Comprehensive Analysis of Genetic and Environmental Factors Influencing Type 2 Diabetes in the Spanish Population with NGS and the SEQENS Algorithm

Celeste Moya-Valera[1], Alba Valverde-Porcar[2], Pedro Pons-Suñer[2], Francisco Lara-Hernandez1, Mariana de Jesús Gallardo-Espinoza1, Maria Elena Quiroz-Rodriguez[1], Joaquim Arlandis[2], J.Ramón Navarro-Cerdán[2], Ana-Barbara Garcia-Garcia[1,3],Felipe Javier Chaves Martínez[1,3]

1 Genomic and Diabetes Unit, INCLIVA Biomedical Research Institute, Valencia 46010, Spain, 2 Instituto Tecnológico de Informática (ITI), Universitat Politècnica de València, Camí de Vera, s/n, Valencia 46022, Spain, 3 CIBERDEM, ISCIII, Madrid, Spain.

Contact:
cmoya@incliva.es
@cmoyavalera.bsky.social

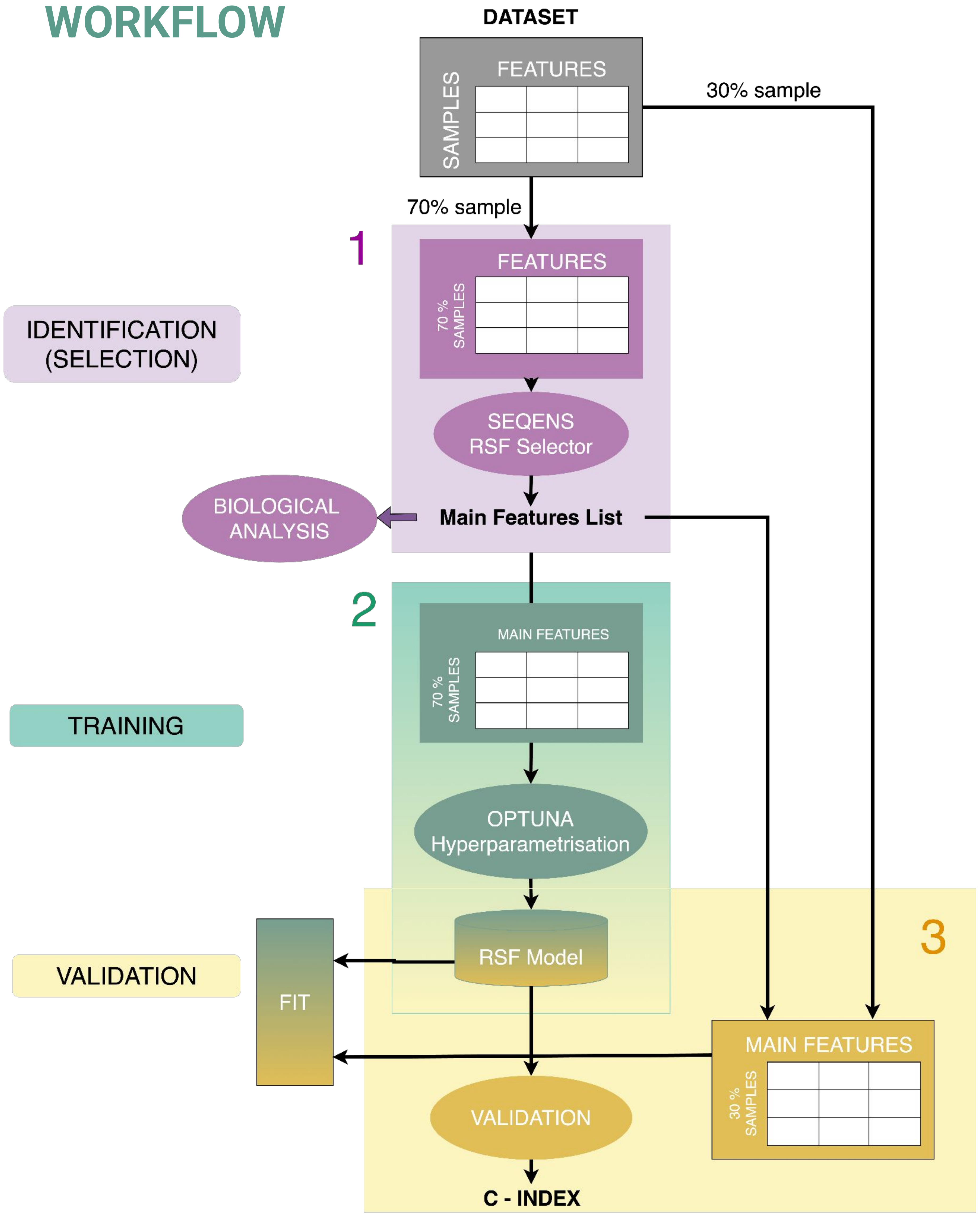
INTRODUCTION

Type 2 diabetes mellitus (T2D) is a major cause of mortality and significantly reduces quality of life due to its progressive effects on cardiometabolic health. While genetic predisposition is estimated to account for over 50% of T2D risk, the majority of contributing genetic variants remain unidentified. The identification of heterogeneous biomarkers, encompassing genetic variants and their interactions with environmental, clinical, and anthropometric factors, is imperative to enhance our comprehension of the etiology of T2D. The objective of this study is to identify genetic variants associated with T2D using Next-Generation Sequencing (NGS) and to explore their interaction with non-genetic variables in a representative cohort of the Spanish population. By leveraging data from the DI@BET.ES study (n=4200), which integrates clinical, environmental, and genetic information, a comprehensive approach is adopted to uncover biomarkers of T2D susceptibility (Table 1). SEQENS, an optimized feature-selection methodology for high-dimensional data, enhances the capture of genetic-phenotypic relationships. It surpasses traditional encoding limitations, improving predictive models for complex diseases like T2D in diverse populations.

Table 1. DI@BET.ES dataset description.

Databases	Nº samples	Nº features
Genetic	4823	540
Environmental	4200	600
Complete (Genetic + Environmental)	4200	1140

WORKFLOW



DATASET

FEATURES

SAMPLES

30% sample

70% sample

1

IDENTIFICATION (SELECTION)

FEATURES

70 % SAMPLES

SEQENS RSF Selector

BIOLOGICAL ANALYSIS

Main Features List

2

TRAINING

MAIN FEATURES

70 % SAMPLES

OPTUNA Hyperparametrisation

3

VALIDATION

FIT

RSF Model

MAIN FEATURES

30 % SAMPLES

VALIDATION

C - INDEX

Figure 1. Workflow diagram illustrating the relevant feature extraction process (utilising sections 1, 2, and 3) and the full feature extraction process (based on sections 2 and 3 only).

CONCLUSIONS

- The use of SEQENS, a methodology optimized for high-dimensional data, has enabled the identification of complex relationships between genetic and environmental factors associated with disease risk.
- Six common variants relevant to the genetic model have been identified: *KCNA1*, *CYP4F34P*, *MUC6*, *GET4*, *PLEKHG6*, and *CACNA2D4*.
- Eight relevant variables of the environmental model have been identified: fasting glucose, first-degree consanguinity, TG, BMI, weight, ggpt, PM2.5 pollutant particles, and HDL.
- Twelve relevant variables in the interaction of environmental and genetic variables: fasting glucose, working outside the home, first-degree consanguinity, weight, waist-to-hip ratio (WHR), fasting insulin, HDL, monthly beer consumption, BMI, LDL, nickel, and TG.

FUNDING - PI17/00544, PI21/00506 and PI24/00338 (Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación and co-funded by the European Union), CIBER -Consorcio Centro de Investigación Biomédica en Red-CB-07/08/0018, Instituto de Salud Carlos III. IMAMCA/2024/11 (IVACE, Generalitat Valenciana).

RESULTS

Genetic model

Table 2. Relevant variables in the genetic model.

Variable	Votes	% Seq.	p-value adj.	Gene	Ontology	MAF
rs2227910	57	62.63	<0.0001	*KCN1*	Ion transport	0.47
rs10220060	40	43.95	<0.0001	*CYP4F34P*	Ion transport	0.15
rs200089063	34	37.36	<0.0001	*MUC6*	Glycosylation	0.47
rs115450168	31	34.06	<0.0001	*GET4*	Protein location	0.14
rs4149651	30	32.96	<0.0001	*PLEKHG6*	GTPase Cycle	0.02
rs2286372	29	31.86	<0.0001	*CACNA2D4*	Ion transport	0.32

Environmental model

Table 3. Relevant variables in the environmental model.

Variable	Votes	% Seq.	p-value adj.
Fasting glucose (mg/dl)	128	100	<0.0001
First-degree consanguinity	94	73.43	<0.0001
TG (mg/dl)	58	45.31	<0.0001
BMI (kg/m2)	41	32.03	<0.0001
Weight (kg)	36	28.12	<0.0001
ggpt (ukat/l)	30	23.43	<0.0001
Particle contamination	28	21.09	0.0014
HDLc (mg/dl)	27	21.09	0.0014

Combined model

Table 4. Relevant variables in the combined model.

Variable	Votes	% Seq.	p-value adj.
Fasting glucose (mg/dl)	200	100	<0.0001
Work outside home	106	53	<0.0001
First-degree consanguinity	64	32	<0.0001
Weight (kg)	46	23	<0.0001
Wrist (cm)	45	22.5	<0.0001
Fasting insulin (mg/dl)	35	17.5	<0.0001
HDLc (mg/dl)	30	15	<0.0001
Monthly beer intake	24	12	<0.0001
BMI (kg/m2)	20	10	<0.0001
LDLc (mg/dl)	19	9.5	<0.0001
rs7816608	17	8.5	<0.0001
Nickel (ni)	17	8.5	<0.0001
TG (mg/dl)	16	8	0.0080

Feature Interaction Test

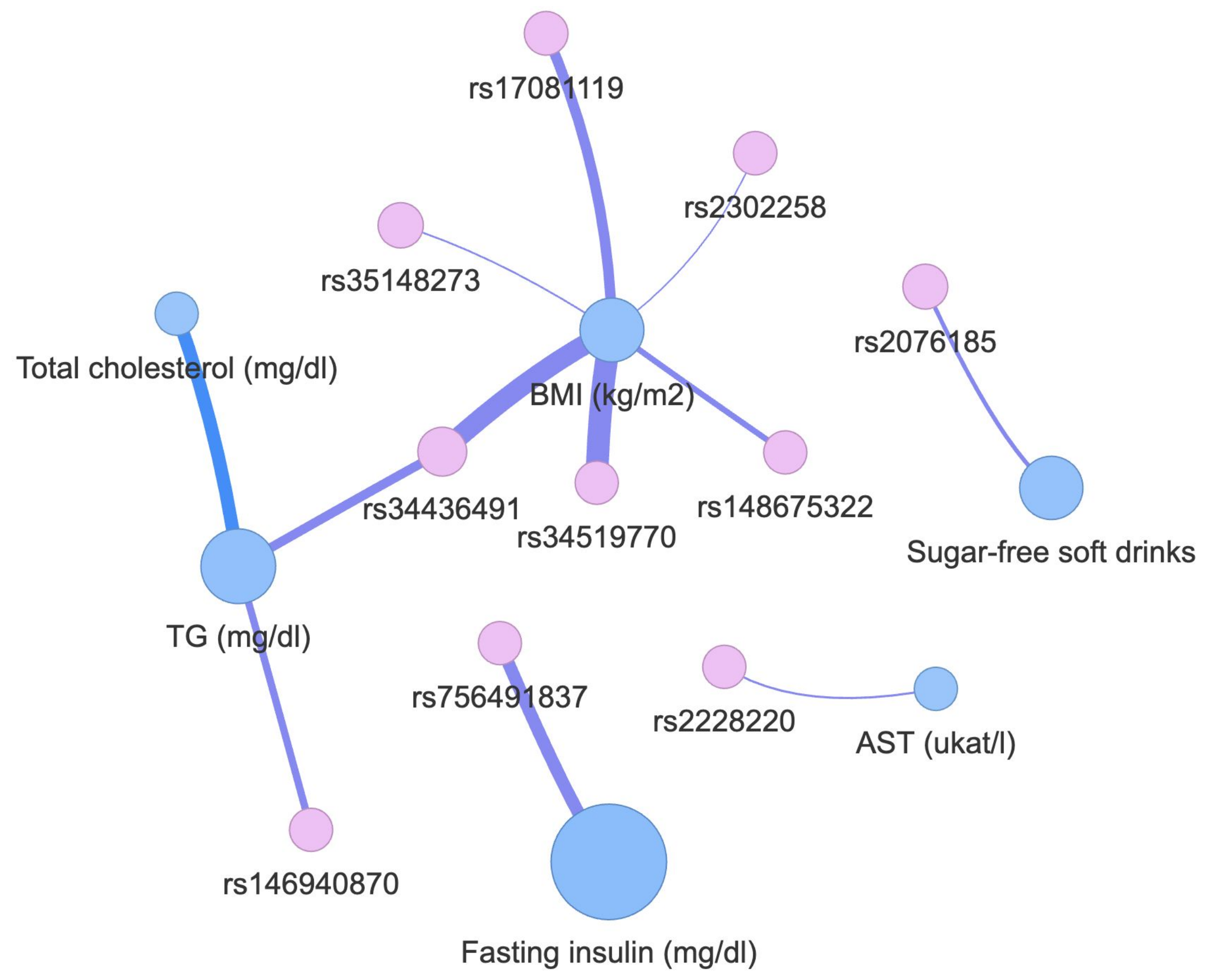


Figure 2. Feature Interaction Test graph of both environmental and genetic variables.

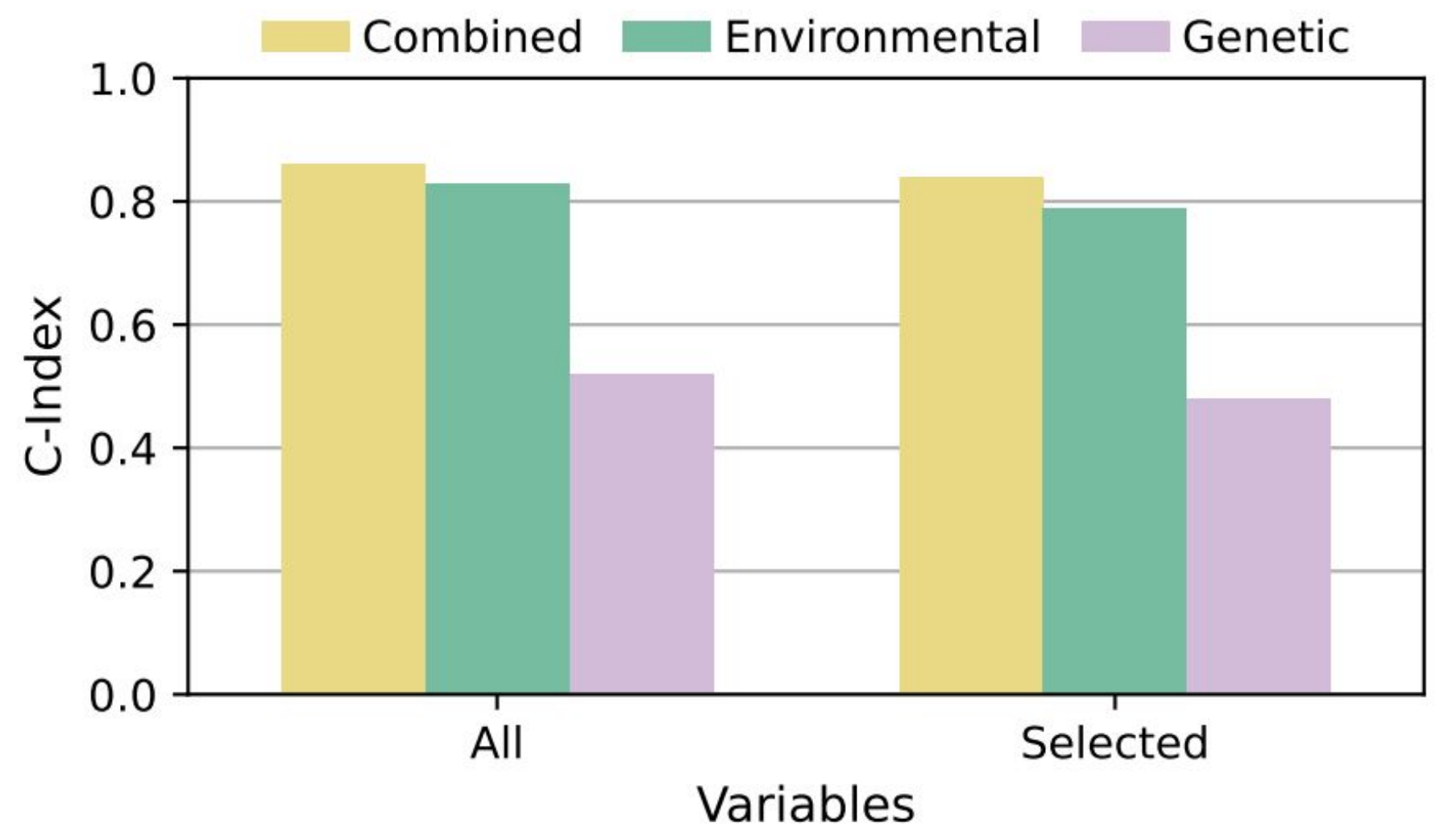


Figure 3. C-index values of combined, environmental and genetic models using all and SEQENS-selected variables.