refGeneSearch: Using Python and MySQL to investigate the breakpoints of structural variants

There are a number of programs for annotation of genetic variations; with the ensembl variants effect predictor (VEP) being one of the most popular ones. VEP has a lot of advantages and produces good and fast results when predicting the effects of single nucleotide polymorphisms, single nucleotide variations and smaller deletions and duplications. However, when it comes to bigger structural variations such as inversions, fusions and translocations VEP does not offer much help. As an example I used a VCF file (example.vcf) with 20 inversions and ran it through VEP. The result was an output (example_vep_output.txt) containing more than 42 000 lines listing all consequences ensembl has listed inside the span of the inversion coordinates. These results do not give any real indication of how the inversion or other structural variants could affect the expression or function of the genes within or surrounding the breakpoints.

The goal was therefore to find a way to investigate if the breakpoints of the inversions overlaps with the exons of any known genes, and if so which these genes are and what exon is affected. At first I started to look at the source code for VEP to try to find a way of tweaking the existing code to get a more favorable result, but as I expected that is a big project that would require a substantial amount of time and effort. The second idea was therefore to try to find a way of using the breakpoints of the inversions and investigating if the breakpoints themselves are impacting the reading frames of any exons.

The first step was to find an easy way of handling the data from the variant data in the VCF. I found a python package called scikit-allel that gives you the ability to read VCF-data into a numpy array and further filter the data and convert into a more readable output in the form of a pandas data frame. I then used a package called MySQL Connector that gives your python program the ability to connect to MySQL databases. As an example I extracted the start and end coordinates of the inversions to query the refGene database to check if the breakpoints are found inside the boundaries of exons. The results of the searches are stored as pandas data frames.

The aim is to further develop the way of investigating the effect of larger structural variants such as inversions.

VEP

https://doi.org/10.1186/s13059-016-0974-4

scikit-allel

https://zenodo.org/badge/latestdoi/7890/cggh/scikit-allel

MySQL Conncetor

https://dev.mysql.com/doc/connector-python/en/