This project was generously funded by the George Henderson and George Brown endowments funds, as well as the Newcastle University Summer Research Scholarships Scheme.

Can Machine Learning Identify Patients with iAMP21, a High Risk of Tumour Relapse Subtype of Childhood Acute Lymphoblastic Leukaemia?

Keaan Amin[1], Dr Matthew Bashton[1], Dr Amir Enshaei[1]

[1]Leukaemia Research Cytogenetics Group, Wolfson Childhood Cancer Research Centre, Newcastle University

Bloodwise
Beating blood cancer since 1960

1. Introduction

Acute Lymphoblastic Leukaemia (ALL) is the most common cancer in children. Leukaemia is a blood cancer, characterised by the accumulation of mutations in the bone marrow. **iAMP21** is a high-risk subgroup of B-cell ALL where patients are **three times more** at **risk** of a **tumour relapse** over other B-cell ALL subgroups (1). Subsequently, reliable detection of this subgroup can enable intensification of chemotherapy and more targeted treatments.

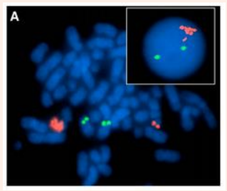

***Figure 1**. Patient cells with iAMP21, detected by RUNX1 probe markers (3).*

iAMP21 subgroup is characterised by a series of multiple genetic alterations within a precise region on Chromosome 21. This is currently detectable by a highly visual gene marker, RUNX1, under microscopy (Figure 1). However, RUNX1 is also known to be poor gene expression-based marker (2). This suggests that further genes are likely to be active (3).

Machine learning (ML) techniques are a form of artificial intelligence capable of detecting subtle patterns, particularly within gene expression-based datasets (4,5). Thus, we can feasibly classify patients with iAMP21 from other ALL patients using gene expression. This could form a valuable tool in diagnostic clinical decision-making processes.

2. Aims

a) Detect a set of active genes of high importance in iAMP21 positive patients.
b) Are these genes predictive of iAMP21 compared to low-risk ALL patients using machine learning?
c) What is the impact of these genes on patient response to therapy?

3. Methods

59 childhood ALL patient samples were acquired at diagnosis *(Figure 2)* from bone marrow biopsies. **16** samples were **iAMP21-subtype** positive and **33** were a diverse **non-iAMP21** group from another major ALL subtype, **B-other**. Each of the samples were processed through RNA-Sequencing, a form of large-scale gene analysis, to generate a measure of gene expression.

Gene Selection → Training + Test Groups → Tuning Parameters → Quality Metrics

I. **Gene Selection** was required to reduce the complexity of our model from 20,375 protein-coding genes. Lasso, a linear regression technique, was most effective in rigorously reducing this uncertainty to provide more reproducible results. Visual clustering analysis was performed (Figure 3).
II. Our 59 patient samples were randomly allocated to our **training** (80%) set to learn patterns and **test** set (20%).
III. Our applied **ML algorithms** included linear-based classifiers, tree-based classifiers (Random Forests), K-nearest neighbours and Support Vector Machines (Linear and Radial).
IV. Our models were **fine-tuned** and **cross-validated** over 30 iterations using Leave Group Out Cross-Validation.
V. To identify the best performing model, we plotted receiver operating characteristic (ROC) curves and boxplots of sensitivity and specificity rates.

4. Results

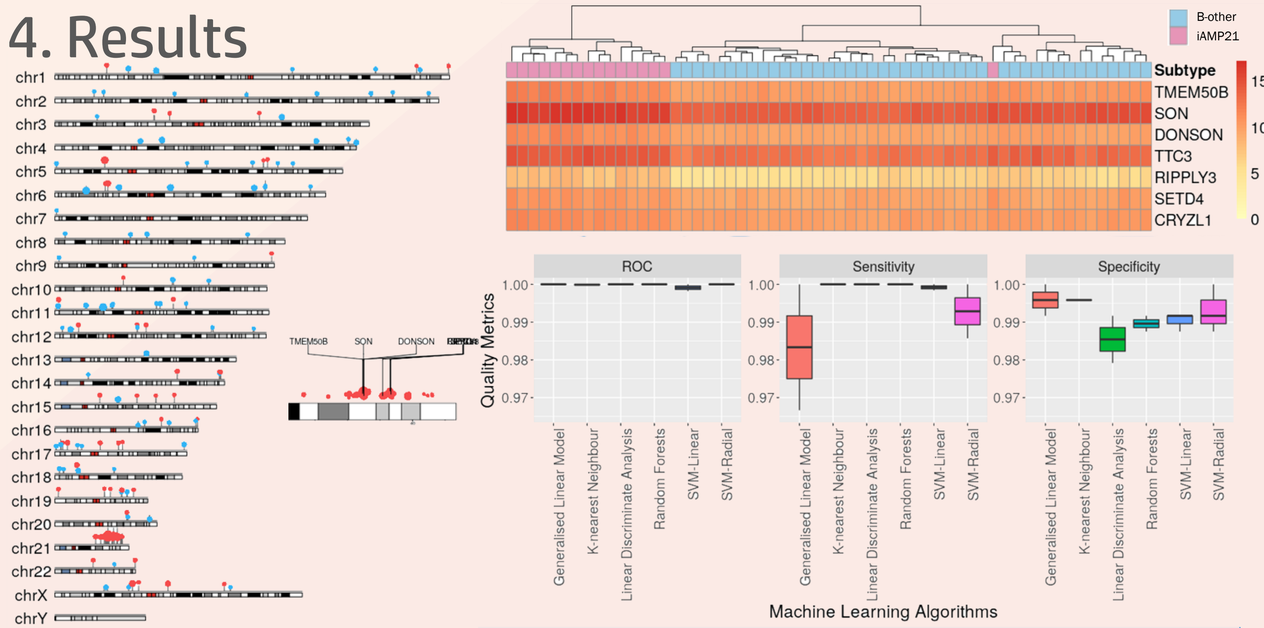

***Figure 2** (top-left). Chromosome location of genes in iAMP21 (over-expressed = red, under-expressed = blue) with our seven gene Chromosome 21 group (labelled). **Figure 3** (top-right). Gene Expression of our selected seven gene Chromosome 21 group. **Figure 4** (bottom-right). Machine Learning Algorithm performance in detecting iAMP21 across variable tuning parameters.*

We successfully detected seven significantly over-expressed genes, located on Chromosome 21 (Figure 2) and this collective gene group is useful in grouping iAMP21 patients (Figure 3).

Through machine learning, this seven gene Chromosome 21 group were extremely predictive of iAMP21 patients with a perfect level of accuracy and sensitivity, through SVM-Radial and Generalised Linear Models (Figure 4). After extensive cross-validation, both models continued to perfectly separate samples.

5. Conclusion

This project identifies a seven gene Chromosome 21 group can perfectly distinguish iAMP21 patients from other B-Cell ALL patients through the application of machine learning.

These seven gene-expression-based markers support the positive detection of iAMP21 patients, known to be at a significantly increased risk of tumour relapse. This could be a usual supportive diagnostic method of detecting iAMP21 patients (a high-risk subgroup of ALL).

The acquisition of further samples from other major leukaemia subtypes would enable the quick re-application of our methods to define further gene groupings. This would further support the potential role of gene-expression based markers and machine learning as an adjunctive diagnostic tool.

References

1. Harrison CJ, et al. An international study of intrachromosomal amplification of chromosome 21 (iAMP21): cytogenetic characterization and outcome. Leukemia. 2014;28(5):1015-21.
2. Moorman AV, et al. Risk-directed treatment intensification significantly reduces the risk of relapse among children and adolescents with acute lymphoblastic leukemia and intrachromosomal amplification of chromosome 21: a comparison of the MRC ALL97/99 and UKALL2003 trials. J Clin Oncol. 2013;31(27):3389-96.
3. Wenric S, Shemirani R. Using Supervised Learning Methods for Gene Selection in RNA-Seq Case-Control Studies. Front Genet. 2018;9:297.
4. Gu Z, Churchman ML, Roberts KG, Moore I, Zhou X, Nakitandwe J, et al. PAX5-driven subtypes of B-progenitor acute lymphoblastic leukemia. Nat Genet. 2019;51(2):296-307.
5. Harrison CJ. Blood Spotlight on iAMP21 acute lymphoblastic leukemia (ALL), a high-risk pediatric disease. Blood. 2015;125(9):1383-6.