Abstract

The variant call format (VCF) is a file format used to represent and store information about DNA variation. Genetic variants in VCF can be represented in multiple ways because the VCF specification allows for ambiguity, which can arise because of different variant calling pipelines or differences in sequence alignment. Ambiguities interfere with the comparison of VCF files and the variants therein, leading to complications in further analysis of variants.

This thesis explores the differences in the representation of genetic variants that can occur, as well as their causes and impacts on further analysis. Furthermore, the normalization of VCF files is addressed and an algorithm for the atomization and deatomization of VCF files is shown.

Keywords: VCF, variant call format, ambiguous variant representation, variant comparison, variant atomization, variant deatomization