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

It is challenging to identify recurrently mutated regions associated with diseases because background mutation rate is usually heterogeneous and severely confounded by many known genomic features. Limited noncoding annotation information further hinders the result interpretability in such population analysis. Here, we address these issues with a Negative binomial regression based Integrative Method for mutation Burden analysis (NIMBus). It first treats mutation rates as a random variable and therefore models the over dispersed mutation count data by a negative binomial distribution. Then, to remove the confounding effect, we regress the mutation counts against 381 features extracted from REMC and ENCODE in all available tissues to accurately estimate the local background mutation rate. Such integrative framework in NIMBus can be immediately extended to accommodate new features conveniently in the future. In addition, we also customize all the noncoding annotations from ENCODE for somatic burden analysis and integrate them 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 NIMBus available as a software tool, and release our results as an online resource (nimbus.gersteinlab.org).

1. Introduction

Population level analysis is still one of the most powerful ways to identify deleterious mutations for diseases[1-3]. Nowadays, the development in whole genome sequencing (WGS) and personal genomics has provided unprecedented statistical power to perform such analysis. The burden test basically looks for recurrent variants, no matter germline or somatic, associated with diseases that occur more frequently than expected, which is called the burden test. 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, the mutation burden test for somatic variants remains a challenge 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 [6]. Hence, , the positional level mutation counts often demonstrate larger than expected variances under such assumption, which is called overdispersion. This

Jing Zhang 1/5/16 12:33 PM
Style Definition ... [1]

Jing Zhang 1/5/16 12:58 PM
Deleted:

Jing Zhang 1/5/16 12:47 PM
Deleted: Identification …dentify of ... [2]

Jing Zhang 1/5/16 12:38 PM
Formatted ... [3]

Jing Zhang 1/5/16 12:38 PM
Deleted: a new computational framework called …NIMBus). It first treats …reats ... [4]

Jing Zhang 1/5/16 12:34 PM
Deleted: With the rapid development of high throughput sequencing technologies, thousands of whole genome sequencing (WGS) data for patients with various diseases considerably 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.…The ... [5]

Jing Zhang 1/5/16 2:14 PM
Deleted: First, many *state-of-the-art* methods are optimally designed for designed for analysis[4] 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[5] in multiple cancer types. Hence, a unified coding and noncoding analysis is needed to annotate the discovered hotspots.
Lastly ... [6]

assumption results in bad fitting and generates numerous false positives [7], so it is necessary to introduce more sophisticated models to handle such mutation heterogeneity.

Second, numerous genomic features have been reported to largely affect the mutation process [4], 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, Lochovsky *et al* only corrected replication timing at relatively low resolution [7]. 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 to estimate local background mutation rate. However, as the number of covariates increases, it is usually difficult to find a meaningful neighborhood in high dimension space.

Lastly, many *state-of-the-art* methods are optimally designed for coding regions analysis [4], which 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. One well-known example is that mutations in TERT promoter were found to be associated with cancer progression [5]. 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 intuitively treats mutation rates as a random variable by scaled gamma distribution, and resultantly models the mutation counts as a type I negative binomial distribution to handle overdispersion. To capture the covariate effects, we integrate 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, we also customized the most comprehensive noncoding annotations from ENCODE to facilitate our results interpretation. Such integrative approach employed in NIMBus enables us to effectively pinpoint mutation hotspots associated with disease progression and to better understand the 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 (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. Then the bins that are overlapped with any of the two blacklist regions will be removed (details in Text S1 Section S2). Then 381 features were extracted from both REMC and ENCODE, and average signal in the bins were calculated (details see Text S1 Section S2). Let $x_{i,j}$ denote the average signal strength in the i^{th} bin and j^{th} covariate, where $i = 1,\cdots,n$ and $j = 1,\cdots,m$.

Jing Zhang 1/5/16 1:38 PM
Deleted: resulting…enerates in ... [7]

Jing Zhang 1/5/16 2:14 PM
Deleted: Lastly…econd, numerous ... [8]

Jing Zhang 1/5/16 2:14 PM
Deleted: 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.

Jing Zhang 1/5/16 2:22 PM
Formatted: No Spacing

Jing Zhang 1/5/16 2:15 PM
Deleted: Binomial …inomial ... [9]

Jing Zhang 1/5/16 2:22 PM
Formatted: Font:(Default) Times New Roman, Font color: Auto, (Asian) Chinese (PRC), Pattern: Clear

Jing Zhang 1/5/16 2:20 PM
Deleted: and use a negative binomial regression …o predict the local mutati ... [10]

Jing Zhang 1/5/16 2:31 PM
Formatted: Heading 3

Jing Zhang 1/7/16 5:50 PM
Deleted: see …n Text S1 Section S2 ... [11]

(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 samples for a specific type of disease d. Let $y_i^{d,s}$ and $\lambda_i^{d,s}$ denote the mutation counts and rate for the i^{th} bin 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 $\forall\, i, d, s$ [6]. Hence $y_i^{d,s}$ follows a Poisson distribution with the probability mass function (PMF) given in equation (1).

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

However, somatic genomes are highly heterogeneous because regions from various disease, samples, and regions of the same genome usually demonstrate considerably different mutation rates, severely violating assumptions in equation (1) [4]. As a result, the fitting of $y_i^{d,s}$ is usually very poor using (1) since a larger than expected variance, the so called overdispersion, is often observed [7]. Simply using the constant mutation rate assumption in (1) will generate numerous false positives. Instead in our model, we first pool all the samples from the same disease d together and count the mutations in region i as

$$y_i^d=\sum\nolimits_{s=1}^{S_d} y_i^{d,s} \qquad (2).$$

Then we try to model y_i^d to better handle overdispersion. Assume that $\lambda_i^d = \sum_s \lambda_i^{d,s}$ represent the overall mutation rate from all samples in region i of disease type d. Different from the constant mutation rate assumption in (1), we instead assume that λ_i^d is a random variable with a scaled gamma distribution (Γ) with parameter μ_i^d and σ_i^d in equation (3).

$$\lambda_i^d=\mu_i^d\gamma_i^d,\quad \gamma_i^d\sim\Gamma\left(1,\frac{1}{\sigma_i^d}\right) \qquad (3)$$

Then the conditional distribution of y_i^d given λ_i^d can be represented as a Poisson distribution with PMF in (4).

$$p_{Y_i^d}\left(y_i^d\middle|\lambda_i^d\right)=\frac{e^{-\lambda_i^d}\left(\lambda_i^d\right)^{y_i^d}}{y_i^d\,!} \qquad (4)$$

By integrating (3) into (4), the marginal distribution of y_i^d can be represented by a type I negative binomial distribution with PMF in (5).

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

The mean and variance of y_i^d can be described as μ_i^d and $\mu_i^d\left(1+\sigma_i^d\right)$ respectively. Our model in equation (5) is convenient with explicit interpretability. First, the gamma distribution in equation (3) conveniently models the mutation rate variability from two sources: i) for each position $p, p = 1, \cdots, l$ within bin i, we allow $\lambda_{i,p}^d$ (mutation rate for

Jing Zhang 1/5/16 2:36 PM
Formatted: Heading 4

Jing Zhang 1/5/16 2:54 PM
Deleted:

Jing Zhang 1/5/16 2:38 PM
Deleted: different

Jing Zhang 1/5/16 2:38 PM
Deleted: sections

Jing Zhang 1/5/16 2:39 PM
Deleted: very

Jing Zhang 1/5/16 2:43 PM
Deleted: This violation usually result in

Jing Zhang 1/5/16 2:47 PM
Deleted: in $y_i^{d,s}$

Jing Zhang 1/5/16 2:44 PM
Deleted: , and this phenomenon is called mutation count overdispersion

Jing Zhang 1/5/16 2:47 PM
Deleted: It

Jing Zhang 1/5/16 2:48 PM
Deleted: and needs to be carefully calibrated in somatic burden analysis

Jing Zhang 1/5/16 2:48 PM
Deleted: In

Jing Zhang 1/5/16 2:59 PM
Deleted:

Jing Zhang 1/5/16 3:14 PM
Formatted: Right

Jing Zhang 1/5/16 3:14 PM
Deleted:

Jing Zhang 1/5/16 3:14 PM
Deleted: it automatically incorporates the sample mutation rate differences. For region i of disease d, the individual contribution of mutation rate $\lambda_i^{d,s}$ from each sample s may vary as a random variable, and resultantly the overall mutation rate λ_i^d in (3) is not a constant anymore

position p in bin i for disease d) to vary from position to position in a small scale, and each $\lambda_{i,p}^{d}$ can be considered as a point on the gamma distribution in (3). ii) We can also interpret that the contribution of from $\lambda_i^{d,s}$ each sample s to λ_i^d is different and such variation can be described by a gamma shape. Second, our model in (5) 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 mutation rate.

(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 $\boldsymbol{X}$ described in 2.2. Again $x_{i,j}$ denote the average signal strength in the i^{th} bin and j^{th} covariate, where $i = 1, \cdots, n$ and $j = 1, \cdots, m$. Because we noticed that the genomic features in the covariate tables are highly correlated, which may introduce multicollinearity if directly used in the regression model. 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

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

Here we used the log function for g_1 and g_2 since y_i^d follows a negative binomial distribution, so the regression model in (6) is also called a negative binomial regression. We used the GAMLSS package in R to estimate the parameters in (6) as $\hat{\alpha}_0^d, \cdots, \hat{\alpha}_m^d, \hat{\beta}_0^d, \cdots, \hat{\beta}_m^d$.

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. They can either be internal noncoding elements such as promoters or enhancers, or elements designed by users. 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 this target region into the length of the training bins (sketch given in Fig. S2 in Text S1). Then within this extended bin, we calculate the average signal for feature j as $x_{k,j}^d, j = 1, \cdots m$, and after PCA projection, $x'^d_{k,j}$ represents 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

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

In reality, l_k, the length of the k^{th} test region that could be an enhancer or promoter is, for example, much shorter than the length of our training bins (l in section 2.2 (A),

Jing Zhang 1/5/16 3:16 PM
Deleted: -

Jing Zhang 1/5/16 3:17 PM
Deleted: in $\lambda_i^{d,s}$

Jing Zhang 1/5/16 3:17 PM
Deleted: across different samples

Jing Zhang 1/5/16 3:22 PM
Formatted: Heading 4

Jing Zhang 1/5/16 3:22 PM
Deleted: .

Jing Zhang 1/1/16 4:33 PM
Formatted: Font:Not Bold

Jing Zhang 1/1/16 4:50 PM
Formatted: No Spacing

Jing Zhang 1/1/16 4:11 PM
Formatted: No Spacing, Right

Unknown
Field Code Changed

Jing Zhang 1/1/16 4:11 PM
Formatted: Not Raised by / Lowered by

Jing Zhang 1/1/16 4:11 PM
Formatted: Not Raised by / Lowered by

Jing Zhang 1/1/16 4:38 PM
Formatted: No Spacing

Jing Zhang 1/1/16 4:28 PM
Formatted: No Spacing, Right

Unknown
Field Code Changed

Jing Zhang 1/1/16 4:28 PM
Formatted: Not Raised by / Lowered by

Jing Zhang 1/1/16 4:28 PM
Formatted: Not Raised by / Lowered by

Jing Zhang 1/1/16 4:50 PM
Formatted: No Spacing

might be up to 1Mb). Hence $\hat{\mu}_k^d$ need to be adjusted by a factor of l_k/l for the length effect. Then $\hat{\sigma}_k^d$ and the adjusted $\hat{\mu}_k^d$ can be used to calculate to p_k^d to indicate how likely it is for this this test region to have more than expected mutations through negative binomial distribution. This optimal scheme is usually computationally expensive because in a typical test there are millions of functional elements to be tested, and to calculate the average signal for all features in the extended target bin takes a long time. Therefore, we proposed an approximation method to approximate 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 several related diseases (or disease subtypes) needs to be analyzed together to provide a combined P value. One typical example is the pan-cancer analysis. In section 2.3 (C), we calculated the P value for disease/disease type d as p_k^d for test region k and used to the Fisher's method to combine P values. Specifically, the test statistic can be calculated in (8)

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

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

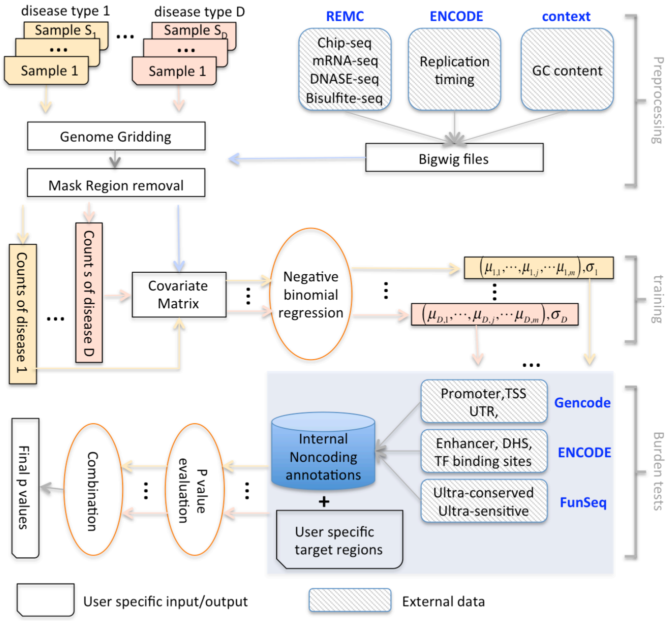
2.4 Noncoding annotations customized for NIMBus

We customized the full list of noncoding annotations from both ENCODE annotations and our previous efforts from our experience in 1000 genomes projects. More details are given in Text S1.

2.5 Flowchart of NIMBus

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

Figure 1. Flowchart of NIMBus



Jing Zhang 1/1/16 5:02 PM
Moved (insertion) [1]

Jing Zhang 1/5/16 4:39 PM
Deleted: *2.6*

Jing Zhang 1/5/16 4:40 PM
Deleted: ,

Jing Zhang 1/5/16 4:40 PM
Deleted: s

Jing Zhang 1/5/16 4:40 PM
Deleted: of different

Jing Zhang 1/1/16 5:08 PM
Formatted: No Spacing, Right

Unknown
Field Code Changed

Jing Zhang 1/1/16 5:10 PM
Formatted: No Spacing

Jing Zhang 1/1/16 5:12 PM
Deleted: 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.
2.6 Combining P values ... [12]

Jing Zhang 1/1/16 5:02 PM
Moved up [1]: 2.6 Combining P values

Jing Zhang 1/1/16 5:12 PM
Deleted: 2.6 Combining P values ... [13]

Jing Zhang 1/1/16 5:02 PM
Deleted: 7

Jing Zhang 1/5/16 10:11 PM
Formatted: Font:10 pt

3. Results

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

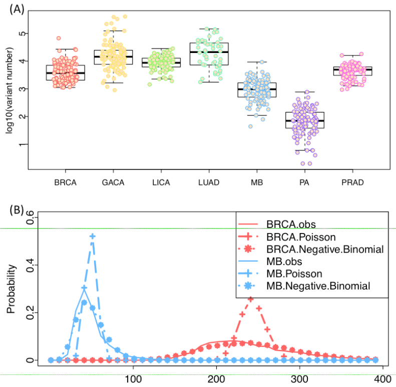


Figure 2. (A) Disease and sample mutation rate heterogeneity; (B) improved fitting by negative binomial distribution

Pioneer genome wide somatic burden analysis usually assumes a homogeneous mutation rate, which consequently uses binomial or Poisson tests to evaluate P values [6]. 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 (Fig. S1).

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. Even within the same disease, mutation rates 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. 3). Therefore, a binomial or Poisson distribution usually provides poor fitting to the real mutation counts data (Fig. 2B, dash 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 (dash lines with star in Fig. 2B).

3.2 Local mutation rate is confounded by other genomic features

Somatic mutation rate has been reported in literature to be confounded by several genomic features [4, 8]. 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. We observed a similar trend in our data. For example, in breast cancer samples, the Pearson correlation between normalized mutation counts and replication timing values is as high as 0.670 in the first 70 1mb bins (Fig. S4 A). Another example is the chromatin organization, which arranges the genome into heterochromatin- and euchromatin-like domains and has a dominant influence on regional mutation-rate variation in human somatic cells [8]. We also find that mutation counts are significantly associated with the DNASeq signal in breast cancer (Pearson correlation=−0.614, P value=1.524e-08, Fig S4 D in Text S1).

Jing Zhang 1/5/16 9:51 PM
Deleted: as

Jing Zhang 1/5/16 9:54 PM
Deleted: shown in

Jing Zhang 1/5/16 9:51 PM
Deleted: 2A

Jing Zhang 1/5/16 9:54 PM
Formatted: No Spacing

Jing Zhang 1/5/16 9:54 PM
Deleted: (Fig. 2B-C)

Jing Zhang 1/5/16 9:56 PM
Deleted: may dramatically change

Jing Zhang 1/5/16 9:59 PM
Deleted: 2D

Jing Zhang 1/5/16 9:52 PM
Deleted: overdispersed

Jing Zhang 1/5/16 9:59 PM
Deleted: Fig. 2D

Jing Zhang 1/5/16 11:42 PM
Deleted: *<sp>*Instead of homogeneously located across the genome in a random way, s

Jing Zhang 1/5/16 11:50 PM
Deleted: 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 \{cite Mutsig and refer}. Furthermore, it has been reported that

Jing Zhang 1/5/16 11:50 PM
Deleted: ,

Jing Zhang 1/5/16 11:45 PM
Deleted: \{cite 22820252}

Jing Zhang 1/5/16 11:51 PM
Deleted: a similar trend in our data analysis

Jing Zhang 1/5/16 11:54 PM
Deleted: 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.

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

Features in matched tissues provides best prediction accuracy but features in unmatched tissue still helps

It has been reported that most accurate local mutation rate prediction can be achieved by using the matched tissue [9]. Hence, we specifically selected variants in two distinct cancer types BRCA and MB and predicted their local mutation rate by 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 (equation s3 in Text S1), was used to compare model performance. Consistent with previous conclusions, we observed 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 is more accurate 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 expertise. 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 the suboptimal but still decent performance during our regression (scatter plots given in Fig. S6B). This is extremely helpful when processing 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 background mutation rate prediction

In light of the correlated nature of covariates, especially those from epigenetics experiments [10], we first performed principle component analysis (PCA) on the covariate matrix to overcome the multi-collinearly problem during regression. 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 (Fig. S7A in Text S1). 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 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 effect of multiple PCs to jointly predict the local mutation rates. Specifically, for each cancer type, we first ranked the individual PCs by their correlations with mutation rates, and then only selected the top 1, 30, and all PCs to predict the local mutation rate. Fig. 3A shows that to use more PCs can boost prediction performance noticeably in all cancer types. 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 suggested to use all PCs for accurate local mutation rate estimation.

Jing Zhang 1/6/16 8:50 AM
Formatted: Heading 4

Jing Zhang 1/6/16 12:01 PM
Formatted: No Spacing

Jing Zhang 1/6/16 1:19 PM
Deleted: ***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 < 2.2*** ... [14]

Jing Zhang 1/6/16 1:24 PM
Deleted: To

Jing Zhang 1/6/16 1:25 PM
Deleted: due to the correlated nature of genomic features

Jing Zhang 1/6/16 1:25 PM
Deleted: , we first performed a principle component analysis (PCA) on the 381 features.

Jing Zhang 1/6/16 3:03 PM
Deleted: 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

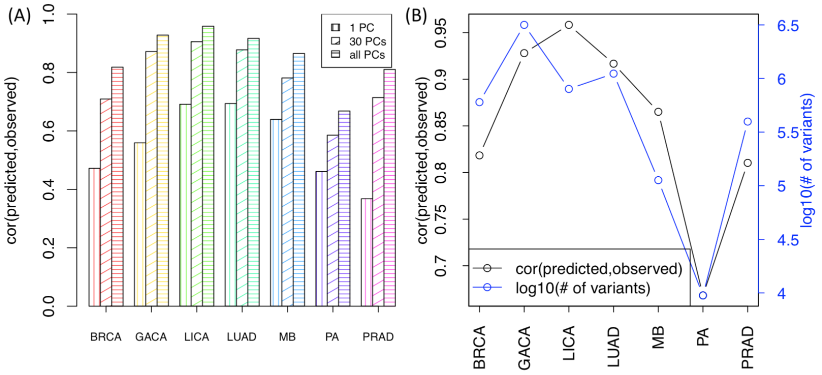
Jing Zhang 1/6/16 3:38 PM
Deleted: 4C

Jing Zhang 1/6/16 3:38 PM
Deleted: in all 7 cancer types, using

Jing Zhang 1/6/16 3:38 PM
Deleted: noticeably

Jing Zhang 1/6/16 3:39 PM
Deleted: used

Figure 3. (A) Regression performance by correcting different number of PCs; (B) Relationship of regression performance against total number of variants used.



As it is shown in Fig. 3B, we obtained good prediction accuracy through regression against all PCs for the covariate matrix 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. S8 in Text S1. It is worth mentioning that although there is no prostate tissue related data in REMC, we can still achieve a very decent correlation of 0.81 with the help of 381 unmatched but still correlated features. It indicates that even when somatic WGS of a disease is given without optimally matched covariates, our model could still achieve acceptable performance.

In addition, the number of available variants obviously affects prediction performance, but it is not the only effective factor. As shown in Fig. 3B, limited number of variants, such as in quiet somatic genomes like PA, usually restricts 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 could also influence the prediction performance considerably. 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. 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 to check its performance. First coding regions were extracted from the genecode annotation v19 and NIMBus was run on both real and simulated datasets (details in Text S1). We found that in all cancer types analyzed, NIMBus provided reasonable P values in real WGS data as compared to the simulated data. For example, in LUAD the P values for real data follows nicely with the uniform P values with a few exceptions as the true signals (black lines in Fig. S4). However, in the simulated data no highly mutated genes were discovered (orange lines in Fig. S4).

We also used Fisher's method to combine P values from all cancer types. In total 15 genes has been discovered to be hyper-mutated and 12 out of them are well documented as related with cancer progression. The top genes are shown in Table 1 and Pubmed ID for related reference was given in the last column. These results showed that NIMBus is

Jing Zhang 1/6/16 3:40 PM
Formatted: Centered

Jing Zhang 1/6/16 3:47 PM
Deleted: Our PCA based negative binomial regression can precisely capture the effect of various genomic features on somatic n ... [15]

Jing Zhang 1/6/16 3:56 PM
Deleted: 4D

Jing Zhang 1/6/16 3:57 PM
Deleted: data

Jing Zhang 1/6/16 3:58 PM
Deleted: somatic mutation data

Jing Zhang 1/6/16 3:58 PM
Deleted: an

Jing Zhang 1/6/16 3:58 PM
Deleted: unknown

Jing Zhang 1/6/16 3:59 PM
Deleted: ,

Jing Zhang 1/6/16 3:59 PM
Deleted: performance without the ... [16]

Jing Zhang 1/6/16 4:00 PM
Deleted: our regression performance ... [17]

Jing Zhang 1/6/16 4:02 PM
Deleted: total

Jing Zhang 1/6/16 4:04 PM
Deleted: variants

Jing Zhang 1/6/16 4:03 PM
Deleted: (bottom right figure in Fig. 4D)

Jing Zhang 1/6/16 4:04 PM
Deleted: Obviously

Jing Zhang 1/6/16 4:04 PM
Deleted: mutations

Jing Zhang 1/6/16 4:05 PM
Deleted: could restrict the prediction ... [18]

Jing Zhang 1/6/16 4:05 PM
Deleted: very

Jing Zhang 1/6/16 4:07 PM
Deleted: Other

Jing Zhang 1/6/16 4:07 PM
Deleted: considerably

Jing Zhang 1/6/16 4:07 PM
Deleted: of our model

Jing Zhang 1/6/16 4:08 PM
Deleted: 4D

Jing Zhang 1/7/16 4:07 PM
Formatted: Normal

Jing Zhang 1/7/16 4:10 PM
Deleted:

able to find sensible mutational hotspots and disease drivers but at the same time effectively control the number of false positives.

Fig. 4. Q-Q plots of P values

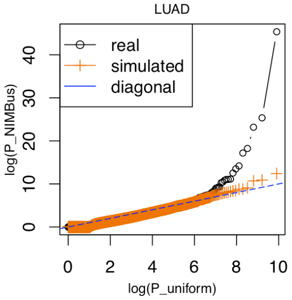


Table 1. Top genes after P value combination

rank	gene	Adjust P	PubMed ID
1	TP53	4.50E-139	17401424
2	DDX3X	3.79E-18	22820256
3	KRAS	2.66E-06	19847166
4	MUC4	4.64E-06	19935676
5	CDH1	2.65E-05	10973239
6	ARID1A	2.10E-4	22037554
7	SMARCA4	3.43E-4	18386774
8	FGFR1	6.86E-4	23817572

3.5. 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. 2B). 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).

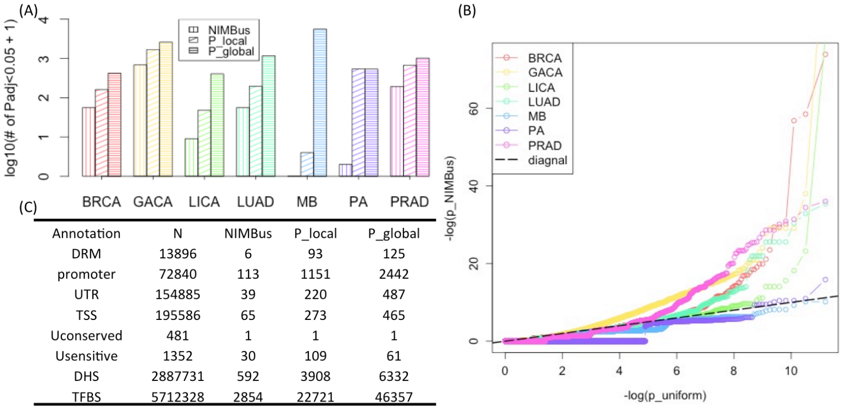
Jing Zhang 1/7/16 4:57 PM
Formatted: No Spacing

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

Jing Zhang 1/7/16 4:57 PM
Deleted: ... [19]

[11]. 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 [5]. The discovery of such validated results proved that NIMBus could serve as an powerful tool for driver events discovery in diseases.

Fig. 4 (A) number of detected prompter 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



Annotation	N	NIMBus	P_local	P_global
DRM	13896	6	93	125
promoter	72840	113	1151	2442
UTR	154885	39	220	487
TSS	195586	65	273	465
Uconserved	481	1	1	1
Usensitive	1352	30	109	61
DHS	2887731	592	3908	6332
TFBS	5712328	2854	22721	46357

4. Discussion

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. 2B).

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,

Jing Zhang 1/7/16 4:55 PM
Deleted: *<sp>*

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 list of comprehensive covariates in multiple tissues 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 [4]. 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 quickly; 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.

Jing Zhang 1/7/16 5:06 PM
Comment [1]: Is this too strong?

Jing Zhang 1/7/16 5:06 PM
Formatted: Highlight

In addition, we also put a lot of effort on NIMBus to explore the most extensive noncoding annotations that is customized for somatic burden analysis. 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. We evaluated the performance of NIMBus on coding regions, which were investigated with much more detail by researchers. A list of well-documented cancer related genes has been discovered by NIMBus (Table 1 and Table S3). Besides, we also repeated the same analysis on simulated dataset and found no significant genes. There results demonstrate that NIMBus is able to find hyer-mutated genes effectively while controlling false positives. Furthermore, 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.

Jing Zhang 1/7/16 5:09 PM
Deleted: As a result

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.

Funding

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. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-8.
5. Vinagre, J., et al., *Frequency of TERT promoter mutations in human cancers.* Nat Commun, 2013. **4**: p. 2185.
6. Weinhold, N., et al., *Genome-wide analysis of noncoding regulatory mutations in cancer.* Nat Genet, 2014. **46**(11): p. 1160-5.
7. 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.
8. 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.
9. Polak, P., et al., *Cell-of-origin chromatin organization shapes the mutational landscape of cancer.* Nature, 2015. **518**(7539): p. 360-4.
10. 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.
11. Park, J. and S. Jeong, *Wnt activated beta-catenin and YAP proteins enhance the expression of non-coding RNA component of RNase MRP in colon cancer cells.* Oncotarget, 2015. **6**(33): p. 34658-68.