Submission type:
Oral communication or Poster (Choose one): **POSTER**

Session(Choose one of the topics):
- Translational, Clinical, and Industry Bioinformatics

Title: Characterization and predictive role of human-specific genes in Acute Lymphoblastic Leukemia

All Authors :
Thomas Sirchi [1]*, Andrea Tonina [1]*, Gloria Lugoboni [1]*, Lorenzo Santarelli [1]*, Matteo Gianesello [1], Federica Ress [1], Emma Busarello [1], Valter Cavecchia [3], Luca Tiberi [1], Toma Tebaldi [1], Enrico Blanzieri [2]

Affiliation:
[1]Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento, via Sommarive n. 9, 38123 Povo, Trento, Italy.
[2]Department of Information Engineering and Computer Science, University of Trento, Via Sommarive 9, 38123 Povo, Trento, Italy
[3] Institute of Materials for Electronics and Magnetism (CNR), via alla Cascata 56/c, 38123, Povo, TN, Italy
* These authors contributed equally

SECTIONS

Motivation: (1136)

Humans and chimpanzees separated approximately 6 million years ago. Since then, rapid evolution and genetic alterations have occurred, resulting in significant differences at the genetic level. The impact of these genetic changes can be observed across various aspects, including diet, immune function, and anatomy. We refer to these features as 'human-specific', corresponding to unique human genes. However, the precise role of these genes remains incompletely understood. Numerous studies have aimed to identify and expand the list of known human-specific genes, with the ultimate goal of understanding their association with human-specific diseases.

Acute Lymphoid Leukemia (ALL) is one of the most common leukemias in children, with 80% of pediatric cases. ALL is a malignant transformation that causes an abnormal proliferation and differentiation of lymphoid progenitor cells and correlates with genetic aberrations and complex events such as rearrangements of multiple chromosomes.

This study uses quantitative methodologies to extend our knowledge of human-specific genes associated with Acute Lymphoblastic Leukemia.

Methods: (1902)
The study is based on 8 databases retrieved from the Gene Expression Omnibus (GEO), composed of expression profiles of mRNAs extracted from bone marrow and blood cells in 794 patients, divided between tumoral ALL samples and controls. From associated metadata, we were able to distinguish patients based on their age (adult and pediatric) and tumor subtype (B, T, PreB and PreT).

Batch correction was performed on count matrices using *Combat-Seq* from the *sva* package in R and sequentially, data was normalized using the *TMM* method.
Differential Gene expression analysis was performed using the R package *EdgeR* Differentially expressed genes (DEGs) were filtered using a p-value < 0.01, highlighting up and down-regulated genes by setting a threshold of ±1.5 on log-transformed fold changes. From the list of DEGs, the human-specific genes were extracted using a reference list derived from the literature.

Enrichment and pathway analysis were performed on DEGs using the packages *clusterProfiler* and *EnrichR* in R. The enrichment was performed using *enrichGO* and focusing on the Biological Process (BP) sub-ontology. The functional enrichment analysis on WikiPathway was performed using *enrichWP*.These analyses were performed on genes that were found to be differentially expressed in tumors compared to controls, in pediatric samples compared to adult samples, and finally, in the various subtypes.

Using the R libraries *tidyverse* and *tidymodels*, we implemented various methods of classification trained on differentially expressed human-specific genes, in detail: an ensemble method, a gradient boosting method and CPU-based deep learning. This section aimed to identify human-specific genes with a predictive value that could serve as markers for tumor subtyping. From the initial data, 24 tumor samples were not associated with any subtype, and we used the classifiers to assign them a possible subtype class. To statistically validate the effectiveness of these genes, the models were trained with a three-fold cross-validation (CV) with two repeats with the ADASYN algorithm for a balanced sampling. The hyperparameters were tuned using the library *finetune.* They were initiated from a Latin hypercube sampling, and the best result was fed to a Simulated Annealing for 25 search iterations. After the classification, we proceeded with extracting the variable importance from our human-specific models and ranked the features.

Results: (1064)

We were able to investigate Acute Lymphoid Leukemia (ALL) in the context of human-specific genes. In particular, among the retrieved differentially expressed genes, it was possible to identify human-specific genes characterizing the tumor and the various subtypes (B, T, PreB and PreT). Enrichment analysis suggests that the human-specific genes are involved in deregulation of immune response, differentiation and splicing, all relevant processes associated with cancer.
Through the differentially expressed human-specific genes we were able to create a consensus classifier capable of associating unknown data to specific tumor subtypes.
The classifier showed good results, particularly an accuracy of over 84%, an F mean score of around 0.74 and a Kappa score of around 0.75. Among human-specific genes with higher importance for our models, we found EBF1, a cancer-related gene connected to the signal

transduction in leukemia, MYO7B, a known proto-oncogenic driver and RAB6C, a member of the RAS oncogene family.

In conclusion, our study highlighted a set of human-specific genes associated with ALL which can be used to characterize patients based on age and subtype. We created a classifier able to define the correct subtype of unknown samples with overall good accuracy and identified significant human-specific genes that can be predictors of ALL tumor subtypes.

Availability:

Supplementary information: the GitHub repository can be found at the following link: Acute_Lymphoid_Leukemia_Project