Data Curation Manual

1 Website

- https://www.ebi.ac.uk/arrayexpress/

2 Filter selection

- By organism - `Homo sapiens`
- By experiment type - `rna assay`

3 Procedure

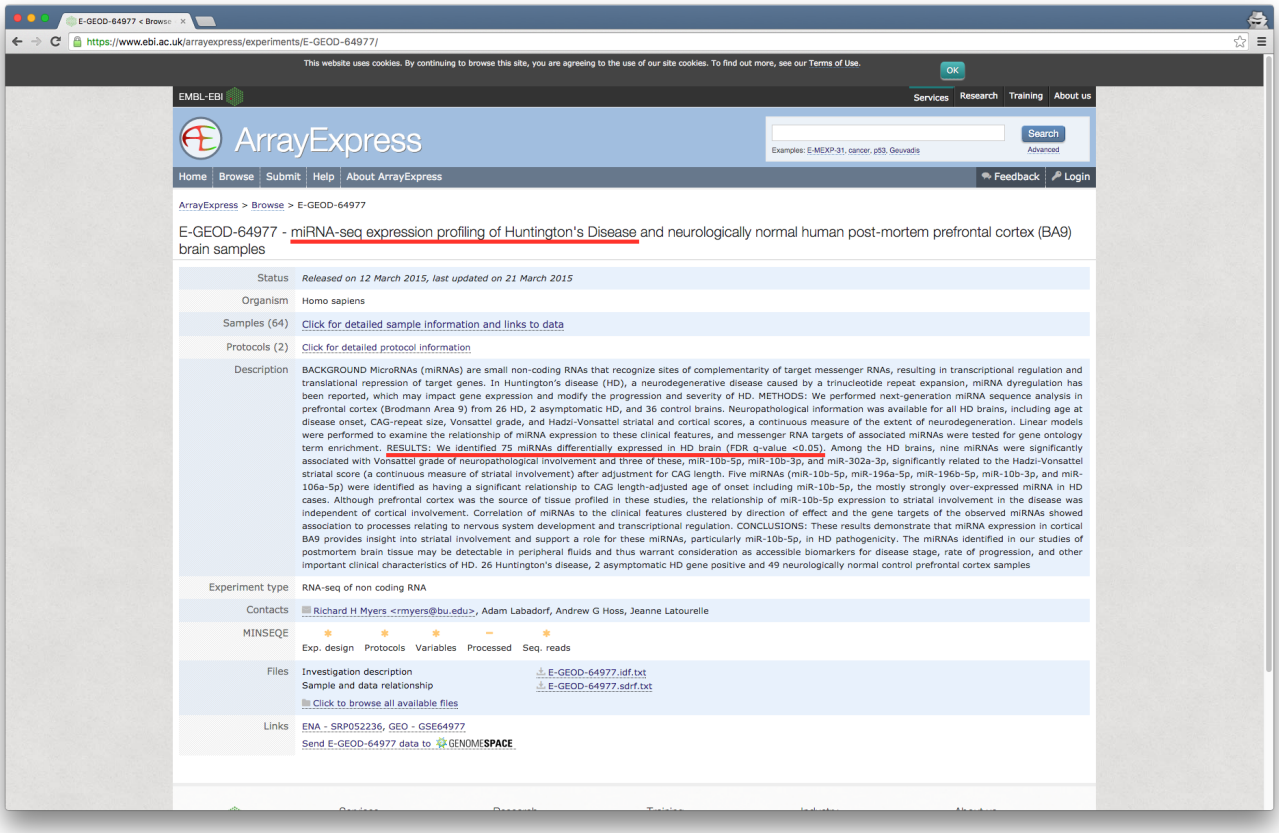
3.1 Search disease name by using the correct filter in `ArrayExpress`

3.2 For each `Accession`, check the following

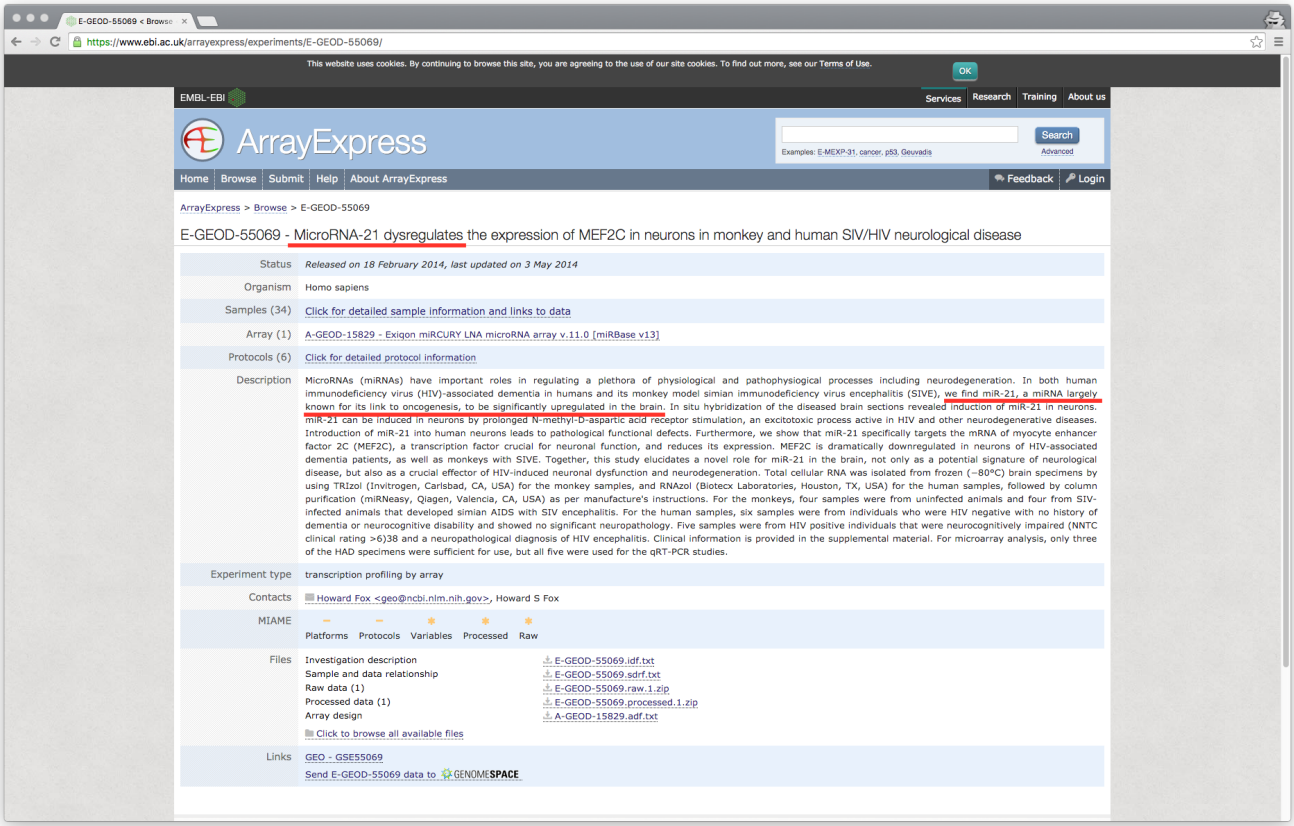
3.2.1 Check if this accession is about `MicroRNA`

The following two datasets are all examples of `MicroRNA`

- Huntington's Disease, E-GEOD-64977

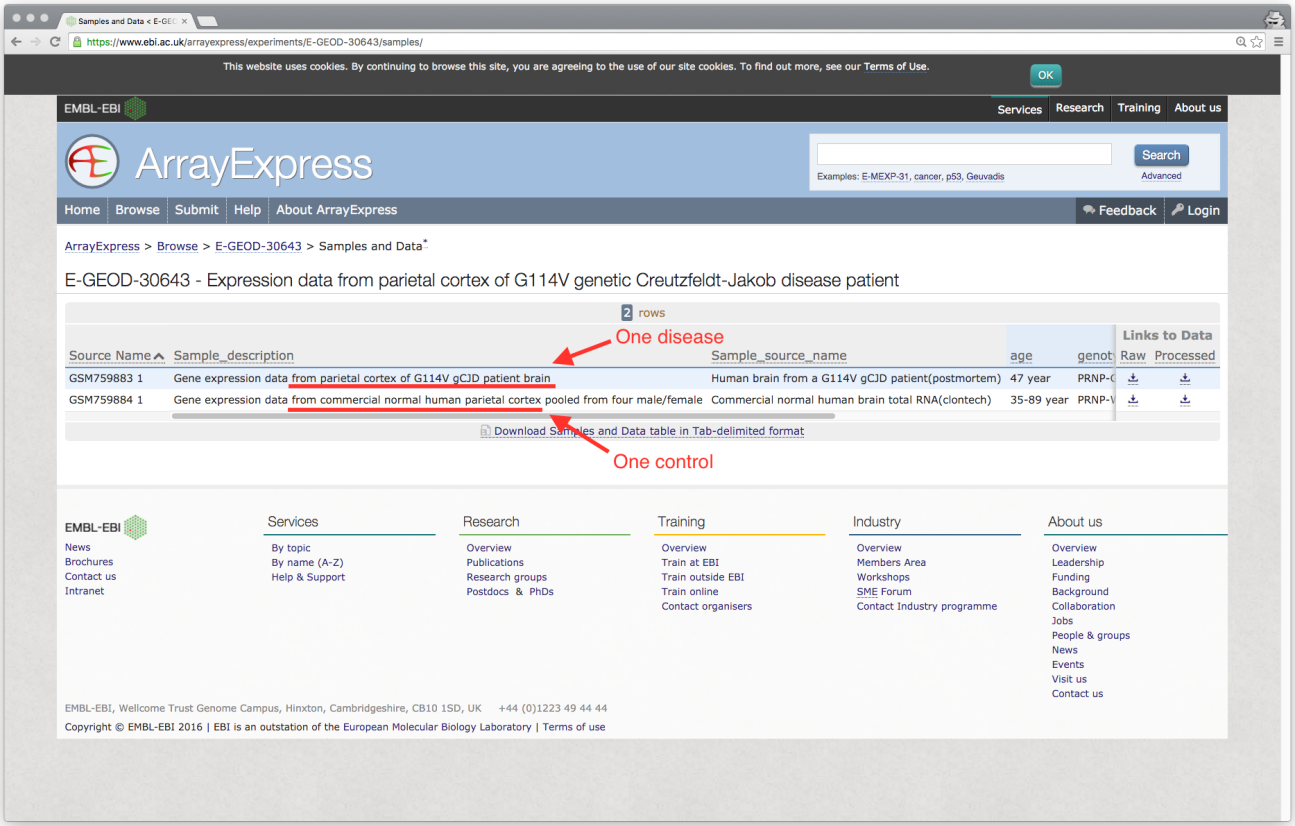


- Encephalitis, E-GEOD-55069



3.2.2 Check if it has enough replicates

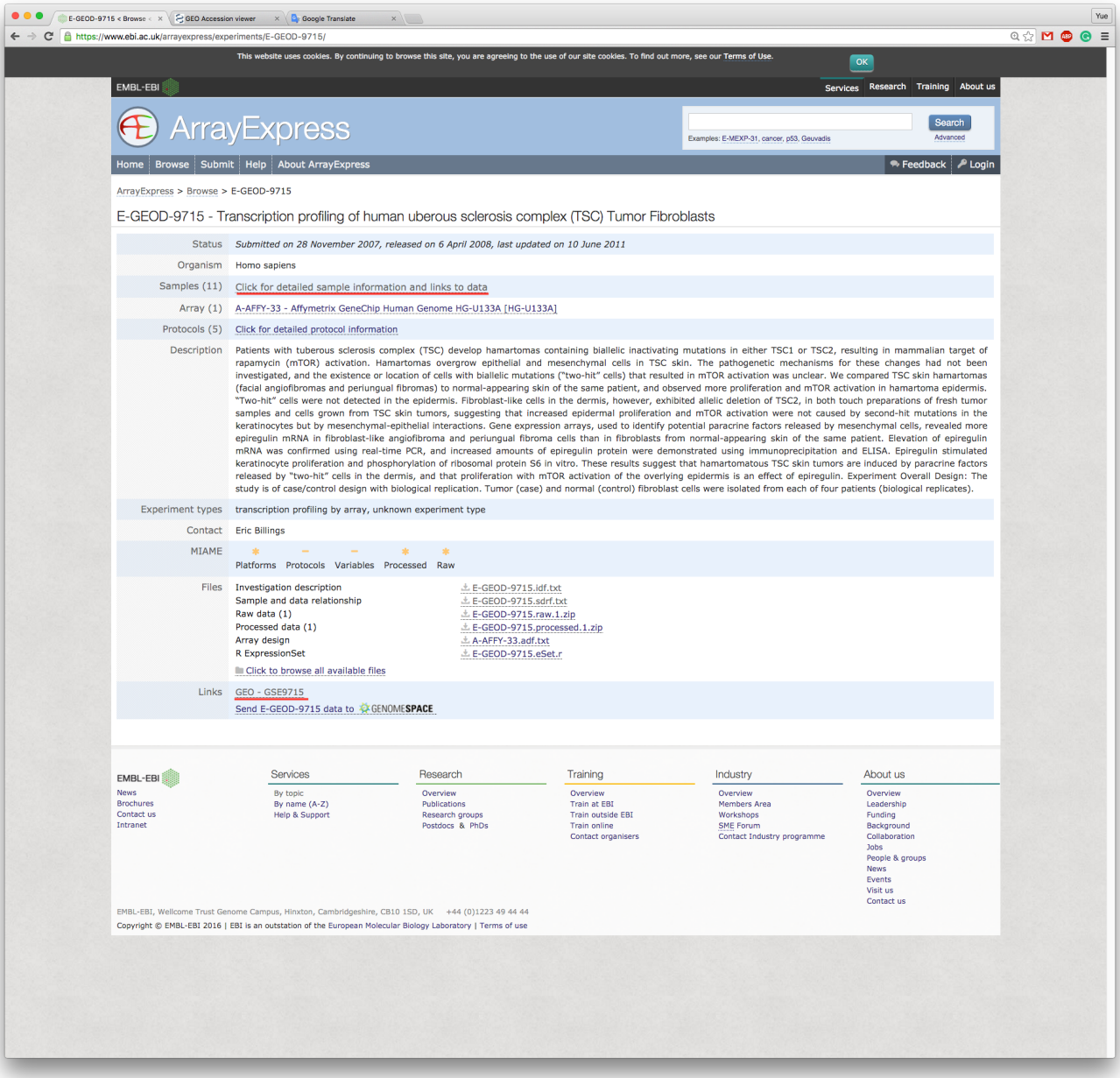
- Creutzfeldt-Jakob Disease, E-GEOD-30643



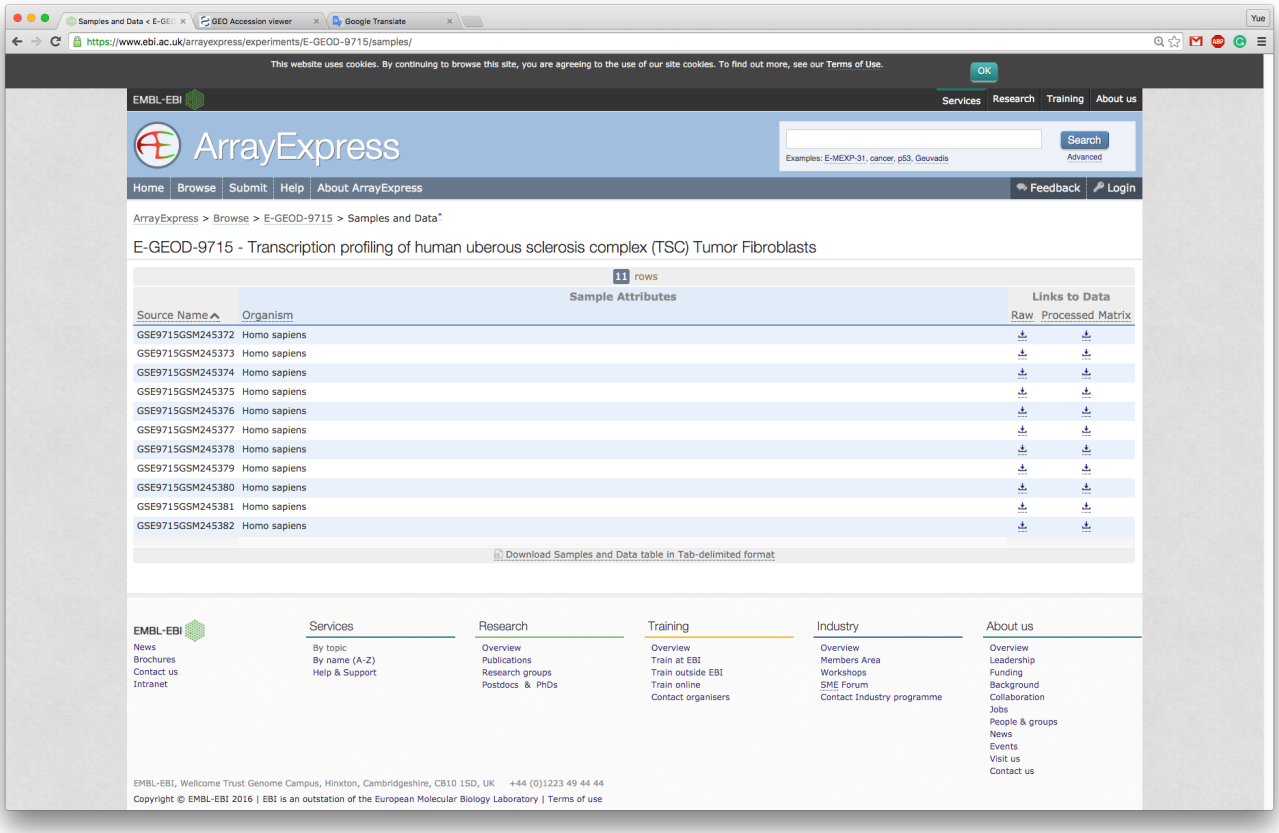
3.2.3 Check if it has disease vs. control

- Tuberous Sclerosis, E-GEOD-9715

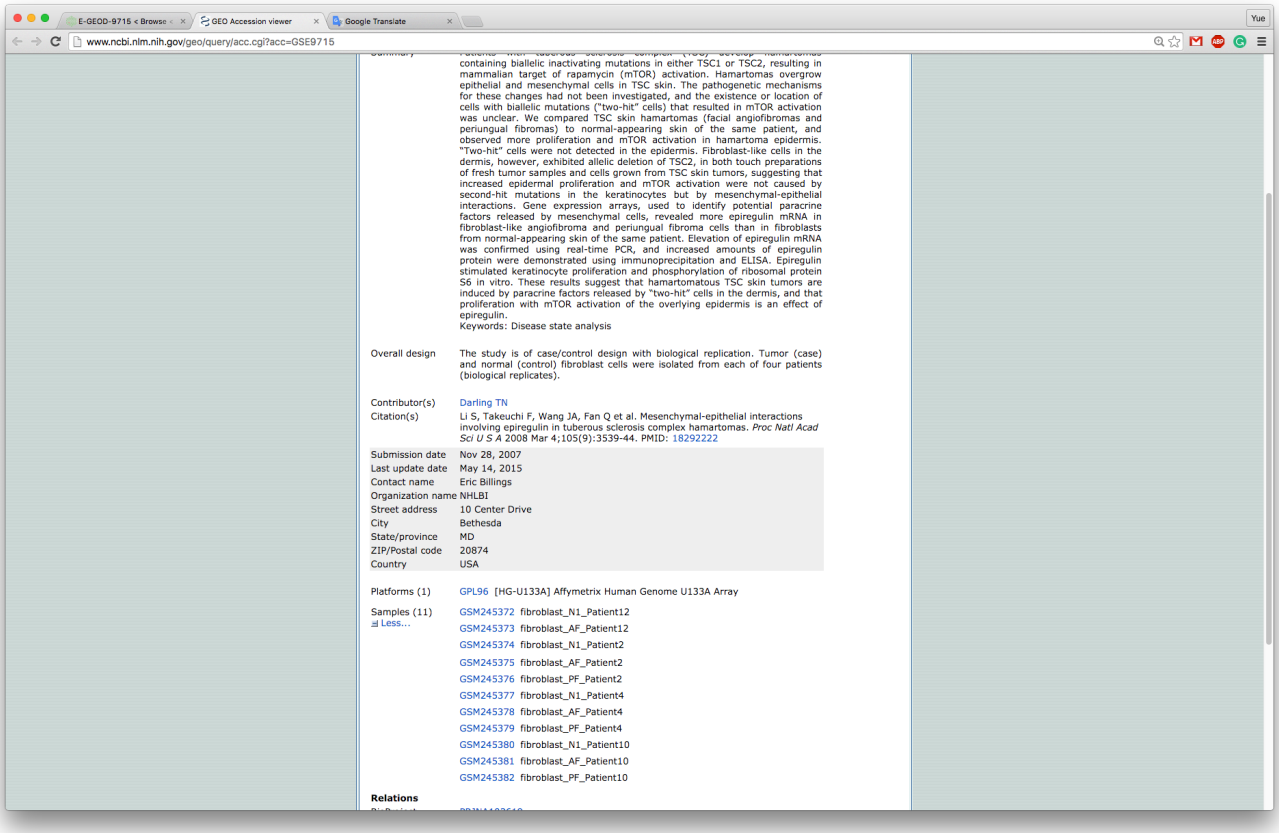
It is hard to determine whether there is disease and control or not sometimes. The following dataset is about a disease called `Tuberous Sclerosis`



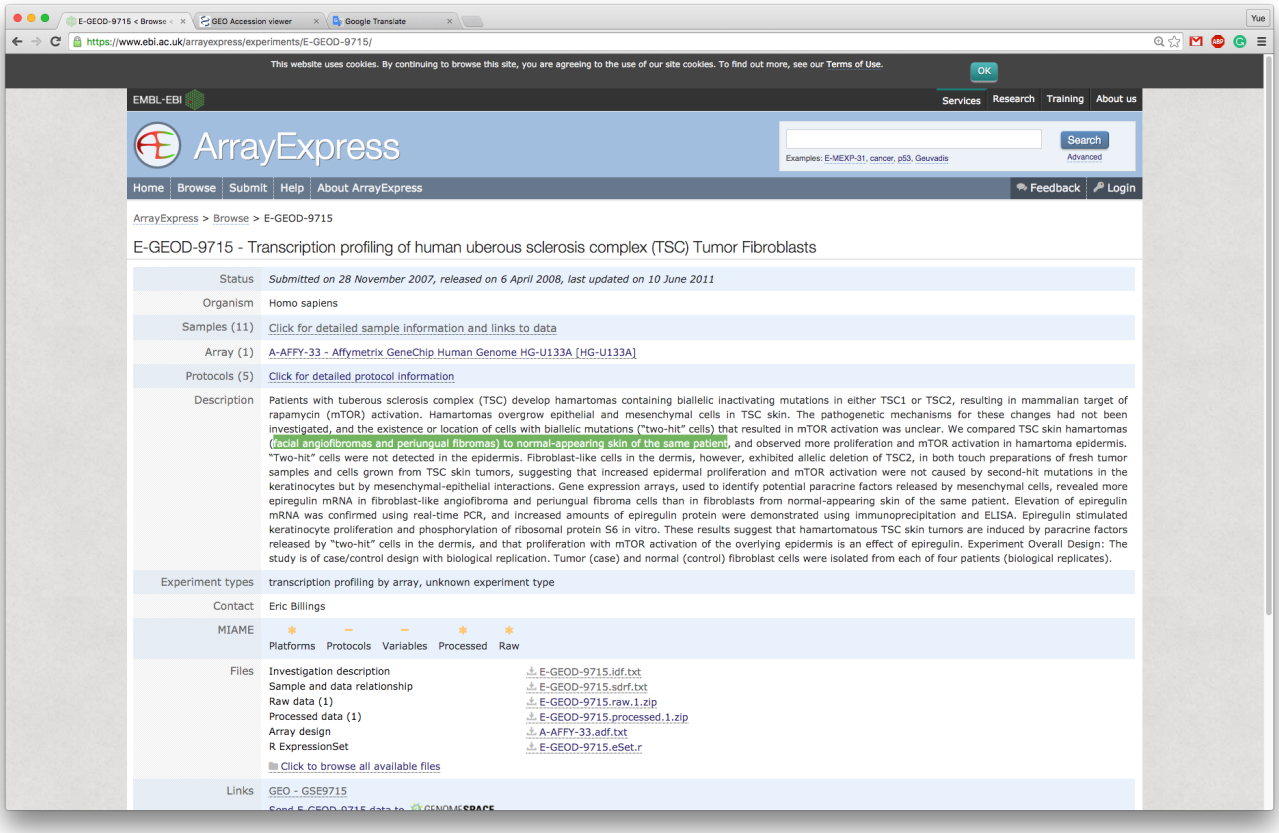
If we check its samples like what we usually do, we will get the following



At this time, we need to check its `GEO` database, by click `Links GEO - GSE9715` at the bottom of its main page



Now, we can know what their samples are come from 'N1', 'AF' and 'PF' patients. To figure out what 'N1', 'AF', 'PF' mean, we need to check it description again

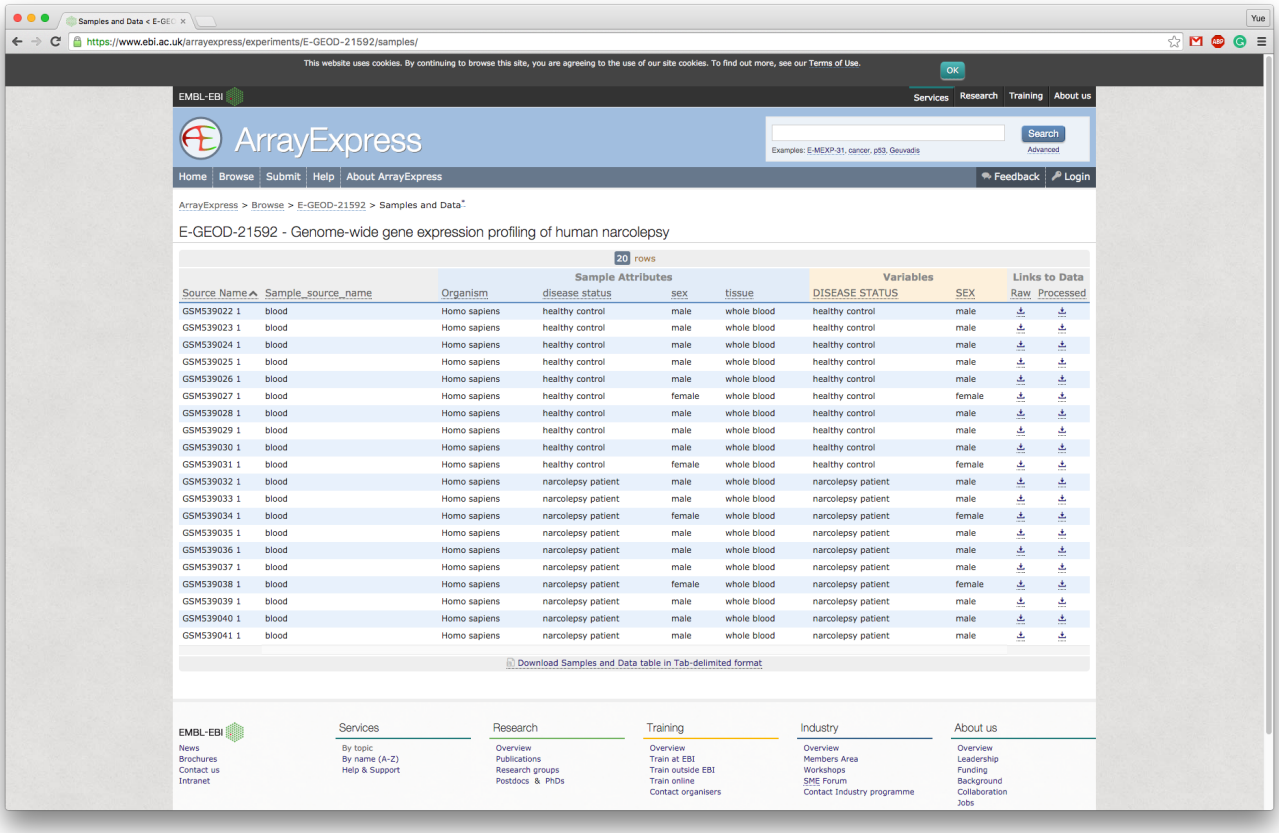


Now, we know they are comparing `facial angiofibromas`(AF) and `periungual fibromas`(PF) to normal-appearing(N1), so it is relevant.

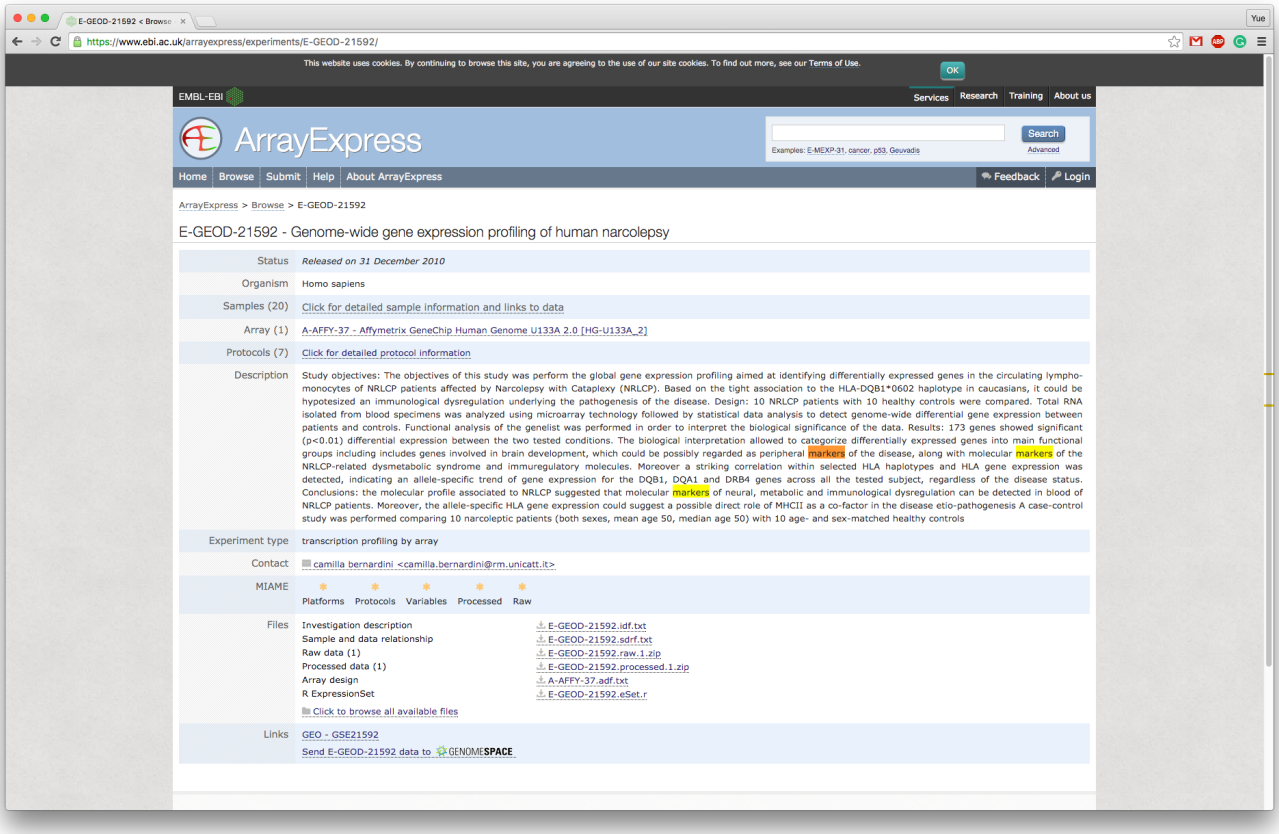
3.2.4 Check if it is `biomarker`

- Narcolepsy, E-GEOD-21592

The following dataset is about `narcolepsy`, which is a neurological disorder. But the dataset source is from blood



So, it is may be a `biomarker`. Check the dataset description now.



You can see the goal of this study is to find the `biomarker`. The disease signature must reflect the underlying biological basis of the disease. For biomarker study, the goal is not to investigate disease mechanisms. Thus, data sets collected for biomarker study can not be used for drug repositioning.

4 Do It Yourself

Check the following datasets:

- Huntington's Disease, E-MTAB-2206
- Dravet Syndrome, E-GEOD-46472
- Tuberous Sclerosis, E-GEOD-6002
- Huntington's Disease, E-GEOD-1751

5 Newly added

- Do not focus too much on confirming whether it is about mRNA. The dataset should be about mRNA by default. Just pay a attention for `MicroRNA(miRNA)`.
- Check whether it is disease vs control:

acute megakaryoblastic leukemia, E-GEOD-16677

Based on the title of this dataset, it is study the Down Syndrome (DS)-AMKL and non-DS AMKL samples. There is no healthy sample.

E-GEOD-16677 - Gene expression profiling of Down Syndrome (DS)-AMKL and non-DS AMKL samples

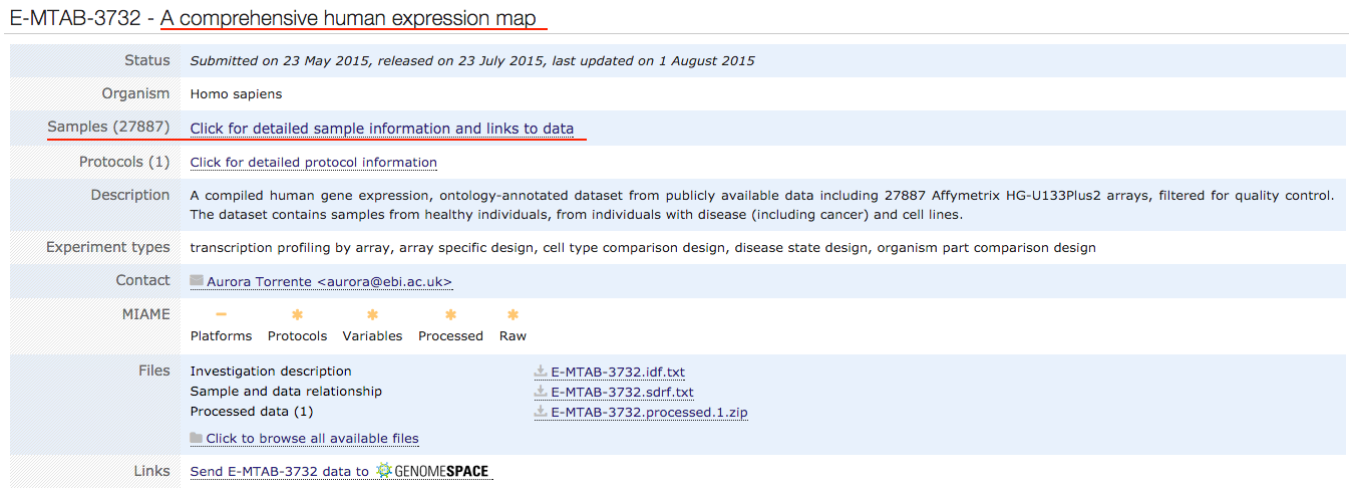
18 rows

					Sample Attributes		Links to Data	
Source Name	Sample_description	Sample_source_name	Organism	cell description	cell line	cell typ	Raw	Processed
GSM417985 1	Gene expression data from sorted primary leukemic cells.	DS-AMKL sample 1, sorted blasts	Homo sapiens	FACS sorted leukemic blasts from patients		megaka		
GSM417986 1	Gene expression data from sorted primary leukemic cells.	DS-AMKL sample 2, sorted blasts	Homo sapiens	FACS sorted leukemic blasts from patients		megaka		
GSM417987 1	Gene expression data from sorted primary leukemic cells.	DS-AMKL sample 3, sorted blasts	Homo sapiens	FACS sorted leukemic blasts from patients		megaka		
GSM417988 1	Gene expression data from sorted primary leukemic cells.	DS-AMKL sample 4, sorted blasts	Homo sapiens	FACS sorted leukemic blasts from patients		megaka		
GSM417989 1	Gene expression data from sorted primary leukemic cells.	DS-AMKL sample 5, sorted blasts	Homo sapiens	FACS sorted leukemic blasts from patients		megaka		
GSM417990 1	Gene expression data from sorted primary leukemic cells.	DS-AMKL sample 6, sorted blasts	Homo sapiens	FACS sorted leukemic blasts from patients		megaka		
GSM417991 1	Gene expression data from sorted primary leukemic cells.	Non-DS AMKL sample 1, sorted blasts	Homo sapiens	FACS sorted leukemic blasts from patients		megaka		
GSM417992 1	Gene expression data from sorted primary leukemic cells.	Non-DS AMKL sample 2, sorted blasts	Homo sapiens	FACS sorted leukemic blasts from patients		megaka		
GSM417993 1	Gene expression data from sorted primary leukemic cells.	Non-DS AMKL sample 3, sorted blasts	Homo sapiens	FACS sorted leukemic blasts from patients		megaka		
GSM417994 1	Gene expression data from sorted primary leukemic cells.	Non-DS AMKL sample 4, sorted blasts	Homo sapiens	FACS sorted leukemic blasts from patients		megaka		
GSM417995 1	Gene expression data from sorted primary leukemic cells.	Non-DS AMKL sample 5, sorted blasts	Homo sapiens	FACS sorted leukemic blasts from patients		megaka		
GSM417996 1	Gene expression data from a non-DS AMKL cell line.	Non-DS AMKL cell line UT-7	Homo sapiens	AMKL cell line	UT-7			
GSM417997 1	Gene expression data from a DS-AMKL cell line.	DS-AMKL cell line CMK	Homo sapiens	DS-AMKL cell line	CMK			
GSM417998 1	Gene expression data from a non-DS AMKL cell line.	Non-DS AMKL cell line UT-7	Homo sapiens	AMKL cell line	UT-7			
GSM417999 1	Gene expression data from a DS-AMKL cell line.	DS-AMKL cell line CMK	Homo sapiens	DS-AMKL cell line	CMK			
GSM418000 1	Gene expression data from a non-DS AMKL cell line.	AMKL cell line Meg01	Homo sapiens	AMKL cell line	MEG01			
GSM418001 1	Gene expression data from a non-DS AMKL cell line.	AMKL cell line M07	Homo sapiens	AMKL cell line	M07			
GSM418002 1	Gene expression data from a erythroleukemia cell line.	Erythroleukemia cell line K562	Homo sapiens	Erythroleukemia cell line	K562			

Download Samples and Data table in Tab-delimited format

- If you see a comprehensive dataset, mark as `comprehensive`.

Acute myeloblastic leukemia with maturation, E-MTAB-3732



E-MTAB-3732 - A comprehensive human expression map

Status	*Submitted on 23 May 2015, released on 23 July 2015, last updated on 1 August 2015*
Organism	Homo sapiens
Samples (27887)	Click for detailed sample information and links to data
Protocols (1)	Click for detailed protocol information
Description	A compiled human gene expression, ontology-annotated dataset from publicly available data including 27887 Affymetrix HG-U133Plus2 arrays, filtered for quality control. The dataset contains samples from healthy individuals, from individuals with disease (including cancer) and cell lines.
Experiment types	transcription profiling by array, array specific design, cell type comparison design, disease state design, organism part comparison design
Contact	Aurora Torrente <aurora@ebi.ac.uk>
MIAME	– Platforms * Protocols * Variables * Processed * Raw
Files	Investigation description E-MTAB-3732.idf.txt Sample and data relationship E-MTAB-3732.sdrf.txt Processed data (1) E-MTAB-3732.processed.1.zip Click to browse all available files
Links	Send E-MTAB-3732 data to GENOMESPACE

- If the dataset is about methylation, it is irrelevant.

Ulcerative colitis, E-GEOD-27899

E-GEOD-27899 - Methylation profiling in Ulcerative colitis

Status	*Released on 27 September 2012, last updated on 2 May 2014*
Organism	Homo sapiens
Samples (40)	Click for detailed sample information and links to data
Arrays (2)	A-GEOD-8490 - Illumina HumanMethylation27 BeadChip (HumanMethylation27_270596_v.1.2) A-GEOD-13269 - Nimblegen Human high-density promoter array CpG island methylation array
Protocols (12)	Click for detailed protocol information
Description	The series was designed to identify the different methylated single CpGs involved in the pathophysiology of ulcerative colitis. A cohort of n=20 monozygotic twins, discordant for ulcerative colitis was selected. Illumina and Nimblegen platforms were used.
Experiment type	methylation profiling by array
Contacts	Zhe Feng <geo@ncbi.nlm.nih.gov>, Robert Häsler
MIAME	* Platforms – Protocols * Variables * Processed – Raw
Files	Investigation description: E-GEOD-27899.idf.txt Sample and data relationship: E-GEOD-27899.sdrf.txt Raw data (2): E-GEOD-27899.raw.1.zip, E-GEOD-27899.raw.2.zip Processed data (1): E-GEOD-27899.processed.1.zip Array designs: A-GEOD-13269.adf.txt, A-GEOD-8490.adf.txt Click to browse all available files
Links	GEO - GSE27899 Send E-GEOD-27899 data to GENOMESPACE