Title

Single nucleotide variant data in patients with thyroid cancer

Authors

Purvi Kanfade,[1] Amanda Bataycan,[2] and Ming-Ying Leung[1,2,3]

[1]Bioinformatics Program, [2]Computational Science Program, and [3]Department of Mathematical Sciences, The University of Texas at El Paso, El Paso, TX

Abstract

According to the American Cancer Society, over 40,000 new cases of thyroid cancer (TC) are diagnosed each year, causing over 2,000 deaths. Although certain specific DNA mutations have been associated with TC, the pathogenesis pathways are not fully understood. The goal of this project is to organize a set of genomic single nucleotide variant (SNV) data obtained from patients with (TC) for downstream bioinformatics analysis to help provide insights into the molecular mechanisms of the disease.

Individual variant call format (VCF) data files with SNV information from both tumor and normal tissue samples of 497 patients with TC were downloaded from The Cancer Genome Atlas (TCGA) resources. These files were converted to more readable CSV files and the SNV information within were subsequently merged into one single data file listing all the distinct SNVs. By integrating essential information from RefFlat gene annotations into our computational workflow, we further organized the SNVs by their genomic locations. While all 497 tumor samples provided useful SNV information, only 477 normal samples yielded reliable data after quality filtering. The resulting dataset contained 98,952 distinct SNVs, of which 95,627 occurred only in tumor samples, 7 only in normal samples, and 3325 in both. The SNVs in tumor samples involved 20,308 unique genes with 21,703 possible transcripts, while those in normal samples involved 2,672 unique genes with 2,718 possible transcripts. Our next step will be to evaluate of the functional effects of the SNVs with the aim of finding potential TC-associated genes.