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

JECS Genetics Research Group

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 since the participant mothers were pregnant. Analyses of genome-wide genotyping for 80,638 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 1,163 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 4,985 common genomic loci (of which 1,885 loci passed the phenome-wide significant $P = 4.3 \times 10^{-11}$), of which 25% of loci represent novel associations not previously reported. Additional longitudinal GWAS of BMI using Gaussian process regression identified 66 novel dynamic genetic associations along child development. In addition, genetic correlation analysis suggested some evidence that maternal environmental exposures during pregnancy influence child traits at birth. Together with studies of environmental exposures and genetic risk factors, across time and multiple outcomes, these demonstrate the uniqueness and value of the JECS data.

Introduction

Genome-wide association studies are the most effective way to identify previously unknown genetic factors influencing common complex traits, such as human height or type 2 diabetes[1]. They also reveal the underlying biological mechanisms by which these genetic variants affect various human phenotypes, particularly disease risks, which offers new therapeutic targets[2], improved diagnosis[3], and better disease prevention[4] or drug side effects[5]. As of 2024, 691,532 genetic associations in 36,643 traits have been reported in the GWAS Catalog[1]. Despite extensive research conducted worldwide over the past two decades, the genetic basis of common complex traits remains unknown. The few successful adult traits, such as height[6] or diseases like diabetes[7] and cardiovascular diseases[8], whose heritabilities are mostly explained by a number of common variants through the collection of millions of cases, are exceptions. These successes are mainly in European populations, and African and Asian populations are underrepresented[9].

A birth cohort study is the ideal genetic research platform for studying genetic variations in human health and development over the life course[10]. 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 heritabilities in complex diseases[11]. This platform also directly enables the partitioning of genetic effects on birth phenotypes into maternal and fetal components[12]. 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 diagnosis[13–15].

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[16]. 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 still being conducted on average 70% of the child participants until the age of 4.

We present the results of genome-wide association studies investigating a wide range of traits related to child health and development, as well as parental environmental exposures. Subsequently, cross-trait LD score regression (LDCS) analysis was conducted for a cross-sectional analysis of child phenotypes. Furthermore, we elucidate age-specific genetic associations with BMI as an exemplar of longitudinal GWAS. Ultimately, we conducted cross-trait LDSC between maternal exposure and child health outcomes to demonstrate the impact of maternal intrauterine conditions on several fetal growth measures. The findings of these analyses, including GWAS summary statistics for 1,163 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[16,17]. 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[17]. Of these, 80,638 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 619,923 genome-wide genotypes 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 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[18] (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 Osaka Regional Centre after Bonferroni correction).

We then employed KING 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. 1a, b**). 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[19] (**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. 1c**). Following the removal of the aforementioned samples, the first two principal components demonstrated the existence of diversity between the East Asian continental population (including Chinese individuals from Beijing and South Koreans) and the Japanese Ryukyu population (**Fig. 1f**), which can be seen to reflect the dual structure of the Japanese population previously reported[20,21].

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 triples 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, the genotype data were expanded in accordance with the record of multiple births for GWAS (**Methods**). This resulted in 80,278 children, including those from multiple births whose sexes are consistent with their genotypes, for downstream analysis.

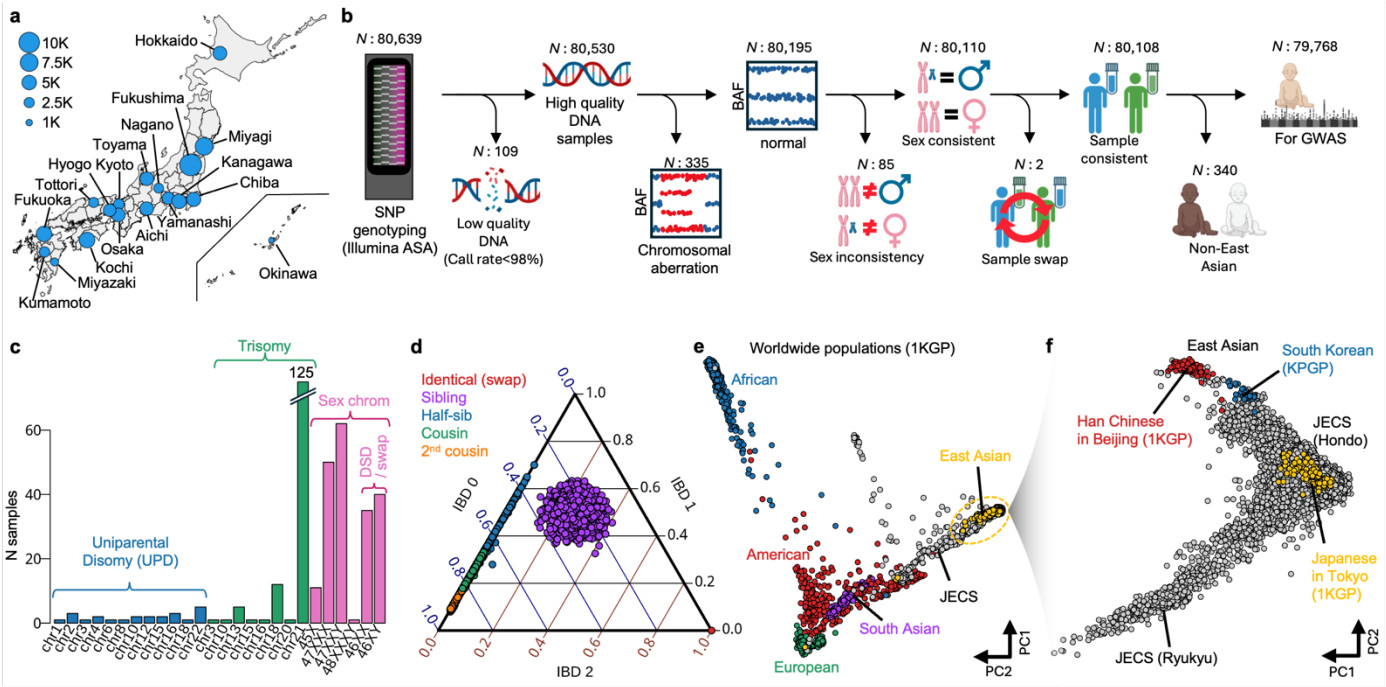


Figure 1. Overview of the JECS genetic study. **a.** Geographic distribution and demographic statistics of the 80,639 participants involved in the JECS genetic study. Dot size on the map indicates the number of participants in each region. **b.** Sample quality control flow for GWAS. Samples were selected by call rate (CR), B allele frequency (BAF) distributions, inferred sex based on genotype data and genotype principal components. **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; green: African; blue: 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).

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 1,033 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 of less than 5% for quantitative traits (**Methods**). The test statistics are often inflated when samples from related populations are analyzed in a GWAS. To address this issue, we employed REGENIE[22] to account for the relatedness between samples (**Methods**). In addition, GWAS were performed on an extra 130 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[23], a viable approach given the observed correlation between child and parent genotypes.

In total, 4,985 GWAS loci were identified across 713 traits (**Supplementary Table 2**), 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 1,885 loci across 263

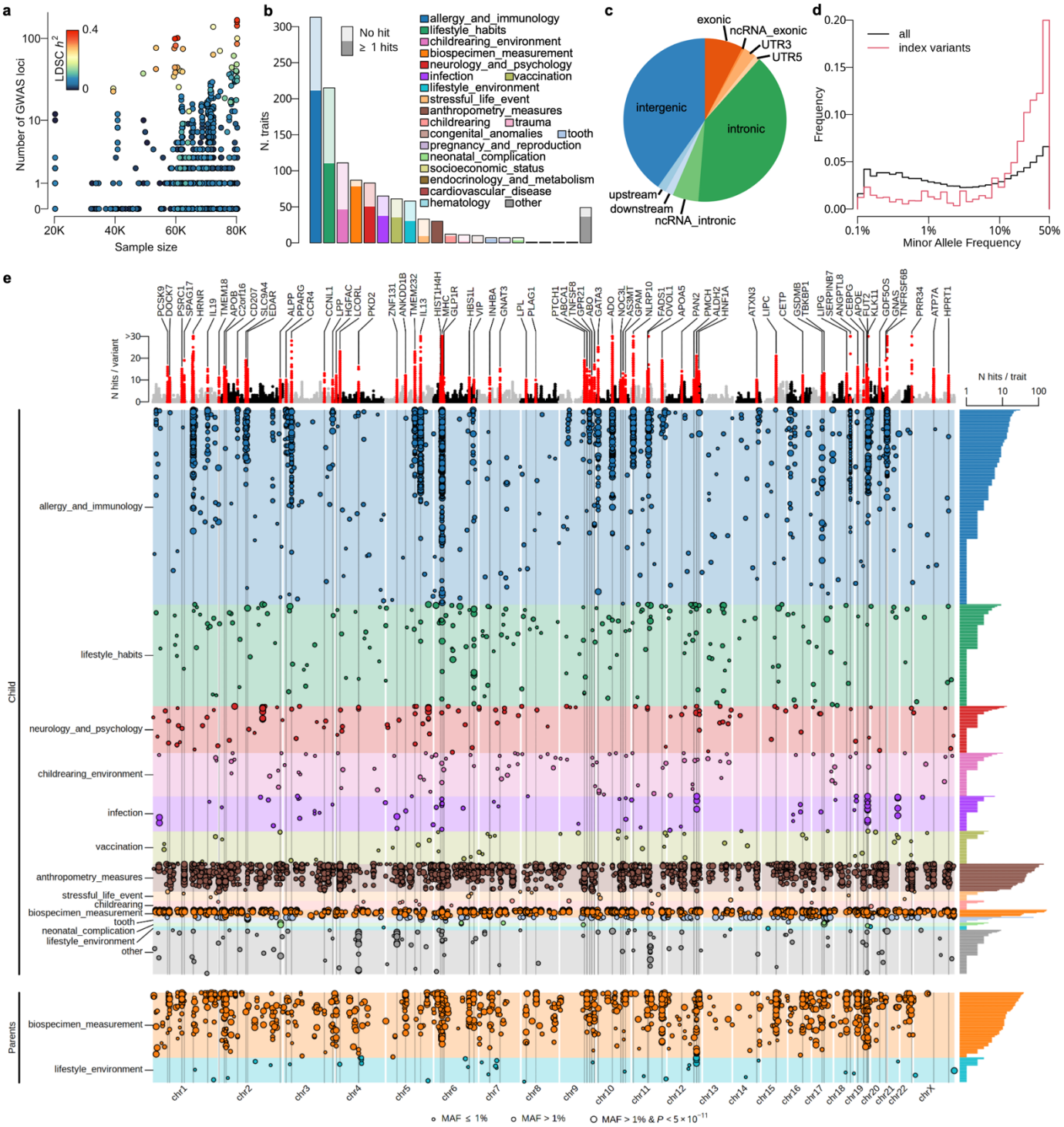


Figure 2. Characteristics of genetic loci and index variants identified by the JECS GWAS. **a.** Scatterplot shows the number of GWAS loci for each trait as a function of sample size. Each dot is colored by heritability estimation from the LD Score Regression (LDSC). **b.** The number of traits in different 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.** Pie chart shows the proportion of variant locations relative to the RefSeq gene (GRCh38). **d.** Distribution of the minor allele frequency of the identified GWAS index variants against those of the entire tested variants. **e.** Genetic associations discovered through GWAS of 1,163 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 right panel shows a barplot of the number of GWAS hits for each trait. The size of each dot indicates the index variant whose minor allele frequency (MAF) is greater than 1%, as well as considering multiple

testing corrections for all the traits ($P < 5.0 \times 10^{-11}$).

traits passed a more stringent threshold of *P*-values after Bonferroni correction (phenome-wide significant $P = 4.3 \times 10^{-11}$). We initially examined the number of GWAS loci identified for each trait as the sample size varied. As expected, the number of loci demonstrated an increase in conjunction with both sample size and heritability (**Fig. 2a**, **Extended Data Fig. 2a**). It should be noted that over 90% of the traits included in the study were analyzed using samples of greater than 60,000 children. 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 (**Fig. 2b**, **Extended Data Fig. 2b**).

According to the RefSeq gene annotation (GRCh38), only 8% of GWAS index variants were found in the coding sequences (**Fig. 2c**). Of these, 330 index variants (6.6% of all) potentially altered protein sequences (**Extended Data Fig. 2c**). 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. By comparing the background allele frequency distribution of all the tested variants, we found that genetic associations are significantly enriched with common variants (**Fig. 2d**; Kolmogorov–Smirnov test $P < 10^{-16}$).

While common complex traits are highly polygenic, the number of variants is finite, which results in single variant affects multiple traits as pleiotropy. The landscape of pleiotropy in European and Japanese populations has been intensively studied[24]. Here we also defined the degree of pleiotropy as the number of significant associations per variant (**Fig. 2e**). Although some pleiotropic loci (e.g., CCR4/IL13) have been identified through GWAS of analogous traits (e.g., food allergy), we also observed loci that are known to be under recent selection pressure[24], including the major histocompatibility complex (MHC), ABO, ALDH2, and APOE loci.

Overall, our discovered loci spanned approximately 7.2% 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[25]. Using our imputed genotype data, we calculated the LD between the index variants and the entire GWAS variants reported in the GWAS catalog, regardless of traits (**Extended Data Fig. 2d**). We identified 996 out of 4,985 index variants (approximately 20% of index variants) that showed weak LD ($r^2 < 0.1$), representing potentially novel genetic loci in a Japanese population. Note that this number doesn't change much (958 or 943 variants) even if we set the cis-window to 500Kb or 1Mb, respectively.

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

We performed the genetic correlation analysis using cross-trait LD score regression[26] between 551 traits with positive heritability estimates ($h^2 > 0$) and one or more genetic loci discovered through GWAS (**Fig. 3, Supplementary Table 3**). We identified 5,113 significant genetic correlations (GCs) after multiple testing corrections (Benjamini-Hochberg $Q < 0.05$). Not surprisingly, they were significantly enriched within the same trait category (4,808 GCs; $OR = 85.2$; $P < 2.2 \times 10^{-16}$). The most frequent trait categories were “allergy and immunology” (3,710 GCs), “neurology and psychology” (554 GCs) and “anthropometry measures” (305 GCs) (**Fig. 3**).

We discovered a modest number of genetic correlations between traits in different trait categories (305 GCs across

trait categories below $Q < 0.05$). For example, we identified GCs between 1) several ISAAC[27,28] traits (food allergy and

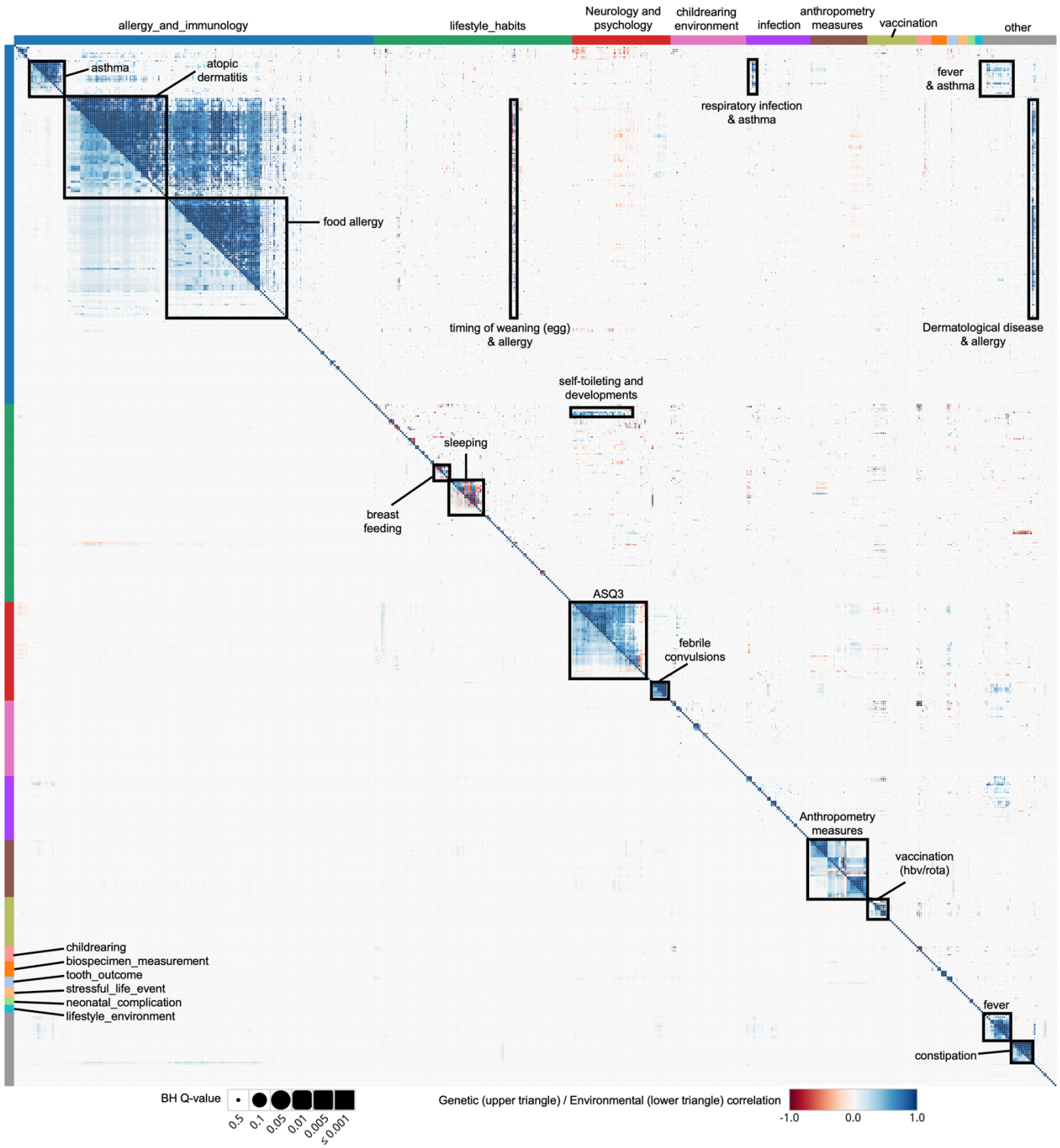


Figure 3. Genetic correlations (upper triangular matrix) and environmental correlations (lower triangular matrix) of 551 child traits estimated by the cross-trait LD score regression. Blue colors indicate positive correlations and red colors indicate negative correlations. The color bars at the top and left of the heatmap indicate trait categories.

atopic dermatitis) and dermatological disease, 2) timing of weaning (eggs) and a part of allergy traits, and 3) asthma and fever/infection traits (**Fig. 3**). There are also reasonable GCs between several developmental traits measured by

the ASQ3[29,30] and a self-toileting trait (**Fig. 3**). Overall, there are no surprising pairs of traits that share genetic architectures. We also note that there is general agreement between environmental correlations (**Methods**) and genetic correlations of trait pairs (**Extended Data Fig. 3a**).

This analysis highlighted not only traits whose correlations are previously unknown, but also similar traits that exhibit different associated genetic loci. An example is the food allergy traits between eggs and milk (**Extended Data Fig. 3b, c**). The GWAS for egg allergy showed a strong association at the MHC locus. In contrast, the GWAS of milk allergy showed associations at the CTLA4 and CCR4 loci, but not in the MHC region. CTLA4 is known as the immunoglobulin superfamily and codes for inhibitory receptors that act as key negative regulators of T-cell responses[31]. CCR4 encodes a receptor for the CC chemokine including TARC; an important biomarker for atopic dermatitis which has an important role in induction and activation of T cells[32]. This suggests that different blood cell types may be involved in different food types. The fact was supported by the fGWAS analysis[25,33–35] (**Methods**) using the various blood cell types in PBMC from the recent single-cell multi-omics data[36] (**Extended Data Fig. 3d**).

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[14]. 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. 4a**).

We first performed standard GWAS of BMI at 11 time points in parallel from birth to 4 years old with a proper adjustment by child age in months. We identified 426 loci ($P < 5.0 \times 10^{-8}$) in the 11 GWAS comprising 248 independent genetic loci after LD pruning between index variants ($r^2 < 0.8$ in our population) (**Fig. 4a**). By comparing with the BMI associations in the GWAS catalog, we identified 204 novel loci whose index variants are in weak LD ($r^2 < 0.8$) with any of the BMI-associated variants previously reported. We then compared the genetic correlations between the GWAS at the 11 time points and the adult BMI GWAS obtained from BioBank Japan[24]. 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. 4b**).

Using all data points simultaneously, we then performed a dynamic genetic mapping using the Gaussian process regression[37] (**Methods**) to discover an additional 73 independent genetic loci, of which 66 loci were novel (**Fig. 4a**). First, to show the statistical calibration of the model, we simulated genotypes from a binomial distribution (MAF ranging from 0.01 to 0.5) and confirmed that the *P*-values follow the uniform distribution (**Extended Data Fig. 4c**; **Methods**). As a proof of concept, we analyzed the FTO locus, which is strongly associated with 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[38]. We clearly showed that the QTL effect was reversed at around 3 years of age (**Extended Data Fig. 4d**), as previously reported[39].

Our novel discoveries were enriched for variants with a significantly lower MAF in a European population (**Fig. 4b**). One

of these novel associations was identified in the TAF2 locus (Fig. 4c). The minimum *P*-value was achieved by an insertion variant (rs397892259:T>TG) in an intron of the TAF2 gene, although 223 additional SNPs and INDELs near this variant

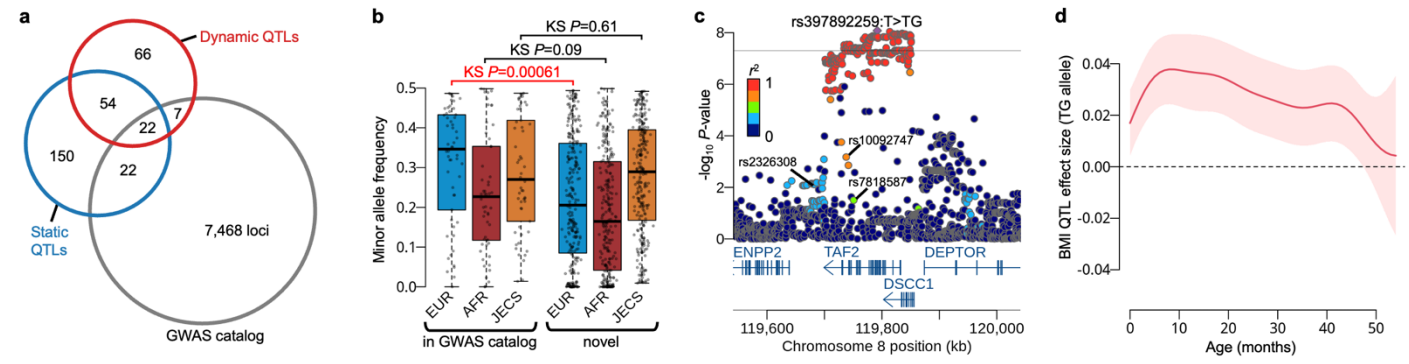


Figure 4. Genetic study of longitudinal traits in JECS. **a.** Snapshot and dynamic genetic effects discovered through longitudinal GWAS of childhood BMI. **b.** Minor allele frequency distributions in European (1KGP EUR), African (1KGP AFR) and Japanese (JECS) populations, stratified by known or novel loci. MAF distributions between known and novel loci are compared by Kolmogorov–Smirnov test for each population. **c.** Locuszoom plot of the TAF2 locus shows the statistical significance of dynamic genetic effects assessed by the score statistics (**Methods**). Points are colored by the LD index value (r^2) to the index variant rs397892259:T>TG. 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 TG allele over the T allele at rs397892259 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.

were in strong LD ($r^2 > 0.8$). The estimated effect size of the TG allele over the T allele showed the dynamic effect, increasing in infants and decreasing in toddlers (**Fig. 4d**). We also found that there are several height-associated variants (rs2326308:C>T, rs10092747:C>T, rs7818587:G>T) in this locus[6]. However, the LD between the index variants and those variants were modest ($0.24 \leq r^2 \leq 0.72$ in our population), therefore the putative causal variants should be different. As there are several putative candidate genes in this locus, we checked the colocalization with eQTLs in GTEx. We found that the index variant (rs397892259) is correlated with the lead variants of 4 different eQTL genes in 6 different tissues (**Extended Data Fig. 4e**). One of them was the ENPP2 eQTL in visceral adipose tissue ($r^2 = 0.96$), and the locuszoom plot was consistent with the GWAS locus (**Fig. 4c** and **Extended Data Fig. 4f**). 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[40].

Impact of maternal exposure during pregnancy on child health and development

Another notable benefit of JECS is its ability to assess environmental exposures of participating parents, including heavy metals[41] and other chemical substances[42], as well as lifestyle factors (such as dietary habits, smoking, and alcohol consumption). GWAS-by-proxy of 130 exposure traits has identified 98 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.

Our study especially highlighted a number of genetic associations to metal (mercury, cadmium, manganese and lead) or metalloid (selenium) exposure with sample sizes of 77,622 which were almost 50 to 100 times larger than those studied previously[43,44]. 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[45]. The locus was detected in maternal blood mercury and lead levels as well as subtypes of per- and polyfluoroalkyl substances[46] (PFAS). We observed the opposite effect direction between

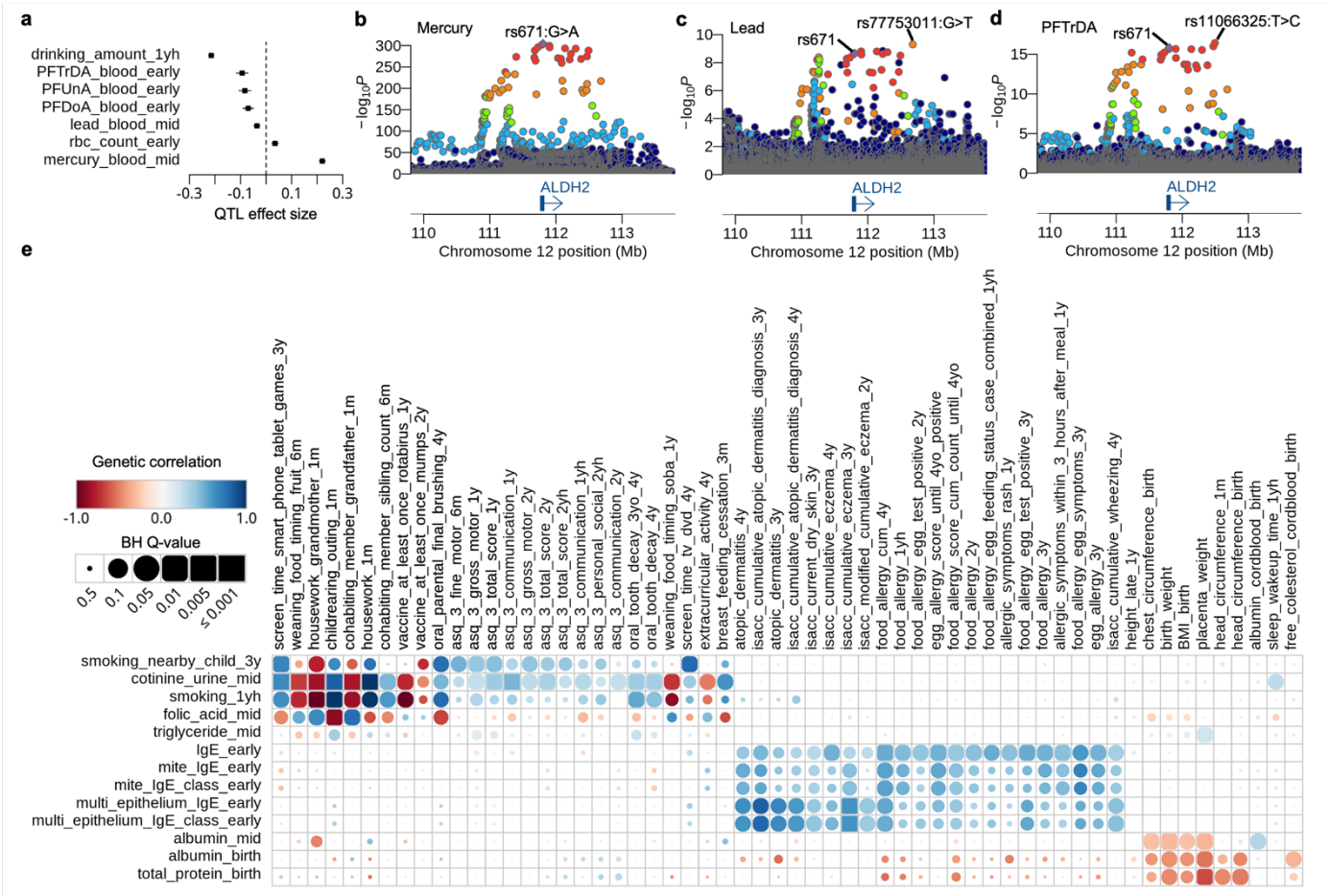


Figure 5. Genetics of environmental exposure traits for participating parents. **a.** QTL effect sizes of the missense SNP rs671:G>A at the ALDH2 locus on several environmental exposure traits. Here, the effect of the A allele was estimated over the G allele. **b.** Locus zoom plot of the blood mercury 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.

alcohol intake and blood mercury level at the missense SNP rs671 (**Fig. 5a**), which suggests alcohol intake reduces the mercury level. This finding aligns with previous reports indicating that alcohol consumption elevates mercury exhaustion in both mice and humans[47,48]. As for the blood lead and PFAS levels, the effect directions are the same as alcohol intake (**Fig. 5a**), while the index variants of the locus are different from alcohol intake (**Fig. 5b-d**) and therefore there can be different mechanisms of metabolism and exhaustion for these substances.

We then performed the cross-trait genetic correlation analysis (**Supplementary Table 3**), which revealed shared genetic architectures between exposure traits (**Extended Data Fig. 5**), such as positive correlations between maternal blood lead level and red blood cell counts (r_g=0.53; Benjamini-Hochberg Q=0.0004), and eosinophil percentage and various IgE levels ($0.48 \leq r_g \leq 0.51$; Q<0.006) during pregnancy. We also observed a negative correlation between urinary cotinine level and folic acid (r_g=-0.79; $Q=2.5\times10^{-10}$). It is worth noting that paternal exposure traits (e.g., cholesterol levels) are

clustered alone and not correlated well with maternal traits, suggesting there may be a sex difference and/or impact of pregnancy which skews genetic associations of these traits.

Finally, we investigated the mediating effect of environmental exposures during pregnancy on child health and development. We identified several maternal IgE levels that should be inherited for some allergic traits in children, such as eczema 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 3**). For example, we found an effect of smoking exposure during pregnancy on tooth decay in children at 4 years of age (r_g=0.35; Q=1.5×10^{-5}). However, we did not observe strong evidence that a specific genetic locus is colocalised 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.

Discussion

In this study, we conducted a large-scale GWAS of child health and development traits in more than 80,000 individuals, greatly expanding our knowledge of common genetic determinants in various early-life traits in an East Asian population. The study included 1,033 traits in children up to the age of four, as well as 130 environmental exposure traits in their parents. We identified 4,985 genetic associations across 21 different trait categories, some of which are highly pleiotropic at MHC or ALDH2 loci. In addition, we discovered a modest number of genetic correlations between children's traits that go beyond their respective categories. Our analysis of longitudinal traits, such as child BMI, revealed both novel and dynamic genetic associations during child development. We also identified genetic correlations, albeit limited in number, between parental environmental exposure traits as well as maternal exposure traits during pregnancy and child traits at birth.

One reason why we observed a modest number of parental environmental effects on child outcomes may be that our GWAS for participating parents were conducted using proxy genotypes, i.e. the genotypes of their children. Because children and parents share a complete haplotype, the genotypes are correlated with an average correlation of 0.5, suggesting that the power to detect genetic associations in the parents' traits is about 1/4 in terms of chi-squared statistics. In addition, it is recognized that there is a bias in genetic correlation analysis when using the result of GWAS-by-proxy[49]. Maternal genotyping is currently underway and will be completed in 2025. At that time, we will have full power to detect genetic associations of parental exposure traits, and we will also be able to fully disentangle genetic effects of children from those of parents on birth outcomes such as birth weight.

We discovered mainly common genetic variants, which may be because statistical power and imputation accuracy were both higher for common variants. We definitely need exome sequencing or whole genome sequencing for future work to uncover rare genetic variants that may contribute to child health and development.

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[50] to identify secondary hits in each locus and integrate genomic annotations, such as ChIP-seq or chromatin accessibility peaks used in fGWAS[33]. We could also colocalize with public eQTL databases,

such as GTEx[51], 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. Freezed phenotype data up to the age of 6 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 the cord blood sample during delivery, and only one sample was taken, even in cases of multiple births. In the case of identical twins or triplets, the common variants genotyped by the SNP array should be identical. However, if the children were dizygotic twins, they were considered genetically siblings. Partially detecting which dizygotic twin pair was genotyped, it was possible to remove dizygotic twin pairs whose sexes were different with the cord blood samples. According to the ratio of monozygotic and dizygotic twins in Japan[52], approximately 100 to 200 individuals of the non-genotyped dizygotic pair were potentially included in the analysis.

DNA extraction from cord blood samples

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

SNP Genotyping

We genotyped 80,639 cord blood samples using the Illumina Asian Screening Array (ASA-24v1-0_E2) BeadChip SNP array (https://www.illumina.com/products/by-type/microarray-kits/infinium-asian-screening.html). We conducted

genotype calling on Illumina GenomeStudio software to generate PED files. These PED files were first converted into the VCF format using PLINK2 (Code availability). We used the fixref plugin of bcftools to align the strand orientation of each SNP (Code availability). For Watson-Crick SNPs (A-T or C-G SNPs), BLAST software (ver 2.5.0) was used to identify the SNP probe orientations a priori (Code availability).

Detection of chromosomal aberrations

Chromosomal aberrations were detected from the signal intensity data for each DNA sample. Nominal XY signal intensities were extracted from Illumina idat files and then converted into B allele frequency (BAF) using the polar coordinate transformation. A mixture of Gaussian mixtures was fitted to detect the number of copies of each autosomal or sex chromosome. The means and variances for the Gaussian mixture distributions were fixed and posterior probability of aberrations were obtained.

Population stratification analysis using principal component analysis

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

Phenotypes

Phenotype data for 4,306 variables were obtained from the JECS questionnaire data of participating children up to the age of 4 (all analyses were conducted using a dataset named jecs-ta-20190930 and jecs-qa-20210401). Each trait was subjectively classified as heritable or non-heritable according to the meaning of the questionnaire, and 1,076 variables were excluded. We then selected 82 variables that were potentially covariates rather than phenotypes. We obtained 3,148 potentially heritable traits, of which we classified the traits available for GWAS as primary data or not and found that 2,259 traits could be analyzed independently. For the remainder (889 traits), we created an additional 601 composite traits, as some traits could not be analyzed independently (e.g., due to copyright) or not calculated a priori (e.g., BMI). The food allergy scores (egg, fish, fruit, milk, nuts, peanuts, rice, sesame, shellfish, soba, soybean and wheat) indicate that, for a certain food, a participant presented with some symptoms after eating or had a positive blood or skin test at some point up until the age of 4 years. We also found that 731 traits were relevant to caregivers but not to children. However, we decided to keep 130 traits that were useful for analyzing environmental exposures during pregnancy or cultural aspects of children.

During data cleaning, appropriate data replacements were applied, because the original data freeze was incomplete in nested questionnaires. In addition, outliers with a high probability of misdescription were manually excluded. For binary and multiple-choice categorical traits, all data with undefined values in the questionnaires were excluded. Categorical traits were converted to a set of binary traits using one-hot encoding. For numerical data (excluding count

data), we excluded data points that were outside the 4 standard deviations (SD). In addition, for some numerical traits, we introduced upper and lower bounds instead of the 4 SD cutoff and excluded only data outside the bounds. These criteria for valid data were discussed and decided by a team including pediatricians and experts in epidemiological and statistical analysis.

When defining the set of traits for the GWAS, 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 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 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 1,163 parental and child traits were analyzed in the GWAS. Detailed information on each trait is summarized in **Supplementary Table 1**.

Genome-wide association study

Genome-wide association study of all the 1,163 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, 16,636,246 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 6 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 6 genotype PCs were used.

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[53] was employed to handle the large number of data points (N=688,752). Model parameters were estimated to maximize the Titsias lower bound[54] of the model using the three stage approach[37]. The score statistics 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 ref[37] for further details.

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 heritability (h^2) was

estimated for each of 1,163 traits using the baseline LD scores. 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) 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 $h^2{}_1$ indicate heritability estimates for the first and the second traits, respectively.

fGWAS analysis

The fGWAS approach[25,33] to determine disease-relevant cell types is described elsewhere[34,35]. The code used for fGWAS plots and for cell type proportion analysis is available in github (Code availability).

Data availability

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 visceral adipose tissue (V8) was downloaded from the eQTL Catalogue (ftp.ebi.ac.uk/pub/databases/spot/eQTL/sumstats/QTS000015/QTD000121/QTD000121.all.tsv.gz). The GWAS summary statistics for all 1,163 traits presented in the manuscript, 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 Beagle (beagle.06Aug24.a91.jar) was used for imputation of the genotyped samples in JECS (https://faculty.washington.edu/browning/beagle/beagle.html). 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/). LDSC (v1.0.1) was used to estimate the trait heritability and genetic correlations between traits (https://github.com/bulik/ldsc). Pairwise Hierarchical Model (PHM) was used to perform fGWAS

analysis (https://github.com/natsuhiko/PHM).

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). 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 and Affiliations

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

Natsuhiko Kumasaka, Hidetoshi Mezawa, Kei Sakamoto, Miori Sato, Mayako Saito-Abe, Limin Yang, Daisuke Harama, Yumiko Miyaji, Minaho Nishizato, Kenji Matsumoto, Kazuhiko Nakabayashi, Masayo Kagami, Kiwako Yamamoto-Hanada, Maki Fukami

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

Natsuhiko Kumasaka

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

Tomohiko Isobe, Makiko Sekiyama, Miyuki Iwai, Yayoi Kobayashi, Mai Takagi, Nozomi Tatsuta, Yu Taniguchi, Shoji Nakayama, Shin Yamazaki

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

Kenji Matsumoto

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

Kazuhiko Nakabayashi

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

Masayo Kagami

Institute of Medicine, University of Tsukuba, Tsukuba, Japan

Emiko Noguchi

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

Akira Hata

Graduate School of Medicine, Gunma University, Maebashi, Japan

Kenichiro Hata

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

Yukinori Okada

Graduate School of Medical Sciences, Kanazawa University, Kanazawa, Japan

Atsushi Tajima

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

Shi Mingyang, Masaru Koido, Yoichiro Kamatani, Koichi Matsuda

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

Yosuke Kawai, Katsushi Tokunaga

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

Reiko Kishi

Graduate School of Medicine, Tohoku University, Sendai, Japan

Chiharu Ota

School of Medicine, Fukushima Medical University, Fukushima, Japan

Koichi Hashimoto

School of Medicine, Chiba University, Chiba, Japan

Chisato Mori

School of Medicine, Yokohama City University, Yokohama, Japan

Shuich Ito

Center for Birth Cohort Studies, University of Yamanashi, Chuo, Japan

Ryoji Shinohara

School of Medicine, University of Toyama, Toyama, Japan

Hidekuni Inadera

Graduate School of Medicine, Kyoto University, Kyoto, Japan

Takeo Nakayama

Graduate School of Medicine, Osaka University, Suita, Japan

Ryo Kawasaki

Pediatrics, Hyogo Medical University, Nishinomiya, Japan

Yasuhiro Takeshima

Faculty of Medicine, Tottori University, Yonago, Japan

Seiji Kageyama, Akemi Morita

Kochi Medical School, Kochi University, Nankoku, Japan

Narufumi Suganuma

Graduate School of Medical Sciences, Kyushu University, Fukuoka, Japan

Shoichi Ohga

Faculty of Life Sciences, Kumamoto University, Kumamoto, Japan

Takahiko Katoh

Graduate School of Medical Sciences, Nagoya City University, Nagoya, Japan

Michihiro Kamijima, Yukihiro Ohya

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

Akihiro Umezawa

JECS Genetics Research Group

Genetics Research Working Group

Natsuhiko Kumasaka, Hidetoshi Mezawa, Kei Sakamoto, Kenji Matsumoto, Kazuhiko Nakabayashi, Masayo Kagami, Emiko Noguchi, Akira Hata, Kenichiro Hata, Yukinori Okada, Atsushi Tajima, Yukihiro Ohya

JECS Group

Reiko Kishi, Chiharu Ota, Koichi Hashimoto, Chisato Mori, Shuich Ito, Ryoji Shinohara, Hidekuni Inadera, Takeo Nakayama, Ryo Kawasaki, Yasuhiro Takeshima, Seiji Kageyama, Akemi Morita, Narufumi Suganuma, Shoichi Ohga, Takahiko Katoh, Michihiro Kamijima, Maki Fukami, Shin Yamazaki

BioBank Japan

Shi Mingyang, Masaru Koido, Natsuhiko Kumasaka, Yukinori Okada, Yoichiro Kamatani, Koichi Matsuda

National Center Biobank Network

Yosuke Kawai, Katsushi Tokunaga

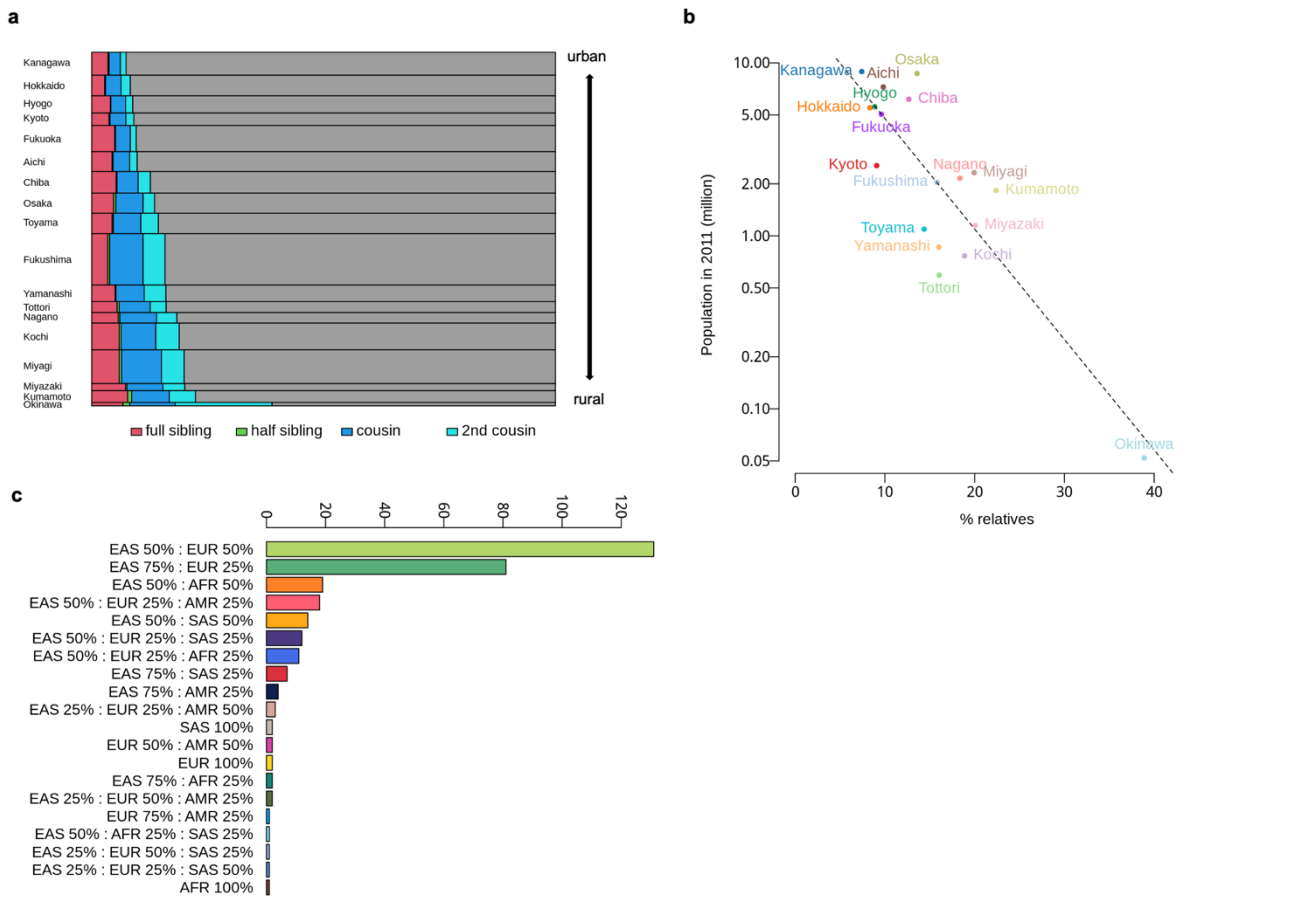
Contributions

All authors reviewed and revised the manuscript. Detailed author contributions are available in the Supplementary Information.

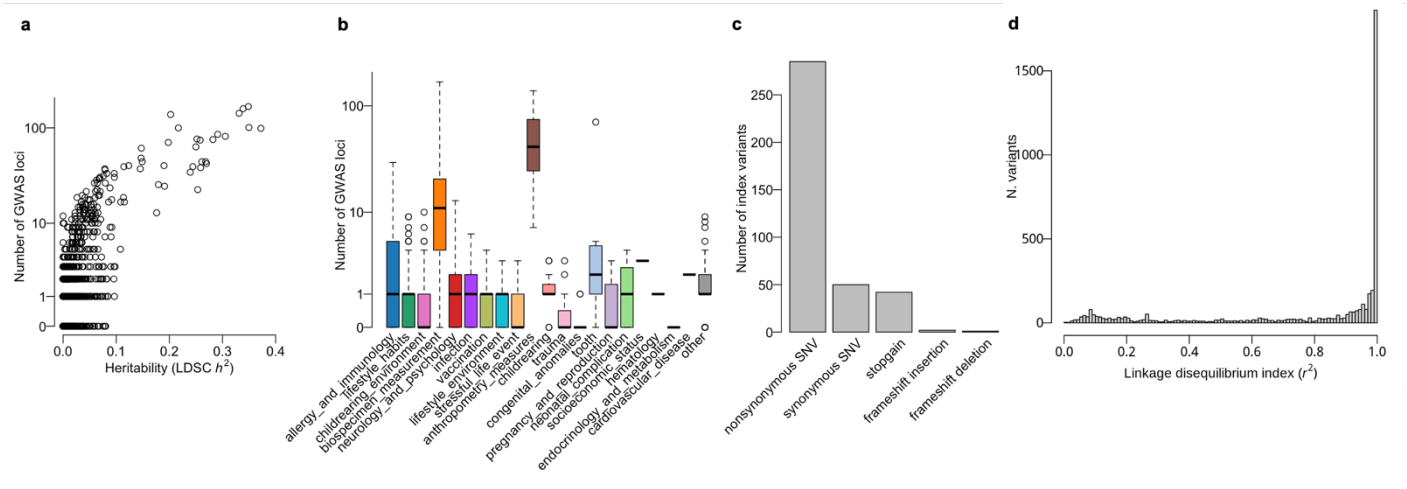
Corresponding authors

Correspondence to Natsuhiko Kumasaka or Shoji Nakayama.

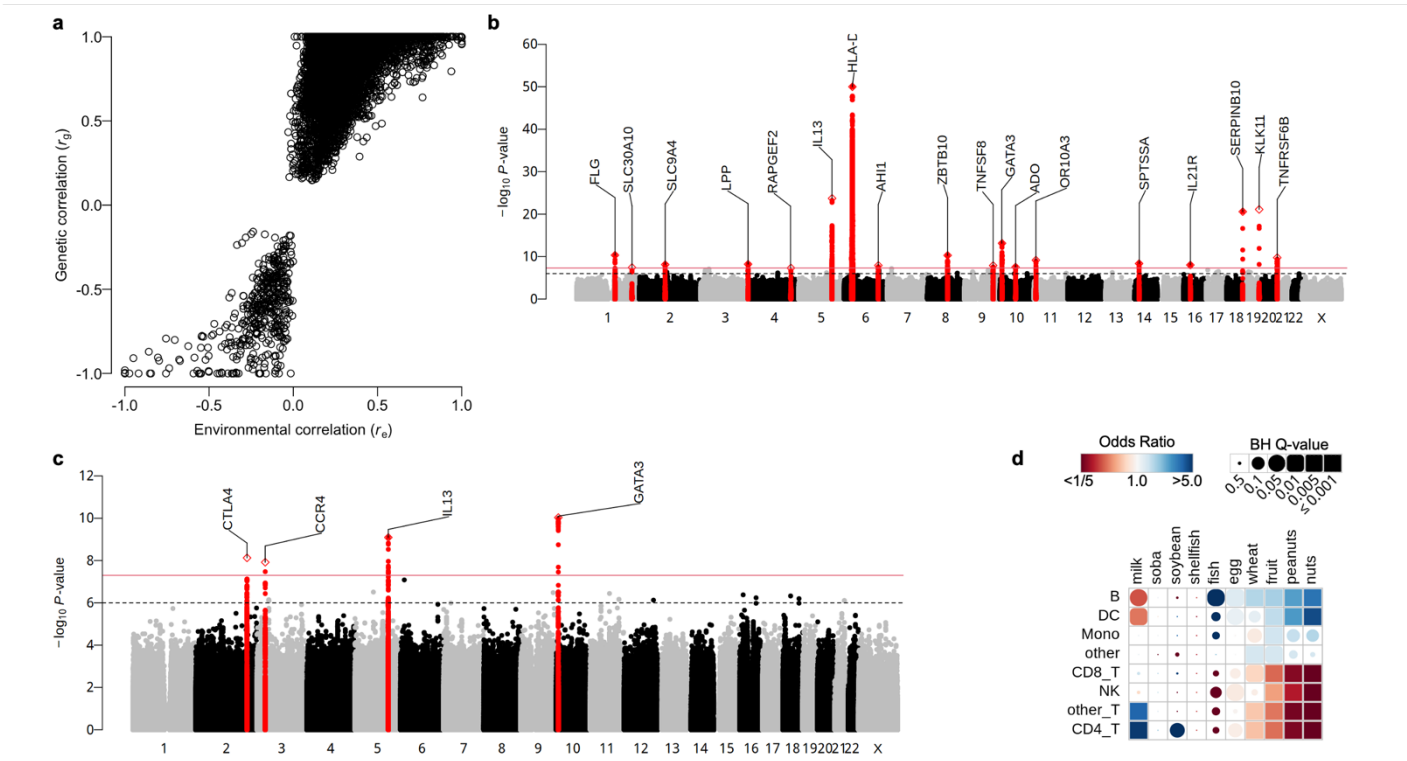
Extended Data Figures



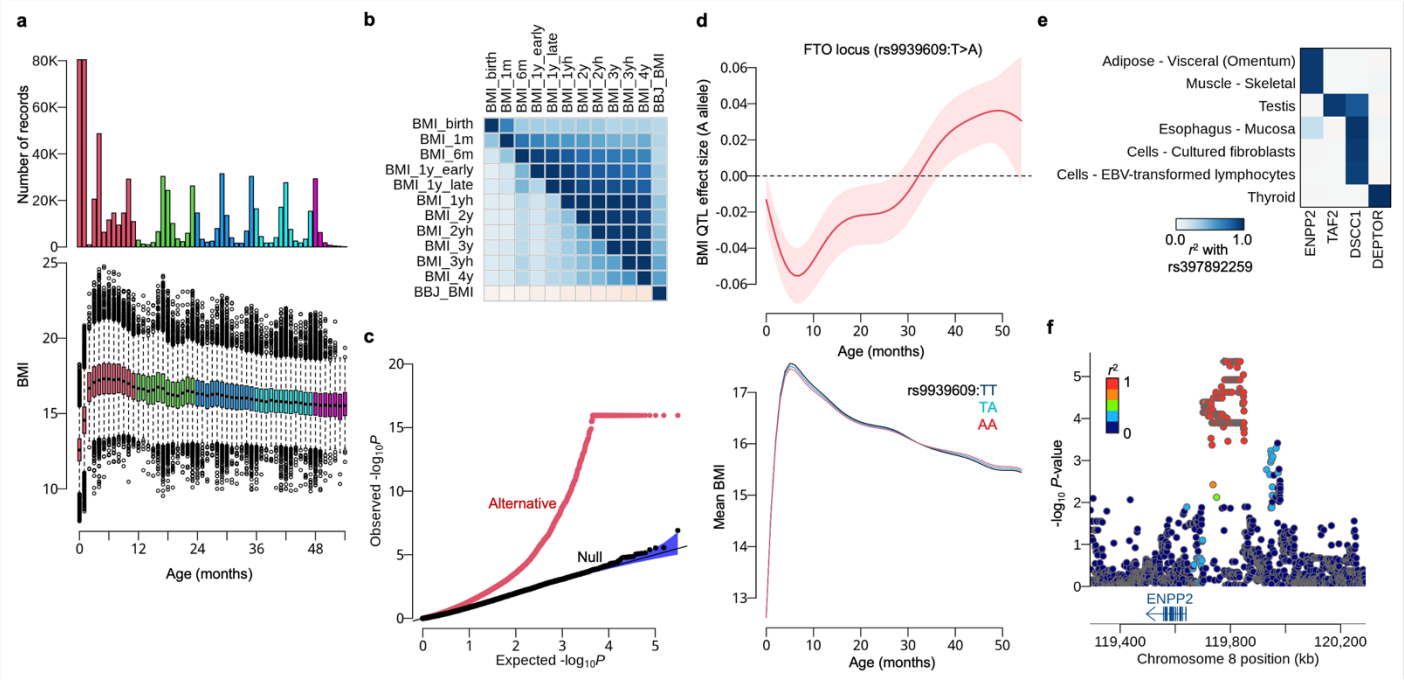
Extended Data Figure 1. a. Mosaicplot shows the percentage of relatives in our data, stratified by the prefectures of Japan where participants were enrolled. **b.** Scatterplot shows the relationship between the percentage of relatives (full/half siblings, cousins and second cousins) and the population (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. **c**. Barplot shows the number of non-East Asian participants with different ancestral backgrounds inferred from the principal component analysis.



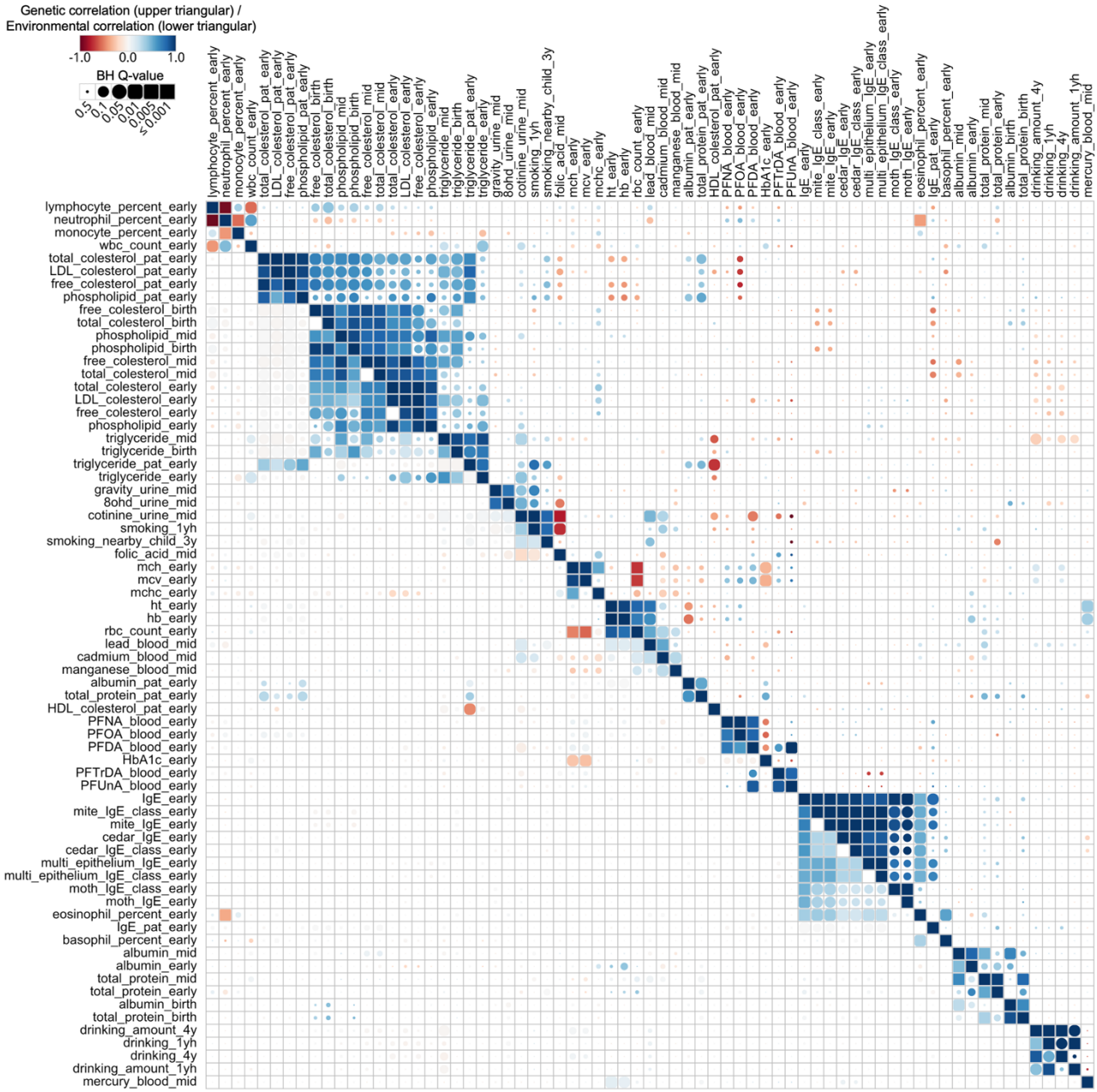
Extended Data Figure 2. a. Scatterplot shows the number of GWAS loci for each trait as a function of heritability estimation obtained by the LD Score regression (LDSC). **b.** Boxplot shows the number of significant GWAS loci ($P < 5.0 \times 10^{-8}$) for each trait stratified by the trait categories. **c.** The number of index variants detected within the coding regions, stratified by the putative functions of variants annotated by the RefSeq gene. **d.** 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. Scatterplot shows the relationship between the environmental correlation (x-axis) and the genetic correlation (y-axis) (**Methods**). **b.** Manhattan plot shows the GWAS of the cumulative egg allergy score until 4 years old. **c.** Manhattan plot shows GWAS of cumulative milk allergy score until 4 years old. **d.** Cell type enrichment analysis using the reference PBMC data[36] and the fGWAS model[25,33–35] (**Methods**).



Extended Data Figure 4. 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. 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[24]) estimated by the cross-trait LD score regression. The color coding is the same as in Fig. 3 in the main text. **c.** QQ-plot of P-values under the null (black) and the alternative (red) hypotheses, where the genotype under the null hypothesis was drawn from a binomial distribution with the probability of success randomly drawn from the uniform distribution $U(0.01,0.99)$. **d.** The top 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. The bottom panel shows the mean BMI along the child's age in month, stratified by the three genotype groups (TT, TA and AA) at rs9939609. **e.** Heatmap shows the LD index (r^2) between the TAF2 index variant (rs397892259:T>TG) and the lead variants of eQTL genes across 6 tissues in GTEx (V8). **f.** The locus zoom plot shows the ENPP2 eQTL in visceral adipose tissue.The summary statistics was obtained from the eQTL Catalogue (**Data availability**). The color of each point indicates the LD index (r^2) to the TAF2 index variant (rs397892259:T>TG). The LD index between the TAF2 index variant (rs397892259:T>TG) and the eQTL lead variant (rs1433953:T>C) was 0.96.



Extended Data Figure 5. 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.

References

1. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
2. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
3. Jung, S. *et al.* Case-case genome-wide association analysis identifying genetic loci with divergent effects on Crohn's disease and ulcerative colitis. *Hum. Mol. Genet.* **32**, 677–684 (2023).
4. Inouye, M. *et al.* Genomic risk prediction of coronary artery disease in 480,000 adults: Implications for primary prevention. *J. Am. Coll. Cardiol.* **72**, 1883–1893 (2018).
5. Giacomini, K. M. *et al.* Genome-wide association studies of drug response and toxicity: an opportunity for genome medicine. *Nat. Rev. Drug Discov.* **16**, 1 (2017).
6. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).
7. Mahajan, A. *et al.* Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nat. Genet.* **54**, 560–572 (2022).
8. Tcheandjieu, C. *et al.* Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. *Nat. Med.* **28**, 1679–1692 (2022).
9. Fitipaldi, H. & Franks, P. W. Ethnic, gender and other sociodemographic biases in genome-wide association studies for the most burdensome non-communicable diseases: 2005-2022. *Hum. Mol. Genet.* **32**, 520–532 (2023).
10. 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).
11. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
12. Huang, S. *et al.* The Born in Guangzhou Cohort Study enables generational genetic discoveries. *Nature* **626**, 565–573 (2024).
13. Helgeland, Ø. *et al.* Genome-wide association study reveals dynamic role of genetic variation in infant and early childhood growth. *Nat. Commun.* **10**, 4448 (2019).
14. Helgeland, Ø. *et al.* Characterization of the genetic architecture of infant and early childhood body mass index. *Nat. Metab.* **4**, 344–358 (2022).
15. Westergaard, D. *et al.* Genome-wide association meta-analysis identifies five loci associated with postpartum hemorrhage. *Nat. Genet.* **56**, 1597–1603 (2024).
16. Kawamoto, T. *et al.* Rationale and study design of the Japan environment and children's study (JECS). *BMC Public Health* **14**, 25 (2014).
17. Michikawa, T. *et al.* Baseline profile of participants in the Japan Environment and Children's Study (JECS). *J.*

Epidemiol. **28**, 99–104 (2018).

18. García-Acero, M., Moreno, O., Suárez, F. & Rojas, A. Disorders of sexual development: Current status and progress in the diagnostic approach. *Curr. Urol.* **13**, 169–178 (2020).
19. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
20. Yamaguchi-Kabata, Y. *et al.* Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. *Am. J. Hum. Genet.* **83**, 445–456 (2008).
21. Yamamoto, K. *et al.* Genetic legacy of ancient hunter-gatherer Jomon in Japanese populations. *Nat. Commun.* **15**, 9780 (2024).
22. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
23. Liu, J. Z., Erlich, Y. & Pickrell, J. K. Case-control association mapping by proxy using family history of disease. *Nat. Genet.* **49**, 325–331 (2017).
24. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**, 1415–1424 (2021).
25. Kumasaka, N., Knights, A. J. & Gaffney, D. J. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. *Nat. Genet.* **51**, 128–137 (2019).
26. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
27. Asher, M. I. *et al.* International Study of Asthma and Allergies in Childhood (ISAAC): rationale and methods. *Eur. Respir. J.* **8**, 483–491 (1995).
28. Weiland, S. K. *et al.* Phase II of the International Study of Asthma and Allergies in Childhood (ISAAC II): rationale and methods. *Eur. Respir. J.* **24**, 406–412 (2004).
29. Squires, J., Twombly, E., Bricker, D. D. & Potter, L. ASQ-3 user's guide. (2009).
30. Mezawa, H. *et al.* Psychometric profile of the Ages and Stages Questionnaires, Japanese translation. *Pediatr. Int.* **61**, 1086–1095 (2019).
31. Pardoll, D. M. The blockade of immune checkpoints in cancer immunotherapy. *Nat. Rev. Cancer* **12**, 252–264 (2012).
32. Catherine, J. & Roufosse, F. What does elevated TARC/CCL17 expression tell us about eosinophilic disorders? *Semin. Immunopathol.* **43**, 439–458 (2021).
33. 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).
34. Elmentaite, R. *et al.* Cells of the human intestinal tract mapped across space and time. *Nature* **597**, 250–255 (2021).

35. Madissoon, E. *et al.* A spatially resolved atlas of the human lung characterizes a gland-associated immune niche. *Nat. Genet.* (2022) doi:10.1038/s41588-022-01243-4.
36. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
37. Kumasaka, N. Genetic association mapping leveraging Gaussian processes. *J. Hum. Genet.* **69**, 505–510 (2024).
38. Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.* **373**, 895–907 (2015).
39. Sovio, U. *et al.* Association between common variation at the FTO locus and changes in body mass index from infancy to late childhood: the complex nature of genetic association through growth and development. *PLoS Genet.* **7**, e1001307 (2011).
40. Sah, J. P. *et al.* Ectonucleotide pyrophosphatase 2 (ENPP2) plays a crucial role in myogenic differentiation through the regulation by WNT/β-Catenin signaling. *Int. J. Biochem. Cell Biol.* **118**, 105661 (2020).
41. Nakayama, S. F. *et al.* Blood mercury, lead, cadmium, manganese and selenium levels in pregnant women and their determinants: the Japan Environment and Children's Study (JECS). *J. Expo. Sci. Environ. Epidemiol.* **29**, 633–647 (2019).
42. Iwai-Shimada, M. *et al.* Questionnaire results on exposure characteristics of pregnant women participating in the Japan Environment and Children Study (JECS). *Environ. Health Prev. Med.* **23**, 45 (2018).
43. Ng, E. *et al.* Genome-wide association study of toxic metals and trace elements reveals novel associations. *Hum. Mol. Genet.* **24**, 4739–4745 (2015).
44. Eom, S.-Y. *et al.* Exome-wide association study identifies genetic polymorphisms of C12orf51, MYL2, and ALDH2 associated with blood lead levels in the general Korean population. *Environ. Health* **16**, 11 (2017).
45. Matoba, N. *et al.* GWAS of 165,084 Japanese individuals identified nine loci associated with dietary habits. *Nature Human Behaviour* (2020) doi:10.1038/s41562-019-0805-1.
46. Nakayama, S. F. *et al.* Poly- and perfluoroalkyl substances in maternal serum: Method development and application in Pilot Study of the Japan Environment and Children's Study. *J. Chromatogr. A* **1618**, 460933 (2020).
47. Dunn, J., Clarkson, T. & Magos, L. Ethanol-increased exhalation of mercury in mice. *Br. J. Ind. Med.* **35**, 241–244 (1978).
48. 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).
49. 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).
50. 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).
51. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
52. Jones, R. E. & Lopez, K. H. Gamete Transport and Fertilization. in *Human Reproductive Biology* 159–173 (Elsevier,

2014).

53. Hensman, J., Fusi, N. & Lawrence, N. D. Gaussian Processes for Big Data. *arXiv [cs.LG]* (2013).

54. Titsias, M. Variational learning of inducing variables in sparse Gaussian processes. *Artificial Intelligence and Statistics* (2009).