WIKI MATHS BIOLOGIE OPTION

AIM:

- Compare the expression of tumor suppressor gene and oncogenes in different healthy human tissues.
- Does the environment and lifestyle of individuals/patients influence expression of their genes?

DATA:

- Gtex
- Cancer Census
- CoLaus

TOOLS:

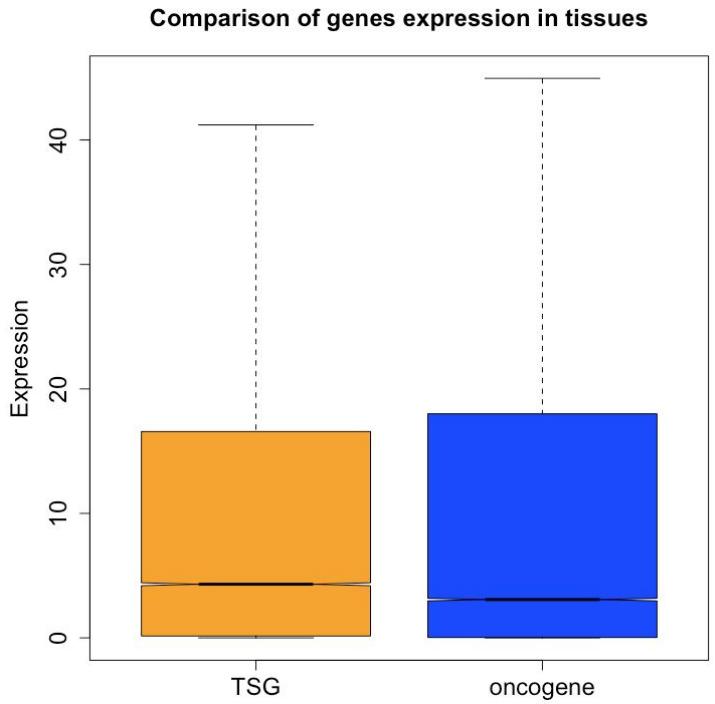
- R logiciel (`wilcox.test`/`cor.test`)
- Excel

Meeting 1

- Topic introduction: TSG are genes which repress cell divisions and oncogenes promote them.
- Discussed our knowledge: pathways, cancer genes, mutation types, command in R.
- Datasets of the project come from Gtext and Cancer Census
- We calculated the mean, the median and sd values for each tissue for the two sets of genes.
- We used the function wilcox.test in R to compare the pooled gene expression of TSG against oncogene
- TSG expression is bigger than the oncogene expression (p-value: $2.2*10^{-16}$). We explain this by the fact that we study healthy individuals and as TSG control the cell cycles and divisions, it makes sense that they are more expressed than oncogenes which promote cell divisions.

	Mean [TSG]	Sd [TSG]	Median [TSG]	Mean [Onanograms]	Sd [Onanograms]	Median [Onanograms]
Adipose - Subcutaneous	28,56948	115,0443	5,553	24,50461	76,9783	4,3195
Adipose - Visceral [Omentum]	24,66789	100,6223	5,022	22,87216	68,14577	3,996
Adrenal Gland	18,27595	65,43048	4,0345	21,79177	152,8392	2,532
Artery - Aorta	28,97127	120,4124	5,011	29,6356	125,6485	4,031
Artery - Coronary	28,09354	109,667	5,64	26,42137	96,36436	4,271
Artery - Tibial	32,2304	126,4282	4,97	28,3972	105,8655	3,335
Bladder	26,92124	98,37946	6,658	24,05283	60,13573	5,212
Brain - Amygdala	10,12665	30,23806	2,467	10,97906	32,39659	1,8695
Brain - Anterior cingulate cortex [BA24]	10,94631	29,51654	2,816	12,23541	38,248	2,205
Brain - Caudate [basal ganglia]	10,52269	28,55927	2,8715	11,88307	38,14007	2,167
Brain - Cerebellar Hemisphere	24,56778	68,08073	5,766	23,8252	61,19691	2,3505
Brain - Cerebellum	22,25047	57,1091	5,832	21,85472	52,84556	2,797
Brain - Cortex	12,79016	30,56126	3,433	13,95085	37,58823	2,535
Brain - Frontal Cortex [BA9]	14.49475	37.31097	3.627	16.25984	51.03073	2.611
Brain - Hippocampus	10.17457	29.4064	2.467	10.68301	27.25485	1.986
Brain - Hypothalamus	11.79781	33.32936	3.231	13.37613	39.79683	2.786
Brain - Nucleus accumbens [basal ganglia]	11.63192	32.80346	3.106	13.0961	38.56814	2.325
Brain - Putamen [basal ganglia]	9.194668	25.87832	2.3735	10.0809	30.5101	1.878
Brain - Substantia nigra	11.50165	37.80381	2.622	12.26337	41.21868	2.3125
Brain - Spinal cord [cervical c-1]						
Breast - Mammary Tissue	26.88569	109.899	5.995	23.00167	62.34893	4.717
Cells - Transformed lymphocytes	36.70534	201.1138	5.311	30.90314	146.7217	1.3625
Cells - Transformed fibroblasts	27.94747	124.0225	5.36	26.07321	85.11357	2.46
Cervix - Ectocervix	30.57809	130.6472	6.744	26.26329	69.15869	4.889
Cervix - Endocervix	33.82682	142.6565	7.918	28.49542	75.83425	5.849
Colon - Transverse	19.00989	72.47766	4.81	16.48882	36.9918	3.841
Esophagus - Mucosa	20.18338	75.2843	4.38	18.78248	45.36853	3.381
Esophagus - Muscularis	22.37935	85.22919	4.855	20.0623	54.18484	3.6295
Fallopian Tube	28.91029	113.0141	7.306	27.01568	79.55456	5.312
Heart - Atrial Appendage	12.40862	42.46935	2.568	11.65692	38.52156	1.979
Heart - Left Ventricle	9.160254	30.68893	1.657	8.613701	41.48662	1.095
Kidney - Cortex	13.31116	47.20261	3.178	11.95037	33.28586	2.523
Kidney - Medulla						
Liver	13.59875	75.4042	1.909	9.422296	29.15854	1.219
Lung	30.46336	130.7939	6.509	28.58495	94.25647	5.761
Minor Salivary Gland	19.24538	74.13581	5.299	17.41709	40.98803	4.253
Muscle - Skeletal	17.63683	97.25411	1.7895	13.69807	101.2711	1.117
Nerve - Tibial	31.87141	107.7036	8.2515	29.82339	96.13515	6.3675
Ovary	36.00509	215.0424	7.061	27.56069	81.66135	4.254
Pancreas	31.41901	630.7947	2.009	8.022838	19.83944	1.4445
Pituitary	19,36548	60,45793	6,4075	20,69549	59,25954	4,272
Prostate	23,87778	87,69146	6,8905	21,2713	46,50901	5,067
Skin - Not Sun Exposed [Suprapubic]	25,59762	107,313	5,129	21,20315	56,48304	4,0665
Skin - Sun Exposed [Lower leg]	25,8711	103,3518	5,391	21,54594	52,21343	4,3655
Spleen	21,17743	91,73515	5,312	18,28641	55,08038	4,466
Stomach	28,09994	132,2603	6,20575	27,49166	114,7699	4,70125
Testis	15,23047	59,57104	3,70325	13,79435	31,93001	2,94925
Thyroid	22,68297	44,74811	8,9505	22,29725	51,4483	5,875
Transformed fibroblasts	27,63089	96,68833	7,96075	26,77107	71,89539	5,58425
Uterus	32,30588	134,3627	7,8425	28,25274	81,75205	5,2685
Vagina	26,72643	106,9609	6,776	23,39426	55,00798	5,4555
Whole Blood	17,65922	108,7121	1,1195	18,98009	90,88153	0,6133

	Wilcoxon
Adipose - Subcutaneous	0,274
Adipose - Visceral [Omentum]	0,3015
Adrenal Gland	0,03225
Artery - Aorta	0,1403
Artery - Coronary	0,1294
Artery - Tibial	0,05304
Bladder	0,07919
Brain - Amygdala	0,09648
Brain - Anterior cingulate cortex [BA24]	0,06287
Brain - Caudate [basal ganglia]	0,08301
Brain - Cerebellar Hemisphere	0,0006338
Brain - Cerebellum	0,0006338
Brain - Cortex	0,03869
Brain - Frontal Cortex [BA9]	0.03223103
Brain - Hippocampus	0.1015366
Brain - Hypothalamus	0.06674195
Brain - Nucleus accumbens [basal ganglia]	0.06131446
Brain - Putamen [basal ganglia]	0.1015581
Brain - Substantia nigra	0.1199954
Brain - Spinal cord [cervical c-1]	
Breast - Mammary Tissue	0.2275324
Cells - Transformed lymphocytes	0.005331283
Cells - Transformed fibroblasts	0.01510465
Cervix - Ectocervix	0.1597437
Cervix - Endocervix	0.0872277
Colon - Transverse	0.1578968
Esophagus - Mucosa	0.1572102
Esophagus - Muscularis	0.1024
Fallopian Tube	0.09768
Heart - Atrial Appendage	0.1074
Heart - Left Ventricle	0.05372
Kidney - Cortex	0.1225
Kidney - Medulla	
Liver	0.01337
Lung	0.3683
Minor Salivary Gland	0.1663
Muscle - Skeletal	0.0185
Nerve - Tibial	0.2237
Ovary	0.01443
Pancreas	0.05007
Pituitary	0,03404
Prostate	0,1042
Skin - Not Sun Exposed [Suprapubic]	0,09489
Skin - Sun Exposed [Lower leg]	0,1181
Spleen	0,1228
Stomach	0,1519
Testis	0,1515
Thyroid	0,001147
Transformed fibroblasts	0,0885
Uterus	0,07318
Vagina	0,2054
Whole Blood	0,06435

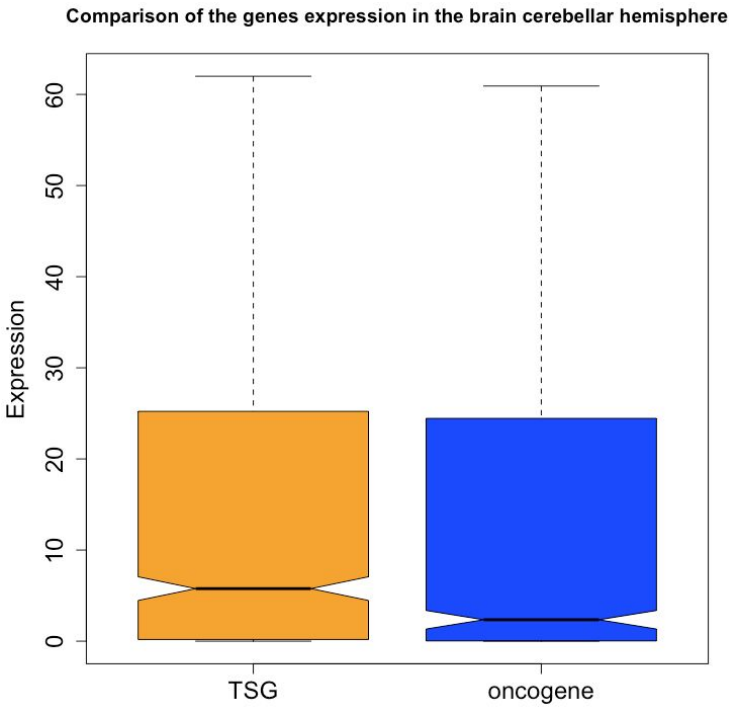


Meeting 2 *(09.0.18)*

- We used a wilcox.test to compare the gene expression between tumor suppressor gene and oncogene for each tissue
- Bonferroni correction of wilcox p.values *(p.values*number of tests (52 here))*
- We found that the brain cerebellar hemisphere showed significant differences:

p.value = 0.033. We explain this significant result by the fact that there is very few new cells in the brain so the ones that are there have to be protected better as in any other tissue.

- We started an across tissue analysis (Brain cerebellar hemisphere)



Meeting 3 *(16.03.18)*

- Because our results in the across tissue analysis weren't coherent (not significant result between brain cerebellar hemisphere and bladder) with the wilcox.test. We did also a correlation test with R (`cor.test`) and we found that there was a significant difference between all the tissues.

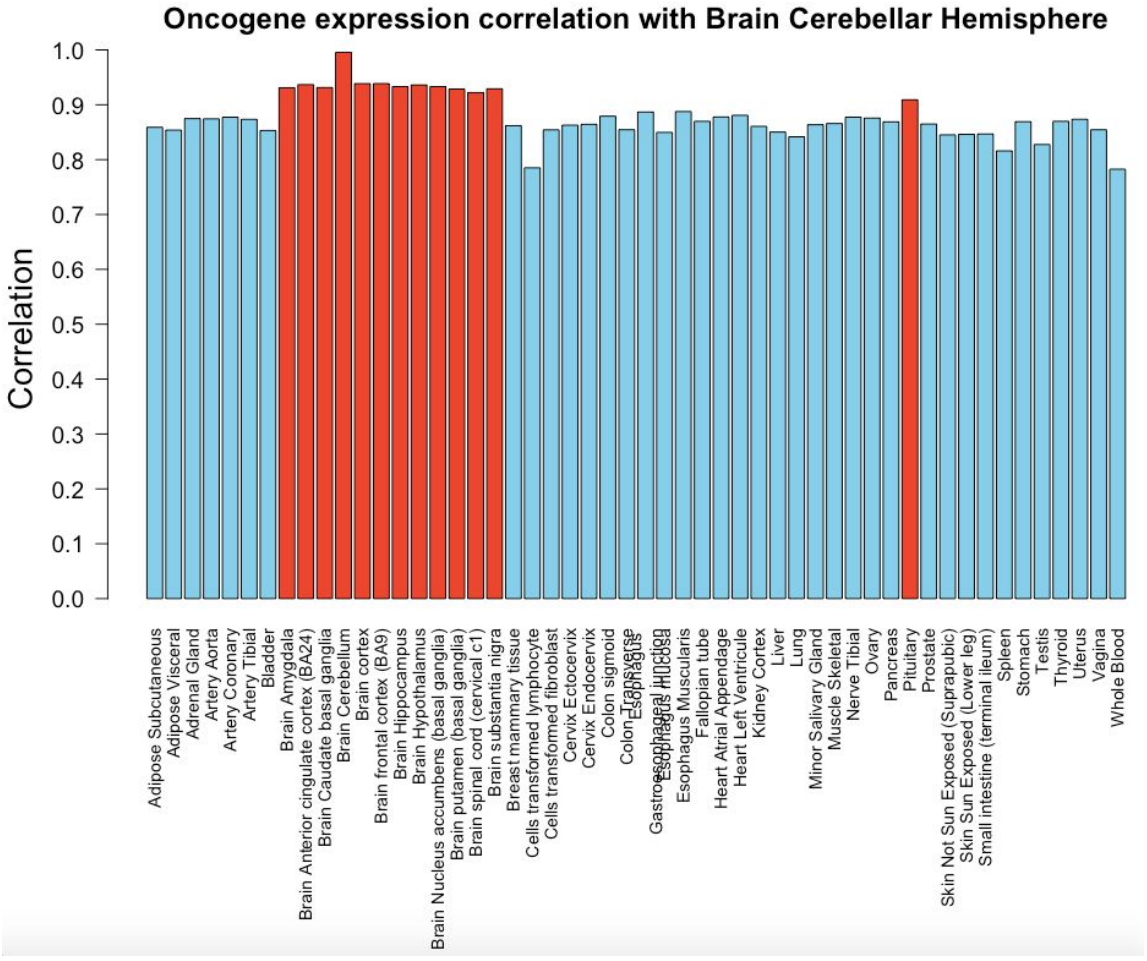
In fact, the wilcox.test measures the difference between expression levels for each tissues. However, the spearman correlation test measures dependence between two variables.

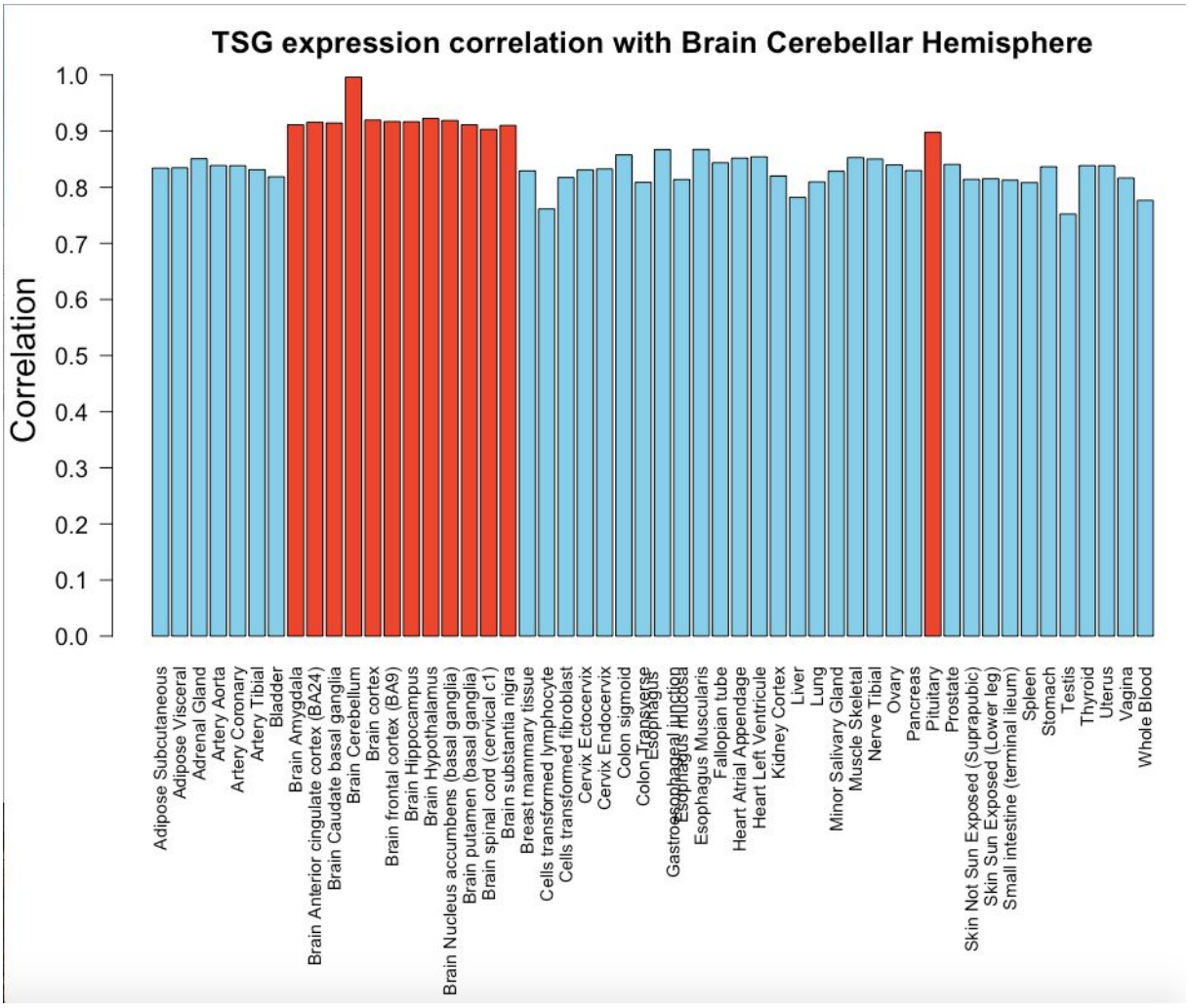
- We have pooled the gene expression values from all tissues and we have made boxplot of them. We have suppressed the outliers.

- We plan to analyze if age, sex and death cause have an effect on gene expression *(the second part of our aim)*.

Meeting 4 *(23.03.18)*

- In red are brain tissues. They have a correlation of over 90%.

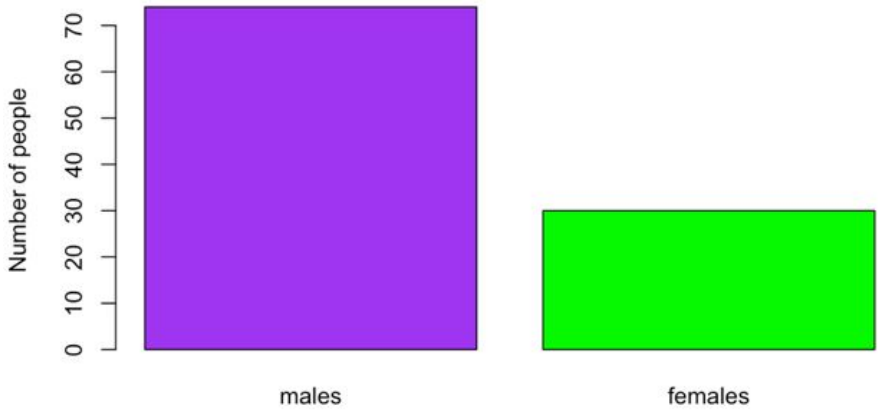




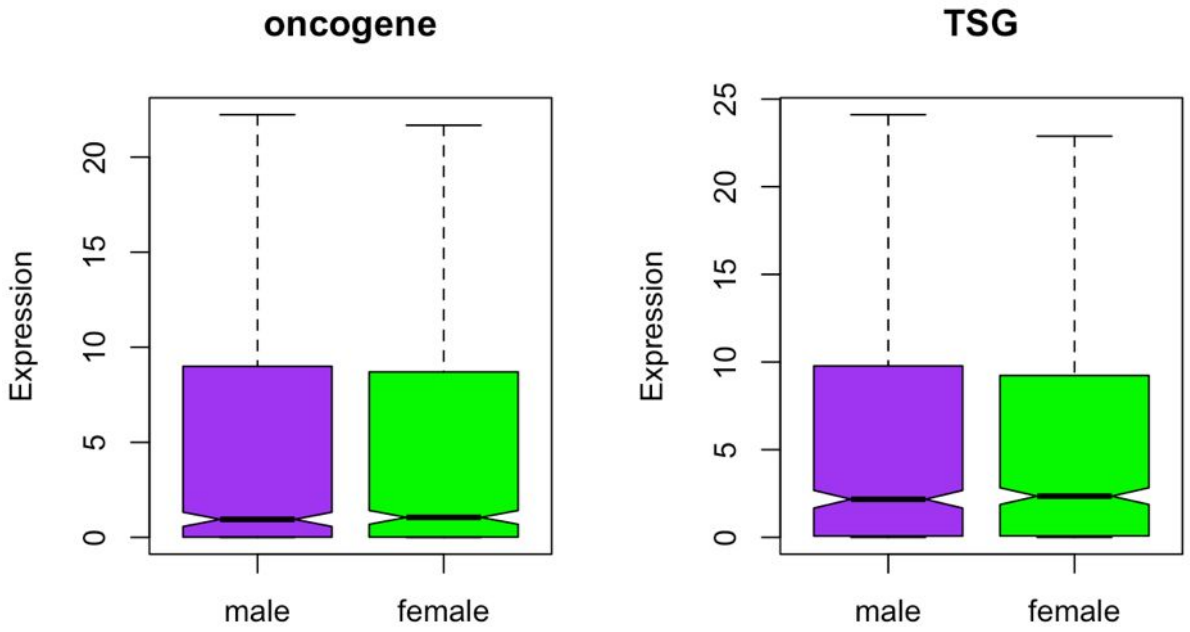
- These graphs show that the DNA has a major importance in the expression of cancer genes as the level of correlation is very high through all the tissues.
- That's the end of our first aim and the beginning of our second aim: we will see if other factors also have an effect on this expression.

Meeting 5

<u>Female/Male</u>



We can see that males and females are not equally distributed.



Next we did the comparison of oncogenes and TSG expression between males and females and we found no significant result.

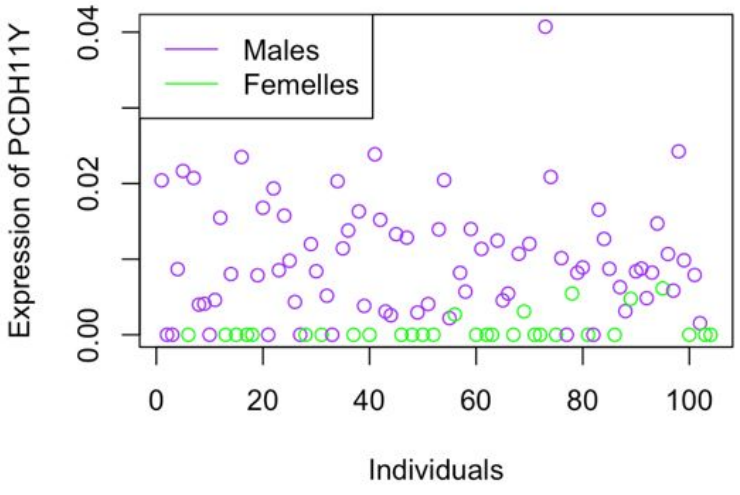
Cancer genes are essential and expressed in every tissue so we don't expect a difference in expression of cancer genes between males and females.

Then, we did a comparison by using wilcox test of each genes expression between male and female. We found four genes (2 oncogenes and 2 TSG) expressed differently between males and females :

TSG: PCDH11Y ; BCORP1

Oncogenes: ARSFP1 ; PARP4P1

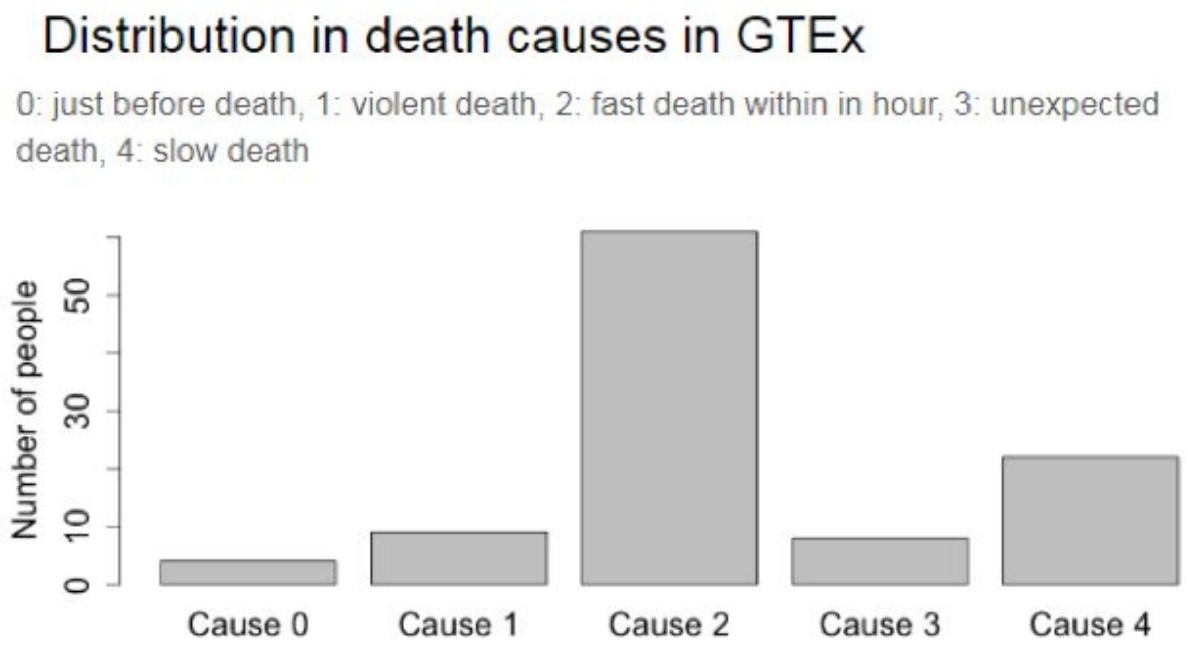
We looked at the distribution of PCDH11Y :



We see that this gene is mainly expressed in males and not female. When we look at the name of this gene (« Protocadherin 11 Y-Linked ») we see that PCDH11Y is located on the Y chromosome and that's why it is expressed only in males.

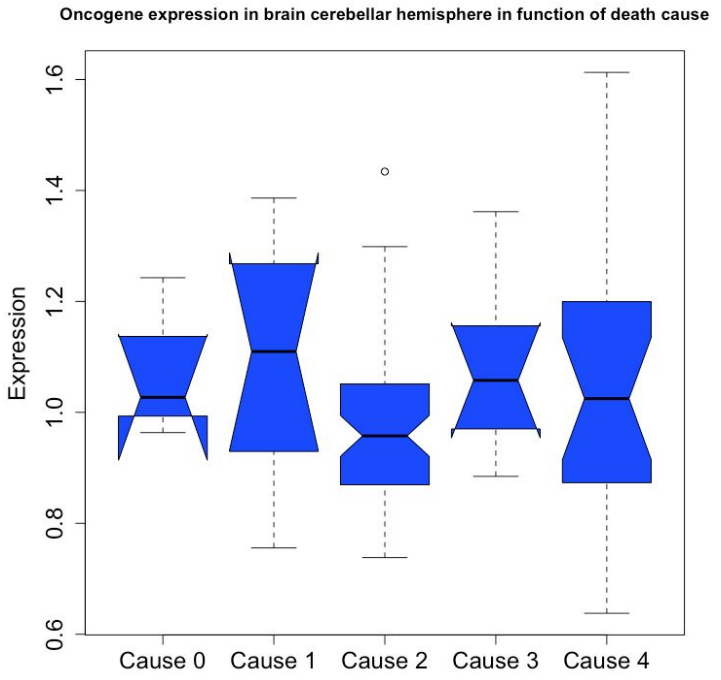
Cause of death

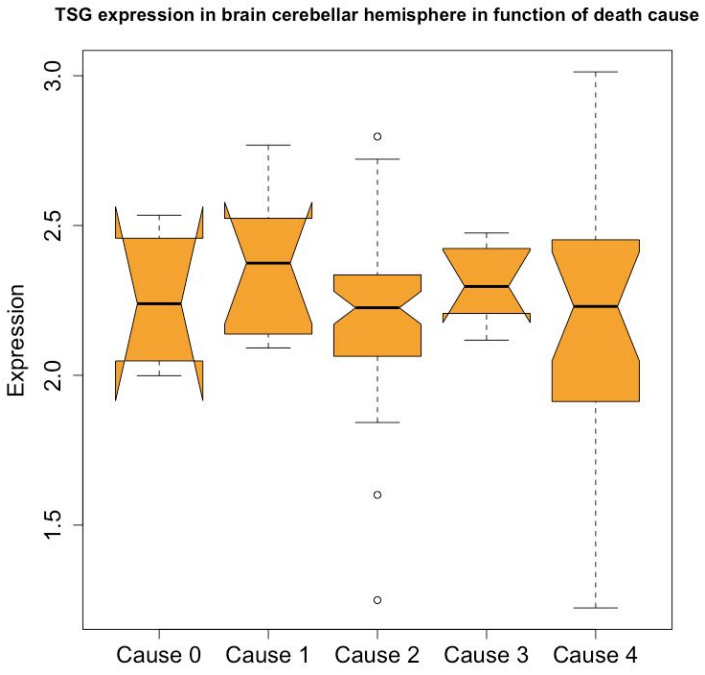
First, we looked through a graph the number of people who died from different causes.



We found that there was between 4 and 9 people who died of cause 0, 1 and 3. 22 people who died of cause 4 and 61 who died of cause 2.

In a second step, we created a boxplot representing to the left the expression of the oncogenes according to the cause of the death, and on the right the expression of the tsg according to the causes of the death.

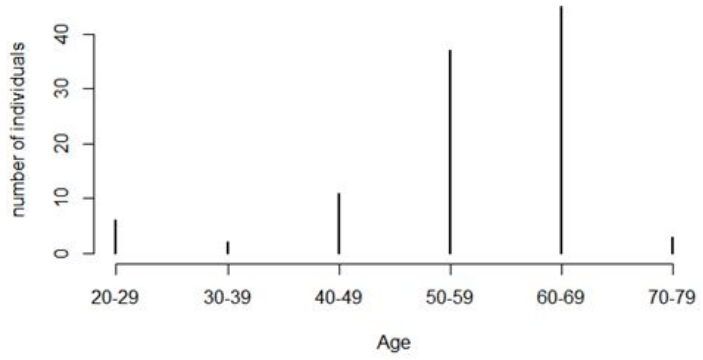




- We observe no obvious difference, and we find no significant result.
- The expression of cancer genes is therefore not affected by the cause of death.

Class of age

We created a graph representing the distribution of the number of people according to the age groups.



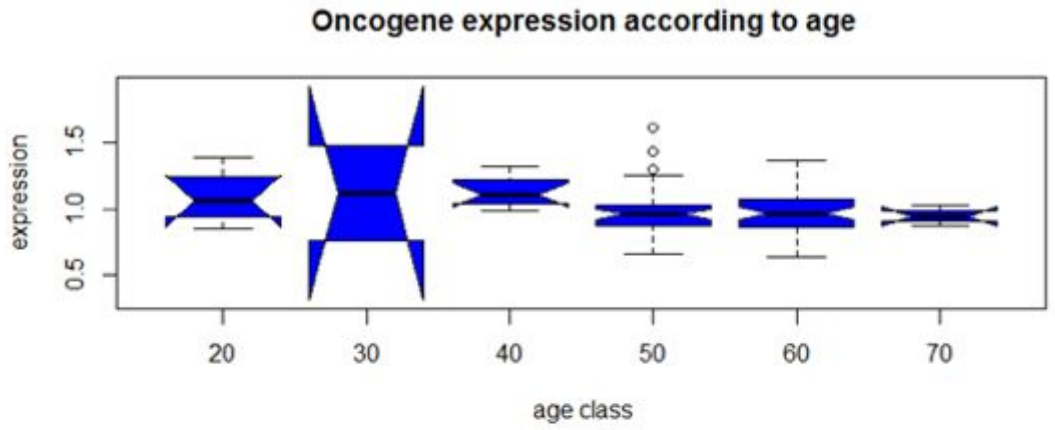
We observe that there are between 2 and 6 people in the age groups of 20 to 29 and 70 to 79 years, while there are between 11 and 45 individuals between 40 and 69 years old

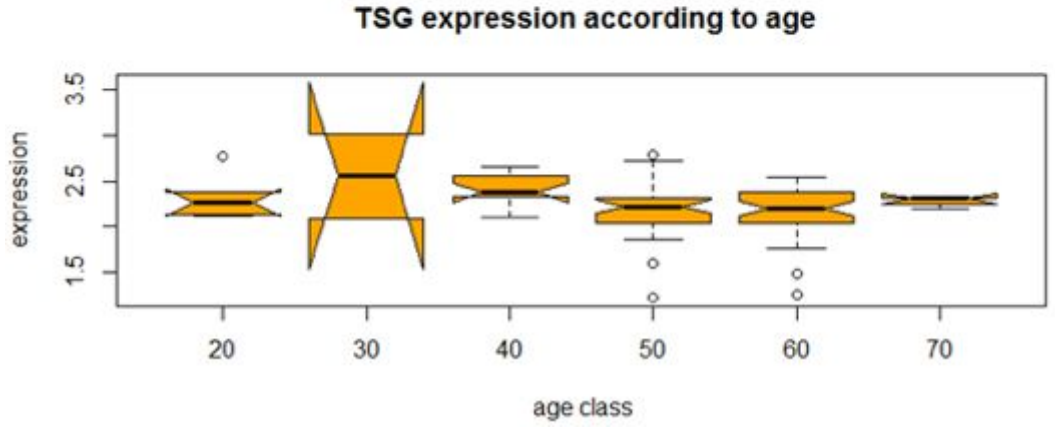
Thus, we expect to find more significant results in age groups between 40 and 69 years old, because the population size for these classes is larger than for the others.

After using a wilcox.test comparing the expression between TSG and oncogene by age group, there is a significant difference for age groups of 40 to 69 years, but not in other population classes (because the size of population is too small).

Tumor suppressor genes are found to be more expressed than oncogenes in these age groups. Which is consistent with the results found previously.

Then, we have created boxplots. They allow us to see if there is a correlation between age and the expression of cancer genes. At a glance these graphs make us think of a negative correlation.





We use a spearman's correlation test to confirm our hypothesis:

- Correlation coefficient = - 0.21
- P-value = 0.03

Spearman's correlation is positive: the expression of cancer genes decreases with age.

After a multiple correlation tests for each gene separately, we found that:

- Spearman correlation test was significant for 22 oncogenes out of 1416

Spearman correlation test was significant for 13 tsg out of 889

We found the function of each of its genes through the Gtex database:

ONCOGENE Name	Function	Correlation	P.value
Aliases for KITLG GENE KIT Ligand	Role in the regulation of cell survival and proliferation, hematopoiesis, stem cell maintenance, gametogenesis	-4.303293	3.841179e-05
Aliases for ERBB3 Gene Erb-B2 Receptor Tyrosine Kinase 3	Epidermal growth	-5.056881	1.861667e-06

Aliases for SMARCE1 Gene SWI/SNF Related, Matrix Associated, Actin Dependent Regulator Of Chromatin, Subfamily E, Member 1	Involved in transcriptional activation and repression of select genes by chromatin remodeling	-4.579035	1.311316e-05
Aliases for ERGIC2 Gene Endoplasmic Reticulum-Golgi Intermediate Compartment Protein 2	Possible role in transport between endoplasmic reticulum and Golgi.	-4.460224	2.093836e-05
Aliases for PLPPR4 Gene Phospholipid Phosphatase Related 4	Hydrolyzes lysophosphatidic acid (LPA). Facilitates axonal outgrowth during development and regenerative sprouting	-5.574220	2.006253e-07
Aliases for BCL2L2 Gene BCL2 Like 2	Anti- and pro-apoptotic regulators. Promotes cell survival.	-4.442850	2.240762e-05
Aliases for RERG Gene RAS Like Estrogen Regulated Growth Inhibitor	A member of the RAS superfamily of GTPases, inhibits cell proliferation and tumor formation	-5.345137	5.454002e-07
Aliases for PDGFRA Gene Platelet Derived Growth Factor Receptor Alpha	Role in the regulation of embryonic development, cell proliferation, survival and chemotaxis	-4.608005	1.168632e-05
Aliases for SCARB2 Gene Scavenger Receptor Class B Member 2	Participate in membrane transportation and the reorganization of endosomal/lysosomal compartment	-5.054206	1.882679e-06
Aliases for IFNAR1 Gene Interferon Alpha And Beta Receptor Subunit 1	Component of the receptor for type I interferons. Type I interferon binding activates signaling cascade	-4.456897	2.121229e-05
Aliases for AFF3 Gene AF4/FMR2 Family Member 3	Putative transcription activator that may function in lymphoid development and oncogenesis.	-4.880886	3.869172e-06
Aliases for MARCH4 Gene Membrane Associated Ring-CH-Type Finger 4	Member of the MARCH family of membrane-bound E3 ubiquitin ligases	-4.782109	5.796904e-06
Aliases for SETD7 Gene SET Domain Containing Lysine Methyltransferase 7	Methylate histone lysinesand Chromatin organization.	-4.263143	4.476752e-05
Aliases for PLPPR1 Gene Phospholipid Phosphatase Related 1	Mediate lipid phosphate phosphatase activity in neurons and are known to be involved in neuronal plasticity This gene is strongly expressed in brain	-5.161622	1.196614e-06
Aliases for USP6NL Gene USP6 N-Terminal Like	Involved in retrograde transport from the endocytic pathway to the Golgi, recycling endosomes to the trans-Golgi network. Required for structural integrity of the Golgi complex.	-4.462358	2.076440e-05
Aliases for GART Gene Phosphoribosylglycinamide Formyltransferase, Phosphoribosylglycinamide	Purine biosynthesis.	-4.664714	9.315698e-06

Synthetase, Phosphoribosylaminoimidazole Synthetase			
Aliases for ARPC5 Gene Actin Related Protein 2/3 Complex Subunit 5	Involved in regulation of actin polymerization	-4.507641	1.738641e-05
Aliases for PSIP1 Gene PC4 And SFRS1 Interacting Protein 1	Involved in neuroepithelial stem cell differentiation and neurogenesis	-4.221471	5.242974e-05
Aliases for PARM1 Gene Prostate Androgen-Regulated Mucin-Like Protein 1	May regulate TLP1 expression and telomerase activity, thus enabling certain prostatic cells to resist apoptosis	-5.884783	5.009928e-08
Aliases for CARNS1 Gene Carnosine Synthase 1	Catalyzes the synthesis of carnosine and homocarnosine which are found mainly in skeletal muscle and the central nervous system	-4.802283	5.339497e-06
Aliases for PTPN11 Gene Protein Tyrosine Phosphatase, Non-Receptor Type 11	PTPs are known to be signaling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation.	-4.539780	1.531808e-05
Aliases for RELN Gene Reelin	Extracellular matrix serine protease that plays a role in layering of neurons in the cerebral cortex and cerebellum. Regulates microtubule function in neurons and neuronal migration	-4.777087	5.916551e-06

TSG Name	function	Correlation	P.value
Aliases for SMARCE1 Gene SWI/SNF Related, Matrix Associated, Actin Dependent Regulator Of Chromatin, Subfamily E, Member 1	Involved in transcriptional activation and repression of select genes by chromatin remodeling	-4.579035	1.311316e-05
Aliases for ATRN Gene Attractin	Involved in the initial immune cell clustering during inflammatory response and may regulate chemotactic activity of chemokines.	-4.528936	1.598783e-05
Aliases for RNF125 Gene Ring Finger Protein 125	E3 ubiquitin-protein ligase that mediates ubiquitination and subsequent proteasomal degradation of target proteins	-4.806180	5.255274e-06
Aliases for RNF141 Gene Ring Finger Protein 141	Involved in protein-DNA and protein-protein interactions	-5.134590	1.341811e-06
Aliases for RNF11 Gene Ring Finger Protein 11	For protein-protein interactions	-4.417911	2.469183e-05
Aliases for FHOD3 Gene Formin Homology 2 Domain Containing 3	This protein is thought to play a role in actin filament polymerization in cardiomyocytes	-4.386180	2.792465e-05
Aliases for NCOA4 Gene	Enhances the androgen receptor	-5.333012	5.747130e-07

Nuclear Receptor Coactivator 4	transcriptional activity in prostate cancer cells		
Aliases for CDH13 Gene Cadherin 13	Protein acts as a negative regulator of axon growth during neural differentiation. It also protects vascular endothelial cells from apoptosis due to oxidative stress, and is associated with resistance to atherosclerosis.	-6.022826	2.674792e-08
Aliases for PTPRK Gene Protein Tyrosine Phosphatase, Receptor Type K	PTPs are known to be signaling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation.	-4.420733	2.442255e-05
Aliases for NRG1 Gene Neuregulin 1	Cell-cell signaling and plays a critical role in the growth and development of multiple organ systems	-5.643546	1.476406e-07
Aliases for PTPRT Gene Protein Tyrosine Phosphatase, Receptor Type T	PTPs are known to be signaling molecules that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle, and oncogenic transformation.	-4.451921	2.162851e-05
Aliases for ZNF277 Gene Zinc Finger Protein 277	May be involved in transcriptional regulation	-4.363116	3.052679e-05
Aliases for ATP1A1-AS1 Gene ATP1A1 Antisense RNA 1	Is an RNA Gene, and is affiliated with the non-coding RNA class.	-4.465195	2.053534e-05

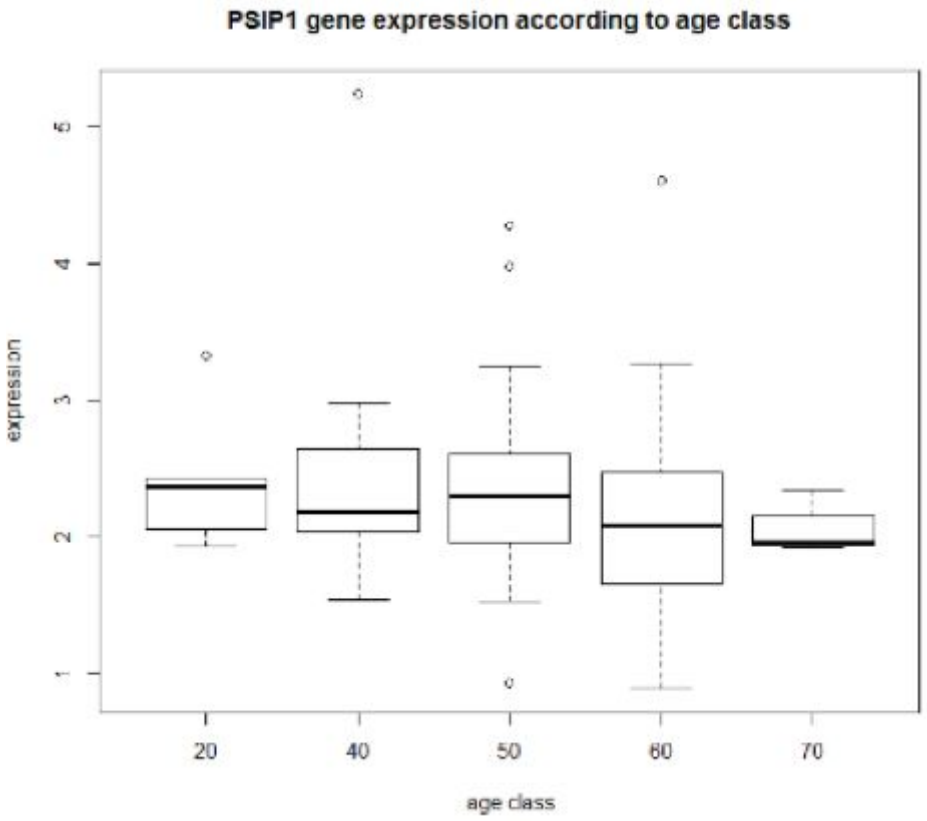
Most of these genes are involved in cell growth, and in neuronal development.

We selected two genes as an example:

- The **PSIP1 gene** that is involved in neuroepithelial stem cell differentiation and neurogenesis. We have traced the expression of this gene according to age using boxplots and we observe that the expression of PSIP1 decreases with age.

  - Correlation test: -4.221471

  - P-value: 5.242974e-05

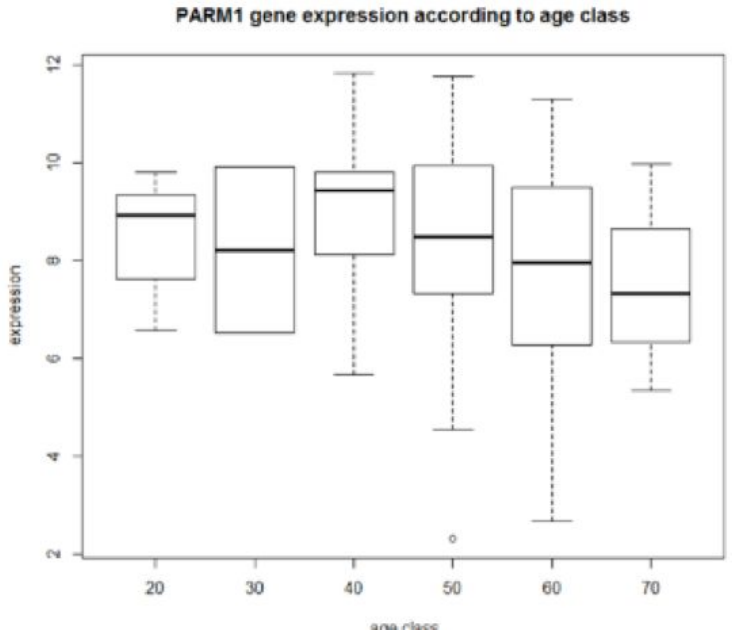


- The **PARM1 gene** that is involved in the regulation of telomeres, and is therefore directly related to aging.

We have traced the expression of this gene according to age using boxplots and we observe that the expression of PARM1 decreases with age.

  - Spearman correlation test: -5.884783

  - P-value: 5.009928e-08



So, we found that:

Many of the genes are involved in cell proliferation.

Many of them are also either expressed in brain or involved in neuronal development

Age does seem to affect cancer gene expression

=> brain tissue loses its ability to proliferate with age

After that, we began to work with CoLaus by choosing a phenotype to work on.

CoLaus - A phenotypic database

After having worked on GTEx data, we decided it would be interesting to test if there were any correlations with cancer gene expression and certain phenotypes.

CoLaus is a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. The study consists of approximately 6000 people in Lausanne with data or around 500 gene expressions. Although the focus of the study is not cancer, it was interesting to see if we would find any significant correlations with oncogenes end TSG gene expressions. In other words, we focused on trying to determine if there were any new genetic determinants associated with Cancer gene expression.

Gene expressions of the CoLaus data are from a lymphoma blastoma cell line. As this is a derived cell line we may not catch environmental effects, but only the effects of mutations, that we checked with many different phenotypes thereafter.

A summary of our findings (phenotype tested, type of test pertaining to the type of data and examples of the significant genes) can be found in the following table.

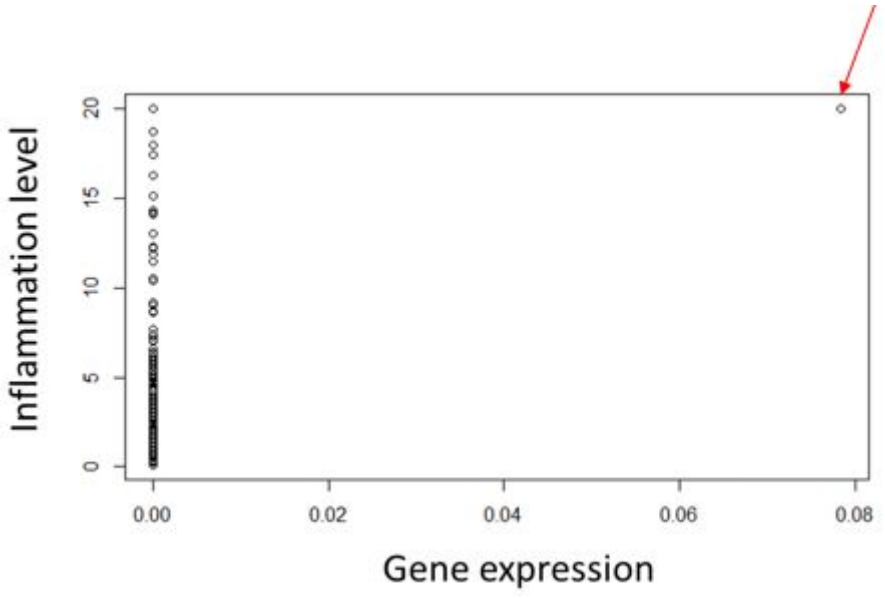
Phenotype	Type of data	Number of significant genes (out of 555 genes)	Examples
Inflammation	Continuous (Correlation test)	2	TARM1: T Cell-Interacting, Activating Receptor on Myeloid Cells NPM1P36: Nucleophosmin 1 Pseudogene
Prostate	Binary (Wilcox test)	1	ZNF299P: zinc finger protein 299, pseudogene
BMI	Continuous (Correlation test)	5	CFH: Complement Factor H PWRN4: Prader-Willi Region Non-Protein Coding RNA 4
Osteoporosis	Binary (Wilcox test)	3	DDX3X: DEAD-Box Helicase 3, X-Linked NF1P6: Neurofibromin 1 Pseudogene
Parkinson's	Binary (Wilcox test)	4	TARM1: T Cell-Interacting, Activating Receptor on Myeloid Cells NF1P5: Neurofibromin 1 Pseudogene
Hayfever	Binary (Wilcox test)	0	
Breast Cancer	Binary (Wilcox test)	6	ARMS2: Age-related Maculopathy susceptibility 2 MYOD1: Myogenic Differentiation 1
Blood Lipid levels	Continuous (Correlation test)	6	KDM6A: Lysine Demethylase 6A DDX3X: DEAD-Box Helicase 3, X-Linked
Diabetes type 1	Binary (Wilcox test)	3	RPL5P2: Ribosomal Protein L5 Pseudogene SCARNA14: Small Cajal Body – Specific RNA
Diabetes type 2	Binary (Wilcox test)	4	NF1P3: Neurofibromin 1 pseudogene PWRN4: Prader-Willi Region Non-Protein Coding RNA 4

For most of the phenotypes that we tested, we found a few significant results. However, the significant genes found were not particularly highly expressed, and were almost never involved in the relevant phenotype: there was a correlation, but no causation.

These results were partially to be expected because, as we mentioned, the way the experiment is designed is that gene expression is only affected by DNA ant not the environment, so we do not expect to see any effect of the phenotype; phenotype is not in blood cells.

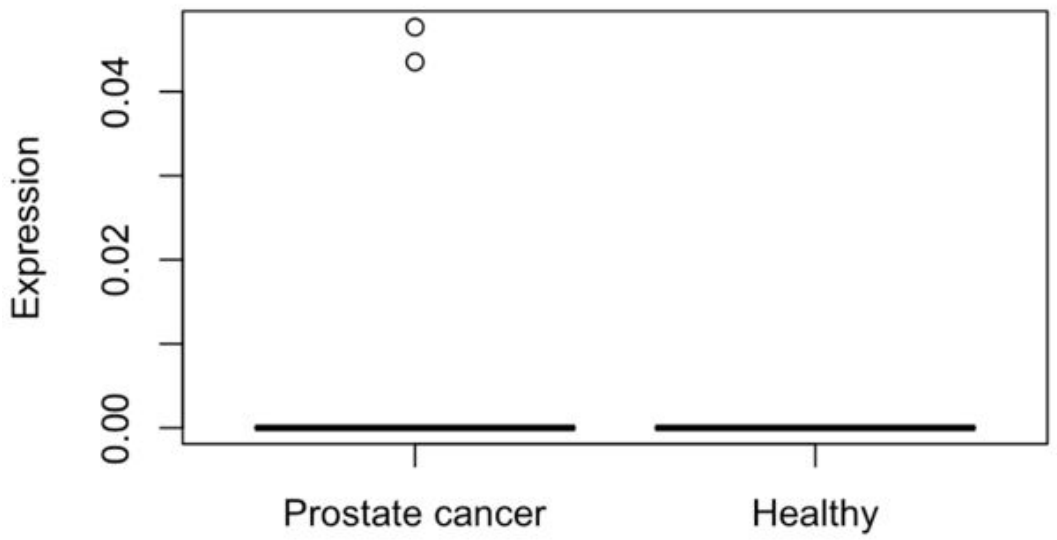
Concerning inflammation, we knew beforehand that there was a known correlation between cancer gene expression and inflammation.

One of the genes that proved significant after our tests was the TARM1 gene. This gene is a T Cell-Interacting, Activating Receptor on Myeloid Cells. When we visualised the gene expression in the form of a graph, we saw that there was only one individual with a significantly higher expression of this gene. This could ultimately be an indication that this specific sample may have something else going on, a confounding factor in the background.



Throughout our tests, this gene revealed itself as significant to other phenotypes as well, thus enforcing our previous hypothesis that one sample may have introduced a bias in our results. It is interesting to note that the original correlation between this gene and oncogenes was a mere 0.22, and no significant correlation was found with Tumour Suppressor Genes.

We also analysed the phenotype « prostate » and found one gene (ZNF299P) that is significantly expressed differently between healthy subject and prostate cancer patient. We looked at the distribution of this gene:



In fact this gene is not expressed in healthy subject but only in 2 cancer patients. And even in these 2 patients the expression is very low. The analysis was thus falsified by these outliers. So we cannot draw of scientific conclusion of this analysis.

Knowing that this gene is a zinc finger protein and by looking in the literature, we saw that certain zinc to finger proteins are known to be associated with prostate cancers.

Conclusion

The TSG are more highly expressed than oncogenes because they control the cell divisions and we study healthy individuals. This difference is especially significant in the brain cerebellar hemisphere as there is few cell renewal. Sex and cause of death don’t seem to affect the expression but the age shows significant results. CoLaus isn’t a good database to see phenotypes effect.