The Convergent Cancer Evolution toward a Single Cellular Destination

Han Chen¹ and Xionglei He*,¹,²
¹Key Laboratory of Gene Engineering of Ministry of Education, College of Ecology and Evolution, School of Life Sciences, Sun Yat-sen University, Guangzhou, China
²Collaborative Innovation Center of High Performance Computing, National University of Defense Technology, Changsha, China
*Corresponding author: E-mail: hexiongl@mail.sysu.edu.cn.
Associate editor: Jianzhi Zhang

Abstract

The essence of Darwin’s theory is that evolution is driven by purposeless mutations that are subsequently selected by natural environments, so there is often no predefined destination in organismal evolution. Using gene expressions of 107 cell types, we built a functional space of human cells to trace the evolutionary trajectory of 18 types of solid tumors. We detected a dominant evolving trend toward the functional status of embryonic stem cells (ESC) for approximately 3,000 tumors growing in distinct tissue environments. This pattern remained the same after excluding known cancer/ESC signature genes (~3,000 genes) or excluding all oncogenic gene sets (~12,000 genes) annotated in MSigDB, suggesting a convergent evolution of the overall functional status in cancers. In support of this, the functional distance to ESC served as a common prognostic indicator for cancers of various types, with shorter distance corresponding to poor prognosis, which was true even when randomly selected gene sets were considered. Thus, regardless of the external environments, cancer evolution is a directional process toward a defined cellular destination, a finding reconciling development and evolution, the two seemingly incompatible philosophies both adopted by the cancer research community, and also raising new questions to evolutionary biology.

Key words: cancer evolution, convergence, expression profile.

Introduction

Cancer is an evolutionary process at the cellular level (Nowell 1976; Greaves and Maley 2012). For cancers of different types, say, liver cancer and lung cancer, they start with functionally very different liver cells and lung cells, respectively, and progress on liver-specific or lung-specific habitats, so it is unlikely for the two types of cancer cells to evolve toward the same functional status (Friedl and Alexander 2011; Greaves and Maley 2012). This idea forms the theoretical basis on which human cancers of different types are treated as nearly completely different diseases in the clinic (http://www.cancer.gov). However, phenotypic commonalities among cancers are frequently observed (Hanahan and Weinberg 2011); also, we recently provided evidence for a reverse evolution from multicellularity to unicellularity during cancer (Chen et al. 2015), and suggested that rules of cancer evolution could be very different from those of typical organismal evolution (Ostrow et al. 2014; Chen et al. 2015). These considerations motivated us to reconsider the evolutionary process of cancer.

Previous studies have identified numerous gene expression signatures that relate to certain cellular processes and are common to cancers (Ramaswamy et al. 2003; Rhodes et al. 2004; Rhodes and Chinnaiyan 2005). In particular, the cancer stem cell hypothesis stimulated a surge of the interests in finding signatures shared between cancer and embryonic stem cells (ESC; Takebe and Ivy 2010; Beck and Blanpain 2013), revealing quite a few such cancer/ESC gene sets (Kim and Orkin 2011). These findings, however, are anecdotal because we are also able to identify expression signatures shared by cancer and many of other cell/tissue types such as thymus gland, tonsil, liver, breast, and so on (supplementary fig. S1, Supplementary Material online). In fact, a variety of systematic gene expression analyses showed that cancer samples are more similar to their corresponding normal tissues than to samples of a different cancer type (Ramaswamy et al. 2001; Hu et al. 2013), suggesting no overall functional convergence in cancers of various types.

Results

To better understand how cancer evolves in the functional space of human cells, we assembled genome-wide mRNA expression profiles of 107 normal human tissues (including organs and cells) (supplementary table S1, Supplementary Material online), with each expression profile representing a functional status, and together forming the functional space of human cells (fig. 1a). We also collected expression profiles of up to 3,000 solid tumor samples of 18 cancer types (supplementary table S2, Supplementary Material online). The Jensen-Shannon divergence (Endres and Schindelin 2003) was used to measure the distance (D) of two expression profiles. We defined the “home” tissue of a cancer sample as the normal tissue on which the cancer originated, and used $D_C$ and $D_H$ to stand for the distance of a cancer and its home tissue, respectively, to one of the 107 targeting normal tissues (fig. 1a). We first calculated the $107 D_C$ for each of the cancer samples and found no specific targeting...
tissues with consistently small $D_C$ for different cancer types. Also, in support of the aforementioned argument that previously identified cancer/ESC signatures are anecdotal, the $D_C$ to ESC ranks from the 1st (prostate cancer) to the 59th (colon cancer) largest, depending on the cancer types considered (fig. 1b). In addition, the $D_C$ to home tissues is generally quite small compared with that of other targeting tissues, a pattern in line with previous studies (Ramaswamy et al. 2001; Hu et al. 2013; fig. 1b).

The lack of commonality between the different cancer types could result from the heterogeneous localization of their home tissues in the functional space. To control for this, we calculated, for a given $D_C$ to a targeting tissue, say, liver, the corresponding $D_{Hi}$ to liver, and used $(D_{Hi} - D_{C})/D_{Hi}$ (or rewritten as $1 - D_C/D_{Hi}$) to measure the extent to which the cancer sample has approached liver compared with its home tissue. We first examined 322 lung cancer samples, and found that the $(D_{Hi} - D_C)/D_{Hi}$ to ESC is larger than that to any other targeting tissues in all the samples examined (fig. 2a). Separation of these samples according to their TNM stages (Sobin and Fleming 1997) revealed that later-stage tumors are generally closer to ESC than earlier-stage ones ($P < 0.0002$; fig. 2b). Because the cancer gene expressions examined here are measured mostly from whole tumors, this finding cannot be explained by the tiny fraction (often <0.1%) of cancer stem cells (Ishizawa et al. 2010). We checked using the same approach the rest 17 cancer types. Interestingly, the same pattern holds for most (15/18 = 85%) of the cancer types examined (fig. 2c). The finding is unlikely due to specific sets of cancer genes, because we observed the same dominant trend after excluding from the analyses approximately 3,000 genes of the known cancer/ESC signatures, or up to approximately 12,000 genes of the 189 oncogenic gene sets annotated in MSigDB (Subramanian et al. 2005; fig. 2d and supplementary fig. S2, Supplementary Material online), suggesting that a property representing the overall functional status was discovered. The results were essentially the same when the distance of two functional statuses was measured using 1-$R$, where $R$ is the Pearson correlation coefficient of two expression profiles (supplementary fig. S3, Supplementary Material online). We further collected two large data sets (batch II and III, supplementary table S2, Supplementary Material online) to control the potential bias caused by data combining. Batch II contained approximately 2,000 tumor samples, 17 ESC samples, and the 106 normal tissues used above. Batch III is an independent data set containing approximately 1,700 tumor samples and 37 normal tissues. The results were largely the same for all batches (supplementary figs. S4 and S5, Supplementary Material online), indicating the robustness against data usage. Notably, the association between $(D_{Hi} - D_C)/D_{Hi}$ to ESC and tumor stage/grade in figure 2b was cancer-type dependent, as we did not observe similar association for breast cancer ($R = 0.11$, $P = 0.055$, $n = 327$) and sarcoma ($R = 0.07$, $P = 0.21$, $n = 310$).

Observation of the same largest $(D_{Hi} - D_C)/D_{Hi}$ in cancers originating from different home tissues suggests a common evolutionary destination located around ESC. To gain a better understanding of the distribution of the normal and cancer samples in the multidimensional functional space, we ran principal coordinate analysis to reduce the dimensions and focused on the genes that show consistent expression changes in the tumor samples to improve the signal-to-noise ratio (supplementary table S3, Supplementary Material online). For each data set containing tumors of the same type we plotted the two largest coordinates, and observed that, for the vast majority of cancer types examined, tumor samples are distributed largely along the line connecting their home tissue with ESC.

**Fig. 1.** Cancer is generally diverged from ESC and similar to its home tissue. (a) Definition of $D_C$ and $D_{Hi}$ in the functional space of human cells. (b) The distribution of 107 $D_C$ values for each of the 18 cancer types. For tumors of a given cancer type, their median $D_C$ to one of the 107 targeting tissues was calculated; each of the resulting 107 values was then normalized by subtracting the median of the 107 values and shown as a data point. The number at the top of each column shows the ranking of the $D_C$ to ESC among the 107 $D_C$ values.
Because the home tissues of, say, glioblastoma and lung tumors, are far away from each other in the functional space (fig. 3), these data well illustrate that cancers with distinct starting positions evolve toward a common destination approximating the functional status of ESC. Notably, we were not able to observe the above pattern for myeloma and prostate tumors (supplementary fig. S6 ag–ah, Supplementary Material online), possibly due to the failure to find the correct coordinates, or the uncertainty of the home tissues for some tumors of these two types (Matsui et al. 2004; Wang et al. 2009).

The estimation of \((D_{t}-D_{c})/D_{h}\) could have been confounded by some deterministic events at tumor initiation, for instance, cell immortalization (Hanahan and Weinberg 2011). If \((D_{t}-D_{c})/D_{h}\) can really capture the properties of tumor progression that is an evolutionary process driven by stochastic factors, one should expect higher malignancy in tumors with larger \((D_{t}-D_{c})/D_{h}\) to ESC, because malignancy is acquired during tumor progression over time. In other words, observation of the seemingly convergent evolution toward ESC in cancers of various types suggested a simple but universal indicator of cancer prognosis. We collected publicly available data sets of five major cancer types containing information necessary for prognosis analysis. Indeed, the \((D_{t}-D_{c})/D_{h}\) associated with tumor stage in a cancer-type dependent manner, it is critical to test whether \((D_{t}-D_{c})/D_{h}\) can provide...
additional prognostic power. Using multiple Cox-regression model, we found that, for all of the four cancer types, \((D_H - D_C)/D_H\) to ESC significantly associated with patient outcomes when used with clinical variables (\(P\) values ranged from \(1.5 \times 10^{-5}\) to \(0.021\), supplementary table S4, Supplementary Material online). We examined for each cancer sample, its distance to the rest 106 targeting tissues, and found that none of them performs better than the distance to ESC in cancer prognosis (supplementary fig. S7, Supplementary Material online), suggesting that ESC represents the evolutionary destination of cancers better than any other tissues in the functional space.

The above result is unusual, considering the previous huge efforts in finding specific gene sets involving in certain cellular processes for cancer prognosis (van de Vijver et al. 2002; Wang et al. 2005). In fact, there has been an unresolved enigma that the prognosis genes identified by different studies have diverse functions, rarely overlap, and in combination show no better performance (Fan et al. 2006), which has motivated a pioneer speculation of an unidentified single cancer prognostic space, with different sections captured by the different sets of genes (Fan et al. 2006; Massague 2007). It seems that the functional distance of cancer to ESC defines the single prognostic space. In the attempts to determine signal-contributing genes, we found that the prognostic performance of randomly selected 500 genes (1,500 genes for breast cancer) is equivalent to that of the total approximately 20,000 human genes (fig. 4e–h), suggesting that the overall functional status, rather than expressions of some specific genes, underlies the prognosis. Notably, as far as the prognosis analysis is concerned, 1) the simple \((D_H - D_C)/D_H\) used in this study is not an optimized measure of the functional distance to ESC; 2) because \(D_H\) is the same given the cancer type, \(D_C\) is sufficient for patient classifications.

**Discussion**

The conventional strategies of searching for cancer commonalities using gene expressions are either focusing on the shared

---

**FIG. 3.** The distribution of cancer and normal samples across the two largest principal coordinates of the multidimensional functional space. For each cancer data set the corresponding two largest coordinates were defined using the principal coordinate analysis, and plotted for 80 glioblastoma (a), 56 lung tumors (b), 41 breast tumors (c), and 35 pancreas tumors (d), respectively, along with the 107 normal tissues. The green arrow connects the home tissue of the tumors with ESC.
FIG. 4. The (D_H-D_C)/D_H to ESC is a common prognostic indicator. For each indicated cancer type (a–d), the patients were equally divided into two groups according to their (D_H-D_C)/D_H to ESC, with larger (D_H-D_C)/D_H corresponding to the poor group. The P value was computed using log-rank test. The (D_H-D_C)/D_H based on a random subset of genes shows a similar prognostic power to that of all human genes, evidenced by their comparable P values (e–h). The red arrow shows the P value based on all genes, and the histogram summarizes the P values from 500 simulations each considering a random gene set with the indicated gene number shown in the parenthesis.
expression differences between cancer and its home tissue at the gene level (Rhodes et al. 2004) or measuring directly the similarity between cancer and normal tissues at the level of overall functional status (i.e., the genome-wide expression profile; Ramaswamy et al. 2001). The former is confounded by the gene-level heterogeneity in different cancers, and the latter is confounded by the heterogeneity of home tissues in different cancers. In this study, we combined the strengths of the two strategies using (D_h - D_c)/D_h, which compares between cancer and its home tissue their overall functional distance to a given target, and successfully detected a dominant convergent evolution toward ESC for a large number of tumors growing in distinct tissue environments. Notably, this work is different from previous anecdotal findings of cancer signatures that often add another layer of complexity to the already super-complex question (Weinberg 2014), because such signatures are composed of a small fraction of human genes and almost always from studies with predefined hypothesis (Kim and Orkin 2011), thus containing ascertainment bias for making a general argument; also, there are diverse cancer signatures, such as epithelial–mesenchymal transition signature (Taube et al. 2010), cell cycle signature (Malumbres and Barbacid 2009), wound-healing signature (Adler et al. 2006), hypoxia signature (Chi et al. 2006), and so on, which can generate various predictions.

The different cancers started from very different home tissues, were driven by stochastic mutations, and, because only primary tumors were considered here, progressed on very different tissue environments that presumably confer distinct selection pressures (Friedl and Alexander 2011; Valastyan and Weinberg 2011; Greaves and Maley 2012). So, observation of such an environment-independent convergent evolution in cancers of distinct types is intriguing, and reminds of a traditional view in the community that cancer is a development-like process driven by fixed cellular programs (Fearon and Vogelstein 1990). The fact of ESC as the destination may suggest a dedifferentiation process for restoring totipotency during cancer. However, although ESC is often known for its totipotency, an underappreciated characteristic of ESC is its rapid and excessive proliferation to form cell colony, a feature typical to unicellular life aiming at maximizing its growth rate (Vander Heiden et al. 2009). It would be more reasonable if the status of being unicellular, instead of being totipotent, is approached in the convergence toward ESC, given the apparent Darwinian process of cancer cells that deprive themselves of the multicellularity-associated features to seek for, instead of the public fitness of the whole organisms, their own private fitness (Nowell 1976). Therefore, this study helps reconcile development and evolution, the two distinct languages in the cancer research community, with the former spoken mostly by cell biologists who advocate fixed cellular programs as driving forces (Hanahan and Weinberg 2011), and the latter spoken by geneticists who witness the apparent stochasticity of driver gene mutations (Garraway and Lander 2013).

The environment-independent convergent evolution explains how the capacity of distant-organ metastasizing evolves in primary tumors despite the often very different environmental requirements between the primary and metastatic sites (Bernards and Weinberg 2002), thus blowing away the remaining “cloud” over the field of cancer evolution (Valastyan and Weinberg 2011), and also justifying future efforts of developing common therapies for cancers of different kinds. Interestingly, results in this study are consistent with a previously proposed model of cancer reverse evolution in which, no matter what the tissue environments are, cancer cells increase their private fitness by knocking down the internal genetic constraints required for the maintenance of multicellularity (Davies and Lineweaver 2011; Chen et al. 2015; fig. 5). This being said, we do not deny the contribution of environmental factors to cancer evolution; instead, commonalities of the different tissue environments are emphasized. Thus, regardless of the external environments, cancer evolution is primarily directional toward a defined cellular destination (Huang 2012), a finding reminiscent of the Lamarck’s theory of evolution (Mayr 1972).

Materials and Methods

Microarray-Based Gene Expression Data

The microarray data of both cancer and normal tissues used in this study were generated on the Affymetrix U133 plus 2.0 platform, which facilitates the data processing and comparison. The CEL files were loaded to the R software affy (Gautier et al. 2004) for probe calling, RMA normalization, and generation of probe-level signals that were further collapsed into gene-level signals using CDF (Sandberg and Larsson 2007). Because there are often multiple samples for each of the 107 normal cell types, we defined the expression level of a gene as the median gene-level signal of all the same-type samples considered.

The data of Human Body Index Project were downloaded from GEO (http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE7307. This data set contains 677...
samples, in which 504 ones are from healthy tissues and used in this study. The 504 samples were manually grouped into 106 tissue types according to the description in the project, with details that can be found in supplementary table S1, Supplementary Material online, which also includes the information of 10 human ESCs downloaded from the NIH stem cell database (http://stemcells.nih.gov/Pages/Default.aspx). Expression profiles of the 107 normal cell/tissue types were used to build the functional space of human cells. We also assembled expressions of 3,086 primary tumors of 18 distinct types, with 1,067 collected by a published meta-analysis of cancer transcriptome (Hu et al. 2013), and 2,019 with clinical records necessary for prognosis study (supplementary table S2, Supplementary Material online). These data were used in all the analyses unless otherwise stated.

To make sure that our result is robust against data set selection, we also examined expression data from the Expression Project of Oncology (expO; http://www.integen.org/). The expO data set contains 2,158 tumor samples. We defined the cancer type of a tumor sample according to its primary site, and excluded samples with unclear cancer type information and cancer types with less than 10 tumor samples, remaining 1,749 tumor samples of 18 cancer types (supplementary table S2, Supplementary Material online). Also, we considered a different expression data set for human ESC, which is from GEO under the accession number of GSE23402 and contains 17 ESC expression profiles (supplementary table S2, Supplementary Material online). The finding of a dominant convergent evolution toward ESC is essentially the same when the new cancer and ESC data sets were analyzed (see supplementary fig. S4, Supplementary Material online).

We further collected an independent data set generated on the Affymetrix U133A platform (supplementary table S2, Supplementary Material online), containing 1,653 tumor samples of 10 cancer types and 37 normal tissues. The CEL files were loaded to the R software affy (Gautier et al. 2004) for probe calling, RMA normalization, and generation of probe-level signals that were further collapsed into gene-level signals using the CDF (Sandberg and Larsson 2007) of Affymetrix U133A platform. Our observation of convergent evolution was supported by this independent data set (supplementary fig. S5, Supplementary Material online).

Identification of Tissue-Specifically Expressed Genes

To facilitate the comparison between tissues (see supplementary fig. S1, Supplementary Material online), we selected in this study. The 504 samples were manually grouped into 106 tissue types according to the description in the project, with details that can be found in supplementary table S1, Supplementary Material online, which also includes the information of 10 human ESCs downloaded from the NIH stem cell database (http://stemcells.nih.gov/Pages/Default.aspx). Expression profiles of the 107 normal cell/tissue types were used to build the functional space of human cells. We also assembled expressions of 3,086 primary tumors of 18 distinct types, with 1,067 collected by a published meta-analysis of cancer transcriptome (Hu et al. 2013), and 2,019 with clinical records necessary for prognosis study (supplementary table S2, Supplementary Material online). These data were used in all the analyses unless otherwise stated.

To make sure that our result is robust against data set selection, we also examined expression data from the Expression Project of Oncology (expO; http://www.integen.org/). The expO data set contains 2,158 tumor samples. We defined the cancer type of a tumor sample according to its primary site, and excluded samples with unclear cancer type information and cancer types with less than 10 tumor samples, remaining 1,749 tumor samples of 18 cancer types (supplementary table S2, Supplementary Material online). Also, we considered a different expression data set for human ESC, which is from GEO under the accession number of GSE23402 and contains 17 ESC expression profiles (supplementary table S2, Supplementary Material online). The finding of a dominant convergent evolution toward ESC is essentially the same when the new cancer and ESC data sets were analyzed (see supplementary fig. S4, Supplementary Material online).

We further collected an independent data set generated on the Affymetrix U133A platform (supplementary table S2, Supplementary Material online), containing 1,653 tumor samples of 10 cancer types and 37 normal tissues. The CEL files were loaded to the R software affy (Gautier et al. 2004) for probe calling, RMA normalization, and generation of probe-level signals that were further collapsed into gene-level signals using the CDF (Sandberg and Larsson 2007) of Affymetrix U133A platform. Our observation of convergent evolution was supported by this independent data set (supplementary fig. S5, Supplementary Material online).

Identification of the Genes with Common Expression Changes in Cancers

We considered 32 cancer data sets each containing more than five tumor/normal paired samples (supplementary table S2, Supplementary Material online). Following previous studies (Rhodes et al. 2004; Lu et al. 2007), we identified in each data set differentially expressed genes with a false discovery rate less than 0.1, using the R software samr (Tusher et al. 2001) with default settings, and defined genes with unidirectional expression changes in more than 40% of the data sets (i.e., more than 12 data sets) as those with common expression changes in cancers (supplementary table S3, Supplementary Material online), resulting in 302 genes. These genes were used for computing the Jensen–Shannon distance (JSD; Endres and Schindelin 2003) between samples, which was used in the analyses for figure 3 and supplementary figure S5, Supplementary Material online. In addition, using a less stringent criterion that requires unidirectional expression changes in more than 30% of the tumor samples examined, we defined 782 genes with common expression changes (550 upregulation and 232 downregulation) in cancers for the analysis in supplementary figure S1, Supplementary Material online.

Calculation of the JSD

Let PX be the probability distribution of the expression profile X:

\[
PX = \left\{ \sum_{i=0}^{n} EX_i, ..., \sum_{i=0}^{n} EX_n \right\}
\]

where \( EX_i \) was the expression level of the ith gene in the sample X.

The Jensen–Shannon divergence between expression profile X and Y (JSD (PX, PY)) was defined as:

\[
JSD(PX, PY) = \frac{1}{2} \text{KLD}(PX, \frac{PX + PY}{2}) + \frac{1}{2} \text{KLD}(PY, \frac{PX + PY}{2})
\]

where the Kullback–Leibler Divergence (Kullback and Leibler 1951) between expression profile X and Y (KLD (PX, PY)) was defined as:

\[
\text{KLD}(PX, PY) = \sum_{i=1}^{n} EX_i \log_2 \left( \frac{EX_i}{EY_i} \right)
\]

The \( I_{gt} \) measures the gth gene's relative expression level in the tth tissue compared with the other 106 tissues. We then computed the Shannon's entropy (Shannon 1997) only for the 1,000 genes with the smallest \( I_{gt} \) values in each of the 107 tissues, and the 200 genes with the smallest Shannon's entropies were defined as specifically highly expressed in the tissue. We defined the 200 specifically lowly expressed genes in each tissue using the same approach, except that the largest \( I_{gt} \) and Shannon's entropy were considered.
The resulting JSD (PX, PY) was then used to calculate the JSD (Endres and Schindelin 2003) between PX and PY (D (PX, PY)):

\[ D(PX, PY) = \sqrt{\text{JSD}(PX, PY)} \]

Calculation of \( D_{H} \), \( D_{C} \), and \( (D_{H}-D_{C})/D_{H} \)

The calculations of \( D_{H} \), \( D_{C} \), and \( (D_{H}-D_{C})/D_{H} \) were demonstrated using the example of a lung tumor with the targeting tissue to be liver. We calculated \( D_{H} \) as the JSD between the targeting tissue (liver) and the home tissue (normal lung in this example); \( D_{C} \) as the JSD between the lung tumor and the targeting tissue (liver). For a cohort of lung tumors, we calculated the \( (D_{H}-D_{C})/D_{H} \) to liver of each tumor and defined the median as the \( (D_{H}-D_{C})/D_{H} \) to liver for the lung cancer. Similarly, we calculated the \( (D_{H}-D_{C})/D_{H} \) of each cancer type to each of the 107 normal tissues to generate figure 2c. Notably, when the targeting tissue is exactly the home tissue \( (D_{H} = 0) \), we artificially set the \( (D_{H}-D_{C})/D_{H} \) to -1, because the cancer cells are LEAVING the targeting tissue (the home tissue), instead of APPROACHING it.

Exclusion of the Cancer/ESC Signatures or All Known Oncogenic Signatures

We collected five cancer/ESC expression signatures reported in previous studies (Ben-Porath et al. 2008; Wong et al. 2008; Kim et al. 2010; Mizuno et al. 2010; Shats et al. 2011), resulting in 2,956 nonredundant genes that are excluded in the analysis for figure 2d and supplementary figure S2a, Supplementary Material online. We used the c6 collection of oncogenic signatures of the Molecular Signature Database (MsigDB) (Subramanian et al. 2005), which comprises 189 oncogenic gene sets with 12,129 nonredundant genes that are excluded in the analysis for figure 2d and supplementary figure S2b, Supplementary Material online.

Visualization of the Convergent Evolution toward ESC by Multidimensional Scaling (or principal coordinate analysis)

We conducted the multidimensional scaling analysis by following a previously described procedure (Arumugam et al. 2011). Specifically, for each cancer data set, we constructed the JSD matrix of all samples, including the tumor samples and the 107 normal tissues. The matrix was classified by \( k \)-medoid clustering and the most optimized \( k \) value was determined by measuring the Calinski–Harabasz index (CHI). We found that for nearly all the data sets examined, CHI meets its maximum when \( k = 3 \), when each of the three clusters largely corresponds to tumor samples, normal neural tissues, and other normal tissues, respectively. We then performed multidimensional scaling analysis using the Jensen–Shannon distance matrix, and plotted all samples using the first and the second largest coordinates, with the three clusters in the plot marked using the R software ade4. There are 38 panels in figure 3 and supplementary figure S5, Supplementary Material online, corresponding to 38 cancer data sets examined, with details in supplementary table S2, Supplementary Material online.

Cancer Prognosis Analysis

We collected five data sets each containing at least 200 patients with accessible CEL files for prognosis analysis (supplementary table S2, Supplementary Material online). Patients of each of the five cancer types were equally divided into two groups according to their \( (D_{H}-D_{C})/D_{H} \) to ESC, with larger \( (D_{H}-D_{C})/D_{H} \) for the poor group. We then used the Cox-regression model to determine the difference of the disease-free survival between the poor group and good group, and computed the \( P \) values using the log-rank test. All the analyses were performed using the R software survival.

Supplementary Material

Supplementary figures S1–S7 and table S1–S4 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Dr J. Zhang and Dr C.-I. Wu for comments and critical reading of the manuscript. This work was supported by the National Basic Research Program of China (#2014CB542005) and the National Natural Science Foundation of China (#91431103 and #31225014).

References


