TCGA Analysis Pipeline Usage Details

0_Download_SNPArray_Table_From_GDC.pl – This script is useful for getting the SNP-Array tables for cancer types. It is also useful if the user just wants the SNP-Array tables and not the download files. To use this script the user will need to specify the cancer type (or types separated by a comma: e.g. OV, PRAD), the experimental strategy which is "Genotyping array", the array type which is Genotypes or "Copy number estimate", the command for downloading (either curl or aria2c) and the full path to the key that was downloaded from GDC.

0_Download_SNPArray_From_GDC.pl – Downloads either SNP-Array data for Genotypes or "Copy number estimate". If no table can be located by the script then it will download and make the table from GDC otherwise it will just start downloading the files as the table already exists. To use this script the user will need to specify the cancer type, the experimental strategy which is "Genotyping array", the array type which is either Genotypes or "Copy number estimate", and the full path to the key that was downloaded from GDC.

1.0_Prep_SNPs_for_Imputation_and_Plink.pl – Gets the Genotypes SNP-Array data ready for imputation by unzipping GenomeWideSNP_6.na35.annot.csv.zip from the Database directory into a new directory called affy6, which will be used in other scripts. After unzipping the archive it parses GenomeWideSNP_6.na35.annot.csv file to outputs the data to snp6.anno.txt. It then performs another parse and outputs the results to a file called snp6.cd.txt. If these files already exist, then this script will not be executed and it will as the user to execute the next script. To use this script, the user will need to specify the cancer type.

1.1_Birdseed_to_ped_and_maps.pl – Parses the Genotype files that were download to the Genotypes directory and creates map and ped files. To use this script, the user will need to specify the cancer type.

1.2_Shapeit_and_Imputation.pl – Performs the shapeit process on the peds and maps directories and prints the output to a new directory called phased. After the shapeit process is finished, it gets the sizes of each chromosome and outputs them to a file called chr_lens_grep_chr. It will then parse that file and get only chr1-22 and X. The imputation is broken into two parts where it will do half the chromosomes then the other have after the first half is finished. It prints the results of the imputation to a directory called phased_imputed_raw_out. After each imputation, this script will remove the unnecessary files in the phased_imputed_raw_out.

2.0_Prep_For_Bad_SNPs_CNVs.pl – Performs lookups on Copy number estimate files to get normal CNVs. The user needs to only specify the cancer type as well as have Copy number estimate data downloaded.

2.1_Get_Bad_SNPs_and_CNVs.pl – Performs routines to get the bad SNPs and CNVs from the Genotypes. Only the cancer type needs to be specified to run this script as well as running all of the previous scripts for this one to work.

3.0_Download_RNASeq_WGS_and_do_Mpileup.pl – Used for downloading either RNA-Seq or WGS data. If the script does not locate a table for the specified type then it will create one and begin the download or it will just start the download if it locates the table in the directory it looks in. The

arguments to use this script are same for RNA-Seq and WGS with WGS only having an additional one. For RNA-Seq the user needs to specify the cancer type, the experimental strategy (e.g. RNA-Seq), and the path to the gdc key. For WGS, just do put in the same arguments as RNA-Seq but put WGS for experimental strategy and specify the path to the VarScan.jar file. There are also optional arguments that can be entered in. These arguments are number, choice and command. Number will take the amount that it specifies and download those number of files before getting the next amount. Choice is either all, download, or mpileup. All will download and do mpileups, download will just download, and mpileup will do mpileups on downloaded bams and will only run if they are downloaded and indexed. The command argument is to specify which downloader the user wants to run. The two download commands that this pipeline supports is curl and aria2c.

3.0_Dwnld_Table_4_RNASeq_WGS_From_GDC.pl – Creates either RNA-Seq or WGS tables for cancer types. To use this script, the user will need to specify the cancer type (or types separated by a comma: e.g. OV, PRAD), the experimental strategy which is either RNA-Seq or WGS, and the full path to the key that was downloaded from GDC with the optional argument being command.

3.1_Gene_Level_ASE_Analysis.pl – Performs gene level ase analysis to get gene level reads. For this script to work the user will need to specify the cancer type and run the previous scripts.

3.2_Export_ASE_Data.pl – Gets data ready to be analyzed in Matlab or other platforms. Only the cancer type needs to be specified.

4.0_Somatic_Variants.pl – Filters through the varscan results to get somatic mutations. Only the cancer type needs to be specified to run this script with read cutoff, tumor frequency and normal frequency being optional.

4.1_Upstream_Downstream_Analysis.pl – Performs upstream and downstream analysis to get annotations and gets the data to ready to be analyzed in Matlab or another platform. Only the cancer type needs to be specified as well as running the previous script.

Flags for Scripts

0_Download_SNPArray_Table_From_GDC.pl:

[--disease_abbr|-d disease_abbr (e.g. PRAD or OV,PRAD)]

[--exp_strat|-e Experimental Strategy (Genotyping array)]

[--array_type|-a array data type (e.g. Genotypes)] [--help|-h]

0_Download_SNPArray_From_GDC.pl:

[--disease_abbr|-d disease_abbr (e.g. PRAD)] [--exp_strat|-e Experimental Strategy (Genotyping array)]

[--array_type|-a array data type (e.g. Genotypes)] [--command|-c curl or aria2c]

[--key|-k path to gdc key] [--help|-h]

1.0 - 2.1, 3.1 - 3.2 and 4.1:

[--disease_abbr|-d disease_abbr (e.g. PRAD)] [--help|-h]

3.0_Dwnld_Table_4_RNASeq_WGS_From_GDC.pl:

[--disease_abbr|-d disease_abbr (e.g. PRAD or OV,PRAD)]

[--exp_strat|-e Experimental Strategy (e.g. WGS/RNA-Seq)]

[--key|-k path to gdc key] [--help|-h]

3.0_Download_RNASeq_WGS_and_do_Mpileup.pl:

[--disease_abbr|-d disease_abbr (e.g. PRAD)]

[--exp_strat|-e Experimental Strategy (e.g. WGS/RNA-Seq)] [--option|-o all, download or mpileups]

[--number|-n number of bams to download for RNA-Seq and number of bam pairs for WGS]

[--command|-c curl or aria2c] [--var_path|-v path to the VarScan jar file (Enter if -e is WGS)]

[--key|-k peht to gdc key] [--help|-h]

4.0_Somatic_Variants.pl:

[--disease_abbr|-d (e.g. PRAD)] [--readcutoff|-r read cutoff (e.g. 20)]

[--tfreq|-t tumor alt frequency (e.g. 0.1)] [--nfreq|-n normal alt frequency (e.g. 0.1)]