Chapter 1

Introduction

1.1 Background

According to the WHO, obesity in an Asian adult is defined as having a body mass index (BMI) value of greater than or equal to 25 (WHO Expert Consultation, 2004). In Indonesia, the prevalence of obesity in adults (>18 years old) had reached 23.4% in 2023 (BKPK, 2023). Obesity has become one of the most focused diseases by the WHO not only because of the high prevalence, but also due to its effect towards raising the potential of other diseases. Even though obesity is categorized by the BMI, there are multiple factors that can affect the obesity potential of someone, including genes (Chooi et al., 2019). One such method to predict obesity is to use polygenic risk scores (PRS). PRS is an analysis which uses genotype and phenotype data to create an association between specific gene alleles and traits, such as risk of getting a certain disease (Lambert et al., 2019; Lewis & Vassos, 2020). PRS can then be used in personalized medicine as a guidance towards possible treatment, prevention, or diagnosis confirmation (Lewis & Vassos, 2020).

To calculate a PRS and draw the conclusion from its results, multiple steps needed to be done, such as input data quality control, ancestry adjustment, and the PRS calculation itself (ICDA PRS Task Force, 2021). As this research was done at a healthcare company in Indonesia, a previous workflow to calculate PRS was already made, with each step involving multiple tools and programming languages, as the analysis involved multiple individuals. The downside of this previous approach was not only the scattered data locations (although all data is stored in the cloud server), but also the possible confusion it could cause for future users attempting to run the analyses. Meanwhile, rerunning the pipeline multiple times might be needed as the sample data grows, which can result in an increase of accuracy. Another possibility is that PRS itself can be used for many different traits, the possibility of re-running the PRS calculation multiple times for each of those traits is high (Lewis & Vassos, 2020). In the case of this research, a pipeline was needed to handle the growing number of data that needed to be analyzed to calculate the PRS of obesity in Indonesian patients.

The standardization of a pipeline was deemed important to ensure the reproducibility of an analysis and the accuracy of results (Leipzig, 2017). One such way is to use Nextflow, which is a script-based workflow engine to develop an automatic analysis. Nextflow is POSIX compatible and uses its own language, based on Groovy, to specify its input paths, output paths, and the script for the process. The script itself is able to be written in various other languages such as Bash, R, Python, etc., which makes it suitable to handle the multiple language problem of the company's previous workflow (Pohl et al., 2024). To handle the multiple tools, packages, and dependencies of the workflow, a Docker container was created which connects to the Nextflow pipeline (Spišaková et al., 2023).

1.2 Objective

This research aims to create a working Nextflow pipeline to calculate the PRS for obesity from Indonesian samples.

1.3 Hypothesis

It is hypothesized that the Nextflow pipeline created for PRS calculation can produce accurate results for multiple batches of data.