Genomic contributions to infant and toddler vocabulary scores: Implications for association with health-, cognition-, and behaviour-related outcomes

Ellen Verhoef[1], Andrea G. Allegrini[2], Philip R. Jansen[3,4,5], Katherine Lange[6,7], Carol A. Wang[8,9], Angela T. Morgan[6,7,10,11], Tarunveer S. Ahluwalia[12,13,14], Christos Symeonides[6,11,15], EAGLE working group, Hans Bisgaard[12], Else Eising[1], Marie-Christine Franken[16], Elina Hypponen[17,18], Toby Mansell[6,7], Mitchell Olislagers[1,19], Emina Omerovic[6], Kaili Rimfeld[2,20], Fenja Schlag[1], Saskia Selzam[2], Chin Yang Shapland[21,22], Henning Tiemeier[3,23], Andrew J.O. Whitehouse[24], Richard Saffery[6,7,25], Klaus Bønnelykke[12], Sheena Reilly[6,7,26], Craig E. Pennell[8,9,27], Melissa Wake[6,7,28], Charlotte A.M. Cecil[3,29,30], Robert Plomin[2], Simon E. Fisher[1,31], Beate St Pourcain[1,21,31]

1. Language and Genetics Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
2. Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
3. Department of Child and Adolescent Psychiatry/Psychology, Erasmus University Medical Center, Rotterdam, The Netherlands
4. Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU University, Amsterdam, The Netherlands
5. Section Clinical Genetics, Department Human Genetics, Amsterdam University Medical Centers, Amsterdam, The Netherlands
6. Murdoch Children's Research Institute, Parkville, VIC, Australia
7. Department of Paediatrics, University of Melbourne, Parkville, VIC, Australia
8. School of Medicine and Public Health, The University of Newcastle, Newcastle, NSW, 2308, Australia
9. Mothers and Babies Research Program, Hunter Medical Research Institute, Newcastle, New South Wales, Australia, NSW 2305
10. Department of Audiology and Speech Pathology, University of Melbourne, Parkville, VIC, Australia
11. Royal Children's Hospital, Melbourne, Victoria, Australia

12. COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, Denmark
13. Steno Diabetes Center Copenhagen, Herlev, Denmark
14. The Bioinformatics Center, Department of Biology, University of Copenhagen, Copenhagen 2200, Denmark
15. Minderoo Foundation, Perth, WA, Australia
16. Erasmus University Medical Center, Sophia Children's Hospital, Department of Otorhinolaryngology, The Netherlands
17. Australian Centre for Precision Health, Unit of Clinical and Health Sciences, University of South Australia, Adelaide Australia, SA5000
18. South Australian Health and Medical Research Institute, Adelaide Australia
19. Department of Urology, Erasmus University Medical Center, Erasmus MC Cancer Institute, Rotterdam, the Netherlands
20. Department of Psychology, Royal Holloway University of London, London, UK
21. MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK
22. Population Health Sciences, University of Bristol, Bristol, UK
23. Harvard, T.H. Chan School of Public Health, Boston, MA, USA
24. Telethon Kids Institute, The University of Western Australia, Perth, Western Australia, Australia, WA 6009
25. Chongqing Medical University, Chongqing, China
26. Menzies Health Institute Queensland, Griffith University, Queensland, Australia
27. Maternity and Gynaecology John Hunter Hospital, Newcastle, New South Wales, Australia, NSW 2305
28. The Liggins Institute, The University of Auckland, Grafton, New Zealand
29. Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands
30. Molecular Epidemiology, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands
31. Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

Corresponding author: Beate St Pourcain

Postal address: Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525 XD Nijmegen, the Netherlands

e-mail: Beate.StPourcain@mpi.nl / Phone: +31 24 3521964 / Fax: +31 24 3521213

Requests for reprints should be sent to the corresponding author.

Key Points

Question: What is the genetic architecture underlying vocabulary acquisition during language development, and does it affect links with later-life outcomes?

Findings: At least two genetic components contribute to vocabulary size, predominantly distinguishing infant expressive from toddler receptive vocabulary. Matching patterns of genetic overlap were found with later-life outcomes: Larger infant expressive but smaller toddler receptive vocabulary size was correlated with higher ADHD risk and/or childhood maltreatment exposure (a behavioural proxy). Consistently, later-life cognition was associated with toddler vocabulary scores only, irrespective of power.

Meaning: The genetic architecture underlying vocabulary acquisition is dynamic, shaping polygenic associations with later-life behaviour and cognition.

Abstract

Importance: The number of words a child produces (expressive vocabulary) and understands (receptive vocabulary) change rapidly during infancy and toddlerhood, partially due to genetic factors. However, the genetic architecture underlying vocabulary development and association patterns with later-life outcomes, have not yet been fully characterised.

Objective: To (i) study the multivariate genetic architecture underlying vocabulary size during infancy and toddlerhood, and (ii) describe polygenic association patterns with childhood behavioural and health measures, as well as adult cognition-related outcomes.

Design: Meta-genome-wide association study (meta-GWAS) of expressive and receptive vocabulary (age: 15-38 months) performed within the Early Genetics and Life Course Epidemiology (EAGLE) Consortium. Structural equation modelling techniques were applied to study multivariate genetic architectures.

Setting: Children of European descent across seven independent population-based cohorts.

Participants: 37,913 observations from 17,298 individuals.

Main Outcome and Measure: Meta-analyses were performed for early-phase expressive vocabulary (15-18 months), late-phase expressive vocabulary (24-38 months), late-phase receptive vocabulary (24-38 months), and combinations thereof. Vocabulary size was assessed by parent report using standardised psychological instruments.

Results: Common genetic variation explained a modest proportion of phenotypic variation across all vocabulary measures (Single-Nucleotide Polymorphism heritability: 0.08(SE=0.01) to 0.24(SE=0.03)). Genetic correlation (r_g) analyses showed that late-phase expressive vocabulary largely shared genetic influences with both early-phase expressive (r_g=0.69(SE=0.14)) and late-phase receptive vocabulary (r_g=0.67(SE=0.16)). However, the latter two measures were genetically unrelated (r_g=0.07(SE=0.10)),

suggesting different underlying genetic factors. Consistently, we observed differences in polygenic association patterns: Larger early-phase expressive vocabulary size was genetically correlated with increased ADHD risk (r_g=0.23(SE=0.08)) and childhood maltreatment exposure (r_g=0.19(SE=0.07)), a behavioural proxy. In contrast, larger late-phase receptive vocabulary size was genetically correlated with lower childhood maltreatment exposure (r_g=-0.33(SE=0.08)). Finally, toddler, but not infant, vocabulary size was linked to cognitive skills (e.g. late-phase expressive vocabulary and intelligence: r_g=0.32(SE=0.08)), despite comparable power.

Conclusions and Relevance:

There are at least two distinct genetic components contributing to vocabulary development during infancy and toddlerhood that shape polygenic association patterns with later-life cognition and ADHD-related traits. Our findings suggest differences in biological mechanisms during a phase where children "learn to speak" (infancy) compared to a phase where children mastered some fluency and "speak to learn" (toddlerhood).

Introduction

Language development in infants and toddlers is often assessed with measures of expressive and receptive vocabulary[1,2]. These constructs relate to language production and understanding, respectively, and can be relatively easily (albeit indirectly) measured through parental reports. The first spoken words, representing one of the milestones in language development, typically emerge between the ages of 10 to 15 months[2]. Receptive vocabulary development, usually, precedes expressive vocabulary development, emerging at six to nine months of age[3]. Consequently, the number of words children understand is often larger than the number of words they produce, and exceeds the latter at least four-fold based on parent-reported measures at 16 months of age[4]. Once children reach an expressive vocabulary size of ~50 words at an age of 12 to 18 months, there is often a period of rapid vocabulary growth around 16 to 22 months of age[5], resulting in a vocabulary size of 100 to 600 words at ~24 months[4]. During the early stages of language learning in infancy (≤18 months of age) children typically produce words in isolation[2], followed by a period of two-word combinations and more complex grammatical structures[4,6].

Individual differences in early-life vocabulary development can, partially, be explained by genetic factors[7–10]. Twin heritability (twin-h^2) estimates for expressive vocabulary range between 10% and 25% (24-36 months)[7–9], reflecting phenotypic variation due to all possible genetic influences. These findings are corroborated by genetic research investigating Single-Nucleotide Polymorphisms (SNPs), with SNP-h^2 estimates of 13% and 14% (15-30 months)[7]. For receptive vocabulary at 14 months of age, a twin-h^2 estimate of 28% has been reported[11]. Evidence for SNP-h^2 at a similar age was poor[12], based on a single-cohort study, but was present at 38 months of age (12% SNP-h^2)[12].

Population-based studies of English-speaking children showed that the genetic architecture of language development spanning infancy to early childhood is complex, with evidence for both stability and change in underlying genetic contributions[7,8,12,13]. At the genome-wide level, genetic correlations (r_g) for measures of expressive vocabulary between 15 and 38 months of age ranged from 0.48 to 0.74[7,8,12],

suggesting moderate-to-strong genetic stability. At the individual SNP-level, a previous meta-genome-wide association study (meta-GWAS, N=8,889) identified a GWAS signal at rs7642482 on chromosome 3p12.3, near the *ROBO2* gene[7] that was associated with expressive vocabulary in infants (age: 15-18 months) but attenuated in toddlers (age: 24-30 months)[7]. These changes might be due to age-specific genetic mechanisms, highlighting the need for more powerful studies to identify and characterise genetic association.

Genetic influences underlying early-life vocabulary are shared with many later childhood abilities. In UK twins, for example, early expressive language skills (24-48 months) were moderately genetically correlated (r_g=0.36) with childhood reading abilities[13]. Similarly, in a UK population-based genomic study, receptive vocabulary scores (38 months) showed moderate-to-strong genetic correlations (r_g=0.58-0.92) with mid-childhood reading skills[14]. Genetic influences underlying early-life vocabulary may also be shared with other childhood behavioural and health measures, including childhood-onset neurodevelopmental conditions as Attention-Deficit/Hyperactivity Disorder (ADHD) and Autism Spectrum Disorder (ASD). For example, children with ADHD often experience difficulties with mastering language and literacy skills[15–17] and poor language skills at the age of three years were found to be predictive of inattention and hyperactive symptoms two years later in life[18]. More specifically, there is evidence for genetic overlap between higher ADHD risk and lower mid-childhood/early-adolescent language- and literacy-related abilities, primarily implicating reading performance[19–22], underscoring the need to study also genetic links with early-life language measures. For children diagnosed with ASD, the phenotypic spectrum is wider, including children with little or no spontaneous spoken language by the time they reach school age[23] as well as individuals with, comparably, few problems in the language domain[24]. Both, children with ADHD and ASD are at increased risk for childhood maltreatment compared to typically developing children[25,26], consistent with moderate genetic trait correlations (ADHD: r_g=0.56, ASD: r_g=0.41)[27]. Such genetic relationships may capture gene-environment correlation, as children's behaviour (partially influenced by genetic factors) elicits

responses from parents and others[27]. In addition, we assess evidence for genetic links between children's vocabulary and head circumference measures, a characteristic that may indicate abnormal development[28] and is highly correlated with MRI brain volume[29,30].

In this study, we aim to elucidate the polygenic architecture of vocabulary acquisition. We investigate developmental changes in genetic contributions at the single-variant and trait covariance level by performing genome-wide association meta-analyses of expressive and receptive vocabulary size at different developmental stages during infancy and toddlerhood, including an early, single-word phase (15-18 months) and a late (24-38 months) phase during which children start using two-word combinations and more complex grammatical structures. Furthermore, we characterise polygenic links with childhood behavioural and health measures, as well as adult cognition-related outcomes, fitting structural models.

Methods

Phenotype selection and study design

Cohorts with quantitative vocabulary scores assessed during the first three years of life and genome-wide genotypes were invited to participate in this study, embedded within the Early Genetics and Life Course Epidemiology (EAGLE) consortium[31] (https://www.eagle-consortium.org/working-groups/behaviour-and-cognition/early-language/). Expressive vocabulary scores were assessed between 15 and 38 months of age and analysed across two developmental stages to allow for age-specific genetic influences: an early phase (15-18 months) and a late phase (24-38 months). Scores for receptive vocabulary were included for the late phase only (24-38 months), due to low measurement availability, low reliability and little evidence for SNP-h^2 during the early phase[12] (eMethods 1).

Up to seven population-based cohorts (eMethods 2) participated in the meta-analyses, of which two had longitudinal vocabulary assessments (Figure 1, eTable 1). Vocabulary scores were ascertained by

parental report using age-specific word lists that were adapted from the MacArthur Communicative Development Inventory (CDI)[9,32–36] or the Language Development Survey (LDS)[37] (eTable 1, eMethods 3). Ethical approval was obtained by the local research ethics committee for each participating study, and all parents and/or legal guardians provided written informed consent (eMethods 2).

Genotyping and imputation

Genotyping within each cohort was conducted using high-density SNP arrays and quality control followed standard procedures[38] (eTable 2). In total, between 440,476 and 608,517 high-quality autosomal genotyped markers were imputed against a Haplotype Reference Consortium (HRC) r1.1 panel[39] (eTable 2).

Single variant association analyses and meta-analyses

Within each cohort, vocabulary scores were adjusted for potential covariates and rank-transformed to achieve normality and to allow for comparisons of genetic association effects across different psychological instruments (eMethods 4). SNP-vocabulary associations were then estimated within each cohort using linear regression of rank-transformed residuals on posterior genotype probability, except for the Longitudinal Study of Australian Children (LSAC) where we analysed best-guess genotypes, assuming an additive genetic model (eTable 2, eMethods 4). Prior to meta-analysis, GWAS summary statistics underwent extensive quality control using the EasyQC R package[40] (v9.2) (eTable 2, eMethods 4).

As part of analysis stage I, single-trait meta-analyses were performed for early-phase expressive vocabulary, late-phase expressive vocabulary and late-phase receptive vocabulary using either METAL[41] and/or multi-trait analysis of genome-wide association (MTAG)[42] software (Figure 1, eMethods 4). As part of analysis stage II, multi-trait meta-analyses were performed with MTAG[42] combining genetically correlated vocabulary traits to increase statistical power (Figure 1).

The number of independent vocabulary measures analysed in this study was 2.38, as estimated with a Matrix Spectral Decomposition (matSpD)[43] based on bivariate genetic correlations (see below) across the three single-trait vocabulary meta-analyses (stage I). This resulted in a multiple-testing-adjusted genome-wide association significance threshold of $P<2.10\times10^{-8}$ ($5\times10^{-8}/2.38$).

FUMA analyses

SNP-vocabulary associations passing the unadjusted genome-wide significance threshold ($P<5\times10^{-8}$) were identified and annotated using Functional Mapping and Annotation of genetic associations[44] software (FUMA v1.3.6). In addition, gene-based genome-wide, gene-set and gene-property analyses were conducted with Multi-marker Analysis of GenoMic Annotation (MAGMA, v1.08) within FUMA[44] (v1.3.6a) (eMethods 5).

SNP-heritability and genetic relationship analyses

Using GWAS summary statistics, SNP-h^2 and bivariate genetic correlations (r_g) were estimated for early-life vocabulary (meta-analysis stages I and II) and/or later-life traits using High-Definition Likelihood[45] (HDL, eMethods 6, see below).

Given evidence of HDL-SNP-h^2 ($P<0.05$), HDL-r_g analyses (eMethods 6) were performed to assess genetic overlap (i) across single-trait vocabulary measures (stage I) and (ii) across early-life vocabulary measures and later-life health-, cognition-, and behaviour-related outcomes (eMethods 6): reading performance in the general population (8-22 years, N=13,027; GWAS summary statistics were created as described in eMethods 7, eTable 3), intelligence[46] (5-98 years, N=279,930), educational attainment[47] (>30 years, N=766,345), infant head circumference[48] (6-30 months, N=10,768), childhood head circumference[49] (6-9 years, N=10,600), childhood aggressive behaviour[50] (1.5-18 years, N=151,741), childhood internalising

symptoms[51] (3-18 years, N=64,641), childhood maltreatment exposure[52] (<18 years, N=150,290), ADHD[53] (N=53,293; N_{cases}=19,099) and ASD[54] (N=46,350; N_{cases}=18,381). The multiple-testing-adjusted threshold for HDL-r_g analyses was defined at 5.32×10^{-3}, reflecting a correction for 9.39 independent traits considered (0.05/9.39), estimated using matSpD[43,55] and a bivariate genetic correlation matrix (eFigure 1).

Polygenic prediction of late-phase expressive vocabulary was carried out using polygenic scoring[56] (eMethods 8).

Structural equation models

We modelled the multivariate genetic architecture between early-life vocabulary measures and genetically associated later-life outcomes (as identified with HDL-r_g analyses, $P<5.32 \times 10^{-3}$) using genomic structural equation modelling (genomic SEM)[57] and genome-wide summary statistics (eMethods 9). Follow-up analyses of phenotypic, genetic and residual correlations of mid-childhood and early adolescent ADHD symptoms with infant and toddler vocabulary measures were performed using individual-level data from the Avon Longitudinal Study of Parents And Children (ALSPAC) cohort and genetic-relationship-matrix structural equation modelling[58] (GRM-SEM) (eMethods 10, eTable 4).

Results

Single-trait and multi-trait meta-GWAS

Single-trait genome-wide association analyses based on children of European descent from seven independent cohorts were performed for early-phase expressive vocabulary (15-18 months, N=8,799), late-phase expressive vocabulary (24-38 months, N=16,615) and late-phase receptive vocabulary (24-38 months, N=6,291) (stage I, Figure 1). There was little evidence for novel SNP association signals with vocabulary size at the multiple-testing-adjusted genome-wide significance level ($P<2.10 \times 10^{-8}$, eFigure 2a-

c). For early-phase expressive vocabulary, a single GWAS signal passed the unadjusted genome-wide significance threshold (rs9854781, $P<5x10^{-8}$), consistent with a known locus (rs764282, LD-r^2=0.78) identified through a previous meta-GWAS studying overlapping samples[7]. Genome-wide gene-based, gene-set and gene-property analyses with MAGMA[59] did not provide evidence for gene-based association passing the multiple-testing-adjusted significance thresholds (eFigure 3, eTable 5).

All vocabulary measures were modestly heritable (Figure 2a, eTable 6), with SNP-h^2 estimates of 0.24(SE=0.02), 0.08(SE=0.01), and 0.20(SE=0.04) for early-phase expressive vocabulary, late-phase expressive vocabulary, and late-phase receptive vocabulary, respectively. Given limited data availability, polygenic prediction (out of the meta-analysis samples) was carried out for late-phase expressive vocabulary only (β=0.04(SE=0.04), P=0.35, R^2=0.14%), though power was low (≤0.11) due to a combination of low SNP-h^2 and low target sample size (N=639). In contrast, genetic correlations between early- and late-phase expressive vocabulary (r_g=0.69(SE=0.14)), as well as between late-phase expressive and receptive vocabulary (r_g=0.67(SE=0.16)) were moderate (Figure 2b), suggesting some stability in genetic factors during development. Genetic influences underlying early-phase expressive vocabulary were, however, largely independent of genetic influences related to late-phase receptive vocabulary (r_g=0.07(SE=0.10)). Given comparable power to detect r_g=0.70 with late-phase receptive vocabulary for both expressive vocabulary measures (power: early-phase=0.83, late-phase=0.71), these findings suggest developmental genetic heterogeneity.

To maximise power for single-variant discovery, genetically correlated vocabulary measures were combined as part of two multi-trait meta-analyses using MTAG (stage II, Figure 1, eTable 7). However, we neither identified further SNP-vocabulary associations (eFigure 2d-e) nor increased evidence for SNP-h^2 (Z-scores, Figure 2a, eTable 6).

Genetic relationships with later-life outcomes

We investigated genetic links between vocabulary traits (stage I) and multiple heritable health-, cognition-, and behaviour-related traits, especially during childhood (see eTable 8 for SNP-h^2) by estimating HDL genetic correlations[45] (multiple-testing-adjusted threshold: $P<5.32\times10^{-3}$). Consistent with the identified heterogeneous genetic architecture (see above), polygenic association patterns were highly divergent for early-phase and late-phase vocabulary (Figure 3a). Larger early-phase expressive vocabulary size was genetically correlated with increased ADHD risk (r_g=0.23(SE=0.08)), exposure to childhood maltreatment (r_g=0.19(SE=0.07)) and, at the nominal level ($P<0.05$), childhood aggressive behaviour (r_g=0.42(SE=0.16)). The direction of the association effect reversed, however, in toddlerhood, such that lower late-phase receptive vocabulary size was genetically related to increased childhood maltreatment exposure (r_g=-0.33(SE=0.08)). In addition, genetic links with cognition-related later-life outcomes became detectable in toddlerhood only (Figure 3a), despite comparable power estimates across infant and toddler measures (eTable 9). Both, larger late-phase expressive and receptive vocabulary size were genetically correlated with higher intelligence across the lifespan (late-phase expressive vocabulary: r_g=0.32(SE=0.08); late-phase receptive vocabulary: r_g=0.36(SE=0.12)) and higher educational attainment (late-phase expressive vocabulary: r_g=0.26(SE=0.05); late-phase receptive vocabulary: r_g=0.37(SE=0.06)).

To integrate genetic covariance patterns into a structural model, we adopted a genomic SEM[57] approach that was informed by exploratory factor analyses (eTable 10). Specifically, we studied all three early-life vocabulary measures and genetically correlated later-life traits (HDL-r_g-$P<5.32\times10^{-3}$: intelligence, educational attainment, exposure to childhood maltreatment and ADHD). The best-fitting model was a correlated 3-factor model (Figure 3b, eTable 11), implicating two uncorrelated genomic dimensionalities. For comparison with estimates presented above, we report, here, unstandardised factor loadings, representing the direction and strength of association between a trait and model-implied latent genetic factor: The first genetic factor ($F1_g$) reflected shared genetic influences between early-phase

(λ=0.25(SE=0.09)) and late-phase expressive vocabulary (λ=0.31(SE=0.10)). This factor was modestly correlated (r_g=0.25(SE=0.09)) with a second genetic factor ($F2_g$) accounting for the majority of genetic variance in educational attainment (λ=0.33(SE=0.01)) and intelligence (λ=0.32(SE=0.01)). The third genetic factor ($F3_g$), which was, largely, independent of $F1_g$ (r_g=-0.09(SE=0.09)), captured the entirety of genetic ADHD liability (λ=0.48(SE=0.04)) and, to a lesser extent, genetic influences contributing to childhood maltreatment exposure (λ=0.13(SE=0.01)). Notably, early-phase expressive (λ=0.16(SE=0.12)) and late-phase receptive (λ=-0.19(SE=0.04)) vocabulary were associated with this factor with an opposite direction of effect. While, largely, consistent with reported SNP-h^2 estimates and genetic relationships above (Figure 2, 3a), the identified model structure lacked genetic overlap between late-phase expressive and receptive vocabulary (eTable 12), underlining the limitations of the model.

To confirm the change in genetic association pattern with direct genotyping data, we studied children of the ALSPAC cohort with genetic, vocabulary and ADHD symptom information and modelled jointly the phenotypic (r_p), genetic (r_g) and residual (r_e) correlations fitting a saturated (Cholesky) structural model using GRM-SEM[58] (eMethods 10). Genetic correlations in ALSPAC (eFigure 4) had the same direction of effect as meta-analytic HDL-derived estimates (Figure 3a), underlining the robustness of our findings. However, 95%-confidence intervals (CIs) were wide and analyses have, thus, an exploratory character only. Increased ADHD symptoms were phenotypically very modestly correlated with smaller vocabulary size (r_p=-0.06(SE=0.02)), irrespective of developmental stage (eFigure 4). Thus, during infancy only, genetic correlations (r_g=0.51(SE=0.26)) may have been masked by opposite residual correlations (r_e=-0.18(SE=0.06)).

Discussion

Here, we present findings from genome-wide association meta-analyses of expressive and receptive vocabulary size across infancy and toddlerhood. Genomic covariance analyses revealed heterogeneity within the genetic architecture underlying vocabulary size across different developmental stages that matched distinct polygenic association patterns with cognitive and behaviour-related traits during later life.

Bivariate genetic correlation patterns and structural models suggested two independent genetic factors contributing to early-life vocabulary size, confirming the heterogeneous multivariate genetic architecture observed in previous reports[7,8,12,13]. Both factors may reflect aetiological differences in vocabulary acquisition stages, given limited genetic overlap between infant expressive and toddler receptive vocabulary size. Genetic influences contributing to utterances in infancy, approximated by early-phase expressive vocabulary size (15-18 months), may capture the first stages of language learning related to emerging speech, where words are usually produced in isolation[2]. During this phase of "learning to speak", children acquire phonological skills to identify phonemes and sequences from speech and store them for future production[60], but also develop oral motor[61] and speech motor skills[62]. Despite sufficient power, there was little evidence for genetic overlap between early-phase vocabulary scores and later-life cognition. Such an association only emerged for late-phase vocabulary scores during toddlerhood (24-38 months), in line with previous work[12] and suggested specificity. The genetic overlap with cognition may reflect the onset of a phase of "speaking to learn", where toddlers have mastered some language fluency and start to use word combinations and more complex grammatical structures[4,6]. Together, our findings highlight rapid changes in the genetic architecture of vocabulary acquisition across a period of only two years, even when assessed with similar psychological instruments. Shared genetic influences across both developmental phases underline the dynamic character of this process, illustrated by the genetic overlap of late-phase expressive vocabulary with both, early-phase expressive and late-phase receptive vocabulary.

The heterogeneity in genetic components contributing to vocabulary acquisition was further reflected by distinct polygenic association patterns with later-life behaviour and related proxies. Larger infant expressive but lower toddler receptive vocabulary size was genetically correlated with increased ADHD risk and/or childhood maltreatment exposure. Consistently, younger age at first walking, another early developmental milestone, has been linked to higher polygenic ADHD risk[63]. During a developmental phase of "learning to speak", potentially involving motor skills that shape children's learning environment and, in turn, behaviour and language learning[64], children with a higher genetic predisposition for ADHD may show larger rather than smaller vocabulary size. In contrast, the direction of genetic overlap between ADHD risk and toddler receptive vocabulary is consistent with the known inverse polygenic association pattern of ADHD risk with child/adolescent verbal and cognitive abilities[22]. Genetic links implicating ADHD and childhood maltreatment exposure were similar, reflecting the genetic trait overlap[27].

This work builds on a previous GWAS effort[7] by increasing the number of studied children by ~50% and adopting a multivariate analysis approach to maximise statistical power while extending the studied phenotypic vocabulary spectrum. However, the power to detect single variant contributions of small effect (e.g. 0.1%) remained low (eMethods 4), especially for receptive vocabulary. Given limited data availability, the study focussed also exclusively on European children and languages. Furthermore, to harmonise vocabulary measures across different developmental stages and instruments, all measures had to be rank-transformed. Finally, although structural models built from genetic summary statistics had an acceptable model fit, the model could not capture all aspects of the underlying multivariate genetic architecture, limiting its interpretation. Future studies may increase the sample size further and boost study power through multivariate analysis of vocabulary with genetically related aspects of language, such as grammatical abilities[8,9], preferably in genetically diverse populations.

In summary, there are at least two distinct genetic factors contributing to vocabulary size during infancy and toddlerhood that match distinct polygenic association patterns with later-life outcomes. Our

findings highlight the importance of studying genetic influences underlying early-life vocabulary acquisition to unravel the aetiological processes that shape future behaviour, health and cognition.

Acknowledgments

We are extremely grateful to all children, parents and caregivers for making this study possible. The project was embedded within the Early Genetics and Life Course Epidemiology (EAGLE) Consortium. Cohort-specific acknowledgements and funding information can be found in eMethods 11. In addition, we would like to thank all cohorts and researchers that made their summary statistics available to us. This includes the Social Science Genetic Association Consortium (SSGAC), Early Growth Genetics (EGG) Consortium, Early Genetics and Life Course Epidemiology (EAGLE) Consortium, Psychiatric Genomics Consortium (PGC) and the Danish Lundbeck Foundation Initiative for Integrative Psychiatric Research (iPSYCH). EVe and BSTP had full access to all summary-statistic level data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

The EAGLE working group that contributed to this manuscript consists of the following members: Ole A. Andreassen[1,2,3], Meike Bartels[4], Dorret Boomsma[4], Philip S. Dale[5], Erik Ehli[6], Dietmar Fernandez-Orth[7], Mònica Guxens[7,8,9,10], Christian Hakulinen[11], Kathleen Mullan Harris[12,13], Simon Haworth[14,15], Vincent Jaddoe[16,17], Liisa Keltikangas-Järvinen, Terho Lehtimäki[18,19,20], Christel Middeldorp[21,22], Josine L. Min[14,23], Pashupati P. Mishra[18,19,20], Pål Rasmus Njølstad[24,25], Jordi Sunyer[7,8,9], Ashley E. Tate[26], Nicholas Timpson[14,23], Camiel van der Laan[4], Martine Vrijheid[7,8,9], Eero Vuoksimaa[27], Alyce Whipp[27], Eivind Ystrom[28,29], ACTION Consortium[30], BIS investigator group[31]

1. NORMENT Centre, Institute of Clinical Medicine, University of Oslo, Oslo, Norway
2. Division of Mental Health and Addiction, Oslo University Hospital, Oslo, Norway
3. KG Jebsen Centre for Neurodevelopmental disorders, University of Oslo, Oslo, Norway
4. Netherlands Twin Register, dept Biological Psychology, Vrije Universiteit Amsterdam, the Netherlands
5. Speech & Hearing Sciences Department, University of New Mexico, Albuquerque, NM, USA
6. Avera Institute for Human Genetics, Sioux Falls, SD, USA
7. ISGlobal, Barcelona, Spain

8. Pompeu Fabra University, Barcelona, Spain
9. Spanish Consortium for Research on Epidemiology and Public Health (CIBERESP), Instituto de Salud Carlos III, Madrid, Spain
10. Department of Child and Adolescent Psychiatry/Psychology, Erasmus MC, University Medical Centre, Rotterdam, The Netherlands
11. Department of Psychology and Logopedics, University of Helsinki, Finland
12. Department of Sociology, University of North Carolina at Chapel Hill, NC, USA
13. Carolina Population Center, University of North Carolina at Chapel Hill, NC, USA
14. Medical Research Council Integrative Epidemiology Unit at the University of Bristol, Bristol, UK
15. Bristol Dental School, University of Bristol, Bristol, UK
16. Department of Pediatrics, Erasmus University Medical Center, Rotterdam, The Netherlands
17. Generation R Study Group, Erasmus University Medical Center Rotterdam
18. Department of Clinical Chemistry, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
19. Finnish Cardiovascular Research Centre, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
20. Department of Clinical Chemistry, Fimlab Laboratories, Tampere, Finland
21. University of Queensland, Child health Research Centre, Brisbane, Australia
22. Child and Youth Mental Health Service, Children's Health Queensland Hospital and Health Service, Brisbane, Australia
23. Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
24. Department of Clinical Science, University of Bergen, NO-5020 Bergen, Norway
25. Children and Youth Clinic, Haukeland University Hospital, NO-5021 Bergen, Norway
26. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
27. Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Finland
28. PROMENTA Research Center, Department of Psychology, University of Oslo, Oslo, Norway
29. Department of Mental Disorders, Norwegian Institute of Public Health, Oslo, Norway

30. eMethods 12
31. eMethods 13

EVe, EEi, SEF and BSTP were funded by the Max Planck Society. TSA was supported by the Novo Nordisk Foundation Grant NNF18OC0052457. ATM is supported by the National Health and Medical Research Council. CS was supported by an Australian Government NHMRC Postgraduate Research Scholarship. EH receives funding from the National Health and Medical Research Council Australia, Australian Research Council, Medical Research Future Fund, and Tour de Cure. MG is funded by a Miguel Servet II fellowship (CPII18/00018) awarded by the Spanish Institute of Health Carlos III. We acknowledge support from the Spanish Ministry of Science and Innovation and the State Research Agency through the "Centro de Excelencia Severo Ochoa 2019-2023" Program (CEX2018-000806-S), and support from the Generalitat de Catalunya through the CERCA Program. SH receives support from the UK National Institute for Health Research through the academic clinical fellowship scheme. KR is supported by a Sir Henry Wellcome Postdoctoral Fellowship. CYS and JLM are supported by the UK Medical Research Council (MRC) Integrative Epidemiology Unit at the University of Bristol (MC_UU_00011/5). OAA is supported by KG Jebsen Stiftelsen, Research Council of Norway (#223273, 273291, 324252). EY receives support from the Research Council of Norway (grant numbers 262177; 288083). PRN was supported by grants from the European Research Council (AdG SELECTionPREDISPOSED #293574), the Bergen Research Foundation ("Utilizing the Mother and Child Cohort and the Medical Birth Registry for Better Health"), Stiftelsen Kristian Gerhard Jebsen (Translational Medical Center), the University of Bergen, the Research Council of Norway (FRIPRO grant #240413), the Western Norway Regional Health Authority (Strategic Fund "Personalized Medicine for Children and Adults"), the Novo Nordisk Foundation (grant #54741), and the Norwegian Diabetes Association. CC receives support from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 848158 (EarlyCause Project).

OAA is a consultant to HealthLytix. All other authors declare no conflict of interest.

Data sharing statement

Derived single-trait (stage I) and multi-trait (stage II) vocabulary summary statistics will be made available upon publication of the manuscript via a data repository.

References

1. Kennison SM. *Introduction to Language Development*. 1st ed. SAGA; 2014.
2. Clark EV. *First Language Acquisition*. Cambridge University Press; 2016.
3. Bergelson E, Swingley D. At 6-9 months, human infants know the meanings of many common nouns. *Proc Natl Acad Sci USA*. 2012;109(9):3253-3258. doi:10.1073/pnas.1113380109
4. Fenson L, Dale PS, Reznick JS, Bates E, Thal DJ, Pethick SJ. Variability in early communicative development. *Monogr Soc Res Child Dev*. 1994;59(5):1-173; discussion 174-85.
5. Goldfield BA, Reznick JS. Early lexical acquisition: rate, content, and the vocabulary spurt. *J Child Lang*. 1990;17(1):171-183.
6. Hoff E. *Language Development*. Cengage Learning; 2013.
7. St Pourcain B, Cents RA, Whitehouse AJ, et al. Common variation near ROBO2 is associated with expressive vocabulary in infancy. *Nature communications*. 2014;5:4831. doi:10.1038/ncomms5831
8. Dionne G, Dale PS, Boivin M, Plomin R. Genetic Evidence for Bidirectional Effects of Early Lexical and Grammatical Development. *Child Development*. 2003;74(2):394-412. doi:10.1111/1467-8624.7402005
9. Dale PS, Dionne G, Eley TC, Plomin R. Lexical and grammatical development: a behavioural genetic perspective. *Journal of Child Language*. 2000;27(03):619-642. doi:null
10. Hayiou-Thomas ME, Dale PS, Plomin R. The etiology of variation in language skills changes with development: a longitudinal twin study of language from 2 to 12 years. *Dev Sci*. 2012;15(2):233-249. doi:10.1111/j.1467-7687.2011.01119.x
11. Reznick JS, Corley R, Robinson J. A longitudinal twin study of intelligence in the second year. *Monogr Soc Res Child Dev*. 1997;62(1):i-vi, 1-154; discussion 155-60.
12. Verhoef E, Shapland CY, Fisher SE, Dale PS, Pourcain BS. The developmental genetic architecture of vocabulary skills during the first three years of life: Capturing emerging associations with later-life reading and cognition. *PLOS Genetics*. 2021;17(2):e1009144. doi:10.1371/journal.pgen.1009144

13. Harlaar N, Hayiou-Thomas ME, Dale PS, Plomin R. Why Do Preschool Language Abilities Correlate With Later Reading? A Twin Study. *J Speech Lang Hear Res*. 2008;51(3):688-705. doi:10.1044/1092-4388(2008/049)
14. Verhoef E, Shapland CY, Fisher SE, Dale PS, Pourcain BS. The developmental origins of genetic factors influencing language and literacy: Associations with early-childhood vocabulary. *Journal of Child Psychology and Psychiatry*. Published online September 14, 2020. doi:10.1111/jcpp.13327
15. Geurts HM, Embrechts M. Language profiles in ASD, SLI, and ADHD. *J Autism Dev Disord*. 2008;38(10):1931-1943. doi:10.1007/s10803-008-0587-1
16. Helland WA, Posserud MB, Helland T, Heimann M, Lundervold AJ. Language Impairments in Children With ADHD and in Children With Reading Disorder. *Journal of Attention Disorders*. Published online October 16, 2012:1087054712461530. doi:10.1177/1087054712461530
17. Germanò E, Gagliano A, Curatolo P. Comorbidity of ADHD and dyslexia. *Dev Neuropsychol*. 2010;35(5):475-493. doi:10.1080/87565641.2010.494748
18. Peyre H, Galera C, van der Waerden J, et al. Relationship between early language skills and the development of inattention/hyperactivity symptoms during the preschool period: Results of the EDEN mother-child cohort. *BMC Psychiatry*. 2016;16. doi:10.1186/s12888-016-1091-3
19. Martin NC, Levy F, Pieka J, Hay DA. A Genetic Study of Attention Deficit Hyperactivity Disorder, Conduct Disorder, Oppositional Defiant Disorder and Reading Disability: Aetiological overlaps and implications. *International Journal of Disability, Development and Education*. 2006;53(1):21-34. doi:10.1080/10349120500509992
20. Willcutt EG, Pennington BF, DeFries JC. Twin study of the etiology of comorbidity between reading disability and attention-deficit/hyperactivity disorder. *Am J Med Genet*. 2000;96(3):293-301. doi:10.1002/1096-8628(20000612)96:3<293::AID-AJMG12>3.0.CO;2-C

21. Willcutt EG, Pennington BF, Olson RK, DeFries JC. Understanding comorbidity: A twin study of reading disability and attention-deficit/hyperactivity disorder. *Am J Med Genet*. 2007;144B(6):709-714. doi:10.1002/ajmg.b.30310
22. Verhoef E, Demontis D, Burgess S, et al. Disentangling polygenic associations between attention-deficit/hyperactivity disorder, educational attainment, literacy and language. *Translational Psychiatry*. 2019;9(1):35. doi:10.1038/s41398-018-0324-2
23. Tager-Flusberg H, Paul R, Lord C, Volkmar FR, Klin A, Cohen D. Handbook of autism and pervasive developmental disorders. Published online 2005.
24. Ozonoff S, South M, Miller JN. DSM-IV-Defined Asperger Syndrome: Cognitive, Behavioral and Early History Differentiation from High-Functioning Autism. *Autism*. 2000;4(1):29-46. doi:10.1177/1362361300041003
25. Warrier V, Baron-Cohen S. Childhood trauma, life-time self-harm, and suicidal behaviour and ideation are associated with polygenic scores for autism. *Mol Psychiatry*. 2021;26(5):1670-1684. doi:10.1038/s41380-019-0550-x
26. Sanderud K, Murphy S, Elklit A. Child maltreatment and ADHD symptoms in a sample of young adults. *Eur J Psychotraumatol*. 2016;7:32061. doi:10.3402/ejpt.v7.32061
27. Warrier V, Kwong ASF, Luo M, et al. Gene-environment correlations and causal effects of childhood maltreatment on physical and mental health: a genetically informed approach. *Lancet Psychiatry*. 2021;8(5):373-386. doi:10.1016/S2215-0366(20)30569-1
28. Harris SR. Measuring head circumference: Update on infant microcephaly. *Can Fam Physician*. 2015;61(8):680-684.
29. Maunu J, Parkkola R, Rikalainen H, Lehtonen L, Haataja L, Lapinleimu H. Brain and Ventricles in Very Low Birth Weight Infants at Term: A Comparison Among Head Circumference, Ultrasound, and Magnetic Resonance Imaging. *Pediatrics*. 2009;123(2):617-626. doi:10.1542/peds.2007-3264

30. Bartholomeusz HH, Courchesne E, Karns CM. Relationship Between Head Circumference and Brain Volume in Healthy Normal Toddlers, Children, and Adults. *Neuropediatrics*. 2002;33(05):239-241. doi:10.1055/s-2002-36735
31. Middeldorp CM, Felix JF, Mahajan A, et al. The Early Growth Genetics (EGG) and EArly Genetics and Lifecourse Epidemiology (EAGLE) consortia: design, results and future prospects. *Eur J Epidemiol*. 2019;34(3):279-300. doi:10.1007/s10654-019-00502-9
32. Fenson L, Pethick S, Renda C, Cox JL, Dale PS, Reznick JS. Short-form versions of the MacArthur Communicative Development Inventories. *Applied Psycholinguistics*. 2000;21(01):95-116. doi:null
33. Zink I, Lejaegere M. N-CDI's: korte vormen, Aanpassing en hernormering van de MacArthur Short Form Vocabulary Checklist van Fenson et al.
34. Reznick JS, Goldsmith L. A multiple form word production checklist for assessing early language. *Journal of Child Language*. 1989;16(01):91-100. doi:10.1017/S0305000900013453
35. Fenson L, Dale P, Reznick JS, et al. *User's Guide and Technical Manual for the MacArthur Communicative Development Inventories*. Singular Publishing; 1993.
36. Bleses D, Vach W, Slott M, et al. The Danish Communicative Developmental Inventories: validity and main developmental trends. *J Child Lang*. 2008;35(3):651-669. doi:10.1017/S0305000907008574
37. Rescorla Leslie. The Language Development Survey. *Journal of Speech and Hearing Disorders*. 1989;54(4):587-599. doi:10.1044/jshd.5404.587
38. Marees AT, Kluiver H de, Stringer S, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*. 2018;27(2):e1608. doi:10.1002/mpr.1608
39. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016;48(10):1279-1283. doi:10.1038/ng.3643

40. Winkler TW, Day FR, Croteau-Chonka DC, et al. Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc*. 2014;9(5):1192-1212. doi:10.1038/nprot.2014.071
41. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 2010;26(17):2190-2191. doi:10.1093/bioinformatics/btq340
42. Turley P, Walters RK, Maghzian O, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nature Genetics*. 2018;50(2):229-237. doi:10.1038/s41588-017-0009-4
43. Nyholt DR. A Simple Correction for Multiple Testing for Single-Nucleotide Polymorphisms in Linkage Disequilibrium with Each Other. *Am J Hum Genet*. 2004;74(4):765-769.
44. Watanabe K, Taskesen E, Bochoven A van, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nature Communications*. 2017;8(1):1826. doi:10.1038/s41467-017-01261-5
45. Ning Z, Pawitan Y, Shen X. High-definition likelihood inference of genetic correlations across human complex traits. *Nat Genet*. 2020;52(8):859-864. doi:10.1038/s41588-020-0653-y
46. Savage JE, Jansen PR, Stringer S, et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature Genetics*. 2018;50(7):912-919. doi:10.1038/s41588-018-0152-6
47. Lee JJ, Wedow R, Okbay A, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*. 2018;50(8):1112-1121. doi:10.1038/s41588-018-0147-3
48. Taal HR, Pourcain BS, Thiering E, et al. Common variants at 12q15 and 12q24 are associated with infant head circumference. *Nat Genet*. 2012;44(5):532-538. doi:10.1038/ng.2238
49. Haworth S, Shapland CY, Hayward C, et al. Low-frequency variation in TP53 has large effects on head circumference and intracranial volume. *Nature Communications*. 2019;10(1):357. doi:10.1038/s41467-018-07863-x

50. Ip HF, van der Laan CM, Krapohl EML, et al. Genetic association study of childhood aggression across raters, instruments, and age. *Transl Psychiatry*. 2021;11(1):1-9. doi:10.1038/s41398-021-01480-x

51. Jami ES, Hammerschlag AR, Ip HF, et al. Genome-wide association meta-analysis of childhood and adolescent internalising symptoms. Published online July 31, 2021:2020.09.11.20175026. Accessed August 12, 2021. https://www.medrxiv.org/content/10.1101/2020.09.11.20175026v2

52. Dalvie S, Maihofer AX, Coleman JRI, et al. Genomic influences on self-reported childhood maltreatment. *Transl Psychiatry*. 2020;10(1):1-12. doi:10.1038/s41398-020-0706-0

53. Demontis D, Walters RK, Martin J, et al. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature Genetics*. Published online November 26, 2018:1. doi:10.1038/s41588-018-0269-7

54. Grove J, Ripke S, Als TD, et al. Identification of common genetic risk variants for autism spectrum disorder. *Nature Genetics*. 2019;51(3):431. doi:10.1038/s41588-019-0344-8

55. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)*. 2005;95(3):221-227. doi:10.1038/sj.hdy.6800717

56. Walters R, Churchhouse C, Neale BM. Insights from estimates of SNP-heritability for >2,000 traits and disorders in UK Biobank. Published September 20, 2017. http://www.nealelab.is/blog/2017/9/20/insights-from-estimates-of-snp-heritability-for-2000-traits-and-disorders-in-uk-biobank#footnote6

57. Grotzinger AD, Rhemtulla M, de Vlaming R, et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature Human Behaviour*. 2019;3(5):513-525. doi:10.1038/s41562-019-0566-x

58. St Pourcain B, Eaves LJ, Ring SM, et al. Developmental changes within the genetic architecture of social communication behaviour: A multivariate study of genetic variance in unrelated individuals. *Biological Psychiatry*. 2017;83:598-606. doi:10.1016/j.biopsych.2017.09.020

59. Leeuw CA de, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Computational Biology*. 2015;11(4):e1004219. doi:10.1371/journal.pcbi.1004219
60. Curtin S, Archer SL. Speech perception. In: Bavin EL, Naigles LR, eds. *The Cambridge Handbook of Child Language*. 2nd ed. Cambridge Handbooks in Language and Linguistics. Cambridge University Press; 2015:137-158. doi:10.1017/CBO9781316095829.007
61. Alcock K. The development of oral motor control and language. *Down Syndrome Research and Practice*. 2006;11(1):1-8. doi:10.3104/reports.310
62. Smith A, Goffman L, Stark RE. Speech Motor Development. *Semin Speech Lang*. 1995;16(2):87-99. doi:10.1055/s-2008-1064112
63. Hannigan LJ, Askeland RB, Ask H, et al. Developmental milestones in early childhood and genetic liability to neurodevelopmental disorders. *Psychological Medicine*. Published online September 21, 2021:1-9. doi:10.1017/S0033291721003330
64. Libertus K, Violi DA. Sit to Talk: Relation between Motor Skills and Language Development in Infancy. *Front Psychol*. 2016;7. doi:10.3389/fpsyg.2016.00475

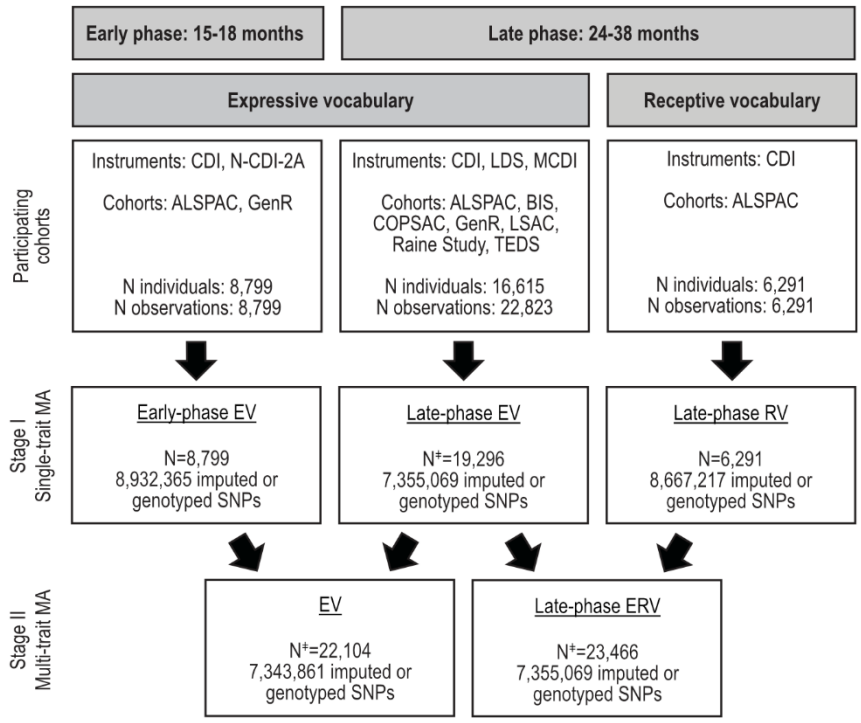

Figure 1: Meta-analysis study design

Vocabulary scores were assessed between 15-38 months of age and divided into an early phase (15-18 months) and late phase (24-38 months) of language acquisition allowing for age-specific genetic influences. Scores for receptive vocabulary were included in the late-phase only. In stage I, three single-trait meta-analyses were conducted: early-phase expressive vocabulary, late-phase expressive vocabulary and late-phase receptive vocabulary. In stage II, multi-trait genome-wide analyses were performed across early-phase and late-phase expressive vocabulary, as well as across late-phase expressive and receptive vocabulary to increase statistical power.

‡ Estimated sample size based on the increase in mean χ^2 statistic using multi-trait analysis of genome-wide association.

Abbreviations: ALSPAC, Avon Longitudinal Study of Parents and Children; BIS, Barwon Infant Study; CDI, Communicative Development Inventory; COPSAC, Copenhagen Prospective Studies on Asthma in Childhood; EV, expressive vocabulary; ERV, expressive and receptive vocabulary; GenR, Generation R Study; LDS; Language Development Survey; LSAC, Longitudinal Study of Australian Children; MA, meta-analysis; RV, receptive vocabulary; TEDS, Twins Early Development Study

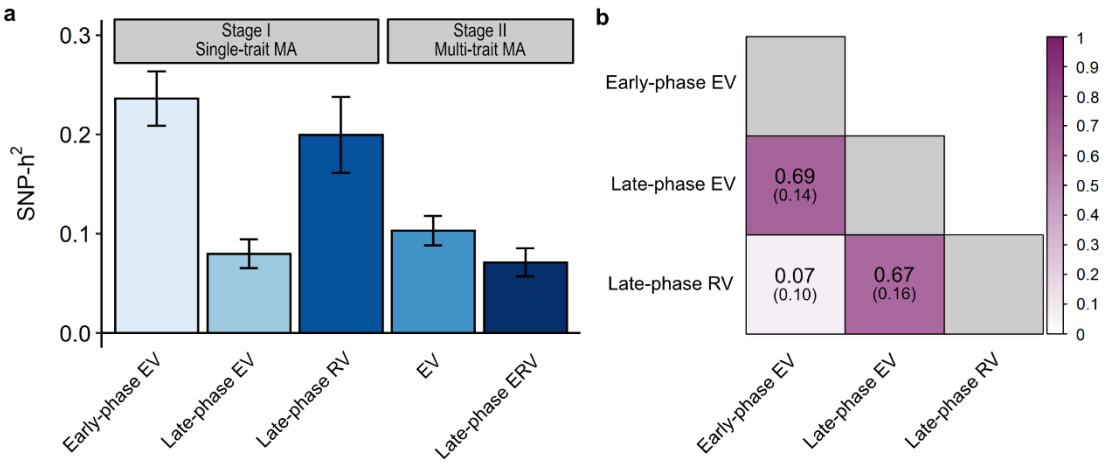

Figure 2: SNP-heritability and genetic correlations of vocabulary traits

(a) SNP-heritability estimates for single- and multi-trait vocabulary summary statistics were estimated with High-Definition Likelihood[45] software. Error bars represent standard errors. **(b)** Genetic correlations (r_g) among single-trait vocabulary summary statistics were estimated with High-Definition Likelihood[45] software. Corresponding standard errors are shown in brackets.

Abbreviations: EV, expressive vocabulary; ERV; expressive and receptive vocabulary; MA, meta-analyses; RV, receptive vocabulary; SNP-h^2, Single-Nucleotide Polymorphism heritability

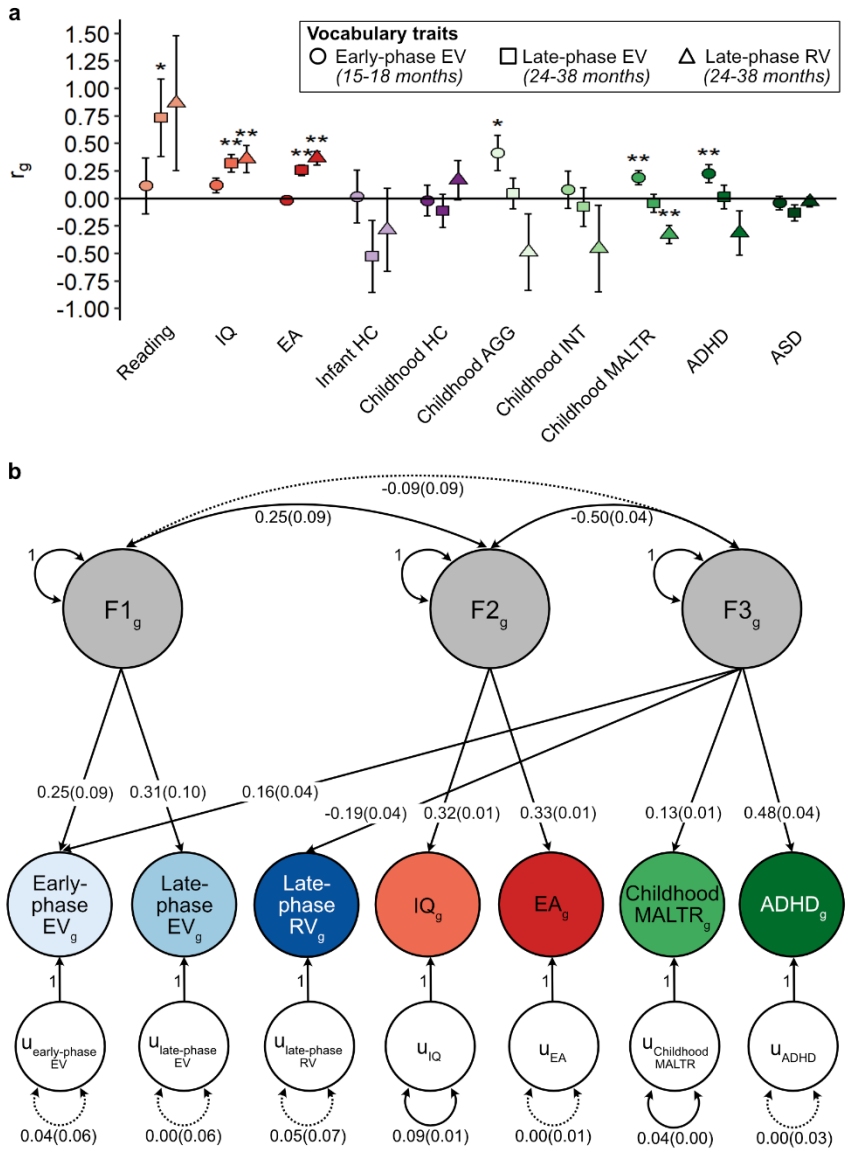

Figure 3: Genetic relationships of vocabulary with several later-life cognitive, health and behavioural outcomes

(a) Genetic correlations (r_g) were estimated using summary statistics and High-Definition Likelihood (HDL)[45]. Bars represent standard errors. ** multiple-testing adjusted $P<5.32\times10^{-3}$; * $P<0.05$. **(b)** Three-factor model fitted to genetic covariance patterns of early-life vocabulary measures and genetically correlated later-life outcomes (identified with HDL, $P(r_g)<5.32\times10^{-3}$) using Genomic SEM[57]. Solid and dashed arrow lines represent factor loadings with $P<0.05$ and $P\geq0.05$, respectively. Unstandardised factor loadings are shown, with standard error in parenthesis. Model fit characteristics are provided in eTable 11.

Abbreviations: ADHD, Attention-Deficit/Hyperactivity Disorder; AGG, aggression; ASD, Autism Spectrum Disorder; EA, educational attainment; EV, expressive vocabulary; HC, head circumference; INT, internalising symptoms; IQ, general intelligence; M, months; MALTR, maltreatment; RV, receptive vocabulary; SDQ, Strengths and Difficulties Questionnaire; Y, years