Deep learning on genetic data with Diet Network and its application to a complex phenotype

BIRS MEETING, JUNE 2022
DEEP LEARNING FOR GENETICS, GENOMICS AND METAGENOMICS:
LATEST DEVELOPMENTS AND NEW DIRECTIONS

Université de Montréal

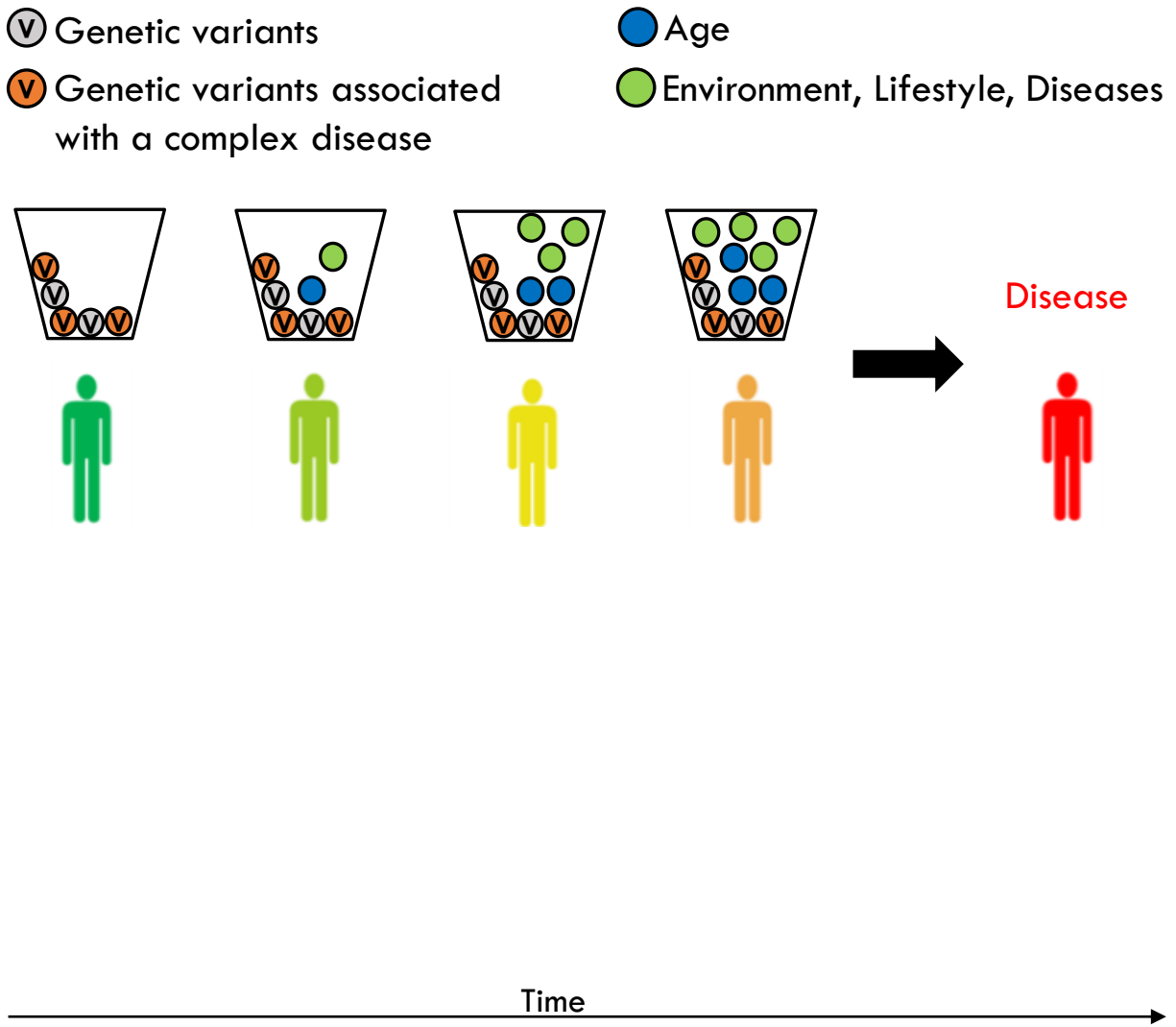
Camille Rochefort-Boulanger
MHI-omics
Julie Hussin and Yoshua Bengio

Context

Complex phenotypes
Interaction of genetic variants and environmental factors

Large number of genetic variants, each making only a small contribution to the final phenotype

Ex: Height, cardiovascular diseases, type II diabetes, ...



Context

Complex phenotypes

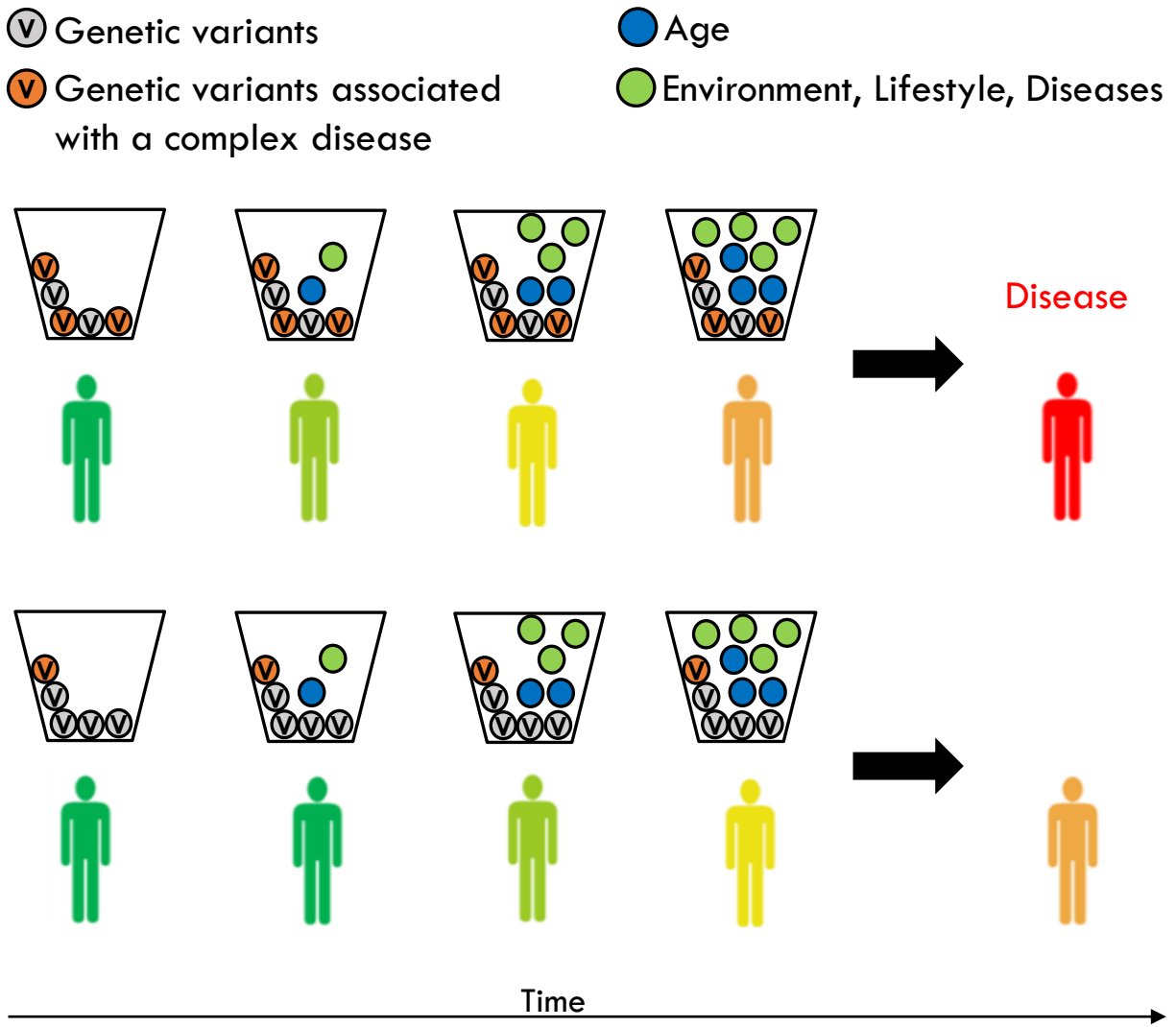
Interaction of genetic variants and environmental factors

Large number of genetic variants, each making only a small contribution to the final phenotype

Ex: Height, cardiovascular diseases, type II diabetes, ...

Genetic susceptibility

Depending on their genetic variants, some people are more or less at risk to develop a given complex disease

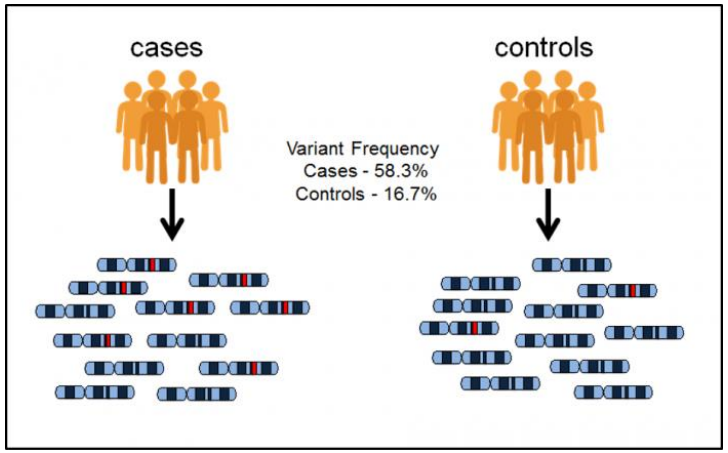


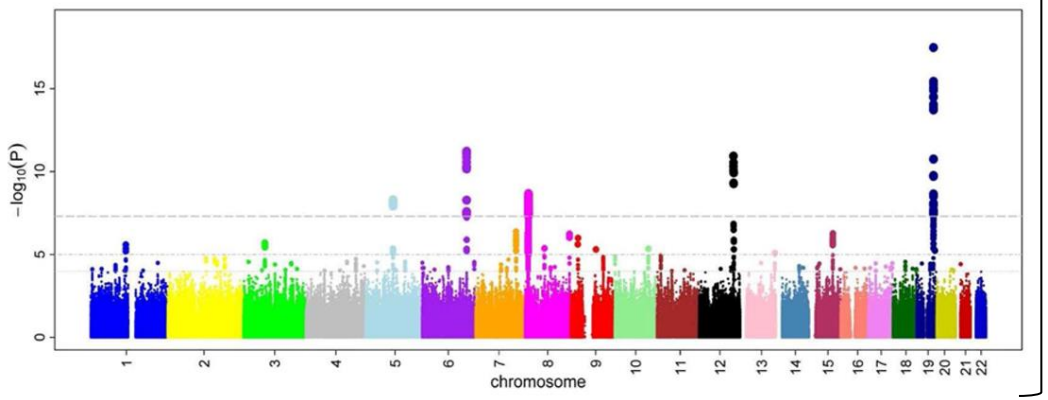
Context

How can we identify people at risk of developing a given complex disease

Genome Wide Association Studies (GWAS)

Statistical test of association between genetic variants (SNPs) and a complex phenotype





SNPs

- P-value of association
- Estimated effect

GWAS Catalog EMBL-EBI 2017

Genetic risk score

Computes the risk of developing a complex disease for an individual

individual

Phenotype associated SNPs (GWAS)

SNPs effect (GWAS)

Individual's gonotype {0,1,2}

$$GRS_i = \sum_{j=0}^{m} \beta_j \cdot X_{ij}$$

Context

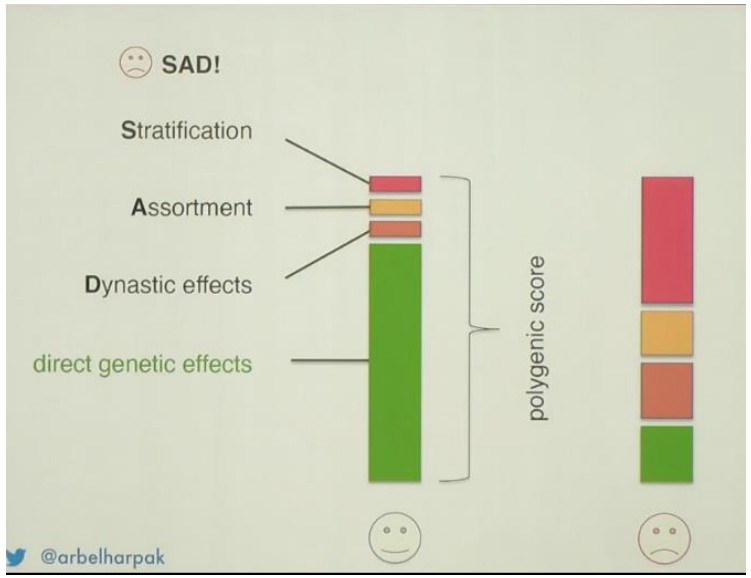
GWAS and Genetic risk scores limitations

GWAS test for genetic effects but are confounded by

- **S**tratification (Population structure)
- **A**ssortative mating
- **D**ynastic (indirect) parental genetic effects

  → SAD effects

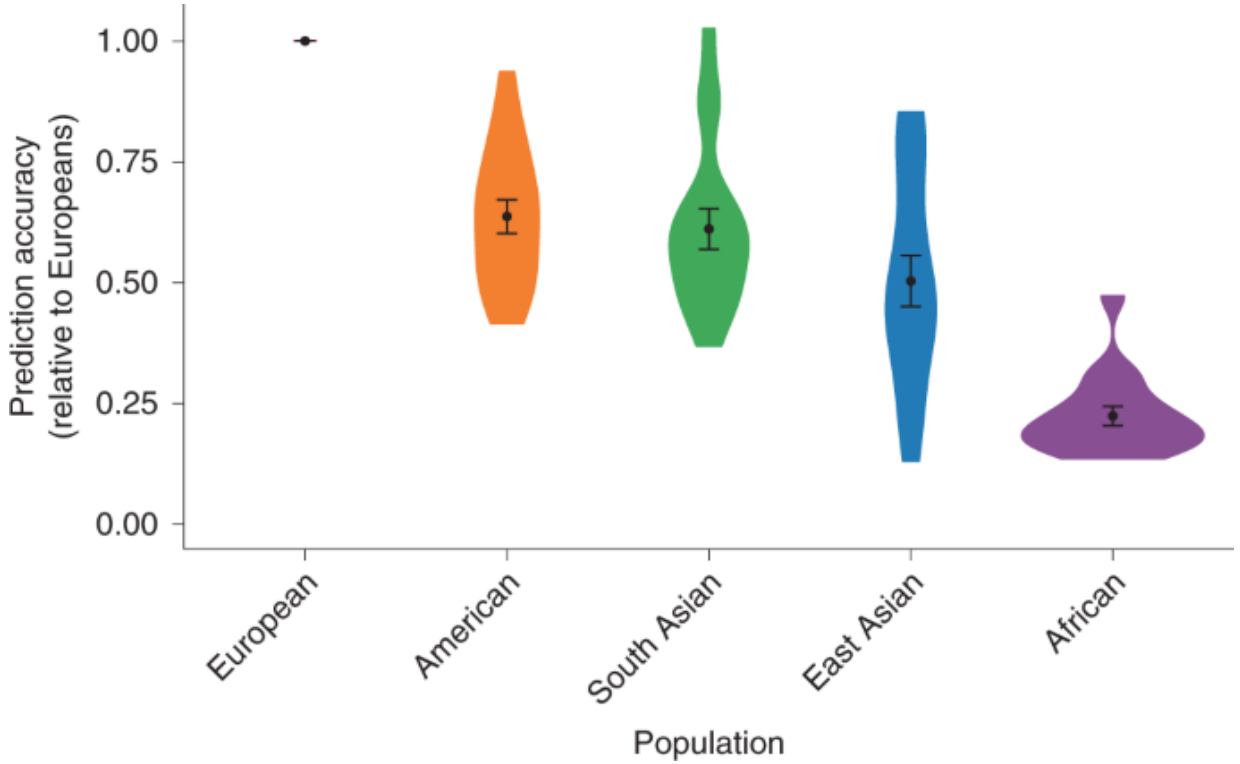
GWAS cannot detect the interaction between genetic variants



Arbel Harpak, Biology of Genomes 2022

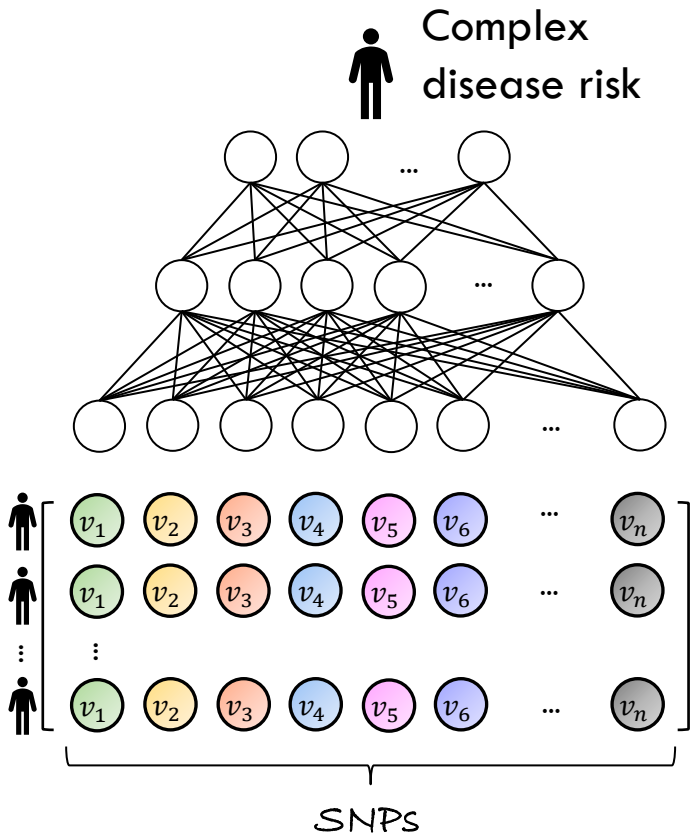
Genetic risk scores are not generalizable across populations

Prediction accuracy relative to European-ancestry individuals across 17 quantitative traits and 5 continental populations in the UKBB



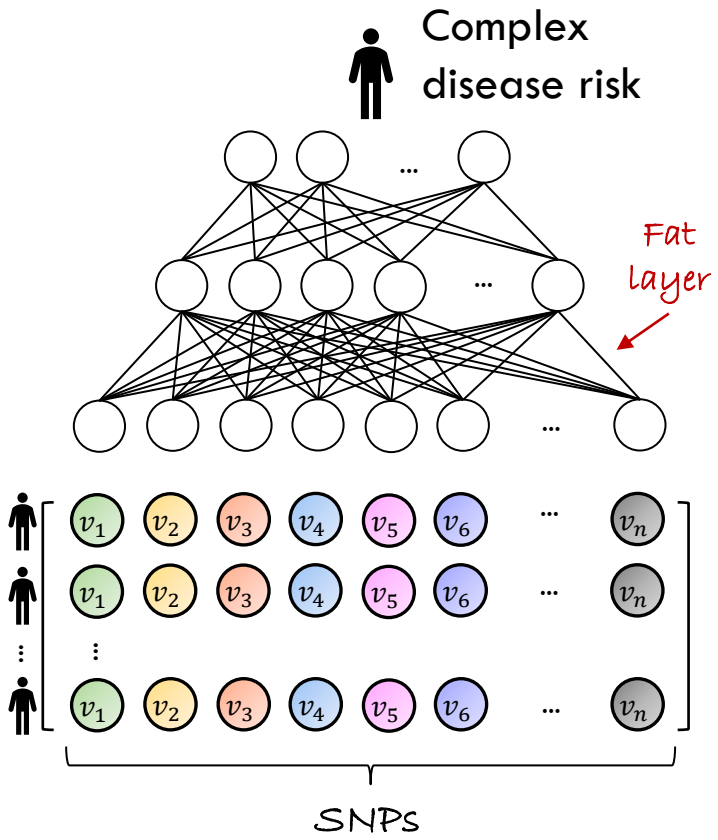
Martin al. Nature Genetics 2019

Deep learning using genotype data



Deep learning using genotype data

Fat data : number of features (SNPs) is order of magnitude higher than the number of samples (individuals) → Overfitting



Diet Network

Romeo et al. ICLR 2017

Deep learning using genotype data

Fat data : number of features (SNPs) is order of magnitude higher than the number of samples (individuals) → Overfitting

Yoshua Bengio

Deep learning using genotype data

Fat data : number of features (SNPs) is order of magnitude higher than the number of samples (individuals) → Overfitting

Diet Networks

Developed and tested on a genetic ancestry classification task in 1000G

3450 individuals, ~300K SNPs, 26 populations

Yoshua Bengio

Diet Networks

Romeo et al. ICLR 2017

Result I : Generalization capability

Can the Diet Network generalize its predictions in independent datasets

Train

1000G

Populations: - African; - American; - East Asian; - European; - South Asian;

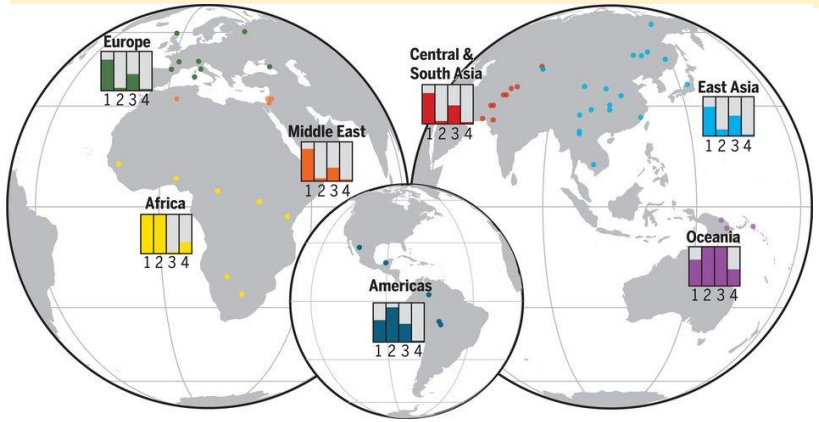
~300K SNPs

3450 individuals

26 populations

Test #1

Human Genome Diversity Project (**HGDP**)

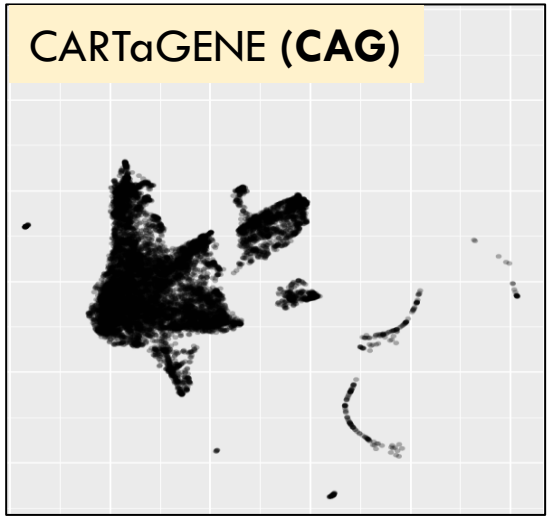


~250K/300K SNPs

Population dataset

Test #2

CARTaGENE (**CAG**)

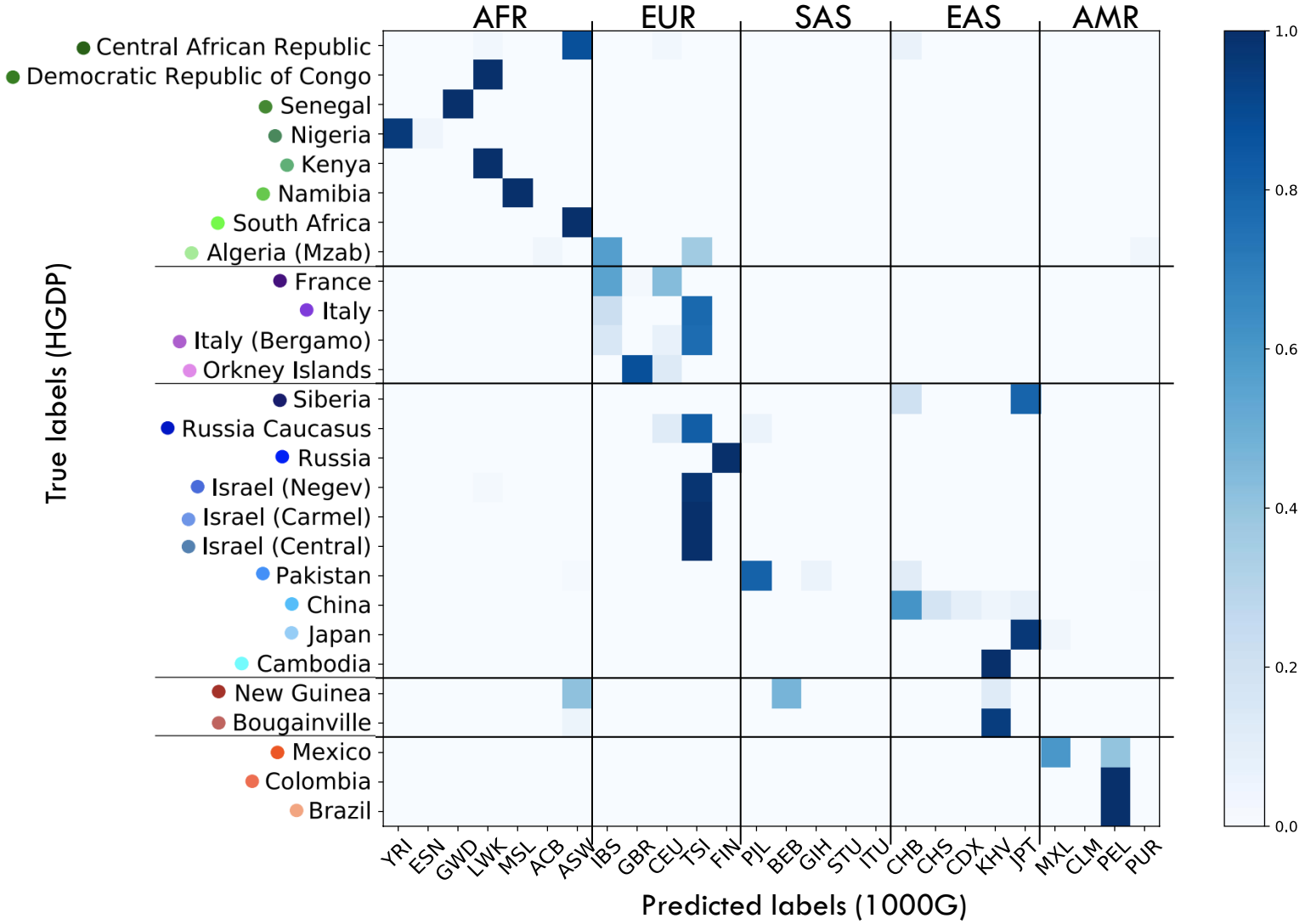


~173K/300K SNPs

Quebec biobank with self-reported ethnicity

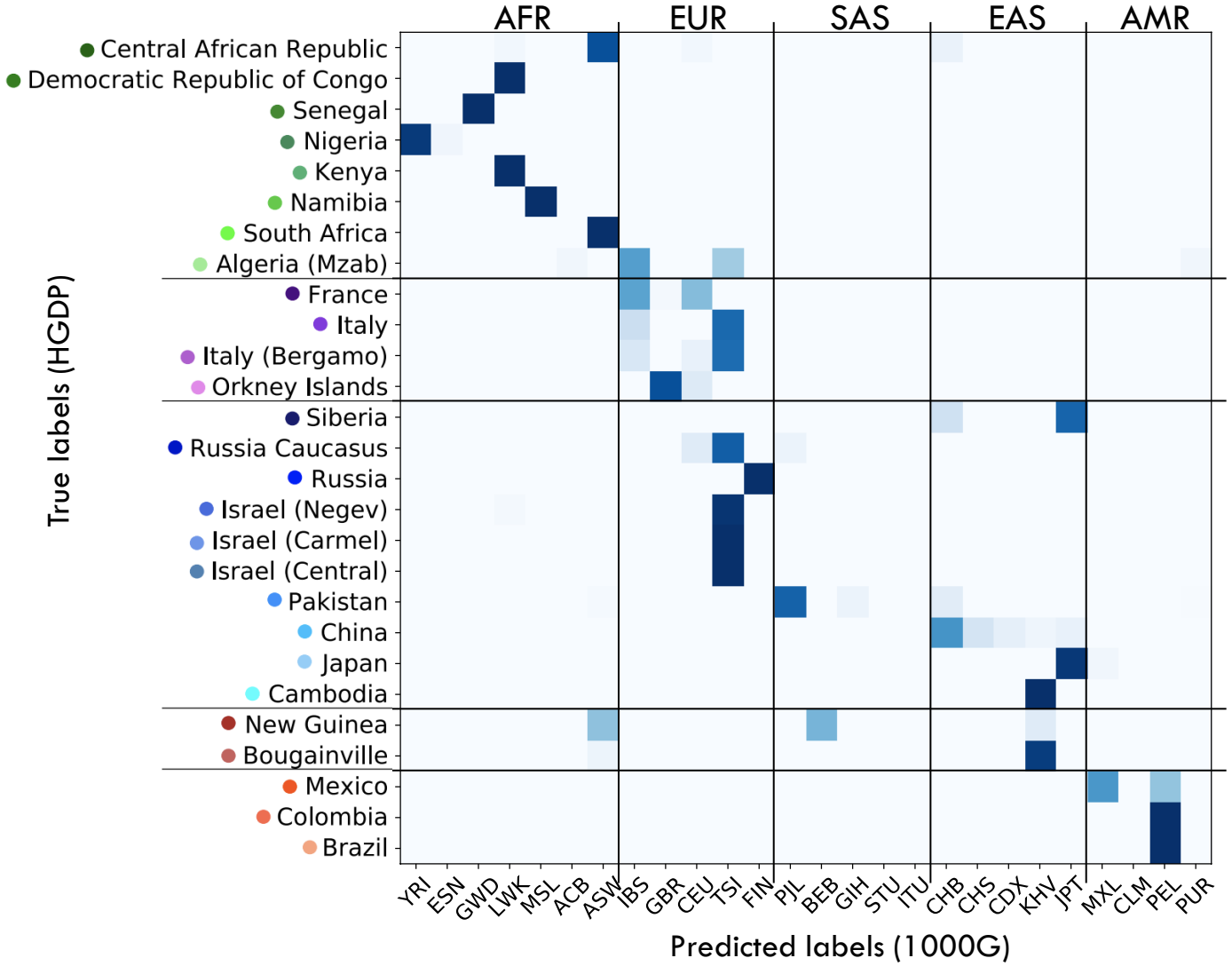
Result I : Generalization capability in HGDP

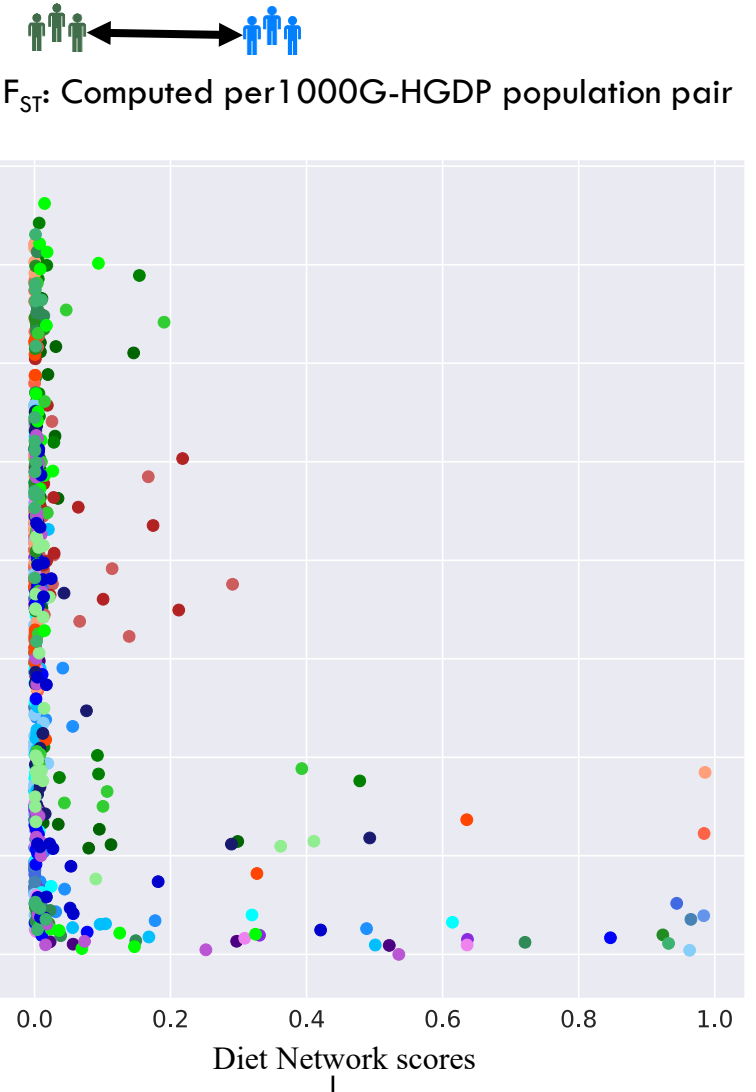
Rochefort-Boulanger et al. MLCB 2019



Result I : Generalization capability in HGDP

Rochefort-Boulanger et al. MLCB 2019





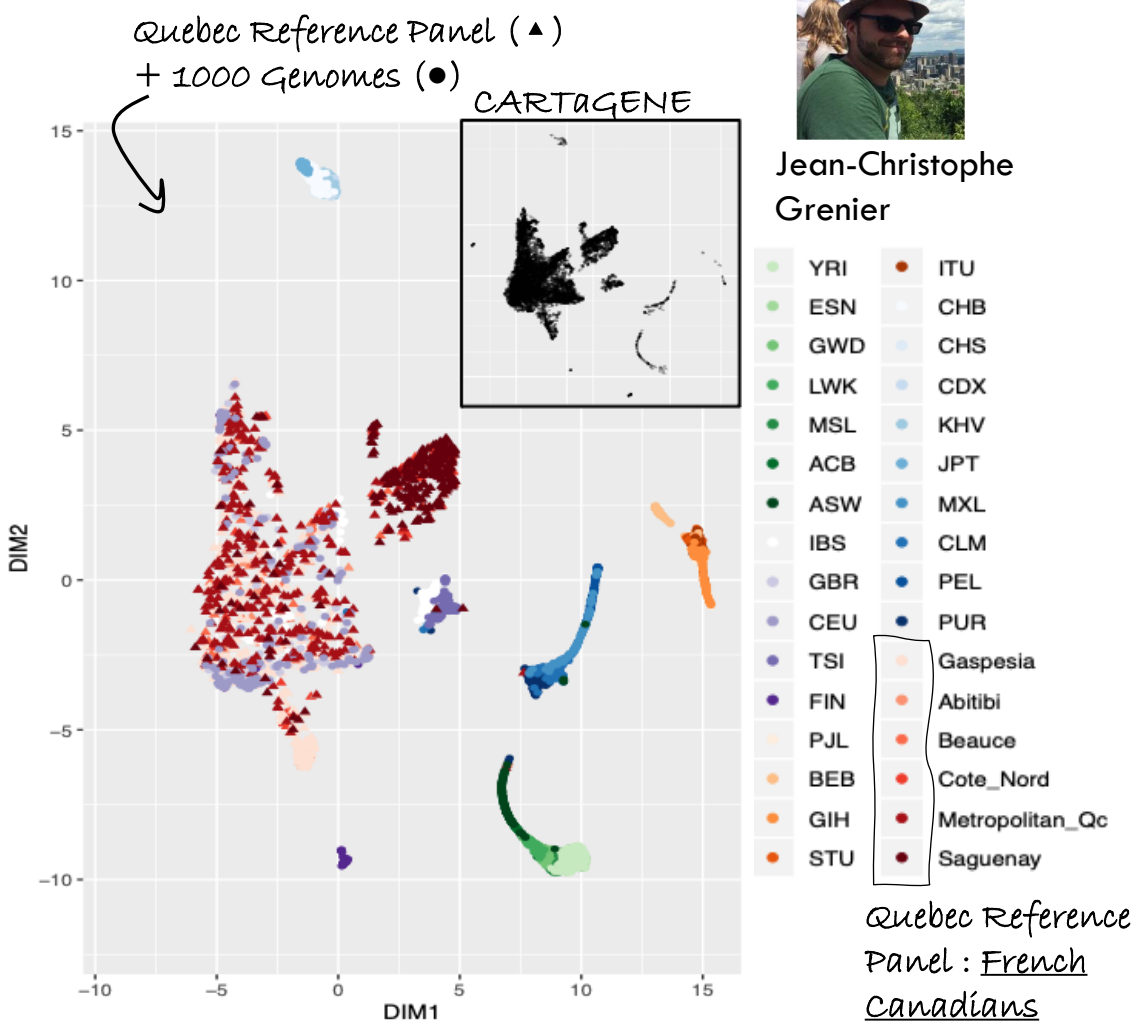
Softmax output averaged by HGDP population

$[0.01, 0.92, 0.003, \ldots, 0.008]$

$[0.91, 0.002, 0.05, \ldots, 0.001]$

Diet Network gives high scores to genetically similar populations

Result I : Generalization capability in CARTaGENE

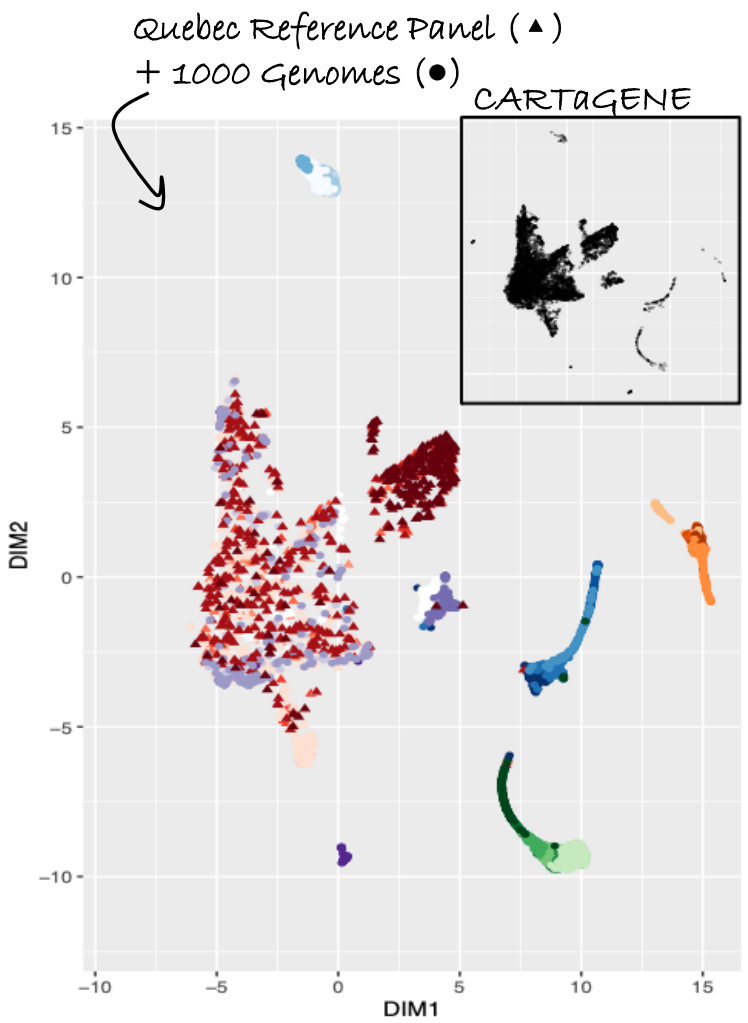


Quebec Reference Panel : French Canadians

Jean-Christophe Grenier

In CARTaGENE, most individuals are French Canadians (founder population from Europe)

Result I : Generalization capability in CARTaGENE



Jean-Christophe Grenier

YRI	ITU
ESN	CHB
GWD	CHS
LWK	CDX
MSL	KHV
ACB	JPT
ASW	MXL
IBS	CLM
GBR	PEL
CEU	PUR
TSI	Gaspesia
FIN	Abitibi
PJL	Beauce
BEB	Cote_Nord
GIH	Metropolitan_Qc
STU	Saguenay

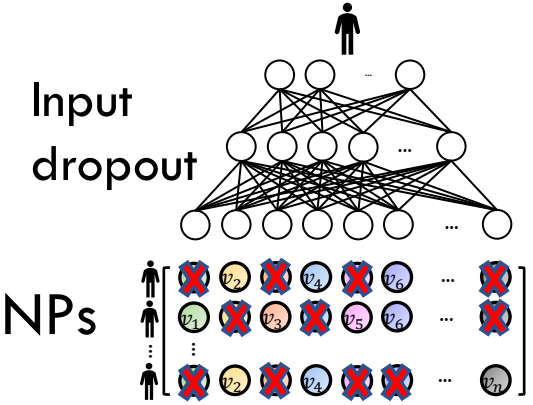
Quebec Reference Panel : French Canadians

In CARTaGENE, most individuals are French Canadians (founder population from Europe)

Diet Network

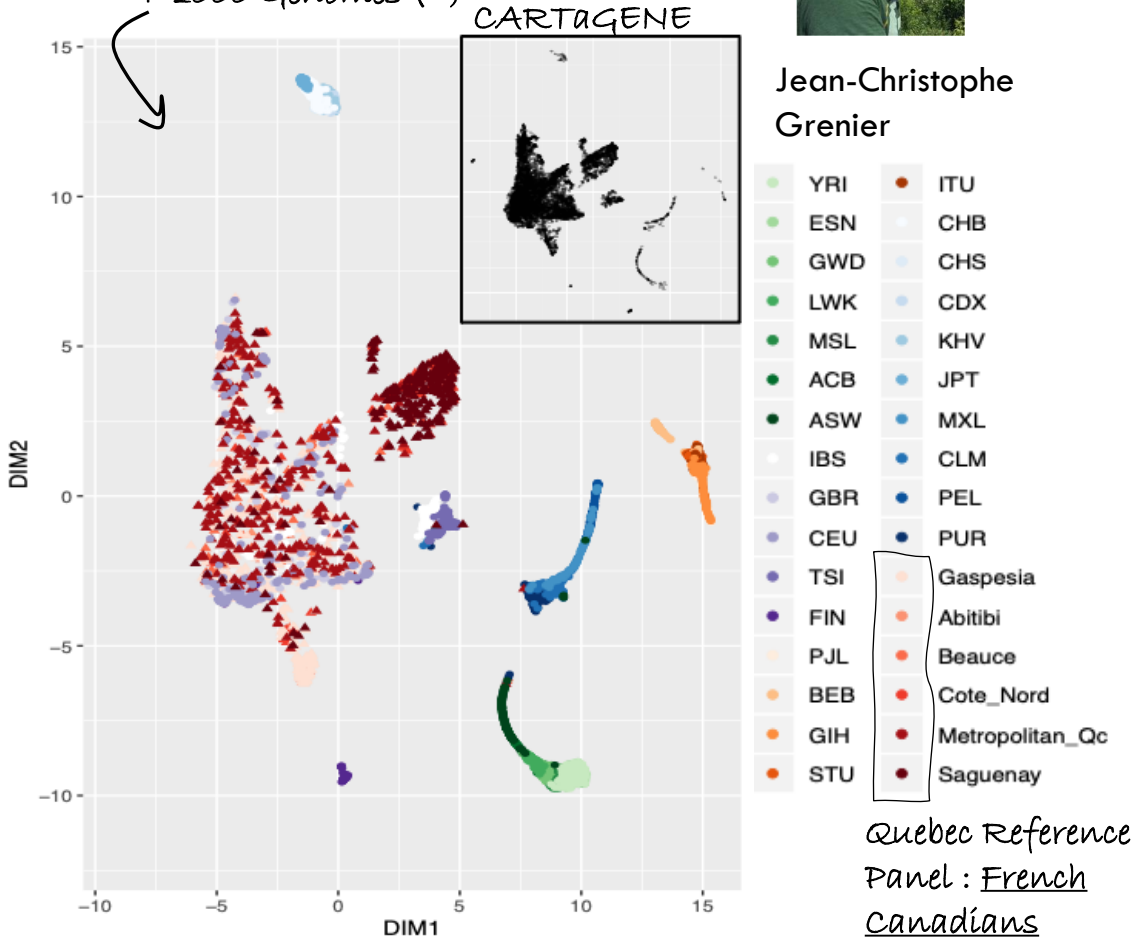
Train in 1000G : 300K SNPs

Test in CARTaGENE : 173K SNPs



Result I : Generalization capability in CARTaGENE

Jean-Christophe Grenier

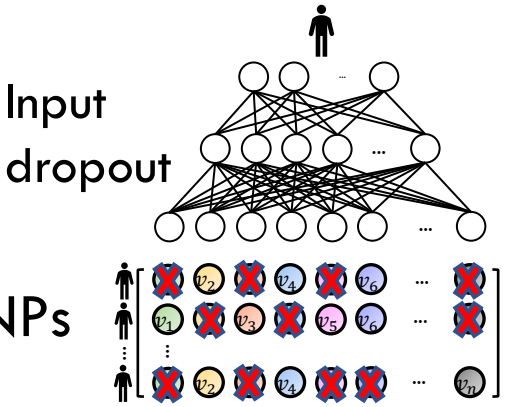


In CARTaGENE, most individuals are French Canadians (founder population from Europe)

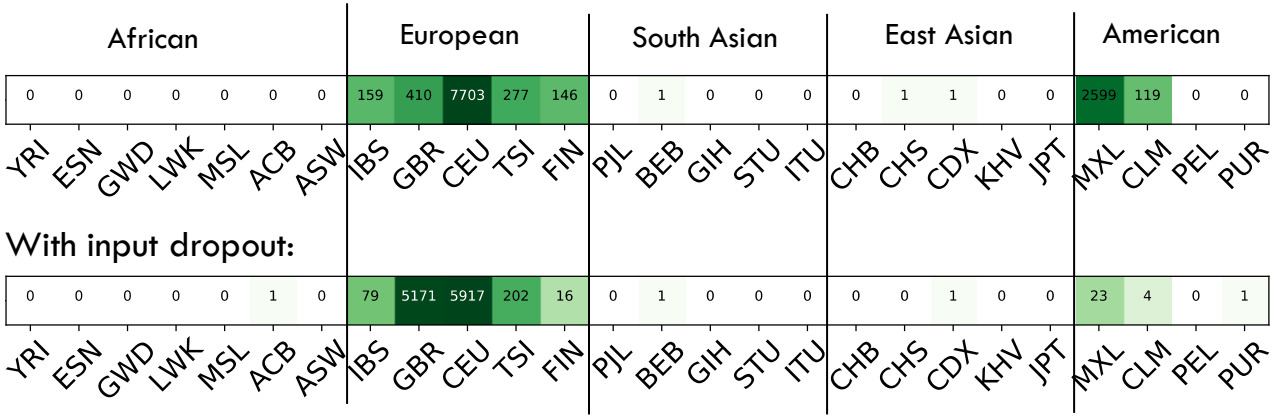
Diet Network

Train in 1000G : 300K SNPs

Test in CARTaGENE : 173K SNPs



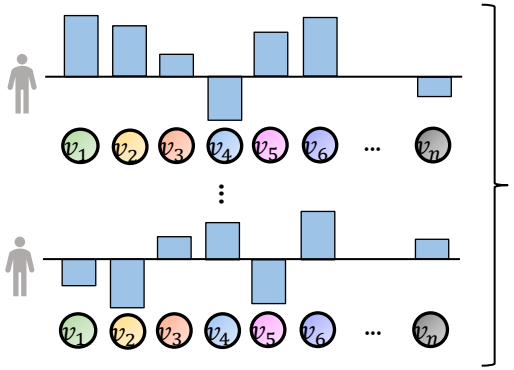
French Canadians classification :



Diet Network predictions are generalizable to a new population never seen in training

Result II : Interpretability

Which SNPs are important in the Diet Network predictions

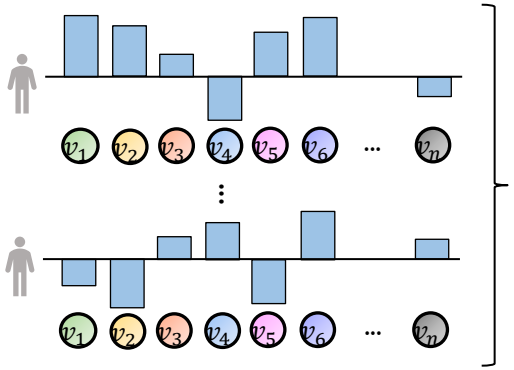


Attribution scores computed with
Integrated Gradients (Sundararajan et al. 2017)

- indicates how useful a feature is
- each sample may have different scores

Result II : Interpretability

Which SNPs are important in the Diet Network predictions

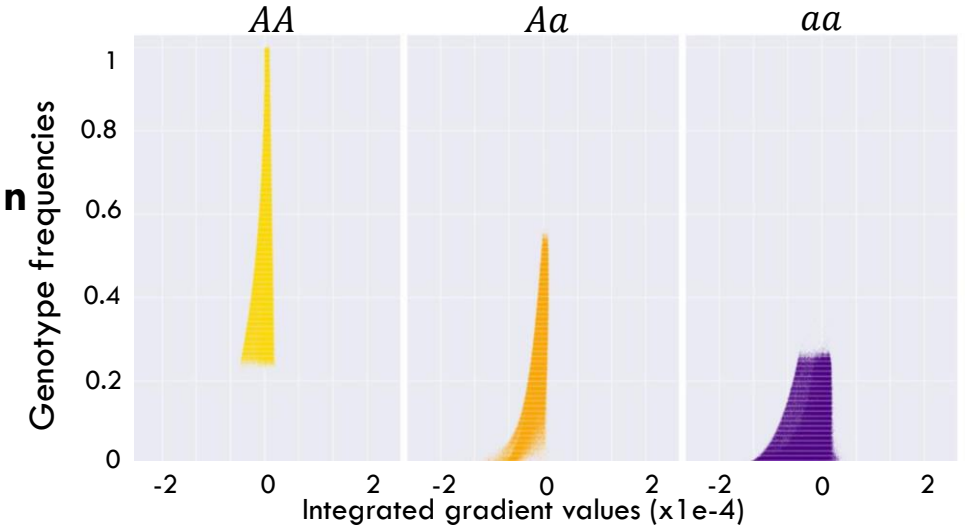


Attribution scores computed with **Integrated Gradients** (Sundararajan et al. 2017)

- indicates how useful a feature is
- each sample may have different scores

Integrated Gradients with binary classification (European and African)

Léo Choinière

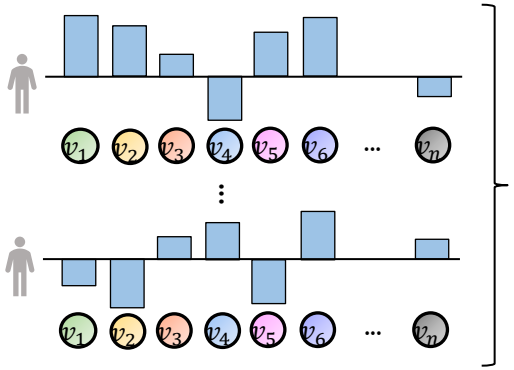


Integrated Gradients show that low frequency SNPs are important in Diet Network's predictions

This is opposite to genetic population methods that use common SNPs to compare populations

Result II : Interpretability

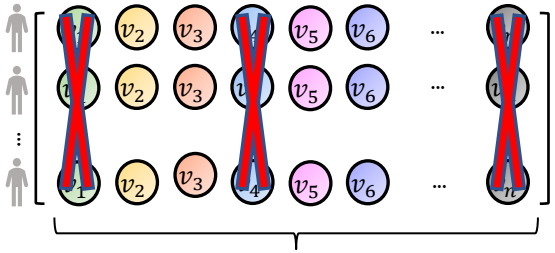
Which SNPs are important in the Diet Network predictions



Attribution scores computed with **Integrated Gradients** (Sundararajan et al. 2017)

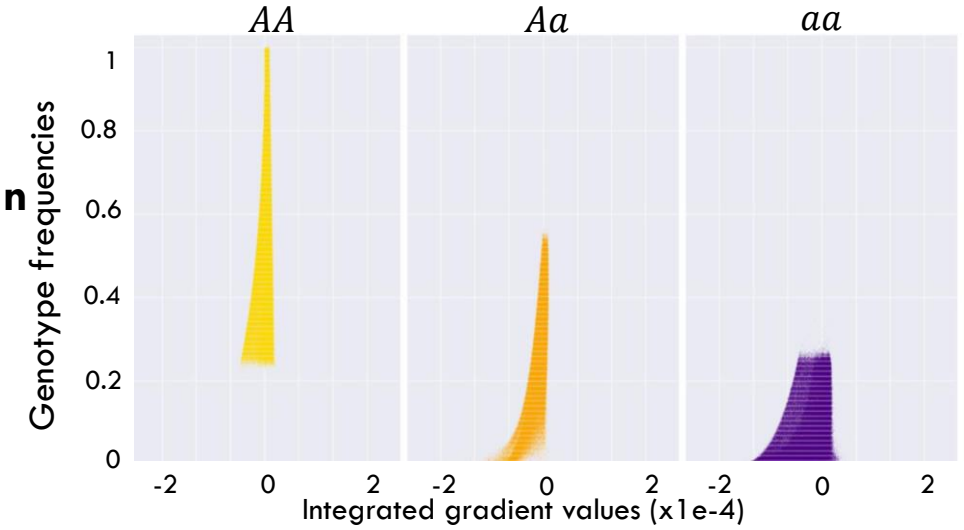
- indicates how useful a feature is
- each sample may have different scores

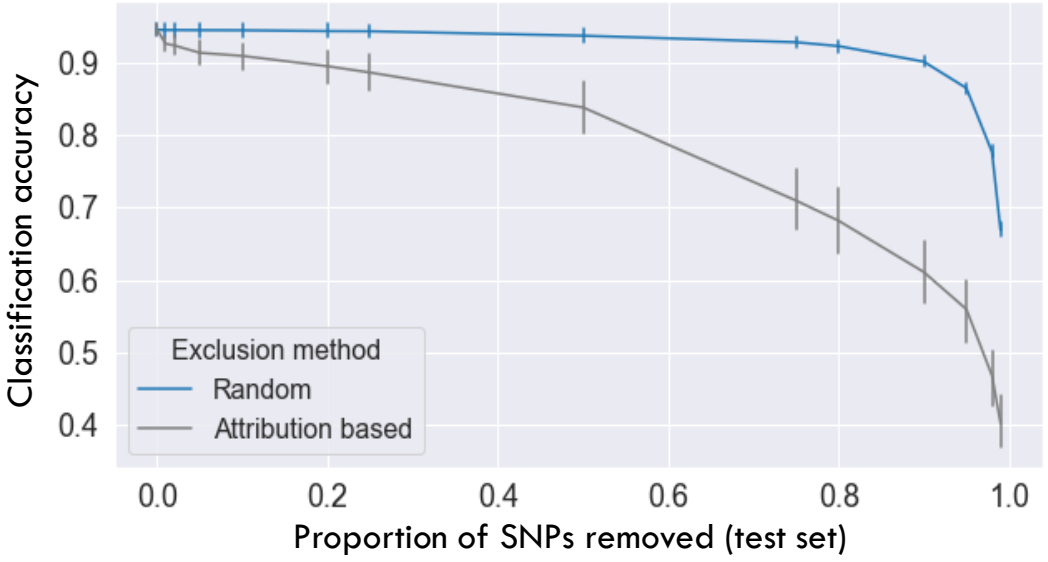
Matthew Scicluna



Integrated Gradients with binary classification (European and African)

Léo Choinière





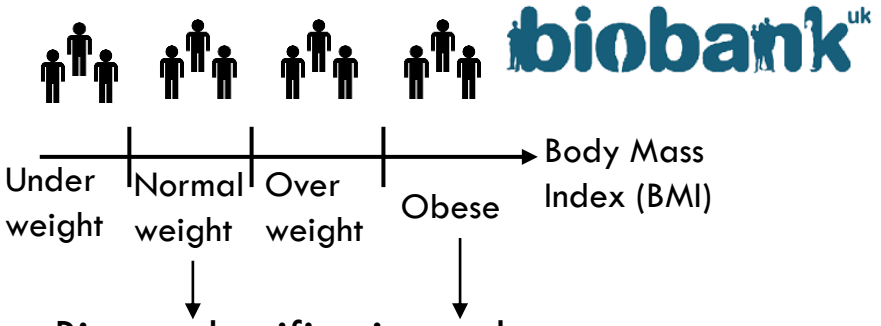
Integrated Gradients show that low frequency SNPs are important in Diet Network's predictions

This is opposite to genetic population methods that use common SNPs to compare populations

Result III : Complex phenotype prediction

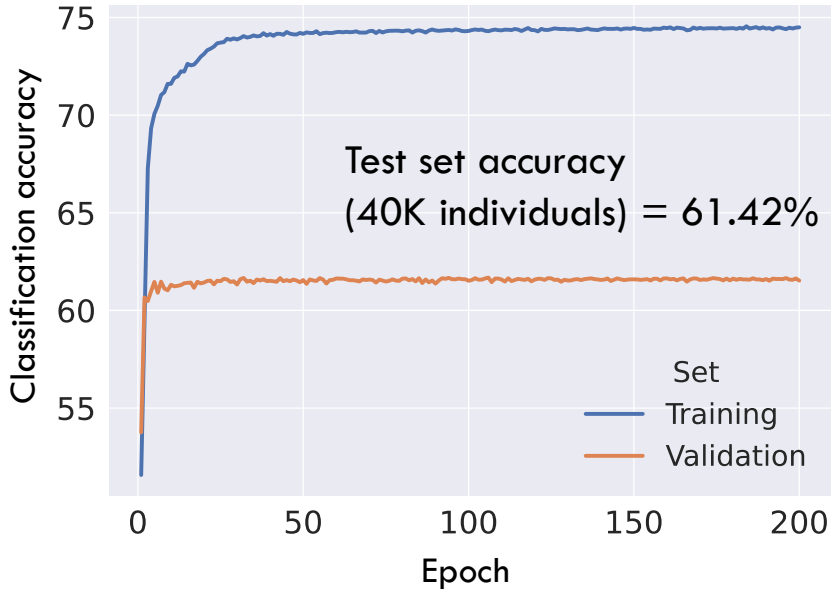
What about real complex phenotypes

Obesity prediction in the UK biobank



Binary classification task

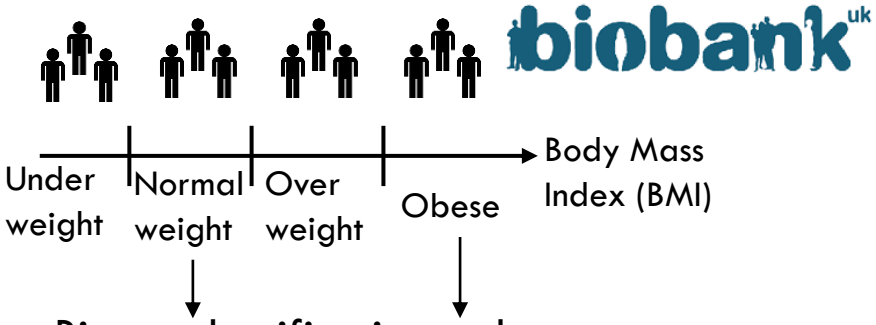
~197K White British and ~408K SNPs



Result III : Complex phenotype prediction

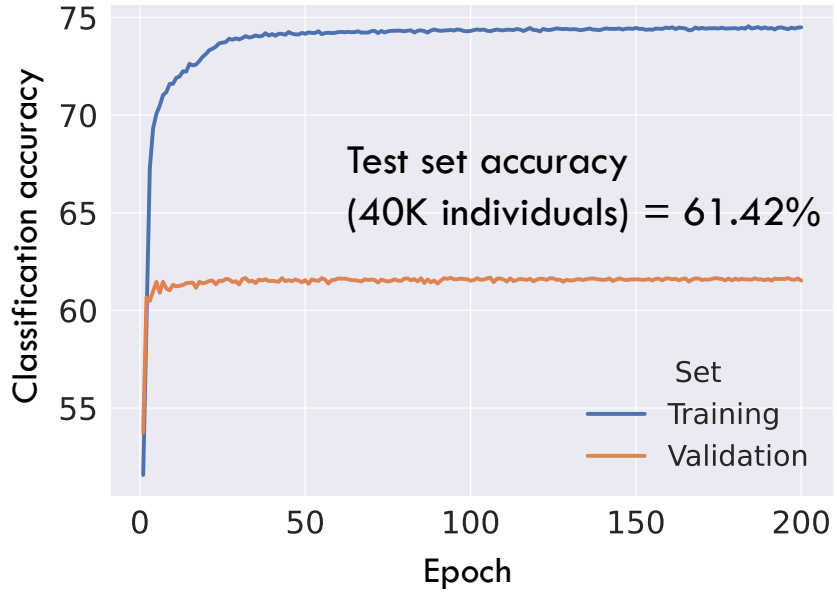
What about real complex phenotypes

Obesity prediction in the UK biobank



Binary classification task

~197K White British and ~408K SNPs

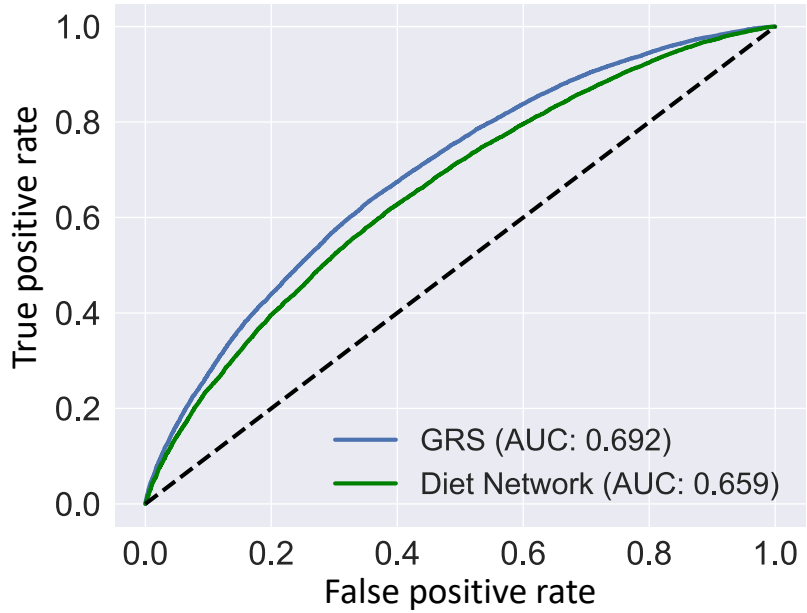


Comparison with a Genetic Risk Score

Khera et al. Cell. 2019

SNPs effects obtained from a previous GWAS (Locke et al. Nature 2015)

GRS created using ~120K UK biobank White British participants (which ones?) and 2.1M SNPs



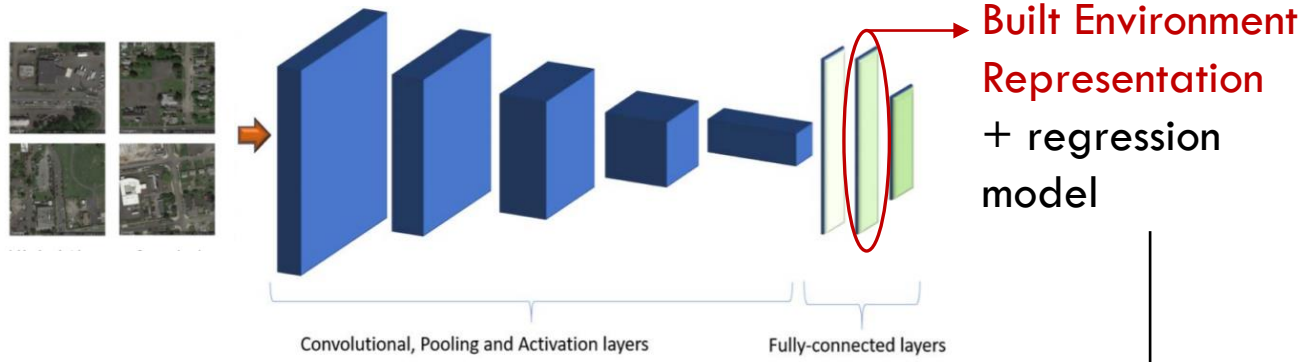
Obesity prediction with Diet Networks in the UK biobank yields similar results to a Genetic Risk Score

Result IV : Environmental factors

How to take into account environmental factors

- Clinical and lifestyle variables available in biobanks
- Information from the built environment

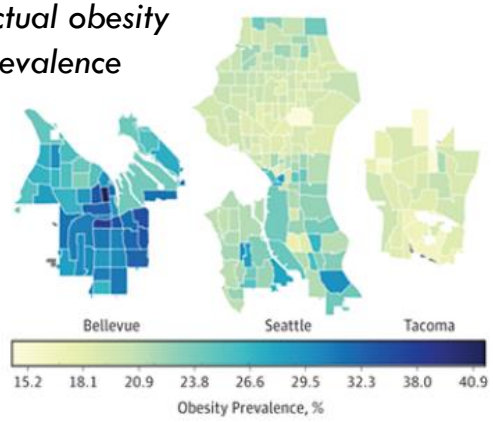
Use of Deep Learning to Examine the Association of the Built Environment With Prevalence of Neighborhood Adult Obesity
Maharana et al. 2018



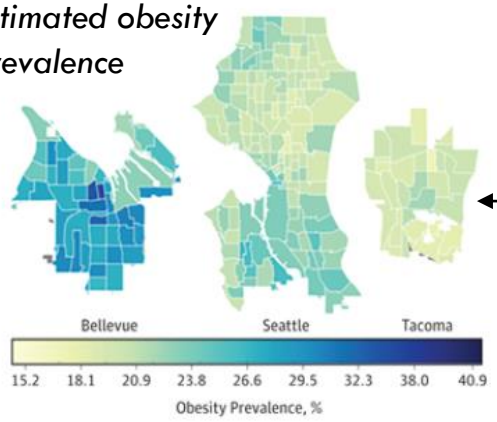
Built Environment Representation + regression model

Seattle, Washington

Actual obesity prevalence



Estimated obesity prevalence

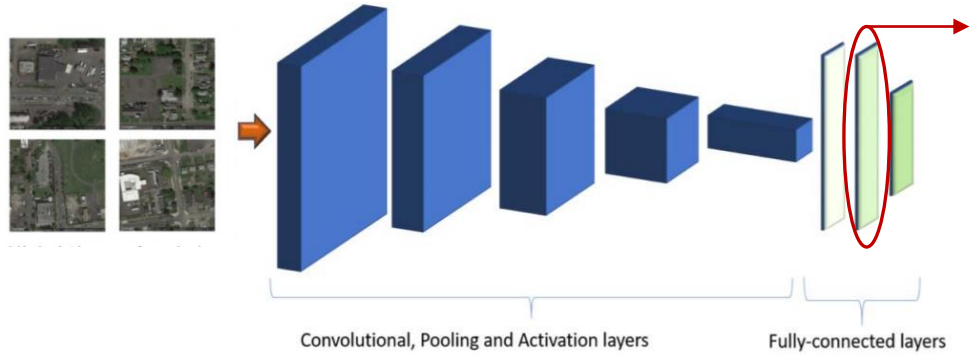


Result IV : Environmental factors

How to take into account environmental factors

- Clinical and lifestyle variables available in biobanks
- Information from the built environment

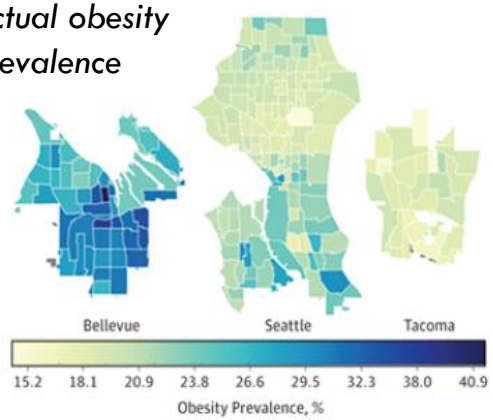
Use of Deep Learning to Examine the Association of the Built Environment With Prevalence of Neighborhood Adult Obesity
Maharana et al. 2018



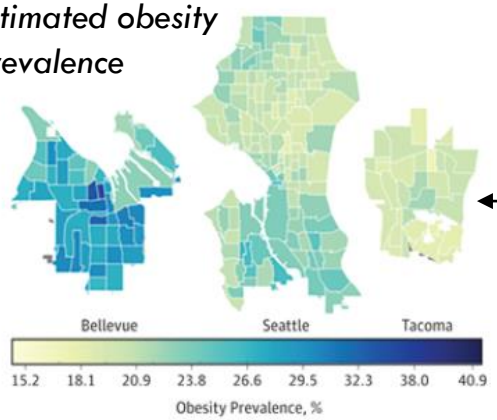
Built Environment Representation + regression model

Seattle, Washington

Actual obesity prevalence

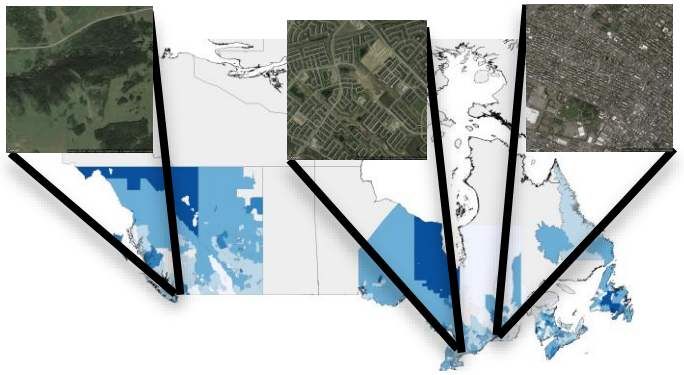


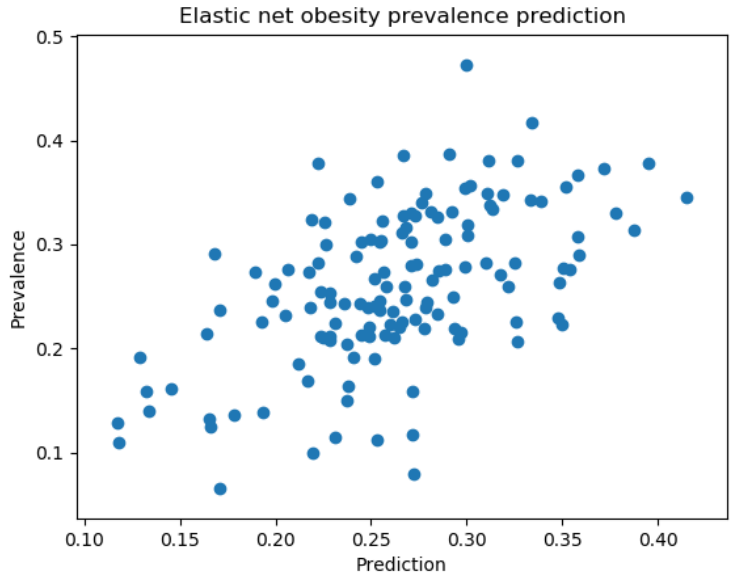
Estimated obesity prevalence



Canadian health cohort
- Body mass index of participants
- Forward Sortation Area (FAS) : 3 first digits of postal codes

Marie-Julie Favé





OHS (participants in Ontario) :
~30% of obesity variation prevalence predicted by regression

Future directions for the Diet Network

Prediction of complex phenotypes

- Regression
- Height (higher heritability)

Multi tasks learning of several complex phenotypes

Information given in SNPs embedding (Auxiliary network input)

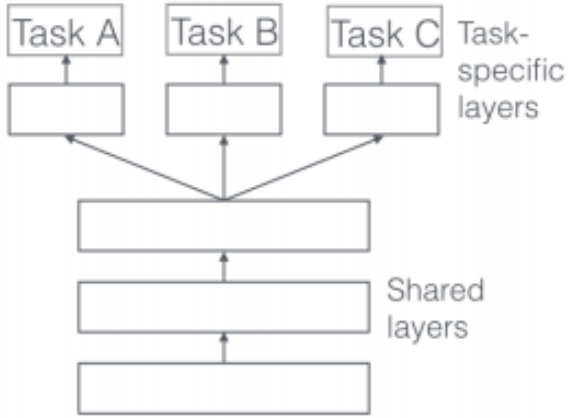
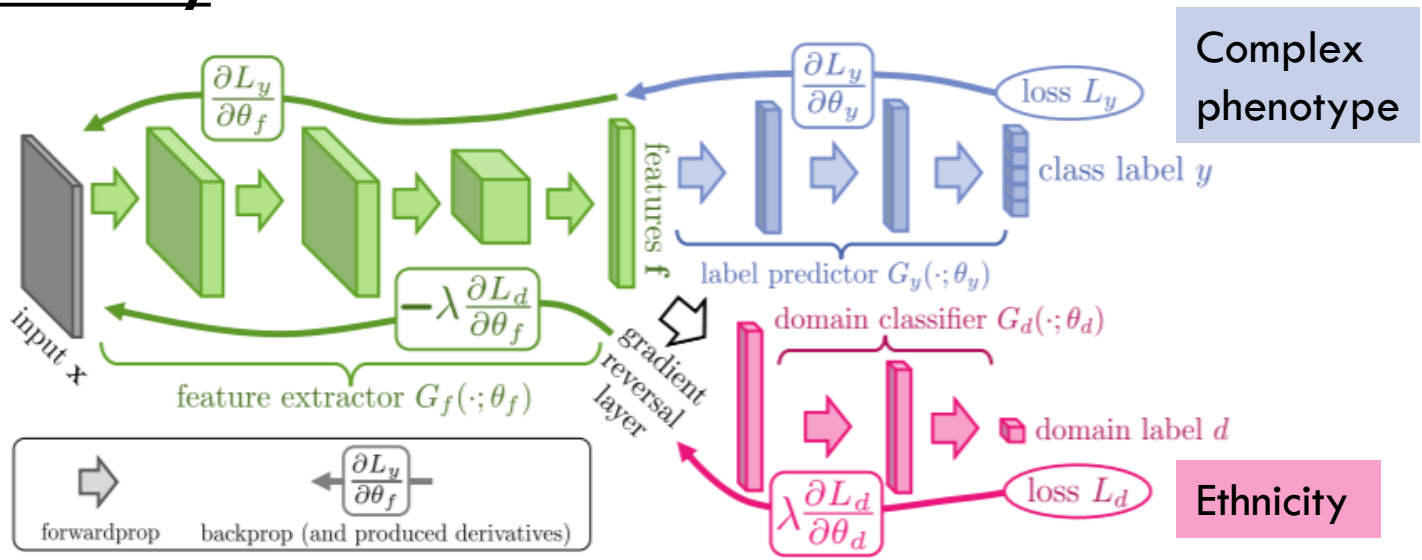


Image : Ruder, S. (2017)

Taking into account genetic ancestry diversity

Domain-adversarial neural network to Penalize the use of SNPs that have a large difference in alleles frequencies between population

Model portability across populations



Ganin et al. (2016)

THANK YOU!

MHI-OMICS

- Julie Hussin
- Matthew Scicluna
- Léo Choinière
- Jean-Christophe Grenier
- Marie-Julie Favé
- Holly Trochet
- Vanessa Bellegarde

MILA

- Yoshua Bengio
- Pierre-Luc Carrier