Proposal Report

Human-specific genes

Andrea Tonina, Thomas Sirchi,
Lorenzo Santarelli, Sabri Kaci, Gloria Lugoboni

October 2023

Contents

1 Introduction

1.1 Background

1.1.1 Human-specific genes

The evolutionary separation between humans and chimpanzees happened circa 6 million years ago. From this point over, we evolved rapidly and new alterations have been acquired. Main differences are found at the genetic level, via a series of aberrations such as rearrangements, duplications, and losses, resulting in orthologous genes and also de-novo ones. The role of these aberrations can be seen at different levels, starting from the diet and immune changes to anatomy (brain and neuroanatomy comparisons, bones, etc). We talk of "human-specific" features, therefore human-specific genes [1].

Thanks to an in-depth analysis it was possible to associate and group these human-specific genes to a restricted set of functions. Specifically, neural functions, metabolic functions (carbohydrate metabolism, adipogenesis pathway, glutamate biosynthesis, etc), immunological functions (parasitism, host-response, phagocytosis, etc), and structural growth and functions at the cell level (cytoskeleton organization, motility, transport, protein modifications, and targeting) were the main terms identified.

These results were obtained via a pipeline that involved the use of:

- GeneTerm Linker (FGNet), an algorithm used to identify sets of associated genes via a process of clustering [2];
- Gene Ontology (GO) Analysis, to perform enrichment analysis on gene sets and identify over-represented (or under-represented) genes [3];
- Ingenuity Pathway Analysis (IPA), a web-based application that allows functional analysis of genes to identify networks or correlations [4];

Human-specific genes are still to be completely uncovered. Correlations between these genes and diseases have been made, identifying so-called human-specific diseases [5].

1.1.2 Pediatric Acute Lymphoid Leukemia (ALL)

In this research project, we decided to focus on Acute Lymphoid Leukemia (ALL) in children. ALL is in fact, one of the most common leukemia in children (80% of patients are children). It is a malignant transformation that provokes an abnormal proliferation and differentiation of lymphoid progenitor cells, attacking mainly B-cells. It correlates with chromosomal aberrations and complex events such as rearrangement of multiple chromosomes. In children, it is often observed associated with other syndromes that have a genetic predisposition, such as Down syndrome, Bloom syndrome, and Fanconi anemia. Different main rearrangement events have been detected and are considered hallmarks of ALL, one example is the translocation BCR-ABL1. The symptoms are usually non-specific and include examples of bone marrow failure. In some cases at the

moment of the diagnosis, involvement of the central nervous system is observed, with cranial nerve deficit or meningismus [6].

1.2 Statement of need

Over the past years, human-specific genes have received increasing attention as potential major contributors responsible for different human-specific diseases. Still, it is a world yet to be discovered and further analyzed. Because of the complex topic and the high connection with the evolution process, it is a complex field to uncover and study.
We believe that thanks to the proposed pipeline, in which human-specific genes are compared to hallmarks of ALL, we will possibly be able to obtain connections between the two and identify if possible, some main drivers of the disease that are linked to human-specific gene pathways.

1.3 Objectives

The main objective of this project is to uncover the possible correlations between human-specific genes and ALL, finding hallmarks of the disease in a way to be able to estimate a treatment sensibility based on the genetic profile.

2 Materials and Methods

2.1 Databases

2.1.1 GEO

The Gene Expression Omnibus (GEO) database is a public resource containing high-throughput gene expression and other functional genomics data sets [7].
It was founded in 2000, rapidly evolving to contain multiple datasets connected to genome methylation data, chromatin structure, genome–protein interactions, whole-genome sequencing or RNA-sequencing.

2.1.2 cBioPortal

cBioPortal is a database containing cancer genomics data set linked to patients and clinical applications. This database enables an easy and direct viability of raw data to the entire cancer research community [8].

2.1.3 PubMed

PubMed is a database containing citations and abstracts of biomedical papers/literature. It is a searching tool that enables a fast and easy retrieving of biomedical and life sciences literature, improving research and viability of information [9].

2.2 Tools

2.2.1 Gene Ontology

A gene ontology is a way to capture biological knowledge for individual gene products in a written and computable form. It has a formal structure and can be defined as a set of concepts and their relationships, arranged in a hierarchy, from a less specific description to a more specific description. In a gene ontology, each component is associated with a specific notion [3].
Three main hierarchies can be defined:

- Molecular Function. This category describes activities that happen at a molecular level, it includes the activities that are involved in an action and do not specify where, when, or in which context the actions happen
- Biological Process. A biological process is a series of events resulting from multiple ordered groups of molecular functions. It can be thought of as a chain of execution.
- Cellular Component. A cellular component is linked to a component of a cell with the condition that is part of a larger object and can be part of an anatomic structure.

2.2.2 EnrichR

Enrichr is an enrichment analysis web-based tool. It is a resource that contains curated gene sets and it can be used as a search engine to visualize and rank enriched terms [10].

2.2.3 Combat and Combat-Seq

In microarray experiment it is often possible to observe technical and non-technical biases called "batch effects". Combat is a robust tool that exploit an empirical Bayes framework to perform a correction of the batch effects in the data, recalibrating them [11].
Combat-Seq is a tool that exploits a negative binomial regression model to remove batch effects from RNA sequencing data. Combat-Seq is also implemented in a way to reduce the number of false positives in differential expression and recover the biological signal in the data. [12].

2.2.4 Differential Gene Expression

A gene expression profile of a cell is the snapshot of which genes are expressed in that cell at the time the sample was taken. Knowing which genes are expressed in a cell at a certain moment allows the identification of new genes or transcripts and the comparison of expression profiles between samples (a typical scenario we are interested in). Gene expression profiles are extremely heterogeneous since they vary based on the individual, tissue, condition, and cells of origin. During a differential gene expression experiment the expression profile of genes is

compared between samples. Comparisons are usually effectuated over different disease states or differences between healthy and diseased individuals [13].
Two main technologies to obtain expression data can be distinguished:

- Micro-Arrays: Microarrays have been introduced at the beginning of the 2000s and were the first high-throughput technology. Microarrays are glass surface matrices on solid supports made of thousands of spots, each containing multiple and identical DNA probes. The probes target different genes or transcripts [14].
- RNA-Sequencing: RNA sequencing is a next-generation sequencing approach that sequences the cDNA from the mRNA component. A whole variety of sequencing machines exist. Based on the biological question a specific machine needs to be used. These machines are distinguished on the throughput, the read length, and the coverage [15].

2.2.5 Principal Component Analysis

Principal component analysis (PCA) is a technique used to reduce the dimensionality of large datasets thanks to the identification of so-defined "principal components". This technique is used to increase interpretability and at the same time minimize information loss. Principal components can be seen as variables that maximize the variance among the data, maximizing the corresponding information brought [16].

2.2.6 OneGenE and FANTOM5

OneGenE [17] is an evolution of the NES2RA algorithm [18], based on the PC-algorithm able to expand Local Gene Network (LGN) using transcriptomic data. Thanks to the gene@home project, a computation project that relies on the use of volunteers' computers to make the computation faster [19]. The term gene network expansion is used to define a process that aims at finding new genes to expand a given known gene network.
FANTOM5 (functional annotation of the mammalian genome 5) is an atlas containing expression profiles and functional annotation of mammalian transcriptomes [20].

2.3 Datasets

2.3.1 Expression data matrices

Expression data matrices can be obtained from two main experiments as explained above. We can divide our dataset into two categories, based on the technology used to gather the information:

Micro-Array			
Dataset Names	Number of samples	PMID	Data types
GSE2604	36	15996926	T
GSE26495	16	21383243	C
GSE27131	7	21781987	C
GSE34465	9	23226513	C
GSE11877	207	20699438	T

RNA-Sequencing			
Dataset Names	Number of samples	PMID	Data types
GSE163634	65	34362951	C-T
GSE133499	42	32204435	T
all_phase2_target	1978	25207766	T
all_stjude_2016	73	27776115	T
all_stjude_2015	93	25730765	T

In the above table, the label C is used to indicate controls while the label T is used to indicate tumor samples. In the cases in which the dataset contains both tumor and control samples, we used the identifier C-T

3 Pipeline

After the retrieval and the merging of the datasets, containing expression data for ALL patients, a normalization process is needed. Based on the technology used to obtain the expression data, a different normalization process is performed, specifically, an RMA normalization for Affymetrix data and a process of intersample/intrasample normalization for RNA sequencing data. Plus, a batch effect correction will be performed using ComBat and ComBat-Seq tools. This process is fundamental to obtain comparable data and not create biases for future analyses.
The pipeline is then followed by a Principal Component Analysis (PCA), used to cluster the samples into different subtypes based on the genetic profile. Based on the retrieved information we will apply a differential gene expression (DGE) analysis both on *Control vs Tumor* samples and on the *different subtypes* of the samples. The idea is then to operate a comparison between the obtained data with the already-known datasets of human-specific genes, in a way to have an idea of the quantity of the human-specific genes characterizing Acute Lymphoid Leukemia. Depending on the results of this procedure, we will choose the best path to proceed with our analysis. We have a branch:

- If the number of human-specific genes is too low, our data will undergo a process of gene expansion using FANTOM5, thanks to the creation of causal relationships. From this expanded network we will make another comparison with the human-specific gene dataset, hoping to obtain more related genes

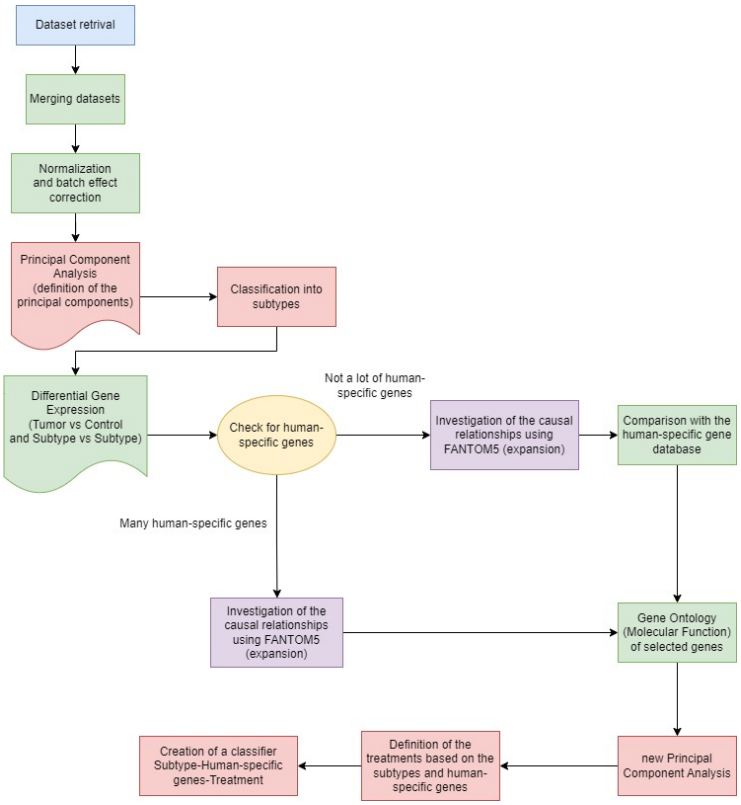

Figure 1: Workflow for the project

- If the number of human-specific genes is large enough, we will proceed with a process of gene expansion, only for the filtered human-specific genes associated with ALL.

In both cases, the data will undergo an enrichment analysis of gene ontology (GO). Thanks to this procedure it will be possible to annotate the data to identify the terms (molecular functions or the biological processes) that are over-represented (or under-represented) in our genes.
After that, we are going to operate another PCA, hoping to have a new insight into the data, after the process of expansion and annotation.
similar analysis using a different approach that is based on an unsupervised method, PCA (Principal Component Analysis), identifying clusters of patients characterized by key features.
The following step in the pipeline is to use the EnrichR tool with the aim of creating a "DrugMatrix" based on the analyzed data, identifying the most used and efficient drugs and therapies.
The final step to the pipeline would be to create a classifier for the definition of the identified subtypes and link each of these to a specific drug or treatment.

The initial plan is, if possible, to use the scikit-learn library in Python [21], to create a simple machine-learning model.

4 Expected outcomes

We hope that we will be able to obtain further insight into the world of human-specific genes and human-specific diseases, finding novel genes that could correlate to this topic. We also hope to be able to define hallmarks and subtypes of ALL, creating a connection between the genetic expression profile and possible treatments.
We have great expectations for this project, if the initial analysis does not bring us the hoped results we are thinking of moving to a different pediatric disease of the blood, like Agammaglobulinemia.

Bibliography

1. Et al., B. M. Genes with human-specific features are primarily involved with brain, immune and metabolic evolution. *BMC Bioinformatics* (2019).
2. Aibar, S., Fontanillo, C., Droste, C. & De Las Rivas, J. Functional Gene Networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering. *Bioinformatics* **31,** 1686–1688. `https://doi.org/10.1093/bioinformatics/btu864` (Jan. 2015).
3. Et al., A. M. Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium. Nat Genet.* (2000).
4. Et al.., K. A. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* (2014).
5. Et al., B. The influence of evolutionary history on human health and disease. *Nat Rev Genet 22* (2021).
6. Inaba H, M. C. Pediatric acute lymphoblastic leukemia. *Haematologica.* (2020).
7. E, C. & T., B. The Gene Expression Omnibus Database. *Methods Mol Biol.* (2016).
8. Cerami E Gao J Dogrusoz U, e. a. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* (2012).
9. NCBI. *PubMed* `https://pubmed.ncbi.nlm.nih.gov/about/`.
10. Et al., C. E. Y. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC bioinformatics vol. 14 128* (2013).
11. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.*

12. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics.*

13. Rodriguez-Esteban R & Jiang, X. Differential gene expression in disease: a comparison between high-throughput studies and the literature. *BMC Med Genomics* (2017).

14. R., B. Overview of DNA microarrays: types, applications, and their future. *Curr Protoc Mol Biol.* (2013).

15. Et al., C. A. A survey of best practices for RNA-seq data analysis. *Genome Biol 17* (2016).

16. T., J. I. & Jorge, C. Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc* (2016).

17. Asnicar, F. *et al. OneGenE: Regulatory Gene Network Expansion via Distributed Volunteer Computing on BOINC* in *2019 27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)* (2019), 315–322.

18. al., P. S. NES2RA: a tool for grapevine transcriptomic data mining. *The First Annual Meeting of COST Action CA17111 INTEGRAPE 2019 - Data Integration as a key step for future grapevine research, Chania, Crete Greece.* `http://hdl.handle.net/10449/54350` (2019).

19. Kalisch, M. & Bühlmann, P. Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *Journal of Machine Learning Research.* `http://jmlr.org/papers/v8/kalisch07a.html` (2007).

20. Consortium, T. F., the RIKEN PMI & (DGT)., C. A promoter-level mammalian expression atlas. *Nature 507* (2014).

21. Python. *Scikit-learn library* `https://scikit-learn.org/stable/`.