The UK Biobank Project: adding genome-wide genetic data on 500,000 individuals

Hugh Watkins
The Wellcome Trust Centre for Human Genetics
University of Oxford

Array-based SNP Genotyping

- Genome-wide
- Robust (data generation & analysis), scalable, affordable
- Good for common variant : common disease

- Genotyping all 500,000 participants

  - need the power
  - uniformity desirable
  - allows nested case:control studies

Array-based SNP Genotyping

Funding secured for genotyping all 500,000 participants (UK MRC, NIHR, BHF).

MRC Medical Research Council

NHS National Institute for Health Research

British Heart Foundation

Two phases: First 50,000 participants: UK BiLEVE Study

Next 450,000: UK Biobank led.

Tender process to select genotyping platform, first for UK BiLEVE, the UK Biobank.

Both selected Affymetrix Axiom platform.

Membership of the UK Biobank Array Design Group

Peter Donnelly (chair), University of Oxford
Jeff Barrett, Wellcome Trust Sanger Institute
Jose Bras, University College London
Adam Butterworth, University of Cambridge
Richard Durbin, Wellcome Trust Sanger Institute
Paul Elliott, Imperial College London
Ian Hall, University of Nottingham
John Hardy, University College London
Mark McCarthy, University of Oxford
Gil McVean, University of Oxford
Tim Peakman, UK Biobank
Nazneen Rahman, The Institute of Cancer Research
Nilesh Samani, University of Leicester
Martin Tobin, University of Leicester
Hugh Watkins, University of Oxford

Design Process

Expert group asked to design the array.

-> specific sets of SNPs for inclusion on the chip, often with further advice from experts in particular areas.

-> Affymetrix then used their imputation-aware algorithms to choose additional SNPs to provide good coverage of the genome in selected categories.

UK Biobank Array Content Summary

http://www.ukbiobank.ac.uk/scientists-3/uk-biobank-axiom-array/

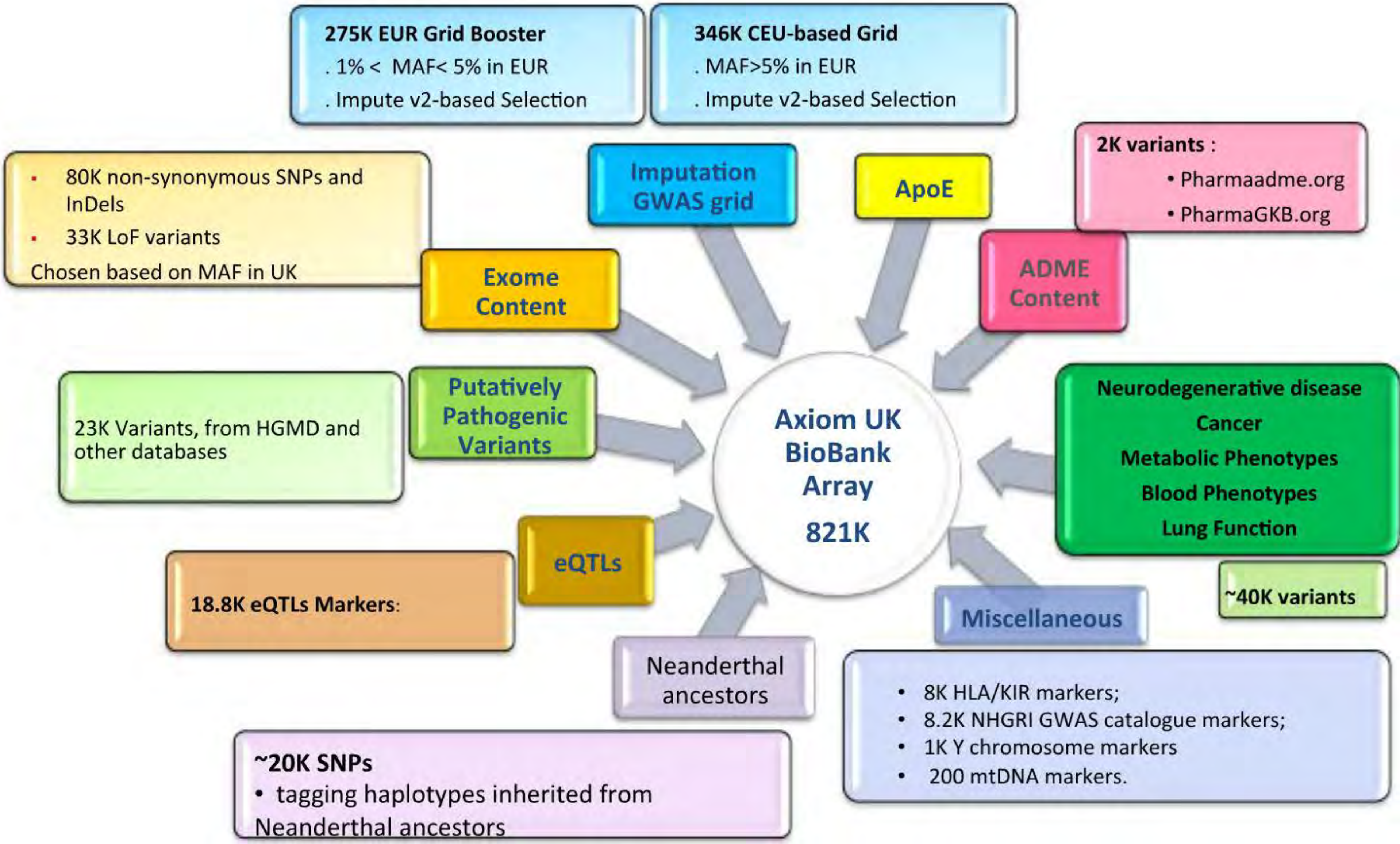
Content summary

Search/enquiry options

Non-disclosure agreement etc

How to order more!

UK Biobank Array Content Summary



Axiom UK BioBank Array

821K

Imputation GWAS grid

275K EUR Grid Booster

. 1% < MAF< 5% in EUR

. Impute v2-based Selection

346K CEU-based Grid

. MAF>5% in EUR

. Impute v2-based Selection

ApoE

ADME Content

2K variants :

- Pharmaadme.org
- PharmaGKB.org

Exome Content

- 80K non-synonymous SNPs and InDels
- 33K LoF variants

Chosen based on MAF in UK

Putatively Pathogenic Variants

23K Variants, from HGMD and other databases

eQTLs

18.8K eQTLs Markers:

Neurodegenerative disease

Cancer

Metabolic Phenotypes

Blood Phenotypes

Lung Function

~40K variants

Miscellaneous

- 8K HLA/KIR markers;
- 8.2K NHGRI GWAS catalogue markers;
- 1K Y chromosome markers
- 200 mtDNA markers.

Neanderthal ancestors

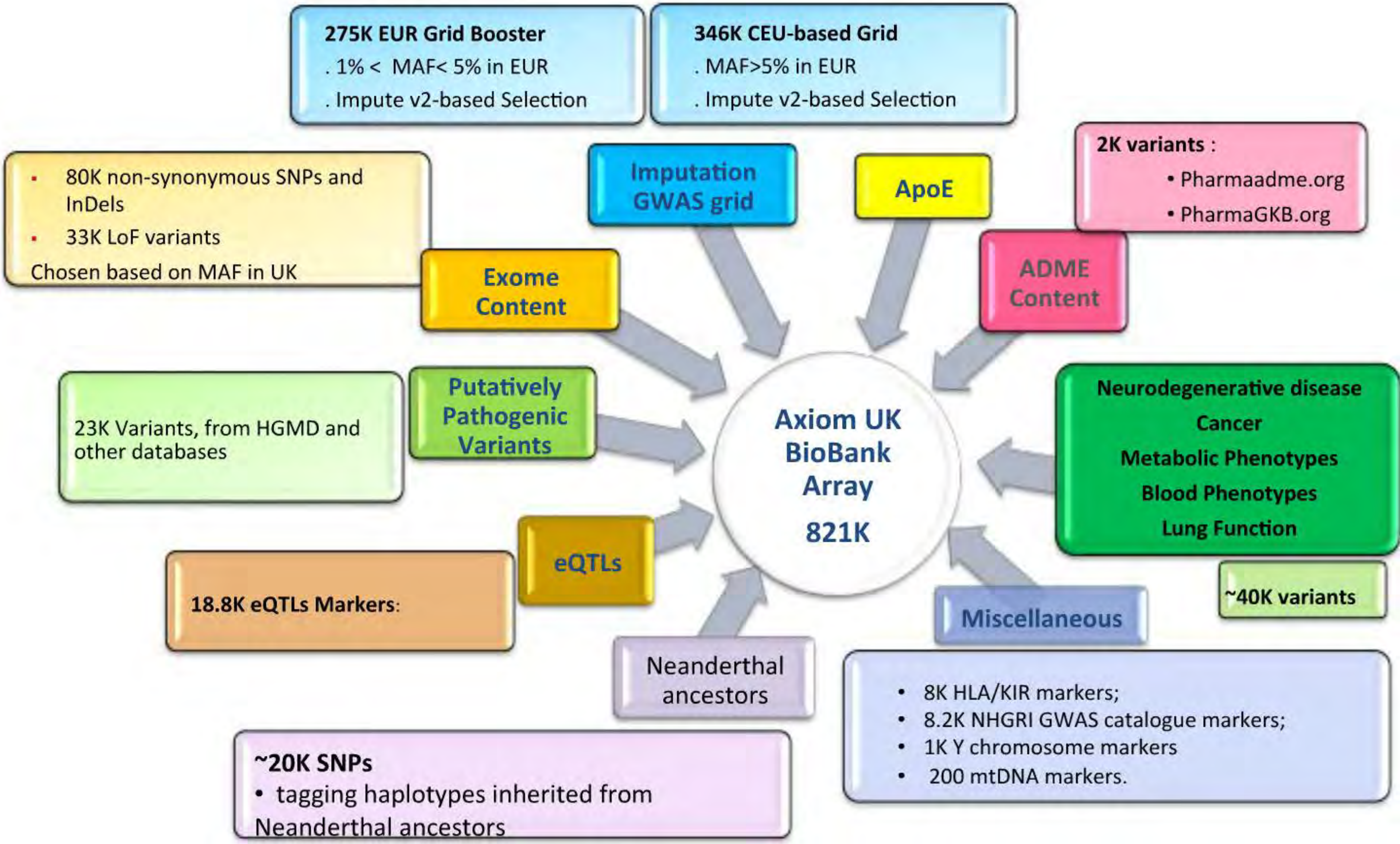
~20K SNPs

- tagging haplotypes inherited from Neanderthal ancestors

UK Biobank Axiom® Array Content Summary

Category	Number of markers
Markers of Specific Interest	
Alzheimer’s Disease	803
ApoE	1,147
Autoimmune/Inflammatory	258
Blood Phenotypes	2,545
Cancer common variants	343
Cardiometabolic	377
eQTL	17,115
Fingerprint	262
HLA	7,348
KIR	1,546
Lung function phenotypes	8,645
Common mitochondrial DNA variants	180
Neurological disease	19,791
NHGRI GWAS catalog	8,136
Pharmacogenetics/ADME	2,037
Tags for Neanderthal ancestry	11,507
Y chromosome markers	807
Rare variants in cancer predisposition genes	6,543
Rare variants in cardiac disease predisposition genes	1,710
Rare, possibly disease causing, mutations	13,729
CNV regions for developmental delay, neuropsychiatric disorders and lung function	2,369

UK Biobank Array Content Summary



275K EUR Grid Booster

. 1% < MAF< 5% in EUR

. Impute v2-based Selection

346K CEU-based Grid

. MAF>5% in EUR

. Impute v2-based Selection

Imputation GWAS grid

ApoE

2K variants :

- Pharmaadme.org
- PharmaGKB.org

ADME Content

- 80K non-synonymous SNPs and InDels
- 33K LoF variants

Chosen based on MAF in UK

Exome Content

23K Variants, from HGMD and other databases

Putatively Pathogenic Variants

Axiom UK BioBank Array

821K

Neurodegenerative disease

Cancer

Metabolic Phenotypes

Blood Phenotypes

Lung Function

~40K variants

eQTLs

18.8K eQTLs Markers:

Miscellaneous

Neanderthal ancestors

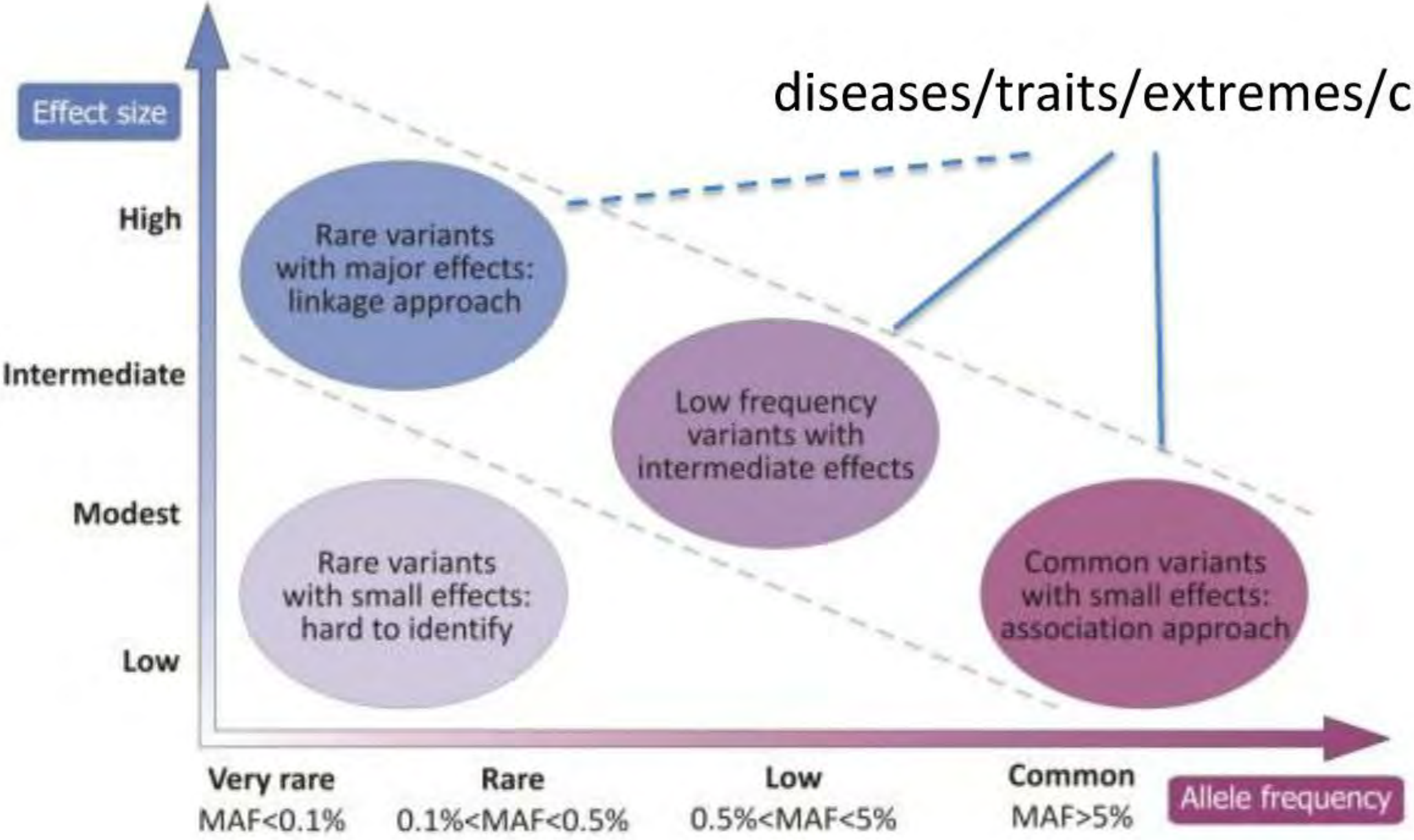
~20K SNPs

- tagging haplotypes inherited from Neanderthal ancestors

- 8K HLA/KIR markers;
- 8.2K NHGRI GWAS catalogue markers;
- 1K Y chromosome markers
- 200 mtDNA markers.

Opportunities

diseases/traits/extremes/controls


Effect size
High
Intermediate
Modest
Low
Rare variants with major effects: linkage approach
Low frequency variants with intermediate effects
Rare variants with small effects: hard to identify
Common variants with small effects: association approach
Very rare MAF<0.1%
Rare 0.1%<MAF<0.5%
Low 0.5%<MAF<5%
Common MAF>5%
Allele frequency

Genotype Imputation

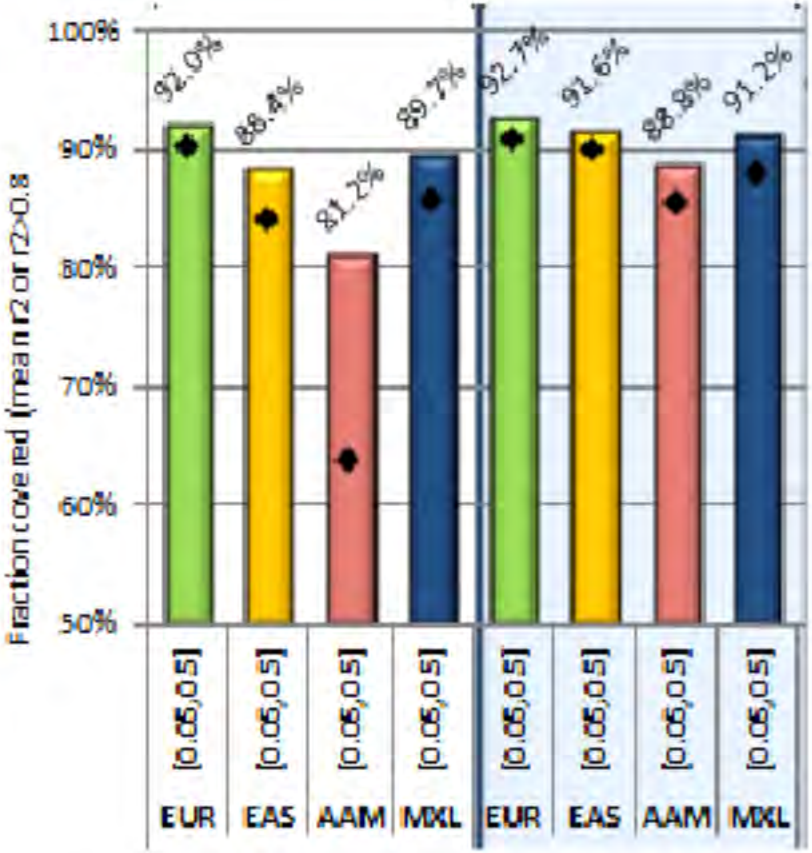
Sample size for UK Biobank offers potential for powerful new imputation methods.

Classical HLA Alleles

Different algorithms also allow imputation of genotypes at the classical HLA loci.

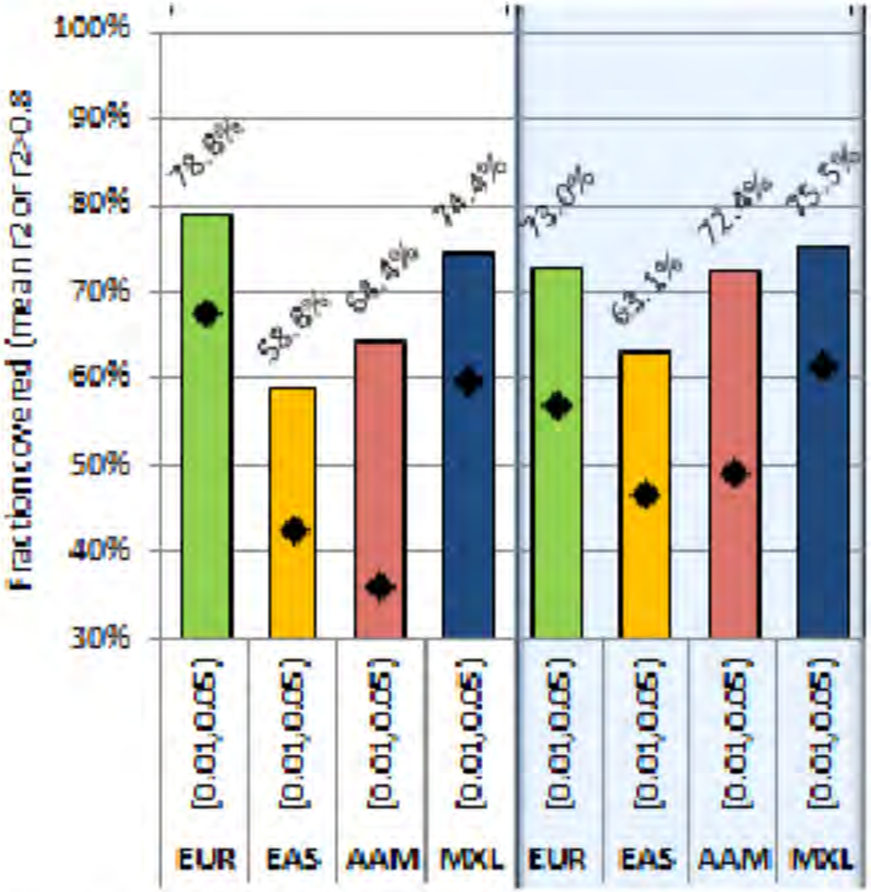
For European ancestry samples, imputation accuracy is good: over 95% at four-digit accuracy.

Exploring changes to the array to increase coverage in other ethnic groups.



Imputation-based coverage of common variation (**MAF > 5%**).
Left: UK Biobank Array; Right: Array which removes 250,000 SNPs (covering 1-5% variation) from UK Biobank Array and replaces them with SNPs to capture common variation in east Asian and African American samples.

Exploring changes to the array to increase coverage in other ethnic groups.



Imputation-based coverage of low frequency variation **1% < MAF < 5%**)
Left: UK Biobank Array; Right: Array which removes 250,000 SNPs (covering 1-5% variation) from UK Biobank Array and replaces them with SNPs to capture common variation in east Asian and African American samples.

Set up timeline

Date	Milestone	Elapsed Time
Fri - Aug-2	• UK Biobank Array Content approved	17 calendar days to manufacture first batch of Axiom UK Biobank Array
Mon – Aug-12	• Start Sample processing	
Mon – Aug-19	• Axiom UK Biobank received in	20 calendar days to genotype Samples on Axiom UK Biobank Array
Thu– Aug-21	• Complete Sample processing • Start Data Analysis	
Mon– Aug-26	• Complete Data Analysis	

~6,000 samples/week from December 2013

Timelines: primary data generation

Affy target schedule for UK Biobank

100,000	28 Apr 2014
200,000	18 Aug 2014
300,000	08 Dec 2014
450,000	22 Jun 2015 (finish)

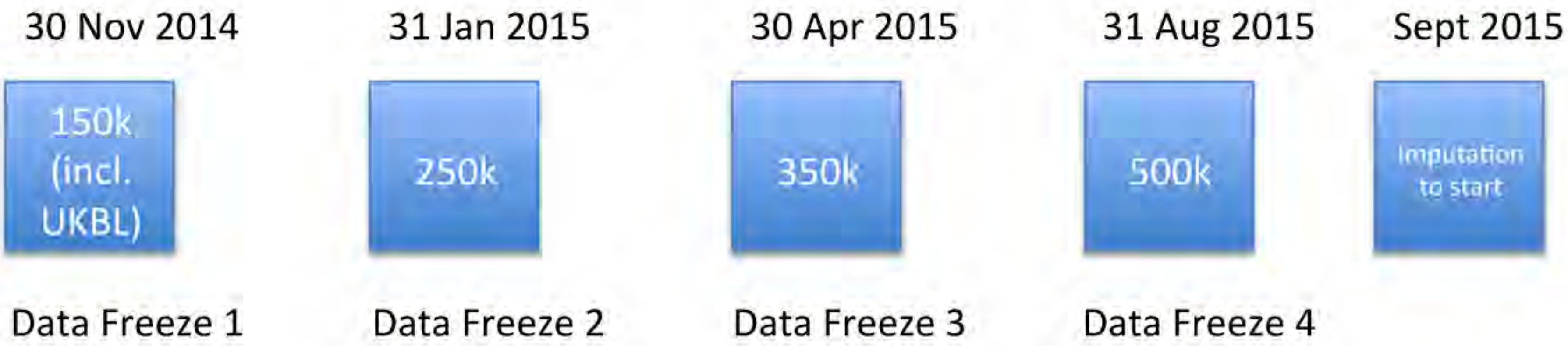
Current data received

~98k	19 June 2014

Expecting +25k by month end June 2014

Possible Timelines: to finish QC

- **First batch of called genotype data for 150k available by the end of 2014**



30 Nov 2014	31 Jan 2015	30 Apr 2015	31 Aug 2015	Sept 2015
150k (incl. UKBL)	250k	350k	500k	Imputation to start
Data Freeze 1	Data Freeze 2	Data Freeze 3	Data Freeze 4	

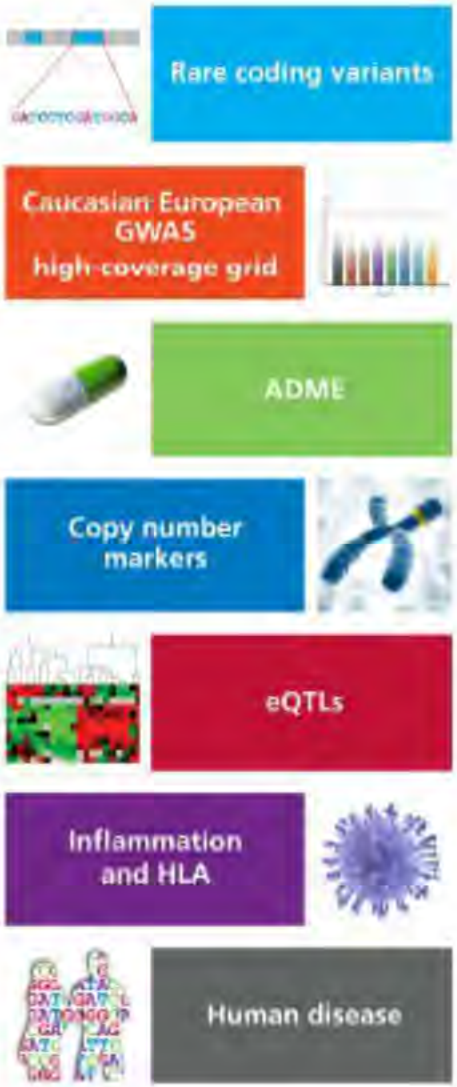
- **All called genotype data for 500k by autumn 2015**

- **Imputation will start autumn 2015 and is expected to be available in 2016**

Presentations

- ASHG 2014 (QC presentation; Donnelly group)

- QC publication



Rare coding variants

Caucasian European GWAS high-coverage grid

ADME

Copy number markers

eQTLs

Inflammation and HLA

Human disease

biobank uk

Improving the health of future generations

UK Biobank Axiom® Array

Content Summary

Membership of the UK Biobank Array Design Group

Peter Donnelly (chair), University of Oxford
Jeff Barrett, Wellcome Trust Sanger Institute
Jose Bras, University College London
Adam Butterworth, University of Cambridge
Richard Durbin, Wellcome Trust Sanger Institute
Paul Elliott, Imperial College London
Ian Hall, University of Nottingham
John Hardy, University College London
Mark McCarthy, University of Oxford
Gil McVean, University of Oxford
Tim Peakman, UK Biobank
Nazneen Rahman, The Institute of Cancer Research
Nilesh Samani, University of Leicester
Martin Tobin, University of Leicester
Hugh Watkins, University of Oxford

http://www.ukbiobank.ac.uk/scientists-3/uk-biobank-axiom-array/

Deliverables

- Genotype calls
  - Original calls plus:
  - Re-calls (from Affymetrix); ApoE, rare variants, batch effects
  - Failed SNPs set to missing per batch (e.g. Affymetrix fails, plate effect, batch effect)
- Additional QC information e.g.
  - PCA (10 PCs)
  - Related Individuals (1st-3rd degree)
  - Sample QC metrics (e.g. missingness, het. rate)
  - SNP QC metrics (e.g. Call rate, MAF, HWE)
- Intensity data (for cluster plots)
- Documentation (including Use Cases)

Additional

- Archive - Affymetrix data (Full set of files; CEL files, original calls etc.)

- Subsets – 'European'
  - Self reported ethnicity 'British/Irish/Other White' individuals; PCA cluster defined with Aberrant
  - Useful as controls for UK GWAS studies or studies that require a large homogeneous population
  - e.g. UK BiLEVE used self-reported ethnicity to choose samples

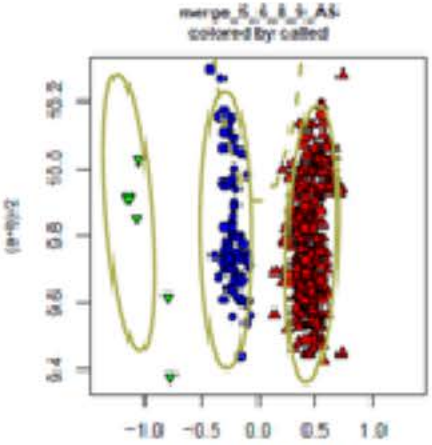
- CNVs – Affymetrix calls ?

rs429358, Apo E SNP implicated in Alzheimer's disease
Performance profile on UKBioBank Axiom array

- rs429358 historically challenging to genotype with Hybridization-based assay
  - High GC content in flanking sequence (>76% FWD in 80% REV strands)
  - Affymetrix advanced Axiom Probe Design algorithms enabled successful genotyping with specific Axiom probe construct

Colored according to calls on UKBiobank array in 270 CEU, GBR, TSI individuals.

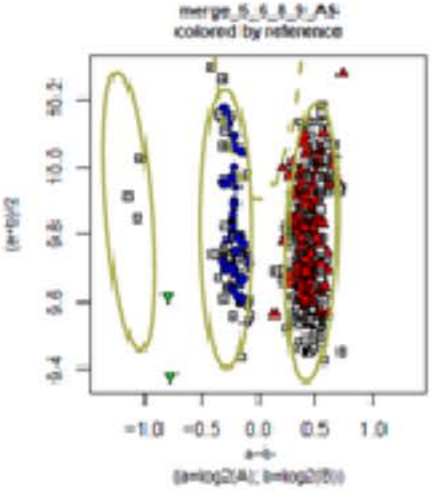
- Call rate: 100%
- Concordance to 1000 Genomes:
  - ❑ - 100% overall
  - ❑ - 100% heterozygote concordance



Colored according to 1000 Genomes (phase 1, March 2012)

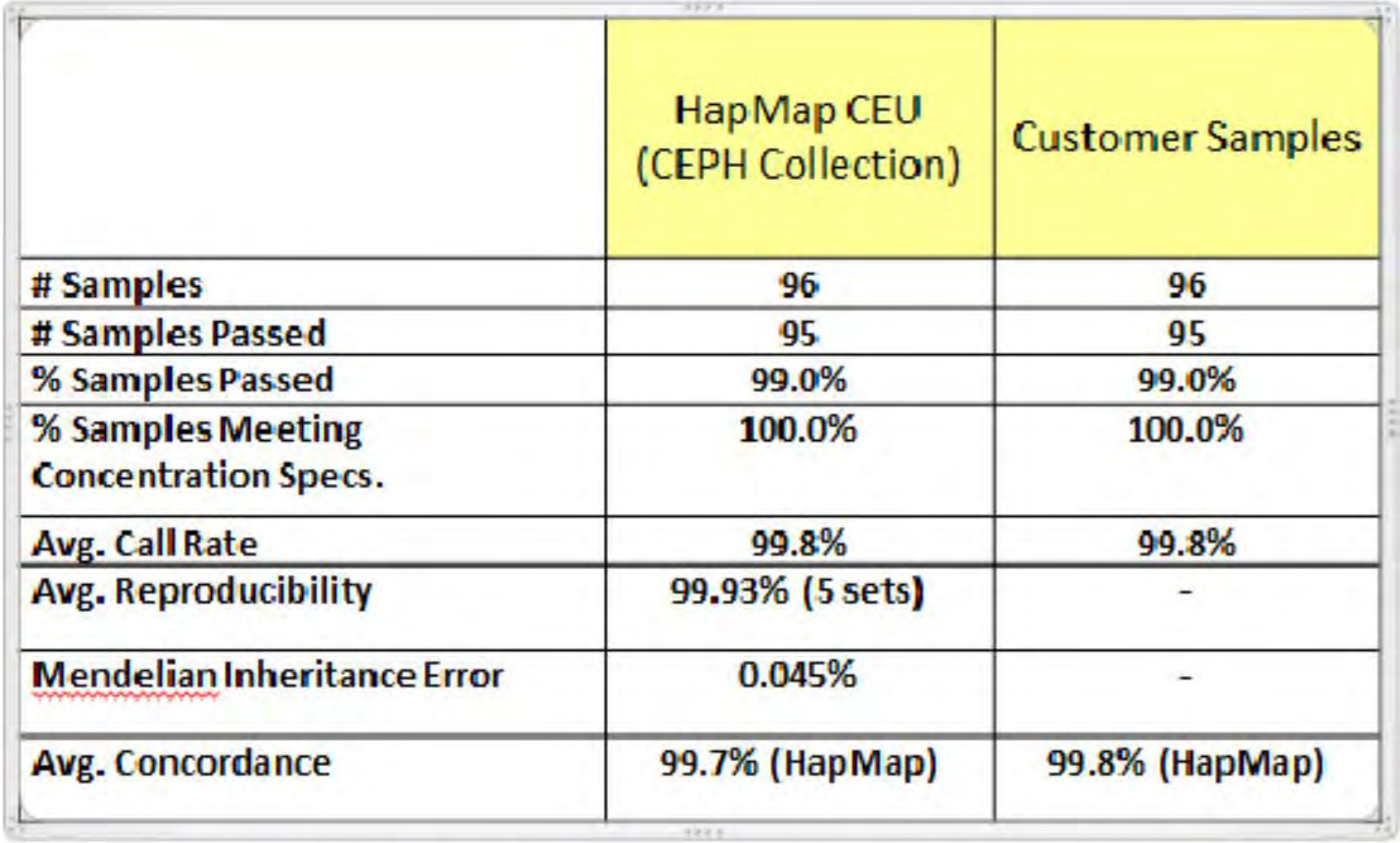
calls in same individuals

(gray indicates individual not included in 1000 Genomes)



affymetrix

Performance Summary – Early Results



	HapMap CEU (CEPH Collection)	Customer Samples
# Samples	**96**	**96**
# Samples Passed	**95**	**95**
% Samples Passed	**99.0%**	**99.0%**
% Samples Meeting Concentration Specs.	**100.0%**	**100.0%**
Avg. Call Rate	**99.8%**	**99.8%**
Avg. Reproducibility	**99.93% (5 sets)**	**-**
Mendelian Inheritance Error	**0.045%**	**-**
Avg. Concordance	**99.7% (HapMap)**	**99.8% (HapMap)**

- Analys is following Affymetrix best practices workflow analysis http://www.affymetrix.com/support/downloads/manuals/axiom_best_practice_supplement_user_guide.pdf
- 3 HapMap controls included on customer plate

affymetrix

Average r^2 of all target markers in the specified MAF range			
MAF Range	Population	Axiom UKBB Array	ILMN HCE
[0.05,0.5]	CEU	0.925	0.869
[0.01,0.05)	CEU	0.767	0.599
[0.05,0.5]	GBR	0.917	0.862
[0.01,0.05)	GBR	0.738	0.581
[0.05,0.5]	CHB	0.877	0.838
[0.01,0.05)	CHB	0.548	0.477
[0.05,0.5]	JPT	0.880	0.841
[0.01,0.05)	JPT	0.543	0.475
[0.05,0.5]	MXL	0.897	0.851
[0.01,0.05)	MXL	0.735	0.624
[0.05,0.5]	YRI	0.812	0.782
[0.01,0.05)	YRI	0.643	0.599
[0.05,0.5]	LWK	0.808	0.783
[0.01,0.05)	LWK	0.636	0.598
[0.05,0.5]	ASW	0.809	0.774
[0.01,0.05)	ASW	0.659	0.599

Genome-wide coverage, via imputation, in different allele frequency ranges, for various populations