Genome-Wide DNA Methylation Model for the Diagnosis of Prostate Cancer

Jason B. Nikas and Emily G. Nikas

1. INTRODUCTION

1.1. Prostate Cancer and Its Diagnosis. Among males in the United States of America, prostate cancer is the most prevalent (approximately 165 000 new cases every year) and the second most lethal malignancy (approximately 29 000 deaths every year)\(^1\). The pathological diagnosis of prostate cancer via a biopsy is associated with an average error of 25−30% in the case of underdetection and an average error of 1.3−7.1% in the case of overdetection\(^2,3\), and in some cases, the underdetection error is considerably higher\(^4\). Furthermore, the accuracy of a Gleason score based on a scale used by pathologists to grade biopsied prostate tumors—its estimated be 61%\(^5\). All those different types of error associated with the current method of prostate cancer diagnosis, in all likelihood, stem from the observation that histopathological analysis of tissue (performed at the tissue level) does not have the requisite resolution to discern those molecular processes and mechanisms, such as epigenetic modifications, that initiate tumorigenesis and drive its early evolution (occurring at the molecular level)\(^6\).

1.2. Epigenetic Modifications. Epigenetic modifications are molecular mechanisms that provide cells with the means to regulate gene expression without altering the DNA, and they include DNA methylation, chromatin remodeling, microRNAs (miRNAs), long intergenic nonprotein coding RNAs (lincRNAs), small nucleolar RNAs (snoRNAs), etc. Chromatin remodeling pertains to covalent modifications of histones, proteins that act as spools around which DNA is wound and compacted. Chromatin-remodeling mechanisms include histone methylation, acetylation, phosphorylation, ubiquitination, etc., and some of them, such as H3K4me3 and H3K9ac, have gene-activating functions, whereas others have gene-inactivating functions, such as H3K9me3 and H3K27me3\(^6\). miRNAs regulate messenger RNA (mRNA) and can suppress gene expression either by degrading mRNA or by inhibiting translation\(^8\).

In addition to regulating mRNA, lincRNAs can activate or inactivate transcription-regulating genes and homeobox genes, and they are involved in transcription, translation, splicing, genomim imprinting, X-chromosome inactivation, etc.\(^9−11\) Dysregulation and aberrant expression of lincRNAs have been observed in the pathogenesis of numerous types of cancer\(^12,13\); lincRNAs have been associated with various cancer subtypes;\(^12\) and overexpression of lincRNAs has been observed to promote epithelial-to-mesenchymal transition and metastasis, both of which are characteristic of the most aggressive tumors.\(^13\) In this study, the lincRNA LINC01091, which was found to be significantly unmethylated in the tumor cells as compared with normal cells, is one of the input variables to the model.

snoRNAs modify mainly ribosomal RNA (rRNA), which is essential for protein synthesis, as well as transfer RNA (tRNA), small nuclear RNA (snRNA), and mRNA, and they are involved in ribosome biogenesis, splicing, and telomere maintenance.\(^14−16\) Just as in the case of lincRNAs, dysregulation of snoRNAs has been linked to various types
of cancer,\textsuperscript{17,18} and the overexpression of some of them has been observed to promote tumorigenesis and to be indirectly proportional to patient survival time.\textsuperscript{18} In this study, one of the five input variables to the model is the snRNA SNORA10, which was found to be significantly unmethylated in the tumor cells, and another input variable is the ribosomal protein RPS15, which was also found to be significantly unmethylated in the tumor cells.

1.3. DNA Methylation. One of the most common and important epigenetic modifications is DNA methylation, which is used by cells for gene-inactivation purposes. It involves the addition of a methyl group (–CH\textsubscript{3}) at specific locations of the DNA molecule, called CpG sites, where a cytosine nucleotide is bound to, and immediately followed by, a guanine nucleotide. CpG sites are highly concentrated in certain DNA areas, called CpG islands, where the vast majority of gene promoters lie; methylation of CpG islands results in stable gene inactivation.\textsuperscript{19} Within a distance of about 2 000 base pairs away from the CpG islands, and extending in both directions, there are DNA areas with a lower concentration of CpG sites, called CpG island shores.\textsuperscript{20} Extending farther, within a distance of about 2 000 base pairs away from the CpG island shores and in both directions, there are DNA areas with even lower concentration of CpG sites, called CpG island shelves.\textsuperscript{21} Finally, in the DNA area which corresponds to the gene body, and which extends beyond the first exon, there are also CpG sites.\textsuperscript{19} In normal cells, DNA methylation plays a most important role in embryonic stem cell differentiation, embryonic development, genomic imprinting, X-chromosome inactivation, transposable element inactivation, etc. In the case of tumor cells, on the other hand, the process of DNA methylation, like other epigenetic modifications and many other cellular processes, is hijacked, and it is used to activate (demethylate) genes that are vital or beneficial to them and to inactivate (methylate) genes that are detrimental to them. Aberrant alterations of DNA methylation, which include, among others, demethylation of genes that promote tumorigenesis and methylation of genes that suppress tumorigenesis, have been observed from CpG islands and the gene promoters to CpG island shelves and the CpG sites in gene bodies in many types of cancer.\textsuperscript{20,22,23} It is interesting to point out here that the average levels of genome-wide DNA methylation begin to decline years before cancer diagnosis.\textsuperscript{24}

1.4. Hypothesis. The hypothesis of this study is that tumor cells have effected significant changes in the methylation profile of a number of genes and that, consequently, these methylation changes can discriminate between tumor cells and normal cells.

1.5. Model Overview. In this study, we were able to develop a multivariable function, whose five input variables were the methylation $\beta$ values of five CpG sites, corresponding to the genes LINCO1091, RPS15, SNORA10, and two unknown DNA areas in chromosome 1 (genomic coordinates 201509316 and 220697615). The model (multivariable function) was developed during the training phase using approximately 70% of all available samples, that is, 329 tumor samples and 35 normal samples, and it was based on the methylation analysis of over 485 000 CpG sites per sample (spanning across gene promoters, CpG islands, shores, shelves, gene bodies, and intergenic and other areas). Following its development, the model was validated with unknown samples (approximately 30% of all available samples, i.e., 140 tumor samples and 15 normal samples), which had been randomly preallocated at the beginning of the study and used only for testing purposes. Subsequently, the model was further tested with a sixfold cross-validation and a leave-one-out cross-validation. The overall performance of the model, including the training and the validation phases, was sensitivity = (447/469) = 95.31% and specificity = (47/50) = 94.00%.

2. METHODS

2.1. Data. All data used in this study were the normalized data obtained from 469 prostatic tumor tissue samples and 50 normal prostatic tissue samples, generated from DNA methylation analysis of the tissue using the Infinium Human Methylation 450 Bead Chip by Illumina, and downloaded from The Cancer Genome Atlas of the National Cancer Institute under the category PRAD.

2.2. Clinical Methods. All tumor samples used in this study were primary tumors obtained from a biopsy used for the initial diagnosis and prior to the administration of any treatment. Furthermore, in this study, tumor samples were included only if they had been obtained from subjects that did not have a history of prior and/or other malignancy. The range of the Gleason scores of all 469 tumor samples was from 6 to 10 (Table S1). The Gleason grading system, with a possible score between 2 and 10, is currently used for the grading of prostate cancer and for the clinical prognosis. Patients with tumors with a Gleason score $\leq$ 6 are considered to have good prognoses (typically, no treatment is recommended),\textsuperscript{24} whereas patients with tumors with the highest scores are considered to have the worst prognoses.

2.3. Statistical Methods. Using only the data from the training set, a methylation analysis was performed. Given that the total number of variables (CpG sites) per sample was 485 577 and using the Bonferroni correction for multiple tests, the statistical significance level for the methylation analysis was set at $\alpha = (0.05/485 577) = 1.03 \times 10^{-7}$ (two-tailed). Therefore, in order for any variable to be deemed statistically significant, it must have a $P$ value $< 1.03 \times 10^{-7}$. The calculation of the $P$ value of a given variable was done in accordance with the individual conditions imposed by various statistical tests. More specifically, for those variables that were normally distributed with respect to both groups and met the equality of variance condition, the independent $t$-test was used to calculate their $P$ value. For those variables that were normally distributed with respect to both groups and failed to meet the equality of variance condition, the Welch unequal-variance $t$-test was used to calculate their $P$ value. Finally, for those variables that were not normally distributed with respect to both groups, the Mann–Whitney $U$ test was used to calculate their $P$ value. Moreover, in the case of the Mann–Whitney $U$ test, although there were no ties in any of the variables, the approximate method with correction was used because of large group sample sizes. The Anderson–Darling test was used to assess normality, and the Levene absolute test for equal variances was used to assess the equality of variance throughout this study. The sixfold cross-validation and the leave-one-out cross-validation were performed using the same methods as in one of our previous studies.\textsuperscript{25} Briefly, in the case of the former, six rounds (six folds) of training and testing were performed. In each round, approximately 5/6 of the total number of samples were randomly selected and used for training purposes, whereas the remaining approximately 1/6 of samples were used for testing. At the end, the misclassification rate and the mean-squared error were calculated based on the total number of misclassifications. In the case of the leave-one-out cross-validation, one sample was removed at a time, and the model was trained and tested on the remaining samples. This procedure was repeated for each sample in the dataset.
validation, the number of rounds was equal to the total number of samples (519). In each round, one sample was randomly left out and used for testing, and it was excluded from being selected again in any of the remaining rounds. At the end, the misclassification rate and the mean-squared error were calculated based on the total number of misclassifications.

In addition to the aforementioned criterion of statistical significance ($P < 1.03 \times 10^{-7}$), a criterion of biological significance was imposed. In order for a CpG site to be deemed significantly differentially methylated by the two groups, namely, $T$ (tumor samples) and $N$ (normal samples), the mean $\beta$ value of one group must be $\beta \geq 0.60$ and, at the same time, the mean $\beta$ value of the other group must be $\beta \leq 0.40$. This would ensure with a great confidence that a particular CpG site was methylated ($\beta \geq 0.60$) by one group and concurrently unmethylated ($\beta \leq 0.40$) by the other group. In summary, both the statistical significance criterion and the biological significance criterion had to be met in order for a particular CpG variable to be deemed significant. The rationale for this follows next.

### 2.4. $\beta$ Value

The methylation $\beta$ value for a CpG site is defined by Illumina\textsuperscript{21} as follows

$$\beta = \frac{M}{M + U + C_u}$$  \hspace{1cm} (1)

In eq 1, $M$ is the methylated signal detected by probes with respect to both alleles at a particular CpG site; $U$ is the unmethylated signal detected by probes with respect to both alleles at a particular CpG site; and $C_u$ is a normalization constant ($C_u > 0$). The scale of the $\beta$ value is continuous, and its range is the interval $[0, 1]$. For all practical purposes, $\beta$ may be considered as the methylation percentage of a particular CpG site with respect to both alleles expressed in the decimal form. A $\beta$ value equal to zero indicates that a CpG site is unmethylated (0% methylated) with respect to both alleles (in this case, $M = 0$ and $U = 1$), whereas a $\beta$ value very close to one indicates that a CpG site is 100% methylated with respect to both alleles (in this case, $M = 1$ and $U = 0$).

In this study, we were solely interested in those CpG sites that were methylated by one group (either $T$ or $N$) and, at the same time, were unmethylated by the other group. To state it differently and more broadly, the goal of this study was (a) to identify all the CpG sites in the human genome in which prostatic tumor cells effected significant methylation alterations and (b) to use those CpG sites to develop a genomic model that could identify with a high accuracy tumor tissue versus normal tissue. To that end, we imposed the aforementioned stringent biological criterion of significance. More specifically, we did not follow the widely used $\beta = 0.50$ point as the criterion of methylation, such that if $\beta > 0.50$, a particular CpG site is considered methylated, and if $\beta < 0.50$, that CpG site is considered unmethylated. In our view, that criterion is very precarious and could easily be unreliable and misleading. The technology is not perfect, and the wrong signal by only a few probes could lead to $\beta = 0.51$ instead of $\beta = 0.49$. In this case, the particular CpG site is erroneously assumed to be methylated. To avoid this type of error and to increase the confidence, we imposed the following and very conservative criterion of methylation: in order for a CpG site to be deemed methylated, it should have $\beta \geq 0.60$, and conversely, in order for a CpG site to be deemed unmethylated, it should have $\beta \leq 0.40$. This further means that in order to arrive at either a false-positive or a false-negative methylation conclusion, more than 20% of the probe signals would have to be erroneous, something which is not likely. Since, as was mentioned above, in this study, we were exclusively interested in those CpG sites that were methylated by one group (either $T$ or $N$) and, at the same time, were unmethylated by the other group, we imposed the aforementioned stringent biological criterion of significance, whereby in order for a CpG site to be deemed biologically significant, that CpG site had to be methylated by one group (with a $\beta \geq 0.60$) and, at the same time, had to be unmethylated by the other group (with a $\beta \leq 0.40$).

### 2.5. Development of the Model

Approximately, 70% of the total number of samples from each group were randomly selected for the training set, and the remaining approximately 30% of the samples were assigned to the validation set. Subsequently, the random selection of the tumor samples was modified in order for the following condition to be fulfilled. The training set had to comprise approximately 70% of the tumor samples with a particular Gleason score. To that end, approximately 70% of the tumor samples with a particular Gleason score were randomly selected and assigned to the training set, and the remaining approximately 30% were assigned to the validation set. Therefore, the training set comprised 329/469 tumor and 35/50 normal samples. In the 329 tumor samples, there were 30 with a Gleason score $GS = 6$; 163 with $GS = 7$; 42 with $GS = 8$; 91 with $GS = 9$; and 3 with $GS = 10$ (Table S2). The validation set comprised 140/469 tumor and 15/50 normal samples. In the 140 tumor samples, there were 13 with $GS = 6$; 69 with $GS = 7$; 18 with $GS = 8$; 39 with $GS = 9$; and 1 with $GS = 10$ (Table S3).

Of the 485 577 variables interrogated during the methylomic analysis using the data from the training set, 2 913 met both of the aforementioned criteria of significance and provided the variable pool for the development of the model. Using the same methodology as in our previous studies,\textsuperscript{25-28} we were able to develop the following multivariable function (eq 2).

Briefly, multiple functions, with at most ten input variables for each one of them, were generated and tested with computer simulations using the training set. Equation 2 exhibited the best performance (highest sensitivity and highest specificity) in the training phase and was selected for validation.

$$F = \ln\left(\left(\frac{(X_1 + 1)(X_2 + 1)^{1/10}(X_3 + 1)^{3/10}(X_4 + 1)}{(X_5 + 1)^{4/10}}\right)C\right)$$  \hspace{1cm} (2)

In eq 2, $X_1$ corresponds to the gene $RPS15$, and $X_2 = \beta_1 	imes 100$ (i.e., the $\beta$ value of $X_1$ expressed in the percentage form); $X_3$ corresponds to the snoRNA $SNORA10$, and $X_4 = \beta_3 	imes 100$; $X_5$ corresponds to the lncRNA $LINC01091$, and $X_6 = \beta_3 	imes 100$; $X_7$ corresponds to the unknown DNA area in chromosome 1 with the genomic coordinate of 220697615, and $X_8 = \beta_4 	imes 100$; and $X_9$ corresponds to the unknown DNA area in chromosome 1 with the genomic coordinate of 20159316, and $X_9 = \beta_5 	imes 100$. $C$ is a constant and is equal to $10^4$.

Since $X_1$ to $X_9$ are the respective $\beta$ values of the aforementioned genes expressed in the percentage form and since the value of each one of them belongs in the continuous interval $[0, 100]$ (eq 1), the domain of $F$ (eq 2) is the interval $[0, 100]$. Moreover, since $X_1 \geq 0$, the function $F$ is continuous and defined throughout its domain $[0, 100]$. From eq 2, one can see that $F$ attains its minimum value when $X_1 = X_2 = X_3 = X_4 = 0$ and $X_5 \approx 100$, and in that case, $F_{\text{min}} \approx 3.22$. Similarly, when
Table 1. DNA Methylation Analysis Results of the Five Input Variables of the Model Based on the Data of the Training Set

| var. | gene symbol | gene ID | gen. coord. | $M_T$ | $SD_T$ | $M_N$ | $SD_N$ | FC | $P$ |
|------|-------------|---------|-------------|-------|--------|-------|--------|-----|-----|
| $X_1$ | RPS15 | 6209 | 1440293 | 0.3830 | 0.1344 | 0.7145 | 0.0958 | $-1.8654$ | $2.41 \times 10^{-19}$ |
| $X_2$ | SNORA10 | 574042 | 2012763 | 0.3970 | 0.1302 | 0.6984 | 0.0960 | $-1.8046$ | $1.28 \times 10^{-19}$ |
| $X_3$ | LINC01091 | 285419 | 124694137 | 0.3255 | 0.1494 | 0.6525 | 0.0905 | $-2.0048$ | $4.74 \times 10^{-19}$ |
| $X_4$ | unknown | 220697615 | 0.2809 | 0.1618 | 0.6273 | 0.1007 | $-2.2332$ | $1.03 \times 10^{-18}$ |
| $X_5$ | unknown | 201509316 | 0.6082 | 0.2435 | 0.0794 | 0.1274 | $7.6578$ | $9.10 \times 10^{-19}$ |

The FC was calculated as follows: $R = M_T/M_N$. If $R \geq 1$, FC = $R$; if $R < 1$, FC = $-1/R$.

Figure 1. (A) ROC AUC curve of the model F in the training phase. (B) Combination graph (box plot, density plot, and dot plot) of the model F in the training phase. Red circles denote statistical outliers. 249 × 111 mm (300 × 300 DPI).

$X_1 = X_2 = X_3 = X_4 \approx 100$ and $X_5 = 0$, $F$ attains its maximum value, and in that case, $F_{\text{max}} \approx 29.01$. Therefore, the range of the function $F$ is the interval $(3.22, 29.01)$. Based on all 519 samples used in this study, the observed range of $F$ was $[10.146, 17.127]$ (Table S1).

The cutoff point (COP) that demarcates a score of a tumor sample from that of a normal sample was determined to be $\text{COP} = 15.258$ using a receiver operating characteristic (ROC) curve analysis on the results of the $F$ function in the training phase, as well as taking into account the respective variances of the two groups (T vs N). An $F$ score < COP signifies a tumor sample, whereas an $F$ score $\geq$ COP signifies a normal sample. The COP can be adjusted in the future within the constraints of ROC curve analysis in order to optimize the performance of the model in a subsequent clinical trial study. The methylation analysis revealed that in the case of tumor cells, $X_1$ was significantly unmethylated (low $\beta$ values) and $X_5$ was significantly methylated (high $\beta$ value) as compared with normal cells (Table 1). Taking those results into consideration, as well as the COP, one can see that the $F_{\text{max}}$ represents, at least theoretically, the worst possible methylomic state in which tumor cells could be. This makes biological sense because in that case, $X_1$ is completely unmethylated, while $X_5$ is completely methylated, and this methylomic state is the farthest from the normal one. In the opposite direction, the $F_{\text{min}}$ represents theoretically the most extreme methylomic state in which normal cells could be: $X_1$ is completely methylated, while $X_5$ is completely unmethylated. There are two more interesting possibilities that need to be considered here. In the case where $X_1 = X_2 = X_3 = X_4 = X_5 = 0$, eq 2 yields $F = 6.91$. According to the COP, this score, which is very low, corresponds to a tumor sample, and this makes biological sense because four out of five markers ($X_1$-$X_4$) exhibit the worst possible methylation profile that tumor cells could have—they are all completely unmethylated. Finally, in the case where $X_1 = X_3 = X_5 = 100$, eq 2 yields $F \approx 25.33$. This score, which is very high, corresponds to a normal sample, and from a biology perspective, it makes sense because four out of five markers ($X_1$-$X_4$) exhibit the most extreme methylomic profile that normal cells could have—they are all completely methylated.

2.6. Principal Component Analysis. Following the same methodology as the one in one of our previous studies and using the five variables ($X_1$-$X_5$), we performed a principal component analysis (PCA). The scores of the five principal components (PC1–PC5) are listed in Table S4.

3. RESULTS

3.1. Methylation Analysis. The methylation results in connection with the five input variables of the model based on the data of the training set appear in Table 1, wherein the variable name, the gene symbol, the NCBI gene ID, the genomic coordinate, the mean $\beta$ value ($M_T$) of the T (tumor) group with its standard deviation (SD$_T$), the mean $\beta$ value ($M_N$) of the N (normal) group with its standard deviation (SD$_N$), the fold change (FC), and the probability of significance ($P$) for all five variables are listed.

3.2. Training. In the training phase, the model (eq 2) identified correctly 312/329 tumor samples and 33/35 normal samples (Table S2). The sensitivity, therefore, was 94.83%, and the specificity was 94.29%. The ROC AUC statistics were as follows: AUC = 0.97647, its standard error was AUC SE = 0.00964, z-value = 49.41, and the 95% confidence interval of the AUC value was $[0.94773, 0.98949]$ (Figure 1A). The T group had a mean $F$ score and a standard deviation of $F_T = 13.5061 \pm 1.2254$, and its median value was 13.5457. Using 100 000 Monte Carlo simulations and 100 000 Bootstrap simulations, the 99% confidence interval of the $F_T$ was calculated to be $[13.3315, 13.6799]$. The N group had a mean $F$ score and a standard deviation of $F_N = 16.2719 \pm 0.5735$, and its median value was 16.3988. Using 100 000 Monte Carlo simulations and 100 000 Bootstrap simulations, the 99% confidence interval of the $F_N$ was calculated to be $[16.0590, 16.5467]$. The Mann–Whitney U test statistics were
as follows: \( U_T = 271, U_N = 11244, \text{z-value} = 9.2698, \) and the approximate probability of significance with correction was \( P = 1.87 \times 10^{-20} \) (Figure 1B).

3.3. Validation with Unknown Samples. Using the 155 preallocated unknown samples (140 tumor and 15 normal samples), the model (eq 2) identified correctly 135/140 tumor samples and 14/15 normal samples (Table S3). The sensitivity, therefore, was 96.43%, and the specificity was 93.33%. The unknown T samples had a mean F score and a standard deviation of \( F_T = 13.5759 \pm 1.1354, \) and their median value was 13.7448. The unknown N samples had a mean F score and a standard deviation of \( F_N = 16.1719 \pm 0.8585, \) and their median value was 16.3709. The Mann–Whitney U test statistics for the two groups of unknown samples in connection with their F scores were as follows: \( U_T = 91, U_N = 2009, \text{z-value} = 5.8011, \) and the approximate probability of significance with correction was \( P = 6.59 \times 10^{-9}. \)

3.4. Overall Performance. Combining the 364 samples (329 tumor and 35 normal) from the training phase with the 155 unknown samples (140 tumor and 15 normal) from the validation phase, the model’s overall performance was as follows: sensitivity = \( (447/469) = 95.31\% \) and specificity = \( (47/50) = 94.00\% \) (Figure 2 and Table S1). The ROC AUC was calculated to be \([13.3848, 13.6706]\). The N group had a mean F score and a standard deviation of \( F_N = 16.2419 \pm 0.6640, \) and its median value was 16.3939. Using 100 00 Monte Carlo simulations and 100 000 Bootstrap simulations, the 99% confidence interval of the \( F_N \) was calculated to be \([16.0343, 16.5067]\). The Mann–Whitney U test statistics were as follows: \( U_T = 691, U_N = 22759, \text{z-value} = 10.9454, \) and the approximate probability of significance with correction was \( P = 6.99 \times 10^{-28}. \) The range of all 519 F scores was \([10.146, 17.127]\) (Figure 3B). Finally, Table 2 shows the DNA methylation results of the five input variables of the model based on the combined data of the training and the validation sets. The variable name, the gene symbol, the NCBI gene ID, the genomic coordinate, the mean \( \beta \) value \( (M_T) \) of the T (tumor) group with its standard deviation (SDT), the mean \( \beta \) value \( (M_N) \) of the N (normal) group with its standard deviation (SDN), the FC, and the probability of significance \( (P) \) for all five variables are listed.

3.5. Cross-Validations. The sixfold cross-validation yielded a misclassification rate of 9.63% \((50/519)\) and a mean-squared error of 0.096. The sensitivity was 0.900, and the specificity was 0.940. The Matthews correlation coefficient (MCC) was 0.643. The training and testing sets for each of the six rounds, along with the confusion matrix and other statistical information, appear in Table S5. Similarly, the leave-one-out cross-validation yielded a misclassification rate of 9.63% \((50/519)\), a mean-squared error of 0.096, a sensitivity of 0.900, a specificity of 0.940, and an MCC of 0.643. The results of all 519 rounds, along with the confusion matrix and other statistical information, appear in Table S6.

4. DISCUSSION

Recent, rapid advances in biotechnology have ushered in a genomic era. For the first time, researchers have acquired the capability to probe and monitor cellular events at the molecular level with unprecedented resolution and accuracy. Considering that the onset of any dysregulation of any cellular function results immediately in molecular changes and considering that dysregulation of any cellular function might take years to manifest itself at the tissue level, the importance of the newly acquired genomic capabilities becomes obvious, and so does the need to transition to genomic medicine.

Utilizing one of those recent advances in biotechnology, namely, genome-wide DNA methylation, we developed and presented in this study a novel genomic test, operating at the molecular level, that could be used in the clinic to aid in the
The FC was calculated as follows: $R = M_T/M_N$. If $R \geq 1$, FC = $R$; if $R < 1$, FC = $-1/R$. The table contains the DNA Methylation Analysis Results of the Five Input Variables of the Model Based on the Combined Data of the Training and the Validation Sets:

| var. | gene symbol | gene ID   | gen. coord. | $M_T$ | $SD_T$ | $M_N$ | $SD_N$ | FC    | $P$     |
|------|-------------|-----------|-------------|-------|--------|-------|--------|-------|---------|
| $X_1$ | RPS15       | 6209      | 1440293     | 0.3865| 0.1315 | 0.7077| 0.1038 | -1.8308| $2.79 \times 10^{-26}$ |
| $X_2$ | SNORA10     | 574042    | 2012763     | 0.3928| 0.1298 | 0.6908| 0.1045 | -1.7586| $7.15 \times 10^{-26}$ |
| $X_3$ | LINC01091   | 285419    | 124694137   | 0.3293| 0.1495 | 0.6534| 0.1053 | -1.9843| $7.54 \times 10^{-26}$ |
| $X_4$ | unknown     |           | 220697615   | 0.2821| 0.1588 | 0.6171| 0.1218 | -2.1876| $6.45 \times 10^{-25}$ |
| $X_5$ | unknown     |           | 201509316   | 0.6055| 0.2369 | 0.0766| 0.1194 | 7.9034 | $6.26 \times 10^{-27}$ |

The authors declare no competing financial interest.

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.9b01613.

Overall performance of the model; training of the model; validation of the model with unknown samples; PCA—scores of principal components PC1–PC5; sixfold cross-validation of the model; and leave-one-out cross-validation of the model (XLSX)

**ASSOCIATED CONTENT**

**Supporting Information**

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsomega.9b01613.

**AUTHOR INFORMATION**

**Corresponding Author**

E-mail: jbnikas@genomix-inc.com.

**ORCID**

Jason B. Nikas: 0000-0001-9703-0422

**Notes**

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

This study was supported by Genomix Inc. The funding agency had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
REFERENCES

(1) American Cancer Society. Cancer Facts & Figures 2018; American Cancer Society, 2018.

(2) Ukimura, O.; Coleman, J. A.; de la Taille, A.; Emberton, M.; Epstein, J. I.; Freedland, S. J.; Giannarini, G.; Kibel, A. S.; Montironi, R.; Ploussard, G.; Roobol, M. J.; Scattoni, V.; Jones, J. S. Contemporary role of systematic prostate biopsies: indications, techniques, and implications for patient care. Eur. Urol. 2013, 63, 214–230.

(3) Graif, T.; Loeb, S.; Roehl, K. A.; Gashi, S. N.; Griffin, C.; Yu, X.; Catalona, W. J. Under diagnosis and over diagnosis of prostate cancer. J. Urol. 2007, 178, 88–92.

(4) Piert, M.; Montgomery, J.; Kunju, L. P.; Siddiqui, J.; Rogers, V.; Rajendiran, T.; Johnson, T. D.; Shao, X.; Davenport, M. S. 18F-Choline PET/MRI: The Additional Value of PET for MRI-Guided Transrectal Prostate Biopsies. J. Nucl. Med. 2016, 57, 1065–1070.

(5) Goodman, M.; Ward, K. C.; Osunkoya, A. O.; Datta, M. W.; Luthringer, D.; Young, A. N.; Marks, K.; Cohen, V.; Kennedy, J. C.; Haber, M. J.; Amin, M. B. Frequency and determinants of disagreement and error in gleason scores: a population-based study of prostate cancer. Prostate 2012, 72, 1389–1398.

(6) Hemberger, M.; Dean, W.; Reik, W. Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington’s canal. Nat. Rev. Mol. Cell Biol. 2009, 10, 526–537.

(7) Vermeulen, M.; Eberl, H. C.; Matarase, F.; Marks, H.; Denissov, S.; Butler, F.; Lee, K. C.; Olsen, J. V.; Hyman, A. A.; Stunnenberg, H. G.; Mann, M. Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. Cell 2010, 142, 967–980.

(8) Bartel, D. P. MicroRNAs. Cell 2004, 116, 281–297.

(9) Yoon, J.-H.; Abdelmohsen, K.; Gorospe, M. Posttranscriptional gene regulation by long noncoding RNA. J. Mol. Biol. 2013, 425, 3723–3730.

(10) Rinn, J. L.; Chang, H. Y. Genome regulation by long noncoding RNAs. Annu. Rev. Biochem. 2012, 81, 145–166.

(11) Woo, C. J.; Kingston, R. E. HOTAIR lifts noncoding RNAs to new levels. Cell 2007, 129, 1257–1259.

(12) White, N. M.; Cabanské, C. R.; Silva-Fisher, J. M.; Dang, H. X.; Govindan, R.; Maher, C. A. Transcriptome sequencing reveals altered long intergenic non-coding RNAs in lung cancer. Genome Biol. 2014, 15, 429.

(13) Gupta, R. A.; Shah, N.; Wang, K. C.; Kim, J.; Horlings, H. M.; Wong, D. J.; Tsai, M.-C.; Hung, T.; Argani, P.; Rinn, J. L.; Wang, Y.; Brzoska, P.; Kong, B.; Li, R.; West, R. B.; van de Vijver, M. J.; Sukumar, S.; Chang, H. Y. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 2010, 464, 1071–1076.

(14) Kiss, T. Small Nucleolar RNAs. Cell 2002, 109, 145–148.

(15) Dragon, F.; Lemay, V.; Trahren, C. snoRNAs: biogenesis, structure and function. Encyclopedia of Life Sciences, 2006.

(16) Kishore, S.; Stamm, S. The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. Science 2006, 311, 230–232.

(17) Williams, G. T.; Farzaneh, F. Are snoRNAs and snoRNA host genes new players in cancer? Nat. Rev. Cancer 2012, 12, 84–88.

(18) Mei, Y.-P.; Liao, J.–P.; Shen, J.; Yu, L.; Liu, B.-L.; Liu, L.; Li, R.-Y.; Ji, L.; Dorsey, S. G.; Jiang, Z.-R.; Katz, R. L.; Wang, J.-Y.; Jiang, F. Small nucleolar RNA 42 acts as an oncogene in lung tumorigenesis. Oncogene 2012, 31, 2794–2804.

(19) Moore, L. D.; Le, T.; Fan, G. DNA methylation and its basic function. Neuropsychopharmacology 2013, 38, 23–38.

(20) Irazarry, R. A.; Ladd-Acosta, C.; Wen, B.; Wu, Z.; Montano, C.; Onyango, P.; Cui, H.; Gabo, K.; Rongione, M.; Webster, M.; Ji, H.; Pothas, J. B.; Sabunciyan, S.; Feinberg, A. P. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat. Genet. 2009, 41, 178–186.

(21) Bibikova, M.; Barnes, B.; Tsan, C.; Ho, V.; Klotzle, B.; Le, J. M.; Delano, D.; Zhang, L.; Schroth, G. P.; Gunderson, K. L.; Fan, J.-B.; Shen, R. High density DNA methylation array with single CpG site resolution. Genomics 2011, 98, 288–295.

(22) Huang, T. H.-M.; Perry, M. R.; Laux, D. E. Methylation profiling of CpG islands in human breast cancer cells. Hum. Mol. Genet. 1999, 8, 459–470.

(23) van Veldhoven, K.; Polidoro, S.; Baglietto, L.; Severi, G.; Sacerdote, C.; Panico, S.; Mattiello, A.; Palli, D.; Masala, G.; Krogh, V.; Agnoli, C.; Tumino, R.; Frasca, G.; Flower, K.; Curry, E.; Orr, N.; Tomczyk, K.; Jones, M. E.; Ashworth, A.; Swerdlow, A.; Chadeau-Hyam, M.; Lund, E.; Garcia-Closas, M.; Sandanger, T. M.; Flanagan, J. M.; Vineis, P. Epigenome-wide association study reveals decreased average methylation levels years before breast cancer diagnosis. Clin. Epigenet. 2015, 7, 67.

(24) American Cancer Society. Initial Treatment of Prostate Cancer, by Stage. https://www.cancer.org/cancer/prostate-cancer/treating/ by-stage.html (accessed 2019).

(25) Nikas, J. B. Inflammation and immune system activation in aging: a mathematical approach. Sci. Rep. 2013, 3, 3254.

(26) Nikas, J. B.; Boylan, K. L.; Skubitza, A. P.; Low, W. C. Mathematical prognostic biomarker models for treatment response and survival in epithelial ovarian cancer. Canc. Inf. 2011, 10, 233–247.

(27) Nikas, J. B. A mathematical model for short-term vs. long-term survival in patients with glioma. Am. J. Cancer Res. 2014, 4, 862–873.

(28) Nikas, J. B. Independent validation of a mathematical genomic model for survival of glioma patients. Am. J. Cancer Res. 2016, 6, 1408–1419.

(29) Nikas, J. B.; Low, W. C. ROC-supervised principal component analysis in connection with the diagnosis of diseases. Am. J. Transl. Res. 2011, 3, 180–196.

(30) Toshimitsu, H.; Hashimoto, K.; Tangoku, A.; Iizuka, N.; Yamamoto, K.; Kawauchi, S.; Oga, A.; Furuya, T.; Oka, M.; Sasaki, K. Molecular signature linked to acquired resistance to cisplatin in esophageal cancer cells. Cancer Lett. 2004, 211, 69–78.

(31) Nakamura, T.; Nakatsu, N.; Yoshiida, Y.; Yamazaki, K.; Dan, S.; Sadahiro, S.; Makuch, H.; Yamori, T. Identification of candidate genes determining chemosensitivity to anti-cancer drugs of gastric cancer cell lines. Biol. Pharm. Bull. 2009, 32, 1936–1939.

(32) Bretones, G.; Álvarez, M. G.; Arango, J. R.; Rodríguez, D.; Nadeu, F.; Prado, M. A.; Valdés-Mas, R.; Puente, D. A.; Paulo, J. A.; Delgado, J.; Villamor, N.; López-Guillermo, A.; Finley, D. J.; Gygi, S. P.; Campo, E.; Quesada, V.; López-Otín, C. Altered patterns of global protein synthesis and translational fidelity in RPS15-mutated chronic lymphocytic leukemia. Blood 2018, 132, 2375–2388.

(33) Landau, D. A.; Tausch, E.; Taylor-Weiner, A. N.; Stewart, C.; Reiter, J. G.; Bahlol, J.; Kluth, S.; Boic, I.; Lawrence, M.; Böttcher, S.; Carter, S. L.; Cibulskis, K.; Mertens, D.; Sougnez, C. L.; Rosenberg, M.; Hess, J. M.; Edelmann, J.; Kless, S.; Kneba, M.; Ritgen, M.; Fink, A.; Fischer, K.; Gabriel, S.; Lander, E. S.; Nowak, M. A.; Döhner, H.; Hallem, M.; Neuberg, D.; Getz, G.; Stilgenbauer, S.; Wu, C. J. Mutations driving CLL and their evolution in progression and relapse. Nature 2015, 526, 525–530.