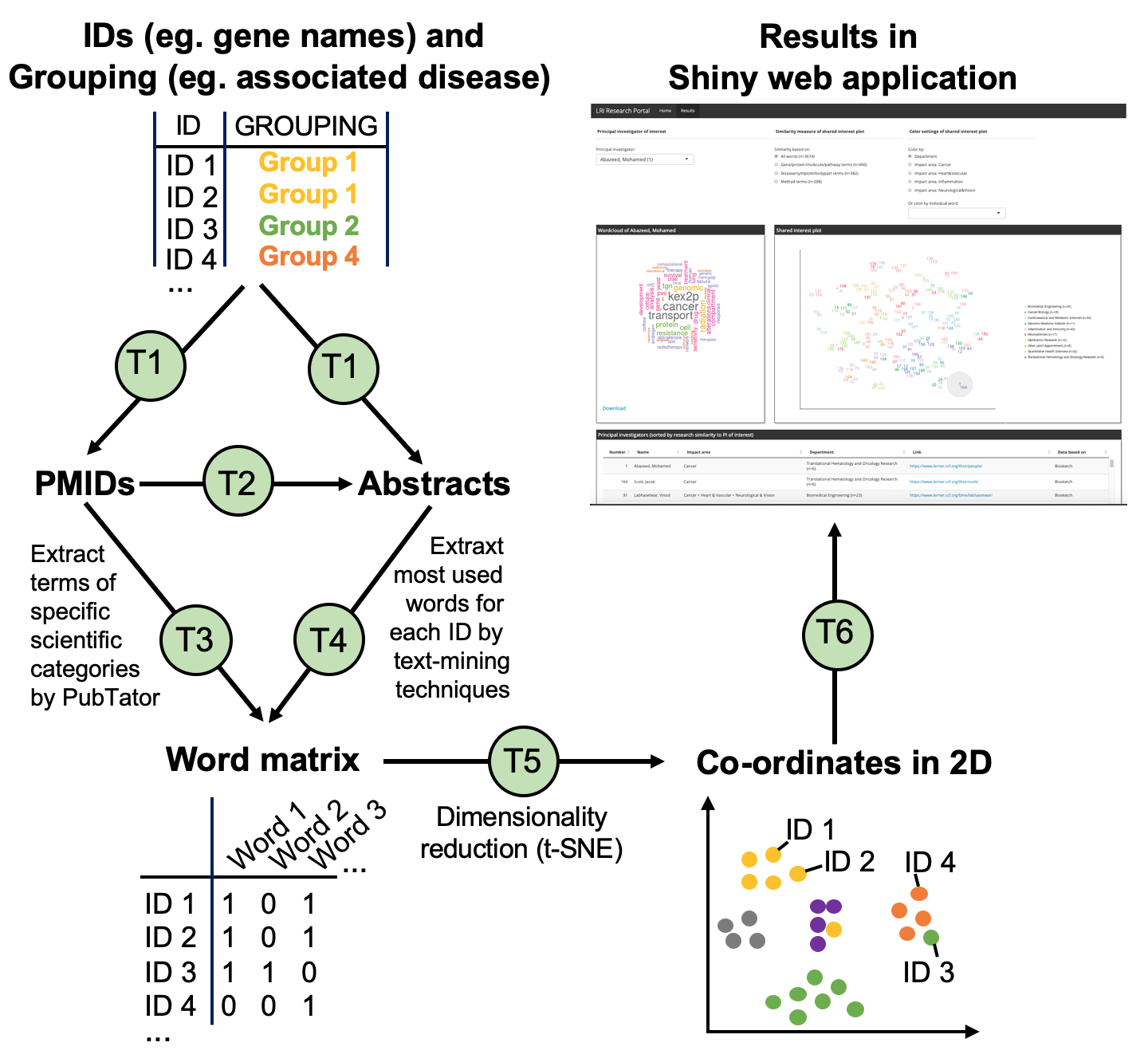
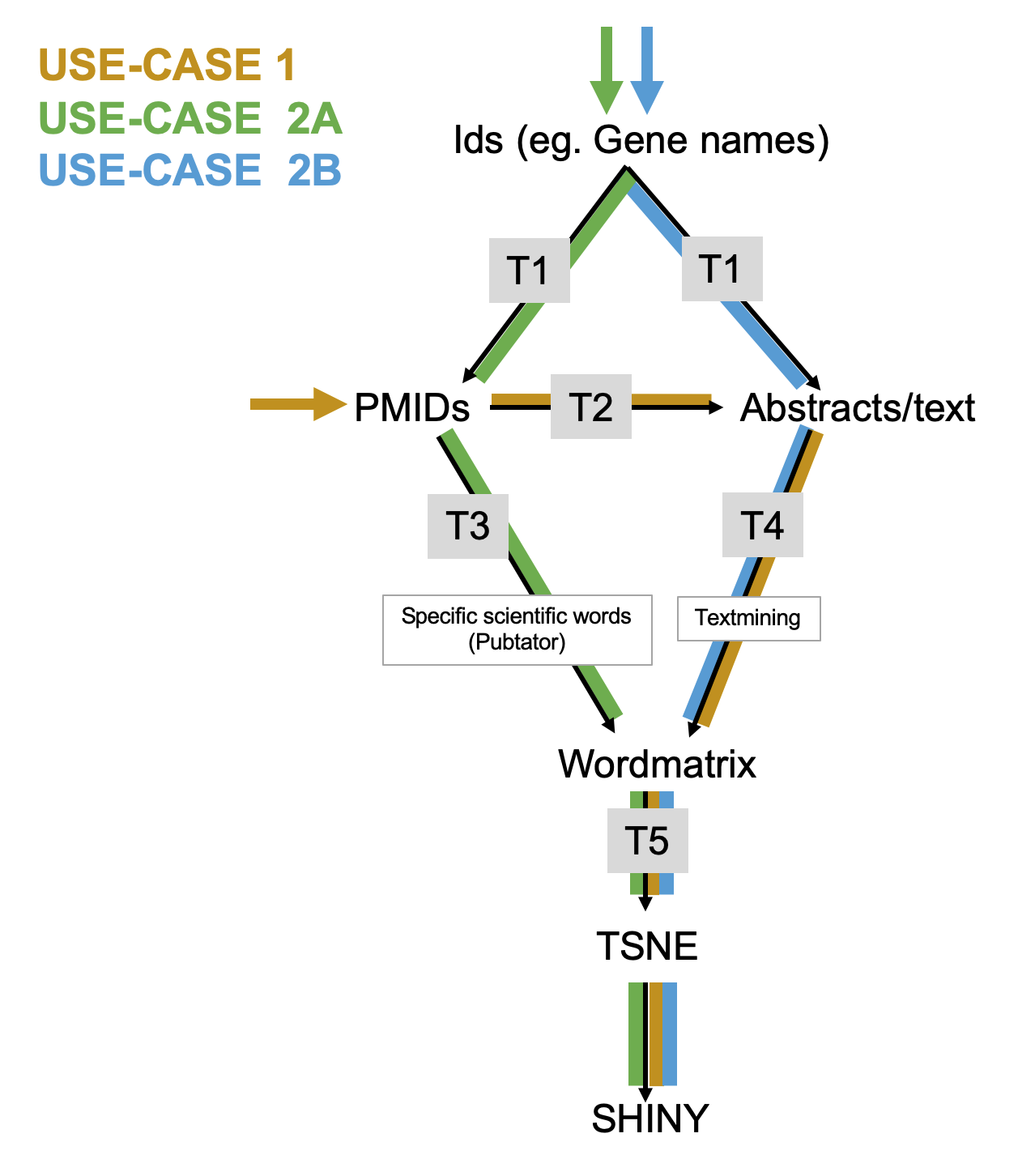
**Toolset overview**

****

**Workflows of different use-cases**

T=Tool

**USE CASE 1:**

**AIM**

Fetch abstracts of PMIDs that were authored by researchers (=IDs) and generate word matrix of researchers\*top100 words. With this matrix perform T-SNE to identify groups of researchers with similar interest based on the words they use in the abstracts.

**REQUIREMENT**

R

**INPUT**

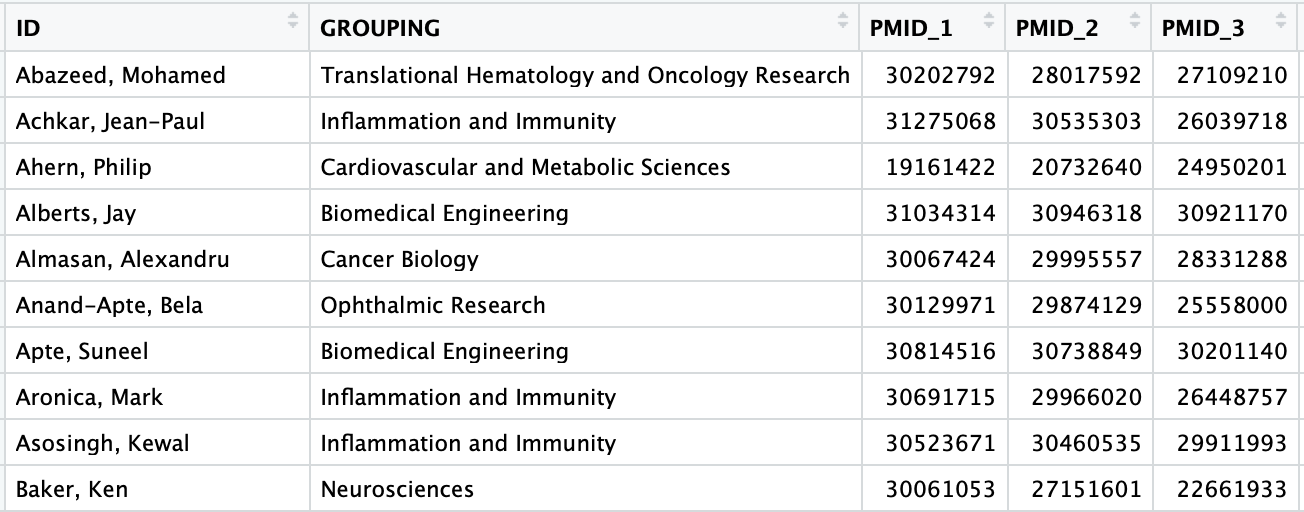
Researcher\_testdata.txt

Columns:

ID= names of researchers (n=182)

GROUPING= department name of researcher

PMID\_1 .. PMID\_10= PMIDs of latest publications per researcher



**STEP 1:**

Tool 2: For each PMID get the according abstract text. Output by default: "T2\_result"

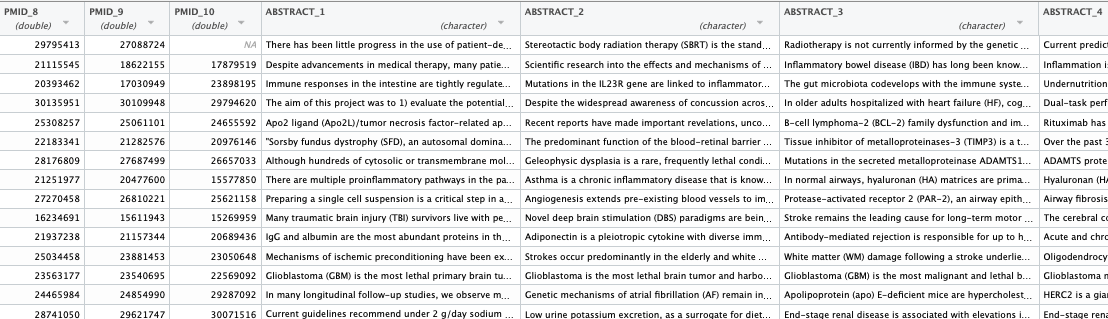
$ cd <path\_to\_folder\_with\_data\_and\_scripts>

$ chmod -x T2\_pubmedsearch\_by\_pmid.R

$ RScript T2\_pubmedsearch\_by\_pmid.R \

--input "Researcher\_testdata.txt"

Cutout of output:



**STEP 2:**

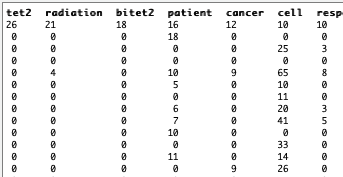
Tool4: Text mining of abstracts/text. By default words in plural and singular are merged to the singular form and stop words are removed. See help for more options. A word matrix based on the top 100 words is generated. Output by default: "T4\_result"

$ chmod -x T4\_text\_to\_wordcountmatrix.R

$ RScript T4\_text\_to\_wordcountmatrix.R \

--input "T2\_result" -n 100

Cutout of output:



**STEP 3:**

Tool5: Dimensionality reduction of binary matrix into coordinates of researchers in 2D.

Here --input is the initial input data and --matrix the generated word matrix, --perplexity is the perplexity value of TSNE, --textsize is the size of the labels in the plot.

$ chmod -x T5\_dimreduction\_of\_wordcountmatrix.R

$ RScript T5\_dimreduction\_of\_wordcountmatrix.R \

--input "Researcher\_testdata.txt" \

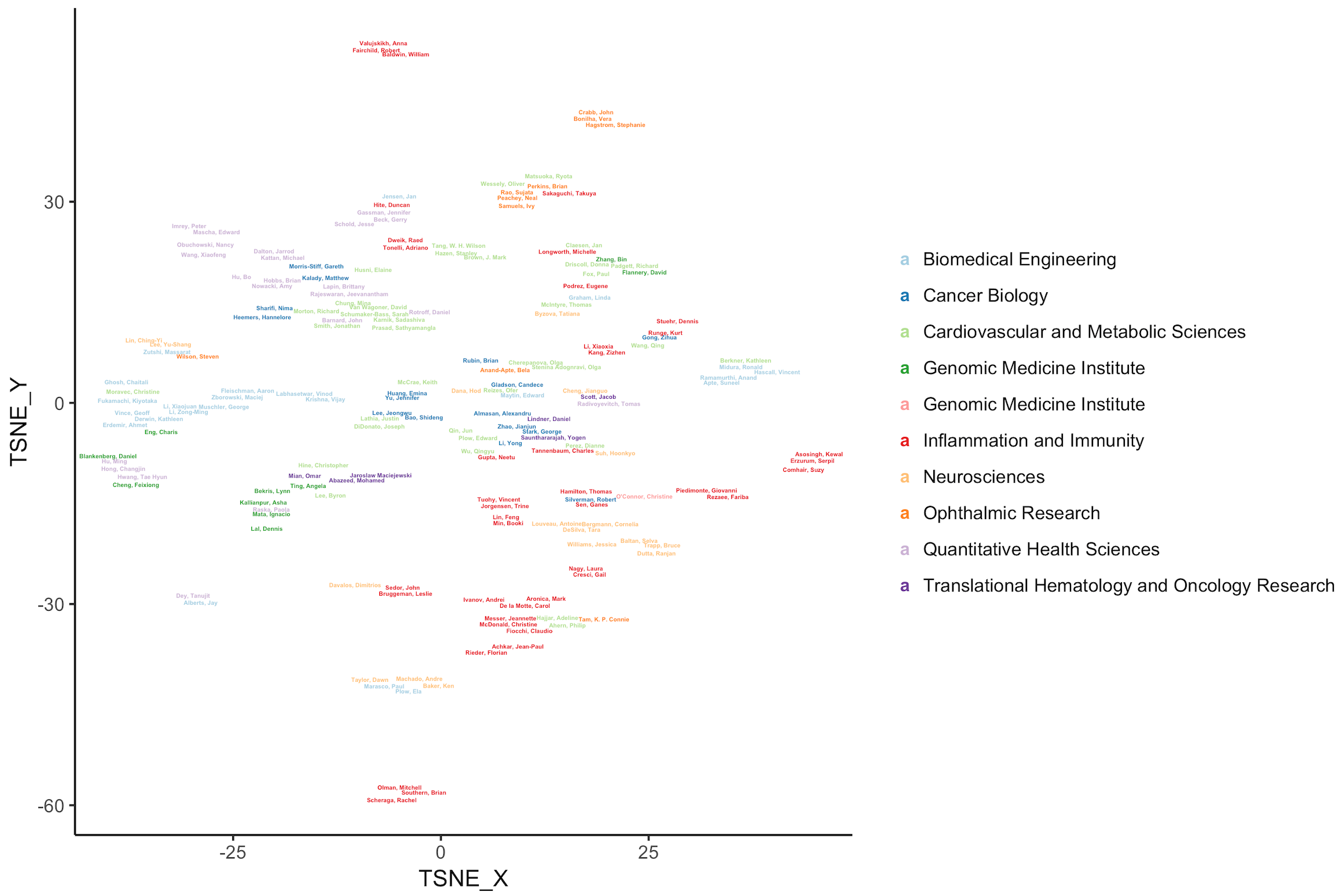
--matrix "T4\_result" \

--output "Researcher\_testdata\_result" \

--perplexity 3 \

--textsize 1

Output: plot and initial input data with coordinates are saved.



**USE CASE 2A:**

**REQUIREMENT**

R

**INPUT**

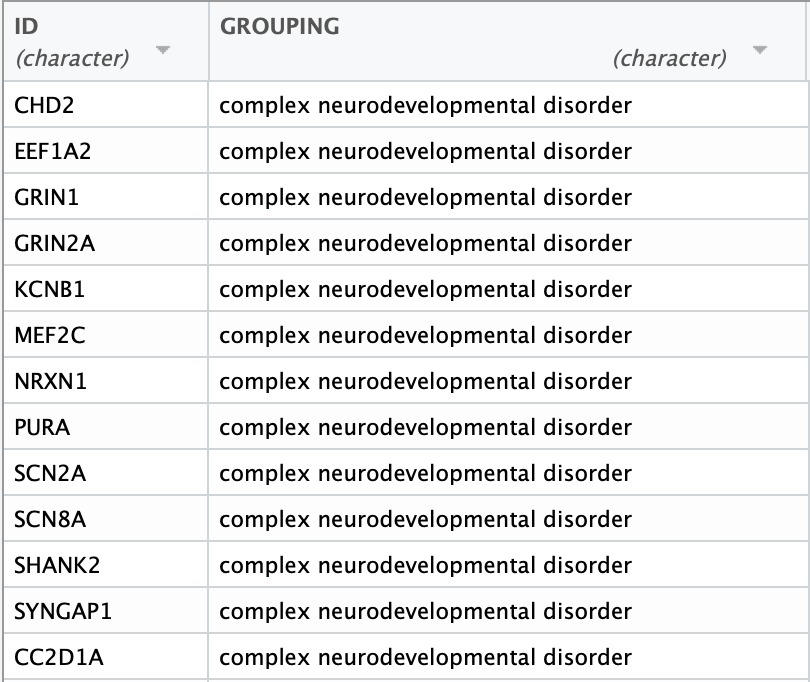
clingen\_testdata

Head of input data:

Columns:

ID= gene names of diseases that have more than 5 associated genes in ClinGen (n=95)

GROUPING= associated diseases (n=6)



**AIM**

Per gene name of diseases that have more than 5 associated genes in ClinGen search Pubmed for the latest 10 abstracts and generate a word matrix of genes\*MeSH terms of abstracts. With this matrix perform T-SNE to identify groups of genes that occur in abstracts with similar MeSH terms.

**STEP 1:**

Tool 1: For each ID term (gene names) get 10 PMIDs. Here –k is used to fetch the PMIDs with an API key from ncbi (not mandatory). Output: "T1\_result".

$ cd <path\_to\_folder\_with\_data\_and\_scripts>

$ chmod -x T1\_pubmedsearch\_by\_keyword.R

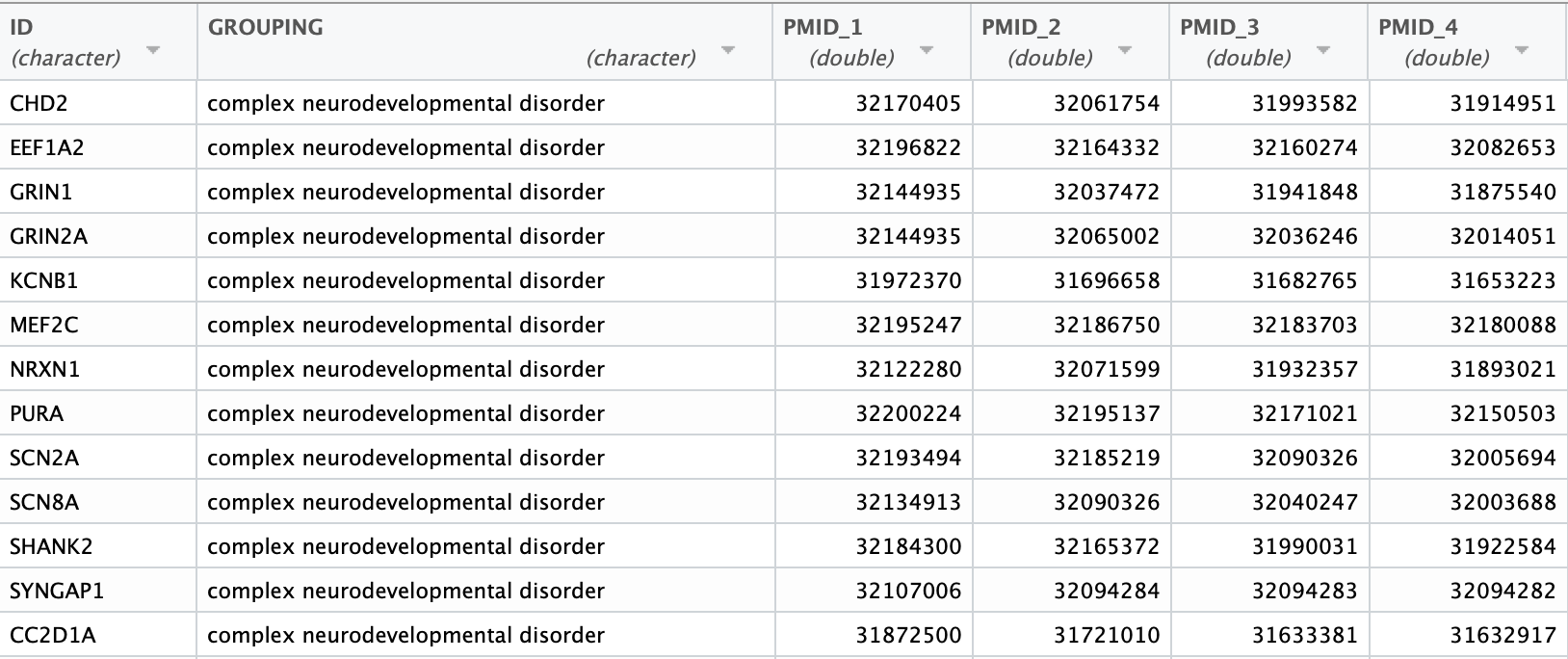
$ RScript T1\_pubmedsearch\_by\_keyword.R \

--input "clingen\_testdata" \

--number 10 \

--key "15f2fa5c9d427fd88d53b5041002aeabd309"

Cutout of output: PMID columns are added



**STEP 2:**

Tool3: Pubtator search of PMIDs. Generation of IDs/MESH terms matrix. Here --categories is used to only extract Genes and Disease MeSH terms. Output: "T3\_result".

$ chmod -x T3\_pmids\_to\_pubtator\_matrix.R

$ RScript T3\_pmids\_to\_pubtator\_matrix.R \

--input "T1\_result" \

--categories Diseases Genes

**STEP 3:**

Tool5: Dimensionality reduction of binary gene name/MESH terms matrix into co-ordinates of gene names in 2D. Here --input is the initial input data and --matrix the generated word matrix, --perplexity is the perplexity value of TSNE, --textsize is the size of the labels in the plot. Output: T-SNE plot and initial input data with co-ordinates of TSNE plot.

$ chmod -x T5\_dimreduction\_of\_wordcountmatrix.R

$ RScript T5\_dimreduction\_of\_wordcountmatrix.R \

--input "clingen\_testdata" \

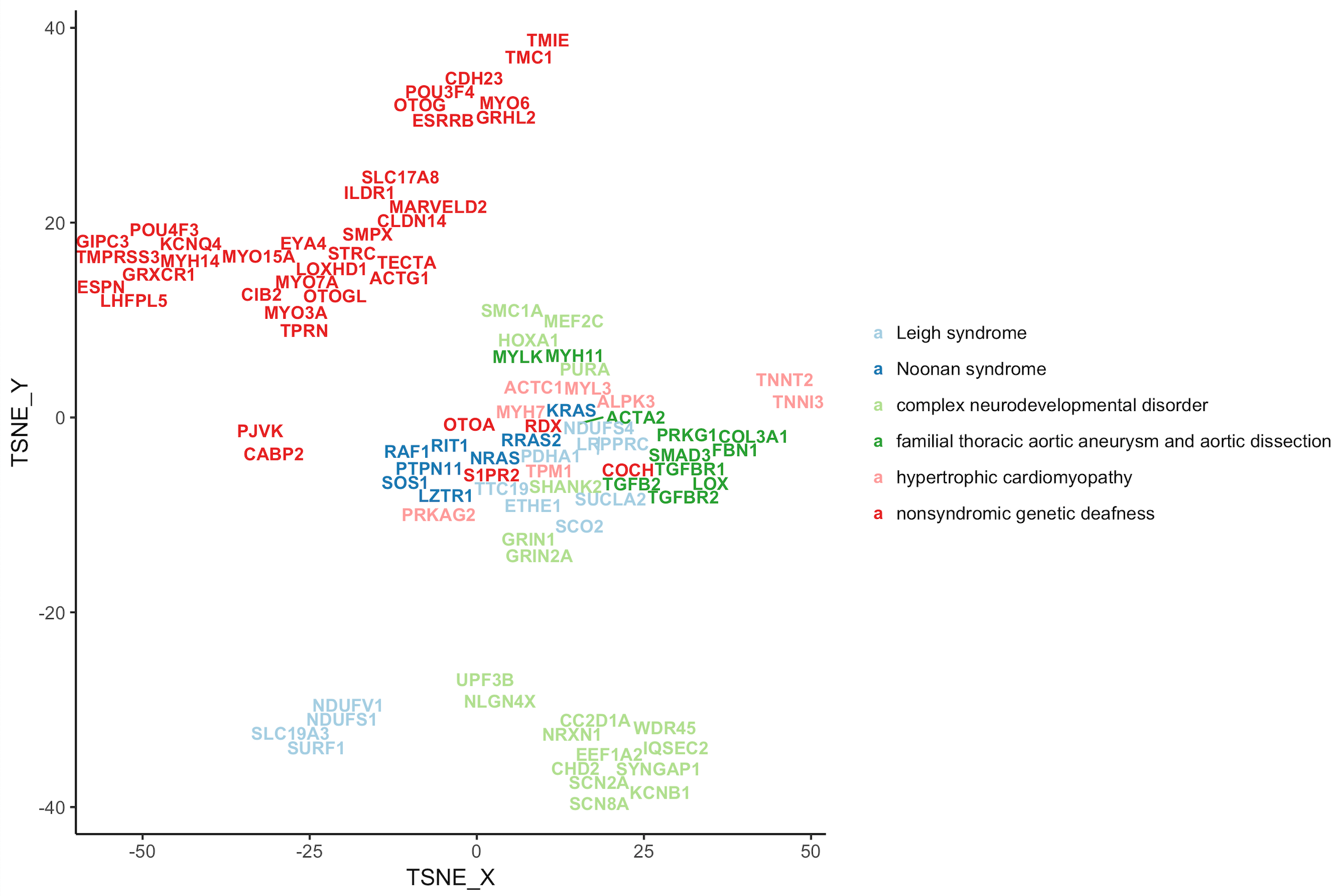
--matrix "T3\_result" \

--output "Clingen\_testdata \_result" \

--perplexity 2 \

--textsize 3

Output:



**USE CASE 2B:**

**REQUIREMENT**

R

**INPUT**

clingen\_testdata

Head of input data:

Columns:

ID= gene names of diseases that have more than 5 associated genes in ClinGen

GROUPING= associated disease

**AIM**

Per gene name of diseases that have more than 5 associated genes in ClinGen get the latest 20 abstracts from PubMed. Next, extract the top 50 words per gene by textmining of all abstracts per gene with the Apply Porter's stemming algorithm. After that, generate a word matrix of genes\*top50words. With this matrix perform T-SNE to identify groups of genes that occur in abstracts with similar words.

**STEP 1:**

Tool 1: For each ID term (gene names) get 20 abstracts. Here –-key is used to fetch the abstracts with an API key from ncbi (not mandatory) and –-abstracts to fetch abstracts and not PMIDs. Output: "T1\_result".

$ cd <path\_to\_folder\_with\_data\_and\_scripts>

$ chmod -x T1\_pubmedsearch\_by\_keyword.R

$ RScript T1\_pubmedsearch\_by\_keyword.R \

--input "clingen\_testdata" \

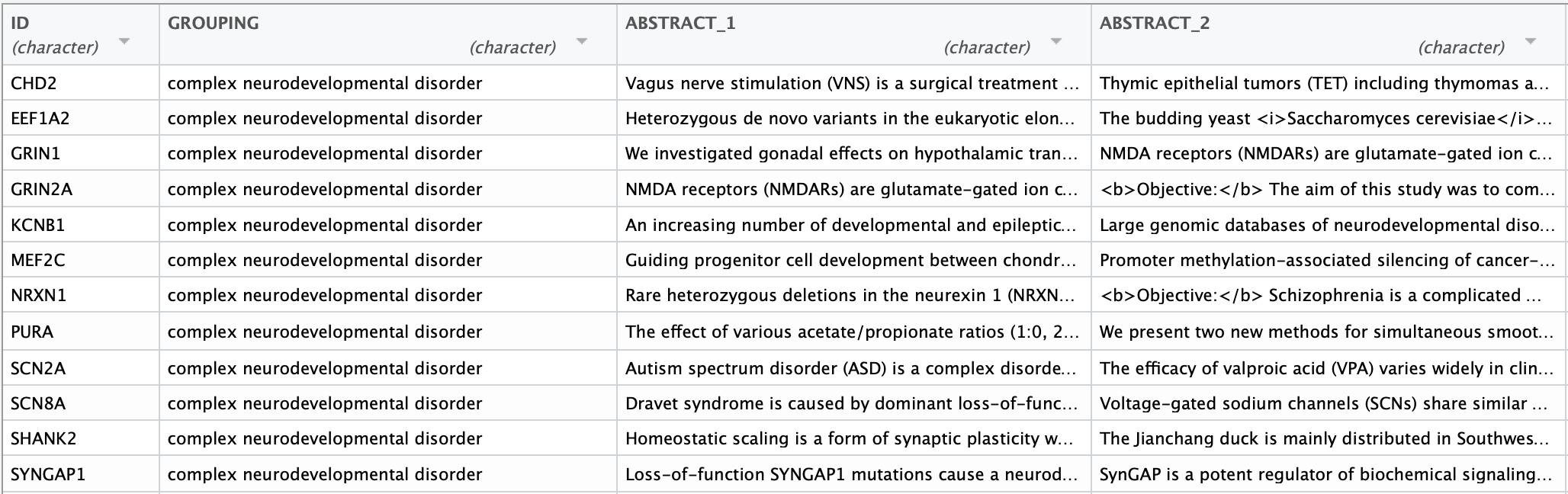
--number 20 \

--key "15f2fa5c9d427fd88d53b5041002aeabd309" \

--abstract \

--output “T1\_result\_b”

Cutout of output:



**STEP 2:**

Tool4: Text mining of abstract text per gene using Apply Porter's stemming algorithm (--stemDoc). Generation of word matrix based on top 50 words. Output: "T4\_result\_b"

$ chmod -x T4\_text\_to\_wordcountmatrix.R

$ RScript T4\_text\_to\_wordcountmatrix.R \

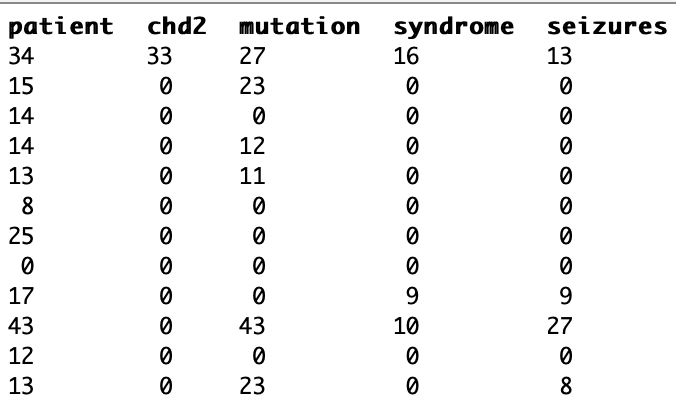
--input "T1\_result\_b" \

--number 50 \

--stemDoc

--output "T4\_result\_b"

Cutout of output:

****

**STEP 3:**

Tool5: Dimensionality reduction of binary word matrix (genes\*words) into co-ordinates of gene names in 2D. Here --input is the initial input data and --matrix the generated word matrix, --perplexity is the perplexity value of TSNE, --textsize is the size of the labels in the plot. Output: plot and initial input data with co-ordinates of TSNE plot.

$ chmod -x T5\_dimreduction\_of\_wordcountmatrix.R

$ RScript T5\_dimreduction\_of\_wordcountmatrix.R \

--input "clingen\_testdata" \

--matrix "T4\_result" \

--output "Clingen\_testdata \_result\_b" \

--perplexity 2 \

--textsize 3

Output: