genomics

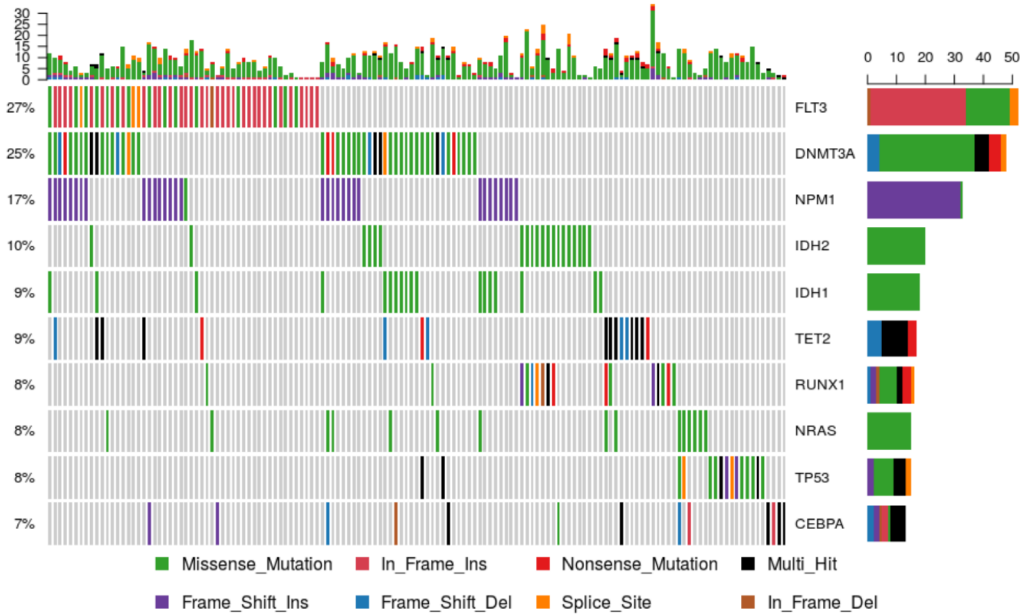
ge·no·mics

A branch of biotechnology concerned with applying the techniques of genetics and molecular biology to the genetic mapping and DNA sequencing of sets of genes or the complete genomes of selected organisms, with organizing the results in databases, and with applications of the data

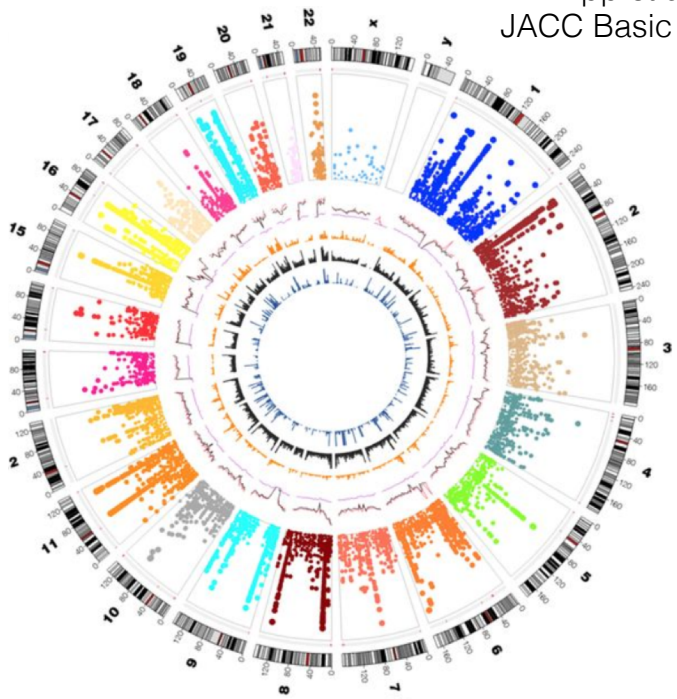
- Quickly developing field
- Big data
- Multidisciplinary
- Heterogeneous skills required
  - Genetics/biology, bioinformatics, statistics, medicine, computer science, mathematics
  - Visualization, communication skills, project management
  - R, Python, cloud computing, Apache Spark, JavaScript…

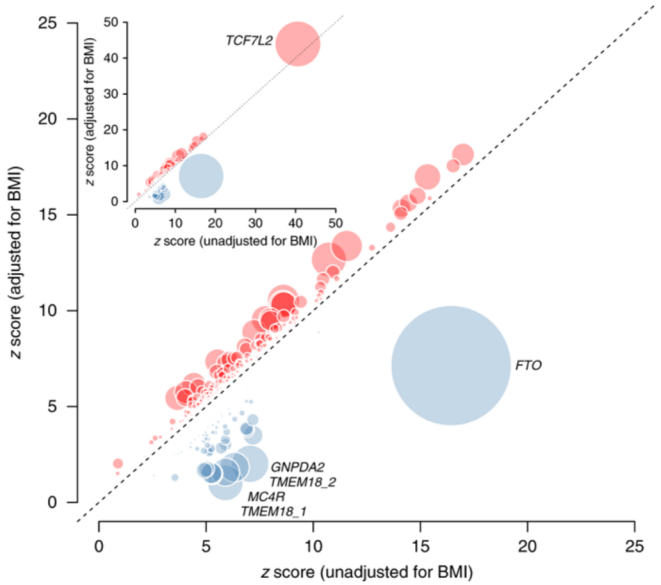
Meaningful, impact on patient care

maftools

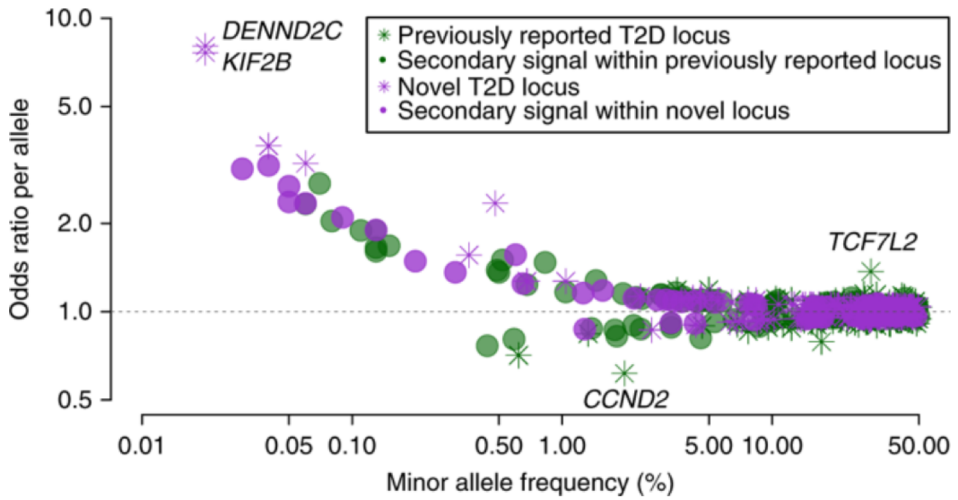


Kipp et al. 2017
JACC Basic Transl Sci.

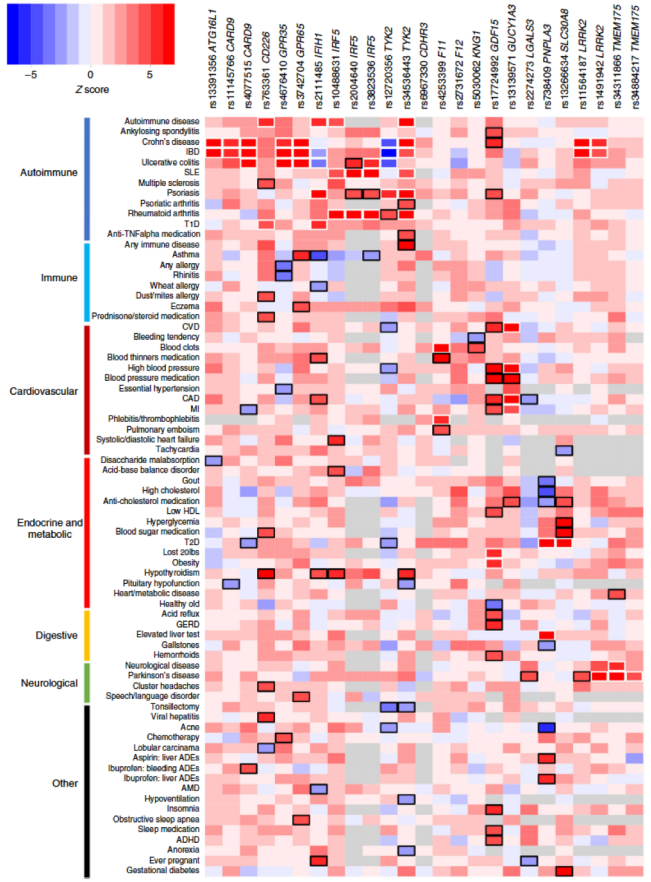




Mahajan et al. 2018 Nature Genetics.

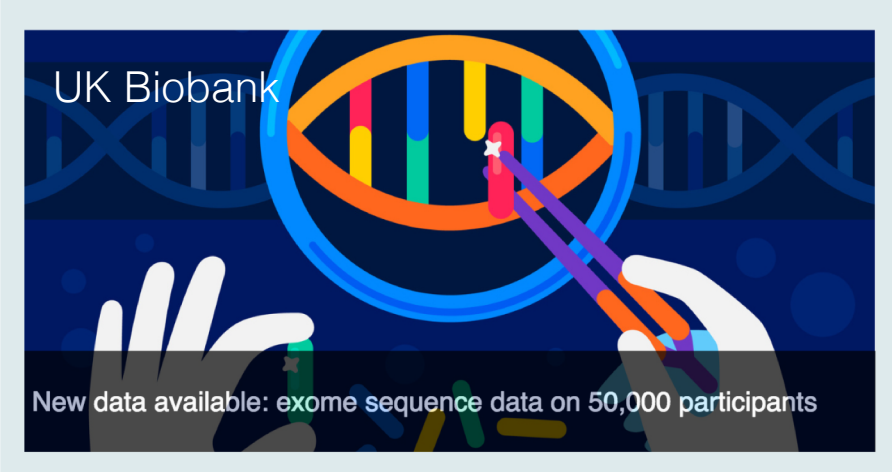


Mahajan et al. 2018 Nature Genetics.

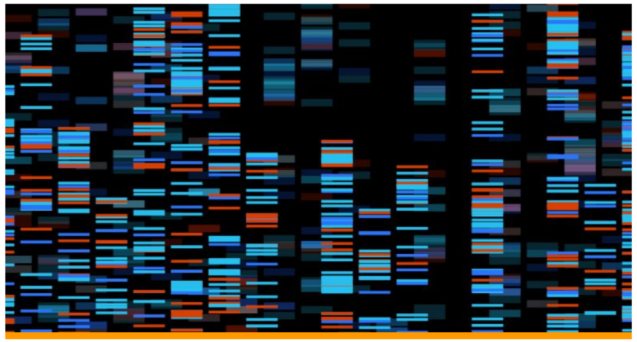


Diogo et al. 2018 Nat. Commun.

Exome sequencing



Whole-genome sequencing

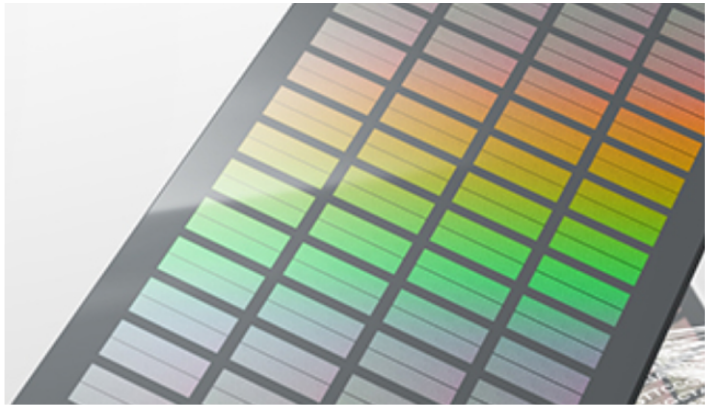


ACHIEVEMENT UNLOCKED

UK Researchers Just Finished Sequencing 100,000 Human Genomes

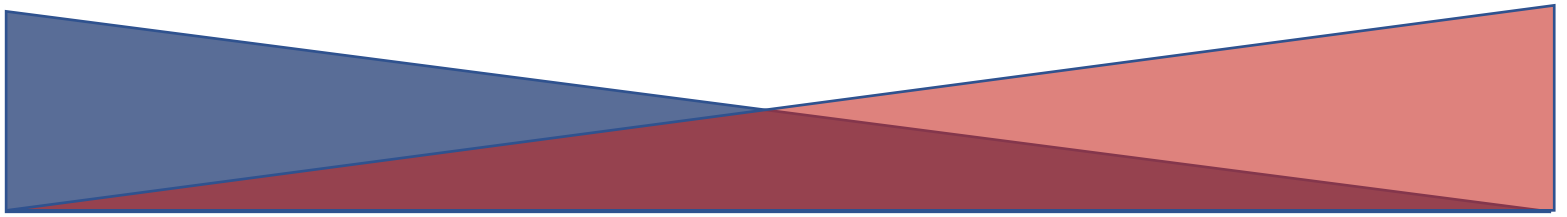
And their next goal will smash that last one.

Genotyping



Monogenic

Polygenic



Rare diseases

Common diseases

Cystic fibrosis

Type 2 diabetes
Coronary artery disease

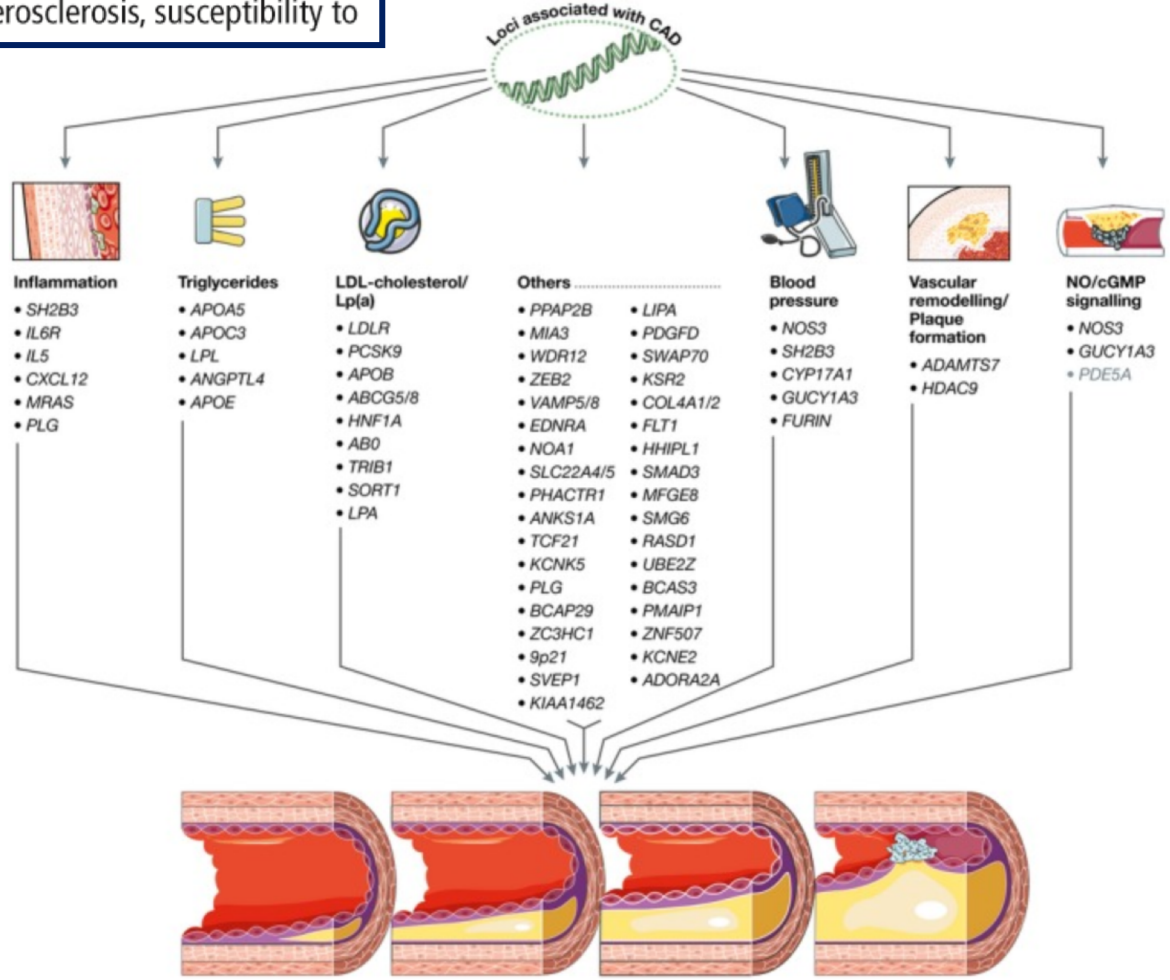
19

63 million base pairs

Coxsackie virus sensitivity
Cyclic hematopoiesis
Fucosyltransferase-6 deficiency
Hypocalciuric hypercalcemia, type II
Leukemia, myeloid/lymphoid or mixed-lineage
Wegener granulomatosis autoantigen
Bleeding disorder
Persistent Mullerian duct syndrome, type I
Mucolipidosis
Glutaricaciduria, type I
Leprechaunism
Rabson-Mendenhall syndrome
Diabetes mellitus, insulin-resistant
Ichthyosis
Leukemia, T-cell acute lymphoblastoid
Liposarcoma
Mycobacterial and salmonella infections, susceptibility to
Eye color, green/blue
Hemiplegic migraine, familial
Episodic ataxia, type 2
Ataxia, spinocerebellar and cerebellar
Leukemia, acute myeloid
Mannosidosis, alpha, types I and II
Alzheimer disease, late onset
Glomerulosclerosis, focal segmental
Deafness, autosomal dominant
Hypercalcemia, familial benign, Oklahoma type, type III
Orofacial cleft
Charcot-Leyden crystal protein
Hemolytic anemia
Hydrops fetalis
Malignant hyperthermia susceptibility
Central core disease
Osteodysplasia, polycystic lipomembranous
Maple syrup urine disease, type Ia
Camurati-Engelmann disease
Myotonic dystrophy
Heart block, progressive familial, type I
Optic atrophy
3-methylglutaconicaciduria, type III
Cystic fibrosis modifier
Meconium ileus in cystic fibrosis, susceptibility to
Cone dystrophy
Leber congenital amaurosis
Retinitis pigmentosa, late-onset dominant
Diabetes mellitus, noninsulin-dependent
Hyperferritinemia-cataract syndrome
Hypogonadism, hypergonadotropic
Retinitis pigmentosa, autosomal dominant
Ectrodactyly, ectodermal dysplasia, cleft lip/palate

Ataxia, cerebellar, Cayman type
Convulsions, familial febrile
Guanidinoacetate methyltransferase deficiency
Muscular dystrophy
Hirschsprung disease
Peutz-Jeghers syndrome
Leukemia, acute lymphoblastic
Atherosclerosis, susceptibility to
Malaria, cerebral, susceptibility to
Sicca syndrome
Glioblastoma
Thyroid carcinoma, nonmedullary
Low density lipoprotein receptor
Hypercholesterolemia, familial
Arteriopathy, cerebral
Pseudoachondroplasia
Epiphyseal dysplasia, multiple
Severe combined immunodeficiency disease
Hair color, brown
Leigh syndrome
MHC class II deficiency
Exostoses, multiple, type 3
Benign familial infantile convulsions
Leukemia/lymphoma, B-cell
Spondylocostal dysostosis, autosomal recessive
Prostate-specific antigen
Spastic paraplegia , autosomal dominant
Cystinuria, types II and III
Nephrosis, congenital, Finnish type
Generalized epilepsy with febrile seizures plus
Ovarian carcinoma
Microcephaly, autosomal recessive
Hyperlipoproteinemia, types Ib and III
Myocardial infarction susceptibility
Cytochrome P450 (coumarin resistance)
Nicotine addiction, protection from
X-ray repair
Excision repair
Xeroderma pigmentosum, group D
Trichothiodystrophy
DNA ligase I deficiency
Polio virus receptor
Herpes virus entry mediator B
Glutaricaciduria, type IIB
Colorectal cancer
Leukemia, T-cell acute lymphoblastic
Shaw-related subfamily genes
Melanoma inhibitory activity
Cardiomyopathy, familial hypertrophic

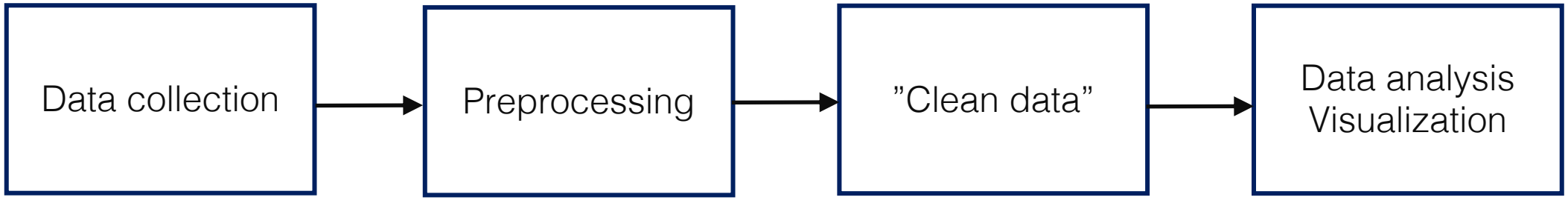
Atherosclerosis, susceptibility to



Kessler et al. EMBO Mol Med 2016

Looks familiar?

80% of the time spent cleaning/preprocessing the data,
(if we don't consider the data collection)



Replicability, scalability

Systematic workflows

May 10, 2018

Working

Genotype imputation workflow v3.0

Kalle Pärn[1], Marita A. Isokallio[1], Javier Nunez Fontarnau[1], Aarno Palotie[2], Samuli Ripatti[2], Priit Palta[2]

[1]FIMM, University of Helsinki, [2]equal contribution; FIMM, University of Helsinki

dx.doi.org/10.17504/protocols.io.nmndc5e

FIMM HumGen Sequencing Informatics

Priit Palta
FIMM, University of Helsinki

4 Ensure that duplicate individuals do not exist in the chip data or between the chip and reference panel as this would compromise the imputation.

Input files:
• <dataset>_SNPID.vcf.gz
• panel_sample_IDs.txt

Outpt file:
• <dataset>_noduplicate_samples.vcf.gz

Note: Confirm the results for example by comparing the input and output file sizes and line counts.

```
# Copy the panel sample ID file with a different name to the w
cp /path/to/panel_sample_IDs.txt duplicate_sample_IDs.txt

# Generate a list of sample IDs from the chip data, keep only
bcftools query -l <dataset>_SNPID.vcf.gz | uniq -d >> duplicat

# Remove the listed sample IDs from the chip data VCF
bcftools view -S ^duplicate_sample_IDs.txt --force-samples <da
```

bcftools query parameters:
-l list of sample IDs
uniq
-d only print duplicate lines
bcftools view parameters:
-S file of sample IDs to include
^ exclusion prefix
--force-samples only warn about unknown subset of samples
-Oz compressed output

5 Ensure that there are no duplicate variants (these might appear in some chip genotype datasets). If duplicate variants are present, they need to be removed before imputation.

Input file:
• <dataset>_noduplicate_samples.vcf.gz

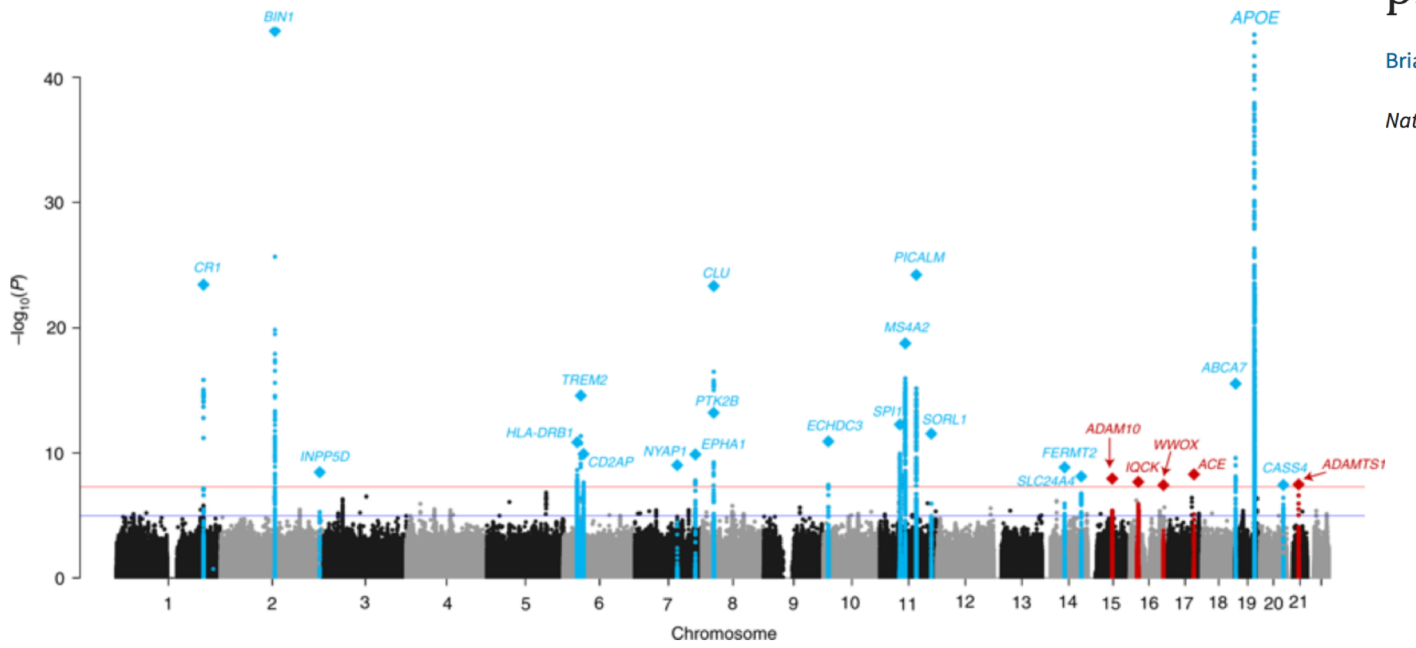
Outpt files:
• <dataset>_duplicate_variants.txt
• <dataset>_noduplicate_variants.vcf.gz

Note: Confirm the results for example by comparing the input and output file sizes and line counts.

```
# Store a list of duplicate positions
bcftools query -f '%ID\n' <dataset>_noduplicate_

# Check whether the file contains any variants
if [ -s <dataset>_duplicate_variants.txt ]; then
   # Then remove the duplicate variants
   bcftools view -e ID=@<dataset>_duplicate_vari
else
  # If the file is empty i.e. no duplicate varia
  mv <dataset>_noduplicate_samples.vcf.gz <datas
fi
```

Genome-wide association study,
statistical association between genetic variation and disease or trait



The threshold for genome-wide significance ($P < 5 \times 10^{-8}$) is indicated by the red line, while the blue line represents the suggestive threshold ($P < 1 \times 10^{-5}$). Loci previously identified by the Lambert et al.[1] IGAP GWAS are shown in blue and newly associated loci are shown in red. Loci are named for the closest gene to the sentinel variant for each locus. Diamonds represent variants with the smallest P values for each genome-wide locus.

nature genetics

Article | Published: 28 February 2019

Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing

Brian W. Kunkle, Benjamin Grenier-Boley, […] Margaret A. Pericak-Vance

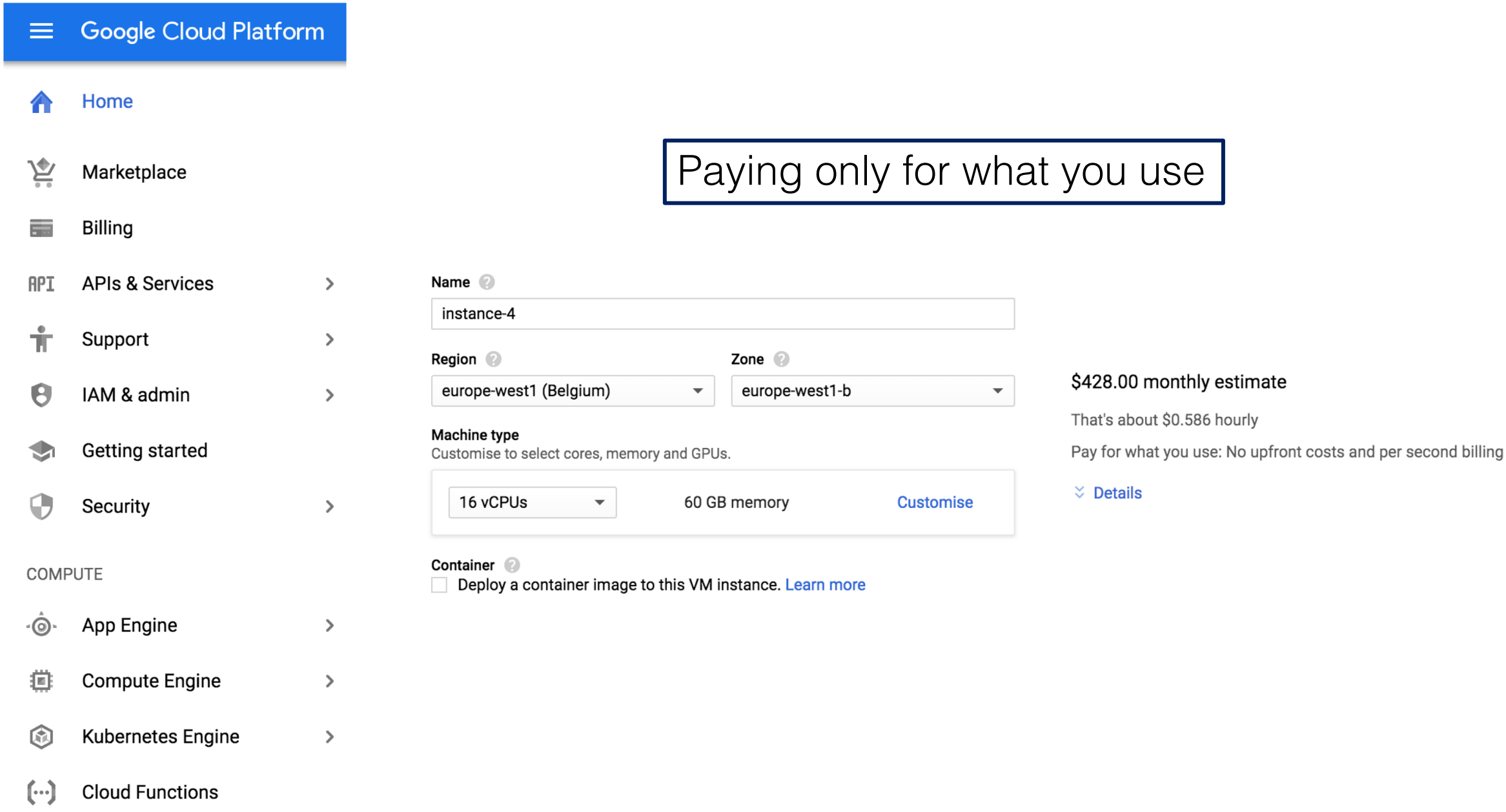
Nature Genetics **51**, 414–430 (2019) | Download Citation

Summary-level data

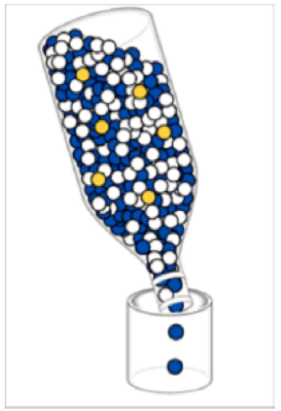
SNP	Chr	Pos	EA	NEA	EAF	Beta	SE	Pvalue	Neff	
1:100000012		1	100000012		T	G	0.25	-0.026	0.0073	4.0e-04 231420
1:10000006		1	10000006		A	G	0.0047	-0.038	0.056	4.9e-01 225429
1:100000135		1	100000135		A	T	0.99	-0.033	0.055	5.5e-01 226311
1:100000436		1	100000436		T	C	1	-0.098	0.19	6.1e-01 185906
1:100000827		1	100000827		T	C	0.3	-0.023	0.0069	7.5e-04 231420
1:100000843		1	100000843		T	C	0.94	0.0006	0.014	9.7e-01 231420
1:100001138		1	100001138		A	G	0.97	-0.0041	0.022	8.6e-01 230619
1:100001201		1	100001201		T	G	0.1	-0.017	0.0099	8.9e-02 231420
1:100001233		1	100001233		T	C	1	-0.45	0.24	5.7e-02 171397
1:100001483		1	100001483		A	G	0.0002	0.7	0.47	1.4e-01 129508
1:100001846		1	100001846		A	G	0.003	0.062	0.087	4.7e-01 216365
1:100001904		1	100001904		T	C	0.0005	0.048	0.26	8.5e-01 127786
1:100002106		1	100002106		C	G	0.0009	0.15	0.15	3.1e-01 194119
1:100002236		1	100002236		T	G	1	-0.38	0.4	3.5e-01 111960
1:100002271		1	100002271		A	G	0.99	0.051	0.049	2.9e-01 230618

23.5M rows

Mahajan et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nature Genetics, 2018.


Google Cloud Platform
Home
Marketplace
Billing
APIs & Services
Support
IAM & admin
Getting started
Security
COMPUTE
App Engine
Compute Engine
Kubernetes Engine
Cloud Functions
Paying only for what you use
Name
instance-4
Region
europe-west1 (Belgium)
Zone
europe-west1-b
Machine type
Customise to select cores, memory and GPUs.
16 vCPUs
60 GB memory
Customise
Container
Deploy a container image to this VM instance. Learn more
$428.00 monthly estimate
That's about $0.586 hourly
Pay for what you use: No upfront costs and per second billing
Details

- Countries with Biobanks
- Universal Healthcare
- Unique personal identity code
- Isolated population
- Recalling made easy



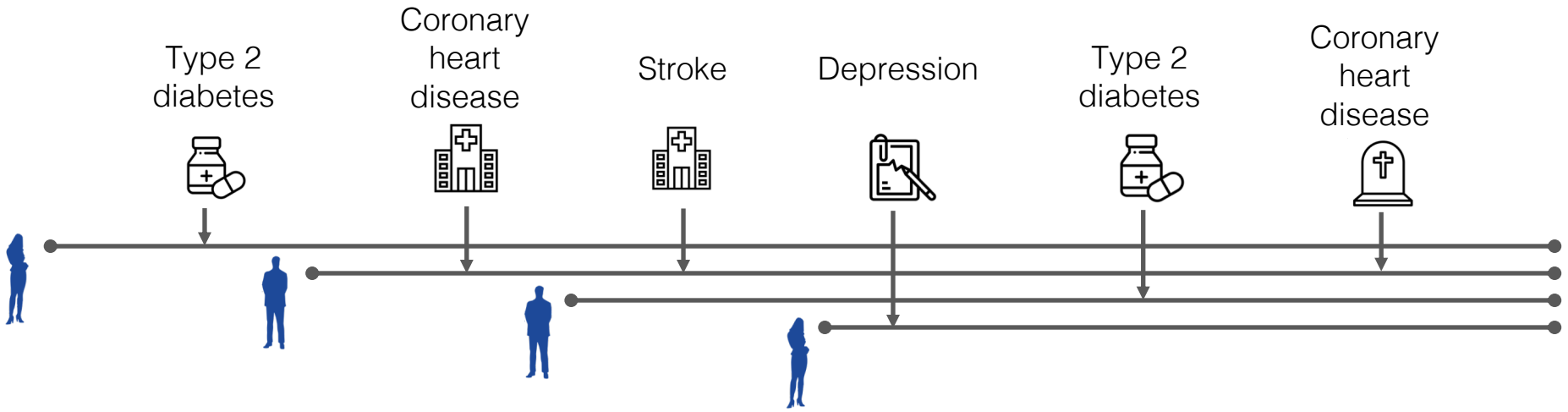
POPULATION ISOLATE

Parliament of Finland has just voted for new regulation regarding #secondaryuse of #healthdata. New authority will be established.

Benefits expected e.g. in healthcare, research,innovation & information management
#datadrivenhealthcare
#personalisedmedicine
@THLorg @MSAH_News

Käännä twiitti

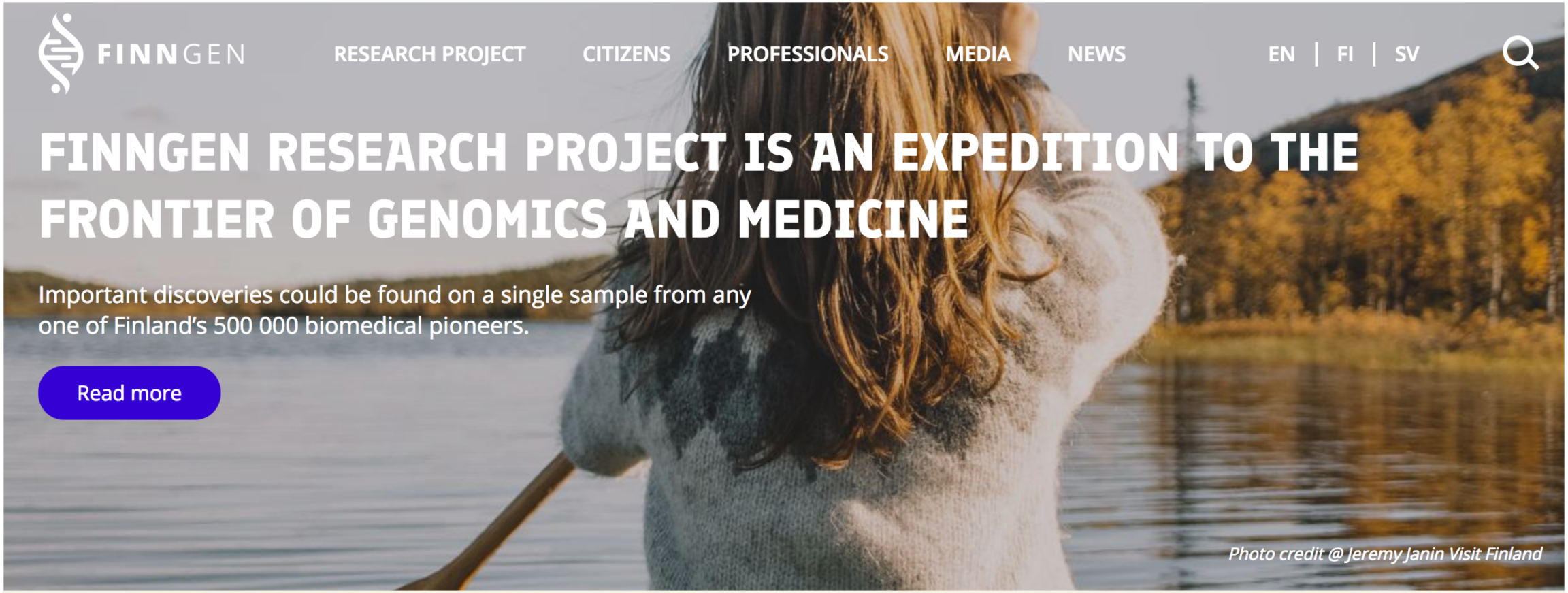
Comprehensive and longitudinal population medical data available for research



Discovering and confirming associations between genetic variation and diseases

Exploring a broad set of outcomes, detecting potential negative consequences of proposed therapeutic hypotheses

Icons: FlatIkon



FINNGEN BRINGS TOGETHER THE NATION-WIDE NETWORK OF FINNISH BIOBANKS.

Every Finn can be a part of the FinnGen study by giving a biobank consent.

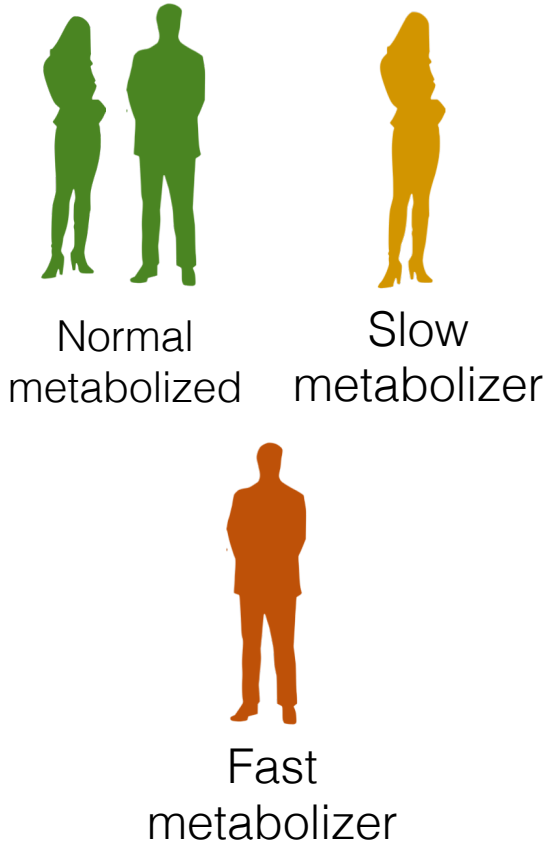
SAMPLES AVAILABLE

234 000

Samples needed by 2023: 500 000

Clinical applications

- Whole-exome sequencing (WES) is being used clinically to diagnose rare monogenic disorders, especially when standard tests have failed
- WES and WGS to detect novel disorders
- Personalized medicine
  - Pharmacogenomics and individualized drug therapy
  - Risk-based screening and prevention of common disases

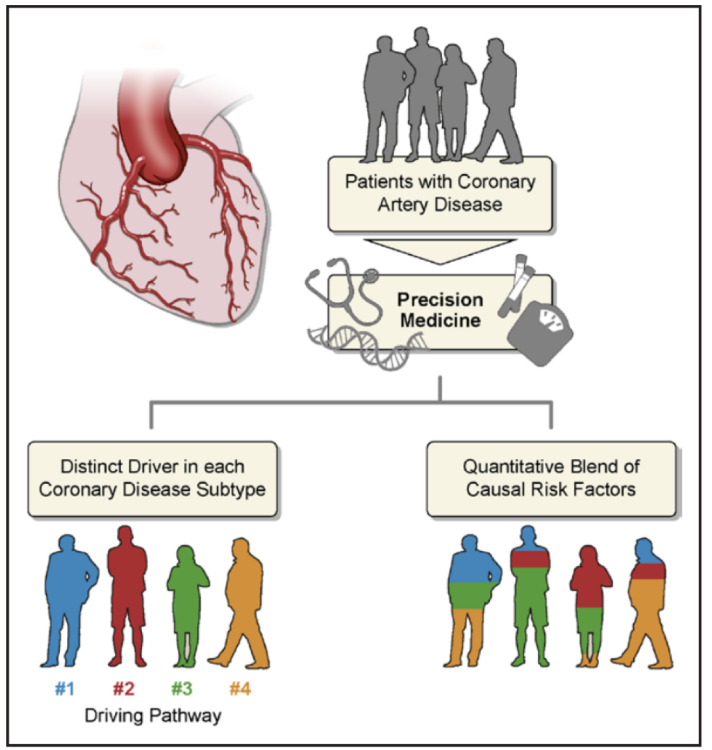


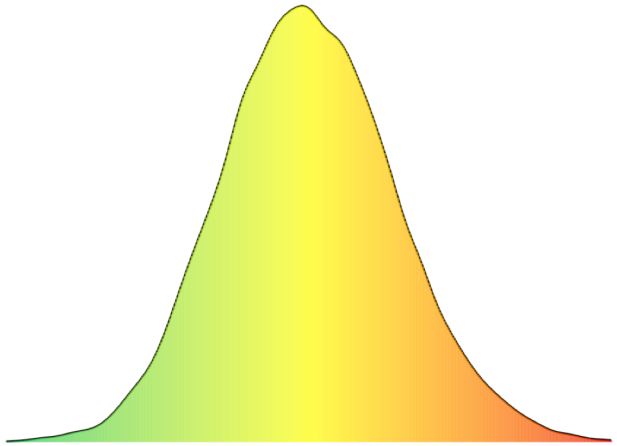
Polygenic risk scores

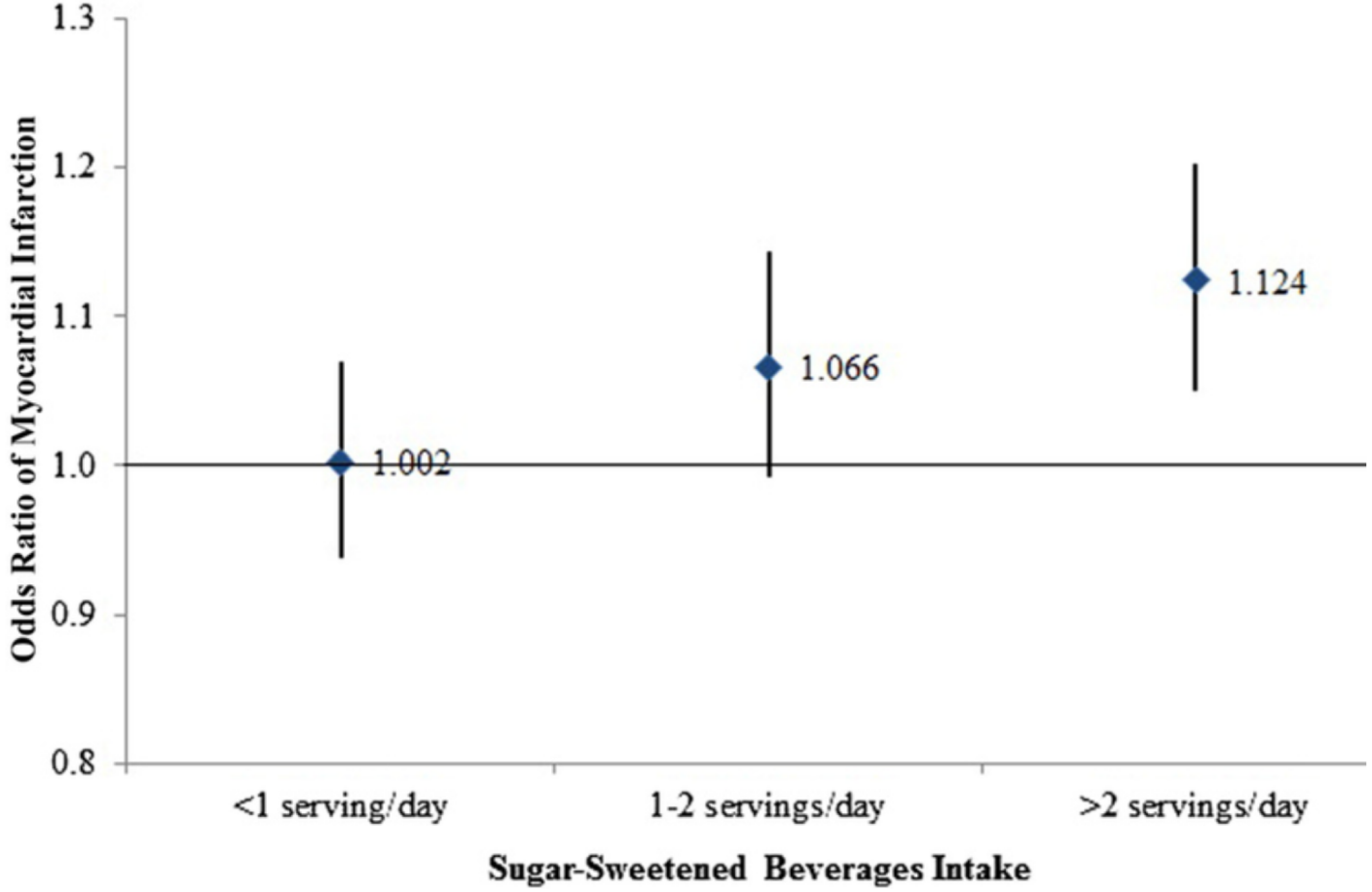
Hundreds of individually modest risk factors can be combined to generate predictive information

Predicting health outcomes and therapeutic responses

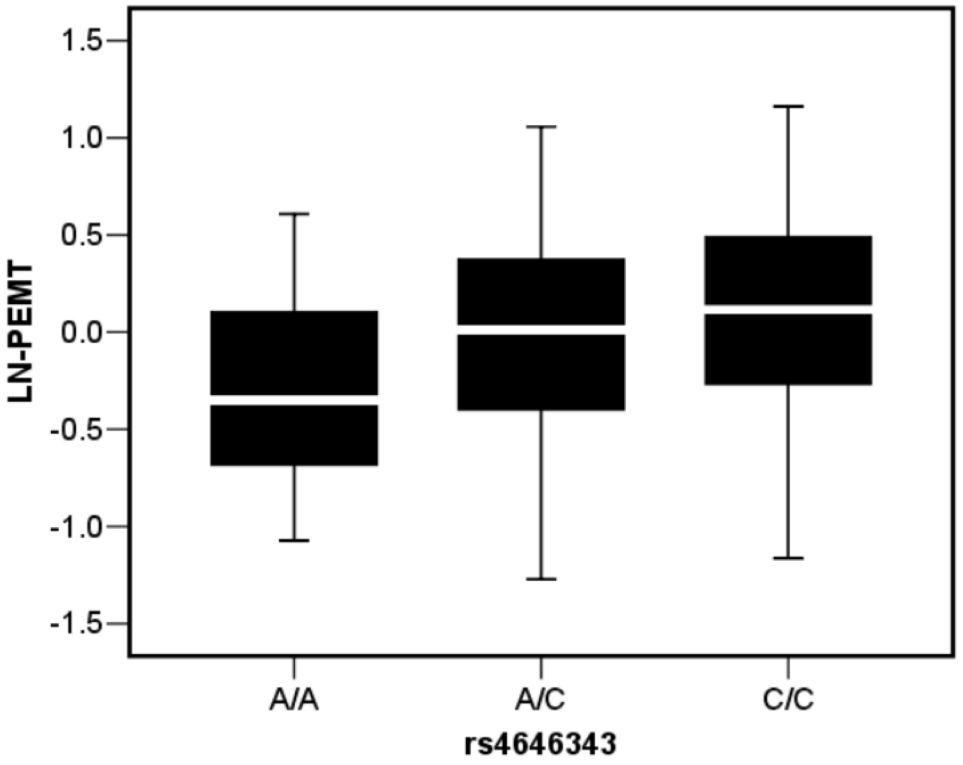
Khera & Kathiresan. Circulation 2017







Zheng et al. AJCN 2016



Sharma et al. PLoS ONE 2013

The New York Times

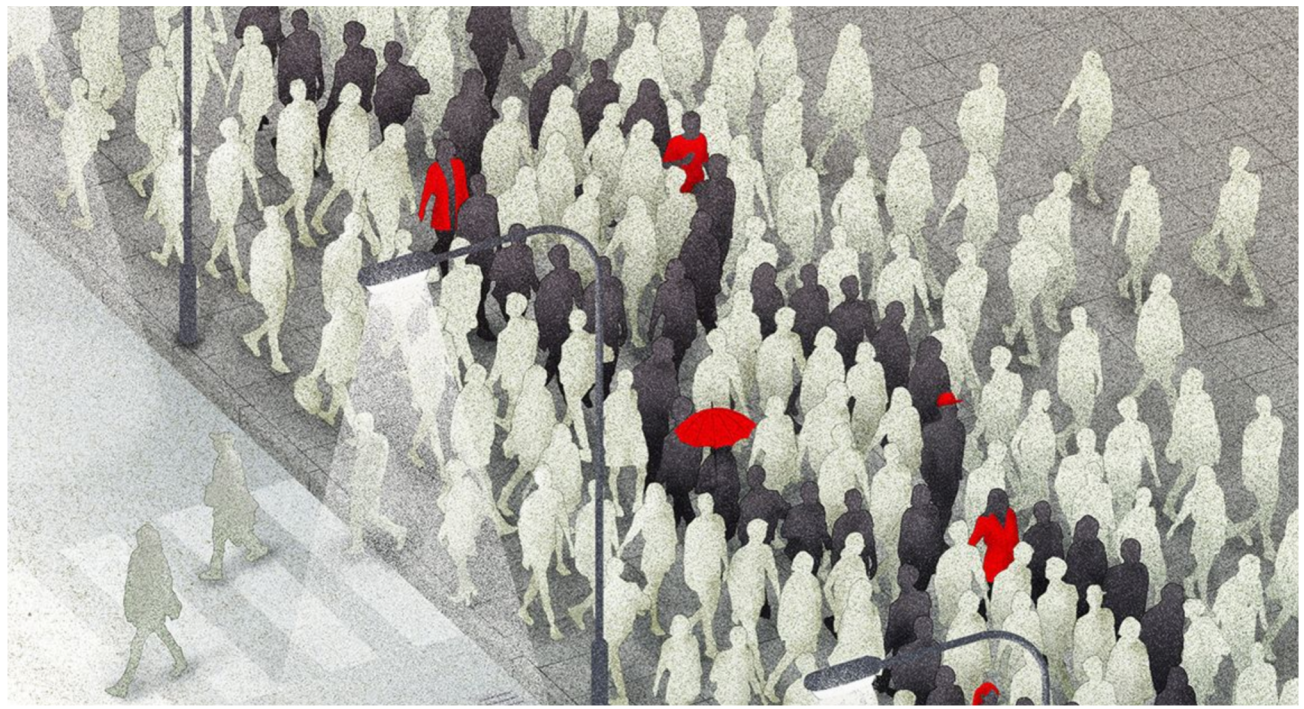
https://nyti.ms/2UyoP6U

MENU
nature
International journal of science

The approach to predictive medicine that is taking genomics research by storm

Polygenic risk scores represent a giant leap for gene-based diagnostic tests. Here's why they're still so controversial.

NEWS FEATURE · 10 OCTOBER 2018



MIT Technology Review

Rewriting Life

23andMe thinks polygenic risk scores are ready for the masses, but experts aren't so sure

Regulation

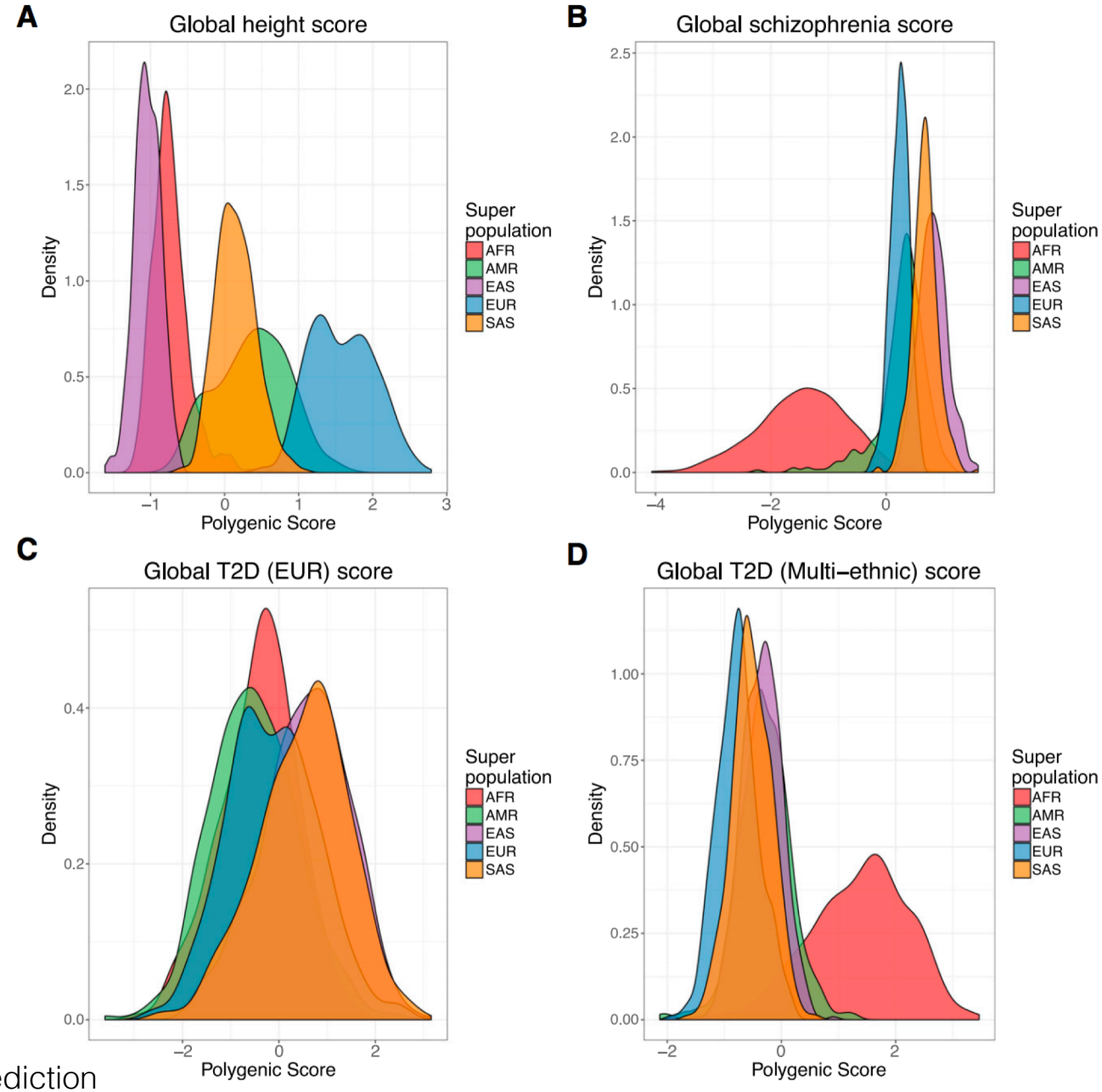
Since then, however, the science of prediction has improved and regulations have loosened. According to 23andMe, the current diabetes report needs no regulation at all. That is because it falls into an exemption for low-risk tests and phone apps that offer only "general wellness" suggestions, not real medical advice or diagnoses.

Impact on behavior and public health?

Prediction in different ancestries?

The genetic predictions are also especially spotty for African-Americans. The company developed its model using DNA from white people of European ancestry, who make up most of its database. The result is that the predictions perform less well for other populations.

Biased genetic discoveries influence disease risk inferences



Martin et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. AJHG 2017