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

Jing Zhang[1,2], Jason Liu[2,3], Lucas Lochovsky[1], Jayanth Krishnan[1], Donghoon Lee[1], Mark Gerstein[1,2,4*]

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

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

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

[4]Department 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

Identifying highly mutated regions is a key way that scientists can use on population scale sequencing to discover key genomic regions associated with complex diseases such as cancer. Nevertheless, it is challenging to identify such regions because severe mutation rate heterogeneity, across different genome regions of the same individual and also across different individuals, gives rise to highly over-dispersed counts of mutations. Moreover, it is known that part of this heterogeneity relates to confounding genomic features, such as replication timing and chromatin organization. Here, we address these issues with a Negative binomial regression based Integrative Method for mutation Burden analysis (NIMBus). This approach uses a Gamma-Poisson mixture model to capture the mutation rate heterogeneity across different individuals and thus models the over dispersed mutation counts by a negative binomial distribution. Furthermore, it regresses the mutation counts against 381 features extracted from REMC and ENCODE to accurately estimate the local background mutation rate. This framework can be readily extended to accommodate additional genomic features in the future. NIMBus was used to analyze 649 whole-genome cancer sequences. It successfully controlled P value inflation and identified well-known coding and noncoding drivers, such as TP53 and the TERT promoter. We make NIMBus available and release our results as an online resource (nimbus.gersteinlab.org).

1. Introduction

Population level analysis, which looks for regions mutated more frequently than expected, is still one of the most powerful ways to identify deleterious mutations for diseases [1-3]. The recent development of whole genome sequencing (WGS) and personal genomics has provided us unprecedented statistical power to perform such analyses. Therefore, an accurate quantification of mutation burden is important to uncover the genetic cause of various diseases, which in turn allows for targeted therapies in clinical studies. However, mutation burden test for somatic variants remains challenging for several reasons.

First, some of the pioneer work analyzing WGS assumed a constant mutation rate across different regions or cancer genomes and ignored that somatic genomes are highly heterogeneous [4]. Hence, the positional level mutation counts often demonstrate larger than expected variance under such assumption, which is called overdispersion. This assumption results in poor data fitting and generates numerous false positives [5], so it is necessary to introduce more sophisticated models to handle mutation rate heterogeneity.

Second, numerous genomic features have been reported to largely affect the mutation process [6], which need to be corrected carefully in burden analysis. Unfortunately, none of the few current methods that considered such effects systematically explored these genomic features in a tissue-specific way, and their models demonstrated very limited extensibility to accommodate new features in the future. For instance, MutSigCV tried to correct the effects of several features, such as expression and replication timing, by only using a small neighborhood of genes with similar covariate values. However, as the covariate number increases, it is usually difficult to find a meaningful neighborhood in a high dimension space.

Lastly, many *state-of-the-art* methods are optimally designed for coding regions analysis [6], which represents less than 2 percent of the human genome. Nowadays, a myriad of studies have shown that noncoding mutations can serve as driver events for diseases. One well-known example is that mutations in the TERT promoter were found to be associated with cancer progression [7]. Hence, a unified coding and noncoding analysis is needed to annotate the discovered hotspots.

In this article, we propose a Negative binomial regression based Integrative Method for mutation Burden analysis (NIMBus) that solves the three problems mentioned above. It first intuitively treats mutation rates from different individuals as random variables with gamma distribution, and resultantly models the pooled mutation counts from a heterogeneous population as a negative binomial distribution to handle the overdispersion. Furthermore, to capture the covariate effects, it integrates the most extensive features in all available tissues from Roadmap Epigenomics Mapping Consortium (REMC) and the Encyclopedia of DNA Elements (ENCODE) project to create a covariate table to predict the local mutation rate with high precision through regression. In addition, it also customizes the most comprehensive noncoding annotations from ENCODE to facilitate results interpretation. This integrative approach employed in NIMBus enables us to effectively pinpoint mutation hotspots associated with disease progression and to better understand the biological mechanisms thereof.

2. Methods

2.1 WGS variants data used

We collected 649 whole genome variant calls from public resources and our collaborators. This data contains a broad spectrum of 7 different cancer types (details in Text S1 section 1).

2.2 Local background mutation rate estimation

(A) Human genome gridding and covariate matrix calculation

First we divided the whole genome into bins with fixed length l. In this stage, l is usually large, such as 1mb. Then the bins overlapped with any of the two blacklist regions were removed. Then 381 features were extracted from both REMC and ENCODE, and average signal in the bins was calculated (details in Text S1 Section S2). Let $x_{i,j}$ denote the average signal strength for the i^{th} bin and j^{th} covariate, where $i = 1, \cdots, n$ and $j = 1, \cdots, m$.

(B) Use negative binomial distribution to handle mutation count overdispersion

Suppose there are $d = 1, \cdots, D$ different diseases (or disease types) in the collected WGS data, and $s = 1, \cdots, s_d$ represents sample for disease (or disease type) d. Let $y_i^{d,s}$ and $\lambda_i^{d,s}$ denote the mutation count and rate for the i^{th} bin defined in section 2.2 (A) for sample s in disease d. In previous efforts, scientists assume that mutation rate $\lambda_i^{d,s}$ is constant across different regions of the human genome, samples, and diseases, so they have $\lambda_i^{d,s} \triangleq \lambda$ for $\forall\, i, d, s$. Hence $y_i^{d,s}$ follows a Poisson distribution with the probability mass function (PMF) given in equation (1).

$$p\left\{Y_i^{d,s} = y_i^{d,s}\right\} = \frac{e^{-\lambda_i^{d,s}}\left(\lambda_i^{d,s}\right)^{y_i^{d,s}}}{y_i^{d,s}!} \triangleq \frac{e^{-\lambda}\lambda^{y_i^{d,s}}}{y_i^{d,s}!} \qquad (1)$$

However, somatic genomes are highly heterogeneous because regions from various diseases, samples, and regions of the same genome usually demonstrate considerably different mutation rates, severely violating the assumption in equation (1). As a result, fitting of $y_i^{d,s}$ is usually very poor because a larger than expected variance, the so called overdispersion, is often observed [5]. Simply using the constant mutation rate assumption will generate numerous false positives. Instead in our model, we assume that different $\lambda_i^{d,s}$ are i.i.d random variables that follow Gamma distribution with probability density function (PDF) as

$$P\left\{\lambda_i^{d,s} = x\right\} = \frac{1}{\Gamma\left(c_i^d\right)\left(\upsilon_i^d\right)^{c_i^d}} x^{\left(c_i^d - 1\right)} e^{-\frac{x}{\upsilon_i^d}} \qquad (2),$$

where $c_i^d > 0$ and $\upsilon_i^d > 0$. In equation (2), c_i^d and υ_i^d are the shape and scale parameters respectively. Assume that $\lambda_i^d = \sum_{s=1}^{s_d} \lambda_i^{d,s}$ is the overall mutation rate from all samples in bin i of disease d. Its distribution can be readily obtained through convolution as

$$P\left\{\lambda_i^d = x\right\} = \frac{1}{\Gamma\left(s_d c_i^d\right)\left(\upsilon_i^d\right)^{s_d c_i^d}} x^{\left(s_d c_i^d - 1\right)} \exp\left(-\frac{x}{\upsilon_i^d}\right) \qquad (3).$$

Let $y_i^d = \sum_{s=1}^{s_d} y_i^{d,s}$ represent the mutation counts in region i of all the samples from disease d. The conditional distribution of y_i^d given λ_i^d can be written into

$$P\left(y_i^d \middle| \lambda_i^d\right) = \frac{\left(\lambda_i^d\right)^{y_i^d} \exp\left(-\lambda_i^d\right)}{\left(y_i^d\right)!} \qquad (4).$$

By integrating (3) into (4), the marginal distribution of y_i^d can be denoted as a negative binomial distribution ([8], page 50 in [9]).

$$P\left(y_i^d \middle| c_i^d, \upsilon_i^d\right) = \left(\frac{1}{1+\upsilon_i^d}\right)^{s_d c_i^d} \frac{\Gamma\left(s_d c_i^d + y_i^d\right)}{\Gamma\left(s_d c_i^d\right)\left(y_i^d\right)!} \left(\frac{\upsilon_i^d}{1+\upsilon_i^d}\right)^{y_i^d} \qquad (5a).$$

Equation (5) is the PDF of a negative binomial distribution with $E\left(y_i^d\right) = s_d c_i^d \upsilon_i^d$ and $Var\left(y_i^d\right) = s_d c_i^d \upsilon_i^d \left(1 + \upsilon_i^d\right)$. To better interpret (5a), we define $\upsilon_i^d = \mu_i^d \sigma_i^d$ and $s_d c_i^d = 1/\sigma_i^d$. Then equation (5a) can be rewritten into (5b).

$$p_{Y_i^d}\left(y_i^d \middle| \mu_i^d,\sigma_i^d\right)=\left(\frac{1}{1+\sigma_i^d\mu_i^d}\right)^{1/\sigma_i^d}\frac{\Gamma\left(y_i^d+1/\sigma_i^d\right)}{\Gamma\left(1/\sigma_i^d\right)\Gamma\left(y_i^d+1\right)}\left(\frac{\sigma_i^d\mu_i^d}{1+\sigma_i^d\mu_i^d}\right)^{y_i^d} \tag{5b}$$

The mean and variance of y_i^d from (5b) can be described as μ_i^d and $\mu_i^d\left(1+\mu_i^d\sigma_i^d\right)$ respectively. Our model in equation (5b) is convenient with explicit interpretability. First, it assumes that the individual mutation rate is heterogeneous by modeling $\lambda_i^{d,s}$ as i.i.d. Gamma distributed random variable. Hence, different from the constant mutation rate assumption where $E\left(y_i^d\right)=Var\left(y_i^d\right)$, it captures the extra variance of y_i^d due to population heterogeneity. Second, our model in (5b) clearly separates the two main parameters μ_i^d and σ_i^d with physically interpretable meanings: the mean and overdispersion. Here a larger σ_i^d indicates a more severe degree of overdispersion, which is usually due to larger difference in mutation rates.

(C) Accurate local background mutation rate estimation by regression

After modeling y_i^d using negative binomial distribution in 2.2 (B), we then tried to estimate the local mutation rate by correcting the covariate table ***X*** described in 2.2 (A). Again $x_{i,j}$ denote the average signal strength in the i^{th} bin and j^{th} covariate, where $i=1,\cdots,n$ and $j=1,\cdots,m$. We noticed that the genomic features in the covariate tables are highly correlated, which may introduce multicollinearity if directly used in regression. We first applied principle component analysis (PCA) to matrix $\boldsymbol{X}$. Let $\boldsymbol{X}'$ represent the covariate matrix after PCA and $x'_{i,j}$ denote each element in $\boldsymbol{X}'$.

A generalized regression scheme is used here. Suppose g_1 and g_2 are two link functions. We then use linear combinations of covariate matrix $\boldsymbol{X}'$ to predict the transformed mean parameter μ_i^d and overdispersion parameter σ_i^d as

$$\begin{aligned} g_1\left(\mu_i^d\right)&=\log\left(\mu_i^d\right)=\beta_0^d+\beta_1^d x'_{i,1}+\cdots+\beta_j^d x'_{i,j}+\cdots+\beta_j^d x'_{i,m} \\ g_2\left(\sigma_i^d\right)&=\log\left(\sigma_i^d\right)=\alpha_0^d+\alpha_1^d x'_{i,1}+\cdots+\alpha_j^d x'_{i,j}+\cdots+\alpha_m^d x'_{i,m} \end{aligned} \quad (6).$$

Here we used the log function for both g_1 and g_2, so the regression model in (6) is also called a negative binomial regression. Note that ***X*** contains 381 genomic features in all available tissues. In the following analysis, we used all features to run the regression in (6) to achieve better performance. We used the GAMLSS package in R to estimate the parameters in (6) as $\hat{\alpha}_0^d,\cdots,\hat{\alpha}_m^d,\hat{\beta}_0^d,\cdots,\hat{\beta}_m^d$. There are usually biological reasons to explain how μ_i^d changes with covariates. For example, single-stranded DNA in the later replicated regions usually suffers from accumulative damage to have larger μ_i^d. But it is difficult to interpret such relationship on σ_i^d. Hence, we can simplify equation (6) by assuming σ_i^d is constant in our real data analysis.

2.3 Somatic burden tests using local background mutation rate

(A) Background mutation rate calculation for target regions

Suppose there are K regions to be tested. We used the local mutation rate to evaluate the mutation burden. For the k^{th} target region ($k=1,\cdots,K$), optimally we should extend it into length l (illustrative figure given in Fig. S2). Then we calculate the average signal for feature j as $x_{k,j}, j=1,\cdots m$ for this extended bin, and after PCA projection let $x'_{k,j}$

represent the value for the j^{th} PC. Then the local mutation parameters $\hat{\mu}_k^d$ and $\hat{\sigma}_k^d$ in the extended bin for the k^{th} target region can be calculated as

$$\begin{aligned}\hat{\mu}_k^d &= \exp\left(\hat{\beta}_0^d + \hat{\beta}_1^d x'_{k,1} + \cdots + \hat{\beta}_j^d x'_{k,j} + \cdots + \hat{\beta}_m^d x'_{k,m}\right)\\ \hat{\sigma}_k^d &= \exp\left(\hat{\alpha}_0^d + \hat{\alpha}_1^d x'_{k,1} + \cdots + \hat{\alpha}_j^d x'_{k,j} + \cdots + \hat{\alpha}_m^d x'_{k,m}\right)\end{aligned} \quad (7).$$

In reality, the length of the k^{th} test region l_k is much shorter than the length of the training bins (up to 1Mb). Hence $\hat{\mu}_k^d$ need to be adjusted by a factor of l_k/l. Then $\hat{\sigma}_k^d$ and the adjusted $\hat{\mu}_k^d$ can be used to calculate the disease specific P value p_k^d. This optimal scheme is usually computationally expensive because there are usually millions of target regions to be tested. Therefore, we proposed an approximation method to replace the optimal $\hat{\mu}_k^d$ and $\hat{\sigma}_k^d$ in our analysis (details see section S4 in Text S1).

(B) Combining P values for multiple disease types

Sometimes it is necessary to analyze several related diseases (or disease types) to provide a combined P value. One typical example is the pan-cancer analysis. In section 2.3 (A), we calculated the P value for disease/disease type d as p_k^d for test region k, Then Fisher's method can be used to combine P values. Specifically, the test statistic can be calculated in

$$T_k = -2\sum_{d=1}^{D} \ln\left(p_k^d\right) \sim \chi^2\left(2D\right) \quad (8).$$

Here T_k follows a centered chi-square distribution with $2D$ degrees of freedom, where D is the total number of diseases/disease types. Then the final P value p_k can be calculated accordingly.

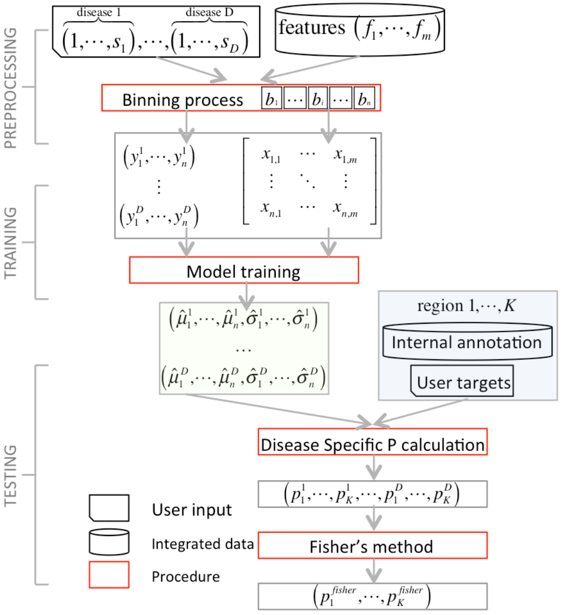
2.4 Noncoding annotations customized for NIMBus

We customized the full list of noncoding annotations from both ENCODE annotations and our previous efforts in the 1000 Genomes Project to make it suitable for burden analysis. More details are given in Text S1.

2.5 Flowchart of NIMBus

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

Figure 1. Flowchart of NIMBus


disease 1
disease D
$(1,\cdots,s_1),\cdots,(1,\cdots,s_D)$
features $(f_1,\cdots,f_m)$
PREPROCESSING
Binning process
b_1 ⋯ b_i ⋯ b_n
$(y_1^1,\cdots,y_n^1)$
$(y_1^D,\cdots,y_n^D)$
$x_{1,1}$ ⋯ $x_{1,m}$
$x_{n,1}$ ⋯ $x_{n,m}$
TRAINING
Model training
$(\hat{\mu}_1^1,\cdots,\hat{\mu}_n^1,\hat{\sigma}_1^1,\cdots,\hat{\sigma}_n^1)$
$(\hat{\mu}_1^D,\cdots,\hat{\mu}_n^D,\hat{\sigma}_1^D,\cdots,\hat{\sigma}_n^D)$
region $1,\cdots,K$
Internal annotation
User targets
TESTING
Disease Specific P calculation
$(p_1^1,\cdots,p_K^1,\cdots,p_1^D,\cdots,p_K^D)$
Fisher's method
$(p_1^{fisher},\cdots,p_K^{fisher})$
User input
Integrated data
Procedure

3. Results

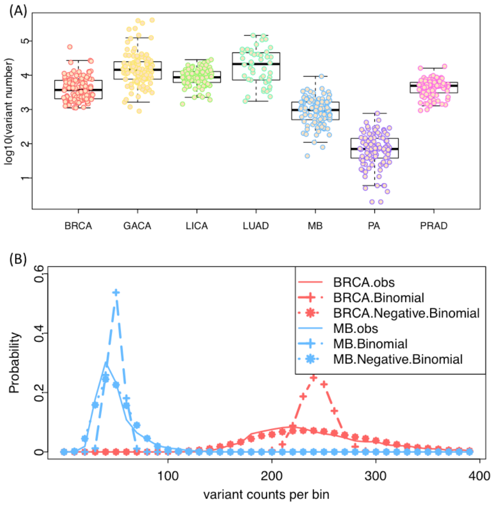
3.1. Heterogeneity from various sources leads to large overdispersion in mutation counts data

Pioneer genome wide somatic burden analysis usually assumes a homogeneous mutation rate per nucleotide, which consequently uses binomial tests to calculate P values [4]. However, we found that mutation count data usually violates this assumption because there is severe mutation rate heterogeneity from various sources. To demonstrate this, we collected

WGS variants from 649 cancer patients and 7 cancer types (Fig. S1).

First, we found that the mutation count per genome varies across diseases and samples. For instance, the median number of variants can be as low as 70 in Pilocytic Astrocytoma (PA) and as high as 21287 in Lung adenocarcinoma (LUAD). Even within the same cancer type, mutation counts vary dramatically from sample to sample (lowest at 1743 and highest at 145500 in LUAD, Fig. 2A). In addition, there are also large regional mutation rate differences within the same sample (Fig. S4). Therefore, distributions based on constant mutation rate assumption usually fit poorly to the real mutation counts data (Fig. 2B, dashed lines with +, Fig. S3 in Text S1). In light of these, we utilized a two-parameter negative binomial distribution to further capture the over-dispersed nature of mutation counts data, which improves fitting to real data significantly (dashed lines with star in Fig. 2B).

Figure 2. (A) Disease and sample mutation rate heterogeneity; (B) improved fitting by negative binomial distribution of mutation counts in 1mb bins in breast cancer (BRCA) and Medulloblastoma (MB)



3.2 Local mutation rate is confounded by many genomic features

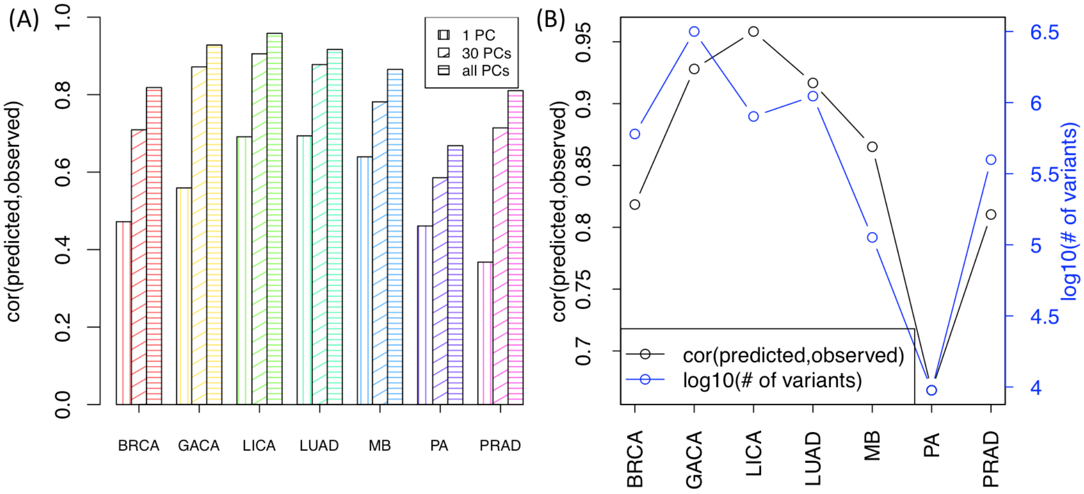
Somatic mutation rate has been reported to be confounded by several genomic features [6, 10]. For example, single-stranded DNA during replication usually suffers from endogenous DNA damage, such as oxidation and deamination. Therefore, the accumulative damage effect in the later replicated regions will result in increased mutation rate. We have observed a similar trend in our data. For example, the Pearson correlation between normalized mutation counts and replication timing values in BRCA is as high as 0.67 in the first 70 1mb bins (Fig. S4A). Another example is 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 [10]. Consistently, we also find that mutation counts are significantly associated with the DNase-seq signal (Pearson correlation=-0.61, P=1.52×10^{-8}, Fig S4B in Text S1). In light of these, it is important to estimate local background mutation rate for accurate mutation burden analysis.

3.3. Negative binomial regression precisely estimates local mutation rates by correcting many genomic features

Features in matched tissues usually provide best prediction accuracy but features in unmatched tissue still help

It has been reported that the most accurate local mutation rate prediction can be achieved by using features from matched tissue [11]. Hence, we specifically selected variants in two distinct cancer types BRCA and MB and predicted their local mutation rates by a few features from matched (or loosely matched) and unmatched tissues (Table S2 in Text S1). Relative error, defined by the normalized difference of observed and predicted value

Figure 3. (A) Regression performance by correcting different number of PCs; (B) Regression performance vs. total number of variants used in all cancer types



(equation s3 in Text S1), was used to compare model performance. Consistent with previous conclusions, we find that features in matched tissues usually outperform those from unmatched tissues. For example, the relative error is only 0.128 by using breast related features to predict BRCA mutation rates, noticeably smaller than 0.195 by using brain related features (Table S3 in Text S1). Similarly, brain related features are more accurate than breast related ones in predicting mutation rates in MB (0.135 VS. 0.183).

However, biologically meaningful tissue matching remains challenging and usually it is not an obvious choice for researchers without enough domain knowledge. Specifically, if samples of distinct hidden subtypes were pooled together for a certain disease, tissue matching will be more difficult. Furthermore, even after the best-matched tissue has been identified, we frequently need to handle missing features in that tissue. We noticed that many genomic features are highly correlated both within and across tissues (correlation plot in Fig. S6A), which leads to suboptimal but still decent regression performance (scatter plots given in Fig. S6B). This is extremely helpful when processing WGS from diseases without matched features. For example, there are no prostate related tissues in REMC, but features in other tissues still help to estimate the local mutation rates.

Pooling features from multiple tissues significantly improves local background mutation rate prediction

In light of the correlated nature of covariates, especially those epigenetic features [12], we first performed principle component analysis (PCA) on the covariate matrix to overcome the multicollinearly problem during regression. The correlation of each PC with the mutation counts data varies significantly across different cancer types (boxplots in Fig. S7B in Text S1). For example, the first PC demonstrates a Pearson correlation as high as 0.653 in LICA, much higher than 0.352 in PRAD. Therefore, it is necessary to run the regression model separately for different cancer types.

Since numerous PCs have been shown to be associated with mutation rates, we tried to investigate the collaborative efforts from multiple PCs to jointly predict the local mutation rates. Particularly, for each cancer type, we first ranked the individual PCs by their correlations with mutation rates, and then selected the top 1, 30, and all PCs to estimate the local mutation rate. Fig. 3A shows that using more PCs can noticeably boost prediction accuracy in all cancer types. For example, in BRCA the Pearson correlation is only 0.472 if 1 PC is used in regression, while the correlation coefficient rises to 0.655 and 0.709 if 15 and 30 PCs are used respectively. And eventually it increases to 0.818 after using all 381 PCs. As a result, in all the following analyses we used all PCs for accurate local mutation rate estimation.

As it is shown in Fig. 3B, we achieved good prediction accuracy through regression against all PCs of the covariate matrix in all cancer types. The Pearson correlations of the observed mutation count and the predicted $\hat{\mu}_i^d$ vary from 0.668 in PA to 0.958 in LICA. Scatter plots are given in Fig. S8 in Text S1. It is worth mentioning that although there is no feature in prostate tissue in REMC, we can still achieve a very high correlation of 0.81 with the help of 381 unmatched but correlated features. This indicates that our model could still provide acceptable performance even when somatic WGS of a disease is given without optimally matched covariates.

In addition, the number of available variants obviously affects prediction performance, but it is not the only factor. As shown in Fig. 3B, limited number of variants, such as those in quiet somatic genomes in PA, can usually restrict our prediction precision (lowest correlation at 0.668 among 7 cancer types). However, other factors, such as number of effective covariates, quality of mutation calls, and molecular similarity of pooled samples of the same disease can also influence the prediction performance considerably. For instance, although there are fewer variants in MB than those in BRCA, our regression in MB still outperforms that in BRCA (0.865 vs 0.818, Fig. 3B).

3.4. Coding region calibration for NIMBus

Since coding regions have been investigated in more detail than the noncoding regions, we first applied NIMBus on coding regions. First, we extracted coding regions from the GENCODE annotation v19 and NIMBus was run on both real and simulated datasets (details in section S11 in Text S1). We found that in all cancer types analyzed, NIMBus effectively controlled P value inflation as compared with the method mentioned in [4]. For example, in LUAD the P values for real data follow nicely with the uniform P values except a few outliers as the true signals (black dots on the right sides in Fig. 4). After P value correction by Benjamini–Hochberg method, only 11 genes has been reported as highly mutation in LUAD, while none was discovered on the simulated data (orange dots

Figure 4. Q-Q plots of P values of real and simulated WGS data

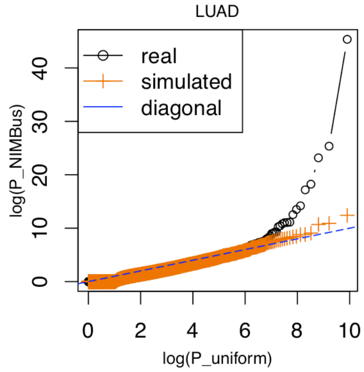


Table 1. Top genes after P value combination

Rank	Gene	Adjust P	PubMed ID
1	TP53	4.33E-139	17401424
2	DDX3X	3.65E-18	22820256
3	KRAS	2.56E-06	19847166
4	MUC4	4.47E-06	19935676
5	CDH1	3.07E-05	10973239
6	ARID1A	2.36 E-04	22037554
7	SMARCA4	3.78 E-04	18386774
8	FGFR1	7.43 E-04	23817572

in Fig. 4). On the other hand, the method with constant mutation rate assumption in [4] reported 6023 genes to be significantly mutated, indicating severe P value inflation.

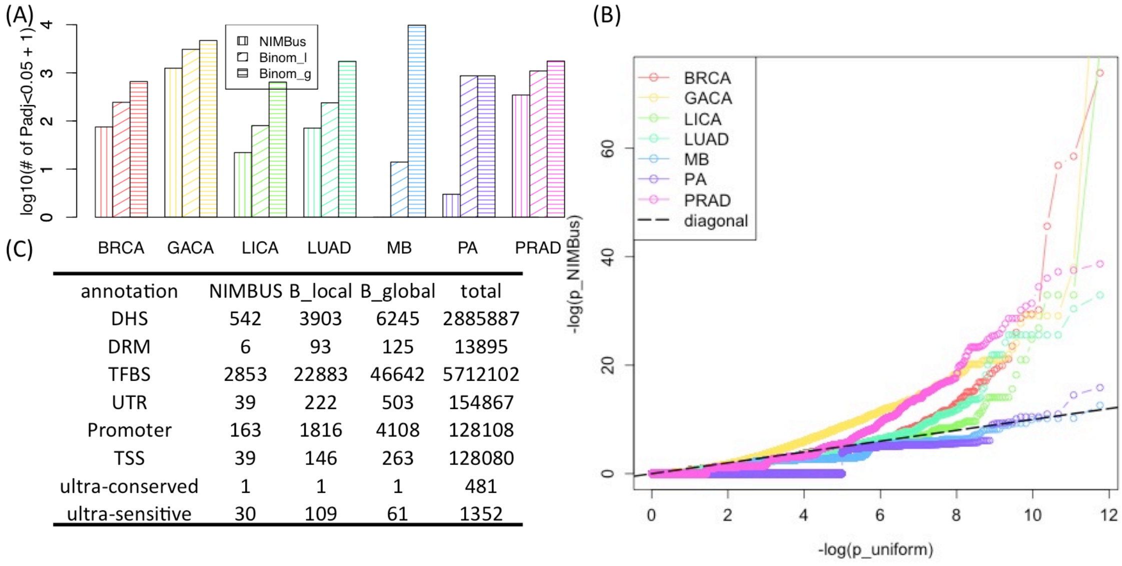
We also used Fisher's method to combine P values from all cancer types. In total, 15 genes were discovered as overly mutated. 12 out of them have been well documented as related with cancer progression. The top genes are shown in Table 1 and PubMed ID for related reference is given in the last column. These results showed that NIMBus is able to find sensible mutational hotspots as cancer drivers.

3.5. NIMBus discovered a list of highly mutated 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 with those from global and local Binomial models (details in Text S1).

Similar to the results in the coding region analysis, both global and local Binomial models generated too many burdened regions in all noncoding annotation categories, as evidenced by the poor fitting in Fig. 2B. For example, in liver cancer after P value correction, NIMBus identified 21 promoters as highly mutated, while local and global binomial models identified 79 and 641 respectively. 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 in promoter regions for all 7 cancer types in Fig. 5B as a quality check. In theory, if no significantly burdened regions are detected, the P values should follow uniform distribution. As shown in Fig. 5B, the majority of our P values follow the uniform assumption with a few outliers as the true signals, indicating reasonable P value distributions for all cancer types. Similar results were also observed in other noncoding annotations (data not shown). We released our burden test results on nimbus.gersteinlab.org as an online resource for the whole community.

Figure 5 (A) number of overly mutated promoter regions in all cancer types; (B) Q-Q plots of P values for promoter regions; (c) total number of burdened regions in all noncoding annotations after merging P values from 7 cancer types. B_local: local Binomial Model, B_global: global Binomial Model



annotation	NIMBUS	B_local	B_global	total
DHS	542	3903	6245	2885887
DRM	6	93	125	13895
TFBS	2853	22883	46642	5712102
UTR	39	222	503	154867
Promoter	163	1816	4108	128108
TSS	39	146	263	128080
ultra-conserved	1	1	1	481
ultra-sensitive	30	109	61	1352

To summarize 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 Binomial models, but are rigorously controlled by NIMBus (table C in Fig. 5). For example, NIMBus reported only 39 transcription start sites (TSS) as burdened, compared with 164 and 263 for the other two methods. Additionally, out of the 39 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 involved in many cancer types, and combined P value for TP53 TSS is ranked second in our analysis (P=4.26×10^{-14}). LMO3 interacts with the tumor suppressor TP53 and regulates its function, and it is ranked fourth in our analysis (P=3.25×10^{-13}). Similar to previous report, we also identified AGAP5 TSS site as a mutation hotspot that ranked third (P=7.07×10^{-14}). Another important example is the TSS sites in TERT, which is ranked fifth in our results (P=1.55×10^{-10}) and has been experimentally validated as associated with multiple types of cancer progression [7]. The discovery of such validated results proves that NIMBus can serve as a powerful tool for driver events discovery in genetic diseases.

4. Discussion

Thousands of somatic genomes are now available due to the fast development of whole genome sequencing technologies, providing us increasing statistical power to scrutinize the somatic mutation landscape. At the same time, thanks to the collaborative effort of big consortia, such as REMC and ENCODE, tens of thousands of functional characteristic experimental results on human genomes have been released for immediate use to the whole community. Hence, integrative frameworks are of urgent need to explore the interplay between WGS data and these 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 to analyze somatic genomes. Due to the heterogeneous nature of various somatic genomes, our method treated the individual mutation rate as a gamma distributed random variable to mimic the varying mutational baseline for different patients. Resultantly, it modeled the mutation counts data using a two parameter negative binomial distribution, which improved data fitting dramatically as compared to previous work (Fig. 2B). Then it uses a negative binomial regression to capture the effect of a widespread list of genomic features on mutation processes for accurate somatic burden analysis.

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 extracted 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 demonstrated three obvious advantages: 1) It incorporates the most comprehensive list of covariates in multiple tissues to achieve accurate 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 of other methods [6]. This is extremely important because the amount of available functional characteristic data is growing rapidly as the time and money cost of sequencing technologies drops quickly; 3) It automatically utilizes the genomic regions with the highest credibility for training purposes, so users are not bothered to perform carefully calibrated training data selection and complex covariate matching processes.

The length of training bins l is an important parameter for NIMBus. On one side, a shorter bin size will be advantageous in the P value evaluation as it can remove the mutational heterogeneity across regions more effectively at a higher resolution. On the other side, smaller l sometimes will result in worse mutation rate prediction performance for two reasons. First, sensible mutation rate quantification is necessary in each single bin for the regression purpose. However, somatic mutations are usually sparsely scattered across the genome due to limited number of disease genomes available at the moment. In the extreme case, when l is so small that most bins have zero mutations, it is difficult for the regression model to capture the interplay between mutations and covariates. Second, some of the covariates are only reported to be functional in a large scale[10], so reducing l will not necessarily boost prediction precision. Optimal bin size selection is still a challenging problem that needs further case-by-case investigation. In our analysis, we used a 1mb bin size for all cancer types.

In addition, noncoding regions represent more than 98% of the whole human genome, and are less investigated mainly due to limited knowledge of their biological functions. NIMBus is also designed to explore the most comprehensive noncoding annotations. Therefore, it collects the up to date full catalog of noncoding annotation of all possible tissues from ENCODE and our previous efforts from in 1000 Genomes Project. Furthermore, it further customizes these annotations specifically for somatic burden analysis. All these integrated internal annotations of NIMBus can be either tested for somatic burden or used to annotate the user specific input regions.

We applied NIMBus to 649 cancer genomes of 7 different types collected from public data and collaborators. The burden test P values for each cancer type were deduced and then Fisher's method was used to calculate the combined P values. We first evaluated the performance of NIMBus on coding regions, which have been investigated with much more detail by researchers. Many well-documented cancer associated genes were discovered by NIMBus (Table 1 and Table S3). Besides, we also repeated the same analysis on simulated dataset and found no significant genes. These results demonstrate that NIMBus is able to find overly mutated genes effectively while rigorously controlling false positives. Furthermore, numerous non-coding elements were also reported to have more mutations than expected by chance (Table C in Fig. 5D). Among these, some well-known regions, such as the TP53, LMO, and TERT TSS, were also reported in our analysis to be overly mutated, proving the effectiveness of NIMBus to identify disease 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 integrates comprehensive genomic features 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 diseases and potentially identify biological targets for drug discovery.

Abbreviation

breast cancer (BRCA), gastric cancer (GACA), liver cancer (LICA), Lung adenocarcinoma (LUAD), prostate cancer (PRAD), Medulloblastoma (MB), and Pilocytic Astrocytoma (PA)

Funding

This work was supported by the National Institutes of Health [5U41HG007000-04].

Reference

1. Kanchi, K.L., et al., *Integrated analysis of germline and somatic variants in ovarian cancer.* Nat Commun, 2014. **5**: p. 3156.
2. Lee, J.H., et al., *De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly.* Nat Genet, 2012. **44**(8): p. 941-5.
3. Lin, M.T., et al., *High aggregate burden of somatic mtDNA point mutations in aging and Alzheimer's disease brain.* Hum Mol Genet, 2002. **11**(2): p. 133-45.
4. Weinhold, N., et al., *Genome-wide analysis of noncoding regulatory mutations in cancer.* Nat Genet, 2014. **46**(11): p. 1160-5.
5. Lochovsky, L., et al., *LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations.* Nucleic Acids Res, 2015. **43**(17): p. 8123-34.
6. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-8.
7. Vinagre, J., et al., *Frequency of TERT promoter mutations in human cancers.* Nat Commun, 2013. **4**: p. 2185.
8. Manton, K.G., M.A. Woodbury, and E. Stallard, *A variance components approach to categorical data models with heterogeneous cell populations: analysis of spatial gradients in lung cancer mortality rates in North Carolina counties.* Biometrics, 1981. **37**(2): p. 259-69.
9. Chiang, C.L., *Introduction to stochastic processes in biostatistics*. Wiley series in probability and mathematical statistics. 1968, New York,: Wiley. xvi, 313 p.
10. Schuster-Bockler, B. and B. Lehner, *Chromatin organization is a major influence on regional mutation rates in human cancer cells.* Nature, 2012. **488**(7412): p. 504-7.
11. Polak, P., et al., *Cell-of-origin chromatin organization shapes the mutational landscape of cancer.* Nature, 2015. **518**(7539): p. 360-4.
12. Ernst, J. and M. Kellis, *Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues.* Nat Biotechnol, 2015. **33**(4): p. 364-76.