UK Biobank Axiom Array

Advancing human health studies with powerful genotyping technology

Array highlights

The Applied Biosystems™ UK Biobank Axiom™ Array is a powerful array for translational research. Designed using imputation-aware SNP selection, this array provides optimized content modules for genome-wide association studies (GWAS) of common and low-frequency variants, biological function, and human disease in populations of European and British ancestry. The comprehensive coverage also includes rare coding variants, pharmacogenomics markers, copy number regions, human leukocyte antigen (HLA) genes, inflammation variants, and expressed quantitative trait locus (eQTL) markers.

This array was designed by leading researchers in the fields of epidemiology, human disease, and population genetics for use by the UK Biobank in one of the largest genotyping studies to date. The UK Biobank's prospective genotyping study of 500,000 individuals is aimed at uncovering the complex interactions between genes, lifestyle, and environment (marker–phenotype interactions). The resulting data will be available to researchers, providing a large reference data set (500,000 samples) that can be useful in assessing potential case associations of rare variants identified in studies of other populations.

Highly informative content categories relevant for translational research offer the best opportunity for identifying phenotype-associated variant candidates. See Table 1 for a detailed list of biological categories included on the array.

Array format

The array plates are available in two formats:

- UK Biobank Axiom Array (catalog)
  - The content is predefined and optimized for the European ancestry population (see Table 1)
  - Available as a catalog product
- Axiom Biobank Plus Genotyping Array (custom)
  - Markers can be added, removed, or replaced in every content module
  - Customizable for any population included in the 1000 Genomes Project
  - Array can be optimized for any disease focus or trait of interest

Array content

GWAS markers were selected using our imputation-aware algorithms [1–3] to provide genome-wide coverage in European ancestry populations of common (EMAF ≥5%, including the set of 246,055 markers on the Axiom Biobank Genotyping Array and low-frequency (1%< EMAF <5%) markers using the EUR panel defined as the GBR, CEU, FIN, IBS, and TSI samples from the KGP [4]. See Table 2 for the imputed genome-wide coverage calculated against the KGP March, 2012 build.

Table 1. Categories of markers for the European ancestry population* on the UK Biobank Axiom Array.

Category	No. of markers
Markers of specific interest	
Alzheimer's disease	803
APOE	1,147
Autoimmune and 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 disorders	19,791
NHGRI GWAS catalog	8,136
Pharmacogenomics/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 on array	**820,967**

* The European ancestry population is the EUR panel defined as the GBR, CEU, FIN, IBS, and TSI samples from the 1000 Genomes Project (KGP) [4]. Markers may be selected for more than one category but only appear on the array once. **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.**

Table 2. Imputed genome-wide coverage in European ancestry populations (EUR panel).

Category	EMAF* range	Mean r^2	% markers with $r^2 > 0.8$
Common	5%≤ EMAF ≤50%	0.920	90.1%
Low frequency	1%< EMAF <5%	0.785	67.1%

* EMAF: estimated minor allele frequency.

While genomic coverage in targeted European populations is high, this advanced GWAS array design also provides comprehensive coverage in diverse populations (Table 3). It is also important to note that this array may be customized to provide genomic coverage for any given population.

Table 3. Imputed genetic coverage metric (mean r^2) of the UK Biobank Axiom Array and a similar commercial array across 1000 Genomes Project populations at specified EMAF ranges.

EMAF range	Population	UK Biobank Axiom Array	Alternative array
[0.05,0.5]	CEU	0.925	0.869
[0.01,0.5]	CEU	0.875	0.783
[0.01,0.05]	CEU	0.767	0.599
[0.05,0.5]	GBR	0.917	0.862
[0.01,0.5]	GBR	0.859	0.771
[0.01,0.05]	GBR	0.738	0.581
[0.05,0.5]	FIN	0.927	0.883
[0.01,0.5]	FIN	0.871	0.799
[0.01,0.05]	FIN	0.764	0.638
[0.05,0.5]	TSI	0.915	0.855
[0.01,0.5]	TSI	0.845	0.751
[0.01,0.05]	TSI	0.711	0.551
[0.05,0.5]	CHB	0.877	0.838
[0.01,0.5]	CHB	0.776	0.728
[0.01,0.05]	CHB	0.548	0.477
[0.05,0.5]	CHS	0.887	0.849
[0.01,0.5]	CHS	0.785	0.738
[0.01,0.05]	CHS	0.572	0.505
[0.05,0.5]	JPT	0.880	0.841
[0.01,0.5]	JPT	0.777	0.729
[0.01,0.05]	JPT	0.543	0.475
[0.05,0.5]	MXL	0.897	0.851
[0.01,0.5]	MXL	0.844	0.777
[0.01,0.05]	MXL	0.735	0.624
[0.05,0.5]	CLM	0.902	0.853
[0.01,0.5]	CLM	0.850	0.784
[0.01,0.05]	CLM	0.745	0.645
[0.05,0.5]	PUR	0.909	0.862
[0.01,0.5]	PUR	0.860	0.797
[0.01,0.05]	PUR	0.772	0.680
[0.05,0.5]	YRI	0.812	0.782
[0.01,0.5]	YRI	0.741	0.706
[0.01,0.05]	YRI	0.643	0.599
[0.05,0.5]	LWK	0.808	0.783
[0.01,0.5]	LWK	0.727	0.695
[0.01,0.05]	LWK	0.636	0.598
[0.05,0.5]	ASW	0.809	0.774
[0.01,0.5]	ASW	0.752	0.708
[0.01,0.05]	ASW	0.659	0.599

Alzheimer's disease and APOE markers in these categories were chosen from a list of variants showing some evidence of association with Alzheimer's disease from a meta-analysis of Alzheimer's association studies [5] and a set of mitochondrial markers suspected to be associated with the disease from an Alzheimer's disease research group. Additionally, a dense set of markers spanning the APOE region is included on this array, including two SNPs, rs429358 and rs7412, which define the APOE isoforms known to be associated with the risk of Alzheimer's disease and other conditions. Both of these SNPs are previously unavailable on any microarray product.

Autoimmune and inflammatory variants were included that show evidence for association with specific autoimmune and inflammatory disorders, including ulcerative colitis, Crohn's disease, type 1 diabetes, Graves' disease, Hashimoto's thyroiditis, and celiac disease.

Blood phenotype markers were chosen from GWAS and candidate gene studies for their association with red blood cell groups, the regulation of formation of red blood cells and platelets, and the regulation of blood homeostasis.

Cancer common variants 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 unpublished cancer-associated SNPs as of June 2013.

Cardiometabolic variants were included that are known to be associated with various cardiometabolic traits, including coronary disease, lipids, anthropometry, glycemic markers, and blood pressure.

eQTL markers were supplied from the GEUVADIS project in advance of their publication [6] together with eQTLs for several other discovery projects, including the MuTHER project, the ALSPAC project, and the GENCORD project.

Fingerprint is a set of SNPs used by the University of Washington and the Broad Institute. These markers are shared among several major genotyping platforms to facilitate sample tracking. The set of Fingerprint markers from the Axiom Biobank Genotyping Array was selected for the UK Biobank Axiom Array.

HLA and killer-cell immunoglobulin-like receptor (KIR) genes in the HLA (chr6) and KIR (chr19) regions are known to be important in immune response; however, they are not easily assayed directly on commercial arrays. Markers have been added to this array to facilitate imputation of the HLA alleles (at the classical Class I and Class II loci) and of KIR genes.

Lung function phenotypes are covered by a set of markers having an established or putative association with lung function, lung disease (asthma, cystic fibrosis, chronic obstructive pulmonary disease, idiopathic pulmonary fibrosis, and lung cancer), and smoking behavior.

Common mitochondrial DNA variants include coding and noncoding mitochondrial DNA (mtDNA) population markers, which can be used as a framework to generate complex phylogenetic networks and conduct mtDNA haplogroup comparisons. Also included on the array are variants with possible associations with the most common mitochondrial disorders.

Neurological disorders markers include a large number of rare mutations with possible association to a variety of neurological diseases, recently identified association hits, and a number of markers chosen from exome sequencing studies. Cohorts of diseases considered include Alzheimer's disease, frontotemporal dementia, progressive supranuclear palsy, amyotrophic lateral sclerosis, and Parkinson's disease.

NHGRI GWAS markers were chosen directly from, or as tags for, markers in the NHGRI Catalog of Published Genome-Wide Association Studies as of January 2013.

Pharmacogenomics/ADME content consists of markers for genetic variants related to absorption, distribution, metabolism, and excretion (ADME) from the Pharmacogenomics Knowledgebase [7] with known relevance to drug metabolism and some markers from the Applied Biosystems™ DMET™ Plus platform. The set of ADME markers from the Axiom Biobank Genotyping Array was selected for the UK Biobank Axiom Array.

Neanderthal ancestry markers aim to tag genetic segments or haplotypes present in the genome of modern Europeans that are thought to be inherited from our Neanderthal ancestors. This small contribution of Neanderthal DNA, still present in modern populations, might be enriched for SNPs with functional effects. Additional markers were 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 define lineages on the male-specific region of the Y chromosome, including markers to identify all of the main branches of the Y phylogeny. Markers that are present but rare in the UK are included to provide information about population frequencies of higher-resolution sub-branches.

Rare variants in categories of disease predisposition genes include markers for rare missense variants in proven cancer predisposition genes and are selected from sources including locus-specific and human disease mutation databases. Markers for cardiac disease predisposition include rare variants in the genes *MYBPC3* and *MYH7*, markers chosen from the ARVD/C locus-specific database [8] to look at variants in the genes *DSC2*, *DSG2*, *DSP*, *JUP*, and *PKP2*, and a set of markers chosen from human disease mutation databases for genes related to cardiac disease and hemochromatosis. Additionally, a set of markers was chosen to allow investigation of the frequency of a set of rare, possibly disease-causing mutations for disorders relevant to lung function, other phenotypes, or that are polymorphic in the Exome Aggregation Consortium (ExAC) data. It is anticipated that analysis of a large unselected population will allow improved stratification into nonpathogenic and pathogenic variants.

Copy number variants (CNVs) cover a set of 67 CNVs, selected for developmental delay phenotypes [9,10]. Markers were added where necessary to ensure dense coverage within the CNV region and to cover breakpoints, where possible. Markers were also added to ensure coverage of a small number of CNVs that 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.

Rare coding variants were selected based on EMAF in the UK and other European populations (using the content from Applied Biosystems™ Axiom™ Exome Genotyping Arrays and exome sequencing experiments). Protein truncating variants (PTV) resulting in premature stop codons, frameshifts, and loss of start codons were chosen, as they are natural candidates to cause loss of function for that gene copy. Other rare coding variants are primarily missense mutations.

Multiallelic markers on the array include 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, and T allele; so specific probes were added to the array for each allele.

Due to the chemistry of the array, multiallelic markers are difficult to interpret and require the development of new algorithms to analyze them. It is recommended that these markers be excluded from standard analyses.

Specifications

Genotyping performance has been evaluated on >380 samples, including samples from the International HapMap Project, against stringent quality control metrics, including average sample call rate, sample concordance, and reproducibility. See Table 4 for performance metrics and validation data.

Table 4. Performance metrics for the UK Biobank Axiom Array.

Metric	Specification	Performance
Number of samples	–	>380
Sample pass rate	≥95%	99.7%
Average call rate	≥99%	99.7%
Reproducibility	≥99.8%	99.9%
Average HapMap concordance	≥99.5%	99.8%

applied biosystems

Imputation-aware marker selection

We utilized proprietary imputation-based marker selection algorithms [1–3] to maximize the genomic coverage for this array. Reference panels from KGP and European ancestry populations coupled with advanced imputation algorithms provide improved genomic coverage over the commonly used pairwise tagging methods alone.

Analysis workflow for UK Biobank Axiom Arrays

The analysis workflow as described in the Axiom Genotyping Solution Data Analysis Guide (P/N 702961) is an advanced analysis technique that provides the greatest flexibility in finding the most informative content for each study. The Analysis Note, UK Biobank Axiom Array (P/N 703267), details additional instructions for customized analysis options and unique markers specific to the UK Biobank Axiom Array.

Acknowledgements

We are very grateful to the UK Biobank Array Design Group and the many scientists and research groups who provided markers and data as part of the design process for the UK Biobank Axiom Array and to the 500,000 UK participants who have generously donated time, medical information, and biological samples to better understand the complex interactions between genes, lifestyle, and environment.

References

1. Hoffman TJ, et al. (2011) Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics* 98(6):422–430.
2. Hoffmann TJ, et al. (2011) Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics* 98(2):79–89.
3. Howie BN, et al. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics* 5(6):e1000529.
4. McVean GA, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56–65.
5. Lambert JC, et al. (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics* 45(12):1452–1458.
6. Lappalainen T, et al. (2013) Transcriptome and genome sequencing uncovers functional variations in humans. *Nature* 501(7468):506–511.
7. The Pharmacogenomics Knowledgebase. http://www.pharmgkb.org/
8. ARVD/C Genetic Variants Database. http://www.arvcdatabase.info/
9. Cooper GM, et al. (2011) A copy number variation morbidity map of developmental delay. Nature Genetics 43(9):838–846.
10. DECIPHER database. http://decipher.sanger.ac.uk

Ordering information

Product	Description	Cat. No.
UK Biobank Axiom Array	Contains one 96-array plate; reagents and GeneTitan Multi-Channel Instrument consumables sold separately	902502
Axiom Biobank Plus Genotyping Array	Contains one customized Biobank array; reagents and GeneTitan Multi-Channel Instrument consumables sold separately	000854
Axiom GeneTitan Consumables Kit	Contains all GeneTitan Multi-Channel Instrument consumables required to process one Axiom array plate	901606
Axiom 2.0 Reagent Kit	Includes all reagents (except isopropanol) for processing 96 DNA samples	901758

Find out more at **thermofisher.com/microarrays**

ThermoFisher SCIENTIFIC