LARVA: an integrative framework for Large-scale Analysis of Recurrent Variants in noncoding Annotations

Lucas Lochovsky[1], Jing Zhang[2] , Yao Fu[1], Ekta Khurana[2], and Mark Gerstein[1,2,3,*]

[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]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

Present Address: Mark Gerstein, Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA

ABSTRACT

Noncoding variants are known to be associated with numerous diseases, but they are barely investigated due to poor functional interpretation. Mutation burden tests are usually used to detect the highly mutated regions, but extensive overdispersion in mutation count data further complicates the significance assessment. We address these issues with the development of a new computational framework called LARVA (Large-scale Analysis of Recurrent Variants in noncoding Annotations). LARVA first integrates a comprehensive set of noncoding regulatory elements from the ENCODE project and then models the mutation count by beta-binomial distribution to handle the overdispersion. Furthermore, it incorporates regional genomic features like replication timing to achieve a local background model for more accurate enrichment assessment. We demonstrate the effectiveness of LARVA by reporting well-known noncoding drivers, such as TERT promoter, on 760 cancer whole genomes. Several novel noncoding regulators were also discovered as potential new driver candidates. We release our data, which includes noncoding regulatory elements, mutation counts and p-values, to serve as a valuable resource for cancer researchers. In addition, although naturally designed for somatic variant analysis, the logic of LARVA can be immediately extended to germline variants to discover the hypermutated regions of complex diseases. LARVA is available at [url].

INTRODUCTION

Genomes of numerous patients have been sequenced (1-5), opening up opportunities to identify the underlying genetic causes for complex disease (6-9) and develop more effective therapies targeted at specific molecular disease subtypes (10). Most of these studies have so far focused on identifying mutations and defects in the protein coding regions, or exomes, of disease genomes (2,11-14). These methods usually search for coding regions with higher than expected mutation frequencies in protein coding genes through rigorous background mutation rate control over a variety of genomic features (11). Such methods have been successfully used on numerous cancer genomes (15). However, the

The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint First

Jing 12/21/2014 4:35 PM
Deleted: their

Jing 12/21/2014 4:26 PM
Deleted: Moreover,

Jing 12/21/2014 4:26 PM
Deleted: the noncoding

Jing 12/21/2014 4:27 PM
Deleted: , due to mutation rate ... [1]

Jing 12/21/2014 4:27 PM
Deleted: in the mutation burden test

Jing 12/21/2014 4:28 PM
Deleted: for both germline and so ... [2]

Jing 12/21/2014 4:28 PM
Deleted: , which span both proxima ... [3]

Jing 12/21/2014 4:38 PM
Deleted: in each regulatory element as

Jing 12/21/2014 4:38 PM
Deleted: a

Jing 12/21/2014 4:38 PM
Deleted: distributed random variable

Jing 12/21/2014 4:39 PM
Deleted: observed

Jing 12/21/2014 4:40 PM
Deleted: more accurate

Jing 12/21/2014 4:39 PM
Deleted: estimate. Consequently, l ... [4]

Lucas Lochovsky 12/22/2014 1:41 PM
Deleted: cc

Jing 12/21/2014 4:32 PM
Deleted: analyzing a set of

Jing 12/21/2014 5:21 PM
Comment [1]: Not sure “are” or “w ... [6]

Jing 12/21/2014 4:32 PM
Deleted: sequences for recurrent ... [5]

Lucas Lochovsky 12/22/2014 1:53 PM
Deleted: are

Lucas Lochovsky 12/22/2014 1:54 PM
Deleted: Besides

Lucas Lochovsky 12/22/2014 1:54 PM
Deleted: the

Jing 12/21/2014 5:21 PM
Comment [2]: Are or were

Lucas Lochovsky 12/22/2014 1:53 PM
Deleted: P

Lucas Lochovsky 12/22/2014 1:54 PM
Deleted: are released

Lucas Lochovsky 12/22/2014 1:43 PM
Deleted: on

Lucas Lochovsky 12/22/2014 8:26 PM
Deleted: -

Lucas Lochovsky 12/22/2014 1:55 PM
Deleted: for

noncoding regions, which comprise more than 98% of the human genome, were rarely investigated, primarily due to the difficulty of functional interpretation of noncoding variants.

Recent genome annotation analysis has revealed a significant portion of the human genome is functional in a certain tissue or development stage (16,17), and several noncoding variants has been implicated in disease (18). For example, several genome-wide association studies (GWASs) studies have discovered the phenotypic effect of common noncoding variants in regulatory regions (19,20). Other studies have reported that noncoding TERT mutations drive cancer progression in multiple tumor types, including melanomas and gliomas (21-23). Moreover, mutations in the promoter regions of PLEKHS1, WDR74 and SDHD were also identified as recurrent driver mutations in some cancer types (24). In another example, analysis of the miRNA-binding sites on BRCA1 and BRCA2, the established drivers of breast cancer, indicated that certain variants in these sites are associated with increased risk of early onset breast cancer (25). Histones also serve as important noncoding regulators, as demonstrated in an analysis of a histone H1 variant linked to oncogene expression in ovarian cancer (26). In light of these discoveries, and the growing availability of whole-genome sequencing data (2,27-32), a statistical framework facilitating the identification of highly mutated noncoding mutations would be useful.

More recently, a genome wide computational effort has been made to discover the noncoding regions with higher mutation burden in cancer genomes (24). The authors called whole genome somatic variants for 863 human tumor sequences from The Cancer Genome Atlas (TCGA) (28), and analyzed the variants that fall into noncoding annotations. A p-value was computed for each annotation reflecting the likelihood that the given annotation had more variants than expected from background mutation processes, which was modelled with a binomial distribution. It successfully identified some known noncoding drivers, such as the TERT promoter, and reported some novel candidates that were not discovered previously. The use of the binomial distribution is based on two assumptions: 1) the mutation rate is homogeneous; 2) variants mutate independently. However, ~~human~~ genomes often violate these assumptions. First, studies on the coding variants already proved that the mutation rates in cancer genomes demonstrate substantial cancer type, sample, and regional heterogeneity (11). Second, some passenger mutations were generated by other driver events, such as structural alterations and mutations in DNA replication or repair genes (33). In the human genome, there are many regions with high linkage disequilibrium (LD), so mutations therein, at least for germline and possibly on somatic also, tend to arise together. Hence, some degree of dependency is to be expected in human germline and somatic mutation landscape. Consistent with these statements, we observed that the somatic mutation counts in the noncoding elements exhibited substantially higher variance than expected, the so-called overdispersion, indicating that a binomial distribution might be potentially inadequate to handle such data, and the resultant p-values might be heavily inflated.

Here, we present a computational system, LARVA (Large-scale Analysis of Recurrent Variants in noncoding Annotations), that identifies highly mutated noncoding regulatory elements using whole genome sequencing (WGS) variant data from multiple genetic disease patients. LARVA treats the

Jing 12/21/2014 4:54 PM
Deleted:

Jing 12/21/2014 4:54 PM
Deleted:

Jing 12/21/2014 4:54 PM
Deleted:

Jing 12/21/2014 4:54 PM
Deleted: of

Jing 12/21/2014 4:55 PM
Deleted: pioneering

Lucas Lochovsky 12/22/2014 1:59 PM
Comment [3]: We might want to indicate that this is both exome and whole genome, because we have a smaller number, but we only include whole genomes.

Jing 12/21/2014 4:56 PM
Formatted: Not Highlight

Jing 12/21/2014 5:05 PM
Deleted: in

Jing 12/21/2014 4:56 PM
Formatted: Not Highlight

Jing 12/21/2014 5:06 PM
Deleted:

Jing 12/21/2014 4:56 PM
Formatted: Not Highlight

Jing 12/21/2014 4:59 PM
Deleted: , it is observed that germline variants, and possibly somatic variants as well,

Jing 12/21/2014 4:56 PM
Formatted: Not Highlight

Jing 12/21/2014 5:08 PM
Deleted: be mutated in a highly associated way

Jing 12/21/2014 4:56 PM
Formatted: Not Highlight

Lucas Lochovsky 12/22/2014 2:01 PM
Deleted: was

Jing 12/21/2014 4:56 PM
Formatted: Not Highlight

Jing 12/21/2014 4:59 PM
Deleted:

Lucas Lochovsky 12/22/2014 2:01 PM
Deleted: so

mutations counts within a given regulatory element as a beta-binomial distributed random variable. This design automatically accommodates the heterogeneous nature of mutation accumulation in cancer genomes and the potential dependency among neighboring loci by allowing the mutation rate to be drawn from a beta distribution. Furthermore, we also divided the whole genome into several local bins and classified them using some known genomic confounders of the mutation rate, such as replication timing, for a more accurate local background mutation model. Such integrative analysis could potentially control the false positive rate in an effective manner. We demonstrate the usefulness of LARVA for finding both well-known and novel noncoding regulators with higher mutation burdens in a set of WGS cancer data that represents all the different types of whole genome sequenced cancers as of this writing (see Methods for details). We release the noncoding annotations, the mutation counts, and the corresponding p-values on the 760 cancer genomes used in this paper as a potentially powerful resource to facilitate cancer researchers for driver events discovery and validation in the future. Although naturally designed for somatic variant analysis, the logic of LARVA can be immediately extended for germline variant analysis in complicated diseases. The following sections describe LARVA's concepts, their applications to the study of genetic disease, and our cancer findings.

Jing 12/21/2014 5:15 PM
Deleted: calculation

Jing 12/21/2014 5:16 PM
Deleted: elements

Lucas Lochovsky 12/22/2014 2:03 PM
Deleted: further

Lucas Lochovsky 12/22/2014 2:03 PM
Deleted: d

Jing 12/21/2014 5:19 PM
Deleted: Besides the improved performance in cancer

Jing 12/21/2014 5:19 PM
Deleted: mutation

Jing 12/21/2014 5:19 PM
Deleted: also

Jing 12/21/2014 5:19 PM
Deleted: used

Jing 12/21/2014 5:20 PM
Deleted: burden tests

MATERIAL AND METHODS

Whole genome cancer variant data

We collected whole genome cancer variant calls from a large number of previously sequenced cancer genomes. The majority of our data came from a set of 507 whole genome cancer samples published in Alexandrov *et al.* (27). This data spans breast cancer, lung cancer, leukemia, pancreatic cancer, pilocytic astrocytoma, medulloblastoma, liver cancer, and lymphoma (Fig **1 A** and supplementary table 1). This was supplemented with a collection of 95 prostate cancer samples we obtained from publications (2,28-30), a set of 26 unpublished glial tumor samples, 32 kidney cancer samples from the TCGA (28), a set of 100 stomach cancer samples from Wang *et al.* (31).

Quality control of the WGS variants

A number of genomic regions are known to have poor read mappability due to sequence phenomena that cause ambiguous mapping results, such as a large number of tandem repeats. These regions are known as *signal artifact blacklist regions* (34). Since it is likely that variant calls in this region are possibly inaccurate, we opted not to use these variants in our mutation rate calculations (details in Fig. S1). Blacklist regions were derived from (34), and downloaded from the UCSC Genome Browser. Variants intersecting these regions, as determined by BEDTools (35), were removed from the analysis.

Noncoding annotation summary

Our analysis covered a range of noncoding regulatory annotations. The GENCODE v16 main annotation file was parsed to derive the coordinates of regulatory annotations close to gene regions,

including promoters and untranslated regions (UTRs)(36). Transcription factor (TF) binding sites were derived from the Chip-seq experiments conducted as part of the ENCODE project (37). We collected the full list of TF binding sites in all possible tissues and cell lines from ENCODE. Distal regulatory modules (DRM) enhancers, which regulate the expression of genes at non-adjacent sites, were derived from (38). Another class of regulators, the Dnase I hypersensitive (DHS) sites (39), were also derived from the ENCODE project. Additionally, we added a set of sites deemed "ultra-conserved" in (40) due to their extremely high level of conservation across many species. Furthermore, we used a set of "ultra-sensitive" sites from (41), so named because they are noncoding regions under higher selective pressure from the population genetics perspective. Finally, similar to the 2500bp promoter sites, we studied the more proximal transcription start sites (TSSes) by extracting the 100bp regions immediately upstream of GENCODE gene coding annotations (36). Table 1 summarizes the noncoding annotations.

Pseudogenes are known hot spots for artifacts due to their high context resemblance to their parent genes. In order to avoid potential variant calling bias, partially due to mapping difficulty, we removed the promoters, TSS, and UTR analyses for pseudogenes in the GENCODE annotation (details in Fig. S2 Text S1 section 1).

Models used for significance evaluation of mutation burden

The mutation counts for each regulatory element were calculated from the 760 cancer genomes mentioned above. For each regulatory element category, three models were used to calculate the mutation rate that would be expected due to background stochastic mutation processes for significance evaluation.

Suppose there are k noncoding regulatory elements (e.g. TF binding sites) to be analyzed. For the i^{th} element, let n_i stand for the total number of nucleotides. x_i and p represent the number of mutations with in element i and the probability of observing a mutation in each position. Some previous models (24,42) assumed that p is constant over the entire genome and mutations occur in an independent way. Hence, in model 1 x_i can be described as a binomial distribution in the following.

$$x_i : Binomial(n_i, p) \quad (1)$$

However, due to the heterogeneous nature of the cancer genomes and the possible dependency among neighboring loci, large overdispersion was found in the mutation count data. As a result we first improved the model 1 into a two-layer hierarchical model 2. Instead of setting p as a constant, we allow it to be drawn from a beta distribution with two parameters μ and σ indicating the average

Lucas Lochovsky 12/22/2014 2:05 PM
Deleted: point of view

Lucas Lochovsky 12/22/2014 2:05 PM
Deleted: for

Jing 12/21/2014 5:24 PM
Deleted: Significance

Jing 12/21/2014 5:25 PM
Deleted: recurrence

Jing 12/21/2014 6:14 PM
Deleted: We used our set of cancer variant calls to derive a "null model" representative of

Jing 12/21/2014 5:45 PM
Formatted: Indent: First line: 0.48 cm

Unknown
Field Code Changed

Unknown
Field Code Changed

Unknown
Field Code Changed

Unknown
Field Code Changed

Unknown
Field Code Changed

Unknown
Field Code Changed

Unknown
Field Code Changed

Jing 12/21/2014 5:45 PM
Deleted: this rate

Unknown
Field Code Changed

Jing 12/21/2014 5:53 PM
Formatted: Right, Indent: First line: 0.48 cm

Unknown
Field Code Changed

Jing 12/21/2014 5:45 PM
Formatted: Indent: First line: 0.48 cm

Unknown
Field Code Changed

Unknown
Field Code Changed

Unknown
Field Code Changed

mutation rate and overdispersion respectively (details in **Text S1**). As a result, the marginal distribution of x_i follows a beta-binomial distribution.

$$x_i | p : \ Binomial(n_i, p)$$
$$p : \ Beta(\mu, \sigma) \quad (2)$$

Furthermore, mutation rates were known to be confounded by a lot of genomic features, such as replication timing (represented by R), so we further divided the noncoding regulatory elements into 10 bins according to the averaged replication timing signal. Within each bin, we assumed that the mutation rate follows the same distribution. Therefore, model 3 can be represented as

$$x_i | p : \ Binomial(n_i, p)$$
$$p : \ Beta(\mu|_R, \sigma|_R) \quad (3)$$
$$\mu|_R, \sigma|_R : \ \text{constant within the same } R \text{ bin}$$

Method of maximum likelihoods was used for model 1. The moment estimator mentioned in (43,44) was used to estimate the parameters in model 2 and 3, and the p-values were calculated accordingly for the three models (for details see section 2 in **Text S1**).

Workflow of LARVA

Workflow of LARVA was given in **Fig. 1B**. The cancer variants in VCF format pass through a quality control filter that includes removing those variants that fall into blacklist regions. The preprocessed variants, along with our collected set of noncoding annotations that do not overlap blacklist regions, are used in the main computation. The main processing step includes counting all variant intersections with the noncoding annotations. DNA replication timing was used in model 3 for local mutation rate corrections. For each annotation category, the background mutation model was calculated using model 1-3 mentioned above, and p-values were given accordingly.

RESULTS

Overview of the annotated noncoding variants on various cancer genomes

We sought to study the whole genome somatic mutation patterns of as many different cancer patients as possible. To that end, we collected whole genome cancer variant call sets from a range of cancer data repositories (27,28) and publications (2,27,29-32). Our data spans 760 genomes, and includes 14 types of cancer (Fig 1A and Supplementary Table S1).

As it is shown in table 1, our noncoding annotation list spans approximately 30% of the human genome. We observed different cancer types demonstrate distinct mutational preferences over these noncoding regions. To illustrate this phenomenon, we used 11 types of cancer from our overall dataset for which there are at least 20 samples and calculated the fraction of WGS mutations within

Jing 12/21/2014 5:58 PM
Formatted: Font:Bold

Unknown
Field Code Changed

Jing 12/21/2014 5:54 PM
Formatted: Right, Indent: First line: 0.48 cm

Unknown
Field Code Changed

Jing 12/21/2014 5:59 PM
Deleted: Hence, the entire genome's expected mutation can be represented as a binomial distribution $B(n,p)$ where n is the number of basepairs (bp) of the regulatory element, and p is the probability that a single nucleotide is mutated. However, due to the heterogeneous nature of the cancer genomes and the possible dependency among neighboring loci, the binomial distribution might be inadequate to describe the mutation count data.

Jing 12/21/2014 5:45 PM
Formatted: Indent: First line: 0.48 cm

Unknown
Field Code Changed

Jing 12/21/2014 5:59 PM
Deleted: Also, taking into consideration the confounding influence of other genomic features, such as the replication timing, we proposed the following hierarchical model.

Jing 12/21/2014 6:07 PM
Formatted: Right

Jing 12/21/2014 6:07 PM
Deleted: where the mutation count is $x_i, i = 1,2,\cdots,k$, and the sample size, mutation probability, and replication ... [7]

Jing 12/21/2014 6:17 PM
Deleted: the beta-binomial distribution

Jing 12/21/2014 6:09 PM
Deleted: against the fitted null model

Jing 12/21/2014 7:31 PM
Formatted: Font:Bold

Jing 12/21/2014 7:32 PM
Formatted: Font:Bold

Jing 12/21/2014 7:33 PM
Formatted: Font:Bold

Jing 12/21/2014 7:32 PM
Formatted: Font:Bold

Jing 12/21/2014 7:44 PM
Deleted: elements

Jing 12/21/2014 7:37 PM
Formatted: Indent: First line: 0 cm

Jing 12/21/2014 7:37 PM
Deleted: . ... [8]

each noncoding element category (boxplots of different color in Fig. 2). The overall nucleotide percentage of each annotation over the genome was used as the background (black dash lines in Fig.2). In one instance representative of the large differences observed between cancer types, variants in kidney cancer was found to be preferentially located in the TF binding site while lung adenocarcinoma is mutation depleted in this region (0.140 average vs. 0.098 average, light green color vs. aquamarine color in Fig. 2). Huge sample effect was also observed in several cancer types. For instance, within Pilocytic Astrocytoma, there are samples that have a TF binding peak mutation fraction as high as 0.252 and as low as 0.011, which represents a ~23-fold difference. Hence, it is important to understand the mutation patterns in these noncoding annotations, and take their unique characteristics into consideration.

REPHRASE

Large cancer type, sample, regional heterogeneity of cancer genomes, and the potential dependency among neighboring regions violate the binomial assumption

In (24), the mutation burden tests are performed based on the binomial distribution, which inherently assumes a constant mutation rate and completely independent mutation events. However, these assumptions might not be appropriate for either somatic or germline variant analysis.

First, in our analysis of hundreds of WGS somatic mutation signatures, we observed huge cancer type, sample, and regional somatic mutation rate heterogeneity. To demonstrate cancer type and sample mutation rate heterogeneity, we selected all cancer types with more than 20 samples in it. We split the human genome into 1 mega basepair (Mbp) size bins, and intersected the individual sample variants from our data set to calculate the mutation rate of each sample. Consistent with the analysis in coding regions (11), we observed huge mutation rate differences between cancer types. For instance, the average whole genome mutation rate in stomach cancer is as high as 11.389 mutations/Mbp (violet red colors in Fig 3.A), which is ~800 times of the mutation rate in medulloblastoma (0.0142, blue colors in Fig 3.A). Furthermore, the whole genome mutation rate also fluctuates wildly across samples, and such changes may go up to 100 times within the same cancer type (0.359 VS. 21.8 in breast cancer for example). Additionally, to illustrate regional mutation rate heterogeneity, we randomly selected 50 one-megabase-length regions to calculate the mean and standard deviation (SD) of the local mutation rate across samples in lung cancer and prostate cancer (Fig 3. B, dashed lines show the SD). As shown in Fig.3 B, the average local mutation may vary from 0 to 50.8 mutations/Mbp across the randomly selected bins, and the SD range is unusually huge for each bin. Similar results were also observed in prostate cancer (red dots and lines in Fig 3. B).

Several biological signatures could partially explain the observed mutation rate heterogeneity. For example, the later replicated regions usually suffer from accumulative DNA damage and therefore are prone to mutations (45). Besides, methylated cytosines in CpG sites are often unstable and undergo deamination to thymine, which yields a C to T transition (33). Hence, there is a noticeable mutation rate difference at CpG and non CpG sites. Several other hypothesis were also proposed and summarized in the excellent review paper (33).

Jing 12/21/2014 7:47 PM
Deleted: Replication timing is one of the well-known genomic features that affects both germline and somatic mutation rate. T

Jing 12/21/2014 7:47 PM
Deleted: replicating

Jing 12/21/2014 7:48 PM
Deleted: accumulate

Jing 12/21/2014 7:51 PM
Deleted: , such as oxidation and deamination, making them

Jing 12/21/2014 8:45 PM
Deleted: of all sources

Second, mutation events might not be independent of each other. For example, in germline mutation analysis, mutations with high LD are prone to mutate together. Additionally, some passenger mutations are generated by other driver mutations. The driver mutation might be a mutation in a DNA replication or repair gene. Moreover, some structural variations, such as long insertions or deletions, might cause problems in pairing during meiosis and thus generate additional point mutations in neighboring regions (46). Consistent with this hypothesis, the mutation rates of the surrounding structural variations are elevated in several eukaryotic-species (46-48).

Due to the violation of these two assumptions, we observed a much higher than expected variance in the mutation count data. For example, at a 10kb bin resolution, the observed mutation count variance is 7.679 times of the expected valued under the binomial assumption. Hence, it is necessary to introduce other statistical models to handle such overdispersion in the mutation count data.

Improved mutation count fitting through a beta-binomial distribution

As discussed in the previous section, a binomial distribution model used in (24), which assumes a constant mutation rate and independent mutation process, could be problematic in more practical data analysis applications when the mutation counts are highly overdispersed. Hence, we first proposed a two-layer model to fit the variant count data (model 2 in the method section). Instead of setting a constant mutation rate, our model treated the mutation rate as a beta distributed random variable, which flexibly provides the underlying mutation rate with desired mean and variance properties. Then the mutation counts within each regulatory element could be easily modelled as a beta-binomial distribution (details in methods).

We fitted the mutation count data at a 10kb bin resolution of the 760 WGS cancer genomes under the fixed (binomial) and variable (beta-binomial) mutation rate assumptions in Fig. 4. We calculated the frequency of the observed mutation count in each bin and compared it with the binomial (model 1) and beta-binomial (model 2) fittings respectively. It is shown in Fig. 4 A that the observed data (black dots) demonstrates much heavier tails than the binomial distribution (purple dots), while the beta-binomial distribution (green dots) fits very well at the right tail. In order to quantitatively exhibit the improved performance of beta-binomial fitting, we utilized Kolmogorov-Smirnov (KS) statistics to compare the two distributions with the observed data in a nonparametric way. A larger KS statistic indicates a higher level of deviation between the two distributions. Specifically, 1000 bins were simulated from beta-binomial and binomial fitted distributions separately to calculate the KS statistic against the randomly sampled 1000 mutation counts from the observed data. This scheme was repeated 1000 times and the cumulative distribution function (C.D.F) of the KS statistics were given in Fig. 4B. The median KS statistic value for the beta-binomial distribution was 0.087, significantly smaller than 0.218 of the binomial distribution (p-value for two-sided Wilcoxon test < 2.2×10^{-16}, boxplots given in Fig. 4C). Different bin sizes were analyzed using the sample method and results were similar (Fig. S3-Fig. S4). In order to avoid overfitting, we utilized half of the data for distribution fitting, and the remaining half as the input to calculate the KS statistic for evaluation. This scheme was repeated 100 times. The beta-binomial distribution still significantly outperforms the binomial

Lucas Lochovsky 12/22/2014 2:22 PM
Deleted: to

Lucas Lochovsky 12/22/2014 2:23 PM
Deleted: structure

Jing 12/21/2014 7:49 PM
Moved (insertion) [1]

Jing 12/21/2014 7:51 PM
Deleted: Consistent with

Jing 12/21/2014 7:50 PM
Deleted: this

Jing 12/21/2014 7:50 PM
Deleted: did

Jing 12/21/2014 7:52 PM
Deleted: with

Jing 12/21/2014 7:53 PM
Deleted: As shown in Fig.3, cancer genomes display large mutational heterogeneity due to various factors

Jing 12/21/2014 7:54 PM
Formatted: Indent: First line: 0 cm

Lucas Lochovsky 12/22/2014 2:24 PM
Deleted: last

Jing 12/21/2014 7:53 PM
Deleted: therefore the theoretical

Lucas Lochovsky 12/22/2014 2:24 PM
Deleted: -

Jing 12/21/2014 7:49 PM
Moved up [1]: Consistent with this assumption, we did observe a much higher than expected variance in the mutation count data. For example, at a 10kb bin resolution, the observed mutation count variance is 7.679 times of the expected valued under the binomial assumption. Hence, it is necessary to introduce other statistical models to handle such overdispersion in the mutation count data.

Jing 12/21/2014 7:54 PM
Deleted: . ... [9]

Jing 12/21/2014 7:57 PM
Deleted: utilized a beta distribution to describe the mutation rate more flexibly. This two-parameter distribution conveniently

Jing 12/21/2014 7:58 PM
Deleted: ,

Jing 12/21/2014 7:58 PM
Deleted: and

Jing 12/21/2014 7:59 PM
Deleted: also

Jing 12/21/2014 7:59 PM
Deleted: the mutation count data within a specific region

Lucas Lochovsky 12/22/2014 8:10 PM
Deleted: for

distribution (0.0821 vs. 0.216, p-value for two sided Wilcoxon test < 2.2×10^{-16}, Fig. S5). Hence, the improved performance of the beta-binomial distribution is due to its enhanced flexibility to handle the overdispersed mutation count data instead of overfitting.

In the significance analysis of recurrent mutations, p-values were usually calculated from the right tails of the null distribution. But the huge deviation of the binomial distribution from the observed one could potentially introduce huge p-value inflation, and consequently result in numerous false positives. We defined the p-values for the observed distribution as the percentage of bins with equal or larger mutation counts. However, the improved fitting of the beta-binomial distribution could solve this problem and provide more accurate p-value assessment.

Local background mutation rate calculation through replication timing correction further controls false positives and false negatives

Recently, several computational efforts have been made to link somatic mutation rates with several genomic features in protein-coding regions (11,33). A particularly well-known example is DNA replication timing. During replication, the single stranded DNA usually accrues endogenous DNA damage, such as oxidation and deamination (45). Hence, DNA that is replicated in a later stage would be susceptible to the effects of accumulative damage, and would be prone to all classes of substitutions. Consistent with this assumption, scientists observed that the later replicating regions demonstrate remarkably higher mutation rate (45). Although replication timing has been used successfully in the coding regions, little work has been done in the noncoding regions in cancer genomics. Hence, we explored the effect of replication timing on the mutation rate calculation (model 3 in method section), and the consequential effect on the p-value evaluation.

Using 1kb bins, we counted the average replication timing value within the bin, and then separated the top and bottom 10% of replication timing bins for mutation rate calculation. As shown in Fig. 5 A, we observed noticeable differences in the mutation rate vis-a-vis the replication timing signal. The average mutation count of the 760 samples was 1.200 for the bottom 10% replicating timing bins, as compared to 4.028 for the top 10 percent counterparts (p-value for two-sided Wilcoxon test < 2.2×10^{-16}). A KS test was performed to determine whether these two sets of mutation counts data follow the same distribution, and the p-value is less than 2.2×10^{-16}, indicating that the two distributions are significantly different.

Moreover, we observed that the mutation counts data for bins with similar replication timing values still shows extensive overdispersion. For example, for the bottom 10% of replication timing bins, the observed variance of mutation counts was 4.168, which is 3.477 times that under the binomial assumption. Consistently, we observed poor fitting of binomial distribution against the observed distribution, especially in the right tails (dark black bars vs. dark purple bars in Fig. 5A). The huge deviation in the right tails would result in huge p-value calculation inflation as shown in Fig. 5B. The p-value for 16 mutations in the bottom replication timing 1kb region from the empirical distribution shows only marginal significance (3.994×10^{-4}), but the binomial distribution could inflate it to 2.585×10^{-13} due

Jing 12/21/2014 7:41 PM
Deleted: overfitting

Lucas Lochovsky 12/22/2014 8:11 PM
Deleted:

Jing 12/21/2014 8:02 PM
Deleted: Replication

Jing 12/21/2014 7:43 PM
Deleted: helps to

Jing 12/21/2014 8:02 PM
Deleted: both

Lucas Lochovsky 12/22/2014 8:12 PM
Deleted: the

Lucas Lochovsky 12/22/2014 8:12 PM
Deleted: the

Lucas Lochovsky 12/22/2014 8:12 PM
Deleted: suffers from

Jing 12/21/2014 7:44 PM
Deleted: s

to its bad fitting of the heavy tails on the right side. But our beta-binomial distribution rigorously controls the p-values through the flexible mutation rate assumption (p-value = 1.002×10^{-3}). We demonstrated the better p-value curve of the beta-binomial distribution in a variety of data points and replication timings, indicating the robustness of our method (Fig. 5B).

Additionally, the replication timing effect correction further improves the p-value calculation to avoid potential false positives and false negatives. For instance, for a region among the top replication timing regions, 8 mutations in 1kb bin would give a p-value at 0.094 after replication correction from beta-binomial model, but might be reported as positive when ignoring replication timing effect (p-value = 0.038 from beta-binomial by mixing the top and bottom 10% replication timing points). Similarly, a p-value of 0.064 would reject 7 mutations within 1kb bin as significant without correction. However, if this point comes from the bottom 10% of replication timing region, the true p-value should be 0.030 due to its relatively lower local mutation rate. Hence, it is important to perform covariate correction before calculating p-values.

Lucas Lochovsky 12/22/2014 8:17 PM
Deleted: with

Lucas Lochovsky 12/22/2014 8:17 PM
Deleted: refuse

LARVA discovered a list of highly recurrent noncoding regulatory regions from WGS data

We first applied LARVA to the 760 genomes' variants, intersecting them with the noncoding regions listed in Table 1. In total, LARVA reported 3964 and 3776 highly mutated regions before and after replication timing corrections, respectively (as shown in Table 2). On the other hand, the binomial distribution models reported at least 30 times more regions as significant because of the aforementioned p-value inflation, giving rise to a high false positive rate. We also tested the immediate 100bp upstream of every possible transcription start site (see methods for details), the results of which are depicted in Fig. 6 B. Forty-five TSSs passed the 0.05 p-value thresholds after p-value adjustment (BH method, (49)). Consistent with previous studies, we observed that the TSS for TERT came up in the top regions (Fig. 6 B), and the oncogene TP53 also ranked second among all sites. LMO3, which ranked third after replication timing correction, is a protein-coding oncogene that is predominantly expressed in brain tissue. It has been reported to be involved in a variety of cancer types, such as lung cancer (50) and neuroblastomas (51). PRRC2B's TSS was reported as the most significantly recurrent region among all TSSes. It is a protein-coding gene that is extensively expressed in brain tissue, but to our best of knowledge, there is no study to show the link of PRRC2B to cancer. Further investigations should be performed for the purpose of validation. Similar results were given for promoters and UTR regions as well. We selected all the genes with highly mutated TSSes, promoters, or UTRs (adjusted p-values after corrections ≤ 0.05) and performed GO analysis (http://amigo.geneontology.org, (52) results in Table S2). The top three enriched GO terms are: "negative regulation of fibrolast proliferation", "regulation of extrinsic apoptotic signaling pathway in absence of ligand", and "regulation of cell growth".

Lucas Lochovsky 12/22/2014 8:18 PM
Deleted: s

In terms of transcription factor binding sites, LARVA claimed 2054 out of the 5,710,954 binding sites as highly recurrent (0.036%). The transcription factor CTCF had 852 binding sites reported as significant (Table 3). CTCF is a multifunctional protein that is linked with multiple cancer types (53).

Specifically, several studies have reported that disruption of CTCF binding sites through mutations or abnormal methylation sites is closely associated with cancer (54,55). Moreover, we found that the oncogene BCL3 has a noticeably higher significant percentage with respect to the average (7.721 times of the average, p-value for two-sided binomial test = 6.762×10^{-13}). Interestingly, BCL3 is a proto-oncogene candidate which is closely associated with progression of diverse solid tumors (56). For example, BCL3 is aberrantly up- and down-regulated in breast cancer and nasopharyngeal carcinomas respectively, and is also reported to be strongly associated with survival in colorectal cancer. However, it is not a highly mutated gene according to our data: BCL3's mutation rate is 1.22 mutations/Mbp while the gene average is 2.52 mutations/Mbp. Our analysis suggests another possibility that the misregulation of BCL3 is possibly due to binding site disruption instead of the changes in the protein itself. Further computational and experimental effort should be made to clarify the mechanism of BCL3 regulation in different cancer types.

Lucas Lochovsky 12/22/2014 8:20 PM
Deleted: mutation

Whole genome recurrent events evaluation

Despite great efforts to annotate noncoding regions, there are still many regions with as yet unknown regulatory roles. In order to evaluate the recurrent events in these regions, LARVA provides all possible p-values, whether before or after adjustment, and with or without replication timing corrections, for high confidence bins on the genome (see methods for details) of variable length. We also compared the results from our beta-binomial model with the binomial models. For example, we randomly sampled 5000 10kb bins from the whole genome and made a Manhattan plot of p-values from both methods. It is obvious that the p-values from the binomial distribution were noticeably inflated (Fig. 7 B), while our beta-binomial model effectively controls the p-values (Fig. 7 A). The p-values obtained from using different bin sizes are provided in Table S4.

Lucas Lochovsky 12/22/2014 8:21 PM
Deleted: confident

DISCUSSION

Due to the rapid decline in time and money involved to perform whole genome sequencing, data is now available for thousands of genomes where previously only a handful were available (57). However, the analyses necessary for finding useful patterns in this data, and making sense of it for clinical benefit, have not kept pace with this sudden increase. Therefore, it is important that new algorithms are developed that can efficiently mine relevant patterns from genome sequence data, and that user interfaces for finding and understanding that data are optimized so that clinicians and biologists, who may not have extensive technical expertise, can use these results effectively in their work.

Compared with the extensive computational and experimental efforts on the mutation patterns in the protein coding regions in the past decade (58), the noncoding regions, which was viewed as 'dark matter', and comprises up to 98% of the human genome, are barely investigated in cancer research studies, partially due to the limited knowledge of noncoding function. However, recently several examples clearly pinpointed the phenotypic effect of mutations in noncoding regulatory regions in a variety of cancer types. For instance, TERT promoter, a well-known example, has been associated

with several cancer types (21-23). Fusions of the 5' UTR of TMPRSS2 with ETS genes frequently observed in prostate cancer, as well as mutations in certain miRNA binding sites (59), can influence the binding affinity at these sites, and thus affect androgen receptor regulation in prostate cancer. Hence, it is important to explore the mutation landscapes of such noncoding regions.

In this paper, we have introduced a new computational framework for exploring patterns of mutation across either somatic or rare germline variants, especially in the noncoding regulatory regions of human genomes. We took advantage of the complete genome annotation efforts of the ENCODE project (16) to extract the most extensive catalog of noncoding regulatory regions to date. We included the TF binding sites and DHS sites from all ENCODE experiments, promoters, UTRs, predicted enhancers, conserved and sensitive noncoding regions from our previous efforts (18). We then explored 760 cancer genomes on this comprehensive list of noncoding annotations to search for the highly mutated regulatory regions as potential noncoding driver candidates.

Moreover, consistent with the highly heterogeneous protein coding regions (11), we observed larger than expected mutation variation across cancer types, samples, and genomic regions (Fig. 3). Therefore, the recently proposed binomial models, which assume a constant mutation rate and independence of mutation events, might be inadequate for the observed data (Fig. 4, Fig. S3-S4). Instead, we set up two hierarchical models to handle mutation count overdispersion (model 2 and model 3 in the method section). First, we flexibly modeled the mutation rate in the regulatory elements as a two-parameter beta distribution, hence the corresponding mutation count could be conveniently described as a beta-binomial distribution. It provided significant improvement over the binomial model. In addition, we found that other genomic features, such as replication timing, would largely affect the background mutation rate (Fig. S6) and consequently generate both false positives and negatives. We corrected the replication timing effect by estimating the local mutation parameters in the beta-binomial distribution for better p-value assessment.

In the 760 cancer whole genomes in our analysis, we discovered 3776 noncoding regulatory regions that have significantly higher mutations than expected and provided the mutation enrichment significance of bins with variable length on the whole genome (Table 2). A list of known noncoding hypomutated regions, such as TERT and TP53 TSS, were also reported by our analysis, which convincingly proved the effectiveness of LARVA in discovering functionally relevant results. We also observed some relatively novel results such as PRRC2B TSS, CTCF and BCL3 binding sites. BCL3 is a known oncogene that is highly associated with several solid tumors (56,60), but this gene itself is not enriched in somatic mutations. Our results advocate an alternate possibility that its involvement in cancer cells is actually in the disruption of its binding sites, rather than the disabling of the protein itself. We released the noncoding annotations, the mutation count in each regulatory element of the 760 cancer genomes, and the resultant p-values assessment to the public, which would potentially serve as a powerful resource for cancer researchers in the future for comparison and assessment of somatic mutation burdens.

Jing 12/21/2014 8:06 PM
Deleted:

Jing 12/21/2014 8:06 PM
Deleted:

Jing 12/21/2014 8:07 PM
Deleted: -

Jing 12/21/2014 8:08 PM
Deleted: a

Lucas Lochovsky 12/22/2014 8:25 PM
Deleted: to compare and access

It is worth pointing out that although LARVA was naturally designed to analyze somatic genomes, it can be immediately extended to discover the hypermutated regions for germline variants. As with somatic variants, the germline mutation landscape demonstrates extensive heterogeneity and dependency, which can't be properly handled by a binomial distribution. Furthermore, unlike GWAS common variants discovery, LARVA could combine both rare and common variants to assess the mutation burden in noncoding regulatory regions. Due to the popularity of rare variants in human genomes, LARVA could potentially serve as a powerful tool to discover hypermutated noncoding regulatory regions.

In summary, LARVA is a powerful computational method to explore a broad range of genome annotations to uncover the ones that are mutated across many samples. LARVA makes it possible to predict putative noncoding drivers of genetic disease, and prioritize these predicted drivers for more rigorous downstream analysis. This may lead to faster identification of important targets that may be used to suppress disease with therapies and drugs.

Jing 12/21/2014 8:13 PM
Moved (insertion) [2]

Lucas Lochovsky 12/22/2014 8:25 PM
Deleted: while

Jing 12/21/2014 8:48 PM
Deleted: to

Jing 12/21/2014 8:13 PM
Deleted: we

Jing 12/21/2014 8:13 PM
Deleted: demonstrated the effectiveness of our method by analyzing

Jing 12/21/2014 8:14 PM
Deleted: LARVA is also potentially suitable to discover noncoding regions under higher germline mutation burden

Lucas Lochovsky 12/22/2014 8:25 PM
Deleted: -

Lucas Lochovsky 12/22/2014 8:26 PM
Deleted: Similar to the

Lucas Lochovsky 12/22/2014 8:27 PM
Deleted: also

Lucas Lochovsky 12/22/2014 8:29 PM
Deleted: studies associated

Lucas Lochovsky 12/22/2014 8:27 PM
Deleted: cc

Lucas Lochovsky 12/22/2014 8:29 PM
Deleted: Because of

Lucas Lochovsky 12/22/2014 8:27 PM
Deleted: -

Jing 12/21/2014 8:15 PM
Deleted: .

Jing 12/21/2014 8:13 PM
Moved up [2]: It is worthwhile to point out that although we demonstrated the effectiveness of our method by analyzing somatic genomes, LARVA is also potentially suitable to discover noncoding regions under higher germline mutation burden.

FUNDING

This work was supported by the National Institutes of Health [5R01CA152057-02]. Funding for open access charge: National Institutes of Health.

REFERENCES

TABLE AND FIGURES LEGENDS

Figure 1. (A) A pie chart representing the distribution of samples in our dataset of collected whole genome sequenced (WGS) cancers. (B) A flowchart of LARVA's procedure for identifying significant highly mutated noncoding elements. Cancer variants in VCF format are passed through quality control filters, and then intersected with our noncoding annotation corpus. After factoring in regional mutation rate corrections, a beta-binomial distribution is fitted to the observed data, which allows the identification of elements with a significant mutational burden.

Figure 2. Mutational heterogeneity between different types of cancer within several prominent classes of noncoding annotations. The percentage of mutations varies widely between noncoding element types, between cancer types, and between samples of the same cancer type.

Figure 3. (A) Between samples of the same cancer type, there is huge mutation rate heterogeneity. For most cancers, the mutation rate spans several orders of magnitude. (B) Variation in the mutation rate across chromosome 1 in lung cancer (top) and prostate cancer (bottom).

Figure 4. (A) The beta-binomial distribution (pink line) provides better fitting to the observed mutation counts at 10kb resolutation (black line) of 760 cancer genomes, especially at the right tail as compared to the binomial distribution (turquoise line). (B) A comparison of the cumulative distribution

function (CDF) of the binomial distribution and the beta-binomial distribution from part A. (C) Boxplots of the Kolmogorov-Smirnov (KS) statistics.

Figure 5. The 1 kb genome bins representing the top 10% and bottom 10% of the DNA replication timing were used to derive an observed distribution of mutation counts, demonstrating the influence of replication timing. The fitted binomial and beta-binomial distributions are plotted as bar plots (A). P-values at different mutation counts were given by the observed, beta-binomial, and binomial distribution.

Figure 6. (A) The number of significant p-values implied by beta-binomial distribution and binomial distribution (with and without DNA replication timing correction). (B) A sorted p-value plot of the top significant TSSes derived from the LARVA analysis.

Figure 7. Manhattan plot of the p-values from 5000 randomly samples 10kb bins from binomial (A) and binomial distribution (B). Binomial distribution might provide heavily inflated p-values due to its inadequacy to capture the extensive overdispersion of the mutation count data.

Table 1. List of noncoding annotations collected for LARVA's analysis.

Table 2. Number of significant recurrently mutated elements in each noncoding annotation class derived by LARVA

Table 3. The top transcription factor binding sites (TFBSes) from LARVA's analysis of our 760 cancer dataset. These findings may point to important regulatory disruptions in cancer.

1. Barbieri, C.E., Baca, S.C., Lawrence, M.S., Demichelis, F., Blattner, M., Theurillat, J.P., White, T.A., Stojanov, P., Van Allen, E., Stransky, N. *et al.* (2012) Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nature genetics*, **44**, 685-689.
2. Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M. *et al.* (2013) Punctuated evolution of prostate cancer genomes. *Cell*, **153**, 666-677.
3. Grasso, C.S., Wu, Y.M., Robinson, D.R., Cao, X., Dhanasekaran, S.M., Khan, A.P., Quist, M.J., Jing, X., Lonigro, R.J., Brenner, J.C. *et al.* (2012) The mutational landscape of lethal castration-resistant prostate cancer. *Nature*, **487**, 239-243.
4. Shi, L., Zhang, X., Golhar, R., Otieno, F.G., He, M., Hou, C., Kim, C., Keating, B., Lyon, G.J., Wang, K. *et al.* (2013) Whole-genome sequencing in an autism multiplex family. *Molecular autism*, **4**, 8.
5. Almasy, L., Dyer, T.D., Peralta, J.M., Jun, G., Wood, A.R., Fuchsberger, C., Almeida, M.A., Kent, J.W., Fowler, S., Blackwell, T.W. *et al.* (2014) Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. *BMC Proceedings*, **8**, S2.
6. Tervasmaki, A., Winqvist, R., Jukkola-Vuorinen, A. and Pylkas, K. (2014) Recurrent CYP2C19 deletion allele is associated with triple-negative breast cancer. *BMC cancer*, **14**, 902.
7. Stefansson, O.A., Moran, S., Gomez, A., Sayols, S., Arribas-Jorba, C., Sandoval, J., Hilmarsdottir, H., Olafsdottir, E., Tryggvadottir, L., Jonasson, J.G. *et al.* (2014) A DNA

methylation-based definition of biologically distinct breast cancer subtypes. *Molecular oncology*.
8. Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S. *et al.* (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature*, **513**, 382-387.
9. Chen, X., Iliopoulos, D., Zhang, Q., Tang, Q., Greenblatt, M.B., Hatziapostolou, M., Lim, E., Tam, W.L., Ni, M., Chen, Y. *et al.* (2014) XBP1 promotes triple-negative breast cancer by controlling the HIF1alpha pathway. *Nature*, **508**, 103-107.
10. Kurtova, A.V., Xiao, J., Mo, Q., Pazhanisamy, S., Krasnow, R., Lerner, S.P., Chen, F., Roh, T.T., Lay, E., Ho, P.L. *et al.* (2014) Blocking PGE-induced tumour repopulation abrogates bladder cancer chemoresistance. *Nature*.
11. Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214-218.
12. Rudd, M.L., Mohamed, H., Price, J.C., AJ, O.H., Le Gallo, M., Urick, M.E., Cruz, P., Zhang, S., Hansen, N.F., Godwin, A.K. *et al.* (2014) Mutational analysis of the tyrosine kinome in serous and clear cell endometrial cancer uncovers rare somatic mutations in TNK2 and DDR1. *BMC cancer*, **14**, 884.
13. Long, G.V., Fung, C., Menzies, A.M., Pupo, G.M., Carlino, M.S., Hyman, J., Shahheydari, H., Tembe, V., Thompson, J.F., Saw, R.P. *et al.* (2014) Increased MAPK reactivation in early resistance to dabrafenib/trametinib combination therapy of BRAF-mutant metastatic melanoma. *Nature communications*, **5**, 5694.
14. Yadav, M., Jhunjhunwala, S., Phung, Q.T., Lupardus, P., Tanguay, J., Bumbaca, S., Franci, C., Cheung, T.K., Fritsche, J., Weinschenk, T. *et al.* (2014) Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature*, **515**, 572-576.
15. Youn, A. and Simon, R. (2011) Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics*, **27**, 175-181.
16. Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.
17. Gerstein, M.B., Rozowsky, J., Yan, K.K., Wang, D., Cheng, C., Brown, J.B., Davis, C.A., Hillier, L., Sisu, C., Li, J.J. *et al.* (2014) Comparative analysis of the transcriptome across distant species. *Nature*, **512**, 445-448.
18. Fu, Y., Liu, Z., Lou, S., Bedford, J., Mu, X., Yip, K.Y., Khurana, E. and Gerstein, M. (2014) FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome biology*, **15**, 480.
19. Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M.R. (2004) A census of human cancer genes. *Nature reviews. Cancer*, **4**, 177-183.
20. Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R. *et al.* (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome research*, **22**, 1589-1598.
21. Vinagre, J., Almeida, A., Populo, H., Batista, R., Lyra, J., Pinto, V., Coelho, R., Celestino, R., Prazeres, H., Lima, L. *et al.* (2013) Frequency of TERT promoter mutations in human cancers. *Nature communications*, **4**, 2185.
22. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J. *et al.* (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**, 1190-1195.
23. Grossman, S.R., Andersen, K.G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D.J., Griesemer, D., Karlsson, E.K., Wong, S.H. *et al.* (2013) Identifying recent adaptations in large-scale genomic data. *Cell*, **152**, 703-713.
24. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. and Lee, W. (2014) Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature genetics*, **46**, 1160-1165.

25. Erturk, E., Cecener, G., Polatkan, V., Gokgoz, S., Egeli, U., Tunca, B., Tezcan, G., Demirdogen, E., Ak, S. and Tasdelen, I. (2014) Evaluation of Genetic Variations in miRNA-Binding Sites of BRCA1 and BRCA2 Genes as Risk Factors for the Development of Early-Onset and/or Familial Breast Cancer. *Asian Pacific journal of cancer prevention : APJCP*, **15**, 8319-8324.
26. Medrzycki, M., Zhang, Y., Zhang, W., Cao, K., Pan, C., Lailler, N., McDonald, J.F., Bouhassira, E.E. and Fan, Y. (2014) Histone h1.3 suppresses h19 noncoding RNA expression and cell growth of ovarian cancer cells. *Cancer research*, **74**, 6463-6473.
27. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L. *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415-421.
28. Cancer Genome Atlas Research, N. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061-1068.
29. Berger, M.F., Lawrence, M.S., Demichelis, F., Drier, Y., Cibulskis, K., Sivachenko, A.Y., Sboner, A., Esgueva, R., Pflueger, D., Sougnez, C. *et al.* (2011) The genomic complexity of primary human prostate cancer. *Nature*, **470**, 214-220.
30. Weischenfeldt, J., Simon, R., Feuerbach, L., Schlangen, K., Weichenhan, D., Minner, S., Wuttig, D., Warnatz, H.J., Stehr, H., Rausch, T. *et al.* (2013) Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. *Cancer cell*, **23**, 159-170.
31. Wang, K., Yuen, S.T., Xu, J., Lee, S.P., Yan, H.H., Shi, S.T., Siu, H.C., Deng, S., Chu, K.M., Law, S. *et al.* (2014) Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nature genetics*, **46**, 573-582.
32. Cancer Genome Atlas Research, N., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C. and Stuart, J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, **45**, 1113-1120.
33. Hodgkinson, A. and Eyre-Walker, A. (2011) Variation in the mutation rate across mammalian genomes. *Nature reviews. Genetics*, **12**, 756-766.
34. Derrien, T., Estelle, J., Marco Sola, S., Knowles, D.G., Raineri, E., Guigo, R. and Ribeca, P. (2012) Fast computation and applications of genome mappability. *PloS one*, **7**, e30377.
35. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.
36. 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.
37. Rozowsky, J., Euskirchen, G., Auerbach, R.K., Zhang, Z.D., Gibson, T., Bjornson, R., Carriero, N., Snyder, M. and Gerstein, M.B. (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology*, **27**, 66-75.
38. Yip, K.Y., Cheng, C., Bhardwaj, N., Brown, J.B., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P., Snyder, M. *et al.* (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome biology*, **13**, R48.
39. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75-82.
40. Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S. and Haussler, D. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321-1325.
41. Khurana, E., Fu, Y., Colonna, V., Mu, X.J., Kang, H.M., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A. *et al.* (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science*, **342**, 1235587.

42. Ding, L., Wendl, M.C., Koboldt, D.C. and Mardis, E.R. (2010) Analysis of next-generation genomic data in cancer: accomplishments and challenges. *Human molecular genetics*, **19**, R188-196.
43. Young-Xu, Y. and Chan, K.A. (2008) Pooling overdispersed binomial data to estimate event rate. *BMC medical research methodology*, **8**, 58.
44. Kleinman, J.C. (1975) Proportions with extraneous variance: two dependent samples. *Biometrics*, **31**, 737-743.
45. Stamatoyannopoulos, J.A., Adzhubei, I., Thurman, R.E., Kryukov, G.V., Mirkin, S.M. and Sunyaev, S.R. (2009) Human mutation rate associated with DNA replication timing. *Nature genetics*, **41**, 393-395.
46. Tian, D., Wang, Q., Zhang, P., Araki, H., Yang, S., Kreitman, M., Nagylaki, T., Hudson, R., Bergelson, J. and Chen, J.Q. (2008) Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature*, **455**, 105-108.
47. Hollister, J.D., Ross-Ibarra, J. and Gaut, B.S. (2010) Indel-associated mutation rate varies with mating system in flowering plants. *Molecular biology and evolution*, **27**, 409-416.
48. McDonald, M.J., Wang, W.C., Huang, H.D. and Leu, J.Y. (2011) Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PLoS biology*, **9**, e1000622.
49. Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological).* **57**, 289-300.
50. Kwon, Y.J., Lee, S.J., Koh, J.S., Kim, S.H., Lee, H.W., Kang, M.C., Bae, J.B., Kim, Y.J. and Park, J.H. (2012) Genome-wide analysis of DNA methylation and the gene expression change in lung cancer. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer*, **7**, 20-33.
51. Isogai, E., Ohira, M., Ozaki, T., Oba, S., Nakamura, Y. and Nakagawara, A. (2011) Oncogenic LMO3 collaborates with HEN2 to enhance neuroblastoma cell growth through transactivation of Mash1. *PloS one*, **6**, e19297.
52. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, **25**, 25-29.
53. Filippova, G.N. (2008) Genetics and epigenetics of the multifunctional protein CTCF. *Current topics in developmental biology*, **80**, 337-360.
54. Ohlsson, R., Renkawitz, R. and Lobanenkov, V. (2001) CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends in genetics : TIG*, **17**, 520-527.
55. Takai, D., Gonzales, F.A., Tsai, Y.C., Thayer, M.J. and Jones, P.A. (2001) Large scale mapping of methylcytosines in CTCF-binding sites in the human H19 promoter and aberrant hypomethylation in human bladder cancer. *Human molecular genetics*, **10**, 2619-2626.
56. Maldonado, V. and Melendez-Zajgla, J. (2011) Role of Bcl-3 in solid tumors. *Molecular cancer*, **10**, 152.
57. Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nature biotechnology*, **26**, 1135-1145.
58. Koch, L. (2014) Cancer genomics: Non-coding mutations in the driver seat. *Nature reviews. Genetics*, **15**, 574-575.
59. Lin, P.C., Chiu, Y.L., Banerjee, S., Park, K., Mosquera, J.M., Giannopoulou, E., Alves, P., Tewari, A.K., Gerstein, M.B., Beltran, H. *et al.* (2013) Epigenetic repression of miR-31 disrupts androgen receptor homeostasis and contributes to prostate cancer progression. *Cancer research*, **73**, 1232-1244.

60. Kim, Y.M., Sharma, N. and Nyborg, J.K. (2008) The proto-oncogene Bcl3, induced by Tax, represses Tax-mediated transcription via p300 displacement from the human T-cell leukemia virus type 1 promoter. *Journal of virology*, **82**, 11939-11947.