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

Jianhai Chen[1,2,*], Patrick Landback[1], Deanna Arsala[1], Alexander Guzzetta[3], Shengqian Xia[1], Jared Atlas[1], Dylan Sosa[1],Yong E. Zhang[4], 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]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 genetic novelties pivotal in mammalian evolution. Their phenotypic impacts and evolutionary pattern over time, however, remain elusive in humans due to the technical and ethical complexities in functional studies. By combining human gene age dating and Mendelian disease phenotyping, our research reveals a gradual increase in disease gene proportions with gene age. Logistic regression modeling indicates that this increase could be related to longer protein lengths and higher burdens of deleterious de novo germline variants (DNVs) for older genes. 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. Notably, young genes show significant enrichment in diseases related to the male reproductive system, indicating strong sexual selection. Young genes also exhibit disease-related functions in tissues and systems 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 due to intensifying selective constraints over time. We propose a “pleiotropy-barrier” model that delineates higher potentials of phenotypic innovation for young genes than for older genes, a process 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 (Clausius 1879; and and Bebenek 2000; Vanchurin et al. 2022). 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 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 (Antonarakis and Beckmann 2006; Fetro and Scherman 2020). 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) (Boycott et al. 2013). These identified disease genes and associated phenotypes could provide critical 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 – have been continually emerging and integrating into the human genome throughout the macroevolutionary process of human lineage (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). 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 (Ciccarelli et al. 2005; Ding et al. 2021). For instance, in yeast, some *de novo* genes play roles in DNA repair process (Cai et al. 2008; Li et al. 2010; Parikh et al. 2022). In *Drosophila* species, lineage-specific genes can 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 the functions of morphological development, a process which 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 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 knock-downs 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 (Marques et al. 2005; Shi and Su 2012; Zhang and Long 2014). 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 (Zhang et al. 2011). 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 (An et al. 2023; Rich and Carvunis 2023). These lines of evidence in recent decades about the functions of new genes contradict the conventional conservation-dominant understanding of human genetics and phenotypes.

It has long been found that there are more disease genes in older genes than in young ones (Domazet-Lošo and Tautz 2008). However, the underlying mechanism remains unclear. In recent years, increasing availability of large-scale sequencing of whole exomes and whole genomes have greatly promoted the identification of deleterious variants underlying rare disorders (Richards et al. 2015). Statistical and medical studies have demonstrated that rare diseases are often caused by rare variants, which have more significant effects on disease phenotypes than common variants (Richards et al. 2015; Wang et al. 2021; Chen et al. 2022; Greene et al. 2023; Jia et al. 2023; Weiner et al. 2023; Chen et al. 2024a). The effect of gene-based rare variant burden—the aggregate impact of rare (including de novo germline) protein-altering variants—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. 2021).

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 allow us to understand the influence of gene ages on phenotypic enrichment, pleiotropy, and selective constraints along evolutionary journey. We revealed higher gene-wise deleterious DNVs burden in older genes than in young genes, suggesting higher susceptibility of older genes to genetic

diseases. This finding is also consistent with stronger purifying selection in older genes to remove deleterious mutations. 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 stable pace of gene integration per million years, young genes impact limited number of disease systems, with an enrichment in the male reproductive system and the nervous system, suggesting their lower pleiotropy and accelerated sexual selection and human-specific adaptive innovation. By contrast, older genes are higher in pleiotropic burden that impacts more anatomical systems and are under stronger purifying selection (selective constraints). These patterns suggest that the gene age as an important factor to influence levels of pleiotropy. 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 (Shao et al. 2019) and gene model annotations from Ensembl v110 (Supplementary Table 1). To ensure sufficient power in statistical comparisons across, we merged evolutionary age groups with a small number of genes (<100) into their adjacent older group (Figure 1a). As a result, we reclassified 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 the Human Phenotype Ontology database (HPO, Sep 2023), which is the *de facto* standard for phenotyping of rare Mendelian rare diseases (Köhler et al. 2018). This repository synthesizes information from diverse databases, including Orphanet (INSERM 1997; Weinreich et al. 2008), DECIPHER (Wright 2015), and OMIM (Hamosh et al. 2000). An intersection of these data sets yielded 4,946 genes annotated with both evolutionary age and organ/tissue/system-specific phenotypic abnormalities (Figure 1a-1b and Supplementary Table 2).

We observed an increase in the proportion of disease genes over evolutionary time (Figure 1a and 1c), suggesting that gene age impacts disease susceptibility—a trend consistent with earlier studies (Domazet-Lošo and Tautz 2008). Recent advances in medical genetics have revealed that Mendelian disorders are predominantly influenced by rare variants due to their significant phenotypic effects, a rational underlying the widespread use of burden tests in the field (Lee et al. 2014; Guo et al. 2018; Kingdom et al. 2024). However, the correlation between gene age and rare variant burden remains elusive. In addition, assuming a random mutation process, the rare variant burden could be a function of protein length. We analyzed the relationship between rare variant burdens, gene age, and protein length, using data from predicted deleterious de novo germline mutations (DNVs) from 46,612 trios (Wang et al. 2022) and rare variants from 76,215 genomes in the gnomAD database (v4.0.0, minor allele frequency (MAF) < 0.0001) (Chen et al. 2024b). Our findings showed strong positive correlations: protein length with DNVs burden (coefficient = 0.57, p < 2.2e-16) and rare variant burden (coefficient = 0.89, p < 2.2e-16). Additionally, gene age (mya) also showed a positive correlation with DNV burden (Pearson's r =

0.13, $p < 2.2e-16$) and rare variant burden (Pearson's r = 0.14, $p < 2.2e-16$). 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 datasets on de novo variant (DNV) burdens (Zhao et al. 2020), we also found the significant correlation between gene age and DNVs burden (Supplementary Figure 1). These correlations could not be inflated by mutation rate difference among genes, as no significant correlation was observed between gene age (*T, mya*) and DNVs per nucleotide (Pearson's product-moment correlation, $p = 0.17$).

The role of variant burdens on disease genes was analyzed based on datasets of gene-wise DNVs burden of 68,404 individuals from the Gene4Denovo database (Zhao et al. 2020) and the ultrarare predicted loss-of-function variant (pLoF, MAF $< 1 \times 10^{-5}$) burden from 394,783 exomes (Weiner et al. 2023). We observed that disease genes consistently exhibit a higher burden across all major evolutionary age groups (Wilcoxon rank sum test, $p < 2.2e-16$, Figure 1d-1e). Disease genes consistently exhibited longer proteins compared to non-disease genes across all major age groups (Figure 1f). To understand quantitative roles of these factors, we assigned binary states ("1", disease genes; "0", non-disease genes) to all genes and performed stratified Logistic regression modeling (Supplementary Table 4). We explored multiple predictors including gene age (*T, mya*), gene length (L_g) or protein length (L), DNVs burden (D) (Wang et al. 2022), and rare variant burden (R) based on gnomAD genomes (Chen et al. 2024b) (Supplementary Table 3). Model comparison was conducted based on likelihood ratio test (LRT) and Akaike information criterion (AIC) (Supplementary Table 4). The optimal model (M9) revealed the significant effects of three significant variables (DNV burden, gene age, and protein length), where the likelihood of being a disease gene increases with protein length (L at logarithm scale, coefficient = 0.15; $p < 0.044$), DNVs burden (D, coefficient = 0.09; $p < 8.9e-07$), and gene age (T, coefficient = 0.005; $p < 2e-16$) (Table 1). The Variance Inflation Factor (VIF) values for these three variables ranged from 1.02 to 1.72, well below the multicollinearity concern thresholds of 5 to 10, suggesting minimal impact on our model.

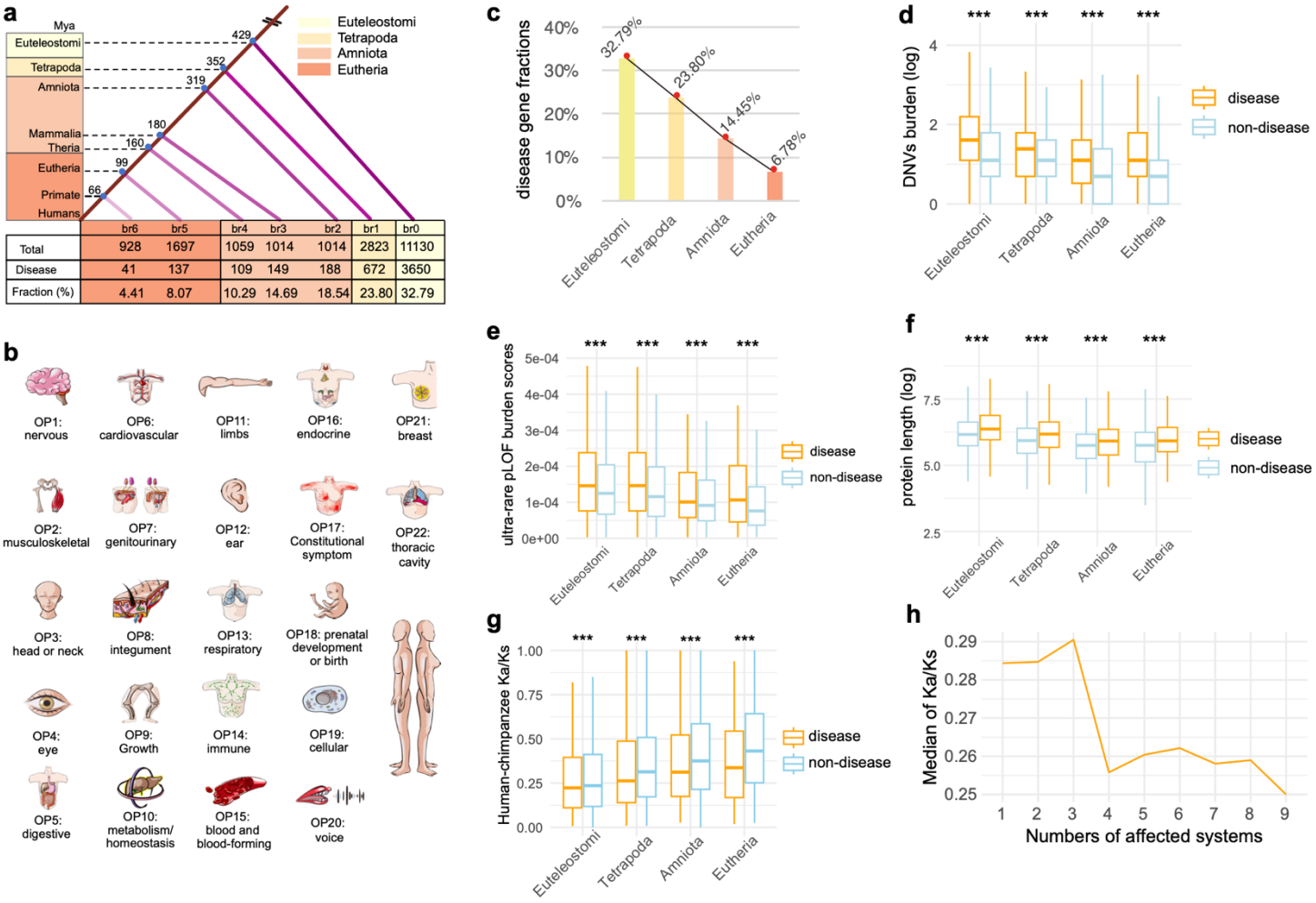
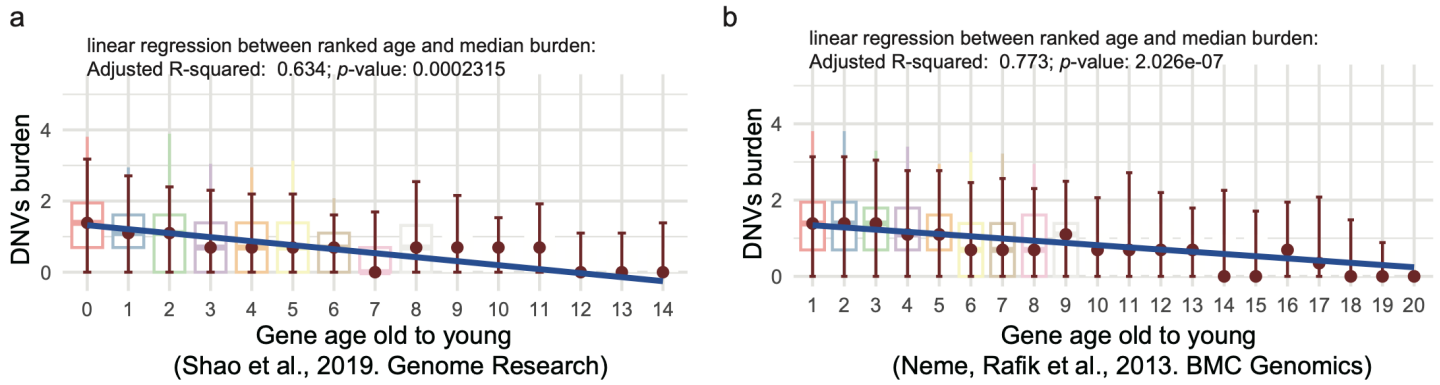


Figure 1. Number distribution and Ka/Ks ratios of genes categorized by ages and disease phenotypes (also organ phenotype genes). (**a**) The phylogenetic framework illustrating gene ages and disease genes associated with organ phenotypes. The phylogenetic branches represent age assignment for all genes and disease genes. The "br" values from br0 to br6 indicate different 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 22 HPO-defined organ/tissue systems, which are ordered based on the proportion of genes among all disease genes. (**c**) The proportions of disease genes at four major evolutionary age groups from Euteleostomi to Eutheria. (**d**) The gene-wise burden based on de novo germline variants from Gene4Denovo database (Zhao et al. 2020) between disease and non-disease genes across four major evolutionary age groups. (**e**) The burden score of ultrarare predicted loss-of-function (pLOF) variants (Weiner et al. 2023) between disease and non-disease genes across four major evolutionary age groups. (**f**) The protein lengths between disease and non-disease genes across four major evolutionary age groups. (**g**) 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). (**h**) The line plot for the number of affected systems for disease genes and their pairwise Ka/Ks ratios for “one to one” orthologs between human and

chimpanzee. Only genes under purifying selection are shown (Ka/Ks < 1). Note: all significance levels are determined using the Wilcoxon rank sum test, comparing disease genes to non-disease genes. The symbol "***" indicates significance level of $p < 0.001$.

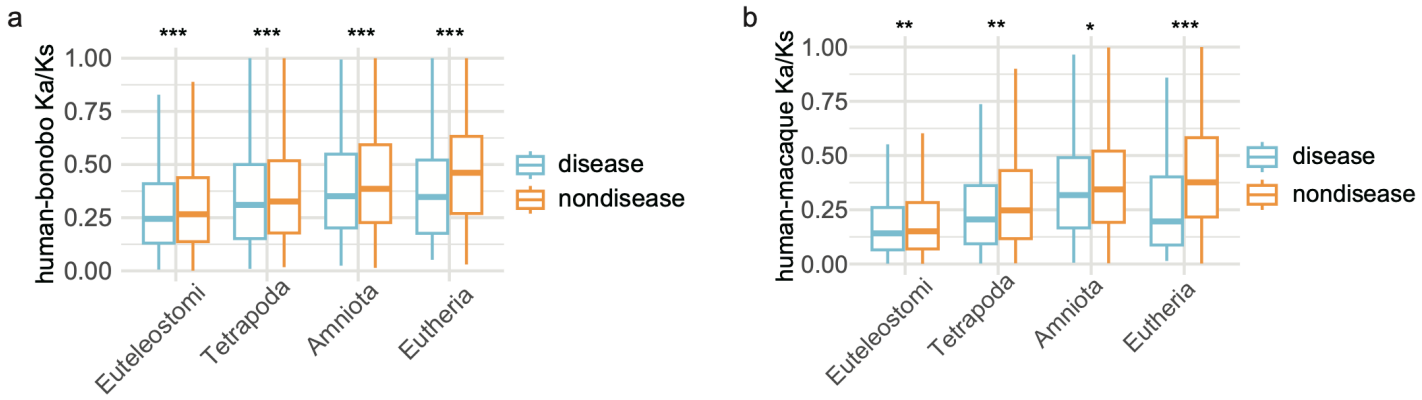


Supplementary Figure 1. The relationship between gene-wise DNVs burden from 68,404 individuals (Zhao et al. 2020) and the previously reported synteny-based gene age (Shao et al. 2019) (**a**) and gene-family based gene age (Neme and Tautz 2013) (**b**). Note: the linear models are between median values and ranked ages.

Purifying selection intensifies with gene age and is stronger in disease genes than in non-disease genes.

To understand 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). Values of Ka/Ks ratio less than 1 suggest the degree of evolutionary constraint (acting against change) (Yang and Bielawski 2000). 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 1g and Supplementary Figure 2), which is consistent with some previous studies (Carvunis et al. 2012;

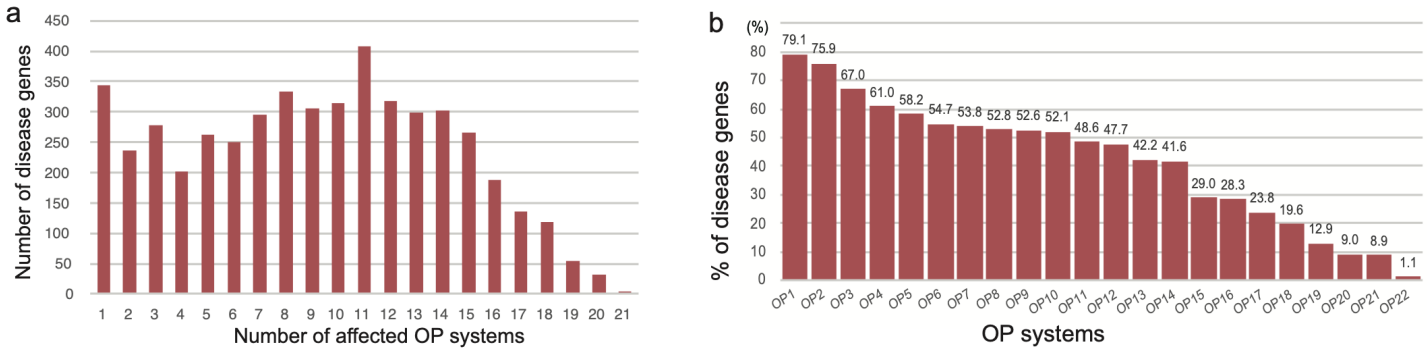
Prabh et al. 2018; Montañés et al. 2023). Notably, despite the relaxation of purifying selection for younger genes, disease genes still tend to show lower Ka/Ks ratio than non-disease genes, suggesting a general pattern of stronger purifying selection in disease genes during evolutionary process.



Supplementary Figure 2. 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.

We observed a heterogeneous distribution of disease genes underlying 22 HPO-defined anatomical systems, suggesting varied genetic complexity for diseases of different systems (Supplementary Figure 3a). None of 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, Supplementary table 2). The nervous system displayed the highest fraction of diseases genes (79%, Supplementary Figure 3b). 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 (Hodgkin 1998; Paul 2000; Lobo 2008). 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 (Pearson correlation coefficient = -0.83, p = 0.0053). This implies that disease genes with higher pleiotropy, which impact multiple anatomical systems, face more stringent evolutionary constraints compared to genes with lower pleiotropy (Figure 1h).



Supplementary Figure 3. The distributions of disease genes across the affected OP systems. (**a**) The distribution of OP system numbers on which disease gene defects could affect. The horizontal axis is the numbers of affected OP systems, and the vertical axis is the number of genes. (**b**) The percentages of disease genes out of all disease genes for different OP systems (OP1 to OP22). The definition of OP systems is consistent with Figure 1b.

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, for genes arising in lineages from Euteleostomi to Primate.

Using the recently reported average human generation time of 26.9 years (Wang et al. 2023), the most updated number of coding genes (19,831 based on Ensembl v110), and assuming the simplified monogenic model (Richards et al. 2015), 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 (Stolk et al. 2006). 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 (Kaessmann 2010; Chen et al. 2013; Xia et al. 2021).

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 (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 3a) is similar with the distribution of gene expression breath across tissues (Supplementary Figure 4a-4c). 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 4d), which also aligns with previously 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 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 over time. We applied a logistic growth function and observed a significant pattern: as evolutionary time increases, the level of pleiotropy rises (Figure 2d, $p < 0.001$). Moreover, the logistic model demonstrates a diminishing marginal growth for pleiotropy over time, indicating that the rate of increase in pleiotropy slows down over evolutionary time. This pattern suggests that pleiotropy is initially lower in new genes but increases at a faster rate compared to older genes. This result is also consistent with the finding that purifying selection gradually increase over evolutionary time from young to old (Figure 1g), which could limit the space of pleiotropy growth.

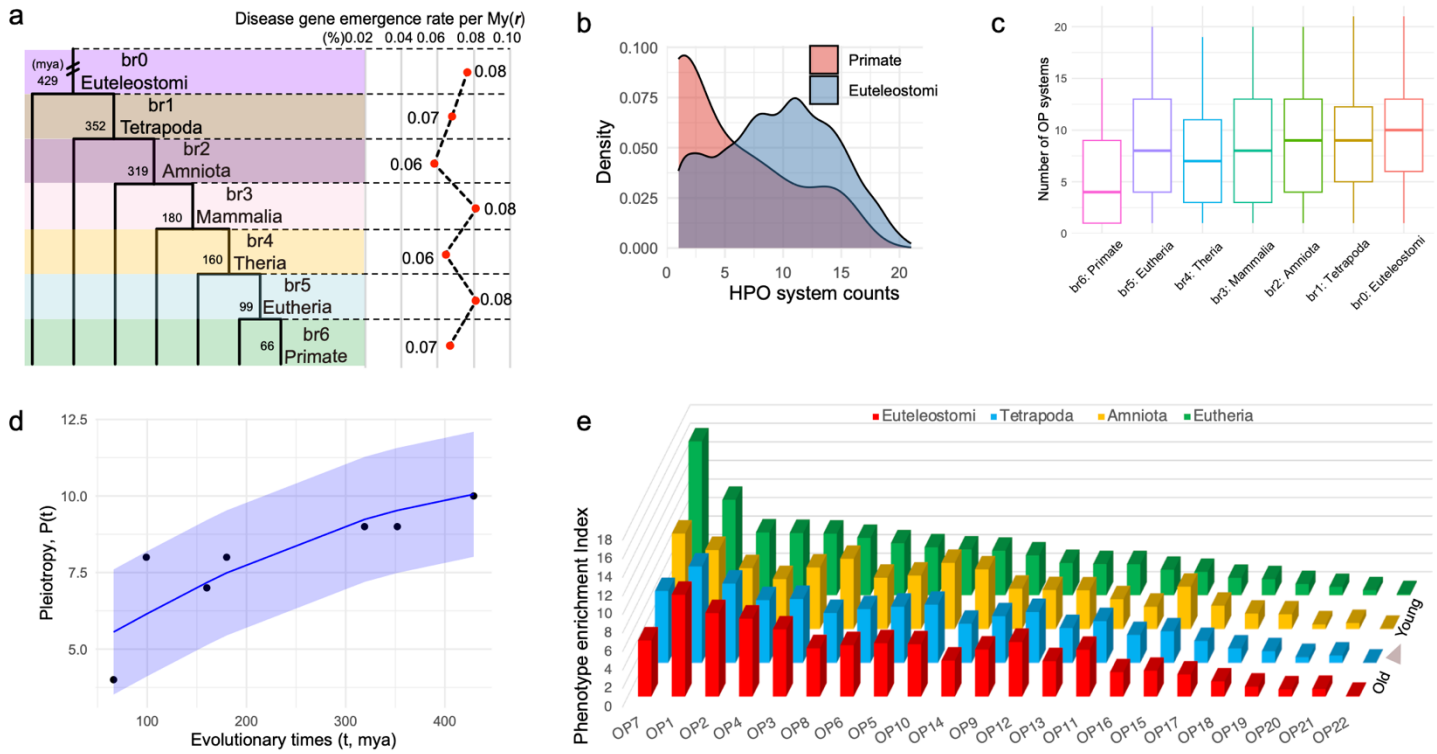


Figure 2. Disease gene emergence rates, phenotypic system coverage, and disease phenotype enrichment index (PEI) along evolutionary age groups. (**a**) The disease-gene emergence rate per million years (*r*) across evolutionary epochs. (**b**) Density distributions showcase the numbers of affected organ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

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

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 (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 OP7 showing a 38.65% higher PEI than OP1 (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 (Kaessmann 2010), 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, their expression patterns tend to broaden, potentially leading to 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 (Zhang et al. 2011). 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.

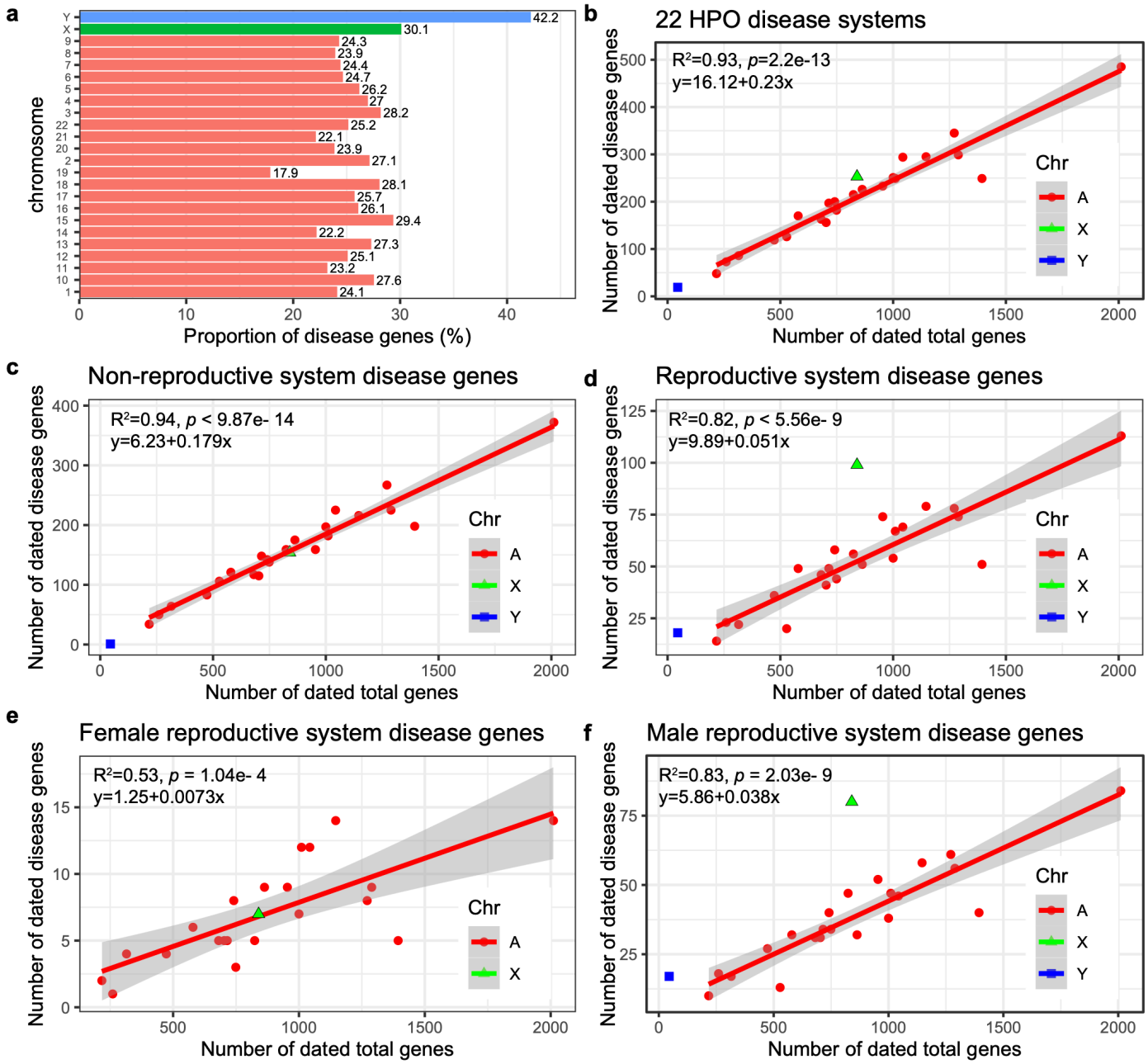


Figure 3. (**a**) The proportions of disease genes across chromosomes. The pink bars represent the autosomes, while green and blue indicate the X and Y chromosomes, respectively. The proportions (%) for different chromosomes are shown above bars. (**b**) The linear regression plotting of disease gene counts against the numbers of total genes with age information on chromosomes. (**c**) Numbers 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**) 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 dated genes 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 displayed 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.

Sex chromosomes are enriched for male-reproductive disease genes: the male hemizygous effect.

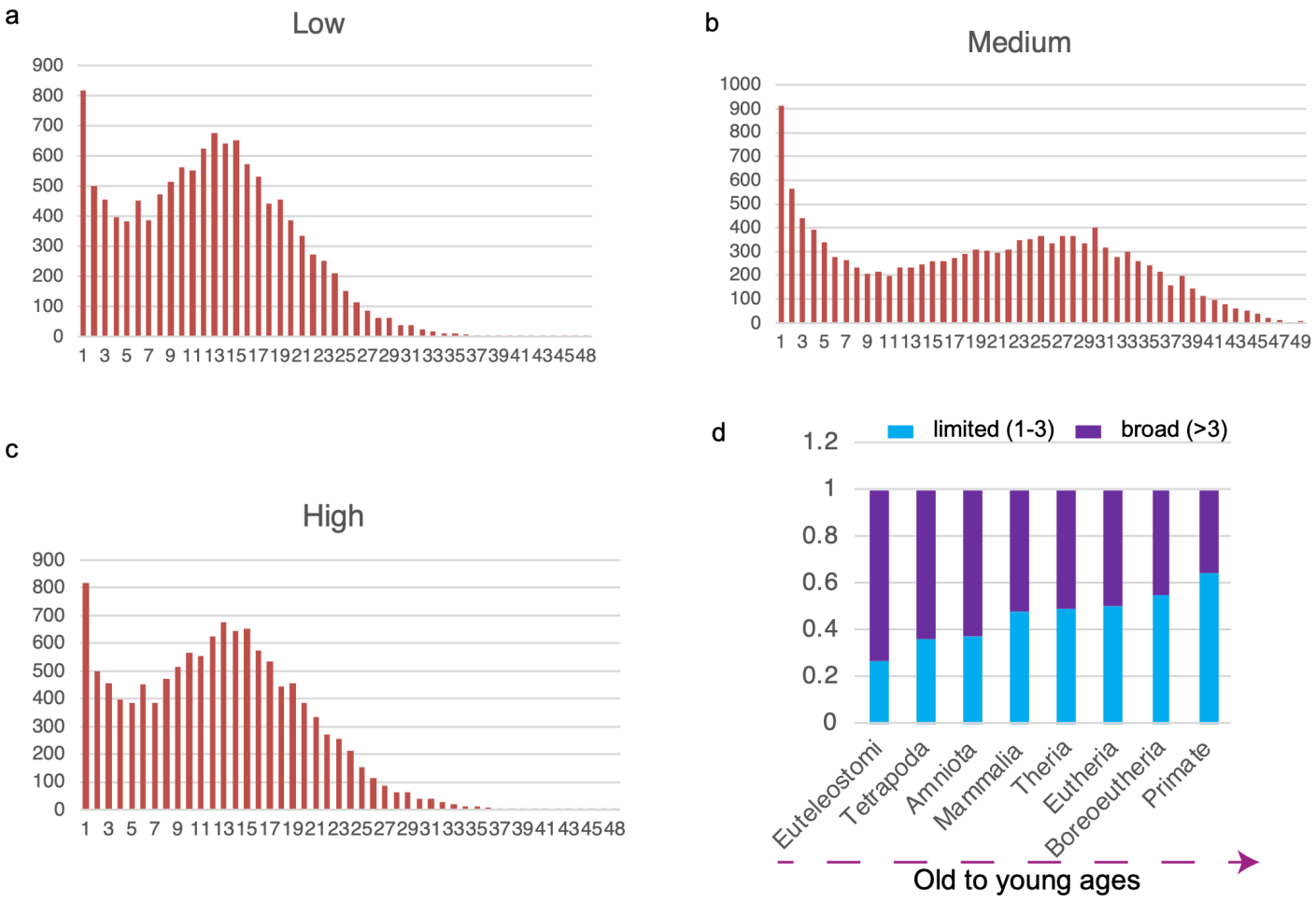
Considering the rapid concentration of the youngest disease genes in the reproductive system (Figure 2e, OP7), we hypothesized that disease genes could have skewed chromosomal distributions. First, we examined the distribution of all disease genes and found a distinct, uneven spread across chromosomes (Figure 3a and Supplementary Table 4). The X and Y chromosomes contain a higher number of disease genes compared to the autosomal chromosomes. While autosomes have a linear slope of 0.23 (Figure 3b, $R^2 = 0.93$; $p = 2.2 \times 10^{-13}$), 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 based on affected organs (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, on the X chromosome, disease genes affecting non-reproductive systems were slightly fewer than expected (-1.65% excess rate, with 154 observed versus 156.59 expected). 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).

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 (Rice 1984; Charlesworth et al. 1987). 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 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.

The genome-wide excess of male reproductive disease genes: the "faster-X" and "faster-male" effects.

To determine which gender (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. 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 6 and 7). 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, for male reproductive disease genes, the X and Y chromosomes stood out compared to autosomes, which followed a linear pattern (R^2 = 0.82, p = 5.56e-9, Figure 3f). The X chromosome contained 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, male reproductive genes outnumbered female ones by a factor of 10.43 (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 (Rice 1984; Charlesworth et al. 1987). Interestingly, a male bias was also observed 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) (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 (Wu and Davis 1993).



Supplementary Figure 4. 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 (only "HIGH" expression genes based on HPA are used).

Excess of male reproductive disease genes at younger regions of X-chromosome.

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 (25.42%) but a deficiency of young genes (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 (Zhang et al. 2010). 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 (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 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 male-biased sexual selection have less evolutionary constraints than older ones (median Ka/Ks ratios 0.35 vs. 0.23).

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 (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 (Ross et al. 2005; McLysaght 2008; Pandey et al. 2013). 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 8). Moreover, analyses of the evolutionary strata, which relies on substitutions method (Lahn and Page 1999a) 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 (Figure 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.

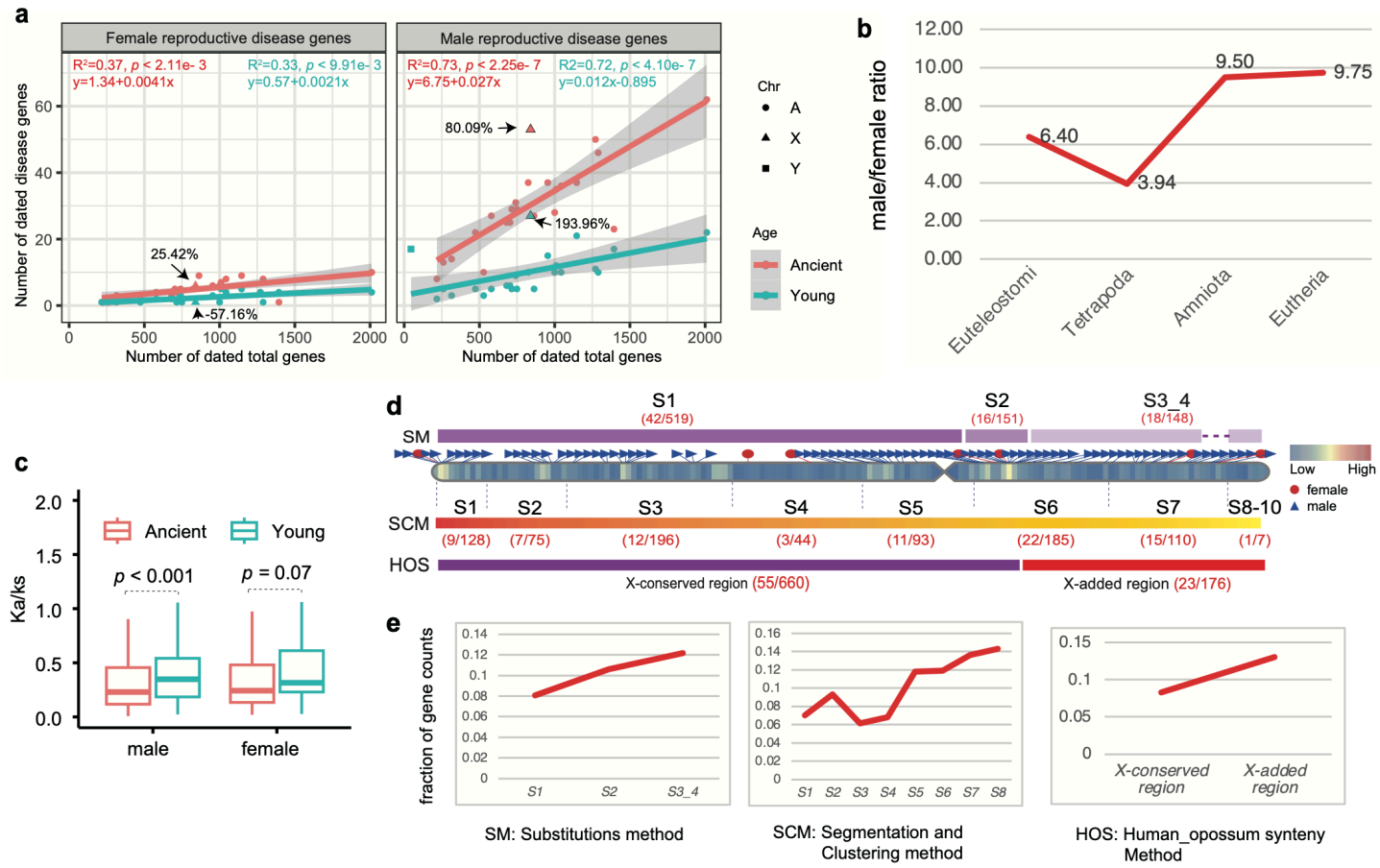


Figure 4. (**a**) Numbers of female-specific (left) and male-specific reproductive disease genes (right) are plotted against all protein-coding genes with gene ages on chromosomes. The linear formulas fitted for autosomal genes at ancient (Euteleostomi) and younger (post-Euteleostomi) stages are shown in red and blue, respectively. (**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 (SM), the segmentation and clustering method (SCM), and the synteny method (orthologous gene order conservation between human and opossum, HOS). (**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.

Discussion

The 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 (Gusella et al. 1983), 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 (Asimit et al. 2012; Boycott et al. 2013; Henn et al. 2015; Krumm et al. 2015; Ji et al. 2016; Study 2017; Krausz and Riera-Escamilla 2018; Wen et al. 2018; Almlöf et al. 2019; Povysil et al. 2019; Thuresson et al. 2019; Guo et al. 2021; Oud et al. 2021). In our study, we utilized the latest disease gene and clinical phenotype data from HPO annotations (Köhler et al. 2018) and incorporated synteny-based gene age dating to account for new gene duplication events (Shao et al. 2019). Our synteny-based gene age dating reveals that younger genes have lower percentages of disease genes than older genes. This result is consistent with the pattern observed in the previous study (Domazet-Lošo and Tautz 2008). By utilizing DNVs and rare variants from large-scale datasets of exomic and genomic sequencing aggregated in recent years (Wang et al. 2022; Greene et al. 2023), we found that evolutionary older genes tend to have higher gene-wise DNVs burden. Logistic regression modeling indicates that protein length, gene age, and DNVs burden are positively correlated with the probability of genes as being disease genes. Thus, our finding suggests that the overrepresentation of disease genes in older evolutionary age groups could be due to the joint effect of deleterious variant burden, gene length, and gene age over evolutionary time under various forms of selection. Despite previous debates on the selective pressure of disease genes (Smith and Eyre-Walker 2003; Domazet-Lošo and Tautz 2008; Chakraborty et al. 2016; Spataro et al. 2017), 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 (Haldane 1937; Keightley 2012).

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” (Rice 1984; Charlesworth et al. 1987), 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 (Migeon 2020). 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 (Wu and Davis 1993). 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 (Kaessmann 2010), 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 (Vinckenbosch et al. 2006; Bekpen et al. 2018). The “out of testis” hypothesis predicts that genes could gain broader expression patterns and higher phenotypic complexity over evolutionary time (Vinckenbosch et al. 2006). 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 (Betrán et al. 2002; Heinen et al. 2009), in both *Drosophila* (Heinen et al. 2009; Gubala et al. 2017; VanKuren and Long 2018) and mammals (Emerson et al. 2004; Jiang et al. 2017). 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) (Emerson et al. 2004; Zhang et al. 2010; Jiang et al. 2017).

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 (Zhang et al. 2011). The brain size enlargement, especially the neocortex expansion over ~50% the volume of the human brain, ranks among the most extraordinary human phenotypic innovations (Rakic 2009; Zhang et al. 2011). 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 (Pirozzi et al. 2018; Lerner et al. 2020; van Schie et al. 2020; Ma et al. 2022). The *NOTCH2NLA*, *NOTCH2NLB,* and *NOTCH2NLC* may promote human brain size enlargement, due to their functions in neuronal intranuclear inclusion disease (NIID), microcephaly, and macrocephaly (Fiddes et al. 2018; Suzuki et al. 2018; Liu et al. 2022). The *RRP7A* is also a microcephaly disease gene evidenced from patient-derived cells with defects in cell cycle progression and primary cilia resorption

(Farooq et al. 2020). The defects of *SMPD4* can lead to a neurodevelopmental disorder characterized by microcephaly and structural brain anomalies (Magini et al. 2019). The *SRGAP2C* accounts for human-specific feature of neoteny and can promote motor and execution skills in mouse and monkey model (Charrier et al. 2012; Dennis et al. 2012; Meng et al. 2023). The *de novo* gene *SMIM45* (Delihas 2023) associates with cortical expansion based on extensive models (An et al. 2023).

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 (Fritsche et al. 2016), *SMPD4* with the retinopathy (Smits et al. 2023), *TUBA3D* with the keratoconus (Hao et al. 2017), *OPN1MW* with loss of color vision (Winderickx et al. 1992; Ueyama et al. 2002), *YY1AP1* with Fibromuscular dysplasia (Guo et al. 2017), *SMN2* with the Spinal Muscular Atrophy (Hahnen et al. 1996), *GH1* with defects in adult bone mass and bone loss (Dennison et al. 2004), *KCNJ18* with thyrotoxicosis complicated by paraplegia and hyporeflexia (Ryan et al. 2010), *TBX5* with the cardiac and limb defects of Holt-Oram syndrome (Basson et al. 1997; Li et al. 1997), and *DUX4* with muscular dystrophy (Lemmers et al. 2012). 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 (Di Stazio et al. 2019). The *TUBB8* defects could lead to complete cleavage failure in fertilized eggs and oocyte maturation arrest (Feng et al. 2016; Yuan et al. 2018; Yao et al. 2022). Interestingly, a previous case study on mice also shows the role of *de novo* genes on female-specific reproductive functions (Xie et al. 2019). 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 (Carroll 1995).

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” hypothesis to explain the relationship between innovation potential and gene ages (Figure 5a). The recognition of pleiotropy

is as early as biology itself as a research field (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 syndromes had already described syndromes with different symptoms and a single "familial" factor (Eckman 1788). It is established that young genes exhibit higher specificity and narrower expression breadth across tissues (Zhang et al. 2012). In this study, we used a broader definition of pleiotropy to understand phenotype evolution (Pyeritz 1989; Lobo 2008; Tyler et al. 2009; Zhang 2023). 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') (Zeng and Hill 1986; Baatz and Wagner 1997; Orr 2000; Wagner and Zhang 2011; Fraïsse et al. 2018; Quiver and Lachance 2022), while low pleiotropy could foster morphological evolution (Carroll 2005; Wray 2007). The inhibitory effect of pleiotropy on novel adaptation aligns with our observations of the strong purifying selection on both high extent of pleiotropy (Wagner and Zhang 2011; Quiver and Lachance 2022) and expression breadth (Zhu et al. 2008). As expected, we observed that multi-system genes and older genes, which exhibit higher pleiotropy, undergo stronger purifying selection (Figure 1g-1h). 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 (Charlesworth et al. 1993). The evolution of new genes, especially gene duplicates, serves as a primary mechanism to mitigate pleiotropic effects through subfunctionalization and neofunctionalization (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 the adaptive conflicts caused by the multifunctionality of parental genes (Des Marais and Rausher 2008). The reduced pleiotropy in young genes could thereby allow for a more diverse mutational space for functional innovation without triggering unintended pleiotropic trade-offs (Dezső et al. 2008).

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. Notably, high pleiotropy could also hamper the resolution of sexual conflict through the sex-limited expression (Judith E. Mank et al. 2008). Thus, the “pleiotropy-barrier” model also applies to the evolution of new genes with male-specific functions under sexual selection (VanKuren and Long 2018). Therefore, new or 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.

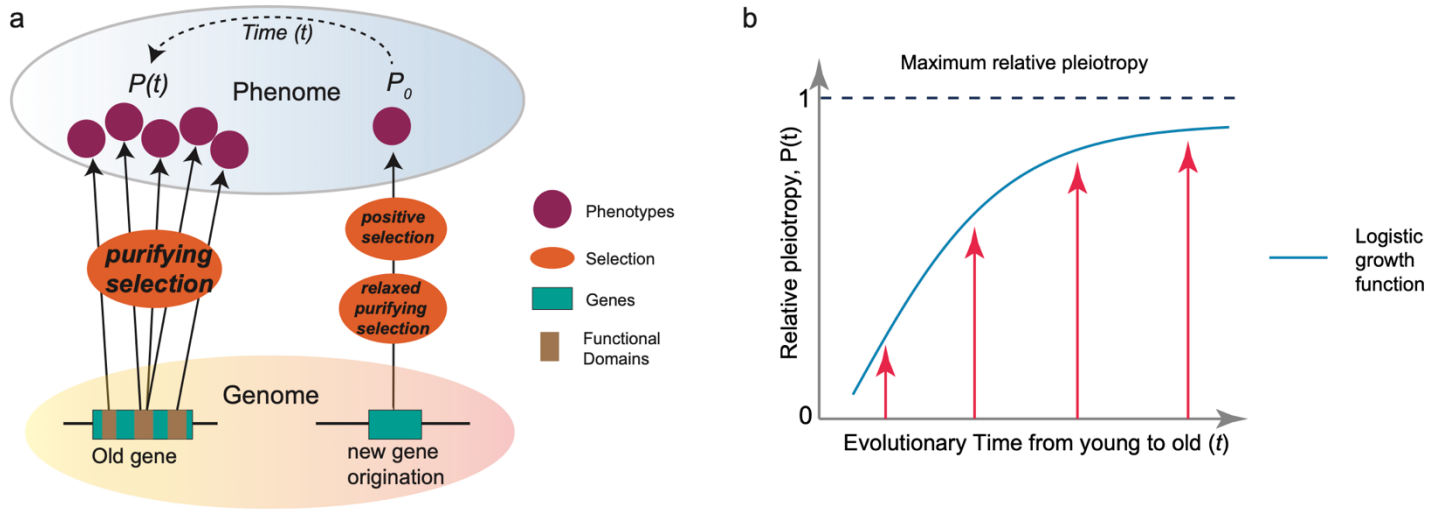


Figure 5. (**a**) The 'pleiotropy-barrier model' posits that new genes evolve adaptively more quickly. It suggests that older genes undergo stronger purifying selection because their multiple functions (usually adverse pleiotropy) act as a barrier to the uptake of mutations that might otherwise be beneficial 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.

Conclusion

By combining human gene age dating and Mendelian disease phenotyping, we revealed an increase trend of disease gene proportions over evolutionary time. This growth pattern could be due to higher gene-wise deleterious variant burdens in older genes. We found that the ratio of genes associated with health-related functions per million years remains relatively consistent across macroevolutionary epochs. 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 (Zhang et al. 2010; Shao et al. 2019) that 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 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 (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 time 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 under-estimation 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 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) (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 inference 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 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).

Logistic regression modeling and model comparison

We retrieved the burden of de novo germline variants from multiple studies, including the Gene4Denovo database (68,404 individuals) (Zhao et al. 2020), UK Biobank exomes (394,783 individuals), and gnomAD database (V4.0, 76,215 individuals) (Wang et al. 2022; Greene et al. 2023). The genotypes and allele frequencies were downloaded from gnomAD (https://gnomad.broadinstitute.org/downloads). Rare variants were extracted based on 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 the effects of multiple predictors and their interactions on the outcome of gene disease states. The disease and non-disease genes were assigned into binary states ("1", disease genes; "0", non-disease genes) as 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), DNVs burden (D) (Wang et al. 2022), and rare variant burden (R) (Chen et al. 2024b). The likelihood ratio test (LRT) and Akaike information criterion (AIC) were used for model comparison (Table 1). 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, the models with logarithm treatment were also incorporated and compared based on AIC criteria. The Variance Inflation Factor (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).

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) (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 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/) (Kumar et al. 2017).

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 (Ross et al. 2005). The "evolutionary strata" on X were based on previous reports of two methods: substitutions method and the Segmentation and Clustering method (Lahn and Page 1999b; McLysaght 2008; Pandey et al. 2013). 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, Dr. Urs C. Schmidt-Ott, and Dr. Greg Dwyer 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, 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.

Data availability: The gene age dating and disease gene phenotyping data were listed in the Supplementary Table. Other important data, including the files of rare variants and allele frequencies across global human populations (76,215 individuals) and the R script for modeling, could be retrieved from Zenodo (https://zenodo.org/uploads/11000269).

References

Almlöf JC, Nystedt S, Leonard D, Eloranta M-L, Grosso G, Sjöwall C, Bengtsson AA, Jönsen A, Gunnarsson I, Svenungsson E. 2019. Whole-genome sequencing identifies complex contributions to genetic risk by variants in genes causing monogenic systemic lupus erythematosus. *Human genetics* **138**: 141-150.

An NA, Zhang J, Mo F, Luan X, Tian L, Shen QS, Li X, Li C, Zhou F, Zhang B et al. 2023. De novo genes with an lncRNA origin encode unique human brain developmental functionality. *Nature Ecology & Evolution* **7**: 264-278.

and TAK, Bebenek K. 2000. DNA Replication Fidelity. *Annual Review of Biochemistry* **69**: 497-529.

Antonarakis SE, Beckmann JS. 2006. Mendelian disorders deserve more attention. *Nature Reviews Genetics* **7**: 277-282.

Asimit JL, Day-Williams AG, Morris AP, Zeggini E. 2012. ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. *Human heredity* **73**: 84-94.

Baatz M, Wagner GP. 1997. Adaptive Inertia Caused by Hidden Pleiotropic Effects. *Theoretical Population Biology* **51**: 49-66.

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

Barton NH. 1990. Pleiotropic models of quantitative variation. *Genetics* **124**: 773-782.

Basson CT, Bachinsky DR, Lin RC, Levi T, Elkins JA, Soults J, Grayzel D, Kroumpouzou E, Traill TA, Leblanc-Straceski J et al. 1997. Mutations in human TBX5 [corrected] cause limb and cardiac malformation in Holt-Oram syndrome. *Nat Genet* **15**: 30-35.

Bekpen C, Xie C, Tautz D. 2018. Dealing with the adaptive immune system during de novo evolution of genes from intergenic sequences. *BMC Evolutionary Biology* **18**: 121.

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

Betrán E, Long M. 2022. Evolutionary New Genes in a Growing Paradigm. *Genes* **13**: 1605.

Betrán E, Thornton K, Long M. 2002. Retroposed New Genes Out of the X in Drosophila. *Genome Research* **12**: 1854-1859.

Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. 2013. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics* **14**: 681-691.

Brosius J. 1991. Retroposons--seeds of evolution. *Science* **251**: 753.

Cai J, Zhao R, Jiang H, Wang W. 2008. De Novo Origination of a New Protein-Coding Gene in Saccharomyces cerevisiae. *Genetics* **179**: 487-496.

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.

Carroll SB. 1995. Homeotic genes and the evolution of arthropods and chordates. *Nature* **376**: 479-485.

Carroll SB. 2005. Evolution at two levels: on genes and form. *PLoS biology* **3**: e245.

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.

Chakraborty S, Panda A, Ghosh TC. 2016. Exploring the evolutionary rate differences between human disease and non-disease genes. *Genomics* **108**: 18-24.

Charlesworth B, Coyne JA, Barton NH. 1987. The relative rates of evolution of sex chromosomes and autosomes. *The American Naturalist* **130**: 113-146.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289-1303.

Charrier C, Joshi K, Coutinho-Budd J, Kim J-E, Lambert N, De Marchena J, Jin W-L, Vanderhaeghen P, Ghosh A, Sassa T. 2012. Inhibition of SRGAP2 function by its human-specific paralogs induces neoteny during spine maturation. *Cell* **149**: 923-935.

Chen J, He X, Jakovlić I. 2022. Positive selection-driven fixation of a hominin-specific amino acid mutation related to dephosphorylation in IRF9. *BMC Ecology and Evolution* **22**: 132.

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. *Journal of Medical Genetics* doi:10.1136/jmg-2024-109896: 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.

Chen S, Krinsky BH, Long M. 2013. New genes as drivers of phenotypic evolution. *Nature Reviews Genetics* **14**: 645-660.

Ciccarelli FD, von Mering C, Suyama M, Harrington ED, Izaurralde E, Bork P. 2005. Complex genomic rearrangements lead to novel primate gene function. *Genome Research* **15**: 343-351.

Clausius R. 1879. *The mechanical theory of heat*. Macmillan.

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.

Conrad B, Antonarakis SE. 2007. Gene Duplication: A Drive for Phenotypic Diversity and Cause of Human Disease. *Annual Review of Genomics and Human Genetics* **8**: 17-35.

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.

Delihas N. 2023. Evolutionary formation of a human de novo open reading frame from a mouse non-coding DNA sequence via biased random mutations. *bioRxiv* doi:10.1101/2023.08.09.552707: 2023.2008.2009.552707.

Dennis MY, Nuttle X, Sudmant PH, Antonacci F, Graves TA, Nefedov M, Rosenfeld JA, Sajjadian S, Malig M, Kotkiewicz H. 2012. Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**: 912-922.

Dennison EM, Syddall HE, Rodriguez S, Voropanov A, Day IN, Cooper C. 2004. Polymorphism in the growth hormone gene, weight in infancy, and adult bone mass. *J Clin Endocrinol Metab* **89**: 4898-4903.

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

Dezső Z, Nikolsky Y, Sviridov E, Shi W, Serebriyskaya T, Dosymbekov D, Bugrim A, Rakhmatulin E, Brennan RJ, Guryanov A. 2008. A comprehensive functional analysis of tissue specificity of human gene expression. *BMC biology* **6**: 1-15.

Di Stazio M, Collesi C, Vozzi D, Liu W, Myers M, Morgan A, D′Adamo PA, Girotto G, Rubinato E, Giacca M et al. 2019. TBL1Y: a new gene involved in syndromic hearing loss. *European Journal of Human Genetics* **27**: 466-474.

Ding D, Nguyen TT, Pang MYH, Ishibashi T. 2021. Primate-specific histone variants. *Genome* **64**: 337-346.

Domazet-Lošo T, Tautz D. 2008. An ancient evolutionary origin of genes associated with human genetic diseases. *Molecular biology and evolution* **25**: 2699-2707.

Eckman OJ. 1788. The description and multiple causes of osteomalaciae sistens.

Emerson J, Kaessmann H, Betrán E, Long M. 2004. Extensive gene traffic on the mammalian X chromosome. *Science* **303**: 537-540.

Farooq M, Lindbæk L, Krogh N, Doganli C, Keller C, Mönnich M, Gonçalves AB, Sakthivel S, Mang Y, Fatima A et al. 2020. RRP7A links primary microcephaly to dysfunction of ribosome biogenesis, resorption of primary cilia, and neurogenesis. *Nat Commun* **11**: 5816.

Feng R, Sang Q, Kuang Y, Sun X, Yan Z, Zhang S, Shi J, Tian G, Luchniak A, Fukuda Y et al. 2016. Mutations in TUBB8 and Human Oocyte Meiotic Arrest. *New England Journal of Medicine* **374**: 223-232.

Fetro C, Scherman D. 2020. Drug repurposing in rare diseases: Myths and reality. *Therapies* **75**: 157-160.

Fiddes IT, Lodewijk GA, Meghan M, Bosworth CM, Ewing AD, Mantalas GL, Novak AM, Anouk VDB, Alex B, Rosenkrantz JL. 2018. Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell* **173**: 1356-1369.e1322.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S. 2014. Ensembl 2014. *Nucleic acids research* **42**: D749-D755.

Fraïsse C, Puixeu Sala G, Vicoso B. 2018. Pleiotropy Modulates the Efficacy of Selection in Drosophila melanogaster. *Molecular Biology and Evolution* **36**: 500-515.

Fritsche LG Igl W Bailey JNC Grassmann F Sengupta S Bragg-Gresham JL Burdon KP Hebbring SJ Wen C Gorski M et al. 2016. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nature genetics* **48**: 134-143.

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. *Nature Medicine* **29**: 679-688.

Gubala AM, Schmitz JF, Kearns MJ, Vinh TT, Bornberg-Bauer E, Wolfner MF, Findlay GD. 2017. The goddard and saturn genes are essential for Drosophila male fertility and may have arisen de novo. *Molecular biology and evolution* **34**: 1066-1082.

Guillaume F, Otto SP. 2012. Gene Functional Trade-Offs and the Evolution of Pleiotropy. *Genetics* **192**: 1389-1409.

Guo DC, Duan XY, Regalado ES, Mellor-Crummey L, Kwartler CS, Kim D, Lieberman K, de Vries BBA, Pfundt R, Schinzel A et al. 2017. Loss-of-Function Mutations in YY1AP1 Lead to Grange Syndrome and a Fibromuscular Dysplasia-Like Vascular Disease. *Am J Hum Genet* **100**: 21-30.

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.

Guo T, Tu C-F, Yang D-H, Ding S-Z, Lei C, Wang R-C, Liu L, Kang X, Shen X-Q, Yang Y-F et al. 2021. Bi-allelic BRWD1 variants cause male infertility with asthenoteratozoospermia and likely primary ciliary dyskinesia. *Human Genetics* **140**: 761-773.

Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY et al. 1983. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**: 234-238.

Hahnen E, Schönling J, Rudnik-Schöneborn S, Zerres K, Wirth B. 1996. 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* **59**: 1057-1065.

Haldane J. 1937. The effect of variation of fitness. *The American Naturalist* **71**: 337-349.

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. *Nature Communications* **11**: 1842.

Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. 2000. Online Mendelian inheritance in man (OMIM). *Human mutation* **15**: 57-61.

Hao XD, Chen P, Zhang YY, Li SX, Shi WY, Gao H. 2017. De novo mutations of TUBA3D are associated with keratoconus. *Sci Rep* **7**: 13570.

He X, Zhang J. 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**: 1157-1164.

Heinen TJ, Staubach F, Häming D, Tautz D. 2009. Emergence of a new gene from an intergenic region. *Current biology* **19**: 1527-1531.

Henn BM, Botigué LR, Bustamante CD, Clark AG, Gravel S. 2015. Estimating the mutation load in human genomes. *Nature Reviews Genetics* **16**: 333-343.

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.

INSERM. 1997. Orphanet: an online database of rare diseases and orphan drugs. French National Institute for Health and Medical Research Paris.

Ji X, Kember RL, Brown CD, Bućan M. 2016. Increased burden of deleterious variants in essential genes in autism spectrum disorder. *Proceedings of the National Academy of Sciences* **113**: 15054-15059.

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. *Clinical Genetics* **103**: 79-86.

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. 2021. Deviation from baseline mutation burden provides powerful and robust rare-variants association test for complex diseases. *Nucleic Acids Research* **50**: e34-e34.

Jiang L, Li T, Zhang X, Zhang B, Yu C, Li Y, Fan S, Jiang X, Khan T, Hao Q. 2017. RPL10L is required for male meiotic division by compensating for RPL10 during meiotic sex chromosome inactivation in mice. *Current Biology* **27**: 1498-1505. e1496.

Judith E. Mank, Lina Hultin-Rosenberg, Martin Zwahlen, Hans Ellegren. 2008. Pleiotropic Constraint Hampers the Resolution of Sexual Antagonism in Vertebrate Gene Expression. *The American Naturalist* **171**: 35-43.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome research* **20**: 1313-1326.

Kaessmann H, Zöllner S, Nekrutenko A, Li WH. 2002. Signatures of domain shuffling in the human genome. *Genome Res* **12**: 1642-1650.

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.

Keightley PD. 2012. Rates and fitness consequences of new mutations in humans. *Genetics* **190**: 295-304.

Kingdom R, Beaumont RN, Wood AR, Weedon MN, Wright CF. 2024. Genetic modifiers of rare variants in monogenic developmental disorder loci. *Nature Genetics* 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.

Köhler S, Carmody L, Vasilevsky N, Jacobsen JO B, Danis D, Gourdine J-P, Gargano M, Harris NL, Matentzoglu N, McMurry JA et al. 2018. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Research* **47**: D1018-D1027.

Krausz C, Riera-Escamilla A. 2018. Genetics of male infertility. *Nature Reviews Urology* **15**: 369-384.

Krumm N, Turner TN, Baker C, Vives L, Mohajeri K, Witherspoon K, Raja A, Coe BP, Stessman HA, He Z-X et al. 2015. Excess of rare, inherited truncating mutations in autism. *Nature Genetics* **47**: 582-588.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Molecular biology and evolution* **34**: 1812-1819.

Lahn BT, Page DC. 1999a. Four evolutionary strata on the human X chromosome. *Science* **286**: 964-967.

Lahn BT, Page DC. 1999b. Four evolutionary strata on the human X chromosome. *Science* **286**: 964-967.

Lee S, Abecasis Gonçalo R, Boehnke M, Lin X. 2014. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *The American Journal of Human Genetics* **95**: 5-23.

Lemmers RJ, Tawil R, Petek LM, Balog J, Block GJ, Santen GW, Amell AM, van der Vliet PJ, Almomani R, Straasheijm KR et al. 2012. Digenic inheritance of an SMCHD1 mutation and an FSHD-permissive D4Z4 allele causes facioscapulohumeral muscular dystrophy type 2. *Nat Genet* **44**: 1370-1374.

Lerner LK, Holzer S, Kilkenny ML, Šviković S, Murat P, Schiavone D, Eldridge CB, Bittleston A, Maman JD, Branzei D et al. 2020. Timeless couples G-quadruplex detection with processing by DDX11 helicase during DNA replication. *The EMBO Journal* **39**: e104185.

Li D, Dong Y, Jiang Y, Jiang H, Cai J, Wang W. 2010. A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Research* **20**: 408-420.

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**.

Li QY, Newbury-Ecob RA, Terrett JA, Wilson DI, Curtis AR, Yi CH, Gebuhr T, Bullen PJ, Robson SC, Strachan T et al. 1997. Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family. *Nat Genet* **15**: 21-29.

Liu Q, Zhang K, Kang Y, Li Y, Deng P, Li Y, Tian Y, Sun Q, Tang Y, Xu K et al. 2022. Expression of expanded GGC repeats within NOTCH2NLC causes behavioral deficits and neurodegeneration in a mouse model of neuronal intranuclear inclusion disease. *Science Advances* **8**: eadd6391.

Lobo I. 2008. Pleiotropy: one gene can affect multiple traits.

Long M, VanKuren NW, Chen S, Vibranovski MD. 2013. New Gene Evolution: Little Did We Know. *Annual Review of Genetics* **47**: 307-333.

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 Biology* **23**: 251.

Magini P, Smits DJ, Vandervore L, Schot R, Columbaro M, Kasteleijn E, van der Ent M, Palombo F, Lequin MH, Dremmen M et al. 2019. Loss of SMPD4 Causes a Developmental Disorder Characterized by Microcephaly and Congenital Arthrogryposis. *Am J Hum Genet* **105**: 689-705.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of Young Human Genes after a Burst of Retroposition in Primates. *PLOS Biology* **3**: e357.

McLysaght A. 2008. Evolutionary steps of sex chromosomes are reflected in retrogenes. *Trends in genetics* **24**: 478-481.

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

Meng X, Lin Q, Zeng X, Jiang J, Li M, Luo X, Chen K, Wu H, Hu Y, Liu C et al. 2023. Brain developmental and cortical connectivity changes in the transgenic monkeys carrying the human-specific duplicated gene srGAP2C. *National Science Review* doi:10.1093/nsr/nwad281.

Migeon BR. 2020. X-linked diseases: susceptible females. *Genetics in Medicine* **22**: 1156-1174.

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* **13**: 968.

Montañés JC, Huertas M, Messeguer X, Albà MM. 2023. Evolutionary Trajectories of New Duplicated and Putative De Novo Genes. *Molecular Biology and Evolution* **40**.

Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent de novoevolution. *BMC Genomics* **14**: 117.

Orr HA. 2000. Adaptation and the cost of complexity. *Evolution* **54**: 13-20.

Oud MS, Houston BJ, Volozonoka L, Mastrorosa FK, Holt GS, Alobaidi BKS, deVries PF, Astuti G, Ramos L, Mclachlan RI et al. 2021. Exome sequencing reveals variants in known and novel candidate genes for severe sperm motility disorders. *Human Reproduction* **36**: 2597-2611.

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 biology and evolution* **5**: 1863-1871.

Parikh SB, Houghton C, Van Oss SB, Wacholder A, Carvunis A-R. 2022. Origins, evolution, and physiological implications of de novo genes in yeast. *Yeast* **39**: 471-481.

Paul D. 2000. A double-edged sword. *Nature* **405**: 515-515.

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.

Pirozzi F, Nelson B, Mirzaa G. 2018. From microcephaly to megalencephaly: determinants of brain size. *Dialogues Clin Neurosci* **20**: 267-282.

Povysil G, Petrovski S, Hostyk J, Aggarwal V, Allen AS, Goldstein DB. 2019. Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nature Reviews Genetics* **20**: 747-759.

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 Research* **28**: 1664-1674.

Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, O'dushlaine C, Chambert K, Bergen SE, Kähler A. 2014. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* **506**: 185-190.

Pyeritz RE. 1989. Pleiotropy revisited: Molecular explanations of a classic concept. *American Journal of Medical Genetics* **34**: 124-134.

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.

Ragsdale EJ, Müller MR, Rödelsperger C, Sommer RJ. 2013. A developmental switch coupled to the evolution of plasticity acts through a sulfatase. *Cell* **155**: 922-933.

Rakic P. 2009. Evolution of the neocortex: a perspective from developmental biology. *Nat Rev Neurosci* **10**: 724-735.

Rice WR. 1984. Sex chromosomes and the evolution of sexual dimorphism. *Evolution*: 735-742.

Rich A, Carvunis A-R. 2023. De novo gene increases brain size. *Nature Ecology & Evolution* **7**: 180-181.

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.

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.

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.

Ryan DP, da Silva MR, Soong TW, Fontaine B, Donaldson MR, Kung AW, Jongjaroenprasert W, Liang MC, Khoo DH, Cheah JS et al. 2010. Mutations in potassium channel Kir2.6 cause susceptibility to thyrotoxic hypokalemic periodic paralysis. *Cell* **140**: 88-98.

Shao Y, Chen C, Shen H, He BZ, Yu D, Jiang S, Zhao S, Gao Z, Zhu Z, Chen X. 2019. GenTree, an integrated resource for analyzing the evolution and function of primate-specific coding genes. *Genome research*.

Shi L, Su B. 2012. Identification and functional characterization of a primate-specific E2F1 binding motif regulating MCPH1 expression. *The FEBS Journal* **279**: 491-503.

Smith NG, Eyre-Walker A. 2003. Human disease genes: patterns and predictions. *Gene* **318**: 169-175.

Smits DJ, Schot R, Krusy N, Wiegmann K, Utermöhlen O, Mulder MT, den Hoedt S, Yoon G, Deshwar AR, Kresge C et al. 2023. SMPD4 regulates mitotic nuclear envelope dynamics and its loss causes microcephaly and diabetes. *Brain* **146**: 3528-3541.

Spataro N, Rodríguez JA, Navarro A, Bosch E. 2017. Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. *Human molecular genetics* **26**: 489-500.

Stolk P, Willemen MJ, Leufkens HG. 2006. Rare essentials: drugs for rare diseases as essential medicines. *Bulletin of the World Health Organization* **84**: 745-751.

Study TDDD. 2017. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**: 433.

Suzuki IK, Gacquer D, Van Heurck R, Kumar D, Wojno M, Bilheu A, Herpoel A, Lambert N, Cheron J, Polleux F. 2018. Human-specific NOTCH2NL genes expand cortical neurogenesis through Delta/Notch regulation. *Cell* **173**: 1370-1384. e1316.

Thuresson AC, Zander CS, Zhao JJ, Halvardson J, Maqbool K, Månsson E, Stenninger E, Holmlund U, Öhrner Y, Feuk L. 2019. Whole genome sequencing of consanguineous families reveals novel pathogenic variants in intellectual disability. *Clinical genetics* **95**: 436.

Tyler AL, Asselbergs FW, Williams SM, Moore JH. 2009. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *Bioessays* **31**: 220-227.

Ueyama H, Kuwayama S, Imai H, Tanabe S, Oda S, Nishida Y, Wada A, Shichida Y, Yamade S. 2002. Novel missense mutations in red/green opsin genes in congenital color-vision deficiencies. *Biochem Biophys Res Commun* **294**: 205-209.

Van Oss SB, Carvunis AR. 2019. De novo gene birth. *PLoS Genet* **15**: e1008160.

van Schie JJM, Faramarz A, Balk JA, Stewart GS, Cantelli E, Oostra AB, Rooimans MA, Parish JL, de Almeida Estéves C, Dumic K et al. 2020. Warsaw Breakage Syndrome associated DDX11 helicase resolves G-quadruplex structures to support sister chromatid cohesion. *Nature Communications* **11**: 4287.

Vanchurin V, Wolf YI, Koonin EV, Katsnelson MI. 2022. Thermodynamics of evolution and the origin of life. *Proceedings of the National Academy of Sciences* **119**: e2120042119.

VanKuren NW, Long M. 2018. Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions. *Nature ecology & evolution* **2**: 705-712.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proceedings of the National Academy of Sciences* **103**: 3220-3225.

Wagner GP, Zhang J. 2011. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nature Reviews Genetics* **12**: 204-213.

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.

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

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. *Proceedings of the National Academy of Sciences* **119**: e2203491119.

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.

Weinreich SS, Mangon R, Sikkens J, Teeuw M, Cornel M. 2008. Orphanet: a European database for rare diseases. *Nederlands tijdschrift voor geneeskunde* **152**: 518-519.

Wen L, Zhu C, Zhu Z, Yang C, Zheng X, Liu L, Zuo X, Sheng Y, Tang H, Liang B. 2018. Exome-wide association study identifies four novel loci for systemic lupus erythematosus in Han Chinese population. *Annals of the rheumatic diseases* **77**: 417-417.

Winderickx J, Sanocki E, Lindsey DT, Teller DY, Motulsky AG, Deeb SS. 1992. Defective colour vision associated with a missense mutation in the human green visual pigment gene. *Nature Genetics* **1**: 251-256.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics* **8**: 206-216.

Wright ES. 2015. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. *BMC bioinformatics* **16**: 1-14.

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

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

Wu D-D, Irwin DM, Zhang Y-P. 2011. De Novo Origin of Human Protein-Coding Genes. *PLOS Genetics* **7**: e1002379.

Xia S, VanKuren NW, Chen C, Zhang L, Kemkemer C, Shao Y, Jia H, Lee U, Advani AS, Gschwend A et al. 2021. Genomic analyses of new genes and their phenotypic effects reveal rapid evolution of essential functions in Drosophila development. *PLOS Genetics* **17**: e1009654.

Xie C, Bekpen C, Künzel S, Keshavarz M, Krebs-Wheaton R, Skrabar N, Ullrich KK, Tautz D. 2019. A de novo evolved gene in the house mouse regulates female pregnancy cycles. *eLife* **8**: e44392.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**: 1586-1591.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution* **15**: 496-503.

Yao Z, Zeng J, Zhu H, Zhao J, Wang X, Xia Q, Li Y, Wu L. 2022. Mutation analysis of the TUBB8 gene in primary infertile women with oocyte maturation arrest. *Journal of Ovarian Research* **15**: 38.

Yuan P, Zheng L, Liang H, Li Y, Zhao H, Li R, Lai L, Zhang Q, Wang W. 2018. A novel mutation in the TUBB8 gene is associated with complete cleavage failure in fertilized eggs. *J Assist Reprod Genet* **35**: 1349-1356.

Zeng ZB, Hill WG. 1986. The selection limit due to the conflict between truncation and stabilizing selection with mutation. *Genetics* **114**: 1313-1328.

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. *Nature Ecology & Evolution* **6**: 273-287.

Zhang J. 2023. Patterns and Evolutionary Consequences of Pleiotropy. *Annual Review of Ecology, Evolution, and Systematics* **54**: 1-19.

Zhang YE, Landback P, Vibranovski M, Long M. 2012. New genes expressed in human brains: implications for annotating evolving genomes. *Bioessays* **34**: 982-991.

Zhang YE, Landback P, Vibranovski MD, Long M. 2011. Accelerated Recruitment of New Brain Development Genes into the Human Genome. *PLOS Biology* **9**: e1001179.

Zhang YE, Long M. 2014. New genes contribute to genetic and phenotypic novelties in human evolution. *Current opinion in genetics & development* **29**: 90-96.

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 biology* **8**: e1000494.

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.

Zhu J, He F, Hu S, Yu J. 2008. On the nature of human housekeeping genes. *Trends in Genetics* **24**: 481-484.

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. *Proceedings of the National Academy of Sciences* **111**: E455-E464.