Visualizing Personalized Cancer Risk Prediction

Maulik R. Kamdar

Biomedical Informatics Training Program, Stanford University
maulikrk@stanford.edu

1 Introduction

Cancer, as we now know, is a genetic disease. This knowledge entails that to enable evidence-based personalized diagnosis for any patient, the location of where the tumor occurs (e.g. brain, liver, etc.) is less relevant than the underlying genetic signature that the cancer cells express (i.e. whether genes are silenced, amplified or mutated). With the advent of high-throughput gene sequencing technologies, it has been possible to sequence the entire human genome of disease-diagnosed or normal patients. As a consequence, large repositories of genomic datasets are now available for data analysis, and identifying the underlying patterns could aid in the diagnosis, prognosis and treatment on a personalized basis. The Cancer Genome Atlas[1] (TCGA) publishes data pertaining to the molecular information (Exon (mRNA) expression, DNA Methylation, Single Nucleotide Polymorphisms (SNP), Copy Number Variations (CNV) etc.) and clinical attributes of around 9000 patients tested across different cancer typologies.

Data from TCGA is of high value for oncologists as it enables matching the genomic evidence found in their own patients with those enrolled in the TCGA project. TCGA data has been widely used for hypothesis-driven translational research as all of its data results are from direct experimental evidence. Previous research has been carried out for the discovery of new tumor bio-markers (genes or single nucleotides) using unsupervised learning algorithms, and classification of patient samples using supervised learning methods. However most of these analyses use TCGA Exon Expression or SNP datasets. It has been shown recently that DNA methylation signatures are robust bio-markers, vastly more stable than mRNA or proteins, and hence will extend our ability to classify cancer and predict outcome beyond what is currently possible. This could lead the development of new approaches for diagnosis and prognosis of different kinds of cancer [1]. However, class prediction based on these patterns is an under-determined problem, due to the extreme high dimensionality of the data compared to the usually small number of available samples. Hence, a reduction of the data dimensionality, through a combination of several feature selection methods, is a necessary pre-processing step for classification performance.

1.1 Motivation

The motivation of this project is to lead towards the development of a diagnostic framework, integrated with a cancer genome visualization tool GenomeSnip [2], which enables a clinical researcher to visually predict cancer risk in new patients using machine learning (ML) classification models built over the TCGA DNA Methylation (DM) and Exon Expression (EE) data. The researcher would be able to upload the genomic data for a new patient, load a set of prior classification models and predict cancer risk.

The specific ML goals for the project are:

1. Obtain the set of genomic features which provide the best evaluation metrics (sensitivity, specificity and error) for each typology through **Genomic Co-occurrence-based feature selection**
2. Compare different **Supervised Learning** algorithms built over DM and EE data to predict whether a new sample is *'tumor'* (1) or *'normal'* (0), and also classify Subtypes of Breast Invasive Carcinoma
3. Perform **Principal Component Analysis** (PCA) for generating Cluster visualizations of the data
4. Explore and discuss how well the best classification model developed for one tumor typology can predict tumor risk in a patient experimentally classified under another typology

The models and features so identified, along with their evaluation metrics, could then be made available through the GenomeSnip web application for personalized risk prediction.

2 Methods and Materials

2.1 Data Processing

The raw TCGA data is available as zipped text files from their data portal, and consists of DNA Methylation ($0 \leq x \leq 1$) or Exon expression ($x \in \mathbb{R}^+$) values mapped to a specific position or a range

[1] https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp

of a human chromosome (e.g. chr19:58858635 indicates position 58858635 of Chromosome 19). A pre-processed version is made available by Broad Institute's FireBrowse[2], where these values are mapped to a specific human gene annotated with the HGNC nomenclature[3]. In this work, I have used the DM and EE data for only 6 tumor typologies, the statistics for which are mentioned in the Table below. The sample type (*'tumor'* or *'normal'*) is contained within the TCGA identifier value[4] of the sample, e.g. TCGA-02-0001-**01**C-01D-0182-01. Tumor types range from 01-09 ('1') and normal types range from 10-19 ('0'). Moreover, Tumor Subtypes for a particular patient sample can be determined from the *'histological_sample_type'* attribute present in the TCGA clinical dataset. There were 7 known Subtypes for Breast Invasive Carcinoma. I had to convert the EE data to Log Scale, as a '10' EE value indicates a 10-times *'increased'* expression of a gene, whereas a '0.1' value indicates a 10-times *'decreased'* expression. The EE values were further normalized for having a 0 mean and Unit Variance.

Tumor Type	DM Data Size (GB)	EE Data Size (GB)	#Patients	'1'	'0'	Train	Test
Breast Invasive Carcinoma-BRAC (T_A)	13.47	6.81	876	777	99	657	219
Bladder Urothelial Carcinoma-BLAC (T_B)	6.28	1.6	407	366	41	305	102
Colorectal Adenocarcinoma-COADREAD (T_C)	6.92	2.67	432	382	50	324	108
Head Squamous Cell Carcinoma-HNSC (T_D)	9.38	2.28	579	529	50	434	145
Lung Adenocarcinoma-LUAD (T_E)	8.57	3.58	546	488	58	409	137
Prostrate Adenocarcinoma-PRAD (T_F)	7.69	3.5	537	485	52	403	134

Table 1. Statistics of the TCGA DM and EE Data for the six tumor typologies used for this work

2.2 Genomic Co-occurrence for Feature Selection

$$Sim(Gene_1, Gene_2) = \alpha * \frac{|Dis(Gene_1) \bigcap Dis(Gene_2)|}{|Dis(Gene_1) \bigcup Dis(Gene_2)|} + \beta * \frac{|Path(Gene_1) \bigcap Path(Gene_2)|}{|Path(Gene_1) \bigcup Path(Gene_2)|} + \gamma * \frac{|Pub(Gene_1) \bigcap Pub(Gene_2)|}{|Pub(Gene_1) \bigcup Pub(Gene_2)|} + \omega * BioInput(Gene_1, Gene_2)$$

There are $>$ 20000 protein-coding genes in the human genome, and other auxiliary genomic segments, which increases our feature space $\mathcal{X}(n >> m)$ drastically. Performing traditional methods like **L1-based feature selection** or **Randomized Sparse Models** on the entire set of human genes, for reduction to a smaller set, still incorporated some noise and unrelated features. This resulted in classifiers which were not able to accurately predict cancer risk in newer patients. Biologically, however, it has been shown that only a small number of these genes are actually implicated in the cancer initiation and propagation. The Cancer Gene Census[5] maintains an updated list of HGNC genes, which have been experimentally proven to be implicated in the different types of tumor typologies, using Genome-wide Association Studies. I used this prior biomedical knowledge to reduce the gene list to only 522 genes. However, this list in itself is not exhaustive, as there may be many genes and biomarkers which may be associated with tumor typologies, but have yet to be discovered. I used a **Genomic Co-occurrence-based method** to obtain the set of genes which co-occur with each of these 522 'oncogenes' in diseases, pathways or publications.

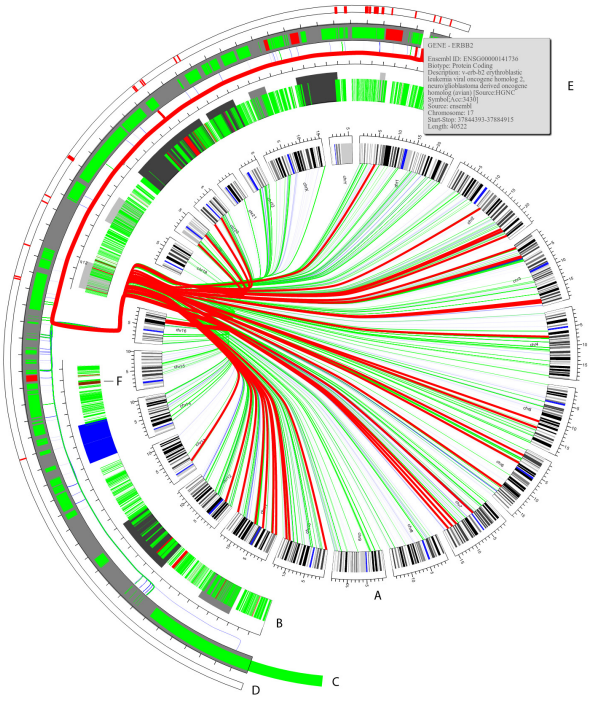

Fig. 1. Genomic Wheel visualization

Using the equation above, I calculate the similarities of each 'oncogene' with the rest of the genes. For instance, the term $\frac{|Dis(Gene_1) \bigcap Dis(Gene_2)|}{|Dis(Gene_1) \bigcup Dis(Gene_2)|}$ indicates the fraction of the diseases where the $Gene_1$

[2] http://firebrowse.org/
[3] http://www.genenames.org/
[4] https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode
[5] http://cancer.sanger.ac.uk/cancergenome/projects/census/

and $Gene_2$ are simultaneously implicated (i.e. their protein products play a major role) with respect to the total number of diseases either of the two genes are implicated. The equation represents a weighted measure by including co-occurrence of genes as observed in pathways and co-mentions in publications. This co-occurrence knowledge pertaining to diseases (*Dis*), pathways (*Path*) and publications (*Pub*) is obtained from OMIM[6] (Online Mendelian Inheritance in Man), KEGG[7] (Kyoto Encyclopedia of Genes and Genomes) and PubMed[8] (knowledgebase of biomedical literature) abstracts. The last term of the equation ($BioInput(Gene_1, Gene_2)$) indicates a user-provided measure, which was obtained by allowing biomedical researchers to interactively click on genes represented in the *'Genomic Wheel'* visualization during the first evaluation of the GenomeSnip platform (Fig. 1) [2], and denote them to be explicitly similar. In the *'Genomic Wheel'*, the *green* and *red* arcs denote normal and cancer genes on the human chromosomes. The chords connecting the different genes, indicate the co-occurrence relations (*red*, *green* and *blue* for disease, pathway and publication co-occurrence respectively). The values of weights used for this work were: $\alpha = 1.0, \beta = 0.33, \gamma = 0.20, \omega = 0.75$. Using a similarity threshold $Sim(Gene_1, Gene_2) > 0.75$, the final number of genes was hence reduced to 1708. Combined with three patient demographic features (gender, age and ethnicity), I had $\#n = 1711$ features. Performing **L1-based feature selection** on this reduced set of genes resulted in a much refined set of features (counts shown in Fig. 2).

2.3 Classification Methods

I primarily have a classification problem for which most of the input features x are continuous-valued random variables and y~Bernoulli(ϕ). I have used four different supervised learning methods:

1. **Gaussian Naive Bayes (GNB)**
   When dealing with continuous data, a typical assumption is that the continuous values of a feature x_j associated with each class k are distributed according to a Gaussian distribution $p(x_j = x|y = k) = \frac{1}{\sqrt{2\pi\sigma_{jk}^2}} \exp\Big[-\frac{1}{2}\Big(\frac{x-\mu_{jk}}{\sigma_{jk}}\Big)^2\Big]$. Training the Classifier over $\{x^{(i)}, y^{(i)}|i = 1, \dots m\}$, I find $P(y = k)$, μ_{jk} and σ_{jk}^2 for each feature x_j and $k \in \{0, 1\}$ ($\{0, 1, \dots, 7\}$ for BRCA Subtypes) using ML estimation.
   $$\hat{\mu}_{jk} = \frac{\sum_i x_j^{(i)} I(y^{(i)} = k) + 1}{\sum_i I(y^{(i)} = k) + m} \quad , \quad \hat{\sigma}_{jk}^2 = \frac{\sum_i (x_j^{(i)} - \hat{\mu}_{jk})^2 I(y^{(i)} = k) + 1}{\sum_i I(y^{(i)} = k) + m}$$
2. **Support Vector Machine (SVM) with a Gaussian Kernel**
   I used the LibLinear Package with L2-Loss L2-Penalty to solve the primal optimization problem.
3. **Decision Tree (DT)**
   Decision Tree over the training set was generated by partitioning data Q at Node m recursively using a candidate split $\theta = (feature\ j, threshold\ t_m)$, i.e. $Q_{left}(\theta) = (x, y)|x_j \leq t_m$, $Q_{right}(\theta) = Q \setminus Q_{left}$. I used the Gini Impurity measure, for $i \in Node\ m$, with N_m examples, and $k \in \{0, 1\}$ (or $\{0, 1, \dots, 7\}$)
   $$p_{mk} = \frac{1}{N_m} \sum_{i \in m} I(y^{(i)} = k) \quad , \quad H(\mathcal{X}_m) = \sum_k p_{mk}(1 - p_{mk})$$
   The classifier finds the optimal $\theta^* = \arg\min_\theta [\frac{n_{left}}{N_m} H(Q_{left}(\theta)) + \frac{n_{right}}{N_m} H(Q_{right}(\theta))]$ and iterates over the partitions until a maximum tree depth of 25 is reached, with at least 5 samples on each leaf.
4. **Ensemble Method Random Forest (RF)**
   For this method, I train $K = 250$ Decision Trees f_k over a sample space $[\mathcal{X}_k y_k]$ obtained by random sampling, with replacement, n training examples from $[\mathcal{X} y]$. For prediction on a new example x', I average the predictions over individual decision trees, i.e. $\hat{f} = \frac{1}{K} \sum_{k=1}^{K} \hat{f}_k(x')$.

2.4 Implementation

TCGA data was downloaded from FireBrowse and processed through three Python scripts - *1)* extract a gene-patient matrix from the DM and EE files for the 1708 genes, *2)* parse the sample types ($y \in \{0, 1\}$) from the identifiers and retrieve the Subtypes and the demographic features from the clinical files, and *3)* generate the final matrix ($[\mathcal{X} y]$). The classification methods were initially tested using Matlab and were re-implemented using Python SciKit Learn. I carried out a **4-fold cross-validation** method and computed the average evaluation metrics (Training error, Test error, Sensitivity and Specificity). The main aim was to achieve a lower Test error and a higher Sensitivity (how accurately the classifier classifies a *'tumor'* sample as '1') for the classifiers. PCA cluster plots were generated using MatPlotLib library.

[6] `http://www.omim.org/`
[7] `http://www.genome.jp/kegg/`
[8] `http://www.ncbi.nlm.nih.gov/pubmed`

3 Results

The training error, test error, sensitivity and specificity were calculated for each Supervised Learning algorithm trained using DM or EE data, across each typology using 4-fold cross validation.

	Training Error (%)						Test Error (%)						Sensitivity (%)						Specificity (%)					
	T_A	T_B	T_C	T_D	T_E	T_F	T_A	T_B	T_C	T_D	T_E	T_F	T_A	T_B	T_C	T_D	T_E	T_F	T_A	T_B	T_C	T_D	T_E	T_F
GNB	1.28	1.38	0.0	0.63	0.41	6.82	1.68	2.07	0.70	1.55	1.04	7.38	99.0	99.0	100	100	100	93.0	97.0	81.0	93.0	86.0	91.0	86.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	1.44	2.07	0.23	1.04	0.62	4.22	99.0	99.0	100	99.0	99.0	97.0	96.0	81.0	99.0	94.0	100	88.0
DT	0.0	0.0	0.0	0.0	0.0	0.0	4.45	4.65	1.64	4.32	2.49	6.33	97.0	98.0	99.0	98.0	98.0	96.0	85.0	57.0	93.0	76.0	84.0	76.0
RF	0.0	0.0	0.0	0.0	0.0	0.0	1.56	2.33	0.70	1.38	0.41	4.85	99.0	100	100	100	100	97.0	91.0	57.0	93.0	84.0	94.0	76.0

Table 2. Evaluation Metrics (%) for Classifiers trained using DNA Methylation Data

The Evaluation metrics (in %) for the classifiers trained using TCGA DNA Methylation Data are shown in Table 3. It can be seen that the SVM and the RF classifiers have minimal Test errors ($< 5.0\%$), 0.0% Training errors, and a very high Test Sensitivity ($> 97.0\%$) across all tumor typologies. Moreover the SVM Classifier has a considerable Specificity also with only two tumors having $< 90.0\%$. Overall, all the classifiers did reasonably well with Test errors $< 8.0\%$ and Sensitivity generally $> 95.0\%$.

	Training Error (%)						Test Error (%)						Sensitivity (%)						Specificity (%)					
	T_A	T_B	T_C	T_D	T_E	T_F	T_A	T_B	T_C	T_D	T_E	T_F	T_A	T_B	T_C	T_D	T_E	T_F	T_A	T_B	T_C	T_D	T_E	T_F
GNB	35.5	1.78	15.62	8.76	25.59	13.72	37.17	4.53	18.36	14.68	21.51	16.2	62.0	98.0	87.0	89.0	76.0	86.0	71.0	42.0	44.0	55.0	57.0	67.0
SVM	0.0	0.0	0.0	0.0	0.0	0.0	10.26	4.53	10.63	6.48	8.78	9.12	94.0	97.0	93.0	96.0	96.0	94.0	59.0	58.0	60.0	74.0	50.0	60.0
DT	0.0	0.0	0.0	0.0	0.0	0.0	19.95	10.67	20.77	14.33	16.82	15.27	88.0	94.0	87.0	92.0	91.0	92.0	17.0	11.0	24.0	32.0	21.0	19.0
RF	0.0	0.0	0.0	0.0	0.0	0.0	11.63	5.07	11.84	10.58	10.79	9.68	100	100	98.0	100	100	100	0.0	0.0	14.0	0.0	0.0	0.0

Table 3. Evaluation Metrics (%) for Classifiers trained using Exon Expression Data

Similar success was not found when training the classifiers using the TCGA Exon Expression Data, as seen in Table 3. GNB had exceptionally huge Training and Test errors for two typologies ($> 20.0\%$), whereas most classifiers had $> 10.0\%$ Test errors and $< 50.0\%$ Specificity. Out of the four, SVM functioned comparatively well with a maximum of ~10.0% Test Error, $> 93.0\%$ Sensitivity and $> 50.0\%$ Specificity.

T_A	DM Training Error (%)	DM Test Error (%)	EE Training Error (%)	EE Test Error (%)
GNB	0.16	0.12	0.19	0.46
SVM	0.04	0.24	0.0	0.11
DT	0.0	0.48	0.0	0.23
RF	0.0	0.24	0.0	0.11

Table 4. Training and Test errors (%) for BRCA Subtype Classification

For Breast Invasive Carcinoma (BRCA) Subtype prediction, SVM and RF classifiers trained using either DM or EE data provided great results. Here y is a discrete-valued variable ($y \in \{0, 1 \dots 7\}$).

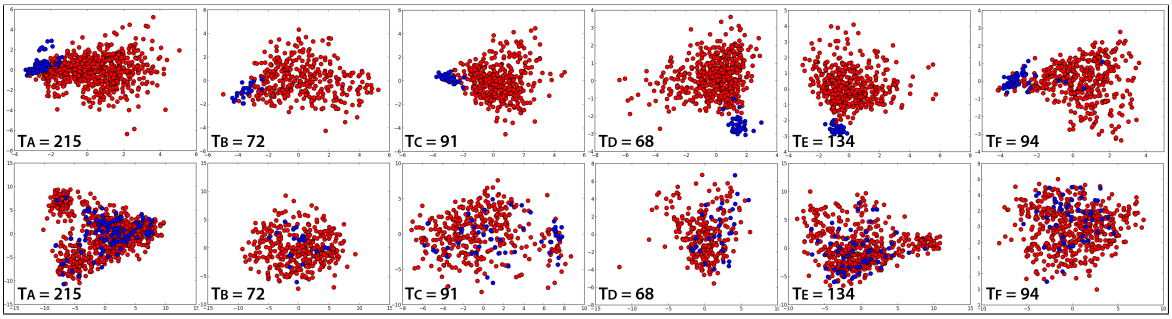

Fig. 2. Cluster Visualization after PCA Dimensionality Reduction

I carried out a Principal Component Analysis for reduction to 2 dimensions to generate Cluster visualizations over the different tumor typologies data (shown in Fig. 2). The first row proves that DNA Methylation serves as an extremely robust marker to separate the *'tumor'* samples (*red* nodes) from *'normal'* samples (*blue* nodes) for all tumor typologies, and distinct clusters are observed. This however is not the case for Exon Expression features, where no discernible separation is observed.

4 Discussion and Conclusion

There is large evidence that DNA methylation patterns *'hyper-'* and *'hypo-methylation'* in *'tumor'* sample tissues are abberant compared to *'normal'* sample tissues, and have been associated with a large number of human malignancies [1]. DNA Methylation[9] involves the addition of a methyl ($-CH_3$) group to the

[9] `http://en.wikipedia.org/wiki/DNA_methylation#In_cancer`

Cytosine nucleotide resulting in *5-methyl Cytosine*, whose accumulation of high levels over time renders a gene transcriptionally silent. Typically there is *'hyper-methylation'* of tumor suppressor genes and *'hypo-methylation'* of *oncogenes*, resulting in increased cancer growth. As a result, better evaluation metrics and PCA clusters are obtained for classifiers trained using DM Data. I believe that the 2 adjoint PCA clusters seen for the T_A using EE data (2^{nd} row, 1^{st} panel) represent 2 Subtypes of BRCA. As such, classifiers trained using TCGA DM and EE data performed well for BRCA subtype classification.

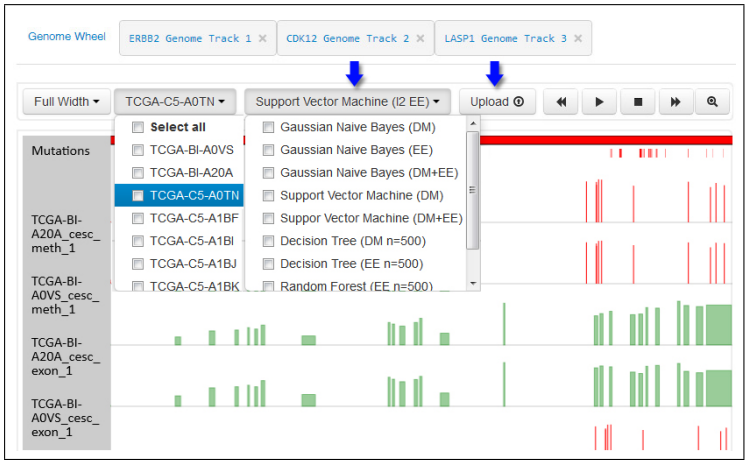

Fig. 3. Genome Tracks View

Feature Selection using Genomic Co-occurrence and then an L1-based approach provided better prediction statistics as compared to implementing an L1-based restriction on the entire $>$ 20000 gene list (results not shown). Hence prior biomedical knowledge and user inputs are essential to build better classifiers. SVM performed best for classification overall, whereas RF which performed better for DM was seen to over-fit EE (0.0% Specificity). This may be due to the case that there are more *'tumor'* ('1') samples as compared to *'normal'* ('0'). Using equal number of samples, or performing **Tree Pruning** could alleviate this. New UI controls could be added to the GenomeSnip Tracks View (Fig. 3) so the researcher can select a specific classification model along with the visualized patient data to predict risk.

My last goal was analogous to the classification of tumors using the TCGA data, and is particularly useful from a biomedical perspective. Drugs approved by regulatory agencies, and many publication resources, are still researched in the context of a single tumor typology. If we have a classification model which could efficiently predict tumor risk for a separate typology, we are essentially linking two tumor typologies through their genomic signature and this could enable drug repurposing. However, I only found an SVM classifier trained using HNSC (T_D) samples provided 99.8% Sensitivity and a desirable 90.0% Specificity while predicting COADREAD (T_C) samples. Other pairs generated skewed metrics (~100.0% Sensitivity and $<$ 50.0% Specificity or vice versa). Hence additional research is required here.

5 Future Work

This work revolves around building classifiers using the entirety of TCGA DM Data and compares the classification results against similar classifiers using EE Data. I would like to build/evaluate classifiers using all the different TCGA molecular datasets (SNP, CNV etc.), independently and in a combination to determine which would provide the best prediction metrics. Also, by the inclusion of clinical features listed for each TCGA patient (such as patient follow up status, drug administered) in our models, we could extend our diagnostic framework to generate prognosis and therapeutic advice. I would like to experiment with Deep Learning Methods and conduct user-driven evaluation of the system. We showed a preliminary approach where clinicians select genes from the *'Genomic Wheel'* to guide Genomic Co-occurrence-based feature selection - the system could be configured to allow real-time selection of the features, adjust the weights in the Similarity Equation, build custom classifiers and generate new predictions. Finally, I would like to do Unsupervised Learning on the datasets for the discovery of new Bio-markers.

Acknowledgements

I would like to acknowledge Dr. Marina Sirota who extensively guided this research and helped me understand TCGA data. This work was funded under the Biomedical Informatics PhD Program.

References

1. Das, P.M., Singal, R.: Dna methylation and cancer. Journal of Clinical Oncology 22(22), 4632–4642 (2004), `http://jco.ascopubs.org/content/22/22/4632.abstract`
2. Kamdar, M.R., Iqbal, A., et al.: GenomeSnip: Fragmenting the genomic wheel to augment discovery in cancer research. In: Conference on Semantics in Healthcare and Life Sciences (CSHALS). ISCB (2014)