Using RNA-Seq data to predict large-scale copy number alterations in Acute Myeloid Leukemia

Carmen Bayly, Lauren Chong, Rod Docking, Fatemeh Dorri, Emily Hindalong, Rebecca Johnston.

Can RNA-seq differentiate the key karyotypes of Acute Myeloid Leukemia?

Hypothesis: The transcriptional profile of Acute Myeloid Leukemia patients is predictive of their inherent cytogenetic abnormalities and prognosis.

Aims:
1. Characterize the differences in RNA expression between different cytogenetic risk groups and patients with or without key large-scale copy number alterations (CNAs);
2. Train classifiers that predict the cytogenetic risk groups in Acute Myeloid Leukemia using only RNA-seq data.

INTRODUCTION

Acute Myeloid Leukemia (AML) is a cancer of the blood cells characterized by recurrent, large-scale chromosomal abnormalities. Detection of these abnormalities is critical for proper treatment stratification[1, 2].

Figure 1: Trisomy 8 karyotype[3]

The current standard of care involves testing of both cytogenetic and molecular markers, assessed through karyotyping and PCR-based tests, respectively. However, as new research identifies more clinically relevant markers, these serial tests are difficult to scale.

RNA-seq data provides an unbiased view of gene expression, and allows the detection of expressed structural variants that may be functionally relevant to this tumour phenotype.

There are **three levels of cytogenetic risk status** according to current treatment guidelines: good, intermediate, and poor risk. These different classes correspond to the presence or absence of specific CNA events.

RISK STATUS	CYTOGENETICS
Good -risk	Inv(16) or t(16;16) t(8;21) t(15;17)
Intermediate-risk	Normal cytogenetics **+8 alone** t(9;11) Other non-defined
Poor-risk	Complex Monosomal karyotype **-5, 5q-, -7, 7q-** 11q23 – non t(9;11) inv(3), t(3;3) t(6;9) t(9;22)

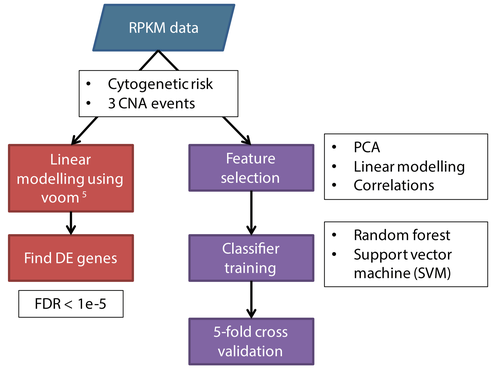
Table 1: Cytogenetic risk groups[1]

METHODOLOGY

Dataset: We have re-analyzed data made available through a recent publication by The Cancer Genome Atlas[4]. The data-set consists of RNA-seq libraries from 179 patients with AML, with matched clinical data noting CNA events observed through standard cytogenetics.

Cytogenetic Risk	Count
Good	33
Intermediate	101
Poor	42
Undetermined	3

CNA Event	Present	Absent
+8	19	160
–5, 5q-	16	163
–7, 7q-	21	158



RESULTS: Machine learning

Unsupervised feature selection

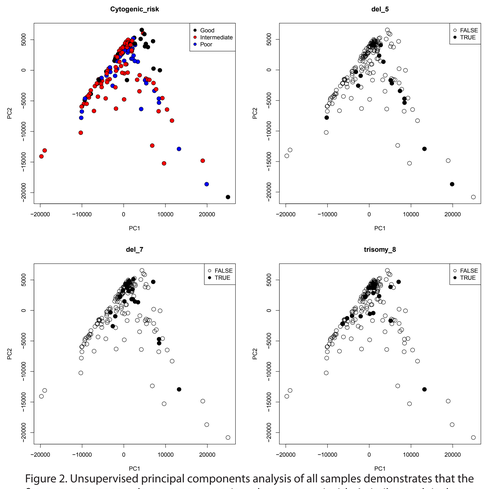


Figure 2. Unsupervised principal components analysis of all samples demonstrates that the first two components do not separate patients by cytogenetic risk. A similar result is shown for the three large-scale copy number alterations (del_5, del_7, trisomy_8).

	Support vector machine			
	Basic PCA		Kernelized PCA	
	Sens.	Spec.	Sens.	Spec.
Poor	0.45	0.84	0.31	0.80
Intermediate	0.70	0.66	0.60	0.67
Good	0.64	0.96	0.58	0.91
Trisomy_8	0.26	0.91	0.11	0.92
Del_5	0.44	0.94	0.13	0.94
Del_7	0.38	0.98	0.29	0.95

Table 2. Kernelized SVMs trained using the first 20 principal components as features showed low sensitivity for predicting cytogenetic risk and copy number alterations.

Supervised feature selection

	Random forest				Support vector machine			
	Linear model		Correlation		Linear model		Correlation	
	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
Poor	0.57	0.94	0.38	0.96	0.55	0.96	0.45	0.94
Intermediate	0.81	0.90	0.61	0.90	0.87	0.87	0.69	0.85
Good	0.85	0.99	0.85	0.99	0.94	0.99	0.82	0.99
Trisomy_8	0.68	0.99	0.68	0.99	0.68	0.97	0.58	0.99
Del_5	0.81	0.99	0.44	0.99	0.63	0.99	0.19	0.99
Del_7	0.71	0.98	0.71	0.98	0.71	0.99	0.67	0.98

Table 3. The sensitivity and specificity of random forests and SVMs trained using two different supervised feature selection methods: 1) Taking the top 25 differentially expressed genes using voom+limma; 2) Taking the 25 genes whose expression are most correlated with the outcomes. The classifiers are trained to classify cytogenetic risk and the three large-scale copy number alterations (del_5, del_7, trisomy_8).

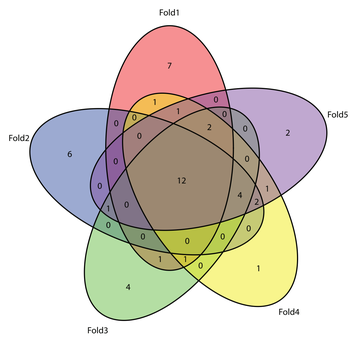


Figure 3. Venn diagram depicting the 25 most differentially expressed genes (voom+limma) across 5-folds to predict the del_5 CNA based on a random forest classifier.

RESULTS: Linear modelling

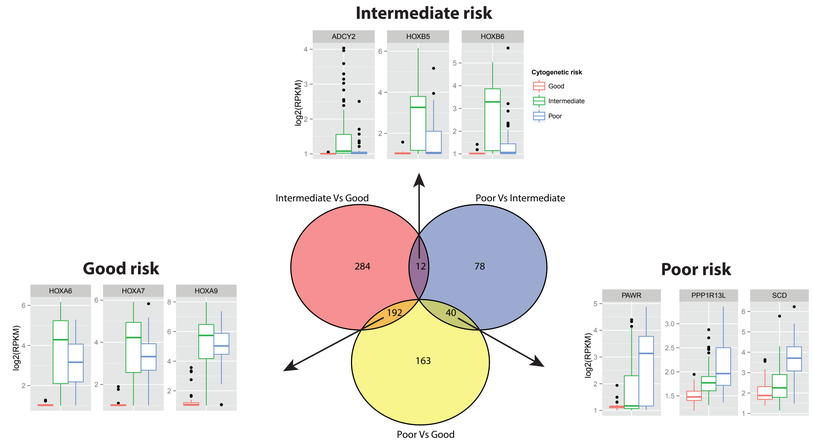


Figure 4: Differentially expressed genes between cytogenetic risk groups. Expression for 3 representative hits from each category are shown.

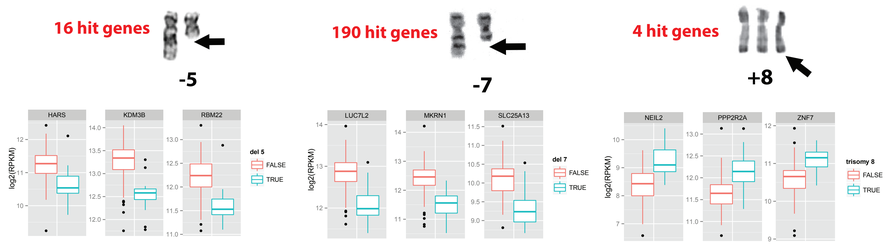


Figure 5: Differentially expressed genes between three key CNA events. Expression for 3 representative hits from each category are shown.

CONCLUSIONS

Differential Expression:
- In the risk-specific differential expression analysis, we identified many genes which exhibited expression patterns that were specific to a single cytogenetic risk status
- Additionally, in the CNA-specific differential expression model, many of the observed hits corresponded to the genomic region of the CNA of interest (*e.g.*, over-expressed genes on chromosome 8 in +8 samples)

Machine Learning:
- Linear model-based feature selection outperformed the correlation-based feature selection, as assessed by the downstream classifier performance
- The classifiers trained to predict 'intermediate' and 'good' cytogenetic risk showed high sensitivity and specificity
- The observed high performance of the 'good' classifiers is expected; this class makes up the most homogeneous set of cases

BIBLIOGRAPHY

1. National Comprehensive Cancer Network (2014). *NCCN Clinical Practice Guidelines in Oncology - Acute Myeloid Leukemia.* Version 2.
2. Grimwade, D *et al.* (2010). Refinement of cytogenetic classification in acute myeloid leukemia: determination of prognostic significance of rare recurring chromosomal abnormalities among 5876 younger adult patients treated in the United Kingdom Medical Research Council trials. *Blood* Jul 22;116(3):354-65.
3. Prebet, T *et al.* (2004). Secondary Philadelphia chromosome after non-myeloablative peripheral blood stem cell transplantation for a myelodysplastic syndrome in transformation. Bone Marrow Transplantation. Jan;33(2):247-9.
4. The Cancer Genome Atlas Research Network (2013). Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *New England Journal of Medicine*. May 30;368(22):2059-74.
5. Law, CW *et al.* (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*. Feb 3;15(2):R29.

CENTRE FOR CLINICAL GENOMICS