**Isolating Somatic Variants - Computational Methods of Somatic Variant Identification using Single Cell RNA Sequencing Data**

Elżbieta Wierciak 1, Kees Van Bergen 2, Szymon Kiełbasa 2, Marek Kisiel-Dorohinicki 1

1*AGH University of Science and Technology*, al. Adama Mickiewicza 30, 30-059 Kraków, Poland

2*Leiden University Medical Centre*, Albinusdreef 2, 2333 ZG Leiden, Netherlands

{*e.wierciak*}@*gmail.com***Keywords**: CLL, chronic lymphocytic leukemia, MBL, monoclonal B-cell lymphocytosis, scRNA-seq, single-cell RNA sequencing, computational medicine, bioinformatics

1. **Introduction**

The development of data and computer science opens exciting avenues for personalized medicine and disease prevention. An example of a disease that could be better understood with the support of computational methods is chronic lymphocytic leukemia (CLL) – the most commonly diagnosed blood cancer among European adults [1]. The exact cause of CLL is still unknown, but attempts to understand its genesis on a genetic level, as well as its connection to monoclonal B-cell lymphocytosis (MBL), have shown promise [2].

1. **Description of the problem**

Chronic lymphocytic leukemia affects B cells, causing them to multiply excessively and accumulate in the blood, bone marrow and lymphoid tissues, thus interfering with normal blood cell production and immune function [1]. Monoclonal B-cell lymphocytosis is a precursor condition to CLL characterized by the presence of small numbers of clonal B cells in the peripheral blood, but without the symptoms seen in full-blown leukemia [2]. In order to understand early CLL development, it is important to find out whether MBL drives the premalignant expansion or the malignant progression of CLL. To answer this question, genetic data from healthy subjects with MBL was sequenced using the Illumina scRNA-seq method. For each subject, each cell was tagged according to the clone it belonged to. The aim was to compare the genetic sequence of expanded clones against cells that do not belong to expanded clones. In order to avoid detecting genetic changes that are the result of inter-individual variability, each studied expanded clone and the reference non-expanded set were both sourced from the same individual.

1. **Related work**

A series of research that sparked the idea for this project was focusing on autonomous B-cell receptor signaling (BCR signaling) [3][4]. In this work it was shown that autonomous BCR signaling operates in MBL analogously to CLL. Subclonal genetic CLL driver mutations in MBL have also been observed [5], which would support the idea of gradual clonal expansion driven by moderate autonomous BCR signaling, potentially resulting in a level of genetic instability that facilitates graduate acquisition of CLL driver mutations.

1. **Solution to the problem**

Thanks to Cyfronet HPC resources and Leiden University Medical Centre, the scRNA-seq data was made available for analysis. Several methods of somatic variant detection were proposed and compared, then the best one selected. It involved creating a germline and expanded set count table, with the information for each position how many of each nucleotide were detected for both sets. Next a statistical test was carried out to determine what zygosity type each position in each set had. It was tested whether the observed alternative reads can be explained by sequencing error (a fixed 5% error rate was assumed) or by a somatic variant, using a binomial likelihood. The GQ score was then used to measure the confidence with which the data supports the somatic variant hypothesis [5][6].



**Fig.1.** Diagram of the evaluation procedure

1. **Conclusions and future work**

Several interesting somatic variant detection approaches were proposed for this problem based on previous similar work. Most genetic changes detected were losses of heterozygosity, and several SNPs were also detected. Insertions and deletions were beyond the scope of this work. Consultations with an immunologist are in progress to determine the impact of each of the detected somatic variants.

**Acknowledgements.** I am thankful for help and support from Leiden University Medical Center and their representatives, dr. Cornelis Van Bergen and dr. Szymon Kiełbasa, for the time and effort they spent sharing their expertise with me. I would also like to extend my sincerest thanks to my thesis supervisor prof. Marek Kisiel-Dorohinicki and my research supervisor Katarzyna Jurkowska for their support in this project. In addition, I am grateful to Polish high-performance computing infrastructure PLGrid for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium through grant plgaghlumc2025.

**References**

1. Julio Delgado Ferran Nadeu D.C., Campo E.: *Chronic lymphocytic leukemia: from molecular pathogenesis to novel therapeutic strategies*. In: Haematologica, vol. 105(9), pp. 2205–2217, 2020. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7556519/pdf/1052205.pdf.
2. Rawstron A.C., Bennett F.L., O’Connor S.J.M., Kwok M., Fenton J.A.L., Plummer M., de Tute R.M., Owen R.G., Richards S.J., Jack A.S., Hillmen P.: Monoclonal B-cell lymphocytosis and chronic lymphocytic leukemia. In: New England Journal of Medicine, vol. 359(6), pp. 575–583, 2008. URL http://dx.doi.org/10.1056/NEJMoa075290.
3. Quinten E, Sepúlveda-Yáñez JH, Koning MT, Eken JA, Pfeifer D, Nteleah V, de Groen RAL, Saravia DA, Knijnenburg J, Stuivenberg-Bleijswijk HE, Pantic M, Agathangelidis A, Keppler-Hafkemeyer A, van Bergen CAM, Uribe-Paredes R, Stamatopoulos K, Vermaat JS, Zirlik K, Navarrete MA, Jumaa H, Veelken H. Autonomous B-cell receptor signaling and genetic aberrations in chronic lymphocytic leukemia-phenotype monoclonal B lymphocytosis in siblings of patients with chronic lymphocytic leukemia. *Haematologica.* 2024;109(3):824–834. doi:10.3324/haematol.2022.282542.
4. Burger J.A., Chiorazzi N.: B cell receptor signaling in chronic lymphocytic leukemia. In: Trends in Immunology, vol. 34(12), pp. 592–601, 2013. URL http://dx.doi.org/10.1016/j.it.2013.06.006.
5. Li, H. (2011)**.** *A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.* Bioinformatics, 27(21), 2987–2993.
6. DePristo, M. A., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics, 43(5), 491–498.