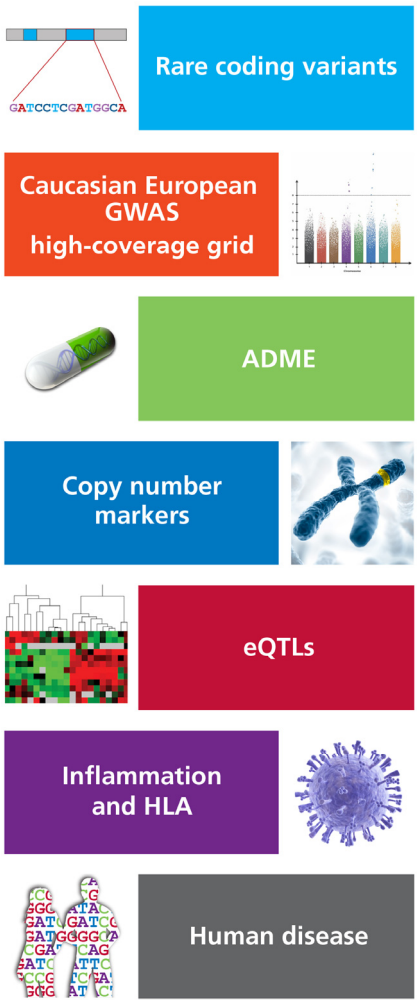
biobank uk
Improving the health of future generations

UK Biobank Axiom® Array
Content Summary

This document provides an overview of the content of UK Biobank Axiom® Array, which was designed by the UK Biobank Array Design Group. Full membership of the design group is listed at the end of this document.

There are 820,967 SNP and indel markers on UK Biobank Axiom Array.



Further details are included below. At the highest level, the general array design philosophy was to:

1. Add markers of particular interest, or tags for such markers, because of known associations or possible roles in phenotypic variation. There are 95,490 markers on the array of this type.

2. Add coding variants (principally missense and protein truncating variants) subject to certain estimated minor allele frequency (EMAF) criteria (which differed for different categories of coding variant). There are 111,904 markers on the array of this type.

3. Choose the remaining content to provide good genome-wide coverage in Caucasian populations in particular EMAF ranges. (This marker selection assumed imputation methods would be used in downstream analysis and used Affymetrix' software tools.) There are 629,368 markers on the array of this type.

There are two closely related arrays, referred to as the UK BiLEVE array and UK Biobank Axiom Array. The UK BiLEVE array was designed first and was run on the first ~50,000 samples genotyped for UK Biobank. UK Biobank Axiom Array was designed subsequently with a view to maximizing marker overlap as much as possible between the arrays. Some minor changes were made to the array to produce UK Biobank Axiom Array, which will be typed on the remaining ~450,000 samples. The UK BiLEVE and UK Biobank Axiom Arrays are extremely similar with over 95% common content.

While UK Biobank Axiom Array is commercially available on the Axiom® platform, Affymetrix also markets an array called "Axiom® Biobank Genotyping Array" in which the design objective is quite different from both the UK BiLEVE and UK Biobank arrays.

More detailed description of specific categories of content follows. The number of markers in each category refers to the number of markers on the array that were selected from a given category. A marker can be selected for more than one category but only appears on the array once. It should be noted that when we refer to the possible disease causing or disease associated variants we do not mean to imply that the status of the variant, or any definitive role in, or association with, disease has been established.

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
Rare coding variants	
Protein truncating variants	30,581
Other rare coding variants	80,581
Genome-wide coverage	
Genome-wide coverage for common variants	348,569
Genome-wide coverage for low frequency variants	280,838
Total number of markers	820,967

1. Markers of Specific Interest

1.1 Markers of particular phenotypic interest

Alzheimer's disease (803 markers)
These markers were chosen from a list of variants showing some evidence for association with Alzheimer's disease from a meta-analysis of Alzheimer's genome-wide association studies (Lambert *et al.*, *Nat Genet.* 2013 Dec;45(12):1452-8.). Additionally, a set of mitochondrial markers which are suspected to be associated with Alzheimer's disease, provided to us by a collaborating Alzheimer's disease group, were added to the array.

ApoE (1,147 markers)
The two SNPs, rs429358 and rs7412, which define the ApoE isoforms known to be associated with risk of Alzheimer's disease and other conditions, were included on the array. In addition, a dense set of markers across the ApoE region were included.

Autoimmune/Inflammatory (258 markers)
Variants were included on the array which had evidence for association with specific autoimmune/inflammatory disorders (Ulcerative colitis, Crohn's disease, Type 1 diabetes, Graves disease, Hashimoto's thyroiditis and Celiac disease). These were identified by studies based on the Illumina Immunochip array (many will not appear in the NHGRI GWAS catalog as the array is not considered to be a genome-wide array).

Blood Phenotypes (2,545 markers)
Markers showing evidence for association with various blood phenotypes were chosen from GWAS and candidate gene studies. The phenotypes included regulation of formation of red blood cells and platelets, regulation of blood homeostasis and red cell blood groups.

Cancer Common Variants (343 markers)
Markers were chosen from the list of published common variants associated with cancer phenotypes identified via GWAS, as per the NHGRI GWAS catalog, as well as some recently published and some unpublished cancer-associated SNPs as at June 2013.

Cardiometabolic (377 markers)
Variants were included on the array which are known to be associated with various cardiometabolic traits (e.g. coronary disease, lipids, anthropometry, glycaemic markers, blood pressure). Many were discovered by GWAS follow-up studies based on the Illumina CardioMetabochip array and hence do not appear in the NHGRI GWAS catalog as the array is not considered to be a genome-wide array.

eQTL (17,115 markers)
eQTL markers were supplied from the GEUVADIS project in advance of their recent publication (Lappalainen *et al*., *Nature*. 2013 Sep 26;501(7468):506-11.), together with eQTLs for several other discovery projects including the MuTHER project, the ALSPAC project and the GENCORD project.

Fingerprint (262 markers)
Fingerprint is a set of SNPs used as fingerprint SNPs by the University of Washington and the Broad Institute. These have been shared among several major genotyping platforms to facilitate sample tracking. The set of Fingerprint markers from Affymetrix' Axiom® Biobank Genotyping Array were adopted for UK Biobank Axiom Array.

HLA (7,348 markers)
Genes in the HLA (chr6) region are known to be important in immune response but are not easily assayed directly on commercial arrays. Markers have been added to the array to facilitate imputation of the alleles at the classical Class I and Class II loci.

KIR (1,546 markers)
Genes in the KIR (chr19) region are known to be important in immune response but are not easily assayed directly on commercial arrays. Markers have been added to the array to facilitate imputation of KIR genes.

Lung function phenotypes (8,645 markers)
A set of markers with established or putative association with lung function, lung disease (including asthma, cystic fibrosis, chronic obstructive pulmonary disease, idiopathic pulmonary fibrosis and lung cancer) and/or smoking behavior were included on the array.

Common mitochondrial DNA variants (180 markers)
A list of coding and non-coding mitochondrial DNA (mtDNA) population markers were included on the array. These could be used as a framework to generate complex phylogenetic networks and conduct mtDNA haplogroup comparisons. Additionally, a number of variants with possible associations with the commonest mitochondrial disorders have been included.

Neurological disease (19,791 markers)
This list is comprised of a large number of rare possibly disease associated mutations for a variety of neurological diseases, as well as recently identified association hits. Cohorts of diseases considered included: Alzheimer's disease, frontotemporal dementia, progressive supranuclear palsy, amyotrophic lateral sclerosis and Parkinson's disease. Additionally, a number of markers were chosen from exome sequencing studies in the diseases above which are not present in controls or publicly available databases. The Neurological disease content on UK Biobank Axiom® Array aims to tag the same content as is on the Illumina NeuroX array.

NHGRI GWAS catalog (8,136 markers)
Markers were chosen directly from, or as tags for, markers in the NHGRI Catalog of Published Genome-Wide Association Studies as at January 2013.

Pharmacogenetics/ADME (2,037 markers)
The ADME category consists of markers from the Pharmacogenomics Knowledgebase (http://www.pharmgkb.org/) with known relevance to drug metabolism, and some markers from Affymetrix' DMET™ Plus platform. The set of ADME markers from Affymetrix' Axiom® Biobank Genotyping Array were adopted for UK Biobank Axiom® Array.

Tags for Neanderthal ancestry (11,507 markers)
There is known to be a small contribution to the genome of modern Europeans from Neanderthal DNA. Genetic segments or haplotypes thought to be inherited from our Neanderthal ancestors, but still present in modern populations, might be enriched for SNPs with functional effects. This set of markers aims to tag some such haplotypes. Markers were chosen from the set of markers identified as Neanderthal-derived that efficiently tag the Neanderthal haplotypes. Markers were also chosen from the set of Neanderthal-derived markers which are much more frequent (>3x) in modern European populations compared to modern African populations.

Y chromosome markers (807 markers)
These markers define lineages on the male-specific region of the Y Chromosome, including markers to identify all the main branches of the Y Phylogeny. Markers were included that are present but rare in the UK to provide information about population frequencies of higher resolution sub-branches.

1.2. Rare variants in disease predisposition genes

Rare variants in cancer predisposition genes (6,543 markers)
The markers in this category are rare missense variants in proven cancer predisposition genes that have been reported in cases of disease. These were selected from sources including HGMD (Human Gene Mutation Database) and locus-specific databases. It is hoped that analysis of a

large unselected population will allow improved stratification into non-pathogenic and pathogenic variants.

Rare variants in cardiac disease predisposition genes (1,710 markers)
Rare variants in the genes, MYBPC3 and MYH7 were included on the array. A second set of markers were chosen from the ARVD/C locus specific database (http://www.arvcdatabase.info/) to look at variants in the genes DSC2, DSG2, DSP, JUP and PKP2. Additionally, some markers were chosen from HGMD for genes related to cardiac disease and haemochromatosis. It is hoped that analysis of a large unselected population will allow improved stratification into non-pathogenic and pathogenic variants.

Rare, possibly disease causing, mutations (13,729 markers)
Markers were chosen to allow investigation of the frequency of a set of rare, possibly disease causing mutations from HGMD, for some disorders relevant to lung function and other phenotypes. Additional content was added to the array from the set of markers in HGMD which are polymorphic in the Exome Aggregation Consortium (ExAC) data.

1.3 Copy number variants (CNVs)

CNV regions for developmental delay, neuropsychiatric disorders and lung function (2,369 markers)

In a set of 67 CNVs, selected for developmental delay phenotypes (see Cooper *et al*., *Nat Genet.* 43(9):838-846 (2011) and the DECIPHER database http://decipher.sanger.ac.uk), markers were added where necessary to ensure dense coverage within the CNV region, and where breakpoints were known, assays were included for the breakpoints if possible. Markers were also added to ensure coverage of a small number of CNVs which have shown suggestive evidence of association with lung function and a small number of candidate CNV regions for which there was interest to test association with lung function and related phenotypes.

2. Rare coding variants

Using the content from Affymetrix' Axiom® Exome Genotyping Arrays as a starting point, data available from the Exome Chip and exome sequencing experiments was used to estimate allele frequencies in the UK population. Where this allele frequency data suggested that even in the large sample size (500,000) in the UK Biobank relatively few copies of the minor allele would be seen, it was expected that there would be limited power to see any effects, and so such markers were not chosen for the array. Protein truncating variants (PTV) had a lower frequency cut-off than other coding variants. The actual cut-offs used involved a trade-off between allele frequency and the amount of real estate on the array needed to genotype the marker. Data from three sources was used to assess allele frequency: UK Illumina Exome Chip data and Exome Aggregation Consortium (ExAC) European exome sequencing data (from the Broad Institute) and the UK10K non-Finnish exome sequencing data. Additional PTV content was added from interrogation of the ExAC data, not all of which data were available at the time of the design of Axiom Exome Genotyping Arrays.

Inclusion criteria based on estimated minor allele frequency (EMAF) are given below.

2.1 Protein truncating variants (30,581 markers)

Variants which truncate proteins (e.g. premature stop, frameshift, loss of start) are natural candidates to cause Loss of Function of that copy of the gene. Variants were included on the array as follows.

All PTV variants with EMAF > 0.0002
PTV variants with 0.00005 < EMAF < 0.0002 which require either 1 or 2 features*.

Additional PTV variants were chosen that were present in the ExAC exomes which were not on the commercial Axiom® Biobank Genotyping Array, with EMAF > 0.0002.

2.2 Other rare coding variants (80,581 markers, primarily Missense)

Variants were included on the array as follows.

Coding variants with EMAF > 0.001
Variants with 0.0004 < EMAF < 0.001 which require either 1 or 2 features*.
Variants with 0.0002 < EMAF < 0.0004 which require only 1 feature*.

3. Genome-wide Coverage

3.1 Genome-wide coverage for common variants (348,569 markers)

348,569 markers were selected using Affymetrix' imputation aware marker choice algorithms (Hoffman et al, Genomics 98 (2011) 422–430) to provide genome-wide coverage in Caucasian European populations of common (EMAF≥5%) markers (using the EUR panel defined as the GBR, CEU, FIN, IBS and TSI samples from 1000G). This explicitly included the set of 246,055 markers on Affymetrix' Axiom Biobank Genotyping Array selected to capture common (EMAF≥5%) variation.

3.2 Genome-wide coverage for low frequency variants (280,838 markers)

280,838 markers were selected using Affymetrix' imputation aware marker choice algorithms to provide genome-wide coverage in Caucasian European populations of low frequency (1%<EMAF<5%) markers (using the EUR panel described above).

Genome-wide imputation coverage in the EUR panel (see above for definition) estimated by Affymetrix:

Category	EMAF range	Mean r^2	% of markers with r^2>0.8
Common	5%≤EMAF≤50%	0.92	90.1%
Low frequency	1%<EMAF<5%	0.785	67.1%

* A feature is the smallest unit of real estate on Affymetrix' Axiom UK Genotyping Array. Typical AT or GC SNPs require 4 features, other SNPs require 2 features. Some previously validated SNPs only require a single feature.

Tagging Strategy
Where markers of particular interest were to be included on the array we first checked whether the marker was included in Affymetrix' Axiom® Genomic Database which had been experimentally validated for successful genotyping with Axiom® Genotyping Solution. Where the marker had been validated we selected it for inclusion on the array. Where the marker was not previously validated, we checked whether there was a good tag (typically r^2>0.7) in the validated set of markers and if one existed we selected the best such tag for inclusion on the array. For content added for UK Biobank Axiom® Array after the UK BiLEVE array was in production we similarly preferred good tags amongst markers already on the UK BiLEVE array.

Multi-allelic Markers
On the array there are 1,360 loci, corresponding to 2,881 markers that have multiple pairs of A/B alleles for the same chromosomal position. Some of these markers correspond to observed or expected rare alternate alleles which were added to the array because of their potential phenotypic interest. For a number of markers in the 'Rare variants in cancer predisposition genes' and 'Rare variants in cardiac disease predisposition genes' categories it was important to know exactly the counts of each possible A,C,G,T allele and so specific probes were added to the array for each allele.
Due to the chemistry of the array, these markers are difficult to interpret and, indeed, some require the development of new calling algorithms to analyse them. It is recommended that these markers are excluded from standard analyses.

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

Acknowledgements
We are very grateful to the many scientists and research groups who provided markers and data as part of the design process for UK Biobank Axiom® Array.

Version
1.0 07 Mar 2014

©2014 UK Biobank. All rights reserved.