BTEP Workshop 2016: Using NCBI's Gene Expression Omnibus (GEO) to Explore Gene Expression; June 6, 2016

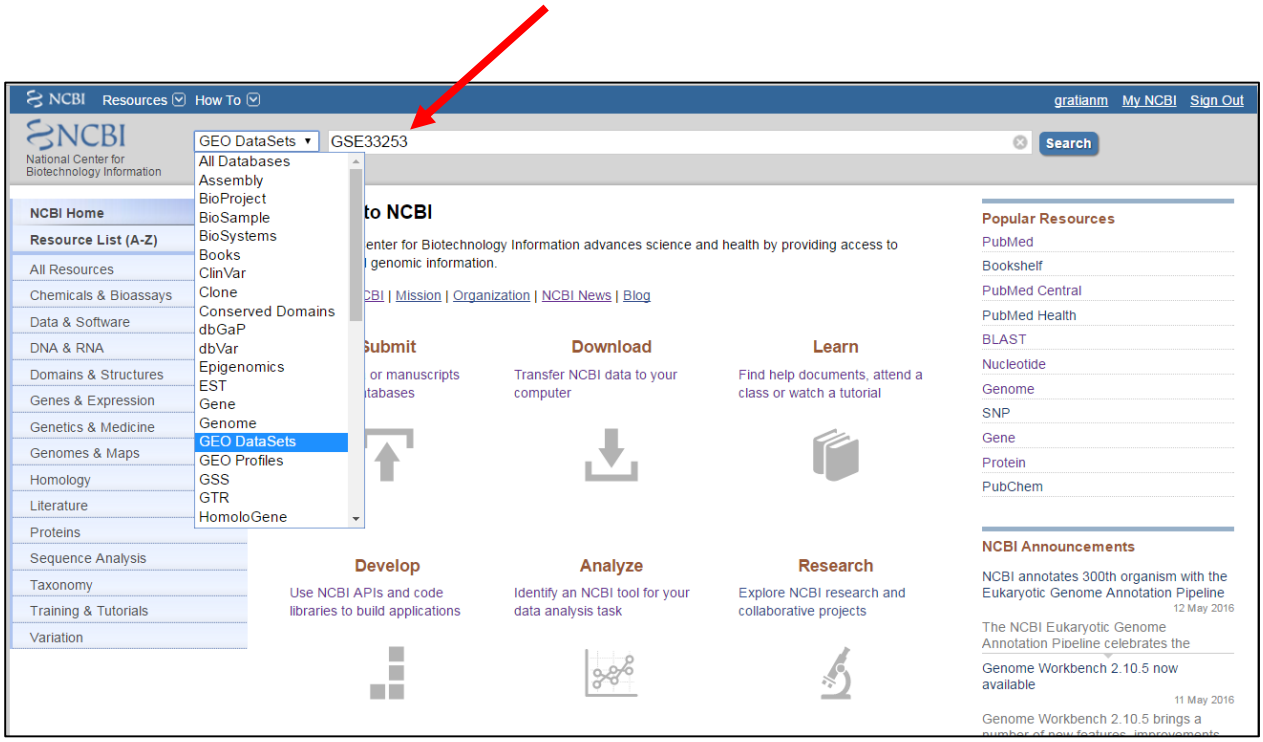
Exercise 1: Explore a Study that was submitted to GEO

(Independent practice)

1. Access the GSE33253 study and answer any questions along the way

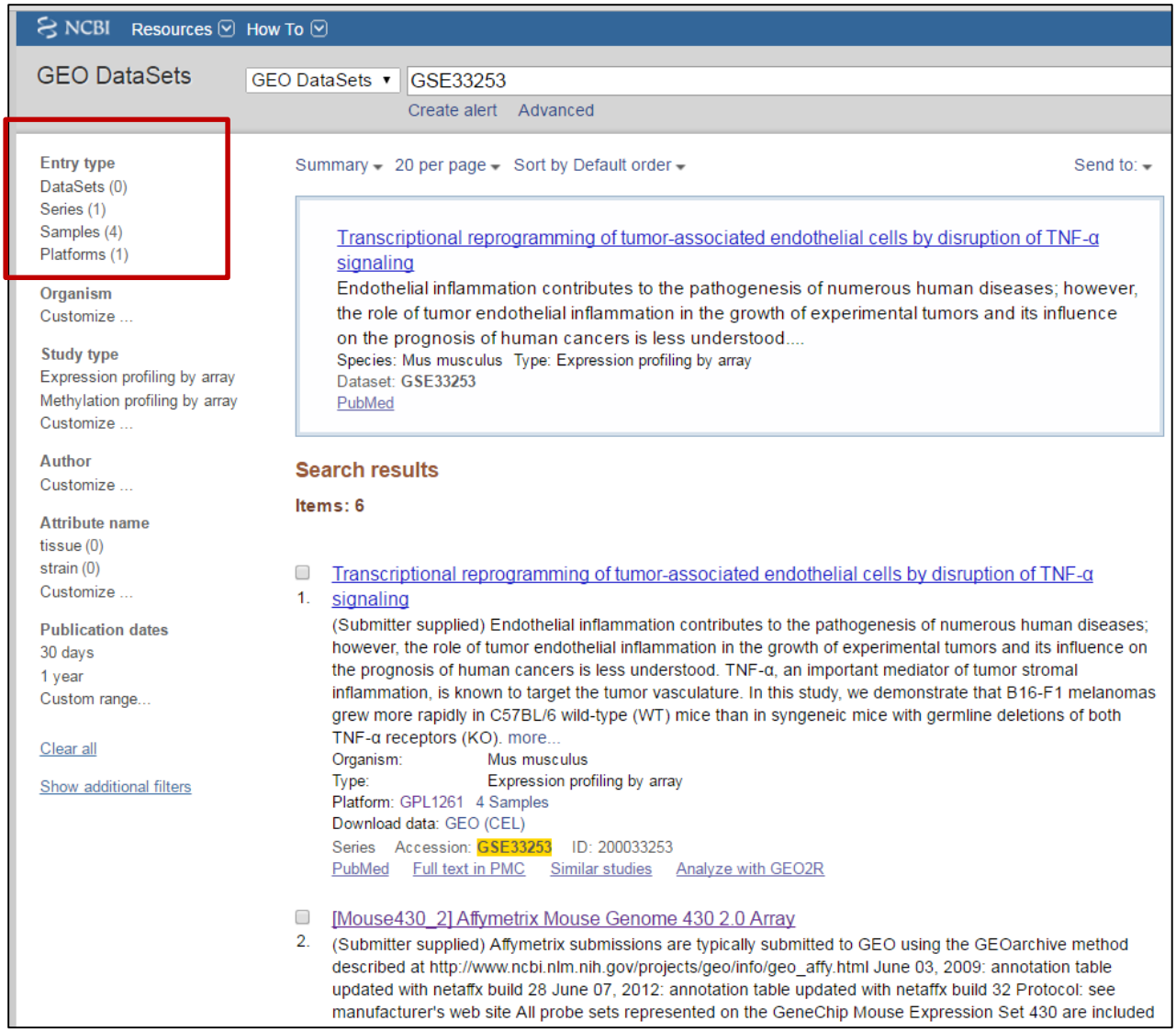
   – On the NCBI home page select the **GEO DataSets** database from the **All Databases** pull down menu.
   Q: What is the name of another GEO database that you notice on the pull down menu? _____
   (You will explore this second database later.)

   – You can search the database with a text term, or a GEO accession number. Here, enter the following GEO accession number: **GSE33253** and click on the **Search** button.



2. Examine your search results

   – How many records have you retrieved in total? _____
   – Note the **Entry type** filter on the left side of the screen.
   Q: What type of entries do you see?_____



3. Examine individual records

a) Series

- Use the **Entry type** filter on the left side of the screen. Click on **Series** to select the series record that describes the overall design of the study.
  Q: Looking at the summary of the series record, can you confirm that the technology used in the study was indeed **Expression profiling by array**?
  _____
- Click on the series record title to open up and examine the record itself.
  Q: Would you agree that the series record describes the overall design of the study? _____

b) Platform

– Go back in your browser to the search page and deselect **Series** and select **Platforms**.
Q: Looking at the summary of the Platform record, what is the Platform Accession number (Hint: it is listed at the bottom of the summary)?
_____

– Click on the Platform record title to examine the record itself.
Q: Is this a commercial array or a custom-made array? _____
Q: Who is the manufacturer of the array?_____
Q: Was this array used in other studies and how many?
(Hint: scroll down the page to find how many series are listed for the platform. For this purpose you can assume 1 series = 1 study but be aware that it is a bit more complex as there are super series that include several series.) _____

– Scroll down the Platform record (past **Relations** that indicate that this particular array relates to many other arrays) to access the **Data table**:

	best associated with the transcribed region being interrogated by the probe set. Refer to the "Sequence Source" field to determine the database used.
Gene Title	Title of Gene represented by the probe set.
Gene Symbol	A gene symbol, when one is available (from UniGene).
ENTREZ_GENE_ID	Entrez Gene Database UID
RefSeq Transcript ID	References to multiple sequences in RefSeq. The field contains the ID and Description for each entry, and there can be multiple entries per ProbeSet.
Gene Ontology Biological Process	Gene Ontology Consortium Biological Process derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence". The description corresponds directly to the GO ID. The evidence can be "direct", or "extended".
Gene Ontology Cellular Component	Gene Ontology Consortium Cellular Component derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence". The description corresponds directly to the GO ID. The evidence can be "direct", or "extended".
Gene Ontology Molecular Function	Gene Ontology Consortium Molecular Function derived from LocusLink. Each annotation consists of three parts: "Accession Number // Description // Evidence". The description corresponds directly to the GO ID. The evidence can be "direct", or "extended".

Data table

ID	GB_ACC	SPOT_ID	Species Scientific Name	Annotation Date	Sequence Type	Sequence Source	Target Description
1415670_at	BC024686		Mus musculus	Jun 9, 2011	Consensus sequence	GenBank	gb:BC024686.1 /DB_XREF=gi:19354080 /FEA=FLmRNA
1415671_at	NM_013477		Mus musculus	Jun 9, 2011	Consensus sequence	GenBank	gb:NM_013477.1 /DB_XREF=gi:7304908 /GEN=Atp6v0
1415672_at	NM_020585		Mus musculus	Jun 9, 2011	Consensus sequence	GenBank	gb:NM_020585.1 /DB_XREF=gi:10181207 /GEN=AB04
1415673_at	NM_133900		Mus musculus	Jun 9, 2011	Consensus sequence	GenBank	gb:NM_133900.1 /DB_XREF=gi:19527115 /GEN=AI480
1415674_a_at	NM_021789		Mus musculus	Jun 9, 2011	Consensus sequence	GenBank	gb:NM_021789.1 /DB_XREF=gi:11140824 /GEN=Sbdn
1415675_at	BC008256		Mus musculus	Jun 9, 2011	Consensus sequence	GenBank	gb:BC008256.1 /DB_XREF=gi:14198389 /FEA=FLmRNA
1415676_a_at	NM_011186		Mus musculus	Jun 9, 2011	Consensus sequence	GenBank	gb:NM_011186.1 /DB_XREF=gi:6755203 /GEN=Psmb5
1415677_at	NM_026819		Mus musculus	Jun 9, 2011	Consensus sequence	GenBank	gb:NM_026819.1 /DB_XREF=gi:13386167 /GEN=1110
1415678_at	BC008595		Mus musculus	Jun 9, 2011	Consensus sequence	GenBank	gb:BC008595.1 /DB_XREF=gi:14250330 /FEA=FLmRNA
1415679_at	NM_025498		Mus musculus	Jun 9, 2011	Consensus sequence	GenBank	gb:NM_025498.1 /DB_XREF=gi:13384921 /GEN=1700
1415680_at	NM_008569		Mus musculus	Jun 9, 2011	Consensus sequence	GenBank	gb:NM_008569.1 /DB_XREF=gi:6678833 /GEN=Mcpr /
1415681_at	NM_053164		Mus musculus	Jun 9, 2011	Consensus sequence	GenBank	gb:NM_053164.1 /DB_XREF=gi:17298675 /GEN=Mrpl4
1415682_at	NM_023045		Mus musculus	Jun 9, 2011	Consensus sequence	GenBank	gb:NM_023045.1 /DB_XREF=gi:12746421 /GEN=Ranbp

- The main purpose of the table is to correspond the microarray identifiers with the gene information (IDs, names, ontology, …) . Each column in the data table contains information as described in the **Data table header descriptions**.
  - Focus on the first two columns. The first column provides the microarray **ID** and the second column tells the accession number of the sequence that was used to design the oligos attached to the spot.
    Q: Can you recognize which of the **GB_ACC** in the second column are the NCBI reference accessions and which one are for primary GenBank records?
    NCBI reference sequence accession example _____
    Primary GenBank record accession example _____
  - Scrolling along the table, find the Gene Symbol and ENTREZ_GENE_ID columns.
- A partial table that corresponds microarray IDs with sequences and genes is copied here (you can refer to this table when working with GEO2R):

Data table

ID	GB_ACC	Gene Symbol	ENTREZ_GENE_ID
1415670_at	BC024686	Copg	54161
1415671_at	NM_013477	Atp6v0d1	11972
1415672_at	NM_020585	Golga7	57437
1415673_at	NM_133900	Psph	100678
1415674_a_at	NM_021789	Trappc4	60409
1415675_at	BC008256	Dpm2	13481
1415676_a_at	NM_011186	Psmb5	19173
1415677_at	NM_026819	Dhrs1	52585
1415678_at	BC008595	Ppm1a	19042
1415679_at	NM_025498	Psenen	66340
1415680_at	NM_008569	Anapc1	17222
1415681_at	NM_053164	Mrpl43	94067
1415682_at	NM_023045	Xpo7	65246
1415683_at	BC016526	Nmt1	18107
1415684_at	AV168389	Atg5	11793
1415685_at	NM_133767	Mtif2	76784
1415686_at	AV339290	Rab14	68365
1415687_a_at	BM212050	Psap	19156
1415688_at	NM_025985	Ube2g1	67128
1415689_s_at	BC007473	Zkscan3	72739

- Back on the web, check the number of rows that are listed for the table.
  Q: If you consider that about 80 rows in the table are dedicated to descriptors and control, approximately how many genes are assayed on this array?_____
- (To find information for your gene of interest you would need to download and parse the full table.)

c) Samples

- From the Platform record, go back in the browser to access the search results and uncheck Platforms and check **Samples**.
  Q: Each of the samples have its own accession (at the bottom of each of the summary. What three letters are used in the sample accession format?
  _____

- To check one of the samples click, for example, on the **Wild-type-2** title to access the information in the record. Confirm that you are able to find the growth- and sample preparation protocols in the record.

- Scroll to the bottom of the record to the **Data table** and note that the table is available for download.

- Click on the **View full table…** button. In most browsers, the table should open in a new tab. Scroll past the "AFFX-…" entries that are various controls, until you come to the spot IDs that you will recognize because you have just seen them in the Platform table.

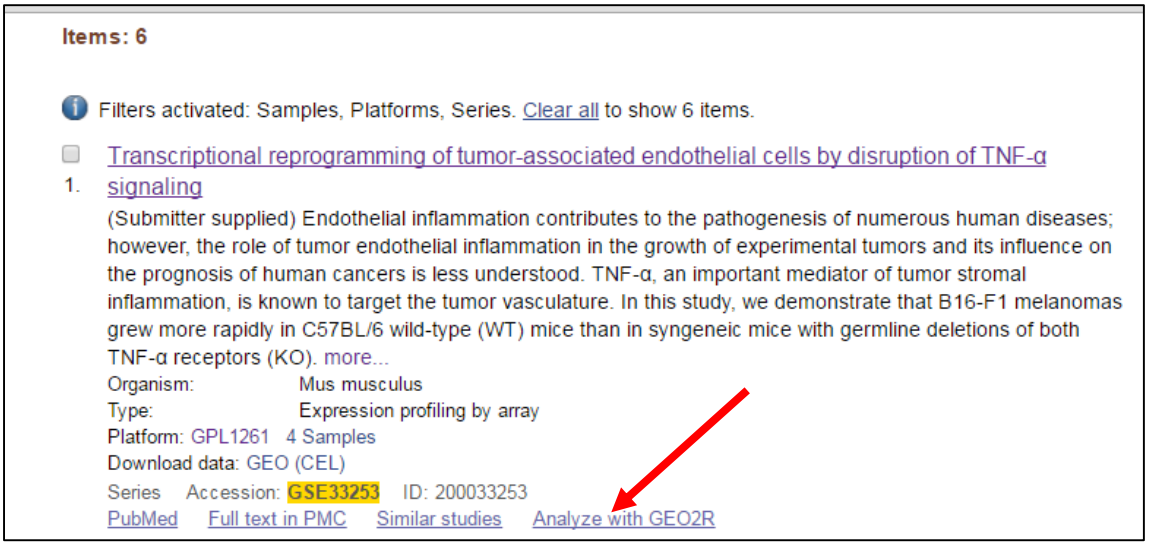
  Q: What was the expression value for the Dpm2 gene? _____
  (Hint: Help yourself with the simplified platform table on page 4 of this document or the platform record on the web)

  Q: Is the Dpm gene expression value for this sample higher or lower than those for the preceding spots/genes in the table?_____

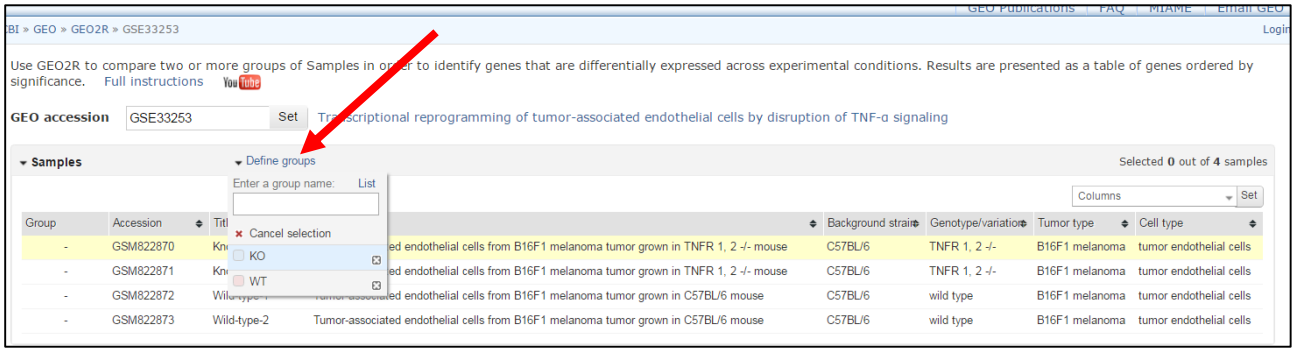
4. Gene expression analysis on demand

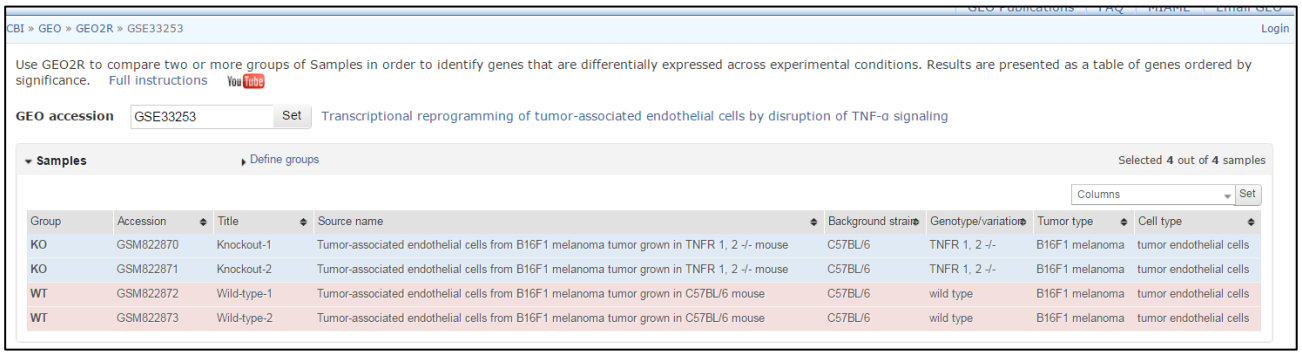
(Guided practice)

– **GSE33253** is not a curated study (you will learn about GEO curation shortly), but you can still analyze the data on demand by using the **GEO2R** web analysis tool. The **GSE33253** series record provides the **Analyze with GEO2R** link to the tool.



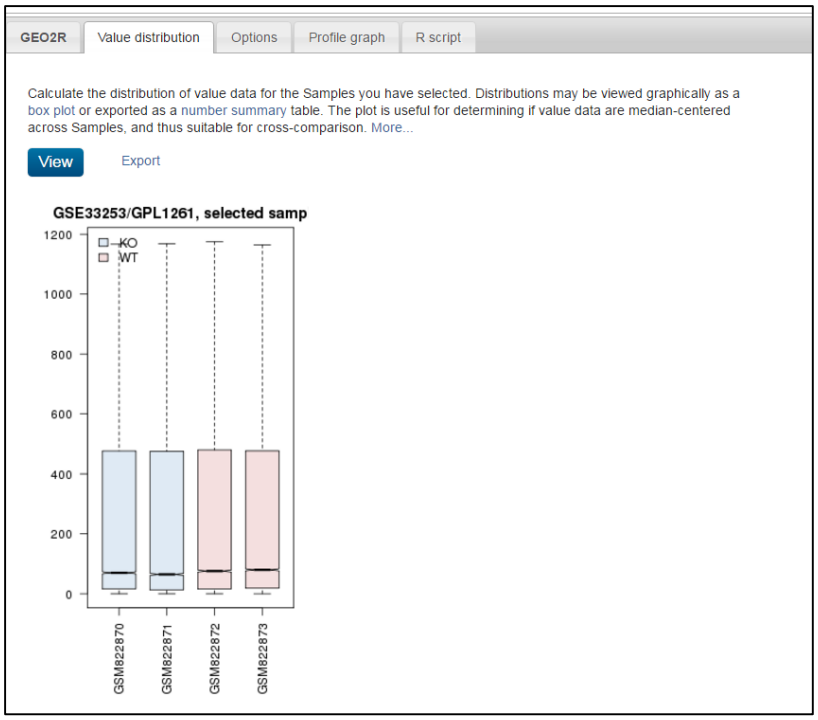
– **Define groups:** Note that this study only have two treatments, so you do not have options. In other studies with several treatments/samples, GEO2R allows you to define your own groups and select samples for comparison.



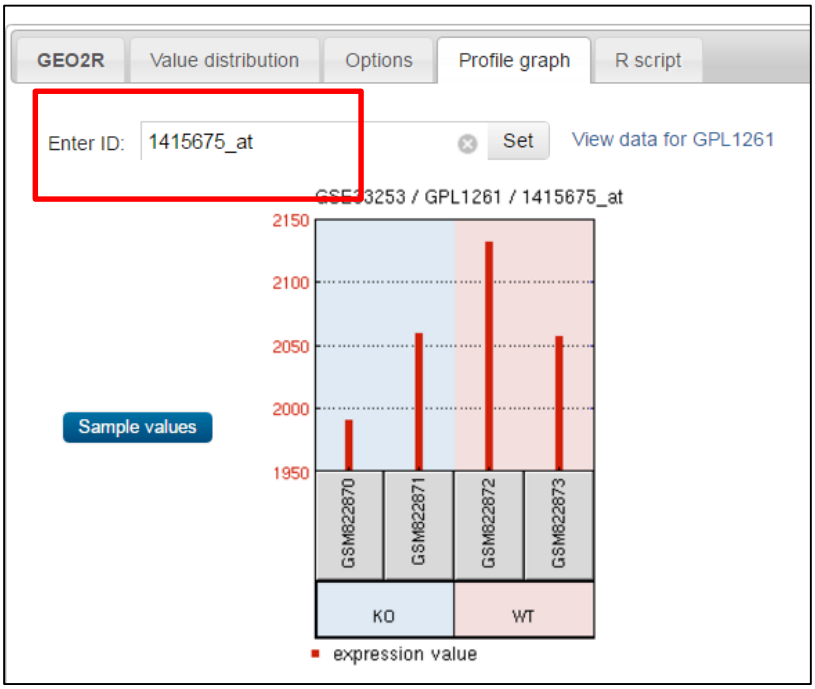


If you are going to use GEO2R in your research, please read Full Instructions!

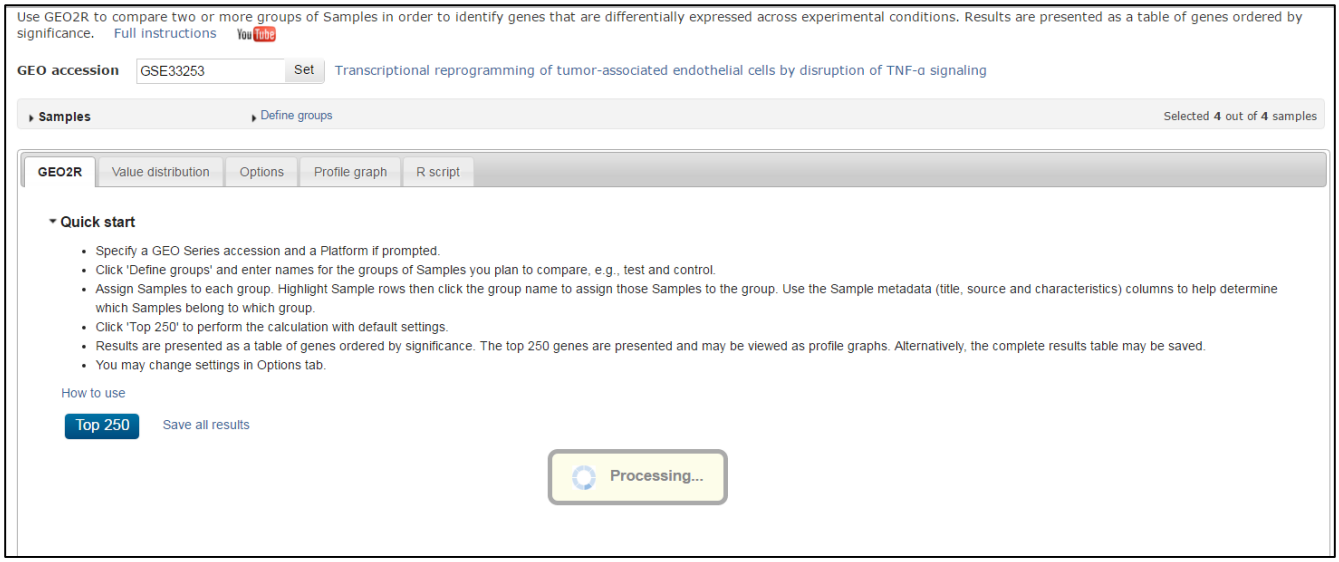
- **View Value distribution:** This tells you if the samples are centered at the median.



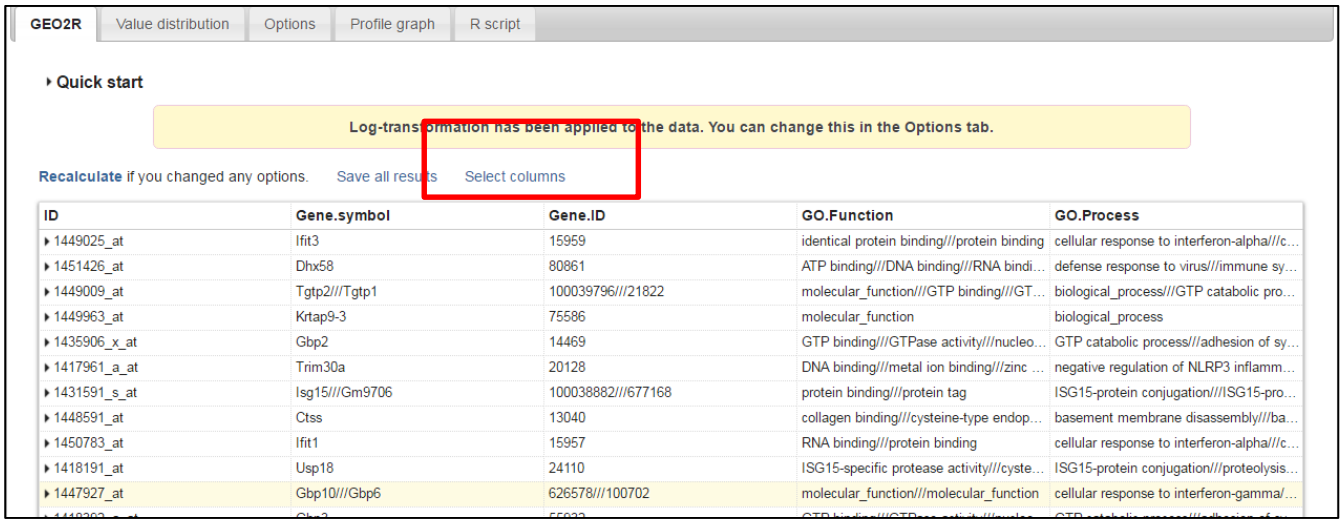
- Check **Options** (refer to Full Instructions)
- Find the Profile graph for the Dpm2 (1415675_at) gene:



– **Run GEO2R**:



– On the results page note that Log-transformation has been applies to the data.
– **Select columns** so that you display the gene symbol and the GO. Function and GO. Process columns



ID	Gene.symbol	Gene.ID	GO.Function	GO.Process
1449025_at	Ifit3	15959	identical protein binding///protein binding	cellular response to interferon-alpha///c…
1451426_at	Dhx58	80861	ATP binding///DNA binding///RNA bindi…	defense response to virus///immune sy…
1449009_at	Tgtp2///Tgtp1	100039796///21822	molecular_function///GTP binding///GT…	biological_process///GTP catabolic pro…
1449963_at	Krtap9-3	75586	molecular_function	biological_process
1435906_x_at	Gbp2	14469	GTP binding///GTPase activity///nucleo…	GTP catabolic process///adhesion of sy…
1417961_a_at	Trim30a	20128	DNA binding///metal ion binding///zinc …	negative regulation of NLRP3 inflamm…
1431591_s_at	Isg15///Gm9706	100038882///677168	protein binding///protein tag	ISG15-protein conjugation///ISG15-pro…
1448591_at	Ctss	13040	collagen binding///cysteine-type endop…	basement membrane disassembly///ba…
1450783_at	Ifit1	15957	RNA binding///protein binding	cellular response to interferon-alpha///c…
1418191_at	Usp18	24110	ISG15-specific protease activity///cyste…	ISG15-protein conjugation///proteolysis…
1447927_at	Gbp10///Gbp6	626578///100702	molecular_function///molecular_function	cellular response to interferon-gamma/…

(Independent practice)

– Many of the top five differentially expressed genes (skip Krtap9-3 for which GO annotation is missing) have at least a couple of things in common.
Q: Is the expression of the top five genes higher in KO or WT? What is the common function of these genes (the GO information is copied below because it is hard to read from the web)_____

"1449025_at" "Ifit3" "15959"
"identical protein binding///protein binding"
"cellular response to interferon-alpha///cellular response to interferon-beta///defense response to virus///immune system process///innate immune response"

"1451426_at" "Dhx58" "80861"
"ATP binding///DNA binding///RNA binding///double-stranded RNA binding///double-stranded RNA binding///helicase activity///hydrolase activity///hydrolase activity, acting on acid anhydrides///metal ion binding///nucleotide binding///single-stranded RNA binding///zinc ion binding"

"defense response to virus///immune system process///innate immune response///negative regulation of MDA-5 signaling pathway///negative regulation of RIG-I signaling pathway///negative regulation of RIG-I signaling pathway///negative regulation of innate immune response///negative regulation of innate immune response///negative regulation of type I interferon production///negative regulation of type I interferon production///positive regulation of MDA-5 signaling pathway///positive regulation of RIG-I signaling pathway///positive regulation of type I interferon production///regulation of innate immune response///response to virus"

"1449009_at" "Tgtp2///Tgtp1" "100039796///21822"
"molecular_function///GTP binding///GTPase activity"
"biological_process///GTP catabolic process///cellular response to interferon-beta///response to interferon-alpha///response to interferon-gamma"

"1449963_at" "Krtap9-3" "75586"
"molecular_function"
"biological_process"

"1435906_x_at" "Gbp2" "14469"

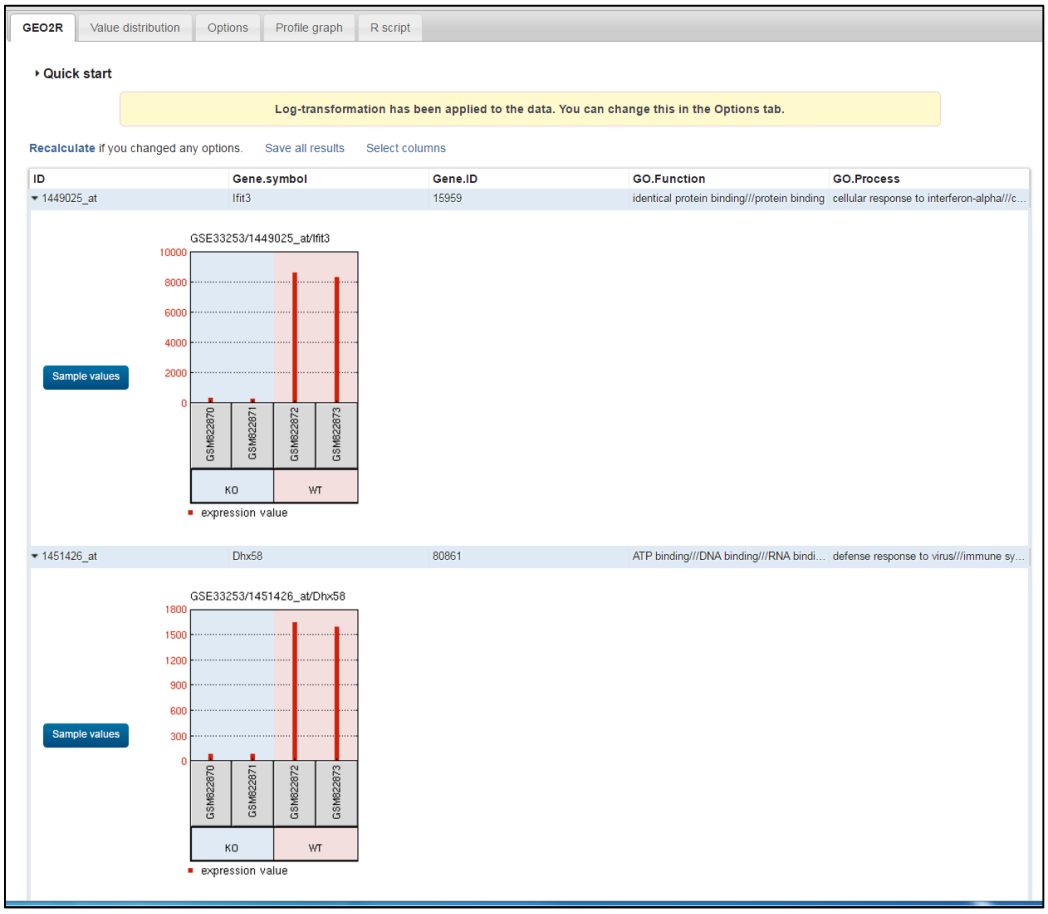
"GTP binding///GTPase activity///nucleotide binding"

"GTP catabolic process///adhesion of symbiont to host///cellular response to interferon-beta///cellular response to interferon-gamma///cellular response to lipopolysaccharide///defense response to Gram-positive bacterium///defense response to protozoan"

"1417961_a_at" "Trim30a" "20128"

"DNA binding///metal ion binding///zinc ion binding"

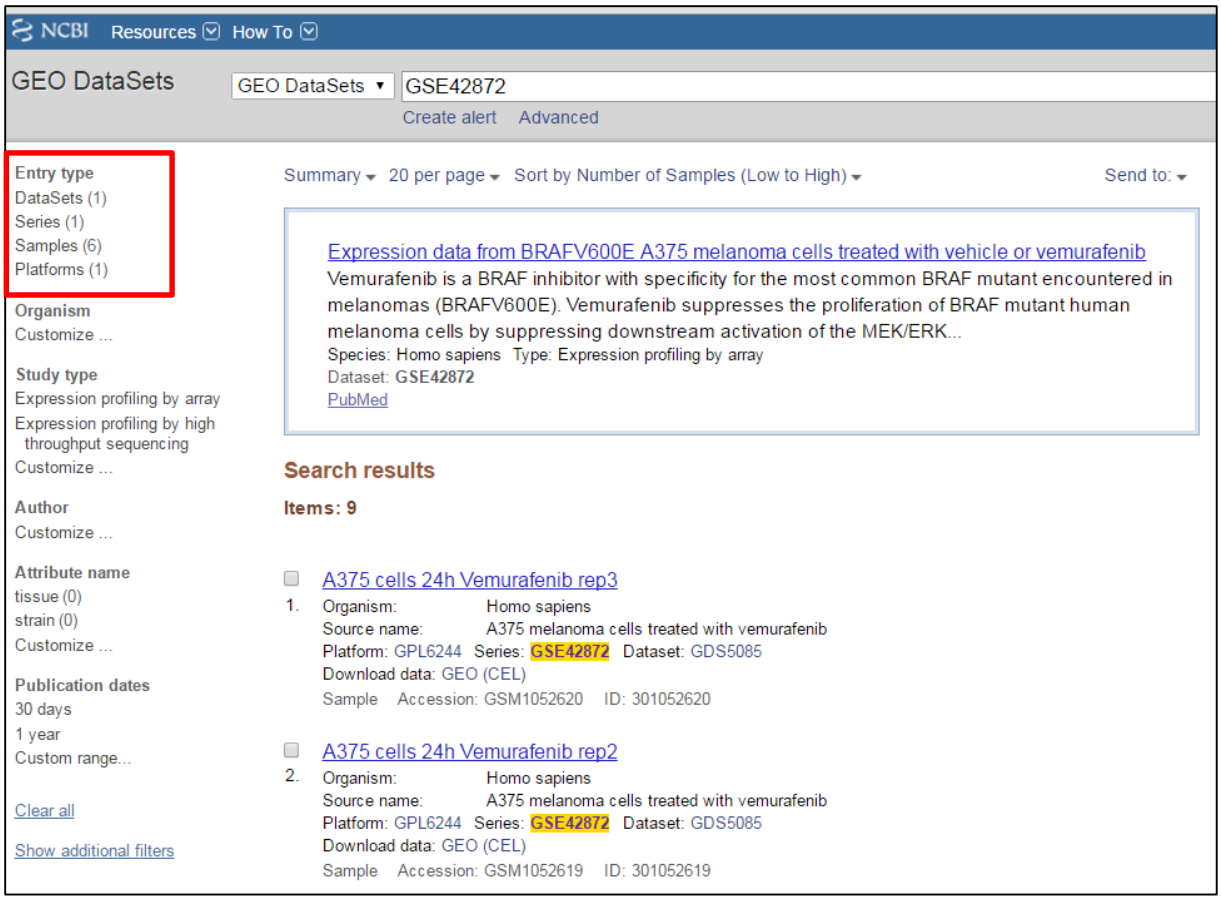
"negative regulation of NLRP3 inflammasome complex assembly///negative regulation of interleukin-6 production///negative regulation of reactive oxygen species metabolic process///negative regulation of toll-like receptor signaling pathway///negative regulation of tumor necrosis factor production///positive regulation of protein catabolic process///positive regulation of viral entry into host cell///protein autoubiquitination///regulation of transcription, DNA-templated///transcription, DNA-templated"



Exercise 2: Explore a Curated GEO DataSets record: Analysis Tools, Individual Gene Profiles, and the Gene- and Biosystems databases

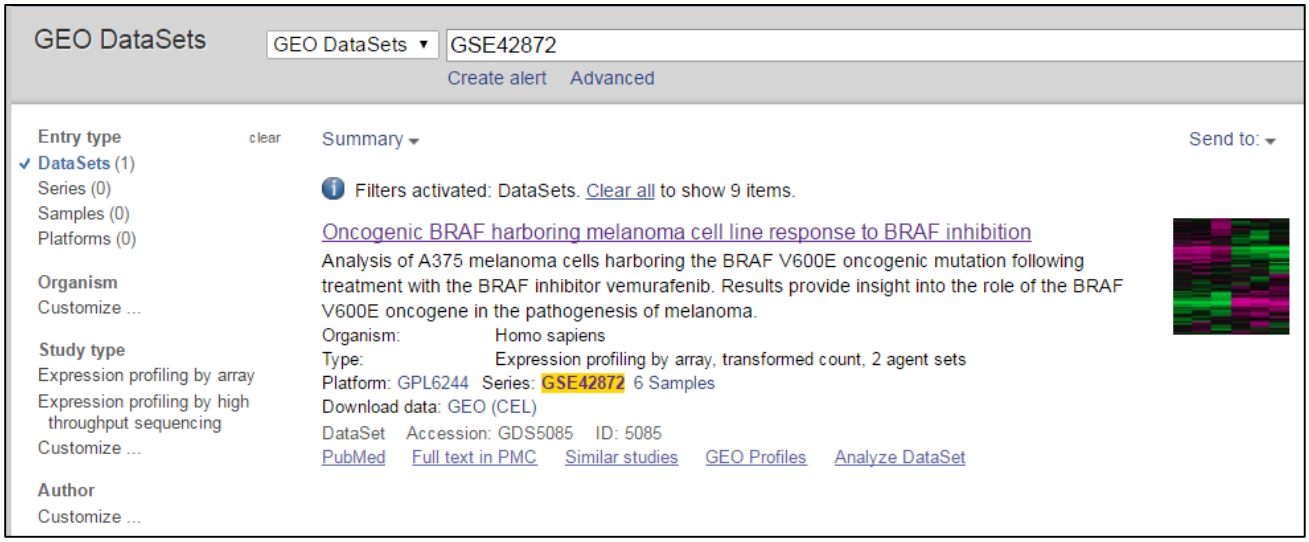
(Guided practice)

1. Search for GSE42872 and examine your search results:



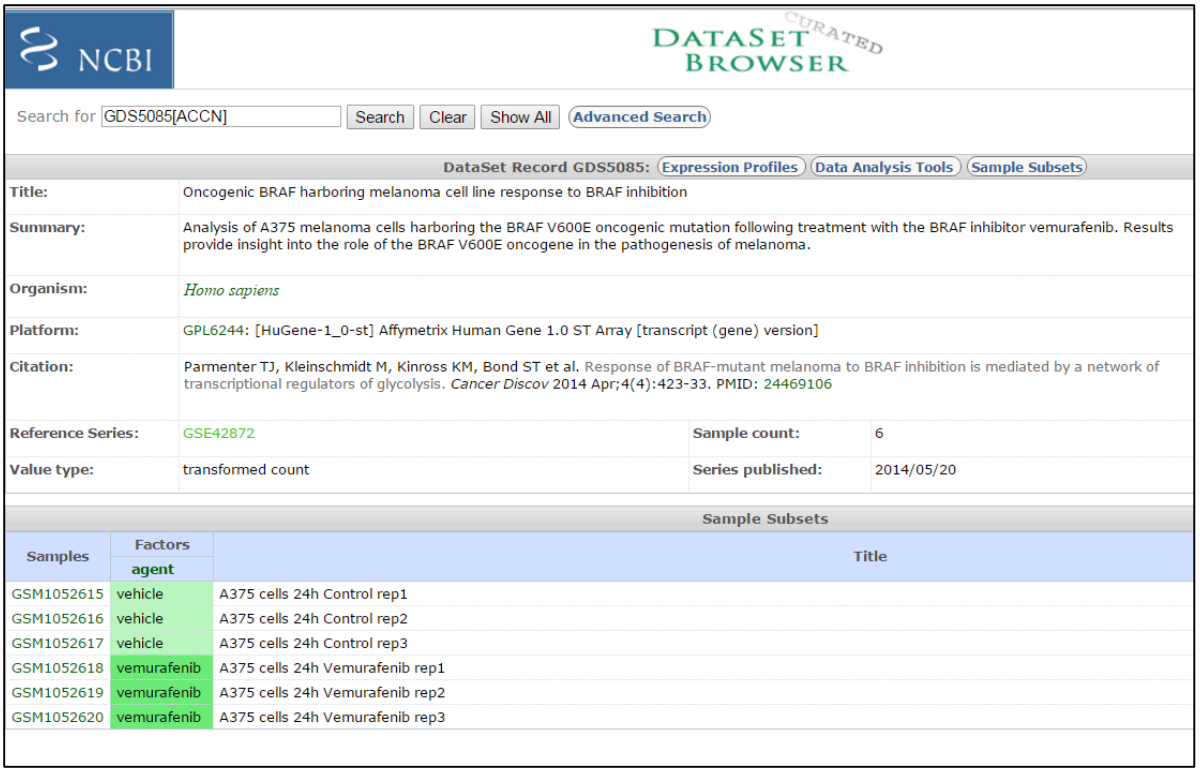
- Clear any of the set filters from the previous search.
- In addition to the submitted records (series, sample, and platform) you will also retrieve a **DataSet** record that was generated as a product of GEO curation.
- The study included six samples. Each of the sample was the A375 cell line that endogenously expresses a mutant BRAF protein. The mutation is designated as BRAF(V600E), meaning that the gene is mutated so that valine at position 600 of the protein expressed by the BRAF gene is replaced by glutamic acid. BRAF(V600E) is an oncogenic mutation: http://www.ncbi.nlm.nih.gov/clinvar/variation/13961/

- In the study, three of the samples were treated with vemurafenib (https://pubchem.ncbi.nlm.nih.gov/compound/42611257) that is an inhibitor of the BRAF(V600E) kinase. (Control in this case means that the mutant kinase stays active in the cell).
- Select the DataSets record:
  - Check the summary display (do not go to the record just yet) and note the accession number of the curated dataset record: **GDS5085**



2. Analysis of a curated set

- Click on the title of the curated study to access and examine the record:
- The **Sample Subset** button displays the sample information.
- **Data Analysis Tools** allows for visualizing clusters of genes with similar expression, and find up and down regulated genes.
- The **Expression Profiles** button leads in the **Geo Profiles** database for either all of the individual-gene profiles in the study, or those that are up/down in a sample comparison.



NCBI DATASET CURATED BROWSER

Search for GDS5085[ACCN] Search Clear Show All Advanced Search

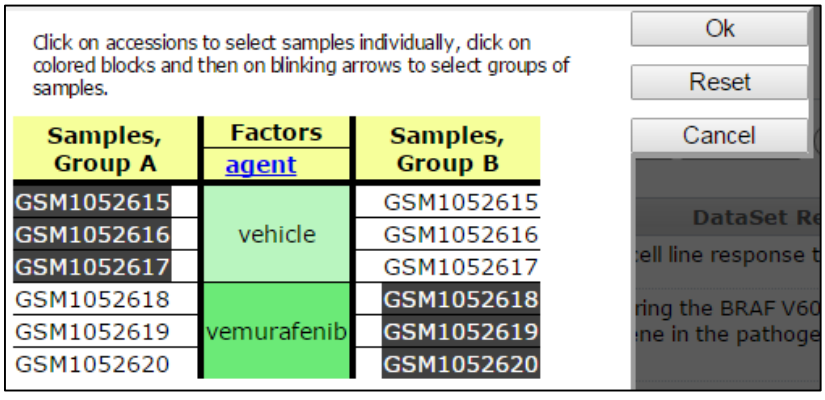
DataSet Record GDS5085: Expression Profiles Data Analysis Tools Sample Subsets

Title:	Oncogenic BRAF harboring melanoma cell line response to BRAF inhibition
Summary:	Analysis of A375 melanoma cells harboring the BRAF V600E oncogenic mutation following treatment with the BRAF inhibitor vemurafenib. Results provide insight into the role of the BRAF V600E oncogene in the pathogenesis of melanoma.
Organism:	*Homo sapiens*
Platform:	GPL6244: [HuGene-1_0-st] Affymetrix Human Gene 1.0 ST Array [transcript (gene) version]
Citation:	Parmenter TJ, Kleinschmidt M, Kinross KM, Bond ST et al. Response of BRAF-mutant melanoma to BRAF inhibition is mediated by a network of transcriptional regulators of glycolysis. *Cancer Discov* 2014 Apr;4(4):423-33. PMID: 24469106
Reference Series:	GSE42872
Sample count:	6
Value type:	transformed count
Series published:	2014/05/20

Sample Subsets

Samples	Factors: agent	Title
GSM1052615	vehicle	A375 cells 24h Control rep1
GSM1052616	vehicle	A375 cells 24h Control rep2
GSM1052617	vehicle	A375 cells 24h Control rep3
GSM1052618	vemurafenib	A375 cells 24h Vemurafenib rep1
GSM1052619	vemurafenib	A375 cells 24h Vemurafenib rep2
GSM1052620	vemurafenib	A375 cells 24h Vemurafenib rep3

- This study has numerous differentially expressed genes, and we will select only those where there is a four-fold difference in expression between “vehicle” (the BRAF V600E kinase is active) and vemurafenib (the BRAF V600E kinase is inhibited):
  - Use the **Compare 2 sets of samples** widget, and from there **Value means difference** and set it to **4+ fold.** In the first round select **higher**.
  - Select “vehicle” as group A by clicking on all of the samples for vehicle. Select all of the vemurafenib samples as Group B and **OK**.

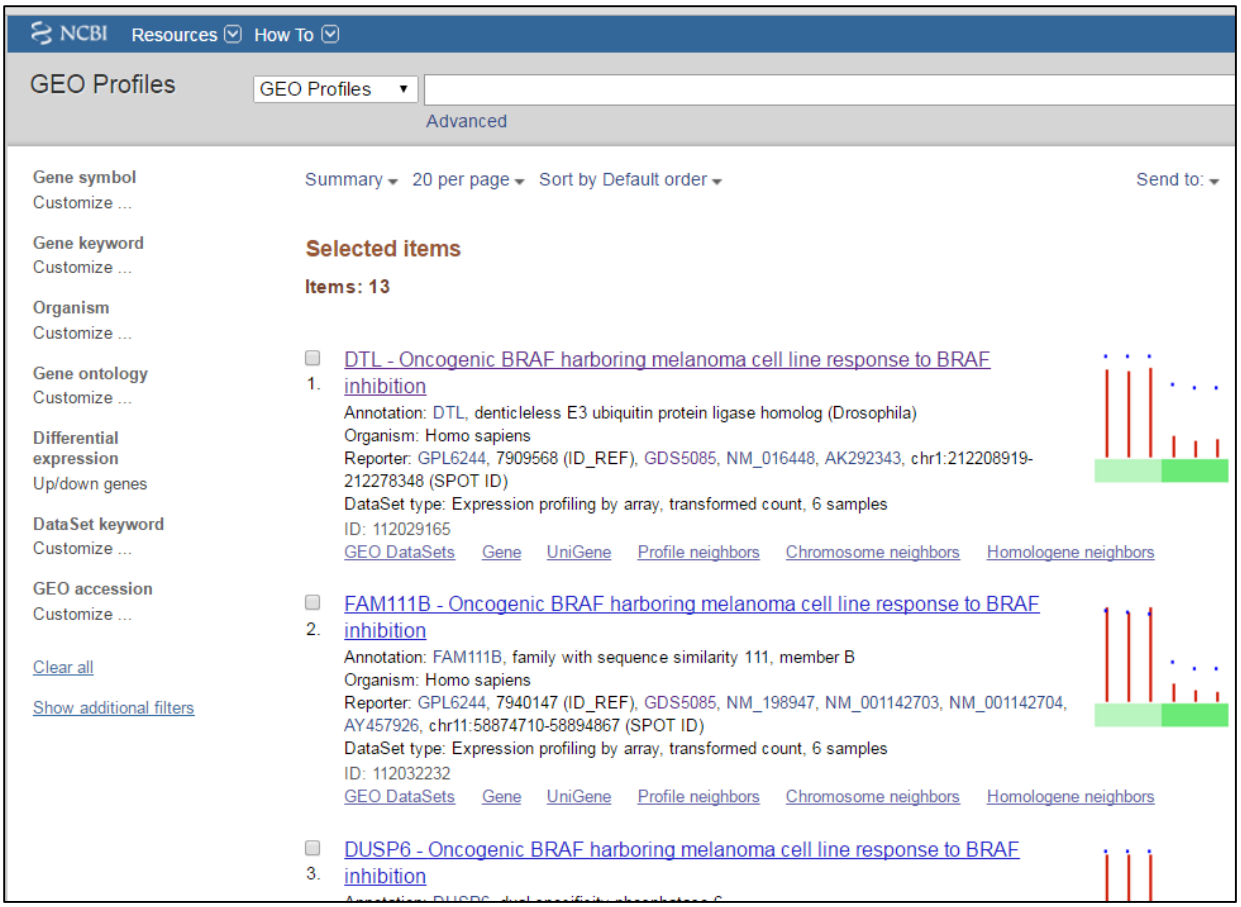


Click on accessions to select samples individually, dick on colored blocks and then on blinking arrows to select groups of samples.

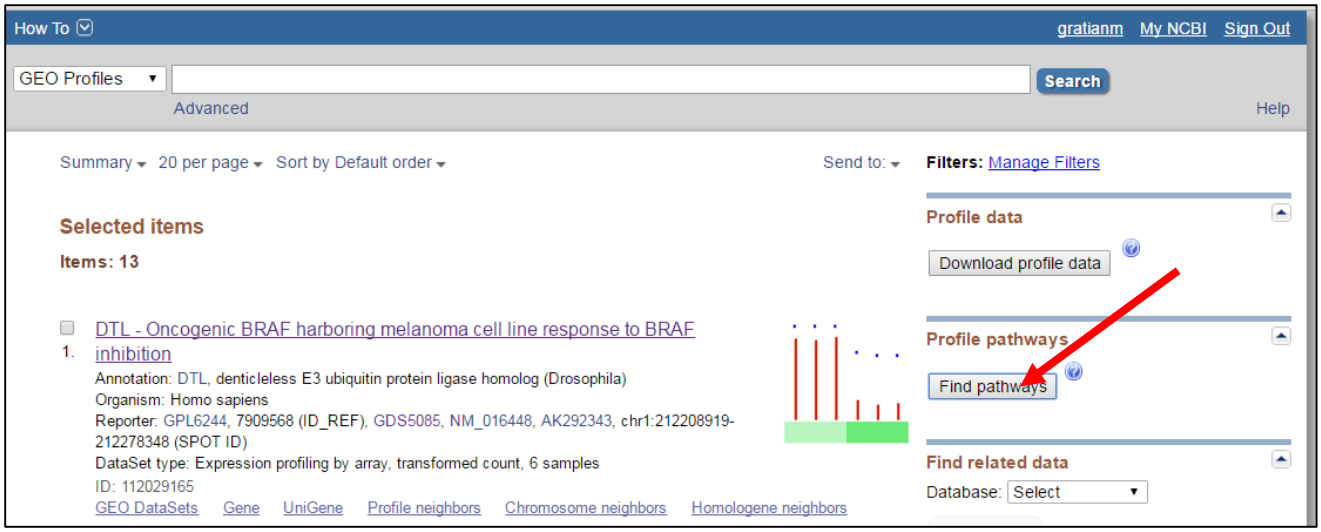
Ok Reset Cancel

Samples, Group A	Factors: agent	Samples, Group B
GSM1052615	vehicle	GSM1052615
GSM1052616		GSM1052616
GSM1052617		GSM1052617
GSM1052618	vemurafenib	GSM1052618
GSM1052619		GSM1052619
GSM1052620		GSM1052620

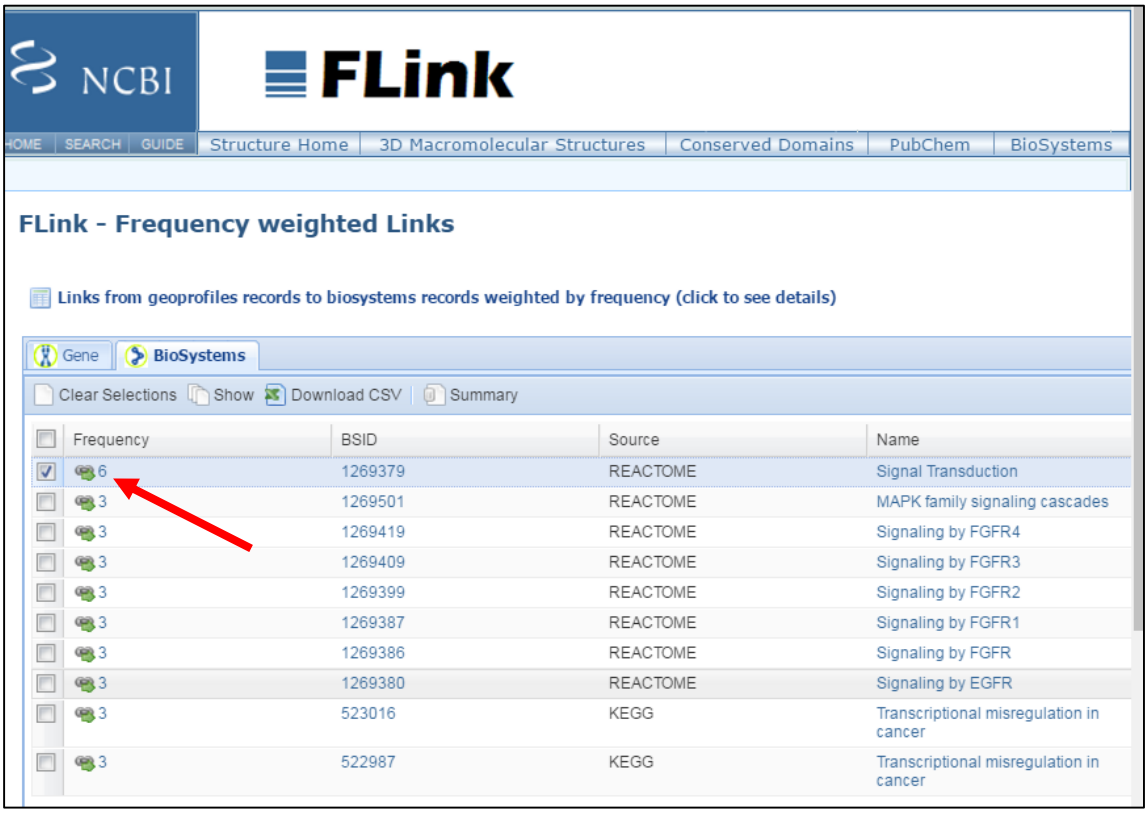
  - Query Group A vs. B.
  - The list of genes that have four fold (or more) higher expression in the cells with active BRAF mutant will be displayed in the GEO Profiles database:



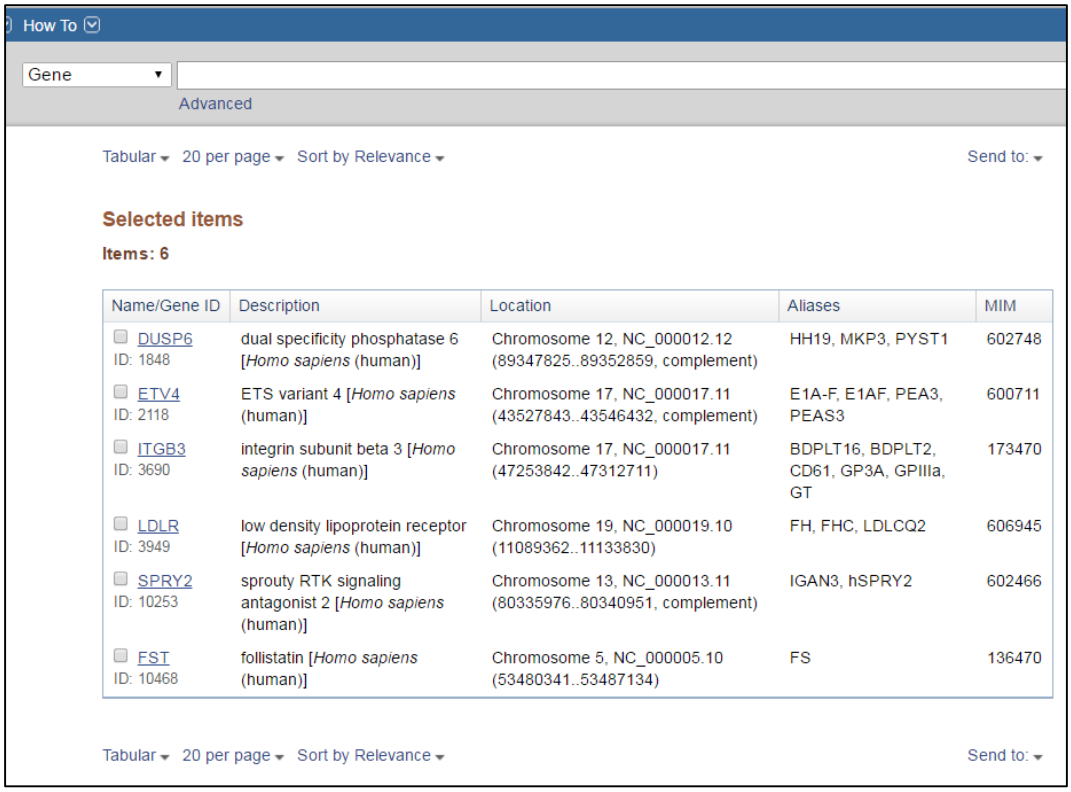
- Check the profile of the DTL gene and its expression values. From there you can link to the individual sample information that states that the values are given as log base 2
- A web-based log calculator (for example http://www.1728.org/logrithm.htm) will help you get the non-log values; where 12.1 is 4,400 and 8.9 is about 500.
- Find the biological pathways in which the 13 genes participate:



– The **FLink** (**Frequency weighted Links**) results indicate that six of the 13 genes are involved in the Signal Transduction pathway, BSID 1269379. The biological pathways in the NCBI **BioSystems** database are imported from external resources, in this case REACTOME. (If you open the BSID 1269379 record, you can get the pathway description. There are over 2500 participating genes, including BRAF and TNF.)



– To see what are the six genes, and to learn more about their function, click on the **Frequency** number to link to the records in the Gene database:



– You can display the summary of the records as text and copy the information to be used in Part II of this module:

Genes with higher expression when mutant BRAF is active

```
1. DUSP6
Official Symbol: DUSP6 and Name: dual specificity phosphatase 6 [Homo sapiens
(human)]
Other Aliases: HH19, MKP3, PYST1
Other Designations: MAP kinase phosphatase 3; dual specificity protein
phosphatase PYST1; mitogen-activated protein kinase phosphatase 3;
serine/threonine specific protein phosphatase
Chromosome: 12; Location: 12q21.33
Annotation: Chromosome 12 NC_000012.12 (89347825..89352859, complement)
MIM: 602748
ID: 1848

2. ETV4
Official Symbol: ETV4 and Name: ETS variant 4 [Homo sapiens (human)]
Other Aliases: E1A-F, E1AF, PEA3, PEAS3
```

Other Designations: EWS protein/E1A enhancer binding protein chimera; adenovirus E1A enhancer-binding protein; ets variant gene 4 (E1A enhancer-binding protein, E1AF); polyomavirus enhancer activator 3 homolog
Chromosome: 17; Location: 17q21.31
Annotation: Chromosome 17 NC_000017.11 (43527843..43546432, complement)
MIM: 600711
ID: 2118

3. ITGB3
Official Symbol: ITGB3 and Name: integrin subunit beta 3 [Homo sapiens (human)]
Other Aliases: BDPLT16, BDPLT2, CD61, GP3A, GPIIIa, GT
Other Designations: antigen CD61; integrin beta 3; integrin, beta 3 (platelet glycoprotein IIIa, antigen CD61); platelet membrane glycoprotein IIIa
Chromosome: 17; Location: 17q21.32
Annotation: Chromosome 17 NC_000017.11 (47253842..47312711)
MIM: 173470
ID: 3690

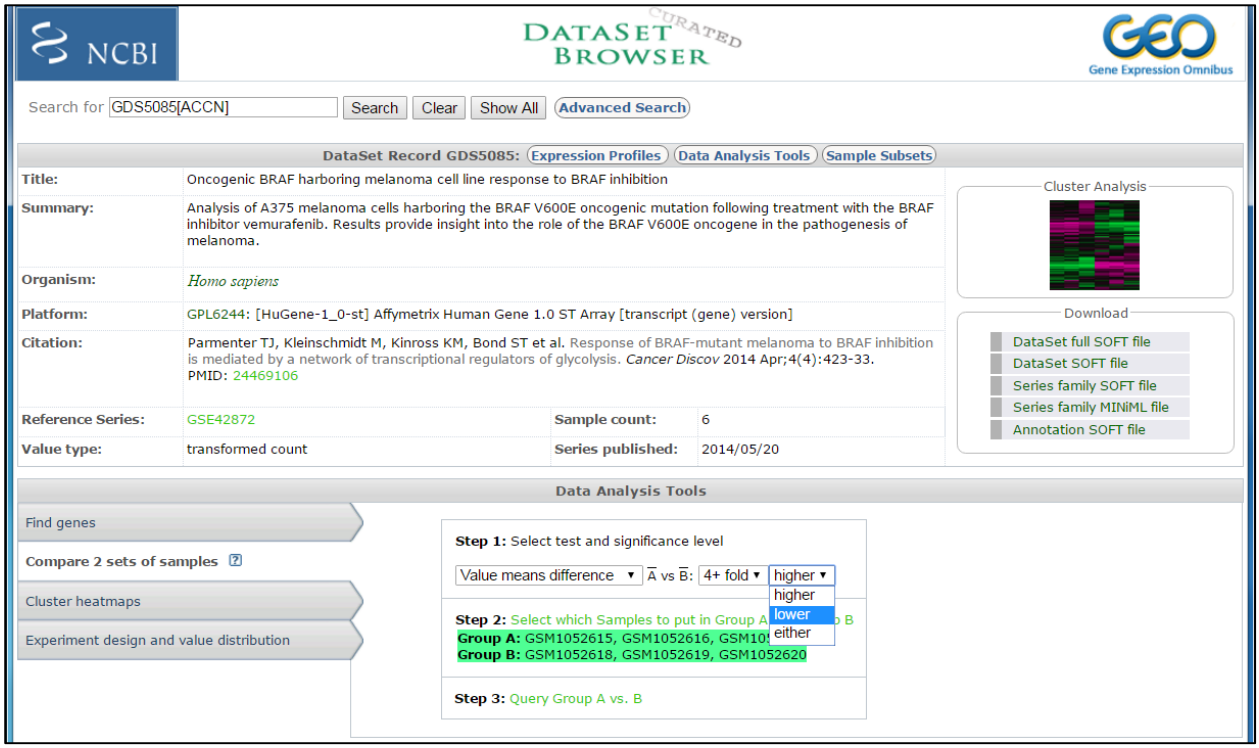
4. LDLR
Official Symbol: LDLR and Name: low density lipoprotein receptor [Homo sapiens (human)]
Other Aliases: FH, FHC, LDLCQ2
Other Designations: LDL receptor; low-density lipoprotein receptor class A domain-containing protein 3
Chromosome: 19; Location: 19p13.2
Annotation: Chromosome 19 NC_000019.10 (11089362..11133830)
MIM: 606945
ID: 3949

5. SPRY2
Official Symbol: SPRY2 and Name: sprouty RTK signaling antagonist 2 [Homo sapiens (human)]
Other Aliases: IGAN3, hSPRY2
Chromosome: 13; Location: 13q31.1
Annotation: Chromosome 13 NC_000013.11 (80335976..80340951, complement)
MIM: 602466
ID: 10253

6. FST
Official Symbol: FST and Name: follistatin [Homo sapiens (human)]
Other Aliases: FS
Other Designations: activin-binding protein; follistatin isoform FST317
Chromosome: 5; Location: 5q11.2
Annotation: Chromosome 5 NC_000005.10 (53480341..53487134)
MIM: 136470
ID: 10468

(Independent practice)

– Return to the DataSet Browser for the GDS5085 record and this time select for those genes with 4+ fold lower expression in vehicle (the BRAF V600E kinase is active):



– One of the genes that you should have retrieved is TCN1.
Q: How does this gene rank among all of the genes that were expressed on the platform? How much does the rank changes when the mutant BRAF is not active?
– From the FLink results, select the three genes that participate in the Metabolism (BSID 1269956)

Gene with higher expression when mutant BRAF is inhibited

```
1. CD36
Official Symbol: CD36 and Name: CD36 molecule [Homo sapiens (human)]
Other Aliases: BDPLT10, CHDS7, FAT, GP3B, GP4, GPIV, PASIV, SCARB3
Other Designations: CD36 antigen (collagen type I receptor, thrombospondin
receptor); CD36 molecule (thrombospondin receptor); GPIIIB; PAS IV; PAS-4
protein; cluster determinant 36; fatty acid translocase; glycoprotein IIIb;
leukocyte differentiation antigen CD36; platelet glycoprotein IV; scavenger
receptor class B, member 3
Chromosome: 7; Location: 7q11.2
Annotation: Chromosome 7 NC_000007.14 (80602188..80679277)
MIM: 173510
ID: 948
```

2. DCT
Official Symbol: DCT and Name: dopachrome tautomerase [Homo sapiens (human)]
Other Aliases: TRP-2, TYRP2
Other Designations: DT; L-dopachrome Delta-isomerase; L-dopachrome isomerase; TRP2; dopachrome delta-isomerase; dopachrome tautomerase (dopachrome delta-isomerase, tyrosine-related protein 2); tyrosinase related protein-2; tyrosinase-related protein 2; tyrosine-related protein 2
Chromosome: 13; Location: 13q32
Annotation: Chromosome 13 NC_000013.11 (94437304..94550265, complement)
MIM: 191275
ID: 1638

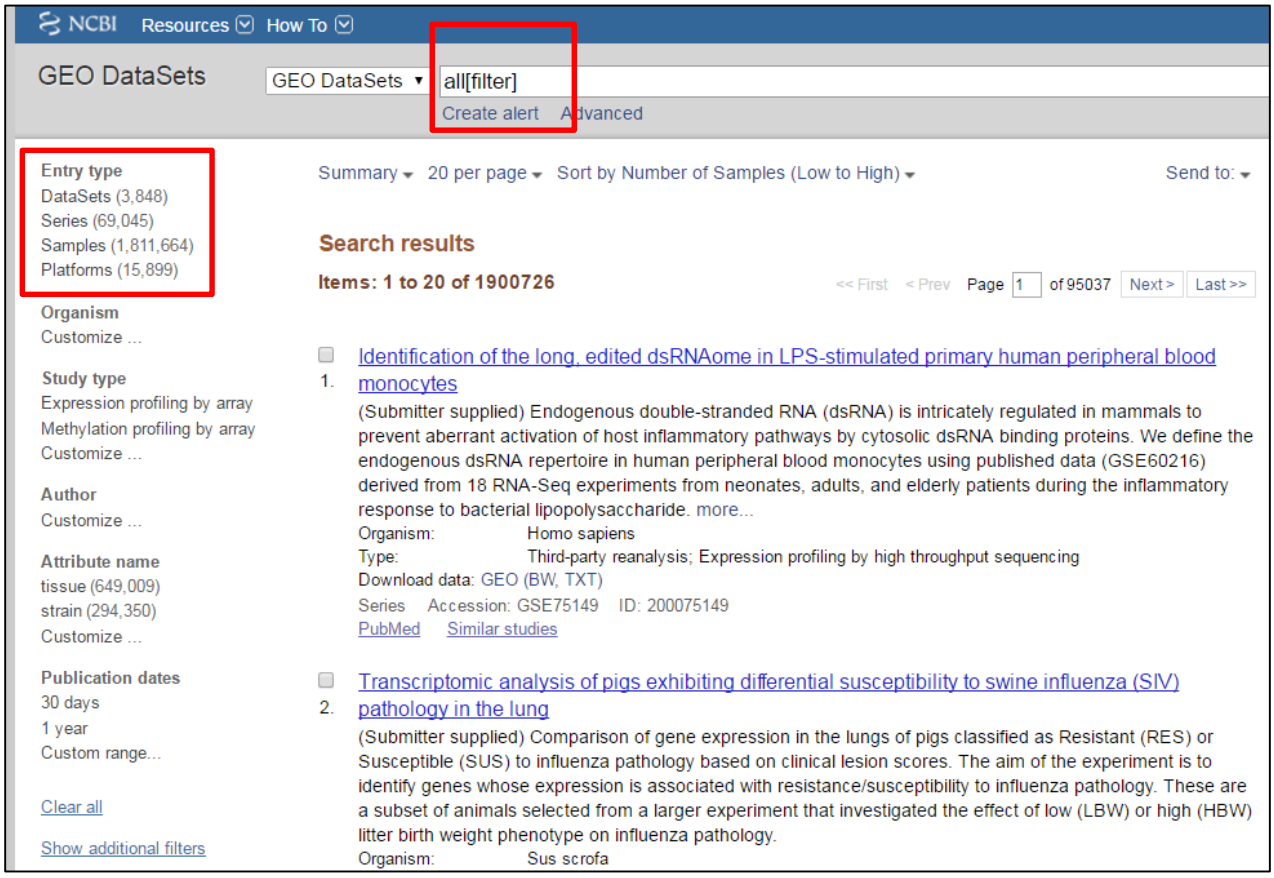
3. TCN1
Official Symbol: TCN1 and Name: transcobalamin 1 [Homo sapiens (human)]
Other Aliases: HC, TC-1, TC1, TCI
Other Designations: haptocorin; haptocorrin; protein R; transcobalamin I (vitamin B12 binding protein, R binder family)
Chromosome: 11; Location: 11q11-q12
Annotation: Chromosome 11 NC_000011.10 (59852808..59866568, complement)
MIM: 189905
ID: 6947

Exercise 3: The contents of the GEO DataSets database and RNA-Seq studies

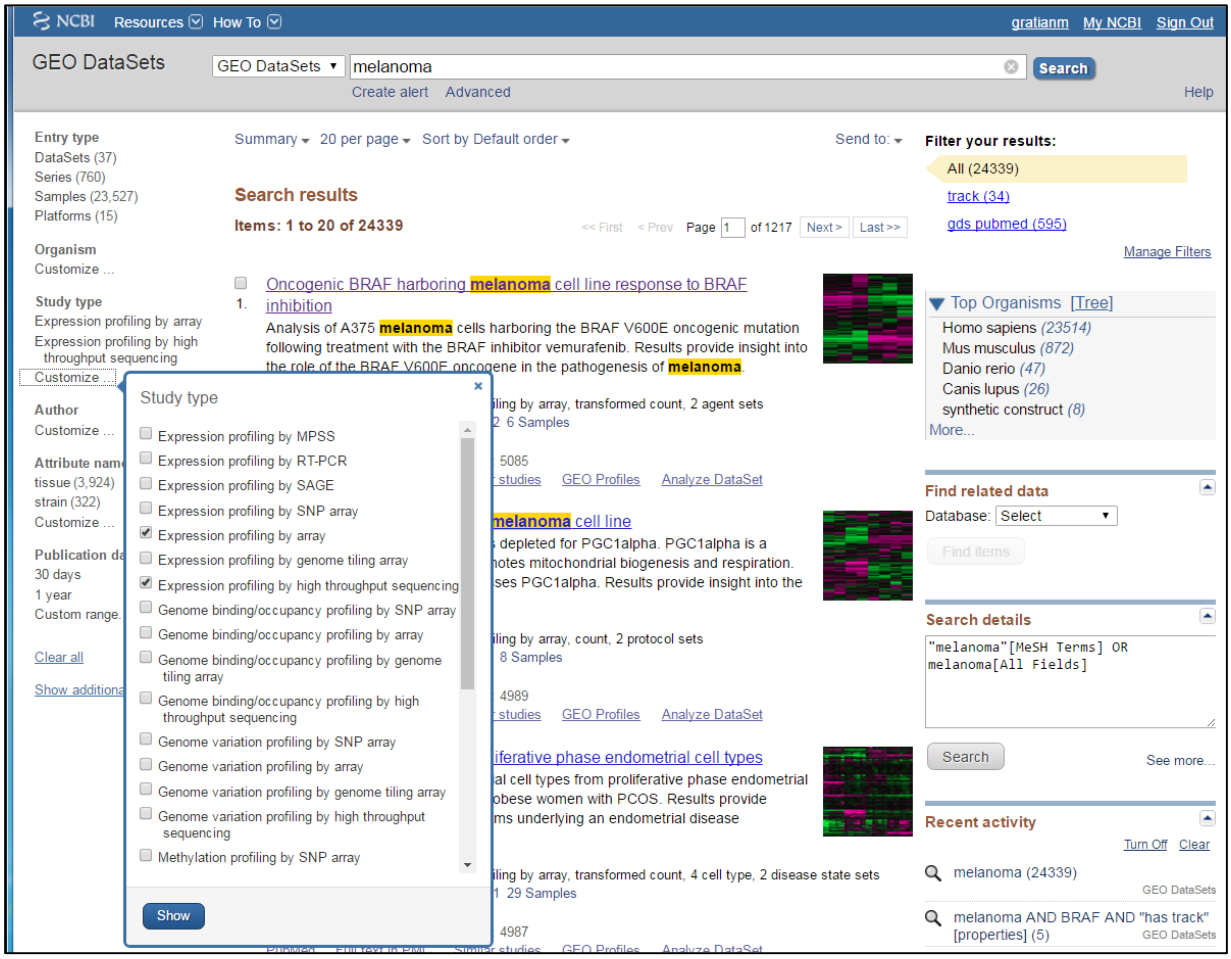
(Guided practice)

1. Study (technology) types in GEO database

   – Return to the GEO DataSets search page and check the total number of records in the database:

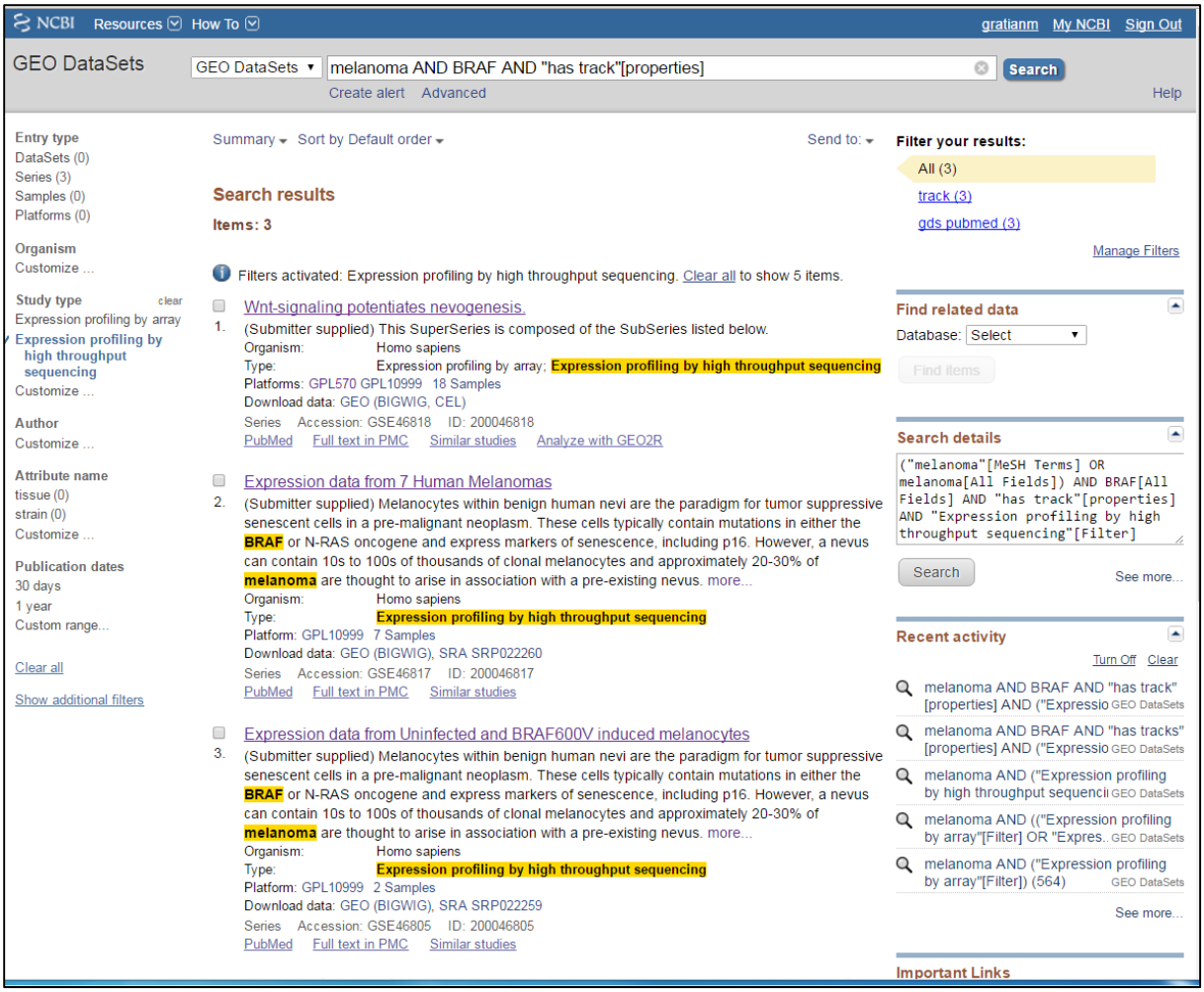


– Search for studies that relate to melanoma:

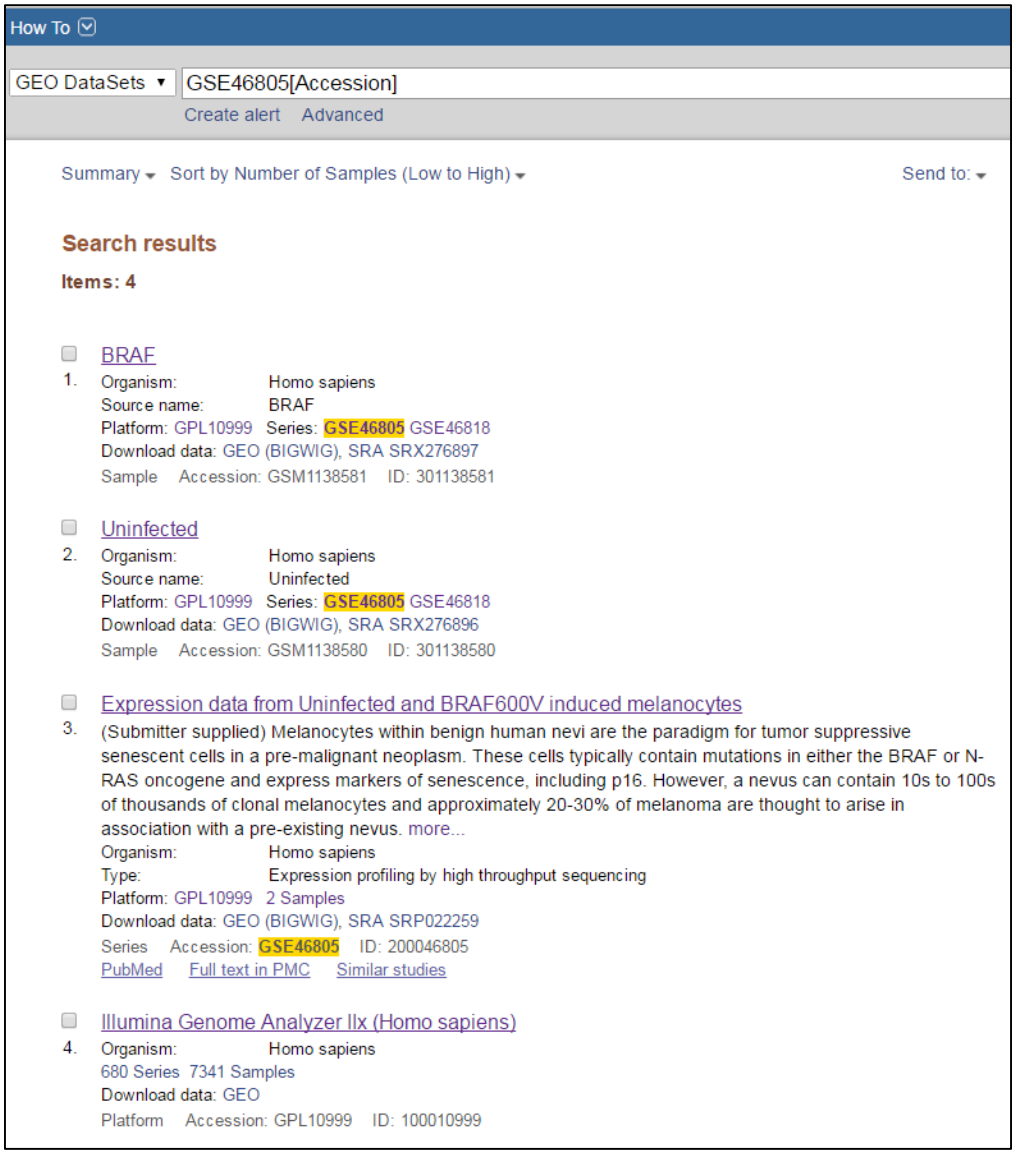


– Customize the left-side **Study Type** filter so that it displays filtering for "**Expression profiling by array**" and also for "**Expression profiling by high throughput sequencing**" (RNA-Seq). These are two of the prevalent study types among several that are accepted to be archived in the GEO DataSet database.
– Select the "Expression profiling by high throughput sequencing". To find out if there are any RNA-Seq studies that deal with melanoma and also specifically with the BRAF mutant, add the term BRAF to the search.
– To be able to do any visualization and examination of the RNA-Seq data on the web, you also need to select a study that "has track":

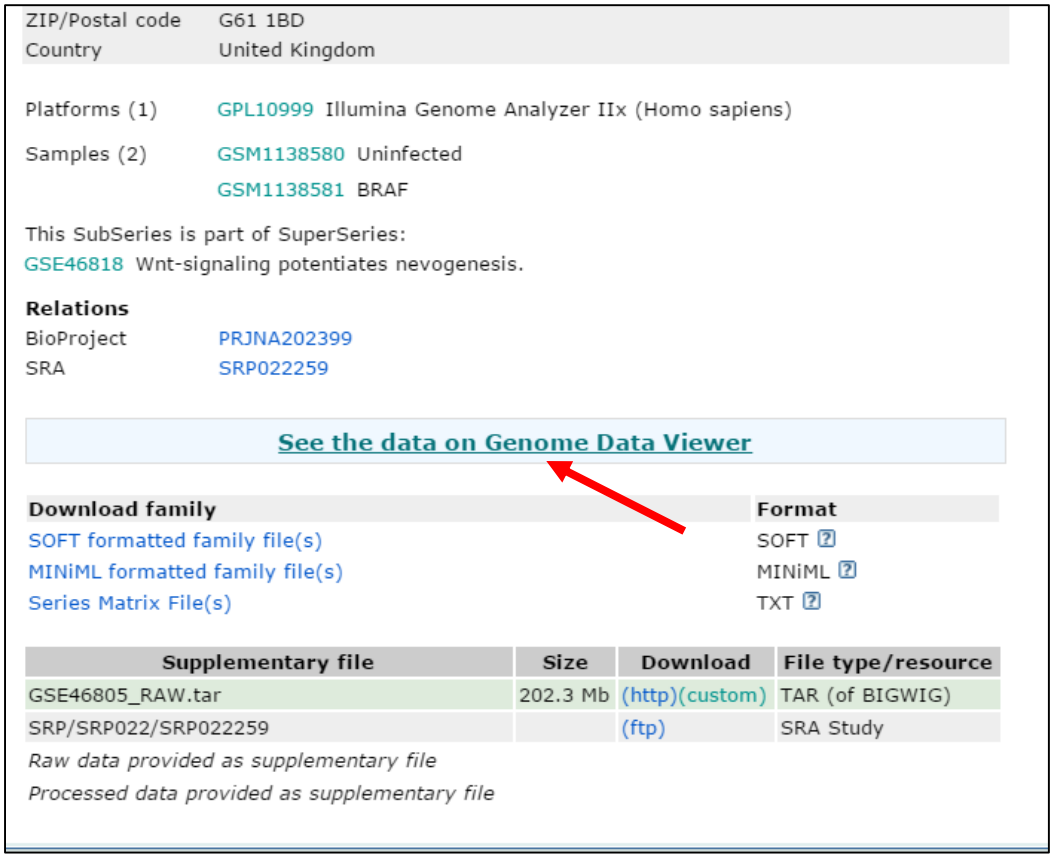
melanoma AND BRAF AND "has track"[properties]



- Even though there are only three studies in your search results, sort the results by **Number of Samples (Low to High).**
- The "Expression data from Uninfected and BRAF600V induced melanocytes" study (GSE46805) has two samples.



- In the “BRAF” sample, human epidermal melanocytes were infected with lentivirus with the BRAF mutant. (This is somewhat comparable with the “vehicle cells” in the A375 cell line as the mutant protein is expressed there as well.) The “uninfected” sample could be considered as an equivalent for BRAF suppression of the A375 cell line.
- The sequencing was on the Illumina Genome Analyzer IIx (Homo sapiens) platform.
- Access the series record and from there use the “**See the data on Genome Data Viewer**” link.



ZIP/Postal code	G61 1BD
Country	United Kingdom

Platforms (1) GPL10999 Illumina Genome Analyzer IIx (Homo sapiens)

Samples (2) GSM1138580 Uninfected
GSM1138581 BRAF

This SubSeries is part of SuperSeries:
GSE46818 Wnt-signaling potentiates nevogenesis.

Relations

BioProject	PRJNA202399
SRA	SRP022259

See the data on Genome Data Viewer

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE46805_RAW.tar	202.3 Mb	(http)(custom)	TAR (of BIGWIG)
SRP/SRP022/SRP022259		(ftp)	SRA Study

Raw data provided as supplementary file
Processed data provided as supplementary file

(Independent practice)

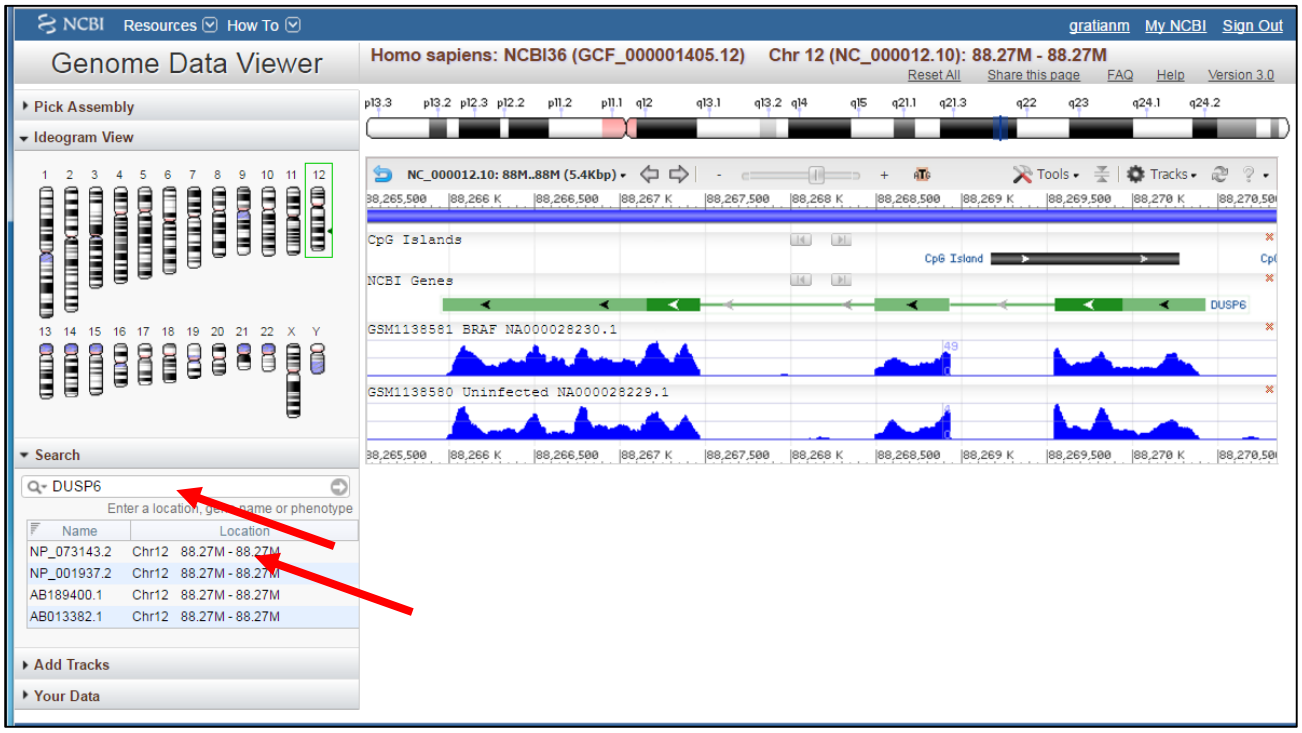
- Use the **Genome Data Viewer** to see if you can corroborate the significance of the differentially expressed genes from Exercise 2 in this study. The genes with the following symbols: DUSP6, ETV4, ITGB3, LDLR, SPRY2, and FST were those with much higher expression in the presence of the active BRAF mutant protein.
- Select one of the genes and search the Genome Data Viewer to locate the gene, so that you can examine its expression tracks.

  Q: What is the genomic assembly to which the RNA-Seq BRAF expression data were mapped? _____

  Q: Is this the current assembly for the human reference genome? _____

  Q: What can you tell about the expression of the gene that you checked?_____

  Q: Does the expression corroborate the findings from the array study in Exercise 2?_____



- The genes with the following symbols: CD36, DCT, and TCN1 were those with much lower expression in the presence of the active BRAF mutant.
- Locate the TCN1 gene in the Genome Data Viewer:

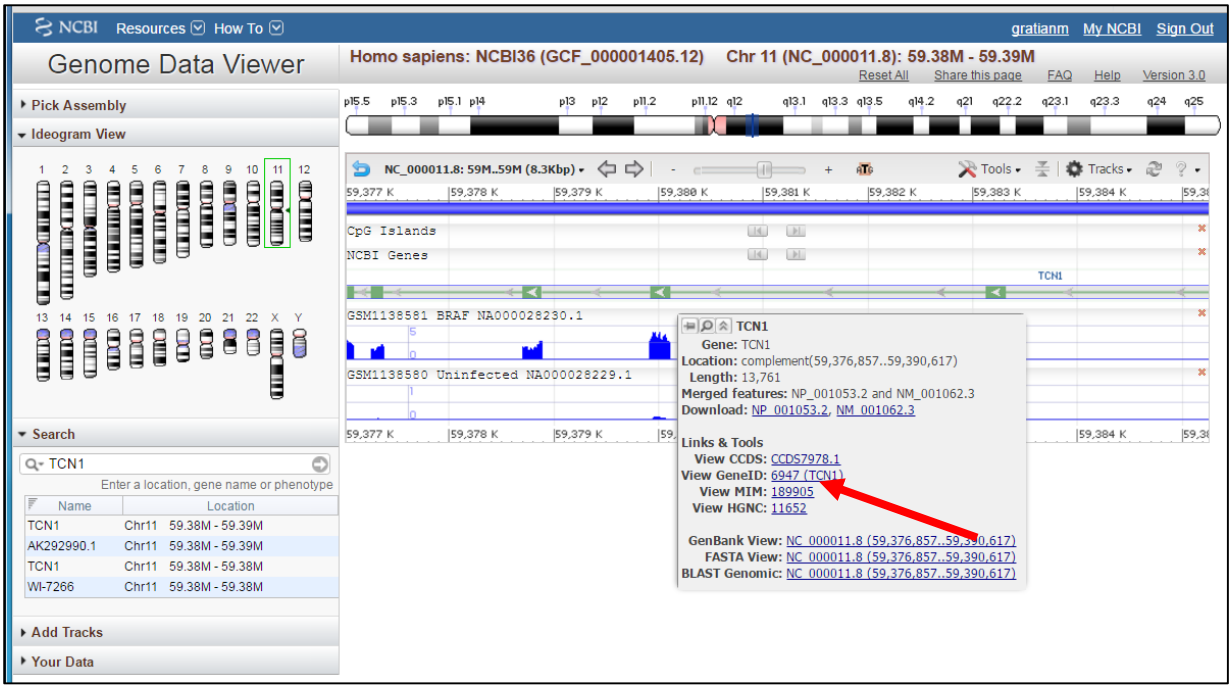
  Q: What can you tell about the expression of TCN1 (transcobalamin 1)?_____

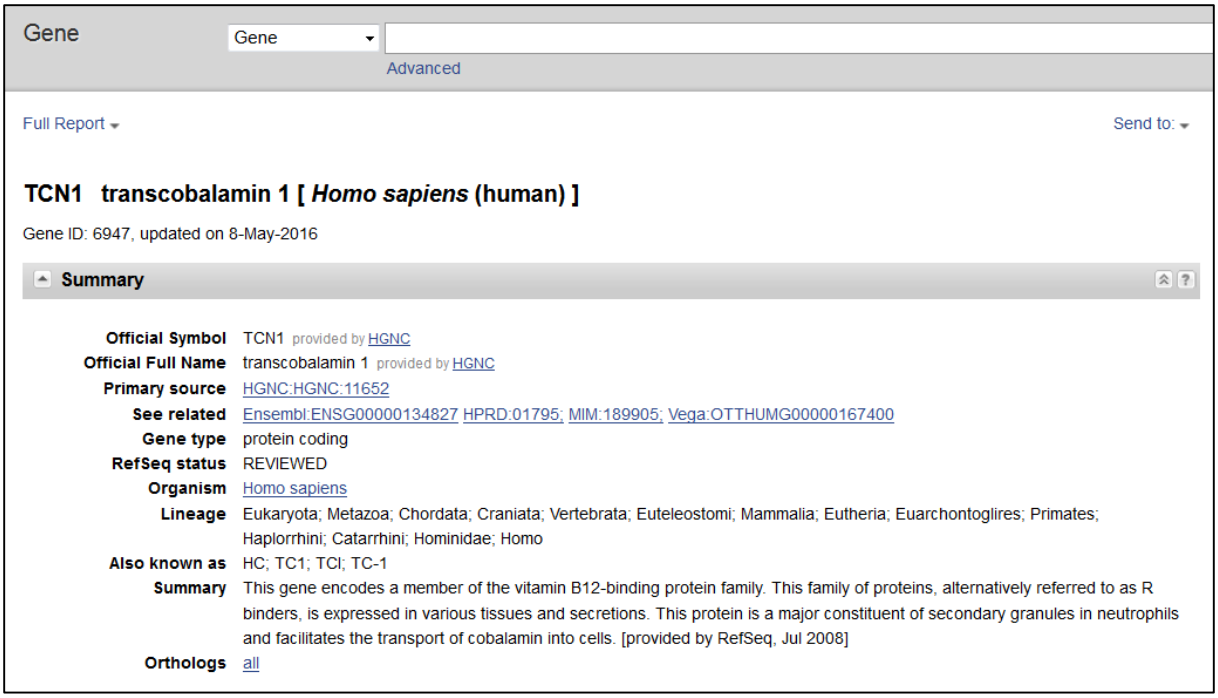
Exercise 4: Processed RNA-Seq data in the Gene database (with a quick detour to UniGene)

(Independent practice)

1. Visualizing gene expression in normal tissues with NCBI Sequence Viewer gene expression tracks

   – From the Genome Data Viewer link to the TCN1 gene record in the Gene database.





– Scroll to the **Genomic context** section of the record.

  Q: What is the chromosome on which the TCN1 gene is annotated?_____

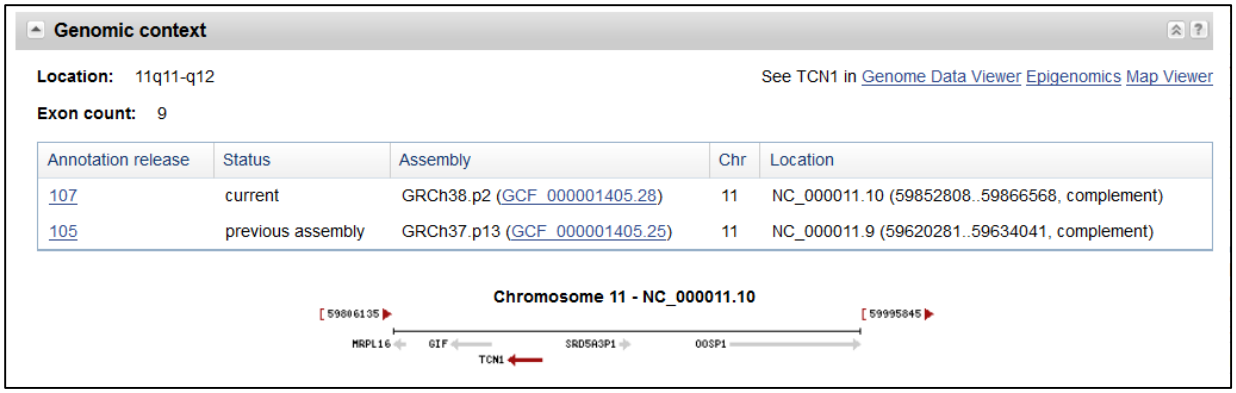
  Q: Is the gene annotated on the plus strand or on the reverse complement?_____

  Q: Are you visualizing annotation on the current human genome reference assembly?_____

  Q: How many transcripts are known for the Gene?_____

  Q: What is the accession number for the transcript (mRNA)?_____

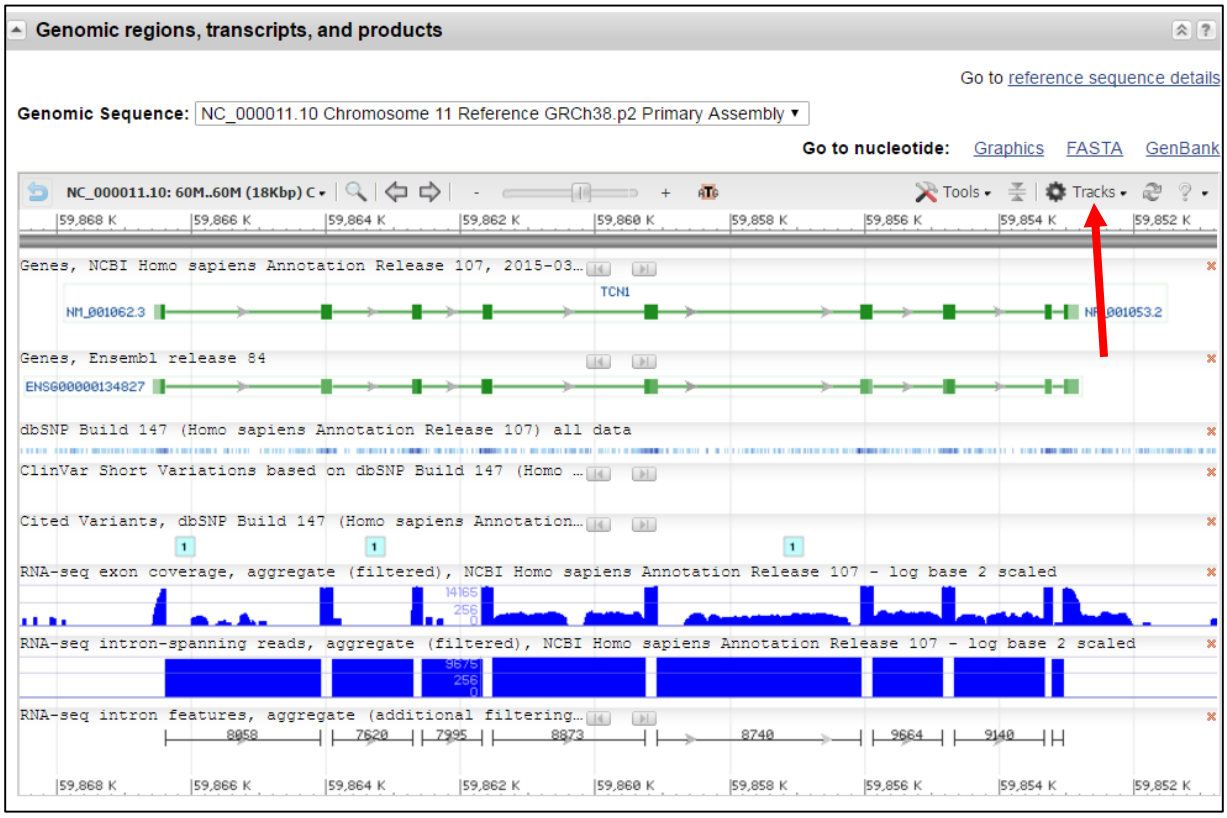
– Verify your answers just below the **Genomic context** image in this document:



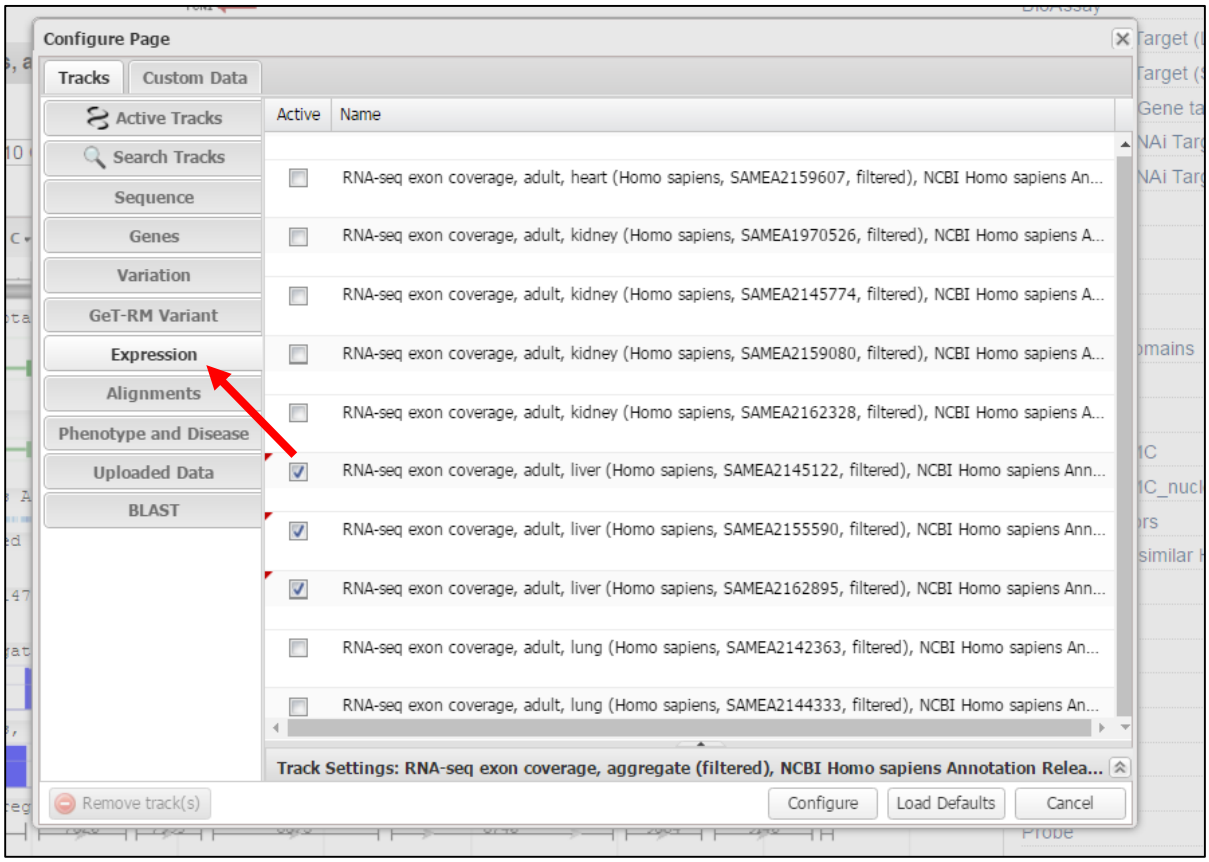
The TCN1 (transcobalamin 1) gene is annotated on chromosome 11 (NC_000011.10). It is located on the reverse complement strand and it is spanning from position 59852808 to 59866568. The annotation that is on display by default is for the current assembly GRCh38.p2. The current annotation release is 107. There is a single transcript (mRNA) that is annotated for this gene, meaning that no splice variants are described. The accession number for the transcript is NM_001062.3 and it encodes the transcobalamin-1 precursor protein, designated with the NP_001053.2 accession.

(Guided practice)

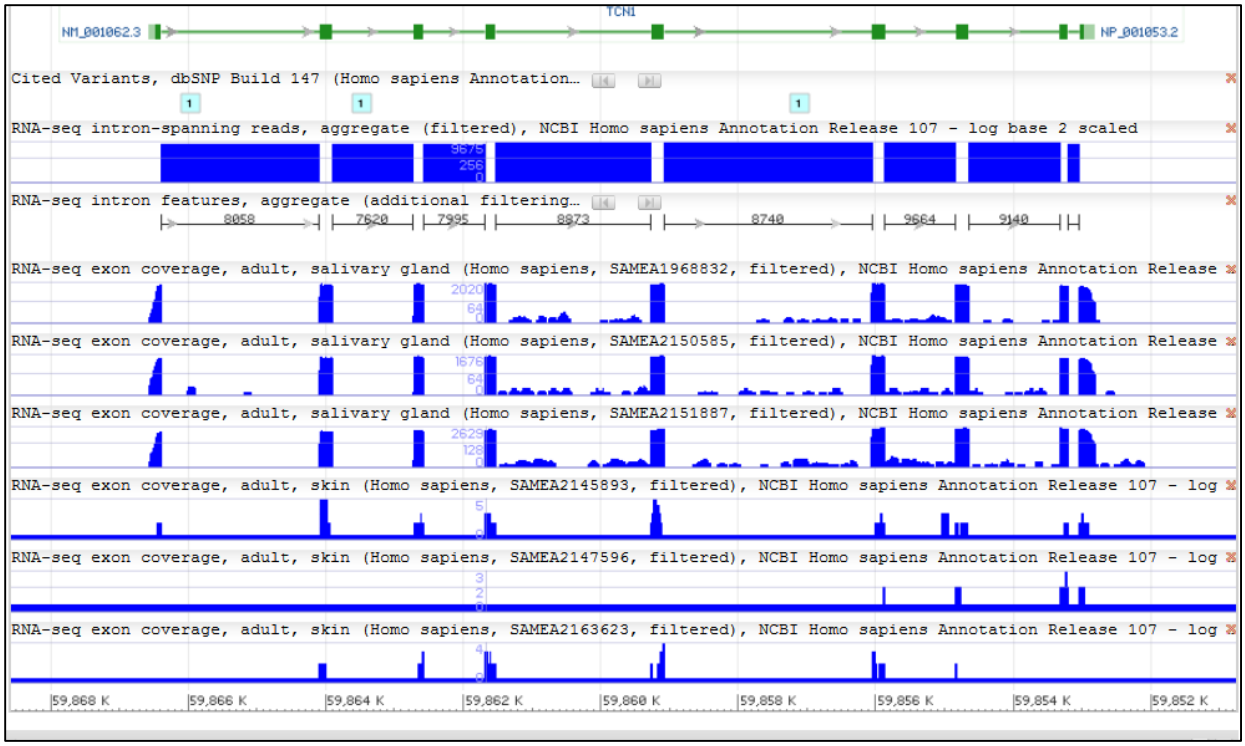
– Scroll to the **Genomic regions, transcripts, and products** section of the record.



- In the Graphics, note the RNA-Seq exon coverage and RNA-Seq intron-spanning reads. This information is from analyzed ("processed") RNA-Seq studies in the **SRA** (**Sequence Reads Archive**) database.
- Use the **Tracks** button to **Configure Expression** tracks. By default some of the aggregate tracks are selected.
- Scroll down to the second category which is RNA-Seq from individual samples. Select the RNA-Seq exon coverage tracks for salivary gland and skin:



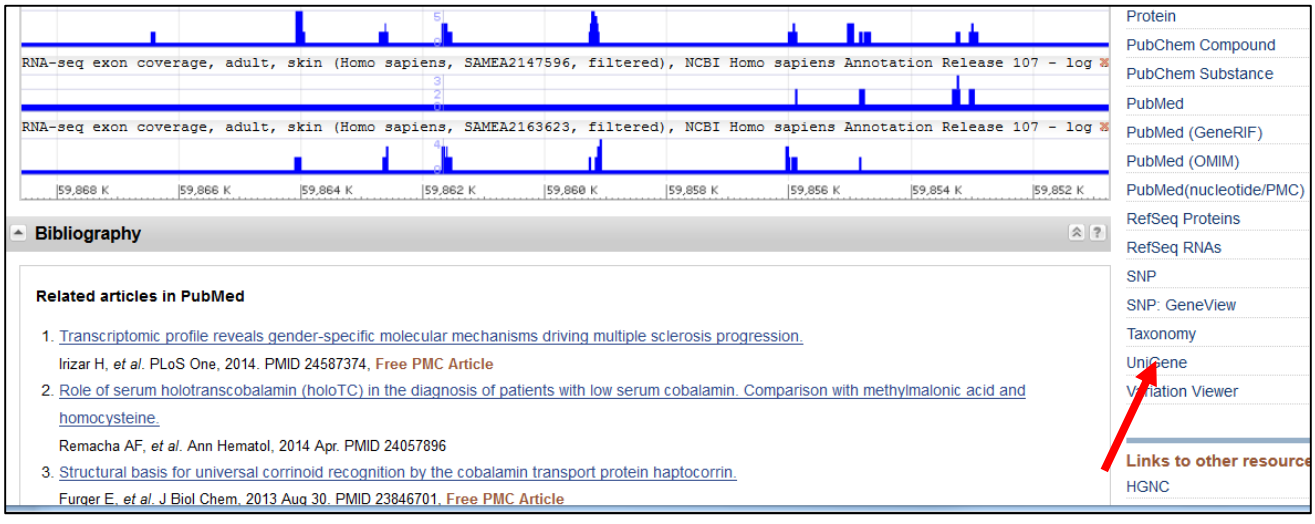
- Once you upload the tracks, also uncheck all of the variation and other tracks that are not relevant for gene expression:



– Compare the expression levels in skin with that in salivary gland.

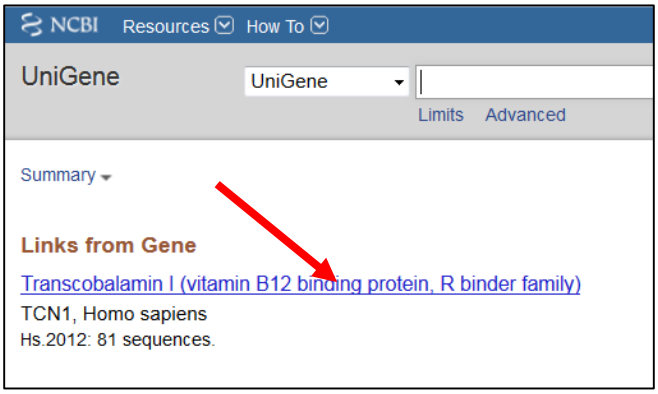
Q: Would you consider expression of this gene in the skin as high?_____

2. Using UniGene as a semi-quantitative assessment of TCN1 expression in various tissues

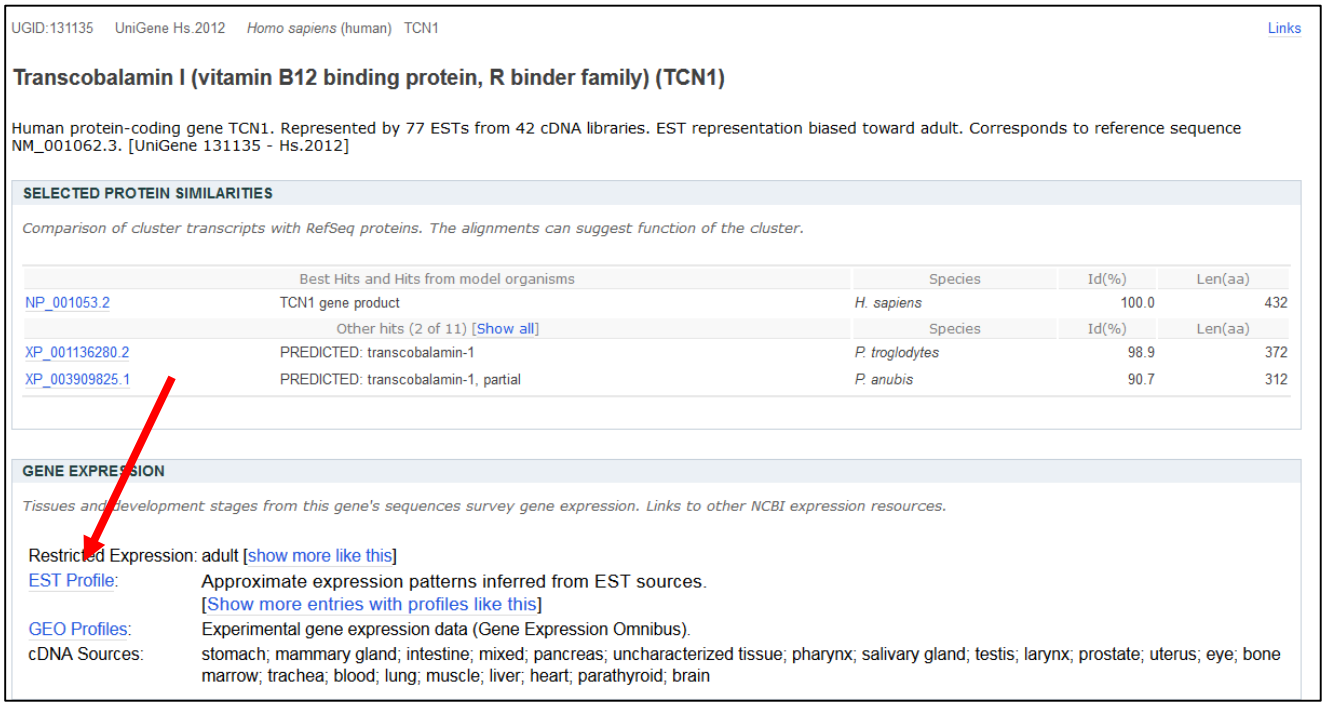


- RNA-Seq data for TCN1 expression in normal tissues indicate a very high expression in salivary gland and a very low expression in skin.

- Check the UniGene database if you could corroborate the RNA-Seq finding with UniGene expression profiles which are based on EST sequences.
- Use the UniGene link from the Gene record that takes you to the Hs.2012 UniGene Cluster. Each UniGene record is a cluster of RNA sequences that are products of the same gene.



- Open up the record and access the **EST Profile**:



prostate	26		5 / 189536
salivary gland	148		3 / 20265
skin	0		0 / 210759
spleen	0		0 / 53397
stomach	114		11 / 95679
testis	2		1 / 435204

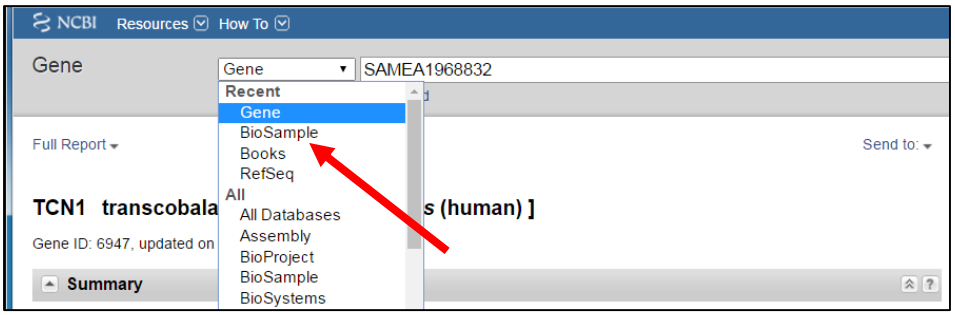
Exercise 5 (Optional): A tour of BioSample-, BioProject-, and SRA databases.

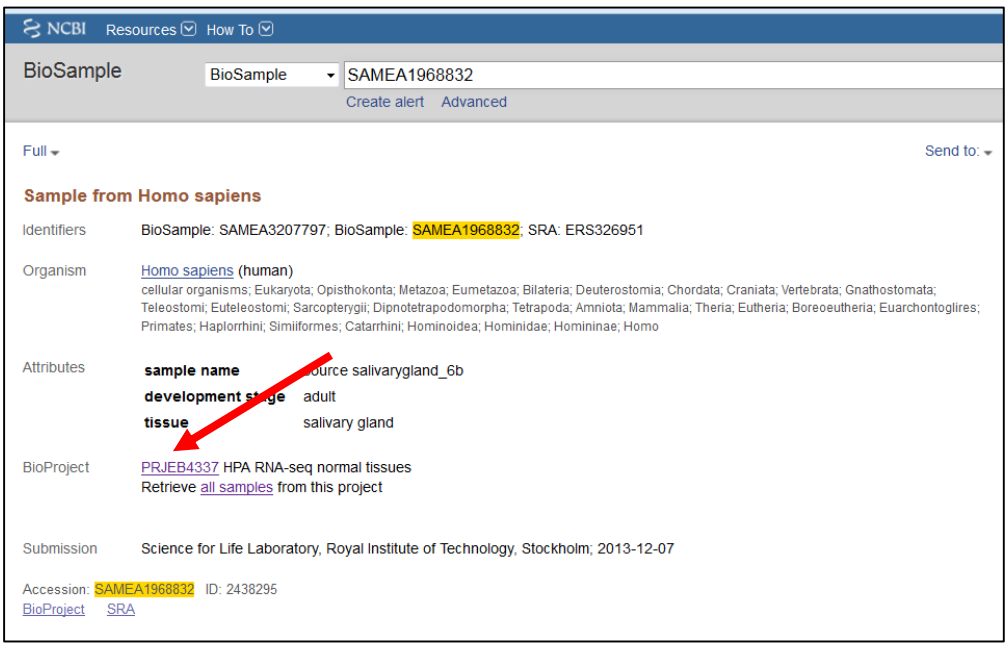
1. RNA-Seq information in Gene: Where is it coming from?

   – Return back to Gene. Each of the tracks in the Sequence Viewer (Graphic Display) is labeled. One of the labels states:

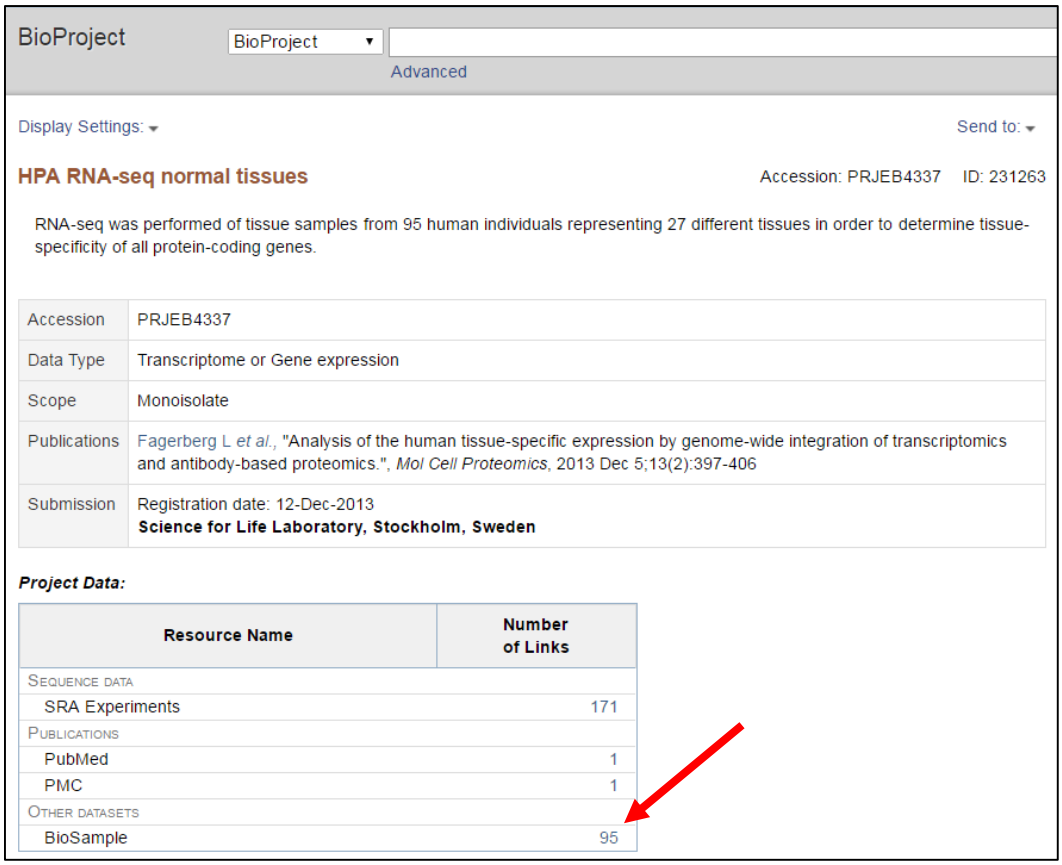
"RNA-seq exon coverage, adult, salivary gland (Homo sapiens, SAMEA1968832, filtered), NCBI Homo sapiens Annotation Release 107 - log 2 scaled"

   – The SAMEA1968832 designation is sample accession of the sample that was used in the SRA experiment.
   – Search the **BioSample** database with the accession. The BioSample database is one of the two meta-databases used for registering large research projects. The second one is the **BioProject** database.

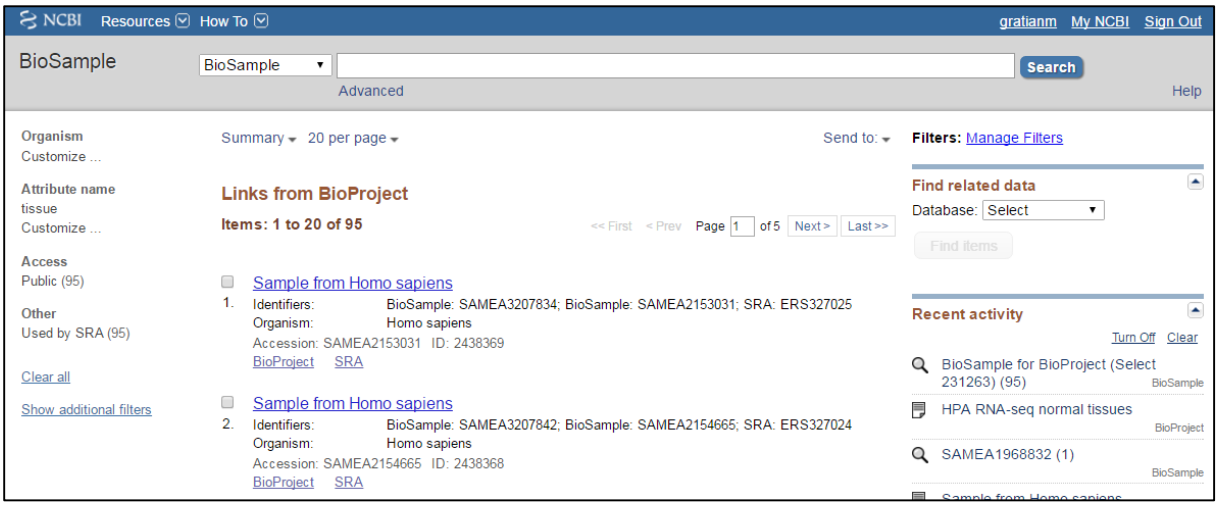




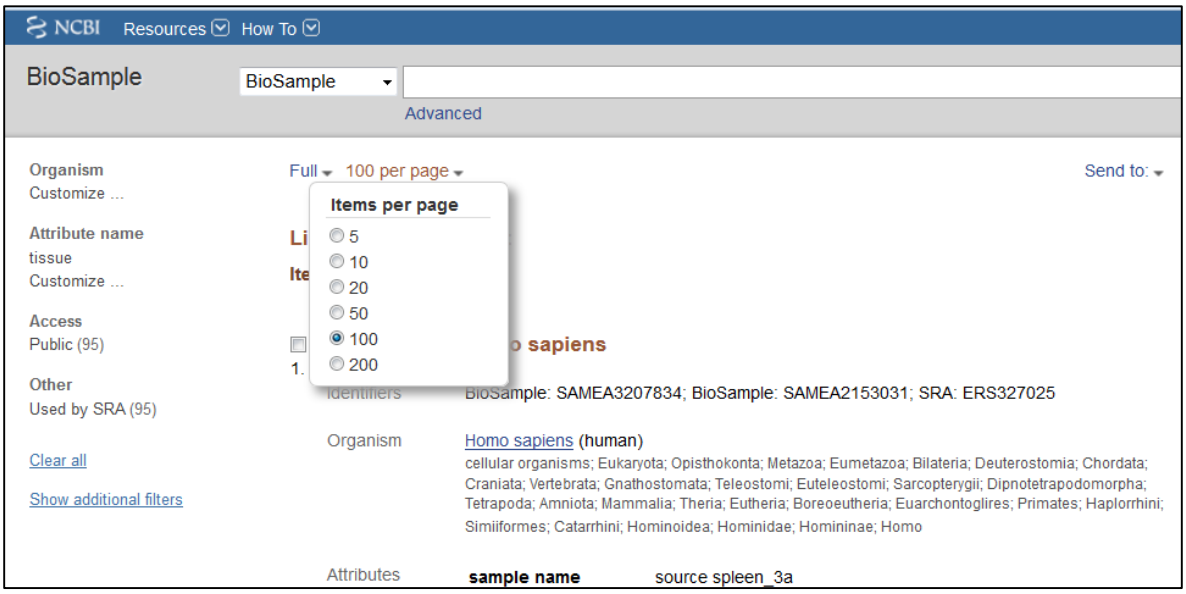
- The SAMEA1968832 sample is one of the several samples that was used in a project designated with the PRJEB4337 accession.
- Link to the project in the BioProject database, to see what the study was all about:



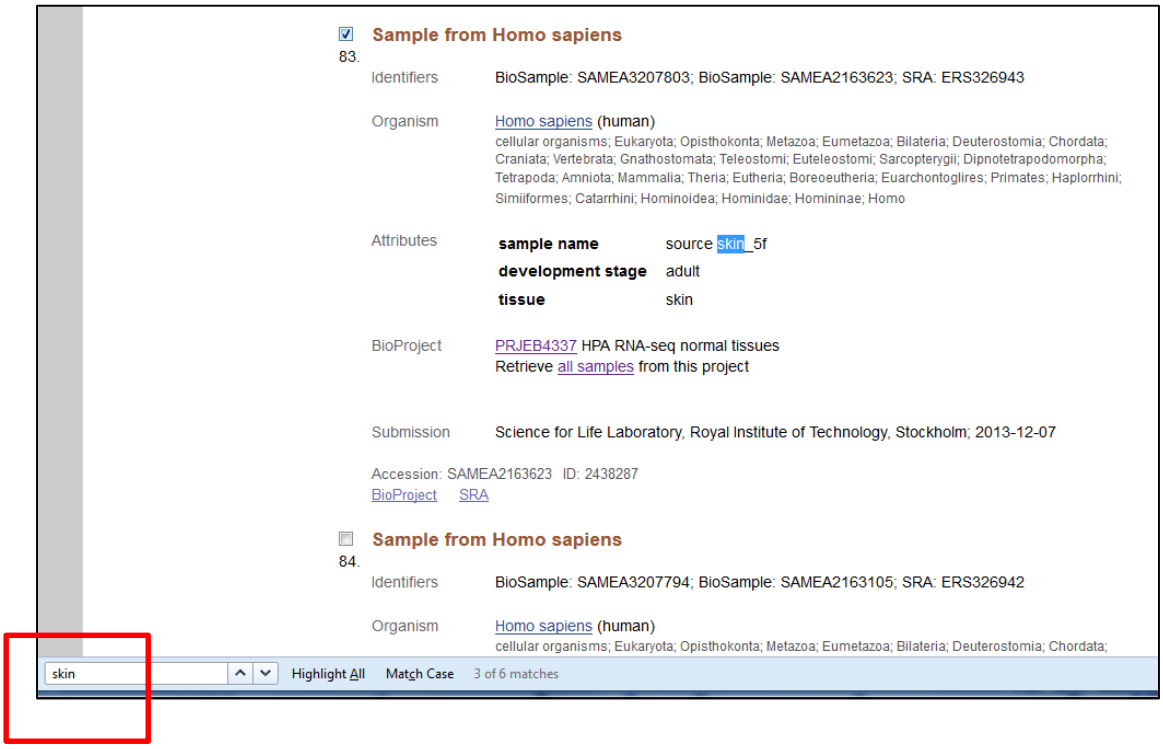
- The BioProject record also links you back to all 95 sample records in the BioSample database:



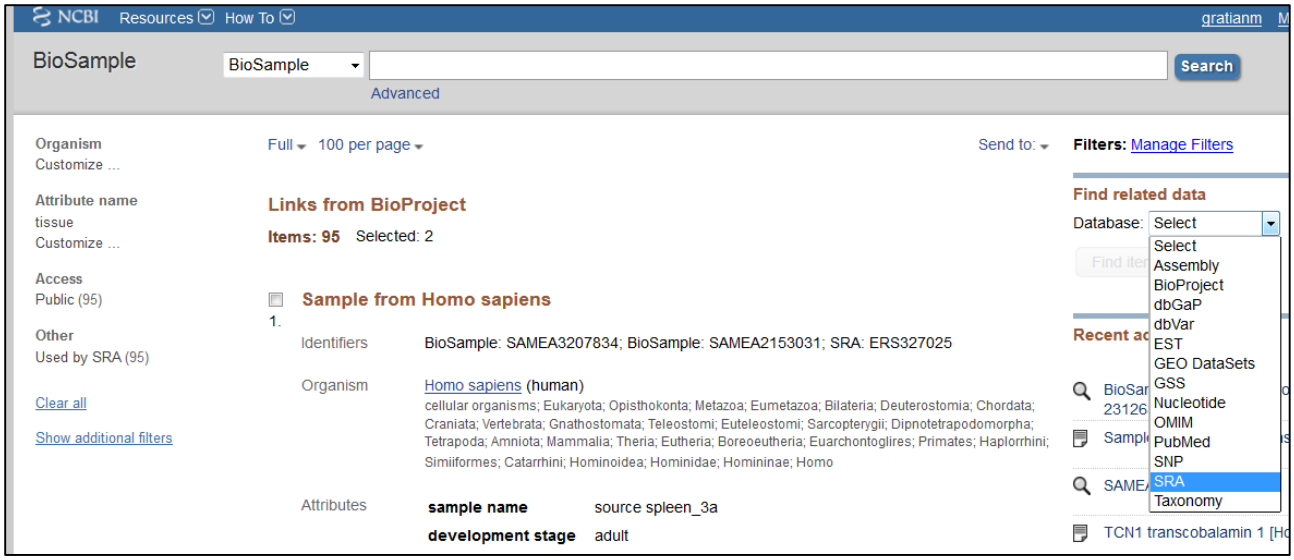
- Configure the display from summary to Full and 100 items per page:



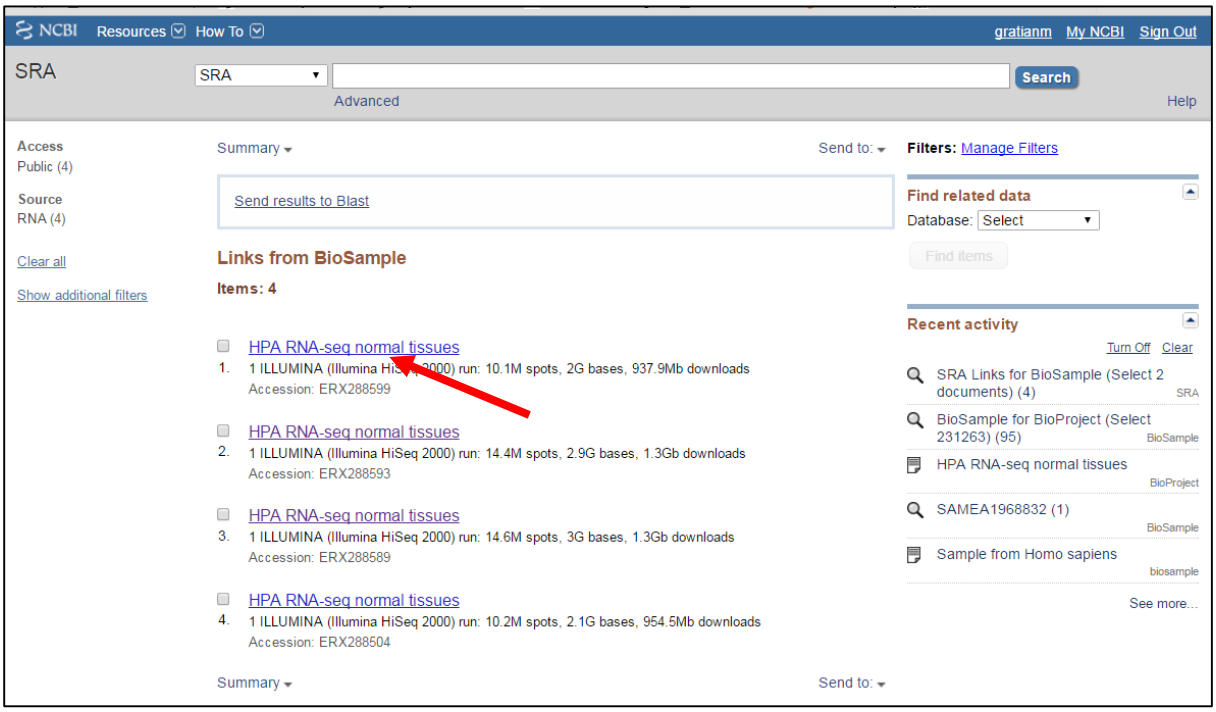
- Use the browser's search function to locate the tissue that you want, for example skin.



- Select (checkmark) the sample of your interest. Here two samples are selected:



– From the samples, link to the SRA experiments in the SRA database: these are the records for RNA-Seq reads that were processed by NCBI and made available:



ERX288599: HPA RNA-seq normal tissues
1 ILLUMINA (Illumina HiSeq 2000) run: 10.1M spots, 2G bases, 937.9Mb downloads

Design: HPA RNA-seq normal tissues

Submitted by: SCIENCE FOR LIFE LABORATORY, STOCKHOLM, SWEDEN

Study: HPA RNA-seq normal tissues
PRJEB4337 • ERP003613 • All experiments • All runs
show Abstract

Sample: Homo sapiens; salivarygland_6c
SAMEA2150585 • ERS326959 • All experiments • All runs
Organism: Homo sapiens

Library:
Name: V240
Instrument: Illumina HiSeq 2000
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: cDNA
Layout: PAIRED
Construction protocol: The use of human tissue samples was approved by the Uppsala Ethical Review Board (Reference #2011/473). Tissues samples, collected within the infrastructure of an established biobank, were embedded in Optimal Cutting Temperature (O.C.T.) compound and stored at -80C. A hematoxylin-eosin (HE) stained frozen section (4um) was prepared from each sample using a cryostat and the CryoJane Tape-Transfer System (Instrumedics, St. Louis, MO, USA). Each slide was examined by a pathologist to ensure proper tissue morphology. Three sections (10um) were cut from each frozen tissue block and collected into a tube for subsequent RNA extraction. The tissue was homogenized mechanically using a 3 mm steel grinding ball (VWR). Total RNA was extracted from cell lines and tissue samples using the RNeasy Mini Kit (Qiagen, Hilden, Germany) according to the manufacturers instructions. The extracted RNA samples were analyzed using either an Experion automated electrophoresis system (Bio-Rad Laboratories, Hercules, CA, USA) with the standard-sensitivity RNA chip or an Agilent 2100 Bioanalyzer system(Agilent Biotechnologies, Palo Alto, USA) with the RNA 6000 Nano Labchip Kit. Only samples of high-quality RNA (RNA Integrity Number 7.5) were used in the following mRNA sample preparation for sequencing. Illumina Truseq RNA v2

Spot descriptor:

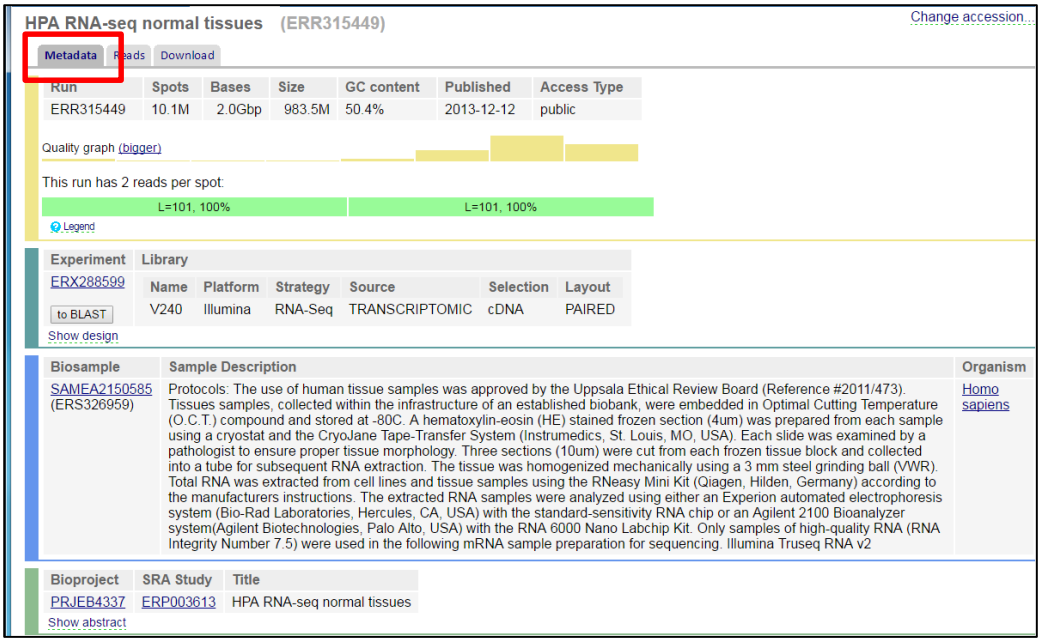
Experiment attributes:
Experimental Factor: organism part: salivary gland
Experimental Factor: individual: V240

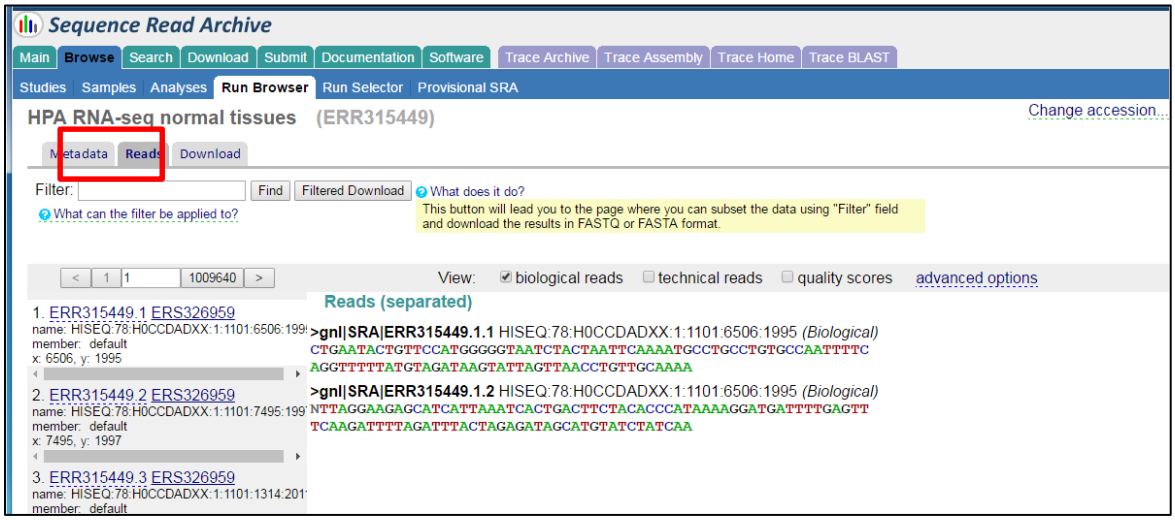
Runs: 1 run, 10.1M spots, 2G bases, 937.9Mb

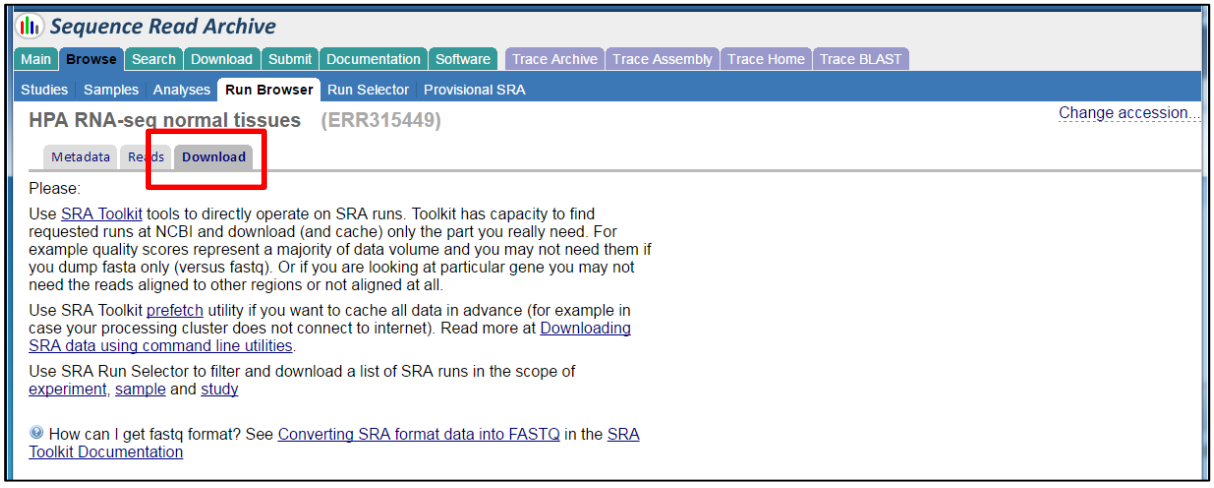
Run	# of Spots	# of Bases	Size	Published
ERR315449	10,096,395	2G	937.9Mb	2013-12-12

ID: 558501

- The table at the bottom of the record takes you to the **Run** (results) information for the experiment:

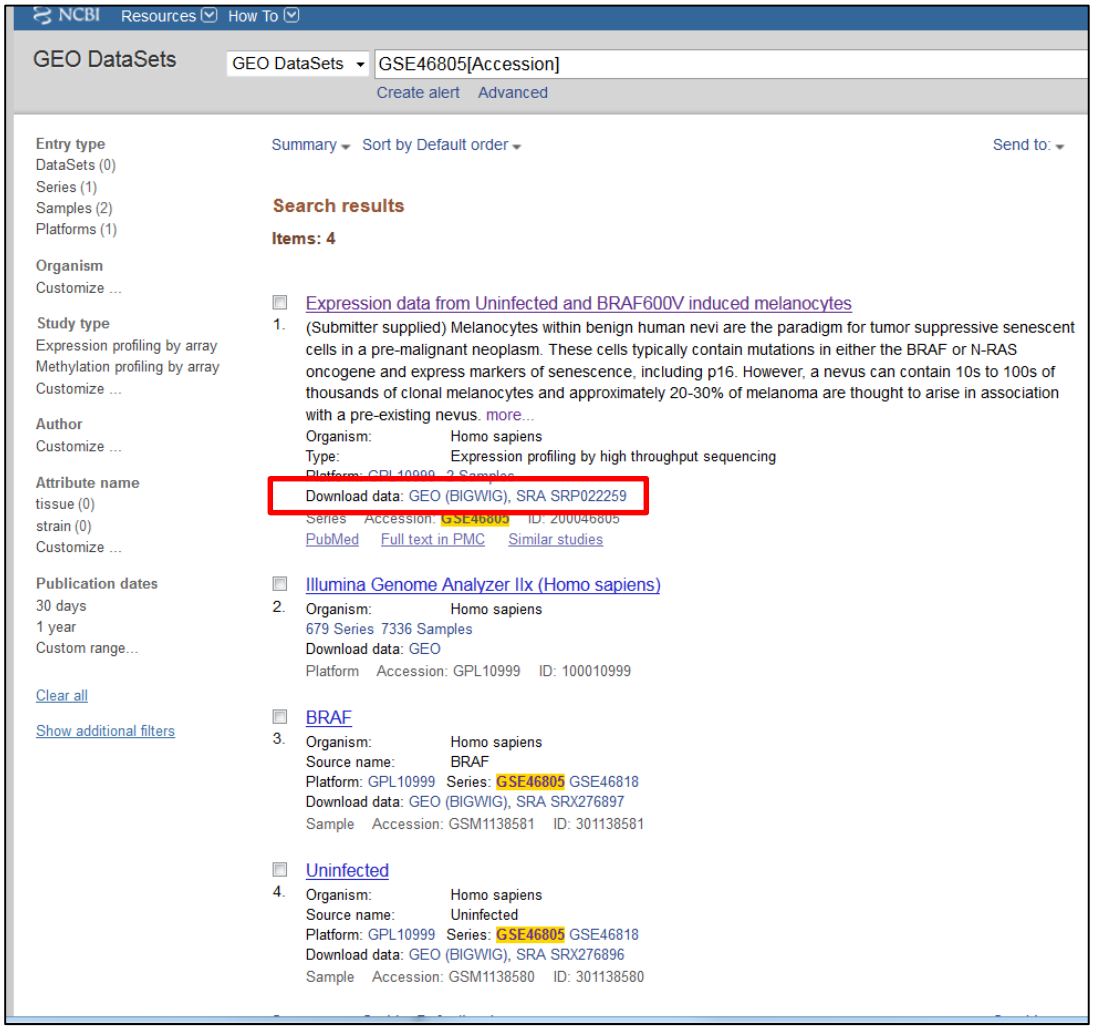

HPA RNA-seq normal tissues (ERR315449)
Change accession...
Metadata
Reads
Download
Run
Spots
Bases
Size
GC content
Published
Access Type
ERR315449
10.1M
2.0Gbp
983.5M
50.4%
2013-12-12
public
Quality graph (bigger)
This run has 2 reads per spot:
L=101, 100%
L=101, 100%
Legend
Experiment
Library
ERX288599
Name
Platform
Strategy
Source
Selection
Layout
to BLAST
V240
Illumina
RNA-Seq
TRANSCRIPTOMIC
cDNA
PAIRED
Show design
Biosample
Sample Description
Organism
SAMEA2150585
(ERS326959)
Protocols: The use of human tissue samples was approved by the Uppsala Ethical Review Board (Reference #2011/473). Tissues samples, collected within the infrastructure of an established biobank, were embedded in Optimal Cutting Temperature (O.C.T.) compound and stored at -80C. A hematoxylin-eosin (HE) stained frozen section (4um) was prepared from each sample using a cryostat and the CryoJane Tape-Transfer System (Instrumedics, St. Louis, MO, USA). Each slide was examined by a pathologist to ensure proper tissue morphology. Three sections (10um) were cut from each frozen tissue block and collected into a tube for subsequent RNA extraction. The tissue was homogenized mechanically using a 3 mm steel grinding ball (VWR). Total RNA was extracted from cell lines and tissue samples using the RNeasy Mini Kit (Qiagen, Hilden, Germany) according to the manufacturers instructions. The extracted RNA samples were analyzed using either an Experion automated electrophoresis system (Bio-Rad Laboratories, Hercules, CA, USA) with the standard-sensitivity RNA chip or an Agilent 2100 Bioanalyzer system(Agilent Biotechnologies, Palo Alto, USA) with the RNA 6000 Nano Labchip Kit. Only samples of high-quality RNA (RNA Integrity Number 7.5) were used in the following mRNA sample preparation for sequencing. Illumina Truseq RNA v2
Homo sapiens
Bioproject
SRA Study
Title
PRJEB4337
ERP003613
HPA RNA-seq normal tissues
Show abstract


Sequence Read Archive
Main
Browse
Search
Download
Submit
Documentation
Software
Trace Archive
Trace Assembly
Trace Home
Trace BLAST
Studies
Samples
Analyses
Run Browser
Run Selector
Provisional SRA
HPA RNA-seq normal tissues (ERR315449)
Change accession...
Metadata
Reads
Download
Filter:
Find
Filtered Download
What does it do?
This button will lead you to the page where you can subset the data using "Filter" field and download the results in FASTQ or FASTA format.
What can the filter be applied to?
1
1009640
View:
biological reads
technical reads
quality scores
advanced options
Reads (separated)
1. ERR315449.1 ERS326959
name: HISEQ:78:H0CCDADXX:1:1101:6506:199
member: default
x: 6506, y: 1995
>gnl|SRA|ERR315449.1.1 HISEQ:78:H0CCDADXX:1:1101:6506:1995 (Biological)
CTGAATACTGTTCCATGGGGGTAATCTACTAATTCAAAATGCCTGCCTGTGCCAATTTTC
AGGTTTTTATGTAGATAAGTATTAGTTAACCTGTTGCAAAA
2. ERR315449.2 ERS326959
name: HISEQ:78:H0CCDADXX:1:1101:7495:199
member: default
x: 7495, y: 1997
>gnl|SRA|ERR315449.1.2 HISEQ:78:H0CCDADXX:1:1101:6506:1995 (Biological)
NTTAGGAAGAGCATCATTAAATCACTGACTTCTACACCCATAAAAGGATGATTTTGAGTT
TCAAGATTTTAGATTTACTAGAGATAGCATGTATCTATCAA
3. ERR315449.3 ERS326959
name: HISEQ:78:H0CCDADXX:1:1101:1314:201
member: default

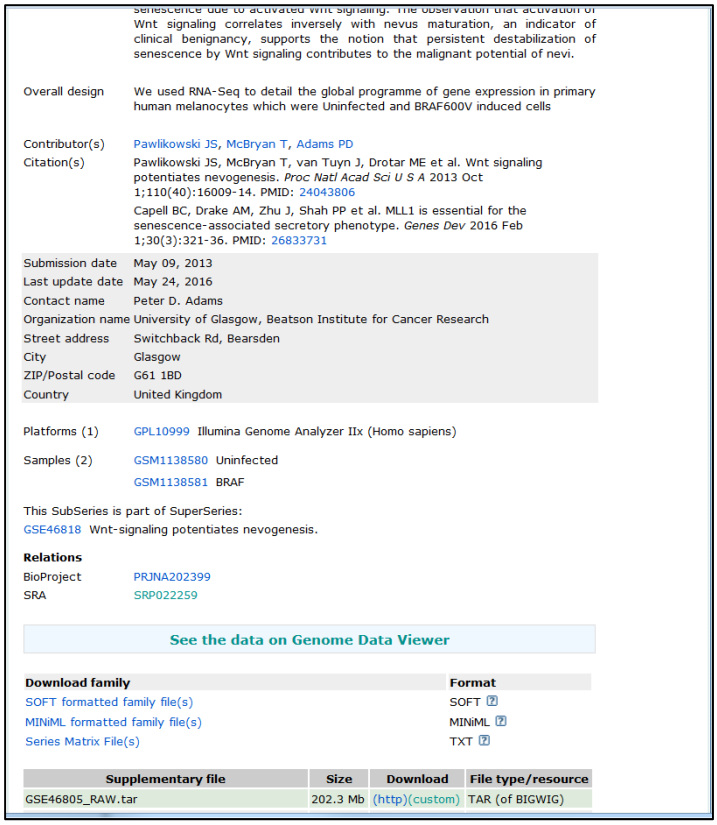

Sequence Read Archive
Main
Browse
Search
Download
Submit
Documentation
Software
Trace Archive
Trace Assembly
Trace Home
Trace BLAST
Studies
Samples
Analyses
Run Browser
Run Selector
Provisional SRA
HPA RNA-seq normal tissues (ERR315449)
Change accession...
Metadata
Reads
Download
Please:
Use SRA Toolkit tools to directly operate on SRA runs. Toolkit has capacity to find requested runs at NCBI and download (and cache) only the part you really need. For example quality scores represent a majority of data volume and you may not need them if you dump fasta only (versus fastq). Or if you are looking at particular gene you may not need the reads aligned to other regions or not aligned at all.
Use SRA Toolkit prefetch utility if you want to cache all data in advance (for example in case your processing cluster does not connect to internet). Read more at Downloading SRA data using command line utilities.
Use SRA Run Selector to filter and download a list of SRA runs in the scope of experiment, sample and study
How can I get fastq format? See Converting SRA format data into FASTQ in the SRA Toolkit Documentation

2. RNA-Seq experiments in GEO also are also registered in BioProject- and BioSample databases:

   – Revisit the BRAF RNA-Seq experiment (GSE46805)
   – The **Download data** link leads to the FTP site:



   – The SRA link in the record takes you to the records on the SRA web:



Wnt signaling correlates inversely with nevus maturation, an indicator of clinical benignancy, supports the notion that persistent destabilization of senescence by Wnt signaling contributes to the malignant potential of nevi.

Overall design	We used RNA-Seq to detail the global programme of gene expression in primary human melanocytes which were Uninfected and BRAF600V induced cells
Contributor(s)	Pawlikowski JS, McBryan T, Adams PD
Citation(s)	Pawlikowski JS, McBryan T, van Tuyn J, Drotar ME et al. Wnt signaling potentiates nevogenesis. *Proc Natl Acad Sci U S A* 2013 Oct 1;110(40):16009-14. PMID: 24043806 Capell BC, Drake AM, Zhu J, Shah PP et al. MLL1 is essential for the senescence-associated secretory phenotype. *Genes Dev* 2016 Feb 1;30(3):321-36. PMID: 26833731
Submission date	May 09, 2013
Last update date	May 24, 2016
Contact name	Peter D. Adams
Organization name	University of Glasgow, Beatson Institute for Cancer Research
Street address	Switchback Rd, Bearsden
City	Glasgow
ZIP/Postal code	G61 1BD
Country	United Kingdom
Platforms (1)	GPL10999 Illumina Genome Analyzer IIx (Homo sapiens)
Samples (2)	GSM1138580 Uninfected GSM1138581 BRAF

This SubSeries is part of SuperSeries:
GSE46818 Wnt-signaling potentiates nevogenesis.

Relations

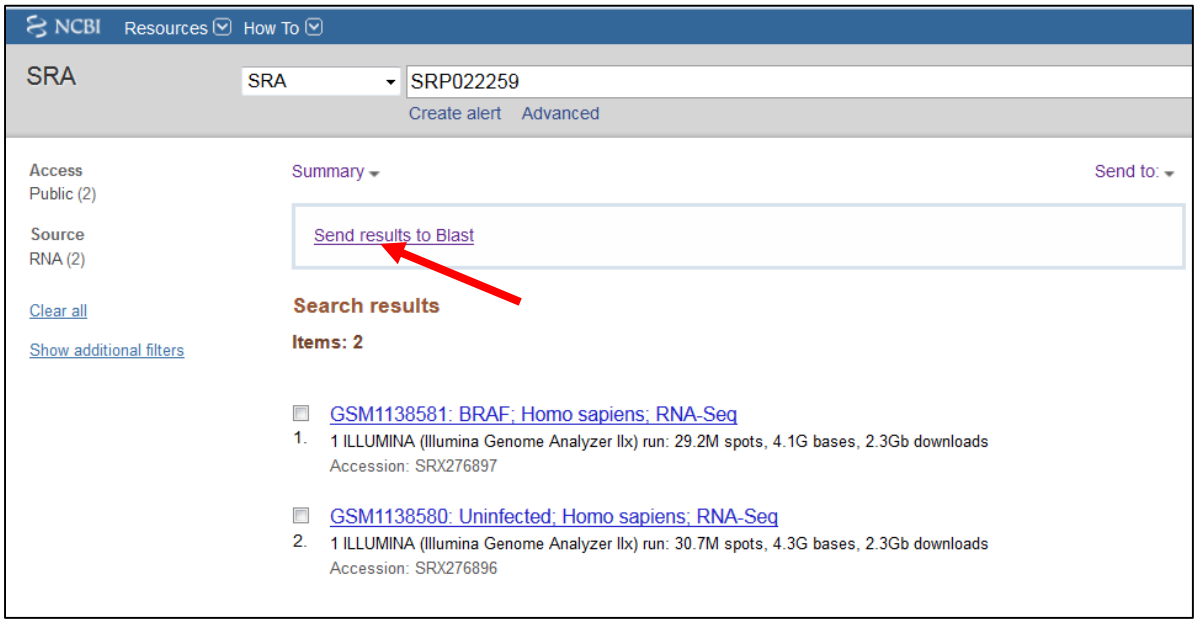
BioProject	PRJNA202399
SRA	SRP022259

See the data on Genome Data Viewer

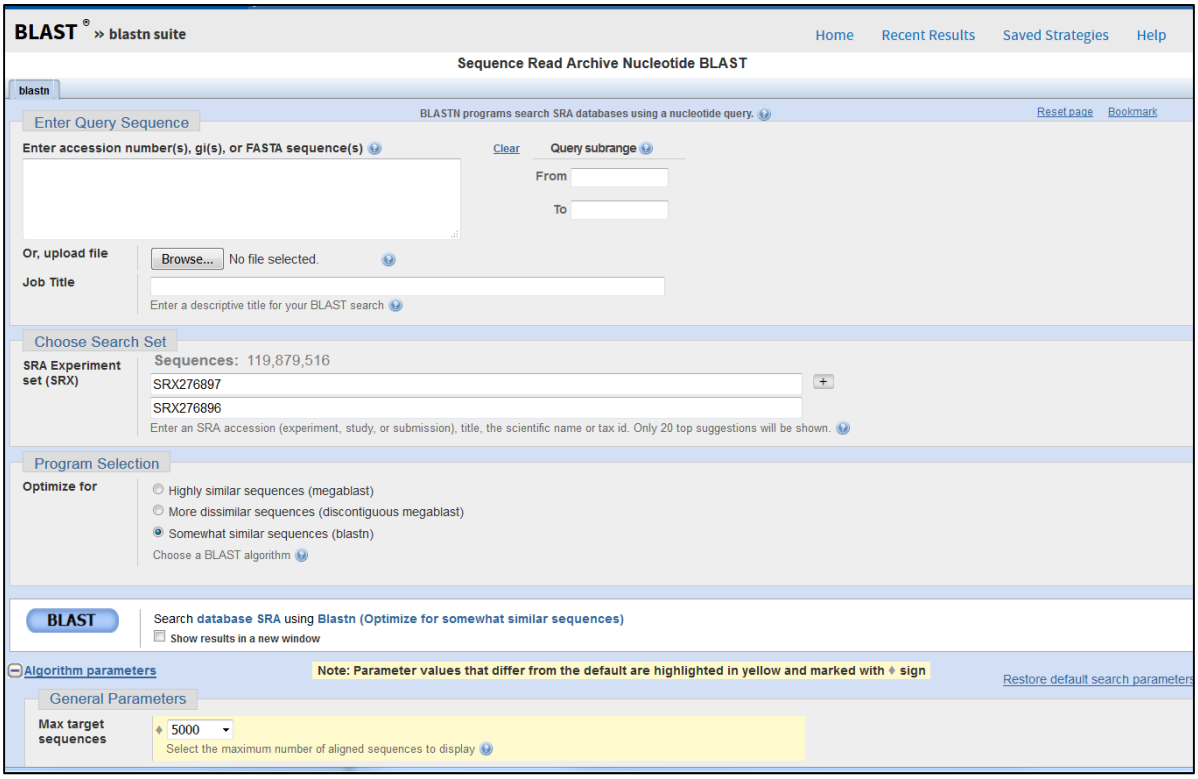
Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE46805_RAW.tar	202.3 Mb	(http)(custom)	TAR (of BIGWIG)

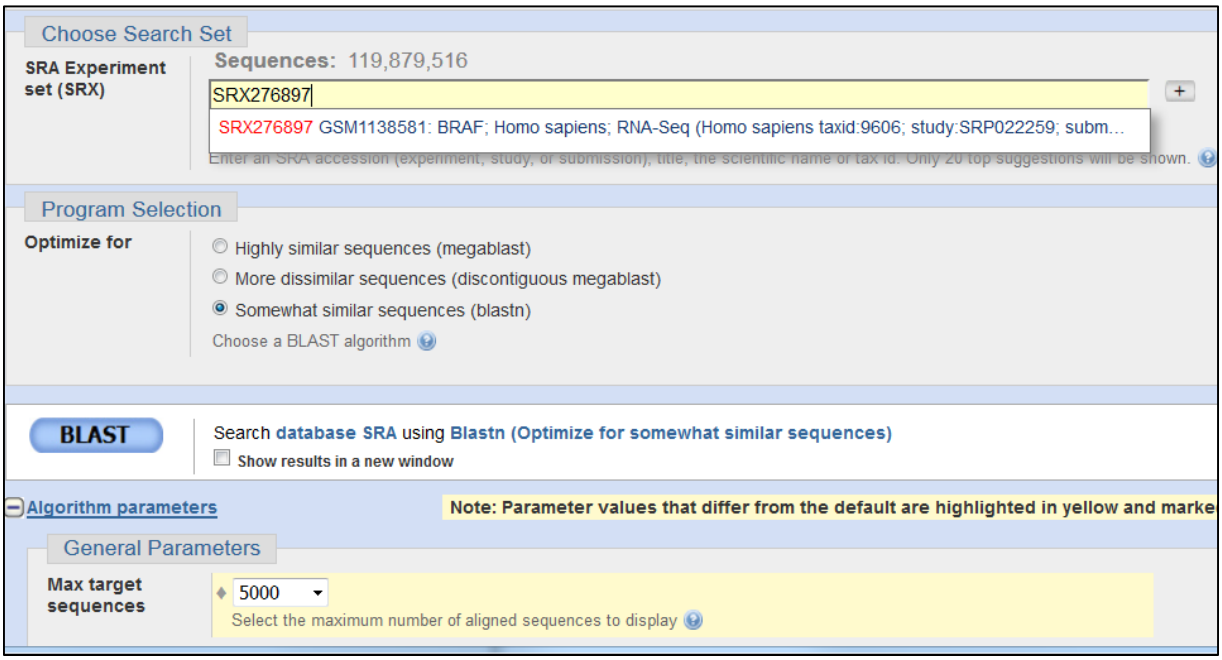
– From the SRA web you can send the results to BLAST:



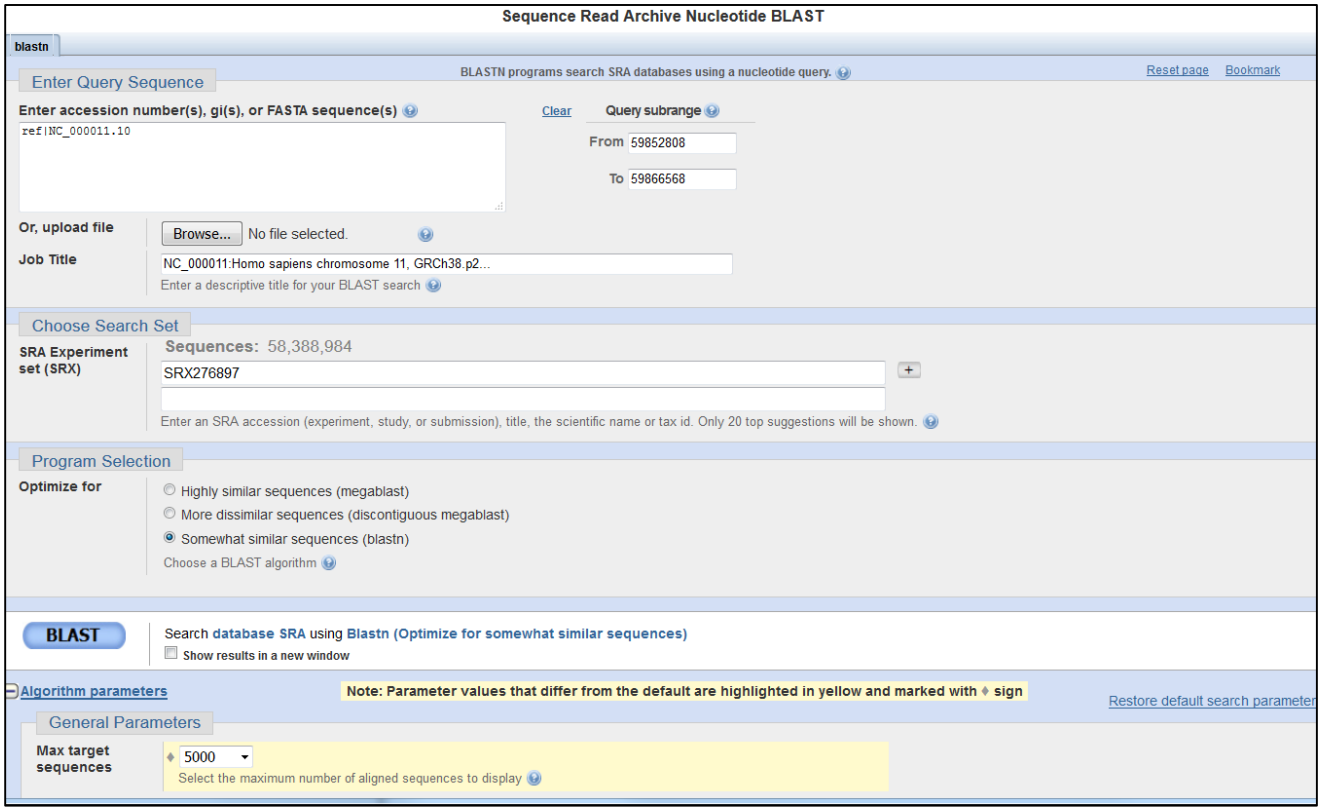
– The two experiments (**SRX276896** and **SRX276897**) are now listed as the database to which you can align your sequence of interest:



– We will work with one experiment at a time, for example **SRX276897**:

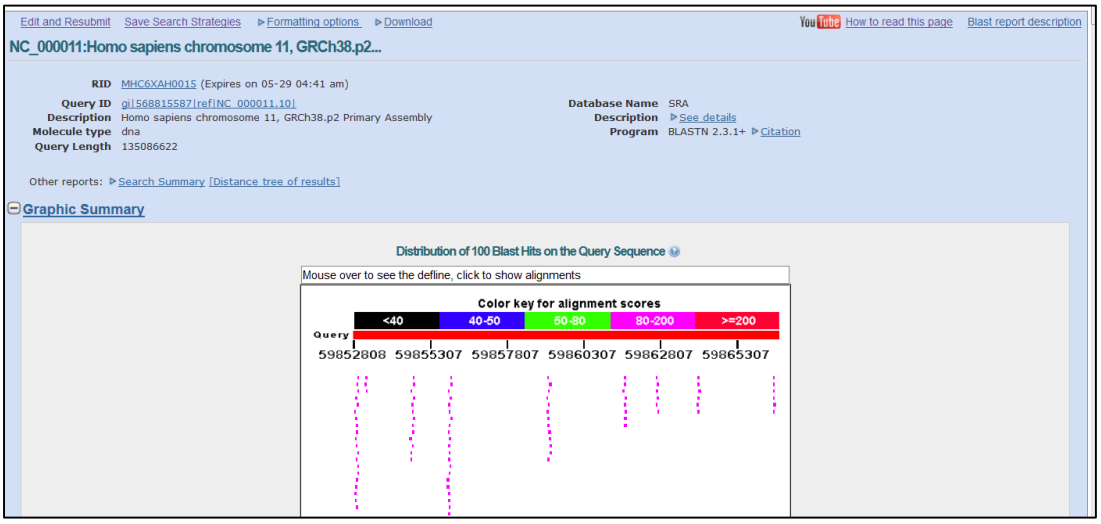


– Our query sequence will be the genomic sequence of the TCN1 gene: **ref|NC_000011.10 (From: 59852808 To: 59866568)**



– See the search results for SRX276897 BLAST:

Q: Can you identify the nine exons on the sequence?_____



– Re-run BLAST for SRX276896 BLAST:

Q: Can you identify the nine exons on the sequence?_____