Development of a classification system for the translational analysis of cancer genomic and imaging data for melanoma prognosis

Georgia Kontogianni[1,2], Olga Papadodima[2] , Irene Liampa[2,3], Hector Xavier de Lastic[2,3], Ilias Maglogiannis[1], and Aristotelis Chatziioannou[2,4]

[1]Department of Digital Systems, School of Information and Communication Technologies, University of Piraeus, Piraeus, Greece
[2]Metabolic Engineering and Bioinformatics Group, Institute of Biology, Medicinal Chemistry and Biotechnology, NHRF, Athens, Greece
[3]Department of Molecular Biology and Genetics, Democritus University of Thrace, Dragana, Greece.
[4]e-NIOS, Kallithea-Athens, Greece

The onset and constant advancement of molecular technologies has enabled the parallel, high-throughput process of millions of sequence reads, thus ushering a new era with numerous, novel applications in basic, applied and clinical research. Here, we present an analysis framework of NGS genomic data, integrating functional and pathway analyses, for the inference of gene signatures with diagnostic and/or prognostic value. In addition, a system for the integration and processing of heterogeneous, multi-layered (omics, histological images and clinical) data, is presented. A pilot analysis was performed including nine patients with cutaneous melanoma, resulting in a short list of candidate genes with a probable causative role in the disease. Since the number of patients analysed was limited, additional samples from public databases and datasets were acquired. Based on the identified gene signature, a classifier, exploiting mutational data, was built. This classifier was able to distinguish melanomas from dysplastic nevi, with high accuracy. In the interests of reproducibility, streamlined tools were used, resulting in standardised workflows. Aiming at the multi-angled depiction of melanoma, our classification system will integrate and exploit dermoscopic and clinical data.

Molecular Data Analysis:

Molecular Data Analysis concerns raw next generation sequencing (NGS) data derived from exome sequencing of melanoma tissue and matched healthy control. The framework of analysis of NGS data (Figure 1) has been previously presented by our team. A pilot analysis was performed including nine patients.

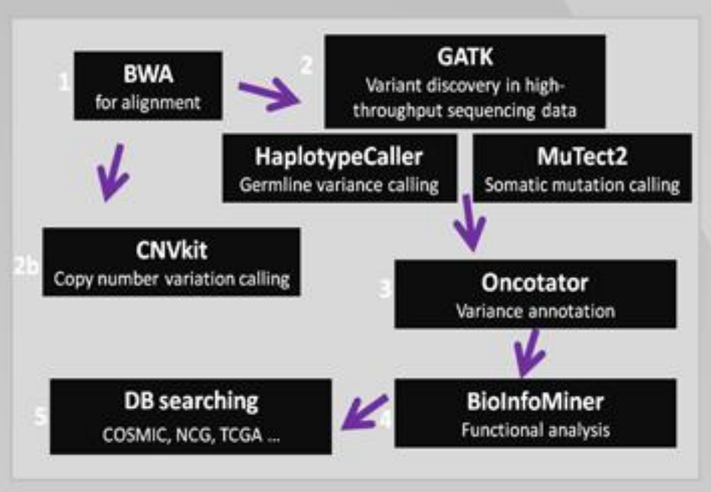

Figure 1. Workflow of analysis for the identification of germ line variance and somatic mutations. The outcome of this analysis is a list of significantly mutated genes characterising melanoma.

Results:

- Identified a total of 10016 somatic mutations in all patients.
- Comparable numbers of somatic mutations (median: 589), with the exception of two patients.
- Median mutation frequency was 12.75 mutations/Mb, consistent with melanoma high mutational burden.

Patients:	3	5	8	10	11	12	13	14	15
Somatic Mutations	522	693	901	27	935	5324	511	589	528
Non-synonymous	227	284	387	18	364	2036	222	224	193
Mutation Frequency	10.4	13.8	17.9	0.5	18.6	105.7	10.1	11.7	10.5
C>T%	42.1	41.4	43.6	29.6	42.5	51	42.9	38.9	42.6
C>T/G>A %	84.3	86.4	86.7	55.6	80.6	94.9	87.5	81.2	84.7

Table 1. Number of somatic mutations after strand-specific artefact removal, mutation frequency per Mb, C>T and C>T & G>A rates

- Functional analysis was performed on the union of non-synonymous mutations for all the patients, corresponding to 1587 genes, excluding the genes that were solely mutated in patient 12 (1303 genes), to avoid patient-specific bias. Table 2 presents the top 30 prioritised genes.

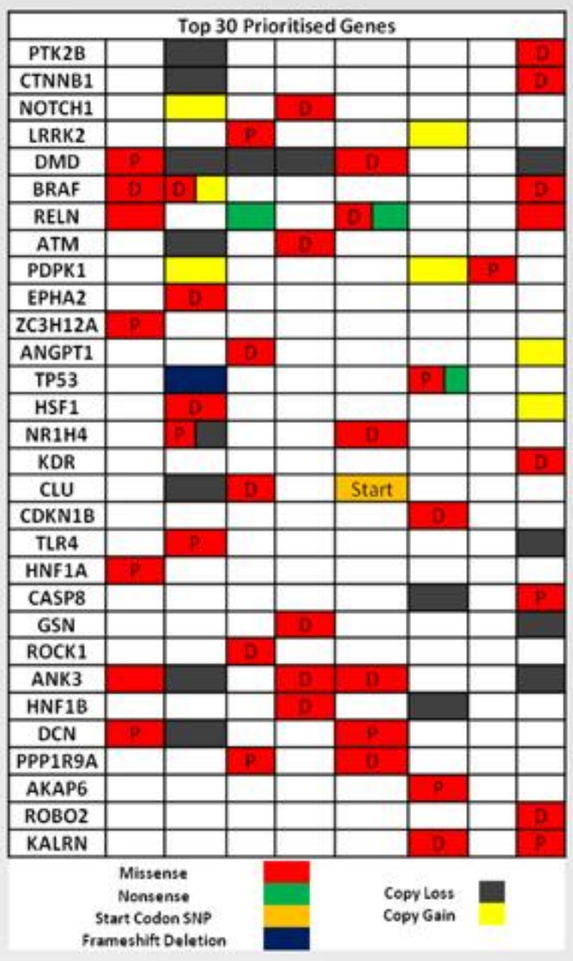

Table 2. BioInfoMiner analysis results, based on their centrality (genes taking part in numerous distinct mechanisms are ranked higher), exploiting semantic information and network analysis: Top 30 prioritised genes, according to their centrality, as described in Gene Ontology & Reactome vocabularies, and the mutations they carry in different patients

Classification

- Since the number of patients analysed was limited, we added samples from cBioPortal. As healthy state (non-melanoma) we used mutational data from dysplastic nevus that were acquired through similar experimental procedure by [Melamed et al. 2017]. On the molecular level, this state holds a considerably lower mutational load compared to melanoma.
- For feature selection, we reduced the list of mutated genes by prioritising them according to their centrality using BioInfoMiner.
- The samples (samples of dysplastic nevus and melanoma) were separated under two labels, dysplastic nevus (represented by DNS) and melanoma (represented by MEL) and each sample is attributed a 157-dimensional binary vector showing if the corresponding gene contains a mutation or not.
- To deal with unbalanced classes, the SMOTE algorithm was utilised to generate synthetic data for the DNS label. Figure 2 graphically shows data assortment for the two classes. Due to the binary type of the classification problem, the Random Forests (RF) algorithm was selected, as an appropriate and effective methodology.

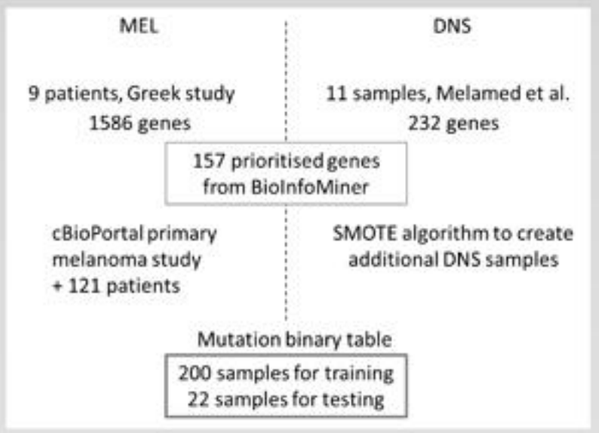

Figure 2. Data assortment for classification

The best performance was reported for the RF classifier with the following parameters:

- 200 samples for training, 22 for testing
- 157 predictors
- 2 classes: 'DNS', 'MEL'
- No pre-processing
- Resampling: Cross-Validated (10 fold, repeated 3 times)
- mtry = 22
- ntree = 100

As a criterion for the cross validation performance, the receiver operating characteristic (ROC) curve was used, which controls the sensitivity with respect to the specificity. The area under the curve (AUC) of the plot gives an unbiased estimation of the classifier's performance at each round. The results are shown in Figure 3.

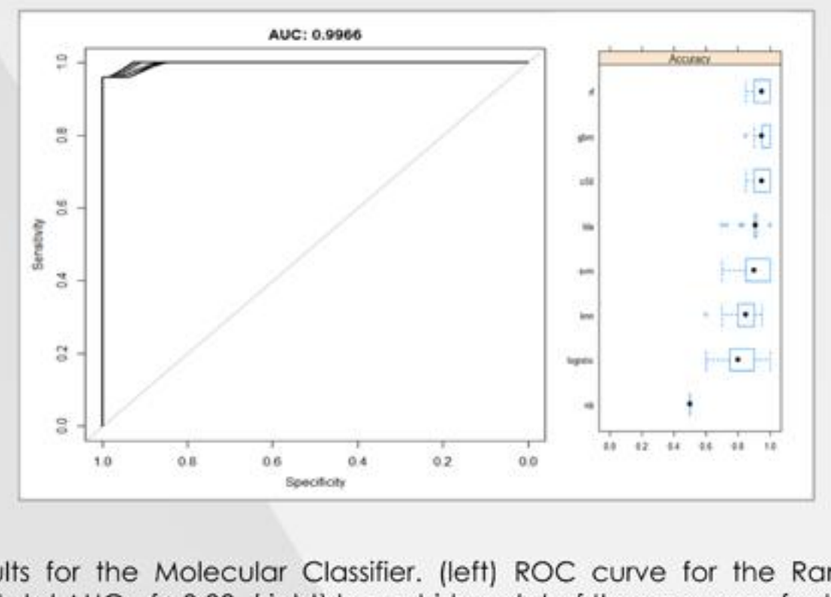

Figure 3. Results for the Molecular Classifier. (left) ROC curve for the Random Forests classifier, with total AUC of >0.99, (right) box-whisker plot of the accuracy for the examined classification algorithms, rf-Random Forests, gbm-Stochastic Gradient Boosting, c50-Decision Trees C5.0, lda-Linear Discriminant Analysis, svm-Support Vector Machines, knn-k Nearest Neighbours, logistic-Logistic Regression, nb-Naive Bayes

The classifier performed very well, reaching a mean accuracy of 0.96. This result justifies the utilisation of this classifier as a model for class prediction (melanoma vs. dysplastic nevus) of unknown samples of mutation data.

Discussion & Future Plans:

- In this study we analyse primary melanomas from Greek patients, by whole exome sequencing, in order to derive their mutational profile, by exploiting various databases to develop a more accurate cancer signature and identify possibly causative genes.
- The fact that diverse genes in each patient are found mutated, targeting the same pathways, predicates upon the possibility of different, potentially pathogenic modes of molecular perturbation, concerning those pathways. A short list of promising genes for further analysis has been identified.
- Integration of molecular features with dermoscopic imaging features, from our previous work. Ultimate aim is to develop a composite signature, based on the 157-molecular & 31-dermoscopic features that have been isolated.
- Constantly update our database through the incorporation of new cases, allowing the accurate patient classification, towards a precision medicine approach. TRANSITION aims at the holistic study of Cutaneous Melanoma through an intelligent, multi-layered, combinatorial analytical strategy integrating demographic, clinical, imaging and molecular data of patients.

References

1.COSMIC database. Available at: http://cancer.sanger.ac.uk. (Accessed: January 2017)
2.Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. in Nat Methods 7, 248–9 (2010).
3.An, O. et al. NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. Nucleic acids research 44, D992-9 (2016).
4.Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. (2012).
5.Chawla, N. V. et al. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357 (2002).
6.Chen, X. & Ishwaran, H. Random forests for genomic data analysis. Genomics 99, 323–9 (2012).
7.Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology 31, 213–9 (2013).
8.Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science signaling 6, pl1 (2013).
9.Hajian-Tilaki, K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian journal of internal medicine 4, 627–35 (2013).
10.Kontogianni, G. et al. Dissecting the Mutational Landscape of Cutaneous Melanoma: An Omic Analysis Based on Patients from Greece. Cancers 10, 96 (2018).
11.Kontogianni, G. et al. Integrative Bioinformatic Analysis of a Greek Epidemiological Cohort Provides Insight into the Pathogenesis of Primary Cutaneous Melanoma. in (2016).
12.Koutsandreas, T. et al. Analyzing and visualizing genomic complexity for the derivation of the emergent molecular networks. International Journal of Monitoring and Surveillance Technologies Research (IJMSTR) 4, 30–49 (2016).
13.Kuhn, M. Caret: classification and regression training. Astrophysics Source Code Library (2015).
14.Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499, 214–8 (2013).
15.Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) 26, 589–95 (2010).
16.McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research 20, 1297–303 (2010).
17.Melamed, R. D. et al. Genomic characterization of dysplastic nevi unveils implications for diagnosis of melanoma. Journal of investigative Dermatology 137, 905–909 (2017).
18.Ramos, A. H. et al. Oncotator: cancer variant annotation tool. Human mutation 36, E2423-9 (2015).
19.Developement, R. Core Team (2011) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. (ISBN 3-900051-07-0. Available: h ttp://www. R-project. org).
20.Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics 12, 77 (2011).
21.Talevich, E. et al. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS computational biology 12, e1004873 (2016).
22.Torgo, L. Data mining with R: learning with case studies. (CRC press, 2016).

We acknowledge support of this work by the project "TRANSITION: **TRANSlating the complexity of melanoma diagnosis into raTional therapeutic stratificatiON**" (MIS 5031295) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

We acknowledge support of this work by the project "ELIXIR-GR: Hellenic Research Infrastructure for the Management and Analysis of Data from the Biological Sciences" (MIS 5002780) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).