

NIMBus: a Negative Binomial Regression based Integrative Method for Mutation Burden analysis

Jing Zhang^{1,2}, Jason Liu^{2,3}, Lucas Lochovsky¹, Jayanth Krishnan¹, Donghoon Lee¹, Mark Gerstein^{1,2,4*}

¹Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA

²Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA

³Program in Applied Math, Yale University, New Haven, Connecticut 06520, USA

⁴Department of Computer Science, Yale University, New Haven, Connecticut 06520, USA

* To whom correspondence should be addressed. Tel: +1 203 432 6105; Fax: +1 360 838 7861; Email: Mark.Gerstein@Yale.edu

ABSTRACT

Identification of somatic mutation hotspots as potential driver candidates in disease genomes is challenging because background mutation rate is usually heterogeneous and severely confounded by many known genomic features. Limited annotation information, especially in noncoding regions, further hinders the interpretability of the discovered candidates. Here, we address these issues with a new computational framework called NIMBus. It treats the varying mutation rates from different samples as a random variable with scaled gamma distribution and therefore models the over dispersed mutation count data by negative binomial distribution. To remove the confounding effect, we extracted 381 features from REMC and ENCODE in all available tissues and utilized a negative binomial regression to accurately estimate background mutation rate. Such integrative framework in NIMBus is flexible and can be immediately extended to accommodate any number of new features in the future. We also provided the most comprehensive noncoding annotation from ENCODE inside NIMBus to help users to better interpret the underlying biological mechanisms for the discovered targets. We applied NIMBus on 649 whole-genome cancer sequences and it successfully identified well-known noncoding drivers, such as the TERT promoter. We make LARVA available as a software tool, and release our results as an online resource (nimbus.gersteinlab.org).

1. Introduction

With the rapid development of high throughput sequencing technologies, thousands of whole genome sequencing (WGS) data for patients with various diseases considerably

EXPL
PRESENT
?

REWORD

2
,

LOGK

increases the statistical power to dissect the mutation landscape at an unprecedented resolution. Frequently, scientists need to decide whether or not a region in a disease genome has more than the expected number of mutations, with somatic burden tests, to discover potential driver events that lead to complex diseases, such as cancer. Therefore, an accurate quantification of mutation burden is important to uncover the genetic cause of various diseases and then allow for targeted therapies in clinical studies. However, mutation burden test for somatic variants remains a challenge for several reasons.

First, many *state-of-the-art* methods are optimally designed for designed for analysis of coding regions, which usually represents less than 2 percent of the human genome. Nowadays, a myriad of studies have shown that noncoding mutations could serve as driver events for diseases. For example, the well-known noncoding mutations in TERT promoter were found to be associated with cancer progression in multiple cancer types. Hence, a unified coding and noncoding analysis is needed to annotate the discovered hotspots.

Second, some of the pioneer work analyzing WGS assumes a constant mutation rate across different regions or cancer genomes. However under this assumption, the mutation count data at the positional level often demonstrates larger than expected variances, resulting in many false positives and negatives. As a result, more sophisticated statistical models need to be introduced to handle such mutation heterogeneity in the somatic burden analysis.

Lastly, numerous functional characteristic features, such as replication timing, chromatin organization, and sequence context information, have been reported to largely affect somatic mutation process. Therefore, accurate mutational burden analysis should remove such covariate effect. Unfortunately, none of the few current methods that integrate limited covariates demonstrates acceptable model extensibility. For example, MutsigCV was one of the first methods to consider covariates in coding region burden tests by estimating local background mutation rate using a small neighborhood of genes with similar covariates. However, as the number of covariates increases, the 'curse of dimensionality' will make it almost impossible to find a meaningful neighborhood in high dimension space. Similarly, Lochovsky *et al* only corrected replication timing with relatively low resolution. A large number of other genomic features were left behind in this methodology, and it is difficult to add them for further analysis under the same framework. Melton *et al* utilized a logistic regression framework to estimate a patient specific mutation rate at a single nucleotide resolution. Similar to MutsigCV, it only considered a limited number of covariates that explain a small part of variation in the mutation count data. Additionally in some quiet genomes with very few mutations such as Pilocytic Astrocytoma, usually there are less than 500 mutations across a single genome, resulting in poor model performance.

In this article, we propose a Negative Binomial Regression based Integrative Method for Mutation Burden analysis (NIMBus) that solves the three challenges mentioned above. It intuitively treats different somatic mutation rates from various disease genomes as random variables with a scaled gamma distribution, and resultantly models the mutation counts by a type I negative binomial distribution to handle overdispersion. In addition, to

DISCUSS

REF

capture the background mutational rate variation due to other genomic features, we integrate the most extensive features in all possible tissues from the Roadmap Epigenomics Mapping Consortium (REMC) to create a covariate table and use a negative binomial regression to predict the local mutation rate with high precision. We also summarized the most comprehensive noncoding annotations from the ENCODE project and used these biologically meaningful blocks as natural units to facilitate the understanding of our results. The integrative approach employed in NIMBus enables us to effectively pinpoint the mutation hotspots that are solely related with disease progression and better interpret their biological mechanisms.

2. Methods

2.1 WGS variants data used

We collected 649 whole genome variant calls from public resources and our collaborators. This data set contains a broad spectrum of 7 different cancer types, including breast cancer (BRCA), gastric cancer (GACA), liver cancer (LICA), lung cancer (LUAD), prostate cancer (PRAD), Medulloblastoma (MB), and Pilocytic Astrocytoma (PA). Among these samples, 100 stomach cancer samples were from \cite{Wang}, and 95 prostate cancer samples were obtained from our collaborators. The remaining comes from samples published in Alexandrov.

2.2 Covariate table

Numerous studies showed that several functional characteristics data will affect the local mutation rate, and such covariate effect should be removed during somatic burden analysis. In step one, we collected all the signal files from major histone modification marks, chromatin status, methylation, and mRNA-seq data from the REMC at a 20 nucleotides resolution. We then processed the signal files into bigWig format (<https://genome.ucsc.edu/goldenpath/help/bigWig.html>) for every available tissue. If multiple replicates were found in their original data, the averaged signal after normalization was used in the final bigwig file. We only used the experimental real data as opposed to imputed data. If some file, for example the chromatin accessibility, is missing in a specific tissue, we skipped it in our final covariate table. Since replication timing has been proved to be associated with mutation rate in several articles, we also collected 8 replication timing bigWig files from the ENCODE project. Lastly, as researchers have observed elevated mutation rates in regions with lower GC content in certain diseases, we also include the GC percentage files in our covariate list.

In step two, we aim to provide effective training of our model that is convenient for users. Different from the calibrated training data selection mentioned in XXX, we divided the whole genome into bins at a fixed length, such as 1mb, 100kb, 50kb, etc. Only autosomal chromosomes 1-22 and chromosome X were included in our analysis to remove the gender imbalance in either the mutation data or the covariates tissues.

Repetitive regions on human genome are known to generate artifacts in high throughput sequencing analysis mainly due to low mappability. We downloaded the mappability consensus excludable table used in the ENCODE project from <http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz>. Any generated fixed length bins that overlap with this table would be removed from the training process. We also downloaded the gap regions of hg19 from the UCSC genome browser, which include gaps from telomere, short_arm, heterochromatin, contig, and scaffold. The fixed length bins that intersect with these gap regions were also removed in the training process.

All of the bigWig files generated in step one were used to calculate the average signal using the bigWigAverageOverBed tool for each fixed length bin we generated in step two. When calculating the GC percentage, if the sequence information is not available at a certain position (such as the Ns), such position will be excluded in the averaging process. In the end, we summarized all the covariates values in each bin into a covariate table, with columns indicating different features and rows representing different training bins.

2.3 Noncoding annotations lists from ENCODE projects

We collected the full list of noncoding annotations, which include promoter regions, transcription start sites (TSS), translated regions (UTR), transcription factor binding sites (TFBS), enhancers, ultra-conserved, and ultra-sensitive sites. Promoters and TSS sites were defined as the 2500 and 100 nucleotides (nt) before the genes that are annotated by GENCODE v16. We also collected all the TFBS and enhancers from all tissues that are uniformly processed through the ENCODE pipeline. In addition, the ultra-conserved and ultra-sensitive sites were defined as those under positive selection during transcription regulations in our previous method FunSeq.

2.4 Negative binomial based burden analysis model

When pulling the mutation counts from disease genomes of the same molecular subtype together, we assume that given the mutation rate λ in a specific region, the mutation counts y in this region should follow a Poisson distribution with the probability mass function (PMF) given in equation (1).

$$p_Y(y|\lambda) = \frac{e^{-\lambda} \lambda^y}{y!} \quad (1).$$

However, as we mentioned in previous sections there is huge heterogeneity in the mutation counts data in different disease genomes. Hence, instead of modeling a constant mutation rate in **XXX**, we assumed that the mutation rate λ itself is a scaled gamma random variable in equation (2).

$$\lambda = \mu\gamma, \gamma \sim \text{Gamma}\left(1, \frac{1}{\sigma^2}\right) \quad (2).$$

SAMPLES

λ k

Therefore, the marginal distribution of y can be represented by a negative binomial distribution of type I with PMF listed in (3).

$$p_Y(y|\mu, \sigma) = \frac{\Gamma(y + 1/\sigma)}{\Gamma(1/\sigma)\Gamma(y+1)} \left(\frac{\sigma\mu}{1+\sigma\mu}\right)^y \left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma} \quad (3).$$

Where the mean and variance of Y can be written into μ and $(1+\delta)\mu$. In order to model the effect of K covariates $x_1, x_2, \dots, x_k, \dots, x_K$ to the mutation rate, we assumed that

$$\begin{aligned} g_1(\mu) &= \log(\mu) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + \dots + \beta_K x_K \\ g_2(\sigma) &= \log(\sigma) = \alpha_0 \end{aligned} \quad (4).$$

Therefore, we run the negative binomial regression using (4) on the training data set. Specifically, suppose there are $i=1, 2, \dots, n$ different fixed length bins selected in section 2.2. For each bin we selected in section 2.2, mutation counts of all genomes of each cancer type is represented as a vector $\mathbf{Y} = \text{vector}(y_1, \dots, y_i, \dots, y_n)$, and matrix \mathbf{X} represents the $n \times 381$ covariate matrix generated in section 2.2. The goal of training is to accurately estimate $(\mu_1, \dots, \mu_i, \dots, \mu_n)$ for each bin and δ as the local mutation rate parameters. For a target region to be tested, we first find its nearest bin μ_{target} and used μ_{target} and δ for P value evaluation. In most cases, the length of the target region is far smaller than the bin length used in the training process, so an offset of region length can be added to represent such exposure effect.

We also used two other simple models to calculate P values as comparisons to NIMBus results. First, in the global Poisson model we assumed that a constant mutation rate across the whole genome and then used Poisson distribution to evaluate somatic burdens in target regions. Also, we assume that mutation rate varies in different regions using our estimated $(\mu_1, \dots, \mu_i, \dots, \mu_n)$. But we treat the mutation rate from different patients as constant to derive the P values. This model is called the local Poisson model.

2.5 Training bin size selection

There are pros and cons to select certain a bin size for model training. On one hand, a shorter bin size will be advantageous in the P value evaluation as it can more effectively remove the heterogeneity in the local mutation process at a higher resolution. On the other hand, it sometimes will result in worse regression performance when estimating μ and δ . One reason is that in order to effectively estimate the mutation rate change with covariates, we need to obtain sensible mutation rate estimation in each single bin. However on the 3 billion bases genome, somatic mutation count data is usually sparse due to limited number of disease genomes available at the moment. In the extreme case, when the bin size is small enough so that most bins have mutations, it is difficult for the regression model to capture the relationship between mutations and covariates. Another reason is that some of the covariates are only reported to be functional in a large scale, so reducing the bin size will not necessarily introduce better prediction precision. The best

NO?

SAMPLE?

anion

MATH?

no?

bin size selection for the training purpose is still a challenging question that needs further case-by-case investigation. In our analysis, we used a 1mb bin size for all cancer types.

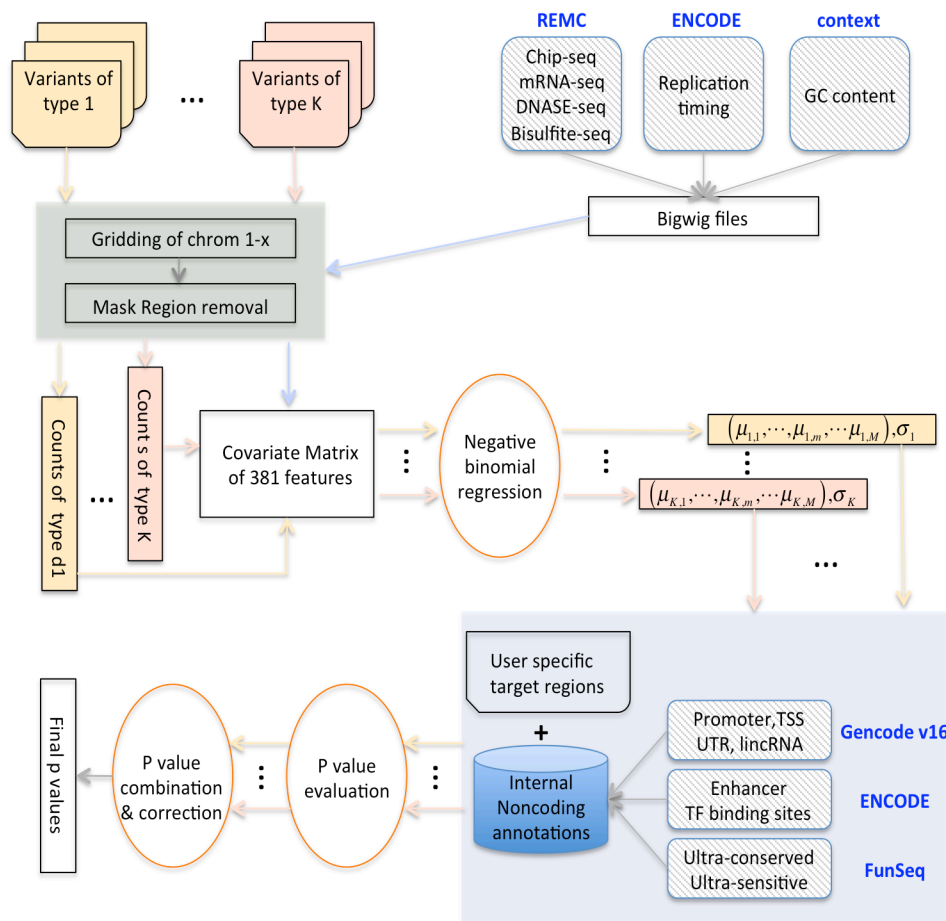
2.6 Combining P values

Sometimes several related diseases, or diseases of different subtypes needs to be analyzed together to provide a combined P value. Since regression is run separately on each single disease/subtype of disease, either Fisher's method can be used to combine the P values together.

2.7 Flowchart of NIMBus

To better illustrate how NIMBus works, its workflow is given in Figure 1

Figure 1. Flowchart of NIMBus

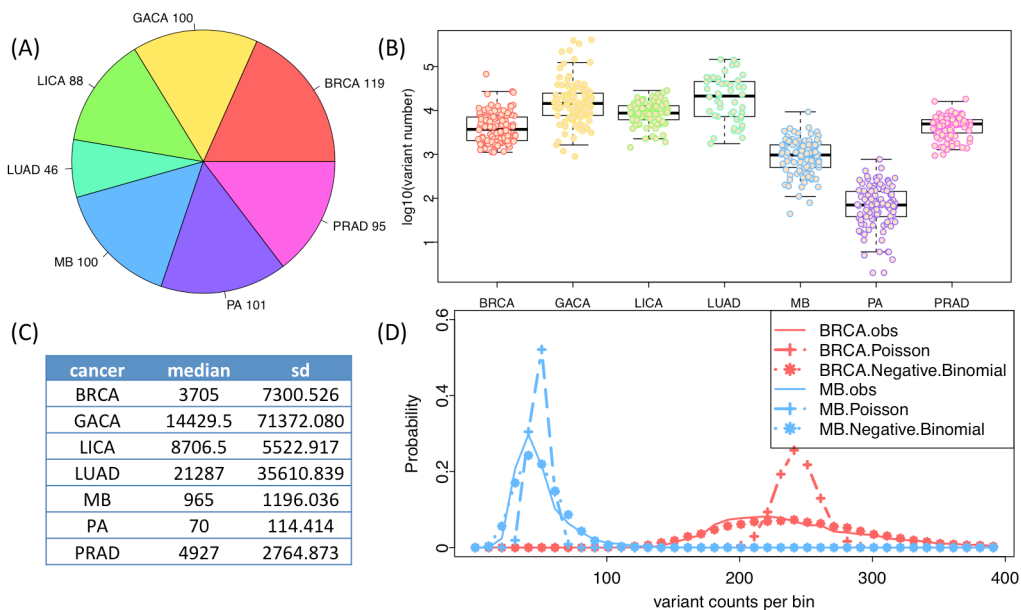


3. Results

3.1. Heterogeneity from various sources leads to large overdispersion in mutation counts data

Pioneer genome wide somatic burden analysis usually assumes a homogeneous mutation rate, which consequently uses binomial or Poisson tests to evaluate P values. However, we found that mutation count data usually violates such an assumption because there are various sources of heterogeneity in the mutation rate. To demonstrate this, we collected WGS variants from 649 cancer patients and 7 cancer types as shown in Fig. 2A. First, we found that mutation count per genome varies from disease to disease. For instance, the median number of variants can be as low as 70 in PA and as high as 21287 in LUAD

Figure 2. Mutation rate heterogeneity from various sources result in huge overdispersion in mutation count data. (A) Sample numbers for each cancer type in our analysis; (B) boxplots and scatter plots of mutation counts per sample; (C) summary of mutation counts; (D) fitting of Poisson and Negative binomial distribution

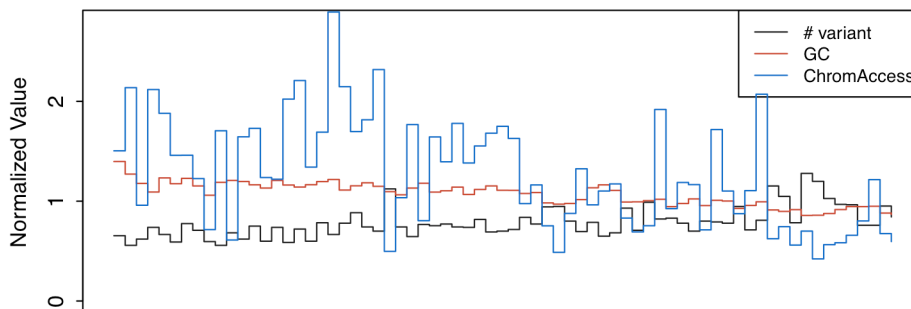


(Fig. 2B-C). Even within the same disease, mutation rate may dramatically change from sample to sample (lowest at 1743 and highest at 145500 in LUAD). In addition, there are also large regional mutation rate differences within the same sample (Fig. 3). Therefore, a binomial or Poisson distribution usually provides poor fitting to the real mutation counts data (Fig. 2D, dash lines with +). In light of these, we utilized a two parameter negative binomial distribution to further capture the overdispersed nature of mutation counts data, which improves fitting to real data significantly (dash lines with star in Fig. 2D).

3.2 Local mutation rate is confounded by other genomic features

Instead of ~~homogeneously located across the genome in a random way~~, somatic mutation rate has been reported in literature to be confounded by several genomic features. For example, single stranded DNA during replication usually suffers from endogenous DNA damage, such as oxidation and deamination. Therefore, the accumulative damage effect in the later replicated regions will result in elevated mutation rate. Another well-known factor is the expression levels. The highly expressed regions often demonstrate lower mutation rate as compared to lowly expressed regions due to the transcription-coupled repair mechanism $\{\text{cite Mutsig and refer}\}$. Furthermore, it has been reported that the chromatin organization, which arranges the genome into heterochromatin- and euchromatin-like domains, has a dominant influence on regional mutation-rate variation in human somatic cells $\{\text{cite 22820252}\}$. We also find a similar trend in our data analysis. For instance, the normalized mutation rate of the first 100 1mb bins extracted in 95 breast cancer samples was plotted in the black line in Fig. 3. It is notably correlated with GC content (red line) and chromatin accessibility (blue line) in matched breast tissue.

Figure 3. Local mutation rate is correlated with many genomic features. The black line is the normalized mutation rate in 95 breast cancer sample. Normalized GC content (red line) and Chromatin accessibility (blue line) were significantly correlated with mutation rate.

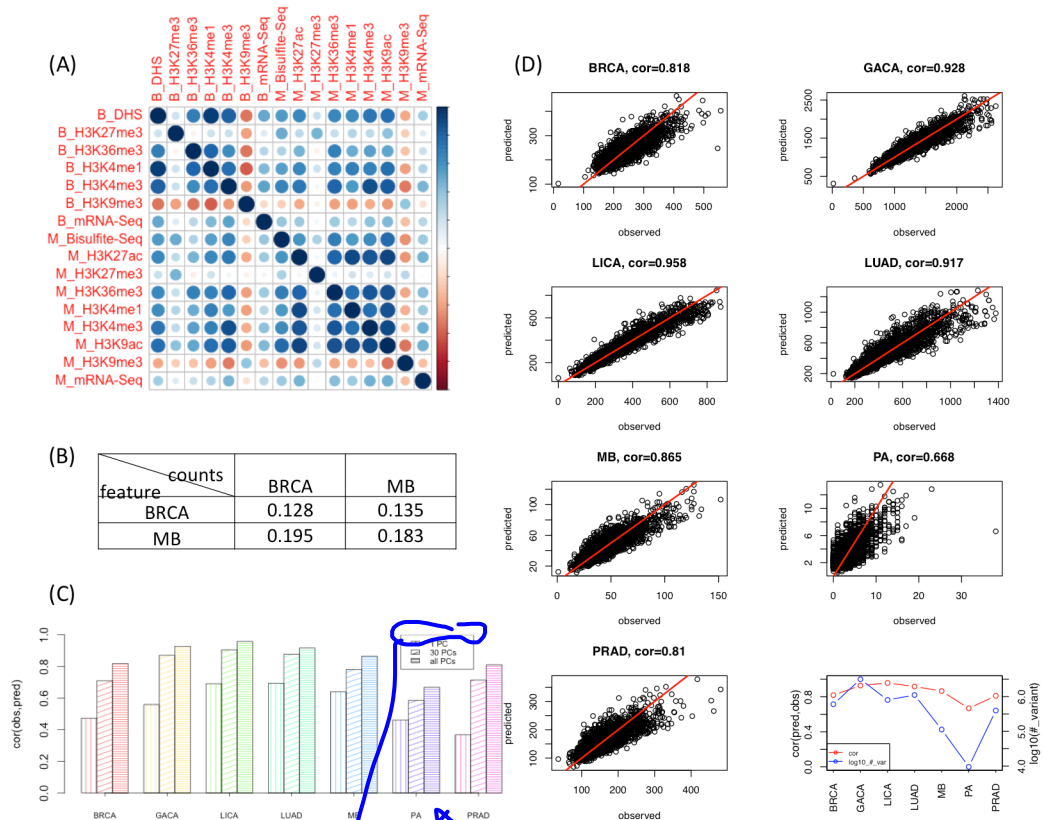


3.3. Negative binomial regression precisely estimates local mutation rates by correcting a list of genomic features

In light of the high correlation of individual genomic features with mutation rate in all cancer types, we try to investigate the joint collaborative effect of multiple features to predict mutation rate. It has been reported that most accurate prediction can be achieved by using the matched tissue. Hence, we specifically selected variants in two distinct cancer types BRCA and MB and predict their local mutation rate by features from different tissues. Relative error of each 1mb bin is defined by the absolute value of observed and predicted difference divided by the observed value and its mean value over all bins is used to compare the regression performance. As expected, regression using matched tissue features provides the lowest relative error. For example, using MB related features to predict variant counts in MB has a mean relative error of 0.183, less than that using BRCA features for MB (Fig. 4B, P value of two sided Wilcoxon test $< 2.2e^{-16}$).

However, non-matched tissue regression still provides insightful information during regression. For example, even when MB features are used to predict BRCA mutation rates, a decent although slightly poorer performance can still be achieved (0.135 for unmatched vs. 0.128 for matched). One explanation is that these genomic features are highly correlated both within and across different tissues. A correlation plot is given in Fig. 4A. Size and darkness of color of each dot inside the matrix represent the absolute value of Pearson correlation, while blue and red represents positive and negative values respectively. Usually histone modification marks within the same tissue provides the highest correlation (e.g H3k4me1, H3k4me3, H36k4me3). However, these features in MB are also significantly associated with those in BRCA. That is why these signals can be imputed using interpolation through data in other tissues [\cite{25690853}](#). This conclusion is important for two reasons: (1) tissue matching is usually extremely difficult for complex disease such as cancer. For example, there are several sets of covariates for breast tissues in REMC, which come from different parts of breast tissue. However, the exact tumor location data is sometimes vaguely described or even missing. (2) Even if the perfectly matched tissue can be identified in some cases, the data matrix from REMC is far from complete. In our example, methylation data for breast tissue of this particular

Figure 4. (A) Genomic features are highly correlated. B_* represents breast cancer related features and M_* indicates Medulloblastoma related features. (B) relative error of average mutation rate in each 1mb bin by different regressions. (C) Regression performance against number of PCs used in the regression. (D) performance of regression for all cancer types.



type is missing (Fig. 4A). Another example is that there is no prostate tissue in REMC, so there is no matched for prostate cancer. We still need to predict the local mutation rate in prostate cancer so correlated features in other related tissues would still provide additional information to boost regression performance.

To overcome the multi-collinearity problem due to the correlated nature of genomic features, we first performed a principle component analysis (PCA) on the 381 features. The first principle component (PC) could explain up to 55.69% of variance in the covariate matrix. It takes at least 15 and 106 PCs to capture 90% and 99% of variance respectively. However, PCA based regression is very sensitive to the PCs selected because sometimes the PCs that can best explain the variance in the mutation count data is not necessarily those that can best explain the variance in the covariate matrix. Hence, for each cancer type, we first ranked the individual PCs, and then only select the top 1, 30, and all PCs to predict the local mutation rate. Fig. 4C shows that in all 7 cancer types, using more PCs can noticeably boost prediction performance. For example, in BRCA the Pearson correlation is only 0.472 if 1 PC is used in regression. However, correlation coefficient can rise to 0.655 and 0.709 if 15 and 30 PCs are used respectively, and it can eventually be increased to 0.818 after using all PCs. As a result, in all the following analysis, we used all PCs for accurate local mutation rate estimation.

Our PCA based negative binomial regression can precisely capture the effect of various genomic features on somatic mutation process and accurately predict the local mutation rate in all cancer types. The Pearson correlations of the observed data and the predicted value vary from 0.668 in PA to 0.958 in LICA. Scatter plots are given in Fig. 4D. It is worth mentioning that although there is no prostate tissue in REMC data, we can still achieve a very decent correlation of 0.81 with the help of 381 unmatched but correlated features. It indicates that when somatic mutation data of an unknown disease is given, our model could still achieve acceptable performance without the knowledge of related tissue information. In addition, our regression performance is affected but not limited to the total number of variants (bottom right figure in Fig. 4D). Obviously limited number of mutations could restrict the prediction precision, such as in very quiet somatic genomes like PA (lowest correlation at 0.668 among 7 cancer types). Other factors, such as number of effective covariates, quality of mutation calls, and molecular similarity of pooled samples of the same disease could also considerably influence the performance of our model. For instance, although there are a smaller number of variants in MB than in BRCA, our regression in MB still outperforms that in BRCA (0.865 VS. 0.818, Fig. 4D).

3.4. NIMBus discovered a list of highly recurrent noncoding regions from cancer WGS data

We applied NIMBus on WGS variant calls on all 7 cancer types to deduce the individual somatic burden P values, and compared with the results from global and local Poisson models (details in section 2.4). As expected, both global and local Poisson models generated obviously too many burdened regions in all noncoding annotation categories because of the poor fitting of Poisson distributions to the mutation count data (Fig. 2D). For example, in the promoter regions, after P value correction, NIMBus provided 8 promoters as highly mutated, while local and global Poisson models identified 47 and

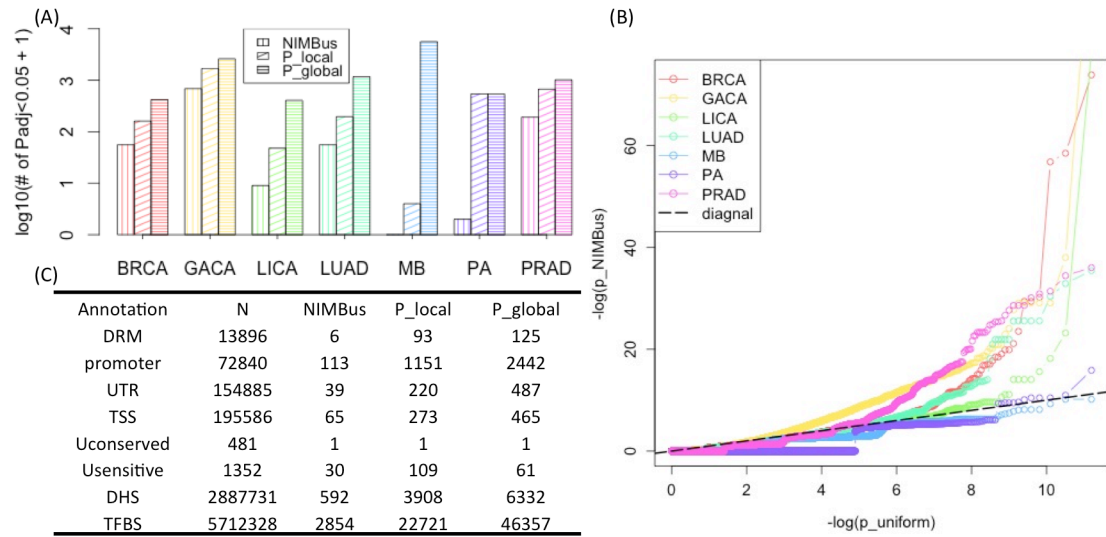
406 respectively. It is very unlikely that in a single tissue, all these 47 or 406 promoters are all linked with tumor progression. Hence, our negative binomial assumption in NIMBus effectively captured the overdispersion and controlled the number of false positives. To further demonstrate this, we provided the Q-Q plots of P values from all 7 cancer types were provided in Fig. 5B as quality check. In theory, if no significantly burdened regions are detected, the P values should follow uniform distribution. As it is seen in Fig. 5B, in all cancer types the majority of our P values for all cancer types follows the uniform assumption with a few outliers as the true signals, indicating reasonable P value distributions. Similar results have also been seen in other noncoding annotations (data not shown).

To summary the mutation burdens from all cancer types, we used Fisher's method to calculate the final P values for all three models. Similar to P values from a single cancer type, the combined P values are severely inflated in both global and local Poisson models, but are rigorously controlled by NIMBus (table C in Fig. 5). Take the TSS as an example, NIMBus reported only 65 sites as burdened, as compared with 273 and 465 for the other two methods. Additional, out of the 65 TSS elements, several of them have been experimentally validated or computationally predicted as associated with cancer in other work. For instance, TP53 is a well-studied oncogene that is related in many cancer types, and combined P value for TP53 TSS is ranked second in our analysis ($P=4.26e-14$). LMO3 interacts with the tumor suppressor TP53 and regulates its function, and it is ranked fourth in our analysis ($P=3.25e-13$). We also found that the fifth ranked gene RMRP ($p=1.36e-10$), which is the RNA component of mitochondrial RNA processing endoribonuclease, has been claimed to be associated with colorectal and breast cancers [26415221]. Another important example is the TSS sites in TERT, which is ranked sixth in our results ($p=1.55e-10$) and has been experimentally validated as associated with multiple types of cancer progression. The discovery of such validated results proved that NIMBus can serve as an powerful tool for driver events discovery in diseases.

4. Discussion

→ P
LOG
?
WHAT
ABOUT
UNK
SUBTYPE
?

Fig. 5 (A) number of detected promoter regions in all cancer types; (B) Q-Q plots of P values for promoter regions; (c) total number of over burdened regions in our noncoding annotations after merging P values from 7 cancer types. P_local: local Poisson Model, P_global: global Poisson Model



Thousands of somatic genomes are now available due to the fast development of whole genome sequencing technologies, providing us with increasing statistical power to scrutinize the somatic mutation landscape. At the same time, thanks to the collaborative effort of big consortiums, such as REMC and ENCODE, tens of thousands of functional characteristic data on human genomes has been released for immediate use to the whole community. Hence, integrative frameworks are of urgent need to explore the interplay between WGS data and the functional characteristic data. It will not only be important to accurately search for mutational hotspots as driver candidates for complex diseases but also to better interpret the underlying biological mechanism for clinicians and biologists.

In this paper, we proposed a new integrative framework called NIMBus that uses a negative binomial regression to capture the effect of a widespread list of genomic features on mutation processes for accurate somatic burden analysis. Due to the heterogeneous nature of various somatic genomes, our model treated the mutation rates as a scaled gamma distribution to mimic the varying mutation baseline for different patients or disease subtypes. Resultantly, it modeled the mutation counts data using a two parameter negative binomial distribution, which improved the mutation counts fitting dramatically as compared to previous work (Fig. 2D).

Unlike previous efforts which use very limited covariates to estimate local mutation rate in very qualitative way, we explored the whole REMC and ENCODE data and searched for 381 features that best describe chromatin organization, expression profiling, replication status, and context effect in all possible tissues to jointly predict the local mutation rate at high precision. In terms of covariate correction, NIMBus demonstrates three obvious advantages: 1) it incorporates the most comprehensive covariates that provides the most accurate, at least to our knowledge, background mutation rate

estimation; 2) it provides an integrative framework that can be extended to any number of covariates and successfully avoids the high dimensionality problem as in other methods $\{cite mutsigCV\}$. This is extremely important since the amount of available functional characteristic data is growing rapidly as the time and money cost of sequencing technologies drops; 3) it automatically utilizes the genomic regions with highest credibility for training purposes so potential users are not bothered to perform carefully calibrated training data selection and complex covariate matching processes.

how
ARC
CIV
WASP
?

In addition, we also put a lot of effort on NIMBus to explore the most extensive noncoding annotations. Noncoding regions represent more than 98% of the whole human genome, and are less investigated mainly due to limited knowledge to understand its biological functions. NIMBus collects the up to date full catalog of noncoding annotation of all possible tissue from the ENCODE project and our previous efforts from population genetics efforts in 1000 Genomes Project. All these included internal annotations of NIMBus can be either tested for somatic mutation burden or used to annotate the user specific input regions.

We applied NIMBus on 649 cancer genomes of 7 different types collected from public data and collaborators. The individual burden test P values for each cancer type have been deduced and then Fisher's method has been used to calculate the combined P values. As a result, a list of non-coding elements has been reported to have more than expected mutations (Table C in Fig. 5D). A list of already well-known regions, such as TP53, LMO, and TERT TSS, has also been reported in our analysis to be hypomutated, proving the effectiveness of NIMBus to identify functionally associated results.

It is worth mentioning that although we demonstrate the effectiveness of NIMBus mostly on somatic mutation analysis, it can be immediately extended to germline variant analysis as well. In summary, NIMBus is the first method that can integrate thousands of functional characteristic experimental data to analyze the mutation burdens in disease genomes. Such external data does not only help to better estimate the background mutation rate for successful false positive and negative control, but also provide the most extensive noncoding annotations for users to interpret their results. It may serve as a powerful computation tool to accurately predict driver events in human genetic disease and potentially identify biological targets for drug discovery.

ASSISTE

Funding

Reference