iScience, Volume *23*

Supplemental Information

Scalable Prediction of Acute Myeloid Leukemia Using High-Dimensional Machine Learning and Blood Transcriptomics

Stefanie Warnat-Herresthal, Konstantinos Perrakis, Bernd Taschler, Matthias Becker, Kevin Baßler, Marc Beyer, Patrick Günther, Jonas Schulte-Schrepping, Lea Seep, Kathrin Klee, Thomas Ulas, Torsten Haferlach, Sach Mukherjee, and Joachim L. Schultze

Figure S1

A

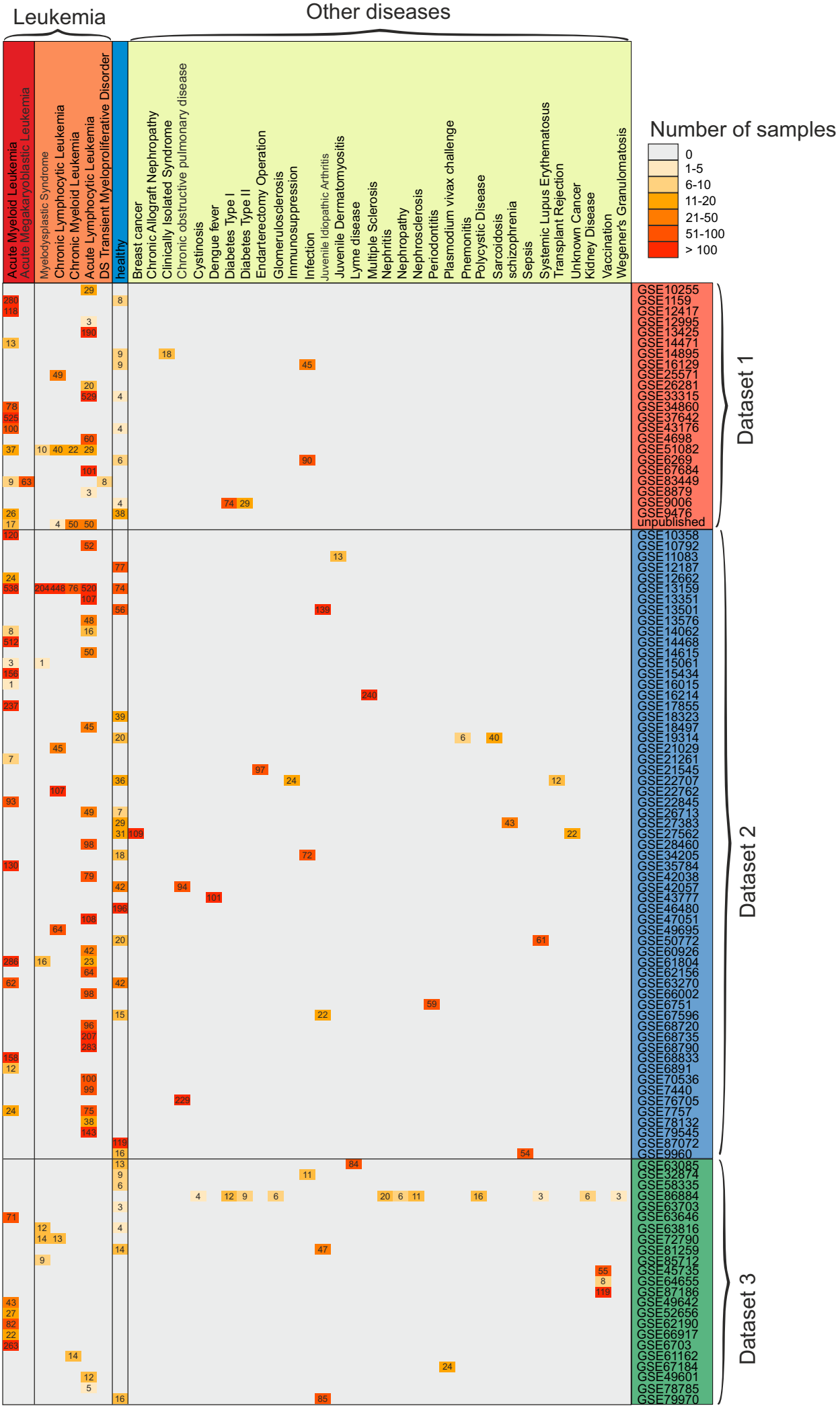
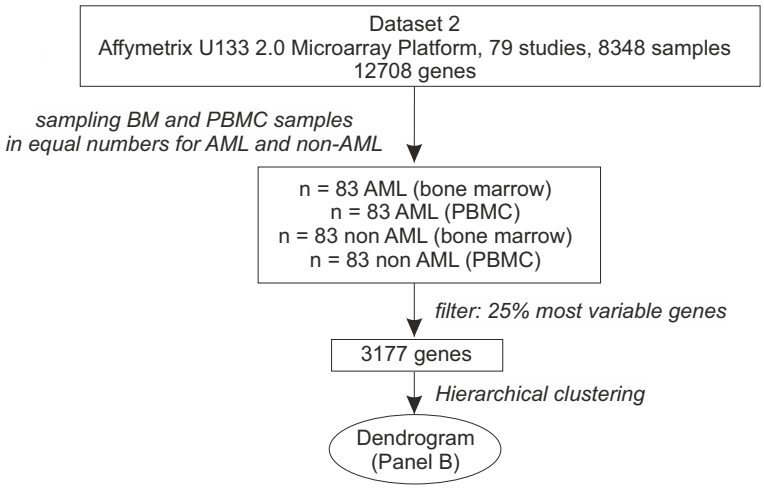

Leukemia
Other diseases
Acute Myeloid Leukemia
Acute Megakaryoblastic Leukemia
Myelodysplastic Syndrome
Chronic Lymphocytic Leukemia
Chronic Myeloid Leukemia
Acute Lymphocytic Leukemia
DS Transient Myeloproliferative Disorder
healthy
Breast cancer
Chronic Allograft Nephropathy
Clinically Isolated Syndrome
Chronic obstructive pulmonary disease
Cystinosis
Dengue fever
Diabetes Type I
Diabetes Type II
Endarterectomy Operation
Glomerulosclerosis
Immunosuppression
Infection
Juvenile Idiopathic Arthritis
Juvenile Dermatomyositis
Lyme disease
Multiple Sclerosis
Nephritis
Nephropathy
Nephrosclerosis
Periodontitis
Plasmodium vivax challenge
Pneumonitis
Polycystic Disease
Sarcoidosis
schizophrenia
Sepsis
Systemic Lupus Erythematosus
Transplant Rejection
Unknown Cancer
Kidney Disease
Vaccination
Wegener's Granulomatosis
Number of samples
0
1-5
6-10
11-20
21-50
51-100
> 100
GSE10255
GSE1159
GSE12417
GSE12995
GSE13425
GSE14471
GSE14895
GSE16129
GSE25571
GSE26281
GSE33315
GSE34860
GSE37642
GSE43176
GSE4698
GSE51082
GSE6269
GSE67684
GSE83449
GSE8879
GSE9006
GSE9476
unpublished
Dataset 1
GSE10358
GSE10792
GSE11083
GSE12187
GSE12662
GSE13159
GSE13351
GSE13501
GSE13576
GSE14062
GSE14468
GSE14615
GSE15061
GSE15434
GSE16015
GSE16214
GSE17855
GSE18323
GSE18497
GSE19314
GSE21029
GSE21261
GSE21545
GSE22707
GSE22762
GSE22845
GSE26713
GSE27383
GSE27562
GSE28460
GSE34205
GSE35784
GSE42038
GSE42057
GSE43777
GSE46480
GSE47051
GSE49695
GSE50772
GSE60926
GSE61804
GSE62156
GSE63270
GSE66002
GSE6751
GSE67596
GSE68720
GSE68735
GSE68790
GSE68833
GSE6891
GSE70536
GSE7440
GSE76705
GSE7757
GSE78132
GSE79545
GSE87072
GSE9960
Dataset 2
GSE63085
GSE32874
GSE58335
GSE86884
GSE63703
GSE63646
GSE63816
GSE72790
GSE81259
GSE85712
GSE45735
GSE64655
GSE87186
GSE49642
GSE52656
GSE62190
GSE66917
GSE6703
GSE61162
GSE67184
GSE49601
GSE78785
GSE79970
Dataset 3

Figure S2

A


Dataset 2
Affymetrix U133 2.0 Microarray Platform, 79 studies, 8348 samples
12708 genes
sampling BM and PBMC samples
in equal numbers for AML and non-AML
n = 83 AML (bone marrow)
n = 83 AML (PBMC)
n = 83 non AML (bone marrow)
n = 83 non AML (PBMC)
filter: 25% most variable genes
3177 genes
Hierarchical clustering
Dendrogram
(Panel B)

B

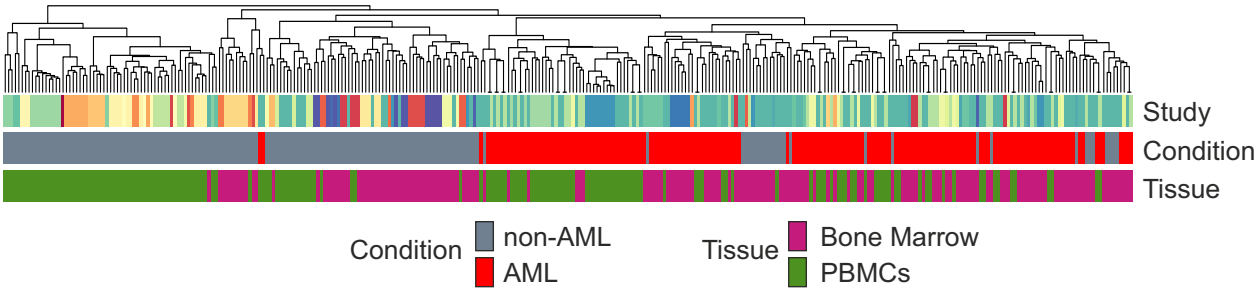
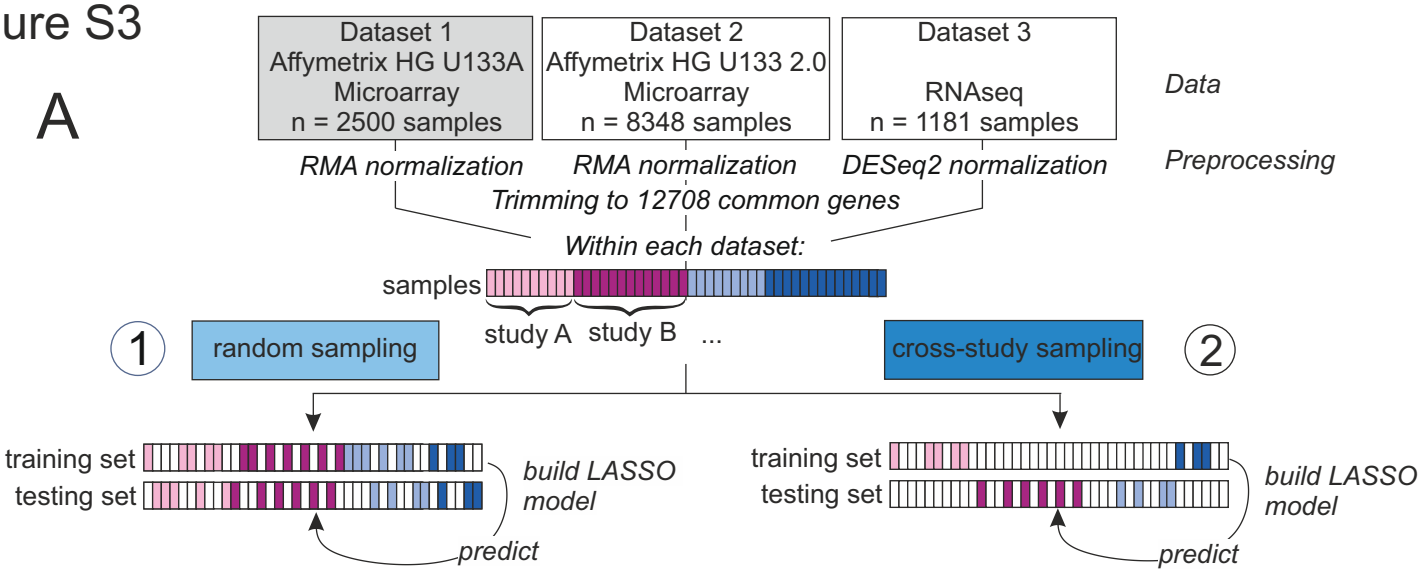
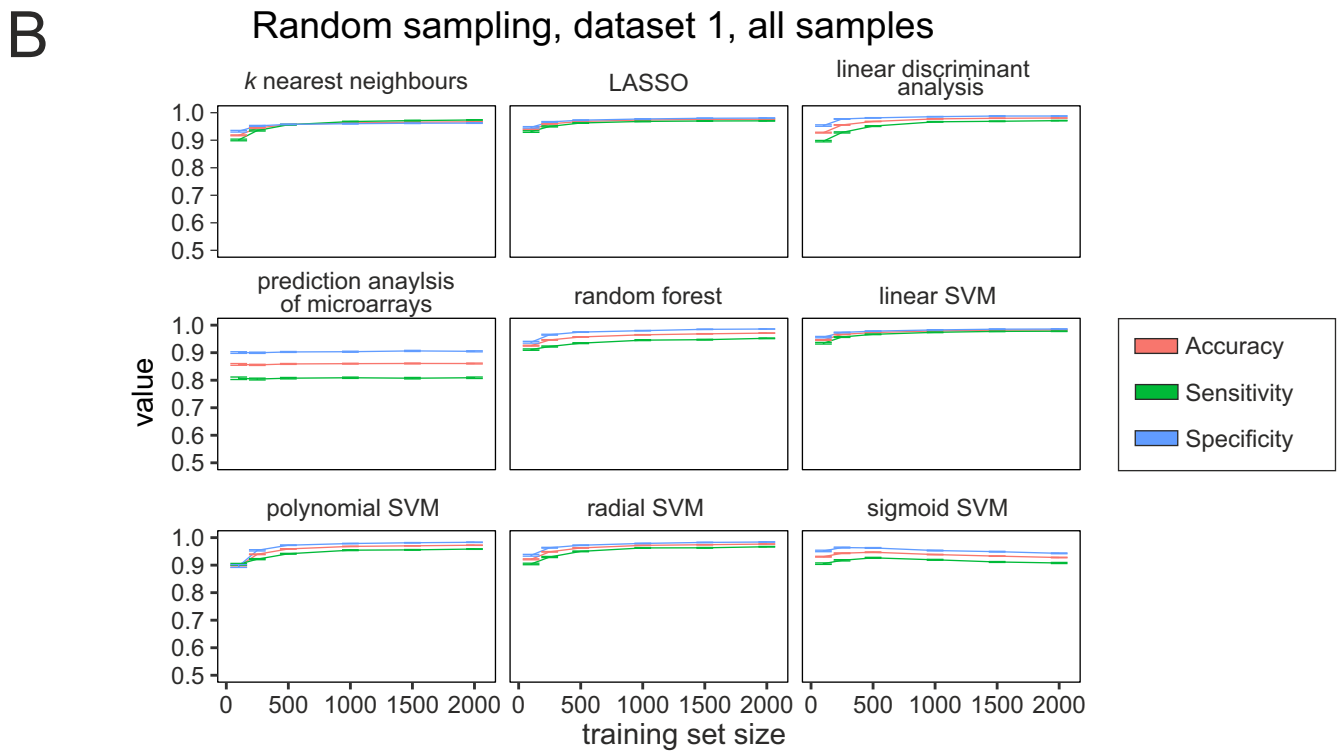

Study
Condition
Tissue
Condition
non-AML
AML
Tissue
Bone Marrow
PBMCs

Figure S3


A
Dataset 1
Affymetrix HG U133A
Microarray
n = 2500 samples
Dataset 2
Affymetrix HG U133 2.0
Microarray
n = 8348 samples
Dataset 3
RNAseq
n = 1181 samples
Data
RMA normalization
RMA normalization
DESeq2 normalization
Preprocessing
Trimming to 12708 common genes
Within each dataset:
samples
study A
study B
...
1
random sampling
cross-study sampling
2
training set
testing set
build LASSO model
predict
training set
testing set
build LASSO model
predict


B
Random sampling, dataset 1, all samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
training set size
Accuracy
Sensitivity
Specificity

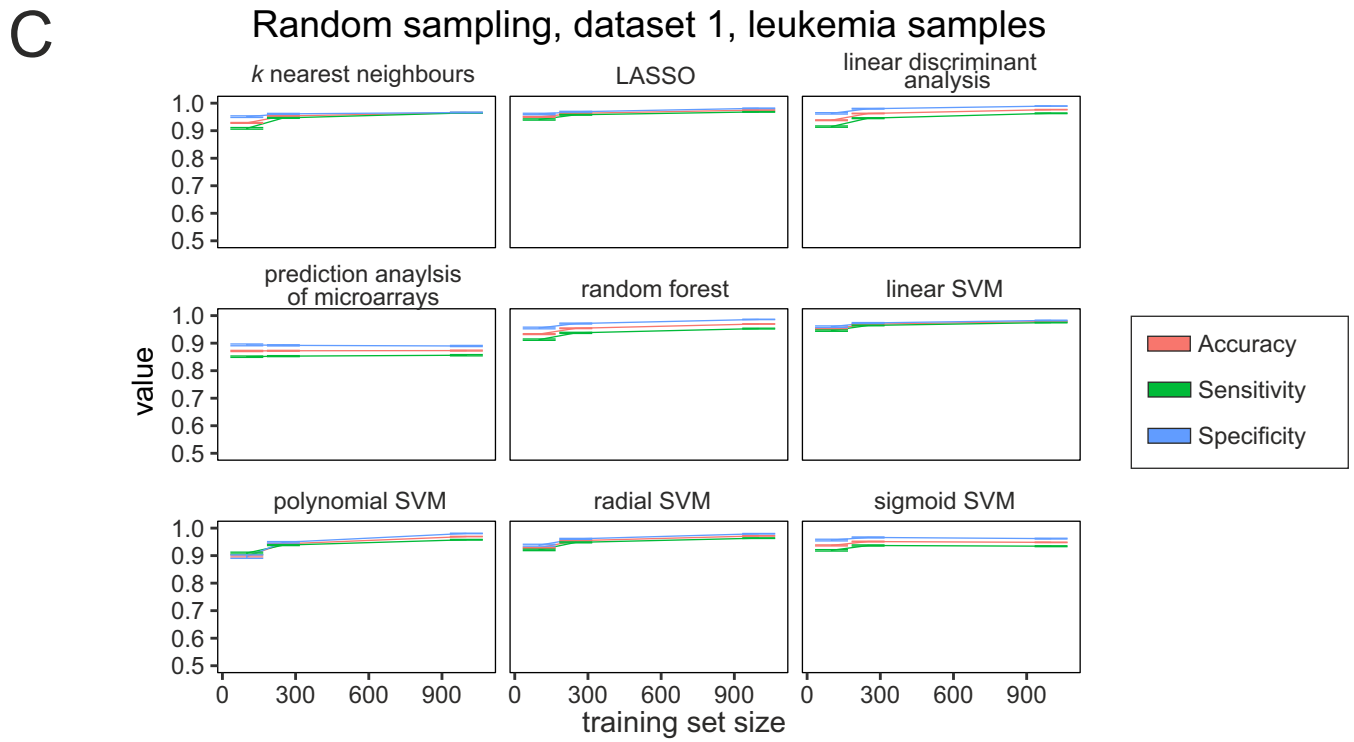

C
Random sampling, dataset 1, leukemia samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
training set size
Accuracy
Sensitivity
Specificity

Figure S4

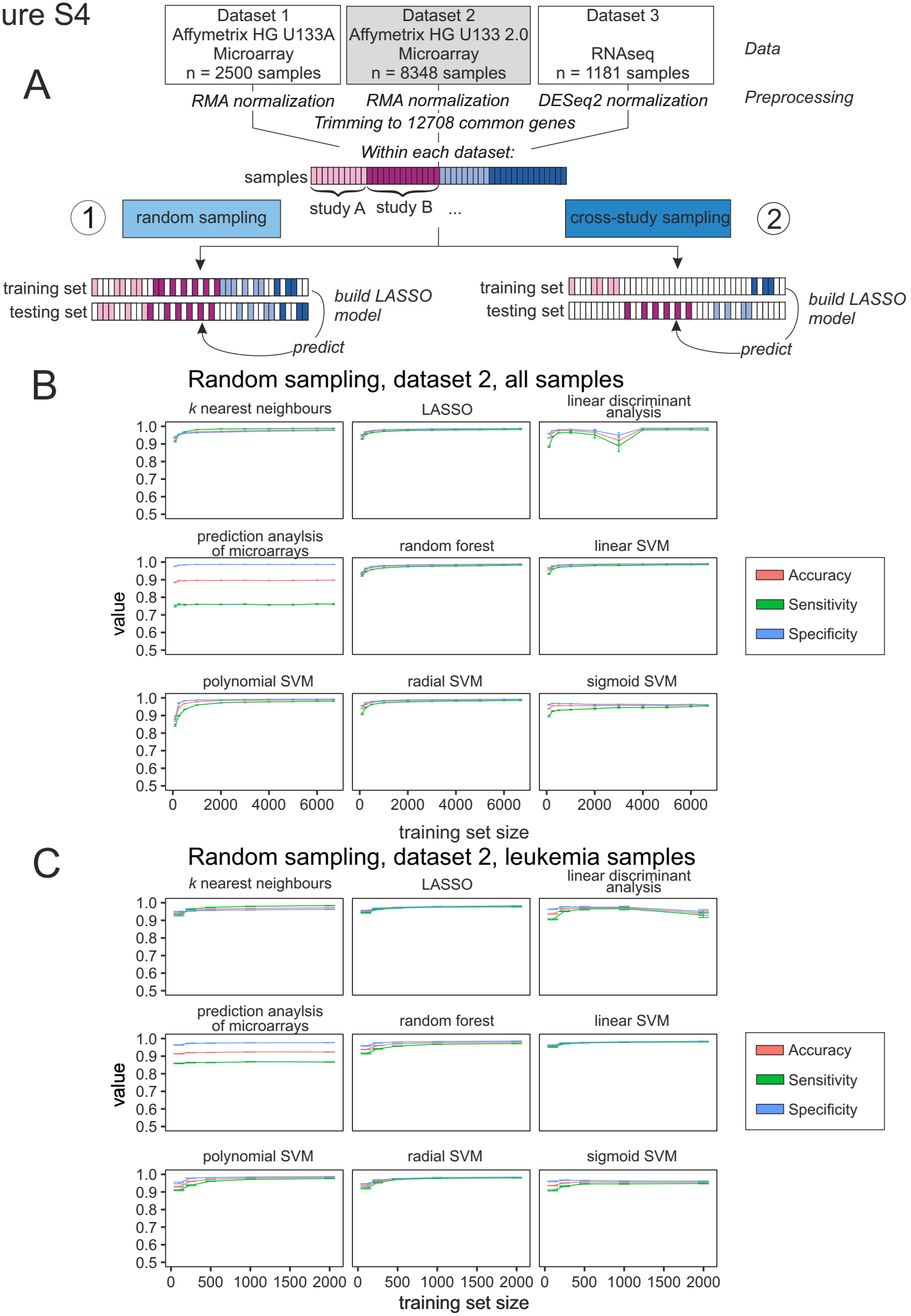
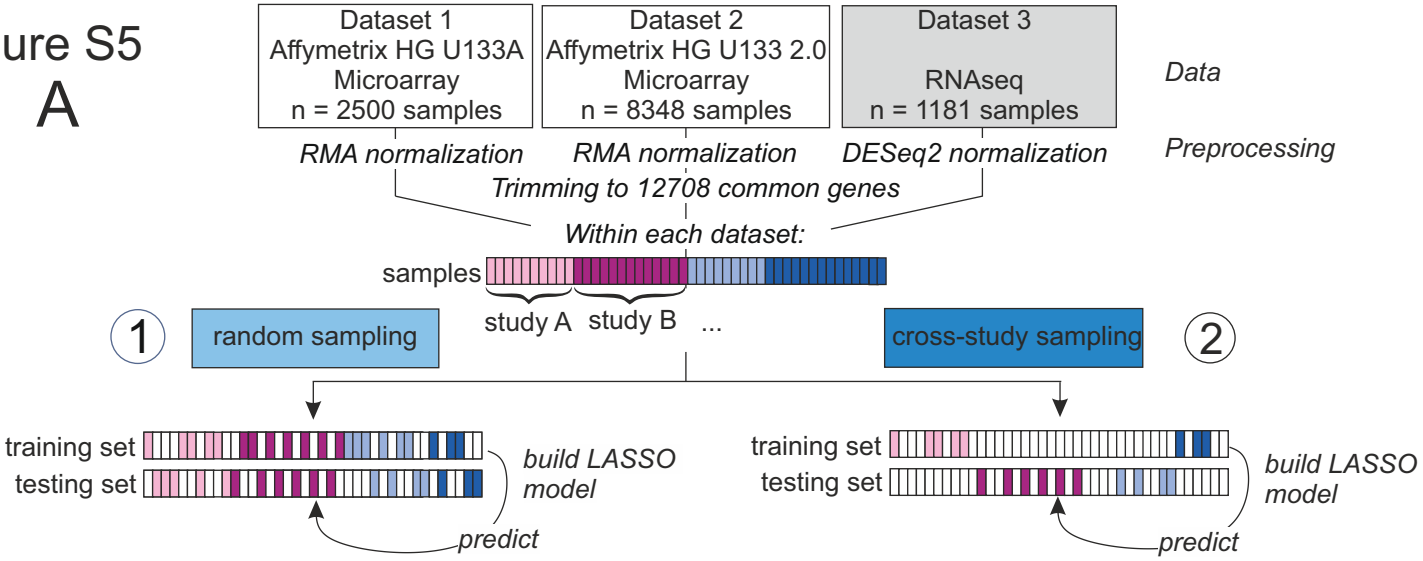

A
Dataset 1
Affymetrix HG U133A
Microarray
n = 2500 samples
Dataset 2
Affymetrix HG U133 2.0
Microarray
n = 8348 samples
Dataset 3
RNAseq
n = 1181 samples
Data
RMA normalization
RMA normalization
DESeq2 normalization
Preprocessing
Trimming to 12708 common genes
Within each dataset:
samples
study A
study B
...
1
random sampling
cross-study sampling
2
training set
testing set
build LASSO model
predict
training set
testing set
build LASSO model
predict
B
Random sampling, dataset 2, all samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
Accuracy
Sensitivity
Specificity
training set size
C
Random sampling, dataset 2, leukemia samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
Accuracy
Sensitivity
Specificity
training set size

Figure S5

A


Dataset 1
Affymetrix HG U133A
Microarray
n = 2500 samples
Dataset 2
Affymetrix HG U133 2.0
Microarray
n = 8348 samples
Dataset 3
RNAseq
n = 1181 samples
Data
RMA normalization
RMA normalization
DESeq2 normalization
Preprocessing
Trimming to 12708 common genes
Within each dataset:
samples
study A
study B
...
1
random sampling
cross-study sampling
2
training set
testing set
build LASSO model
predict
training set
testing set
build LASSO model
predict

B

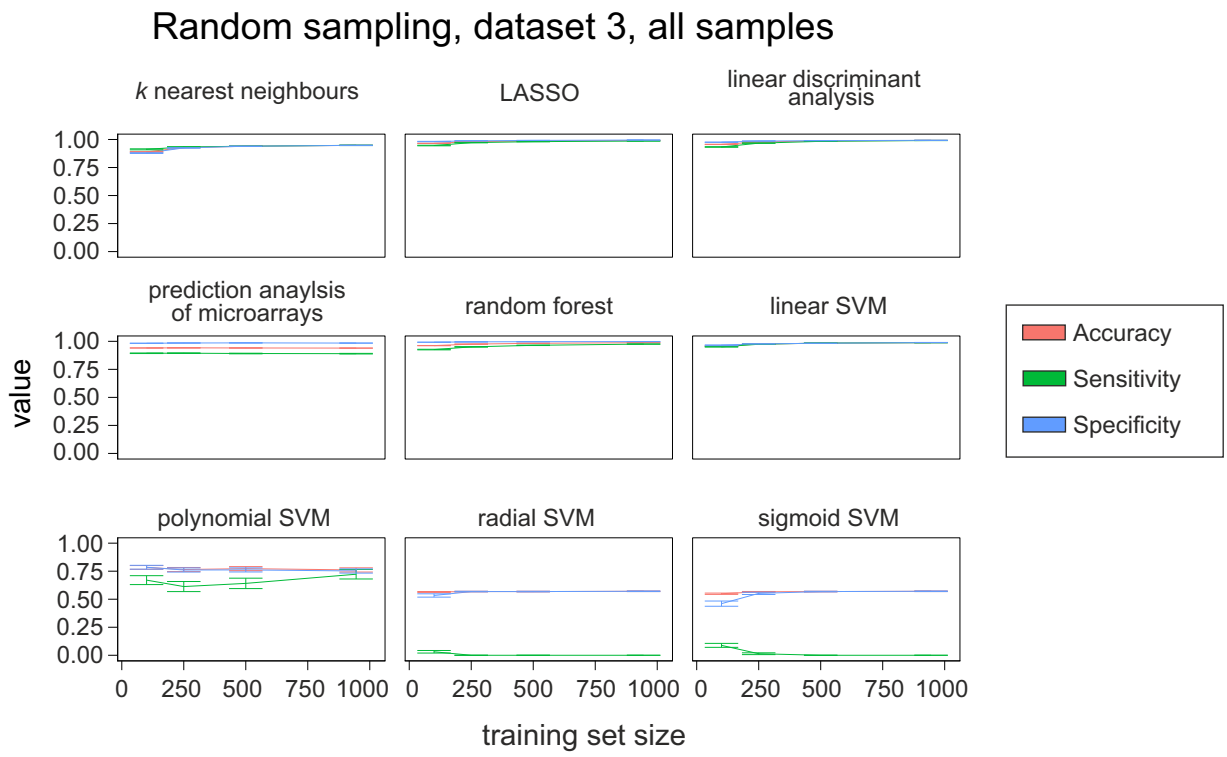
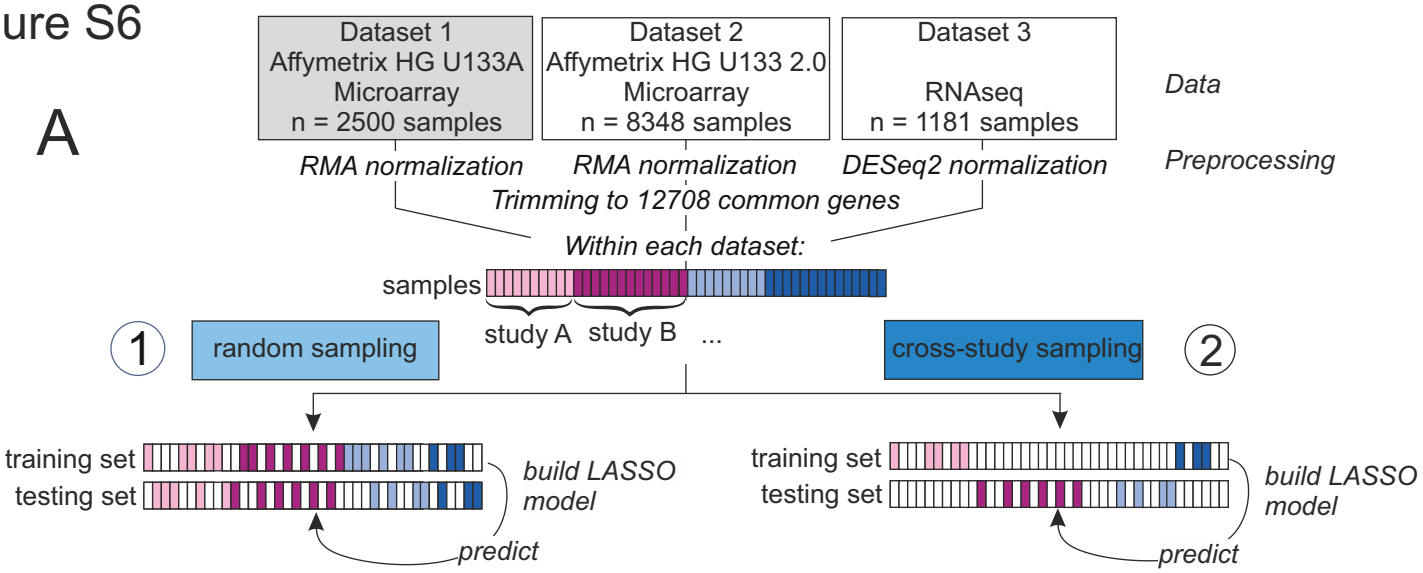
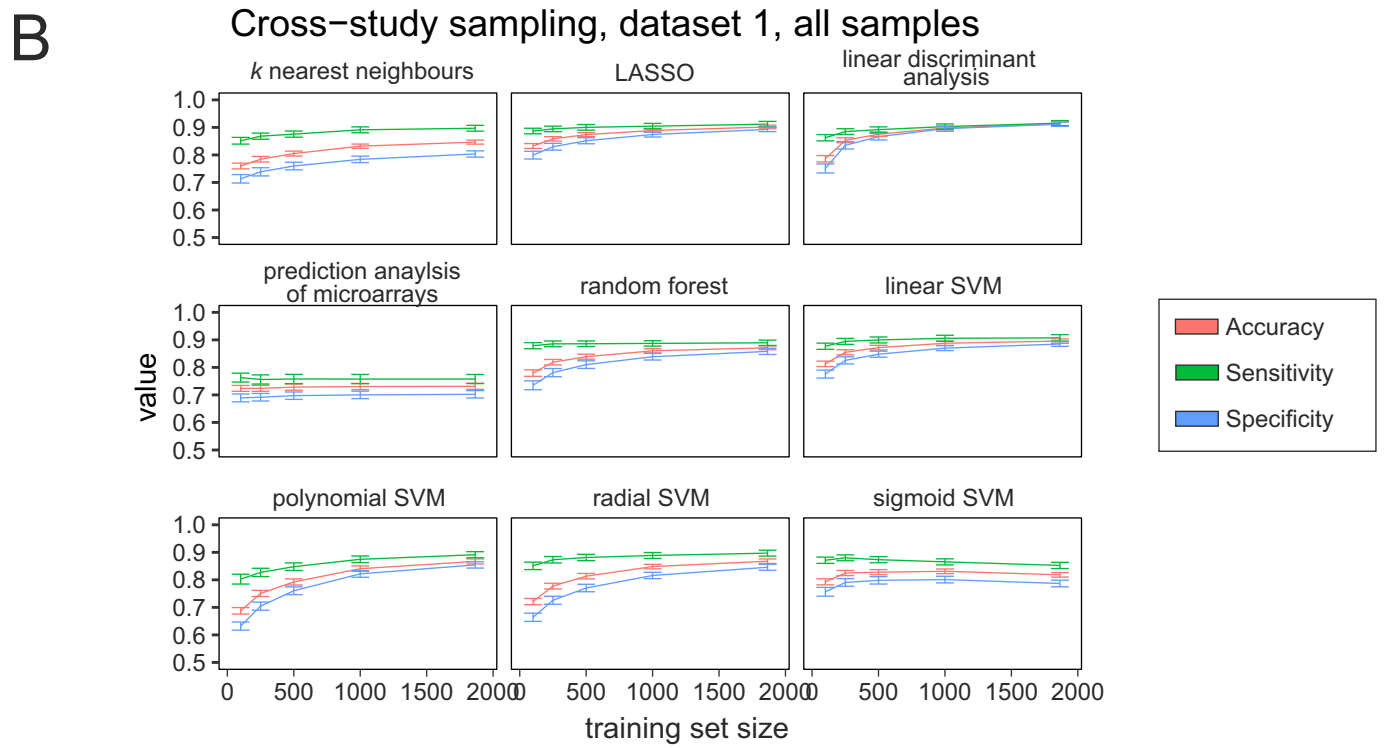

Random sampling, dataset 3, all samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
training set size
Accuracy
Sensitivity
Specificity

Figure S6


A
Dataset 1
Affymetrix HG U133A
Microarray
n = 2500 samples
Dataset 2
Affymetrix HG U133 2.0
Microarray
n = 8348 samples
Dataset 3
RNAseq
n = 1181 samples
Data
RMA normalization
RMA normalization
DESeq2 normalization
Preprocessing
Trimming to 12708 common genes
Within each dataset:
samples
study A
study B
...
1
random sampling
cross-study sampling
2
training set
testing set
build LASSO model
predict
training set
testing set
build LASSO model
predict


B
Cross−study sampling, dataset 1, all samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
training set size
Accuracy
Sensitivity
Specificity

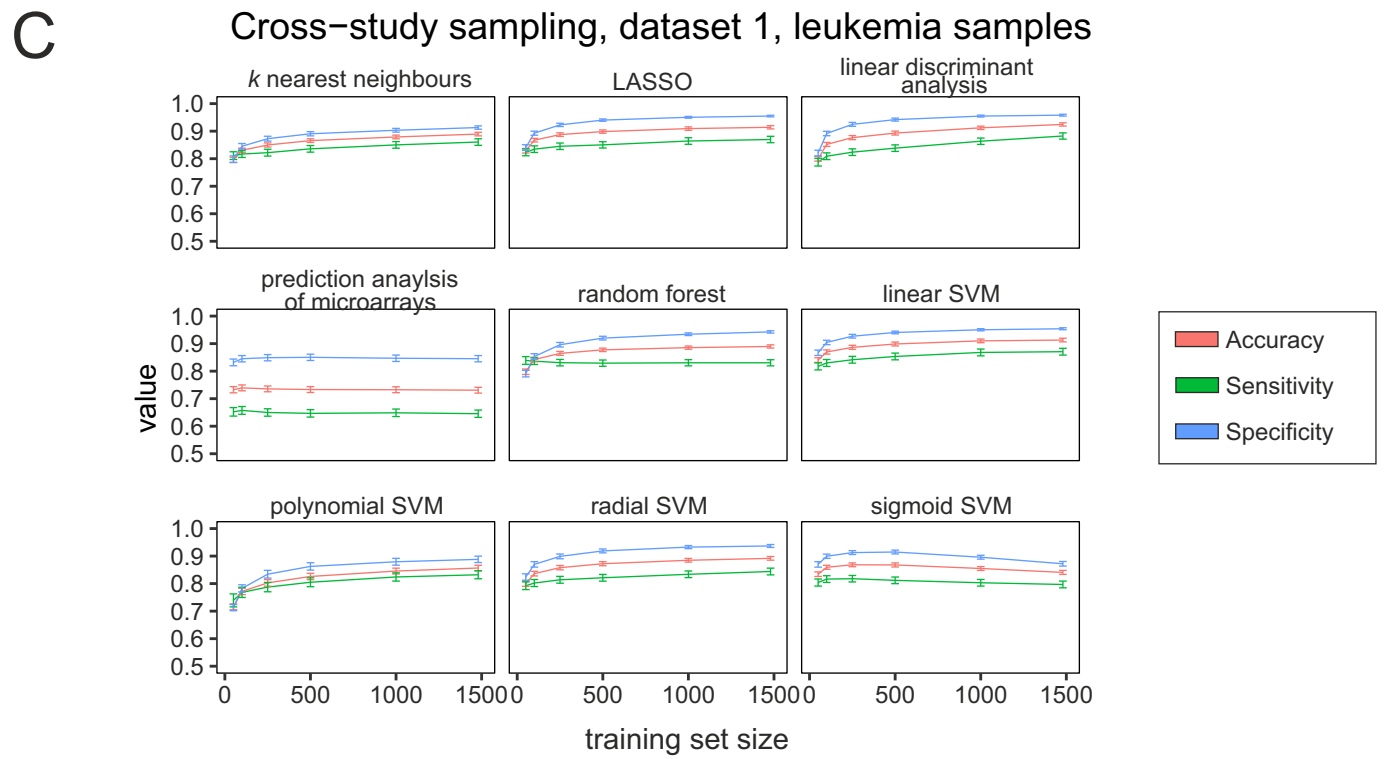
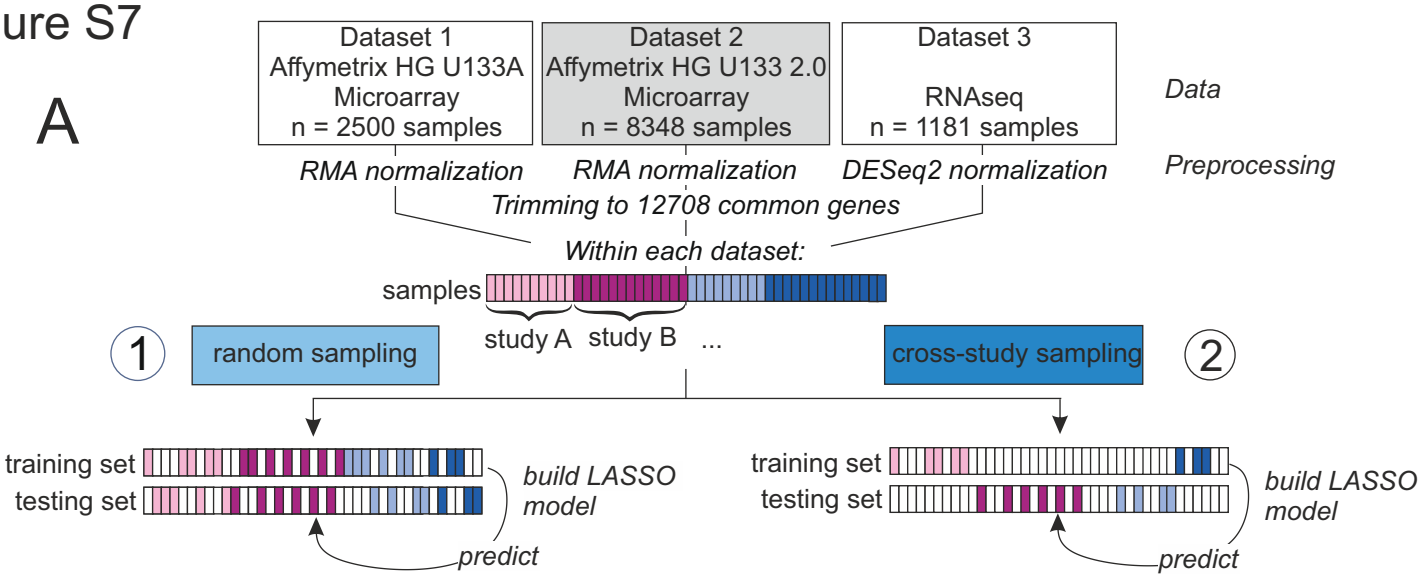
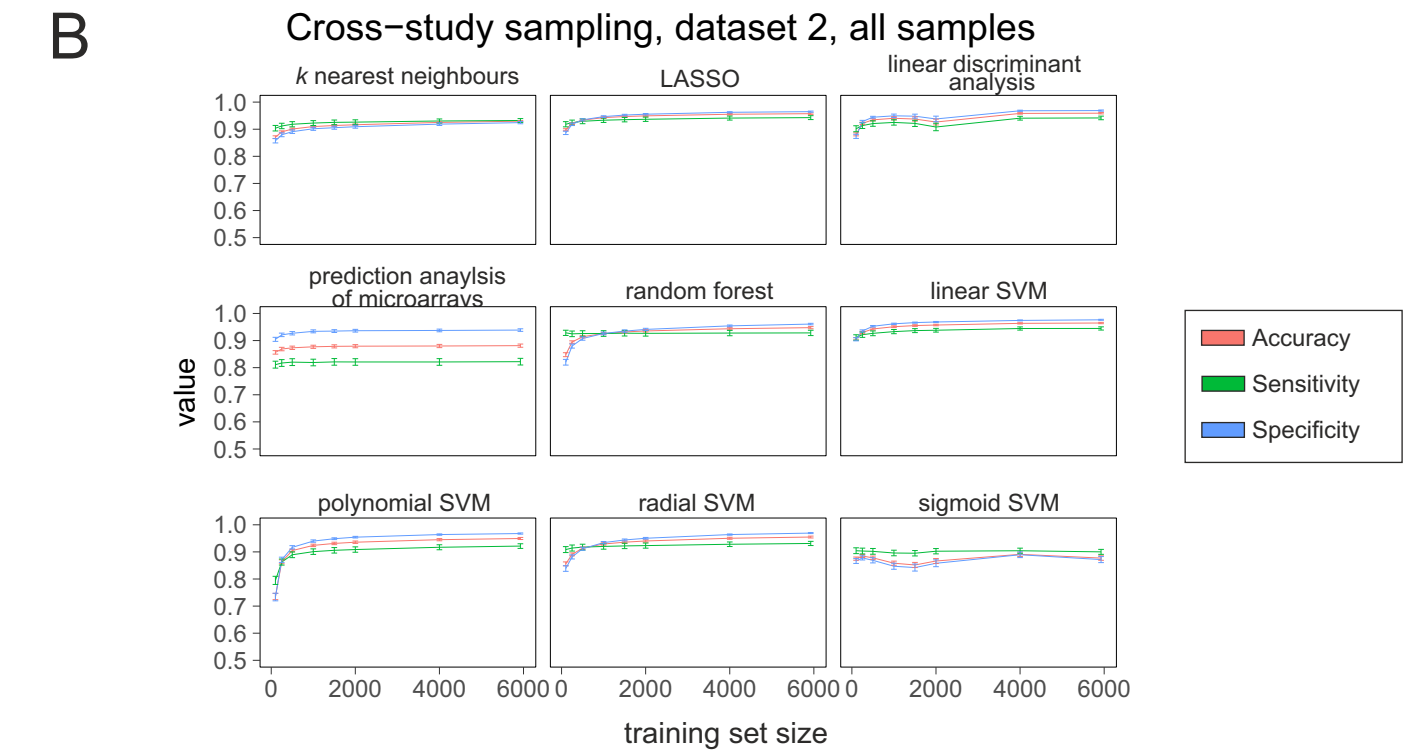

C
Cross−study sampling, dataset 1, leukemia samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
training set size
Accuracy
Sensitivity
Specificity

Figure S7


A
Dataset 1
Affymetrix HG U133A
Microarray
n = 2500 samples
Dataset 2
Affymetrix HG U133 2.0
Microarray
n = 8348 samples
Dataset 3
RNAseq
n = 1181 samples
Data
RMA normalization
RMA normalization
DESeq2 normalization
Preprocessing
Trimming to 12708 common genes
Within each dataset:
samples
study A
study B
...
1
random sampling
cross-study sampling
2
training set
testing set
build LASSO model
predict
training set
testing set
build LASSO model
predict


B
Cross−study sampling, dataset 2, all samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
training set size
Accuracy
Sensitivity
Specificity

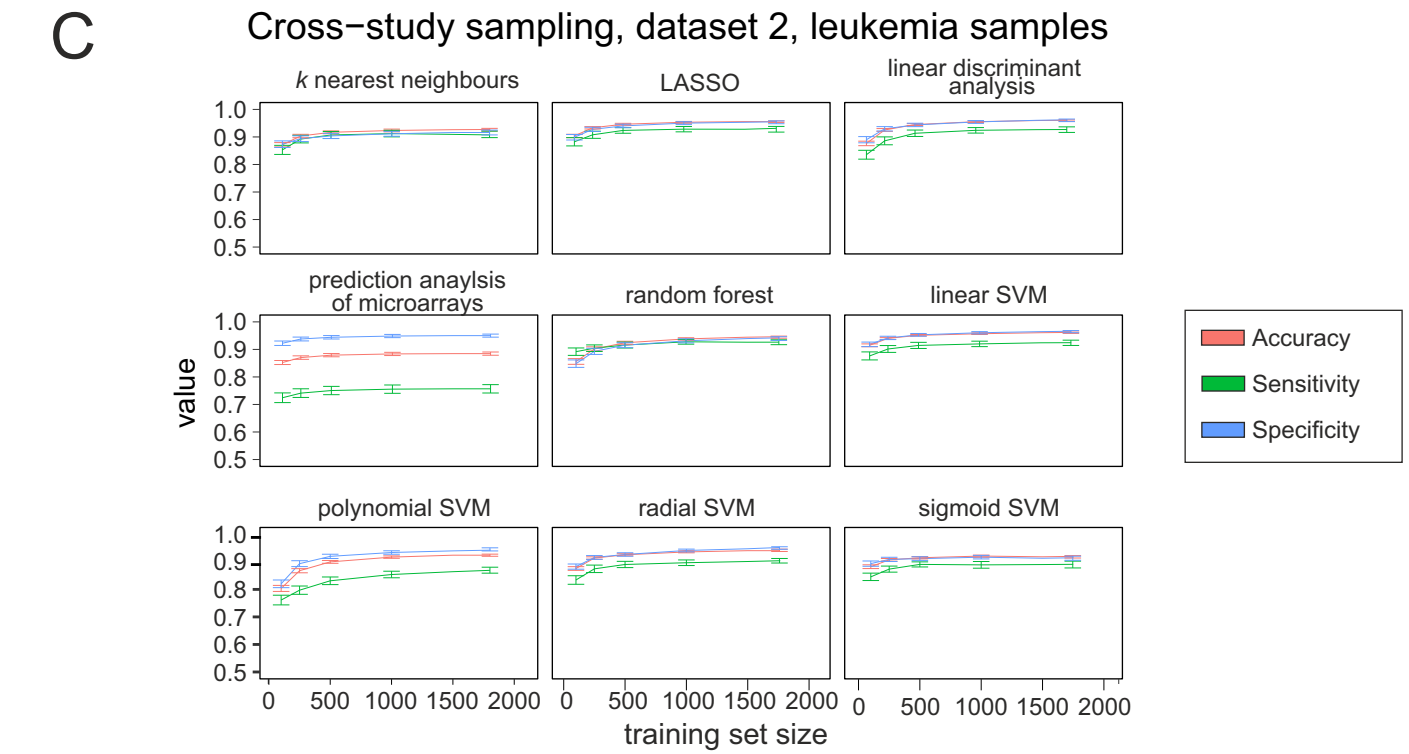
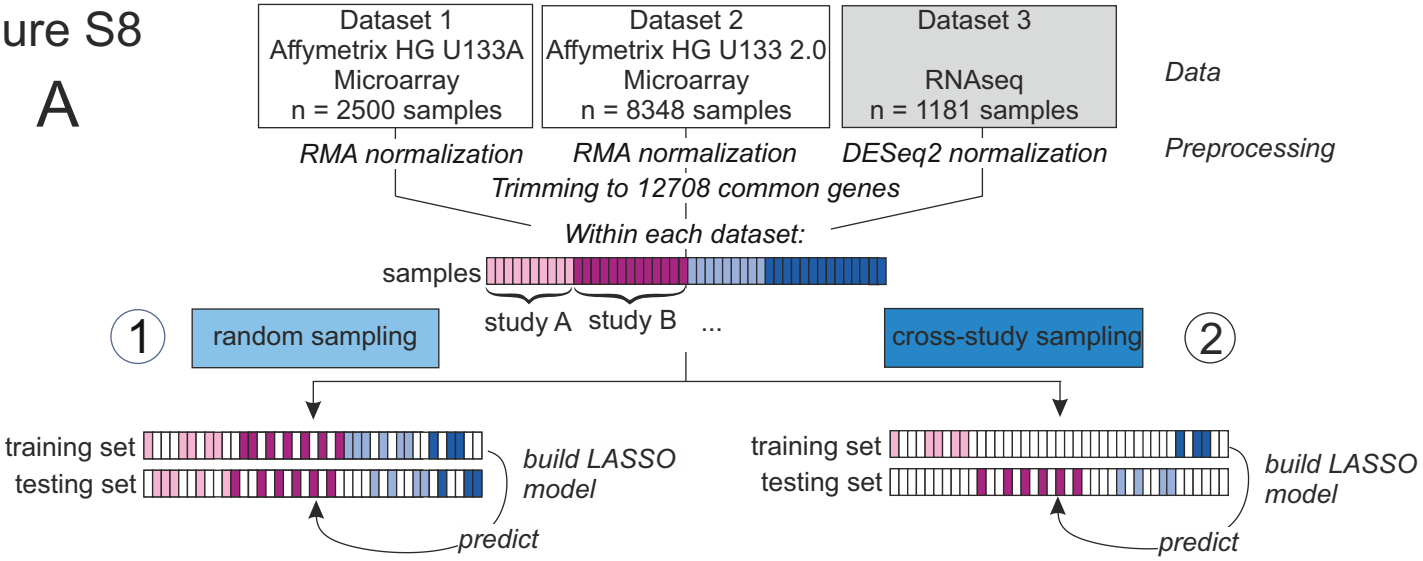

C
Cross−study sampling, dataset 2, leukemia samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
training set size
Accuracy
Sensitivity
Specificity

Figure S8

A


Dataset 1
Affymetrix HG U133A
Microarray
n = 2500 samples
Dataset 2
Affymetrix HG U133 2.0
Microarray
n = 8348 samples
Dataset 3
RNAseq
n = 1181 samples
Data
RMA normalization
RMA normalization
DESeq2 normalization
Preprocessing
Trimming to 12708 common genes
Within each dataset:
samples
study A
study B
...
1
random sampling
cross-study sampling
2
training set
testing set
build LASSO model
predict
training set
testing set
build LASSO model
predict

B

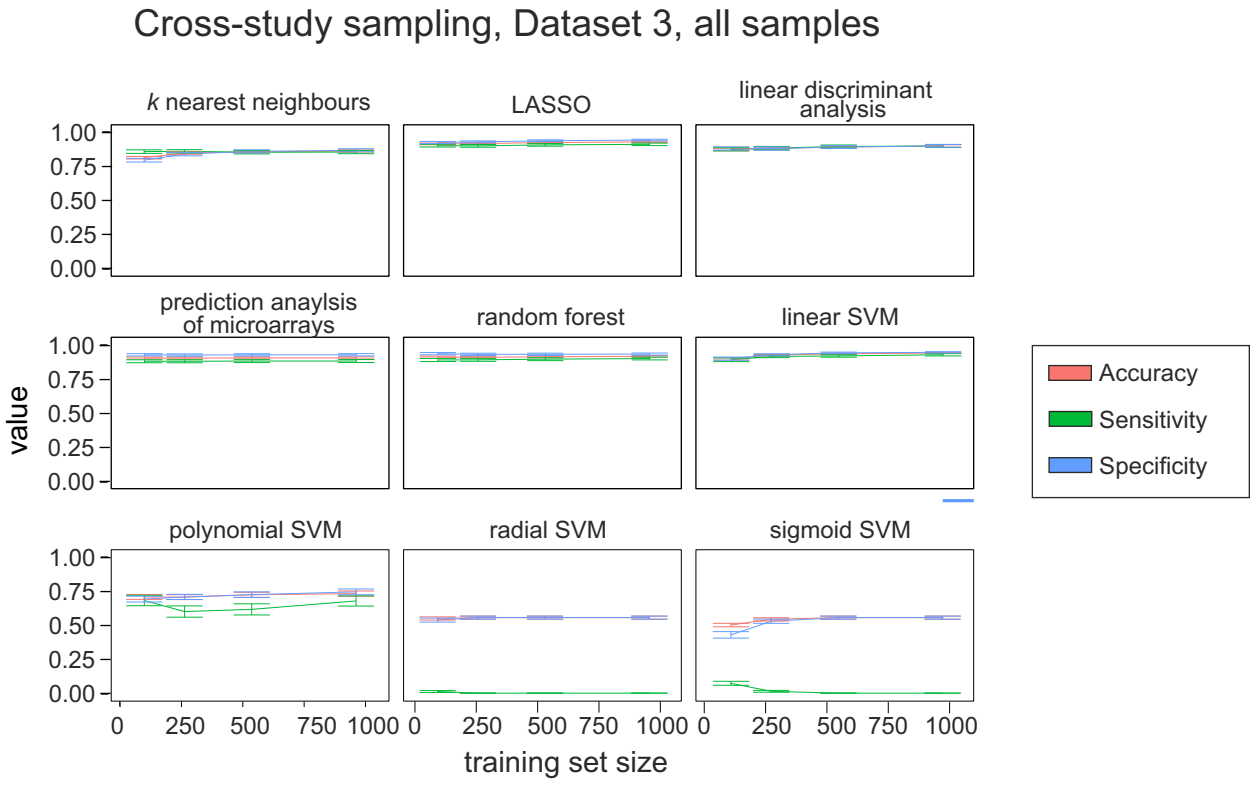
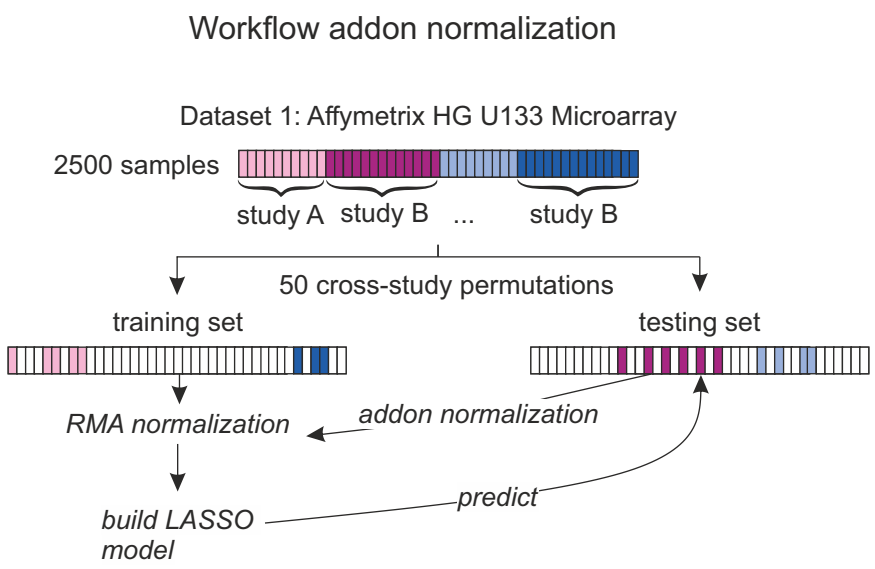

Cross-study sampling, Dataset 3, all samples
k nearest neighbours
LASSO
linear discriminant analysis
prediction anaylsis of microarrays
random forest
linear SVM
polynomial SVM
radial SVM
sigmoid SVM
value
1.00
0.75
0.50
0.25
0.00
0
250
500
750
1000
training set size
Accuracy
Sensitivity
Specificity

Figure S9

A


Workflow addon normalization
Dataset 1: Affymetrix HG U133 Microarray
2500 samples
study A
study B
...
study B
50 cross-study permutations
training set
testing set
RMA normalization
addon normalization
build LASSO model
predict

B

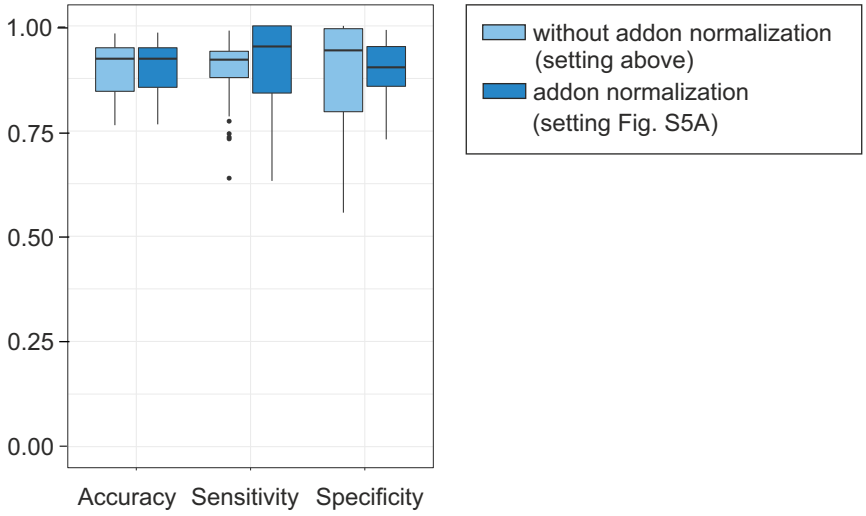
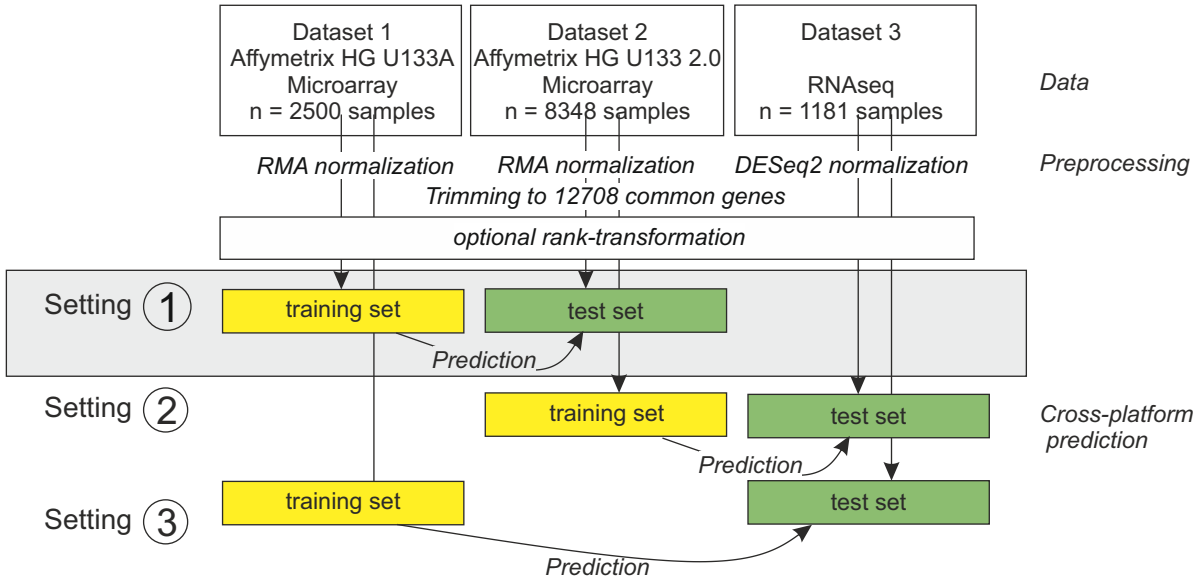
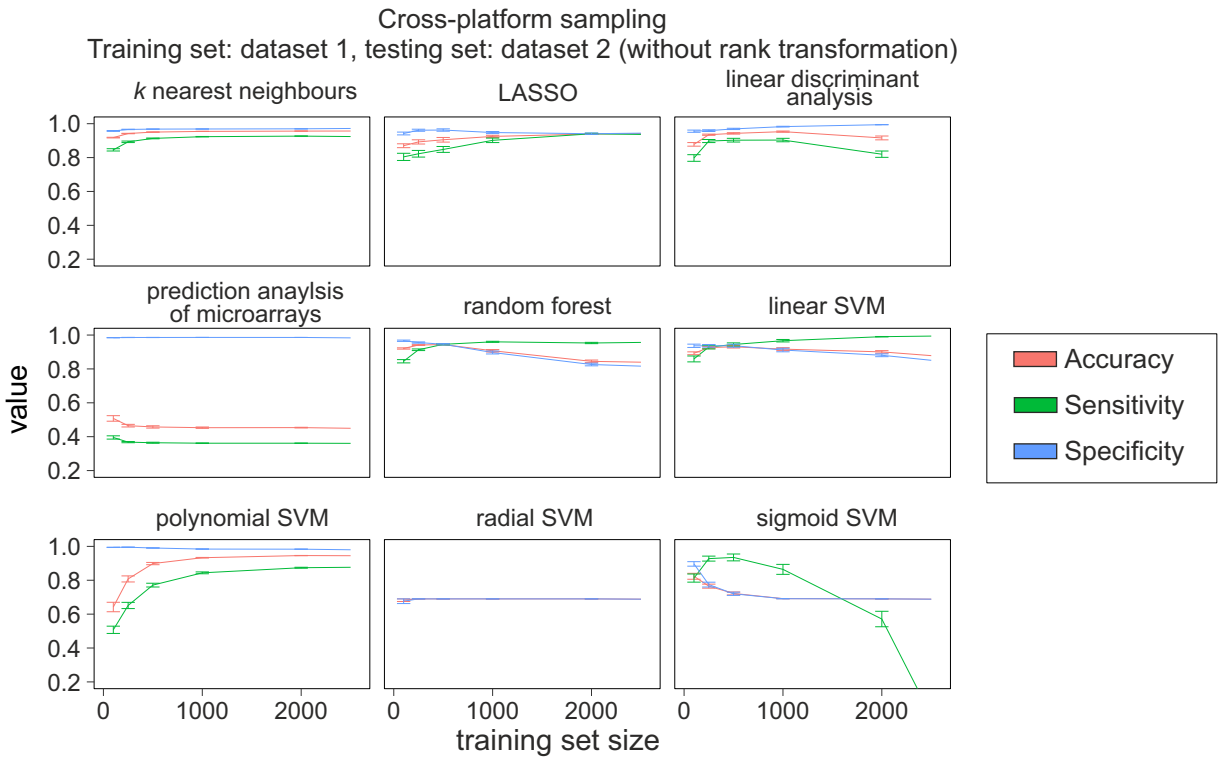

1.00
0.75
0.50
0.25
0.00
Accuracy
Sensitivity
Specificity
without addon normalization (setting above)
addon normalization (setting Fig. S5A)

Figure S10

A



B



C

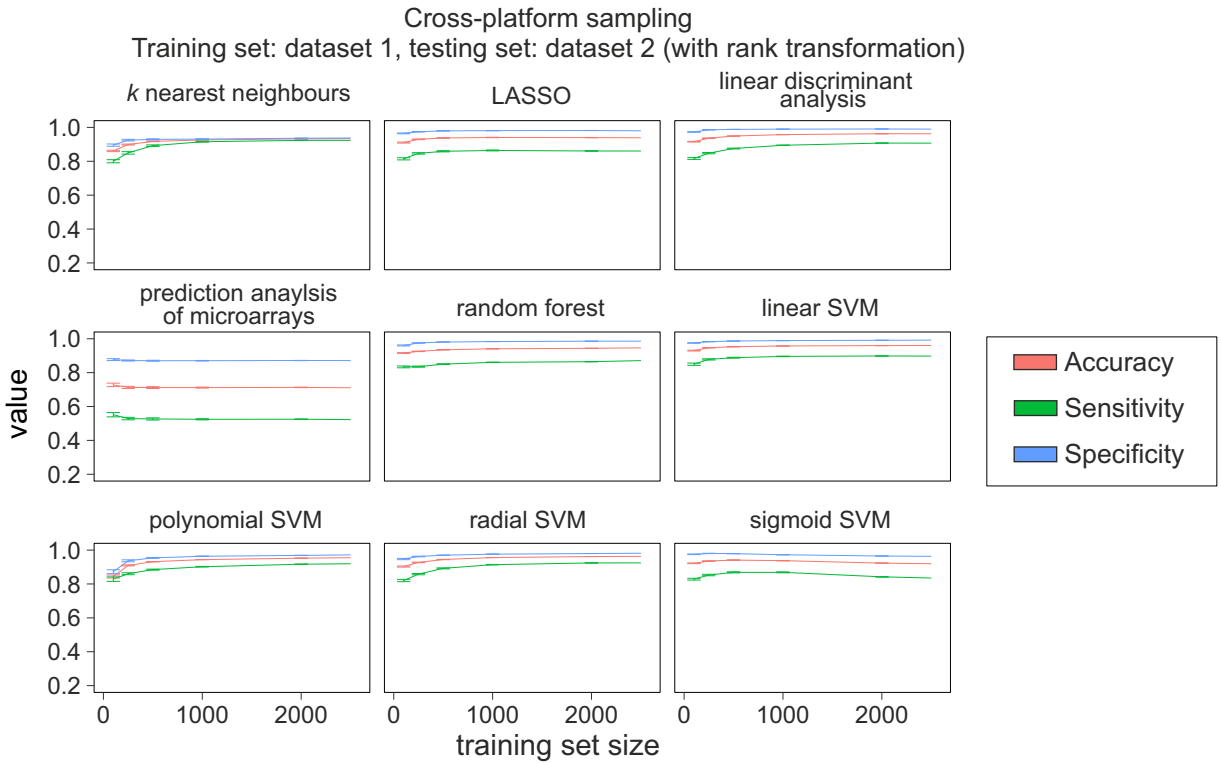
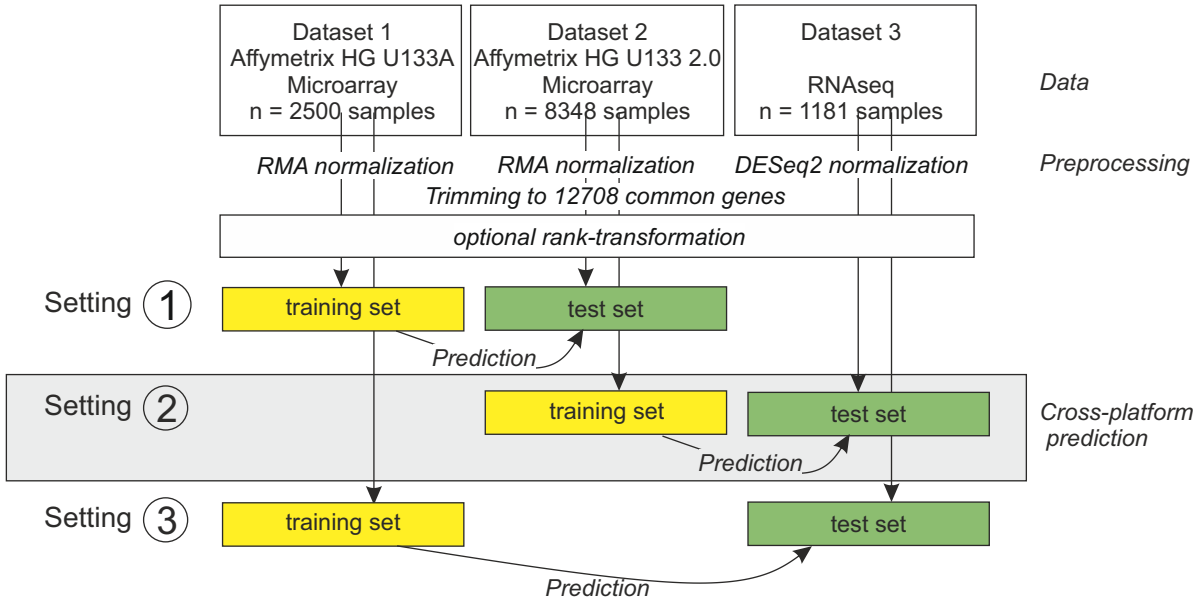
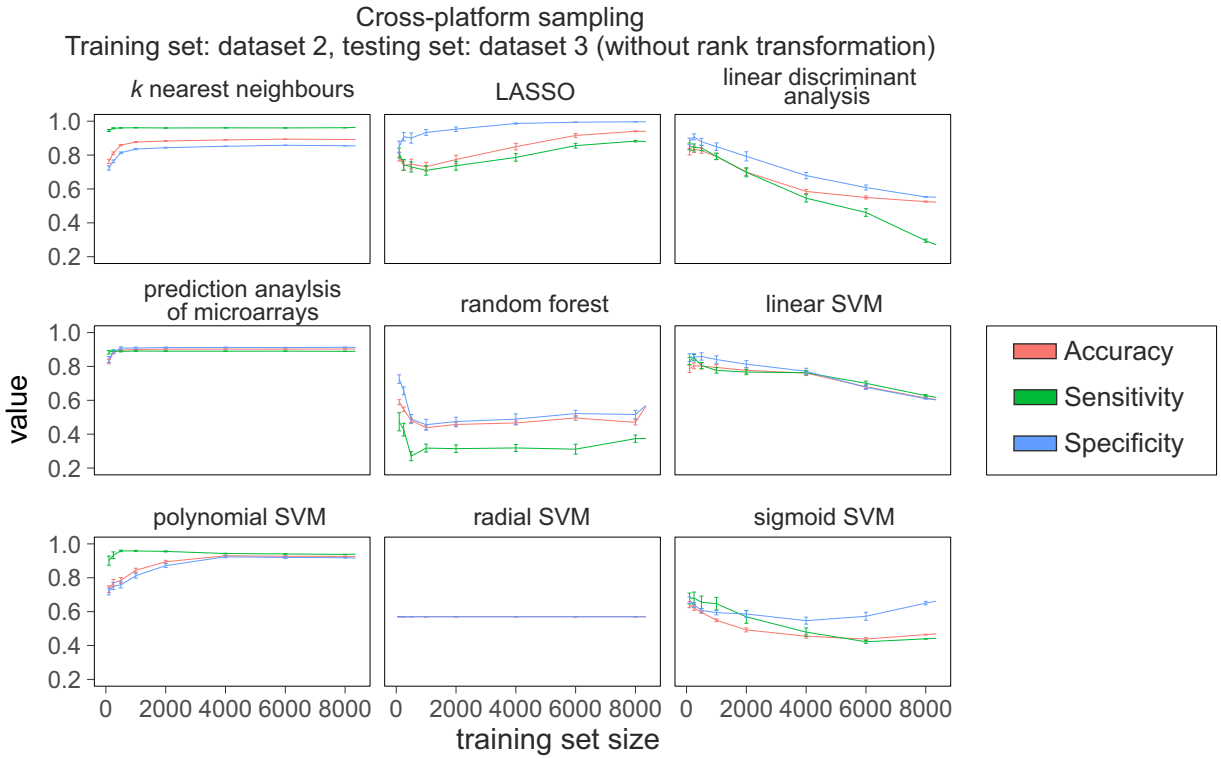


Figure S11

A



B



C

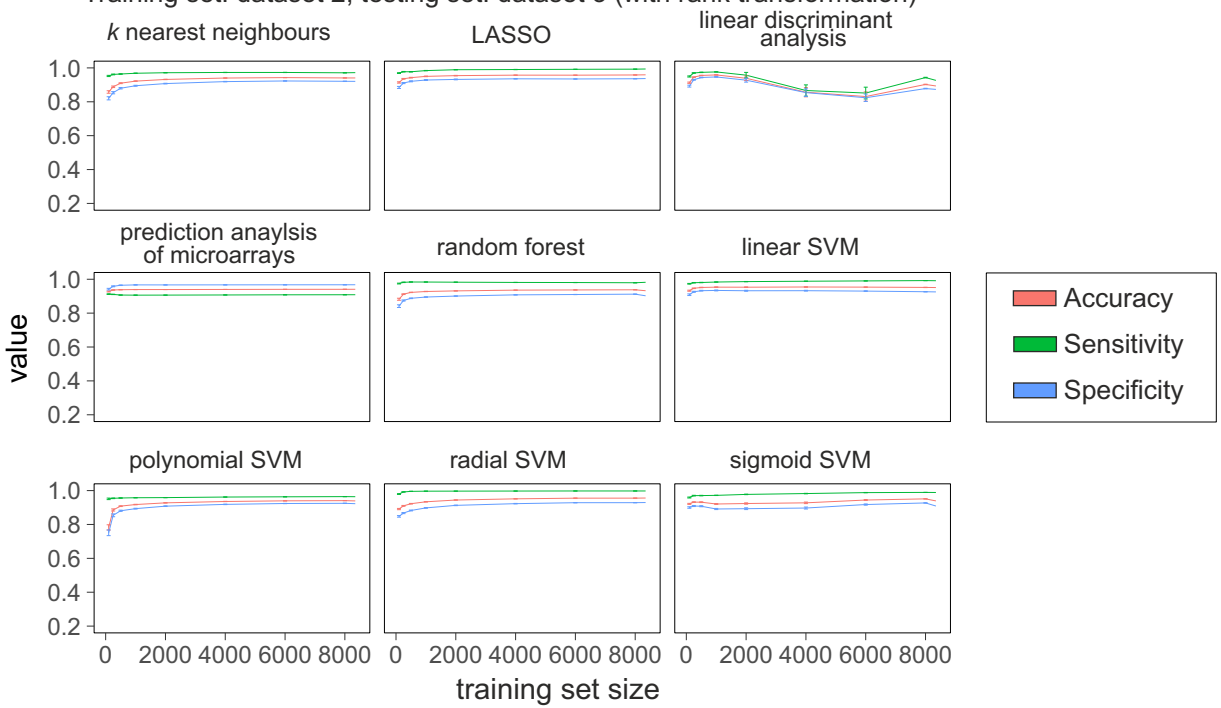
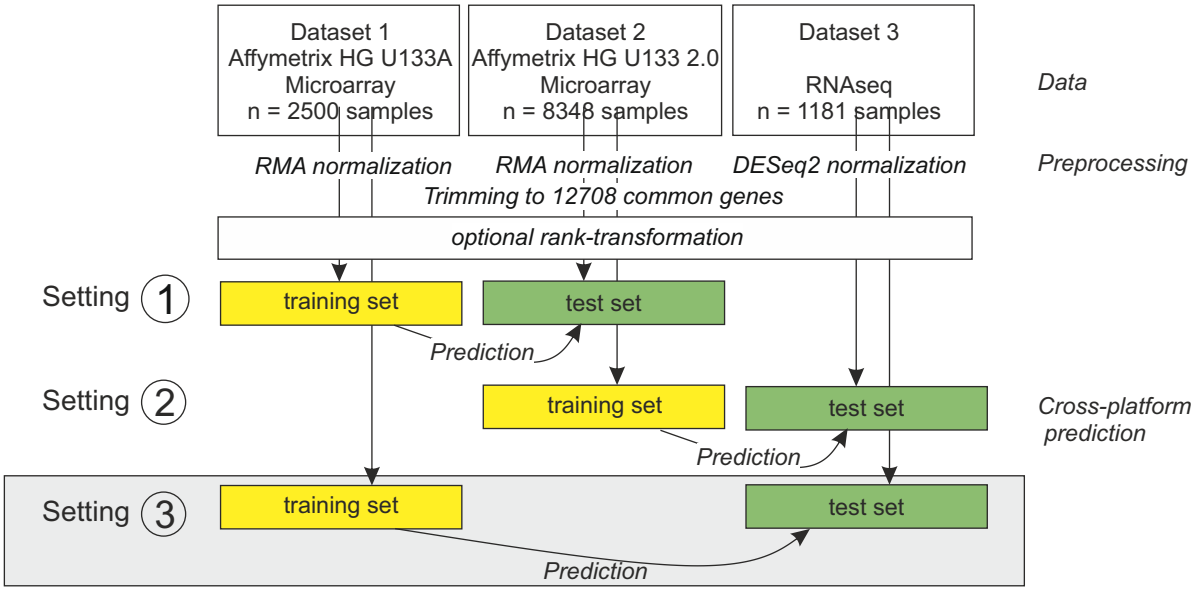
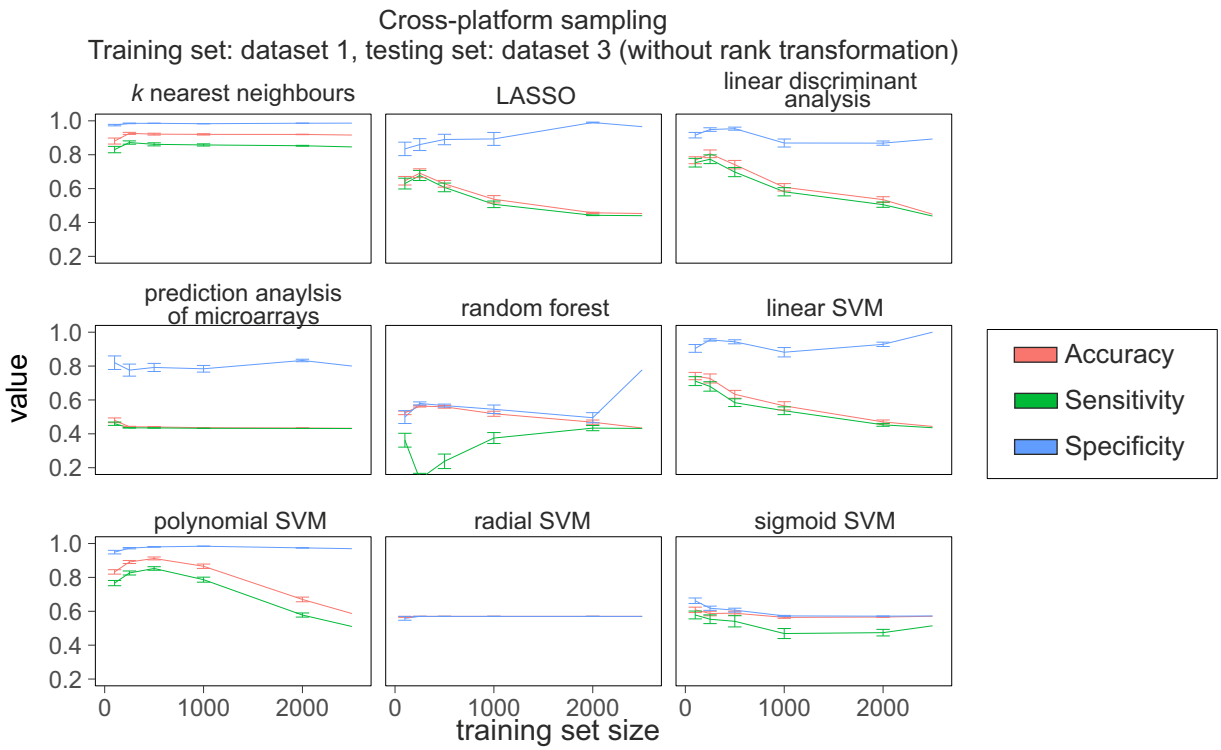


Figure S12

A



B



C

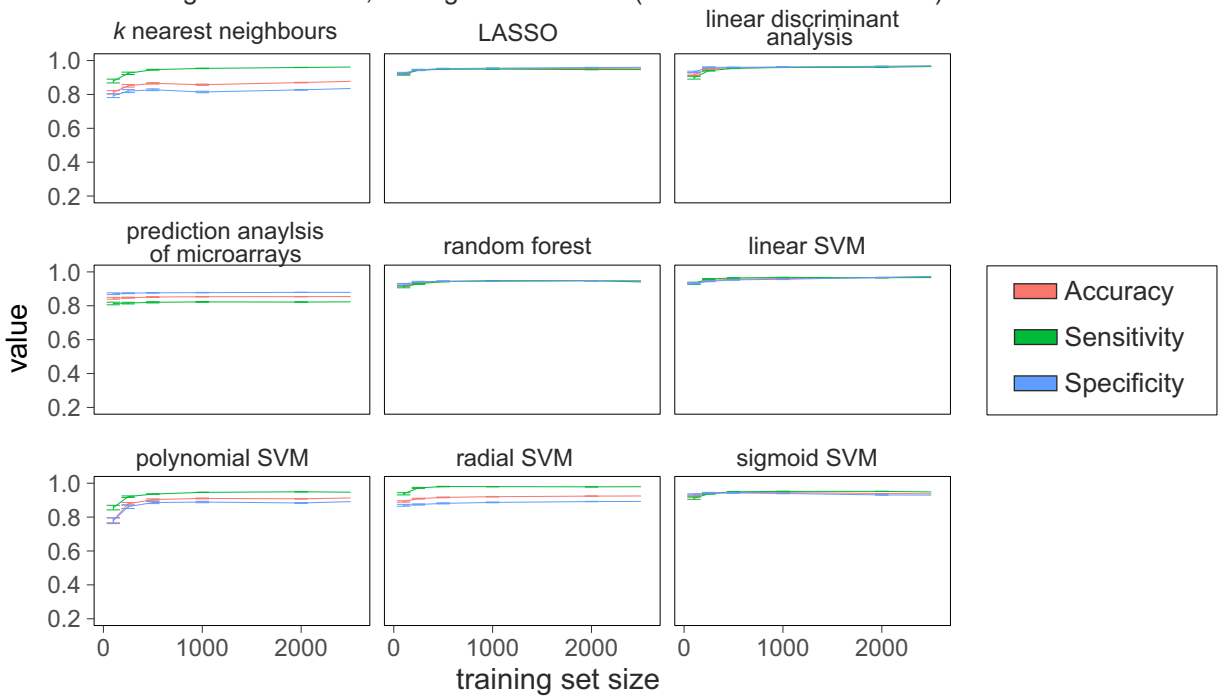
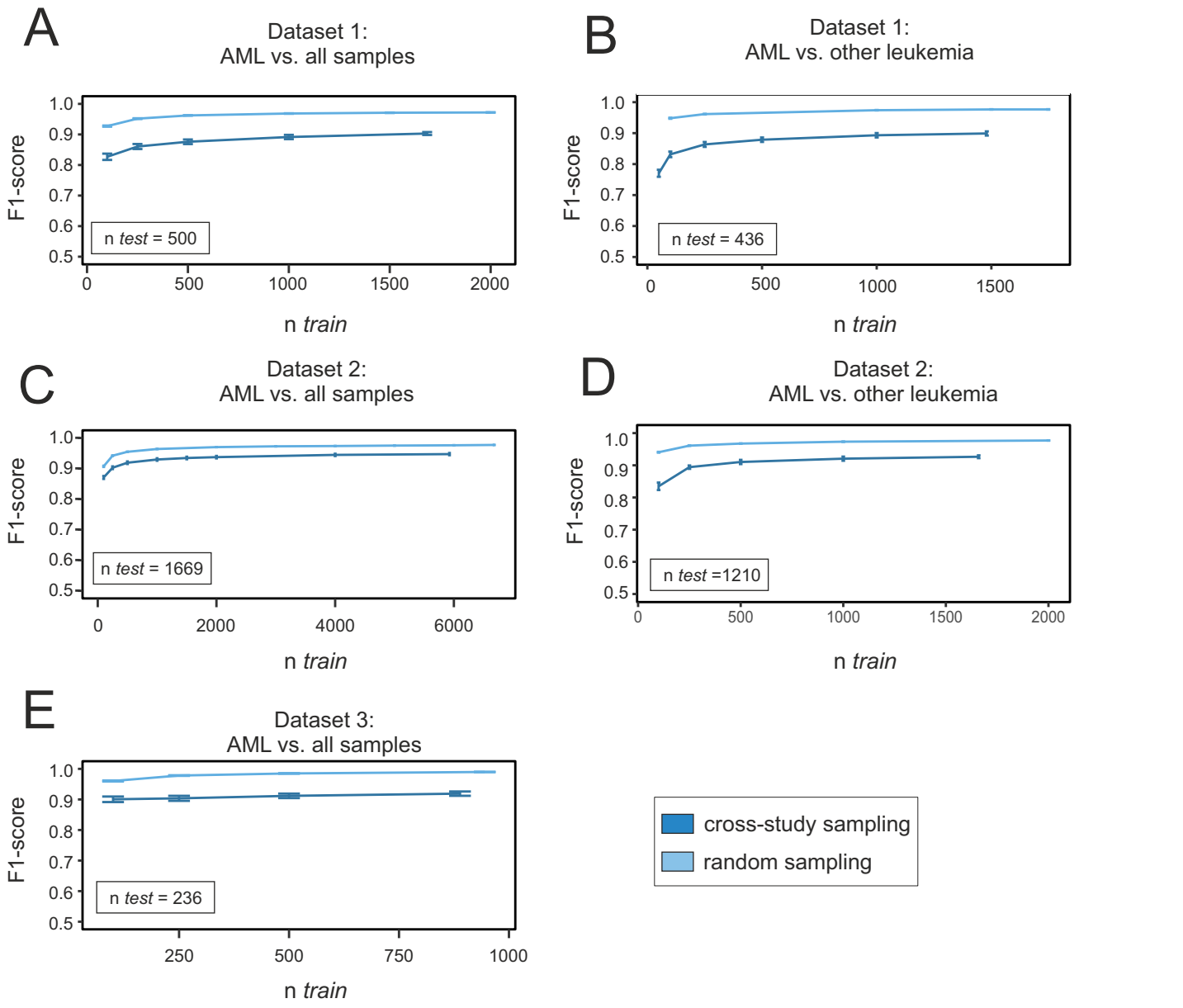


Figure S13


A
Dataset 1:
AML vs. all samples
F1-score
n test = 500
n train
B
Dataset 1:
AML vs. other leukemia
F1-score
n test = 436
n train
C
Dataset 2:
AML vs. all samples
F1-score
n test = 1669
n train
D
Dataset 2:
AML vs. other leukemia
F1-score
n test =1210
n train
E
Dataset 3:
AML vs. all samples
F1-score
n test = 236
n train
cross-study sampling
random sampling

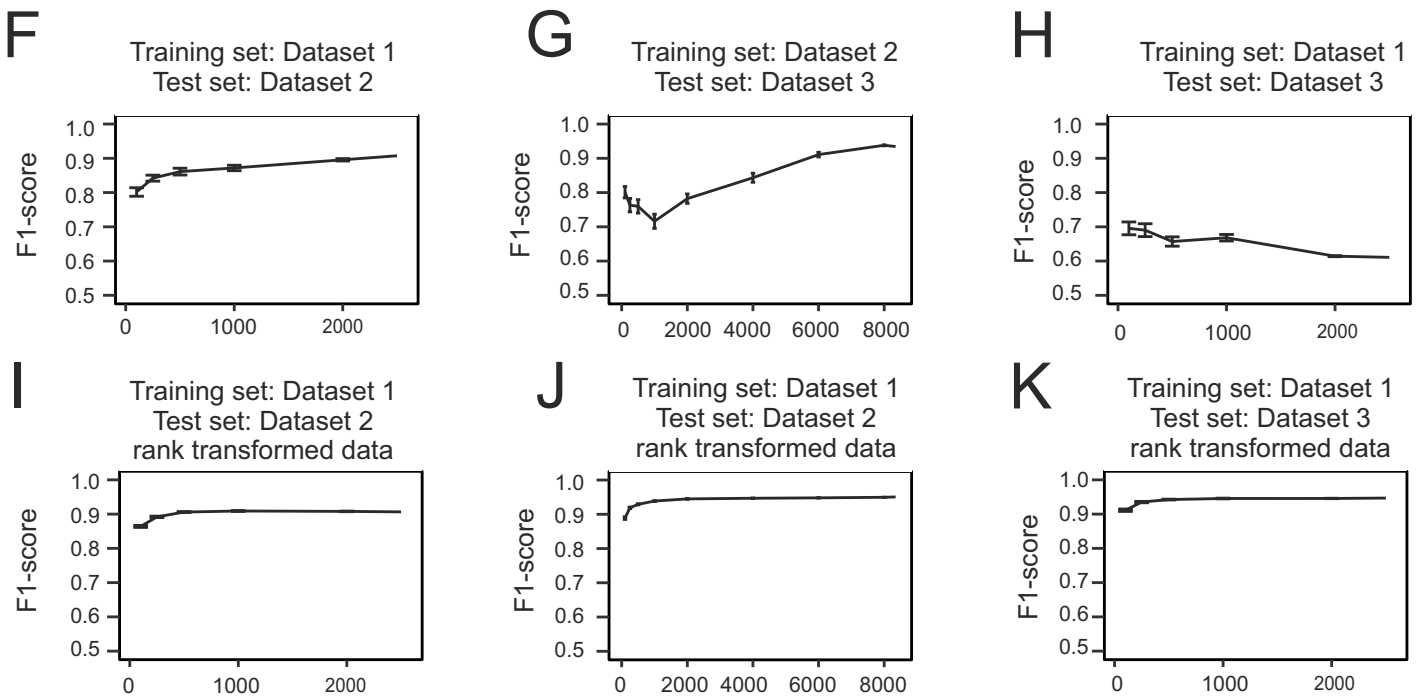
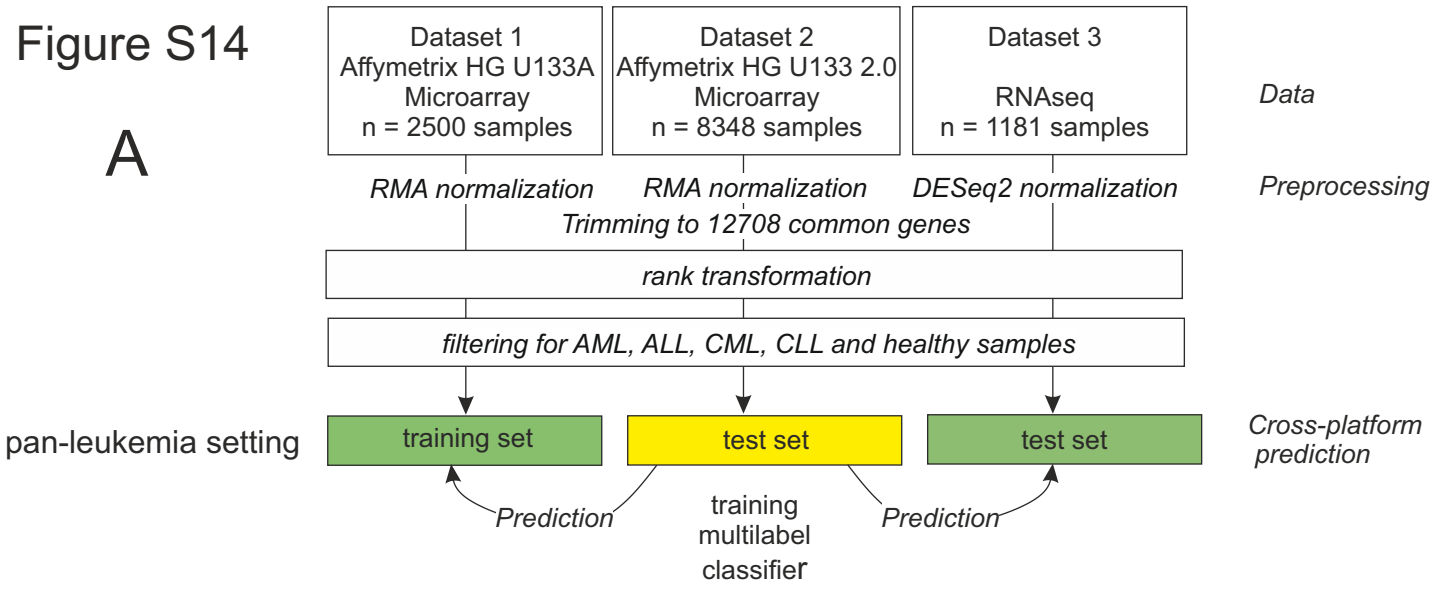
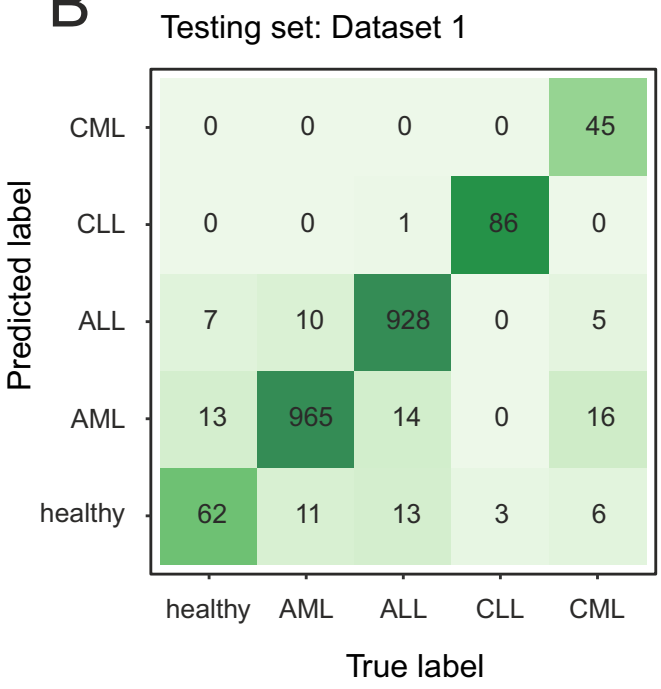

F
Training set: Dataset 1
Test set: Dataset 2
G
Training set: Dataset 2
Test set: Dataset 3
H
Training set: Dataset 1
Test set: Dataset 3
I
Training set: Dataset 1
Test set: Dataset 2
rank transformed data
J
Training set: Dataset 1
Test set: Dataset 2
rank transformed data
K
Training set: Dataset 1
Test set: Dataset 3
rank transformed data
F1-score

Figure S14

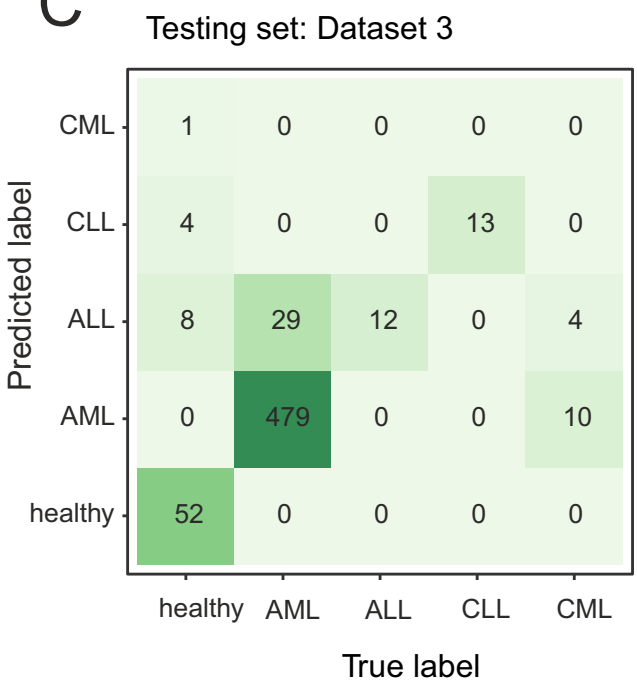
A



B



C



D

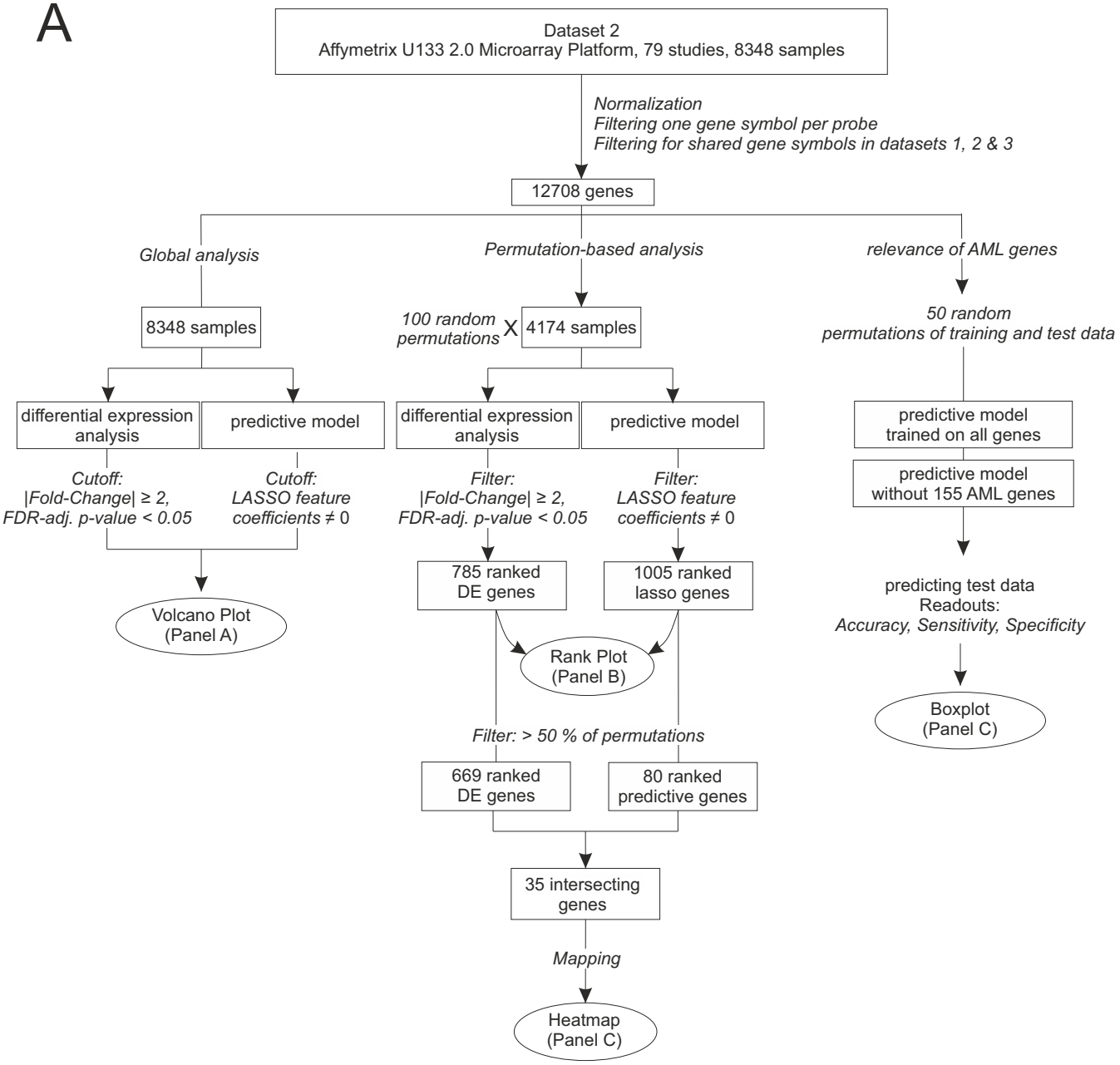
	healthy	AML	ALL	CLL	CML
bal. Accuracy	0.87	0.97	0.98	0.98	0.81
Sensitivity	0.76	0.98	0.97	0.97	0.63
Specificity	0.98	0.96	0.98	>0.99	0.63

E

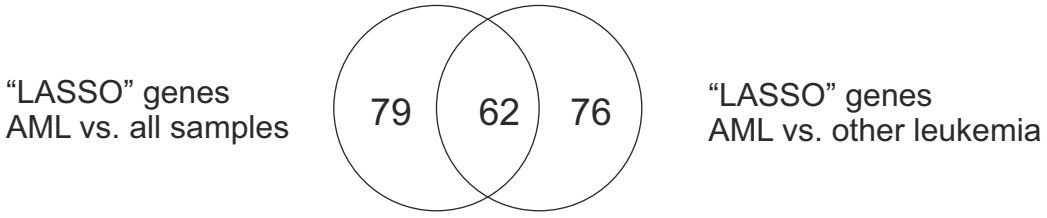
	healthy	AML	ALL	CLL	CML
bal. Accuracy	0.90	0.92	0.97	0.99	0.49
Sensitivity	0.80	0.94	1.00	1.00	0
Specificity	1.00	0.90	0.93	0.99	0.99

Figure S15

A



B



Supplemental Figure Legends

Figure S1: Sample overview, related to Figure 2

(**A**) Overview of the sample and study composition of all three datasets. The GSE number and the number of samples per disease are depicted for each study.

Figure S2: Comparison of bone marrow and PBMC samples, related to Figure 1

(**A**) Workflow: Dataset 2 was used to sample bone marrow and PBMC samples of AML patients and controls in equal numbers. **(B)** The resulting dataset of 332 samples was scaled and gene expression values of the top 25% variable genes were clustered and shown in a dendrogram.

Figure S3: Prediction of AML in random sampling scenarios (dataset 1), related to Figure 2

(**A**) Workflow: Dataset 1 (Affymetrix HG-U133 A) was RMA normalized and subjected to 100 times random sampling of training and test data, with training data samples from n_{train} = 100 to n_{train} = 2000 samples and test data of n_{test} = 500 samples. (**B**) Accuracy, sensitivity and specificity for nine different prediction algorithms on the whole dataset 1. (**C**) Accuracy, sensitivity and specificity for nine different prediction algorithms on leukemia samples of dataset 1 (AML, ALL, CML, CLL, MDS and down syndrome transient myeloproliferative disorder). Errorbars depict the standard deviation.

Figure S4: Prediction of AML in random sampling scenarios (dataset 2), related to Figure 2

(**A**) Workflow: Dataset 2 (Affymetrix HG-U133 2.0) was RMA normalized and subjected to 100 times random sampling of training and test data, with training data samples from n_{train} = 100 to n_{train} = 6679 samples and test data of n_{test} = 1669 samples. (**B**) Accuracy, sensitivity and specificity for nine different prediction algorithms on the whole dataset 2. (**C**) Accuracy, sensitivity and specificity for nine different prediction algorithms on leukemia samples of dataset 1 (AML, ALL, CML, CLL, MDS). Errorbars depict the standard deviation.

Figure S5: Prediction of AML in random sampling scenarios (dataset 3), related to Figure 2

(**A**) Workflow: Dataset 3 (RNA-seq) was normalized using DESeq2 and subjected to 100 times random sampling of training and test data, with training data samples from n_{train} = 100 to n_{train} = 945 samples and test data of n_{test} = 236 samples. (**B**) Accuracy, sensitivity and specificity for nine different prediction algorithms on the whole dataset 3. Prediction of leukemia samples only was not possible due to small sample sizes (see Figure S1). Errorbars depict the standard deviation.

Figure S6: Prediction of AML in cross-study sampling scenarios (dataset 1), related to Figure 2

(**A**) Workflow: Dataset 1 (Affymetrix HG-U133 A) was RMA normalized and subjected to 100 times cross-study sampling of training and test data. (**B**) Accuracy, sensitivity and specificity for nine different prediction algorithms on cross-study sampling on the whole dataset 1, with training data samples from n_{train} = 100 to n_{train} = 1865 (mean) samples and test data of n_{test} = 500 samples. (**C**) Accuracy, sensitivity and specificity for nine different prediction algorithms on cross-study sampling of leukemia samples of dataset 1 (AML, ALL, CML, CLL, MDS and down syndrome transient myeloproliferative disorder), with training data samples from n_{train} = 100 to n_{train} = 1480 (mean) samples and test data of n_{test} = 436 samples. Errorbars depict the standard deviation.

Figure S7: Effective prediction of AML in cross-study sampling scenarios (dataset 2), related to Figure 2

(**A**) Workflow: Dataset 2 (Affymetrix HG-U133 2.0) was RMA normalized and subjected to 100 times cross-study sampling of training and test data. (**B**) Accuracy, sensitivity and specificity for nine different prediction algorithms on cross-study sampling on the whole dataset 1, with training data samples from n_{train} = 100 to n_{train} = 5926 (mean) samples and test data of n_{test} = 1669 samples. (**C**) Accuracy, sensitivity and specificity for nine different prediction algorithms on cross-study sampling of leukemia samples of dataset 1 (AML, ALL, CML, CLL and MDS), with training data samples from n_{train} = 100 to n_{train} = 1750 (mean) samples and test data of n_{test} = 1210 samples. Errorbars depict the standard deviation.

Figure S8: Effective prediction of AML in cross-study sampling scenarios (dataset 3), related to Figure 2

(**A**) Workflow: Dataset 3 (RNA-seq) was normalized using DESeq2 and subjected to 100 times cross-study sampling of training and test data. (**B**) Accuracy, sensitivity and specificity for nine different prediction algorithms on cross-study sampling on the whole dataset 3, with training data samples from n_{train} = 100 to n_{train} = 889 (mean) samples and test data of n_{test} = 236 samples. Prediction of leukemia samples only was not possible due to small sample sizes (see Figure S1). Errorbars depict the standard deviation.

Figure S9: Addon RMA normalization, related to Figure 2

(A) Schema for addon RMA normalization on dataset 1. The 2500 samples were subjected to 50 times cross-study sampling, which corresponds to the first 50 permutations in Figure 5SA. Different to the aforementioned approach, the data was not normalized beforehand, but after splitting the samples into training and test data. Training data was RMA-normalized and testing data was normalized “onto” the training data using addon normalization. **(B)** Accuracy, sensitivity and specificity of addon normalization as shown in **(A)** (light blue), compared to performance of the “standard” cross-study sampling approach as described in Figure 5SA.

Figure S10: Translating predictive signature across technological platforms (setting 1), related to Figure 4

(**A**) Workflow: Datasets were normalized individually and trimmed to 12,708 common genes. The predictors were trained on subsamples of different sizes on dataset 1 and tested on all samples of dataset 2. (**B**) Accuracy, sensitivity and specificity of lasso prediction trained on dataset 1 with training sample size from n_{train} = 100 to n_{train} = 2500 and tested on the full dataset 2 (n_{test} = 8348). (**C**) Accuracy, sensitivity and specificity of lasso prediction trained on rank transformed dataset 1 with training sample size from n_{train} = 100 to n_{train} = 2500 and tested on the full dataset 2 (n_{test} = 8348, rank transformed). Errorbars depict the standard deviation.

Figure S11: Translating predictive signature across technological platforms (setting 2), related to Figure 4

(A) Workflow: Datasets were normalized individually and trimmed to 12708 common genes. The predictors were trained on subsamples of different sizes on dataset 2 and tested on all samples of dataset 3. (**B**) Accuracy, sensitivity and specificity of lasso prediction trained on dataset 2 with training sample size from n_{train} = 100 to n_{train} = 8348 and tested on the full dataset 3 (n_{test} = 1181). (**C**) Accuracy, sensitivity and specificity of lasso prediction trained on rank transformed dataset 2 with training sample size from n_{train} = 100 to n_{train} = 8348 and tested on the full dataset 3 (n_{test} = 1181, rank transformed). Errorbars depict the standard deviation.

Figure S12: Translating predictive signature across technological platforms (setting 3), related to Figure 4

(**A**) Workflow: Datasets were normalized individually and trimmed to 12708 common genes. The predictors were trained on subsamples of different sizes on dataset 1 and tested on all samples of dataset 3. (**B**) Accuracy, sensitivity and specificity of lasso prediction trained on dataset 1 with training sample size from n_{train} = 100 to n_{train} = 2500 and tested on the full dataset 3 (n_{test} = 1181). (**C**) Accuracy, sensitivity and specificity of lasso prediction trained on rank transformed dataset 1 with training sample size from n_{train} = 100 to n_{train} = 2500 and tested on the full dataset 3 (n_{test} = 1181, rank transformed). Errorbars depict the standard deviation.

Figure S13: F1 scores of AML prediction in random sampling, cross-study and cross-platform scenarios, related to Figures 2 and 5

F1 scores of prediction results in random and cross-study sampling scenarios in dataset 1, all samples (A), dataset 1, leukemia samples only (B), dataset 2, all samples (C), dataset 2, leukemia samples only (D), and dataset 3, all samples (E). F1 scores for cross-platform prediction results for the settings depicted in Figure 5. (F-K).

Figure S14: Pan-leukemia classification across platforms, related to Figure 4

(**A**) Workflow: Datasets were normalized individually and trimmed to 12708 common genes and samples were filtered to include only AML, ALL, CML, CLL and healthy samples. A multilabel logistic regression model was fit on dataset 2 and then tested on the independently normalized datasets 1 and 3. **(B,C)** Confusion matrices comparing predicted labels to true labels for all tested leukemia types for testing on dataset 1 and 3, respectively. **(D,E)** Balanced accuracy, sensitivity and specificity of the multiclass prediction on dataset 1 and 3.

Figure S15: Workflow: Comparing differentially expressed and predictive genes, related to Figure 5

(**A**) Workflow to Figure 5: Dataset 2 was used to compare DE and the sparse predictive models. First, a global analysis of DE genes and lasso genes was performed and visualized in a heatmap. Second, dataset 2 was permuted and 35 genes that appeared at least 50 out of 100 times as “DE gene” or “lasso gene” were visualized in a heatmap. Third, predictive signatures were trained on all 12708 genes, with and without 155 known AML genes (genes included in DO and KEGG terms). Results were visualized in a boxplot. **(B)** Comparison of “lasso genes” of the prediction AML vs. all samples and AML vs. other leukemia samples of dataset 2 (same prediction setting as in Figures 2D, E).

Transparent Methods

Study search strategy

All data sets published in the National Center for Biotechnology Information Gene Expression Omnibus (GEO, (Edgar, 2002)) on 20 September 2017 were reviewed for inclusion in the present study. Basic criteria for inclusion were the cell type under study (human peripheral blood mononuclear cells (PMBCs) and/or bone marrow samples) as well as the species (*Homo sapiens*). Both tissues are considered equivalent in the diagnosis of AML. We compared bone marrow and PBMC samples of dataset 2 and did not identify overall differences in gene expression (Figure S2) and therefore did not differentiate between bone marrow and PBMC samples throughout the study. Furthermore, we excluded GEO SuperSeries to avoid duplicated samples (Table S1). We filtered the datasets for data generated with Affymetrix HG-U133 A microarrays, Affymetrix HG-U133 2.0 microarrays and high-throughput RNA sequencing (RNA-seq) and excluded studies with very small sample sizes (< 50 samples for microarray and < 10 samples for RNA-seq data). We then applied a disease-specific search, in which we filtered for acute myeloid leukemia, other leukemia and healthy or non-leukemia-related samples.

The results of this search strategy were then internally reviewed and data were excluded based on the following criteria: (i) exclusion of duplicated samples, (ii) exclusion of studies that sorted single cell types (e.g. T cells or B cells) prior to gene expression profiling, (iii) exclusion of studies with inaccessible data. Other than that, no studies were excluded from our analysis (see also Table S1). In addition, we included one unpublished dataset (in dataset 1). The above steps gave rise to the data referred to above as **dataset 1** (Affymetrix HG-U133 A microarrays), **dataset 2** (Affymetrix HG-U133 2.0 microarrays) and **dataset 3** (RNA-seq). The RNA-seq data contained was not filtered for any particular protocol and contained paired and well as single-end data of different sequencing depth. AML subtype annotations were taken from the respective metadata-files on GEO. Subgroups of FAB-classifications were combined to represent the major FAB class (e.g. AML M3 and AML M3v were combined to AML M3).

Pre-processing

All raw data files were downloaded from GEO. For normalization, we considered all platforms independently, meaning that normalization was performed separately for the samples in dataset 1, 2 and 3, respectively. Microarray data (datasets 1 and 2) were normalized using the robust multichip average (RMA) expression measures (Irizarry et al., 2003), as implemented in the R package affy (Gautier et al., 2004). RNA-seq data (dataset 3) was preprocessed using kallisto (Bray et al., 2016) and normalized with the R package DESeq2 using standard parameters (Love et al., 2014). In order to keep the datasets comparable, we filtered the data for genes annotated in all three datasets, which resulted in 12,708 genes. No filtering of low-expressed genes was performed. All scripts used in this study for pre-processing are provided as a docker container on Docker Hub (https://hub.docker.com/r/schultzelab/aml_classifier).

Prediction

Prior to classification, data sets were split into non-overlapping training and test data. For the comparisons of AML vs. all samples, all non-AML samples were used as controls, which would in clinical terms, reflect finding a diagnosis. For the prediction of AML vs. other leukemia, all non-AML leukemias, namely chronic myeloid leukemia (CML), acute lymphoblastic leukemia (ALL), chronic lymphoblastic leukemia (CLL), Myelodysplastic syndrome (MDS) and down syndrome transient myeloproliferative disorder were used as non-AML labels, which would be the equivalent of finding a differential diagnosis between different leukemias. All main classification tasks were performed in the programming language R (R Core Team, 2016). All main results were obtained using l_1-penalized logistic regression using the package glmnet (Friedman et al., 2010). Non-zero coefficients were extracted for feature ranking (Figure 4). The regularization parameter was set using 10-fold cross-validation (using training set data only). To assess predictive performance, accuracy, sensitivity, specificity and F1 score were calculated as well as positive predictive value (PPV) under several prevalence scenarios. For assessing the performance of support vector machines (SVMs), we used the R package e1071 for SVMs (linear, radial, polynomial and sigmoid kernels) (Meyer et al., 2015). The R package randomForest was used for random forest classification (Shi et al., 2004). K nearest neighbors classification was done using the knn function implemented in the class package in R (Venables and Ripley, 2002). Linear discriminant analysis was performed with the lda function implemented in the R package MASS (Venables and Ripley, 2002). For RNA-seq data, features with zero variance were excluded for LDA. Prediction analysis of microarrays was done with the pamr package

(Hastie et al., 2014). Neural networks were built using Keras (Chollet et al., 2017) with a Tensorflow backend (10 layers, ~7×10^6 parameters). Unless otherwise noted, default settings were used for tuning parameters as implemented in the respective packages.

Rank transformation to normality

As an example of a simple data transformation that would facilitate translation between gene expression platforms, we performed a rank transformation to normality. For this, gene expression values were transformed from microarray intensities (dataset 1 & 2) or RNAseq counts to their respective ranks. This was done gene-wise, meaning all gene expression values per gene were given a rank based on ordering them from lowest to highest value. The rankings were then turned into quantiles and transformed via the inverse cumulative distribution function of the Normal distribution. This leads to all genes following the exact same distribution (that is, a standard Normal with a mean of 0 and a standard deviation of 1) across all samples (Zwiener et al., 2014).

Differential expression analysis

For differential expression analysis of dataset 2 the R package limma was used (Ritchie et al., 2015). A linear model was fit on the data with inclusion of the study as a factor. Differentially expressed genes were called using an FDR-corrected p-value < 0.05 and a minimum fold change of +/- 2. For the permutation-based approach, 4174 samples were randomly drawn 100 times from the dataset. In each subset, DE genes were called as before, but without correcting for any batch in the model. The number of times each gene was called was summed up over all 100 permutations. Genes were ranked according to their overall DE count.

In addition to that a l_1-penalized logistic regression was performed using the package glmnet (Friedman et al., 2010) on the whole dataset and on each of the permutations. Genes were called to be of predictive importance if features had non-zero coefficients. The number of times each feature was of predictive importance was summed up, which resulted in a feature ranking of all “lasso genes”.

Hierarchical Clustering

35 genes which had a stability of > 50% over 100 permutations for lasso and DE genes were visualized using the R package pheatmap (Kolde, 2015) (Figure 6B). The data was z-scaled and columns clustered according to Euclidean distance. Rows were ordered according to diseases. Two gene clusters were visualized.

Exclusion of gene sets from prediction

In order to evaluate the robustness of our classification results (Figure 6C), we excluded 155 genes present in either the KEGG or the disease ontology term “Acute Myeloid Leukemia” and compared this to the results achieved when all 12078 genes of the dataset are included (random sampling, dataset 2).

Supplemental references

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

Chollet, F., Allaire, J.J., and others (2017). R Interface to Keras.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. *33*, 1–22.

Hastie, T., Tibshirani, R., Narasimhan, B., and Chu, G. (2014). pamr: Pam: prediction analysis for microarrays.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249–264.

Kolde, R. (2015). pheatmap: Pretty Heatmaps.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2015). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.

R Core Team (2016). R: A Language and Environment for Statistical Computing.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Shi, T., Seligson, D., Belldegrun, A.S., Palotie, A., and Horvath, S. (2004). Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. Mod Pathol *18*, 547–557.

Venables, W.N., and Ripley, B.D. (2002). Modern Applied Statistics with S (Springer).