Comparative analysis of patients with metastatic and non-metastatic cases of uveal melanoma: an attempted replication

Lukas Kuderna[1,*], Eric March[1,*], Maitena Tellaetxe[1,*], Jesse Willis[1,*]
1 Universitat Pompeu Fabra, MSc Bioinformatics, Barcelona, Catalonia
∗ **E-mail: Corresponding lukas.kuderna01@estudiant.upf.edu, eric.march01@estudiant.upf.edu, maitena.tellaetxe01@estudiant.upf.edu, jesse.willis01@estudiant.upf.edu**

Abstract

Metastatic tumors are common in uveal melanoma patients, and it has been suggested by Laurent et al. [**?**] that the gene Protein tyrosine phosphatase type IV A member 3 (PTP4A3/PRL3) plays a significant role in metastasis. They performed microarray analyses of gene expression on 63 tumor samples obtained after enucleation of the eye. This is an attempt to reproduce part of the results of their study using a sample of their original data. Moderated t-statistics were implemented to look for differentially expressed genes between 12 uveal melanoma samples from patients who developed metastases within three years of enucleation and 8 samples that did not develop metastases or did so more than three years after enucleation. With this portion of the original 63, no differentially expressed genes were found, possibly due to a small sample size, but gene set enrichment analysis using biological pathways in the KEGG database yields a list of 48 pathways that appear to be enriched in the first group, many of which, such as oxydative phosphorilation or cytokine-cytokine receptor interaction, are important in the process of metastasis.

Author Summary

Patients with uveal melanoma, a cancer of the pigment cells that give color to the eye, commonly develop metastatic tumors, and it has been suggested by Laurent et al.1 that the gene Protein tyrosine phosphatase type IV A member 3 (PTP4A3/PRL3) plays a significant role in metastasis. Here we attempt to reproduce the expression analysis in their study using a sample of their original data. We implement a number of statistical tools to look for differentially expressed genes between 8 uveal melanoma samples without metastases and 12 from patients who developed metastases within three years of obtaining the samples. With our portion of the original 63, we find no differentially expressed genes, possibly due to our small sample, but an analysis of gene sets, as opposed to individual genes, yields a list of 48 biological pathways that appear to be enriched in the metastasis developing group, many of which are important in the mentioned process.

Introduction

Uveal melanoma is the most common malignant tumour localized in the eye. In adult patients, around a 50% develop metastases within a median of 36 months, with a median survival of 6 months after metastasis [**?**]. It can have severe repercussions on visual capability and, sometimes, complete removal of the eye is needed to achieve healing. The most frequent symptoms are loss of visual acuity and defects on the field of vision, and, although it can be detected by those symptoms, the most common method of diagnosis is exploring the depth of the eye to ascertain if there is an intraocular mass. Loss of chromosome 3 and gains of 8q and 6p have been described as the most frequent imbalances that lead to uveal melanoma [**?**]. A set of features have been found to correlate with patient survival. Among these features there is patient age (>60), location of the tumor, tumor cell histology, tumor diameter,

mitotic activity and chromosome 3 monosomy. Some gene expression profiling studies [**?**] have identified two molecular classes strongly associated with metastatic risk, however, not all metastatic tumors can be classified using these markers. Here we re-analyse the difference in gene expression levels of 12 patients with uveal melanoma that develop metastasis within 3 years of diagnosis (meta1 group) and 8 patients with late or no development of metastasis (meta0 group) to try to reproduce a recently established relationship between protein tyrosine phosphatase type IV A member 3 (PTP4A3) and early metastasis onset [**?**], [**?**].

Materials and Methods

Tumor samples and clinical data

The original study, accessible through GEO Series accession number GSE22138 (`http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22138`) analyzed 63 uveal melanoma tumors, of which 35 developed metastasis and 28 did not in the period they were studied. In order to make the work manageable, only 20 of those samples were selected, so that 12 were cases and 8 controls, maintaining a similar case-control ratio to the original dataset. As batch effect i.e. non-biological variability in microarrays processed at different days, could be observed in the original data, sampling could not randomly be done and therefore a similar number of samples was selected from each of the batches in order to keep a balanced distribution (See supplementary materials for further information). Clinical data of the patients included attributes such as age; gender; eye; tumor location, thickness and diameter and so on, but data was not available for all the samples in many of these features. From the complete ones, age and gender were just selected since it was believed that they could account for variability regarding metastasis.

Array preprocessing

Microarray CEL files corresponding to scanner images must be converted into measures of expression. Robust Multi-array average (RMA) algorithm was employed for this aim, which consists in a background correction, log2 transformation and quantile normalization of raw intensity values, followed by a linear model fit to this normalized data in order to obtain an expression measure for every probeset in the array. Batch effect identification and removal was subsequently performed in order to remove the variation arisen from the environmental differences of the days when the assays were done. Finally, from those probesets targeting common genes, only the one with the largest interquartile range (IQR) value was retained (See supplementary materials for further information).

Differential expression analysis

Data for the differential expression (DE) analysis between the meta1 and meta0 groups consisted of 20 samples of which expression data of 20109 genes and clinical data including age and gender was known. For the purpose of finding those genes with changing expression values, several approaches were tried: **(a)** model based on just gene expression and *1.* no other adjustment *2.* adjusting for batch *3.* adjusting for batch and age *4.* adjusting for batch, age and gender, **(b)** surrogate variable analysis (SVA) with gene expression data and *1.* adjusting for batch and age *2.* adjusting for batch, age and gender, **(c)** surrogate variable analysis with gene expression data, adjusting for batch and age and removing *1.* the genes with the lowest 40% of variability *2.* the lowest 50% of variability and *3.* the lowest 70% of variability and **(d)** surrogate variable analysis with gene expression data, adjusting for batch and age, removing probesets with expression values below 3.5 (log2 scale) in all samples and discarding *1.* the genes with the lowest 40% of variability *2.* the lowest 50% of variability and *3.* the lowest 70% of variability. A moderated t-test was performed in each of the mentioned cases and adjustment for multiple testing was applied so

that the expected false discovery rate (FDR) was kept below 0.05. The number of samples and features used in each analysis are summarized in Table 1.

Enrichment analysis

In order to give a biological sense to the results obtained in the previous step, we performed Gene Set Enrichment Analysis (GSEA) in KEGG pathways by means of z-tests. Notice that this method does not require a list of DE genes as an input and assesses DE directly at gene set level, so that even if no DE are found due to the expression changes being small, pathway enrichment can still be observed. Adjustment for multiple testing was again applied at 5% FDR.

Results

No single overexpressed factor was found to be significantly associated with metastatic uveal melanoma tumors

Initial DE analysis between patients with early metastatic onset and those with late or no metastasis yielded no significantly differentially expressed probe sets ($p < 0.05$, FDR=5%) in all the models. Out of all considered ones, the model correcting for surrogate variables, age and batch yielded the most powerful results, albeit below the significance threshold. The results also stayed the same after filtering out probes with consistently low variability as well as consistently low expression values (See Materials and Methods section). The distribution of raw p-values from the t-tests clearly shows a peak around zero suggesting a non random influence in many of the analyses, however, correcting for multiple hypothesis testing renders these values insignificant (See Figure 1).

Simple gene set enrichment analysis hints at possible pathway candidates

After being unable to find DE genes GSEA was performed. In order to avoid too many overlaps, only KEGG pathways were chosen from the Molecular Signatures Database to see if any of them were enriched in our data. Roughly 78% of the pathway annotations overlapped by less than 50% genes and none of them overlapped completely (See supplementary materials). Performing the analysis by calculating z-scores, a total of 48 enriched pathways out of 182 possible ones were obtained(See Table 2). Although these pathways have some overlapping, the annotations give an idea of the biological differences between the two groups.

Plotting the quantiles of the distribution of the z-scores against the probability distribution of the null hypothesis also supports the idea of having many enriched pathways in the meta1 group (See Figure 2).

Discussion

With our set of samples, we were unable to reproduce the results from the original paper. We could not find any of the 983 probe sets the authors reported to be differentially expressed in our analysis. One possible reason for this could be our small sample size, using just about a third of the available data, combined with possibly small effects, idea which is supported by the histograms of the raw p-values in many of the models, which look really promising. Indeed, a quick general analysis of the whole dataset without any filtering showed roughly 500 probesets to be differentially expressed at 5% FDR (data not shown).

Significant improvement in the p-value histograms was obtained when surrogate variable adjustment was included, as expected since with this methodology at least a fraction of the non-biological variability

is taken into account. Adjusting for age also improves the models but it seems that adjusting for gender is not helping. Filtering by expression and/or variability was expected to improve the analysis since we are ruling out genes that have poor quality or are very unlikely to be called DE but, even though the smallest adjusted p-values are lower than those obtained in the analyses in which no filtering is carried out, the histograms look worse and we still do not find any DE gene.

Since our analysis based on this subset of the data did not produce any results, we moved on to perform a gene set enrichment analysis (GSEA) in order to assess differential expression at the gene set level, which could account for small changes previously undetected. Of the 48 pathways found to be enriched, some are clearly important for metastasis, such as oxidative phosphorylation and cytokine-cytokine receptor interaction, the two most significantly enriched pathways. However, there are also many with less obvious importance, including the next most significantly enriched pathways: Parkinsons disease, Alzheimers disease and Huntingtons disease. One reason for this could be that the factors accounting for the enrichment of those pathways are rather promiscuous regulators involved in a variety of cellular processes. Furthermore, since there is overlap it is unsurprising to find significant results for different pathways. These results are, however, to be taken with precaution as simple GSEA methods are highly sensitive. When performing the same type of analysis using a Chi Squared test, we found no enriched pathways at all.

Acknowledgments

References

1. Laurent C, Valet F, Planque N, Silveri L, Maacha S, et al. (2011) High PTP4A3 Phosphatase Expression Correlates with Metastatic Risk in Uveal Melanoma Patients. Cancer Res 71: 666-674.

2. Gargoudas ES, Egan KM, Seddon JM, Glynn RJ, Walsh SM, et al. (1991) Survival of patients with metastases from uveal melanoma. Ophthalmology 98: 3839.

3. Trolet J, Hup e P, Huon I, Lebigot I, Decraene C, et al. (2009) Genomic profiling and identification of high risk uveal melanoma by array-CGH analysis of primary tumors and liver metastases. Invest Ophthalmol Vis Sci 50: 257280.

4. Onken MD, Worley LA, Tuscan MD and Harbour JW. (2010) An accurate, clinically feasible multi-gene expression assay for predicting metastasis in uveal melanoma. J Mol Diag 12: 4618.

Figure Legends

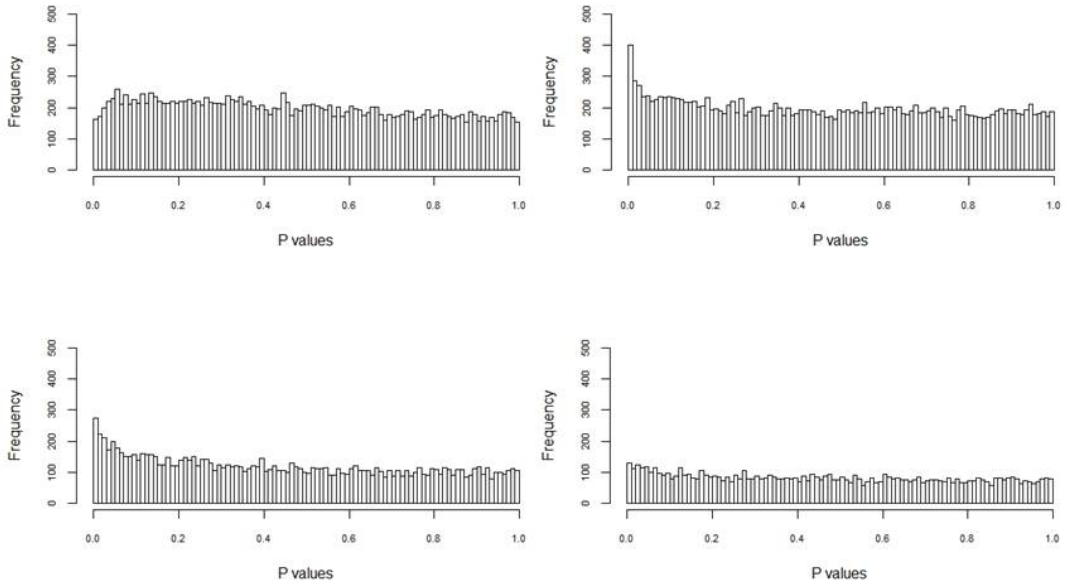

Figure 1. Histograms of raw p-values of a.1, b.1, c.1 and d.1 moderated t-test analyses, respectively. In a.1 (top left), p-values were obtained from a model based on gene expression alone, with no other adjustments. In b.1 (top right), SVA was performed with adjustments for batch and age. In c.1 (bottom left), SVA was performed with adjustments for batch and age after removing those genes with the lowest 40% of variability. In d.1 (bottom right), the process was the same as in c.1, with additional removal of those probesets with expression values below 3.5 (log2 scale) in all samples. Whereas a.1 and d.1 do not suggest promising results, b.1 and c.1 show a peak near 0, implying that there is a probability higher than random to find differentially expressed genes between the two groups.

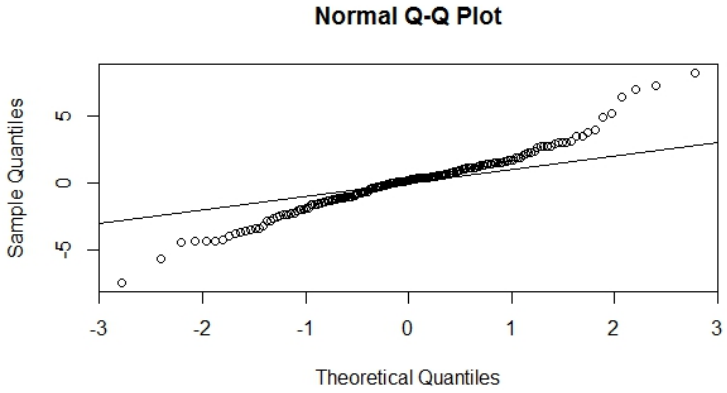

Figure 2. QQplot of the z-statistics in GSEA. The plots shows values of theoretical versus empirical quantile values of z-scores obtained from the gene-set enrichment analysis. The solid line shows the theoretical slope under the null hypothesis of no difference between the meta0 and meta1 group (mean=0, sd=1). The plotted values do not correspond to the null hypothesis, the empirical distribution is more dispersed indicating up and down regulated gene sets.

Tables

Table 1. Total number of samples and microarray probesets and genes reported at each stage of the analysis.

Analysis	Number of samples	Number of probesets	Probeset:gene relationship
a.1	18	19907	1:1
a.2	18	19907	1:1
a.3	18	19907	1:1
a.4	18	19907	1:1
b.1	18	19907	1:1
b.2	18	19907	1:1
c.1	18	11944	1:1
c.2	18	9953	1:1
c.3	18	5972	1:1
d.1	18	8161	1:1
d.2	18	6801	1:1
d.3	18	4081	1:1

Different models that vary on the employed probeset numbers were created. The table summarizes how many samples and probesets were used in each model. Notice that as all probeset-gene relationships are 1:1, the number of probesets and genes is the same.

Table 2. Enriched KEGG pathways in GSEA between meta0 and meta1 groups.

KEGG pathway	5% FDR corrected p-value	KEGG pathway	5% FDR corrected p-value
Oxydative phosporilation	2.04e-14	Cytokine cytokine receptor interaction	2.84e-12
Parkisons disease	7.75e-12	Alzheimers disease	5.57e-11
Huntingtons disease	1.82e-09	Chemokine signaling pathway	1.87e-07
Proteasome	1.99e-06	Oocyte meiosis	1.05e-05
Hematopoietic cell lineage	8.20e-05	ECM receptor interaction	9.12e-05
JAK stat signaling pathway	9.12e-0.5	Pathways in cancer	9.12e-05
T cell receptor signaling pathway	1.22e-04	Cardiac muscle contraction	3.55e-04
B cell receptor signaling pathway	3.55e-04	Nod like receptor signaling pathway	7.90e-04
Glycolysis-Gluconeogenesis	3.02e-04	Small cell lung cancer	9.62e-04
Natural killer cell mediated cytotoxicity	1.66e-03	Citrate cycle TCA cycle	1.90e-03
Arginine and proline metabolism	1.90e-03	Focal adhesion	1.90e-03
Leukocyte transendothelial migration	2.21e-03	Drug metabolism cytochrome P450	2.67e-03
Ribosome	4.64e-03	Terpenoid backbone biosynthesis	6.36e-03
Antigen processing and presentation	7.13e-03	Alanine aspartate and glutamate metabolism	8.03e-03
Long term depression	8.03e-03	Amyotrophic lateral sclerosis als	1.08e-02
Glutathione metabolism	1.37e-02	P53 signaling pathway	1.38e-02
Pyruvate metabolism	1.44e-02	Pentose phosphate pathway	1.51e-02
Protein export	1.56e-02	Fructose and mannose metabolism	1.72e-02
Primary immunodeficiency	1.90e-02	Purine metabolism	1.91e-02
Selenoamino acid metabolism	2.61e-02	Metabolism of xenobiotics by cytochrome P450	2.61e-02
Hedgehog signaling pathway	3.53e-02	Axon guidance	3.64e-02
Taste transduction	3.64e-02	Toll like receptor signaling pathway	3.91e-02
Nicotinate and nicotinamide metabolism	4.12e-02	Complement and coagulation cascades	4.47e-02
Prostate cancer	3.89e-02	Apoptosis	4.05e-02
Nicotinate and nicotinamide metabolism	4.16e-02	Nitrogen metabolism	4.28e-02
Acute myeloid leukemia	4.44e-02	Chronic myeloid leukemia	4.74e-02
Long term potentiation	4.47e-02	Nitrogen metabolism	4.70e-02

Enriched KEGG pathways in GSEA between meta0 and meta1 groups. Only pathways with an associated corrected p-value (5% FDR) smaller than 0.05 are shown. These results suggest that these mechanisms change between the two conditions.