Mutation-Profile-Based Methods for Understanding Selection Forces in Cancer Somatic Mutations: A Comparative Analysis

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

[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, School of Life Sciences, 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 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.

42 **Keywords:** Cancer somatic mutations, Natural selection, Cancer essential gene,

43 Evolution

Introduction

Since the pioneering work of Cairns and Nowell[1,2], the evolutionary concept of cancer progression has been widely accepted[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[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[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 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[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[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[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[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[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 [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[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[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[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[25], 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[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[28]. The different mutation profiles may lead to different biological progressions in carcinogenesis, which have been depicted in several publications[17,26]. Thus, it is inappropriate to use conventional methods such as NG86[25] 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

the mpNG method

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[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 χ^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 (χ^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 (χ^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 (χ^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 (χ^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[29]. 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[14] or

198 the increased positive selection of globally expressed genes[15]. Our results show that
199 the number of genes under positive selection increased, whereas the number of genes
200 under negative selection decreased, in cancer genomes compared with germline
201 genomes. This result indicates that both the relaxation of purifying selection on
202 passenger mutations and the positive selection of driver mutations may contribute to
203 the increased dN/dS values of human genes in cancer genomes.
204
205 **Relaxation of purifying selection for human genes in cancer cells**
206 In this study, we used the mpNG method with exonic mutation profiles to
207 estimate the dN/dS values for germline substitutions and cancer somatic mutations.
208 The Cancer Gene Census[30,31] contains more than 500 cancer genes that have been
209 reported in the literature to exhibit mutations and that are causally implicated in
210 cancer development. Of those, 503 genes were included in the 18,602 genes we tested.
211 These known cancer genes had significantly lower dN/dS values for germline
212 substitutions (Wilcoxon test, $P<10^{-16}$), but slightly greater dN/dS values (Wilcoxon
213 test, P=0.01) for cancer somatic mutations than those of other genes (Table 2A). For
214 selection over longer time scales, we extracted the dN/dS values between
215 human-mouse orthologs from the Ensembl database (Release 73)[32,33]. The known
216 cancer genes had significantly lower human-mouse dN/dS values than other human
217 genes. Among the cancer genes, oncogenes (OGs) had significantly lower dN/dS
218 values than non-cancer genes (Wilcoxon test, $P<10^{-15}$), whereas the mean dN/dS
219 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.*[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 (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[35] and cancer common essential genes[21]. We extracted 2,452 human essential genes from DEG10 (the Database of Essential Genes)[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[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[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)[38]. We set the strengthened filter for detecting positively and negatively selected genes at $P<10^{-3}$ 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[30]. Fourteen of the 61 genes are human essential genes, and seven are orthologs of mouse or yeast essential genes, according to the DEG10[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[39-42]. Additionally, ACTG1 is associated with cancer cell migration[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[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[45,46]. Overexpression of BCL2L12 has been detected in several cancer types, and BCL2L12 is therefore considered a molecular prognostic biomarker in these cancers[47-50]. 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 bladder cancer grade. Additionally, silencing MAP4 can efficiently disrupt the

microtubule cytoskeleton, inhibiting the invasion and migration of bladder cancer cells[51]. 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[52]. 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[26]. 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[16-18,53], 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[54]. 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[55], however, can only be performed in cell lines *in vitro* and are limited to the analysis of genes important for proliferation and survival[21,56]. 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 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[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[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[29]. 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

396 genes have previously been reported as prognostic biomarkers for cancers. Because
397 cancer-specific negatively selected genes are more evolutionarily constrained in
398 cancer cells than in normal cells, identification of cancer-specific negatively selected
399 genes would inform the potential options for cancer therapeutic targets or diagnostic
400 biomarkers. However, cancer somatic mutations vary greatly among different cancer
401 types and even among individual cancer genomes[17,18,26,57], therefore, further studies
402 will be needed to better understand the evolution of human cancer.
403

404 **Methods**

405 **Datasets**

406 Cancer somatic mutation data from 7,042 primary cancers corresponding to 30
407 different classes were extracted from the work of Alexandrov *et al.*[26], which includes
408 4,938,362 somatic substitutions and small insertions/deletions from 507 WGSs and
409 6,535 WESs. Data on rare human protein-coding variants (minor allele frequency
410 <0.01%) from 6,500 human WESs (ESP6500) were extracted from the ANNOVAR
411 database [58] based on the NHLBI GO Exome Sequencing Project. A total of 522
412 known cancer genes were extracted from the Cancer Gene Census
413 (http://cancer.sanger.ac.uk/cancergenome/projects/census/, COSMIC v68)[30,31].

414 Human gene sequences and annotations were extracted from the Ensembl
415 database (Release 73)[32,33]. For each gene, we only chose the longest sequence to avoid
416 duplicate records of each single substitution. The HGNC (HUGO Gene Nomenclature
417 Committee) database [59] (http://www.genenames.org/) and the Genecards database [60]

418 (http://www.genecards.org) were also used to map the gene IDs from different
419 datasets. DAVID v6.7 was utilized for the functional annotation analysis[36,37].
420
421 **Calculating mutation rate profiles**
422 We calculated the mutation rate profiles using the 96 substitution
423 classifications[26,27], which not only show the base substitution but also include
424 information on the sequence context of each mutated base. We counted all somatic
425 substitutions in the protein-coding regions of the 7,042 tumor-normal paired WESs, as
426 well as all the protein-coding variants of the ESP6500 dataset. We also counted the
427 total number of each trinucleotide type for the exonic, intronic, and intergenic regions
428 in the human genome. We calculated the mutation rate of each substitution type as the
429 number of substitutions per trinucleotide type per patient. The mutation profiles were
430 depicted as the mutation rate of each mutation type according to the 96 substitution
431 classifications.
432
433 **Detection of positive and negative selections**
434 ANNOVAR was utilized to perform biological and functional annotations of the
435 cancer somatic mutations and germline substitutions [58]. Substitutions within
436 protein-coding genes were classified as either nonsynonymous or synonymous. We
437 counted the number of nonsynonymous (n) and synonymous (s) substitutions for each
438 gene across all somatic mutations in the 7,042 tumor-normal pairs. Somatic mutations
439 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[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: $\frac{dN}{dS} = \frac{n/\mathrm{N}}{(\mathrm{s}+0.5)/(S+0.5)}$.

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

A χ^2 test was used to compare the 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,

462 selection. The false discovery rate was estimated using the qvalue package from
463 Bioconductor[38]. A Wilcoxon test was performed to compare dN/dS values between
464 cancer somatic substitutions and germline substitutions, as well as between known
465 cancer genes and all other genes. The software tool R was used for statistical analysis
466 (http://www.r-project.org/).

467

468 **References**

469 1 Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255**, 197-200 (1975).
470 2 Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23-28 (1976).
471 3 Crespi, B. J. & Summers, K. Positive selection in the evolution of cancer. *Biol. Rev. Camb. Philos.*
472 *Soc.* **81**, 407-424, doi:10.1017/S1464793106007056 (2006).
473 4 Merlo, L. M., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological
474 process. *Nat. Rev. Cancer* **6**, 924-935, doi:10.1038/nrc2013 (2006).
475 5 Podlaha, O., Riester, M., De, S. & Michor, F. Evolution of the cancer genome. *Trends Genet.* **28**,
476 155-163, doi:10.1016/j.tig.2012.01.003 (2012).
477 6 Yates, L. R. & Campbell, P. J. Evolution of the cancer genome. *Nat. Rev. Genet.* **13**, 795-806,
478 doi:10.1038/nrg3317 (2012).
479 7 Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306-313,
480 doi:10.1038/nature10762 (2012).
481 8 Luo, J., Solimini, N. L. & Elledge, S. J. Principles of cancer therapy: oncogene and non-oncogene
482 addiction. *Cell* **136**, 823-837, doi:10.1016/j.cell.2009.02.024 (2009).
483 9 Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646-674
484 (2011).
485 10 Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719-724,
486 doi:10.1038/nature07943 (2009).
487 11 Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-158,
488 doi:10.1038/nature05610 (2007).
489 12 Beckman, R. A. & Loeb, L. A. Negative clonal selection in tumor evolution. *Genetics* **171**,
490 2123-2131, doi:10.1534/genetics.105.040840 (2005).
491 13 McFarland, C. D., Korolev, K. S., Kryukov, G. V., Sunyaev, S. R. & Mirny, L. A. Impact of
492 deleterious passenger mutations on cancer progression. *Proc. Natl. Acad. Sci. USA* **110**, 2910-2915,
493 doi:10.1073/pnas.1213968110 (2013).
494 14 Woo, Y. H. & Li, W. H. DNA replication timing and selection shape the landscape of nucleotide
495 variation in cancer genomes. *Nat. Commun.* **3**, 1004, doi:10.1038/ncomms1982 (2012).
496 15 Ostrow, S. L., Barshir, R., DeGregori, J., Yeger-Lotem, E. & Hershberg, R. Cancer Evolution Is
497 Associated with Pervasive Positive Selection on Globally Expressed Genes. *PLoS Genet.* **10**,
498 e1004239, doi:10.1371/journal.pgen.1004239 (2014).
499 16 Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and

shape the cancer genome. *Cell* **155**, 948-962, doi:10.1016/j.cell.2013.10.011 (2013).

17 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).

18 Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501, doi:10.1038/nature12912 (2014).

19 Chen, H., Xing, K. & He, X. The dJ/dS Ratio Test Reveals Hundreds of Novel Putative Cancer Drivers. *Mol. Biol. Evol.*, doi:10.1093/molbev/msv083 (2015).

20 Jordan, I. K., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res.* **12**, 962-968, doi:10.1101/gr.87702 (2002).

21 Luo, B. *et al.* Highly parallel identification of essential genes in cancer cells. *Proc. Natl. Acad. Sci. USA* **105**, 20380-20385, doi:10.1073/pnas.0810485105 (2008).

22 Endo, T., Ikeo, K. & Gojobori, T. Large-scale search for genes on which positive selection may operate. *Mol. Biol. Evol.* **13**, 685-690 (1996).

23 Arbiza, L., Dopazo, J. & Dopazo, H. Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Comput. Biol.* **2**, e38, doi:10.1371/journal.pcbi.0020038 (2006).

24 Ovens, K. & Naugler, C. Preliminary evidence of different selection pressures on cancer cells as compared to normal tissues. *Theor. Biol. Med. Model* **9**, 44, doi:10.1186/1742-4682-9-44 (2012).

25 Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418-426 (1986).

26 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

27 Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46**, 944-950, doi:10.1038/ng.3050 (2014).

28 Hodis, E. *et al.* A landscape of driver mutations in melanoma. *Cell* **150**, 251-263, doi:10.1016/j.cell.2012.06.024 (2012).

29 Yang, Z. & Bielawski, J. P. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15**, 496-503, doi:10.1016/S0169-5347(00)01994-7 (2000).

30 Futreal, P. A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177-183, doi:10.1038/nrc1299 (2004).

31 Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945-950, doi:10.1093/nar/gkq929 (2011).

32 Kinsella, R. J. *et al.* Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)* **2011**, bar030, doi:10.1093/database/bar030 (2011).

33 Flicek, P. *et al.* Ensembl 2014. *Nucleic Acids Res.* **42**, D749-755, doi:10.1093/nar/gkt1196 (2014).

34 Thomas, M. A. *et al.* Evolutionary dynamics of oncogenes and tumor suppressor genes: higher intensities of purifying selection than other genes. *Mol. Biol. Evol.* **20**, 964-968, doi:10.1093/molbev/msg110 (2003).

35 Luo, H., Lin, Y., Gao, F., Zhang, C. T. & 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.* **42**, D574-580, doi:10.1093/nar/gkt1131 (2014).

36 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44-57, doi:10.1038/nprot.2008.211

(2009).

37 Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1-13, doi:10.1093/nar/gkn923 (2009).

38 Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440-9445, doi:10.1073/pnas.1530509100 (2003).

39 Kamal, M. *et al.* Loss of CSMD1 expression is associated with high tumour grade and poor survival in invasive ductal breast carcinoma. *Breast Cancer Res. Treat.* **121**, 555-563, doi:10.1007/s10549-009-0500-4 (2010).

40 Zhang, R. & Song, C. Loss of CSMD1 or 2 may contribute to the poor prognosis of colorectal cancer patients. *Tumour Biol.* **35**, 4419-4423, doi:10.1007/s13277-013-1581-6 (2014).

41 Chang, G. *et al.* Hypoexpression and epigenetic regulation of candidate tumor suppressor gene CADM-2 in human prostate cancer. *Clin. Cancer Res.* **16**, 5390-5401, doi:10.1158/1078-0432.CCR-10-1461 (2010).

42 He, W. *et al.* Aberrant methylation and loss of CADM2 tumor suppressor expression is associated with human renal cell carcinoma tumor progression. *Biochem. Biophys. Res. Commun.* **435**, 526-532, doi:10.1016/j.bbrc.2013.04.074 (2013).

43 Luo, Y. *et al.* Loss of ASAP3 destabilizes cytoskeletal protein ACTG1 to suppress cancer cell migration. *Mol. Med. Rep.* **9**, 387-394, doi:10.3892/mmr.2013.1831 (2014).

44 Roy, R., Chun, J. & Powell, S. N. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat. Rev. Cancer* **12**, 68-78, doi:10.1038/nrc3181 (2012).

45 Stegh, A. H. *et al.* Bcl2L12-mediated inhibition of effector caspase-3 and caspase-7 via distinct mechanisms in glioblastoma. *Proc. Natl. Acad. Sci. USA* **105**, 10703-10708, doi:10.1073/pnas.0712034105 (2008).

46 Stegh, A. H. *et al.* Glioma oncoprotein Bcl2L12 inhibits the p53 tumor suppressor. *Genes Dev.* **24**, 2194-2204, doi:10.1101/gad.1924710 (2010).

47 Thomadaki, H. *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.* **45**, 1362-1367, doi:10.1016/j.clinbiochem.2012.06.012 (2012).

48 Karan-Djurasevic, T. *et al.* Expression of Bcl2L12 in chronic lymphocytic leukemia patients: association with clinical and molecular prognostic markers. *Med. Oncol.* **30**, 405, doi:10.1007/s12032-012-0405-7 (2013).

49 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.* **32**, 39 e29-36, doi:10.1016/j.urolonc.2013.04.005 (2014).

50 Tzovaras, A. *et al.* BCL2L12: a promising molecular prognostic biomarker in breast cancer. *Clin. Biochem.* **47**, 257-262, doi:10.1016/j.clinbiochem.2014.09.008 (2014).

51 Ou, Y. *et al.* Activation of cyclic AMP/PKA pathway inhibits bladder cancer cell invasion by targeting MAP4-dependent microtubule dynamics. *Urol. Oncol.* **32**, 47.e21-e28, doi:10.1016/j.urolonc.2013.06.017 (2014).

52 Guo, X. *et al.* Potential predictive plasma biomarkers for cervical cancer by 2D-DIGE proteomics and Ingenuity Pathway Analysis. *Tumour Biol.* **36**, 1711-1720, doi:10.1007/s13277-014-2772-5 (2015).

588 53 Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T. & Lehner, B. Synonymous Mutations Frequently
589 Act as Driver Mutations in Human Cancers. *Cell* **156**, 1324-1335, doi:10.1016/j.cell.2014.01.051
590 (2014).
591 54 Dobbelstein, M. & Moll, U. Targeting tumour-supportive cellular machineries in anticancer drug
592 development. *Nat. Rev. Drug Discov.* **13**, 179-196, doi:10.1038/nrd4201 (2014).
593 55 Root, D. E., Hacohen, N., Hahn, W. C., Lander, E. S. & Sabatini, D. M. Genome-scale
594 loss-of-function screening with a lentiviral RNAi library. *Nat. Methods* **3**, 715-719,
595 doi:10.1038/nmeth924 (2006).
596 56 Marcotte, R. *et al.* Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer*
597 *Discov.* **2**, 172-189, doi:10.1158/2159-8290.CD-11-0224 (2012).
598 57 Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**,
599 333-339, doi:10.1038/nature12634 (2013).
600 58 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from
601 high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164, doi:10.1093/nar/gkq603 (2010).
602 59 Gray, K. A. *et al.* Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res.* **41**, D545-552,
603 doi:10.1093/nar/gks1066 (2013).
604 60 Safran, M. *et al.* GeneCards Version 3: the human gene integrator. *Database (Oxford)* **2010**,
605 baq020, doi:10.1093/database/baq020 (2010).

606

607

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.

646 **Tables**

647 **Table 1.** The dN/dS values and number of human genes under positive or negative
648 selection in germline and cancer based on the NG86 and mpNG methods with
649 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

650 Note: *P<0.05

651

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

654 (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

655 (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	1.6×10^{-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

656

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