Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/5/8/eaaw7195/DC1

Supplementary Materials for

Data-driven phenotype discovery of *FMR1* premutation carriers in a population-based sample

Arezoo Movaghar, David Page, Murray Brilliant, Mei Wang Baker, Jan Greenberg, Jinkuk Hong, Leann Smith DaWalt, Krishanu Saha, Finn Kuusisto, Ron Stewart, Elizabeth Berry-Kravis, Marsha R. Mailick*

*Corresponding author. Email: marsha.mailick@wisc.edu

Published 21 August 2019, *Sci. Adv.* **5**, eaaw7195 (2019)
DOI: 10.1126/sciadv.aaw7195

The PDF file includes:

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/5/8/eaaw7195/DC1)

Supplementary Text

Burden of disease

We used various age thresholds (40, 60, 80 and lifetime) to discover the variables that differentiated premutation carriers from controls, and created an independent model for each threshold. Then we used mean decrease in impurity based on Gini score (MDG), to identify the variables that are more influential in creating each model. We examined the participant EHR data in terms of three indicators of the burden of disease: 1) the percentage of cases and controls who received these diagnoses, 2) the number of medical encounters for those conditions, and 3) the age of participants when they received the diagnoses for the first time. Tables provided in Data S1 and S2 list the values for the burden of disease criteria for the 100 variables with highest MDG scores identified in each model. Examination of the results indicated that for all of these variables, premutation carriers had either a higher percentage of receiving the diagnosis, a higher frequency of medical encounters, or were diagnosed at a younger age compared to the controls. Figures S1-S3 show the distribution of three indicators of the burden of disease for the 100 variables with the highest MDG for male and female participants.

Phenotypic associations

We used Phenome-Wide Association Study (PheWAS) approach, separately for females and for males, to examine the phenotypic association of clinical phenotypes and *FMR1* premutation. We used the same age thresholds as above for this analysis (40, 60, 80 and lifetime) and created an independent model for each threshold. Data D3-S4 show the phecodes that had significant associations ($p<0.05$) with *FMR1* premutation. Figures S4-S6 illustrate the Manhattan plot of unadjusted p-values for phecodes. The phecodes with p-value <0.01 are annotated in the figures and the list of phecodes that had significant association with premutation (p-value between 0.01-0.05) are listed in the captions.

Genetic Data

The number of *FMR1* CGG repeats was determined for all samples using a PCR-based protocol that incorporated reagents developed and manufactured by Celera Corporation (Alameda, CA). The protocol combined gene-specific primers that flank the CGG repeat region of the *FMR1* gene with gender-specific primers, a polymerase mixture, and a reaction buffer that is optimized for amplification of GC-rich DNA. The PCR reactions were carried out in 96-well plates, and each well contained a 20 µl reaction volume that consisted of 13 µl of High GC PCR buffer, 0.8 µl of *FMR1* primers, 0.6 µl of gender primers, 1.2 µl of TR PCR Enzyme Mix, 1.4 µl of water, and 3 µl of 5–10 ng/µl DNA template. All reagents and the reaction plate were placed on ice throughout the duration of assay setting-up. PCR was performed on a ABI Veriti thermal cycler (Applied Biosystems, Grand Island, NY) using two sets of cycling parameters that were automatically implemented serially by the system. Conditions for the first set of 15 cycles were: 98.5 °C for 10 sec, 58 °C for 60 sec, and 75 °C for 6 min. Conditions for the second set of 15 cycles were: 98.5 °C with 0.1 °C increment per cycle for 10 sec, 56 °C for 60 sec, and 75 °C for 6 min.

Upon completion of PCR, 3 µl of CleanUp Enzyme Mix was added to 2 µl of PCR product to reduce the stutter (n – 1) signal typically observed with the amplification of GC-rich DNA targets. The mixture was incubated at 75 °C for 10 min followed by the addition of 17 µl of Hi-Di™ Formamide (Applied Biosystems) and 3 µl of ROX™ 1000 Size Standard (Celera Corporation). The function of ROX™ 1000 Size Standard was to accurately size all PCR products between 75 and 1,000 bp. This size standard consists of 20 single-stranded DNA fragments of known sizes each labeled with X-Rhodamine (ROX). The DNA fragment sizes are as follows: 50, 75, 100, 200, 300, 350, 400, 450, 475, 500, 550, 600, 650, 700, 750, 800, 850, 900, 950, and 1,000 bp. Samples were subsequently denatured at 93 °C for 30 sec before undergoing electrophoresis on an ABI 3730xl with POP-7® polymer using a 50-cm array. The sample was injected twice in order to accommodate optimal capillary electrophoresis conditions for a wide range of CGG repeats. ABI 3730xl run module settings are similar to those previously described*(20)*; the major differences are Injection Time (1 sec in injection 1 and 22 sec in injection 2 vs. 8 sec) and Run Time (4,000 sec in injection 1 and 5,500 sec in injection 2 vs. 2,000 sec). To ensure the consistency of assay performance, in each PCR/CE run, a size control DNA sample was also processed along with the testing samples. The size control samples used in this study are Coriell NA06891, NA06910, and NA20239.Data from each run were analyzed using GeneMapper® v. 4.0 (Applied Biosystems). CGG triplet repeats were calculated using the following formula: number of CGG repeats = (peak size – 193)/3.

The protocol also detects the presence of X and Y chromosomes within a sample. This information enabled sex confirmation and helped identify female samples with a single detectable CGG repeat (apparent homozygosity). Those samples were further evaluated by agarose gel electrophoresis for the presence of large full mutation alleles.

A genomic Coriell DNA sample with 645 CGG triplet repeat (NA04025) was incorporated in each PCR plate, and verified on agarose gel electrophoresis.

List of abreviations

AUROC: Area under the receiver operating characteristic curve
CGG: cytosine-guanine-guanine
FMR1: Fragile X mental retardation 1
FXS: Fragile X syndrome
FXTAS: Fragile X associated tremor/ataxia syndrome
FXPOI: Fragile X associated primary ovarian insufficiency
EHR: Electronic health record
ICD: International classification of disease
MDG: Mean decrease in impurity based on Gini
PMRP: Personalized medicine research project
PheWAS: Phenome-wide association study
ROC: Receiver operating characteristic curve
SNOMED: Systematized nomenclature of medicine

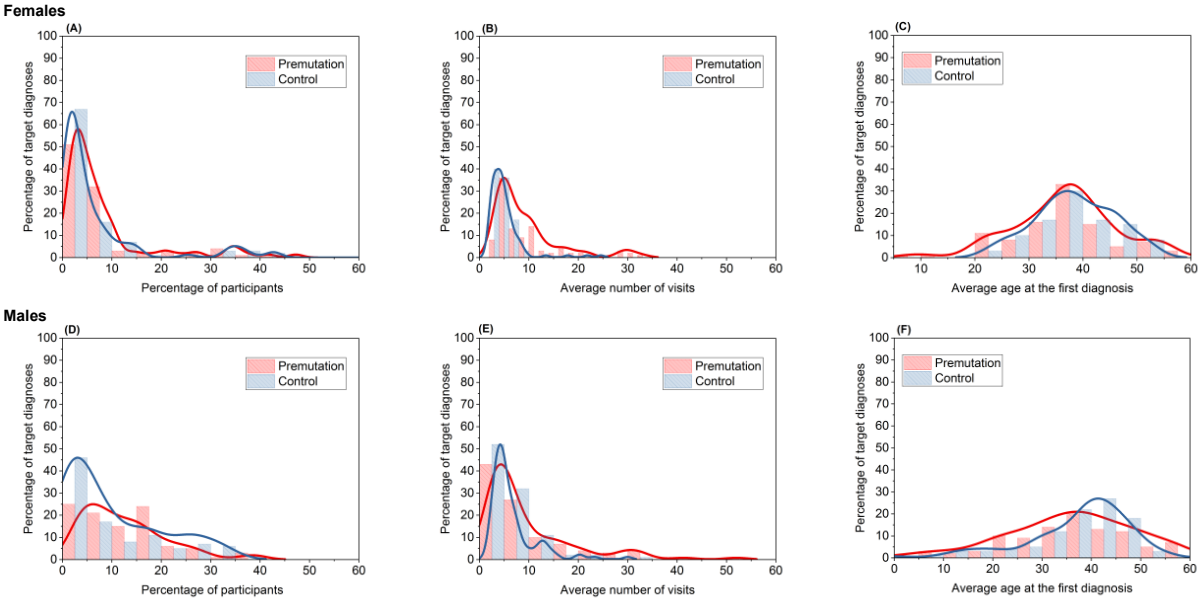


Fig. S1. The distribution of the target diagnoses in premutation carriers and controls for codes received before age 60. A) The prevalence of target diagnosis codes in females, with premutation carriers having higher percentage of receiving the codes. **B)** The average frequency of visit for target diagnoses in females, showing that premutation carriers had a higher number of visits for the differentiating diagnoses compared to the normal population. **C)** The average age in which participants received the target codes for the first time in females, illustrating that premutation carriers started to experience health/illness symptoms at younger ages than controls. **D)** The prevalence of target diagnosis codes in males, with premutation carriers having higher percentage of receiving multiple codes. **E)** The average frequency of visit for target diagnoses in males, showing that premutation carriers had a higher number of visits for the differentiating diagnoses compared to the normal population. **F)** The average age in which participants received the target codes for the first time in males illustrating premutation carriers started to experience health/illness symptoms at younger ages than controls.

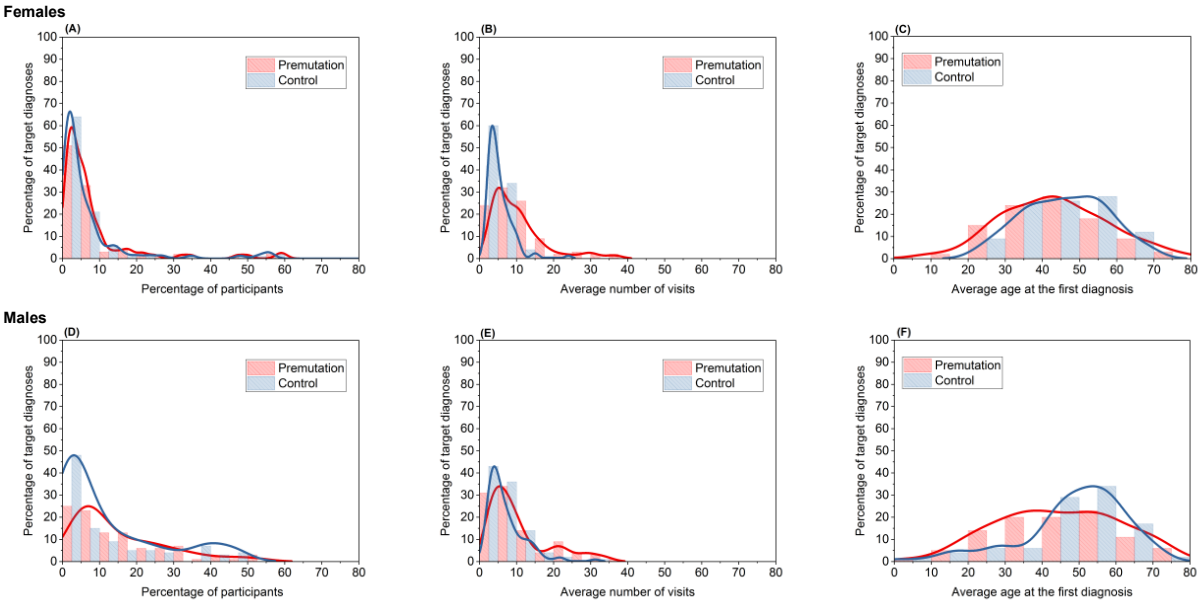


Fig. S2. The distribution of the target diagnoses in premutation carriers and controls for codes received before age 80. A) The prevalence of target diagnosis codes in females, with premutation carriers and controls having similar percentage of receiving the codes. **B)** The average frequency of visit for target diagnoses in females, showing that premutation carriers had a higher number of visits for the differentiating diagnoses compared to the normal population. **C)** The average age in which participants received the target codes for the first time in females, illustrating that premutation carriers started to experience health/illness symptoms at younger ages than controls. **D)** The prevalence of target diagnosis codes in males, with premutation carriers having higher percentage of receiving multiple codes. **E)** The average frequency of visit for target diagnoses in males, showing that premutation carriers had a higher number of visits for the differentiating diagnoses compared to the normal population. **F)** The average age in which participants received the target codes for the first time in males illustrating premutation carriers started to experience health/illness symptoms at younger ages than controls.

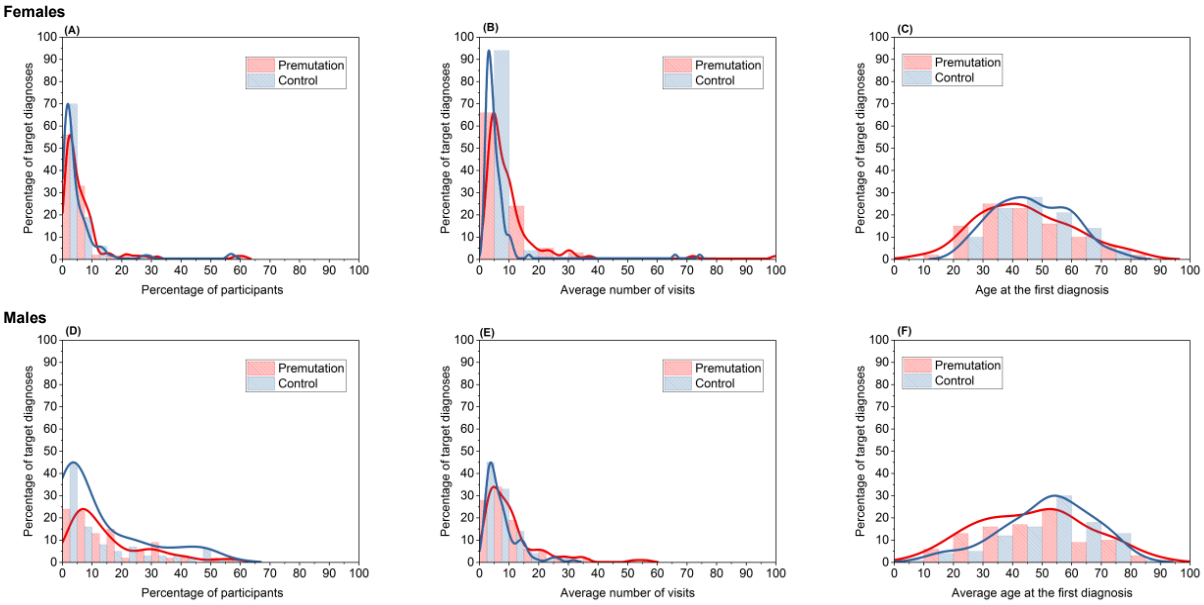


Fig. S3. The distribution of the target diagnoses in premutation carriers and controls based on lifetime diagnoses. A) The prevalence of target diagnosis codes in females, with premutation carriers having higher percentage of receiving the codes. **B)** The average frequency of visit for target diagnoses in females, showing that premutation carriers had a higher number of visits for the differentiating diagnoses compared to the normal population. **C)** The average age in which participants received the target codes for the first time in females, illustrating that premutation carriers started to experience health/illness symptoms at younger ages than controls. **D)** The prevalence of target diagnosis codes in males, with premutation carriers having higher percentage of receiving the codes. **E)** The average frequency of visit for target diagnoses in males, showing that premutation carriers had a higher number of visits for the differentiating diagnoses compared to the normal population. **F)** The average age in which participants received the target codes for the first time in males illustrating premutation carriers started to experience health/illness symptoms at younger ages than controls.

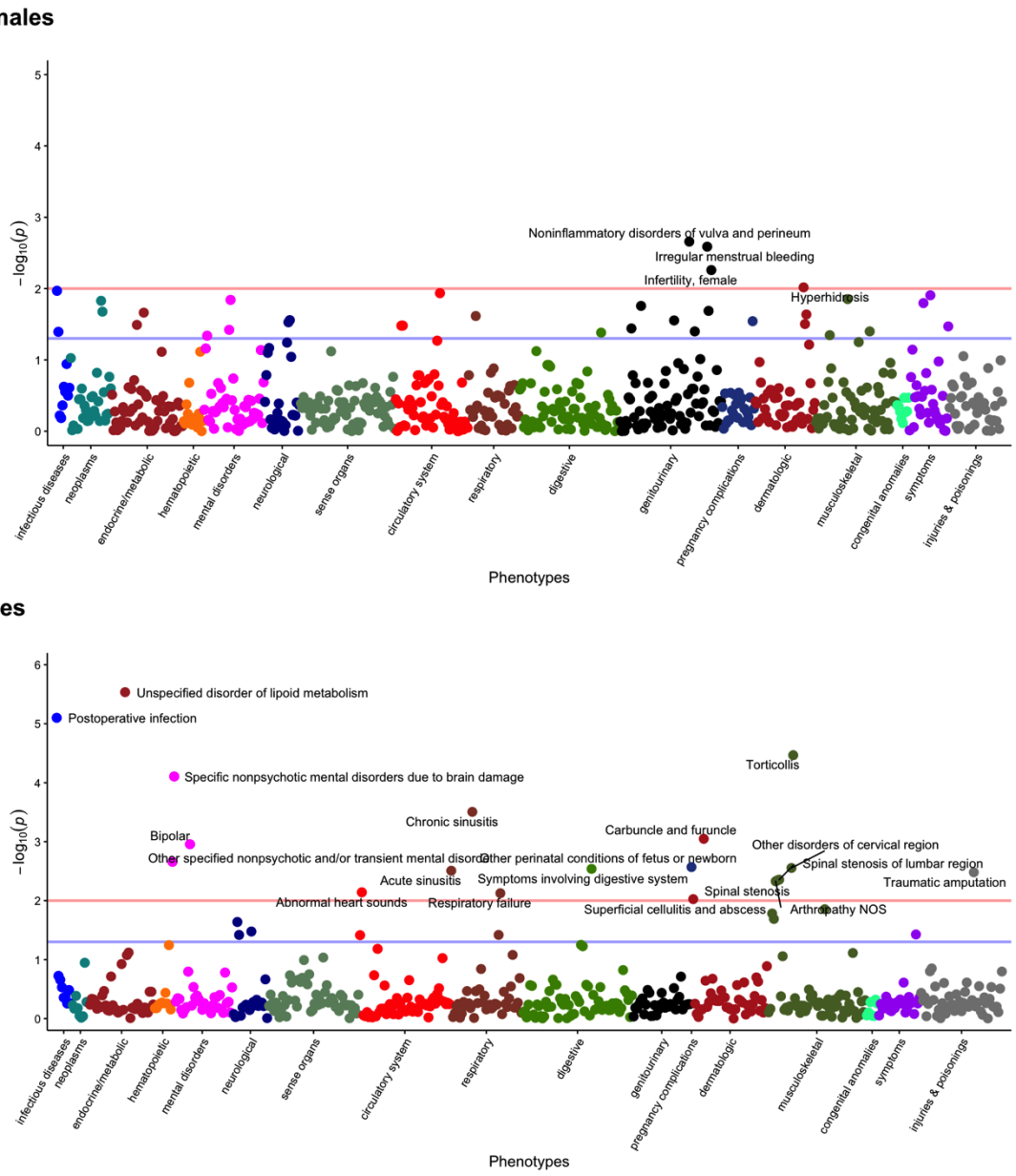
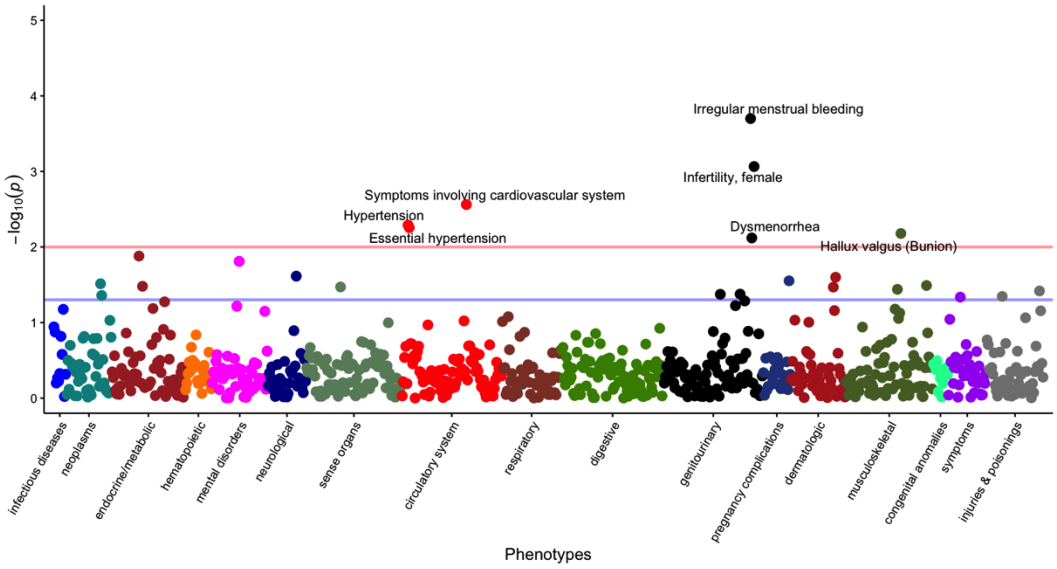


Fig. S4. Manhattan plots of unadjusted $-\log_{10}$ (*P* values) for phecodes observed before age 60. Each point shows one individual phecode. The horizontal red and blue lines show the significant thresholds for p-values less than 0.01 and 0.05. All association with p-value< 0.01 are annotated. For females, the conditions with p-values between 0.05-0.01 include bacterial infection, symptoms involving cardiovascular system, symptoms involving nervous and musculoskeletal systems, synovitis and tenosynovitis, obsessive-compulsive disorders, benign neoplasm of uterus, swelling of limb, other disorders of bladder, dysmenorrhea, uterine leiomyoma, protein-calorie malnutrition, acne, nasal polyps, abnormality of gait, inflammatory diseases of uterus [except cervix], complications of labor and delivery, abnormal movement, diseases of sebaceous glands, other abnormal glucose, essential hypertension, hypertension, malaise and fatigue, cystitis and urethritis, agorophobia, social phobia, and panic disorder, hallux valgus (bunion), hypertrophy of female genital organs, postoperative infection, nonspecific elevation of levels of transaminase or lactic acid dehydrogenase, displacement of intervertebral disc, and alteration of consciousness. For males, the conditions with p-values between 0.05-0.01 include other disorders of bone and cartilage, other arthropathies, unspecified monoarthritis, sleep apnea, migraine, other abnormal blood chemistry, respiratory failure, insufficiency, arrest, obstructive sleep apnea, and heart valve disorders.

Females



Males

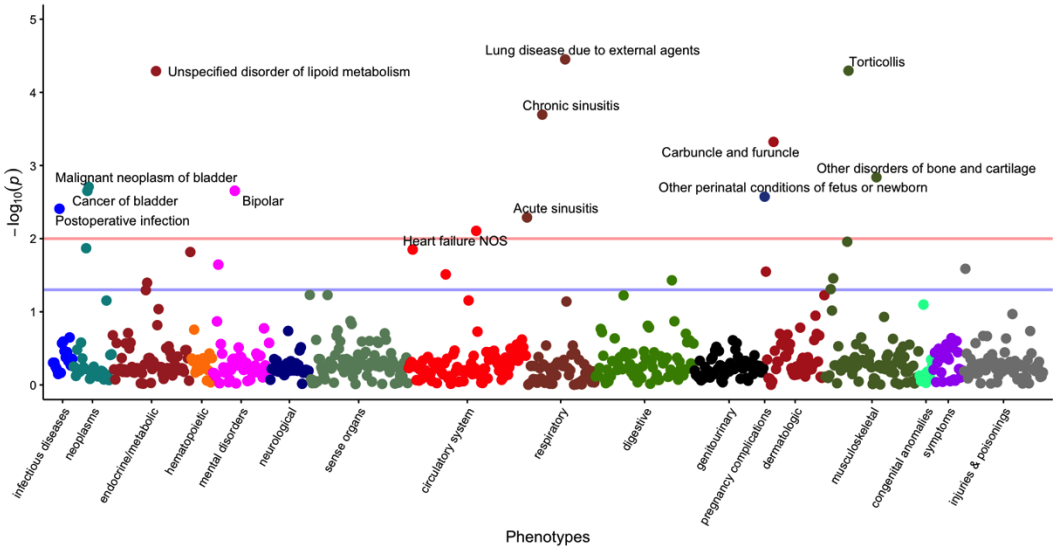


Fig. S5. Manhattan plots of unadjusted $-\log_{10}$ (*P* values) for phecodes observed before age 80. Each point shows one individual phecode. The horizontal red and blue lines show the significant thresholds for p-values less than 0.01 and 0.05. All association with p-value< 0.01 are annotated. For females, the conditions with p-values between 0.05-0.01 include disorders of the pituitary gland and its hypothalamic control, obsessive-compulsive disorders, lack of coordination, acne, complications of labor and delivery, benign neoplasm of uterus, other derangement of joint, protein-calorie malnutrition, anisometropia, hyperhidrosis, acquired toe deformities, poisoning by agents primarily affecting skin & mucous membrane, ophthalmological, otorhinolaryngological, & dental drugs, hypertrophy of female genital organs, inflammatory diseases of uterus [except cervix], uterine leiomyoma, fracture of unspecified bones, and swelling of limb. For males, the conditions with p-values between 0.05-0.01 include other disorders of cervical region, cancer of urinary organs (incl. kidney and bladder), abnormal heart sounds, other immunological findings, specific nonpsychotic mental disorders due to brain damage, fracture of neck of femur, superficial cellulitis and abscess, secondary/extrinsic cardiomyopathies, arthropathy, other disorders of liver, disorders of plasma protein metabolism, and other arthropathies.

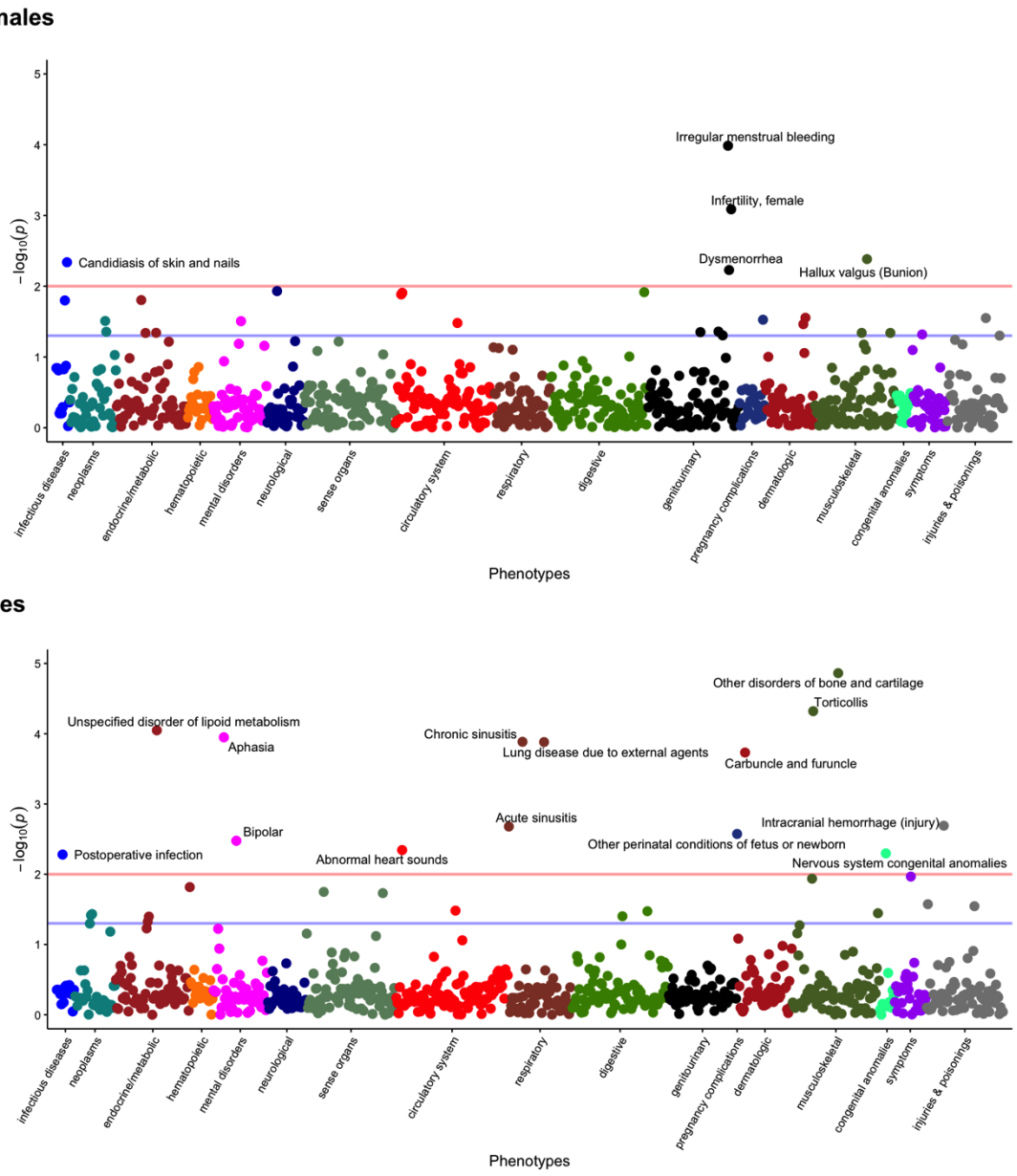


Fig. S6. Manhattan plots of unadjusted −log$_{10}$ (*P* values) for lifetime phecodes. Each point shows one individual phecode. The horizontal red and blue lines show the significant thresholds for p-values less than 0.01 and 0.05. All association with p-value< 0.01 are annotated. For females, the conditions with p-values between 0.05-0.01 include parkinson's disease, hemorrhage of rectum and anus, essential hypertension, hypertension, disorders of the pituitary gland and its hypothalamic control, dermatophytosis of the body, acne, complication of amputation stump, complications of labor and delivery, benign neoplasm of uterus, obsessive-compulsive disorders, symptoms involving cardiovascular system, hyperhidrosis, hypertrophy of female genital organs, uterine leiomyoma, inflammatory diseases of uterus [except cervix], hyperlipidemia, acquired foot deformities, other derangement of joint, protein-calorie malnutrition, swelling of limb, disorders of menstruation and other abnormal bleeding from female genital tract, and poisoning by agents primarily affecting skin & mucous membrane, ophthalmological, otorhinolaryngological, & dental drugs. For males, the conditions with p-values between 0.05-0.01 include hypothermia/chills, other disorders of cervical region, other immunological findings, corneal opacity and other disorders of cornea, otalgia, fracture of neck of femur, injuries to the nervous system, arrhythmia (cardiac), other disorders of liver, pathologic fracture of vertebrae, malignant neoplasm of bladder, cancer of bladder, inguinal hernia, paraproteinemia, and disorders of plasma protein metabolism.

Table S1. Phenotypic association with *FMR1* premutation reported in published literature, identified by KinderMiner. "Articles" column shows the total number of articles where each specific phecode and "*FMR1* premutation" co-occurred.

Phecode	Description	Group	Articles	P-value	Ratio
291	Other specified nonpsychotic and/or transient mental disorders	Mental disorders	39	2.11E-67	0.00425254
300.12	Agorophobia, social phobia, and panic disorder	Mental disorders	7	9.81E-10	0.00132051
626.8	Infertility, female	Genitourinary	5	7.48E-07	0.00105865
296	Mood disorders	Mental disorders	13	1.87E-13	0.00070217
271	Disorders of carbohydrate transport and metabolism	Endocrine/metabolic	3	5.23E-04	0.00066637
723.1	Torticollis	Musculoskeletal	1	6.13E-02	0.00053476
300.3	Obsessive-compulsive disorders	Mental disorders	2	1.28E-02	0.00039944
350	Abnormal movement	Neurological	2	1.42E-02	0.00037864
332	Parkinson's disease	Neurological	15	2.54E-09	0.00025998
296.22	Major depressive disorder	Mental disorders	5	5.72E-04	0.00025887
301	Personality disorders	Mental disorders	3	1.06E-02	0.00022708
798	Malaise and fatigue	Symptoms	1	1.43E-01	0.00021867
296.2	Depression	Mental disorders	39	1.52E-15	0.0001662
327.3	Sleep apnea	Neurological	3	3.72E-02	0.00013985
340	Migraine	Neurological	2	1.61E-01	0.00009466
338.2	Chronic pain	Neurological	5	9.02E-02	0.00007174
615	Endometriosis	Genitourinary	1	3.88E-01	0.00006879
401	Hypertension	Circulatory system	13	9.07E-02	0.00005163
599.4	Urinary incontinence	Genitourinary	1	4.94E-01	0.0000497