BRIEF COMMUNICATION OPEN

The Egyptian Collaborative Cardiac Genomics (ECCO-GEN) Project: defining a healthy volunteer cohort

Yasmine Aguib[1,2,12]✉, Mona Allouba[1,2,3,12], Alaa Afify[1,12], Sarah Halawa[1,4], Mohamed El-Khatib[1], Marina Sous[1], Aya Galal[1], Eslam Abdelrahman[1], Nairouz Shehata[1], Amr El Sawy[1], Mohamed Elmaghawry[1], Shehab Anwer[1,5], Omnia Kamel[1], Wesam El Mozy[1,6], Hadir Khedr[1], Ahmed Kharabish[1], Nagwa Thabet[1], Pantazis I. Theotokis[2,3], Rachel Buchan[2,3], Risha Govind[2,3], Nicola Whiffin[2,3,7], Roddy Walsh[2,3,8], Heba Aguib[1], Ahmed Elguindy[1,2], Declan P. O'Regan[7], Stuart A. Cook[2,3,9,10], Paul J. Barton[2,3], James S. Ware[2,3,7] and Magdi Yacoub[1,2,11]

The integration of comprehensive genomic and phenotypic data from diverse ethnic populations offers unprecedented opportunities toward advancements in precision medicine and novel diagnostic technologies. Current reference genomic databases are not representative of the global human population, making variant interpretation challenging, especially in underrepresented populations, such as the North African population. To address this, the Egyptian Collaborative Cardiac Genomics (ECCO-GEN) Project launched a study comprising 1000 individuals free of cardiovascular disease (CVD). Here, we present the first 391 Egyptian healthy volunteers recruited to establish a pilot phenotyped control cohort. All individuals underwent detailed clinical investigation, including cardiac magnetic resonance imaging (MRI), and were sequenced using a targeted panel of 174 genes with reported roles in inherited cardiac conditions. We identified 1262 variants in 27 cardiomyopathy genes of which 15.1% were not captured in current global and regional genetic reference databases (here: gnomAD and Great Middle Eastern Variome). The ECCO-GEN project aims at defining the genetic landscape of an understudied population and providing individual-level genetic and phenotypic data to support future studies in CVD and population genetics.

npj Genomic Medicine (2020)5:46 ; https://doi.org/10.1038/s41525-020-00153-w

INTRODUCTION

Cardiovascular disease (CVD) is a major cause of death and disability worldwide[1,2], and its prevalence continues to increase in low- and middle-income countries toward epidemic proportions[3,4]. Effective tailored strategies for the prevention and treatment of disease depend on thorough understanding of the molecular determinants and mechanisms involved, which may differ between specific populations. The rapid evolution of genomic and personalized precision medicine offers unprecedented opportunities in this regard[5,6]. These, however, are critically dependent on defining the genetic landscape of different populations, their individuals and the relation to their dynamic phenotype[7,8]. In-depth information is lacking in populations with a high disease burden, and yet they continue to be grossly understudied[9–13]. Notable initiatives and studies have recently begun in the Middle East and North Africa (MENA) region aiming at data collection and harmonization, paving the way for large-scale genomic studies[14,15]. To directly address these issues, the Egyptian Collaborative Cardiac Genomics (ECCO-GEN) Project is recruiting 1000 Egyptian healthy volunteers (EHVols) from the general population, who consent to be recalled to future research and are simultaneously establishing a regional biobank that hosts a broad range of biological samples for prospective studies. Participants are fully phenotyped with respect to cardiovascular health. The full dataset of 1000 EHVols will aid in distinguishing between incidental and medically actionable variants, and thus enhance diagnostic and treatment strategies. Here, we describe the protocol of the study and examine background genetic variation in genes previously shown to be involved in inherited cardiac conditions (ICCs), with a special focus on hypertrophic cardiomyopathy (HCM) and dilated cardiomyopathy (DCM).

RESULTS

Characteristics of the EHVol Study Population

We initially screened 440 self-reported healthy individuals of whom 11% were excluded for not meeting the first round of inclusion criteria (see "Methods", Supplementary Table 1 and Fig. 1). The remaining healthy individuals; the EHVols ($n = 391$), were genetically characterized using the ICC gene panel[16]. The general and cardiac characteristics of the study population are summarized in Table 1. The study population comprised 166 females (42.5%) and 225 males (57.5%). The mean age (years) was 33.2 (SD 9.5).

Representation of EHVol genetic variation in reference genetic datasets

We assessed the representation of the EHVol cohort in the Genome Aggregation Database (gnomAD) (https://gnomad.broadinstitute.org; version v2.1.1). GnomAD is the largest population reference dataset to date. It comprises genetic data from

[1]Aswan Heart Centre, Aswan, Egypt. [2]National Heart and Lung Institute, Imperial College London, London, UK. [3]Cardiovascular Research Centre, Royal Brompton and Harefield NHS Foundation Trust, London, UK. [4]Biotechnology Graduate Program, American University in Cairo, New Cairo, Egypt. [5]University Heart Center Zürich, Zürich, Switzerland. [6]Radiology department, Kasr Alainy School of Medicine, Cairo University, Cairo, Egypt. [7]MRC London Institute of Medical Sciences, Imperial College London, London, UK. [8]Department of Experimental Cardiology, Amsterdam Cardiovascular Sciences, Amsterdam UMC, Amsterdam, Netherlands. [9]National Heart Centre Singapore, Singapore, Singapore. [10]Duke-National University of Singapore, Singapore, Singapore. [11]Harefield Heart Science Centre, Harefield, Middlesex, UK. [12]These authors contributed equally: Yasmine Aguib, Mona Allouba, Alaa Afify. ✉email: yaguib@gmail.com

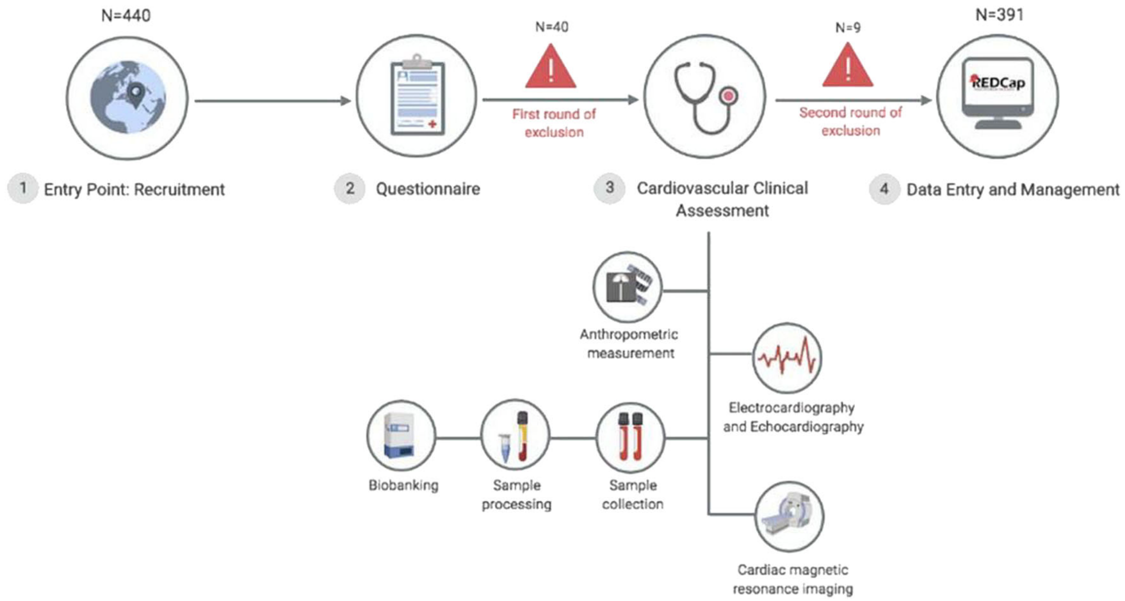
N=440
N=40
N=9
N=391
First round of exclusion
Second round of exclusion
REDCap
1 Entry Point: Recruitment
2 Questionnaire
3 Cardiovascular Clinical Assessment
4 Data Entry and Management
Anthropometric measurement
Electrocardiography and Echocardiography
Biobanking
Sample processing
Sample collection
Cardiac magnetic resonance imaging

Fig. 1 Workflow of the ECCO-GEN EHVol study. Study participants (1) were recruited from the general population via announcements (brochures, flyers, and public events); (2) completed a questionnaire providing demographic data, family, and clinical history; and (3) underwent detailed cardiovascular clinical assessment and blood sampling. (4) All data were recorded and managed in a local REDCap database. Rounds of exclusion: first round of exclusion is based on the basis of demographic and general health questionnaires as described under "study protocol and data collection". A second round of exclusion was based on detailed cardiovascular phenotyping as described under "cardiovascular phenotyping" in the "Methods" section. ECCO-GEN The Egyptian Collaborative Cardiac Genomics, EHVol Egyptian healthy volunteers. Created with BioRender.

unrelated individuals from five main populations: African/African American, Admixed American, East Asian, Non-Finnish European, and South Asian, while data from the MENA region was lacking in the dataset. Individuals were sequenced by whole-exome sequencing (WES; $n = 125{,}748$) and whole-genome sequencing (WGS; $n = 15{,}708$)[17] (see "Methods"). A total of 1262 variants were identified in the EHVol cohort in validated cardiomyopathy (CM) genes (see Supplementary Table 2 for transcript details), as well as genes implicated in syndromes that may present with isolated left ventricular hypertrophy (LVH)[18] (Supplementary Table 3; see "Methods"). Variants identified in the EHVol cohort were categorized into three classes: "loss-of-function" (LoF; predicted truncating variants, i.e., frameshift, splice acceptor, splice donor, and nonsense), "synonymous", and "other protein altering" (i.e., missense, inframe deletion, inframe insertion,…etc.). We calculated the proportion of EHVol variants present in gnomAD within each corresponding observed allele count (AC) bin. Of the 1262 EHVol CM variants, 1062 (84.2%) were captured in gnomAD, including all the variants that were common (>1%) in the EHVol cohort ($n = 283$; Fig. 2a and Table 2). The majority of non-gnomAD variants (EHVol only; $n = 200$) were singletons (AC bin 1) and were mostly classified as "other protein altering" (Fig. 2a, b). Among these singletons, LoF represented 2% of variants absent from gnomAD, as compared to those present (0.2%), indicating that underrepresented populations can potentially contribute to pathogenicity interpretation. Differences in sequencing coverage between WES data (gnomAD) and gene panel data (EHVol) are expected; thus, we assessed the coverage of the positions of non-gnomAD variants ($n = 200$) in gnomAD (Supplementary Fig. 2). We then compared the proportion of non-gnomAD variants in our cohort to the proportion of non-gnomAD variants in an ethnic control group that is generally represented in gnomAD, the Caucasian healthy volunteers (CHVols) cohort (see Methods). The CHVols comprised 1028 CHVols who were also sequenced using the Trusight Cardio Sequencing Kit for ICC[16,19]. As expected, the proportion of EHVol non-gnomAD variants was significantly higher (p value = 1.9e−06) than that of the CHVol cohort (15.9% and 10.3%, respectively; Table 2). The proportion of controls with LoF variants in CM genes was comparable between the EHVol (0.48%) and CHVol (0.3%) cohorts.

In addition, we assessed the representation of the EHVol variants in another reference dataset, the Great Middle Eastern Variome (GME; $n = 1111$), as it includes WES data from the ethnically matched North African subpopulations[14]. Only ten non-gnomAD EHVol variants (0.79%) were exclusively present in the GME dataset and 524 (41.5%) variants were absent from it (Table 2). EHVol variants that were only present in GME were rare (AC = 1 or 2) and were mostly "other protein-altering" variants (data not shown).

High frequency of rare variation in the EHVol cohort

We compared the frequency of rare protein-altering variation in the EHVol cohort with the CHVol cohort, in order to study the background genetic variation in CM-associated genes (see "Methods"). Since we had individual-level data for both control cohorts, we were able to report and compare the proportion of individuals with rare variants in DCM (gnomAD filtering allele frequency (FAF) popmax $\leq 8.4 \times 10^{-5}$ (ref. [20])) and HCM (gnomAD FAFpopmax $\leq 4.0 \times 10^{-5}$ (ref. [20])) genes (Fig. 3). Individuals were counted more than once if they carried more than one variant type in a gene. In both cohorts, *TTN*, *DSP*, *RBM20*, *MYH7*, and *SCN5A* accounted for the majority of rare variation in DCM genes (Fig. 3a). The predominant HCM genes *MYBPC3*, *MYH7*, and the syndromic LVH gene *CACNA1C* accounted for the majority of rare variation in both cohorts (Fig. 3b).

Comparison of the distribution of rare *TTN* and *MYH7* variants between the different control cohorts

We investigated the distribution of rare variants in *TTN* (ENST00000589042) and *MYH7* (ENST00000355349), as different positional distributions of rare variants between cases and

Table 1. Summary of general self-reported and cardiac characteristics of the EHVol ($n = 391$).

General characteristics	EHVols *n* (%)	Available data (*n*)
Age (years) Mean (SD)	33.2 (9.5)	391
Gender		391
Males	225 (57.5%)	
Females	166 (42.5%)	
Offspring of consanguineous marriage (self-reported)	64 (16.5%)	388
Smoking		390
Former smoker	11 (2.8%)	
Current smoker	83 (21.3%)	
Nonsmoker	296 (75.9%)	
Alcohol		389
1 glass/week	3 (0.8%)	
2–3 glasses/week	1 (0.3%)	
Unspecified	3 (0.8%)	
Family history of CVD	64 (16.4%)	391
BMI		352
<18.5	14 (4.0%)	
18.5–25	127 (36.1%)	
25–30	128 (36.4%)	
>30	83 (23.6%)	
Heart rate (beat/min) mean (SD)	75.7 (13.1%)	319

Cardiac characteristics	Mean (standard deviation)	Available data (*n*)
ECG variables		340
PR interval (ms)	147.3 (21.8)	
QRS duration (ms)	91.1 (47.8)	
QTc interval (ms)	401.8 (44.4)	
CMR variables		391
LVEF (%)	62.9 (5.4)	
LVMi (g/m^2)	46.3 (13.8)	
LVESVi (ml/m^2)	27.7 (6.9)	
LVEDVi (ml/m^2)	74.5 (11.8)	

EHVols Egyptian healthy volunteers.

controls have been previously reported[21,22]. The analysis was restricted to rare LoF variants in *TTN*, and both missense and LoF variants in *MYH7* (Fig. 4), as they may be putative pathogenic[23,24]. Across all non-diseased cohorts, the majority of *TTN* variants were located in exons that are not constitutively expressed in the myocardium (Fig. 4a). One LoF variant identified in an EHVol (0.26%; p.Ser34842ProfsTer9) and six LoF variants identified in CHVols (0.58%; p.Glu35478Ter, p.Gln16235Ter, p.Gln15575Ter, c.44015-1 G > T, p.Pro4353GlnfsTer14, and p.Gln3243Ter) resided in cardiac constitutive exons, which would have been interpreted as "likely pathogenic" (LP) were they identified in individuals presenting with DCM[25]. Such variants are of unknown penetrance and frequently observed in individuals with no disease[26,27]. The LoF variant p.Ser34842ProfsTer9 identified in the EHVol cohort, is absent from both gnomAD and CHVol controls. According to the guidelines by the American College of Medical Genetics and the American Association of Pathology (ACMG/AMP), the p.Ser34842ProfsTer9 should be classified as LP. This classification is based on the activation of two ACMG/AMP rules, which are: the strong PVS1 "absent in population databases" rule and moderate PM2 rule "predicted null variant in a gene where LOF is a known mechanism of disease"[28]. However, the individual harboring this variant, who was clinically assessed at the age of 33, did not show a clinical DCM phenotype. Regular clinical follow-up will be performed for this individual. Moreover, across all populations, the majority of rare variants in *MYH7* (popmax FAF $\leq 4.0 \times 10^{-5}$) were located outside of the functional domains that typically harbor HCM-associated variants (protein residues: 167–931; EHVol 4 (66.7%); CHVol 9 (75%); gnomAD 163 (71.5%); Fig. 4b). In addition, the *MYH7* LoF variant (p.Leu781ArgfsTer3) identified in one EHVol control did not fulfill the ACMG/AMP criteria for a "benign" or "pathogenic" classification, and thus its default ACMG/AMP classification was "variant of uncertain significance".

DISCUSSION

The availability of large-scale population genetic databases, such as gnomAD, has opened up new possibilities in identifying putative disease-causing variants. Current datasets, however, do not adequately represent genetic variation in the MENA region, which hinders accurate variant interpretation in the respective populations. Several recent studies address this shortcoming. Scott et al., for instance, studied rare genetic variation of the GME region, including North East- and North West Africa[14]. Individuals recruited to the aforementioned study were from the general population (self-reported healthy volunteers), and thus may not necessarily be free of disease. To our knowledge, our study represents the first in the region to include high-coverage sequencing data from a clinically phenotyped healthy cohort. All EHVol controls were consented for recall for future studies to allow us to assess the effect of particular genotypes of interest on the clinical phenotype.

EHVol variants that were represented in both reference datasets comprised 57.69% ($n = 728$) of the variants. Given that our population is represented in the GME dataset, we expected more variants to be captured in it. This highlights the underrepresentation of the Egyptian population in current global and regional reference genetic databases. Overall, the frequency of rare variants in CM-associated genes was higher among EHVols compared to CHVols. However, it is important to note here that "rare" was defined using the frequency reported in gnomAD. Hence, observing this excess of variation in a CM free cohort likely reports variants that are more common in the Egyptian population rather than a higher burden of disease-related rare variation. This is supported by the phenotyping data. Our findings emphasize the importance of examining the rate of background rare variation in CM genes in ethnically distinct cohorts to allow us to discriminate the subset of variants that may be associated with an increased risk of disease.

Although the cohort size reported is still relatively small, this study provides preliminary insights into genetic variation of the underrepresented multi-ethnic Egyptian population. Expanding the ECCO-GEN project and further integration of ancestry-specific genetic and phenotypic data will deepen our understanding of African diversity, support variant interpretation and variant discovery. This will provide important information on disease-related mechanisms and a vital tool for precision medicine in ethnically related populations and in Africa.

METHODS

Study protocol and data collection

This is a population-based study to recruit 1000 EHVols from across Egypt. All participants provided informed consent, and the study was reviewed and approved by the local research ethics committee (20151125MYFAHC_Hvol). Here, we report data from an initial cohort of 391 volunteers recruited from December 2015

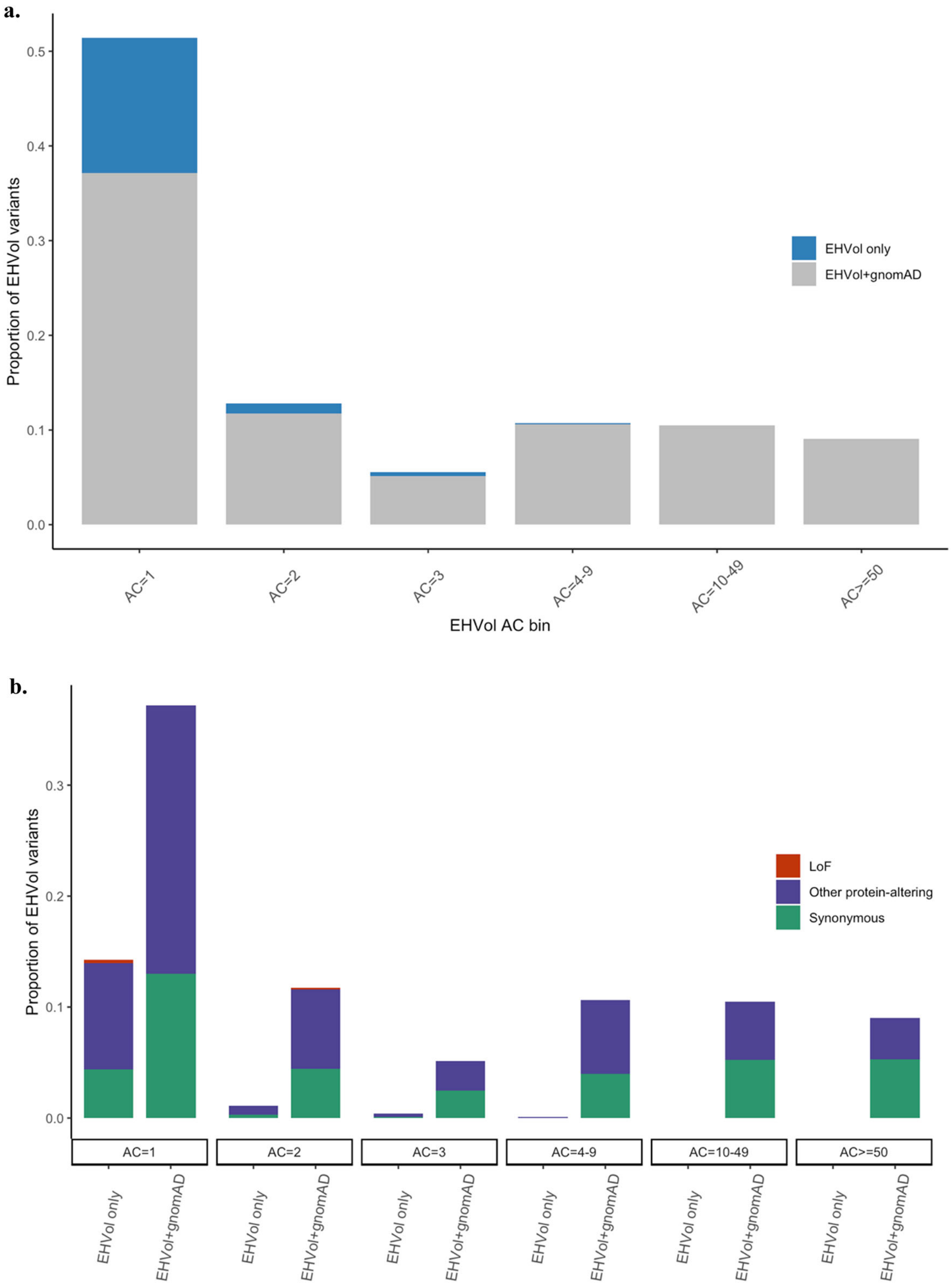

Fig. 2 All of the variants absent from gnomAD were rare in the EHVol cohort (seen in <1% of the cohort). a Bar graph showing the distribution of observed ACs (binned) of variants identified in the EHVol cohort. The proportion of variants that were present in gnomAD (all populations) is shown in gray and the proportion of variants that were only present in our cohort is represented in blue. **b** Bar graph showing the classification of the EHVol variants in each AC bin for variants that were captured in gnomAD (EHVol + gnomAD) and variants that were absent from gnomAD (EHVol only). Variants included in this analysis were collated by LoF, other protein-altering, and synonymous variants (color-coded in figure). EHVol Egyptian healthy volunteer, AC: allele count.

to June 2018 via national advertisement (brochures, flyers, and public events) and assessed at the Aswan Heart Centre (AHC). A paper-based questionnaire was conducted to gather information regarding demographics, health status, smoking and drinking habits, medical and surgical history, medication, consanguineous marriage (defined as self-reported first-cousin marriage) and family history (defined as having one or more first-degree relatives with self-reported heart disease). In the first round of exclusion,

Table 2. Proportion of non-gnomAD variants in the EHVol and CHVol cohorts.

Cohort	Cohort size	Total no. of variants in CM and syndromic genes	Total no. of non-gnomAD variants[a]	Prop. of non-gnomAD variants	Total no. of non-GME variants	Prop. of non-GME variants
EHVols	391	1262	200	15.9%*	524	41.5%
CHVols	1028	2344	241	10.3%*	2339	99.8%

EHVols Egyptian healthy volunteers, *CHVols* Caucasian healthy volunteers.
*Fisher's exact test (two-sided) p value = 1.943e−06.
[a]Variants found in gnomAD in low quality were included in the analysis.

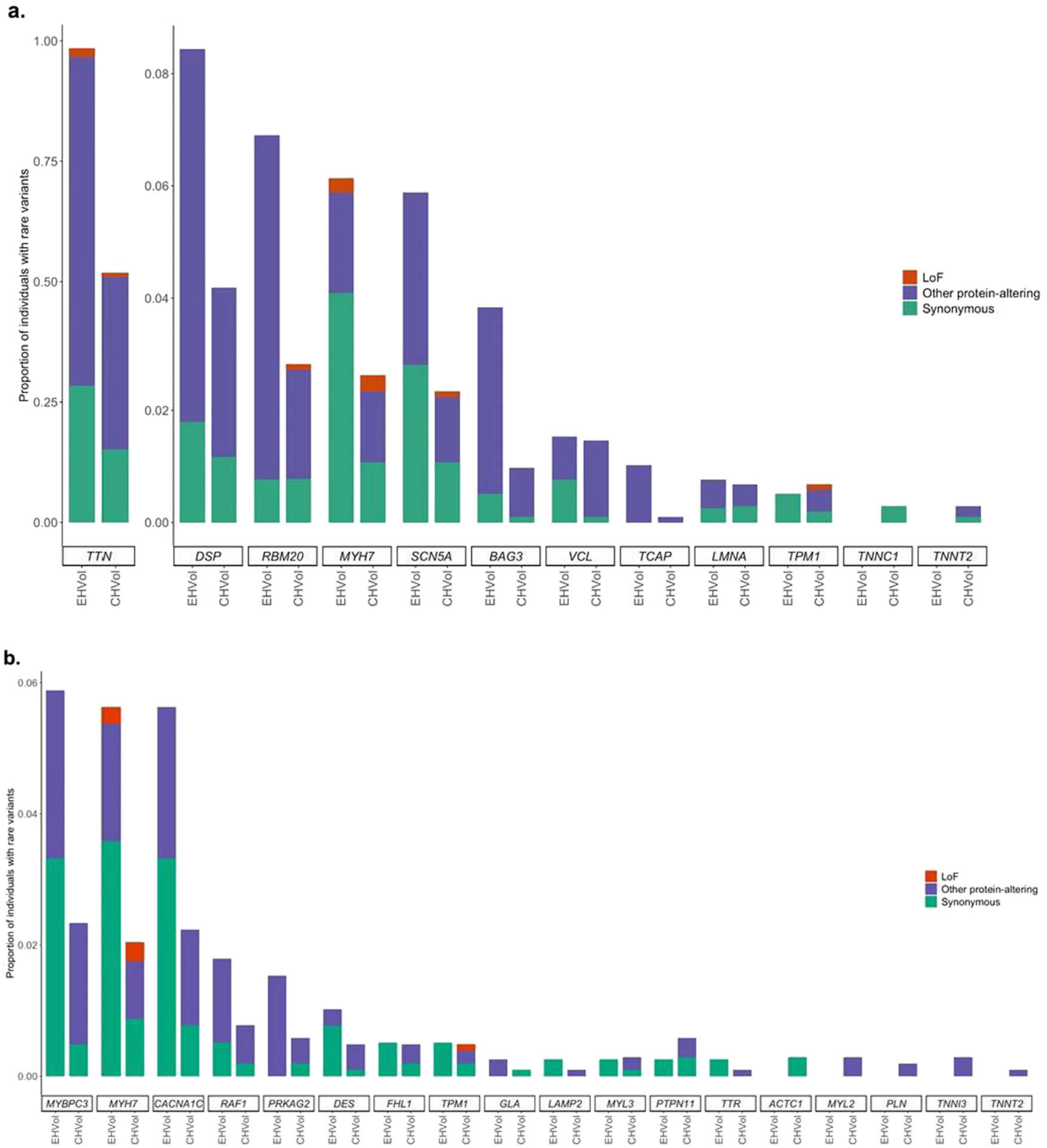

Fig. 3 Rare variation in CM genes in EHVol and CHVol controls. Bar charts represent the proportion of individuals with rare variants in **a** DCM and **b** HCM, and LVH syndromic genes. The first and second bars represent the CHVol and EHVol cohorts, respectively. Variants are collated by variant class (LoF, synonymous, and other protein altering). CM cardiomyopathy, EHVol Egyptian healthy volunteer, CHVol Caucasian healthy volunteers, DCM dilated cardiomyopathy, HCM hypertrophic cardiomyopathy, LVH left ventricular hypertrophy.

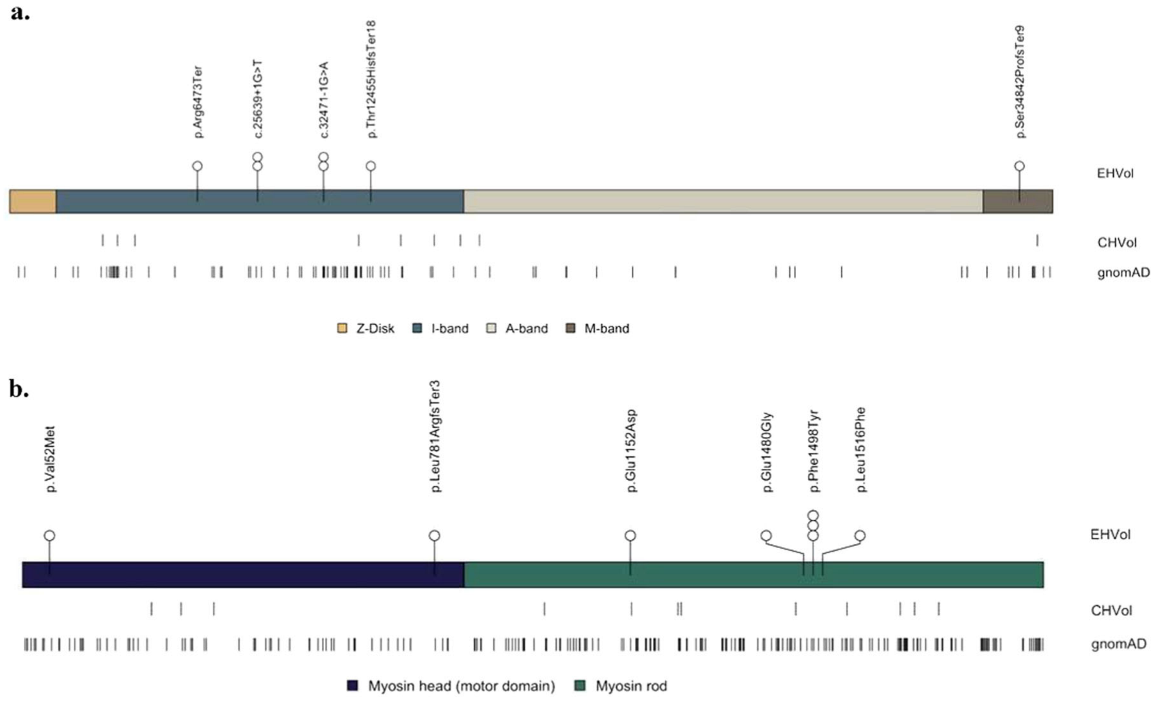

Fig. 4 Distribution of rare *TTN LoF* and *MYH7* missense and LoF variants in the EHVol, CHVol, and gnomAD cohorts. Variant distribution is shown relative to a schematic representation of the **a** *TTN* protein, with sarcomere regions delimited and the **b** *MYH7* protein with myosin domains delimited. The number of circles represent the number of individuals carrying the rare variant. EHVol Egyptian healthy volunteer, CHVol Caucasian healthy volunteers, *TTN* (ENST00000589042); *MYH7* (ENST00000355349).

individuals were excluded if they met any of the following criteria: <18 years of age, non-Egyptian nationality, pregnancy, presentation with known cardiovascular or collagen vascular disease, communication difficulties or contraindication to CMR. Data were entered into a Research Electronic Data Capture (REDCap) database and was stored, along with all appropriate documentation, on an access-controlled server[29].

Cardiovascular phenotyping

Eligible individuals underwent detailed cardiovascular phenotyping, including clinical examination, 12-lead electrocardiogram and CMR (Fig. 1). CMR was performed with a 1.5 T scanner (Siemens Magnetom Aera, Erlangen, Germany) using retrospective ECG triggering to capture the heart during the cardiac cycle. Steady-state free precession end-expiratory breath-hold cine images were acquired in the short axis orientation covering the whole heart. Standard parameters were repetition, time/echo time 3.6/1.8 ms; sense factor 2, flip angle, 60°; section thickness, 8 mm; no slice gap, matrix, 160 × 256; field of view, 300 mm; pixel size, 1.6 × 1.6 mm; number of phases 30 and phase percentage 67%. Phase-contrast images were acquired at different aortic levels for flow mapping. At the same acquisition levels, 4D flow was performed to assess flow patterns. T1 mapping was performed on base, mid and apical heart levels for fibrosis assessment. The 3D tagging acquisition was done at base, mid, and apical levels for strain assessment. Detailed structural and functional analysis on the CMR acquisitions was performed retrospectively using dedicated post processing and in-house software. Following phenotyping, a second round of exclusions, on the basis of specific cardiovascular diagnostic criteria, was applied (Table 3). Isolated apical non-compaction with normal ECG and normal CMR was not excluded.

Sample collection and biobanking

A total of 20 ml whole venous blood were withdrawn from each participant for biomarker measurement (hemoglobin A1c and troponin I), and biobanking of serum, plasma, and DNA. For DNA extraction, blood samples were transferred to K3EDTA tubes to avoid clotting. Blood samples were stored at 4 °C (max. 5 days) prior to DNA extraction. DNA was extracted using Wizard® Genomic DNA Purification Kit (Promega, Catalog No. A1620) according to manufacturer's instructions. Concentration of 1 µl DNA sample was determined using the NanoDrop 2000 (Thermo Scientific) spectrophotometer. A 260/280 and 260/230 nm ratios were used to assess the DNA quality. All samples were stored centrally in the AHC Biobank.

DNA sequencing

EHVols were sequenced with the Illumina Miseq and Nextseq platforms using the Trusight Cardio Sequencing Kit (Illumina, Catalog No. FC-141-1010 (MiSeq) and FC-141-1011 (NextSeq)) comprising 174 genes with reported roles in ICCs[16]. Sequencing was performed following the manufacturer's protocol. The concentration and quality of the DNA libraries were evaluated using Qubit (Invitrogen) and TapeStation 4200 (Agilent Technologies).

Bioinformatics pipeline

Raw data were subject to quality control using FastQC v0.10.1 (ref. [30]) and low-quality reads were trimmed via prinseq-lite v0.20.4 (ref. [31]). Trimmed reads were then mapped to hg19 (ref. [32]) using the Burrows–Wheeler Aligner v0.7.10-r789 (ref. [33]). After alignment, removal of duplicate reads was performed using picard v1.117 ("Picard Tools," n.d.). GATK v3.2-2-gec30cee[34] was then used for realignment of insertions or deletions (indels), as well as base quality score recalibration. Variant calling was performed using GATK's HaplotypeCaller. Joint genotyping was performed with GATK v4.0.8.1 and hard filters were applied based on GATK best practices workflow for germline short variant discovery[34]. Only variants marked as "PASS", with a quality of depth score ≥4 and an allelic balance ≥ 0.20 were retained. We attained high coverage of the target region (>99%) at ≥20× read depth

Table 3. Second round of exclusion.

Second round of exclusion criteria
Electrocardiogram (ECG) findings:
• Abnormal QT interval
• High-degree atrioventricular block
• Type I Brugada ECG
• Atrial fibrillation
• Non-sustained VT
Cardiac magnetic resonance (CMR) findings[a]:
• Left ventricular mass index (LVM_i) ≥ 99.5 g/m^2 in males and 85.5 g/m^2 in females
• Left ventricular end-systolic volume ($LVESV_i$) ≥ 44.5 ml/m^2 in males and 40.1 ml/m^2 in females
• Ejection fraction < 52.5% in males and 53.2% and in females
• LV wall thickness ≥ 15 mm (measured at any segment)
• Non-compacted-to-compacted myocardium thickness of ≥2.3:1 affecting mid and/or basal segment (i.e., not limited to the apex/apical segments)

Participants were excluded if any of these findings were identified.

[a] Cutoff thresholds were determined as three standard deviations from the mean to be more inclusive of extremes, especially in the absence of normal CMR reference values specific to the Egyptian population[37].

(Supplementary Fig. 1). The Ensembl Variant Effect Predictor[35] was used for variant annotation.

Data analysis

Analysis of background genetic variation in selected cardiomyopathy genes. To examine the background genetic variation among genes involved in inherited CMs; specifically, DCM and HCM, we selected genes that were recently validated (Supplementary Table 2). The following 12 DCM genes were analyzed[36]: *BAG3, DSP, LMNA, MYH7, RBM20, SCN5A, TCAP, TNNC1, TNNT2, TPM1, TTN*, and *VCL*. For HCM genes, the following eight genes were analyzed: *MYBPC3, MYH7, TNNT2, TNNI3, TPM1, ACTC1, MYL2, MYL3* as they have been reported to be definitively associated with the disease[18]. In addition, nine genes associated with syndromes that include cardiac hypertrophy and can present with isolated LVH (*CACNA1C, DES, FHL1, GLA, LAMP2, PRKAG2, PTPN11, RAF1*, and *TTR)* were analyzed, and also *PLN*, which is associated with a variable cardiomyopathic phenotype[18].

Comparison of genetic variation between EHVol and gnomAD controls. We downloaded WES data from the gnomAD database[17] (https://gnomad.broadinstitute.org; version v2.1.1) and used it to assess genetic variation in CM genes between the EHVol and gnomAD controls. We restricted our analysis to protein-coding regions only. Identified variants were categorized into three classes: LoF (predicted truncating variants, i.e., frameshift, splice acceptor, splice donor, and nonsense), synonymous and other protein-altering (missense, inframe deletion, inframe insertion,…etc.).

Comparison of rare variation between EHVol and CHVol controls. The CHVols were recruited prospectively via advertisement for the UK Digital Heart Project at Imperial College London. All volunteers underwent CMR to confirm the absence of cardiac disease. The frequency of rare variation in the selected CM genes was compared between the EHVol and CHVol cohorts. The threshold maximum credible population allele frequency (MAF) for putative disease-causing variants in DCM ($\leq 8.4 \times 10^{-5}$) and HCM ($\leq 4.0 \times 10^{-5}$) were defined using the statistical framework developed by Whiffin et al.[20] as denoted by the following equation[19]: MAF = (disease prevalence × maximum allelic contribution (i.e., genetic heterogeneity)/penetrance). The MAF was then compared with the FAF available in gnomAD, which is the precomputed AF threshold calculated for all variants in gnomAD, as described in ref. [20]. For an HCM/DCM variant to be considered potentially disease causing, its MAF should be lower than the popmax FAF. The "popmax FAF" is the highest calculated AF among the five main gnomAD subpopulations. The frequency of rare variants per gene in the EHVol/CHVol cohorts was calculated by counting the number of rare variants per gene and dividing this by the cohort size (EHVols: $n = 391$, CHVols: $n = 1028$).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

Sequencing data (ICC panel) from this study were submitted to the European Genome-phenome Archive (EGA), which is hosted by The European Bioinformatics Institute (EBI) and the Centre for Genomic Regulation (CRG) under the accession code: EGAD00001006160. All other data generated or analyzed during this study are included in this published article (and its Supplementary information files).

CODE AVAILABILITY

Processing of raw sequence files, variant calling, variant annotation, and all other analyses described in this study were performed using available and open source software, as described in the "Methods" section. No custom code was therefore used, apart from simple scripts integrating commands for such programs.

Received: 18 September 2019; Accepted: 31 August 2020;
Published online: 23 October 2020

REFERENCES

1. Murray, C. J. L. & Lopez, A. D. Measuring global health: motivation and evolution of the Global Burden of Disease Study. *Lancet* **390**, 1460–1464 (2017).
2. World Health Organization (WHO). *Cardiovascular Diseases (CVDs)*. https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (2017).
3. Roth, G. A. et al. Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *J. Am. Coll. Cardiol.* **70**, 1–25 (2017).
4. Yusuf, S. et al. Cardiovascular risk and events in 17 low-, middle-, and high-income countries. *N. Engl. J. Med.* **371**, 818–827 (2014).
5. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
6. O'Donnell, C. J. & Nabel, E. G. Genomics of cardiovascular disease. *N. Engl. J. Med.* **365**, 2098–2109 (2011).
7. Lau, E. & Wu, J. C. Omics, big data, and precision medicine in cardiovascular sciences. *Circ. Res.* **122**, 1165–1168 (2018).
8. Leopold, J. A. & Loscalzo, J. Emerging role of precision medicine in cardiovascular disease. *Circ. Res.* **122**, 1302–1315 (2018).
9. Need, A. C. & Goldstein, D. B. Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* **25**, 489–494 (2009).
10. Bustamante, C. D., Burchard, E. G. & De la Vega, F. M. Genomics for the world. *Nature* **475**, 163–165 (2011).
11. Owolabi, M. O. et al. Understanding the rise in cardiovascular diseases in Africa: harmonising H3Africa genomic epidemiological teams and tools: cardiovascular topic. *Cardiovasc. J. Afr.* **25**, 134–136 (2014).
12. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* **538**, 161–164 (2016).
13. Keates, A. K., Mocumbi, A. O., Ntsekhe, M., Sliwa, K. & Stewart, S. Cardiovascular disease in Africa: epidemiological profile and challenges. *Nat. Rev. Cardiol.* **14**, 273–293 (2017).
14. Scott, E. M. et al. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nat. Genet.* **48**, 1071–1076 (2016).
15. Owolabi, M. O. et al. Data resource profile: cardiovascular H3Africa innovation resource (CHAIR). *Int. J. Epidemiol.* **48**, 366–367 (2019).
16. Pua, C. J. et al. Development of a comprehensive sequencing assay for inherited cardiac condition genes. *J. Cardiovasc. Transl. Res.* **9**, 3–11 (2016).
17. Karczewski, K. J. et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. Preprint at https://doi.org/10.1101/531210 (2019).
18. Ingles, J. et al. Evaluating the clinical validity of hypertrophic cardiomyopathy genes. *Circ. Genomic Precis. Med.* **12**, e002460 (2019).

19. de Marvao, A. et al. Precursors of hypertensive heart phenotype develop in healthy adults: a high-resolution 3D MRI study. *Jacc. Cardiovasc. Imaging* **8**, 1260–1269 (2015).
20. Whiffin, N. et al. Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet. Med.* **19**, 1151–1158 (2017).
21. Pugh, T. J. et al. The landscape of genetic variation in dilated cardiomyopathy as surveyed by clinical DNA sequencing. *Genet. Med.* **16**, 601–608 (2014).
22. Homburger, J. R. et al. Multidimensional structure-function relationships in human β-cardiac myosin from population-scale genetic variation. *Proc. Natl Acad. Sci.* **113**, 6701–6706 (2016).
23. Mazzarotto, F. et al. The genetic architecture of left ventricular non-compaction reveals both substantial overlap with other cardiomyopathies and a distinct aetiology in a subset of cases. Preprint at https://doi.org/10.1101/2020.01.03.19015602 (2020).
24. Allouba, M. et al. Analysis of HCM in an understudied population reveals a new mechanism of pathogenicity. Preprint at https://doi.org/10.1101/2020.03.24.20037358 (2020).
25. Roberts, A. M. et al. Integrated allelic, transcriptional, and phenomic dissection of the cardiac effects of titin truncations in health and disease. *Sci. Transl. Med.* **7**, 270ra6–270ra6 (2015).
26. Schafer, S. et al. Titin-truncating variants affect heart function in disease cohorts and the general population. *Nat. Genet.* **49**, 46–53 (2017).
27. Pirruccello, J. P. et al. Titin truncating variants in adults without known congestive heart failure. *J. Am. Coll. Cardiol.* **75**, 1239–1241 (2020).
28. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–423 (2015).
29. Harris, P. A. et al. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inform.* **42**, 377–381 (2009).
30. Simon Andrews. *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
31. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
32. Kuhn, R. M. et al. The UCSC genome browser database: update 2009. *Nucleic Acids Res.* **37**, D755–D761 (2009).
33. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
34. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
35. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 1–14 (2016).
36. Walsh, R. et al. Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet. Med.* **19**, 192–203 (2017).
37. Kawel-Boehm, N. et al. Normal values for cardiovascular magnetic resonance in adults and children. *J. Cardiovasc. Magn. Reson.* **171**, 29 (2015).

ACKNOWLEDGEMENTS

We would like to thank the Egyptian volunteers for their generous contribution to the study and the Genome Aggregation Database; including all the contributing groups that provided exome variant data for comparison (full list can be found at: http://gnomad.broadinstitute.org/about). We also thank Dr. Mohamed Zakareya, Aswan Heart Centre Hospital director, for his support throughout the project, our cardiologists, Dr. Peter Philip and Dr. Kerolos Wagdy, our lab chemist Fiesal Mohamed, and our lead nurse, Mr. Mahmoud Abdel Hay, for their commitment, cooperation, and coordination. We would also like to acknowledge our research secretary, Ms. Omayma Alm El Din, for her dedication and diligence. In addition, we would like to express our gratitude to Eng. Yehia Raef for his support with the raw data deposition to the EGA repository. This study was supported by the Science and Technology Development Fund (STDF) government grant, the Wellcome Trust (107469/Z/15/Z), the Medical Research Council (UK), the NIHR Biomedical Research Unit in Cardiovascular Disease at Royal Brompton & Harefield NHS Foundation Trust, and Imperial College London, the NIHR Imperial College Biomedical Research Centre, the Fondation Leducq (11 CVD-01), and a Health Innovation Challenge Fund award from the Wellcome Trust and Department of Health, UK (HICF-R6–373). M.A. and S.H. are funded by Al Alfi Foundation to support their PhD. degrees at Imperial College London and American University in Cairo respectively. N.W. is supported by a Rosetrees and Stoneygate Imperial College Research Fellowship, and D.R. is funded by the British Heart Foundation (NH/17/1/32725 and RG/19/6/34387).

AUTHOR CONTRIBUTIONS

M.Y., Y.A., A.E.G. and H.A. conceived and designed the ECCO-GEN study. M.M., O.K., W.E.M., A.K., H.K. and N.T. performed clinical assessments, A.K. was responsible for coordinating the phenotyping stations. M.A., A.A., S.H., M.E.-K., M.S., A.G., E.A., N.S., A.E.S. performed experiments and contributed to data acquisition. P.I.T. contributed to the bioinformatics analysis of the raw sequencing data. Contributors to relevant set ups were: R.G. (bioinformatics pipeline), R.B. (NGS Lab) and S.A. (RedCap). Y.A., M.A. and A.A. contributed to data analysis and interpretation and drafted the manuscript. S.A.C. initiated and designed the CHVol study. M.Y., Y.A., P.J.B., J.S.W., R.W. AND N.W. contributed to critical revision of the manuscript for intellectual content.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41525-020-00153-w.

Correspondence and requests for materials should be addressed to Y.A.

Reprints and permission information is available at http://www.nature.com/reprints

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