

The Human genetics programme employs cutting-edge methods of genome analysis to study genetic variation in the human genome, with the goals of understanding how genetic variation impacts health and disease, and how human evolution has shaped this variation. In parallel with experimental approaches, which ultimately aim to provide insights into the biological processes underlying human disease, the programme also develops software tools that underpin the analysis of these large genetic datasets.

- Statistical genetics
- Statistical and computational genetics
- Metabolic diseases
- Chromosome rearrangements
- Genetics of complex traits
- 1000 Genomes Project
- Genomic mutation and genetic disease
- Genetics of common neurological disorders
- Genetic epidemiology
- Genetics of quantitative variation
- Human evolution
- Applied statistical genetics



Kate Whitley, Wellcome Images

### Genome-wide association studies

Genome-wide association studies, many involving Sanger Institute researchers, have uncovered numerous loci contributing to disease susceptibility, and relevant phenotypic traits, and provided biological insights into complex diseases. To increase the power of these studies, international consortia have been created to pool data and jointly analyse very large datasets. Sanger Institute researchers have contributed to and led many such international efforts, which have identified numerous novel loci, including:

- 17 type 2 diabetes risk loci and 26 loci influencing glycaemic traits

- 22 loci affecting blood cell-related traits and a novel locus impacting heart disease and autoimmunity
- the first reported genetic association, on chromosome 8q, with migraine with aura
- three novel loci for ulcerative colitis, which have suggested that loss of integrity of the intestinal wall may be an important disease mechanism
- one of the first examples of a low-frequency susceptibility locus affecting a complex trait (rheumatoid arthritis) detected by pooling data across rare variants
- 95 loci affecting blood lipids, shedding new light on lipid metabolism
- 18 new loci associated with body mass index and fat distribution

## Faculty members

*Inês Barroso, Joint acting Head*  
*Richard Durbin, Joint acting Head*  
*Carl Anderson*  
*Jeff Barrett*  
*Nigel Carter*  
*Panos Deloukas*  
*Matthew Hurles*  
*Aarno Palotie*  
*Manj Sandhu*  
*Nicole Soranzo*  
*Chris Tyler-Smith*  
*Eleftheria Zeggini*

To look more deeply at likely candidate variants influencing disease risk, and to provide increased resolution to identify the variants underlying association signals (fine-mapping), investigators have begun to develop custom-made chips, each with approximately 200 000 variants. Two of our research teams have helped to develop an 'ImmunoChip', which is being tested on 150 000 patients worldwide with autoimmune disease, while 'MetaboChip' data from thousands of participants with cardiometabolic phenotypes are now being analysed at the Institute.

Analysis of genetic data also depends on novel analytical tools and software. Important contributions from Institute scientists include methods to analyse data across genes and multiple phenotypes simultaneously; Evoker software, which provides easy access to genotype data; and the analysis of pooled rare variants across a locus.

## The roots of diabetes

### Type 2 diabetes is a highly complex disease, but large-scale genetic studies are providing important insight into its origins.

Type 2 diabetes, one of the most common and fastest growing diseases of the modern world, stems from a failure to adequately regulate blood glucose levels. Up to 2010, large-scale genetic studies had identified around 20 loci affecting risk of diabetes. During 2010, Sanger Institute researchers played key roles in international consortia that identified 17 additional susceptibility loci, as well as genetic influences on physiological traits relevant to diabetes.

Studies performed by the DIAGRAM and MAGIC consortia, involving more than 100 000 individuals, assessed the contribution of genetic variants on disease risk and on physiological factors relevant to diabetes. Seventeen loci influencing diabetes risk were identified. Consistent with previous results, these variants affected mostly pancreatic beta cell function, confirming the importance of genetic influences on insulin production from beta cells as an important aspect of diabetes.

MAGIC investigators also assessed genetic influences on the regulation of blood sugar levels after an oral intake of glucose. Three loci were shown to influence this process; of particular biological interest was the association with the *GIPR* gene. This gene encodes a receptor for a hormone, GIP, which is secreted from the gut in response to food intake and acts on the pancreas to increase insulin secretion.

Glycated (glucose-bonded) haemoglobin levels are influenced by blood sugar levels and have been used to monitor, and recently proposed to diagnose, diabetes. Studies performed by MAGIC revealed ten loci associated with levels of glycated haemoglobin, of which seven appeared to not exert their effects via glucose effects. Nevertheless, the percentage of people that would be incorrectly diagnosed as having diabetes on the basis of their glycated haemoglobin levels is very small, supporting its use as a diagnostic.

These studies provide insight into genetic factors influencing the biological mechanisms that regulate blood sugar levels and diabetes risk. Some of these loci, or the pathways that they regulate, may be the target of further studies to discover possible new therapeutic approaches for diabetes.

Voight BF et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature Genet* 2010; 42(7):579–89.

Dupuis J et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genet* 2010; 42(2):105–16.

Saxena R et al. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nature Genet* 2010; 42(2):142–8.

Soranzo N et al. Common variants at 10 genomic loci influence hemoglobin A1(C) levels via glyemic and nonglycemic pathways. *Diabetes* 2010; 59(12):3229–39.



Wellcome Images

Recently it has been proposed that many of the association signals identified via genome-wide studies could arise from chance clustering of rare variants, the so-called 'synthetic association' model.

Although this model is theoretically possible, our analyses suggest that it is not likely to account for much of the missing heritability, nor is it likely to account for most of the reported signals.

### Impact of human variation in health and disease

In addition to single nucleotide variants, the human genome also contains structural variants, larger stretches of DNA that are either deleted or multiplied in different individuals. In the past year, our researchers have led efforts to develop detailed maps of structural variation in diverse global populations. Association studies suggest that structural variants are unlikely to contribute

significantly to the heritability of common diseases, although rare structural variants can cause severe early-onset obesity and other conditions.

An important development this year has been the funding of the Deciphering Developmental Disorders (DDD) project. Identification of the mutation underlying a set of symptoms is important as it provides a diagnosis for patients and their families. The DDD project aims to identify such mutations in 12 000 UK children and translate diagnostic findings into clinical practice. Results will be shared with the wider scientific and clinical community via the DECIPHER database.

Data being generated by the 1000 Genomes Project with contributions from several Sanger Institute groups, have provided a largely unbiased dataset of human genetic

## 1000 Genomes – the pilot

### Even in its pilot phase, the 1000 Genomes Project has already provided researchers with a valuable resource.

The 1000 Genomes Project, an international project jointly led by Richard Durbin at the Sanger Institute and David Altshuler from the Broad Institute, with contributions from several Sanger Institute groups, is generating a view of human genetic variation in unprecedented detail. It will provide a much more complete and consistent picture of human genetic variation than ever before, creating a resource of fundamental value to researchers studying genetic influences on health.

Its pilot phase comprised three complementary projects, which used next-generation sequencing technologies to gather data from several global populations. Collectively, the projects generated 4.9 terabases (4900 billion bases) of sequence information and documented more than 95 per cent of the currently identifiable variants present in any individual. Eventually this will expand to more than 99 per cent of variants.

One finding so far is that, on average, each person carries 250–300 genetic changes that disable a gene, as well as 50–100 genetic variations previously associated with disease. Sequencing of families revealed that each child had approximately 60 new mutations not present in either parent.

As data are freely released, researchers can immediately begin to use the newly identified variants and assess their linkage to disease and disease traits. In addition, when association with a region of the genome has been identified, the data provide a more comprehensive list of variants that could be responsible for the observed association.

Perhaps most significantly, though, the first phase validated the methods developed for acquiring and analysing genome-wide sequence data on many individuals, in preparation for the main analysis of some 2500 genomes.

This preparatory work has also paved the way for an even more ambitious study: the £10.5m UK10K Project, funded through a Wellcome Trust Strategic Award made jointly to the Sanger Institute and multiple UK clinical collaborators. The UK10K Project will analyse 10 000 genomes over the next three years – 4000 from healthy individuals in the Avon Longitudinal Study of Parents and Children, ALSPAC, and TwinsUK cohorts and 6000 from patients with conditions thought to have a genetic origin. Such intensive sequencing should reveal even extremely rare variants contributing to disease processes.

1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467(7319):1061–73. PMID: 20981092



Genome Research Limited

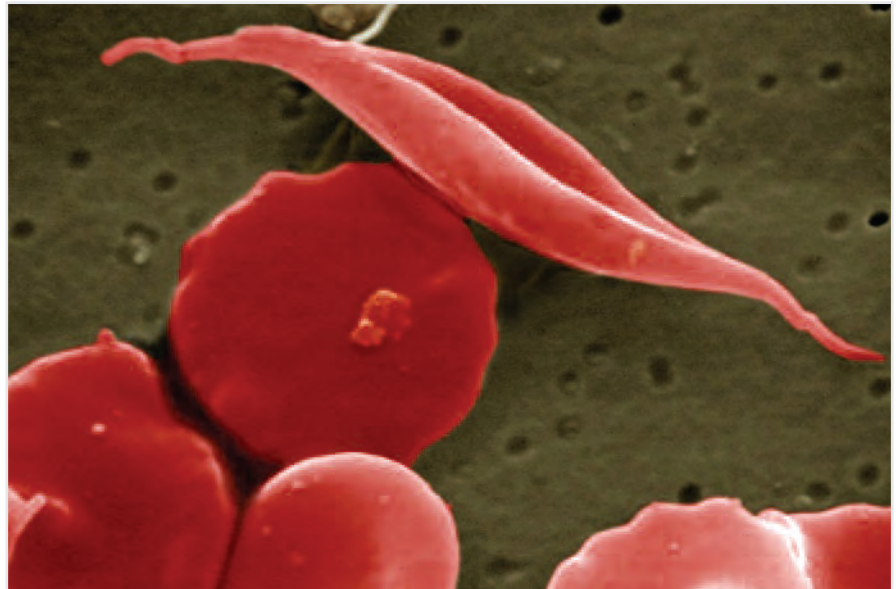
The three-year UK10K Project will build on the work of the 1000 Genomes Project by analysing 10 000 genomes to enable increased understanding of the link between low-frequency and rare genetic changes with human disease.

variation in diverse populations. Three pilot projects generated more than 15 million genetic variants and provided valuable data to validate methods for identification of variants. New variants have already been included in the ImmunoChip and MetaboChip.

As well as its impact on health, human genetic variation also contains important information about the evolutionary forces that have shaped our genomes. Early data from the 1000 Genomes Project have fed into the largest study to date of the impact of positive selection on the genome. This work has also revealed that a surprisingly high number of genes – approximately 7 per cent – are inactive in one or more participating individual.

**Future directions**

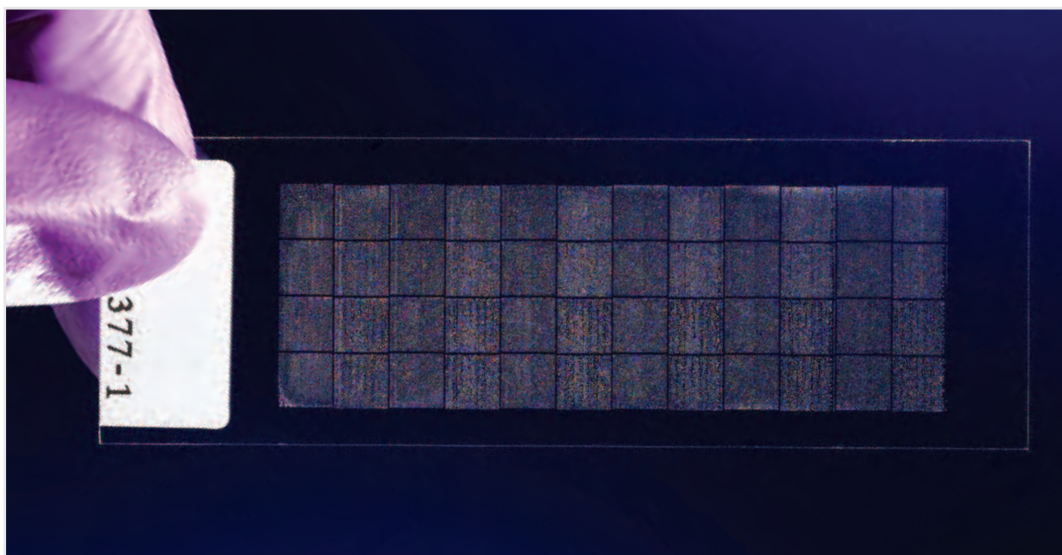
Despite the successes of genome-wide studies, for most complex human diseases a substantial proportion of heritability remains unexplained. Initiatives such as the UK10K Project with its focus on whole-genome and whole-exome re-sequencing, will provide essential data to assess the role of rare variants in the missing heritability. Equally important will be analytical tools to identify and analyse rare variants from both custom arrays (such as ImmunoChip and MetaboChip) and sequence data.



EM Unit, UCL Medical School, Royal Free Campus/Wellcome Images

Most genetic studies to date have focused on populations of European ancestry. In the coming years increasing efforts will be placed on building partnerships in Africa to understand genetic diversity in this continent, as well as to study diseases (for example sickle cell disease and non-communicable disease) impacting the health of its peoples. A key role of these partnerships will be to strengthen research infrastructure and the capacity to undertake genomics studies in Africa.

Sickled cell and red blood cells. We will be seeking to build partnerships in Africa to strengthen genomic research into the genetic basis of sickle cell disease and other non-communicable diseases.



Wellcome Images

The wealth of data that will be generated by using ImmunoChip and MetaboChip custom genotyping arrays will require the Human genetics programme to develop new methodologies capable of analysing the rare variants associated with disease.

## We are developing statistical tools for the analysis of high-throughput genetic datasets and applying them to large cohorts to elucidate the mechanisms underlying common human disease.

Genome-wide association studies (GWAS) have identified many loci underlying common human disease. In spite of these successes, for most complex human diseases, a substantial proportion of heritability remains unexplained. We are developing and applying methods to identify the rare variants that are believed to underlie some of this missing heritability.

One way of increasing power to detect association is by imputation of ungenotyped markers using densely genotyped reference panels, derived from HapMap or 1000 Genomes Project data. We quantified the power to detect association following imputation using several different reference panels and found that HapMap 2 is better powered than HapMap 3 for the detection of rare variant associations. We also quantified the gain in power that imputation with 1000 Genomes Project data will provide, and demonstrated that the increase will be substantial.

The proportion of heritability that is explained by rare variants is related to the genetic architecture of a trait. The recently proposed 'synthetic association' model suggests that many of the association signals identified via GWAS could arise from chance clustering of rare variants. In collaboration with statistical genetics groups within the Human genetics department, we showed that this model, while theoretically possible, is unlikely to account for many GWAS signals and probably does not explain much missing heritability.



Professor P. Motta & G. Familiari, Science Photo Library

Coloured scanning electron micrograph of collagen fibres (green) underlying the surface of an endometriotic ovarian cyst. Endometriosis occurs when fragments of the uterine lining become attached to other sites in the body. We have discovered a locus that affects the severity of this disease.

We conducted a meta-analysis of two GWAS of endometriosis, a common gynaecological disease where endometrial-like tissue grows in sites outside the womb. This study, the largest genetic analysis of the disease to date, identified one locus that significantly increases risk of disease and its severity. We also showed that severe disease is more heritable than mild disease.

Next year we will be focusing on developing tools to identify rare variant genotypes and applying these tools to bespoke genotyping arrays, such as the 'ImmunoChip', to identify loci underlying autoimmune diseases. We will also be conducting large-scale meta-analyses and sequencing experiments aimed at identifying rare disease-associated variants, with a focus on immune-related traits such as Inflammatory Bowel Disease, primary biliary cirrhosis and primary sclerosing cholangitis.



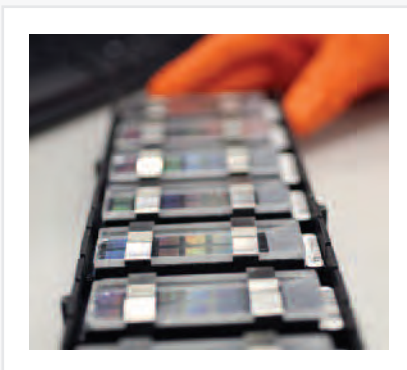
Genome Research Limited

Several lines of evidence suggest that most genome-wide association signals are driven by common causal variants.

Anderson CA et al. Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS Biol* 2011; 9(1):e1000580

Meta-analysis identifies a locus on chromosome 7 contributing to endometriosis.

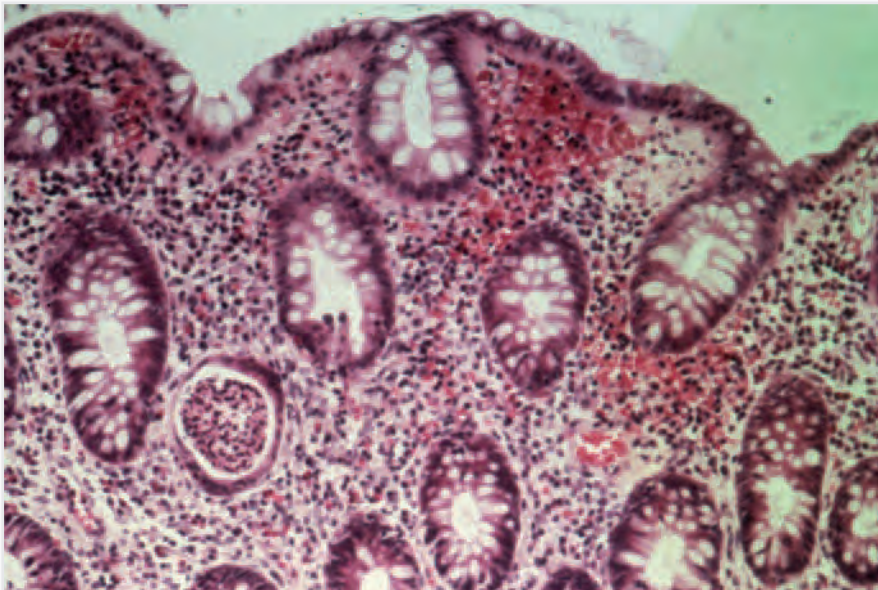
Painter JN et al. Genome-wide association study identifies a locus at 7p15.2 associated with the development of endometriosis. *Nat Genet* 2011; 43(1):51-4



Genome Research Limited

Bespoke genotyping arrays, such as 'ImmunoChip', should facilitate the identification of loci associated with autoimmune diseases.

**We combine increasingly detailed surveys of genetic variation with new statistical techniques to uncover how this variation contributes to human health and disease.**



Wellcome Images

Light micrograph section through the colon, showing ulcerative colitis. The condition presents as an inflammation of the colon characterised by deep ulcers, which can cause diarrhoea, severe pain, and require surgery in extreme cases. Genome-wide association studies of affected individuals have linked three new loci with intestinal epithelial wall integrity.

Genome-wide association studies (GWAS) have generated huge amounts of data and yielded many insights into the biology of human disease. We take data from these existing studies, drawn from many related phenotypes (especially those related to inflammatory conditions), and combine them with each other and with public datasets (including 1000 Genomes Project data). We also develop new tools and methods to maximise the power of these studies.

A central focus of our work recently has been development of the 'ImmunoChip' (in collaboration with Panos Deloukas). The 'ImmunoChip' includes some 200 000 SNPs drawn from regions of the genome found to be associated with a dozen autoimmune and inflammatory diseases in GWAS, as well as extensive lists of unconfirmed loci, submitted by a global network of researchers. The chip will be genotyped on 150 000 individuals worldwide, generating data for future analyses of autoimmunity.

Bringing these data together requires new methods of analysis. We have developed a technique that combines information across

genes (rather than treating each variant independently) and allows us to consider multiple distinct phenotypes simultaneously. This approach is more powerful for detecting genes that have either multiple causal variants or subtly different effects in different diseases.

In order to maximise the value of this methodological work, we also develop tools for public release. The recently released 'Evoker', for example, provides rapid and easy access to genotype data from GWAS and similar projects, and is designed to minimise false-positive associations. Evoker is distributed as an open-source project, enabling the community to obtain maximum benefits and extend its functionality.

The avalanche of data relating to genetic variation and human health continues to grow, through use of experimental approaches such as the 'ImmunoChip' and next-generation sequencing. The coming challenge will be to use computational tools and new methodologies to translate these data into deeper understanding of disease biology.



Wellcome Library, London

➤ **A new open-source tool for visualising genotype data, designed to filter out false positive associations.**

Morris JA et al. Evoker: a visualization tool for genotype intensity data. *Bioinformatics* 2010; 26(14):1786-7.

➤ **Genome-wide scan identifies three new ulcerative colitis loci and suggests that loss of integrity of the intestinal epithelial wall may be contributing to disease.**

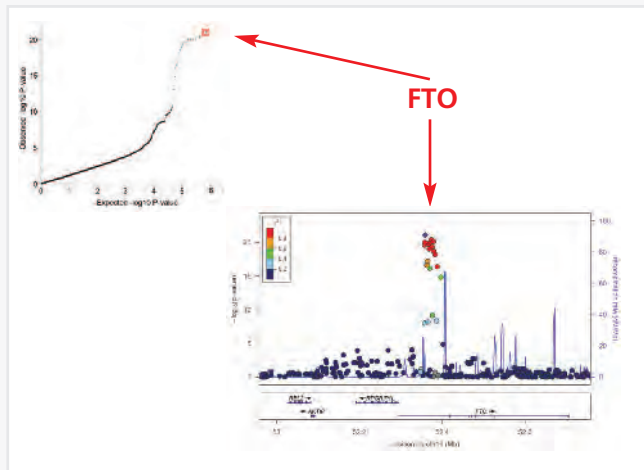
UK IBD Genetics Consortium et al. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the *HNF4A* region. *Nat Genet* 2009; 41(12):1330-34.



stock.xchng

We are committed to making our tools open source so they will be as useful as possible to as many people as possible as quickly as possible.

We are searching for loci influencing risk of type 2 diabetes, obesity and related traits and, using whole-exome sequencing, identifying mutations underlying extreme forms of metabolic disease.



Unpublished data, Genome Research Limited

Plot of the *FTO* locus, previously shown to associate with body mass index in the general population, here associated with extreme childhood obesity.

As part of our continued efforts to identify loci influencing risk of type 2 diabetes, obesity and underlying quantitative traits, we have participated in and led large-scale international collaborations that have identified more than 20 novel loci influencing diabetes risk (within MAGIC and DIAGRAM+ consortia) or related glycaemic traits (such as fasting and post-challenge glucose levels).

Following on from these results, we have begun looking for associations between risk loci and particular aspects of beta cell function and other biological processes relevant to diabetes.

We have also been part of the multinational GIANT consortium, which has identified more than 30 loci influencing body mass index and fat distribution.

Currently, we are analysing data obtained with the 'Metabochip', a custom-made chip with around 200 000 variants concentrated at loci showing some evidence of association with various cardiometabolic traits. This targeted approach will facilitate further replication and fine-mapping of susceptibility loci.

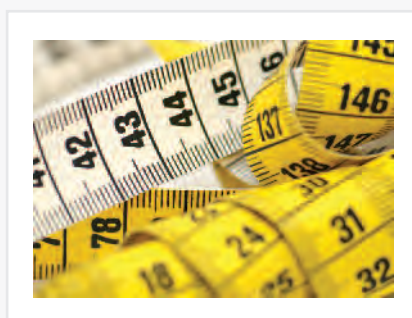
Results from first phase genome-wide association studies have identified several loci predisposing to severe early-onset childhood obesity, including some loci already seen in population-based studies. Other associations have not been seen before, and appear specifically to increase

the risk of extreme early-onset obesity. We are now attempting to replicate these findings in other populations.

To identify rare mutations causing severe early-onset childhood obesity, we are sequencing the exomes of affected individuals as part of the UK10K project.

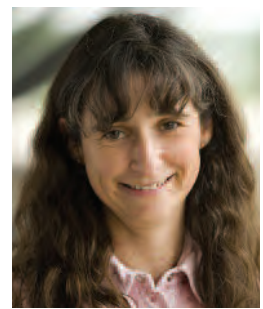
This year whole-exome sequencing of patients with syndromes of insulin resistance has begun to yield fruit, while also highlighting some of the challenges of the approach, particularly in demonstrating that the mutations identified are actually causing disease.

In the forthcoming year large-scale sequencing approaches are likely to become the main focus of the group, along with preliminary exploration of the biological consequences of the genetic changes identified.



stock.xcimg

At least 30 genetic variants affecting body mass index and fat distribution have now been found.



Wellcome Library, London

A meta-analysis of 21 studies identifies five loci affecting risk of type 2 diabetes among 16 influencing control of blood glucose and beta cell function and two linked to insulin resistance.

Dupuis J et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet* 2010; 42:105–16.

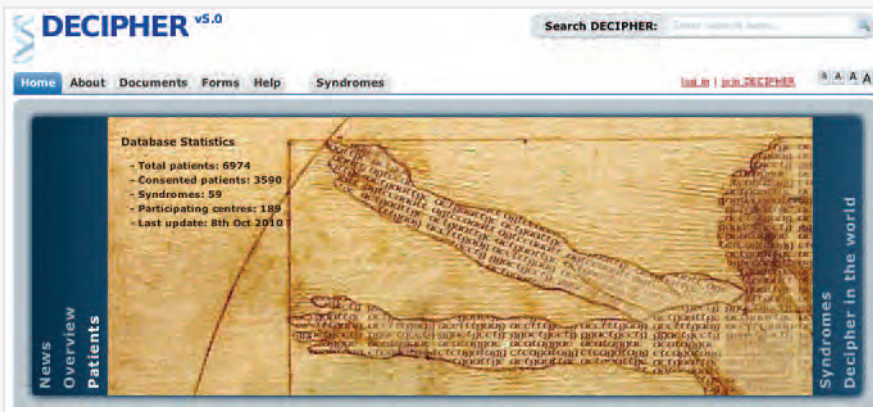
The DIAGRAM+ international consortium identifies 12 novel risk loci for type 2 diabetes.

Voight BF et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 2010; 42: 579–589.

Large international study confirms 14 known loci and identifies 18 new sites associated with body mass index.

Speliotes EK et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* 2010; 42(11):937–48.

Our aim is to investigate the role of genome variants in human developmental disease and to translate advances in research technology into clinical practice.



DECIPHER

The DECIPHER database contains information about chromosomal changes in more than 8000 individuals with developmental disorders. These data have been gathered from 200 participating clinical genetics centres worldwide.

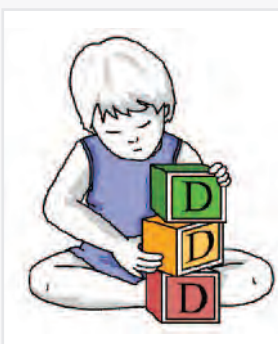
A key development during the past year has been the funding of a major UK-wide research project, Deciphering Developmental Disorders (the DDD project). Every year in the UK, thousands of babies are born with errors in their genetic makeup that affect their development. Currently, only a small minority of children – less than 3 per cent – receive a diagnosis, based on a clinician's ability to interpret the appearance of a child and the pattern of his or her symptoms, supplemented by the use of microscopes to identify large chromosomal rearrangements.

The latest molecular testing methods identify previously undetectable changes in chromosomes, enabling new diagnoses to be made in 20 per cent of children. Next-generation sequencing technologies could improve this figure still further. However, clinical use is hampered by the limited availability and inconsistent application of these technologies, and by lack of basic knowledge to link genetic changes directly to symptoms.

The DDD project, funded from the Health Innovation Challenge Fund and from institutional funds, brings together the comprehensive clinical resources of the 23 UK Regional Genetics Services with the expertise in high-throughput genomic technologies, data interpretation and dissemination at the Wellcome Trust Sanger Institute.

In the DDD project, we will apply the latest molecular testing technologies to 12 000 UK children with developmental disorders. Results will be shared with the wider clinical community through the DECIPHER database. These results will provide a unique online catalogue of genetic changes and their associated symptoms, to help clinicians make more precise diagnoses.

Furthermore, we will design more efficient and cheaper diagnostic assays for relevant genetic testing to be offered to all such patients in the UK, thereby transforming clinical diagnosis for children with developmental disorders.



Deciphering Developmental Disorders

Only 3 per cent of children with developmental disorders currently receive a definitive diagnosis. The Deciphering Developmental Disorders (DDD) project aims to use next-generation sequencing to drastically improve this figure.



Wellcome Library, London

### The most detailed map to date of structural variation in the human genome.

Conrad DF et al. Origins and functional impact of copy number variation in the human genome. *Nature* 2010; 464(7289):704–12.

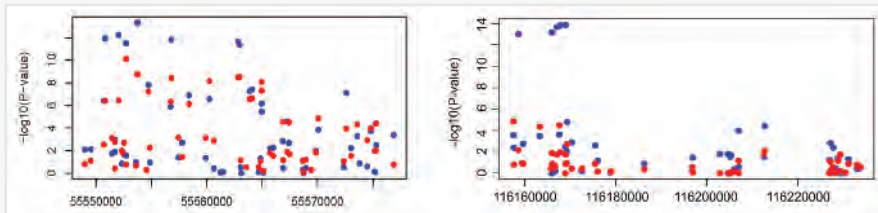
### New method for calling absolute copy number variants uncovers common CNVs in Asian populations.

Park H et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat Genet* 2010; 42(5):400-5. Epub 2010 Apr 4

### A literature review confirms that microarrays are far superior to microscopy-based methods for identifying chromosomal abnormalities in clinical diagnosis.

Miller DT et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am J Hum Genet* 2010; 86(5):749–64. Review.

We are investigating the molecular basis of complex traits in humans, in particular coronary artery disease and its main complication, myocardial infarction, alongside quantitative traits linked to established risk factors (e.g. lipid levels).



Genetic architecture can vary between populations, as shown by association signals at the *CETP* (left) and *APOE5* (right) loci associated with high-density lipoprotein cholesterol and triglycerides, respectively, in Pakistanis (blue) and Europeans (red).

Saleheen D et al. *Circ Cardiovasc Genet.* 2010; 3(4):348–57  
doi:10.1161/CIRCGENETICS.109.966780

Large-scale meta-analyses have identified substantial numbers of loci underlying anthropometric and other quantitative traits, such as blood lipid levels and haematological variables.

Through the GIANT consortium, we identified 180 loci influencing adult height, while the Global Lipids Genetics Consortium found evidence for 95 loci affecting blood lipid levels. Having identified 22 loci for haematological traits, the HaemGen consortium is assessing platelet-related phenotypes in more than 50 000 individuals. We identified 31 loci affecting coronary artery disease in two large meta-analyses involving 22 000 and 15 000 patients (CARDIoGRAM and C4G, the Coronary Artery Disease Genetics, respectively).

The results suggest that genetic predisposition to coronary artery disease is mediated in part by multiple common genetic variants of small to moderate effect size, many of which appear to act independently of traditional risk factors (e.g. lipids). Interestingly, several loci also affect other complex traits.

Most association signals from genome-wide scans need to be refined to identify variants for functional evaluation. In addition, meta-analyses have yielded many loci with moderate evidence of association. To dissect potential risk loci, we have worked with international collaborators to develop custom tools for large-scale genetic studies of cardiometabolic traits and autoimmune disease phenotypes – the ‘Metabochip’ and ‘ImmunoChip’, each containing 200 000 markers. For coronary artery disease, a new initiative, CARDIoGRAM-Plus, is validating a set of 6400 markers in a two-stage analysis involving 50 000 patients in total.

We are also using experimental techniques to characterise protein binding in regions of association. We have found that a common variant at the *PIK3CG* locus, associated with mean platelet volume, alters a transcription factor-binding site. Differential binding at the risk allele is associated with higher *PIK3CG* expression.

We are also assessing gene expression in multiple tissues to determine whether association signals reflect regulatory variation. We have also begun to characterise methylation patterns in regions of association, to identify further possible sources of regulatory variation.



Wellcome Library, London

### Several genetic variants found in European populations also affect blood lipid levels in Pakistanis.

Saleheen D et al. Genetic determinants of major blood lipids in Pakistanis compared with Europeans. *Circ Cardiovasc Genet.* 2010; 3(4):348–57.

### Meta-analysis reveals loci previously implicated in lipid level variation as well as others never previously linked to lipoprotein metabolism.

Teslovich TM et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 2010; 466(7307): 707–13.

### Meta-analysis identifies 22 loci affecting blood-cell characteristics relevant to health, and a chromosomal region linked to heart disease.

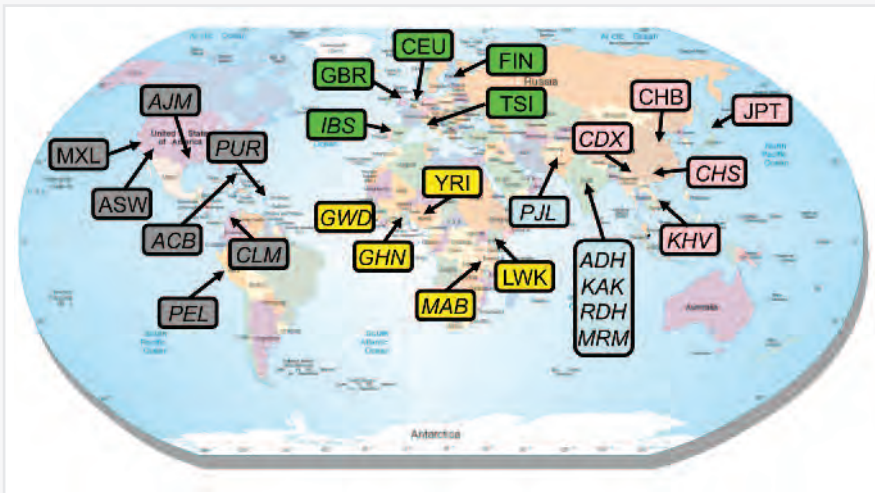
Soranzo N et al. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat Genet* 2009; 41(11):1182–90.



Annie Cavanagh, Wellcome Images

Using experimental techniques, we have found that a common variant at the *PIK3CG* locus, associated with mean platelet volume, alters a transcription factor-binding site.

The 1000 Genomes Project, an international consortium jointly led by the Sanger Institute, is producing a detailed catalogue of single nucleotide and structural genetic variation in multiple human populations.



Genome Research Limited

Building on the three pilot studies, the 1000 Genomes Project has been extended to include 27 populations across the globe (26 are shown above).

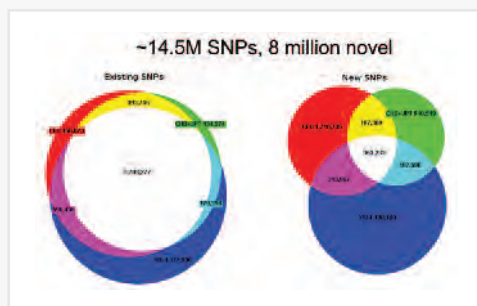
Current genetic methods such as genome-wide association studies are based on only partial knowledge of genetic variation in humans. With the advent of new sequencing technologies, it is now possible to sequence the whole genomes of thousands of people to provide a much more complete baseline resource. This is the goal of the 1000 Genomes Project, to which multiple groups at the Sanger Institute are contributing.

During 2009/10 we published the results from three pilot projects. First, 179 people from three population groups (European, West African and East Asian ancestry) were sequenced genome-wide at low coverage. Although this approach does not allow every variant present in an individual to be identified, by combining information across participants we were able to identify almost all variants present in five or more people, and say with reasonable accuracy which individuals contain which variants. Deeper sequencing was carried out on six individuals genome-wide and 900 genes in 697 people.

In all we found approximately 14.5 million single nucleotide polymorphisms (SNPs) and more than one million other genetic variants, most of which had not been seen before. The results validate the project design and

computational methods we have developed, and have already created a unique resource for medical and evolutionary genetics research.

Beyond the pilot, the Project has now sequenced more than 1100 individuals at, on average, over 4x depth, generating around ten times as much sequence data as in the pilots. An initial analysis of approximately half these data has revealed 10 million additional variants, the overwhelming majority of them novel. We have also finalised the main project sampling design, and will sequence 2500 people in total from 27 populations in five population groups, both at low coverage genome-wide and at high depth in the protein-coding exome. Many new projects at the Sanger Institute and elsewhere are already using both the data from the Project and the methods developed for it.



Genome Research Limited

The 1000 Genomes Project low coverage pilot found approximately 14.5 million genetic variants, many of them novel.



Wellcome Library, London

**Collaborating Faculty**

Jeff Barrett  
Matthew Hurles  
Aarno Palotie  
Chris Tyler-Smith

➤ **The results of three pilot projects, validating methods and identifying more than 15 million genetic variants.**

1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* 2010; 467:1061–73.

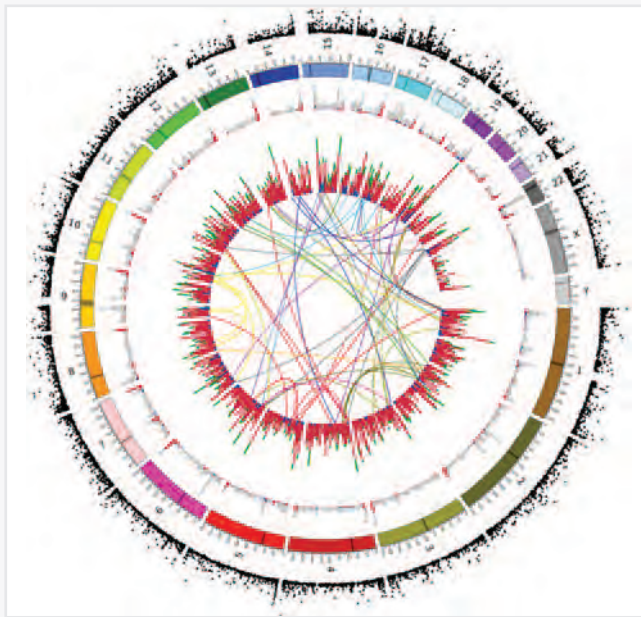
➤ **A new method developed to identify short insertions and deletions (indels) in the 1000 Genomes Project pilots.**

Albers CA et al. Dindel: Accurate indel calls from short-read data. *Genome Res* 2010 Oct 27, doi: 10.1101/gr.112326.110.

➤ **A method for accurately calling variants from low-coverage sequencing data in population samples.**

Le SQ, Durbin R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res* 2010 Oct 27, doi: 10.1101/gr.113084.110.

Our aim is to decipher the genetic architecture of rare genetic diseases and to quantify the contribution of the tens of new genomic mutations that are introduced every generation.



Conrad DF et al. *Nature* 2010; 464(7289):704–12  
doi:10.1038/nature08516

Comprehensive catalogue of copy number variants: The innermost circle shows lines connecting the origin and the new location of 58 putative interchromosomal duplications. The next circle out shows a stacked histogram representing the number of deletions (red), duplications (green) and multiallelic (blue) loci. Chromosomes 1-22, X and Y are represented in the outermost circle.

We develop novel experimental and analytical methods to characterise genetic variation in families and populations, and to identify disease-causing variants in patients. Such studies shed light on the mutational mechanisms underlying human genetic variation and their contributions to disease. Gaining a deeper understanding of population variation increases our ability to identify disease-causing variants and to improve molecular diagnosis of genetic diseases.

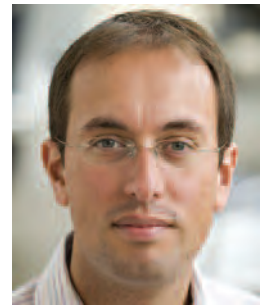
The past year has seen the publication of the results of several large projects in high-impact journals, including six *Nature* papers. These include the comprehensive characterisation by the Genome Structural Variation consortium of common copy number variants in European and African populations at an unprecedented resolution, the Wellcome Trust Case Control Consortium's analysis of common copy number variants and susceptibility to eight common diseases, the International HapMap 3 consortium's map of sequence and structural variation in over 1000 individuals drawn from diverse global populations, and, in collaboration with Sadaf Farooqi's group in Cambridge, our identification of rare copy number variants causing severe early-onset obesity.

Another notable study is the bioinformatic prediction of the subset of genes for which having both copies functional is critical for healthy development. These predictions are already being used to improve clinical interpretation of rare chromosomal deletions, through the DECIPHER browser.

We have also played a major role in analysing the data generated during the pilot phase of the 1000 Genomes Project, leading the analyses of structural variation, and gaining new insights into sex-specific base substitution mutation rates.

Finally, we have investigated the value of exome sequencing for elucidating the genetic causes of rare diseases, to inform new large-scale collaborative studies such as the UK10K project and the Deciphering Developmental Disorders project.

During the next five years we plan to expand our focus from structural variation to incorporate all forms of genetic variation made accessible by new sequencing technologies, and to translate our better understanding into improved clinical practices.



Wellcome Library, London

Common copy number variants are unlikely to contribute in a major way to genetic susceptibility to eight common diseases.

Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* 2010; 464(7289):713–20.

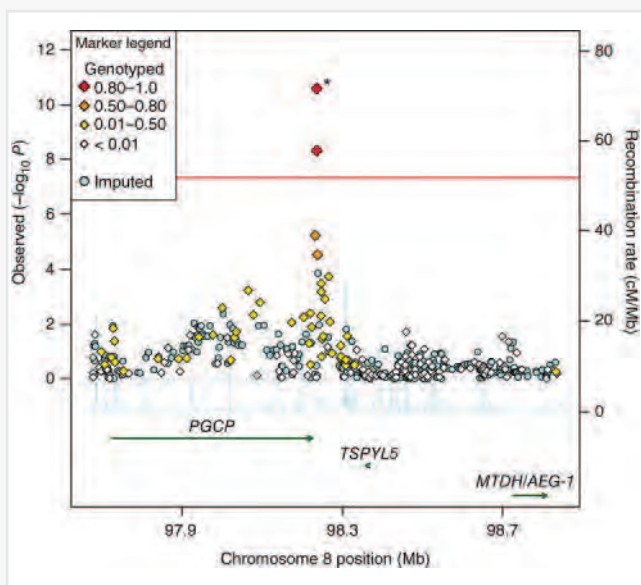
A comprehensive catalogue of copy number variants, their impact on genes, the mechanisms that generate them and their distribution among global populations.

Conrad DF et al. Origins and functional impact of copy number variation in the human genome. *Nature* 2010; 464(7289):704–12.

Rare chromosomal deletions are a cause of severe early-onset obesity in some patients, especially those with developmental delay.

Bochukova EG et al. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* 2010; 463(7281):666–70.

Our goal is to improve understanding of the pathogenetic mechanisms underlying neurological diseases such as migraine, epilepsy, schizophrenia and autism.



International Headache Genetics Consortium et al. Nat Genet 2010; 42: 869-73  
doi:10.1038/ng.652

Association of two loci on chromosome 8q22.1 with migraine. The red line denotes the threshold for genome-wide significance.

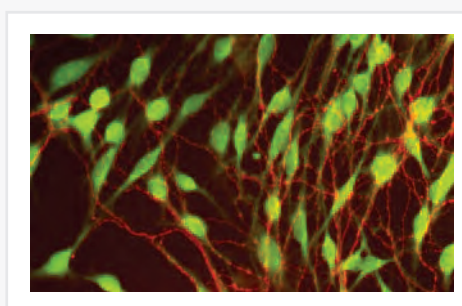
We aim to improve the understanding of the genetic architecture of neurodevelopmental disorders by combining genome-wide information of both single nucleotide polymorphism and structural variants across classical diagnostic phenotypes.

Much of our work draws on clinical and population-based data from Finland. The country's long-established and relatively isolated population, shaped by famines, internal migrations and recent explosive growth, along with its excellent health infrastructure and strong traditions in epidemiological research, make it a particularly interesting location for genetic epidemiology.

We recently published results from a genome-wide association study of migraine with aura, in collaboration with six major headache research centres in Europe and Australia. This study identified a susceptibility variant on chromosome 8q that is potentially linked to glutamate neurotransmitter regulation. We are now following up this initial result in different migraine subtypes, migraine sufferers from population cohorts and individuals suffering from chronic pain.

To identify potential high-penetrance variants predisposing to familial forms of migraine and epilepsy, we have begun whole-exome sequencing of 2000 familial cases. This multicentre collaboration should greatly improve our understanding of the genetic architecture underlying these nervous system disorders.

Unique populations have also shed light on possible cellular differences in autism. By combining data from genome-wide association studies and expression profiles of peripheral blood leukocytes from autism patients from an isolated region of Finland, we identified two biological pathways associated with autism: nervous system development and cell-to-cell signalling and interaction. Whole-exome sequencing of individuals from these families is now being carried out as part of the UK10K project.



A. J. Irving, Wellcome Images

Glutamate receptors in the brain: a locus gene on chromosome 8, associated with glutamate regulation, is the first to be linked to migraine susceptibility.



## ➤ Identification of the first gene variant associated with common forms of migraine.

International Headache Genetics Consortium et al. Genome-wide association study of migraine implicates a common susceptibility variant on 8q22.1. Nat Genet 2010; 42: 869-73.

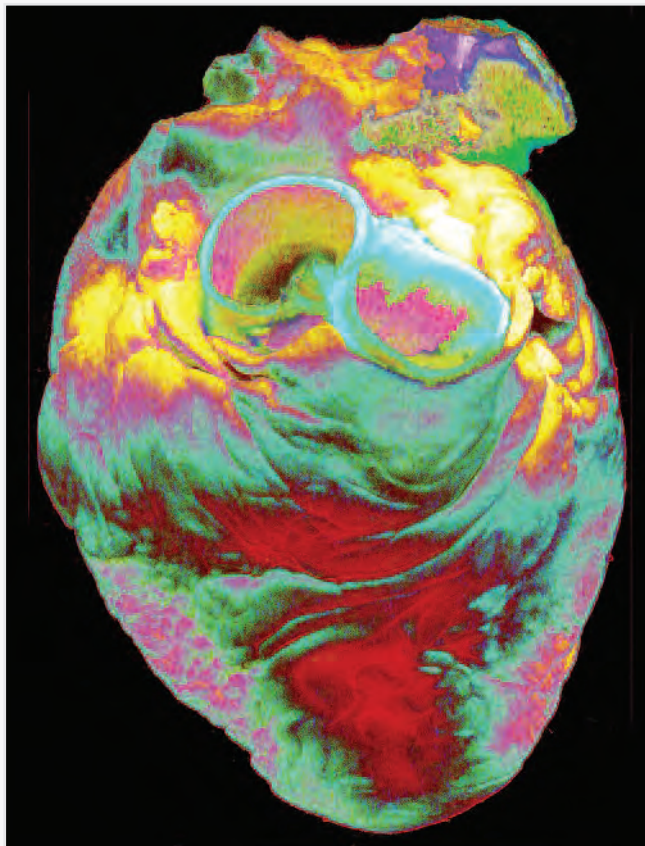
## ➤ Analysis of longitudinal birth cohort data reveals that variants in the LIN28B gene affect different phases of growth.

Widén E et al. Distinct variants at LIN28B influence growth in height from birth to adulthood. Am J Hum Genet 2010; 86(5):773-82.

## ➤ Structural genomic variants affect the risk of complications after haematopoietic stem cell transplantation.

McCarroll SA et al. Donor-recipient mismatch for common gene deletion polymorphisms in graft-versus-host disease. Nat Genet 2009; 41(12):1341-4.

We are exploring genomic diversity and its impact on disease risk factors among populations, to gain new insights into disease aetiology, prevention and treatment.



Gordon Museum, Wellcome Images

Non-communicable diseases such as heart disease and diabetes are rapidly becoming major killers in developing countries.

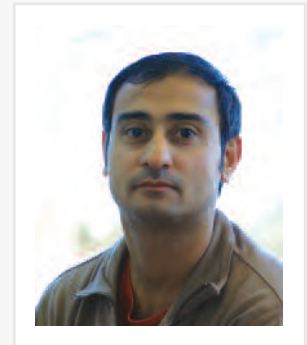
We are working to understand the relation between different blood lipids and the risk of developing heart disease. Over the past year, we have been using genome-wide approaches to identify novel, common genetic loci that explain variation in blood lipid levels. We are now using a combination of whole-genome and exome sequencing of individuals from the global human population to locate more accurately the variants responsible for these differences, and to identify rarer variants that may shed light on the biological role of these loci in lipid regulation and risk of heart disease. We have also applied statistical techniques (such as Mendelian randomisation) to assess whether these, or other biomarkers, are causal risk factors for diseases such as heart disease and diabetes.

We are also studying the roles of infectious and non-infectious risk factors in determining the patterns of non-communicable diseases in populations across sub-Saharan Africa.

These new population-based studies will help support public health planning, including implementation of preventative and control strategies in this region for diseases such as heart disease and type 2 diabetes.

In parallel, studies of the marked genomic diversity across sub-Saharan Africa will provide new opportunities to understand the aetiology of these diseases. In collaboration with colleagues in Uganda (Medical Research Council/Ugandan Virus Research Institute Unit), Malawi (Malawi-Liverpool-Wellcome Trust Research Unit), and other centres across sub-Saharan Africa, we are developing large-scale population-based studies to assess the burden and aetiology of non-communicable diseases throughout the region.

These studies will be underpinned by a partnership aimed at strengthening research infrastructure and the capacity to undertake studies of the genomics of non-communicable diseases in sub-Saharan Africa.



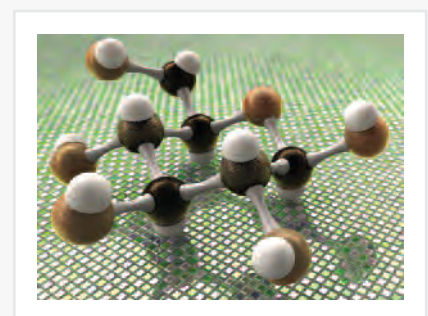
Genome Research Limited

A genome-wide association study in more than 100 000 individuals identifies 95 loci affecting blood lipids and provides insight into the biological processes underlying lipid metabolism.

Teslovich TM et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 2010; 466:707–13.

Variation in the gene encoding the phospholipid transfer protein is associated with higher levels of HDL-cholesterol and lower risk of heart disease.

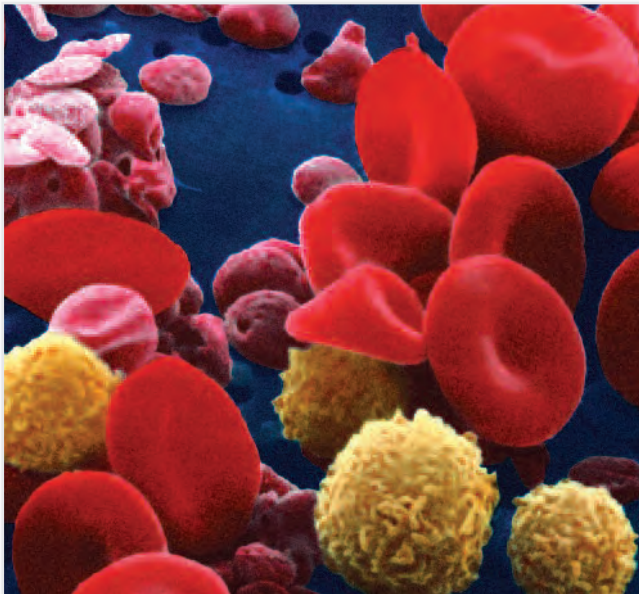
Vergeer M et al. Genetic variation at the phospholipid transfer protein locus affects its activity and high-density lipoprotein size and is a novel marker of cardiovascular disease susceptibility. *Circulation* 2010; 122:470–7.



Anna Tarnczos, Wellcome Images

Glucose molecule. High blood glucose levels experienced in type 2 diabetes raise the risk of heart disease. Studying the role of genetic factors in the patterns of type 2 diabetes across sub-Saharan Africa may help in understanding the causes of this disease.

**We study the population and statistical genetics of quantitative traits that are risk factors for cardiometabolic disease, aiming to link genetic and clinical findings to evolutionary and functional evidence.**



National Cancer Institute, Science Photo Library

Red blood cells, white cells (leucocytes) and platelets. Using a combination of genetic analyses and functional investigations in model organisms, we are working with colleagues in the HaemGen Consortium to understand the mechanisms of blood cell differentiation.

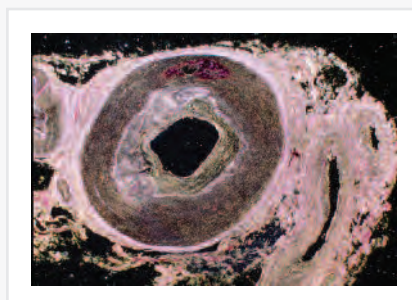
Over the past 12 months we have contributed to large-scale international meta-analyses of genome-wide data for a broad range of quantitative traits that are risk factors for cardiometabolic diseases such as type 2 diabetes and myocardial infarction. This has led to the discovery of more than 200 novel loci associated with diverse traits such as obesity, glycaemic levels and several blood biomarkers, including inflammatory and haematological traits. We have also assessed the association of these loci with disease risk, to identify novel disease risk factors and to shed light on physiological mechanisms underlying disease risk.

For some traits, we are seeking to explore in greater depth the biological mechanisms underlying novel genetic associations. In collaboration with colleagues in the HaemGen Consortium, we are investigating the mechanisms of blood cell differentiation (haematopoiesis), using a combination of genetic analysis, surveys of gene expression and functional assays in model organisms such as zebrafish and fruit fly.

In parallel with this work on established biomarkers, we are exploring the value of novel biomarkers for cardiometabolic disease. Using high-throughput technologies, we are

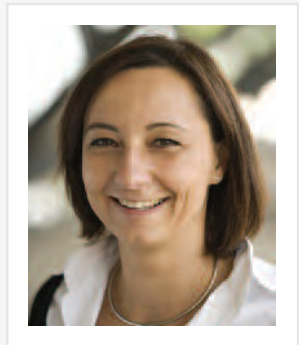
trying to understand the genetic basis of human biochemical individuality, for example by identifying genetic loci associated with variation in cellular metabolite levels. Furthermore, we are using statistical methods for causal inference to identify the contribution these traits make to established biomarkers and disease.

Finally, we are beginning to explore the contribution of intermediate- to low-frequency variants identified through large-scale sequencing of phenotyped individuals from cohort studies. This information, coupled with our current understanding of common variants, should contribute to a more comprehensive picture of genetic predisposition to cardiometabolic disease as well as phenotypic variability in healthy individuals.



Eye of Science, Science Photo Library

Scanning electron microscopy of artery with atherosclerosis. Genetic studies have now linked more than 200 loci to traits relevant to cardiometabolic disease.



Wellcome Library, London

➤ **As well as 22 loci affecting blood cell-related traits, a meta-analysis identifies a novel risk locus for heart disease and autoimmunity that shows signs of having undergone positive selection.**

Soranzo N et al. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat Genet* 2009; 41(11):1182–90.

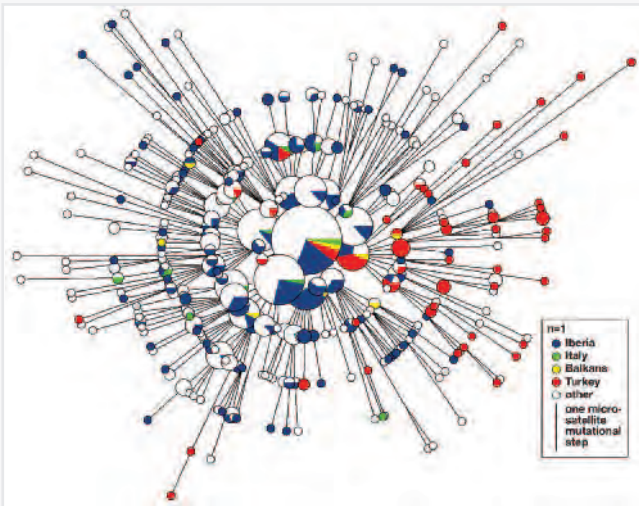
➤ **The first comprehensive survey of genetic variation affecting variation in small molecule metabolite levels.**

Illig T et al. A genome-wide perspective of genetic variation in human metabolism. *Nat Genet* 2010; 42(2):137–41.

➤ **A large collaborative effort identifies ten genetic determinants of glycated haemoglobin levels, a marker for type 2 diabetes.**

Soranzo N et al. Common variants at ten genomic loci influence hemoglobin A<sub>1c</sub> levels via glycaemic and non-glycaemic pathways. *Diabetes* 2010; 59(12):3229–39

Our aim is to understand human genetic history, the patterns of variation and signatures of natural selection in our genome, and their implications for health and disease.



Balaresque P et al. PLoS Biol 2010; 8:e1000285  
doi:10.1371/journal.pbio.1000285

Y chromosome lineage. There is a single origin (centre of network) for this lineage in the Near East (red), and the time depth is less than 11 000 years. Circles represent different haplotypes, with area proportional to frequency and coloured according to population, showing that farmers entered Europe at beginning of the Neolithic and replaced indigenous populations, just as the carriers of advanced technologies have always done.

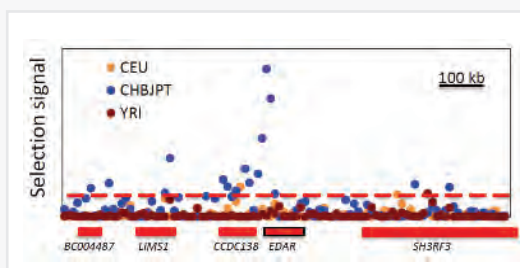
Human genetic variation carries extensive information about our evolutionary history. Using new sequencing technologies, we can now generate large quantities of raw data, and begin to interpret its evolutionary and functional significance. This analysis has been greatly aided by the international 1000 Genomes Project, which for the first time has produced genome-wide nearly unbiased data on genetic variation on a population scale.

One output from this work is a comprehensive view of neutral markers – mitochondrial DNA and the Y chromosome – which can be used to investigate female-specific and male-specific aspects of genetic history, respectively. Variation is apparent even within a single individual's mitochondrial DNA (heteroplasmy), while the first truly representative Y chromosome tree reveals a striking expansion of one lineage, R1b, in Europeans. An independent piece of work has linked this expansion to the arrival of Neolithic farmers, the R1b lineage replacing earlier Paleolithic Y chromosomes.

The 1000 Genomes Project has also provided the most comprehensive survey yet of positive selection in the human genome. Thousands of selected sites show up, often with unanticipated levels of complexity, which will be highly challenging to interpret.

The data also provide a first glimpse of the full extent of variation in functional gene content. More than 1400 genes (7 per cent of the total) are inactive in one or more of the participating individuals (members of the general public). Between them, the participants also carry 671 variants implicated in inherited Mendelian disorders.

Over the next year, we plan to analyse 1000 Genomes Project data in more depth. This work will provide a more detailed view of how human history has shaped present-day genetic variation in Europe, East Asia and Africa, and the first view in the Americas. We will also extend these studies to additional populations. In a collaboration with UCL, we have begun sequencing Native American genomes to compare with genomic data from Americans participating in the 1000 Genomes Project, and we hope it will be possible to begin sequencing individuals from South Asia as well.



Genome Research Limited

Signals of positive selection in the human genome. Red boxes represent genes in 1Mb of chromosome 2. The peak of blue dots reveals a strong selection signal in the EDAR gene in East Asians.



Wellcome Library, London

The first direct measurement of the base substitution mutation rate in humans reveals that roughly one mutation occurs each time the Y chromosome is transmitted.

Xue Y et al. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Curr Biol* 2009; 19:1453–7.

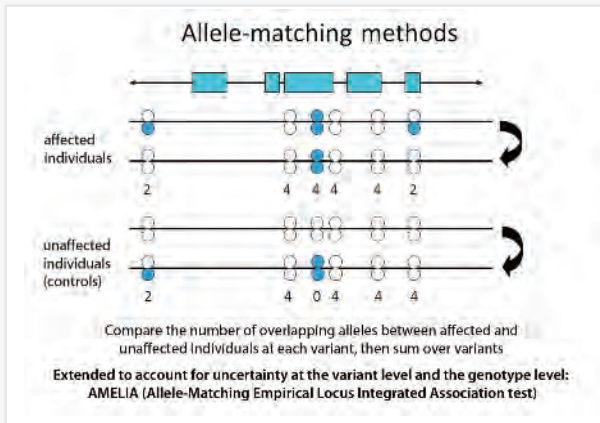
Most European men descend from the first farmers arriving from the Near East after 10 000 years ago.

Balaresque P et al. A predominantly Neolithic origin for European paternal lineages. *PLoS Biol* 2010; 8:e1000285.

The most detailed map of human copy number variation thus far, including over 11 000 CNVs.

Conrad DF et al. Origins and functional impact of copy number variation in the human genome. *Nature* 2010; 464:704–12.

Our group develops strategies and conducts large-scale association and resequencing studies to identify novel genetic determinants of complex traits.



Genome Research Limited

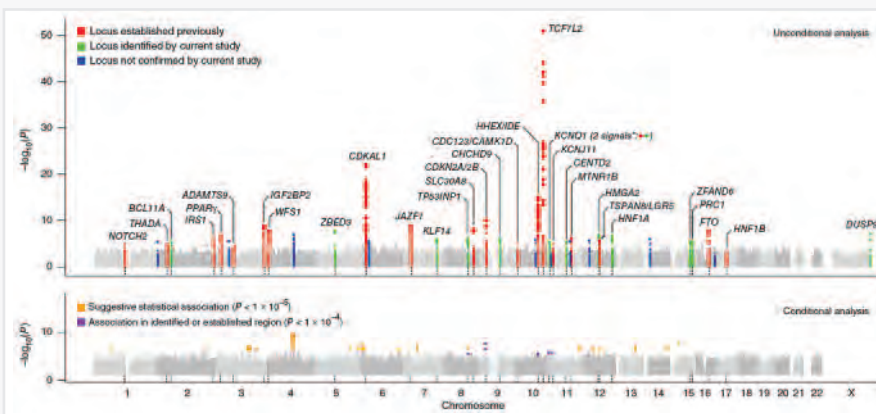
Novel statistical techniques can be used to identify rare genetic variants affecting human health.

Our overarching aim is to elucidate the genetic architecture of complex traits and thereby shed light on the origins and mechanisms of human disease. We are using high-throughput genotyping and next-generation sequencing technologies on large cohorts, in conjunction with publicly available catalogues of human genetic variation, to identify variants affecting a wide range of traits. These include metabolic, anthropometric, musculoskeletal and cardiovascular phenotypes.

We evaluate study design strategies and develop analytical methods, for example to enable results to be compared across different ethnic groups and to facilitate large-scale meta-analyses. We have also established new deeply phenotyped sample collections from population isolates. Much of our work is carried out as part of large and widely distributed collaborator networks, and we place an emphasis on hosting visits and training colleagues.

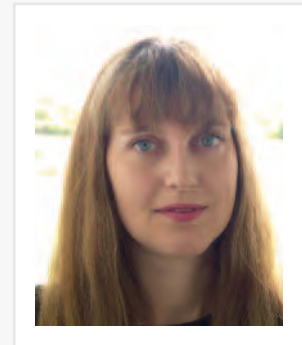
We have led the analysis of more than 10 international genome-wide meta-analyses, leading to the identification of novel genetic associations underlying diverse traits, including 12 new type 2 diabetes variants. We have developed and evaluated several methods for analysing rare variants in complex traits and have identified susceptibility loci containing low-frequency polymorphisms, for example *TNFAIP3* in rheumatoid arthritis.

Looking forward, we are establishing strategies for next-generation genetic association studies using sequence data (for example through the UK10K project). In addition, we are gearing up to make use of the high degree of genetic heterogeneity (and lower degree of inter-marker correlation) in African populations that can help dissect complex traits. We are also embarking on studies of isolated populations, in which the identification of disease loci is empowered by the fact that rare risk variants may have risen in frequency, coupled with typically high levels of linkage disequilibrium.



Voight BF et al. Nat Genet 2010; 42:579-89 doi:10.1038/ng.609

Analysis of data from more than 140 000 people has revealed 12 new genetic (green) variants linked to type 2 diabetes.



Genome Research Limited

One of the first examples of data pooling on rare variants ('variant collapsing') to detect a low-frequency susceptibility locus affecting rheumatoid arthritis.

Bowes J et al. Rare variation at the *TNFAIP3* locus and susceptibility to rheumatoid arthritis. *Hum Genet* 2010; 128(6):627-33.

International collaboration (DIAGRAM+ consortium) culminating in the identification of 12 novel and robustly replicating loci for type 2 diabetes.

Voight BF et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 2010; 42:579-89.

Statistical methods for the analysis of rare variants in complex traits.

Asimit J, Zeggini E. Rare variant association analysis methods for complex traits. *Annu Rev Genet* 2010; 44:293-308. Review.