Ancient regulatory evolution shapes individual language abilities in present-day humans

Lucas G. Casten[1†], Tanner Koomar[1†], Taylor R. Thomas[2], Jin-Young Koh[3], Dabney Hofammann[1], Savantha Thenuwara[4], Allison Momany[5], Marlea O'Brien[6], Jeffrey C. Murray[5], J. Bruce Tomblin[6], Jacob J. Michaelson[*1,6]

[1]Department of Psychiatry, University of Iowa, Iowa City, 52242, USA.

[2]Center for Genomic Medicine, Massachusetts General Hospital, Boston, 02114, USA.

[3]Department of Otorhinolaryngology, University of Maryland, College Park, 20742, USA.

[4]Carver College of Medicine, University of Iowa, Iowa City, 52242, USA.

[5] Stead Family Department of Pediatrics, University of Iowa, Iowa City, 52242, USA.

[6] Department of Communication Science and Disorders, University of Iowa, Iowa City, 52242, USA.

[*]Corresponding author. Email: jacob-michaelson@uiowa.edu

[†]These authors contributed equally to this work.

Language is a defining feature of our species, yet the genomic changes enabling it remain poorly understood. Despite decades of work since *FOXP2*'s discovery, we still lack a clear picture of which regions shaped language evolution and how variation contributes to present-day phenotypic differences. Using an evolutionary stratified polygenic score approach, we find that Human Ancestor Quickly Evolved Regions (HAQERs) are associated with spoken language abilities (discovery N = 350, total replication N > 100,000). HAQERs evolved before the human–Neanderthal split, giving hominins increased binding of Forkhead and Homeobox transcription factors, and show evidence of balancing selection across the past 20,000 years. Remarkably, language-associated variants in HAQERs appear more prevalent in Neanderthals and HAQER-like sequences show convergent evolution across vocal-learning mammals. Our results reveal how ancient innovations continue shaping human language.

Language is one of our species' most remarkable cognitive innovations, yet the genetic mechanisms underlying this ability remain elusive. While the human genome differs by only 1-5% from our closest primate relatives (*1–3*), these modest genetic changes enabled the evolution of our species' unique capacity for complex language. Understanding how these relatively small genomic differences produced profound cognitive differences represents a central challenge in evolutionary genetics, with implications for language disorders, human cognitive diversity, and the origins of human-specific traits (*4*).

The discovery that mutations in *FOXP2* cause speech and language disorders provided the first clear example of a single gene with significant effects on language, reinforcing early expectations for simple genetic architectures (*5, 6*). However, *FOXP2*'s contribution to typical variation in language ability proved limited, with subsequent studies failing to find associations between common *FOXP2* variants and individual differences in language skills (*7, 8*). This limitation shifted research toward polygenic models emphasizing a large number of regulatory elements scattered throughout the genome that collectively influence language development. Genome-wide association studies have since identified numerous loci contributing to reading abilities, stuttering, rhythm, and vocabulary development, supporting a highly polygenic architecture (*9–14*). Cross-species studies have revealed that language-related traits (vocal learning and rhythm) show convergent evolution

across mammalian lineages, with distributed regulatory networks rather than single genes controlling complex vocal behaviors (*15–20*). However, this polygenic model has left critical evolutionary questions unanswered: how did language-relevant regulatory elements change during human evolution, when did humans acquire language-promoting functions, and how do ancient evolutionary changes translate into present-day individual differences in language abilities?

To address these questions, we analyzed 65 million years of primate evolutionary history to trace the origins of language-relevant genetic variation. As part of our approach, we developed an evolutionary stratified polygenic score (ES-PGS) method that partitions genetic effects based on the evolutionary origins of their sequence context. We applied this approach across multiple datasets with detailed language phenotyping (discovery N = 350, total replication N > 100,000), combined with molecular analysis, ancient DNA analysis, and cross-species genomic comparisons. This multi-modal framework enabled us to directly connect ancient genomic innovations with modern individual differences in language ability, identify neurobiological mechanisms supporting language evolution, and revealed how evolutionary trade-offs have shaped human cognitive variation.

Results

Dimensions of language ability

To quantify dimensions of developmental language abilities, we analyzed 17 longitudinal cognitive and language assessments administered from kindergarten through 4th grade for 350 children sampled from a community-based cohort (*21*), which we refer to as the "EpiSLI" cohort. This analysis revealed seven factors representing distinct aspects of language ability (Figure 2A). The first factor (F1), primarily driven by spoken sentence repetition scores, represents general language ability. Sentence repetition strongly indicates overall language capacity, making F1 a key measure of general language competence (*22, 23*). The second factor (F2) relates to receptive vocabulary and listening comprehension, covering broad receptive language skills. The third factor (F3) specifically reflects nonverbal IQ, aligning with performance IQ at both kindergarten and 2nd grade. Factor F4 captures pre-literacy language skills, incorporating all kindergarten scores except performance IQ. Its slight correlation to F1 and F2 (r = 0.13 and 0.12), but not F3, suggests specificity to language

(Figure 2B). Factor F5, which we call "talkativeness", mainly reflects the number of clauses produced in a narrative task. Factor F6, based on a comprehension of concepts and directions assessment, indexes mastery of directive language (i.e., task-based instructions). Factor F7 spans a variety of assessments, with specific loading on vocabulary and grammar-related tasks, suggesting a broad, crystallized knowledge of language.

We generated high-coverage whole-genome sequencing data for all EpiSLI participants, enabling investigation of both common and rare genetic variation associated with language abilities. Most of our preliminary investigation of these factors suggested that Factors 1, 2, and 3 carried the most genetic association signal (Figure 2D, Supplementary Table 2). We also find pervasive associations with F1-F3 and measures of mental health in our sample (N = 241, Figure S1, Supplementary Table 1).

Evolutionary Stratified Polygenic Score Analysis

To investigate the genetic origins of language ability, we developed an evolutionary stratified polygenic score (ES-PGS) approach that systematically examines how genetic variants from different evolutionary periods contribute to traits. ES-PGS builds on the conceptual framework of partitioned heritability and pathway-based polygenic score methods, which have successfully partitioned genetic effects across functional genomic regions (*24–27*). The central idea is straightforward: if language capability evolved during a particular period of human evolution, then genetic changes that occurred during that period should show stronger associations with language ability today than genetic changes from other evolutionary periods. By comparing multiple evolutionary periods, we can trace the origins of language-relevant genetic variation across deep evolutionary time. For example, if regions that differentiate modern humans from Neanderthals predict language ability better than other genomic regions, this would provide evidence that genetic changes unique to anatomically modern humans were particularly important for language evolution.

Our approach works by partitioning the genome based on these evolutionary annotations and then testing whether variants within each annotation predict language ability more strongly than would be expected by chance. Critically, we compare each evolutionary annotation not only against the rest of the genome but also against carefully matched control regions that share similar genomic

properties (same size, chromosome, GC content, distance to genes, overlap with coding regions, etc.) but did not undergo the same evolutionary changes. This two-fold comparison ensures that any effects we detect are specific to the evolutionary history of these regions rather than other genomic properties. By using individual-level data rather than summary statistics, ES-PGS enables direct association testing in deeply phenotyped cohorts, which is particularly valuable for specialized studies like ours with extensive language assessments.

Human-specific genomic regions predict individual differences in language ability

To trace the evolutionary origins of language abilities we applied ES-PGS using a cognitive performance polygenic score (CP-PGS) from a large-scale study that examined both cognitive and educational traits (*28*). This polygenic score captures general cognitive function and enables valid comparisons across multiple cognitive domains, including language and nonverbal IQ. We first confirmed that the CP-PGS showed the expected associations in our EpiSLI sample, the genome-wide CP-PGS showed significant associations with both sentence repetition (F1, $r = 0.22$, FDR adjusted p-value = 0.001) and receptive language ability (F2, $r = 0.19$, FDR adjusted p-value = 0.01), (Figure 2D). Finally, we partitioned CP-PGS with across five evolutionary annotations spanning approximately 65 million years of primate and human evolution, ranging from ancient primate-conserved regions to sequences differentiating modern humans from Neanderthals, to trace which genomic regions and evolutionary periods contributed to different aspects of human cognition (*29–33*).

Human Ancestor Quickly Evolved Regions (HAQERs (*31,34*)) emerged as the most compelling finding from this analysis. Despite comprising less than 0.1% of the human genome, HAQERs showed associations with four of the seven factors (F1, F2, F4, and F6, Figure 3B, Supplementary Table 3). HAQER CP-PGS demonstrated the strongest association with sentence repetition, a measure of general language ability (F1, $r = 0.23$, ES-PGS model $\beta = 0.18$, p-value = 1.2×10^{-4}, FDR adjusted p-value = 0.004, Figures 3A-C), while showing no association with nonverbal IQ (F3, ES-PGS model $\beta = 0.06$, p-value = 0.19, FDR adjusted p-value = 0.61). HAQERs are largely non-coding sequences that began rapidly evolving after the human-chimpanzee split (approximately

6 million years ago) but before human-Neanderthal divergence (approximately 600,000 years ago), acquiring novel regulatory functions in the human lineage (*31, 34*).

The predictive power of HAQERs is striking. While the background and matched CP-PGS together utilized approximately 300,000 independent SNPs and explained 3.7% of variance in sentence repetition ability, adding HAQER CP-PGS (comprising only 1,763 independent SNPs) increased explained variance to 7.7%. This indicates that an average HAQER SNP carries 188 times more predictive power for language than SNPs elsewhere in the genome, with HAQERs alone explaining slightly more variance of sentence repetition scores in the EpiSLI cohort (R^2 gain of 4%) than the remaining >99.9% of the human genome (R^2 of 3.7%).

In contrast, Human Accelerated Regions (HARs), which are deeply conserved regulatory elements that acquired human-specific changes (*32, 35*), showed no comparable signal using ES-PGS (F1 β = -0.05, p-value = 0.27, FDR adjusted p-value = 0.7, Figure S2). This distinction suggests that human language ability emerged through novel regulatory innovations (HAQERs) rather than modifications to existing functional elements (HARs). The association between HAQERs and language factors, with no observed association with nonverbal IQ, suggests a distinct evolutionary trajectory for verbal abilities compared to general cognition. Supporting this distinction, nonverbal IQ (F3) was most strongly associated with genomic regions that underwent rapid changes across all great apes (*30*) (ES-PGS β = 0.13, p-value = 0.004, FDR adjusted p-value = 0.07, Figure S2).

HAQERs influence language ability across multiple cohorts and variant types

We validated HAQER effects on language across multiple independent cohorts using both common and rare genetic variants. In the SPARK autism dataset (N > 30,000) (*36*), HAQER CP-PGS predicted verbal language capability ("Able to talk using short phrases or sentences", ES-PGS β = 0.05, p-value = 0.008, N = 29,266) and language disorder diagnoses in parents without autism (model improvement p-value = 7.9×10^{-5}, N = 713) but not psychiatric conditions (model improvement p-value = 0.58, N = 713), confirming language specific effects (Figure 4A-B, Supplementary Table 5). Clinical records showed HAQER CP-PGS is associated with verbal IQ (ES-PGS β = 2.08, p-value = 0.022, N = 620) but not nonverbal IQ (ES-PGS β = 0.59, p-value = 0.49, N = 620, Figure S3), further supporting specificity.

To test HAQER effects through an orthogonal approach independent of our ES-PGS method, we analyzed rare genetic variation in SPARK whole genome sequencing data (N > 2,000). We examined rare "reversions", variants that revert from the human-specific version to their human-chimp ancestral state, reasoning that if HAQERs evolved to support human language, reversions should impair language function, which would support the observed polygenic score associations. Individuals carrying more reversions in HAQERs showed increased likelihood of developmental language disorder ($\beta = 0.16$, p-value = 6.5×10^{-4}) and delayed language developmental milestones, but no association with age started walking or intellectual disability (Figure 4C, Supplementary Table 6). Notably, HAQERs showed higher rates of reversions compared to HARs and matched random sequences (Figure 4D-F).

Further independent validation in the ABCD developmental cohort (*37*) showed the HAQER CP-PGS is associated with the Rey Auditory Verbal Learning Test performance, a measure of spoken word recall (ES-PGS β = 0.24, p-value = 0.048, N = 5,625, Supplementary Table 7). We also tested HAQER specificity in the UK Biobank using a word reading ability polygenic score (WR-PGS) rather than our primary cognitive performance score, as the latter included UK Biobank participants in its discovery cohort (*9, 28*). Despite the substantially smaller discovery sample for word reading GWAS (N ≈ 30,000) compared to cognitive performance GWAS (N ≈ 250,000), the HAQER WR-PGS showed significant associations with verbal working memory (ES-PGS β = 0.01, p-value = 0.002, N = 84,433), a task requiring spoken recall of number sequences shown to be relevant to language (*38*), and educational attainment (ES-PGS β = 0.004, p-value = 0.005, N = 370,830), but not with matrix reasoning (ES-PGS β = 0.005, p-value = 0.44, N = 26,051), a nonverbal fluid intelligence measure (Figure S4). The detection of HAQER effects across diverse spoken language measures despite less-than-optimal validation data provides additional support for the robustness and specificity of HAQER associations with verbal abilities. These converging results across cohorts and variant types strongly support that genetic variation within HAQERs has a significant effect on spoken language abilities in contemporary humans.

HAQERs evolved stronger binding affinity for language-relevant transcription factors

To investigate the molecular mechanisms underlying HAQERs' association with language development, we analyzed how rare genetic variants affect transcription factor binding sites in these regions. We compared two classes of rare variants in the EpiSLI cohort: hominin-chimpanzee ancestral allele reversions versus other rare variants, using position weight matrices to quantify how these variants alter predicted transcription factor binding affinity. By comparing the effects of reversions with other rare variants, we could detect systematic evolutionary changes in hominin-specific transcription factor binding associated with language phenotypes.

In HAQERs, hominin-gained transcription factor motif binding showed significant correlation with individual sentence repetition ability (N = 350 individuals, β = 0.14, p-value = 5.6×10^{-4}, Figure 5A), suggesting that hominins evolved increased transcription factor binding affinity in these regions that may support spoken language abilities. In contrast, regions under sequence conservation with human-specific changes (HARs, β = 0.01, p-value = 1, Figure 5B) or neutral evolution (RAND sequences, β = 0, p-value = 1, Figure 5C) showed no relationship between motif integrity and language ability, highlighting HAQERs' unique selection for regulatory function during hominin evolution.

Analysis of specific transcription factor families revealed striking enrichment of Homeobox and Forkhead box transcription factors associated with both enhanced binding affinity in HAQERs and improved language performance (Figure 5D). The Homeobox family displayed the strongest enrichment among all transcription factor families (odds ratio = 11.58, p-value = 2.6×10^{-34}), followed by the Forkhead box family (which includes *FOXP2*, odds ratio = 7.28, p-value = 5.3×10^{-6}, Figure 5E, Supplementary Table 16). These results suggest that hominin-gained binding of Homeobox and Forkhead box families within HAQERs may have played a crucial role in the evolution of human language capability.

HAQERs regulate language-relevant brain cell types through human-specific chromatin accessibility

To determine which types of brain cells are regulated by HAQERs, we analyzed their overlap with candidate cis-regulatory elements (cCREs) identified by single-nucleus chromatin accessibility profiling (snATAC-seq) of human and mouse brain cell types (*39*). We tested whether HAQERs preferentially associate with human-specific versus evolutionarily conserved chromatin accessible regions, reasoning that HAQERs should show increased overlap with human-specific regulatory elements if they provide novel functions in the human lineage.

HAQERs demonstrated significant enrichment for human-specific cCREs across brain cell types (Figure S7A), with the strongest enrichment in medium spiny neurons (MSNs, p-value = 9.8×10^{-8}). MSNs comprise over 90% of neurons in the human striatum, a circuit that plays essential roles in vocal learning across species (*17, 40*) and was the only part of the brain robustly associated with developmental language disorders in a recent meta-analysis (*41*). HAQERs also showed enrichment around human-specific cCREs in *FOXP2*-expressing neurons (p-value = 5.3×10^{-4}), providing independent evidence linking HAQERs to *FOXP2* regulatory networks beyond our transcription factor binding findings (Figure 5E). In contrast, HARs showed minimal enrichment for human-specific chromatin accessibility but strong enrichment for evolutionarily conserved regions (strongest in VIP neurons, p-value = 2.7×10^{-10}, Figure S7B). These results are consistent with HAQERs providing novel regulatory functions specific to human brain development that may support language capabilities.

Selective pressures acting on language and general cognition

Having multiple lines of evidence associating HAQERs with human language evolution, we next examined how selective pressures may have influenced language-related genetic variation over the past 20,000 years of human history using the Allen Ancient DNA Resource (AADR) (*42*). The AADR is the largest genotyped collection of ancient humans, providing harmonized genotype and metadata for each sample (like radiocarbon dating based sample ages). We identified ancient west Eurasians, then correlated their HAQER CP-PGS and the background CP-PGS with sample age (N = 3,244 individuals with remains dated between 18,775 to 150 years ago passing quality

control). We see that the polygenic score for general cognition (background CP-PGS) has been subject to positive selection and has increased substantially over time (selection coefficient = 0.088, p-value = 2.1×10^{-12}), Figure 6A. Unexpectedly, we found that HAQER CP-PGS has been stable throughout human history, indicating that ancient and modern humans carry similar numbers of language-related alleles in HAQERs (selection coefficient = -0.004, p-value = 0.71, Figure 6A).

The presence of HAQERs in archaic humans provided a unique opportunity to describe genetically predicted cognitive traits across human species. To do this, we computed HAQER CP-PGS and background CP-PGS in archaic humans (N = 10) and compared them to ancient (N = 3,244) and modern humans (N = 503). Remarkably, the ten archaic human genomes (eight Neanderthals and two Denisovans) showed elevated HAQER CP-PGS (mean z-score = 0.91, median z-score = 1.23), while having reduced background CP-PGS (mean z-score = -3.02, median z-score = -3.01, Figure 6A). In contrast, a set of random matched control regions showed no differences in CP-PGS across groups (Supplementary Figure S5A-C, Supplementary Table 9). While this data should be interpreted with caution due to challenges applying polygenic scores across populations (*43*), the elevated HAQER CP-PGS we observed aligns with arguments that archaic humans were capable of complex language (*44–46*).

Evidence of balancing selection from modern genomes

The stability of HAQER CP-PGS throughout human evolutionary history led us to hypothesize that HAQERs have been maintained through balancing selection. Multiple population genetic analyses in the EpiSLI sample provided support for this hypothesis. HAQERs exhibited significantly more heterozygosity compared to both HARs (t-statistic = 110, p-value = 5.3×10^{-273}) and matched control sequences (t-statistic = 147, p-value = 1.6×10^{-315}), suggesting that heterozygosity at HAQER loci provided a selective advantage (Figure 6C). Additionally, we observed an enrichment of intermediate frequency variants (MAF 30-50%) in HAQERs (Figure 6B). These patterns could arise from several mechanisms, including elevated mutation rates or balancing selection (*47, 48*). However, together with the stability of HAQER polygenic scores over 20,000 years and the obstetric trade-offs (described below), our findings are most consistent with balancing selection maintaining genetic variation in these regions at intermediate frequencies throughout human history.

HAQERs influence prenatal brain development

The evolutionary analysis revealed a puzzling pattern: while other cognitive variants show recent positive selection, the HAQER CP-PGS has remained stable for at least the past 20,000 years of human history despite the cognitive benefits associated with these variants. This stability suggests ongoing fitness costs that counterbalance the advantages of increased language ability. Given HAQERs' established role in neurodevelopment (*34*), we investigated whether these variants create pleiotropic effects on prenatal brain development that could explain their evolutionary constraints. First, we investigated temporal and cell-type enrichment for these regions to identify if they were likely to influence birth related neurodevelopmental traits. HAQERs showed broad enrichment for variants affecting prenatal brain gene expression when intersected with single-cell quantitative trait loci (scQTLs) from developing midbrain neurons (*49*). The strongest enrichment observed was at the late prenatal time point, which corresponds to when the human brain most rapidly expands (*50*) (Figure S7C). Critically, HAQERs showed no enrichment when we examined adult brain regulatory elements (*51*), confirming prenatal neurodevelopmental effects (Figure S7D). HAQERs were also significantly enriched around genomic loci associated with head circumference at birth, a proxy for brain size (*52*) (p-value = 4.4×10^{-4}, Figure S7E).

HAQERs link language evolution to the obstetric dilemma

The evidence for HAQER effects on prenatal brain development and head circumference at birth suggests a potential mechanism for their evolutionary stability: the obstetric dilemma. Enhanced fetal brain development may create reproductive costs through the obstetric dilemma, where increased brain size complicates birth in bipedal humans with narrow pelvises (*53,54*). To test whether HAQERs contribute to this trade-off, we analyzed brain imaging, cognitive, and birth outcome data in the ABCD cohort (*37*).

A canonical correlation analysis revealed two distinct composite phenotypic axes associated with HAQER CP-PGS, providing evidence for the obstetric dilemma that could plausibly drive the observed balancing selection (Figure 6D). The first canonical component captured variance primarily from birth complication-related variables, while the second component loaded predominantly on cognitive performance measures (including a measure of verbal language learning; Figure S6).

Critically, both composite scores showed positive correlations with HAQER CP-PGS (r = 0.69 and r = 0.66 respectively; p-value $< 2.2 \times 10^{-16}$ for both), and the two composite scores were themselves positively correlated (r=0.32, p-value $< 2.2 \times 10^{-16}$). This pattern indicates that genetic variants contributing to higher HAQER CP-PGS simultaneously increase both cognition (a trait under positive selection) and birth complication risk (a trait under negative selection). The consistent positive relationship between HAQER CP-PGS and both phenotypic domains provides a plausible mechanistic explanation for the balancing selection signatures specific to HAQERs. While the correlation magnitudes should be interpreted cautiously given our optimization procedure (see Materials and Methods), the qualitative finding of antagonistic pleiotropy is robust and aligns with the evolutionary hypothesis that the cognitive benefits conferred by HAQER variants are counterbalanced by obstetric costs, maintaining genetic diversity at these loci through balancing selection. This trade-off between cognitive ability and increased birth complications is a fundamental constraint that may have shaped the evolution of human language.

HAQER-like sequences show convergent evolution in vocal learning mammals

To test whether HAQER functions extend beyond humans, we analyzed homologous sequences across 170 non-primate mammalian species, including 49 vocal learners and 121 non-vocal learner species (*15*). Vocal learner species can acquire and modify vocalizations through experience, contrasting with species restricted to innate vocalizations. We computed genome-wide "HAQER-like" and "HAR-like" sequence similarity scores for each species by comparing homologous genomic regions to the human reference genome using whole genome alignment data (*55*), with these scores measuring how similar each species' sequences are to human HAQERs and HARs. We tested for associations between these similarity scores and vocal learning ability while controlling for phylogenetic relatedness across species (*56, 57*).

Vocal learner species showed significantly higher HAQER-like sequence similarity compared to non-vocal learners (phylogenetic regression β = 1.41, p-value = 1×10^{-4}, Figure 6E). HAQERs were also enriched around previously identified mammalian vocal learner enhancer regions (p-value = 0.028, Figure S7F) (*15*). While HARs showed a similar enrichment around vocal learner enhancer regions (p-value = 0.04), HAR sequence similarity was not associated with vocal learner

classification (phylogenetic regression β = -0.15, p-value = 0.67, Figure 6E). Given the independent evolution of vocal learning across mammalian lineages, these results suggest HAQER-like sequences may be a fundamental genetic mechanism for complex vocal communication that has been repeatedly utilized across evolutionary history.

Remarkably, HAQER-like sequences also associated with brain size across species (phylogenetic regression $\beta = 0.42$, p-value = 0.006, N = 116 species) and larger relative birth weights (phylogenetic regression β = 0.44, p-value = 0.001, N = 115 species), mirroring the human obstetric dilemma pattern (Figure 6F-G). This convergent evidence across independent evolutionary lineages supports a link between the genetic architecture of vocal learning, brain development, and reproductive constraints. Together, these results suggests the trade-offs we observe in human language evolution may represent a broader biological phenomenon that support complex vocal communication.

Discussion

Our evolutionary stratified polygenic score analysis identifies genomic regions that disproportionately contributed to human language evolution and continue to influence individual differences in language abilities observed in present-day humans. While previous research demonstrated that rare mutations in *FOXP2* can cause language disorders (*5, 58*), common variants in *FOXP2* show minimal association with typical language variation (*7, 8*), prompting GWAS studies and polygenic models of language-related traits (*9, 11–14, 59*). However, these models left critical questions unanswered: when did language-associated variation evolve, and how do these molecular changes influence language development? Our analysis reveals that HAQERs (*31, 34*), mostly non-coding regions that rapidly evolved before the human-Neanderthal split and represent less than 0.1% of the human genome, harbor variants with disproportionate effects on language. Individual SNPs in HAQERs carry 188 times more impact on language than variants elsewhere in the genome. While we observed consistent associations between HAQERs and language-related measures across multiple independent cohorts, the relatively modest replication effect sizes (UK Biobank p-value = 0.002, minimum SPARK p-value = 7.9×10^{-5}, and ABCD p-value = 0.048) likely reflect phenotypic heterogeneity, with the EpiSLI discovery cohort undergoing extensive, standardized language assessments compared to briefer measures in replication samples. In contrast, we observed no as-

sociation between HAQERs and nonverbal cognition across multiple independent cohorts, though continued investigation of HAQER effects across diverse cognitive domains will be valuable for understanding the full extent of their phenotypic specificity. These potent regulatory effects distributed across thousands of small genomic elements support the polygenic architecture of human language while pointing to its specific evolutionary origins.

We find that HAQERs evolved across hominins to increase binding for Forkhead (including *FOXP2*) and Homeobox transcription factors (TFs), with motif integrity correlating with individual language scores. Supporting this regulatory mechanism, HAQERs show significant enrichment for human-specific chromatin accessible regions across brain cell types, with strong enrichment in medium spiny neurons and *FOXP2*-expressing neurons, while HARs primarily overlap with evolutionarily conserved regulatory elements. This suggests *FOXP2* influences language primarily through its regulatory networks (*60*) rather than protein-coding changes, explaining how rare mutations in a single TF can produce profound effects on language development while common variants in the same TF shows minimal associations (*7, 8*). The observed 11.58-fold and 7.28-fold enrichment of hominin-gained Homeobox and Forkhead binding sites that positively correlate with language scores in HAQERs suggests that these developmentally essential TF families may have played a central role in human language evolution (*61–63*).

Cross-species triangulation provides independent support that HAQERs are functional regulatory elements for language-related traits. Consistent with our human evidence, vocal learner mammals show significantly higher HAQER-like sequence similarity than non-vocal learners after controlling for phylogenetic relationships, with parallel associations for brain size and birth weight. Importantly, HAQERs show enrichment around established mammalian vocal learning enhancer regions (*15*), consistent with previous reports of convergent evolution for complex vocal communication (*15–20*). This convergent evolution of HAQER-like sequences across vocal learning lineages provides strong independent support that these regulatory elements are fundamental genetic mechanisms for complex vocal communication.

Ancient DNA analysis reveals that while general cognitive variants show positive selection over 20,000 years, the language-related HAQER polygenic score has remained stable, suggesting balancing selection maintains genetic variation in these language-relevant regions. Analysis of birth outcomes, brain imaging, and cognitive data provided a potential mechanism for this

unexpected evolutionary constraint: individuals with higher HAQER CP-PGS were more likely to have larger heads and birth complications, indicating trade-offs between language capability and reproductive costs. This pattern connects HAQERs to the obstetric dilemma, the evolutionary trade-off between narrower pelvises supporting upright walking and larger fetal brains enabling complex cognition (*53, 54*), potentially explaining why variants in HAQERs persist at intermediate frequencies rather than reaching fixation. Further supporting this neurodevelopmental mechanism, we find that HAQERs are enriched for genetic variants associated with prenatal gene expression and head circumference at birth (*49, 52*). The prenatal brain regulatory activity of HAQERs aligns with established evidence that early developmental processes critically influence later language outcomes (*64, 65*). Intriguingly, despite substantial methodological limitations in cross-population polygenic score applications (*43*), the available Neanderthal and Denisovan genomes show higher HAQER CP-PGS than modern humans, although this requires careful interpretation. This finding needs further validation, and future research should also investigate morphological and obstetric differences between archaic and modern humans that may have enabled archaic populations to maintain higher HAQER polygenic scores. These results point to the obstetric dilemma as an ongoing evolutionary constraint that specifically limits language-related genetic variation from reaching fixation, creating a fundamentally different selection landscape for vocal communication compared to general cognitive abilities.

These findings demonstrate how ancient regulatory innovations continue to shape human language variation through evolutionary constraints that balance cognitive benefits against reproductive costs. The independent evolution of HAQER-like sequences in vocal learning mammals reveals fundamental genetic mechanisms for complex communication that have emerged repeatedly across lineages. While general cognitive variants show positive selection over time, language variants remain stable at intermediate frequencies, suggesting that individual differences in spoken language abilities reflect ongoing evolutionary trade-offs. These evolutionary constraints reveal why individual differences in language abilities persist despite the importance of communication skills. Further investigation of HAQER regulatory networks will be essential to translate these evolutionary insights into approaches for supporting those with language disorders.

Materials and methods summary

We developed an evolutionary stratified polygenic score (ES-PGS) approach to trace the origins of language-relevant genetic variation across 65 million years of primate evolutionary events. Language abilities were assessed through factor analysis of 17 longitudinal cognitive and language assessments from kindergarten through 4th grade in 350 children from a community-based cohort (EpiSLI) (*21*). We applied ES-PGS using cognitive performance (*28*) polygenic scores partitioned across five evolutionary annotations (*29–33*), testing whether genomic regions from specific evolutionary periods contribute disproportionately to language versus general cognitive functions.

Validation was performed across multiple independent cohorts using both common and rare genetic variants. In the SPARK autism dataset (N > 30,000) (*36*), we tested HAQER effects on verbal language ability and language disorder diagnoses. We analyzed rare "reversions" in SPARK whole genome sequencing data (N > 2,000 individuals), variants reverting from human-specific to human-chimpanzee ancestral states (*34*), reasoning that if HAQERs evolved to enhance language, reversions should impair language function. Additional validation used unrelated individuals from the ABCD cohort (N = 5,625) for spoken word recall performance, cognitive test scores, brain size, and birth traits (*37*). Finally, we validated the effects of HAQERs on language-related traits in the UK Biobank (*66*). Given that cognitive trait GWASs use the UK Biobank, we opted to use the largest quantitative language-related GWAS available (for word reading) in our ES-PGS analysis (*9*). We focused our analysis on a verbal working memory trait, which was previously shown to be highly relevant to language, and educational attainment (*38*). For comparison, we also analyzed matrix reasoning in the UK Biobank.

To investigate molecular mechanisms supporting language evolution, we analyzed how rare genetic variants in evolutionarily significant regions affect transcription factor binding sites using position weight matrices for 633 human transcription factors from JASPAR2020 (*67*). We compared hominin-chimpanzee ancestral allele reversions versus other rare variants to detect systematic evolutionary changes in transcription factor binding associated with language phenotypes. Additionally, we looked for enrichment of evolutionarily significant regions across human-specific or conserved chromatin accessible regions (*39*), birth head circumference GWAS loci (*52*), neurodevelopmental related scQTLs datasets (*49, 51*), and mammalian vocal learning genomic regions (*15*).

Ancient DNA analysis examined selective pressures using the Allen Ancient DNA Resource (*42*), correlating HAQER polygenic scores with sample age in 3,244 ancient west Eurasians (18,775-150 years ago). We computed polygenic scores in archaic humans (8 Neanderthals, 2 Denisovans) and compared them to ancient and modern humans. Cross-species validation analyzed HAQER-like sequence similarity across 170 non-primate mammalian species (49 vocal learning, 121 non-vocal learning) (*15*) using phylogenetic regression to control for evolutionary relatedness (*56, 57*) in analyses of vocal learning, brain size, and a birth:adult weight ratio (used as a rough proxy for the obstetric dilemma) (*68*).

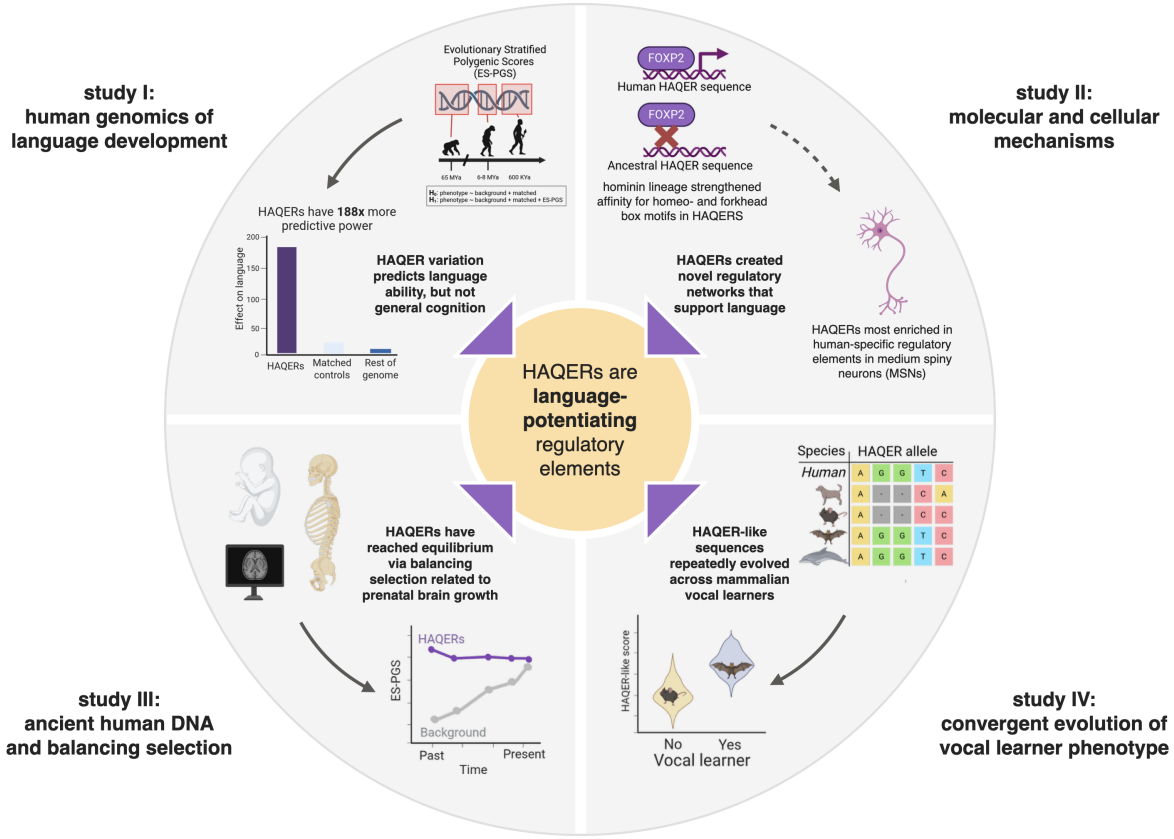
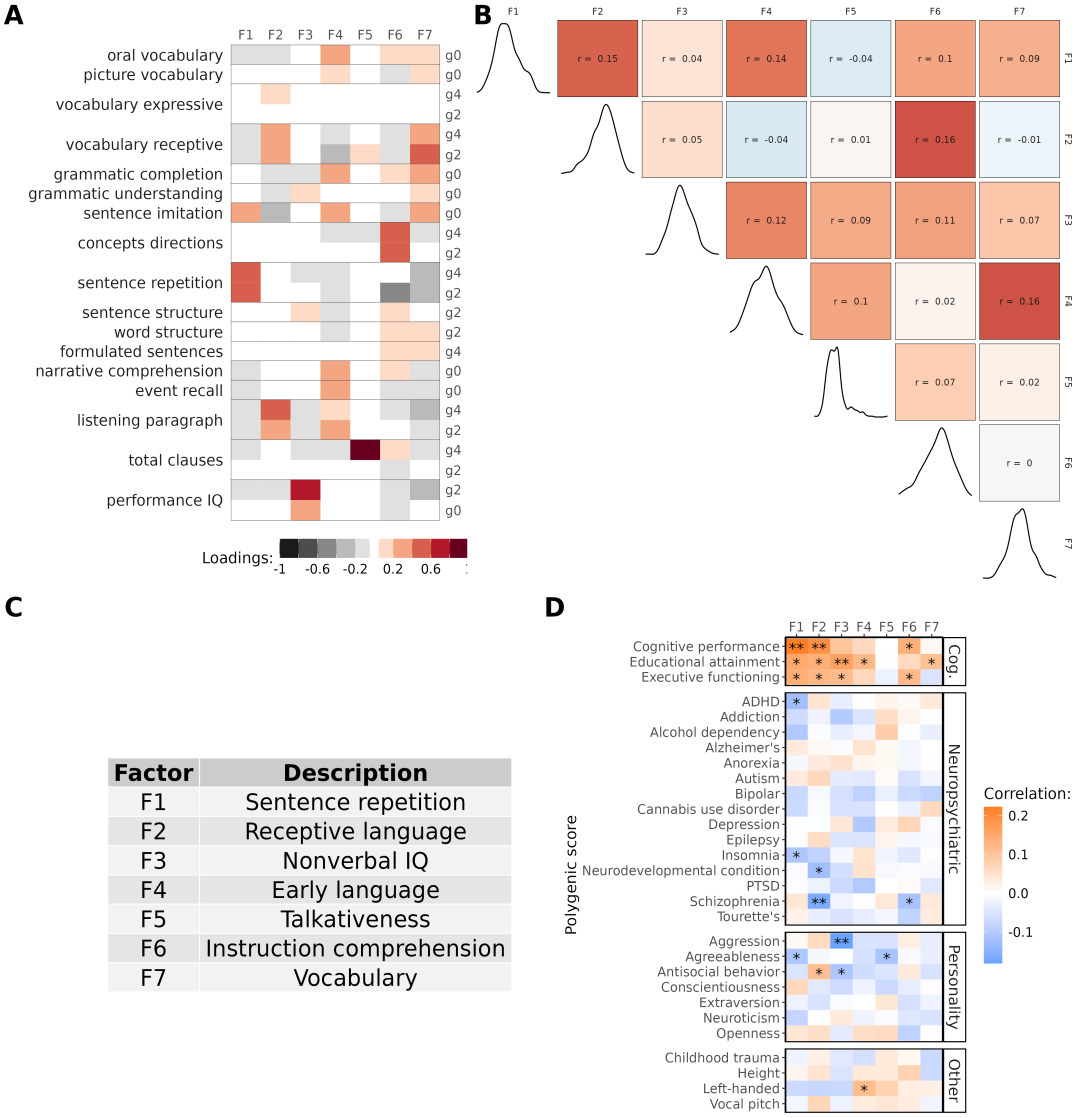


Figure 1: **Overview of this study and key findings**



Factor	Description
F1	Sentence repetition
F2	Receptive language
F3	Nonverbal IQ
F4	Early language
F5	Talkativeness
F6	Instruction comprehension
F7	Vocabulary

Figure 2: **Factor loadings and genetic associations**

A Loadings of cognitive and language assessments onto the seven language factors. g0 = Kindergarten (age 5-6), g2 = 2nd grade (age 7-8), g4 = 4th grade (age 9-10). **B** Pearson correlations for language factors (upper triangle) and distribution of each factor (diagonal). **C** Interpretations of the language factors based on their loadings. **D** Pearson correlations for each factor with genome-wide PGS. ** indicates FDR adjusted p-value < 0.05 and * indicates unadjusted p-value < 0.05.

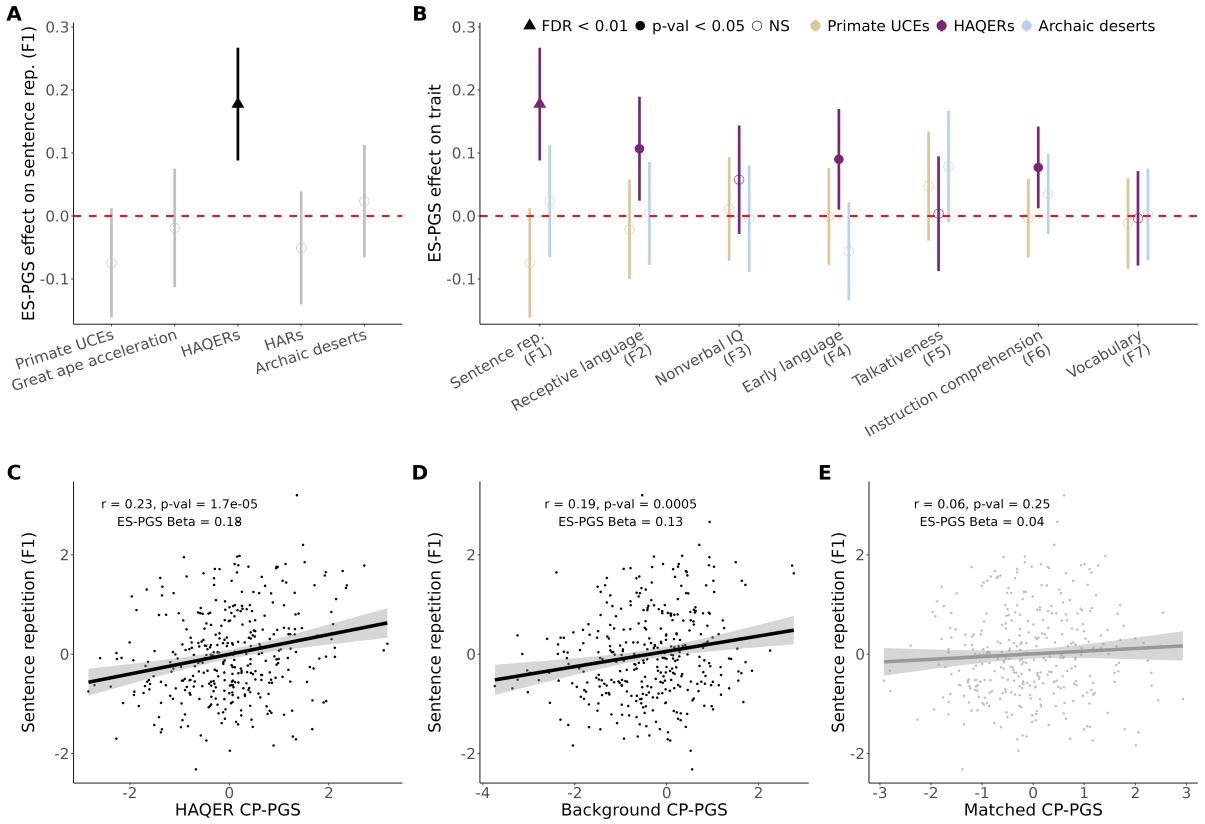


Figure 3: **HAQERs are associated with language ability**

A Comparison of evolutionary events effect on sentence repetition ability in EpiSLI (N = 350). Points represent the β provided from the ES-PGS models for each evolutionary annotation, while the ranges represent the 95% confidence interval. Solid points indicate p-value < 0.05. **B** Comparison of 3 evolutionary (oldest = Primate UCEs, middle = HAQERs, and youngest = archaic deserts) events on the 7 factor scores in EpiSLI (N = 350). Points represent the β provided from the ES-PGS models for each evolutionary annotation, while the ranges represent the 95% confidence interval. Solid points indicate p-value < 0.05. **C** Scatterplot of HAQER CP-PGS with sentence repetition scores (F1) in the EpiSLI sample. **D** Scatterplot of background CP-PGS with sentence repetition scores (F1) in the EpiSLI sample. **E** Scatterplot of biologically matched control regions CP-PGS with sentence repetition scores (F1) in the EpiSLI sample (matched to HAQERs).

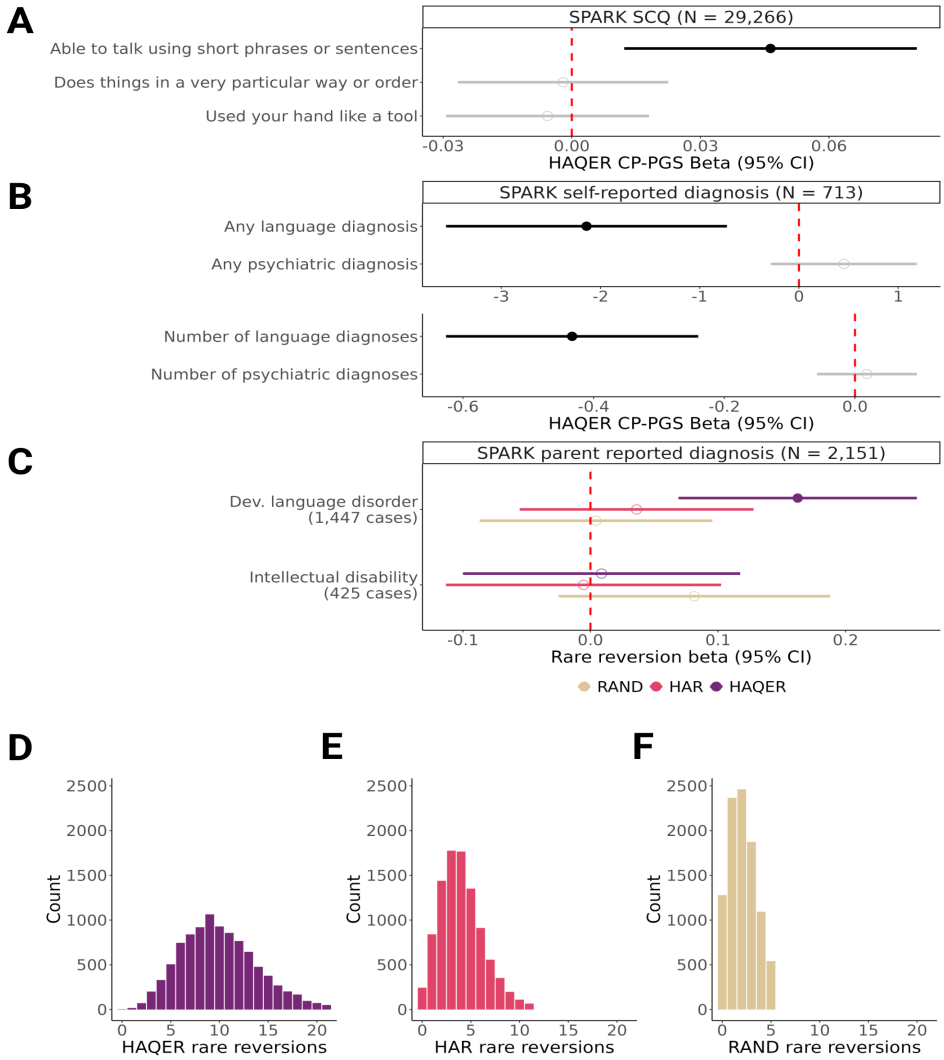


Figure 4: **Large-scale validation of HAQERs association with language ability**

A Points represent the β provided from the ES-PGS models for the HAQER CP-PGS on Social Communication Questionnaire (SCQ) items in SPARK, while the ranges represent the 95% confidence interval. Solid points indicate p-value < 0.05. **B** Points represent the β provided from the ES-PGS models for the HAQER CP-PGS on self-reported language and psychiatric diagnosis in SPARK, while the ranges represent the 95% confidence interval. Solid points indicate p-value < 0.05. **C** Points represent the β provided from the regression models for the rare reversions within 10Kb of HAQERs, HARs, or RAND (random matched) sequences, while the ranges represent the 95% confidence interval. Solid points indicate p-value < 0.05. **D-F** Distributions of rare reversions counts from the SPARK whole genome sequencing data within 10Kb of the following regions: HAQERs (D), HARs (E), and random sequence (RAND, F).

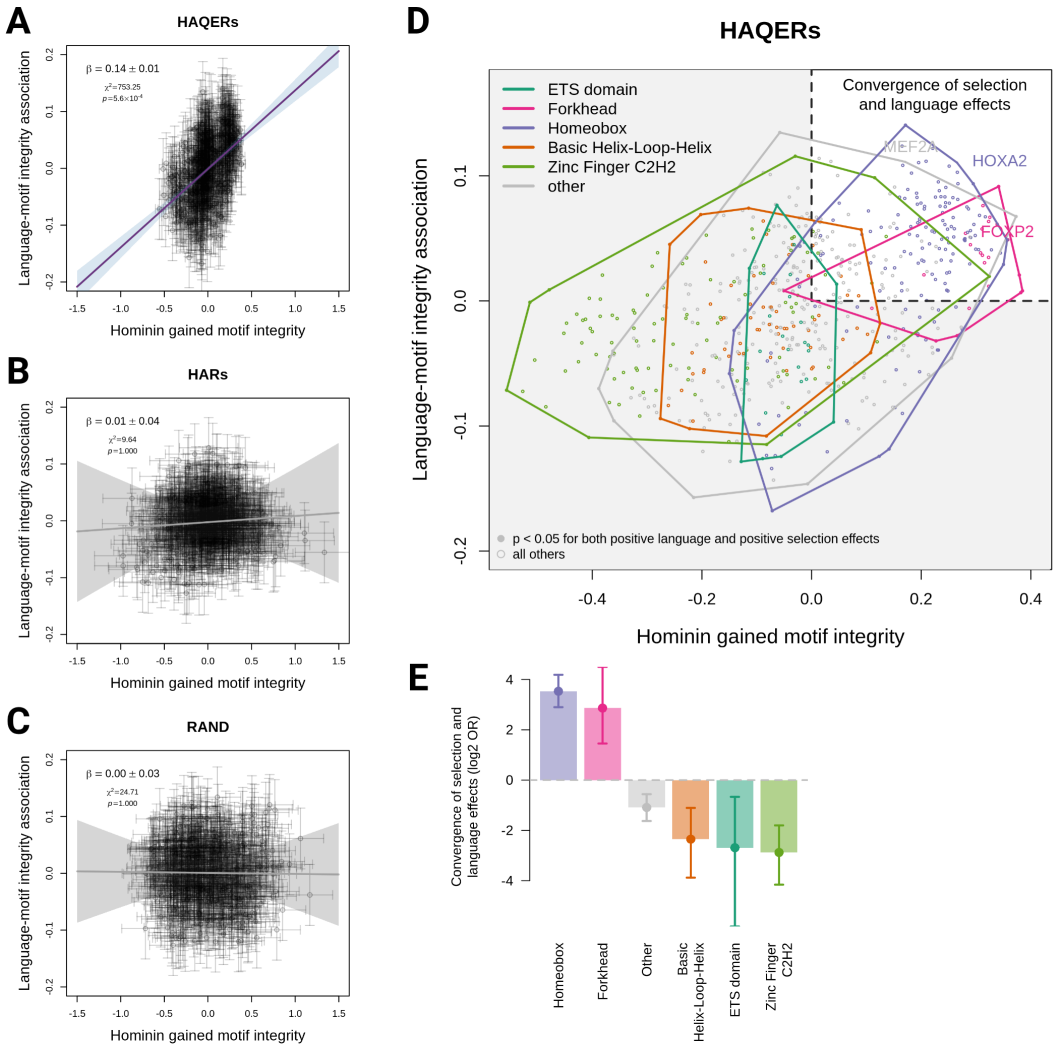


Figure 5: **Hominin gained transcription factor binding in HAQERs influences language**

A-C Relationship between selection for transcription factor motif integrity (x-axis) and motif association with language ability (y-axis) in (A) HAQERs, (B) HARs, and (C) random genomic regions. Each point represents one transcription factor motif. Error bars indicate ±1 standard error. Purple line (or gray for non-significant fits) shows York regression fit with 95% confidence interval (shaded); regression coefficient (β), chi-squared statistic (χ^2), and p-values are shown. **D** Detailed view of motif effects in HAQERs colored by transcription factor family. Solid points indicate motifs with $p < 0.05$ for both positive selection and positive language association. Colored polygons show convex hulls for each transcription factor family. **E** Enrichment analysis of transcription factor families for concordant positive selection and language effects, shown as log2 odds ratios. Error bars indicate 95% confidence intervals. Solid points indicate $p < 0.05$.

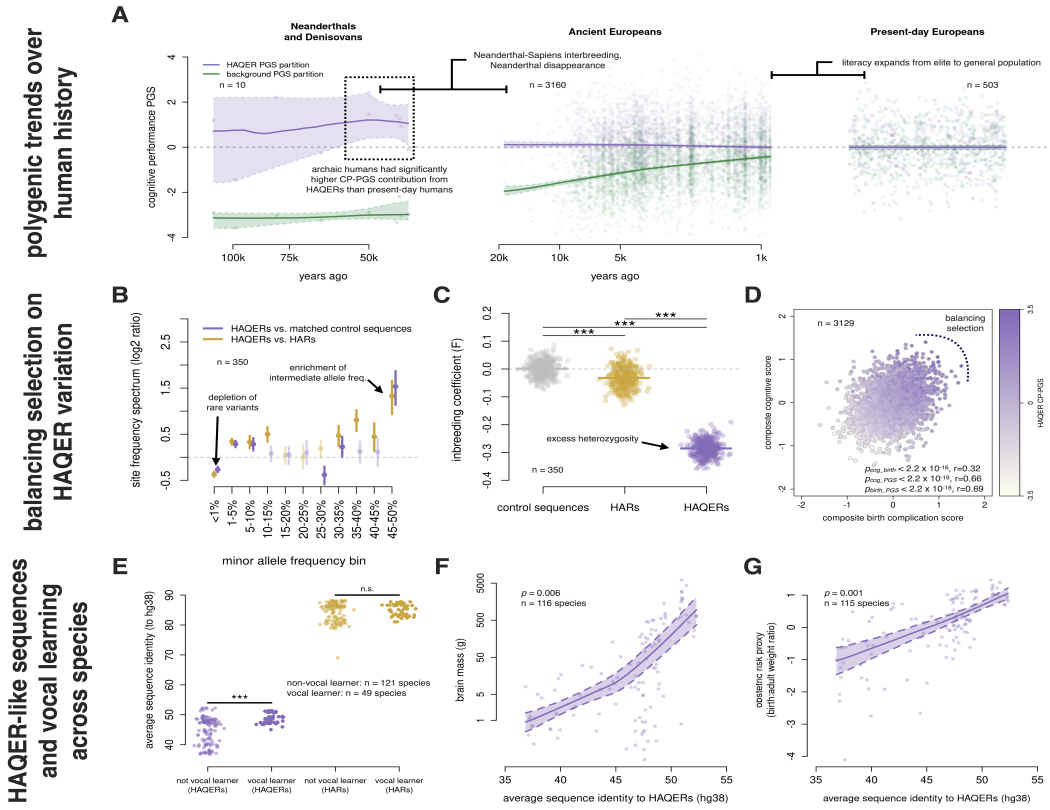


Figure 6: **Selective pressures on human cognition and convergent evolution of vocal learning** **A** CP ES-PGS across hominin evolution. HAQER PGS (purple) and background PGS (green) plotted against sample age for Neanderthals/Denisovans, Ancient Europeans, and Present-day Europeans. LOESS fits with 95% CIs shown. **B** Site frequency spectrum comparing HAQERs to matched controls and HARs. Log_2 ratio of proportion of variants across allele frequency bins; positive values indicate HAQER enrichment. Error bars: 95% bootstrap CIs. **C** Inbreeding coefficient (F-statistic) across sequence types. Lower F-statistics indicate excess heterozygosity. HAQERs show significantly lower F-statistics than HARs and control regions (*** indicates p-value < 0.001). **D** Canonical correlation analysis linking HAQER CP-PGS to cognitive and birth complication traits in the ABCD cohort. Each point is an individual, colored by their HAQER CP-PGS. X-axis represents a composite birth complication score; Y-axis captures cognitive performance measures (see Figure S6 for variable loadings). **E** HAQER-like and HAR-like sequence similarity scores in non-vocal learning (N = 121 species) and vocal learning (N = 49 species) mammals. Phylogenetic logistic regression statistics are shown. **F-G** HAQER-like sequence similarity correlates with brain mass (N = 116 species, **F**) and birth:adult weight ratio (N = 115 species, **G**) across mammals. LOESS fits with 95% CIs.

References and Notes

1. M. V. Suntsova, A. A. Buzdin, Differences between human and chimpanzee genomes and their implications in gene expression, protein functions and biochemical properties of the two species. *BMC genomics* **21**, 535 (2020), doi:10.1186/s12864-020-06962-8.

2. R. J. Britten, Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 13633–5 (2002), doi:10.1073/pnas.172510699.

3. A. Varki, T. K. Altheide, Comparing the human and chimpanzee genomes: searching for needles in a haystack. *Genome research* **15**, 1746–58 (2005), doi:10.1101/gr.3737405.

4. S. E. Fisher, Evolution of language: Lessons from the genome. *Psychon. Bull. Rev.* **24** (1), 34–40 (2017).

5. C. S. L. Lai, S. E. Fisher, J. A. Hurst, F. Vargha-Khadem, A. P. Monaco, A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature* **413**, 519–523 (2001), doi:10.1038/35097076.

6. S. E. Fisher, C. Scharff, FOXP2 as a molecular window into speech and language. *Trends in Genetics* **25**, 166–177 (2009), doi:10.1016/j.tig.2009.03.002.

7. E. G. Atkinson, *et al.*, No evidence for recent selection at FOXP2 among diverse human populations. *Cell* **174** (6), 1424–1435.e15 (2018).

8. K. L. Mueller, *et al.*, Common Genetic Variants in FOXP2 Are Not Associated with Individual Differences in Language Development. *PLOS ONE* **11**, e0152576 (2016), doi: 10.1371/journal.pone.0152576.

9. E. Eising, *et al.*, Genome-wide analyses of individual differences in quantitatively assessed reading- and language-related skills in up to 34,000 people. *Proceedings of the National Academy of Sciences* **119** (2022), doi:10.1073/pnas.2202764119.

10. H. G. Polikowsky, *et al.*, Large-scale genome-wide analyses of stuttering. *Nat. Genet.* **57** (8), 1835–1847 (2025).

11. E. Verhoef, *et al.*, Genome-Wide Analyses of Vocabulary Size in Infancy and Toddlerhood: Associations With Attention-Deficit/Hyperactivity Disorder, Literacy, and Cognition-Related Traits. *Biological Psychiatry* **95**, 859–869 (2024), doi:10.1016/j.biopsych.2023.11.025.

12. G. Alagöz, *et al.*, The shared genetic architecture and evolution of human language and musical rhythm. *Nat. Hum. Behav.* (2024).

13. C. Doust, *et al.*, Discovery of 42 genome-wide significant loci associated with dyslexia. *Nature genetics* **54**, 1621–1629 (2022), doi:10.1038/s41588-022-01192-y.

14. M. Niarchou, *et al.*, Genome-wide association study of musical beat synchronization demonstrates high polygenicity. *Nature Human Behaviour* **6**, 1292–1309 (2022), doi: 10.1038/s41562-022-01359-x.

15. M. E. Wirthlin, *et al.*, Vocal learning-associated convergent evolution in mammalian proteins and regulatory elements. *Science* **383** (6690), eabn3263 (2024).

16. I. M. Kaplow, *et al.*, Relating enhancer genetic variation across mammals to complex phenotypes using machine learning. *Science* **380** (6643), eabm7993 (2023).

17. E. D. Jarvis, Evolution of vocal learning and spoken language. *Science* **366** (6461), 50–54 (2019).

18. R. L. Gordon, *et al.*, Linking the genomic signatures of human beat synchronization and learned song in birds. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **376** (1835), 20200329 (2021).

19. J. A. Cahill, *et al.*, Positive selection in noncoding genomic regions of vocal learning birds is associated with genes implicated in vocal learning and speech functions in humans. *Genome Res.* **31** (11), 2035–2049 (2021).

20. M. Sebastianelli, *et al.*, A genomic basis of vocal rhythm in birds. *Nat. Commun.* **15** (1), 3095 (2024).

21. J. B. Tomblin, The EpiSLI database: a publicly available database on speech and language. *Lang. Speech Hear. Serv. Sch.* **41** (1), 108–117 (2010).

22. M. Klem, *et al.*, Sentence repetition is a measure of children's language skills rather than working memory limitations. *Dev Sci* **18** (1), 146–154 (2015).

23. I. Rujas, S. Mariscal, E. Murillo, M. Lázaro, Sentence repetition tasks to detect and prevent language difficulties: A scoping review. *Children (Basel)* **8** (7), 578 (2021).

24. S. H. Lee, *et al.*, Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics* **44**, 247–250 (2012), doi:10.1038/ng.1108.

25. H. K. Finucane, *et al.*, Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47**, 1228–1235 (2015), doi:10.1038/ng.3404.

26. X. Wei, *et al.*, The lingering effects of Neanderthal introgression on human complex traits. *Elife* **12** (2023).

27. S. W. Choi, *et al.*, PRSet: Pathway-based polygenic risk score analyses and software. *PLoS genetics* **19**, e1010624 (2023), doi:10.1371/journal.pgen.1010624.

28. J. J. Lee, *et al.*, Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics* **50**, 1112–1121 (2018).

29. L. F. K. Kuderna, *et al.*, Identification of constrained sequence elements across 239 primate genomes. *Nature* **625**, 735–742 (2024), doi:10.1038/s41586-023-06798-8.

30. X. Bi, *et al.*, Lineage-specific accelerated sequences underlying primate evolution. *Science advances* **9**, eadc9507 (2023), doi:10.1126/sciadv.adc9507.

31. D. Yoo, *et al.*, Complete sequencing of ape genomes. *Nature* **641** (8062), 401–418 (2025).

32. J. A. Capra, G. D. Erwin, G. McKinsey, J. L. R. Rubenstein, K. S. Pollard, Many human accelerated regions are developmental enhancers. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **368**, 20130025 (2013), doi:10.1098/rstb.2013.0025.

33. S. Peyrégne, M. J. Boyle, M. Dannemann, K. Prüfer, Detecting ancient positive selection in humans using extended lineage sorting. *Genome Res.* **27** (9), 1563–1572 (2017).

34. R. J. Mangan, *et al.*, Adaptive sequence divergence forged new neurodevelopmental enhancers in humans. *Cell* **185**, 4587–4603.e23 (2022), doi:10.1016/j.cell.2022.10.016.

35. K. S. Pollard, *et al.*, Forces shaping the fastest evolving regions in the human genome. *PLoS genetics* **2**, e168 (2006), doi:10.1371/journal.pgen.0020168.

36. P. Feliciano, *et al.*, SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. *Neuron* **97**, 488–493 (2018), doi:10.1016/j.neuron.2018.01.015.

37. K. M. Lisdahl, *et al.*, Adolescent brain cognitive development (ABCD) study: Overview of substance use assessment methods. *Developmental cognitive neuroscience* **32**, 80–96 (2018).

38. G. A. Jonsdottir, *et al.*, Genetic propensities for verbal and spatial ability have opposite effects on body mass index and risk of schizophrenia. *Intelligence* **88** (101565), 101565 (2021).

39. Y. E. Li, *et al.*, A comparative atlas of single-cell chromatin accessibility in the human brain. *Science* **382** (6667), eadf7044 (2023).

40. L. Xiao, *et al.*, Expression of FoxP2 in the basal ganglia regulates vocal motor sequences in the adult songbird. *Nat. Commun.* **12** (1), 2617 (2021).

41. M. T. Ullman, *et al.*, The neuroanatomy of developmental language disorder: a systematic review and meta-analysis. *Nature Human Behaviour* **8**, 962–975 (2024), doi: 10.1038/s41562-024-01843-6.

42. S. Mallick, *et al.*, The Allen Ancient DNA Resource (AADR) a curated compendium of ancient human genomes. *Sci. Data* **11** (1), 182 (2024).

43. Y. Ding, *et al.*, Polygenic scoring accuracy varies across the genetic ancestry continuum. *Nature* **618** (7966), 774–781 (2023).

44. D. Dediu, S. C. Levinson, Neanderthal language revisited: not only us. *Curr. Opin. Behav. Sci.* **21**, 49–55 (2018).

45. M. Conde-Valverde, *et al.*, Neanderthals and Homo sapiens had similar auditory and speech capacities. *Nature ecology evolution* **5**, 609–615 (2021), doi:10.1038/s41559-021-01391-6.

46. J. Krause, *et al.*, The derived FOXP2 variant of modern humans was shared with Neandertals. *Curr. Biol.* **17** (21), 1908–1912 (2007).

47. V. Llaurens, A. Whibley, M. Joron, Genetic architecture and balancing selection: the life and death of differentiated variants. *Mol. Ecol.* **26** (9), 2430–2448 (2017).

48. B. D. Bitarello, D. Y. C. Brandt, D. Meyer, A. M. Andrés, Inferring balancing selection from genome-scale data. *Genome Biol. Evol.* **15** (3) (2023).

49. J. Jerber, *et al.*, Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nat. Genet.* **53** (3), 304–312 (2021).

50. R. A. I. Bethlehem, *et al.*, Brain charts for the human lifespan. *Nature* **604** (7906), 525–533 (2022).

51. P. S. Emani, *et al.*, Single-cell genomics and regulatory networks for 388 human brains. *Science* **384** (6698), eadi5199 (2024).

52. S. Vogelezang, *et al.*, Genetics of early-life head circumference and genetic correlations with neurological, psychiatric and cognitive outcomes. *BMC Med. Genomics* **15** (1), 124 (2022).

53. N. D. S. Grunstra, *et al.*, There is an obstetrical dilemma: Misconceptions about the evolution of human childbirth and pelvic form. *Am. J. Biol. Anthropol.* **181** (4), 535–544 (2023).

54. L. Xu, *et al.*, The genetic architecture of and evolutionary constraints on the human pelvic form. *Science* **388** (6743), eadq1521 (2025).

55. L. F. K. Kuderna, *et al.*, A global catalog of whole-genome diversity from 233 primate species. *Science* **380** (6648), 906–913 (2023).

56. A. R. Ives, T. Garland, Jr, Phylogenetic logistic regression for binary dependent variables. *Syst. Biol.* **59** (1), 9–26 (2010).

57. L. s. T. Ho, C. Ané, A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst. Biol.* **63** (3), 397–408 (2014).

58. S. C. Vernes, *et al.*, A Functional Genetic Link between Distinct Developmental Language Disorders. *New England Journal of Medicine* **359**, 2337–2345 (2008), doi:10.1056/NEJMoa0802828.

59. V. M. Rajagopal, *et al.*, Genome-wide association study of school grades identifies genetic overlap between language ability, psychopathology and creativity. *Scientific Reports* **13**, 429 (2023), doi:10.1038/s41598-022-26845-0.

60. S. L. Hickey, S. Berto, G. Konopka, Chromatin decondensation by FOXP2 promotes human neuron maturation and expression of neurodevelopmental disease genes. *Cell Rep.* **27** (6), 1699–1711.e9 (2019).

61. S. A. Lambert, *et al.*, The human transcription factors. *Cell* **172** (4), 650–665 (2018).

62. O. Hobert, Homeobox genes and the specification of neuronal identity. *Nat. Rev. Neurosci.* **22** (10), 627–636 (2021).

63. M. Co, A. G. Anderson, G. Konopka, FOXP transcription factors in vertebrate brain development, function, and disorders. *Wiley Interdiscip. Rev. Dev. Biol.* **9** (5), e375 (2020).

64. B. Mariani, *et al.*, Prenatal experience with language shapes the brain. *Sci. Adv.* **9** (47), eadj3524 (2023).

65. S. Madigan, M. Wade, A. Plamondon, D. Browne, J. M. Jenkins, Birth weight variability and language development: Risk, resilience, and responsive parenting. *J. Pediatr. Psychol.* **40** (9), 869–877 (2015).

66. C. Sudlow, *et al.*, UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12** (3), e1001779 (2015).

67. O. Fornes, *et al.*, JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48** (D1), D87–D92 (2020).

68. K. E. Jones, *et al.*, PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90** (9), 2648–2648 (2009).

69. J. B. Tomblin, N. L. Records, X. Zhang, A System for the Diagnosis of Specific Language Impairment in Kindergarten Children. *Journal of Speech, Language, and Hearing Research* **39** (6), 1284–1294 (1996), doi:10.1044/jshr.3906.1284, https://doi.org/10.1044\%2Fjshr.3906.1284.

70. J. B. Tomblin, M. A. Nippold, eds., *Understanding Individual Differences in Language Development Across the School Years* (Psychology Press) (2014), doi:10.4324/9781315796987, https://doi.org/10.4324\%2F9781315796987.

71. P. L. Newcomer, D. D. Hammill, *Test of Language Development – 2 Primary* (Pro-Ed Austin, TX) (1988).

72. D. Wechsler, Wechsler preschool and primary scale of intelligence–revised (1989), title of the publication associated with this dataset: PsycTESTS Dataset.

73. B. Culatta, J. L. Page, J. Ellis, Story retelling as a communicative performance screening tool. *Lang. Speech Hear. Serv. Sch.* **14** (2), 66–74 (1983).

74. R. W. Woodcock, Woodcock reading mastery tests-revised (1987), title of the publication associated with this dataset: PsycTESTS Dataset.

75. H. W. Catts, The relationship between speech-language impairments and reading disabilities. *J. Speech Lang. Hear. Res.* **36** (5), 948–958 (1993).

76. T. M. Achenbach, *Child Behavior Checklist* (Springer New York), pp. 546–552 (2011), doi:10.1007/978-0-387-79948-3_1529.

77. R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2020), https://www.R-project.org/.

78. F. E. Harrell Jr, with contributions from Charles Dupont, many others., *Hmisc: Harrell Miscellaneous* (2020), https://CRAN.R-project.org/package=Hmisc, r package version 4.4-0.

79. M. Mayer, *missRanger: Fast Imputation of Missing Values* (2019), https://CRAN.R-project.org/package=missRanger, r package version 2.1.0.

80. S. Andrews, *et al.*, FastQC, Babraham Institute (2012).

81. B. Chapman, *et al.*, bcbio/bcbio-nextgen: v1.1.6 (2019), doi:10.5281/zenodo.3564939, https://doi.org/10.5281/zenodo.3564939.

82. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* (2013).

83. A. McKenna, *et al.*, The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297–1303 (2010), doi:10.1101/gr.107524.110.

84. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing. *arXiv* (2012).

85. A. Rimmer, *et al.*, Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics* **46**, 912–918 (2014), doi:10.1038/ng.3036.

86. A global reference for human genetic variation. *Nature* **526** (7571), 68–74 (2015), doi:10.1038/nature15393, https://doi.org/10.1038\%2Fnature15393.

87. W. McLaren, *et al.*, The Ensembl Variant Effect Predictor. *Genome Biology* **17** (1) (2016), doi:10.1186/s13059-016-0974-4, https://doi.org/10.1186\%2Fs13059-016-0974-4.

88. K. J. Karczewski, *et al.*, The mutational constraint spectrum quantified from variation in 141,456 humans (2019), doi:10.1101/531210, https://doi.org/10.1101\%2F531210.

89. E. M. Smigielski, dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Research* **28** (1), 352–355 (2000), doi:10.1093/nar/28.1.352, https://doi.org/10.1093\%2Fnar\%2F28.1.352.

90. B. S. Pedersen, R. M. Layer, A. R. Quinlan, Vcfanno: fast, flexible annotation of genetic variants. *Genome Biology* **17** (1) (2016), doi:10.1186/s13059-016-0973-5, https://doi.org/10.1186\%2Fs13059-016-0973-5.

91. S. W. Choi, T. S.-H. Mak, P. F. O'Reilly, Tutorial: a guide to performing polygenic risk score analyses. *Nature protocols* **15**, 2759–2772 (2020), doi:10.1038/s41596-020-0353-1.

92. F. Privé, C. Albiñana, J. Arbel, B. Pasaniuc, B. J. Vilhjálmsson, Inferring disease architecture and predictive ability with LDpred2-auto. *The American Journal of Human Genetics* **110**, 2042–2055 (2023), doi:10.1016/j.ajhg.2023.10.010.

93. D. Demontis, *et al.*, Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature genetics* **51** (1), 63–75 (2019).

94. A. S. Hatoum, *et al.*, Multivariate genome-wide association meta-analysis of over 1 million subjects identifies loci underlying multiple substance use disorders. *Nat. Ment. Health* **1** (3), 210–223 (2023).

95. R. K. Walters, *et al.*, Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nature Neuroscience* **21**, 1656–1669 (2018), doi: 10.1038/s41593-018-0275-1.

96. D. P. Wightman, *et al.*, A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nature Genetics* **53**, 1276–1282 (2021), doi:10.1038/s41588-021-00921-z.

97. J. Grove, *et al.*, Identification of common genetic risk variants for autism spectrum disorder. *Nature genetics* **51** (3), 431–444 (2019).

98. H. J. Watson, *et al.*, Genome-wide association study identifies eight risk loci and implicates metabo-psychiatric origins for anorexia nervosa. *Nature genetics* **51**, 1207–1214 (2019).

99. N. Mullins, *et al.*, Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nature Genetics* **53**, 817–829 (2021), doi:10.1038/s41588-021-00857-4.

100. E. C. Johnson, *et al.*, A large-scale genome-wide association study meta-analysis of cannabis use disorder. *Lancet Psychiatry* **7** (12), 1032–1045 (2020).

101. D. M. Howard, *et al.*, Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nature neuroscience* **22**, 343–352 (2019).

102. International League Against Epilepsy Consortium on Complex Epilepsies, GWAS meta-analysis of over 29,000 people with epilepsy identifies 26 risk loci and subtype-specific genetic architecture. *Nat. Genet.* **55** (9), 1471–1482 (2023).

103. K. Watanabe, *et al.*, Genome-wide meta-analysis of insomnia prioritizes genes associated with metabolic and psychiatric pathways. *Nature Genetics* **54**, 1125–1132 (2022), doi:10.1038/s41588-022-01124-w.

104. Q. Q. Huang, *et al.*, Examining the role of common variants in rare neurodevelopmental conditions. *Nature* **636** (8042), 404–411 (2024).

105. C. M. Nievergelt, *et al.*, International meta-analysis of PTSD genome-wide association studies identifies sex- and ancestry-specific genetic risk loci. *Nature Communications* **10**, 4558 (2019), doi:10.1038/s41467-019-12576-w.

106. V. Trubetskoy, *et al.*, Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022), doi:10.1038/s41586-022-04434-5.

107. D. Yu, *et al.*, Interrogating the genetic determinants of Tourette's syndrome and other tic disorders through genome-wide association studies. *Am. J. Psychiatry* **176** (3), 217–227 (2019).

108. C. A. Rietveld, *et al.*, Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proceedings of the National Academy of Sciences* **111** (38), 13790–13794 (2014), doi:10.1073/pnas.1404623111, `https://doi.org/10.1073\%2Fpnas.1404623111`.

109. A. Okbay, *et al.*, Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics* **54**, 437–449 (2022), doi:10.1038/s41588-022-01016-z.

110. A. S. Hatoum, *et al.*, Genome-wide Association Study Shows That Executive Functioning Is Influenced by GABAergic Processes and Is a Neurocognitive Genetic Correlate of Psychiatric Disorders. *Biological psychiatry* **93**, 59–70 (2023), doi:10.1016/j.biopsych.2022.06.034.

111. J. de la Fuente, G. Davies, A. D. Grotzinger, E. M. Tucker-Drob, I. J. Deary, A general dimension of genetic sharing across diverse cognitive traits inferred from molecular data. *Nature human behaviour* **5**, 49–58 (2021), doi:10.1038/s41562-020-00936-2.

112. H. F. Ip, *et al.*, Genetic association study of childhood aggression across raters, instruments, and age. *Transl. Psychiatry* **11** (1), 413 (2021).

113. J. J. Tielbeek, *et al.*, Uncovering the genetic architecture of broad antisocial behavior through a genome-wide association study meta-analysis. *Molecular psychiatry* **27**, 4453–4463 (2022), doi:10.1038/s41380-022-01793-3.

114. V. Warrier, *et al.*, Genome-wide analyses of self-reported empathy: correlations with autism, schizophrenia, and anorexia nervosa. *Translational psychiatry* **8**, 35 (2018), doi:10.1038/s41398-017-0082-6.

115. P. Gupta, *et al.*, A genome-wide investigation into the underlying genetic architecture of personality traits and overlap with psychopathology. *Nature Human Behaviour* (2024), doi:10.1038/s41562-024-01951-3.

116. W. D. Hill, *et al.*, Genome-wide analysis identifies molecular systems and 149 genetic loci associated with income. *Nature communications* **10**, 5741 (2019), doi:10.1038/s41467-019-13585-5.

117. B. M. Neale, UK Biobank GWAS Round 2 results (2018), `http://www.nealelab.is/uk-biobank/`.

118. K. L. Grasby, *et al.*, The genetic architecture of the human cerebral cortex. *Science (New York, N.Y.)* **367** (2020), doi:10.1126/science.aay6690.

119. E. Tissink, *et al.*, Genome-wide association study of cerebellar volume provides insights into heritable mechanisms underlying brain development and mental health. *Communications Biology* **5**, 710 (2022), doi:10.1038/s42003-022-03672-7.

120. E. Tissink, *et al.*, The Genetic Architectures of Functional and Structural Connectivity Properties within Cerebral Resting-State Networks. *eNeuro* **10** (2023), doi:10.1523/ENEURO.0242-22.2023.

121. L. Yengo, *et al.*, A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022), doi:10.1038/s41586-022-05275-y.

122. S. Dalvie, *et al.*, Genomic influences on self-reported childhood maltreatment. *Transl. Psychiatry* **10** (1), 38 (2020).

123. R. S. Gisladottir, *et al.*, Sequence variants affecting voice pitch in humans. *Science advances* **9**, eabq2969 (2023), doi:10.1126/sciadv.abq2969.

124. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841–2 (2010), doi:10.1093/bioinformatics/btq033.

125. E. S. Davis, *et al.*, matchRanges: generating null hypothesis genomic ranges via covariate-matched sampling. *Bioinformatics* **39** (5) (2023).

126. R. M. Kuhn, D. Haussler, W. J. Kent, The UCSC genome browser and associated tools. *Briefings in bioinformatics* **14**, 144–61 (2013), doi:10.1093/bib/bbs038.

127. T. R. Thomas, *et al.*, Polygenic Scores Clarify the Relationship Between Mental Health and Gender Diversity. *Biological Psychiatry Global Open Science* **4**, 100291 (2024), doi:10.1016/j.bpsgos.2024.100291.

128. L. G. Casten, *et al.*, Lingo: an automated, web-based deep phenotyping platform for language ability. *medRxiv : the preprint server for health sciences* (2024), doi:10.1101/2024.03.29.24305034.

129. K. Prüfer, *et al.*, The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505** (7481), 43–49 (2014).

130. F. Mafessoni, *et al.*, A high-coverage Neandertal genome from Chagyrskaya Cave. *Proc. Natl. Acad. Sci. U. S. A.* **117** (26), 15132–15136 (2020).

131. K. Prüfer, *et al.*, A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science* **358** (6363), 655–658 (2017).

132. M. Meyer, *et al.*, A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338** (6104), 222–226 (2012).

133. J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** (1), 1–22 (2010).

134. D. York, N. M. Evensen, M. L. Martınez, J. De Basabe Delgado, Unified equations for the slope, intercept, and standard errors of the best straight line. *Am. J. Phys.* **72** (3), 367–375 (2004).

135. T. Paysan-Lafosse, *et al.*, InterPro in 2022. *Nucleic Acids Res.* **51** (D1), D418–D427 (2023).

136. D. Szklarczyk, *et al.*, The STRING database in 2025: protein networks with directionality of regulation. *Nucleic Acids Res.* **53** (D1), D730–D737 (2025).

137. A. Akbari, *et al.*, Pervasive findings of directional selection realize the promise of ancient DNA to elucidate human adaptation. *bioRxiv* (2024).

138. S. Purcell, *et al.*, PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics* **81** (3), 559–575 (2007), doi: 10.1086/519795, `https://doi.org/10.1086\%2F519795`.

139. J. Yang, S. H. Lee, M. E. Goddard, P. M. Visscher, GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics* **88**, 76–82 (2011), doi:10.1016/j.ajhg.2010.11.011, `https://doi.org/10.1016/j.ajhg.2010.11.011`.

140. H. Perdry, C. Dandine-Roulland, *gaston: Genetic Data Handling (QC, GRM, LD, PCA) Linear Mixed Models* (2023), `https://CRAN.R-project.org/package=gaston`, r package version 1.6.

141. N. R. Karcher, D. M. Barch, The ABCD study: understanding the development of risk for mental and physical health outcomes. *Neuropsychopharmacology* **46** (1), 131–142 (2021).

142. D. Risso, J. Ngai, T. P. Speed, S. Dudoit, Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32** (9), 896–902 (2014).

143. A. Tenenhaus, *et al.*, Variable selection for generalized canonical correlation analysis. *Biostatistics* **15** (3), 569–583 (2014).

144. E. H. Au, *et al.*, Gonomics: uniting high performance and readability for genomics with Go. *Bioinformatics* **39** (8) (2023).

145. P. J. A. Cock, *et al.*, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25** (11), 1422–1423 (2009).

Acknowledgments

General: We are grateful for the contributions of the EpiSLI cohort, their families, and the research team. We also appreciate the effort of participants, families, and research team members in ABCD, SPARK, SPARK Research Match, Tempus, and the UK Biobank. We appreciate obtaining access to genetic and phenotypic data for SPARK data on SFARI Base. We are incredibly appreciative of the open access datasets used in this study, including the researchers who put together the AADR dataset, the Zoonomia consortium datasets, and PanTHERIA. We also greatly appreciate Dr. Kristi Hendrickson and Dr. Susan Shen for valuable and encouraging discussions about this project.

Funding: This work was funded by NIH grant DC014489 to J.J.M, some additional funding came from an NIH Predoctoral Training Grant to L.G.C.(T32GM008629). Additionally, J.J.M. is supported by the Roy J. Carver Charitable Trust.

Author contributions: Conceptualization: L.G.C, T.K., J.B.T, J.J.M.; Methodology: L.G.C, T.K., T.R.T, J.B.T., J.J.M.; Investigation: L.G.C, T.K., T.R.T, JY.K, A.M., J.B.T, J.J.M.; Visualization: L.G.C, T.K., J.J.M.; Funding acquisition: J.J.M.; Project administration: D.H., S.T., M.O'B., J.C.M, J.B.T, J.J.M.; Supervision: J.C.M, J.B.T, J.J.M.; Writing – original draft: L.G.C., T.K., J.J.M; Writing – review & editing: L.G.C., T.K., T.R.T, JY.K, S.T, A.M, J.C.M., J.B.T., J.J.M

Competing interests: There are no competing interests to declare.

Data and materials availability: Custom code for ES-PGS and all analyses in this paper is available here, including example code for ES-PGS that can be applied to other research problems:
https://github.com/lucasgcasten/language_evolution

The EpiSLI whole genome sequencing data described here is available to qualified researches via dbGaP (study accession = phs002255.v1.p1):
https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002255.v1.p1

SPARK genetic and phenotype data is available to qualified researchers at SFARI base:
https://base.sfari.org/

ABCD is available to qualified researchers at:
https://nda.nih.gov/abcd/request-access

1000 Genomes Phase 3 data is available at:
https://www.internationalgenome.org/data/

UK Biobank data are available to approved researchers at:
https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access

Allen Ancient DNA Resource:
https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-p

Neanderthal and Denisovan genomes:
https://www.eva.mpg.de/genetics/genome-projects/

Cross-species sequence alignment data:
https://hgdownload.soe.ucsc.edu/goldenPath/hg38/cactus447way/

PanTHERIA:
https://figshare.com/articles/dataset/Full_Archive/3531875

We used publicly available tools for data processing and analysis:
bcftools:
https://samtools.github.io/bcftools/bcftools.html

PLINK:
https://www.cog-genomics.org/plink

GCTA:

https://yanglab.westlake.edu.cn/software/gcta

LDpred2:

https://privefl.github.io/bigsnpr/articles/LDpred2.html

PRSet:

https://choishingwan.github.io/PRSice/quick_start_prset/

bedtools:

https://bedtools.readthedocs.io

Biopython:

https://biopython.org/

gonomics:

https://github.com/vertgenlab/gonomics

R:

https://www.r-project.org/

Supplementary materials

Materials and Methods

Figs. S1 to S5

References *(65-145)*

Supplementary Materials for

Ancient regulatory evolution shapes individual language abilities in present-day humans

Lucas G. Casten[†],Tanner Koomar[†],Taylor R. Thomas,Jin-Young Koh,Dabney Hofammann,Savantha Thenuwara,Allison Momany,Marlea O'Brien,Jeffrey C. Murray,J. Bruce Tomblin,Jacob J. Michaelson[*]

[*]Corresponding author. Email: jacob-michaelson@uiowa.edu

[†]These authors contributed equally to this work.

This PDF file includes:

Materials and Methods

Figures S1 to S7

Materials and Methods

Human Subjects and IRB

All EpiSLI subjects in this study were minors who assented to participation under the University of Iowa IRB# 200511767. Secondary genetic analyses for EpiSLI subjects were approved and carried out under the University of Iowa IRB #201406727. SPARK is approved under WIRB #201703201. Approval for data collection and analysis of the SPARK Research Match study described here was conducted under the the University of Iowa IRB #201705739. This research has been conducted using the UK Biobank Resource under Application Number #46053.

EpiSLI Cohort

The discovery cohort of this study, referred to as EpiSLI, was originally recruited as an epidemiological study of language impairments in Kindergarteners in the state of Iowa. A more detailed description of the recruitment scheme can be found in the initial (*69*), and subsequent (*21, 70*) publications. In brief, 7,218 kindergarteners, sampled to be representative of Iowa's population were screened for language impairment with a rapid, 40-item subset (*69*) of the TOLD-2P (*71*). Children who failed the TOLD-2P screener (i.e., have poor language ability) were over-sampled to capture a broad range of language ability in the primary cohort (n = 1,929 children). Children with autism spectrum disorder and/or intellectual disability were excluded. The entirety of the primary cohort received a more complete battery of language and cognitive assessments, while teachers and parents completed scholastic and behavioral questionnaires about the children 2nd, 4th, 8th, and 10th grades. Language and cognitive assessments included: TOLD-2P (*71*), WPPSI (*72*), a narrative story task (*73*), Woodcock Reading Mastery Tests-Revised (*74*), Word-Sound Deletion task (*75*), Random Animals-Colors task (*75*). We selected a total of 390 children from the primary cohort for whole genome sequencing. This sample represents a broad spectrum of language abilities, these 390 children were chosen by sampling from the tails of the distribution of a composite language score inward until we reached our final sample size (*8, 69*).

Behavioral assessment scores came from normed T-scores for the summary scales of the Child Behavior Checklist (CBCL) developed by ASEBA (*76*). CBCLs were completed at the 2nd grade timepoint of the study (241 of the 350 children used in this study had complete CBCL data for the

analysis in Supplementary Table 2).

Factor Analysis

An exploratory maximum-likelihood factor analysis was carried out on the language and cognitive assessment scores using the `factanal` function from the `stats` package in R (*77*), with the default "varimax" rotation. A number of factors from 1-10 were evaluated, with 7 - the largest number of factors with a nominal χ^2 p-value less than 0.05 - chosen for subsequent analyses. These seven factors accounted for an estimated 62.3% of the total variance in the original data.

EpiSLI Cognitive measures

Language assessment scores and IQ were normalized via the general scheme described in (*69*) to account for the over-sampling of individuals with low language ability. Scores were corrected for age as the residual of a linear regression model using `lm()` in R (*77*). The age-corrected language scores were then normalized with respect to the population using the `wtd.mean()` and `wtd.var()` functions from the `Hmisc` package (*78*) in R.

Cognitive data Imputation

Most individuals for whom whole genome sequencing was generated had scores from assessments given in 2nd grade, while a minority had only 4th grade (41/390) or kindergarten (39/390) scores. Missing scores were imputed using chained random forests and predictive mean matching as implemented the `missRanger` R package (*79*) with 5,000 trees and `pmm.k = 20`.

EpiSLI Whole Genome Sequencing

Sample Collection and whole genome sequencing

DNA was collected and extracted from either blood or saliva samples for 390 individuals (350 of which were used in the final analysis after quality control). DNA concentration of all sequenced samples was quantified with with Qubit 2.0 Fluorometer (Life Technologies Corporation). 390 DNA samples were sheared on a E220 Focused-ultrasonicator (Covaris) to an average size of 400 bp. Sequencing libraries were generated with a Kapa Hyper Prep kit (Kapa Biosystems) according

to the manufacturer protocol. All samples had an average genome-wide coverage of at least 20X and were sequenced on a HiSeq4000 (Illumina) with 150-bp Paired End chemistry.

Post-sequencing QC

All sequencing data was analysed with Fastqc (v0.11.8) (*80*), where no samples failed the run modules. After genome alignment (see below), some samples were found to have significantly better coverage than others, in excess of 40x. To help control for ascertainment bias in these samples, all samples with an average coverage > 35x (based on initially mapped BAM) were randomly down-sampled, leading to the distribution of genome-wide coverage and average insert-size outlined in Supplementary Table 17.

Genome Alignment and Variant Calling

Reads were processed with bcbio (v1.1.6) (*81*) and mapped to hg19 via BWA-mem (v0.7.17) (*82*). SNVs and InDels were then called with all samples in a pool with three variant callers: GATK (v4.1.2.0) (*83*), FreeBayes (v1.1.0.46) (*84*), and Platypus (v0.8.1.2) (*85*).

Variants from each caller were filtered according to caller-specific quality metrics. For GATK, variants needed to pass all VQSR tranche thresholds. For Platypus, variants were filtered based on the goodness of fit of genotype calls, excessive region-based haplotype scores, root-mean-square mapping quality, variant quality and its ratio with read depth, low complexity sequence context, allele bias, region-based read quality, neighboring homopolymers, and strand bias. For FreeBayes, variants were filtered based on a combination of allele frequency, read depth, and overall quality. The thresholds used for filtering Platypus and FreeBayes calls were the default set by bcbio. An ensemble of all three callers was generated and used in all subsequent analyses in order to achieve improved specificity in the detection of rare variants. All variants in the ensemble callset were called by GATK as well as either Platypus OR FreeBayes. The majority of variants (87%; 25,775,508) were called by all three callers, while a minority were called by GATK and only one other caller (13%; 3,937,146). All variants were filtered to have a QUAL and individual quality score ≥ 20. Qualified researchers can access the individual level genotypes from dbGaP at (study accession = phs002255.v1.p1): `https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002255.v1.p1`

Final sample QC

Starting from the initial sample size of 390, a total of 14 samples were flagged for removal from all association analyses because of population stratification. Specifically, these samples did not cluster with 1,000 Genomes Europeans (*86*), or were more than 3 standard deviations away from the rest of the EpiSLI cohort based on the top 10 multidimensional scaling components calculated from SNPs found at or above a 0.05 minor allele frequency. An additional 26 samples were dropped due to relatedness or limited phenotypic data, leaving a final sample size of 350 unrelated European individuals with complete data for all genomic analyses.

Variant Annotations

Variants were annotated with the Ensembl Variant Effect Predictor tool (VEP v109) (*87*). Reference population allele frequencies came from 1000 Genomes Phase 3 samples (*86*) and GnomAD (*88*), rsIDs of variants were dbSNP (*89*) (v151), these additional annotations were added using VCFanno (*90*) (v20190119).

Polygenic Scores (PGS)

Genotypes for PGS

To derive a set of SNPs suitable for PGS analysis, we merged our dataset with 1000 Genomes Europeans to use as a reference sample. Based on widely used recommendations (*91*), we extracted SNPs with a minor allele frequency $\geq 1\%$, Hardy-Weinberg equilibrium p-value $> 1 \times 10^{-6}$, and a missingness rate $< 2\%$ in both samples (leaving 7,719,665 SNPs total for PGS calculation).

PGS Calculation

LDpred2 was used to calculate a genome-wide PGS for all traits with the infinitesimal model using the provided UK Biobank LD reference panel and HapMap3+ variant set (*92*). PGS were calculated using GWAS summary statistics for neurodevelopmental and psychiatric traits: ADHD (*93*), addiction (*94*), alcohol dependency (*95*), Alzheimer's (*96*), Autism (*97*), Anorexia (*98*), bipolar disorder (*99*), cannabis use disorder (*100*), depression (*101*), epilepsy (*102*), insomnia (*103*), neurodevelopmental conditions (*104*), PTSD (*105*), schizophrenia (*106*), and Tourette's (*107*).

GWAS summary statistics for cognitive traits included: cognitive performance (*108*), educational attainment (*109*), executive functioning (*110*), and the "g Factor" (*111*). Additional PGS were calculated for the following behavioral and socioeconomic status related traits: childhood aggression (*112*), antisocial behavior (*113*), empathy (*114*), the BIG5 personality traits (*115*), income (*116*), and the Townsend Deprivation Index (a measure of material deprivation) (*117*). PGS were also calculated for brain structural (*118, 119*) and functional connectivity phenotypes (*120*). Finally, we computed PGS for miscellaneous traits: left-handedness (*117*), height (*121*), childhood trauma (*122*), and vocal pitch (*123*). It is important to note we did not compute PGS in EpiSLI for reading based traits, because our sample was part of the discovery cohort of the largest reading related GWAS to date (*9*). Associations for all of these PGS with our language factors can be found in Supplementary Table 3.

To account for population stratification, we corrected PGS for the first 5 genetic principal components. PGS were normed to the 1000 Genomes Europeans reference sample.

ES-PGS Calculation

We developed evolutionary stratified polygenic scores (ES-PGS) to identify genomic regions from specific evolutionary periods that disproportionately contribute to modern phenotypic variation. ES-PGS partitions polygenic score effects based on evolutionary annotations, testing whether variants from particular evolutionary epochs show disproportionately strong associations with traits compared to the broader genome.

The method uses three components: (1) annotation-specific PGS calculated from variants within evolutionary regions of interest, (2) background PGS calculated from all remaining genome-wide variants, and (3) matched control PGS from biologically similar regions to test specificity. Matched control regions were generated by sampling 1,000 random genomic segments for each annotation region (e.g., 10 annotation region would produce 10,000 random regions for matching), these regions are then matched for chromosome, size, GC content, repeat content, distance to nearest gene, and overlap with promoter or coding sequences. Control regions were required to be geq 100kb from any annotation region to ensure independence.

We implemented ES-PGS using nested model comparisons. The reduced model tests background genomic effects and matched control effects:

$$y = \beta_0 + \beta_1 \text{PGS}_{\text{background}} + \beta_2 \text{PGS}_{\text{matched}} + \epsilon \tag{S1}$$

The full model adds the annotation-specific term:

$$y = \beta_0 + \beta_1 \text{PGS}_{\text{background}} + \beta_2 \text{PGS}_{\text{matched}} + \beta_3 \text{PGS}_{\text{annotation}} + \epsilon \tag{S2}$$

Where y represents the phenotype (e.g., language factor scores), $\text{PGS}_{\text{background}}$ includes all variants except those in the focal annotation, $\text{PGS}_{\text{matched}}$ represents the matched control regions, and $\text{PGS}_{\text{annotation}}$ contains only variants within the evolutionary annotation. An ANOVA comparing these models tests whether the annotation contributes significantly beyond background genetic effects and matched controls. Additional covariates (genetic principal components, demographics) can be incorporated as needed.

PRSet (*27*) was used to calculate stratified PGS using human genome annotations in BED formatted files, computing a clumping and thresholding based ES-PGS for each annotation and a background PGS for the rest of the genome. Background regions for each annotation were identified using the "complement" function from bedtools (v2.26.0) (*124*). Matched control regions were matched using the `nullranges` R package (*125*). As recommended by the authors of PRSet, we used a p-value threshold of 1 and the 1000 Genomes Europeans as the LD reference (*86*). We corrected for population stratification using the same steps as described in the genome-wide PGS analysis.

We focused our analysis on genome annotations related to primate and human evolution, annotations included: primate ultra conserved regions (primate UCEs, Supplementary Data 5 from the reference) (*29*), great ape lineage accelerated regions (*30*)(Supplementary Data S1 filtered to "Hominidae" from the reference), Human Ancestor Quickly Evolved Regions (HAQERs, Supplementary Table 68 from the reference) (*31*), Human Accelerated Regions (HARs) (*32*), and ancient human selective sweep regions (Supplemental File S2 from the reference) (*33*). All annotations were converted to BED format and lifted over to match genome builds when necessary, using the UCSC liftOver tool (*126*). All annotation coordinates and matched regions can be found in our repository: LINK XX

ES-PGS replication in SPARK

To validate our ES-PGS results in a separate sample we used the SPARK cohort, a large genetic study of individuals with autism and their family members (*36*). For the replication, we utilized imputed SNP array data that we have previously described (*127*) to compute an ES-PGS using the same workflow as in our EpiSLI sample. The self-reported language and psychiatric diagnosis phenotypes, came from a questionnaire as part of a previously described research match where we had > 1,000 adults with autism or parents of children with autism complete an online language battery (*128*). We grouped all self-reported language-related diagnoses (language impairment, dyslexia, hearing impairment, stuttering, and speech impediments), and psychiatric diagnoses together (major depression, generalized anxiety disorder, bipolar, schizophrenia, substance abuse, OCD, and Tourette's). Given the large number of people without any language or psychiatric diagnosis, we used zero inflated Poisson regression models to determine the relationship of the HAQER CP-PGS with having any language or psychiatric diagnosis, as well as the number of language or psychiatric diagnoses. We used the log-ratio test to determine if adding the HAQER CP-PGS term significantly improved prediction accuracy in the zero inflated Poisson regression models. All other phenotypes came from the SPARK December 2014 phenotype release. All analyses included age, sex, background CP-PGS, HAQER matched control CP-PGS, and the first 5 genetic principal components as covariates.

SPARK rare ancestral reversion analysis

To explore the effects of rare genetic variation in evolutionary significant regions on language ability, we used the whole genome sequencing data from the previously described SPARK cohort (max N = 11,545) (*36*). Briefly, we merged the whole genome sequencing data provided by SPARK (WGS batches 1-4) using bcftools, then used the same processing pipeline as we did for the EpiSLI cohort: filtered to variants with a QUAL and individual quality score ≥ 20, and annotated the variants with VEP (v109) (*87*). To identify rare variants we then filtered variants in both datasets to have a maximal reference population allele frequency < 1% and an allele frequency < 1% in SPARK. Ancestral alleles were identified using those provided in the original HAQER manuscript (*34*). We identified all reversion variants within 10Kb of HAQERs from the updated set (*31*), HARs (*32*),

or random non-coding (RAND) (*34*) sequence and counted the number of rare ancestral reversions each sample had in each of these elements. We removed outlier samples who had > 2.5 median absolute deviations away from the median value for either HAQER, HAR, or random reversions, but had consistent phenotypic associations even when including these outliers in the analysis. We then used these reversion counts for association with speech and language phenotypes in SPARK.

ES-PGS replication in the UK Biobank

To further validate the effects of HAQERs on language abilities, we analyzed data from the UK Biobank cohort under Application Number #46053 (*66*). While the UK Biobank does not collect deeply phenotyped language data, they collect cognitive test scores relevant to language-related skills. For example, previous research has identified a verbal working memory task, where participants hear a sequence of digits and are asked to verbally repeat as many as they can, as highly relevant to language ability (*38*). In addition to this task, we analyzed educational attainment (a trait correlated with language abilities) and matrix reasoning, a measure of nonverbal fluid intelligence comparable to the F3 factor in our EpiSLI cohort.

We computed ES-PGS using the same pipeline established for the EpiSLI cohort, using the imputed v3 genotype data (filtered to SNPs with an imputation INFO score > 0.6 and a MAF > 1%), incorporating the first 20 genetic principal components (PCs) as covariates to mitigate the effects of population stratification. Our analysis focused on "white British European" samples; we removed all related individuals by identifying pairs with kinship score > 0.0442 using UK Biobank-provided relatedness data and randomly retaining one individual per pair, as well as removing anyone related to multiple participants. This filtering resulted in final samples of N = 84,433 unrelated individuals for the analysis of verbal working memory, N = 370,830 for educational attainment, and N = 26,051 for matrix reasoning scores. Because most large-scale cognitive GWASs include the UK Biobank in their discovery cohorts (including the cognitive performance GWAS used in our primary analyses (*28*)), for the ES-PGS we utilized the largest available GWAS on a quantitative language-related trait that excludes UK Biobank participants: a word reading ability GWAS (*9*). Although the word reading GWAS is substantially less powerful than the cognitive performance GWAS used in our primary analyses (N ≈ 30,000 versus N ≈ 250,000), it provided an independent cohort to further validate the effects of HAQERs on language-related skills. We applied rank inverse normal

transformation to all phenotypes and included year of birth, age, age^2, sex, and the first 20 genetic principal components (PCs) as covariates in the ES-PGS models.

Transcription Factor Analysis

Variant selection and annotation

We analyzed transcription factor binding sites in three distinct genomic contexts: Human Accelerated Regions (HARs), Human-Accelerated Quickly Evolved Regions (HAQERs), and matched random genomic regions (RAND, taken from (*31, 34*)). Position weight matrices (PWMs) for 633 human transcription factors were obtained from the JASPAR2020 database (*67*), with pseudocounts adjusted according to base frequency distributions.

Variants were filtered using strict quality control criteria as described above. For this analysis, we retained only biallelic single nucleotide variants (SNVs) located within feature boundaries that exhibited minor allele frequencies below 1% across all reference populations provided by VEP (*87*) and a minor allele frequency below 5% in our smaller EpiSLI sample. Complete great ape allele information was required for each variant, including data from the human reference (hg19), Neanderthals and Denisovans, Chimpanzee, Bonobo, Gorilla, and Orangutan genomes (*34, 129–132*). Briefly, archaic hominin genotype data came from VCFs with genotypes for the three high coverage Neanderthals (Altai (*129*), Chagyrskaya (*130*), and Vindija (*131*)) and one Denisovan (*132*) produced from high-coverage while genome sequencing were downloaded from the Max Planck Institute for Evolutionary Anthropology website and merged with our data. The final dataset comprised genotype information from 15,729 rare variant sites across 350 individuals, with corresponding variant annotations.

Reversion Status Determination

To characterize the evolutionary trajectory of variants, we developed a machine learning approach to impute hominin (human reference and most common allele from Neanderthal and Denisovan genomes matched) to human-chimp ancestral reversion status where direct determination was not available. We implemented an elastic net regression model (α = 0.9) using the glmnet package in R (*133*), incorporating great ape allele states as predictors of reversion status as given in (*34*). To

address class imbalance, we applied weights to ensure equal representation across sequence context types (HAQER, HAR, random). Reversion status was assigned using a probability threshold of 0.95 (i.e., at least a 95% predicted probability of being a human-chimp ancestral reversion), with known states preserved for training data.

Sequence and Motif Analysis

For each variant, we extracted 51-base pair genomic windows centered on the variant position from the human reference genome (hg19). Alternative sequences were generated by substituting variant alleles into the reference background. We then calculated maximal motif scores for both reference and alternative sequences across all JASPAR2020 human transcription factor motifs. Scores were computed on both forward and reverse complement strands, with the maximum score retained for reference and alternate alleles of each variant-motif pair.

Language Association Analysis

Sentence repetition ability, a proxy for general spoken language abilities, was assessed using the F1 measure as described above in the factor analysis. We computed burden scores for each transcription factor motif by combining variant effects weighted by genotype status of reversion sites. Linear regression models were used to estimate the associations between individual context-specific (aggregate) motif scores and language ability.

Transcription Factor Motif Score Selection Analysis

For each transcription factor, the (Z-scaled) difference in reference and alternate allele motif scores was modeled as a linear function of variant sequence context (HAQER, HAR, or random) and a binary reversion status indicator. Separate reversion effects were estimated for HAR, HAQER, and random region variants as a sequence context by reversion interaction term. Estimated beta coefficients from these terms, as well as their standard errors, were extracted for use in downstream analyses.

Joint Selection-Language Enrichment Analysis

We used York regression analysis (*134*) to examine relationships between hominin-divergent motif integrity and language-related effects. This approach accounts for uncertainty in both variables (i.e., language association betas and hominin-divergent binding effect betas). Prior to the analysis, the sign of the hominin-divergent binding binding betas were flipped such that positive values indicate human-gained motif integrity scores (i.e., when reversions to the human-chimp ancestral allele tend to decrease motif scores). York regression betas, their standard errors, a Chi-squared goodness of fit statistic and its p-value were extracted to interpret the significance of the overall relationship between TF motif integrity (i.e., motif score) and motif score effect on individual differences in language ability, for each sequence context. Individual TF motifs of interest are those with nominal significance ($p < 0.05$) for both selection for motif integrity and for positive association of aggregate motif integrity with higher sentence repetition (F1) scores. Transcription factors were classified into families using InterPro annotations (*135*) of representative families found to be significantly associated with a reversion effect in either direction (using www.string-db.org) (*136*). To identify transcription factor families showing convergent patterns of hominin selection and language association, we performed 2x2 Fisher's exact tests comparing the proportion of motifs with concordant effects (hominin gained binding AND positive language association vs. all other combinations) across families. Odds ratios with 95% confidence intervals and p-values were computed for each TF family.

Ancient DNA

DNA data and sample age information for ancient *Homo sapiens* came from the Allen Ancient DNA Resource (AADR) version 54. We downloaded the publicly available EIGENSTRAT formatted files and converted them to PLINK format using the EIGENSOFT tool. We then merged the ancient genomes with our EpiSLI and 1000 Genomes Europeans dataset to ensure we were using comparable SNPs for our ES-PGS selection analysis as we did in our discovery sample. We identified ancient west Eurasians using the same criteria as a recent large-scale selection analysis (*137*). Briefly, we filtered to samples found between longitude 25W and 60E and latitude 35N to 80N, samples passing quality control with an assessment labeled as "PASS", and sample ages > 0 but < 20,000 years old.

We then computed ES-PGS for CP in HAQERs using the same methodology as we did in the EpiSLI, SPARK, and ABCD samples for use in our polygenic selection analysis. Given the challenges of accounting for population structure in ancient DNA, we opted to use a LMM based approach instead of a traditional PC based approach as recommended by Akbari et al., 2024 (*137*). With the ancient west Eurasian subsample, we then identified independent SNPs using 1000 Genomes as the LD reference with PLINK's "–indep-pairwise" function (window size = 1000bp, step size = 1bp, r^2 threshold = 0.05, MAF ≥ 5%) (*138*). Next, we identified samples and SNPs with low missingness for GRM calculation (samples missing < 50% of independent SNPs, and SNPs missing in < 10% of those samples). Finally, we computed the genetic relatedness matrix (GRM) with GCTA (*139*) with the QC passing samples and SNPs and removed duplicate/twin samples for subsequent analysis (GCTA "grm-cutoff" of 0.9).

We then used the 3,244 QC passing samples and the GRM based on the 12,146 QC passing SNPs for the ES-PGS analysis. We implemented the LMM based polygenic selection analysis with the `gaston` (*140*) R package (`lmm.aireml` function), allowing us to account for the GRM which reflects population structure and relatedness of the sample. We used log10(sample age) as the outcome variable and HAQER CP-PGS and background CP-PGS as the independent variables (similar to the ES-PGS analysis we used in our other samples).

Additionally, we computed HAQER CP-PGS in the 10 available that are part of this release of the AADR. We compared these archaic hominins to ancient humans (described above), and 1000 Genomes Europeans, all data was processed together using the same set of SNPs to ensure polygenic scores would be comparable.

Detecting balancing selection in modern genomes

To detect signatures of balancing selection, we analyzed the WGS data we generated in EpiSLI. First, we subset to regions of interest (HAQERs, HARs, RAND, and HAQER matched control sequences). Then we computed a site frequency spectrum (SFS) for each sequence class, calculating the proportion of variants in each minor allele frequency bin. We compared HAQERs SFS to the SFS of HARs, RAND, and matched control sequences, to identify whether HAQERs had a relative enrichment of intermediate frequency variants -which can indicate balancing selection (or ongoing selection). To create 95% confidence intervals for HAQERs we generated 1,000 random bootstrap

samples of HAQERs (sampled with replacement) and computed the allele frequency proportions within in each sampled HAQER set, These confidence intervals allowed us to more confidently determine whether there was enrichment or depletion across MAF bins in HAQERs when compared to other sequence types.

Next, to more formally test our balancing selection hypothesis we identified common independent SNPs in these regions using PLINK (*138*), using the "–indep-pairwise" function (window size = 200bp, step size = 50bp, r^2 threshold = 0.5, MAF $\geq$ 5%) (*138*). This identified common independent SNPs in HAQERs, HARs, RAND sequences. We then computed individual level method of moments F-coefficients for each class of variation using the "–het" function in PLINK (*138*), which derives statistics based on the expected number of homozygotes and observed homozygotes (with more negative values indicating there are more heterozygotes than expected). We compared values between classes using paired t-tests to determine if there was excess heterozygosity in HAQERs, a signature of balancing selection.

ES-PGS analysis in ABCD

To explore the effects of HAQER CP-PGS in prenatal development we analyzed the ABCD cohort, a large longitudinal study of adolescent development with genetic, cognitive, brain imaging, and developmental phenotypes (*141*). All phenotypes used in ABCD came from the v4.0 data release. Similar to the SPARK replication analysis, we utilized imputed SNP array data that we have previously described (*127*) to compute ES-PGS using the same workflow as in our EpiSLI and SPARK samples. We computed genetic relatedness using GCTA (*139*) in the merged ABCD and SPARK dataset. Using the relatedness matrix, we identified unrelated individuals for ES-PGS analysis with brain imaging, cognitive, and birth phenotypes (genetic relatedness < 0.05). All evolutionary stratified polygenic scores were adjusted for genetic principal components, similar to the SPARK and EpiSLI cohorts. We included age, sex, and genetic principal components as covariates in the ES-PGS analyses.

To investigate potential evolutionary trade-offs underlying the balancing selection observed in HAQERs, we examined the relationship between HAQER CP-PGS and both cognitive performance and birth complications in ABCD. We employed a two-stage approach to address confounding and maximize signal detection. First, we applied a modified Remove Unwanted Variation using

residuals (RUVr) approach (*142*) to remove variance components in the phenotypic data that were orthogonal to the HAQER CP-PGS signal. Specifically, we identified and removed k = 15 principal components from the phenotype matrix that captured variance attributable to measured confounders (age, sex, ancestral background from genetic principal components) and unmeasured technical or biological factors not associated with HAQER CP-PGS. This denoising step preserved only phenotypic variance potentially related to our polygenic score of interest while removing confounding sources of variation. Second, we performed canonical correlation analysis (CCA) (*143*) on the RUVr-adjusted phenotypes to identify composite phenotypic dimensions that maximally correlated with HAQER CP-PGS. The CCA simultaneously analyzed two phenotypic domains (each having a block): cognitive performance measures (including memory, executive function, and processing speed tasks) and birth complication-related variables (including cesarean section delivery, intracranial volume measured from MRIs, and other obstetric risk factors). Sparsity constraints (λ = 0.75) were applied to the phenotypic loadings to improve interpretability. While we present correlation statistics for completeness, we acknowledge these may be inflated due to the optimization procedure; our primary inference focuses on the qualitative pattern of associations rather than specific effect sizes.

HAQER enrichment analysis

To determine enrichment between loci of interest (prenatal scQTLs, birth head circumference GWAS loci, and mammalian vocal learning enhancers) with our evolutionary regions of interest (HAQERs, HARs, and random sequence), we used the "intervalOverlap" function from gonomics (*144*). This allowed us to compute expected versus actual overlaps (based on size of the human genome and the provided annotations) providing an enrichment p-value for each region.

Evolutionary annotations

Given the two sets of HAQER annotations available (*31, 34*), created using different reference genomes and sequencing methodology we opted to use a conservative set of HAQERs for enrichment analyses to ensure any associations were not due to artifacts. The conservative set of HAQERs was identified by overlapping regions in the two separate HAQER sets using the bedtools "intersect" function (*124*), leaving nearly 900 overlapping HAQER regions. We included the set of HARs and

random sequence (matched to basic biological properties in HAQERs like size and non-coding sequence) from the original HAQER paper (*34*) for all enrichment analyses, allowing us to directly compare annotations. Similar to the other genomic analyses, we restricted analysis of all annotations to autosomal regions.

Defining human-specific and cell-type specific chromatin accessible regions

To identify whether HAQERs provided humans with novel regulatory mechanisms in specific neural cell-types, we examined data from snATAC-seq of human and mouse brains (*39*). To identify human-specific and cell-type specific cCREs, we used Suppelemental Table 6 provided by the authors to get hg38 coordinates of cCREs and filtered to "humanSpec" cCREs using data provided in Supplementary Table 20. Similarly, to identify human-mouse conserved cCREs, we used Suppelemental Table 6 provided by the authors and filtered to "CA_cons" cCREs using data provided in Supplementary Table 20. We added a 100Kb flank around all chromatin accessible regions and merged overlapping regions using the bedtools "merge" function, allowing us to test whether evolutionary annotations are enriched around cell-type specific chromatin accessibility regions.

Defining neurodevelopmental scQTLs

To determine if HAQERs, HARs, and RAND sequences significantly influence gene expression in prenatal and postnatal brains, we leveraged scQTLs from two studies. (1) a study of stem cell derived neurons from > 200 individuals, meant to mimic early prenatal cells in the midbrain (*49*) and (2) a study of nearly 400 postmortem adult brains from psychENCODE2 (*51*). For scQTL study (1), we defined scQTLs as all common variants (MAF > 5%) with a p-value $< 5 \times 10^{-4}$, for scQTL study (2) we utilized all provided scQTLs defined as significant by the authors downloaded from `https://psychscreen.wenglab.org/psychscreen/downloads`.

Defining birth head circumference regions

We identified birth head circumference associated genomic loci with a genome-wide suggestive p-value (p-value $< 5 \times 10^{-5}$) and a MAF > 5%, using the largest GWAS available to date (*52*).

Given the influence of LD on human traits, we added a 100Kb flank around all associated SNPs and merged overlapping regions using the bedtools "merge" function, allowing us to test whether evolutionary annotations are enriched around associated SNPs.

Defining mammalian vocal learning enhancer regions

We gathered previously established mammalian vocal learning enhancer regions from a recent cross-species analysis (*15*), which identified 50 open chromatin regions in the motor cortex strongly associated with vocal learning. Since these regions were provided in mouse genome coordinates (mm10), we lifted these over to hg19 and we added 1Mb flanks around these enhancer regions which were then used to test whether evolutionary annotations are enriched near established vocal learning enhancers.

Convergent evolution of vocal learning analysis

To test for evidence of convergent evolution of "HAQER-like" sequences in vocal learning species, we utilized a large dataset of > 400 species with whole genome data aligned to the human reference genome (hg38) (*55*). We parsed the alignments with Biopython (*145*) to subset to regions of interest. For each species and sequence type (HAQERs or HARs), we computed a sequence similarity (using the number of bases matching the human reference genome in these regions). We then used this "HAQER-like" sequence similarity to predict vocal learning status in 170 non-primate species that have been previously described (*15*). To determine statistical significance and account for species relatedness we used phylogenetic logistic regression (*56*) with the `phylolm` package (*57*) in R. We did similar analyses for brain size and birth:adult weight ratio using species level phenotype data from the PanTHERIA dataset (*68*).

Statistical analysis

All statistical analysis was done in R (version 4.3.1) (*77*).

Approach to multiple testing correction and triangulating evidence

To address multiple testing concerns, we employed False Discovery Rate (FDR) correction for analyses testing more than 20 hypotheses, allowing us to limit type 1 errors (false positives). This was applied to the EpiSLI CBCL mental health score analysis, genome-wide PGS analysis, and the ES-PGS analysis where we conducted many statistical tests.

We adopted a comprehensive approach to validate our key findings through multiple lines of evidence. We established confidence in our results through: (1) replication across independent cohorts (e.g., EpiSLI, SPARK, ABCD, and the UK Biobank); (2) identifying convergent evidence across different analytical approaches (e.g., common variant, rare variant, and transcription factor binding analyses); (3) demonstration of specificity, showing HAQERs' associations with language but not nonverbal IQ; (4) evolutionary support from ancient DNA and cross-species analyses; and (5) mechanistic support through analysis of transcription factor binding, enrichment for human-specific and cell-type specific chromatin accessible regions, and enrichment of variants influencing prenatal gene regulation. This multi-faceted approach allowed us to distinguish robust biological signals from statistical noise, even when some individual analyses showed moderate statistical significance. Our findings' consistency across diverse data types, species, and analytical methods provides stronger evidence than would be achieved through any single statistical test, regardless of its p-value.

Supplementary Figures

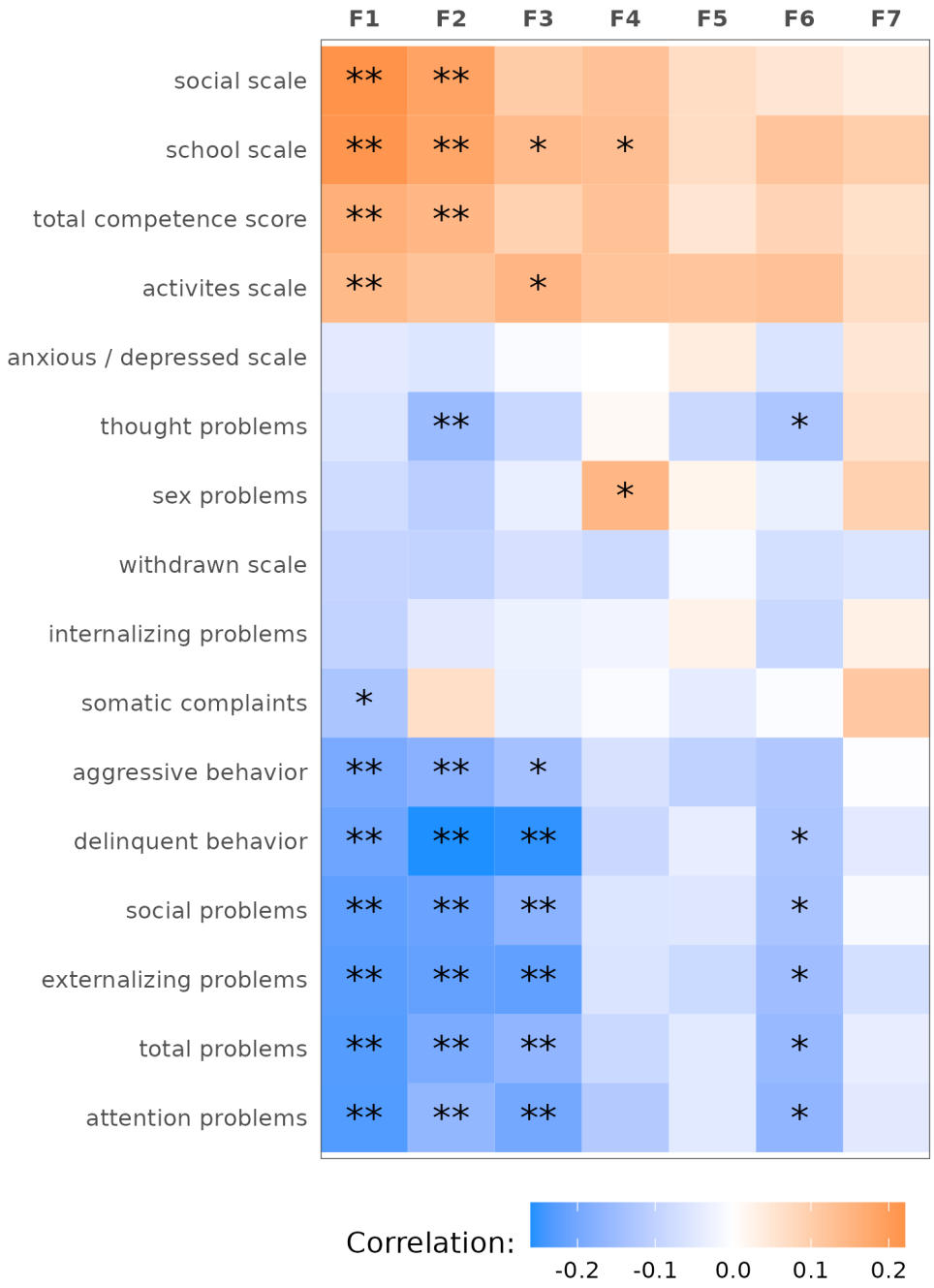


Figure S1: **Mental health correlates of EpiSLI factors**

Heatmap of Pearson correlation coefficients (r) of CBCL scales with EpiSLI factor scores (N = 241 individuals). Color indicates the association between that CBCL subscale score and the EpiSLI factor score (orange = positive correlation, blue = negative correlation. "**" indicates FDR < 0.05, and "*" indicates unadjusted p-value < 0.05.

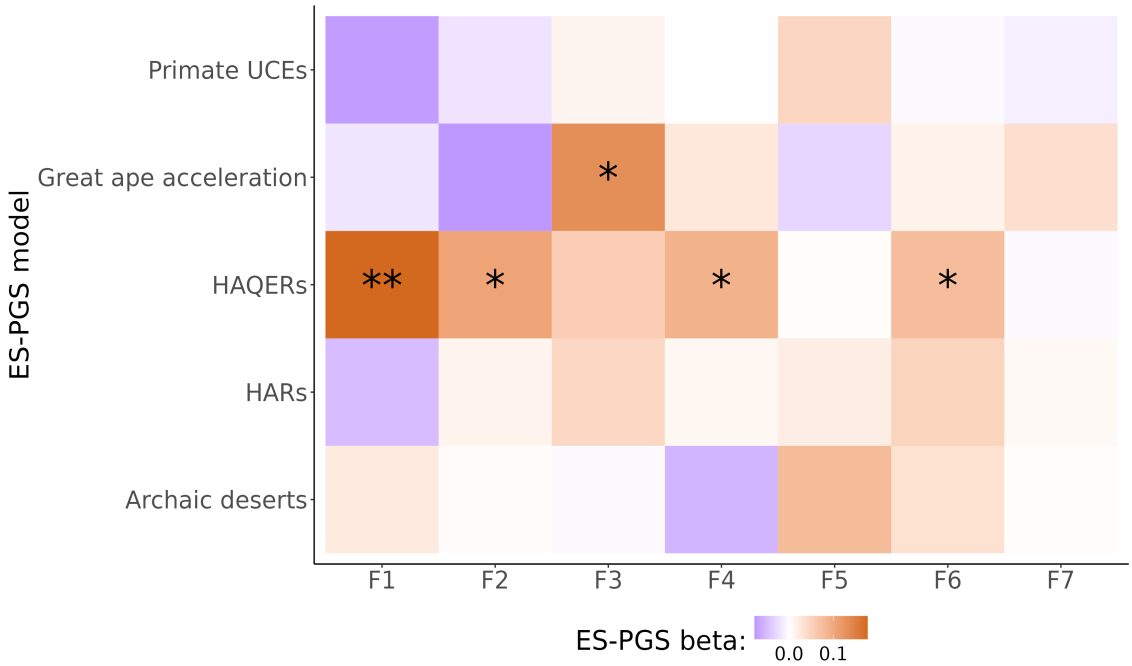


Figure S2: **Evolutionary history of EpiSLI factors**

Heatmap of ES-PGS β's stratifying the CP-PGS based on evolutionary annotations. Color indicates the association between that evolutionary annotation and the EpiSLI factor score (orange = positive ES-PGS β, purple = negative ES-PGS β). "**" indicates FDR < 0.01, and "*" indicates unadjusted p-value < 0.05.

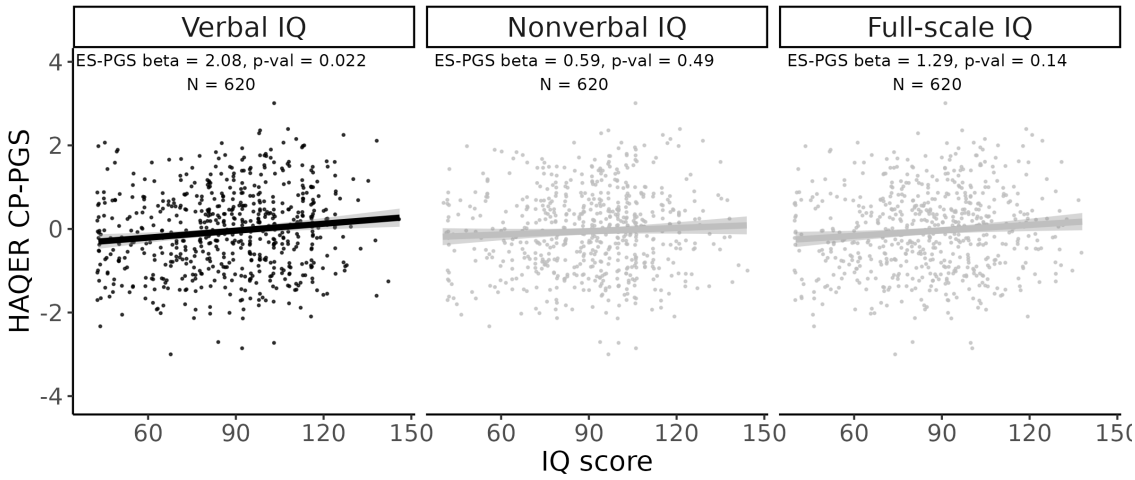


Figure S3:

HAQER CP-PGS is associated with verbal IQ in an independent cohort

ES-PGS results of HAQER CP-PGS and clinical IQ assessment scores from the SPARK dataset (N = 620).

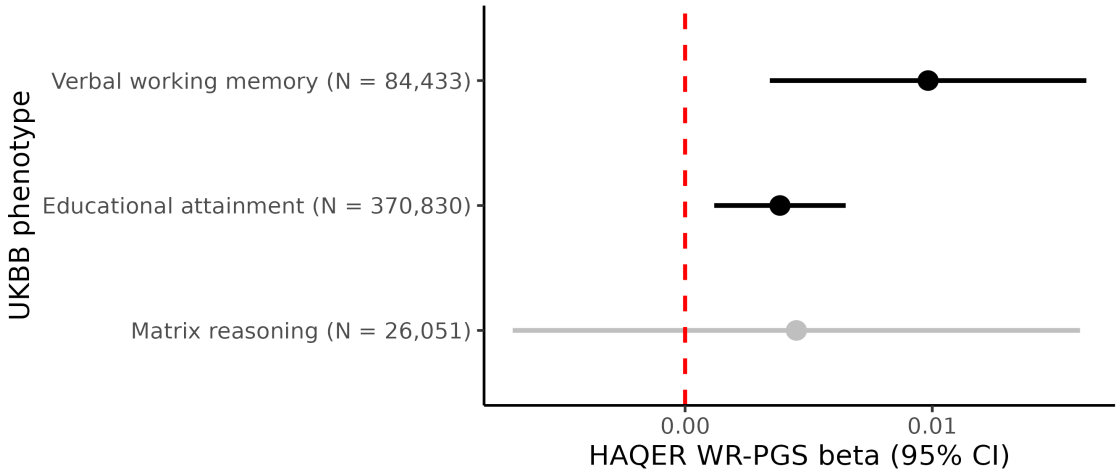


Figure S4:

HAQER WR-PGS is associated with verbal working memory in an independent cohort

ES-PGS results of the HAQER word reading polygenic score (WR-PGS) and relevant cognition-related traits in the UK Biobank.

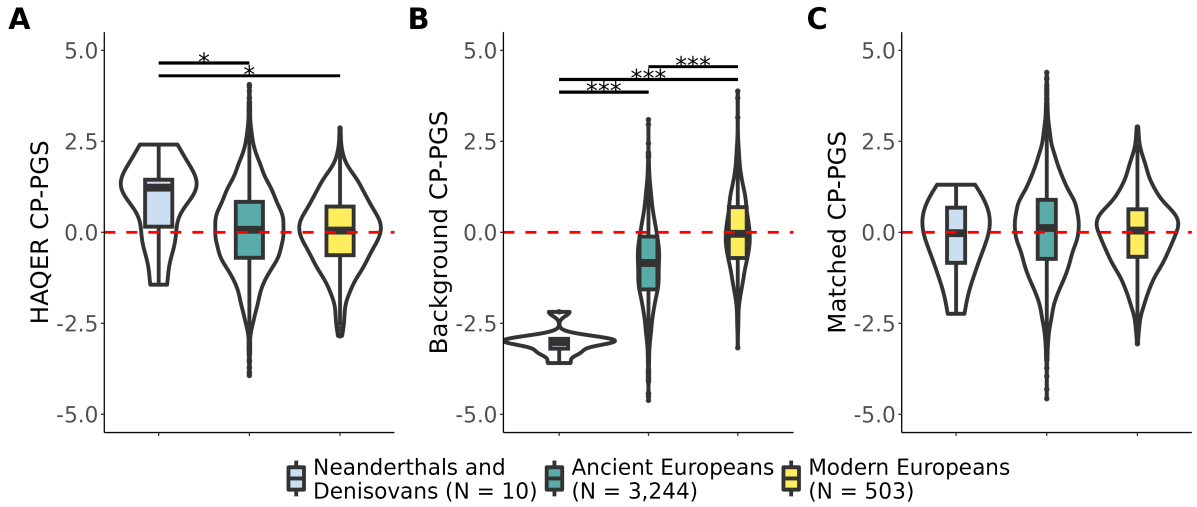


Figure S5:

ES-PGS differences across human history

A-C Distribution of HAQER CP-PGS (**A**), background CP-PGS (**B**), or random matched control regions (**C**) across archaic humans (neanderthals and denisovans from the AADR), ancient anatomically modern humans (AADR), and modern Europeans (1000 Genomes dataset). "*" indicates a p-value of < 0.05 from a t-test comparison, and "***" indicates p-value < 0.001.

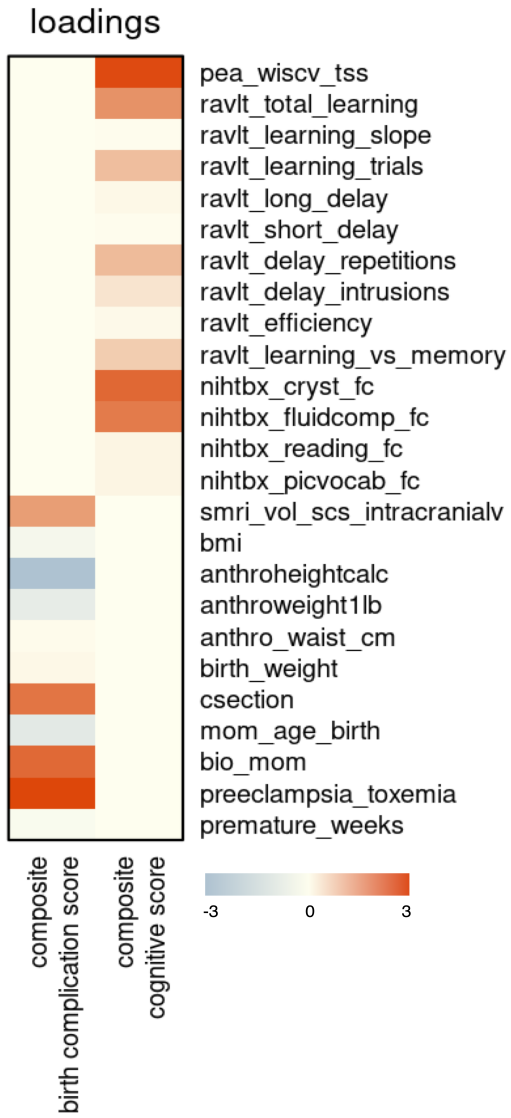


Figure S6: **Phenotypic dimensions of cognition and birth complications**

Composite phenotypic scores derived from ABCD cohort data using regularized canonical correlation analysis with HAQER CP-PGS. This method identifies linear combinations of variables maximally correlated with HAQER CP-PGS while enforcing sparsity (parameter = 0.75). Block 2: 14 cognitive measures (RAVLT, NIH Toolbox cognition, reading, vocabulary, matrix reasoning). Block 3: 11 birth/development measures (intracranial volume, anthropometrics, birth weight, delivery mode, maternal factors, prematurity). Heatmap shows sparse canonical loadings; Component 1 = composite birth complication score, Component 2 = composite cognitive score. Orange = positive loadings onto component, while blue = negative loading.

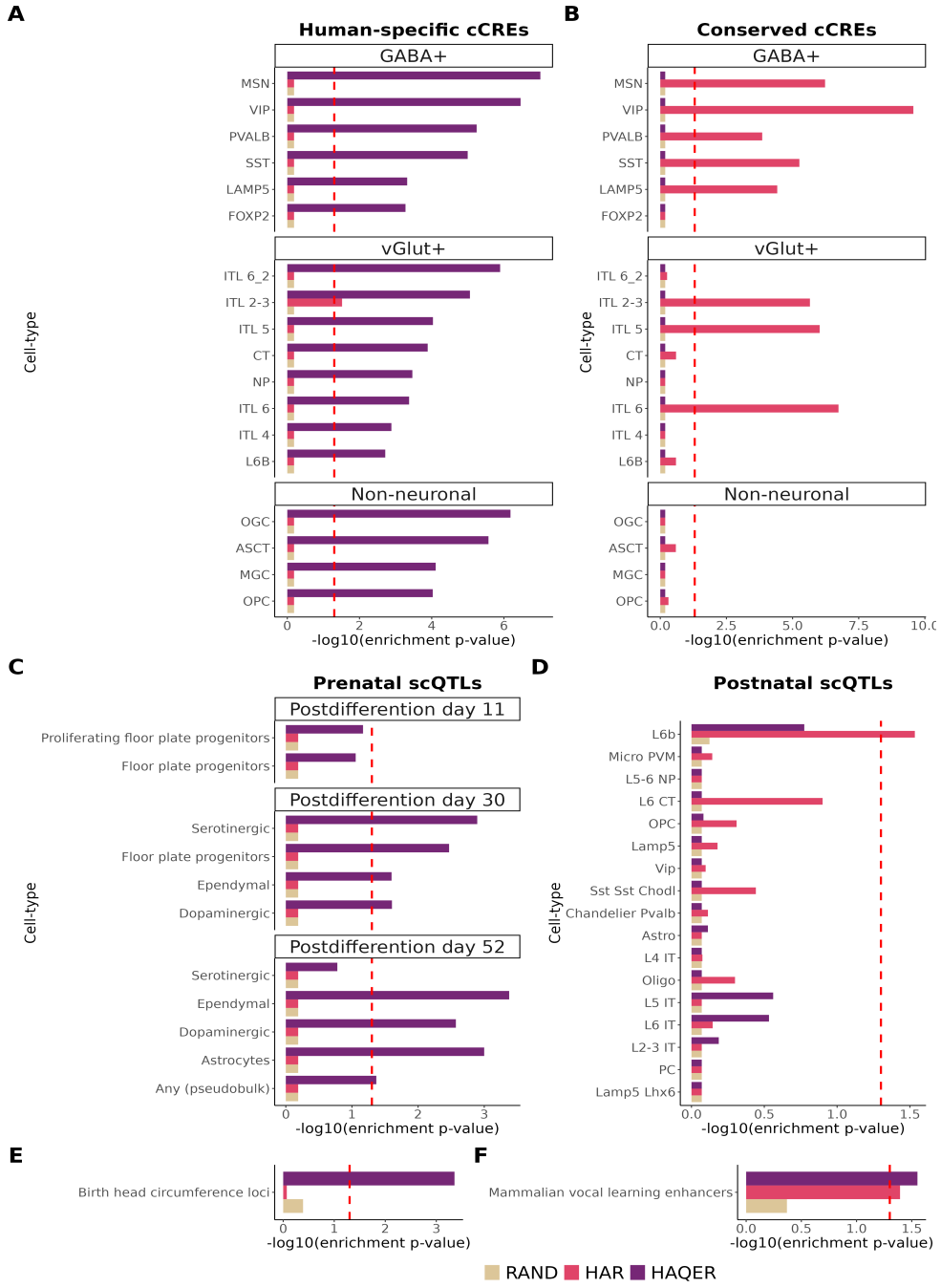


Figure S7: **HAQERs influence prenatal brain development**

A-B Enrichment of HAQERs, HARs, and RAND for brain cCREs from human-specific (**A**) or human-mouse conserved (**B**) elements. **C-D** Enrichment for regulatory variants (scQTLs) across timepoints and cell-types in newly differentiated neurons (**C**) or adult post mortem brains (**D**). **E** Enrichment for common SNPs (MAF > 5%) associated with head circumference at birth. **F** Enrichment for enhancer regions associated with vocal learning across mammals (Wirthlin et al., 2024).