Supplementary material

Batch-normalization of cerebellar and medulloblastoma gene expression datasets utilizing empirically defined negative control genes

Holger Weishaupt[†*], **Patrik Johansson**[†], **Anders Sundström**[†], **Zelmina Lubovac-Pilav**[§], **Björn Olsson**[§], **Sven Nelander**[†], and **Fredrik J. Swartling**[†*]

[†]*Department of Immunology, Genetics amd Pathology, Science for Life Laboratory, Rudbeck Laboratory, Uppsala University, Uppsala, Sweden*
[§]*School of Bioscience / The Systems Biology Research Centre, University of Skövde, Skövde, Sweden*

[*]*Corresponding authors: holger.weishaupt@igp.uu.se, fredrik.swartling@igp.uu.se*

Contents

1. Approach

In order to establish a large-scale, integrated, and batch-corrected dataset comprising both normal cerebellar and MB samples, the present study was carried out along four main phases as follows:

- The literature was screened for relevant microarray gene expression datasets containing MB and/or normal cerebellar samples. A preprocessing framework was implemented to merge the data from different studies and platforms and establish MB subgroup affiliations for samples with missing information.
- The collected data was used to empirically define negative control genes, i.e. genes with low observed variation between or within phenotypes.
- Empirically defined negative control genes were employed to batch-correct the merged dataset using the Removal of Unwanted Variation (RUV) method (Gagnon-Bartsch and Speed, 2012; Jacob *et al.*, 2016). A range of RUV related regularization parameters were tested.
- Various metrics were implemented and utilized to evaluate the performance of the different batch-effect removal configurations and ultimately select a normalized dataset.

2. Supplementary methods

2.1 Collection of gene expression datasets

For purposes of mergability, only samples hybridized to the *HG-U133 Plus 2*, *Human Exon 1.0 ST*, *Human Gene 1.0 ST*, and *Human Gene 1.1 ST* Affymetrix arrays were selected, which encompassed the most frequently used platforms and included the majority of data. Datasets for these platforms were obtained from the Gene Expression Omnibus (GEO) (Barrett *et al.*, 2012) and ArrayExpress (AE) (Kolesnikov *et al.*, 2014) repositories, and each dataset will in the following be referred to by the respective GEO or AE accession codes. To restrict the number of batches that would have to be considered in the merging process, only datasets with more than five samples were considered. Furthermore, datasets, which were composed solely of samples already selected from another study, were also discarded from the collection.
The selected datasets were further pruned as follows. 15 SHH samples in GSE73038 were suspected to be present in GSE49243 and were subsequently excluded from GSE49243. There was also substantial overlap between GSE50765, GSE37382, and GSE85217. A total of 71 Samples were removed from GSE50765, because they were duplicated in GSE37382 and/or GSE85217. In turn, 235 samples were excluded from GSE37382, which were duplicated in GSE85217. Duplicated samples were mainly identified due to identical samples names or patient IDs, while a minority was removed due to an artificially high correlation with the gene expression profile (plus an agreement with clinical information) of already included samples.
As we only sought to include primary MB tumor samples, we excluded 2 relapse samples from GSE74195.

2.2 Processing of gene expression datasets

For all selected samples, raw CEL files were downloaded from GEO or AE. Subsequently, all raw CEL files from the same platform were processed together using the R/Bioconductor package *oligo* (Carvalho *et al.*, 2010) in conjunction with the RMA algorithm (Irizarry *et al.*, 2003). The *Human Gene 1.0 ST* and *Human Gene 1.1 ST* arrays were analysed at the *core* level, while the *Human Exon 1.0 ST* arrays were processed at the *extended* level. Subsequently, we mapped the identifiers of the *HG-U133 Plus 2* and *Human Exon 1.0 ST* to *Human Gene 1.0/1.1 ST* identifiers using *'Best Match'* information from Affymetrix (https://www.affymetrix.com/support/technical/byproduct.affx?product=hugene-1_0-st-v1). In addition, to increase the overlap between the *Human Exon 1.0 ST* and *Human Gene 1.0/1.1 ST* data we also inspected and added probe mappings from the *'Good Match'* and *'Complex Match'* files, including probes for the genes *MYCN*, *PTCH1*, *NPR3*, *UNC5D*, *DKK2*, and *GABRA5*. After mapping of probe identifiers within each platform, multiple rows mapping to the same identifier were collapsed using the mean value. Subsequently, all platform datasets were merged on probe identifiers, and gene symbols were assigned using the *hugene11sttranscriptcluster.db* package.
Multiple rows mapping to the same gene or multiple columns mapping to the same patient were collapsed using the mean value. Finally, the resulting gene expression matrix was quantile normalized using the respective function in the *preprocessCore* package.

2.3 Subgroup classification of MB samples

Classifications were conducted in R using the *Prediction Analysis for Microarrays* (PAM) classifier, via the respective implementation in the Bioconductor/R package *pamr*, and an *ElasticNet* classifier via the *glmnet* package.
Specifically, the PAM classifications were conducted on a set of 100 genes comprising 25 empirically defined signature genes for each MB subgroup. These classifier genes were estimated as follows. First, within each of the four datasets, GSE10327, GSE21140, GSE37418, GSE85217, the differential gene expression between a subgroup and each other subgroup was investigated using the *limma* package (Ritchie *et al.*, 2015) and for each gene the maximum FDR corrected p-value (q-value) and minimal fold change (FC) across all comparisons were recorded and used for further analyses. Secondly,

genes significantly upregulated ($q < 0.05, FC > 1.3$) in the subgroup as compared to the other subgroups were extracted and q-values converted to ranks. Thirdly, for each subgroup, the intersection of significantly upregulated genes was extracted from the four studies, re-ranked according to the mean rank across the four datasets, and the 25 top-ranking genes were extracted as signature genes for that subgroup. The subsequently established PAM classifiers were always trained on only these 100 signature genes.
The ElasticNet classifier was implemented by setting the penalty $\alpha = 0.9$ and was instead applied to all genes in the dataset.

2.3.1 Cross-validation

In order to evaluate the existing MB subgroup labels in the dataset, a cross-validation analysis was conducted. For this purpose, the PAM and ElasticNet classifiers were trained and applied to all MB samples with an existing subgroup label, using leave-one-out classifiers to obtain a class prediction for each individual sample. Samples, which were correctly classified by both PAM and ElasticNet classifiers, were considered most reliable. For these samples, the supplied subgroup affiliations were retained, while subgroup labels of samples, which were incorrectly classified by at least one of the two classifiers, were removed for the downstream analyses, thus leaving those samples effectively unlabeled.

2.3.2 Classification of MB samples with unknown subgroup affiliation

After removing the subgroup labels from samples with unreproducible subgroup labels, new PAM and ElasticNet classifiers were trained on all samples with retained MB subgroup labels, and the classifiers were applied to the 137 samples, for which no subgroup was originally supplied. Samples obtaining the same class prediction by both classifiers were labeled with the respective class, while other samples were left unlabeled.

2.4 Visualization of batch-effects in merged data

To inspect the existence of batch-effects present in the dataset after merging of studies and platforms, three visualization tools were used on the gene expression data: (i) Relative Log Expression (RLE) plots (Brettschneider *et al.*, 2008; Gandalfo *et al.*, 2017) , (ii) a scatter plot of data after dimensional reduction through Multi-Dimensional Scaling (MDS) (Cox *et al.*, 2008), and (iii) hierarchical clustering (HC) (Anderberg *et al.*, 2014).

2.4.1 Relative log expression (RLE) plots

The Relative log expression (RLE) denotes a measure of deviation of the $\log_2$ expression value of a gene in a single sample from the median of $\log_2$ transformed expression values of that gene computed across all samples in an expression matrix (Brettschneider *et al.*, 2008; Gandalfo *et al.*, 2017). Formally, let x_{is} denote the expression of gene i in sample s given in normal scale, and let $\mathbf{x}_{i*}$ be the vector holding the expression values of gene i from all samples in the expression matrix. Then the RLE of x_{is} is defined as (Gandalfo *et al.*, 2017)

$$RLE(x_{is}) = \log_2(x_{is}) - \text{median}(\log_2(\mathbf{x}_{i*})).$$

RLE plots are then typically shown as box-plots, where each box corresponds to the distribution of RLE values within one sample. Thus, this method represent a visual tool employed to illustrate the heterogeneity or variation of gene expression distributions between samples. In a batch-free dataset, samples are generally expected to show comparable RLE distributions, while the presence of batches might cause discernible differences between distributions.
However, to allow a visualization for the large number of samples gathered in this project, a simplified version was implemented, in which only the median RLE value, the region between the first ($Q1$) and third (Q_3) quantiles, and the range between minimum and maximum RLE values (excluding outliers) were computed following Tukey's box plot paradigm (Frigge *et al.*, 1989). Specifically, Q_1 and Q_3 were taken to be the 25% percentile and 75% percentile of the RLE values in a sample, respectively, and the minimum and maximum RLE values were computed as $Q1 - 1.5 \cdot IQR$ and $Q_3 + 1.5 \cdot IQR$, respectively, where $IQR = |Q3 - Q1|$ is the interquartile region.

2.4.2 Multi-dimensional scaling (MDS)

A Multi-dimensional scaling (MDS) of the data down to two or three dimensions was computed via the *isoMDS* function from the R package *MASS* (Venables *et al.*, 2002), using all samples and the 1200 genes with highest standard deviation across samples. A scatter plot of the MDS dimensions was then utilized to inspect the overall clustering of the data based on platforms, studies and phenotypes.

2.4.3 Hierarchical clustering (HC)

While the MDS was employed to illustrate similarities between all samples, Hierarchical clusterings (HC) was utilized more specifically to evaluate whether the MB samples clustered according to their tumor subgroup in accordance with previous class discovery studies (Cho *et al.*, 2011; Kool *et al.*, 2008; Northcott *et al.*, 2011; Thompson *et al.*, 2006). To this end, HC

was conducted only on MB samples, again using the 1200 most variable genes as measured by the standard deviation of expression values across these samples. Specifically, clustering was performed using the *hclust* function from the *fastcluster* (Müllner *et al.*, 2013) package in R, using Euclidian distances and complete linkage. Results were visualized using the *heatmap.3* package (Zhao *et al.*, 2014).

2.5 Batch-effect removal

The *naiveRandRUV* function of the R/Bioconductor package *RUVnormalize* (Jacob *et al.*, 2016) was used to correct for batch effects in the merged gene expression data. Appart from a matrix holding the raw expression data, the function takes as input the column indices of negative control genes and three regularization parameters: the regulization strength (*nu.coeff*), the assumed number of independent sources of unwanted variation (k), and a tolerance parameter (*tol*).
Negative control genes were estimated empirically, as described below. To identify a suitable selection of regularization parameters a range of values for $nu.coeff \in \{i \cdot 10^j; i \in \{1,2,\ldots,10\}, j \in \{-5,-4,-3\}\}$ and $k \in \{3,4,\cdots,23\}$ were used, while the default value for tol was used. For each combination of regularization parameters, the expression data was processed using the *naiveRandRUV* method, and the performance of the batch-effect removal was quantified according to different metrics (described in section 2.7).

2.6 Negative control gene identification

Negative controls gene in the RUV sense are genes, which are expected to show almost no changes in expression over the conditions of interest (Gagnon-Bartsch and Speed, 2012). Thus, variations in expression levels of such genes between different datasets can be utilized as a means to detect and correct for batch effects.
While housekeeping genes have been suggested as a potential source of negative control genes (Gagnon-Bartsch and Speed, 2012), it is unclear how applicable such genes are for MB, considering that housekeeping genes are typically derived from adult tissues under normal conditions (Eisenberg and Levanon, 2013). To obtain a set of negative controls to be used in the present project we thus aimed here at empirically defining such controls by identifying genes with low variation of gene expression within and between any of the investigated phenotypes.
Specifically, we established three rank-scores, referred to as F_W, which measures the amount of expression variation within a phenotype, F_{B_1}, which measures the amount of expression variation between MB subgroups, and F_{B_2}, which measures expression variation between MB and normal brain (The indices B_1 and B_2 are here used to distinguish between the first and second type of *between-phenotype* variance measures). The computation of the three measures is described below and illustrated in **Supp. Fig. 3**. Given the set $G = \{1,2,3,...,g\}$ containing the indices of all genes in the dataset, the overall score F_{total} for gene $i \in G$ was then obtained from the three individual scores as

$$F_{total}(i) = \frac{F_W(i) + F_{B_1}(i) + F_{B_2}(i)}{3}.$$

A low score corresponds to a generally lower amount of gene expression variation within and between subgroups, while a high score implies more variation. Accordingly, negative control genes were selected as the genes with lowest values of F_{total}.

2.6.1 Measuring expression variation within phenotypes

In order to score genes based on how stable their expression is within phenotypes, we calculated one measure of dispersion within each set of samples belonging to the same phenotype and study. To avoid a bias towards genes with low average expression, dispersion was here computed in terms of the Relative Mean absolute Deviation (RMD) defined as

$$RMD(i) = \frac{\frac{1}{n}\sum_{j=1}^{n} |x_i(j) - \bar{x_i}|}{|\bar{x_i}|},$$

where $x_i(j)$ measures the gene expression of gene i in sample j, and n denotes the total number of samples.
Now, let p be an index over the different phenotypes and s be an index over all studies. For every combination of s and p that includes at least 5 samples, we define $RMD_s^p(i)$ as the RMD of gene i across all samples belonging to study s and phenotype p; for combinations not satisfying this criterion we set $RMD_s^p(i) = 0$. Then the dispersion measure for gene i in study s was obtained as

$$RMD_s(i) = \max_p(RMD_s^p(i)).$$

Subsequently, the final score F_W was obtained by first calculating the maximum dispersion across all studies

$$RMD_{max}(i) = \max_s(RMD_s(i)),$$

and ranking genes based on the RMD measure as

$$F_W(i) = \left(\text{rank}^{\uparrow}(RMD_{max})\right)_i,$$

where $\text{rank}^{\uparrow}(\cdot)$ denotes the fractional rank assigned to values in increasing order. The final score $F_W(i)$ for a gene i measures variation within subgroups, with a low rank reflecting low variation and a high rank reflecting large variation.

2.6.2 Measuring expression variation between MB subgroups

In order to score genes with respect to how stable their expression was across MB subgroups, we instead performed a one-way analysis of variance (ANOVA) between subgroup specific expression means within each study, which contains at least 5 samples of each subgroup. The datasets in question were GSE10327, GSE21140, GSE37418, GSE73038, and GSE85217. Specifically, for each of those studies s the ANOVA related score for a gene i was computed as

$$AOV_s^{B_1}(i) = -\log_{10}(p(i)),$$

where $p(i)$ denotes the p-value of the ANOVA for gene i and the label B_1 was simply added to distinguish this measure from the computation of AOV scores between normal controls and MB (see below). To account for unequal variances, we employed the *oneway.test* implementation of ANOVA in R.
Subsequently, the maximum AOV score across all studies s was computed as

$$AOV_{max}^{B_1}(i) = \max_s(AOV_s^{B_1}(i)),$$

and the final score was obtained by ranking genes with respect to the maximum AOV scores as

$$F_{B_1}(i) = \left(\text{rank}^{\uparrow}(AOV_{max}^{B_1})\right)_i,$$

with a low rank implying that gene i shows relatively little variation between subgroups.

2.6.3 Measuring expression variation between MB and normal brain

The scoring of genes with respect to variations between MB and normal brain gene expression was conducted similar as above. Specifically, in studies that contained at least five MB samles and five normal brain samples, we performed an ANOVA through the *oneway.test* function (which in this case is equivalent to conducting a Welch's t-test) comparing the mean expression of MB samples (regardless of subgroup) against the mean expression of normal brain samples. The datasets in question were EMTAB292 and GSE74195.
In each study s, the ANOVA related score for a gene i was computed as

$$AOV_s^{B_2}(i) = -\log_{10}(p(i)),$$

where $p(i)$ again denotes the p-values of the ANOVA for gene i. Subsequently, the maximum AOV score across all studies s was computed as

$$AOV_{max}^{B_2}(i) = \max_s(AOV_s^{B_2}(i)),$$

and the final score was obtained by ranking genes with respect to the maximum AOV scores as

$$F_{B_2} = \left(\text{rank}^{\uparrow}(AOV_{max}^{B_2})\right)_i,$$

with a low rank implying that gene i shows relatively little variation between MB and cerebellum.

2.6.4 Comparison with house-keeping genes

The empirically derived control genes were compared to house-keeping genes with respect to: (1) total overlap of genes and (2) performance when used as negative controls in the RUV normalization. For that purpose, a set of 575 house-keeping genes (Eisenberg and Levanon, 2003), which we will refer to as HKG2003, and a set of 3804 house-keeping genes (Eisenberg and Levanon, 2013)), which we will refer to as HKG2013, were downloaded from https://www.tau.ac.il/~elieis/HKG/. For the HKG2003 set, RefSeq identifiers were mapped to approved gene symbols using the HUGO Gene Nomenclature Committee (HGNC, https://www.genenames.org/). After mapping the house-keeping gene sets to the genes retained in the merged expression datasets, a total of 314 (HKG2003) and 3074 (HKG2013) house-keeping genes were available for downstream comparisons.

2.7 Evaluation of normalizations

In order to estimate the existence of batch effects in the raw data and to determine how well such batch-effects have been removed by a particular configuration of the RUV normalization, a number of evaluation metrics were employed.

2.7.1 Standard deviation of median RLE values (σ_{mRLE})

To obtain a quantitative metric measuring one aspect of heterogeneity in RLE plots, first the median RLE value was computed for each sample and then the standard variation across those median values (σ_{mRLE}) was considered.

2.7.2 Intra- to inter-group distances (IIGD)

In addition to clustering, we also aimed to investigate the overall similarities of gene expression profiles of samples within the same and between phenotypes. Specifically, if batch effects lead to artificial differences between samples within the same phenotype or an artificial clustering of samples due to platform rather than phenotype, then batch-effect removal might cause expression profiles between samples of the same phenotype to become more similar, and/or the ratio of distances of expression profiles within the same phenotype to the distances between phenotypes to decrease. To quantify such properties, we calculate two types of mean distances for each phenotype, where distance is measured in terms of Euclidean distance. Specifically, let $\{p_k;\ k = 1,2,\cdots,l\}$ denote the l unique phenotypes, let S_k denote the set of samples belonging to phenotype p_k, and let S_k^C denote the set of samples not belonging to phenotype p_k. Let $x_k^u(i)$ further denote the expression of gene i in sample u. We then calculated first the mean Euclidean distance of expression profiles between pairs of samples within the same phenotype p_k as

$$\bar{D}_W(p_k) = \frac{2}{|S_k|(|S_k|-1)} \sum_{t=1}^{|S_k|-1} \sum_{w=t+1}^{|S_k|} \sqrt{\sum_{i=1}^{g} (x_k^t(i) - x_k^w(i))^2},$$

where g denotes the number of genes in the dataset. The mean Euclidean distance of expression profiles of samples from phenotype p_k to samples of other phenotypes was equivalently calculated as

$$\bar{D}_B(p_k) = \frac{1}{|S_k||S_k^C|} \sum_{t \in S_k} \sum_{w \in S_k^C} \sqrt{\sum_{i=1}^{g} (x_k^t(i) - x_k^w(i))^2}.$$

The final metric, denoted as IIGD (Inter to Intra Group Distances), was then computed as the ratio

$$IIGD = \frac{1}{l} \sum_{k=1}^{l} \frac{\bar{D}_W(p_k)}{\bar{D}_B(p_k)}.$$

2.7.3 K-means clustering and Adjusted Rand Index (ARI)

In order to evaluate how well the actual clustering of MB samples corresponds to the ideal clustering, in which all samples belonging to the same subgroup would fall into one distinct cluster, we employed the Adjusted Rand Index. Specifically, a clustering into $k = 4$ clusters was performed using the *kmeans* function in R. Subsequently, the *adjustedRandIndex* from the *mclust* package was utilized to compare the cluster affiliation of samples to the ideal sequence of labels, in which each cluster only contained samples from one unique subgroup.

2.7.4 Entropy

We hypothesized that the ARI for a particular clustering might produce a good result, even if within a subgroup specific cluster samples would agglomerate due to platform. However, in the batch-effect free scenario and assuming a uniform distributions of subgroups across platforms, we would not expect such a clustering due to platform. To distinguish between such cases, we employed here a measure of entropy applied on the sequence of platform labels obtained from a hierarchical clustering.
Specifically, let $L = (l_1, l_2, \ldots, l_{n-1}, l_n)$ denote the sequence of platform labels ordered based on the sample ordering in the dendrogram established by the hierarchical clustering. Let L_i with $i = (1,2,\ldots,n-50,n-49)$ be subsequences obtained by applying a sliding window on L, such that $L_i = (l_i, l_{i+1}, \ldots, l_{i+48}, l_{i+49})$. Let $\{k_i^j;\ j = 1,2,\cdots,m\}$ denote the m unique platform labels included in L_i. Then Shannon's entropy for subsequence L_i is defined as

$$H(L_i) = -\sum_{j=1}^{m} P(k_i^j) \log_2(P(k_i^j)),$$

which we have calculated in R using the *entropy* package.

The final metric $\bar{H}$ is then obtained as

$$\bar{H} = \sum_{i=1}^{n-49} H(L_i).$$

A high value of $\bar{H}$ corresponds then to a more uniform distribution of platform labels across the clustered samples, while a low value of $\bar{H}$ implies an agglomeration of samples due to platform in the hierarchical clustering.

2.7.5 Accuracy of Support Vector Machine classifications (SVM)

The accuracy of phenotype classifications within the merged data was evaluated as follows. 50 training samples, encompassing 10 normal brain samples, 10 WNT, 10 SHH, 10 G3, and 10 G3 samples, were randomly selected from the merged dataset and used to train a SVM classifier using the *e1071* package in R. 50 additional samples with the same number of phenotype labels were selected randomly without repetition and served as test data. The fraction of correct class predictions were recorded and averaged over ten classification runs with different random training and test sets and initial configurations.

2.7.6 Overlap of differentially expressed genes with positive control genes (OPG)

The differential expression of positive control genes in the raw and batch-corrected data was tested as follows. Positive control genes for each MB subgroup were estimated through differential gene expression analyses against samples from other MB subgroups using the *limma* package in R. Specifically, for every subgroup one list of upregulated genes was obtained from each of five studies (GSE10327, GSE21140, GSE37418, GSE73038, and GSE85217), which contained at least 8 samples from each of the four MB subgroups. The intersection of the five lists was then considered the set of positive control genes for the subgroup. Subsequently, an analogous differential expression analysis, but including all MB samples regardless of study or platform, was conducted on the batch normalized data. For each subgroup the fraction of significantly upregulated positive control genes was calculated and the final metric was obtained as the mean across the four subgroup related fractions.

2.8 Evaluation of overall strategy on independent training and test datasets

In the main text, the NCGs used for the RUV normalization were derived from the same dataset, to which the normalization was applied. In ensure the feasibility of this strategy, we sought to validate the approach by instead using two independent datasets for NCG identification and batch effect removal. For this purpose, we split the merged dataset comprising the 1641 samples into two separate datasets, one of which was only used to define NCGs (and positive control genes as described further below), while the other dataset was then normalized via the RUV method and using the NCGs determined from the first dataset. The details and individual steps of this validation experiment are outlined in the following.

2.8.1 Splitting of datasets into training and testing data

In order to obtain two independent datasets for validation, we proposed to split the 23 included studies into two separate sets, such that one set could be used for NCG identification, while the other was to be batch-corrected using the derived NCGs. The former dataset will in the following be referred to as training dataset, while the latter will be referred to as testing dataset. The proposed method for NCG selection is based on the calculation of three different metrics, i.e. gene expression variations (i) within phenotypes, (ii) between MB subgroups, and (iii) between MB and normal cerebellar controls, which are computed within individual studies to avoid batch effects. Accordingly, the training dataset was required to contain studies with the relevant composition of MB subgroups and studies that included both MB and normal cerebellar samples. Furthermore, in order to ensure that the testing data displays sufficient levels of batch effects, it should also comprise studies broadly distributed across all technical platforms. To satisfy these requirements, we selected a total of seven studies (including 958 samples) to represent the training dataset, while the remaining 16 studies (619 samples) were assigned to the testing dataset (**Supp. Table 5**).

2.8.2 Selection of negative control genes

Subsequently, NCGs were selected according to the strategy outlined in **section 2.6** of the supplementary methods and **Supp. Fig. 3**, with the exception that only the datasets GSE37418 and GSE85217 were used for the step *'Measuring expression variation between MB subgroups'*. To inspect whether these NCGs, derived from the training dataset, displayed the expected expression profiles across phenotypes, studies, and platforms in the testing data, we generated two plots. Specifically, we selected the two NCGs with the highest MAD across all samples in the testing data and plotted their expression levels for all samples separated based on phenotype, study, and phenotype in the testing dataset (**Supp. Fig. 6A**). Furthermore, we calculated and plotted for each NCG a measure of expression variation between phenotypes within the same study, between studies within the same platform, and between platforms (**Supp. Fig. 6B**). Both results suggest that the NCGs, despite having been independently derived from the training data, show even in the testing data comparably little variation between phenotypes within studies, but increasingly more variation between studies and platforms. Thus, while the NCGs

have been identified from the training data, they display the desired expression pattern to be considered feasible NCGs for batch-correcting the testing dataset.

2.8.3 Evaluation of normalization performance

To evaluate the independently derived NCGs with respect to the RUV batch-normalization of the testing data, we started again by performing the respective RUV-normalization using the same range of parameters as described in **section 2.5**. Subsequently, we computed the six performance metrics described in **section 2.7** for the raw data and the batch-corrections, which were conducted using either the empirically derived NCGs or three different types of controls: (i) 314 house keeping genes proposed by Eisenberg and Levanon (2003) and retained in the merged data (*HKG*), (ii) the 372 genes with the lowest expression RMD values calculated across all samples in the testing dataset (*Ctrl1*), and (iii) 372 genes chosen randomly (*Ctrl2*). The positive control genes, needed for the computation of the OPG metric, were extracted as described above, but only using the studies GSE37418 and GSE85217 included in the training dataset. Considering especially the results for the ARI, SVM, and OPG metrics, the analysis suggested that the RUV-normalization using the NCGs, despite the fact that they had been derived from the independent training dataset, produced generally better results on the testing data than the other three sets of control genes (**Supp. Fig. 7**).

2.8.4 Visualization of batch-effects in raw and RUV-normalized datasets

By utilizing the independently derived NCGs for the RUV-normalization and employing the four metrics, ARI, IIGD, Entropy, and σ_{mRLE}, and visual inspections, we obtained a final batch-corrected version of the testing data. By visualizing, via the use of the RLE, MDS and HC plots, and comparing the batch effects between the raw testing data (**Supp. Fig. 8**) and the final batch-correction of the testing data (**Supp. Fig. 9**), we found that the strategy was capable to remove a large amount of batch effects in the testing data, thus validating the proposed batch-correction strategy.

3. Availability of normalized data

The batch-corrected data have been deposited in the Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/, together with (i) the original MB subgroup affiliations, (ii) the reclassified MB subgroup labels assigned in this study, and (iii) all originally supplied clinical information for individual samples. The data is available through the GEO accession number GSE124814.

4. Supplementary tables and figures

Supplementary table 1

Supp. Table 1: Selected gene expression datasets

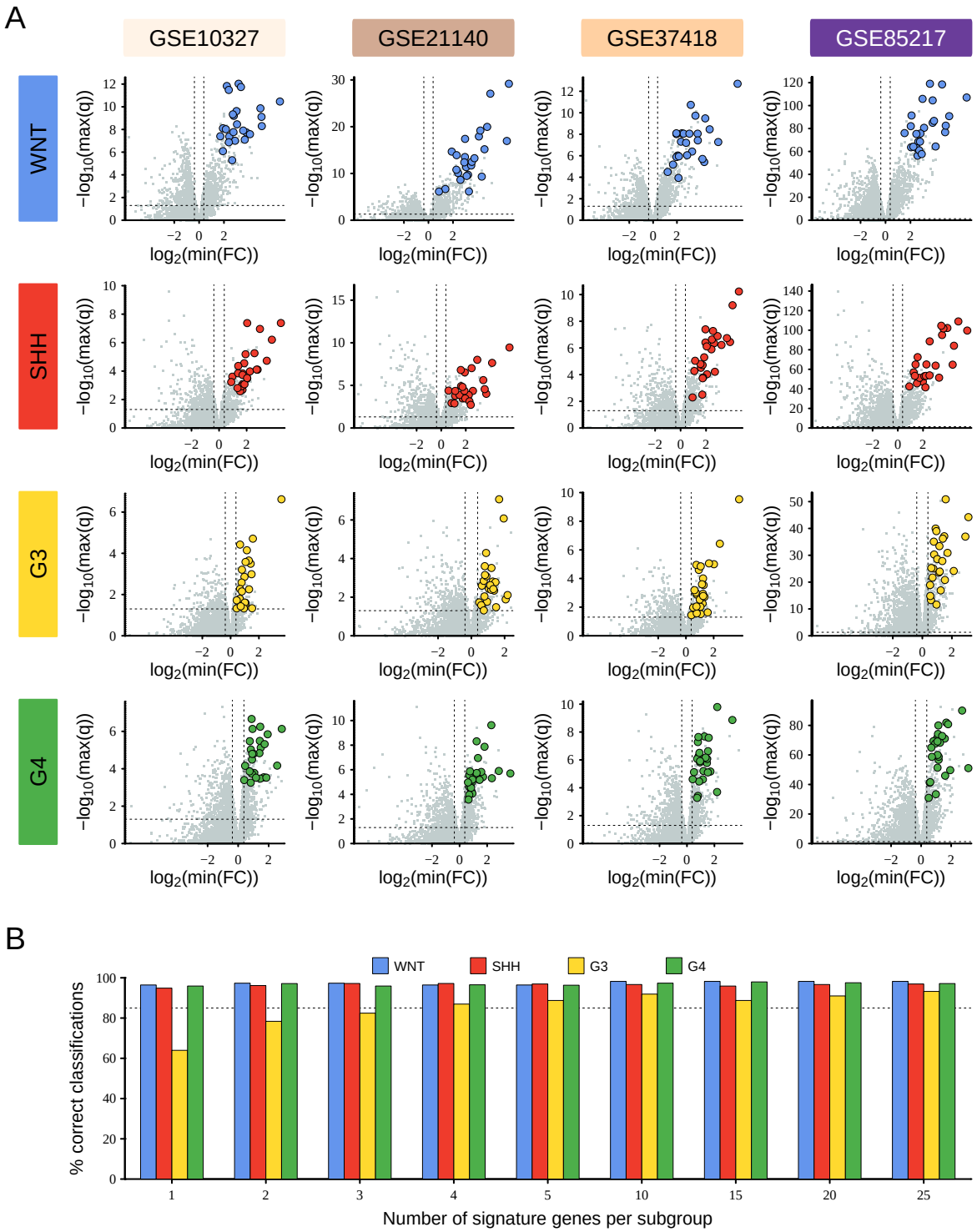
		Number of samples*							
GEO/AE ID	**Platform**	**Normal†**	**WNT**	**SHH**	**G3**	**G4**	**MB: unk.§**	**Tot.**	**Reference**
GSE3526	HG-U133-Plus-2	9						9	(Roth *et al.*, 2006)
GSE4036	HG-U133-Plus-2	14						14	-
GSE10327	HG-U133-Plus-2		9	15	11	27		62	(Kool *et al.*, 2008)
GSE12992	HG-U133-Plus-2						40	40	(Fattet *et al.*, 2009)
GSE37418	HG-U133-Plus-2		8	10	16	39	3	76	(Robinson *et al.*, 2012)
GSE44971	HG-U133-Plus-2	9						9	(Lambert *et al.*, 2013)
GSE49243	HG-U133-Plus-2			58				58	(Kool *et al.*, 2014)
GSE50161	HG-U133-Plus-2	2		8	4	7	3	24	(Griesinger *et al.*, 2013)
GSE67850	HG-U133-Plus-2						22	22	(Ho *et al.*, 2015)
GSE73038	HG-U133-Plus-2		10	16	10	10		46	(Sturm *et al.*, 2016)
GSE74195	HG-U133-Plus-2	5					25	30	(deBont *et al.*, 2008)
EMTAB292	HuEx-10	5					14	19	(Menghi *et al.*, 2011)
GSE21140	HuEx-10		8	33	27	35		103	(Northcott *et al.*, 2011)
GSE25219	HuEx-10	51						51	(Kang *et al.*, 2011)
GSE60862	HuEx-10	130						130	(Ramasamy *et al.*, 2014; Trabzuni *et al.*, 2013)
GSE22569	HuGene-10	22						22	(Liu *et al.*, 2012; Somel *et al.*, 2011)
GSE30074	HuGene-10						30	30	(Park *et al.*, 2011)
GSE35974	HuGene-10	44						44	(Chen *et al.*, 2013)
GSE41842	HuGene-10		6	3	2	8		19	(Gokhale *et al.*, 2010)
GSE37382	HuGene-11			11	6	33		50	(Northcott *et al.*, 2012a)
GSE50765	HuGene-11			12				12	(Vanner *et al.*, 2014)
GSE62803	HuGene-11		1	1	2	4		8	(Morrissy *et al.*, 2017)
GSE85217	HuGene-11		70	223	144	326		763	(Cavalli *et al.*, 2017)

*: Where the number of samples corresponds to the number of unique patients in a study, ignoring duplicated samples for the same patient.

†: Samples labeled as *"Normal"* designate normal cerebellar controls.

§: Samples labeled as *"MB: unk."* refer to MB samples, for which no subgroup label could be obtained from the respective dataset record or the accompanying publication.

Supplementary figure 1



Supp. Fig. 1. Selection of classifier genes. **A)** Volcano plots showing the results of differential expression analyses between MB subgroups (rows) within four different datasets (columns). For a given subgroup and dataset, three *limma* analyses were performed, comparing the subgroup to each of the other three subgroups in this dataset. The plots depict the maximum q-value (FDR corrected p-value) and minimal FC across the three analyses. The horizontal reference line indicates the $q = 0.05$ threshold, while the vertical lines indicate $\log FC$ thresholds of $-\log_2(1.3)$ and $\log_2(1.3)$, respectively. Colored data points indicate the 25 subgroup specific classifier genes used for classification analyses. **B)** The percentages of samples correctly classified by the PAM classifier as a function of the number of top signature genes used for the classification. The dashed reference line indicates the 85% level.

Supplementary table 2

Supp. Table 2: Signature genes identified through differential expression analyses

WNT	SHH	G3	G4
WIF1	PDLIM3	GABRA5	SH3GL3
TMEM51	CYYR1	SORBS2	RBM24
ADAMTSL1	KIAA0922	NXPH4	RND1
GAD1	SFRP1	SMARCD3	KIAA0319
RUNX2	EYA1	ARL6	PTPN5
ZNRF3	NDP	PNPLA3	CAP2
TMEM132C	PPP2R2C	NPR3	ANKS1B
P4HA2	NDST3	PCOLCE2	SLC10A4
TNFRSF19	ZC3H12C	GABRB3	PDZD4
TNC	ATOH1	TRIP10	SPTAN1
FZD10	HHIP	DOCK9	MID2
NKD1	NRIP2	SSX2IP	KIAA2022
ADAM12	PREX1	TDRP	SH3BP5
LRP4	PRLR	PALMD	NEUROD2
PYGL	ANKRD6	RABGAP1L	THRA
FBXL7	PBX4	MCF2L2	ST18
LAMP5	TEX15	GPD1L	RPH3A
MAF	SRGAP1	ARL4D	TMEM35A
RASL11B	CPLX1	INHBB	RAPGEFL1
RTTN	ARHGEF26	PYY	STXBP1
DKK2	PTX3	NRL	MPP3
OSR2	ABCB4	RALGPS2	SLC9A6
RAI2	ZNF516	TSHZ3	RALGPS1
TMEM2	TMEM144	TBX21	RNF144A
IFT57	KIF21A	ARHGAP9	BLCAP
PGM5	GRIA4	FGF11	GPR12

Genes highlighted in color overlap with a previously published set of 22 signature genes (Northcott *et al.*, 2012b).

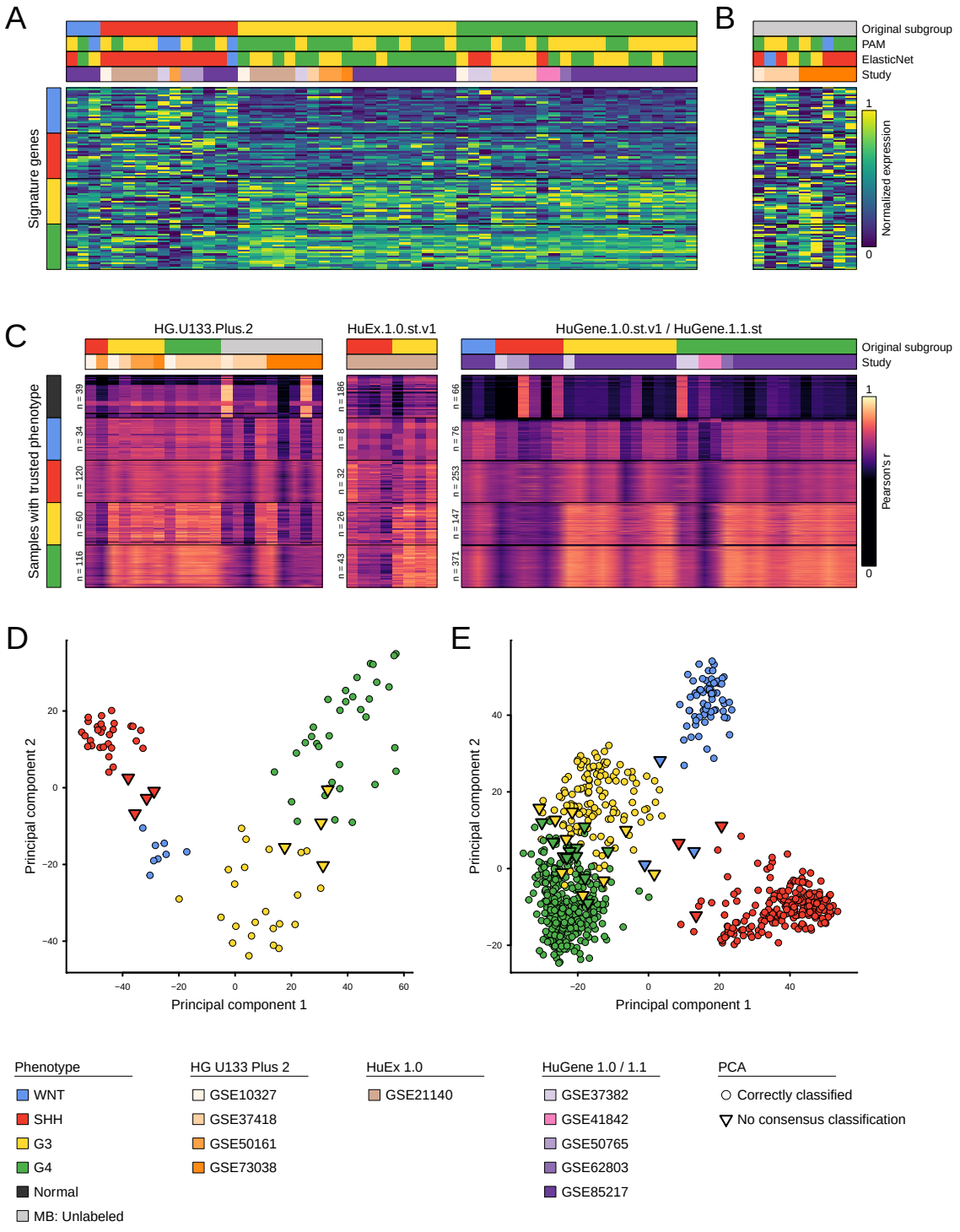
Supplementary table 3

Supp. Table 3: Coefficients used by the *Elastic Net* classifier

WNT	SHH	G3	G4
ADAMTSL1	CYYR1	C2orf71	BARHL1
AMHR2	EHD1	DCT	EPHB1
DLX3	EYA1	DENND1B	EXPH5
FZD10	NDP	EML1	GRM8
GAD1	NRIP2	GABRA5	HTR2C
MYH15	PDLIM3	GSG1	KIAA0319
NKD1	SFRP1	GUCA1C	KLRD1
OSR2	ZNF516	HLX	LINC01105
PGM5		LMO1	NEUROD2
RUNX2		NPFFR2	NID2
TGFA		NPR3	PTPN5
TMEM2		NXPH4	RAPGEF2
WIF1		PYY	RAPGEFL1
		RHO	RBM24
		RIMS2	RPH3A
		TBX21	SH3GL3
		TRIP10	SHC4
		USP2	SIX6
			SNCAIP
			STOX2
			SYCP1
			TES
			TFAP2D
			TMEM192
			TSPAN2

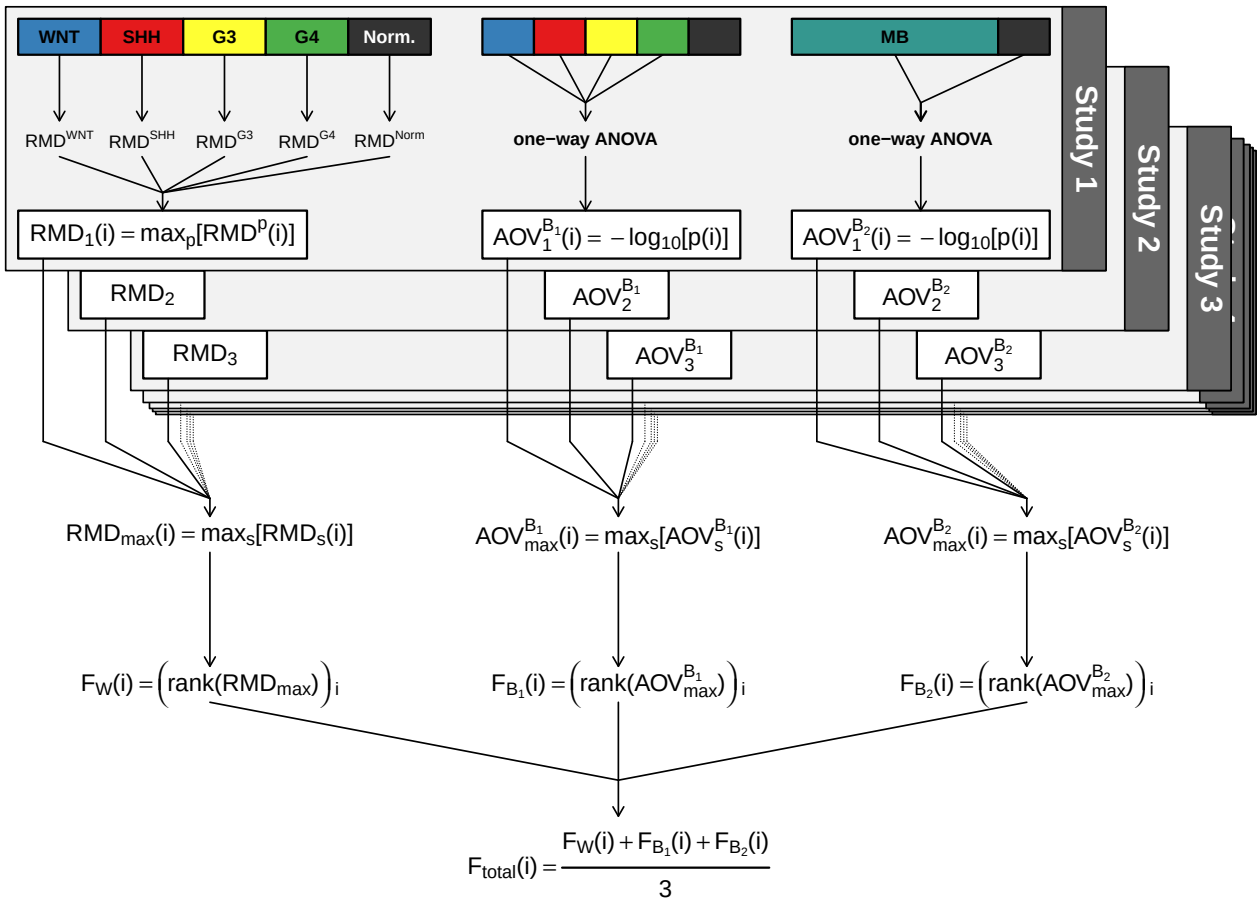
Genes highlighted in color overlap with a previously published set of 22 signature genes (Northcott *et al.*, 2012b).

Supplementary figure 2



Supp. Fig. 2. Inspection of samples with lacking consensus MB subgroup prediction. **A)** Heat map showing the expression of 100 signature genes in 55 MB cases, which exhibited subgroup affiliations but could not be correctly classified by both the PAM and Elastic Net classifiers. **B)** Heat map showing the signature gene expression in 9 MB samples, which lacked previous subgroup affiliations and could not be robustly classified by the two classifiers. **C)** Heatmap showing pairwise Pearson's correlation coefficients comparing each of the 64 samples, which could not be robustly classified by the two classifiers, against all those samples, which belonged to the same platform and exhibited trustworthy phenotype labels (either normal cerebellar controls or correctly classified MB subgroup labels). **D-E)** Biplots showing the results of principal component analyses performed on all samples in the GSE21140 **(D)** or the GSE85217 **(E)** datasets and utilizig the 1200 most variable genes in each dataset, respectively. Samples from these datasets, which could not be robustly classified by both PAM and ElasticNet classifiers, are drawn using triangles.

Supplementary figure 3



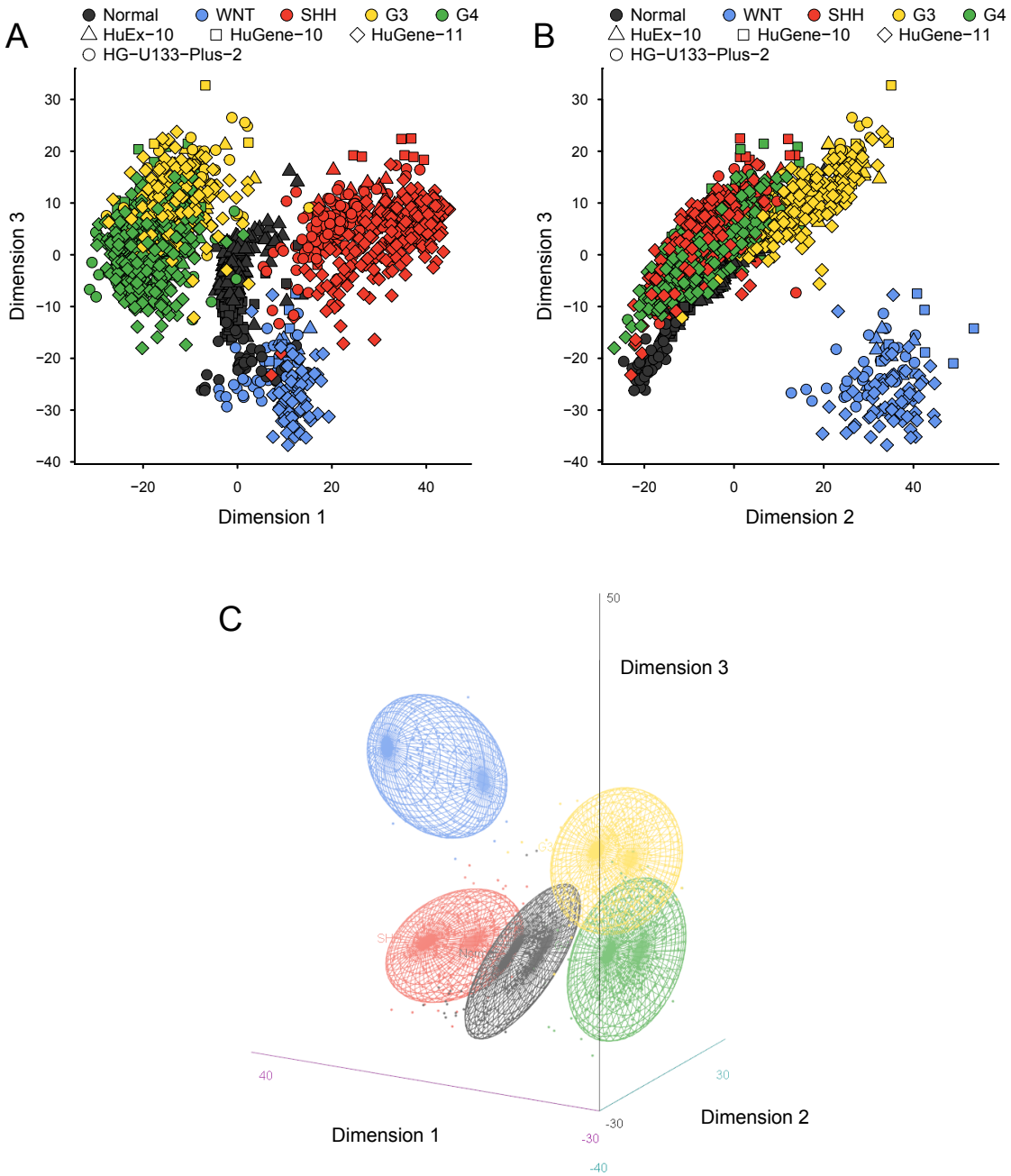
Supp. Fig. 3. Flow-chart visualizing the procedure and metrics employed to select empirical negative control genes. For more details, please refer to supplementary methods section.

Supplementary table 4

Supp. Table 4: List of 372 empirically defined negative control genes

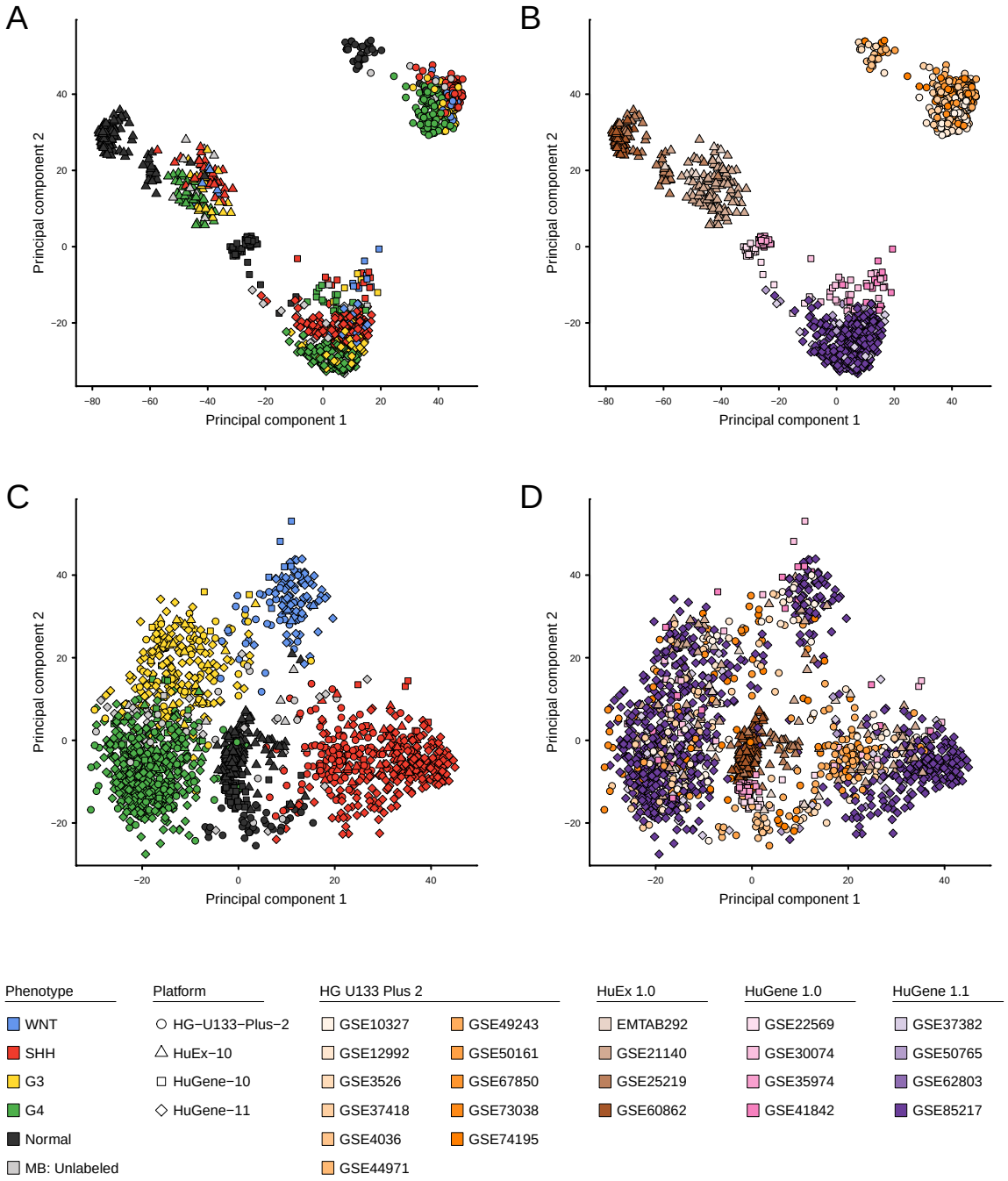
A4GNT	CCDC12	ETV7	KRT24	NLRP13	S100PBP	TNF
ABCB11	CCDC127	EXOSC6	KRT4	NLRP8	SCART1	TNP2
ABCB5	CCDC130	FAM110D	KRT72	NPHS2	SCGB2A2	TPP1
ABHD11	CCER1	FAM129C	KRTAP3-1	NR0B1	SEBOX	TPPP2
ABI3	CCL1	FAM205BP	LAD1	NR1I2	SERPINB7	TRAF3IP3
ACO2	CCL13	FAM49B	LAYN	NSFL1C	SH3GL1	TREML1
ACTL8	CCL21	FAM71A	LEXM	NSUN2	SIGLEC11	TRIM29
ACTR10	CD1E	FAM91A1	LGALS8-AS1	NUDT18	SIRPB2	TRIM41
ACVR1B	CD244	FAM9C	LILRA1	NXF1	SLA2	TRIM42
ADCY10	CD40LG	FANCD2OS	LILRA5	ODF3L2	SLC16A8	TRIML1
ADGRE3	CD7	FASLG	LINC00301	OGFR	SLC22A18AS	TRMU
ADGRG5	CDC42SE2	FBXO39	LINC00304	OGG1	SLC25A25	TRPC4AP
ADRA1D	CDSN	FCRL2	LINC00523	OPN4	SLC28A1	TSACC
ADRA2B	CEACAM7	FCRL3	LINC00598	OPRPN	SLC28A2	TSPY26P
ADRM1	CELA2B	FER1L6-AS1	LINC00638	OR51B2	SLC5A2	TTLL10
AFAP1-AS1	CELA3B	FETUB	LINC01366	OR8D1	SLC7A6OS	TTTY11
AFG3L2	CFL2	FFAR2	LINC01565	PARP10	SMDT1	TYW5
AGAP3	CLDN18	FGA	LINC01620	PATE1	SMG9	UBE2G2
AGR2	CLEC4C	FGB	LIPC	PDYN	SMIM12	UBE2J2
AKR1D1	CLNK	FGL1	LMF2	PIGO	SMNDC1	UBL4B
ALG12	CLTA	FGR	LMNA	PIK3CD-AS1	SON	UCMA
ANXA9	CNBD2	FLJ31713	LMOD3	PIN1	SPATA16	UMOD
AP1G1	COG7	FLJ40288	LOC100127955	PLA2G2A	SPATA19	UPK3B
AP1M1	COPS7A	FOXP3	LOC101927051	PLEKHH3	SPATA8	UROS
AP1M2	CPSF7	FRMD8	LOC254028	PLG	SPHK2	USP21
AP3D1	CRKL	G6PC	LOC645261	POF1B	SSBP4	UTF1
APOB	CSF2	GAS2L2	LOC93622	POLL	SSMEM1	UTP18
APTX	CSF3R	GBA3	LRPAP1	POLR3H	SSU72	WFDC11
ARFRP1	CST11	GNLY	LRRC47	POM121L2	STAR	WFDC8
ARMC5	CST8	GOLPH3	LYL1	PPP1R35	SUGP1	VIL1
ARPP19	CTAGE1	GP2	MAS1L	PPP2R1A	SUN1	WNT2
ART5	CTBP1	GPKOW	MATN1	PPP6R3	SUN5	WNT8B
ASAH1	CUL3	GPSM3	MCM3AP	PRDM9	SYVN1	WNT9B
ASB1	CWH43	GPX5	MECOM	PRG3	TACSTD2	VPS18
ATE1	CXCR6	GSDMC	MED15	PRKACG	TAF2	VPS37D
ATF2	CXorf36	HIST1H4G	MEMO1	PRR15L	TAT	YTHDC1
ATP13A3	CYP1A1	HMGXB3	MKRN2	PRR30	TBC1D10B	YY1
ATP6V1F	CYP2B7P	HRG	MLF2	PRRX2	TBC1D22A-AS1	ZAP70
ATP8B5P	CYP4B1	HSD11B1L	MNT	PRSS54	TCF21	ZC3H7B
AURKAIP1	CYP4F2	HSPA12B	MOV10L1	PSMD1	TESMIN	ZCCHC13
BAAT	DDI1	HYAL4	MPG	PTCD2	TESPA1	ZNF142
BAP1	DDX17	HYI	MRGPRX2	PUDP	TFPT	ZNF251
BMX	DMP1	IAH1	MROH2B	R3HCC1L	THAP3	ZNF343
BPIFB1	DNAJB12	ICAM5	MS4A3	RAB7A	TIAL1	ZNF554
BTLA	DNAJC11	IGLL1	MTG2	RAF1	TIMM10B	ZNF576
C11orf16	DRD3	IL1RN	MTHFSD	REG3A	TM4SF5	ZNF584
C12orf42	DRG1	IL20	MTUS2-AS1	RETNLB	TMCO2	ZNF629
C1orf116	DVL2	IL36B	MUL1	RHBDD3	TMCO4	ZNF696
C20orf141	DYM	IL37	MYH4	RHOT2	TMCO5A	
C7orf77	EARS2	IRX4	NAA38	RNF114	TMEM198	
CALCR	ELF3	ITGAX	NDST1	RNH1	TMEM225	
CAPN9	ERP29	KCNK16	NDUFS7	RP9	TMEM40	
CASR	ERVH48-1	KRBA1	NFX1	RTP1	TMEM41A	
CCAR2	ETFBKMT	KRT20	NLRC4	RTP3	TMEM8A	

Supplementary figure 4



Supp. Fig. 4. Scatter plots of MDS results. **A)** Scatter plot of the first and third component of a MDS reduction of the batch-corrected dataset down to three dimensions. **B)** Scatter plot of the second and third component of the MDS results. **C)** Three dimensional scatter plot comparing all 3 dimensions of the MDS results.

Supplementary figure 5



Supp. Fig. 5. Biplots illustrating results of principal component analyses on all samples in the merged dataset before and after RUV-normalization, respectively. **A-B)** Scatter plots on the first and second principal components obtained for the raw data. **C-D)** Scatter plots on the first and second principal components obtained for the RUV-normalized data. In each of the two datasets, the 1200 most variable genes were used for the PCA. For each sample, the microarray platform is indicated via the shape of the respective datapoint, while colors reflect either the phenotye **(A, C)** or study affiliation **(B, D)**.

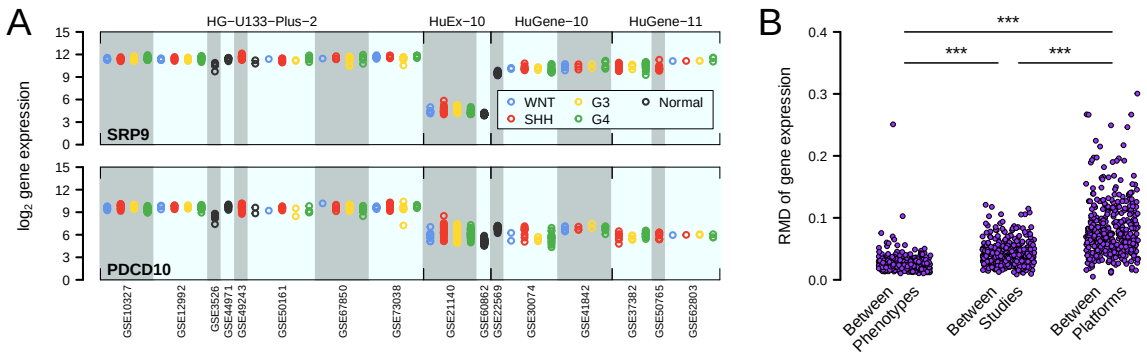
Supplementary table 5

Supp. Table 5: Datasets used for validation

	Number of samples						Used for*	
Dataset	**WNT**	**SHH**	**G3**	**G4**	**Normal**	**Total**	**Training**	**Testing**
GSE3526	0	0	0	0	9	9		X
GSE4036	0	0	0	0	14	14	X	
GSE10327	9	14	10	26	0	59		X
GSE12992	4	7	8	20	0	39		X
GSE37418	8	10	15	35	0	68	X	
GSE44971	0	0	0	0	9	9		X
GSE49243	0	58	0	0	0	58		X
GSE50161	1	9	2	7	2	21		X
GSE67850	1	5	9	7	0	22		X
GSE73038	10	16	9	10	0	45		X
GSE74195	1	1	7	11	5	25	X	
EMTAB292	0	3	3	8	5	19	X	
GSE21140	8	29	23	35	0	95		X
GSE25219	0	0	0	0	51	51	X	
GSE60862	0	0	0	0	130	130		X
GSE22569	0	0	0	0	22	22		X
GSE30074	2	9	3	16	0	30		X
GSE35974	0	0	0	0	44	44	X	
GSE41842	6	3	2	6	0	17		X
GSE37382	0	10	5	31	0	46		X
GSE50765	0	10	0	0	0	10		X
GSE62803	1	1	2	3	0	7		X
GSE85217	67	220	135	315	0	737	X	
Training	76	234	160	369	119	958	X	
Testing	42	171	73	161	172	619		X

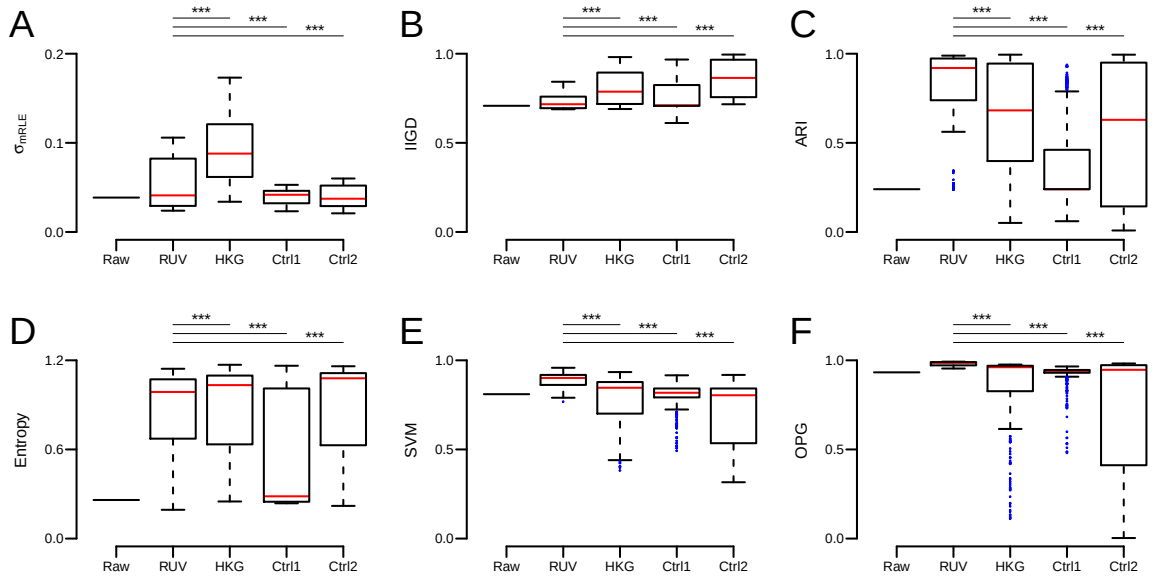
* The *training* dataset was only employed to extract the NCGs, which were then utilized to normalize the *testing* dataset.

Supplementary figure 6



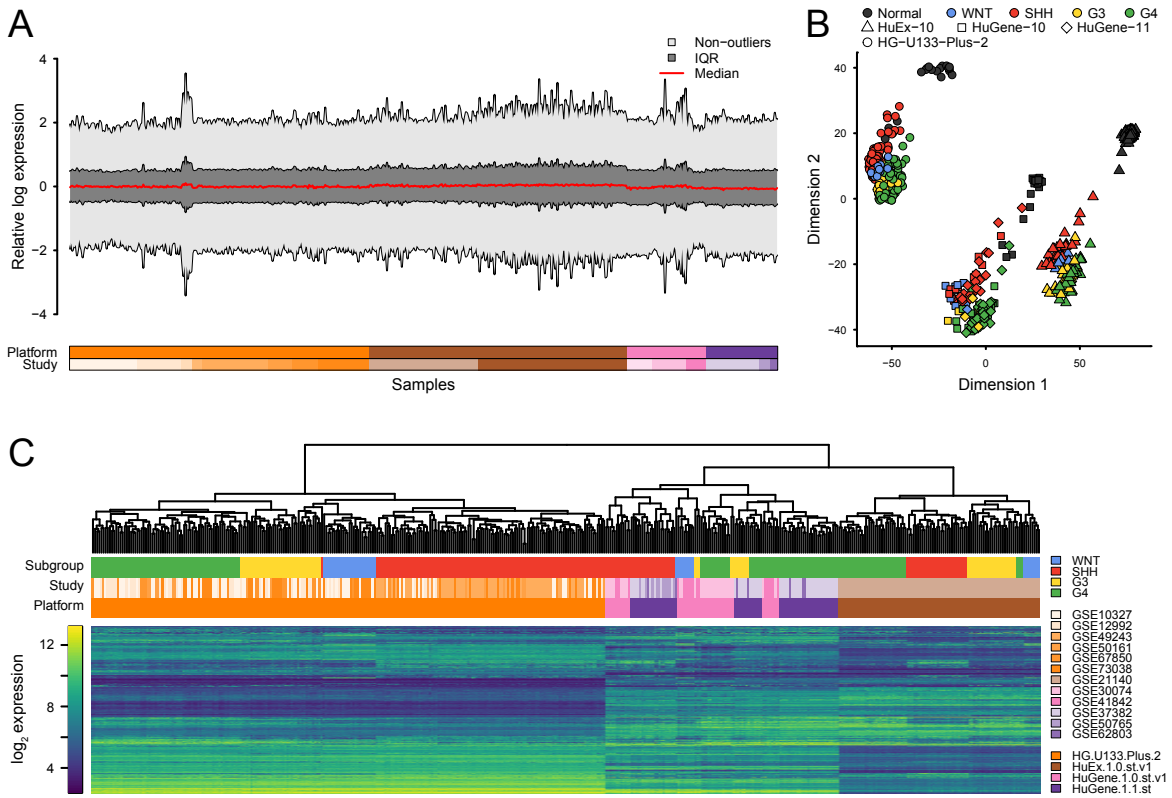
Supp. Fig. 6. Inspection of NCGs defined in a training dataset (7 studies, 958 samples) and illustrated in an independent test dataset (16 studies, 619 samples). **A)** Strip chart showing the gene expression across all 619 samples in the test data for the NCGs with the largest (top panel) and second largest (bottom panel) MAD score across all samples in the test dataset. **B)** Strip chart depicting the variation of expression values between phenotypes, between studies, and between platforms within the test dataset, as calculated for the empirically defined NCGs (one dot per gene). For each gene, the variation between phenotypes was calculated within each study as the RMD across phenotype means and the maximum RMD across studies was utilized as the final value. Similarly, the variation between studies was calculated on study means within each platform and the maximum across platforms was recorded. The variation between platforms was calculated as the RMD across platform mean expression values. ***: $p < 0.001$ (Wilcoxon signed-rank test).

Supplementary figure 7



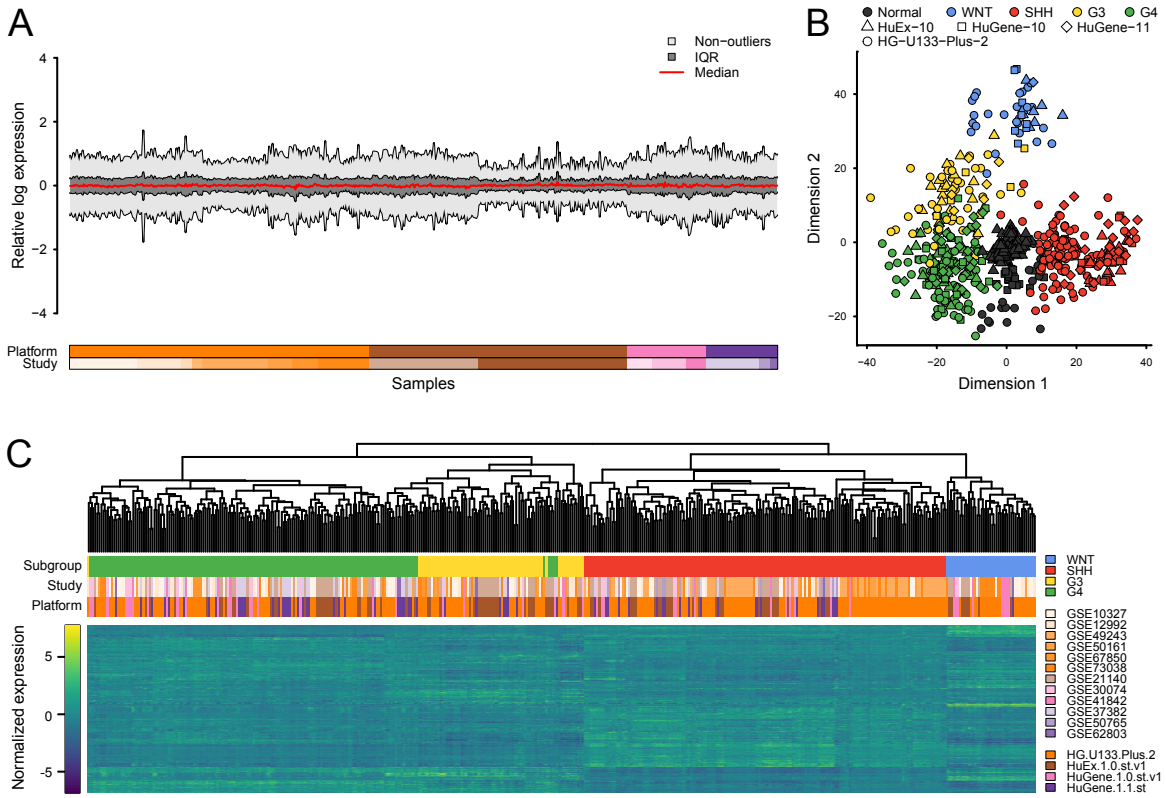
Supp. Fig. 7. Evaluation of batch-effect removal on the test dataset (16 studies, 619 samples) using NCGs identified from an independent training dataset (7 studies, 958 samples). **A-F)** Box plots depicting the distribution of σ_{mRLE} (**A**), IIGD (**B**), ARI (**C**), Entropy (**D**), SVM (**E**), and OPG (**F**) scores obtained from the raw expression data or after batch normalization over a range of regularization parameters and using either empirically defined NCGs (RUV) or three reference sets of control genes (HKG: 314 Housekeeping genes, Ctrl1: the 372 genes with the lowest RMD across all samples in the test dataset, Ctrl2: 372 randomly sampled genes). ***: $p < 0.001$ (Wilcoxon rank sum test).

Supplementary figure 8



Supp. Fig. 8. Visualization of batch effects in raw, merged dataset comprising 619 samples from 16 studies (GSE3526, GSE10327, GSE12992, GSE44971, GSE49243, GSE50161, GSE67850, GSE73038, GSE21140, GSE60862, GSE22569, GSE30074, GSE41842, GSE37382, GSE50765, GSE62803). **A)** Modified RLE plot showing the median, interquartile region (IQR), and non outlier ranges of each sample's RLE distribution. **B)** Scatter plot showing the result of a two-dimensional MDS analysis utilizing the top 1200 most variable genes. **C)** Hierarchical clustering of MB samples and the 1200 most variable genes.

Supplementary figure 9



Supp. Fig. 9. Visualization of batch effects in RUV-normalized dataset comprising 619 samples from 16 studies (GSE3526, GSE10327, GSE12992, GSE44971, GSE49243, GSE50161, GSE67850, GSE73038, GSE21140, GSE60862, GSE22569, GSE30074, GSE41842, GSE37382, GSE50765, GSE62803). The negative control genes, which were utilized in the RUV normalization, were empirically determined in an independent dataset comprising 958 samples from 7 studies (GSE4036, GSE37418, GSE74195, EMTAB292, GSE25219, GSE35974, GSE85217). **A)** Modified RLE plot showing the median, interquartile region (IQR), and non outlier ranges of each sample's RLE distribution. **B)** Results of a two-dimensional MDS analysis utilizing the top 1200 most variable genes. **C)** Hierarchical clustering of MB samples and the 1200 most variable genes.

References

Anderberg, M. R. (2014). Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks (Vol. 19). Academic press.

Barrett, T., *et al*. (2012). NCBI GEO: archive for functional genomics data sets—update. Nucleic acids research, 41(D1), D991-D995.

Brettschneider, J., *et al*. (2008). Quality assessment for short oligo-nucleotide microarray data. Technometrics, 50(3), 241-264.

Carvalho, B. S., and Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. Bioinformatics, 26(19), 2363-2367.

Cavalli, F. M., *et al*. (2017). Intertumoral Heterogeneity within Medulloblastoma Subgroups. Cancer Cell, 31(6), 737-754.

Chen, C., *et al*. (2013). Two gene co-expression modules differentiate psychotics and controls. Molecular psychiatry, 18(12), 1308-1314.

Cho, Y. J., *et al*. (2011). Integrative genomic analysis of medulloblastoma identifies a molecular subgroup that drives poor clinical outcome. Journal of Clinical Oncology, 29(11), 1424-1430.

Cox, M. A., and Cox, T. F. (2008). Multidimensional scaling. Handbook of data visualization, 315-347.

de Bont, J. M., *et al*. (2008). Differential expression and prognostic significance of SOX genes in pediatric medulloblastoma and ependymoma identified by microarray analysis. Neuro-oncology, 10(5), 648-660.

Eisenberg, E., and Levanon, E. Y. (2013). Human housekeeping genes, revisited. Trends in Genetics, 29(10), 569-574.

Eisenberg, E., and Levanon, E. Y. (2003). Human housekeeping genes are compact. Trends in Genetics, 19(7), 362-365.

Fattet, S., *et al*. (2009). Beta-catenin status in paediatric medulloblastomas: correlation of immunohistochemical expression with mutational status, genetic profiles, and clinical characteristics. The Journal of pathology, 218(1), 86-94.

Frigge, M., *et al*. (1989). Some implementations of the boxplot. The American Statistician, 43(1), 50-54.

Gagnon-Bartsch, J. A., and Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. Biostatistics, 13(3), 539-552.

Gandolfo, L. C., and Speed, T. P. (2017). RLE Plots: Visualising Unwanted Variation in High Dimensional Data. arXiv preprint arXiv:1704.03590.

Gokhale, A., *et al*. (2010). Distinctive microRNA signature of medulloblastomas associated with the WNT signaling pathway. Journal of cancer research and therapeutics, 6(4), 521.

Griesinger, A. M., *et al*. (2013). Characterization of distinct immunophenotypes across pediatric brain tumor types. The Journal of Immunology, 191(9), 4880-4888.

Ho, D. M. T., *et al*. (2015). Integrated genomics has identified a new AT/RT-like yet INI1-positive brain tumor subtype among primary pediatric embryonal tumors. BMC medical genomics, 8(1), 32.

Irizarry, R. A., *et al*. (2003). Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Accepted for publication in Biostatistics.

Jacob, L., *et al*. (2016). "Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed." Biostatistics 17.1: 16-28.

Kang, H. J., *et al*. (2011). Spatio-temporal transcriptome of the human brain. Nature, 478(7370), 483-489.

Kolesnikov, N., *et al*. (2014). ArrayExpress update—simplifying data submissions. Nucleic acids research, 43(D1), D1113-D1116.

Kool, M., *et al*. (2008). Integrated genomics identifies five medulloblastoma subtypes with distinct genetic profiles, pathway signatures and clinicopathological features. PloS one, 3(8), e3088.

Kool, M., *et al*. (2014). Genome sequencing of SHH medulloblastoma predicts genotype-related response to smoothened inhibition. Cancer cell, 25(3), 393-405.

Lambert, S. R., *et al*. (2013). Differential expression and methylation of brain developmental genes define location-specific subsets of pilocytic astrocytoma. Acta neuropathologica, 126(2), 291-301.

Liu, X., *et al*. (2012). Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques. Genome research, 22(4), 611-622.

Menghi, F., *et al*. (2011). Genome-wide analysis of alternative splicing in medulloblastoma identifies splicing patterns characteristic of normal cerebellar development. Cancer research, 71(6), 2045-2055.

Morrissy, A. S., *et al*. (2017). Spatial heterogeneity in medulloblastoma. Nature Genetics, 49(5), 780-788.

Müllner, D. (2013). fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. Journal of Statistical Software, 53(9), 1-18.

Northcott, P. A., *et al*. (2011). Medulloblastoma comprises four distinct molecular variants. Journal of Clinical Oncology, 29(11), 1408-1414.

Northcott, P. A., *et al*. (2012a). Subgroup-specific structural variation across 1,000 medulloblastoma genomes. Nature, 488(7409), 49-56.

Northcott, P. A., *et al*. (2012b). Rapid, reliable, and reproducible molecular sub-grouping of clinical medulloblastoma samples. Acta neuropathologica, 123(4), 615-626.

Park, A. K., *et al*. (2011). Prognostic classification of pediatric medulloblastoma based on chromosome 17p loss, expression of MYCC and MYCN, and Wnt pathway activation. Neuro-oncology, nor196.

Ramasamy, A., *et al*. (2014). Genetic variability in the regulation of gene expression in ten regions of the human brain. Nature neuroscience, 17(10), 1418-1428.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research, 43(7), e47-e47.

Robinson, G., *et al*. (2012). Novel mutations target distinct subgroups of medulloblastoma. Nature, 488(7409), 43-48.

Roth, R. B., *et al*. (2006). Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. neurogenetics, 7(2), 67-80.

Somel, M., *et al*. (2011). MicroRNA-driven developmental remodeling in the brain distinguishes humans from other primates. PLoS biology, 9(12), e1001214.

Sturm, D., *et al*. (2016). New brain tumor entities emerge from molecular classification of CNS-PNETs. Cell, 164(5), 1060-1072.

Thompson, M. C., *et al*. (2006). Genomics identifies medulloblastoma subgroups that are enriched for specific genetic alterations. Journal of Clinical Oncology, 24(12), 1924-1931.

Trabzuni, D., *et al*. (2013). Widespread sex differences in gene expression and splicing in the adult human brain. Nature communications, 4, 2771.

Vanner, R. J., *et al*. (2014). Quiescent Sox2+ cells drive hierarchical growth and relapse in sonic hedgehog subgroup medulloblastoma. Cancer cell, 26(1), 33-47.

Venables, W. N., and Ripley, B. D. (2002). Modern applied statistics with S. Statistics and computing.

Zhao, S., *et al*. (2014). Heatmap3: an improved heatmap package with more powerful and convenient features. BMC Bioinformatics, 15(10), P16.