Multi-dimensional Cohorts in Precision Medicine

Klaudia Walter

Wellcome Sanger Institute

18-10-2019

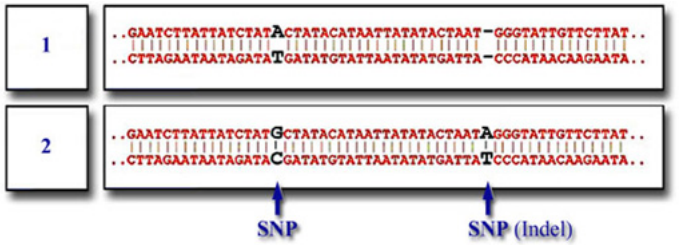
Overview

- Applications and value of human complex trait genetics in precision medicine
- Progress of whole-genome sequencing projects to study human complex traits
- Value of different high-throughput phenotyping platforms

The case for studying complex human traits

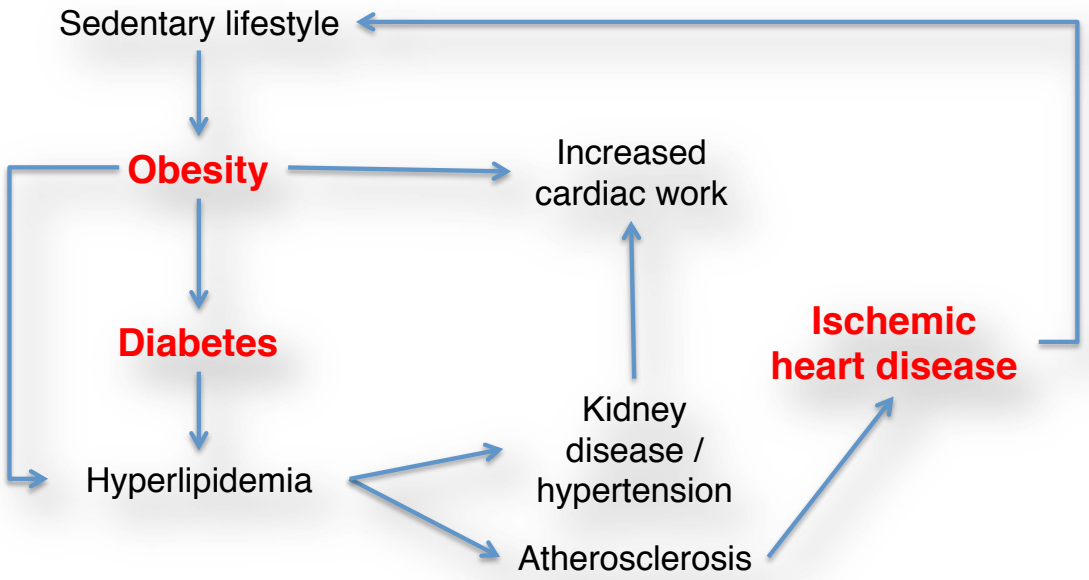
Populations as ‘living laboratories’

Populations as reservoirs of genetic variation



- Many human characteristics can be represented by heritable, quantitative measurements (e.g. height, glucose levels, BMI, muscle strength, etc.)
- Phenotypic variation is a readout of complex biological processes
- Disease is an extreme of a ‘normal’ spectrum of phenotype or function
- Many genes/pathways/processes are shared between normal phenotypic variation and disease
- **Use genetics to study fundamental processes at the basis of human biology in health and disease**

Use of genetics in precision medicine



- Discover new **disease genes/processes**
- Characterise differences between **healthy and disease** state and between different diseases
- Quantify the burden of **genetic vs environmental** risk
- Identify individuals in the population at **greatest risk of developing disease**
- Discover or prioritise **new drugs** for treating disease

UK Cohorts Studies

- **ALSPAC** – Birth cohort with 14,500 families
- **TwinsUK** – Twins registry with 14,000 identical and non-identical twins
- **UK10K cohorts** – 3,781 whole genome sequence samples
  - ALSPAC – 1,927 samples
  - TwinsUK – 1,854 samples
- **INTERVAL** – Blood donor cohort with 50,000 samples
- **UKBioBank** – Prospective cohort study with 500,000 samples

Lessons learned from UK10K

UK10K: 10,000 UK Genomes (2010-2013)

- **Design**
  - 10.4M GBP strategic award grant by the Wellcome Trust
  - 164 researchers from 51 institutions
  - Sequence 10,000 samples from UK and Finland

- **Goals**
  - Exhaustive discovery of rare and low frequency variants
  - Direct association of sequenced samples
  - Provide a sequence and phenotype variation resource for the community

Project arms

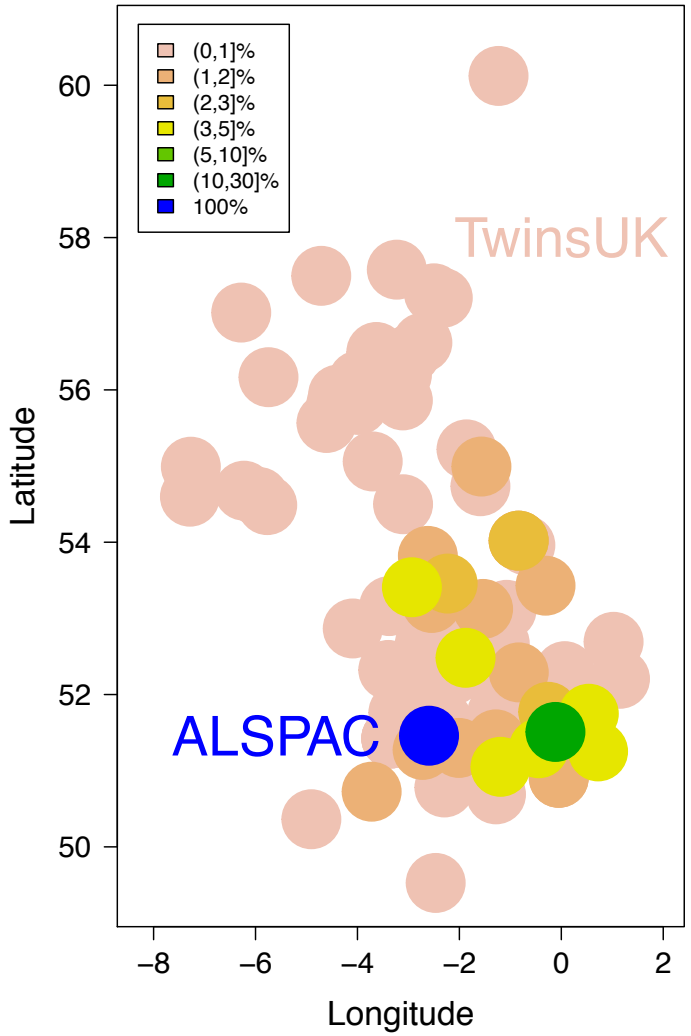
Case-control
Trios
Families
Genetic isolates

Severe childhood obesity
Population extremes

Congenital Heart Disease*
Thyroid disorders
Familial Hypercholesterolaemia
Coloboma*
Severe Insulin Resistance*
Intellectual Disability
Ciliopathies*
Neuromuscular*

* Several distinct clinical phenotypes

UK10K cohorts: ALSPAC and TwinsUK



- **ALSPAC (http://www.bristol.ac.uk/alspac)**
  - Avon Longitudinal Study of Parents and Children, also known as Children of the 90s
  - Birth cohort study
  - Recruitment between April 1991 and December 1992
  - Bristol area
  - 1,927 children in UK10K cohorts (~18 years old)

- **TwinsUK (https://twinsuk.ac.uk)**
  - UK's largest adult twin registry
  - Ages between 16 and 100
  - 1,854 female twins in UK10K cohorts (median age 46 years)

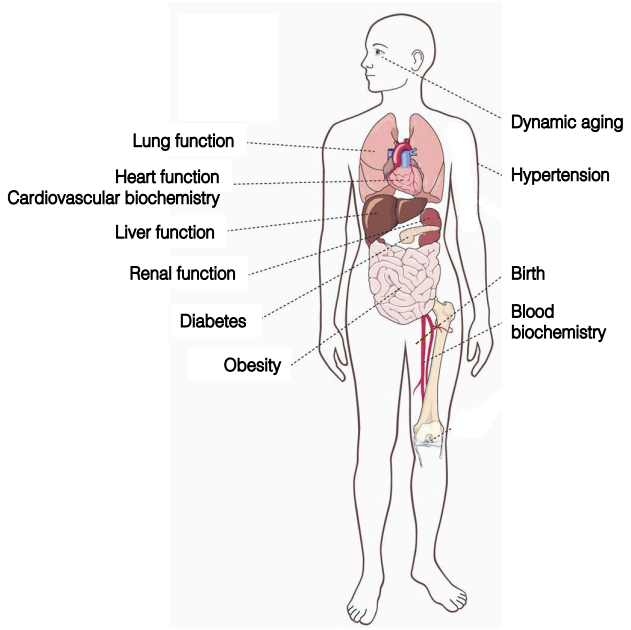
Both cohorts with deep genetic and phenotype coverage (clinical, questionnaire, molecular)

Rare variants in complex traits

UK10K cohorts project
Co-leads: Nicole Soranzo, Nicholas Timpson, Brent Richards

Study design

- 3,781 low-read depth whole-genome sequencing
- 64 quantitative phenotypes
- Combination of single-variant and rare variant tests
- Locus discoveries
- Properties of allelic spectrum and optimal designs



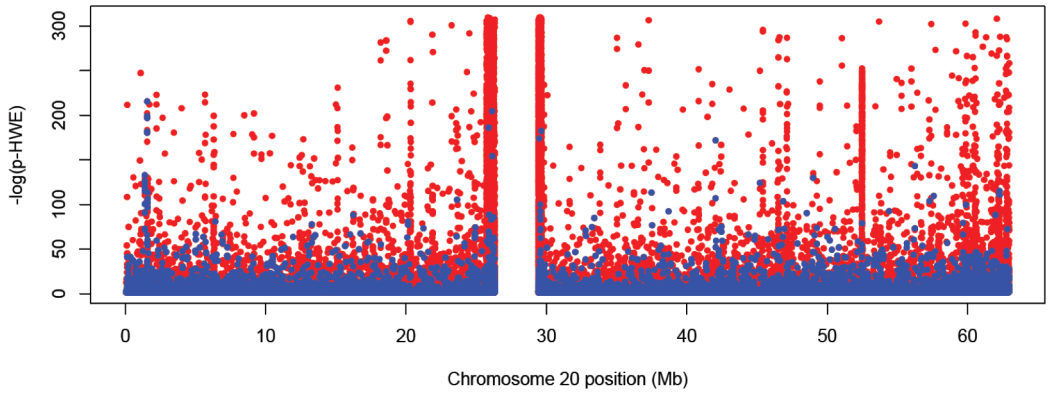
Walter, *et al.* Nature 2015

Resources

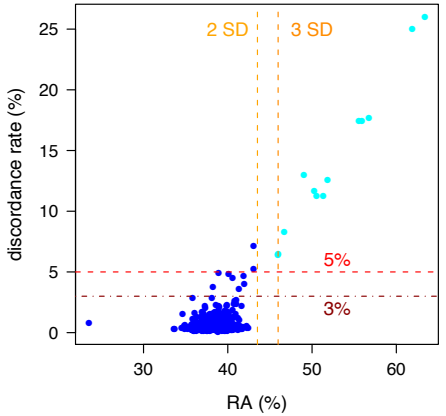
- UK10K Cohorts dataset (3,781 WGS and 64 phenotypes released to the European Genome-phenome archive)
- UK10K Haplotype reference panel (imputation reference panel for 3,781 individuals released to the European Genome-phenome archive; it has been used as a backbone for imputation of the UK Biobank Phase1 study alongside 1000 Genomes Project reference set)
- UK10K Cohorts Associations Results Browser (http://www.uk10k.org/dalliance/)

Quality control is important

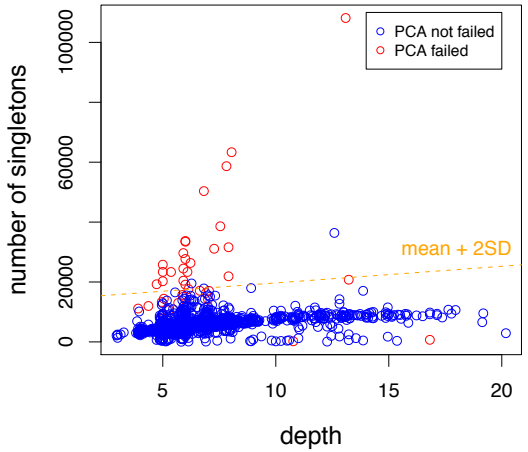
Variant quality score recalibration enables efficient filtering of variants, *e.g.* sites that are not in Hardy-Weinberg equilibrium



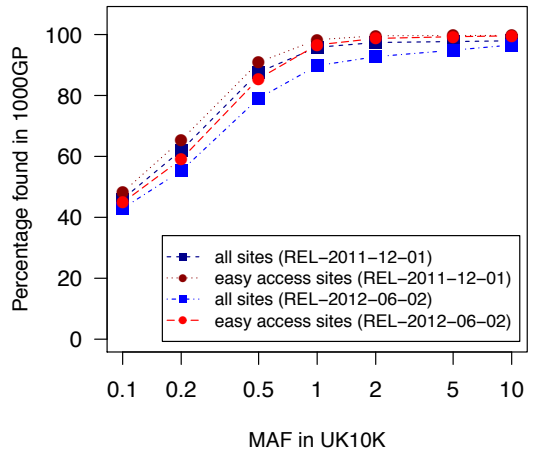
Heterozygous/homozygous genotype ratio agrees with array genotype discordance rate



Number of singleton calls increases with read coverage, but singleton outliers also likely to be population outliers

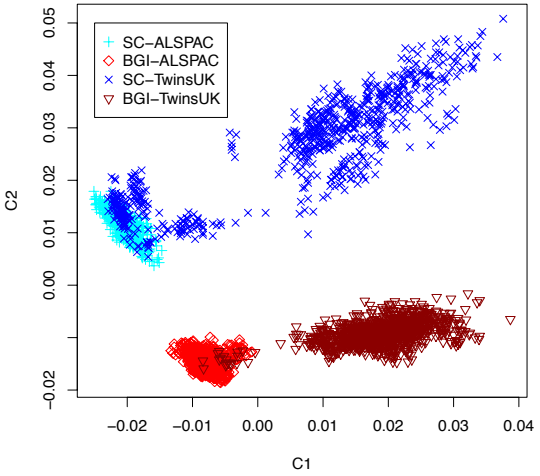


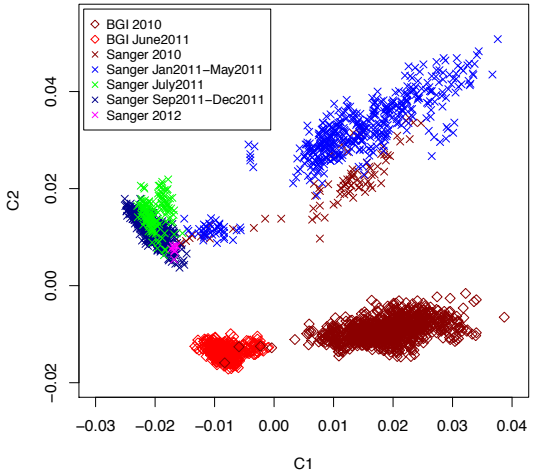
Common variant sites are shared with published data sets, *e.g.* the 1000 Genomes Project



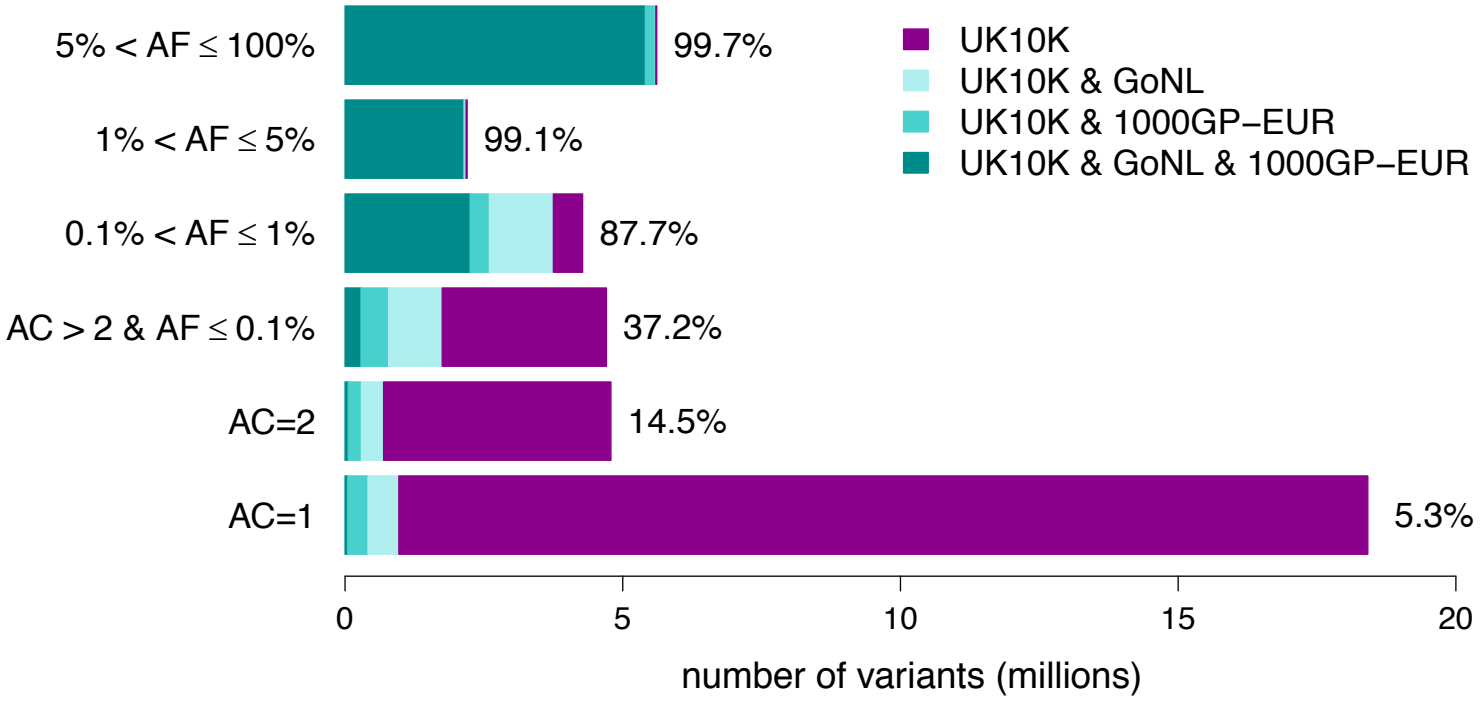
	2011-12-01	2012-06-02
MAF	overlap(%)	overlap(%)
0.1	46.0	42.9
0.2	62.0	55.2
0.5	87.6	79.1
1.0	95.8	89.8
2.0	97.4	92.7
5.0	97.7	94.9
10.0	98.0	96.6

PCA shows batch effects by cohort and by sequencing centre





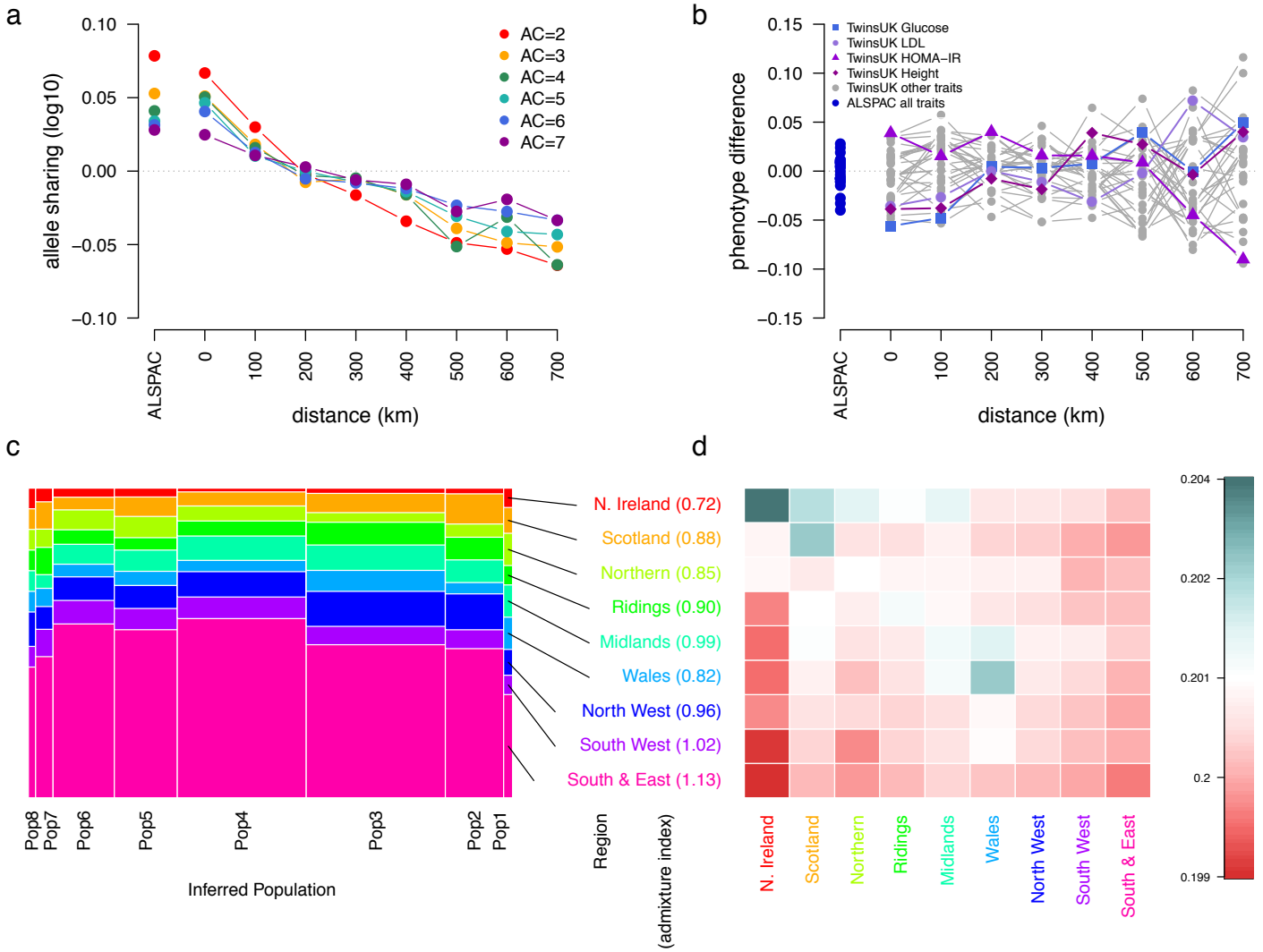
Allele sharing between populations



Comparison of variants in UK10K with Genome of the Netherlands (GoNL) and the European samples of the 1000 Genomes Project

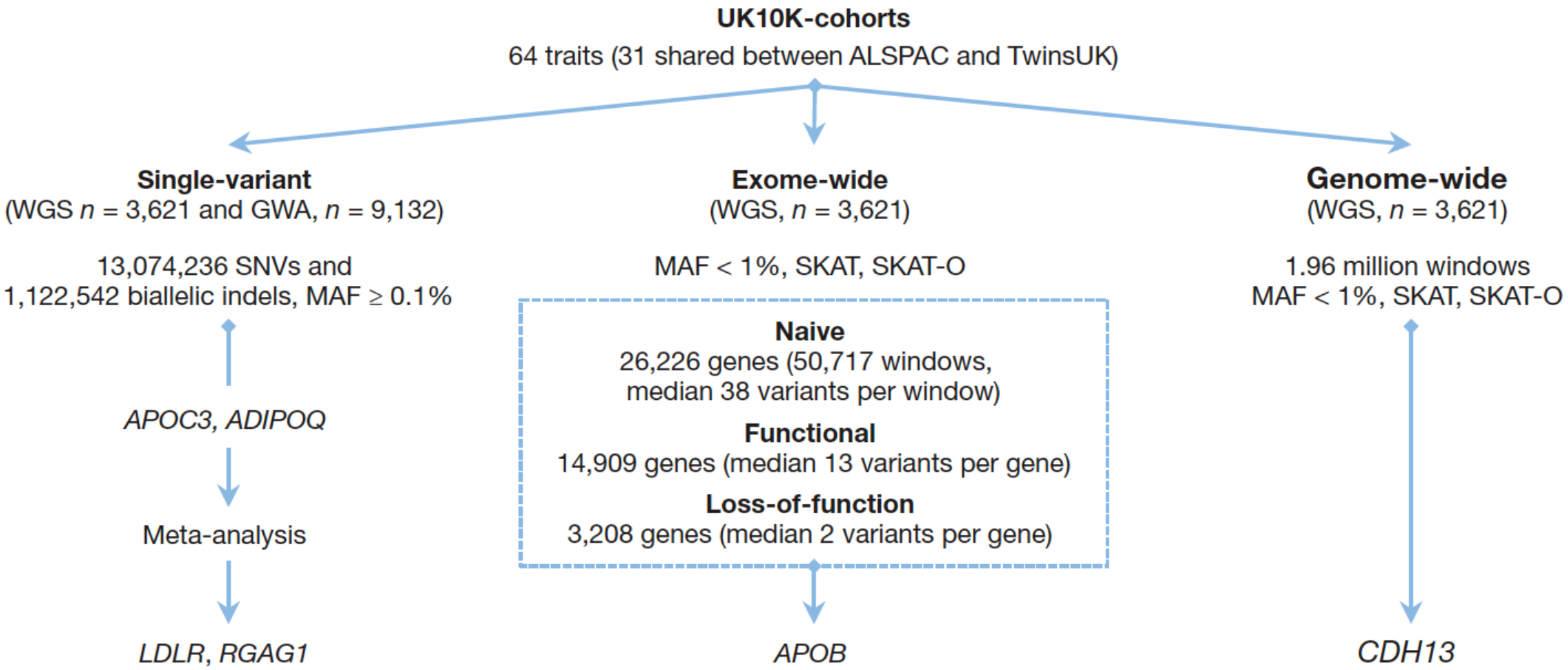
- 42 million SNPs, 3.5 million INDELs
- Common variants (allele frequency AF >5%) are mostly shared between European populations
- Most variants are rare, *i.e.* almost half of the variants are singletons (only found in one sample)

Population structure in UK10K



- Rare genetic variants showed excess allele sharing at distances smaller than about 200 km, and reduced sharing for more than about 300 km
- Fine structure analysis suggested that the identified populations were not strongly geographically defined

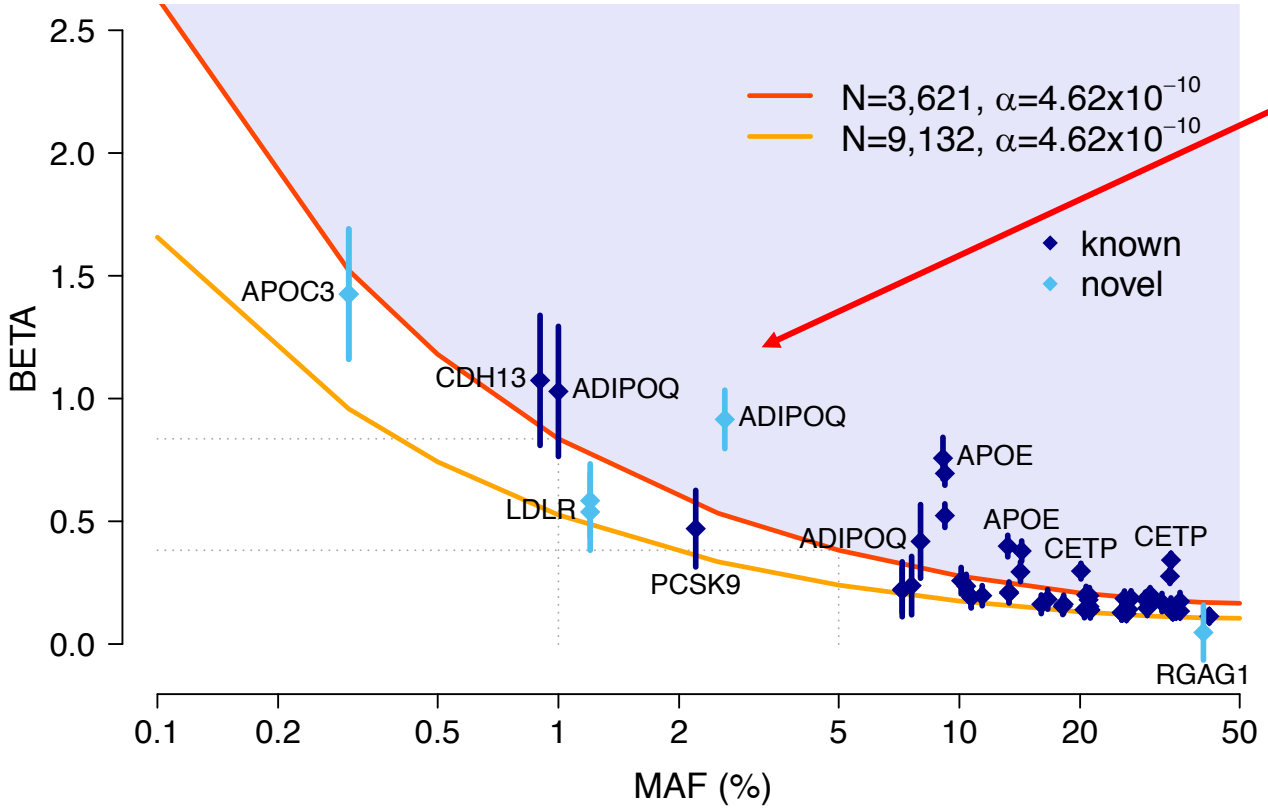
Analysis strategy in UK10K



Walter, *et al.* Nature 2015

Rare variants in complex traits

Summary of single-variant associations
(31 traits, N=3,621)



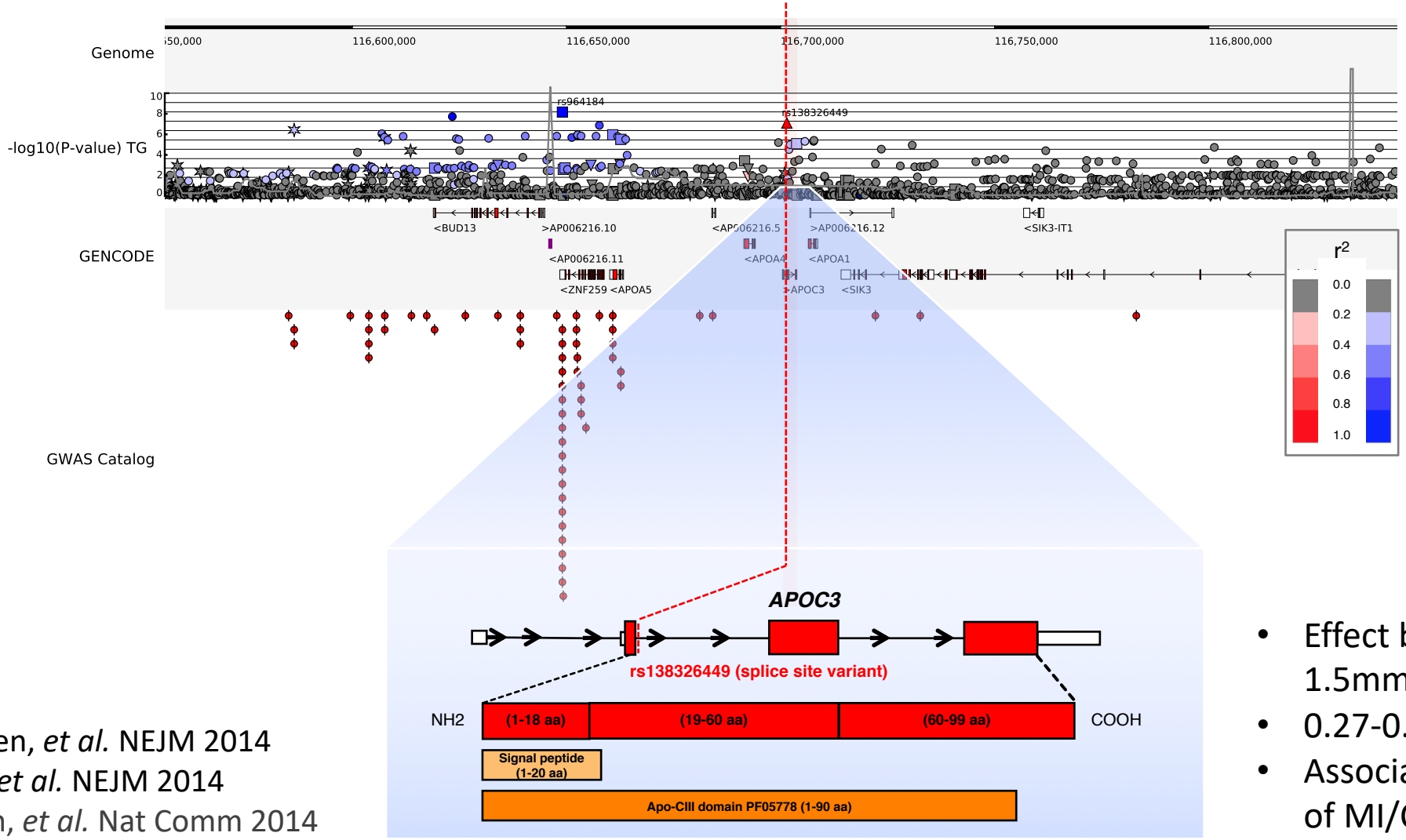
No low-frequency alleles of high impact

Main insights

- Classical lipid variants define extremes of risk for 31 traits
- Studies of thousands of participants underpowered for complex traits
- Imputation panel enables rare variants association studies

Timpson, *et al.* Nat Comm 2014
Huang, *et al.* Nat Comm 2015
Geihs, *et al.* Bioinformatics 2015
Walter, *et al.* Nature 2015
Iotchkova, *et al.* Nat Genet 2016

A rare *APOC3* splice variant has a clinically-significant effect size and is associated with heart disease



- Effect between 0.5 and 1.5mmol/L per allele copy
- 0.27-0.39% variance explained
- Associated with decreased risk of MI/CAD

Jørgensen, *et al.* NEJM 2014
Crosby, *et al.* NEJM 2014
Timpson, *et al.* Nat Comm 2014

Discovery and refinement of genetic loci associated with cardiometabolic risk using dense imputation maps

Trait	Marker	Nearest gene	Allele 1	Allele 2	MAF (WGS)	β (joint)	SE (joint)	P value (joint)	n (joint)	Signal	Annotation
PCV	rs10008637	SHROOM3	C	T	0.463	0.032	0.004	1.08 . 10−14	124,890	Primary	Intronic
PLT	rs2546979	FABP6	C	G	0.291	−0.049	0.004	1.81 . 10−31	134,858	Primary	Intergenic
WBC	rs3130725	ZNF311	G	T	0.131	−0.008	0.001	2.70 . 10−26	121,238	Primary	Intergenic
WBC	rs113164910	HLA-DRA	AAC	A	0.327	0.008	0	4.19 . 10−54	122,412	Primary	Intergenic
PLT	rs61750929	S1PR3	T	C	0.059	−0.081	0.008	2.20 . 10−21	134,858	Primary	Intergenic
PLT	rs150813342	GFI1B	T	C	0.004	−0.408	0.026	4.73 . 10−57	111,278	Primary	Synonymous
PLT	rs113373353	RASSF3	T	C	0.111	0.055	0.006	1.76 . 10−17	134,858	Primary	Intronic
PLT	rs575505283	TP53BP1	AT	A	0.014	−0.160	0.019	6.89 . 10−17	121,073	Primary	Intronic
PLT	rs1801689	APOH	C	A	0.033	0.106	0.012	3.92 . 10−19	134,858	Primary	Nonsynonymous
PLT	rs75570992	TRABD–MOV10L1	C	G	0.072	0.096	0.008	7.75 . 10−32	134,377	Primary	Intronic
PLT	rs41315846	GCSAML	C	T	0.479	0.048	0.004	3.03 . 10−34	134,858	Secondary	Intronic
PLT	rs78565404	THPO	T	C	0.057	0.136	0.009	1.65 . 10−50	134,858	Secondary	3′ UTR
Uric acid	rs56223908	SLC2A9	C	A	0.08	0.137	0.018	9.21 . 10−15	26,727	Secondary	Intronic
WBC	rs2442735	HLA-B	G	A	0.14	−0.010	0.001	1.93 . 10−46	121,528	Secondary	Intergenic
MCV	rs112233623	CCND3	T	C	0.011	0.723	0.049	5.65 . 10−49	107,036	Secondary	Intronic
HDL	rs3824477	ABCA1	A	G	0.026	0.122	0.016	1.43 . 10−13	56,306	Secondary	Intronic
MCH	rs117747069	NPRL3	C	G	0.037	−0.172	0.024	4.20 . 10−13	119,687	Secondary	Intronic

- Genotype imputation based on UK10K and 1000 Genomes Project into 35,981 samples of European ancestry
- GWA using 20 quantitative cardiometabolic and hematological traits
- 17 new associations signals
- Functional enrichment analysis (GARFIELD)
- Fine-mapping combined with regulatory information improves understanding of risk factors

Iotchkova, *et al.* Nat Gen 2016, Nat Gen 2019

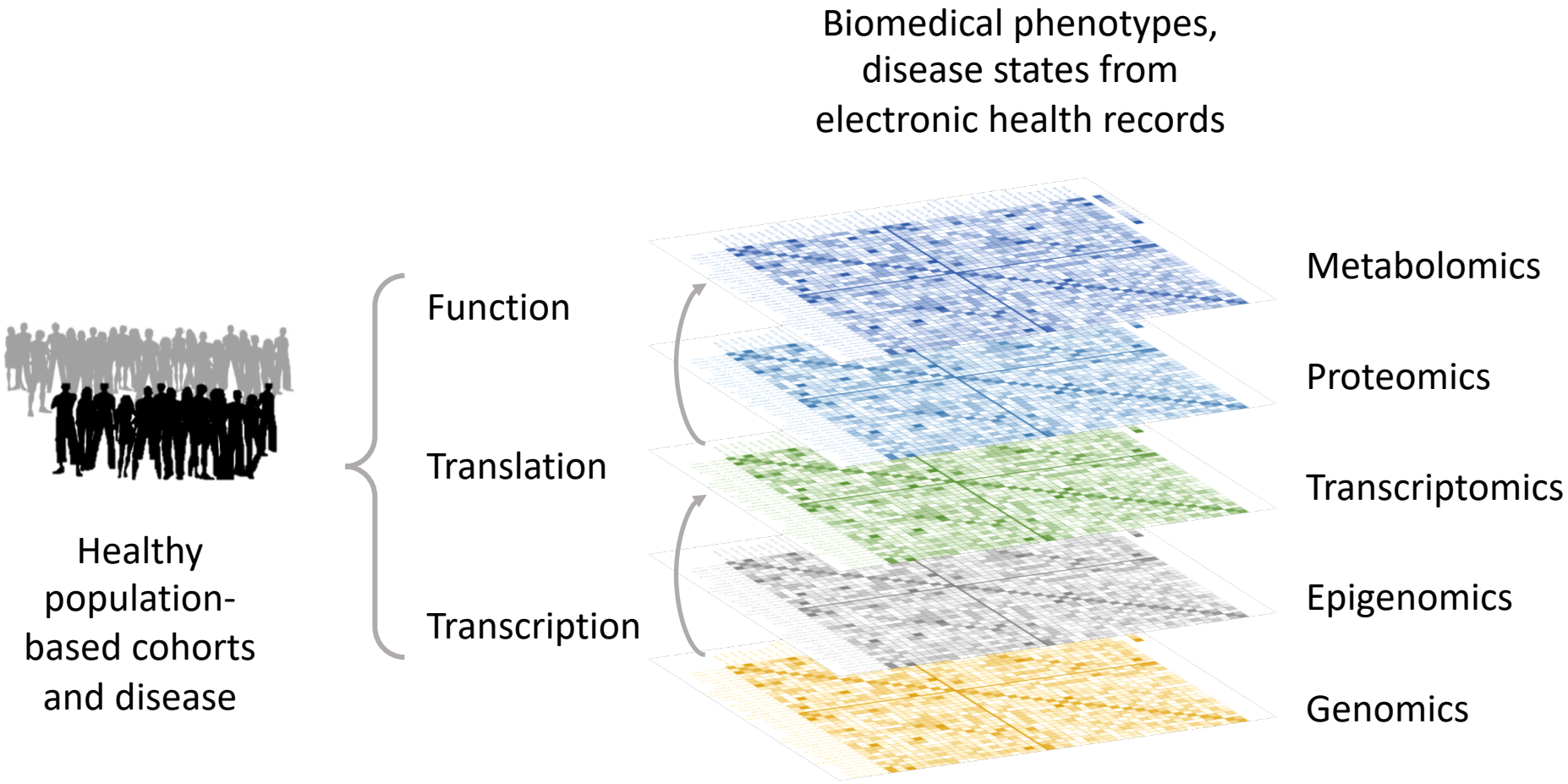
First examples of rare variant associations

Gene	Variant ID	Trait/Disease	Samples	AF (cases/controls)	Beta/OR (SE)/(CI)	Type	Study	Population
TMEM161B	rs774396010	HDL cholesterol	3621	0.001	-1.887 (0.378)	WGS	UK10K	British
APOC3	rs138326449	Triglycerides	3734	0.001	NA	WES	ESP of NHLBI	European or African
LGR4	hg18_chr11:27369242_A	BMD	95085	0.00174	-0.75 (0.16)	WGS + Imputation	deCODE	Icelandic
PDX1	hg18_chr13:27396636delT	Type 2 diabetes	278254	0.00198	2.47	WGS + Imputation	deCODE	Icelandic
APOC3	rs138326449	Triglycerides	3621	0.003	-1.425 (0.265)	WGS	UK10K	British
APOC3	rs138326449	VLDL	3621	0.003	-1.426 (0.265)	WGS	UK10K	British
FNBP1	rs528899443	FEV1/FVC	3621	0.004	1.078 (0.204)	WGS	UK10K	British
GFI1B	rs150813342	PLT	114753	0.004	-0.406 (0.026)	WGS + Imputation	UK10K and others	European
VWF	rs150077670	VWF antigen	4468	0.0044	-34.5 (12.7)	WES	ESP of NHLBI	European or African
APP	rs63750847	Alzheimer's disease	71743	0.00467	0.189	WGS + Imputation	deCODE	Icelandic
C3	rs147859257	AMD	52578	0.0055	3.13 (1.99–4.91)	WGS + Imputation	deCODE	Icelandic
ANK3	rs141471070	FEV1/FVC	3621	0.006	0.739 (0.164)	WGS	UK10K	British
TREM2	rs75932628	Alzheimer's disease	110050	0.0063	2.26	WGS + Imputation	deCODE	Icelandic
VWF	rs61750625	VWF antigen	4468	0.00763	-39.6 (9.71)	WES	ESP of NHLBI	European or African
GFI1B	rs150813342	PLT	13744	0.008	-0.402 (0.07)	WES	6 Cohort Studies	European and African American
VWF	rs149424724	VWF antigen	4468	0.008	-40.3 (10.0)	WES	ESP of NHLBI	European or African
CDH13	rs12051272	Adiponectin	3621	0.009	-1.074 (0.156)	WGS	UK10K	British
ADIPOQ	rs17366653	Adiponectin	3621	0.01	-1.029 (0.150)	WGS	UK10K	British
GLP1R	rs10305492	Fasting glucose	60564	0.01	-0.09 (0.013)	GWAS + WES	CHARGE	African and European

Bomba, Walter and Soranzo, Genome Biology 2017

INTERVAL Study

From complex disease to molecular mechanisms



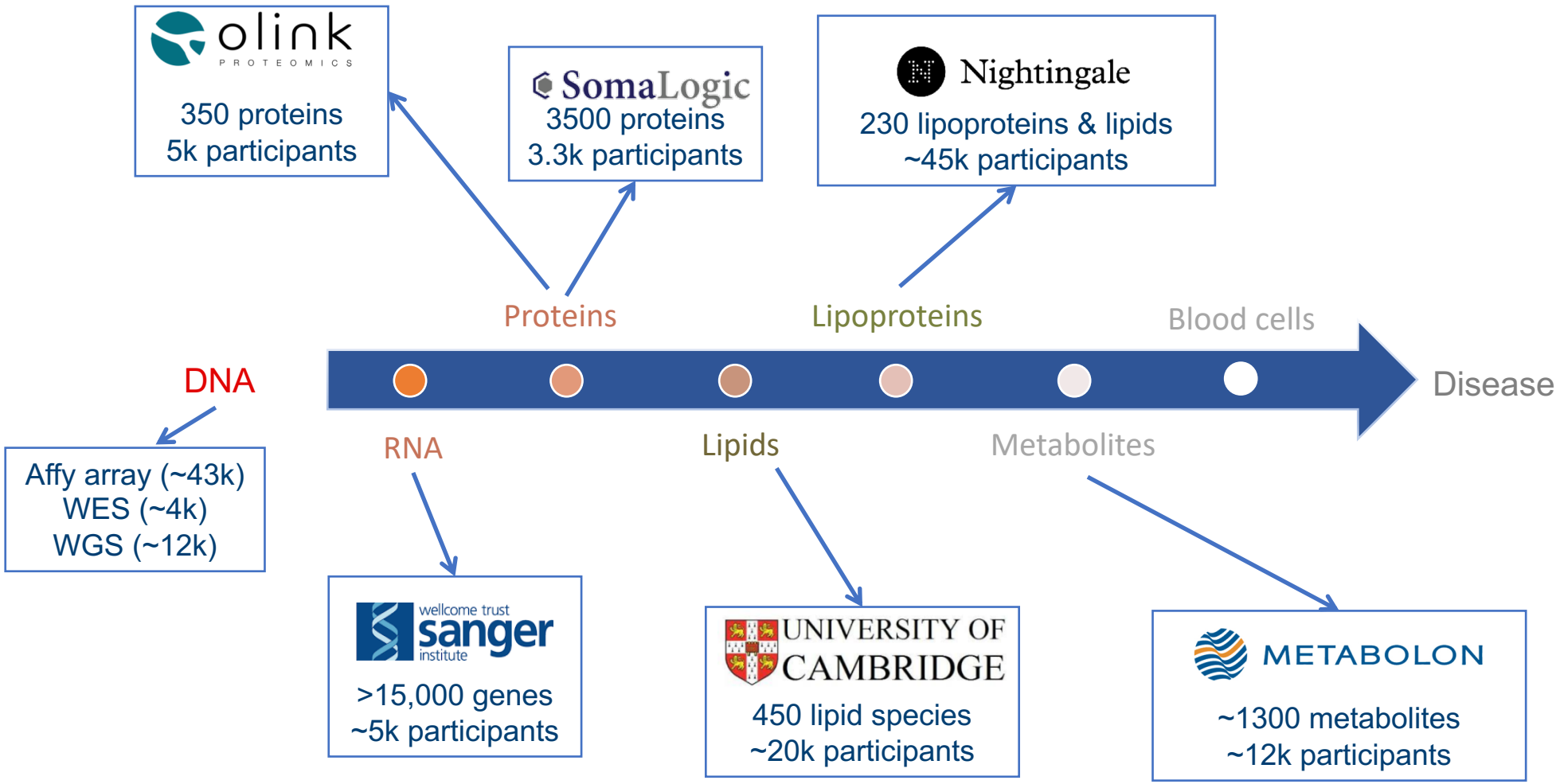
Nature 2011a, Nature 2011b, Nature 2012, Nat Genet 2014, Nature 2015, Nat Genet 2015, Cell 2016a, Cell 2016b, Nature 2018, Nat Genet 2019

INTERVAL Study

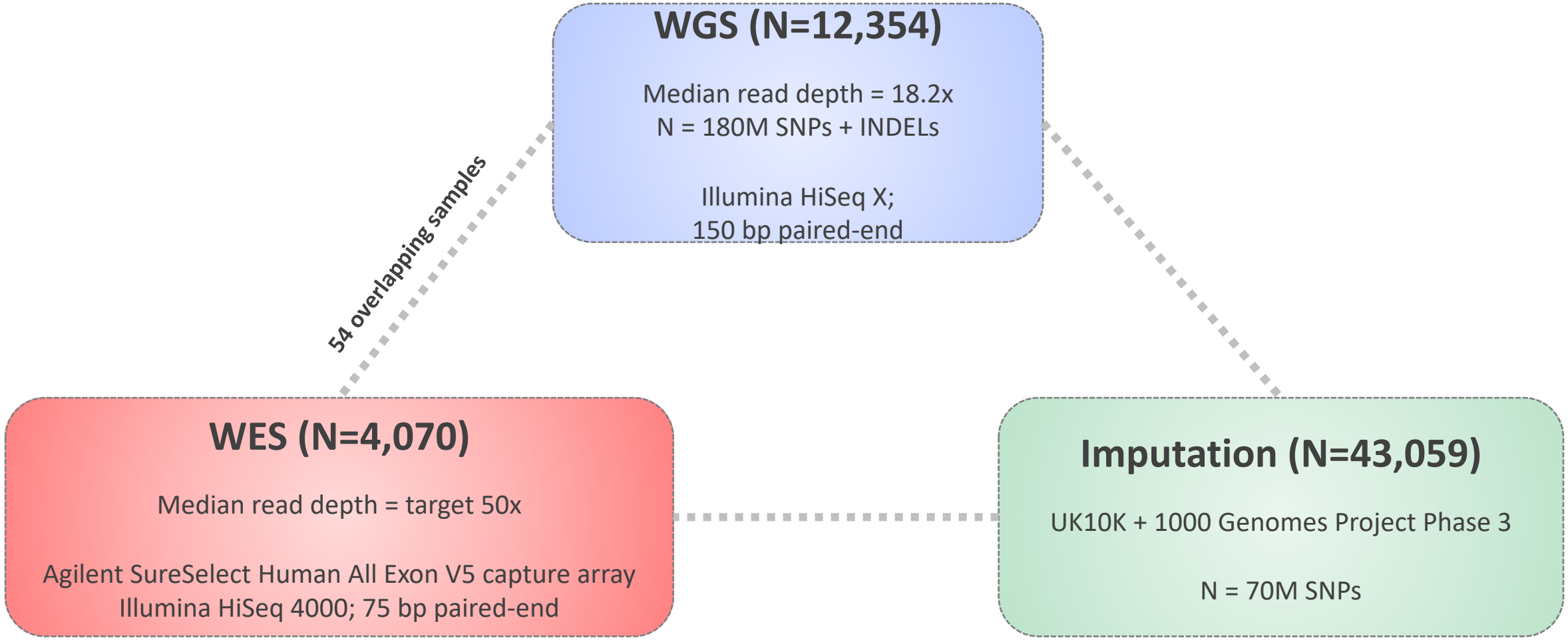
- 25,000 men and 25,000 women recruited between June 2012 and June 2014
- NHS Blood and Transplant (NSHBT) blood donation centres across England
- Men donate every 12, 10 or 8 weeks, and women every 16, 14 or 12 weeks
- Compare the amount of blood donated and measures of well-being in people at standard intervals versus more frequently
- Over a 2-year period, intervals for whole blood donation can be safely reduced to meet blood shortages
- After 4 years, donors had decreased haemoglobin concentrations and more self-reported symptoms compared with the initial 2 years of the trial
- Findings suggest that blood collection services could safely use shorter donation intervals and more intensive reminders to meet shortages, for donors who maintain adequate haemoglobin concentrations and iron stores

Moore, *et al.* Trials 2014
Di Angelantonio, *et al.* Lancet 2017
Kaptoge, *et al.* Lancet Haematol 2019

INTERVAL ‘high dimensional’ data

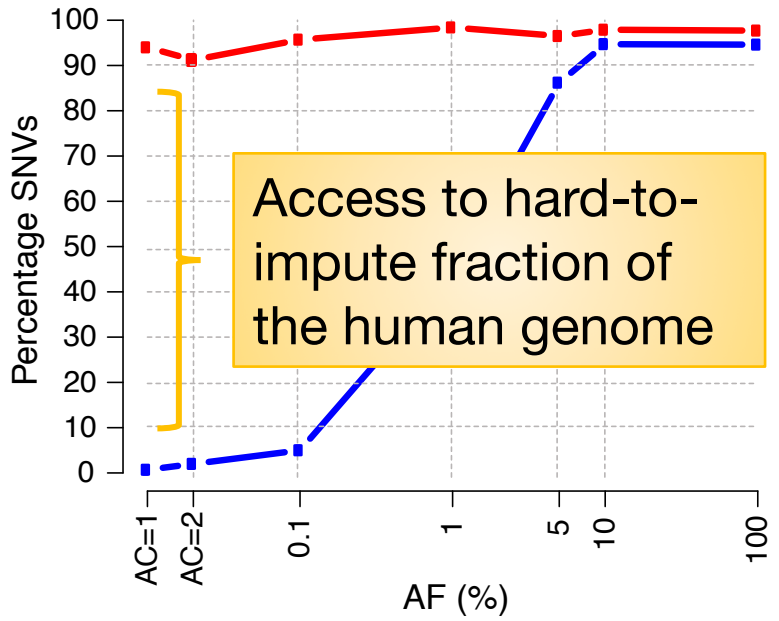


Empirical evaluation of variant sets from WES and WGS



Studies at whole-genome sequencing resolution

Variant discovery vs UK10K + 1000GP3 imputation

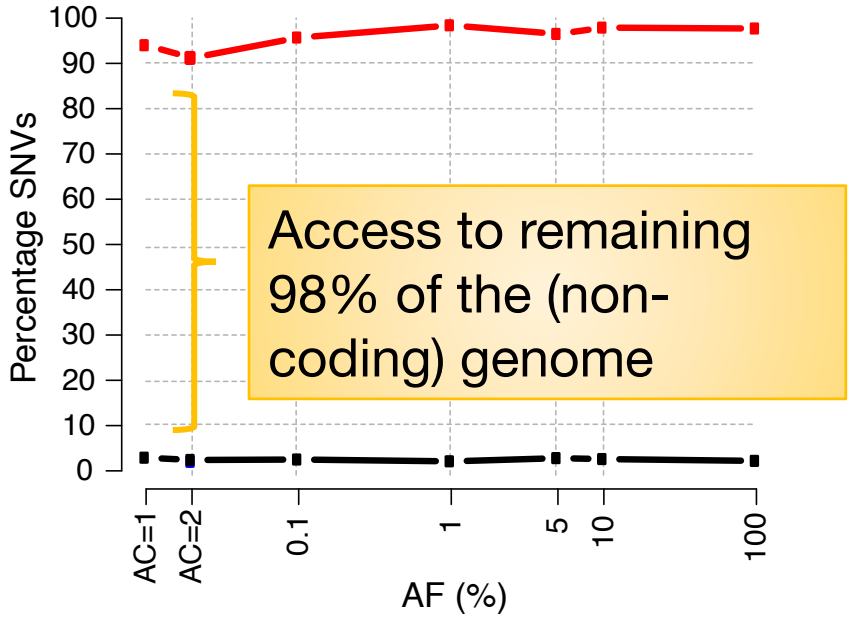


Substantial power gains at low allele frequencies

Comparison of: 50x WES 18x WGS UK10K + 1000GP3 imputation

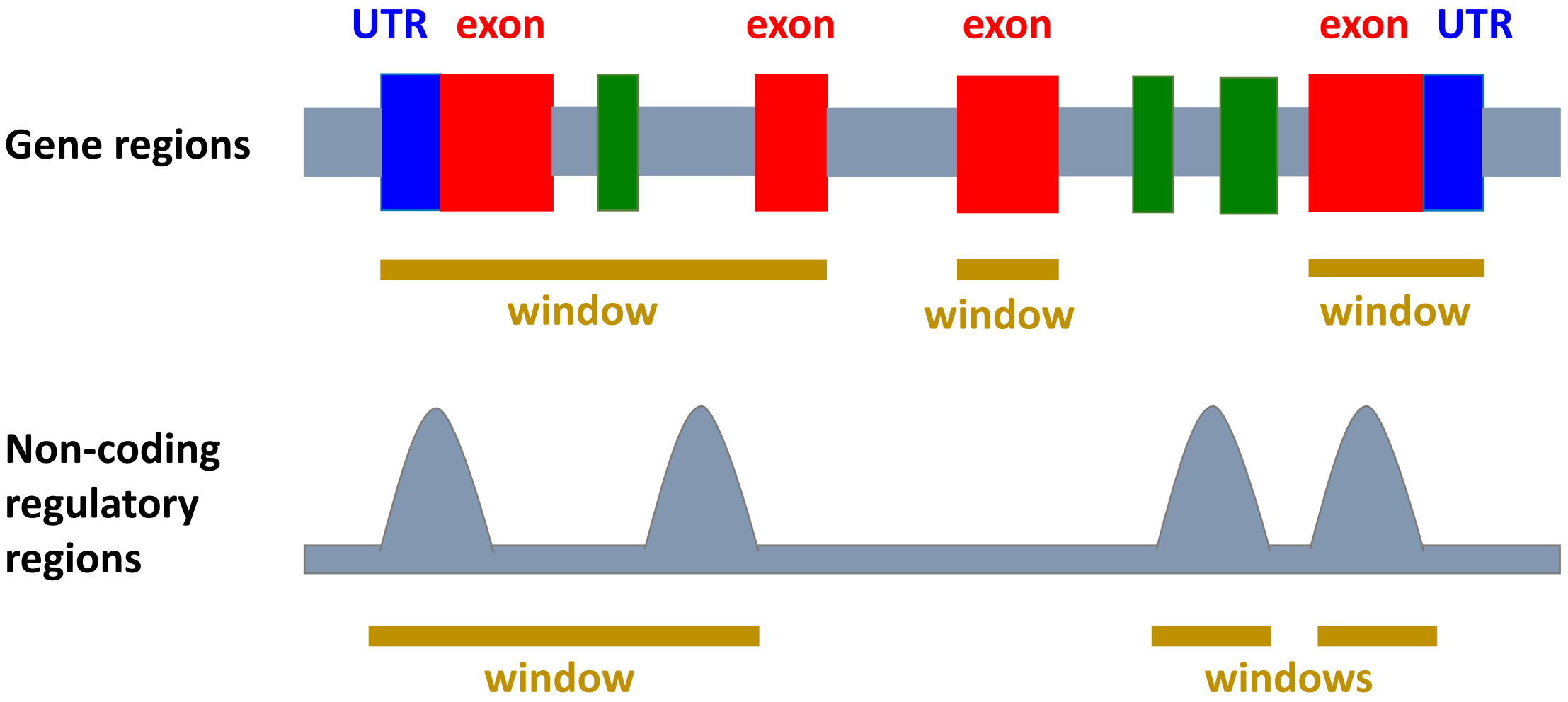
Sequence variant discovery

Variant discovery vs ‘standard’ exome capture



Comparison of: 50x WES 18x WGS UK10K + 1000GP3 imputation

Association tests aggregating rare variants over functional domains



Rare variant analysis strategy for INTERVAL WES data

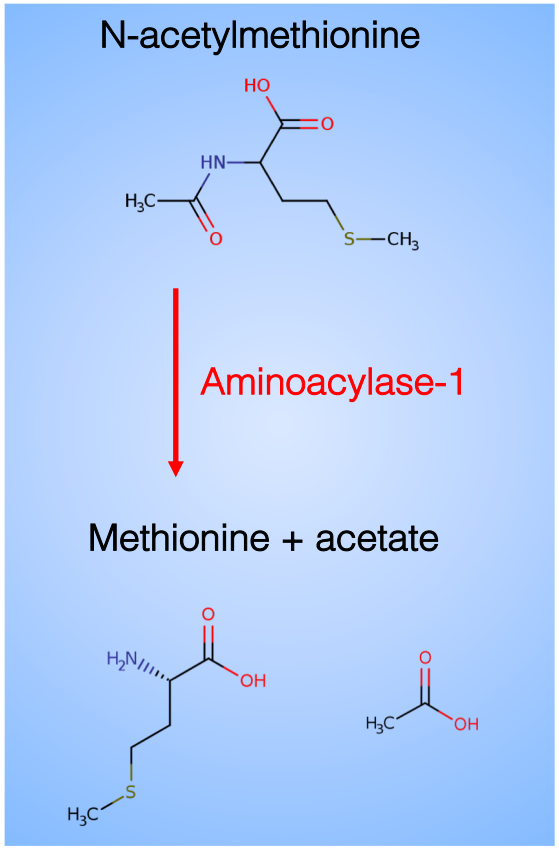
- **Variant selection**
  - MAF ≦ 0.1%
  - < 20 variants per window
  - Preserving exon structure

- **Rare variant tests**
  - Regression-based: Sequence Kernel Association Test (SKAT)
  - Burden family: Madsen and Browning (MB), Variable Threshold (VT) and Burden

- **Strategy**
  - Naïve approach (23,864 genes / 52,024 analysis windows)
  - Functional approach (20,835 genes / 32,534 analysis windows)
  - LoF approach (9,385 genes / 9,428 analysis windows)

Aminoacylase-1 (ACY1) and N-acetylmethionine

Association. Burden test of 14 rare driver variants associated with increased metabolite levels and with decreased enzyme levels (Somalogic platform)



- Implicated in the breakdown of proteins by removal of the acetyl group from certain amino acids
- Causes Aminoacylase-1 deficiency (ACY1D) is an autosomal recessive inborn error of metabolism
- Also associated with N-acetylalanine, N-formylmethionine, N-acetylvaline, N-acetylthreonine, N-acetylglutamate and N-acetylserine

- **UKBioBank PheWAS.** rs770702363 associated with heart/cardiac problems

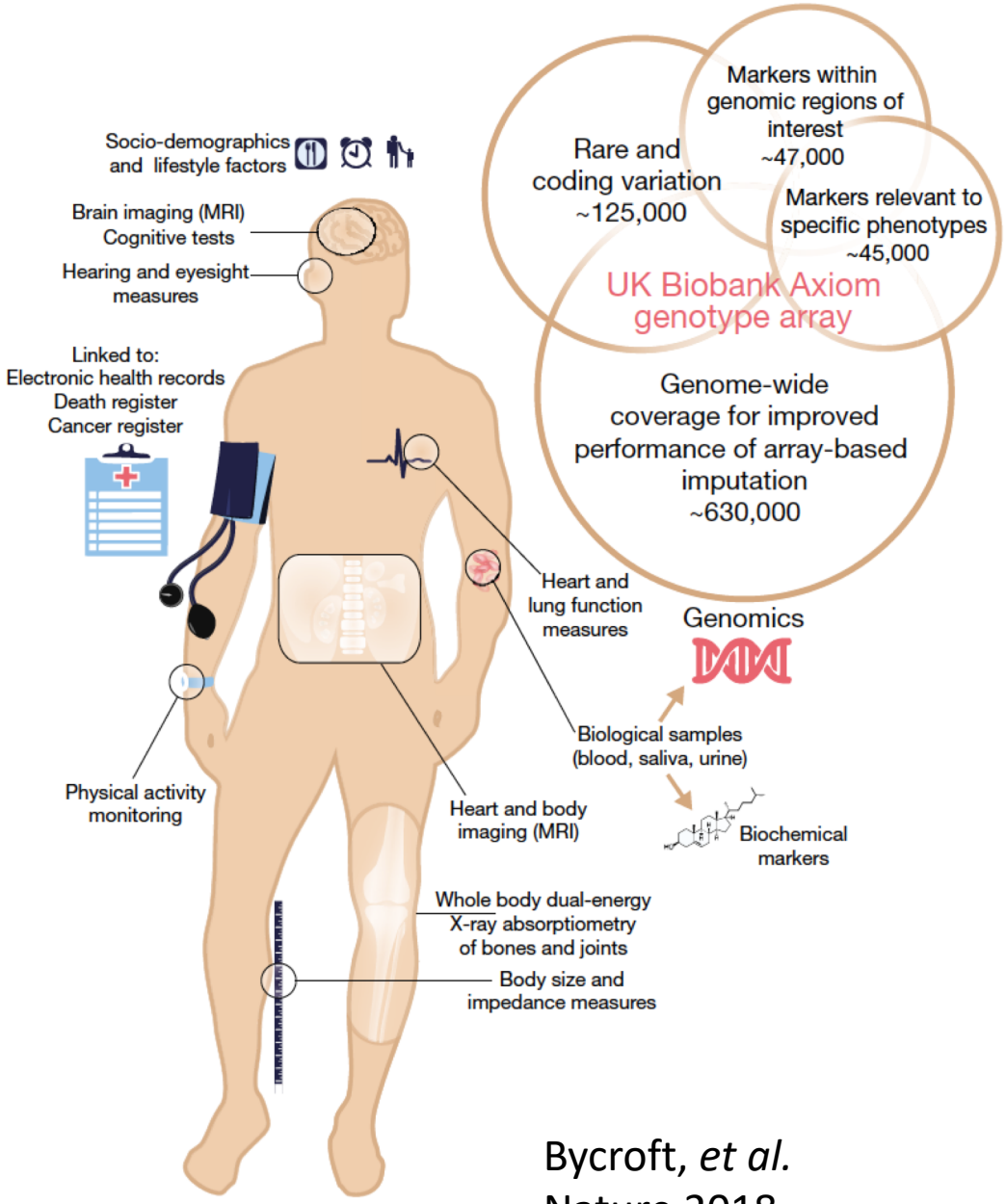
Lorenzo Bomba

UKBioBank

UKBioBank

- The UK Biobank project is a prospective cohort study with deep genetic and phenotypic data
- Recruited 500,000 people aged between 40-69 years in 2006-2010 from across the country
- Aim to improve prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses – including cancer, heart diseases, stroke, diabetes, arthritis, osteoporosis, eye disorders, depression and forms of dementia
- Participants have undergone measures, provided blood, urine, saliva, detailed information about themselves, and health follow-up

https://www.ukbiobank.ac.uk/



Bycroft, *et al.*
Nature 2018

UKBioBank

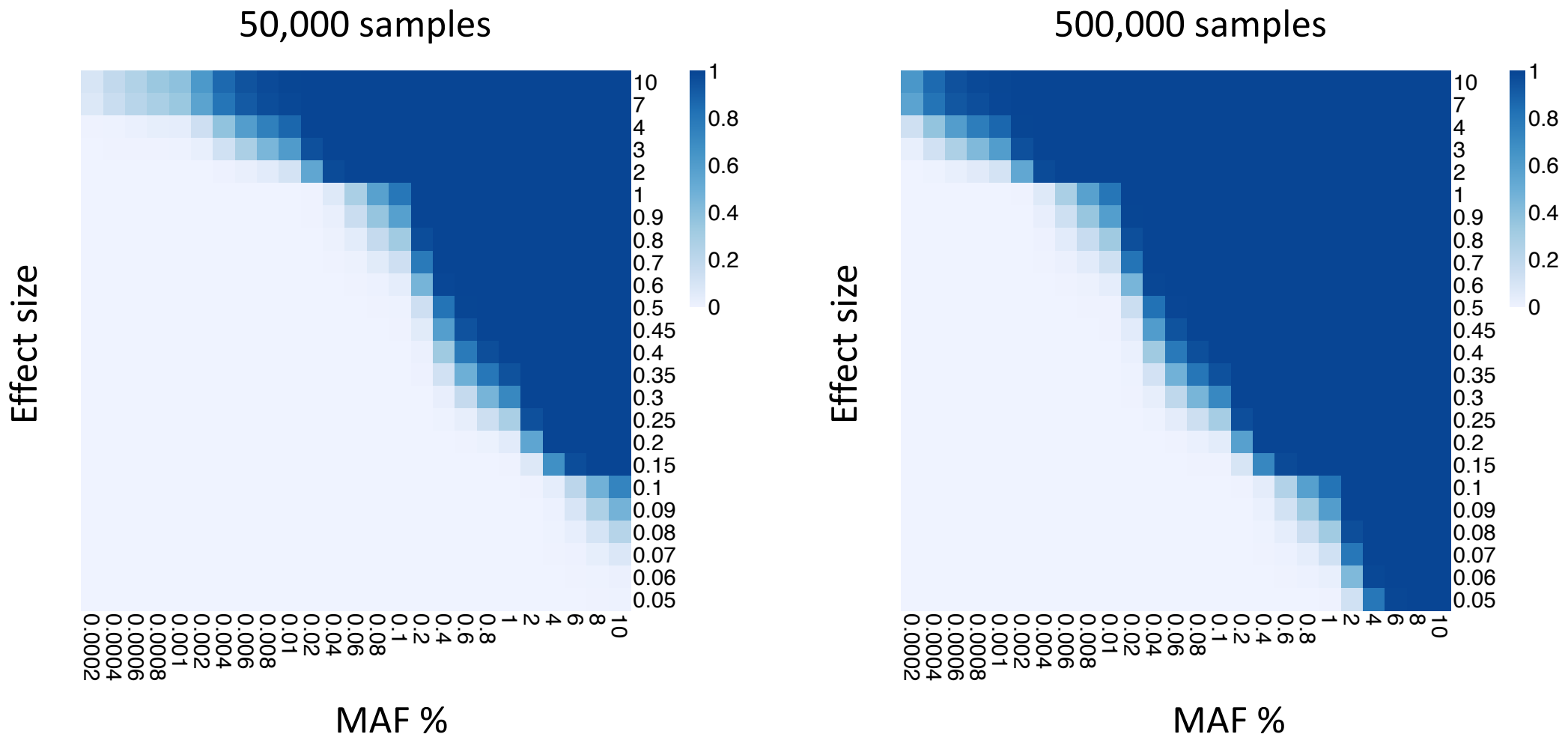
- 96 million variants after efficient phasing and genotype imputation
- **Genome-wide genotyping data** for ~850,000 variants and all 500,000 participants (Affymetrix UK BiLEVE Axiom and Affymetrix UK Biobank Axiom®)
- **Exome sequencing** of 50,000 UK Biobank participants by Regeneron already available (Van Hout, *et al.* bioRxiv 2019), and the remaining 450,000 participants by the end of 2020
- **Whole genome sequencing** of all 500,000 UK Biobank participants will be undertaken by the Wellcome Sanger Institute and deCODE genetics and the release expected by early 2023
- Since 2012 over 10,000 registrations were approved from researchers working in over 1,375 institutes in 68 countries

UK Biobank WGS: Vanguard Phase

- £30M MRC funding
- 50,000 participants
- 30x whole genome sequences, >85 Gb/genome
- Illumina short-read technology NovaSeq 6000 with S4 flow-cells
- 151-bp PE, PCR free libraries
- Sequencing start Aug 2018, at Sanger Institute
- 18 months turnaround for sequencing
- 4.5 petabases of sequence

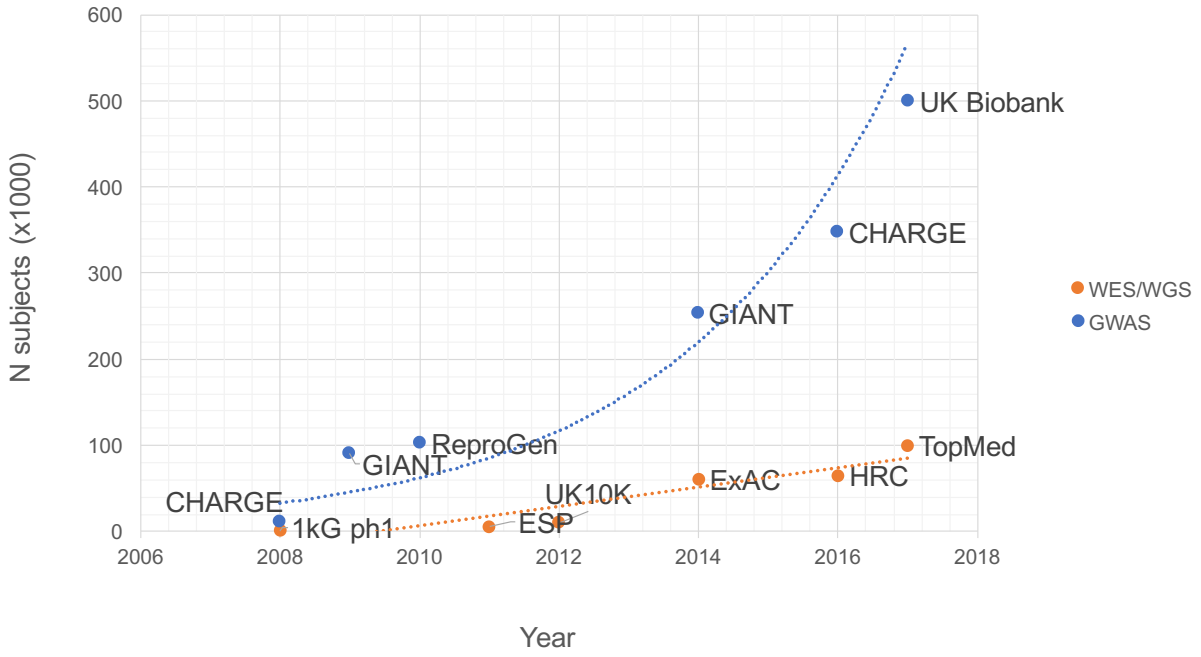
- Innovation pilots inform Phase 2

Power to discover associations by aggregated rare variants



Aleksejs Sazonovs

The future: an explosion of genomic data



Coding variation. Near-complete catalog of rare and private coding variants

Rare regulatory variation. Access to remaining 98% of the genome, integration with functional genome annotation

Genome tools. Better tools for genome assembly, graphs, genotype imputation

Structural and copy number variation

Telomere length variants

Mitochondrial DNA copy number variants

Heterochromatin structure and repeats

Pathogen genomes

Summary

- No low frequency variants with high effect sizes
- Imputation does not capture rare variants
- Rare variants tests are harder
- Increasing sample size increases power
- Most GWA signals are found in non-coding regions
- Detect and account for batch effects and other biases

Discussion

- Analysing potential population structure is useful
- Are there variants with different allele frequencies in the cohort, *i.e.* rare variants with higher AF?
- Opportunity to study the effect of region-specific environmental factors and region-specific genotypes

Acknowledgments

Wellcome Sanger Institute

Team Soranzo

Team Soranzo
Kousik Kundu

University of Cambridge
Adam Butterworth
John Danesh

Biomarin
Lorenzo Bomba

MRC WIMM
Valentina Iotchkova

Wellcome Sanger Institute
Aleksejs Sazonovs

INTERVAL Whole Genome Sequence

- **12,354** participants, 18 to 70+ years old
- **15x** whole genome sequences, >40 Gb/genome
- **180 million** SNPs + INDELs, ~8Tb of data
- Illumina short-read technology **HiSeq X**
- Sequencing done at the Sanger Institute in three phases:
  - Phase 1 - PCR libraries (N = 5,093)
  - Phase 2 - PCR free libraries (N = 5,570)
  - Phase 3 - PCR free libraries (N = 1,691)

Whole genome sequencing in the UK Biobank

- Part of the UK Government's Industrial Strategy Challenge Fund (ISCF) for the 'Data to Early Diagnosis and Precision Medicine' initiative
- Aim to produce deep characterisation of whole-genome sequences for the entire UK Biobank
- All data will be fully released to the scientific community
- 5 year landscape for sequence data generation
- Two phases:
  - Pilot phase ("Vanguard"): first 50,000 participants
  - Main phase: remaining 450,000 participants

biobank uk

Improving the health of future generations

Strategies to enrich rare variants signals

- Genomic tools for assessing low-frequency and rare variants
  - Imputation
  - Custom genotyping arrays
  - Exome or whole genome sequencing
- Optimal methods for association analysis with low frequency and rare variants
  - Burden tests
  - Variance-component tests
  - Combined tests
- Study designs for enriching or prioritising rare variants
  - Population isolates
  - Loss-of-function
- Power, replication and confounding affecting rare variant association tests