A New Mutation-Profile-Based Method for Understanding the Evolution of Cancer Somatic Mutations

Zhan Zhou[1,2,4,†], Yangyun Zou[1,2,†], Gangbiao Liu[1,2], Jingqi Zhou[1,2], Shiming Zhao[1], Zhixi Su[1,2,*], Xun Gu[1,2,3,*]

[1]State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai, China,

[2]Ministry of Education Key Laboratory of Contemporary Anthropology, Center for Evolutionary Biology, Fudan University, Shanghai, China,

[3]Department of Genetics, Development and Cell Biology, Program of Bioinformatics and Computational Biology, Iowa State University, Ames, Iowa, USA,

[4]College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang, China

[†]These authors have contributed equally to this work.

[*]Correspondence to:

Xun Gu, **email:** xgu@iastate.edu

Zhixi Su, **email:** zxsu@fudan.edu.cn

Abstract

Human genes perform different functions and exhibit different effects on fitness in cancer and normal cell populations. Here, we present an evolutionary approach to measuring the selective pressure on human genes in cancer and normal cell genomes using the well-known dN/dS (nonsynonymous to synonymous substitution rate) ratio. We develop a new method called the mutation-profile-based Nei-Gojobori (mpNG) method, which applies sample-specific nucleotide substitution profiles instead of conventional substitution models to calculating dN/dS ratios in cancer and normal populations. Using 7,042 exome sequences from tumor-normal pairs, and germline variations from 6,500 exome sequences (ESP6500) as references, we found a significant relaxation of selective constraint for human genes in cancer cells. Compared with previous studies that focused on positively selected genes in cancer genomes, which potentially represent the driving force behind tumor initiation and development, we employed an alternative approach to identifying cancer constrained genes that strengthen negative selection pressure in tumor cells. As a conservative estimate of positively and negatively selected genes in cancer, we found 45 genes under intensified positive selection and 16 genes under strengthened purifying selection in cancer cells compared with germline cells. The cancer-specific positively selected genes are enriched for cancer genes and human essential genes, while several cancer-specific negatively selected genes have been reported as prognostic biomarkers for cancers. Therefore, our computation pipeline used to identify cancer positively and negatively genes may provide useful information for understanding the evolution of cancer somatic mutations.

Keywords: Cancer somatic mutations, Natural selection, Cancer essential gene, Evolution

Introduction

Since the pioneering work of Cairns and Nowell [1, 2], the evolutionary concept of cancer progression has been widely accepted [3-7]. Essentially, cancer cells evolve through random somatic mutations and epigenetic changes that may alter several crucial pathways, a process that is followed by the clonal selection of the resulting cells. Consequently, cancer cells can survive and proliferate under deleterious circumstances [8, 9]. Hence, knowledge of evolutionary dynamics will benefit our understanding of cancer initiation and progression. For instance, there are two types of somatic mutations in cancer genomes: driver mutations and passenger mutations [10, 11]. Driver mutations are those that confer a selective advantage on cancer cells, as indicated by statistical evidence of positive selection. However, some passenger mutations undergo purifying selection because they would have potentially deleterious effects on cancer cells [12, 13]. Between these two cases are passenger mutations that are usually considered to be neutral in cancer.

Analyses of large-scale cancer somatic mutation data have revealed that the effects of positive selection on cancer cells are much stronger than on germline cells [14, 15]. Given that many of the positively selected genes in tumor development act as the driving force behind tumor initiation and progression, it is understandable that almost all previous studies focused on the positively selected genes in cancer genomes [3, 16-19]. We have realized that an alternative approach, i.e., identifying cancer constrained genes that are highly conserved in tumor cell populations (under purifying selection), is also valuable. As we have known that essential genes are more evolutionary conserved [20], it would be feasible to identify cancer essential genes from the genes that are evolutionary conserved in cancer cells. As cancer essential genes may be not the driver genes for carcinogenesis, but are crucial for cancer cell proliferation and survival [21], this idea may be advantageous in addressing issues related to drug resistance in cancer therapies, especially in cancers with high intratumor heterogeneity.

Many previous studies used the ratio of nonsynonymous and synonymous substitution rate (dN/dS) to identify genes that might be under strong positive selection both in organismal evolution and tumorigenesis (e.g., [22-24, 14, 25, 15, 11, 26]). However, most of these studies applied well-known methods that are usually based on simple nucleotide mutation/substitution models, e.g., every mutation or substitution pattern having the same probability [27]. However, this may not be a realistic biological model because many recent cancer genomics studies have shown that mutation profiles are quite different between different cancer samples [28, 15]. In addition, context-dependent mutation bias, that is to say, base-substitution profiles that consider the flanking 5’ and 3’ bases of each mutated base, should be taken into consideration [28, 29].

In this study, we describe a new method, called the mutation-profile-based Nei-Gojobori (mpNG) method, to estimate the selective constraint in cancer somatic mutations. Simply stated, mpNG method removes an unrealistic assumption inherent in the original NG method (named NG86), wherein each type of nucleotide change has the same mutation rate [27], which can lead to nontrivial biased estimations when this assumption is violated considerably. Instead, mpNG implements an empirical nucleotide mutation model that simultaneously takes into account several factors, including single-base mutation patterns, local-specific effects of surrounding DNA regions, and tissue/cancer types. Using 7,042 tumor-normal paired whole-exome sequences (WESs), as well as rare germline variations from 6,500 exome sequences (ESP6500) as references, we used the mpNG method to identify the selective constraint of human genes in cancer cells. The potential for our computational pipeline to identify cancer constrained genes may provide useful information for identifying promising drug targets or prognostic biomarkers.

Results

The mutation profiles in cancer genomes and human populations are different

Estimating evolutionary selective pressure on human genes is a practicable method of inferring the functional importance of genes to a specific population. By comparing selective pressures on genes in cancer cell populations with those in normal cell populations, we may identify different functional and fitness effects of human genes in cancer and normal cells. The conventional method for measuring selective pressure is to calculate the dN/dS ratio using the NG86 method [27], which assumes equal substitution rates among different nucleotides. In our study, we used the cancer somatic mutations from 7,042 tumor-normal pairs as well as rare variations from 6,500 exome sequences from the National Heart, Lung, and Blood Institute (NHLBI) Grant Opportunity (GO) Exome Sequencing Project (ESP6500) as a reference, in order to compared the relative mutation probabilities from cancer somatic mutations and germline substitutions for all possible base substitutions, considering the identities of the bases immediately 5' and 3' of each mutated bases and depicted the mutation profiles as 96 substitution classifications [28, 29]. The mutation profiles exhibits the prevalence of each substitution pattern for somatic point mutations, which present not only the substitution types but also the sequence context (see Materials and Methods) [29]. The exonic mutation profiles of cancer somatic substitutions and germline substitutions were differed from one another, and the intronic and intergenic mutation profiles were quite different from the exonic mutation profile of cancer cells (Figure 1). We also calculated the exonic mutation profiles of four different cancer types: colon adenocarcinoma (COAD), lung adenocarcinoma (LUAD), skin cutaneous melanoma (SKCM), and breast carcinoma (BRCA). These cancer types varied considerably not only in mutation rates but also in mutation patterns. Especially, the mutation rate of SKCM was much higher than other three types and the mutation profiles of SKCM were highly enriched in the C-to-T substitution pattern (Figure 1), indicating a direct mutagenic role of ultraviolet (UV) light in SKCM pathogenesis [30]. The different mutation profiles may present different biological progresses in carcinogenesis, which have been depicted in several publication [17, 28]. Hence, it is inappropriate to use conventional methods such as NG86 [27] to measure selective pressure by means of dN/dS calculation, which ignores the mutation bias of different nucleotide substitution types.

Measuring selective pressure on human genes in cancer and germline cells using the mpNG method

We therefore formulated an evolutionary approach that were designed specifically to estimate the selective pressure imposed on human genes in cancer cells and to then identify genes that had undergone positive and purifying selection in cancer cells rather than in normal cells (see Figure 2 for illustration). We developed the mpNG method to estimate the dN/dS ratio of each human gene based on the mutation profiles of cancer somatic mutations and germline substitutions. In contrast to the NG86 method [27], our method considered the substitution rate difference and took the overall mutation profile as the weight matrix (Figure 1).

We calculated the expected number of nonsynonymous and synonymous sites based on the exonic mutation profiles and counted the number of nonsynonymous and synonymous substitutions in protein-coding region of each human gene for all cancer/normal samples. A χ^2 test was performed to identify the genes whose dN/dS values were significantly greater than one or less than one, which indicates positive or negative (purifying) selection, respectively. Of the 18,602 genes with at least one germline substitution and cancer somatic substitution, the overall dN/dS value for cancer somatic substitutions (mean±s.e.=1.367±0.009) is much greater than that of germline substitutions (mean±s.e.=0.903±0.006) (Wilcoxon test, $P<10^{-16}$) (Table 1, Supplementary Table S1). In cancer genomes, 1,230 genes have dN/dS values significantly greater than one and 326 genes have dN/dS values significantly less than one (χ^2 test, $P<0.05$). In contrast, the germline substitutions include only 306 genes with dN/dS values significantly greater than one, whereas 4,357 genes have dN/dS values significantly less than one (χ^2 test, P<0.05) (Table 1). Among these, 1,191 genes exhibit positive selection in cancer genomes but non-positive selection in germline genomes and 275 genes exhibit negative selection in cancer genomes but non-negative selection in germline genomes. These genes may therefore be under different selective pressure in cancer and normal genomes.

Considering that different models might provide varying estimates, we used the NG86 method [27] as the simplest model to calculate the numbers of nonsynonymous and synonymous sites. The overall dN/dS value for cancer somatic substitutions (mean±s.e.=0.990±0.006) is greater than that for germline substitutions (mean±s.e.=0.624±0.004) for the 18,602 genes (Supplementary Table S1), whereas it is less than that calculated using mpNG method (Wilcoxon test, $P<10^{-16}$) (Table 1). Consequently, for both germline and cancer somatic substitutions, the number of genes with dN/dS values >1 (χ^2 test, P<0.05) is much lower, whereas the number of genes with dN/dS values <1 (χ^2 test, P<0.05) is much greater, than those calculated using the exonic mutation profiles (Table 1). We further used the intergenic and intronic somatic mutation profiles of 507 cancer samples with whole-genome sequences (WGSs) within the 7,042 tumor-normal pairs as a contrast. The overall dN/dS values calculated using these mutation profiles were between those obtained using the NG86 method and the exonic mutation profiles, as were the number of genes under positive and negative selection (Table 1, Table S1). Different models show different properties of single-nucleotide substitution, which resulted in the different list of candidate genes under positive and negative selection. However, the genes under positive and negative selection calculated using different models are almost overlapped (Figure 3A,B). The NG86 method ignores the mutation rate bias between different substitution types, leading to underestimation of the dN/dS ratio. Therefore, the NG86 method is strict with regard to detecting positive selection, and is relaxed about detection of negative selection [31]. Whereas the mpNG method takes the mutation bias, which can be depicted as the internal variance between mutation rates of different substitution types, into consideration. Thus, the mpNG method could recover the underestimation of the true dN/dS ratio estimated by NG86 method, while would increase the sampling errors and false discovery rates (FDRs). It would increase the false positive results for detecting positively selected genes, whereas be more conserved for detecting negatively selected genes. The mutation bias does not affect the detection of genes under strong selection pressure, while may affect the detection of genes under weak selection pressure. The mutation bias could be depicted by the internal variance of different substitution types. The exonic mutation profile has greater internal variance (σ=0.015) than that of intronic (σ=0.008) and intergenic (σ=0.008) mutation profiles, leading to the maximum estimation of dN/dS ratios.

Regardless of the method used to calculate the dN/dS values for germline and cancer somatic substitutions, we found that the dN/dS value for cancer somatic substitutions is much greater than that for germline substitutions. Previous studies have attributed the elevated dN/dS values to the relaxation of purifying selection [14] or the increased positive selection of globally expressed genes [15]. Our results show that the number of genes under positive selection increased, whereas the number of genes under negative selection decreased in cancer genomes compared with germline genomes. This result indicates that both the relaxation of purifying selection on passenger mutations and the positive selection of driver mutations may contribute to the increased dN/dS values of human genes in cancer genomes.

Relaxation of purifying selection for human genes in cancer cells

In this study, we used the mpNG method with exonic mutation profiles for estimation the dN/dS values for germline substitutions and cancer somatic mutations. The Cancer Gene Census [32, 33] contains more than 500 cancer genes that have been reported in the literatures to exhibit mutations and that are causally implicated in cancer development, of which 503 genes were included in the 18,602 genes we tested. These known cancer genes have significantly lower dN/dS values for germline substitutions (Wilcoxon test, $P<10^{-16}$), but slightly greater dN/dS values (Wilcoxon test, P=0.01) for cancer somatic mutations than those of other genes (Table 2A). For selection over longer time scales, we extracted the dN/dS values between human-mouse orthologs from the Ensembl database (Release 73) [34, 35]. The known cancer genes have significantly lower human-mouse dN/dS values than other human genes. Among the cancer genes, oncogenes (OGs) have significantly lower dN/dS values than non-cancer genes (Wilcoxon test, $P<10^{-15}$), whereas the mean dN/dS values of tumor suppressor genes (TSGs) are not significantly different from those of non-cancer genes (Wilcoxon test, P=0.89). These results support the work of Thomas *et al.* [36], who showed that known cancer genes may be more constrained and more important than other genes at the species and population levels, especially for oncogenes. In contrast, known cancer genes are more likely to gain functional somatic mutations in cancer relative to all other genes. However, within the known cancer genes, only 53 genes exhibited positive selection (χ^2 test, P<0.05) for cancer somatic substitutions, which suggests that positive selection for driver mutations is obscured by the relaxed purifying selection of passenger mutations.

We also examined human essential genes [37] and cancer common essential genes [21]. We extracted 2,452 human essential genes from DEG10 (the Database of Essential Genes) [37]. These genes are critical for cell survival, and are of cause more conserved than other genes in species and population levels. Human essential genes have significantly lower dN/dS values of human-mouse orthologs and germline substitution, while similar dN/dS values for cancer somatic mutations, comparing to non-essential genes (Table 2A). Cancer essential genes were identified by performing genome-scale pooled RNAi screens. RNAi screens with the 45ksh RNA pool in 12 cancer cell lines, including small-cell lung cancer, non-small-cell lung cancer, glioblastoma, chronic myelogenous leukemia, and lymphocytic leukemia, revealed 268 common essential genes [21]. These cancer essential genes also have significantly lower dN/dS values of human-mouse orthologs and germline substitutions, while similar dN/dS values for cancer somatic mutations, comparing to other human genes (Table 2A).

The cancer positively selected genes showed similar pattern with the cancer genes, cancer common essential genes, and human essential genes. These genes have lower dN/dS values for human-mouse orthologs (Wilcoxon test, $P=4.5\times10^{-4}$) and germline substitutions (Wilcoxon test, P=0.01), while significantly greater dN/dS values for cancer somatic mutations (Wilcoxon test, $P<10^{-16}$). However, the cancer negatively selected genes showed the different pattern, which have greater dN/dS values for human-mouse orthologs (Wilcoxon test, $P=7.3\times10^{-4}$) and germline substitutions (Wilcoxon test, $P=2.3\times10^{--4}$), while significantly lower dN/dS values for cancer somatic mutations (Wilcoxon test, $P<10^{-16}$). These results indicate that the positively selected genes may include the cancer associate genes or human essential genes, while the negatively selected genes may include genes strengthened selective constraint in cancer cells than that in normal cells.

We further tested the correlation of dN/dS values of human genes for human-mouse orthologs, germline substitutions and cancer somatic mutations, to compare selective pressures among species, population and cancers (Table 2B). For different gene sets, the dN/dS values between human-mouse orthologs show a weak positive correlation with those of germline substitutions, but no correlation with those of cancer somatic substitutions. The dN/dS values for human germline and cancer somatic substitutions show different correlation patterns between different gene sets. The tumor suppressor genes and cancer positively selected genes show weak positive correlation, while other gene sets have no correlation.

Roles of cancer positively and negatively selected genes in cancer cells

We then tested the genes under positive or purifying selection for their roles in cancer. Functional annotation analysis based on the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 [38, 39] showed an enrichment of genes involved in cell morphogenesis and pathways in cancer for cancer positively selected genes (Table 3A), while an enrichment of genes involved in sensory perception for cancer negatively selected genes (Table 3B) . As we only used a relaxed filter ($P<0.05$) for detecting cancer positively or negatively genes, which would lead to high FDRs. We further calculated the FDR for each P-value, using the qvalue (Table S1) [40]. We set the strengthened filter for detecting positively and negatively selected genes as $P<10^{-3}$ and $FDR<0.25$. There are only 61 genes meeting this requirement, which include 45 cancer positively selected genes and 16 cancer negatively selected genes (Table S2).

Among the 45 cancer positively selected genes, there are three oncogenes (GANP, NFE2L2, RHOA) and five tumor suppressor genes (TP53, CSMD1, CDKN2A and SPOP), according to the Cancer Gene Census [32]. Fourteen out of these genes are human essential genes, and seven genes are orthologs of mouse or yeast essential genes, according to the DEG10 [37]. Besides these genes, four cancer positively selected genes (IKBIP, TEX13A, FZD10 and PGAP2) also have dN/dS values significantly greater than one ($P<0.01$, $FDR<0.05$) for germline substitutions. Six genes show negative selection ($P<0.01$, $FDR<0.5$) in germline substitutions. Among these six genes, CAMD2, CSMD1 and CSMD3 have been reported as candidate tumor suppressor genes [41-44]. ACTG1 is associated with cancer cell migration [45]. There are also thirteen cancer positively selected genes show neutral selection for germline substitutions. It would be interesting to investigate the roles to cancer of these cancer-specific positively selected genes and four human essential genes which are not identified to be cancer related genes.

Among the 16 cancer negatively selected genes, there are two human essential genes, which are oncogene (FUS) and tumor suppressor gene (APC). These two genes are also under negative selection ($P<0.02$, $FDR<0.06$) for germline substitutions. The BRCA1 mutations, which would increase cancer risk for breast and ovarian cancer, can also be germline mutations, besides somatic mutations [46]. The other thirteen genes show strengthened selective constraint in cancer cells than in normal cells. It would attract much interest to uncover the roles of these evolutionary conserved genes in cancer cells. Several of these genes were reported to be required for the survival and proliferation of cancer cells and might therefore serve as potential drug targets or prognostic biomarkers. For example, BCL2L12 is a member of the BCL2 family and is an anti-apoptotic factor that can inhibit the p53 tumor suppressor as well as caspases 3 and 7 [47, 48]. Overexpression of BCL2L12 has been detected in several cancer types, and BCL2L12 can be considered as a molecular prognostic biomarker in these cancers [49-52]. MAP4 is a major non-neuronal microtubule-associated protein that promotes microtubule assembly. Ou *et al.* have reported that the protein level of MAP4 is positively correlated with the bladder cancer grade. And silencing MAP4 can efficiently disrupt the microtubule cytoskeleton, inhibiting the invasion and migration of bladder cancer cells [53]. EPPK1 is a member of plakin family, which plays a role in the organization of cytoskeletal architecture. Guo *et al.* identified EPPK1 as a predictive plasma biomarker for cervical cancer by proteomics [54]. These cancer-specific negatively selected genes are more conserved in cancer cells than in normal cells, indicating they may be crucial for the basic cellular processes of cancer cells.

Discussion

A key goal of cancer research is to identify cancer-related genes, such as OGs and TSGs, whose mutation might promote the occurrence and progression of tumors [28]. There are also cancer essential genes that are important for the growth and survival of cancer cells [21]. Different methods are needed to identify different types of cancer-related genes. In contrast to recent studies focused on the detection of driver mutations [16-18, 55], we aimed to detect cancer essential genes using a molecular evolution approach. Advances in the understanding of positively selected cancer drivers, as well as the severe side effects of classical chemotherapy and radiation therapies that target DNA integrity and cell divisions, have fueled efforts to develop anticancer drugs with more precise molecular targeting and fewer side effects. Though personalized therapeutic approaches that target genetically activated drivers have greatly improved patient outcomes in a number of common and rare cancers, the rapid acquisition of drug resistance due to high intra-tumor heterogeneity is becoming a challenging problem [56]. In other words, driver mutations may differ considerably among tumor sub-clones. Instead of looking for cancer-causing genes with multiple driver mutations, an alternative approach is to identify cancer essential genes that are highly conserved in tumor cell populations because they are crucial for carcinogenesis, progression and metastasis. To some extent, this idea may overcome drug resistance in targeted cancer therapies, as mutations in cancer essential genes are deleterious in tumor populations.

Several approaches can be utilized to identify cancer essential genes suitable for targeting with drugs, including siRNA-mediated knockdown of specific components and genetic tumor models. The genome-wide pooled shRNA screens promoted by the RNAi Consortium [57], however, can only be performed in cell lines in vitro and are limited to the analysis of genes important for proliferation and survival [21, 58, 59]. Thus, these screens will miss certain classes of genes that may function only in the proper *in vivo* tumor environment. Furthermore, siRNA screens may not be sensitive to target genes whose products are components of the cellular machinery. These types of targets may be frequently stabilized by their participation in complexes with a long biological half-life. Indeed, this longevity may be the reason why not all such targets seem to be essential for cancer cells in standard short-term siRNA screens [8]. Genetic tumor models can also enable screening strategies within an entire organism to identify cancer essential genes. However, this method is not suitable for large-scale screening. With the explosive increase in cancer somatic mutation data from cancer genome sequencing, it is now possible to investigate the natural selection of each human gene in cancer genomes using evolutionary genomics methods [8]. One major aim is to identify genes that was significantly strengthened purifying selective constraint from normal to cancer cells, which would suggest that these genes are cancer-specific essential genes.

Through analyses of large-scale cancer somatic mutation data derived from The Cancer Genome Atlas (TCGA) or International Cancer Genome Consortium (ICGC), previous studies found important differences between the evolutionary dynamics of cancer somatic cells and whole organisms [14, 6, 16]. However, these studies applied canonical nucleotide substitution models to identify the molecular signatures of natural selection in cancer cells or human populations, which neglected the apparently different mutation profiles between these cell types. Here, we developed a new mutation-profile-based Nei-Gojobori method (mpNG) to calculate the dN/dS values of 18,602 human genes for both cancer somatic and normal human germline substitutions.

Two prerequisites are crucial to apply the mpNG method properly. First, a large number of samples with similar mutation profiles is necessary to increase the power of selection pressure detection. Second, a subset of nucleotide substitutions should be chosen to represent the background neutral mutation profiles among the samples. In this study, because of the limitation of the number of cancer samples, especially the number of whole-genome sequenced cancer-normal tissue pairs, we pooled all the samples to analyze pan-cancer-level selection pressures. Mutation profiles are well known to be heterogeneous, even for samples with the same tissue origin [28, 17]. As an increasing number of cancer genomes are sequenced in the near future, we can classify cancer samples by their specific mutation profiles and infer evolutionarily selective pressures using the mpNG method. With respect to background neutral mutation profiles, it will be appropriate to calculate them based on intergenic regions from the corresponding samples. However, only a small number of cancer-normal paired WGSs are currently available. Therefore, in this study, we assume that most exonic somatic mutations in the cancer samples do not have significant effects on the fitness of cancer cells. Under this assumption, we can apply the mutation profiles of WESs to approximate the background. The exonic mutation profiles used in our mpNG method consider the weight of the 96 substitution classifications within the cancer exomes, which may reflect the mutation bias of different substitution types within the protein-coding regions. This method would recover the underestimation of the dN/dS value by the NG86 method [31]. Using the mpNG method, the detection of positive selection would be relaxed, whereas the detection of negative selection would be conservative when comparing to the NG86 method. Were more cancer-normal WGSs available, it would be better to choose suitable mutation profiles for the mpNG method. With the expansion of these data in the future, we may apply more precise methods to identify neutral background mutation properties.

As a conservative estimate of positively and negatively selected genes in cancer, we found 45 genes under intensified positive selection and 16 genes under strengthened purifying selection in cancer cells compared with germline cells. The cancer-specific positively selected genes are enriched for known cancer genes and/or human essential genes, while several cancer-specific negatively selected genes have been reported as prognostic biomarkers for cancers. As cancer-specific negatively selected genes are more evolutionarily constrained in cancer cells than in normal cells, identification of cancer-specific negatively selected genes would inform the potential resource of therapeutic targets or diagnostic biomarkers for cancers. However, cancer somatic mutations vary greatly among different cancer types and even among individual cancer genomes [28, 60, 17, 18], further studies will be needed to better understand the evolution of human cancer.

Methods

Datasets

Cancer somatic mutation data from 7,042 primary cancers corresponding to 30 different classes were extracted from the work of Alexandrov *et al.* [28], which includes 4,938,362 somatic substitutions and small insertions/deletions from 507 WGSs and 6,535 WESs. Data on human rare protein-coding variants (minor allele frequency < 0.01%) from 6,500 human WESs (ESP6500) were extracted from the ANNOVAR database [61] based on the NHLBI GO Exome Sequencing Project. A total of 522 known cancer genes were extracted from the Cancer Gene Census (http://cancer.sanger.ac.uk/cancergenome/projects/census/, COSMIC v68) [32, 33].

Sequences and annotations of human genes were extracted from the Ensembl database (Release 73) [34, 35]. For each gene, we only chose the longest sequence to avoid duplicate records of each single substitution. The HGNC (HUGO Gene Nomenclature Committee) database [62] (http://www.genenames.org/) and the

Genecards database [63] (http://www.genecards.org) were also used to map the gene IDs from different datasets. DAVID (Database for Annotation, Visualization and Integrated Discovery) v6.7 was utilized for the functional annotation analysis [38, 39].

Calculating mutation rate profiles

We calculated the mutation rate profiles using the 96 substitution classifications [28, 29], which not only show the base substitution but also include information on the sequence context of each mutated base. We counted all the somatic substitutions in protein-coding regions of the 7,042 cancer-normal paired WESs as well as all the protein-coding variants of the ESP6500 data set. We also count the total number of each trinucleotide type for the exonic, intronic, and intergenic regions in human genome. We calculated the mutation rate of each substitution type as the number of substitution per trinucleotide type per patient. The mutation profiles were depicted as the mutation rate of each mutation type according to the 96 substitution classifications.

Detection of positive and negative selections

ANNOVAR was utilized to perform biological and functional annotations of the cancer somatic mutations and germline substitutions [61]. Substitutions within protein-coding genes were classified as nonsynonymous and synonymous. We counted the numbers of nonsynonymous (n) and synonymous (s) substitutions for each gene among all the somatic mutations of 7,042 cancer-normal pairs. Somatic mutations at the same site and with the same mutation type that occurred in different patients were counted as different substitutions, as these substitutions, unlike germline evolution, occurred independently.

We further calculated the numbers of nonsynonymous (N) and synonymous (S) sites in each human protein-coding gene utilizing different models. The simple method of Nei and Gojobori was used [27]. We also considered cancer somatic mutation profiles, which were depicted as the percentage of each mutation type according to the 96 substitution classifications. For each gene, we calculated the proportion of substitutions that would be nonsynonymous or synonymous for each protein-coding site, as the probability of mutation types for each site was determined according to the mutation profiles. Then, we added up the proportions to calculate the total number of nonsynonymous (N) and synonymous (S) sites for each gene.

After counting the numbers of nonsynonymous (n) and synonymous (s) substitutions, as well as the numbers of nonsynonymous (N) and synonymous (S) sites for each gene, we calculated the ratio of the rates of nonsynonymous and synonymous substitutions (dN/dS) for each human gene as follows: $\frac{dN}{dS} = \frac{n/\mathrm{N}}{(\mathrm{s}+0.5)/(S+0.5)}$.

The dN/dS for germline substitutions was calculated using the same approach.

A χ^2 test was used to compare the numbers of nonsynonymous and synonymous substitutions to the numbers of nonsynonymous and synonymous sites for each gene in order to test the statistical significance of the difference between the dN/dS values and one. The genes with dN/dS values significantly greater than one were classified as being under positive selection in tumors, whereas the genes with dN/dS values significantly less than one were classified as being under negative, or purifying, selection. The false discovery rate was estimated using the qvalue package from Bioconductor [40]. A Wilcoxon test was performed to compare dN/dS values between cancer somatic substitutions and germline substitutions as well as between known cancer genes and all other genes. The software tool R was used for the statistical analysis (http://www.r-project.org/).

Acknowledgements

We are grateful to Leng Han, Siyuan Zheng, and Xiaopu Wang, for their help with the manuscript preparation. We would like to thank the NHLBI GO Exome Sequencing Project and its ongoing studies which produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926) and the Heart GO Sequencing Project (HL-103010). We also gratefully acknowledge the TCGA Research Network and the Broad Institute TCGA GDAC Firehose for referencing the TCGA datasets.

This work was supported by a grant from the Ministry of Science and Technology China (2012CB910101), grants from the National Natural Science Foundation of China (31272299, 31301034), a grant from the Zhejiang Provincial Natural Sciences Foundation of China (LY15C060001), grants from Fudan University and Iowa State University to XG, the Shanghai Pujiang Program (13PJD005) to ZS, and a General

Financial Grant from China Postdoctoral Science Foundation (2013M531117) to ZZ.

References

1. Cairns J. Mutation selection and the natural history of cancer. Nature. 1975;255(5505):197-200.

2. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;194(4260):23-8.

3. Crespi BJ, Summers K. Positive selection in the evolution of cancer. Biol Rev Camb Philos Soc. 2006;81(3):407-24.

4. Merlo LM, Pepper JW, Reid BJ, Maley CC. Cancer as an evolutionary and ecological process. Nat Rev Cancer. 2006;6(12):924-35.

5. Podlaha O, Riester M, De S, Michor F. Evolution of the cancer genome. Trends Genet. 2012;28(4):155-63.

6. Yates LR, Campbell PJ. Evolution of the cancer genome. Nat Rev Genet. 2012;13(11):795-806.

7. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012;481(7381):306-13.

8. Luo J, Solimini NL, Elledge SJ. Principles of cancer therapy: oncogene and non-oncogene addiction. Cell. 2009;136(5):823-37.

9. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011;144(5):646-74.

10. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458(7239):719-24.

11. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007;446(7132):153-8.

12. Beckman RA, Loeb LA. Negative clonal selection in tumor evolution. Genetics. 2005;171(4):2123-31.

13. McFarland CD, Korolev KS, Kryukov GV, Sunyaev SR, Mirny LA. Impact of deleterious passenger mutations on cancer progression. Proc Natl Acad Sci U S A. 2013;110(8):2910-5.

14. Woo YH, Li WH. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. Nat Commun. 2012;3:1004.

15. Ostrow SL, Barshir R, DeGregori J, Yeger-Lotem E, Hershberg R. Cancer Evolution Is Associated with Pervasive Positive Selection on Globally Expressed Genes. PLoS Genet. 2014;10(3):e1004239.

16. Davoli T, Xu AW, Mengwasser KE, Sack LM, Yoon JC, Park PJ et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell. 2013;155(4):948-62.

17. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214-8.

18. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014;505(7484):495-501.

19. Chen H, Xing K, He X. The dJ/dS Ratio Test Reveals Hundreds of Novel Putative Cancer Drivers. Mol Biol Evol. 2015.

20. Jordan IK, Rogozin IB, Wolf YI, Koonin EV. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 2002;12(6):962-8.

21. Luo B, Cheung HW, Subramanian A, Sharifnia T, Okamoto M, Yang X et al. Highly parallel identification of essential genes in cancer cells. Proc Natl Acad Sci USA. 2008;105(51):20380-5.

22. Endo T, Ikeo K, Gojobori T. Large-scale search for genes on which positive selection may operate. Mol Biol Evol. 1996;13(5):685-90.
23. Messier W, Stewart CB. Episodic adaptive evolution of primate lysozymes. Nature. 1997;385(6612):151-4.
24. Arbiza L, Dopazo J, Dopazo H. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. PLoS Comput Biol. 2006;2(4):e38.
25. Ovens K, Naugler C. Preliminary evidence of different selection pressures on cancer cells as compared to normal tissues. Theor Biol Med Model. 2012;9:44.
26. Greenman C, Wooster R, Futreal PA, Stratton MR, Easton DF. Statistical analysis of pathogenicity of somatic mutations in cancer. Genetics. 2006;173(4):2187-98.
27. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 1986;3(5):418-26.
28. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415-21.
29. Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM et al. A framework for the interpretation of de novo mutation in human disease. Nat Genet. 2014;46(9):944-50.
30. Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP et al. A landscape of driver mutations in melanoma. Cell. 2012;150(2):251-63.
31. Yang Z, Bielawski JP. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 2000;15(12):496-503.
32. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R et al. A census of human cancer genes. Nat Rev Cancer. 2004;4(3):177-83.
33. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 2011;39(Database issue):D945-50.
34. Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford). 2011;2011:bar030.
35. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S et al. Ensembl 2014. Nucleic Acids Res. 2014;42(Database issue):D749-55.
36. Thomas MA, Weston B, Joseph M, Wu W, Nekrutenko A, Tonellato PJ. Evolutionary dynamics of oncogenes and tumor suppressor genes: higher intensities of purifying selection than other genes. Mol Biol Evol. 2003;20(6):964-8.
37. Luo H, Lin Y, Gao F, Zhang CT, Zhang R. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. Nucleic Acids Res. 2014;42(Database issue):D574-80.
38. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57.
39. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13.
40. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100(16):9440-5.
41. Kamal M, Shaaban AM, Zhang L, Walker C, Gray S, Thakker N et al. Loss of CSMD1 expression is associated with high tumour grade and poor survival in invasive ductal breast carcinoma. Breast Cancer Res Treat. 2010;121(3):555-63.

42. Zhang R, Song C. Loss of CSMD1 or 2 may contribute to the poor prognosis of colorectal cancer patients. Tumour Biol. 2014;35(5):4419-23.
43. Chang G, Xu S, Dhir R, Chandran U, O'Keefe DS, Greenberg NM et al. Hypoexpression and epigenetic regulation of candidate tumor suppressor gene CADM-2 in human prostate cancer. Clin Cancer Res. 2010;16(22):5390-401.
44. He W, Li X, Xu S, Ai J, Gong Y, Gregg JL et al. Aberrant methylation and loss of CADM2 tumor suppressor expression is associated with human renal cell carcinoma tumor progression. Biochem Biophys Res Commun. 2013;435(4):526-32.
45. Luo Y, Kong F, Wang Z, Chen D, Liu Q, Wang T et al. Loss of ASAP3 destabilizes cytoskeletal protein ACTG1 to suppress cancer cell migration. Mol Med Rep. 2014;9(2):387-94.
46. Roy R, Chun J, Powell SN. BRCA1 and BRCA2: different roles in a common pathway of genome protection. Nat Rev Cancer. 2012;12(1):68-78.
47. Stegh AH, Kesari S, Mahoney JE, Jenq HT, Forloney KL, Protopopov A et al. Bcl2L12-mediated inhibition of effector caspase-3 and caspase-7 via distinct mechanisms in glioblastoma. Proc Natl Acad Sci U S A. 2008;105(31):10703-8.
48. Stegh AH, Brennan C, Mahoney JA, Forloney KL, Jenq HT, Luciano JP et al. Glioma oncoprotein Bcl2L12 inhibits the p53 tumor suppressor. Genes Dev. 2010;24(19):2194-204.
49. Thomadaki H, Floros KV, Pavlovic S, Tosic N, Gourgiotis D, Colovic M et al. Overexpression of the novel member of the BCL2 gene family, BCL2L12, is associated with the disease outcome in patients with acute myeloid leukemia. Clin Biochem. 2012;45(16-17):1362-7.
50. Karan-Djurasevic T, Palibrk V, Zukic B, Spasovski V, Glumac I, Colovic M et al. Expression of Bcl2L12 in chronic lymphocytic leukemia patients: association with clinical and molecular prognostic markers. Med Oncol. 2013;30(1):405.
51. Foutadakis S, Avgeris M, Tokas T, Stravodimos K, Scorilas A. Increased BCL2L12 expression predicts the short-term relapse of patients with TaT1 bladder cancer following transurethral resection of bladder tumors. Urol Oncol. 2014;32(1):39 e29-36.
52. Tzovaras A, Kladi-Skandali A, Michaelidou K, Zografos GC, Missitzis I, Ardavanis A et al. BCL2L12: a promising molecular prognostic biomarker in breast cancer. Clin Biochem. 2014;47(18):257-62.
53. Ou Y, Zheng X, Gao Y, Shu M, Leng T, Li Y et al. Activation of cyclic AMP/PKA pathway inhibits bladder cancer cell invasion by targeting MAP4-dependent microtubule dynamics. Urol Oncol. 2014;32(1):47.e21-e8.
54. Guo X, Hao Y, Kamilijiang M, Hasimu A, Yuan J, Wu G et al. Potential predictive plasma biomarkers for cervical cancer by 2D-DIGE proteomics and Ingenuity Pathway Analysis. Tumour Biol. 2015;36(3):1711-20.
55. Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. Cell. 2014;156(6):1324-35.
56. Dobbelstein M, Moll U. Targeting tumour-supportive cellular machineries in anticancer drug development. Nat Rev Drug Discov. 2014;13(3):179-96.
57. Root DE, Hacohen N, Hahn WC, Lander ES, Sabatini DM. Genome-scale loss-of-function screening with a lentiviral RNAi library. Nat Methods. 2006;3(9):715-9.
58. Koh JL, Brown KR, Sayad A, Kasimer D, Ketela T, Moffat J. COLT-Cancer: functional genetic screening resource for essential genes in human cancer cell lines. Nucleic Acids Res. 2012;40(Database issue):D957-63.

59. Marcotte R, Brown KR, Suarez F, Sayad A, Karamboulas K, Krzyzanowski PM et al. Essential gene profiles in breast, pancreatic, and ovarian cancer cells. Cancer Discov. 2012;2(2):172-89.

60. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502(7471):333-9.

61. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.

62. Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2013. Nucleic Acids Res. 2013;41(Database issue):D545-52.

63. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M et al. GeneCards Version 3: the human gene integrator. Database (Oxford). 2010;2010:baq020.

Figure Legends

Figure 1. Mutation profiles of cancer somatic substitutions and germline substitutions, including the exonic mutation profile of 7,042 cancer samples, the exonic mutation profile of ESP6500, the intronic mutation profile of 507 cancer whole genomes, the intergenic mutation profile of 507 cancer whole genomes, and the exonic mutation profiles of breast carcinoma (BRCA), lung adenocarcinoma (LUAD), colon adenocarcinoma (COAD), and skin cutaneous melanoma (SKCM).

Figure 2. The pipeline used to identify cancer positively and negatively selected genes with mpNG method.

Figure 3. The overlap of positive selection (A) and negative selection (B) genes based on different models.

Tables

Table 1. The dN/dS values and number of human genes under positive or negative selection in germline and cancer based on NG86 and mpNG methods with different mutation profiles. P-values are according to a χ^2 test.

	dN/dS	# Positive selection*	# Negative selection*
Germline (NG86)	0.624 ± 0.004	42	9093
Cancer (NG86)	0.990 ± 0.006	306	2330
Cancer (intergenic)	1.240 ± 0.008	697	722
Cancer (intronic)	1.281 ± 0.008	822	624
Germline (exonic)	0.903 ± 0.006	264	4357
Cancer (exonic)	1.367 ± 0.009	1230	326

Note: *P<0.05

Table 2. The dN/dS values (A) and Correlation of dN/dS values (B) of different gene sets for human-mouse orthologs, germline and cancer somatic substitutions.

(A)

	Human-Mouse	Germline	Cancer
All genes	0.155 ± 0.006	0.903 ± 0.006	1.367 ± 0.009
Known cancer genes	0.111 ± 0.005	0.675 ± 0.017	1.350 ± 0.033
Oncogenes	0.101 ± 0.006	0.665 ± 0.020	1.336 ± 0.038
Tumor suppressor genes	0.151 ± 0.014	0.732 ± 0.039	1.350 ± 0.066
Human essential genes	0.093 ± 0.002	0.704 ± 0.013	1.288 ± 0.015
Cancer essential genes	0.089 ± 0.007	0.698 ± 0.032	1.413 ± 0.067
Positively selected genes	0.136 ± 0.004	0.918 ± 0.029	3.216 ±0.091
Negatively selected genes	0.172 ± 0.008	0.915 ± 0.023	0.479 ±0.009

(B)

	Human-Mouse vs Germline		Human-Mouse vs Cancer		Germline vs Cancer	
	r	P-Value	r	P-Value	r	P-Value
All genes	0.04	3.3×10^{-7}	-0.01	0.47	0.10	$<10^{-16}$
Known cancer genes	0.45	$<10^{-16}$	-0.02	0.72	0.11	0.02
Oncogenes	0.43	$<10^{-16}$	-0.01	0.85	0.04	0.43
Tumor suppressor genes	0.52	1.0×10^{-8}	0.04	0.66	0.36	$1.6\times10^{--4}$
Human essential genes	0.19	$<10^{-16}$	-0.05	0.01	0.06	1.4×10^{-3}
Cancer essential genes	0.30	5.7×10^{-7}	-0.07	0.29	0.03	0.65
Positively selected genes	0.17	2.4×10^{-9}	-0.02	0.57	0.23	$<10^{-16}$
Negatively selected genes	0.22	6.3×10^{-5}	0.10	0.07	0.04	0.60

Table 3. Functional enrichment of positively and negatively selected genes in cancer genomes (P<0.01, FDR<10%).

(A)

Category	Term	P-Value	FDR (%)
GOTERM_BP_FAT	GO:0032989~cellular component morphogenesis	7.42×10^{-4}	1.34
GOTERM_BP_FAT	GO:0043009~chordate embryonic development	2.40×10^{-3}	4.28
GOTERM_BP_FAT	GO:0009792~embryonic development ending in birth or egg hatching	2.89×10^{-3}	5.13
GOTERM_BP_FAT	GO:0000902~cell morphogenesis	3.28×10^{-3}	5.80
GOTERM_BP_FAT	GO:0030098~lymphocyte differentiation	4.90×10^{-3}	8.55
GOTERM_BP_FAT	GO:0051276~chromosome organization	5.19×10^{-3}	9.02
KEGG_PATHWAY	hsa05200:Pathways in cancer	4.23×10^{-3}	0.52
KEGG_PATHWAY	hsa05215:Prostate cancer	5.88×10^{-4}	0.72
KEGG_PATHWAY	hsa05213:Endometrial cancer	1.46×10^{-3}	1.78
KEGG_PATHWAY	hsa05210:Colorectal cancer	2.27×10^{-3}	2.75
KEGG_PATHWAY	hsa05216:Thyroid cancer	2.74×10^{-3}	3.32

(B)

Category	Term	P-Value	FDR (%)
GOTERM_BP_FAT	GO:0007600~sensory perception	1.35×10^{-3}	2.20
GOTERM_BP_FAT	GO:0050890~cognition	3.11×10^{-3}	5.00
GOTERM_BP_FAT	GO:0007608~sensory perception of smell	4.27×10^{-3}	6.80
GOTERM_BP_FAT	GO:0007606~sensory perception of chemical stimulus	4.67×10^{-3}	7.41
KEGG_PATHWAY	hsa04740:Olfactory transduction	1.03×10^{-3}	1.11

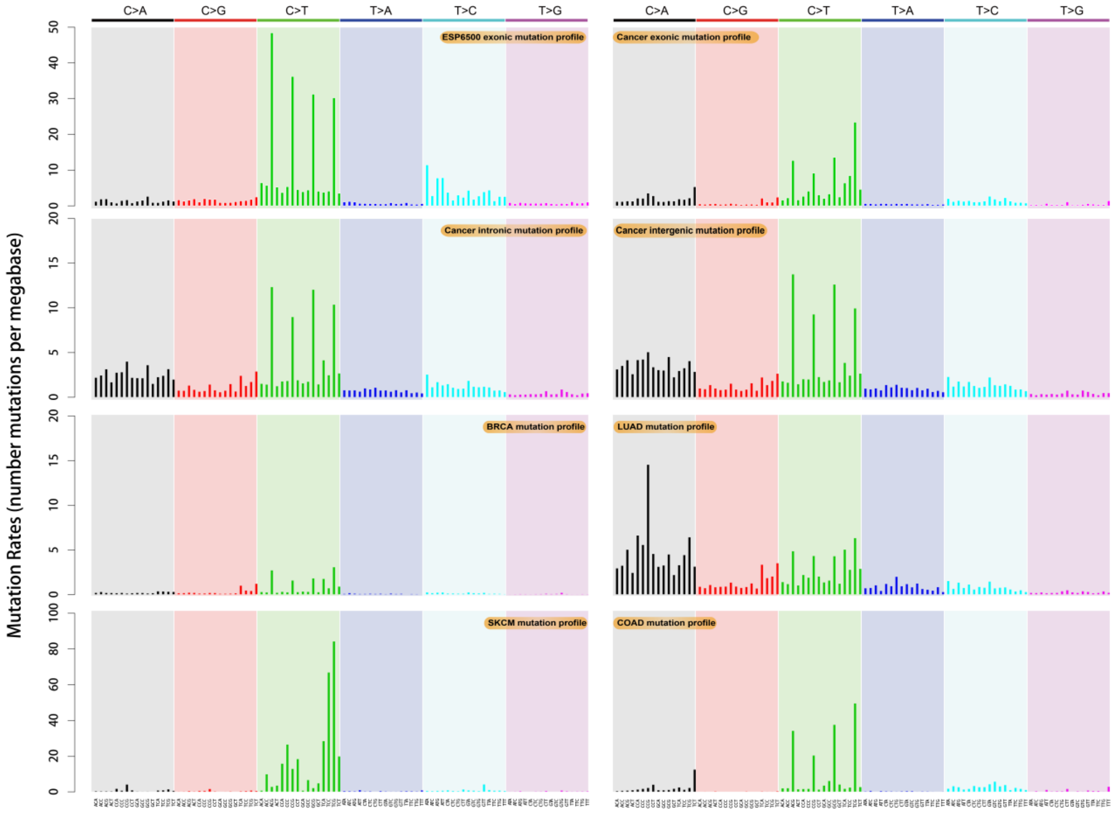
C>A
C>G
C>T
T>A
T>C
T>G
C>A
C>G
C>T
T>A
T>C
T>G
ESP6500 exonic mutation profile
Cancer exonic mutation profile
Cancer intronic mutation profile
Cancer intergenic mutation profile
BRCA mutation profile
LUAD mutation profile
SKCM mutation profile
COAD mutation profile
Mutation Rates (number mutations per megabase)
0
10
20
30
40
50
0
5
10
15
20
0
5
10
15
20
0
20
40
60
80
100

Figure 1

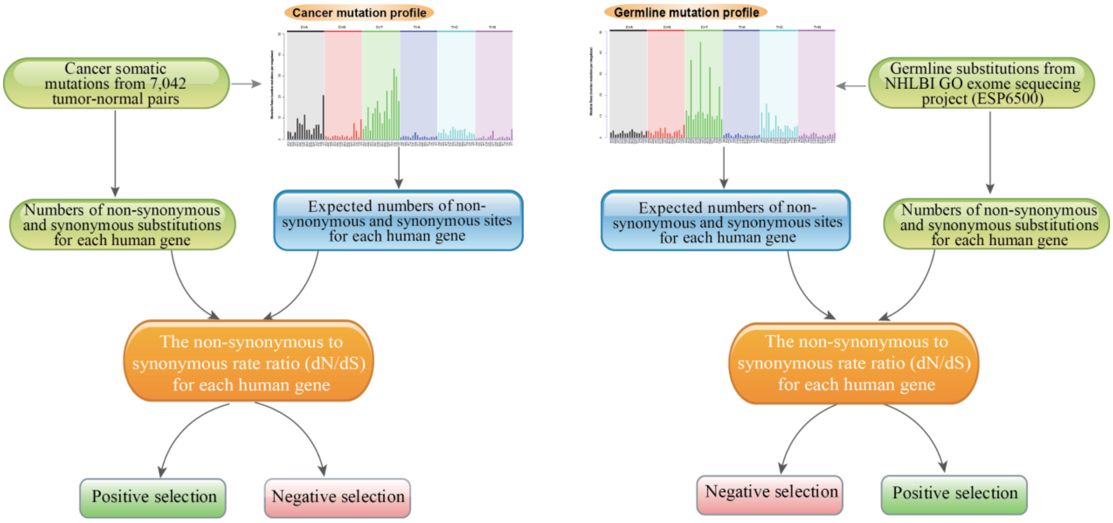
Cancer mutation profile
Germline mutation profile
Cancer somatic mutations from 7,042 tumor-normal pairs
Germline substitutions from NHLBI GO exome sequecing project (ESP6500)
Numbers of non-synonymous and synonymous substitutions for each human gene
Expected numbers of non-synonymous and synonymous sites for each human gene
Expected numbers of non-synonymous and synonymous sites for each human gene
Numbers of non-synonymous and synonymous substitutions for each human gene
The non-synonymous to synonymous rate ratio (dN/dS) for each human gene
The non-synonymous to synonymous rate ratio (dN/dS) for each human gene
Positive selection
Negative selection
Negative selection
Positive selection

Figure 2

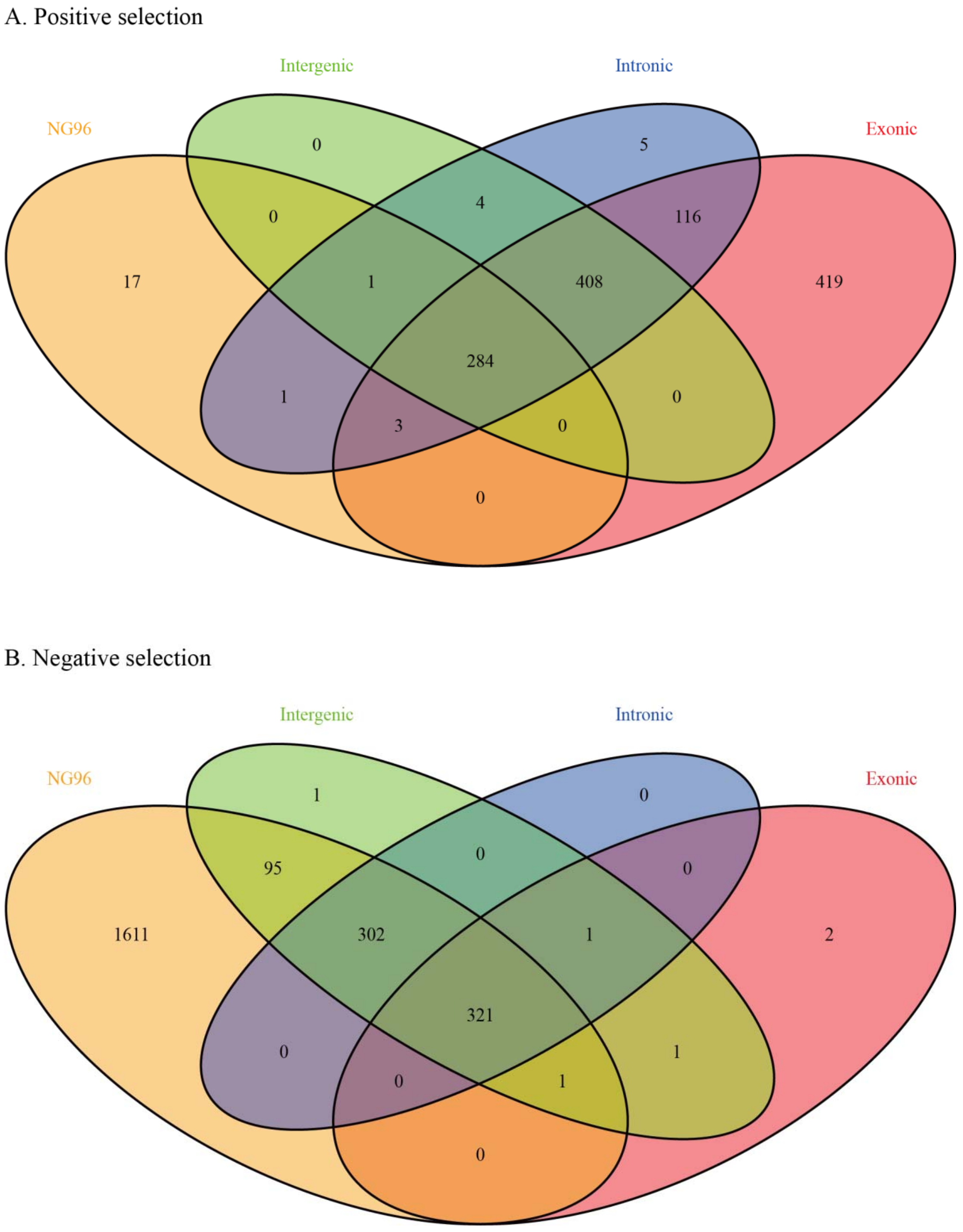
A. Positive selection
Intergenic
Intronic
NG96
Exonic
0
5
4
0
116
17
1
408
419
284
1
0
3
0
0
B. Negative selection
Intergenic
Intronic
NG96
Exonic
1
0
0
95
0
1611
302
1
2
321
0
1
0
1
0

Figure 3