GENIUS: GEnome traNsformatIon and spatial representation of mUltiomicS data

Authors

Mateo Sokač[1,2,3], Lars Dyrskjøt[1,2], Benjamin Haibe-Kains[4], Hugo J.W.L. Aerts[5,6,7], Nicolai J Birkbak[1,2,3*]

Affiliations

1 Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark
2 Department of Clinical Medicine, Aarhus University, Aarhus, Denmark
3 Bioinformatics Research Center, Aarhus University, Aarhus, Denmark
4 Princess Margaret Cancer Centre, University Health Network, Temerty Faculty of Medicine, University of Toronto, Toronto, ON, Canada
5 Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, MA
6 Departments of Radiation Oncology and Radiology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA
7 Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, The Netherlands

*Correspondence to:

Nicolai J Birkbak

Department of Molecular Medicine (MOMA), Aarhus University Hospital

Science Center Skejby, Brendstrupgårdsvej 21A

8200 Aarhus N, Denmark

E-mail: nbirkbak@clin.au.dk

Phone number: (+45) 78455347

Running Title: GENIUS

ABSTRACT

The application of next-generation sequencing (NGS) has transformed cancer research. As costs have decreased, NGS has increasingly been applied to generate multiple layers of molecular data from the same samples, covering genomics, transcriptomics, and methylomics. Integrating these types of multi-omics data in a combined analysis is now becoming a common issue with no obvious solution, often handled on an ad-hoc basis, with multi-omics data arriving in a tabular format and analyzed using computationally intensive statistical methods. These methods particularly ignore the spatial orientation of the genome and often apply stringent p-value corrections that likely result in the loss of true positive associations. Here, we present GENIUS (GEnome traNsformatIon and spatial representation of mUltiomicS data), a framework for integrating multi-omics data using deep learning models developed for advanced image analysis. The GENIUS framework is able to transform multi-omics data into images with genes displayed as spatially connected pixels and successfully extract relevant information with respect to the desired output. Here, we demonstrate the utility of GENIUS by applying the framework to multi-omics datasets from the Cancer Genome Atlas. Our results are focused on predicting the development of metastatic cancer from primary tumours, and demonstrate how through model inference, we are able to extract the genes which are driving the model prediction and likely associated with metastatic disease progression. We anticipate our framework to be a starting point and strong proof of concept for multi-omics data transformation and analysis without the need for statistical correction.

INTRODUCTION

The recent advent of Next Generation Sequencing (NGS) has revolutionized research and has been applied extensively to investigate complex biological questions. As the cost of sequencing continues to drop, it has become increasingly common to apply NGS technology to investigate complementary aspects of the biological processes on the same samples, particularly through analysis of DNA to resolve genomic architecture and single nucleotide variants, RNA to investigate gene expression, and methylation to explore gene regulation and chromatin structure. Such multi-omics data provides opportunities to perform integrated analysis, which investigates multiple layers of biological data together. Over the years, this has resulted in the generation of an incredible amount of very rich data derived from the genome itself, either directly or indirectly. The genome is spatially organized, with genes positioned on chromosomes sequentially and accessed by biological processes in blocks based on chromatin organization[1]. However, genome-derived NGS data is usually stored in and analyzed from a tabular format, where the naturally occurring spatial connectivity is lost. Furthermore, while genomic data is rich, the feature space is generally much larger than the number of samples. As the number of features to evaluate in statistical tests increases, the risk of chance associations increases as well. To correct for such multiple hypothesis testing, drastic adjustments of p-values are often applied which ultimately leads to the rejection of all but the most significant results, likely eliminating a large number of weaker but true associations. While this is a significant issue when analyzing a single type of data, the problem is exacerbated with performing multi-omics analysis where different types of data are combined, often in an ad-hoc manner tailored to specific use cases. Importantly, a common theme in multi-omics analytical approaches is that observations are processed individually, thereby discarding potential spatial information that may originate from the organization of genes on individual chromosomes.

Using artificial intelligence methods may help overcome this problem. Over the past decade, the development of artificial intelligence methods, particularly within deep learning architectures, has thoroughly revolutionized several technical fields, such as computer vision, voice recognition, advertising, and finance. Within the medical field, the roll-out of AI-based technologies has been slower, hampered in part by considerable regulatory hurdles that have proven difficult for machine-learning applications where the systems may accurately classify

patients or samples by some parameter, but the logical reason behind this is unclear[2]. Nevertheless, AI systems have proven successful in a multitude of medical studies, and in recent years some AI-powered tools have started to move past testing to deployment[3]. A major benefit of deep neural networks is that it can capture nonlinear patterns in the data without necessitating correction for multiple hypothesis testing. However, the problem with such models is not the analysis itself but the interpretation of deep learning models[4]. Simpler models tend to have high interpretability; however, they are unable to capture complex nonlinear connections in data. This often leads to the utilization of "black box" models at the cost of interpretability [5,6]. "Black box" models are popular in the artificial intelligence industry, especially in computer vision applications, where immense progress is being made in technologies such as self-driving cars and computer interpretation of images. However, in many of those applications, the interpretability of models is not as important as in medicine [7,8].

In medicine, the interpretability of models is crucial since there is a need for discovering new biomarkers as well as identifying underlying biological processes[9]. In addition to advancements in artificial intelligence and NGS, a vast amount of research has been conducted to interpret highly complex machine learning models; frameworks such as DeepLIFT[10], Integrated Gradients[11,12], and DeepExplain [10,13,14] were developed in recent years with the purpose of debugging complicated machine learning models[15]. These frameworks enable the usage of deep learning models for integrated multi-omics analysis through their ability to evaluate input attribution in models that are traditionally considered a "black box". In multiomics analysis, this means that it is possible to combine the entirety of the data from multiple data sources into a high-dimensional data structure and process it with deep learning models without losing interpretability. As output, an attribution score can be produced for every input, which may be interpreted as the relative importance of the feature in the model and used for further analysis.

Here, we present a framework for multi-omics analysis based on a convolutional deep learning network to find hidden, non-linear patterns in spatially connected feature-rich multi-layered data. The spatial connection of the data is made by transforming the data into a multi-channel image in such a way that spatial connections between genes are captured and analyzed using

convolutional layers. Using spatial connections between the data showed superior performance when compared to non-spatially data transformations. Furthermore, the trained model is combined with Integrated Gradients, which allows us to evaluate the relative contribution of individual features and thus decipher the underlying biology that drives the classification provided by the deep learning models. We named the framework GENIUS (GEnome traNsformatIon and spatial representation of mUltiomicS data), and the methodology may be split into two parts, classification and interpretation. First, the key feature of GENIUS is that for classification, multi-omics data is transformed into multi-channel images where each gene is presented as a pixel in an image that covers the whole genome (Figure 1A-B). We then incorporate multiple types of omics data, such as mutation, expression, methylation, and copy-number data, into the image as distinct layers. These layers are then used as input into the deep learning model for training against a binary or continuous outcome variable. Next, for interpretation, an attribution score is assigned to each feature using integrated gradients, allowing the user to extract information about which feature or features may drive a specific prediction based on deep learning analysis of input from multiple-omics data sources. In this work, we describe the development of the GENIUS framework and demonstrate its utility in predicting the development of metastatic cancer, patient age, chromosomal instability, cancer type, and as proof of concept, loss of TP53.

All based on multi-omics input through the GENIUS framework, users may train multi-omics data against a specified endpoint tailored to the user's choice. The GENIUS framework thus overcomes the issue of multiple hypothesis testing and may provide new insights into the biology behind classification by deep learning models. The GENIUS framework is made available as a GitHub repository and may be used without restrictions to develop stratification models and inform about genome biology using multi-omics input.

METHODS

GENIUS Model Architecture and Hyperparameters

We designed a four-part convolutional neural network with the purpose of extracting the features from multi-dimensional data (Figure 1C). The structure of the network is similar to an autoencoder architecture; however, the reconstruction of the genome image is not penalized. The motivation behind the implemented network structure is to use an encoder in order to learn how to compact genomic information into a small vector, L, forcing the network to extract relevant information from five data sources. The next module reconstructs the image from vector L, learns which features are important, reorganizes them optimally, and removes noise. The final module of the network uses a series of convolutions and max pooling layers in order to extract information from the reconstructed image and, finally, predicts the outcome variable using a fully connected dense network.

The first part of the network is called the encoder, as its purpose is to encode the entire image to a vector of size 128, representing the latent representation of the input data, "L". Next, the original image is reconstructed from L into its original size using a decoder module in the network. In this step, since we are not using the reconstruction loss, the network reconstructs the image of a genome which is optimal for information extraction. This is followed by the extractor module containing convolution and max-pooling layers aiming to extract relevant information from the reconstructed image. The final part of the network flattens the learned features obtained from previous layers, concatenates them with the L vector, and forwards it to a fully connected dense feed-forward network where the final prediction is made (Figure 1C) [16]. During training, the last module of this model was adopted to predict qualitative as well as quantitative types of data.

All models were trained with Adagrad optimizer with the following hyperparameters: starting learning rate = 9.9e-05 (including learning rate scheduler and early stopping), learning rate decay and weight decay = 1e-6, batch size = 256, except for memory-intensive chromosome images where the batch size of 240 was used. Adding chromosome interaction information to the data transformation showed improvement during training; Next question was whether we should

penalize the reconstruction of genome image during the training process. After multiple training scenarios and hyperparameter exploration, we concluded that by forcing the network to reconstruct genome images in the process of learning, we are limiting network performance. Instead, we used the appropriate loss function for prediction and allowed the network to reconstruct genome images that are optimal for making predictions.

Evaluating input image design

To evaluate the performance of GENIUS with an image-based transformation of input omics data, we tested four different image layouts of the genome. For each layout, we created a set of images where each sample is represented by one multichannel image and each channel represented a specific type of omics data (gene expression, methylation, mutation, deletion, amplification) (Supplementary Figure 2A). Each data type was encoded for each gene as a continuous value, where each gene was defined by a single pixel in each layer. We then tested the performance of the deep neural network on four different image layouts. First, we assembled the genome as a square image, measuring 198 x 198 pixels in total. Here, all genes were placed on the image sequentially according to their chromosomal locations, and individual chromosomes were organized by how close they were oriented in 3D space[17]. Second, we tested an image organized by 24 x 3760 pixels, with 3760 pixels representing the most gene-rich chromosome, and each chromosome placed below the other on the image following the same order as in 198x198 images. Chromosomes containing fewer than 3760 genes had black pixels added to the end to create a rectangular image. Third, we tested a random 2D location, with each gene placed as a random pixel in a 198x198 pixel square image. Lastly, we tested an image of a single vector with all genes placed in a randomly ordered sequence. Data transformation we performed and tested:

1. Square image (198 x 198 pixels), each gene represented by one pixel ordered by chromosome position. Chromosomes are ordered by interaction coefficient based on Hi-C sequencing [17].
2. Square image (198 x 198 pixels), each gene is represented by one pixel located on the image in random order; thus, the 2D location carries no information.

3. Rectangular image (24 x 3760 pixels), each gene represented by one pixel ordered by chromosome position. Chromosomes are ordered by interaction coefficient based on Hi-C sequencing [17]
4. A flat, one-dimensional vector containing all features from the five data sources in random order.

By using different image layouts, we wanted to investigate the spatial dependency of observations. Images were created by making a matrix for each source of data where each cell was represented by a single gene (Figure 1A, Supplementary Figure 2A-B). The genes in 198x 198 and 24 x 3760 images were ordered by position as well as by chromosome interaction coefficients resulting in the following order of chromosomes: 4, X, 7, 2, 5, 6, 13, 3, 8, 9, 18, 12, 1, 10, 11, 14, 22, 19, 17, 20, 16, 15, 21. Finally, newly created observations for each data source were merged as a multi-channel image where each channel represents a single source of data (Supplementary Figure 2A).

Samples & Training Data

We obtained gene expression, exome mutation, methylation and copy number data from six cancer types from the Cancer Genome Atlas (TCGA). These were picked to filter out cancer types with less than 400 samples. Next, cancer types with an extremely high proportion of metastatic samples (0.85 < Proportion > 0.15) were removed, resulting in ovarian serous cystadenocarcinoma (OV), colon adenocarcinoma (COAD), uterine corpus Endometrial carcinoma (UCEC), kidney renal clear cell carcinoma (KIRC), urothelial bladder carcinoma (BLCA) and stomach adenocarcinoma (STAD) (Supplementary Figure 1). RNAseq was obtained from the UCSC Toil pipeline [18] and summarized to transcript per million (TPM) on the gene level. SNP6 copy number data were segmented using ASCAT v2.4[19,20] and converted into a ploidy and purity normalized logR value by dividing the total copy number with ploidy and taking the log2 value. The weighted genome integrity index (wGII)[21] was calculated on the available segmented copy number data, as previously described. Mutation calls were annotated using Polyphen2 to assess the mutation's impact on the protein structure. Methylation was summarized by the mean methylation score for each gene.

Validation cohorts acquisition and processing

Two independent cohorts of bladder cancer patients were used for validation. The UROMOL cohort [22,23] contains molecular data from 535 tumors from patients with early-stage bladder cancer (Ta and T1) and was included to evaluate the progression to muscle-invasive bladder cancer. The Mariathasan cohort[24] contains molecular data from 348 tumors from patients with advanced or metastatic bladder cancer (stage III and IV), treated with checkpoint immunotherapy. This cohort was included to evaluate the ability of the GENIUS framework to predict the likelihood of developing metastatic disease.

For both cohorts, RNAseq data was aligned against hg38 using STAR[25] version 2.7.2 and processed to generate count and transcript per million (TPM) expression values with Kallisto [26] version 0.46.2. Whole exome sequence (WES) data was processed using GATK [27] version 4.1.5 and ASCAT version 2.4.2 to obtain mutation and allele-specific copy number, purity, and ploidy estimates.

Data transformation

For easier and more stable training of machine learning models, we generated mathematically equivalent values for each data source ranging from 0 to 1. The simple linear transformation was performed on gene expression and methylation data. All mutations were ranked by PolyPhen scores. LogR segmented copy number data was analyzed as deletion and amplification separately. Copy number deletion was defined as logR scores < log2 of 0.5/2, copy number amplification was defined as logR scores > log2 of 5/2. All data types were defined on the gene level. For copy number alterations, we defined genes as amplified if the entirety of the gene was found within the amplified DNA segment. Genes were defined as deleted if they were partially within the deleted DNA segment (Figure 1A, Supplementary Figure 2A-B).

Training scenarios

We used the GENIUS framework to make models predicting the following conditions:

1. Metastatic cancer (binary classification), defined as Stage IV vs Stages I, II and III
2. TP53 mutation (binary classification), where the TP53 mutation was removed from the input data and used only as a binary outcome label
3. The tissue of origin (multi-class classification)

4. Age (continuous variable)
5. Weighted genome integrity index (wGII)[21], a chromosomal instability marker (continuous variable)
6. Randomized tissue of origin (multi-class variable), negative control

In order to adapt the network for predicting different variables, we simply changed the output layer and loss function for training. Binary classifications and the multi-class classification used softmax as the output layer and the cross entropy loss function. When predicting continuous values, the model used the output from the activation function with the mean squared error loss function.

Evaluating Model performance

Every model output variable was trained until we observed no change in loss function or until validation loss values started increasing, indicating overfitting, which we handled by implementing early stopping. We compared the performance metrics of the models using four different data transformations: Square (198x198) images, Chromosome Images (24x3760), Random Square (198x198) Images, and a single flat vector including all features from all data sources.

Latent representation of genome

The purpose of latent vectors is to capture the most significant information from the entire genome data and compress it into a vector of size 128. This vector was later appended into a feed-forward network when making the final prediction. This way, the model had access to extracted information before and after image reconstruction. After successful model training, we extracted the latent representations of each genome and performed the Uniform Manifold Approximation and Projection (UMAP) of the data. The UMAP projected latent representations into two dimensions which could then be visualized. In order to avoid modeling noise, this step was used to address model accuracy and inspect if the model is distinguishing between variables of interest. We observed that all training scenarios successfully utilized genome images to make predictions with the exception of Age and randomized cancer type (negative control), where the model performed poorly (Figure 2B). Information in latent vectors extracted from Age-Model

and randomized cancer type showed no obvious patterns, which is likely the cause of poor performance.

Identifying genes relevant to the tested outcome

Once the model was trained on the data, the appropriate loss function and output layer, including the model weights, were stored in a .pb file. The model and final weights were analyzed using the Integrated Gradients method (IG) implemented by Capture [12]. IG is an attribution method that assigns an "attribution score" to each feature of the input data based on predictions the model makes. The attribution score is calculated based on the gradient associated with each feature of each image channel with respect to the output of the model. This information indicates to the neural network the extent of weight decrease or increases needed for certain features during the backpropagation process. Next, the created attribution images are used to extract information for each image channel and for every pixel. Since the pixels represent individual genes, this information can be reformatted and filtered to show the most important genes from every data source included in the analysis. All attribution scores were scaled using a min-max scaler for every cancer type to address biological differences between cancer types.

Code Availability

All code is available on the public GitHub repository (https://github.com/mxs3203/genome_mapped_to_image), where the framework is easily available for analysis of private or public data. The framework provides tools to transform gene-oriented data into an image, train a model using existing model architecture and infer the most informative genes from the model using integrated gradients.

RESULTS

Building genome images to utilize spatial connections in genomic data

We endeavored to present genomic data as an image with genes represented as individual pixels to be processed by our deep-learning architecture. To evaluate the relevance of the spatial orientation of the genes relative to model performance, we tested four different image layouts (Methods, Figure 2A): (1) Square image (198 x 198 pixels), each gene represented by one pixel ordered by chromosome position. Chromosomes are ordered by interaction coefficient based on Hi-C sequencing [9]. (2) Square image (198 x 198 pixels), each gene is represented by one pixel located on the image in random order; thus, the 2D location carries no information. (3) Rectangular image (24 x 3760 pixels), each gene represented by one pixel ordered by chromosome position. Chromosomes are ordered by interaction coefficient based on Hi-C sequencing [9]. (4) A flat, one-dimensional vector containing all features from the five data sources in random order. To evaluate the image layout, we used each type of layout to train against six biological states: (1) metastatic disease (stage IV vs. I-III), (2) cancer type, (3) burden of copy number alterations (defined by the weighted genome instability index, wGII), (4) patient age, (5) TP53 status (where the TP53 pixel was set to "0" for all samples), and (6) randomized tissue type (negative control).

While predicting metastatic disease, we observed that the Square Image data transformation outperformed all other data transformations, reaching a validation AUC of 0.87. Chromosome image and shuffled squared image performed similarly with AUC of 0.72 and 0.70, respectively. Interestingly, the flat vector of features scored validation AUC around 0.84; however, the loss function started increasing as training epochs increased, indicating that the model overfitted (Figure 2B, Supplementary Figure 4 C-D). In the second scenario, we tested multiclass prediction using six cancer types in our dataset. The performance measure was defined by the F1 score, a standard measure for multiclass classification that combines the sensitivity and specificity scores, and is defined as the harmonic mean of its precision and recall. Square Image outperformed other image layouts, reaching an F1 score of 0.81. Chromosome Image followed with an F1 score of 0.74 and flat vector of features performed similarly to the random square image, reaching F1 scores of 0.66 and 0.71, respectively (Supplementary Figure 4 E-F). In order

to address the framework's capabilities for predicting numeric output variables, we used wGII and patient age. Predicting wGII showed that the flat vector of features reached the least favorable RMSE score of 0.22, where chromosome image, shuffled square image, and square image reached similar RMSE scores of 0.16, 0.15, and 0.14, respectively (Supplementary Figure 5 C-D).

These results suggest that data layout does not play a major role when predicting wGII as the number of events in the genome would be predictive regardless of location. The age prediction model using square image data transformation outperformed other data transformations and obtained a validation RMSE of 0.19. The shuffled image performed the worst, reaching an RMSE of 0.49, while the chromosome-organized image and flat vector of features scored similar RMSE values of 0.38 and 0.31, respectively. Additionally, the flat vector was inconsistent during training, but it did outperform the chromosome image (Supplementary Figure 5 A-B). In the fifth scenario, we predicted TP53 mutation status but removed the TP53 mutation itself from the data. Square Image performed the best, reaching a validation AUC of 0.83, whereas no major difference could be observed between a flat vector of features and Chromosome Image, reaching a validation AUC of 0.75 (Supplementary Figure 4 A-B). Finally, we tested the framework by predicting randomized cancer types as the negative control. All data transformations had similar and poor results (Supplementary Figure 6). For each output variable, we trained four different models utilizing the four data transformations. In all cases, the square image (198x198, ordered by chromosomes) outperformed the other transformations and was chosen as the layout for the final GENIUS framework, which was used for all subsequent analyses.

Latent representation of genome captures relevant biology

The model architecture contains an encoder and decoder connected by a latent vector of size 128 (L), which provides the opportunity to inspect model performance (Figure 1). The L vector is considered the latent representation of the genome data because it extracts and captures the most relevant data with respect to the output variable. This implies that an optimally trained model would show a perfect latent representation of the genome when overlaid with the output variable. Furthermore, this vector was later appended into a feed-forward network when making the final prediction. This way, the model had access to extracted information before and after image

reconstruction. In order to inspect model performance, we extracted the latent representations of each genome and performed the Uniform Manifold Approximation and Projection (UMAP) of the data to project it into two dimensions. We observed that all training scenarios successfully utilized genome images to make predictions that clustered into distinct groups, with the exception of Age and randomized cancer type (negative control), where the model performed poorly (Figure 2C). Information in latent vectors extracted from Age-Model and randomized cancer type showed no obvious patterns, which is likely the cause of poor performance.

GENIUS classification identifies tumors likely to become metastatic

To explore the utility of the GENIUS framework to classify tumors from multiomics data and to interpret the biological drivers behind the classification, we further investigated the GENIUS model trained against metastatic disease using the TCGA datasets (Figure 2B). This analysis included primary tumors from six cancer types, a total of 2307 tumors, with 53 percent progressing to metastatic disease (bladder cancer [BLCA, 277 metastatic/133 not-metastatic], ovarian cancer [OV, 535 metastatic/47 not-metastatic], colon adenocarcinoma [COAD, 196 metastatic/254 not-metastatic], stomach adenocarcinoma [STAD, 230 metastatic/189 not-metastatic], kidney clear cell renal cancer [KIRC, [208 metastatic/326 not-metastatic], uterine endometrial cancer [UCEC, 117 metastatic/394 not-metastatic]). The omics data types included somatic mutations, gene expression, methylation, copy number gain and copy number loss. Using cross-validation, we observed a generally high performance of GENIUS, with a mean validation AUC of 0.83 (Figure 2B). The GENIUS framework allows us to explore the attribution of individual data layers to the final prediction. Across the cohort, gene expression and methylation data were generally the most informative data layers when it comes to classifying metastatic disease (Figure 3A). We noted that expression and methylation overall ranked the highest in terms of mean scaled attribution, with the exception of OV, which showed enrichment in methylation followed by copy number gain and loss. The same analysis was performed for cancer type, wGII, patient age, TP53 status and randomized tissue type, (Supplementary Figures 7 - 8).

Interpreting the GENIUS model classifying metastatic cancer biology

The most informative data type overall regarding the development of metastatic disease was methylation (Figure 3A). To identify the individual genes driving the prediction, we pulled the 100 genes with the highest methylation attribution according to the GENIUS classification. We observed that many methylated regions overlapped between the six cancer types. These regions included methylation on specific regions of chromosomes 1, 6, 11, 17, and 19 (Supplementary Figure 9). Additionally, OV showed a unique methylation pattern spanning most of chromosome 7, while KIRC, COAD, and BLCA displayed regions of overlapping methylation on chromosome 22. To determine the genes that overall across the multi-omics data analysis contributed the most to the GENIUS classification of metastatic disease, we normalized gene attribution by cancer type and compared the top 50 genes for each cancer type (total of 152 genes, Figure 3B, Supplementary Table 1). Unsurprisingly, we observed that TP53 mutations held the highest attribution score, followed by mutations to VHL. Both of these genes are well-established drivers of cancer and were previously reported as enriched in metastatic cancer [28,29], likely representing a more aggressive disease. However, of the 152 top genes, we noted only 11 genes previously reported as either oncogenes or tumor suppressor genes in the COSMIC cancer gene census (Figure 3B, indicated with a star), leaving 141/152 as potentially novel cancer genes. The highest-scoring gene not previously associated with cancer was SLC3A1, the expression of which was found to be strongly associated with metastatic disease in clear-cell renal cancer. SLC3A1 gene is a protein-coding gene associated with the transportation of amino acids in the renal tubule and intestinal tract, and aberrations in this gene have been associated with cystinuria, a metabolic disorder of the kidneys, bladder, and ureter [30,31]. Furthermore, we identified PLVAP, often involved in MAPK cascades as well as in cellular regulatory pathways and the tumor necrosis factor-mediated signaling pathway. In BLCA, one of our most significant findings was increased expression of KRT17, a gene associated with a cytoskeletal signaling pathway, glucocorticoid receptor regulatory network, and MHC class II receptor activity [32,33]. KRT17 has previously been reported as a potential cancer gene, but with an uncertain role[34]. Across cancer types, TOP3A was found to be commonly methylated in BLCA, COAD, STAD and UCEC. TOP3A is associated with homology-directed repair and methylation may lead to increased chromosomal instability, a hallmark of cancer[35]. The top

ten most important events driving the prediction of every output variable included in the study are summarized in Supplementary Table 2-3.

Validation of bladder cancer metastasis-associated genes in an independent cohort of advanced and metastatic bladder cancer

The top 10 expressed/methylated genes from the TCGA analysis with the highest attribution score associated with the development of metastatic disease in bladder cancer were NBPF9, RBMX, COL7A1, RPL26L1, KRT17, FAM84B, JUP, WIPI2, PRKACB, and TOP3A. To validate that these genes are indeed associated with metastatic bladder cancer, we utilized an immunotherapy-treated predominantly late-stage (mainly stage III and IV) bladder cancer cohort with gene expression data available for 348 tumors[24]. For this analysis, we considered only the methylation and gene-expression associated genes from the TCGA analysis, and given that methylation negatively correlates with gene expression, these were considered together. First, we implemented 10 multivariate cox proportional hazard models (one for each gene), including available clinical parameters such as tumor stage, gender, neoantigen burden and baseline performance status (Supplementary Table 4). This showed that in multivariate analysis, 6/10 genes had a significant association with outcome (Figure 4A). To evaluate the results of this analysis, we compared it to an identical model run 1,000 times, but where the 10 genes were randomly picked. Only in 1 of 1,000 runs did at least 6 genes emerge significant ($P = 0.001$) (Supplementary Figure 10A). The median percentage of significant tests is reported in Figure 4B. Next, we performed two independent analyses, comparing the expression values of the top 10 genes between either (1) tumors defined as stage IV versus stage I-III, and (2) patients that responded to immunotherapy (CR and PR) versus patients that did not respond to immunotherapy (SD and PD). In both analyses, we observed 5/10 genes significantly enriched after correction (Figure 4C). In stage IV samples, the genes TOP3A and NBPF9 were enriched, while in stage I-III tumors, JUP, KRT17, and PRKACB were found to be most significant. While comparing responses to immunotherapy, TOP3A, RBMX, and WIPI2 were more expressed in CR/PR while KRT17 and COL7A1 were expressed in SD/PD. Conversely, we observed that TOP3A was enriched in stage IV samples, suggesting a role in metastatic disease, yet we also observed that the same gene was more expressed in samples that responded to immunotherapy. This suggests that TOP3A is associated with the development of metastatic disease, but it may

result in the development of a bladder cancer phenotype that is more sensitive to checkpoint immunotherapy.

Validation of metastasis-associated genes in an independent cohort of early-stage bladder cancer

To investigate if the metastasis-associated genes found through the GENIUS framework also plays a role in the development of aggressive features in early stage bladder cancer, we acquired the UROMOL dataset[23], which includes gene expression data from 535 low-stage tumors. We again investigated the top 10 methylated or expressed genes found in the TCGA analysis of BLCA, using the gene expression data from UROMOL. First, we performed Cox proportional hazard analysis with progression-free survival (PFS) using the top 10 genes found by the GENIUS framework, creating 10 individual models containing the selected genes and available clinical factors such as age, tumour stage, and sex (Figure 5A). The results were compared with cox proportional hazard models utilizing random sets of ten genes, repeated 1,000 times. Of these, 33 runs showed at least seven significant genes ($P = 0.033$) (Supplementary Figure 10B). Next, we computed the median percentage of significant genes and compared it to the top 10 genes picked by the GENIUS framework. By random chance, 20% of genes overall were found to be significantly associated with PFS compared to 70% of GENIUS genes (Figure 5B). To further investigate the top 10 genes picked by GENIUS, we compared the mean expression of each gene between different clinical risk groups (EORTC status [36]) and tumor grade. The same seven significant genes from proportional hazard models (RBMX, PRKACB, RPL26L1, KRT17, FAM84B, NBPF9, TOP3A) were reported significantly in this analysis (Figure 5C, Supplementary Table 5). When comparing tumor grade, FAM84B was no longer significant (Figure 5D).

DISCUSSION

In this work, we explored multiple options on how to transform multi-omics data into an image, leading to the utilization of deep learning models, which are often described as “black box” models. The model architecture was evaluated in six different training scenarios, with a focus on validating the prediction of metastatic cancer. In this process, we also evaluated four different image layouts, concluding that of these, projecting the genome into a 198x198 square image with genes organized based on chromosome interaction[17] performed the best. Given that spatial organisation improved the prediction, we recognize that there may exist a more optimal representation of multi-omics data. For this reason, we included an auto-encoder in the network to recreate the input information without reconstruction loss. In this manner, the model can reconstruct the image of a genome in a manner that is optimal for the prediction it is trying to make. The model also produces a latent representation of multi-omics data in a shape of a vector of a size 128 (L), which is later concatenated in a model when making final predictions. In order to investigate training effectiveness, we performed a UMAP clustering analysis of the L vector, where we compared the 2D representation of L with the variables of interest (Figure 2C). It is clear from this analysis that the L vector itself holds information that may be particularly relevant for multi-class prediction, but further analysis is needed to decipher what information is encoded in the L vector.

The main purpose behind the study was to demonstrate the feasibility of leveraging the power of deep learning techniques optimised for image analysis to interpret genome-derived multi-omics data. A key element of this approach includes the transformation of genomic data into images with genes arranged as pixels organised by chromosomal location. Beyond the readout from multi-omics data, this approach provides spatial information to the deep-learning framework, which significantly improves the performance of the models (Figure 2B). To the best of our knowledge, we are the first to demonstrate the utility of spatial information and to provide a ready-to-use framework that incorporates spatial information and deep learning for the analysis of genome-derived multi-omics data. Furthermore, within the GENIUS framework, we facilitate the interpretation of the trained model in order to explore the biology behind the prediction without the need for data preprocessing and multiple hypothesis correction. This was achieved

by combining a deep learning network with Integrated Gradients[12], allowing us to infer the attribution score for the input, resulting in non-parametric, ready-to-analyze output.

For every cancer type included in the dataset, we listed the top 10 genes driving metastatic disease and investigated in detail genes associated with BLCA metastasis and aggressiveness. For this, we used two independent cohorts, one representing late-stage and metastatic cancer, and one representing early stage cancer. In both cohorts, we tested if methylation and expression of genes found by the GENIUS framework were associated with survival at higher rates than when compared to randomly picked genes. In the late stage BLCA cohort, six out of 10 genes were significantly associated with overall survival, while in the early stage BLCA cohort, we found that seven out of 10 were significantly associated with progression-free survival. Interestingly, not all 10 genes seem to play the same role in early-stage and in late-stage BLCA (Table 1). The TOP3A and KRT17 genes showed a dual role in metastatic disease development. It is known that TOP3A has an important role in homology-directed repair and loss may be associated with chromosomal instability, which has shown a positive association with immunotherapy response [37–39], potentially offering a likely explanation for this finding. Similarly, we observed that KRT17 was enriched in stages I-III, suggesting it may be associated with a less aggressive disease type. However, in the immunotherapy-treated cohort, KRT17 is associated with poor response to immunotherapy. In previous studies, KRT17 has been reported as associated with the development of metastatic disease, MCH type II receptor activity and angiogenesis [34,40]. This indicates that the KRT17 gene plays an important role as tumor suppressor gene in early-stage cancer, and that loss may further promote the development of aggressive, metastatic disease. While further research in this field is required to properly assess the utility of the reported genes, this work provides a framework that unlocks powerful machine-learning for more direct analysis of multi-omics data.

Taken together, we provide here the GENIUS framework along with analysis demonstrating the utility in multi-omics analysis. While we have focused on cancer analysis here, we believe GENIUS may find utility in a diverse range of genome-based multi-omics analyses. We have provided a git-hub repository that can be used to transform data into images and train the same

model predicting variables of user's interest and inferring the importance of input with respect to the desired output.

ACKNOWLEDGEMENTS

N.J.B. is a fellow of the Lundbeck Foundation (R272-2017-4040), and acknowledges funding from Aarhus University Research Foundation (AUFF-E-2018-7-14), and the Novo Nordisk Foundation (NNF21OC0071483). The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

FIGURE LEGENDS

Figure 1. Study overview. (A) The study utilised 2332 tumour samples representing six cancer types (bladder, uterine, stomach, ovarian, kidney and colon) and transformed multiomics data into images using chromosome interaction network. After the model was trained, we validated found genes with two independent cohorts representing early stage BLCA (UROMOL) and late stage BLCA (Mariathasan) (B) The validation included looking at the most important genes driving metastatic disease, similar/different methylation patterns between cancer types, latent representation of genome data and looking at survival data. (C) The model architecture where the first part of the network encodes genome data into latent vector, L, followed by decoding where image is reconstructed. Next layers aim to extract information from the reconstructed image, concat it with L and make a final prediction.

Figure 2. Data transformation overview. (A) The multiomics genome data was transformed into four image types; square image organised by chromosome interaction network, chromosome image organised by chromosome interaction network, randomly organised image and flat vector containing all multiomics data. (B) All four image types were used in training for metastatic disease prediction and the square image organised by chromosome interaction network resulted in the best model performance (green colour). The bar plot shows the performance within each cancer type included in study (C) 2-dimensional representation of vector L using UMAP for each predicted variable. Colours indicate the output variable which was used in the specific run

Figure 3. The most important events in metastatic disease development. (A) Pieplot showing the relative importance of each data source when predicting metastatic disease for each cancer type included in the study. (B) Top 50 genes for every cancer type scale by cancer type. The star symbol below the gene names indicates that gene is part of COSMIC gene consensus. The color of the gene name indicates the data source and colour of the bar indicates the cancer type.

Figure 4. Validation on late stage immunotherapy treated bladder cancer (Mariathasan). (A) Comparison of randomly selected genes vs. genes picked by GENIUS in cox proportional hazard model. (B) Forest plot showing top 10 expressed/methylated genes in multivariate cox

proportional hazard model. (C) Volcano plot showing top 10 expressed/methylated genes and their enrichment in two comparisons; Stages I-III vs Stage IV and immunotherapy binary response.

Figure 5. Validation on early stage bladder cancer (UROMOL). (A) Comparison of randomly selected genes vs genes picked by GENIUS in cox proportional hazard model. (B) Forest plot showing top 10 expressed/methylated genes picked by GENIUS for BLCA. (C) Volcano plot showing the top 10 expressed/methylated genes when compared between EORTC-Low and EORTC-High groups. (D) Volcano plot showing the top 10 expressed/methylated genes when compared between low grade and high grade BLCA tumours.

REFERENCES

1. Franke M, Ibrahim DM, Andrey G, Schwarzer W, Heinrich V, Schöpflin R, et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature. 2016;538: 265–269.

2. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. Nature Medicine. 2019. pp. 1337–1340. doi:10.1038/s41591-019-0548-6

3. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med. 2020;3: 118.

4. Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nat Mach Intell. 2019;1: 206–215.

5. Elmarakeby HA, Hwang J, Arafeh R, Crowdis J, Gang S, Liu D, et al. Biologically informed deep neural network for prostate cancer discovery. Nature. 2021;598: 348–352.

6. Wolfe JC, Mikheeva LA, Hagras H, Zabet NR. An explainable artificial intelligence approach for decoding the enhancer histone modifications code and identification of novel enhancers in Drosophila. Genome Biol. 2021;22: 308.

7. Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Inf Fusion. 2022;77: 29–52.

8. Petch J, Di S, Nelson W. Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. Can J Cardiol. 2022;38: 204–213.

9. Picard M, Scott-Boyer M-P, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021;19: 3735–3746.

10. Shrikumar A, Greenside P, Kundaje A. Learning Important Features Through Propagating Activation Differences. 2017 [cited 8 Feb 2023]. doi:10.48550/arXiv.1704.02685

11. Ancona M, Ceolini E, Öztireli C, Gross M. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. 2017 [cited 8 Feb 2023]. doi:10.48550/arXiv.1711.06104

12. Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. 2017. Available: http://arxiv.org/abs/1703.01365

13. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K-R. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer Nature; 2019.

14. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W. On Pixel-Wise

Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS One. 2015;10: e0130140.

15. Despraz J, Gomez S, Satizábal HF, Peña-Reyes CA. Towards a Better Understanding of Deep Neural Networks Representations using Deep Generative Networks. Proceedings of the 9th International Joint Conference on Computational Intelligence. 2017. doi:10.5220/0006495102150222

16. LeNail A. NN-SVG: Publication-Ready Neural Network Architecture Schematics. Journal of Open Source Software. 2019. p. 747. doi:10.21105/joss.00747

17. Sarnataro S, Chiariello AM, Esposito A, Prisco A, Nicodemi M. Structure of the human chromosome interaction network. PLoS One. 2017;12: e0188201.

18. Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, Novak A, et al. Toil enables reproducible, open source, big biomedical data analyses. Nat Biotechnol. 2017;35: 314–316.

19. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7. doi:10.1038/nmeth0410-248

20. Raine KM, Van Loo P, Wedge DC, Jones D, Menzies A, Butler AP, et al. ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. Curr Protoc Bioinformatics. 2016;56: 15.9.1.

21. Burrell RA, McClelland SE, Endesfelder D, Groth P, Weller M-C, Shaikh N, et al. Replication stress links structural and numerical cancer chromosomal instability. Nature. 2013;494: 492–496.

22. Zuiverloon TCM, Beukers W, van der Keur KA, Nieuweboer AJM, Reinert T, Dyrskjot L, et al. Combinations of urinary biomarkers for surveillance of patients with incident nonmuscle invasive bladder cancer: the European FP7 UROMOL project. J Urol. 2013;189: 1945–1951.

23. Lindskrog SV, Prip F, Lamy P, Taber A, Groeneveld CS, Birkenkamp-Demtröder K, et al. An integrated multi-omics analysis identifies prognostic molecular subtypes of non-muscle-invasive bladder cancer. Nat Commun. 2021;12: 2301.

24. Mariathasan S, Turley SJ, Nickles D, Castiglioni A, Yuen K, Wang Y, et al. TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. Nature. 2018;554: 544–548.

25. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21.

26. Ayers M, Lunceford J, Nebozhyn M, Murphy E, Loboda A, Kaufman DR, et al. IFN-γ-related mRNA profile predicts clinical response to PD-1 blockade. J Clin Invest.

2017;127: 2930–2940.

27. Van der Auwera GA, O'Connor BD. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. O'Reilly Media; 2020.

28. Pandey R, Johnson N, Cooke L, Johnson B, Chen Y, Pandey M, et al. TP53 Mutations as a Driver of Metastasis Signaling in Advanced Cancer Patients. Cancers. 2021. p. 597. doi:10.3390/cancers13040597

29. Christensen DS, Ahrenfeldt J, Sokač M, Kisistók J, Thomsen MK, Maretty L, et al. Treatment represents a key driver of metastatic cancer evolution. Cancer Res. 2022. doi:10.1158/0008-5472.CAN-22-0562

30. Jiang Y, Cao Y, Wang Y, Li W, Liu X, Lv Y, et al. Cysteine transporter SLC3A1 promotes breast cancer tumorigenesis. Theranostics. 2017;7: 1036–1046.

31. Woodard LE, Welch RC, Veach RA, Beckermann TM, Sha F, Weinman EJ, et al. Metabolic consequences of cystinuria. BMC Nephrol. 2019;20: 227.

32. Wu J, Xu H, Ji H, Zhai B, Zhu J, Gao M, et al. Low Expression of Keratin17 is Related to Poor Prognosis in Bladder Cancer. Onco Targets Ther. 2021;14: 577–587.

33. Li C, Su H, Ruan C, Li X. Keratin 17 knockdown suppressed malignancy and cisplatin tolerance of bladder cancer cells, as well as the activation of AKT and ERK pathway. Folia Histochem Cytobiol. 2021;59: 40–48.

34. Zhang H, Zhang Y, Xia T, Lu L, Luo M, Chen Y, et al. The Role of Keratin17 in Human Tumours. Front Cell Dev Biol. 2022;10: 818416.

35. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144: 646–674.

36. European Organisation For Research And Treatment Of Cancer. In: EORTC [Internet]. 17 Jan 2017 [cited 13 Jul 2022]. Available: https://www.eortc.org/

37. Bakhoum SF, Cantley LC. The Multifaceted Role of Chromosomal Instability in Cancer and Its Microenvironment. Cell. 2018;174: 1347–1360.

38. Chen M, Linstra R, van Vugt MATM. Genomic instability, inflammatory signaling and response to cancer immunotherapy. Biochim Biophys Acta Rev Cancer. 2022;1877: 188661.

39. Sokač M, Ahrenfeldt J, Litchfield K, Watkins TBK, Knudsen M, Dyrskjøt L, et al. Classifying cGAS-STING Activity Links Chromosomal Instability with Immunotherapy Response in Metastatic Bladder Cancer. Cancer Research Communications. 2022;2: 762–771.

40. Ji R, Ji Y, Ma L, Ge S, Chen J, Wu S, et al. Keratin 17 upregulation promotes cell metastasis

and angiogenesis in colon adenocarcinoma. Bioengineered. 2021;12: 12598–12611.

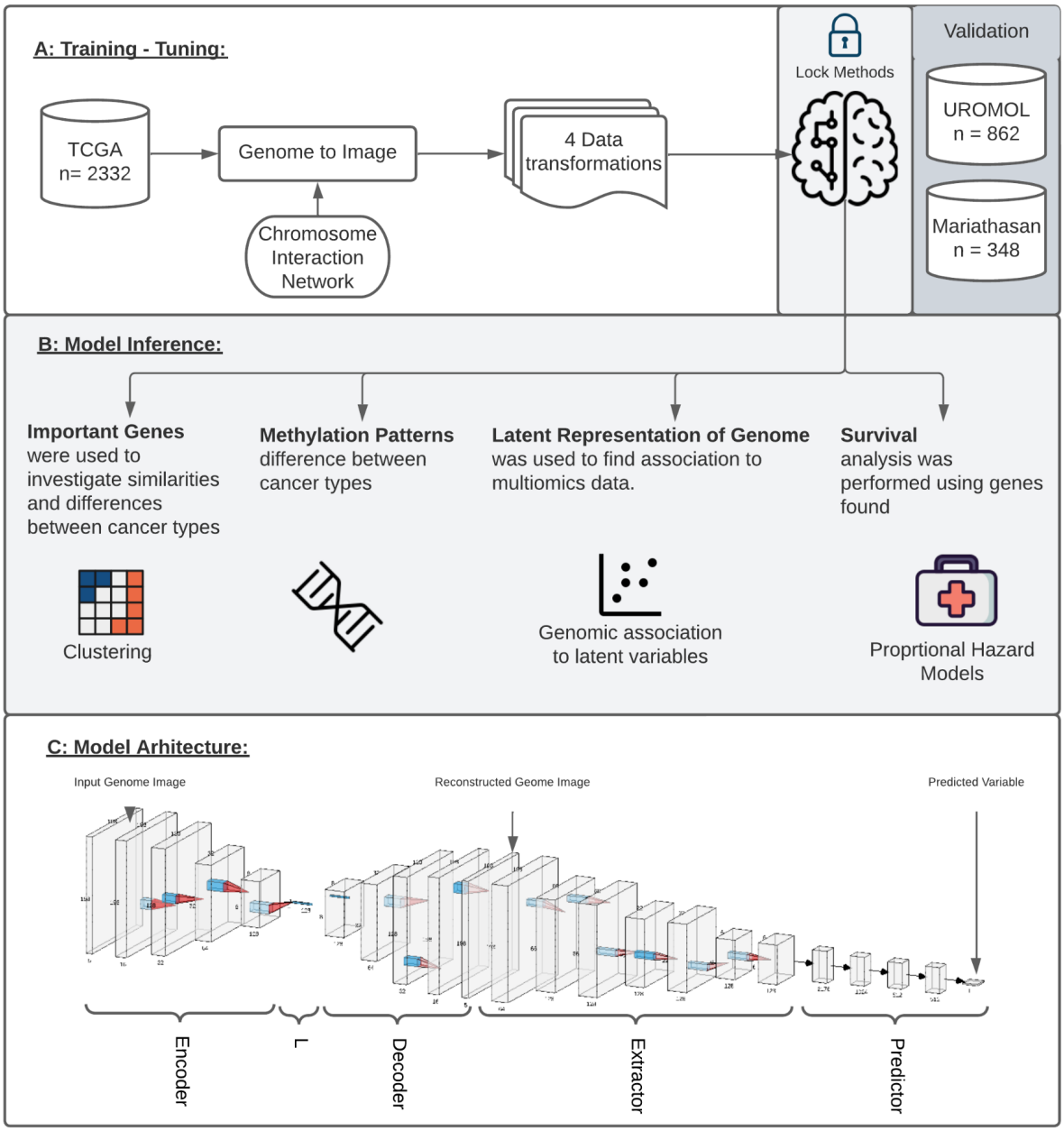


Figure 1. Study overview. (A) The study utilised 2332 tumour samples representing six cancer types (bladder, uterine, stomach, ovarian, kidney and colon) and transformed multiomics data into images using chromosome interaction network. After the model was trained, we validated found genes with two independent cohorts representing early stage BLCA (UROMOL) and late stage BLCA (Mariathasan) (B) The validation included looking at the most important genes driving metastatic disease, similar/different methylation patterns between cancer types, latent representation of genome data and looking at survival data. (C) The model architecture where the first part of the network encodes genome data into latent vector, L, followed by decoding where image is reconstructed. Next layers aim to extract information from the reconstructed image, concat it with L and make a final prediction.

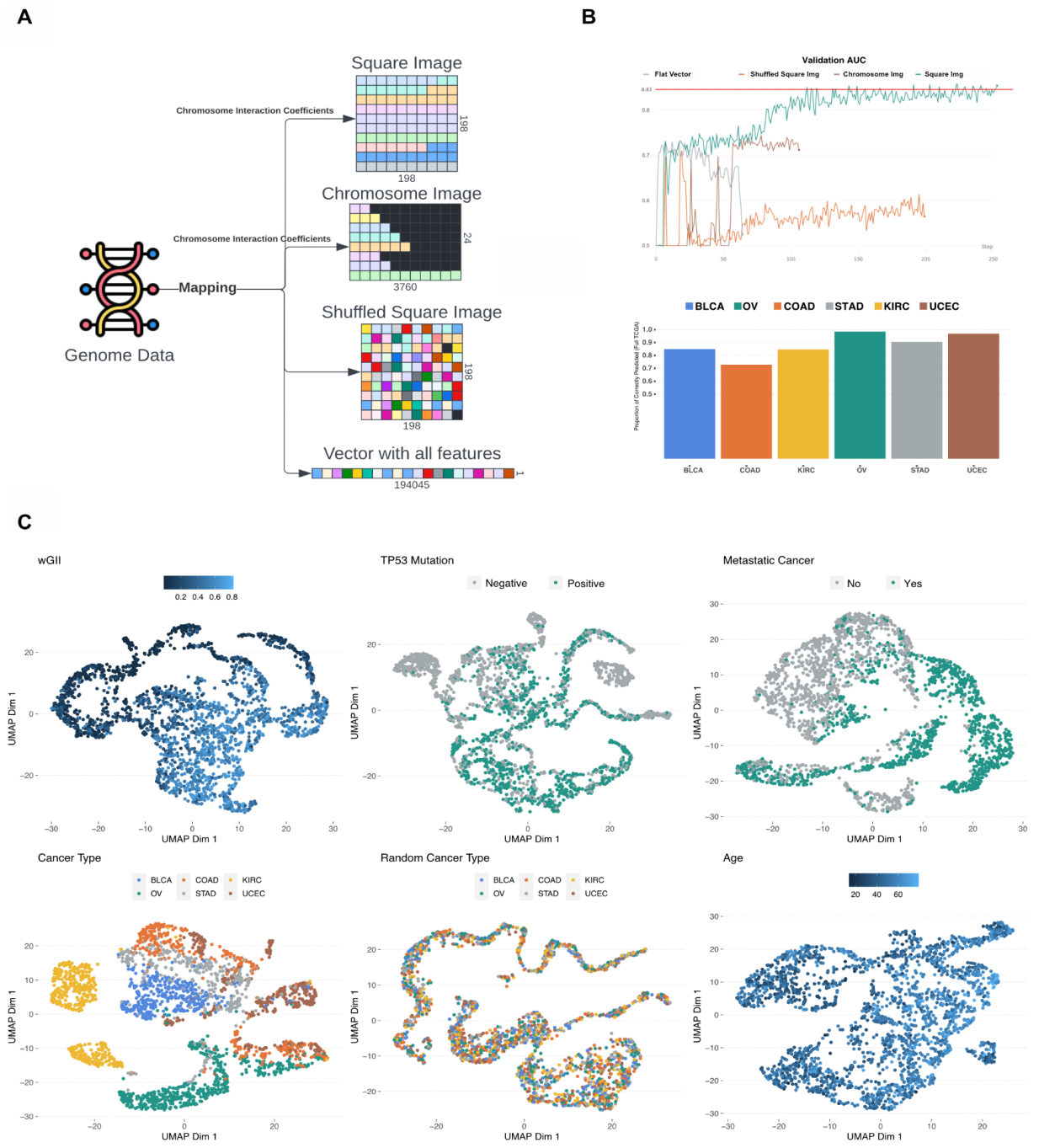
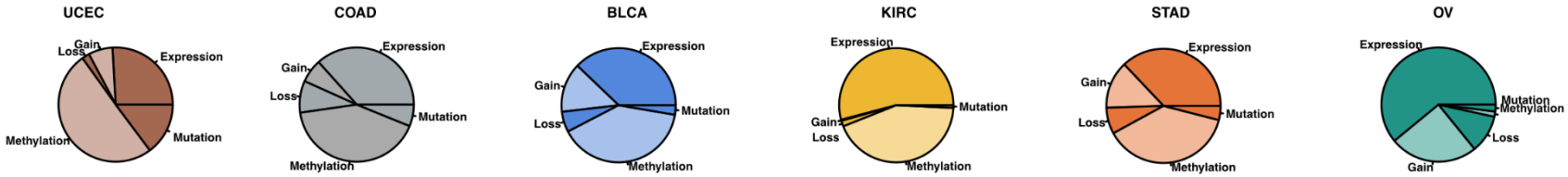


Figure 2. Data transformation overview. (A) The multiomics genome data was transformed into 4 image types; square image organised by chromosome interaction network, chromosome image organised by chromosome interaction network, randomly organised image and flat vector containing all multiomics data. (B) All four image types were used in training for metastatic disease prediction and the square image organised by chromosome interaction network resulted in best model performance (green colour). The bar plot shows the performance within each cancer type included in the study (C) 2-dimensional representation of vector L using UMAP for each predicted variable. Colours indicate the output variable which was used in the specific run

A



B

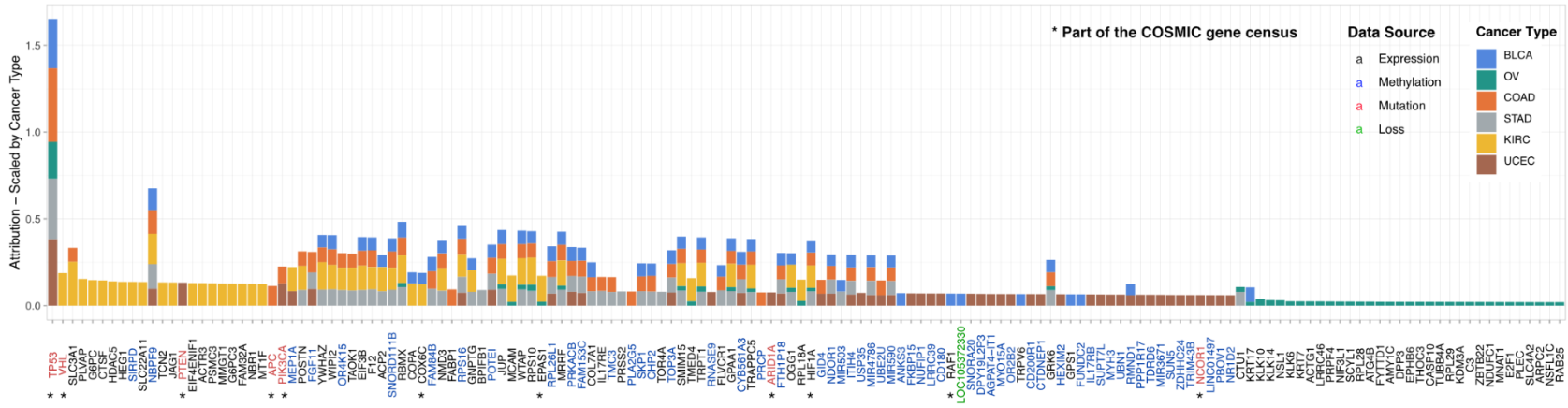


Figure 3. The most important events in metastatic disease development. (A) Pieplot showing the relative importance of each data source when predicting metastatic disease for each cancer type included in the study. (B) Top 50 genes for every cancer type scale by cancer type. The star symbol bellow the gene names indicate that gene is part of COSMIC gene consensus. The color of the gene name indicates the data source and color of the bar indicates the cancer type.

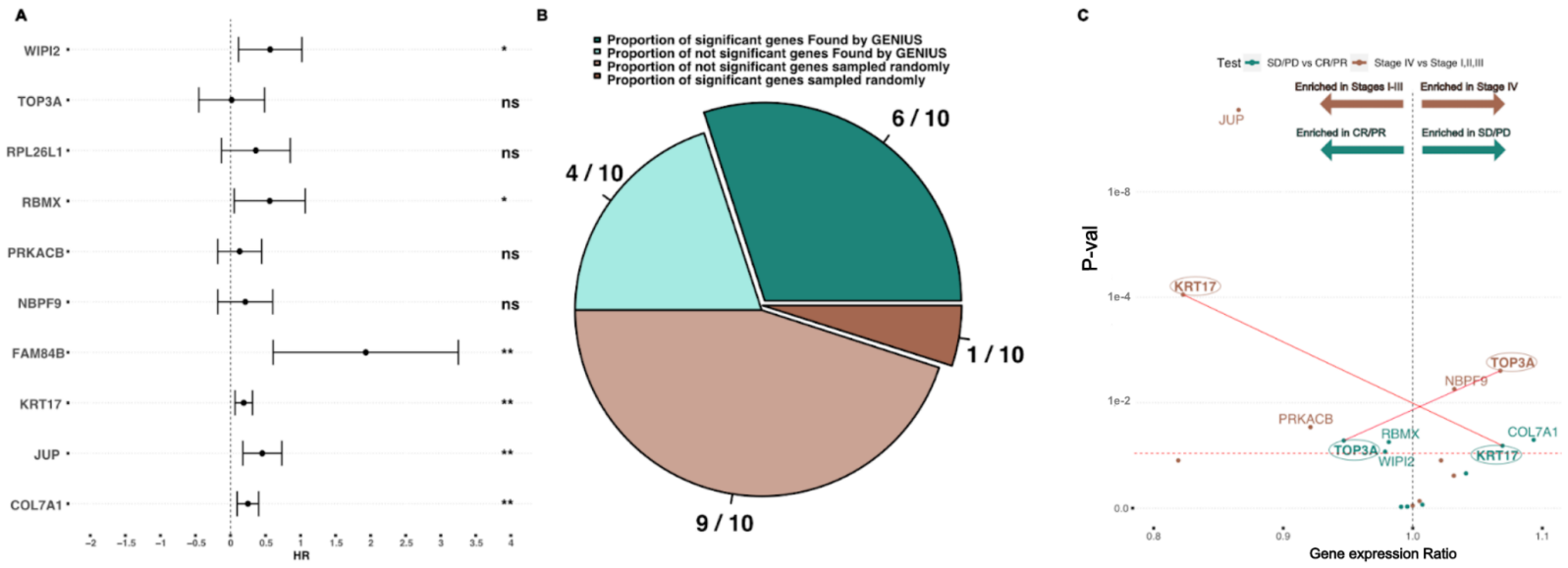


Figure 4. Validation on late stage immunotherapy treated bladder cancer (Mariathasan). (A) Comparison of randomly selected genes vs genes picked by GENIUS in cox proportional hazard model. (B) Forest plot showing top 10 expressed/methylated genes in multivariate cox proportional hazard model. (C) Volcano plot showing top 10 expressed/methylated genes and their enrichment in two comparisons; Stages I-III vs Stage IV and immunotherapy binary response.

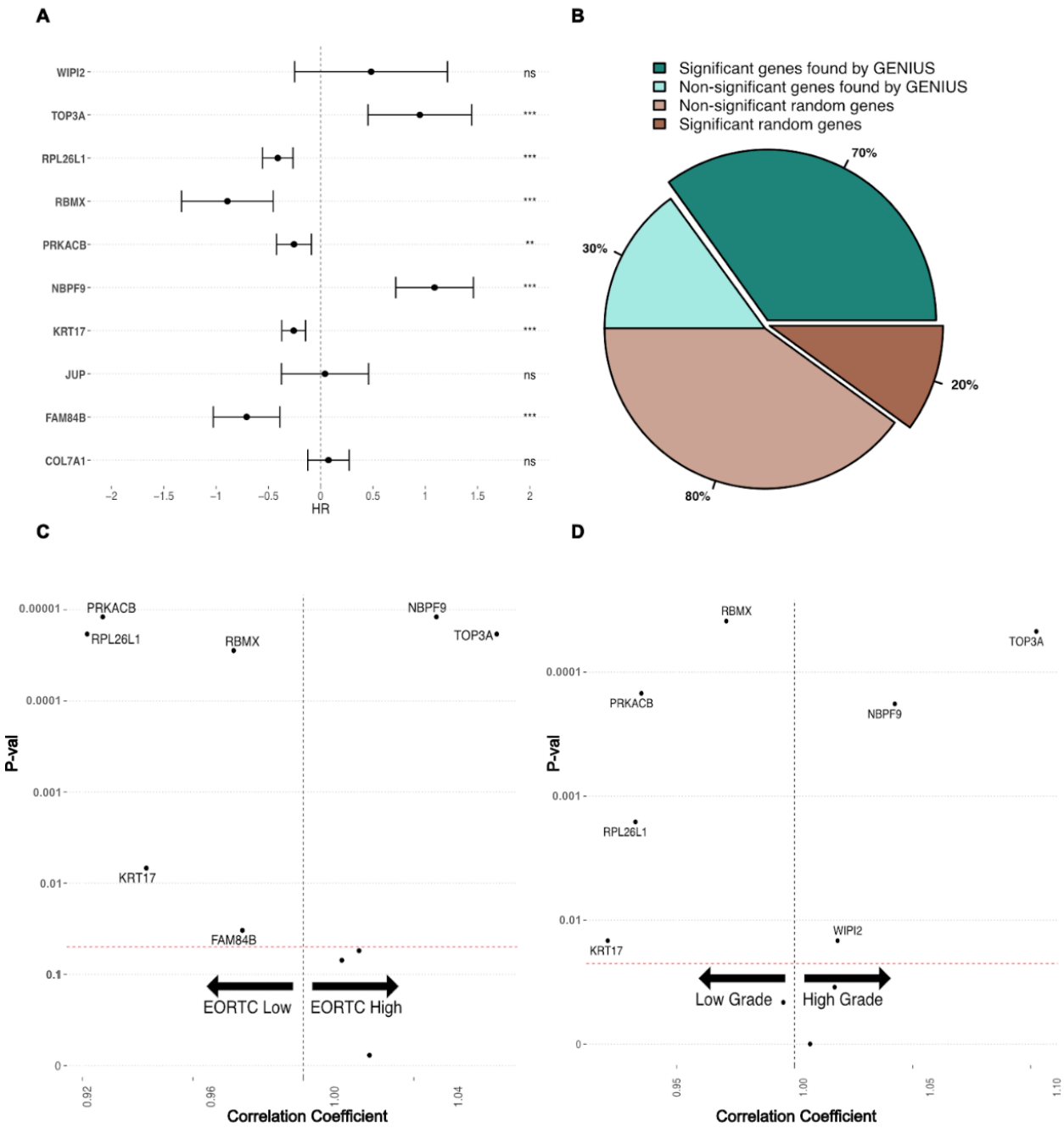


Figure 5. Validation on early stage bladder cancer (UROMOL). (A) Comparison of randomly selected genes vs genes picked by GENIUS in cox proportional hazard model. (B) Forest plot showing top 10 expressed/methylated genes picked by GENIUS for BLCA. (C) Volcano plot showing the top 10 expressed/methylated genes when compared between EORTC-Low and EORTC-High groups. (D) Volcano plot showing the top 10 expressed/methylated genes when compared between low grade and high grade BLCA tumours.

Table 1: Summary of BLCA genes in two validation cohorts

Gene	Early Stage	Late Stage (Immunotherapy)	Description
TOP3A	HR > 0 (PFS), High Grade, High EORTC	HR = 0 (OS), Enriched in Stage IV, Enriched in CR/PR	Catalyses the transient breaking and rejoining of a single strand of DNA, involved in regulation of recombination and homology directed repair. Positive association to OS in OV [79]
RBMX	HR < 0 (PFS), Low Grade, Low EORTC	HR > 0 (OS), Enriched in CR/PR	Associated with translational control and DNA damage pathways. Reported to be negatively correlated with tumour stage, histological grade and poor patient prognosis in BLCA [80]
PRKACB	HR < 0 (PFS), Low Grade, Low EORTC	HR = 0 (OS), Enriched in Stages I-III	Catalytic subunit of cAMP-dependent protein kinase, which mediates signalling through cAMP. Associated with cell proliferation. Associated with poor prognosis in colorectal cancer [81] Downregulated in NSCLC [82]
KRT17	HR < 0 (PFS) Low Grade, Low EORTC	HR > 0 (OS), Enriched in Stages I-III Enriched in SD/PD	Associated with structural molecule activity and MHC class II receptor activity. Associated with metastasis and angiogenesis in variety of tumour types [77,78]
WIPI2	HR = 0 (PFS), High Grade	HR > 0 (OS), Enriched in CR/PR	Component of the autophagy machinery that controls the major intracellular degradation process. WIPI2 is suggested as a biomarker for predicting colorectal cancer prognosis [83]
RPL26L1	HR < 0 (PFS), Low Grade, Low EORTC	HR = 0 (OS)	Associated with peptide chain elongation pathways
NBPF9	HR > 0 (PFS), High Grade, High EORTC	HR = 0 (OS), Enriched in Stage IV	Characterised by tandemly repeated copies of DUF1220 protein domains and implicated in urinary tract anomalies as well as in cancer.
JUP	HR = 0 (PFS)	HR > 0 (OS) Enriched in Stages I-III	Common junctional plaque protein. Controversial role in different malignancies. Knockdown of JUP in epithelium-like GC cells causes EMT and promotes GC cell migration and invasion [84]
FAM84B	HR < 0 (PFS), Low EORTC	HR > 0 (OS)	Associated with multiple cancer types [85–88]. Studies suggest a collaboration of FAM84B with Myc [89]
COL7A1	HR = 0 (PFS)	HR > 0 (OS) Enriched in SD/PD	Associated with metabolism of proteins and integrins in angiogenesis. Aberrant gene expression is associated with distinct tumour environment, metastasis and survival in multiple cancer types [90–92]