Driver-ASE 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" and the array type, which is Genotypes or "Copy number estimate".

0_Download_SNPArray_From_GDC.pl – Downloads either SNP-Array data for Genotypes or "Copy number estimate". If the script cannot locate the table, 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", the full path to the key that was downloaded from GDC and/or the download command.

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 the GenomeWideSNP_6.na35.annot.csv file and outputs the data to a file called snp6.anno.txt. It then performs another parse and outputs the results to a file called snp6.cd.txt. If these files already exist, the script will not be executed and it will ask 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 – uses software called plink to parse 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 and/or enter the path/command to the plink software as well as the input type to tell the program to only get tumor, normal or tumor and normal samples in the processing.

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. To use this script, the user will need to specify the cancer type with the option to enter in the path/command shapeit and/or impute2.

1.3_Make_het_cds – Processes and keep heterozygous SNPs as well as make cds sorted files using them. To use this script, the user will need to enter the cancer type with the option of entering the path/command for plink.

2.0_Prep_For_Bad_SNPs_CNVs.pl – An optional script that performs lookups on Copy number estimate files to get normal CNVs. If the user wishes to include CNV data in their analysis, they will need to download the Copy number estimate files from GDC using script 0_Download_SNPArray_From_GDC.pl. To use this script, the user will need to enter in the cancer type.

2.1_Get_Bad_SNPs_and_CNVs.pl – An optional script that 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. Once the selected number of bams have been downloaded and indexed, it will run mpileups on them, if the user doesn't specify to only download bams for the choice argument. 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 enter in the same arguments as RNA-Seq with the addition of the path to the VarScan jar file. There are also optional arguments that can be entered in. These arguments are number, option, intersect, download and samtools. Number will take the amount that it specifies and download those number of files before getting the next amount. Intersect will intersect the RNA-Seq, WGS and Genotypes tables to get matching TCGA IDs. Option 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 download argument is to specify which downloader the user wants to run. The two download commands that this pipeline supports are curl (default) and aria2c. The samtools argument is just the path/command for samtools.

3.0_Dwnld_RNASeq_WGS_Table_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, the path to the gdc key and the options of inputting the command for downloading (curl (default) or aria2) and/or overlap to overlap the table files.

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 with the option of inputting the command/path for overlapSelect. There is also an option to overlap the IDs of RNA-Seq with the IDs of WGS and Genotypes and only have those overlapped IDs be used in the analysis.

3.2_Export_ASE_Data.pl – Gets data ready to be analyzed in Matlab or other platforms. If the user is overlapping the RNA-Seq, WGS, and Genotype id, they will need the tables that have the required ids to overlap. Also, if the user is including bad SNPs and CNVs in their analysis, they will need to run script 0_Download_SNPArray_From_GDC.pl to download copy number data from GDC and run both script 2.0 and 2.1 to process it. For command line options, only the cancer type needs to be specified with the option to archive, delete files, include bad SNPs/CNVs, include duplicate ids, overlap to archiving only overlapped data at the end of the script and/or the path/command for overlapSelect.

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. There is also an option to overlap the IDs of WGS with the IDs of RNA-Seq and Genotypes and only have those overlapped IDs be used in the analysis.

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 with the option of archiving files, removing not needed files, overlap for archiving only overlapped data at the end of the script and/or the path/command for overlapSelect.

Flags for Scripts

0_Download_SNPArray_Table_From_GDC.pl:

[--cancer|-c cancer_type (e.g. PRAD)] [--Expstrategy|-E Experimental strategy (Genotyping array)]

[--arraytype|-a array data type (e.g. Genotypes or "Copy number estimate")] [--help|-h]

0_Download_SNPArray_From_GDC.pl:

[--cancer|-c cancer_type (e.g. PRAD)] [--Expstrategy|-E Experimental strategy (Genotyping array)]

[--arraytype|-a array data type (e.g. Genotypes or "Copy number estimate")]

[--download|-d curl or aria2c] [--key|-k path to gdc key] [--help|-h]

1.0_Prep_SNPs_for_Imputation_and_Plink.pl, 2.0_Prep_For_Bad_SNPs_CNVs.pl and 2.1_Get_Bad_SNPs_and_CNVs.pl:

[--cancer|-c cancer_type (e.g. PRAD)] [--help|-h]

1.1_Birdseed_to_ped_and_maps.pl:

[--cancer|-c cancer_type (e.g. PRAD)] [--plink|-p path to plink]

[--sampletype|-s 0(normal)|1(tumor)|2(normal/tumor)] [--help|-h]

1.2_Shapeit_and_Imputation.pl:

[--cancer|-c cancer_type (e.g. PRAD)] [--shapeit|-s path to shapeit] [--impute|-i path to impute2]

[--help|-h]

1.3_Make_het_cds.pl:

[--cancer|-c cancer_type (e.g. PRAD)] [--plink|-p path to plink] [--help|-h]

3.0_Dwnld_RNASeq_WGS_Table_From_GDC.pl:

[--cancer|-c cancer_type (e.g. PRAD)] [--Expstrategy|-E Experimental strategy (e.g. WGS/RNA-Seq)]

[--download|-d curl or aria2c] [--overlap|-o (n|no|y|yes)] [--key|-k path to gdc key] [--help|-h]

3.0_Download_RNASeq_WGS_and_do_Mpileup.pl:

[--cancer|-c cancer_type (e.g. PRAD)] [--Expstrategy|-E Experimental strategy (e.g. WGS/RNA-Seq)]

[--option|-o all, download or mpileups] [--intersect|-i (y|yes|n|no)]

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

[--download|-d curl or aria2c] [--samtools|-s path to samtools]

[--VarScan|-V path to the VarScan jar file (Enter if -E is WGS)] [--key|-k path to gdc key] [--help|-h]

3.1_Gene_Level_ASE_Analysis.pl:

[--cancer|-c cancer_type (e.g. PRAD)] [--overlapSelect|-S path to overlapSelect]

[--overlap|-o (y|yes|n|no)] [--help|-h]

3.2_Export_ASE_Data.pl:

[--cancer|-c cancer_type (e.g. PRAD)] [--overlapSelect|-S path to overlapSelect]

[--badsnpcnvs|-b (y|yes|n|no)] [--overlap|-o (y|yes|n|no)]

[--intersecttype|i 0 for long format or 1 for summary] [--duplicates|-d (y|yes|n|no)]

[--archive|-a (y|yes|n|no)] [--remfiles|-r (y|yes|n|no)] [--help|-h]

4.0_Somatic_Variants.pl:

[--cancer|-c (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)]

4.1_Upstream_Downstream_Analysis.pl:

[--cancer|-c cancer_type (e.g. PRAD)] [--overlap|-o (y|yes|n|no)] [--remfiles|-r (y|yes|n|no)]

[--archive|-a (y|yes|n|no)] [--overlapSelect|-S path to overlapSelect] [--help|-h]