Finding differentially expressed genes in skin cancer

Neus Amat Sorando, Aleix Canalda Baltrons, Maria Díaz Ros, Judit Tella Vila

Contents

1 Abstract

Cancers can be divided in different stages depending on the tumour progression. The aim of this study is to analyse the differential RNA-seq expression between early and late stages of cancer patients. Several under or overexpressed genes were obtained and then there analysed to find a correlation between them and skin cancer. Finally, eight genes where the most probably related with the disease.

2 Introduction

Skin cancer is an abnormal growth of skin cells. It generally develops in areas that are exposed to the sun, but it can also form in places that do not normally get sun exposure. There are three major types of skin cancer:

- Basal cell carcinoma: the most common type.
- Squamous cell carcinoma: it develops in the outer layers of the skin and it is more aggressive than the previous one.
- Melanoma: it is the less common but the most dangerous one and it causes the majority of skin cancer-related deaths each year.

Except in rare instances, most skin cancers arise from DNA mutations induced by ultraviolet light affecting cells of the epidermis. Therefore, skin cancers typically arise in areas of the skin exposed to the sun repeatedly over many years such as on the face and nose, ears, back of the neck, and the bald area of the scalp. After someone is diagnosed with cancer, doctors will try to figure out if it has spread, and if so, how far. This process is called staging. The stage of a cancer describes how much cancer is in the body. It helps determine how serious the cancer is and how best to treat it. The stage is based on 3 key pieces of information:

- The size of the tumor and if it has grown deeper into nearby structures or tissues, such as a bone.
- If the cancer has spread to nearby lymph nodes.
- If the cancer has spread (metastasized) to distant parts of the body.

he earliest stage of skin cancer is stage 0 (also called carcinoma in situ, or CIS). The other stages range from I (1) through IV (4). As a rule, the lower the number, the less the cancer has spread. A higher number, such as stage IV, means cancer has spread more. In this study, we will consider stages I and II as early tumours and satges III and IV as late tumours.

3 Methods

In this study we performed a RNA-seq expression analysis on RNA-seq data of patients on different stages of skin cancer. The objective is to find genes with different expression depending on the stage the patients are.

3.1 Packages and tools used

For the analysis of our data we have used the following `packages` on R and/or R studio:

- **summarizedExperiment**. It contains one or more assays, each represented by a matrix-like object of numeric or other mode. The rows typically represent genomic ranges of interest and the column represent samples.
- **edgeR**. It assists in doing empirical analysis of digital gene expression data. It implements a range of statistical methodology based on the negative binomial distributions. It can be applied to different types of analysis such as RNA-seq, ChIP-seq, SAGE and CAGE.
- **DESeq2**. It estimates variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution.
- **tweeDEseq**. It is used to study differential expression analysis of RNA-seq using the Poisson-Tweedie family of distributions.
- **tweeDEseqCountData**. It is employed to illustrate the use of the Poisson-Tweedie family of distributions with the tweeDEseq package.
- **GOstats**. It is set of tools for interacting with GO and microarray data. It contais a variety of basic manipulation tools for graphs, hypothesis testing and other simple calculations.
- **annotate**. It provides an environment to do annotations for microarrays.
- **org.Hs.eg.db**. Genome wide annotation for Human, primarily based on mapping using Entrez Gene identifiers.
- **biomaRt**. It provides an interface to a growing collection of databases implementing the BioMart software suite. It enables retrieval of large amounts of data in a uniform way without the need to know the underlying database schemas or write complex SQL queries. Examples of BioMart databases are Ensembl, COSMIC, Uniprot, HGNC, Gramene, Wormbase and dbSNP mapped to Ensembl. These major databases give biomaRt users direct access to a diverse set of data and enable a wide range of powerful online queries from gene annotation to database mining.
- **ggplot2**. It's a system to create graphs. You have to provide the data and specify how to map variables to aesthetics.
- **ggrepel**. Provides text and label geoms for 'ggplot2' that help to avoid overlapping text labels.

3.2 Data description

Firstly, we must download all the information related to the patients and their skin cancer which will be saved inside an object in our R environment. The number of columns represent the number of total patients, which would be 473, while the number of rows tells us how many genes we're looking at, which would be 58.037 genes.

What is of utmost interest is to eliminate all those patients that don't present information regarding their gene expression, in other words, that present an "NA". Therefore, the next move was to eliminate all those patients with an "NA", which comprised 60 of them. That means that we were left with 413 patients.

In order to determine which genes are over or under expressed we have to create two separate groups, those patients in an early cancer stage and those in a late cancer stage, and compare the genes between these two groups. Specifically for our case of skin cancer, 218 patients were inside the early stage group while 195 patients were in the later stage group.

3.3 Quality control

After taking a look at our data, the next step is to normalize the RNA-seq information. There are three main reasons why we have to do that:

1. The number of counts is related to sequencing depth. In other words, a gene can seem to be more expressed than another, when in reality it has simply been sequenced more deeply.
2. The number of counts is related to transcript length. If a transcript is longer than another, it will have more counts associated to it, and may seem that both transcripts present the same expression if both of them present the same amount of counts, but in reality the shorter transcript is more expressed since in less length it has more counts.
3. The number of counts is proportional to the mRNA level. There are some variables that can skew some normalization methods, that's why it's important to take into account the RNA composition.

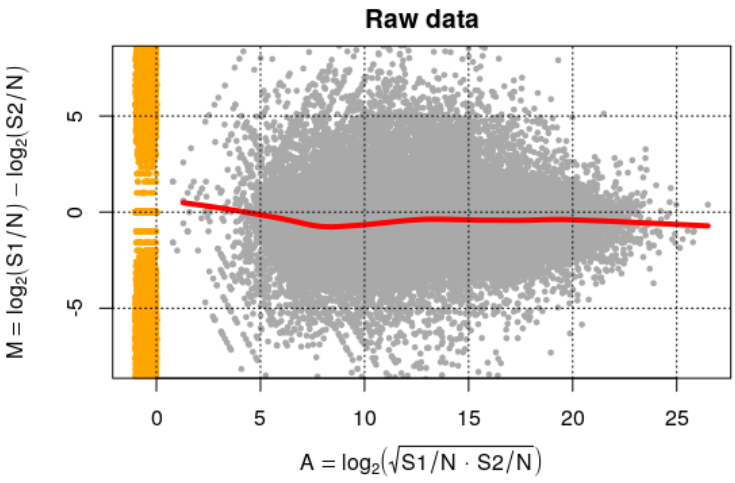

Figure 1 – Raw data representing the differential expression without normalization.

As can be seen in Figure X, without a normalization our raw data is not as reliable since it doesn't overlap with 0, it's skewed and makes it seem that most genes are underexpressed, when what we should expect is for genes to not be differentially expressed at all. However we can fix this by using one of two normalization methods:

- RPKM (Reads Per Kilobase Million) This method only corrects for sequencing depth and gene length but not for the third reason given above. It's based on dividing counts by the transcript length (kb) times the total number of millions of mapped reads. The result of this normalization in our data can be observed in figure Z.

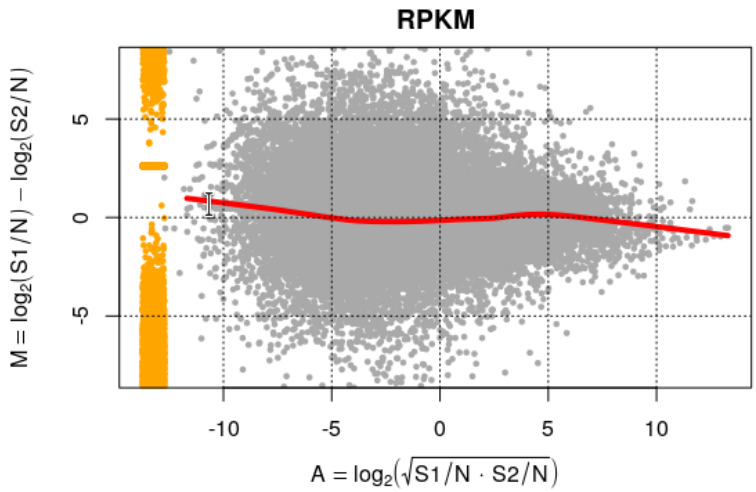

Figure 2 – Data after normalization using the RPKM method.

- TMM (Trimmed Mean of M values) Unlike RPKM, this method accounts for all three reasons given above, and we can observe in Figure Y that the normalization is at its best when this method is used. There is a clear improvement when comparing all three methods, especially when we compare the raw data with the TMM normalization.

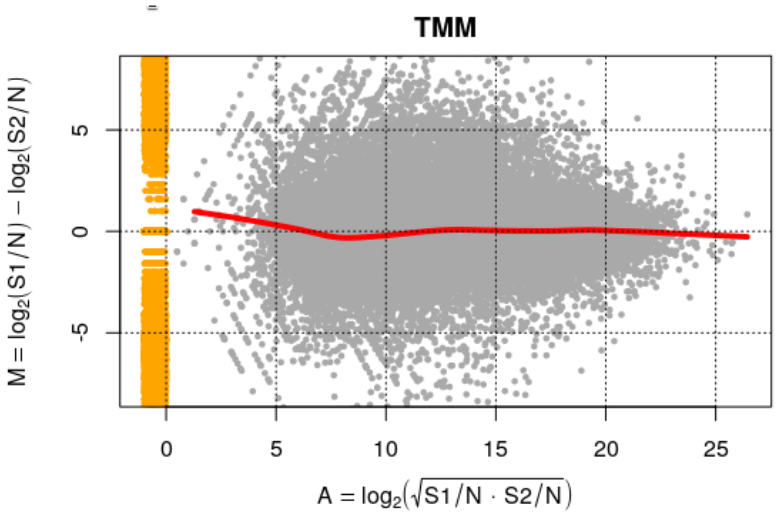

Figure 3 – Data after normalization using the TMM method.

4 Results

After normalizing the data, the next step is to do a differential expression analysis. The R package used to do this is DESeq2 and it starts with a count matrix with one row for each gene and one column for

each sample. Later, we represent the data in a plot where the Y axis is the log of the fold-change and the X axis is the mean of normalized counts of all the samples. The red dots represent the genes that have an adjusted p-value under 0.1. Also, points which fall out of the window are represented as triangles pointing either up or down.

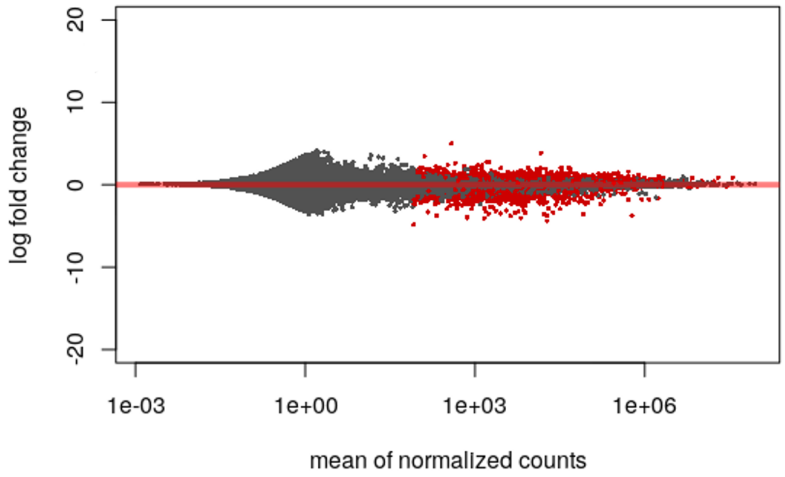

Figure 4 – Differentially expressed genes in early vs late skin cancer.

With this analysis is possible to find genes that are significantly over or underexpressed in late tumours. To do so, three filters will be applied:

1. Keep genes whose adjusted p-value is lower than 0.001. With this filter we keep 391 genes.
2. Keep genes that have a 10 log2fold-change. This is a more strict criteria and only 12 genes are kept.

If we apply these criteria we end up having 3 significantly overexpressed genes and 9 significantly underexpressed genes.

After all this procedure, the data of the post RNA-seq analysis has to be visualized. In order to do this, we have created a volcano plot, which is a type of scatter plot used to identify changes in large data sets. In the y-axis there is represented the negative log of the P-value and in the X-axis there is the log of the fold change between two conditions. This means that the higher the dot is, the more significant it is.

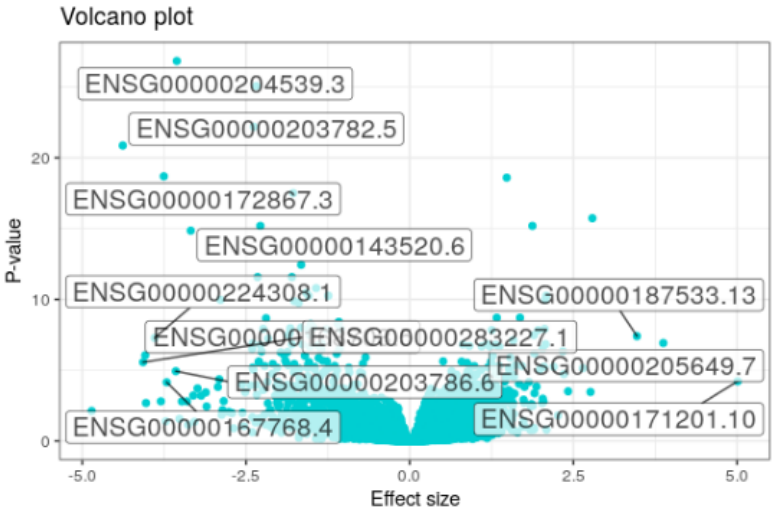

Figure 5 – Volcano plot for genes expressed differently in skin cancer.

As we can see in the plot, there are 11 significant SNPs: ENSG00000204539.3, ENSG00000203782.5, ENSG00000172867.3, ENSG00000143520.6, ENSG00000224308.1, ENSG00000283227.1, ENSG00000203786.6, ENSG00000167768.4, ENSG00000187533.13, ENSG00000205649.7 and ENSG00000171201.10.

The final step is performing a gene set enrichment analysis based on the functional annotation of the differentially expressed genes. This allows to identify if the genes found have association with a certain biological process or molecular function. In this case, we have used the Ensembl database to analyse the genes. We obtain a dataset and, after eliminating the repeated genes, we have ??? genes. Then the association with Gene Ontology terms is tested and the results are classified in the plot below.

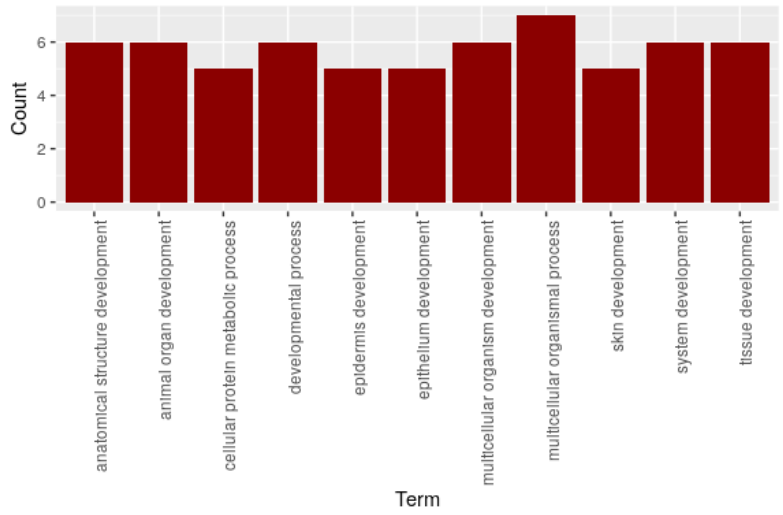

Figure 6 – Gene Ontology analysis for the differentially expressed genes.

5 Discussion

After obtaining the results, the next step will consist on searching all the genes in the Ensembl genome browser.

The first gene is **ENSG00000204539.3** which is located in chromosome 6 (31,082,867-31,088,223) on the reverse strand. This gene is 2552 bp long and it's protein is 529 aminoacids long. The gene codes for the corneodesmosin (*CDSN*) and has 1 allele (splice variant) and 1 transcript which has 70 orthologues and is associated with 4 phenotype.

The second gene is **ENSG00000203782.5** which is located in chromosome 1 (153,232,176-153,234,598) on the forward strand. This gene is 1230 bp long and it's protein is 312 aminoacids long. The gene codes for the loricrin (*LOR*) and has 1 transcript (splice variant) which has 2 exons, 2 orthologues and is associated with 4 phenotypes.

The third gene is **ENSG00000172867.3** which is located in chromosome 12 (53,038,342-53,045,948) on the reverse strand. This gene is 2403 bp long and it's protein is 639 aminoacids long. This gene codes for the keratin 2 (*KRT2*) and has 2 transcripts (splice variants; 1 protein coding and 1 retained intron), 67 orthologues, 69 paralogues, and is associated with 2 phenotypes.

The fourth gene is **ENSG00000143520.6** which is located in chromosome 1 (152,348,735-152,360,006) on the reverse strand. This gene is 9124 bp long and it's protein is 2391 aminoacids long. This gene codes for the filaggrin family member 2 (*FLG2*) and has 1 transcript (splice variant) which has 3 exons, 96 orthologues, 3 paralogues and is associated with 2 phenotypes.

The fifth gene is **ENSG00000224308.1** which is located in chromosome 1 (152,930,040-152,949,210) on the reverse strand. This gene is 1997 bp long and doesn't encode for any protein. This gen is a long intergenic non-protein coding RNA, in particular the 1527 (*LINC01527*). This gene has 1 transcript that has 2 exons.

The sixth gene is **ENSG00000283227.1** which is located in the chromosome 1 (152,947,154-152,949,258) on the forward strand. This gene is 817 bp long and it's protein is 108 aminoacids long. This gene codes for the small proline rich protein 5 (*SPRR5*). This gene has 1 transcript (splice variant) that has 2 exons and 5 orthologues.

The seventh gene is **ENSG00000203786.6** which is located in the chromosome 1 (152,759,561-152,762,052) on the forward strand. This gene is 2492 bp long and it's protein is 579 aminoacids long. This gene codes for the keratinocyte proline rich protein (*KPRP*). It has 1 transcript that has 1 exon (splice variant), 81 orthologues and 14 paralogues.

The eighth gene is **ENSG00000167768.4** which is located in the chromosome 1 (52,674,736-52,680,407) on the reverse strand. This gene is 2451 bp long and it's protein is 644 aminoacids long. This gene codes for keratin 1 (*KRT1*). It has 2 transcripts (splice variants; 1 protein coding and 1 retained intron), 94 orthologues, 69 paralogues and is associated with 14 phenotypes.

The ninth gene is **ENSG00000187533.13** which is located in the chromosome 4 (70,133,616-70,176,799) on the forward strand. This gene is 5451 bp long and it's protein is 219 aminoacids long. This gene codes for proline rich 27 (*PRR27*). This gene has 5 transcripts (splice variants; 2 protein coding, 2 nonsense mediated decay and a processed transcript), 1 gene allele and 43 orthologues.

The tenth gene is **ENSG00000205649.7** which is located in the chromosome 4 (70,028,413-70,036,538) on the forward strand. This gene is 775 bp long and it's protein is 51 aminoacids long. This gene codes for histatin 3 (*HTN3*). This gene has 7 transcripts (splice variants; 4 protein coding and 3 related to a retained intron), 1 gene allele, 19 orthologues and 2 paralogues.

The eleventh gene is **ENSG00000171201.10** which is located in the chromosome 4 (70,370,093-70,390,244) on the forward strand. This gene is 783 bp long and it's protein is 79 aminoacids long. This gene codes for submaxillary gland androgen regulated protein 3B (*SMR3B*). This gene has 4 transcripts (splice variants; 3 protein coding and 1 retained intron) and 5 orthologues.

We have searched these genes in NCBI so as to look for any relationship between them and skin cancer.

- *CDSN* encodes a protein found in corneodesmosomes, which localize to human **epidermis** and other cornified squamous epithelia. This gene is associated in protein homodimerization activity function and involved in processes as **cell adhesion**, **corneocyte desquamation**, **cornification**, **epidermis development**, **keratinocyte differentiation**, negative **regulation of cornification** and **skin morphogenesis**.

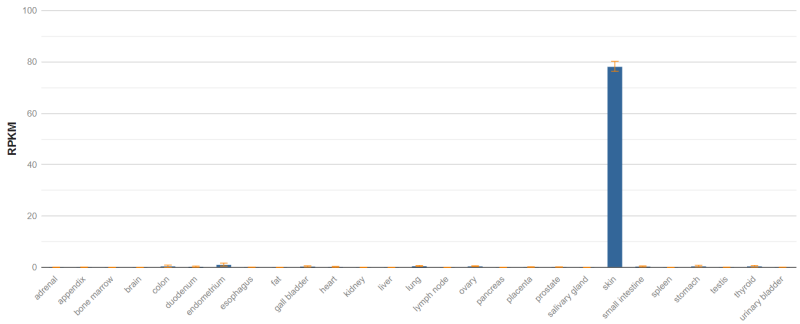

Figure 7 – Differential *CDSN* expression between tissues.

This gene is highly polymorphic (located in major histocompatibility complex class I region, a highly polymorphic genes) in human populations, and variation has been associated with **skin diseases** such as psoriasis, hypotrichosis and peeling skin syndrome. As there is a relation between skin diseases and this gene it may also have a relevance in skin cancer.

- *LOR* encodes loricrin, a major protein component of the cornified cell envelope found in terminally differentiated epidermal cells. This gene is associated in functions as protein binding, structural constituent of cytoskeleton and **structural constituent of epidermis**. The biological processes are **cornification**, cytoskeleton organization, **keratinocyte differentiation** and peptide cross-linking.

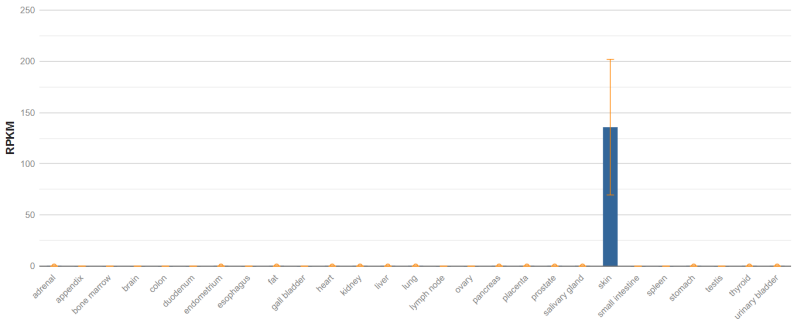

Figure 8 – Differential *LOR* expression between tissues.

Mutations in this gene are associated with Vohwinkel's syndrome and progressive symmetric erythrokeratoderma, both **inherited skin diseases**. As there is a relation between skin diseases and this gene it may also have a relevance in skin cancer.

- *KRT2* encodes for type 2 cytokeratin, this type of cytokeratins consist of basic or neutral proteins which are arranged in pairs of heterotypic keratin chains coexpressed during differentiation of simple and stratified epithelial tissues. The gene is associated in cytoskeletal protein binding, protein binding, structural constituent of cytoskeleton and structural constituent of epidermis functions. The biological processes in which are involved the gene are the followings: cornification, **epidermis development**, intermediate filament organization, **keratinization**, **keratinocyte activation**, **keratinocyte development**, **keratinocyte migration**, **keratinocyte proliferation**, peptide cross-linking and positive **regulation of epidermis development**.

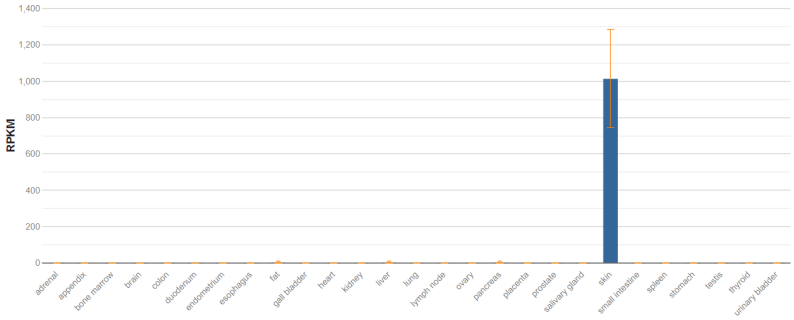

Figure 9 – Differential *KRT2* expression between tissues.

This type II cytokeratin is expressed largely in the upper spinous layer of epidermal keratinocytes and mutations in this gene have been associated with bullous congenital ichthyosiform erythroderma. As there is a relation between **skin diseases** and this gene it may also have a relevance in skin cancer.

- *FLG2* encodes for filaggrin-like protein which is upregulated by calcium, proteolyzed by calpain 1, and is involved in **epithelial homeostasis**. Its functions are calcium ion binding, structural molecule activity and transition metal ion binding. The biological processes are **cell adhesion**, **epidermis morphogenesis**, **establishment of skin barrier** and neutrophil degranulation.

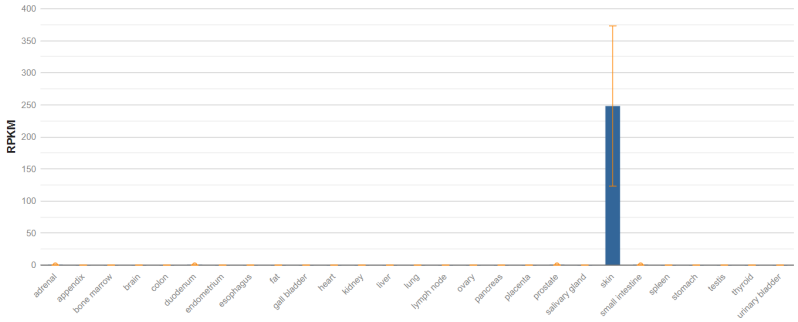

Figure 10 – Differential *FLG2* expression between tissues.

The encoded protein is required for proper cornification in skin, with defects in this gene being associated with **skin diseases**. This protein also has a function in skin barrier protection. In fact, in addition to providing a physical barrier, C-terminal fragments of this protein display antimicrobial activity against P. aeruginosa. As there is a relation between skin diseases and this gene it may also have a relevance in skin cancer.

- *LINC01527* as this gene codifies for a long intergenic non-protein coding RNA, it is the ideal one we would think of regarding skin cancer. However, as it is expressed in skin we cannot just discard it.

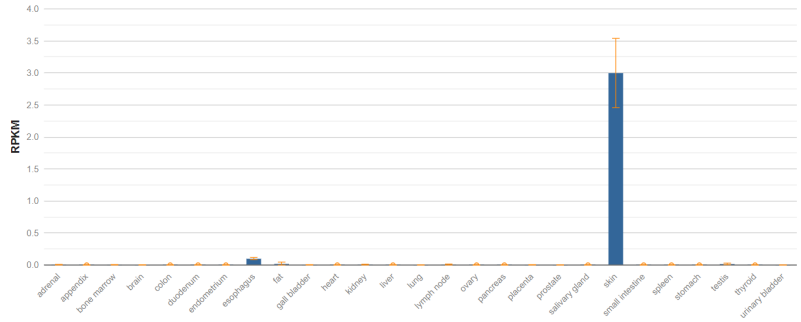

Figure 11 – Differential *LINC01527* expression between tissues.

- *SPRR5* codes for small proline rich protein 5 (there isn't a graph about differential expression between tissues). We found a relationship between this gene and a long non-coding RNA (LINC00941) which is a crucial **regulator of human epidermal homeostasis**. It is enriched in progenitor keratinocytes and acts as a repressor of keratinocyte differentiation.

  Furthermore, LINC00941 represses SPRR5, a previously uncharacterized molecule, which functions as an essential positive regulator of keratinocyte differentiation. Interestingly, half of the genes repressed in SPRR5-deficient epidermal tissue are induced in LINC00941-depleted organotypic epidermis, suggesting a common mode of action for both molecules. As there is a relation between keratinocytes differentiation and this gene it may also have a relevance in skin cancer.

- *KPRP* encodes a proline-rich skin protein (with protein binding function) possibly involved in keratinocyte differentiation. As there is a relation between **keratinocytes differentiation** and this gene it may also have a relevance in skin cancer.

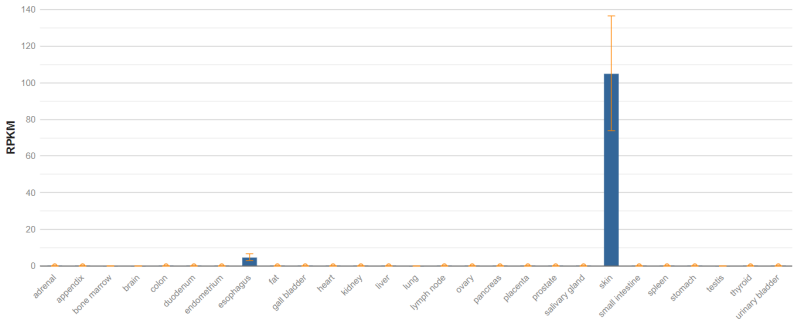

Figure 12 – Differential *KPRP* expression between tissues.

- *KRT1* encodes for a member of the keratin gene family. The gene functions are carbohydrate binding, protein binding, protein heterodimerization activity, signaling receptor activity and structural constituent of epidermis. The biological processes in which are involved are complement activation, **cornification**, **establishment of skin barrier**, fibrinolysis, **keratinization**, negative regulation of **inflammatory response**, neutrophil degranulation, peptide cross-linking, protein heterotetramerization, regulation of **angiogenesis**, response to oxidative stress and retina homeostasis.

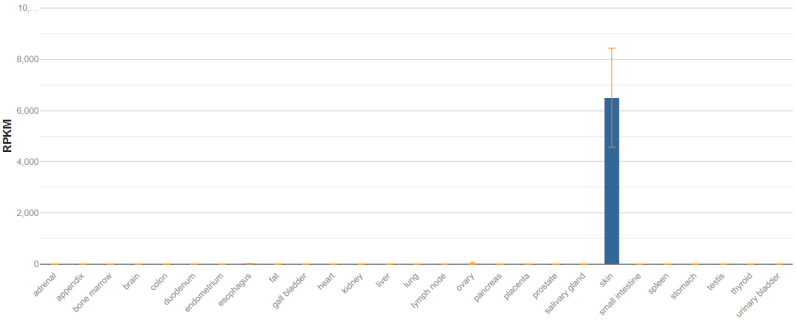

Figure 13 – Differential *KRT1* expression between tissues.

This cytokeratin is specifically expressed in the spinous and granular layers of the epidermis with family member KRT10 and mutations in these genes have been associated with bullous congenital ichthyosiform erythroderma. As there is a relation between skin diseases and this gene it may also have a relevance in skin cancer.

- *PRR27* encodes for proline rich 27. As this gene has restricted expression toward salivary gland we believe that it is not involved in skin cancer.

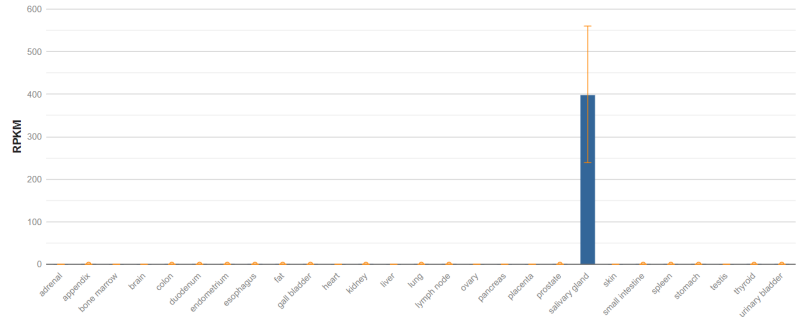

Figure 14 – Differential *PRR27* expression between tissues.

- *HTN3* encodes for a member of the histatin family of small, histidine-rich, cationic proteins (protein and metal ion binding). Involved in processes as antimicrobial humoral response, biomineral tissue development, defense response to bacterium, defense response to fungus and killing of cells of other organism.

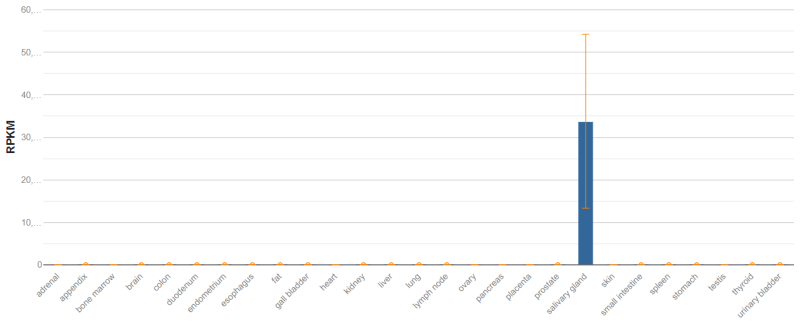

Figure 15 – Differential *HTN3* expression between tissues.

They function as antimicrobial peptides and are important components of the innate immune system. Histatins are found in saliva and exhibit antibacterial, antifungal activities and function in wound healing. As this gene has restricted expression toward salivary gland (and without functions correlated with DNA repair, cell division, apoptosis...) we believe that it is not involved in skin cancer.

- *SMR3B* encodes for submaxillary gland androgen regulated protein 3B. This gene has endopeptidase inhibitor activity, molecular function and protein binding functions and biological processes of negative regulation of endopeptidase activity and regulation of sensory perception of pain.

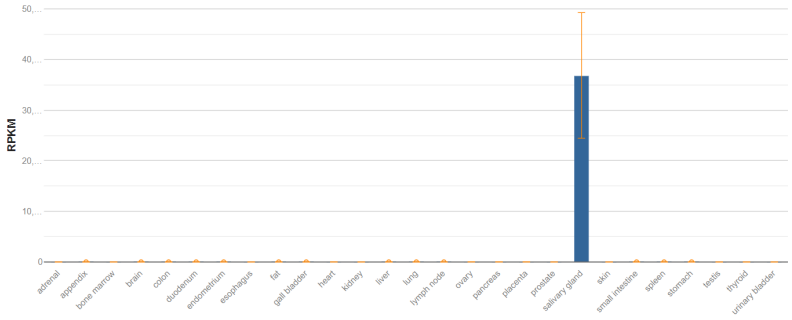

Figure 16 – Differential *SMR3B* expression between tissues.

As this gene has restricted expression toward salivary gland (and without functions correlated with angiogenesis, cell differentiation, cell proliferation...) we believe that it is not involved in skin cancer.

With the available information we can point out that the first 8 genes might be involved in skin cancer because of their relationship with skin diseases but we cannot know it for sure yet. More studies are needed to provide real evidence, for example, we could compare the expression of these gens in skin cancer patients and controls.