

To the Senate of Medical University of Bialystok

### **Pre-examination statement of the PhD thesis by Bence Gálik**

It is a pleasure for me to write this pre-examination report on the PhD thesis titled “Development and implementation of original bioinformatics pipelines for NGS data analysis in cancer and assisted reproduction research” by Bence Gálik. This thesis represents a timely and important endeavour. Currently, the generation of large-scale and complex biological omics datasets is fast and cost-effective, making the availability of efficient and robust data analysis a critical limiting factor in the modern data driven research projects. The PhD thesis by Bence Gálik responds to this existing limitation very well by establishing data analysis pipelines allowing the robust and efficient analysis of data in cancer and assisted reproduction research related projects.

Four of the cases studies included in this thesis have gone through peer-review process and been published in scientific journals, while one of the studies is independently included in the thesis. Bence Gálik is the shared first author in two, second author in one and fourth author in one of the published studies. His contribution to these four publications and the independent study is clearly invaluable, thereby making his contribution to the dissertation sufficient for a doctoral thesis.

The first case study concentrates on chronic lymphocytic leukemia (CLL) which is characterized by substantial clinical and genetic heterogeneity. Here a targeted analysis of 30 recurrently mutated genes on 20 patients, before and during ibrutinib treatment. Characterization of ibrutinib resistance mechanisms can potentially lead to novel clinical advances. The data analysis of the generated custom amplicon sequencing data was performed by Bence Gálik using an in-house generated data analysis workflow consisting of widely used tools for the different analysis steps.

The second study aimed at improving the molecular classification of primary central nervous system lymphoma patients, the results of which could potentially lead to more precise patient stratification in diagnostic use. A gene expression based NanoString assay was compared here with the standard immunohistochemical algorithm. The same bioinformatics workflow developed for study I was used to analyze the data in addition to which further statistical analysis was performed.

In the third case study glioblastoma samples were studied in order to improve the understanding of the molecular mechanisms related to the development of the disease. This was carried out by analyzing DNA CpG methylation patterns in the samples using the reduced-representation bisulfite sequencing method. The analysis of the sequencing data were carried out using bioinformatics workflow developed by Bence Gálik. The workflow

consists of tools that are widely used for similar purposes. In addition, further statistical testing was performed on the data.

Fourth study focuses on non-small cell lung cancer that represents a heterogeneous group of malignancies comprised mainly of adenocarcinomas and squamous cell carcinomas. Here, miRNA-sequencing data were generated in order to investigate how the miRNAs could help to improve the characterization of disease subtypes and be potentially used as biomarkers for targeted therapies. The sequencing data were analyzed by the bioinformatics workflow developed by Bence Gálik in addition to which extensive statistical analysis and predictive modelling were conducted.

The last study deals with reproduction research, where the goal was to develop an approach for non-invasive evaluation of embryonic DNA content that could be used before implantation in in vitro fertilization, supplementary to existing methods. DNA-sequencing was performed and the resulting data were analysed to detect copy number variations using the bioinformatics workflow developed by Bence Gálik. The data analysis steps relied on widely used bioinformatics tools.

The dissertation comprehensively presents the background of each study, as well related wet-lab and dry-lab materials and methods. Analysis results of each study are also described in a detailed and comprehensive manner. Bioinformatics and NGS applications in general are described, somewhat concisely, in the introductory part of the dissertation. Given that the title of the thesis highlights the bioinformatics-oriented contribution of the PhD candidate in the conducted studies, it would have been interesting to see a wider presentation of the bioinformatics data analysis in general, and related tools and practices for building data analysis workflows. In addition, the technical aspects of the developed workflows could have been described in greater detail including how the run parameters for the different tools were selected when deviating from the default values.

The clinical aspects of the different case studies included in the thesis are comprehensively discussed at the end of the thesis, illustrating the solid understanding of the conducted research by the PhD candidate. However, the discussion section could have profited from a deeper consideration of the relevance of generating bioinformatics workflows within research projects in general, and also on the potential future prospect on the reuse of such workflows in later studies. The thesis would also benefit from further language editing to improve the style and spelling.

There is no doubt that the thesis of Bence Gálik has generated new knowledge within the fields of cancer and reproduction research by enabling the data analysis of several real life projects using the bioinformatics workflows established and data analysis approaches applied during the thesis work. The submitted thesis complies with the requirements of the doctoral regulations of the Medical University of Białystok to obtain the academic doctoral degree. **Thus, I recommend that this dissertation moves to further stages of the doctorate process and may be subjected to a public defense.**

Sincerely,  
Dr. Asta Laiho

