Dear candidate,

We’d like you to build the prototype of an annotation system that can handle SNV data.

Your inputs are the following:

* Links to a dbVar (https://www.ncbi.nlm.nih.gov/dbvar/content/ftp\_manifest): NCBI’s repository for pathogenicity flags associated to CNVs, reported by clinicians in the literature

Your system should incorporate the following elements:

* An SQL database that stores the following information:
  + A unique variant identifier
  + Genomic positions of the SNV variants seen in dbVar. Note that SNV variants are defined by start-end positions, but that varying levels of certainty can occur (hence, you’ll get “outer start”, “start” and “inner start” telling you whether we now precisely or not the exact boundaries of the SNV. Respectively, the “Outer/Inner” terms define the range within which the duplication boundary sits, the absence of that term states we know exactly the boundary location).
  + Copy number status (deletion, duplication or insertion)
  + Metadata: original variant identifier, data origin and pathogenicity assessment (if any)
  + The data model should be optimized for fast access via genomic ranges (e.g. query all variants that overlap with a segment defined by position X and Y)
* The scripts used for extracting, transforming and loading the relevant data into your database
* A script (or SQL functions) that allow searching for variants by genomic range (either exact matches or closest hit with matching criteria to be defined and justified) in the database. Those queries must convey the certainty / confidence we have in the returned matches. Please provide a few examples of queries.

For the sake of time and to keep this assignment within reasonable limits, you can limit your analysis to chromosome 1

Please return a structured report along with scripts and materials needed to reproduce your analysis.

Thank you for your time and efforts on this assignment, we wish you the best.