

RESEARCH

Open Access



# The adaptive evolution of cancer driver genes

Langyu Gu<sup>1\*</sup>, Canwei Xia<sup>2</sup>, Shiyu Yang<sup>3</sup> and Guofen Yang<sup>4\*</sup>

## Abstract

**Background** Cancer is a life-threatening disease in humans; yet, cancer genes are frequently reported to be under positive selection. This suggests an evolutionary-genetic paradox in which cancer evolves as a secondary product of selection in human beings. However, systematic investigation of the evolution of cancer driver genes is sparse.

**Results** Using comparative genomics analysis, population genetics analysis and computational molecular evolutionary analysis, the evolution of 568 cancer driver genes of 66 cancer types were evaluated at two levels, selection on the early evolution of humans (long timescale selection in the human lineage during primate evolution, i.e., millions of years), and recent selection in modern human populations (~ 100,000 years). Results showed that eight cancer genes covering 11 cancer types were under positive selection in the human lineage (long timescale selection). And 35 cancer genes covering 47 cancer types were under positive selection in modern human populations (recent selection). Moreover, SNPs associated with thyroid cancer in three thyroid cancer driver genes (CUX1, HERC2 and RGD3) were under positive selection in East Asian and European populations, consistent with the high incidence of thyroid cancer in these populations.

**Conclusions** These findings suggest that cancer can be evolved, in part, as a by-product of adaptive changes in humans. Different SNPs at the same locus can be under different selection pressures in different populations, and thus should be under consideration during precision medicine, especially for targeted medicine in specific populations.

**Keywords** Cancer driver genes, Positive selection, Primates, Modern human populations, Thyroid cancer

## Background

Cancer is one of the most life-threatening diseases with a high mortality rate in humans. Cancer gene mutations at the individual level are usually deleterious to the organism and negatively affect fitness [1, 2]. Thus, purifying selection of cancer gene mutations is necessary to maintain fitness. Surprisingly, studies have continuously found cancer genes to be under positive selection [3, 4], indicating their adaptive roles during evolution. Several hypotheses have been proposed to explain the evolutionary trade-off between positive selection and cancer risks, such as sexual selection, pathogen-host interactions, and genomic compensation [3, 5–8]. Studies have reported positively selected cancer genes in the ancestral lineage leading to human and other species [3, 7, 9], but few focused on positively selected cancer genes in the human

\*Correspondence:

Langyu Gu  
guly6@mail.sysu.edu.cn; langyugu@gmail.com  
Guofen Yang  
yangguof@mail.sysu.edu.cn

<sup>1</sup> State Key Laboratory for Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, Guangdong 510275, China

<sup>2</sup> Ministry of Education Key Laboratory for Biodiversity and Ecological Engineering, College of Life Sciences, Beijing Normal University, Beijing 100875, China

<sup>3</sup> The Affiliated Brain Hospital, Guangzhou Medical University, Guangzhou 510180, Guangdong, China

<sup>4</sup> Department of Gynecology, First Affiliated Hospital, Sun Yat-Sen University, Guangzhou 510060, Guangdong, China



© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

lineage itself. The extent to which positive selection of cancer genes occurs along the human lineage during primate evolution remains unclear.

Notably, positive selection detection across the primate phylogeny usually detects selection in the earlier evolution of the human lineage on a large timescale, i.e., millions of years (*long timescale selection*) [10]. In contrast, detection at the population level reflects selection on a more recent time scale (*recent selection*) starting at the neolithic demographic transition and migration out of Africa in the last 100,000 years. During this period, modern humans have experienced significant changes in lifestyle and living environments [11, 12]. Diversified living environments and food resources, accompanied by new infectious diseases, represent new selective pressures on modern human populations [13]. These selective pressures work on new genotypes that are better adapted to novel environments. Thus, the last 100,000 years encompasses one of the most interesting time periods in human history [10, 14, 15]. A number of genes involved in local adaptations have been identified, such as genes involved in skin pigmentation [14], malaria response [16], hair morphology [17], and lactase persistence [18]. Genes related to human diseases have also been reported to be under recent positive selection due to their pleiotropic effects, such as genes related to neurodegenerative disease (HTT [19], APOE [20]). Although some cancer genes have also been reported to be under recent positive selection, such as the prostate cancer-related gene MLPH [21], systematic investigation of the role of recent selection on the adaptive evolution of cancer genes remains sparse.

Existing studies on the adaptive evolution of cancer genes have mainly focused on genes in specific cancer types or in limited species across the phylogeny [6, 7, 9, 22]. In addition, studies have not differentiated between selection of cancer driver genes and cancer passenger genes, which may be under different selection pressures [23]. With the rapidly developed next generation sequencing technologies and bioinformatics analysis, a recent study systematically analyzed more than 28,000 tumors of 66 cancer types and identified 568 cancer driver genes which is published in *Nature Reviews Cancer* [24]. In-depth evolutionary analyses of these cancer driver genes covering different cancer types across the phylogeny and in modern human populations may give us a more comprehensive picture of cancer gene evolution.

To obtain a deep understanding of cancer evolution, the adaptive evolution of 568 cancer driver genes in 66 different cancer types across the primate phylogeny (*long timescale selection*), as well as in modern human populations retrieved from the 1000 Genomes Project (*recent*

*selection*), was evaluated. The main study aim was to determine the adaptive evolution of cancer driver genes at two levels, i.e. *selection on the early evolution of Homo sapiens* (detection across the primate phylogeny, i.e., millions of years timescale) and *recent selection* (detection among modern human populations, i.e., 100,000 years timescale). Findings from this study can have profound implications for understanding the evolution of cancer driver genes and provide clues for further precision medicine.

## Results

### Positive selection detection across the primate phylogeny

The transcript IDs of 568 cancer driver genes were retrieved from the literature published in *Nature Reviews Cancer* [24]. To understand the evolutionary dynamics of these 568 cancer driver genes across the primate phylogeny, orthologous sequences in 23 primates were obtained using Oma inference [25]. Positive selection was detected in the main primate lineages under the branch-site model using PAML [26, 27]. Eight genes covering 11 cancer types were determined to be under positive selection in the human lineage under the branch-site model in PAML (Table 1), including the neuronal stress related gene, DROSHA [28], the immune related gene, LY75-CD302 [29], the neurodegenerative diseases related gene, RBFOX1 [30], the neurocognitive related gene, NRG1 [31], the viral and bacterial response related gene, STAT3 [32], the brain development related gene, NIN [33], the zinc finger protein gene, ZNF814, and the immune response and antiviral response related gene, TRAF3 [34]. Among which, three genes, including NIN, NRG1 and RBFOX1 have been reported under positive selection [33, 35, 36]. Compared to the human lineage, more genes were under positive selection in ancestral lineages leading to human and other primates (Fig. 1).

### Recent positive selection in modern human populations

Signals of recent positive selection were detected across the genome based on several methods. We used the haplotype-based test [38] first. Briefly, the frequency of selection-favored alleles can arise quickly, such that long-range association with the surrounding loci does not have time to be eliminated by recombination. The selected allele may thus be located at an unusually long haplotype with low diversity, which is in sharp contrast with the unselected genomic background. Therefore, in the haplotype-based method, various core haplotypes (haplotypes at a locus of interest) present at a single locus are used as internal controls to adjust for heterogeneity in the local recombination rate [15, 38], limiting the confounding effects of demography [39]. In addition, recently admixed populations and populations with close

**Table 1** The cancer driver genes under positive selection in the human lineage detected using PAML

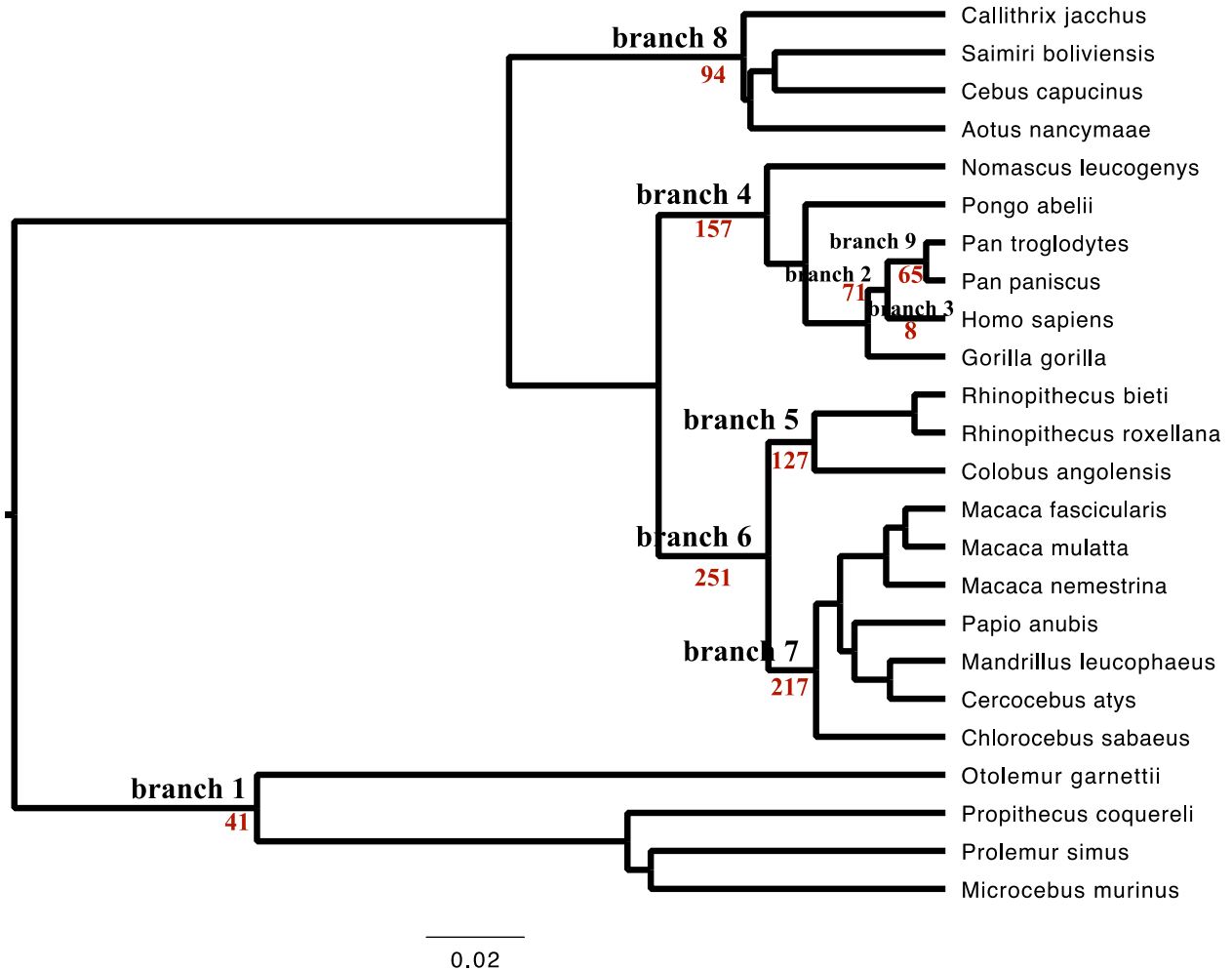
Gene	Branch-site model				P/2	Sites under positive selection Uner BEB model with a posterior probability > 0.95	Cancer types
	Model A	Model B	2ln(A-B)	$\omega$			
DROSHA	-9569.2	-9593.43	48.44	202.65	<0.01	142E, 150H, 1050 M, 1058E	LUSC,WT
LY75-CD302	-15,467	-15,535	136	200	<0.01	1383Y,1388I,1393L,1394S,1402L,1412S,1413 V, 1417Q,1418G,1419 V,1423E,1428S,1429F,1430Q	AML
RBFOX1	-2118.24	-2147.37	58.27	270.69	<0.01	297L,299Q,304A,305L,307P	PRAD
NRG1	-5366.86	-5379.37	25.02	94.63	<0.01	2 K	ST
STAT3	-5563.45	-5577.78	28.65	92.15	<0.01	N/A	DLBCL
NIN	-17,468.17	-17,477.49	18.64	791.16	<0.01	N/A	BRCA,COREAD,MESO,ST
ZNF814	-4723.37	-4734.49	22.25	5.37	<0.01	243H,268H	VV
TRAF3	-4022.31	-4033.59	22.56	999	<0.01	N/A	MM

Model A: the alternative model in the branch-site test of positive selection in PAML

Model B: the null model in the branch-site test of positive selection in PAML

$\omega$ : the ratio of nonsynonymous (dN) to synonymous (dS) substitutions

p/2: the null distribution is a mixture distribution, and thus have to use p divided by 2 to test the hypothesis (PAML manual)



**Fig. 1** Cancer driver genes under positive selection under the branch-site model in PAML across the primate phylogeny. Target foreground branches were labeled. The number of genes under positive selection was given in red. The species tree was retrieved from the literature [37]. The unrooted tree was used for PAML analysis

geographical proximity were excluded, similar to the strategy applied by the literature [10]. Since genome-wide empirical distributions were used for standardization in these haplotype-based tests, there was no formal significance test [38]. Two tests were applied, i.e. the integrated haplotype score (iHS) method and the cross-population extended haplotype homozygosity (xpEHH) method. The iHS method allows for discovery of more recent positively selected variants that have not reached fixation in the population [38], whereas the xpEHH method allows discovery of variants near/at fixation on long haplotypes in one population that remain polymorphic in other populations [14]. Under positive selection, SNPs with large absolute haplotype scores tended to be clustered together [38]. Therefore, to identify sites under positive selection, we counted the number of SNPs for which  $|iHS|$  or  $|xpEHH| > 2$  in a 51-SNPs window. SNPs with  $|iHS|$  or  $|xpEHH| > 2$  located at the top 1% 51-SNPs windows were considered under positive selection (thresholds of top 1% windows were given in the Supplementary File 1).

We also test for differences in allele frequencies between populations ( $F_{ST}$ ) [40]. Since  $F_{ST}$  and allele frequency are highly correlated, to define a high  $F_{ST}$ , we grouped  $F_{ST}$  values in different allele frequency bins and contrasted them to their own frequency classes. The top 5%  $F_{ST}$  values in its own frequency bin were statistically significantly high. Therefore, we applied different  $F_{ST}$  thresholds to identify positively selected sites for alleles falling into different frequency bins (Supplementary File 2). SNPs under positive selection detected by both haplotype-based tests (iHS or xpEHH) mentioned above, and meanwhile with high  $F_{ST}$  values were considered under recent positive selection [14].

The haplotype-based tests have already limited the confounding effects of demography as we mentioned above. We also excluded recent admixed populations and populations with close geographical proximity to further minimize demographic effects. To make our results more convinced, we further applied the population branch statistics (PBS) test to account for demography. PBS value represents the amount of allele frequency change at a given locus in a population since its divergence from the other two populations [41]. SNPs with statistically significantly high PBS values (top 5%) were considered under positive selection (PBS thresholds can be found in the Supplementary File 3).

Because the main goal of this study was to explore the evolution of coding regions, only those SNPs located in exons under positive selection were thus chosen for further study. Seven genes across 14 cancer types were identified under positive selection using the iHS test and the  $F_{ST}$  test (Fig. 2a). One pigmentation related gene, *HERC2*, has been reported under positive selection before [42, 43].

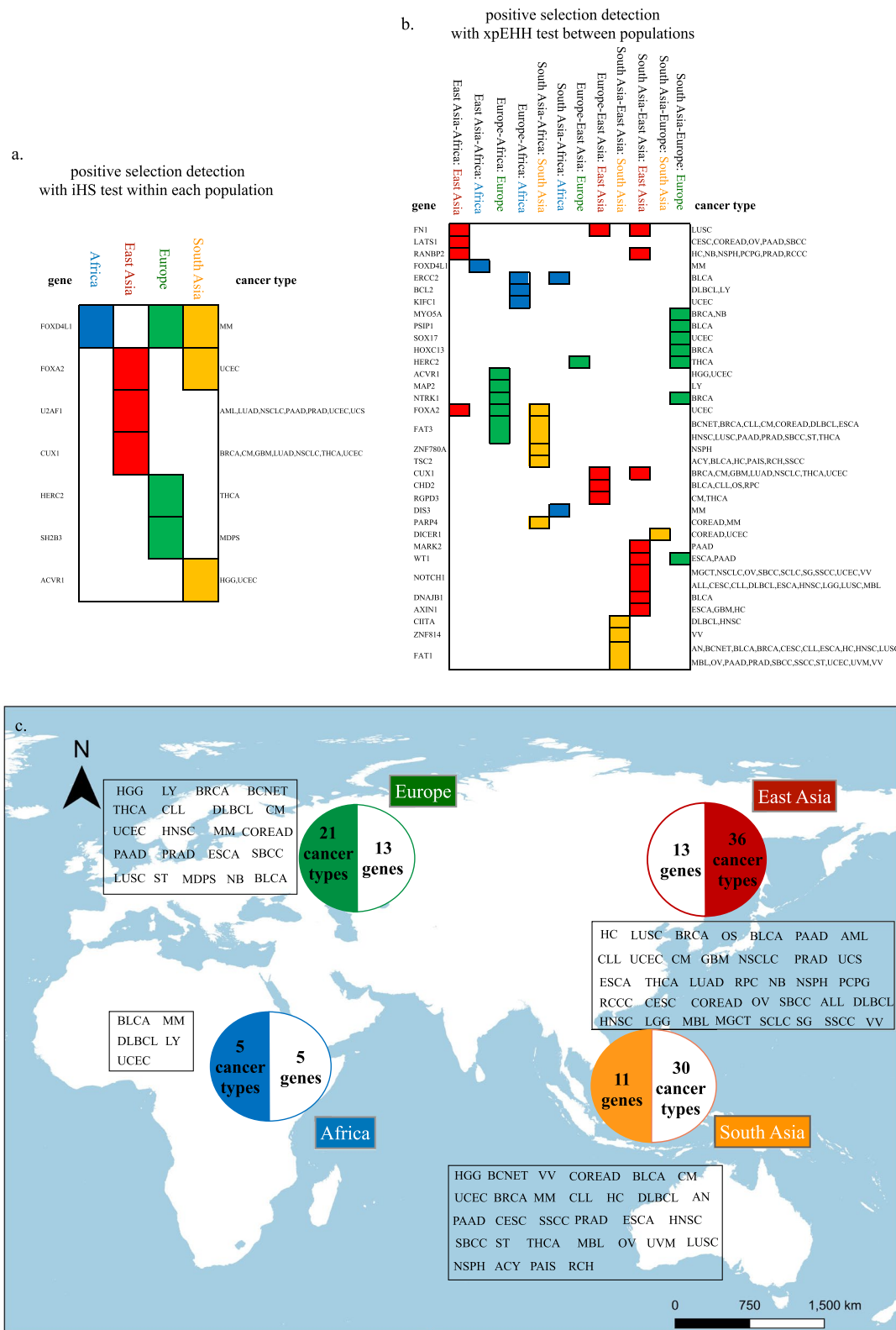
Gene *SH2B3* which functions in anti-bacteria defense also has been reported under positive selection [44]. Using the xpEHH test and the  $F_{ST}$  test, 33 genes across 44 cancer types were identified under positive selection (Fig. 2b). Three genes had been previously reported to be under positive selection, including the gene *HERC2* mentioned above, the psychiatric disorder-relevant gene, *FAT1* [45], and the human skull shape and morphology determination-related gene, *RGPD3* [46]. Most positively selected genes we identified here are related to immune and neurodevelopment (Supplementary File 4). The site frequency spectrum of each locus under positive selection was plotted (Supplementary File 5). Abbreviations of cancer types can be found in the Supplementary File 6.

PBS tests showed that except three SNPs of *FN1*, one SNP of *CHD2*, one SNP of *HERC2*, one SNP of *CIITA*, one SNP of *ZNF814*, one SNP of *SOX17*, one SNP of *PSIP1*, and one SNP of *DICER1* (highlighted in the Supplementary File 7), other positively selected SNPs detected based on the haplotype-based methods (iHS and xpEHH) and the  $F_{ST}$  method were all with significantly high PBS values. And this only affected six genes (*CHD2*, *CIITA*, *ZNF814*, *SOX17*, *PSIP1* and *DICER1*) in the final positively selected gene list. Therefore, signals we identified with the haplotype-based methods (iHS and xpEHH) and the  $F_{ST}$  test were very likely due to positive selection instead of demography. Since the PBS test cannot detect positively selected signals in the African population (because always as the outgroup), we chose positively selected signals identified based on both the haplotype-based tests as well as the  $F_{ST}$  test, but also presented corresponding PBS values (Supplementary File 7).

Cancer driver genes under positive selection exhibited population-specific patterns (Fig. 2, Supplementary File 7). There were 13 genes (*CUX1*, *U2AF1*, *FOXA2*, *RANBP2*, *LATS1*, *FN1*, *CHD2*, *RGPD3*, *DNAJB1*, *AXIN1*, *NOTCH1*, *WT1* and *MARK2*) corresponded to 36 cancer types under positive selection in East Asia; five genes (*FOXD4L1*, *ERCC2*, *BCL2*, *KIFC1* and *DIS3*) corresponded to five cancer types under positive selection in Africa; 13 genes (*HERC2*, *FOXD4L1*, *SH2B3*, *ACVR1*, *MAP2*, *FOXA2*, *NTRK1*, *FAT3*, *HOXC13*, *SOX17*, *WT1*, *PSIP1* and *MYO5A*) corresponded to 21 cancer types under positive selection in Europe; and 11 genes (*FOXA2*, *FOXD4L1*, *ACVR1*, *TSC2*, *PARP4*, *ZNF780A*, *FAT3*, *CIITA*, *ZNF814*, *FAT1* and *DICER1*) corresponded to 30 cancer types under positive selection in South Asia.

### THCA-associated SNPs under positive selection

To see whether sites under positive selection happen to be cancer-associated sites, we screened all positively selected SNPs we identified here in the COSMIC database, the largest and most comprehensive resources

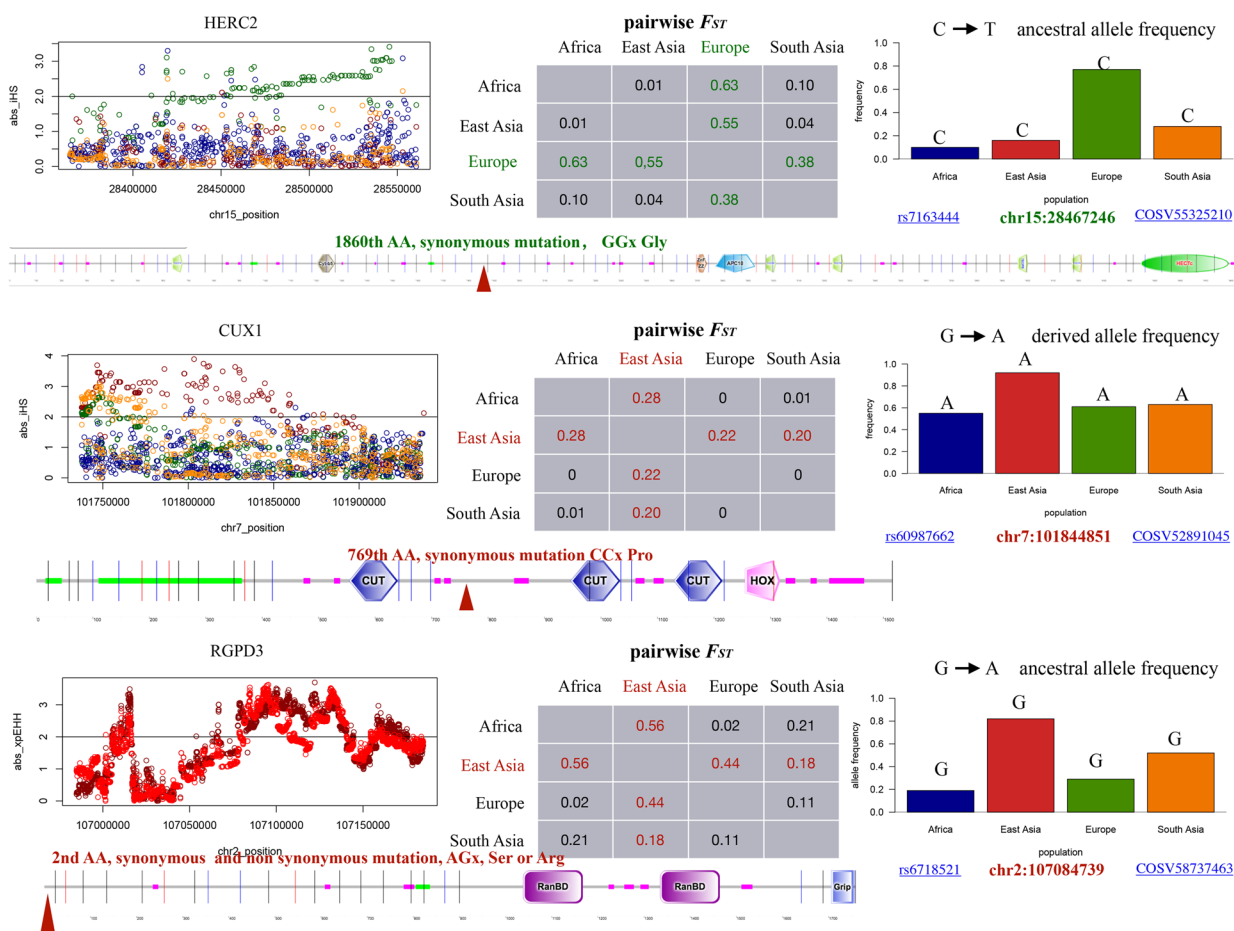


**Fig. 2** Recent positive selection detection in modern human populations. **a** Positive selection detection using the iHS test within each population. Positively selected cancer driver genes and corresponding cancer types were given. **b** Positive selection detection using the xpEHH test between populations. Positively selected cancer driver genes and corresponding cancer types were given. **c** Corresponding cancer types of positively selected cancer driver genes in different populations. Abbreviations of cancer types can be found in the Supplementary File 6



of somatic mutations in human cancer [47]. Interestingly, SNPs associated with thyroid cancer (THCA) in three THCA driver genes were found to be under positive selection in the East Asian population (CUX1 and RGPD3) and the European population (HERC2) (Fig. 3). The locus chr7:101844851 of the CUX1 gene was located within clusters showing an  $|iHS| > 2$  and high pairwise  $F_{ST}$  values between the East Asian and other populations. Allele frequencies were consistent with these results, i.e., there was a higher frequency of derived SNP that was positively selected in the East Asian population. CUX1 belongs to the homeodomain transcription factor family, and is involved in various physiological events, including tissue development and tumorigenesis [48], but has not been previously reported to be under positive selection. Similarly, the locus chr15:28467246 of the HERC2 gene was located within clusters showing an  $|iHS| > 2$  and high pairwise  $F_{ST}$  value between the European and other

populations (Fig. 3). Allele frequencies were consistent with these results, i.e. a higher frequency of the positively selected ancestral SNPs was present in the European population. HERC2 has been reported to be under positive selection and is related to pigmentation [43]. Mutations in the SNPs of these two genes are synonymous and are not located on any known protein domains (Fig. 3). A cross-population test using xpEHH found the the locus chr2:107084739 of the RGPD3 gene, the human skull shape and face morphology determination-related gene [46], was located within clusters with an  $|xpEHH| > 2$  in the East Asian population when compared to both the European and South Asian populations. Mutation of this SNP in RGPD3 can be synonymous or nonsynonymous (Ser or Arg), and is also not located on any known protein domains (Fig. 3). Thus, THCA-associated SNPs under positive selection were identified in populations of East Asia and Europe.



**Fig. 3** Thyroid cancer (THCA) associated SNPs of three THCA driver genes under positive selection in the East Asian population and the European population. High  $iHS$  or  $xpEHH (> 2)$  scores plottings, pairwise  $F_{ST}$  comparisons, allele frequencies, and positions of SNPs on the corresponding secondary protein structure were given. THCA-associated SNP of the driver gene HERC2 showed positive selection signals in the European population, and the THCA-associate SNPs of driver genes CUX1 and RGPD3 showed positive selection signals in the East Asian population. SNP IDs in Ensembl and COSMIC databases were given

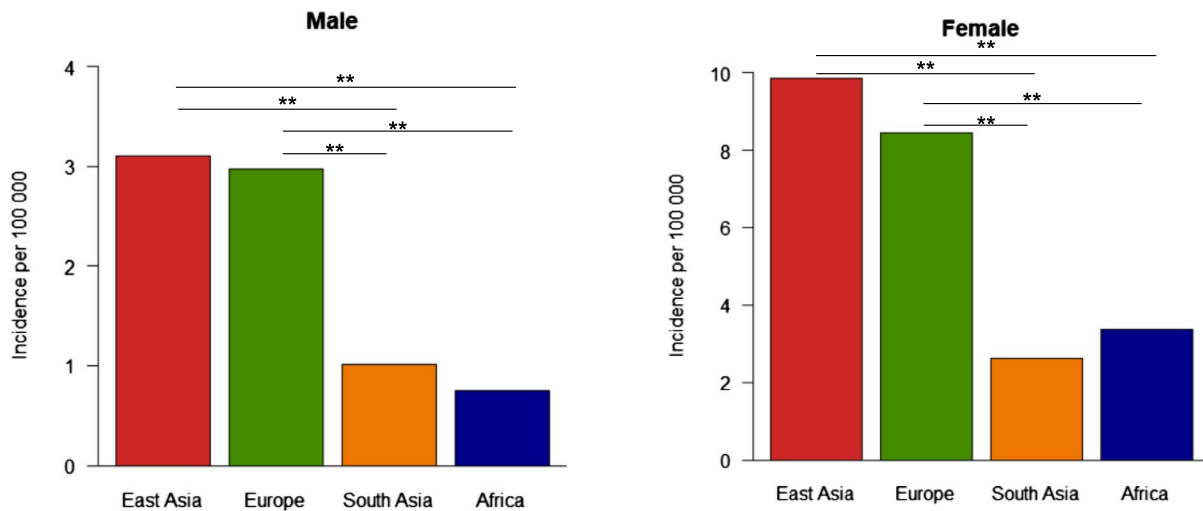
### THCA incidence in different populations

Identification of THCA-associated SNPs under positive selection in East Asian and European populations could lead to higher incidence of THCA in these populations. To determine whether this is the case, global THCA incidence was evaluated. THCA incidence varied among populations (Fig. 4, Supplementary File 8). For males, there were high incidences in the East Asian population (3.11 per 100 000 persons) and in the European population (2.97 per 100 000 persons), and relative low incidences in the South Asian population (1.01 per 100 000 persons) and the African population (0.75 per 100 000 persons). The pattern in females was similar. For females, there were high incidences in the East Asian population (9.85 per 100 000 persons) and the European population (8.44 per 100 000 persons), and relative low incidences in the African population (3.36 per 100 000 persons) and the South Asian population (2.61 per 100 000 persons). The THCA incidences were significantly different between high incidence

group (East Asia and Europe) and low group (Africa and South Asia) in both males and females.

### Discussion

A few studies have already conducted a molecular evolutionary analysis of cancer genes across primates and found many cancer genes were under positive selection in primates [7]. We also found hundreds of genes under positive selection in nonhuman primates (Fig. 1). This indicates that many cancer driver genes play adaptive roles during the early evolution of primates. The genetic background of many cancer genes can be traced back long before the human lineage. For example, some cancer driver genes are essential biological genes controlling basic biology, such as DNA damage repair, cell division [49–51]. As we mentioned in the Introduction section, several hypotheses have already been proposed to explain this phenomenon. For example, viruses can exert a powerful selective pressure and be the key driver of adaptive mutations in proteins [8], which has already been demonstrated in cancer driver genes [7].



population	East Asian	Europe	South Asian	Africa
East Asian	/	1.35	17.73	6.90
Europe	0.088	/	16.78	6.63
South Asian	<0.001	<0.001	/	1.20
Africa	<0.001	<0.001	0.115	/

population	East Asian	Europe	South Asian	Africa
East Asian	/	7.53	34.71	10.71
Europe	<0.001	/	29.80	9.04
South Asian	<0.001	<0.001	/	2.28
Africa	<0.001	<0.001	0.011	/

**Fig. 4** Thyroid cancer (THCA) incidence comparisons among different populations by Z test. The incidence were significantly higher in both the East Asian population and the European population than other populations. Z statistics were given in the upper triangle, while p values were given in the lower triangle

However, when focusing on positive selection in the human lineage itself, only eight cancer driver genes were under positive selection (Table 1, Fig. 1). This may be because many cancer driver genes served adaptive roles in the early evolution of primates, but did not work in the human lineage. Massive changes occurred during the evolution of human beings, such as upright walking [52–54] and brain development [55], which can induce different selection pressures in the evolution of the human lineage. Indeed, most genes under positive selection in the human lineage we identified here are related to neurocognitive, brain development and immune responses. In addition, nonhuman primates may possess alternative cancer genetic bases or anticancer mechanisms, as these are not rare in mammals and can be seen in mole rats, elephants, and whales [56].

In modern human populations, there were 35 cancer driver genes covering 47 cancer types under recent positive selection. Genetic disease may occur as a by-product of evolution due to genetic pleiotropic effects. For instance, the seemingly human-specific disease of schizophrenia and the greater human susceptibility to Alzheimer's disease may be a by-product of human specialization for higher cognitive function [36, 57]. Significant changes in modern human populations but not in long timescale evolution are also found in the evolution of testis-related genes [10]. Similar alterations may be observed in human cancers. Identification of cancer driver genes under recent positive selection supports adaptive roles for these genes and may help survival of these cancer driver genes and accumulation of cancer-related mutations during evolution.

We further found THCA-associated SNPs in three THCA driver genes, CUX1, HERC2 and RGD3, were determined to be under positive selection in East Asian and European populations. SNP mutations of CUX1 and HERC2 are synonymous mutations. Studies have already shown that synonymous SNPs under positive selection are pervasive in mammals; for example, they affect mRNA levels [58] and translation via mRNA destabilization [59] or create internal promoter sites [60]. Synonymous mutations frequently function as driver mutations in human cancers [61] and can be under positive selection [62]. Functions of THCA-associated SNPs under positive selection we identified here requires further study. We also found a higher incidence of THCA in East Asian and European populations compared to other populations. Both environmental conditions [63] and genetic components [64] are important factors to affect cancer occurrence. Although we cannot exclude other effects on the high incidence rates in East Asian and European populations, the THCA-associated SNPs under positive

selection we identified here at least provides one genetic clue to explain the high incidence of THCA in corresponding populations.

## Conclusions

Cancer can be evolved, in part, as a by-product of adaptive changes in humans. Different SNPs at the same locus can be under different selection pressures in different populations, and thus should be under consideration during precision medicine, especially targeted medicine in specific populations.

## Methods

### Orthologous sequences inference

The transcript IDs of 568 cancer driver genes were retrieved from the literature published in *Nature Reviews Cancer* [24] (Supplementary File 9). To understand the evolutionary dynamics of these 568 cancer driver genes across the primate phylogeny, orthologous sequences of 23 primates were retrieved from Ensembl Biomart (<https://www.ensembl.org/index.html>). Oma was used to perform further orthologous sequence inference with default parameters [25]. The species tree retrieved from the literature [37] was used for orthologous sequences inference using oma, with *Mus musculus* as the outgroup.

### Positive selection detection across the phylogeny of primates

Positive selection detection was based on the rates of protein evolution calculated using the branch-site model with codeml in PAML [26, 27]. The ratio of nonsynonymous (dN) to synonymous (dS) substitutions ( $\omega$ ) can vary over sites and time. We thus used the branch-site model which considers the variation in  $\omega$  both among sites and across branches to detect positive selection affecting sites along the target lineage (foreground branch). If a nonsynonymous mutation is less likely to be tolerated during evolution,  $\omega$  will be  $< 1$ , meaning that purifying selection takes effects. If there is no selection pressure (neutral selection),  $\omega$  will be equal to 1. If changes of nonsynonymous mutations have beneficial effects and are favored by selection,  $\omega$  will be  $> 1$ , indicating that they are under positive selection. Multiple sequence alignment for codon alignment was conducted using MAFFT within T-coffee [65] using default parameters. An unrooted primate phylogenetic tree retrieved from the literature [37] was used for the calculation (Fig. 1).

Nine branches representing the main primate lineages were set as foreground branches to run the branch-site model separately (Fig. 1). Parameter settings and test for statistical significance were the same as those previously published [66]. Briefly,



initial branch lengths were estimated using the M0 model, and fixed branch lengths ( $\text{fix\_blength}=2$ ) were used for downstream analysis. We removed alignment gaps and ambiguity characters by setting  $\text{Cleandata}=1$ . Comparisons were made between the modified ModelA ( $\text{model}=2$ ,  $\text{NSsites}=2$ ) and the ModelB (a null model with  $\omega=1$  fixed ( $\text{fix\_omega}=1$  and  $\text{omega}=1$ )). A likelihood ratio test (LRT) was then used to calculate a *chi*-square approximation, and the  $p/2$  value was used to consider mixture distribution [26, 27]. Bonferroni correction was used for multiple test corrections [67]. Sites under positive selection were identified under BEB (Bayes empirical Bayes) with a posterior probability  $> 0.95$  [26, 27].

### Recent positive selection detection in modern human populations

Phased genetic data of the 1000 Genomes Project of modern human populations were retrieved from [ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502). The populations used were the same as those in the recently published literature [10], which excluded recently admixed populations or populations in close geographic proximity. Details of these populations can be found in Supplementary File 10. 1) Populations of African ancestry ( $n=311$ ) included Gambians in Western Division, The Gambia (GWD,  $n=113$ ), Luhya in Webuye, Kenya (LWK,  $n=99$ ) and Esan in Nigeria (ESN,  $n=99$ ); 2) Populations of East Asian ancestry ( $n=306$ ) included Han Chinese in Beijing, China (CHB,  $n=103$ ), Japanese in Tokyo, Japan (JPT,  $n=104$ ), and Kinh in Ho Chi Minh City, Vietnam (KHV,  $n=99$ ); 3) Populations of European ancestry ( $n=297$ ) included British people in England and Scotland, United Kingdom (GBR,  $n=91$ ), Finnish people in Finland (FIN,  $n=99$ ), and Toscani people in Italy (TSL,  $n=107$ ); and 4) Populations of South Asian ancestry ( $n=284$ ) included Bengali people in Bangladesh (BEB,  $n=86$ ), Indian Telugu people in the United Kingdom (ITU,  $n=102$ ), and Punjabi people in Lahore, Pakistan (PJL,  $n=96$ ).

PLINK and VCFtools were used to process variant call format (VCF) files for autochromosomes [68]. SNPs with indels were removed. The GRCh37/hg19 reference genome from Ensembl was used to retrieve genomic location and identify SNPs in coding sequences of cancer driver genes here. The iHS and xpEHH methods, implemented in selscan were used to detect genome-wide positive selection using default parameters [69]. Biallelic SNPs with minor allele frequency (MAF)  $\geq 0.05$  were considered for positive selection detection. Ancestral alleles have already been annotated in the vcf files [70]. Unstandardized iHS

and xpEHH scores were normalized in frequency bins across the entire genome using the script *norm* in the selscan program. Pairwise  $F_{ST}$  was calculated using the Weir & Cockerham  $F_{ST}$  calculation in VCFtools. Allele frequencies in each population were calculated using VCFtools with parameters  $-\text{keep}$  and  $-\text{frequency}$ . The VCF files were indexed and target regions were extracted with tabix [71]. Then the site frequency spectrums of each locus under positive selection were plotted for different populations.

Details about PBS tests were described in the literature [41]. Briefly, the  $F_{ST}$  for each SNP was first log-transformed [72]:

$$T = -\log(1 - F_{ST})$$

Then the PBS of test population since the divergent from the sister population was obtained as:

$$PBS = \left( T^{\text{test-sister}} + T^{\text{test-Africa}} - T^{\text{sister-Africa}} \right) / 2$$

We thus calculated PBS values across the genome as follows, with the African population as the outgroup.

1) the European population diverged from the East Asian population;

$$PBS^1 = T^{\text{Europe-EastAsia}} + T^{\text{Europe-Africa}} - T^{\text{EastAsia-Africa}} / 2$$

2) the European population diverged from the South Asian population;

$$PBS^2 = T^{\text{Europe-SouthAsia}} + T^{\text{Europe-Africa}} - T^{\text{SouthAsia-Africa}} / 2$$

3) the East Asian population diverged from the South Asian population;

$$PBS^3 = T^{\text{EastAsia-SouthAsia}} + T^{\text{EastAsia-Africa}} - T^{\text{SouthAsia-Africa}} / 2$$

4) the East Asian population diverged from the European population;

$$PBS^4 = T^{\text{EastAsia-Europe}} + T^{\text{EastAsia-Africa}} - T^{\text{Europe-Africa}} / 2$$

5) the South Asian population diverged from the European population;

$$PBS^5 = T^{\text{SouthAsia-Europe}} + T^{\text{SouthAsia-Africa}} - T^{\text{Europe-Africa}} / 2$$

6) the South Asian population diverged from the East Asian population.

$$PBS^6 = T^{\text{SouthAsia-EastAsia}} + T^{\text{SouthAsia-Africa}} - T^{\text{EastAsia-Africa}} / 2$$

We then plotted the distribution of PBS values across the genome in each comparison and top 5% PBS values were considered as statistically significantly high. Since we have to use the African population as the outgroup,

we thus cannot calculate PBS values for African population-related comparisons.

### Protein domain prediction

Protein domains were predicted using the Simple Modular Architecture Research Tool (SMART) with default parameters [73].

### Cancer incidence comparisons

THCA data were retrieved from the Cancer Incidence in Five Continents Vol. XI [74]. To accompany the population genetic analysis above, the same populations or nearby populations were used in cancer incidence rate analysis. The age-standardized rates (ASR) were used for comparisons. In Africa, the THCA incidence ranges from 0 to 1 per 100 000 males, and 1.9 to 4.4 per 100 000 females from 3 registries involving 5 465 978 persons; in East Asia, the THCA incidence ranges from 0.3 to 7.6 per 100 000 males, and 0.5 to 22 per 100 000 females from 45 registries involving 107 646 738 persons; in Europe, the THCA incidence ranges from 1.5 to 10.6 per 100 000 males, and 3.3 to 38.5 per 100 000 females from 41 registries involving 101 816 542 persons; in South Asia, the THCA incidence ranges from 0.1 to 3.8 per 100 000 males, and 0.8 to 9.7 per 100 000 females from 16 registries involving 50 983 915 persons; Z test was used to compare the THCA incidence among different populations (Africa, East Asia, South Asian, Europe) in males and females, respectively. The statistical Z is calculated by  $\frac{\left| \frac{c_1}{n_1} - \frac{c_2}{n_2} \right|}{\sqrt{\left( \frac{c_1+c_2}{n_1+n_2} \right) \left( 1 - \frac{c_1+c_2}{n_1+n_2} \right) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$ , where  $c_1$  and  $c_2$

represent the number of THCA patients in population 1 and population 2, respectively;  $n_1$  and  $n_2$  represent the population size in population 1 and population 2, respectively.

### Abbreviations

dN	Nonsynonymous
dS	Synonymous
BEB	Bayes empirical Bayes
iHS	The integrated haplotype score
xpEHH	The cross-population extended haplotype homozygosity
PBS	Population branch statistics
THCA	Thyroid cancer
LRT	Likelihood ratio test
GWD	Gambians in Western Division, The Gambia
LWK	Luhya in Webuye, Kenya
ESN	Esan in Nigeria
CHB	Han Chinese in Beijing, China
JPT	Japanese in Tokyo, Japan
KHV	Kinh in Ho Chi Minh City, Vietnam
GBR	British people in England and Scotland, United Kingdom
FIN	Finnish people in Finland
TSI	Toscani people in Italy
BEB	Bengali people in Bangladesh
ITU	Indian Telugu people in the United Kingdom
PJL	Punjabi people in Lahore, Pakistan
VCF	Variant call format

MAF	Minor allele frequency
SMART	The Simple Modular Architecture Research Tool
ASR	Age-standardized rates

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-023-09301-9>.

- Additional file 1.
- Additional file 2.
- Additional file 3.
- Additional file 4.
- Additional file 5.
- Additional file 6.
- Additional file 7.
- Additional file 8.
- Additional file 9.
- Additional file 10.

### Acknowledgements

We thank Prof. Haosen Li, Prof. Muhua Wang, Dr. Cheng Huang and Chenqing Zheng for the valuable suggestions and supports. We also appreciate the valuable suggestions and comments from reviewers. Data computation was supported by National Supercomputer Center in Guangzhou, China. The English language was edited by Elsevier Language Editing Services.

### Authors' contributions

LG and GY designed the study. LG did the orthologous sequences inference and positive selection detection at the species- and population- levels. CX and SY did statistical analyses of global cancer incidence rates comparison. CX also provides valuable code support. LG wrote the manuscript and all authors commented on it. All authors read and approved the final manuscript.

### Funding

This work was supported by the National Natural Science Foundation of China (32171465) to GY and Fundamental Research Funds for the Central Universities (20lgpy109) to LG. Data computation was supported by National Supercomputer Center in Guangzhou, China.

### Availability of data and materials

All data generated or analysed during this study are included in this published article and its supplementary information files. Five hundred sixty-eight cancer driver genes we analyzed were retrieved from [24] (Supplementary File 9). Homologous sequences were retrieved from Ensembl database <https://asia.ensembl.org/index.html>. Phased genetic data of the 1000 Genomes Project of modern human populations were retrieved from <ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>.

### Declarations

**Ethics approval and consent to participate**  
Not applicable.

**Consent for publication**  
Not applicable.

**Competing interests**  
The authors declare no competing interests.

Received: 11 August 2022 Accepted: 8 April 2023  
Published online: 25 April 2023

## References

- Fischer A, Greenman C, Mustonen V. Germline fitness-based scoring of cancer mutations. *Genetics*. 2011;188:383–93.
- Shendure J, Akey JM. The origins, determinants, and consequences of human mutations. *Science*. 2015;349:1478–83.
- Vicens A, Posada D. Selective Pressures on Human Cancer Genes along the Evolution of Mammals. *Genes*. 2018;9(12):582.
- Kang L, Michalak P. The evolution of cancer-related genes in hominoids. *J Mol Evol*. 2015;80:37–41.
- Crespi BJ, Summers K. Positive selection in the evolution of cancer. *Biol Rev Camb Philos Soc*. 2006;81:407–24.
- Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, et al. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol*. 2005;3:e170.
- Lou DI, McBee RM, Le UQ, Stone AC, Wilkerson GK, Demogines AM, et al. Rapid evolution of BRCA1 and BRCA2 in humans and other primates. *BMC Evol Biol*. 2014;14:155.
- Enard D, Cai L, Gwennap C, Petrov DA. Viruses are a dominant driver of protein adaptation in mammals. *Elife*. 2016;5:e12469.
- Huttley GA, Easteal S, Southey MC, Tesoriero A, Giles GG, McCredie MR, et al. Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Australian Breast Cancer Family Study. *Nat Genet*. 2000;25:410–3.
- Schaschl H, Wallner B. Population-specific, recent positive directional selection suggests adaptation of human male reproductive genes to different environmental conditions. *BMC Evol Biol*. 2020;20:27.
- Scerri EML, Chikhi L, Thomas MG. Beyond multiregional and simple out-of-Africa models of human evolution. *Nat Ecol Evol*. 2019;3:1370–2.
- Stewart JR, Stringer CB. Human evolution out of Africa: the role of refugia and climate change. *Science*. 2012;335:1317–21.
- Benton ML, Abraham A, LaBella AL, Abbot P, Rokas A, Capra JA. The influence of evolutionary history on human health and disease. *Nat Rev Genet*. 2021;22:269–83.
- Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. *Nature*. 2007;449:913–8.
- Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*. 2002;419:832–7.
- Tishkoff SA, Varkonyi R, Cahinhinan N, Abbas S, Argyropoulos G, Destro-Bisol G, et al. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science*. 2001;293:455–62.
- Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, Batubara L, et al. A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum Mol Genet*. 2008;17:835–43.
- Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, et al. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet*. 2004;74:1111–20.
- Peng J, Redman CM, Wu X, Song X, Walker RH, Westhoff CM, et al. Insights into extensive deletions around the XK locus associated with McLeod phenotype and characterization of two novel cases. *Gene*. 2007;392:142–50.
- Lappalainen T, Salmela E, Andersen PM, Dahlman-Wright K, Sistonen P, Savontaus M-L, et al. Genomic landscape of positive natural selection in Northern European populations. *Eur J Hum Genet*. 2010;18:471–8.
- Ermini L, Francis JC, Rosa GS, Rose AJ, Ning J, Greaves M, et al. Evolutionary selection of alleles in the melanophilin gene that impacts on prostate organ function and cancer risk. *Evol Med Public Heal*. 2021;9:311–21.
- Wildman DE, Uddin M, Liu G, Grossman LI, Goodman M. Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: enlarging genus *Homo*. *Proc Natl Acad Sci U S A*. 2003;100:7181–8.
- Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*. 2017;171:1029–41.e21.
- Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. *Nat Rev Cancer*. 2020;20:555–72.
- Altenhoff AM, Levy J, Zarowiecki M, Tomiczek B, Warwick Vesztrocy A, Dalquen DA, et al. OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Res*. 2019;29:1152–63.
- Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007;24:1586–91.
- Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*. 1997;13:555–6.
- Xu H, Liu X, Li W, Xi Y, Su P, Meng B, et al. p38 MAPK-mediated loss of nuclear RNase III enzyme Drosha underlies amyloid beta-induced neuronal stress in Alzheimer's disease. *Aging Cell*. 2021;20:e13434.
- Butler M, Morel A-S, Jordan WJ, Eren E, Hue S, Shrimpton RE, et al. Altered expression and endocytic function of CD205 in human dendritic cells, and detection of a CD205/DCL-1 fusion protein upon dendritic cell maturation. *Immunology*. 2007;120:362–71.
- Raghavan NS, Dumitrescu L, Mormino E, Mahoney ER, Lee AJ, Gao Y, et al. Association Between Common Variants in RBFox1, an RNA-Binding Protein, and Brain Amyloidosis in Early and Preclinical Alzheimer Disease. *JAMA Neurol*. 2020;77:1288–98.
- Yang J, Kang C, Wu C, Lin Y, Zeng L, Yuan J, et al. Pharmacogenetic associations of NRG1 polymorphisms with neurocognitive performance and clinical symptom response to risperidone in the untreated schizophrenia. *Schizophr Res*. 2021;231:67–9.
- Wu J, Jin Y-Y, Gong R-L, Yang F, Su X-Y, Chen T-X. Genome-Wide ChIP-seq and RNA-seq Analyses of STAT3 Target Genes in TLRs Activated Human Peripheral Blood B Cells. *Front Immunol*. 2022;13:821457.
- Montgomery S, Mundy N. Positive selection on NIN, a gene involved in neurogenesis, and primate brain evolution. *Genes Brain Behav*. 2012;11(8):903–10.
- Liu Y, Gokhale S, Jung J, Zhu S, Luo C, Saha D, et al. Mitochondrial Fis3on Factor Is a Novel Interacting Protein of the Critical B Cell Survival Regulator TRAF3 in B Lymphocytes. *Front Immunol*. 2021;12:670338.
- Ayub Q, Yngvadottir B, Chen Y, Xue Y, Hu M, Vernes SC, et al. FOXP2 targets show evidence of positive selection in European populations. *Am J Hum Genet*. 2013;92:696–706.
- Crespi B, Summers K, Dorus S. Adaptive evolution of genes underlying schizophrenia. *Proceedings Biol Sci*. 2007;274:2801–10.
- Upham NS, Esselstyn JA, Jetz W. Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLoS Biol*. 2019;17:e3000494.
- Voight BF, Kudaravalli S, Wen X, Pritchard JK. A Map of Recent Positive Selection in the Human Genome. *Hurst L, editor. PLoS Biol* 2006;4:e72.
- Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. Recent and ongoing selection in the human genome. *Nat Rev Genet*. 2007;8:857–68.
- Innan H, Kim Y. Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. *Genetics*. 2008;179:1713–20.
- Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. *Science*. 2010;329:75–8.
- Cheng X, Xu C, DeGiorgio M. Fast and robust detection of ancestral selective sweeps. *Mol Ecol*. 2017;26:6871–91.
- Wilde S, Timpson A, Kirsanow K, Kaiser E, Kayser M, Unterländer M, et al. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proc Natl Acad Sci U S A*. 2014;111:4832–7.
- Zhernakova A, Elbers CC, Ferwerda B, Romanos J, Trynka G, Dubois PC, et al. Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. *Am J Hum Genet*. 2010;86:970–7.
- Sato DX, Kawata M. Positive and balancing selection on SLC18A1 gene associated with psychiatric disorders and human-unique personality traits. *Evol Lett*. 2018;2:499–510.
- Wu W, Zhai G, Xu Z, Hou B, Liu D, Liu T, et al. Whole-exome sequencing identified four loci influencing craniofacial morphology in northern Han Chinese. *Hum Genet*. 2019;138:601–11.
- Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*. 2019;47:D941–7.
- Liu N, Sun Q, Wan L, Wang X, Feng Y, Luo J, et al. CUX1, A Controversial Player in Tumor Development. *Front Oncol*. 2020;10:738.
- Wu Z, Li S, Tang X, Wang Y, Guo W, Cao G, et al. Copy Number Amplification of DNA Damage Repair Pathways Potentiates Therapeutic Resistance in Cancer. *Theranostics*. 2020;10:3939–51.

50. O'Connell MJ. Selection and the cell cycle: positive Darwinian selection in a well-known DNA damage response pathway. *J Mol Evol.* 2010;71:444–57.
51. Zahir N, Sun R, Gallahan D, Gatenby RA, Curtis C. Characterizing the ecological and evolutionary dynamics of cancer. *Nat Genet.* 2020;52:759–67.
52. Niemitz C. The evolution of the upright posture and gait—a review and a new synthesis. *Naturwissenschaften.* 2010;97:241–63.
53. Zirkle D, Lovejoy CO. The hominid ilium is shaped by a synapomorphic growth mechanism that is unique within primates. *Proc Natl Acad Sci U S A.* 2019;116:13915–20.
54. Gruss LT, Schmitt D. The evolution of the human pelvis: changing adaptations to bipedalism, obstetrics and thermoregulation. *Philos Trans R Soc Lond B Biol Sci.* 2015;370:20140063.
55. Kanton S, Boyle MJ, He Z, Santel M, Weigert A, Sanchís-Calleja F, et al. Organoid single-cell genomic atlas uncovers human-specific features of brain development. *Nature.* 2019;574:418–22.
56. Seluanov A, Gladyshev VN, Vijg J, Gorbunova V. Mechanisms of cancer resistance in long-lived mammals. *Nat Rev Cancer.* 2018;18:433–41.
57. Vamathevan JJ, Hasan S, Emes RD, Amrine-Madsen H, Rajagopalan D, Topp SD, et al. The role of positive selection in determining the molecular cause of species differences in disease. *BMC Evol Biol.* 2008;8:273.
58. Shen X, Song S, Li C, Zhang J. Synonymous mutations in representative yeast genes are mostly strongly non-neutral. *Nature.* 2022;606:725–31.
59. Resch AM, Carmel L, Mariño-Ramírez L, Ogurtsov AY, Shabalina SA, Rogozin IB, et al. Widespread positive selection in synonymous sites of mammalian genes. *Mol Biol Evol.* 2007;24:1821–31.
60. Lebeuf-Taylor E, McCloskey N, Bailey SF, Hinz A, Kassen R. The distribution of fitness effects among synonymous mutations in a gene under directional selection. *Elife.* 2019;8:e45952.
61. Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell.* 2014;156:1324–35.
62. Bin Y, Wang X, Zhao L, Wen P, Xia J. An analysis of mutational signatures of synonymous mutations across 15 cancer types. *BMC Med Genet.* 2019;20:190.
63. Bahnassy AA, Abdellateif MS, Zekri A-RN. Cancer in Africa: Is It a Genetic or Environmental Health Problem? *Front Oncol.* 2020;10:2508.
64. Worch J, Matthay KK, Neuhaus J, Goldsby R, DuBois SG. Ethnic and racial differences in patients with Ewing sarcoma. *Cancer.* 2010;116:983–8.
65. Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol.* 2000;302:205–17.
66. Gu L, Xia C. Cluster expansion of apolipoprotein D (ApoD) genes in teleost fishes. *BMC Evol Biol.* 2019;19:9.
67. Anisimova M, Yang Z. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Mol Biol Evol.* 2007;24:1219–28.
68. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. *Bioinformatics.* 2011;27:2156–8.
69. Szpiech ZA, Hernandez RD. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol.* 2014;31:2824–7.
70. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. *Nature.* 2015;526:68–74.
71. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics.* 2011;27:718–9.
72. Cavalli-Sforza LL. Human Diversity Proc 12th Int Congr. Genet. 1969;2:405–16.
73. Letunic I, Doerks T, Bork P. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.* 2015;43:D257–60.
74. Bray F, Colombet M, Mery L, Piñeros M, Znaor A, Zanetti R, et al. Cancer Incidence in Five Continents, Vol. XI (electronic version). Lyon: International Agency for Research on Cancer; 2017.

## Publisher's Note

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

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

