The 14th Annual International Conference On Critical Assessment of Massive Data Analysis (CAMDA)

Djork-Arné Clevert	Johannes Kepler University Linz
Joaquín Dopazo	Centro de Investigación Príncipe Felipe (CIPF)
Sepp Hochreiter	Johannes Kepler University Linz
Lan Hu	Exosome Diagnostic
David P. Kreil	Boku University

CAMDA focuses on the analysis of massive data in life sciences. The conference presents new techniques in the field of bioinformatics, data analysis, and statistics for the handling and processing large data sets, the combination of multiple data sources, and computational inference. An essential part of CAMDA is its open-ended data analysis challenge which focuses on big heterogeneous data sets. The winners of the annual challenges are voted by the delegates at the end of the conference. CAMDA has a track record as a well-recognized annual conference going back to the year 2000, and it has become an ISMB Satellite Meeting since 2011.

http://www.camda.info

Keynote Speakers

Des Higgins, PhD - Making and using extremely large multiple sequence alignments.

Short Bio: Des Higgins is professor of Bioinformatics in University College Dublin, Ireland and has been working on sequence alignment and molecular evolution since the mid 1980s. He originated the widely used Clustal package and continues to develop and maintain multiple sequence alignment algorithms and software. He also works on the analysis of high throughput genomics data, especially in the application of multivariate analysis methods for data integration. He has published more than 130 peer-reviewed articles in bioinformatics, sequence alignment and genomics with an h-index of 54.

Christopher E. Mason, PhD - Leveraging short and long reads for optimal RNA-Sequencing with CAMDA data set #1.

Short Bio: Dr. Mason founded his laboratory as an assistant professor at Weill Cornell Medical College in the Department of Physiology and Biophysics and at the Institute for Computational Biomedicine. Professor Mason also holds an appointment in the Tri-Institutional Program on Computational Biology and Medicine between Cornell, Memorial Sloan-Kettering Cancer Center and Rockefeller University and he also has an appointment at the Weill Cornell Cancer Center and the Brain and Mind Research Institute. In 2013, he won the Hirschl-Weill-Caulier Career Scientist Award. In 2014, he won the Vallee Foundation Young Investigator Award, the CDC Honor Award for Standardization of Clinical Testing, and he was just named as one of the “Brilliant Ten” Scientists in the world by Popular Science magazine.

Friday July 10th 2015

07:30 - 09:00	ISMB / CAMDA Registration
09:00 - 09:15	CAMDA Welcome
09:15 - 10:15	**Keynote: Des Higgins**, **University College Dublin, Ireland** Making and using extremely large multiple sequence alignments
10:15 - 10:45	*Morning break*
10:45 - 11:10	**Djork-Arné Clevert, Johannes Kepler University Linz, Austria** Setting the context
11:10 - 11:50	**Hubert Rehrauer, ETH Zurich and University of Zurich, Switzerland** Prognostic value of cross-omics screening for cancer survival
11:50 – 12:30	**Marta R. Hidalgo, Centro de Investigación Príncipe Felipe (CIPF), Spain** Functional hallmarks in clear cell renal cell carcinoma grade and stage progression revealed by changes in signalling circuit activities
12:30 - 13:30	*Lunch break*
13:30 - 14:10	**Jari Björne, University of Turku, Finland** Cancer progression classification for mutation analysis
14:10 - 14:50	**Sinjini Sikdar, University of Louisville, USA** Exploring the importance of cancer pathways by meta-analysis of differential protein expression networks in three different cancers
14:50 - 15:30	**Pramila Tata, Agilent Technologies, USA** Multi-Omics analysis for understanding the molecular basis of Lung Adenocarcinoma
15:30 - 16:00	*Afternoon break*
16:00 - 16:40	**Alejandra Cervera, University of Helsinki, Finland** A pipeline for exploratory and pathway analysis of NGS data
16:40 - 17:20	**Jose Carbonell, Centro de Investigación Príncipe Felipe (CIPF), Spain** Integrative Gene Set Analysis of mRNA and miRNA expression data
17:20 - 18:00	**Olivier Gevaert, Stanford University, USA** Multi-omics fusion for cancer data
19:00	CAMDA Dinner. Tickets are available via ISMB website!

Saturday July 11th 2015

09:05 - 09:15	CAMDA Welcome
09:15 - 10:15	**Keynote: Christopher E. Mason, Weill Cornell Medical College, USA** Leveraging short and long reads for optimal RNA-Sequencing with CAMDA data set #1
10:15 - 10:45	*Morning break*
10:45 - 11:10	**Weida Tong, US Food and Drug Administration, USA** Setting the context
11:10 - 11:50	**Cankut Cubuk, Centro de Investigación Príncipe Felipe (CIPF), Spain** Signalling circuit activities as mechanism-based features to predict mode of action of chemicals
11:50 – 12:30	**Chathura Siriwardhana, University of Louisville, USA** Inter-platform concordance of gene expression data for the prediction of chemical mode of action
12:30 - 13:30	*Lunch break*
13:30 - 13:50	**Serghei Mangul, UCLA, USA** Examining lost reads to survey the microbiome and immune components of the human body across 43 human sites from 175 individuals
13:50 - 14:30	**Pawel P. Labaj, Boku University Vienna, Austria** Sensitivity, specificity and reproducibility of RNA-Seq differential expression calls
14:30 - 14:50	**Najmeh Abiri, Lund University, Sweden** Unbiased Optimization of Microarray Pre-processing.
14:50 - 15:30	**Udo Gieraths, ETH Zurich, Switzerland** Genome-wide detection of intervals of genetic heterogeneity associated with complex traits
15:30 - 16:00	*Afternoon break*
16:00 - 16:20	**Aleksandra Gruca, Silesian University of Technology, Poland** New Gene Ontology term similarity measure - comparison and performance evaluation based on DNA microarray data
16:20 - 17:00	**Menno J. Witteveen, ETH Zurich, Switzerland** In silico phenotyping via co-training for improved phenotype prediction from genotype
17:00 - 17:30	Closing address and CAMDA contest awards

The 14th Annual International Conference On Critical Assessment of Massive Data Analysis (CAMDA)

Extended Abstracts

Prognostic value of cross-omics screening for cancer survival

Slavica Dimitrieva [1,*], Ralph Schlapbach [1] and Hubert Rehrauer [1]

Functional Genomics Center Zurich, ETH Zurich and University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

* Correspondence to: slavica.dimitrieva@fgcz.ethz.ch

Introduction

Large-scale molecular profiling of cancers offers a great potential to advance our understanding of the development and progression of this disease. Systematic cancer genomics projects, like The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC), have applied high-throughput genome analysis techniques to generate genomics, transcriptomics, epigenomics and clinical data for several cancers. These data can be informative for multiple aspects ranging from discovering of new markers for more accurate cancer diagnosis and prognosis, to development of new therapeutics and personalized treatments.

The overall goal of our study, as a response to one of the CAMDA 2015 Challenges, is to gain novel biological insights into three less well studied cancers: Lung Adenocarcinoma (LUAD), Kidney Renal Clear Cell Carcinoma (KIRC) and Head and Neck Squamous Cell Carcinoma (HNSC). We performed a systematic analysis of genome-wide molecular datasets provided from the ICGC Data Portal (miRNA, mRNA and protein expression, somatic copy-number variation (CNV) and DNA methylation profiles) to investigate underlying mechanisms of cancer initiation and progression. Cancer is an extremely complex disease and it is of no surprise that previous genomics analyses have revealed extensive tumor heterogeneity[1]. As consequence, the identification of molecular signatures from genomics analyses that can give accurate prediction and prognosis of response to therapy is still a major challenge. In the last few years, extensive efforts have been made to incorporate diverse molecular information for better prognosis and treatment plans[2,3]. However, due to the high cost of large-scale molecular profiling, in practice clinicians are mainly focusing on a small number of selected genes or are using only single-platform genomic data. Therefore, with our study we want to understand how and to what extent different molecular profiling data can be useful in cancer diagnosis and prognosis. Using miRNA and mRNA expression, somatic copy-number variation, DNA methylation and somatic mutation profiles we have identified genes that are frequently altered in each of the selected cancers and are linked to patient survival. Some of the biological markers that we identified have already been reported in previous studies, but few of them are yet to be examined. In addition, we assessed which of the molecular dataset, as a standalone platform is the most informative for patient diagnostic and survival prediction.

Results

Molecular signatures for discrimination between normal and cancer tissues

First, we were interested in finding molecular signatures that can discriminate neoplastic from normal tissue in the selected cancer cohorts. For this purpose, we used a classification approach based on LASSO regression model[4]. In this analysis. only molecular data from normal tissue that is adjacent to primary tumor was used; the molecular data from blood derived normal tissue was not considered in order to avoid building models based on genes that can discriminate between blood and the corresponding solid tissue (lung, kidney or head/neck). The classification performance of the selected models was measured using the AUC ("Area Under Curve") statistic, which can be interpreted as a probability that the classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example[5]. The AUCs values of the selected models for discrimination between normal and cancer populations range from 0.95 – 1.00 (see Table 1). Almost perfect performance can be reached easily, which suggest that there are radical molecular changes in

cancerous cells compared to normal cells. Interestingly, the best (and perfect) classifier performance was achieved based on DNA methylation data for the LUAD and KIRC cohorts (Table 1). It is a well-known fact that DNA methylation can alter the expression of genes and several recent studies have shown that it also plays a crucial role in the development of nearly all types of cancer[6,7]. In the HNSC cohort, miRNA and mRNA expression data had equal performance with DNA methylation data in discriminating between normal and cancer tissue. With CNV data, we observed the worst performance in each cancer cohort.

Building a model that can discriminate whether a sample comes from a tumor that will go into remission or from one that will progress until the donor's death has proven to be a much more difficult task. For this task the above approach based on LASSO regression gave poor prognostic results (AUC values in range 0.5 – 0.76).

Cancer Type	Analyzed Data	AUC	Number of Selected Features
Lung Adenocarcinoma	miRNA expression	0.98	16
	mRNA expression	0.99	24
	CNV	0.95	64
	DNA Methylation	1.00	30
Kidney Renal Clear Cell Carcinoma	miRNA expression	0.97	12
	mRNA expression	0.98	36
	CNV	0.98	76
	DNA Methylation	1	120
Head and Neck Squamous Cell Carcinoma	miRNA expression	0.99	29
	mRNA expression	0.99	33
	CNV	0.93	66
	DNA Methylation	0.99	23

Table 1. Classification performance of the supervised learning models for discrimination between normal and cancer tissues

Molecular biomarkers associated with overall patient survival

To identify molecular signatures linked to patient survival for each cancer cohort, we asked whether low or high levels of a particular measured entity (expression, CNV or methylation) are significantly correlated with patients overall survival. In particular, in each cancer cohort, for a given miRNA, mRNA, protein, CNV and methylation probe, we separated the patients into quartiles based on the measured levels of the particular entity (miRNA/mRNA/protein expression, CNV or methylation values respectively). Then, using a log-rank statistical test we compared the overall survival of the patient group characterized by low levels of the particular measured entity (ie. values below the first quartile) to the survival of the patient group with high levels of that particular measured entity (values above the third quartile) (see Figure 1). The patients were split into training and validation sets and all statistical tests were conveyed on the training datasets. Based on this "quartile" approach, we could identify miRNAs, protein-coding genes, CNV and methylation probes whose extreme measured values were statistically linked to overall patients survival (p-value of log-rank test < 0.05). For further analyses, we kept only those that were significantly associated to the overall survival also in the validation dataset. Next, in each molecular dataset we clustered the selected genes/probes from the "quartile" test using non-negative matrix factorization[8] and selected best representatives from each cluster. To build prognosis models for each molecular dataset and each cancer cohort, we performed a multivariate Cox regression[9] on the selected genes/probes. For each signature, coefficients from a multivariate Cox regression analysis on the training cohort were used to compute a risk on the validation cohort. The accuracy of the prognosis methods was assessed through a concordance index, which is a non-parametric measure that quantifies the fraction of pairs of patients whose predicted survival times are correctly ordered among all pairs that can actually be ordered[10]. The best performing models for each cancer cohort are shown in Table 2. Using only 3 or 4 genes/probes from each molecular dataset, we could achieve concordance correlation coefficient greater than 0.7 in the validation cohorts (see the red bars on Figure 2).

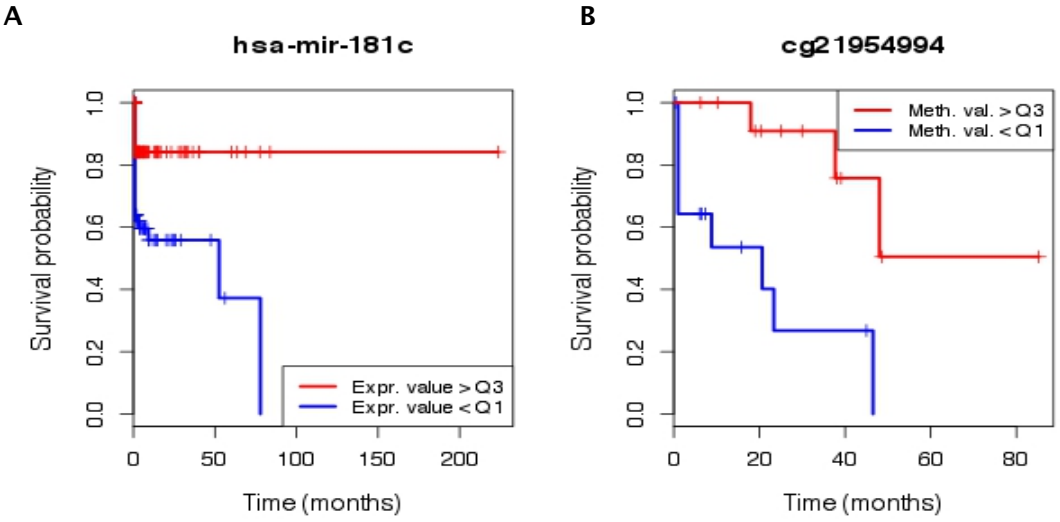


Figure 1. Quartile-based selection of features associated to overall survival. A) Differences in the survival probability between patients with high expression values of "hsa-mir-181c" (>Q3) and patients with low expression values (<Q1). B) Differences in the survival probability between patients with high methylation values (>Q3) of the "cg21954994" methylation probe and patients with low methylation values (<Q1).

Cancer Type	Analyzed Data	Survival Concordance Index	Molecular Signatures for Survival Prognosis
Lung Adenocarcinoma	miRNA expression	0.70	hsa-mir-23b; hsa-mir-181c; hsa-mir-1976
	mRNA expression	0.71	ATP8A2; FOXM1; LCN10
	CNV	0.57	HP1BP3; MLLT3; GDPD3; RP11-778D9.13
	Methylation	0.73	cg06602857; cg21954994; cg19213569
Kidney Renal Clear Cell Carcinoma	miRNA expression	0.72	hsa-mir-21; hsa-mir-183; hsa-mir-3942; hsa-let-7b
	mRNA expression	0.69	BARX1; ITPKA; NKX2-5
	CNV	0.65	IFNA5; CDKN2A; RP11-399D6.2
	Methylation	0.77	cg09635053; cg14898260; cg23368159
Head and Neck Squamous Cell Carcinoma	miRNA expression	0.68	hsa-mir-520g; hsa-mir-29b-1; hsa-mir-144; hsa-mir-137
	mRNA expression	0.64	AQP5; CAMKV; SNAP25
	CNV	0.52	RP11-419C19.2; HOXD3; BRIX1
	Methylation	0.67	cg14526044; cg15716405; cg17720011; cg12042587

Table 2. Molecular signatures for cancer survival prognosis and their performance on the validation datasets for each cancer cohort.

Next, we wanted to test whether the molecular profiles that are distinctive for normal and cancer tissues are also correlated with patient survival. Using the selected genes/probes from the normal vs cancer tissue classification, we built multivariate Cox regression prognostic models and assessed their prediction performance through a concordance index (green bars on Figure 2). Our results show that even though one can well discriminate between normal and cancer tissues using selected features, the same features are not necessarily good survival predictors. In fact, only very few genes selected from the normal vs cancer classification appear to be predictive for survival. For example, the miRNA "hsa-mir-21", an "oncomir" associated with a wide variety of cancers[11], is predictive for survival in KIRC cohort, but it is also selected as a discriminatory feature in the normal vs cancer tissue prediction in the KIRC and LUAD cohorts.

To further assess the power of our selected molecular signatures, we built multivariate Cox regression prognostic models using randomly selected genes/probes. Figure 2 shows that our prognostic markers selected from the different molecular datasets (miRNA, mRNA, CNV, methylation) are largely superior to randomly chosen genes/probes in the three cancer cohorts.

We extended the analyses to include survival prediction based on somatic mutations profiles (SNP data), which we obtained from the TCGA Data Portal. For each gene we split the patients into two groups: patients having a somatic mutation in that particular gene, and patients with no somatic

mutations in that gene. If the difference in survival between the two patient groups is significant ($p<0.01$), we included the corresponding gene in the multivariate Cox model. Again we split the set of patients on training and validation sets. The Cox model built on the training set was used to predict the survival on the validation dataset. For each particular gene, we required that at least 10 patients have a mutation in that gene. The survival prognosis signatures from SNPs data were superior over the signatures from the other datasets in LUAD and HNSC cohorts. Only in the KIRC cohort the signature from the methylation data gave the best performance. Next, we integrated the prediction signatures from the different "-omics" data together with clinical variables (donor age, sex and donor icd10 diagnosis) to build a "multi-omics" Cox survival prediction model. The addition of variables into the model was assessed through a forward model selection procedure (Aikake information criterion) combined with a Cox regression. However, the prognostic performance of this "multi-omics" prediction model has not improved.

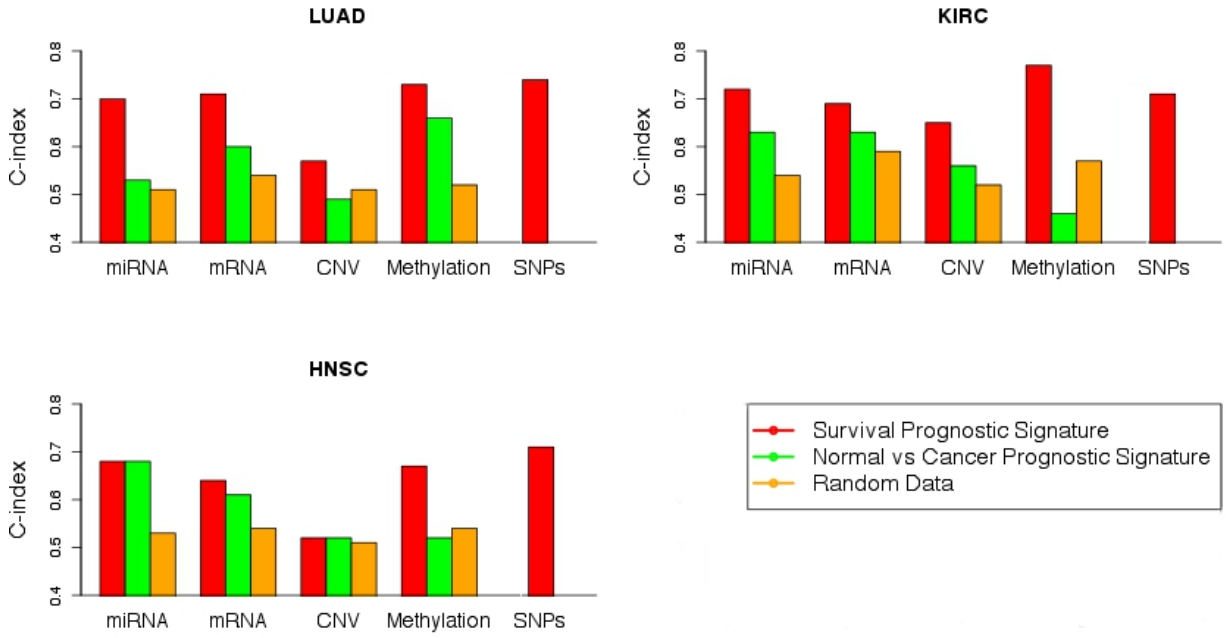


Figure 2. Performance assessment of several prognosis signatures on the validation datasets in A) Lung Adenocarcinoma, B) Kidney Renal Clear Cell Carcinoma and C) Head and Neck Squamous Cell Carcinoma. Red: Survival prognostic using molecular signatures listed in Table 2. Green: Survival prognostic using molecular signatures from normal-cancer classification. Orange: Survival prognostic using randomly chosen molecular data.

Discussion

In this work we evaluated patient survival prediction from different molecular data types and described potential prognostic signatures across three cancer types. Currently, only a few gene expression signatures are routinely used in the clinical practice for these three cancers[12]. In LUAD and HNSC cancer cohorts, somatic mutation profiles (SNP data) appear to be the most informative resources for prognostics, while DNA methylation profiles are the most informative in the KIRC cohort. Using a quartile-based selection we identified features that are prognostic for at least a subset of patients. This approach inherently supports heterogeneity, in contrast to classification methods. Some of the prognostic signatures that we identified are well studied in the literature: eg. the FOXM1 gene has been shown to promote tumor metastasis in non-small cell lung cancer patients and is associated with chemotherapy resistance[13,14]. But we also identified prognostic signatures that have not been reported as linked to cancer progression. For example, the gene ATP8A2, member of aminophospholipid transporter family, is associated with several diseases, but not with cancer. However, another gene from the same family, ATP11A, was recently identified as a predictive marker for metastasis in colorectal cancer[15].

The fact that we can relatively easily discriminate normal from tumor tissue suggests that cancer consistently alters the molecular machinery. However, cancer malignancy is heterogeneously

defined within cancer type, and as a consequence molecular signatures do not perfectly predict survival. Different molecular data types have different predictive values in cancer types, which suggests that cancer malignancy relies on different mechanisms across cancers. Our analyses do not necessarily identify the cancer causal changes; they rather identify molecular markers that are affected by causal changes and are associated with survival. They offer new prospects for further investigations of cancer pathogenesis.

Methods

Data

We used preprocessed mRNA expression (mRNA-seq), miRNA expression, protein expression, somatic CNV (all them downloaded from the ICGC Data Portal, release 17) and DNA methylation data (ICGC, release 18). The LUAD dataset contains molecular profiles of 473 patients, KIRC dataset contains molecular profiles of 515 patients, and HNSC 422 patients. The data comes from 3 tissue types: primary tumor solid tissue, normal tissue adjacent to primary and normal blood derived tissue. Expression data are the most commonly and consistently available ICGC data type. Training and validation sets were created from each cancer cohort in a ratio 2:1, meaning that two-thirds of the corresponding data set was used for building the models and one-third of it for validating the models. No bias in tumor stage, age, overall survival, or gender distribution was observed between the training and validation sets.

Identification of prognostic signatures

For each molecular profile (i.e. for each miRNA, mRNA and protein) in the training dataset two groups of patients were constructed based on expression levels of the miRNA, mRNA or protein respectively: lower than the 25% quartile and higher than the 75% quartile. A log-rank test was then applied to determine if the difference in terms of overall survival between the two groups was significant (p-value < 0.05). Clustering of significant survival-associated genes (probes) was performed through a non-negative matrix factorization (NMF) with ranks tested from 2 to 6. Representative genes (probes) for each cluster were selected based on their basis coefficient. All possible combinations of representative genes (probes), such that to have only one representative per cluster, were tested to obtain the signature. A multivariate Cox regression analysis on miRNA expression values was used to compute a risk for each combination. For each signature, coefficients from a multivariate Cox regression analysis on the training cohort were used to compute a risk on the validation cohort. Performance was assessed through a concordance index (c-index). To test the significance of a particular molecular signature, we selected random genes (probes) from the ICGC datasets and trained a Cox model using these genes (probes). The number of the randomly selected genes in each test was equal to the size of the particular molecular signature. Sampling was performed over 1000 iterations to obtain an average C-index and its standard deviation.

References

1. Vogelstein B et al.. (2013). Cancer Genome Landscapes. Science. vol. 339 no. 6127 pp. 1546-1558
2. Gerlinger M et al.. (2014) Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. Nat. Genet. 2014/02/04 ed. Nature Publishing Group; 2014; 46:1–12.
3. The Cancer Genome Atlas Research Network. 2014. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511, 543–550
4. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., Vol. 58, No. 1
5. Fawcett, Tom (2006). "An Introduction to ROC Analysis". Pattern Recognition Letters 27 (8): 861–874.
6. Craig, JM; Wong, NC (editor) (2011). Epigenetics: A Reference Manual. Caister Academic Press. ISBN 978-1-904455-88-2.
7. Kulis M1, Esteller M. 2010. DNA methylation and cancer. Adv Genet. 70:27-56.
8. Brunet, J-P et al. (2004). Metagenes and molecular pattern discovery using matrix factorization. PNAS. USA 101(12)
9. Cox, D. R.; Oakes, D. (1984). Analysis of Survival Data. New York: Chapman & Hall. ISBN 041224490X
10. Lawrence I-Kuei Lin (1989). "A concordance correlation coefficient to evaluate reproducibility". Biometrics (International Biometric Society) 45 (1): 255–268.
11. Zheng J, Xue H, Wang T et al. (2011). "miR-21 downregulates the tumor suppressor P12(CDK2AP1) and Stimulates Cell Proliferation and Invasion". J. Cell. Biochem. 112 (3): 872–80.
12. Yuan Y et al. 2014. Assessing the clinical utility of cancer genomic and proteomic data across tumor types.
13. Nuo Xu e al. 2013. FoxM1 Is Associated with Poor Prognosis of Non-Small Cell Lung Cancer Patients through Promoting Tumor Metastasis. Plos ONE. DOI: 10.1371/journal.pone.0059412
14. Wang et al. 2013. FoxM1 expression is significantly associated with cisplatin-based chemotherapy resistance and poor prognosis in advanced non-small cell lung cancer patients. Lung Cancer. 2013 Feb;79(2):173-9
15. Miyoshi, N.et al. ATP11A is a novel predictive marker for metachronous metastasis of colorectal cancer. Oncol. Rep 2010, 23, 505–510.

Functional hallmarks in clear cell renal cell carcinoma grade and stage progression revealed by changes in signalling circuit activities.

Marta R. Hidalgo[1.2], Cankut Cubuk[1], Jose Carbonell-Caballero[1], and Joaquín Dopazo[1,2,3]

1. Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain.
2. Functional Genomics Node, (INB) at CIPF, Valencia, Spain.
3. Bioinformatics of Rare Diseases (BIER), CIBER de Enfermedades Raras (CIBERER), Valencia, Spain.

Abstract

The acquisition of the cancer phenotype is a process largely dominated by changes in cell signalling that can hardly be interpreted as the consequence of isolated changes in gene activity but rather as the results of complex interactions among these. Here we propose the use of a transformation of individual gene expression data into numerical descriptors of signalling pathway activities that are further used to understand the evolution of the disease across the different tumour grades. We have studied the clear cell renal cell carcinoma (ccRCC) data from the ICGC Cancer Genome Consortium Challenge.

Introduction

Complex traits, including most diseases, are associated with complex changes in biological pathways rather than being the direct consequence of single gene alterations. In particular, the hallmarks of cancer, which include sustaining proliferative signalling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis [1] are all directly or indirectly related to pathologically altered signalling processes. The idea of using the information contained in different biological pathways to understand complex traits, such as diseases, is recently gaining acceptance [2]. Signaling pathways provide a formal representation of the processes by which the cell triggers actions in response to stimulus through a network of intermediate gene products that configure signaling circuits. Interestingly, such circuits can directly be related to cell functionalities. Recently some methods have developed that focus particularly on the estimation of the activity of these stimulus-response signaling circuits from gene expression data [3, 4]. Here we show how to use gene expression values in the context of signaling circuits to understand the molecular mechanisms underlying the evolution of tumor grade and tumor stage.

Method

We evaluated the pathological signal transduction changes in ccRCC by analysing the TCGA ccRCC samples (https://dcc.icgc.org/repository/release_18/Projects/KIRC-US) [5] over a set of selected previously cancer related pathways that includes PI(3)K/AKT and mTOR signalling pathways taken from KEGG. The pathways are decomposed into elementary signalling circuits that connect receptor proteins with effector proteins, whose mission in the cell is triggering functional responses to the stimuli received by the receptors (see [4] for details). Activation-inactivation relationships between nodes (proteins) along the circuits enabled us to use a graph traversal methodology for updating signal intensity at each visited node and finally computing a global value of signal transduction for the circuit (thereinafter signalling circuit activity or SCA).
Both tumour grade (TG) and tumour stage (TS) status per sample were obtained from clinical data from the ccRCC page. Patients were stratified according to their status (TG and TS) and normal samples were grouped into a single initial state (s0). Then, a chronogram with the precise sequence of pathologic events that occurs after reaching each tumour status was reconstructed by comparing SCA at each stage or grade against all the precedent ones (eg. G3 against G2, G1 and G0). Here we focused only into monotonically increasing or decreasing behaviours and only significant differences were reported.

Results and discussion

When samples are clustered on the basis of their SCA patterns a clear separation between cases and controls is observed (Figure 1) which provides an initial evidence of the relationship of these values to the biological progression of cancer.

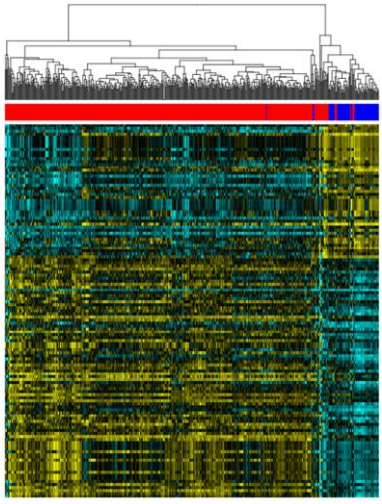


Figure 1. *Hierarchical clustering of SCA values. Normal samples are coloured in blue and cases in red*

When the SCA values are compared across tumour developmental phases, several systematic activations or deactivations of signalling circuits across TGs and TSs is observed.

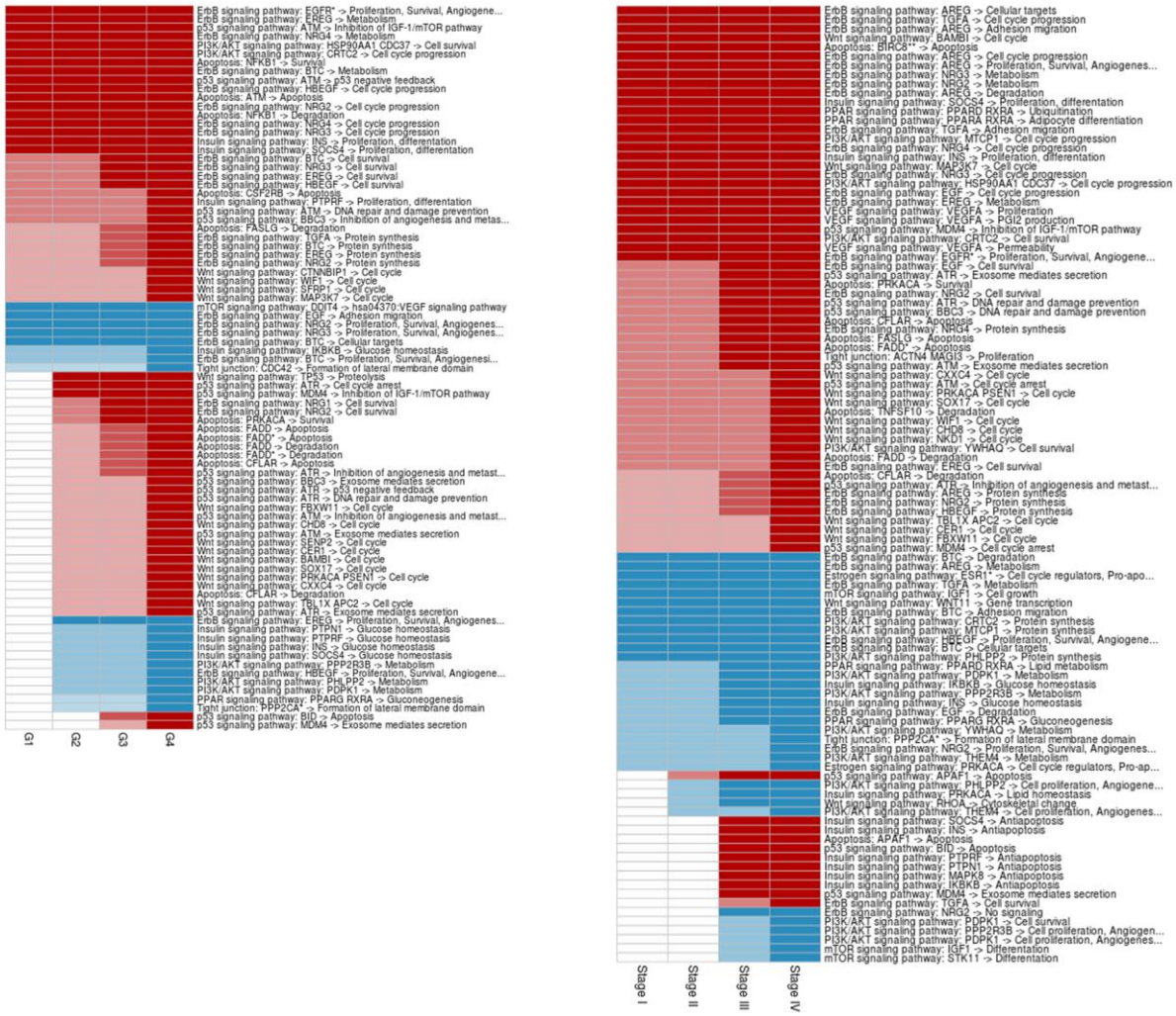


Figure 2*: Evolution of SCA values corresponding to different circuits with progressive behaviour. Upregulated circuit/grades (left) and circuit/stages (right) are coloured in red, and downregulated in blue. The different red/blue intensities describe an increase in up/down regulation in late cancer phases.*

Interestingly, activated and deactivated functions triggered by signalling circuits have a direct relationship to cancer progression. Thus, biological processes such as *cell cycle*, *survival*, *angiogenesis*, *proliferation, antiapoptosis* or *cell survival* are systematically activated as TG and TS progresses. On the other hand, *protein synthesis*, *metabolism*, *glucose homeostasis* and, in general, *differentiation* processes are inhibited, as expected from the indiferentiation process that occurs in cancer. Other functions, like *cell adhesion* are also deactivated, favouring thus invasion and metastasis.

Conclusions

We concluded that gene expression data can be transformed into measurements of SCA values that account for cell functionalities. Such measurements cell functionalities can be related to cancer progression, in particular TG and TS. The cancer hallmarks already described [1] can be considered the consequences of a series of functional hallmark that are elegantly described in the approach proposed here.

We propose that approaches that model cell functionalities will be not only more accurate in predicting phenotypic traits, such as the disease progression, but will also provide insights into the molecular mechanisms that account for such phenotype.

References

1. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation.** *Cell* 2011, **144:**646-674.
2. Davis MJ, Ragan MA: **Understanding cellular function and disease with comparative pathway analysis.** *Genome Med* 2013, **5:**64.
3. Sebastian-Leon P, Carbonell J, Salavert F, Sanchez R, Medina I, Dopazo J: **Inferring the functional effect of gene expression changes in signaling pathways.** *Nucleic Acids Res* 2013, **41:**W213-217.
4. Sebastian-Leon P, Vidal E, Minguez P, Conesa A, Tarazona S, Amadoz A, Armero C, Salavert F, Vidal-Puig A, Montaner D, Dopazo J: **Understanding disease mechanisms with models of signaling pathway activities.** *BMC Syst Biol* 2014, **8:**121.
5. CancerGenomeAtlasResearchNetwork: **Comprehensive molecular characterization of clear cell renal cell carcinoma.** *Nature* 2013, **499:**43-49.

Cancer Progression Classification for Mutation Analysis

Jari Björne and Tapio Salakoski

Department of Information Technology, University of Turku
Turku Centre for Computer Science (TUCS)
Joukahaisenkatu 3-5, 20520 Turku, Finland
firstname.lastname@utu.fi

1 Introduction

The International Cancer Genome Consortium is a global project aiming to produce a description of the genomic, transcriptomic and epigenomic changes in 50 different cancer types. The current 18th release of the ICGC data provides varied biochemical analyses for a set of 12,807 cancer patients. The CAMDA conference concerns the development of computational approaches for analysis of large-scale biomedical datasets. The ICGC dataset has been used in the shared "CAMDA challenge" first in 2014 and now in 2015. Having participated in the 2014 challenge we now extend our work of analysing cancer datasets through functional classification followed by analysis of the genetic basis behind the classification. In our 2014 entry we studied the ICGC cancers as separate datasets and evaluated various approaches for feature selection [Björne et al., 2014]. In the current work we expand our analysis to the whole ICGC cross-cancer dataset. With the increased scale of the data, we are able to utilize the somatic mutation data that illuminates the underlying causes of the cancer. With the effective embedded feature analysis of an ensemble classifier we can evaluate the type of mutations that affect the clinical outcome of a patient's cancer. Comparison with the COSMIC cancer gene census shows that genes central in causing cancer are also central for predicting its progression.

2 Materials and Methods

2.1 The ICGC cancer data

We use the publicly available parts of release 18 of the ICGC project. Files for the 55 cancer projects were downloaded from the ICGC data portal[1] [Zhang et al., 2011]. For use in experiments the TSV-formatted files were converted into an SQLite database approximately 30 Gb in size.

2.2 COSMIC cancer gene census

The COSMIC database (Catalogue Of Somatic Mutations In Cancer) is a collection of somatic mutations present in cancers, developed by the Wellcome Trust Sanger Institute[2]. The COSMIC cancer gene census is a list of genes known to be causally implicated in cancer [Futreal et al., 2004]. It thus represents a conservative set of the most strongly cancer related genes. The version of the census used in these experiments was downloaded on May 18th 2015 and consists of 572 genes.

2.3 Machine learning methods

For machine learning we use the scikit-learn library, version 0.16.1. We perform binary classification, using the Linear SVC (support vector machine) and Extra Trees classifiers [Geurts et al., 2006]. In the current 0.16.1 release of scikit-learn both of these methods support sparse matrices allowing efficient processing of large data sets.

[1] https://dcc.icgc.org
[2] http://www.sanger.ac.uk/cosmic

For estimating classification performance we use the scikit-learn implementation of the AUC-metric (area under the ROC curve). The AUC is a robust and largely class-distribution independent performance measure, whose results are in the range 0.5 (completely random) to 1.0 (perfect classification).

3 Experimental Setup

3.1 Division of Data

In performing classification experiments we optimize parameters using five-fold cross-validation on a *training* dataset. Final results are produced on a separate *hidden* dataset left aside for this purpose. We divide ICGC cancer samples by patient into training and hidden sets in a 7:3 ratio. The sets are divided on a pseudorandom distribution seeded with the *ICGC donor id*, ensuring that the same patient always belongs to either the training or the hidden set regardless of the selection of patients for a particular experiment.

3.2 Classification

Our goal is to develop a classification system for predicting the prognosis of a patient's cancer based on the available biochemical data. The prognosis is of interest as a classification task in itself, but also as a preliminary step for the feature analysis that aims to uncover the genetic basis of the prognosis.

The primary classification we perform is the division of the ICGC cancers that go into "complete remission" (disappearance of all signs of cancer) vs. those that progress to the death of the patient. These represent the two, opposite end-points for a cancer patient. This division follows our per-cancer classification task from our 2014 entry, and applied for the whole ICGC dataset, for samples with SSM (simple somatic mutation) data, this division results in a set of 3491 examples for complete remission and 1307 examples for progression until death.

In optimizing the parameters powers of ten in the range -10 to 10 are evaluated for the C-parameter of the SVM and values 10, 100 and 1000 are tested for the number of trees in the Extra Trees Classifier. In previous experiments we have seen performance increases when using up to 10,000 trees with ensemble methods but due to the large size of the cross-cancer datasets this is not feasible in the current experiments.

3.3 Features

Unlike our entry from 2014, in this work we use only one ICGC data type per classification experiment. This is primarily due to the number of cases where only some of the data types are available for a patient. For example, the SSM data is available for 7,908 patients whereas the EXP-A data is available for only 3,135 patients.

The primary data type used in our experiments is the SSM (simple somatic mutations). When generating features based on SSM the primary challenge is the sparsity of the data. Even the most common SSM in the ICGC data, MU62030 (a single base A>T substitution in the gene BRAF), occurs in only 405 donors across all the ICGC projects. Therefore, individual mutations have to be grouped if they are to be used as machine learning features. We first experimented with simply grouping all mutations within one gene, that is, we used simply the binary mutation status of a gene as a feature. While reaching decent classification performance, such features are not very interesting for the analysis of the mutational basis behind a certain classification. Therefore, in the final experiments we grouped mutations both by gene and by their functional impact on that gene (e.g. exon variant, intron variant, missense etc). This feature type mostly preserved the classification performance of gene-level features, while providing a more interesting feature set for the subsequent mutation analysis.

As a point of comparison for the SSM mutations we tested gene expression levels. Gene expression is commonly used a sort of "fingerprint" for the phenotype of a particular cancer. In a machine learning context gene expression levels are easier to work with than the SSM, as at least some value is present for each gene in each analysed sample, but the expression features are of course "one step removed" from the underlying genetic causes of the cancer. For the expression data we chose the sequencing based expression (EXP-S) as that is more commonly available for the ICGC samples than the array based expression (EXP-A).

3.4 Feature Analysis

The feature analysis is based on the embedded feature importance ranking provided by the Extra Trees Classifier [Breiman et al.]. In ensemble methods the relative rank (depth) of a feature contained in a decision tree can be used as a measure of the importance of that feature in performing the classification. In our 2014 CAMDA entry we have shown that compared with e.g. greedy forward selection and recursive feature elimination the embedded feature importance estimation results in relatively stable performance progression. To determine the relevance of the selected features we compare them against the genes in the COSMIC Cancer Gene Census.

4 Results and Discussion

4.1 Classification Performance

The classification results are shown in Table 1. The primary feature set of SSM results in a decent performance of slightly above 0.7 AUC. Both the support vector machine (SVM) as well as the extra trees classifier (ETC) provide similar performance. Unlike in our 2014 entry using the ETC does not result in higher performance, perhaps due to the larger class sizes. SSM-based classification with the SVM is generally faster and has slightly higher performance, but does not provide the embedded feature analysis. With both the SSM and EXP-S feature sets we observe a notable increase in performance compared with the five-fold cross-validation of the training set and the final classification of the hidden set. We speculate this may be due to the size of the datasets and the additional 20% of training data available when classifying the hidden set. A learning curve experiment should be done in the future to evaluate this assumption.

4.2 Feature Analysis

As seen in Figure 1 known cancer genes are more common among the features selected as most important. This alludes to some biological relevance behind the automatically learned classification and the automatic selection of features.

Table 2 shows the top 20 features from the feature importance ranking produced with the extra trees classifier for the SSM feature set. The most important feature turns out to be any mutation in an intergenic region. As such a feature has no gene name it becomes very common and is possibly slightly correlated with one of the two classes.

The more interesting features are those generated for mutations within known genes, as they define both a gene name and a functional consequence (depending on the mutated site). The top three genes, EGFR, KRAS and TERT are traditional, well known cancer genes. Mutations in the *epidermal growth factor receptor* (EGFR or ErbB-1) can lead to uncontrolled cell division and as such it is a central gene in a number of cancers [Lynch et al., 2004]. The *V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog* (KRAS) is a regulator of growth-related signaling and its mutation is essential for the growth of many tumours [Kranenburg, 2005]. Mutations in *Telomerase reverse transcriptase* allow telomerase to remain active in somatic cells and thus leads to immortal cancer cells [Zhang et al., 1999].

Table 1: Classification performance. Example counts are shown for the two classes followed by the AUC_T and AUC_H scores for the (t)raining and (h)idden sets.

features	classifier	remission	progression	AUC_T	AUC_H
SSM	Extra Trees	3491	1307	0.612 ± 0.056	0.704
SSM	Linear SVC	3491	1307	0.611 ± 0.030	0.722
EXP-S	Extra Trees	4592	1210	0.565 ± 0.033	0.819
EXP-S	Linear SVC	4592	1210	0.602 ± 0.044	0.758

Table 2: The most important features. Each feature is the id of the gene combined with the mutation consequence. The census column indicates whether the gene is among the known cancer genes in the COSMIC census.

#	gene id	gene name	consequence	census
1			intergenic region	
2	ENSG00000146648	EGFR	missense variant	•
3	ENSG00000133703	KRAS	missense variant	•
4	ENSG00000164362	TERT	upstream gene variant	•
5	ENSG00000121879	PIK3CA	missense variant	•
6	ENSG00000175826	CTDNEP1	stop gained	
7	ENSG00000023516	AKAP11	missense variant	
8	ENSG00000187172	BAGE2	intron variant	
9	ENSG00000141510	TP53	intron variant	•
10	ENSG00000169031	COL4A3	intron variant	
11	ENSG00000096968	JAK2	intron variant	•
12	ENSG00000182185	RAD51B	intron variant	
13	ENSG00000149531	FRG1B	stop gained	
14	ENSG00000141510	TP53	exon variant	•
15	ENSG00000115896	PLCL1	intron variant	
16	ENSG00000210154	MT-TD	downstream gene variant	
17	ENSG00000174473	GALNTL6	intron variant	
18	ENSG00000229981	LINC01435	intron variant	
19	ENSG00000130226	DPP6	intron variant	
20	ENSG00000140945	CDH13	intron variant	

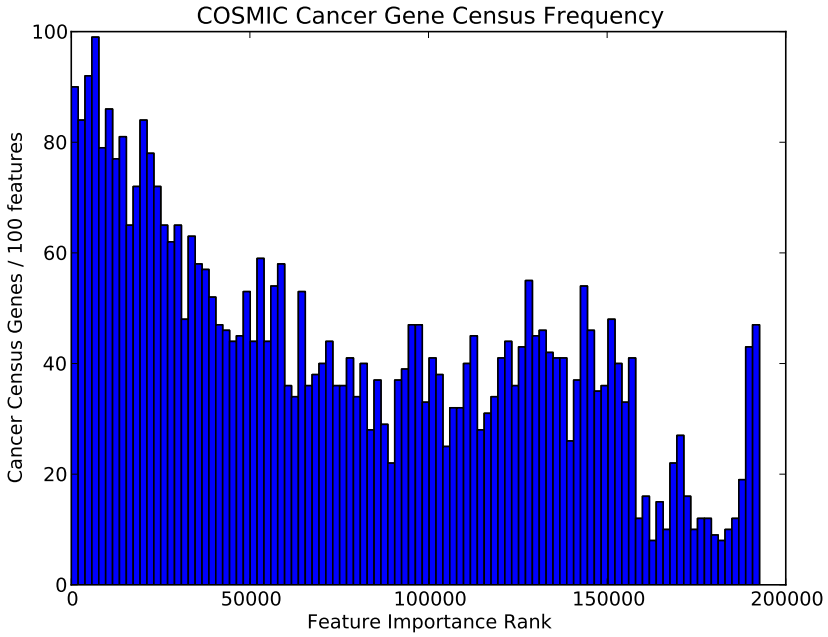


Figure 1: Known cancer genes among the selected features. Known cancer genes from the COSMIC census are more common among the features automatically selected for classifying cancer progression.

5 Conclusions

We have extended our classification-based cancer analysis approach to the entire ICGC cross-cancer dataset. With the increased size of the dataset, sparse feature groups such as the SSM become usable as classification features, and can achieve decent classification performance for predicting cancer progression. The SSM represents the most causally relevant feature set for understanding the nature of the ICGC cancers, as these individual mutations form the driving force of many tumours. While analysis of the feature selection results shows a correlation with known cancer genes, more work is needed to evaluate the role of the less known mutations.

The classification into complete remission or progression until death is a classification where all examples are positive for being cancers. However, common cancer genes present in the COSMIC census rank highly as features relevant for predicting the progression of cancer. We speculate that genes commonly mutated in cancer are also among the strongest drivers of cancerous growth, making them good indicators for the severity of progression, with mutations in several such genes being more likely to result in a fatal cancer.

As future work we hope to find ways to better utilize the mutation data on the level of individual mutations, to provide the kind of analysis required for the current biomedical research of separating the important driver mutations from the random passenger ones. As with our earlier project, we will publish all of our experimental code under an open source license[3].

References

J. Björne, A. Airola, T. Pahikkala, and T. Salakoski. Classification and feature selection across the ICGC Cancer Projects in the CAMDA 2014 Challenge. 2014.

L. Breiman, J. Friedman, and R. Olshen. Stone, cj (1984) classification and regression trees. *Wadsworth, Belmont, California*.

P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–183, 2004.

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

O. Kranenburg. The kras oncogene: past, present, and future. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1756(2):81–82, 2005.

T. J. Lynch, D. W. Bell, R. Sordella, S. Gurubhagavatula, R. A. Okimoto, B. W. Brannigan, P. L. Harris, S. M. Haserlat, J. G. Supko, F. G. Haluska, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. *New England Journal of Medicine*, 350(21):2129–2139, 2004.

J. Zhang, J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty, M. Wong-Erasmus, L. Yao, and A. Kasprzyk. International cancer genome consortium data portal – a one-stop shop for cancer genomics data. *Database*, 2011, 2011.

X. Zhang, V. Mar, W. Zhou, L. Harrington, and M. O. Robinson. Telomere shortening and apoptosis in telomerase-inhibited human tumor cells. *Genes & development*, 13(18):2388–2399, 1999.

[3] `https://github.com/jbjorne/CAMDA2015`

Exploring the Importance of Cancer Pathways by Meta-Analysis of Differential Protein Expression Networks in Three Different Cancers

Sinjini Sikdar, Somnath Datta, Susmita Datta[1]

Department of Bioinformatics and Biostatistics, University of Louisville, KY 40202, USA

1 INTRODUCTION

Landscape of most cancers involve twelve important pathways ("target pathways") that regulate three core cellular processes "cell fate", "cell survival" and "genome maintenance"; the "driver" genes, which are responsible for the formation of tumors, function through these signaling pathways [1]. We undertake a novel investigation of the roles of these pathways using a differential network analysis of the protein expression datasets on three cancers (Head and Neck Squamous Cell Carcinoma, Lung Adenocarcinoma and Kidney Renal Clear Cell Carcinoma). These datasets were available to us from International Cancer Genomic Consortium (ICGC) as part of the CAMDA 2015 challenge data. We pursue a meta-analysis of protein expressions to investigate whether each of these target pathways plays a significant role in these three cancers in the sense that the proteins associated in these pathways interact differently between two clinical groups ("progression" or "complete remission") of patients. From our analysis of the protein expression data, overall, RAS and PI3K signaling pathways appear to play the most significant roles in these three cancers. This analysis suggests that these two signaling pathways should be investigated further for their roles in cancers. It is interesting to note that these two main pathways are related to "cell survival" function.

2 DATASETS

We have analyzed the preprocessed challenge datasets for CAMDA 2015 provided by the International Cancer Genomic Consortium (ICGC). For our study, we have considered the protein expression and the clinical profiles of the patients for three cancers, Head and Neck Squamous Cell Carcinoma (HNSC), Lung Adenocarcinoma (LUAD), and Kidney Renal Clear Cell Carcinoma (KIRC). A set of 132 proteins is present in the protein expression profiles of each of the three cancers; the patient sample sizes of HNSC, LUAD and KIRC were 212, 237 and 454 patients, respectively. The clinical profile of each of the cancer type represents the disease status (progression or complete remission) of each patient. In summary, we have two groups of patients for each cancer type and the set of recorded protein expression values of 132 proteins on each of them.

[1] to whom correspondence should be addressed (susmita.datta@louisville.edu)

3 METHODOLOGY

3.1 Pathway analysis: From a recent study [1], it has been found that there are 140 "driver" genes/proteins which can promote the formation of tumors if affected by intragenic mutations. These "driver" genes can be classified into twelve signaling pathways which are: TGF − β, MAPK, STAT, PI3K, RAS, Cell Cycle/Apoptosis, NOTCH, HH, APC, Chromatin modification, Transcriptional regulation and DNA damage control. Among these, TGF − β, MAPK, STAT, PI3K, RAS and Cell Cycle/Apoptosis regulate "cell survival"; NOTCH, HH, APC, Chromatin modification and Transcriptional regulation regulate "cell fate"; while the DNA damage control signaling pathway regulates "genome maintenance". We refer to these twelve signaling pathways as "target pathways".

We separately analyze the protein profiles of the three cancer types using "GO" clustering tool [2, 3], and group the proteins according to their biological pathways. Out of the pathways obtained, we only considered the proteins included in the "target pathways" for our analysis.

3.2 Differential network analysis: In order to identify whether the network structures of the "target pathways" have changed from the complete remission group to the progression group, we performed differential network analysis [4] using the R package *dna* [5]. This differential network analysis for each pathway is conducted based on connectivity scores between the proteins in these target pathways. Initially, to get an idea about the network structures in each of the two groups, graphical networks are constructed by connecting each pair of proteins for which the connectivity scores exceed a threshold. The difference in connectivity between the two groups (progression versus complete remission) is computed mathematically, using the following statistic:

$$\Delta(\mathcal{F}) = \frac{1}{k(k-1)} \sum_{p \neq p' \in \mathcal{F}} \left| s_{pp'}^{pr} - s_{pp'}^{cr} \right|, \qquad (1)$$

where $\mathcal{F}$ denotes the set of proteins present in a "target pathway" and k denotes the number of proteins in $\mathcal{F}$. Here $s_{pp'}^{pr}$ and $s_{pp'}^{cr}$ are the connectivity scores between the proteins p and p' in the progression and complete remission groups, respectively. For our analysis, the connectivity score of a protein pair in a network is taken to be the Pearson's correlation coefficients of the expression values of the two proteins in the corresponding sample data. A permutation test is then carried out using the test statistic $\Delta(\mathcal{F})$ and the corresponding observed level of significance (p-value) is obtained.

In addition to testing the overall pathway significance, we also test whether the connectivity of each single protein has changed between the two groups (progression versus complete remission) using the following statistic:

$$d(p) = \frac{1}{f-1} \sum_{p' \in \mathcal{G},\ p' \neq p} \left| s^{pr}_{pp'} - s^{cr}_{pp'} \right|, \qquad (2)$$

where $\mathcal{G}$ denotes the set of all proteins and f is the number of proteins in $\mathcal{G}$. Once again, a permutation test is carried out for each protein using the test statistic $d(p)$ and the corresponding p-value is obtained.

3.3 Rank Aggregation: The p-values, obtained using the test statistic given in (1), are used to obtain ranked lists of the "target pathways" for each cancer type. Here, ranking is done in such a way that the "target pathway" with the lowest p-value gets rank 1, the next one gets rank 2 and so on. Since, these ranked lists vary according to the cancer type; we need to aggregate them in a meaningful way to get an overall ranked list which would then rank the pathways by their global order of importance. In other words, this overall ranked list may provide us with the most important "target pathways" in all the three cancers. The R package RankAggreg [6], which is based on Cross-entropy Monte Carlo algorithm [7], is used to get this overall ranked list.

For our second analysis at the individual protein level, the p-values obtained using the test statistic given in (2), are used to rank the set of 132 individual proteins. An overall ranked list of these proteins is obtained using the R package RankAggreg [6].

4 RESULTS

We find representation of five out of twelve "target pathways" in our sample of 132 proteins; they are the PI3K signaling pathway, Cell Cycle, Apoptosis, RAS signaling pathway and MAPK signaling pathway. Based on our differential network analysis [4, 5] between the two groups of patients (progression vs complete remission) using the test statistic given in (1), with Pearson's correlation coefficients as scores and absolute distance measure carried out for each of the 3 cancer types, we have the following findings: the RAS signaling pathway is highly significant (p-value = 0.026) and MAPK signaling pathway is marginally significant (p-value = 0.082) in HNSC; for LUAD, PI3K signaling pathway is highly significant (p-value = 0.013). Table 1 shows the overall ordering of the 5 "target pathways" for the three cancers along with the rank aggregated list. Thus overall, the RAS signaling pathway appears to be most important followed by the PI3K signaling pathway, based on our meta-analysis of the available data on three cancers.

Table 1: Target pathways ordered by statistical significance (p-values) for each cancer type along with the overall ordering by rank aggregation.

Cancer Type	Pathway Ordering by p-values
HNSC	R, M, P, A, C
LUAD	P, C, A, R, M
KIRC	R, A, M, P, C
Overall	R, P, M, A, C

R: RAS Signaling pathway
M: MAPK Signaling pathway
P: PI3K Signaling pathway
A: Apoptosis
C: Cell Cycle

A graphical representation of the network structure of the proteins in the two groups of patients for RAS signaling pathway in HNSC is shown in Figure 1. In this figure, two proteins are connected if the connectivity score between them is significantly large. Different colors and shades in the figure represent positive or negative correlations and the thickness of the lines represents the strength of the associations. A visual inspection reveals some obvious differences in the network connectivity between the two groups of patients. Notably, GAB2, MAPK1, MET, and BAD show noticeably different activities in the two networks. The corresponding genes are known oncogenes; e.g., GAB2 – melanoma, MAPK1 – multiple cancers, MET - papillary carcinoma, BAD - pancreatic cancer, prostate cancer.

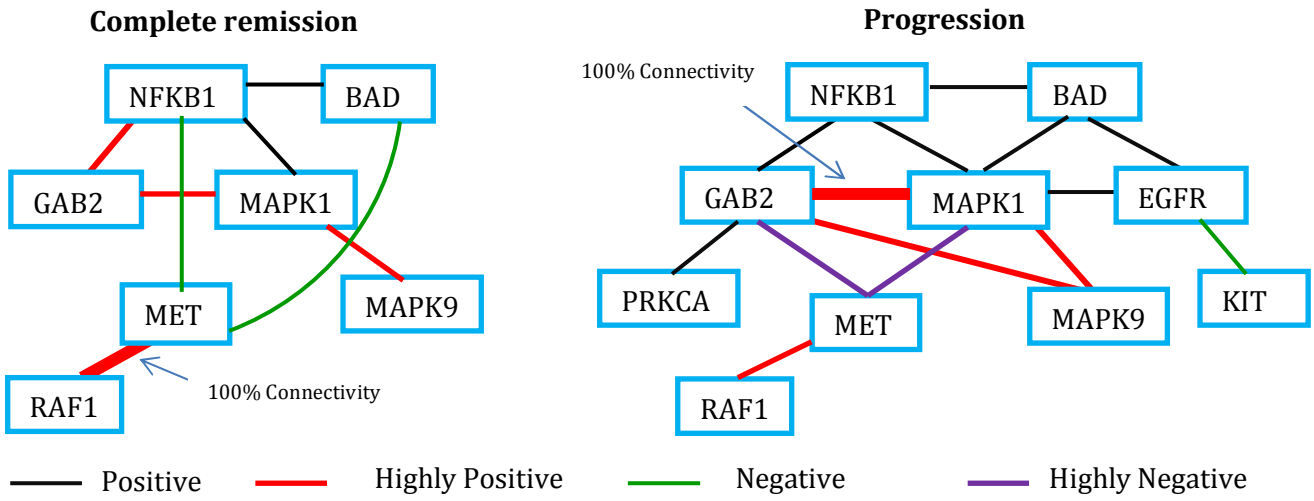


Figure 1: Network structure for RAS signaling pathway in Head and Neck Squamous Cell Carcinoma (HNSC)

Our analysis of individual proteins using the test statistic (2) produces Figure 2. The pie charts represent the proportions of top fifty differentially connected proteins for each of these pathways in the three cancers and in the overall aggregated list of proteins. Once again, PI3K and RAS take the top two most important spots in terms of differential network connectivity.

5 DISCUSSION

It is known that for most cancers with solid tumors the genes in the above mentioned "target pathways" display somatic mutations and change their protein products [1]. Here in this purely quantitative analysis of the existing protein expression data of three different cancers also reveals the significant alteration of the proteins in PI3K and RAS pathways. It is interesting to know that PI3K is a regulatory subunit, which binds to cell-surface receptors and to the RAS protein. Genes and proteins in PI3K and RAS have been investigated as therapeutic targets for many cancers ([8], [9] etc). Our findings are consistent with this and suggest that continued future efforts be made in this direction.

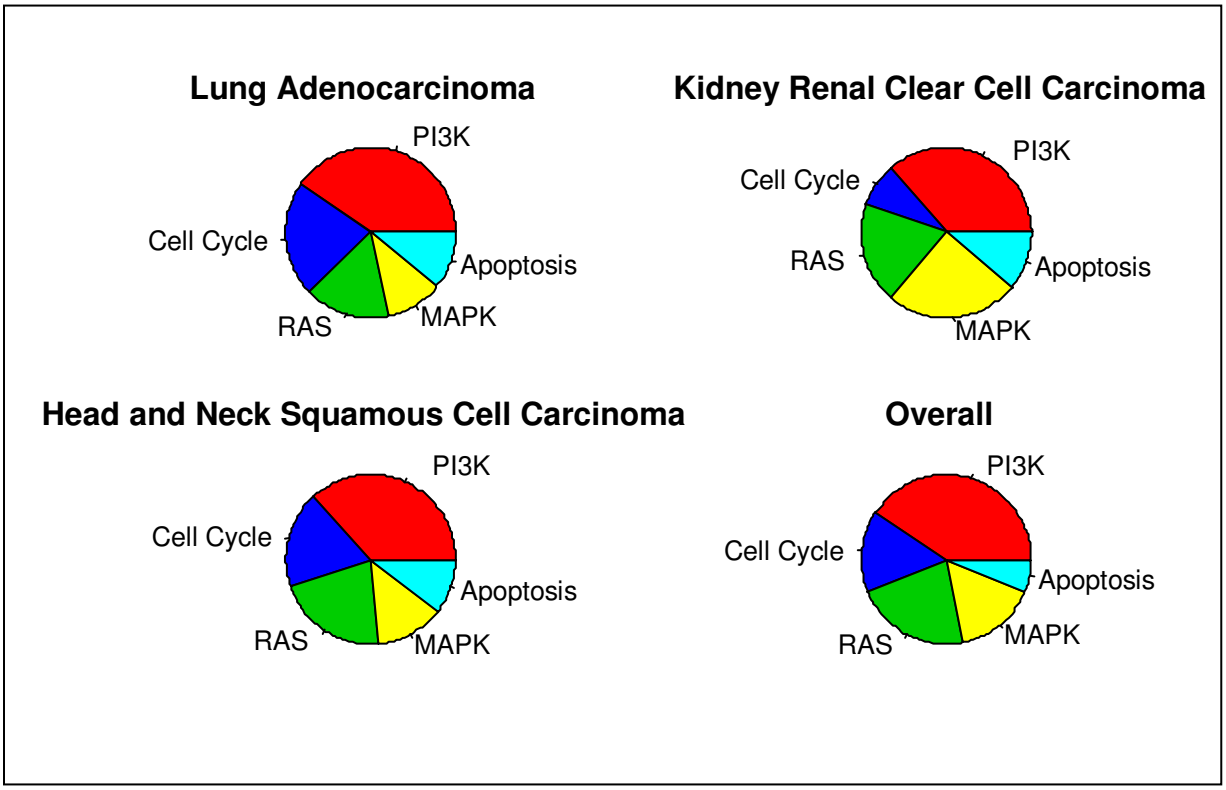


Figure 2: Relative contributions of the 5 "target pathways" in each of the three cancers separately as well as for all the three cancers combined.

References

[1] Vogelstein B, Papadopoulos N, Velculescu VE et al. Cancer Genome Landscapes. Science 2013; 339 (6127): 1546-58.
[2] Reimand J, Kull M, Peterson H et al. g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucl. Acids Res. 2007; 35(2): W193-W200.
[3] Reimand J, Arak T, Vilo J. g: Profiler—a web server for functional interpretation of gene lists (2011 update). Nucl. Acids Res. 2011; 39(2): W307-W315.
[4] Gill R, Datta S, Datta S. A statistical framework for differential network analysis from microarray data. BMC Bioinformatics 2010; 11(1): 95.
[5] Gill R, Datta S, Datta S. dna: An R package for differential network analysis. Bioinformation 2014; 10(4): 233–34.
[6] Pihur V, Datta S, Datta S. Weighted rank aggregation of cluster validation measures: a monte carlo cross-entropy approach. Bioinformatics 2007; 23(13): 1607-15.
[7] Rubinstein R. The cross-entropy method for combinatorial and continuous optimization. Methodology and Computing in Applied Probability 1999; 1: 127-90.
[8] Knight ZA, Shokat KM. Chemically targeting the PI3K family. Biochem Soc Trans.2007;35:245-249.
[9] Gysin, S., Salt, M., Young, A and McCormick, F. Therapeutic Strategies for Targeting Ras Proteins. Genes Cancer. 2011; 2(3): 359–372.

Title: -Multi-Omics analysis for understanding the molecular basis of Lung Adenocarcinoma.

Presented by: Agilent Technologies and Strand Life Sciences

Speaker: TBD

Abstract:

High throughput data from large cohorts of cancer patients is being generated by various consortia and is also being made accessible for use by other researchers. These data sources span multiple platforms and are a valuable resource for understanding the molecular basis of cancer. Key oncogenic and tumor suppressor players have been shown to undergo genomic copy number changes in lung adenocarcinoma patients [Nature 2014, 511:543-550]. In an effort to identify the oncogenes which are triggered by amplifications, we examined the expression levels of these genes in the context of their genomic aberrations (i.e. amplifications and deletions) and regulation by miRNA. Expression patterns of genes differentially expressed in tumors were further correlated with clinical and pathological metadata, as well as with mutational profiles of critical and known drivers of oncogenesis. An independent cohort of lung adenocarcinoma patients were evaluated for de-regulated pathways in the light of the mutational and genomic copy number findings from the TCGA cohort.

A pipeline for exploratory and pathway analysis of NGS data

Alejandra Cervera[1], Antonio Neme[2], Sampsa Hautaniemi[1]
[1]Research Programs Unit, Genome-Scale Biology, and Medicum, Biochemistry and Developmental Biology, Faculty of Medicine, University of Helsinki, Finland
[2]School of Medicine/ Institute of Biomedicine, University of Eastern Finland, Kuopio
alejandra.cervera@helsinki.fi, antonio.nemecastillo@uef.fi, sampsa.hautaniemi@helsinki.fi

We implemented a pipeline for processing and analysis of high-throughput sequencing and microarray level 3 data from ICGC Cancer Genome Consortium Challenges with the aim of finding the driver mechanisms in three cancers: Lung Adenocarcinoma (LUAD), Kidney Renal Clear Cell Carcinoma (KIRC), and Head and Neck Squamous Cell Carcinoma (HNSC).
We used a random forest to identify the genes that are most relevant for classifying the samples into normal tissue and tumor. The top 50 genes for each of the three cancers served as input for the SOMs showed in Figure 1. It can be observed that 50 genes are enough to achieve a fairly good separation of classes. DESeq2 was used to produce a list of differentially expressed genes (DEGs). Heatmaps using the genes with the greatest fold change from the DEGs are shown in Figure 2. It can be observed that the DEGs also achieve a fairly good separation of classes. The gene lists obtained from the random forest classification and the DEGs were used to identify possible relevant pathways for each cancer. From the top most represented pathways all genes belonging to those pathways were extracted. In the three cancers the Metabolic pathway (hsa01100) was enriched, and Salivary Secretion (has:04970), Cell Adhesion (hsa04514), and PI3K-Akt signaling pathways were enriched for HNSC, KIRC, and LUAD respectively. The new gene list obtained from the pathways was queried against the other levels of data: proteomics, miRNAs, copy number variation, and methylation. In the case of proteomics, we checked for indication of phosphorylation in any of the genes involved. Furthermore, DESeq was also used for obtaining a set of differentially expressed miRNAs and miRbase was used to find the known target genes. Figure 3 shows the miRNAs pathways in Cancer from where we found MIR17HG expressed.
The pipeline for preprocessing, analysing, and integrating all the datasets was implemented in Anduril; all steps are automated and it is available upon request (and soon made available in Anduril's website). Anduril is a framework for scientific data analysis that automates parallelization making it ideal for working with large datasets.

a) LUAD b) KIRC c) HNSC

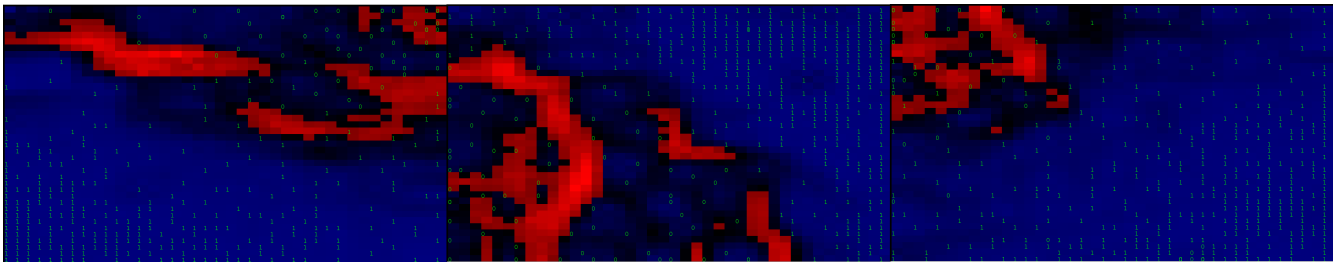


Figure 1. SOM using top50 most relevant genes from random forest classification of normal (0) vs tumor (1) samples. The color represents the distance between the points in the lattice (closer → blue, farther → red).

Figure 2. Heatmaps from gene expression of differentially expressed genes for each cancer.

a) LUAD b) KIRC c) HNSC

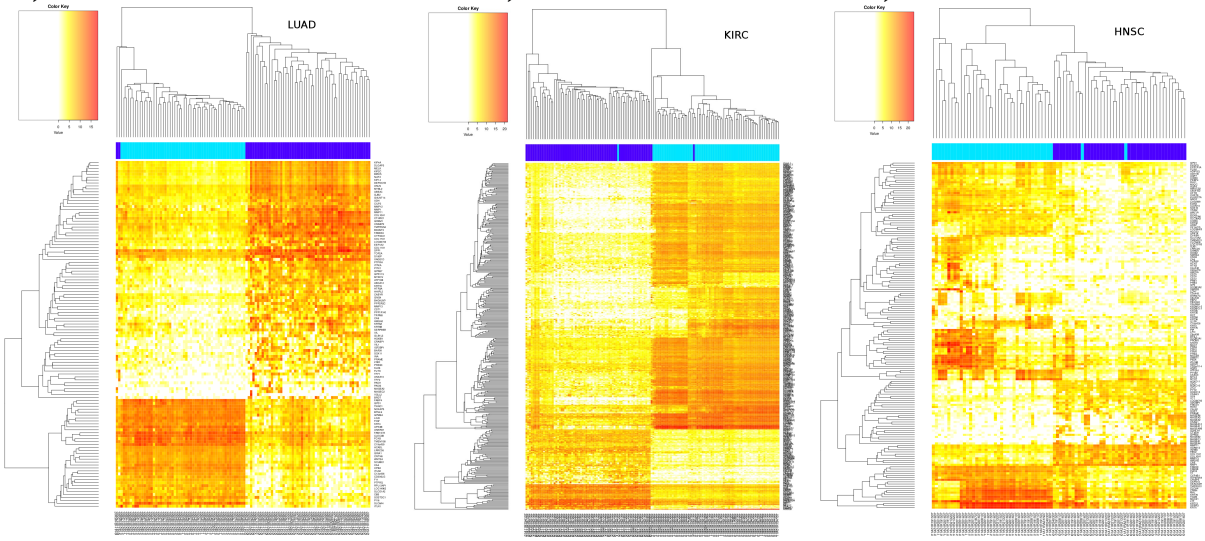
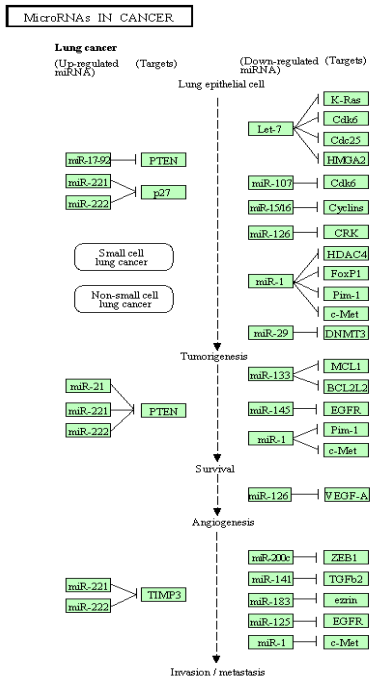


Figure 3. MiRNAs pathways in Cancer: miR-17~92 known as oncomiR-1 is known to be disregulated in cancer and we found MIR17HG (the primary transcript of the cluster) to be expressed in our samples.



INTEGRATIVE GENE SET ANALYSIS OF GENE AND MIRNA EXPRESSION DATA

FRANCISCO GARCIA-GARCIA[1] , JOAQUIN DOPAZO[1,2,3], DAVID MONTANER[1]

[1]Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF)
[2] Spain Bioinformatics of Rare Diseases (BIER), CIBER de Enfermedades Raras (CIBERER)
[3] Spain Functional Genomics Node, (INB) at CIPF

Introduction: From a systems biology perspective, gene set analysis (GSA) allows us to understand the molecular basis of a genome-scale experiments. Employing a systems biology approach that includes several genome-scale measurements gets a better functional interpretation. In this work we present a multidimensional method to the functional profile of mRNA and miRNA studies which integrates both expression data.

Methods: We downloaded 20 datasets from The Cancer Genome Atlas (http://cancergenome.nih.gov/), containing tumoral and normal samples. Differential expression analysis was carried out for mRNA and miRNA levels (Bioconductor library edgeR). Information from miRNA was transferred to gene level by adding its effects and generating a new index which ranks genes according to their differential inhibition by miRNA activity across biological conditions. Given both ranking statistics of mRNA and miRNA, for each functional class, we apply the logistic regression models for GSA. P-values were corrected for multiple testing using the method Benjamini and Yekutieli.

Results: This new approach has allowed to obtain a genomic functional profiling for different cancers when using an integrated approach with mRNA and miRNA data. In our study we used Gene Ontology terms (http://www.geneontology.org/) to define gene sets, obtaining detailed functional results for each ontology (biological process, cellular component and molecular function).

Discussion: Integrative Gene Set Analysis of mRNA and miRNA expression data constitutes a novel approach of functional profiling which allows us to detect interactions between gene and miRNA that account for functional roles dependent on several genomic properties or measurements. From this method, we can differentiate several patterns for functional modules to understand and discover of new cell functionalities with complex dependences.

Conclusion: This method may be successfully applied in genomic functional profiling, transferring miRNA data to gene level and integrating mRNA and microRNA data at the same level, so that GSA can be properly used. Functional results take advantage of the knowledge already available in biological databases and can help to understand large-scale experiments from a systems biology perspective.

References:

Benjamini, Y., Yekutieli, D. (2001). *The control of the false discovery rate in multiple testing under dependency*. The Annals of Statistics, 29(4), 1165–1188.
Montaner, D., Dopazo, J. (2010). *Multidimensional gene set analysis of genomic data.* PLoS One. Apr 27;5(4):e10348. doi: 10.1371/journal.pone.0010348.
Montaner, D., Minguez, P., Al-Shahrour, F., Dopazo, J., (2009) *Gene set internal coherence in the context of functional profiling*. BMC Genomics 10: 197.
Robinson, MD., McCarthy, DJ., Smyth, G K. (2010).
edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139–140.
Sartor, MA., Leikauf, GD., Medvedovic, M. (2009) *LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data.* Bioinformatics, 25(2):211–217.
Smyth, GK. (2004). *Linear models and empirical bayes methods for assessing differential expression in microarray experiments*. Stat Appl Genet Mol Biol 3: Article3.

Multi-omics data fusion for cancer data

Magali Champion, Olivier Gevaert

Stanford Center for Biomedical Informatics Research (BMIR), Department of Medicine, Stanford University, CA, USA

E-mail : mchampio@stanford.edu
ogevaert@stanford.edu

1 Introduction

Life sciences have been highly transformed by the emergence of the so-called "big data" era, synonimous of the large and multi-omics data sets now available. The increasing availability of such data provides a real challenge: integrate them to improve our understanding of biological concepts. As an example, the The Cancer Genome Atlas (TCGA) project aims at improving our ability to diagnose, treat and prevent cancer by analysing large numbers of human tumors, using gene expression, copy number, microRNA and DNA methylation data [1, 2]. In this contribution, the main goal consists of taking advantage of these multi-omics data to identify cancer driver genes (e.g. oncogenes) and to understand their roles within the genome. Previous work has focused on incorporating copy number data to filter potential regulators in a Bayesian module network analysis [3] whereas others have added mutation data for studying driver genes [4].

We recently developed AMARETTO, an algorithm that integrates copy number, DNA methylation and gene expression data to identify a set of driver genes by analysing both cancer and normal samples, and constructs a module network to connect them to clusters of co-expressed genes [5] and applied AMARETTO on several single cancer sites. Here, we propose a pancancer AMARETTO analysis. To accomplish this, we cluster the modules of co-expressed genes in communities according to their similarities to identify pancancer driver genes. This will allow the identification of master regulators across all cancers associated with common pathways across different types of tumors, and eventually may lead to the identification of pancancer drug targets.

2 Materials and methods

2.1 Data preprocessing

We used gene expression, copy number and DNA methylation data from TCGA for 11 cancer sites, namely bladder cancer (BLCA), breast cancer (BRCA), colon and rectal cancer (COADREAD), glioblastoma (GBM), head and neck squamous carcinoma (HNSC), clear cell renal carcinoma (KIRC), acute myeloid leukemia (LAML), lung adeno carcinoma (LUAD), lung squamous carcinoma (LUSC), serous ovarian cancer (OV) and endometrial carcinoma (UCEC) (for more details on these data sets, see Table 1). All data sets are available at the ICGC [6] and TCGA data portals [7].

The gene expression data were produced using Agilent microarrays for GBM and ovarian cancer, and RNA sequencing for all other cancer sites. Preprocessing was then done by log-transformation and quantile normalization of the arrays. The DNA methylation data were generated using the Illumina Infinium Human Methylation 27 Bead Chip. DNA methylation was quantified using β-values ranging from 0 to 1 according to the DNA methylation levels. We removed CpG sites with more than 10% of missing values in all samples. We used the 15-K nearest neighbour algorithm to

TCGA Cancer Site	Copy number data		DNA methylation data		Gene expression data	
	Samples	Genes	Samples	Genes	Samples	Genes
BLCA	178	1,974	123	472	181	15,432
BRCA	968	1,523	887	890	985	16,020
COADREAD	578	2,523	570	522	589	15,533
GBM	481	1,561	321	395	501	17,811
HNSC	365	2,184	308	753	371	15,828
KIRC	501	3,052	497	567	509	16,123
LAML	166	1,681	170	613	173	14,296
LUAD	487	3,585	367	678	489	16,092
LUSC	487	2,592	355	679	490	16,219
OV	528	1,499	540	510	541	17,814
UCEC	500	2074	496	821	508	15,706

Table 1: Overview of the number of samples and genes for each cancer site.

estimate the remaining missing values in the data set [8]. Finally, the copy number data we used are produced by the Agilent Sure Print G3 Human CGH Microarray Kit 1M×1M platform. This platform has high redundancy at the gene level, but we observed high correlation between probes matching the same gene. Therefore, probes matching the same gene were merged by taking the average. For all data sources, gene annotation was translated to official gene symbols based on the HUGO Gene Nomenclature Committee (version August 2012). Due to the size of TCGA data, the TCGA samples are analysed in batches and a significant batch effect was observed based on a one-way analysis of variance in most data modes. We applied Combat to adjust for these effects [9].

2.2 AMARETTO: multi-omics data fusion

Our approach for analysing TCGA cancer data is based on AMARETTO, a novel algorithm devoted to construct a module network of co-expressed genes through the integration of multi-omics data [5]. More precisely, AMARETTO is a two-step algorithm that (i) identifies a set of potential cancer driver genes by integrating copy number, DNA methylation and gene expression data, (ii) connects these cancer driver genes to their regulated modules of co-expressed genes using a penalized regulatory program. We describe in details these two steps below:

- Step 1: To establish a list of cancer driver genes, we investigate the linear effects of copy number and DNA methylation on gene expression through a linear regression model performed on each gene independently. Then we integrate DNA copy number and DNA methylation data to reduce the list of candidates. This will restrict the cancer driver genes to genes with either copy number or DNA methylation alterations. These alterations are detected using the GISTIC [10, 11] and MethylMix [12] algorithms for copy number and DNA methylation data respectively.
- Step 2: Given the cancer driver genes identified in Step 1, Step 2 aims at connecting them to their regulated targets to construct the module network. First, the filtered data are clustered in modules of co-expressed genes using a k-means algorithm with 100 clusters. Then, we regress independently all cancer driver gene expression values using as regressors every module's mean expression, i.e. each module is written as a linear combination of cancer driver genes. In order to induce sparsity, we choose to focus on the elastic net regularization [13]. The module network is finally constructed by running iteratively the two following steps: (i) reassigning genes based on closed match to new regulatory programs, (ii) performing the regulatory program, until less than 1% of the cancer driver genes are assigned to new modules.

2.3 Pancancer module communities

The pancancer analysis we perform is based on a careful comparison between the module networks constructed using AMARETTO for all considered tumor types. More precisely, we evaluate whether there is a significant association between all pairs of modules through a hyper-geometric test. We correct for multiple hypothesis testing using the false discovery rate [14]. We consider the association to be major if both of the following conditions are satisfied: (i) the adjusted p-value is < 0.05 and (ii) the overlap between two modules is larger than 5 genes. This defines a module network according to a score, measured through the minus log-transformation of the adjusted p-value. We used the open-source platform Cytoscape to visualize this network [15].

We finally cluster the module network in communities of modules using the clustering algorithm OH-PIN [16], implemented in Cytoscape. This algorithm has already proven to be powerful for identifying both overlapping and hierarchical modules in Protein-Protein Interaction Networks (PPI networks). To run it, we need to define an overlapping maximal score that limits the overlap between two communities (usually set to 0.5 [17]) and a threshold that controls the size of the communities (set to 2).

2.4 Gene set enrichment analysis

To assign biological meaning to these communities of modules, we perform gene set enrichment analysis based on the databases GeneSetDB [18] and MSigDB [19]. For the latter, we restrict the enrichment to hallmark (H), curated (C2), GO (C5), oncogenic (C6) and immunologic signatures (C7) gene sets, which are best suited for our study. The enrichment is evaluated by performing multiple hyper-geometric tests, corrected using the false discovery rate (FDR) [14].

3 Results

Running AMARETTO on the 11 cancer sites and performing pancancer analysis as described leads to a module network with 1673 edges between 592 nodes (Figure 1). Given this network, the clustering algorithm OH-PIN then identified 28 communities containing between 3 to 81 modules each. An example of such a community is highlighted in red in Figure 1.

Analysing more precisely the community represented in Figure 1, we found 35 regulators from 6 modules and representing 5 different cancer sites, namely BLCA, HNSC, LUAD, LUSC (two modules) and UCEC. The top two genes in this community are GPX2 and NQO1, with GPX2 present as a regulator in all modules and NQO1 in half of the modules. GPX2 is expressed at crypt bases of the intestinal epithelium and in tumour tissues. It also has been shown to be involved in cell proliferation [20]. NQO1 has been shown to be involved in the regulation of inflammatory mediators associated with prostate tumorigenesis [21].

Next, we used gene set enrichment analysis to investigate which pathways are enriched in this community. We found that chronic inflammation pathways were highly enriched in this community of modules. This included the NFE2L2 transcription factor [22]. This gene has proven to be critical in the lung's defense mechanism against oxydants, providing more precisely protection against chemical carcinogenesis, chronic inflammation or asthma [23]. In addition, a gene expression signature related to the response to cigaretto smoking is enriched in this community [24] and is also relevant for the pathogenesis of Chronic Obstructive Pulmonary Disease (COPD), a risk factor for lung cancer.

4 Discussion

We have presented a multi-omics data fusion framework that combines gene expression, DNA methylation and DNA copy number data across 11 cancer sites. Our goals are to find common

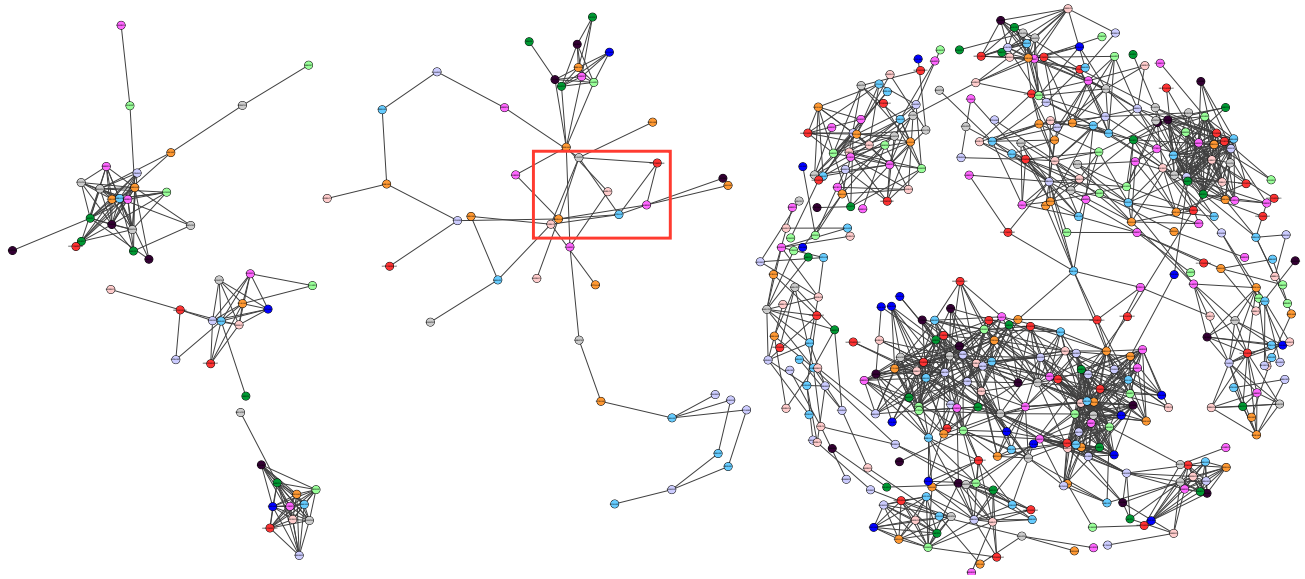


Figure 1: Visualization of the module network. The nodes of the graph are the modules of all cancers (represented using different colors according to the cancer type). An edge between two modules stands for a significant association between them. One of the community detected through OH-PIN is represented in red.

regulators across different types of tumors independent of anatomical location based on our hypothesis that tumors are more similar when considering their molecular makeup compared to their clinical profile. Our results show that pancancer communities of modules exist with common cancer driver genes. We highlight one community that is linked with chronic inflammation across carcinoma with a squamous nature including bladder cancer (BLCA), head and neck carcinoma (HNSC), lung cancers (LUAD and LUSC) and also including endometrial cancer (UCEC). More specifically, we identified two genes, GPX2 and NQO1, as pancancer regulators of chronic inflammation in these tumors.

Acknowledgements: Research reported in this publication was supported by the National Institute Of Biomedical Imaging And Bioengineering of the National Institutes of Health under Award Number R01EB020527, and by the National Cancer Institute under Award Numbers U01CA176299 and R01CA184968. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012.

[2] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.

[3] U.D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H.C. Causton, P. Pochanard, E. Mozes, L.A. Garraway, and D. Pe'er. An integrated approach to uncover drivers of cancer. *Cell*, 143:1005–1017, 2010.

[4] G. Ciriello, E. Cerami, C. Sander, and N. Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.*, 22:398–406, 2011.

[5] O. Gevaert, V. Villalobos, B.I. Sikic, and S.K. Plevritis. Identification of ovarian cancer driver genes by using module network integration of multi-omics data. *Interface Focus*, 3(4):20130013, 2013.

[6] ICGC data portal. https://dcc.icgc.org/repository/release_18/projects/.
[7] TCGA data portal. https://tcga-data.nci.nih.gov/tcga/.
[8] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R.B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17:520–525, 2001.
[9] W.E. Johnson, C. Li, and A. Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8:118–127, 2007.
[10] C.H. Mermel, S.E. Schumacher, B. Hill, M.L. Meyerson, R. Beroukhim, and G. Getz. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, 12:R41, 2011.
[11] B.S. Taylor, J. Barretina, N.D. Socci, P. Decarolis, M. Ladanyi, M. Meyerson, S. Singer, and C. Sander. Functional copy-number alterations in cancer. *PLoS ONE*, 3:e3179, 2008.
[12] O. Gevaert, R. Tibshirani, and S. Plevritis. Pancancer analysis of DNA methylation-driven genes using MethylMix. *Genome Biology*, 16:17, 2015.
[13] H. Zou and T. Hastie. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society*, Series B:301–320, 2005.
[14] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57:289–300, 1995.
[15] Cytoscape. http://www.cytoscape.org/.
[16] G.D. Bader and C.W.V. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(1):2, 2003.
[17] J. Wang, J. Ren, M. Li, and W. Fang-Xiang. Identification of hierarchical and overlapping functional modules in ppi networks. *IEEE transactions on nanobioscience*, 11(4):386–393, 2012.
[18] A.C. Culhane, T. Schwarz, R. Sultana, K.C. Picard, T.H. Lu, K.R. Franklin, S.J. French, G. Papenhausen, M. Correll, and J. Quackenbusch. GeneSigDB - a curated database of gene expression signatures. *Nucleic Acids Res.*, 38:D716–D725, 2010.
[19] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, and J.P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, 102:15545–15550, 2005.
[20] Taku Naiki, Aya Naiki-Ito, Makoto Asamoto, Noriyasu Kawai, Keiichi Tozawa, Toshiki Etani, Shinya Sato, Shugo Suzuki, Tomoyuki Shirai, Kenjiro Kohri, et al. Gpx2 overexpression is involved in cell proliferation and prognosis of castration-resistant prostate cancer. *Carcinogenesis*, 35(9):1962–1967, 2014.
[21] Dinesh Thapa, Peng Meng, Roble G Bedolla, Robert L Reddick, Addanki P Kumar, and Rita Ghosh. Nqo1 suppresses nf-κb–p300 interaction to regulate inflammatory mediators associated with prostate tumorigenesis. *Cancer research*, 74(19):5644–5655, 2014.
[22] A.J. Sandford, D. Malhotra, H.M. Boezen, M. Siedlinski, D.S. Postma, V. Wong, L. Akhabir, J.Q. He, J.E. Connett, N.R. Anthonisen, P.D. Paré, and S. Biswal. NFE2L2 pathway polymorphisms and lung function decline in chronic obstructive pulmonary disease. *Physiol Genomics*, 44(15):754–763, 2012.
[23] D. Malhotra, E. Portales-Casamar, A. Singh, S. Srivastava, D. Arenillas, C. Happel, C. Shyr, N. Wakabayashi, T.W. Kensler, W.W. Wasserman, and S. Biswa. Global mapping of binding sites for Nrf2 identifies novel targets in cell survival response through ChIP-Seq profiling and network analysis. *Nucleic Acids Res.*, 38(17):5718–5734, 2010.
[24] B. Harvey, A. Heguy, P.L. Leopold, B.J. Carolan, B. Ferris, and R.G. Crystal. Modification of gene expression of the small airway epithelium in response to cigarette smoking. *Journal of Molecular Medicine*, 85(1):39–53, 2007.

Signalling circuit activities as mechanism-based features to predict mode of action of chemicals.

Cankut Cubuk[1], Marta R. Hidalgo[1], Jose Carbonell-Caballero[1], and Joaquín Dopazo[1,2,3]

1. Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, Spain.
2. Bioinformatics of Rare Diseases (BIER), CIBER de Enfermedades Raras (CIBERER), Valencia, Spain.
3. Functional Genomics Node, (INB) at CIPF, Valencia, Spain.

Abstract

Most biological phenotypes are too complex to be described as consequences of the activities of individual genes but rather as a complex interaction among these. Here we propose the transformation of individual gene expression data into numerical descriptors of signalling pathway activities and its use to predict the mode of action (MOA) of chemicals. Here we addressed the Challenge 1a: SEQC Rat TGx - rat liver response to chemicals data, Topics 1 and 2. Our results show how the performance of the transformed values is quite good and how the predictions derived from RNA-seq seem to be better that the ones derived from microarrays.

Introduction

Many complex traits, as drug response, are associated with complex changes in biological pathways rather than being the direct consequence of single gene alterations. Actually, the idea of using the information contained in different biological pathways to understand complex traits, such as disease or drug mode of action is gaining acceptance [1]. Signaling pathways provide a formal representation of the processes by which the cell triggers actions in response to particular stimulus through a network of intermediate gene products. In particular, specific sub-networks (or circuits) that connect stimulus reception proteins to proteins that produce the consequent cell response can directly be related to cell functionalities. Recently some methods have developed that focus particularly on the estimation of the activity of these stimulus-response signaling circuits from gene expression data [2, 3].

Method:

We obtained *Rattus norvegicus* (rat) signalling pathway information from KEGG database. A total of 23 signalling pathways were examined. Each pathway was split up into their elementary signalling circuits, as described previously [3]. Activation-inactivation relationships between nodes (proteins) along the circuits enabled us to use

a graph traversal methodology for updating signal intensity at each visited node and finally compute a global value of signal transduction for the circuit, that we call signalling circuit activity therein. Unlike in previous methods [2, 3], the algorithm used here for the calculation of these signalling circuit activities is platform independent and can use gene expression data either from microarrays or from RNA-seq.

The microarray and RNAseq datasets (GSE55347,GSE47792) were downloaded from the GEO database. The raw microarray data were normalized by RMA method. The probe IDs were converted into Entrez Gene IDs. The probe expression values were summarized into gene expression values by 90 quantile.

The RNAseq data were already normalized, as provided by the MAGIC pipeline and we used them directly, and annotated with Entrez gene IDs (duplicated gene IDs were excluded).

In total, 1334 genes were used to calculate signalling circuit activities for the 867 sub-pathways that compose the 23 signalling pathways studied here. These signalling circuit activities and normalized gene expression values were used to compare their respective prediction accuracies.

ANOVA was used to detect the differential expressed genes and signalling circuits. All training and test set groups were used together for ANOVA.

For the prediction, support vector machine (SVM) with radial basis function (RBF) kernel was used [4]. Two parameters for an RBF kernel were used: cost and sigma. Best sigma and cost parameters were selected among different values tested. The model optimized with 10 fold cross validation.

(MOAs were used as endpoints for training the model as follows:

Training Set:

"PPARA", "CAR/PXR", "CONTROL", "UNKNOWN"(AhR, Cytotoxoc, DNA Damage)

Test Set:

"PPARA", "CAR/PXR", "CONTROL","UNKNOWN"(ER, HMGCOA)

Results and discussion

For both platforms the prediction accuracy obtained using signalling circuit activities as classification variables was reasonable and better than the corresponding accuracy obtained when using genes alone (see Figure 1).

It must be taken into account that not all the chemicals studied are acting at the level of the signalling pathways and therefore some MOAs will probably be deficiently predicted using only information on signalling. For example, HMGCOA (all) and AHR (LEFLUNOMIDE) MOAs are known to act at the level of metabolic pathways.

RNA-seq

Signalling Circuits

```
Confusion Matrix and Statistics

          true
pred       CAR_PXR PPARA UNKNOWN VEHICLE
  CAR_PXR        2     2       3       0
  PPARA          0     3       0       0
  UNKNOWN        7     2      15       0
  VEHICLE        0     2       0       6

Overall Statistics

               Accuracy : 0.619
                 95% CI : (0.4564, 0.7643)
    No Information Rate : 0.4286
    P-Value [Acc > NIR] : 0.01

                  Kappa : 0.4372
 Mcnemar's Test P-Value : NA
```

Model accuracy= 72.7%

Gene expression

```
Confusion Matrix and Statistics

          true
pred       CAR_PXR PPARA UNKNOWN VEHICLE
  CAR_PXR        0     0       0       0
  PPARA          0     0       0       0
  UNKNOWN        0     0       0       0
  VEHICLE        9     9      18       6

Overall Statistics

               Accuracy : 0.1429
                 95% CI : (0.0543, 0.2854)
    No Information Rate : 0.4286
    P-Value [Acc > NIR] : 1

                  Kappa : 0
 Mcnemar's Test P-Value : NA
```

Model accuracy = 40.90909

Microarray

Signalling Circuits

```
Confusion Matrix and Statistics

          true
pred       CAR/PXR Control PPARA UNKNOWN
  CAR/PXR        1       1     4       3
  Control        4       4     5       8
  PPARA          3       0     0       4
  UNKNOWN        1       1     0       3

Overall Statistics

               Accuracy : 0.1905
                 95% CI : (0.086, 0.3412)
    No Information Rate : 0.4286
    P-Value [Acc > NIR] : 0.99970

                  Kappa : -0.0171
 Mcnemar's Test P-Value : 0.00796
```

Model accuracy = 60.0%

Gene expression

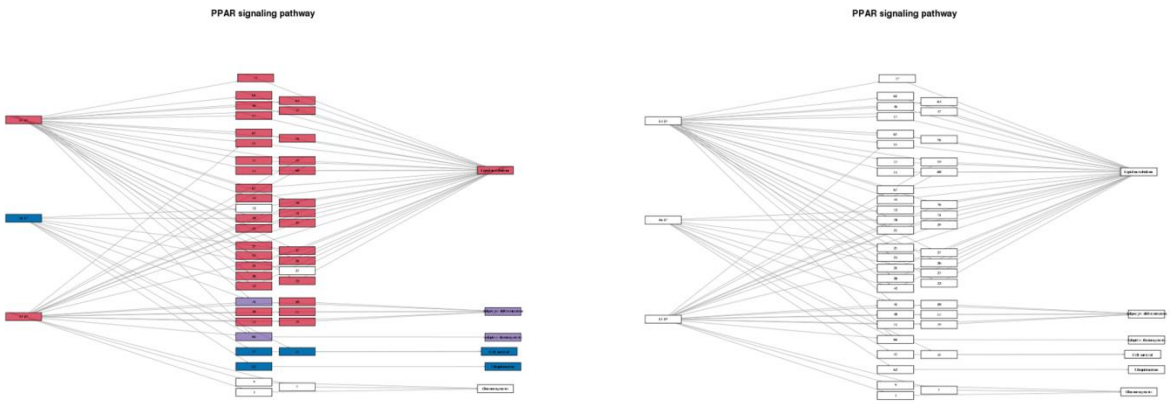
```
Confusion Matrix and Statistics

          true
pred       CAR/PXR Control PPARA UNKNOWN
  CAR/PXR        0       1     1       0
  Control        5       5     8      16
  PPARA          4       0     0       2
  UNKNOWN        0       0     0       0

Overall Statistics

               Accuracy : 0.119
                 95% CI : (0.0398, 0.2563)
    No Information Rate : 0.4286
    P-Value [Acc > NIR] : 1

                  Kappa : -0.0444
 Mcnemar's Test P-Value : NA
```

Model accuracy=48.9%

***Figure 1**. Prediction accuracy obtained using signalling circuit activities or gene expression values as classification variables obtained for RNA-seq and microarray data.*

An example in which the different effect of chemicals over pathways is obvious is depicted in Figure 2. It presents the analysis results of the PPARA signalling pathway

among the two MOA groups. Common MOAs groups of the test and training sets were merged for this analysis.
In the PPARA group (the group which exposed to PPARA agonists) the PPAS signalling pathway present a clear alteration in the lipid metabolism, while the AHR (and actually the other MOA groups, data not shown) have the pathway unaltered. These analyses were carried out for both, RNAseq and microarray data, rendering highly correlated results.



PPARA group AHRgroup

Figure 2. *Differential activities in the PPAR signalling pathway in presence of chemicals belonging to the PPAR (left) and AHR (right) groups.*

Conclusions

The method presented here shows that transforming the gene expression data into mechanism-based biomarkers within the context of signalling pathways is useful to predict molecular phenotypes that are controlled by signalling pathways. In addition, we were able to distinguish different phenotypes using signalling circuit activities.
We propose that approaches that model cell functionalities will be not only more accurate in predicting phenotypic traits, such as the drug response, but will also provide insights into the molecular mechanisms that account for such phenotype.

References

1. Davis MJ, Ragan MA: **Understanding cellular function and disease with comparative pathway analysis.** *Genome Med* 2013, **5:**64.

2. Sebastian-Leon P, Carbonell J, Salavert F, Sanchez R, Medina I, Dopazo J: **Inferring the functional effect of gene expression changes in signaling pathways.** *Nucleic Acids Res* 2013, **41:**W213-217.
3. Sebastian-Leon P, Vidal E, Minguez P, Conesa A, Tarazona S, Amadoz A, Armero C, Salavert F, Vidal-Puig A, Montaner D, Dopazo J: **Understanding disease mechanisms with models of signaling pathway activities.** *BMC Syst Biol* 2014, **8:**121.
4. Wang C, Gong B, Bushel PR, Thierry-Mieg J, Thierry-Mieg D, Xu J, Fang H, Hong H, Shen J, Su Z, et al: **The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance.** *Nat Biotech* 2014, **32:**926-932.

Inter-platform Concordance of Gene Expression Data for the Prediction of Chemical Mode of Action

Chathura Siriwardhana, Susmita Datta, Somnath Datta*

Department of Bioinformatics and Biostatistics,University of Louisville, KY 40202, USA

1. Introduction

The purpose of this study is two fold: (i) develop a classifier that has high accuracy in both microarray and RNA-seq platforms and (ii) study the concordance of multiple standard classifiers in the two platforms. We use seven standard classifiers and an adaptive ensemble classifier built around them to achieve these goals. The dataset for our study is resulted from a Rat liver experiment conducted by the FDA SEQC consortium to assess the performance of modern gene transcript expression profiling methods and released as part of 2015 Critical Assessment of Massive Data Analysis (CAMDA) challenges. The Rat liver experiment was designed for developing predictive models to predict the chemical Mode of Action (MOA). A previous comprehensive analysis (Wang et al. 2014) of the above gnomic data suggested weak classification accuracies for a set of classifiers applied to multiple platforms.

2. Data

The dataset consists of gene expression responses profiled by Affymetrix microarray and Illumina RNA-seq in rat liver tissues from 105 male Sprague-Dawley Rats, which were exposed to 27 different chemicals represented by 9 different MOAs. Microarray and RNA-seq platforms contain gene expression measurements of nearly 31,000 and 46,000 genes, respectively. In the original experiment, a training set is formed with 45 rats, which were treated with 15 chemicals corresponding to MOAs of "PPARA",

*somnath.datta@louisville.edu

"CAR/PXR", "AhR", "Cytotoxic", "DNA damage", and 18 controls. Test set contains data on 36 rats which were treated with 12 chemicals corresponding to "PPARA", "CAR/PXR", "ER", "HMGCOA" and 6 controls. We noticed that two MOAs, "ER" and "HMGCOA", are presented only in the test set. Due to duplication and removal of some initial samples, the data set profiled by RNA-seq contains 116 samples, which causes imbalance between training sets among platforms. We further noticed, approximately 22,253 average expressions per sample in RNA-seq data were recorded as "NA", where it indicates an insufficient number of reads mapped onto the gene to provide a reliable gene expression estimate. As a result, around 16,133 expression measurements remained, once all "NA"s were removed.

3. Methodology

Support Vector Machine (SVM), Random Forest (RF), Neural Network (NN), Linear and Quadric Discernment Analysis (LDA, QDA) are some examples of standard techniques widely applied in classification problems. For high dimensional data, these classifiers are often combined with dimension reduction, variable selection, or penalization techniques such as Partial Least Squares (PLS), Principle Component Analysis (PCA), Random Forest (RF) based importance measures, L_1 regularization, etc., for greater applicability and improved prediction accuracy (Boulesteix, 2004, Dai, 2006). However, the accuracies of these individual classifiers are highly variable and dependent on the true underlying data structures of various classes. Datta et al. (2010) described an optimal adaptive ensemble classifier via bagging and rank aggregation to offer a classification solution that has good performance across multitude of data structures. The ensemble classifier we used is developed with a set of seven standard classifiers, namely, SVM, RF, LDA, PLS+RF (Random Forest using the PLS terms), PLS+LDA (linear discriminant analysis using the PLS terms) , PCA+RF (Random Forest using the principal components), PCA+LDA (LDA using the principal components), and Recursive Partitioning (RPART).

We conducted three different analyses to study the performances of our classifiers in classifying the MOAs: (1) Classifiers trained and tested on each individual platforms; (2) Classifiers trained in one platform and tested on the other platform; (3) Classifiers trained on the perturbed training set with permuted gene expressions for each platform followed by accuracy calculation for identification of important variables (genes).

In general, there is no established criteria to define prediction for an unknown class that was not represented in the training data. Thus, we performed the 1st analysis after removing all test samples belonging to two classes of "ER" and "HMGCOA". However, for the 2nd and the 3rd analyses we were able to retain all classes and data since in effect the the classifiers were trained on the union of training and testing data in each platform. We used normalized expression levels that came from microarray

data using Robust Multi-Array Average (RMA) expression measurements (Irizarry et al., 2003), whereas data obtained for RNA-seq was already normalized via the Magic normalization. We felt that it would be more meaningful to perform an analysis with a common set of genes represented in both platforms for a comparative study. To that end the expression data for 8336 unique common genes were used in building our classifiers.

In the first analysis, we developed a set of classifiers directly using the training data with different classification algorithms and made predictions on the given test dataset in the same platform. However, since the classifier needed to run on both platforms for the 2nd analysis, each gene expression measurement was standardized, separately for both platforms, prior to the 2nd analysis. We performed a 10-fold cross validation for each individual classifier to select the number of components for PLS and PCA methods, separately for two platforms. We employed the same number of components to build the ensemble classifier. For the third analysis, we permute the expression of a single gene in the training set and fit a classifier on the modified training set followed by accuracy calculation on the test set. This was done for each of the gene common to both platforms. The reduction in accuracy as compared to the original (unperturbed) training set is a measure of importance of a given gene in the classification process. In order to reduce the computational burden, we did not use the ensemble classifier for this purpose. Instead the component classifier PLS+LDA which had an overall accuracy close to that of the ensemble classifier was used. The genes are then ranked according to their importance for both platforms.

4. Results

The results of analyses 1 and 2 are summarized in Figure 1. The left panel shows that the performance of each classifier is similar in both platforms since all the data points are fairly close to the diagonal line (Pearson's r =0.92). The accuracy of individual classifier varies from 17% to 75%, and as to be expected, the performance of the ensemble classifier is the best in both platforms. The overall accuracy of the optimal classification method is slightly better in microarray compared to RNA-seq (75% vs 67%). On the other hand, we observe a lower prediction accuracy for the class "PPARA" in RNA-seq (55.56%), compared to the microarray (88.87%) platform (not shown in the figure). Results of the second analysis summarized on the right hand plot shows even greater agreement between the prediction accuracies of the classifiers trained on a bigger training set in one platform and used to predict using the bigger test data on the other platform (Pearson's r =0.99). Remarkably, the ensemble classifier was able to provide 100% accurate predictions for both cases, regardless of the additional complexity caused by 8 varieties of classes. In this analysis, the component classifier PLS+LDA also performed similarly to the ensemble classifier in both cases yielding 100% accurate class predictions. Clearly, between the two types of dimension reduction methods, PLS performs better

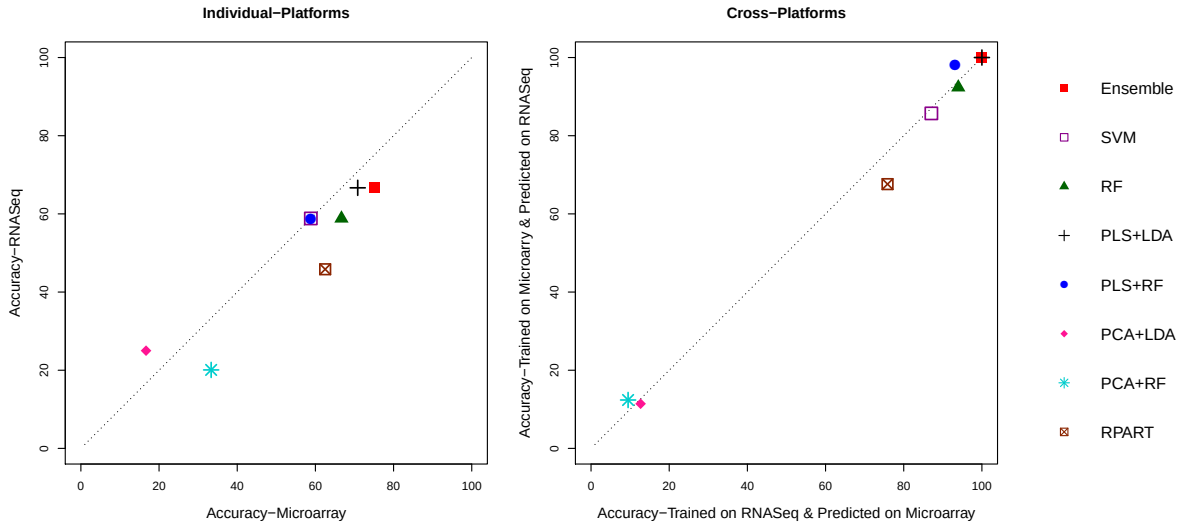


Figure 1: Plots between predication accuracies of RNAseq vs microarray test sets, by eight different classification techniques, for classifiers trained and predicted on individual platforms and cross platforms.

than PCA throughout this study. The performances of classifiers integrated with PCA are clearly the weakest among all individual classifiers in each scenario.

From the third analysis, we observed that five of ten most important genes for classification (Cyp1a1, Fam111a, Ugt2b, Akr1b8, and Hbb) were common between the two platforms. From literature search we found that CYP1A1 encodes a member of the cytochrome P450 superfamily of enzymes which catalyze many reactions involved in drug metabolism. Likewise, Ugt2b belongs to a large family of proteins capable of detoxifying a wide variety of both endogenous and exogenous substrates such as biogenic amines, steroids, bile acids, phenolic compounds and various other pharmacologically relevant compounds, including numerous carcinogens, toxic environmental pollutants and prescription drugs. Mutations in Hbb have been implicated in a number of blood disorders.

5. Discussion

In this study, we developed an ensemble classifier built on a set of standard classifiers to predict MOAs in Rat liver experiment data profiled by microarray and RNA-seq. The newly constructed ensemble classifier performed reasonably well in both platforms separately; we observed comparable overall predictability of MOAs in both test sets with 75% and 67% accuracies for microarray and RNA-seq, respectively. In an earlier classification approach applied on the same data, Wang et al. (2014) reported averaged overall

accuracies of 58% and 61% for microarray and RNA-seq, suggesting a slightly better predictability in RNA-seq. However outcomes of these two studies are somewhat incomparable due to the differences in test data sets used. For example, we omit two unknown classes present in original test sets after including controls as a separate class, whereas in their analysis, two unknowns were considered as another class while discarding controls. Interestingly, once we trained classifiers to make predictions on cross platforms, the ensemble classifier provided 100% accurate predictions for all 8 classes presented in the whole experiment. This result exhibits a perfect cross platform concordance in view of classification. Clearly, throughout the whole analysis, none of the individual classifiers outperformed the ensemble classifier with respect to the overall accuracy. However, PLS+LDA performs equally well in many cases. We observe widely different classification performances among standard classifiers, which reflects the unreliability of restricting to a single classifier in case of high dimensional classification problems. On the other hand, this also proves the utility of the adaptive ensemble classifier which is expected to perform as good or better than the individual classifiers.

References

1. Boulesteix, A. (2004), "PLS dimension reduction for classification", *Statistical Applications in Genetics and Molecular Biology with Microarray Data*, Vol 3(1), pp. 1-30.

2. Dai, J.J., Lieu L., Rocke, D., (2006), "Dimension reduction for classification with gene expression microarray data", *Statistical Applications in Genetics and Molecular Biology with Microarray Data*, Vol 5(1), 1544-6115.

3. Datta, S., Datta, S., Pihur, V., (2010) "An adaptive optimal sensemble classifier via bagging and rank aggregation with application to high dimensional data", *BMC Bioinformatics*, 11:427.

4. Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., Speed, T.P., (2003), "Summaries of Affymetrix GeneChip probe level data", *Nucleic Acids Research*, Vol 31, No. 4 e15.

5. Wang, C., Gong, B., Bushel, P.R., et al. (2014), "The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance", *Nature Biotechnology* 32, pp. 926-932.

Examining lost reads to survey the microbiome and immune components of the human body across 43 human sites from 175 individuals

Serghei Mangul[1], Nicolas Strauli[2], Ryan Hernandez[2], Roel Ophoff[3], Eleazar Eleazar Eskin[1,3], Noah Zaitlen[4]

[1]UCLA, Computer Science, Los Angeles, CA, [2]UCSF, Department of Bioengineering and Therapeutic Sciences, San Francisco, CA, [3]UCLA, Human Genetics, Los Angeles, CA, [4]UCSF, Department of Medicine, San Francisco, CA

Contact: smangul@ucla.edu

Advances in RNA sequencing technology and the ability to generate deep coverage data in the form of millions of reads provide an unprecedented opportunity to probe the universe of gene expression. Standard RNA-seq analysis protocols map reads against a host reference genome to determine the placement of the reads on the genome. Mapping-based protocols are complemented by assembly procedures to accurately profile the origin of reads condensed into isoform transcripts. Many reads are discarded by these protocols and the possibility that reads originate outside of the extant genome is usually ignored. In this work we aim to profile the origin of every last read delivered by RNA sequencing, in order to identify shortcomings of existing technologies as well as identify novel uses of RNA-Seq data. Our study reveals that the vast majority of unmapped reads are human reads discarded by the mapping protocol. Many unmapped human reads correspond to novel exon junctions from previously unknown isoforms. Another significant source of discarded human reads are sequences originating from the recombined Ig locus of B and T lymphocytes (BCR and TCR sequences). In addition to human DNA, the human body harbors a diverse microbial ecosystem, and we identified a substantial number of reads mapping to non-human sequence. Careful analysis of the BCR and TCR sequences in conjunction with the microbial communities provides an opportunity to profile immune system function across tissues directly from RNA-seq data.

We use 1641 RNA-Seq samples corresponding to 175 individuals and 43 sites from GTEx project: 29 solid organ tissues, 11 brain subregions, whole blood, and two cell lines, LCL and cultured fibroblasts from skin. Illumina Hiseq 2000 platform was used to produce Illumina RNA-Seq data sets. RNA-Seq libraries were prepared from total RNA using poly(A) enrichment of the mRNA. We use the unmapped reads to obtain a detailed profile of the microbial and immune components of the human body (unmapped reads were extracted from the .bam files download from the gtex storage). Wee obtained 6.77 +-1.60 million 76bp unmapped reads per sample. First we filtered out 54.68%+7.28% of the unmapped reads, which were low-quality and/or low complexity (using FASTX and SEQCLEAN). We attempted realignment of remaining reads to the human reference sequences using the bowtie2 aligner (up to 10 mismatches were allowed). Bowtie2 was able to identify 33.18%+-4.82% of the reads compatible with human genome and transcriptome reference (ENSEMBL hg19 build, ENSEMBL GRCh37 transcripome).

The remaining high-quality unique reads are used to perform a survey of the microbiome and immune components. We used phylogenetic marker genes to assign candidate microbial reads to the bacterial and archeal taxa. We use Phylosift to perform taxonomic classification of the samples and compare it across the tissues. The Phylosift approach uses hypervariable taxa-specific gene families to provide the precise resolution for the bacteria and archaea community assemblages. Hyper-variable regions from gene families are previously identified to be nearly universal and present in a single copy allowing differentiating between species and taxa. Reads are also mapped to a reference database of viral (n = 1,401), bacterial (n = 1,980) and fungal (n = 32) genomes downloaded from NCBI. To study the distribution of of B and T cells cross individuals and tissues we use reads mapped to the V(D)J regions of the Ig loci. Those recombinations correspond to early stages of T and B cell maturation.

A total of 713 taxa were assigned with Phylosift, with 8 taxa on the phylum level. Most of the taxa we observe are bacterial and a smaller portion is archeal .We observed no evidence of the presence of nonhuman eukaryotes. We observe all tissues to be dominated by Proteobacteria. No microbial organisms were observed in heart, pituitary and adrenal gland. All other tissues contain at least onebacterial or archeal phyla (0.79+-0.55 phyla per sample). We observe two viruses harbored in multiple tissues. EBV virus is present in 20% of the skin samples and 66% of the liver samples and it is not present in any of the brain samples. Enterobacteria phage phiX174 virus is present in 20% of the skin samples and is not present in liver and brain tissues.

Examining immune and microbial genes in GTEx can help define typical profiles for a healthy tissue. It is essential to monitor microbial and immune diversity, and this work may eventually help diagnose immune and microbiome imbalance in a tissue specific manner.

Sensitivity, specificity and reproducibility of RNA-Seq differential expression calls

Paweł P. Łabaj and David P. Kreil
Chair of Bioinformatics Research Group, Boku University Vienna, Austria
pawel.labaj@boku.ac.at

Introduction The MAQC[1] and SEQC[2, 3] projects have introduced a key resource for testing future developments of microarray and RNA-seq analysis tools, as required in clinical and regulatory settings. In this study, based on SEQC data set, we investigate the sensitivity, specificity and reproducibility of RNA-Seq differential expression calls. Going beyond general results of the original SEQC studies[2, 3] I will extend and complement the comparative analysis by considering differential expression tests that are closer to typical 'real world' experiments. In particular I will concentrate on comparisons of samples A and C, where C consists of 3 parts of sample A and 1 part of sample B.[1, 2] This pair of samples has the smallest average effect strength ('signal') amongst the different possible pair-wise comparisons of the MAQC/SEQC ABCD samples. Exploring the effect of RNA-Seq pipeline choices, we now also consider all 55,674 known AceView genes,[17] rather than the 23,437 genes of the originally published comparison with Affymetrix HGU133Plus2.0 microarrays. A key result of our study is thus as comprehensive benchmark of alternative methods for gene expression estimation and differential expression calling, representative of the wide range of tools now available and reflecting the rapid development of the field. The presented metrics assess sensitivity, specificity, and reproducibility for both genome wide analysis and the identification of top candidates for further follow-up.

Results Comparing the SEQC samples A and C we are expecting more genes with a stronger expression in sample C because it contains, in addition to RNA from sample A, also RNA of genes expressed in sample B. Benchmark results compare a wide range of tools for expression estimation (**EE**), including r-make[4], Subread[5], TopHat2[6]/Cufflinks2[7], SHRiMP2[8]/BitSeq,[9] and kallisto[10], in combination with a range of established tools for differential expression calls (**DEC**), including limma[11], edgeR,[12] and DESeq2.[13] Tools were selected to provide a good overview of the current state of the art in RNA-Seq data analysis.

Depending on the methods for expression estimation and DE calling, the number of detected differentially express genes vary roughly between 7,000 – 10,000 (Fig. 1). Sensitivity in general depends less on the method for differential expression calling, while more variation is observed for the different approaches in estimation of expression levels. Remarkably, there nevertheless is only limited agreement of the lists of genes identified by different methods for differential expression calls, with a typical pairwise agreement of 56–67%. To investigate these discrepancies we examined *M(A)* plots, where genes are represented by dots coloured according to which methods identified them as differentially expressed (Fig. 2). In the left panel (for AvsC) we can identify areas where different DEC methods are particularly sensitive. Variation in the sensitivity of DEC methods for different effect strengths (*M*) and gene abundances (*A*) reflects the range of approaches to data normalization and statistics used for DE calling. Among the examined DEC methods, DESeq2 appears to be the most conservative in calling DE genes of low abundance (low average expression). This may be appropriate considering the relatively high variance of low count data that is characteristic of weakly expressed genes in RNA-Seq.[14] Also weakly expressed genes might be relatively more affected by site-specific variation arising during library preparation[3], as seen in the right panel of the Fig. 2, which shows a same–same sample comparison – genes identified there as 'differentially expressed' are false positives (FPs) in a search for biologically relevant differences.

The SEQC study design[1 – 3] provides us the unique possibility to further examine the site-specific effects. In particular, we can calculate an eFDR (empirical False Discovery Rate) by comparing the cross-site sensitivities for AvsC, CvsC and AvsA (Fig. 3, and Fig. 4 left panel). The number of false positives (FPs) can be reduced when appropriate methods[15, 16] are applied to remove the unwanted variation by analysing the experiment in context of similar experiments obtained from the public repositories. In our study we can use different sequencing site to mimic such 'context'. We have applied the PEER tool[16], which has performed the best in the SEQC study[3], to remove the unwanted variation. The eFDR has been

reduced noticeably from typical eFDR reaching up to 50% to not crossing in general the 20% (Fig. 4 left vs middle panel). As the eFDR is strongly dependent on the combination of EE method and DEC method, even after PEER some sequencing site pairs obtain more than 60% eFDR (outlier sites for kallisto). As the eFDR level is still not satisfactory the further filtering is needed as was shown in MAQC[1] and SEQC[2, 3] studies. In terms of RNA-Seq, unlike for microarrays, in addition to filter for small changes also the filter for small expression levels is required (see Methods for threshold details; reasoning, approach and consequences will be extended in the full version of the manuscript). This is the direct consequence of the sampling nature of NGS[14]. Application of the dedicated filters which fix the EE+DEC pipelines sensitivity for intra-site AvsC comparison to about 3000 differentially expressed genes reduced the eFDR for a typical site pair below 2.5% for almost all cases. Just for SHRiMP2/BitSeq and kallisto used together with edgeR the typical eFDR is higher but still below 5 %. Adding filtration by removing the FPs not only lower the eFDR but also increase the agreement between DEC methods as now the method specific FPs has been removed. The agreement has increased from 60-67% (after PEER correction) to the level of 86-94% depending on site, EE and DEC method.

In medical and life sciences the goal is to produce the accurate gene signatures – lists of differentially expressed genes which can be reproduced in the other laboratory. This challenge can be seen in different ways depending on the study design and the next steps which will be taken with the provided gene signatures. In terms of the whole genome studies the interest is in the accuracy of the list of all differentially expressed genes. Based on our study we can conclude that agreement between sequencing sites depends strongly on the selected DEC method when no addition filtering is used: typically 50-68% for limma, 66-72% for edgeR and 72-78% for DESeq2. Application of the additional filtering, although reduce the sensitivity, increase the agreement and makes that all DEC method have more similar ranges: typically 77-80% for limma, 81-83% for edgeR and 82-84% for DESeq. There are studies, however, where the full list of differentially express genes is not of interest. More important is list of ‘top’ candidates which can be further tested in follow-up studies. As here not the sensitivity but rather specificity is of the main concern, the filtering is more than welcome. In Figure 5 the summarized agreement (on y-axis in %) between topN (where N is on x-axis) differentially expressed genes (sorted by the effect strength) is shown. The different panels represent different DEC methods while different colours in violin plots represents different expression estimation methods (as specified in the legend). For the short top lists the agreement is strongly dependent on combination of EE and DEC methods. These differences are getting smaller when lists are getting longer with almost all combinations reaching 80% agreement for top200 and crossing 90% agreement for top1000. For a good performance for the short lists the solution might be to use of even stronger filters on average expression, but then the ‘true’ candidates with weaker average expression can be filtered out, what for many studies can be a big loss. That is why a better approach is to consider a different combination of EE and DEC method, eg. SHRiMP2/BitSeq + edgeR).

Conclusions Going beyond the general comparison presented in SEQC study[2, 3], we present here the benchmarking for scenarios which better represent the effect sizes of typical experiments. We have in details examined the sensitivity, specificity, and reproducibility of the RNA-Seq differential expression calls for a comparison of the SEQC samples A and C. We have shown that application of appropriate procedures and filters improves the reproducibility of both the genome wide analysis as well as the identification of top candidates. We also have shown that it is important to benchmark different analysis tools and pick the one which fits the best for our scenario.

In particular, it is crucial to analyse results in the context of similar experiments, as such an approach allows to apply tools like PEER which can identify and remove hidden confounders, having a great influence on the eFDR without changing the overall landscape of sensitivity. We have shown, however, that further filtering of FPs is required to obtain acceptable level of eFDR. A cost of an improved specificity is the decreased sensitivity. The good news is, however, that both for the genome wide studies as well as for ones when the top candidates are identified the results have been improved. When we consider the full list of genes called as differentially expressed, both the agreement between sites for the same DEC method as well as the agreement between different DEC methods improves

noticeably with filtering, making analysis results more robust and easier to reproduce across laboratories. The improvement from filtering can also be seen for the top ranked candidates with the strongest expression change. Here we can recommend in general the use of DESeq2 tool for DE calling especially in combination with BitSeq. This combination performed particularly well for the shorter lists of the most highly-ranked 50–200 differentially expressed genes. Different aspects of performance, however, vary across tool combinations. In general, pipelines relying on Tophat2/Cufflinks2 for estimation of expression levels performed the worst, while newer tools such as BitSeq (or kallisto) performed better.

Future work for the full manuscript: For the conference presentation and the full proceedings manuscript, the analysis will be extended to explore DE calling by dedicated methods for BitSeq and Cufflinks, and examine BitSeq DE calling for kallisto bootstrapping results.

Methods In this study the SEQC data set has been used (which is described in details and summarized elsewhere)[2]. Here the sequencing data of samples A and C of six Illumina HiSeq 2000 sites have been used.

The expression profiles of AceView[17] genes has been assessed by selected tools representing the state of the art approaches for expression profiling. The gene expression profiles were assessed in the form of read counts. R-make (based on STAR) and Subread perform the alignment to the genome what is followed by counting the reads which are falling into the gene regions. The TopHat2 with the G option represents the hybrid approach, where reads are first aligned to the virtual transcriptome and then mapped back to the genome. The gene and transcript expressions are then estimated with Cufflinks2 based on the genome based alignments. BitSeq uses directly the transcriptome alignments (here provided by SHRiMP2) to assess the transcripts abundances. These where then sum up per gene to obtain the read count estimates for genes. Kallisto represents the alignment free approach, where transcript abundances are assessed directly from reads based on graph pre-built with use of the transcript sequences. Also here the transcript expression estimates where sum up per gene to obtain the read count estimates for genes.

Gene expression estimates for all samples were used to detect latent variables using PEER package[16]. The covariates associated with sample type were included for inference and the inferred hidden confounders were removed from the signal.

Differential expression analysis has been performed with use of three dedicated R packages: limma, edgeR and DEseq2. In all three cases the suggested way of analysis has been performed (in terms of limma it includes TMM+voom pre-processing). The Benjamini-Hochberg adjustment for multiple testing has been performed. The genes were called differentially expressed when q-val <0.05. When filtering has been applied in addition: gene effect strength has to be higher than 2 (abs(log2FC) > 1) and the Average Expression has to be higher than dedicated threshold. Average expression threshold was selected for each combination of expression estimation and DE calling method separately in order to fix the average intra-site AvsC sensitivity at level of 3000 genes. On average 45th percentile with SD of 2.3 has been used (lowest for limma than DESeq2 and edgeR; lowest for Subread, then kallisto, TH2G, BitSeq and r-make). The same thresholds have been applied to inter-site DE calling. The DE analysis has been focused on down-regulated genes in A versus C comparison, as the strength of the up-regulated signal is limited by design of sample C as 3 parts of A and one part of B.

Overall agreement between lists of differentially expressed genes has been calculated as ratio of lists intersection and lists union. Agreement of topN candidates has been calculated as ratio of intersection of compared topN lists and the N, where differentially expressed candidates have been order by the change strength.

Bibliography

1. Shi, L. *et al*. Nat. Biotechnology 24, 1151–1161 (2006)
2. Su, Z. *et al.* Nature Biotechnology 32, 903–914 (2014)
3. Li, S. *et al.* Nature Biotechnology 32, 888–895 (2014)
4. Dobin, A. *et al*. Bioinformatics 29, 15–21 (2013)
5. Liao, Y. *et al*. Nucleic Acids Res. 41, e108 (2013)
6. Trapnell, C. *et al*. Nat. Biotechnol. 31, 46–53 (2013)
7. Kim, D. *et al*. Genome Biol. 14, R36 (2013)
8. David, M. *et al.* Bioinformatics 27, 1011–1012 (2011)
9. Glaus, P. *et al.* Bioinformatics 28, 1721–1728 (2012).
10. Bray, N. *et al.* arXiv:1505.02710 (2015)
11. Smyth, G.K. *et al.* 397–420 (Springer, New York, 2005).
12. Robinson, M.D. *et al.* Bioinformatics 26, 139–140 (2010).
13. Love, M.I. *et al.* Genome Biology, 15, pp. 550
14. Łabaj, P.P. *et al*. Bioinformatics 27, i383–i391 (2011)
15. Leek, J.T. *et al*. Bioinformatics 28, 882–883 (2012).
16. Stegle, O. *et al*. PLoS Comput. Biol. 6, e1000770 (2010).
17. Thierry-Mieg, D. *et al.* Genome Biol. 7, S12 (2006).

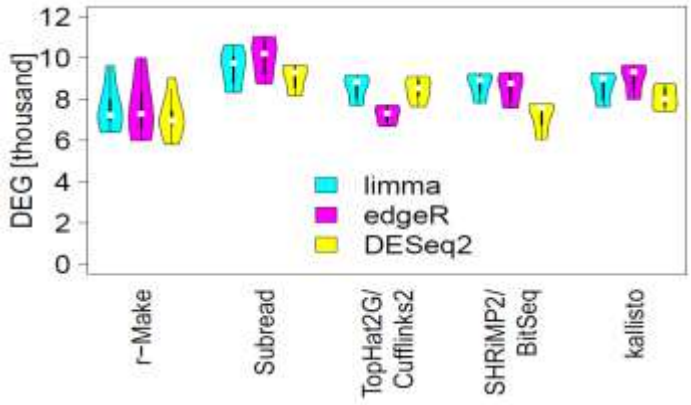


Figure 1. Intra-site differential expression call sensitivity. For each expression estimate method (x-axis) and each DE calling method (colour) all intra-site A versus C comparisons are presented in a form of the violin plot. Y-axis represents the sensitivity as a number of differentially expressed genes (with q.val<0.05).

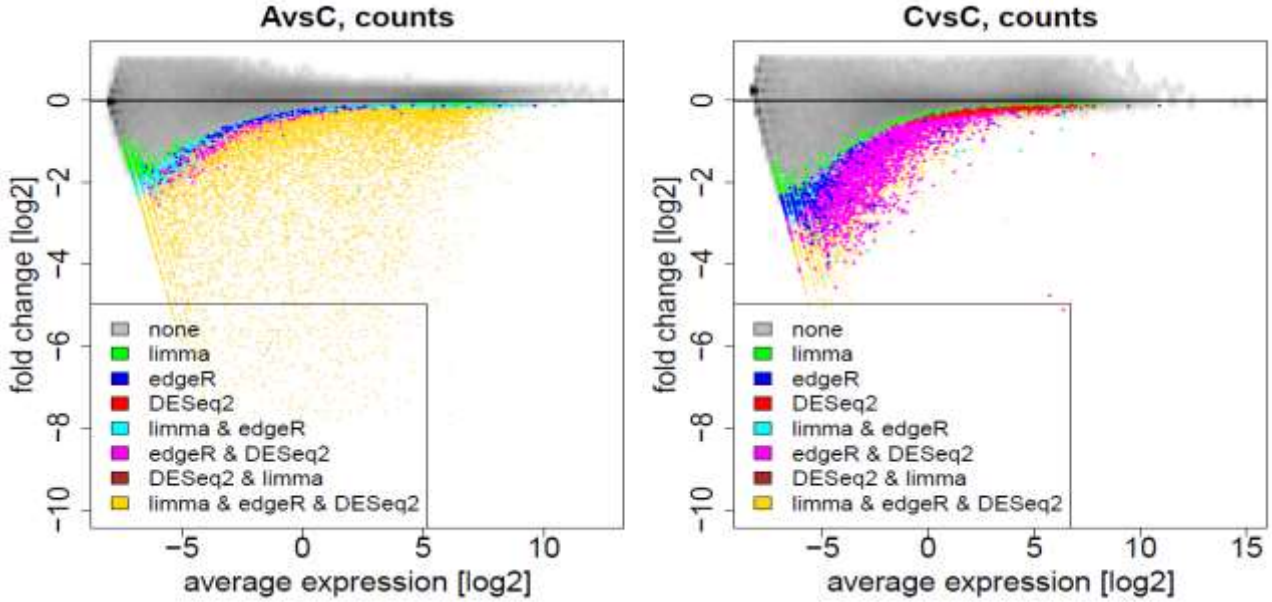


Figure 2. Left panel represents the overlap of the DE calling by different DEC methods for AvsC intra-site comparison, while right panel shows the results for the inter-site AvsA comparison. The overlap between calling as DE by different DEC methods is encoded by different colours. Grey clouds represent not down-regulated genes.

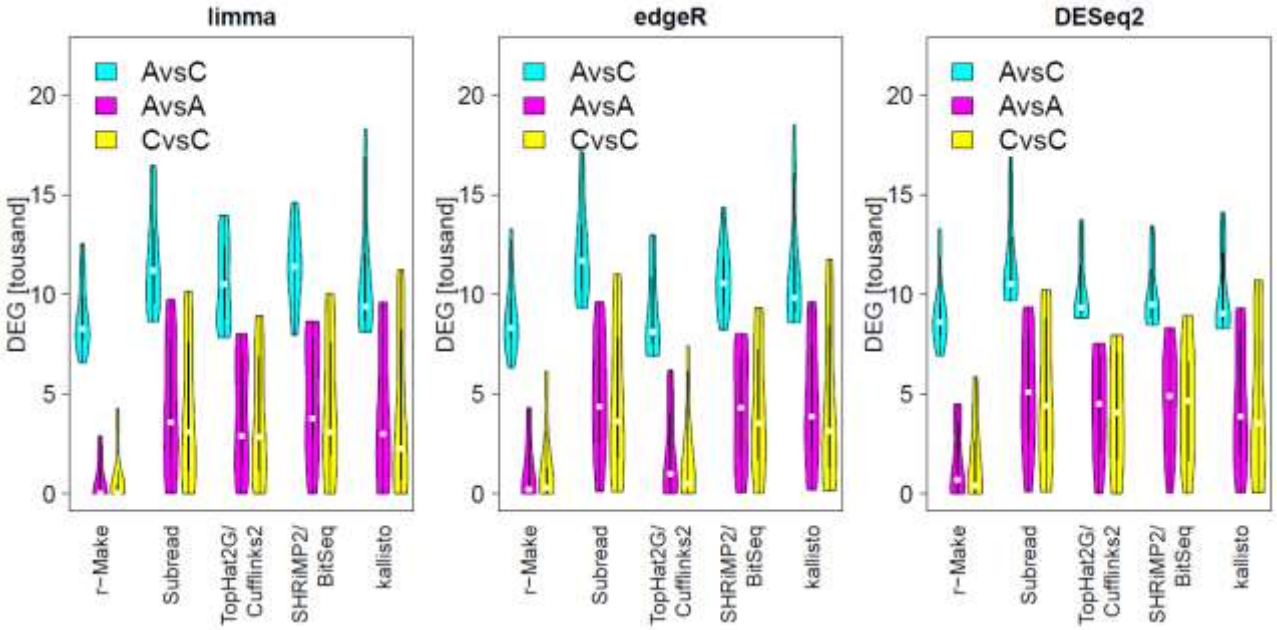


Figure 3. Inter-site differential expression call sensitivity, including false-positives from same–same comparisons. For each expression estimate method (*x*-axis) and each DE calling method (panel) all inter-site A versus C comparisons (cyan) as well as A versus A (magenta) and C versus C (yellow) are presented in a form of violin plots. The same–same comparisons show the sensitivity of methods to picking up false positives.

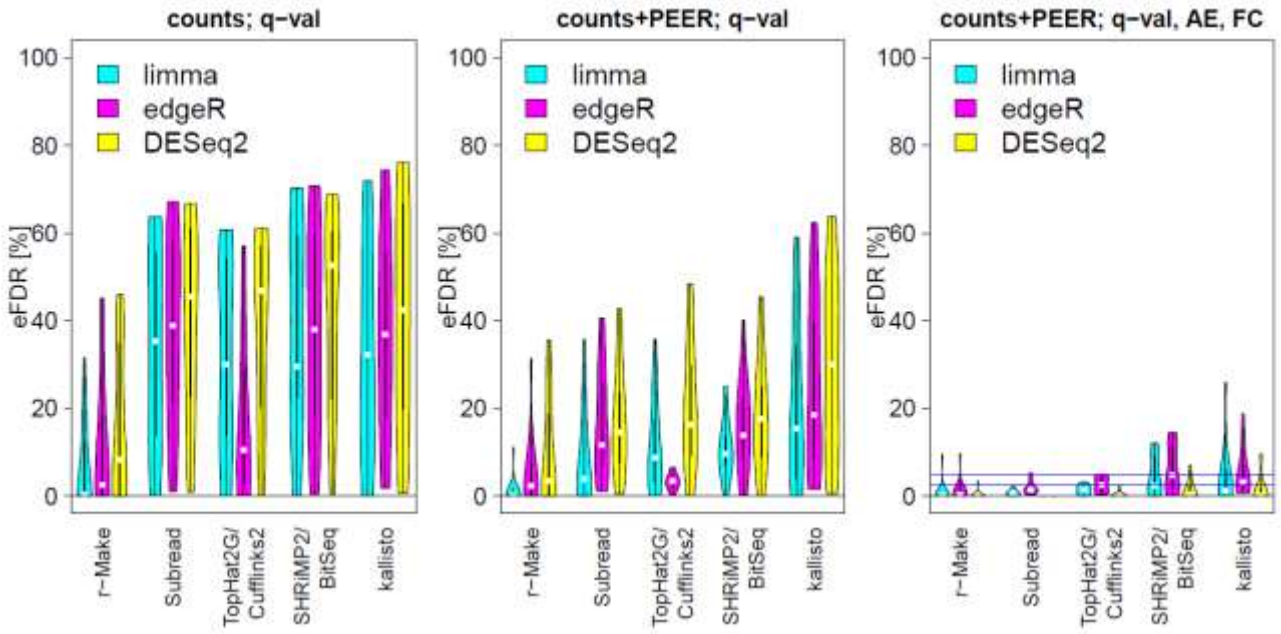


Figure 4. eFDR. For each expression estimate method (x-axis) and each DE calling method (colour) eFDR has been estimated as ratio between inter-site A versus A plus C versus C and A versus C. The left panel represents results based on not corrected counts with DE calling by q-val threshold. In the middle panel hidden confounders have been removed by PEER from count expression estimates. In the right panel additional DE calling filters has been applied (as described in methods).

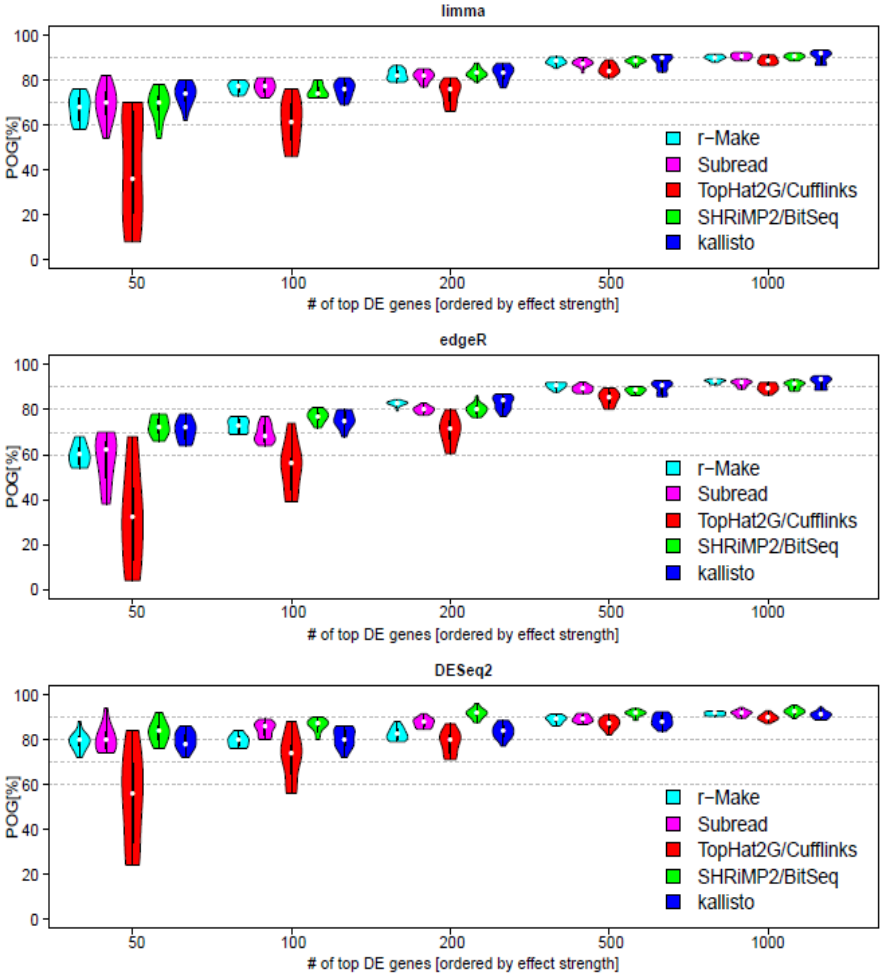


Figure 5. Inter-site reproducibility of differential expression calls. Comparing the identities and the directions of change for DEGs across sites, agreement is plotted for lists including the top-ranked genes as sorted by effect size (*x*-axis). The observed response violin plots depend on expression estimate pipeline, DE calling pipeline and filter choice, showing more variation and lower agreement levels for shorter lists. Results for BitSeq and DESeq2 seems to be the most robust. Agreement for top1000 genes cross 90% irrespectively from the pipeline choices.
Presented results were obtained based on expression estimates after removing the hidden confounders by PEER. For DE calling additional filters for average expression and effect strength have been applied.

Unbiased Optimization of Microarray Pre-processing

Najmeh Abiri[1], Payam Delfani[2], Mattias Ohlsson[1], Christer Wingren[2], Patrik Edén[1]
1: Computational Biology & Biological Physics, Dept. of Astronomy and Theoretical Physics, Lund University
2: Affinity Proteomics, Dept. of Immunotechnology, Lund University

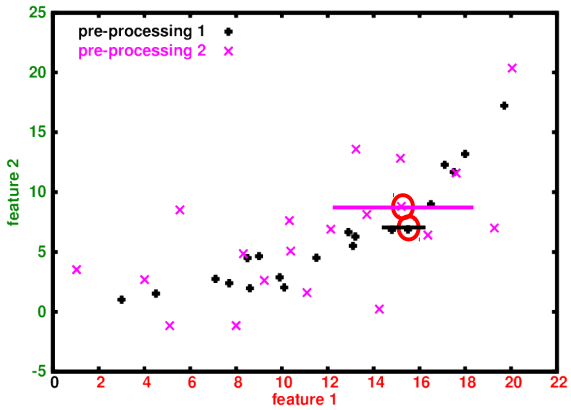
Corresponding author: Patrik Edén, patrik@thep.lu.se

The objective
Standard statistical methods, preferably involving test sets, can control false discovery rates in the enormously flexible microarray data analysis. However, it is normally assumed that a similar flexibility in pre-processing (e.g. quality control, normalization and variance filter) was not exploited with knowledge of sample annotations. This leaves the typical research group with the unpleasant choice to either abstain from pre-processing optimization or lose formal control of their statistical tests.

We develop new computational tools that optimizes pre-processing without any use of sample annotations, or any use of sample cluster structure.

The Tool: Validated Imputation
High-throughput microarray data is expected to be rich on correlations. This explains the success of imputation algorithms, that exploit the correlations to estimate missing values. Imputation algorithms are usually evaluated by artificially removing known data, impute them, and check the error to the true values.

Our approach, which we call **Validated Imputation (VI)**, is to use this imputation test "backwards". Instead of testing imputation algorithms with benchmark data, we test pre-processing options with benchmark imputation algorithms.

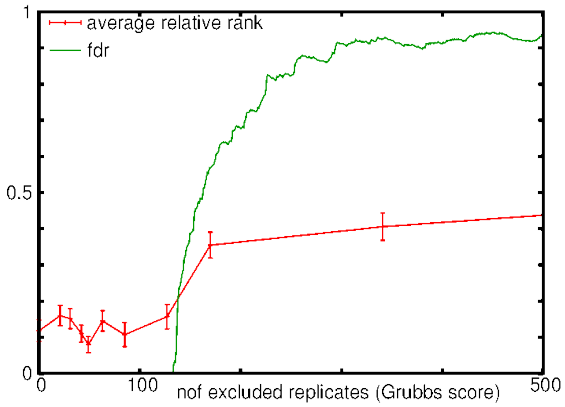


The principal idea is that proper quality filters and noise reduction give better imputation.

Artificial example: The same data in two pre-processings. The feature 1 value of the encircled sample is artificially removed (for both pre-processings) and re-imputed, ending up somewhere on the line. In the noisy (purple) option, that may be far away from the correct value.

A Test Case
Our protein affinity array includes 3-8 replicate spots, where technical errors may appear as outliers. This introduces an outlier threshold as a pre-processing parameter. Outliers among triplicates can be estimated using the Grubbs score (maximal distance to sample mean, divided by sample estimate of standard deviation).

Under normality assumption, the *p*-value and a false-discovery rate (fdr) is calculable. High fdr implies that many points rejected as outliers actually carried useful information.

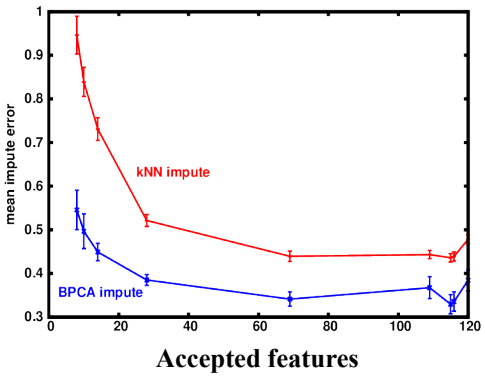


We have run validated imputation on data with various outlier thresholds, and checked which alternative that most often performed best. All threshold options imputed the same set of 5% artificially removed values, and they were ranked according to mean squared error. The rank was scaled to a number between 0 and 1 (relative rank). This was done multiple times, and the average relative rank was recorded. As seen in the figure, VI agrees in conclusion with the fdr analysis. Both methods suggest that the outlier threshold should be set to assign roughly 120 outliers. With a more stringent threshold, the fdr curve shows that a large fraction of excluded data carries useful information, and the VI test shows that imputation becomes less successful.

Discussion

The test case is a promising result, suggesting that validated imputation works: it can rank different pre-processing options. In this case, the suggested optimum agrees with an alternative approach (the fdr based on normality assumption). Note that VI provides slightly more information than the statistical test: The fdr curve shows when a large fraction of extra excluded values carries useful information, but fdr cannot tell if it is worth the prize, in order to exclude a few more technical errors. VI settles the question.

One other merit with VI is that it can be used also when there is no simple statistical test available, and it can be used to evaluate more or less heuristic procedures for, *e.g.*, background correction or normalization. For the protein recombinant antibody array in particular, the relatively low number of features measured means that one must look for normalization strategies other than standard approaches for mRNA and DNA arrays.



We have examined if VI can be run with "any" benchmark imputation algorithm, by comparing the relatively simple and quick knn-impute algorithm [1] with the more elaborate but slow bpca-impute algorithm [2]. Here, we examined a variance filter, removing low-varying features. Overall, BPCA performs better, but the important message is that both algorithms agree on the *conclusion*. (In this case that a variance filter is not really needed, except perhaps to remove 3-5 of 120 features.)

Conclusion and outlook

Validated Imputation is a promising tool to allow pre-processing optimization of high-throughput data without being influenced by any final data analysis results. We will continue to develop it within the framework of protein antibody array data, to examine quality control filters and normalization strategies. The basic method is in principle applicable to any high-throughput data with inherent correlations, for which imputation algorithms outperform simple row-average imputation.

References:

[1] Troyanskaya et al., Bioinformatics 2001, 17 p.520
[2] Oba et al., Bioinformatics 2003, 19 p.2088

Genome-wide detection of intervals of genetic heterogeneity associated with complex traits*

Felipe Llinares López [1], Dominik G. Grimm [1], Dean A. Bodenham [1], Udo Gieraths [1], Mahito Sugiyama [2,3], Beth Rowan [4], Karsten M. Borgwardt [1]

[1]*Machine Learning and Computational Biology Lab, Department of Biosystems Science and Engineering, ETH Zurich, Switzerland* [2]*The Institute of Scientific and Industrial Research, Osaka University, Japan* [3]*JST, PRESTO, Japan* [4]*Department of Molecular Biology, Max Planck Institute for Developmental Biology, Germany*

contact email address: udo.gieraths@bsse.ethz.ch

*accepted for oral presentation at ISMB 2015

ABSTRACT Genetic heterogeneity is the phenomenon that several distinct sequence variants may give rise to the same phenotype (Burrell *et al.*, 2013). This phenomenom is of the utmost importance to the exploration of the genetic basis of complex phenotypes, as most of them have been found to be affected by numerous loci, rather than a single locus (McClellan and King, 2010).

Current approaches for finding regions in the genome that exhibit genetic heterogeneity suffer from at least one of two shortcomings: 1) they require the definition of an exact interval in the genome that is to be tested for genetic heterogeneity, potentially missing intervals of high relevance, or 2) they suffer from an enormous multiple hypothesis testing problem due to the large number of potential candidate intervals being tested, which results in either many false positive findings or a lack of power to detect true intervals.

To illustrate the scale of this multiple testing problem in genetic heterogeneity search: When one considers all possible intervals in a genome in a dataset with 10^6 SNPs, the number of tests one performs is quadratic in the number of SNPs, that is approximately $5 \cdot 10^{11}$ candidate intervals. When ignoring the multiple testing problem, one will obtain billions of false positives. If one performs the standard Bonferroni correction (Bonferroni, 1936), which divides the significance threshold α (typically 0.05 or 0.01) by the number of tests, then the corrected threshold will be so low that hardly any finding will be statistically significant.

We propose an algorithm for genome-wide detection of contiguous intervals that may exhibit genetic heterogeneity with respect to a given binary phenotype. More specifically, we search for genomic intervals in which the occurrence of at least one type of sequence variant (e.g. a point mutation or minority allele) is significantly more frequent in one of the two phenotypic classes. Figure 1 illustrates this matter.

Our algorithm, Fast Automatic Interval Search (`FAIS`), automatically finds the starting and end positions of these intervals, while properly correcting for multiple hypothesis testing and preserving statistical power. Central to this algorithm is an approach by Tarone (Tarone, 1990), which allows one to reduce the Bonferroni correction factor for multiple hypothesis testing. Additionally we extended `FAIS` to a Westfall-Young permutation based version called `FAIS-WY`. In practice, `FAIS-WY` is more computationally demanding than `FAIS` but has increased statistical power.

We employ our novel algorithms on simulated data as well as on *Arabidopsis thaliana* GWAS data. In the simulations our algorithms outperform in terms of

power the brute force approach using Bonferroni correction as well as an approach using univariate Fisher's Exact Test (UFE) that only checks for a significant difference in single SNPs. For the *Arabidopsis thaliana* GWAS data, out of 21 binary phenotypes we were able to discover intervals of SNPs that are associated with 14 of these phenotypes, but could not be found with previous methods. The comparison is done to the univariate Fisher's Exact Test (UFE) and a state-of-the-art linear mixed model (LMM) to account for confounding due to population structure (Lippert *et al.*, 2011). The Proportion of novel intervals among all intervals found by `FAIS-WY`, across all phenotypes is visualized in figure 2.

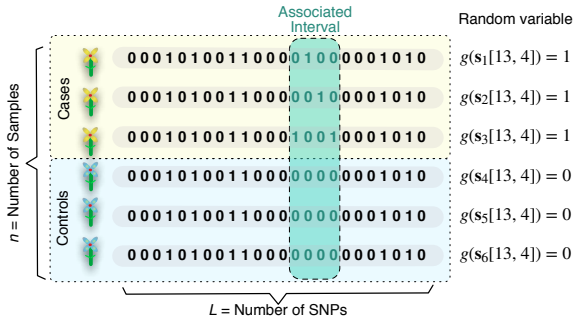


Figure 1: Schematic illustration of the problem of detecting genomic intervals that may exhibit genetic heterogeneity. $\mathbf{s}_i[\tau; l]$: interval of length l, starting at index τ of the i-th genomic binary sequence , $g(\mathbf{s}_i[\tau; l]) = \mathbf{s}_i[\tau] \vee \mathbf{s}_i[\tau+1] \vee \ldots \vee \mathbf{s}_i[\tau+l-1]$, where $\vee$ denotes the binary OR operator. The problem to solve is that of finding all intervals (τ, l) with $l = 1, \ldots, L$ and $\tau = 0, \ldots, L-l$ such that the random variable $g(\mathbf{s}[\tau; l])$ is statistically associated with the phenotype $y \in \{\text{Cases}, \text{Controls}\}$ after correction for multiple hypothesis testing.

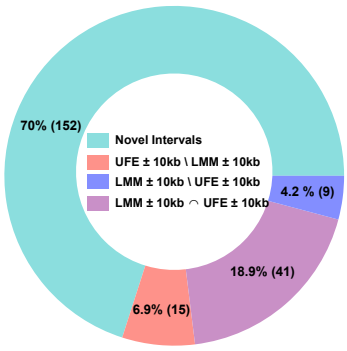


Figure 2: Proportion of novel intervals among all intervals found by `FAIS-WY`, across all phenotypes. The green part shows the proportion of novel intervals found by `FAIS-WY`. The red part (UFE ± 10kb \LMM ± 10kb) are intervals containing an UFE hit or are in close proximity (± 10kb) to one and the hit could not be found with a LMM. The blue part (LMM ± 10kb \UFE ± 10kb) are intervals containing a LMM hit or are in close proximity (± 10kb) to one and the hit could not be found with an UFE. The purple part (LMM ± 10kb ∩ UFE ± 10kb) are intervals that contain both, a hit (± 10kb) found with an UFE and a LMM.

References:

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3–62.

Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, **501**(7467), 338–345.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nat Meth*, **8**(10).

McClellan, J. and King, M.-C. (2010). Genetic heterogeneity in human disease. *Cell*, **141**(2), 210–217.

Tarone, R. E. (1990). A modified bonferroni method for discrete data. *Biometrics*, **46**(2), 515–522.

New Gene Ontology term similarity measure - comparison and performance evaluation based on DNA microarray data

Aleksandra Gruca [a], Michał Kozielski [b]
[a], *Institute of Informatics, Silesian University of Technology, Gliwice, Poland*
[b], *Institute of Electronics, Silesian University of Technology, Gliwice, Poland*
aleksandra.gruca@polsl.pl

Rapid evolution of high-throughput technologies provides us with more and more data and development of automated tools for data interpretation is necessary in order to process and understand results of such experiments. The main goal of the presented research is to analyze if and how information derived from the Gene Ontology (GO) database can be useful in the automated process of interpretation of gene groups obtained in expression level analysis.
A number of gene similarity measures based on Gene Ontology can be found in the literature but there is still a lack of complete studies that compare their performance. The first objective of this work is to propose new relatives-based GO terms similarity measure based on a granular approach and which allow comparing genes on a more general level. The second objective is to analyze existing similarity measures, compare them and evaluate in terms of clustering and correlation quality. We assume that good and efficient measure should reflect biological dependences among genes, therefore our conclusions are based on comparison with expression data from two different microarray experiments.

Following GO term similarity measures were analyzed and compared:

- Semantic term similarity
  - Information content
  - Jiang-Conrath
  - Lin
  - GraSM
  - G-SESAME
  - **Group and Group-soft relatives-based granular term similarity – new measures proposed**
- Path-based term similarity
- Binary Similarity
  - Jaccard measure
  - Czekanowski measure

Group and Group-soft are two new methods of Gene Ontology term similarity calculation based on the idea of granular analysis in order to compare ontology terms on more abstract and general level. In proposed approach, not a pair of terms is compared, but a pair of granules (sets) related to these terms is analyzed.

In the presented research we compare gene similarity in two representations: gene expression values and Gene Ontology graph. The rationale leading to such comparison is that genes that act in the same way (fact translating into similar expression patterns) should be similar in other representations, e.g., annotations to Biological Process Gene Ontology. Two types of analysis were performed:

- correlation of gene similarity in gene expression representation and GO representation,
- clusterability of the Gene Ontology data and comparison of clustering results in both representations,

From the clustering results perspective, gene similarity measures were used as a similarity/distance measures. Such analysis can show which similarity/distance measure gives the values making data objects more cohesive within a group and more easily separable between the groups, in other words, which measure gives a more clusterable data representation.

Two DNA microarray datasets were analysed: Eisen (Eisen et al, PNAS 1998) and Iyer (Iyer et al., Sciene 1999). The correlation and clustering quality results are presented in Table 1.

Table 1. Results of correlation and clustering analysis

	Correlation analysis		Clustering quality (NMI)	
	Eisen	*Iyer*	*Eisen*	*Iyer*
Binary Czekanowski	0.483	0.102	0.468	0.092
Binary Jaccard	0.475	0.119	0.468	0.092
Group	**0.571**	**0.124**	**0.569**	**0.103**
Group Soft	**0.572**	**0.136**	**0.705**	**0.109**
GSezame	0.522	0.088	0.526	0.072
Jiang-Conrath	0.412	0.104	0.518	0.109
Jiang-Conrath GraSM	0.427	0.112	0.597	0.121
Lin	0.36	0.088	0.444	0.123
Lin GraSM	0.385	0.103	0.526	0.103
Path	0.572	0.136	0.603	0.095
Resnik	0.458	0.085	0.45	0.092
Resnik GraSM	0.467	0.091	0.544	0.073
Weighted Czekanowski	0.477	0.11	0.592	0.094
Weighted Jaccard	0.461	0.125	0.592	0.094

Finally, to verify if the gene clusters obtained for the best measure (**Group Soft**) do have biological meaning, we analyzed their gene composition and compared the results with the reference partition for Eisen DNA microarray dataset. Analysis shows that our clusters have similar gene composition. In case of original cluster C (described by Eisen keyword Proteasome) and our group 7 we obtained identical partition. For other groups, differences were more visible, however typically it was not more than a few genes. In several cases we obtained group that consisted of reference groups merged together – for example gene composition of our group 1 is: CDC10, HTB2, HTB1, HHF1, HHF2, HTA2, HHT1, HHT2, HTA1, MCM7, DBF2, MCM4, MCM3 which mostly covers two Eisen groups: H which consist of genes: HTB2, HTB1, HHF1, HHF2, HTA2, HHT1, HHT2, HTA1 and J which consists of genes: MCM7, DBF2, MCM4, MCM3, MCM2. This result can be explained by the following facts. If we analyze the original dendrogram we can notice that genes composing clusters J and H are placed next to each other, therefore depending on selected cut-off value we can obtain one or two clusters. Another explanation of merging two clusters can be found by analyzing genes function. Original cluster H was described by Eisen by a keyword *chromatin structure* and includes, among others, genes HHF1, HHF2, HHT1, HHT2 that contribute to telomeric silencing. If we analyze biological function of MCM3 and MCM7 genes we can see that they also play a role in silencing and interact with the essential silencing chromatin factor, SIR2

In silico phenotyping via co-training for improved phenotype prediction from genotype*

Damian Roqueiro [1†], Menno J. Witteveen [1†], Verneri Anttila [2,3,4], Gisela M. Terwindt [5], Arn M.J.M. van den Maagdenberg [5,6], Karsten Borgwardt [1]

[1] *Machine Learning and Computational Biology Lab, Dept. of Biosystems Science & Engineering, ETH Zurich, Switzerland* [2] *Analytical and Translational Genetics Unit, Dept. of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA* [3] *Program in Medical and Population Genetics and* [4] *Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA* [5] *Dept. of Neurology and* [6] *Dept. of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands*

*Accepted for oral presentation at ISMB 2015. Contact: menno.witteveen@bsse.ethz.ch
†Both authors contributed equally to this work

ABSTRACT

Predicting disease phenotypes from genotypes is a key challenge in medical applications in the postgenomic era. Large training datasets of patients that have been both genotyped and phenotyped are the key requisite when aiming for high prediction accuracy. With current genotyping projects producing genetic data for hundreds of thousands of patients, large-scale phenotyping has become the bottleneck in disease phenotype prediction.

Here we present an approach for imputing missing disease phenotypes given the genotype of a patient. Our approach is based on *co-training*, which predicts the phenotype of unlabeled patients based on a second class of information, e.g. clinical health record information. Augmenting training datasets by this type of *in silico* phenotyping can lead to significant improvements in prediction accuracy. We demonstrate this on a dataset of patients with two diagnostic types of migraine, termed migraine with aura and migraine without aura, from the International Headache Genetics Consortium.

Imputing missing disease phenotypes for patients via co-training leads to larger training datasets and improved prediction accuracy in phenotype prediction.

Motivation Co-training (Blum and Mitchell, 1998) is an instance of semi-supervised learning, which is often employed in scenarios where the number of labeled examples ($\mathcal{L}$) is low and the number of unlabeled instances ($\mathcal{U}$) is large. The reason for this imbalance is simply due to the high cost of labeling the data. The co-training method benefits from a natural split of the feature space. An instance x is described by the set $\mathcal{X}$ of all features, comprised of two mutually exclusive "views" $\mathcal{X}_1$ and $\mathcal{X}_2$. A labeled object x is referenced as $((x_1, x_2), y)$ where x_1 and x_2 are the values for the features in $\mathcal{X}_1$ and $\mathcal{X}_2$, and y is the class label. The algorithm then learns two classifiers h_1 and h_2, one for each view of $\mathcal{L}$, followed by an iterative bootstrapping in which instances of $\mathcal{U}$ are labeled and the most confident ones are moved to $\mathcal{L}$.

In this study, and following the spirit of co-training, the two exclusive views of the data are the clinical covariates and the genotype data of patients with one of two different types of migraine: a) migraine with aura and b) migraine without aura.

Results Our analysis was conducted by partitioning the entire dataset into three groups: Set **I**, the *training dataset*: contains a subset of the patients for which all available information is present, i.e.: a disease phenotype, a set of clinical covariates and genotype data in the form of single-nucleotide polymorphisms (SNPs); Set **II**, the *co-training dataset*: similar to the training set but with a much larger number of patients. Here the patients lack a disease phenotype (unlabeled); Set **III**, the *evaluation dataset*: is used to evaluate the method. It does not contain clinical covariates. This is depicted in Fig. 1.a

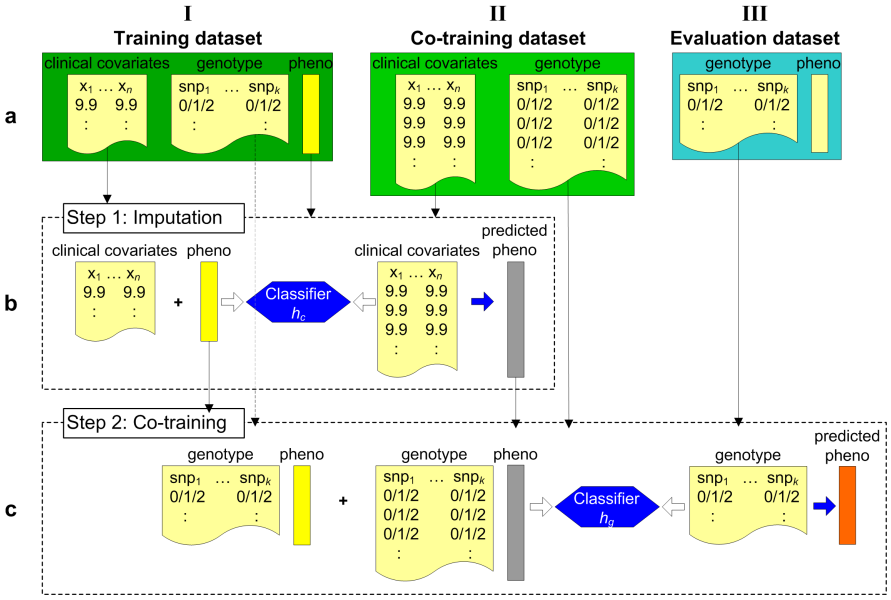


Figure 1: Data partioning and the proposed two-stage approach to co-training

The algorithm was then applied in two sequential steps (Fig. 1.b-c): Step 1: predict a disease phenotype for the patients in set II by learning a classifier h_c from the clinical covariates of the patients in set I; Step 2: the previous predictions are used to augment the pool of labeled examples. Then, a genotype classifier h_g is constructed via co-training. Finally, h_g is tested on III to obtain an AUC score.

Four metrics were used to compare the prediction performance of the algorithm. These metrics corresponded to different cases that ranged from using the least possible amount of data for training (to compute a lower bound) to using all available data (upper bound). Between these two ranges, the actual prediction performance was reported and all these values are shown in Table 1.

Table 1: Bounds and prediction performance of *in silico* phenotyping. Partition of the data into: set I = 10%, set II = 70% and set III = 20%; 100 random folds.

	AUC scores	
Metric	μ	σ
Lower bound, training only on I	0.574	0.034
Univariate featire selection on I, training on I+II	0.608	0.035
In silico phenotyping (co-training)	0.646	0.029
Upper bound, I+II with true labels	0.689	0.025

References

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pages 92–100, New York, NY, USA. ACM.