Mutation-Profile-Based Methods for Understanding Selection Forces in Cancer

Somatic Mutations: A Comparative Analysis

Zhan Zhou1,4,†, Yangyun Zou2,†, Gangbiao Liu1, Jingqi Zhou1, Jingcheng Wu4,
Shimin Zhao1, Zhixi Su2,*, Xun Gu3,2,*

1State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai, China,
2Ministry of Education Key Laboratory of Contemporary Anthropology, Center for Evolutionary Biology, School of Life Sciences, Fudan University, Shanghai, China,
3Department of Genetics, Development and Cell Biology, Program of Bioinformatics and Computational Biology, Iowa State University, Ames, Iowa, USA,
4College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang, China

†These authors contributed equally to this work
*Corresponding authors

Xun Gu: xgu@iastate.edu
Zhixi Su: 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 both 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. 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. In cancer cells, we found a conservative estimate of 45 genes under intensified positive selection and 16 genes under strengthened purifying selection relative to germline cells. The cancer-specific positively selected genes were enriched for cancer genes and human essential genes, while several cancer-specific negatively selected genes were previously reported as prognostic biomarkers for cancers. Thus, our computation pipeline used to identify positively and negatively selected genes in cancer 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\textsuperscript{1,2}, the evolutionary concept of cancer progression has been widely accepted\textsuperscript{3-7}. In this model, 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\textsuperscript{8,9}. Therefore, knowledge of evolutionary dynamics will benefit our understanding of cancer initiation and progression. For example, there are two types of somatic mutations in cancer genomes: driver mutations and passenger mutations\textsuperscript{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\textsuperscript{12,13}. Between these two extremes 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 are much stronger on cancer cells than on germline cells\textsuperscript{14,15}. Given that many of the genes positively selected for 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\textsuperscript{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. Because essential genes are more evolutionarily conserved\textsuperscript{20}, it
would be feasible to identify cancer essential genes from the genes that are evolutionarily conserved in cancer cells. Because cancer essential genes may not be the driver genes for carcinogenesis but are crucial for cancer cell proliferation and survival\textsuperscript{21}, using evolutionary conservation to identify relevant genes may be advantageous in addressing therapeutic issues related to drug resistance, especially in cancers with high intratumor heterogeneity.

Many previous studies have used the ratio of nonsynonymous to synonymous substitution rates (dN/dS) to identify genes that might be under strong positive selection both in organismal evolution and tumorigenesis\textsuperscript{11,14,15,22-24}. However, most of these studies applied well-known methods that are usually based on simple nucleotide mutation/substitution models, where every mutation or substitution pattern has the same probability \textsuperscript{25}. Unfortunately, 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\textsuperscript{15,26}. In addition, context-dependent mutation bias (i.e., base-substitution profiles that consider the flanking 5’ and 3’ bases of each mutated base) should be taken into consideration\textsuperscript{26,27}.

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, the 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\textsuperscript{25}. This assumption can lead to nontrivial biased estimations when it is significantly violated. In contrast, 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 of cancer genomes and human populations are different

Estimating evolutionary selective pressure on human genes is a practical method for 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 be able to 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, 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. We used these data to compare 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 base. We then depicted the mutation profiles as 96 substitution classifications\textsuperscript{26,27}. The mutation profiles exhibit the prevalence of each substitution pattern for somatic point mutations, which present not only the substitution types but also the sequence context (see Methods). The exonic mutation profiles of cancer somatic substitutions and germline substitutions differed from one another, and the intronic and intergenic mutation profiles were quite different from the exonic mutation profile of cancer cells (Fig. 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 their mutation rates but also in their mutation patterns. Specifically, the mutation rate of SKCM was much higher than that of the other three types. Additionally, the mutation profiles of SKCM were highly enriched in the C-to-T substitution pattern (Fig. 1). These data indicated a direct mutagenic role for ultraviolet (UV) light in SKCM pathogenesis\textsuperscript{28}. The different mutation profiles may lead to different biological progressions in carcinogenesis, which have been depicted in several publications\textsuperscript{17,26}. Thus, it is inappropriate to use conventional methods such as NG\textsuperscript{86} to measure selective pressure by means of dN/dS calculation because this approach ignores the mutation bias of different nucleotide substitution types.

Measuring selective pressure on human genes in cancer and germline cells using
We therefore formulated an evolutionary approach that was designed specifically to estimate the selective pressure imposed on human genes in cancer cells and then to identify genes that had undergone positive and purifying selection in cancer cells rather than in normal cells (see Fig. 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\textsuperscript{25}, our method considered the difference in substitution rate and took the overall mutation profile as the weight matrix (Fig. 1).

We calculated the expected number of nonsynonymous and synonymous sites based on the exonic mutation profiles. We then counted the number of nonsynonymous and synonymous substitutions in the protein-coding region of each human gene for all cancer somatic mutations or germline substitutions. A $\chi^2$ test was performed to identify the genes whose dN/dS values were either 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) was 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 the cancer genomes, 1,230 genes had dN/dS values significantly greater than one, and 326 genes had dN/dS values significantly less than one ($\chi^2$ test, $P<0.05$). In contrast, the germline substitutions included only 306 genes with dN/dS values...
significantly greater than one, whereas 4,357 genes had dN/dS values significantly
less than one ($\chi^2$ test, P<0.05) (Table 1). Of these cancer positively selected genes,
1,191 genes exhibited positive selection in cancer genomes but non-positive selection
in germline genomes. Additionally, 275 genes exhibited 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 $^{25}$ 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) was greater than that for germline substitutions
(mean±s.e.=0.624±0.004) for the 18,602 genes (Supplementary Table S1), whereas
the ratio was 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 ($\chi^2$ test, P<0.05) was much lower than those
calculated using the exonic mutation profiles, whereas the number of genes with
dN/dS values <1 ($\chi^2$ test, P<0.05) was much greater (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 the
values obtained using the NG86 method and the exonic mutation profiles, as was the
number of genes under positive and negative selection (Table 1, Supplementary Table
S1). Different models show different single-nucleotide substitution properties, which
resulted in a different list of candidate genes under positive and negative selection. However, the genes under positive and negative selection calculated using different models almost overlapped (Fig. 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, but it is relaxed about detection of negative selection. In contrast, 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 the NG86 method, which would increase the sampling errors and false discovery rates (FDRs). It would also increase the false positive results for detecting positively selected genes, but be more conservative in detecting negatively selected genes. The mutation bias does not affect the detection of genes under strong selection pressure, while it 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 had 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 was much greater than that for germline substitutions. Previous studies have attributed the elevated dN/dS values to the relaxation of purifying selection or...
the increased positive selection of globally expressed genes. 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 to
estimate the dN/dS values for germline substitutions and cancer somatic mutations.
The Cancer Gene Census contains more than 500 cancer genes that have been
reported in the literature to exhibit mutations and that are causally implicated in
cancer development. Of those, 503 genes were included in the 18,602 genes we tested.
These known cancer genes had 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). The known
cancer genes had significantly lower human-mouse dN/dS values than other human
genomes. Among the cancer genes, oncogenes (OGs) had 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) were not significantly different from those
of non-cancer genes (Wilcoxon test, P=0.89). These results support the work of
Thomas et al.\textsuperscript{34}, 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 ($\chi^2$ 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\textsuperscript{35} and cancer common essential genes\textsuperscript{21}.

We extracted 2,452 human essential genes from DEG10 (the Database of Essential
Genes)\textsuperscript{35}. These genes are critical for cell survival and are therefore more conserved
than other genes at species and population levels. Here, we found that human essential
genes had significantly lower dN/dS values of human-mouse orthologs and germline
substitution, and similar dN/dS values for cancer somatic mutations, relative to the
values for non-essential genes (Table 2A). Cancer essential genes were identified by
performing genome-scale pooled RNAi screens. RNAi screens with the 45k shRNA
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\textsuperscript{21}. Compared to other human genes, these
cancer essential genes had significantly lower dN/dS values of human-mouse
orthologs and germline substitutions, and similar dN/dS values for cancer somatic
mutations (Table 2A).
The cancer positively selected genes displayed a pattern similar to that of the cancer genes, cancer common essential genes, and human essential genes. These genes had lower dN/dS values for human-mouse orthologs (Wilcoxon test, $P=4.5\times10^{-4}$) and germline substitutions (Wilcoxon test, $P=0.01$), but significantly greater dN/dS values for cancer somatic mutations (Wilcoxon test, $P<10^{-16}$). However, the negatively selected cancer genes displayed a different pattern, with greater dN/dS values for human-mouse orthologs (Wilcoxon test, $P=7.3\times10^{-4}$) and germline substitutions (Wilcoxon test, $P=2.3\times10^{-4}$), and 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 associated genes or human essential genes, while the negatively selected genes may include genes under greater selective constraints in cancer cells than 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, in order to compare selective pressures among species, populations and cancers (Table 2B). For different gene sets, the dN/dS values between human-mouse orthologs showed a weak positive correlation with those of germline substitutions, but no correlation with the values for cancer somatic substitutions. The dN/dS values for human germline and cancer somatic substitutions displayed different correlation patterns between different gene sets. The tumor suppressor genes and positively selected cancer genes showed weak positive correlation, while other gene sets had no correlation.
Roles of cancer positively and negatively selected genes in cancer cells

We next 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\textsuperscript{36,37} showed an enrichment of genes involved in cell morphogenesis and pathways in cancer for positively selected genes (Table 3A). Additionally, we found an enrichment of genes involved in sensory perception for cancer negatively selected cancer genes (Table 3B). It is important to note that we only used a relaxed filter (P<0.05) for detecting cancer positively or negatively selected genes, which would lead to high FDRs. We further calculated the FDR for each P-value, using the qvalue (Supplementary Table S1)\textsuperscript{38}. We set the strengthened filter for detecting positively and negatively selected genes at P<10\textsuperscript{-3} and FDR<0.25. Only 61 genes met this requirement, which included 45 cancer positively selected genes and 16 cancer negatively selected genes (Supplementary Table S2).

Among the 45 cancer positively selected genes, there were three oncogenes (GANP, NFE2L2, RHOA) and five tumor suppressor genes (TP53, CSMD1, CDKN2A and SPOP), according to the Cancer Gene Census\textsuperscript{30}. Fourteen of the 61 genes are human essential genes, and seven are orthologs of mouse or yeast essential genes, according to the DEG10\textsuperscript{35}. In addition, four positively selected genes (IKBIP, TEX13A, FZD10 and PGAP2) also had dN/dS values significantly greater than one (P<0.01, FDR<0.05) for germline substitutions. Six genes showed negative selection (P<0.01, FDR<0.5) in germline substitutions. Among those six genes, CAMD2,
CSMD1 and CSMD3 have been reported as candidate tumor suppressor genes\textsuperscript{39-42}. Additionally, ACTG1 is associated with cancer cell migration\textsuperscript{43}. There were also 13 cancer positively selected genes displaying neutral selection for germline substitutions. It would be interesting to investigate the roles in cancer of these cancer-specific positively selected genes and the four human essential genes that are not known to be cancer-related.

Among the 16 cancer negatively selected genes, there were two human essential genes: an oncogene (FUS) and a tumor suppressor gene (APC). Both were also under negative selection for germline substitutions (P<0.02, FDR<0.06). BRCA1 mutations, which would increase cancer risk for breast and ovarian cancer, can be germline mutations as well as somatic mutations\textsuperscript{44}. The other 13 genes showed greater selective constraint in cancer cells than in normal cells. It would be quite valuable to uncover the roles of these evolutionarily 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 and caspases 3 and 7\textsuperscript{45,46}. Overexpression of BCL2L12 has been detected in several cancer types, and BCL2L12 is therefore considered a molecular prognostic biomarker in these cancers\textsuperscript{47-50}. MAP4 is a major non-neuronal microtubule-associated protein that promotes microtubule assembly. Ou \textit{et al.} have reported that the protein level of MAP4 is positively correlated with bladder cancer grade. Additionally, silencing MAP4 can efficiently disrupt the
microtubule cytoskeleton, inhibiting the invasion and migration of bladder cancer
cells. EPPK1 is a member of the plakin family, which plays a role in the
organization of cytoskeletal architecture. Guo et al. used proteomics to identify
EPPK1 as a predictive plasma biomarker for cervical cancer. These negatively
selected, cancer-specific 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, the mutation of which might promote the occurrence and progression of
tumors. There are also cancer essential genes that are important for the growth and
survival of cancer cells. Different methods are needed to identify different types of
cancer-related genes. In contrast to recent studies focused on the detection of driver
mutations, 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 division, have fueled efforts to develop
anticancer drugs with more precise molecular targeting and fewer side effects.
Although 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. In other words, driver mutations may differ considerably
between 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, however, can only be performed in cell lines in vitro and are

limited to the analysis of genes important for proliferation and survival. 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. 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. One major
aim is to identify genes that are under significantly increased purifying selective
constraint in cancer cells relative to normal cells, which would suggest that they 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\textsuperscript{6,14,16}. 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 properly apply the mpNG method. 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 of the samples. In this
study, because of the limited number of cancer samples, especially the number of
whole-genome sequenced tumor-normal tissue pairs, we pooled all of 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\textsuperscript{17,26}. As an increasing
number of cancer genomes are sequenced in the near future, we will be able to
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 tumor-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 that occurs with the NG86 method. Using the mpNG method, the detection of positive selection would be relaxed, whereas the detection of negative selection would be conservative relative to the NG86 method. Were more tumor-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 set of cancer-specific positively selected genes was enriched for known cancer genes and/or human essential genes, while several of the cancer-specific negatively selected
genes have previously been reported as prognostic biomarkers for cancers. Because 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 options for cancer therapeutic targets or diagnostic biomarkers. However, cancer somatic mutations vary greatly among different cancer types and even among individual cancer genomes^{17,18,26,57}. therefore, 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.*^{26}, which includes 4,938,362 somatic substitutions and small insertions/deletions from 507 WGSs and 6,535 WESs. Data on rare human protein-coding variants (minor allele frequency <0.01%) from 6,500 human WESs (ESP6500) were extracted from the ANNOVAR database^{58} 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)^{30,31}.

Human gene sequences and annotations were extracted from the Ensembl database (Release 73)^{32,33}. For each gene, we only chose the longest sequence to avoid duplicate records of each single substitution. The HGNC (HUGO Gene Nomenclature Committee) database^{59} (http://www.genenames.org/) and the Genecards database^{60}
were also used to map the gene IDs from different datasets. DAVID v6.7 was utilized for the functional annotation analysis\textsuperscript{36,37}.

### Calculating mutation rate profiles

We calculated the mutation rate profiles using the 96 substitution classifications\textsuperscript{26,27}, which not only show the base substitution but also include information on the sequence context of each mutated base. We counted all somatic substitutions in the protein-coding regions of the 7,042 tumor-normal paired WESs, as well as all the protein-coding variants of the ESP6500 dataset. We also counted the total number of each trinucleotide type for the exonic, intronic, and intergenic regions in the human genome. We calculated the mutation rate of each substitution type as the number of substitutions 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\textsuperscript{58}. Substitutions within protein-coding genes were classified as either nonsynonymous or synonymous. We counted the number of nonsynonymous (n) and synonymous (s) substitutions for each gene across all somatic mutations in the 7,042 tumor-normal pairs. Somatic mutations at the same site and with the same mutation type that occurred in different patients
were counted as different substitutions because they, unlike germline evolution, occurred independently.

We further calculated the number 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\textsuperscript{25}. 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 number of nonsynonymous (n) and synonymous (s) substitutions, as well as the number 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: 
\[ \frac{dN}{dS} = \frac{n + 0.5}{N + 0.5} \times \frac{n + 0.5}{S + 0.5}. \]

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

A $\chi^2$ test was used to compare the number of nonsynonymous and synonymous substitutions to the number 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\textsuperscript{38}. 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 statistical analysis (http://www.r-project.org/).

References

16 Davoli, T. et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and


Acknowledgements

We are grateful to Xiaopu Wang for his 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.

Authors Contributions

ZZ, ZS and XG conceived of the study. ZZ, YZ, GL, JZ and ZS performed the data analyses. SZ contributed ideas and tools for the analysis and edited the manuscript. ZZ, YZ, ZS and XG wrote the manuscript. All authors read and approved the final manuscript.
Competing financial interests

The authors declare that they have no competing financial interests.
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 whole cancer genomes, the intergenic mutation profile of 507 whole cancer 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 positively and negatively selected cancer genes with the mpNG method.

**Figure 3.** The overlap of positively selected (A) and negatively selected (B) genes based on different models.
### Table 1. The dN/dS values and number of human genes under positive or negative selection in germline and cancer based on the NG86 and mpNG methods with different mutation profiles. P-values are according to a $\chi^2$ test.

<table>
<thead>
<tr>
<th></th>
<th>dN/dS</th>
<th># Positive selection $^*$</th>
<th># Negative selection $^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germline (NG86)</td>
<td>0.624 ± 0.004</td>
<td>42</td>
<td>9093</td>
</tr>
<tr>
<td>Cancer (NG86)</td>
<td>0.990 ± 0.006</td>
<td>306</td>
<td>2330</td>
</tr>
<tr>
<td>Cancer (intergenic)</td>
<td>1.240 ± 0.008</td>
<td>697</td>
<td>722</td>
</tr>
<tr>
<td>Cancer (intrinsic)</td>
<td>1.281 ± 0.008</td>
<td>822</td>
<td>624</td>
</tr>
<tr>
<td>Germline (exonic)</td>
<td>0.903 ± 0.006</td>
<td>264</td>
<td>4357</td>
</tr>
<tr>
<td>Cancer (exonic)</td>
<td>1.367 ± 0.009</td>
<td>1230</td>
<td>326</td>
</tr>
</tbody>
</table>

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, and for germline and cancer somatic substitutions.

(A)

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

(B)

<table>
<thead>
<tr>
<th>Gene Type</th>
<th>Human-Mouse vs Germline</th>
<th>Human-Mouse vs Cancer</th>
<th>Germline vs Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r</td>
<td>P-Value</td>
<td>r</td>
</tr>
<tr>
<td>All genes</td>
<td>0.04</td>
<td>3.3x10^-7</td>
<td>-0.01</td>
</tr>
<tr>
<td>Known cancer genes</td>
<td>0.45</td>
<td>&lt;10^-16</td>
<td>-0.02</td>
</tr>
<tr>
<td>Oncogenes</td>
<td>0.43</td>
<td>&lt;10^-16</td>
<td>-0.01</td>
</tr>
<tr>
<td>Tumor suppressor genes</td>
<td>0.52</td>
<td>1.0x10^-8</td>
<td>0.04</td>
</tr>
<tr>
<td>Human essential genes</td>
<td>0.19</td>
<td>&lt;10^-16</td>
<td>-0.05</td>
</tr>
<tr>
<td>Cancer essential genes</td>
<td>0.30</td>
<td>5.7x10^-7</td>
<td>-0.07</td>
</tr>
<tr>
<td>Positively selected genes</td>
<td>0.17</td>
<td>2.4x10^-9</td>
<td>-0.02</td>
</tr>
<tr>
<td>Negatively selected genes</td>
<td>0.22</td>
<td>6.3x10^-5</td>
<td>0.10</td>
</tr>
</tbody>
</table>
Table 3. Functional enrichment of positively and negatively selected genes in cancer genomes (P<0.01, FDR<10%).

(A)

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

(B)

<table>
<thead>
<tr>
<th>Category</th>
<th>Term</th>
<th>P-Value</th>
<th>FDR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0007600~sensory perception</td>
<td>1.35×10⁻³</td>
<td>2.20</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0050890~cognition</td>
<td>3.11×10⁻³</td>
<td>5.00</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0007608~sensory perception of smell</td>
<td>4.27×10⁻³</td>
<td>6.80</td>
</tr>
<tr>
<td>GOTERM_BP_FAT</td>
<td>GO:0007606~sensory perception of chemical stimulus</td>
<td>4.67×10⁻³</td>
<td>7.41</td>
</tr>
<tr>
<td>KEGG_PATHWAY</td>
<td>hsa04740:Olfactory transduction</td>
<td>1.03×10⁻³</td>
<td>1.11</td>
</tr>
</tbody>
</table>