Appendix A

Supplementary Tables and Figures

Several files that were generated over the course of the dissertation were either too large to be included in print or were descriptions of processes. These files have been included in an accompanying compact disk and are described in this Section.

A.1 Results from association analysis for all genes in the Cancer Gene Census

File name in CD : Supplementary_Table_6_Association_results_for_CGCL_genes.xlsx

This table includes the results from the association analysis for all the genes that are present in the Cancer Gene Census. The columns included in the table are the Ensembl id of the gene, the symbol of the gene, the chromosome where the gene is present, the original p-value generated for the gene and the false discovery rate corrected p-value of the gene.

A.2 Results from association analysis for all protein coding genes as designated on Ensembl

File name in CD : Supplementary_Table_7_Association_results_for_all_protein_coding_genes.xlsx

This table includes the results from the association analysis for all the genes that are present in all protein coding genes. The complete list of protein coding genes were obtained from Ensembl. The columns included in the table are the Ensembl id of the gene, the symbol of the gene, the chromosome where the gene is present, the original p-value generated for the gene and the false discovery rate corrected p-value of the gene.

A.3 Parameters used for the different steps involved in the execution of pVAAST for the joint association-linkage analysis

File name in CD : Supplementary_Table_8_parameters_for_running_pVAAST.xlsx

Multiple commands are executed to generate the pVAAST results from the input VCF file containing the jointly called variants across all the samples in the dataset. This file includes all the different steps and softwares that are a part of the pVAAST package which need to be executed for generating the final results. The parameters necessary for running these softwares along with the required input files are also specified in this file.

A.4 Results from pVAAST using the default background file from the 1000 genomes project

File name in CD : Supplementary_Table_9_pVAAST_results_default_1000G_background.xlsx

This table includes the final results from pVAAST using the default background or control file provided by pVAAST which is obtained from the 1000 genomes project. The columns included in the table are the symbol of the gene, the CLRT score generated by the association analysis, the p-value generated during the association analysis and the LOD score generated by the linkage analysis.

A.5 Results from pVAAST using INTERVAL exomes background file

File name in CD : Supplementary_Table_10_pVAAST_results_INTERVAL_exomes_background.xlsx

This table includes the final results from pVAAST using the secondary background file that became available over the duration of the project which was obtained from the INTERVAL exomes project. The columns included in the table are the symbol of the gene, the CLRT score generated by the association analysis, the p-value generated during the association analysis and the LOD score generated by the linkage analysis.

A.6 Complete list of variants with high segregation in cases

File name in CD : Supplementary_Table_11_variants_with_high_segregation_in_cases.xlsx

This table includes the list of variants where the variant almost completely segregates with the diseases in the families each variant is present in.

A.7 Complete list of variants associated with cancer in ClinVAR that are present in the dataset

File name in CD : Supplementary_Table_12_ClinVAR_cancer_variants.xlsx

This table includes all the variants annotated as playing a role in different types of cancer as annotated by ClinVAR which are also present within the cases in the dataset.

A.8 Results from the association analysis for the transcription factor binding motif variants for genes in the Cancer Gene Census

File name in CD : Supplementary_Table_13_Association_results_for_CGCL_genes_TFMOTIF_analysis.xlsx

This table includes the results from the association analysis for all the genes that are present in the Cancer Gene Census where the variants are also present in known transcription factor binding motif locations. The columns included in the table are the Ensembl id of the gene, the symbol of the gene, the chromosome where the gene is present, the original p-value generated for the gene, the false discovery rate corrected p-value of the gene and a final column showing the number of samples affected in each family to represent the maximum percentage of segregation of the variant with the disease for each gene.

A.9 Parameters used for the generation of structural variants

File name in CD : Supplementary_Table_14_parameters_for_running_LUMPY.xlsx

This file includes the parameters used in the generation of structural variants using Lumpy and their respective descriptions. All required input files are also specified here.

A.10 Complete list of filtered and annotated structural variants

File name in CD : Supplementary_Table_15_Structural_variants_results.xlsx

This table includes the complete list of filtered and annotated structural variants.

A.11 Supplementary Figure 1

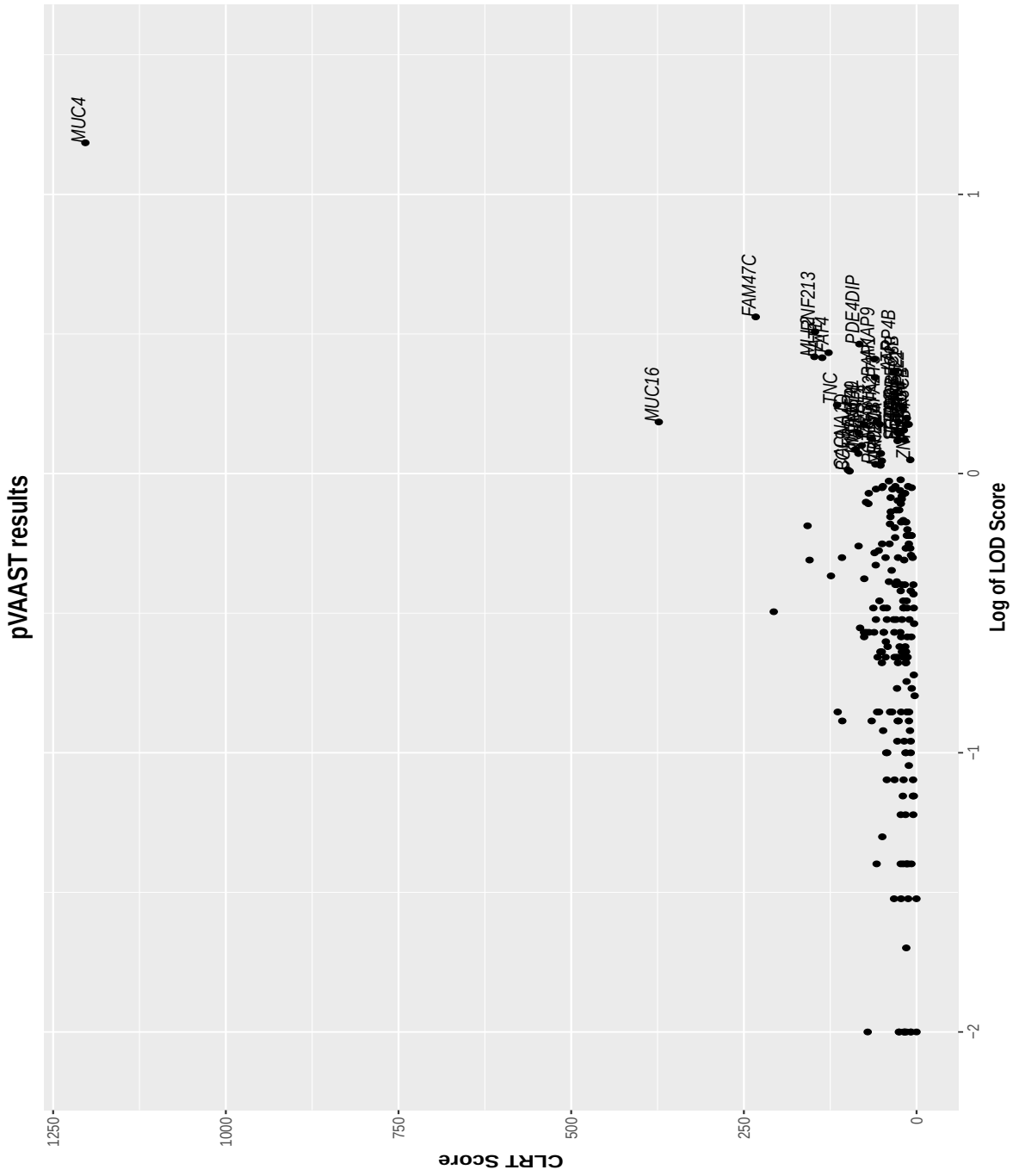

Figure A.1: Original results from pVAAST using the INTERVAL exomes as the background. The y-axis represents the CLRT score for each gene while the x-axis represents the log10 value of the LOD score. Genes with CLRT score>50 or log10 LOD score>0 are represented with their names while the other genes are represented as points.

A.12 Supplementary Figure 2

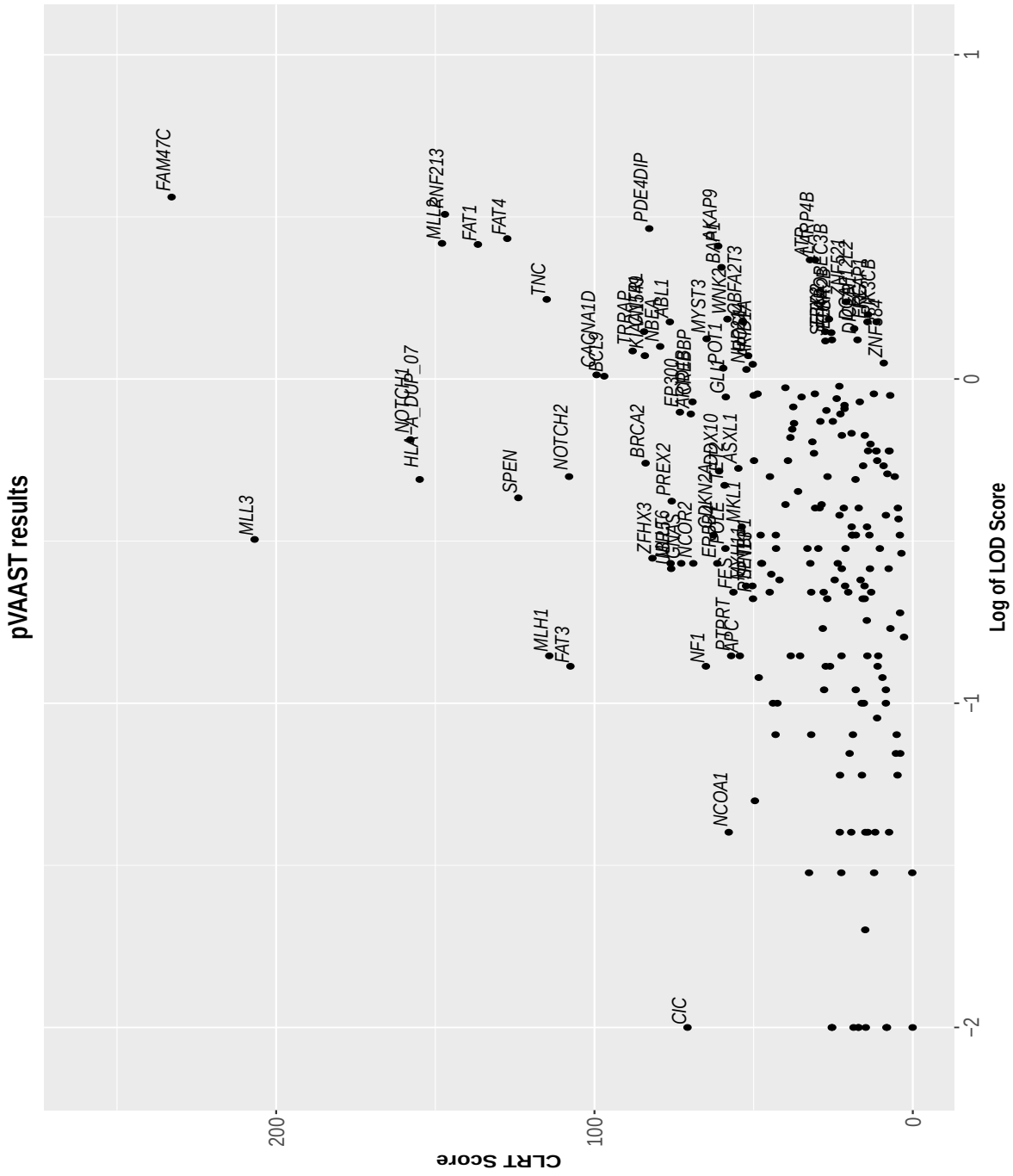

Figure A.2: Results from pVAAST for all genes in the Cancer Gene Census excluding *MUC4* and *MUC16* using the INTERVAL exomes as the background. The y-axis represents the CLRT score for each gene while the x-axis represents the log10 value of the LOD score. Genes with CLRT score>50 or log10 LOD score>0 are represented with their names while the other genes are represented as points.