8 Summary

The thesis "The elucidation of the causes of neurogenetic diseases by the MPS data analysis using advanced algorithms" is focused on processing the massively parallel sequencing (MPS) data from a gene panel, whole-exome sequencing (WES) and whole-genome sequencing (WGS). The aim of the study was to develop a suitable pipeline to evaluate at least 250 MPS gene panel data, 150 WES data and 20 WGS data in order to improve molecular genetic testing of rare neurogenetic disorders. Associated data management and database implementation is also described.

Targeted gene panel sequencing A custom-designed gene panel consisting of genes previously associated with the disease was used. In the Epileptic Encephalopathy (EE) panel, two prerequisites need to be met for inclusion into the panel: the gene has to have been published in at least two independent publications OR at least in one publication but in multiple independent families. In the case of the EE panel, 112 genes were included. The targeted gene panel sequencing was then performed on 257 patients with EE. Pathogenic or likely pathogenic (according to ACMG criteria) variants have been found in 28% of patients (72 out of 257). Further analysis of the pathogenic or likely pathogenic variants was performed (76 in total); the variants were grouped by gene and then the genes were grouped by inheritance and origin of the variants. Out of 112 genes included in the custom gene panel, pathogenic or likely pathogenic variants were found in 33 genes. According to the segregation analysis, we were able to determine the origin of the variant in 68 patients out of 72. De novo origin was confirmed in 70.3% of variants. A relationship between the age of onset of the epileptic seizures and clarification rate was also demonstrated; the number of solved cases is almost two times higher in patients with onset of seizures in the first four weeks of life [Staněk et al. 2018]. In addition, two co-authored publications presenting case reports of patients with EE are included [Neupauerová et al. 2017] and [Štěrbová et al. 2018].

Whole exome sequencing (WES) For WES data analysis, three different bioinformatic pipelines were compared (GATK, SureCall, NextGene). Twenty-four WES samples of EE patients were used as a test dataset. It was shown that the optimal results were obtained with both GATK and SureCall. Therefore, these two algorithms were chosen to be the method of choice in the DNA lab.

Two main approaches (de novo and singleton model) were used for variant evaluation.

Based on the analysis of WES samples in Trio (proband and both parents), several different tools were tested for the de novo model. The DeNovoGear tool produced the best results for de novo variant detection and cases solved by using the de novo model have been reported, [Sedláčková et al. 2018] and [Neupauerová et al. 2018].

The second option for WES variant evaluation is a singleton model. This analysis is based on manual filtering followed by a genotype-phenotype association tool Exomiser (using HPO terms describing the phenotype). The singleton model was used to identify variants in the SBF2 gene – presented in the co-authored publication [Laššuthová et al. 2018]. We also describe some other cases solved by using proposed methods.

At the end of WES chapter, the CNV analysis workflow is presented. At first, we performed the analysis of available and suitable tools and finally we managed to introduce an optimized methodology for germline CNV in WES data by using GATK 4 beta pipeline. We tested the methodology on two previously confirmed CNV cases and with the GATK4 tool and the custom virtual panel tool (implemented in the DNA laboratory) we were able to detect both of the CNV in the samples.

Bioinformatic databases The first database is a variant database of 222 WES samples (all samples in our workplace). By using GATK, we detected 300,111 variants in 17,512 genes. For these variants we calculated their allele frequency in our subpopulation, listed their type and performed gene analysis. The variants were divided into five classes according to their allele frequency in our database and compared to gnomAD allele frequency. Further analysis of variants, that are more common in our subpopulation rather than in the gnomAD, was performed.

The protein domain database is available at url: *www.prot2hg.com*. Here we introduce a resource that addresses the question of whether a particular variant falls onto an annotated protein domain. When applied to patient genetic data, we found that rare (<1%) variants in the gnomAD were significantly more annotated onto a protein domain in comparison to common (>1%) variants. Similarly, variants described as pathogenic or likely pathogenic in ClinVar were more likely to be annotated onto a domain. In addition, we tested a dataset consisting of 60 causal variants in a cohort of patients with epileptic encephalopathy, and found that 71% of them (43 variants) were propagated onto protein domains (presented as a poster at the ASHG conference 2018; poster is shown in the supplementary section; manuscript has been submitted).

The Inherited Neuropathy Variant Browser is a database of CMT variants developed in an international collaboration with the Hussman Institute for Human Genomics in Miami (USA). It is a community-driven database of CMT variants where users can add, evaluate and share variants. The results of this project were published in a co-authored publication [Saghira et al. 2018]. My role in the project was to help to design the database structure, to load data into the database and to implement the component that projects protein domain location onto the CMT genes.

Data management The final aim of the thesis was to design a sustainable data management for our DNA laboratory. Therefore, we designed a simple and clear system allowing the long-term storage of data in an easily traceable form (precisely defined tree structure). Data is backed up on the NAS devices and other remote servers, and, with the defined methodology, we minimized redundancy of data and maximized data safety.