Supplementary Material for NIMBus

S1. WGS data collection

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 adenocarcinoma (LUAD), prostate cancer (PRAD), Medulloblastoma (MB), and Pilocytic Astrocytoma (PA). Fig. S1 gives the pie chart of sample numbers for these cancer types.

Figure S1. Pie chart of sample numbers of the WGS data

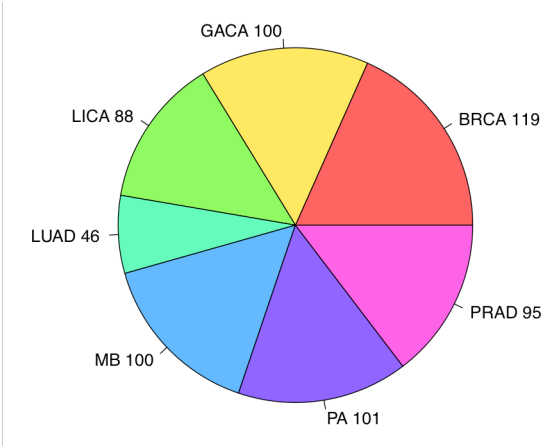


Table S1. Summary of WGS data

cancer	median	sd
BRCA	3705	7300.526
GACA	14429.5	71372.080
LICA	8706.5	5522.917
LUAD	21287	35610.839
MB	965	1196.036
PA	70	114.414
PRAD	4927	2764.873

Among these samples, 100 stomach cancer samples were from Wang *et al* [1] and 95 prostate cancer samples were obtained from our collaborators. The remaining comes from samples published by Alexandrov *et al* [2].

S2. Covariate table generation

Numerous studies showed many genomic features severely affect the mutation process, and such covariate effect should be removed for somatic burden analysis [3, 4]. We created the covariate table in three steps.

Step 1. Covariate collection and bigWig file generation

We first collected all the signal track files from major histone modification marks, chromatin status, methylation, and mRNA-seq data from the REMC. Only the experimental real data, as opposed to the imputed data, was used at this step. For each feature, we then processed these signal files into bigWig format (https://genome.ucsc.edu/goldenpath/help/bigWig.html) at 20nt resolution. If some file, for example the chromatin accessibility, is missing in a specific tissue, we skipped it in our feature selection step. If multiple replicates were found in their original data, the averaged signal after normalization was used in the final bigWig file.

Since replication timing has been proved to be associated with mutation rate in several articles [3-5], we also collected 8 replication timing bigWig files from the ENCODE project. Lastly, as researchers have observed elevated mutation rates in regions wither lower GC content in certain diseases, we also include the GC percentage files in our covariate list and generated its corresponding bigWig files.

Step 2. Genome gridding

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 [6], we divided the whole genome into bins with fixed length, such as 1mb, 100kb, 50kb, etc. Only autosomal chromosomes and chromosome X were included in our analysis to remove the gender imbalance in mutation data or covariates.

Repetitive regions on human genome are known to generate artifacts in high throughput sequencing analysis mainly due to their 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 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 our analysis.

Step 3: covariate table creation

All 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 above. 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.

S3. Noncoding annotations

We collected the full list of noncoding annotations to the best of our knowledge and customized it suitable for burden analysis. This list includes 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 of known protein coding genes were defined as the 2500 and 100 nucleotides (nt) before the transcripts annotated by GENCODE v19. We also collected all the TFBS and enhancers

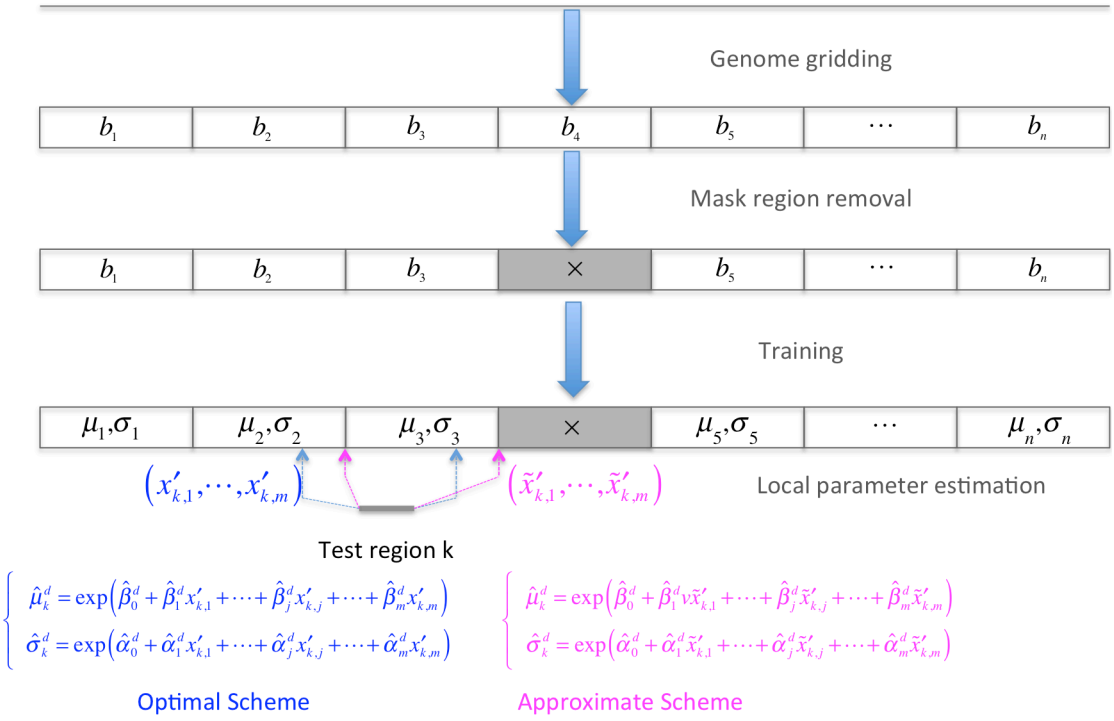
from all tissues that are uniformly processed through the ENCODE pipeline. In addition, the ultra-conserved and ultra-sensitive sites were defined as those under positive selection during transcription regulations in our previous method FunSeq [7].

S4. Optimal and approximate local background mutation rate estimation

After the training process through equation (6) in the main manuscript, the estimates of parameters for negative binomial regression can be represented by $\hat{\alpha}_0^d, \cdots, \hat{\alpha}_m^d, \hat{\beta}_0^d, \cdots, \hat{\beta}_m^d$. To obtain the optimal local mutation rate for test region k, which may be either an internal noncoding annotation such as enhancer or a user-defined element, we should first extend this region into the training bin length l centered at the center of test region k (blue parts in Fig. S2). Then the covariates values after PCA projection in this extended bin should be calculate as $(x'_{k,1}, \cdots, x'_{k,m})$. Hence in this scheme, the local mutation parameters should be calculated as

$$\begin{aligned} \hat{\mu}_k^d &= \exp\left(\hat{\beta}_0^d + \hat{\beta}_1^d x'_{k,1} + \cdots + \hat{\beta}_j^d x'_{k,j} + \cdots + \hat{\beta}_m^d x'_{k,m}\right) \\ \hat{\sigma}_k^d &= \exp\left(\hat{\alpha}_0^d + \hat{\alpha}_1^d x'_{k,1} + \cdots + \hat{\alpha}_j^d x'_{k,j} + \cdots + \hat{\alpha}_m^d x'_{k,m}\right) \end{aligned} \quad \text{(s1).}$$

Figure S2. Schematic sketch of optimal and approximate local mutation rate estimation



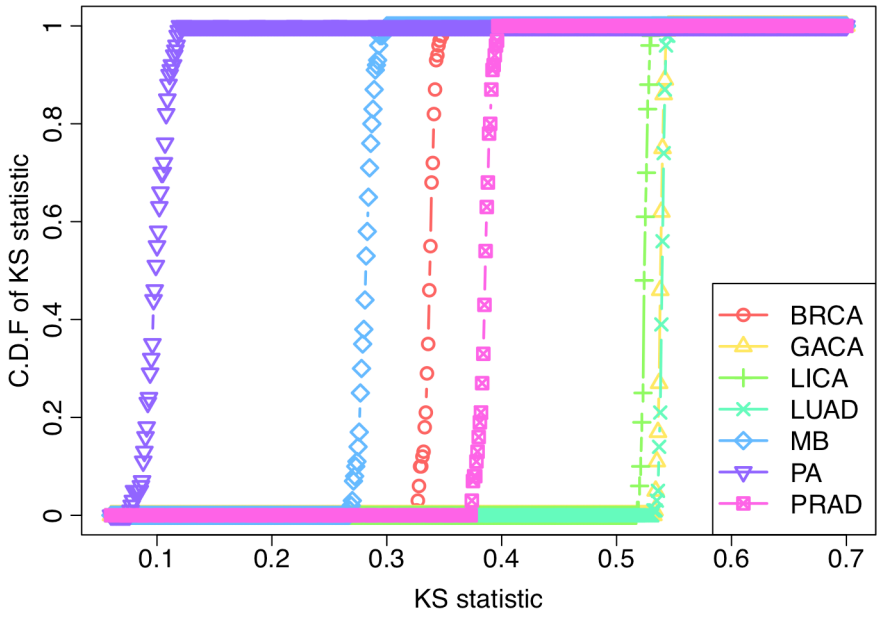
However, in real data analysis there are usually millions of regions to be tested and for each region it needs to process 381 features. Hence, the above optimal scheme is usually computational expensive. Here we proposed an approximation scheme to calculate $\hat{\mu}_k^d$ and $\hat{\sigma}_k^d$. Instead of using covariates for the extended bin centered at target region k, we used the values for the nearest training bin $(\tilde{x}'_{k,1}, \cdots, \tilde{x}'_{k,m})$ (magenta parts in

Fig. S2), and burden tests are performed after length adjustment. Since $(\tilde{x}'_{k,1}, \cdots, \tilde{x}'_{k,m})$ has already been pre-calculated during the training process, our approximation scheme significantly reduced the computation burden for tests.

S5. Kolmogorov–Smirnov (KS) statistic to compare mutation counts in 1mb bins again fitted ones using Binomial distribution

In order to check the degree of overdispersion in the mutation counts by Binomial assumption in [8], we compared the observed and fitted mutation count data by Binomial distribution and provided the KS statistic in each cancer type. Specifically, we counted the number of mutations y_i^d in n 1mb bins generated in section S2 step 2. Then the maximum likelihood estimate of mutation rate $\hat{\lambda}_d$ per position under the constant mutation rate assumption is calculated for cancer type d. Then we randomly generated n simulated mutation counts $\acute{y}_i^d$ with $\hat{\lambda}_d$ and calculated the KS statistic. We repeated the above process 100 times and plot the cumulative density function (C.D.F) of these KS statistics. A large KS statistic near 1 indicates larger overdispersion in the mutation count data. From Fig. S3, we showed that in all 7 cancer types, Binomial model provides poor fitting.

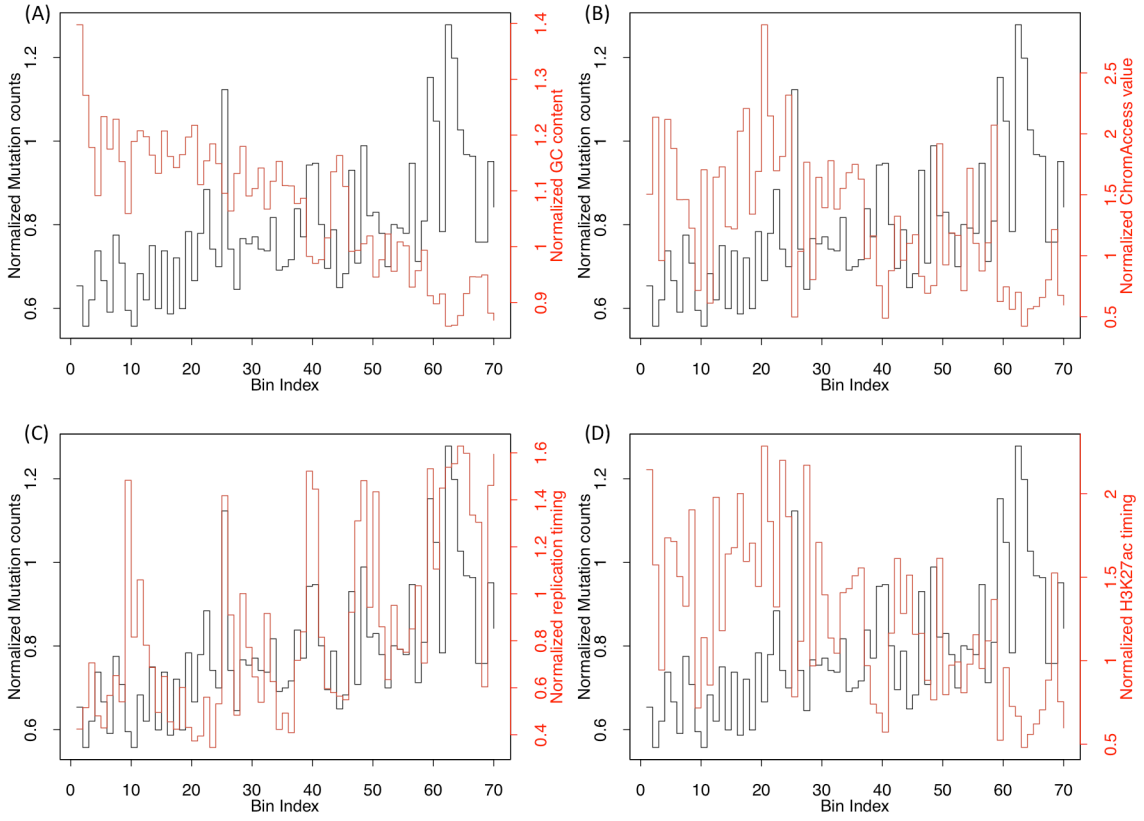
Figure S3. KS statistic of the observed and fitted mutation counts using Binomial distribution



S6. Local mutation rates are highly correlated with many genomic features

It has been reported that local mutation rates are associated with many well-known genomic features, such as mRNA expression, GC content, replication timing, and chromatin organization [3]. We found that the WGS data in our datasets also demonstrated similar characteristics. For example, Fig. S4 shows how mutation counts at a 1mb resolution (the first 70 bins on chromosome 1) are correlated with several genomic features.

Figure S4. Mutation rates are severely affected by many genomic features in breast cancer in the first 70 1mb bins on chromosome 1. Mutation counts (black line and left y axis) and other genomic features (red line and right y axis) are normalized in a genome-wide way.



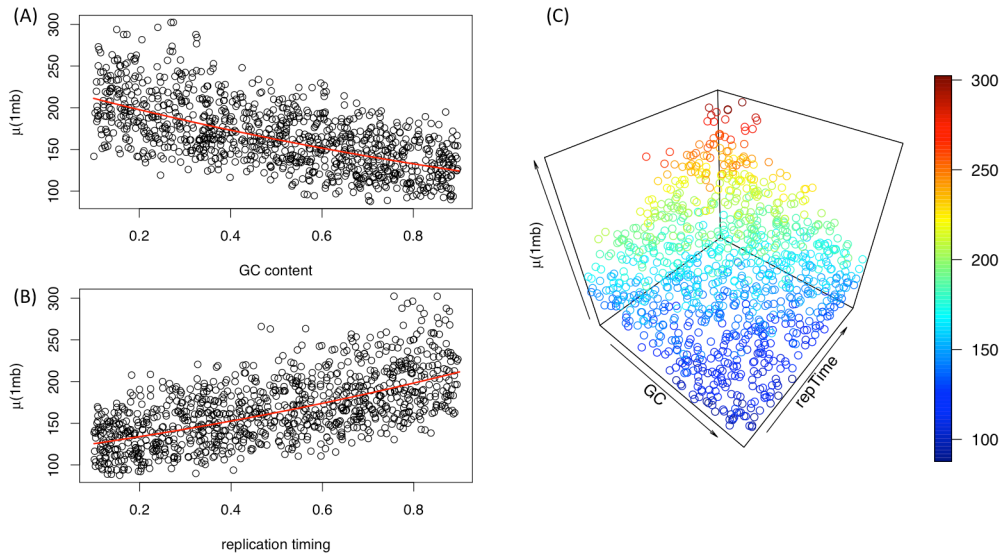
S7. Toy example to show how the regression runs with multiple features

In order to illustrate how the negative binomial regression works, we provided a toy example in Fig. S5. Suppose in a specific disease d, let μ_i^d represent the mean values of mutation counts in the i^{th} bin, and $x_{i,1}^d$ and $x_{i,2}^d$ denote the GC content and replication timing in the same bin. In this example, we suppose that

$$log(\mu_i^d) = 0.670x_{i,1} - 0.683x_{i,2} + \varepsilon_i \quad \text{(s2)},$$

where ε_i is a Gaussian random variable with mean 0.6 and standard deviation 0.1. Suppose $x_{i,1}^d$ and $x_{i,2}^d$ are two uniformly distributed random variables between [0.1,0.9], we generated 1000 points and showed how $x_{i,1}^d$ and $x_{i,2}^d$ jointly affect the local mutation rate through Fig. S5

Figure S5. Toy example to show how two parameters jointly affect the mutation rate. (A-B): scatter plot of mutation rate vs. covariates. Red line is the fitted line using only one covariate; (C) 3D scatter plot of mutation rates and two covariates



S8. Regression Performance comparison using features from matched and unmatched tissues

Specifically, we represented mutations rates in BRCA and MB as μ_i^B and μ_i^M for the i^{th} bin 1mb bin. 7 genomic features in breast related features were extracted from REMC, including DNASeq, H3K27me3, H3K36me3, H3K4me3, H3K9me3, mRNA-seq and methylation data (features starting with B_ in Fig. S6A), denoted by $x_{i,1}^B, \cdots, x_{i,7}^B$. Similarly we also got 8 features in brain related tissues for MB denoted by $x_{i,1}^M, \cdots, x_{i,8}^M$ (H3K27me3, H3K27ac, H3K36me3, H3K4me1, H3K4me3, H3K9me3, mRNA-seq and methylation, features starting with M in Fig. S6A). We found that these features are highly correlated both within and across tissues (as shown in the correlation plot in Fig. S6A).

To compare the performance of regressions using (loosely) matched and unmatched tissues, four regression models can be run as shown in Table S2. The scatter plots of the observed and predicted values were given in Fig. S6B. To compare model performance, we defined the relative error e_i^d as

$$e_i^d = \frac{|\mu_i^d - \hat{\mu}_i^d|}{\mu_i^d} \quad \text{(s3)}.$$

Relative errors for these four models were given in Table S3.

Figure. S6. (A) Features in breast and brain related tissues are highly correlated; (B) scatter plot of observed and predicted mutations by regression for all four models. Red line represents the diagonal line and Pearson correlations are also given.

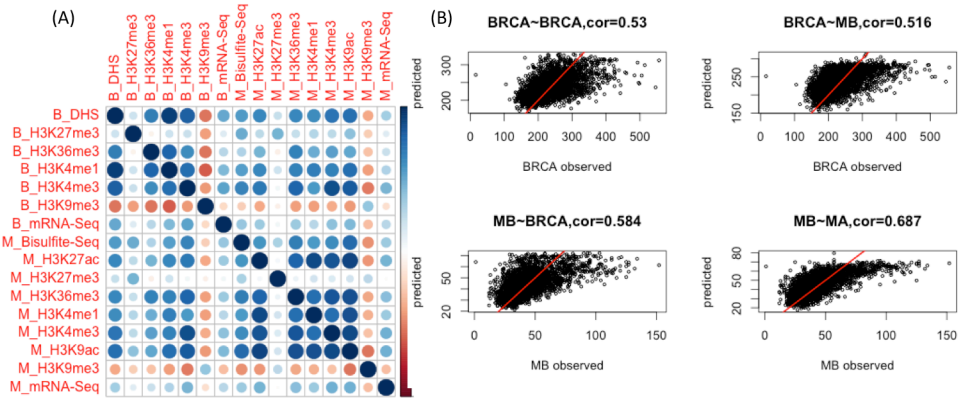


Table S2. Four regression models using matched and unmatched genomic features

Mutations \ Covariates	BRCA	MB
BRCA	$\log(\mu_i^B) = \beta_0 + \sum_{j=1}^{7} \beta_{i,j} x_{i,j}^B$	$\log(\mu_i^B) = \beta_0 + \sum_{j=1}^{8} \beta_{i,j} x_{i,j}^M$
MB	$\log(\mu_i^M) = \beta_0 + \sum_{j=1}^{7} \beta_{i,j} x_{i,j}^B$	$\log(\mu_i^M) = \beta_0 + \sum_{j=1}^{8} \beta_{i,j} x_{i,j}^M$

Table S3. Relative error for models using matched and unmatched genomic features

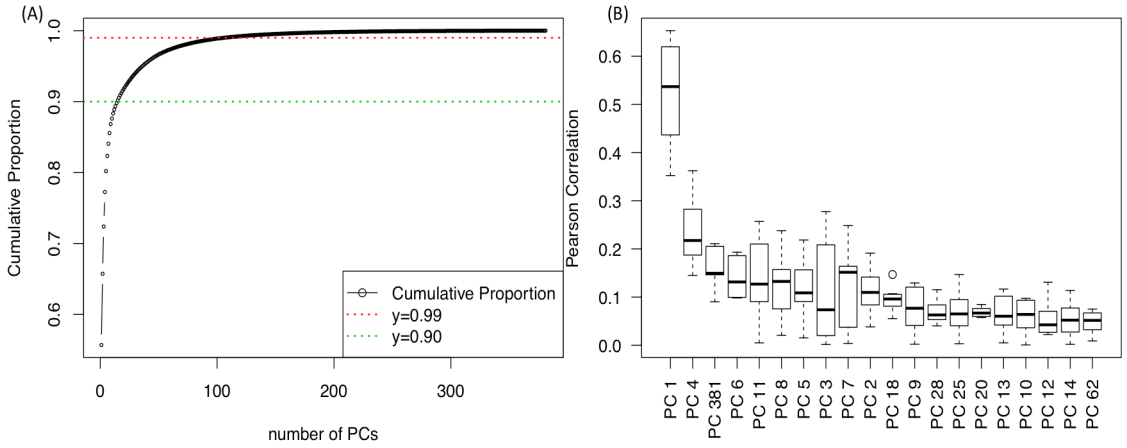
Mutations \ Covariates	BRCA	MB
BRCA	0.128	0.135
MB	0.195	0.183

S9. PCA analysis of the covariate matrix

It has been reported that many genomic signal tracks demonstrate noticeable correlations across features and tissues [9]. Hence we first centered and scaled the covariate matrix X and then performed PCA on it to obtain $\acute{X}$. Then the cumulative proportion of variance explained by the PCs was given in Fig. S7 A. As expected, there is lots of redundancy in the covariate table. The first PC may explain as much as 55.69% of variance. And it takes up to 15 and 106 PCs to capture 90% and 99% of variance.

We also calculated the Pearson correlation of PC j with mutation counts in cancer type d as ρ_j^d. Then the absolute correlation value $\left|\rho_j^d\right|$ were averaged over different cancer types as $\widehat{\rho_j}$ to rank the PCs. The top 20 PCs with highest $\widehat{\rho_j}$ were selected and boxplot for each of the PCs was given in Fig. 7B.

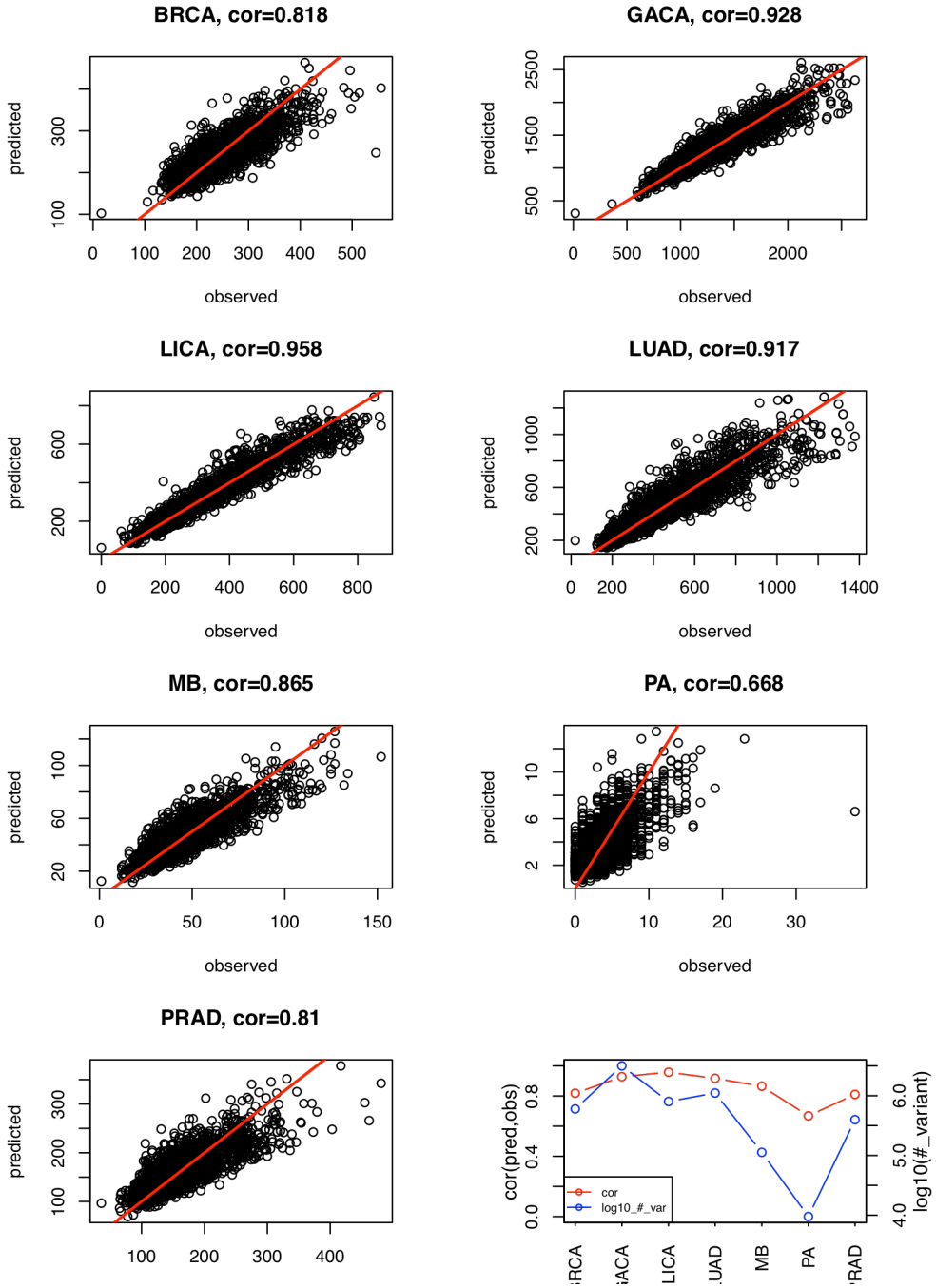
Figure S7. (A) Cumulative proportion of variance explained by the number of PCs; (B) Boxplot of Pearson correlations of top PCs to mutation counts data in different cancer types.



S10. Performance of local background mutation rate estimation by correcting all PCs for the covariate matrix

For each cancer type, we tried to predict the local mutation rate by correcting the covariate matrix after PCA projection. Then the Pearson correlation of the predicted and observed mutation rates are given in Fig. S8.

Figure S8. Scatter plot of observed and predicted mutation rate. Red line represents the diagonal line.

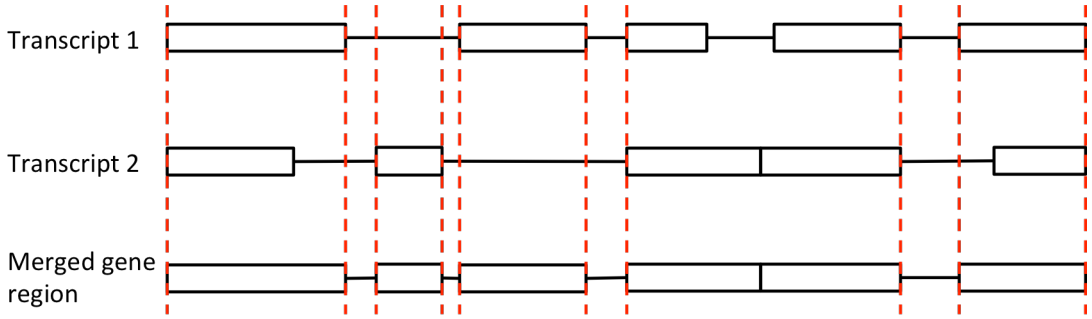


S11. Coding region analysis on both real and simulated data

Coding region extraction

We first extracted all the coding regions from the Gencode v19 annotation. For annotation accuracy, we only selected the protein coding genes with gene_status labeled as "KNOWN" from the annotation. Then all the protein coding transcripts of the selected genes were selected. We merged multiple transcripts to get the final protein coding gene annotation as shown in Fig. S9. In total, 19,291 known protein-coding genes haven been used in this analysis.

Figure S9. Example of protein coding gene merging



Simulated variants for all cancer types

For each variant in a set of whole genome sequencing data, we tried to find a new position in a 100kb neighboring region (50k and 50k up and downstream each). Then we tested all the coding genes defined above on the original and simulated data set. Since the permuted size 100kb is relatively large as compared to the test region, a better method is supposed to give less or even no positives on the permuted data set. The Q-Q plots of P values of protein coding genes in both real and simulated data were given in Fig. S10.

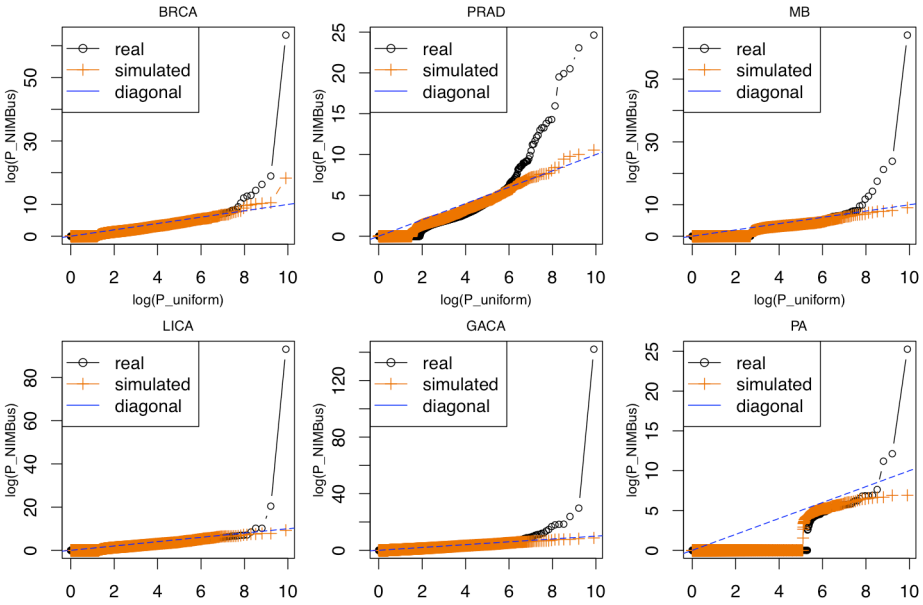
Significant genes after multi-test correction for the combined P values

Table S4 significant 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	2.56097E-06	19847166
4	MUC4	4.46636E-06	19935676
5	CDH1	3.06519E-05	10973239
6	ARID1A	0.000236255	22037554
7	SMARCA4	0.000377726	18386774
8	FGFR1	0.000742868	23817572
9	OTOP1	0.003694811	NA

10	SPOP	0.004493795	22610119
11	STK11	0.004493795	15021901
12	PTEN	0.004812088	9697695
13	SMO	0.012684045	9422511
14	TAS2R31	0.033983962	NA
15	TBC1D29	0.034276895	NA

Figure S10. Q-Q plots of gene regions on real and simulated data



S12. Comparison with local and global binomial models

In [8], after pooling samples from a certain disease, a constant mutation rate was assumed at each single nucleotide over the genome. Hence, the number of mutations y_i^d within a region with length l_i follows a Binomial distribution as

$$P\left\{Y_i^d = y_i^d\right\} = \begin{pmatrix} n_i \\ p_i^d \end{pmatrix} \left(p_i^d\right)^{y_i^d} \left(1 - p_i^d\right)^{n_i - y_i^d} \quad \text{(s4)},$$

where p_i^d is the mutation rate at a single nucleotide. In a global Binomial model, $p_i^d \equiv p$ is assumed, and p is calculated in a genome-wide way. To remove the covariate effect, we may also assume a local Binomial model by using different p_i^d for different regions. Specifically, p_i^d can be approximated by the length normalized μ_i^d in NIMBus.

S13. Reference

1. Wang, K., et al., *Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer.* Nat Genet, 2014. **46**(6): p. 573-82.
2. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-21.
3. 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.
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. 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. Melton, C., et al., *Recurrent somatic mutations in regulatory regions of human cancer genomes.* Nat Genet, 2015. **47**(7): p. 710-6.
7. Khurana, E., et al., *Integrative annotation of variants from 1092 humans: application to cancer genomics.* Science, 2013. **342**(6154): p. 1235587.
8. Weinhold, N., et al., *Genome-wide analysis of noncoding regulatory mutations in cancer.* Nat Genet, 2014. **46**(11): p. 1160-5.
9. 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.