Exploring cancer transcriptomes for novel driver events

Abstract

High-throughput sequencing has revealed the transcriptome as an effective representation of the cellular state and useful for investigating events that may drive oncogenesis. Identification of these events has had much benefit historically towards the pursuit of precision oncology and diagnostics. However, further research is required to identify the extent of molecular heterogeneity in cancer and develop methods to use this knowledge clinically. This bioinformatics PhD applies machine learning to classify B-Cell Acute Lymphoblastic Leukemia (B-ALL) samples into a set of 19 subtypes, is investigating the molecular landscape of T-ALL, and applies a novel visualisation to rare splice variants.

Introduction

Overview

Fundamental to every human cell is Deoxyribonucleic Acid (DNA), a molecule that encodes the information necessary to maintain healthy function. Though the fidelity of cellular division is considered high, mutations in DNA are still thought to occur at a rate of one per billion nucleotides [1]. These mutations can then proliferate within subpopulations of cells, which in turn, may experience additional aberration. The consequences of accumulating mutations range from relatively benign to malignant [1–5]. Though mutation may enable oncogenes, the cell has multiple repair and self destruct mechanisms to prevent tumorigenesis. However, mutations can also occur within tumour suppressing genes which assist in disabling these mechanisms. This scenario enables mutations to proceed unchecked and the cell to ultimately exhibit uncontrolled proliferation and other hallmarks of cancer.

Investigating mutations in tumours has led to many initiatives to consider their implication. In oncology, sets of mutations that are causally linked to the progression of disease are termed *drivers* [6,7]. Those that exist within the disease state, but do not contribute to its persistence, are termed *passengers* [6,7]. Distinguishing between pathogenic driver and benign passenger events is an active focus in cancer research, as driver mutations have historically been successful targets for pharmacological intervention [8,9]. The combined sets of both types of events can infer the existence of distinct disease states. And identification of these can potentially be used clinically to guide treatment [10]. Sequencing technologies have rapidly transformed much of genetics into a quantitative science, with bioinformatics pursuing research and clinical application in genomics, transcriptomics, and proteomics. With respect to mutations, transcriptomics has proven to be of particular use. Microarray Profiling, RNA Sequencing (RNA-Seq), and Single-Cell RNA Sequencing (scRNA-Seq), have granted insight into how genes across samples are functionally affected by mutation [11]. In addition, pathogenic states that are described by distinct and recurrent patterns of gene expression provide an opportunity to learn statistical models that can identify them [10]. As such, machine learning, a suite of algorithms that learn through exposure to data, has been widely used within the literature for both exploration and clinical application.

Cancer Transcriptomes

Identifying the molecular events that drive cancer remains a key challenge within oncology. This information is used to both guide treatment and contribute to the understanding of the initiation and progression of cancer. These mutations can take many forms. From single nucleotide changes to large structural variation, such as: insertions, deletions, inversions, translocations, duplications. Entire chromosomes can even break apart and fuse to one another, potentially resulting in the fusion of genes. Therefore, it is clear that the mutational landscape is as complex as it is hard to detect. Next generation sequencing has somewhat relieved this latter concern, with various technologies now available that can provide data in various forms that is representative of an individual's genomic sequence. The challenge is how to interpret this information. Recent cross-institutional efforts to elucidate such mechanisms from exome sequencing, sourced from the Cancer Genome Atlas (TCGA), reveal the enormity of such a task and its importance. In this case, 299 driver genes were proposed across 9,097 samples through a consensus between 26 various bioinformatics tools and expert curation, 60-85% of which were experimentally validated [12]. Other efforts within the same cohort, but through different sequencing technologies, have highlighted the significance of known cancer hallmark pathways, noting that 89% of the 9,125 tumours studied harbour an alteration within these networks, 57% targetable with current treatment options [13]. Such discoveries have seen large clinical utility in treating cancer effectively. For example, patients identified to have the BCR-ABL1 fusion gene can experience an improved outcome with small molecule inhibitors of ABL1 [14].

Oncogenic mechanisms can be investigated within the abundance and makeup of a cell's transcriptome - the complete set of RNA within the cell [11]. As DNA is transcribed into RNA, mutations that occur within the cancer genome can reverberate through to the transcriptome, potentially resulting in a complex expression profile. This information is critically important when interpreting the state of an oncogenic sample and may only be visible through RNA sequencing.

One of the most transformative applications of transcriptomics in cancer has been the realisation that there exists recurrent patterns in the gene expression of tumour samples. This has granted the ability to delineate between healthy and malignant samples and between tumours that originate from the same cell type but with distinct causal mechanisms [15–20]. Transcriptomics has also led to methods that identify perturbation within known cancer hallmarks pathways [15]. And has allowed further refinement in the understanding that distinct genomic lesions can lead to an apparent convergence of phenotype [14,21]. The transcriptome is also an excellent medium for investigating specific mutational events. Not only does it demonstrate the relative quantities of genes, exons, and transcript isoforms, but also the transcribed consequences of variation - such as fusion genes [22]. Indeed, fusion events are known to be a significant class of drivers within many cancer types. They are particularly prevalent in Acute Lymphoblastic Leukemia (ALL), where studies have claimed that around 65% of samples contain a gene fusion [18,23]. RNA-Seq has enabled a comprehensive discovery and succinct visualisation of fusion genes which can then be subject to downstream analysis for interpretation [22,24–26].

Despite ongoing success research in cancer bioinformatics is a complex undertaking. Large and meticulous data curation efforts like the TCGA are not without systemic issues, with analyses revealing technical artifacts that may hamper efforts to analyse such efforts [27]. Coupled with this, the methods that many rely upon to remove such technical effects may also impose bias of their own [28]. Technical aspects aside, the biological theories of how tumorigenesis arises and is maintained is also diverse across the literature [5,29,30]. This is naturally apparent when observing the many methods being developed to make new discoveries [12,31,32].

Given the heterogeneity between and within cancer types, the relatively small sample sizes available and the variation of methods/environments used when obtaining them, the various standards for describing this data, and the difficulty in predicting which information will lead to successful outcomes if pursued, improving cancer outcomes seems daunting. Nonetheless, given the potential reward for discovery, it is relentlessly pursued.

Machine Learning in Cancer Transcriptomes

Machine learning describes a suite of algorithms that are capable of learning statistical models. Through an increasing exposure to data these algorithms refine models and improve their performance towards a task as measured by a pre-defined metric. Although the approach and complexity of these algorithms differ greatly they can be generally segmented into four larger classes: Supervised, Unsupervised, Self-Supervised, and Reinforcement Learning [33]. Each of these classes have a distinct application.

Relevant to this project, supervised learning methods generally receive two inputs: a set of samples that are each described by a set of consistent features and a list of labels which specify the expected output of each sample. The algorithm then has the objective of learning a model that can predict which label an unseen sample belongs to based on the values of its features. This is better known as classification and is one of the most widely used applications of machine learning. Unsupervised methods receive similar input to their supervised counterparts but with the exception of the true labels. These methods attempt to find patterns between samples with the objectives of either distinguishing clusters of samples or discovering a lower-dimensional manifold upon which the original data can be projected. Generally these methods attempt to measure similarity of samples through correlation or a distance metric.

One of the first steps a bioinformatician takes when receiving a new dataset, aside from querying its source, is to perform exploratory analysis. This may include interrogating the quality of the samples, evaluating the significance of batch effects, finding biological signals inherent within the data, and evaluating which samples within the set may share these signals. This latter example is generally a powerful use case for unsupervised machine learning methods. This is perhaps best exemplified in a series of seminal papers at the turn of the century that deduced a connection between the diversity in gene expression profiles and breast cancer subtypes [19,20]. In these studies, hierarchical clustering on gene expression profiles stratified breast cancer samples according to clinically relevant subgroups. In addition, follow up samples that were successfully treated were subsequently clustered with normal samples [20]. This is now an accepted method for exploring tumour heterogeneity which is demonstrated in various large scale cohort studies using Microarray,

bulk RNA-Seq, or scRNA-Seq [16,18,34]. As machine learning has itself further diversified over the decades, largely attributable to the rise of deep learning, so has its new methods been trialed and adopted within bioinformatics. Whereas probabilistic models typically make a single transformation per sample, deep learning merely utilises multiple sequential transformations, typically within a neural network architecture [33]. One exciting question of these methods is whether there is biological significance between transformations. Numerous studies have revealed that these successive transformations have represented different levels of biological processes [35]. Therefore, inspection of trained models could contribute to understanding.

Although discovery is a large application of machine learning in cancer transcriptomes, methods that can be deployed clinically are also highly desirable. Broadly, this would include pursuits in both precision oncology and precision diagnostics. Precision oncology can be defined as matching an individual cancer patient with an efficacious treatment, that typically, is to be based in part on genomic evidence [36–38]. Given that tumours have a variety of driver mutations, some treatments that can specifically target these events have resulted in significant prognostic improvement over conventional chemotherapy. Criticisms of this approach claim that clinical studies have revealed that successful cases are the exception and not the rule [36,38]. However, this may be due to a lack of good drugs being available [36,37]. Supervised machine learning has been used towards progressing this barrier to translation. Deep learning on drug perturbed transcriptomes has been used to both investigate the repurposing of existing drugs and predict drug-target interactions [39,40].

Precision diagnostics is defined to be the use of genomic information to diagnose a patient, for example, segmenting samples according to their transcriptional profile and subsequently, a defined tumour subtype. Prasad et al. [38] outlines four firm benchmarks for success for these methods. Of particular note, that precision diagnostics must be able to reclassify cases that were misclassified or indeterminate using conventional methods. RNA-seq based classifiers have outperformed standard testing when identifying DUX4-rearranged tumour samples [26]. Though, conventional methods have also been shown to also outperform tools applied to RNA-Seq [26]. This may offer the conclusion that newer technologies should not be seen as complete replacements of current methods, but rather complementary, providing a higher level of assurance to clinicians and patients.

Projects

Project 1: B-ALL Acute Lymphoblastic Leukemia Classifier / ALLSorts

Pre B-Cell Acute Lymphocytic Leukemia (B-ALL), is a form of Acute Lymphoblastic Leukemia that arrests maturation of lymphoblasts that leads to their accumulation and subsequent malignancy. The five-year survival of ALL patients has increased from 10% to over 90% over the past 60 years, mainly due to modern chemotherapy, precision oncology, and risk stratification of patients [8,9,41]. In particular, the stratification of B-ALL patients according to tumour subtype can also subsequently segment them according to the aggressiveness and risk that the subtype conveys [18,42,43]. In the case of B-ALL, current standard of care protocols incorporate subtype classification when assigning patients to low, high, standard, and very high risk groups [26,42]. These groups then assist in guiding treatment, attempting to find balance between cure and toxicity [20,41,42]. Although the complete catalogue of B-ALL subtypes are not included within existing risk stratification protocols, indeed further clinical studies are required to validate their utility, it is clear that classifying patients accordingly will further discovery and clinical application.

The application of supervised machine learning algorithms to subtype classification has a rich history in B-ALL. Seminal work at the turn of the century demonstrated that hierarchical classifiers trained on Microarray profiles could predict B-ALL subtypes with high accuracy [44,45]. Building on these datasets, Li et al. [41] constructed six binary SVM classifiers and validated their efficacy on a new, held out dataset that reflected a distinct ethnicity. This study found that though performance did decrease across batches it still exceeded the performance of conventional methods stated in Ross et al. [45]. Over the subsequent years, newer data types have been explored for this purpose. Nordlund et al. [46] used the Nearest Shrunken Centroid method on methylation data and explored classification of samples into multiple subtypes. Lilljebjörn et al. [47] constructed a classifier using RNA-Seq and fusion gene information. More recent efforts have used deep learning models trained on bone marrow microscopy images that could segment samples according to an older designation of subtypes [48].

Classification is not a trivial task as there are many challenges encountered applying these algorithms. For instance, it is interesting to note that most of the historical applications above have required a feature selection step prior to training their classifiers. This is likely required to combat the curse of dimensionality - the phenomenon where the number of samples required for training increases exponentially with the number of features available. In addition, patient cohorts can be subject to batch effects which can alter a classifiers performance in real world application. The Li et al. [41] study explored this as they realised that a model should not be retrained to account for new samples. In addition to the technical challenges of developing a robust classifier, there are insights from the underlying biology that can be explored. For instance, a sample with B-ALL can have multiple subtypes associated, i.e. B-ALL is a *multi-class* and *multi-label classification* problem. This was explored by Nordlund et al. [46] and showed that classifiers may be able to perform this task. Lilljebjörn et al. [47] also demonstrated improved performance by including fusion gene information into their model.

Coinciding with the exploration of methods and generation of new datasets has been the increase in proposed subtypes. Although the World Health Organisation have formally recognised nine distinct subtypes in B-ALL, recent studies by St. Jude Children's Research Hospital, have revealed that their cohorts could be described by at least 23 with associated prognostic risk [18,43].

There is an opportunity to add to this effort by: investigating a B-Cell ALL classifier using the updated 23 subtype designations, exploring the effect of multiple different batches on classifiers performance, attempt to classify known samples with multiple subtypes and known normal samples, and finally explore the effect of integrating domain knowledge into the model. Such a task is now possible with the availability of large cohorts of sequencing information with validated diagnosis.

Project 2: T-ALL Acute Lymphoblastic Leukemia Classifier

T-Cell Acute Lymphoblastic Leukemia (T-ALL) constitutes 10-15% of paediatric and 25% of adult ALL cases, with cure rates of 75% and 50%, respectively [49,50]. However, prognosis is poor for patients who relapse or harbour tumours resistant to current therapies [49]. In addition, intensive chemotherapy regimes to combat these aggressive forms can cause long term harm to patients [8,49]. Therefore, like B-ALL, there is a desire to stratify patients according to cytogenetic subtypes. However, unlike B-ALL, there is only a single provisional subtype listed under WHO's formal designation - ETP-ALL [43]. However, similar to B-ALL, multiple studies have begun to notice distinct genomic lesions and the risks that they correlate to [14,49,50].

Preliminary investigations have revealed no substantial effort to develop a classifier that can stratify T-ALL patients according to a cytogenetic lesion. Prior to the methods applied in Project 1, it would first be necessary to explore the transcriptional landscape to identify whether common subtypes can be defined across cohorts. This presents a significant unsupervised learning challenge. For example, if a potential subtype solely exists within a single cohort, is it possible to delineate between the biological signal and that of the batch? In addition, as is common within B-ALL, can a hierarchy of subtypes, or phenocopies, be systematically identified? Finally, can bulk RNA-Seq gene expression fulfill this task solely, or is it likely that a single cell resolution or complementary sequencing technologies are required? These topics will be investigated further during this project.

Project 3 - Aberrant Splicing Visualisation / Slinker

Visualisation is of critical importance in any quantitative field. If performed correctly it is impactful, it's meaning obvious, and can reflect a greater context through juxtaposition with other information. Bioinformatics is a field of large data and results requiring significant work in visualisation tools to convey biological significance to clinicians and researchers. The Integrative Genomics Viewer is a widely used tool for visualising samples against other sources of information, such as a reference genome and its associated annotations [51]. Typically this would involve aligning data obtained through sequencing to a reference genome. However, RNA-Seq is a reflection of the transcriptome, in which introns are not typically retained. This leads to unnecessary sparsity in the visualisation which can subsequently lead to difficulty in interpretation.

Previous work has developed solutions to this problem. The superTranscript is a novel reference that exclusively consists of concatenated exons and alignment to this resolves the sparsity problem [52]. Further to this, I previously had developed Clinker to provide users with a clear visualisation of fusion genes and their biological context [25]. It achieves this through the concatenation of multiple superTranscripts and generates a custom annotation to describe each. There is an opportunity to improve upon this method. Aberrant splicing events in rare diseases can result in intronic and intergenic sequences being retained within the transcriptome [53]. Therefore, the pursuit of this project is to follow this line of work and develop sample specific superTranscriptomes that can reveal these events in a similar manner as Clinker.

Progress

Project 1: B-ALL Acute Lymphoblastic Leukemia Classifier

The objective of this project is to construct a machine learning classifier that can take gene expression counts as input (samples x genes). The learned model should then provide some measure of confidence that an individual sample has membership with a subtype.

1. Acquisition of Data

Six distinct B-ALL datasets with readily available gene expression counts data were acquired to use within this project. These datasets were used to validate different aspects of the classification. Table 1 outlines their purpose and indicates which data was used for training and what was to be used for testing.

Dataset	Train (%)	Test (%)	Test (dilution) (%)	Test (Multi) (%)	Split Purpose
StJudes (1988)	80	20	0	0	Class stratification
Lund (X)	80	20	0	0	Class stratification
GTEx (Blood)	0	100	0	0	Non-cancer effect
TARGET	0	100	0	0	Batch effects
RCH	0	100	0	0	Batch effects
PeterMac	0	100	0	0	Age effects (Adult)
RCH (dilution)	0	0	100	0	Tumour purity effect
StJudes (Multi)	0	0	0	100	Multi-class efficacy

Table 1. Distribution of six distinct B-ALL datasets across intended purpose and training/testing allocation.

2. Exploratory Data Analysis

It was assumed that batch would have a significant effect in a combined cohort. Therefore, the question was whether a subset of genes could be selected that would reveal the biological signalling expected from the various subtypes. Figure 1 depicts differential gene expression, as calculated by edgeR and visualised in a heatmap, using only the training data. This step validated the idea that subtypes could be segmented by gene expression alone, which is consistent with the literature.

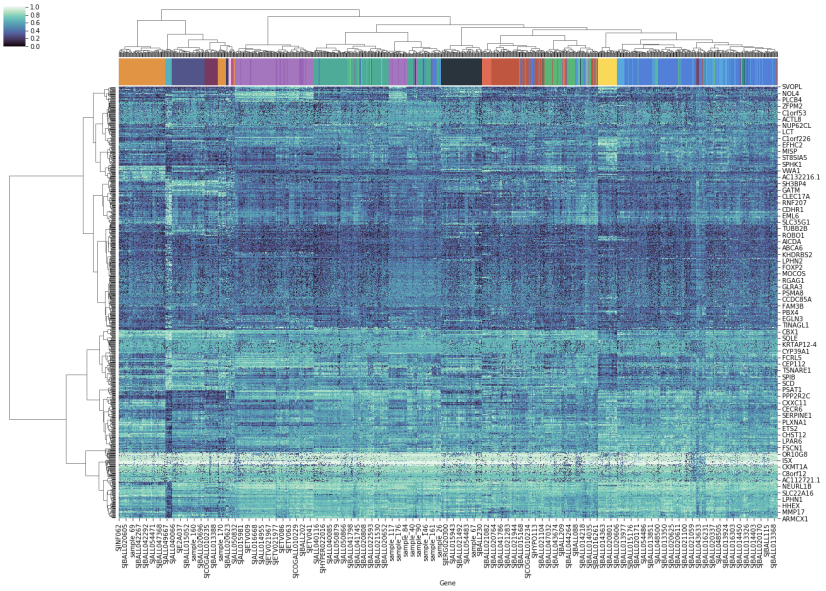


Figure 1. Differential Gene Expression of training data visualised with clustermap - ward linkage (70% StJudes and Lund combination). There is clear segmentation between most of the subtypes when inspecting the clustering of expression of genes with over 1.5 log fold change. Genes (rows) and samples (columns) are a subset chosen by this heatmap that are representative of the clusters.

With these genes selected, batch effects were visualised through Principal Component Analysis (PCA). It is clear that subtype signals are a larger source of variance than batch when relevant genes are selected (Figure 2).

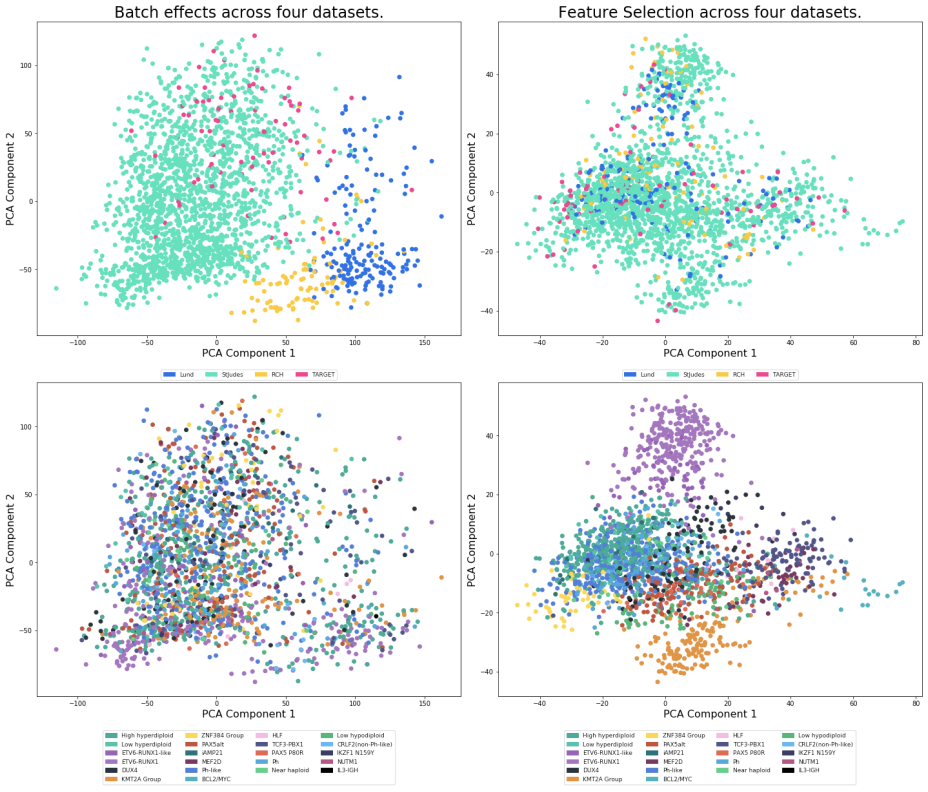


Figure 2. Samples coloured according to batch (top row) and subtypes (bottom row). Log normalised counts, genes filtered with DGE prior to PCA reveal biological signals as the main source of variance (right column).

3. Feature Selection

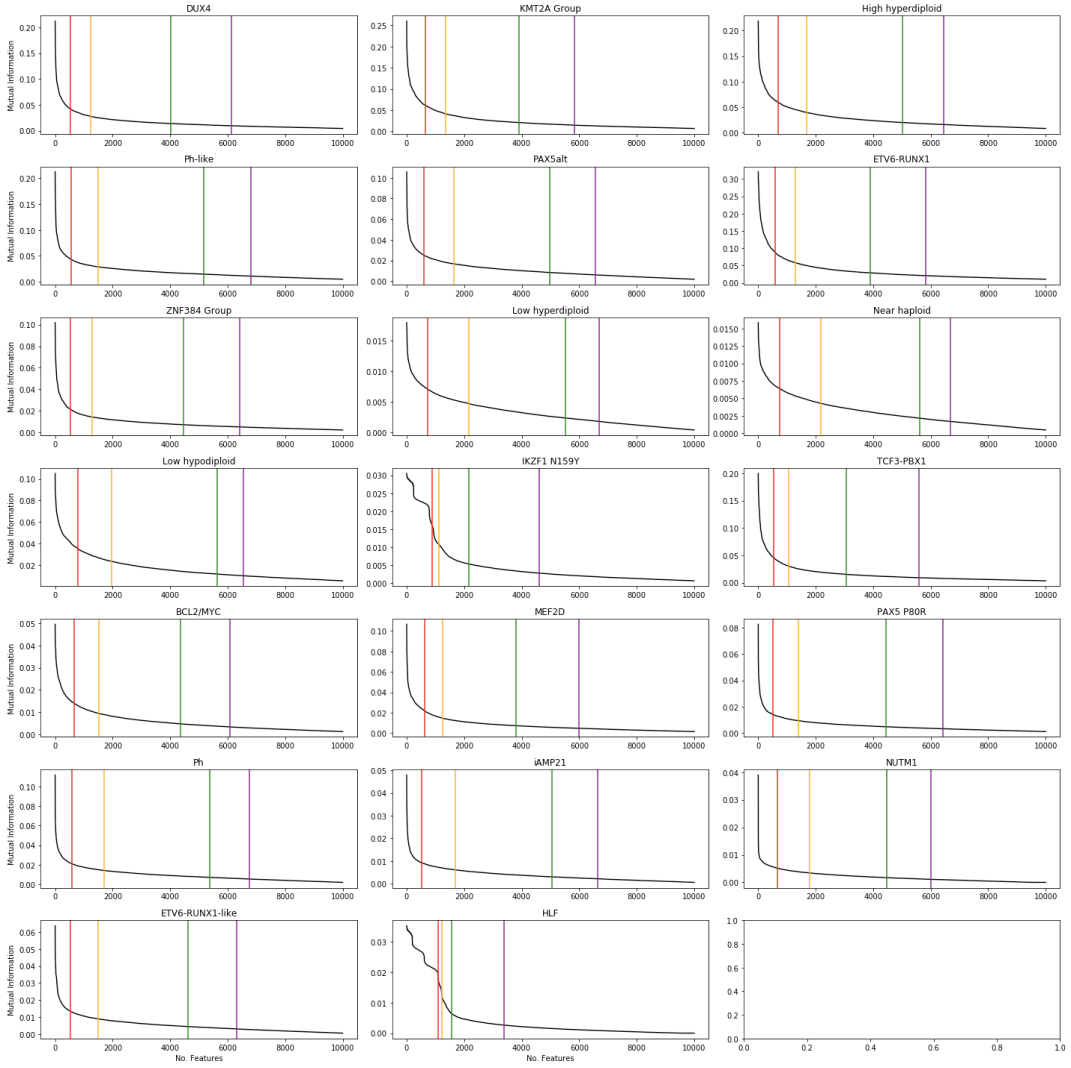


Figure 3. Genes ranked by mutual information calculated between genes and subtype. Vertical lines depict mean (purple) and increasing standard deviations from the mean (red:3, yellow:2, green:1).

Figure 3 explored whether this was also true for feature selection methods that are commonly used within machine learning pipelines. In this case, mutual information was used to quantify the dependency between genes and subtype labels using sklearns *mutual_info_classif*. This function returns a ranked list of genes, with high values indicative of dependency and 0 implying complete independence. Distributions of these scores revealed genes known to be aberrant within a subtype to have relatively high scores. Interestingly, multiple subtypes did not have such clear features. Low hyperdiploid, for example, reveals a more gradual increase in scores rather than the sharp jump seen in subtypes like DUX4 (Figure 3). This perhaps is indicative of individual features describing some subtypes and a collection for others. Therefore, a cutoff of two standard deviations was chosen to be incorporated into the model.

4. Feature Creation

Though there existed a set of features that accurately segmented many subtypes, most ploidy subtypes (High/Low hyperdiploid and Near haploid) and iAMP21 were not as distinct. In an attempt to remedy this custom features were created. First an IAMP21_ratio feature was created which reflected the recurring pattern of expression described in the literature [54]. In short, this single feature is a division of medians between the regions with high and low expression across this characteristic pattern. Figure 4 suggests that this feature is distinguishable in iAMP21 from most subtypes, apart from some outliers which might be attributable to 21+ cases.

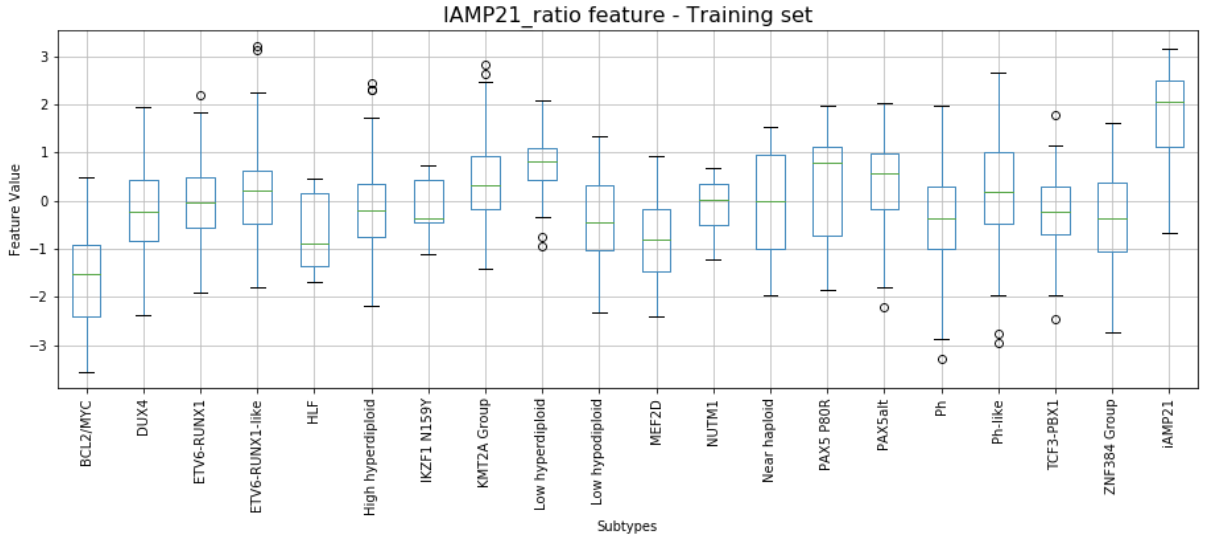


Figure 4. Box and whisker plot of standardised expression of each subtype relative to the new IAMP21_ratio.

In an attempt to characterise the ploidy subtypes, 23 new features were created that reflected the expression of each chromosome. The median absolute deviation was calculated across all training samples and median iterative smoothing was applied to each chromosome - both techniques have seen utility in the literature towards this problem [18,55] (Figure 5). Though some chromosomes across samples are accurately portrayed, many are not. This might be indicative of how recurrent these chromosomes are modified in the ploidy subtypes.

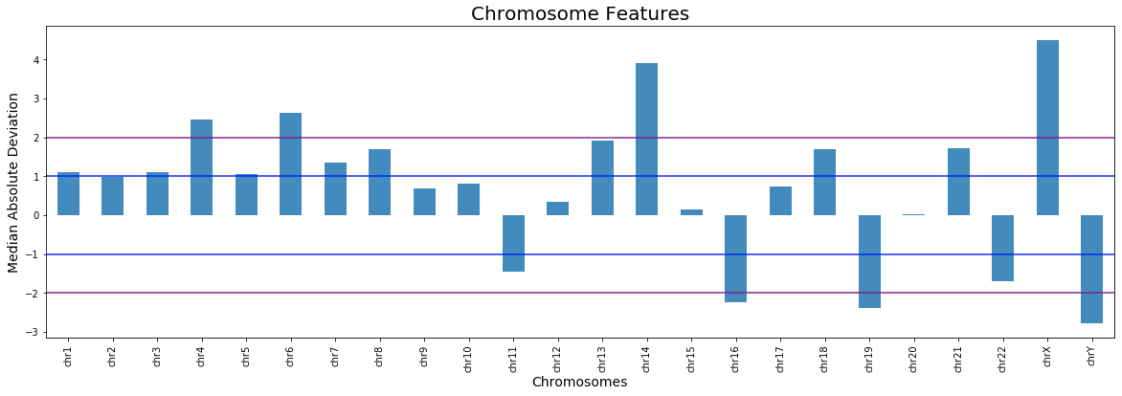


Figure 5 Customised features representing chromosome expression level. The karyotype of this sample is (57,XX,+X,+4,+5,+6,+10,+14,+15,+17,+18,+19,+21[cp3]/46,XX[17]).

With these feature selection and creation methods, clear segmentation of subtypes emerged (Figure 6). Hence, these methods will be incorporated into the training algorithm.

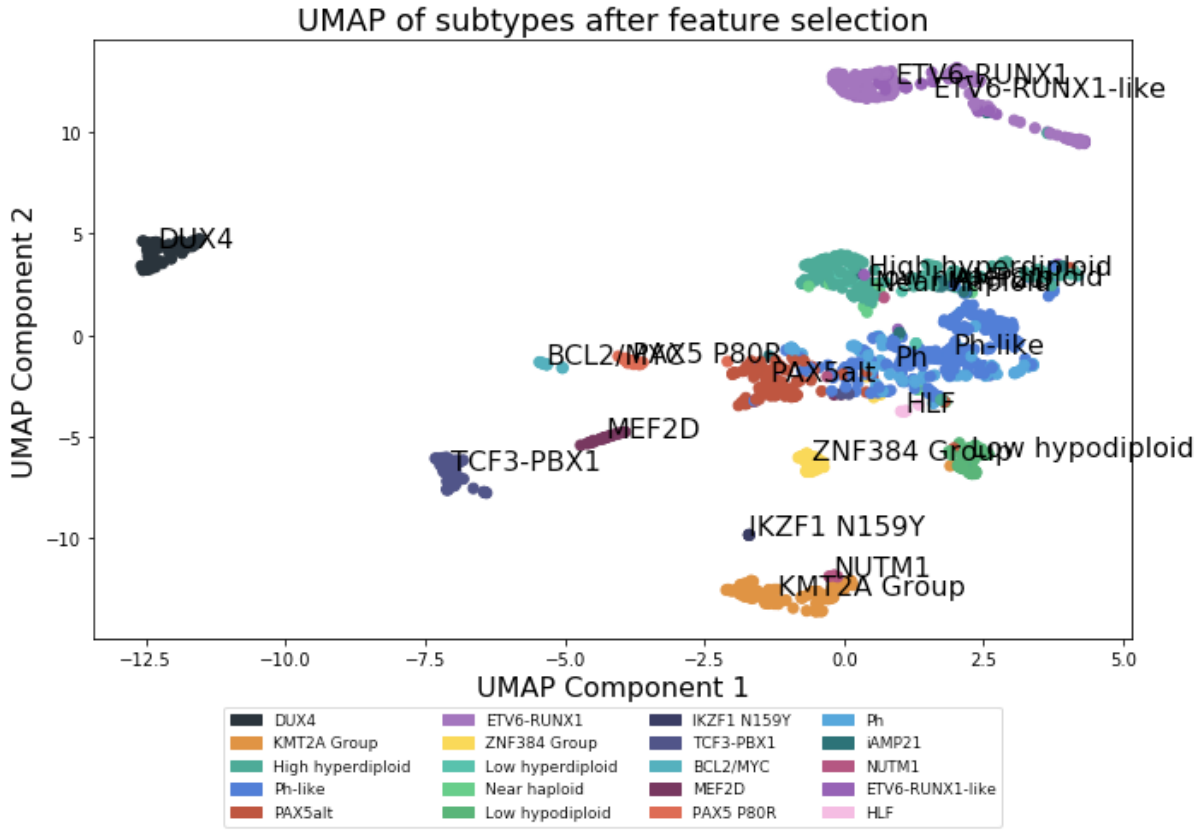


Figure 6. UMAP of samples after log transformation and feature selection using Mutual Information.

5. Training Algorithm

In a similar method to historical B-ALL classifiers, a hierarchical classification approach was utilised [44,45,56]. This was due to the recognition that many of the subtypes are phenocopies. For example, Ph and Ph-like have distinct causal mechanisms, however, they exhibit a similar transcriptional profile. In this paradigm, subtypes are arranged in parent and child relationships, with a classifier trained at each parent. The intention here is that this may aid in interpretability of the model. For instance, which genes are useful in classifying downstream processes of Ph and Ph-like? Could a classifier trained on these downstream events catch novel driving events that were not included within the training data? This could be seen if a child classifier failed to attribute a class but the parent was capable of doing so.

For brevity, the internal mechanisms of the ALLSorts algorithm will not be dissected in major detail here - the results will provide as much insight. Suffice to say that at its core is a Logistic Regression (LR) classifier for every subtype and meta-subtype, limited to classifying siblings at the level in which it has membership (Figure 7). LR was chosen as it was tested to be as high performing as tuned Support Vector Machines (SVM) and Random

Forests (RF), but far less complex. In addition, the probabilistic output is useful for downstream analysis and does not require further calibration like the SVM or RF to achieve.

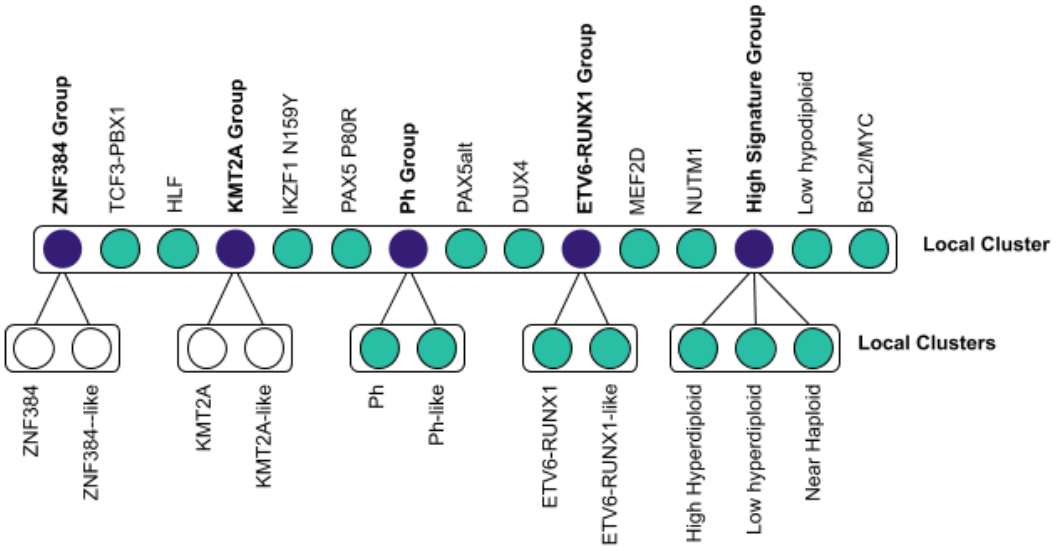


Figure 7. Overview of the ALLSorts classification strategy. For each subtype a Logistic Regression classifier is trained relative to the combined siblings per local group. In purple are the novel meta-subtypes that represent the downstream events that the children converge to or have overlapping overlap. Green nodes are terminal subtypes. White nodes exist in the hierarchy, but classification terminates at the parent node - this is due to a lack of samples. CRLF2(Non Ph-like) is not included in this classification as its identification is better suited to downstream analysis. IL3-IGH is also not included given only a single case across all cohorts.

This hierarchy was compiled from various sources in the literature [14,18,23,57].

Cross-validation is used to tune the probability cutoffs and the regularisation strength for each LR model. This is further explained below.

5. Results

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	Purpose
		Average over individual subtype results.			
Validation Score	89.5618	88.621	79.8566	83.2299	Overfitting
StJudes & Lund	90.1119	97.9397	81.5456	86.8291	Class stratification
GTEx (Blood)	59.2138	100.0	59.2138	74.3827	Normal Effect
TARGET	72.7273	85.5385	67.0455	72.9294	Batch effects
RCH	83.75	84.1667	77.7381	77.5426	Batch effects
PeterMac	TBA	TBA	TBA	TBA	Age effects (Adult)
RCH (dilution)	TBA	TBA	TBA	TBA	Dilution effects
	Two correct (%)	**One subtype (%)**	**Unclassified (%)**	**At least one (%)**	

StJudes (Multi)	19.39	61.22	19.39	80.61	Multi-class efficacy

Table 2. Classifiers performance for each of the datasets listed in Table 1. Note, TBA results are still to be finalised. Precision, Recall, and F1 scores are the average across each subtype. Validation score was found through 5 cross-fold validation. Multi-class results were evaluated on accuracy alone using: Prediction matched both true labels, matched only one label, was considered “unclassified”, and at least one class predicted.

Interpretation of results: Royal Children’s Hospital example

To better understand these, a significant feature of ALLSorts is the visualisation of the results. Figures 8 and 9 offer further guidance in interpreting these Royal Children’s Hospital results in Table 2. Figure 8 depicts the distribution of probabilities for every sample per subtype. The significance of this visualisation is that a quick overview on the performance of the classifier can be considered with regards to a new cohort.

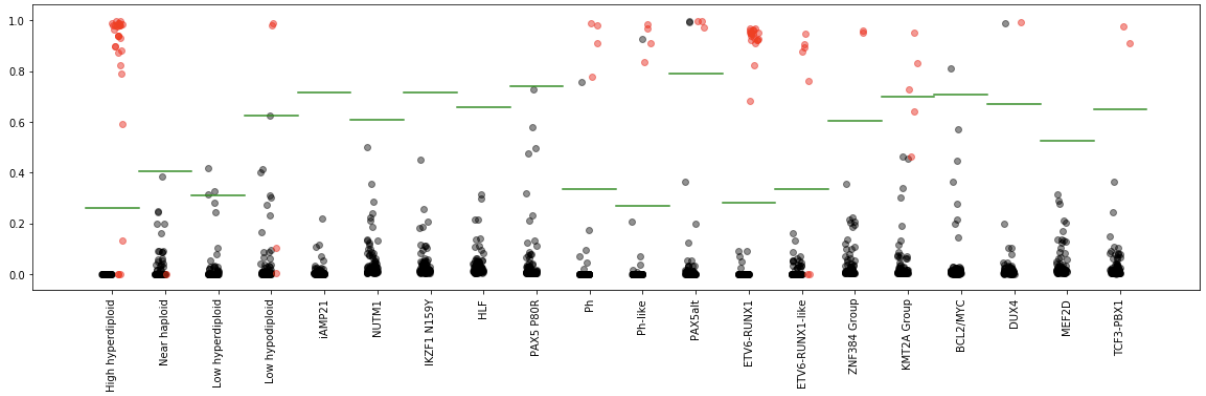


Figure 8. Probabilities distribution of RCH samples per subtype. Black dots are samples that are the negative label for that subtype, red is the positive. The green lines are probability thresholds which are calculated through cross validation based on F1 score or maximal distance between highest negative label and lowest positive label. Samples that have parents which do not meet threshold have probabilities set to 0.0.

Figure 9 depicts ALLSorts “waterfall” plot. The value of this visualisation is that it demonstrates how effective a classification is relative to all the others in that subtype.

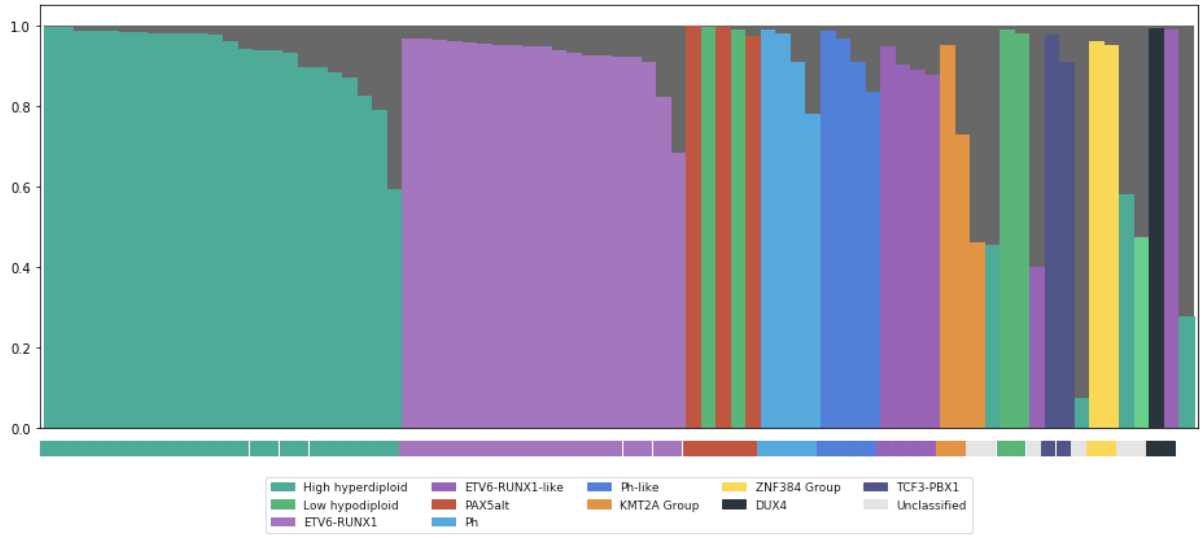


Figure 9. Waterfall plot of RCH data. The X-axis depicts the predicted class, the Y axis is the probability of belonging to a subtype. The colours within the plot represent the true label. Note: This does not show subtypes that have multiple labels associated, for example, Figure 8 has misclassified Ph over threshold, but this has also been classified as KMT2A and is thus not shown in this plot - this will need to be depicted in another form.

Though the precision in the RCH data seems relatively high, these plots quickly deduce that there may be some mislabelling within the dataset. This is understandable given that these were classified using a previous version of ALLSorts which had far fewer classes. Figures 8 and 9 quickly demonstrate that there are PAX5alt and DUX4 samples that may need relabelling, as they seem to have a similar probability to other members in the class. On the other hand, the KMT2A subtypes surrounding the threshold show less confidence and should be explored further.

Project 2: T-ALL Acute Lymphoblastic Leukemia Classifier

A natural follow on from B-ALL is to approach a more difficult problem, but in the same domain. T-ALL is starting to reveal itself as having a heterogenous molecular landscape, with recurrent lesions being correlated with prognosis. Therefore, the objective of this project will attempt to explore this possibility and, if successful, perhaps construct the first T-ALL classifier to my knowledge.

Three datasets are readily available to begin exploring subtypes within T-ALL: St Judes Hospital, the Royal Children's Hospital, and Peter MacCallum. The first step in this was to explore whether an unsupervised learning algorithm could infer groupings between samples that somewhat resembled proposed subtypes. Figures 10 and 11 demonstrate the allocation of subtypes within the StJudes cohort according to the K-means clustering algorithm and expert curation, respectively. Some of the proposed subtypes are clearly distinguishable from the others, yet others appear more challenging. This provides some merit to the idea that transcriptomes can be used to investigate this problem.

UMAP coloured with K-Means Clustering

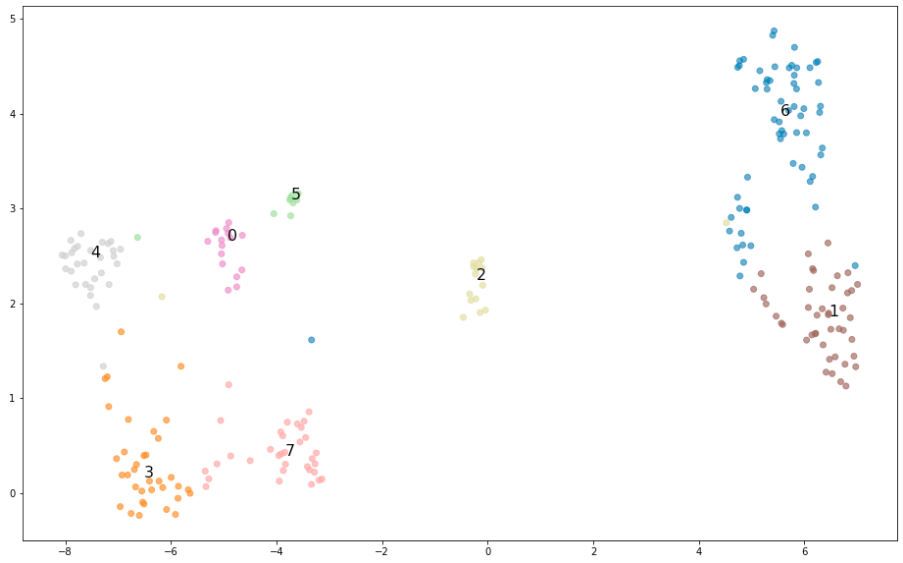


Figure 10. UMAP plot of T-ALL samples sourced from TARGET, coloured according to K-means marked clusters.

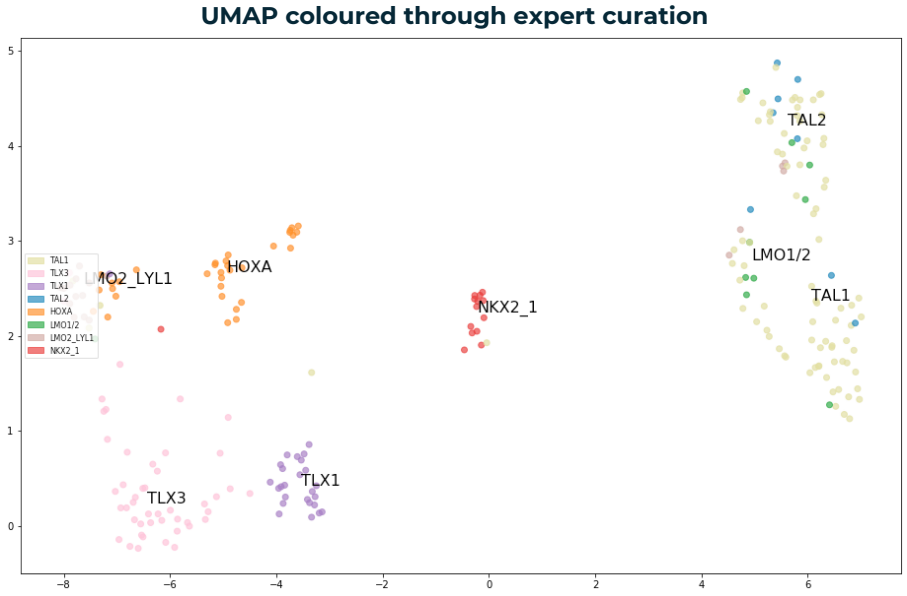


Figure 11. UMAP plot of T-ALL samples coloured according to expert curation [50].

Project 3: Aberrant Splicing Visualisation / Slinker

Previous work featured the development of Clinker, a visualisation tool that aligns RNA-Seq reads to superTranscripts and generates a visualisation of the result. A superTranscript is the collection of all transcripts for a gene, collapsed into a single reference, that exclusively consists of concatenated exons [25]. However, in many circumstances rare diseases will result from a novel splice variant that may include intronic sequence [53]. Figure 12 depicts such a scenario. Although this visualisation is effective, it is not typical. In general intronic regions are significantly longer than exons, resulting in very sparse visualisations. However, this can be resolved with the superTranscript paradigm. As such, a pipeline has been developed that can perform genome-guided assembly of

transcripts from alignments, convert these into superTranscripts, and then perform a visualisation process similar to Clinker [25]. This method will allow these novel regions to be included within the superTranscript despite existing outside referenced exon boundaries. Figures 12 and 13 demonstrate the differences between the standard visualisation after alignment to the genome and the superTranscriptome, respectively. The novel event is clearly displayed in the latter case through the clear contrast in coverage, the novel exon annotation, and the presence of a splice junction between those positions. Further work could also include the genomic alignment, with lines tracking to the relevant regions in the superTranscript.

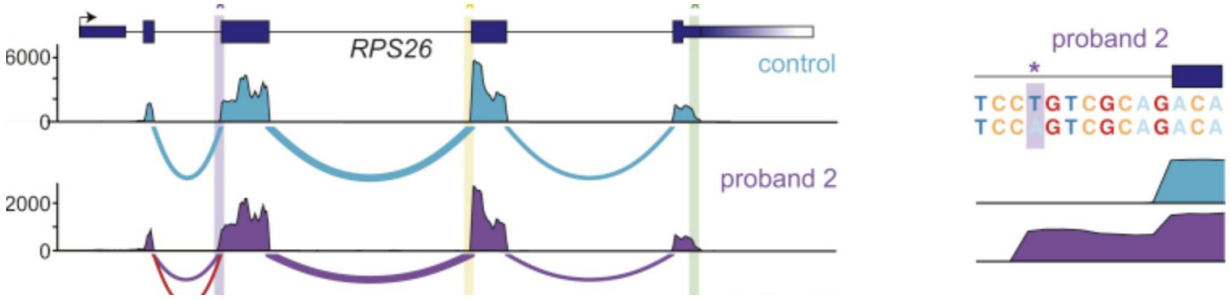


Figure 12. Plot from recent paper depicting non-canonical splice variants [53]. In this example, reads are aligned to the reference genome and a sashimi plot is generated. Whilst informative, the introns are a redundancy that can potentially be removed without invalidating the meaning.

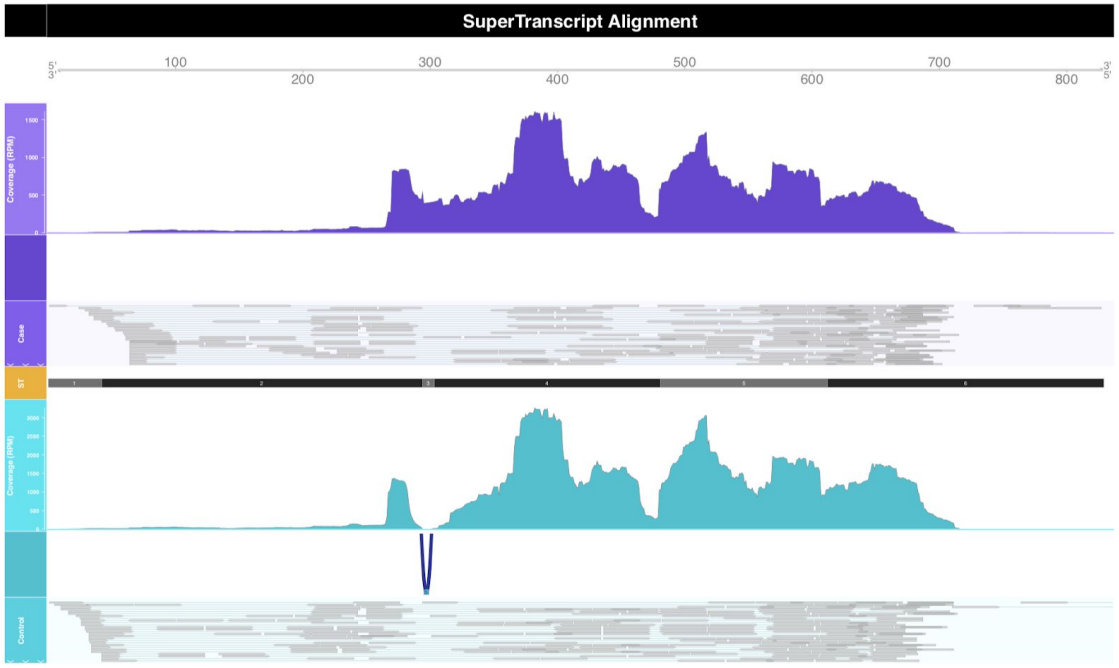


Figure 13. Work in progress of Slinker - a method to align to a bespoke superTranscript which includes novel inclusions of sequence. Without the intronic sequence, the distinct drop in coverage between the two samples is evident.

Further work is required to distinguish the novel event from a reference, but it is clearly effective in this early stage of development.

Thesis and Research Plan

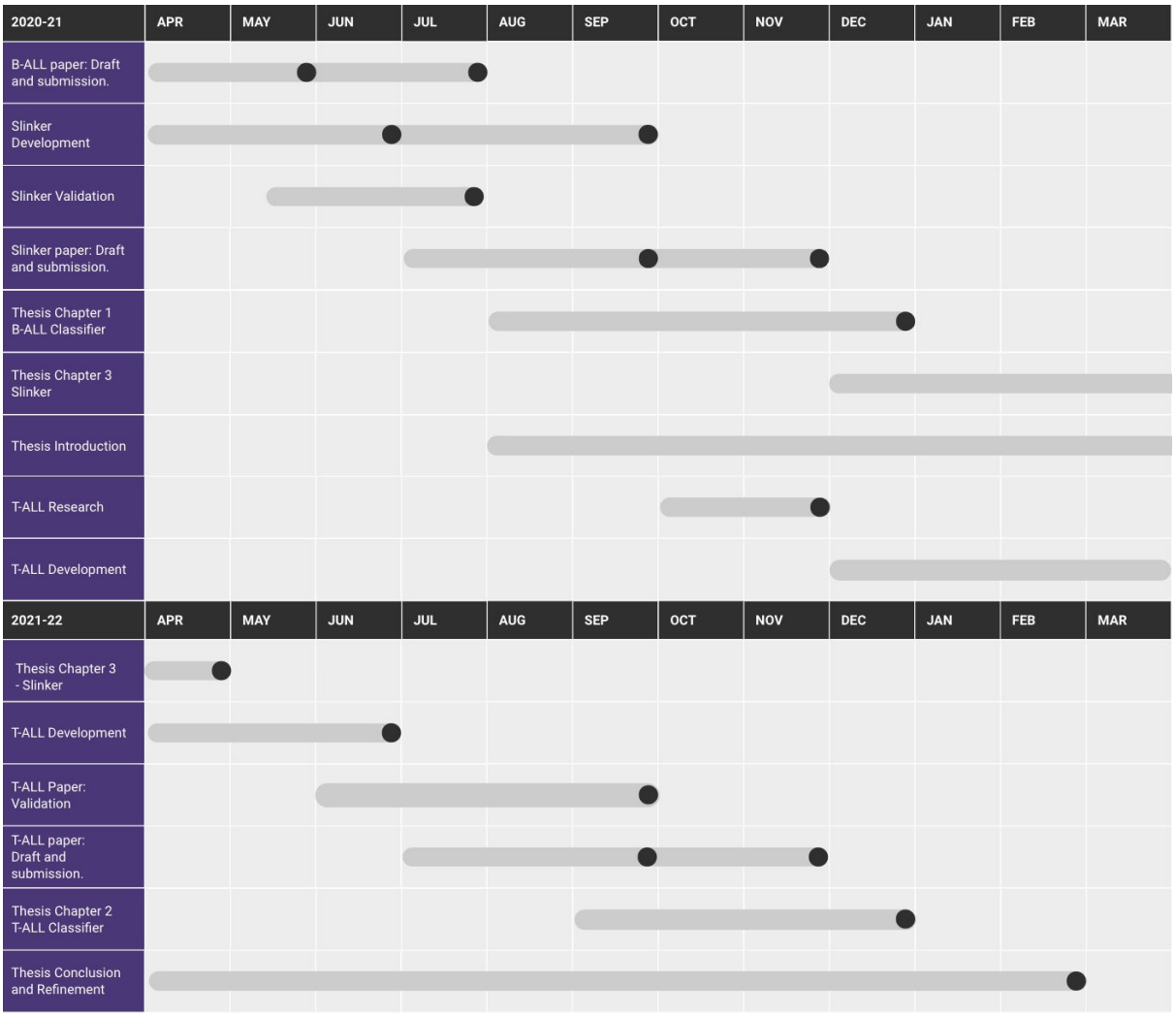


Figure 14. GANTT chart indicating priorities and timeline over next 24 months. The top half is for 20/21, the bottom for 21/22. This is an ideal projection based on completing the overall PhD in three years.

The objective over the next two years is “write soon and write often”. I will endeavour to begin writing my thesis in earnest by mid year, with hopefully a published paper and draft guiding the ongoing work. Currently I have conservatively estimated the three projects taking up the majority of my output, however, if time allows I will consider additional projects of varying scales depending on the circumstance. I believe this timeline is achievable and will provide enough content for a successful thesis and PhD completion.

1. Alberta, B. *et al. Molecular Biology of the Cell*. (Garland Pub - Usa, 2008).
2. McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168**, 613–628 (2017).
3. Finkel, T., Serrano, M. & Blasco, M. A. The common biology of cancer and ageing. *Nature* **448**, 767–774 (2007).
4. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
5. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
6. McFarland, C. D., Mirny, L. A. & Korolev, K. S. Tug-of-war between driver and passenger mutations in cancer and other adaptive processes. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 15138–15143 (2014).
7. Pon, J. R. & Marra, M. A. Driver and passenger mutations in cancer. *Annu. Rev. Pathol.* **10**, 25–50 (2015).
8. Adamson, P. C. Improving the outcome for children with cancer: Development of targeted new agents. *CA Cancer J. Clin.* **65**, 212–220 (2015).
9. Brown, L. M. *et al.* Different Classes of ABL1 Fusions Activate Different Downstream Signalling Nodes. *Blood* **132**, 2628–2628 (2018).
10. Salvadores, M., Mas-Ponte, D. & Supek, F. Passenger mutations accurately classify human tumors. *PLoS Comput. Biol.* **15**, e1006953 (2019).
11. Cieślik, M. & Chinnaiyan, A. M. Cancer transcriptome profiling at the juncture of clinical translation. *Nat. Rev. Genet.* **19**, 93–109 (2018).
12. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **174**, 1034–1035 (2018).
13. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337.e10 (2018).
14. Hunger, S. P. & Mullighan, C. G. Redefining ALL classification: toward detecting

high-risk ALL and implementing precision medicine. *Blood* **125**, 3977–3987 (2015).

15. Way, G. P. *et al.* Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. *Cell Rep.* **23**, 172–180.e3 (2018).
16. Marisa, L. *et al.* Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med.* **10**, e1001453 (2013).
17. Young, J. D., Cai, C. & Lu, X. Unsupervised deep learning reveals prognostically relevant subtypes of glioblastoma. *BMC Bioinformatics* **18**, 381 (2017).
18. Gu, Z. *et al.* PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. *Nat. Genet.* **51**, 296–307 (2019).
19. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 10869–10874 (2001).
20. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).
21. Li, J.-F. *et al.* Transcriptional landscape of B cell precursor acute lymphoblastic leukemia based on an international study of 1,223 cases. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E11711–E11720 (2018).
22. Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D. & Craig, D. W. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.* **17**, 257–271 (2016).
23. Lilljebjörn, H. & Fioretos, T. New oncogenic subtypes in pediatric B-cell precursor acute lymphoblastic leukemia. *Blood* **130**, 1395–1401 (2017).
24. Davidson, N. M., Majewski, I. J. & Oshlack, A. JAFFA: High sensitivity transcriptome-focused fusion gene detection. *Genome Med.* **7**, 43 (2015).
25. Schmidt, B. M. *et al.* Clinker: visualizing fusion genes detected in RNA-seq data. *Gigascience* **7**, (2018).

26. Brown, L. M. *et al.* The application of RNA sequencing for the diagnosis and genomic classification of pediatric acute lymphoblastic leukemia. *Blood Adv* **4**, 930–942 (2020).
27. Rasnic, R., Brandes, N., Zuk, O. & Linial, M. Substantial batch effects in TCGA exome sequences undermine pan-cancer analysis of germline variants. *BMC Cancer* **19**, 783 (2019).
28. Nygaard, V., Rødland, E. A. & Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* **17**, 29–39 (2016).
29. Huang, S., Ernberg, I. & Kauffman, S. Cancer attractors: a systems view of tumors from a gene network dynamics and developmental perspective. *Semin. Cell Dev. Biol.* **20**, 869–876 (2009).
30. Batlle, E. & Clevers, H. Cancer stem cells revisited. *Nat. Med.* **23**, 1124–1134 (2017).
31. Zhang, W., Chien, J., Yong, J. & Kuang, R. Network-based machine learning and graph theory algorithms for precision oncology. *NPJ Precis Oncol* **1**, 25 (2017).
32. Danaee, P., Ghaeini, R. & Hendrix, D. A. A DEEP LEARNING APPROACH FOR CANCER DETECTION AND RELEVANT GENE IDENTIFICATION. *Pac. Symp. Biocomput.* **22**, 219–229 (2017).
33. Chollet, F. *Deep Learning with Python*. (Manning Publications Company, 2017).
34. Karaayvaz, M. *et al.* Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nat. Commun.* **9**, 3588 (2018).
35. Chen, L., Cai, C., Chen, V. & Lu, X. Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. *BMC Bioinformatics* **17 Suppl 1**, 9 (2016).
36. Letai, A. Functional precision cancer medicine—moving beyond pure genomics. *Nat. Med.* **23**, 1028–1035 (2017).
37. Mody, R. J., Prensner, J. R., Everett, J., Parsons, D. W. & Chinnaiyan, A. M. Precision

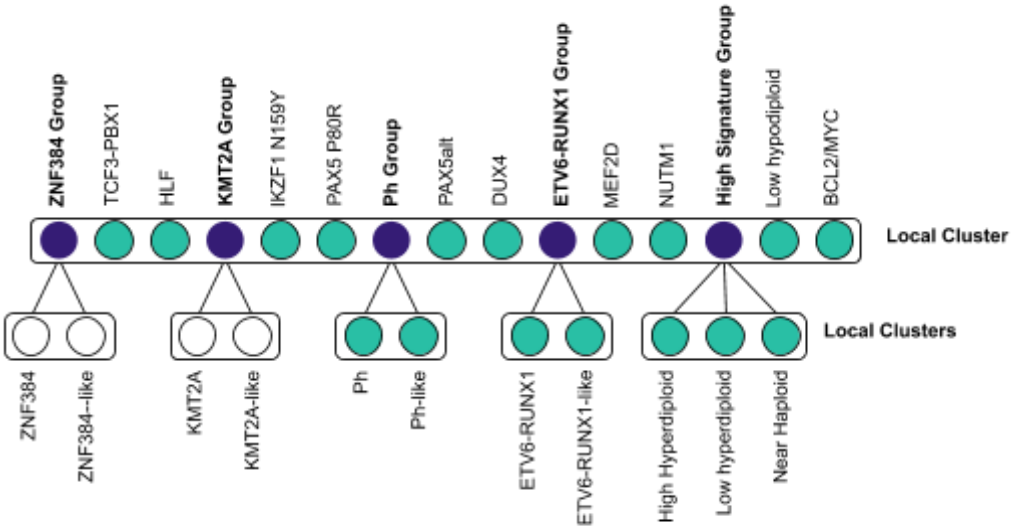
medicine in pediatric oncology: Lessons learned and next steps. *Pediatr. Blood Cancer* **64**, (2017).

38. Prasad, V., Fojo, T. & Brada, M. Precision oncology: origins, optimism, and potential. *Lancet Oncol.* **17**, e81–e86 (2016).
39. Aliper, A. *et al.* Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. *Mol. Pharm.* **13**, 2524–2530 (2016).
40. Xie, L., He, S., Song, X., Bo, X. & Zhang, Z. Deep learning-based transcriptome data classification for drug-target interaction prediction. *BMC Genomics* **19**, 667 (2018).
41. Li, Z. *et al.* Gene expression–based classification and regulatory networks of pediatric acute lymphoblastic leukemia. *Blood* **114**, 4486–4493 (2009).
42. Schultz, K. R. *et al.* Risk- and response-based classification of childhood B-precursor acute lymphoblastic leukemia: a combined analysis of prognostic markers from the Pediatric Oncology Group (POG) and Children's Cancer Group (CCG). *Blood* **109**, 926–935 (2007).
43. Arber, D. A. *et al.* The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405 (2016).
44. Yeoh, E.-J. *et al.* Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* **1**, 133–143 (2002).
45. Ross, M. E. *et al.* Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood* **102**, 2951–2959 (2003).
46. Nordlund, J. *et al.* DNA methylation-based subtype prediction for pediatric acute lymphoblastic leukemia. *Clin. Epigenetics* **7**, 11 (2015).
47. Lilljebjörn, H. *et al.* Identification of ETV6-RUNX1-like and DUX4-rearranged subtypes in paediatric B-cell precursor acute lymphoblastic leukaemia. *Nat. Commun.* **7**, 11790

(2016).

48. Rehman, A. *et al.* Classification of acute lymphoblastic leukemia using deep learning. *Microsc. Res. Tech.* **81**, 1310–1317 (2018).
49. Van Vlierberghe, P. & Ferrando, A. The molecular basis of T cell acute lymphoblastic leukemia. *J. Clin. Invest.* **122**, 3398–3406 (2012).
50. Liu, Y. *et al.* The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia. *Nat. Genet.* **49**, 1211–1218 (2017).
51. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
52. Davidson, N. M., Hawkins, A. D. K. & Oshlack, A. SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes. *Genome Biol.* **18**, 148 (2017).
53. Ulirsch, J. C. *et al.* The Genetic Landscape of Diamond-Blackfan Anemia. *Am. J. Hum. Genet.* **104**, 356 (2019).
54. Tsuchiya, K. D., Davis, B. & Gardner, R. A. Is intrachromosomal amplification of chromosome 21 (iAMP21) always intrachromosomal? *Cancer Genet.* **218-219**, 10–14 (2017).
55. Serin Harmanci, A., Harmanci, A. O. & Zhou, X. CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. *Nat. Commun.* **11**, 89 (2020).
56. Silla, C. N. & Freitas, A. A. A survey of hierarchical classification across different application domains. *Data Min. Knowl. Discov.* **22**, 31–72 (2011).
57. Safavi, S. & Paulsson, K. Near-haploid and low-hypodiploid acute lymphoblastic leukemia: two distinct subtypes with consistently poor prognosis. *Blood* **129**, 420–423 (2017).

Drawings


ZNF384 Group
TCF3-PBX1
HLF
KMT2A Group
IKZF1 N159Y
PAX5 P80R
Ph Group
PAX5alt
DUX4
ETV6-RUNX1 Group
MEF2D
NUTM1
High Signature Group
Low hypodiploid
BCL2/MYC
Local Cluster
Local Clusters
ZNF384
ZNF384-like
KMT2A
KMT2A-like
Ph
Ph-like
ETV6-RUNX1
ETV6-RUNX1-like
High Hyperdiploid
Low hyperdiploid
Near Haploid