SCIENTIFIC REPORTS
nature research

OPEN

Oncogenes, tumor suppressor and differentiation genes represent the oldest human gene classes and evolve concurrently

A. A. Makashov[1,2,5], S. V. Malov[3,4] & A. P. Kozlov[1,2,5,6*]

Earlier we showed that human genome contains many evolutionarily young or novel genes with tumor-specific or tumor-predominant expression. We suggest calling such genes Tumor Specifically Expressed, Evolutionarily New (*TSEEN*) genes. In this paper we performed a study of the evolutionary ages of different classes of human genes, using homology searches in genomes of different taxa in human lineage. We discovered that different classes of human genes have different evolutionary ages and confirmed the existence of *TSEEN* gene classes. On the other hand, we found that oncogenes, tumor-suppressor genes and differentiation genes are among the oldest gene classes in humans and their evolution occurs concurrently. These findings confirm non-trivial predictions made by our hypothesis of the possible evolutionary role of hereditary tumors. The results may be important for better understanding of tumor biology. *TSEEN* genes may become the best tumor markers.

We are interested in the possible evolutionary role of tumors. In previous publications[1–5] we formulated the hypothesis of the possible evolutionary role of hereditary tumors, i.e. tumors that can be passed from parent to offspring. According to this hypothesis, hereditary tumors were the source of extra cell masses which could be used in the evolution of multicellular organisms for the expression of evolutionarily novel genes, for the origin of new differentiated cell types with novel functions and for building new structures which constitute evolutionary innovations and morphological novelties. Hereditary tumors could play an evolutionary role by providing conditions (space and resources) for the expression of genes newly evolving in the DNA of germ cells. As a result of expression of novel genes, tumor cells acquired new functions and differentiated in new directions, which might lead to the origin of new cell types, tissues and organs[5]. The new cell type was inherited in progeny generations due to genetic and transgenerational epigenetic mechanisms similar to those for pre-existing cell types[5–7].

Our hypothesis makes several nontrivial predictions. One of predictions is that tumors could be selected for new functional roles beneficial to the organism. This prediction was addressed in a special work[5,8], in which it was shown that the "hoods" of some varieties of gold fishes such as Lionhead, Oranda, etc. are benign tumors. These tumors have been selected by breeders for hundreds of years and eventually formed a new organ, the "hood".

The other prediction of the hypothesis is that evolutionarily young and novel genes should be specifically expressed in tumors. This prediction was verified in a number of papers from our laboratory[9–19]. We have described several evolutionarily young or novel genes with tumor-predominant or tumor-specific expression, and even the evolutionary novelty of the class of genes – cancer/testis genes – which consists of evolutionary young and novel genes expressed predominantly in tumors (reviewed in)[9]. We suggest calling such genes Tumor Specifically Expressed, Evolutionarily New (*TSEEN*) genes[5,9]. *TSEEN* genes may become the best tumor markers[9,10].

In this paper, we performed a systematic study of the evolutionary ages of different functional classes of human genes in order to verify one more nontrivial prediction of the hypothesis of the possible evolutionary

[1]Biomedical Center, Viborgskaya str. 8, Saint-Petersburg, 194044, Russia. [2]Peter the Great St. Petersburg Polytechnic University, Politekhnicheskaya ul., 29, St. Petersburg, 195251, Russia. [3]Theodosius Dobzhansky Center for Genome Bioinformatics, St.-Petersburg State University, 41A, Sredniy av., St. Petersburg, 199004, Russia. [4]Department of Algorithmic Mathematics, St.-Petersburg Electrotechnical University, 5, Prof. Popova str, St. Petersburg, 197376, Russia. [5]Research Institute of Ultra Pure Biologicals, 7 Pudozhskaya str., St. Petersburg, 197110, Russia. [6]Vavilov Institute of General Genetics, 3 Gubkina str., Moscow, 119333, Russia. *email: contact@biomed.spb.ru

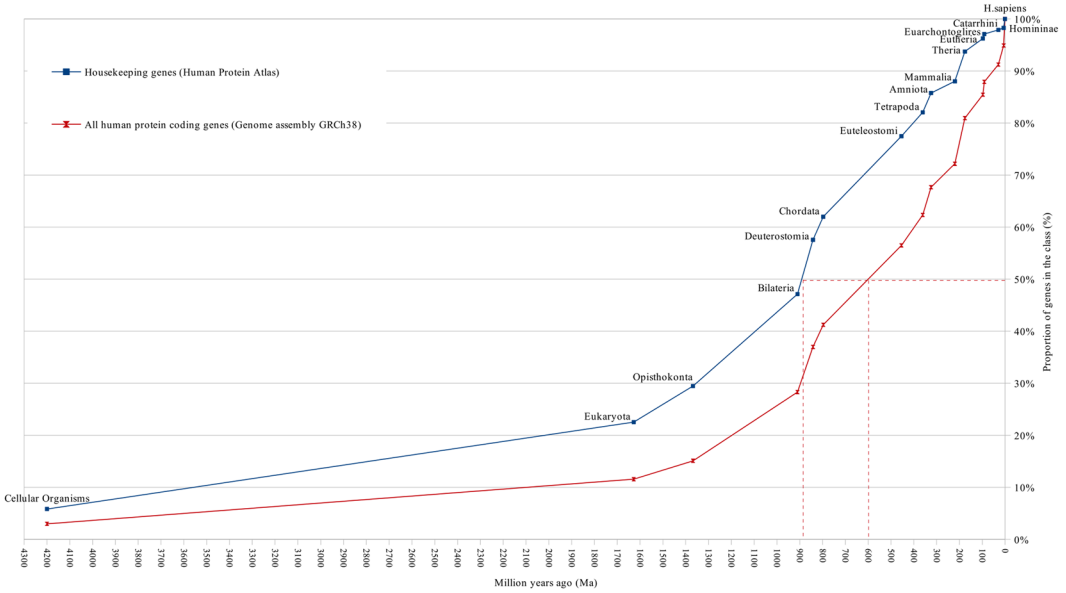

Figure 1. Distribution of human housekeeping genes genes and all protein coding genes according to their evolutionary ages. The evolutionary ages of the gene classes are measured numerically in million years at the median of distribution, i.e. at the time point on the human evolutionary timeline that corresponds to the origin of 50% of genes in this class.

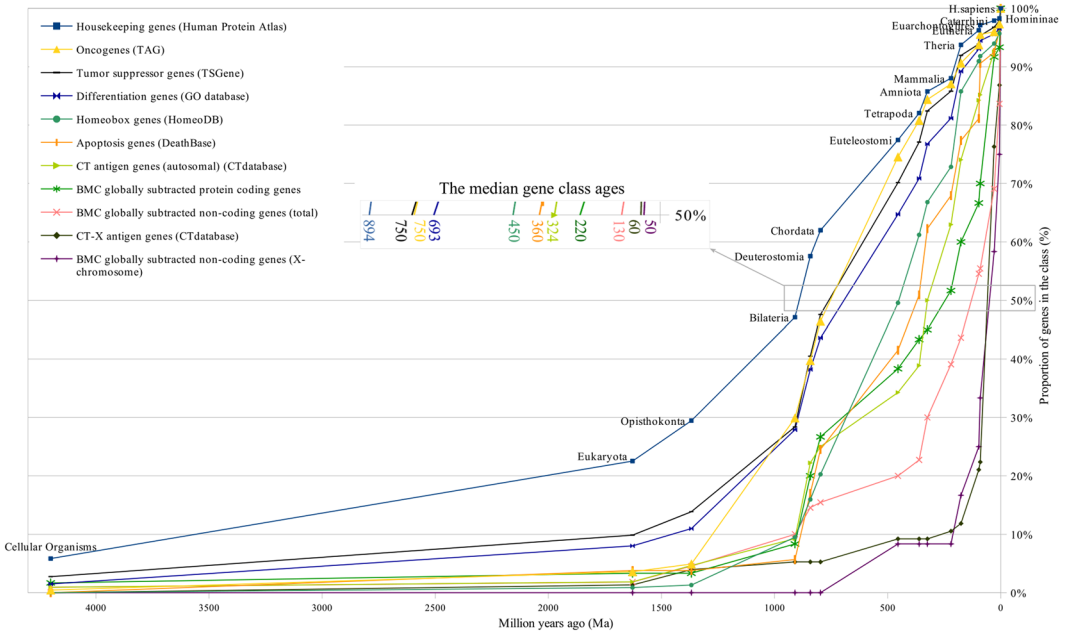

Figure 2. Gene age distributions of different classes of human genes.

role of hereditary tumors, i.e. the prediction of concurrent evolution of oncogenes, tumor suppressor genes and differentiation genes[2,3,5].

Results

The curves of gene age distribution for different classes of human genes obtained by the ProteinHistorian tool are represented in Figs 1–7.

These figures show curves sloping upward from left to right. The uppermost curve describes the gene age distribution of human housekeeping genes. The evolutionary age of this gene class, defined by the median position of the curve, is 894 million years (Ma) (Fig. 1). The curve of all human protein-coding genes has evolutionary age of 600 Ma (Figs 1 and 3–5). These curves were used as control curves in our study. Some curves are located

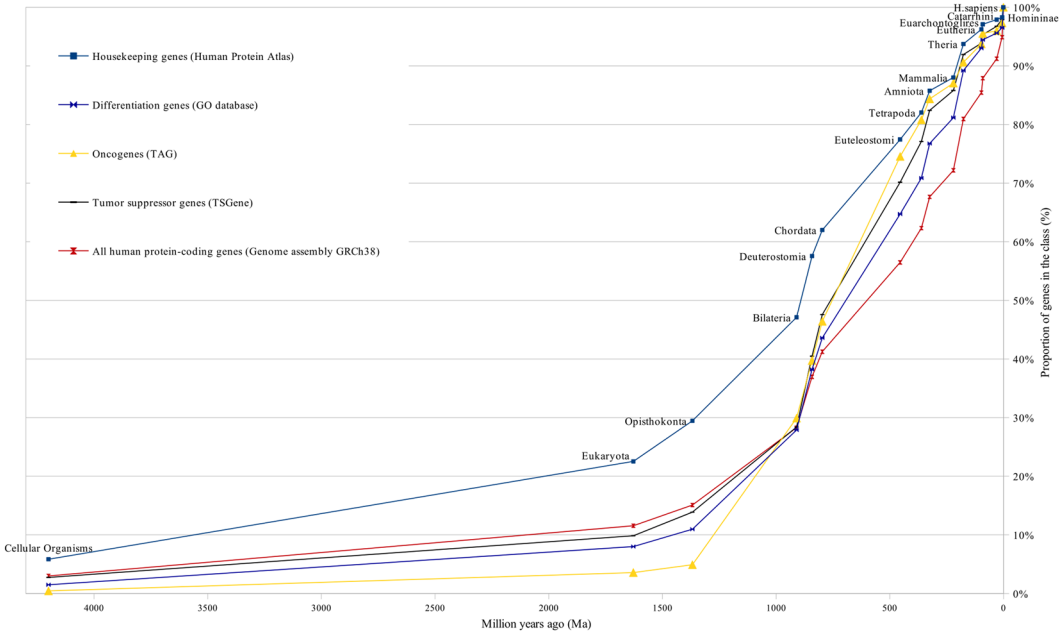

Figure 3. Cluster I of gene age distribution and control curves.

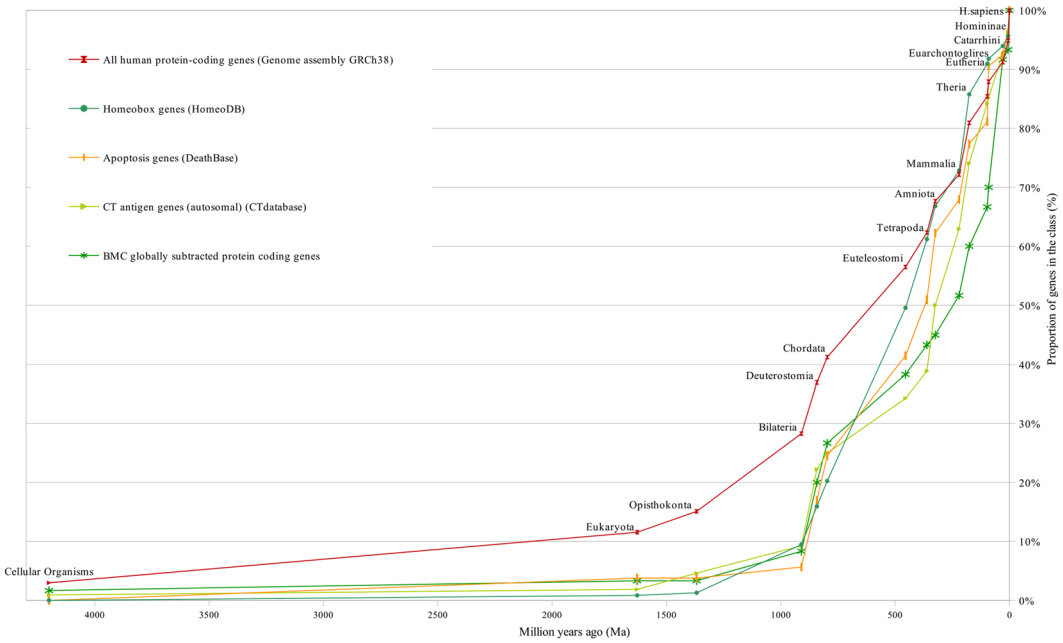

Figure 4. Cluster II of gene age distribution and control curve.

mainly between the control curves (Fig. 3), others are located below the second control curve (Figs 4 and 5). The median ages of other groups of genes are the following: oncogenes (750 Ma), tumor suppressor genes (750 Ma), differentiation genes (693 Ma), homeobox genes (450 Ma), apoptosis genes (360 Ma), canser/testis (CT) antigen genes (autosomal) (324 Ma), Biomedical Center globally subtracted, tumor-specifically expressed (BMC GSTSE) protein-coding genes (220 Ma), BMC GSTSE non-coding sequences (130 Ma), CT antigen genes located on X chromosome (CT-X) (60 Ma) and BMC GSTSE non-coding sequences located on X chromosome (BMC GSTSE-X non-coding sequences) (50 Ma) (Fig. 2). In most of the cases the pairwise differences in the age distributions of genes belonging to different classes are statistically significant (see Supplementary Dataset 1).

As follows from Figs 2–5 the curves are organized in clusters. The existence of the clusters is supported by hierarchical cluster analysis (Fig. 8). The Kolmogorov-Smirnov distance classification demonstrates moderate bootstrap reliability. If we remove from consideration housekeeping genes (control), and replace BMC GSTSE non-coding sequences with GSTSE-X non-coding sequences, the Kolmogorov-Smirnov distance classification

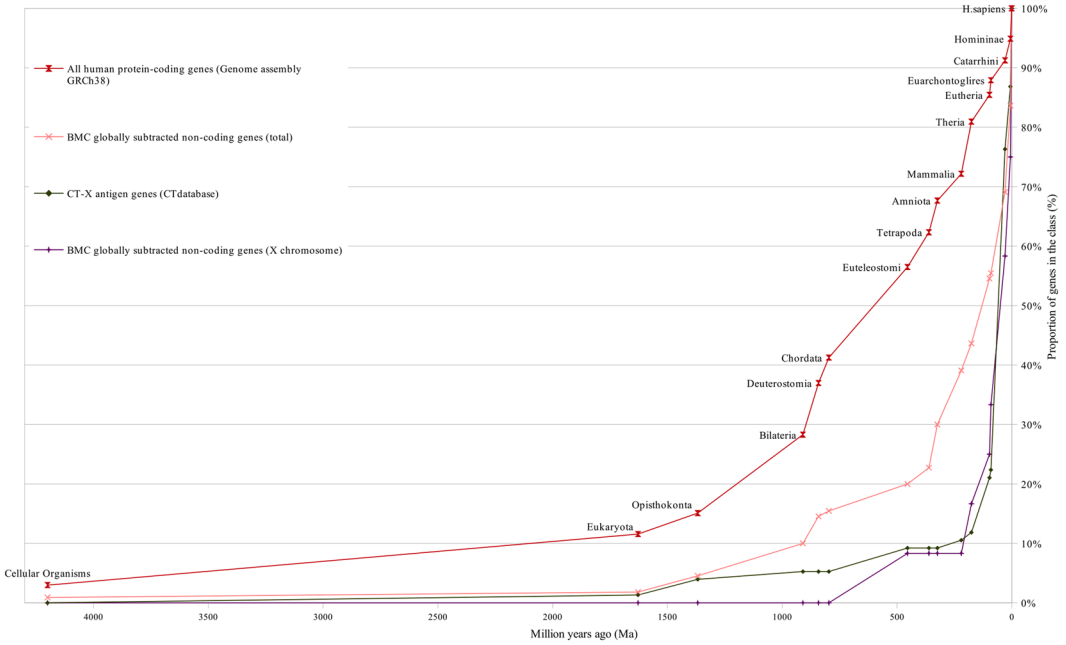

Figure 5. Cluster III of gene age distribution and control curve.

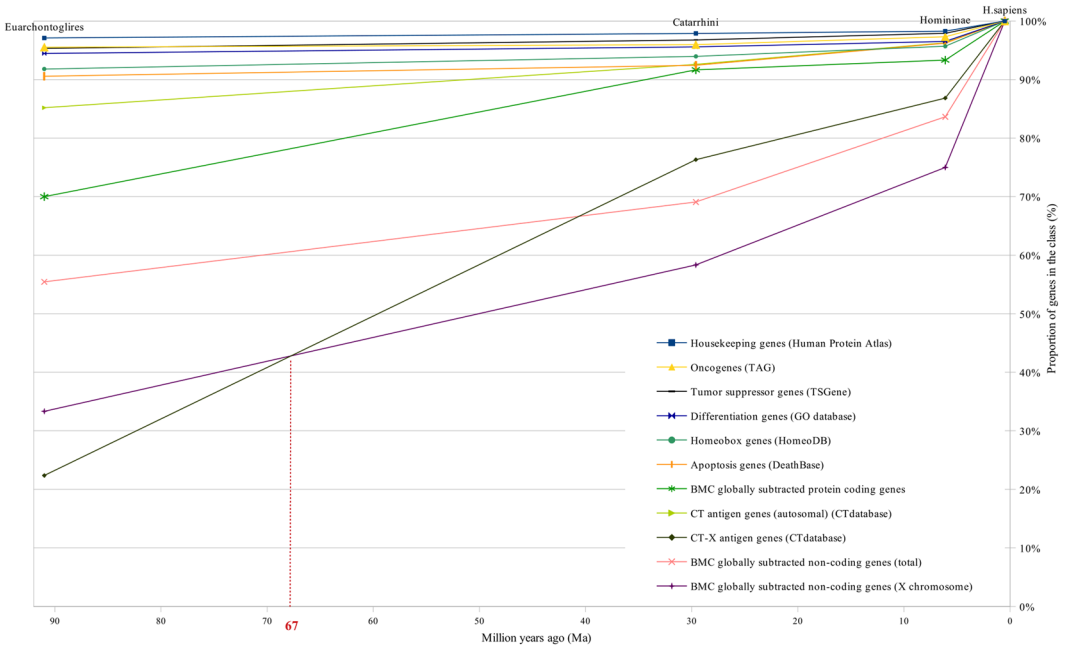

Figure 6. Gene age distribution for different classes of human genes between Euarchontoglires and H. sapiens.

demonstrates perfect bootstrap reliability (Fig. 9). The difference between the three clusters' evolutionary ages is statistically significant (chi square P-value not exceed $1 * 10^{-300}$; $X^2 = 1756$ under 30 df) as well as the pairwise difference of the ages of each pair of clusters (see Supplementary Dataset 2).

Cluster I includes the gene age distribution curves of human housekeeping genes, oncogenes, tumor suppressor genes and differentiation genes. It is located mainly between the control curves (Fig. 3). Below the all protein-coding genes curve is the larger part of cluster II (including the following: homeobox genes, apoptosis genes, autosomal CT antigen genes and BMC GSTSE protein-coding sequences, Fig. 4). The lowest position is occupied by cluster III, which includes curves of gene age distribution of the BMC GSTSE and BMC GSTSE-X non-coding sequences, and CT-X antigen genes orthologs (Fig. 5).

The curves which belong to cluster I demonstrate growth starting from > 4000 Ma. In *Bilateria* (910 Ma) they reach a proportion of 30%. The oncogene age distribution curve stays almost flat until *Opisthokonta* (1368 Ma),

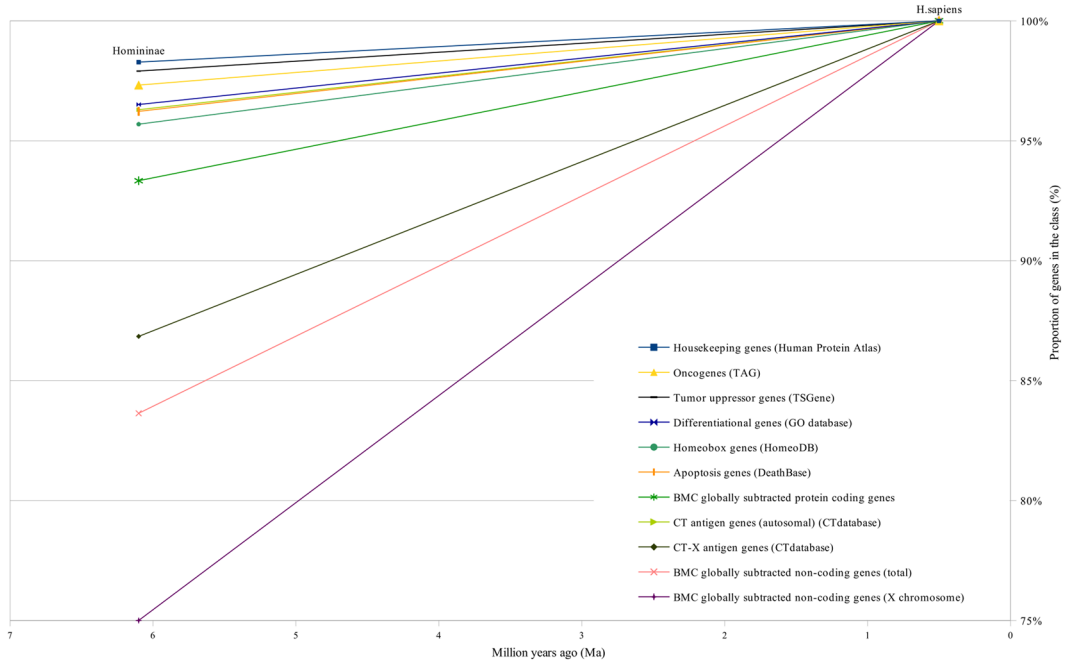

Figure 7. The proportion of different classes of human genes originated between Homininae and H. sapiens.

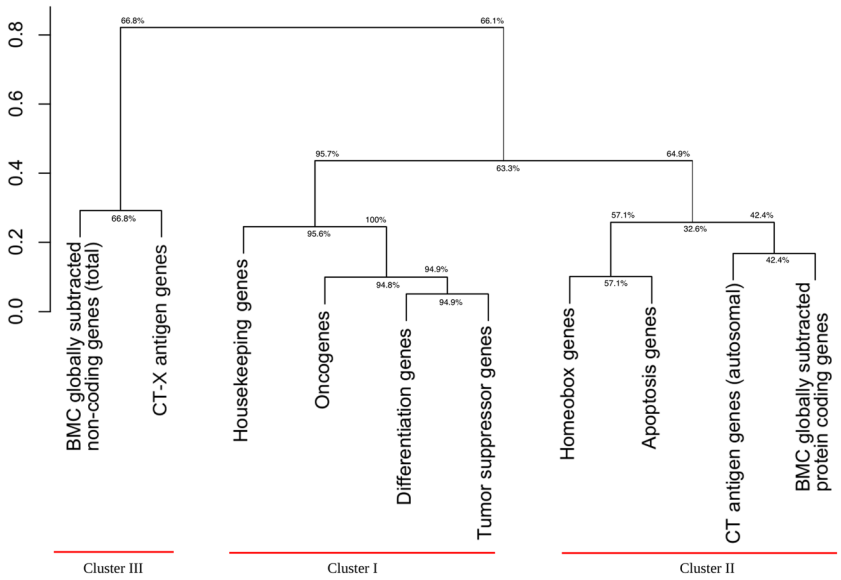

Figure 8. Hierarchical classification of 10 classes of human genes (Kolmogorov-Smirnov, complete linkage).

but after *Opisthokonta* goes upward and in *Bilateria* reaches 30% like other curves of cluster I. Between *Bilateria* and *Chordata* all curves of cluster I show a steep increase to 50%, and after *Chordata* (797 Ma) keep an almost constant slope up to 100% (Figs 2 and 3). The curve of housekeeping gene ages reaches 23% in *Eukaryota*, 29% in *Opisthokonta*, 47% in *Bilateria*, and makes a similar jump of 15% between *Bilateria* and *Chordata* (Fig. 1).

The curves of cluster II are slightly sloping until *Opisthokonta*, then slowly grow between *Opisthokonta* and *Bilateria*, and then demonstrate the 20% jump between *Bilateria* and *Chordata*, similarly to the curves of cluster I. The curve of homeobox genes ages, which belongs to cluster II, demonstrates almost constant slope between *Bilateria* and *Eutheria* (Figs 2,4).

The curves of CT-X antigen genes and BMC GSTSE-X non-coding sequences are characterized by the highest growth (as compared to other curves) of 78% and 67%, respectively, during the last 90 mln years (Figs 2 and 5–6). The gene ages curve of BMC GSTSE-X non-coding sequences occupies the lowest position during the period of the last 67 Ma (53%) and shows the maximum slope during the period of the last 6 Ma (25%), when the majority of other curves stop increasing (Figs 6 and 7).

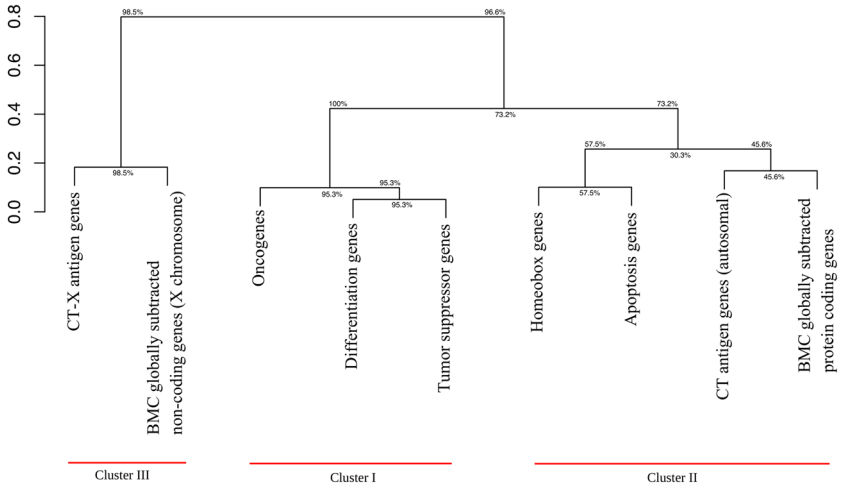

Figure 9. Hierarchical classification of 9 classes of human genes (Kolmogorov-Smirnov, complete linkage).

The CT-X antigen gene class was stochastically younger than the housekeeping gene class (two sided test P-value 0.027) and tumor suppressor gene class (two sided test P-value was 0.049), but after correction for multiple testing, simultaneously these results are not significant (see Supplementary Dataset 3 for complete pairwise relative evolutionary novelty analysis for different gene classes). Moreover, we discovered that the class of the BMC GSTSE non-coding sequences was stochastically younger than the class of housekeeping genes (two sided test P-value $2.6 * 10^{-4}$) and that the differentiation gene class was stochastically younger than the housekeeping gene class (two sided test P-value $5 * 10^{-6}$). The bootstrap rates of the stochastically younger cases agree with these hypotheses (see Supplementary Dataset 4).

We also found that cluster III was stochastically younger than cluster I (two sided test P-value is $1.7 * 10^{-5}$) and the combination of clusters I and II (two sided test P-value is $1.9 * 10^{-5}$). Moreover, cluster III was stochastically younger than all protein-coding genes (P-value 0.0015) (Supplementary Dataset 5, see also Supplementary Dataset 6 for the bootstrap agreement).

Many genes that we studied are in two or more classes (Supplementary Dataset 7). As far as we are interested in co-evolution of differentiation genes, oncogenes and tumor-suppressor genes, we examined the gene age distributions of pairwise intersections of these gene classes (Supplementary Fig. 1) and of their pairwise subtractions (Supplementary Fig. 2). We found that curves of overlapping gene subclasses (diff x onco, diff x TSG, and onco x TSG) and subtracted gene subclasses (diff-onco, diff-TSG, onco-diff, onco-TSG, TSG-diff and TSG-onco) have similar shapes, and the ages of gene subclasses are similar to the ages of original gene classes (i.e. differentiation genes, oncogenes and tumor suppressor genes) (Supplementary Figs 1–3). The curves of pairwise gene subclasses fit in the same cluster, i.e. cluster I (Supplementary Fig. 4).

Discussion

To study different functional classes of genes we used publicly available gene databases describing different gene classes – The Human Protein Atlas (housekeeping genes); Tumor-Associated Gene database (TAG database) (oncogenes); TSGene (tumor suppressor genes); CTDatabase (cancer/testis (CT) antigen genes); HomeoDB (HomeoBox genes); DeathBase (apoptosis genes); GeneOntology (differentiation genes); Biomedical Center Database (BMC GSTSE protein-coding genes and BMC GSTSE non-coding sequences). All annotated human protein coding genes (Genome assembly GRCh38) and housekeeping genes were used as controls. Although we understand the limitations of such an approach connected with differing philosophies of the authors of databases and continuing upgrading of databases, we were able to obtain meaningful results. The results were also reproducible for different versions of databases with curves corresponding to different versions almost overlapping (see Supplementary Fig. 5).

We decided to study the ages of different gene classes in order to verify the predictions which stem from the hypothesis of the possible evolutionary role of heritable tumors formulated by one of us[5]. According to this hypothesis, hereditary tumors were the source of extra cell masses, which might be used in the evolution of multicellular organisms for the expression of evolutionarily novel genes and for the origin of new differentiated cell types with novel functions.

The evolutionary role of cellular oncogenes might consist in sustaining certain level of autonomous proliferative processes in the evolving populations of organisms and in promoting the expression of evolutionarily new genes. After the origin of a new cell type, the corresponding oncogene should have turned into a cell type-specific regulator of cell division and gene expression. If true, the number of cellular oncogenes should correspond to the number of cell types in higher animals[2,3,5].

If tumors and cellular oncogenes played a role in evolution as proposed, then the evolution of oncogenes, tumor suppressor genes, differentiation genes and cell types should proceed concurrently[5].

We found that any functional gene class includes genes with different evolutionary ages. This means that genes with similar functions originated during different periods of evolution. The age of a gene was defined by the most

recent common ancestor on the human evolutionary timeline [20,21]] containing genes with similar sequences, i.e. with a significant BLAST score (or HMMER E-value).

The age of a functional gene class (or the age of the cluster) was described by distribution of ages of genes belonging to this gene class (i.e. particular gene database). For convenience, the age of the gene class can be measured numerically in million years at the median of distribution, i.e. at the time point on the human evolutionary timeline that corresponds to the origin of 50% of genes in this class. We found that different functional classes of human genes have different evolutionary ages ranging from 894 millions years for housekeeping genes to 50 million years for BMC GSTSE-X non-coding sequences. This reflects the different evolutionary history of different functional gene classes.

The curves of the older gene classes occupy the higher-left position and those of younger gene classes occupy the lower-right position on distribution curves (Figs 1–7). The slope of curves changes along the evolutionary timeline. This suggests that the rate of novel genes origin is different during different periods of evolution. Thus, the slope of all curves of clusters I and II, including the housekeeping gene ages distribution curve, increases sharply during the period between the origin of *Bilateria* and the origin of *Chordata* when many new cell types and morphological novelties originated.

About 20% of all orthologs emerge during this period. Trends of the curves during the period of the Cambrian explosion (~543–~508 Ma), when most major animal phyla appeared in the fossil record[22], suggest that this radiation was preceded and followed by the extensive origin of novel genes (Figs 2–5). We see the last considerable increase in the origin of new genes 6 Ma ago, between *Homininae* and *H. sapiens*, when 15% of CT-X antigen genes, 10% of BMC GSTSE protein-coding genes, 17% of BMC GSTSE non-coding sequences and 25% of BMC GSTSE-X non-coding sequences originated (Figs 6 and 7).

It is known that housekeeping genes represent the oldest gene class in existing cells and evolve more slowly (according to their Ka/Ks rates) than tissue-specific genes[23,24]. We found that the class of human housekeeping genes as described previously in[25] also contains evolutionarily younger genes, i.e. housekeeping genes continue to originate in the course of evolution, although at relatively slower rate than genes in other functional gene classes (see the slope of the corresponding curve). But as far as the class of housekeeping genes is large (7367 genes according to Uhlen *et al.*[25]), even in humans 117 housekeeping genes originated, according to our data.

The intensive increase in the number of oncogenes began between *Opisthokonta* and *Bilateria* (25% of oncogenes), which coincided with the origin of multicellularity. This suggests a role for oncogenes in the origin of multicellular organisms. The other important jumps in the origin of oncogenes occur between *Bilateria* and *Chordata* (26%) and between *Chordata* and *Euteleostomi* (30%), which were periods of great morphological changes. Thus 83% of oncogenes originated between *Opisthoconta* and *Mammalia*.

Our data correspond with results of phylostratigraphic tracking of cancer genes which suggest a link to the emergence of multicellularity[26]. But our data also show considerable increase in the proportion of oncogenes and tumor suppressor genes before and beyond the emergence of vertebrates (Figs 2 and 3), while Domaset-Loso and Tautz described significantly lower origination of founder genes related to cancer beyond the emergence of vertebrates. This difference may be due to difference in methodology: Domaset-Loso and Tautz studied the emergence of cancer related domains while ProteinHistorian tool, which we used, studies the origin of the full-size proteins, in our case oncoproteins and tumor suppressor proteins.

While the origin of oncogene class, according to our data, is related to the origin of multicellularity, many differentiation genes were co-opted from unicellular ancestors (Fig. 3). Today, genes that control metazoan development and differentiation are found in *Opisthokonta* suggesting that multicellularity evolved from unicellular opisthokont ancestors[27–29]. The slope of the differentiation gene ages distribution curve supports this notion. According to our data, 11% of human differentiation genes are conserved in *Opisthokonta* (Fig. 3).

The gene classes studied in this paper form three clusters visually and based on hierarchical cluster analysis. Each cluster contains curves with the least difference in gene age distributions.

The first cluster includes gene age distribution curves of housekeeping genes, oncogenes, tumor suppressor genes, and differentiation genes. This cluster is the oldest with evolutionary ages of gene classes from 894 Ma (housekeeping genes) to 693 Ma (differentiation genes). It is not homogeneous because the curve of housekeeping gene ages is separate from the other curves of the cluster, and differentiation gene class is stochastically younger than housekeeping gene class. On the other hand, gene age distribution curves of oncogenes, tumor suppressor genes and differentiation genes almost overlap. The removal of housekeeping gene class from bootstrap analysis does not destroy cluster I, but even increases its bootstrap reliability (Figs 8 and 9).

It was known for a long time that there are oncogenes, which are very ancient[30–34]. But to our knowledge this paper is the first indication in the literature that oncogenes represent the most ancient class of genes in human genome with the exception of housekeeping genes. The other interesting piece of data is that tumor suppressor genes and differentiation genes coevolve with oncogenes. The fact that orthologs of oncogenes, tumor suppressor genes and differentiation genes belong to the same cluster and their distribution curves almost overlap means that they evolve concurrently, as predicted earlier[2–5].

Moreover, we found that differentiation, onco-, and tumor suppressor gene classes partially overlap (Supplementary Dataset 7), and pairwise intersection and subtraction gene subclasses co-evolve with the main gene classes (Supplementary Figs 1–3). Overlapping of gene classes means that some genes have two (or more) functions, and may belong to two (or more) functional gene classes. It is known that a gene may function in several processes and contain exons that determine diverse molecular functions and biological processes[35]. The existence of diff x onco and diff x TSG subclasses confirmes our prediction on co-evolution of differentiation, onco-, and tumor suppressor functions even on a single gene level.

As example of gene with dual function could be TGFb. It is known that TGFb may function as tumor promoter or tumor suppressor. This phenomenon is known as "TGFb paradox"[36]. In gene classes studied in this paper, TGFb is found in oncogene, differentiation and tumor suppressor gene classes. Actually, it is a triple function.

The other example is *Wnt* gene. It was discovered as proto-oncogene[37]. On the other hand, the *Wnt* gene family encodes a group of cell-signaling molecules that participate in vertebrate and invertebrate development. Wnt protein sequence have been conserved during a billion years of evolution[38]. *Wnt* gene is found in differentiation gene and oncogene classes that we studied in this paper.

The existence of such dual-function genes and other data support our hypothesis that hereditary tumors at early or intermediate stages of progression might participate in the evolutionary origin of new differentiated cell types[4,5]. Our prediction that there should be a general correspondence between the number of oncogenes and the number of cell types is also supported by the other existing data. Thus, the TAG database, which we used in this study, currently contains 245 human oncogenes, of which 224 are found by ProteinHistorian. Domaset-Loso and Tautz used other data sets (Sanger Cosmic, NCBI Entrez section in CancerGenes, the CancerGenes and the Network of Cancer Genes (NCG)). They found 380 oncogenes in these databases[26]. On the other hand, the current estimate of the number of the cell types in humans produced the numbers from 240[39,40] to 411 cell types[41]. Supplementary Dataset 8 contains a table of correspondence of the number of oncogenes and cell types in different multicellular organisms (Supplementary Dataset 8). That is, the general correspondence between the number of cell types and the number of oncogenes does exist, as was predicted in[2,3.] It is noteworthy that when such correspondence was first predicted in 1987, only 20 oncogenes have been described[42], and by 1996 – only 70 oncogenes[43].

We further hypothesized that at least three different classes of genes are necessary for the origin of a new cell type in evolution: oncogenes, tumor suppressor genes, and evolutionarily novel genes, which determine a new function[5]. The existence of cluster I supports our hypothesis of co-evolution of differentiation, onco-, and tumor suppressor genes[5]. The bootstrap values are always the highest for differentiation, onco-, and tumor suppressor genes. This strongly supports the existence of cluster I and co-evolution of differentiation, onco-, and tumor suppressor gene classes, although the number of protein coding tumor suppressor genes (TSGene database, 1018 genes) and differentiation genes (Gene Ontology, 3697 genes) is higher than the number of oncogenes (TAG database, 245 genes). The existence of cluster I and particulary clasterization of differentiation genes and tumor suppressor genes also supports the differentiation theory of cancer[44]. According to this theory, cancer is abnormal programming of gene function during cell differentiation. The loss of tissue-specific functions (e.g. due to mutations of corresponding genes) is connected with tumors. Terminal differentiation is incompatible with tumors, i.e. has a tumor suppressor function.

The second cluster occupies the intermediate position between cluster I and custer III with evolutionary ages of gene classes between 450 Ma (homeobox genes) and 220 Ma (BMC GSTSE protein-coding genes). Cluster II locates mainly below the second control curve, i.e. the curve of all protein coding genes.

It is extremely interesting that in the evolutionary timeline the distribution curves of gene ages of homeobox and apoptosis genes are separated from those of differentiation genes by the period of several hundred millions years, i.e. evolutionarily the origin of genes responsible for differentiation and organogenesis are widely separated. Thus, before *Bilateria*, almost 30% of differentiation genes originated, and only 10% of homeobox genes. Half of differentiation genes originated at 643 million years, and half of homeobox genes – at 450 million years. In M*ammalia* 87% of differentiation genes and 73% of homeobox genes are represented. Indeed, the processes of differentiation and organogenesis are separated in evolution. For example, the thyriod gland was diffuse in the common ancestor of vertebrates and still has a diffuse nature and lacks the capsule in cyclostomes and in teleostean fishes[45–48]. In mammalians and humans diffuse endocrine system and diffuse, unencapsulated bundles of lymphatic cells still exist. Nevertheless, during certain periods of the evolutionary timeline the curves of cluster I and cluster II behave in a similar manner. E.g. between *Bilateria* and *Chordata* the curves of cluster I and cluster II demonstrate similar jump of about 20%, although in cluster II this jump starts from much lower level.

Finally, the third cluster is the youngest with evolutionary ages between 130 Ma and 50 Ma. This cluster includes gene classes expressed predominantly in tumors – CT-X genes, BMC GSTSE and BMC GSTSE-X non-coding sequences. Genes belonging to this cluster continue to originate during last 90 Ma, and even during the last 6 Ma, as shown in Figs 6 and 7. They also evolve more rapidly than other gene classes (reviewed in[5]). The youngest during the last 6 Ma period are tumor-specifically expressed non-coding sequences located on X chromosome, discovered at the Biomedical Center by global subtraction of cDNAs of all known normal libraries from cDNAs of all known tumor libraries[10,12].

We already described the evolutionary novelty of CT-X antigen gene class earlier[19]. Later other authors reproduced our results with appropriate reference to our original paper[49]. Here we confirmed the evolutionary novelty of CT-X gene class using the current upgraded database of CT genes – CTDatabase, and with another method – ProteinHistorian. In this paper, we also described the new class of *TSEEN* genes – BMC GSTSE ncRNA genes. In our other work we discovered a new long non-coding RNA (lncRNA) – *OTP-AS1* (*OTP*- antisense RNA 1)[50], which belongs to cancer/testis sequences.

Statistical analysis supported the existence of two classes of *TSEEN* genes – CT-X gene class and BMC GSTSE ncRNA gene class (Supplementary Dataset 1), which constitute cluster III. Cluster III was stochastically younger than the combination of two clusters I and II (Supplementary Datasets 5 and 6). Reduced cluster III composed of BMC GSTSE-X ncRNA and CT-X genes demonstrates perfect bootstrap reliability (Fig. 9).

Thus at least three evolutionary categories of gene classes are expressed in human tumor cells: evolutionarily old (e.g. oncogenes), evolutionarily young or novel (e.g. CT-X genes and BMC GSTSE non-coding sequences) and intermediate age gene classes (e.g. BMC GSTSE protein-coding genes). But even evolutionarily older gene classes contain evolutionarily novel genes, for example, oncogenes *CT45A1* and *TBC1D3*[51–53] (see also discussion of evolutionarily novel housekeeping genes above). On the contrary, even evolutionarily younger gene classes contain evolutionarily older genes (10% of all genes in CT-X and BMC GSTSE-X ncRNA gene classes).

The data presented in this paper support and extend the concept of tumor-specifically expressed, evolutionarily novel (*TSEEN*) genes, formulated in[3–5], and confirmed in[9–19]. From the data presented in this paper we can

see that even different classes of genes (e.g. CT-X antigen genes and BMC GSTSE non-coding sequences) could be tumor-predominantly expressed and evolutionarily young or novel.

Thus the data presented in this paper confirm two predictions of our hypothesis of the possible evolutionary role of tumors, i.e. concurrent evolution of oncogenes, tumor suppressor genes and differentiation genes, and the existence of tumor specifically expressed, evolutionarily novel (*TSEEN*) gene classes. This may be important for better understanding of tumor biology, in particular of the possible evolutionary role of tumors as described in[5].

Methods

The following public databases were used as a source of human gene classes in this study: housekeeping genes – The Human Protein Atlas; oncogenes – TAG database; tumor suppressor genes – TSGene; differentiation genes – GeneOntology; HomeoBox genes – HomeoDB; apoptosis genes – DeathBase; cancer-testis (CT) antigen genes – CTDatabase; BMC GSTSE protein-coding genes and non-coding sequences – Biomedical Center Database; and all annotated human protein coding genes – Genome assembly GRCh38 (21694 genes). CT antigen genes were divided into two groups: autosomal genes and genes located on X chromosome. BMC GSTSE non-coding sequences located on X chromosome were also separately studied. This was done because X chromosome contains relatively more evolutionarily novel genes than autosomes[5].

Housekeeping genes are 7367 genes expressed in all analyzed tissues in the Human Protein Atlas[25]. This database contains information for a large majority of all human protein-coding genes regarding the expression and localization of the corresponding proteins based on both RNA and protein data. The Atlas contains information about 44 different human tissues and organs[25].

The TAG database (Tumor Associated Genes Database) (245 oncogenes) was designed to utilize information from well-characterized oncogenes and tumor suppressor genes to facilitate cancer research. All target genes were identified through text-mining approach from the PubMed database. A semi-automatic information retrieving engine collects specific information of these target genes from various resources and store in the TAG database. At the current stage, TAG database includes 245 oncogenes[54], which were used in ProteinHistorian analysis (see below). The database we used was modified for the last time on 2014.10.03.

TSGene 2.0 database contains 1217 human tumor suppressor genes (1018 coding and 199 non-coding genes) curated from a total of over 5700 PubMed abstracts[55]. In ProteinHistorian analysis we used only 1018 protein-coding tumor suppressor genes.

Differentiation genes (3697 genes) were obtained by manual search for "differentiation" in the Gene Ontology database[35].

Homeobox gene database (HomeoDB2) (333 genes) is a manually curated database of homeobox genes and their classification. HomeoDB2 includes all homeobox loci from 10 animal genomes (human, mouse, chicken, frog, zebrafish, amphioxus, nematode, fruitfly, beetle and honeybee) plus tools for downloading sequences, comparison between different species and BLAST search[56,57]. We used the database, which was updated for the last time on 2011.08.08.

Deathbase (53 genes) is a database of proteins involved in different cell death processes. It is aimed to compile relevant data on the function, structure and evolution of this important cellular proccess in several organisms (human, mouse, zebrafish, fruitfly and worm). Information contained in the database is subject to manual curation[58]. The database was updated for the last time in 2011.

CTdatabase (286 genes) provides basic information including gene names and aliases, RefSeq accession numbers, genomic location, known splicing variants, gene duplications and additional family members. Gene expression at the mRNA level in normal and tumor tissues has been collated from publicly available data obtained by several different technologies. Manually curated data related to mRNA and protein expression, and antigen-specific immune responses in cancer patients are also available, together with links to PubMed for relevant CT antigen articles[59]. We used the update of 2017.

To construct the BMC database of sequences that are expressed in tumors but not in normal tissues, the normal EST set was subtracted *in silico* from the tumorous EST set. This approach is known as computer-assisted differential display (CDD). In total, 4564 cDNA libraries categorized as "tumorous" and 2304 "normal" libraries were used in CDD experiments. 251 EST clusters with tumor predominant expression were described in[10], and 196 clusters – in[12]. From these clusters 60 protein-coding genes and 121 non-coding sequences were selected for analysis.

All annotated human protein coding genes (21694 genes) were obtained from Genome assembly GRCh38[60] with Ensembl tool[61]. The genome assembly was submitted on 2013.12.17.

The ProteinHistorian tool was used to perform homology search in genomes of different taxa.

The ProteinHistorian tool is an integrated web server, database and a set of command line tools which estimates the phylogenetic age of proteins based on a species tree, several external datasets of protein family predictions from the Princeton Protein Orthology Database (PPOD)[62] and two algorithms for ancestral family reconstruction (Dollo and Wagner parsimony)[63]. The ProteinHistorian tool searches the orthologs in 34 completely sequenced eukaryotic and prokariotic genomes from 16 taxa in the human lineage (Cellular Organisms, Eukaryota, Opisthokonta, Bilateria, Deuterostomia, Chordata, Euteleostomi, Tetrapoda, Amniota, Mammalia, Theria, Eutheria, Euarchontoglires, Catarrhini, Homininae, and *H. sapiens*).

The species tree used in analysis is presented in Supplementary Fig. 6. Divergence time is estimated in millions of years ago (Ma) for each internal node in the species tree. It is important to note that a protein could have appeared at any time along the branch to which it is assigned, so the divergence time estimate reported is a lower bound. The ages are taken from the TimeTree database[20]. Time tree database collects estimation of time of divergence among species data from publications in molecular evolution and phylogenetics. These included phylogenetic trees scaled to time (timetrees) and occasionally tables of time estimates and regular text. The data was collected from more than 2300 studies that have been published since 1987[20].

The ProteinHistorian tool detected the following gene numbers in databases mentioned above: The Human Protein Atlas (housekeeping genes) – 6789 genes; The TAG database (oncogenes) – 224 genes; TSGene (tumor suppressor genes) – 984 genes; GeneOntology (differentiation genes) – 3697 genes; HomeoDB (homeobox genes) – 231 genes; DeathBase (apoptosis genes) – 53 genes; CTDatabase (CT-antigen genes) – 187 genes, including 109 autosomal and 78 X-chromosome located genes; Biomedical Center Database – 60 protein-coding genes; Genome assembly GRCh38 (all protein-coding genes) – 19911 genes.

The nucleotide BLAST algorithm, HMMER tool and the original Python script were used to analyze the ages of non-coding sequences. The orthologs were searched in 25 completely sequenced eukaryotic and prokaryotic genomes (Supplementary list 1).

The processing of datasets obtained with ProteinHistorian tool was carried out with Python script and Grep tool.

The age of the gene is defined by the most recent common ancestor on human evolutionary timeline containing genes with similar sequences, i.e. with a significant BLAST score (or HMMER E-value)[21].

The age of the functional gene class (or cluster) is described by distribution of ages of genes belonging to this gene class. For convenience, the age of the gene class can be measured numerically in million years at the median of distribution, i.e. at the time point on the human evolutionary timeline which corresponds to the origin of 50% of orthologs of the functional gene class (Fig. 1).

A probability distribution is stochastically smaller then another one if its cumulative distribution function is larger than the cumulative distribution function of the another one for each value of the argument. We say that a class of genes is stochastically younger than another one, if the age of this class is stochastically smaller than the age of the another class. Thus, we associate stochastically younger property of the gene class with its relative evolutionary novelty.

Before statistically analyze the relative evolutionarily novelty of gene classes we first evaluated stochastic difference in the age of gene classes using Kolmogorov-Smirnov distance to specify clusters based on the complete linkage, and performed pairwise comparative statistical analysis by using the Kolmogorov-Smirnov and Chi-square tests to discover statistically significant differences between the evolutionary ages of gene classes.

We used appropriate contrasts and Sheffe S-method of multiple comparison to verify stochastic order in the evolutionarily ages of different genes classes observed in all the time points (taxons) from cellular organisms to humans. Thus, we apply covariance-adjusted method to create efficient joint confidence intervals for differences of the empirical distribution functions in all the time points available with the covariance obtained from the weak convergence of centered difference of the empirical distribution functions to the Brownian bridge process. The distribution of maximum modulus of correlated normal distributions required for the covariance-adjusted joint confidence interval was obtained by using Monte Carlo method with 10^6 (before clustering) and 10^7 (after clustering) replications.

In order to check bootstrap reliability of the obtained results we bootstrapped independently from the original classes the same size classes 10000 times and performed exploratory analysis for each of the genes age curves, including the bootstrapped mean value, mean square error, median and quartiles for all the taxon break points (Supplementary Dataset 9). At each of the replications we obtain the hierarchical classification based on the Kolmogorov-Smirnov distance, and report the bootstrapped rates for all classes and for all nodes of the initial trees. Moreover, at each of the replications for each pair of the bootstrapped genes age curves we checked for intersections, and finally report the bootstrapped rates of stochastically larger and stochastically smaller cases.

Some genes are included in several databases. This was taken into account in statistical analysis.

To investigate intersections of gene classes, for each pair of gene classes we report the observed number of genes belonging to both classes and the corresponding expected counts, which were calculated under assumption of independent attendance of genes to classes (Supplementary Dataset 7.1). More precisely, with each of gene classes we associate a binary variable taking value 1 if the corresponding gene belongs to the gene class and 0 otherwise. The independent attendance of genes to a pair of gene classes means that the corresponding variables are independent. Moreover, for each pair of gene classes we create the 2x2 contingency table and report P-values of Chi-square and Fisher's exact tests (Supplementary Dataset 7.2). Several genes belong to three or even four gene classes. For triple and quadruple intersections of gene classes we report their counts and share of the intersection in each of the classes (Supplementary Dataset 7.3).

In order to check reliability of the classification by age of gene classes with respect to the dual functionality we use additional six subclasses to classification: the subclass of genes belonging to both differentiation and tumor suppressor gene classes (diff x TSG); the subclass of genes belonging to both differentiation and oncogene classes (diff x onco); the subclass of genes belonging to differentiation but not to tumor suppressor gene classes (diff-TSG); the subclass of genes belonging to differentiation but not to oncogene classes (diff-onco); the subclass of genes belonging to oncogenes but not to differentiation gene classes (onco-diff); the subclass of genes belonging to tumor suppressor gene but not to differentiation gene classes (TSG-diff).

Received: 11 December 2018; Accepted: 24 October 2019;
Published online: 11 November 2019

References

1. Kozlov, A. P. Evolution of living organisms as a multilevel process. *J. Theor. Biol.* **81**, 1–17 (1979).
2. Kozlov, A. P. Gene competition and the possible evolutionary role of tumours and cellular oncogenes. In: Presnov, E. V., Maresin, V. M. & Zotin, A. I. eds. *Theoretical and Mathematical Aspects of Morphogenesis. Moscow: "Nauka", 1987: 136–140 (in Russian)* (1987).
3. Kozlov, A. P. Gene competition and the possible evolutionary role of tumors. *Med. Hypotheses* **46**, 81–84 (1996).
4. Kozlov, A. P. The possible evolutionary role of tumors in the origin of new cell types. *Med. Hypotheses* **74**, 177–185 (2010).
5. Kozlov, A. P. Evolution by Tumor Neofunctionalization: The role of Tumors in the Origin of New Cell Types, Tissues and Organs. (Academic Press/Elsevier, 2014).

6. Hanson, M. A. & Skinner, M. K. Developmental origins of epigenetic transgenerational inheritance. *Environ. Epigenet.* **2**(1), dvw002 (2016).
7. Bohace, K. J., Engman, O., Germain, P.-L., Schelbert, S. & Mansuy, I. M. Transgenerational epigenetic inheritance: from biology to society—Summary Latsis Symposium Aug 28–30, 2017, Zürich, Switzerland. *Environ Epigenet.* **4**(2), dvy012, https://doi.org/10.1093/eep/dvy012 (2018).
8. Kozlov, A. P. *et al*. Hyperplastic skin growth on the head of goldfish comparative oncology aspects. *Probl. Oncol. (Voprosi Oncologii)* **58**, 387–393 (2012).
9. Kozlov, A. P. Expression of evolutionarily novel genes in tumors. *Infect. Agents Cancer* **11**, 34 (2016).
10. Baranova, A. V. *et al*. *In silico* screening for tumor-specific expressed sequences in human genome. *FEBS Lett.* **508**, 143–148 (2001).
11. Kozlov, A. P. *et al*. Evolutionarily new sequences expressed in tumors. *Infect. Agent. Cancer* **1**, 8 (2006).
12. Galachyants, Y. & Kozlov, A. P. CDD as a tool for discovery of specifically-expressed transcripts. Russ. J. AIDS, Cancer. *Public Health* **13**(2), 60–61 (2009).
13. Krukovskaya, L. L., Baranova, A., Tyezelova, T., Polev, D. & Kozlov, A. P. Experimental study of human expressed sequences newly identified in silico as tumor specific. *Tumor Biol.* **26**, 17–24 (2005).
14. Samusik, N. A., Galachyants, Y. P. & Kozlov, A. P. Analysis of evolutionary novelty of tumor-specifically expressed sequences. *Russ. J. Genet. Applied Res.* **1**, 138–148 (2011).
15. Samusik, N., Krukovskaya, L., Meln, I., Shilov, E. & Kozlov, A. P. PBOV1 is a human de novo gene with tumor-specific expression that is associated with a positive clinical outcome of cancer. *Plos One* **8**, e56162 (2013).
16. Polev, D. E., Nosova, J. K., Krukovskaya, L. L. & Kozlov, A. P. Expression of transcripts corresponding to cluster Hs.633957 in human healthy and tumor tissues. *Mol. Biol. (Mosk)* **43**, 88–92 (2009).
17. Polev, D. E., Krukovskaya, L. L. & Kozlov, A. P. Expression of the locus Hs.633957 in human digestive system and tumors. *Probl. Oncol. (Voprosy Onkologii)* **57**, 48–49 (2011).
18. Polev, D. E., Karnaukhova, J. K., Krukovskaya, L. L. & Kozlov, A. P. ELFN1-AS1 A a novel primate gene with possible microRNA function expressed predominantly in tumors. *BioMed Res.Internatl.* **2014**, e398097 (2014).
19. Dobrynin, P., Matyunina, E., Malov, S. V. & Kozlov, A. P. The novelty of human cancer/testis antigen encoding genes in evolution. *Int. J. Genomics* **2013**, e105108 (2013).
20. Hedges, S. B., Marin, J., Suleski, M., Paymer, M. & Kumar, S. Fast Track Tree of Life Reveals Clock-Like Speciation and Diversification. *Mol. Biol. Evol.* **32**(4), 835–845 (2015).
21. Koonin, E. V. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet.* **39**, 309–38 (2005).
22. Zhuravlev, A. Y. & Wood, R. A. The two phases of the Cambrian Explosion. *Sci Rep.* **8**, 16656 (*2018*).
23. Jacob, F. Evolution and tinkering. *Science* **196**(4295), 1161–1166 (1977).
24. Zhou, Q. *et al*. On the origin of new genes in Drosophila. *Genome Res.* **18**, 1446–1455 (2008).
25. Uhlen, M. *et al*. Tissue-based map of the human proteome. *Science* **347**(6220), 1260419 (2015).
26. Domaset-Loso, T. & Tautz, D. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biology* **8**, 66 (*2010*).
27. King N. The unicellular ancestry of animal development. *Dev. Cell*, 7: 313-325.
28. Steenkamp, E. T., Wright, J. & Baldauf, S. L. (2006) The protistan origins of animals and fungi. *Mol. Biol. Evol.* **23**, 93–106 (2004).
29. Mikhailov, K. V. *et al*. The origin of Metazoa: a transition from temporal to spatial cell differentiation. *Bioessays* **31**(7), 758–68, https://doi.org/10.1002/bies.200800214 (2009).
30. Shilo, B. Z. & Weinberg, R. A. DNA sequences homologous to vertebrate oncogenes are conserved in Drosophila melanogaster. *Proc. Natl. Acad. Sci. USA* **78**, 6789–6792 (1981).
31. Bishop, J. M. Cellular oncogenes and retroviruses. *Annu. Rev. Biochem.* **52**, 301–354 (1983).
32. Atchley, W. R. & Fitch, W. M. Myc and Max: molecular evolution of a family of proto-oncogene products and their dimerization partner. *Proc. Natl. Acad. Sci. USA* **92**, 10217–10221 (1995).
33. 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 (2003).
34. Srivastava, M. *et al*. The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature* **466**, 720–726 (2010).
35. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**(1), 25–29 (2000).
36. Massague, J. TGFb in Cancer. *Cell* **134**, 215–230 (2008).
37. Nusse, R. & Varmus, H. E. Many tumors induced by the mouse mammary tumor virus contain a provirus integrated in the same region of the host genome. *Cell* **31**(1), 99–109 (1982).
38. Nusse, R. & Varmus, H. E. Wnt genes. *Cell* **69**, 1073–1087 (1992).
39. Bell, G. & Mooers, A. O. Size and complexity among multicellular organisms. *Biological Joumal of the Linnean Society* **60**, 345–363 (1997).
40. Chen, L., Stephen, J. B., Jaime, M. T.-C., Atahualpa, C.-M. & Araxi, O. U. Correcting for Differential Transcript Coverage Reveals a Strong Relationship between Alternative Splicing and Organism Complexity. *Mol. Biol. Evol.* **31**(6), 1402–1413 (2014).
41. Vickaryous, M. K. & Hall, B. K. Human cell type diversity, evolution, development, and classification with special reference to cells derived from neural crest. *Biol. Rev. Camb. Philos. Soc.* **81**, 425–455 (2006).
42. Duesberg, P. H. Latent cellular oncogenes: the paradox dissolves. *J. Cell Sci. Suppl.* **7**, 169–187 (1987).
43. Marx, J. Oncogenes reach a milestone. *Science* **266**(5193), 1942–1944 (1994).
44. Markert, C. L. Neoplasia: a disease of cell differentiation. *Cancer Res.* **28**, 1908–1914 (1968).
45. Harshbarger, J. C. Pseudoneoplasms in ectothermic animals. *Natl. Cancer Inst. Monogr.* **65**, 251–273 (1984).
46. Hoover, K. L. Hyperplastic thyroid lesions in fish. *Natl. Cancer Inst. Monogr.* **65**, 275–289 (1984).
47. Grizzle J. M. and Goodwin A. E. Neoplasms and Related Lesions. In: Leatherland, J. F., Woo, P. T. K. (Eds), Fish Diseases and Disorders 2: 37-104 (1998).
48. Chanet, B. & Meunier, F. J. The anatomy of the thyroid gland among "fishes": phylogenetic implications for the Vertebrata. *Cybium: international journal of ichthyology* **38**(2), 90–116 (2014).
49. Zhang, Y. & Long, M. New genes contribute to genetic and phenotypic novelties in human evolution. *Curr. Opin. Genet. Dev.* **29**, 90–96 (2014).
50. Karnaukhova, I. K. *et al*. A new cancer-testis long noncoding RNA, the OTP-AS1 RNA. RNA *The journal* (in press) (2018).
51. Hodzic, D. *et al*. TBC1D3, a hominoid oncoprotein, is encoded by a cluster of paralogues located on chromosome 17q12. *Genomics* **88**, 731–736 (2006).
52. Wainszelbaum, M. J. *et al*. The hominoidspecific oncogene TBC1D3 activates ras and modulates epidermal growth factor receptor signaling and trafficking. *J. Biol. Chem.* **283**, 13233–13242 (2008).
53. Shang, B. *et al*. CT45A1 acts as a new proto-oncogene to trigger tumorigenesis and cancer metastasis. *Cell Death Dis.* **5**, e1285, https://doi.org/10.1038/cddis.2014.244 (2014).
54. Chen, J. S., Hung, W. S., Chan, H. H., Tsai, S. J. & Sunny Sun, H. In silico identification of oncogenic potential of fyn-related kinase in hepatocellular carcinoma. *Bioinformatics* **29**, 420–427 (2013).
55. Zhao, M., Kim, P., Mitra, R., Zhao, J. & Zhao, Z. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Res. Jan 4* **44**(D1), D1023–1031 (2013).

56. Zhong, Y., Butts, T. & Holland, P. W. H. HomeoDB: a database of homeobox gene diversity. *Evolution & Development* **10**(5), 516–518 (2008).
57. Zhong, Y. & Holland, P. W. H. HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evolution& Development* **13**(6), 567–568 (2011).
58. Díez, J., Walter, D., Muñoz-Pinedo, C. & Gabaldón, T. DeathBase: a database on structure, evolution and function of proteins involved in apoptosis and other forms of cell death. *Cell Death Differ.* **17**(5), 735–736 (2010).
59. Almeida, L. G. *et al.* CTdatabase: a knowledge-base of high-throughput and curated data on cancer-testis antigens. *Nucleic Acids Res.* **37**(suppl_1), D816–D819 (2009).
60. Guo, Y. *et al.* Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics* **109**(2), 83–90 (2017).
61. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30**(1), 38–41 (2002).
62. Heinicke, S. *et al.* The Princeton Protein Orthology Database (P-POD): A Comparative Genomics Analysis Tool for Biologists. *PLoS ONE* **2**, e766, https://doi.org/10.1371/journal.pone.0000766 (2007).
63. Capra, J. A. *et al.* ProteinHistorian: Tools for the Comparative Analysis of Eukaryote Protein Origin. *PLoS Computational Biology* **8**(6), e1002567, https://doi.org/10.1371/journal.pcbi.1002567 (2012).

Acknowledgements

This study was supported by 5-100-2020 program grant at Peter the Great St. Petersburg Polytechnic University and a grant No833 from Ministry of Education and Science of the Russian Federation to A.P.K. We thank staff the Department of Information and Control Systems of Peter the Great St. Petersburg Polytechnic University for providing the access to processing power required to perform the analysis of the non-coding sequences. We thank D. Frishman, P. Drobintsev, and I. Selin for their help during this work.

Author contributions

A.P.K. is an author of original hypothesis of the evolutionary role of tumors and *TSEEN* genes, and general design of experiments. A.A.M. performed bioinformatics analysis, and S.V.M. performed statistical analysis. All authors contributed to writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-52835-w.

Correspondence and requests for materials should be addressed to A.P.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.