Supplementary

Cohort profile update: Tehran Cardiometabolic Genetic Study

Table of contents

Study design and sampling

Tehran lipid and glucose study (TLGS)

Tehran lipid and glucose study (TLGS) is a large prospective population cohort study in Tehran, Iran. A detailed description of the survey can be found elsewhere[1,2]. The cohort included 19118 participants (mean age 46.2 and female 51%) who were initiated by 1999. Researchers did the baseline assessments (demographics, risk factors, anthropometry, and biospecimens) between 1999 and 2001, and they continued the follow-up for six phases with periodic intervals of 3 years (1999-until now). The major cardio-metabolic-related health events, such as myocardial infarction, stroke, diabetes mellitus, hypertension, obesity, hyperlipidemia, and familial hypercholesterolemia, were followed up over the last 22 years. All participants provided written informed consent at the start of the study and follow-up visits, and the Medical Ethics Committee approved the study.

Tehran Obesity Treatment Study (TOTS)

Tehran Obesity Treatment Study (TOTS) commenced in March 2013 to undertake a prospective study evaluating and comparing several surgical bariatric procedures in an Iranian population of morbidly obese patients presenting to a specialized bariatric center. This study is a longitudinal prospective cohort in 5500 consecutive patients undergoing bariatric surgery (mean age xxx and female 79.7%). The TOTS study, with 457 genotyped participants (women%) and at least nine years follow–up, was included in TCGS to examine the genomic study. DNA was obtained from

consenting participants. Additional details on the study protocol have previously been described [3].

Thyroid cancer Project (TCP)

In 2006, a genetic study was started on patients with medullary thyroid cancer (MTC). First, according to American Thyroid Association (ATA) guidelines, the study aimed to assess and compare mutations of two exons 10 and 11 of the RET proto-oncogene in patients of the Iranian population with reports from other populations [4]. With the discovery and identification of 5 specific Mutations in the Iranian population in the NCBI, the decision was made to conduct a comprehensive genetic study on these patients. In the next step, exons 10, 11, 15, 16, 17, 18, and 19 were added to this study. From this clinical registry, we included 416 participants for genotyping.

Clinical Follow-up Project (CFP)

Kawsar Human Genetics Research Center (KHGRC) focuses on the genetic diagnosis of more than 130 genetic diseases in Iran. KHGRC center has a family-based DNA biobank with more than 10,000 participants[5]. From this center, 320 genotyped participants were selected based on the birthplace of the last three generations and included in TCGS.

What has been measured?

Physical Examination (P/E)

Anthropometric Measurements

Anthropometric measurements are taken with shoes removed and the participants wearing light clothing. Weight and height are measured according to the standard protocol. Body mass index (BMI) is calculated by dividing the weight in kilograms by the square of height in meters. Waist circumference is measured at the level of the umbilicus, and hip circumference is measured over light clothing at the widest girth of the hip.

Ankle-arm Blood Pressure

After 15 minutes of participant arrival, a qualified physician measures blood pressure two times using a standard mercury sphygmomanometer calibrated by the Iranian Institute of Standards and Industrial Researches. The participants remained seated for 15 minutes. A pediatric, regular adult, or large cuff is used based on the participant's arm's circumference. The cuff is placed on the right arm at the heart level and inflated as high as possible increments until the cuff pressure is 30 mmHg above the level at which the radial pulse disappeared. There is at least a 30-second interval between those two separate measurements, and the mean of the two measurements is recorded as the participant's blood pressure. Systolic Blood Pressure (SBP) is the first sound's appearance. Diastolic Blood Pressure (DBP) is defined as the disappearance of the sound while deflating the cuff at a 2–3 mm per second decrement rate.

Laboratory Blood Test

After 12–14 hours of overnight fasting, a venous blood sample was drawn and centrifuged within 30–45 minutes of collection. All blood sampling was done between 7.00 and 9.00 A.M, and all measurements were completed on the sampling day. A Selectra 2 autoanalyzer (Vital Scientific, Spankeren, Netherlands) was used in the TLGS research laboratory on the day of blood collection to analyze samples.

Basic Metabolic Panel

Fasting and 2h blood sugar (FBS and 2h BS) were measured by the enzymatic colorimetric glucose oxidase method (Pars Azmoon Inc., Tehran, Iran); inter-and intra-assay coefficients of variation (CV) at baseline and follow-up phases were both less than 2.2%.

Serum creatinine levels were measured by the kinetic colorimetric Jaffe method. Both inter and intra-assay CVs were <5%.

Lipid Panel

The enzymatic colorimetric test was used with cholesterol esterase and cholesterol oxidase to evaluate total cholesterol (TC). Triglyceride (TG) was assayed using an enzymatic colorimetric method with glycerol phosphate oxidase. High-density lipoprotein cholesterol (HDL-C) was measured after the precipitation of the lipoprotein-B-containing lipoproteins with phosphotungstic acid. These analyses were performed using commercial kits (Pars Azmoon Inc., Tehran, Iran) and a Selectra 2 auto-analyzer (Vital Scientific, Spankeren, The Netherlands). Low-density lipoprotein cholesterol (LDL-C) was calculated from serum TC, TG, and HDL-C

concentration according to the modified Friedewald formula. Non-HDL-C was calculated by subtracting HDL-C from TC; TC/HDL-C and TG/HDL-C were calculated by dividing TC and TG by HDL-C, respectively. In all baseline and follow-up assays of lipid profile, the inter and intra-assay coefficients of variation were less than 1.9, 2.1, and 3% for TC, TG, and HDL-C.

Advanced Lipid Panel

Apolipoprotein A1 (Apo A1) and Apo B were measured by immunoturbidimetry methods (Pars Azmoun Co, Tehran, Iran), and Apo A4 and Apo C3 were measured by the enzyme-linked immunosorbent assay (ELISA) method.

Inflammatory Panel

Interleukin-6 (IL-6) and IL-10 levels were measured on the frozen samples using the sensitive ELISA method (Diaclone Co. French). Adiponectin level was also measured using sensitive ELISA (Adipogen Co. Korea). Enzyme conversion method, sensitive ELISA, and Axis-Shield kit were used to measure homocysteine (Hcy) level (Diazyme co. USA). Plasma C-reactive protein (hs-CRP) was measured using highly sensitive immunoenzymometric assays (IEMA) (Diagnostic Biochem, Co. Canada). The samples were analyzed by Sunrise ELISA reader (TECAN Co, Salzburg, Austria). The measurements' intra- and inter-assay coefficients of variations (CVs) were <5%.

Thyroid Function Test

Thyroid-stimulating hormone (TSH) and free thyroxine (fT4) were measured on serum samples by the electrochemiluminescence immunoassay (ECLIA) method, using Roche Diagnostics kits and Roche/Hitachi Cobas e- 411 analyzers (GmbH, Mannheim, Germany). Thyroid peroxidase antibody (TPO-Ab) was measured using IEMA by the related kit (Monobind, Costa Mesa, CA, USA) and the Sunrise ELISA reader (Tecan Co., Salzburg, Austria). Lyophilized quality control material (Lyphochek Immunoassay plus Control, Bio-Rad Laboratories) was used to monitor the accuracy of assays. Intra- and inter-assay CVs were 1.3% and 3.7% for fT4 and 1.5% and 4.5% for TSH determinations, respectively. Intra- and inter-assay CVs for TPO-Ab were 3.9% and 4.7%, respectively(1). Thyroid hormone resistance was calculated by the Thyroid Feedback Quantile-based Index (TFQI) and Iranian-referenced Parametric TFQI (PTFQI) and compared with two other indices: Thyrotroph T4 Resistance Index (TT4RI) and TSH Index.

Liver Function Test

Liver function tests (LFTs), including aspartate aminotransferase (AST), alanine aminotransferase (ALT), gamma-glutamyltransferase (GGT), ALT/AST ratio, alkaline phosphatase (ALP), and lactate dehydrogenase (LDH) were measured using enzymatic colorimetric methods. AST and ALT were analyzed by enzymatic photometry and GGT by enzymatic colorimetric methods. Their intra- and inter-assay CVs were 2.8 and 3.8% for AST, 2.2 and 3.8% for ALT, and 2.9 and 3.0% for GGT, respectively. Both inter-and intra-assay CVs were <5%. These analyses were performed using

commercial kits (Pars Azmon Inc., Tehran, Iran) and a Selectra 2 auto-analyzer (Vital Scientific, Spankeren, The Netherlands) at the research laboratory of the TLGS.

Hormones

Aliquots of serum samples were stored at -70 °C and transferred to the Research Institute for Endocrine Sciences (RIES) hormone laboratory for the following assays. Leptin level was determined by an enzyme immunoassay (EIA) method. A sandwich ELISA kit evaluated cortisol levels via two antigen-specific antibodies, a purified immunoglobulin bound to a solid phase, and enzyme-linked detection antibodies. Then cortisol concentration was measured spectrophotometrically at 450 nm in a microplate reader. Anti-Mullerian hormone (AMH) level was measured by the two-site EIA method using the Gen II kit (Beckman Coulter, Inc., CA, USA) and the Sunrise ELISA reader (Tecan Co., Salzburg, Austria) at the time of recruitment. The intra- and inter-assay CVs were 1.9% and 2.0%, respectively. Fasting insulin level was measured through the electrochemiluminescence immunoassay method using Roche Diagnostics kits and Roche/Hitachi Cobas e-411 analyzer (GmbH, Mannheim, Germany). Lyophilized quality control material (Lyphochek Immunoassay Plus Control; Bio-Rad Laboratories, Irvine, CA, USA) was used to monitor the precision of assays. Intra- and inter-assay CVs were 1.2 and 3.5%, respectively.

Others

The serum level of 25-OH-D was assessed using the EIA method (25-OH-D EIA kit, DRG, Marburg, Germany); the expected range of the kit was 10–50 ng/ml. The assay sensitivity was 5.6

nmol/L, and intra-inter assay CVs were 2.7 and 3.9%, respectively. Serum zinc concentration was determined by flame atomic absorption spectrometry (CTA-2000, Chem-Tech, Analytical Co., Kempston, UK). Serum total nitrite and nitrate (NOx) levels were measured by a rapid and straightforward spectrophotometric method by the Griess reaction(2), which was validated in the RIES laboratory(3,4). Absorbance was read at 540 nm using the ELISA reader (Sunrise, Tecan, Austria). The concentration of NOx in serum samples was determined from the standard linear curve established by 0–100 μM sodium nitrate. The measurements' inter-and intra-assay CVs were 5.2% and 4.4%, respectively. Oxidative stress was assessed using serum levels of Malondialdehyde (MDA) and total antioxidant capacity (TAC). MDA concentration and lipid peroxidation assay kit (Abcam, Cambridge, CA, USA) were used. The total antioxidant capacity of plasma was measured using a TAC assay kit (Abcam, Cambridge, CA, USA) according to manufacturing protocol. All samples were analyzed under the internal quality control supervision of the acceptable criteria.

Traits

The TCGS cohort gathered traits and clinical information from 4 different well-designed projects. Each project is designed to catch the information through paper-based or web-based questionnaires, Patients' self-reports after being seen by a doctor, and hospitalization information.

Questionnaires

The TLGS committee gathered information through different questionnaires.

Demographic information

Includes age, sex, marital status, familial relationship, level of education, employment status, history of living in Tehran, etc.

Ethnicity

This information was gathered based on self-reported ethnicity, parents, and grandparents' place of birth.

Past Medical History

Consist of the history of any risk factors, such as cardiovascular disease, thyroid disease, diabetes, any major diseases, hospitalization, medication usage, family history of cardiovascular disease, etc.

Smoking and Physical activity

These questionnaires have separate sections for adults and adolescents. A smoking questionnaire is about ever using a cigarette, pipe, or hookah, duration of use, quitting period and amount of smoking, the pattern of secondhand smoke exposure, etc. A physical activity (PA) questionnaire, including the Lipid Research Clinic (LRC) questionnaire in phase 1 and the Modifiable Activity Questionnaire (MAQ) in phase 2, determines the amount of PA during work and outside the workplace and sports activities.

Two methods were used for PA data collection: LRC, used in the initial phase of TLGS (phase I), and a simple measurement containing four questions. The data obtained via LRC were subjective and not valid for the Iranian culture, so, then, the PA questionnaire was changed to a Persian-modified MAQ questionnaire(5), which measures leisure time, occupational, and household

activities and calculates the metabolic equivalent (MET) based on min/wk. High reliability (98%) and moderate validity (47%) were reported for the Persian version of MAQ(6,7). MAQ evaluated the PA pattern of adolescents aged 12-18 years. Individuals reported their PA during the past 12 months by identifying the frequency and duration of each activity. The MET was obtained by each activity weighted by its relative intensity. One MET is set at 3.5 ml of oxygen consumed per kg of body weight per minute and represents the resting metabolic rate. Then, the obtained MET was multiplied by the time spent at each level for all activity levels. MET-time calculated the average daily PA level from each level added to the total 24-hour MET-time.

Information about PA for adults was also collected using a reliable and validated Iranian version of MAQ, and the measurement details were described above. Total numbers of minutes/year for all leisure time PA were summed up and then divided by 52 to estimate total leisure-time PA in min/week. MET of total leisure-time PA for each person was then calculated by multiplying the number of min/week of each leisure time activity by its MET. Employed persons were asked to indicate how many hours a week they usually worked. According to the questionnaire, individuals had to identify the months and hours they participated in PA at work (standing, housework, work activities) over the past year. The assessment of occupational activity was based on summing up the number of hours per week of light, moderate and vigorous-intensity activities and then multiplying by 60 to express minutes per week of occupational activity over the past year. Finally, occupational (MET-min/week) activity was calculated by multiplying the number of minutes per

week of each of the three categories of occupational activity by MET values. The total PA was reached by adding leisure time PA to occupational activity.

Obstetric and Gynecology

Comprehensive questionnaires on reproductive lifespan, including menarche, menopause, menstrual regularity, parity, abortion, type and duration of contraception usage, infertility, and lactation, were collected through face-to-face interviews with trained staff.

Dietary assessment

Usual dietary intake was assessed using a valid and reliable 168-item semi-quantitative food frequency questionnaire (FFQ). FFQ comprised a list of foods with a standard serving size commonly consumed by Iranians. Trained dietitians collected the dietary data during private face-to-face interviews; they asked subjects to report the frequency of a given serving of each food item intake during the previous year on a daily, weekly, or monthly basis. The described frequency of each food item was converted to a daily intake. Portion sizes of consumed foods were converted to grams using measuring cups and spoons.

The Iranian food composition table (FCT) is incomplete and only provides data on a few nutrients, so we used the United States Department of Agriculture (USDA) FCT to analyze food and beverages for energy and nutrients. Some Iranian food items (e.g., dairy products such as kashk) were not included in the USDA FCT, so we used the Iranian FCT. For mixed items (e.g., pizza), nutrients were estimated based on described basic ingredients and usual restaurant recipes [67].

The health-related quality of life (HRQoL)

The health-related quality of life (HRQoL) questionnaires assess individuals' perceptions regarding different aspects of their health status. For adults, HRQoL information was collected using a reliable and validated Iranian version of the short-form 12-item health survey version 2 (SF-12v2)[8]. The SF-12v2 consists of 12 items and eight subscales, including physical functioning (2 items), role physical (2 items), bodily pain (1 item), general health (1 item), vitality (1 item), social functioning (1 item), role emotional (2 items), and mental health (2 items). The subscale scores range from 0 to 100; a higher score indicates a better health condition [9].

The Iranian version of the short-form of Keyes social well-being questionnaire was used to assess social well-being in adults [10]. This questionnaire has been developed by Keyes and encompasses 15 items and five dimensions, including social integration (3 items), social acceptance (3 items), social coherence (3 items), social contribution (3 items), and social actualization (3 items). The answering choices ranged from 1=Agree strongly to 7= Disagree strongly for each question. The subscale scores range from 0 to 15, and a higher score shows better social well-being [11].

The Spiritual health inventory in Muslim adults (SHIMA-48) was used to assess spiritual health in adult participants. It contains 48 items and two components, entitled cognitive-emotional and behavioral. The items were rated on a 5-point Likert scale anchored at 1 to 5 ("1= Strongly disagree to 5= Strongly agree for insight and tendency" subscales, and "1= Always to 5= Never"

for behavioral subscale). The scores ranged from 0 to 100, with higher scores indicating better spiritual health[12].

Medical complication

Self-reports

Trained general physicians visited participants in person and asked them three main questions during the verbal interview in their native language (Persian): 1-concerning their current chief complaint, 2. about any hospitalizations in their lifetimes, and 3. about any hospitalizations in the last three months. In response, participants explained their past medical, surgical, drug, habitual, and family history. All information based on the patient's statements in Persian is gathered in a text file.

Hospitalization

If any hospitalization occurred for participants, a follow-up questionnaire was filled out to record any symptoms during hospitalization. Every participant was followed up for any medical event during the prior year by telephone. They were asked for any medical conditions from a trained nurse, and if a related event had occurred, a trained physician collected complementary data during a home visit. If necessary, the physician visited the respected hospital and collected medical files. In the case of mortality, a local physician collected data from hospitals or death certificates. An outcome committee of related specialists labeled the disease for the participants who experienced hospitalization while participating in their parents' study. This committee

gathered hospitalization documents and completed a follow-up questionary (Supplementary excel sheet) to evaluate the final diagnosis. Also, all final diagnoses were gathered in a text file.

ICD 11 coding

We opted to use the International Statistical Classification of Diseases and Related Health Problems (ICD) as a guide to organize all complications and translate them to a worldwide standard language which includes 55000 unique codes for the injuries, diseases, and causes of death in 25 main chapters [13]

The created text files were comprised of different keywords and sentences, in English or Persian, with some dictation errors, different accents, and some slangy expressions or poor literature. At first, a bioinformatics group breaks down all the sentences in each text file into keywords. Then a group of physicians reviewed the keywords, filtered the medical terms, and translated them into English to create a native local dictionary. Each Persian keyword or phrase with a diagnostic value in the created dictionary is labeled with an ICD11 code. Bioinformaticists assigned an ICD-11 code to each complication with in-house python programming. The frequency of each condition was assessed.

Biobanking

The concept of designing a genomic bank from TLGS samples was first presented to the endocrine research center (ERC) and the Iranian molecular medicine network. It was funded by FA and MSD [grant number 147] (2004) and [grant number 265] (2008). Funding for the main study began in June 2012 with an agreement between RIES and DeCode genetic company

(Reikjavik, Iceland) with FA and MSD as main investigators. In 2008, a project for drawing pedigrees according to their genetic relationship [grant number 321] to MSD and AAM as principal investigators was funded by ERC. The final protocol for the genetic study was written by (FA, MSD, MSF, and DK) and submitted to the Ministry of Health and Medical Education (MOHME) in August 2012, and was approved by the National Committee for Ethics in biomedical research in December 2012. Among TLGS, 17495 participants (888 independent persons and the rest family members (3100 family min=3; max=56)) were included in TCGS.

Genotyping

DNA extraction

The blood collected in EDTA containing a test tube was immediately sent to the genomic laboratory. The blood sample was drawn based on international protocols (9). DNA was extracted by the Proteinase K, salting out the standard method [14]. The quality and quantity of extracted DNA were evaluated. Thermo Scientific NanoDrop™ 1000 Spectrophotometer qualified samples were aliquot in 1.5 ml tubes and stored in -80°C ultra-freezers for future studies.

A total of 16226 TCGS participants have been genotyped using Illumina Human OmniExpress-24-v1-0 bead chip containing 652,919 single nucleotide polymorphisms (SNPs) loci at the deCODE genetics/Amgen company (Iceland) according to the manufacturer's specifications (Illumina Inc., San Diego, CA, USA)2. 578 SNPs out of 652,919 SNPs in the entire dataset are monomorphic and fixed at a single allele. Approximately 96 % of 652,341 SNPs had a missing call rate of less than 0.02 (Table 2 supplementary). More than 99.8% of TCGS participants had less than 20% missing

genotype SNPs, (Table 3 supplementary). In TCGS genotyped data, Over 90% of SNPs have MAFs greater than 0.05, and the minimum minor allele frequency (MAF) is 3.3710e-05. The distribution of MAF of genetic variants in the TCGS population is depicted in (Table 4 supplementary). Figure 1 supplementary also illustrates the distribution of MAF and its relation to MAC. According to this boxplot, the population minor allele count (MAC) is appropriate, considering MAF less than 0.001. Out of 652,919 SNPs, 20,108 SNPs have P-value$_{hwe}$ <10E-7. The distribution of -log10 (p-value) was shown in Figure 2 supplementary. The right histogram is for all SNPs, and the left belongs to SNPs with p-value$_{hwe}$ <10E-7. The distribution of missing call rates for variants with MAF<0.05 and MAF>0.05 is depicted in Figure 3 supplementary. This graph demonstrates no relationship between the missingness pattern throughout the genome and the MAF in TCGS data.

The Impact of Variants: We grouped genetic variants into four categories: (i) LOW, including synonymous variants 3' and 5' UTR variants (N=9,118); (ii) MODERATE, including missense variants and splice region variants (N=9,370), (iii) LOF, loss of function including start/stop gained, frameshift, donor/acceptor splice, and initiator codon variants (N=318), and (iv) OTHER, including intronic and intergenic variants (N=633,512). The sequence variants were annotated using the last version of Variant Effect Predictor (VEP, ver 105)8 (Table 2 supplementary). The distribution of MAF by the impact is shown in Figure 4 supplementary.

Familial relationship

A trained nurse interviewed participants to ask for their genealogical information in all included projects, including their kinship, marital status, and consanguineous of their marriage. In each

follow-up, the project manager let the son-in-law, daughter-in-law, and new child (age over three years) enter the study and updated family information. A code was dedicated to each family group as a cluster. This information was gathered in a file and sent to the TCGS team to draw biological ties with standard guidelines for human pedigree nomenclature (8). Each individual was coded with a unique identification number (ID), cluster code, gender, and paternal and maternal ID. Some families with consanguineous marriages were assigned a new family identification number (FID) rather than a cluster code (Figure 5). The statisticians checked the ties using Statistical Analysis for Genetic Epidemiology (SAGE) software (9). Next, individual identity-by-descent (IBD) was calculated pairwise using PLINK software and was used to double-check the family relationships (10). Kinship relations and accuracy of drawing pedigrees were also controlled by the Family-Based Association Tests (11)-Toolkit V 1.7.3.

The genetic data management system Progeny Clinical Version 7 saved and processed family data, pedigree information, phenotypic data, and genotype data (Progeny Software LLC, Delray Beach, FL)(12).

Consequently, some ID numbers were modified or their consanguineous status revised; hence, some FID numbers were changed. The family separation, remarriage, child adoption, and name change were double-checked till there were no more missing data in any FID.

We used the SAGE software to obtain the number of family members, generations number, and available pairwise relationships, including parent-offsprings, sibling-sibling, grandparents-grandchildren, avuncular, half-sibships, and cousins in each pedigree(9).

Ethnicity

Iran has a land area of 1,648,195 km2, making it the second-largest country in the Middle East and the 18th largest globally, with 82.8 million [15]. The majority of Iranians are Persian. However, millions of ethnic groups live in the region, including Turkic, Kurdish, Gilaki, Mazandarani, Lurs, Tats, Talysh, Arabs, and Baloch. The Central Intelligence Agency's (CIA) World Fact-book has estimated that around 79% of the population of Iran is a diverse Indo-European ethno-linguistic group that comprises speakers of the Iranian languages [16]. By 2008, the main ethnic groups in Iran were Persians (65%) (incl. Mazenderanis and Gilaks). The other groups consist of Azerbaijani Turks (16 %), Kurds (7 %), Lurs (6 %), Arabs (2 %), Baluchis (2 %), Turkmens (1 %), Turkish tribal groups such as the Qashqai (1 %) and non-Persian, non-Turkic groups such as Armenians, Assyrians, and Georgians (less than 1 %) [17].

In this study, the TCGS team submitted a questionnaire to the TLGS and CFS data centers to collect ethnicity data by categorizing individuals according to their ethnicity. In these two projects, individuals were sampled across all geographic regions in Iran, representing the original 31 provinces of the country in 5 governmental regions (Supplementary Table 1). Selected individuals should have Iranian ancestry, requiring that their parents were born in Iran. Most of them were born in Tehran (the TLGS population gathered in the east of Tehran), but their parents

or grandparents were born in different cities of different ethnicities. Participants of different ethnicities were coded according to their place of birth in the last three generations. If birthplace information was unavailable, the province status was based on residence.

Obstetric and Gynecology assessment

Two projects were established to research women's health according to the data extracted from women's questionnaires, hospitalization information, self-reports, comprehensive physical examination, and universal biochemical and hormonal assessment. **Androgen Excess Project(AEP)** is a clinical and genetic study initiated on women with androgen excess disorders who were referred to the Reproductive Endocrinology Research Center (2009). Androgen excess manifestations were evaluated using valid tools. The hirsutism scores were evaluated using the modified Ferriman-Gallwey (mFG) scoring scale. Acne was assessed based on its type, number, and distribution. AEP provides a valuable data set for investigating some gaps in knowledge in the context of a population-based cohort. This project is ongoing, and up to now, data on 2100 women have been collected. **The premature Ovarian Insufficiency Project (POIP)** is a clinical and genetic study initiated on women with Premature Ovarian Insufficiency who were referred to the Reproductive Endocrinology Research Center (2010). Over previous assessments, for hormonal estimation, fasting venous blood sampling was collected on the third day of the spontaneous or progesterone withdrawal menstrual period. Moreover, transvaginal ultrasound scans of the ovaries were performed for non-virgin participants by an experienced specialist on the same day

the blood samples were obtained. This project is continuous and has so far gathered data from 1190 women.

Infrastructure

The dynamic and rapidly evolving nature of different bioinformatics and biostatistics research imposes special requirements on a suitable computational infrastructure wherein all hardware, software, and network. This infrastructure has to be optimized to gain more effectiveness.

To this aim, we proposed hosting our data on an isolated local network of Virtual desktop (VM) infrastructure (Horizon View) with the capacity to support 15 zero clients. The Servers are two processing servers with 136 cores and 384 GB of memory. The hard disk Storage is 200 TB (HPE MSA Storage server). Operating systems include Linux Centos7 – Microsoft Windows. Moreover, our researchers are content with the technical support and maintenance of this design and the architecture of computational infrastructure.

Management

Organization: The TCGS unit is located in the Research Institute for Endocrine Sciences (RIES), Shahid Beheshti University of Medical Sciences (Tehran, Iran). The building consists of many units: a laboratory, bioinformatics section, admission, information, examination rooms, nutrition, and social worker units. An administrator and additional staff such as physicians, recruitment coordinators, electrocardiogram technicians, laboratory personnel, social workers, nurses, data collectors, and others ensure the following of the study protocol.

Steering committee

This committee consists of a principal investigator, program coordinators, additional researchers, a manager, and the section director in the TCGS units, all meeting regularly. This committee is responsible for approving the study design, policies, and decisions and oversees the administrative aspects of the TLGS Research Group.

References

1. Azizi, F. *et al.* Prevention of non-communicable disease in a population in nutrition transition: Tehran Lipid and Glucose Study phase II. *Trials* **10**, 5 (2009).
2. Azizi, F., Zadeh-Vakili, A. & Takyar, M. Review of Rationale, Design, and Initial Findings: Tehran Lipid and Glucose Study. *Int J Endocrinol Metab* **16**, e84777 (2018).
3. Barzin, M. *et al.* Bariatric Surgery for Morbid Obesity: Tehran Obesity Treatment Study (TOTS) Rationale and Study Design. *JMIR Res Protoc* **5**, e8 (2016).
4. Hedayati, M., Zarif Yeganeh, M., Sheikholeslami, S. & Afsari, F. Diversity of mutations in the RET proto-oncogene and its oncogenic mechanism in medullary thyroid cancer. *Crit Rev Clin Lab Sci* **53**, 217–227 (2016).
5. Mahdieh, N., Rabbani, B., Wiley, S., Akbari, M. T. & Zeinali, S. Genetic causes of nonsyndromic hearing loss in Iran in comparison with other populations. *Journal of Human Genetics 2010 55:10* **55**, 639–648 (2010).
6. Esfahani, F. H., Asghari, G., Mirmiran, P. & Azizi, F. Reproducibility and relative validity of food group intake in a food frequency questionnaire developed for the Tehran Lipid and Glucose Study. *J Epidemiol* **20**, 150–158 (2010).
7. Mirmiran, P., Hosseini Esfahani, F., Mehrabi, Y., Hedayati, M. & Azizi, F. Reliability and relative validity of an FFQ for nutrients in the Tehran lipid and glucose study. *Public Health Nutr* **13**, 654–662 (2010).
8. Montazeri, A. *et al.* The 12-item medical outcomes study short form health survey version 2.0 (SF-12v2): a population-based validation study from Tehran, Iran. *Health Qual Life Outcomes* **9**, (2011).
9. Loewen, H. LibGuides: Assessment Tools: How to score version 2 of the SF-12 health survey (with a supplement documenting version 1).

10. SHAYEGHIAN, Z., AMIRI, P., VAHEDI-NOTASH, G., KARIMI, M. & AZIZI, F. Validity and Reliability of the Iranian Version of the Short Form Social Well Being Scale in a General Urban Population. *Iran J Public Health* **48**, 1478–1487 (2019).

11. Keyes, C. L. M. Social well-being. *Soc Psychol Q* **61**, 121–137 (1998).

12. P., A. *et al.* Designation And Psychometric Assessment Of A Comprehensive Spiritual Health Questionnaire For Iranian Populations. vol. 8 25–55 Preprint at (2015).

13. ICD-11 for Mortality and Morbidity Statistics. https://icd.who.int/browse11/l-m/en.

14. Truett, G. E. *et al.* Preparation of PCR-quality mouse genomic DNA with hot sodium hydroxide and tris (HotSHOT). *Biotechniques* **29**, 52,54 (2000).

15. Iran At a Glance. https://www.investiniran.ir/en/Iran-At-a-Glance.

16. Iranians in the World. https://www.atlasofhumanity.com/iranians.

17. The International Handbook of the Demography of Race and Ethnicity - Google Books.

Table 1: Participants by region, province, ethnicity, and chip data

*Region	Province	Population	Total Ethnicity	Chip data Ethnicity
1	Alborz Golestan Mazandaran Qazvin Qom Semnan Tehran	2712400 1868819 3283582 1273761 1292283 702360 13267637	Total N: 2255 Arab(2), Arab-Persian(5), Gilak(114), Kurd(1), Lur-Persian(1), Persian(2072), Tat(10), Tat-Turk(28), Turk(12), Turkaman(6), Others(4)	Total N:1351 Arab(2), Arab-Persian(3), Gilak(80), Kurd(1), Lur-Persian(1),Persian(1218), Tat(10), Tat-Turk(18), Turk(11), Turkaman(3), Others(4)
2	Bushehr Chaharmahal and Bakhtiari Fars Hormozgan Isfahan	1163400 947763 4851274 1776415 5120850	Total N: 1221 Arab(26), Arab-Persian(2), Lur(26), Persian(1157), Qashqai(3), Turk(2), Others(5)	Total N: 743 Arab(22), Arab-Persian(2), Lur(23), Persian(688), Qashqai(2), Turk(1), Others(5)
3	Ardabil East Azerbaijan Gilan Kordestan West Azerbaijan Zanjan	1270420 3909652 2530696 1603011 3265219 1057461	Total N: 810 Gilak(138), Kurd(18), Lur(1), Mix(1), Persian(9), Persian-Turk(2), Tat(1), Turk(633), Others(7)	Total N: 504 Gilak(65), Kurd(17), Lur(1), Mix(1), Persian(6), Persian-Turk(2), Tat(1), Turk(404), Others(7)
4	Hamadan Ilam Kermanshah Khuzestan Lorestan Markazi	1738234 580158 1952434 4710509 1760649 1429475	Total N: 1040 Arab(50), Gilak(3), Kurd(40), Lur(106), Lur-Lak(17), Lur-Persian(3), Persian(803), Tat-Turk(1), Turk(9), Others(8)	Total N: 632 Arab(23), Gilak(2), Kurd(30), Lur(71), Lur-Lak(10), Lur-Persian(3), Persian(476), Tat-Turk(1), Turk(9), Others(7)
5	Kerman North Khorasan Razavi Khorasan Sistan and Baluchestan South Khorasan Yazd	3164718 863092 6434501 2775014 768898 1138533	Total N: 458 Balouch(28), Kurd(10), Persian(414), Tat(2), Turk(1), Turkaman(3)	Total N: 291 Balouch(24), Kurd(7), Persian(254), Tat(2), Turk(1), Turkaman(3)
Foreigner	Afghanistan Azerbaijan Iraq Russia Turkmenistan Yemen		Total N: 36 Afghan(3), Arab(6), Arab-Persian(5), Mix(4), Russia(7), Turk(6), Turkaman(5)	Total N: 24 Afghan(3), Arab(4), Arab-Persian(4), Mix(4), Russia(6), Turk(3)
	Unknown		Total N: 357 Persian(274), Turk(83)	Total N: 306 Persian(234), Turk(72)
Total			6177	3851

* Five main governmental regions in Iran: 1: Tehran,2: Isfahan,3: Tabriz,4: Kermanshah, 5: Mashhad.

Table 2: Missing call rate of 652,341 SNPs among TCGS participants

Missing call rate	Frequency	Cumulative Frequency	Cumulative Percentage
<0.01	598,955	598,955	0.9182
0.01 - 0.02	27,281	626,236	0.9600
0.02 - 0.03	8,672	634,908	0.9733
0.03 - 0.04	3,921	638,829	0.9793
0.04 - 0.05	2,132	640,961	0.9826
0.05 - 0.06	1,275	642,236	0.9845
0.06 - 0.07	797	643,033	0.9857
0.07 - 0.08	597	643,630	0.9867
0.08 - 0.09	398	644,028	0.9873
0.09 - 0.10	298	644,326	0.9877
0.1 - 0.2	958	645,284	0.9891
> 0.2	7,057	652,341	1
Total SNPs	652,341		

Table 3: Per-individual missingness among 14,835 TCGS participants

Individuals missing call rate	Frequency	Cumulative Frequency	Cumulative Percentage
<0.01	14,297	14,297	0.9637
0.01 - 0.02	514	14,811	0.9984
0.02 - 0.2	2	14,813	0.9985
> 0.2	22	14,835	1
Total	14,835		

Table 4: Missing call rate of 652,341 SNPs among TCGS participants

MAF categories	Frequency	Cumulative Frequency	Cumulative Percentage
<0.0001	602	602	0.0009
0.0001 - 0.001	8,785	9,387	0.0143
0.001 - 0.01	18,718	28,105	0.0431
0.01 - 0.05	47,231	75,336	0.1155
0.05 - 0.1	68,919	144,255	0.2211
> 0.1	508,086	652,341	1
Total SNPs	652,341		

Table 5: Relationship between MAF and variants' impact among genotyped data

MAF	LOW (Synonymous, stop retained, 3'/5' UTR)	MODERATE (missense, splice region)	LoF (Frameshift, start/stop gained, splice acceptor/ donor, initiator codon)	Other (Intronic, intergenic)	Total
<0.0001	199 (2.265 %)	304 (3.460 %)	6(0.0683 %)	8276(94.206 %)	8785
0.0001 - 0.001	518 (2.768 %)	651 (3.478 %)	21(0.112 %)	17524(93.641 %)	18714
0.001 - 0.01	1062 (2.249 %)	1279 (2.7081 %)	28(0.059 %)	44860(94.984 %)	47229
0.01 - 0.05	1060 (1.538 %)	1201 (1.743 %)	39(0.056 %)	66612(96.662 %)	68912
0.05 - 0.1	12 (1.993 %)	34 (5.648 %)	2(0.332 %)	554(92.026 %)	602
> 0.1	6267 (1.233 %)	5901 (1.161 %)	222(0.043 %)	495686(97.561 %)	508076
Total	9118	9370	318	633512	652318

Table 6: Top five ICD-11 chapters with the most reported mortality and morbidity conditions in the TCGS cohort

Diseases of the circulatory system	N=11190, 17.70%
Hypertensive heart disease (BA01)	22.2%
Angina pectoris (BA40)	17.9%
Diseases of the coronary artery (BA8Z)	13.8%
Pregnancy childbirth or the puerperium	N=10194, 16.12%
Delivery (JB22, JB2Z)	63.36%
Diabetes mellitus in pregnancy (JA63)	25.95%
Gestational hypertension (JA23)	7.37%
Endocrine nutritional or metabolic diseases	N=10194, 15.36%
Hyperlipoproteinemia (5C80.Z)	21.74%
Nontoxic goiter (5A01)	15.51%
Hypothyroidism (5A00)	14.09%
Diseases of the digestive system	N=5209, 7.85%
Appendicitis (DB10)	18.0%
Gastritis (DA42)	13.7%
Hernias (DD5Z)	10.9%
Genitourinary system	N=4794, 7.22%
Calculus of the kidney (GB70.0Z)	17.10%
Hyperplasia of the prostate (GA90)	10.20%
Unspecified ovarian cysts (GA18.6)	8.74%

Table 7: Characteristics of traits included in ROH analysis

Trait Group	Full trait name	Trait	Units	Required covariates	N
Anthropometric	Waist-Hip ratio	whr	-	sex + age + age^2 + bmi	11710
Anthropometric	Body mass index	bmi	kg / m2	sex + age + age^2	11716
Anthropometric	Weight	weight	kg	sex + age	11724
Anthropometric	Height	height	m	sex + age	11731
Behavioural	Ever smoked	ever_smoked_glm	TRUE / FALSE	sex + (yob-1950) + (yob-1950)^2 + (yob-1950)^3	11678
Blood Lipids	LDL cholesterol	ldl	mmol/L	sex + age + age^2	11459
Blood Lipids	Triglycerides	log.triglyc	log(mmol/L)	sex + age + age^2	11678
Blood Lipids	Total cholesterol	tot_chol	mmol/L	sex + age + age^2	11690
Blood Lipids	HDL cholesterol	hdl	mmol/L	sex + age + age^2	11760
Blood pressure	Diastolic blood pressure	bp_dia	mmHg	sex + age + age^2 + bmi	11655
Blood pressure	Systolic blood pressure	bp_sys	mmHg	sex + age + age^2 + bmi	11657
Cognition	Education attained	edu	years	sex + (yob-1950) + (yob-1950)^2 + (yob-1950)^3	7251
Electrocardiography	Heart rate	hr	bpm	sex + age + age^2	9735
Female reproductive	Age at menarche	age_menarche	years	(yob-1950) + (yob-1950)^2 + (yob-1950)^3	4158
Glycaemic	Fasting plasma glucose	fpg	mmol/L	sex + age + age^2 + bmi	9873
Inflammatory	Interleukin-6	log.il6	log(pg/mL)	sex + age + age^2 + bmi	320
Inflammatory	Hs-CRP	log.hscrp	log(nmol/L)	sex + age	653
Renal function	eGFR	log.egfr	log(mL/min/1.73 m)	sex + age + age^2	11077

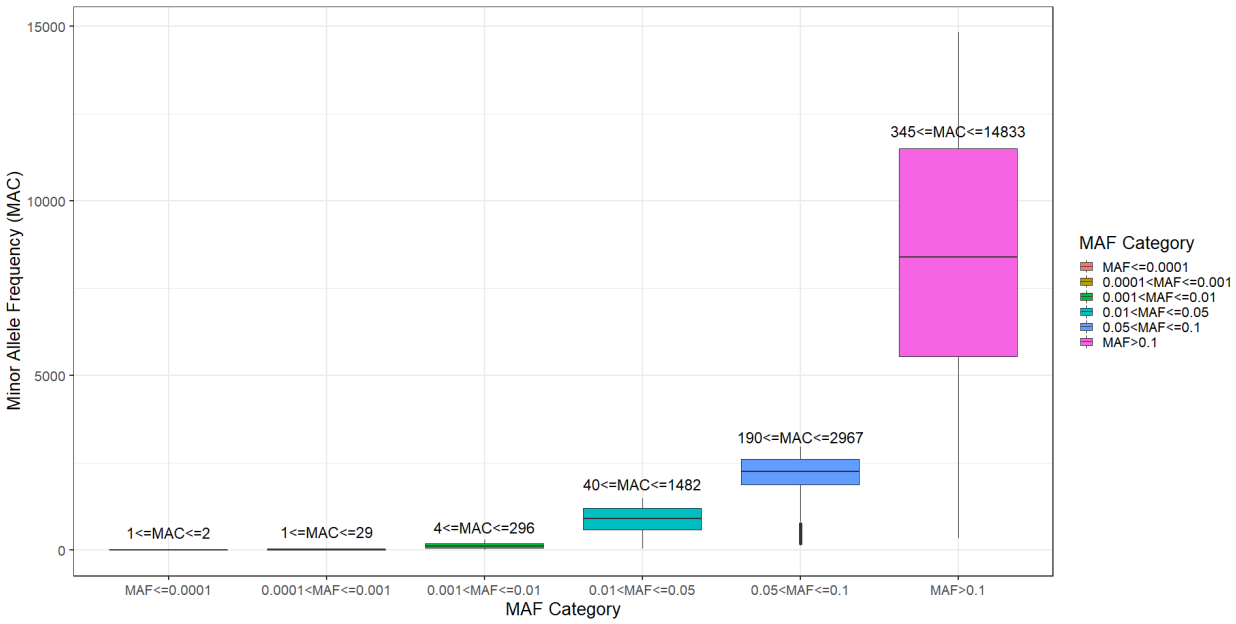

Figure 1: Relationship between MAF and MAC in TCGS genotyped data

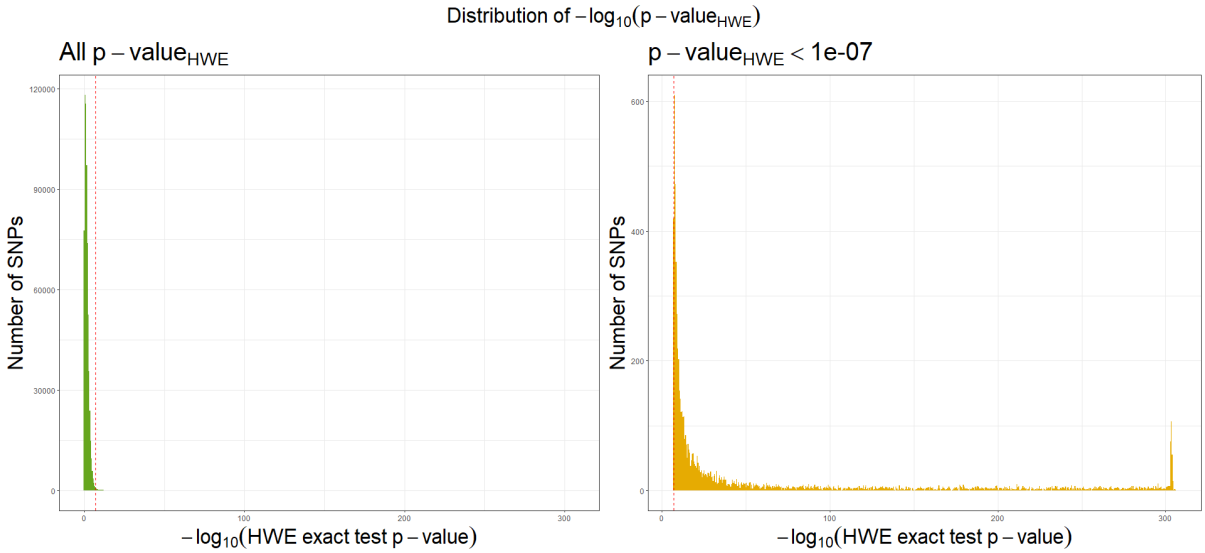

Figure 2: Hardy-Weinberg Equilibrium assumption in TCGS

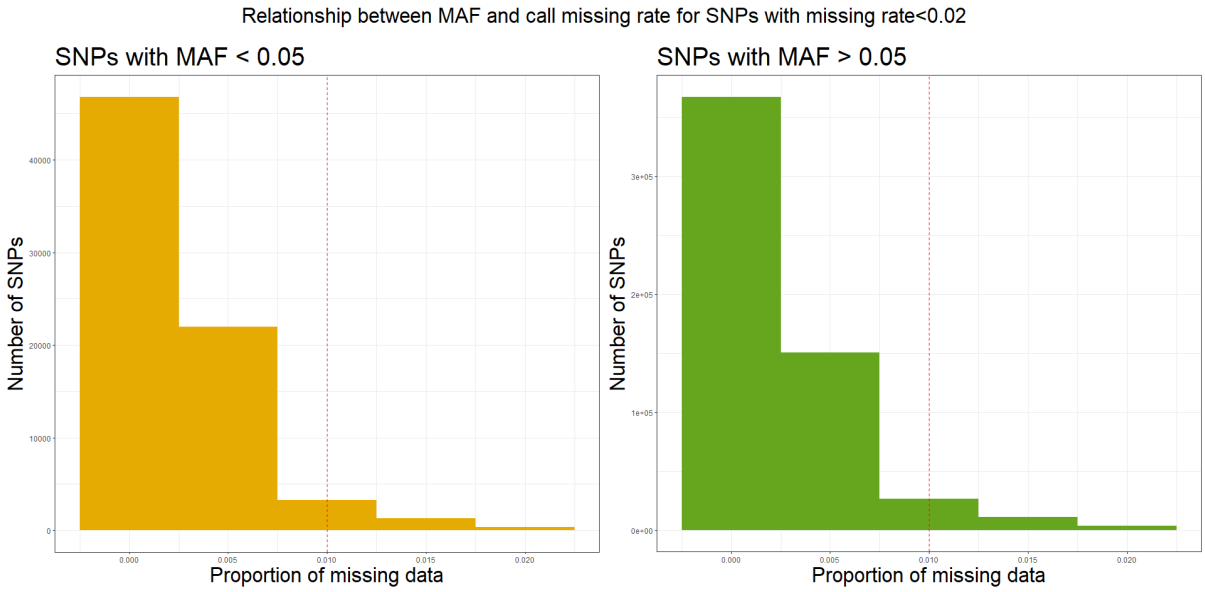

Figure 3: Missingness and Minor Allele Frequency (MAF)

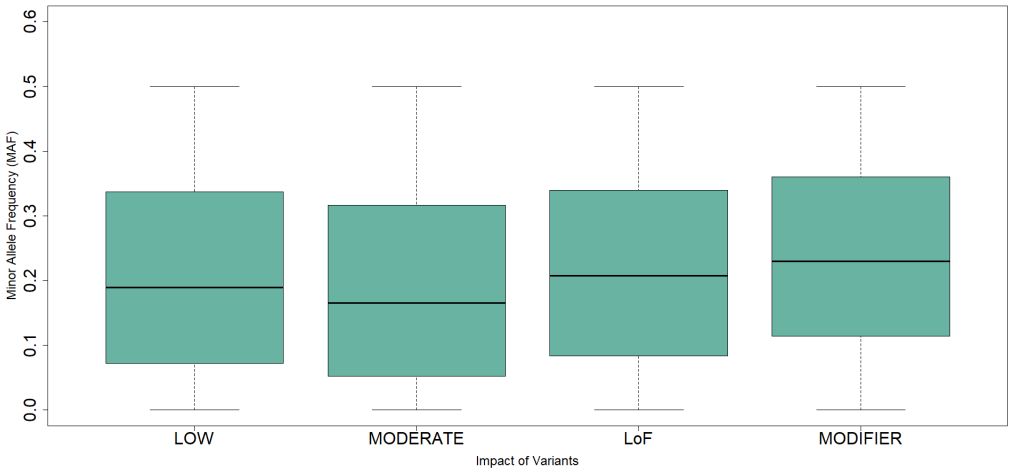

Figure 4: Distribution of MAF through the different impact of the variants

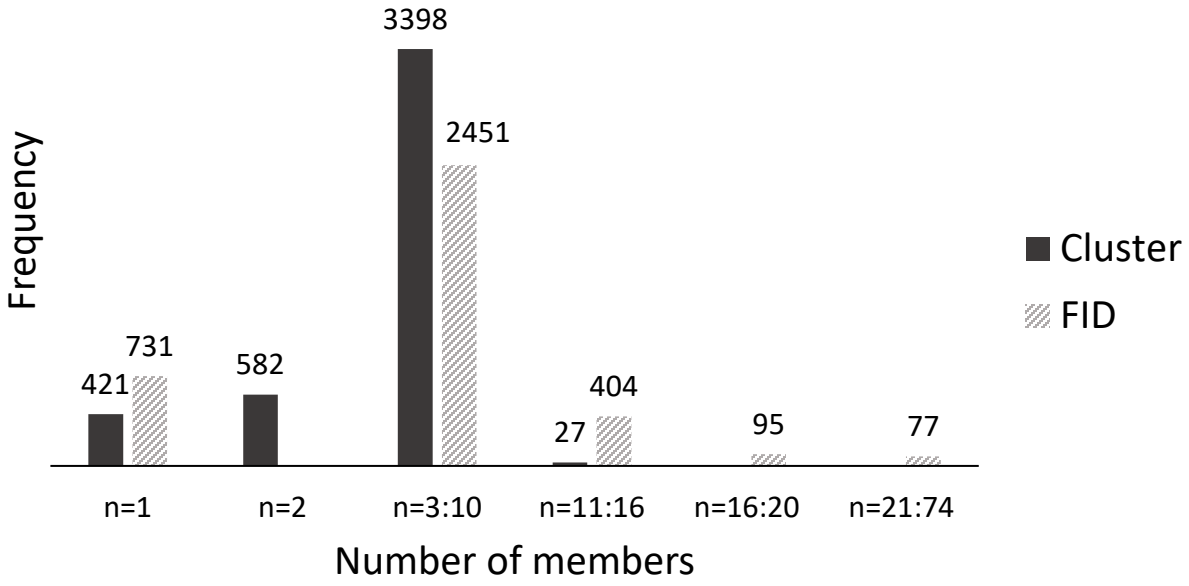

Figure 5: Comparison between family identification numbers and clusters

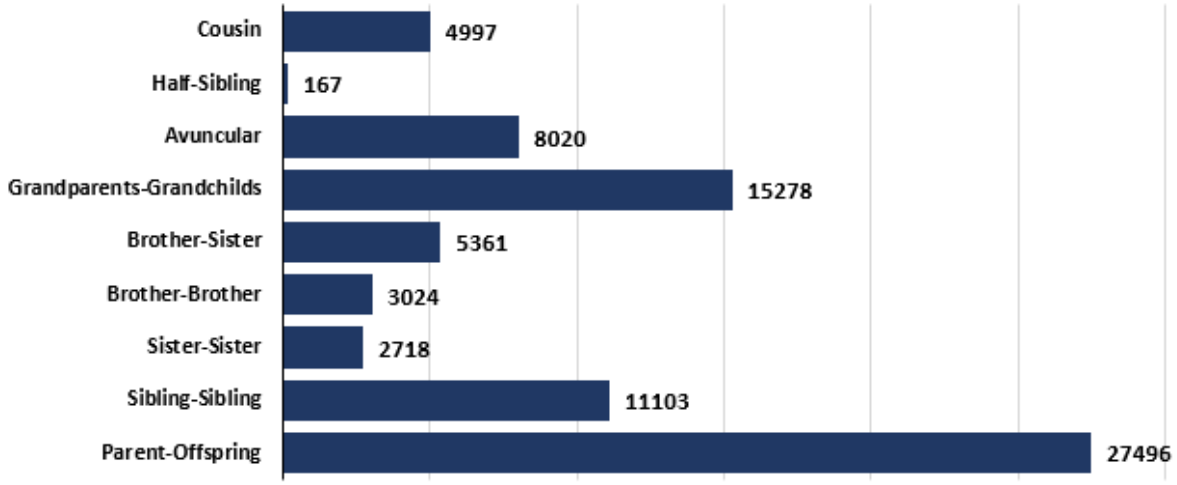

Figure 6: The relationship distribution in TCGS families

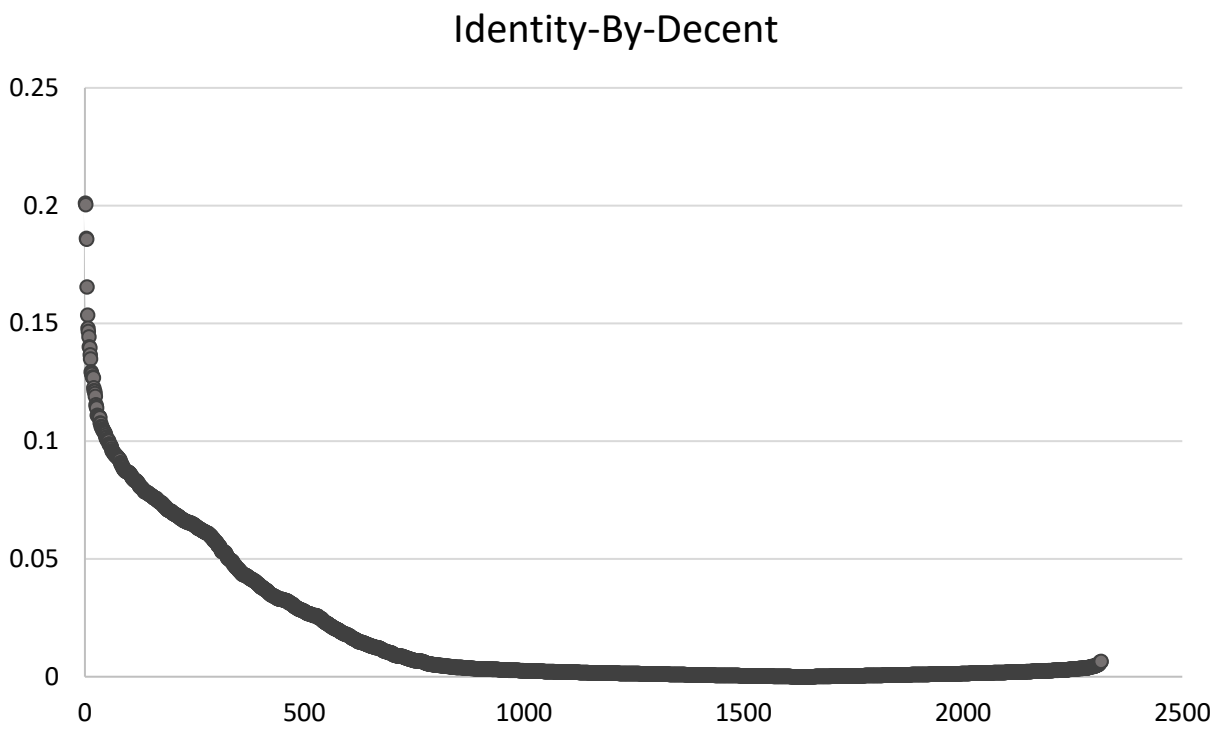

Figure 7. The genomic relationship matrix of spouses