Publicly Available Data

Jessica Murphy

April 16, 2019

GEOsearch

The following table shows the number of datasets available at different filtering steps using GEOsearch in R. The first step browsed for the original search term: bladder cancer, alcohol use disorder, or type 1 diabetes. The second step selected datasets from only human samples. The third step selected datasets that used specific microarray techologies: Illumina human-6 v2.0 expression beadchip, Illumina HumanHT-12 V4.0 expression beadchip, and Affymetrix Human Genome U133 Plus 2.0 Array or the Affymetrix Human Genome U133A 2.0 Array. The fourth step selected datasets with a sample size of more than 100 people.

	Bladder Cancer	Alcohol Use Disorder	Type 1 Diabetes
Step 1: Original Search	222	70	4
Step 2: Only in humans	201	34	3
Step 3: Specific Arrays	23	4	3
Step 4: Certain Sample Size	2	2	1

GEOquery

Data was downloaded from GSE13507 using GEOquery. There were 43,148 probe sets and 256 samples. The gender statistics from the metadata showed there were 30 females, 135 males, and 91 missing values. The age statistics showed the average age was 65.18 with a standard deviation of 11.97 and 91 missing values.

TCGAbiolinks

Using TCGAbiolinks, a search was performed for bladder cancer sequencing data that used Illumina technology and had gene expression from primary solid tumors. There were 1,249 resulting samples. Of these, 414 were from RNA-Seq experiments, 417 were from miRNA-Seq experiments, and 418 were from whole exome sequencing (WXS) experiments.