Supplementary Method

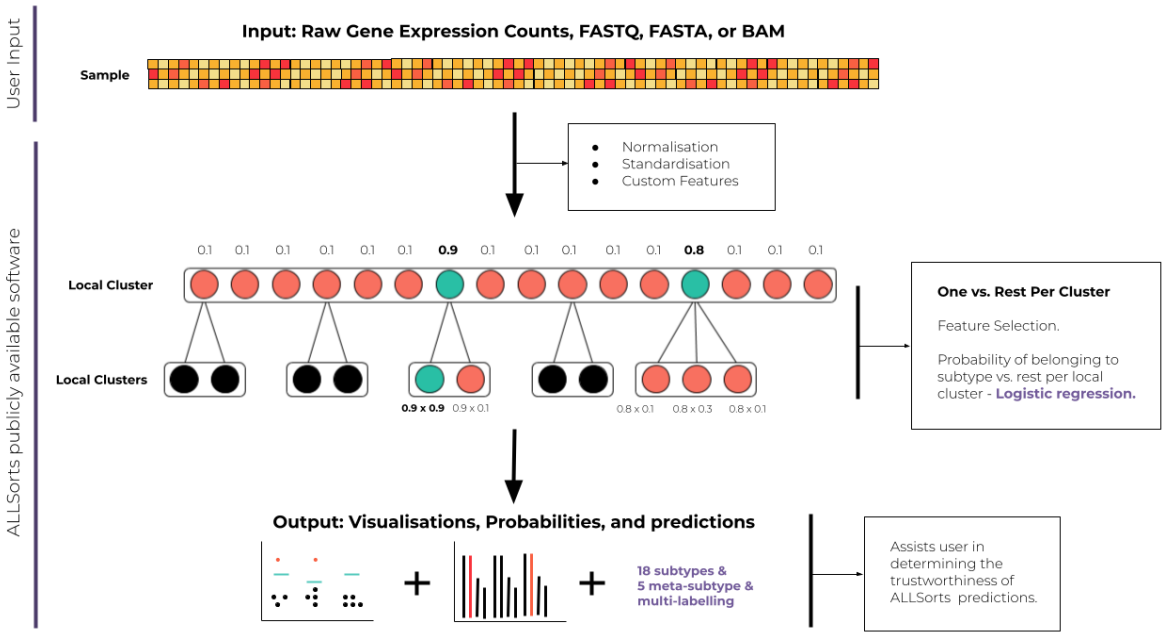
Overview

A combination of RNA-Seq, raw gene expression counts, and clinical information were obtained for 322 pediatric, 68 adult and 1988 mixed age B-ALL patients (Supplementary Table S1). After gene expression counts were obtained for each dataset, gene identifiers were converted to symbols for consistency.

Training of a machine learning classifier was broken down into four key steps: Preprocessing, Feature Creation, Standardisation, and Model Creation. These were encapsulated within a 10-fold cross validation (with replacement) and a nested grid search for optimal hyperparameter selection [14].

The preprocessing involved filtering for genes, expression transformation and normalisation. The Feature Creation step generates additional derived features from the gene expression. A standardisation step was then applied to the preprocessed counts and new features by calculating a z-score to create a consistent scale across features, aiding in a model's interpretability.

Finally the resulting counts matrix was input into a hierarchically organised set of logistic regression classifiers which were trained using the One Versus Rest method [18]. Applying the pre-trained model to new samples follows a similar sequence of preprocessing as described above (Supplementary Figure 1). Where any subtype's probability exceeds the threshold, ALLSorts classifies the sample accordingly.



Supplementary Figure 1 and visual abstract. Overview of the ALLSorts classification strategy for new input. Green circles are where the probability exceeds threshold. No probabilities are calculated for the black circles as classification terminates at their meta-subtypes. In this example, two meta-subtypes exceed their thresholds at the first level. However, only one nested subtype succeeds. This would result in a multi-label classification consisting of the deepest subtypes/meta-subtypes that exceeded their respective thresholds.

Data used in this study

Gene expression counts with associated clinical information were obtained/created for four B-ALL cohorts and a dilution series (Supplementary Table 1).

Cohort	No. Samples	Train (%)	Test (%)	Purpose	Source
St. Jude Children Hospital	1847	70	30	Train & Hold out	Gu et al. (2019)
Lund University	195	70	30	Train & Hold out	Lilljebjörn et al. (2016)
Royal Children's Hospital	127	0	100	Paediatric	Brown et al. (2020) & Children's Cancer

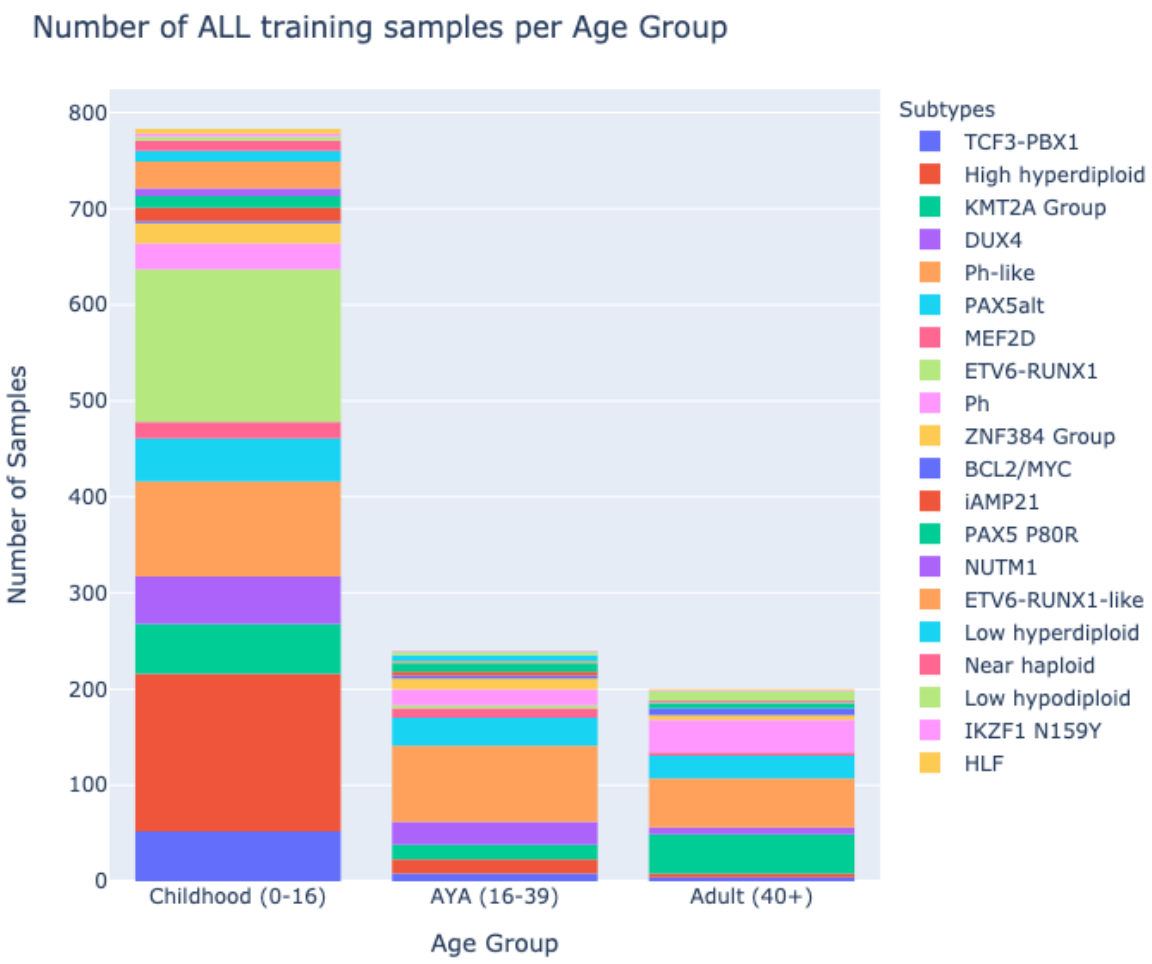
					Centre Tissue Bank at the Murdoch Children's Research Institute and The Royal Children's Hospital
Peter MacCallum Cancer Centre	68	0	100	Adult	Molecular Haematology Diagnostic Laboratory, Peter MacCallum Cancer Centre
Royal Children's Hospital (dilution)	16	0	100	Purity	Brown et al. (2020)
St. Jude Children Hospital and Lund (Multi-label)	117	0	100	Multiple Subtypes	Gu et al. (2019)

Supplementary Table S1. Datasets used for training and validating the ALLSorts method. Datasets that are split across training and test are stratified by subtype. The age breakdown across training datasets includes 783 childhood (under 16), 240 young adult (16-39), and 200 adult (40+) samples. Note: Samples previously subtyped as "Other" or multi-label were removed in each set apart from the Multi-labels samples, which were tested separately. A breakdown of the samples used in the final training set are listed in Supplementary Table 5.

Raw counts for 1988 samples from a recent St. Jude Children's Research Hospital study were available for public download through the St. Jude Cloud's visualisation community (Gu et al., 2019; McLeod et al., 2021). Raw sequencing reads from 195 samples were obtained from Lilljebjörn et al. (2016) (Lund - accession id: EGAD00001002112), 127 paediatric samples from the Children's Cancer Centre Tissue Bank at The Royal Children's Hospital (RCH), Melbourne, Australia (Brown et al., 2020) and 68 adult samples from the Molecular Haematology Laboratory, Peter MacCallum Cancer Centre, Melbourne, Australia (PM). The Lund, RCH, and PM datasets were mapped to the human reference genome version hg19 using STAR 2.7.3a_2020-01-23 in 2 pass mode with quantMode set to Gene, with otherwise default options. Gene expression counts for the RCH, Lund, and PM dataset were provided as output from STAR, using the GRCh37.87 annotation obtained from the ensembl FTP

(ftp://ftp.ensembl.org/pub/grch37/current/gtf/homo_sapiens/Homo_sapiens.GRCh37.87.chr.gtf.gz). After gene expression counts were obtained for each dataset, Ensembl gene identifiers were converted to gene symbols. Identifiers with multiple copies had their counts combined. After this, genes other than those with biotypes of protein coding or recognised as Immunoglobulin (Ig) variable chain or T-cell receptor (TcR) genes were discarded. Resulting in a final 20652 of the 57773 original genes being used in the training data.

The St. Jude samples were labelled according to the 23 subtypes outlined in Gu et al. (Gu et al., 2019). However, the Lund samples were assigned labels according to the following classes: *High hyperdiploidy, ETV6-RUNX1*, *Ph-like*, *MLL*, *TCF3-PBX1*, *DUX4-rearranged*, *BCR-ABL1*, *dic(9;20)*, *ETV6-RUNX1-like*, *B-other with fusion*, *B-other*, *without fusion*, *Hypodiploid*, *Near Tetraploid*, and *iAMP21*. As karyotype data was also available, the aneuploid samples were distributed across *High hyperdiploid* (58) and *Low hypodiploid* (1) accordingly. *MLL*, *DUX4-rearranged,* and *BCR-ABL1* were renamed *KMT2A*, *DUX4*, and *Ph*, respectively, to reflect the St. Jude naming conventions where fusion information was available. Samples labelled *dic(9;20)* and *Other* were removed and explored using the trained classifier. Samples that were not labelled as having multiple subtypes, but showed signs according to associated clinical information (i.e. *iAMP21* being mentioned in karyotype but not mentioned as a subtype) were discarded from the training data. In addition, St. Jude labelled aneuploid samples that did not have concordance with the karyotype were also discarded. In all, 168 samples were marked for exclusion from the training data across both the St. Jude and Lund cohorts. Finally, training and test sets were segmented according to Supplementary Table S1 and the training samples listed in Supplementary Table 5. The training data had a range in ages from pediatric to adult (Supplementary Figure 2).



Supplementary Figure 2. Distribution of subtypes in the Lund/St Jude training set segmented by age. The training set consists of 783 childhood (under 16), 240 young adult (16-39), and 200 adult (40+) samples

Pre-Processing

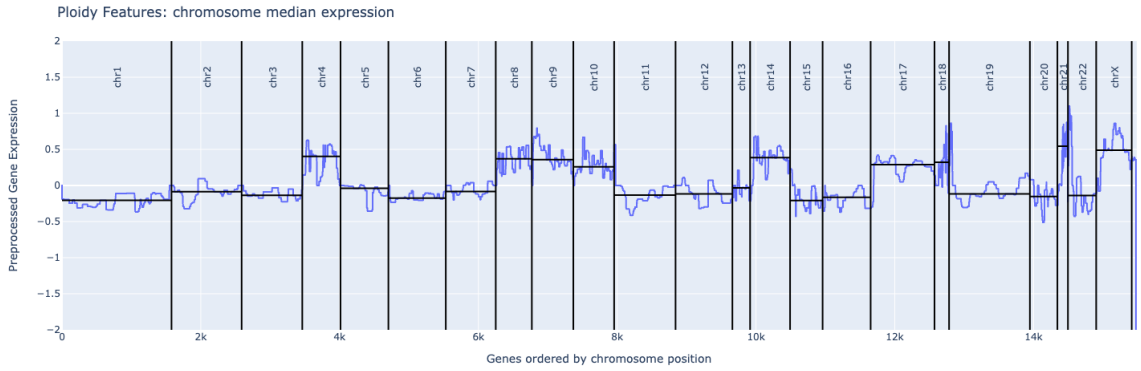
The first step of pre-processing is to filter lowly expressed genes as they contribute little information about the biology fundamental to the classification problem. To achieve this, the method outlined in Chen et al. (2016) was adopted. That is, the training set was transformed into counts per million (CPM) prior to filtering. This is to ensure that genes that are lowly expressed due to smaller library sizes are not naively filtered. Genes are retained if there are at least 10/L (where L is the smallest library size) in at least as many samples as the subtype with the lowest membership. Once lowly expressed genes have been identified they are stored for later removal in new samples input into the classifier. The training data is then normalised using a Python implementation of the Trimmed Mean of M-values (TMM) normalization method (Robinson & Oshlack, 2010). This method is preferable over methods such as fragments per kilobase million (FPKM) as it accounts for library composition and is

therefore suitable for inter-sample comparisons (Conesa et al., 2016). The reference used to calculate TMM scaling factors is then stored for later use when ALLsorts is applied to a new dataset. The filtered raw counts are then transformed to log2 counts per million (CPM) to scale samples by library size. Further scaling is then applied using the factors calculated from the Trimmed Mean of M-values (TMM) method.

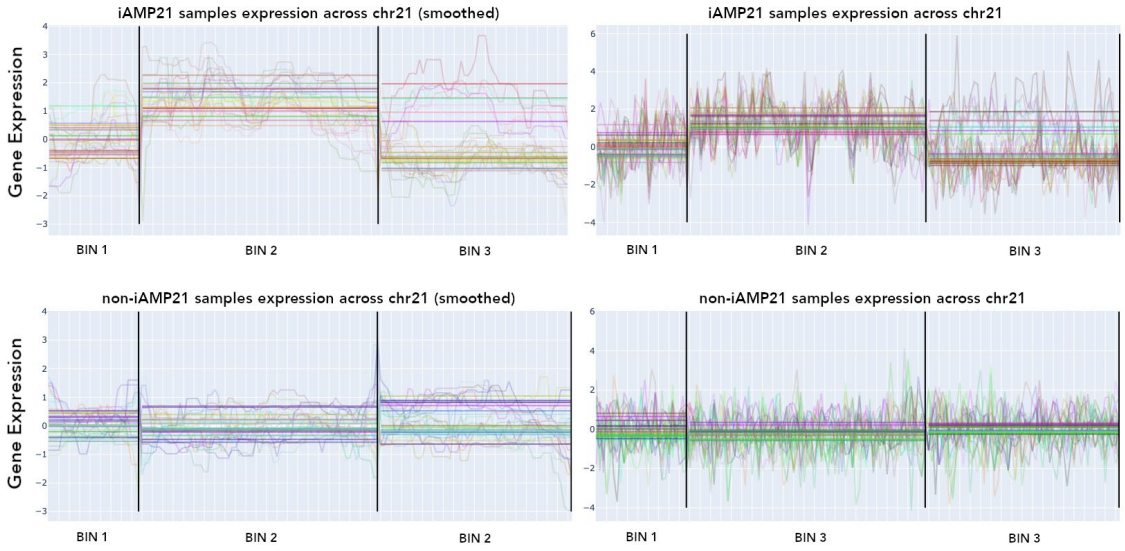
Feature Creation

ALLSorts uses four sets of manually crafted features to represent the biology of B-ALL represented in the literature. The first is a set of features that represent known fusion genes that are highly relevant to some subtypes: *ETV6-RUNX1*, *BCR-ABL1*, *TCF3-PBX1*, and *TCF3-HLF*. The resulting features are simply the log difference between the two partner genes. This feature is important to include as it encapsulates the relationship within a single feature. The second set of features represent the relative expression of each chromosome per sample. This was calculated in a similar fashion to existing solutions to calculating ploidy from RNA-seq (Gu et al., 2019; Serin Harmanci, Harmanci, & Zhou, 2020). The genes in the training set are first scaled by median absolute deviation. Median iterative filtering is then applied across genes in each chromosome, smoothing the signal across the chromosome. A final median is then selected per chromosome and a set of 27 features are created from this (chr 1-22, X, Y, median across the chromosomes, and two representing how many are highly expressed and how many are low). A visualisation of these features can be seen in Supplementary Figure 3. Thirdly, an iAMP21_ratio feature is also created based on the knowledge of its distinct signal across chromosome 21 (Tsuchiya et al., 2017). For each sample chromosome 21 is divided into four bins and is scaled by median absolute deviation (Supplementary Figure 4). Fourthly, in an attempt to capture non-linear relationships associated with a subtype, a feature that represents the euclidean distance towards a subtype's centroid in a nonlinear projection is calculated for each sample. Concretely, for each local cluster of subtypes (Figure 1) a random forest machine learning classifier is trained in a one-versus-rest fashion. From this, the top 20 features as measured by feature importance are attributed to each of the subtypes. A Kernel PCA projection is then created for each subtype using these genes, hopefully delineating between the subtype and the rest. The centroid is then calculated across each of the true subtype samples in this projection and the euclidean distance to this point is then calculated per sample. The final feature is constructed from the difference between the bin at the highest and lowest points. Finally, a single B-ALL feature is created as the log sum of CD19, CD34, CD22, DNTT, and CD79A. These genes are known markers for B-ALL in the literature (Chiaretti, Zini, & Bassan, 2014; Cobaleda & Sánchez-García, 2009). The purpose of this feature is not to be necessarily

used in the following classification stage, but rather as a filter to remove false positives when classifying healthy samples or other cancers.



Supplementary Figure 3. St. Jude Children's Research Hospital sample with ***'56,XX,+X,+4,+6,+8,+10,+14,+17,+18,+21,+21[18]/46,XX[2]'*** karyotype. The preprocessed gene expression values are first ordered according to their genomic position. In each sample, per chromosome, the ordered gene expression values undergo median iterative smoothing. The median of the smoothed expression values is then calculated per chromosome. These medians (chr1-22, X, Y) are then added to the list of features presented to the model prior to feature selection.



Supplementary Figure 4. iAMP21 as depicted by the ordered genes in chromosome 21 is divided over three bins. Y-axis represents the difference in expression for the sample and the median of the cohort. This difference is further divided by the median absolute deviation value for each gene. Samples with confirmed iAMP21 (top) vs non-iAMP21 (bottom). Left plots use iterative median smoothing over genes, right does not. Each coloured line is a sample. The horizontal lines represent the median for that sample over each bin. A clear motif across bins is apparent between those samples with iAMP21 and those without. Each bin is included as a feature within ALLSorts, along with the log difference between bins 2 and 3 - iAMP21 Ratio.

Feature Standardisation

Features are standardised through transformation into a z-score, by subtracting the mean and dividing by the standard deviation in a feature-wise manner. The end result of this is a set of feature distributions, each with a mean of 0 and a standard deviation of 1. Standardisation was shown to perform better when utilised within a hierarchical architecture (Supplementary Table 6). In addition, standardisation by z-score allows the coefficients within a linear model to be equally considered in terms of importance, that is, higher coefficients are relative to the corresponding features importance.

Hierarchical Classification

Given the large set of input genes (~20000) relative to the number of training samples (~2000), a subset of genes needed to be selected to reduce dimensionality and the risk of overfitting of a trained model. During training, three feature selection methods were competed to find the best performing, each would then be followed by a further selection using L1 Regularisation embedded within liblinear's implementation of the logistic regression model, the core classification algorithm used within ALLSorts. L1 regularisation is used to encourage sparsity within the model by penalising uninformative genes by trending their coefficients to zero (Pedregosa et al., 2011). This can be considered a multivariate feature selection method and is tuned with the parameter C - lower values will result in more aggressive regularisation. The first competing method is a univariate statistical test, mutual information, that is calculated per subtype (Pedregosa et al., 2011). This list is then rank-ordered and subjected to a two standard deviation from the mean cutoff with the intention of completely omitting a set of uninformative genes, whilst retaining genes with high value. The other method used a L1 regularised logistic regression classifier to pre-train the model and select the genes useful for classification, passing them on to the final classifier. The final method has no feature selection prior to the classifier, depending only on the single regularisation method embedded in liblinear. The best combination of hyperparameter values was determined by minimizing log loss over 3 fold cross validation (Supplementary Table 6). Although more folds would have been ideal, there were too few samples in many of the subtypes. Balanced accuracy was used as the metric of success, as opposed to purely accuracy, as this accounts for the imbalance of samples allocated across subtypes. Balanced accuracy is calculated according to:

$$(1)\ balanced\ accuracy\ = \frac{1}{2}\left(\frac{TP}{FP + TP} + \frac{TP}{FP + TP}\right)$$

The winning feature selection method was the single L1 regularised logistic regression model, without prior feature selection. The results of hyperparameter selection can be seen in Supplementary Table 6.

An important aspect of classification is assigning probabilities to a discrete class. This is typically performed by setting a 50% threshold per subtype. Though this is appropriate for binary classification problems it is not necessarily appropriate for multi-class classification problems. In these cases, the probability is distributed amongst multiple classes and may not exceed 50% in any. Many algorithms therefore surpass this problem by choosing the maximum probability from the result. However, this is not appropriate for cases where there is a chance that a sample belongs to a new class entirely. To attempt to resolve this problem, ALLSorts determines the probability thresholds using the cross-validation results from the training data. Thresholds are determined for each subtype, using one of two methods. In the first case, if the positive and negative samples for that subtype separate cleanly on probability, i.e. the highest probability for a negative sample is lower than the lowest probability of a positive sample, the midpoint between these two points is chosen. However, if this is not the case and the positive and negative samples overlap in probability, a threshold that maximises the F1 score is chosen. The threshold for any child subtype is weighted by the parent threshold through multiplication.

Prediction

To predict B-ALL subtype, raw gene expression counts are input into the trained ALLSorts algorithm, this data is pre-processed and has features created using values acquired through the training set. These processed counts are then filtered by the genes selected during the training of the algorithm. Finally, they are input into the hierarchical classifier and their probabilities of belonging to each subtype is calculated. Furthermore, children probabilities are the multiplication between the child probability and its parent. After this, ALLSorts has various visualisation and prediction methods that help users in understanding their results.

ALLSorts cross validation

10-Fold Cross Validation was performed during training to score the method and select optimal thresholds, whilst 3-fold cross validation was implemented in an inner loop to select the hyperparameters of the model. The scores averaged across the outer loop were then used as a benchmark for comparison to other datasets to consider overfitting.

Each of the following stages was performed within each fold to prevent the leakage of data between the resulting training and validation splits. The trained classifier was also

applied to held-out test sets, which were split from the cohorts prior to training (Supplementary Table S1).

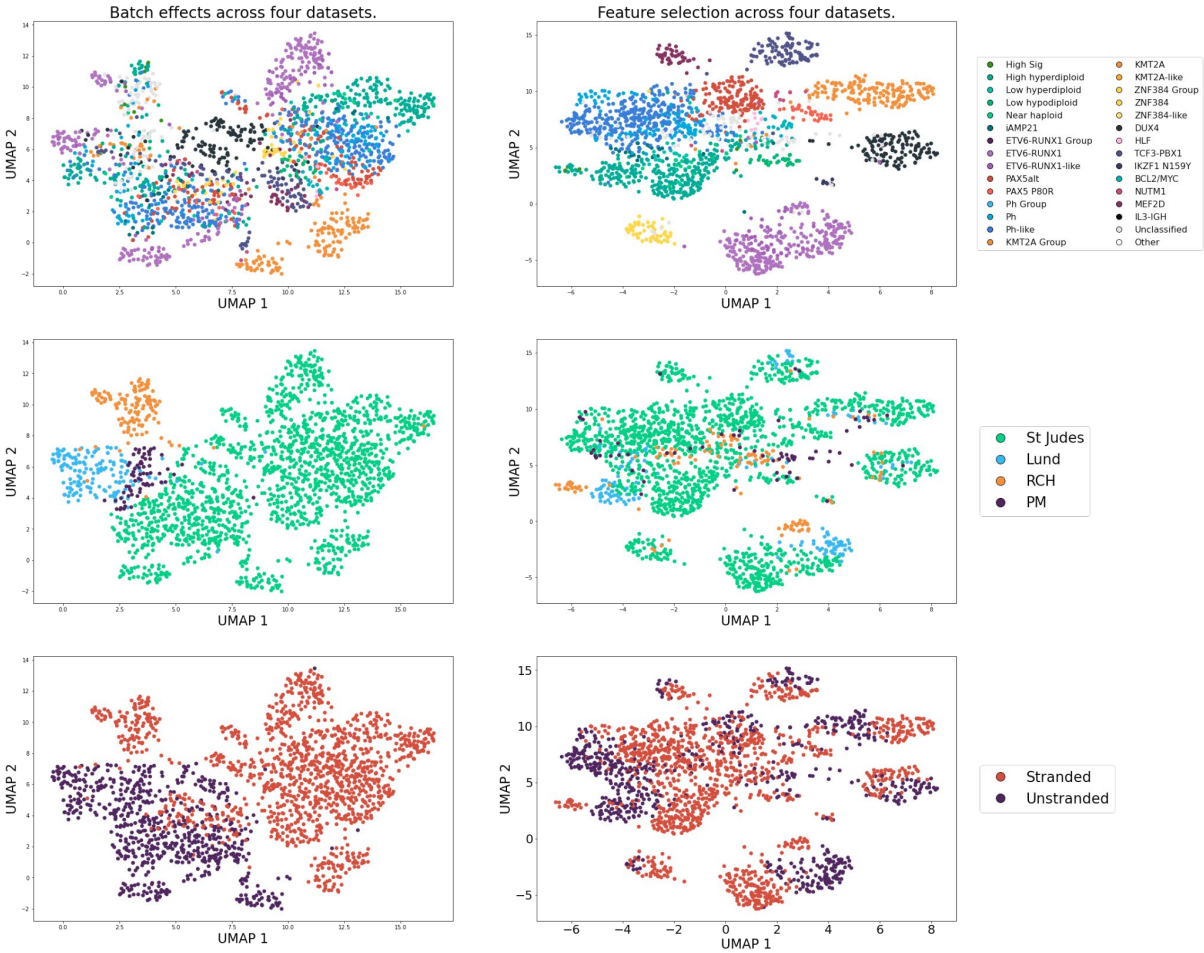
Four standard statistical metrics were used in the evaluation of the classifier: Accuracy, Precision, Recall, and F1 Score. Accuracy is the proportion of samples that were predicted correctly. Precision and recall are complementary, measuring the proportion of true positives and false negatives, respectively. Finally, the F1 score reflects the balance between precision and recall. These are calculated for each subtype and then aggregated by weighting the proportion of samples in each subtype.

Dataset	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**
		Weighted average		
Cross-Validation (avg over 10 fold)	90	96	91	93
St. Jude's & Lund hold out	92	97	92	94

Supplementary Table S2. ALLSorts performance in 10 fold cross validation determined during training and performance of held-out test sets from the St. Jude's and Lund cohorts.

Having both the cross-validation and held-out test set results from the same cohorts used for training allows us to determine whether the model is underfit or overfit. The held-out test set has a higher precision, recall, F1 score, and a slightly lower accuracy than the cross-validation result (Supplementary Table 6).

Typically, ALLSorts will be applied to new samples which include technical differences in the acquisition and processing of the samples compared to the training data (Supplementary Figure 5). To test whether ALLSorts is robust to such effects we applied it to paediatric and adult B-ALL cohorts from the Royal Children's Hospital (RCH) and Peter MacCallum Cancer Centre (PM), respectively. Each cohort had different sequencing and library preparation protocols making them an effective representation of a typical input with batch effects. The batch effects were found to be less influential when using the features selected with the training data (Supplementary Figure 5)



Supplementary Figure 5. Each point represents a sample and their gene expression values projected via UMAP. Plots in the left column are constructed from all genes, plots in the right column are filtered by the gene features that are selected by the ALLSorts classifier. The top row is coloured by subtype, middle row by source, bottom row by RNA library preparation of total RNA (stranded) or mRNA (unstranded). The RNA preparation is a large source of variation across the cohort. We found the gene features we selected are robust to batch effects resulting in the subtype becoming the largest source of variation. While batch effects are still present, they are less influential.

The results of the ALLsorts on the validation cohorts can be broken into four categories: match with ground truth, new classification into a subtype, reclassification to another subtype(s), and subtype to unclassified. Assuming the matched samples are correct (109 samples or 56%), only samples described by the latter three categories required further exploration.

Of the 74 samples that were previously Unclassified, 61 (82%) were newly classified into one of the 18 subtypes or five meta-subtypes offered by ALLSorts. Of these, 46 were evaluated to be plausible, two were incorrect, one had a blast % under 10%, and no definitive evidence could be found for 12. Reclassification to a new subtype accounted for 10 samples. Of these, eight matched the same meta-subtype as the previous label. One sample was incorrectly called High Sig instead of iAMP21. However, this sample had a tumour purity

of only 13% which could account for this misclassification. Finally, one contained a novel *ETV6* fusion but was predicted as being DUX4. The reason for this is currently unknown. The most important misclassifications to explore were the 15 samples (7.7%) previously labelled as a distinct subtype which ALLSorts assigned as Unclassified. Six of these samples had tumour purities of less than 10%, which may account for misclassifications in these cases. Of the remaining nine, three were previously labelled as KMT2A rearranged of which each had cytogenetic evidence of the relevant fusion genes. However, these samples exhibited low expression for genes such as MEIS1, a typical target of KMT2A fusions which ALLSorts weights highly in KMT2A Group classification. Four High Sig samples with a tumour purity above 10% did not reclassify according to their ground truth. However, two had high probabilities of being ETV6-RUNX1 Group and had an associated ETV6-BCL2L14 fusion discovered through Arriba. As these two samples also had relatively high probabilities for High Sig (over 39%), it is possible that these are multi-subtype samples. The remaining newly Unclassified samples were labeled according to cytogenetics only or had a lower tumour purity (~16%).

A full list of samples that had unexpected classifications with potential causative variants found is provided (Supplementary Table 7). Of these 86 samples, 62% had a plausible explanation that the ALLSorts classification was correct at least to the meta-subtype level, 10% were incorrect, 19% remained ambiguous in terms of evidence supporting or dismissing plausibility of the call, and 9% were defined as having tumour purity too low for concrete classification (less than 10%).

In summary, the overall accuracy of ALLsorts on the combined RCH and PM validation cohort was between 84% and 92% depending on if the ambiguous samples were considered incorrect by ALLsorts or correct (Supplementary Table 7).

Ambiguous Samples	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 (%)**
		Weighted average		
Marked as incorrect	84	93	95	93
Marked as correct	92	98	97	97

Supplementary Table S3. ALLSorts performance in the RCH and PM cohorts once orthogonal evidence gave plausibility to the calls. Two sets of summary statistics are presented, representing where ambiguous samples have been marked as correct or as incorrect - indicating the boundaries of the classifiers performance on these datasets.

ALLSorts classifies samples with multiple subtypes

Without specifically training ALLSorts to recognise samples exhibiting multiple subtypes, this cohort was used to investigate the capacity for multi-label classification.

ALLSorts prediction on samples labeled with multiple subtypes	%	No. Samples
Both subtypes called	19.65	23
One subtype and one meta-subtype	5.98	7
Two meta-subtypes	0	0
One subtype called	60.68	71
One meta-subtype	4.27	5
Neither subtype called	9.4	11

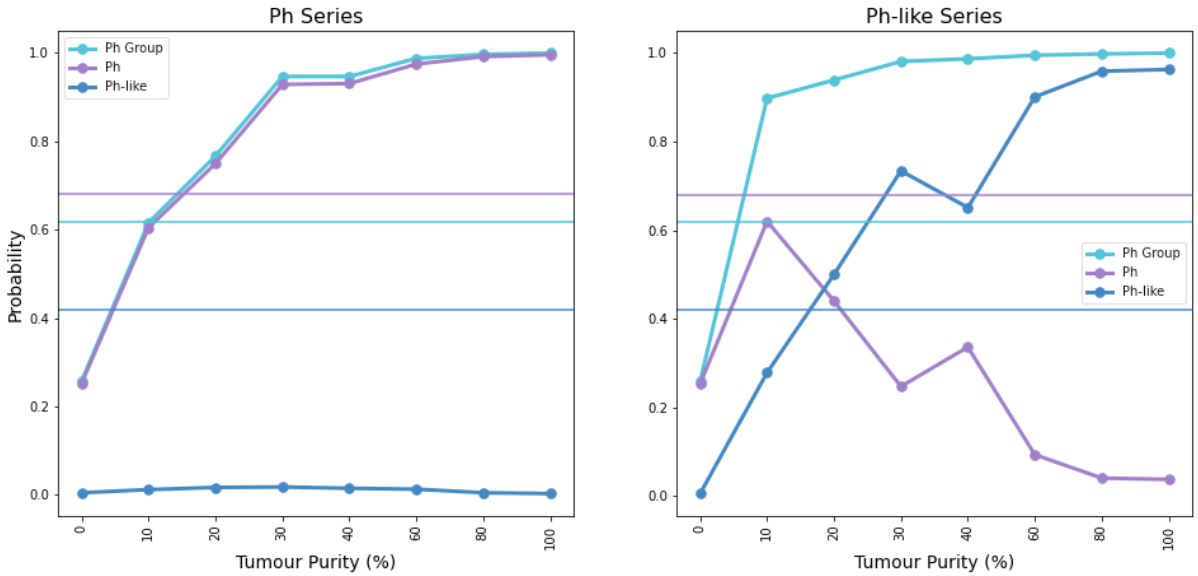
Supplementary Table S4. Breakdown of multi-label predictions.

We found the probability of getting at least a single subtype correct is 86.31% and 90.5% if including meta-subtypes (Supplementary Table S4). Given this is similar accuracy to the single subtype benchmarks, multi-label classification can be added without reducing single subtype classification accuracy. However, we only predicted both subtypes 26% of the time. Interestingly, within the held-out test set thought to be composed of samples with only a single subtype, nine samples were predicted as having two. Similarly, the PM and RCH combined cohort had six samples (3%) classified with two subtypes. Of these six samples five were found to have evidence pointing to the accuracy of both calls from fusion calling and karyotyping (Supplementary Table 7).

ALLSorts is robust with tumour purity above 20%

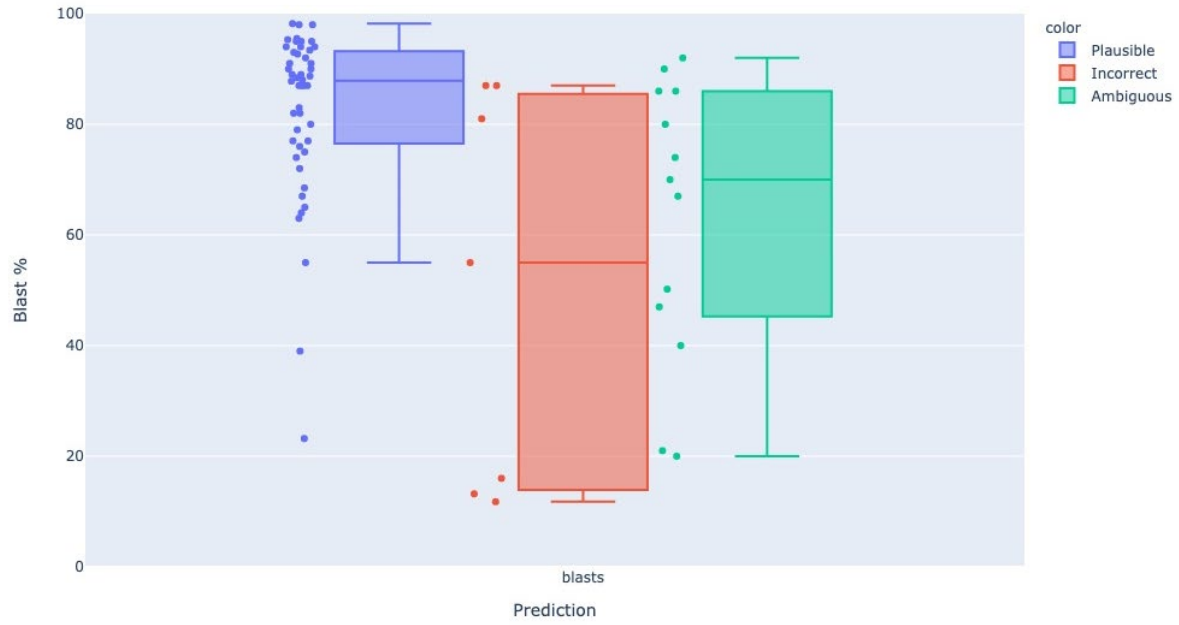
To test the effect of tumour purity on classification performance, two dilution series were used. These were created from a mixture of RNA extracted from the lymphoblastoid cell line from NA12878 with either a Ph (*BCR-ABL1* fusion) tumour sample or a Ph-like (*PAX5-JAK2* fusion) tumour sample (Brown et al., 2020). The tumour RNA was mixed in the proportions: 0%, 10%, 20%, 30%, 40%, 60%, 80% and 100% and each mixture was classified with ALLSorts. As expected, purer tumour samples corresponded to higher probabilities (Figure 4). For the Ph+ dilution series ALLsorts was able to classify the correct subtype down to a purity of 10%. The Ph-like dilution series classified correctly down to 20% purity and at 10% deferred to being classified as Ph Group. Though a higher tumour purity will result in more confident classifications, these results suggest ALLSorts is robust to

tumour proportions of above 20% in these subtypes. However, this would need to be tested for each subtypes in order for a claim of general robustness to tumour purity could be made.



Supplementary Figure 6. Probabilities from ALLsorts of the Ph Group meta subtype and Ph and Ph-like subtypes in response to tumour dilution. Subtype probability thresholds are indicated by the horizontal lines.

To visualise this with real patient data, samples with available blast percentage within both the PM and RCH cohorts were plotted against the plausible, ambiguous, and incorrect categories discussed in Results. Seven samples had a blast percentage of less than 30%. The single plausible sample was called a PAX5alt. Three ambiguous samples were classified as Ph-like, of which one was also called High hyperdiploid. The three incorrect classifications were: a DUX4 (IGH-DUX4 fusion) that was unclassified, a sample with normal cytogenetics called High hyperdiploid, and an iAMP21 misclassified as High hyperdiploid.



Supplementary Figure 7. Blast % versus Plausible/Incorrect/Ambiguous categorisation allocated during investigation into known driver events.