Evolutionarily new genes in humans with disease phenotypes reveal functional enrichment patterns shaped by adaptive innovation and sexual selection.

Jianhai Chen[1,*], Patrick Landback[1], Deanna Arsala[1], Alexander Guzzetta[3], Shengqian Xia[1], Jared Atlas[1], Chuan Dong[4], Dylan Sosa[1],Yong E. Zhang[5], Jingqiu Cheng[2], Bairong Shen[2,*], Manyuan Long[1,*]

[1]Department of Ecology and Evolution, The University of Chicago, 1101 E 57th Street, Chicago, IL 60637

[2]Institutes for Systems Genetics, West China University Hospital, Chengdu 610041, China

[3]Department of Pathology, The University of Chicago, 1101 E 57th Street, Chicago, IL 60637

[4]School of Forestry and Biotechnology, Zhejiang A&F University, Hangzhou 311300, China

[5]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China

*Correspondence:

jianhaichen@uchicago.edu;

mlong@uchicago.edu;

bairong.shen@scu.edu.cn;

Abstract

New genes (or young genes) are structural novelties pivotal in mammalian evolution. Their phenotypic impact on humans, however, remains elusive due to the technical and ethical complexities in functional studies. Through combining gene age dating with Mendelian disease phenotyping, our research reveals that new genes associated with disease phenotypes steadily integrate into the human genome at a rate of ~0.07% every million years over macroevolutionary timescales. Despite this stable pace, we observe distinct patterns in phenotypic enrichment, pleiotropy, and selective pressures between young and old genes. Notably, young genes show significant enrichment in the male reproductive system, indicating strong sexual selection. Young genes also exhibit functions in tissues and systems potentially linked to human phenotypic innovations, such as increased brain size, bipedal locomotion, and color vision. Our findings further reveal increasing levels of pleiotropy over evolutionary time, which accompanies stronger selective constraints. We propose a “pleiotropy-barrier” model that delineates different potentials for phenotypic innovation between young and older genes subject to natural selection. Our study demonstrates that evolutionary new genes are critical in influencing human reproductive evolution and adaptive phenotypic innovations driven by sexual and natural selection, with low pleiotropy as a selective advantage.

Keywords: New genes, Pleiotropy, Young genes, Phenotypic innovation, Sexual selection, Natural selection

Introduction

The imperfection of DNA replication serves as a rich source of variation for evolution and biodiversity [1-3]. Such genetic variations underpin the ongoing evolution of human phenotypes, with beneficial mutations being conserved under positive selection, and detrimental ones being eliminated through purifying selection. In medical terminology, this spectrum is categorized as "case and control" or "disease and health," representing two ends of the phenotypic continuum [4]. Approximately 8,000 clinical types of rare Mendelian disorders, affecting millions worldwide, are attributed to deleterious DNA mutations in single genes (monogenic) or a small number of genes (oligogenic) with significant effects [5, 6]. To date, over 4,000 Mendelian disease genes have been identified, each contributing to a diverse array of human phenotypes (https://mirror.omim.org/statistics/geneMap) [7]. These identified genes and associated phenotypes could provide critical insights into the evolutionary trajectory of human traits [8].

Evolutionarily new genes – such as *de novo* genes and gene duplicates – have been continually emerging and integrating into the human genome throughout the macroevolutionary process of human lineage [9-15]. Previous reports revealed that human disease genes tend to be primarily ancient, with many tracing back to the last common ancestor of eukaryotes [16]. This conclusion aligns with the deep conservation of many critical biological processes shared among cells, such as DNA replication, RNA transcription, and protein translation, which emerged early in the tree of life. Consequently, it may be inferred that new genes play less or no important role in biomedical processes. However, decades of genetic studies in non-human systems have provided extensive evidence contradicting this intuitive argument. New genes can integrate into biologically critical processes, such as transcription regulation, RNA synthesis, and DNA repair [17, 18]. For instance, in yeast, some *de novo* genes (*BSC4* and *MDF1*) play roles in DNA repair process [19-21]. In *Drosophila* species, lineage-specific genes can control the key cytological process of mitosis [22]. New genes (*Nicknack* and *Oddjob*) have also been found with roles in early larval development of *Drosophila* [23]. In *Pristionchus* Nematodes, some lineage-specific genes could serve as the developmental switch determining mouth morphology [24]. Moreover, in multiple insect lineages, embryonic development of body plans, which was long believed to be governed by deeply conserved genetic mechanisms, was found to be driven by newly arising genes [25]. These studies from model species reveal various important biological functions of new genes.

Compared to non-human model organisms, where gene functions can be characterized through genetic knockdowns and knockouts, interrogating functions of human genes in their native context is unfeasible. Despite this limitation, numerous omics data and in vitro studies in human genes have suggested the potential roles of evolutionary young genes in basic cellular processes and complex phenotypic innovations [26-28]. Brain transcriptomic analysis has revealed that primate-specific genes are enriched among up-regulated genes early in human development, particularly within the human-specific prefrontal cortex [29]. The recruitment of new genes into the transcriptome suggests that human anatomical novelties may evolve through the contribution of new gene evolution. Recent studies based on organoid modeling also support the importance of de novo genes on human brain size enlargement [30, 31]. These lines of evidence in recent decades about the functions of new genes contradict the conventional conservation-dominant understanding of human genetics and phenotypes.

In this study, we tackled the complexities of human phenotypic evolution and the underlying genetic basis by integrating gene age dating with analyses of Mendelian disease phenotypes. As a direct indicator of functional effects, the anatomical organ/tissue/system phenotypes (OP) affected by causal genic defects can allow us to understand the influence of gene ages on phenotypic enrichment, pleiotropy, and selective constraints along evolutionary journey. We aimed to understand include whether, what, and why human anatomical/physiological/cellular phenotypes could be affected by human evolutionary new genes. Notably, disease gene emergence rates per million years were found to be similar among different macroevolutionary stages, suggesting the continuous integration of young genes into biomedically important phenotypes. Despite the consistent pace of gene integration per million years, younger disease genes, with lower pleiotropy score, display accelerated sexual selection and human-specific adaptive innovations. By contrast, older genes are higher in pleiotropic burden that impacts more anatomical systems and are thus under stronger selective constraints. These patterns suggest that new genes can rapidly become the genetic bases of human critical phenotypes, especially the reproductive and innovative traits, a process likely facilitated by their low pleiotropy.

Results

Ages and organ/tissue phenotypes of human genetic disease genes

We determined the ages of 19,665 non-redundant genes, following the phylogenetic framework of the GenTree database [32] and gene model annotations from Ensembl v110 (Supplementary table 1). To ensure comparable gene numbers across different age groups, we merged evolutionary age groups with a small number of genes (<100) into their adjacent older group. As a result, we classified these genes into seven ancestral age groups, ranging from Euteleostomi (or more ancient) nodes to modern humans (br0-br6, Figure 1a). These evolutionary groups have been further categorized into four evolutionary age epochs, starting from the oldest, Euteleostomi, to progressively younger stages of Tetrapoda, Amniota, and Eutheria, each containing over 2000 genes. Disease gene data were sourced from Human Phenotype Ontology database (HPO, Sep 2023), which is the *de facto* standard for phenotyping of rare Mendelian diseases [33]. This repository synthesizes information from diverse databases, including Orphanet [34, 35], DECIPHER [36], and OMIM [37]. An intersection of these data sets yielded 4,946 genes annotated with both evolutionary age and organ/tissue/system-specific phenotypic abnormalities (Figure 1a and Supplementary Table 2). Contrasting earlier estimates which suggest that only 0.6% of young genes arising in Eutherian lineage could contribute to human disease genes, we observed nearly 10 times higher percentage of disease genes in this age group (6.67%, Figure 1a and Supplementary Table 2). This indicates that the role of younger genes as disease genes might have been significantly underestimated.

To better ascertain if disease genes evolve under different evolutionary pressures compared to non-disease genes, we compared the metric of Ka/Ks ratio, which is the ratio of the number of nonsynonymous substitutions per nonsynonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks). We retrieved the “one to one” human-chimpanzee orthologous genes and the corresponding pairwise Ka/Ks ratios (12830 genes) from Ensembl database. We also evaluated whether the pattern is consistent with Ka/Ks ratios of human-bonobo and human-macaque orthologs. To include more orthologous genes, we did not use Ka/Ks ratios based on more distant species (such as the test of branch-model). Interestingly, Ka/Ks ratios were consistently lower in disease genes than in non-disease genes for human-chimpanzee orthologs (0.250 vs. 0.321), human-bonobo orthologs (0.273 vs. 0.340), and human-macaque orthologs (0.161 vs. 0.213) (Wilcoxon rank sum test, $p < 2.2e-16$ for all three datasets). These results revealed that disease genes are under

significantly stronger purifying selection than non-disease genes, suggesting the important component of selective pressure in constraining the sequence evolution of disease genes. In addition, we observed an increase in Ka/Ks ratios (<1) for genes from older to younger stages, suggesting a trend of relaxed purifying selection in young genes (Figure 1b and Supplementary figure 1). Notably, despite the relaxation of purifying selection for younger genes, disease genes still tend to show lower Ka/Ks ratio than non-disease genes, suggesting a general pattern of stronger purifying selection in disease genes during evolutionary process.

We observed a heterogeneous distribution of disease genes underlying 22 HPO-defined anatomical systems, suggesting varied genetic complexity for diseases of different systems (Figure 1c-1d). None of disease genes was found to impact all 22 systems. In contrast, 6.96% of disease genes (344/4946) were specific to a single system's abnormality. Notably, four systems – the genitourinary system (with 81 genes), the eyes (68 genes), the ears (63 genes), and the nervous system (55 genes) – collectively represented 77.62% of these system-specific genes (267/344, Supplementary table 2). The nervous system displayed the highest fraction of diseases genes (79%, Figure 1d). A significant 93.04% of genes were linked to the abnormalities of at least two systems (4602/4946), indicating broad disease impacts or pleiotropy for human disease genes on multiple anatomical systems. This phenotypic effect across systems might arise from the complex clinical symptoms of rare diseases that manifests in multiple organs, tissues, or systems, which could indicate the levels of pleiotropy [38-40]. Hence, the comprehensive and deep phenotyping offered by HPO delivers a more systematic perspective on the functional roles of human disease genes, compared to the commonly used functional inferences based on human gene expression profile or *in vitro* screening. Interestingly, we discovered a significant negative correlation between the median Ka/Ks ratios and the number of affected anatomical systems in disease genes (the Pearson correlation coefficient $\rho = -0.83$, $p = 0.0053$). This implies that disease genes exhibiting higher pleiotropy, impacting multiple anatomical systems, are subject to more stringent evolutionary constraints compared to genes with low pleiotropy (Figure 1e).

Disease gene emergence rate per million years is similar across macroevolutionary epochs.

To comprehend whether different evolutionary epochs have different emergence rate for disease genes, we assessed the disease gene emergence rate per million years across macroevolutionary stages from Euteleostomi to Primate (μ_{da}). Considering the sampling space variations at different age group, we calculated μ_{da} as the fraction of disease genes per

million years at each stage (Figure 2a). Although the proportions of disease genes were found to gradually increase from young to old age groups (Figure 1a), the rate μ_{da} is nearly constant ~0.07% per million years for different age groups (Figure 2a). This constant disease gene emergence rate suggests a continuous and similar fraction of genes evolving to have significant impacts on health. Using the recently reported average human generation time of 26.9 years [41], the most updated number of coding genes (19,831 based on Ensembl v110), and assuming the simplified monogenic model [42], we estimated the number of casual genes for rare diseases per individual per generation (μ_d) as 3.73 x 10^{-4} (= 19,831 x 26.9 x 0.07 x 10^{-8}). Using this rate, we can derive the rare disease prevalence rate (r_{RD} = 10,000 x μ_d), which equates to approximately 4 in 10,000 individuals. This prevalence agrees remarkably well with the EU definition of rare disease rate prevalence of 5 in 10,000 people [43]. The constant parameter highlights the idea that young genes continually acquire functions vital for human health, which agrees with previous observations of young genes and their importance in contributing to phenotypic innovations [44-46].

Young genes are highly enriched into phenotypes of the reproductive and nervous system.

Despite the nearly constant integration of young genes (Figure 2a), it remains uncertain if gene age could influence disease phenotypic spectrums (or pleiotropy). The overall distribution of OP system counts for disease genes (Supplementary figure 2) is similar with the distribution of gene expression breath across tissues (Supplementary figure 3a-3c). The distribution for the numbers of OP systems showed that young genes have lower peak and median values than older genes (Figure 2b-2c). This pattern was consistent with the results that younger genes tend to express in a limited range of tissues, while older genes exhibit a broader expression profile (Supplementary figure 3d), which also aligns with previously reported expression profiles [11, 47-49]. We found an increasing trend for the median numbers of OP systems from young to old evolutionary epochs (Figure 2c). Interestingly, the increase rates ($\frac{\Delta OP_median}{\Delta t}$) are higher at the younger epochs than other older ones (0.12/mya at Eutherian stage vs. 0.05/mya at older stages on average, Supplementary table 4a), suggesting a non-linear and restricted growth model for the level of pleiotropy. We applied a logistic growth function and observed a significant pattern: as evolutionary time increases, the level of pleiotropy rises (Figure 2d). Moreover, the model demonstrates a diminishing marginal effect, indicating that the rate of increase in pleiotropy slows down as evolutionary time continues to grow. This pattern suggests that pleiotropy is initially lower in new genes but increases at a faster

rate compared to older genes. In addition, the higher pleiotropy in older genes is attributed to the cumulative effects over evolutionary history, rather than being inherently high from the outset.

To understand a finer-scale pattern of disease phenotypes for young and old genes, we introduced a metric of the disease phenotype enrichment index (PEI), which accounts for the range of phenotypes on multiple systems (see method for details). Our findings revealed that the most ancient genes, specifically from the Euteleostomi and Tetrapoda periods, had the strongest PEI association with the nervous system (OP1). Conversely, young genes from Amniota and Eutheria epochs tend to display the highest PEI for disease phenotypes of the genitourinary system (OP7) and the nervous system (OP1), with the former showing a 38.65% higher PEI than the latter (Figure 2e, Supplementary table 4). Among the 22 disease phenotype systems, only the reproductive system (OP7) was unique in showing a steady rise in PEI from older epochs to younger ones (Figure 2e). There were smaller variations in PEI for the older epochs when compared to the more recent Eutheria epoch (~2.79 vs. 3.67), hinting that older disease genes impact a greater number of organ systems, as also shown in Figure 2c. This finding is consistent with the “out-of-testis” hypothesis [45], which was built on many observations where the expression patterns of young genes are limited to the testes and can have vital roles in male reproduction. As genes evolve over time, their expression tends to broaden potentially leading to increased phenotypic effects that impact multiple organ systems.

Apart from the reproductive system (OP7), we found that the nervous system (OP1) showed the second highest PEI for Eutherian young disease genes (Figure 2e). Moreover, 42% of the 19 Primate-specific disease genes with diseases affecting the nervous system (OP1) correlate with phenotypes involving brain size or intellectual development (*CFC1*, *DDX11*, *H4C5*, *NOTCH2NLC*, *NOTCH2NLA*, *NPAP1*, *RRP7A*, and *SMPD4*. Supplementary table 2 and Discussion), consistent with the expectation of previous studies based on gene expression [29]. Furthermore, young genes emerging during the primate stage are connected to disease phenotypic enrichment in other adaptive systems, particularly in the HPO systems of the head, neck, eyes, and musculoskeletal structure (Figure 2e). Overall, the Primate-specific disease genes could impact phenotypes from both reproductive and non-reproductive systems, particularly the genitourinary, nervous, and musculoskeletal systems (Supplementary table 2), supporting their roles in both sexual and adaptive evolution.

Sex chromosomes are enriched for disease-associated genes.

Young gene duplicates with a bias toward male expression show chromosomal shifts between sex chromosomes and autosomes [50]. This movement might be an adaptation to address sexual conflicts in gamete formation or to avoid meiotic sex chromosome inactivation (MSCI) in spermatogensis [50-54]. Considering the rapid concentration of the youngest disease genes in the reproductive system (Figure 2e, OP7), we hypothesized that disease genes related to various organs or tissues could have skewed chromosomal distributions. First, we examined the distribution of all disease genes and found a distinct, uneven spread across chromosomes (Figure 3a and Supplementary table 5). The X and Y chromosomes have more disease genes than autosomal ones. While autosomes have a linear slope of 0.23 (Figure 3b, R^2 = 0.93; p = 2.2 x 10^{-13}), the Y chromosome's disease gene proportion is 82.61% higher at 0.42. Meanwhile, the X chromosome's proportion is 30.43% more than autosomes, sitting at 0.301.

To understand if the differences between sex chromosomes and autosomes relate to reproductive functions, we divided disease genes into reproductive (1285 genes) and non-reproductive (3661 genes) categories (See Supplementary tables 6 and 7). By fitting the number of disease genes against all dated genes on chromosomes, we observed that the X chromosome exhibited a bias towards reproductive functions. Specifically, the X chromosome had slightly fewer disease genes affecting non-reproductive systems compared to autosomes (excess rate -1.65%, observed number 154, expected number 156.59). In contrast, the X chromosome displayed a significant surplus of reproductive-related disease genes (observed number 99, expected number 52.73, excess rate 87.75%, p < 5.56e-9) (Figure 3d). This result highlights the prominent difference in functional distribution between the X chromosome and autosomes, which might be attributed to the X chromosome's unique role in reproductive functions. Given the sex-imbalanced mode of inheritance for the X chromosome, theoretical models have predicted that purifying selection would remove both dominant female-detrimental mutations and recessive male-detrimental mutations [55, 56]. We determined that the ratio of male to female reproductive disease genes ($M_{disease}/F_{disease}$ or α_d) is considerably higher for the X chromosome (80/9 = 8.89) than for autosomes on average (38/21 = 1.81, odds ratio = 16.08, 95% CI: 6.73-38.44, p < 0.0001). This suggests a disproportionate contribution of disease genes from the male hemizygous X chromosome compared to the female homozygous X. Thus, our analysis indicates that the abundance of disease genes on the X chromosome compared to autosomes might largely stem from male-specific functional effects. These data also hint that the overrepresentation of disease genes on the X chromosome is

driven primarily by the recessive X-linked inheritance affecting male phenotypes rather than the dominant X-linked effect that impacts both genders.

Sexual selection drives the uneven chromosomal distribution of reproductive disease genes.

To determine which gender might influence the biased distribution of reproductive-related genes on different chromosomes, we focused on genes specific to male and female reproductive disease. Based on the HPO terms of abnormalities in reproductive organs and gene age dating, we retrieved 154 female-specific and 945 male-specific disease genes related to the reproductive system with age dating data (Supplementary table 5 and 6). Through linear regression analysis, we assessed the number of gender-specific reproductive disease genes against the total counted genes for each chromosome. We observed strikingly different patterns that are dependent on gender and chromosomes.

For female reproductive disease genes, the X chromosome did not differ from autosomes, adhering to a linear autosomal pattern (R^2 = 0.53, p = 1.04e-4, Figure 3e). However, when examining male reproductive disease genes, the X and Y chromosomes starkly stood out compared to autosomes, which followed a linear pattern (R^2 = 0.82, p = 5.56e-9, Figure 3f). The X chromosome held an 111.75% more male reproductive genes than expected. Moreover, compared to autosomes (averaging 38/853), the sex chromosomes, Y (17/45) and X (80/840), demonstrated significantly higher ratios of male reproductive disease genes, with odds ratios of 8.48 (95% CI: 4.45 – 16.17, p < 0.0001) and 2.14 (95% CI: 1.44 to 3.18, p = 0.0002), respectively. On the X chromosome, the fraction of male reproductive genes was 10.43 times greater than that of female reproductive genes (80/840 vs. 7/840). This observation is consistent with the “faster-X hypothesis”, where purifying selection is more effective in eliminating recessive deleterious mutations on the X chromosome due to the male hemizygosity of the X chromosome [55, 56]. Interestingly, we also observed a male-bias in reproductive disease gene density on autosomes, where the slope of the autosomal linear model for males was approximately 4.21 times steeper than for female (0.038 vs. 0.0073) (Figure 3e and 3f). Thus, our observed excess of male reproductive disease genes is not caused solely by the “faster-X” effect. It might also be influenced by the “faster-male” effect, postulating that the male reproductive system evolves rapidly due to heightened sexual selection pressures on males [57].

Excess of young genes with male reproductive disease phenotypes

While we observed a male-bias in reproductive disease genes, the influence of gene ages as a factor on this excess remains unclear. We compared gene distribution patterns between older (or ancient, stage Euteleostomi) and younger (post-Euteleostomi) stages. For female-specific reproductive disease genes, the X chromosome has an excess of ancient genes but a deficiency of young genes (25.42% vs. -57.16%, Figure 4a). Conversely, for male-specific reproductive disease genes, younger genes exhibited a higher excess rate than ancient genes (193.96% vs. 80.09%) (Figure 4a). These patterns suggest an age-dependent functional divergence of genes on the X chromosome, which is consistent with gene expression data. The X chromosome is “masculinized” with young, male-biased genes and old X chromosomal genes tend to be “feminized,” maintaining expression in females but losing it in males [52]. On autosomes, the linear regression slope values were higher for male reproductive disease genes than for female ones, both for ancient (0.027 vs. 0.0041) and young genes (0.012 vs. 0.0021) (Figure 4a). The ratio of male to female reproductive disease gene counts (α_d) showed a predominantly male-biased trend across epochs, with a higher value in the most recent epoch of Eutheria (9.75) compared to the ancient epochs Euteleostomi and Tetrapoda (6.40 and 3.94, Figure 4b). Selection pressure comparison between young and ancient genes revealed no significant difference for female-specific reproductive disease genes, but significant difference for male-specific ones (Figure 4c, the Wilcoxon rank-sum test, $p < 0.0001$), indicating that young genes under sexual selection have less evolutionary constraints than older ones (median Ka/Ks ratio 0.35 vs. 0.23).

Structurally, the eutherian hemizygous X chromosome comprises an ancestral X-conserved region and a relatively new X-added region [58]. The ancestral X-conserved region is shared with the marsupial X chromosome, whereas the X-added region originates from autosomes (Figure 4d). To understand which human X chromosome regions might contribute differentially to human genetic disease phenotypes, we compared genes within the X-conserved and X-added regions, based on previous evolutionary strata and X chromosome studies [59-61]. After excluding genes on X-PAR (pseudoautosomal regions) regions (Ensembl v110), we found that the proportion of male-specific reproductive disease genes in X-added region (13.07%, 23/176) exceeds that in the X-conserved region (8.33%, 55/660) (Figure 4d and 4e, Supplementary table 7). Moreover, analyses of the evolutionary strata, which relies on substitutions method [62, 63] and the segmentation and clustering method [64], consistently showed higher fractions of male-specific reproductive disease genes in younger evolutionary strata than in older ones (Figure 4e). These observations indicate that, on the X chromosome, young genes could be more susceptible to the forces of sexual selection than old genes, despite their nearly identical hemizygous environment.

Discussion

The underestimated roles of young genes in human biomedically important phenotypes and innovations.

After the discovery of the first disease gene in 1983, which was based on linkage mapping for a Huntington's disease with pedigree [65], there has been a rapid advancement in medical genetics research. As of now, this field has identified approximately 20% of human genes (~4000-5000 genes) with phenotypes of the rare or "orphan" diseases [7, 66-77]. In our study, we utilized the latest disease gene and clinical phenotype data from HPO annotations [33] and incorporated synteny-based gene age dating to account for new gene duplication events [32]. Contradicting the prior belief that only a tiny fraction of Eutherian young genes are related to diseases [16], our synteny-based gene age dating reveals almost a tenfold increase, suggesting the substantial role of young genes in human biomedical phenotypes. Despite previous debates on the selective pressure of disease genes [16, 78-80], our comparative analyses of Ka/Ks ratios between humans and primates consistently show stronger purifying selection on disease genes than non-disease genes, indicating evolutionary constraints to remove harmful mutations. The epoch-wise estimates of the emergence rate of disease genes per million years reveal a steady integration of genes into disease phenotypes, which supports Haldane's seminal 1937 finding that new deleterious mutations are eliminated at the same rate they occur [81, 82].

Young genes rapidly acquire phenotypes under both sexual and natural selection.

The chromosomal distribution of all disease genes shows the excess of disease genes in X chromosome (Figure 3), which supports the “faster-X effect” [55, 56], that male X-hemizygosity could immediately expose the deleterious X chromosome mutations to purifying selection. Conversely, the X-chromosome inactivation (XCI) in female cells could lessen the deleterious phenotypes of disease variants on the X chromosome [83]. The X chromosome excess of disease genes is attributed predominantly to that of the male reproductive disease genes (Figure 3). This male-specific bias was not limited to the sex chromosome but also detectable in autosomes (Figure 3). These findings align with the “faster-male” effect, where the reproductive system evolves more rapidly in males than in females due to heightened male-specific sexual selection [57]. Intriguingly, of the 22 HPO systems, young genes are enriched in disease phenotypes affecting the reproductive-related system. As genes age, there's a marked decline in both PEI (phenotype enrichment index) and (the male-to-female ratio of reproductive disease gene numbers). These patterns are consistent with the “out of testis” hypothesis [45],

which describes the male germline as a birthplace of new genes due to factors including the permissive chromatin state and the immune environment in testis [84, 85]. The “out of testis” hypothesis predicts that genes could gain broader expression patterns and higher phenotypic complexity over evolutionary time [85]. Consistently, we observed a pattern where older sets of disease genes have phenotypes over a much broader anatomical systems compared to younger genes which tend to impact limited systems. The strong enrichment of male reproductive phenotypes for young genes is also consistent with findings from model species that new genes often exhibit male-reproductive functions [50, 86], in both *Drosophila* [53, 86, 87] and mammals [51, 88]. Some new gene duplicates on autosomes are indispensable during male spermatogenesis, to preserve male-specific functions that would otherwise be silenced on the X chromosome due to the meiotic sex chromosome inactivation (MSCI) [51, 52, 88].

Apart from the reproductive functions, new genes are also enriched for adaptive phenotypes. Previous transcriptomic studies indicate that new genes have excessive upregulation in the human neocortex and under positive selection [29]. The brain size enlargement, especially the neocortex expansion over ~50% the volume of the human brain, ranks among the most extraordinary human phenotypic innovations [29, 89]. Here, we found that at least 42% of primate-specific disease genes affecting the nervous systems could impact phenotypes related to brain size and intellectual development. For example, *DDX11* is critical in pathology of microcephaly [90-93]. The *NOTCH2NLA*, *NOTCH2NLB,* and *NOTCH2NLC* may promote human brain size enlargement, due to their functions in neuronal intranuclear inclusion disease (NIID), microcephaly, and macrocephaly [94-96]. The *RRP7A* is also a microcephaly disease gene evidenced from patient-derived cells with defects in cell cycle progression and primary cilia resorption [97]. The defects of *SMPD4* can lead to a neurodevelopmental disorder characterized by microcephaly and structural brain anomalies [98]. The *SRGAP2C* accounts for human-specific feature of neoteny and can promote motor and execution skills in mouse and monkey model [99-101]. The *de novo* gene *SMIM45* [102] associates with cortical expansion based on extensive models [31].

New genes were also found with enrichment in other adaptive phenotypes, particularly involving the head and neck, eye, and musculoskeletal system. Some examples of these primate-specific disease genes encompass *CFHR3* associated with macular degeneration [103], *SMPD4* with the retinopathy [104], *TUBA3D* with the keratoconus [105], *OPN1MW* with loss of color vision [106, 107], *YY1AP1* with Fibromuscular dysplasia [108], *SMN2* with the Spinal Muscular Atrophy [109], *GH1* with defects in adult bone mass and bone loss [110], *KCNJ18* with thyrotoxicosis complicated by paraplegia

and hyporeflexia [111], *TBX5* with the cardiac and limb defects of Holt-Oram syndrome [112, 113], and *DUX4* with muscular dystrophy [114]. Additionally, some other specific functions have also been reported for these young genes. For example, the Y chromosome gene *TBL1Y* could lead to male-specific hearing loss [115]. The *TUBB8* defects could lead to complete cleavage failure in fertilized eggs and oocyte maturation arrest [116-118]. Interestingly, a previous case study on mice also shows the role of *de novo* genes on female-specific reproductive functions [119]. These emerging studies independently support the importance of new genes in phenotypic innovation and sexual selection, refuting previous assumptions that new genes contribute little to phenotypic innovation [120].

New genes underlying rapid phenotypic innovations: low pleiotropy as a selective advantage.

Our findings raise the question of why new genes can quickly enrich into phenotypic traits that are crucial for both sexual evolution and adaptive innovation. This question could not be fully addressed by previous hypotheses. The "out of testis" theory, as well as the "male-driven," "faster-X," and "faster-male" theories, do not offer specific predictions regarding the propensity of new or young genes to be involved in adaptive traits. Here, we proposed a “pleiotropy-barrier” model to explain the relationship between innovation potential and gene ages (Figure 5a). The evidence of extensive pleiotropy was found early in the history of genetics [121-123]. It is established that young genes exhibit higher specificity and narrower expression breadth across tissues [48]. In this study, we used a broader definition of pleiotropy to understand phenotype evolution [38, 124-126]. We reveal a pattern that older genes tend to impact more organs/systems, while young genes display phenotype enrichment in specific organs (Figure 2c). Therefore, both phenotype pattern and expression trend across evolutionary epochs suggest lower pleiotropy for young genes, compared to the progressively higher pleiotropy observed in older genes.

Numerous theoretical and genomic studies have revealed that pleiotropy impedes evolutionary adaptation (a so-called ‘cost of complexity’) [121, 127-130], while low pleiotropy could foster more morphological evolutions [131, 132]. The inhibitory effect of pleiotropy on novel adaptation aligns with our observations of the strong purifying selection on both high extent of pleiotropy [127, 128] and expression breadth [133]. As expected, we observed that multi-system genes and older genes, which exhibit higher pleiotropy, undergo stronger purifying selection (Figure 1b-1e). This evolutionary constraint suggests a restricted mutation space to introduce novel traits for old genes due to the “competing interests” of

multifunctionality (Figure 5). The inhibitory pressure could also reduce genetic diversity due to background selection [134]. The evolution of new genes, especially gene duplicates, serves as a primary mechanism to mitigate pleiotropic effects through subfunctionalization and neofunctionalization [135, 136] and avoid adverse pleiotropy in ancestral copies [137]. The tissue-specific functions of new genes, as a general pattern in numerous organisms, could circumvent the adaptive conflicts caused by the multifunctionality of parental genes [138]. The reduced pleiotropy in young genes could thereby allow for a more diverse mutational space for functional innovation without triggering unintended pleiotropic trade-offs [139].

The “pleiotropy-barrier” model predicts that the capacity for phenotypic innovation is limited by genetic pleiotropy under nature selection (Figure 5a). Over evolutionary time, the pleiotropy increase follows a logistic growth pattern, where the speed of growth could be higher for younger genes but lower for older genes (Figure 5b). The multifunctional genes could encounter an escalating “barrier” toward the pleiotropy maximum. This barrier arises because more functions necessitate stronger selective constraints, which could in turn reduce mutational space of beneficial mutations for novel phenotypes. In contrast, low or absent pleiotropy in new genes allows for a higher and tunable mutation space under the relaxed purifying selection. The permissive environment provides a fertile ground for beneficial mutations to appear with novel functions. Such innovations, initially as polymorphisms within a population, can become advantageous phenotypes and ready responder in certain environment under positive selection. Therefore, young genes, with lower pleiotropic effect as a selective advantage, not only spurs molecular evolution under sexual and natural selection but, from a medical standpoint, also are promising targets for precise medicine, warranting deeper investigation.

Conclusion

In this study, we unveil a remarkable pattern of new gene evolution with vital pathogenic functions shaped by the non-neutral selection. Although the ratio of genes associated with health-related functions per million years remains relatively consistent across macroevolutionary epochs, we note an enrichment pattern of disease systems for young genes. Importantly, young genes are preferentially linked to disease phenotypes of the male reproductive system, as well as systems that undergone significant phenotypic innovations in primate or human evolution, including the nervous system, head and neck, eyes, and the musculoskeletal system. The enrichment of these disease systems points to the driving

forces of both sexual selection and adaptive evolution for young genes. As evolutionary time progresses, older genes display fewer specialized functions compared to their young counterparts. Our findings highlight that young genes are likely the frontrunners of molecular evolution, being actively selected for functional roles by both adaptive innovation and sexual selection, a process aided by their lower pleiotropy. Therefore, young genes play a pivotal role in addressing a multitude of questions related to the fundamental biology of humans.

Materials and Methods

Gene age dating and disease phenotypes

The gene age dating was conducted using an inclusive approach. For autosomal and X chromosomal genes, we primarily obtained gene ages (or branches, origination stages) from the GenTree database [32, 52] that is based on Ensembl v95 of human reference genome version hg38 [140]. We then trans-mapped the v95 gene list of GenTree into the current release of Ensembl gene annotation (v110). The gene age inference in the GenTree database relied on genome-wide synteny and was based on the presence of syntenic blocks obtained from whole-genome alignments between human and outgroup genomes [11, 32, 52]. The most phylogenetically distant branch at which the shared syntenic block was detected marks the time when a human gene originated. In comparison to the method based on the similarity of protein families, namely the phylostratigraphic dating [141], this method employed in GenTree is robust to recent gene duplications [32], despite its under-estimation of the number of young genes [90]. We obtained gene age for human Y genes through the analysis of 15 representative mammals [142]. Notably, Y gene ages are defined as the time when these genes began to evolve independently of their X counterpart or when they translocated from other chromosomes to the Y chromosome due to gene traffic (transposition/translocation) [142]. For the remaining Ensembl v110 genes lacking age information, we dated them using the synteny-based method with the gene order information from ENSEMBL database (v110), following the inference framework of GenTree [32]. These comprehensive methods resulted in the categorization of 19,665 protein-coding genes into distinct gene age groups, encompassing evolutionary stages from Euteleostomi to the human lineage, following the phylogenetic framework of the GenTree database. The HPO annotation used in this study for phenotypic abnormalities contains disease genes corresponding to 23 major organ/tissue systems (09/19/2023, https://hpo.jax.org/app/data/annotations). After filtering out mitochondrial genes, unplaced genes, RNA genes, and genes related to neoplasm ontology, we obtained with gene ages and phenotypic abnormalities (across 22 categories) for 4946 protein-coding genes. The

reproductive system disease genes were retrieved from the “phenotype to genes.txt” file based on “reproduct”, “male”, “female” keywords (neoplasm-related items were removed).

Ka/Ks ratio

Ka/Ks is widely used in evolutionary genetics to estimate the relative strength of purifying selection (Ka/Ks < 1), neutral mutations (Ka/Ks = 1), and beneficial mutations (Ka/Ks > 1) on homologous protein-coding genes. Ka is the number of nonsynonymous substitutions per non-synonymous site, while Ks is the number of synonymous substitutions per synonymous site that is assumed to be neutral. The pairwise Ka/Ks ratios (human-chimpanzee, human-bonobo, and human-macaque) were retrieved from the Ensembl database (v99) [140], as estimated with the Maximum Likelihood algorithm [143].

Disease gene emergence rate per million years (r)

To understand the origination tempo of disease genes within different evolutionary epochs, we estimated the disease gene emergence rate per million years *r* for disease genes, which is the fractions of disease genes per million years for each evolutionary branch. The calculating is based on the following formula:

$$r_i = \frac{O_i}{A_i T_i}$$

where r_i represents the phenotype integration index for ancestral branch *i*. The O_i indicates the number of disease genes with organ phenotypes in ancestral branch *i*. The denominator A_i is the number of genes with gene age information in branch *i*. The T_i represents the time obtained from the Timetree database (http://www.timetree.org/) [144].

Pleiotropic modeling with logistic growth function

For each evolutionary epoch (t), we estimated the median numbers of OP systems that genic defects could affect, which serve as the proxy of pleiotropy over evolutionary time (P(t)) for regression analysis. The logistic growth function was used to fit the correlation with the Nonlinear Least Squares in R.

Phenotype enrichment along evolutionary stages

The phenotype enrichment along evolutionary epochs was evaluated based on a phenotype enrichment index (PEI). Specifically, within "gene-phenotype" links, there are two types of contributions for a phenotype, which are "one gene, many phenotypes" due to potential pleiotropism as well as "one gene, one phenotype". Considering the weighting differences between these two categories, we estimated the $PEI_{(i,j)}$ for a given phenotype (p_i) within an evolutionary stage (br_j) with the following formula.

$$\mathrm{PEI}_{(i,j)} = \frac{\sum_{i=1}^{n} \frac{1}{m_i}}{\sum_{j=1}^{l} \sum_{k=1}^{n_j} \frac{1}{m_k}}$$

The m indicates the number of phenotype(s) one gene can affect, n represents the number of genes identified for a given phenotypes, and l is number of phenotypes within a given evolutionary stage. Considering the genetic complexity of phenotypes, the enrichment index (*PEI*) firstly adjusted the weights of genes related to a phenotype with the reciprocal value of m, *i.e.*, $\frac{1}{m}$. Thus, the more phenotypes a gene affects, the less contributing weight this gene has. Here, m_i is the number of phenotypes affected by the *i*-th gene, n is the total number of genes associated with the specific phenotype p_i, n_j is the number of genes associated with the *j*-th phenotype within the evolutionary stage, and m_k is the number of phenotypes affected by the *k*-th gene within the *j*-th phenotype. Then, we can obtain the accumulative value (*p*) of the adjusted weights of all genes for a specific phenotype within an evolutionary stage. Because of the involvement of multiple phenotypes within an evolutionary stage, we summed weight values for all phenotypes ($\sum_{j=1}^{l} p$) and finally obtained the percentage of each phenotype within each stage ($\frac{p}{\sum_{j=1}^{l} p}$) as the enrichment index.

The linear regression and excessive rate

The linear regression for disease genes and total genes on chromosomes was based on the simple hypothesis that the number of disease genes would be dependent on the number of total genes on chromosomes. The linear regression and statistics were analyzed with R platform. The excessive rate was calculated as the percentages (%) of the vertical difference between a specific data point, which is the number of gene within a chromosome (*n*), and the expected value based on linear model (*n*-*e*) out of the expected value ($\frac{n-e}{e}$).

The X-conserved and X-added regions

The Eutherian X chromosome is comprised of the pseudoautosomal regions (PAR), X-conserved region, and X-added region. The regions of two PAR were determined based on NCBI assembly annotation of GRCh38.p13 (X:155701383-156030895 and X:10001-2781479). The X-boundary between X-conserved and X-added regions was determined with Ensembl biomart tool. The "one to one" orthologous genes between human and opossum were used for gene synteny identification. The X-conserved region is shared between human and opossum, while X-added region in human has synteny with the autosomal genes of opossum [61]. The "evolutionary strata" on X were based on previous reports of two methods:

substitutions method and the Segmentation and Clustering method [59, 60, 145]. The coordinates of strata boundaries were up-lifted into hg38 genome with liftover tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver).

Acknowledgments:

Manyuan Long was supported by the John Simon Guggenheim Memorial Fellowship for Natural Sciences (2022) and the University of Chicago Division of Biological sciences, the National Institutes of Health (1R01GM116113-01A1) and the National Science Foundation (NSF2020667). Deanna Arsala was supported by National Institutes of Health (F32GM146423). We greatly appreciate the constructive discussions with Dr. Stefano Allesina of the University of Chicago, Dr. Anne O'Donnell Luria of the Broad Institute of M.I.T. and Harvard, Dr. Cheng Deng of Western China Hospital, Dr. Chengchi Fang from the Chinese Academy of Sciences, Dr. Shuaibo Han from Zhejiang Agriculture and Forestry University, and Dr. Chuanzhu Fan from Wayne State University. Special acknowledgment is given to Xuefei He from Western China Hospital for designing the silhouettes. The authors extend their gratitude to the maintainers and contributors of the HPO data.

Competing Interests: The authors have declared that no competing interests exist.

References

1. Clausius R. The mechanical theory of heat: Macmillan; 1879.
2. Vanchurin V, Wolf YI, Koonin EV, Katsnelson MI. Thermodynamics of evolution and the origin of life. Proceedings of the National Academy of Sciences. 2022;119(6):e2120042119. doi: doi:10.1073/pnas.2120042119.
3. and TAK, Bebenek K. DNA Replication Fidelity. Annual Review of Biochemistry. 2000;69(1):497-529. doi: 10.1146/annurev.biochem.69.1.497. PubMed PMID: 10966467.
4. Pavličev M, Wagner GP. The value of broad taxonomic comparisons in evolutionary medicine: Disease is not a trait but a state of a trait. MedComm (2020). 2022;3(4):e174. Epub 20220922. doi: 10.1002/mco2.174. PubMed PMID: 36186235; PubMed Central PMCID: PMCPMC9495303.
5. Fetro C, Scherman D. Drug repurposing in rare diseases: Myths and reality. Therapies. 2020;75(2):157-60.
6. Antonarakis SE, Beckmann JS. Mendelian disorders deserve more attention. Nature Reviews Genetics. 2006;7(4):277-82. doi: 10.1038/nrg1826.
7. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nature Reviews Genetics. 2013;14(10):681-91. doi: 10.1038/nrg3555.
8. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, et al. A brief history of human disease genetics. Nature. 2020;577(7789):179-89. doi: 10.1038/s41586-019-1879-7.

9. Wu D-D, Irwin DM, Zhang Y-P. De Novo Origin of Human Protein-Coding Genes. PLOS Genetics. 2011;7(11):e1002379. doi: 10.1371/journal.pgen.1002379.
10. Van Oss SB, Carvunis AR. De novo gene birth. PLoS Genet. 2019;15(5):e1008160. Epub 20190523. doi: 10.1371/journal.pgen.1008160. PubMed PMID: 31120894; PubMed Central PMCID: PMCPMC6542195.
11. Long M, VanKuren NW, Chen S, Vibranovski MD. New Gene Evolution: Little Did We Know. Annual Review of Genetics. 2013;47(1):307-33. doi: 10.1146/annurev-genet-111212-133301. PubMed PMID: 24050177.
12. Conrad B, Antonarakis SE. Gene Duplication: A Drive for Phenotypic Diversity and Cause of Human Disease. Annual Review of Genomics and Human Genetics. 2007;8(1):17-35. doi: 10.1146/annurev.genom.8.021307.110233. PubMed PMID: 17386002.
13. Zhang D, Leng L, Chen C, Huang J, Zhang Y, Yuan H, et al. Dosage sensitivity and exon shuffling shape the landscape of polymorphic duplicates in Drosophila and humans. Nature Ecology & Evolution. 2022;6(3):273-87. doi: 10.1038/s41559-021-01614-w.
14. Kaessmann H, Zöllner S, Nekrutenko A, Li WH. Signatures of domain shuffling in the human genome. Genome Res. 2002;12(11):1642-50. doi: 10.1101/gr.520702. PubMed PMID: 12421750; PubMed Central PMCID: PMCPMC187552.
15. Betrán E, Long M. Evolutionary New Genes in a Growing Paradigm. Genes. 2022;13(9):1605. PubMed PMID: doi:10.3390/genes13091605.
16. Domazet-Lošo T, Tautz D. An ancient evolutionary origin of genes associated with human genetic diseases. Molecular biology and evolution. 2008;25(12):2699-707.
17. Ding D, Nguyen TT, Pang MYH, Ishibashi T. Primate-specific histone variants. Genome. 2021;64(4):337-46. doi: 10.1139/gen-2020-0094 %M 33245240.
18. Ciccarelli FD, von Mering C, Suyama M, Harrington ED, Izaurralde E, Bork P. Complex genomic rearrangements lead to novel primate gene function. Genome Research. 2005;15(3):343-51. doi: 10.1101/gr.3266405.
19. Cai J, Zhao R, Jiang H, Wang W. De Novo Origination of a New Protein-Coding Gene in Saccharomyces cerevisiae. Genetics. 2008;179(1):487-96. doi: 10.1534/genetics.107.084491.
20. Li D, Dong Y, Jiang Y, Jiang H, Cai J, Wang W. A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. Cell Research. 2010;20(4):408-20. doi: 10.1038/cr.2010.31.
21. Parikh SB, Houghton C, Van Oss SB, Wacholder A, Carvunis A-R. Origins, evolution, and physiological implications of de novo genes in yeast. Yeast. 2022;39(9):471-81. doi: https://doi.org/10.1002/yea.3810.
22. Ross BD, Rosin L, Thomae AW, Hiatt MA, Vermaak D, de la Cruz AFA, et al. Stepwise evolution of essential centromere function in a Drosophila neogene. Science. 2013;340(6137):1211-4.
23. Kasinathan B, Colmenares SU, III, McConnell H, Young JM, Karpen GH, Malik HS. Innovation of heterochromatin functions drives rapid evolution of essential ZAD-ZNF genes in Drosophila. eLife. 2020;9:e63368. doi: 10.7554/eLife.63368.
24. Ragsdale EJ, Müller MR, Rödelsperger C, Sommer RJ. A developmental switch coupled to the evolution of plasticity acts through a sulfatase. Cell. 2013;155(4):922-33.
25. Klomp J, Athy D, Kwan CW, Bloch NI, Sandmann T, Lemke S, et al. A cysteine-clamp gene drives embryo polarity in the midge Chironomus. Science. 2015;348(6238):1040-2.
26. Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. Emergence of Young Human Genes after a Burst of Retroposition in Primates. PLOS Biology. 2005;3(11):e357. doi: 10.1371/journal.pbio.0030357.
27. Zhang YE, Long M. New genes contribute to genetic and phenotypic novelties in human evolution. Current opinion in genetics & development. 2014;29:90-6.

28. Shi L, Su B. Identification and functional characterization of a primate-specific E2F1 binding motif regulating MCPH1 expression. The FEBS Journal. 2012;279(3):491-503. doi: https://doi.org/10.1111/j.1742-4658.2011.08441.x.
29. Zhang YE, Landback P, Vibranovski MD, Long M. Accelerated recruitment of new brain development genes into the human genome. PLoS biology. 2011;9(10):e1001179.
30. Rich A, Carvunis A-R. De novo gene increases brain size. Nature Ecology & Evolution. 2023;7(2):180-1. doi: 10.1038/s41559-022-01942-5.
31. An NA, Zhang J, Mo F, Luan X, Tian L, Shen QS, et al. De novo genes with an lncRNA origin encode unique human brain developmental functionality. Nature Ecology & Evolution. 2023;7(2):264-78. doi: 10.1038/s41559-022-01925-6.
32. Shao Y, Chen C, Shen H, He BZ, Yu D, Jiang S, et al. GenTree, an integrated resource for analyzing the evolution and function of primate-specific coding genes. Genome research. 2019.
33. Köhler S, Carmody L, Vasilevsky N, Jacobsen JO B, Danis D, Gourdine J-P, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Research. 2018;47(D1):D1018-D27. doi: 10.1093/nar/gky1105.
34. INSERM. Orphanet: an online database of rare diseases and orphan drugs. French National Institute for Health and Medical Research Paris; 1997.
35. Weinreich SS, Mangon R, Sikkens J, Teeuw M, Cornel M. Orphanet: a European database for rare diseases. Nederlands tijdschrift voor geneeskunde. 2008;152(9):518-9.
36. Wright ES. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. BMC bioinformatics. 2015;16(1):1-14.
37. Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. Online Mendelian inheritance in man (OMIM). Human mutation. 2000;15(1):57-61.
38. Lobo I. Pleiotropy: one gene can affect multiple traits. 2008.
39. Paul D. A double-edged sword. Nature. 2000;405(6786):515-. doi: 10.1038/35014676.
40. Hodgkin J. Seven types of pleiotropy. Int J Dev Biol. 1998;42(3):501-5. PubMed PMID: 9654038.
41. Wang RJ, Al-Saffar SI, Rogers J, Hahn MW. Human generation times across the past 250,000 years. Sci Adv. 2023;9(1):eabm7047. Epub 20230106. doi: 10.1126/sciadv.abm7047. PubMed PMID: 36608127; PubMed Central PMCID: PMCPMC9821931.
42. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-24. Epub 20150305. doi: 10.1038/gim.2015.30. PubMed PMID: 25741868; PubMed Central PMCID: PMCPMC4544753.
43. Stolk P, Willemen MJ, Leufkens HG. Rare essentials: drugs for rare diseases as essential medicines. Bulletin of the World Health Organization. 2006;84(9):745-51.
44. Chen S, Krinsky BH, Long M. New genes as drivers of phenotypic evolution. Nature Reviews Genetics. 2013;14(9):645-60.
45. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome research. 2010;20(10):1313-26.
46. Xia S, VanKuren NW, Chen C, Zhang L, Kemkemer C, Shao Y, et al. Genomic analyses of new genes and their phenotypic effects reveal rapid evolution of essential functions in Drosophila development. PLOS Genetics. 2021;17(7):e1009654. doi: 10.1371/journal.pgen.1009654.

47. Carelli FN, Hayakawa T, Go Y, Imai H, Warnefors M, Kaessmann H. The life history of retrocopies illuminates the evolution of new mammalian genes. Genome Res. 2016;26(3):301-14. Epub 20160104. doi: 10.1101/gr.198473.115. PubMed PMID: 26728716; PubMed Central PMCID: PMCPMC4772013.

48. Zhang YE, Landback P, Vibranovski M, Long M. New genes expressed in human brains: implications for annotating evolving genomes. Bioessays. 2012;34(11):982-91.

49. Miller D, Chen J, Liang J, Betrán E, Long M, Sharakhov IV. Retrogene Duplication and Expression Patterns Shaped by the Evolution of Sex Chromosomes in Malaria Mosquitoes. Genes. 2022;13(6):968. PubMed PMID: doi:10.3390/genes13060968.

50. Betrán E, Thornton K, Long M. Retroposed New Genes Out of the X in Drosophila. Genome Research. 2002;12(12):1854-9. doi: 10.1101/gr.604902.

51. Emerson J, Kaessmann H, Betrán E, Long M. Extensive gene traffic on the mammalian X chromosome. Science. 2004;303(5657):537-40.

52. Zhang YE, Vibranovski MD, Landback P, Marais GA, Long M. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. PLoS biology. 2010;8(10):e1000494.

53. VanKuren NW, Long M. Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions. Nature ecology & evolution. 2018;2(4):705-12.

54. Wu C-I, Yujun Xu E. Sexual antagonism and X inactivation – the SAXI hypothesis. Trends in Genetics. 2003;19(5):243-7. doi: https://doi.org/10.1016/S0168-9525(03)00058-1.

55. Rice WR. Sex chromosomes and the evolution of sexual dimorphism. Evolution. 1984:735-42.

56. Charlesworth B, Coyne JA, Barton NH. The relative rates of evolution of sex chromosomes and autosomes. The American Naturalist. 1987;130(1):113-46.

57. Wu C-I, Davis AW. Evolution of postmating reproductive isolation: the composite nature of Haldane's rule and its genetic bases. The American Naturalist. 1993;142(2):187-212.

58. Bellott DW, Skaletsky H, Pyntikova T, Mardis ER, Graves T, Kremitzki C, et al. Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. Nature. 2010;466(7306):612-6.

59. Pandey RS, Wilson Sayres MA, Azad RK. Detecting evolutionary strata on the human X chromosome in the absence of gametologous Y-linked sequences. Genome biology and evolution. 2013;5(10):1863-71.

60. McLysaght A. Evolutionary steps of sex chromosomes are reflected in retrogenes. Trends in genetics. 2008;24(10):478-81.

61. Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, et al. The DNA sequence of the human X chromosome. Nature. 2005;434(7031):325-37. doi: 10.1038/nature03440.

62. Lahn BT, Page DC. Four evolutionary strata on the human X chromosome. Science. 1999;286(5441):964-7. doi: 10.1126/science.286.5441.964. PubMed PMID: 10542153.

63. McLysaght A. Evolutionary steps of sex chromosomes are reflected in retrogenes. Trends Genet. 2008;24(10):478-81. Epub 20080905. doi: 10.1016/j.tig.2008.07.006. PubMed PMID: 18774619.

64. Pandey RS, Wilson Sayres MA, Azad RK. Detecting evolutionary strata on the human x chromosome in the absence of gametologous y-linked sequences. Genome Biol Evol. 2013;5(10):1863-71. doi: 10.1093/gbe/evt139. PubMed PMID: 24036954; PubMed Central PMCID: PMCPMC3814197.

65. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, et al. A polymorphic DNA marker genetically linked to Huntington's disease. Nature. 1983;306(5940):234-8. doi: 10.1038/306234a0.

66. Henn BM, Botigué LR, Bustamante CD, Clark AG, Gravel S. Estimating the mutation load in human genomes. Nature Reviews Genetics. 2015;16(6):333-43.
67. Asimit JL, Day-Williams AG, Morris AP, Zeggini E. ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. Human heredity. 2012;73(2):84-94.
68. Povysil G, Petrovski S, Hostyk J, Aggarwal V, Allen AS, Goldstein DB. Rare-variant collapsing analyses for complex traits: guidelines and applications. Nature Reviews Genetics. 2019;20(12):747-59.
69. Krumm N, Turner TN, Baker C, Vives L, Mohajeri K, Witherspoon K, et al. Excess of rare, inherited truncating mutations in autism. Nature Genetics. 2015;47(6):582-8. doi: 10.1038/ng.3303.
70. Ji X, Kember RL, Brown CD, Bućan M. Increased burden of deleterious variants in essential genes in autism spectrum disorder. Proceedings of the National Academy of Sciences. 2016;113(52):15054-9. doi: 10.1073/pnas.1613195113.
71. Thuresson AC, Zander CS, Zhao JJ, Halvardson J, Maqbool K, Månsson E, et al. Whole genome sequencing of consanguineous families reveals novel pathogenic variants in intellectual disability. Clinical genetics. 2019;95(3):436.
72. Almlöf JC, Nystedt S, Leonard D, Eloranta M-L, Grosso G, Sjöwall C, et al. Whole-genome sequencing identifies complex contributions to genetic risk by variants in genes causing monogenic systemic lupus erythematosus. Human genetics. 2019;138(2):141-50.
73. Wen L, Zhu C, Zhu Z, Yang C, Zheng X, Liu L, et al. Exome-wide association study identifies four novel loci for systemic lupus erythematosus in Han Chinese population. Annals of the rheumatic diseases. 2018;77(3):417-.
74. Oud MS, Houston BJ, Volozonoka L, Mastrorosa FK, Holt GS, Alobaidi BKS, et al. Exome sequencing reveals variants in known and novel candidate genes for severe sperm motility disorders. Human Reproduction. 2021;36(9):2597-611. doi: 10.1093/humrep/deab099.
75. Krausz C, Riera-Escamilla A. Genetics of male infertility. Nature Reviews Urology. 2018;15(6):369-84. doi: 10.1038/s41585-018-0003-3.
76. Guo T, Tu C-F, Yang D-H, Ding S-Z, Lei C, Wang R-C, et al. Bi-allelic BRWD1 variants cause male infertility with asthenoteratozoospermia and likely primary ciliary dyskinesia. Human Genetics. 2021;140(5):761-73. doi: 10.1007/s00439-020-02241-4.
77. Study TDDD. Prevalence and architecture of de novo mutations in developmental disorders. Nature. 2017;542(7642):433.
78. Smith NG, Eyre-Walker A. Human disease genes: patterns and predictions. Gene. 2003;318:169-75.
79. Chakraborty S, Panda A, Ghosh TC. Exploring the evolutionary rate differences between human disease and non-disease genes. Genomics. 2016;108(1):18-24.
80. Spataro N, Rodríguez JA, Navarro A, Bosch E. Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. Human molecular genetics. 2017;26(3):489-500.
81. Keightley PD. Rates and fitness consequences of new mutations in humans. Genetics. 2012;190(2):295-304. doi: 10.1534/genetics.111.134668. PubMed PMID: 22345605; PubMed Central PMCID: PMCPMC3276617.
82. Haldane J. The effect of variation of fitness. The American Naturalist. 1937;71(735):337-49.
83. Migeon BR. X-linked diseases: susceptible females. Genetics in Medicine. 2020;22(7):1156-74. doi: 10.1038/s41436-020-0779-4.
84. Bekpen C, Xie C, Tautz D. Dealing with the adaptive immune system during de novo evolution of genes from intergenic sequences. BMC Evolutionary Biology. 2018;18(1):121. doi: 10.1186/s12862-018-1232-z.

85. Vinckenbosch N, Dupanloup I, Kaessmann H. Evolutionary fate of retroposed gene copies in the human genome. Proceedings of the National Academy of Sciences. 2006;103(9):3220-5.

86. Heinen TJ, Staubach F, Häming D, Tautz D. Emergence of a new gene from an intergenic region. Current biology. 2009;19(18):1527-31.

87. Gubala AM, Schmitz JF, Kearns MJ, Vinh TT, Bornberg-Bauer E, Wolfner MF, et al. The goddard and saturn genes are essential for Drosophila male fertility and may have arisen de novo. Molecular biology and evolution. 2017;34(5):1066-82.

88. Jiang L, Li T, Zhang X, Zhang B, Yu C, Li Y, et al. RPL10L is required for male meiotic division by compensating for RPL10 during meiotic sex chromosome inactivation in mice. Current Biology. 2017;27(10):1498-505. e6.

89. Rakic P. Evolution of the neocortex: a perspective from developmental biology. Nat Rev Neurosci. 2009;10(10):724-35. doi: 10.1038/nrn2719. PubMed PMID: 19763105; PubMed Central PMCID: PMCPMC2913577.

90. Ma C, Li C, Ma H, Yu D, Zhang Y, Zhang D, et al. Pan-cancer surveys indicate cell cycle-related roles of primate-specific genes in tumors and embryonic cerebrum. Genome Biology. 2022;23(1):251. doi: 10.1186/s13059-022-02821-9.

91. van Schie JJM, Faramarz A, Balk JA, Stewart GS, Cantelli E, Oostra AB, et al. Warsaw Breakage Syndrome associated DDX11 helicase resolves G-quadruplex structures to support sister chromatid cohesion. Nature Communications. 2020;11(1):4287. doi: 10.1038/s41467-020-18066-8.

92. Lerner LK, Holzer S, Kilkenny ML, Šviković S, Murat P, Schiavone D, et al. Timeless couples G-quadruplex detection with processing by DDX11 helicase during DNA replication. The EMBO Journal. 2020;39(18):e104185. doi: https://doi.org/10.15252/embj.2019104185.

93. Pirozzi F, Nelson B, Mirzaa G. From microcephaly to megalencephaly: determinants of brain size. Dialogues Clin Neurosci. 2018;20(4):267-82. doi: 10.31887/DCNS.2018.20.4/gmirzaa. PubMed PMID: 30936767; PubMed Central PMCID: PMCPMC6436952.

94. Suzuki IK, Gacquer D, Van Heurck R, Kumar D, Wojno M, Bilheu A, et al. Human-specific NOTCH2NL genes expand cortical neurogenesis through Delta/Notch regulation. Cell. 2018;173(6):1370-84. e16.

95. Fiddes IT, Lodewijk GA, Meghan M, Bosworth CM, Ewing AD, Mantalas GL, et al. Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. Cell. 2018;173(6):1356-69.e22.

96. Liu Q, Zhang K, Kang Y, Li Y, Deng P, Li Y, et al. Expression of expanded GGC repeats within NOTCH2NLC causes behavioral deficits and neurodegeneration in a mouse model of neuronal intranuclear inclusion disease. Science Advances. 2022;8(47):eadd6391. doi: doi:10.1126/sciadv.add6391.

97. Farooq M, Lindbæk L, Krogh N, Doganli C, Keller C, Mönnich M, et al. RRP7A links primary microcephaly to dysfunction of ribosome biogenesis, resorption of primary cilia, and neurogenesis. Nat Commun. 2020;11(1):5816. Epub 20201116. doi: 10.1038/s41467-020-19658-0. PubMed PMID: 33199730; PubMed Central PMCID: PMCPMC7670429.

98. Magini P, Smits DJ, Vandervore L, Schot R, Columbaro M, Kasteleijn E, et al. Loss of SMPD4 Causes a Developmental Disorder Characterized by Microcephaly and Congenital Arthrogryposis. Am J Hum Genet. 2019;105(4):689-705. Epub 20190905. doi: 10.1016/j.ajhg.2019.08.006. PubMed PMID: 31495489; PubMed Central PMCID: PMCPMC6817560.

99. Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, et al. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. Cell. 2012;149(4):912-22.

100. Charrier C, Joshi K, Coutinho-Budd J, Kim J-E, Lambert N, De Marchena J, et al. Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. Cell. 2012;149(4):923-35.

101. Meng X, Lin Q, Zeng X, Jiang J, Li M, Luo X, et al. Brain developmental and cortical connectivity changes in the transgenic monkeys carrying the human-specific duplicated gene srGAP2C. National Science Review. 2023. doi: 10.1093/nsr/nwad281.
102. Delihas N. Evolutionary formation of a human de novo open reading frame from a mouse non-coding DNA sequence via biased random mutations. bioRxiv. 2023:2023.08.09.552707. doi: 10.1101/2023.08.09.552707.
103. Fritsche LG, Igl W, Bailey JNC, Grassmann F, Sengupta S, Bragg-Gresham JL, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nature genetics. 2016;48(2):134-43. doi: 10.1038/ng.3448. PubMed PMID: 26691988.
104. Smits DJ, Schot R, Krusy N, Wiegmann K, Utermöhlen O, Mulder MT, et al. SMPD4 regulates mitotic nuclear envelope dynamics and its loss causes microcephaly and diabetes. Brain. 2023;146(8):3528-41. doi: 10.1093/brain/awad033.
105. Hao XD, Chen P, Zhang YY, Li SX, Shi WY, Gao H. De novo mutations of TUBA3D are associated with keratoconus. Sci Rep. 2017;7(1):13570. Epub 20171019. doi: 10.1038/s41598-017-13162-0. PubMed PMID: 29051577; PubMed Central PMCID: PMCPMC5648796.
106. Winderickx J, Sanocki E, Lindsey DT, Teller DY, Motulsky AG, Deeb SS. Defective colour vision associated with a missense mutation in the human green visual pigment gene. Nature Genetics. 1992;1(4):251-6. doi: 10.1038/ng0792-251.
107. Ueyama H, Kuwayama S, Imai H, Tanabe S, Oda S, Nishida Y, et al. Novel missense mutations in red/green opsin genes in congenital color-vision deficiencies. Biochem Biophys Res Commun. 2002;294(2):205-9. doi: 10.1016/s0006-291x(02)00458-8. PubMed PMID: 12051694.
108. Guo DC, Duan XY, Regalado ES, Mellor-Crummey L, Kwartler CS, Kim D, et al. Loss-of-Function Mutations in YY1AP1 Lead to Grange Syndrome and a Fibromuscular Dysplasia-Like Vascular Disease. Am J Hum Genet. 2017;100(1):21-30. Epub 20161208. doi: 10.1016/j.ajhg.2016.11.008. PubMed PMID: 27939641; PubMed Central PMCID: PMCPMC5223026.
109. Hahnen E, Schönling J, Rudnik-Schöneborn S, Zerres K, Wirth B. Hybrid survival motor neuron genes in patients with autosomal recessive spinal muscular atrophy: new insights into molecular mechanisms responsible for the disease. Am J Hum Genet. 1996;59(5):1057-65. PubMed PMID: 8900234; PubMed Central PMCID: PMCPMC1914839.
110. Dennison EM, Syddall HE, Rodriguez S, Voropanov A, Day IN, Cooper C. Polymorphism in the growth hormone gene, weight in infancy, and adult bone mass. J Clin Endocrinol Metab. 2004;89(10):4898-903. doi: 10.1210/jc.2004-0151. PubMed PMID: 15472182.
111. Ryan DP, da Silva MR, Soong TW, Fontaine B, Donaldson MR, Kung AW, et al. Mutations in potassium channel Kir2.6 cause susceptibility to thyrotoxic hypokalemic periodic paralysis. Cell. 2010;140(1):88-98. doi: 10.1016/j.cell.2009.12.024. PubMed PMID: 20074522; PubMed Central PMCID: PMCPMC2885139.
112. Basson CT, Bachinsky DR, Lin RC, Levi T, Elkins JA, Soults J, et al. Mutations in human TBX5 [corrected] cause limb and cardiac malformation in Holt-Oram syndrome. Nat Genet. 1997;15(1):30-5. doi: 10.1038/ng0197-30. PubMed PMID: 8988165.
113. Li QY, Newbury-Ecob RA, Terrett JA, Wilson DI, Curtis AR, Yi CH, et al. Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family. Nat Genet. 1997;15(1):21-9. doi: 10.1038/ng0197-21. PubMed PMID: 8988164.
114. Lemmers RJ, Tawil R, Petek LM, Balog J, Block GJ, Santen GW, et al. Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. Nat Genet. 2012;44(12):1370-4. Epub 20121111. doi: 10.1038/ng.2454. PubMed PMID: 23143600; PubMed Central PMCID: PMCPMC3671095.
115. Di Stazio M, Collesi C, Vozzi D, Liu W, Myers M, Morgan A, et al. TBL1Y: a new gene involved in syndromic hearing loss. European Journal of Human Genetics. 2019;27(3):466-74. doi: 10.1038/s41431-018-0282-4.

116. Yuan P, Zheng L, Liang H, Li Y, Zhao H, Li R, et al. A novel mutation in the TUBB8 gene is associated with complete cleavage failure in fertilized eggs. J Assist Reprod Genet. 2018;35(7):1349-56. Epub 20180427. doi: 10.1007/s10815-018-1188-3. PubMed PMID: 29704226; PubMed Central PMCID: PMCPMC6063842.

117. Yao Z, Zeng J, Zhu H, Zhao J, Wang X, Xia Q, et al. Mutation analysis of the TUBB8 gene in primary infertile women with oocyte maturation arrest. Journal of Ovarian Research. 2022;15(1):38. doi: 10.1186/s13048-022-00971-9.

118. Feng R, Sang Q, Kuang Y, Sun X, Yan Z, Zhang S, et al. Mutations in TUBB8 and Human Oocyte Meiotic Arrest. New England Journal of Medicine. 2016;374(3):223-32. doi: 10.1056/NEJMoa1510791. PubMed PMID: 26789871.

119. Xie C, Bekpen C, Künzel S, Keshavarz M, Krebs-Wheaton R, Skrabar N, et al. A de novo evolved gene in the house mouse regulates female pregnancy cycles. eLife. 2019;8:e44392. doi: 10.7554/eLife.44392.

120. Carroll SB. Homeotic genes and the evolution of arthropods and chordates. Nature. 1995;376(6540):479-85. doi: 10.1038/376479a0.

121. Baatz M, Wagner GP. Adaptive Inertia Caused by Hidden Pleiotropic Effects. Theoretical Population Biology. 1997;51(1):49-66. doi: https://doi.org/10.1006/tpbi.1997.1294.

122. Wright S. Evolution and the genetics of populations, volume 4: variability within and among natural populations: University of Chicago press; 1984.

123. Barton NH. Pleiotropic models of quantitative variation. Genetics. 1990;124(3):773-82.

124. Pyeritz RE. Pleiotropy revisited: Molecular explanations of a classic concept. American Journal of Medical Genetics. 1989;34(1):124-34. doi: https://doi.org/10.1002/ajmg.1320340120.

125. Tyler AL, Asselbergs FW, Williams SM, Moore JH. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. Bioessays. 2009;31(2):220-7.

126. Zhang J. Patterns and Evolutionary Consequences of Pleiotropy. Annual Review of Ecology, Evolution, and Systematics. 2023;54(1):1-19. doi: 10.1146/annurev-ecolsys-022323-083451.

127. Quiver MH, Lachance J. Adaptive eQTLs reveal the evolutionary impacts of pleiotropy and tissue-specificity while contributing to health and disease. HGG Adv. 2022;3(1):100083. Epub 20211224. doi: 10.1016/j.xhgg.2021.100083. PubMed PMID: 35047867; PubMed Central PMCID: PMCPMC8756519.

128. Wagner GP, Zhang J. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. Nature Reviews Genetics. 2011;12(3):204-13. doi: 10.1038/nrg2949.

129. Zeng ZB, Hill WG. The selection limit due to the conflict between truncation and stabilizing selection with mutation. Genetics. 1986;114(4):1313-28. doi: 10.1093/genetics/114.4.1313. PubMed PMID: 3803917; PubMed Central PMCID: PMCPMC1203042.

130. Orr HA. Adaptation and the cost of complexity. Evolution. 2000;54(1):13-20.

131. Carroll SB. Evolution at two levels: on genes and form. PLoS biology. 2005;3(7):e245.

132. Wray GA. The evolutionary significance of cis-regulatory mutations. Nature Reviews Genetics. 2007;8(3):206-16.

133. Zhu J, He F, Hu S, Yu J. On the nature of human housekeeping genes. Trends in Genetics. 2008;24(10):481-4. doi: https://doi.org/10.1016/j.tig.2008.08.004.

134. Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134(4):1289-303. doi: 10.1093/genetics/134.4.1289. PubMed PMID: 8375663; PubMed Central PMCID: PMCPMC1205596.

135. He X, Zhang J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics. 2005;169(2):1157-64.

136. Guillaume F, Otto SP. Gene Functional Trade-Offs and the Evolution of Pleiotropy. Genetics. 2012;192(4):1389-409. doi: 10.1534/genetics.112.143214.
137. Hoekstra HE, Coyne JA. The locus of evolution: evo devo and the genetics of adaptation. Evolution. 2007;61(5):995-1016. doi: 10.1111/j.1558-5646.2007.00105.x.
138. Des Marais DL, Rausher MD. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. Nature. 2008;454(7205):762-5.
139. Dezső Z, Nikolsky Y, Sviridov E, Shi W, Serebriyskaya T, Dosymbekov D, et al. A comprehensive functional analysis of tissue specificity of human gene expression. BMC biology. 2008;6(1):1-15.
140. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. Nucleic acids research. 2014;42(D1):D749-D55.
141. Neme R, Tautz D. Phylogenetic patterns of emergence of new genes support a model of frequent de novoevolution. BMC Genomics. 2013;14(1):117. doi: 10.1186/1471-2164-14-117.
142. Cortez D, Marin R, Toledo-Flores D, Froidevaux L, Liechti A, Waters PD, et al. Origins and functional evolution of Y chromosomes across mammals. Nature. 2014;508(7497):488-93. doi: 10.1038/nature13151.
143. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular biology and evolution. 2007;24(8):1586-91.
144. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: a resource for timelines, timetrees, and divergence times. Molecular biology and evolution. 2017;34(7):1812-9.
145. Lahn BT, Page DC. Four evolutionary strata on the human X chromosome. Science. 1999;286(5441):964-7.

Figure and legend

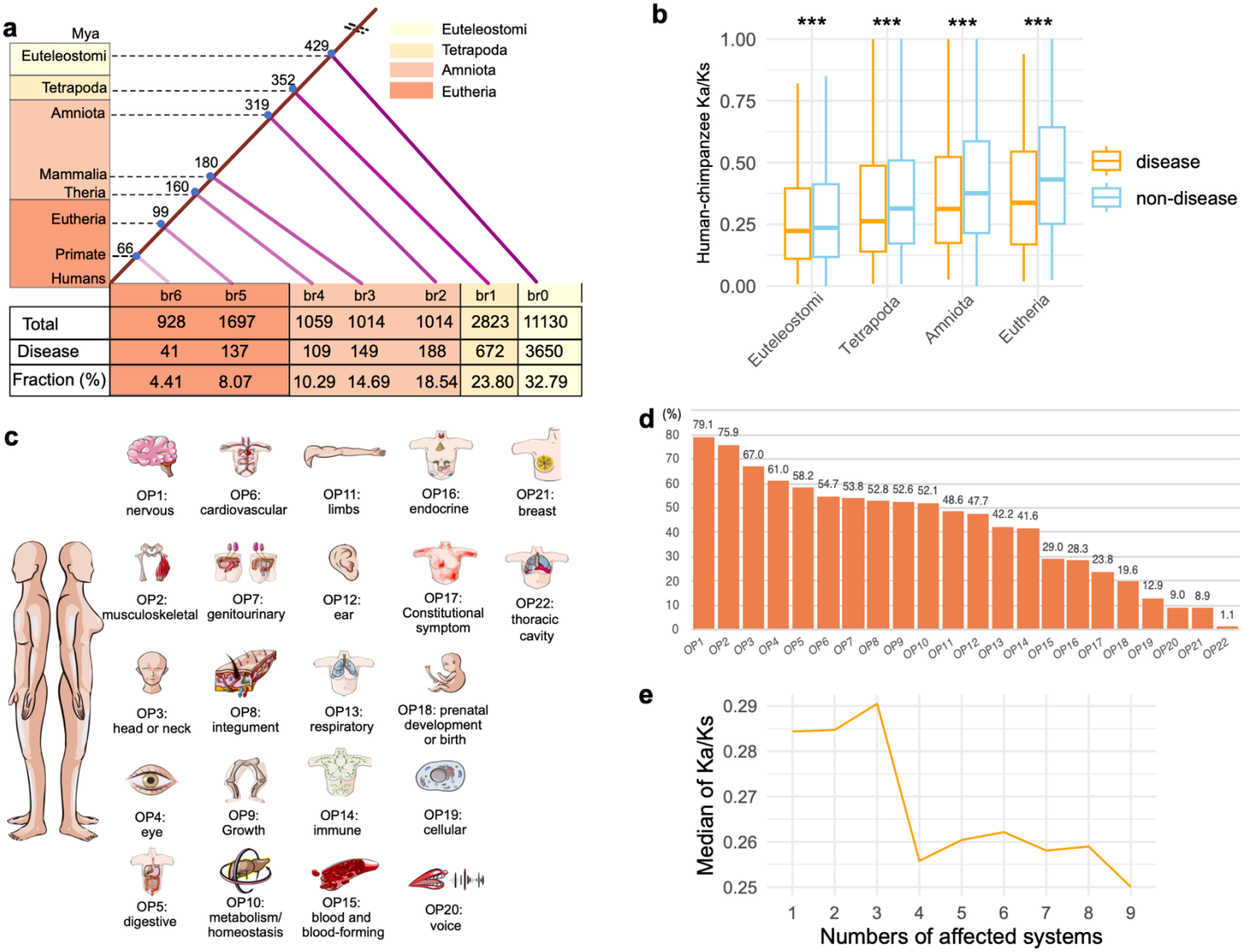


Figure 1. Number distribution and Ka/Ks ratios of genes categorized by ages and disease phenotypes (also organ phenotype genes). (**a**) The phylogenetic framework illustrating gene ages and disease genes associated with organ phenotypes. The phylogenetic branches represent age assignment for all genes and disease genes. The "br" values from br0 to br7 signify ancestral age groups (or branches). These are further categorized into four evolutionary age stages. The vertical axis depicts the divergence time sourced from the Timetree database (July 2023). The numbers of total genes and disease genes and their ratios are shown for each evolutionary age stage. (**b**) The pairwise Ka/Ks ratios from Ensembl database based on Maximum Likelihood estimation for “one to one” orthologs between human and chimpanzee. Only genes under purifying selection are shown (Ka/Ks < 1). The significance levels are determined using the Wilcoxon rank sum test, comparing disease genes to non-disease genes. The symbol “***” indicates significance level of $p < 0.001$. (**c**) The 22 HPO-defined organ/tissue systems, which are ordered based on the proportion of genes among all disease genes. (**d**) Percentages representing disease genes affecting various organs/tissues/systems in relation to the total number of disease genes.

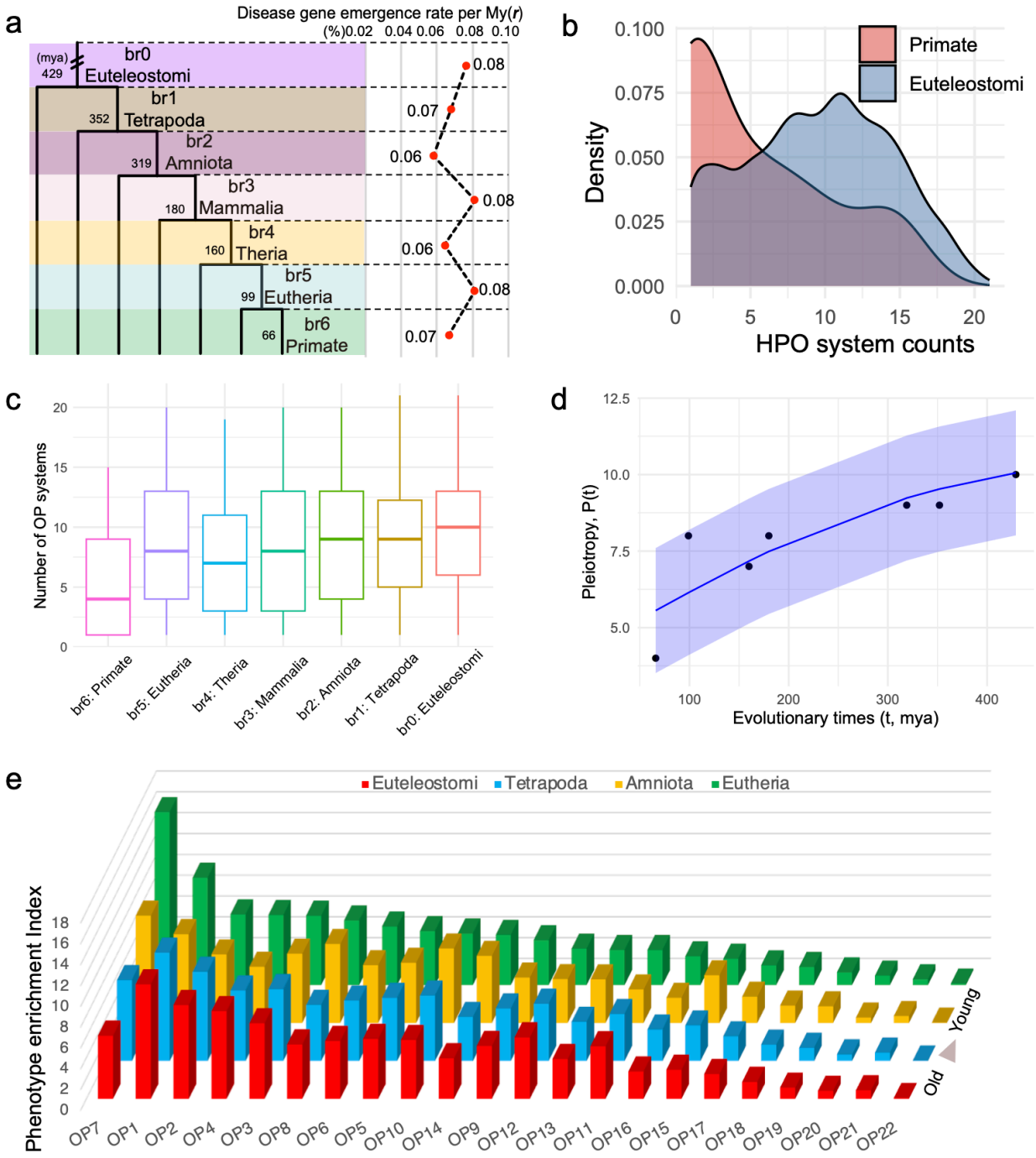


Figure 2. Disease gene emergence rates, phenotypic system coverage, and disease phenotype enrichment index (PEI) along evolutionary age groups. (**a**) The disease-gene emergence rate per million years (*r*) across evolutionary epochs. (**b**) Density distributions showcase the numbers of affected organ phenotypic systems (OPs) for genes originated at primate

and Euteleostomi stage. (**c**) Boxplot distributions showcase the numbers of affected organ phenotypic systems (OPs) for genes grouped by their evolutionary age (median values are 4, 8, 7, 8, 9, 9, 10, from left to right). (**d**) The nonlinear least squares (NLS) regression between pleiotropy score (P) and evolutionary times *t* with the logistic growth function (P(*t*) = $\frac{P_max}{1+\frac{P_max-P_0}{P_0}e^{-k(t)}}$, *k* = 1.66, *p* = 0.000787, 95% confidence interval is shown shade. P_max and *P_0* are empirical medians 10 and 4, respectively) (**e**) The distribution of age and phenotype for the phenotype enrichment index (PEI). The bar plots, colored differently, represent various age epochs, namely Euteleostomi, Tetrapoda, Amniota, and Eutheria, in ascending order of age. The organ phenotypes (OP) are displayed on the horizontal axis and defined in Figure 1c. The standard deviations of PEI are 3.67 for Eutherian epochs and approximately 2.79 for older epochs.

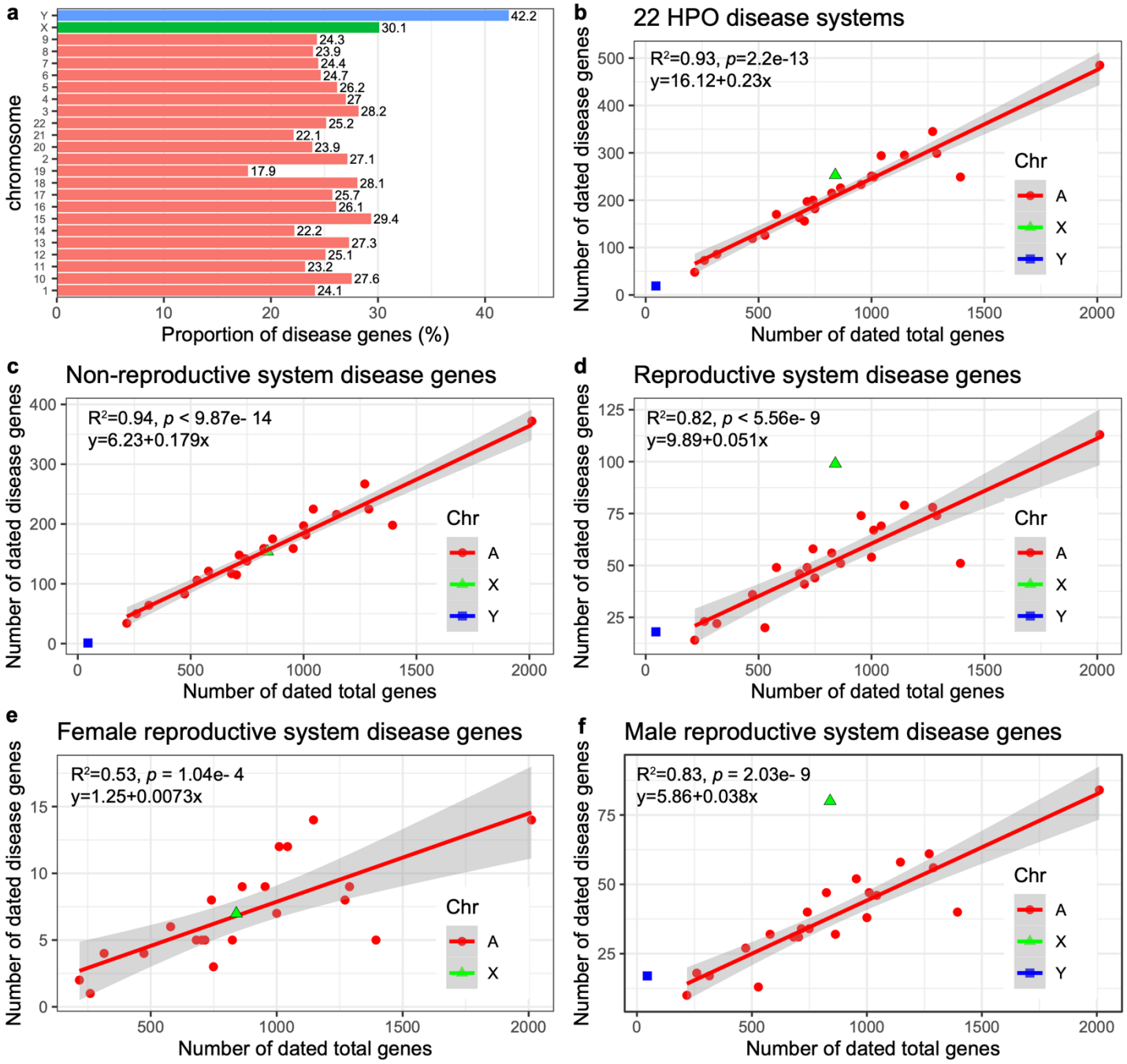


Figure 3. (**a**) The proportions of disease genes across chromosomes. The pink bars are the autosomes, while green and blue indicate the sex chromosome of X and Y respectively. The proportions (%) for different chromosomes are shown above bars. (**b**) The linear regression plotting of disease gene counts against the numbers of total genes with age information on chromosomes. (**c**) The number of genes related to the abnormality of non-genitourinary system (non-reproductive system) are plotted against all protein-coding genes on chromosomes with gene age information. (**d**) The number of genes related to the abnormality of genitourinary system (the reproductive system) are plotted against all protein-coding genes on chromosomes with

gene age information. (**e**) Linear regression of dated disease gene counts against the total numbers of dated gene on chromosomes for female-specific reproductive disease genes. (**f**) Linear regression of dated disease gene counts against the total numbers of dated gene on chromosomes for male-specific reproductive disease genes. The autosomal linear models are shown on the top left corner. Note: All linear regression formulas and statistics pertain only to autosomes. “A”, “X”, and “Y” indicate autosomes, X and Y chromosomes, respectively.

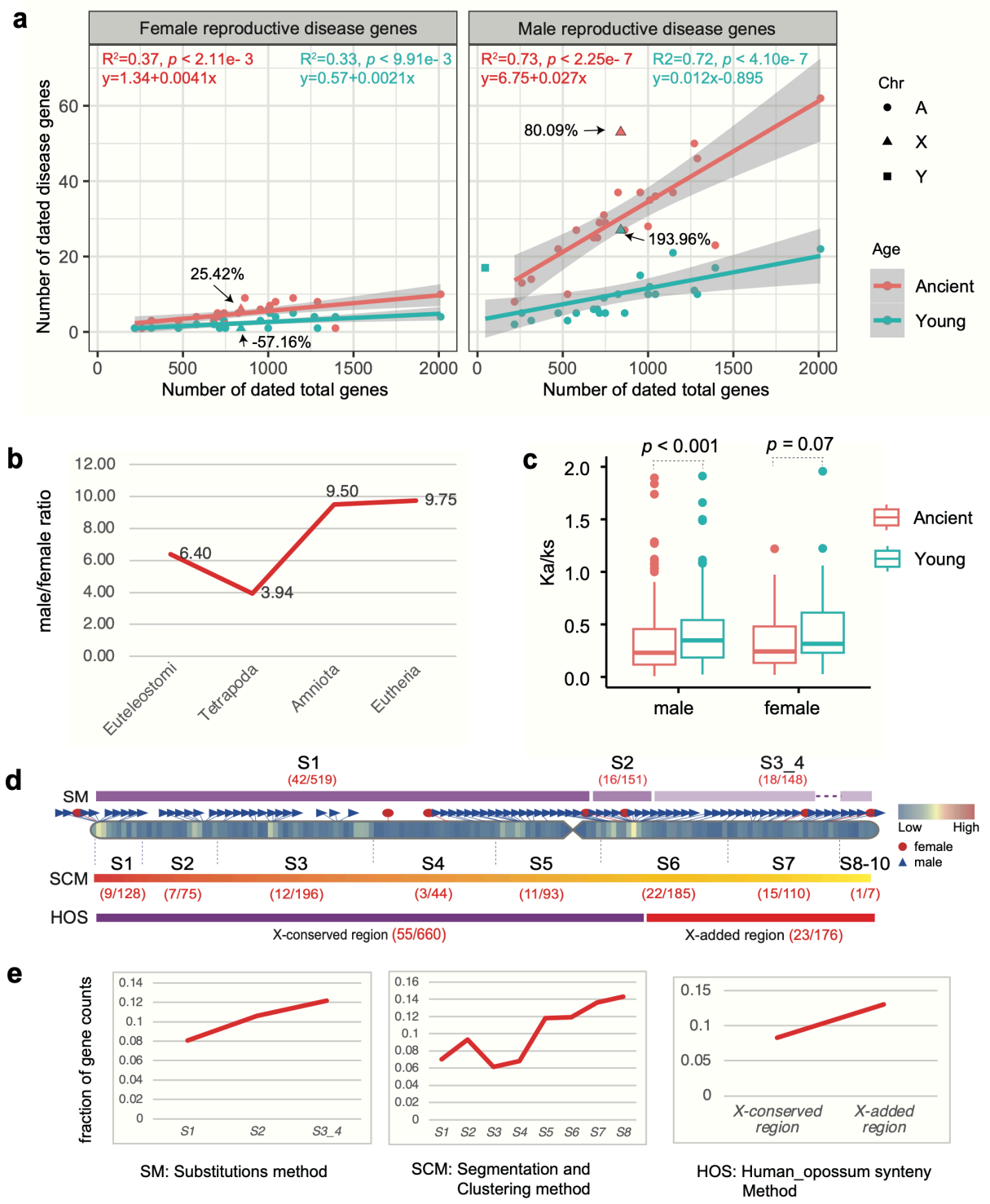


Figure 4. (**a**) The numbers of female-specific (left) and male-specific reproductive disease genes (right) are plotted against all protein-coding genes with gene ages on chromosomes. The linear formulas fitted for autosomal genes at ancient (Euteleostomi) and younger (post-Euteleostomi) stages are shown in red and blue, respectively. (**b**) The ratios of male to female reproductive disease gene numbers (α_d) across four evolutionary epochs. (**c**) The

comparison of selection pressure (human-chimpanzee pairwise Ka/Ks) for sex-specific reproductive disease genes between the ancient (stage Euteleostomi) and younger (post-Euteleostomi) epochs. Only the autosomal comparison is shown, with *p* value from the Wilcoxon test. (**d**) The numbers of male-specific reproductive disease genes (*m*) and the background genes (*b*) within the subregions from old to young on the X chromosome are provided, with the numbers displayed within round brackets for each subregion (*m*/*b*). SM, SCM, and HOS denote three classification methods for X chromosome structure: the substitutions method, the segmentation and clustering method, and the synteny method (orthologous gene order conservation between human and opossum). (**e**) The fraction of disease genes with male-specific reproductive disease phenotypes within each stratum or subregion, as illustrated in (**d**), is presented. The gene coordinates have been updated based on the hg38 reference with liftover tool. “A”, “X”, and “Y” indicate autosomes, X and Y chromosomes, respectively.

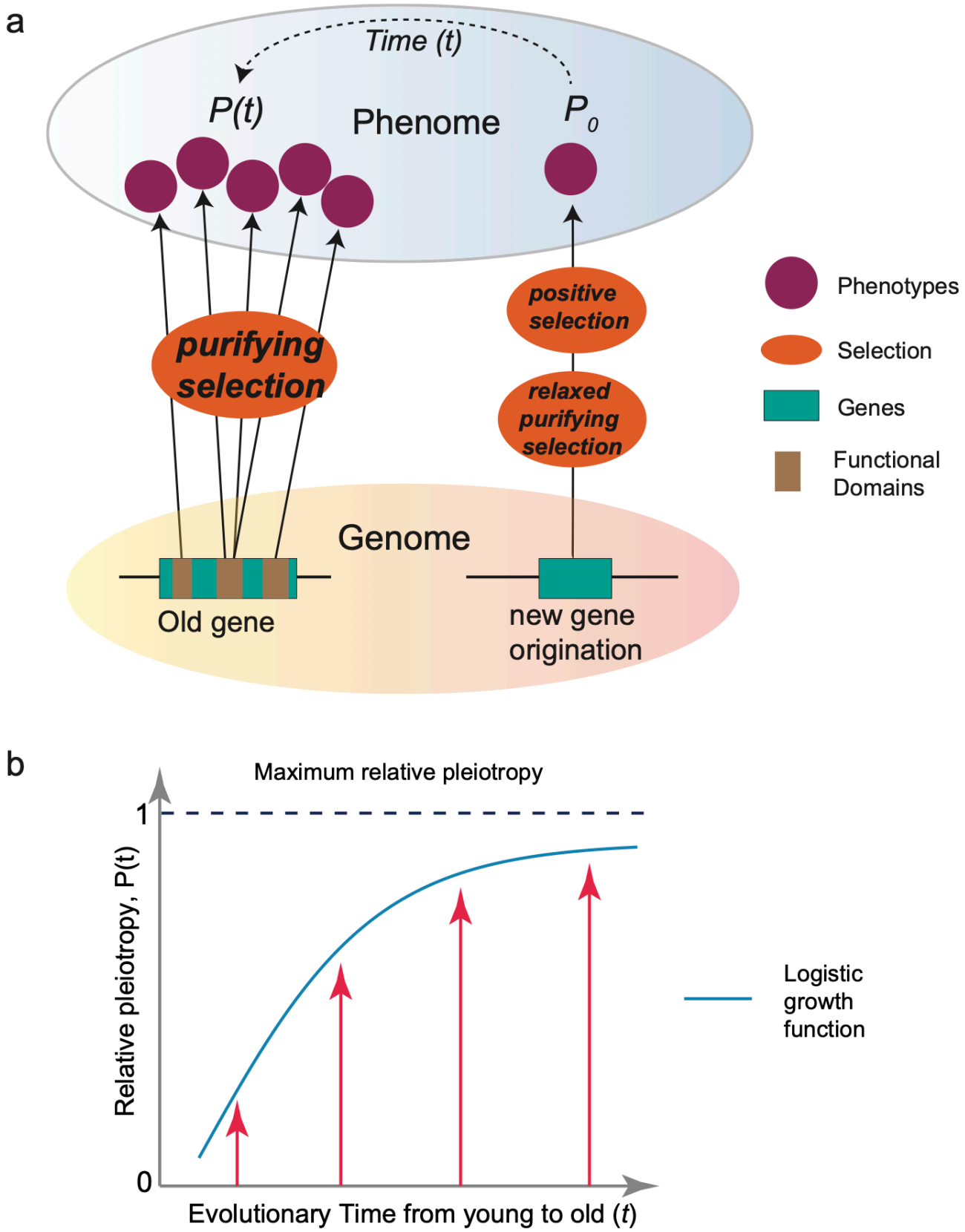
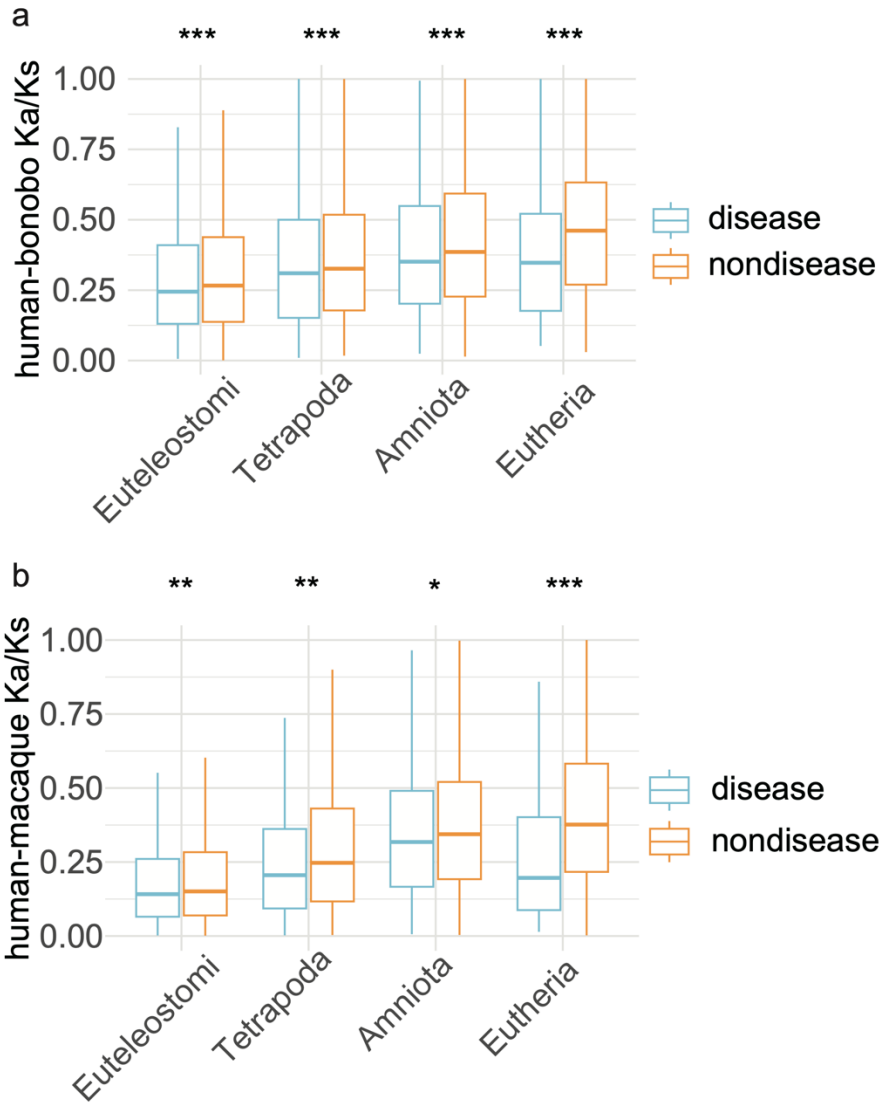
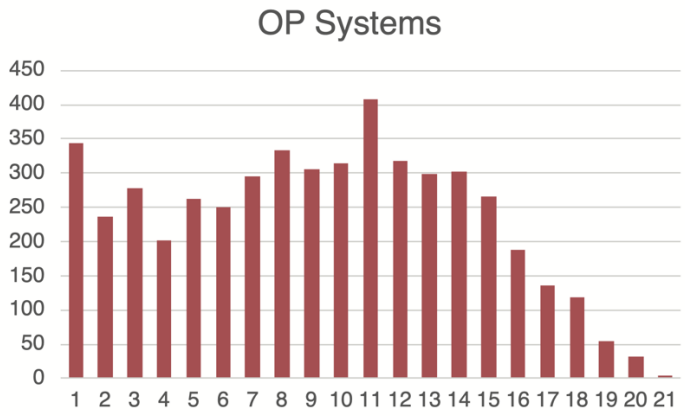

a
Time (t)
P(t)
Phenome
P0
positive selection
purifying selection
relaxed purifying selection
Phenotypes
Selection
Genes
Functional Domains
Genome
Old gene
new gene origination
b
Maximum relative pleiotropy
1
0
Relative pleiotropy, P(t)
Logistic growth function
Evolutionary Time from young to old (t)

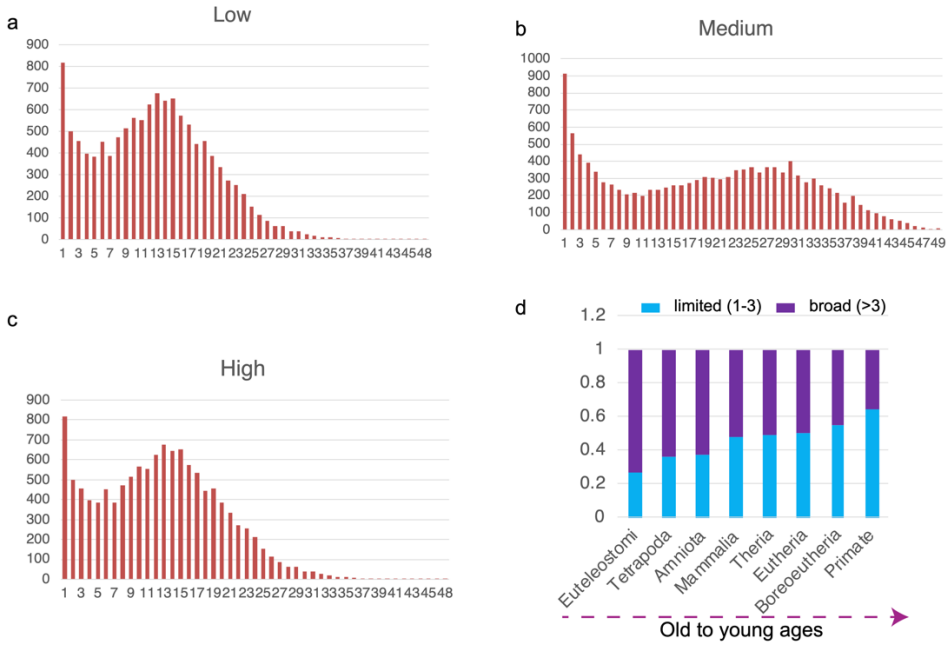
Figure 5. (**a**) The 'pleiotropy-barrier model' posits that new genes evolve adaptively more quickly. It suggests that older genes undergo stronger purifying selection because their multiple functions (usually adverse pleiotropy) act as a barrier to the uptake of mutations that might otherwise be beneficial for novel phenotypes. (**b**) The logistic function between relative pleiotropy P(*t*) and evolutionary time *t*, $P(t) = \frac{P_max}{1+e^{-kt}}$, where *P_max* represents the maximum relative pleiotropy. The *k* is the growth rate parameter, which controls how quickly the phenomenon approaches the maximum value. A higher k value means faster growth initially.



Supplementary Figure 1. The pairwise Ka/Ks ratios from Ensembl database based on the Maximum Likelihood estimation for “one to one” orthologs between human and bonobo (**a**) and between human and macaque (**b**). Only genes under purifying selection are visualized (Ka/Ks < 1). Note: the significance levels are based on the Wilcoxon rank sum test comparing disease genes and non-disease genes (one tail test). ”*”, ”**”, “***” indicate $p < 0.05$, < 0.01, < 0.001, respectively.



Supplementary Figure 2. The distribution of counts of OP system on which disease gene defects could affect. The horizontal axis is the numbers of OP systems, and the vertical axis is the number of genes.



Supplementary Figure 3. The distribution for the numbers of tissues with RNASEQ expression (HPA normal tissues). (**a**) The distribution of counts of normal tissues with low levels of gene expression (HPA annotation). (**b**) The distribution of counts of normal tissues with medium levels of gene expression (HPA annotation). (**c**) The distribution of counts of normal tissues with high levels of gene expression (HPA annotation). (**d**) The distribution percentages of genes expressed in multiple tissues (>3) compared to those expressed in a limited number of tissues (<=3) across different evolutionary ages (normal tissue with high gene expression estimated from HPA).