0000-0003-4366-5662

Abstract

Background: Recent technological advances have resulted in an unprecedented increase in publicly available biomedical data, yet the reuse of the data is often precluded by experimental bias and a lack of annotation depth and consistency. Missing annotations makes it impossible for researchers to find datasets specific to their needs. **Findings:** Here, we investigate RNA-sequencing metadata prediction based on gene expression values. We present a deep-learning–based domain adaptation algorithm for the automatic annotation of RNA-sequencing metadata. We show, in multiple experiments, that our model is better at integrating heterogeneous training data compared with existing linear regression–based approaches, resulting in improved tissue type classification. By using a model architecture similar to Siamese networks, the algorithm can learn biases from datasets with few samples. **Conclusion:** Using our novel domain adaptation approach, we achieved metadata annotation accuracies up to 15.7% better than a previously published method. Using the best model, we provide a list of >10,000 novel tissue and sex label annotations for 8,495 unique SRA samples. Our approach has the potential to revive idle datasets by automated annotation making them more searchable.

Keywords: RNA-seq metadata; data reusability; automated annotation; machine learning; domain adaptation; bias invariance; deep learning; computational biology; bioinformatics

Introduction

Next-generation RNA-sequencing (RNA-seq) has been a pillar of biomedical research for many years [1, 2]. It allows researchers to simultaneously quantify and compare the expression of tens of thousands of genomic transcripts. A continuous decrease in cost makes RNA-seq a widely available method of choice to uncover the molecular basis of biological development and diseases [3, 4]. As a result, recent years have seen a strong growth in publicly accessible RNA-seq data. The actual reuse and integration of this data, however, have been largely limited by the lack of consistent metadata annotation and individual dataset bias [5, 6]. The lack of metadata annotation for RNA-seq samples, such as tissue of origin, disease, or sex phenotype, prohibits experimenters from finding data that are relevant to their research. Moreover, dataset biases [7] due to differences in protocols and technologies [8] or of a biological nature hinder integration and comparative analysis.

To allow for efficient data reuse, publicly available data have to be harmonized and well annotated with standardized metadata and subsequently be made accessible (and searchable) [9]; this practice is followed by the Genotype-Tissue Expression Project (GTEx) [10] and The Cancer Genome Atlas (TCGA). Nevertheless, the primary database for next-generation sequencing projects, the SRA [11], stores raw sequencing information that lacks rigorous standards of curation, which limits the reusability of its data.

Efforts to predict missing or sparse metadata in public RNA-seq resources have shown promising results. For instance, recently published studies used text mining approaches to retrieve missing annotation from associated abstracts or free text annotations in the data sources [12–14]. Others have used RNA-seq expression values for phenotype prediction. For example, machine learning (ML) has successfully been applied to disease and cell type classification [15, 16] or survival outcomes on TCGA data [17]. Others have taken advantage of prior domain knowl-

Received: 16 December 2020; **Revised:** 11 June 2021; **Accepted:** 1 September 2021

edge such as gene regulatory networks for enhanced feature selection [18, 19]. Recently a linear regression model fitted to GTEx data has been presented for the prediction of tissue, sex, and other phenotypes of SRA and TCGA samples [20]. These efforts provide evidence that missing RNA-seq metadata can be successfully predicted on the basis of genomic expression values using ML approaches.

Artificial neural networks (ANNs) in their various forms and functions consistently outperform classical ML approaches in a large variety of biological tasks, including classification, data generation, and segmentation [21–24]. Given large training datasets, these algorithms can learn complex representations of data by automatically weighting and combining features nonlinearly. This has led us to hypothesize that ANN-based models could increase the performance in metadata prediction beyond that of classical ML approaches such as linear regression. Of special interest in this context is domain adaptation (DA) [25], a subfield of ML that aims to specifically alleviate problems conferred by dataset bias [26]. The aim of DA is to build and train ANNs on a source domain in such a way that the model performs well on a biased target domain.

Here, we present a DA approach capable of leveraging a number of dataset biases, boosting generalizability of phenotype prediction. We developed the model using 3 data sources (GTEx, TCGA, and SRA) of different size and with a different degree of bias. To validate our approach we compare it to a previously suggested linear model (LIN) [20], as well as a standard supervised multi-layer perceptron (MLP) for prediction of tissue of origin, sex, and sample source. Importantly, we find that our DA network is able to integrate heterogeneous training data such that classification accuracy is up to 15.7% higher for tissue classification compared to the supervised LIN model. We subsequently apply trained models to generate and make available new metadata for 8,495 unique SRA samples.

Methods

Data acquisition

To train and test models we gathered data from 3 different sources (i.e., GTEx, TCGA, and SRA), each with a different level of heterogeneity (Supplementary Fig. S1). We measure data source heterogeneity by the number of unique datasets (or studies) in the source. Each dataset (or study) is believed to have a unique bias. Biases stem from the unique circumstances, protocols, and reagents used, as well as biological factors of the study [7, 8]. Here we define a dataset as all the RNA-seq samples from 1 study on the basis of the assumption that they were obtained and processed under equal conditions. To avoid additional biases by the use of different bioinformatic alignment pipelines [27] all data were downloaded from recount2 (release 13.09.19) [28]. Recount2 aggregates raw RNA-seq data from different sources and reruns the data through the Rail-RNA alignment pipeline [29]. The RSE V2 files of all available RNA-seq projects (n = 2,036) from recount2 were downloaded using the recount R package (v 1.11.13). The downloaded data were separated into 3 different data sources according to their origin. Figure 1A gives a general overview of the data obtained, the pre-processing steps, and dataset preparation.

GTEx

GTEx (v6) comprises 9,662 samples from 554 healthy donors across 31 tissues. GTEx strives to build a highly homogeneous dataset with strict guidelines on donor selection, biopsy, and sequencing methodology [31]. We considered the GTEx data source to have a single dataset bias.

SRA

From the SRA, a total of 2,034 studies containing a total of 49,657 samples were downloaded from recount2. Every SRA study was potentially processed at a different site by a different technician following different standards. In addition, the underlying biological condition of the samples is often unclear. We assume each study to have a unique dataset bias, which makes the SRA a highly heterogeneous data source. In addition, data annotation is not standardized, resulting in sparse metadata with low fidelity.

TCGA

RNA-seq data for TCGA were downloaded, consisting of 11,284 samples spanning 26 tissues. While there are 740 samples of healthy donors across 20 tissues, >90% of the samples are tumor biopsies from different tissues and different stages of tumor progression. TCGA accepts sequence data from different locations using different sequencing technologies. Despite the high level of standardization and reliability of metadata information, heterogeneity is also inherent to the TCGA dataset due to the biological context (cancers, stages) albeit not as pronounced as in the SRA.

Pre-processing of SRA data source

In this study, we focus on bulk mRNA-seq data because they are by far the most frequent RNA type in each of the 3 data sources used. The following approaches were used to remove data from single-cell and small RNA-seq studies from further analysis: First, we identified small RNA-seq data on the basis of the total fraction of small RNA counts and protein coding RNAs. Specifically, we considered a subset of the Gencode gene types (i.e., protein_coding and processed_pseudogene vs ribosomal RNA, microRNA, misc_RNA, small nuclear RNA, and long intervening noncoding RNA). Every sample that had its maximum total count fraction not allocated to either protein_coding or processed_pseudogene was removed from further analysis (Supplementary Fig. S2). Second, we removed single-cell RNA-seq studies by scanning titles and abstracts for variations of the words "single cell" and manually validated and excluded the identified samples. In addition to this semi-automatic validation step, we manually validated the 50 largest projects within the SRA data source and removed samples that did not qualify as bulk RNA-seq data. Most importantly, we noticed numerous technical replicates in the remaining SRA data. Using technical replicates to train and test a classification model inflates the reported metrics. Therefore, only samples with a unique experiment accession (SRX) were retained. From the 49,657 SRA samples downloaded initially, 29,685 samples and 1,833 unique studies passed our pre-processing steps.

Metadata

We considered 3 different phenotypes for expression-based prediction. Explicitly, we predicted the tissue of origin of a biopsy (e.g., heart, lung, kidney, ovary), the patient's sex, and sample source (denoting whether the sample was from a patient biopsy or a laboratory-grown cell line) (Fig. 1A).

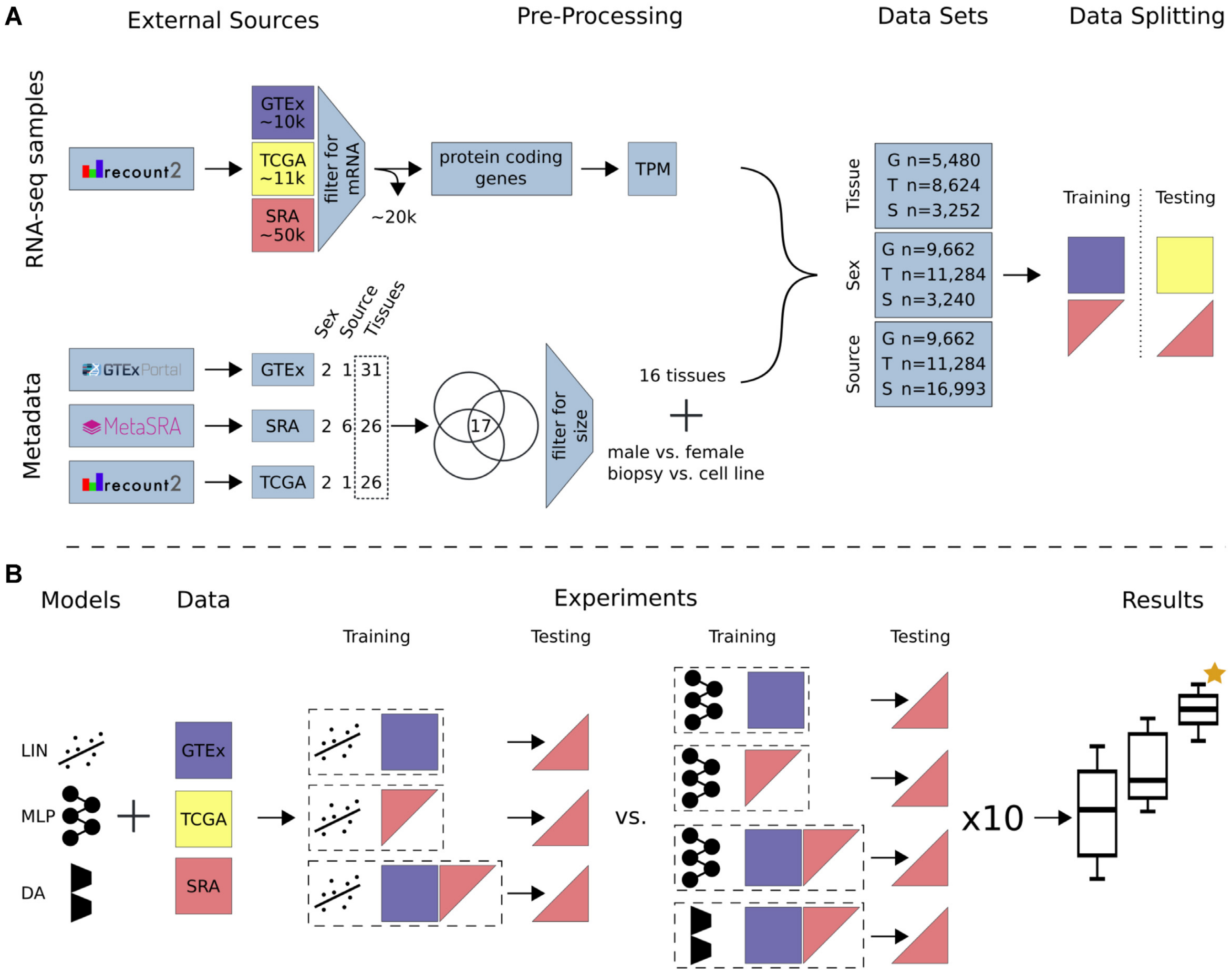

Figure 1: Study overview. (A) All data available on recount2 were downloaded and split into 3 data sources: (i) GTEx, (ii) TCGA, and (iii) SRA. Single-cell and small RNA samples as well as technical replicates were removed from the SRA data. Protein-coding genes were selected from the gene count tables and TPM normalized. Metadata for tissue of origin (e.g., heart), source (e.g., biopsy), and sex phenotype were collected, if available. A subset of 17 tissues (common to GTEx, TCGA, and SRA) was selected and filtered for class size, resulting in 16 tissue classes. For sample source, the 2 largest classes in SRA were selected. Samples were subsequently annotated and training and testing datasets were created. GTEx was only used for model training unless stated otherwise. TCGA was only used for model testing. SRA was split such that samples from 1 study are exclusively in the train or test set. (B) We compare 3 models: LIN (linear model), MLP (multi-layer perceptron), and DA (novel domain adaptation algorithm). Experiments are different combinations of models and data sources. Here, an exhaustive list of experiments for tissue and sex classification tested on SRA data is depicted. Each configuration (dashed box) is made up of a model and training data. The previously published LIN model served as a benchmark for our MLP and DA model. Each model configuration was trained 10 times with different seeds to give an estimation of uncertainty. The best model (orange star) was chosen by comparing average performance across all seeds. After determination of the best model, all available data were used for model training. Previously unlabeled SRA data were automatically annotated with the appropriate metadata. A list of all new metadata can be downloaded with the Supplementary Material [30].

GTEx and TCGA

Tissue and sex annotation for GTEx were extracted from the official sample annotation table as provided by GTEx (GTEx_Data_V6_Annotations_SampleAttributesDS.txt, from https://gtexportal.org/ [31]). An annotation file for TCGA was provided by recount2. For tissue and sex annotation we took columns gdc_cases.project.primary_site and gdc_cases.demographic.gender, respectively. Sample source was assumed to be of type biopsy for all GTEx (n = 9,662) and TCGA (n = 11,284) samples.

SRA

For the SRA samples, we relied on normalized metadata provided by MetaSRA [14]. Available SRA identifiers were downloaded through the GUI on http://metasra.biostat.wisc.edu by searching for all 31 GTEx tissues (site accessed on 9 November 2019). Supplementary Table S1 lists assumed mappings from GTEx tissue names to MetaSRA tissue names where no direct mapping was available. Of the 31 tissues available for GTEx we were able to identify samples for 26 in MetaSRA, resulting in 6,183 annotated SRA samples. Sample identifiers for sex were accessed through the same GUI by searching for male organism and female organism + *Homo sapiens* cell line, which resulted in 3,240 annotated SRA samples. Sample source was determined using the sqlite file provided by MetaSRA (metasra.v1-5.sqlite, [32], column sample_type), resulting in 28,043 annotated samples across 6 sample source categories.

Tissue label harmonization

GTEx, TCGA, and SRA have 17 common tissue types (Supplementary Fig. S3). Bladder was removed owing to its small sample size (GTEx n = 11). We kept samples of comparable size in SRA (adrenal gland n = 14, testis n = 14, pancreas n = 17 in the SRA training data) because the SRA training data are mainly used for bias injection, such that size was not considered an exclusion criterion. This resulted in 5,480, 8,624, and 3,252 tissue annotated samples across 16 tissues for GTEx, TCGA, and SRA, respectively (Supplementary Tables S2 and S3).

Dimensionality reduction and normalization

The downloaded gene count table provided counts for 58,037 genes (Gencode v25, GRCh38, 07.2016). First standard $\log_2$ transcript per million (TPM) normalization was applied to normalize for gene length (Gencode v25, GRCh38, 07.2016) and library size. We next reduced the number of input features (genes), aiming to keep features that contain information and removing potentially uninformative features. First, all non-protein-coding genes were removed, reducing the gene set by 65.5% to 19,950 genes. For sex classification, only protein-coding genes on the X and Y chromosome (n = 913) were selected. For retaining only genes that show significant dispersion across tissues, we computed the Gini coefficient [15, 33, 34] for all remaining genes across all GTEx samples. Housekeeping genes, for example, are known to be expressed similarly across tissues and would score a low Gini coefficient (i.e., high dispersion). Low and high cut-offs were applied during hyperparameter optimization. For tissue classification, genes with Gini coefficients g between 0.5 and 1 were retained, resulting in a feature space of dimension $d = 6{,}974$. For sex classification, genes with $0.4 < g < 0.7$ were used ($d = 190$). Sample source classification included genes with $0.3 < g < 0.8$ (d = 8,679) (Supplementary Table S2, list of input features in Supplementary Material [30]).

Dataset preparation

Phenotype classification experiments

Tissue: To ensure that dataset biases are not shared between training and test sets, SRA data were always split on the study level. For tissue of origin prediction, the 2 largest SRA studies per class were put in the training set. This ensured maximal bias variability in the remaining test data, ensuring a realistic test score. Of the 178 SRA studies containing tissue annotated samples, 30 studies were selected for the training set (n = 1,721) and 148 studies for the test set (n = 1,531) (Supplementary Tables S2 and S3).

Sex: We noticed SRA samples identified as female by MetaSRA to have a significant amount of reads mapped to chrY (Supplementary Fig. S4). All samples labeled as female with a total normalized count ≥2 and all samples labeled male with a total normalized count <2 were removed. In total, 149 SRA studies contained samples annotated with male and or female by MetaSRA. These studies were combined into the training set (studies = 73, n = 2,017) and test set (studies = 76, n = 791) (Supplementary Tables S2 and S3). For model validation, GTEx was randomly split into training and test sets with an 80:20 ratio for both sex and tissue classification.

Sample Source: A confidence cut-off of ≥0.7 was applied (provided by MetaSRA), reducing the total amount of annotated samples for SRA from 23,651 to 17,343. MetaSRA provided 6 different types of sample source. The 2 largest classes, TISSUE and CELL LINE, were selected. In this study we renamed the MetaSRA label TISSUE to biopsy to avoid confusion with the phenotype tissue (e.g., heart, lung, skin). For each of the 2 selected categories we sorted all available studies by number of samples, placed the first third of studies into the training (studies = 420, n = 12,725), the second third into the test (studies = 422, n = 3,144), and the last third into the SRA validation set (studies = 418, n = 1,124) (Supplementary Tables S2 and S3). A list of the sample IDs and corresponding labels is available in the Supplementary Material [30].

Metadata annotation

After determining the best model for each phenotype, we retrained the models for automated metadata annotation. The same datasets as defined above were used for the sex metadata annotation. Tissue: We followed the same pipeline as described above. Samples from a tissue class other than the original 16 classes were pooled together into a “catch-all” class, resulting in 17 classes. In total, 44 SRA studies were selected for the training set (n = 3,370) and 203 studies for the test set (n = 2,813). Sample Source: Contrary to before, for metadata annotation we used all available classes in the SRA data source. All classes that are not of type biopsy were grouped into a single “catch-all” class, while the same cut-off as before was applied. The training set (n = 16,463) is made up of 974 SRA studies and the test set (n = 3,707) of 492 studies.

Multilayer perceptron

MLPs use fully connected neural network layers to learn nonlinear features from a raw input space [35] and constitute the most basic form of ANNs. All our ANN-based models were developed and trained on tf.keras (Tensorflow 2.1). The hyperparameters for each prediction task were determined using exhaustive iterative random search (keras tuner 1.0.1) (Supplementary Table S4). In case of approximately equal accuracy on the validation set, the least complex model was chosen. A single hidden layer was used in each case with 128, 128, and 32 nodes for tissue, sample source, and sex prediction, respectively (Supplementary Table S5, Supplementary Fig. S5). Each network was trained for 10 epochs with a batch size of 64. Performance was quantified by mean sample accuracy and mean class accuracy and subsequently used to benchmark our DA approach.

Domain adaptation model

Many DA models correct bias between 2 domains, a source and a target domain. In biological research, however, one is often confronted with many small datasets, each potentially with its unique dataset bias. Therefore, we specifically designed our DA model to be able to learn from very few data by using a Siamese network architecture [36]. The Siamese network learns bias from pairs or triplets of training samples by exposing each sample in multiple relationships to the model. We distinguished 3 different types of input data for our model. The source domain is a large single-bias dataset used to learn the feature embedding for the classification task (in our case: GTEx). The bias domain contains labeled samples from multiple smaller datasets (in our case: SRA) each with its own bias. The target domain refers to unlabeled and biased datasets that we want to classify (unlabeled SRA or TCGA data).

Model architecture

Our DA architecture is based on the Siamese network architecture. A Siamese network usually shares the weights between 2 equal networks. Here, however, we do not use weight sharing. Weight sharing and other types of architecture did not prove to be applicable to this problem (see Methods section Other Models). It consists of 3 modules: A source mapper (SM) and bias mapper (BM), which correspond to the Siamese part of the model, as well as a classification layer (CL). These modules give rise to 3 different configurations, i.e., 2 training cycles and a prediction configuration (see Supplementary Fig. S6 for a brief illustration). In the first training cycle, the SM and the CL are combined to form an MLP (Fig. 2A). The task of the SM is to learn a mapping from the input space to an embedding space from which the CL can predict phenotype classes. The SM-CL module is trained with a batch size of 64 for 10 epochs. Because the SM-CL MLP is trained on a large single-bias dataset, it will likely overfit and thus not readily generalize to other datasets (Fig. 2B). For the second training cycle, the SM and the CL are separated and their weights frozen. Frozen weights are not updated during the second training cycle. The bias mapper is created by copying the architecture and weights of the trained source mapper. SM and BM are trained on triplets drawn from the source and the bias domain (Fig. 2C). Samples from the source domain are passed through the SM, and samples from the bias domain through the BM at the same time. Each triplet is made up of an anchor (*a*) sampled from the bias domain, and a positive (*p*) and a negative sample (*n*) from the source domain. The anchor and the positive sample have equal class labels, whereas the negative sample is from a randomly selected different class. The triplet loss function [37] was used to optimize the model during training:

$$\mathcal{L} = \max(d(a, p) - d(a, n) + m, 0),$$

where $d(i, j)$ are the distances between the constricted embedding space of the SM and the bias mapping into that space of the BM on samples i and j. For improved training time and robustness, our model is trained on semi-hard triplets [37]

$$d(a, p) < d(a, n) < d(a, p) + m,$$

with a margin parameter m. The distances are defined as Euclidean distances in embedding space:

$$d(a, p) = \|\sigma(\mathrm{BM}(a)) - \sigma(\mathrm{SM}(p))\|$$
$$d(a, n) = \|\sigma(\mathrm{BM}(a)) - \sigma(\mathrm{SM}(n))\|$$

σ is the sigmoid activation function for the embedding vector. Triplets are mined online, meaning that they are newly generated for each batch [37]. The SM-BM module was trained for 10 epochs with a batch size of 64. Hyperparameters were determined as described above (Supplementary Table S5, Supplementary Fig. S5). As this training cycle proceeds, the BM learns to map its output onto the SM embedding space. After training, the bias mapper and the classification layer are combined to a BM-CL MLP and can be used for prediction of the target domain (Fig. 2D). The source code as well as an example are available at the project Git repository [38].

Linear regression model

We used the metadata prediction performance of the LIN model described in Ellis et al. [20] as a point of reference. The LIN model was optimized on the same data as all other models (see Data section of Methods). For each experimental set-up, the following steps were conducted in R version 3.6.3 in order to build the corresponding phenotype predictor and evaluate its accuracy based on the test data:

(1) calculating the coverage matrix for the training samples based on the regions reported in Ellis et al. [20] by using the function “coverage_matrix_bwtool” (R package recount.bwtool version 0.99.31).
(2) building the model by running “filter_regions” and “build_predictor” (R package phenopredict version 0.99.0) with the same parameters used in Ellis et al. [20]
(3) testing the model on the test samples with “extract_data,” “predict_pheno,” “test_predictor” (R package phenopredict version 0.99.0)

Notably, our experiments differ from the original work [20] solely by applying additional pre-processing steps to the samples (see Methods), which may be responsible for observed small differences in performance. For implementation details and code examples for the aforementioned functions, see the documentation [39].

Nomenclature of experiments

Each experiment was named after the model, the training, and the test data used. The possible models are LIN (linear model [20]), MLP (multi-layer perceptron), and DA (novel DA approach). The data sources are named G (GTEx), T (TCGA), and S (SRA). If only the SRA training data are used (i.e., if the model is evaluated on the SRA test data), we write $\mathrm{S}_{\mathrm{small}}$. If the SRA train and test sets are combined for training, we write $\mathrm{S}_{\mathrm{large}}$. For instance, an experiment using an MLP, trained on a mix of GTEx and SRA and evaluating on SRA data, would be named MLP G+$\mathrm{S}_{\mathrm{small}}$-S.

Impact of data diversity and quantity on model performance

To analyze the effect of training data diversity on prediction accuracy, the following experiments were designed. First, MLP S-S models for sample source prediction were trained with an increasing number of unique SRA studies in the training data, systematically increasing bias diversity. Only SRA studies containing >100 samples for either class were considered. To control for training set size, each SRA study was subsampled to 50 samples before training. Six iterations of this training process were conducted, starting with 1 study (i.e., 1 bias) per class (biopsy vs cell line). At each step 1 additional SRA study per class was subsampled, ending with 6 SRA biases and 350 samples in the training set per class. As a control experiment, we chose the largest SRA study available for each class to create a training set with a single bias per class. Starting with 50 samples per class in 6 iterations, we subsampled an additional 50 samples, ending with 350 samples, thereby assessing the effect on performance that can be attributed to the dataset size. Subsampling and random selection of SRA studies were repeated 10 times with different seeds, and each configuration was trained on 10 different seeds, yielding an estimate of uncertainty.

Test for overfitting

We have identified mislabeled samples for the sex phenotype (see Methods). The following experiment was designed to test the ANN-based model’s susceptibility to overfitting on mislabeled training data. An MLP model was trained on GTEx data on 4 tissue classes (i.e., brain, esophagus, lung, and skin). A range of

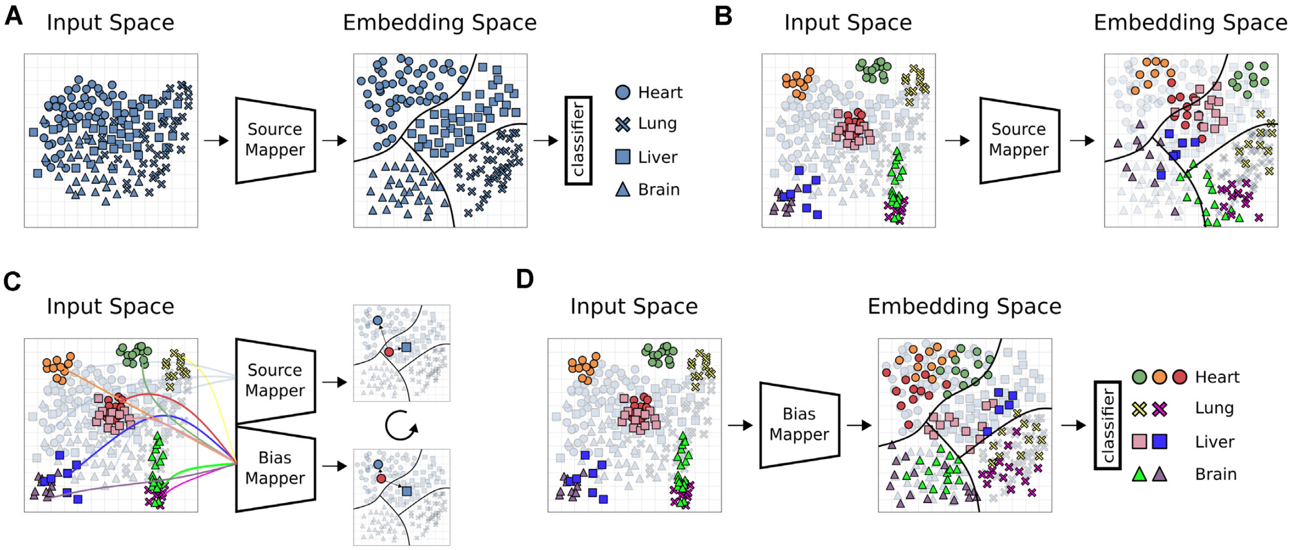

Figure 2: Overview of domain adaptation model. Illustration of our DA model architecture and training. Shapes of (hypothetical) data points represent classes; colors are datasets with unique biases. Source mapper (SM), bias mapper (BM), and classifier layer (CL) are ANN modules. (A) First training cycle: The SM is trained on a single bias dataset, the source domain (SD). In this step, the SM learns a feature embedding. The CL learns how to partition this embedding space into classifiable regions and draws decision boundaries (black lines). (B) For biased test data (colored sample data points), the same classes may occupy distinct regions in input space. In this case, the source mapper may not be able to map the samples to the correct region of embedding space, compromising classification performance of the CL. (C) To learn the mapping of different biases to the embedding learned in A, a BM is created by copying the SM, and trained weights of the SM are fixed. In this second training cycle, triplets of samples are passed through the SM-BM configuration, consisting of an anchor from the bias domain and 2 samples from the source domain, 1 of them with a matching label. The triplet loss function is defined to minimize distance of like labels in embedding space and to maximize distance of opposite labels. This process is repeated until the SM has learned to map all known biases into the previously learned embedding space. (D) The BM is now able to map data points from previously unseen datasets into the embedding space, where the CL can classify them.

fractions of the brain samples were randomly assigned to skin tissue (i.e., 0.01, 0.025, 0.05, 0.1, 0.20, 0.5, and 0.8). The model was then trained on GTEx samples of the 4 classes, including the mislabeled brain samples. We tested the model's overfitting capabilities by letting it predict the label of the mislabelled brain samples. If the model overfits, these samples should be predicted to be from skin tissue. The same experiment was conducted for the sex phenotype by mislabeling male samples as female.

Metrics

We report micro and macro accuracy, which are equivalent to mean sample accuracy (msa) and mean class accuracy (mca), respectively. Sample accuracy is a measure of absolute performance on the test data. It reports the fraction of correctly classified samples over all classes:

$$\text{msa} = \frac{\sum_i^N \mathbb{1}_{y_i}(\hat{y}_i)}{N},$$

where N is the number of samples, y the true label, and $\hat{y}$ the predicted label, and $\mathbb{1}$ is the indicator function. Given the large class imbalance in some of our experiments, an increase in accuracy in a small class will not be captured by this metric. Average class accuracy, on the other hand, reports the average sample accuracy per class, weighing each class equally and thereby capturing local improvements of the models:

$$\text{mca} = \frac{\sum_{j=1}^C \frac{1}{M_j} \sum_{i=1}^{M_j} \mathbb{1}_{y_{ij}}(\hat{y}_{ij})}{C}.$$

Here, C is the number of classes, M_j is the number of samples for class j, and y_{ij} and $\hat{y}_{ij}$ are the true and predicted values, and $\mathbb{1}$ is the indicator function.

Statistical tests

Accuracy distributions were tested for significance using the non-parametric Mann-Whitney *U* test (scipy.stats.mannwhitneyu v 1.3.1).

Other models

While developing our DA model, we did a thorough literature search and implemented and tested multiple architectures and strategies. Here, we give a brief overview of the models that we found not suitable for the problem of bias-invariant RNA-seq metadata annotation. The first strategy that was tested was interpolation between source and target domain by training feature extractors on an increasing ratio of target to source domain data [40]. The second strategy was adversarial training by applying 2 loss functions. The first loss function forces the model to learn weights for the class prediction task, while the second forces the model to learn to ignore differences between the source and target domain [41]. We also implemented the adaptation of this idea by Tzeng et al. [42], proposing a model using a separate source and target encoder, using them as "real" and generator input for a generative adversarial network [43] that is capable of ignoring bias. These models ultimately failed owing to the hundreds of dataset biases in the SRA data and their relatively small sample size (data not shown). For the case of scarce target data, an approach was previously proposed using Siamese networks [36, 44]. The trained model achieved msa of 0.83 and mca of 0.79 for tissue classification on SRA data. The mca achieved is comparable to the results of the MLP model, however, the msa score is 6% lower than even the LIN model. The more challenging task of learning to map the bias embedding into the pre-learned class embedding, as presented in this article, finally resulted in the desired outcome.

Results

Experimental set-up

This study aims to find the best model for RNA-seq metadata annotation based on gene expression. Three different data sources were selected for which phenotype data were available (Fig. 1A). Each of the 3 data sources comes with a different number of dataset biases. Briefly, GTEx is a large homogeneous dataset containing healthy samples following a strict centralized standard protocol. TCGA contains pooled samples from different cancers, disease stages, and sequencing centers. Our SRA data comprise hundreds of individual studies following no centralized standard, resulting in the largest number of biases of all 3 data sources. Bias in a test dataset that has not been learned by a model can severely compromise performance. We hypothesized that exposing classification models to a sufficient number of dataset biases will enable them to learn a generalized internal feature representation. Such a model would be able to classify data with previously unseen biases. To test and benchmark our models we selected the classification tasks of (i) tissue of origin of a given RNA-seq sample, (ii) biopsy vs cell line origin of a sample (i.e., sample source), and (iii) sample sex (Fig. 1A).

Three different ML models were compared (Fig. 1B). First, a fully connected ANN (MLP) was tested because of its capability to create novel latent features (see Methods for model details). Second, we developed a DA approach (Fig. 2), a subfield of ML dealing with dataset biases. Last, the LIN model trained on GTEx data, proposed in Ellis et al. [20], was used as the baseline for all tissue and sex classification experiments.

Models were trained on either GTEx or a mix of GTEx and SRA data and tested on TCGA and SRA data. Uncertainties for MLP and DA models were estimated from 10 training runs with different random seeds (Fig. 1B).

Domain adaptation outperforms other models on tissue classification

We first tested the performance of the LIN, MLP, and DA algorithms to predict the tissue of origin on GTEx ($n = 5{,}480$), TCGA ($n = 8{,}624$), and SRA (train $n = 1{,}721$, test $n = 1{,}531$) datasets. A subset of 16 tissue labels was chosen that is common to all 3 data sources (see Methods, Supplementary Fig. S3, Supplementary Table S3). First, we conducted a single-bias experiment, i.e., MLP G-G (see Nomenclature of Experiments in Methods). The nearly perfect score of msa 0.996 and mca 0.99 (data not shown) confirmed that the MLP yielded highly accurate results when trained and tested on a single-bias dataset (for details on model training, validation, and testing see Methods).

Prediction of SRA tissue

Metadata prediction on SRA was the most challenging and interesting task owing to the potentially large number of different biases in the data source. We retrained and tested LIN G-S on our datasets and achieved msa of 0.893 and mca of 0.765 for the 16 tissues (Fig. 3A). Of note is the significantly higher accuracy achieved with LIN G-S compared to that reported by Ellis et al. [20] (0.519 msa). MLP G-S (msa: 0.872, mca: 0.77) had a higher mca but a lower msa than the corresponding LIN model (Fig. 3A). In the next step we investigated the effect of adding bias to the training dataset on prediction performance. In particular, we first predicted SRA tissue from S_{small} data. MLP S_{small}-S (msa: 0.894, mca: 0.746) matched the base model's msa score but performed slightly worse using the mca metric. Similarly, the LIN S_{small}-S model matched the msa of LIN G-S but showed an increased performance for mca (msa: 0.893, mca: 0.795). Notably, by only using the small SRA training dataset, we lose the advantage of the large sample size of GTEx. Based on this, we hypothesized that by combining SRA and GTEx in the training data, we may be able to leverage both sample size and diversity.

The LIN G+S_{small}-S model increased its msa to 0.908 and mca to 0.785, which in turn is 1% lower than the LIN S_{small}-S model. The 2 best performing models were MLP G+S_{small}-S and DA G+S_{small}-S, outperforming LIN G-S on msa by 2.5% and mca 5.5% (MLP G+S_{small}-S msa: 0.915, mca: 0.817 and DA G+S_{small}-S msa: 0.922, mca: 0.821). No significant difference in the mean performance was detected between these 2 models (msa $P > 0.02$, mca $P > 0.4$, Mann-Whitney). Crucially, however, DA G+S_{small}-S exhibited the lowest standard deviation (std = 0.003 for msa and std = 0.009 for mca) of all models tested (Supplementary Table S6). For this reason, DA G+S_{small}-S was considered the best model for the prediction of tissue on the highly heterogeneous SRA test data. The best model increased the msa score by 3.6% compared to LIN G+S_{small}-S and mca by 5.6% compared to the baseline LIN S_{small}-S, the best performing linear models for the respective metrics.

Prediction of TCGA tissue

Next, model performance on TCGA data was assessed (Fig. 3B). The baseline model LIN G-T achieved msa 0.718 and mca 0.638. Applying the MLP model on the same data resulted in a decrease in msa and mca of 2.4% and 3.3%, respectively (MLP G-T msa: 0.684, mca: 0.605). For TCGA tissue prediction, we used S_{large} for training, essentially doubling the SRA training data (SRA train + SRA test set: $n = 3{,}252$). LIN S_{large}-T improved accuracy by 6.6% for msa and 8.6% for mca to 0.784 and 0.724, respectively. In comparison, MLP S_{large}-T increased model performance by 11.4% to 0.832 (by 11.7% to 0.755) for msa (mca) with respect to LIN G-T. Combining GTEx and SRA training data reduced LIN G+S_{large} performance to msa 0.725 and mca 0.651. The best accuracy was achieved by our MLP G+S_{large} (msa: 0.842, mca: 0.773) and DA G+S_{large} (msa: 0.875, mca: 0.813) models. The DA model had thus a 15.7% and 9.1% performance increase for msa compared to LIN G-T and LIN S_{large}-T, respectively. In addition to being the top performer, DA G+S_{large}-T also was the most robust model for this task, having the lowest variation in its results (std = 0.004 for msa and std = 0.006 for mca) (Supplementary Table S6).

We repeated the prediction for TCGA with the models trained for SRA tissue prediction (previous section), i.e., on S_{small}, which allows us to assess the influence of the amount of bias injection on model performance. Whereas the addition of more SRA data to the training dataset had little influence on LIN models (except for a slight increase of $\sim$0.2% for G-S_{large}-T), both MLP and DA model accuracies improved significantly (by between 5% and 9%) upon addition of additional SRA data (Supplementary Table S6).

Notably, adding 5,480 GTEx training samples to MLP S_{small} (MLP-S_{small} $\longrightarrow$ MLP G+S_{small} increased msa from 0.748 to 0.764 and msa from 0.688 to 0.716 on the TCGA test set. On the other hand, adding 1,531 SRA samples (MLP-S_{small} $\longrightarrow$ MLP S_{large} increased msa to 0.832 and msa to 0.755, underlining our model's ability to incorporate multiple biases for better generalization (Supplementary Table S6).

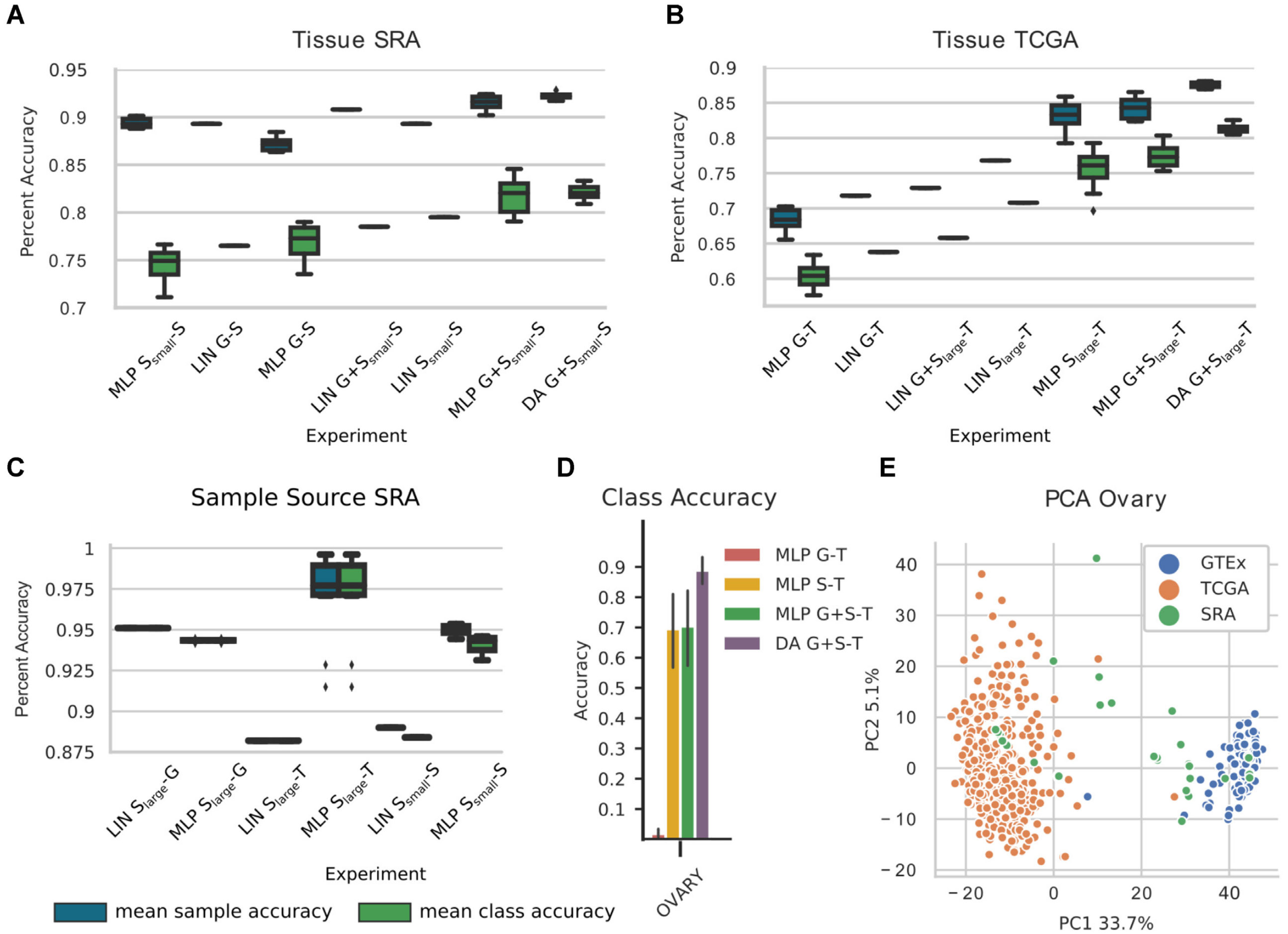

Figure 3: Phenotype prediction results for (A, B) prediction of tissue of origin on SRA (S) and TCGA (T) (16 classes) and (C) prediction of sample source on SRA (2 classes). Indices "small" and "large" refer to the different size of SRA training data used due to splits of the dataset in SRA prediction. Box plots represent model uncertainty of ANN-based models, estimated from training with different random seeds (n = 10) and show the the minimum, the maximum, the sample median, and the first and third quartiles. Mean sample accuracy and mean class accuracy were calculated for each seed. For (A–C) LIN G-X is the baseline proposed in [20]. (D) Accuracy of each ANN model predicting ovary tissue on the TCGA data source, error bars show plus/minus one standard deviation, and (E) a principal component analysis (PCA) plot of the gene expression values for the ovary tissue samples. A domain shift (i.e., bias) is clearly visible between GTEx (G) and TCGA, leading to the poor performance of multilayer perceptron (MLP) G-T on ovary. SRA data in the training set help to establish a good accuracy. DA: domain adaptation; LIN: linear regression.

Expression-based prediction of sample source

SRA data stem from multiple different sources, from which we selected the 2 largest, namely, either biopsy or (immortalized) cell lines, whereas GTEx and TCGA data are exclusively from biopsies. Starting from the hypothesis that fundamental differences do show on an expression level, we set out to train LIN and MLP models on SRA data to predict the sample source of SRA, GTEx, and TCGA. Of note, while we were able to approximately reproduce the original results for LIN S_{small}-G and LIN S_{small}-S, we were not able to do so for LIN S_{small}-T (msa: 0.998 reported in [20]). LIN S_{large}-G (msa/mca 0.951) did slightly better than MLP S_{large}-G (msa and mca of 0.943). MLP S_{large}-T achieved msa and mca 0.971, outperforming LIN S_{large}-T (msa and mca of 0.882). MLP S_{small}-S achieved msa 0.95 and mca 0.941, outperforming LIN S_{small}-S with msa 0.89 and mca of 0.884 (Fig. 3C).

Multi-bias data enhance tissue classification on TCGA

For tissue classification on TCGA, mean class accuracy increased by 16.8% between MLP G-T and MLP G+S_{large}-T. This confirms our hypothesis that the homogeneity of the GTEx data did not allow the MLP G-T model to generalize to TCGA data, while the addition of SRA training data in MLP G+S_{large}-T resulted in a model with significantly improved generalization. To further investigate this result, we took a closer look at the per class accuracy for the TCGA tissue prediction (Fig. 3D, Supplementary Fig. S7). MLP G-T was unable to predict samples for 3 tissues, namely, bone marrow (msa: 0.08), ovary (msa: 0.02), and uterus (msa: 0.07), whereas all our other models achieved accuracies between 0.7 and 1.0 on these tissues. Adding SRA data to the training set enabled the model to achieve per tissue sample accuracy of 1.00, 0.704, and 0.67 for bone marrow, ovary, and uterus, respectively. We used principal component analysis (PCA) to visualize the dataset bias for ovary tissue (Fig. 3E). Interestingly, the GTEx-ovary and TCGA-ovary data points show little overlap in the PCA plot, while the SRA-ovary data overlap with GTEx- as well as TCGA-ovary data, forming a "bridge."

Linear model sufficient for sex classification

For sex classification, only genes on the X and Y chromosome were used as input features ($d = 190$). We first tested the trivial case MLP G-G by splitting GTEx into training and test sets, achieving sample and class accuracy of 0.995 (data not shown).

Prediction of TCGA sex

Sex phenotype prediction on TCGA data was the only task where the linear model outperformed the ANN models. The baseline LIN G-T, as well as the other linear models LIN S_{large}-T and LIN G+S_{large}-T, achieved almost perfect accuracy on the TCGA data (msa/mca 0.989 for LIN G-T and LIN G+S_{large}-T, msa 0.988 and mca 0.987 for LIN S_{large}-T). Our best model, based on the data annotation provided by MetaSRA, was MLP G+S_{large}-T with msa 0.964 and mca 0.962 (Supplementary Fig. S8).

Prediction of SRA sex

All linear models for the prediction of sex for SRA data achieved high accuracy (msa: 0.98 and mca: 0.98 for LIN G-S and LIN G+S_{small}-S, msa: 0.979 and mca: 0.979 for LIN S_{small}-S). This result is significantly better than what was previously reported (msa: 0.863 [20]). The MLP G-S model (msa: 0.971 and mca: 0.979) did, on average, perform worse than all the linear models. While adding SRA data to the training set did not improve the LIN model, it increased the performance of the MLP and DA models, DA G+S_{small}-S (msa: 0.99 and mca: 0.987), MLP S_{small}-S (msa: 0.994 and mca: 0.994), and MLP G+S_{small}-S (msa: 0.993 and mca: 0.992). Results are shown in Supplementary Fig. S8.

According to MetaSRA all our training and testing data for sex prediction on SRA stem from patient biopsies. However, ≥2 of the largest misclassified SRA studies in the test set are clearly cultured cell lines. For example, SRP056612 is a study on the effect of the MERS coronavirus on cultured kidney and lung cells [45] and SRP045611 is a study involving HEK cells, which lack the Y chromosome but are annotated as male by MetaSRA [46]. These are 2 examples of errors in the MetaSRA. Clearly, mislabeled data can compromise classifier accuracy, either by providing the wrong ground truth for training or by reporting the false label at the point of prediction. As described in the Methods section, obviously mislabeled samples have been removed.

Training data diversity outweighs quantity

Our experiments on phenotype classification seem to indicate that increased training data diversity might enhance classification performance. To learn more about the relationship between the amount of training data and model performance, MLP G-S was trained on an increasingly large subset of the GTEx training data for tissue classification. We observed a limited effect on model performance with increased training dataset size. The msa reaches its peak with one-third of the available training data, while the mca saturates at approximately half of the available training data (Supplementary Fig. S9).

To test the effect of bias in the training data, an MLP S_{small}-S for sample source classification was trained on an increasing number of biases in the training set. As a control experiment, an MLP was trained with the same amount of data but drawn from a single-bias source. We observed a positive correlation between msa and the number of biases in the training set (Fig. 4A). Contrary to that, increasing the number of training samples by the same amount but from a single-bias source did not lead to better model performance (Fig. 4B), validating our assumptions. Both experiments support our assumption that ANN-based models can integrate different biases in the training set and translate them into better model performance compared to other methods.

ANN models can correct mislabeling in MetaSRA

Given the difficulties with metadata standards in SRA data, mislabeling in MetaSRA is to be expected. To understand whether and when ANN models would overfit on mislabeled MetaSRA data, we trained an MLP on partially mislabeled samples (see Methods). Supplementary Fig. S10 shows that the MLP model correctly predicts brain samples, even if they were presented as skin samples during model training. A decrease in this accuracy was observed if >20% of all brain samples were mislabeled as skin. A similar observation was made for the sex phenotype (Supplementary Fig. S10). We concluded that our models are robust if <20% mislabeled data are present during training. More importantly, these models can be used to correct mislabeled MetaSRA data.

In the specific case of sex classification, the MLP G+S was used to predict the true corrected label for the SEX samples that were removed from training due to low sex-chromosome counts (see Methods). For 82% of the 132 filtered samples, the MLP model predicted the opposite of the presumably wrong MetaSRA labels. However, our MLP model was able to confirm the MetaSRA label for 24 samples. These samples had a mean chrY count sum of 2.4 (i.e., close to the cut-off value). Manual confirmation revealed a high model accuracy. For example, SRR1164833, SRR1164787, and SRR1164842 are samples from a prostate cancer study labeled as male by MetaSRA. Our MLP model correctly classified these samples despite the fact that their chrY total sum count was between 0.4 and 1.4. On the other hand, SRR16076 54/56/61/62/64/65/70/71 are annotated as female by MetaSRA and the MLP but had a chrY total sum count of 2–5.3. We see the correct classification of these borderline cases as further evidence that no overfitting is taking place.

A list of all SRA samples for which the MetaSRA labels and the predicted labels mismatched is available in the Supplementary Material [30].

Prediction and availability of novel metadata

We have used our best models to predict high-quality metadata for published SRA samples lacking information on tissue, sex, or sample source. Prediction of sex is straightforward because our models were trained on all possible biological categories. For tissue and sample source, however, our models were trained on a subset of all potential classes in the unlabeled data. If, for example, we try to label a sample of a tissue type unknown by the model, the model will force 1 of the learned classes onto that sample. To deal with this in the best possible way for sample source classification, we modified the classification task into one vs all. Specifically, we first trained a new MLP model to identify the sample source biopsy vs all other sample sources available in the SRA data as defined by MetaSRA. This model (i.e., MLP S_{small}-S) achieved msa 0.947 and mca 0.93 on a test set (data not shown) and MLP S_{large} was subsequently used to identify all of our yet unannotated SRA samples of source type biopsy. At a probability cut-off of 0.5 we identified 1,072 new SRA samples as originating from a biopsy.

Second, we extended the tissue classification task to 17 classes by adding a “catch-all” class. To this end, we extended the training data to all GTEx (n = 9,366) and SRA (n = 6,183) data with tissue labels and assigned the placeholder class for every sample that did not belong to the original set of 16 tissues. That way, we ensure that the learned model will not force known classes on every tissue type. With this approach, the

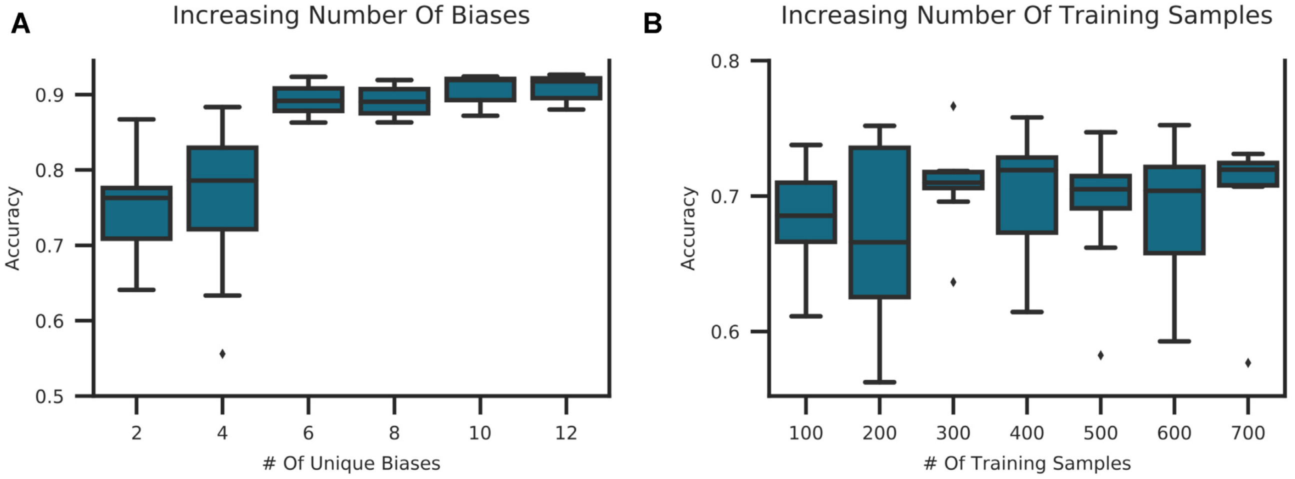

Figure 4: Increasing bias vs increasing sample size in training data. (A) An MLP S_{small} for sample source prediction on SRA data was trained by randomly sampling an increasing number of SRA studies per class. Each study was subsampled to 50 samples. Studies were drawn from all SRA studies with $n > 100$ for either sample source tissue or cell line. (B) To differentiate the effect of increased bias vs increased sample size, the same model was trained by randomly subsampling the largest available SRA study per class. At each step an additional 50 samples were added to the training set per class. Models were run with 10 different seeds and the mean sample accuracy was computed. Box plots are produced by 10 random sampling iterations and show the the minimum, the maximum, the sample median, and the first and third quartiles. We observe a positive correlation between training data diversity and accuracy.

DA G+S_{small} model achieved msa 0.912 and mca 0.787 (data not shown). Training and test datasets were subsequently combined to train DA G+S_{large} for annotation prediction of unlabeled SRA samples. We predicted the tissue of origin for all SRA samples of source type biopsy for which no entry on MetaSRA was available ($n = 2,818$).

Third, 8,495 SRA biopsy samples with missing sex information were predicted using MLP G+S_{large}. Supplementary Fig. S11 shows the true-positive rate for each phenotype and each class on the test set. We provide this information such that users can make their own decision on probability cut-offs applied to each class. We provide the full list of all classified SRA samples as well as the probability output of the classifier in the Supplementary Material [30].

Discussion

We developed a novel deep-learning–based DA approach for automated bias-invariant metadata annotation. To our knowledge this is the first time that DA has been applied to this problem. We were able to outperform the current best model [20] on tissue prediction by 2.9% for SRA and 15.7% for TCGA data on mean sample accuracy. We can confirm, as was previously reported [17], that ANNs trained on single-bias training data do not perform better than linear models. Given multi-bias training data, however, we showed that MLPs, and especially our DA algorithm, have an advantage over standard ML approaches (e.g., linear regression). Our current models help researchers to verify the sex, tissue, and sample type of RNA-seq samples in the presence of bias. This metadata information is currently rarely given for datasets downloaded from the SRA but can be of crucial importance.

The main strength of our method is its ability to incorporate dataset bias from datasets with only a few samples by applying a Siamese network-like architecture. The model learns to ignore bias by repeated exposure to (a few) samples in (many) different contexts, i.e., as triplets. In addition, it does not rely on feature selection but uses normalized gene count tables and lets the network learn which features carry important information.

Different types of experiments showed the importance of training models on a multi-bias dataset. First, we showed for every phenotype classification that models that had SRA samples included in the training data performed better than models trained only on GTEx data. For tissue classification, we further showed that the effect of adding SRA samples to the training data outweighs adding 3.2 times as much GTEx data (MLP S_{small} → MLP S_{large} vs MLP S_{small} → MLP G-S_{small}). Second, for SRA tissue classification, we showed that there is a limit of accuracy that can be achieved irrespective of the size of the training set. Our experiment showed that peak accuracy is already reached by using 50% of the available data. Last, for sample source classification, we directly compared the relationship between the number of biases in the training data, the number of samples, and the model performance. We found a positive correlation between the diversity of the training data and the accuracy achieved by that model.

A major concern with our experiments is the potential misclassification in the MetaSRA-annotated ground truth. The MetaSRA pipeline serves mainly as a normalizer for already existing metadata and is therefore susceptible to human error. Systematic annotation errors create signals in the training data that a model can learn and then replicate on the test set. We approximated a systematic error by randomly mislabeling training data from a single class. We showed that our models are robust to overfitting if ≤20% of the training samples per class are mislabeled. Our models are able to predict the correct class of a sample, even if the sample was mislabeled during model training. This property of our models was exploited for the correction of wrongly annotated metadata in the MetaSRA and made publicly available.

Last, we generated novel metadata for SRA samples using our best performing models, adding >10,000 new metadata entries for 8,495 SRA samples. The newly generated metadata are now publicly available and can be used for future research. We see this as a first and important step in the general direction of an effort to make publicly available data more accessible and reusable in an automated way.

We observed some limitations to our DA approach. Our experiments showed that the DA model does not perform as well as the MLP for classification tasks with a low number of classes (e.g., sex). At least for the TCGA tissue classification, it seems that a minimum of roughly 8 classes is needed for the DA model to be able to unfold its full potential consistently. Our experiments indicate that the difference between DA and MLP performance will keep increasing, in favor of the DA model, the more classes we add (Supplementary Fig. S12). Adding more tissue classes to our model is an important next step. Another limitation is posed by the need for labeled data to train the bias mapper.

Whereas currently the scope of our predictive models has been limited by the availability of data (e.g., intersecting tissue types between datasets, limited size of datasets), the approach is ready to incorporate more data, biases, classes, and more phenotypes, and there is reason to believe that this will confer increased performance of ANN-based models, in particular DA models. At the same time, automated annotation ensures that the vast amount of data currently lying idle in online repositories and institutional data centers can indeed be leveraged. We believe that this synergy is capable of producing a large and comprehensive body of annotated biological data that will boost knowledge discovery for biomedical research.

Availability of Supporting Source Code and Requirements

Project name: Bias invariant RNA-seq metadata annotation
Project home page: https://github.com/imsb-uke/rna_augment
Operating system: Platform independent
Programming language: Python
Other requirements: TensorFlow 2.1
License: MIT

Data Availability

The input data as described in Supplementary Table S2, as well as a copy of the Git repository, are available in the GigaScience Database [30].

Additional Files

Supplementary Figure S1: Visualizing dataset bias. GTEx is a single-study data source, while SRA is a multi-study data source. (A) t-SNE plot of gene expression values of GTEx and (B) SRA samples, belonging to 5 different tissues. The GTEx data are more coherently clustered compared to the SRA data. The individual studies in the SRA data appear to form less homogeneous clusters, indicating a larger within-variance in the data source.
Supplementary Figure S2: t-SNE on fraction of total gene count per gene type. The fraction of the total log TPM normalized counts per gene type was calculated for all types that can be associated with messenger RNA or small RNA. t-SNE was applied on the resulting vectors of fraction per gene type. Samples with their maximum fraction in a gene type belonging to a small RNA category were labeled orange, else blue. The scatter plot shows that samples labeled as small RNA-seq all cluster together, suggesting a valid approach.
Supplementary Figure S3: Tissue label overlap between GTEx, TCGA, and SRA. GTEx v6 provides samples for 31 tissues and TCGA for 26. MetaSRA provided labels for 26 of the 31 GTEx tissues. This figure depicts the 40 tissues that form the union between the 3 data sources, a black square indicating that a tissue is present in the respective dataset. Seventeen tissues are shared between GTEx, TCGA, and SRA, 16 of which were used for tissue prediction.
Supplementary Figure S4: Misclassification in MetaSRA. Histogram of the total sum of normalized counts mapped to the chrY for GTEx, TCGA, and SRA. Male and female clearly overlap in SRA, indicating mislabeling by MetaSRA.
Supplementary Figure S5: Graphical representation of architectures for ANN-based models. (A) MLP models for tissue, sex, and sample source. (B) The (1) SM-CL MLP, (2) SM-BM Siamese Network, and (3) BM-CL prediction models for tissue and (C) sex. Each rectangle represents a layer in the neural network and is colored according to the type of layer that has been used. BM: bias mapper; CL: classification layer; d: input dimension; n: number of nodes; p: dropout probability; SM: source mapper. B2 and C2 show the SM to have frozen weights.
Supplementary Figure S6: Supplementary. Samples are indicated according to their classes (circles, squares, triangles) and their bias (blue: source domain, other colors: bias domain, target domain). The model is ready for prediction after 2 training steps: (A) A source mapper is trained on single bias data together with a classification layer. (B) A bias mapper is created as a duplicate of the source mapper; the weights of the source mapper are fixed. Triplets are passed through the source mapper and bias mapper configuration to learn a bias mapping. (C) The bias mapper, equipped with a classification layer, can be used to predict data from previously unseen datasets.
Supplementary Figure S7: Per class accuracy for TCGA tissue classification. Mean sample accuracy for each tissue and all ANN-based models is shown. The error bar shows the standard deviation across 10 random seeds. The plot demonstrates the varied tissue classification performance of different tissues. For instance, it seems to be difficult to identify adrenal gland or pancreas with any of the models. In particular, the bad classification performance of MLP G-T for bone marrow, ovary, and uterus is especially noticeable, along with the observation that performance can be salvaged by addition of (biased) SRA data to the training dataset. This highlights the strength of ANN-based models in capturing bias from training data.
Supplementary Figure S8: Sex phenotype results. (A) SRA and (B) TCGA test data. DA: domain adaptation; G: GTEx; LIN: linear model; MLP: multilayer perceptron; S: SRA; T: TCGA. ANN-based models yielded consistently worse results than the baseline model, until newly annotated data were incorporated into the training set.
Supplementary Figure S9: Dependence of prediction performance on increasing training dataset sizes for MLP G-S. MLP models were trained on subsets of the GTEx data for SRA tissue classification on 10 seeds and averaged. At each step, the subset was increased by 250 samples. Box plots from 20 iterations for the msa and mca are shown in blue and green, respectively. Mean sample accuracy reaches its peak with only 25% of the training data, while 50% of the data is sufficient for the mean class accuracy to saturate.
Supplementary Figure S10: Test of overfitting. An MLP model was trained on GTEx data. An increasing fraction of 1 class was assigned a wrong class label (e.g., brain to skin). The model was trained on the partially mislabeled data and the mislabeled data were predicted by the model after training. We quantify the model's susceptibility to overfitting by letting it correct the mislabeled training data. The MLP model was able to correct all mis-

labeled data up to a mislabeling fraction of 20%. We conclude that the ANN models are very robust in dealing with mislabeled data.

Supplementary Figure S11: True-positive rate for test data predicted with annotation models. (A) Sample source, (B) sex, and (C) tissue classification.

Supplementary Figure S12: Relationship between number of classes and DA performance in DA G+S-T. The 16 tissues were sorted by sample size in GTEx, and at each step 1 tissue was added to the classification problem, starting with the largest 2. MLP and DA were trained as described above for 10 seeds each and tested on TCGA data. The mean sample accuracy for each seed (top panel) or mean class accuracy (bottom panel) is shown. Each dot shows the difference in accuracy (DA-MLP) at each step for each seed. Seaborn regplot was used to fit a regression line. While, on average, MLP performs better for lower number of classes, the performance gain by the DA model with respect to MLP increases with the number of classes.

Supplementary Table S1: Mapping from GTEx tissue names to MetaSRA tissue names.

Supplementary Table S2: Summary of the datasets used for each phenotype after pre-processing.

Supplementary Table S3: Number of samples per class for phenotype classification experiments.

Supplementary Table S4: Hyperparameters considered during model tuning and their initial range.

Supplementary Table S5: Summary of the hyperparameters used for each model.

Supplementary Table S6: Sample and class accuracy given are the mean over n = 10 seeds

Abbreviations

ANN: artificial neural network; BM: bias mapper; CL: classification layer; DA: domain adaptation; GTEx: Genotype-Tissue Expression Project; GUI: graphical user interface; ML: machine learning; MLP: multi-layer perceptron; mca: mean class accuracy; msa: mean sample accuracy; RNA-seq: RNA-sequencing; SM: source mapper; SRA: Sequence Read Archive; TCGA: The Cancer Genome Atlas; TPM: transcripts per million; t-SNE: t-distributed stochastic neighbor embedding.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) grants CRU 306 P-C, 296 P8, and CRC 1286 Z2 for S.H., H.W., and S.B., respectively.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

SB and HW designed the project. HW designed and implemented the models and conducted the analyses. SH implemented the LIN model and conducted the corresponding analyses. SB, HW, and KK wrote the manuscript.

Acknowledgments

We thank the members of the Institute for Medical Systems Biology for helpful discussions and comments. In particular, we would like to thank Sergio Oller for help with IT problems and support. In addition, we thank ZMNH IT for their continuous support. We would like to acknowledge TCGA Research Network (https://www.cancer.gov/tcga) for data provision.

References

1. Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. Wiley Interdiscip Rev RNA 2017; **8**(1),doi:10.1002/wrna.1364.
2. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 2016; **17**(6):33.
3. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;**10**(1):57–63.
4. Costa V, Aprile M, Esposito R, et al. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. Eur J Hum Genet 2013;**21**(2):134–42.
5. 't Hoen PAC, Friedländer MR, Almlöf J, et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat Biotechnol 2013;**31**(11):1015–22.
6. Li S, Łabaj PP, Zumbo P, et al. Detecting and correcting systematic variation in large-scale RNA sequencing data. Nat Biotechnol 2014;**32**(9):888–95.
7. Torralba A, Efros AA. Unbiased look at dataset bias. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011), Colorado Springs. IEEE; 2011:1521–28.
8. Taub MA, Bravo HC, Irizarry RA. Overcoming bias and systematic errors in next generation sequencing data. Genome Med 2010;**2**(12):87.
9. Stephens ZD, Lee SY, Faghri F, et al. Big data: astronomical or genomical? PLoS Biol 2015;**13**(7):e1002195.
10. Lonsdale J, Thomas J, Salvatore M, et al. The genotype-tissue expression (GTEx) project. Nat Genet 2013;**45**(6):580–5.
11. Leinonen R, Sugawara H, Shumway M, et al. The Sequence Read Archive. Nucleic Acids Res 2010;**39**(suppl1):D19–21.
12. Shah NH, Jonquet C, Chiang AP, et al. Ontology-driven indexing of public datasets for translational bioinformatics. BMC Bioinformatics 2009;**10**(S2):S1.
13. Galeota E, Pelizzola M. Ontology-based annotations and semantic relations in large-scale (epi) genomics data. Brief Bioinform 2017;**18**(3):403–12.
14. Bernstein MN, Doan A, Dewey CN. MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. Bioinformatics 2017;**33**(18):2914–23.
15. Hatje K, Rahman RU, Vidal RO, et al. The landscape of human mutually exclusive splicing. Mol Syst Biol 2017;**13**(12): 959.
16. Bernstein MN, Ma Z, Gleicher M, et al. CellO: Comprehensive and hierarchical cell type classification of human cells with the Cell Ontology. Science 2021;**24**(1):101913.
17. Smith AM, Walsh Long J, JR, et al. Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. BMC Bioinformatics 2020;**21**(1):119.
18. Zarringhalam K, Degras D, Brockel C, et al. Robust phenotype prediction from gene expression data using differential shrinkage of co-regulated genes. Sci Rep 2018;**8**(1): 1237.
19. Kang T, Ding W, Zhang L, et al. A biological network-based regularized artificial neural network model for robust phenotype prediction from gene expression data. BMC Bioinformatics 2017;**18**(1):656.
20. Ellis SE, Collado-Torres L, Jaffe A, et al. Improving the value of public RNA-seq expression data by phenotype prediction.

Nucleic Acids Res 2018;**46**(9):e54.

21. Marouf M, Machart P, Bansal V, et al. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. Nat Commun 2020;**11**(1):166.
22. Menden K, Marouf M, Oller S, et al. Deep learning–based cell composition analysis from tissue expression profiles. Sci Adv 2020;**6**(30):eaba2619.
23. Mamoshina P, Vieira A, Putin E, et al. Applications of deep learning in biomedicine. Mol Pharm 2016;**13**(5):1445–54.
24. Wainberg M, Merico D, Delong A, et al. Deep learning in biomedicine. Nat Biotechnol 2018;**36**(9):829–38.
25. Csurka G. Domain adaptation for visual applications: A comprehensive survey. arXiv 2017: 1702.05374.
26. Tommasi T, Patricia N, Caputo B, et al. A deeper look at dataset bias. arXiv 2017: 1505.01257.
27. Arora S, Pattwell SS, Holland EC, et al. Variability in estimated gene expression among commonly used RNA-seq pipelines. Sci Rep 2020;**10**(1):2734.
28. recount2, 13.09.2019. https://jhubiostatistics.shinyapps.io/recount/.
29. Nellore A, Collado-Torres L, Jaffe AE, et al. Rail-RNA: scalable analysis of RNA-seq splicing and coverage. Bioinformatics 2017;**33**(24):4033–40.
30. Wartmann H, Heins S, Kloiber K, et al. Supporting data for "Bias invariant RNA-seq metadata annotation." GigaScience Database 2021, http://dx.doi.org/10.5524/100920.
31. Project GTEx, 13.09.2019. https://www.gtexportal.org/.
32. MetaSRA Download, 11.09.2019. http://metasra.biostat.wisc.edu/download.html.
33. Ceriani L, Verme P. The origins of the Gini index: extracts from Variabilità e Mutabilità (1912) by Corrado Gini. J Econ Inequal 2012;**10**(2):421–43.
34. Zhang JD, Hatje K, Sturm G, et al. Detect tissue heterogeneity in gene expression data with BioQC. BMC Genomics 2017;**18**(1):277.
35. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015; **521**(7553):436–44.
36. Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. In: Computer Vision and Pattern Recognition, 2005 (CVPR 2005). IEEE; 2005:539–46.
37. Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2015:815–23.
38. Wartmann H, Heins S, Kloiber K, et al. Git repository for "Bias invariant RNA-seq metadata annotation." 2021. https://github.com/imsb-uke/rna_augment.
39. Phenopredict, 20.08.2019. http://rdrr.io/github/ShanEllis/phenopredict/.
40. Chopra S, Balakrishnan S, Gopalan R. Dlid: Deep learning for domain adaptation by interpolating between domains. In: ICML Workshop on challenges in representation learning, vol. **2**; 2013.doi/10.5555/3045118.3045244.
41. Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks. J Mach Learn Res 2016; **17**(1):2030–96.
42. Tzeng E, Hoffman J, Saenko K, et al. Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 2017:7167–76.
43. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Adv Neural Inform Process Syst 2014;**27**.doi:10.1145/3422622.
44. Motiian S, Jones Q, Iranmanesh S, et al. Few-shot adversarial domain adaptation. Adv Neural Inform Process Syst 2017:**30**
45. Yeung ML, Yao Y, Jia L, et al. MERS coronavirus induces apoptosis in kidney and lung by upregulating Smad7 and FGF2. Nat Microbiol 2016; **1**(3):16004.
46. Kravtsova-Ivantsiv Y, Shomer I, Cohen-Kaplan V, et al. KPC1-mediated ubiquitination and proteasomal processing of NF-kB1 p105 to p50 restricts tumor growth. Cell 2015;**161**(2):333–47.