Large scale genomewide association analysis of multiple disease phenotypes: the Wellcome Trust Case Control Consortium

Peter Donnelly for the Wellcome Trust Case Control Consortium

Disease samples

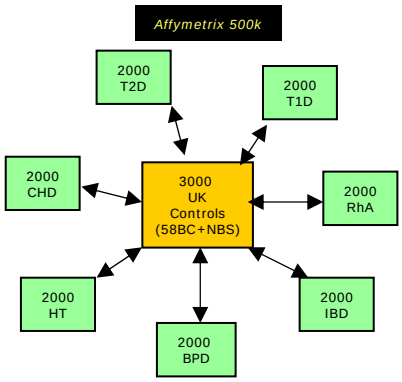
Disease	Co-Principal Applicants	Cohort Abbreviation
Disease cohorts		
Type 1 diabetes	John Todd & David Clayton	T1D
Type 2 diabetes	Mark McCarthy & Andrew Hattersley	T2D
Inflammatory bowel disease	Miles Parkes & Chris Mathew	IBD
Breast cancer	Michael Stratton & Nanzeen Rahmad	BC
Coronary heart disease	Alistair Hall & Nilesh Samani	CHD
Hypertension	Mark Caulfield & Martin Farrall	HT
Bipolar disorder	Nick Craddock	BD
Rheumatoid arthritis	Jane Worthington	RA
Multiple sclerosis	Alastair Compston	MS
Ankylosing spondylitis	Matthew Brown	AS
Autoimmune thyroid disease	Stephen Gough	ATD
Tuberculosis	Adrian Hill, Melanie Newport & Giorgio Sirugo	TB
Control cohorts		
1958 Birth Cohort	Peter Shepherd, Alan Silman, Marcus Pembrey, David Strachan	58BC
National Blood Service	Willem Ouwehand	NBS

AIMS

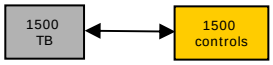
To accelerate efforts to identify variants contributing to susceptibility to diseases of major global importance

To develop and validate informatic and analytical solutions appropriate to the scale and nature of the project

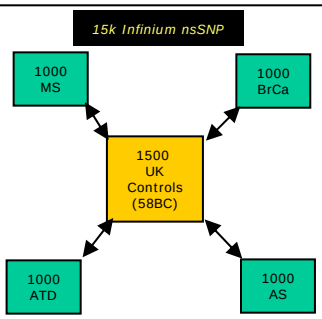
To answer important methodological and biological questions relevant to large-scale association studies



Main study with national cases/controls



Study in Gambian cases and controls



nsSNP experiment 15k Infinium

Sample preparation

- >24,000 DNA samples imported to WTSI
- Requantified and QC at WTSI and JDRF/WT Diabetes and Inflammation Lab (DIL), Cambridge
- African samples (TB) → whole genome amplification
- iPLEX coding at WTSI
- Gender check
- Shipped to California for genotyping

Genotyping

- Affymetrix 500k arrays
- Typed at Affymetrix facility in California
- DM & BRLMM calls in California

Data transfer

- Genotype calls transferred electronically → WTSI
- .CEL files shipped on hard drives → WSTI
- BRLMM & CHIAMO calls in UK

QC and analysis

- Data QC and initial analysis by Analysis Group chaired by Professor David Clayton at the DIL, and Professor Lon Cardon at the Wellcome Trust Centre for Human Genetics (Oxford).
- Disease PIs have access to the genotypic data of their case series and all the corresponding controls.

Progress

- nsSNP study 5500 samples completed
- Main study ~16000 ex 17000 samples genotyped
- TB study - due fall 2007

Key design features

- **Main study** of 7 diseases with common controls
- All cases and controls of UK Europid origin, collected without particular regional focus (ie "national" collections)
- Controls include 1500 individuals from the British Birth Cohort of 1958 and 1500 from a National Blood Service collection
- **Study of TB** uses Gambian cases and controls
- **nsSNP study** includes 1500 BC58 controls and 1000 samples from each of 4 additional diseases. These have been typed on a custom-made Infinium assay which includes 14000 nsSNPs and 1200 tags for the MHC region.
- **See POSTER # 1737 (Deloukas et al) for further details of the nsSNP study**

Some questions the WTCCC will help to address

Technical

- Alternative allele calling methods (see top right)
- Impact of misclassification bias
- Optimal data management
- Optimised QC for large-scale association data
- Optimized analysis of large-scale association data

Analytical

- Comparisons between the two control groups
- Extent of population stratification in UK samples
- Identification of markers informative for structure
- Comparison of alternative approaches for dealing with stratification
- Value of using of other case groups as additional "controls"
- Development of methods for imputing untyped SNPs (cross-platform and beyond)

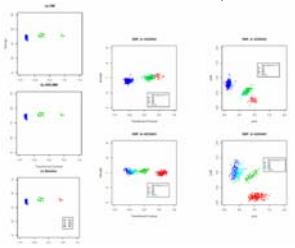
Biological

- Overlap in susceptibility between related diseases
- Role of copy number variation
- Allelic architecture of multiple complex traits
- Disease-gene mapping in European and African samples

Genotype calling

	Pools across individuals?	Uses mismatch info?	Relies on DM calls?
DM	NO	YES	YES
B-RLMM	YES	NO	YES
CHIAMO	YES	YES	NO

Affymetrix **DM** calling inadequate due to preferential heterozygote loss; **BRLMM** calling substantial improvement but can generate erroneous calls for markers where the DM call leads to poorly-calibrated cluster centers and covariance matrices; **CHIAMO** uses scale that improves cluster definition for the 1-2% of SNPs that show poor clustering



		15000 random SNPs			223 cluster shifted SNPs		
		call	conc	overall	Call	Conc	Overal
DM	0.33	97.66	99.08	96.76	94.18	95.03	89.50
BRLMM	0.50	99.51	99.33	98.65	93.08	92.40	86.01
CHIAMO	0.99	99.76	99.15	98.91	98.55	97.82	96.40

Data release

The Consortium anticipates that data generated will be used by others to develop new analytical methods, to understand patterns of variation and to guide selection of markers to map genes involved in specific diseases. **Release of cleaned, raw and summary data from the main and nsSNP studies to qualified investigators is planned 6 months after completion** (i.e. mid 2007). WTCCC generated data on the two control groups will be released before this (late 2006).

Anonymized genotype data (from several calling algorithms) will be made available via a database at the WTSI. This will also provide case/control status, broad geographical origin of the samples, gender and age group (10y intervals). More detailed information on subjects and more comprehensive phenotypic data is held by the respective individual disease and control investigators.

Requests for access will be evaluated by the Consortium Data Access Committee. CDAC (cdac@wellcome.ac.uk) will assess researcher status but not peer-review scientific proposals. Once approved, the researcher will enter into a Data Access Agreement that specifies the terms of access. Users of the data will be required to acknowledge the role of the Consortium and the relevant primary collections and their funders, or the published paper from which the information derives. Users should note that the Consortium bears no responsibility for the further analysis or interpretation of these data, over and above that published by the Consortium.

Principal Investigators

Matthew Brown	Institute of Musculoskeletal Sciences, University of Oxford
Lon Cardon	Wellcome Trust Centre for Human Genetics, Oxford
Mark Caulfield	William Harvey Research Institute, London
David Clayton	JDRF/WT Diabetes and Inflammation Laboratory, Cambridge Institute Medical Research
Alastair Compston	Department of Clinical Neurosciences University of Cambridge
Nick Craddock	Department of Psychological Medicine, University of Wales College of Medicine
Panos Deloukas	The Wellcome Trust Sanger Institute, Cambridge
Peter Donnelly	Department of Statistics, University of Oxford
Martin Farrall	The Wellcome Trust Centre for Human Genetics, Oxford
Stephen Gough	University of Birmingham
Alistair Hall	Institute for Cardiovascular Research, Leeds General Infirmary
Andrew Hattersley	Diabetes and Vascular Medicine, Peninsula Medical School
Adrian Hill	The Wellcome Trust Centre for Human Genetics, Oxford
Dominic Kwiatkowski	The Wellcome Trust Centre for Human Genetics, Oxford
Mark McCarthy	Oxford Centre for Diabetes, Endocrinology and Metabolism (OCDEM)
Christopher Mathew	Department of Medical and Molecular Genetics, Guy's Hospital, London
Willem Ouwehand	Haematology, University of Cambridge & National Blood Service
Miles Parkes	Gastroenterology Unit, Addenbrooke's Hospital, Cambridge
Marcus Pembrey	ALSPAC Director of Genetics
Nazneen Rahman	Institute of Cancer Research
Nilesh Samani	Department of Cardiovascular Sciences, University of Leicester
Michael Stratton	The Wellcome Trust Sanger Institute, Cambridge
John Todd	University of Cambridge
Jane Worthington	School of Epidemiology & Health Sciences, The University of Manchester
David Strachan	St George's Hospital, Medical School

Acknowledgements

DNA: Sarah Nutland, Pamela Whittaker, Sue Bumpstead; **Affymetrix**; **Data**: Neil Walker, Simon Potter, Sarah Hunt, Jonathan Marchini, Jeff Barrett, YY Teo, David Evans, Mike Inouye, Ralph McGinnis, Rob Lawrence, Andrew Morris; Disease investigators and their teams (**T2D**: Ele Zeggini, Will Rayner, Kate Elliott, Mike Weedon, Tim Frayling, Hanna Lango; **BC58**: Sue Ring, Wendy McArdle, Richard Jones, David Strachan; **HT**: Pat Munroe, Anna Dominiczak, John Connell, Morris Brown); Audrey Duncanson (Wellcome Trust)

wellcome trust