

The human genetics programme focuses on the genetic basis of common diseases and other important biological traits. A better understanding of the nature and origins of human genetic variation is allowing us to probe ever more deeply into the genetic basis of disease susceptibility. As well as pointing the way to possible new treatments and novel diagnostic approaches, such work is also shedding light on human evolution.

- **Statistical and computational genetics**
- **Metabolic diseases**
- **Chromosome rearrangements**
- **Genetics of complex traits**
- **Cancer Genome Project**
- **Genome dynamics and evolution**
- **Structural Variation Project**
- **Genetics of common neurological diseases**
- **Genome-wide profiling of common diseases**
- **X-linked mental retardation & diseases**
- **Human evolution**
- **Applied statistical genetics**
- **Medical sequencing**
- **Gorilla Genome Project**

Genome-wide association studies

This year has seen a major focus on meta-analyses of genome-wide association studies. Pooling data from multiple studies increases the statistical power of such studies and, alongside new statistical methodologies, has greatly increased the number of putative disease genes identified. Sanger Institute researchers have led or contributed to several major data-pooling collaborations exploring a range of physiological factors and disease states, including body mass and glycaemic traits (Inês Barroso), height and haematological traits (Panos Deloukas), and autoimmune disorders such as type 1 diabetes and Crohn's disease (Jeff Barrett).

The Sanger Institute serves as the centralised genotyping pipeline for the 15 disease and other quantitative traits, including reading ability and mathematical skills, being analysed in phase 2 of the Wellcome Trust Case Control Consortium (WTCCC2). Some 90 000 individuals are being genotyped, with data from individual projects being transferred to the Wellcome Trust Centre for Human Genetics in Oxford for centralised analysis as well as to principal investigators of individual studies. Jeff Barrett and Eleftheria Zeggini are contributing to the data analyses of various WTCCC2 projects.

Sanger Institute researchers are playing important roles in several other large-scale international projects. With Mark McCarthy in Oxford, Leena Peltonen is leading the EU-funded ENGAGE project, which has identified 22 genes critical to the regulation of serum lipids in a population cohort. This study has now been expanded to create a 'global lipid consortium' combining numerous cohorts in the USA and Europe. A more comprehensive genetic risk profile is being produced using data from more than 120 000 individuals (see box: Fat chance).

Panos Deloukas has played a central role in the WTCCC and in large international networks aiming to identify genetic risk profiles behind common diseases including coronary artery disease and myocardial infarction. His team is also analysing genome-wide data in well-phenotyped population cohorts (TwinsUK, 1958 British Birth Cohort) and has reported new loci for bone mineral density and height.



Wellcome Library, London

Faculty members

Leena Peltonen, Head
Jeff Barrett
Inês Barroso
Nigel Carter
Panos Deloukas
Andy Futreal
Matthew Hurles
Aarno Palotie
Michael Stratton
Chris Tyler-Smith
Eleftheria Zeggini

Rare disease-associated variants can also be identified through work on specific populations. In a genome-wide association study of a full birth cohort from northern Finland, Leena Peltonen's group identified a rare variant having a major impact on serum lipid levels. Similarly, studies of Finnish populations with a high incidence of disease have identified genes contributing to multiple sclerosis and schizophrenia. In addition, Mike Stratton's recently completed systematic sweep of the X chromosome has revealed a multitude of rare alleles causing mental retardation (see box: The X factor).

Although genome-wide association studies identify specific genetic variation associated with a condition, that variation may not itself be the critical or causative variant. Additional work is needed to identify critical variants, making use of more detailed maps being generated by the 1000 Genomes Project or in some cases targeted resequencing of genomic regions. Functional studies can then begin to reveal the biological role of critical variants, an area in which the Sanger Institute's model organism research can play an important role. The collaboration between Inês Barroso and Derek Stemple, exploring the role of *FTO* in zebrafish, indicates the likely direction this work will take.

Causes and impact of genome variation

As well as assessing the impact of genetic variation, research also focuses on its origins and evolution. Chris Tyler-Smith's work on Y-chromosome (paternal) and mitochondrial (maternal) DNA is providing reliable estimates of single nucleotide polymorphism (SNP) mutation rates in humans. Deep sequence information from large regions is also allowing his team to investigate the genomic regions that have undergone positive selection, leading to a greatly improved understanding of natural selection in human populations. One curious discovery is that an unexpectedly high number of genes contain premature stop codons and thus appear to be dispensable to normal life.

In addition to SNP variation, humans also show copy number variants (CNV), stretches of DNA that have been lost or duplicated during evolution. Thanks to studies spearheaded by Matthew Hurles, Chris Tyler-Smith and Nigel Carter, the Sanger Institute holds a leading position in research in this area. With Manolis Dermitzakis, this group has shown that CNV can affect gene expression, good evidence that it does play a role in human phenotypic variation.

An important advance this year has been a new high-quality, high-resolution map of common CNV in European and African populations, which identifies 20 times more CNVs per genome than previous datasets. This map forms the basis of new-generation arrays for genetic association studies, including WTCCC2 studies. A growing body of evidence suggests that CNV underlies a wide diversity of common human disease phenotypes.

The Hurles group has also developed methods to pinpoint deletion and duplication breakpoints in sequence data, shedding light on their relative mutation rates. Such changes are also found in healthy individuals and this information can be used to develop statistical tools to predict which genome regions are likely to be haploinsufficient, producing data that will aid the clinical interpretation of genomic rearrangements.



Wellcome Photo Library

DNA/protein in bar-coded sample tubes.



Loading samples in the genotyping facility.

Fat chance

Genome-wide studies of population cohorts have uncovered new genes affecting blood lipid levels and risk of heart disease.

Most genome-wide association studies have been carried out on populations affected by particular conditions. An alternative approach is to look at population cohorts, which can be considered a cross-section of the general population. They can provide an unbiased analysis of genes affecting physiological traits, some of which may be linked to disease risk.

Working with a range of international consortia, Leena Peltonen has been using this approach to identify genetic factors affecting blood lipid levels. Using samples from 16 European population cohorts, Prof. Peltonen and colleagues identified 22 loci affecting total cholesterol, low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol and triglycerides, including six novel loci. As older individuals contributed samples – the eldest was from a 104-year-old – variants affecting later-life pathology could be identified.

Interestingly, several of the loci showed sex differences in their effects – including HMGCoA reductase, the target of statins. In terms of overall risk, the complement of genetic risk factors provided a slightly greater predictive ability than conventional risk factors.

Further insight has come from a genome-wide analysis based on the Northern Finland Birth Cohort 1966, representing individuals from the most genetically isolated regions of Finland. As well as confirming previous associations, the study identified nine novel loci, several of which mapped close to metabolic genes. Of particular interest, the study identified a rare variant affecting LDL levels, highlighting how studies of isolated populations can identify rare alleles of high impact.

Even with these new findings, a significant proportion of trait variability remains unaccounted for – suggesting that many more risk loci remain to be discovered.

Aulchenko YS et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet.* 2009 Jan;41(1):47-55. PMID: 19060911

Sabatti C et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet.* 2009 Jan;41(1):35-46. PMID: 19060910

The critical importance of CNVs and other chromosomal rearrangements is also illustrated by their substantial contribution to severe congenital developmental diseases. Nigel Carter has developed comparative genomic hybridisation tools to identify previously undetectable rearrangements in patients with a variety of congenital disorders. More recently his team has begun using high-density commercial oligonucleotide arrays, which can accurately detect small chromosomal aberrations, applicable to prenatal diagnosis and screening.

This area is also where our research has most immediate clinical impact. The DECIPHER database (<https://decipher.sanger.ac.uk>), developed by Nigel Carter and Helen Firth at Addenbrooke's Hospital, Cambridge, collates genome rearrangement and clinical information, enabling clinicians and geneticists worldwide to share and interpret data. DECIPHER currently holds detailed phenotypic and molecular data from over 3000 patients submitted by over 150 groups worldwide.

The X factor

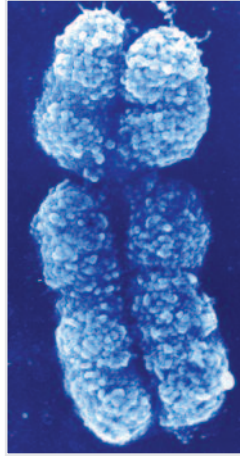
Systematic sequencing of genes on the X chromosome has revealed new genes causing X-linked mental retardation.

Large-scale resequencing has been suggested as a way to identify rare disease-causing mutations in common conditions. As a step towards this goal, Mike Stratton, Andy Futreal and colleagues have carried out systematic resequencing of all the coding exons on the X chromosome in more than 200 families affected by X-linked mental retardation – the largest direct screen carried out to date.

A total of 26 genes were found to carry truncating mutations and nine were identified as likely causes of mental retardation, each being rare or even family-specific.

Although identifying a genetic cause in around 25 per cent of cases, the study highlighted some of the challenges inherent in this approach. An individual's chromosome might contain several variants, and it may not be immediately apparent which (if any) is responsible for the condition. Indeed, around one per cent of X chromosome genes can be lost seemingly without any ill-effects.

Tarpey PS et al. A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nat Genet.* 2009 May;41(5):535-43. PMID: 19377476



Indigo Instruments (www.indigo.com)

X chromosome

A global view of the cancer genome

Through the Cancer Genome Project, the Sanger Institute has played a leading role in large-scale systematic genome-wide searches for somatic mutations in human cancers. The goal is to contribute to a comprehensive cataloguing of cancer genes that drive the abnormal behaviour of cells and also to understand the environmental and other factors that influence the acquisition of somatic mutations.

This year the Cancer Genome Project fully sequenced the exons of 3700 protein-coding genes, identifying *UTX* as a new recessive cancer gene contributing to myeloma and renal and oesophageal cancer. Using next-generation sequencing technology, the project identified over 2000 somatic rearrangements in 24 breast cancers.

Most recently, high coverage sequence has been generated for two cancer genomes, yielding a complete catalogue of all classes of somatic mutations. This study provided the scientific basis for a much larger international collaboration, the International Cancer Genome Consortium, which plans to undertake complete genomic analyses of more than 20 000 human cancers from at least 50 cancer types.

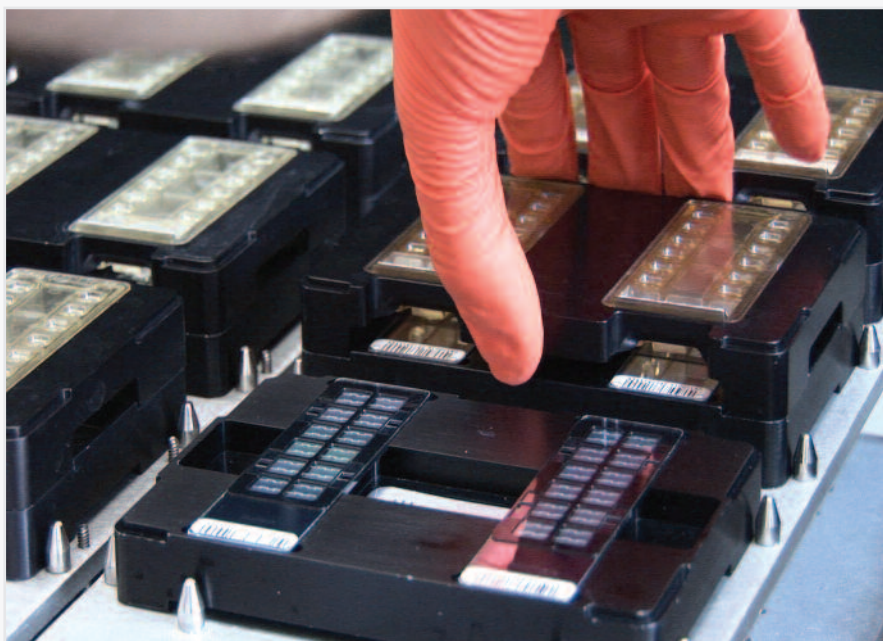
Looking forward

In the next year, the major challenges and opportunities will lie in: (1) development of novel statistical strategies to define critical pathways and networks underlying human diseases, (2) follow up of selected initial gene findings in large population cohorts and by functional studies, (3) the efficient integration of large-scale sequencing datasets from the 1000 Genomes Project and other studies, and (4) the use of model organisms in functional studies of particular disease genes and pathways.

Statistical analysis will benefit greatly from this year's recruitment of two new faculty members, Jeff Barrett and Eleftheria Zeggini. Two junior faculty members have been recruited to career development positions: Carl Anderson is a statistical geneticist and Nicole Soranzo has worked on the genetics of cardiovascular diseases. These recruitments further strengthen human genetics research in strategically critical areas.

We are developing statistical and computational methods and applying them to large datasets to detect disease gene associations and then to dissect the signals to understand how they influence disease biology.

Jeff Barrett joined the Sanger Institute in November 2008.



Loading slides in the genotyping facility.

Taking advantage of the huge quantities of new genetic association data requires large-scale computational and statistical resources. In particular, pooling of data from multiple studies can increase the statistical power to detect significant signals. Recent successes include meta-analyses of international genome-wide association studies in both Crohn's disease and type 1 diabetes, which have each yielded 20 new associated regions. Large-scale studies such as these have opened up new pathways and potential mechanisms underlying these traits.

Making connections between discoveries from related phenotypes has also helped to reveal similarities and differences among diseases. For instance, our group, along with the Wellcome Trust Centre for Human Genetics in Oxford, is leading the analysis of a genome-wide association study in ulcerative colitis (as part of the second phase of the Wellcome Trust Case Control Consortium), which shares autoimmune aetiology with Crohn's and type 1 diabetes. Data from other autoimmune traits, including psoriasis and multiple sclerosis, will also feed into our overall analyses.



Computational power is essential to transform genetic information into improved health care.

Parallel to ongoing genome-wide association studies we are working on incorporating data from large-scale sequencing projects, including the 1000 Genomes Project, to probe for new types of genetic associations. Whereas genome-wide association studies have focused almost exclusively on common variation, use of the 1000 Genomes Project

as a reference sample for genotype imputation – prediction of ungenotyped SNPs in disease samples based on a more completely characterised reference – should allow us to detect rare alleles contributing to common disease for the first time, enhancing our understanding of the allelic architecture of human disease.



Wellcome Library, London

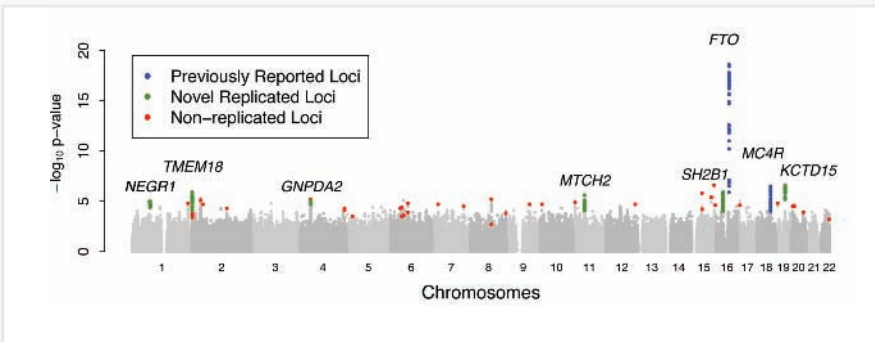
Several interleukin genes are among those identified as risk factors for type 1 diabetes in a genome-wide association study and meta-analysis of two other large studies.

Barrett JC et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet* 6: 703 (2009)

More than 20 new genes affecting risk of Crohn's disease are identified in a meta-analysis of three genome-wide association studies.

Barrett JC et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* 8: 955 (2008)

We are dissecting the genetic aetiology of type 2 diabetes and obesity, closely related metabolic disorders (extreme childhood obesity and syndromes of insulin resistance), and quantitative traits. We are also investigating the function of newly discovered risk loci.



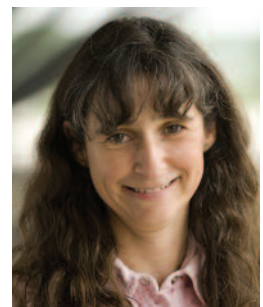
Locations of novel association of six regions of the genome (green circles) with body mass index. Chromosomes are alternating light and dark grey. Known associations (FTO, MC4R) are shown as blue circles. Willer CJ et al. *Nature Genetics* 2009 41(1): 25-34 doi:10.1038/ng.287

During the last 12 months we have focused on meta-analysis of genome-wide data across multiple studies, to increase power, as part of large international consortia. This has directly led to the discovery of nine novel loci associated with body mass and central adiposity (within the GIANT or Genetic Investigation of Anthropometric Traits consortium) and more than 15 loci associated with glycaemic traits, some of which also influence diabetes risk (MAGIC or Meta-Analysis of Glucose and Insulin-related Traits Consortium). Our data have also contributed to the discovery of novel loci associated with height, blood pressure and timing of puberty.

Although genome-wide association studies have provided a wealth of novel loci implicated in type 2 diabetes, obesity risk and related quantitative traits, the mechanisms leading to disease risk generally remain unknown. Furthermore, most of the trait variance remains unexplained, suggesting that additional risk alleles exist. Further studies are required to fine-map signals, identify causal variants, and definitively identify the genes affected by these variants.

We have continued to examine the functional impact of variants on underlying whole-organism physiology. In collaboration with Derek Stemple and Philip Beales (University College London) we have transiently knocked down *FTO* in developing zebrafish embryos and undertaken detailed phenotyping.

Until recently most of our work on extreme phenotypes (childhood obesity and syndromes of insulin resistance) had relied on candidate gene re-sequencing to identify causative mutations. This year we have begun to re-sequence all exons on a limited number of probands with syndromes of insulin resistance, to identify putative causative mutations. These results will inform our future re-sequencing strategy and larger scale experiments.



Wellcome Library, London

A meta-analysis of 15 genome-wide association studies, and follow up in nearly 60 000 individuals, reveals six new loci affecting body mass index, several of which probably act in the central nervous system.

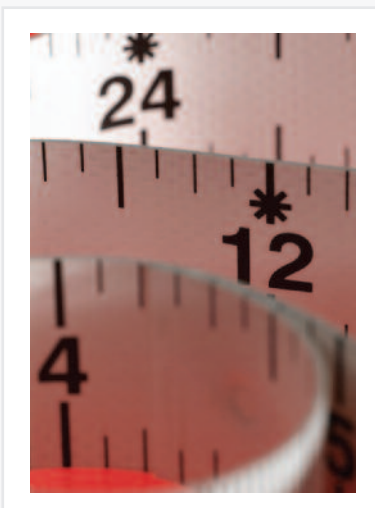
Willer CJ et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet* 41: 25 (2009)

An analysis of 10 genome-wide association studies reveals that variation in the gene encoding melatonin receptor 1B influences blood glucose levels and increases the risk of type 2 diabetes.

Prokopenko I et al. Variants in *MTNR1B* influence fasting glucose levels. *Nat Genet* 41: 77 (2009)

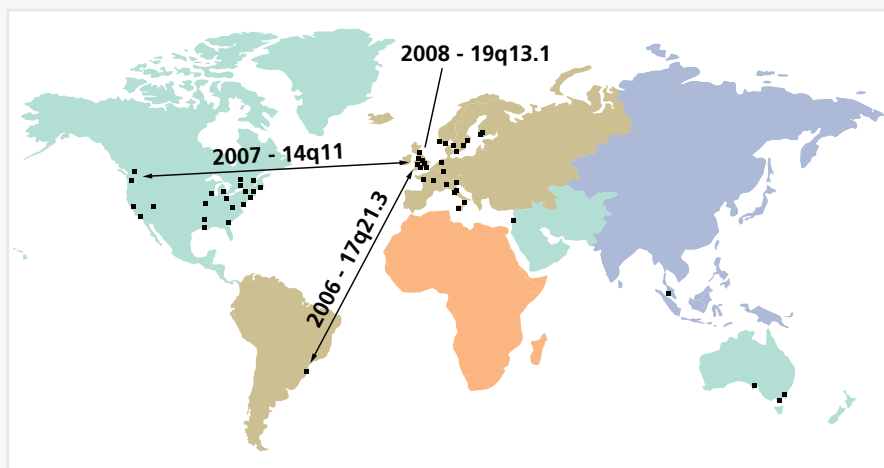
The risk of type 2 diabetes conferred by variation in the promoter region of the *HNF4A* gene differs significantly between UK and Ashkenazi populations.

Barroso I et al. Population-specific risk of type 2 diabetes conferred by *HNF4A* P2 promoter variants: a lesson for replication studies. *Diabetes* 57: 3161 (2008)



Wellcome Library, London

Our aim is to investigate the role of chromosome rearrangement in human developmental disease, moving increasingly from array-based methods to new tools based on next-generation sequencing technologies.



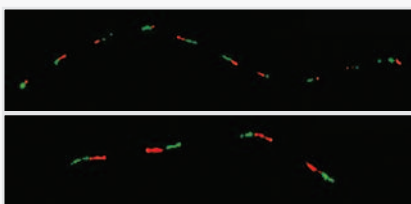
New syndromes defined with the help of DECIPHER. International collaboration centres are represented by black dots.

Our understanding of copy number changes in patients was greatly aided by the first comprehensive map of normal CNV in the human genome (a collaborative project with Matthew Hurles and Chris Tyler-Smith, as well as Charles Lee in Boston and Steve Scherer in Toronto). With data from normal individuals, we can be more certain that observed copy number changes are actually responsible for a patient's phenotype.

We have extended this project to identify smaller CNVs and to include more diverse populations, and have begun an analysis of CNV in great apes. We have also developed microarray assays which allow CNVs to be genotyped in genome-wide association studies.

In collaboration with the group of Philippos Patsalis in Cyprus, we are also assessing whether arrays can be used in non-invasive prenatal diagnosis, based on analysis of free fetal DNA in the maternal circulation. Using high-resolution oligonucleotide arrays and immunoprecipitation specific for methylated DNA, we have identified numerous regions differentially methylated between placenta (fetally derived) and peripheral blood. We are developing methods to enrich fetal DNA from these regions, with a view to using allele-specific assays of chromosome copy number to identify aneuploidy in the fetus from a maternal blood sample.

The DECIPHER database which interacts with the Ensembl genome browser, enables researchers worldwide to share clinical cases, compare their findings with normal chromosome architecture, and define new disease syndromes (as was done for the 17q21.3 and 14q11.2 deletion syndromes). DECIPHER currently holds information on over 3000 patients submitted by more than 150 groups. In the past year, we have added new tools to the database, such as advanced text mining tools to help associate affected genes with particular phenotypic features. As well as improving our knowledge of gene function, such phenotype-genotype associations can also shape patient management. For more information on DECIPHER please visit <https://decipher.sanger.ac.uk>



Copy number variation can be extreme: this individual has ten copies of the amylase 1 gene on one chromosome copy and four on the other. Each copy is marked by a red and a green signal. Perry GH et al. *Nature Genetics* 2007 39(10):1256-60 doi:10.1038/ng2123



Wellcome Library, London

Strategies for whole genome screening and high-resolution breakpoint mapping of copy number changes by array-CGH.

Redon R and Carter NP. Comparative genomic hybridization: microarray design and data interpretation. *Methods Mol Biol* 529: 37 (2009)

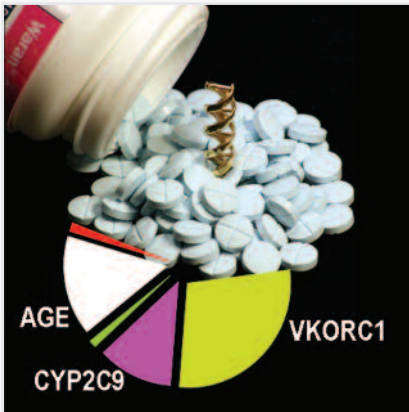
DECIPHER, a database linking chromosomal abnormalities and phenotypic changes in human developmental disorders.

Firth HV et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet* 84: 524 (2009)

Identification of sites on four chromosomes where maternal and fetal DNA is differentially methylated.

Papageorgiou EA et al. Sites of differential DNA methylation between placenta and peripheral blood: molecular markers for noninvasive prenatal diagnosis of aneuploidies. *Am J Pathol* 174: 1609 (2009)

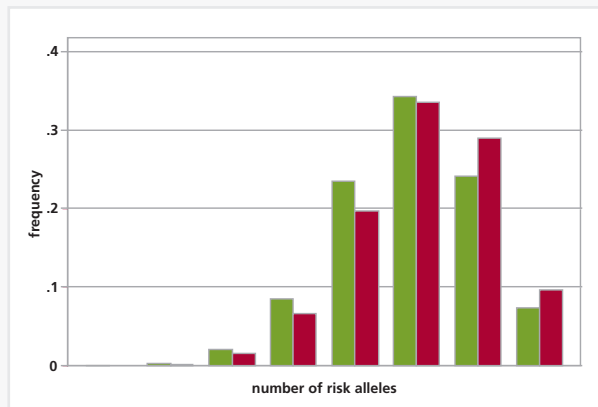
We are investigating the molecular basis of complex traits in humans, in particular coronary artery disease and its main complication, myocardial infarction, as well as relevant cardiovascular traits (e.g. lipid levels).



Genetic and non-genetic predictors of therapeutic warfarin dose.

The availability of multiple genome-wide association studies in both disease and population-based cohorts has triggered large-scale meta-analyses, boosting our ability to detect smaller effects. For example, in a meta-analysis of 19 000 individuals we identified 17 height loci explaining 2.3% of total variation. Through the GIANT (Genetic Investigation of Anthropometric Traits) consortium, we are now assessing adult height in over 100 000 subjects. We have also initiated a new consortium (Haematology Genetics, HaemGen) to look at haematological traits; meta-analyses in 15 000 individuals identified 22 loci reaching genome-wide significance.

Higher numbers of risk alleles for coronary artery disease or heart disease are more commonly found in patients (red) than in apparently healthy people (green). *Coronary Artery Disease Consortium. Arteriosclerosis, Thrombosis and Vascular Biology 2009 29: 774-80 doi:10.1161/atvbaha.108.181388 © 2009 American Heart Association, Inc. All rights reserved. Unauthorized use prohibited.*



As well as our work with the Cardiogenics consortium, we are undertaking a large meta-analysis of 23 000 cases and 60 000 controls (Cardiogram). We analysed lipid levels in 2000 myocardial infarct cases and 2000 controls from the Pakistan Risk of Myocardial Infarction Study typed with a SNP array covering 2000 vascular genes and found very similar allelic architecture in loci affecting lipid levels in Caucasians. We are currently combining similar data from six studies (26 000 individuals).

Through our HaemGen genome-wide association studies of eight haematological traits we discovered that the platelet count locus on 12q24 is also strongly associated with coronary artery disease and myocardial infarction, with the risk allele increasing platelet count. This allele is found on a long-range haplotype spanning 1.6 Mb and has arisen from a selective sweep unique to European populations. Interestingly, this haplotype carries risk alleles for type 1 diabetes and celiac disease.

A major challenge is to identify causative variants within regions of confirmed association. In the Wellcome Trust Case Control Consortium, we undertook targeted resequencing followed by fine mapping in regions of confirmed association but refined the initial signal in only half of the tested loci. Integration of copy number variation and data from the 1000 Genomes Project should streamline such efforts, but increased resolution is more likely to come by fine mapping in populations with lower levels of linkage disequilibrium (e.g. African groups).



Wellcome Library, London

➤ **An association analysis across 11 000 cases confirms effects of four loci on risk of coronary artery disease.**

Coronary Artery Disease Consortium et al. Large-scale association analysis of novel genetic loci for coronary artery disease. *Arterioscler Thromb Vasc Biol* 29: 774 (2009)

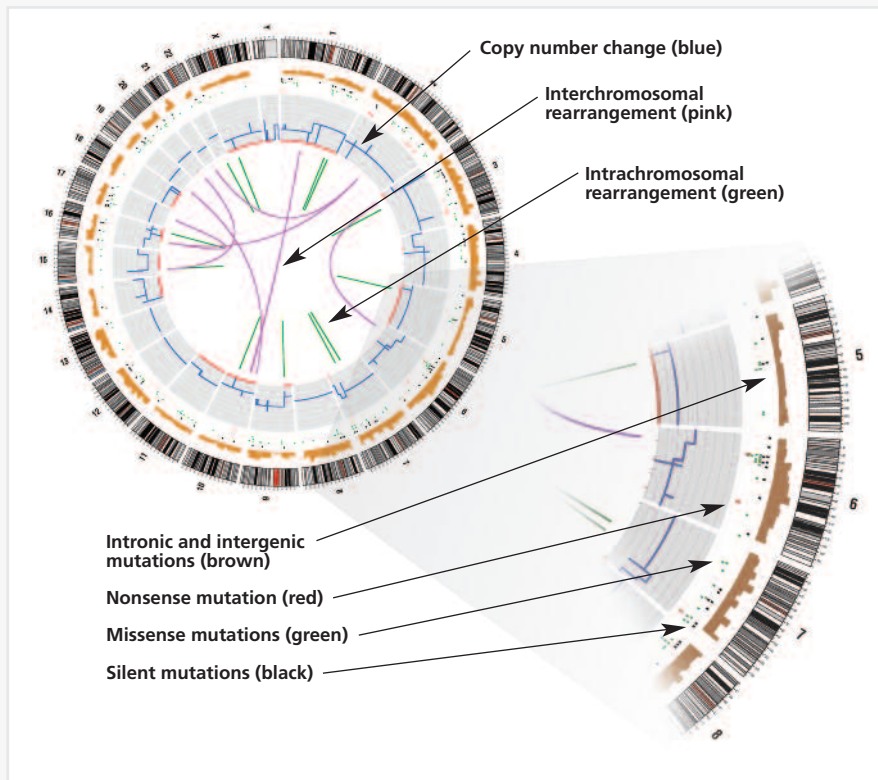
➤ **Two novel loci affecting coronary artery disease and myocardial infarction are identified – MRAS on chromosome 3 and a locus on 12q24.31 (HNF1A, C12orf43).**

Erdmann J et al. New susceptibility locus for coronary artery disease on chromosome 3q22.3. *Nat Genet* 41: 280 (2009)

➤ **A genome-wide association study for optimal dose of the anticoagulant drug warfarin in 1600 Swedish patients identifies common sequence variants in three genes – VKORC1, CYP2C9 and CYP4F2.**

Takeuchi F et al. A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet* 5: e1000433 (2009)

The Cancer Genome Project aims to characterise somatic mutations in human cancer through systematic, genome-wide mutational searches.



Figurative representation of a complete catalogue of somatic mutations obtained by sequencing the genome of a human malignant melanoma. This cancer has approximately 30 000 somatic mutations.

Cancers are clonal cellular proliferations characterised by dysregulation of cell growth, differentiation and death. They are caused by somatic mutations, changes in DNA that occur in dividing cells throughout an individual's lifetime. Through our mutation searches, we aim to identify new cancer genes and to reveal underlying patterns of somatic mutations, which reflect the mutational processes and environmental exposures that have influenced the evolution of each cancer.

We have carried out a high-resolution copy number analysis of 800 cancer cell lines, analysing the results specifically for homozygous deletions. This revealed that recessive cancer genes and regions of genomic fragility are responsible for previously unexplained clusters of homozygous deletions in the cancer genome.

We have completed the first large-scale screen using second generation sequencing technology for somatic rearrangements in

cancer. We identified over 2000 rearrangements in 24 breast cancers, including more than 20 in-frame fusion genes, and discovered substantial diversity of rearrangement phenotypes.

We continue to update the Cancer Cell Line project, which provides online access to our genomic, transcriptomic and mutational data from around 1000 publicly available cancer cell lines of diverse classes. Over the next five years we plan to test the sensitivities of these lines to some 400 therapeutic agents and correlate drug sensitivity with their genomic profiles.

With others, we helped to found the International Cancer Genome Consortium, which aims to sequence 25 000 cancer genomes over the next 5 to 10 years. As a step towards this aim, we have generated high coverage sequence from two cancer genomes, yielding essentially complete catalogues of all classes of somatic mutations.



Wellcome Library, London



Wellcome Library, London

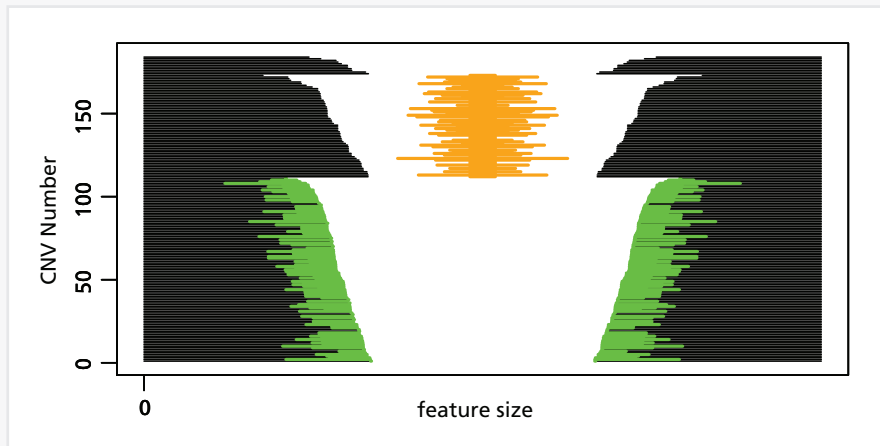
➤ **Sequencing of the coding exons of ~3700 protein-coding genes identifies the H3K27 demethylase, *UTX*, as a new recessive cancer gene; *UTX* shows inactivating mutations in multiple myeloma, renal and oesophageal cancers.**

van Haafden G et al. Somatic mutations of the histone H3K27 demethylase gene *UTX* in human cancer. *Nat Genet* 41: 521 (2009)

➤ **A perspective on the background, goals and potential of cancer genomics.**

Stratton MR et al. The cancer genome. *Nature* 458: 719 (2009)

Our aim is to document the relative contributions of different mutation processes, and to understand the genetic and environmental factors that influence mutation rates, particularly processes that generate structural variation.



Signatures at ~200 deletion breakpoints reveal three classes of deletion: those with sequence inserted (orange), those with similarities at the breaks (green) and those with neither (at top).

Each generation around 100 new mutations are introduced into every new human genome. These changes are generated by a diverse set of mutation processes, from the substitution of a single base to structural changes affecting thousands of bases. A detailed understanding of human mutation processes will allow us to predict disease-causing mutations not yet seen clinically and will inform clinical decision-making, helping to distinguish benign and pathogenic variants and to identify the likely risk that a pathogenic mutation will be passed on to future generations.

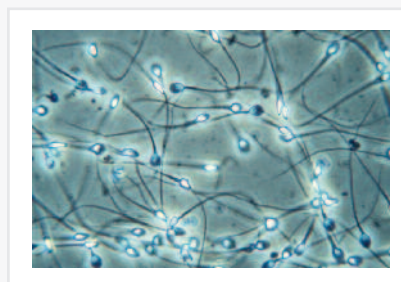
Novel experimental methods for haplotyping single DNA molecules have enabled us to assess mutation rate directly in sperm genomes, shedding light on the relative rates of deletion and duplication. Using new sequencing technologies to characterise large numbers of breakpoints, we have identified signatures of at least five different mutation processes.

We have been developing biochemical methods to identify recombination hotspots in duplication sequences. Knowing the location of these hotspots will allow us to develop locus-specific assays to measure rates of novel chromosomal rearrangements.

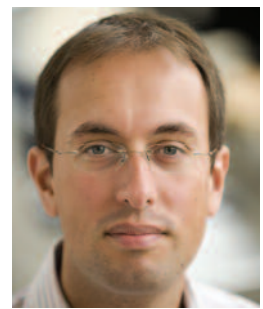
We have also begun studies to measure the heritability of mutation rates in twins and to assess the effect of age on chromosomal rearrangement.

By analysing rearrangements in thousands of apparently healthy individuals, we have developed a statistical tool for predicting which genes are likely to be haploinsufficient. We are now assessing whether rare deletions causing abnormal fetal development or severe childhood obesity tally with our predictions of dosage-sensitive genes.

Finally, using data from the 1000 Genomes Project we are for the first time estimating the rate of germline base substitution in an individual family – a step towards an ultimate goal of identifying all mutations arising in an individual generation.



Light microscope picture of human sperm.



Wellcome Library, London

A new, highly versatile and high-throughput method for genotyping large chromosomal rearrangements and haplotyping SNPs separated by large distances.

Turner DJ et al. Long-range, high-throughput haplotype determination via haplotype-fusion PCR and ligation haplotyping. *Nucl Acids Res* 36: e82 (2008)

Whole-genome sequencing of a male from the Yoruba people in Nigeria reveals 400 000 structural variants, many previously unknown.

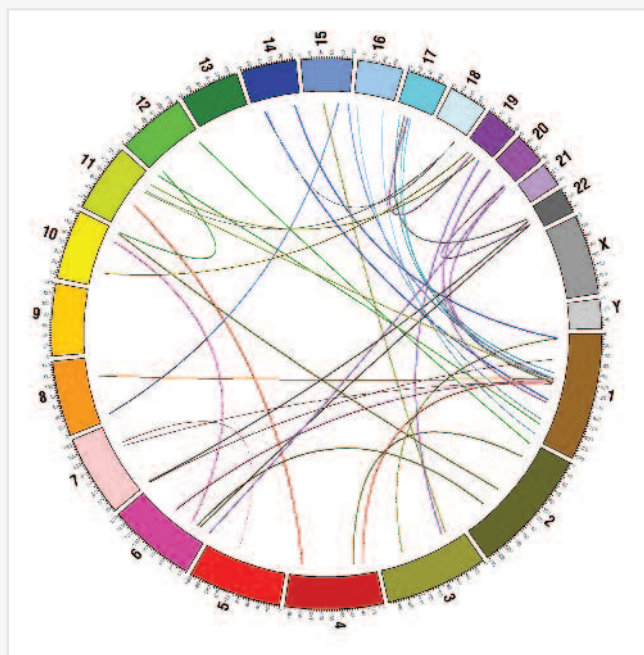
Bentley DR et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456: 53 (2008)

A statistical method for identifying associations in case-control studies that minimises false positives.

Barnes C et al. A robust statistical method for case-control association testing with copy number variation. *Nat Genet* 40: 1245 (2008)

Joyce Harper, UCL, Wellcome Images

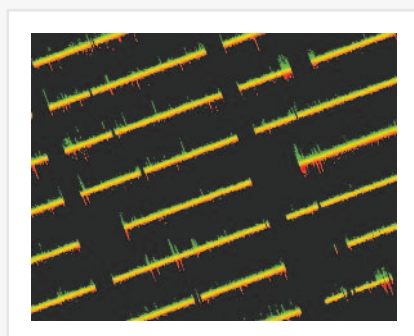
We are using a range of different genomic technologies to map structural variation throughout the human genome and assess its biological impact.



Jumping duplications. Each line represents a sequence that has duplicated from one genomic location to another. The lines are coloured according to the chromosomal origin of the duplication.

Structural variation – deletions and duplications of DNA (copy number variants, CNVs), as well as balanced rearrangements such as inversions – has been implicated in a growing number of diseases as diverse as spinal muscular atrophy, lupus and malaria. Yet its full impact on health remains unclear.

Following on from our widely used first-generation map of large-scale CNV, published in 2006, we have generated a high-quality, high-resolution map of 11 700 CNVs in European and African populations. We have genotyped these variants in 450 individuals from three different populations, providing an unprecedented resource for human genetic studies.



The distinctive patterning of gains and losses of DNA along chromosomes.

This map reveals potential causal variants underlying more than 10 published common disease associations, as well as insights into the different mutation mechanisms underlying this form of variation. This map has been fast-tracked into human genetic association studies, and these CNVs are currently being genotyped in around 100 000 samples as part of the Wellcome Trust-funded Case Control Consortium and other association studies. We have also developed algorithms and software tools to enable robust association testing with the resultant data.

We are establishing a lower resolution CNV genotype resource in a more geographically diverse set of populations through the HapMap3 collaborative project. We have also developed and published methods to identify structural variation from new generation sequencing technologies, which we are implementing as part of the 1000 Genomes Project.

A comparison of points of copy number variation in humans and chimps, and copy number differences between the two species, sheds light on their phenotypic differentiation.

Perry GH et al. Copy number variation and evolution in humans and chimpanzees. *Genome Res* 18: 1698 (2008)

Whole-genome sequencing of a male from the Yoruba people in Nigeria reveals 400 000 structural variants, many previously unknown.

Bentley DR et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456: 53 (2008)

A statistical method for identifying associations in case-control studies that minimises false positives.

Barnes C et al. A robust statistical method for case-control association testing with copy number variation. *Nat Genet* 40: 1245 (2008)

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



Slides in the genotyping facility.

We are conducting the first international genome-wide association study of migraine with aura, in collaboration with six major headache research centres (Brisbane, Cologne, Helsinki, Leiden, Munich and Oslo). Genotype data from 2900 cases from three countries and 10 000 population-specific controls are currently being analysed.

In multiple sclerosis, we have focused on a unique founder population – a patient group from western Finland, all genealogically traced to two ancestral couples. Using genome-wide data we identified a rare, high-impact locus containing the complement C7 gene on chromosome 5p. We are also providing one of the largest individual multiple sclerosis cohorts (980 Finnish samples) to the Wellcome Trust Case Control Consortium. Genotyping has been completed and statistical analysis is in progress.

A genome-wide association study of 2000 patients, part of the EU-funded SGENE (Schizophrenia Gene) consortium, provided evidence for the role of recurrent microdeletions in schizophrenia. A study of 200 related cases from a high-risk area of north-eastern Finland, founded by 40 families in the 16th century, identified a rare 200kb deletion on chromosome 22 shared

by affected individuals. The findings illustrate how discrete well-characterised populations can help identify rare alleles behind common diseases and provide further evidence of the importance of copy number variation in neuropsychiatric disorders.

Unique populations have also shed light on possible cellular abnormalities 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.

The value of founder populations was also emphasised by our identification of mutations in *GLE1* as the cause of recessively inherited lethal congenital contracture syndrome 1 (LCCS1), which disrupts development of anterior motor neurons in the fetus. This discovery identified a novel pathway potentially affected in more common motor neuron diseases such as amyotrophic lateral sclerosis.



Genome-wide association data reveal population structure and founder effects in regions across Finland.

Jakkula E et al. The genome-wide patterns of variation expose significant substructure in a founder population. *Am J Hum Genet* 83: 787 (2008)

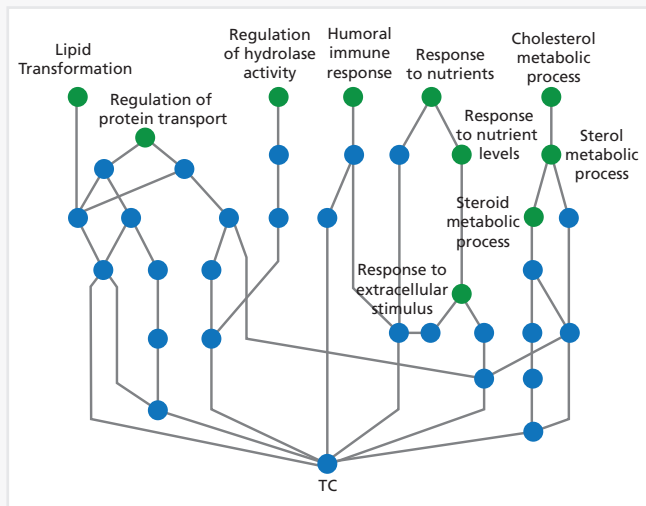
Studies of an isolated Finnish population identify a gene coding for a complement protein as a possible risk factor for multiple sclerosis.

Kallio SP et al. Use of a genetic isolate to identify rare disease variants: C7 on 5p associated with MS. *Hum Mol Genet* 18: 1670 (2009)

A screen of 150 ion channel genes finds no evidence that any contribute to common migraine in European populations.

Nyholt DR et al. A high-density association screen of 155 ion transport genes for involvement with common migraine. *Hum Mol Genet* 17: 3318 (2008)

Our aim is to understand the mechanisms underlying common cardiovascular diseases using special population resources, including founder populations.



Pathways containing an enrichment of the most strongly associated genes (green circles) with total cholesterol (TC) and their connections. Aulchenko YS et al. *Nature Genetics* 2009 41(1): 47-55 doi:10.1038/ng.269

Our strategy is to use European population resources, including founder populations such as Finland, to identify disease genes and to define their function and significance at the population level.

Our research draws upon several European-wide efforts, including the initial EU-funded GenomEUtwin (data from more than 300 000 twin pairs, including UK twins, www.genomeutwin.org) and the EU-funded ENGAGE (European network for genetic and genomic epidemiology, pooling more than 130 000 genome-wide association studies and over 1 million DNA and serum samples, www.euengage.org).

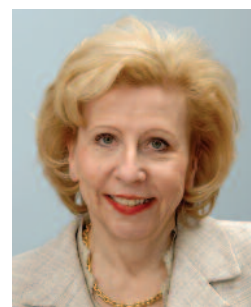
Unlike most genome-wide association studies, which have been carried out on patient groups, we have used European population cohorts, identifying more than 20 genes affecting serum lipid levels at the population level. The first genome-wide association study on a complete birth cohort (1966 Northern Finnish Birth Cohort) demonstrated the value of a founder population cohort to identify rare variants with high impact on serum lipid levels. From these studies, we hope to identify critical variants and evaluate their potential for assessing cardiovascular risk at the population level.

In cardiovascular disease, we contributed to several large international efforts characterising serum lipid traits, myocardial infarct, human height, weight, hypertension and other traits. Such large study samples, with over 100 000 participants, will in future facilitate analyses of potential sex-specific genes as well as gene-environment interactions.

Finally, we have begun functional genomics studies of certain genes, such as those identified in our genomic analyses of monozygotic twins who differ markedly in their levels of obesity. Using this approach we identified a decrease in mitochondrial copy number in fat tissue of obese twins as well as a potential new obesity gene, encoding factor XIII.



A slide in the genotyping facility.



Genome-wide profiling of 25 000 healthy Europeans identifies 22 genes regulating serum lipid levels, known risk factors for cardiovascular diseases.

Aulchenko YS et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet* 41: 47 (2009)

The first genome-wide association study of a complete birth cohort identifies rare variants regulating serum lipid levels.

Sabatti C et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet* 41: 35 (2009)

This project is elucidating the underlying genetic basis of mental retardation associated with genes on the X chromosome (X-linked), to provide a foundation for future healthcare planning in the commonest class of human congenital disease.

Mental retardation affects approximately 2 to 3% of live births and is the most common single feature of congenital human syndromes. The genetic causes of mental retardation are complex and diverse, ranging from deletions affecting many loci to defects in single genes. Many X-linked mental retardation genes have been identified, but others remain undiscovered.

In collaboration with clinical geneticists from the UK, Australia and the USA, we embarked on a direct mutational screen of 205 families, resequencing all protein-coding exons on the X chromosome (around 720 genes and 9000 exons).

As well as identifying a suite of new X-linked mental retardation genes (including *DLG3*, *ZDHHC9*, *CUL4B*, *AP1S2*, *SLC9A6*, *UPF3B* and *BRWD3*), our results also suggested that a remarkable 1% of genes on the X chromosome are not required for normal existence.

Outside mental retardation, we identified *FRMD7* mutations underlying congenital nystagmus, *EDA* mutations in X-linked dominant incisor hypodontia and mutations in *PCDH19* as responsible for a highly unusual condition in which epilepsy and mental retardation are restricted to females.

In the past year, we concluded our study, following up a further 19 truncating mutations. About half appeared to be implicated in mental retardation, and to investigate these further we screened six in a further 1000 families and 1000 control samples. Three new X-linked mental retardation genes were identified: *SYP*, *ZNF711* and *CASK*.

Overall, this study has identified 11 new mental retardation genes and two other disease genes on the X chromosome. It has therefore made a major contribution to the genetics of X-linked mental retardation, to X-linked disease genetics generally, and to an understanding of the patterns and impact of sequence variation on the X chromosome.



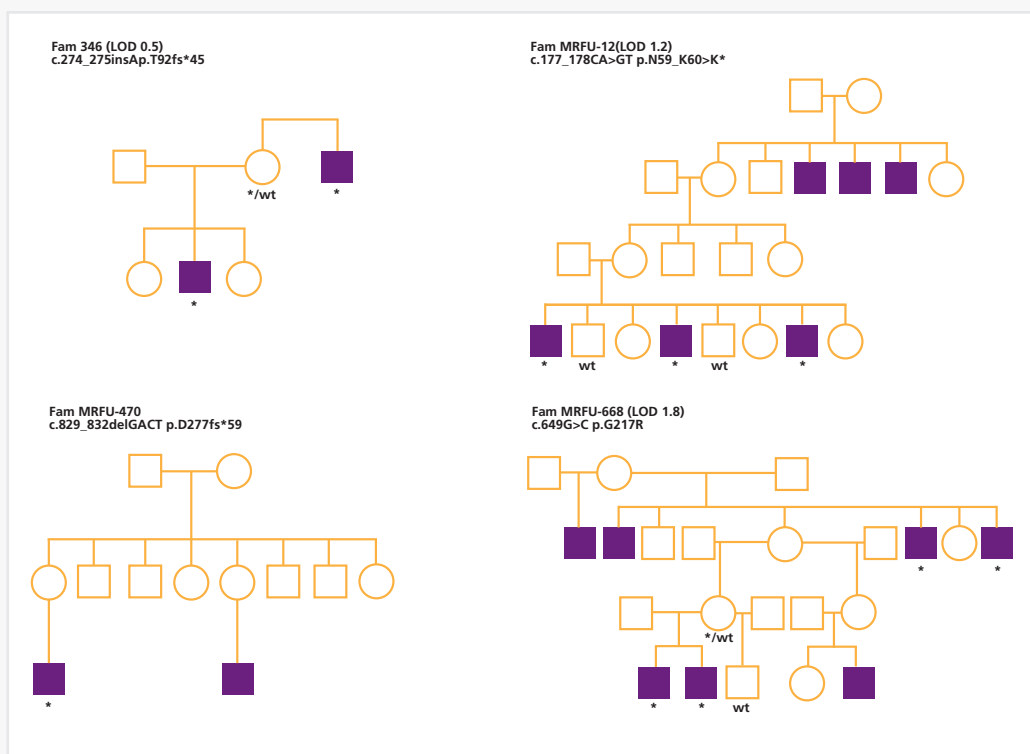
Wellcome Library, London



Wellcome Library, London

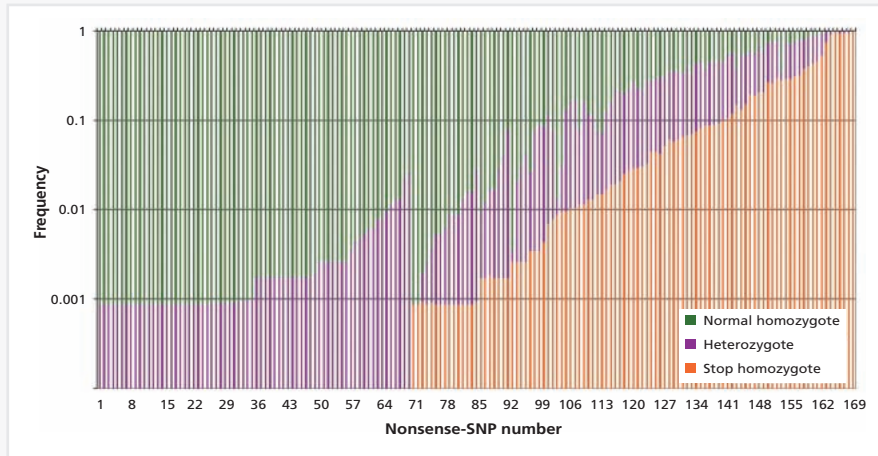
➤ **The largest direct screen for disease-causing mutations identifies nine genes causing X-linked mental retardation.**

Tarpey PS et al. A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nat Genet* 41: 535 (2009)



Pedigrees of families with likely deleterious variants in the *SYP* gene. Shaded symbols indicate individuals with mental retardation and open symbols indicate individuals who are unaffected. An asterisk indicates the presence of the disease-causing allele. lod scores are shown in parentheses. Tarpey PS et al. *Nature Genetics* 2009 41(5): 535-43 doi:10.1038/ng.367

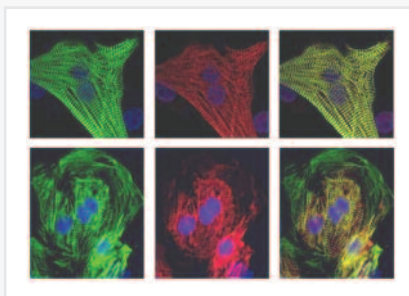
Our aim is to understand human genetic history and the patterns of natural selection in our genome, and their implications for health and disease.



A large number of stop or nonsense-SNPs are found in healthy individuals. Yngvadottir B et al. *American Journal of Human Genetics* 2009 84(2): 224-34 doi:10.1016/j.ajhg.2009.01.008

Every episode of natural or 'positive' selection implies that some people have died or failed to reproduce, so there is an intimate connection between evolution and medicine. These events are written into the human genome, and our main aim is to link present-day patterns of genetic variation to events in our evolutionary history.

In studies of basic evolutionary processes, we have resequenced two entire Y chromosomes and made the first direct measurement of the human base-substitution mutation rate. Reassuringly, it is similar to previous indirect estimates – around 3×10^{-8} nucleotide/generation – and we can now begin to investigate more subtle variations. We have continued our analyses of human migrations, finding that Y chromosomes from the long-gone Phoenician civilisation make up 6% of modern Y-chromosomal lineages in their descendants.



A 25bp deletion in the *MYBPC3* gene disrupts heart muscle structure (compare top, wild type with bottom, deletion) and causes heart failure, yet is present at the high frequency of ~4% in populations from the Indian subcontinent.

Dhandapany PS et al. *Nature Genetics* 2009 41(2): 187-91 doi:10.1038/ng.309

In collaboration with Richard Durbin's group, we are looking systematically at regions of the genome with particularly recent or ancient 'coalescences' or origins and obtaining our first glimpse of the genetic changes underlying the evolution of modern human characteristics.

As part of the 1000 Genomes Project, we have carried out the first scans of the whole genome for selected regions based on full sequence data. We have also shown that it is possible to home in very precisely on targets of selection by resequencing, even when our only starting information is their approximate location. In one case, this strategy identified a microRNA, which may regulate a set of downstream genes, as a target of selection.

One unexpected finding has been the wide variation in the number of functional genes in healthy individuals. A systematic investigation identified 169 genes inactivated by stop codons; for 99, both copies could be lost without any obvious phenotypic consequences (although loss was slightly deleterious over evolutionary time).

In contrast, in around 60 million people from the Indian subcontinent, the heart protein gene *MYBPC3* is effectively inactivated by a small deletion in an intron, resulting in a high risk of heart failure. This strongly disadvantageous variant probably arose in the subcontinent around 30 000 years ago and survives at such a surprisingly high frequency because its impact is felt relatively late in life.



Wellcome Library, London

As a consequence of their genetic history, some 4% of South Asian people carry a mutation affecting a heart protein, predisposing to heart failure.

Dhandapany PS et al. A common *MYBPC3* (cardiac myosin binding protein C) variant associated with cardiomyopathies in South Asia. *Nat Genet* 41: 187 (2009)

Loss of the *UGT2B17* gene, which may affect steroid hormone metabolism, has been positively selected in East Asia but is subject to balancing selection in Europe.

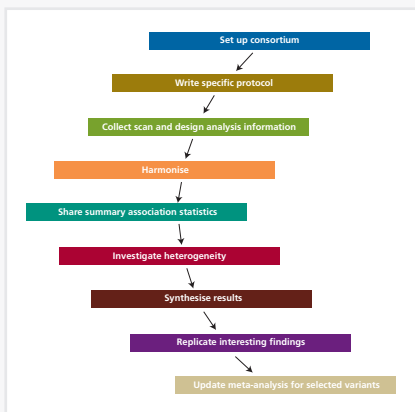
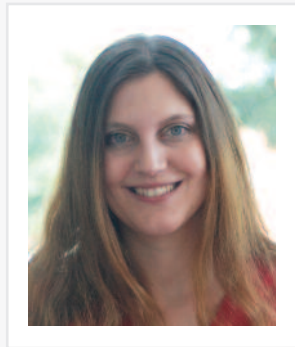
Xue Y et al. Adaptive evolution of *UGT2B17* copy-number variation. *Am J Hum Genet* 83: 337 (2008)

A surprisingly large number of genes contain stop codons without any apparent adverse effects.

Yngvadottir B et al. A genome-wide survey of the prevalence and evolutionary forces acting on human nonsense SNPs. *Am J Hum Genet* 84: 224 (2009)

Our aim is to evaluate, develop and apply strategies for the design, analysis and interpretation of large-scale association and resequencing studies.

Eleftheria Zeggini joined the Sanger Institute in November 2008



Typical workflow for conducting a meta-analysis of genome-wide association data sets.

Advances in high-throughput genomics have not been matched by equivalent advances in the analytical field. We seek to identify methodological needs and evaluate existing strategies, in order to make recommendations and fill any gaps. Importantly, our work has an applied component, with a focus on complex diseases and related traits such as osteoarthritis, hypertension, rheumatoid arthritis, type 2 diabetes, obesity, blood pressure, and glycaemic and anthropometric traits.

We have recently completed stage 1 of the largest genome-wide scan for osteoarthritis, as part of the arcOGEN consortium, which has identified robustly replicating knee and hip osteoarthritis loci. We are also involved in large-scale meta-analyses, for example in type 2 diabetes, through the DIAGRAM+ (DIAbetes Genetics Replication And Meta-analysis) consortium, which has identified at least 13 novel diabetes loci.

We have implemented a method for analysing rare variation from existing genome-wide association scan data and from emerging sequencing data, for both quantitative and dichotomous traits. We have compared the method to other existing analytical solutions for rare variants and applied it to data from international consortia working on a range of phenotypes, identifying several promising signals.

To identify the best follow-up strategies for these experiments, we have carried out a resequencing study design comparison. We have also evaluated the added information afforded by imputation approaches based on 1000 Genomes Project data.

We have analysed the effect of different genotype quality control procedures on the accuracy of imputation (deducing untyped SNPs based on genotyping data and detailed recombination maps). We have also evaluated the use of publicly available control genotypes in genetic association studies where the case samples have been typed using a different platform.



X-ray showing osteoarthritis in the knee.

Wellcome Photo Library, Wellcome Images

Simulations suggest that new statistical methods will be able to identify novel genes contributing to complex traits that conventional analyses would miss.

Morris AP and Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol in press* (2009)

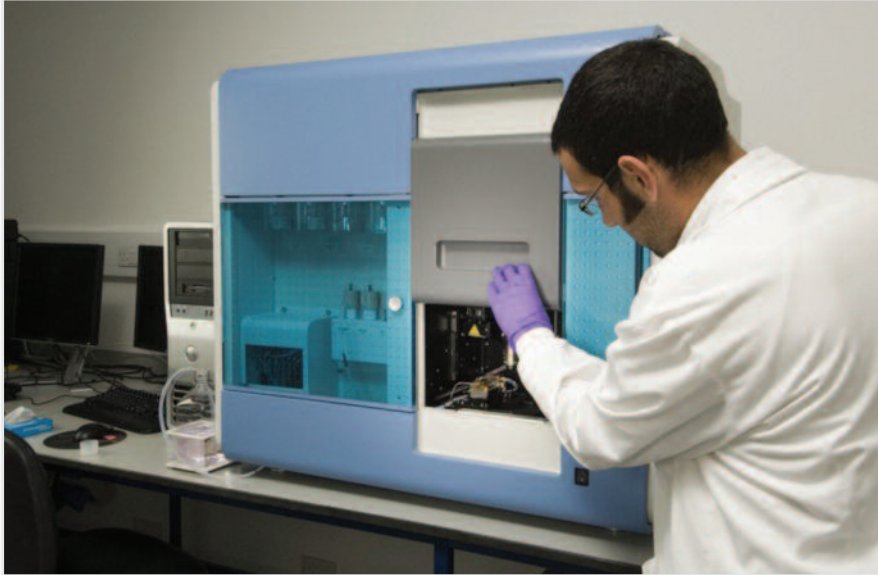
An analysis of type 2 obesity risk genes finds limited support for the thrifty gene hypothesis (that risk alleles have been positively selected in times of food stress).

Southam L et al. Is the thrifty genotype hypothesis supported by evidence based on confirmed type 2 diabetes- and obesity-susceptibility variants? *Diabetologia* 52: 1846 (2009)

A review of statistical techniques useful in meta-analysis of genome-wide association studies.

Zeggini E and Ioannidis JPA. Meta-analysis in genome-wide association studies. *Pharmacogenomics* 10: 191 (2009)

We are responsible for targeted sequencing projects in areas of medical relevance.



Wellcome Library, London

Scientist at a high-throughput DNA sequencing machine.

New sequencing technologies have provided opportunities to unravel disease-associated sequence variations on a larger scale than previously possible. One important development has been the introduction of new 'hybrid capture' techniques, which enable us to sequence continuous genomic regions, such as disease-associated regions, and/or specific elements genome-wide, for example all exons, all microRNAs or specific regulatory regions.

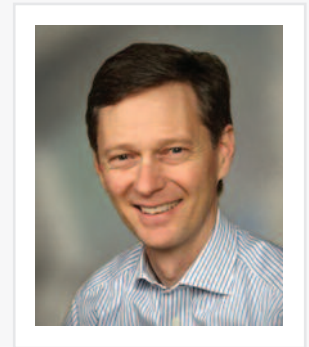
High-throughput hybrid capture has been applied to a 1000 Genomes pilot project, cataloguing sequence variations in all exons of 1000 genes from 200 HapMap individuals. We have also adopted this technique for projects aiming to uncover highly penetrant variants in complex traits or deleterious mutations in Mendelian traits.

As part of the Wellcome Trust Case Control Consortium fine mapping project, we analysed 25 disease-associated areas, aiming to uncover allelic variation in genomic regions associated with five disease traits (Crohn's disease, type 1 diabetes, rheumatoid arthritis, type 2 diabetes and ankylosing spondylitis). We identified over 9000 novel SNPs, which are now undergoing statistical analysis.

In our first large-scale complete exome sequencing study, we are sequencing the exome of 500 individuals from a Swiss population cohort. Individuals showing extreme fasting glucose levels have been selected for sequencing, aiming to identify penetrant, possibly relatively rare variants affecting this trait. This is a collaboration between the University of Lausanne, GlaxoSmithKline and the Sanger Institute (Inês Barroso).

Commercially available products for sequence capture of the whole exome are based on the consensus coding sequence set of genes. We are developing a product that would allow sequence capture of the whole exome as determined at the Sanger Institute.

We are also testing sequence capture technology for use with pooled samples, indexed samples and whole-genome amplified DNA.



X-ray of a hand with rheumatoid arthritis.

Wellcome Photo Library, Wellcome Images

The Gorilla Genome Project is producing a reference genome sequence for the gorilla.



Pieter de Jong, BACPAC Resources, Children's Hospital Oakland Research Institute

Kamilah (left), the western lowland gorilla whose genome sequence is being studied.

After the chimpanzee/ bonobo group, the gorilla is the closest relative of humans, and in some parts of the genome (up to 15%) it is more closely related to humans than chimpanzees are. We are producing a high-quality draft reference sequence for the gorilla genome, by sequencing a western lowland (*G. gorilla gorilla*) individual to high depth and adding data from an eastern lowland (*G. beringei graueri*) animal.

We have completed genomic sequence, comprising 2x coverage by capillary sequencing and 35x final coverage by next-generation Illumina short read sequencing. An early draft sequence based on a PHUSION assembly of the capillary data, corrected by mapped Illumina reads, was released in Ensembl in 2008. Using a combination of assembly techniques with help from Zemin Ning's group at the Sanger Institute (PHUSION) and Ewan Birney's group at the European Bioinformatics Institute (VELVET), and considerable new software development, we are close to a final assembled sequence using entirely *G. gorilla* reads. We expect the final assembly to contain approximately 2.8Gb of sequence.

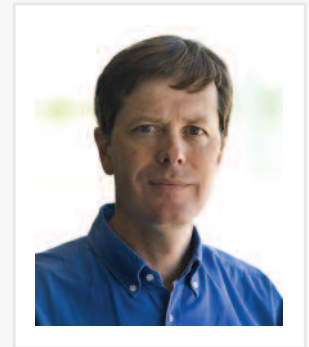
We will remap all the reads from both gorilla species to the final assembly to identify within-gorilla variants, and align our reference to human and chimpanzee to identify between-species differences. We are also Illumina sequencing an RNA sample, to aid transcript annotation and to support studies of gene evolution.

We are initiating genome-wide analyses to shed light on human and gorilla evolution and we expect to submit a primary manuscript in the next year.

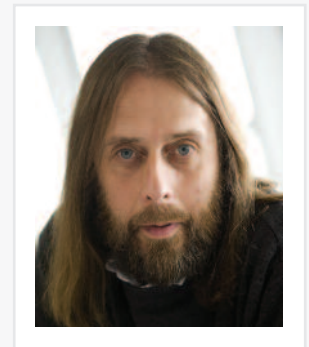


Wellcome Library, London

Rows of different individuals' DNA sequences aligned at the same position in the genome.



Wellcome Library, London



Wellcome Library, London