RNA expression analysis for skin cancer

Katharina Alexa Lang , Nico Enghardt

January 7, 2025

Abstract

The RNA expression levels in 420 patient samples are compared between early- and late-stage melanoma. Differentially expressed genes are detected and their correlations with cancerous behavior discussed. The findings reveal that advanced melanomas exhibit dedifferentiation and a disturbed protein stress response.

Contents

1 Methods

1.1 Packages and tools used

All Data analysis was performed in R. The following r packages were used:

- SummarizedExperiment
- edgeR
- DESeq2
- tweeDEseq
- tweeDEseqCountData
- GOstats
- annotate
- org.Hs.eg.db
- biomaRt
- ggplot2
- ggrepel
- dplyr
- tibble
- clusterProfiler
- pathview
- pheatmap

1.2 Data

Data underlying this analysis was published by The Cancer Genome Atlas (TCGA) project and made available on recount2 (link) in an analysis-ready format.
For this work, the RangedSummarizedExperiment (RSE) object for skin cancer (link) was downloaded directly from recount2. The RSE contains mRNA-Seq data of cutaneous melanoma samples from 473 patients mapped to 58037 distinctive Ensembl IDs. RNA-Seq data was obtained using synthesis based sequencing (Illumina HiSeq) and subsequently saved as raw sequencing data in the counts matrix of the RSE object.
Additional to the counts matrix, metadata is provided for both the samples and Ensemble IDs.
For each Ensembl ID, the length in base pairs and the corresponding gene symbol are provided. Gene symbol is NA for 32511 of the Ensembl IDs, probably due to those IDs corresponding to non-protein-coding genes without an assigned gene symbol. Some IDs share the same symbol, with 385 gene symbols being mapped to more than one ID. This could be due to splicing variants. Ensemble IDs without a corresponding symbol were excluded from the analysis.
Metadata for the samples covers a wide range of data points on patients (e.g., age, gender, time of diagnosis, treatments and response to treatment) as well as on samples (e.g., tissue of origin, sampling method, mRNA-Seq Method, total reads). For this analysis, the cancer stage of each patient was obtained from the metadata. To allow differential gene expression analysis between early vs late stage skin cancer, patients were grouped by disease stage as follows:

Table 1 – Mapping of cancer stages

early	stage 0, stage i, stage ia, stage ib, stage ii, stage iia, stage iib, stage iic
late	stage iii, stage iiia, stage iiib, stage iiic, stage iv

With this mapping, all samples without a given cancer stage where excluded. 420 samples remained for further analysis, 225 for early and 195 for late stages.

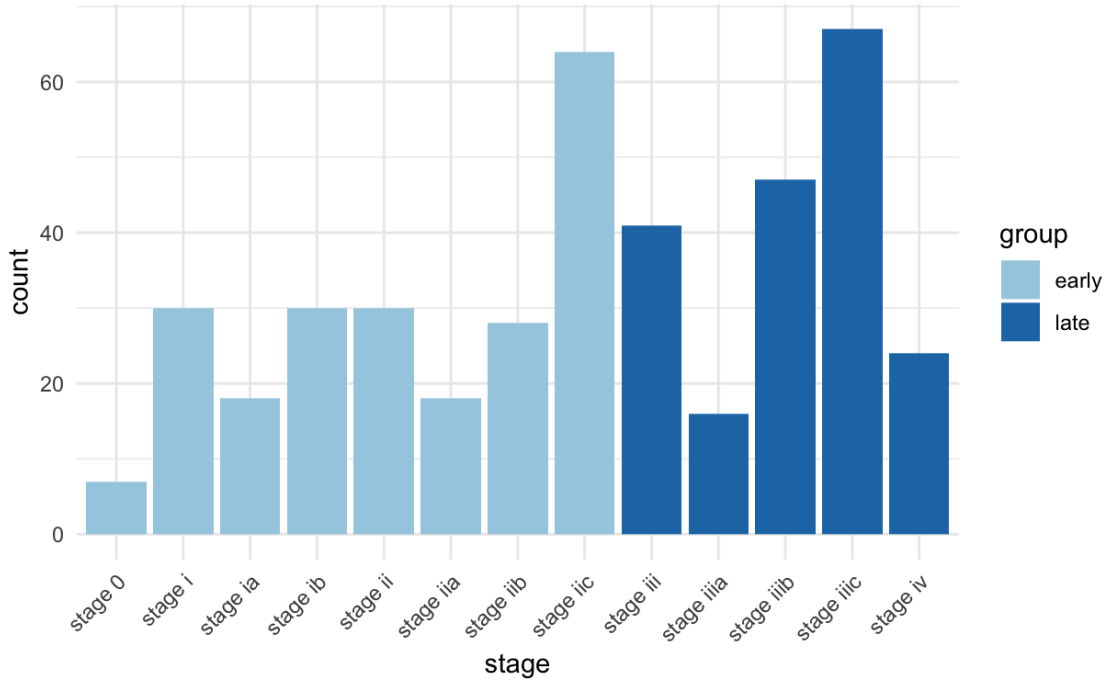


Figure 1 – Count and mapping of samples by stage

2 Statistical analysis

2.1 DEG analysis

Differentially expressed genes are identified by comparing the count of mRNA reads per gene between two datasets. DEG analysis was performed using DESeq2 r package [15].
`DESeq()` takes raw RNA count matrices such as the one in the RSE skin as input. Differences in expression between the datasets are evaluated in comparison to variation within each dataset. Only if the inter-group variation is significantly higher than the intra-group variation, it can be concluded that the change in expression correlates with the distinctive properties of the two groups. `DESeq()` performs a sequence of calculations and estimations to determine expression changes of each gene i over the samples j:
First `DESeq()` estimates s (normalization factor correcting effects occurring due to differently sized samples) and α (expected dispersion for each gene). α defines how widely the expression of each specific gene usually varies from the mean expression.
Next, the counts K are modeled using a negative binomial distribution and the estimated α. The resulting model contains μ as fitted mean per gene and sample. μ is divided by s (sample size factor) and thereby converted into q which reflects the share of counts for gene i in the specific sample j instead of the total counts represented by μ.

$$K \sim NB(\mu_{ij}, \alpha_i) \tag{1}$$

$$\mu_{ij} = s_j q_{ij}$$

The average q for each gene is calculated for the two compared groups. $Log2$ of the change between them (β) is given as a result per gene and defined sample group (X).

$$log2(q_{ij}) = \beta_i X_j$$

To determine the statistical significance of expression changes, the corresponding p-values are calculated based on the Wald test. The Walden test first determines z by dividing (β) by the standard deviation for (β) which it obtains from the variance of observed counts (K_{ij}) and the model fit. A higher $|z|$ suggests a significant change.
The p-value is the probability of z occurring under the condition that z is normally distributed. A low probability (<0.05) suggest that the change in expression is statistically significant instead of a result of normal variation. `results()` is applied the output of `DESeq()`, extracting among other data the Log2FoldChange, p-value and adjusted p-value for each gene. Adjusted p-values are needed due to the high numbers of tested genes. Known as the "multiple testing issue", unadjusted p-values in large number of tests lead to a high probability of false positives [12].
Adjusted p-values were calculated by the false discovery rate method "fdr" established by Benjamini and Hochberg. First p-values are ranked in ascending order and subsequentially each p-value is multiplied by the total number of p-values (m) and divided by its rank (k)

$$p.adjusted = p * \frac{m}{k}$$

2.2 Enrichment analysis

`enrichGO()` and `enrichKEGG()` were applied for enrichment analysis of gene ontology terms and KEGG terms (Kyoto Encyclopedia of Genes and Genomes) respectively. Both of these databases annotate genes with specific terms: GO terms are sorted into three categories that contain information on molecular function (MF), biological process (BP), and cellular compartment (CC). KEGG annotates genes with pathways in which they are involved.
Input for these functions is a vector of genes. For this analysis DEGs with an adjusted p-value < 0.001 and a Log2FoldChange > 1.5 or < -1.5 were included. Furthermore, the `qvaluecutoff` was set at 0.05, therefore only including terms whose enrichment was statistically significant with an adjusted p-value <0.05. The key output of the enrichment functions is the `GeneRatio` and the `qvalue` (= adjusted p-value) per term.
`GeneRatio` of a term indicates the share of genes annotated with this term in the given gene set.

$$GeneRatio = \frac{n}{m}$$

The p-value is the probability of having at least m genes annotated to this term in the input gene list whilst the Null-Hypothesis of no enrichment is true. Hypergeometric testing is applied for this [1]:

$$p = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

N = Total number of genes in database
M = Number of genes in database annotated with the term
n = Number of genes in input
m = Number of genes in input annotated with the term

An adjusted p-value is needed due to the "multiple testing issue" as detailed for the DEG analysis. P-values were adjusted according to the fdr method and are included in the output of `enrich()` functions as `qvalue`.

3 Results

3.1 DEG Analysis

DEG Analysis was applied on the mRNA-seq counts for each gene and each sample in the dataset with reference to the early or late development stage of the sampled cancer. 150 genes were found to be differently expressed (p.adjusted < 0.001 and |Log2FoldChange| > 1.5) between the groups, with 41 overexpressed and 109 underexpressed in the late-stage samples. All genes with absolute log2FoldChange > 3 and p.adjusted < 10^{-4} are outlined in table 2.

Protein Code	log2foldChange	p.adjusted
LORICRIN	- 4.4	$1.4 \cdot 10^{-21}$
SPRR5	- 4.1	$2.7 \cdot 10^{-6}$
WFDC12	- 4.0	$1.8 \cdot 10^{-7}$
LINC01527	-3.9	$5.3 \cdot 10^{-8}$
KRT2	- 3.8	$2.1 \cdot 10^{-19}$
KRT1	- 3.7	$7.1 \cdot 10^{-5}$
KPRP	- 3.6	$1.2 \cdot 10^{-5}$
CDSN	- 3.6	$1.5 \cdot 10^{-27}$
FLG2	- 3.3	$1.4 \cdot 10^{-15}$
SMR3B	5.0	$6,0 \cdot 10^{-5}$
HTN3	3.9	$1.2 \cdot 10^{-7}$
PRR27	3.5	$4.0 \cdot 10^{-8}$

Table 2 – Selection of DGE results

3.2 Enrichment Analysis

Enrichment Analysis was applied to the set of 150 genes with significant differential expression (p.adjusted < 0.001 and |Log2FoldChange| > 1.5). The dominant molecular functions (see 3 a) attributed to the genes with expression changes are **antigen binding** (overexpressed) and **peptidase inhibitor activity**. Genes for the latter category show a variety of expression levels; the overexpressed SMR3B (log2FC.=5.0) and the underexpressed WFDC12 (log2FC.=-4.0) are both in this category. The dominant biological function (see 3 b) attributed to differently expressed genes are **skin development** and **keratinocyte differentiation** specifically. Relevant Cellular components (see 3 c) or advanced melanoma are the **immunoglobulin complex** (place for antigen binding), the **extracellular matrix** and **cornified envelope**.

The Kegg pathways (see 3 c) with highest degree of differential expression are **Salivary secretion** interaction (overexpressed), the Staphylococcus aureus infection (underexpressed) and **Vascular smooth muscle contraction** (underexpressed).

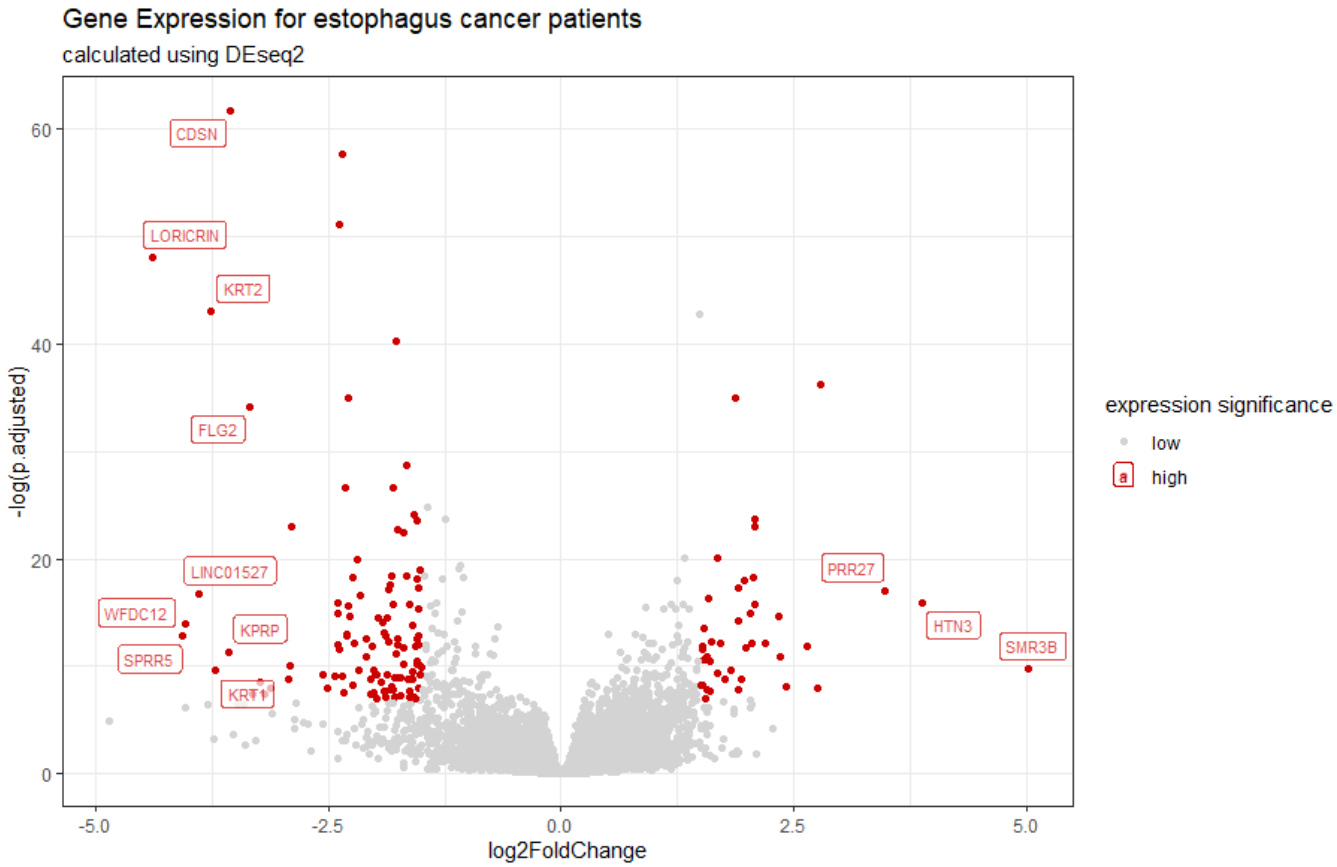


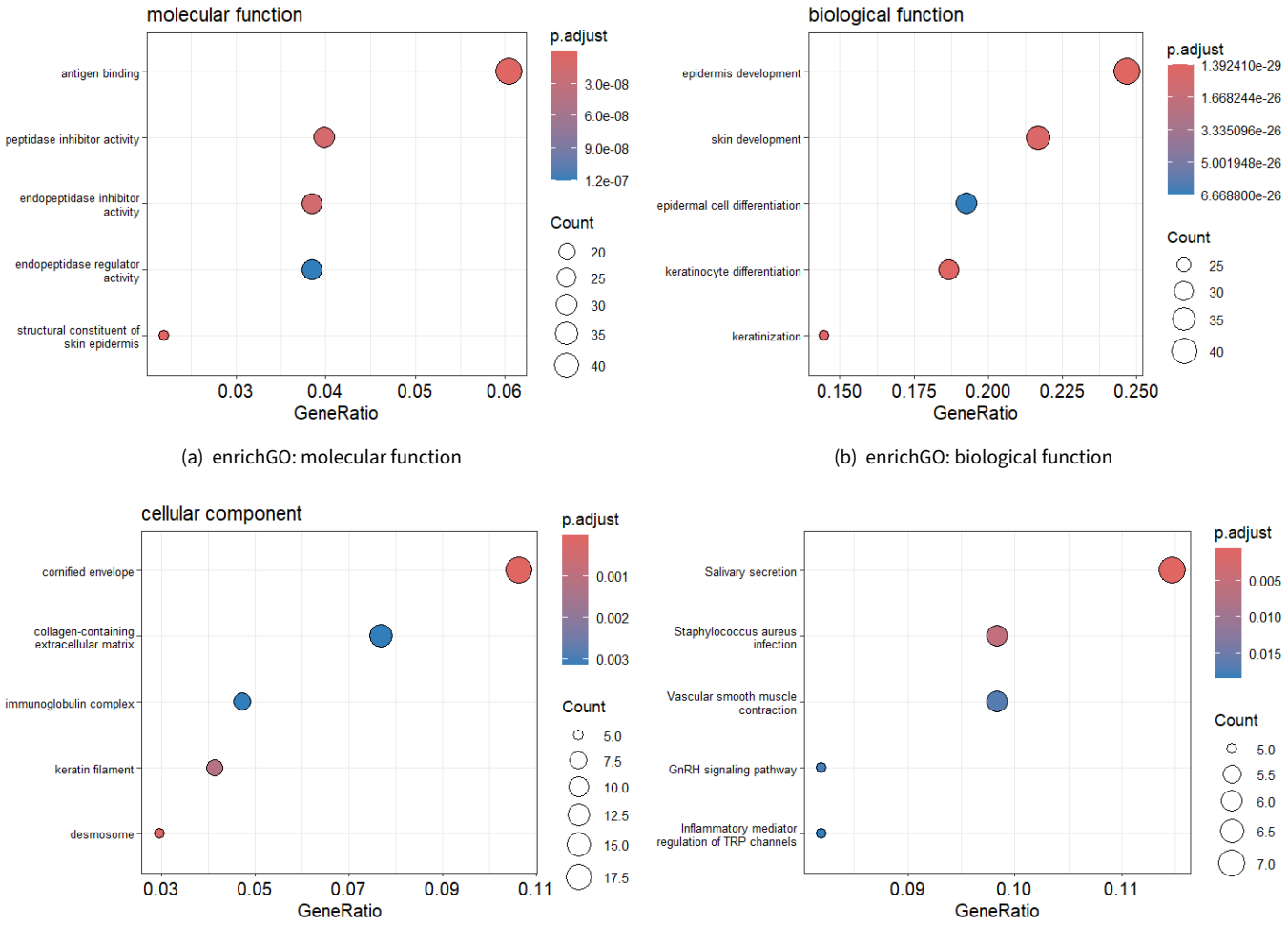
Figure 2 – Results of the DEG Analysis

3.3 Discussion

First, the genes with high degree of differential expression shall be discussed.

LORICRIN is a major component of the cornified envelope of the epidermis [8]. Keratin-1 (KRT1) and Keratin-2 (KRT2) are both structural proteins that constitute intracellular filaments. [4; 5; 8]. Corneodesmosin (CDSN) formes a strong intercellular link to handle mechanical stress in the skin [9]. Filaggrin's (FLG, log2FC=-2.4) purpose is to aggregate and compact Keratin filaments to support the cornification of the outer skin layer [7]. Filaggrin-2 (FLG2) is a intercellular link in this cornified skin layer [2]. All these genes encoding proteins that contribute to a strong skin barrier, are underexpressed in late-stage melanoma cells, which are prone to detach and metastate. This requires mobility, which would be hindered by proteins that link to other cells (CDSN and FLG2) or by a stronger intracellular skeleton (Keratin). SPRR5 which positively regulates the differentiation of ceratinozytes [16], supporting all above mentionend processes, is also underexpressed. The underexpressed WFDC12 also downregulates formation of Keratin filaments [13].

Furthermore these proteins are markers for the differentiation of skin cells. Their underexpression could indicate that a dedifferentiaion process takes place. Dedifferentiation of cancer cells into cancer stem cells is a common phenomenon associated with increased drug resistance and metastatic potential [14]. The importance of this process was recently highlighted, as "unlocking phenotypic plasticity", which is



(a) enrichGO: molecular function

(b) enrichGO: biological function

(c) enrichGO: cellular component

(d) enrichKegg: pathways

Figure 3 – Enrichment Results

enabled via dedifferentiation, was added to the hallmarks of cancer [10].

SMR3B inhibits the destruction of proteins, concretely via inhibition of peptidases and endopeptidases [6]. Therefore, the overexpressed SMR3B disturbs the protein stress response and can thereby contribute to a malignant metabolism.

Histatins (HTN3) and PRR27 are components of human saliva. Their overexpression in late-stage melanoma has no known relation to cancerous behaviour. Likewise, no effect can be attributed to the underexpression of RNA for the non-coding gene LINC01527.

Only a limited number of highly differentially expressed genes could be discussed specifically, the 150 other significantly differentially expressed genes will be discussed in groups that have been identified in the enrichment analysis.

Genes associated to **Antigen binding** showed higher expression in late-stage melanoma. Many of these genes are members of the families IGHV and IGKV, which constitute immunglobuline complexes of leukocytes [3]. IGKV genes have been found present in breast cancer cells as well.

The genes of the DEFB family cytotoxins necessary for fighting a **Staphylococcus aureus infection**. Their underexpression in late-stage melanoma disables the celular immune reaction to S.aureus, which are known to create a cancer-friendly environment [11].

4 Summary

The comparison of gene expression levels between early- and late-stage melanoma yielded interpretable results. Most dominant was the decline in Keratin and other epidermis-constituting proteins which point to the dedifferentiation of the cancerous epidermis cells. Furthermore, the protein stress response was surpressed in late-stage melanoma. Some genes that were intensely underexpressed or overexpressed could could not be attributed any influence on the cancerous behaviour.

5 References

[1] Biomedical knowledge mining using gosemsim and clusterprofiler | yulab-smu.top/biomedical-knowledge-mining-book/enrichment-overview.htmlora-algorithm.

[2] FLG2 Gene - GeneCards | genecards.org/cgi-bin/carddisp.pl?gene=FLG2.

[3] IGKV Gene - GeneCards | genecards.org/cgi-bin/carddisp.pl?gene=IGKV.

[4] KRT1 Gene - GeneCards | genecards.org/cgi-bin/carddisp.pl?gene=KRT1.

[5] KRT2 Gene - GeneCards | genecards.org/cgi-bin/carddisp.pl?gene=KRT2.

[6] SMR3B Gene - GeneCards | genecards.org/cgi-bin/carddisp.pl?gene=FLG2.

[7] M. Armengot-Carbo, Hernández-Martín, and A. Torrelo. The role of filaggrin in the skin barrier and disease development. *Actas Dermo-Sifiliográficas (English Edition)*, 106(2):86–95, 2015.

[8] E. Candi, R. Schmidt, and G. Melino. The cornified envelope: a model of cell death in the skin. *Nature Reviews Molecular Cell Biology 2005 6:4*, 6:328–340, 2005.

[9] D. Garrod and M. Chidgey. Desmosome structure, composition and function. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1778(3):572–587, 2008. Apical Junctional Complexes Part I.

[10] D. Hanahan. Hallmarks of cancer: New dimensions. *Cancer Discovery*, 12(1):31–46, 01 2022.

[11] M. E. Hosen, S. J. Supti, S. Akash, M. E. Rahman, M. O. Faruqe, M. Manirujjaman, U. K. Acharjee, A.-R. Z. Gaafar, L. Ouahmane, B. Sitotaw, M. Bourhia, R. Zaman, K. Ahmad, M. Murahari, K. K. M. Darwish, J. S. Supti, and G. A-rz. Mechanistic insight of staphylococcus aureus associated skin cancer in humans by santalum album derived phytochemicals: an extensive computational and experimental approaches. *Front. Chem*, 11:1273408, 2023.

[12] M. Jafari and N. Ansari-Pour. Why, when and how to adjust your p values? *Cell Journal (Yakhteh)*, 20(4):604–607, 2018.

[13] P. Kalinina, V. Vorstandlechner, M. Buchberger, L. Eckhart, B. Lengauer, B. Golabi, M. Laggner, M. Hiess, B. Sterniczky, D. Födinger, E. Petrova, A. Elbe-Bürger, L. Beer, A. Hovnanian, E. Tschachler, and M. Mildner. The whey acidic protein wfdc12 is specifically expressed in terminally differentiated keratinocytes and regulates epidermal serine protease activity. *The Journal of investigative dermatology*, 141:1198–1206.e13, 5 2021.

[14] J. Li and B. Z. Stanger. How tumor cell dedifferentiation drives immune evasion and resistance to immunotherapy. *Cancer Research*, 80(19):4037–4041, 10 2020.

[15] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15:1–21, 2014.

[16] C. Ziegler, J. Graf, S. Faderl, J. Schedlbauer, N. Strieder, B. Förstl, R. Spang, A. Bruckmann, R. Merkl, S. Hombach, and M. Kretz. The long non-coding rna linc00941 and sprr5 are novel regulators of human epidermal homeostasis. *EMBO Reports*, 20:e46612, 2 2019.