Genome-wide association study on longitudinal and cross-sectional traits of child health and development in a Japanese population

JECS Genetics Research Group

22.04.2026

Abstract

Understanding the influence of both genetics and environment on human health, especially early in life, is essential for shaping long-term health. Here, we utilize a nationwide prospective birth cohort, the Japan Environment and Children's Study (JECS), to conduct a large-scale population-based genetic study using biannual questionnaire surveys and biological and physical measurements collected from both parents and their children up to the age of 4, since the participant mothers were pregnant. Analyses of genome-wide genotyping for 80,639 child participants with parental consent and sufficient DNA from cord blood samples represent the genetic diversity of the general population in Japan. Systematic genome-wide association studies of 1,255 child health and developmental traits (including, e.g., food allergy, anthropometry measurements or ASQ-3 developmental screening) and parental environmental exposure traits (including, e.g., mercury or PFAS exposure) identify 6,073 common genomic loci (2,332 of which passed the phenome-wide significance threshold $P = 4.0 \times 10^{-11}$), of which 18-20% of loci represent novel associations not previously reported, including loci with potential therapeutic relevance. Additional longitudinal GWAS of BMI, using Gaussian process regression, identified 29 novel dynamic genetic associations throughout child development. These associations can also be used to predict an individual's growth trajectory based on their genetic background. In addition, genetic correlation analysis suggested potential associations between maternal environmental exposures during pregnancy and child traits after birth. Together with studies of genetic risk factors and environmental exposures, across time and multiple outcomes, these demonstrate the uniqueness and value of the JECS data.

Introduction

A birth cohort study is the ideal genetic research platform for studying genetic variations in human health and development over the life course[1]. While the number of cases is not as large as in a disease case-control study, it is possible to study a wide range of phenotypes, especially in infant and childhood participants, or environmental exposures as intermediate phenotypes that may account for missing heritability in complex diseases[2]. This platform, particularly once parental genotypes are available, will enable partitioning of genetic effects on birth phenotypes into maternal and fetal components[3]. Furthermore, a biobank alliance between a birth cohort and a case-control study allows us to project the adult disease risks in early life outcomes for early intervention or a potential route to new therapies and diagnostics[3–5].

The Japan Environment and Children's Study (JECS) is a large-scale birth cohort study funded by the Ministry of the Environment, Japan, to evaluate the effects of environmental chemicals on children's health and development[6]. Over 100,000 pregnant women were enrolled at 15 geographically different Regional Centres across Japan, representing the comprehensive genetic diversity of the Japanese population. Detailed data from questionnaires, biological and physical measurements have been collected from both parents and their children since the participant mothers were pregnant, with additional surveys every six months being conducted on average 80% of the child participants until the age of 4. Furthermore, JECS is actively engaged in the collection of samples and data pertaining to the pervasive exposure to environmental pollutants, including heavy metals[7,8] and per- and polyfluoroalkyl substances (PFAS)[9,10].

Genome-wide association studies (GWAS) are powerful tools for uncovering genetic factors behind complex traits and diseases[11], enabling better diagnosis[12], prevention[13], and treatment[14]. As of 2024, over 691,532 associations across 36,643 traits have been cataloged[11]. However, the genetic basis of many traits remains unclear, with most successful findings limited to European populations[15] and a few traits like adult height[16] or diabetes[17]. We here present the results of GWAS investigating a wide range of traits related to child health and development, as well as parental environmental exposures. Subsequently, cross-trait LD score regression (LDSC) analysis was conducted for a cross-sectional analysis of child phenotypes, such as various allergic traits. The analysis suggests pleiotropic associations in similar traits and highlights an East Asian specific variant of potential therapeutic relevance. Furthermore, in a novel and distinctive attempt, we elucidate age-specific genetic associations with the anthropometric measurement BMI using a Gaussian process regression as an example of a longitudinal GWAS. This analysis uncovers novel genetic associations in a Japanese population and creates personalized growth curves, which could significantly improve the assessment and management of growth disorders and obesity. Lastly, cross-trait LDSC analysis revealed genetic correlations for some trait pairs (e.g. maternal IgE and childhood allergy) between maternal exposures and offspring traits. The findings of these analyses, including GWAS summary statistics for 1,255 traits, are available via the JECS GWAS summary statistics request form (https://forms.office.com/r/qipvY8nNTK).

The JECS genetic study of 100K children born in Japan

JECS is a nationwide government-funded birth cohort study in Japan. The study design of JECS has been published elsewhere[6,18]. Briefly, 103,099 pregnancies were enrolled at 15 Regional Centres across Japan through Co-operating Health Care Providers and/or local government offices between 2011 and 2014 (**Fig. 1a**, **Methods**). The previous publication reported that the children in JECS covered approximately 45% of the total live births in the study area[18]. Of these, 80,639 child participants with parental consent and sufficient DNA from the cord blood samples have been genotyped using the Illumina Asian Screening Array (**Methods**). The result yielded 617,557 genome-wide SNP variants after fixing strand orientations and matching alleles to the reference sequence (GRCh38) (**Methods**). A total of 109 samples with low DNA quality, as defined by a sample call rate less than 0.98 (**Fig. 1b**), were excluded from further analysis.

Subsequently, chromosomal aberrations were identified from the Illumina B allele frequency of signal intensity data (**Methods**) to yield a total of 335 participants exhibiting one or more autosomes that are uniparental disomy or trisomy (**Fig. 1b,c**). Additionally, some participants exhibited loss of the X/Y chromosome, as well as duplication of the X chromosome (**Fig. 1c**). Following the selection of normal karyotypes, we examined discrepancies between the inferred sex from the genotype and that from the questionnaire, resulting in the removal of 85 samples (**Fig. 1b**). Given that we collected 100K samples from a general Japanese population, it is plausible that some of the sex-inconsistent samples may represent disorders of sexual development[19] (DSDs) rather than sample swapping. Note that the potential sample swapping was not enriched at certain Regional Centres (the minimum Fisher exact test *P*-value was 0.60 at a Regional Centre after Bonferroni correction).

Following the removal of participants containing chromosomal aberrations, MoChA[20,21] was additionally used to identify participants with mosaic chromosomal alterations (mCA). This analysis revealed a prevalence of 1.9% for autosomal mCA and 1.1% and 1.6% for mosaic loss of the X and Y chromosomes, respectively (**Methods**; **Extended Data Fig. 1a-c**). For comparison, the corresponding frequencies in an adult population obtained from the BioBank Japan (BBJ), combining the first and second BBJ cohorts, were, on average, 11.3%, 37.0% and 25.3% (**Methods**). Note that we did not remove any participants carrying mCA because the mCA events were relatively small and the cell fractions were low (**Extended Data Fig. 1b**).

We then employed KING[22] to examine the relatedness of our samples (**Methods**). The analysis revealed a high degree of relatedness in the data, with 4,131 pairs of half- and full siblings, and 9,628 pairs of cousins (**Fig. 1d**). We also found the number of related samples tends to be lower in urban areas and higher in rural areas of Japan (**Extended Data Fig. 1d, e**). This discrepancy is likely due to the effective sample size in each area. Additionally, two samples were identified as genetically identical (**Fig. 1d**), which is unlikely to occur due to the sampling scheme of cord blood (**Methods**). Therefore, these samples were removed due to a potential sample swap (**Fig. 1b**).

Finally, a population stratification analysis was performed using principal component analysis[23] and ADMIXTURE[24] (**Methods**). Due to the recruitment criteria (at least one parent must be literate in Japanese), some participants may not be of Japanese origin. In fact, 340 non-East Asian children were found to have African, South American, South Asian and/or European ancestries in varying degrees (**Fig. 1e**, **Extended Data Fig. 1f**). Following the removal of the

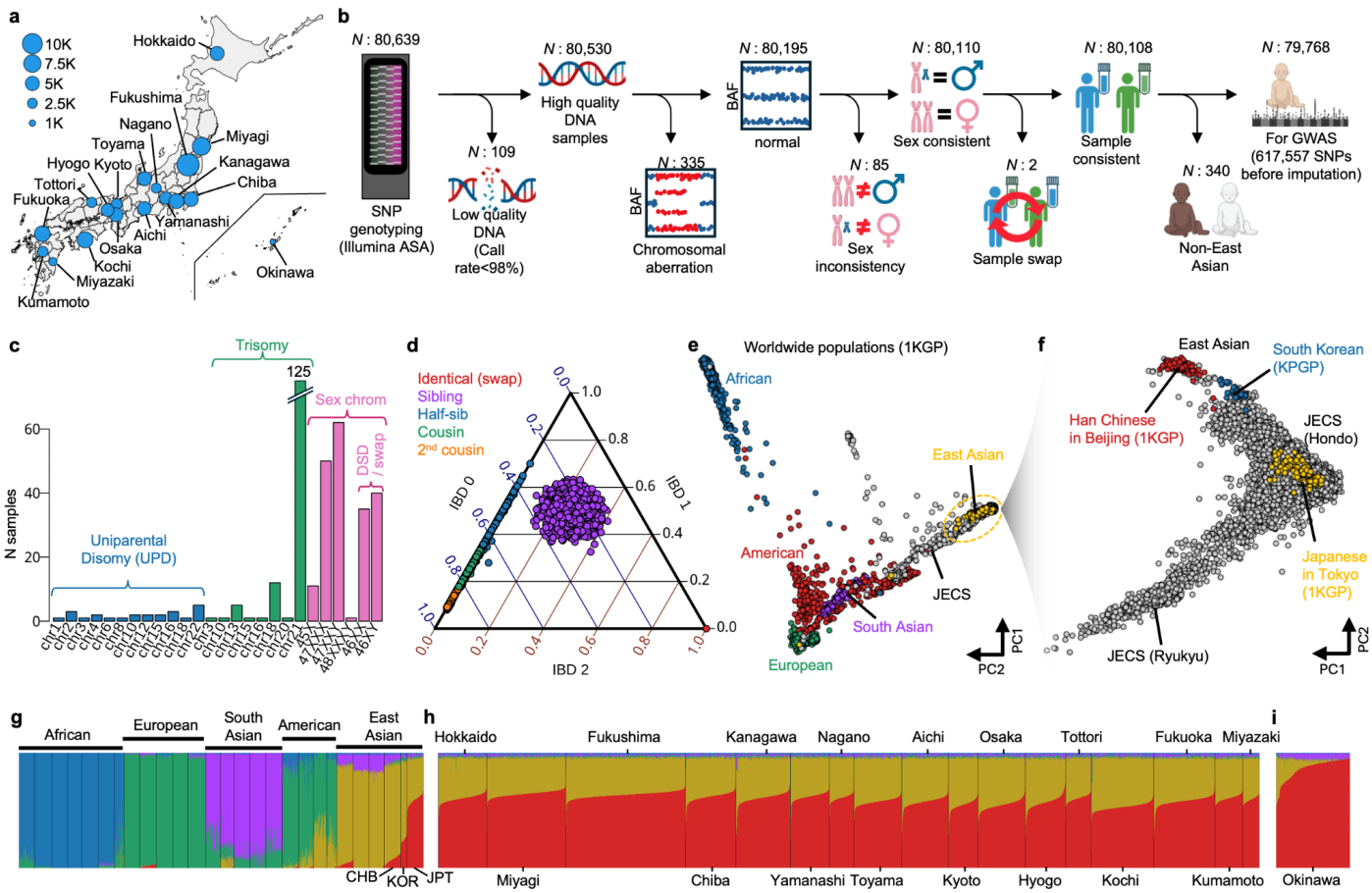


Figure 1. Overview of the JECS genetic study. **a.** Geographic distribution and demographic statistics of the 80,639 participants involved in the JECS genetic study. Dot size on the map indicates the number of participants in each region. **b.** Sample quality control flow for GWAS. Samples were selected by call rate (CR), B allele frequency (BAF) distributions, inferred sex based on genotype data and genotype principal components (**Methods**). **c.** Number of chromosomal aberrations for each chromosome. **d.** Ternary plot shows proportions of identity-by-descent (IBD) segments for each pair of participating children. **e.** The first two genotype principal components of JECS samples with worldwide populations obtained from the 1000 Genomes Project (1KGP) and Korean Personal Genome Project (KPGP). Dots are colored by different populations (gray: JECS; blue: African; green: European; red: American; purple: South Asian; yellow: East Asian). **f.** The first two principal components of East Asian populations (gray: JECS; red: Han Chinese in Beijing; blue: South Korean; yellow: Japanese in Tokyo). **g-i.** The stacked bar plots show the result of ADMIXTURE analysis for worldwide populations in 1KGP with KPGP (*N*=2,587) and JECS participants including non-East Asian populations (*N*=80,108). The result is split into three panels, showing the populations from 1KGP with KPGP (panel g), participants enrolled in mainland Japan (Hondo; *N*=79,353; panel h) and those enrolled in Okinawa Regional Centre (Ryukyu Islands; *N*=755; panel i). The y-scales differ between the panels: the panel g and i are 10 times wider than the panel h, to clearly show the ancestry differences. JPT: Japanese in Tokyo, Japan; CHB: Han Chinese in Beijing, China; and KOR: South Korean (KPGP).

aforementioned samples, the first two principal components demonstrated the existence of genetic diversity from the mainland Japanese (Hondo) to the East Asian continental population (including Chinese individuals from Beijing and South Koreans) and the Japanese Ryukyu ancestry (**Fig. 1f**), which can be seen to reflect the dual structure of the Japanese population previously reported[25,26]. Notably, compared to the other East Asian populations (**Fig. 1g**), the JECS participants exhibited a higher proportion of the Japanese-specific ancestry (red component in **Fig. 1h,i**), which was particularly high among participants enrolled at the Okinawa Regional Centre (**Fig. 1i**). However, this increase is gradual

rather than discrete; the ancestry proportions shift continuously along the geographical gradient from Hondo to Okinawa without forming a clearly distinct cluster (**Fig. 1f and i**). In other words, the Japanese population structure appears clinal, reflecting continuous genetic variation rather than sharp categorical boundaries between Ryukyu and Hondo. In fact, in previous large-scale GWAS conducted in BBJ, Japanese participants have commonly been analyzed as a single ancestry group, while subtle population structure was controlled by excluding non–East Asian outliers and including genetic principal components as covariates[28–31].

A total of 79,768 samples have been determined to meet the quality control standards (**Fig. 1b**). It should be noted that JECS collected the cord blood sample during delivery, and only one sample was taken, even in cases of multiple births (**Methods**). While there were 502 twins and 4 triplets in the genotyped samples, one child was genotyped while the other pair (or the other two children in a triplet) was not. In contrast, phenotypes were observed for each child independently, irrespective of whether they were twins or triplets, and the sample size was not negligible (1,014 twin and triplet children). Therefore, we assigned the same genotype data of a twin to the co-twin(s) when a multiple birth was recorded and the reported sex of the child matched the genotyped sample. This resulted in 80,278 children, including those from multiple births whose sexes are consistent with their genotypes, for downstream analysis. We acknowledge that this approach may introduce some degree of non-independence due to shared genetic and environmental factors among siblings, which was formally evaluated in the sensitivity analysis section below.

GWAS identify common non-coding genetic variants as major genetic determinants of child health and development

Large scale GWAS of more than 80K children were then conducted on 1,046 child traits selected from questionnaires, biological and physical measurements for children up to the age of four (see **Supplementary Table 1**). Some traits were not available for GWAS in their original form; therefore, transformed or combined traits were generated from the original phenotype data (**Methods**). To increase the power and reduce the type I error in GWAS, we filtered out traits with a case-control ratio of less than 1% for binary traits and a complexity (an index of how many identical values/answers in the trait) of less than 5% for quantitative traits (see **Methods** for more details). The test statistics are often inflated when samples from related populations are analyzed in a GWAS. To address this issue, we employed REGENIE[27] to account for the relatedness between samples (**Methods**).

In addition, GWAS was performed on an extra 209 parental environmental exposure traits (see **Supplementary Table 1**). Here the parental (directly observed) traits were regressed onto the child (proxy) genotypes to assess genetic associations, proposed as GWAS-by-proxy[28], a viable approach given the observed correlation between child and parent genotypes (**Methods**). Although this approach was initially used in a case-control study[29], this framework can be extended beyond case‑control traits to quantitative traits of various environmental exposures, because the core idea is that an offspring genotype provides noisy information on the parental genotype, regardless of the scale of the phenotype. The expected effect estimate in this design is approximately half of the direct effect that would be obtained

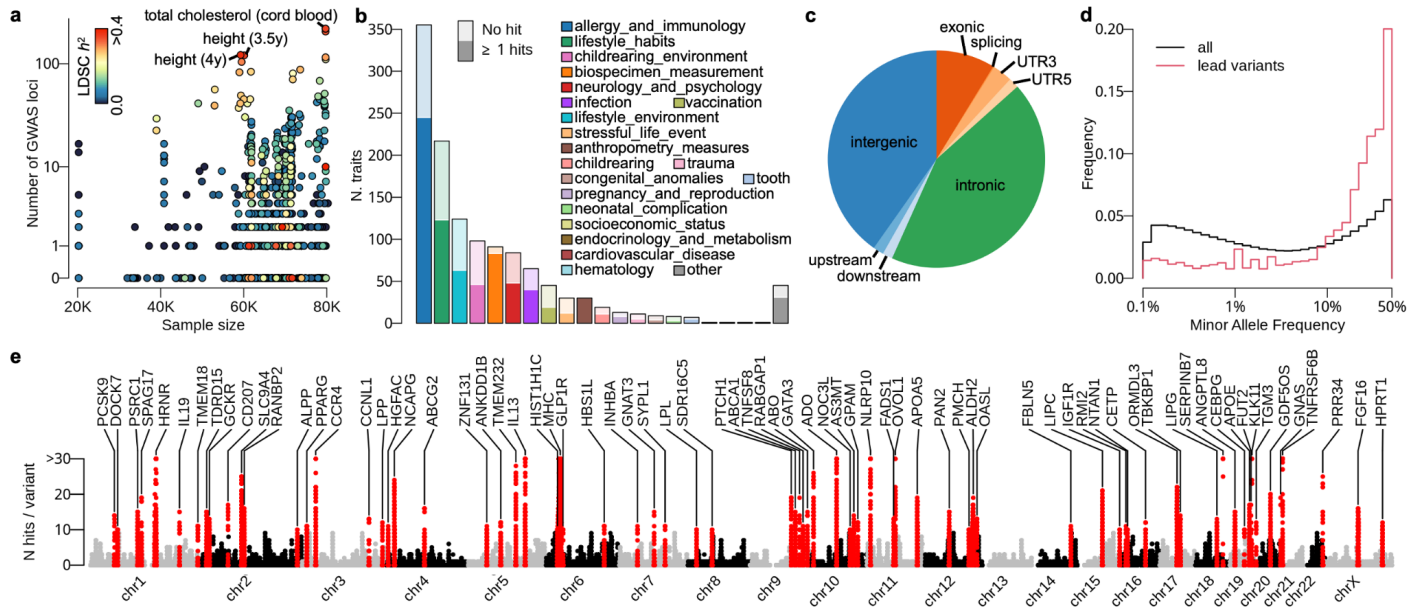


Figure 2. Characteristics of genetic loci and index variants identified by the JECS GWAS. **a.** Scatterplot shows the number of GWAS loci for each trait as a function of sample size. Each dot is colored by the additive SNP heritability estimation from the LD Score Regression (LDSC). **b.** The number of traits in different trait categories segregated by the existence of significant associations. Bars are split and darker colors show the number of traits with one or more genome-wide significant associations ($P < 5.0 \times 10^{-8}$). **c.** Piechart shows the proportion of variant locations relative to the Ensembl gene (version 113; GRCh38). **d.** Distribution of the minor allele frequency of the identified GWAS index variants against those of the entire tested variants. **e**. Summary of the genetic associations discovered through GWAS of 1,255 traits in JECS. Manhattan-like plot shows the number of GWAS hits ($P < 5.0 \times 10^{-8}$) across 1,255 traits at each variant with the nearest gene annotation. Peaks are defined by variants shared across 10 or more traits.

using parental genotypes, assuming no additional indirect effects or assortative mating. We here emphasize that our primary interest lies in locus discovery for parental traits rather than in precise effect-size estimates.

For each trait, 17,454,650 common variants with a minor allele frequency (MAF) greater than 0.1% and an imputation quality score (R^2) greater than 0.3 were tested (**Methods**). In total, 6,073 GWAS loci (including 821 secondary hits in the same region; **Methods**) were identified across 776 traits (**Supplementary Table 2; Extended Data Fig. 2a**), with the index variant of each locus exhibiting a *P*-value of less than the common genome-wide threshold ($P = 5.0 \times 10^{-8}$), of which 2,332 loci across 288 traits passed a more stringent threshold of *P*-values after Bonferroni correction for the 1,255 traits (phenome-wide significant $P = 5.0 \times 10^{-8}/1255 = 4.0 \times 10^{-11}$). We initially examined the number of GWAS loci identified for each trait as the sample size varied (**Fig. 2a**). As expected, the number of loci demonstrated an increase in conjunction with both sample size and the additive SNP heritability h^2 (**Fig. 2a**, **Extended Data Fig. 2b**; **Methods**). It should be noted that over 90% of the traits included in the study were analyzed using samples of greater than 60,000 children, and 34% of the traits showed the additive SNP heritability greater than 5% (**Supplementary Table 1**). We then examined the data to identify highly heritable trait categories and found that the majority of biospecimen and anthropometry measurements identified at least one genetic association, such as height at 4 years old (**Fig. 2b**, **Extended Data Fig. 2c**).

According to the Ensembl gene annotation (version 113; GRCh38), only 9% of GWAS index variants were found in the coding sequences (**Fig. 2c**). Of these, 468 index variants (7.7% of all) potentially altered protein sequences (**Extended Data Fig. 2d**). Consequently, over 90% of variants were identified in non-coding regions (**Fig. 2c**), indicating that the discovered variants are likely to be involved in the regulatory cascades of neighboring genes, thereby influencing phenotypes. Within the range of variants tested (MAF > 0.1%), index variants were significantly enriched toward the common end of the allele frequency spectrum (MAF>10%) relative to the background distribution (**Fig. 2d**; Kolmogorov–Smirnov test $P < 10^{-16}$), consistent with greater statistical power and imputation accuracy for common variants.

While common complex traits are highly polygenic, the number of variants is finite, which results in a single variant affecting multiple traits, a phenomenon known as pleiotropy. The landscape of pleiotropy in European and Japanese populations has been intensively studied, and these regions are recognized as being under recent selection pressure[29]. Here we also defined the degree of pleiotropy as the number of significant associations per variant (**Fig. 2e**), and replicated various known regions under selection pressure[29], including the major histocompatibility complex (MHC), ABO, ALDH2, and APOE loci[30].

Overall, our discovered loci spanned approximately 13% of the entire genome when taking the 200Kb cis-window centered on each index variant, where more than 90% of gene regulatory interactions may occur[30]. We used our imputed genotype data to calculate the LD r^2 index between the index variants of our GWAS and the entire variants reported in the GWAS catalog (**Methods**), regardless of traits (**Extended Data Fig. 2e**). We identified 1,221 out of 6,073 index variants (approximately 20% of index variants) that showed weak LD ($r^2 < 0.1$), representing potentially novel genetic loci or distinct causal variants at known loci in a Japanese population. Alternatively, some of these loci may appear 'novel' simply because the corresponding phenotypes have not yet been examined in GWAS that mostly focus on adult phenotypes. Note that this number doesn't change much (1,167 or 1,123 variants) even if we set the cis-window to 500Kb or 1Mb, respectively. We also repeated the analysis using the genotype data from the three major populations (AFR, EUR and EAS) in the 1000 Genomes Project, and identified 1,674, 1,613 and 1,112 variants that showed weak LD ($r^2 < 0.1$) in 1Mb cis-window. This suggests that using a population-matched LD index is the most conservative approach and best tags previously reported GWAS hits. We note that all the tagged GWAS hits ($r^2 > 0.1$) are also reported in **Supplementary Table 2**.

HLA imputation contributes to improved biological insights into allergy and immunological traits

The allele-level HLA analysis is particularly informative for immune-related and allergy traits, as well as for interpreting associations in the major histocompatibility complex (MHC) region. We performed HLA imputation using DEEP*HLA[36] and retrieved 2,103 classical HLA alleles and amino acid variants in both classical and non-classical HLA genes. We discovered 177 independent associations for the standard GWAS threshold, of which 106 passed the phenome-wide threshold (**Supplementary Table 3**). Of 183 traits with significant lead signals in the MHC region (from 28.5 to 33.5 Mb

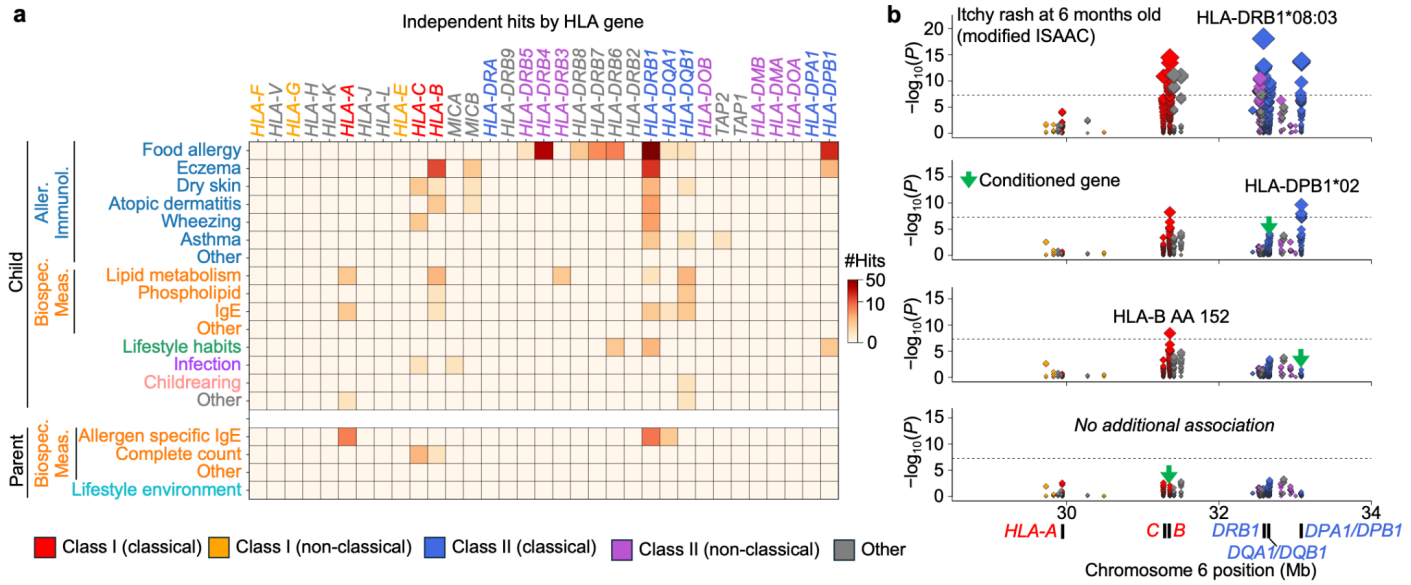


Figure 3. HLA imputation and fine-mapping in the MHC region. **a.** Heatmap showing the independent association counts of HLA genes and traits grouped by trait category and target. When the lead signals were *HLA-DRB2–9* variants in strong LD with *HLA-DRB1* variants (R2 > 0.9), *HLA-DRB1* was counted together. **b.** Four side-by-side panels showing the MHC regional association plots from the stepwise conditional regression analysis of itchy rash at 6 months old, based on the modified ISAAC questionnaire (c6m_0120001). The panels show nominal results, followed by results sequentially conditioned on *HLA-DRB1*, *HLA-DPB1*, and *HLA-B*. HLA gene labels in panel a and the variant markers in panel b are colored according to the HLA gene categories shown at the bottom.

on chromosome 6), 133 traits were linked to HLA alleles (**Supplementary Table 1**).

Associations with food allergy-related traits were predominantly driven by class II HLA genes, with *HLA-DRB1* most frequently emerging as the primary signal (**Fig. 3a**). This observation underscores the central role of class II-mediated antigen presentation to CD4+ T cells in food sensitization. In contrast, some other traits showed associations with both class I and II genes; for instance, *HLA-DRB1*, *HLA-DPB1*, and *HLA-B* were found to be independently associated with self-reported itchy rash, a proxy for eczema (**Fig. 3b**). This is in line with the established involvement of both CD4+ and CD8+ T-cell pathways in the pathogenesis of eczema or atopic dermatitis, where class I HLA-restricted effector cells contribute to cutaneous inflammation[37].

Sensitivity analyses addressing potential confounding from relatedness and population structure

Previous studies have shown that association test statistics can remain inflated in GWAS analyses, including those using REGENIE, when individuals sharing environmental influences are included. To assess the robustness of our results with respect to relatedness in our cohort, we performed GWAS in a subsample (*N*=70,751) excluding individuals related up to the third degree (**Methods**). In this analysis, we identified 5,126 GWAS loci (including secondary signals) across 737 traits at the conventional genome-wide significance threshold ($P = 5.0 \times 10^{-8}$), of which 1,946 loci across 275 traits passed the phenome-wide significance threshold ($P = 4.0 \times 10^{-11}$) (**Supplementary Table 2**). We observed high concordance with the main analysis: 82% of loci (4,193 loci) were identical (either the same index variants or proxies

in LD $r^2 > 0.1$) at the genome-wide threshold, and 97% (1,879 loci) were identical at the phenome-wide threshold (**Extended Data Fig. 3a**). The reduction in the number of detected loci was consistent with the 12% decrease in sample size (**Methods; Extended Data Fig. 3b-c**), suggesting that the observed differences are largely attributable to reduced power rather than confounding. We further compared attenuation ratios from LD score regression and observed a modest reduction (approximately 0.067 on average) after excluding related individuals (**Extended Data Fig. 3d**), indicating that shared familial effects may contribute slightly to the observed signal. However, this effect was small in magnitude, and the overall consistency of association results supports the robustness of our main findings.

Although Japan is often described as a genetically homogeneous, single-ethnicity population (**Fig. 1g–i**), lifestyle, childrearing, and social traits are strongly influenced by socio-cultural, economic, and policy factors[38,39]. Therefore, genetic associations identified in our study could, in principle, reflect genetic drift within specific ancestral subgroups. To address this possibility, we first compared genomic control lambda (λ_{GC}) and LDSC intercepts between lifestyle-related and other traits and found no significant differences (**Extended Data Fig. 3e–f**). Next, we performed GWAS along the two major axes of ancestry variation (**Extended Data Fig. 3g**; **Methods**), capturing differences between (1) Japanese and other East Asian populations (e.g., Chinese and South Korean individuals) and (2) the Japanese Hondo and Ryukyu ancestries, as represented by the first two principal components (see **Fig. 1f**). We treated these axes as quantitative traits and conducted GWAS to identify variants associated with population structure and potential genetic drift (**Extended Data Fig. 3h–i**). Finally, we conducted genetic correlation analyses using the cross-trait LD score regression[31] (**Methods**) and found no traits showing significant genetic correlation with these ancestry axes, even when stratifying traits into lifestyle and non-lifestyle categories (**Extended Data Fig. 3j**; **Supplementary Table 1**).

Genetic correlation analysis uncovers shared genetic architectures within/between trait categories

We then selected 585 child traits with positive SNP heritability estimates (additive SNP $h^2 > 0$) and one or more genetic loci discovered through GWAS (**Supplementary Table 1**), and performed the genetic correlation analysis using cross-trait LD score regression[31] (**Fig. 4**). We identified 8,057 significant genetic correlations (GCs) after multiple testing corrections (Benjamini-Hochberg $Q < 0.05$; **Supplementary Table 4**). To rigorously define the multiple testing correction threshold, we included all traits with a heritability *Z*-score greater than 0. Restricting the analysis to highly heritable traits (h^2 *Z*-score > 2 or 4) led to inflation of genetic correlation *P*-values (**Extended Data Fig. 4a**). Among these significant GCs, there were 7,672 GCs for which both traits had heritability *Z*-scores greater than 2, and 3,126 GCs for which both traits had heritability *Z*-scores greater than 4, consistent with the threshold recommended by LD score regression to ensure reliable genetic correlation estimates[31]. Heritability estimates, their standard errors, and corresponding *Z*-scores for all traits analyzed here are provided in both **Supplementary Table 1** and **4**.

Unsurprisingly, discovered GCs were significantly enriched within the same trait category (7,367 GCs; $OR = 50.7$; $P < 2.2 \times 10^{-16}$). The most frequent trait categories were "allergy and immunology" (6,312 GCs), "neurology and psychology" (472 GCs) and "anthropometry measures" (313 GCs) (**Fig. 4**). We discovered a modest number of genetic

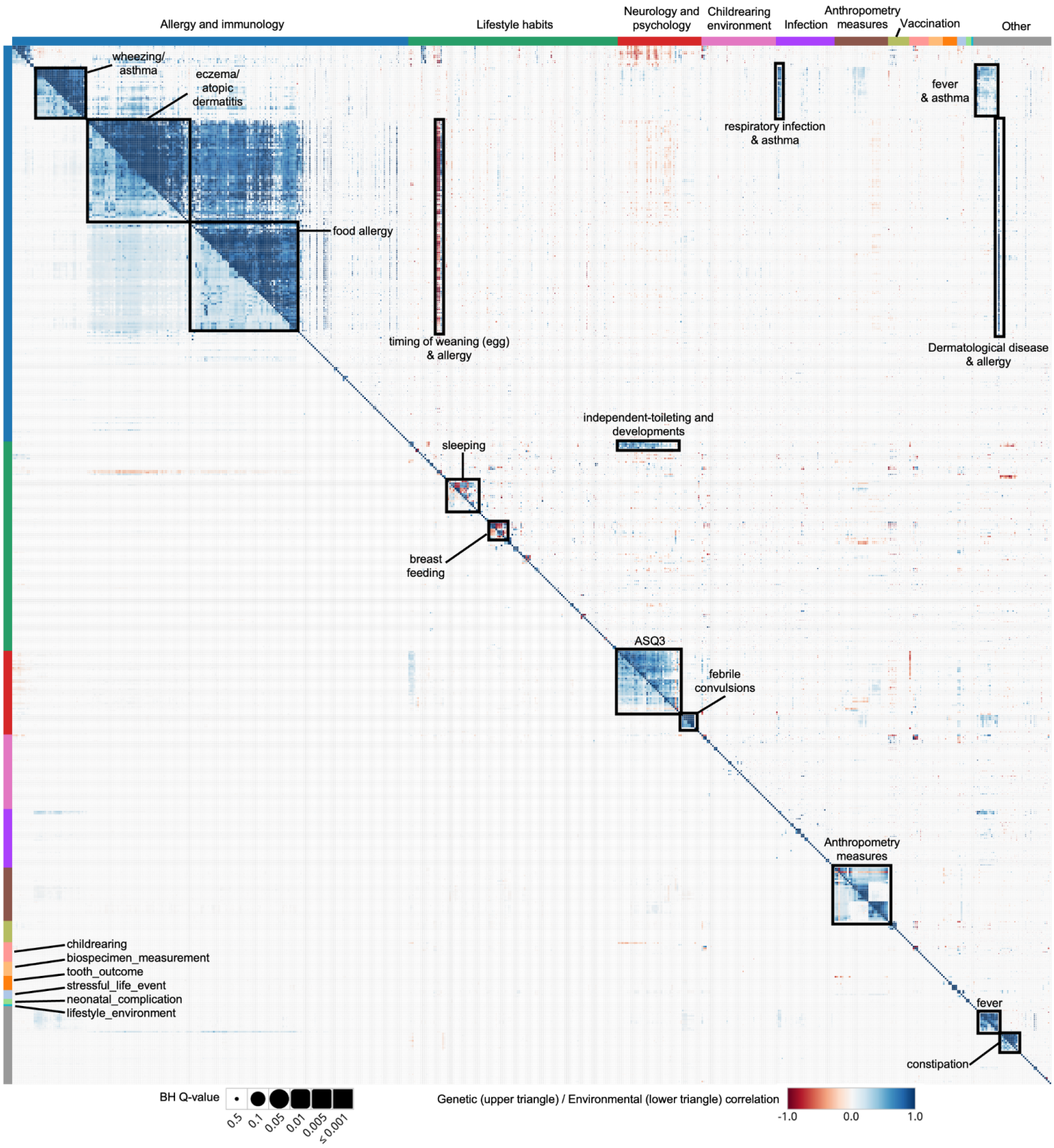


Figure 4. Genetic correlations (upper triangular matrix) and environmental correlations (lower triangular matrix) of 555 child traits estimated by the cross-trait LD score regression. Blue colors indicate positive correlations and red colors indicate negative correlations. The color bars at the top and left of the heatmap indicate trait categories.

correlations between traits in different trait categories (690 GCs across trait categories below $Q < 0.05$). For example,

263 we identified GCs between wheezing/asthma traits and fever/infection traits (**Fig. 4**). There were also reasonable GCs
264 between several developmental traits measured by the ASQ3[32,33] and independent-toileting traits (**Fig. 4**). Overall, there
265 were no surprising pairs of traits that share genetic architectures. We also note that there was general agreement
266 between environmental correlations (**Methods**) and genetic correlations of trait pairs (**Extended Data Fig. 4b**). The
267 environmental correlations are not the phenotypic correlations and computed from the output of LDSC software (see
268 **Methods** for the details).

269 This analysis not only highlighted traits whose correlations were previously unknown, but also similar traits that
270 exhibited both shared and unshared genetic loci. One such example is allergies. These often begin in early life, first
271 manifesting as eczema in young children. The progression from eczema to the subsequent development of atopic
272 dermatitis, food allergies, allergic rhinitis and asthma is known as the *allergic march*[34]. Indeed, a trait of itchy rash at
273 one year old and a trait of egg allergy score at four years old showed a very high genetic correlation ($r_g = 0.83;\ P =$
274 5.4×10^{-30}), sharing multiple genetic loci involved in skin barrier function, such as the FLG/HRNR and SERPINB7 loci
275 (**Extended Data Fig. 4c**). FLG loss-of-function variants are the most robust genetic risk factors for eczema and atopic
276 dermatitis and also increase susceptibility to food allergy and childhood-onset asthma[35,36], largely through impaired
277 skin-barrier function and enhanced skin sensitization. Although functional variants of the FLG have been extensively
278 studied in European and East Asian populations[37–39], we identified a Japanese-specific nonsynonymous variant
279 (rs540453626:G>C also known as p.S2889X) associated with both itchy rash (as a proxy for eczema) and egg allergy for
280 the first time. The clade B serpin (SERPINB) gene cluster at 18q21.3 was previously reported by a GWAS of pediatric
281 food allergy in a European population[40]. We have identified the SERPINB7 gene containing a novel East-Asian specific
282 stop-gain mutation (rs142859678: C>T) that is significantly associated with both itchy rash and egg allergy (**Extended**
283 **Data Fig. 4d**). It is known that the homozygous mutant of the SNP leads to an autosomal recessive inherited skin
284 disorder, Nagashima-type palmoplantar keratosis[41] (NPPK). This suggests that the mutant allele may also cause skin
285 barrier defects and sensitization to various allergens, even in heterozygous individuals. Interestingly, Gentamicin, an
286 antibacterial ointment, has been proposed as a potential read-through therapy for NPPK[45–47], although its clinical utility
287 remains uncertain. While highly speculative, approaches that restore skin barrier function may have downstream
288 effects on allergic sensitization in heterozygous carriers of stop-gain mutations.

289 Another example of genetically correlated traits is the food allergy traits between different foods (e.g., eggs and milk).
290 The GWAS for egg allergy showed a strong association at the MHC locus (**Extended Data Fig. 4c**), consistent with
291 previous studies showing that the MHC region represents the most robust and reproducible genetic risk locus for food
292 allergy, although prior GWAS have been limited by relatively small sample sizes, often including only a few hundred to
293 a few thousand cases[42]. In contrast, a trait of milk allergy at four years of age showed an association at the CTLA4 locus,
294 but not in the MHC region (**Extended Data Fig. 4c**). Notably, the sample size in our study is more than 4-fold larger than
295 that of previous GWAS of food allergy, including recent large-scale meta-analyses in European populations[43] (*N*=14,234
296 children), thereby providing increased statistical power to detect additional loci beyond the MHC region. Recent studies
297 have demonstrated that GWAS signals can be used to prioritize relevant cell types by integrating genetic associations
298 with functional genomic and single-cell data[44,45]. In this context, the observed locus-specific differences raise the

possibility that distinct immune cell types may contribute to the development of allergies to different foods. In line with this, scDRS[48] analysis using PBMC data from the recent single-cell multi-omics data[49] showed distinct enrichment patterns for egg and milk allergy across cell types. Specifically, significant enrichment in dendritic cells was observed only for egg allergy, and CD4+ T cells and monocytes showed a trend toward opposite directions of enrichment for egg and milk allergy (**Methods**; **Extended Data Fig. 4e,f**).

GWAS of longitudinal traits reveal age related genetic determinants

One of the potential benefits of using the birth cohort platform to study the genetics of child traits is that it allows us to track participants at multiple time points, which could help us to analyze any changes in their health and development over time[4]. One such trait that we are able to study in great detail is anthropometric traits, such as height or body mass index (BMI). As we collect the records every 6 months from birth and participants can enter them at their convenience, we have been able to retrieve the trajectory of BMI in month resolution (**Extended Data Fig. 5a**).

We initially performed standard GWAS of BMI at 11 age strata from birth to 4 years old with a proper adjustment by child age in months. We identified 454 loci with the common GWAS threshold (296 loci after phenome-wide Bonferroni correction at $P = 4.5 \times 10^{-9}$ in the 11 GWAS), comprising 197 independent genetic loci after LD pruning between index variants ($r^2 < 0.1$ in our population) (**Fig. 5a**). By comparing with the BMI associations in the GWAS catalog, we identified 108 novel loci whose index variants are in weak LD ($r^2 < 0.1$) with any of the BMI-associated variants previously reported. Because the overlap of loci is modest, we compared the effect sizes between the JECS and summary statistics from the GWAS results from up to 28,681 children in the Norwegian Mother, Father, and Child Cohort Study[4] (MoBa) which were performed within the same age strata (**Methods**). Although the effect sizes are significantly correlated (Pearson *R*=0.76), most of the associations in the smaller MoBa study were below the GWAS threshold (**Extended Data Fig. 5b**). We then compared the genetic correlations between the GWAS at the 11 time points and the adult BMI GWAS obtained from BBJ[29]. We observed strong correlations within childhood GWAS, but modest correlations between childhood and adult GWAS, suggesting that the genetic determinants of childhood obesity may be less shared with adult obesity (**Extended Data Fig. 5c**).

Incorporating all data points simultaneously, we then performed a dynamic genetic mapping using the Gaussian process regression[46] (**Fig. 5b**; **Methods**). This method explicitly models the genetic effect as a non-linear function of child age, regardless of whether the effect changes sign, changes magnitude, or appears only during part of the age range. It is more flexible than modeling only the constant mean shift of a data trajectory[47] or a linear genetic effect along age[48], or summarizing trajectories into a small set of derived phenotypes[49] (e.g., mean or slope). First, to show the statistical calibration of the model, we examined the score test *P*-values on the QQ-plot (**Extended Data Fig. 5d**) and computed the attenuation ratio of LDSC regression, which was sufficiently low ($Ratio = 0.19$), despite the high genomic control lambda ($\lambda_{GC} = 1.59$) (**Methods**). Next, we simulated genotypes with a MAF ranging from 0.001 to 0.5 (**Methods**) and confirmed that the *P*-values follow the uniform distribution (**Extended Data Fig. 5d**). We also performed

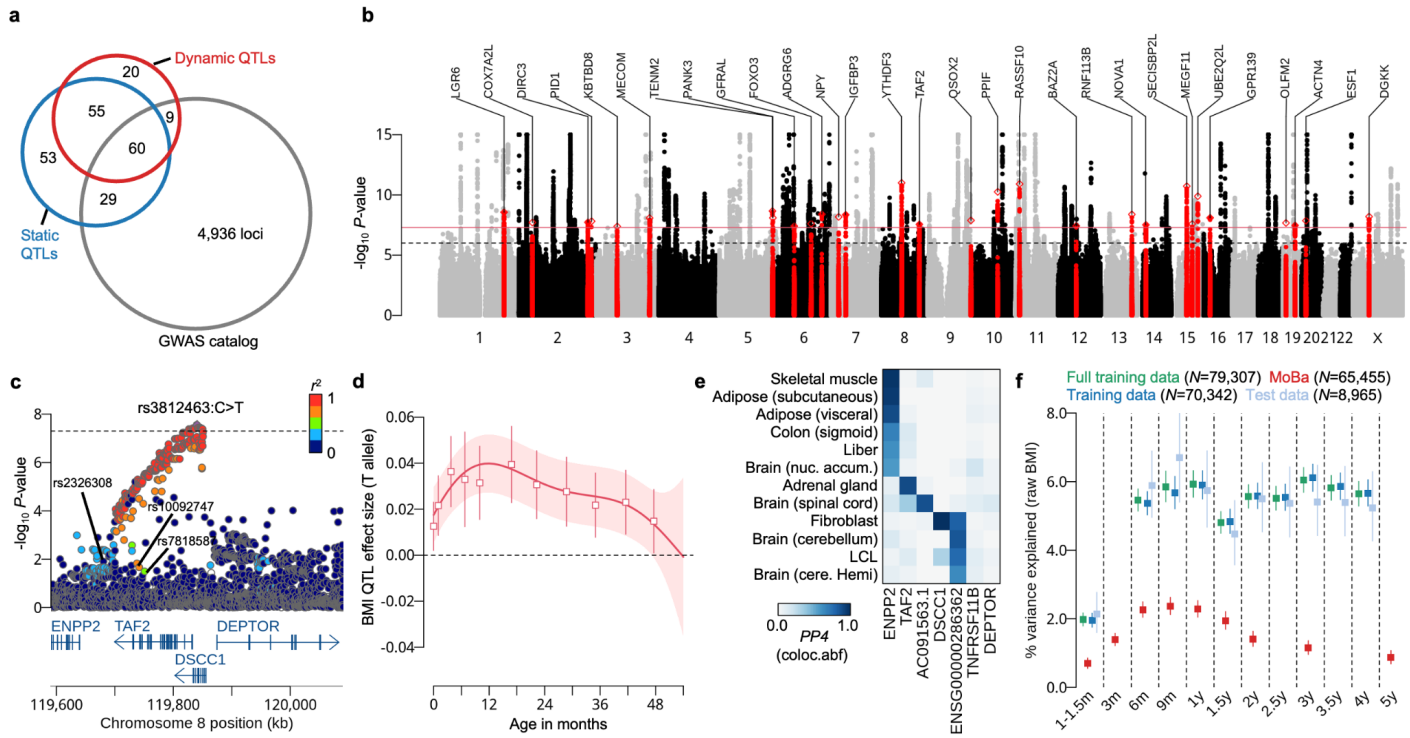


Figure 5. Genetic study of childhood BMI as a longitudinal trait in JECS. **a.** Static and dynamic genetic loci discovered through longitudinal GWAS of childhood BMI compared with BMI-related loci reported in the GWAS catalog. **b.** Manhattan plot shows the score test *P*-values obtained from the GWAS of the dynamic genetic effects of BMI from birth until 4 years old. The 29 novel loci (panel a) are highlighted by red and the nearest genes are annotated. **c.** Locuszoom plot of the TAF2 locus shows the statistical significance of dynamic genetic effects. Points are colored by the LD index value (r^2) to the index variant rs3812463:C>T. The three height-associated SNPs (rs2326308:C>T, rs10092747:C>T, rs7818587:G>T) are also shown in this plot. **d.** Dynamic genetic effect of the T allele over the C allele at rs3812463 estimated by the Gaussian process regression along the child age. The red line shows the posterior mean with the 95% credible interval of the dynamic effect. The static QTL effects and their 95% confidence intervals for the 11 age strata were superimposed as well. **e.** Heatmap shows the colocalization posterior probability (PP4) obtained from the coloc package in R. We used the summary statistics of GTEx eQTLs (V8) across 48 tissues in 1Mb cis-window centered at each gene TSS. We found 12 tissues showed PP4 greater than 0.5 in one of 5 different eQTL genes. We haven't discovered any colocalized eQTLs for TNFRSF11B or DEPTOR. **f.** The performance of PGS in our data and in the MoBa data. Each dot shows the variance in raw BMI explained by PGS at each age stratum, along with its 95% confidence interval. The green dots show the performance using all participants (*N*=79,307) in JECS, and the blue dots show the performance using the training data (*N* = 70,342), excluding relatives beyond the third degree of kinship. Light blue dots show the performance of the remaining participants (test dataset; *N* = 8,965), excluding those used for the training data. The red dots show the performance of MoBa using European participants (*N* = 65,455).

a permutation test using the real genotype data (**Methods**) and demonstrated that the *P*-values follow a uniform distribution. (**Extended Data Fig. 5d**).

Using this approach, we discovered an additional 29 independent genetic loci ($r^2 < 0.1$ with loci discovered through the static GWAS of 11 age strata), of which 20 loci were novel and not in LD ($r^2 < 0.1$) with any BMI-related variant (**Fig. 5a,b**; **Supplementary Table 5**). As a proof of concept, we first examined the FTO locus, which is strongly linked to adult obesity. The FTO locus is known to be associated with infant BMI with an opposite effect size to adults because of the difference in the ratio of brown to white adipose tissue in infants and adults[50]. We clearly showed that the QTL effect was reversed at around 3 years of age (**Extended Data Fig. 5e,f**), as previously reported[51]. Then, we examined

our novel discoveries, which were enriched for variants with a significantly lower minor allele frequency (MAF) in a European population (Kolmogorov–Smirnov test $P = 0.02$; **Extended Data Fig. 5g**). One of these novel associations was identified in the TAF2 locus (**Fig. 5c**). The minimum *P*-value was achieved by a SNP (rs3812463:C>T) in an intron of the TAF2 gene, although 207 additional SNPs and INDELs near this SNP were in strong LD ($r^2 > 0.8$). The estimated effect size of the T allele over the C allele showed the dynamic effect, increasing in infants and decreasing in toddlers (**Fig. 5d**). Note that, no genome-wide significant associations were observed in the GWAS of 11 distinct age strata (**Fig. 5d;** the minimum $P = 5.7 \times 10^{-6}$ was attained at 1.5 years old). We also found that there are several height-associated index variants (rs2326308:C>T, rs10092747:C>T, rs7818587:G>T) in this locus[16] (**Fig. 5c**). However, the LD between the TAF2 index variant and those variants was modest ($0.25 \leq r^2 \leq 0.64$ in our population). Therefore, distinct association signals may be present, although population-specific LD patterns also allow for a shared causal variant. As there are several putative candidate genes in this locus, we checked the colocalization with eQTLs in GTEx (V8; **Data Availability**). We found that the locus is colocalized with 5 different eQTL genes in 12 different tissues (colocalization posterior probability *PP4*>0.5) (**Fig. 5e**). One of them was the ENPP2 eQTL in skeletal muscle tissue ($PP4 = 0.87$), and the locuszoom plot was consistent with the GWAS locus (**Fig. 5c** and **Extended Data Fig. 5h**). ENPP2 encodes an enzyme that converts lysophosphatidyl choline to lysophosphatidic acid (LPA) leading to growth factor-like responses such as stimulation of cell proliferation or chemotaxis[52].

Together with the set of dynamic genetic effects estimated at each locus (*e.g.* TAF2 locus in **Fig. 5d**) combined with individual genotype data, a dynamic polygenic score (PGS) can be constructed along the child's age (**Methods**). This information reveals the extent to which the genetic influence deviates from the population mean and the degree to which the genetic prediction aligns with the observed child's growth (**Extended Data Fig. 5i-l**). In collaboration with MoBa, we carefully assessed the performance of the dynamic PGS model, comprising 226 variants using both internal and external data. First, we trained the dynamic PGS model using all our samples with BMI observations at least once in JECS (*N* = 79,307). Next, we evaluated the model's performance in each age stratum of the trained data (**Methods**). This demonstrated that the average raw BMI explained by the PGS was 5.3% (ranging from 2.0% to 6.0% across the age range) in JECS (**Fig. 5f**). When we applied the model to an independent cohort dataset from MoBa (*N*=65,455; Norwegian children of Northern European descent), we found that the PGS prediction based on 194 variants available in MoBa, on average, explained 1.6% of the variance (ranging from 0.7% to 2.4% across the age range) in this European population. Notably, the PGS model still explained 0.9% ($P = 3.2 \times 10^{-68}$) of the variance at age 5 in MoBa (outside the range of the trained model) (**Fig. 5f**). We also split our data by excluding relatives beyond the third degree, using the resulting subset as training data (*N*=70,342) and the remainder as test data (*N*=8,965). This approach yielded similar performance compared with the full dataset. The constructed dynamic PGS model with the full sample data is available on our GitHub page (**Software availability**).

Potential effect of maternal exposure during pregnancy on child health and development

Adverse environmental conditions and pollution have been identified as significant contributors to the high rates of

childhood mortality, morbidity and disability[53]. Another notable benefit of JECS is its ability to assess environmental exposures of participating parents, including environmental pollutants such as heavy metals[8] and other chemical substances[54], as well as lifestyle factors (such as dietary habits, smoking, and alcohol consumption). GWAS-by-proxy of 209 exposure traits has identified 138 traits with at least one genome-wide association ($P < 5.0 \times 10^{-8}$) (**Supplementary Table 1**; **Fig. 2e**), which could be a valuable addition to the knowledge of genetic determinants for children.

Even after accounting for the approximate four-fold reduction in effective sample size under the GWAS-by-proxy design[28], our study detected a number of genetic associations to metal (mercury, cadmium, manganese and lead) or metalloid (selenium) exposure. The nominal sample size of 77,614 for those traits was almost 50 to 100 times larger than those studied previously[55,56]. We have observed pleiotropic effects among those traits. One of the outstanding effects was observed in ALDH2 locus (**Fig. 5a**), which is implicated in the metabolism of aldehydes, modulating drinking habits, and highly polymorphic in East Asian populations[57]. The locus was detected in maternal blood mercury and lead levels as well as subtypes of PFAS[58]. We observed the opposite effect direction between alcohol intake and blood mercury level at the missense SNP rs671 (**Fig. 5a**), suggesting that higher alcohol intake is associated with lower mercury levels. Although this pattern may reflect pleiotropic effects of ALDH2, this finding aligns with previous reports indicating that alcohol consumption elevates mercury excretion in both mice and humans[59,60]. As for the blood lead and PFAS levels, the effect directions are the same as alcohol intake (**Fig. 5a**). In addition, the index variant for the lead level is different from alcohol intake (**Fig. 5b-d**), there can be different mechanisms of metabolism and excretion for these substances. Furthermore, we sought to provide empirical support for the proxy-based findings at the ALDH2 locus, directly genotyped participating mothers (*N* = 63,528), including rs671 (**Methods**), and confirmed consistent effect directions and magnitudes (**Fig. 6a**).

We then performed the cross-trait genetic correlation analysis, which revealed shared genetic architectures between exposure traits (**Extended Data Fig. 6; Supplementary Table 4**), such as positive correlations between maternal blood lead level and red blood cell counts ($r_g = 0.52; P = 2.0 \times 10^{-6}$), and eosinophil percentage and various IgE levels ($0.42 \leq r_g \leq 0.52$; $P < 2.0 \times 10^{-4}$) during pregnancy. It is worth noting that paternal exposure traits (e.g., cholesterol levels) tend to be strongly clustered and not correlated well with maternal traits. This may reflect biological differences between the sexes and/or the influence of pregnancy on maternal physiology. It may also capture the contribution of fetal genetic effects specific to maternal physiology, as well as differences in trait measurement within the GWAS-by-proxy framework.

Finally, we investigated potential mediating effects of maternal environmental exposures on the health and developmental traits of their children. Using genetic correlations obtained by the cross-trait LDSC regression, we identified several maternal IgE levels that were positively correlated with various allergic traits in children, such as itchy rash and food allergy (**Fig. 5e**). We also discovered maternal albumin and total protein levels during pregnancy, which are negatively correlated with the anthropometric traits of children at birth (**Fig. 5e**). Surprisingly, we found very little evidence that other maternal environmental exposures affect child health and development (**Supplementary Table 4**).

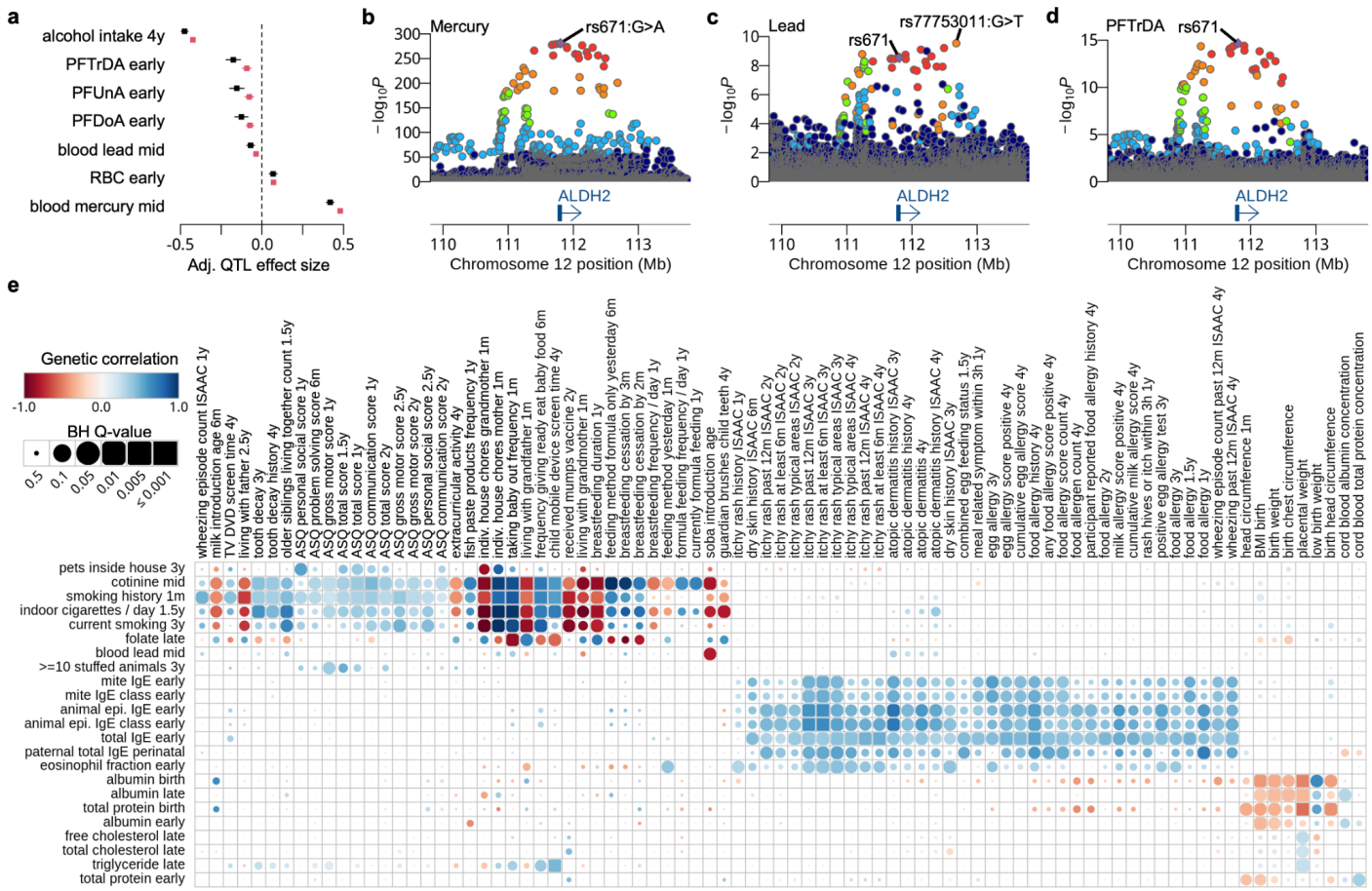


Figure 6. Genetics of environmental exposure traits for participating parents. **a.** QTL effect sizes of the missense SNP rs671 (G>A) at the ALDH2 locus across multiple environmental exposure traits (black: GWAS-by-proxy; red: GWAS using directly genotyped maternal data). Effect sizes are reported for the A allele relative to the G allele. Here the effect sizes from GWAS-by-proxy were rescaled ($\times$ 2) for visualization purposes to be comparable to those from GWAS using direct maternal genotyping. **b.** Locus zoom plot of the blood mercury level GWAS at the ALDH2 locus. **c.** Locus zoom plot of the blood lead level GWAS at the ALDH2 locus. **d.** Locus zoom plot of the blood PFTrDA level GWAS at the ALDH2 locus. **e.** Cross-trait LDSC analysis of parental environmental exposure traits on child health and developmental traits. Only trait combinations with Q-value<0.05 are shown in the heat map.

For example, we found a potential effect of smoking exposure during pregnancy on tooth decay in children at 4 years of age ($r_g = 0.38$; $Q = 1.1 \times 10^{-5}$). However, we did not observe strong evidence that a specific genetic locus is colocalized between mother and child traits or shares a putative causal variant. It is unlikely that there is a shared gene regulatory cascade at the molecular level. Moreover, genetic correlations estimated in the GWAS-by-proxy framework may be influenced by pleiotropy, fetal genetic effects, and correlated environmental or socio-behavioral factors, and therefore do not necessarily reflect direct or causal relationships between maternal exposures and child outcomes.

Discussion

In this study, we conducted a large-scale GWAS of child health and development traits in more than 80K individuals,

441 greatly expanding our knowledge of common genetic determinants in various early-life traits in an East Asian
442 population. The study included 1,046 traits in children up to the age of four, as well as 209 environmental exposure
443 traits in their parents. We identified 5,685 genetic associations across 21 different trait categories, some of which are
444 highly pleiotropic at MHC or ALDH2 loci. In addition, we discovered a modest number of genetic correlations between
445 children's traits that go beyond their respective categories. The analysis also identified a stop-gain variant in *SERPINB7*,
446 previously reported in Nagashima-type palmoplantar keratosis (NPPK), as a candidate locus shared between egg allergy
447 and skin barrier dysfunction, suggesting a possible connection to therapeutic hypotheses.

448 Our analysis of longitudinal child BMI traits revealed novel and dynamic genetic associations during development,
449 enabling us to construct individualized growth curves in the form of dynamic polygenic scores. Typically, the child's
450 growth is benchmarked against the population average. These personalized reference values can improve the early
451 detection and management of growth disorders and obesity by distinguishing between environmental influences and
452 genetic predisposition. Recognizing such deviations could ultimately facilitate the early identification of children whose
453 growth falls outside the genetically expected trajectory, thereby not only enabling the prediction and prevention of
454 lifestyle-related diseases, but also prompting further evaluation for underlying etiologies, including rare congenital
455 disorders[66,67]. This approach could also help to reduce overdiagnosis and overtreatment, as well as healthcare costs,
456 while providing reassurance when growth is within expected ranges.

457 A fundamental limitation of the GWAS-by-proxy (GWAX) framework is that offspring genotypes capture a mixture of
458 direct and indirect genetic effects. Associations with parental phenotypes may therefore reflect a combination of
459 parental genetic effects, fetal genetic effects acting on maternal physiology, and indirect effects such as genetic nurture.
460 Because offspring genotypes are correlated with both maternal and paternal genotypes (correlation approximately 0.5),
461 these components cannot be disentangled without joint parent-offspring genotyping, and the observed associations
462 should be interpreted as composite signals rather than purely parental genetic effects. In pregnancy-related traits, fetal
463 genetic influences on maternal physiology add further complexity, such that differences between maternal and
464 paternal analyses cannot be unambiguously attributed to maternal-specific effects. In addition, recent studies[61,62] have
465 shown that GWAX estimates may be biased due to imperfect proxying, measurement error, residual confounding, and
466 assortative mating, which may be particularly relevant for behavioral and environmental traits. Genetic correlations
467 derived from this framework should therefore be interpreted with caution. Nevertheless, for loci with well-established
468 biological functions, such as ALDH2, the observed associations are consistent with known physiological pathways,
469 although contributions from fetal and indirect effects cannot be excluded. Overall, GWAX enables large-scale analyses
470 in the absence of parental genotypes, but does not allow definitive separation of maternal, paternal, and fetal effects,
471 warranting cautious interpretation.

472 Some of the traits analyzed in this study, particularly lifestyle-related, childrearing, and social traits, are strongly shaped
473 by environmental, cultural, and systemic factors. Although we found no evidence that these traits were driven by the
474 major ancestry axes within the cohort, genetic associations involving such phenotypes should be interpreted with
475 particular caution. These signals may reflect indirect pathways, correlated socio-behavioral factors, or residual

476 confounding, rather than direct biological effects on the measured trait itself. We therefore regard these findings
477 primarily as hypothesis-generating and as a means to characterize the breadth of phenotypes captured in this birth
478 cohort, rather than as evidence of genetic determinants.

479 We also identified a limited number of genetic correlations between parental environmental exposure traits as well as
480 maternal exposure traits during pregnancy and child traits after birth. One possible explanation for the modest number
481 of observed correlations is that our GWAS-by-proxy approach reduces power to detect genetic associations in parental
482 traits, approximately to one quarter in terms of chi-squared statistics under simple assumptions. Maternal genotyping
483 is currently underway and will be completed in 2026. At that time, we will have substantially increased power to detect
484 genetic associations of parental exposure traits, and we will also be able to disentangle genetic effects of children from
485 those of parents on birth outcomes such as birth weight.

486 The analysis of mCA revealed that approximately 1% of study participants carried mCA in their cord blood samples.
487 Although several previous studies have examined chromosomal mosaicism in early life, these investigations were
488 limited by either small sample sizes or insufficient sensitivity. For example, Zhou et al.[63] analyzed a total of 4,512 fetal
489 samples derived from chorionic villus sampling, amniotic fluid, and cord blood combined, but the number of cord blood
490 samples was considerably smaller than in our study. In addition, Li et al.[64] defined chromosomal mosaicism using a
491 ≥10% mosaic cell-fraction threshold, which inherently misses low–cell-fraction events. As a consequence, prior studies
492 were unable to comprehensively characterize low-level mosaicism or to distinguish autosomal mCAs, LOX, and LOY
493 with high precision. By using a large cohort of more than 80,000 cord blood samples and a high-resolution haplotype-
494 based method (MoChA), our study provides the first robust, population-level estimate of neonatal mCA frequencies
495 that is directly comparable to adult datasets such as BBJ. As mCAs are strongly associated with clonal hematopoiesis
496 and increased risks of cancer[64] and infertility[65] in adulthood, it is imperative to follow up participants over the course
497 of their lives to ascertain whether mCA in early childhood develop diseases in later life. However, two important caveats
498 should be noted. First, because the prevalence of mCA in cord blood is expected to be low[63,66], the positive predictive
499 value of the MoChA program is likely lower than in adult samples, implying a higher false-positive rate. Second, the
500 higher LOY prevalence observed in this study compared with previous reports[67–69] is likely due to the absence of a lower
501 cutoff for cell fraction and the inclusion of partial mCA events (**Methods**).

502 Our analyses primarily identified common genetic variants, which is consistent with the well-known properties of
503 genome-wide association studies, where statistical power and imputation accuracy are generally higher for common
504 alleles[70]. Future studies incorporating exome or whole-genome sequencing will be essential to capture the contribution
505 of rare variants, which are often poorly tagged by genotyping arrays[71]. In addition, the integration of both short- and
506 long-read sequencing technologies will enable more comprehensive detection of complex structural variants, which
507 are increasingly recognized as important contributors to human traits and disease[72]. Such approaches will be
508 particularly valuable for elucidating the genetic architecture of child health and developmental traits, which may
509 involve a broader spectrum of genetic variation than can be captured by array-based GWAS alone.

To fully connect variants to their biological functions, we also need to fine-map each genomic locus discovered through GWAS. For example, we could use SuSiE[73] to identify secondary hits in each locus and integrate genomic annotations, such as ChIP-seq or chromatin accessibility peaks used in sLDSC[44] or fGWAS[74]. We could also colocalize with public eQTL databases, such as GTEx[75], to annotate functional genes in non-coding GWAS loci. These works are left for further investigation.

JECS will continue until the participating children reach the age of 18 and beyond. The phenotype data freeze up to the age of 8 will soon be available to further increase the number of phenotypes analyzed in GWAS. We are continuing to uncover the genetic determinants of various childhood traits in early life, which may have the potential to prevent various adult diseases at a much earlier stage or even during pregnancy.

Methods

Ethical compliance

JECS protocol was reviewed and approved by the Ministry of the Environment's Institutional Review Board on Epidemiological Studies (No.100910001) and the Ethics Committees of all participating institutions. Written informed consent was obtained from all children's legal guardians. Following informed consent, participants were given further information about SNP genotyping using microarray technology and the possibility to withdraw from the genetic study at any time.

Data source

The present analyses were conducted using the jecs-ta-20190930 dataset, which encompasses follow-up data until 3 years of age for the children, maternal/cord blood laboratory test results, maternal blood concentrations of five elements (lead, cadmium, mercury, selenium, and manganese), and maternal urinary cotinine concentrations. Additionally, the jecs-qa-20210401 dataset (including 103,057 pregnancies and 100,300 live births) was used, extending follow-up to 4 years of age and including data on maternal plasma concentrations of PFAS.

Collection of cord blood samples

JECS collected a single cord blood sample at delivery, even in cases of multiple births. For monozygotic twins or triplets, genotypes obtained from SNP arrays are expected to be identical for common variants, whereas dizygotic twins are genetically equivalent to siblings. When multiple-birth records were linked to a single cord blood genotype, we assigned the same genotype data of a twin to the co-twin(s), unless the genotype and clinical sex were inconsistent, and we excluded the child whose reported sex was inconsistent with the genotype-inferred sex. This procedure allowed partial resolution of dizygotic twin pairs when the co-twins were of opposite sex. However, same-sex dizygotic twins could not be distinguished from monozygotic twins using genotype data alone. Based on the reported ratio of monozygotic to dizygotic twins in Japan[76], we estimate that approximately 100–200 non-genotyped dizygotic co-twins may remain

541 represented in the analysis.

542 ## DNA extraction from cord blood samples

543 DNA from whole cord blood was extracted using the QIAsymphony SP system and QIAsymphony DSP DNA Midi Kit
544 (QIAGEN). Genomic DNA was quantitated on the SpectraMax M2e Microplate Reader (MOLECULAR DEVICE) and the
545 Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific).

546 ## SNP genotyping

547 We genotyped 80,639 cord blood samples using the Illumina Asian Screening Array (ASA-24v1-0_E2) BeadChip SNP
548 array (https://www.illumina.com/products/by-type/microarray-kits/infinium-asian-screening.html). We conducted
549 genotype calling on Illumina GenomeStudio software to generate PED files. We excluded samples with sample call rate
550 <0.98. These PED files were first converted into the VCF format using PLINK2 (**Code availability**). We used the fixref
551 plugin of bcftools to align the strand orientation of each SNP (**Code availability**). For palindromic SNPs (A-T or C-G SNPs),
552 BLAST software (ver 2.5.0) was used to identify the SNP probe orientations a priori (**Code availability**). After fixing the
553 reference alleles, variants with a variant call rate less than 0.99, a HWE *P*-value less than 10^{-6} and an allele frequency
554 that deviated from four times the standard error (SE) of the allele frequency estimate for the BioBank Japan (BBJ) or
555 the Tohoku Medical Megabank (ToMMo) samples (**Data availability**) were excluded. We estimate that the BBJ and
556 ToMMo allele frequencies have a SE of $\sqrt{p(1-p)/(2N)}$ with sample sizes of 13,406 and 54,000, respectively. The
557 genotype data was then pre-phased using the Eagle software (**Code availability**) without reference haplotype data
558 because the sample size was sufficient for self-inference of population-specific haplotypes. Finally, the phased
559 genotype data (with 617,557 variants) was harmonized with the 1000 Genome Project VCF (**Data availability**) and then
560 missing genotypes were imputed using the Minimac4 software (**Code availability**).

561 ## Detection of chromosomal aberrations

562 Chromosomal aberrations were detected from the signal intensity data for each DNA sample. Nominal XY signal
563 intensities were extracted from Illumina idat files and then converted into B allele frequency (BAF) using the polar
564 coordinate transformation. Here, A and B denote the generic allele labels used by Illumina microarrays, and the BAF
565 intuitively represents the estimated proportion of the B allele at each locus. After hybridization and scanning, the
566 fluorescence intensities for these two alleles are measured. Note that, the A/B alleles are defined in the manifest file
567 of Illumina SNP genotyping arrays (**Data availability**), where each SNP is designed with two probes corresponding to
568 the reference ("A") and the alternative ("B") allele. However, this definition is irrelevant for the "Ref" and "Alt" alleles
569 based on the human reference genome sequence. A mixture of Gaussian mixtures was fitted to detect the number of
570 copies of each autosomal or sex chromosome. The means and variances for the Gaussian mixture distributions were
571 fixed and posterior probability of aberrations were obtained.

572 ## MoChA analysis for detecting mosaic chromosomal alterations

573 Illumina idat files were processed to generate VCF files including BAF and log R ratio (LRR) values for each variant. To
574 detect somatic mosaic chromosomal alterations (mCAs), we further analyzed the BAF and LRR data using the MoChA
575 (**Code availability**), with GRCh38 as the reference genome. Haplotype phasing was performed with SHAPEIT5 (**Code**
576 **availability**), following the established protocol of MoChA. To minimize the inclusion of germline aberrations, events
577 with the cell fraction greater than 0.5 were excluded. However, we did not introduce a lower cutoff value for the cell
578 fraction. Additionally, we filtered out events with relative coverage (rel_cov) ≥ 2.5, as these may represent germline
579 duplications or technical artifacts.

580 To compare with adult samples, we additionally analyzed participants from the first (BBJ1) and second (BBJ2) cohorts
581 of the BioBank Japan Project[77,78]. The mean age of BBJ1 participants was 62.7 (N = 179,254), and the mean age of BBJ2
582 participants was 65.2 (N = 71,469).

583 Population stratification analysis using principal component analysis and ADMIXTURE

584 Population stratification analysis was performed by principal component analysis (PCA) of genotype dosages with
585 reference populations obtained from the 1000 Genomes Project (**Data availability**) and Korean Personal Genome
586 Project (**Data availability**). Korean genotype data (referred to as KOR) based on whole genome sequencing was
587 harmonized with the 1000 Genome Project VCF and then missing genotypes were imputed using BEAGLE software
588 (**Code availability**). SNP loci were chosen based on our sample statistics with MAF ≥ 1% and HWE *P*-value > 1.0×10^{-6}.
589 Then, linkage disequilibrium pruning was performed using "--indep-pairwise 50 10 0.1" in PLINK2, yielding 165,756
590 biallelic SNPs. PCA was performed using the PLINK2 "--pca" command with the "allele-wts" modifier to obtain the first
591 20 principal components (PCs) as well as SNP loadings. The first two components were shown in **Fig. 1e**.

592 We then clustered JECS participants in principal component space by anchoring the known worldwide populations AFR,
593 EUR, AMR, SAS, and EAS (including KOR) using the first 7 genotype principal components. A Gaussian mixture model
594 was fitted to identify individuals with admixed ancestry, including those with approximately half or one-quarter
595 ancestry from a given reference population (for example, Japanese–European mixed ancestry), and only samples
596 assigned to the East Asian cluster were retained. We next performed principal component analysis in the East Asian
597 subset together with JPT, CHB, and KOR reference samples from the 1000 Genomes Project and the Korean Personal
598 Genome Project; the first two components are shown in **Fig. 1f**. Finally, we selected the first 7 PCs as covariates, as
599 their SNP loadings were uniformly distributed across the genome.

600 The ADMIXTURE analysis (**Code availability**) was also applied to the same data set including the worldwide populations
601 (**Fig. 1e**). We ran the ADMIXTURE software with K=5 and the default parameter settings.

602 Phenotypes

603 Phenotype data were derived from the Japan Environment and Children's Study (JECS), a large-scale birth cohort with
604 longitudinal measurements collected from pregnancy through early childhood. We extracted 4,460 variables from 23
605 data sheets encompassing questionnaire responses, clinical measurements, and environmental exposures in children

up to 4 years of age and their parents (all analyses were conducted using datasets named jecs-ta-20190930 and jecs-qa-20210401). Variables were first screened for usability based on predefined criteria, excluding metadata, duplicate entries, variables outside the scope of the study, and those with ambiguous definitions. The remaining variables were classified into candidate traits, composite trait source variables, and covariate source variables according to their role in downstream analyses.

All retained variables underwent standardized quality control and transformation procedures. For continuous variables, outliers were excluded using a four-standard-deviation (4SD) criterion unless clear physiological or measurement-based bounds were defined a priori (e.g., laboratory measurements or questionnaire-derived scores with fixed ranges), in which case lower and upper bounds were applied instead. The choice between 4SD filtering and bounded exclusion was determined based on clinical plausibility and measurement characteristics, as reviewed by domain experts. Binary and categorical variables were harmonized to ensure consistent encoding, and additional processing included conditional correction of nested questionnaire responses and exclusion of non-maternal respondents for respondent-dependent traits. The specific QC criteria applied to each variable are provided in **Supplementary Table 6**.

We then constructed traits from these variables, including both directly analyzable variables and composite traits derived from multiple source variables (e.g., BMI, allergy-related traits, and developmental measures). Variables that were not suitable for independent analysis (e.g., due to low interpretability, copyright restrictions, or distributional properties) were incorporated into composite traits where appropriate. Covariates were selected from a subset of variables representing participant metadata or potential confounders (e.g., gestational age), with redundancy removed where necessary.

After trait construction, we considered several traits unsuitable because they had too many missing responses (low call rate) or were too low in complexity (*i.e.* too many identical values/answers in the trait), leading to high false positives. We calculated the call rate of a trait, such as

```
(call rate) = 1 - (number of valid data) / (number of questionnaires collected
at the given age of the children),
```

and excluded traits with more than 40% missing responses from the analysis. The complexity score, defined as

```
(complexity) = 1.0 - (frequency of the most frequent value in the phenotype data)
/ (number of observations without missing values),
```

was also calculated for each trait (note that this score should be the case ratio for a case-control trait). We removed binary traits with a score of less than 0.01 and quantitative traits with a score of less than 0.05 to reduce potential false positives in GWAS. A total of 1,255 parental and child traits were analyzed in the GWAS.

To ensure transparency and reproducibility, we systematically documented the full mapping from raw variables to derived traits, including excluded variables, covariates, and composite trait definitions, in the **Supplementary Notes 1** and **Supplementary Table 6** and **7**.

Genome-wide association study (GWAS)

GWAS of all the 1,255 traits was performed using REGENIE (**Code availability**). In the first step, 546,576 SNPs directly genotyped on the Illumina ASA chip were used to obtain the leave-one-chromosome-out (LOCO) predictions. In the second step, 17,454,650 imputed variants (MAF greater than 0.1% and imputation score R2 greater than 0.3) were used for GWAS. Quantitative and ordered categorical traits were transformed using the rank-based inverse normal transformation for genome-wide quantitative trait locus (QTL) mapping. Non-ordered categorical traits were transformed into a set of binary traits using the one-hot encoding, and the Firth logistic regression was used to perform GWAS in conjunction with other binary phenotypes (such as disease onset).

Covariates used for each trait were determined based on the child's age in months at the time the questionnaire was filled out. For the traits at the birth, we used (1) mother's age at birth, (2) gestational duration (in days), (3) infant's sex, (4) the first 7 genotype PCs. For the traits from 6 months to 4 years old of children, we also included the child's age in months as covariates. For the caregivers traits, age of the caregivers and the first 7 genotype PCs were used.

Gene annotation was applied using ANNOVAR (**Code availability**) with the Ensembl 113 gene annotation (**Data availability**). The rsID of each variant was assigned using ANNOVAR with the dbSNP version 151 as the default (**Code availability**).

A GWAS locus was defined by a variant with a *P*-value below the standard GWAS threshold of 5.0×10^{-8} with 500 kb of upstream and downstream sequence (1Mb window). If there are multiple variants with *P*-values below the standard GWAS threshold within the 1Mb window, we group them as a same locus, and extend the window by 500Kb upstream and downstream for the set of variants until no variants exceed the standard GWAS threshold. The index variant in the locus was then defined as the variant with the minimum *P*-value within the window. The secondary independent variants within the same region were discovered by the REGENIE' s conditions analysis given the primary index variant(s) as the covariate(s). We iteratively performed the conditional analysis until no variants exceed the standard GWAS threshold within the window.

We used our imputed genotype to compute the linkage disequilibrium (LD) r^2 index between all index variants in our study and variants in the GWAS Catalog Release 2024-11-03 (**Data availability**). We computed the LD index only for pairs of variants within 1 Mb distance between our index variant and the GWAS catalog variant. We omitted any GWAS catalog variant that was not imputed in our genotype data. A total of 231,405 out of 296,928 GWAS Catalog variants were covered by the analysis, excluding genotype-by-genotype interactions.

HLA imputation and association analysis

We performed HLA imputation using a Japanese population–specific reference panel (n = 1,118) constructed in a previous study[79]. We applied Deep*HLA (**Code availability**), a deep learning-based HLA imputation method[80], to the genotyped SNPs in the MHC region (25 to 35 Mb on chromosome 6) and imputed the classical HLA alleles (one-, two-, and three-field [two-, four-, and six-digit]) and amino acid variants of the classical and non-classical HLA genes. HLA

variants imputed with an imputation quality score (R2 in 10-fold cross-validation) > 0.7 and MAF > 0.005 were used for further analyses.

Associations between HLA variants and traits were evaluated using REGENIE (**Code availability**), following the same framework as the GWAS. To identify independently associated HLA loci, we performed step-wise conditional analyses by additionally including all classical HLA alleles of the HLA gene to which the lead variant belonged as covariates[79]. When the lead signals were *HLA-DRB2⁻9* variants and were in strong LD with *HLA-DRB1* variants (R2 > 0.9), we jointly included *HLA-DRB1* classical alleles in the model, considering that signals in the *HLA-DRB2-9* region are often driven by strong LD with *HLA-DRB1*, which is the primary functional locus in this region.

GWAS with unrelated participants

To assess the robustness of our GWAS results to relatedness, we constructed a subset of unrelated individuals by excluding samples with kinship coefficients greater than 0.0442, corresponding to third-degree relatives or closer, as estimated using KING[22]. From each related pair, one individual was preferentially retained based on higher genotype call rate. The resulting subset (*N*=70,751) was used to perform GWAS using the same analytical framework as in the full dataset.

In the subsample analysis, association test statistics were interpreted under the assumption that χ^2 statistics for non-null variants follow a non-central chi-square distribution with one degree of freedom. Under this model, the non-centrality parameter is proportional to the sample size, such that the expected test statistics in the subsample are scaled by the ratio of sample sizes relative to the full dataset. A full derivation of the underlying statistical framework, including all relevant equations, is provided in **Supplementary Notes 2**, Appendix 2.1 and 2.2.

GWAS for ancestry axes

To assess the influence of population structure, we conducted GWAS using genotype principal components (PCs) as quantitative traits (ancestry axes). The first two PCs, which capture the major axes of genetic variation within the cohort, were used as the basis for defining ancestry-related variation. To facilitate biological interpretation, we applied a linear rotation of these PCs using a rotation matrix to derive orthogonal axes corresponding to variation between (i) Japanese and other East Asian populations, and (ii) mainland Japanese (Hondo) and Ryukyu ancestries.

GWAS was then performed on these rotated axes using REGENIE, treating each axis as a quantitative phenotype. The analyses were conducted using the same pipeline as in the primary GWAS; however, genotype principal components were not included as covariates in these models, as they constitute the outcome variables. This approach enables identification of genetic variants associated with ancestry-related variation and allows evaluation of potential confounding due to population structure.

GWAS of parental traits using proxy genotypes (GWAS-by-proxy)

We first removed any parental trait that was observed from a mixture of mother or father or other family members

(*e.g.* the question: *An estimate of cigarettes smoked inside the house per day*). For each parental trait, we regressed the maternal (or paternal) phenotype on the child's imputed genotype dosages using REGENIE's mixed-model framework (**Code availability**), analogous to the child-trait GWAS (see above). We used the same quality-filtered variants and covariates except for the child sex and the gestational duration, and we restricted to parent–child pairs with available phenotypes. Here, the offspring genotype acts as a proxy for the parental genotype, with an expected correlation of 0.5. The effect estimates were then attenuated by half, relative to a genome-wide association study (GWAS) based on parental genotypes under the assumption of no additional indirect genetic effects or assortative mating. In the **Supplementary Table 2**, we reported unadjusted effect sizes from the proxy analyses (except for Fig. 6a for visualization purposes), focusing on the presence and direction of associations. Therefore, caution is required when interpreting the effect sizes.

Validation of rs671 GWAS-by-proxy associations using subset of maternal genotype data

The maternal genotyping using blood samples for the first and the second batches (*N*=80,964) was performed under conditions identical to those applied to the children. Genotype QC included exclusion of six samples with a call rate < 0.98. We then removed 11 samples with sex inconsistency, as all maternal samples were expected to be female. Next, we used KING[22] to verify parent-child relationships and identify duplicate samples, particularly because multiple siblings from the same family participated in the study, which resulted in some mothers appearing more than once in the dataset. We retained mothers genetically confirmed to be linked to children included in the GWAS, with duplicate maternal samples consolidated by preferentially keeping the sample with the higher call rate, leaving 63,257 final maternal samples for the rs671 association analysis. The analysis was performed using REGENIE in the same manner as described above for the children, except that the second step was performed only for rs671. Covariates included maternal age at the child's birth and the first eight PCs of maternal genotypes. As this analysis focused on a directly genotyped variant, we did not perform whole-genome imputation or genome-wide analyses in the maternal dataset.

Effect size comparison with MoBa's childhood BMI GWAS

We downloaded the childhood BMI GWAS results from the Norwegian Mother, Father and Child Cohort (MoBa) at eight different ages (at birth, 6 weeks, 6 months, 8 months, 1 year, 1.5 year, 2 years and 3 years old, sample size ranging from 16,902 to 28,681 children) from the NIPH website (**data availability**) and linked them with our GWAS age strata (0 month, 1 month, 6 months, 1 year (early), 1 year (late), 1.5 years, 2 years and 3 years old). We used the UCSC Liftover tool to map GRCh37 coordinates to GRCh38 coordinates of GWAS results from MoBa. Then, we extracted the effect size and its standard error for each index variant discovered in our GWAS of the specific age stratum. Of the 454 index variants in our GWAS, 208 matched the coordinates, reference, and alternative alleles. After checking the *P*-values, we found that only 18 variants passed the standard GWAS threshold of $P = 5 \times 10^{-8}$.

Longitudinal analysis using Gaussian processes

A Gaussian process (GP) regression model was used to map dynamic genetic associations along a child's age in months. The technique of inducing points in the sparse GP approach[81] was employed to handle the large number of data points

(*N*=678,704) from 79,307 children who had at least one BMI record from ages 0 to 4 (approximately 8 time points per participant). We included all the covariates used in our standard GWAS, as well as the location where the child's height and weight were measured (community health center, nursery, hospital or home). Model parameters were estimated to maximize the Titsias lower bound[82] of the model using the three stage approach[46]. During the parameter tuning, we found that the length-scale parameter of the quadratic exponential kernel used in the Gaussian process regression model was suboptimal. Because this parameter controls the smoothness and correlation structure of the longitudinal trajectory, its inappropriate specification led to overly optimistic association statistics and, consequently, an excess of significant findings. To address this issue, we re-estimated the model with improved calibration of the kernel length-scale parameter with an inverse-Gamma prior distribution.

The dynamic QTL effect was defined as any effect that exhibits a deviation from zero over the course of time. The null hypothesis is the genetic effect being equivalent to zero for the duration of the modeling period (in this case, from 0 to 54 months of age). The dynamic effect and its standard error were estimated as the posterior distribution of the Gaussian process modeling the QTL effect. The score statistics to capture the dynamic QTL effects were computed without estimating the variance parameters under the alternative hypotheses, which dramatically reduces the computational time. The momentchi2 package implemented on R was used to compute the *P*-value of a weighted sum of independent chi-square statistics (see the **Supplementary Notes 2** and ref[46] for further details). We identified, in total, 144 dynamic QTL loci that passed the genome-wide significance threshold. The FTO locus is a special case where the effect reverses signs within the modeling period; many other dynamic loci show an effect that either emerges or attenuates over time but does not change sign (effect size plots for all the associated variants are shown in the figures in **Supplementary Notes 2**).

The LDSC regression was performed using the score test *P*-values, from which *Z*-scores were computed. Here, the QTL effect beta was set to a constant value of 1.0 for the analysis, because it was essentially not a single value in the model; rather, it varied across each data point in the model as a Gaussian process. Here, the standard LDSC regression does not account for the sign or magnitude of beta in the analysis, and the approximation of the Z-score is valid because the sample size is large.

The simulated genotype data for one biallelic SNP locus were sampled 79,307 times independently from a binomial distribution Bin(2; *p*) with a probability parameter *p*. The probability parameter, indicating the minor allele frequency of a SNP locus, was also sampled from a uniform distribution U(0.001, 0.5). We generated, in total, genotype data of 17,454,650 SNP loci for testing under the null hypothesis. The number of loci is identical to the number of tests performed in the actual GWAS for BMI. We have also performed a permutation test using the real genotype data. In this analysis, we applied one identical sample-label permutation to a block of 5000 variants, then a different permutation to the next block to preserve LD structures.

LD pruning and dynamic polygenic score calculation

We computed the LD index between the 574 index variants discovered through the static and dynamic GWAS (11 age

strata and the dynamic QTL mapping) using our imputed genotype data. We then converted the 574-dimensional LD matrix R into the squared distance matrix $I - R^2$ to perform the hierarchical clustering. Setting the cutoff at 0.9 (*i.e.* LD pruning at $r^2 < 0.1$) allowed us to merge loci, resulting in 226 clusters. We selected the index variant with the minimum *P*-value to represent each cluster of loci. Finally, we estimated the dynamic QTL effects at the 226 selected index variants. Here, the posterior distribution of the dynamic QTL effect at each month of child age is given by the GP regression model. Then the individual's genotype was weighted by the effect size to obtain the polygenic score (PGS) along the child age in month. The prediction interval of the PGS was computed using the standard error of the QTL effects across all loci and the estimated residual variance. See the **Supplementary Notes 2** for further details. We have also developed an R package called *DynamicPGS* to predict the dynamic PGS given any genotype data, uploaded to our GitHub page (**Software availability**).

PGS replication in the Norwegian Mother, Father & Child Cohort study (MoBa)

The Norwegian Mother, Father and Child Cohort Study (MoBa) is a pregnancy-based ongoing cohort study consisting of approximately 114,500 children, 95,200 mothers and 75,000 fathers, enrolled from 50 hospitals across Norway from 1999 to 2008[83]. Children's height and weight values used to compute BMI in this study were measured by trained nurses during routine checkups at 6 weeks, at 3, 6 and 8 months, and at 1, 1.5, 2, 3, and 5 years of age and reported by parents in questionnaires. Genotyping was conducted from samples donated by participants[84] by multiple research groups on various genotyping platforms. Genotyping data are managed by the Norwegian Institute of Public Health (NIPH). Information on the different genotyping batches is available from the NIPH (github.com/folkehelseinstituttet/mobagen). This study was conducted on the quality controlled, imputed, and harmonized set of genotypes available from the NIPH version 2025.09.25 (github.com/fhi-beta/mobaGenetics-qc). Only participants clustering with the CEU samples from the HapMap project in a principal component analysis (PCA) were considered (n = 65,455 children). Only variants with a minor allele frequency >0.1% were included.

Out of 226 index variants identified by our data, 194 could be matched to the genotype set in MoBa. Out of the 194, 136 were the same index variant in JECS and 58 were proxy variants in high LD ($r^2 > 0.8$) based on our imputed genotype data (**Supplementary Table 8**). Using these variants, the dynamic PGS was used to predict the BMI of children in MoBa at the age strata corresponding to measurements. The variance explained in measured BMI by the PGS was evaluated using a linear model. The ethics, acknowledgements, and data availability for MoBa are available as **Supplementary Notes 3**.

LD score regression

The python package of LD Score regression (LDSC) was downloaded from github (**Code availability**). The pre-computed LD Scores in an East Asian population (v2.2) was downloaded from Zenodo (**Data availability**). The additive SNP heritability (h^2) was estimated for each of 1,255 traits using the baseline LD scores (the heritability estimates are provided in **Supplementary Table 1**). For case-control traits, we computed the effective sample size (N_{eff}) following the standard LD score regression (LDSC) framework, using $N_{eff} = 4/((1/N_{case}) + (1/N_{control}))$, and applied these values

consistently in all LDSC analyses. We now provide the additive SNP heritability, its standard error, *Z*-score and N_{eff} value for each trait in the **Supplementary Table 1**.

For the cross-trait LDSC analysis between pairs of child and parent traits to compute genetic correlations (r_g), those traits with h^2 greater than 0.0 and one or more associations in GWAS were used. The environmental correlation (r_e), which is the residual correlation after subtracting genetic correlation from the phenotypic correlation, was computed as follows

$$r_e = \frac{\rho - \sqrt{h^2{}_1}\sqrt{h^2{}_2}r_g}{\sqrt{1-h^2{}_1}\sqrt{1-h^2{}_2}}$$

where ρ indicates the total covariance, $h^2{}_1$ and $h^2{}_2$ indicate the additive SNP heritability estimates for the first and the second traits, respectively. Note that the genetic covariance is computed as $\sqrt{h^2{}_1}\sqrt{h^2{}_2}r_g$. We provide the full detail of the derivation in the Appendix section 2.3 of the **Supplementary Notes 2**. The computed genetic correlations and environmental correlations for all the analyzed traits are available in **Supplementary Table 4**.

scDRS

scDRS[85] analysis was performed to assess cell type-specific enrichment of genetic risk across PBMC cell types from recent single-cell multi-omics data[86]. Gene-level association statistics were first obtained from GWAS summary statistics using MAGMA[87] with the 1000 Genomes Phase 3 East Asian reference panel and were then used to construct scDRS input gene sets, as described previously[85]. Cell type-level enrichment was evaluated by the group-level downstream analysis.

Data availability

The manifest file of the Illumina Asian Screening Array (ASA-24v1-0_E2.csv) was available from the product web page (https://support.illumina.com/downloads/infinium-asian-screening-array-v1-0-product-files.html). The allele frequency data of Tohoku Medical Megabank (ToMMo) samples was obtained from jMorp (https://jmorp.megabank.tohoku.ac.jp/). The latest GWAS results were downloaded from the GWAS Catalog (https://www.ebi.ac.uk/gwas/api/search/downloads/gwas_catalog_v1.0-associations_e113_r2024-11-03.tsv). Two imputation reference panels were obtained from the 1000 Genomes Project (ALL.chr*.shapeit2_integrated_snvindels_v2a_27022019.GRCh38.phased.vcf.gz and 1kGP_high_coverage_Illumina.chr*.filtered.SNV_INDEL_SV_phased_panel.vcf.gz). The variant VCF of a South Korean population was obtained from the Korean Personal Genome Project (KPGP) (http://kpgp.kr/). The pre-computed LD scores using the 1000 Genomes Project East Asian Population (1000G_Phase3_EAS_baselineLD_v2.2_ldscores.tgz) was downloaded from ZENODO (https://zenodo.org/records/10515792). GTEx eQTL summary statistics of the skeletal muscle tissue (V8) was downloaded from the eQTL Catalogue

(ftp.ebi.ac.uk/pub/databases/spot/eQTL/sumstats/QTS000015/QTD000121/QTD000121.all.tsv.gz). Ensemble gene annotation (version 113) was downloaded from BioMart (https://www.ensembl.org/). GWAS results of childhood BMI from MoBa were downloaded from the NIPH website (https://www.fhi.no/en/ch/studies/moba/for-forskere-artikler/gwas-data-from-moba/). Raw genotype data for child participants in JECS have not been uploaded to a publicly accessible repository due to parental consent, ethical restrictions, and legal framework of the Act on the Protection of Personal Information (Act No. 57 of 30 May 2003, amendment on 1 June 2025). Access to the raw data can be granted for scientific purposes only at the JECS facilities located in Japan. All inquiries about data access should be sent to: jecs-en@nies.go.jp. The person responsible for handling enquiries sent to this e-mail address is Dr Shoji F. Nakayama, JECS Programme Office, National Institute for Environmental Studies. The GWAS summary statistics for all 1,255 traits, are available via the JECS GWAS summary statistics request form (https://forms.office.com/r/qipvY8nNTK).

Code availability

The fixref plugin of bcftools to align strand orientations of SNP probes is available online (https://samtools.github.io/bcftools/howtos/plugin.fixref.html). BLAST+ was downloaded from the NCBI FTP site (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/). KING (version 2.3.2) was used to estimate kinship between the genotyped samples in JECS (https://www.kingrelatedness.com/). Beagle (beagle.22Jul22.46e.jar) was used for imputation of the South Korean samples and Minimac4 (v4.1.6) was used for imputation of the genotyped samples in JECS (https://faculty.washington.edu/browning/beagle/beagle.html). ADMIXTURE (v1.3.0) was used to estimate mixed ancestries in our participants (https://dalexander.github.io/admixture/index.html). EAGLE software (v2.4.1) was used to pre-phase the genotype data prior to the imputation (https://alkesgroup.broadinstitute.org/Eagle/). SHAPEIT5 was used to perform haplotype phasing prior to call chromosomal mosaicisms. (https://github.com/odelaneau/shapeit5). MoChA (version 2024-09-27) was used to detect mosaic chromosomal alterations (https://github.com/freeseek/mocha/). Plink2 (v2.00a5.10LM 64-bit Intel) was used to perform PCA and genotype data preparations. REGENIE (ver 3.4.1) was used to perform GWAS (https://rgcgithub.github.io/regenie/). ANNOVAR (2020-06-07) was obtained online (https://annovar.openbioinformatics.org/en/latest/). LDSC (v1.0.1) was used to estimate the trait (additive and SNP-based) heritability and genetic correlations between traits (https://github.com/bulik/ldsc). DynamicPGS was developed and uploaded on the GitHub page (https://github.com/natsuhiko/DynamicPGS/tree/main). DEEP*HLA was used to perform HLA imputation (https://github.com/tatsuhikonaito/DEEP-HLA)

References

1. Manolio, T. A., Bailey-Wilson, J. E. & Collins, F. S. Genes, environment and the value of prospective cohort studies. *Nat. Rev. Genet.* **7**, 812–820 (2006).
2. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
3. Helgeland, Ø. *et al.* Genome-wide association study reveals dynamic role of genetic variation in infant and early

868 childhood growth. *Nat. Commun.* **10**, 4448 (2019).
869 4. Helgeland, Ø. *et al.* Characterization of the genetic architecture of infant and early childhood body mass index.
870 *Nat. Metab.* **4**, 344–358 (2022).
871 5. Westergaard, D. *et al.* Genome-wide association meta-analysis identifies five loci associated with postpartum
872 hemorrhage. *Nat. Genet.* **56**, 1597–1603 (2024).
873 6. Kawamoto, T. *et al.* Rationale and study design of the Japan environment and children's study (JECS). *BMC Public*
874 *Health* **14**, 25 (2014).
875 7. Jarup, L. Hazards of heavy metal contamination. *British Medical Bulletin* **68**, 167–182 (2003).
876 8. Nakayama, S. F. *et al.* Blood mercury, lead, cadmium, manganese and selenium levels in pregnant women and
877 their determinants: the Japan Environment and Children's Study (JECS). *J. Expo. Sci. Environ. Epidemiol.* **29**, 633–
878 647 (2019).
879 9. Brunn, H. *et al.* PFAS: forever chemicals—persistent, bioaccumulative and mobile. Reviewing the status and the
880 need for their phase out and remediation of contaminated sites. *Environ. Sci. Eur.* **35**, (2023).
881 10. Hasegawa, K. *et al.* Maternal exposure to per- and polyfluoroalkyl substances and offspring chromosomal
882 abnormalities: The japan environment and children's study. *Environ. Health Perspect.* **132**, (2024).
883 11. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**,
884 D977–D985 (2023).
885 12. Jung, S. *et al.* Case-case genome-wide association analysis identifying genetic loci with divergent effects on
886 Crohn's disease and ulcerative colitis. *Hum. Mol. Genet.* **32**, 677–684 (2023).
887 13. Inouye, M. *et al.* Genomic risk prediction of coronary artery disease in 480,000 adults: Implications for primary
888 prevention. *J. Am. Coll. Cardiol.* **72**, 1883–1893 (2018).
889 14. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381
890 (2014).
891 15. Fitipaldi, H. & Franks, P. W. Ethnic, gender and other sociodemographic biases in genome-wide association
892 studies for the most burdensome non-communicable diseases: 2005-2022. *Hum. Mol. Genet.* **32**, 520–532
893 (2023).
894 16. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–
895 712 (2022).
896 17. Mahajan, A. *et al.* Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for
897 discovery and translation. *Nat. Genet.* **54**, 560–572 (2022).
898 18. Michikawa, T. *et al.* Baseline profile of participants in the Japan Environment and Children's Study (JECS). *J.*
899 *Epidemiol.* **28**, 99–104 (2018).
900 19. García-Acero, M., Moreno, O., Suárez, F. & Rojas, A. Disorders of sexual development: Current status and
901 progress in the diagnostic approach. *Curr. Urol.* **13**, 169–178 (2020).
902 20. Loh, P.-R. *et al.* Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**,
903 350–355 (2018).
904 21. Loh, P.-R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inheritance become instruments for clonal

905 selection. *Nature* **584**, 136–141 (2020).
906 22. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**,
907 2867–2873 (2010).
908 23. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies.
909 *Nat. Genet.* **38**, 904–909 (2006).
910 24. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals.
911 *Genome Res.* **19**, 1655–1664 (2009).
912 25. Yamaguchi-Kabata, Y. *et al.* Japanese population structure, based on SNP genotypes from 7003 individuals
913 compared to other ethnic groups: effects on population-based association studies. *Am. J. Hum. Genet.* **83**, 445–
914 456 (2008).
915 26. Yamamoto, K. *et al.* Genetic legacy of ancient hunter-gatherer Jomon in Japanese populations. *Nat. Commun.*
916 **15**, 9780 (2024).
917 27. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat.*
918 *Genet.* **53**, 1097–1103 (2021).
919 28. Liu, J. Z., Erlich, Y. & Pickrell, J. K. Case-control association mapping by proxy using family history of disease. *Nat.*
920 *Genet.* **49**, 325–331 (2017).
921 29. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**,
922 1415–1424 (2021).
923 30. Kumasaka, N., Knights, A. J. & Gaffney, D. J. High-resolution genetic mapping of putative causal interactions
924 between regions of open chromatin. *Nat. Genet.* **51**, 128–137 (2019).
925 31. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–
926 1241 (2015).
927 32. Squires, J., Twombly, E., Bricker, D. D. & Potter, L. ASQ-3 user's guide. (2009).
928 33. Mezawa, H. *et al.* Psychometric profile of the Ages and Stages Questionnaires, Japanese translation. *Pediatr. Int.*
929 **61**, 1086–1095 (2019).
930 34. Spergel, J. M. & Paller, A. S. Atopic dermatitis and the atopic march. *J. Allergy Clin. Immunol.* **112**, S118–27
931 (2003).
932 35. Weidinger, S. *et al.* Filaggrin mutations, atopic eczema, hay fever, and asthma in children. *J. Allergy Clin.*
933 *Immunol.* **121**, 1203–1209.e1 (2008).
934 36. Astolfi, A. *et al.* Filaggrin loss-of-function mutations are risk factors for severe food allergy in children with atopic
935 dermatitis. *J. Clin. Med.* **10**, 233 (2021).
936 37. Sandilands, A. *et al.* Comprehensive analysis of the gene encoding filaggrin uncovers prevalent and rare
937 mutations in ichthyosis vulgaris and atopic eczema. *Nat. Genet.* **39**, 650–654 (2007).
938 38. Akiyama, M. FLG mutations in ichthyosis vulgaris and atopic eczema: spectrum of mutations and population
939 genetics. *Br. J. Dermatol.* **162**, 472–477 (2010).
940 39. Kono, M. *et al.* Comprehensive screening for a complete set of Japanese-population-specific filaggrin gene
941 mutations. *Allergy* **69**, 537–540 (2014).

942 40. Marenholz, I. *et al.* Genome-wide association study identifies the SERPINB gene cluster as a susceptibility locus
943 for food allergy. *Nat. Commun.* **8**, 1056 (2017).
944 41. Kubo, A. *et al.* Mutations in SERPINB7, encoding a member of the serine protease inhibitor superfamily, cause
945 Nagashima-type palmoplantar keratosis. *Am. J. Hum. Genet.* **93**, 945–956 (2013).
946 42. Arnau-Soler, A. *et al.* Food allergy genetics and epigenetics: A review of genome-wide association studies. *Allergy*
947 **80**, 106–131 (2025).
948 43. Maier, L. *et al.* Meta-analysis of genome-wide association studies of food allergy and IgE sensitization. *J. Allergy*
949 *Clin. Immunol.* (2026) doi:10.1016/j.jaci.2026.02.012.
950 44. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary
951 statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
952 45. Skene, N. G. *et al.* Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825–833
953 (2018).
954 46. Kumasaka, N. Genetic association mapping leveraging Gaussian processes. *J. Hum. Genet.* **69**, 505–510 (2024).
955 47. Ko, S. *et al.* GWAS of longitudinal trajectories at biobank scale. *Am. J. Hum. Genet.* **109**, 433–445 (2022).
956 48. Sikorska, K., Lesaffre, E., Groenen, P. J. F., Rivadeneira, F. & Eilers, P. H. C. Genome-wide Analysis of Large-scale
957 Longitudinal Outcomes using Penalization -GALLOP algorithm. *Sci. Rep.* **8**, 6815 (2018).
958 49. Burrows, K. *et al.* A framework for conducting GWAS using repeated measures data with an application to
959 childhood BMI. *Nat. Commun.* **15**, 10067 (2024).
960 50. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.* **373**,
961 895–907 (2015).
962 51. Sovio, U. *et al.* Association between common variation at the FTO locus and changes in body mass index from
963 infancy to late childhood: the complex nature of genetic association through growth and development. *PLoS*
964 *Genet.* **7**, e1001307 (2011).
965 52. Sah, J. P. *et al.* Ectonucleotide pyrophosphatase 2 (ENPP2) plays a crucial role in myogenic differentiation
966 through the regulation by WNT/β-Catenin signaling. *Int. J. Biochem. Cell Biol.* **118**, 105661 (2020).
967 53. Children's environmental health. *World Health Organization* https://www.who.int/health-topics/children-
968 environmental-health.
969 54. Iwai-Shimada, M. *et al.* Questionnaire results on exposure characteristics of pregnant women participating in the
970 Japan Environment and Children Study (JECS). *Environ. Health Prev. Med.* **23**, 45 (2018).
971 55. Ng, E. *et al.* Genome-wide association study of toxic metals and trace elements reveals novel associations. *Hum.*
972 *Mol. Genet.* **24**, 4739–4745 (2015).
973 56. Eom, S.-Y. *et al.* Exome-wide association study identifies genetic polymorphisms of C12orf51, MYL2, and ALDH2
974 associated with blood lead levels in the general Korean population. *Environ. Health* **16**, 11 (2017).
975 57. Matoba, N. *et al.* GWAS of 165,084 Japanese individuals identified nine loci associated with dietary habits.
976 *Nature Human Behaviour* (2020) doi:10.1038/s41562-019-0805-1.
977 58. Nakayama, S. F. *et al.* Poly- and perfluoroalkyl substances in maternal serum: Method development and
978 application in Pilot Study of the Japan Environment and Children's Study. *J. Chromatogr. A* **1618**, 460933 (2020).

59. Dunn, J., Clarkson, T. & Magos, L. Ethanol-increased exhalation of mercury in mice. *Br. J. Ind. Med.* **35**, 241–244 (1978).
60. Sällsten, G., Kreku, S. & Unosson, H. A small dose of ethanol increases the exhalation of mercury in low-level-exposed humans. *J. Toxicol. Environ. Health Part A.* **60**, 89–100 (2000).
61. Wu, Y. *et al.* Pervasive biases in proxy genome-wide association studies based on parental history of Alzheimer's disease. *Nat. Genet.* 1–8 (2024).
62. Pedersen, E. M., Wimberley, T. & Vilhjálmsson, B. J. A cautionary tale for Alzheimer's disease GWAS by proxy. *Nat. Genet.* **56**, 2590–2591 (2024).
63. Zhou, L. *et al.* Characteristics and mechanisms of mosaicism in prenatal diagnosis cases by application of SNP array. *Mol. Cytogenet.* **16**, 13 (2023).
64. Watson, C. J. & Blundell, J. R. Mutation rates and fitness consequences of mosaic chromosomal alterations in blood. *Nat. Genet.* **55**, 1677–1685 (2023).
65. Snyder, E. A. *et al.* Genetic counseling for women with 45,X/46,XX mosaicism: Towards more personalized management. *Eur. J. Med. Genet.* **64**, 104140 (2021).
66. Li, S. *et al.* Prenatal diagnosis of chromosomal mosaicism in over 18,000 pregnancies: A five-year single-tertiary-center retrospective analysis. *Front. Genet.* **13**, 876887 (2022).
67. Zhou, W. *et al.* Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nat. Genet.* **48**, 563–568 (2016).
68. Jakubek, Y. A. *et al.* Mosaic chromosomal alterations in blood across ancestries using whole-genome sequencing. *Nat. Genet.* **55**, 1912–1919 (2023).
69. Sun, M. *et al.* Impact of clonal hematopoiesis on cardiovascular outcomes in cancer patients of the UK Biobank. *ESMO Open* **10**, 105539 (2025).
70. Visscher, P. M. *et al.* 10 years of GWAS discovery: Biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
71. Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527–532 (2021).
72. Schloissnig, S. *et al.* Structural variation in 1,019 diverse humans based on long-read sequencing. *Nature* **644**, 442–452 (2025).
73. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the 'Sum of Single Effects' model. *PLoS Genet.* **18**, e1010299 (2022).
74. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
75. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
76. Jones, R. E. & Lopez, K. H. Gamete Transport and Fertilization. in *Human Reproductive Biology* 159–173 (Elsevier, 2014).
77. Sato, G. *et al.* Pan-cancer and cross-population genome-wide association studies dissect shared genetic

backgrounds underlying carcinogenesis. *Nat. Commun.* **14**, 3671 (2023).
78. Sonehara, K. *et al.* Germline variants and mosaic chromosomal alterations affect COVID-19 vaccine immunogenicity. *Cell Genom.* **5**, 100783 (2025).
79. Hirata, J. *et al.* Genetic and phenotypic landscape of the major histocompatibilty complex region in the Japanese population. *Nat. Genet.* **51**, 470–480 (2019).
80. Naito, T. *et al.* A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes. *Nat. Commun.* **12**, 1639 (2021).
81. Hensman, J., Fusi, N. & Lawrence, N. D. Gaussian Processes for Big Data. *arXiv [cs.LG]* (2013).
82. Titsias, M. Variational learning of inducing variables in sparse Gaussian processes. *Artificial Intelligence and Statistics* (2009).
83. Brandlistuen, R. E. *et al.* Cohort profile update: The Norwegian mother, father and child cohort (MoBa). *Int. J. Epidemiol.* **54**, dyaf139 (2025).
84. Paltiel, L. *et al.* The biobank of the Norwegian Mother and Child Cohort Study – present status. *Nor. Epidemiol.* **24**, (2014).
85. Zhang, M. J. *et al.* Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *Nat. Genet.* **54**, 1572–1580 (2022).
86. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
87. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).

Acknowledgement

We thank all the JECS participants. We would like to also thank all staff members of the JECS and the informatics group at National Center for Child Health and Development. We thank the members of the Kawasaki disease medical record transcription board (Drs. Sayaka Fukuda and Tohru Kobayashi). We thank Kazuyuki Nakazono, Naoko Miyagawa and Satoshi Nagashima for their contribution to the GWAS pipeline construction. This study was funded by the Ministry of the Environment, Japan. The findings and conclusions of this article are solely the responsibility of the authors and do not represent the official views of the above government.

Author information

Authors

Natsuhiko Kumasaka[1,2,*], Hidetoshi Mezawa[1], Tatsuhiko Naito[1], Kei Sakamoto[1], Kazuhisa Akiba[1], Atsushi Hattori[1], Tomohiko Isobe[3], Miori Sato[1], Mayako Saito-Abe[1], Aya Yoshikawa[1], Limin Yang[1], Daisuke Harama[1], Yumiko Miyaji[1], Minaho Nishizato[1], Makiko Sekiyama[3], Miyuki Iwai-Shimada[3], Yayoi Kobayashi[3], Mai Takagi[3], Nozomi Tatsuta[3], Yu Taniguchi[3], Kenji Matsumoto[4], Kazuhiko Nakabayashi[5], Masayo Kagami[6], Emiko Noguchi[7], Akira Hata[8], Kenichiro Hata[9],

Atsushi Tajima[10], Shi Mingyang[11], Masaru Koido[11], Yoichiro Kamatani[11], Go Sato[12], Koichi Matsuda[11], BioBank Japan, Yosuke Kawai[13], Katsushi Tokunaga[13], National Center Biobank Network, Mikko Baur[14], Nicolas Fragoso-Bargas[14], Marc Vaudel[14], Stefan Johansson[14,15], Reiko Kishi[16], Chiharu Ota[17], Koichi Hashimoto[18], Chisato Mori[19], Shuichi Ito[20], Ryoji Shinohara[21], Hidekuni Inadera[22], Takeo Nakayama[23], Ryo Kawasaki[24], Yasuhiro Takeshima[25], Seiji Kageyama[26], Akemi Morita[26], Narufumi Suganuma[27], Shouichi Ohga[28], Takahiko Katoh[29], Michihiro Kamijima[30], Kiwako Yamamoto-Hanada[1], Akihiro Umezawa[31], Maki Fukami[1], Shin Yamazaki[3], Yukinori Okada[12], Shoji F. Nakayama[3,*], Yukihiro Ohya[30]

Affiliations

1. Medical Support Center for the Japan Environment and Children's Study, National Center for Child Health and Development, Tokyo, Japan

2. Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan

3. Programme Office for the Japan Environment and Children's Study, National Institute for Environmental Studies, Tsukuba, Japan

4. Department of Allergy and Clinical Immunology, National Center for Child Health and Development Research Institute, Tokyo, Japan

5. Department of Maternal-Fetal Biology, National Center for Child Health and Development Research Institute, Tokyo, Japan

6. Department of Molecular Endocrinology, National Center for Child Health and Development Research Institute, Tokyo, Japan

7. Institute of Medicine, University of Tsukuba, Tsukuba, Japan

8. Graduate School of Medicine, Chiba University, Chiba, Japan / Chiba Foundation for Health Promotion & Disease Prevention, Chiba, Japan

9. Graduate School of Medicine, Gunma University, Maebashi, Japan

10. Graduate School of Advanced Preventive Medical Sciences, Kanazawa University, Kanazawa, Japan

11. Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan

12. Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

13. Genome Medical Science Project, National Center for Global Health and Medicine Research Institute, Tokyo, Japan

14. Mohn Center for Diabetes Precision Medicine, Department of Clinical Science, University of Bergen, Bergen, Norway.

15. Department of Pediatrics and Adolescent Medicine, Haukeland University Hospital, Bergen, Norway

16. Center for Environmental and Health Sciences, Hokkaido University, Sapporo, Japan

17. Graduate School of Medicine, Tohoku University, Sendai, Japan

18. School of Medicine, Fukushima Medical University, Fukushima, Japan

.081 19. School of Medicine, Chiba University, Chiba, Japan

.082 20. School of Medicine, Yokohama City University, Yokohama, Japan

.083 21. Center for Birth Cohort Studies, University of Yamanashi, Chuo, Japan

.084 22. School of Medicine, University of Toyama, Toyama, Japan

.085 23. Graduate School of Medicine, Kyoto University, Kyoto, Japan

.086 24. Graduate School of Medicine, University of Osaka, Suita, Japan

.087 25. Pediatrics, Hyogo Medical University, Nishinomiya, Japan

.088 26. Faculty of Medicine, Tottori University, Yonago, Japan

.089 27. Kochi Medical School, Kochi University, Nankoku, Japan

.090 28. Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan

.091 29. Faculty of Life Sciences, Kumamoto University, Kumamoto, Japan

.092 30. Graduate School of Medical Sciences, Nagoya City University, Nagoya, Japan

.093 31. Medical Research Center for Japan Environment and Children's Study, National Center for Child Health and
.094 Development Research Institute, Tokyo, Japan

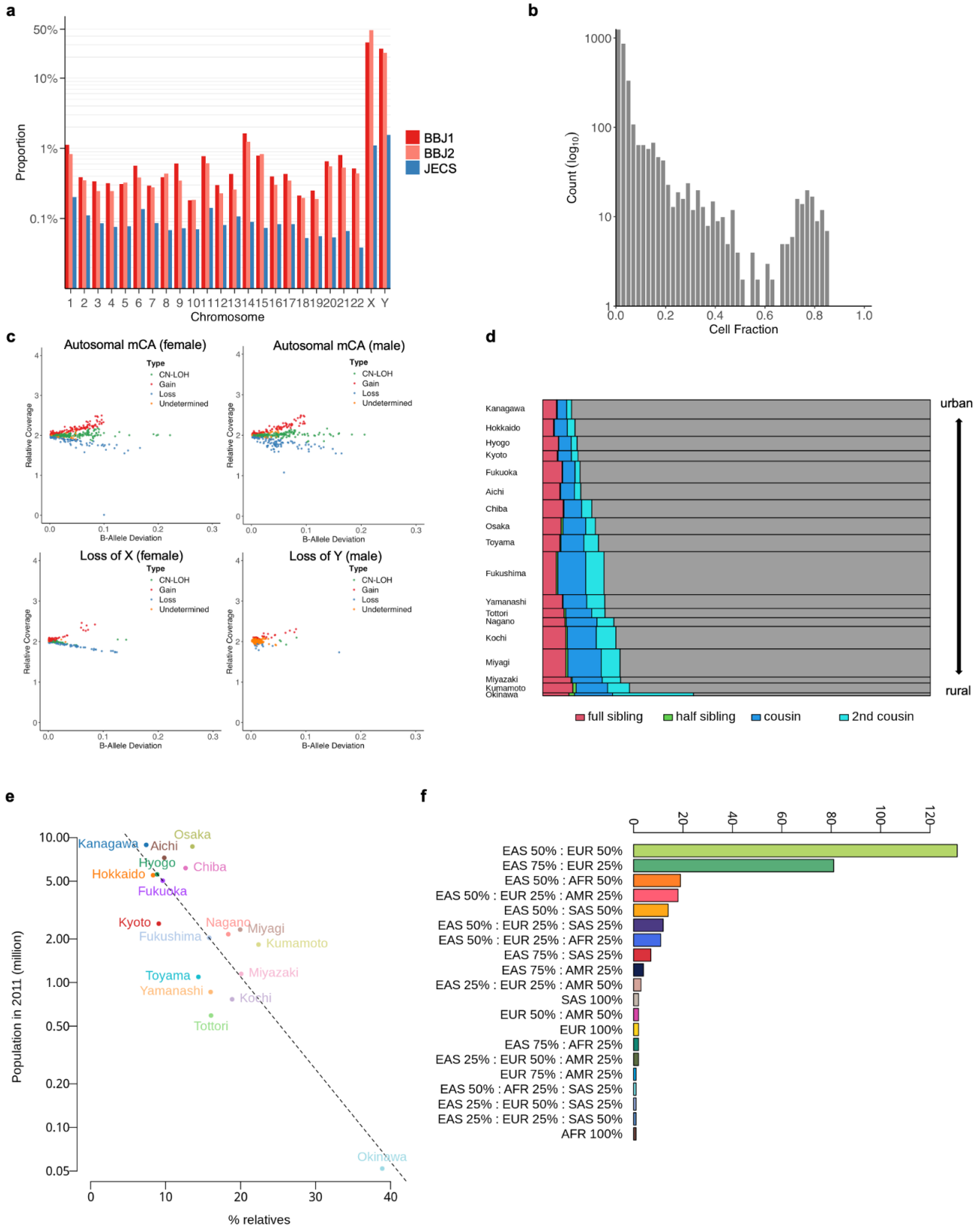
.095

.096 ## Corresponding authors

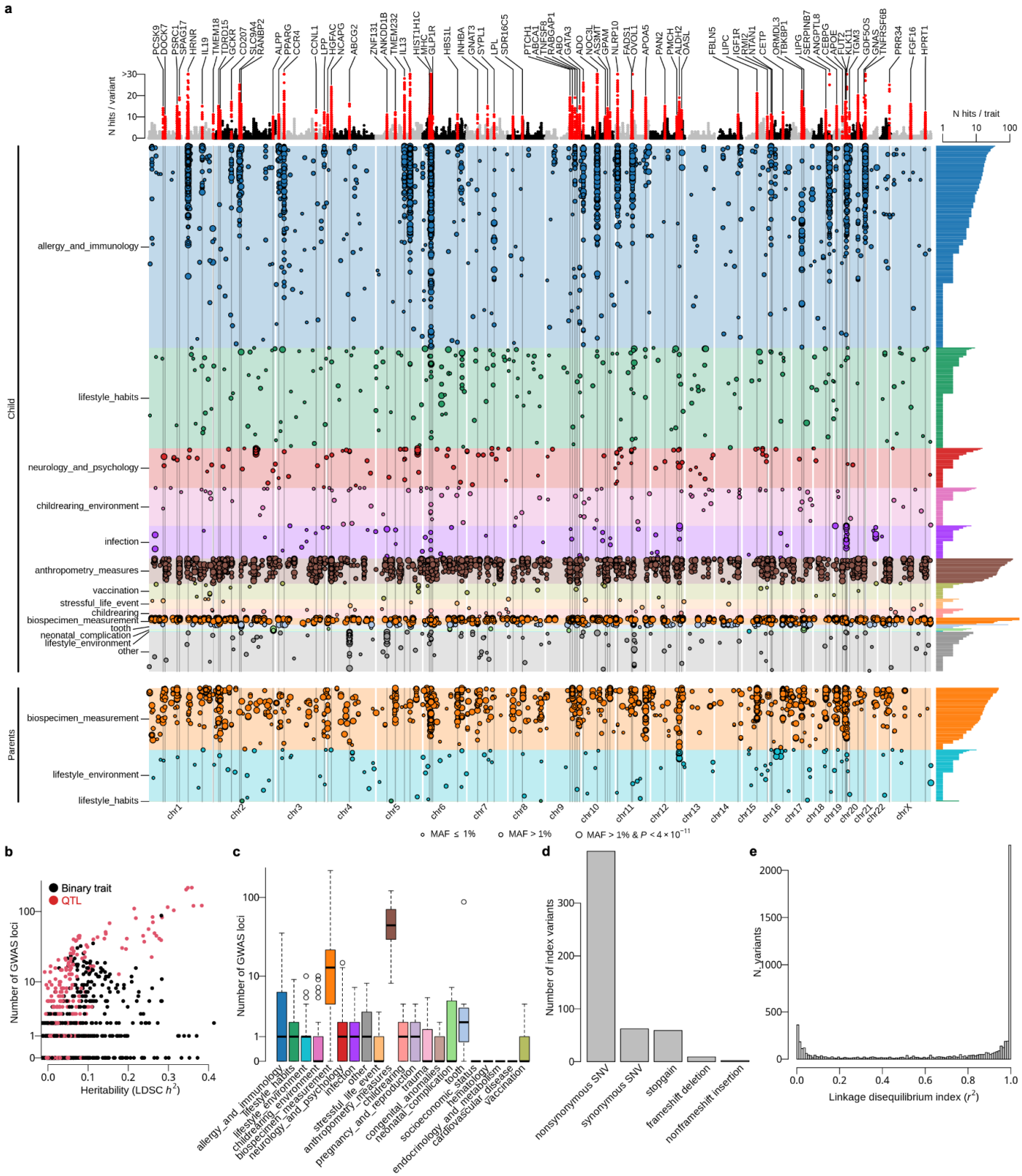
.097 Correspondence to Natsuhiko Kumasaka (kumasaka@hgc.jp) or Shoji F. Nakayama (fabre@nies.go.jp).

.098

Extended Data Figures

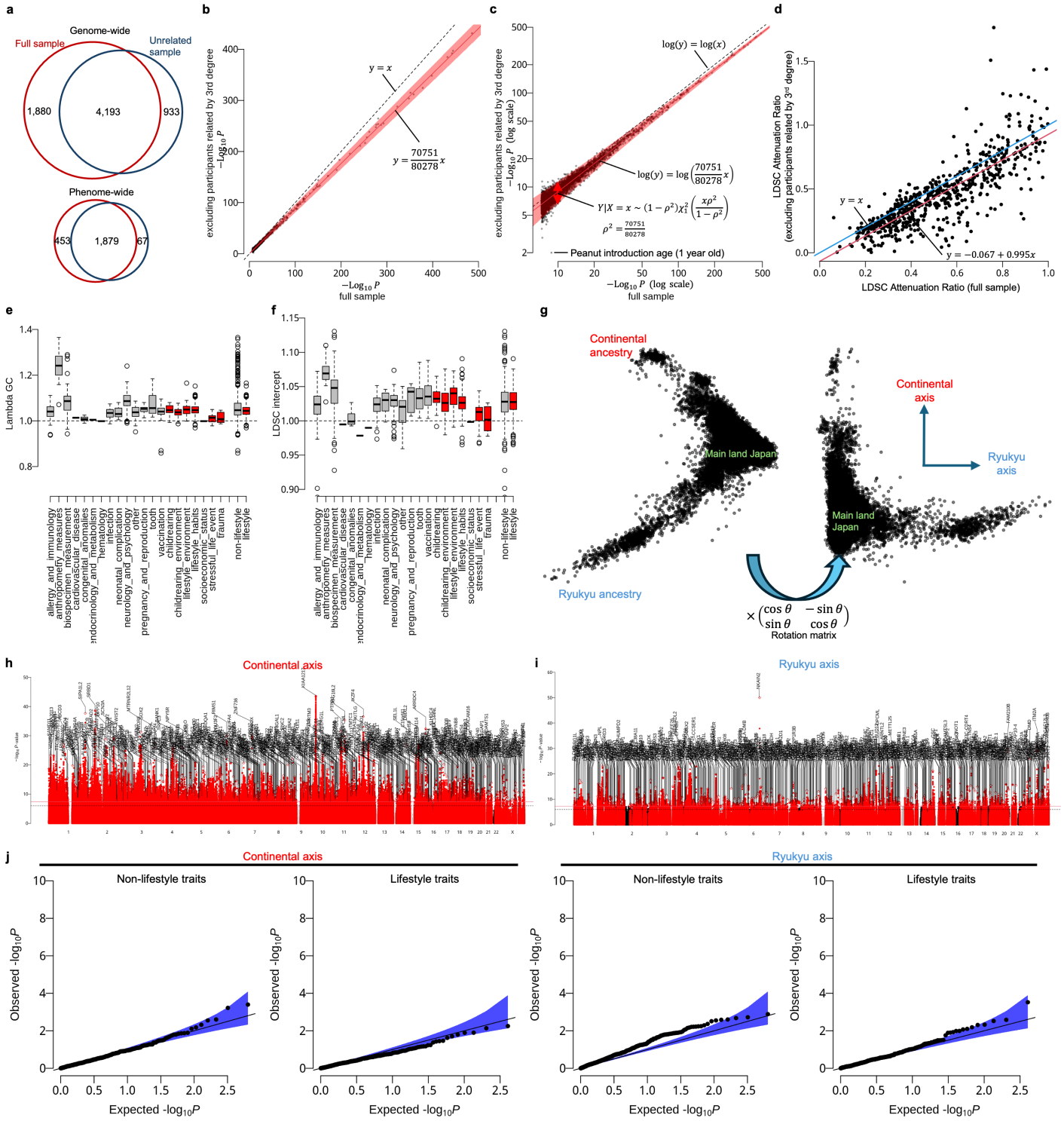


Extended Data Figure 1. a. Barplot shows the proportion of mosaic chromosomal alterations (mCAs) detected in each chromosome across the BBJ1 (red), BBJ2 (orange) cohorts, and JECS (blue). The y-axis is shown on a logarithmic scale. **b.** Histogram shows the distribution of cell fractions among all detected mCA events. The y-axis is plotted on a logarithmic scale. **c.** Scatter plots show the distribution of mosaic chromosomal alterations by B-allele deviation (x-axis) and relative coverage (y-axis) in autosomes and sex chromosomes. Each dot represents a single event, color-coded by event type: CN-LOH (green), Gain (red), Loss (blue), and Undetermined (orange). (Top left) Autosomal mCAs in females. (Top right) Autosomal mCAs in males. (Bottom left) Loss of X in females. (Bottom right) Loss of Y in males. **d.** Mosaicplot shows the percentage of relatives in our data, stratified by the prefectures of Japan where participants were enrolled. **e.** Scatterplot shows the relationship between the percentage of relatives (full/half siblings, cousins and second cousins) and the population (million people) of each prefecture in 2011 obtained from e-Stat in Japan (https://www.e-stat.go.jp/). Note that the population of Okinawa (bottom right point) reflects the number of people in Miyakojima City (but not the total population of Okinawa Prefecture), where all pregnant mothers in the Okinawa area were enrolled. The dotted line shows a linear regression line with an adjusted R-square value of 0.72 and a P-value of 4.7×10^{-6}. Even if we remove the data from Okinawa, the R-squared value is 0.43 with a P-value of 0.0026. **f**. Barplot shows the number of non-East Asian participants with different ancestral backgrounds inferred from the principal component analysis.



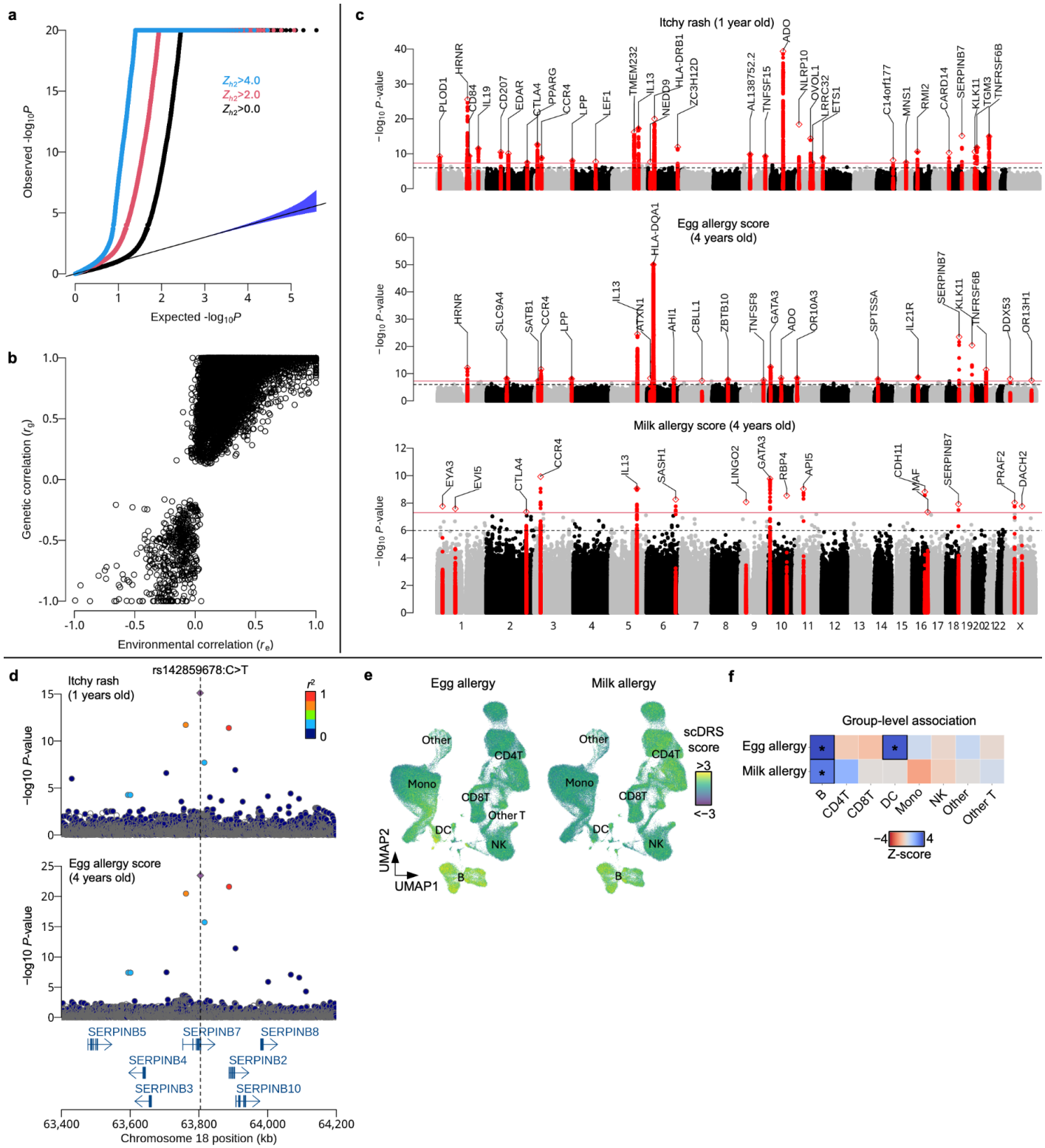
Extended Data Figure 2. a. Genetic associations discovered through GWAS of 1,255 traits in JECS. The top panel is a Manhattan-like plot showing the number of GWAS hits ($P < 5.0 \times 10^{-8}$) at each variant with the nearest gene annotation. Peaks are defined by variants shared across 10 or more traits. The main panel shows locations of index variants for each trait segregated by the trait category. The size of each dot indicates the

magnitude of the association as well as the minor allele frequency (MAF) of the respective index variant. The largest dot indicates an index variant that passed the phenome-wide threshold of $P < 4.0 \times 10^{-11}$, regardless of its MAF. If an index variant does not pass the phenome-wide threshold, it is classified by its MAF greater than 1% (mid-size dot) or less than 1% (smallest dot). The right panel shows a barplot of the number of GWAS hits for each trait. **b.** Scatterplot shows the number of GWAS loci for each trait as a function of the additive SNP heritability estimation obtained by the LD Score regression (LDSC). **c.** Boxplot shows the number of significant GWAS loci ($P < 5.0 \times 10^{-8}$) for each trait stratified by the trait categories. **d.** The number of index variants detected within the coding regions, stratified by the putative functions of variants annotated by the Ensembl gene (version 113; GRCh38). **e.** The histogram shows the distribution of linkage disequilibrium index (r^2) between the GWAS index variants in our study and the entire variants reported in the GWAS catalog. We took the maximum r^2 value for each of our index variants.



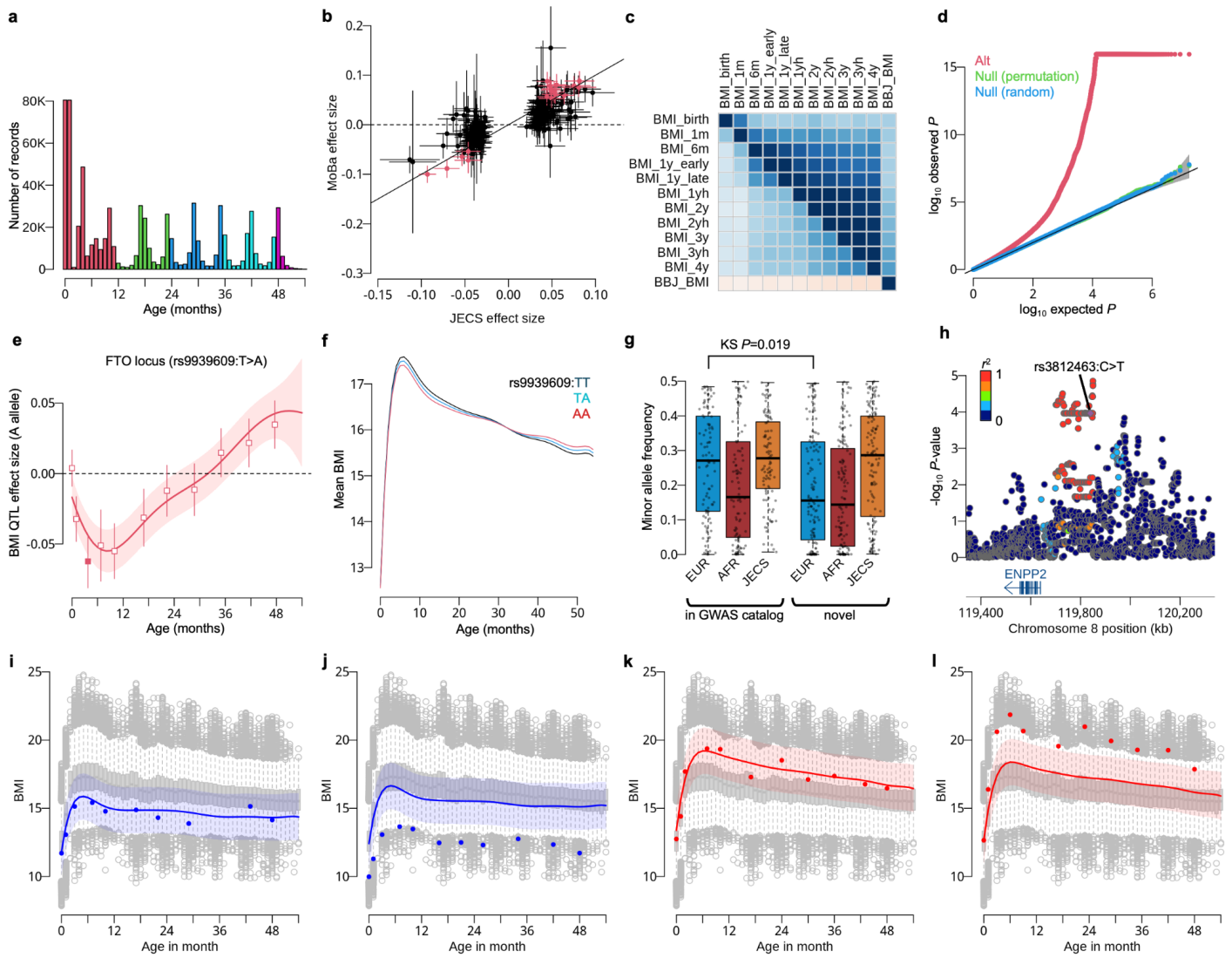
Extended Data Figure 3. a. The number of GWAS loci shared by the full sample (*N*=80,278) and the unrelated sample (*N*=70,751), excluding participants related by third degree. The top Venn diagram showed loci passed the standard GWAS threshold ($P < 5.0 \times 10^{-8}$) and the bottom Venn diagram showed loci passed the phenome-wide threshold ($P < 4.0 \times 10^{-11}$). **b.** Scatter plot of P-values for the 6,073 associations identified in the full-sample GWAS compared with those from the GWAS of unrelated individuals. Association signals were evaluated at the index variants identified in the full-sample GWAS. The dashed line indicates the diagonal (*y* = *x*), and the red line represents the expected reduction in statistical

power when subsampling 70,751 individuals from the full sample of 80,278. The expected slope is given by 70,751/80,278, corresponding to the conditional mean of a non-central chi-square distribution. The red shaded area denotes the 95% confidence interval, also derived from the non-central chi-square distribution. See Methods for details. **c.** Same as in panel b, but with both axes shown on a logarithmic scale. The dashed line, red line, and shaded area are identical to those in panel b. **d.** Scatter plot of LDSC attenuation ratios for traits with heritability Z-score > 0 and at least one GWAS hit. The x-axis shows attenuation ratios from the full-sample GWAS, and the y-axis shows those from the GWAS of unrelated individuals. The blue line indicates the diagonal ($y = x$), and the red line represents the least-squares fit ($y = a + bx$). The negative intercept ($a = -0.067$) reflects a modest reduction in attenuation ratio after excluding individuals related up to the third degree. **e.** Boxplot shows the distribution of genomic control lambda (λ_{GC}) for the 1,255 traits analyzed, stratified by trait categories (red boxes indicate the lifestyle trait categories). The last two boxes compare the distribution of aggregated non-lifestyle and lifestyle traits, in which the lifestyle traits combine *childrearing*, *childrearing environment*, *lifestyle environment*, *lifestyle habits*, *socioeconomic status*, *stressful life events*, and *trauma*. **f.** Boxplot shows the distribution of the LDSC intercept for the 1,255 traits analyzed, stratified by trait categories. **g.** The 2-dimensional rotation approach to extract the continental and Ryukyu axes. **h.** Manhattan plot shows GWAS of the continental axis extracted in the panel f. The continental axis is analyzed as a QTL trait. **i.** Manhattan plot shows GWAS of the Ryukyu axis extracted in the panel f. The Ryukyu axis is analyzed as a QTL trait. **j.** QQ plots of cross-trait LD score regression (LDSC) *P*-values between ancestry axes and all analyzed traits, stratified into lifestyle and non-lifestyle categories. No traits passed the trait-wide significance threshold ($P < 0.05/1255$).



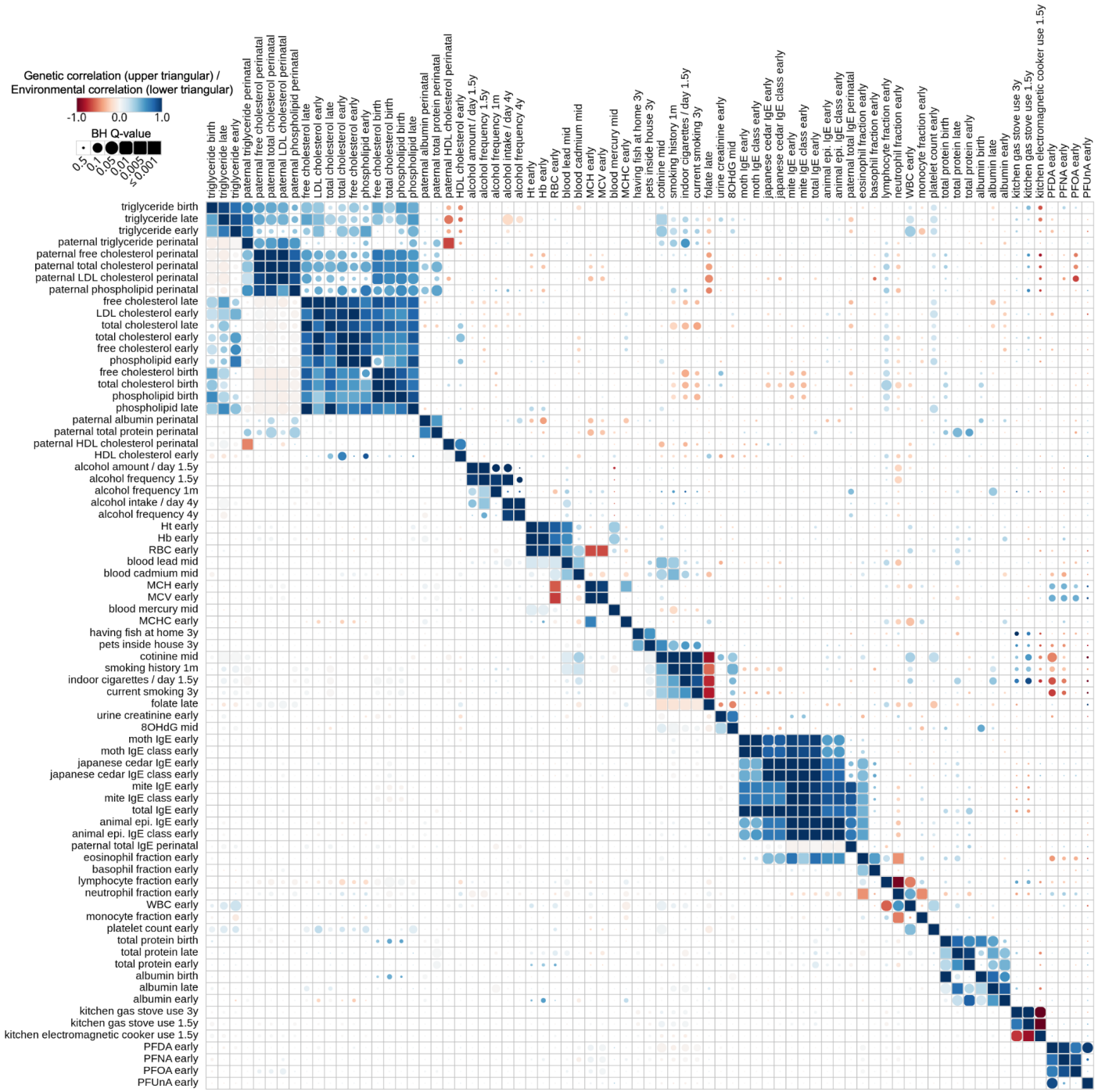
Extended Data Figure 4. a. QQ-plot shows *P*-values of genetic correlations with different cutoff of heritability Z-score (black: $Z_{h^2} > 0$; red $Z_{h^2} > 2.0$; blue: $Z_{h^2} > 4.0$) **b.** Scatterplot shows the relationship between the environmental correlation (x-axis) and the genetic correlation (y-axis) (**Methods**). **c.** Manhattan plots show the GWAS results of the itchy rash status at 1 year old (top), the egg allergy cumulative status until 4 years old (middle) and the milk allergy cumulative status until 4 years old (bottom). **d.** The Locuszoom plots show the GWAS associations around the

SERPINB7 gene. The top panel shows the result for itchy rash at 1 year old and the bottom panel shows the result for egg allergy (both data extracted from the panel b of this figure). The location of the index variant (rs142859678:C>T), which is known to be the pathogenic variant of NPPK (Nagashima-Type Palmoplantar Keratosis), is indicated by a dashed line. **e.** UMAP plots showing scDRS[85] score distributions for cumulative egg and milk allergy scores up to 4 years of age in the reference PBMC data.[86] **f**. Heatmap showing cell type-level scDRS association Z-scores for cumulative egg and milk allergy scores up to 4 years of age. Asterisks indicate significant enrichment after correction for multiple testing across the tested cell types (FDR < 0.05).



Extended Data Figure 5. a. The histogram shows the number of samples along the child's age in month (top) and the box plot shows the distribution of BMI along the child's age in month (bottom). The color indicates the child's age in years (red: 0 years old; green: 1 year old; blue: 2 years old; light blue: 3 years old; purple: 4 years old). **b.** The scatter plot compares QTL effect sizes between JECS and MoBa at variants discovered by the JECS GWAS of 11 different strata. The crossed lines indicate the 95% confidence intervals, which were estimated using standard error estimates. The red color indicates the variant was significant ($P < 5.0 \times 10^{-8}$) in both JECS and MoBa. **c.** Heatmap shows genetic correlations (upper triangular matrix) and environmental correlations (lower triangular matrix) of BMI GWAS at the 11 time points along with BMI in adult (obtained from BioBank Japan[29]) estimated by the cross-trait LD score regression. The color coding is the same as in Fig. 3 in the main text. **d.** QQ-plot of P-values under the null (green and blue) and the alternative (red) hypotheses. The genotype under the two null hypotheses was obtained either (1) by permuting the real genotype data or (2) by drawing from a binomial distribution with a randomly selected probability of success from a uniform distribution (U(0.001, 0.5)) (**Methods**). We obtained the *P*-values from the alternative hypothesis through the dynamic genetic mapping (shown in Fig. 5b of the main text). **e.** The panel shows the estimated dynamic genetic association at the FTO locus, where the mean effect size of T allele compared with A allele at rs9939609 is shown by the red line with the 95% credible interval. **f.** The panel shows the mean BMI along the child's age in month, stratified by the three genotype groups (TT, TA and AA) at rs9939609. **g.** Minor allele frequency distributions in European (1KGP EUR), African (1KGP AFR) and Japanese (JECS) populations, stratified by known (*i.e.*, in GWAS Catalog) or novel loci. **h.** The locus zoom plot shows the ENPP2 eQTL in skeletal muscle tissue. The summary statistics was obtained from the eQTL Catalogue which is based on GTEx V8 (**Data availability**). The color of each point indicates the LD index (r^2) to the TAF2 index variant (rs3812463:C>T). The LD index

between the TAF2 index variant (rs3812463:C>T) and the eQTL lead variant (rs1433953:T>C) was 0.97. **i-l**. The lines show the estimated dynamic polygenic scores for four different participants, with dots indicating the actual observations for each participant. Light color bands around the estimated polygenic scores show the prediction intervals. The color indicates whether the participant is underweight (blue) or overweight (red) compared to the population average, as shown by the boxplot.



Extended Data Figure 6. Cross-trait LDSC result of parent exposure traits. The upper triangular matrix shows the genetic correlations (r_g) and the lower triangular matrix shows the environmental correlations (r_e). Blue colors indicate positive correlations and red colors indicate negative correlations.