

SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

X chromosome genetic data in a Spanish children cohort, dataset description and analysis pipeline

Augusto Anguita-Ruiz ^{1,2,3,4}, Julio Plaza-Diaz^{1,2,3}, Francisco Javier Ruiz-Ojeda^{1,3,9}, Azahara I. Rupérez ^{1,6,7}, Rosaura Leis^{4,5}, Gloria Bueno^{4,6,7}, Mercedes Gil-Campos^{4,8}, Rocío Vázquez-Cobela^{4,5}, Ramón Cañete^{4,8}, Luis A. Moreno⁶, Ángel Gil ^{1,2,3,4} & Concepción María Aguilera ^{1,2,3,4}

Received: 19 November 2018

Accepted: 24 May 2019

Published online: 22 July 2019

X chromosome genetic variation has been proposed as a potential source of missing heritability for many complex diseases, including obesity. Currently, there is a lack of public available genetic datasets incorporating X chromosome genotype data. Although several X chromosome-specific statistics have been developed, there is also a lack of readily available implementations for routine analysis. Here, we aimed: (1) to make public and describe a dataset incorporating phenotype and X chromosome genotype data from a cohort of 915 normal-weight, overweight and obese children, and (2) to deeply describe a whole implementation of the special X chromosome analytic process in genetics. Datasets and pipelines like this are crucial to get familiar with the steps in which X chromosome requires special attention and may raise awareness of the importance of this genomic region.

Background & Summary

Overweight and obesity in children are a public health problem that has raised concern worldwide¹. Childhood obesity is characterized by an expansion of the adipose tissue (AT)¹ and plays an important role in the development of cardiometabolic alterations during early adulthood, thereby increasing morbidity and mortality². According to twin and family studies, around 40–70% of the interindividual variability in body mass index (BMI) has been attributed to genetic factors^{3–5}. Despite this, known single-nucleotide polymorphisms (SNPs) explain <2% of BMI variation⁶, a phenomenon termed ‘missing heritability’. Potential sources explaining this missing heritability include epigenetic components, the existence of low frequency and rare variants as well as the presence of X chromosome genetic variation.

Analysis in current genetic association studies is usually focused on autosomal variants while the sex chromosomes, and specially the X chromosome, are often neglected. Among the reasons, it highlights a lower gene density on the X chromosome, a lower coverage of the region in current genotyping platforms and a number of technical hurdles including complications in genotype calling, imputation and selection of test statistics⁷. According to a previous report, only 242 out of all 743 GWAS conducted from 2005 to 2011 considered the X chromosome in their analyses⁷. The proportion was similar when only family-based GWAS were considered.

¹Department of Biochemistry and Molecular Biology II, School of Pharmacy, University of Granada, Granada, 18011, Spain. ²Institute of Nutrition and Food Technology “José Mataix”, Center of Biomedical Research, University of Granada, Avda. del Conocimiento s/n, Granada, 18016, Spain. ³Biosanitary Research Institute of Granada (IBS. GRANADA), University Clinical Hospital San Cecilio, Av. de la Investigación, s/n, Granada, 18016, Spain. ⁴CIBEROBN, (Physiopathology of Obesity and Nutrition CB12/03/30038), Institute of Health Carlos III (ISCIII), Madrid, 28029, Spain. ⁵Unit of Investigation in Nutrition, Growth and Human Development of Galicia, Pediatric Department (USC). Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), University Clinical Hospital, Santiago de Compostela, Spain. ⁶Growth, Exercise, Nutrition and Development (GENUD) Research Group, Universidad de Zaragoza, Zaragoza, Spain. ⁷Instituto Agroalimentario de Aragón (IA2) and Instituto de Investigación Sanitaria de Aragón (IIS Aragón), Zaragoza, Spain. ⁸Department of Pediatric Endocrinology, Reina Sofia University Clinical Hospital, Institute Maimónides of Biomedicine Investigation of Córdoba (IMIBIC), University of Córdoba, Avda. Menéndez Pidal s/n, 14004, Córdoba, Spain. ⁹Present address: RG Adipocytes and metabolism, Institute for Diabetes and Obesity, Helmholtz Diabetes Center at Helmholtz Center Munich, Munich, Germany. Correspondence and requests for materials should be addressed to A.A.-R. (email: augustoanguitaruiz@gmail.com)

There is therefore a lack of available public datasets including X chromosome genotype data for analysis. On the other hand, although several X chromosome-specific statistical tests and guidelines have now become available, there is also a lack of readily available implementations and user-friendly apps incorporating them for routine analysis^{8,9}.

The majority of the technical hurdles faced when analysing X chromosomal data rise from two of its main particularities. The first one is the fact of women having two allele copies while males having only one. As a consequence, if males are included in the analysis, special caution must be taken. Particularly, the study design process should be performed carefully, trying to maintain a balanced female/male ratio across experimental conditions. Otherwise, many available statistical tests will suffer from type I errors as soon as sex-specific allele frequencies occur, which is typically observed for a great number of variants. Other problems derived from an unbalanced sex ratio in the study sample include problems during the genotype calling process, as the signal intensities obtained from standard array genotyping platforms will be always lower in males than for females (who carry two alleles). The second uniqueness motivating X-chromosome specific analyses lies in the X chromosome inactivation (XCI) process, through which most of the cells of females express only one X chromosome allele in order to compensate the genetic dosage with regard to males. Before selecting a particular statistical approach, it should be mandatory to carefully investigate the concrete XCI model to assume for a gene in a particular tissue. Depending on the XCI model assumed, we should proceed one way or another during the selection of the test statistics. These and other particularities must be addressed as long as X chromosomal data are included into genetic studies.

In relation to obesity, only a few studies have reported association with markers on the X chromosome. One of the most remarkable findings involves the tenomodulin (*TNMD*) gene, a Xq22 located locus encoding a type II transmembrane glycoprotein. First time associated with adult obesity at the genetic level¹⁰, its presence in adult human AT has been demonstrated showing higher expression in obesity and lower expression after diet-induced weight loss. Regarding children population, our research group found that *TNMD* expression was five fold-times up-regulated in visceral adipose tissue (VAT) of children with obesity, compared with their normal-weight counterparts (Gene Expression Omnibus GSE9624)^{11,12}. Recently, we have reported new associations between *TNMD* SNPs and childhood obesity and metabolic alterations in a Spanish children population¹³. Interestingly, our study has been the first to analyse and detect associations between X chromosome *TNMD* genetic variants and obesity in a children cohort. Similarly, SNPs in the *SLC6A14* gene, also located in the X chromosome, have shown evidence of association with obesity¹⁴. As a whole, these *TNMD* and *SLC6A14* reports support the fact that X chromosome genetic variants may be not only useful early life risk indicators of obesity but also an interesting source of missing heritability¹³.

Given the lack of public available genetic datasets incorporating X chromosome genetic variants and the still prevalent statistical hurdles that make the X chromosome a difficult region to be tested in functional genetics, we here aimed: (1) to make public and describe a dataset incorporating X chromosome genotype data from a children cohort^{13,15}, and (2) to outline a whole implementation of the special X chromosome analytic process in genetics. The presented research dataset includes X-chromosomal SNP data (mapping the genes *TNMD* and *SLC6A14*) from a children cohort composed of 915 normal-weight, overweight and obese subjects. Some topics covered in this paper include dataset sharing and description, explanation of sample design, genotype calling, quality control, and test statistics selection procedures. Additionally, a short section explaining and interpreting findings obtained after analysing the dataset with a specific X chromosome analytic approach is presented.

Methods

Experimental design and study population. These methods are an expanded version of descriptions in our related work and general characteristics of the dataset have been previously described¹³. Briefly, in this case-control multicentre study, 915 Spanish children (438 males and 477 females) were recruited from three national health institutions: Lozano Blesa University Clinical Hospital, Santiago de Compostela University Clinical Hospital and Reina Sofía University Clinical Hospital. According to specific X-chromosomal analytic requirements, the female/male ratio of the study sample was perfectly balanced.

Childhood obesity status was defined according to the International Obesity Task Force (IOTF) reference for children¹⁶ which is based on the application, on children population, of the widely used cut-off points of BMI for adults (25 and 30 kg/m², for overweight and obesity respectively). Particularly, these criteria constitute a range of age and sex specific cut-off points for children that have been extracted from solid percentile tables constructed on 97876 boys and 94851 girls ranging from 2 to 18 years. After the application these specific cut-off points, the dataset was composed of 480 children in the obesity group, 177 in the overweight group and 258 in the normal-BMI group. Children were allocated into two experimental conditions according to their obesity status; the affected group (cases) composed of both children with obesity or overweight and the control group composed of normal-weight children. An unbalanced female/male ratio across cases and controls has been proven to heavily affect the power of some specific X chromosome association tests¹⁷. In our study, a balanced female/male ratio was maintained across each experimental condition (122/136 in controls and 355/302 in cases) (Fig. 1).

Inclusion criteria were European-Caucasian heritage and the absence of congenital metabolic diseases. Otherwise, the exclusion criteria were non-European Caucasian heritage, the presence of congenital metabolic diseases (e.g., diabetes or hyperlipidemia), undernutrition, and the use of medication that alters blood pressure, glucose or lipid metabolism.

Ethical statement. All procedures in the study were conducted in accordance with the Declaration of Helsinki (Edinburgh 2000 revised), and followed the recommendations of both the Good Clinical Practice of the CEE (Document 111/3976/88 July 1990) and the legally enforced Spanish regulation for clinical investigation of human beings (RD 223/04 about clinical trials). The Ethics Committees on Human Research of all participant institutions approved all experiments and analyses with registration Code: “2011/198”. All parents or guardians provided written informed consent and children gave their assent.

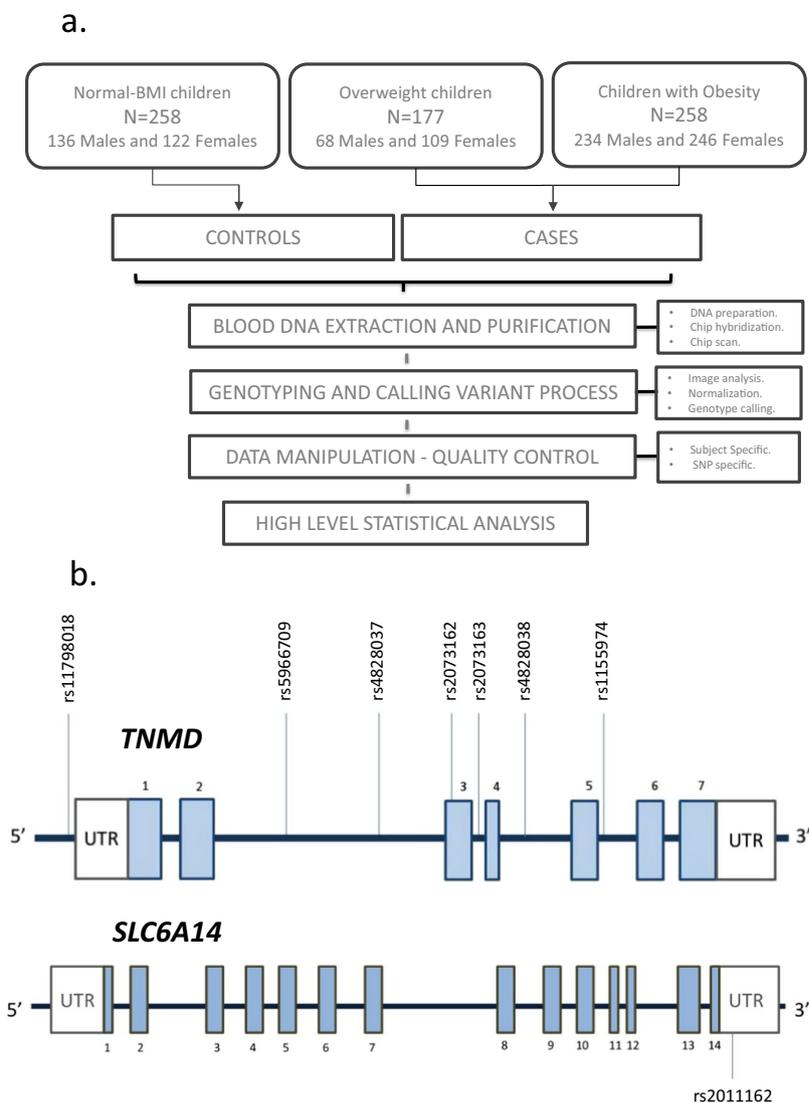


Fig 1 Study design and characteristics. **(a)** Experimental workflow used to generate and analyse the data. **(b)** Genomic context of selected markers; light blue boxes represent exons, while the connecting lines are introns. Abbreviations; rs, reference SNP; UTR, untranslated region.

DNA extraction, processing and analysis. The presented dataset consists on genotype data for eight target SNPs mapping the X-chromosomal genes *TNMD* and *SLC6A14* in the study population. Details regarding SNP selection and molecular analyses are briefly covered here since they have already been fully detailed in our previous work¹³. On the contrary, we pay special attention in the explanation of X chromosomal particularities, data description as well as in the summarization of each data analysis and processing step.

Seven SNPs located at the *TNMD* locus and one located at the *SLC6A14* were selected for genotyping analysis. Genomic DNA was extracted from peripheral white blood cells using two automated kits, the Qiaamp DNA Investigator Kit for coagulated samples and the Qiaamp DNA Mini & Blood Mini Kit for non-coagulated samples (QIAGEN Systems, Inc., Valencia, CA, USA). All extractions were purified using the DNA Clean and Concentrator kit from Zymo Research (Zymo Research, Irvine, CA, USA). Genotyping was performed by TaqMan allelic discrimination assay using the QuantStudio 12 K Flex Real-Time PCR System (Thermo Fisher Scientific, Waltham, MA, USA). Given the X-chromosomal location, it is recommendable to analyse females and males in separate plates during the genotyping process or, at least, maintain a balanced female/male ratio by plate.

Once genotyping was accomplished, we checked candidate SNPs for sex-specific allele frequencies, which can induce type I errors in some statistical X-chromosome analyses (especially in the case of unbalanced designs). Tested by means of the Fisher exact test, all SNPs in the *TNMD* showed no significant P-values and thus equal allele frequencies across sex groups (Table 1). On the contrary, the SNP in the *SLC6A14* did not ($P = 0.01$). This fact should be taken into consideration when selecting an appropriate test for high-level statistical analyses unless a balanced sex ratio across experimental conditions is presented in the population (which is our case). Information regarding minor allele frequencies (MAFs) stratified by experimental condition for all candidate

CHR	SNP	BP	A1	MAF (All)	MAF (Females)	MAF (Males)	A2	P	OR
23	rs11798018	100584572	A	0.26	0.26	0.27	C	0.84	0.97
23	rs5966709	100589508	T	0.32	0.32	0.31	G	0.75	1.05
23	rs4828037	100590686	C	0.34	0.34	0.33	T	0.53	1.08
23	rs2073162	100594019	A	0.45	0.45	0.42	G	0.37	1.12
23	rs2073163	100594053	C	0.45	0.46	0.43	T	0.35	1.13
23	rs4828038	100596678	T	0.44	0.45	0.42	C	0.31	1.13
23	rs1155974	100598283	T	0.44	0.44	0.42	C	0.59	1.07
23	rs2011162	116459132	C	0.34	0.37	0.30	G	0.02*	1.36

Table 1. Allele frequencies in the whole study population and each sex group. P and OR columns correspond to P-values and odd ratios obtained by means of the Fisher exact test for sex-specific allele frequencies (sex differences in allele frequency per SNP). *P < 0.05. Abbreviations; CHR, Chromosome; SNP, Single Nucleotide Polymorphism; BP, Base Pair; A1, Minor Allele; MAF, Minor Allele Frequency; A2, Alternative Allele.

CHR	SNP	GROUP	A1	MAF
23	rs11798018	OVERWEIGHT	A	0.26
23	rs11798018	OBESITY	A	0.27
23	rs11798018	NORMAL-BMI	A	0.28
23	rs5966709	OVERWEIGHT	T	0.29
23	rs5966709	OBESITY	T	0.34
23	rs5966709	NORMAL-BMI	T	0.30
23	rs4828037	OVERWEIGHT	C	0.32
23	rs4828037	OBESITY	C	0.36
23	rs4828037	NORMAL-BMI	C	0.31
23	rs2073162	OVERWEIGHT	A	0.44
23	rs2073162	OBESITY	A	0.46
23	rs2073162	NORMAL-BMI	A	0.43
23	rs2073163	OVERWEIGHT	C	0.45
23	rs2073163	OBESITY	C	0.47
23	rs2073163	NORMAL-BMI	C	0.43
23	rs4828038	OVERWEIGHT	T	0.43
23	rs4828038	OBESITY	T	0.46
23	rs4828038	NORMAL-BMI	T	0.43
23	rs1155974	OVERWEIGHT	T	0.42
23	rs1155974	OBESITY	T	0.45
23	rs1155974	NORMAL-BMI	T	0.43
23	rs2011162	OVERWEIGHT	C	0.34
23	rs2011162	OBESITY	C	0.36
23	rs2011162	NORMAL-BMI	C	0.34

Table 2. Allele frequencies in the study population stratified by experimental condition. Abbreviations; CHR, Chromosome; SNP, Single Nucleotide Polymorphism; BMI, Body Mass Index; A1, Minor Allele; MAF, Minor Allele Frequency.

markers is presented in (Table 2). Linkage disequilibrium (LD) status of the *TNMD* gene was studied using the Haploview Software separately in males and females^{13,18}.

Data Records

The complete research dataset (genotype and phenotype data) has been uploaded into the European Genome-Phenome archive (EGA). The work can be found online with the title “X chromosomal genetic variants are associated with childhood obesity” or with the identifier EGAS00001002738 (2018)¹⁵. Online data are sorted and presented according to obesity status; the affected group (cases) composed of both children with obesity or overweight (EGA reference EGAD00010001482 (2018)) and the control group composed of normal-weight children (EGA reference EGAD00010001481 (2018)). Three files by-experimental condition (a total of six) are available online (*bed*, *bim* and *fam* files). The *bed* files contain raw genotype data while the *bim* files describe information relative to target SNPs (chromosome number, SNP identifier, genetic distance in morgans (set as 0 for all markers), base-pair position and coding alleles). Instead, the *fam* files contain information relative to subjects (sample identifiers, family and paternal identifiers (here set as 0), sex (1 for males and 2 for females) and experimental group (1

for control and 2 for cases)). All presented formats can be easily readable in PLINK 1.9 software using the *-bfile* command option and further transformed into a more standard file format with the *-dosage* option¹⁹.

The complete data set in the current study complies with the requirements of the EGA archive. Detailed information about each sample and shared data files is presented in Online-only Tables 1 and 2, and Supplementary File 1. Specifically, DOI and descriptions for each shared file are provided in the Online-only Table 2.

Technical Validation

X chromosome particularities. Before introducing further steps, we here list two issues making the X chromosome a difficult region for genetic analyses. These particularities will determine important decisions related to genotype calling, data imputation and statistical analysis. It is important to note, however, that all here-described particularities are only applicable to those X chromosomal loci outside the pseudo-autosomal region of the X chromosome (which is the case of *TNMD* and *SLC6A14*).

The first noticeable uniqueness of the X chromosome is the fact of women having two allele copies while males having only one. As a result, while females can present the standard three possible allele combinations (AA, AB and BB), males are homozygous and have only two distinct possible genotypes (A- and B-). For this reason, standard autosomal association tests, such as the Cochran-Armitage trend test^{20,21}, are not immediately applicable to X chromosome data. The second particularity affecting the X-chromosome analysis lies in the X chromosome inactivation (XCI) process, through which the transcription from one of the two X chromosome copies in female mammalian cells is silenced in order to balance the expression dosage between XX females and XY males. XCI is, however, incomplete in humans: with up to one-third of the X-chromosomal genes escaping from this silencing epigenetic mechanism. The degree of 'escape' from inactivation has been reported to strongly vary between genes, tissues and individuals^{22,23}, with three possible scenarios at the gene level: complete XCI, partial XCI or total escape from XCI^{24,25}. Depending on the XCI model assumed for a certain gene, we should proceed one way or another during the selection of the test statistics (see section 'High-Level Analysis: Statistical Analysis' for further details). The assumption of a particular XCI model is therefore a process that must be performed carefully.

Until date, the extent to which XCI is shared between cells and tissues remains poorly characterized and there is a lack of standardized criteria nor well-established databases to check if a gene escapes or not from XCI in a concrete situation. In order to do so, an exhaustive search in PUBMED and other scientific databases should be performed looking for particular studies supporting a certain XCI hypothesis. Currently, the most similar resource to a standardized database on this regard is the initiative carried out by the Genotype-Tissue Expression (GTEx) consortium⁹ in 2017, which describes a systematic survey of XCI, integrating over 5500 transcriptomes from 449 Individuals, spanning 29 tissues from the GTEx (v6p release) and 940 single-cell transcriptomes, combined with genomic sequence data. Particularly, they show that XCI at 683 X-chromosomal genes is generally uniform across human tissues and that incomplete XCI affects at least 23% of X-chromosomal genes. Overall, this work presents an updated catalogue of XCI across human tissues which may be of great utility during the selection of a particular XCI model for a gene. Other available resources also include the work of Slavney *et al.*²⁶, which gathers the main XCI insights from previous studies on X-chromosome gene expression datasets.

By way of example, we here illustrate the whole process followed for the identification of the optimal XCI model in the case of *TNMD*. First, we interrogated the Slavney *et al.* (2016) work²⁶, where no evidence of escape from XCI was reported. In order to get more information about this fact, we further studied in detail the three works summarized in the Slavney *et al.*²⁶ paper. The first work on which the paper is based is a study from Carrel *et al.*²², in which we could not identify any probe covering the *TNMD* region. Instead, a few surrounding regions were mapped; among which the *SRPX2*, *ZD89B07* and the *SYTL4* reported escaping from the XCI process. In spite of it, this study was based on a fibroblast cell model and thus not applicable to our adipose tissue context. Regarding the second revised article²⁷, again, there were not available probes covering *TNMD*. Thus, neither conclusions nor new information could be extracted. In relation to the third included article²⁸, we were not able to find any table or supplemental material showing an output list of the analysed regions. Next, we investigated the well-established work from the GTEx consortium⁹ and found that the XCI status of the *TNMD* region remains catalogued as *unknown* (Supplementary Tables S2 and S13 of *this paper*). As a complementary approach, we performed a search in PUBMED looking for individual studies focused on the gene expression status of *TNMD* from different sexes. As a result, we found a work reporting higher basal expression of *TNMD* in women than in men²⁹, which could indicate that *TNMD* escapes from the XCI.

Taking all this into consideration and given the lack of agreement, both possibilities ('escape from XCI' and 'XCI') should be tested in the case of *TNMD*. A searching process like this is highly recommendable to be done for any X chromosome locus before the selection of a particular statistical approach.

Raw data processing. The primary step of the data analysis consisted on the extraction of genotype calls from fluorescence array data and the construction of work data files for data manipulation and analysis. Details regarding the exact procedure for genotype calling, which is an important procedure in X-chromosomal analyses, are listed below ('Genotype Calling' section).

Once we obtained genotype calls for the 915 individuals, we generated standard format files (*.ped* and *.map*) transforming the ThermoFisher cloud-derived outputs from long to wide format using an own script in R environment³⁰. Finally, data were imported into PLINK 1.9 software¹⁹ and converted into binary format files using the *-make-bed* flag. These binary formats (*.bed*, *.bim* and *.fam*) are a more compact representation of the data that saves space and speeds up subsequent analyses.

Genotype calling. This is the first step of any primary genotype analysis and consists on the extraction of genotype calls from fluorescence array data at the SNP and individual level. Along with the test statistics selection procedure, the genotype calling process is an analytical step heavily affected by X chromosome particularities.

CHR	SNP	MISS FREQ (Males)	MISS FREQ (Females)	MISS FREQ (Males-Females)	P
23	rs11798018	0.11	0.04	0.07	2.01e-05
23	rs5966709	0.06	0.004	0.06	3.11e-07
23	rs4828037	0.06	0.01	0.05	3.46e-05
23	rs2073162	0.08	0.008	0.07	2.86e-08
23	rs2073163	0.14	0.09	0.05	0.02
23	rs4828038	0.07	0.002	0.07	4.19e-10
23	rs1155974	0.08	0.002	0.07	9.73e-11
23	rs2011162	0.07	0.02	0.05	9.30e-05

Table 3. Missing frequency quality control (QC) in our selected markers stratified by sex. P column correspond to differential missingness test between sex groups. Asymptotic p-values were obtained by means of Fisher's exact test. SNPs in bold did pass the QC recommended filters. Abbreviations; CHR, Chromosome; SNP, Single Nucleotide Polymorphism; MISS FREQ, Missing Frequency.

CHR	SNP	TEST	A1	GENO	O(HET)	E(HET)	P
23	rs11798018	ALL	A	34/177/249	0.38	0.39	0.72
23	rs11798018	AFF	A	25/133/187	0.38	0.39	0.89
23	rs11798018	UNAFF	A	9/44/61	0.39	0.40	0.81
23	rs5966709	ALL	T	67/175/234	0.37	0.44	0.0005
23	rs5966709	AFF	T	55/128/171	0.36	0.45	0.0005
23	rs5966709	UNAFF	T	12/46/63	0.38	0.41	0.38
23	rs4828037	ALL	C	75/178/219	0.38	0.45	0.0003
23	rs4828037	AFF	C	63/129/158	0.37	0.46	0.0001
23	rs4828037	UNAFF	C	12/48/61	0.40	0.42	0.66
23	rs2073162	ALL	A	132/172/170	0.36	0.50	4.59e-09
23	rs2073162	AFF	A	99/128/124	0.36	0.50	6.78e-07
23	rs2073162	UNAFF	A	33/43/46	0.35	0.49	0.002
23	rs2073163	ALL	C	129/148/156	0.34	0.50	6.92e-011
23	rs2073163	AFF	C	98/112/113	0.35	0.50	3.83e-08
23	rs2073163	UNAFF	C	31/35/43	0.32	0.49	0.0002
23	rs4828038	ALL	T	133/171/173	0.36	0.50	1.56e-09
23	rs4828038	AFF	T	100/128/127	0.36	0.50	2.49e-07
23	rs4828038	UNAFF	T	33/43/46	0.35	0.49	0.002
23	rs1155974	ALL	T	127/172/178	0.36	0.49	4.18e-09
23	rs1155974	AFF	T	94/128/132	0.36	0.49	4.02e-07
23	rs1155974	UNAFF	T	33/43/46	0.35	0.49	0.002
23	rs2011162	ALL	C	84/179/207	0.38	0.47	0.0001
23	rs2011162	AFF	C	62/136/150	0.39	0.47	0.003
23	rs2011162	UNAFF	C	22/42/57	0.35	0.46	0.01

Table 4. Genotype counts and Hardy-Weinberg test statistics for each SNP in the female group. Each SNP has three entries showing results for either ALL individuals, AFF (overweight and children with obesity) or UNAFF (normal-BMI children only). Hardy Weinberg analysis was performed with the exact test described and implemented by Wigginton *et al.*⁴². Abbreviations; CHR, Chromosome; SNP, Single Nucleotide Polymorphism; A1, minor allele; GENO, genotype counts; O(HET), observed heterozygosity; E(HET), expected heterozygosity; P, hardy weinberg P-value.

Specifically, the main X chromosome uniqueness affecting this process is the dosage imbalance between males and females. Since males carry only one X allele, signal intensities obtained from the Real-Time PCR System are lower in males than for females and thus a correction should be implemented. On this matter, calling algorithms which apply different models to male and female samples (e.g. Illuminus and CRLMM) have been proven to generally perform better than methods which do not (e.g. GenCall and GenoSNP)³¹.

Here, we employed the Applied Biosystems qPCR app module (Thermo Fisher Cloud software) and the auto-calling method for genotype calling. According to literature recommendations, the sex information for each sample was supplied to the software and genotype calling was performed separately in both sexes. In this regard, although genotyped plates did not consist on only boys or girls, the balanced sex ratio of our population (477 females and 438 males) favoured a better performance of the algorithm. Five signal clusters were identified (three in the case of females and two in the case of males). Then, sex information and scatter of the clusters were used to call the genotypes (AA, AB and BB for females, and A- and B- for males). Since the employed software also allows

the option of applying user-definable boundaries for data analysis, those samples classified as undetermined by the autocalling method were recalled using the manual option. A set of controls were used to deduce these questionable genotype calls. Outliers were omitted from the analysis.

Data QC. Prior to high-level statistical analyses, the quality control (QC) process is an important step in any genetic analysis and especially in the X-chromosome analysis. Specific QC guidelines for X chromosome genotype data have been previously reviewed in detail⁸. All these criteria can help us to detect genotype errors or not reliable SNPs which should be excluded from analysis.

Here, the whole QC process was implemented in PLINK 1.9 software¹⁹. According to literature, two criteria concerning missing frequency were employed (the sex-specific missing frequency and the differential missingness between sexes)^{32,33}. As genotype calling was performed separately in males and females (that is, no heterozygote calls in males were allowed), the proportion of heterozygote calls in males, proposed as a filter criterion by Ling and Ziegler *et al.*^{32,33}, was not considered in our QC process. All SNPs (with exception of the rs11798018 and the rs2073163 from the *TNMD* gene) passed the recommended missing frequency filter in females ($\leq 2\%$) (Table 3). On the other hand, none SNP passed the filter in males. Regarding the differential missingness test, the SNPs (rs11798018, rs4828037 and rs2073163) from the *TNMD* and the rs2011162 from the *SLC6A14*, passed the recommended filter ($P \geq 10^{-7}$). The other SNPs, instead, evidenced a marked differential missingness between sex groups. This test was performed in PLINK software using the flag “*test-missing*” and replacing the phenotype column of the *ped* file by the sex information (Table 3).

Regarding additional MAF quality checks, all SNPs showed appropriated frequencies $>1\%$ by sex groups (Table 1). When analysing the Hardy Weinberg equilibrium (HWE) in girls belonging to the normal-BMI group, all SNPs reported proper values ($P \geq 10^{-4}$) (Table 4). According to this QC process, we ensured that there were not important genotyping errors and that our genetic data were reliable for further analyses.

On this point, it is important to note that since genotyping array technologies are not specially designed for sexual chromosomes, quality is always hoped to be lower on X chromosome genetic variants compared to autosomal data.

High-level analysis: statistical analysis. As we previously mentioned, most of available test statistics for performing genetic association analyses have been designed for autosomal variants and thus they are not applicable to X chromosome data (especially when dealing with mixed-sex samples). In these cases, testing for association on the X chromosome raises unique challenges that have motivated the development of X-specific statistical tests in the literature^{34,35}. Association tests on the X chromosome should incorporate into their models not only the fact of dosage imbalance between males and females but also, depending on the analysed locus, a specific XCI model. Some of available approaches include:

- Clayton Tests (2008)³⁴. Clayton tests are two X chromosome specific versions of the common autosomal tests that explicitly account for the XCI process and allow the inclusion of males and females together. In the case of different allele frequencies in males and females, Clayton statistics have inflated type I error frequencies. These tests are available in the R package *snpMatrix*³⁶ with the names:
 - S1³⁴: It is analogous to a Cochran-Armitage trend test of a combined male and female genotype contingency table; it follows a χ^2 distribution on one degree freedom (df) under the null hypothesis.
 - S2³⁴: It is analogous to a Pearson's χ^2 test on 2 df of a combined male and female genotype contingency table, it follows a χ^2 distribution on 2 df under the null hypothesis.
- Zheng tests (2007)³⁵. They are a set of six different statistics that apply to the same SNP and from which a minimum *P*-value is computed, needing to be adjusted according to the correlation between the test statistics. Zheng *et al.*³⁵ showed that the optimal choice of statistic among the six tests depends on whether HWE holds at the locus and whether males and females have the same risk allele. For example, in the case there is departure from HWE in females, the Zheng (Z^2_{mFG}) test has been presented a good choice. For further information regarding test statistic selection, we recommend to read works^{8,35}. Of note is that the Zheng's tests do not explicitly account for the XCI process.

As previously mentioned, an unbalanced female/male ratio between cases and control would affect the relative power of both Zheng and Clayton statistics. If combined with sex-specific allele frequencies, these tests will suffer from increased type I errors.

- Traditional methods easily implementable in PLINK 1.9 or R environment:
 - Ignore males entirely and analyse female data using conventional autosomal tests (a genotypic-based Cochran-Armitage trend test or an allele-based χ^2 by Pearson with 1 df). The problem related to this approach is that we are missing all data from male subjects and therefore losing statistical power. The Cochran-Armitage trend test is the default test employed when a naive analysis of X chromosome data is run in PLINK using the flag *-model*¹⁹. Regarding males, an allele-based test accounting for the number of A- and B- alleles between experimental conditions should be employed apart.
 - Linear or logistic regression analyses on all the samples adjusting by sex. This approach further has the advantage of adjusting the model by covariates of interest. Here, if we assume that the locus of interest escapes from XCI, females should be coded as 0, 1, or 2, according to 0, 1, or 2 number of SNP risk alleles, and males should be coded as 0 or 1 according to 0 or 1 allele copies. On the contrary, if XCI is

SNP	N	Chi.squared.1.df	Chi.squared.2.df	P.1df	P.2df
HOMA-IR					
rs11798018	811	0.10	1.68	0.74	0.43
rs5966709	849	0.14	2.71	0.70	0.25
rs4828037	844	0.35	2.91	0.55	0.23
rs2073162	841	5.48	6.34	0.01	0.04
rs2073163	773	4.78	5.93	0.02	0.05
rs4828038	849	6.00	7.24	0.01	0.02
rs1155974	844	4.22	6.68	0.03	0.03
rs2011162	839	0.48	0.91	0.48	0.63
Glucose (mg/dl)					
rs11798018	844	0.004	1.06	0.94	0.58
rs5966709	881	0.55	0.59	0.45	0.74
rs4828037	876	1.22	1.25	0.26	0.53
rs2073162	873	5.17	8.13	0.02	0.01
rs2073163	804	2.8	4.006	0.09	0.13
rs4828038	880	4.78	6.42	0.02	0.04
rs1155974	876	3.94	4.74	0.04	0.09
rs2011162	871	0.92	3.55	0.33	0.16
BMI z-score					
rs11798018	845	0.97	1.15	0.32	0.56
rs5966709	881	0.77	1.21	0.37	0.54
rs4828037	877	0.51	0.51	0.47	0.77
rs2073162	872	8.61	9.59	0.003	0.008
rs2073163	803	7.09	8.60	0.007	0.01
rs4828038	877	9.02	10.38	0.002	0.005
rs1155974	875	7.75	8.69	0.005	0.01
rs2011162	871	3.31	5.33	0.06	0.06

Table 5. Association between X chromosome SNPs and HOMA-IR, Glucose and BMI z-score in our dataset. SNPs in bold showed statistically significant associations with presented phenotypes under Clayton Statistics. This test explicitly accounts for random X-inactivation and allows the inclusion of females and males together, increasing thereby the statistical power. P.1df and Chi.squared.1.df columns corresponds to Clayton S1 increasing results while P.2df and Chi.squared.2.df corresponds to Clayton S2 statistic. Abbreviations; SNP, Single Nucleotide Polymorphism; N, number of included subjects in the analysis; HOMA-IR, homeostasis model assessment for insulin resistance; BMI z-score, body mass index adjusted by sex and age.

assumed to occur, females should be coded as 0, 1, or 2, according to 0, 1, or 2 number of SNP risk alleles, and males should be coded as 0 or 2 according to 0 or 1 allele copies. By default, the application of the “-dosage” flag to X chromosome input data files (*.bed*, *.bim* and *.fam*) in PLINK will produce a codification which assumes escape from XCI. For XCI to be considered, new allele code numbers should be manually replaced in male samples with a standard text editor (e.g: gedit software).

In general, the selection of the most suitable test among the presented choices will depend on three different criteria; the XCI model assumed for the locus of interest, deviation from HWE of analysed markers and the existence of sex-specific allele frequencies in the study population, which would be a substantial problem in the case of an unbalanced female/male ratio. Regarding XCI, if inactivation is assumed to occur, then either the Clayton's statistics or regression models (with males coded as 0 and 2 (for 0 and 1 risk allele, respectively)) would be the tests of choice. On the contrary, in the case of a locus ‘escaping’ from XCI, Zheng's tests or regression models (with males coded as 0 and 1 (for 0 and 1 risk allele, respectively)) should be employed. In the case of sex-specific allele frequencies, independently of the XCI assumed model, the Zheng's test (Z^2_{mfc}) has been presented a better choice over the Clayton approach. On the other hand, in the case of an adjustment for covariates is required, only regression models can be applied. Of note is that most of the test statistics and analysis considerations covered here are available to implement in the command-line toolset *XWAS* developed by Keinan A. and collaborators^{37–39}.

Although for the analysis of our dataset both possibilities (‘escape from XCI’ and ‘XCI’) were tested in the original work¹³, we here only present results under the XCI assumption. As we have previously seen, selected markers in our sample did not exhibit HWE deviations nor sex-specific allele frequencies. Moreover, the female/male ratio was balanced across experimental groups. For these reasons, and following published recommendations^{8,17,34,40}, Clayton test was here selected to perform the main statistical analysis. According to an *in silico* simulation work, the Clayton's S1 statistic has shown the best performance among all X-specific introduced tests across a wide range of disease models, sex ratios and allele frequencies⁴⁰. Moreover, it allows the inclusion of females and males together, increasing thereby the statistical power.

In Table 5, results derived from the application of Clayton's S1 and S2 statistics to three different continuous phenotypes of the population are presented. All these phenotype data have also been shared and are available in the metadata file (Online-only Table 1). The implementation of this process was performed in R, using the `snpStats` R package and the code have been shared online⁴¹. All reported associations in our previous work¹³ were here replicated under XCI assumption. These findings support therefore a good performance of the Clayton statistics as well as ensure the reliability of the present dataset.

In conclusion, we here share a genetic dataset and present a whole implementation of the special X chromosome analytic process in genetics. Altogether, the pipeline and the shared data will allow researchers to get familiar with the X chromosome particularities and should encourage them to include X chromosome into their genetic studies. Closing this gap is crucial to elucidate the genetic background of complex diseases, especially of those with sex-specific features.

Code Availability

All custom R codes employed in this work have been shared online in a GitHub repository (10.5281/zenodo.2578182)⁴¹. Two short scripts are available online; “`script_from_long_to_wide.r`” and “`Clayton_analysis_code.r`”.

The first one (named “`script_from_long_to_wide.r`”) is a short script designed for loading a genetic dataset (genotype calls) derived from OpenArray technology and transforming it into a handy-format file, which can be further imported into PLINK software. Basically, this script carries out a dataset manipulation and transformation from long to wide format. In order to run the script, users will need an input file derived from OpenArray technology containing information in the long format arranged into three columns (NCBI_SNP_Reference, Sample_ID and Genotype_Call).

The second script shared (named “`Clayton_analysis_code.r`”) gathers functions and R commands required for the application of the X-chromosome specific statistical tests developed by Clayton and collaborators^{34,36} (see section ‘High-Level Analysis: Statistical Analysis’ for further details).

References

1. Collaborators, G. B. D. O. *et al* Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *The New England journal of medicine* **377**, 13–27, <https://doi.org/10.1056/NEJMoa1614362> (2017).
2. Jones, R. E., Jewell, J., Saksena, R., Ramos Salas, X. & Breda, J. Overweight and Obesity in Children under 5 Years: Surveillance Opportunities and Challenges for the WHO European Region. *Frontiers in public health* **5**, 58, <https://doi.org/10.3389/fpubh.2017.00058> (2017).
3. Maes, H. H., Neale, M. C. & Eaves, L. J. Genetic and environmental factors in relative body weight and human adiposity. *Behavior genetics* **27**, 325–351 (1997).
4. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *American journal of human genetics* **90**, 7–24, <https://doi.org/10.1016/j.ajhg.2011.11.029> (2012).
5. Zaitlen, N. *et al*. Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS genetics* **9**, e1003520, <https://doi.org/10.1371/journal.pgen.1003520> (2013).
6. Locke, A. E. *et al*. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206, <https://doi.org/10.1038/nature14177> (2015).
7. Wise, A. L., Gyi, L. & Manolio, T. A. eXclusion: toward integrating the X chromosome in genome-wide association analyses. *American journal of human genetics* **92**, 643–647, <https://doi.org/10.1016/j.ajhg.2013.03.017> (2013).
8. König, I. R., Loley, C., Erdmann, J. & Ziegler, A. How to include chromosome X in your genome-wide association study. *Genetic epidemiology* **38**, 97–103, <https://doi.org/10.1002/gepi.21782> (2014).
9. Tukiainen, T. *et al*. Landscape of X chromosome inactivation across human tissues. *Nature* **550**, 244–248, <https://doi.org/10.1038/nature24265> (2017).
10. Tolppanen, A. M. *et al*. Tenomodulin is associated with obesity and diabetes risk: the Finnish diabetes prevention study. *Obesity* **15**, 1082–1088, <https://doi.org/10.1038/oby.2007.613> (2007).
11. Aguilera, C. M. *et al*. Genome-wide expression in visceral adipose tissue from obese prepubertal children. *International journal of molecular sciences* **16**, 7723–7737, <https://doi.org/10.3390/ijms16047723> (2015).
12. Aguilera, C. M. *et al*. Differential gene expression in omental adipose tissue from obese children. *Gene Expression Omnibus*, <https://identifiers.org/geo/GSE9624> (2018).
13. Ruiz-Ojeda, F. J. *et al*. Effects of X-chromosome Tenomodulin genetic variants on obesity in a children's cohort and implications of the gene in adipocyte metabolism. *Scientific Reports*. <https://doi.org/10.1038/s41598-019-40482-0> (2019).
14. Suviolahti, E. *et al*. The SLC6A14 gene shows evidence of association with obesity. *The Journal of clinical investigation* **112**, 1762–1772, <https://doi.org/10.1172/JCI17491> (2003).
15. Anguita-Ruiz, A., Ruiz-Ojeda, F. J. & Aguilera, C. M. X chromosomal genetic variants are associated with childhood obesity. *European Genome-phenome Archive* <https://identifiers.org/ega.study:EGAS00001002738> (2018).
16. Cole, T. J., Bellizzi, M. C., Flegal, K. M. & Dietz, W. H. Establishing a standard definition for child overweight and obesity worldwide: international survey. *Bmj* **320**, 1240–1243, <https://doi.org/10.1136/bmj.320.7244.1240> (2000).
17. Loley, C., Ziegler, A. & König, I. R. Association tests for X-chromosomal markers—a comparison of different test statistics. *Human heredity* **71**, 23–36, <https://doi.org/10.1159/000323768> (2011).
18. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265, <https://doi.org/10.1093/bioinformatics/bth457> (2005).
19. Purcell, S. *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559–575, <https://doi.org/10.1086/519795> (2007).
20. Cochran, W. G. Some methods for strengthening the common χ^2 tests. *Biometrics* **10**, 417–451, <https://doi.org/10.2307/3001616> (1954).
21. Armitage, P. Tests for Linear Trends in Proportions and Frequencies. *Biometrics* **11**, 375–386, <https://doi.org/10.2307/3001775> (1955).
22. Carrel, L. & Willard, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**, 400–404, <https://doi.org/10.1038/nature03479> (2005).
23. Cotton, A. M. *et al*. Analysis of expressed SNPs identifies variable extents of expression from the human inactive X chromosome. *Genome biology* **14**, R122, <https://doi.org/10.1186/gb-2013-14-11-r122> (2013).
24. Chow, J. C., Yen, Z., Ziesche, S. M. & Brown, C. J. Silencing of the mammalian X chromosome. *Annual review of genomics and human genetics* **6**, 69–92, <https://doi.org/10.1146/annurev.genom.6.080604.162350> (2005).

25. Amos-Landgraf, J. M. *et al.* X chromosome-inactivation patterns of 1,005 phenotypically unaffected females. *American journal of human genetics* **79**, 493–499, <https://doi.org/10.1086/507565> (2006).
26. Slavney, A., Arbiza, L., Clark, A. G. & Keinan, A. Strong Constraint on Human Genes Escaping X-Inactivation Is Modulated by their Expression Level and Breadth in Both Sexes. *Molecular biology and evolution* **33**, 384–393, <https://doi.org/10.1093/molbev/msv225> (2016).
27. Cotton, A. M. *et al.* Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Human molecular genetics* **24**, 1528–1539, <https://doi.org/10.1093/hmg/ddu564> (2015).
28. Schultz, M. D. *et al.* Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature* **523**, 212–216, <https://doi.org/10.1038/nature14465> (2015).
29. Kolehmainen, M. *et al.* Weight reduction modulates expression of genes involved in extracellular matrix and cell death: the GENOBIN study. *International journal of obesity* **32**, 292–303, <https://doi.org/10.1038/sj.ijo.0803718> (2008).
30. R Development Core Team. R: a language and environment for statistical computing, <https://doi.org/3-900051-07-0> (2011).
31. Ritchie, M. E. *et al.* Comparing genotyping algorithms for Illumina's Infinium whole-genome SNP BeadChips. *BMC bioinformatics* **12**, 68, <https://doi.org/10.1186/1471-2105-12-68> (2011).
32. Ling, H., Hetrick, K., Bailey-Wilson, J. E. & Pugh, E. W. Application of sex-specific single-nucleotide polymorphism filters in genome-wide association data. *BMC proceedings* **3**(Suppl 7), S57, <https://doi.org/10.1186/1753-6561-3-S7-S57> (2009).
33. Ziegler, A. Genome-wide association studies: quality control and population-based measures. *Genetic epidemiology* **33**(Suppl 1), S45–50, <https://doi.org/10.1002/gepi.20472> (2009).
34. Clayton, D. Testing for association on the X chromosome. *Biostatistics* **9**, 593–600, <https://doi.org/10.1093/biostatistics/kxn007> (2008).
35. Zheng, G., Joo, J., Zhang, C. & Geller, N. L. Testing association for markers on the X chromosome. *Genetic epidemiology* **31**, 834–843, <https://doi.org/10.1002/gepi.20244> (2007).
36. Clayton, D. *snpStats*: SnpMatrix and XSnpmatrix classes and methods. R package version 1.32.0. (2018).
37. Gao, F. *et al.* XWAS: A Software Toolset for Genetic Data Analysis and Association Studies of the X Chromosome. *The Journal of heredity* **106**, 666–671, <https://doi.org/10.1093/jhered/esv059> (2015).
38. Chang, D. *et al.* Accounting for eXcentricities: analysis of the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases. *PLoS one* **9**, e113684, <https://doi.org/10.1371/journal.pone.0113684> (2014).
39. Ma, L., Hoffman, G. & Keinan, A. X-inactivation informs variance-based testing for X-linked association of a quantitative trait. *BMC genomics* **16**, 241, <https://doi.org/10.1186/s12864-015-1463-y> (2015).
40. Hickey, P. F. & Bahlo, M. X chromosome association testing in genome wide association studies. *Genetic epidemiology* **35**, 664–670, <https://doi.org/10.1002/gepi.20616> (2011).
41. Anguita-Ruiz, A. R scripts for the manipulation, transformation and statistical analysis of Openarray genotype datasets. (Version 1.0. 2). *Zenodo*, <https://doi.org/10.5281/zenodo.2578182> (2019).
42. Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A note on exact tests of Hardy-Weinberg equilibrium. *American journal of human genetics* **76**, 887–893, <https://doi.org/10.1086/429864> (2005).

Acknowledgements

This paper will be part of Augusto Anguita-Ruiz's doctorate, which is being completed as part of the “Nutrition and Food Sciences Program” at the University of Granada, Spain. The authors would like to thank the children and parents who participated in the study. This work was supported by the Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica (I+D+I), Instituto de Salud Carlos III-Fondo de Investigación Sanitaria (FONDOS FEDER) Projects numbers PI051968, PI1102042 and PI1600871, Redes temáticas de investigación cooperativa RETIC (Red SAMID RD12/0026/0015) and the Mapfre Foundation. The authors also acknowledge the Institute of Health Carlos III for personal funding: “Contratos i-PFIS: doctorados IIS-empresa en ciencias y tecnologías de la salud de la convocatoria 2017 de la Acción Estratégica en Salud 2013–2016, Project number: IFI17/00048”.

Author Contributions

A.G. and C.A.G. contributed to the study concept and design. R.L., G.B., M.G.C., R.V.C., L.M. and R.C. participated in the child recruitment and anthropometric measures. A.I.R. and C.A.G. selected the SNPs and revised the DNA extraction. A.A.R. did the all data processing and analysis steps and shared the dataset in the EGA repository. All authors took part in the interpretation of data, the drafting of the manuscript and the critical revision of the manuscript. A.G., C.A.G., R.L., G.B. and R.C. obtained funding. A.A.R., F.J.R.O. and J.P.D. wrote the manuscript.

Additional Information

Supplementary Information is available for this paper at <https://doi.org/10.1038/s41597-019-0109-3>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2019