Supplementary Material for LARVA

1. Pseudogene UTR, TSS, and promoter sites removal

Pseudogenes are known to be hotspots of artifact in numerous genomics analyses. It is partially because that read mapping in pseudogenes might be complicated due to their context similarity with their parent genes. In order to analyze the mutation events in the pseudogene regions, we extracted all the pseudogenes from the Gencode annotation (version 19) and calculated the average mutation counts from the pooled samples in gene and pseudogene regions, and also the up- and downstream 2kb region of all pseudogenes. Possibly due to the shorter length of pseudogenes, a larger variance of the mutation rate was observed in the pseudogene, although two-sided Wilcoxon test shows no significant difference (P = 0.453). However, we observed a noticeable elevated mutation rate in the up- and downstream regions of pseudogenes (Fig. S2). In order to exclude artifacts, such as variant calling difficulties, we excluded the pseudogenes from the Gencode gene list in our analysis.

2. Details of model fittings

The constant mutation rate assumption and the resultant binomial distribution

The underlying assumption of the binomial model is that the mutation rate within the given region is constant. Suppose the target region has n bases in length, and the homogeneous mutation rate is p. Then mutation count x inside this region falls into a binomial distribution with the probability mass function as

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad (1.1).$$

Given the mutation count data, the maximum likelihood estimator of the mutation rate is just

$$\hat{p} = \frac{\sum_{i=1}^{k} x_i}{\sum_{i=1}^{k} n_i} \quad (1.2)$$

where k represents the total number of regions and i is the region index.

The beta-binomial distribution used in LARVA

Instead of the fixed mutation rate assumption, we provided more flexibility of the mutation rate by allowing it to follow a beta distribution

Unknown
Field Code Changed
Unknown
Field Code Changed
Unknown
Field Code Changed
Unknown
Field Code Changed
Unknown
Field Code Changed

$$\pi(p|\alpha,\beta) = Beta(\alpha,\beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{Beta(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \quad (1.3).$$

Suppose the mutation count is $x_i, i = 1,2,\cdots,k$, and the sample size and binomial probability can be expressed as n_i and p_i. Instead of assuming the mutation in all the bins is a constant, we can set up a two-stage model

$$\begin{aligned} x_i | p_i &\sim Binomial(n_i, p_i) \\ p_i &\sim Beta(\alpha,\beta) \end{aligned} \quad (1.4).$$

Then the total number of mutations within the bin with length *n* follows the beta binomial distribution as in (1.5)

$$\Pr\{X = x_i\} = \binom{n_i}{x} \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+x_i)\Gamma(\alpha+n_i-x_i)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n_i)} \quad (1.5)$$

To estimate the parameters in beta-binomial distribution we used the scheme described in (1,2). When the target bin length is fixed, resulting in $n_i = n, i = 1,2,\cdots k$, we the mean and variance of mutation counts can be written into

$$\begin{aligned} E[X] &= n\frac{\alpha}{\alpha+\beta} = n\mu \\ \text{var}[X] &= n\mu(1-\mu)\sigma, \\ \sigma &= \frac{1}{\alpha+\beta+1} \end{aligned} \quad (1.6)$$

For simplicity, we directly estimate μ and σ instead of α and β. Hence the moment estimator can be immediately got from equation (1.6).

When the target region length is variable, estimation is a little bit more complicated. Define additional parameters for mathematical convenience as in (1.7).

$$\begin{aligned} \hat{p} &= \frac{\sum_{i=1}^{k} w_i \hat{p}_i}{w} \\ w_i &= \frac{n_i}{1+\sigma(n_i-1)} \\ w &= \sum_{i=1}^{k} w_i \\ S &= \sum_{i=1}^{k} w_i (\hat{p}_i - \hat{p})^2 \end{aligned} \quad (1.7)$$

We can have the moment estimator in (1.8)

$$\mu = \hat{p} = 1 - \hat{q}$$
$$\sigma = \frac{S - \hat{p}\hat{q}\left[\sum_{i=11}^{k} \frac{w_i}{n_i}\left(1 - \frac{w_i}{w}\right)\right]}{\hat{p}\hat{q}\left[\sum_{i=1}^{k} w_i\left(1 - \frac{w_i}{w}\right) - \sum_{i=1}^{k} \frac{w_i}{n_i}\left(1 - \frac{w_i}{w}\right)\right]} \quad (1.8)$$

Unknown
Field Code Changed

However, from (1.8), w_i is also a function of σ which is to be estimated, and there is no analytical solution to it. Hence as suggested in (2), we initially assigned the w_i proportional to n_i to get a rough estimate of γ. Then w_i was updated with this estimate to obtain a more accurate estimation of σ.

3. Coding Region Mutation Burden Analysis

We evaluated LARVA's ability to identify statistically significant mutation burdens in genes. Exome variant data was obtained from The Cancer Genome Atlas (TCGA) Data Portal (3). The complete set of exome variant calls includes 20 cancer types and 5032 samples in total. A detailed graph of the collected data is provided in Fig S8.

Gene annotation data was derived from the GENCODE v19 annotation files (4). All complete protein-coding transcripts were extracted, and all the transcripts for each gene were merged, as demonstrated in Fig S9. This data spanned 19,822 genes, and a total of 252,356,877 nucleotides. We plotted the distribution of gene lengths in Fig S10. The total number of mutations falling into the merged gene regions is 3,547,350, and the average mutation rate is 0.0141 for the pooled samples. As with the noncoding regions, huge mutation heterogeneity was observed in the coding regions (Fig S11).

We removed the genes with length less than the bottom 5% of gene lengths for higher annotation confidence, and then compared the performance of LARVA and the binomial test. After *p*-value adjustment, LARVA found 7 genes that are potentially under higher mutation burden (Table S3). For each of these genes, we searched for literature supporting cancer association. Except for KRTAP4-11, we found all the remaining genes are clearly documented with some cancer association. Note that we reported only one Pubmed ID per gene, even if there are many more supporting references. Our findings effectively demonstrate that LARVA is capable of finding meaningful results on protein coding regions. On the other hand, the *p*-values for the binomial test method were heavily inflated. After *p*-value adjustment, there are 6759 out of 18,826 genes, roughly 35.90%, with *p*-value less than 0.05. It is very unlikely that all such genes are associated with cancer.

P-values given by LARVA and binomial test are given in Fig. S12. It is shown that the p-value distribution from the binomial test severely violates the uniform distribution assumption, which is consistent with its bad fitting of the data. On the other hand, the p-values from the LARVA method (Fig. S12, left hand side) roughly follow a uniform distribution. It is worth mentioning that after replication timing correction, the p-values from LARVA method have improved concordance with the theoretical distribution, indicating the importance of correction.

Lucas Lochovsky 4/20/2015 4:29 PM
Deleted: values

Lucas Lochovsky 4/20/2015 4:30 PM
Deleted: the

Supplementary figures

Figure S1

Boxplot of mutations count in 10k, 100k, and 1mb regions with or without overlapping with the blacklist region. P-values were calculated from the two-sided Wilcoxon tests ($P<2.2\times10^{-16}$ for 10kb bin, $P=4.767\times10^{-5}$ and 0.473 for the 100k, and 1mb bins). In the smaller bin regions (10k and 100k), regions overlapped with blacklists demonstrates significantly higher mutation rate.

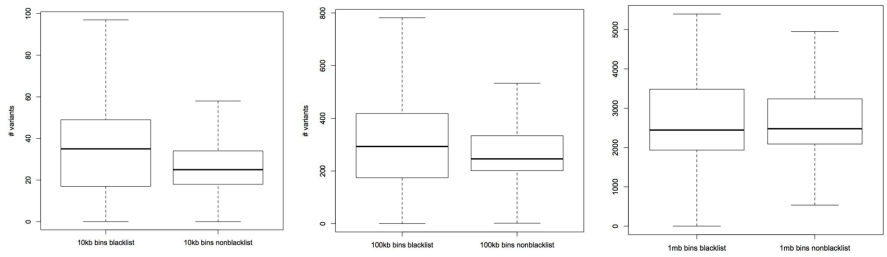


Figure S2

Average mutation rate estimation from gene, pseudogene, and regions before and after pseudogene regions.

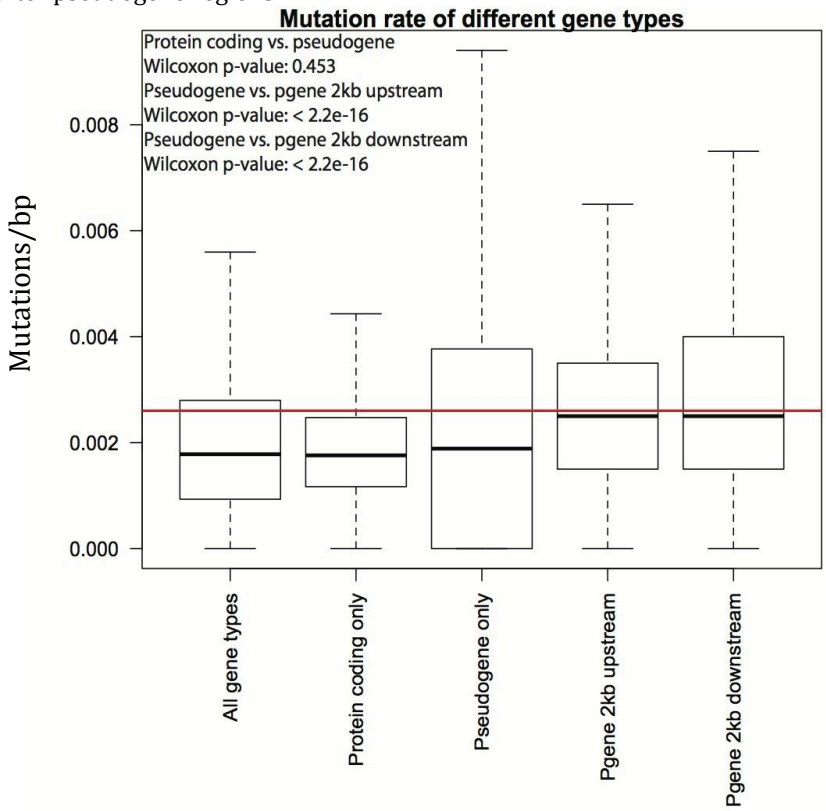


Figure S3

Fitting comparison between beta-binomial and binomial distribution. (A) Density plot of the beta-binomial, binomial, and empirical distribution of read count data in 100kb bins; (B) C.D.F curve of the KS statistics of beta-binomial and binomial generated counts VS. random samplings in the observed counts; (C) Boxplots of the KS statistics.

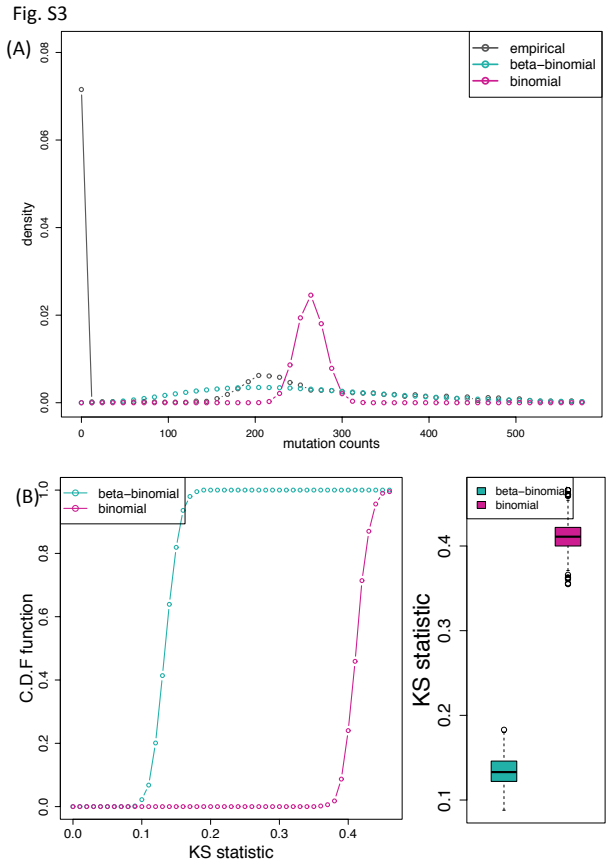


Figure S4

Fitting comparison between beta-binomial and binomial distribution. (A) Density plot of the beta-binomial, binomial, and empirical distribution of read count data in 1kb bins; (B) C.D.F curve of the KS statistics of beta-binomial and binomial generated counts VS. random samplings in the observed counts; (C) Boxplots of the KS statistics.

Fig. S4

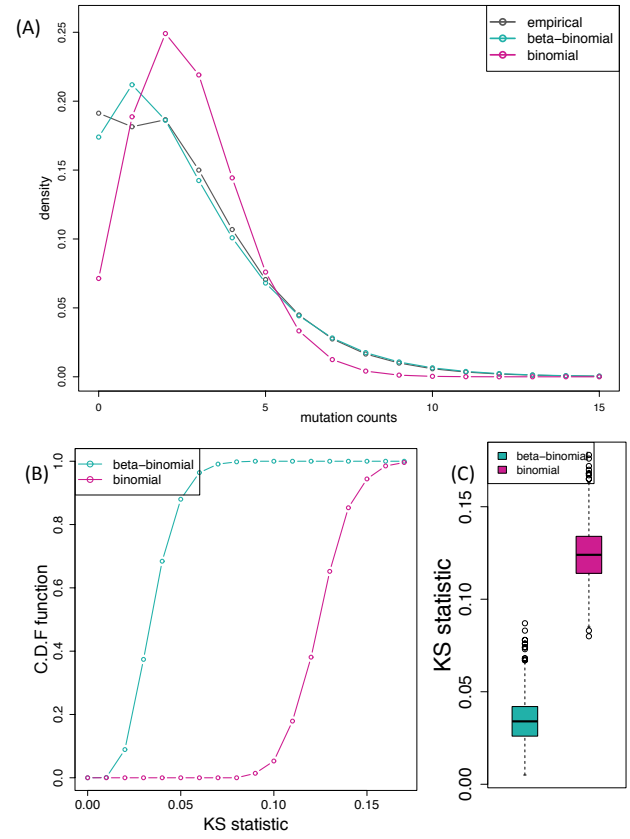


Figure S5

Half of the observed data is used for model fitting of both beta-binomial and binomial distribution, and the remaining half was used to calculate the KS statistics with generalizations from the fitted distributions. Boxplots of 100 repeats were given below.

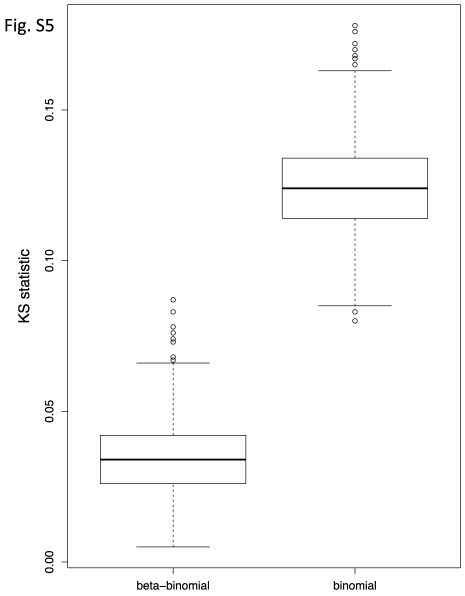


Figure S6

The smooth scatter plots of the mutations count in all tumor samples within 1kb bins vs. its averaged replication timing value. A linear regression was fitted and the R-squared values are up to 0.124.

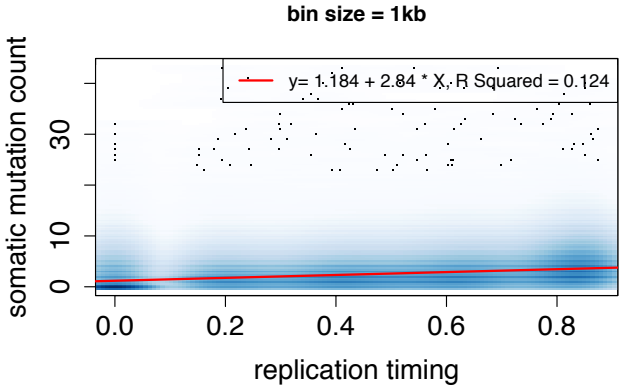


Figure S7

The fitted μ and σ were plotted for each the 10 used replication timing bins.

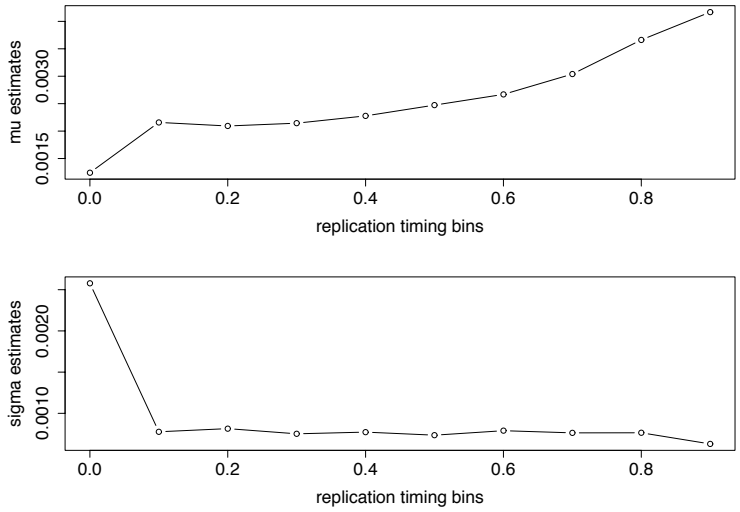


Figure S8

TCGA Whole Exome Sequencing samples by cancer types.

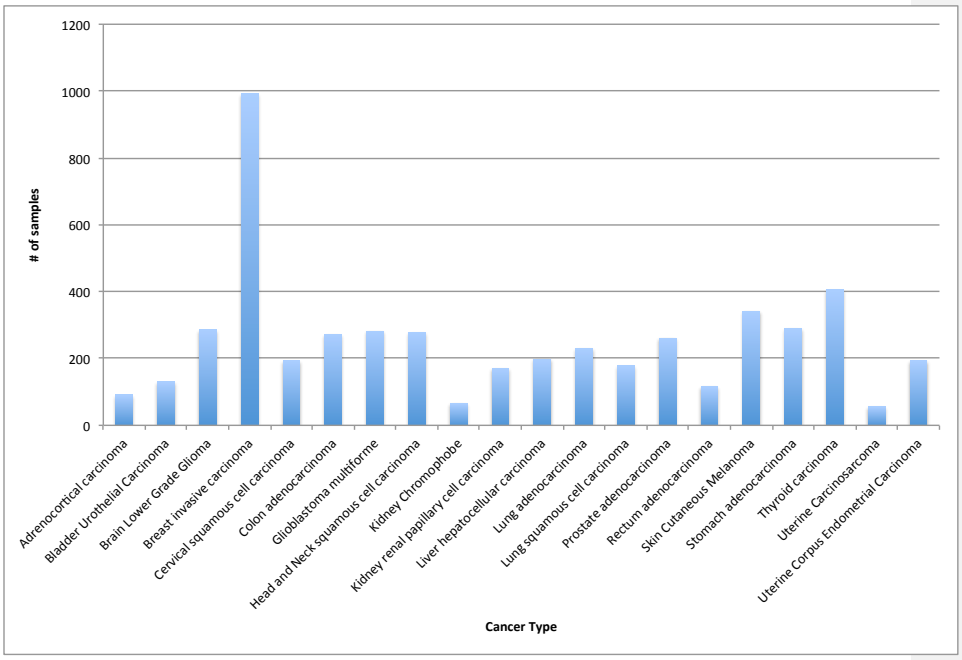


Figure S9

Details of gene regions definition. Note that only coding transcripts were used for the Whole Exome Sequencing data analysis.

Figure S10
Distribution of gene length

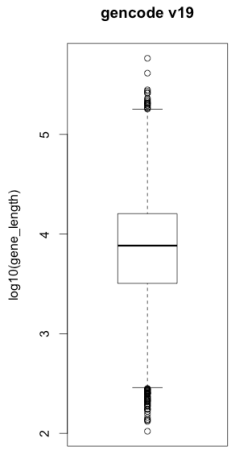


Figure S11
Distribution of the pooled mutation rates

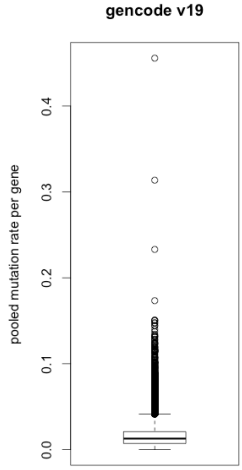


Figure S12

QQ plots of calculated P values VS. uniform theoretical ones of the coding region analysis.

Lucas Lochovsky 4/20/2015 4:31 PM
Formatted: Font color: Text 1

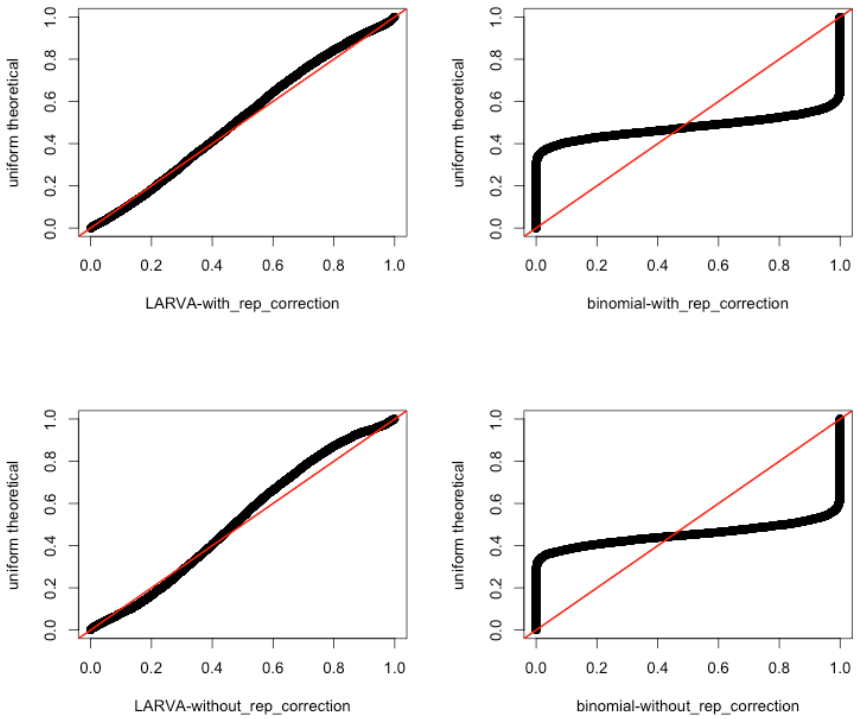
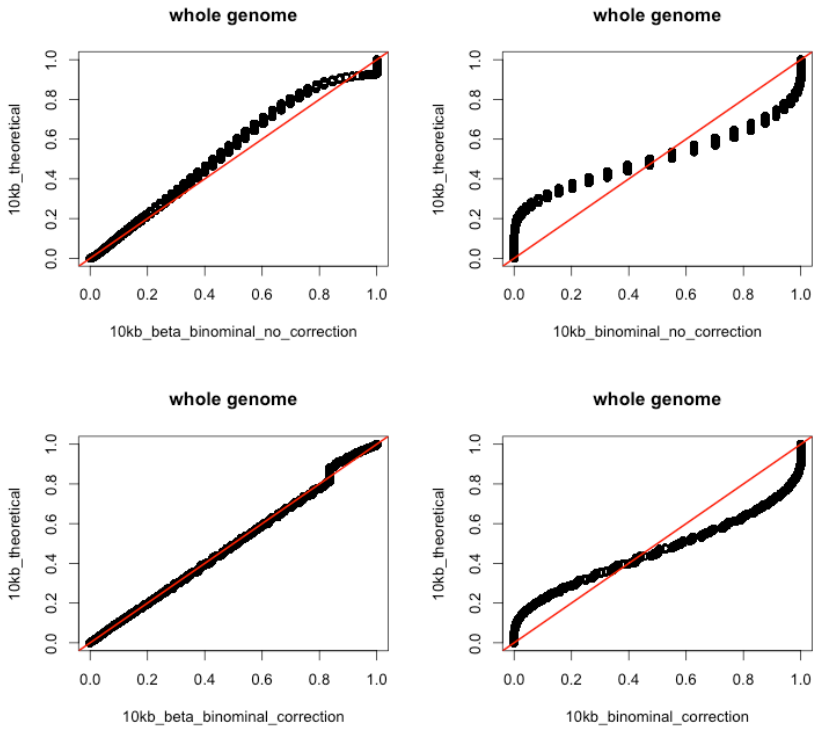


Figure S13: QQ plots of calculated P values VS. uniform theoretical ones of the whole genome 10kb bins analysis.

Lucas Lochovsky 4/20/2015 4:31 PM
Formatted: Font color: Text 1



Supplementary tables

Table S1

Summary of the whole genome sequencing cancer data used in this study

Cancer Type	# of Samples
Acute Lymphoblastic Leukemia	1
Acute Myeloid Leukemia	7
Breast Cancer	119
Chronic Lymphocytic Leukemia	28
Glial Tumor	26
Kidney Carcinoma	32
Liver Cancer	88
Lung Adenocarcinoma	24

Lymphoma B-cell	24
Medulloblastoma	100
Pancreatic Cancer	15
Pilocytic Astrocytoma	101
Prostate Cancer	95
Stomach Cancer	100
Sum	760

Table S2

Percentage of coding mutations in all samples (attached in supplementary file)

Table S3

Genes with significant mutation burden, according to LARVA's exome analysis.

Gene	Adjusted P value	Simple annotation	Supporting Reference
TP53	0	Well-known oncogene	PMID:20182602
BRAF	2.332696e-04	B-Raf proto-oncogene	http://ghr.nlm.nih.gov/gene/BRAF
KRTAP4-11	3.323269e-03	Unknown	
IDH1	3.323269e-03	Glioblastomas, astrocytomas, oligodendroglial tumors	PMID:19435942
FRG1B	4.860527e-03	lineage-specific mutation patterns in many cancer types	PMID: 24465236
CDKN2A	9.842880e-03	pancreatic cancer	PMID: 21150883
PRSS1	2.341413e-02	pancreatic cancer	PMID:22379635

References

1. Young-Xu, Y. and Chan, K.A. (2008) Pooling overdispersed binomial data to estimate event rate. *BMC medical research methodology*, **8**, 58.
2. Kleinman, J.C. (1975) Proportions with extraneous variance: two dependent samples. *Biometrics*, **31**, 737-743.
3. Cancer Genome Atlas Research, N. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061-1068.
4. Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*, **22**, 1760-1774.