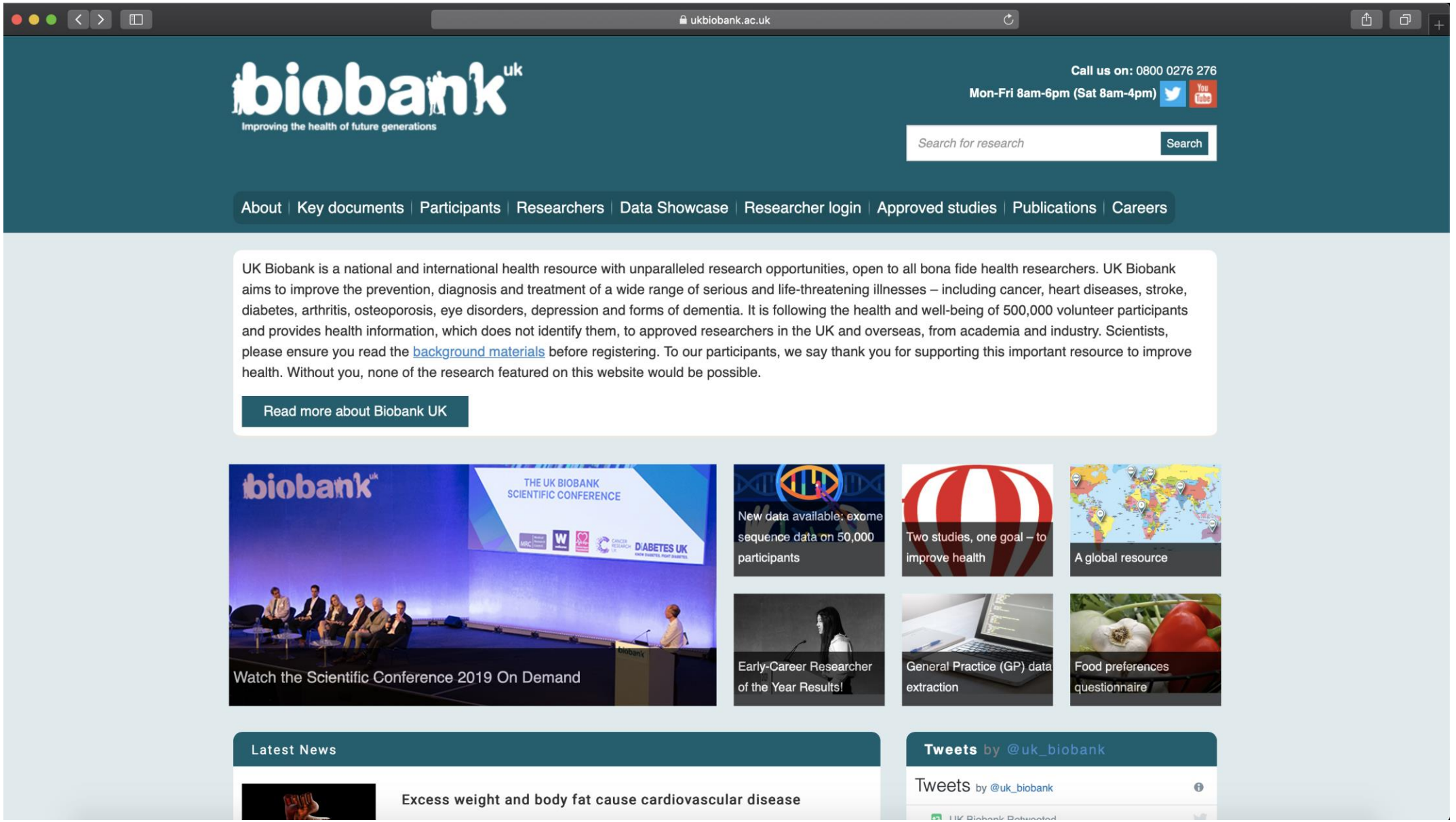
Update from UK Biobank Scientific Conference 2019 & An Introduction on SAIGE

Xiao Jiang, PhD

Biostatistician
Prof. Maciej Tomaszewski group
Division of Cardiovascular Sciences
School of Medical Sciences
The University of Manchester

MANCHESTER 1824
The University of Manchester

biobank uk
Improving the health of future generations



Regularly updated information on a wide range of diseases from NHS datasets in all three countries:

Regularly updated information on a wide range of diseases from NHS datasets in all three countries:

- Death: cause and date of death
  all participants: >20,000 cause-specific deaths
- Cancers: site, stage, grade and date of cancer
  all participants: >120,000 site-specific cancer cases
- Hospital discharges: diagnosis, procedure and date
  all participants: 1000's of disease cases
- Primary care data: diagnosis, prescription, laboratory
  only HALF of the participants: 1000's more cases

Regularly updated information on a wide range of diseases from NHS datasets in all three countries:

- Death: cause and date of death
  all participants: >20,000 cause-specific deaths
- Cancers: site, stage, grade and date of cancer
  all participants: >120,000 site-specific cancer cases
- Hospital discharges: diagnosis, procedure and date
  all participants: 1000's of disease cases
- Primary care data: diagnosis, prescription, laboratory
  only HALF of the participants: 1000's more cases
  - Wales (~95%): Data obtained for 21,000 participants from EMIS and Vision via SAIL
  - Scotland (~75%): Data obtained for 27,000 participants from EMIS and Vision via Albasoft
  - England (~40%): Data obtained for 167,000 and 19,000 participants from TPP and from Vision via Apollo

Primary care

"Primary care includes general practice, community pharmacy, dental, and optometry (eye health) services. This data gives valuable insight into how services are provided and is used to plan, deliver and monitor services."

- NHS Digital

Primary care

"Primary care includes general practice, community pharmacy, dental, and optometry (eye health) services. This data gives valuable insight into how services are provided and is used to plan, deliver and monitor services."

- NHS Digital

- Primary care data available time:
  - An interim release of primary care records will be available in early Q3 (covering ~234,000 participants)
    - >120 million event records
    - >50 million prescribed items

Country	Data Provider	Number of participant records	Coding System
Wales	SAIL	21,000	Read v2
Scotland	Albasoft	27,000	Read v2, BNF
England	TPP	167,000	CTV3 (Read v3)
	Vision	19,000	Read v2, DM+D
	EMIS	-	Read v2, DM+D
	Total	**234,000**	

Primary care

"Primary care includes general practice, community pharmacy, dental, and optometry (eye health) services. This data gives valuable insight into how services are provided and is used to plan, deliver and monitor services."

- NHS Digital

- Primary care data available time:
  - An interim release of primary care records will be available in early Q3 (covering ~234,000 participants)
    - >120 million event records
    - >50 million prescribed items

Country	Data Provider	Number of participant records	Coding System
Wales	SAIL	21,000	Read v2
Scotland	Albasoft	27,000	Read v2, BNF
England	TPP	167,000	CTV3 (Read v3)
	Vision	19,000	Read v2, DM+D
	EMIS	-	Read v2, DM+D
	Total	**234,000**	

- Will contain coded data only e.g.
  - Clinical events
  - Prescriptions
  - Lab test results
  - Immunisation records
- Data will be provided in their native coding system
- Only limited data curation applied

- How will primary care data be made available?
  - This will be an interim release prior to completing linkage across the cohort; subsequent releases are likely during 2020. (plan provide refreshes of linked data on an annual basis starting from 2020)
  - Record level data will be available via the **online Data Portal** (similar to the existing Hospital Episodes Statistics data)
  - The UK Biobank intent to provide algorithmically derived outcome fields for a large number of diseases (currently >1,000 separate codes).

- How will primary care data be made available?
  - This will be an interim release prior to completing linkage across the cohort; subsequent releases are likely during 2020. (plan provide refreshes of linked data on an annual basis starting from 2020)
  - Record level data will be available via the **online Data Portal** (similar to the existing Hospital Episodes Statistics data)
  - The UK Biobank intent to provide algorithmically derived outcome fields for a large number of diseases (currently >1,000 separate codes).
  - Timelines:
    - Record-level data available early Q3 2019; and
    - Derived outcomes either coincident with the interim release, or shortly thereafter

Access to UK Biobank (UKB) data

- Access is described in their **Access Policy** and is managed via their **Access Management System (AMS):**
  - **Registration**: The UKB will check the identity of each researcher to confirm the applicant's bona fides as part of their approval
  - **Application**: The UKB assess whether the application meets the criteria of being for health-related science that is in the public interest
  - **MTA and Payment**: there is a small cost-recovery access fee and researchers must enter into a Material Transfer Agreement (MTA) before data (and samples) are released.

Access to UK Biobank (UKB) data

- Access is described in their **Access Policy** and is managed via their **Access Management System (AMS):**
  - **Registration**: The UKB will check the identity of each researcher to confirm the applicant's bona fides as part of their approval
  - **Application**: The UKB assess whether the application meets the criteria of being for health-related science that is in the public interest
  - **MTA and Payment**: there is a small cost-recovery access fee and researchers must enter into a Material Transfer Agreement (MTA) before data (and samples) are released.

- On average, how long does it take from application to access data now?

Access to UK Biobank (UKB) data

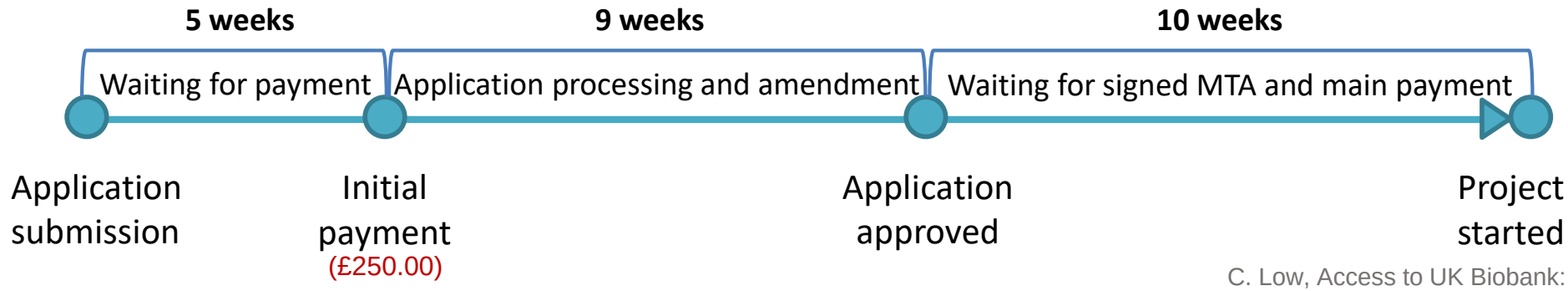
- Access is described in their **Access Policy** and is managed via their **Access Management System (AMS):**
  - **Registration**: The UKB will check the identity of each researcher to confirm the applicant's bona fides as part of their approval
  - **Application**: The UKB assess whether the application meets the criteria of being for health-related science that is in the public interest
  - **MTA and Payment**: there is a small cost-recovery access fee and researchers must enter into a Material Transfer Agreement (MTA) before data (and samples) are released.

- On average, how long does it take from application to access data now? 24 weeks

Access to UK Biobank (UKB) data

- Access is described in their **Access Policy** and is managed via their **Access Management System (AMS):**
  - **Registration**: The UKB will check the identity of each researcher to confirm the applicant's bona fides as part of their approval
  - **Application**: The UKB assess whether the application meets the criteria of being for health-related science that is in the public interest
  - **MTA and Payment**: there is a small cost-recovery access fee and researchers must enter into a Material Transfer Agreement (MTA) before data (and samples) are released.

- On average, how long does it take from application to access data now? 24 weeks

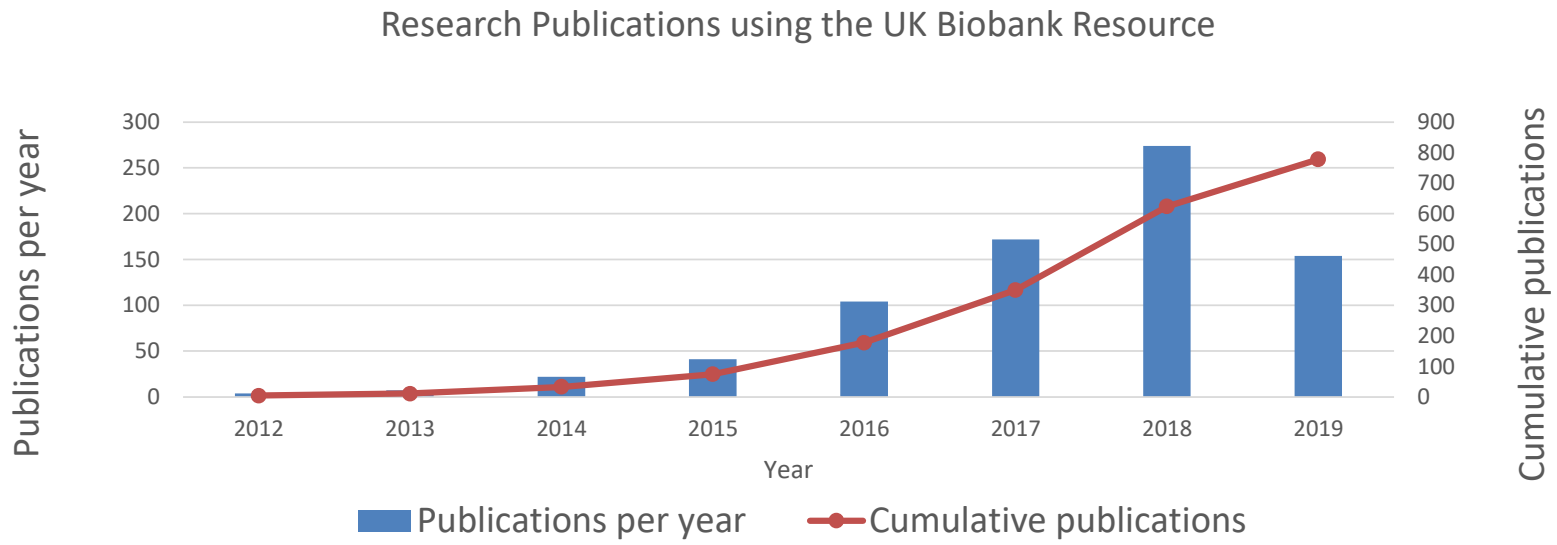


C. Low, Access to UK Biobank: Streamlining the Process

- Upcoming process improvements that the UKB are making:
  - Removal of initial payment (combine with main payment) – early Q3 2019
  - Introduction of a standardised basket of data fields selection (to reduce the application amendment time) – early Q3 2019
  - Improved signposting and help/support - creating tutorial videos to guide the stages of the access process including signing the returning MTAs – by end of Q4 2019
- UKB's goal: 12 weeks turnaround time in 2020

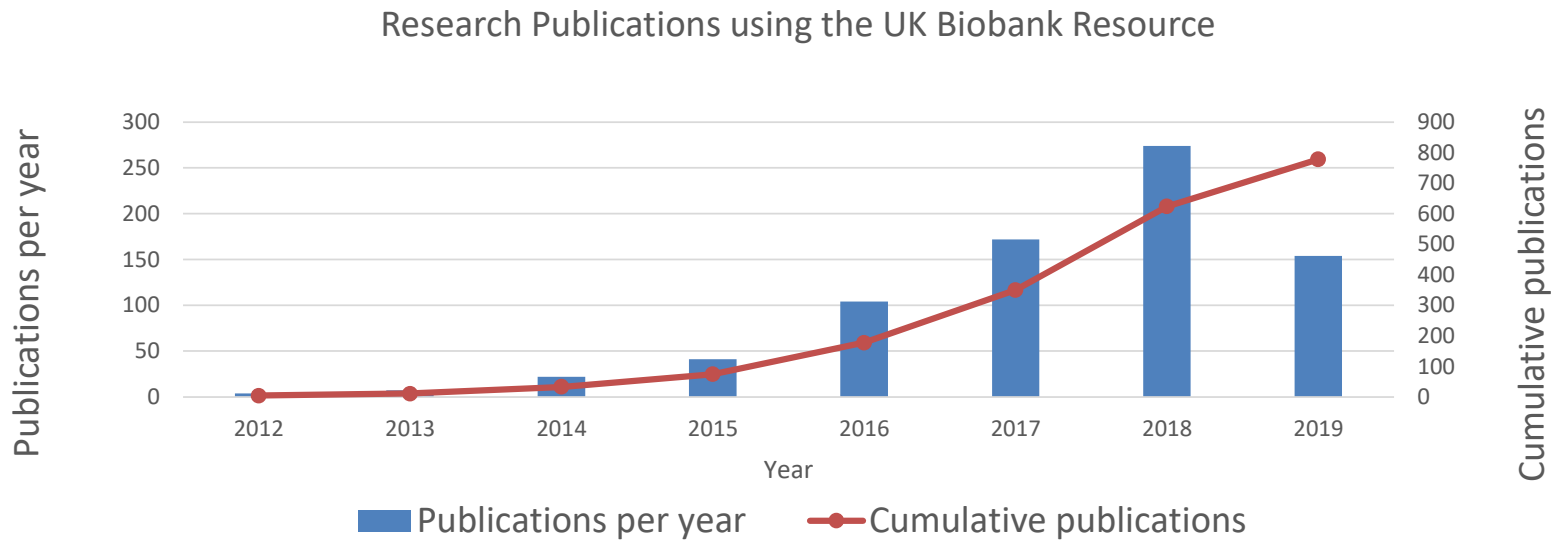
- Upcoming process improvements that the UKB are making:
  - Removal of initial payment (combine with main payment) – early Q3 2019
  - Introduction of a standardized basket of data fields selection (to reduce the application amendment time) – early Q3 2019
  - Improved signposting and help/support - creating tutorial videos to guide the stages of the access process including signing the returning MTAs – by end of Q4 2019
- UKB's goal: 12 weeks turnaround time in 2020
- UKB's suggestions for us to speed up:
  - Pay the fees online rather than make payment through bank
  - Involve the legal team early
  - Be responsible to queries from the Access team
  - Sign and return the MTA electronically

Over 750 published research papers using the UK Biobank resource



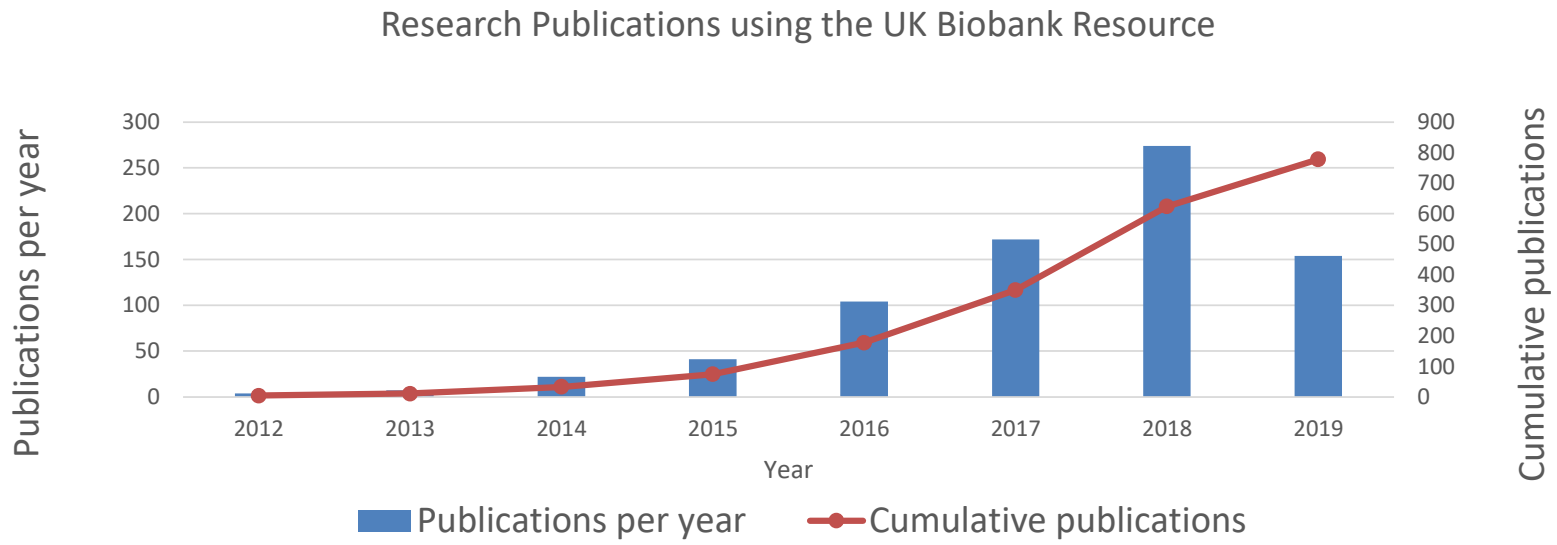
- The UKB ask users to report all papers **2 weeks ahead** of publication for the UKB to ensure that the scope of the application using the UKB resources in line with the content of the research paper.

Over 750 published research papers using the UK Biobank resource



- The UKB ask users to report all papers **2 weeks ahead** of publication for the UKB to ensure that the scope of the application using the UKB resources in line with the content of the research paper.
- Presentation of research at conferences should be reported along with a copy of abstract.

Over 750 published research papers using the UK Biobank resource



- The UKB ask users to report all papers **2 weeks ahead** of publication for the UKB to ensure that the scope of the application using the UKB resources in line with the content of the research paper.
- Presentation of research at conferences should be reported along with a copy of abstract.
- Researchers are obliged to return results to the UKB **within 6 months** of publication so that they can incorporate the data into the resource and make them available to others, also with the purpose of growing the UKB data resource.

MANCHESTER 1824
The University of Manchester

biobank uk
Improving the health of future generations

J. Bell

Highlights of the Life Sciences Strategy

- Strengthen Science base, support higher risk science
- Enhance Clinical trials and translational science
- Facilitate the scaling of innovative companies
- Establish capacity for discovery and manufacturing new generation of therapeutics (cells, viral vectors, nucleic acid based therapy)
- Resolve the adoption problem in the NHS
- Create three new industries in Life Sciences in the UK

MANCHESTER 1824
The University of Manchester

biobank uk
Improving the health of future generations

J. Bell

‘ establish the basis for three new Life Science Industries in the UK’

GENOMICS:
SequencingTechnology
Precision medicine
Target Discovery
Genotype ; Phenotype
Public Health

Investment £600m

DIGITAL HEALTH:
Large Scale clinical datasets optimised for applications in clinical trials RWD, AI and alogorithm for drug discovery, improved diagnostics and increasing efficiency of health care

Investments: £2billion for further enabling health data sets

EARLY DIAGNOSIS
New approach to medicine enabled by new technology and completely new care pathways. Relevant to Diagnostics, Digital and Pharma. Can only be developed in a single payer environment. Enabled by range of new technologies

Investments: £300m to create infrastructure.

MANCHESTER 1824
The University of Manchester

biobank uk
Improving the health of future generations

RELEASED

- Biomarkers that have been measured in all 500,000 participants (in blood unless indicated otherwise).

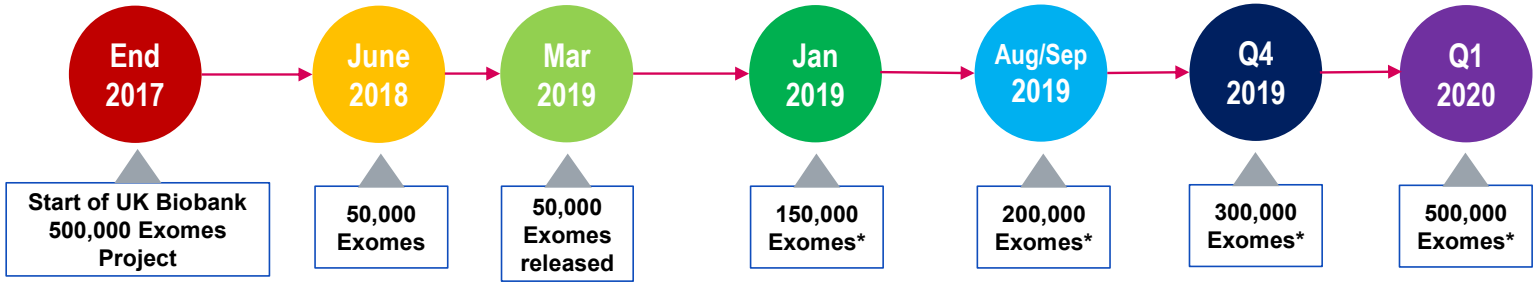
Cardio-metabolic	Bones and joints	Cancer	Renal	Liver
Cholesterol	Vitamin D	SHBG	Cystatin C	Albumin
Direct LDL-c	Rh. factor	Testosterone	Total Protein	Direct Bilirubin
HDL-c	ALP	Oestradiol	Urea	Total Bilirubin
Triglyceride	Calcium	IGF-1	Phosphate	GGT
ApoA			Urate	ALT
ApoB			Urinary:	AST
CRP			Albumin	
Lp(a)			Creatinine	
HbA1c			Potassium	
Glucose			Sodium	

Haematological assays were conducted during the recruitment phase, and those assay data have already been made available to researchers

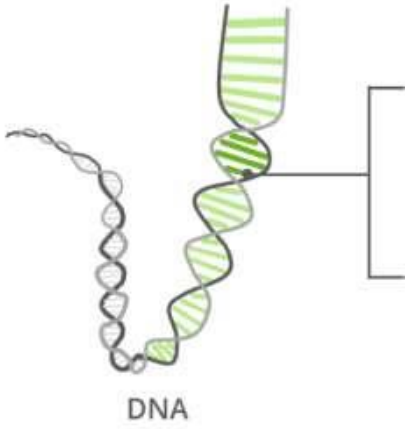
R. Collins, Opening talk

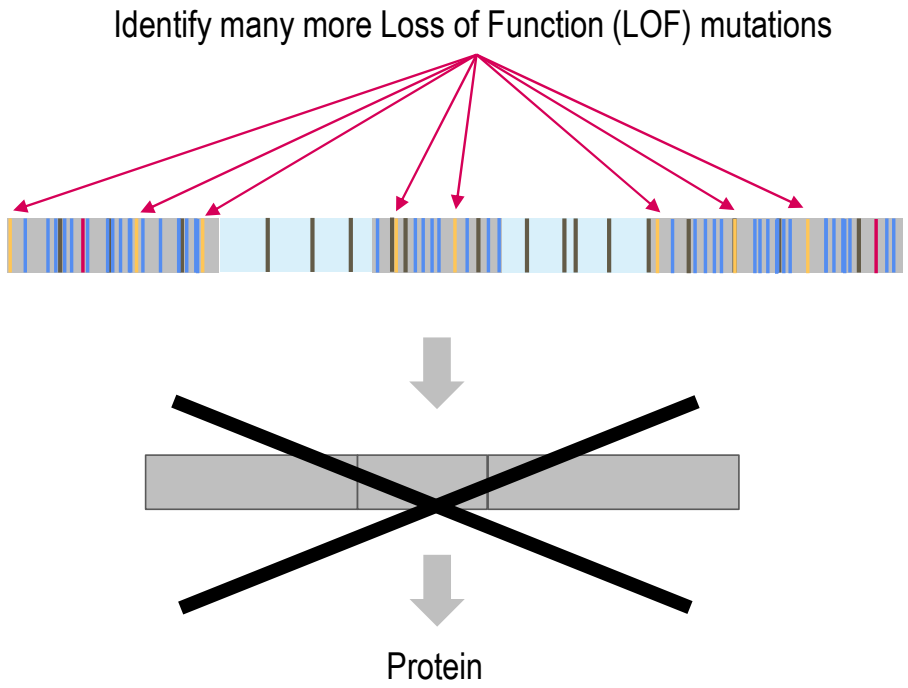
J. Marchini, Regeneron

- UKB 500,000 Exomes Project: Timelines

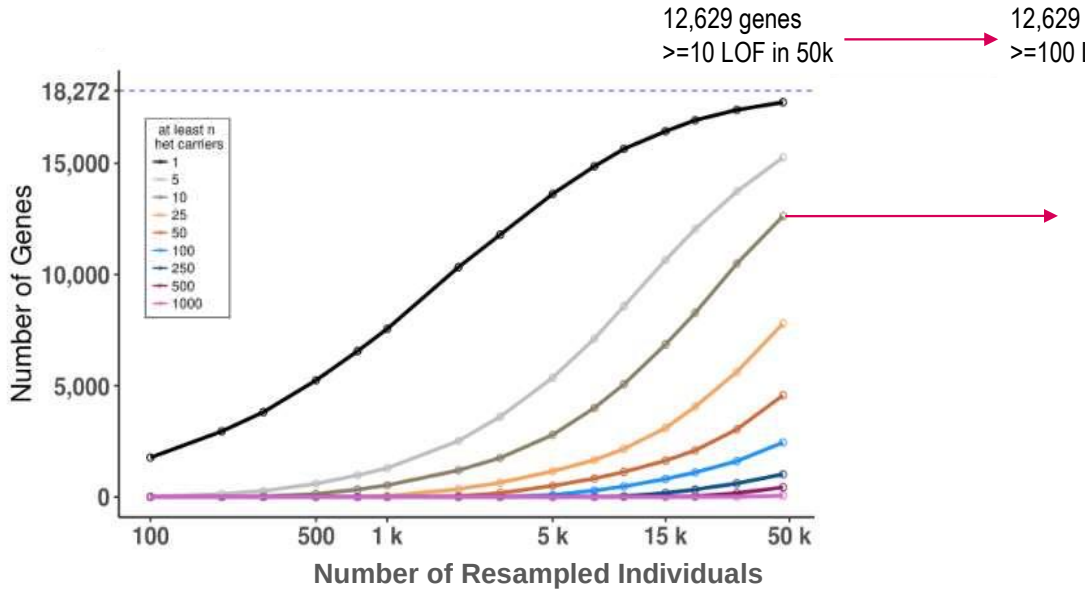


* Release of data via UKB will occur 12 months after listed dates





J. Marchini, Regeneron



LOF definition
- SNPs + indels
- stop_gained, start_lost, splice_donor, splice_acceptor, stop_lost and frameshift
- count losses in any transcript
- MAF < 1%

LOFs

~12 fold increase Exome vs Imputed

	AAF	WES	Imputed 50k	Both
LOF	All	235,915	19,451	12,488
	<1%	234,716	18,162	11,639
	1%	1,199	1,289	849

Missense variants

~6 fold increase Exome vs Imputed

	AAF	WES	Imputed 50k	Both
Missense	All	2,518,075	420,194	317,623
	<1%	2,485,710	389,343	288,599
	1%	32,365	30,851	29,024

Only 13.7% of Exome variants exist in the UKB imputed data

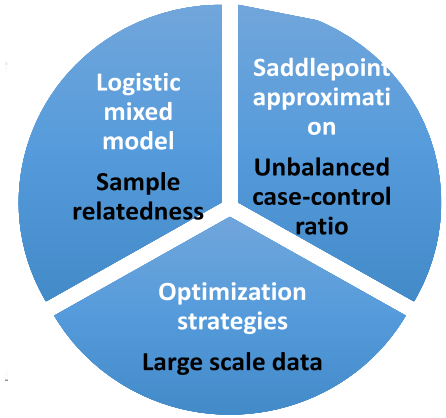
Scalable and Accurate Implementation of Generalised mixed model (SAIGE)

nature genetics
ANALYSIS
https://doi.org/10.1038/s41588-018-0184-y

Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies

Wei Zhou[1,2], Jonas B. Nielsen[3], Lars G. Fritsche[2,4,5], Rounak Dey[2,5], Maiken E. Gabrielsen[4], Brooke N. Wolford[1,2], Jonathon LeFaive[2,5], Peter VandeHaar[2,5], Sarah A. Gagliano[2,5], Aliya Gifford[6], Lisa A. Bastarache[6], Wei-Qi Wei[6], Joshua C. Denny[6,7], Maoxuan Lin[3], Kristian Hveem[4,8], Hyun Min Kang[2,5], Goncalo R. Abecasis[2,5], Cristen J. Willer[1,3,9,10*] and Seunggeun Lee[2,5,10*]

In genome-wide association studies (GWAS) for thousands of phenotypes in large biobanks, most binary traits have substantially fewer cases than controls. Both of the widely used approaches, the linear mixed model and the recently proposed logistic mixed model, perform poorly; they produce large type I error rates when used to analyze unbalanced case-control phenotypes. Here we propose a scalable and accurate generalized mixed model association test that uses the saddlepoint approximation to calibrate the distribution of score test statistics. This method, SAIGE (Scalable and Accurate Implementation of GEneralized mixed model), provides accurate *P* values even when case-control ratios are extremely unbalanced. SAIGE uses state-of-art optimization strategies to reduce computational costs; hence, it is applicable to GWAS for thousands of phenotypes by large biobanks. Through the analysis of UK Biobank data of 408,961 samples from white British participants with European ancestry for > 1,400 binary phenotypes, we show that SAIGE can efficiently analyze large sample data, controlling for unbalanced case-control ratios and sample relatedness.



- To adjust for sample relatedness, the leave-one-chromosome-out (LOCO) scheme is used.

Zhou et al., Nature Genetics, 2018; 50: 1335 - 1341

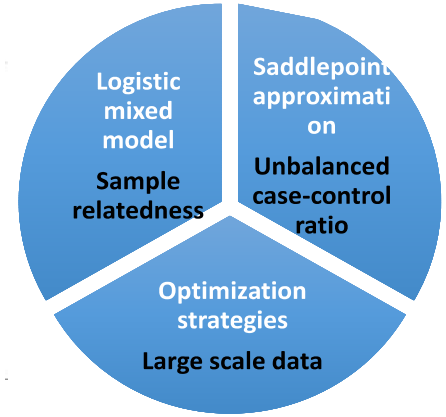
Scalable and Accurate Implementation of Generalised mixed model (SAIGE)

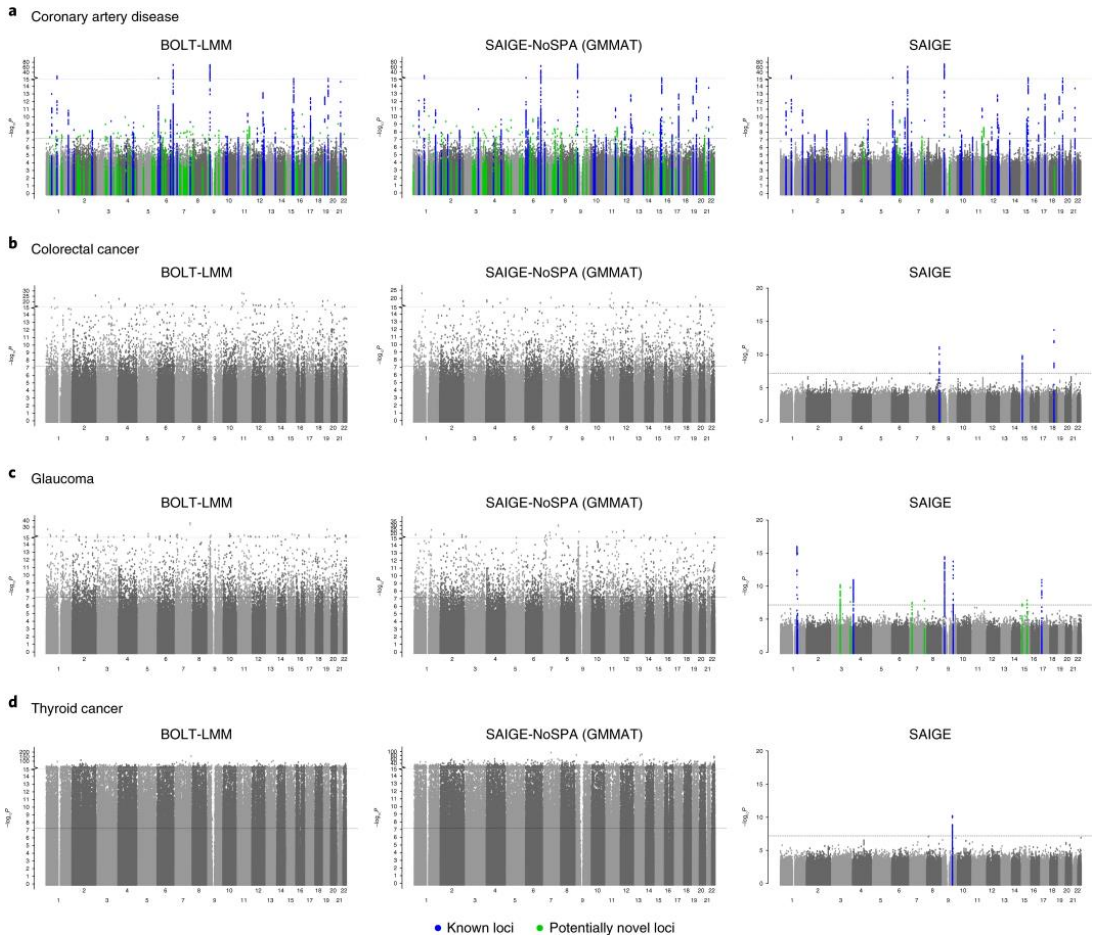
nature genetics
ANALYSIS
https://doi.org/10.1038/s41588-018-0184-y

Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies

Wei Zhou[1,2], Jonas B. Nielsen[3], Lars G. Fritsche[2,4,5], Rounak Dey[2,5], Maiken E. Gabrielsen[4], Brooke N. Wolford[1,2], Jonathon LeFaive[2,5], Peter VandeHaar[2,5], Sarah A. Gagliano[2,5], Aliya Gifford[6], Lisa A. Bastarache[6], Wei-Qi Wei[6], Joshua C. Denny[6,7], Maoxuan Lin[3], Kristian Hveem[4,8], Hyun Min Kang[2,5], Goncalo R. Abecasis[2,5], Cristen J. Willer[1,3,9,10*] and Seunggeun Lee[2,5,10*]

In genome-wide association studies (GWAS) for thousands of phenotypes in large biobanks, most binary traits have substantially fewer cases than controls. Both of the widely used approaches, the linear mixed model and the recently proposed logistic mixed model, perform poorly; they produce large type I error rates when used to analyze unbalanced case-control phenotypes. Here we propose a scalable and accurate generalized mixed model association test that uses the saddlepoint approximation to calibrate the distribution of score test statistics. This method, SAIGE (Scalable and Accurate Implementation of GEneralized mixed model), provides accurate *P* values even when case-control ratios are extremely unbalanced. SAIGE uses state-of-art optimization strategies to reduce computational costs; hence, it is applicable to GWAS for thousands of phenotypes by large biobanks. Through the analysis of UK Biobank data of 408,961 samples from white British participants with European ancestry for > 1,400 binary phenotypes, we show that SAIGE can efficiently analyze large sample data, controlling for unbalanced case-control ratios and sample relatedness.





Case: control = 1:12

Case: control = 1:84

Case: control = 1:89

Case: control = 1:1,138

- To adjust for sample relatedness, the leave-one-chromosome-out (LOCO) scheme is used.

Zhou et al., Nature Genetics, 2018; 50: 1335 - 1341

Empirical type 1 error rates for SAIGE, SAIGE-NoSPA, GMMAT, and BOLT-LMM

Estimated based on 10^9 simulated data sets. BOLT-LMM: compute non-infinitesimal association statistics; BOLT-LMM_ImmInfOnly: compute mixed model association statistics under the infinitesimal model

Variance Component Parameter τ	Case:Control	Test	Empirical Type 1 Error Rates	
			$\alpha = 5\times10^{-4}$	$\alpha = 5\times10^{-8}$
1	1:1	SAIGE	5.11×10^{-4}	5.45×10^{-8}
		SAIGE-NoSPA	4.71×10^{-4}	4.00×10^{-8}
		GMMAT	4.66×10^{-4}	3.81×10^{-8}
		BOLT-LMM_ lmmInfOnly	4.83×10^{-4}	4.83×10^{-8}
		BOLT-LMM	4.95×10^{-4}	4.99×10^{-8}
	1:9	SAIGE	4.43×10^{-4}	4.01×10^{-8}
		SAIGE-NoSPA	6.72×10^{-4}	7.82×10^{-7}
		GMMAT	7.30×10^{-4}	1.00×10^{-6}
		BOLT-LMM_ lmmInfOnly	9.01×10^{-4}	2.73×10^{-6}
		BOLT-LMM	9.03×10^{-4}	2.71×10^{-6}
	1:99	SAIGE	3.82×10^{-4}	1.44×10^{-8}
		SAIGE-NoSPA	2.93×10^{-3}	9.76×10^{-5}
		GMMAT	3.31×10^{-3}	1.26×10^{-4}
		BOLT-LMM_ lmmInfOnly	4.02×10^{-3}	2.10×10^{-4}
		BOLT-LMM	4.02×10^{-3}	2.10×10^{-4}

Zhou et al., Nature Genetics, 2018; 50: 1335 - 1341

Algorithm details

- Step 1. Fitting the logistic mixed model under the null hypothesis
  - 1.1 Generalised linear mixed model and penalised quasi-likelihood.
  - 1.2 Estimate parameters using Average Information – REstricted Maximum Likelihood (AI-REML).

```
[m      2@hlogin2 [csf3] Albumin_SAIGE]$ cat step1_alb.qsub
#!/bin/bash --login
#$ -cwd
#$ -V

#$ -pe smp.pe 24

#$ -t 1


Rscript /mnt/iusers01/bk01-icvs/mbbxkxj2/scratch/apps/SAIGE/SAIGE-master/extdata/step1_fitNULLGLMM.R \
--plinkFile=./genotype_data/ukb_cal_v4_noprune_comb \
--phenoFile=ukb_phenotype_bi_mialb_326472.txt \
--phenoCol=MiAlb \
--covarColList=Age,Gender,Geno_Array,PC1,PC2,PC3,PC4,PC5,PC6,PC7,PC8,PC9,PC10 \
--sampleIDColinphenoFile=IID \
--traitType=binary \
--outputPrefix=./output/biAlb \
--nThreads=24 \
--LOCO=TRUE > ./SAIGE_bi_MiAlb_${SGE_TASK_ID}.log 2>&1
```

Algorithm details

- Step 1. Fitting the logistic mixed model under the null hypothesis
  - 1.1 Generalised linear mixed model and penalized quasi-likelihood.
  - 1.2 Estimate parameters using Average Information – REstricted Maximum Likelihood (AI-REML).

```
[m      2@hlogin2 [csf3] Albumin_SAIGE]$ cat step1_alb.qsub
#!/bin/bash --login
#$ -cwd
#$ -V

#$ -pe smp.pe 24

#$ -t 1


Rscript /mnt/iusers01/bk01-icvs/mbbxkxj2/scratch/apps/SAIGE/SAIGE-master/extdata/step1_fitNULLGLMM.R \
--plinkFile=./genotype_data/ukb_cal_v4_noprune_comb \
--phenoFile=ukb_phenotype_bi_mialb_326472.txt \
--phenoCol=MiAlb \
--covarColList=Age,Gender,Geno_Array,PC1,PC2,PC3,PC4,PC5,PC6,PC7,PC8,PC9,PC10 \
--sampleIDColinphenoFile=IID \
--traitType=binary \
--outputPrefix=./output/biAlb \
--nThreads=24 \
--LOCO=TRUE > ./SAIGE_bi_MiAlb_${SGE_TASK_ID}.log 2>&1
```

- Step 2. Single variant score tests with SaddlePoint Approximation (SPA)
  - 2.1 Calculate score test statistics based on logistic mixed model.
  - 2.2 Estimate the variation of the score test statistics.
  - 2.3 Estimate effect size (β), which equals to the natural logarithm of the odds ratio.
  - 2.4 Calculate P-value using SPA.

```
[m      2@hlogin2 [csf3] Albumin_SAIGE]$ cat step2_alb.qsub
#!/bin/bash --login
#$ -cwd
#$ -V

#$ -e /dev/null
#$ -o /dev/null

#$ -l mem512

#$ -t 1-22

module load tools/env/ukbiobank-full-release-2018

Rscript /mnt/iusers01/bk01-icvs/mbbxkxj2/scratch/apps/SAIGE/SAIGE-master/extdata/step2_SPAtests.R \
--bgenFile=${UKBB_IMPUTATION_DIR}/ukb_imp_chr${SGE_TASK_ID}_v3.bgen \
--bgenFileIndex=${UKBB_IMPUTATION_DIR}/ukb_imp_chr${SGE_TASK_ID}_v3.bgen.bgi \
--minMAF=0.0001 \
--minMAC=4 \
--sampleFile=./samplefileforbgen_487410samples.txt \
--GMMATmodelFile=./output/biAlb.rda \
--varianceRatioFile=./output/biAlb.varianceRatio.txt \
--SAIGEOutputFile=./output/biAlb_SAIGE_chr${SGE_TASK_ID}.bgen.txt \
--numLinesOutput=2 \
--IsOutputAFinCaseCtrl=TRUE > ./step2_log/SAIGE_bi_MiAlb_step2_${SGE_TASK_ID}.log 2>&1
```

Con sides of SAIGE

- Time Consuming. (320k samples → 1 week running time)
- Maybe false negative when case and control approximate balanced, and slight conservative when extremely unbalanced.
- Require manual effort on input data preparation.
- SAIGE assumes that the effect sizes of genetic markers are following standard Normal distribution. (i.e. N(0,1))
- The variant component estimate τ from SAIGE may be biased as it is estimated by penalised quasi-likelihood function.

Zhou et al., Nature Genetics, 2018; 50: 1335 - 1341

SAIGE-driven resources on UK Biobank data

- The GWAS results for ~1,400 binary phenotypes using 400,000 White British samples and 28 million imputed variants are currently available: https://www.leelabsg.org/resources
- Gene-based rare variant test (MAF < 0.01) for 53 UK Biobank quantitative phenotypes were carried out using SAIGE-GENE. Missense and stop-gain variants were used in the analysis. Summary statistics are available via: ftp://share.sph.umich.edu/UKBB_SAIGE_GENE_HRC
- MichiganPheWeb: http://pheweb.sph.umich.edu/SAIGE-UKB/

Zhou et al., Nature Genetics, 2018; 50: 1335 - 1341

Global challenge for scalable outcome phenotyping

- Background: To make sense of the linked healthcare data and the potential for all the other data in UK Biobank to drive the process of phenotyping the health outcomes of the participant.
- Aim: Contributing data-driven scalable methods to create phenotype that helps define new diseases and recognise *novo* disease clusters to enhance the discovery potential of UK biobank.
- Who will be applicable to this challenge?
  - geeks,
  - computational health scientists,
  - bioinformaticians, and
  - anyone else who are interested.
- Formally release time: Q3-Q4, 2019
- Action point: Register your interest online and being updated
- A specific platform will be built to host this challenge

MANCHESTER 1824
The University of Manchester

biobank uk
Improving the health of future generations

Thank you for your attention!

Kidney Research UK
Funding research to save lives

British Heart Foundation