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

Jian-Hai Chen,[1,2] Patrick Landback,[1] Deanna Arsala,[1] Alexander Guzzetta,[3] Shengqian Xia,[1] Jared Atlas,[1,4] Dylan Sosa,[1] Yong E. Zhang,[5] Jingqiu Cheng,[2] Bairong Shen,[2] and Manyuan Long[1]

[1]Department of Ecology and Evolution, The University of Chicago, Chicago, Illinois 60637, USA; [2]Institutes for Systems Genetics, West China University Hospital, Chengdu 610041, China; [3]Department of Pathology, The University of Chicago, Chicago, Illinois 60637, USA; [4]Committee on Genetics, Genomics and Systems Biology, The University of Chicago, Chicago, Illinois 60637, USA; [5]Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China

New genes (or young genes) are genetic novelties pivotal in mammalian evolution. However, their phenotypic impacts and evolutionary patterns over time remain elusive in humans owing to the technical and ethical complexities of functional studies. Integrating gene age dating with Mendelian disease phenotyping, we reveal a gradual rise in disease gene proportion as gene age increases. Logistic regression modeling indicates that this increase in older genes may be related to their longer sequence lengths and higher burdens of deleterious de novo germline variants (DNVs). We also find a steady integration of new genes with biomedical phenotypes into the human genome over macroevolutionary timescales (~0.07% per million years). Despite this stable pace, we observe distinct patterns in phenotypic enrichment, pleiotropy, and selective pressures across gene ages. Young genes show significant enrichment in diseases related to the male reproductive system, indicating strong sexual selection. Young genes also exhibit disease-related functions potentially linked to human phenotypic innovations, such as increased brain size, musculoskeletal phenotypes, and color vision. We further reveal a logistic growth pattern of pleiotropy over evolutionary time, indicating a diminishing marginal growth of new functions for older genes owing to intensifying selective constraints over time. We propose a "pleiotropy-barrier" model that delineates higher potential for phenotypic innovation in young genes compared to older genes, a process under natural selection. Our study demonstrates that evolutionarily 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.

[Supplemental material is available for this article.]

The imperfection of DNA replication serves as a source of variations for evolution and biodiversity (Nei 2013). Such genetic variations underpin the ongoing evolution of human phenotypes, with beneficial mutations being fixed by 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 (Pavličev and Wagner 2022). Approximately 8000 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 (Antonarakis and Beckmann 2006; Fetro and Scherman 2020). To date, more than 4000 Mendelian disease genes have been identified, each contributing to a diverse array of human phenotypes (Boycott et al. 2013; https://mirror.omim.org/statistics/geneMap). These disease genes and associated phenotypes could provide insights into the evolutionary trajectory of human traits (Claussnitzer et al. 2020).

Evolutionarily new genes—such as de novo genes, chimeric genes, and gene duplicates—integrate into the human genome throughout microevolutionary processes (Brosius 1991; Kaessmann et al. 2002; Conrad and Antonarakis 2007; Baertsch et al. 2008; Wu et al. 2011; Long et al. 2013; Van Oss and Carvunis 2019; Betrán and Long 2022; Zhang et al. 2022). New genes can be integrated into essential bioprocesses, such as transcriptional regulation, RNA synthesis, and DNA repair (Ciccarelli et al. 2005; Ding et al. 2021). In *Drosophila* species, lineage-specific genes may control the key cytological process of mitosis (Ross et al. 2013). New genes have also been found with roles in early larval development of *Drosophila* (Kasinathan et al. 2020). In nematodes, insects, and fish, some lineage-specific genes are thought to be involved in morphological development, a process that was long believed to be governed by deeply conserved genetic mechanisms (Ragsdale et al. 2013; Klomp et al. 2015; Li et al. 2021). These studies reveal important biological functions of new genes.

Corresponding authors: jianhaichen@uchicago.edu, mlong@uchicago.edu, bairong.shen@scu.edu.cn

Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.279498.124. Freely available online through the *Genome Research* Open Access option.

Compared to nonhuman model organisms, in which gene functions can be characterized through genetic knockdowns and knockouts, investigating functions of human genes in their native context is impractical. Despite this limitation, accumulating omic data and in vitro studies of human genes have suggested the potential roles of evolutionarily young genes in basic cellular processes and complex phenotypic innovations (Marques et al. 2005; Shi and Su 2012; Zhang and Long 2014). Brain transcriptomic analysis has revealed that upregulated genes early in human development are enriched with primate-specific genes, particularly within the human-specific prefrontal cortex (Zhang et al. 2011). The recruitment of new genes into the transcriptome suggests that human anatomical novelties may evolve with the contribution of new gene evolution. These findings contradict the conventional conservation-dominant view of human genetics and phenotypes.

It has long been observed that there are more genetic disease genes among older genes than among young ones (Domazet-Lošo and Tautz 2008). However, the underlying mechanism remains unclear. In recent years, medical studies have identified deleterious variants causing rare disorders, "orphan" diseases, and rare forms of common diseases (Richards et al. 2015). Rare diseases are often caused by rare variants, which have greater effects than common variants (Richards et al. 2015; Wang et al. 2021; Chen et al. 2022, 2024a; Greene et al. 2023; Jia et al. 2023; Weiner et al. 2023). The effect of gene-based rare variant burden, the aggregate impact of rare (including de novo germline) genetic variants in cohorts, has been confirmed in many genetic disorders (Purcell et al. 2014; Zuk et al. 2014; Guo et al. 2018; Halvorsen et al. 2020; Jiang et al. 2022).

In this study, we analyzed the anatomical organ/tissue/system phenotypes (OPs) of human genetic diseases to understand questions about gene ages, phenotypic enrichment, pleiotropy, and selective constraints. First, are older genes more likely to be disease genes, and if so, why? Second, does the rate of disease gene emergence (per million years) differ across macroevolutionary history? Third, do young genes show a phenotypic preference for certain disease systems? Fourth, do young genes exhibit a different pattern of pleiotropic effects compared with older genes? Finally, are these differences driven by selection?

Results

Young genes have lower fractions of disease genes with OPs than do older genes

We determined the evolutionary ages (phylostrata) for 19,665 genes shared between the GenTree database (Shao et al. 2019) and Ensembl (v110) (Supplemental Table S1). These genes were then categorized into two types, "disease genes" or "nondisease genes," based on disease annotations for a total of 5006 genes from the Human Phenotype Ontology database (HPO; September 2023), which is the de facto standard for phenotyping rare Mendelian diseases (Köhler et al. 2019). Among these disease genes, 60 genes lacked gene age information. Thus, an intersection of data sets yielded 4946 genes annotated with both evolutionary age and OP abnormalities (Fig. 1A,B; Supplemental Table S2). To ensure sufficient statistical power in comparisons, we merged evolutionary age groups with a small number of genes (fewer than 100) with their adjacent older group (Fig. 1A). As a result, we reclassified these genes into seven ancestral age groups, ranging from Euteleostomi (and more ancient) nodes to modern humans (br0–br6) (Fig. 1A). We observed an increase in the proportion of disease genes over evolutionary time (Supplemental Fig. S1; Fig. 1A,C), suggesting that gene age impacts disease susceptibility, a trend qualitatively consistent with earlier studies (Domazet-Lošo and Tautz 2008).

The likelihood of being disease genes positively correlates with gene age, protein length, and DNV burden

To quantify factors contributing to the likelihood of a gene being classified as a disease gene, we assigned binary states to all genes ("1," disease genes; "0," nondisease genes) and performed stratified logistic regression modeling (Supplemental Table S4). We explored multiple predictors, including gene age (T; mya), gene length (L_g) or protein length (L), gene-wise burden of deleterious DNVs (D) from 46,612 trios (Wang et al. 2022), and rare variant burden (R) based on gnomAD genomes (Supplemental Table S3; Chen et al. 2024b). Sequence lengths were considered because, assuming a random mutation process, longer sequences would be expected to accumulate more deleterious variants in a cohort. Additionally, metrics of rare variant burdens were considered because Mendelian disorders and rare forms of common diseases are predominantly influenced by rare variants owing to their significant phenotypic effects (Gibson 2012; Guo et al. 2018; Kingdom et al. 2024). The burden of rare variants, which is often based on a collapsed information of rare loss-of-function or deleterious variants within causal genes at the cohort level, has been actively utilized in studying rare diseases (Lee et al. 2014).

Here, the burden scores of rare variants were based on two data sets: gene-wise burden of predicted deleterious DNVs from 46,612 trios (Wang et al. 2022) and gene-wise burden of rare variants from 76,215 genomes in the gnomAD database (v4.0.0, minor allele frequency [MAF] < 0.0001) (Chen et al. 2024b). Model comparison was conducted based on likelihood ratio test (LRT) and Akaike information criterion (AIC) (Supplemental Table S4). By comparing models of different variable combinations, we identified an optimal model, M9, which supports significant effects of three variables (DNV burden, protein length, and gene age) and the interaction between protein length and DNV burden (Chi-square test, $P = 4.59 \times 10^{-66}$) (for details, see Supplemental Table S4). The likelihood of being a disease gene increases with four terms: DNV burden (D; coefficient = 0.29; $P < 2 \times 10^{-16}$), protein length (L on the logarithmic scale; coefficient = 0.22; $P = 0.00031$), gene age (T; coefficient = 0.0041; $P < 2 \times 10^{-16}$), and the interaction between DNV burden and the logarithmic protein length (coefficient = −0.027; $P = 1.27 \times 10^{-9}$) (Supplemental Table S4 and Supplemental Fig. S2). The logarithmic scale effect of protein length on the likelihood suggests that normalizing extreme lengths could improve model fit. The negative coefficient for the interaction term between DNV burden and log-transformed protein length indicates an underlying trade-off between the two factors, suggesting that the impact of mutation burden on the likelihood of a gene being classified as a disease gene decreases as the protein length increases (Supplemental Fig. S4). In other words, although de novo mutation burden generally increases the probability of a gene being a disease gene, this effect is moderated by protein sequence length, with longer proteins showing a diminished impact of mutation burden. The values of variance inflation factor (VIF) for the variables range from 1.03 to 1.55, well below the multicollinearity concern thresholds of five to 10, suggesting minimal impact of multicollinearity among these variables on our model.

Consistent with the implications of this model, we found that disease genes tend to be significantly longer than nondisease genes in most branches (Wilcoxon rank-sum test, $P < 0.05$),

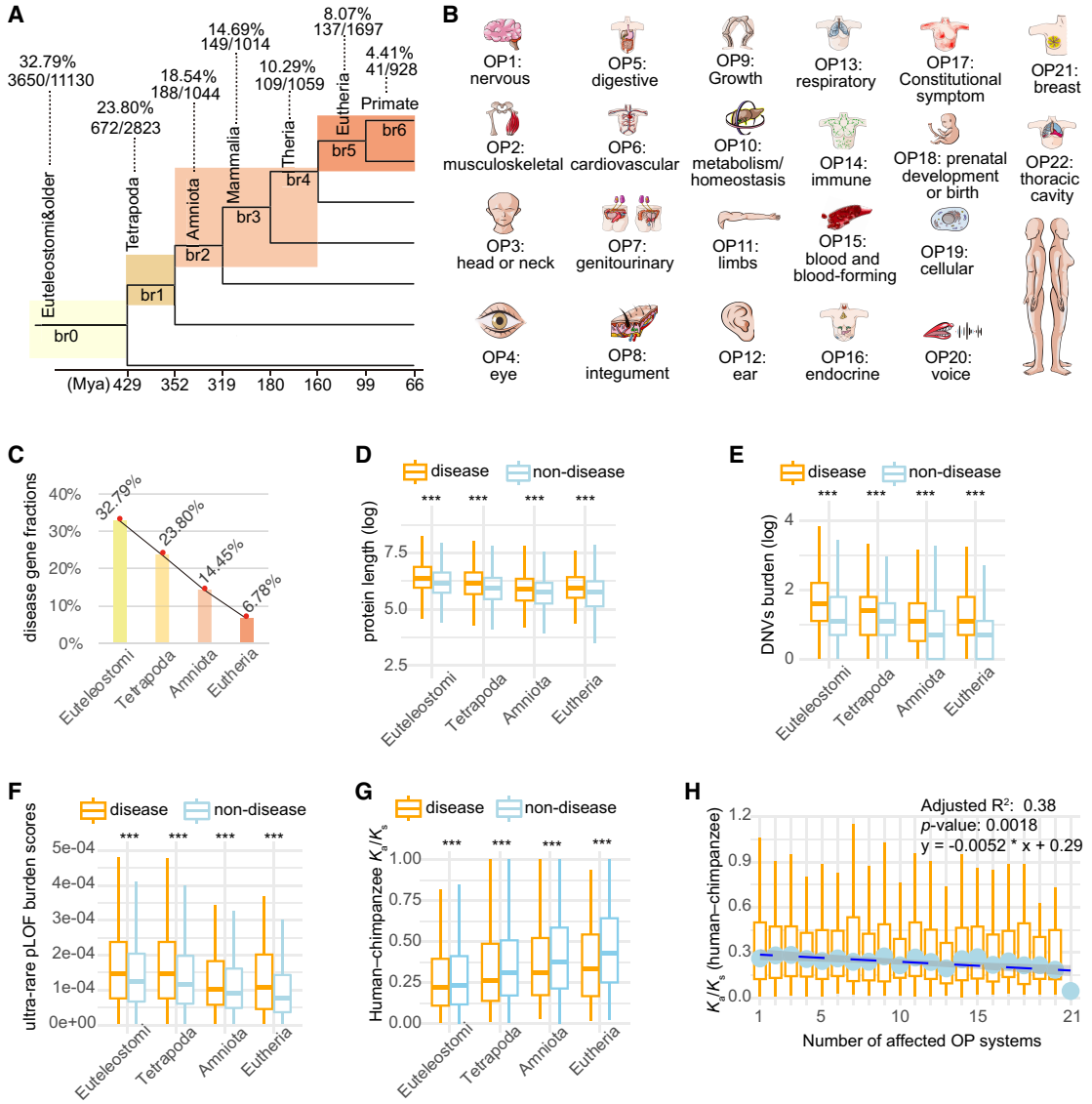

Figure 1. Number of genes, burdens of deleterious rare variants, and K_a/K_s ratios for genes categorized by gene ages (phylostrata) and disease systems (OPs). (*A*) The phylogenetic framework illustrating phylostrata and disease genes associated with disease systems (OPs). The phylogenetic branches represent phylostratum assignment for all genes and disease genes. The "br" values from br0 to br6 indicate different age groups (or branches). These are further categorized into four phylostrata to resolve the small number problem in br6 for some analyses. The horizontal 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 phylostratum. (*B*) The 22 HPO-defined OPs, which are ordered based on the fraction of disease genes in a certain system out of all disease genes. (*C*) The fractions of disease genes at four major phylostrata from Euteleostomi to Eutheria. (*D*) The protein lengths between disease and nondisease genes across four phylostrata. (*E*) The gene-wise burden based on de novo germline variants from the Gene4Denovo database (Zhao et al. 2020) between disease and nondisease genes across four phylostrata. (*F*) The burden score of ultrarare pLOF variants (Weiner et al. 2023) between disease and nondisease genes across four phylostrata. (*G*) The pairwise K_a/K_s ratios from the Ensembl database based on maximum likelihood estimation for "one-to-one" orthologs between human and chimpanzee. Only genes under purifying selection are shown ($K_a/K_s < 1$, 3226 genes). (*H*) The box plot for the number of affected systems for disease genes and their pairwise K_a/K_s ratios for "one-to-one" orthologs between human and chimpanzee (K_a/K_s ratios for all 3369 genes). The linear regression is based on median values and the number of affected tissues, with the statistical details and formula displayed in the *upper right* corner. Note that all significance levels of comparisons between disease genes and nondisease genes are determined using the Wilcoxon rank-sum test. (***) $P < 0.001$.

except for br4 and br6 (Supplemental Fig. S3A). The nonsignificant difference in these two branches may be because of the limited power of comparison caused by the lower sample sizes. We further compared between disease and nondisease genes, using the reported burdens of DNVs (Zhao et al. 2020) and ultrarare predicted loss-of-function variants (pLOF; MAF $< 1 \times 10^{-5}$) from 394,783 exomes (Supplemental Fig. S3B,C; Weiner et al. 2023). For pLOF comparison, we found significant lower burdens of deleterious variants for all branches (Wilcoxon rank-sum test, $P < 0.05$). For DNVs comparison, only the primate-specific branch (br6) was not significant, probably owing to small sample size. To increase the statistical power of comparison between disease and nondisease genes across different phylostrata, we further collapsed the number of phylostrata to four, starting from the oldest, Euteleostomi, to the progressively younger stages of Tetrapoda, Amniota, and Eutheria (Fig. 1A). Based on these four phylostrata,

we confirmed that disease genes tend to have significantly longer lengths and higher burdens of deleterious variants than do nondisease genes across all major phylostrata (Wilcoxon rank-sum test, $P < 2.2 \times 10^{-16}$) (Fig. 1D–F).

Young genes have lower burdens of DNVs and rare variants than older genes

We further examined the statistical correlation between gene age and burden of potentially deleterious variants. We used data sets from predicted deleterious de novo germline variations (DNVs) (Wang et al. 2022) and genome-wide rare variants from the gnomAD database (Chen et al. 2024b). We found that gene age (mya) positively correlates with DNV burden (Spearman's rank correlation $\rho = 0.181$, $P < 2.2 \times 10^{-16}$) and rare variant burden (Spearman's rank correlation $\rho = 0.306$, $P < 2.2 \times 10^{-16}$). When using different methods for human gene age dating, gene family–based (Neme and Tautz 2013) and synteny-based method (Shao et al. 2019), alongside different data sets on DNV burdens (Zhao et al. 2020), we also found significant correlations between gene age and DNV burden (Supplemental Fig. S4A,B).

New genes could originate from both "copying" and "noncopying" mechanisms (Vonica et al. 2020; Luria et al. 2023; Fleck et al. 2024), with the former from the processes of gene duplication and gene traffic (Emerson et al. 2004) and the latter occurring from exclusively sequence evolution (Domazet-Lošo et al. 2007). Thus, genetic novelty of new genes could be either "novelty-by-synteny" or "novelty-by-similarity." If the novelty resides in the protein sequence itself, then new genes by synteny should have burdens equivalent to the ancient genes that they were copied from. Considering that retrogenes are known genes with "novelty-by-synteny" owing to their random insertions into different genomic regions, we chose retrogenes and their parental genes to test this possibility. Retrogenes and their parental genes could be easily detected owing to their clear structural changes, such as intron loss during retro-transposition and insertion hallmarks (Long and Langley 1993; Miller et al. 2022). We retrieved known pairs of retrogenes and their parental genes from the GenTree database (Supplemental Table S5). We found that over half of the retrogenes have lower burdens of DNVs and burden scores of pLOF variants compared with their parental genes. In contrast, <36% of retrogenes showed higher burdens than their parental genes. Thus, retrogenes are different from their parental genes in terms of deleterious variant burden, supporting that new genes by synteny could lead to functional novelty. Moreover, we compared the primate-specific genes identified only from the synteny-based method and only from the similarity-based method, which are also young genes by synteny and those by similarity, respectively. We found significantly higher burdens of DNVs and pLOF in young genes by synteny than those by similarity (Wilcoxon rank-sum test, $P < 0.001$). Together, our results suggest that genes with "novelty-by-synteny" are more likely to have disease functions than those with "novelty-by-similarity" but are less likely than their parental genes. These results are consistent with our Supplemental Figure S4, A and B, which indicate young genes tend to have lower burden of deleterious DNVs than older genes based on a different data set of DNV burden (Zhao et al. 2020).

Purifying selection intensifies with gene age and is stronger in disease genes than in nondisease genes

To understand if disease genes evolve under different evolutionary pressures compared with nondisease genes, we compared the K_a/K_s ratio, which is the ratio of the number of nonsynonymous substitutions per nonsynonymous site (K_a) to the number of synonymous substitutions per synonymous site (K_s). Values of K_a/K_s ratios less than one suggest a degree of evolutionary constraint (acting against change) (Yang and Bielawski 2000). To ensure similar evolutionary backgrounds, we retrieved the "one-to-one" human–chimpanzee orthologous genes and the corresponding pairwise K_a/K_s ratios (12,830 genes) from the Ensembl database (Supplemental Table S6). We also evaluated whether the pattern is consistent with K_a/K_s ratios of human–bonobo and human–macaque orthologs (Supplemental Table S6). K_a/K_s ratios were consistently lower in disease genes than in nondisease 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, $P < 2.2 \times 10^{-16}$ for all three data sets). These results revealed that disease genes are under significantly stronger purifying selection than nondisease genes, suggesting an important component of selective pressure in constraining the sequence evolution of disease genes. We observed that K_a/K_s ratios (less than one) increase inversely proportional to gene age, suggesting a trend of relaxed purifying selection on young genes (Fig. 1G; Supplemental Fig. S5), which is consistent with some previous studies (Carvunis et al. 2012; Prabh et al. 2018; Montañés et al. 2023). Across the seven gene age groups, disease genes showed significantly lower K_a/K_s than nondisease genes in five out of seven groups (Supplemental Fig. S5). The nonsignificant difference in the primate-specific branch (br6) and the therian-specific branch (br4) could be because of the lower number of disease genes with K_a/K_s ratios (13 genes in br6 and 56 genes in br4). Notably, despite the relaxation of purifying selection for younger genes, disease genes still tend to show lower K_a/K_s ratios than nondisease genes for gene ages of four phylostrata, suggesting a pattern of stronger purifying selection in disease genes (Fig. 1G; Supplemental Fig. S5).

We observed a heterogeneous distribution of disease genes underlying 22 HPO-defined anatomical systems, suggesting varied genetic complexity for diseases of different systems (Supplemental Fig. S6A). None of the disease genes were 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) (Supplemental Table S2). The nervous system displayed the highest fraction of disease genes (79%) (Supplemental Fig. S6B). A large proportion of disease genes (93.04%) were linked to abnormalities of at least two systems (4602/4946), indicating that most human disease genes may have broad phenotypic impacts or pleiotropy across multiple anatomical systems. These phenotypic effects across systems might arise from the complex clinical symptoms of rare diseases manifesting in multiple organs, tissues, or systems, indicating considerable levels of pleiotropy (Hodgkin 1998; Paul 2000; Lobo 2008). Compared with commonly used functional inferences based on human gene expression profiles or in vitro screening, the comprehensive and deep phenotyping offered by HPO provides a more systematic perspective on the functional roles of human disease genes. We found a moderate but statistically significant inverse correlation between the median K_a/K_s ratios and the numbers of affected anatomical systems in disease genes (linear correlation adjusted $R^2 = 0.38$, $P = 0.0018$) (Fig. 1H). This implies that disease genes with higher pleiotropy, which impact multiple

anatomical systems, face stronger evolutionary constraints compared with genes with lower pleiotropy (Fig. 1H).

Disease gene emergence rate per million years is similar across macroevolutionary phylostrata

To understand whether different phylostrata have different emergence rates for disease genes, we assessed the disease gene emergence rate per million years across phylostrata from Euteleostomi to Primate (μ). Considering the sampling space variations at different age groups, we calculated μ as the fraction of disease genes per million years at each phylostratum (Fig. 2A). Although the proportions of disease genes were found to gradually increase from young to old phylostrata (Fig. 1A), the rate μ is nearly constant at ~0.07% per million years for different phylostrata (Fig. 2A). This constant emergence rate of disease genes suggests a continuous and similar fraction of genes evolving to have significant impacts on health during evolutionary history.

Using the recently reported average human generation time of 26.9 years (Wang et al. 2023) and the most updated number of coding genes (19,831 based on Ensembl v110) and assuming a simplified monogenic model (Richards et al. 2015), we estimated the number of causal genes for rare diseases per individual per generation (μ_d) as 3.73×10^{-4} (=$19{,}831 \times 26.9 \times 0.07 \times 10^{-8}$). Using this rate, we can derive the rare disease prevalence rate ($r_{RD} = 10{,}000 \times \mu_d$), which equates to approximately four in 10,000 individuals. This prevalence agrees well with the European Union definition of rare disease rate prevalence of five in 10,000 people (Stolk et al. 2006). This constant emergence rate highlights the idea that young genes continually acquire functions vital for human health and phenotypic innovations (Kaessmann 2010; Chen et al. 2013; Xia et al. 2025).

Pleiotropy growth rate is faster in younger genes following a logistic growth pattern

Despite the nearly constant integration of young genes into crucial biological functions (Fig. 2A), it remains uncertain if gene age could influence disease phenotypic spectra (or pleiotropy). The overall distribution of OP counts for disease genes (Supplemental Fig. S6A) is similar to that of gene expression breadth across tissues (Supplemental Fig. S7A–C). The distribution for OP counts showed that young genes have lower peak and median values than older genes (Fig. 2A–C). This pattern is consistent with the results that younger genes tend to be expressed in a limited range of tissues, whereas older genes exhibit a broader expression profile (Supplemental Fig. S7D), which also aligns with reported expression profiles (Zhang et al. 2012; Long et al. 2013; Carelli et al. 2016; Miller et al. 2022). We found an increasing trend for median OP numbers from young to old phylostrata (Fig. 2C). The increase rates $\left(\frac{\Delta \text{OP\_median}}{\Delta t}\right)$ were higher among the younger phylostrata than the older ones (0.12/mya at the Eutheria vs. 0.05/mya at older phylostrata on average) (Supplemental Table S7), suggesting a nonlinear and restricted growth model for the level of pleiotropy over time. We applied a logistic growth function and observed a significant pattern: As evolutionary time increases, the level of pleiotropy rises ($P < 0.001$) (Fig. 2D). Moreover, the logistic model demonstrates a diminishing marginal growth for pleiotropy over time, indicating that the rate of increase in pleiotropy slows down over

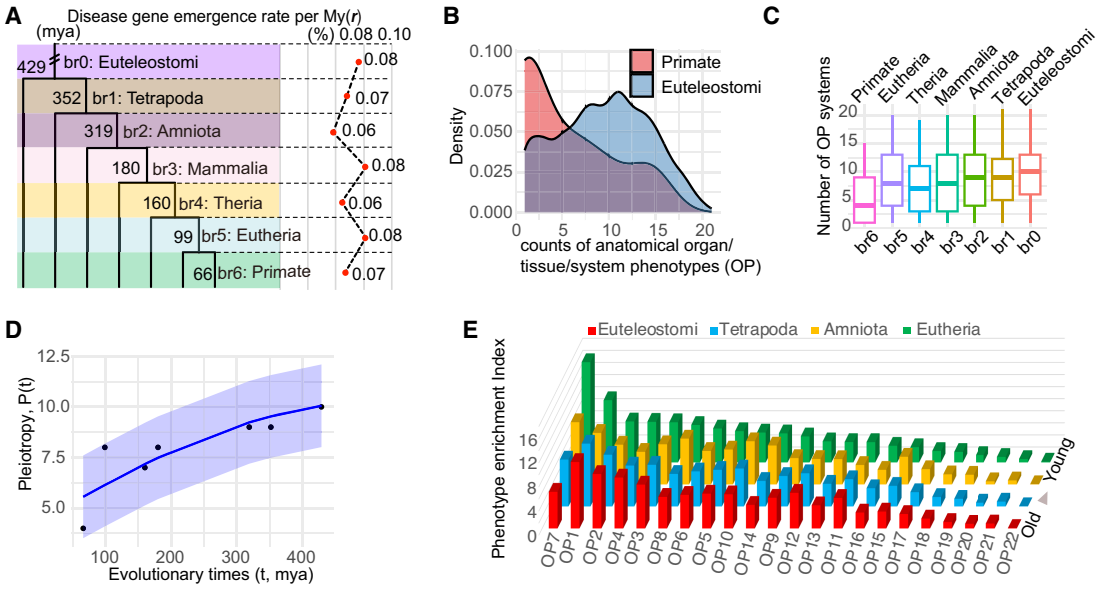

Figure 2. Disease gene emergence rates along phylostrata, OP counts comparison between the youngest and oldest phylostrata, and disease PEI along phylostrata. (*A*) The disease–gene emergence rate per million years (*r*) along phylostrata. (*B*) Density distributions showcase numbers of affected organ phenotypic (OP) systems for genes originated at primate and Euteleostomi phylostrata. (*C*) Box plot distributions showcase the numbers of affected OP systems for genes grouped by their phylostrata (median values are four, eight, seven, eight, nine, nine, 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: $P(t) = \frac{P\_\max}{1 + \frac{P_{\max} - P\_0}{P\_0} e^{-k(t)}}$, $k = 1.66$, $P = 0.000787$. The 95% confidence interval is shown shade. *P*_max and *P*_0 are empirical medians 10 and four, respectively. (*E*) The distribution of phylostrata and OP for PEI. The bar plots, colored differently, represent phylostrata, namely, Euteleostomi, Tetrapoda, Amniota, and Eutheria, in ascending order of evolutionary ages. The disease systems (OPs) are displayed on the horizontal axis and defined in Figure 1C. The SDs of PEI are 3.67 for Eutheria and approximately 2.79 for older phylostrata.

evolutionary time. This pattern suggests that although pleiotropy is initially lower in new genes, it increases more rapidly compared with in older genes. This result is consistent with the finding that purifying selection gradually increases over evolutionary time (Fig. 1G), limiting the space of pleiotropy growth in older genes.

Young genes are highly enriched in reproductive and nervous system diseases

We found a significant positive correlation between tissue expression breadth and the number of affected disease systems (Spearman's rank correlation, $\rho = 0.138$, $P < 2.2 \times 10^{-16}$) (Supplemental Table S8). To understand the enrichment pattern of disease phenotypes for young and old genes, we introduced a metric of the disease phenotype enrichment index (PEI), which quantifies the range of phenotypes across multiple systems (for details, see Methods). Our findings revealed that the most ancient genes, specifically from Euteleostomi and Tetrapoda, had the strongest PEI association with the nervous system (OP1). Conversely, young genes from Amniota and Eutheria exhibit the highest PEI for disease phenotypes of the genitourinary system (OP7) and the nervous system (OP1), in which the genitourinary system (OP7) shows a 38.65% higher PEI than the nervous system (OP1) (Fig. 2E; Supplemental Table S9). Among the 22 disease phenotype systems, only the reproductive system (OP7) showed a steady rise in PEI from older phylostrata to younger ones (Fig. 2E). There are smaller variations in PEI for older phylostrata compared with the more recent Eutheria (about 2.79 vs. 3.67), suggesting that older disease genes impact a greater number of organ systems, as shown in Figure 2C. This finding is consistent with the "out-of-testis" hypothesis (Kaessmann 2010) that the expression patterns of young genes are biased to the testes and may have vital roles in male reproduction. As genes evolve, their expression patterns tend to broaden, potentially leading to phenotypic effects impacting 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 (Fig. 2E). Moreover, 42% of the 19 Primate-specific disease genes with diseases affecting the nervous system (OP1) correlated with phenotypes involving brain size or intellectual development (*CFC1*, *DDX11*, *H4C5*, *NOTCH2NLC*, *NOTCH2NLA*, *NPAP1*, *RRP7A*, and *SMPD4*) (Supplemental Table S2; Supplemental Material), consistent with the expectation of previous studies based on gene expression (Zhang et al. 2011). Furthermore, the primate-specific disease genes show phenotypic enrichment in other adaptive systems, particularly in the HPO systems of the head, neck, eyes, and musculoskeletal structure (Fig. 2E). In summary, the primate-specific disease genes could impact phenotypes from both reproductive and nonreproductive systems, particularly the genitourinary, nervous, and musculoskeletal systems (Supplemental Table S2), supporting their roles in both sexual and adaptive evolution.

Sex chromosomes are enriched for male-reproductive disease genes: the "male X-hemizygosity" effect

Considering the concentration of the youngest disease genes in the reproductive system (Fig. 2E, OP7), we hypothesized that the distribution of disease genes could be skewed across chromosomes. First, we examined the distribution of all disease genes and found a distinct, uneven spread across chromosomes (Fig. 3A; Supplemental Table S10). The X and Y Chromosomes contain higher fractions of disease genes compared with autosomes. Although autosomes have a linear slope of 0.23 ($R^2 = 0.93$; $P = 2.2 \times 10^{-13}$) (Fig. 3B), the proportion of Y Chromosomal disease genes is 82.61% higher, at 0.42. Meanwhile, the proportion of X Chromosomal disease genes is 30.43% higher than that of autosomes, sitting at 0.301. To understand whether the differences between sex chromosomes and autosomes are related to the reproductive functions, we divided disease genes into reproductive (1285 genes) and nonreproductive (3661 genes) categories based on affected organs (Supplemental Table S11). By fitting the number of disease genes against all genes with gene age information, we observed that the X Chromosome has a bias toward reproductive functions. Specifically, on the X Chromosome, disease genes affecting nonreproductive systems were slightly fewer than expected (−1.65% excess rate, with 154 observed vs. 156.59 expected) (Fig. 3C). The X Chromosome displayed a significant surplus of reproductive-related disease genes (observed number 99, expected number 52.73, excess rate 87.75%, $P < 5.56 \times 10^{-9}$) (Fig. 3D).

Given the sex-imbalanced mode of inheritance for the X Chromosome, a theoretical model has predicted that purifying selection would remove both dominant female-detrimental mutations and recessive male-detrimental mutations (Rice 1984; Charlesworth et al. 1987). We determined that the ratio of male to female reproductive disease genes ($M_{disease}/F_{disease}$, or α) 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 representation of disease genes from the male hemizygous X Chromosome compared with the female homozygous X, consistent with a recent medical study (Chen et al. 2024a). Thus, our analysis indicates that the abundance of disease genes on the X Chromosome compared with autosomes is likely owing to male-specific functional effects. These results suggest that the overrepresentation of disease genes on the X Chromosome primarily results from recessive X-linked inheritance affecting males rather than dominant effects impacting both sexes.

Genome-wide excess of male reproductive disease genes: the "faster-X" and "faster-male" effects

To determine which sex (male or female) might influence the biased distribution of reproductive-related genes on different chromosomes, we focused on genes specific to male and female reproductive disease. We retrieved 154 female-specific and 945 male-specific disease genes related to the reproductive system (Supplemental Table S11). Through linear regression analysis, we assessed the number of sex-specific reproductive disease genes against the total gene numbers for each chromosome. We observed strikingly different patterns dependent on sex and chromosomal locations.

For female reproductive disease genes, the X Chromosome followed a linear pattern and did not differ significantly from autosomes ($R^2 = 0.53$, $P = 1.04 \times 10^{-4}$) (Fig. 3E). In contrast, male reproductive disease genes on the X and Y Chromosomes showed significant deviations from the autosomes of a linear pattern ($R^2 = 0.82$, $P = 5.56 \times 10^{-9}$) (Fig. 3F). The X Chromosome contained 111.75% more male reproductive genes than expected. Moreover, the Y (17/45) and X (80/840) Chromosomes had significantly higher ratios of male reproductive disease genes compared with autosomes (averaging 38/853), with odds ratios of 8.48 (95% CI: 4.45–16.17, $P < 0.0001$) and 2.14 (95% CI: 1.44–3.18, $P = 0.0002$), respectively. On the X Chromosome, male reproductive genes outnumbered female ones by a factor of 10.43 (80/840 vs. 7/840). This

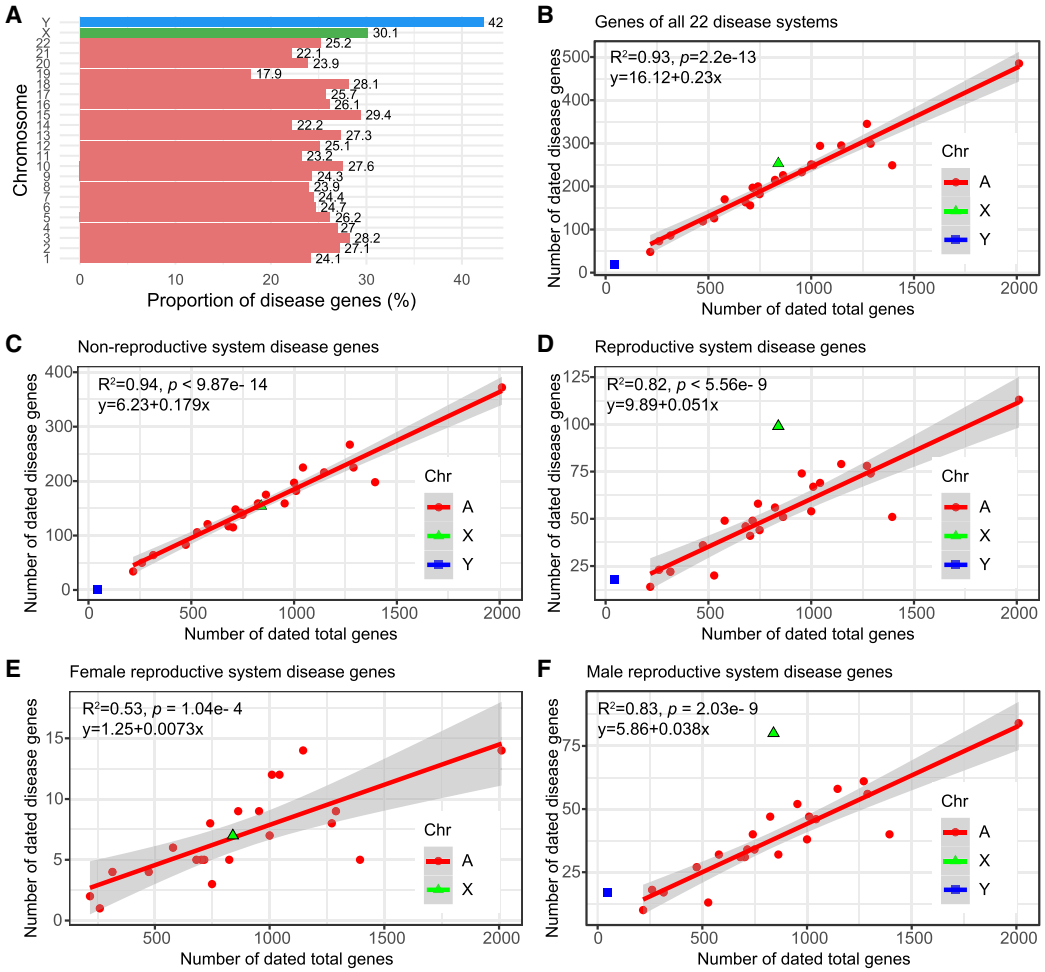

Figure 3. Chromosomal analyses for disease genes. (*A*) The proportions of disease genes across chromosomes. The pink bars represent the autosomes, and green and blue indicate the X and Y Chromosomes, respectively. The proportions (%) for different chromosomes are shown *above* the bars. (*B*) The linear regression plotting of disease gene counts against the numbers of total genes with age information on chromosomes. (*C*) Numbers of genes related to the abnormality of nongenitourinary system (nonreproductive system) are plotted against all protein-coding genes on chromosomes with gene age information. (*D*) Numbers 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 genes on chromosomes for female-specific reproductive disease genes with gene age information. (*F*) Linear regression of disease gene counts against the total numbers of genes on chromosomes for male-specific reproductive disease genes with gene age information. The autosomal linear models are displayed on the *top left* corner. Note that all linear regression formulas and statistics pertain only to autosomes. (A) Autosomes, (X) X Chromosomes, and (Y) Y Chromosomes.

observation is consistent with the "faster-X hypothesis," which suggests that purifying selection is more effective at eliminating recessive deleterious mutations on the X Chromosome owing to the male hemizygosity of the X Chromosome (Rice 1984; Charlesworth et al. 1987). A male bias is also evident in reproductive disease genes on autosomes, with the male linear model slope being approximately 4.21 times steeper than that for females (0.038 vs. 0.0073) (Fig. 3E,F). Thus, the observed excess of male reproductive disease genes is not solely owing to the "faster-X" effect. It might also be influenced by the "faster-male" effect, in which the male reproductive system evolves rapidly owing to heightened sexual selection pressures on males (Wu and Davis 1993).

Excess of male reproductive disease genes in younger regions of the X Chromosome

Although we observed a male bias in reproductive disease genes, the influence of gene age on this excess remains unclear. We compared gene distribution patterns between older (or ancient, Euteleostomi) and younger (post-Euteleostomi) phylostrata. For female-specific reproductive disease genes, the X Chromosome has an excess of ancient genes (25.42%) but a deficiency of young genes (57.16%) (Fig. 4A). Conversely, among male-specific reproductive disease genes, younger genes exhibited a higher excess rate than ancient ones (193.96% vs. 80.09%) (Fig. 4A). These patterns suggest an age-dependent functional divergence of genes on the X Chromosome, which is consistent with gene expression data (Zhang et al. 2010). The X Chromosome is "masculinized" with young, male-biased genes, whereas old X Chromosomal genes tend to be "feminized," maintaining expression in females (Zhang et al. 2010). 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 (0.012 vs. 0.0021) genes (Fig. 4A). The ratio of male to female reproductive disease gene counts (α) showed a predominantly male-biased trend

across phylostrata, with a higher value in the most recent Eutheria (9.75) compared to the ancient phylostrata Euteleostomi and Tetrapoda (6.40 and 3.94) (Fig. 4B). A comparison of selection pressure between young and ancient genes revealed no significant difference for female-specific reproductive disease genes, but a significant difference for male-specific ones (Wilcoxon rank-sum test, $P < 0.0001$) (Fig. 4C), indicating that young genes under male-biased sexual selection have fewer evolutionary constraints than older ones (median K_a/K_s ratios 0.35 vs. 0.23).

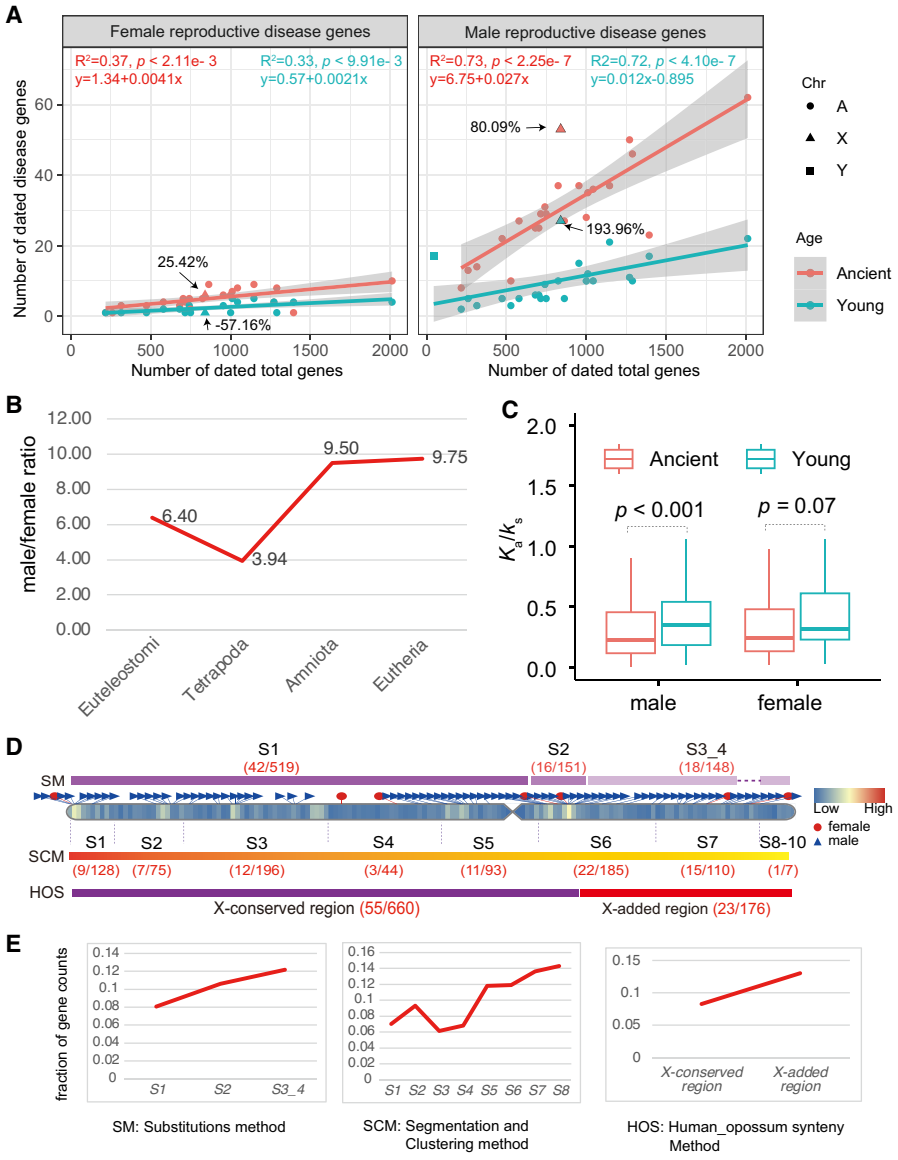

Figure 4. The X Chromosomal analyses for disease genes. (*A*) Numbers of female-specific (*left*) and male-specific reproductive disease genes (*right*) are plotted against all protein-coding genes with gene ages on chromosomes. The linear formulas fitted for autosomal genes at ancient (Euteleostomi) and younger (post-Euteleostomi) stages are shown in red and blue, respectively. The X Chromosome is shown with triangles. (*B*) The ratios of male to female reproductive disease gene numbers (α) across four phylostrata. (*C*) The comparison of selection pressure (human–chimpanzee pairwise K_a/K_s ratios) for sex-specific reproductive disease genes between the ancient (Euteleostomi or older) and younger (post-Euteleostomi) phylostrata. 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 in 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: (SM) the substitutions method (Lahn and Page 1999; McLysaght 2008), (SCM) the segmentation and clustering method (Pandey et al. 2013), and (HOS) the synteny method (orthologous gene order conservation between human and opossum) (Ross et al. 2005). (*E*) The fraction of disease genes with male-specific reproductive disease phenotypes within each X Chromosomal subregion, as illustrated in *D*, is presented. The gene coordinates have been updated based on the hg38 reference with liftOver. (A) Autosomes, (X) X Chromosomes, and (Y) Y Chromosomes.

Structurally, the eutherian hemizygous X Chromosome comprises an ancestral X-conserved region and a relatively new X-added region (Bellott et al. 2010). The ancestral X-conserved region is shared with the marsupial X Chromosome, whereas the X-added region originates from autosomes (Fig. 4D). To understand whether these regions of the human X Chromosome might contribute differently to human genetic disease phenotypes, we compared genes within the X-conserved and X-added regions, based on previous studies of evolutionary strata on the X Chromosome (Ross et al. 2005; McLysaght 2008; Pandey et al. 2013). After excluding genes in pseudoautosomal regions (X-PARs; Ensembl v110), we found that the proportion of male-specific reproductive disease genes in the X-added region (13.07%, 23/176) exceeds that in the X-conserved region (8.33%, 55/660) (Fig. 4D,E; Supplemental Table S12). Moreover, analyses of the evolutionary strata, based on the substitutions method (Lahn and Page 1999) and the segmentation and clustering method (Pandey et al. 2013), consistently showed higher fractions of male-specific reproductive disease genes in younger evolutionary strata than in older ones (Fig. 4E). These observations indicate that, on the X Chromosome, young genes could be more susceptible to male-biased sexual selection than old genes, despite their nearly identical hemizygous environment. Thus, the higher α in X-linked younger regions could not be attributed exclusively to the "male-driven," "faster-X," "faster-male," "male X-hemizygosity" effects, as these impact X-linked older and young genes similarly. Instead, the higher α in X-linked young regions may be driven by lower pleiotropy among young genes, allowing novel male-related functions to emerge faster in young genes than in older ones.

Discussion

By combining human gene age dating and Mendelian disease phenotyping, we reveal a trend of disease gene proportions increasing over evolutionary time. This growth pattern is attributed to higher burdens of deleterious variants in older

genes. The ratio of health-related genes per million years remains relatively consistent across macroevolutionary phylostrata. Young genes are preferentially linked to disease phenotypes in the male reproductive system, as well as in systems that have undergone significant phenotypic innovations during 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, which tend to display specialized functions. Our findings highlight that young genes are likely at the forefront of frontrunners of molecular evolution for phenotypic innovation (see Supplemental Material).

Our findings raise a question of why new genes can quickly contribute to 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 does not offer specific predictions regarding the propensity of new or young genes to be involved in adaptive traits. Other theories, such as the "male-driven," "faster-X," "male X-hemizygosity," and "faster-male" theories, cannot explain the finding that male-biased functions are more prevalent in young X-linked genes compared to older ones. Here, we propose that different phenotypic patterns between young and older genes could be related to differences in their pleiotropy.

The phenomenon of pleiotropy has been recognized or suggested for a considerable time (Mendel 1866; Wright 1984; Barton 1990; Baatz and Wagner 1997). Mendel's classic paper in 1866 suggests a single factor controls three characters of Pisum (Mendel 1866). Before Mendel, many medical professionals had already described syndromes characterized by different symptoms and a single "familial" factor (Eckman 1788). In the context of new gene evolution, it is established that young genes exhibit higher specificity and narrower expression breadth across tissues (Zhang et al. 2012). In our study, pleiotropy is more relevantly defined as involving anatomical systems and critical phenotypes (Pyeritz 1989; Lobo 2008; Tyler et al. 2009; Zhang 2023). We reveal a pattern that older genes tend to impact disease phenotypes of more organs/systems, whereas young genes largely display phenotype enrichment in specific organs. Therefore, both phenotypic pattern and expression trends across phylostrata suggest that young genes may have lower pleiotropy compared with older genes.

Pleiotropy impedes evolutionary adaptation, often referred to as the "cost of complexity" (Zeng and Hill 1986; Baatz and Wagner 1997; Orr 2000; Wagner and Zhang 2011; Fraïsse et al. 2019; Quiver and Lachance 2022), whereas low pleiotropy could foster morphological evolution (Carroll 2005; Wray 2007). The inhibitory effect of pleiotropy on novel adaptation aligns with our observations of the stronger purifying selection on older genes with higher pleiotropy (Wagner and Zhang 2011; Quiver and Lachance 2022) and broader expression patterns (Zhu et al. 2008). This evolutionary constraint suggests a restricted mutation space to introduce novel traits for old genes owing to the "competing interests" of multifunctionality (Fig. 5). The inhibitory pressure could also reduce genetic diversity owing to background selection (Charlesworth et al. 1993). The evolution of new genes, especially gene duplicates, serves as a primary mechanism to mitigate pleiotropic effects (He and Zhang 2005; Guillaume and Otto 2012) and avoid adverse pleiotropy in ancestral copies (Hoekstra and Coyne 2007). The tissue-specific functions of new genes, as a general pattern in numerous organisms, could circumvent adaptive conflicts caused by the multifunctionality of parental genes (Des Marais and Rausher 2008). The reduced pleiotropy in young genes may therefore provide a more diverse mutational space for functional innovation, minimizing unintended pleiotropic trade-offs (Dezső et al. 2008).

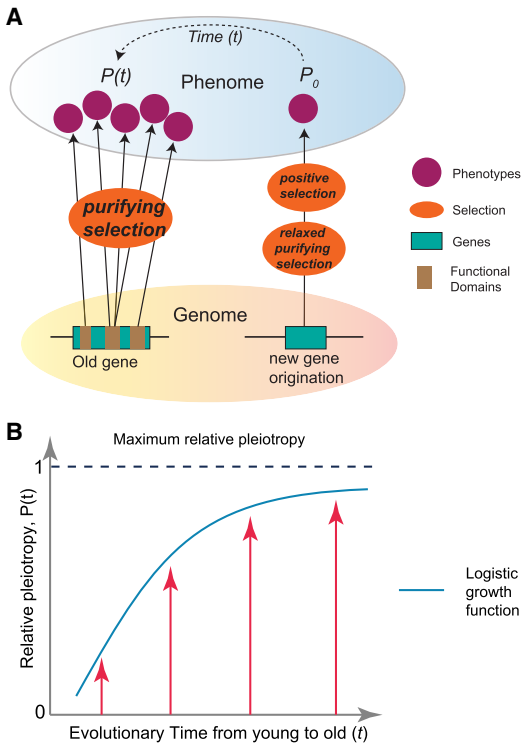

Figure 5. The "pleiotropy-barrier" model. (*A*) The "pleiotropy-barrier" model posits a dynamic process that new genes evolve adaptively more quickly compared with older genes. It suggests that older genes undergo stronger purifying selection because their multiple functions (usually adverse pleiotropy) act as a "barrier-like" factor to hinder fixations 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.

Here, we propose a "pleiotropy-barrier" hypothesis to explain the relationship between innovation potential and gene ages (Fig. 5A). This model predicts that the capacity for phenotypic innovation is limited by genetic pleiotropy under natural selection, suggesting that genes with lower pleiotropy may have greater potential for functional innovation. Over evolutionary time, the pleiotropy increase follows a logistic growth pattern, in which the growth could be faster for younger genes but slower for older genes (Fig. 5B). The multifunctional genes could encounter an escalating "barrier" or "resistance" to the pleiotropy growth. This barrier arises because more functions necessitate stronger selective constraints, which could, in turn, reduce the mutational space of beneficial mutations for novel phenotypes. In contrast, low or absent pleiotropy in new genes allows for a broader and more tunable mutation space under 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 responders in certain environment under

positive selection. For phenotypes under sexual selection, high pleiotropy may limit the potential to resolve sexual conflict through sex-limited expression (VanKuren et al. 2024). Thus, the "pleiotropy-barrier" model also applies to the evolution of new genes with male-specific functions driven by sexual selection (VanKuren and Long 2018).

The "pleiotropy-barrier" model does not predict a static saturation of pleiotropy in older genes, rather it emphasizes a continuous and dynamic process between gene age and innovation potential, in which gene pleiotropy imposes selective constraints that limit further innovation. This evolving constraint creates a "barrier" that diminishes the potential for genes to acquire new functions, as any beneficial mutations must navigate the complex interplay of existing functions and selective pressures. In contrast, younger genes, which start with low or no pleiotropy, have a greater capacity for evolutionary change. Their low multifunctionality allows them to exploit a wider mutational space, facilitating the development of novel traits and functions. This dynamic process is evident in systems that have undergone significant phenotypic innovations in human evolution, such as the nervous system, the musculoskeletal system, the male reproductive system, and the immune system.

Therefore, new or young genes with a lower pleiotropic effect as a selective advantage not only spur molecular evolution under sexual and natural selection but also, from a medical standpoint, are promising targets for precise medicine, warranting deeper investigation.

Methods

Gene age dating and disease phenotypes

The gene age dating was conducted using an inclusive approach for both autosomes and sex chromosomes (Chromosome X and Y). Specifically, for autosomal and X Chromosomal genes, we primarily obtained gene ages (phylostrata, branches, or origination stages) from the GenTree database (Zhang et al. 2010; Shao et al. 2019), which is based on Ensembl v95 of human reference genome version hg38 (Flicek et al. 2014). We then *trans*-mapped the v95 gene list of GenTree into 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 (Zhang et al. 2010; Long et al. 2013; Shao et al. 2019). The most phylogenetically distant branch at which the shared syntenic block was detected marks the latest possible time range when a human gene originated. In comparison to the method based on the similarity of protein families, namely, the phylostratigraphic dating (Neme and Tautz 2013), this method employed in GenTree is robust to recent gene duplications (Shao et al. 2019), despite its underestimation of the number of young genes (Ma et al. 2022). We obtained gene age for human Y genes through the analysis of 15 representative mammals (Cortez et al. 2014). Notably, Y gene ages are defined as the time when they began to evolve independently from their X counterpart or when they translocated from other chromosomes to the Y Chromosome owing to gene traffic (transposition/translocation) (Cortez et al. 2014). 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 phylogenetic framework of GenTree (Shao et al. 2019). These comprehensive methods resulted in the categorization of 19,665 protein-coding genes into distinct gene age groups, encompassing evolutionary stages from Euteleostomi (br0) to the human lineage (br14). We did not differentiate Euteleostomi-specific genes from more ancient genes because the synteny-based method reveals a faster decay of synteny than sequence similarity in ancient genes (Arendsee et al. 2019), making it complicated to determine accurate ancient branches for ancient genes. We merged Euteleostomi-specific and more ancient genes into "br0" (labeled with "Euteleostomi" in this study), following the synteny-based method in previous studies (Zhang et al. 2010; Shao et al. 2019).

The HPO annotation used in this study for phenotypic abnormalities contains disease genes corresponding to organ/tissue systems (September 19, 2023; https://hpo.jax.org/app/data/annotations). This repository synthesizes information from diverse databases, including Orphanet (Weinreich et al. 2008), DECIPHER (Wright 2015), and OMIM (Hamosh et al. 2000). After filtering out mitochondrial genes, unplaced genes, RNA genes, and genes related to neoplasm ontology, we obtained 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 by using a grep shell script with the keywords "reproduct," "male," and "female" (neoplasm-related items were removed).

Logistic regression modeling and model comparison

We retrieved the gene-wise burdens of rare de novo germline variants from multiple studies, including the Gene4Denovo database (68,404 individuals) (Zhao et al. 2020; http://www.genemed.tech/gene4denovo/uploads/gene4denovo/All_De_novo_mutations_1.2.txt), and burden scores of ultrarare loss-of-function variants from UK Biobank exomes (394,783 individuals) (Supplemental Table 1 in a previous study; Weiner et al. 2023). The gene-wise burdens of rare variants at the population level were estimated with data from the whole-genome sequencing genotypes and allele frequencies of gnomAD database (version 4.1.0 of 76,215 individuals) (Wang et al. 2022; Greene et al. 2023; all chromosome VCF files from https://gnomad.broadinstitute.org/downloads). Rare variants were extracted based on a MAF lower than 0.0001 across all major human populations (all human male, all human female, African population, non-Finish European population, East Asian population, South Asian population, and Latino/Admixed American population).

We conducted the stratified logistic regression to account for effects of multiple predictors and their interactions on the outcome of gene disease states. The disease and nondisease genes were assigned into binary states ("1," disease genes; "0," nondisease genes) as a response variable. A step-by-step procedure was performed for multiple predictors, which include gene age (T; mya), gene length (L_g) or protein length (L), DNV burden (D) (Wang et al. 2022), and rare variant burden (R) (Chen et al. 2024b). The LRT and AIC were used for model comparison. Lower AIC was preferred if the same degree of freedom was detected. The model with significant LRT P-value ($P < 0.05$) was chosen when comparing nested models. To account for different scales in variables and potential influence of extreme values, the variables with logarithm treatment were also incorporated in some models. The VIF values were used to account for the multicollinearity among variables. The "glm" package (binomial model) in R platform was used for computing the models (https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm).

K_a / K_s ratio

K_a/K_s is widely used in evolutionary genetics to estimate the relative strength of purifying selection ($K_a/K_s < 1$), neutral mutations

($K_a/K_s = 1$), and potentially beneficial mutations ($K_a/K_s > 1$) on homologous protein-coding genes. K_a is the number of nonsynonymous substitutions per nonsynonymous site, and K_s is the number of synonymous substitutions per synonymous site that is assumed to be neutral. The pairwise K_a/K_s ratios (human–chimpanzee, human–bonobo, and human–macaque) were retrieved from the Ensembl database (v99) (Flicek et al. 2014), as estimated with the maximum likelihood algorithm (Yang 2007).

Disease gene emergence rate per million years (*r*)

To understand the origination tempo of disease genes within different evolutionary phylostrata, 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 calculation was 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, and O_i indicates the number of disease genes with OPs 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 (Kumar et al. 2017; http://www.timetree.org).

Pleiotropic modeling with logistic growth function

For each evolutionary phylostratum (t), we estimated median OP numbers 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 (R Core Team 2023).

Phenotype enrichment along evolutionary stages

The phenotype enrichment along phylostrata 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" owing 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:

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

where m indicates the number of phenotype(s) one gene can affect, n represents the number of genes identified for a given phenotype, and l is number of phenotypes within a given evolutionary stage. Considering the genetic complexity of phenotypes, PEI firstly adjusted the weights of genes related to a phenotype with the reciprocal value of m, that is, $\frac{1}{m}$. 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 ith 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 jth phenotype within the evolutionary stage, and m_k is the number of phenotypes affected by the kth gene within the jth 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 $\left(\sum_{J=1}^{l} P\right)$ and finally obtained the percentage of each phenotype within each stage $\left(\frac{P}{\sum_{J=1}^{l} P}\right)$ as the enrichment index.

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 $\left(\frac{n-e}{e}\right)$.

Analyses on X-conserved and X-added regions

The Eutherian X Chromosome is composed of the pseudoautosomal regions (PARs), X-conserved region, and X-added region. The regions of two PARs were determined based on NCBI assembly annotation of GRCh38.p13 (X:155701383–156030895 and X:10 001–2781479). The X-boundary between X-conserved and X-added regions was determined with the 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, whereas the X-added region in human has synteny with the autosomal genes of opossum (Ross et al. 2005). The "evolutionary strata" on X were based on previous reports of two methods: the substitutions method and the segmentation and clustering method (Lahn and Page 1999; McLysaght 2008; Pandey et al. 2013). The coordinates of strata boundaries were up-lifted into hg38 genome with liftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver).

Software availability

All rare variants data files (MAF lower than 0.0001) generated in this study and the R script for modeling have been submitted to Zenodo (https://zenodo.org/uploads/11000269). The R script is also available as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

M.L. 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). D.A. was supported by National Institutes of Health (F32GM146423). We greatly appreciate the constructive discussions with Dr. Stefano Allesina, Dr. Urs C. Schmidt-Ott, and Dr. Greg Dwyer of the University of Chicago; Dr. Anne O'Donnell Luria of the Broad Institute of Massachusetts Institute of Technology and Harvard; Dr. Cheng Deng of Western China Hospital; and Dr. Chuanzhu Fan from Wayne State University. Special acknowledgment is given to Xuefei He from Western China Hospital for designing the

silhouettes. We thank the maintainers and contributors of the HPO data.

Author contributions: J.-H.C., P.L., and M.L. conceived the study. J.-H.C. and M.L. designed the methodology. P.L., D.A., and A.G. performed data curation and investigation. J.-H.C. and P.L. conducted analyses and wrote the original draft. M.L. and B.S. supervised the study, and all authors reviewed and edited the manuscript.

References

Antonarakis SE, Beckmann JS. 2006. Mendelian disorders deserve more attention. *Nat Rev Genet* **7:** 277–282. doi:10.1038/nrg1826

Arendsee Z, Li J, Singh U, Bhandary P, Seetharam A, Wurtele ES. 2019. *fagin*: synteny-based phylostratigraphy and finer classification of young genes. *BMC Bioinformatics* **20:** 440. doi:10.1186/s12859-019-3023-y

Baatz M, Wagner GP. 1997. Adaptive inertia caused by hidden pleiotropic effects. *Theor Popul Biol* **51:** 49–66. doi:10.1006/tpbi.1997.1294

Baertsch R, Diekhans M, Kent WJ, Haussler D, Brosius J. 2008. Retrocopy contributions to the evolution of the human genome. *BMC Genomics* **9:** 466. doi:10.1186/1471-2164-9-466

Barton NH. 1990. Pleiotropic models of quantitative variation. *Genetics* **124:** 773–782. doi:10.1093/genetics/124.3.773

Bellott DW, Skaletsky H, Pyntikova T, Mardis ER, Graves T, Kremitzki C, Brown LG, Rozen S, Warren WC, Wilson RK, et al. 2010. Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature* **466:** 612–616. doi:10.1038/nature09172

Betrán E, Long M. 2022. Evolutionary new genes in a growing paradigm. *Genes (Basel)* **13:** 1605. doi:10.3390/genes13091605

Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. 2013. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet* **14:** 681–691. doi:10.1038/nrg3555

Brosius J. 1991. Retroposons: seeds of evolution. *Science* **251:** 753. doi:10.1126/science.1990437

Carelli FN, Hayakawa T, Go Y, Imai H, Warnefors M, Kaessmann H. 2016. The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome Res* **26:** 301–314. doi:10.1101/gr.198473.115

Carroll SB. 2005. Evolution at two levels: on genes and form. *PLoS Biol* **3:** e245. doi:10.1371/journal.pbio.0030245

Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and de novo gene birth. *Nature* **487:** 370–374. doi:10.1038/nature11184

Charlesworth B, Coyne JA, Barton NH. 1987. The relative rates of evolution of sex chromosomes and autosomes. *Am Nat* **130:** 113–146. doi:10.1086/284701

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134:** 1289–1303. doi:10.1093/genetics/134.4.1289

Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. *Nat Rev Genet* **14:** 645–660. doi:10.1038/nrg3521

Chen J, He X, Jakovlić I. 2022. Positive selection-driven fixation of a hominin-specific amino acid mutation related to dephosphorylation in IRF9. *BMC Ecol Evol* **22:** 132. doi:10.1186/s12862-022-02088-5

Chen J, Jia Y, Zhong J, Zhang K, Dai H, He G, Li F, Zeng L, Fan C, Xu H. 2024a. Novel mutation leading to splice donor loss in a conserved site of *DMD* gene causes Duchenne muscular dystrophy with cryptorchidism. *J Med Genet* **61:** 741–749. doi:10.1136/jmg-2024-109896

Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, Alföldi J, Watts NA, Vittal C, Gauthier LD, et al. 2024b. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **625:** 92–100. doi:10.1038/s41586-023-06045-0

Ciccarelli FD, von Mering C, Suyama M, Harrington ED, Izaurralde E, Bork P. 2005. Complex genomic rearrangements lead to novel primate gene function. *Genome Res* **15:** 343–351. doi:10.1101/gr.3266405

Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, Kathiresan S, Kenny EE, Lindgren CM, MacArthur DG, et al. 2020. A brief history of human disease genetics. *Nature* **577:** 179–189. doi:10.1038/s41586-019-1879-7

Conrad B, Antonarakis SE. 2007. Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu Rev Genomics Hum Genet* **8:** 17–35. doi:10.1146/annurev.genom.8.021307.110233

Cortez D, Marin R, Toledo-Flores D, Froidevaux L, Liechti A, Waters PD, Grützner F, Kaessmann H. 2014. Origins and functional evolution of Y chromosomes across mammals. *Nature* **508:** 488–493. doi:10.1038/nature13151

Des Marais DL, Rausher MD. 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* **454:** 762–765. doi:10.1038/nature07092

Dezső Z, Nikolsky Y, Sviridov E, Shi W, Serebriyskaya T, Dosymbekov D, Bugrim A, Rakhmatulin E, Brennan RJ, Guryanov A, et al. 2008. A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol* **6:** 49. doi:10.1186/1741-7007-6-49

Ding D, Nguyen TT, Pang MYH, Ishibashi T. 2021. Primate-specific histone variants. *Genome* **64:** 337–346. doi:10.1139/gen-2020-0094

Domazet-Lošo T, Tautz D. 2008. An ancient evolutionary origin of genes associated with human genetic diseases. *Mol Biol Evol* **25:** 2699–2707. doi:10.1093/molbev/msn214

Domazet-Lošo T, Brajković J, Tautz D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet* **23:** 533–539. doi:10.1016/j.tig.2007.08.014

Eckman OJ. 1788. *The description and multiple causes of osteomalaciae sistens*. Uppsala, Sweden.

Emerson J, Kaessmann H, Betrán E, Long M. 2004. Extensive gene traffic on the mammalian X chromosome. *Science* **303:** 537–540. doi:10.1126/science.1090042

Fetro C, Scherman D. 2020. Drug repurposing in rare diseases: myths and reality. *Therapies* **75:** 157–160. doi:10.1016/j.therap.2020.02.006

Fleck K, Luria V, Garag N, Karger A, Hunter T, Marten D, Phu W, Nam KM, Sestan N, O'Donnell-Luria AH, et al. 2024. Functional associations of evolutionarily recent human genes exhibit sensitivity to the 3D genome landscape and disease. bioRxiv doi:10.1101/2024.03.17.585403

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Res* **42:** D749–D755. doi:10.1093/nar/gkt1196

Fraïsse C, Puixeu Sala G, Vicoso B. 2019. Pleiotropy modulates the efficacy of selection in *Drosophila melanogaster*. *Mol Biol Evol* **36:** 500–515. doi:10.1093/molbev/msy246

Gibson G. 2012. Rare and common variants: twenty arguments. *Nat Rev Genet* **13:** 135–145. doi:10.1038/nrg3118

Greene D, Pirri D, Frudd K, Sackey E, Al-Owain M, Giese APJ, Ramzan K, Riaz S, Yamanaka I, Boeckx N, et al. 2023. Genetic association analysis of 77,539 genomes reveals rare disease etiologies. *Nat Med* **29:** 679–688. doi:10.1038/s41591-023-02211-z

Guillaume F, Otto SP. 2012. Gene functional trade-offs and the evolution of pleiotropy. *Genetics* **192:** 1389–1409. doi:10.1534/genetics.112.143214

Guo MH, Plummer L, Chan YM, Hirschhorn JN, Lippincott MF. 2018. Burden testing of rare variants identified through exome sequencing via publicly available control data. *Am J Hum Genet* **103:** 522–534. doi:10.1016/j.ajhg.2018.08.016

Halvorsen M, Huh R, Oskolkov N, Wen J, Netotea S, Giusti-Rodriguez P, Karlsson R, Bryois J, Nystedt B, Ameur A, et al. 2020. Increased burden of ultra-rare structural variants localizing to boundaries of topologically associated domains in schizophrenia. *Nat Commun* **11:** 1842. doi:10.1038/s41467-020-15707-w

Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. 2000. Online Mendelian inheritance in man (OMIM). *Hum Mutat* **15:** 57–61. doi:10.1002/(SICI)1098-1004(200001)15:1<57::AID-HUMU12>3.0.CO;2-G

He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169:** 1157–1164. doi:10.1534/genetics.104.037051

Hodgkin J. 1998. Seven types of pleiotropy. *Int J Dev Biol* **42:** 501–505.

Hoekstra HE, Coyne JA. 2007. The locus of evolution: evo devo and the genetics of adaptation. *Evolution* **61:** 995–1016. doi:10.1111/j.1558-5646.2007.00105.x

Jia Y, Chen J, Zhong J, He X, Zeng L, Wang Y, Li J, Xia S, Ye E, Zhao J, et al. 2023. Novel rare mutation in a conserved site of *PTPRB* causes human hypoplastic left heart syndrome. *Clin Genet* **103:** 79–86. doi:10.1111/cge.14234

Jiang L, Jiang H, Dai S, Chen Y, Song Y, Tang Clara S-M, Pang SY-Y, Ho S-L, Wang B, Garcia-Barcelo M-M, et al. 2022. Deviation from baseline mutation burden provides powerful and robust rare-variants association test for complex diseases. *Nucleic Acids Res* **50:** e34. doi:10.1093/nar/gkab1234

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20:** 1313–1326. doi:10.1101/gr.101386.109

Kaessmann H, Zöllner S, Nekrutenko A, Li WH. 2002. Signatures of domain shuffling in the human genome. *Genome Res* **12:** 1642–1650. doi:10.1101/gr.520702

Kasinathan B, Colmenares SU III, McConnell H, Young JM, Karpen GH, Malik HS. 2020. Innovation of heterochromatin functions drives rapid evolution of essential ZAD-ZNF genes in *Drosophila*. *eLife* **9:** e63368. doi:10.7554/eLife.63368

Kingdom R, Beaumont RN, Wood AR, Weedon MN, Wright CF. 2024. Genetic modifiers of rare variants in monogenic developmental disorder loci. *Nat Genet* **56:** 861–868. doi:10.1038/s41588-024-01710-0

Klomp J, Athy D, Kwan CW, Bloch NI, Sandmann T, Lemke S, Schmidt-Ott U. 2015. A cysteine-clamp gene drives embryo polarity in the midge *Chironomus*. *Science* **348:** 1040–1042. doi:10.1126/science.aaa7105

Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine J-P, Gargano M, Harris NL, Matentzoglu N, McMurry JA, et al. 2019. Expansion of the human phenotype ontology (HPO) knowledge base and resources. *Nucleic Acids Res* **47:** D1018–D1027. doi:10.1093/nar/gky1105

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol* **34:** 1812–1819. doi:10.1093/molbev/msx116

Lahn BT, Page DC. 1999. Four evolutionary strata on the human X chromosome. *Science* **286:** 964–967. doi:10.1126/science.286.5441.964

Lee S, Abecasis Gonçalo R, Boehnke M, Lin X. 2014. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* **95:** 5–23. doi:10.1016/j.ajhg.2014.06.009

Li H, Chen C, Wang Z, Wang K, Li Y, Wang W. 2021. Pattern of new gene origination in a special fish lineage, the flatfishes. *Genes (Basel)* **12:** 1819. doi:10.3390/genes12111819

Lobo I. 2008. Pleiotropy: one gene can affect multiple traits. *Nat Educ* **1:** 10.

Long M, Langley CH. 1993. Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. *Science* **260:** 91–95. doi:10.1126/science.7682012

Long M, VanKuren NW, Chen S, Vibranovski MD. 2013. New gene evolution: little did we know. *Annu Rev Genet* **47:** 307–333. doi:10.1146/annurev-genet-111212-133301

Luria V, Ma S, Shibata M, Pattabiraman K, Sestan N. 2023. Molecular and cellular mechanisms of human cortical connectivity. *Curr Opin Neurobiol* **80:** 102699. doi:10.1016/j.conb.2023.102699

Ma C, Li C, Ma H, Yu D, Zhang Y, Zhang D, Su T, Wu J, Wang X, Zhang L, et al. 2022. Pan-cancer surveys indicate cell cycle-related roles of primate-specific genes in tumors and embryonic cerebrum. *Genome Biol* **23:** 251. doi:10.1186/s13059-022-02821-9

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol* **3:** e357. doi:10.1371/journal.pbio.0030357

McLysaght A. 2008. Evolutionary steps of sex chromosomes are reflected in retrogenes. *Trends Genet* **24:** 478–481. doi:10.1016/j.tig.2008.07.006

Mendel JG. 1866. Experiments in plant hybridization. *Verhandlungen des naturforschenden Vereines in Brunn* **4:** 3–47.

Miller D, Chen J, Liang J, Betrán E, Long M, Sharakhov IV. 2022. Retrogene duplication and expression patterns shaped by the evolution of sex chromosomes in malaria mosquitoes. *Genes (Basel)* **13:** 968. doi:10.3390/genes13060968

Montañés JC, Huertas M, Messeguer X, Albà MM. 2023. Evolutionary trajectories of new duplicated and putative de novo genes. *Mol Biol Evol* **40:** msad098. doi:10.1093/molbev/msad098

Nei M. 2013. *Mutation-driven evolution*. Oxford University Press, Oxford.

Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics* **14:** 117. doi:10.1186/1471-2164-14-117

Orr HA. 2000. Adaptation and the cost of complexity. *Evolution* **54:** 13–20. doi:10.1111/j.0014-3820.2000.tb00002.x

Pandey RS, Wilson Sayres MA, Azad RK. 2013. Detecting evolutionary strata on the human X chromosome in the absence of gametologous Y-linked sequences. *Genome Biol Evol* **5:** 1863–1871. doi:10.1093/gbe/evt139

Paul D. 2000. A double-edged sword. *Nature* **405:** 515. doi:10.1038/35014676

Pavličev M, Wagner GP. 2022. The value of broad taxonomic comparisons in evolutionary medicine: Disease is not a trait but a *state of a trait*! *MedComm (2020)* **3:** e174. doi:10.1002/mco2.174

Prabh N, Roeseler W, Witte H, Eberhardt G, Sommer RJ, Rödelsperger C. 2018. Deep taxon sampling reveals the evolutionary dynamics of novel gene families in *Pristionchus* nematodes. *Genome Res* **28:** 1664–1674. doi:10.1101/gr.234971.118

Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, O'Dushlaine C, Chambert K, Bergen SE, Kähler A, et al. 2014. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506:** 185–190. doi:10.1038/nature12975

Pyeritz RE. 1989. Pleiotropy revisited: molecular explanations of a classic concept. *Am J Med Genet* **34:** 124–134. doi:10.1002/ajmg.1320340120

Quiver MH, Lachance J. 2022. Adaptive eQTLs reveal the evolutionary impacts of pleiotropy and tissue-specificity while contributing to health and disease. *HGG Adv* **3:** 100083. doi:10.1016/j.xhgg.2021.100083

Ragsdale EJ, Müller Manuela R, Rödelsperger C, Sommer Ralf J. 2013. A developmental switch coupled to the evolution of plasticity acts through a sulfatase. *Cell* **155:** 922–933. doi:10.1016/j.cell.2013.09.054

R Core Team. 2023. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Rice WR. 1984. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* **38:** 735–742. doi:10.2307/2408385

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, et al. 2015. 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* **17:** 405–424. doi:10.1038/gim.2015.30

Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M, Howell GR, Burrows C, Bird CP, et al. 2005. The DNA sequence of the human X chromosome. *Nature* **434:** 325–337. doi:10.1038/nature03440

Ross BD, Rosin L, Thomae AW, Hiatt MA, Vermaak D, de la Cruz AFA, Imhof A, Mellone BG, Malik HS. 2013. Stepwise evolution of essential centromere function in a *Drosophila* neogene. *Science* **340:** 1211–1214. doi:10.1126/science.1234393

Shao Y, Chen C, Shen H, He BZ, Yu D, Jiang S, Zhao S, Gao Z, Zhu Z, Chen X, et al. 2019. GenTree, an integrated resource for analyzing the evolution and function of primate-specific coding genes. *Genome Res* **29:** 682–696. doi:10.1101/gr.238733.118

Shi L, Su B. 2012. Identification and functional characterization of a primate-specific E2F1 binding motif regulating MCPH1 expression. *FEBS J* **279:** 491–503. doi:10.1111/j.1742-4658.2011.08441.x

Stolk P, Willemen MJ, Leufkens HG. 2006. Rare essentials: drugs for rare diseases as essential medicines. *Bull World Health Organ* **84:** 745–751. doi:10.2471/BLT.06.031518

Tyler AL, Asselbergs FW, Williams SM, Moore JH. 2009. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *BioEssays* **31:** 220–227. doi:10.1002/bies.200800022

VanKuren NW, Long M. 2018. Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions. *Nat Ecol Evol* **2:** 705–712. doi:10.1038/s41559-018-0471-0

VanKuren NW, Chen J, Long M. 2024. Sexual conflict drive in the rapid evolution of new gametogenesis genes. *Semin Cell Dev Biol* **159-160:** 27–37. doi:10.1016/j.semcdb.2024.01.005

Van Oss SB, Carvunis AR. 2019. De novo gene birth. *PLoS Genet* **15:** e1008160. doi:10.1371/journal.pgen.1008160

Vonica A, Bhat N, Phan K, Guo J, Iancu L, Weber JA, Karger A, Cain JW, Wang ECE, DeStefano GM, et al. 2020. Apcdd1 is a dual BMP/Wnt inhibitor in the developing nervous system and skin. *Dev Biol* **464:** 71–87. doi:10.1016/j.ydbio.2020.03.015

Wagner GP, Zhang J. 2011. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nat Rev Genet* **12:** 204–213. doi:10.1038/nrg2949

Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, Tachmazidou I, Vitsios D, Deevi SVV, Mackay A, Muthas D, et al. 2021. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597:** 527–532. doi:10.1038/s41586-021-03855-y

Wang T, Kim CN, Bakken TE, Gillentine MA, Henning B, Mao Y, Gilissen C, Consortium TS, Nowakowski TJ, Eichler EE, et al. 2022. Integrated gene analyses of de novo variants from 46,612 trios with autism and developmental disorders. *Proc Natl Acad Sci* **119:** e2203491119. doi:10.1073/pnas.2203491119

Wang RJ, Al-Saffar SI, Rogers J, Hahn MW. 2023. Human generation times across the past 250,000 years. *Sci Adv* **9:** eabm7047. doi:10.1126/sciadv.abm7047

Weiner DJ, Nadig A, Jagadeesh KA, Dey KK, Neale BM, Robinson EB, Karczewski KJ, O'Connor LJ. 2023. Polygenic architecture of rare coding variation across 394,783 exomes. *Nature* **614:** 492–499. doi:10.1038/s41586-022-05684-z

Weinreich SS, Mangon R, Sikkens JJ, Teeuw ME, Cornel MC. 2008. Orphanet: a European database for rare diseases. *Ned Tijdschr Geneeskd* **152:** 518–519.

Wray GA. 2007. The evolutionary significance of *cis*-regulatory mutations. *Nat Rev Genet* **8:** 206–216. doi:10.1038/nrg2063

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

Wright ES. 2015. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics* **16:** 322. doi:10.1186/s12859-015-0749-z

Wu C-I, Davis AW. 1993. Evolution of postmating reproductive isolation: the composite nature of Haldane's rule and its genetic bases. *Am Nat* **142:** 187–212. doi:10.1086/285534

Wu D-D, Irwin DM, Zhang Y-P. 2011. De novo origin of human protein-coding genes. *PLoS Genet* **7:** e1002379. doi:10.1371/journal.pgen.1002379

Xia S, Chen J, Arsala D, Emerson JJ, Long M. 2025. Functional innovation through new genes as a general evolutionary process. *Nat Genet* **57:** 295–309. doi:10.1038/s41588-024-02059-0

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24:** 1586–1591. doi:10.1093/molbev/msm088

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol (Amst)* **15:** 496–503. doi:10.1016/S0169-5347(00)01994-7

Zeng ZB, Hill WG. 1986. The selection limit due to the conflict between truncation and stabilizing selection with mutation. *Genetics* **114:** 1313–1328. doi:10.1093/genetics/114.4.1313

Zhang J. 2023. Patterns and evolutionary consequences of pleiotropy. *Annu Rev Ecol Evol Syst* **54:** 1–19. doi:10.1146/annurev-ecolsys-022323-083451

Zhang YE, Long M. 2014. New genes contribute to genetic and phenotypic novelties in human evolution. *Curr Opin Genet Dev* **29:** 90–96. doi:10.1016/j.gde.2014.08.013

Zhang YE, Vibranovski MD, Landback P, Marais GA, Long M. 2010. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol* **8:** e1000494. doi:10.1371/journal.pbio.1000494

Zhang YE, Landback P, Vibranovski MD, Long M. 2011. Accelerated recruitment of new brain development genes into the human genome. *PLoS Biol* **9:** e1001179. doi:10.1371/journal.pbio.1001179

Zhang YE, Landback P, Vibranovski M, Long M. 2012. New genes expressed in human brains: implications for annotating evolving genomes. *BioEssays* **34:** 982–991. doi:10.1002/bies.201200008

Zhang D, Leng L, Chen C, Huang J, Zhang Y, Yuan H, Ma C, Chen H, Zhang YE. 2022. Dosage sensitivity and exon shuffling shape the landscape of polymorphic duplicates in *Drosophila* and humans. *Nat Ecol Evol* **6:** 273–287. doi:10.1038/s41559-021-01614-w

Zhao G, Li K, Li B, Wang Z, Fang Z, Wang X, Zhang Y, Luo T, Zhou Q, Wang L, et al. 2020. Gene4Denovo: an integrated database and analytic platform for de novo mutations in humans. *Nucleic Acids Res* **48:** D913–D926. doi:10.1093/nar/gkz923

Zhu J, He F, Hu S, Yu J. 2008. On the nature of human housekeeping genes. *Trends Genet* **24:** 481–484. doi:10.1016/j.tig.2008.08.004

Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, Daly MJ, Neale BM, Sunyaev SR, Lander ES. 2014. Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci* **111:** E455–E464. doi:10.1073/pnas.1322563111

Received April 21, 2024; accepted in revised form February 6, 2025.

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

Jian-Hai Chen, Patrick Landback, Deanna Arsala, et al.

Genome Res. 2025 35: 379-392 originally published online February 14, 2025
Access the most recent version at doi:10.1101/gr.279498.124