A three-state model for multidimensional genomic data integration

Baca-Lopez K[1], *Correa-Rodriguez MD*[1], *Flores-Espinosa R*[1], *Garcia-Herrera R*[1], *Hernandez-Armenta CI*[1], *Hidalgo-Miranda A*[2], *Huerta-Verde AJ*[1], *Imaz-Rosshandler I*[1], *Martinez-Rubio AV*[1], *Medina-Escareno A*[1], *Mendoza-Smith R*[1], *Rodriguez-Dorantes M*[2], *Salido-Guadarrama I*[2], *Hernandez-Lemus E* [PI] *and Rangel-Escareno C*[PI]
(1) Computational Genomics Department, (2) Cancer Genomics Lab, National Institute of Genomic Medicine Mexico

All authors contributed equally alpha-ordered by last name, PI - Principal Investigator *corresponding author: crangel@inmegen.gob.mx

Introduction

Modern high-throughput genomic technologies have allowed large-scale characterization of living organisms, involving the generation and interpretation of data at an unprecedented scale. Computational tools and mathematical algorithms have been created aiming to integrate, organize and mine the wealth of information generated. Technologies for the detection of different types of genomic alterations have been developed and applied to analyses of almost any living organism, but also of cancer genomes. In cancer research in particular, It is clear now that studies based on a single technology platform are limited compared with the extent of knowledge that can be acquired when using different platforms all together. Hence, there is a need for systematic methodologies to facilitate data management, visualization and integration. Such methodologies should aim to permit a proper analysis of the biological implications for the findings, without sacrificing mathematical and statistical rigour and computational efficiency. The present *three-state model for multidimensional data integration* (**3-MDI**, in short), a data driven approach, has been designed and implemented with such purpose in mind.

Materials and Methods

The Cancer Genome Atlas Network (TCGA) offers a variety of datasets including expression profiling, targeted sequencing and copy number of a large number of tumor and normal samples of Glioblastoma Multiforme (GBM). Low-level analyses, classification and selection for all platforms including quality control, background correction analysis and normalization were performed using [R] and BioConductor.

Level 1 mRNA Low level analysis of raw gene expression microarray data generated from Affymetrix HGU133A of 495 tumor samples and 10 controls were normalized using quantile normalization [1] and summarized using median polish, both methods from the `affy` library. Classification was based on log fold-change, B-statistic and adjusted p-values using `limma` package.

Level 2 miRNA Agilent miRNA 8x15K, of 245 tumor samples and 10 controls. According to TCGA portal data were background corrected using *RMA* and quantile normalized. It was not clear whether the controls and the tumors were normalized together and since boxplots of the data showed a discrepancy between the groups all samples were normalized using *median absolute deviation* (MAD)

Level 3 CNVs Data for 461 samples processed with array CGH technology. Data reported to be lowess normalized. Regions of gain and loss were identified using Circular Binary Segmentation algorithm.

Level 2 Meth Methylation data from 291 tumor samples and 1 control with 6 replicates were normalized and processed using genome wide Infinium HumanMethylation27 BeadChip Array (Illumina, Inc., San Diego, CA, USA) with 27,578 CpG sites. Beta-values and confidence p-values were further examined. Missing beta-values were calculated using the signal intensity (M) and the unmethylated signal intensity (U).

Level 3 NGS Sequencing data of somatic nucleotide alteration data for 143 samples in 3 databases were analyzed. The three databases were combined and relevant mutations were selected. The final database contained 1032 unique gene-mutation pairs, for 500 different genes and 7 different mutation types: Missense, Frame_Shift, Silent, Nonsense, In_Frame, Splice_Site and Unknown.

Strategy for integrative analysis

To construct an integrated view of genomic alterations, that here we apply to the glioblastoma genome; we propose a data driven combinatorial approach that lists all possible scenarios of genomic alterations in a N-platform integrative analysis based on a three-state model applied to statistically significant genes (3-MDI). Each scenario is represented as a

sequence $S_1, S_2, ...S_N$ of states, where S_k denotes the state of a gene for platform k. Each state is defined to take values in {-1, 0, 1} interpreted as {*Down, NoChange, Up*}. This list represents the universe of hypotheses that describe structural variations in the genome as well as transcription activity in coding and non-coding regions. Hypotheses can be chosen for their clear biological relevance but also for their quantitative importance. We may find that a large set of genes follow a particular scenario or that genes commonly share a set of more specific scenarios leading to other important questions to be answered.

For every platform, low-level analysis was carefully performed, genes were classified according to the proper significance levels for each technology. The platforms (P) selected included {*mRNA, miRNA, Methylation, NGS, CNV*}, a list of highly relevant genes was generated from each one. Genes in every list were coded according to two of the three states in the three-state model. In every list of top-m_k genes can be either up {1} or down {-1}. All k lists are combined and basic set theory $P_i \cup P_j = (P_i \cap P_j) \cup (P_i \setminus P_j) \cup (P_j \setminus P_i)$ adds zeros when $(P_i \setminus P_j)$ indicates a zero in P_j. Using these approach the combinatorial analysis can classify all possible scenarios searching for genes in 2,3,4, or all 5 platforms simultaneously.

Under this approach there exist 3^k possible scenarios for a k-platform analysis assuming a three-state model. If we wish to perform a more exhaustive analysis adding levels of information by including a new platform at every level of the analysis, we would have $\sum_{i=1}^{h} \binom{k}{h}$ possible combinations of (k) platforms, giving us a total of

$$\sum_{h=1}^{h} 3^k \binom{k}{h}$$

Even though the hypothesis space grows it is clear that we may not find the full list of scenarios for various reasons:

- There are scenarios whose sequence of states is non-informative. For instance, any of the {0, 00,000, 0000, 00000} which basically reports no change in any platform.
- There will be scenarios with an empty set
- Finally, some scenarios may not have a true biological meaning. For which in further detail may be used to detect possible false positives

Finite alphabet classification and supervised learning

It should be noted that the discretization scheme proposed here (a three-state model) is, within the limitations of such model, a finite alphabet classification scheme. This means that all of the possible states of the system correspond to a realization or string constructed within the lexicon of such finite alphabet (here the states are labelled -1 meaning sub-regulated, hypomethylated or deleted, 0 meaning no change and $+1$ meaning over-expressed, hypermethylated or duplicated (multiplicated)) are considered, i.e. the sample space is exhaustive. Finite alphabet classification schemes have been shown to be equivalent to a class of machine learning algorithms called *Supervised learning* [7], this means that most (in theory, all) of the supervised machine learning techniques and algorithms could be applied to train our three-state model with either experimental or simulated data. This could result highly useful when assessing the findings of an integrative genomics study either with other experimental sources or with synthetic data.

Results

Level 1 and Level 2 data sets were pre-processed and targets were selected based on the following statistics:

	Meth	mRNA	miRNA
No.	972	2971	91
logFCh	(0,1)	(-4.6, 3.4)	(-0.85, 1.32)
B-stat	(0,266)	(0, 412)	na
adj-p	(3E-117, 7.3E-4)	(1.4E-180, 2.5E-4)	na
p-value	(3.4E-121, 6.9E-5)	na	(0.0068, 0.406)

Selected targets for mRNA and methylation were coded with the three-state model and combined with the level 3 somatic mutations data set into a single list. The list of somatic mutations was re-arranged by genes and coded in a 2-state format $\{0, 1\}$ indicating presence or absence of mutation to avoid an *ad-hoc* threshold for a classification of hypo/hyper mutated.

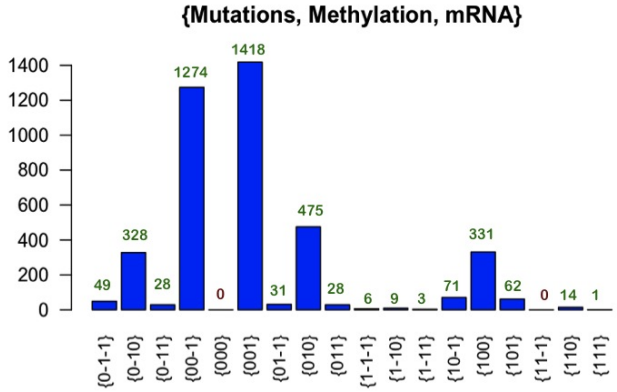

Figure 1. Each bar represents the number of genes present in each scenario, labels on the x-axis represent the status of genes in each of the three platforms chosen. So, {1,0,-1} indicates that 71 genes have mutations, do not report changes in methylation and were down-regulated.

Using these three platforms we could find up to 18 possible scenarios for further analysis. As can be seen in Figure 1, differential changes by single platform take the larger counts, and numbers are lower when platforms are combined.
Scenarios with presence of somatic mutations were selected and that led to 6 different ones with only one with 0 targets reported.

Mutation	Methylation	mRNA	Total of genes
1	0	1	62
1	0	-1	71
1	1	1	1
1	-1	-1	6
1	-1	1	3
1	1	-1	0

The combined list with 143 genes was furthered analyzed for mutation type, mutation rate, miRNAs associated, copy number aberrations and enrichment analysis. The top 10 in the list ordered by mutation rate included: *TP53(38), EGFR(17), RB1(11), PIK3R1(11), SYNE1(7), BCL11A(6), FN1(5), PRKDC(5), COL3A1(4), MSHG(4)*

Visualization of large-scale data becomes a key aspect of the analysis, it allows to distinguish possible biological hypotheses. A circos plot helps us to identify the most mutated chromosomes, identifying genes in each chromosome and hence combine the differential expression and methylation for the set under study.

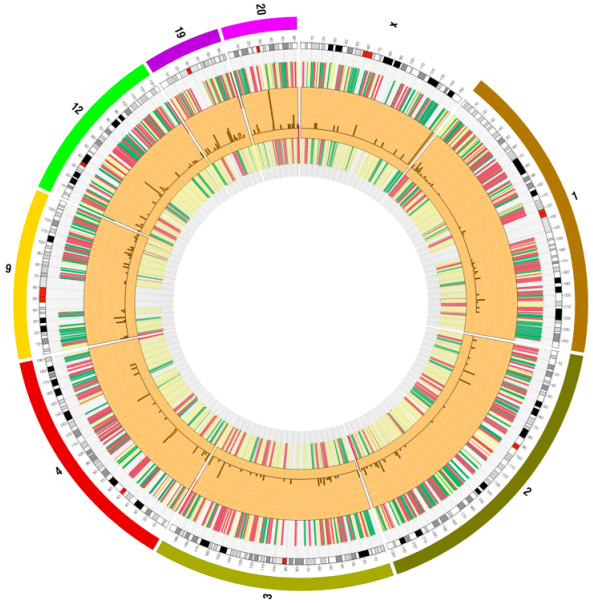

Figure 2. Circos plot showing the most mutated chromosomes. The outside ring shows the cytogenetic bands of the chromosome, the second ring shows mRNA differential expression. The third ring shows mutation rate histograms and the last one differential expression of methylated genes.

The list of genes specifying the mutation types as well as differential expression and changes in methylation is displayed in Figure 3. The correlation of the three platforms is now combined with the different type of mutations shown in a color heatmap. The barplot on the left shows the log fold-changes of differential expression for the significant genes. The right barplot shows the log fold-change for methylated genes. On the bottom are patients which allows us to identify individuals with a large number of genes with mutations.

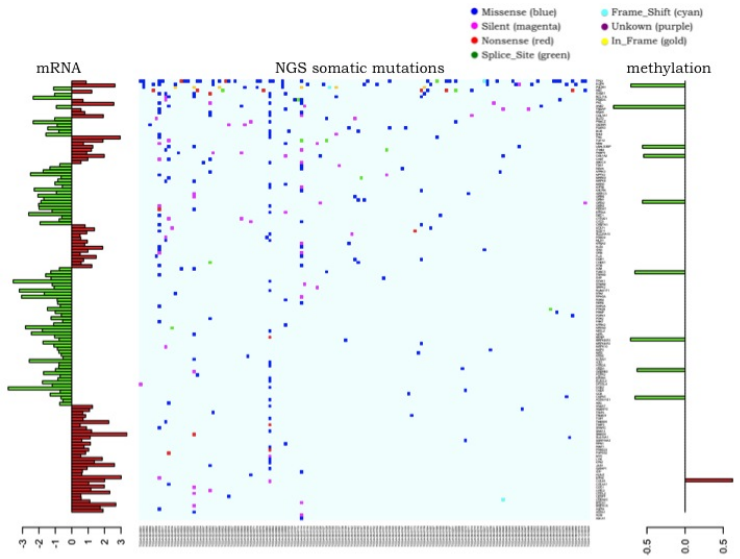

Figure 3. Graphical representation of changes in gene expression, methylation and mutation type per gene, per tumor.

An extra dimension with clinical data can be added on the top of the heatmap in Figure 3 providing with more layers of information in a single plot.

DNA Methylation-miRNA Network analysis

3-MID model explore the relationship between miRNA expression levels and DNA gene methylation status in CpG islands comparing different conditions (tumor vs normal). The Network analysis shows eleven miRNAs highly related (putative Target) to genes over-represented with changes in the CpG methylation status. Four databases were used to map mRNA genes with miRNA targets: Targetscan, Miranda, Pictar and Mirbase; relationships with consistency of 3 of 4 databases were selected. Pathway enrichment analysis shows only a few pathways significantly related to biological processes likely involved in neuronal functions (eg. Axon guidance, synaptical transmission). The 3-MDI model presents several options of novel configurations for genomics analysis using the state matrix design. Some of them may not have an apparent biological interpretation but perhaps the combination of some could lead to new hypotheses.

In this case, configuration of the 3-MDI models could suggest that DNA methylation and modifying expression by miRNA are closely related by a particular states combination which corresponding to a biological characteristic pattern. Biological information could be drawn from this analysis showing that different intersections can be established for new approaches to the glioblastoma studies. For example new gene pathways regulated by concerted mechanism between miRNAS and DNA methylation.

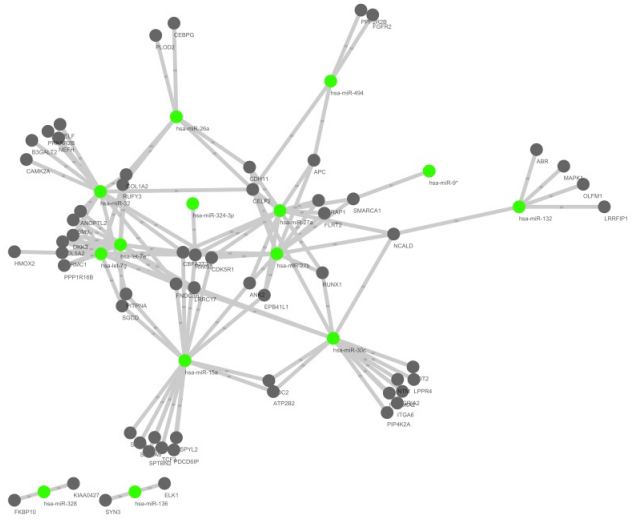

Figure 4. Network of the selected set of 143 genes. Green nodes represent miRNAs and gray nodes represent genes

Enrichment analysis

Pathway enrichment analysis was studied by means of statistical over-representation of Pathway entries in the Reactome databases for the set of 143 genes. Significance assessment was made by means of 'urn model hypergeometric distribution tests. Here, testing a pathway amounts to drawing the genes annotated at it from the urn (gene universe) and classifying them as to whether they belong to a certain pathway or not. Then by counting and calculating the resulting proportions we can perform a significance test, in this case the hypergeometric test (which is equivalent to a one-tailed Fisher's Exact test). In order to correct for multiple testing we used the Benjamini-Hochberg algorithm or False Discovery rate (FDR) [8]. Only associations with corrected p-values below 0.05 were considered significant. The most significant pathways (p-values in parenthesis) included *Hemostasis (1.75E-05), Formation of Platelet plug (5.64E-05), Cell surface interactions at the vascular wall (7.71E-05), ATM mediated response to DNA double-strand break (3.55E-05), Axon guidance (0.0127)* among a larger list.

Discussion

3-MDI as a tool for finding integrative effects in genomic regulation

As we already stated 3-MDI can predict several kinds of novel configurations for genomics analysis, incorporating two or more sceneries taken from the state matrix design. Hence, biological information could be drawn from this analysis showing that different intersections can be established for new approaches to the glioblastoma studies. For example, new gene pathways regulated by *concerted* mechanism between miRNAS and DNA methylation.
By using 3-MDI we have explored the combined effect that miRNA differential expression levels and DNA gene methylation status in CpG islands have on differential gene expression by comparing different conditions (tumor vs normal) in a network based analysis. Results (as showed in Figure 4 could suggest that DNA methylation and miRNA transcriptional regulation are closely related for a particular state-vector representing a novel characteristic pattern. For instance, this analysis shows eleven miRNAs related (as putative targets) to genes over-represented with respect to changes in the CpG methylation status. That is, genes whose methylation profiles and miRNA targeting status may potentially affect their corresponding mRNA expression levels. Pathway enrichment analysis using GO for this set of 11 genes shows only a few pathways significantly enriched in biological processes. It is interesting to notice that they were mainly involved in neuronal functions (eg. Axon guidance, synaptical transmission).

Noise classification

In order to build proper schemes of data integration and analysis, an important issue (especially for statistical and probabilistic modeling) is that of inferring the role that different kinds of noise will play and how these effects aggregate in the integrated setting. It has been possible to distinguish at least five different sources of noise/variability in the considered *'omics* experiments:
• Technical processing noise
• Biological intersample variability
• Batch effects
• Bimodal distributions between already normalized cases/controls
• Dynamic range incommensurability between different technologies
A proper characterization of noise in these instances would greatly improve the performance of integrative schemes for massive datasets such as the ones considered here.

Relative importance of the findings and probability measures

In order to compare the relative effect that different genomic variations could have on different phenotypes (say, diseased Vs healthy conditions, tumor progression stages, types of tumor and so on) we need to take into account the fact that the dynamic ranges of the quantitative measurements differ dramatically. For example, in the case of gene expression experiments, intensity levels or even gene expression signals show a great variability with regards to its range, including positive and negative values of the indicators (e.g. $\log_2$-ratios of fold-change) under a non-symmetrical distribution, whereas in the case of methylation profiles, these are usually characterized by using a β-value as an indicator. β-values are Borel-normalized (i,e, $\beta \in \mathscr{B}[0,1]$). In the other hand, mutation rates and copy number variations also present highly dissimilar dynamic ranges. It appears obvious that these quantitative measurements could not be compared directly with each other so as to establish relative importance for a given phenotype under a systematic (computationally tractable) integrative scheme.
Our purpose is to normalize the different indicators by using the experiment-wise empirical distributions, i.e. by setting the scale such that if x is a non-normalized indicator with realizations or measurements x_i, then its normalized version would have a scale ranging between $0 \sim \frac{\min x_i}{\max x_i}$ and $1 \sim \frac{\max x_i}{\max x_i}$ and then normalizing all other values by homogeneous distribution over this scale.

Normalized mutation rates

Most of the methods used to characterize mutation rates in NGS experiments, report mutations on an individual gene or chromosomal segment basis. Yet, if we consider the wide range of size of genes and chromosomes, it may be more appropriate to report mutation rates on a size-normalized scale. As other groups already take this into account.

CpG islands and methylation profiles

An interesting analysis that could be done to undercover different patterns of epigenetic activity, would be the third-way combined study of gene expression, methylation profiles and CpG island structure (from NGS experiments). This could be implemented either for a set of well known *epigenetically driven cancer genes* or on a whole-genome basis. However due to the fact that CAMDA-challenge available datasets due not include raw sequence or even BAM files, but just level 3 data (regions classified either as mutated or non-mutated) we have not performed (yet?) this quite interesting analysis.

Conclusions

The use of data driven approaches instead of biological hypothesis motivated ones, explores a multidimensional genomic analysis that might be able to find relevant genes that are likely to be overlooked mostly due to the fact that they remain largely unexplored. It is known, that each additional genomic dimension increases both, the amount of information and, consequently the biological and computational complexity of the analysis. We present a model, 3-MDI that integrates several technological platforms visualizing and prioritizing different biological scenarios thus enabling the researcher to pursue in an educated way some of or all these possibilities.

References

[1] Bolstad, B.M., Irizarry R. A., Astrand M., and Speed, T.P. **(2003)**. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics* 19(2):185-193.

[2] Gentleman R., Carey V., HUber W., Irizarry R., Dudoit S. Bioinformatics and Computational Biology Solutions Using R and Bioconductor **2005** Statistics for Biology and Health - Springer.

[3] Irizarry Rafael A. , Bolstad Benjamin M., Collin Francois , Cope Leslie M., Hobbs Bridget and Speed Terence P. **(2003)**. Summaries of Affymetrix GeneChip probe level data *Nucleic Acids Research* 31(4):e15.

[4] Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP **(2003)** Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics* Vol. 4, Number 2: 249-264.

[5] Wit Ernest and McClure John. **(2004)**. Statistics for Microarrays Design, Analysis and Inference. *John Wiley & Sons Ltd.*

[6] Zhijin Wu, Rafael A. Irizarry, Robert Gentleman, Francisco Martinez Murillo, and Forrest Spencer **(2004)** A Model Based Background Adjustment for Oligonucleotide Expression Arrays. *Johns Hopkins Univ, Dept. of Biostatistics Working Papers. Working Paper 1*.

[7] Vapnik, V. N. *The Nature of Statistical Learning Theory* (2nd Ed.), Springer Verlag, 2000

[8] Benjamini, Y.; Hochberg, Y., Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B (Methodological)* 57 ,1, 289-300, (1995).

[9] Martin I Krzywinski, Jacqueline E Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: An information aesthetic for comparative genomics. *Genome Res.* June 18, 2009 doi:10.1101/gr.092759.109