Human Gene Age Dating Reveals an Early and Rapid Evolutionary Construction of the Adaptive Immune System

Li Zhang [1,2,3,4], Jonathan J. Park[1,2,3,4], Matthew B. Dong[1,2,3,4,5,6], Deanna Arsala [7], Shengqian Xia [7], Jianhai Chen[7], Dylan Sosa [7], Jared E. Atlas[7,8], Manyuan Long [7,*], and Sidi Chen[1,2,3,4,5,9,10,*]

[1]System Biology Institute, Integrated Science & Technology Center, West Haven, Connecticut, USA
[2]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut, USA
[3]Center for Cancer Systems Biology, Integrated Science & Technology Center, West Haven, Connecticut, USA
[4]Yale M.D.-Ph.D. Program, New Haven, Connecticut, USA
[5]Immunobiology Program, The Anlyan Center, New Haven, Connecticut, USA
[6]Department of Immunobiology, The Anlyan Center, New Haven, Connecticut, USA
[7]Department of Ecology and Evolution, The University of Chicago, Chicago, Illinois, USA
[8]Committee on Genetics, Genomics and Systems Biology, The University of Chicago, Chicago, Illinois, USA
[9]Yale Comprehensive Cancer Center, New Haven, Connecticut, USA
[10]Yale Stem Cell Center, Yale University School of Medicine, New Haven, Connecticut, USA

*Corresponding authors: E-mails: sidi.chen@yale.edu (S.C.); mlong@uchicago.edu (M.L.).

Accepted: 02 May 2023

Abstract

T cells are a type of white blood cell that play a critical role in the immune response against foreign pathogens through a process called T cell adaptive immunity (TCAI). However, the evolution of the genes and nucleotide sequences involved in TCAI is not well understood. To investigate this, we performed comparative studies of gene annotations and genome assemblies of 28 vertebrate species and identified sets of human genes that are involved in TCAI, carcinogenesis, and aging. We found that these gene sets share interaction pathways, which may have contributed to the evolution of longevity in the vertebrate lineage leading to humans. Our human gene age dating analyses revealed that there was rapid origination of genes with TCAI-related functions prior to the Cretaceous eutherian radiation and these new genes mainly encode negative regulators. We identified no new TCAI-related genes after the divergence of placental mammals, but we did detect an extensive number of amino acid substitutions under strong positive selection in recently evolved human immunity genes suggesting they are coevolving with adaptive immunity. More specifically, we observed that antigen processing and presentation and checkpoint genes are significantly enriched among new genes evolving under positive selection. These observations reveal evolutionary processes of TCAI that were associated with rapid gene duplication in the early stages of vertebrates and subsequent sequence changes in TCAI-related genes. The analysis of vertebrate genomes provides evidence that a "big bang" of adaptive immune genes occurred 300-500 million years ago. These processes together suggest an early genetic construction of the vertebrate immune system and subsequent molecular adaptation to diverse antigens.

Key words: T cell adaptive immunity, new gene evolution, sequence substitution, vertebrate genomes, aging.

Significance

The origination of the adaptive immune system may have played a pivotal role in the evolution of longevity in vertebrate groups. However, the evolutionary history of the genes involved in the T cell adaptive immune (TCAI) system is still unclear. We found that following the origination of the adaptive immune system, lifespans, and age to sexual maturation steadily increased in vertebrates. Gene age dating revealed that the TCAI system experienced multiple bursts of new genes and that many of these new genes encode negative regulators. Although we did not identify any new TCAI members following the divergence of placental mammals, many of their orthologs are evolving under positive selection. Our results show an early evolutionary build-up of the TCAI genetic toolkit in vertebrates.

Introduction

Prior to the development of the adaptive immune system, the innate immune system was the sole immune strategy. Innate immunity is one of the two immune strategies in vertebrates, but remains the sole immune strategy in plants, fungi, and insects (Mushegian and Medzhitov 2001). The innate immune system is considered the first line of defense when an animal is exposed to a pathogen (Medzhitov and Janeway 2000). After infection, the innate immune system generates a broad and nonspecific inflammatory response, wherein infected cells will release factors to recruit immune cells to the site of infection to phagocytose and/or directly attack and eliminate the pathogen (Medzhitov and Janeway 2000; O'Connor and Cornwallis 2022). Although the innate immune system is complex and sufficient in many animal systems, its nonspecificity and short-term immunological memory can incur costs on host cells (Chatzinikolaou et al. 2014; McDade et al. 2016; O'Connor and Cornwallis 2022).

As animals recurrently encounter pathogens, they experience repeated bouts of inflammation, which can become chronic due to the nonspecific and short-lived immunological memory of the innate immune system. Multiple studies have demonstrated a clear link between inflammation and the aging process in vertebrates (Singh et al. 2019; Sorci and Faivre 2009). Chronic inflammation has been shown to shorten telomeres (Jurk et al. 2014; Kordinas et al. 2016), create oxidative stress (Hardbower et al. 2013), and induce DNA damage (Chatzidoukaki et al. 2021; Kawanishi et al. 2017)—all of which are known to shorten lifespans and are associated with the aging process.

To achieve longevity, animal immune systems need to specifically neutralize pathogens while minimizing collateral damage incurred to host cells with each recurrent exposure. How animals achieve longevity is still heavily debated, but evolutionary immunologists speculate that the origination of the adaptive immune system may have permitted the longevity of vertebrates (O'Connor and Cornwallis 2022).

The adaptive immune system originated in jawed vertebrates ~500 million years ago (Ma) following major evolutionary events: the invasion of recombination-activating genes (RAG) and two rounds of whole genome duplication (Boehm and Swann 2014; Cooper and Alder 2006; Flajnik and Kasahara 2010; Ohno 2013). Unlike the innate immune system, the adaptive immune system is highly specific to pathogens and confers long-term immunological memory after an initial immune response to an antigen, which is mediated by B and T cells (Cooper and Alder 2006).

Progress in investigating the evolutionary history of the adaptive immune system has focused primarily on the major histocompatibility complex (MHC), which comprised a highly polymorphic, gene-rich region that encodes cell surface proteins. The origination of MHC genes has been described as a "Big Bang" (Abi Rached et al. 1999) or an accordion-like (Klein et al. 1998) process, primarily driven by gene duplication with subsequent gene loss in some lineages (Klein et al. 1998). MHC genes are divided into three major classes with Class I and Class II genes encoding proteins that bind small protein fragments and display them on the cell surface for recognition by T cells, which are specialized cell types that can eliminate infected or cancerous cells (Kaufman 2018). T cells specifically recognize peptides presented by MHC Class I and Class II molecules, which is essential for triggering an immune response, where they are stimulated into "cytotoxic" $CD8^+$ or "helper" $CD4^+$ T cell types, respectively (Kaufman 2018). The genes expressed within T cells make up a subset of the adaptive immune system referred to as T cell adaptive immunity (TCAI), which mediates a cytotoxic response by inducing cell death or aid in the activation of antibody-secreting cells to destroy infected cells (Bartl et al. 2003; Kaufman 2018; Muller et al. 2018).

Although extensive evolutionary studies have been done to characterize the origination patterns of the MHC genes, the origination patterns of TCAI genes and the evolutionary forces acting upon them are less understood. Previous work has shown that 9 out of 15 genes involved in T cell activation are evolving under positive selection (Forni et al. 2013). Because vertebrates have achieved different longevities following the emergence of the adaptive immune system, we

hypothesize that genes underlying TCAI, cell proliferation, and senescence have coevolved. As lifespans extend, rates of cancer increase while immunological memory and specificity of the adaptive immune system decline, which suggests that these processes are interdependent (DePinho 2000; Effros 2007). Compared with the genes involved in TCAI, the genes involved in cancer, including proto-oncogenes (OG) and tumor suppressor (TS) genes, whose products regulate cell proliferation, growth, and aging, are ancient, predating vertebrates and are evolving under a high degree of purifying selection (de Magalhaes and Church 2007; Thomas et al. 2003).

Here, we set out to understand the origination patterns and evolutionary forces acting upon the TCAI genes, aging, and carcinogenesis using a large-scale, comparative genomic approach. First, we show that longevity and the age to sexual maturity increased after the origination of the adaptive immune system. Next, using publicly available vertebrate genomes, we identified known genes involved in TCAI, longevity, and oncogenesis that are distinct to each process but function in similar pathways. Finally, we show that there were multiple bursts of new TCAI genes emerging through gene duplication just prior to the divergence of placental mammals, which is similar to a pattern observed for other genetic components of the adaptive immune system. We found no new members of TCAI genes after the divergence of placental mammals but identified subsequent rapid sequence evolution of gene duplicates involved in TCAI. Our results highlight a similar "Big Bang" pattern during the expansion process of genes underlying TCAI, as has been previously shown for MHC genes.

Results

Increased Lifespan and Age to Sexual Maturity in Vertebrates Following the Evolution of the Adaptive Immune System

The evolution of the adaptive immune system may have been essential for the evolution of longevity in long-lived vertebrates, such as mammals. To investigate whether lifespans have extended in jawed vertebrates, we computed the correlation between the evolution of longevity in vertebrates and the origination of TCAI. As a control, we also investigated the evolution of adult weight. We extracted the corresponding parameters for animal species from AnAge database (Tacutu et al. 2018) and interspecies divergence time from TimeTree database (Hedges et al. 2015).

Using humans as the focal species, we observed that the age to sexual maturity and length of lifespan increased with divergence time following the origination of the adaptive immune system 500 Ma, with a small but significant increase in body weight (fig. 1*A* and supplementary table S11, Supplementary Material online). When characterizing mammals by their respective phylogenetic order, we find that primates have the oldest age to sexual maturity in both males and females relative to other mammal orders (the Student's *t*-test to Artiodactyla, $P < 0.0001$). Primates also have a relatively high level of longevity and moderate adult weight (fig. 1*B*). As expected, we observed a similar pattern for cetaceans (fig. 1*B*).

TCAI, Carcinogenesis, and Aging Genes Operate in Shared Pathways

The increase in longevity and age to sexual maturity in vertebrates following the origination of TCAI suggests that the genes underlying cell proliferation and senescence may have coevolved with those underlying the adaptive immune system. We investigated whether the genes previously known to regulate cell proliferation and senescence interact with TCAI genes. Based on the available species with whole-genome sequences at the time of study, we focused on the well-studied human system and collected a set of TCAI genes from multiple sources: 1) 69 genes involved in antigen processing and presentation and 101 T cell receptor signaling (TCRS) from KEGG database (Kanehisa et al. 2017); 2) 55 genes involved in seven steps of TCAI against cancer (TCAI-AC) (Chen and Mellman 2013); 3) 30 checkpoint genes curated from reviews (Cogdill et al. 2017; Pardoll 2012); 4) 386 infiltrating T regulatory cell (ITRC) genes; and 5) 77 exhausted T cell genes from single cell RNA-seq of liver cancer samples (Zheng et al. 2017). In total, we identified 616 nonredundant TCAI genes in humans. We also collected 248 OGs ($q < 0.22$) and 292 TSs ($q < 0.18$) as previously defined (Davoli et al. 2013). Additionally, we collected 160 cellular senescence and 89 longevity regulating genes from KEGG database (Kanehisa et al. 2017), representing 226 unique aging genes. The list of genes from each of the described categories can be found in supplementary table S2, Supplementary Material online.

We found that of the annotated TCAI, TS, OG, and aging genes, the majority (90.0% TCAI, 88.7% TS, 89.5% OG, and 72.1% aging) of these annotated genes were unique to each category, suggesting that TCAI, carcinogenesis, and aging are controlled by distinct gene sets. Surprisingly, we found that roughly 19.5% (total of 44) of aging genes are also TCAI genes. The shared TCAI and aging genes predate the origination of the adaptive immune system, which suggests that some of the genetic components regulating senescence and longevity were subsequently coopted by the adaptive immune system.

To investigate whether the genes underlying TCAI interact with those genes involved in aging and cell proliferation, we retrieved Reactome pathways and NHGRI GWAS Catalogs for TCAI, TS, OG, and aging genes using the KOBAS 3.0 server. Although we observed no overlaps

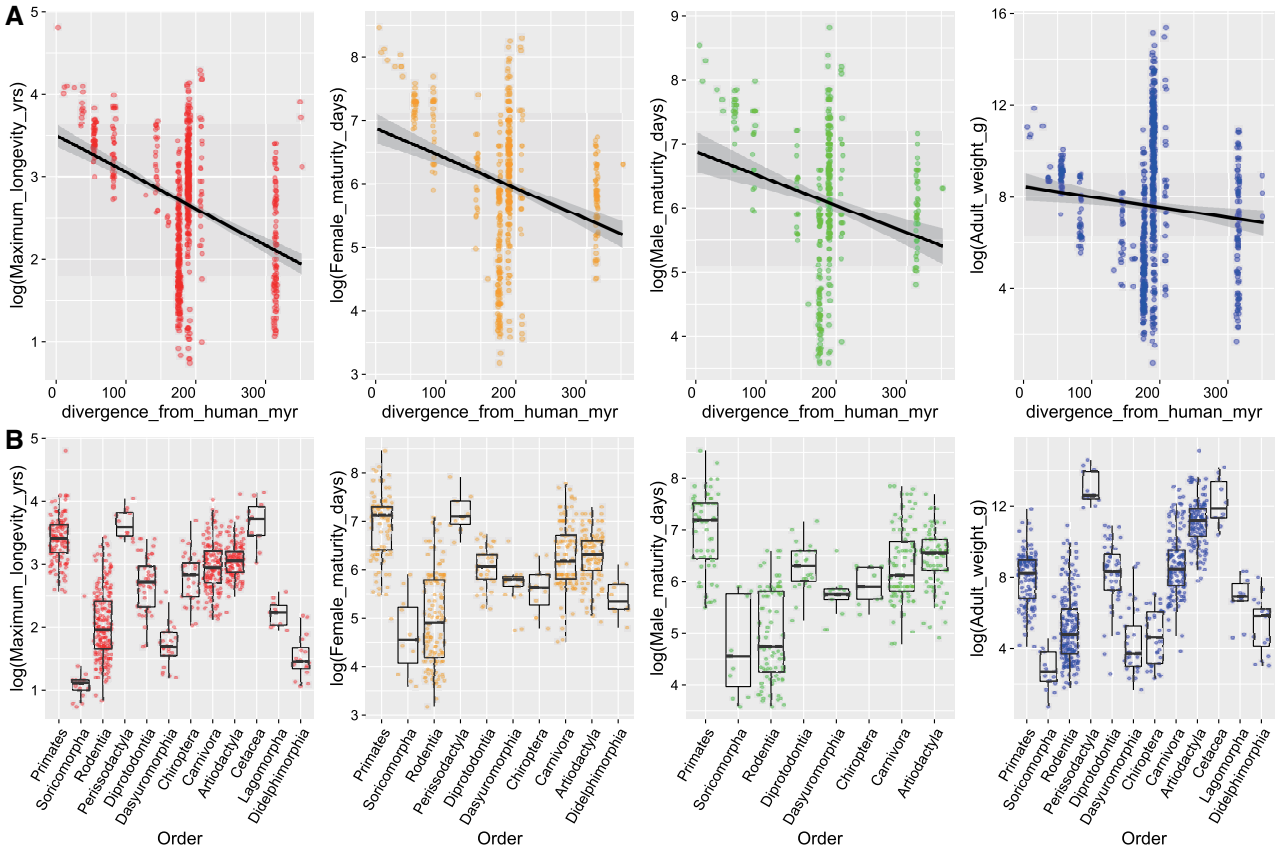

Fig. 1.—The longevity, age of maturation, and adult weight of animal species. The log value of longevity (years), female maturation age (days), male maturation age (days), and adult weight (grams) are plotted against the divergence of the corresponding species to human in (*A*). Otherwise, animal species are clustered in different orders and plotted (*B*).

between their known roles in human diseases (fig. 2*C*), we identified shared pathways among these gene sets (fig. 2*B* and supplementary tables S3 and S10, Supplementary Material online).

The Early Origination of TCAI Is Characterized by a Rapid Gain of Negative Regulators

TCAI originated in basal vertebrates 500 Ma based on the analysis of TCRS-related genes (Boehm and Swann 2014; Cooper and Alder 2006). However, it is unclear whether that is representative of the broad spectrum of TCAI genes. To understand the evolutionary history of TCAI genes, we dated human genes to different vertebrate species groups. We retrieved human genes and their orthologs identified through syntenic alignments and dated them using a phylogeny of 28 vertebrates (supplementary fig. S2, Supplementary Material online). Zebrafish was used as an outgroup species and diverged 435 Ma (Hedges et al. 2015) (supplementary table S4, Supplementary Material online). Out of the 20,198 human genes that contained at least 30 amino acids, 2,949 new genes were absent in both *Xenopus* and zebrafish, suggesting that they arose more recently in vertebrates. When examining these 2,949 newly evolved genes, we found that 37% (1,117) of them do not have any detectable paralogs within the human genome and are thus classified as orphan genes. We further retrieved the functional domains for these 1,117 orphan genes and found that 313 (28.0%) of these genes lack domains with known functions (supplementary table S5, Supplementary Material online). Some of these orphan genes may include ancient genes that arose much earlier in vertebrates (or prior) but experienced rapid sequence changes to escape homology detection. However, even if we assumed that all orphan genes with unknown domains are ancient genes that evolved rapidly, this would account for no >10.6% of the original 2,949 orphan genes.

We found that the vast majority of TCAI (88.3%), TS (96.7%), OG (95.0%), and aging (99.0%) genes originated before the divergence of western clawed frog (fig. 3*A*), which are significantly more conserved than the observed 85.4% of total human genes (chi-square test, $P < 0.001$; supplementary table S6, Supplementary Material online). In line with our expectations, we found that TCAI has significantly

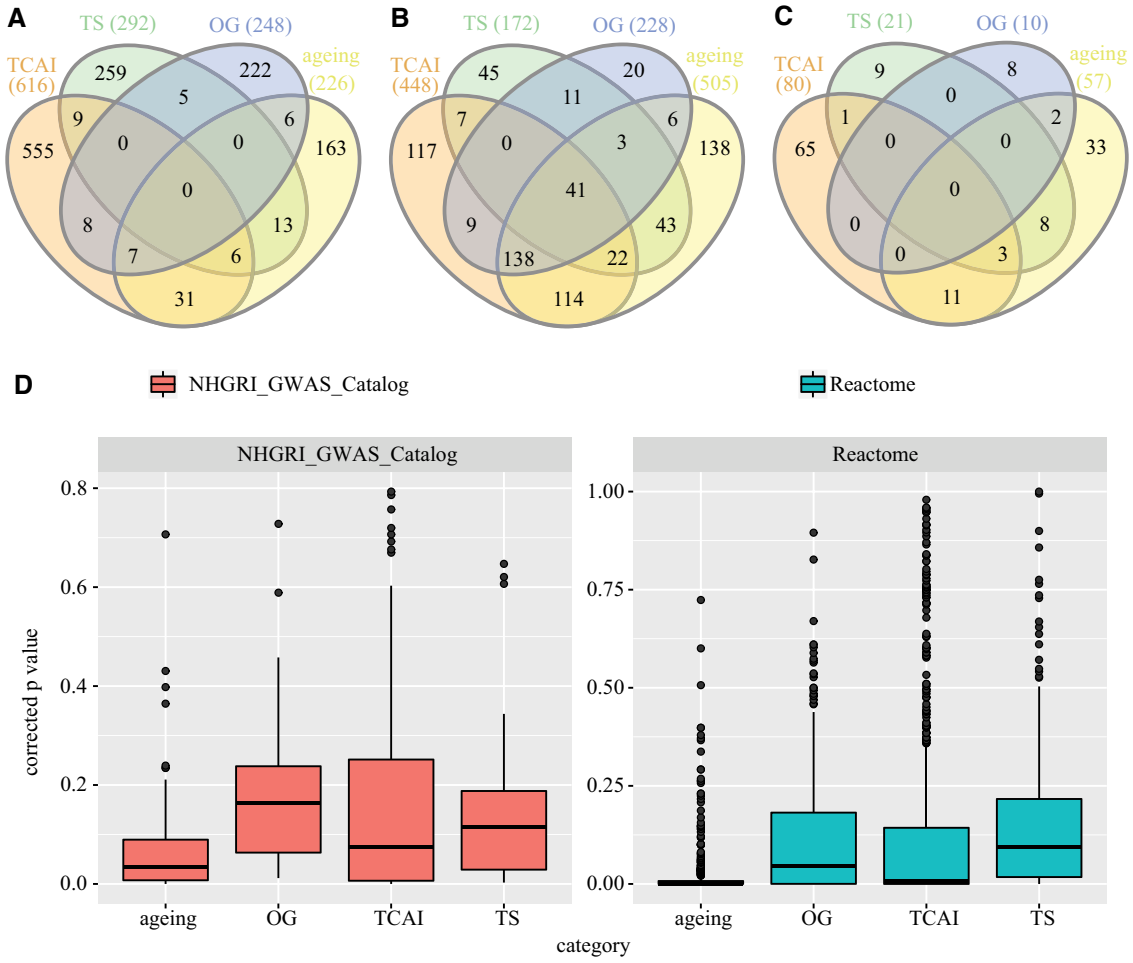

Fig. 2.—The collection of TCAI, TS, OG, and aging genes. There was little categorical overlap of TCAI, TS, OG, and aging genes (*A*). Reactome pathway overlap of TCAI, TS, OG, and aging genes (*B*). NHGRI GWAS catalog overlap (*C*). For details on the Reactome and NHGRI GWAS overlap, see supplementary table S10, Supplementary Material online. (*D*) Enrichment results for Reactome pathway and NHGRI GWAS Catalog.

more new genes than the three other sets of genes we examined (chi-square test, $P < 0.001$; supplementary table S5, Supplementary Material online). However, we found that all new TCAI genes originated before the divergence of placental mammals (fig. 3*A*). Compared with the total number of genes, more TCAI genes originated in branch 2 (supplementary table S7, Supplementary Material online) and a second burst of TCAI genes (as well as TS, OG, and aging) just prior to the divergence of placental mammals in branch 5. Among the 71 TCAI genes that originated after the divergence of western clawed frog, 39 genes are ITRC genes, which suggest an important role in negative regulation of TCAI genes during evolution.

Additionally, all four sets of genes correlate very well with the total genes in terms of their presence in each species (table 1), indicating that there were no biased gene origination events associated with the four major systems. Finally, most TCAI genes are present in the most basally branching vertebrates, which supports the hypothesis that the TCAI system was derived from the innate immune system during the 2R genome duplication in early vertebrates and was soon fixed after its origination (Muller et al. 2018).

New genes often originate through duplication, recombination, or fission (Long et al. 2013). At times, the new gene may undergo rapid evolution, acquiring significant sequence changes from the parent gene(s) and leaves no detectable homology, as is the case for orphan genes. In mammals (Khalturin et al. 2009; Zhang et al. 2011) and *Drosophila* (Khalturin et al. 2009; Zhang et al. 2010), ~10–20% of new genes are orphans, and although in some species, the proportion of orphan genes can be much higher, such as in *Oryza*, with a current proportion of 55% (Zhang et al. 2019). Unexpectedly, we found that compared with the general orphan gene proportion in humans (10%), the majority of TCAI new genes relative to duplicates are orphans (orphan/duplicate: 57/37) (supplementary table S2, Supplementary Material online). Although the numbers are much smaller, we did observe a similar trend for TS (5

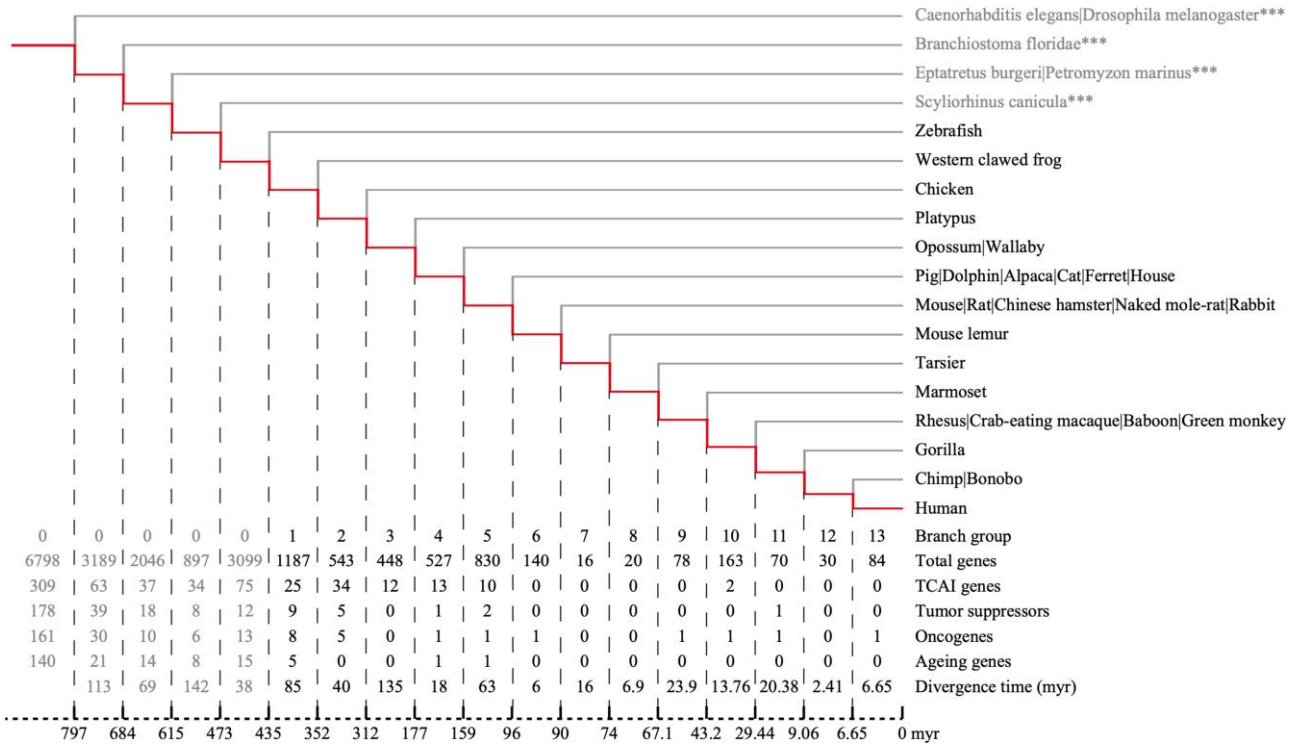

Fig. 3.—The age dating of human genes. Human genes are grouped into different categories including total, TCAI, TS, OG, and aging genes and assigned to different branch groups based on divergence time. Compared with total genes, significantly less or equal amount of TCAI, TS, OG, and aging genes emerged in each branch. ***Gene numbers are slightly different from synteny-based gene age assignment.

orphans/4 duplicates) and aging (2 orphans/0 duplicates) genes while most OG new genes are duplicate genes (1 orphan/11 duplicates). The origination pattern of four example orphan genes can be found in supplementary figure S3, Supplementary Material online as well as the corresponding alignments with their orthologous genes (supplementary figs. S4–7, Supplementary Material online).

Continuous Adaptation of Immunity through Sequence Substitutions in Stationary Gene Evolution of Placentals

We were unable to detect TCAI genes that were unique to placental mammals. Despite the similar genetic makeup of the adaptive immune system among placental mammals, the orthologs and paralogs from prior gene duplication events have divergent sequences, which may reflect species-specific adaptations to their environments. We expect that TCAI genes are continuously evolving under position selection. We applied the site-model test of phylogenetic analysis by maximum likelihood (PAML) to homologous gene clusters and successfully retrieved test results for 17,206 human genes (Yang 2007). We required the positively selected genes to be significant in all four tests with different parameter sets (described in supplementary table S8, Supplementary Material online). We were able to identify 1,131 genes evolving under positive selection, which represents 6.6% of the total number of human genes. Compared with total genes, no significant differences were found for TCAI (49/561, Fisher's exact test, $P = 0.0729$), TS (12/254, Fisher's exact test, $P = 0.3044$), OG (8/204, Fisher's exact test, $P = 0.1929$), and aging genes (7/193, Fisher's exact test, $P = 0.1375$). However, we found that antigen processing and presentation (12/57, Fisher's exact test, $P = 0.001$) and checkpoint genes (9/27, Fisher's exact test, $P = 0.0003$) were significantly enriched for positive selection (table 2). The top 18 enriched KEGG pathways (corrected $P < 0.01$) for 1,131 positively selected genes are all related to immunity (supplementary table S9, Supplementary Material online). These observations were in line with our expectations and suggest that the immune system is an adaptive trait of vertebrate evolution that is undergoing continuous change.

We next investigated whether human-specific genes that are evolving under positive selection are also associated with immune functions. We were able to identify 224 human-specific genes undergoing positive selection, which account for 1.5% of the total genes in the human genome. Compared with total genes, we found significant enrichment for TCAI (84/524, Fisher's exact test, $P << 0.01$) and aging

Table 1
The Enrichment of Positively Selected Genes for Various Gene Categories

Species	Total	TCAI	TS	OG	Aging	MHC	TLR	KIR	PDCD1	CTLA4	CD4	CD8A	CD8B
Human	1	1	1	1	1.00	1	1	1	1	1	1	1	1
Chimp	0.98	0.99	0.99	0.98	1.00	1	1	1	1	1	1	1	1
Bonobo	0.94	0.98	0.97	0.95	0.99	1	1	0.63	0	1	1	1	1
Gorilla	0.96	0.98	0.99	0.95	0.97	1	1	0.63	1	1	1	1	0
Green monkey	0.94	0.98	0.98	0.95	0.99	1	1	0.63	1	1	1	1	1
Crab-eating macaque	0.96	0.99	0.98	0.95	0.99	1	1	0.75	1	1	1	1	1
Rhesus	0.95	0.99	0.98	0.96	1.00	1	1	1	1	1	1	1	1
Baboon	0.96	0.98	0.98	0.96	1.00	0.95	1	1	1	1	1	1	1
Marmoset	0.93	0.97	0.98	0.95	0.99	0.95	1	0.13	1	1	1	1	1
Tarsier	0.84	0.91	0.87	0.86	0.94	0.79	1	0.25	1	1	1	0	1
Mouse lemur	0.9	0.95	0.97	0.91	0.97	0.84	1	0.13	1	1	1	1	1
Mouse	0.93	0.96	0.99	0.96	0.98	0.89	0.9	0.25	1	1	1	1	1
Rat	0.91	0.97	0.97	0.91	0.99	1	1	0.25	1	1	1	1	1
Chinese hamster	0.9	0.96	0.99	0.94	0.98	0.95	1	0.13	1	1	1	1	1
Naked mole-rat	0.85	0.89	0.95	0.91	0.98	0.63	0.9	0	1	1	1	0	0
Rabbit	0.85	0.9	0.96	0.89	0.93	0.84	0.9	0.25	1	1	1	0	0
Pig	0.92	0.96	0.99	0.93	0.98	0.89	1	0.25	1	1	1	1	1
Dolphin	0.64	0.75	0.59	0.63	0.77	0.47	1	0	0	0	1	0	1
Alpaca	0.54	0.59	0.54	0.59	0.65	0.37	1	0	0	1	0	0	0
Horse	0.89	0.96	0.96	0.94	0.96	1	0.9	0.38	1	1	1	1	1
Ferret	0.87	0.94	0.97	0.9	0.98	0.84	1	0.25	1	1	1	1	1
Cat	0.89	0.95	0.96	0.91	0.99	0.79	1	0	1	1	1	1	1
Wallaby	0.38	0.43	0.26	0.36	0.51	1	0.6	0	0	1	0	1	0
Opossum	0.85	0.9	0.94	0.91	0.96	0.84	1	1	1	1	1	1	0
Platypus	0.71	0.78	0.86	0.81	0.87	0.53	0.8	0	0	1	0	0	1
Chicken	0.73	0.82	0.89	0.86	0.90	0.79	0.9	0.75	1	1	1	1	0
Western clawed frog	0.75	0.79	0.91	0.85	0.92	0.74	1	0	0	0	0	1	0
Zebrafish	0.8	0.84	0.93	0.91	0.97	1	0.9	0	0	1	0	0	0
Pearson correlation with total genes	NA	0.99	0.93	0.96	0.95	0.56	0.66	0.54	0.70	0.31	0.72	0.42	0.51

Table 2
The Enrichment of Positively Selected Genes for Various Gene Categories

Category	Genes	Positively Selected Genes	Percentage	*P* (Fisher's Exact Test)
Total	17,206	1131	6.6%	NA
TCAI	561	49	8.7%	0.0729
TS	254	12	4.7%	0.3044
OG	204	8	3.9%	0.1929
Aging	193	7	3.6%	0.1375
APP	57	12	21.1%	0.001
TCRS	100	7	7.0%	0.8392
TCAI-AC	50	6	12.0%	0.1585
CP	27	9	33.3%	0.0003
ITRC	353	23	6.5%	1
ETC	68	6	8.8%	0.463
CS	134	6	4.5%	0.4783
LR	77	1	1.3%	0.0924

APP, antigen processing and presentation; CP, checkpoint; ETC, exhausted T cell genes; CS, cellular senescence, LR, longevity regulating.

genes (23/177, Fisher's exact test, $P << 0.01$) but not for TS (3/223, Fisher's exact test, $P = 1$) and OG (2/183, Fisher's exact test, $P = 1$). When we took a closer look at subcategories of genes involved in TCAI and aging, we found no significant differences (table 3).

Discussion

After the emergence of the adaptive immune system in jawed vertebrates, we discovered that lifespan and age to sexual maturity have lengthened in many vertebrate groups (fig. 1*A*). Investigating whether the adaptive immune system had a central role in the evolution of vertebrate lifespans requires an in-depth characterization of the origination patterns and evolutionary forces acting upon the genes underlying the adaptive immune system. Although this has been characterized extensively in MHC genes, this has not been done in genes involved in TCAI.

Significant work has been to characterize MHC gene evolution. MHC genes are critical for the TCAI to facilitate adaptive immunity and have followed a "Big Bang" pattern of origination, primarily driven by gene duplications (Abi Rached et al. 1999; Klein et al. 1998) that expand and contract in number. As vertebrates have achieved different longevities, we hypothesized that genes involved in the adaptive immune system may interact with genes that control cell

Table 3
The Enrichment of Human-specific Positively Selected Genes for Various Gene Categories

Category	Genes	Positively Selected Genes	Percentage	*P* (Fisher's Exact Test)
Total	15,340	224	1.5%	NA
TCAI	524	84	16.0%	<<0.01
TS	224	3	1.3%	1
OG	183	2	1.1%	1
Aging	177	23	13.0%	<<0.01
APP	48	0	0.0%	1
TCRS	88	1	1.1%	1
TCAI-AC	45	0	0.0%	1
CP	24	0	0.0%	1
ITRC	320	4	1.3%	1
ETC	63	1	1.6%	1
CS	114	0	0.0%	1
LR	74	0	0.0%	1

APP, antigen processing and presentation; CP, checkpoint; ETC, exhausted T cell genes; CS, cellular senescence; LR, longevity regulating.

senescence and proliferation. To investigate this, we characterized the origination pattern of TCAI genes, the evolutionary forces acting upon them, and whether these new TCAI genes have functional overlaps with genes that have known roles in longevity and carcinogenesis.

Using a collection of genome assemblies of 28 species and their respective gene annotations, we identified a general pattern of TCAI gene origination in vertebrate evolution that is similar to what was previously described for MHC genes (Abi Rached et al. 1999; Klein et al. 1998). We found an initial "Big Bang" of TCAI gene components driven by gene duplications—many of which are continuously evolving under positive selection. An initial series of gene duplications quickly expanded the TCAI genetic toolkit in the basally branching vertebrates, and we were unable to detect any new TCAI genes following the radiation of placental mammals, suggesting an early completion of a functional TCAI genetic system in mammals. We also note that the burst of TCAI genes during branches 2 and 5 coincides with the divergence of monotremes (Capuco and Akers 2009) and placental mammals, respectively. Interestingly, the divergence of these two phylogenetic clades are associated with the emergence of traits such as lactation and pregnancy (Lynch et al. 2015; Stadtmauer and Wagner 2020), the latter of which coincides with the expression of negative immune regulators, which may play an important role in pregnancy (Abu-Raya et al. 2020; Koldehoff et al. 2013).

Despite having no new genes that arose exclusively within the placental TCAI system, we found that TCAI genes in placental mammals are under positive selection: ~9% of TCAI genes, higher than the genome-wide ratio of 6% (supplementary table S8, Supplementary Material online), are under positive selection. We think this is an underestimate due to the conservative statistical methods we implemented, which require a Ka/Ks = 1 as a standard of neutrality. Additionally, 6.6% of human genes exhibit strong signals of positive selection and most have known immune functions. Many of these genes participate in antigen processing and presentation and immune checkpoints. Human-specific genes that were evolving under positive selection were also largely involved in immune function.

Because TCAI originated between the Agnathan and Gnathostome split (Boehm et al. 2018; Cooper and Alder 2006), we used gene age dating to assess when TCAI genes arose across a larger timespan. We found that roughly 79% of TCAI genes originated before the divergence of Agnathans and Gnathostomes (supplementary fig. S1, Supplementary Material online) and that 60% of TCAI genes can be detected in insects. The early presence of these TCAI genes does not indicate the presence of a functioning adaptive immune system prior to the Agnathan/Gnathostome split; on the contrary, they were coopted to facilitate immune functions in jawed vertebrates (Cooper and Alder 2006).

The origination of the adaptive immune system depended on the evolution of new genes through two major events: two rounds of whole-genome duplication events and by the insertion of RAG-like transposon that was either of viral or bacterial origin (Flajnik and Kasahara 2010). In the lineage leading to placental mammals, we found that TCAI genes have expanded through several gene duplication events, and many of them are negative regulators of immunity. As mammals have evolved to adapt to nearly all terrestrial and aquatic habitats, we expected to find an abundant number of new TCAI genes that may reflect their habitat diversity. Surprisingly, no new TCAI genes have emerged since the Cretaceous radiation of placental mammals, suggesting the genetic prototype of TCAI that arose nearly 145 Ma was sufficient to help mammals adapt to nearly every habitat on the planet. However, TCAI is not a "one size fits all" machine for mammalian immunity, but a set of highly versatile parts that can be continuously modified. TCAI genes were likely fine-tuned in sequence and function by natural selection as mammals diversified and continued to adapt to their ever-changing environments, which is supported by our observation that TCAI genes are under positive selection. These observations reveal an initial and complete construction of TCAI genes associated with rapid gene duplication in the early stages of vertebrates, and subsequent sequence changes in the TCAI members.

Materials and Methods

Gene Annotation, Genome Assembly, and Genome Alignment

We investigated the genomes of 28 vertebrates ranging from humans, other mammals, birds, and reptiles to

identify the homologues of TCAI, TS, OG, and genes related to this study in all these species. We conducted between-species reciprocal best whole genome alignments for *Homo sapiens* (Ensembl version 92) and 27 other species, including *Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla gorilla*, *Macaca mulatta*, *Macaca fascicularis*, *Papio anubis*, *Chlorocebus sabaeus*, *Callithrix jacchus*, *Tarsius syrichta*, *Microcebus murinus*, *Mus musculus*, *Rattus norvegicus*, *Cricetulus griseus CHO K1*, *Heterocephalus glaber*, *Oryctolagus cuniculus*, *Sus scrofa*, *Tursiops truncates*, *Vicugna pacos*, *Felis catus*, *Mustela putorius furo*, *Equus caballus*, *Monodelphis domestica*, *Macropus eugenii*, *Ornithorhynchus anatinus*, *Gallus gallus*, *Xenopus (Silurana) tropicalis*, and *Danio rerio*, were downloaded from UCSC (Kent et al. 2002). Genome assembly and gene annotation were downloaded from Ensembl version 92 (Zerbino et al. 2018). Annotated genes that had ambiguous "N" in their sequences or were shorter than 30 amino acids or located in patched sequences (https://www.ncbi.nlm.nih.gov/grc) were removed. Finally, there were 20,198 human genes retained for subsequent analyses. Later, we updated our analysis to address the following: 1) to include a broader species spectrum as outgroup species (*Caenorhabditis elegans*, *Drosophila melanogaster*, *Branchiostoma floridae*, *Eptatretus burger*, *Petromyzon marinus*, and *Scyliorhinus canicula*) and ensure our annotations were reliable; 2) to repeat and check the consistency of human gene annotation analysis using updated annotation (Ensembl version 104). We finally got a total of 20,165 genes, which is 33 genes less than the formal number 20,198 of human genes. For the six outgroup species, the UCSC does not include the public gene-syntenic data sets aligned to the human genome; we thus adopted sequence similarity Blast to compare the six genomes with the updated genome version of human. The NCBI Taxonomy node names and Tax IDs are available in supplementary table S12, Supplementary Material online.

Longevity, Sexual Maturation, and Body Mass Analyses

Phenotypic data sets for animal species were downloaded from AnAge database (https://genomics.senescence.info/species/dataset.zip) (de Magalhaes and Costa 2009) and are also available in our supplementary table S11, Supplementary Material online. Then, the phylogenetic relationships of different species were retrieved from TimeTree database (Hedges et al. 2015). Subsequently, unknown, low-quality, and questionable data were removed.

Pathway Enrichment Analyses

Genes were uploaded to KOBAS 3.0 website (Xie et al. 2011) for gene-list enrichment analyses. Enrichment results for KEGG and Reactome pathway and NHGRI GWAS Catalog was retrieved. Only pathways or catalogs with corrected *P* value smaller than predefined values were considered as significant and used for subsequent analyses.

Human Orthologous Gene Clusters

Between-species orthologous genes for each human gene were retrieved by protein sequence similarity of orthologous sequences calculated from BlastP (Camacho et al. 2009) matches. This is a four-step process: 1) If a query human gene matched the target gene with *e*-value ≤ 1e–6, identity percentage ≥20%, and the length of identical amino acids ≥30 at orthologous, syntenic loci, an orthologous gene was confirmed; 2) otherwise, if the query human gene matched the target gene with *e*-value ≤ 1e–6, identity ≥20%, and the length of identical amino acids ≥30 at non-orthologous and nonsyntenic loci, an orthologous gene was confirmed; 3) otherwise, if the query human gene matched an unannotated orthologous DNA sequence with *e*-value ≤ 1e–6, identity ≥50%, and the length of identical amino acids ≥30, one orthologous match was confirmed; 4) otherwise orthologous gene was supposed to be absent in target species. Pairwise orthologous genes were then clustered based on human genes. Human paralogous genes are defined based on the best protein match with parameters including *e*-value ≤ 1e–6 and the length of identical amino acids ≥30.

Sequence Alignments of Orthologous Genes

The longest isoform of each annotated gene in each orthologous cluster was extracted while cases of gene fusion or fission were excluded. Protein sequence alignments were generated by MAFFT (Katoh et al. 2002), which were then used to generate coding sequence (CDS) alignments by RevTrans (Wernersson and Pedersen 2003). Spurious sequences and columns were further removed by TrimAl (Capella-Gutierrez et al. 2009) with parameters "-resoverlap 0.75 -seqoverlap 85".

Phylogenetic Analyses

CDS alignments for orthologous clusters were concatenated into one joint 28-way sequence alignment, which was fed into RAxML to build a species tree with parameters "-f a -m GTRGAMMA -p 12345 -x 12345 -# 100".

PAML Analyses

For each gene family or superfamily, we estimated substitution rates at synonymous and nonsynonymous sites using the maximum likelihood analysis developed in the PAML package (Yang 2007). The CDS alignment for each orthologous cluster and the species tree were fed into PAML to detect positive selection signals (Yang 2007) with four different parameter sets: 1) M1 (neutral) versus M2 (selection) (Nielsen and Yang 1998; Yang et al. 2000),

F3 × 4; 2) M1 versus M2, F61; 3) M8 (β and ω) versus M8a (β and ω = 1) (Swanson et al. 2003), F3 × 4; and 4) M8 versus M8a, F61. The "F61" and "F3 × 4" are codon frequency modes. Only genes with $P < 0.05$ in all four tests were considered significant.

HyPhy Analyses

The CDS alignment for each orthologous cluster and the species tree were fed into HYPHY to detect signals of human-specific positive selection (Weaver et al. 2018) with three different models using default parameter settings: 1) BUSTED, 2) aBSREL, and 3) MEME. For BUSTED, only genes under negative selection or neutrally evolved in nonhuman species were kept.

Tracing TCAI Genes in a Large Time Span

The orthologs in seven distantly related species (*C. elegans*, *D. melanogaster*, *B. floridae*, *P. marinus*, *S. canicular*, *D. rerio*, and *E. burgeri*) of each human gene were retrieved by reciprocal best BlastP (supplementary table S1, Supplementary Material online). For each interspecies BlastP result, effective hits were retained with parameters (e-value ≤ 1e–6, identity ≥20%, and length of identical amino acids ≥30) and sorted by the length of identical amino acids. The gene pair with the longest identical amino acids was recognized as an effective orthologous pair, and BlastP hits containing these two genes were removed before subsequent analyses. The process was repeated until no orthologous pairs can be found. All orthologous genes were clustered referring to the corresponding human genes for gene age dating.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

We thank all members in the Chen laboratory for technical assistance and discussion, as well as various colleagues in Department of Genetics, Systems Biology Institute, Cancer Systems Biology Center, MCGD Program, Immunobiology Program, BBS Program, Cancer Center, and Stem Cell Center at Yale for assistance and/or discussion. We thank the Center for Genome Analysis, High Performance Computing Center, and Keck Biotechnology Resource Laboratory at Yale, for technical support. We thank Zihan Liang from Chinese Institute of Brain Science, Beijing, for assistance of manuscript preparation. This study is supported by the Yale Discretionary fund to S.C. This study was supported by U.S. The National Institutes of Health grant R01GM116113-01A1, awarded to M.L. D.A. was supported by F32GM146423.

Author Contributions

L.Z. conceptualized and designed the study with input from S.C. and M.L. L.Z. wrote the original manuscript, with help from J.J.P., M.B.D., D.A., D.S., J.C., J.E.A., and S.X. and revised by M.L. and S.C. L.Z. carried out the analyses and interpreted the data with input from D.A., S.X., J.J.P., D.S., J.C., and M.B.D. M.L. and S.C. acquired the funding.

Data Availability

Data within this article are readily available in the online supplementary material. Genomic data were obtained from public databases (UCSC and Ensembl) as described in our Materials and Methods.

Literature Cited

Abi Rached L, McDermott MF, Pontarotti P. 1999. The MHC big bang. Immunol Rev. 167:33–45.
Abu-Raya B, Michalski C, Sadarangani M, Lavoie PM. 2020. Maternal immunological adaptation during normal pregnancy. Front Immunol. 11:575197.
Bartl S, Baish M, Weissman IL, Diaz M. 2003. Did the molecules of adaptive immunity evolve from the innate immune system? Integr Comp Biol. 43:338–346.
Boehm T, et al. 2018. Evolution of alternative adaptive immune systems in vertebrates. Annu Rev Immunol. 36:19–42.
Boehm T, Swann JB. 2014. Origin and evolution of adaptive immunity. Annu Rev Anim Biosci. 2:259–283.
Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.
Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. Trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.
Capuco AV, Akers RM. 2009. The origin and evolution of lactation. J Biol. 8:37.
Chatzidoukaki O, et al. 2021. R-loops trigger the release of cytoplasmic ssDNAs leading to chronic inflammation upon DNA damage. Sci Adv. 7:eabj5769.
Chatzinikolaou G, Karakasilioti I, Garinis GA. 2014. DNA damage and innate immunity: links and trade-offs. Trends Immunol. 35: 429–435.
Chen DS, Mellman I. 2013. Oncology meets immunology: the cancer-immunity cycle. Immunity 39:1–10.
Cogdill AP, Andrews MC, Wargo JA. 2017. Hallmarks of response to immune checkpoint blockade. Br J Cancer. 117:1–7.
Cooper MD, Alder MN. 2006. The evolution of adaptive immune systems. Cell 124:815–822.
Davoli T, et al. 2013. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. Cell 155:948–962.
de Magalhaes JP, Church GM. 2007. Analyses of human-chimpanzee orthologous gene pairs to explore evolutionary hypotheses of aging. Mech Ageing Dev. 128:355–364.
de Magalhaes JP, Costa J. 2009. A database of vertebrate longevity records and their relation to other life-history traits. J Evol Biol. 22: 1770–1774.
DePinho RA. 2000. The age of cancer. Nature 408:248–254.
Effros RB. 2007. Role of T lymphocyte replicative senescence in vaccine efficacy. Vaccine 25:599–604.

Flajnik MF, Kasahara M. 2010. Origin and evolution of the adaptive immune system: genetic events and selective pressures. Nature Reviews Genetics 11:47–59.
Forni D, et al. 2013. A 175 million year history of T cell regulatory molecules reveals widespread selection, with adaptive evolution of disease alleles. Immunity 38:1129–1141.
Hardbower DM, de Sablet T, Chaturvedi R, Wilson KT. 2013. Chronic inflammation and oxidative stress: the smoking gun for Helicobacter pylori-induced gastric cancer? Gut Microbes 4: 475–481.
Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. 2015. Tree of life reveals clock-like speciation and diversification. Mol Biol Evol. 32: 835–845.
Jurk D, et al. 2014. Chronic inflammation induces telomere dysfunction and accelerates ageing in mice. Nat Commun. 5:4172.
Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 45:D353–D361.
Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.
Kaufman J. 2018. Unfinished business: evolution of the MHC and the adaptive immune system of jawed vertebrates. Annu Rev Immunol. 36:383–409.
Kawanishi S, Ohnishi S, Ma N, Hiraku Y, Murata M. 2017. Crosstalk between DNA damage and inflammation in the multiple steps of carcinogenesis. Int J Mol Sci. 18:1808.
Kent WJ, et al. 2002. The human genome browser at UCSC. Genome Res. 12:996–1006.. Article published online before print in May 2002.
Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. 2009. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 25:404–413.
Klein J, Sato A, O'HUigin C. 1998. Evolution by gene duplication in the major histocompatibility complex. Cytogenet Cell Genet. 80: 123–127.
Koldehoff M, Cierna B, Steckel NK, Beelen DW, Elmaagacli AH. 2013. Maternal molecular features and gene profiling of monocytes during first trimester pregnancy. J Reprod Immunol. 99:62–68.
Kordinas V, Ioannidis A, Chatzipanagiotou S. 2016. The telomere/telomerase system in chronic inflammatory diseases. Cause or effect? Genes (Basel) 7:60.
Long M, VanKuren NW, Chen S, Vibranovski MD. 2013. New gene evolution: little did we know. Annu Rev Genet. 47:307–333.
Lynch VJ, et al. 2015. Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. Cell Rep. 10:551–561.
McDade TW, Georgiev AV, Kuzawa CW. 2016. Trade-offs between acquired and innate immune defenses in humans. Evol Med Public Health. 2016:1–16.
Medzhitov R, Janeway C Jr. 2000. Innate immune recognition: mechanisms and pathways. Immunol Rev. 173:89–97.
Muller V, de Boer RJ, Bonhoeffer S, Szathmary E. 2018. An evolutionary perspective on the systems of adaptive immunity. Biol Rev Camb Philos Soc. 93:505–528.
Mushegian A, Medzhitov R. 2001. Evolutionary perspective on innate immune recognition. Journal of Cell Biology 155:705.
Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936.
O'Connor EA, Cornwallis CK. 2022. Immunity and lifespan: answering long-standing questions with comparative genomics. Trends Genet. 38(7):650–661.
Ohno S. 2013. Evolution by gene duplication. Berlin; New York: Springer Science & Business Media.
Pardoll DM. 2012. The blockade of immune checkpoints in cancer immunotherapy. Nat Rev Cancer. 12:252–264.
Singh PP, Demmitt BA, Nath RD, Brunet A. 2019. The genetics of aging: a vertebrate perspective. Cell 177:200–220.
Sorci G, Faivre B. 2009. Inflammation and oxidative stress in vertebrate host-parasite systems. Philos Trans R Soc Lond B Biol Sci. 364: 71–83.
Stadtmauer DJ, Wagner GP. 2020. The primacy of maternal innovations to the evolution of embryo implantation. Integr Comp Biol. 60:742–752.
Swanson WJ, Nielsen R, Yang Q. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. Mol Biol Evol. 20:18–20.
Tacutu R, et al. 2018. Human ageing genomic resources: new and updated databases. Nucleic Acids Res. 46:D1083–D1090.
Thomas MA, et al. 2003. Evolutionary dynamics of oncogenes and tumor suppressor genes: higher intensities of purifying selection than other genes. Mol Biol Evol. 20:964–968.
Weaver S, et al. 2018. Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. Mol Biol Evol. 35:773–777.
Wernersson R, Pedersen AG. 2003. Revtrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 31:3537–3539.
Xie C, et al. 2011. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 39: W316–W322.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449.
Zerbino DR, et al. 2018. Ensembl 2018. Nucleic Acids Res. 46: D754–D761.
Zhang L, et al. 2019. Rapid evolution of protein diversity by de novo origination in *Oryza*. Nat Ecol Evol. 3:679–690.
Zhang YE, Landback P, Vibranovski MD, Long M. 2011. Accelerated recruitment of new brain development genes into the human genome. PLoS Biol. 9:e1001179.
Zhang YE, Vibranovski MD, Krinsky BH, Long M. 2010. Age-dependent chromosomal distribution of male-biased genes in *Drosophila*. Genome Res. 20:1526–1533.
Zheng C, et al. 2017. Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. Cell 169:1342–1356.e16.

Associate editor: Mar Alba