A site specific model and analysis of the neutral somatic mutation rate in whole-genome cancer data

Johanna Bertl [*‡] Qianyun Guo [†‡] Malene Juul Rasmussen [*]
Søren Besenbacher [*] Morten Muhlig Nielsen [*] Henrik Hornshøj [*]
Jakob Skou Pedersen [*§] Asger Hobolth [†§]

Key words: Logistic regression, site-specific model, somatic cancer mutations

Abstract

Understanding and modelling the neutral mutational process in cancer cells is crucial for identifying the mutations that drive cancer development. The neutral mutational process is very complex: Whole-genome analyses have revealed that the mutation rate differs between cancer types, between patients, and along the genome depending on the genetic and epigenetic context. Therefore, methods that predict the number of different types of mutations in regions or specific genomic elements must consider local genomic explanatory variables. A major drawback of most methods is the need to average the explanatory variables across the entire region or genomic element. This procedure is particularly problematic if the explanatory variable varies dramatically in the element under consideration. Instead, we model the probabilities of the different types of mutations for each position in the genome by multinomial logistic regression. We apply our site-specific model to a data set of 505 cancer genomes from 14 different cancer types. We show that for 1000 randomly selected genomic positions, the site-specific model predicts the mutation rate much better than regional based models. We use a forward selection procedure to identify the most important explanatory variables. The procedure identifies site-specific conservation (phyloP), replication timing, and expression level as the best predictors for the mutation rate. Finally, our model confirms certain well-known mutational signatures. Our site-specific multinomial regression model can serve as the neutral null model for the mutational process; regions that deviate from the null model are candidates for elements that drive cancer development.

[*]Department of Molecular Medicine, Aarhus University, Denmark
[†]Bioinformatics Research Centre, Aarhus University, Denmark
[‡]These authors contributed equally to this work.
[§]These authors contributed equally to this work.

1 Introduction

Cancer is driven by somatic mutations that convey a selective advantage to the cell. However, in most cases the somatic mutation rate in cancer cells is considerably higher than in healthy tissues, of which only a small fraction of the mutations are thought to be associated with cancer development [Stratton et al., 2009]. The majority of the mutations are neutral and is caused by perturbed cell division, maintenance and repair or over-expression of mutagenic proteins (e.g. the APOBEC gene family [Bacolla et al., 2014]).

A comprehensive framework of the random mutation process in cancer cells is key to identify the regions, pathways and functional units that are under positive selection during cancer development. The mutation rate varies along the genome, depending on genomic properties of the position such as the sequence context (e. g. the 5' and 3' nucleotides; [Alexandrov et al., 2013b]), chromatin organisation [Polak et al., 2015] or replication timing [Lawrence et al., 2013]. Many studies have investigated what determines the mutation rate and what kind of models should be used. Lawrence et al. [2013] modelled the mutation heterogeneity using local regression with expression level and replication timing as explanatory variables. Melton et al. [2015] implemented a Poisson-binomial model on 50 kb windows using the annotation of the transcription factor binding site and replication timing. Polak et al. [2015] applied random forest regression on mutation counts in 1Mb windows using histone modification and density of DNase I hypersensitive sites. Lochovsky et al. [2015] predicted the number of mutations per element by a beta-binomial distribution using replication timing and noncoding annotations such as promoter, UTR, ultra-conserved sites, etc. All of these studies segmented the genome into regions according to the explanatory variables and estimated separate models for different regions. In this paper, we propose a framework based on site-specific multinomial regression where this division is not necessary. We can use site-specific explanatory variables (e.g. CpG) without dividing the data into subsets. In this way, we use the full dataset to estimate the regression coefficients for all the explanatory variables.

Generalized linear models also provide an interpretable framework for modeling and hypothesis testing. For example, we can easily estimate the mutation rates of CpG sites within and outside CpG islands, and test if they are different, and a patient-specific intercept allows us to take the large variation of mutation rates between patients into account.

Here, we analyse 505 cancer genomes from 14 different cancer types (Fredriksson et al. [2014b]). We compare the performance in predicting mutation probabilities for region-based Poisson models, site-specific binomial models and site-specific multinomial models. The site-specific multinomial model can predict both the overall mutation rate and mutation types accurately. We use a forward model selection procedure to compare and identify the explanatory variables that best explain the heterogeneity of the site-specific

mutational process (Figure 1).

The forward model selection procedure is implemented using 2% of the data while the final model fit is conducted using the remaining 98%. We find that site-specific conservation (phyloP), replication timing and expression level are the best predictors for the mutation rate. In general, the framework allows formal testing for inclusion of explanatory variables and also the interaction terms. The impact of different explanatory variables can be inferred from the parameter estimates of our final model as the multiplicative changes in mutation rate. Our model also confirms some known mutational signatures and can be used as the null model for cancer driver detection.

2 Results

2.1 Heterogeneity of the mutation rate

We observed significant heterogeneities of the mutation rate at multiple levels (Figure 2). Mutation rate changes among cancer types (Figure 2a). Skin cutaneous melanoma, colorectal cancer and lung adenocarcinoma are the cancer types with the highest mean mutation rates ($5-10$ mut/patient/Mb) while thyroid carcinoma, prostate adenocarcinoma and low-grade glioma are the lowest ($0.5-1$ mut/patient/Mb). Mutation rate also varies for samples from the same cancer type, with the largest variation seen for skin cutaneous melanoma (the mutation rate ranges from 1 mut/patient/Mb to 150 mut/patient/Mb).

Mutation rate varies for different genomic contexts (Figure 2b). As previously shown, we find that mutational signatures are cancer type specific by looking into the mutation rate for different trinucleotide contexts. Mutation rate at TT sites is significantly higher in skin cutaneous melanoma than in any other cancer types, and the mutation rates at CpG sites are elevated for all the cancer types. In colorectal cancer, we observe more mutations at TCG sites and in glioblastoma more mutations at ACG sites.

Mutation rate also changes for different genomic environments defined by the explanatory variables (Figure 2b, 2c). Coding regions tend to have fewer mutations for all cancer types. Mutation rates are elevated for simple repeat regions, which might be related to technical reasons during mutation calling. The effect of CpG islands varies for different cancer types. The mutation rate for CpG islands are significantly higher than in regions outside for thyroid carcinoma, prostate adenocarcinoma, low-grade glioma and kidney chromophobe, while for colorectal cancer, lung adenocarcinoma, lung squamous cell carcinoma and skin cutaneous melanoma the situation is reversed. Regions that are late replicated, GC rich, evolutionarily less conserved, inside DNase1 peaks and lowly expressed have an elevated mutation rate. The importance of the explanatory variables varies across cancer types as shown by the regression lines in Figure 2c.

2.2 Granularity of regression models

We model the mutation probability in cancer genomes using regression models. The most coarse-grained description of the number of mutations in a region is a Poisson regression model and the most fine-grained is a binomial or multinomial site-specific regression model. Here we describe and investigate in detail the (dis)advantages between these three models. The concepts of the models are shown in Figure 3a.

Poisson count regression model In Poisson regression, the number of mutations in a genomic region of fixed length is modeled. The whole genome is divided into regions of pre-fixed length or according to the value of explanatory variables (e.g. segmented by genomic element types). Regression modeling is facilitated by summing both the mutation counts and the annotations over the region.

We model the mutation count $N_{r,sam}$ in the r-th genomic region with length L_r in sample sam. Furthermore, can is the cancer type of the sample. Mutations arise randomly with probability $p_{r,sam}$. As $p_{r,sam}$ is small and L_r is large, we have the Poisson approximation of the binomially distributed number of mutations

$$N_{r,sam} \sim \mathrm{Bi}(L_r, p_{r,sam}) \approx \mathrm{Po}(L_r p_{r,sam}).$$

Assuming J explanatory variables, the expected mutation count $\lambda_{r,sam} = L_r p_{r,sam}$ can be written with a Poisson regression model:

$$\log \lambda_{r,sam} = \mu_{sam} + \beta_{1,can} x_{r,1} + \cdots + \beta_{J,can} x_{r,J}$$

where for the jth explanatory variable, $x_{r,j}$ is the average value of the annotations across the region r if the explanatory variable is continuous and for categorical explanatory variables, $x_{r,j}$ is derived from the proportions of different levels of the annotations in region r, $j = 1, \cdots, J$.

Site-specific binomial regression model In site-specific regression models, the mutation probability is modeled in each position of the genome. We enable regression modeling by binning the continuous annotations, such that we are able to sum mutation counts over positions with the same combination of annotations, and thereby reduce the size of the data set. We consider both site-specific binomial and multinomial regression models.

We model the mutation probability $p_{i,sam}$ at a site i in sample sam of cancer type can. With a logit link, the mutation probability can be written with a logistic regression model:

$$\log \frac{p_{i,sam}}{1 - p_{i,sam}} = \mu_{sam} + \beta_{1,can} x_{i,1} + \cdots + \beta_{J,can} x_{i,J}$$

where $x_{i,j}$ is the value of the jth explanatory variable at site i.

Site-specific multinomial regression model for strand-symmetric mutation types We model the mutation probability for different mutation types. Assuming strand-symmetry, we are not distinguishing between e.g. A>G mutations and T>C mutations, $p^{A>G} = p^{T>C}$. We consider the strand with the C or T nucleotide, and the mutation probability matrix is

$$\begin{array}{c} \\ \mathbf{A} \\ \mathbf{G} \\ \mathbf{C} \\ \mathbf{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathbf{A} & \mathbf{G} & \mathbf{C} & \mathbf{T} \end{array} \\ \begin{pmatrix} p^{T>T} & p^{T>C} & p^{T>G} & p^{T>A} \\ p^{C>T} & p^{C>C} & p^{C>G} & p^{C>A} \\ p^{C>A} & p^{C>G} & p^{C>C} & p^{C>T} \\ p^{T>A} & p^{T>G} & p^{T>C} & p^{T>T} \end{pmatrix} \end{array}$$

with only 6 types of mutations.

We model these mutation probabilities by setting up two independent multinomial logistic regression models, one for mutations from (G:C) base pairs and one for mutations from (A:T) base pairs. The (G:C) basepairs are modelled with probabilities

$$\left(p_{i,\text{sam}}^{C>A}, p_{i,\text{sam}}^{C>G}, p_{i,\text{sam}}^{C>C}, p_{i,\text{sam}}^{C>T}\right)$$

for (G:C) position i in sample *sam*. With J explanatory variables, the mutation probability at G:C position i in sample *sam* of cancer type *can* can be written as:

$$\begin{aligned}
\log \frac{p_{i,\text{sam}}^{C>A}}{p_{i,\text{sam}}^{C>C}} &= \mu_{\text{sam}}^{C>A} + \beta_{1,\text{can}}^{C>A} x_{i,1} + \ldots + \beta_{k,\text{can}}^{C>A} x_{i,J} \\
\log \frac{p_{i,\text{sam}}^{C>G}}{p_{i,\text{sam}}^{C>C}} &= \mu_{\text{sam}}^{C>G} + \beta_{1,\text{can}}^{C>G} x_{i,1} + \ldots + \beta_{k,\text{can}}^{C>G} x_{i,J} \\
\log \frac{p_{i,\text{sam}}^{C>T}}{p_{i,\text{sam}}^{C>C}} &= \mu_{\text{sam}}^{C>T} + \beta_{1,\text{can}}^{C>T} x_{i,1} + \ldots + \beta_{k,\text{can}}^{C>T} x_{i,J}.
\end{aligned}$$

Note that the probability for no mutation $p_{i,sam}^{C>C}$ is the reference. Similarly, (A:T) basepairs are modeled with probabilities

$$\left(p_{i,\text{sam}}^{T>A}, p_{i,\text{sam}}^{T>G}, p_{i,\text{sam}}^{T>C}, p_{i,\text{sam}}^{T>T}\right)$$

and the reference is the probability for no mutation $p_{i,sam}^{T>T}$. We compare the performance of the three models on 2% of the whole genome data. The setting for the three models is shown in Figure 3a. Each model is

trained with replication timing, phyloP, and context information from the reference genome. For the region-based Poisson model, continuous annotation values are averaged over the selected region. GC percentages are calculated for each region. For the site-specific models, we use the site-specific annotations for each site. Continuous values are discretized to simplify the estimation process.

The results are shown in Figure 3b. When predicting mutation counts in large windows (100 kb), the three regression models perform similarly. For prediction in a randomly selected small number of sites (1 kb), the two site-specific models out-perform the region-based model (Figure 3c). The site-specific models can capture the mutational heterogeneities between sites and provide a more accurate mutation probability in any resolution. This is in contrast to the region-based model where a large number of sites are required for accurate predictions. In addition to predicting the probability for mutation events, the multinomial regression model can also predict the mutation probabilities for different mutation types (Figure 3d, 3e).

2.3 Model selection

We consider the site-specific multinomial regression model to predict mutation probabilities for different mutation types at a single site. We implement a forward model selection procedure to determine the explanatory variables in the final model (Figure 1). In each step, we add all possible new variables to the previous model in turn and rank the resulting new models. We identify and include the explanatory variable with the best fit and iterate the procedure several times. Cross-validation is used to assess the fit of a model while avoiding overfitting. The fit improves along the model selection procedure. The McFadden's pseudo R^2 increases and deviance loss calculated from cross validation decreases. We stop the selection procedure when the fit is not improving by adding more explanatory variables to the model (Figure 4a).

As a reference model, we start out with a single mutation rate for the whole genome in all samples (Model 1). This model cannot explain any of the variation in the mutation rate between samples and positions, so McFadden's pseudo $R^2 = 0$. After including the six strand-symmetric mutation types in the model (Model 2), we add cancer and sample specific intercepts (Model 3 and Model 4) to make sure that we account for sample-specific mutational signatures. In the next step, we include the left and right neighboring base-pair for each cancer type(Model 5). We observe an increase in McFadden's pseudo R^2 and a decrease in the deviance loss obtained by cross-validation.

Starting with Model 5, additional annotations are added using forward model selection procedure. We consider the phyloP score, replication timing, expression, genomic segments, GC content in 1 kb, CpG islands, simple repeats and DNase1 hypersensitivity. For each of these variables, cancer specific regression coefficients are estimated to allow for differences in the mutational process between cancer types. For expression, we

use data directly obtained from matching tumor types, so we also take expression differences between cancer types into account.

The annotation with the largest decrease in the deviance loss function is the phyloP score (Model 6). Subsequently, replication timing (Model 7) and expression (Model 8) are added. Detailed results for each step of the forward selection procedure are provided in the Appendix.

While the trinucleotide context and the phyloP score vary on a basepair scale, both replication timing and expression vary on kilo-base scale. Even though much of the per-base-pair variation in the mutation rate is already captured in Models 5 and Model 6, the long-range variation is considerably better explained in Model 7 and Model 8, as illustrated in Figure 4b.

We can see that adding replication timing in Model 7 considerably changes the predicted mutation rate obtained from Model 6 that is relatively uniform in the 10 kb scale shown in the figure, by taking the replication timing gradient in the region into account. Finally, in Model 8, the lower mutation rate in highly expressed regions is taken into account, which again improves the prediction.

2.4 Estimation results

Upon determining the final model from the model selection procedure, we estimate parameters for the multinomial logistic regression model on the remaining 98% genomic regions. To study the difference between cancer types, all position-specific explanatory variables are stratified by cancer type. The coefficients are interpreted as multiplicative changes in mutation rate. Our results confirm the large differences in mutation pattern both between samples and cancer types, but also between different genomic and epigenomic regions.

We observe that the association of replication timing and mutation rate varies. As previously described [Lawrence et al., 2013], we always find a positive association (see regression lines in Figure 2c; later replicating regions have more mutations), but the regression coefficient varies significantly between the different cancer types (Figure 5a). This can either point to different mutational mechanisms, or it shows that the replication timing dataset at use (Chen et al., 2010a; measured on HeLa cell lines) is not equally representative for different types of cancer tissues. We are currently exploring replication timing measured on different cell lines.

We find that the mutation rates differ between the different types of mutations, but also between the specific contexts that we consider: CpG, to capture the pattern of spontaneous deamination, and TpCpW, to capture the APOBEC signature. We find that the $C > T$ mutation rate is higher in CpG sites than in other sites in all cancer types except in skin cutaneous melanoma (Figure 5b), which is related to the elevated $TT > CC$ mutation rate in skin cutaneous melanoma due to UV light signature. We observe

elevated APOBEC mutation rates in breast cancer, bladder urothelial carcinoma, head and neck squamous cell carcinoma, lung adenocarcinoma, lung squamous cell carcinoma, thyroid carcinoma and uterine corpus endometrial carcinoma.

3 Discussion

We use a multinomial logistic regression model to analyse the somatic mutation rate in each position of a cancer genome. We consider various genomic features, such as the local base composition, the functional impact of a region and replication timing. Because of the site-specific formulation, the model is the most fine-grained description of the mutation rate.

The parameters can be interpreted as multiplicative mutation rate changes. By using interaction terms, it is straightforward to analyse and test differences in mutation rate between cancer types, samples or specific genomic regions of interest. Furthermore, generalized linear models come with a natural framework for hypothesis testing. It would be interesting to compare our model-based description of the mutation rate to unsupervised learning of mutational signatures by matrix factorization [Alexandrov et al., 2013a].

In the logistic regression model, we can estimate cancer type specific regression coefficients to capture the differences between cancer types. We can also include tissue specific explanatory variables if there are measurements for the corresponding cancer tissue or matching healthy tissue available. This is of particular interest for epigentic measurements like histone modifications, where the tissue of origin has been shown to be informative [Polak et al., 2015].

Patient-specific characteristics can also be added as explanatory variables. The age of the patients could be used to study clock-like mutational processes [Alexandrov et al., 2015]. Known somatic or germline mutations that are associated with specific mutational processes or repair pathways can also be used as explanatory variables, e. g. a germline deletion of *APOBEC3B* that fuses APOBEC3A with the 3' UTR of APOBEC3B has been found to be associated with an increased number of APOBEC-type mutations [Nik-Zainal et al., 2014]. The impact of this mutation could be studied by including an interaction term with the TpCpW positions.

Our model is very flexible and versatile and the explanatory variables can be customized according to different applications. An important application of the model is the prediction of the somatic mutation probability under the assumption of neutrality. In our driver detection method ncdDetect [Rasmussen et al., 2017], we use the position-specific predictions that we obtain from the multinomial model to evaluate if the mutation rates or their distributions are significantly different from the expectation under neutrality. This allows a flexible analysis of regions of any size. Even non-contiguous regions with very different properties

than the overall genomic patterns can be investigated.

4 Methods

4.1 Data

4.1.1 Somatic mutation dataset

We use SNV calls from 505 tumor-normal samples across 14 different cancer types from Fredriksson et al. [2014b]. We build our data set based on the UCSC hg19 assembly. We removed regions with low mappability, ultra-high mutation rates and lacking annotation. Problematic regions for NGS alignments identified for the ENCODE project (Bernstein et al. [2012]) were subtracted. Low mappability regions, defined by the GEM tool (Derrien et al. [2012]) and CRG Alignability track from UCSC with mappability less than 0.5 in 100-mers, were also subtracted. Hyper-mutated genomic segments containing Immunoglobulin/T-cell receptor (IG/TR) genes defined by GENCODE, were excluded from analysis. We also excluded sites on ChrX and ChrY, because some of the annotation files we use lack information for one or both of the sex chromosomes.

A total number of $14,200,393$ SNVs are observed in the subtracted regions for 505 samples across 14 cancer types.

Genomic annotation We divided the genome into six genomic annotations: coding, 5' UTR, 3'UTR, ncRNA, intron and intergenic. Based on the GENCODE v.19 transcripts, coding regions, 5' UTR regions and 3' UTR regions as well as introns were defined for protein-coding transcripts. Non-coding RNA regions were defined as all remaining regions in the full transcript set. All remaining bases were categorized as intergenic.

GC content We calculate the percentage of G:C base pairs (strong sites) in 1 kb windows based on the reference genome. Regions with GC percentage less than 10% are annotated with value 0. Other regions are discretized into quartiles.

CpG islands We segmented the genome by presence or absence of CpG islands. The CGI Mountain annotation from `http://cgihunter.bioinf.mpi-inf.mpg.de/anno_wrapper.php` was used. The CGI Mountain score quantifies if a region is a CpG island. Scores above zero indicate a CpG island. We use a dummy variable derived from the CGI Mountain score in our analysis indicating whether the CGI Mountain score is larger than zero.

Simple repeats We annotated the simple repeat regions in the genome according to RepeatMasker (`http://www.repeatmasker.org`), which defines the interspersed and low-complexity repeats in hg19. We use a dummy variable to indicate whether a genomic site is in a region masked as simple repeats.

DNase I peaks We defined DNase I peaks according to the DNAse I annotation HoneyBadger2 from the Roadmap Epigenomics project (`http://www.broadinstitute.org/~meuleman/reg2map/`). We use the score from HoneyBadger2 to indicate the DNase I signal strength for regions with a DNase I peak. The regions not annotated in the HoneyBadger2 were masked as "no peak" regions in our analysis.

PhyloP score The conservation score phyloP (phylogenetic p-values) is part of the PHAST package (`http://compgen.bscb.cornell.edu/phast/`). We used the score from the multiple alignments of 99 vertebrate genomes to the human genome [Pollard et al., 2010]. We use the version of phyloP100way which covers 99.8% of the subtracted regions.

Replication timing We adjust the replication timing data from Chen et al. [2010b] to the hg19 assembly. The replication timing value ranges from 0 to 1, indicating the earlier to later stages in replication process. Annotation of replication timing covers 91.2% of the subtracted regions.

Gene expression We define the gene expression level according to TCGA RNAseq expression data. Expression data was $\log_2(x+1)$ transformed. For each cancer type, the median expression was calculated for all genes. If multiple annotations of a gene exist, the longest annotation is used. For overlapping genes, the expression is a cumulative sum.

As in Fredriksson et al. [2014a], we collapse colon (COAD) and rectal carcinoma (READ) to a joint cancer type CRC by averaging across the expression values.

4.1.2 Preparation of the analytical data table

In order to facilitate the model fitting procedure, we summarized the genomic data into count tables. In Poisson count models, each row in the count table represents a pre-defined region. For continuous explanatory variables, such as replication timing, the annotations are averaged over the region. For categorical explanatory variables, the annotations are transformed to the percentage for different levels of the variables, e.g, the binary explanatory variable indicating whether the site is a (G:C) base pair or not is transformed to GC content of the window. In site-specific regression models, we discretized continuous variables into bins according to quartiles, quintiles or quantiles. Each row in the count table represents the counts of mutations under a certain combination of levels of all the explanatory variables. As we are summarizing the whole genome in

the count table, we expect to see all the combinations of the levels for all the explanatory variables. Thus, the sizes of the count tables grow significantly when adding new explanatory variables. The generation of the count tables also takes much longer time with more explanatory variables. Because of the space and time consumption, a large count table including all the explanatory variables is computationally infeasible. We implemented the forward model selection procedure to avoid many huge count tables. For each iteration in the forward model selection procedure, we made new count tables only for the selected sets of explanatory variables from previous step and one new candidate explanatory variable. We then added the best candidate in the explanatory variable set and repeated for the next step. We used 2% of the whole genome in the model selection and made a final count table with the remaining 98% sites only for the important explanatory variables that were determined from the model selection procedure. The generation of the count table for the final model takes $6,000$ CPU hours on our cluster (2.5GHz CPUs).

4.2 Multinomial regression model

4.2.1 Estimation

The observations in the regression model are indexed by the genomic position $i(1 \leq i \leq 2.56 \cdot 10^9)$ for all positions on chr1 to chr22 after excluding problematic regions and the samples $sam(1 \leq sam \leq 505)$, so the total number of observations is $1.3 \cdot 10^{12}$. The overall, sample specific intercepts for the six mutation types sum up to $6 \cdot 505$ parameters. The regression coefficients for explanatory variables are indexed by the 14 cancer type, which sum up to 574. Thus, in total, 3604 parameters are estimated in this model.

To reduce memory usage and computation time, the parameters are estimated in three separate binary logistic regression models, but the variance-covariance matrix of the parameters is estimated for the multinomial model [Begg and Gray, 1984]. Estimation is conducted in R [R Core Team, 2014] using the function `glm4` from the contributed package `MatrixModels` [Bates and Maechler, 2015b] and the estimation of the variance-covariance matrix is implemented using the package `Matrix` [Bates and Maechler, 2015a].

4.2.2 Dirichlet prior and pseudo counts

If a model with many explanatory variables and interaction terms among them is estimated (e. g. sampleID $\times$ neighbors $\times$ strong), it can easily occur that for a certain combination of levels of categorical variables, there have been no mutations of a certain type observed. This case is especially likely, if the sampleID is involved. This causes numerical problems in the maximum likelihood estimation [Agresti, 2002].

To solve this problem while still obtaining a positive mutation probability, we added pseudo counts to the observed mutation counts. This is equivalent to specifying a Dirichlet prior for the multinomial model

[Durbin et al., 1998]. We obtain the pseudo counts from the mutation counts of the corresponding cancer type for the exact same combination of categorical variables. The mutation count from the sample and from the cancer type are equally weighted.

4.2.3 McFadden's pseudo R^2

To assess the fit of a model, we report McFadden's pseudo R^2

$$R^2_{\text{McFadden}} = 1 - \frac{\log L_M}{\log L_0}$$

where L_M is the likelihood of the model under investigation and L_0 is the likelihood of a model with no predictors [McFadden, 1974]. To measure the improvement of a model in comparison with the binomial model where there is no distinction between the mutation types, we use the same mutation probability for each mutation type in the model without predictors.

4.3 Forward variable selection

To speed up data preparation and estimation, variable selection is conducted on a subset of approximately 2% of the genome. It is constructed by randomly selecting $60,000$ windows of size 1 kb. The cancer types kidney chromophobe, low-grade glioma, prostate adenocarcinoma and thyroid carcinoma have very low mutation counts, so they are disregarded during variable selection. We used both FcFadden's pseudo R^2 and cross-validation for forward variable selection.

Starting with Model 5,

$$\text{logit}(p^{\text{mut. type}}_{i,sam,can}) = \mu^{\text{mut. type}}_{sam} + \beta^{\text{mut. type}}_{\text{context},can} x_{\text{context},i}$$

additional terms of the form

$$\beta^{\text{mut. type}}_{\text{k},can} x_{\text{k},i}$$

with explanatory variable $\text{k} \in \{$phyloP, replication timing, genomic segment, expression, GC content, DNase1, simple repeats, CpG island$\}$, are selected following the forward selection scheme.

4.3.1 Cross validation

An alternative assessment of the fit of a model can be obtained by cross validation. Five-fold cross validation is used to select the annotation with the highest explanatory power. To this end, the 1 kb windows of the variable selection subset are divided randomly distributed among 5 sets. In turn, 4 of them are joined as the

training set, on which the multinomial model is estimated, and the remaining set is used as the validation set to estimate the loss function.

The deviance loss function for the observed site i in sample sam is defined as

$$D_{i,sam} = -2 \sum_{k} \mathbb{1}(y_{i,s} = k) \log \hat{p}(y_{i,s} = k)$$

where k denotes all possible mutation events at site i, $y_{i,sam}$ the observed event at site i in sample sam and $\hat{p}(y_{i,sam} = k)$ the probability of event k estimated under the multinomial regression model [Hastie et al., 2001]. Thus, it measures the prediction accuracy of the multinomial regression model. The total deviance loss of a model is

$$D = \sum_{i=1}^{N} \sum_{sam=1}^{505} D_{i,sam}$$

with N is the number of genomic positions.

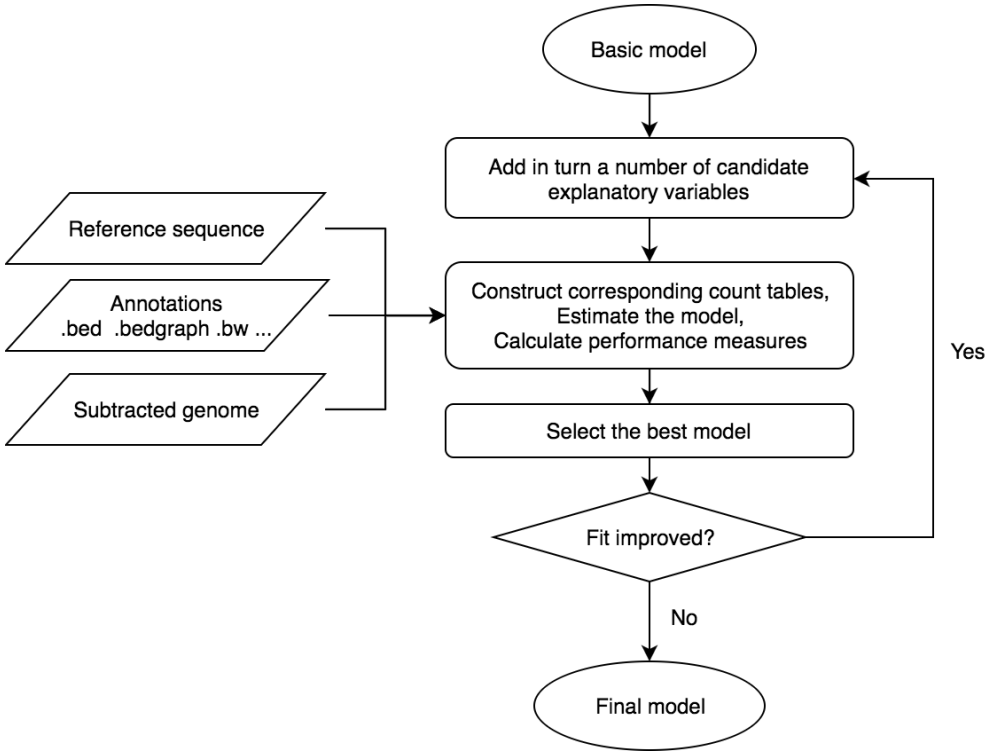
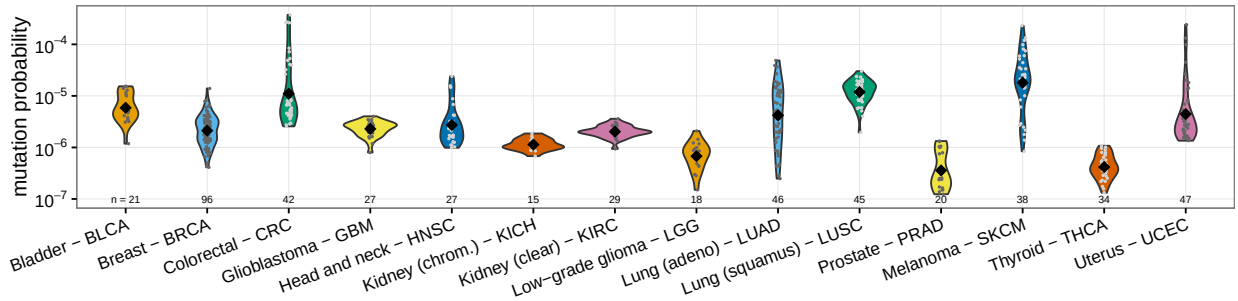
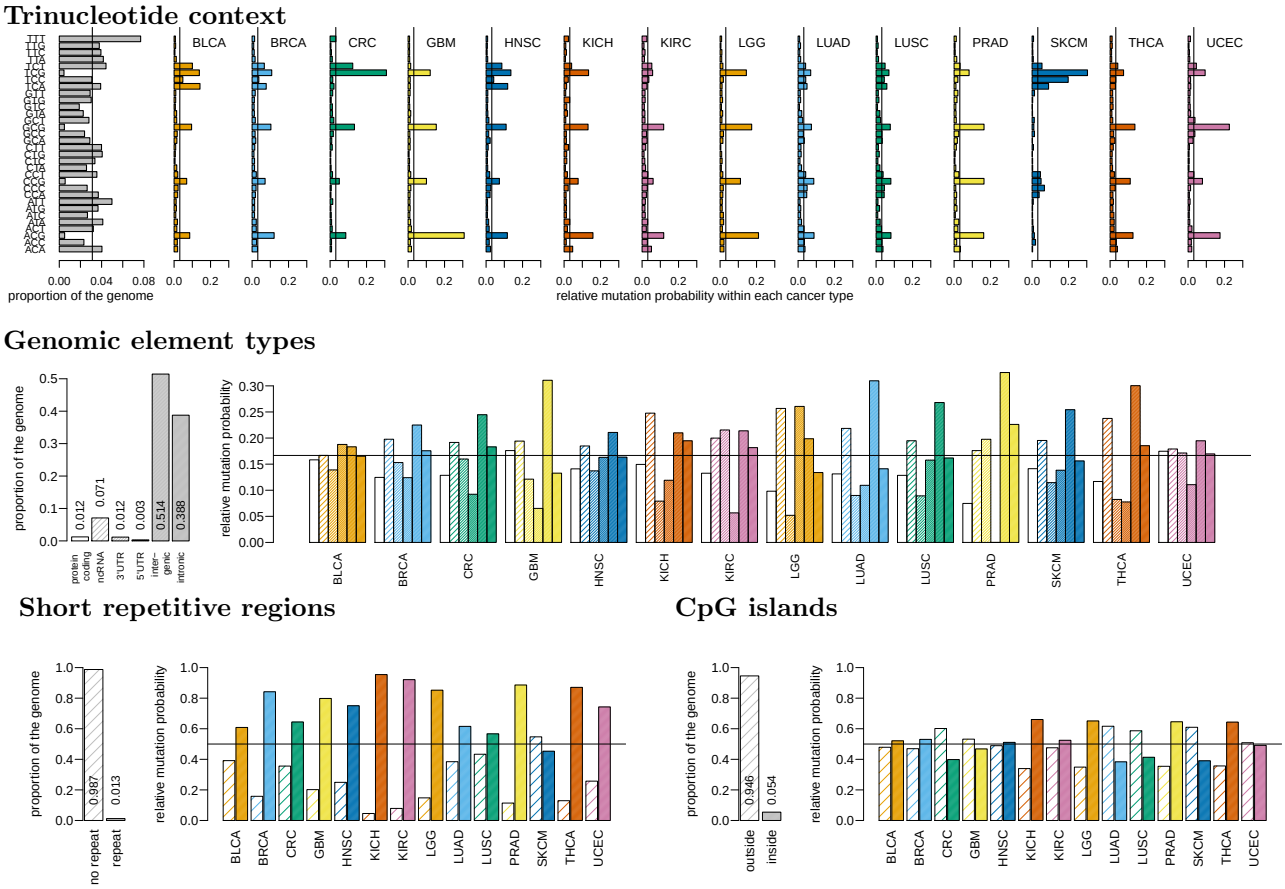


Figure 1: Workflow of the forward model selection procedure. The forward model selection is implemented on 2% of the data to determine the explanatory variables included in the final model. In each iteration of the model selection, data tables are generated to summarize the site-specific annotations. The performance of the models are measured with McFadden's pseudo R^2 and deviance loss obtained by cross-validation. The explanatory variable with the best performance is included in the parameter set for the next iteration. The model selection procedure terminates when the performance of the new models is not improving anymore. Parameter estimation for the final model is based on the remaining 98% of the data.

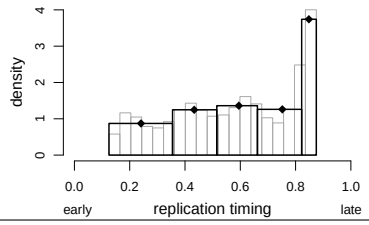
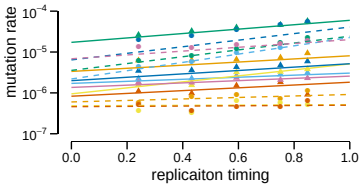
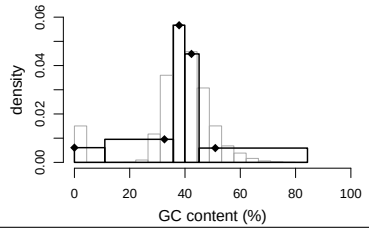
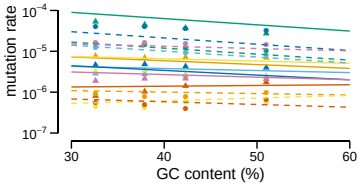
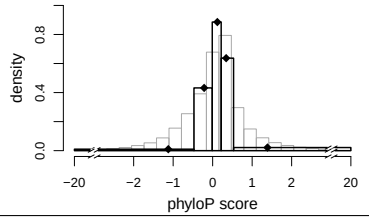
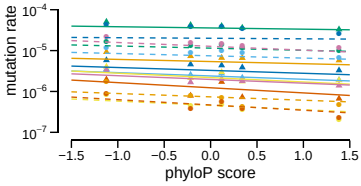
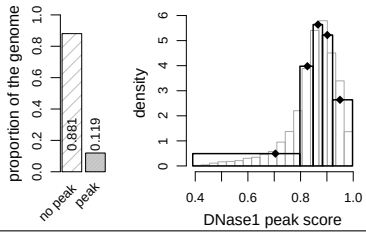
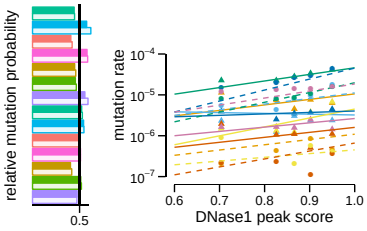
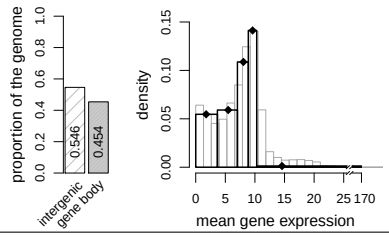
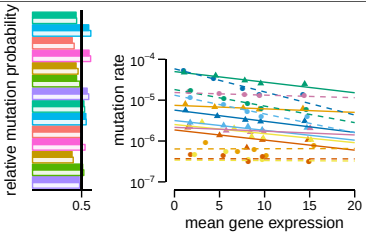


(a) Heterogeneity among cancer types and samples. Violin plot for the mutation probability for 14 cancer types.

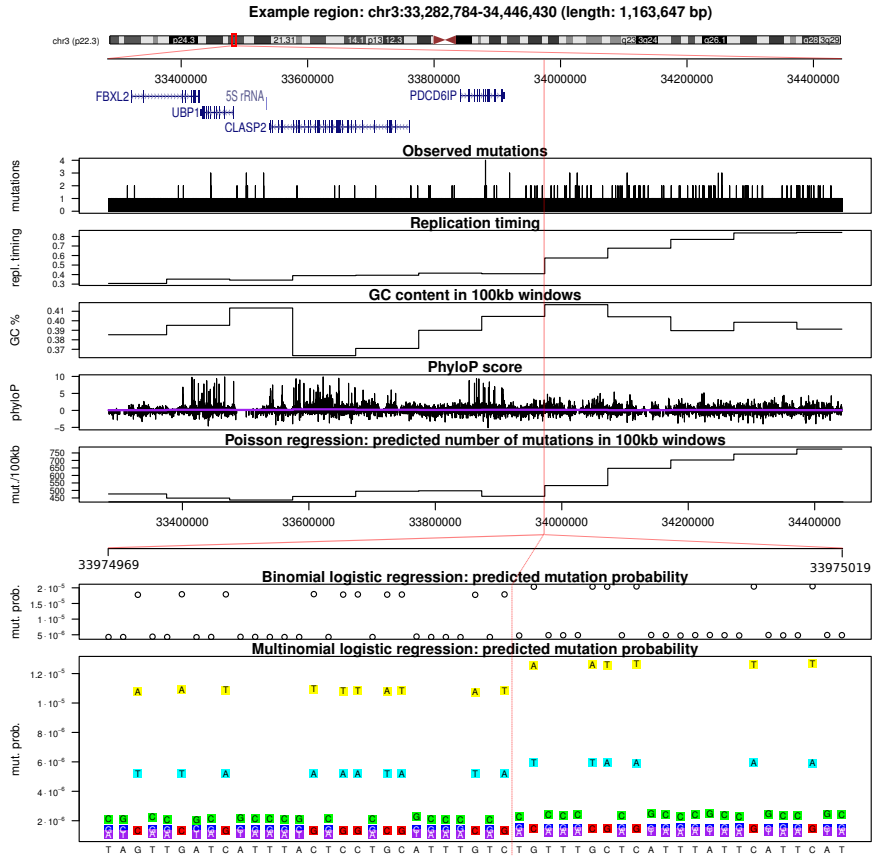


(b) Heterogeneity along the genome and the correlation with categorical explanatory variables. Relative mutation probability from nucleotide C or T in the neighboring context A,G,C,T ($2 \cdot 4 \cdot 4 = 32$ possibilities), relative mutation probability of six different genomic elements, and relative mutation probability within and outside repeat regions or CpG islands.

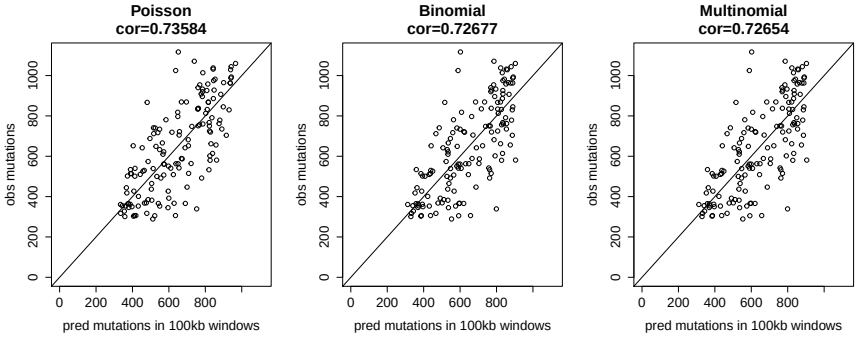
Figure 2: Heterogeneity of the mutation rate and explanatory variables. (continue to next page)

Annotation	distribution along the genome	mutation distribution
Replication timing Chen et al. [2010a] *quintile means*		
GC content $GC\% \leq 10\% \Rightarrow 0$ $GC\% > 10\% \Rightarrow$ *quartile means*		
phyloP score [Pollard et al., 2010] *quintile means*		
DNase1 peaks [Roadmap Epigenomics Consortium, 2015] *quintile means*		
Gene expression [The Cancer Genome Atlas Network, 2012] *quintile means*		

(c) Heterogeneity correlated with continuous variables. Left column: continous variables. Middle column: The continuous annotations are discretized into bins according to quartiles, quintiles or quantiles for site-specific regression models. Each bin is represented by the mean value within the bin. Grey histograms: proportion of the genome covered by the annotation DNase1 peaks and gene expression, respectively. Grey transparent histograms: distribution of the continuous values of the annotation along the genome. Black transparent histograms: distribution of the discrete bins of the annotation (binning scheme in italics in the column "Annotation"). Black diamonds: Discrete value used for the binning. Right column: Predicted (solid or dashed lines) and observed (points) mtuation rate for each cancer type (using the same colour scheme as in (a)) and explanatory variables. The regression lines are generated under a multinomial logistic regression model using only the correspoding explantory variable.

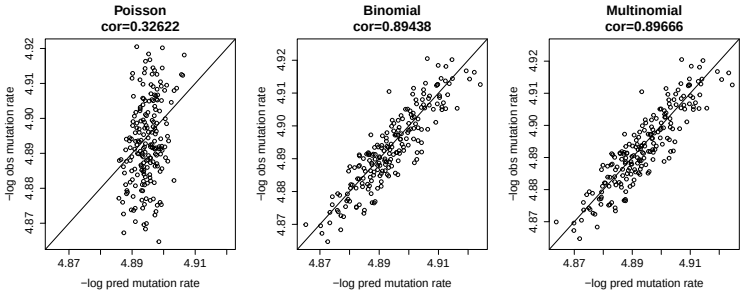


(a) Motivation (site-specificity) and conceptual explanation of the different models. Consider a 1.2 Mb region on Chromosome 3. We observe a number of mutations and the value of the explanatory variables replication timing, GC content and phyloP score. Given the values of the explanatory variables we use Poisson, site-specific binomial logistic regression or site-specific multinomial logistic regression to predict the number of mutations in a region (Poisson), the probability of a mutation in a single site (binomial) or even the probability of the three types of mutation in a single site (multinomial).

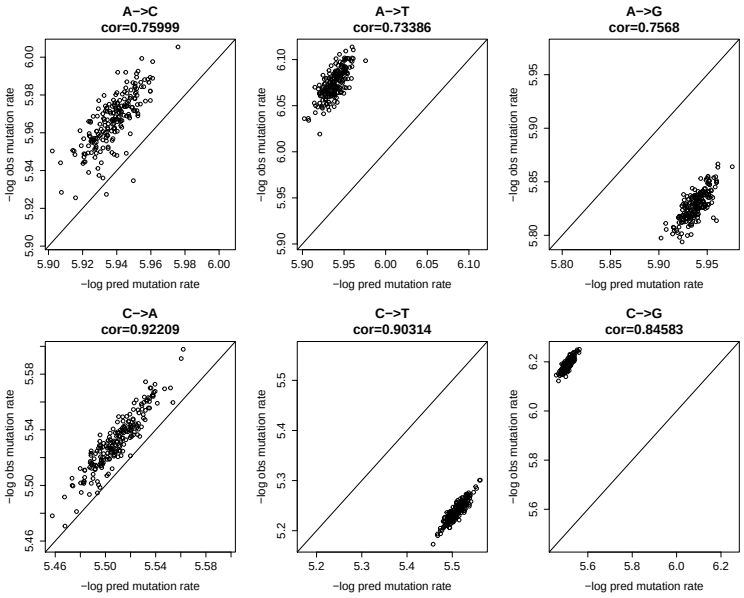


(b) Predicted versus observed number of mutatios for the three models for 100 kb regions.

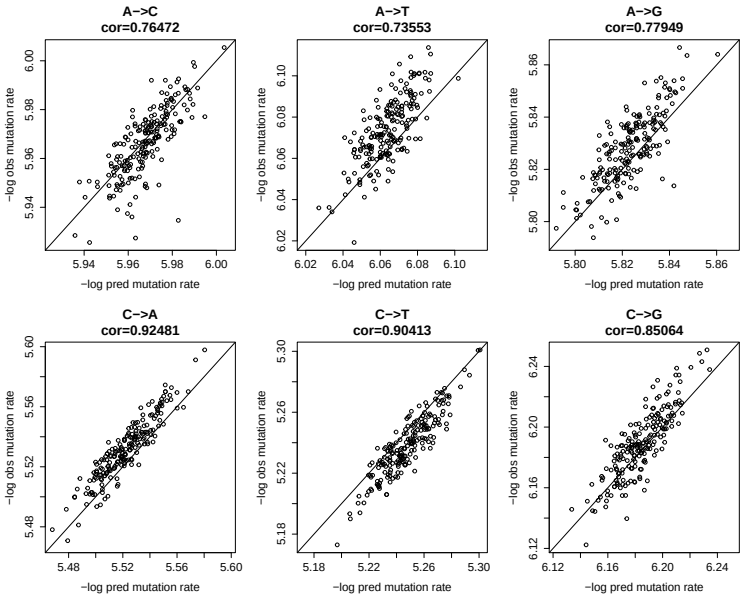
Figure 3: Comparison of Poisson regression model, site-specific binomial logistic regression model and site-specific multinomial logistic regression model.(continue to next page)



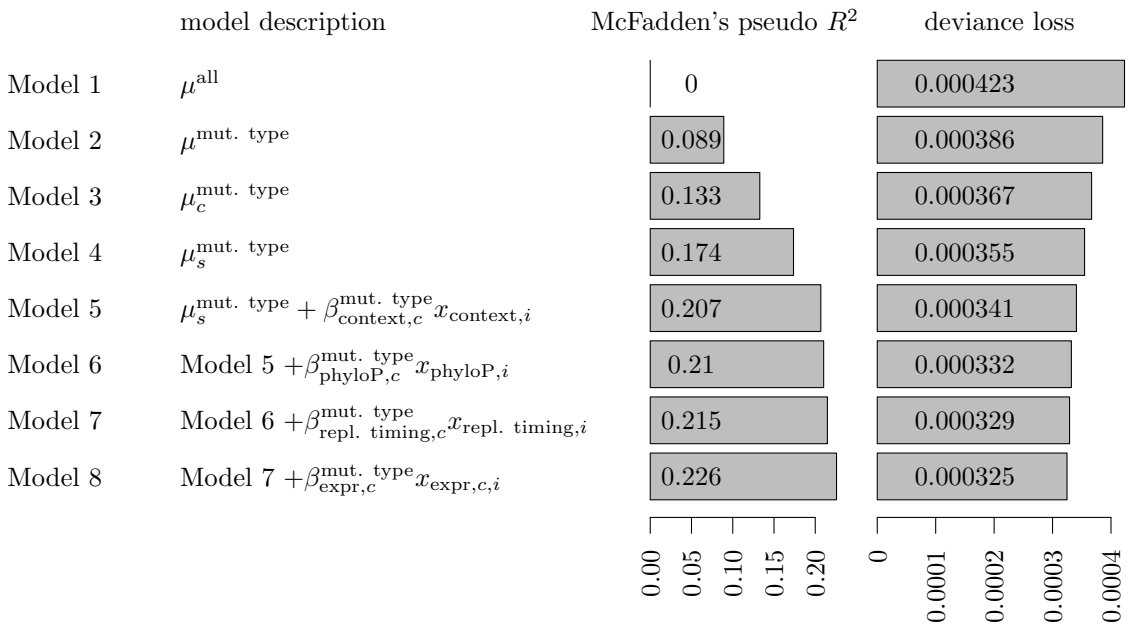
(c) Site-specific models perform substantially better in 1000 randomly selected sites.



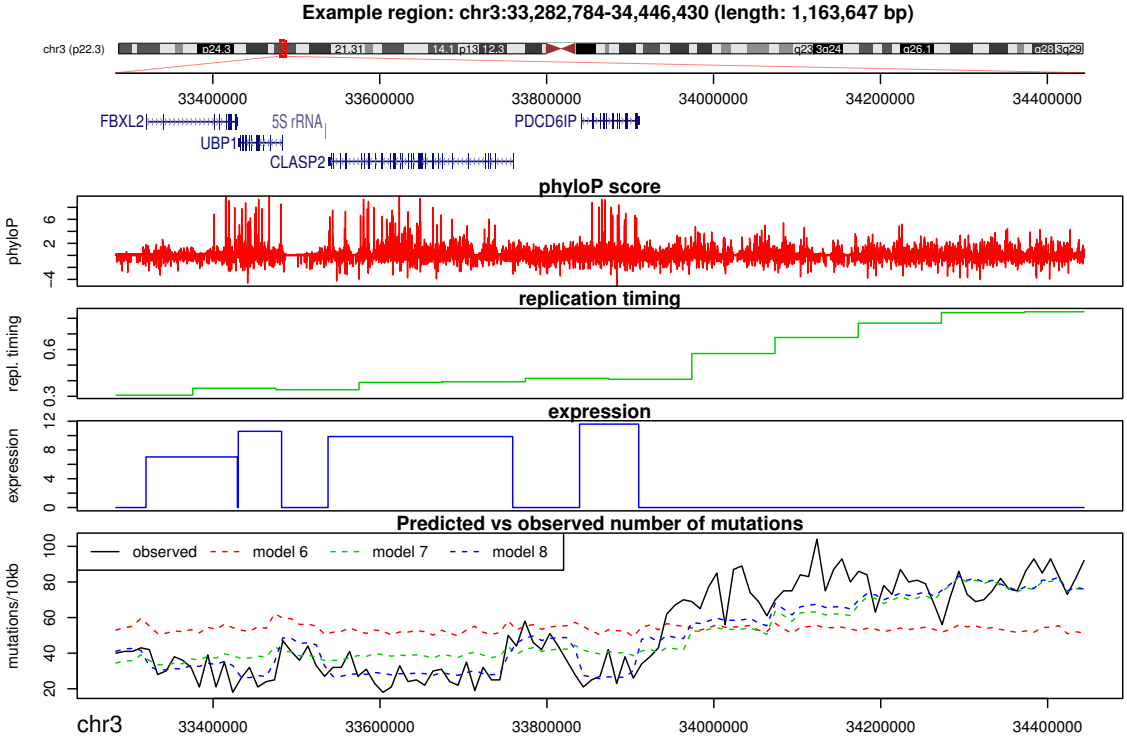
(d) The prediction for different mutation types with binomial logistic regression model in 1000 randomly selected sites.



(e) The prediction for different mutation types with multinomial logistic regression model in 1000 randomly selected sites.

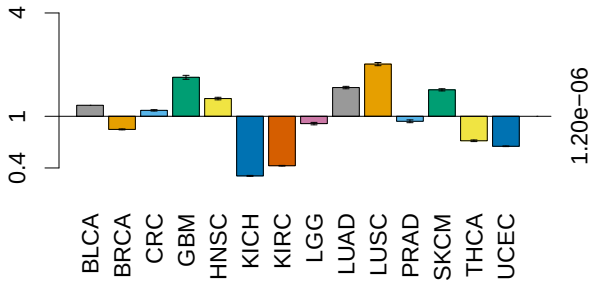


(a) The fit improves during the forward model selection procedure. We use McFadden's pseudo R^2 and deviance loss from cross validation to determine which explanatory variables to include in the final model.

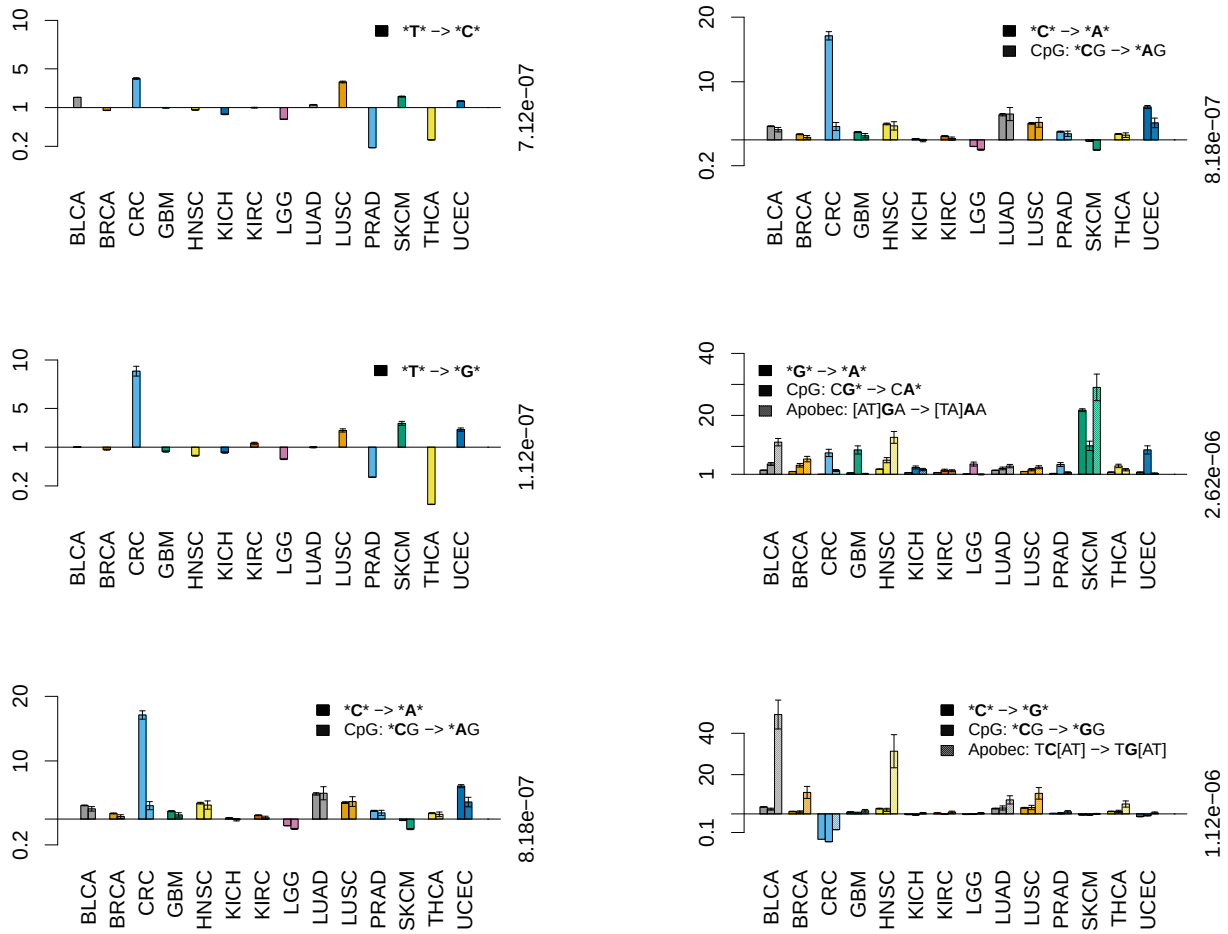


(b) Fit along the genome

Figure 4: Model selection results. Note that the observed number of mutations is better predicted for the more complex models.



(a) The impact of replication timing inferred from parameter estimation as multiplicative change in mutation rate for different cancer types. The horizontal line indicates the average mutation rate across all the cancer types. The height of the bars shows the multiplicative change of the mutation rate for each cancer type.



(b) Multiplicative changes in mutation rate for different mutation signatures. The horizontal line indicates the average mutation rate for a perticular mutation type across all the cancer types. The specific contexts that we consider are CpG and TpCpW.

Figure 5: Parameter estimation results.

Appendix: Forward model selection results

In this appendix, we provide the details for the forward model selection procedure.

Step 1

Model	deviance loss	McFadden's pseudo R^2
Model 5 $+\beta^{\text{mut. type}}_{\text{phyloP},c} x_{\text{phyloP},i}$	0.0003322	0.2102
Model 5 $+\beta^{\text{mut. type}}_{\text{repl. timing},c} x_{\text{repl. timing},i}$	0.0003359	0.2124
Model 5 $+\beta^{\text{mut. type}}_{\text{element},c} x_{\text{element},i}$	0.0003391	0.2089
Model 5 $+\beta^{\text{mut. type}}_{\text{expr.},c} x_{\text{expr.},c,i}$	0.0003395	0.2100
Model 5 $+\beta^{\text{mut. type}}_{\text{GC cont.},c} x_{\text{GC cont.},i}$	0.0003400	0.2087
Model 5 $+\beta^{\text{mut. type}}_{\text{DNase1},c} x_{\text{DNase1},i}$	0.0003405	0.2076
Model 5 $+\beta^{\text{mut. type}}_{\text{repeat},c} x_{\text{repeat},i}$	0.0003406	0.2075
Model 5 $+\beta^{\text{mut. type}}_{\text{CGI},c} x_{\text{CGI},i}$	0.0003406	0.2074

Table 1: Deviance loss and McFadden's pseudo R^2 for each of the models tested in step 1 to obtain model 6.

Step 2

Model	deviance loss	McFadden's pseudo R^2
Model 6 $+\beta^{\text{mut. type}}_{\text{repl. timing},c} x_{\text{repl. timing},i}$	0.0003291	0.2147
Model 6 $+\beta^{\text{mut. type}}_{\text{expr.},c} x_{\text{expr.},c,i}$	0.0003310	0.2132
Model 6 $+\beta^{\text{mut. type}}_{\text{GC cont},c} x_{\text{GC cont},i}$	0.0003315	0.2119
Model 6 $+\beta^{\text{mut. type}}_{\text{element},c} x_{\text{element},i}$	0.0003316	0.2117
Model 6 $+\beta^{\text{mut. type}}_{\text{DNase1},c} x_{\text{DNase1},i}$	0.0003319	0.2109
Model 6 $+\beta^{\text{mut. type}}_{\text{repeat},c} x_{\text{repeat},i}$	0.0003319	0.2108
Model 6 $+\beta^{\text{mut. type}}_{\text{CGI},c} x_{\text{CGI},i}$	0.0003320	0.2107

Table 2: Deviance loss and McFadden's pseudo R^2 for each of the models tested in step 2 to obtain model 7.

Step 3

Model	deviance loss	McFadden's pseudo R^2
Model 7 $+\beta^{\text{mut. type}}_{\text{expr.},c} x_{\text{expr.},c,i}$	0.0003248	0.2258
Model 7 $+\beta^{\text{mut. type}}_{\text{element},c} x_{\text{element},i}$	0.0003279	0.2154
Model 7 $+\beta^{\text{mut. type}}_{\text{repeat},c} x_{\text{repeat},i}$	0.0003289	0.2154
Model 7 $+\beta^{\text{mut. type}}_{\text{CG cont.},c} x_{\text{CG cont.},i}$	0.0003289	0.2152
Model 7 $+\beta^{\text{mut. type}}_{\text{DNase1},c} x_{\text{DNase1},i}$	0.0003290	0.2150
Model 7 $+\beta^{\text{mut. type}}_{\text{CGI},c} x_{\text{CGI},i}$	0.0003291	0.2150

Table 3: Deviance loss and McFadden's pseudo R^2 for each of the models tested in step 3 to obtain model 8.

References

Alan Agresti. *Categorical Data Analysis*. Wiley, 2002.

Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolo Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013a.

Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Samuel AJR Aparicio, Sam Behjati, Andrew V Biankin, Graham R Bignell, Niccolò Bolli, Ake Borg, Anne-Lise Børresen-Dale, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, 2013b.

Ludmil B. Alexandrov, Philip H. Jones, David C. Wedge, Julian E. Sale, Peter J. Campbell, Serena Nik-Zainal, and Michael R. Stratton. Clock-like mutational processes in human somatic cells. *Nature Genetics*, 2015.

Albino Bacolla, David N. Cooper, and Karen M. Vasquez. Mechanisms of base substitution mutagenesis in cancer genomes. *Genes (Basel)*, 5(1):108–146, 2014. doi: 10.3390/genes5010108. URL `http://dx.doi.org/10.3390/genes5010108`.

Douglas Bates and Martin Maechler. *Matrix: Sparse and Dense Matrix Classes and Methods*, 2015a. URL `http://CRAN.R-project.org/package=Matrix`. R package version 1.2-2.

Douglas Bates and Martin Maechler. *MatrixModels: Modelling with Sparse And Dense Matrices*, 2015b. URL `http://CRAN.R-project.org/package=MatrixModels`. R package version 0.4-1.

Colin B. Begg and Robert Gray. Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*, 71(1):11–18, 1984. doi: 10.1093/biomet/71.1.11. URL `http://biomet.oxfordjournals.org/content/71/1/11.abstract`.

BE Bernstein, E Birney, I Dunham, ED Green, C Gunter, and M Snyder. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489:57–74, 2012.

Chun-Long Chen, Aurélien Rappailles, Lauranne Duquenne, Maxime Huvet, Guillaume Guilbaud, Laurent Farinelli, Benjamin Audit, Yves d'Aubenton Carafa, Alain Arneodo, Olivier Hyrien, and Claude Thermes. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Research*, 20(4):447–457, 2010a. doi: 10.1101/gr.098947.109. URL `http://genome.cshlp.org/content/20/4/447.abstract`.

Chun-Long Chen, Aurélien Rappailles, Lauranne Duquenne, Maxime Huvet, Guillaume Guilbaud, Laurent Farinelli, Benjamin Audit, Yves d'Aubenton Carafa, Alain Arneodo, Olivier Hyrien, et al. Impact of replication timing on non-cpg and cpg substitution rates in mammalian genomes. *Genome research*, 20 (4):447–457, 2010b.

Thomas Derrien, Jordi Estellé, Santiago Marco Sola, David G Knowles, Emanuele Raineri, Roderic Guigó, and Paolo Ribeca. Fast computation and applications of genome mappability. *PloS one*, 7(1), 2012.

R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological sequence Analysis*. Cambridge University Press, 1998.

Nils J Fredriksson, Lars Ny, Jonas A Nilsson, and Erik Larsson. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nature Genetics*, 46:1258–1263, 2014a.

Nils J Fredriksson, Lars Ny, Jonas A Nilsson, and Erik Larsson. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nature genetics*, 46(12):1258–1263, 2014b.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., 2001.

Michael S Lawrence, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, Chip Stewart, Craig H Mermel, Steven A Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013.

Lucas Lochovsky, Jing Zhang, Yao Fu, Ekta Khurana, and Mark Gerstein. LARVA: An integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Research*, 2015.

D. McFadden. Conditional logit analysis of qualitative choice behaviour. In P. Zarembka, editor, *Frontiers in Econometrics*. Academic Press, 1974.

Collin Melton, Jason A. Reuter, Damek V. Spacek, and Michael Snyder. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature Genetics*, 47(7):710–716, 2015. doi: 10.1038/ng.3332. URL `http://dx.doi.org/10.1038/ng.3332`.

Serena Nik-Zainal, David C Wedge, Ludmil B Alexandrov, Mia Petljak, Adam P Butler, Niccolo Bolli, Helen R Davies, Stian Knappskog, Sancha Martin, Elli Papaemmanuil, et al. Association of a germline copy number polymorphism of apobec3a and apobec3b with burden of putative apobec-dependent mutations in breast cancer. *Nature genetics*, 46(5):487–491, 2014.

Paz Polak, Rosa Karlic, Amnon Koren, Robert Thurman, Richard Sandstrom, Michael S. Lawrence, Alex Reynolds, Eric Rynes, Kristian Vlahovicek, John A. Stamatoyannopoulos, and Shamil R. Sunyaev. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature*, 518(7539):360–364, Feb 2015. ISSN 0028-0836. URL `http://dx.doi.org/10.1038/nature14221`. Letter.

Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1):110–121, 2010.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org/`.

Malene Juul Rasmussen, Johanna Bertl, Qianyun Guo, Morten Muhlig Nielsen, Michal Switnicki, Henrik Hornshoj, Tobias Madsen, Asger Hobolth, and Jakob Skou Pedersen. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *elife, accepted for publication*, 2017.

Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature*, 518: 317–330, 2015.

Michael R Stratton, Peter J Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239): 719–724, 2009.

The Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 187:330–337, 2012.