Integrative Bioinformatic Analysis of a Greek Epidemiological Cohort Provides Insight into the Pathogenesis of Primary Cutaneous Melanoma

Georgia Kontogianni, Olga Papadodima, Ilias Maglogiannis, Konstantina Frangia-Tsivou, Aristotelis Chatziioannou

HAL Id: hal-01557640
https://hal.inria.fr/hal-01557640

Submitted on 6 Jul 2017

Integrative bioinformatic analysis of a Greek epidemiological cohort provides insight into the pathogenesis of primary cutaneous melanoma

Georgia Kontogianni[1, 2], Olga Papadodima[1], Ilias Maglogiannis[2], Konstantina Frangia-Tsivou[3] and Aristotelis Chatziioannou[1]

[1]Metabolic Engineering and Bioinformatics Group, Institute of Biology, Medicinal Chemistry and Biotechnology, National Hellenic Research Foundation, Athens, Greece
(gkontogianni,opapadod,achatzi)@eie.gr
[2]Department of Digital Systems, School of Information and Communication Technologies, University of Piraeus, Piraeus, Greece
imaglo@unipi.gr
[3]HistoBio Diagnosis, Athens, Greece
dfrangia@otenet.gr

Abstract. Melanoma is the most lethal type of skin cancer. In this study for the first time we analyze a Greek cohort of primary cutaneous melanoma biopsies, subjected to whole exome sequencing, in order to derive their mutational profile landscape. Moreover, in the context of big data analytical methodologies, we integrated the results of the exome sequencing analysis with transcriptomic data of cutaneous melanoma from GEO, in an attempt to perform a multi-layered analysis and infer a tentative disease network for primary melanoma pathogenesis. The purpose of this research is to incorporate different levels of molecular data, so as to expand our understanding of cutaneous melanoma and the broader molecular network implicated with this type of cancer. Overall, we showed that the results of the integrative analysis offer deeper insight in the underlying mechanisms affected by melanoma and could potentially contribute to the valuable effective epidemiological characterization of this disease.

Keywords: Data Integration· Next Generation Sequencing· Functional Analysis· Skin Cancer

1   Introduction

Melanoma is the most dangerous form of skin cancer [1]. Cutaneous melanoma (or melanoma of the skin), the most common type of melanoma, is a complex multi-factorial disease as both environmental and genetic factors are involved in its manifestation [2]. It is often a fatal neoplasm, derived from melanocytes, that accounts for most skin cancer deaths. In the advanced stages of this cancer, therapeutic intervention usually fails to improve survival despite recent advances in immunotherapy. According to the World Health Organization, 132,000 melanoma skin cancers occur globally each year and the global incidence of melanoma continues to increase, with a main predisposal factor; sun exposure.

The complexities of cellular metabolism and regulatory pathways involved have, until recently, obstructed the formulation of a unified description for melanoma [3]. Thus, despite the descent of gene signatures for various cancers, e.g. breast or colon cancer, a similar progress remains elusive for malignant melanoma. This could be attributed to the intricate nature of the molecular basis of cutaneous melanoma, which needs neatly stratified epidemiological cohorts to effectively address the issue of the high heterogeneity of this disease. In any case, genomic studies are limited by the shortage of similar melanoma cohorts, collecting and maintaining frozen tumor tissue, therefore rendering gene expression profiling studies of melanoma relatively scarce [4]. Still, efforts have been made to overcome any issues and shed some light on the underlying mechanisms involved with melanoma pathogenesis and metastases [4, 5]. A number of important emerging biological pathways and gene targets recently identified in melanoma are reported in [6]. Key biological pathways, where several significant genes (e.g. CDKN2A, CDK4, RB1) are involved, include proliferation, transcriptional control, extracellular matrix remodeling, glutamate signaling, and apoptosis.

In this study, we have focused on integrating different levels of molecular data through functional analysis to improve our understanding of the underlying mechanisms involved with melanoma. We incorporated established microarray datasets with next generation sequencing mutational data creating a potential disease network for melanoma.

Section 2 describes the techniques and methodology used in this study for the analysis of next generation sequencing and transcriptomic data. Then, in section 3 we present the results derived from the two datasets, in separate subsections, followed by another subsection for data integration. Finally, we conclude this study in section 4.

2   Materials & Methods

2.1   Analysis of Next Generation Exome Sequencing Data

The data analyzed in this section derived from Whole Exome Sequencing (WES) data of paired tumor and adjacent normal tissue from 9 patients with cutaneous melanoma (manuscript under preparation). The framework for the analysis of Next Generation Sequencing (NGS) data includes various state-of-the-art tools and has been previously presented by our team [7].

We first align the reads to the reference genome (hg19, version b37), using BWA (Burrows-Wheeler Aligner) [8] for DNA reads, adjusted for paired-end sequencing and run in consecutive steps for finding the correct coordinates and generate the final alignment in proper format. Then, we preprocess reads using Picard [9], for marking duplicate reads and sorting sequences according to the reference, to allow further processing with GATK (Genome Analysis Toolkit) [10], so as to ensure the quality of reads (all reads are given quality scores and can be dismissed if needed) and perform realignments and recalibrations based on the scores and references, to optimize the output reads and permit the following variance and somatic mutation investigation. Inspection for variance is performed with MuTect [11], which exploits statistical methodologies (Bayesian classifiers) and identifies sites of somatic mutations in paired datasets (tumor vs. normal). To annotate these sites, we use Oncotator [12], which utilizes several databases to link the sites to specific genes. Finally, we perform functional analysis to identify the molecular pathways affected by the specific mutations, and gene prioritization, so as to highlight genes with central role, implicated in diverse and major mechanisms in the Gene Ontology tree. These are performed using BioInfoMiner [13], which combines the StRAnGER2 [14] and GOrevenge [15] algorithms. Figure 1 presents the workflow used here.
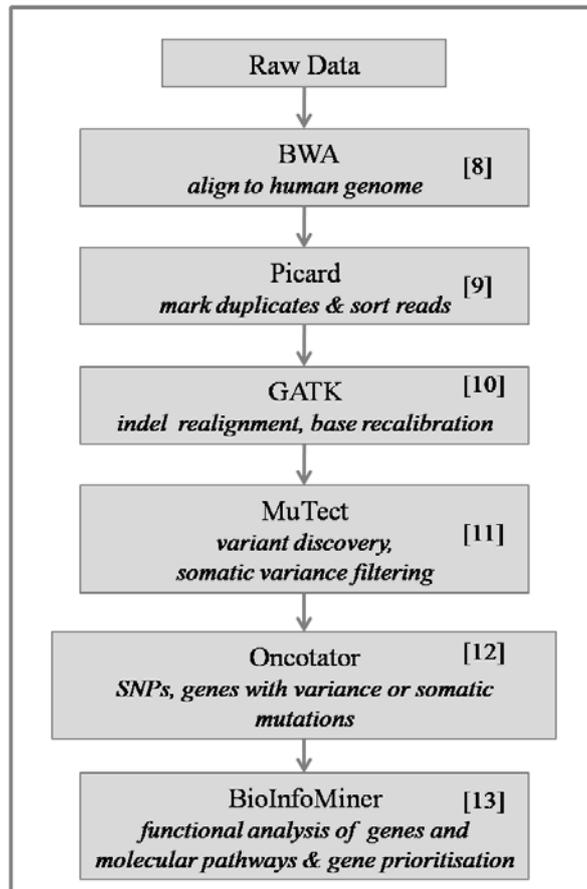
Fig. 1. Workflow of analysis for the identification of variance and somatic mutations

The complexity of NGS data is high, due to the high amount of information contained in each separate sample (compressed ~10Gb per sample/ 20Gb per patient/ ~150Gb for all) and the fact that several distinct parameters need to be adjusted at each step, so as to optimize the performance and the quality of the results (i.e., BWA needs to be adjusted for paired-end sequencing and run in consecutive steps for finding the correct coordinates and generate the final alignment in proper format).

2.2   Analysis of Transcriptomic Data

Transcriptomic analyses among different groups allow the exploration and identification of alterations in gene expression profiles between them. The data used in this section were previously analyzed in [16]. Briefly, the microarray dataset was taken from the Gene Expression Omnibus (GEO) [17], with accession number GDS1375. RNA was isolated from 45 primary melanoma, 18 benign skin nevi, and 7 normal skin tissue specimens and was analyzed for gene expression analysis, using the Affymetrix Hu133A microarray chip containing 22,000 probe sets. Following global normalization, gene expression values across all categories were log transformed, and the mean values of all genes in the normal skin were calculated. Afterwards, the mean gene vector concerning the normal skin categories was subtracted from all replicate vectors of the other two categories. The differentially expressed gene values of the melanoma versus skin, and nevi versus skin, were then analyzed. A false discovery rate for multiple testing adjustment, p-value 0.001 and a 2-fold change threshold for significant differential expression were applied and finally, 1425 unique genes were statistically selected, as being differentially expressed between melanoma and the normal state.

3 Results & Discussion

3.1 Mutational Data derived from Exome Sequencing

WES data derived from tumor and normal samples were aligned to the human genome, with an average sequence coverage of > 100x (number of reads aligning to known reference bases), ideal for achieving the mutational profile required. Overall, the individual samples have depth of coverage > 90, with only 3 samples achieving a lower score. Still this lower score is found only in normal samples, which does not affect further analysis, since high coverage is necessary mainly by the tumor samples to overcome endogenous heterogeneity. Table 1 contains the alignment rates and coverage for all samples that were examined. After the processing of individual samples for analysis of variance based on the reference genome, pairs of data from each patient (tumor vs. normal) are jointly analyzed, so as to identify somatic point mutations. Table 2 shows the number of putative sites of somatic mutations, after the MuTect analysis, as well as the count of missense and nonsense mutations for each patient. These mutations affect gene products, by amino acid substitutions or protein truncation, and require further analysis as candidate genetic biomarkers. It is worth noting that the complete workflow for a pair of samples (tumor and normal samples from one patient) needs approximately 35 hours running time on a 64Gb RAM/ 12 processor cluster server (finally, summarizing the results in ~10Mb).

Table 1. Alignment rates and coverage

patient	Normal alignment rate	coverage	Tumor alignment rate	coverage
1	96	129.8	95.4	118.9
2	84.2	70.5	89.5	91
3	66.7	35.9	93.6	102.2
4	93.8	103.9	92.6	111.5
5	96.8	101.5	96.7	111.4
6	96.8	123.3	96.2	104.8
7	97.4	117.4	96.8	111.9
8	87.9	88.7	88.8	92.3
9	95.6	128.4	92	120.7

Table 2. Number of somatic mutations, missense/nonsense mutations, and unique genes affected per patient

Patient	Sites of somatic mutations	Missense/Nonsense mutations	Unique genes affected
1	855	224	214
2	1134	309	295
3	826	281	265
4	73	10	10
5	944	275	265
6	5985	1811	1474
7	812	226	200
8	922	224	219
9	1111	224	214

In order to discover the molecular pathways affected by the specific mutations, after annotating the mutations to specific genes, we performed functional analysis of the union of affected genes from all the patients (2685 unique genes), which revealed 40 statistically significant biological processes (p-value < 0.05), shown in table 3.

Table 3. Table of the significant biological processes influenced by the mutated genes. Enrichment represents the ratio of the number of genes in the input list annotated with a GO term to the total number of genes annotated to this specific term

Term id	Term Definition	Enrichment	Hypergeometric p-value	Corrected p-value
GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	69/150	4.33E-20	0.0014
GO:0007155	cell adhesion	148/531	2.17E-15	0.0027
GO:0050911	detection of chemical stimulus involved in sensory perception of smell	105/389	1.68E-10	0.0037
GO:0030198	extracellular matrix organization	84/313	1.38E-08	0.0048
GO:0086010	membrane depolarization during action potential	17/30	1.10E-07	0.0063
GO:0007411	axon guidance	95/375	3.64E-08	0.0068
GO:0006811	ion transport	82/319	1.53E-07	0.0101
GO:0022617	extracellular matrix disassembly	39/117	2.98E-07	0.0108
GO:0006814	sodium ion transport	37/106	1.59E-07	0.0115
GO:0055085	transmembrane transport	162/767	7.19E-07	0.012
GO:0007608	sensory perception of smell	61/224	7.06E-07	0.0125
GO:0019228	neuronal action potential	16/31	1.42E-06	0.0144
GO:0035725	sodium ion transmembrane transport	30/89	5.11E-06	0.0145
GO:0007268	synaptic transmission	97/428	5.75E-06	0.0178
GO:0042391	regulation of membrane potential	36/117	6.79E-06	0.0195
GO:0007186	G-protein coupled receptor signaling pathway	192/976	9.61E-06	0.0198
GO:0030574	collagen catabolic process	26/74	9.12E-06	0.0203
GO:0007605	sensory perception of sound	39/133	1.03E-05	0.0223
GO:0034765	regulation of ion transmembrane transport	35/118	2.18E-05	0.0257
GO:0060080	inhibitory postsynaptic potential	8/11	2.26E-05	0.0257
GO:0070588	calcium ion transmembrane transport	38/129	1.19E-05	0.0258
GO:0018108	peptidyl-tyrosine phosphorylation	37/130	3.53E-05	0.0287
GO:0016339	calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules	13/27	3.64E-05	0.0306
GO:0070509	calcium ion import	13/28	5.89E-05	0.0323
GO:0007018	microtubule-based movement	24/74	8.72E-05	0.0331
GO:0001539	cilium or flagellum-dependent cell motility	6/7	5.96E-05	0.034
GO:0032228	regulation of synaptic transmission, GABAergic	7/10	0.0001	0.0353
GO:0007399	nervous system development	72/322	0.0001	0.0376
GO:0007169	transmembrane receptor protein tyrosine kinase signaling pathway	33/119	0.0002	0.0382
GO:0034220	ion transmembrane transport	65/286	0.0002	0.0395
GO:0001964	startle response	10/20	0.0002	0.0399
GO:0050907	detection of chemical stimulus involved in sensory perception	28/96	0.0002	0.0405
GO:0007416	synapse assembly	17/47	0.0002	0.0445
GO:0071625	vocalization behavior	7/12	0.0006	0.0447
GO:2000821	regulation of grooming behavior	4/4	0.0005	0.0455
GO:0016337	single organismal cell-cell adhesion	30/109	0.0003	0.0465
GO:0030534	adult behavior	12/29	0.0004	0.0468
GO:0034332	adherens junction organization	14/38	0.0006	0.0476
GO:0034329	cell junction assembly	22/76	0.001	0.0492
GO:0015721	bile acid and bile salt transport	11/27	0.0009	0.0493

3.2 Transcriptomic Data

The transcriptomic analysis from [16] revealed 1425 unique differentially expressed genes. Enrichment analysis showed 36 statistically significant biological processes (p-value < 0.05), which are presented in Table 4.

Table 4. Table of the significant biological processes influenced by the differentially expressed genes. Enrichment represents the ratio of the number of genes in the input list annotated with a GO term to the total number of genes annotated to this specific term

Term id	Term Definition	Enrichment	Hypergeometric p-value	Corrected p-value
GO:0030198	extracellular matrix organization	66/313	0.00000676	0.0014
GO:0008544	epidermis development	31/109	0.00000027	0.0033
GO:0030216	keratinocyte differentiation	19/56	0.000003067	0.0043
GO:0006094	gluconeogenesis	16/48	0.00002341	0.0053
GO:0048013	ephrin receptor signaling pathway	21/91	0.0005	0.0078
GO:0060512	prostate gland morphogenesis	4/4	0.0001	0.0079
GO:0033599	regulation of mammary gland epithelial cell proliferation	4/5	0.0006	0.0094
GO:0045861	negative regulation of proteolysis	9/26	0.0011	0.0114
GO:0061436	establishment of skin barrier	7/17	0.0012	0.0116
GO:0060326	cell chemotaxis	15/57	0.0008	0.0132
GO:0071230	cellular response to amino acid stimulus	13/48	0.0013	0.0155
GO:0051591	response to cAMP	14/54	0.0013	0.0157
GO:0048538	thymus development	12/45	0.0022	0.0182
GO:0045669	positive regulation of osteoblast differentiation	14/57	0.0023	0.0199
GO:0001954	positive regulation of cell-matrix adhesion	8/23	0.0019	0.021
GO:0042060	wound healing	20/95	0.0024	0.022
GO:0007155	cell adhesion	78/531	0.0028	0.0235
GO:0061036	positive regulation of cartilage development	6/15	0.0032	0.0236
GO:0022617	extracellular matrix disassembly	23/117	0.003	0.025
GO:0045765	regulation of angiogenesis	9/30	0.0033	0.027
GO:0071526	semaphorin-plexin signaling pathway	7/20	0.0036	0.0292
GO:0048661	positive regulation of smooth muscle cell proliferation	13/54	0.004	0.0298
GO:0050773	regulation of dendrite development	5/11	0.0038	0.0313
GO:0048678	response to axon injury	9/32	0.0053	0.0337
GO:0010951	negative regulation of endopeptidase activity	26/144	0.0056	0.0343
GO:0061621	canonical glycolysis	8/27	0.0059	0.0346
GO:0070373	negative regulation of ERK1 and ERK2 cascade	12/50	0.0057	0.0374
GO:0055086	nucleobase-containing small molecule metabolic process	16/78	0.008	0.0402
GO:0007160	cell-matrix adhesion	18/92	0.0084	0.0402
GO:0060441	epithelial tube branching involved in lung morphogenesis	6/17	0.0066	0.0405

GO:0030032	lamellipodium assembly	9/33	0.0066	0.0407
GO:0030324	lung development	20/106	0.0086	0.0435
GO:0002009	morphogenesis of an epithelium	7/23	0.0084	0.045
GO:0043153	entrainment of circadian clock by photoperiod	6/18	0.009	0.0454
GO:0007266	Rho protein signal transduction	13/59	0.0087	0.0465
GO:0030855	epithelial cell differentiation	16/79	0.009	0.0483

3.3   Data Integration

To facilitate a deeper examination of our datasets, we compared the gene lists from the mutational and transcriptomic analyses. Figure 2 illustrates the total unique and common genes, from the two types of datasets. Only 5% of the total genes were common between the two sets. Nevertheless, among the highly ranked processes, presented in tables 3 and 4, cell adhesion, extracellular matrix organization and extracellular matrix disassembly, containing a large number of genes, are found as significantly affected in both cases.
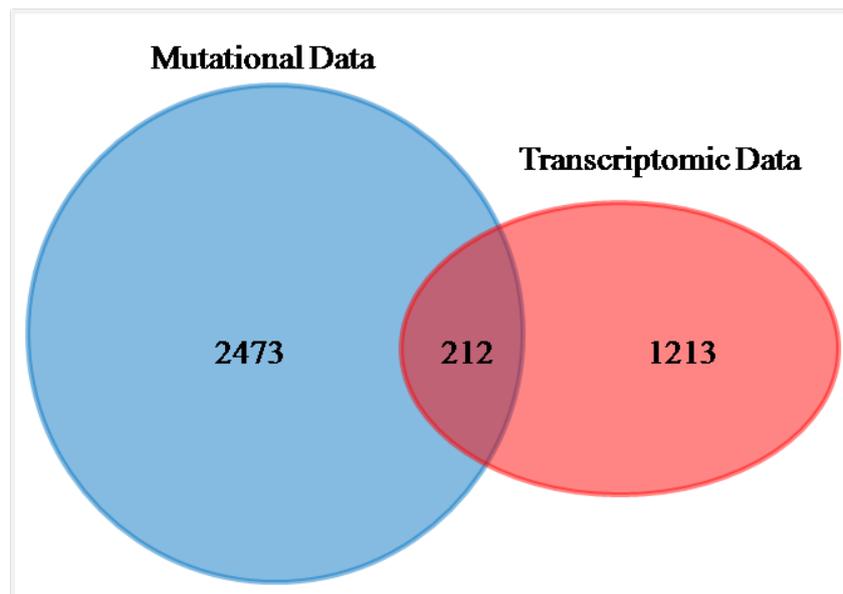
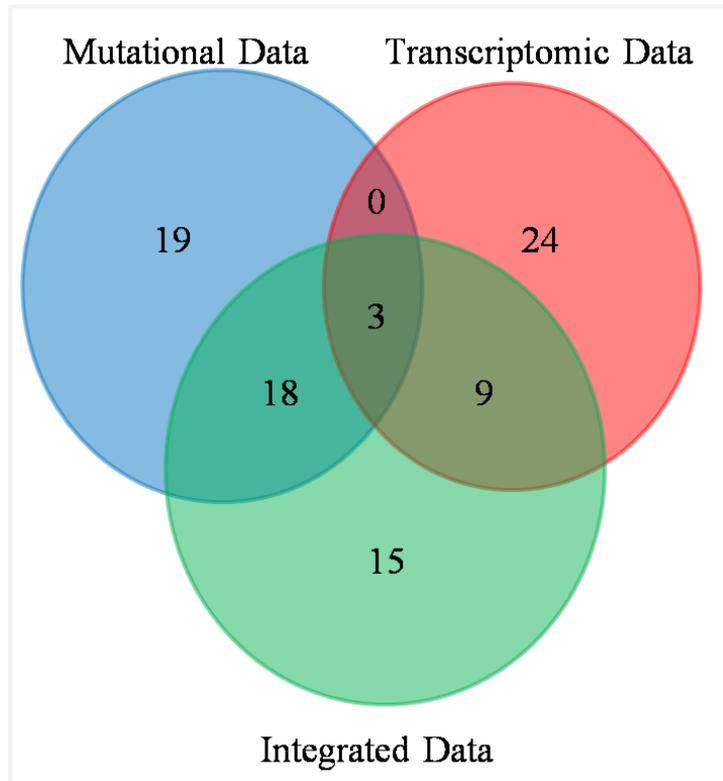**Fig. 2.** Venn diagram for the significant gene lists from the two analyses

In order to create a feasible disease network for melanoma, we merged the previous results, and carried out an additional functional analysis. This enrichment analysis revealed 45 statistically significant biological processes (p-value < 0.05), presented in Figure 3, ranked according to their corrected p-values.
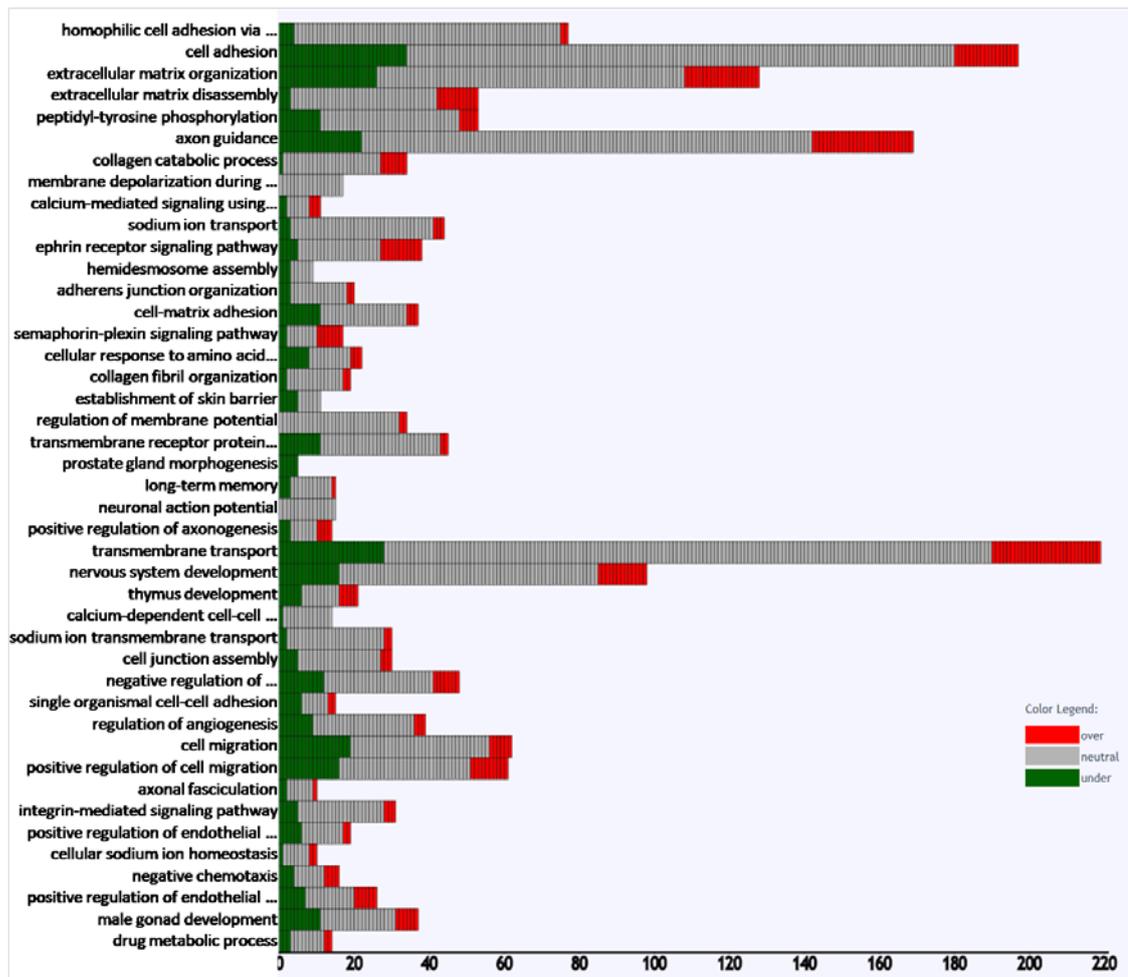
Fig. 3. Bar plot of significant terms with the number of associated genes (x-axis). Terms are ranked using the corrected p-value. The colors of the genes specify their expression fold change, green -on the left- for under-expressed genes and red -on the right- for over-expressed genes, with neutral indicating somatic mutation

This potential Disease Network revealed several mechanisms with known significance, consistent with melanoma. Enrichment of GO terms, such as epithelial tube branching involved in lung morphogenesis, morphogenesis of an epithelium, epithelial cell differentiation, and regulation of mammary gland epithelial cell proliferation reflects the topological origin of cutaneous melanoma [18, 19]. Furthermore, cell-matrix procedures (organization, adhesion) have been previously reported as significantly altered in tumors [20, 21], as well as lamellipodium assembly, an essential structure for cell migration, which plays an important part in cell invasion and metastasis of cancer [22, 23]. In relation to the ephrin receptor and Rho protein signaling pathways, the Eph receptor tyrosine kinases and their ephrin ligands have specific expression patterns in cancer cells [24], while Rho-like GTPase have been identified as key regulators of epithelial architecture and cell migration, both correlated to cancer development [25, 26].

As expected, the previously discussed significant pathways from tables 3 and 4 are complemented by the additional data, incorporating an increased number of genes, with considerable implication in melanoma manifestation and progression. Among the significant processes are several previously highlighted by the distinct datasets, but also a number of newly generated, after data integration. Figure 4 indicates the unique and common pathways in each case.

Fig. 4. Venn diagram for the significant pathway lists from the two distinct analyses, as well as their integration

4 Conclusions & Future Work

In this study, we sought to export the broader molecular network implicated with cutaneous melanoma. We integrated molecular data of different levels in order to identify the important mechanisms that are involved in this type of cancer. This integration advanced our understanding about the mechanisms implicated with melanoma, by observing the correlation between different sets and levels of data. More importantly, it allowed the manifestation of additional mechanisms previously concealed by the statistical cut-offs, thus enhancing the disease network and our general understanding of the phenomenon.

Our future aim is to expand our current dataset, including data from more patients. Apart from that, we want explore additional methodologies for data integration. Furthermore, our goal is to integrate the molecular data with imaging data from dermoscopy, to improve feature selection and classification techniques, concerning melanoma.

Acknowledgements

References

1. The Skin Cancer Foundation http://www.skincancer.org/
2. Rossi, C. R., Foletto, M., Vecchiato, A., Alessio, S., Menin, N., & Lise, M. (1997). Management of cutaneous melanoma M0: state of the art and trends. *European Journal of Cancer*, *33*(14), 2302-2312.
3. Dummer, R., & Hoek, K., Human Melanoma: From Transcriptome to Tumor Biology, Forschungsdatenbank der Universität Zürich, 2004-2008
4. Winnepenninckx, V., Lazar, V., Michiels, S., Dessen, P., Stas, M., Alonso, S. R., ... & Eggermont, A. M. (2006). Gene expression profiling of primary cutaneous melanoma and clinical outcome. *Journal of the National Cancer Institute*, *98*(7), 472-482.
5. Raskin, L., Fullen, D. R., Giordano, T. J., Thomas, D. G., Frohm, M. L., Cha, K. B., ... & Gruber, S. B. (2013). Transcriptome profiling identifies HMGA2 as a biomarker of melanoma progression and prognosis. *Journal of Investigative Dermatology*, *133*(11), 2585-2592.
6. Dutton-Regester, K., & Hayward, N. K. (2012). Reviewing the somatic genetics of melanoma: from current to future analytical approaches. *Pigment cell & melanoma research*, *25*(2), 144-154.
7. Maglogiannis, I., Goudas, T., Billiris, A., Karanikas, H., Valavanis, I., Papadodima, O., Kontogianni, G., & Chatziioannou, A. (2015, September). Redesigning EHRs and Clinical Decision Support Systems for the Precision Medicine Era. In *Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS)* (p. 14). ACM.
8. Li H. & Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. Bioinformatics, Epub. [PMID: 20080505]
9. Picard tools http://picard.sourceforge.net
10. McKenna, A.,. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, *20*(9), 1297-1303.
11. Cibulskis, K., *et al.* (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, *31*(3), 213-219.
12. Ramos, A. H., Lichtenstein, L., Gupta, M., Lawrence, M. S., Pugh, T. J., Saksena, G., ... & Getz, G. (2015). Oncotator: cancer variant annotation tool. *Human mutation*, *36*(4), E2423-E2429.
13. e-Nios BioInfoMiner tool https://bioinfominer.com/#/welcome
14. Pilalis, E.D., Chatziioannou, A.A., (2013)"Prioritized functional analysis of biological experiments using resampling and noise control methodologies," Bioinformatics and Bioengineering (BIBE), 2013 IEEE 13th International Conference on , vol., no., pp.1,3, doi: 10.1109/BIBE.2013.6701558
15. Moutselos, K., (2011). GOrevenge: a novel generic reverse engineering method for the identification of critical molecular players, through the use of ontologies. *Biomedical Engineering, IEEE Transactions on*, *58*(12), 3522-3527.
16. Moutselos, K., Maglogiannis, I., & Chatziioannou, A. (2014). Integration of High-Volume Molecular and Imaging Data for Composite Biomarker Discovery in the Study of Melanoma. *BioMed research international*, *2014*.
17. Barrett T., Troup D.B, Wilhite S.E. et al., "NCBI GEO: archive for functional genomics data sets - 10 years on," Nucleic Acids Res, vol. 39, no. Database issue, pp. D1005-10, Jan, 2011.Chevalier, R.L.: Obstructive nephropathy: towards biomarker discovery and gene therapy. Nat. Clin. Pract. Nephrol. 2(3), 157--168 (2006).
18. Martin-Belmonte F. & Perez-Moreno M., "Epithelial cell polarity, stem cells and cancer", *Nature Reviews Cancer*, January 2012, 12, 23-38.
19. Jögi, A., Vaapil, M., Johansson, M., & Påhlman, S. (2012). Cancer cell differentiation heterogeneity and aggressive behavior in solid tumors. *Upsala journal of medical sciences*, *117*(2), 217-224.
20. Saladi SV, et al, "Modulation of extracellular matrix/adhesion molecule expression by BRG1 is associated with increased melanoma invasiveness.", Mol Cancer. 2010 Oct 22;9:280.
21. Hart, I. R., Birch, M., & Marshall, J. F. (1991). Cell adhesion receptor expression during melanoma progression and metastasis. *Cancer and Metastasis Reviews*, *10*(2), 115-128.
22. Machesky, L. M. (2008). Lamellipodia and filopodia in metastasis and invasion. *FEBS letters*, *582*(14), 2102-2111.
23. Kato, T., Kawai, K., Egami, Y., Kakehi, Y., & Araki, N. (2014). Rac1-dependent lamellipodial motility in prostate cancer PC-3 cells revealed by optogenetic control of Rac1 activity. *PloS one*, *9*(5), e97749.
24. Pasquale, E. B. (2010). Eph receptors and ephrins in cancer: bidirectional signalling and beyond. *Nature Reviews Cancer*, *10*(3), 165-180.
25. Ridley, A. J. (2004). Rho proteins and cancer. *Breast cancer research and treatment*, *84*(1), 13-19.
26. Sander, E. E., & Collard, J. G. (1999). Rho-like GTPases: their role in epithelial cell–cell adhesion and invasion. *European Journal of Cancer*, *35*(9), 1302-1308.