VarDetect

Anthony Ton, A16841070, a1ton@ucsd.edu
Maddie Ritter, A16907429, m1ritter@ucsd.edu
Ethan Xu, A16747623, ecxu@ucsd.edu

We intend to implement a tool called VarDetect that does variant detection, similar to Varscan which we used in Lab 1. We will be coding in python and using Github for version control. At a high level, we intend to implement the same workflow as VarScan. We will take in BAM files from our samples, then use the pileup file we generate to call variants. Our program will use statistical testing to determine if these are true variants or if they are simply sequencing errors, and the program will generate a variant call format (VCF) as output that contains all variants and their respective statistical tests.

We plan to benchmark VarDetect against varscan by comparing it to both the provided data from Lab 1 as well as publicly available data on the IGSR. By taking a well understood data set, we can be more confident in our analysis when comparing output with Varscan. To benchmark, we will compare the total number of variants found, the alternate allele at each variant, and the respective p-value associated with each variant.

We plan to apply our tool to one of the below datasets found on the 1000 Genomes Project website. These datasets are obtained from individuals from all over the world. Each dataset indicates the sex and the ethnicity of the individual, as well as which city and nation they live in. All three of the below datasets are also sequenced using the same technology that the datasets in Lab 1 were, ensuring that our benchmarking plan has as few differences as possible for accurate benchmarking.

Female Han Chinese from Beijing, China:
https://www.internationalgenome.org/data-portal/sample/NA18555
Male British in Britain and Scotland:
https://www.internationalgenome.org/data-portal/sample/HG00145
Male Punjabi in Lahore, Pakistan:
https://www.internationalgenome.org/data-portal/sample/HG02603