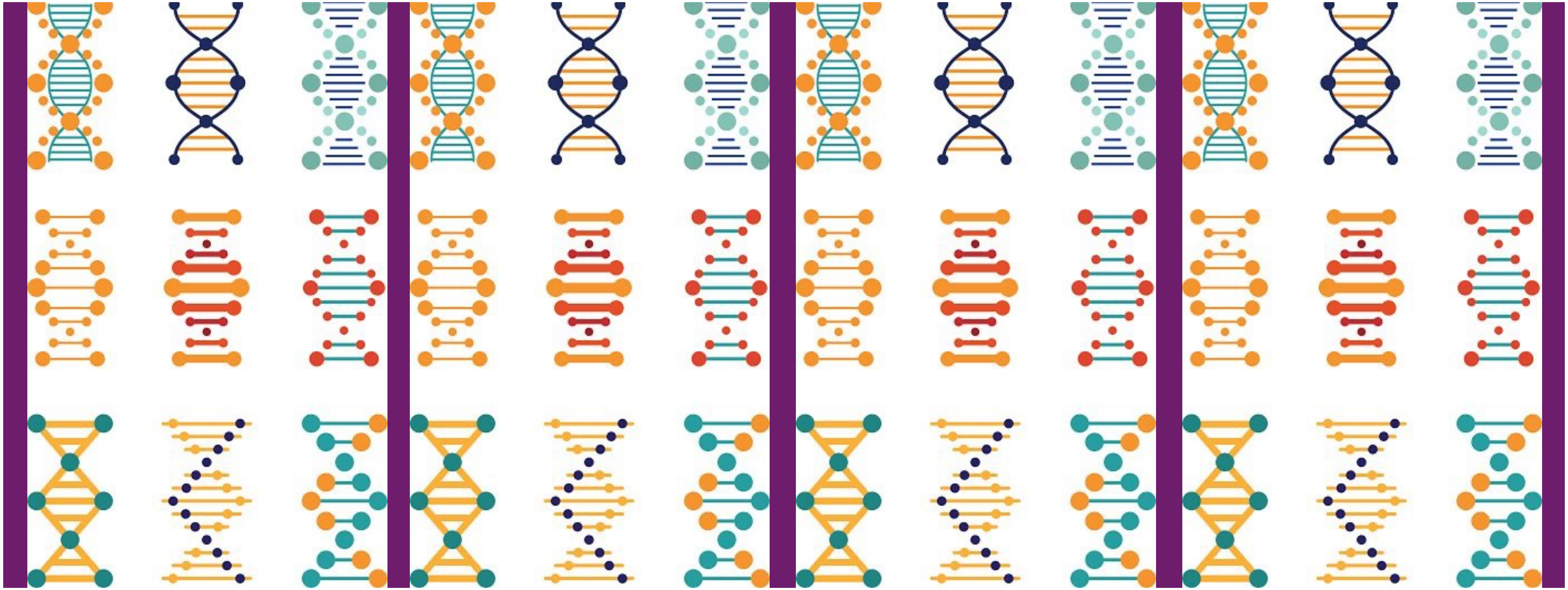


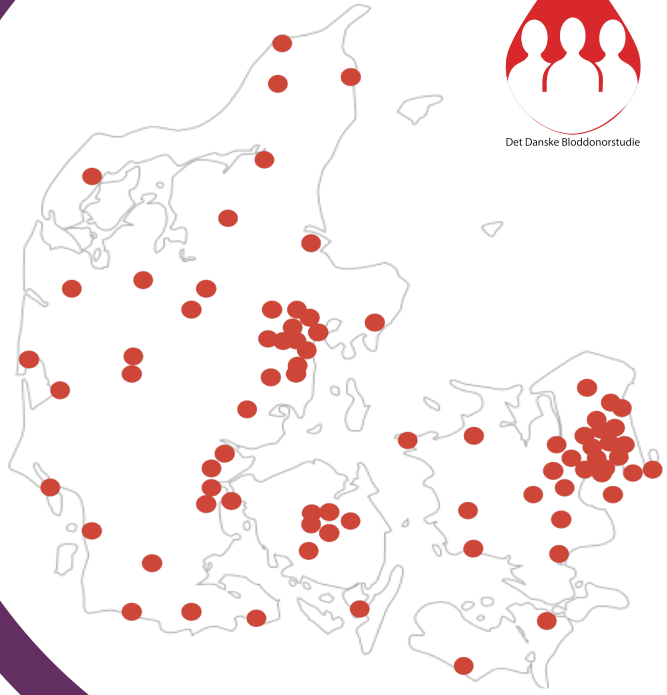
APPLYING ML IN LARGE-SCALE COMMON COMPLEX GENETICS

Karina Banasik
Associate Professor
Novo Nordisk Foundation Center for Protein Research,
University of Copenhagen

THE BIGGEST BIOBANKS WITH GENETICS IN DENMARK

THE DANISH BLOOD DONOR STUDY

- ✓ Nationwide recruitment of blood donors via Danish Blood Banks
- ✓ 206 donation points across Denmark
- ✓ High participation rate: 95%
- ✓ Builds on existing infrastructure
- ✓ 135K blood donors recruited since 2010
- ✓ Donors donate up to four times per year (10 times if plasma donors)
- ✓ EDTA plasma stored in biobank at each donation
- ✓ Data collected on a dedicated secure cloud



Det Danske Bloddonorstudie

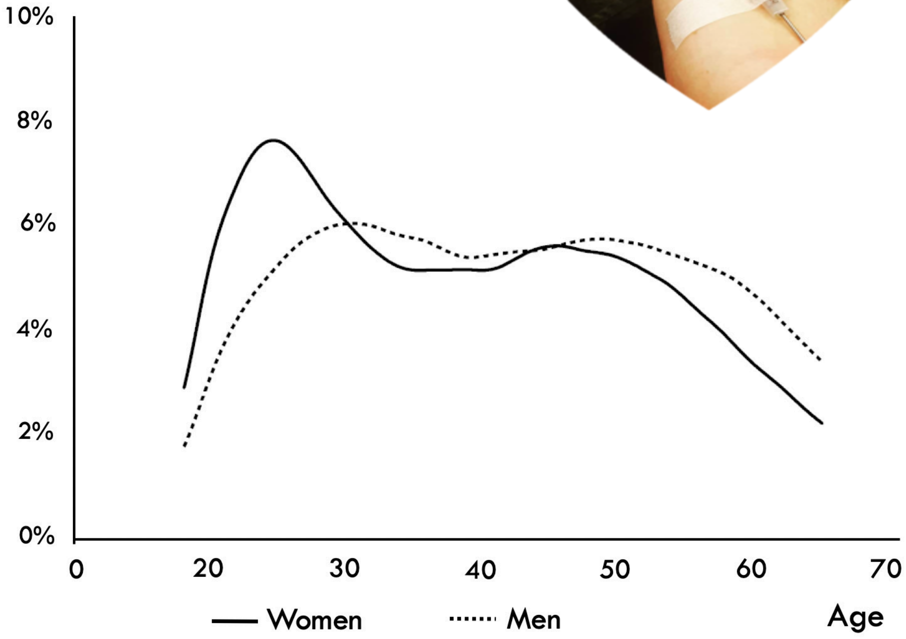
Novo Nordisk Foundation
Center for Protein Research

WHO DONATES BLOOD?

✓ Blood donation is voluntary and unpaid

✓ Inclusion criteria

  - ✓ Physically well
  - ✓ Between 18 and 67 years old
  - ✓ Weigh more than 50 kilos
  - ✓ Speak Danish, have a Danish social security number, and have lived in Denmark for minimum one year

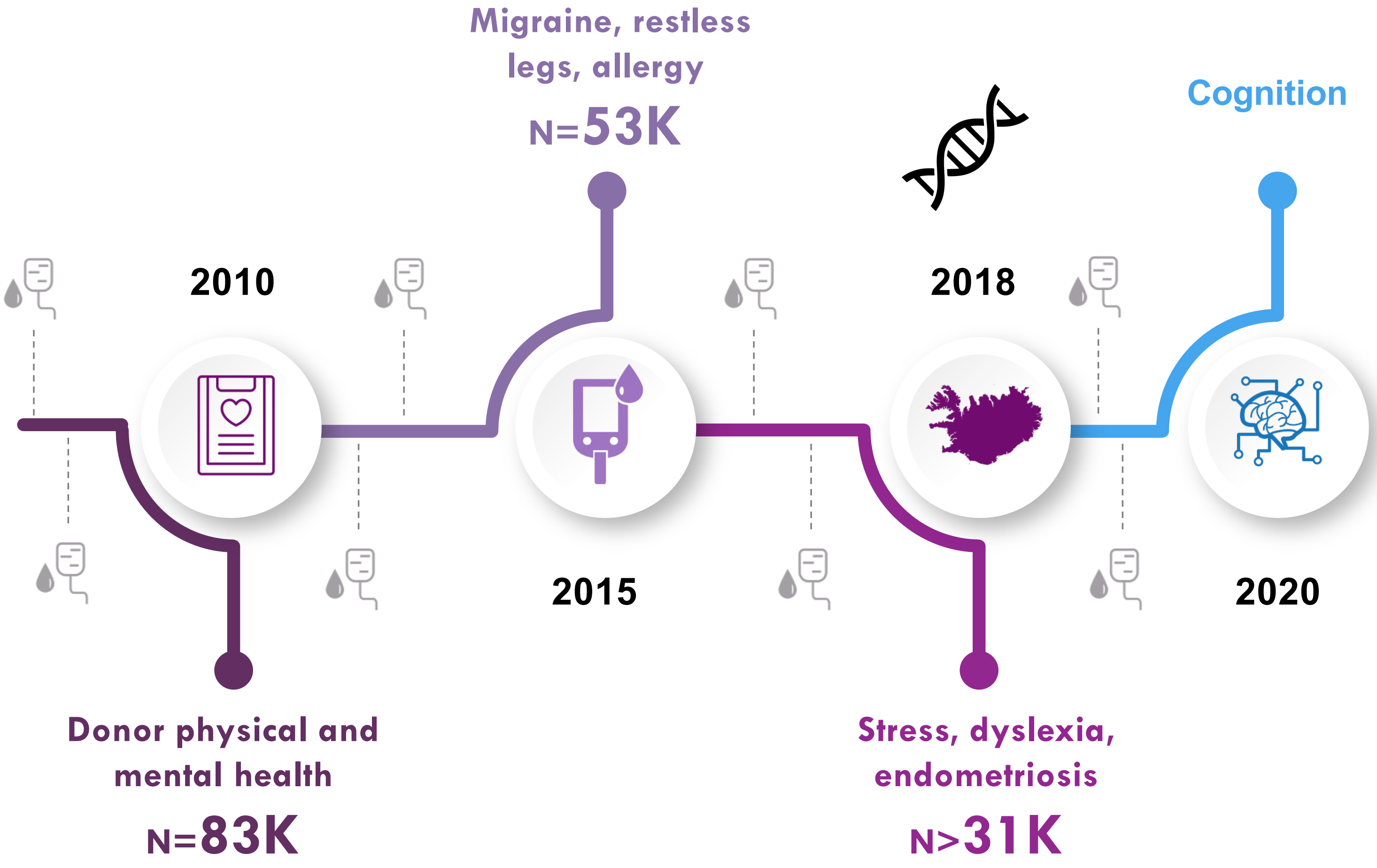


[Burgdorf *et al.* 2017]

Novo Nordisk Foundation
Center for Protein Research

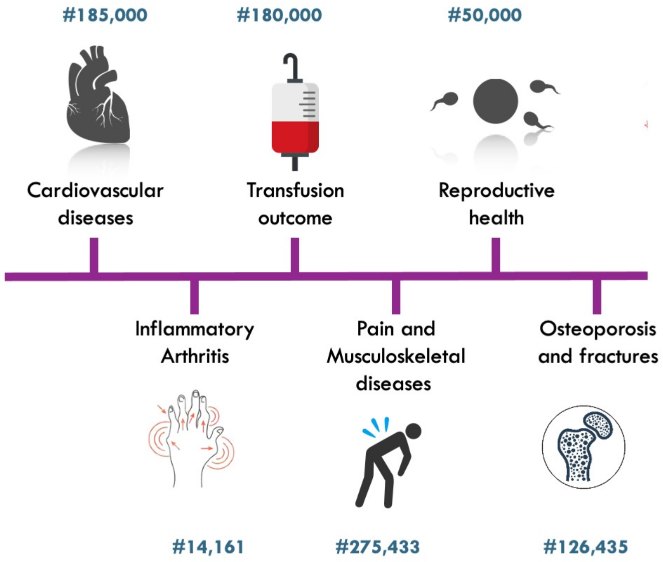
A RICH DATA COLLECTION

- ✓ Illumina GSA chip on 110K in collaboration with deCODE genetics
- ✓ Three questionnaires (basic info: weight, height, smoking etc in all three)
- ✓ 270K unique donations with Sysmex haematology measures
- ✓ >500K EDTA plasma samples in the biobank


Migraine, restless legs, allergy
N=53K
Cognition
2010
2018
2015
2020
Donor physical and mental health
N=83K
Stress, dyslexia, endometriosis
N>31K

COPENHAGEN HOSPITAL BIOBANK

- ✓ Biobank with >430K patients
- ✓ Initiated at Rigshospitalet in Copenhagen in 2009
- ✓ Collection of surplus of EDTA whole blood from the Clinical Immunology department
- ✓ Average age of patients at inclusion >45 years
- ✓ Linkable with Danish national registries and electronic health records
- ✓ Illumina GSA chip in collaboration with deCODE genetics



#185,000
Cardiovascular diseases

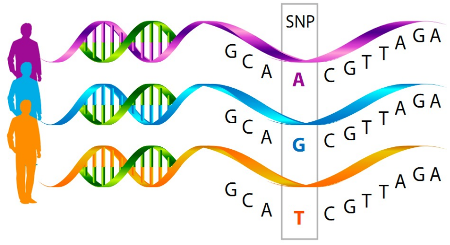
#180,000
Transfusion outcome

#50,000
Reproductive health

Inflammatory Arthritis
#14,161

Pain and Musculoskeletal diseases
#275,433

Osteoporosis and fractures
#126,435



REGION H

Novo Nordisk Foundation
Center for Protein Research

DATA INFRASTRUCTURE ON COMPUTEROME 2.0

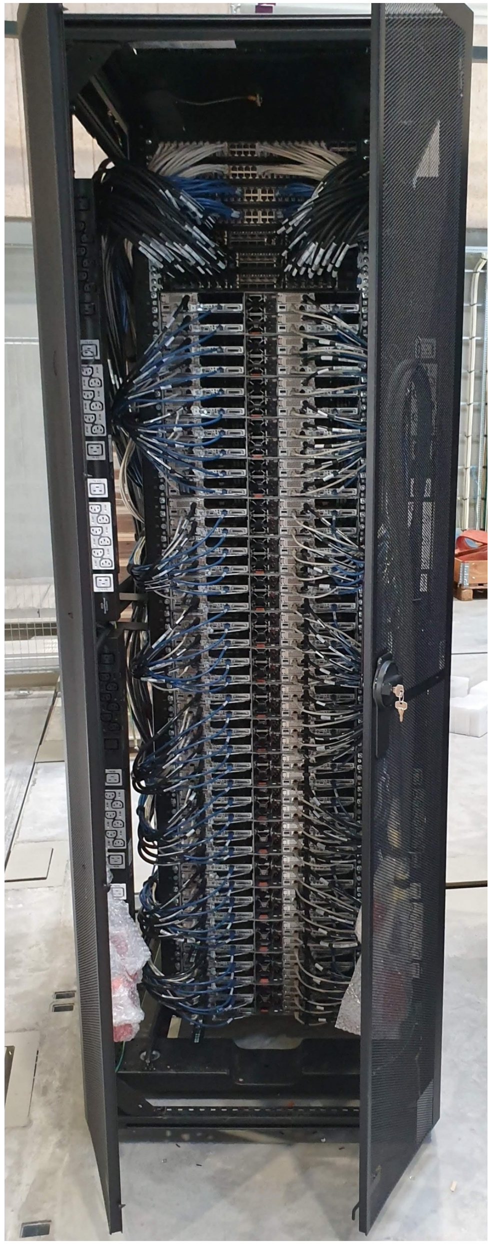
The system is designed to process large amounts of healthcare data with the highest I/O performance using state of the art storage and memory technology

- ✓ 90% of the system is cooled using warm water liquid cooling
- ✓ Heat is re-used
- ✓ Wind power

30K
CPUs

30-40
GPU nodes for machine learning

55
"Fat" nodes with 1.5TB of RAM



GENOME-WIDE ASSOCIATION (GWAS) ANALYSES IN PRECISION MEDICINE

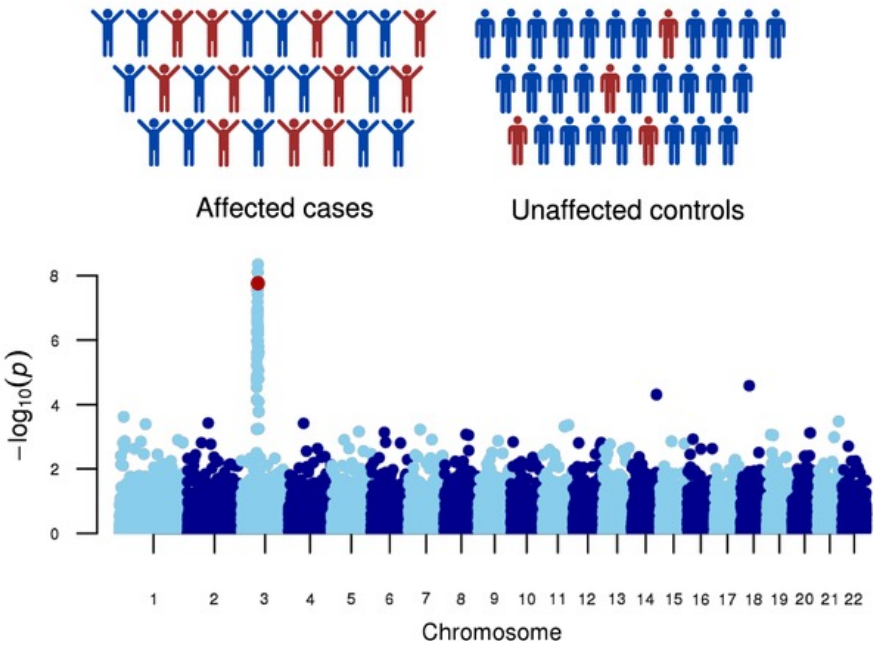
…GATG**C**TTGG…

…GATG**T**TTGG…

An experimental design to discover new genetic variants associated with diseases and traits

In GWAS, allele or genotype frequencies are compared between individuals with a disease (cases) and unaffected individuals (controls)

Since the genetics of complex diseases are driven by a combination of variants with moderate effect sizes, large sample sizes are extremely important!



ADVANTAGES OF BIOBANK-DRIVEN LARGE-SCALE COMMON COMPLEX GENETICS

✓ GWAS can help identify genes that contribute to disease or variation in traits that in turn can be used to develop better prevention and treatment strategies

✓ The biobanks are not unbiased populations, but serves well for making inference about the effects of genetics on disease

✓ The large collection of individuals in the biobanks together with linkable life-course registry data and electronic health records holds great promise in precision medicine in Denmark

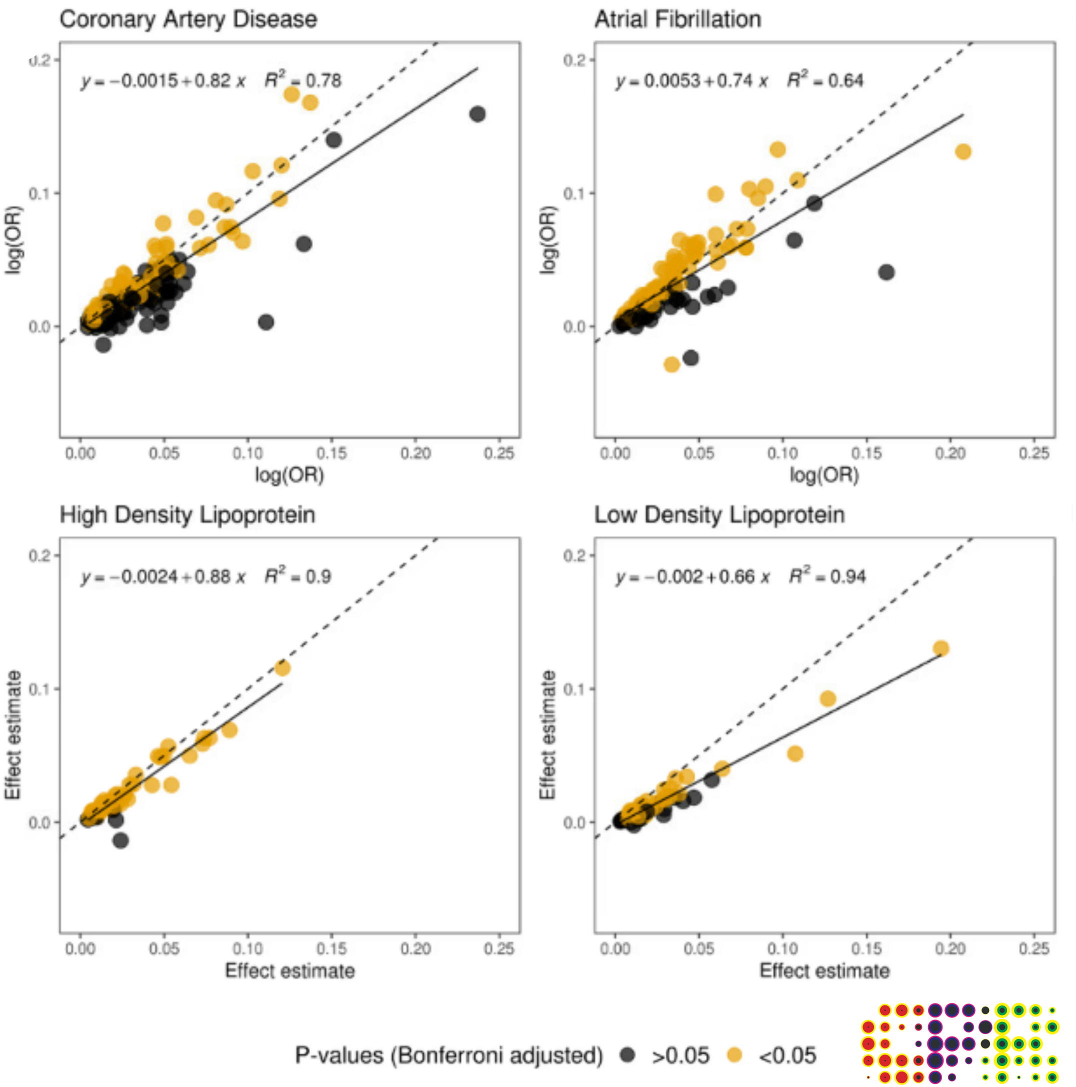
Novo Nordisk Foundation
Center for Protein Research

WHAT HAVE WE FOUND SO FAR?

PERFORMANCE MEASURES

Comparison of GWAS effect sizes (weighted by risk allele frequencies) between our study population and a reference population (usually the largest published European GWAS).

Our results are comparable to results from other large genetic cohorts!



Novo Nordisk Foundation
Center for Protein Research

[Laursen et al. in review]

GWAS META-ANALYSIS OF IRON HOMEOSTASIS

Det Danske Bloddonorstudie

Iron is essential for many biological functions and iron deficiency and overload have major health implications – *what makes a great blood donor?*

Sample size is important!

Iceland
(deCODE genetics)

~ 155K samples, ~32 million variants

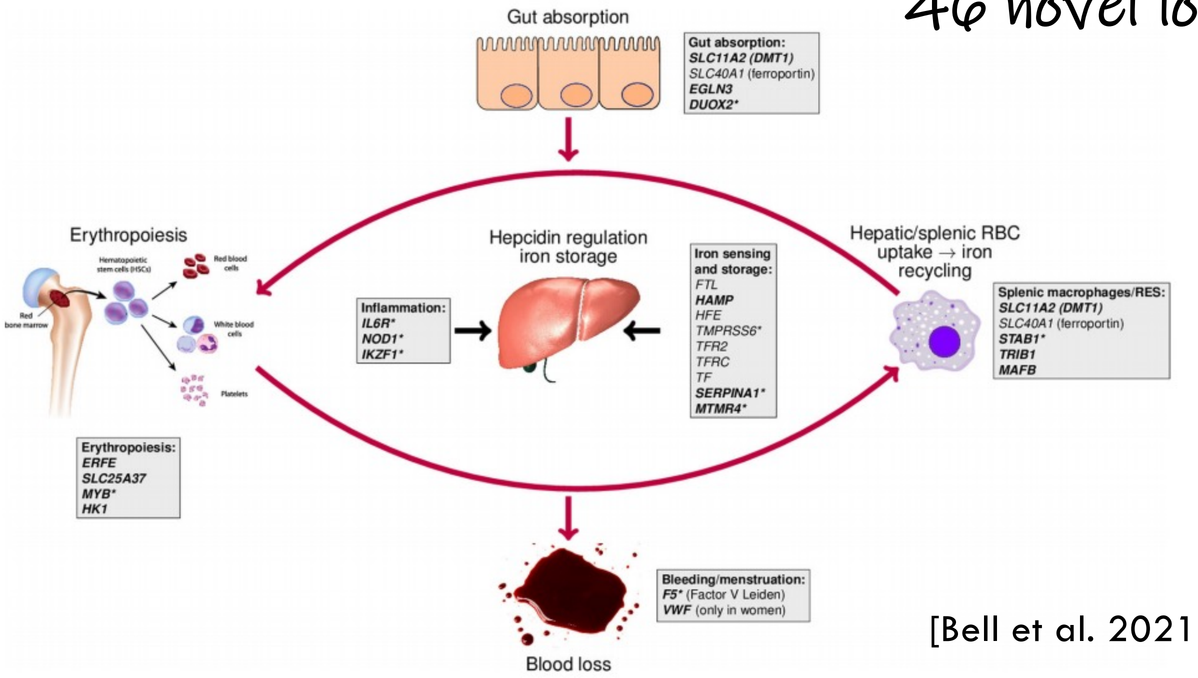
UK
(INTERVAL study)

~ 43K samples, ~19 million variants

Denmark
(Danish Blood Donor Study)

~ 84K samples, ~19 million variants

46 novel loci



Gut absorption:
SLC11A2 (DMT1)
SLC40A1 (ferroportin)
EGLN3
*DUOX2**

Erythropoiesis:
ERFE
SLC25A37
*MYB**
HK1

Inflammation:
*IL6R**
*NOD1**
*IKZF1**

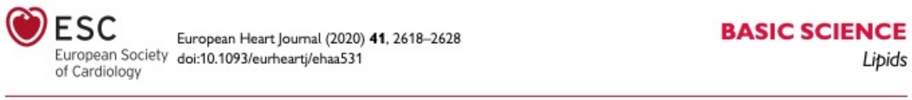
Iron sensing and storage:
FTL
HAMP
HFE
*TMPRSS6**
TFR2
TFRC
TF
*SERPINA1**
*MTMR4**

Splenic macrophages/RES:
SLC11A2 (DMT1)
SLC40A1 (ferroportin)
*STAB1**
TRIB1
MAFB

Bleeding/menstruation:
*F5** (Factor V Leiden)
VWF (only in women)

[Bell et al. 2021]

GENETIC VARIABILITY IN THE ABSORPTION OF DIETARY STEROLS AFFECTS THE RISK OF CORONARY ARTERY DISEASE

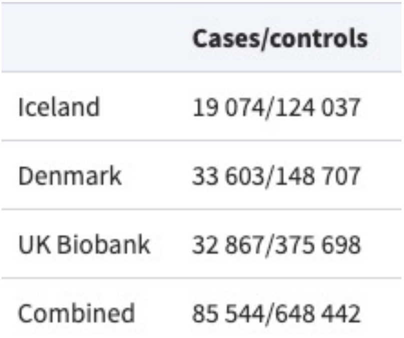

ESC
European Society of Cardiology

European Heart Journal (2020) **41**, 2618–2628
doi:10.1093/eurheartj/ehaa531

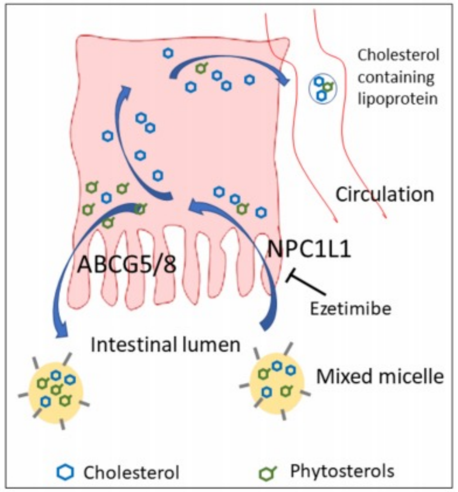
BASIC SCIENCE
Lipids

Genetic variants that decrease function of the ABCG5/8 transporter ***increase uptake*** of both dietary cholesterol and phytosterols into the circulation

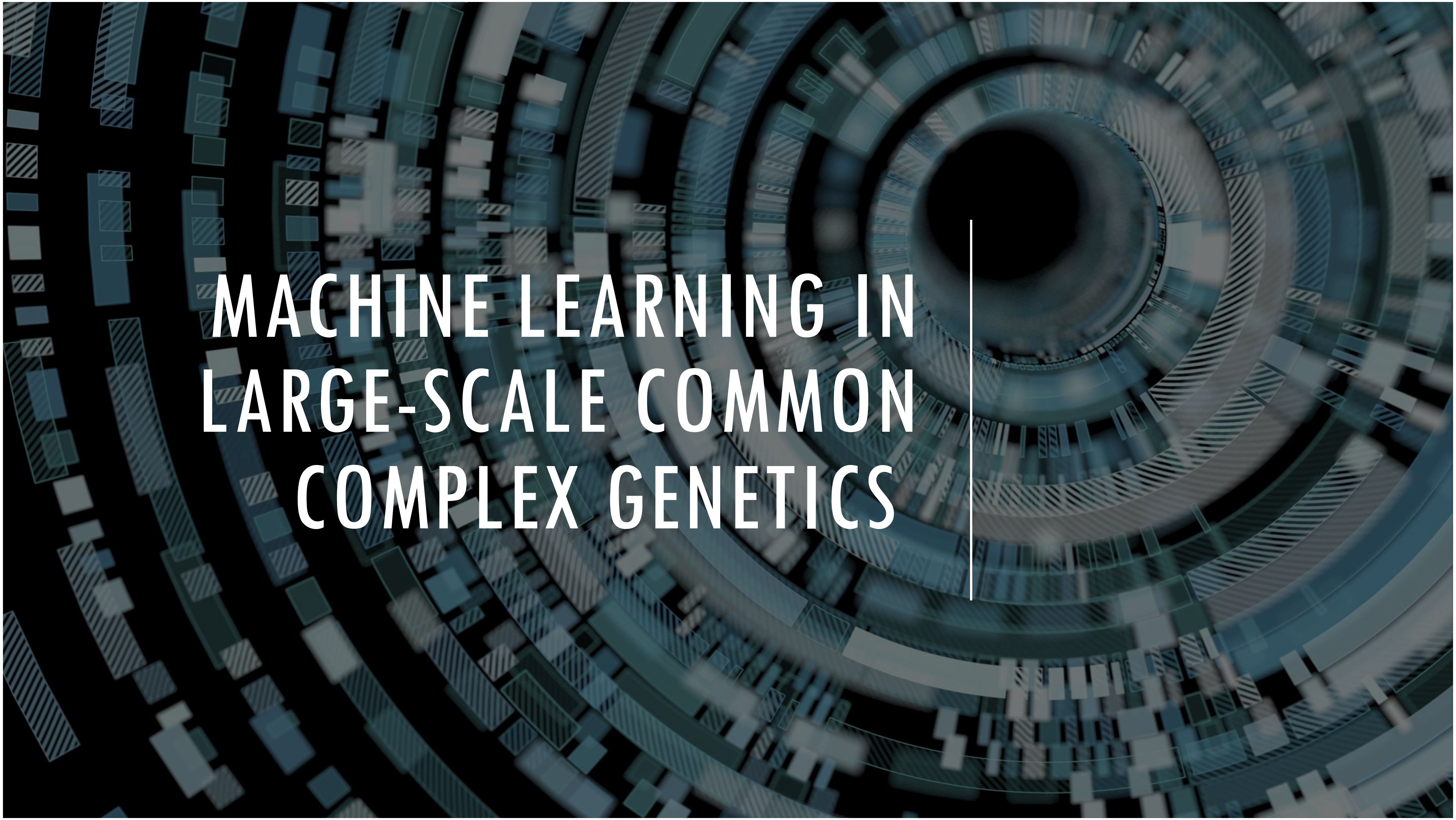
→ increase the risk of coronary artery disease.



	Cases/controls
Iceland	19 074/124 037
Denmark	33 603/148 707
UK Biobank	32 867/375 698
Combined	85 544/648 442



[Helgadottir et al. 2020]



MACHINE LEARNING IN LARGE-SCALE COMMON COMPLEX GENETICS

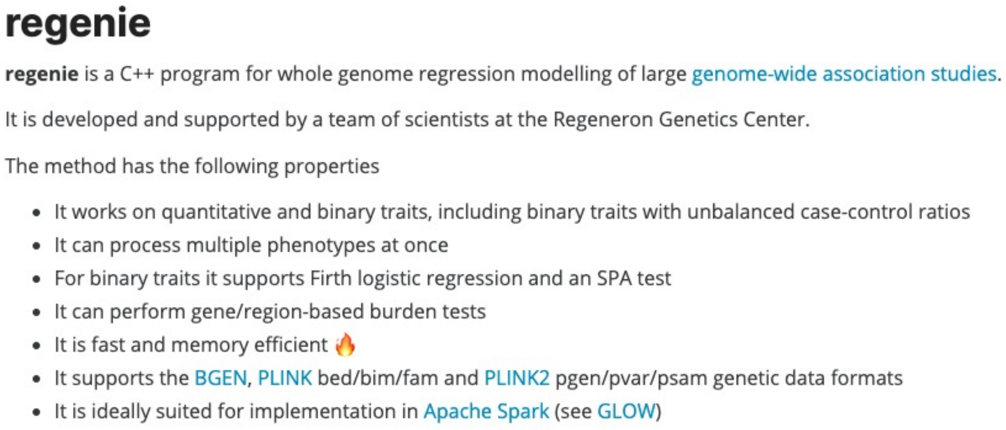
REGENERON®
GENETICS CENTER

ANALYSIS SPEED AND SCALING

- ✓ A GWAS is usually conducted with ~7-32 mio variants (dependendent on imputation depth and allele frequency cut-offs)
- ✓ Very computational demanding in biobank settings with hundreds of thousands of samples!

Computational requirements of running GWAS on 50 binary traits from the UK Biobank with different case-control ratios and distinct missing data patterns:

	CPU (hours)	Elapsed time (hours)
SAIGE	369,417	89,865
REGENIE	46,204	3,348



regenie

regenie is a C++ program for whole genome regression modelling of large genome-wide association studies.

It is developed and supported by a team of scientists at the Regeneron Genetics Center.

The method has the following properties

- It works on quantitative and binary traits, including binary traits with unbalanced case-control ratios
- It can process multiple phenotypes at once
- For binary traits it supports Firth logistic regression and an SPA test
- It can perform gene/region-based burden tests
- It is fast and memory efficient 🔥
- It supports the BGEN, PLINK bed/bim/fam and PLINK2 pgen/pvar/psam genetic data formats
- It is ideally suited for implementation in Apache Spark (see GLOW)

REGENIE is 8x faster in CPU hours

& 26.8x faster in elapsed time

https://rgcgithub.github.io/regenie/

Novo Nordisk Foundation
Center for Protein Research

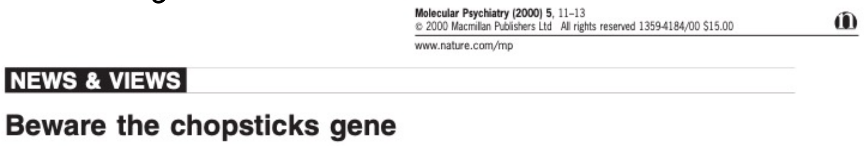
UNSUPERVISED CLUSTERING IN POPULATION ADMIXTURE

✓ Allele frequencies can differ between subpopulations

✓ Population stratification can lead to false positive associations and/or mask true associations.

Why do some people eat with chopsticks and others do not?

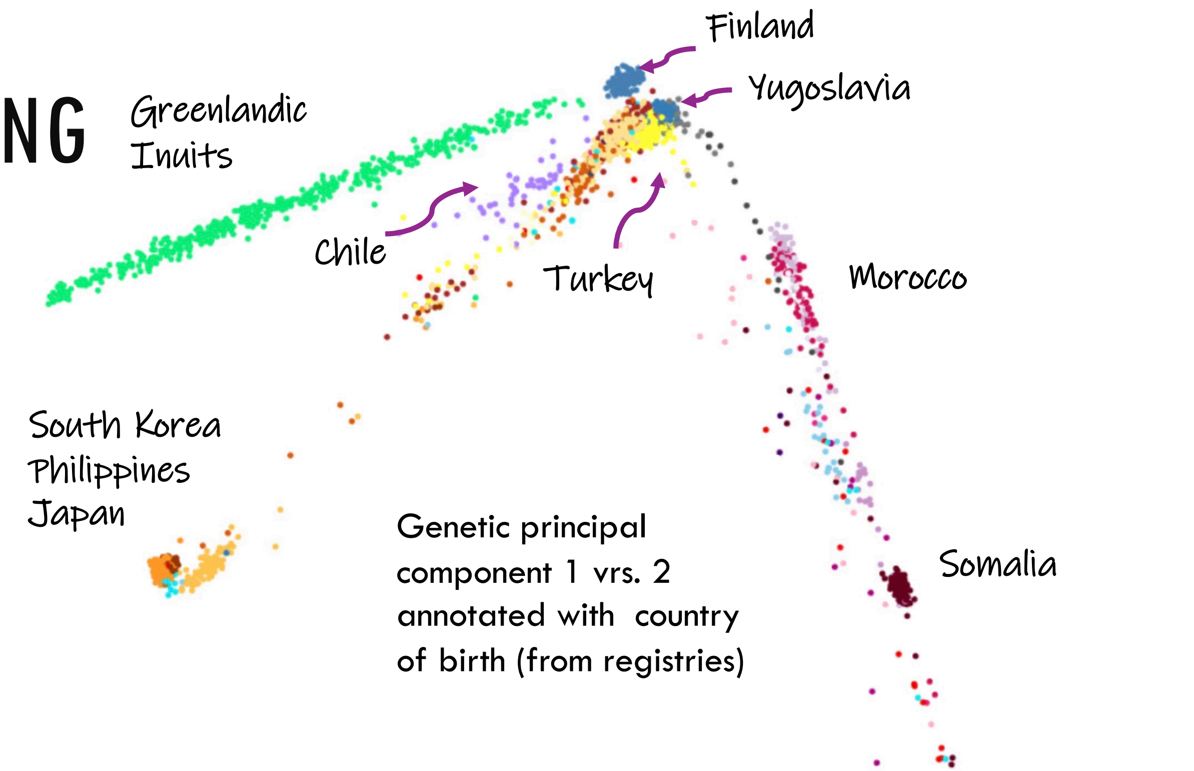
Differences in allele frequencies in Asians and Caucasians, led to the discovery of the 'successful-use-of-selected-handinstruments gene' (SUSHI) which was driven by the fact that these two populations differ in chopstick use for purely cultural rather than biological reasons.



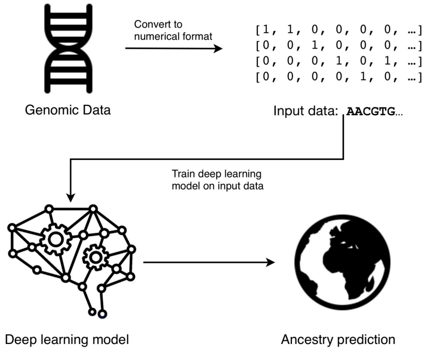
Molecular Psychiatry (2000) 5, 11–13
© 2000 Macmillan Publishers Ltd All rights reserved 1359-4184/00 $15.00
www.nature.com/mp

NEWS & VIEWS

Beware the chopsticks gene



Genetic principal component 1 vrs. 2 annotated with country of birth (from registries)

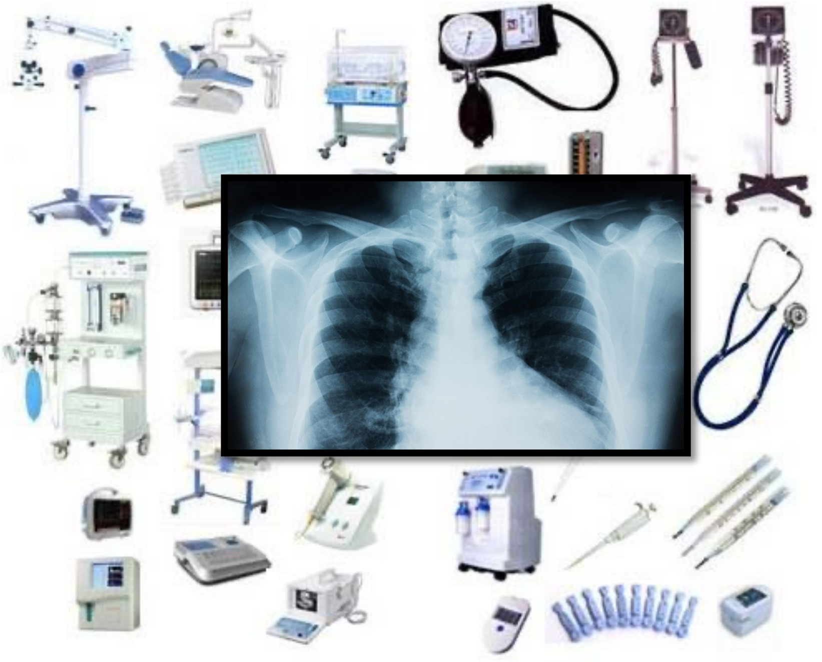


Novo Nordisk Foundation
Center for Protein Research

DEEP PHENOTYPING FOR PATIENT CHARACTERIZATION

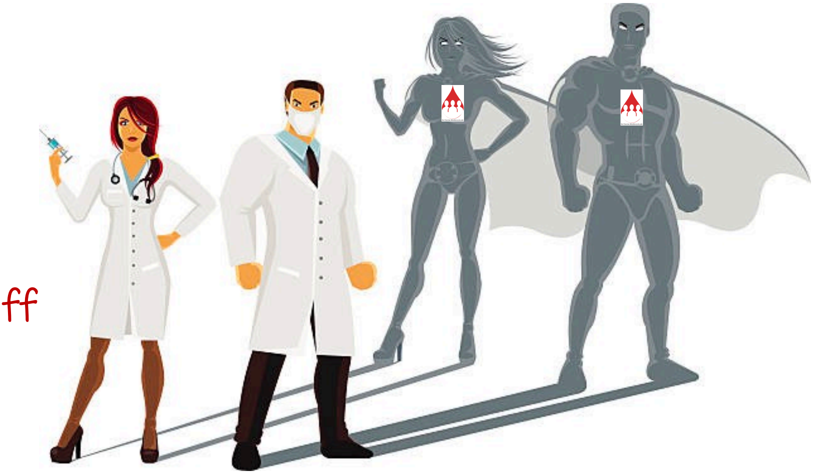
Unstructured data – "the other 80 percent of all data"

Systematically extract and preprocess features from data generated by machines



ACKNOWLEDGEMENTS

Blood Bank staff



Erik Sørensen
Sisse Rye Ostrowski
Henrik Ullum
Ole Pedersen
Kristoffer Burgdorf

Novo Nordisk Foundation
Center for Protein Research

Søren Brunak
Piotr Jaroslaw Chmura
David Westergaard
Thomas Folkmann Hansen

deCODE genetics

Kári Stefánsson
Daníel F. Guðbjartsson
Hreinn Stefánsson
Hilma Hólm
Bragi Walters

...and even more importantly:

Thanks to all the generous
blood donors who donate much more
than blood and truly make a difference!

novo nordisk fonden

Novo Nordisk Foundation
Center for Protein Research

Thank you for the attention!