

METHOD

Open Access



# Flimma: a federated and privacy-aware tool for differential gene expression analysis

Olga Zolotareva<sup>1,2\*†</sup> , Reza Nasirigerdeh<sup>3,8†</sup>, Julian Matschinske<sup>2</sup>, Reihaneh Torkzadehmahani<sup>3</sup>, Mohammad Bakhtiari<sup>2</sup>, Tobias Frisch<sup>4</sup>, Julian Späth<sup>2</sup>, David B. Blumenthal<sup>5</sup>, Amir Abbasinejad<sup>1,7</sup>, Paolo Tieri<sup>6,7</sup>, Georgios Kaissis<sup>3,8,9,10</sup>, Daniel Rückert<sup>3,8,9</sup>, Nina K. Wenke<sup>2</sup>, Markus List<sup>1</sup> and Jan Baumbach<sup>2,4</sup>

\*Correspondence:

[olya.zolotareva@gmail.com](mailto:olya.zolotareva@gmail.com)

†Olga Zolotareva and Reza Nasirigerdeh contributed equally to this work.

<sup>1</sup>Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Freising, Germany

<sup>2</sup>Institute for Computational Systems Biology, University of Hamburg, Hamburg, Germany  
Full list of author information is available at the end of the article

## Abstract

Aggregating transcriptomics data across hospitals can increase sensitivity and robustness of differential expression analyses, yielding deeper clinical insights. As data exchange is often restricted by privacy legislation, meta-analyses are frequently employed to pool local results. However, the accuracy might drop if class labels are inhomogeneously distributed among cohorts. *Flimma* (<https://exbio.wzw.tum.de/flimma/>) addresses this issue by implementing the state-of-the-art workflow *limma voom* in a federated manner, i.e., patient data never leaves its source site. *Flimma* results are identical to those generated by *limma voom* on aggregated datasets even in imbalanced scenarios where meta-analysis approaches fail.

**Keywords:** Differential expression analysis, Federated learning, Privacy of biomedical data, Meta-analysis

## Background

The identification of differentially expressed genes or transcripts, e.g., in diseases or in response to treatment, is a standard but important task in molecular systems medicine.

Differential gene expression analysis compares the expression profiles of two or more groups of samples to reveal genes with significant differences between the groups. Technologies for high-throughput gene expression profiling include microarrays and RNA sequencing, the latter being more widely used in clinical research today. Both are intrinsically different, e.g., signal- vs. count-based measurement, and their results subject to platform-specific biases [1, 2]. Many bioinformatics tools for identifying differentially expressed genes from such data have been developed [3–9]. These methods differ with respect to the assumptions about data distribution (e.g., normal vs. Poisson or negative binomial distribution), the data normalization strategies, and in the test statistic used to detect differentially expressed genes [10–12].



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

One major challenge of differential expression studies is the lack of robustness due to the high technical and biological variability of the data [13–15], which can be addressed using various strategies [16–18]. The simplest and most effective way would be to increase the sample size [16], which is non-trivial, as data collection is expensive and time-consuming, sample availability may be limited (e.g. metastatic cancer or healthy tissue samples are difficult to obtain), or because existing data can not be shared and pooled as they are subject to personal data protection laws. The latter is of particular concern for next-generation sequencing data, from which the sample donor can be identified under certain conditions [19–21]. Although several human-derived expression profiles are nowadays publicly available, their utility (in particular in clinical settings) is often still limited for inherent privacy issues. The statistical analysis of expression data may require relevant clinical metadata, e.g., patient sex, age, weight, ethnicity, and disease status, which may be identifying when combined. In addition, recent works suggest that patient genotypes can be predicted from RNA-seq data, making patients identifiable through expression profiles or eQTL data obtained from open-access sources [22–24]. Schadt et al. have shown that genotypes can be inferred from expression levels of eQTL-controlled genes and sensitive information — for instance, medical history, phenotypic traits, and family relationships can be revealed [22]. Matching predicted genotypes to known ones allowed for identifying individuals with an accuracy of up to 99% in optimal settings, i.e., when the microarray platform, tissue type, and ancestry were the same for expression and eQTL datasets. Harmanci and Gerstein proposed a measure of individual-characterizing information leakage and investigated its dependence on genotype predictability given the expression dataset [23]. They developed a framework to assess privacy risks before publishing the data. They have also presented a simplified but effective attack scheme where homozygous genotypes were predicted from extreme gene expression values.

To control the exchange of sensitive molecular profiling data from, e.g., next-generation sequencing experiments, databases, such as dbGaP [25] or EGA [26], restrict access to authorized users affiliated with organizations willing to guarantee the legal and secure use of personal data. Nevertheless, the application procedure needs to be repeated per study and per database, making this a difficult and time-consuming process, which is also error-prone if a priori unknown confounder variables are not requested and can thus not be corrected for in the downstream analysis. Alternatively, when direct access to raw data is not possible, researchers can combine the results of several studies using meta-analysis techniques such as Fisher's method [27], Stouffer's method [28], RankProd [29], or the random effects model [30] (REM). Meta-analysis is widely adopted for aggregation of genome-wide association studies (GWAS) [31] and differential gene expression analysis results [32, 33] (cf. the “[Meta-analysis approaches](#)” section in the “[Methods](#)” section for details). The main disadvantage of meta-analysis tools is that their underlying assumptions about the distribution of  $p$ -values or effect sizes may not be realistic. Furthermore, meta-analysis largely ignores possible differences between cohorts (e.g., class imbalance or heterogeneity of covariate distributions) [34] or data processing steps (e.g., normalization) [35], which may have a significant impact on the results [34].

Privacy-aware techniques, such as federated learning (FL) [36], differential privacy (DP) [37], homomorphic encryption (HE) [38], and secure multi-party computation (SMPC) [39], have recently moved into the focus of research for tasks involving privacy-sensitive patient data [40]. Note that in this paper, the term *privacy-aware* [41] designates the

techniques that avoid sharing raw personal data between collaborating parties. Such approaches are usually not in conflict with privacy legislation and may thus legally and practically be applied to real-world medical data. We call such a privacy-aware approach *privacy-preserving* (e.g., DP [37]), if it provides a formally proven privacy guarantee that captures the risks associated with each sample of the dataset.

FL has become increasingly popular in bioinformatics for GWAS [42, 43], survival analysis [44], and additional challenges in patient data processing [40, 45]. FL implies collaborative model training by multiple participants without disclosing private data to any other party [36]. Instead, each participant only shares intermediate model parameters while keeping the private data in the local environment (e.g. the legally safe harbors of the hospitals' IT system). The local parameters from the clients are aggregated at the server iteratively to compute a globally optimal model.

DP perturbs the data or results by adding noise to them. Although DP is privacy-preserving and complementary to FL, it might dramatically impact the accuracy of the results. HE performs computation on the encrypted data from the participants. It suffers from two practical disadvantages [46]: it supports a limited number of operations such as addition and multiplication, and consequently, it requires approximations to compute non-linear operations (e.g., computing the inverse of covariance matrix in gene expression analysis), leading to accuracy loss in the final results. More importantly, it is computationally expensive because a single machine performs operations on a large amount of encrypted data and might require a sizable amount of memory to process large datasets [47]. In SMPC, each participant computes secret shares from the data and shares them with the computing parties. The computing parties calculate the intermediate results and exchange them among each other to compute the final results. Because SMPC-based methods send secret shares of the data from the participants to computing parties, they consume a huge amount of network traffic [48].

FL is a promising alternative to SMPC and HE in terms of performance and scalability. Unlike HE, it does not increase computational cost much compared to the centralized method, and unlike SMPC, it transfers only a small number of model parameters through the network. Similar to HE, SMPC, and meta-analyses, FL is not privacy-preserving like DP, i.e., the server might reconstruct the raw data using the model parameters obtained from the clients [49–51]. However, the approaches based on pure HE and SMPC provide stronger privacy compared to FL-based approaches because they reveal less information to the third parties.

The privacy of federated methods can be enhanced by applying HE or SMPC on the shared model parameters. In comparison to purely SMPC- or HE-based methods, hybrid approaches are computationally efficient because heavy computations are distributed across the clients. Additionally, they offer enhanced privacy compared to pure FL, because the original values of the local parameters remain hidden from the server, and only global parameters are revealed to the server and the clients.

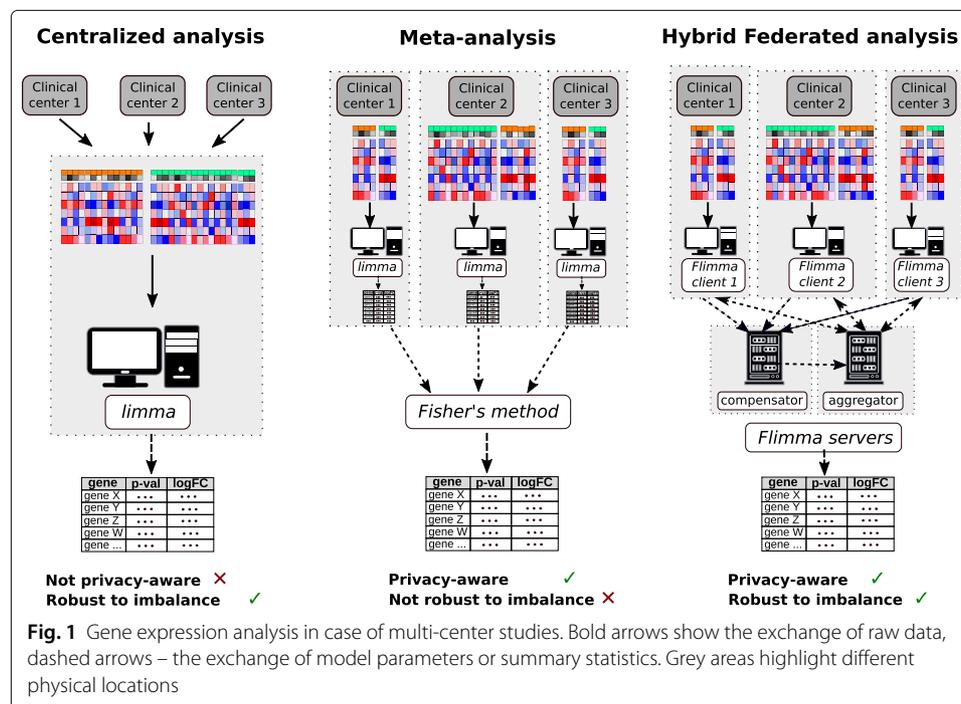
In this paper, we introduce *Flimma* (federated *limma*), a novel federated privacy-aware tool for the identification of differentially expressed genes called *Flimma*. Our new tool represents a federated implementation of the popular differential expression analysis workflow *limma voom* [52], one of the standard pipelines widely applied in the field for expression analyses. We have chosen *limma voom* among other popular count-based methods, because it is comparably fast without sacrificing accuracy [5]. Besides that,

*Flimma* could be easily modified for handling microarray data, since the *limma* method was originally designed for such data [53] and only later extended to RNA-seq data via *voom* [5].

*Flimma* is based on *HyFed* [54], a hybrid FL framework, which applies additive secret sharing-based SMPC method to avoid disclosing the local model parameters to the server (see the “Methods” for details). It provides several advantages over the existing approaches for gene expression analysis (Fig. 1). Unlike *limma voom*, *Flimma* enhances the privacy of the data in the cohorts since the expression profiles never leave the local execution sites and only aggregated parameters are revealed to the server and the other local sites. In contrast to meta-analysis approaches, *Flimma* is particularly robust against heterogeneous distributions of data (in particular of confounders and class labels) across the different cohorts, which makes it a powerful alternative for multi-center studies where patient privacy is a key concern.

### Results

We applied *Flimma* and four meta-analyses approaches on two real-world datasets: a breast cancer expression dataset from TCGA [55] and a skin dataset from GTEx [56]. To assess *Flimma*'s power, we model the multi-party setting by randomly partitioning both datasets into virtual cohorts, while introducing different levels of imbalance w.r.t. target class labels and covariate distributions. For both datasets, we simulated three realistic scenarios leading to different levels of sample distribution heterogeneity between local cohorts. We split the breast cancer dataset such that three virtual cohorts yield different frequencies of the LumA subtype to simulate an imbalanced distribution of disease subtypes collected at different clinical centers (Table 1). In addition, we partitioned the TCGA breast cancer dataset according to tissue source sites. Similarly, GTEx skin dataset was split by the mean ischemic time to illustrate the effect of potential confounders such as



**Table 1** Characteristics of three scenarios for the TCGA-BRCA dataset. The distributions of ages and tumor stages were balanced

	Cohort sizes			Frequency of basal subtype			Frequency of Luma subtype		
	Cohort 1	Cohort 2	Cohort 3	Cohort 1	Cohort 2	Cohort 3	Cohort 1	Cohort 2	Cohort 3
No imbalance	283	283	284	0.20	0.20	0.20	0.57	0.57	0.58
Mild imbalance	121	242	487	0.10	0.30	0.17	0.40	0.50	0.66
Strong imbalance	65	196	589	0.25	0.50	0.09	0.14	0.50	0.65

differences in sample collection and/or processing between the participating laboratories (Table 2).

We then compared *Flimma* with popular meta-analysis tools using the *limma voom* results on the pooled datasets as gold standard. In summary, *Flimma* obtained the same results as *limma voom* in all tests. Across all experiments, the maximal absolute difference for log-transformed  $p$ -values and log-fold-change values computed by *Flimma* and *limma voom* did not exceed 0.1 (Additional file 1: Table S1). In contrast, the results of the meta-analysis methods diverged from the results of *limma voom*, and this effect was especially pronounced in imbalanced scenarios.

One of the main pitfalls of gene expression analysis is the presence of strong batch effects in the data. Even for technical replicates, gene expression levels measured in two laboratories may drastically differ due to the difference in sample preparation and library construction protocols, sequencing platforms, chemical reagents, and many other known and unknown experimental factors. To demonstrate that *Flimma* is robust to experimental batch effects, we applied it to three independent breast cancer datasets generated at different laboratories.

#### Evaluation on artificial dataset splits

We compared negative log-transformed  $p$ -values computed by all privacy-aware approaches (i.e., *Flimma* and meta-analysis methods) with the results obtained by running *limma voom* on the combined dataset. For the privacy-aware approaches, we computed the root mean square error (RMSE), the precision, the recall, the F1 score, the Pearson and the Spearman correlation w.r.t. the results of the aggregated analysis with *limma voom*, which we treated as ground truth.

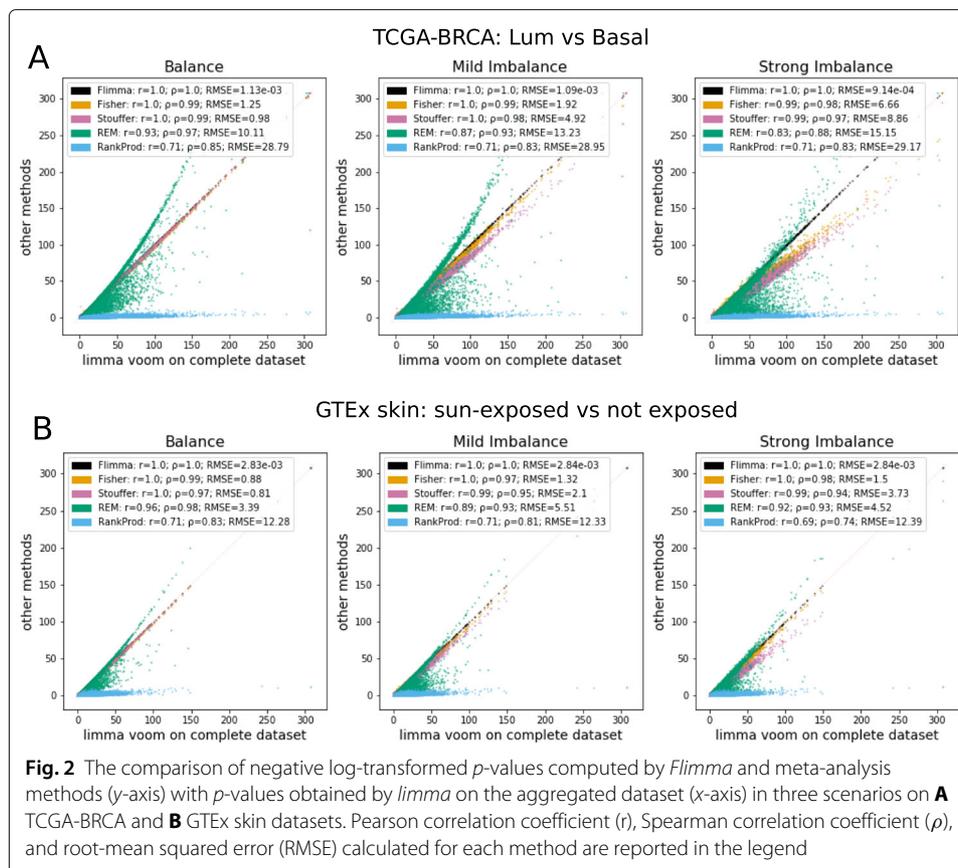
As shown in Fig. 2, Tables 3-4, Additional file 2: Table S2, and Additional file 3: Table S3, *Flimma* produces the same  $p$ -values as the aggregated analysis with *limma voom* in all scenarios, including the imbalanced ones. This implies that *Flimma* is robust against heterogeneous data distributions across the clients. However, this is not the case for the meta-analysis approaches. In general, their RMSEs increase (and Pearson correlations decrease) as the scenarios become more imbalanced, and they introduce false positives and false negatives even in the balanced scenario. In spite of the difference in  $p$ -values calculated by all meta-analysis methods, their gene rankings were quite similar to the ranking produced by the aggregated *limma voom* (the Spearman correlation varied between 0.74 to 0.99 in all experiments).

#### Performance for top-ranked genes

Since some research tasks such as biomarker discovery require the identification of a small number of significantly differentially expressed genes, we investigated how the performance of the methods varies with altered numbers of selected top differentially expressed genes after sorting by  $p$ -value (Fig. 3 and Additional file 4: Figures S1-2). Again, *Flimma* perfectly reproduced the results of aggregated *limma voom* in all scenarios and outperformed all meta-analysis approaches. Fisher's and Stouffer's methods demonstrated almost perfect performance in the balanced scenario, but their performance decreased in the imbalanced ones.

**Table 2** Characteristics of the scenarios for the GTEx skin dataset. The frequencies of samples obtained from male and female individuals were similar in all cohorts (between 30 and 34% samples from females in all scenarios)

	Cohort sizes			Fraction of sun-exposed skin samples			Mean ischemic time, min		
	Cohort 1	Cohort 2	Cohort 3	Cohort 1	Cohort 2	Cohort 3	Cohort 1	Cohort 2	Cohort 3
No imbalance	425	425	427	0.53	0.53	0.53	629	636	636
Mild imbalance	181	363	733	0.4	0.65	0.51	490	620	676
Strong imbalance	97	293	887	0.8	0.4	0.54	347	646	661



### Splitting TCGA-BRCA by sample source site

TCGA is a multi-center project and tumor samples of TCGA-BRCA datasets were collected at 37 different clinical centers, which can result in some between-center variability. Therefore, we also evaluated *Flimma* and its baselines on a more realistic scenario, where TCGA-BRCA dataset was split according to the sample source sites, but we kept only 14 of the 37 cohorts, such that each cohort contained at least 3 samples of LumA and basal subtype.

We selected 3, 5, 7, 10, and 14 cohorts such that subtype frequencies, mean stage, and age are dissimilar across the selected cohorts (cf. Additional file 5: Table S4 for details). We also added additional terms in linear models to account for possible cohort effects. Similar to the previous experiments, *Flimma* clearly outperforms all meta-analysis approaches in terms of RMSE, precision, and recall (Table 5 and Additional file 6: Table S5).

### Robustness to batch effects

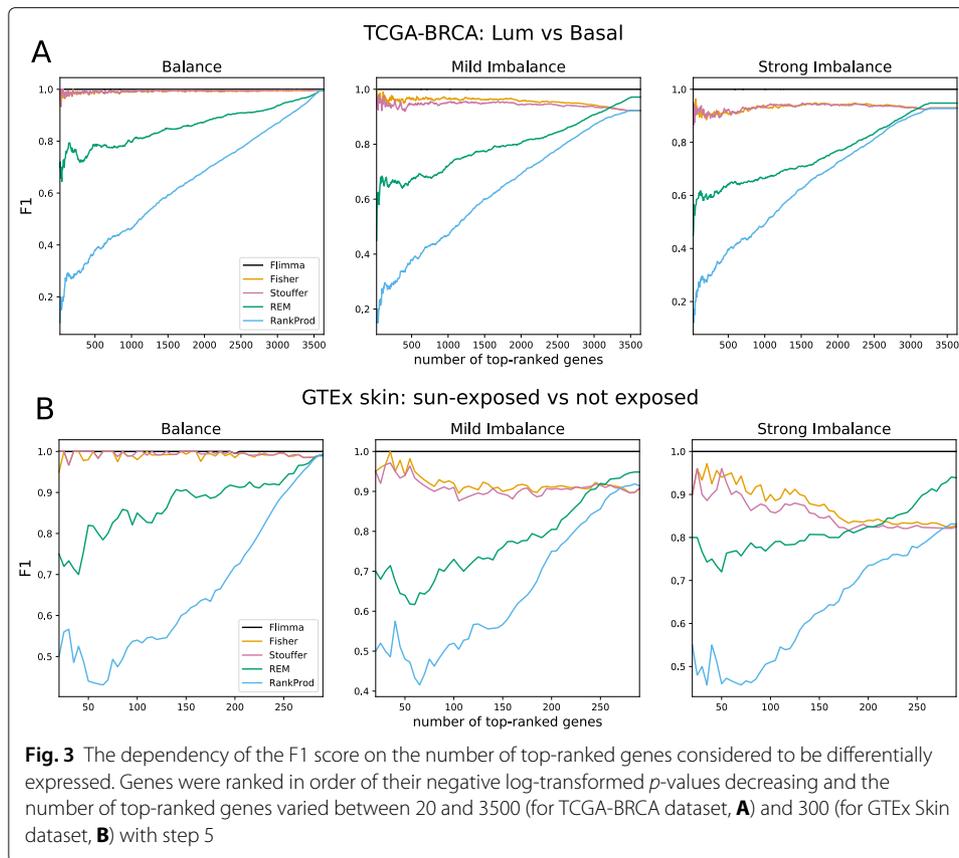
To demonstrate the robustness of *Flimma* towards experiential batch effects, we applied it on three additional publicly available breast cancer cohorts from GEO: GSE129508 [57], GSE149276 [58], and GSE58135 [59]. These datasets were independently collected and sequenced at three different laboratories and subjected to various experimental biases related to sample preparation, library construction, and sequencing platform (Additional file 7: Table S6). However, we assumed that collaborating partners can agree to use the

**Table 3** F1 score, the number of false positives (FP) and the number of false negatives (FN) obtained on TCGA-BRCA dataset in three scenarios. Values corresponding to the best performance over all methods are italicized. All calculated performance measures are reported in Additional file 2: Table S2

Scenario	F1			FP			FN		
	Balanced	Mildly imbalanced	Strongly imbalanced	Balanced	Mildly imbalanced	Strongly imbalanced	Balanced	Mildly imbalanced	Strongly imbalanced
Flimma	<i>1.00</i>	<i>1.00</i>	<i>1.00</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>
Fisher	<i>1.00</i>	0.92	0.93	14	248	192	8	290	265
Strouffer	<i>1.00</i>	0.92	0.93	14	245	189	9	290	265
REM	<i>1.00</i>	0.97	0.95	12	80	121	17	119	215
RankProd	<i>1.00</i>	0.92	0.93	14	243	193	12	295	274

**Table 4** F1 score, the number of false positives (FP), and the number of false negatives (FN) obtained on GTEx skin dataset in three scenarios. Values corresponding to the best performance over all methods are italicized. All calculated performance measures are reported in Additional file 3: Table S3

Scenario	F1			FP			FN		
	Balanced	Mildly imbalanced	Strongly imbalanced	Balanced	Mildly imbalanced	Strongly imbalanced	Balanced	Mildly imbalanced	Strongly imbalanced
Flimma	<i>1.00</i>	<i>1.00</i>	<i>1.00</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>	<i>0</i>
Fisher	0.99	0.91	0.83	4	32	67	0	18	33
Stouffer	0.99	0.91	0.83	4	32	67	0	18	33
REM	0.99	0.95	0.94	4	15	21	2	14	12
RankProd	0.99	0.91	0.83	4	32	67	0	18	33



same quantification pipeline and therefore obtained uniformly (in silico) preprocessed raw read counts from ARCHS<sup>4</sup> [60].

In contrast to TCGA-BRCA, cohort-specific batch effects in the GEO datasets were much more pronounced. Principal component analysis revealed that the differences between samples from different cohorts were much larger than the differences between subtypes within the same cohort (Fig. 4). In this case, effective adjustment for batch effect before testing for differential expression is crucial [61]. This can be done in two ways, either via subtracting the variation explained by batch from the data or via the inclusion of additional variables accounting for batch effects to the model. With *Flimma*, we implemented the second approach, as it is preferable for downstream statistical analysis [62]. Below, we will demonstrate that this approach effectively handles the batch effects in our breast cancer data sets and gives almost identical results. Several methods for batch effect correction exist, but not all of them are compatible with *limma voom* because the latter is computing count-based statistics. A recently published modification of the state-of-the-art batch-effect correction method *ComBat* [63], namely *ComBat-Seq* [64], is developed specifically to handle read count data. Hence, we utilized the results of *limma voom* obtained on the centralized GEO cohort after the removal of laboratory-specific effects by *ComBat-Seq* as a gold standard in the following experiments.

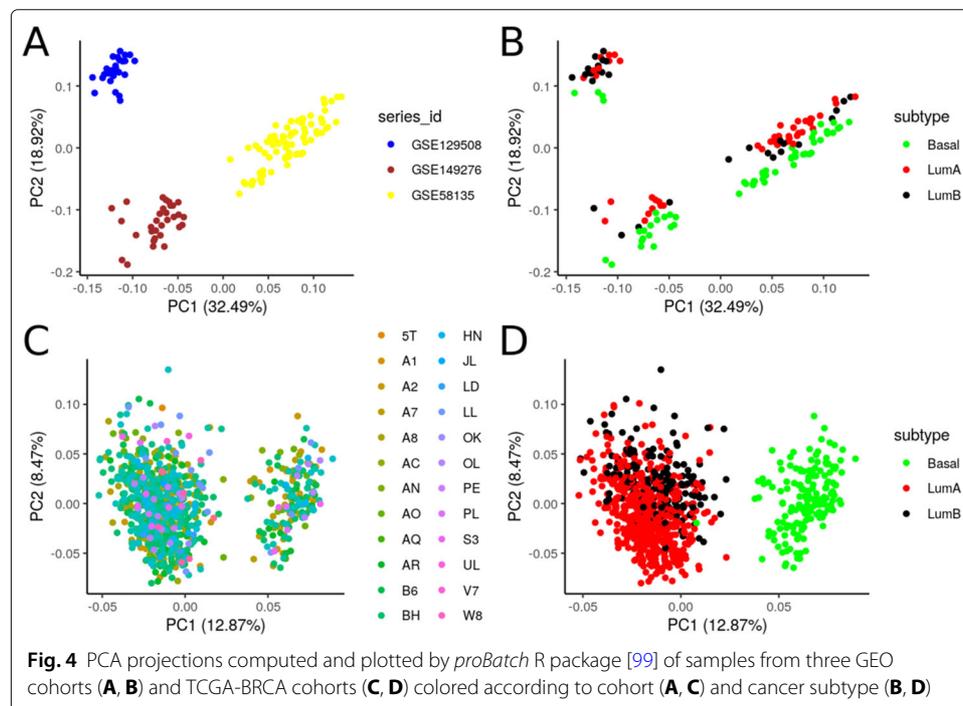
In *Flimma*, we model the batch effects of datasets by adding  $m - 1$  binary covariates to the linear model, where  $m$  is the number of datasets. Despite the strong batch effects in

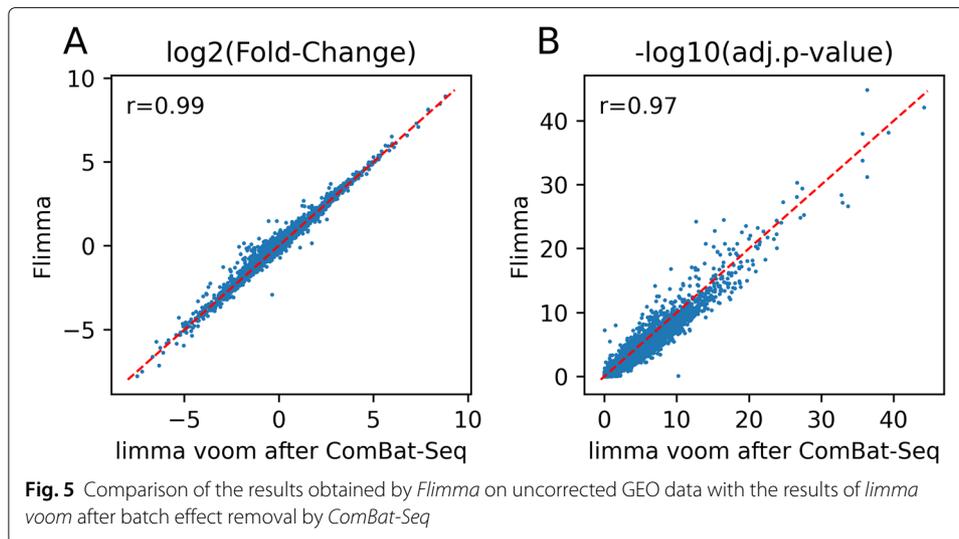
**Table 5** RMSE, precision, and recall obtained by *Flimma* and the meta-analysis tools on TCGA-BRCA datasets split by tissue source sites

The number of cohorts	3	5	7	10	14
RMSE					
Flimma	<i>0.0008</i>	<i>0.0007</i>	<i>0.0008</i>	<i>0.0017</i>	<i>0.0012</i>
Fisher	0.94	1.82	2.53	3.86	5.37
Stouffer	1.47	2.21	2.87	4.26	5.68
REM	2.73	3.68	4.75	7.21	8.50
RankProd	5.16	8.19	11.32	18.92	23.50
Precision					
Flimma	<i>1.00</i>	<i>1.00</i>	<i>1.00</i>	<i>1.00</i>	<i>1.00</i>
Fisher	0.85	0.88	0.90	0.93	0.95
Stouffer	0.85	0.88	0.91	0.93	0.95
REM	0.93	0.94	0.95	0.97	0.97
RankProd	0.92	0.87	0.90	0.93	0.95
Recall					
Flimma	<i>1.00</i>	<i>1.00</i>	<i>1.00</i>	<i>1.00</i>	<i>1.00</i>
Fisher	0.92	0.95	0.95	0.96	0.97
Stouffer	0.89	0.93	0.94	0.96	0.97
REM	0.93	0.96	0.97	0.98	0.98
RankProd	0.87	0.96	0.96	0.96	0.97

Values corresponding to the best performance over all methods are italicized

the GEO data, *Flimma* returned nearly the same fold-changes and BH-adjusted *p*-values as *limma voom* run on the same data after batch effect removal by *ComBat-Seq* (Fig. 5). Moreover, our results suggest that the approach used by *Flimma* gives better results than batch effect correction based on one or several first principal components (Additional file 4: Supplementary Text and Additional file 8: Table S7).





## Discussion

In this work, we presented *Flimma*, a privacy-aware tool for differential expression analysis. While *Flimma* results are mathematically equivalent to *limma voom*, *Flimma* can operate on distributed cohorts without the disclosure of sensitive data. To enhance data privacy, *Flimma* uses a hybrid federated approach, where the local parameters of the clients are hidden from the server and only global parameters resulting from the aggregation are disclosed. We employed *HyFed* to implement *Flimma* because unlike similar methods such as [65], it is an open source framework with a Python API (application programming interface) to develop hybrid federated tools. Moreover, it supports federated mode, in which different components can securely communicate over the Internet using the HTTPS protocol.

In this work, we have demonstrated that *Flimma* is superior to meta-analyses in imbalanced scenarios when the distributions of class labels or covariates are not identical between cohorts. We have also shown that *Flimma* is robust to technical batch effects.

One limitation of this work is the absence of a gold standard for the evaluation of differential expression analysis results. ABCD mixtures used in RNA-seq benchmark projects [1, 66] are not suitable for this study, since only five or less replicates of each mixture are sequenced by each participant. Although these projects are multi-center studies, such a small number of samples per participating center would not be realistic for mimicking modern biomedical studies involving human patients. Moreover, with these artificial mixtures, we could not model biological variability which is intrinsic of real-world patient-derived data. Therefore, we have only tested *Flimma* on patient-derived expression datasets, split them into parts modeling independent cohorts if necessary and considered the results of *limma voom* obtained on the combined datasets as ground truth.

## Remaining privacy risks

Although *Flimma* greatly enhances data privacy compared to centralized analysis, it does not provide a perfect privacy guarantee which quantifies the risk associated with the individual samples in the dataset. *Flimma* assumes non-colluding parties, e.g., the aggregator or compensator never exchanges the individual noisy parameters or noise values

from the clients with each other, and there are more than two clients participating in the study. Another assumption is honest-but-curious parties, which stick with the protocol and follow it but try to reconstruct the data from the model parameters.

One possible scenario of such a reconstruction attack is the recovery of the global  $X$  from  $X^T X$  by the aggregator, if the number of samples is close to the number of covariates. However, this is not realistic for differential expression analysis because the former should be much larger than the latter for a reliable analysis.

Another potential threat is the presence of a column with all 0 but one 1 in the global design matrix  $X$ . In this case,  $X^T Y$  reveals the expression profile of a sample with a non-zero value in that column. This is also an unlikely scenario because covariate columns that contain just a single non-zero element are not informative for differential expression analysis and should not be included in the model.

Since it is impossible to oversee all potentially possible scenarios where reconstruction might be feasible, the users should be aware that *Flimma* cannot fully exclude the risk of reconstruction attacks at intermediate results. Providing a privacy guarantee using DP to capture the privacy risks of patients in the dataset while preserving the accuracy of the results in a satisfactory level remains the direction for future research. Note that the risk of reconstruction attack is not excluded for meta-analysis methods. Although local  $p$ -values and effect sizes appear to be less prone to reconstruction attack than the aforementioned intermediate global parameters computed by *Flimma*, no formal proof of this intuition is provided. Despite that, meta-analyses remain popular approaches that are not in conflict with privacy legislation. In addition to a reasonable protection of the raw data, *Flimma* offers better accuracy than meta-analysis methods.

### Future directions

While *limma voom* is a state-of-the-art method for differential expression analysis that performs favorably in benchmarks [5], other methods for normalization (e.g., quantile normalization [67]) and differential expression analysis (such as *edgeR* [3], *DESeq2* [6], or *sleuth* [9]) exist and may yield different results depending on the dataset used. We thus consider extending *Flimma* with federated implementations of alternative methods in the future.

Another prospective direction for future work is the development of accessory tools for gene expression analysis. This includes for example, federated principle component Analysis (PCA), useful for quality control, or federated batch effect correction methods, such as *ComBat* or *RUVSeq* [68]. Although we have shown that the current version of *Flimma* effectively handles batch effects, other analyses of expression data such as clustering or classification might require transformed data.

Although *limma* has been initially developed for differential gene expression analysis, it is widely used for the analysis of various omics data types, e.g., proteomics [69, 70], metabolomics [71], and microbiomics [72]. Therefore, we plan the development of *Flimma* modifications suitable for the analysis of other omics data types in the future.

### Conclusions

*Flimma* is a privacy-aware tool for the federated identification of differentially expressed genes. It is user-friendly and publicly available at <https://exbio.wzw.tum.de/flimma/> including tutorials and a video documentation on its principle and application to real data.

While *Flimma* results are mathematically equivalent to *limma voom*, *Flimma* operates on distributed cohorts without the disclosure of sensitive data. To enhance data privacy, *Flimma* uses a hybrid federated approach, where the local parameters of the clients are hidden from the server and only global parameters resulting from the aggregation are disclosed. In contrast to meta-analysis approaches, *Flimma* is robust against heterogeneous distribution of data across the different sites and to technical batch effects. In summary, *Flimma* is a promising alternative to meta-analysis methods for multi-center gene expression projects, as it enhances patient privacy while providing the same results as a centralized analysis.

## Methods

### The *limma voom* workflow

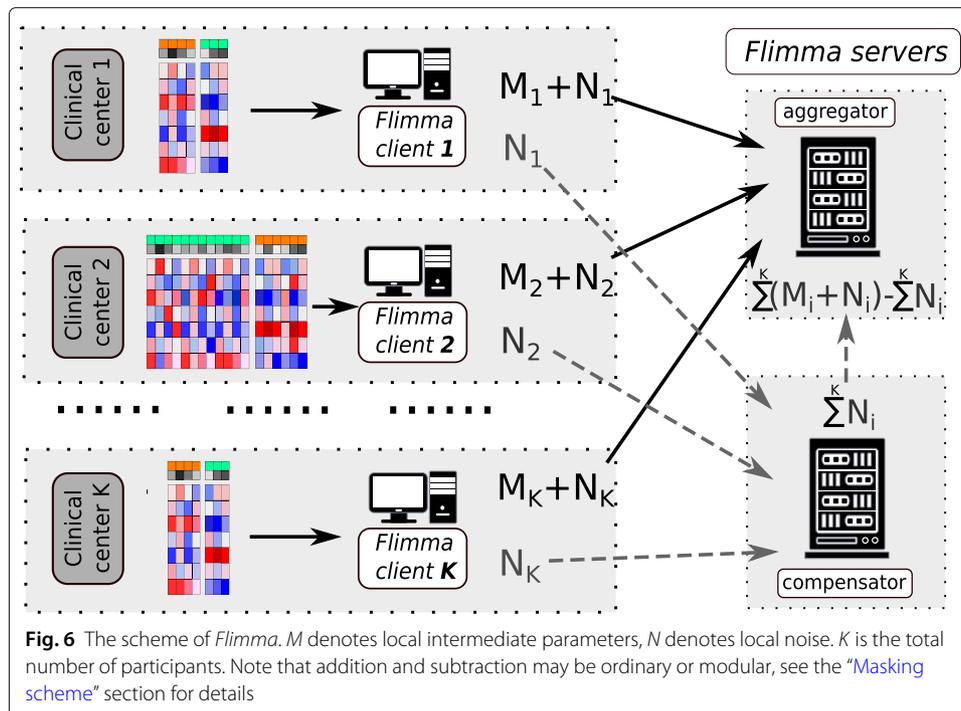
*limma voom* is the state-of-the-art method for differential expression analysis. Initially designed for microarrays [53], it was extended by the *voom* function, which removes the mean-variance trend from RNA-seq data and makes it suitable for analysis by *limma* [5]. Recently, the authors of *limma* published an updated guideline on the recommended *limma voom* workflow [52]. Data preprocessing steps of this workflow include removal of weakly expressed genes using the *filterByExpr* function from the *edgeR* package, conversion of raw read counts to log<sub>2</sub>-transformed counts per million (log-CPM), and normalization of gene expression distributions. We only differ from this workflow by using the upper-quartile (UQ) normalization [35] instead of the trimmed mean of M-values (TMM) normalization [73], since the latter would require disclosing one of the sample profiles to all participants. Although UQ is not the only normalization method that could be implemented in a federated fashion, we have chosen it because it is one of the most widely used in the field [68, 74]. Since no normalization method outperforms others in all cases [75, 76], we are going to implement more federated normalization methods in the future. Furthermore, given the matrix of normalized log-CPM values and the design matrix, *voom* computes precision weights, which compensate for the mean-variance bias that is typical for RNA-seq data and thus makes them suitable for use in *limma*.

### *Flimma*

#### Implementation

*Flimma* is based on *HyFed* (<https://github.com/tum-aimed/hyfed>) [77], a hybrid federated framework implementing an SMPC-like approach to hiding the original values of the local parameters from the server (Fig. 6) [54]. *HyFed* comprises four software components: an aggregator server, a compensator server, a client app, and a web interface.

To start the project, the coordinating user signs into the web interface, creates the project, sets its parameters (e.g., confounding factors, etc.), and invites the participants. Each participant receives a token and a project ID from the coordinator and locally runs the client app to join the study and to select the local dataset. The computations are orchestrated by the aggregator server, which coordinates the clients, aggregates their local model parameters to global parameters, and returns global parameters to clients. Unlike in FL, with *HyFed*, clients mask their local parameters with noise before sending them to the aggregator to enhance the data privacy. The noise matrix has the same shape as the parameter matrix and contains random numbers. The approach to random number



generation depends on the data type of the masked matrix and is described in detail in the next section. The noise matrix is sent to the compensator server, which aggregates the noise received from all clients and shares the global noise matrix with the aggregator. The aggregator calculates noisy global parameters and denoises them, by subtracting the global noise matrix provided by the compensator from the noisy global parameters. The proposed hybrid approach provides improved privacy, because a reconstruction attack would require compromising two servers in this case. The aggregator and compensator server components should run in separate machines at distant physical locations. Ideally, to minimize the risk of reconstruction attacks, they should be controlled by third-party organizations not connected to any of the study participants. Currently, the publicly available *Flimma* web tool is using the aggregator running at the Chair of Experimental Bioinformatics, Technical University of Munich (Germany), while the compensator is hosted at the Department of Mathematics and Computer Science at the University of Southern Denmark (Denmark).

As the original *limma voom*, each *Flimma* client accepts a matrix of read counts and a design matrix, specifying class labels and covariates for each sample. *Flimma* outputs a table with  $p$ -values, fold-changes, and moderated  $t$  statistics for each gene.

*Flimma* is publicly available at <https://exbio.wzw.tum.de/flimma/>. The “HowTo” page provides a quick-start guide for *Flimma* along with test data and describes input file formats.

### Masking scheme

*Flimma* employs the local parameter masking approach of *HyFed*, which treats non-negative integer-valued parameters and real-valued parameters differently. For masking non-negative integers, it applies the standard additive secret sharing scheme based on modular arithmetic over the finite field  $\mathbb{Z}_p = \{0, 1, p - 1\}$ , where  $p$  is a prime number

[39]. The elements of noise matrix  $N_i$  are drawn from  $\mathbb{Z}_p$  and added to parameters matrix  $M_i$  using modular addition over  $\mathbb{Z}_p$ , i.e.,  $M'_i = (M_i + N_i) \bmod p$ . The compensator and aggregator also use modular addition to compute global noise  $N = \left( \sum_{i=1}^K N_i \right) \bmod p$  and global noisy parameters  $M' = \left( \sum_{i=1}^K M'_i \right) \bmod p$ . Finally, the aggregator removes the global noise from global noisy parameters  $M = (M' - N) \bmod p$ .

Real-valued parameters are protected by the secret sharing approach based on Gaussian distribution [78, 79]. Noise values are drawn from  $\mathcal{N}(0, \sigma^2)$  added to local parameters:  $M'_i = M_i + N_i$ . Noise aggregation and compensation is performed using ordinary addition and subtraction operations, respectively:  $N = \sum_{i=1}^K N_i$ ,  $M = \sum_{i=1}^K M'_i - N$ .

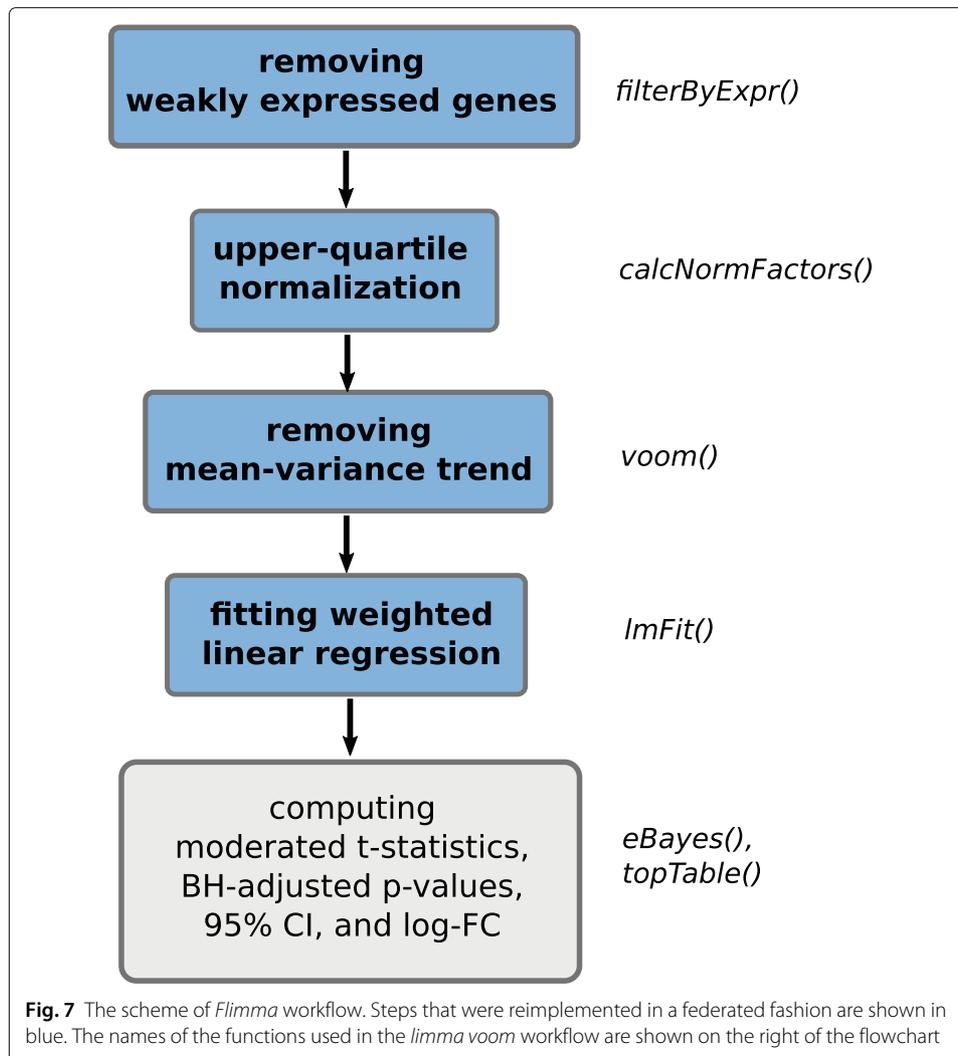
The theoretical analysis of information leakage for additive secret sharing based on modular arithmetics [39] and the real value secret sharing based on Gaussian distribution [79] using the mutual information criterion [80] are provided in the literature [54]. The mutual information measures the reduction in uncertainty about one random variable (e.g., the original values of local parameters  $M_i$ ) given the knowledge of another random variable (e.g., noisy local parameters  $M'_i$ ). Regarding the original and noisy local parameters with non-negative integer values, it has been shown that the mutual information between them is zero, and thus, the noisy local parameters leak no information about the original local parameters [39]. For real-valued local parameters, however, the upper-bound on mutual information between  $M_i$  and  $M'_i$  is:  $\frac{1}{2} \log_2 \left( 1 + \frac{\sigma_{M_i}^2}{\sigma^2} \right)$ , where  $\sigma_{M_i}^2$  and  $\sigma^2$ , indicate the variance of the original values of the local parameters and the variance of the Gaussian noise, respectively. That is, the maximum amount of information about  $M_i$  disclosed by  $M'_i$  depends on  $\frac{\sigma_{M_i}^2}{\sigma^2}$ .

In practice, *Flimma* sets  $p$  equal to  $2^{54} - 33$ , the largest prime number that can fit in a 54-bit integer, and  $\sigma^2 = 10^{12}$ , which is large enough for typical gene expression from the privacy perspective. The mean of the Gaussian noise generator has no significant impact on privacy [79], and therefore, *Flimma* sets it to zero. To ensure the correctness of the results for non-negative local parameters, overflow must not occur during the computation of the aggregated noise, aggregated noisy local parameters, and  $\sum_{i=1}^K M_i < p$ . The value of  $p$  can be set to larger values to support larger integers but at the cost of supporting a fewer number of clients [54]. Likewise, too large values of  $\sigma^2$  might impact the precision of the results. However, we confirmed that with default values of  $p$  and  $\sigma^2$ , the differences between  $p$ -values and  $t$  statistics computed by *Flimma* with and without masking the local parameters never exceeded the  $10^{-8}$ .

### Workflow

*Flimma* implements a federated version of the *limma voom* workflow, allowing privacy-aware detection of differentially expressed genes. The scheme of the *Flimma* workflow is presented in Fig. 7.

First, genes that do not have sufficient counts for further statistical analysis are removed. For this, we implemented a federated version of the *filterByExprs* function [81] from the *edgeR* package, which employs two filters: *min\_total\_count* filter and CPM cutoff. The first filter removes genes whose sum of counts over all samples does not exceed



*min\_total\_count* threshold. The second filter excludes genes expressed in insufficient number of samples. It keeps only genes where at least *min\_n\_samples* samples pass the CPM cutoff. This cutoff is calculated as a ratio of *min\_count* over the median library size multiplied by  $10^6$ , where *min\_n\_samples* is defined by the smallest group size in the design matrix. The function parameters *min\_count* and *min\_total\_count* are set to 10 and 15 by default and can be adjusted by the user.

UQ normalization performed in the second step of the pipeline requires the exchange of scaled normalization factors which cannot be used to reveal any private data. The third and the fourth steps of the workflow resemble the *voom* and *lmFit* functions from the *limma* package, which are fitting linear regression models. For training the linear regression model in the federated fashion, *Flimma* utilizes the same approach described by [82]. For each gene, each of  $n$  clients compute local noisy results  $(X^i)^T X^i + N_{X^T X}^i$  and  $(X^i)^T Y^i + N_{X^T Y}^i$ , where  $X^i$  is a real-valued design matrix,  $Y^i$  is the vector of normalized log2-CPM values for the gene,  $N_{X^T X}^i$  and  $N_{X^T Y}^i$  are the noise matrices, and  $i$  is the index of a client, and sends them to the server. The compensator summarizes noise from clients

to global noise

$$N_{X^T X} = \sum_{i=1}^K N_{X^T X}^i, N_{X^T Y} = \sum_{i=1}^K N_{X^T Y}^i,$$

and shares it with the aggregator. The aggregator computes global noisy results  $X^T X$  and  $X^T Y$  and denoises them:

$$X^T X = \sum_{i=1}^K \left( (X^i)^T X^i + N_{X^T X}^i \right) - N_{X^T X},$$

$$X^T Y = \sum_{i=1}^K \left( (X^i)^T Y^i + N_{X^T Y}^i \right) - N_{X^T Y}.$$

The denoised  $X^T X$  and  $X^T Y$  are used to compute  $\beta$ , and unscaled standard errors of the coefficients:

$$\beta = \left( X^T X \right)^{-1} X^T Y,$$

$$uSE_{\beta} = \text{diag} \left( X^T X \right).$$

Global coefficients  $\beta$  are sent back to the clients, which locally compute fitted log-CPM

$$\hat{Y}^i = X^i \beta,$$

and the noisy sums of squared errors

$$SSE^i = \sum_{s=1}^{m^i} \left( y_s^i - \hat{y}_s^i \right)^2 + N_{SSE}^i,$$

where  $s$  is sample index and  $m^i$  is the total number of samples in the  $i$ th client.

The aggregator collects noisy  $SSE^i$  from clients, receives global noise

$$N_{SSE} = \sum_i N_{SSE}^i$$

from the compensator, and computes estimated residual standard deviations for each gene:

$$\sigma = \sqrt{\frac{\sum_i SSE^i - N_{SSE}}{\left( \sum_i m^i \right) - k}}$$

The fifth step involves only  $\beta$ ,  $\sigma^2$ , and unscaled standard errors, and therefore does not require to be federated. All subsequent computations are performed on the side of the aggregator in the same way as done by the original *limma voom*.

### Meta-analysis approaches

Three classes of meta-analysis approaches can be distinguished: effect size combination methods,  $p$ -value combination methods, and non-parametric methods [33]. Effect size combination methods estimate variances of effect sizes for every gene and compute global effect sizes as a weighted sum of local effect sizes divided by the sum of all weights. This class includes the fixed effects model (FEM) and the random effects model (REM), which differ in the way they compute weights [30]. FEM calculates the weights as the inverse of the within-study variance. REM assumes that total variance includes within-study and

between-study variance components and calculates the inverse of their sum. Both methods calculate  $p$ -values given global effect sizes and assuming their normal distribution. We chose REM since it is more robust to data heterogeneity than FEM and more widely used [83].

$P$ -value combination methods are based on the assumption that the sum, minimum or maximum of log-transformed  $p$ -values obtained in independent studies follow a certain distribution [33]. These methods are thought to be more suitable for imbalanced scenarios than effect size combination methods [84]. From this class of methods, we chose Fisher's method [27] because it is most sensitive to small  $p$ -values [85] and Stouffer's method (also known as  $z$ -method) [28] since it was shown to be superior to Fisher's method in some cases [86].

Non-parametric rank-based methods estimate global permutation-based  $p$ -value, by comparing the sum or the product of ranks obtained for the observed matrix of ranks with the same summary statistics calculated on shuffled rank matrices. Although the Rank Product method [29] is much more computationally expensive than the Rank Sum, the first gives more robust results [87].

In this work, we used the REM and Fisher's method from *metaVolcanoR* package [88], the implementation of Stouffer's method from *MetaDE* package [89] and *RankProd* package [90] for Rank Product method. For all selected meta-analysis methods except REM, global fold change was calculated as a mean of local fold changes.

## Evaluation

The main result of differential expression analysis is a list of genes with  $p$ -values and log-fold changes, reflecting the significance and the strength of differential expression, respectively. To validate the results of *Flimma* and demonstrate its advantage over meta-analysis approaches, we compared the *Flimma* and meta-analysis results obtained on artificial dataset splits to the results of *limma voom* applied on the aggregated datasets.

We chose two large datasets comprising RNA-seq gene expression profiles of human-derived samples. The first dataset included 850 expression profiles of human breast tumors from TCGA-BRCA cohort [55], classified as luminal or basal subtypes and annotated with patient age and tumor stage. We searched for genes differentially expressed between luminal and basal subtypes and included the age of diagnosis and tumor stage as covariates. The second dataset comprised 1277 skin expression profiles from GTEX [56] with sun exposure as target class label and patient age and sex as covariates. Each dataset has been divided into cohorts to model the multi-party setting under various scenarios (see the "Datasets" section for details).

In all tests, we applied *limma voom* on the complete dataset and on each of its partitions independently. The  $p$ -values and effect sizes computed by *limma voom* on the aggregated datasets were treated as ground truth, and those obtained on cohorts were used as input for the meta-analysis methods, which aggregated them to the global  $p$ -values.

To avoid manual execution of *Flimma* GUI for every test, we used a script performing exactly the same computations as the web version of *Flimma*. The code for running *Flimma* and its baselines, and the instructions for data download and preprocessing are available at GitHub (<https://github.com/ozolotareva/flimma>) [91] and at Zenodo (doi:<https://doi.org/10.5281/zenodo.5711972>) [92] under the terms of the Apache 2.0

license. *Flimma* and the methodology of its evaluation are described in AIME registry [93] at <https://aime-registry.org/report/v6v9dj>.

For each method, we considered a gene determined as differentially expressed, if it has  $|\log(FC)| > 1$ , and BH-adjusted  $p$ -value  $< 0.05$ . For the results produced by each method, we computed the RMSE, the precision, the recall, the F1 score, the Pearson, and the Spearman correlation. Since only a small number of the most significantly differentially expressed genes is of interest for some research tasks, we have also investigated how the performance of the methods varies with the numbers of top-ranked genes selected.

## Datasets

### *TCGA breast cancer data*

Unprocessed read counts summarized to gene-level and clinical annotations of samples were downloaded from [https://gdac.broadinstitute.org/runs/stddata\\_\\_2016\\_01\\_28/data/BRCA/20160128](https://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/BRCA/20160128). 850 expression profiles classified as luminal, or basal-like subtypes and annotated with the age of diagnosis and tumor stage were kept. Although breast cancer samples are classified into 4–6 subtypes [94–96], we focused on the most frequent subtypes for evaluation purposes. Luminal and basal subtypes are well distinguishable at the level of gene expression [55, 94] (Additional file 4: Figure S3A). We searched for genes differentially expressed between these subtypes and included the age of diagnosis and tumor stage as covariates. The luminal subtype is subdivided into luminal A (LumA) and luminal B (LumB) subtypes [95]. However, the LumA subtype was not included in the model as a covariate and we modeled the presence of an unknown disease subtype in our experiments.

### *GTEX skin data*

Raw read counts per gene were obtained from the GTEx v8 portal website (<https://www.gtexportal.org/home/datasets>). Expression profiles of sun-exposed and non-sun-exposed skin samples annotated with mean ischemic time and sex were kept. The resulting dataset comprises 1277 expression profiles of 677 sun-exposed and 600 non-sun-exposed skin samples, also annotated with sex and ischemic time. In contrast to the TCGA-BRCA dataset, a smaller fraction of genes was differentially expressed between sun-exposed and non-exposed skin samples (Additional file 4: Figure S3B). Besides patient age and sex, samples were annotated with ischemic time, i.e. the time between patient death or sample withdrawal and sample fixation, or freezing. Ischemic time was not included in linear models but varied between cohorts in imbalanced scenarios, thus serving as an unknown confounder related to differences in sample preprocessing.

### *Generation of artificially distributed and heterogeneous datasets*

To demonstrate the robustness of *Flimma*, we split both datasets differently in a balanced, a mildly imbalanced, and a strongly imbalanced scenario. In the balanced scenario, each sample was randomly assigned to one of three equal-sized cohorts with a similar distribution of covariates. In the imbalanced scenarios, the fractions of target classes and the distributions of some covariates differed among cohorts. Cohort sizes were unequal and related as 1:2:4 and 1:3:9 for the mildly and the strongly imbalanced scenarios, respectively. In the TCGA-BRCA dataset, we introduced an imbalance of luminal and basal subtype frequencies and, in addition, changed the frequency of the LumA subtype (Table 1). In the GTEx skin dataset, the fraction of sun-exposed skin samples and

the median of mean ischemic times were made unequal between cohorts in imbalanced scenarios (Table 2).

### **GEO datasets**

Raw read counts for three breast cancer cohorts from GSE129508 [57], GSE149276 [58], and GSE58135 [59] were obtained from ARCHS<sup>4</sup> [60] (<https://maayanlab.cloud/archs4/>). ARCHS<sup>4</sup> collected raw reads from publicly available human and mouse GEO datasets and uniformly preprocessed them. Raw reads from each human-derived sample were pseudo-aligned against the GRCh38 human reference genome and quantified by *kallisto* [97]. Since in our experiment we searched for genes differentially expressed between human breast cancer subtypes, we have chosen datasets comprising patient-derived breast tumor samples and excluded xenografts and cell lines. We also excluded samples annotated as cell lines from GSE58135 and post-intervention samples from GSE129508. Intrinsic breast cancer subtypes were predicted using the *genefu* R package [98]. Same as before, we searched for genes, differentially expressed between the luminal and basal subtypes. Luminal A subtype and the sequencing center were added to the model as covariates.

## **Supplementary Information**

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-021-02553-2>.

**Additional file 1:** Table S1.  
**Additional file 2:** Table S2  
**Additional file 3:** Table S3  
**Additional file 4:** Supplementary Text and Figures S1-S3  
**Additional file 5:** Table S4  
**Additional file 6:** Table S5  
**Additional file 7:** Table S6  
**Additional file 8:** Table S7  
**Additional file 9:** Review history.

### **Review history**

The review history is available as Additional file 9.

### **Peer review information**

Barbara Cheifet was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

### **Authors' contributions**

O.Z., M.L., and J.B. conceived and designed the study. O.Z. preprocessed the data, performed the experiments, and analyzed the results. O.Z. and R.N. developed the federated algorithms. R.N. and M.B. implemented the client and server components. J.M. implemented the web interface. G.K. and D.R. contributed to the design and implementation of the tool from the privacy aspect. All authors provided critical feedback and helped in the interpretation of data, manuscript writing, and approved the final version.

### **Authors' information**

Twitter handles: @julian\_spaeth (Julian Späth); @dbblumenthal, @bionetslab (David B. Blumenthal); @GKaissis (Georgios Kaissis); @danielrueckert (Daniel Rückert); @itivalist (Markus List); @janbaumbach (Jan Baumbach).

### **Funding**

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of "CLINSPECT-M" (grant FKZ161L0214A) and within the framework of the \*e:Med \*research and funding concept "Sys\_CARE" (\*grant 01ZX1908A\*). This project has received funding from the European Union's Horizon2020 research and innovation programme under grant agreement No 826078 (FeatureCloud) and No 777111 (REPO-TRIAL). This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. JB was partially funded by his VILLUM Young Investigator Grant No 13154. Open Access funding enabled and organized by Projekt DEAL.

**Availability of data and materials**

All data analyzed during this study were obtained from publicly available resources. Gene expressions of breast tumors from TCGA-BRCA dataset [55] were obtained from Broad GDAC Firehose ([https://gdac.broadinstitute.org/runs/stddata\\_2016\\_01\\_28/data/BRCA/20160128/](https://gdac.broadinstitute.org/runs/stddata_2016_01_28/data/BRCA/20160128/)). Uniformly preprocessed gene-level read counts for three additional independently generated breast cancer datasets GSE129508 [57], GSE149276 [58], and GSE58135 [59] were downloaded from ARCHS4 [60] (<https://maayanlab.cloud/archs4/>). Gene expression of sun-exposed and not sun-exposed skin were downloaded from GTEx [56] website (<https://gtexportal.org/home/datasets>). The code for running *Flimma* and its baselines, and the instructions for data download and preprocessing are available at GitHub (<https://github.com/ozolotareva/flimma>) [91] and Zenodo (<https://zenodo.org/record/5711972#.YZrXm5HMIrM>) [92].

**Declarations****Ethics approval and consent to participate**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.

**Author details**

<sup>1</sup>Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Freising, Germany. <sup>2</sup>Institute for Computational Systems Biology, University of Hamburg, Hamburg, Germany. <sup>3</sup>AI in Medicine and Healthcare, Technical University of Munich, Munich, Germany. <sup>4</sup>Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark. <sup>5</sup>Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany. <sup>6</sup>CNR National Research Council, IAC Institute for Applied Computing, Rome, Italy. <sup>7</sup>Sapienza University of Rome, Rome, Italy. <sup>8</sup>Klinikum rechts der Isar, Technical University of Munich, Munich, Germany. <sup>9</sup>Biomedical Image Analysis Group, Imperial College London, London, UK. <sup>10</sup>OpenMined, Oxford, UK.

Received: 23 November 2020 Accepted: 22 November 2021

Published online: 14 December 2021

**References**

1. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat Biotechnol.* 2014;32(9):903–14.
2. Oshlack A, Wakefield M. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct.* 2009;4(1):14.
3. Robinson M, McCarthy D, Smyth G. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010;26(1):139–40.
4. Hardcastle T, Kelly K. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinforma.* 2010;11(1):422.
5. Law C, Chen Y, Shi W, Smyth G. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 2014;15(2):29.
6. Love M, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15(12):.
7. Ritchie M, Phipson B, Wu D, Hu Y, Law C, Shi W, Smyth G. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43(7):47.
8. Tarazona S, Furió-Tarí P, Turrà D, Pietro A, Nueda M, Ferrer A, Conesa A. Data quality aware analysis of differential expression in RNA-seq with NOISeq r/bioc package. *Nucleic Acids Res.* 2015;711:e140.
9. Pimentel H, Bray N, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat Methods.* 2017;14(7):687–90.
10. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason C, Socci N, Betel D. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* 2013;14(9):95.
11. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinforma.* 2013;14:91.
12. Costa-Silva J, Domingues D, Lopes F. RNA-Seq differential expression analysis: an extended review and a software tool. *PLoS ONE.* 2017;12(12):0190152.
13. Zhang M, Yao C, Guo Z, Zou J, Zhang L, Xiao H, Wang D, Yang D, Gong X, Zhu J, Li Y, Li X. Apparently low reproducibility of true differential expression discoveries in microarray studies. *Bioinformatics.* 2008;24(18):2057–63.
14. McIntyre L, Lopiano K, Morse A, Amin V, Oberg A, Young L, Nuzhdin S. RNA-seq: technical variability and sampling. *BMC Genomics.* 2011;12:293.
15. Shi L, Jones W, Jensen R, Harris S, Perkins R, Goodsaid F, Guo L, Croner L, Boysen C, Fang H, Qian F, Amur S, Bao W, Barbacioru C, Bertholet V, Cao X, Chu T-M, Collins P, Fan X-H, Frueh F, Fuscoe J, Guo X, Han J, Herman D, Hong H, Kawasaki E, Li Q-Z, Luo Y, Ma Y, Mei N, Peterson R, Puri R, Shippy R, Su Z, Sun Y, Sun H, Thorn B, Turpaz Y, Wang C, Wang S, Warrington J, Willey J, Wu J, Xie Q, Zhang L, Zhang L, Zhong S, Wolfinger R, Tong W. The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. *BMC Bioinforma.* 2008;9(Suppl 9):10.
16. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci.* 2006;103(15):5923–5928. <https://doi.org/10.1073/pnas.0601231103>.
17. Łabaj P, Kreil D. Sensitivity, specificity, and reproducibility of RNA-Seq differential expression calls. *Biol Direct.* 2016;11(1):66.
18. Papin J, Mac Gabhann F, Sauro H, Nickerson D, Rampadarath A. Improving reproducibility in computational biology research. *PLoS Comput Biol.* 2020;16(5):1007881.

19. Gymrek M, McGuire A, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. *Science*. 2013;339(6117):321–4.
20. Sweeney L, Abu A, Winn J. Identifying Participants in the Personal Genome Project by Name (A Re-identification Experiment). 2013. <https://arxiv.org/abs/1304.7605>.
21. Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. *Nat Genet*. 2020;52(7):646–54.
22. Schadt E, Woo S, Hao K. Bayesian method to predict individual SNP genotypes from gene expression data. *Nat Genet*. 2012;44(5):603–8.
23. Harmanci A, Gerstein M. Quantification of private information leakage from phenotype-genotype data: linking attacks. *Nat Methods*. 2016;13(3):251–6.
24. Shi X, Wu X. An overview of human genetic privacy. *Ann NY Acad Sci*. 2017;1387(1):61–72.
25. Tryka K, Hao L, Sturcke A, Jin Y, Wang Z, Ziyabari L, Lee M, Popova N, Sharopova N, Kimura M, Feolo M. NCBI's database of genotypes and phenotypes: dbGaP. *Nucleic Acids Res*. 2014;42(Database issue):975–9.
26. Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding J, Ur-Rehman S, Saunders G, Kandasamy J, Caccamo M, Leinonen R, Vaughan B, Laurent J, Rowland F, Marin-Garcia P, Barker J, Jokinen P, Torres A, de Argila J, Llobet O, Medina I, Puy M, Alberich M, de la Torre S, Navarro A, Paschall J, Flicek P. The European genome-phenome archive of human data consented for biomedical research. *Nat Genet*. 2015;47(7):692–5.
27. Fisher RA. Statistical methods for research workers. In: *Breakthroughs in statistics*. Springer; 1992. p. 66–70.
28. Stouffer S, Suchman E, Devinney L, Star S, Williams RMbsuffixJ. The American soldier: adjustment during army life. (*studies in social psychology in World War II*) vol. 1. 1949;1:599.
29. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett*. 2004;573(1-3):83–92.
30. Choi J, Yu U, Kim S, Yoo O. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*. 2003;19(Suppl 1):84–90.
31. Zeggini E, Ioannidis J. Meta-analysis in genome-wide association studies. *Pharmacogenomics*. 2009;10(2):191–201.
32. Hong F, Breitling R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*. 2008;24(3):374–82.
33. Toro-Domínguez D, Villatoro-García J, Martorell-Marugán J, Román-Montoya Y, Alarcón-Riquelme M, Carmona-Sáez P. A survey of gene expression meta-analysis: methods and applications. *Brief Bioinform*. 2020;22(2): 1694–1705.
34. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Stat Med*. 2002;21(11):1539–1558.
35. Bullard J, Purdom E, Hansen K, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinforma*. 2010;11:94.
36. McMahan B, Moore E, Ramage D, Hampson S, y Arcas B. Communication-efficient learning of deep networks from decentralized data. Fort Lauderdale, FL, USA: Proc Mach Learn Res; 2017. p. 1273–82.
37. Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Halevi S, Rabin T, editors. *Theory of cryptography*. Berlin, Heidelberg: Springer; 2006. p. 265–84. [https://doi.org/10.1007/11681878\\_14](https://doi.org/10.1007/11681878_14).
38. Gentry C. Fully homomorphic encryption using ideal lattices. In: *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing, STOC '09*. New York, NY, USA: Association for Computing Machinery; 2009. p. 169–78. <https://doi.org/10.1145/1536414.1536440>.
39. Cramer R, Damgård I, Nielsen J. *Secure multiparty computation and secret sharing*. Cambridge: Cambridge University Press; 2015.
40. Torkzadehmahani R, Nasirigerdeh R, Blumenthal DB, Kacprowski T, List M, Matschinske J, Späth J, Wenke NK, Bihari B, Frisch T, et al. Privacy-preserving Artificial Intelligence Techniques in Biomedicine. *arXiv preprint arXiv:2007.11621*. 2020. <https://arxiv.org/abs/2007.11621>.
41. Lyu L, Yu H, Yang Q. Threats to federated learning: A survey. *arXiv preprint arXiv:2003.02133*. 2020. <http://arxiv.org/abs/2003.02133>.
42. Nasirigerdeh R, Torkzadehmahani R, Matschinske J, Frisch T, List M, Späth J, Weiß S, Völker U, Heider D, Wenke NK, et al. sPLINK: a federated, privacy-preserving tool as a robust alternative to meta-analysis in genome-wide association studies. *BioRxiv*. 2020.
43. Wu X, Zheng H, Dou Z, Chen F, Deng J, Chen X, Xu S, Gao G, Li M, Wang Z, Xiao Y, Xie K, Wang S, Xu H. A novel privacy-preserving federated genome-wide association study framework and its application in identifying potential risk variants in ankylosing spondylitis. *Brief Bioinform*. 2020;22(3):.
44. Andreux M, Manoel A, Menuet R, Saillard C, Simpson C. Federated Survival Analysis with Discrete-Time Cox Models. *arXiv preprint arXiv:2006.08997*. 2020. <http://arxiv.org/abs/2006.08997>.
45. Rieke N, Hancox J, Li W, Milletari F, Roth H, Albarqouni S, Bakas S, Galtier M, Landman B, Maier-Hein K, Ourselin S, Sheller M, Summers R, Trask A, Xu D, Baust M, Cardoso M. The future of digital health with federated learning. *NPJ Digit Med*. 2020;3:119.
46. Chialva D, Dooms A. Conditionals in homomorphic encryption and machine learning applications. *arXiv preprint arXiv:1810.12380*. 2018. <https://arxiv.org/abs/1810.12380>.
47. Blatt M, Gusev A, Polyakov Y, Goldwasser S. Secure large-scale genome-wide association studies using homomorphic encryption. *Proc Nat Acad Sci*. 2020;117(21):11608–13. <https://doi.org/10.1073/pnas.1918257117>.
48. Cho H, Wu D, Berger B. Secure genome-wide association analysis using multiparty computation. *Nat Biotechnol*. 2018;36(6):547–51. <https://doi.org/10.1038/nbt.4108>.
49. Nasirigerdeh R, Torkzadehmahani R, Baumbach J, Blumenthal D. On the privacy of federated pipelines. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*. New York: ACM; 2021. <https://doi.org/10.1145/3404835.3462996>.
50. Melis L, Song C, De Cristofaro E, Shmatikov V. Exploiting unintended feature leakage in collaborative learning. In: *2019 IEEE Symposium on Security and Privacy (SP)*. New York: IEEE; 2019. p. 691–706.
51. Zhu L, Han S. *Deep leakage from gradients*. Cham: Springer; 2020, pp. 17–31.

52. Law C, Alhamdoosh M, Su S, Dong X, Tian L, Smyth G, Ritchie M. RNA-seq analysis is easy as 1-2-3 with limma, glimma and edgeR. *F1000Res*. 2016;5: <https://pubmed.ncbi.nlm.nih.gov/27441086/>.
53. Smyth G. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3:3.
54. Nasirigerdeh R, Torkzadehmahani R, Matschinske J, Baumbach J, Rueckert D, Kaissis G. HyFed: A Hybrid Federated Framework for Privacy-preserving Machine Learning. arXiv preprint arXiv:2105.10545. 2021. <http://arxiv.org/abs/2105.10545>.
55. Liu J, Lichtenberg T, Hoadley K, Poisson L, Lazar A, Cherniack A, Kovatich A, Benz C, Levine D, Lee A, Omberg L, Wolf D, Shriver C, Thorsson V, Cancer Genome Atlas Research Network, Hu H. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*. 2018;173(2):400–41611.
56. GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020;369(6509):1318–30.
57. Ligibel J, Dillon D, Giobbie-Hurder A, McTiernan A, Frank E, Cornwell M, Pun M, Campbell N, Dowling R, Chang M, Toloney S, Chagpar A, Yung R, Freedman R, Dominici L, Golshan M, Rhei E, Taneja K, Huang Y, Brown M, Winer E, Jeselsohn R, Irwin M. Impact of a pre-operative exercise intervention on breast cancer proliferation and gene expression: results from the pre-operative health and body (PreHAB) study. *Clin Cancer Res*. 2019;25(17):5398–406. <https://doi.org/10.1158/1078-0432.ccr-18-3143>.
58. Park S, Lee E, Park S, Lee S, Nam S, Kim S, Lee J, Yu J-H, Kim J-Y, Ahn J, Im Y-H, Park W-Y, Park K, Park Y. Clinical characteristics and exploratory genomic analyses of germline BRCA1 or BRCA2 mutations in breast cancer. *Mol Cancer Res*. 2020;18(9):1315–25. <https://doi.org/10.1158/1541-7786.mcr-19-1108>.
59. Varley K, Gertz J, Roberts B, Davis N, Bowling K, Kirby M, Nesmith A, Oliver P, Grizzle W, Forero A, Buchsbaum D, LoBuglio A, Myers R. Recurrent read-through fusion transcripts in breast cancer. *Breast Cancer Res Treat*. 2014;146(2):287–97. <https://doi.org/10.1007/s10549-014-3019-2>.
60. Lachmann A, Torre D, Keenan A, Jagodnik K, Lee H, Wang L, Silverstein M, Ma'ayan A. Massive mining of publicly available RNA-seq data from human and mouse. *Nat Commun*. 2018;9(1): <https://doi.org/10.1038/s41467-018-03751-6>.
61. Leek J, Scharpf R, Bravo H, Simcha D, Langmead B, Johnson W, Geman D, Baggerly K, Irizarry R. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*. 2010;11(10):733–9. <https://doi.org/10.1038/nrg2825>.
62. Nygaard V, Rødland E, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*. 2015;17(1):29–39. <https://doi.org/10.1093/biostatistics/kxv027>.
63. Johnson W, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118–27.
64. Zhang Y, Parmigiani G, Johnson W. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics Bioinforma*. 2020;2(3): <https://doi.org/10.1093/nargab/lqaa078>.
65. Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan H, Patel S, Ramage D, Segal A, Seth K. Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security CCS '17. New York, NY, USA: Association for Computing Machinery; 2017. p. 1175–91. <https://doi.org/10.1145/3133956.3133982>.
66. Li S, Tighe S, Nicolet C, Grove D, Levy S, Farmerie W, Viale A, Wright C, Schweitzer P, Gao Y, Kim D, Boland J, Hicks B, Kim R, Chhangawala S, Jafari N, Raghavachari N, Gandara J, Garcia-Reyero N, Hendrickson C, Roberson D, Rosenfeld J, Smith T, Underwood J, Wang M, Zumbo P, Baldwin D, Grills G, Mason C. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat Biotechnol*. 2014;32(9):915–25.
67. Amaratunga D, Cabrera J. Analysis of data from viral DNA microchips. *J Am Stat Assoc*. 2001;96(456):1161–1170.
68. Risso D, Ngai J, Speed T, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol*. 2014;32(9):896–902. <https://doi.org/10.1038/nbt.2931>.
69. Kammers K, Cole R, Tiengwe C, Ruczinski I. Detecting significant changes in protein abundance. *EuPA Open Proteomics*. 2015;7:11–9. <https://doi.org/10.1016/j.euprot.2015.02.002>.
70. Zhu Y, Orre L, Tran Y, Mermelekas G, Johansson H, Malyutina A, Anders S, Lehtiö J. DEqMS: a method for accurate variance estimation in differential protein expression analysis. *Mol Cell Proteomics*. 2020;19(6):1047–57. <https://doi.org/10.1074/mcp.tir119.001646>.
71. Myint L, Kleensang A, Zhao L, Hartung T, Hansen K. Joint bounding of peaks across samples improves differential analysis in mass spectrometry-based metabolomics. *Anal Chem*. 2017;89(6):3517–23. <https://doi.org/10.1021/acs.analchem.6b04719>.
72. Zhang X, Nieuwduin M, Groen A, Zwinderman A. Statistical evaluation of diet-microbe associations. *BMC Microbiol*. 2019;19(1): <https://doi.org/10.1186/s12866-019-1464-0>.
73. Robinson M, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11(3):25.
74. Li X, Cooper N, O'Toole T, Rouchka E. Choice of library size normalization and statistical methods for differential gene expression analysis in balanced two-group comparisons for RNA-seq studies. *BMC Genomics*. 2020;21(1): <https://doi.org/10.1186/s12864-020-6502-7>.
75. Zypych-Walczak J, Szabelska A, Handschuh L, Górczak K, Klamecka K, Figlerowicz M, Siatkowski I. The impact of normalization methods on RNA-seq data analysis. *BioMed Res Int*. 2015;2015:1–10. <https://doi.org/10.1155/2015/621690>.
76. Evans C, Hardin J, Stoebel D. Selecting between-sample RNA-seq normalization methods from the perspective of their assumptions. *Brief Bioinform*. 2017;19(5):776–92. <https://doi.org/10.1093/bib/bbx008>.
77. Nasirigerdeh R, Torkzadehmahani R, Matschinske J, Baumbach J, Rueckert D, Kaissis G. HyFed: hybrid federated framework for privacy-preserving machine learning. GitHub. 2021. <https://github.com/tum-aimed/hyfed>.
78. Dibert A, Csirmaz L. Infinite secret sharing – examples. 2014;8(2): <https://doi.org/10.1515/jmc-2013-0005>.

79. Tjell K, Wisniewski R. Privacy in distributed computations based on real number secret sharing. *CoRR*. 2021;abs/2107.00911. <http://arxiv.org/abs/2107.00911>.
80. Cover T, Thomas J. Elements of information theory. Inc.: John Wiley & Sons; 1991.
81. Chen Y, Lun A, Smyth G. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using rsubread and the edger quasi-likelihood pipeline. *F1000Res*. 2016;5(1438):1438.
82. Karr A, Lin X, Sanil A, Reiter J. Secure regression on distributed databases. *J Comput Graph Stat*. 2005;14(2):263–79.
83. Siangphoe U, Archer K. Estimation of random effects and identifying heterogeneous genes in meta-analysis of gene expression studies. *Brief Bioinform*. 2017;18(4):602–18.
84. Marot G, Foulley J-L, Mayer C-D, Jaffrézic F. Moderated effect size and p-value combinations for microarray meta-analyses. *Bioinformatics*. 2009;25(20):2692–9.
85. Heard NA, Rubin-Delanchy P. Choosing between methods of combining *p*-values. *Biometrika*. 2018;105(1):239–246.
86. Whitlock M. Combining probability from independent tests: the weighted z-method is superior to fisher's approach. *J Evol Biol*. 2005;18(5):1368–73.
87. Breitling R, Herzyk P. Rank-based methods as a non-parametric alternative of the t-statistic for the analysis of biological microarray data. *J Bioinform Comput Biol*. 2005;3(5):1171–89.
88. Prada C, Lima D, Nakaya H. MetaVolcanoR: Gene expression meta-analysis visualization tool. R Package version 1.8.0. 2019;1. <https://doi.org/10.18129/B9.BIOC.METAVOLCANOR>, <https://bioconductor.org/packages/MetaVolcanoR>.
89. Wang X, Kang D, Shen K, Song C, Lu S, Chang L-C, Liao S, Huo Z, Tang S, Ding Y, Kaminski N, Sibille E, Lin Y, Li J, Tseng G. An R package suite for microarray meta-analysis in quality control, differentially expressed gene analysis and pathway enrichment detection. *Bioinformatics*. 2012;28(19):2534–6.
90. Hong F, Breitling R, McEntee C, Wittner B, Nemhauser J, Chory J. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*. 2006;22(22):2825–7.
91. Zolotareva O, Nasirigerdeh R, Matschinske J, Torkzadehmahani R, Bakhtiari M, Frisch T, Späth J, Blumenthal D, Abbasinejad A, Tieri P, Kaissis G, Rückert D, Wenke N, List M, Baumbach J. Flimma: a federated and privacy-aware tool for differential gene expression analysis. *GitHub*. 2021. <https://github.com/ozolotareva/flimma>.
92. Zolotareva O, Nasirigerdeh R, Matschinske J, Torkzadehmahani R, Bakhtiari M, Frisch T, Späth J, Blumenthal D, Abbasinejad A, Tieri P, Kaissis G, Rückert D, Wenke N, List M, Baumbach J. Flimma: a federated and privacy-aware tool for differential gene expression analysis. *Zenodo*. 2021. <https://doi.org/10.5281/zenodo.5711972>.
93. Matschinske J, Alcaraz N, Benis A, Golebiewski M, Grimm D, Heumos L, Kacprowski T, Lazareva O, List M, Louadi Z, Pauling J, Pfeifer N, Röttger R, Schwämmle V, Sturm G, Traverso A, Steen K, de Freitas M, Silva G, Wee L, Wenke N, Zanin M, Zolotareva O, Baumbach J, Blumenthal D. The AI-Me registry for artificial intelligence in biomedical research. *Nat Methods*. 2021;18(10):1128–31. <https://doi.org/10.1038/s41592-021-01241-0>.
94. Perou C, Sørlie T, Eisen M, van de Rijn M, Jeffrey S, Rees C, Pollack J, Ross D, Johnsen H, Akslen L, Fluge O, Pergamenschikov A, Williams C, Zhu S, Lønning P, Børresen-Dale A, Brown P, Botstein D. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747–52.
95. Sørlie T, Perou C, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen M, van de Rijn M, Jeffrey S, Thorsen T, Quist H, Matese J, Brown P, Botstein D, Lønning P, Børresen-Dale A. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA*. 2001;98(19):10869–10874.
96. Herschkowitz J, Simin K, Weigman V, Mikaelian I, Usary J, Hu Z, Rasmussen K, Jones L, Assefnia S, Chandrasekharan S, Backlund M, Yin Y, Khrantsov A, Bastein R, Quackenbush J, Glazer R, Brown P, Green J, Kopelovich L, Furth P, Palazzo J, Olopade O, Bernard P, Churchill G, Van Dyke T, Perou C. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol*. 2007;8(5):76.
97. Bray N, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34(5):525–7. <https://doi.org/10.1038/nbt.3519>.
98. Gendoo D, Ratanasirigulchai N, Schröder M, Paré L, Parker J, Prat A, Haibe-Kains B. Genefu: an r/bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics*. 2015;32(7):1097–9. <https://doi.org/10.1093/bioinformatics/btv693>.
99. Čuklina J, Lee C, Williams E, Sajic T, Collins B, Martínez M, Sharma V, Wendt F, Goetze S, Keele G, Wollscheid B, Aebbersold R, Pedrioli P. Diagnostics and correction of batch effects in large-scale proteomic studies: a tutorial. 2021;17(8). <https://doi.org/10.15252/msb.202110240>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.