# From genetic to genomic and to functional genomic. Goals and future perspectives

✉ Marta Dueñas, Marcelo Nazábal, Lidia I Novoa
Center for Genetic Engineering and Biotechnology, 31 Ave. / 158 and 190,
AP 6162, CP 10600, Havana, Cuba

In comparison with most other disciplines of science, the field of genetics is still in its youth. The majority of scientific work in genetics has been done in the past 150 years. The successful preliminary sequencing of the human genome was announced in 2001. Nonetheless, interest in heredity and in other concepts within the field of genetics has existed since the beginning of humanity.

Even though not recorded, there is reason to believe that the concept of genetics was contemplated by the first human beings. When specific traits were noted to be shared by parent and child, this likely raised questions about the phenomenon of inheritance (although it was not termed as such then). The same phenomenon of shared traits between parent and offspring was then applied in the areas of plant cultivation and animal husbandry. Decision making based on the concept that traits could be passed on from generation to generation laid the foundation for the modern genetics that we practice today.

Like their predecessors, geneticists of today are influenced by trends and developments in science. Currently, substantial energy and resources are being directed into efforts to sequence the human genome. A complete sequence of the human genome holds the potential to revolutionize science and medicine.

## Laying the foundation for the sequencing of the human genome

In 1953, Watson and Crick elucidated the double helical structure of DNA. Their proposed DNA structure was published in Nature on April 25, 1953, as a brief article that barely exceeded 1 page. This article heralded a new age of discovery in genetics and laid the foundation for the sequencing of the human genome.

Today the term genome is widely understood to be the genetic material of an organism. However, this common term existed well before modern genetics. The word genome was coined by Winkler in 1920 and is defined as the haploid set of chromosomes and all the genes they contain.

Genome is derived from parts of 2 other important words: gene and chromosome. With advances in genetic research, we now further define a genome to be composed of a series of nitrogenous DNA bases (adenine [A], guanine [G], thymine [T], or cytosine [C]). In each organism, these bases are arranged in a specific order, and this order is the genetic code of the organism. In humans, the genome is made up of approximately 3 billion such bases. In 2001, a first draft sequence of the entire human genome was completed and made available to the public for study and research.

The first draft of the human genome was greeted with great anticipation by geneticists and nongeneticists alike. Understandably, every person had some vested interest in this scientific endeavor because the material that the researchers were working on was the genetic code of the human species. The genetics community devoured detailed accounts of the project through record-length journal articles with numerous acronyms. The general public was informed of the sequencing of the human genome through extensive media coverage. Although the geneticists could understand the details and the future implications of this feat better than laypersons, the importance of this achievement was not lost on anyone, regardless of his or her professional training. It was understood that the sequence of the human genome held the answers to our uniqueness as a species and that this information would likely be the basis for important future biomedical achievements.

The formal Human Genome Project (HGP) was conceived fewer than 20 years ago, yet in that short time frame, it has become one of humankind's crowning achievements.

Soon after the initiation of the HGP in 1987 in USA, the Italian National Research Council started a pilot project on genome research composed of 15 groups throughout Italy. The United Kingdom launched its project in February 1989. Then in 1990, the European Commission initiated a 2 year human genome project. France's government project followed months later in June 1990.

## Formal initiation of the HGP

Despite its relatively brief history and skepticism that it would fail, the HGP has seen several remarkable successes. When the HGP was first proposed in the 1980, there was little support for pouring billions of dollars into another big science project with few obvious benefits. Because the human genome was estimated to consist of approximately 3 billion bases, completion of the project was likely to be expensive and labor intensive. At that time, the cost of a finished sequence was approximately $10 per base. A well equipped laboratory could produce about 500 bases of sequence a day, and solid data indicated that more than 90% of the 6 feet of DNA in every cell was "junk" (ie, repetitive sequences with either no known function or noncoding "spacer" DNA). Despite well-founded skepticism at that time, it is now difficult to find someone who admits to opposing the project in the early days. An engaging summary of the personalities

and controversies surrounding the early days of the project was published in the genome issue of Science (Science. 2001; 291: 1145-1434).

The international effort to sequence the human genome was initiated in 1990 and was named the Human Genome Project. The HGP originally developed a 15-year plan to map and sequence the human genome. The plan outlined several goals including (1) development of high-resolution genetic and physical maps of the human genome; (2) determination of the complete DNA sequence of humans and several other model organisms; (3) development of the capability for storing, analyzing, and interpreting these data (now called bioinformatics); and (4) development of the technology necessary to meet these goals, as well as assessment of the ethical, legal, and social implications of genomics. Objectives relating to research training and technology transfer were also elaborated at that time. Rapid progress (particularly in constructing genetic and physical maps) led to the extension and revision of the original goals, and a "5-year plan" was released in October 1993. These revisions were mostly incremental, detailing revised and more ambitious goals, including the completion of 80 million bases of sequence (<3% of the total size of the human genome) for all targeted organisms by 1998.

The first human chromosome to be completely sequenced was chromosome 22, completed in a collaborative effort by scientists at the Sanger Institute in England, at the University of Oklahoma and at Washington University in St Louis in the United States, and at Keio University in Japan. Its sequence was published in December 1999.

Five months later, in May 2000, the sequence of chromosome 21 was published from a collaborative effort by German and Japanese groups. Chromosome 20 was sequenced by the end of December 2001. Like chromosome 22, chromosome 20 was also sequenced by collaborators at the Sanger Institute.

## Achievements

By any measure, the genome project has been an outstanding success, surpassing the most optimistic projections of progress and costing far less than originally expected. The completion of the draft human genome sequence by both the public project and the Celera project was announced in special issues of Nature and Science in February 2001.

By the end of 2001, nearly half of the genome had been deposited in "finished" form (<1 error every 10000 base pairs), and all but 1.5% of the targeted sequence is present in GenBank (the public database) as either finished or draft form. Data from the public project are accessible through the National Center for Biotechnology Information Web site (http://www.ncbi.nlm.nih.gov/).

Other goals of the HGP also have been met or exceeded. Complete genomic sequences for Escherichia coli, Saccharomyces cerevisiae (bakers' yeast), C elegans (roundworm), D melanogaster (fruit fly), and many other microorganisms have been determined and published. As of this writing, the mouse genome is 96% complete in draft form, and projects to complete the rat and zebra fish sequences have been initiated. Impressive progress has been made in detailing the nature and extent of sequence variation in humans. The initial goal for the cataloging of single nucleotide polymorphisms (SNPs) (variations in single bases of DNA detected in the population at large) was to create a map with 100.000 SNPs by 2003. As of October 2001, dbSNP (the repository for these data) had more than 4 million SNPs, which are actively being used by researchers in both academia and industry to identify genes that contribute to disease susceptibility and drug response. These developments promise to dramatically change the way in which medicine is practiced in the 21st century.

## Future goals

The next important challenge is to make sense of all the data generated by the genome project. One of the major surprises from early analysis was that previous estimates of gene number appear to be inaccurate by a substantial margin.

Most pregenome estimates agreed that humans had between 60000 and 100000 genes (although some estimates ranged as high as 120000). Early analysis of the complete genome sequence suggests that the true number of genes required to make a person may be only 30000 to 40000. The difficulty in finding actual gene sequences in 3 billion bases of complex data is underscored by a recent analysis that revealed the 31780 genes predicted by the public project only partially overlap with the 39114 genes identified by Celera. Adding to the controversy is a recent computational reanalysis of the "completed" public sequence that concluded there might be as many as 75000 genes in humans.Thus, there is reason to believe that the human gene number may be greater than 30000. Clearly, one of the immediate goals is to improve the informatics capability to allow more accurate and meaningful analysis of the sequence data.

Whatever the final number of genes in the human genome, a more important goal is to understand the function of each gene. The function of approximately 15000 such gene products is known. Thus, despite years of study, we understand the function of fewer than half of the human genes. Do some of the unknown genes code for products that will help us better understand cancer or heart disease? What new drug targets remain to be discovered? Now that the DNA sequence is largely complete, there is interest in trying to understand not only the function of genes (genomics) but also the function of proteins (proteomics) and how groups of proteins in common pathways combine to produce physiological responses (metabolomics). These efforts are already under way and will lead to insights about cell physiology.

Another area of intensive work is to understand how differences in human DNA sequences affect our daily lives. It is clear that common minor changes in DNA sequence can affect not only susceptibility to rare genetic disease but also many more common conditions. As a preview of what lies ahead, information from the genome project was recently used to identify susceptibility genes for non-insulin-dependent diabetes mellitus and inflammatory bowel disease.

Applications will not be limited to prognostic indicators and diagnosis. The NIH recently launched a program to correlate sequence variations with drug responses (pharmacogenomics). The long-term goal of the effort is to tailor drug therapies to the individual patient. The wealth of genetic information has raised some concerns as well, including the confidentiality of genetic information and the possible stigmatization of people based on their DNA sequences. These issues deserve continued study and scrutiny.

## Conclusions

Humankind's interest and study of genetic-related concepts are recorded throughout history. However, the greatest advances in the field have occurred within the past 150 years. The future promises more progress in genetics and in related areas of medicine and science. An early statement by a well-known geneticist, Alfred Sturtevant, in regard to the future of genetics is as applicable now as it was when he wrote it in the last century: "We should like to conclude the account with an outline of the future of genetics; but one of the things that makes science interesting is that new developments are likely to come in unexpected directions. One thing can safely be predicted; genetics will continue to develop and to include new fields of work-it is not static".

The completion of the sequencing of the human genome will surely lead to further progress in genetics, science, medicine, and other related fields. The greatest benefits from the sequencing of the human genome are yet to be realized. Although continued advances in genetic research will be recognized most immediately by those working in genetics, the achievements and their social and ethical implications will affect all humanity. An understanding of the history of genetics and how we have arrived at where we are today will help prepare us to meet tomorrow's challenges.

## Biotechnology Habana 2003

Last November in celebration of the 50th Anniversary of the discovery of the double helix by Watson and Crick, a very outstanding biotechnology conference was held in Havana at the Center for Genetic Engineering and Biotechnology. This one week conference was organized by the symposia of biomedical research areas and Functional Genomics was one of them. This symposium was divided into two sessions one more focused on populational studies of complex traits, such as arterial hypertension and cancer in which the chairpersons were Dr. Lidia I. Novoa (Cuba) and Dr. Pancaj Sharman (UK) and a second session more related to HLA studies and HGP developments in which the chairpersons were Dr. Marcelo Nazábal (Cuba) and Dr. Derek Middleton (UK). In each session five conferences were presented and all the abstracts are included in this summary.

## *The genetics of cerebrovascular disease*

Pankaj Sharma MD PhD

Department of Clinical Pharmacology, University of Cambridge,Addenbrooke's Hospital, Cambridge CB2 2QQ, UK

Successes in mapping disease genes in monogenic disorders has not been matched in the common diseases such as stroke, heart disease and hypertension, partly hampered by the heterogenic nature of these disorders and the difficulty in recruiting large numbers of well characterised patients who often present late in life.

Stroke is the third commonest cause of death in the developed world and is rapidly becoming a problem for the developing world. Progress has been made in the genetics of the rare causes of stroke such as those that cause familial cerebral haemorrhages, as well as for ischaemic stroke e.g. mitochondrial disorders (MELAS), and cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL). Mutations in *Notch3* gene are responsible for CADASIL. The development of a high throughput method for quickly analyzing all the known *Notch3* mutations has shown that the gene is unlikely to play a major role in the aetiology of common ischaemic stroke.

Although investigators throughout the world have undertaken candidate gene based allelic-association studies in an attempt to identify stroke susceptibility genes, many of these studies are underpowered and have provided conflicting results. Combining the results of all these studies in a systematic meta-analysis allows many of the problems with small studies to be overcome. To date, 70.000 subjects have been included in such an analysis. Nearly half of the candidate genes assessed demonstrated a significant odds ratio for developing ischaemic stroke, with no one gene contributing an OR of >1.5. The study demonstrates the virtue of candidate gene based strategies but also highlights the small risks conferred by each likely susceptibility gene, although the population effects may be much higher.

Identifying stroke risk genes not only increases our understanding of the mechanistic processes underlying cerebrovascular disease but may help to open as yet unpredictable therapeutic avenues.

## *Immortalized Human B-lymphocyte Cell Lines in Translational Cancer Research: The Genomic Instability Project Model*

Stephen Oglesbee

UNC Lineberger Comprehensive Cancer Center, Tissue Culture Facility University of North Carolina at Chapel Hill, CB 7295, 318 Lineberger, Chapel Hill, NC 27599, 919-966-4324, 919-966-5782 (fax), tcf@ucn.edu (direct email). Corresponding Author: Dr. Lisa Carey, MD

Immortalization techniques permit genotypic and phenotypic examinations by multiple collaborating investigators to be performed on a uniform sample received and processed by a centralized dedicated facility. Such studies have become increasingly desirable with expanding understanding of germline variations with risk of cancer development and a proliferation of novel and specific assays of DNA stability and repair. Complications of immortalization and the impact it has on specific types of studies include variations in receipt and processing of tissue samples, alterations in cell line phenotype with serial passage, and issues related to confidentiality of data derived from germline DNA. Modern techniques permit effective B-cell immortalization in with success rates exceeding 99% with an emphasis on uniformity and consistency in the resulting cell lines.

An ongoing case-control study examining DNA stability and repair in breast cancer patients will be used to illustrate these themes. In this study, germline DNA and tumor tissue is obtained from women with a new diagnosis of breast cancer and an age- and race-matched group of control women without cancer. Data regarding exposure history and other clinical variables are obtained from questionnaires and the medical charts.

Blymphocytes from the separated buffy coat are immortalized with EBV. Samples are provided to several collaborating laboratories with complementary interests in DNA repair gene polymorphisms, functional studies of induced hypermutability, and induced oxidative stress patterns. Effective communication between various laboratories and their results, particularly in light of clinical variables including relapse and survival, required extensive a priori planning. Data security and patient confidentiality is crucial in germline DNA studies with clinical correlates, and a model incorporating a single database with multiple layers of access depending upon investigator and/or clinician requirements will be presented

## *The Challenge of Connecting the Human Genome to Clinical Phenotypes: Technological Platforms Required for Collecting and Managing Data in Large Community and Population Genomics Projects*

Daniel Gaudet, Steve Arsenault

Community Genomic Medicine Center, Chicoutimi Hospital, 305 St-Vallier, Chicoutimi Qc CANADA, G7H 5H6

The distinctive feature of large-scale genomics projects is the significant number of people (tens of thousands) involved in these research projects. Such feature yields several challenges which need to be dealt with. One of these challenges is the huge amount of DNA samples and phenotypic/genotypic data to be managed. The implementation of a costeffective high throughput automated system for extracting, storing and managing tens of thousands of high quality DNA samples constitutes a major asset for such projects. In addition, no population project of this size and importance can be envisaged without an integrated and secure data management system. The informatics platform must include hardware and software required for privacy protection (highly secure validated infrastructure with "comprehensive access control" capability), integrated clinical genetic research management (informed consent, clinical data capture, sample request and tracking, de-identification and anonymization), genetic sample banking (sample logistics and management, inventory tracking, request generation and approval, interfacing with LIMS), and genetic and clinical data management (secured genetic and clinical data transfer and storage, data fencing and authorization-based access). The data management system is also essential to the knowledge transfer process, including the assessment, prioritization and validation of the integration of genetic determinants of health into health services. Another challenge lies in the quality of phenotypic information which could be linked to genotypic information. In a context where clinical phenotypic evaluation is conducted on a large number of people, the selection of parameters as well as the quality and standardization of clinical procedures for assessing such parameters must be optimized.

## *Sampling and Data Collection for a Genetic Epidemiology Study of Arterial Hypertension on Cuban Population*

Marta Dueñas, Marcelo Nazábal, Alfredo Dueñas, Lester Leal, Osmel Campanioni, Juan Roca, Hanlet Camacho, Alberto Cintado, Annia Ferrer, Adelaida Villarreal, Tamara Díaz, Jesús Benítez, Racmar Casalvilla and Lidia I. Novoa

Center for Genetic Engineering and Biotechnology, P.O.Box 6162, Cubanacán, Playa, La Habana 10600, Cuba

Genetic epidemiology is a young field in every respect, from study design, to laboratory methods, to techniques for data analysis. New directions on this kind of research have been suggested by many investigators such as association studies, particularly those relying on linkage disequilibrium (LD) using high resolution haplotypes, can have greater statistical power under appropriate circumstances. Given recent advances in genotyping and epidemiologic study design, the opportunity now exists.

In Cuba for more than ten years epidemiological studies have been conducted that have shown the relevance of the control of risk factors on the maintenance of normal levels of blood pressure. Investigators in this study are involved in the "Family Blood Pressure Program" - a large multi-center MINSAP project -. High Blood Pressure (HBP), is a mayor risk factor for: Ischaemic heart disease, Stroke, Cardiac and renal failure and Peripheral vascular disease. HBP is very common, almost 30% of the adult population has high blood pressure, and only 30% of the known hypertensive are under control. The main causes of this low rate of patients under control are low detection rate, inadequate treatment or no adherence to treatment. Whereas the individual treatment requires a selection of the most effective drug, given preferably once a day with fewer side effects. The detection of 90% of the hypertensive patients and the control of 80%, can have fantastic results in reducing mortality of one of the main causes of death in the population.

We have now a bank of approximately 600 samples of DNA, blood and sera from several families of hipertensive individuals with personal and familial data collected and processed in a relational database to further evaluate the Cuban population in terms of genetic polymorphisms of some candidate gene could improve the diagnosis of the real cause of high blood pressure and direct the clinicians to the correct medication. The immediate priorities are first to characterize the two candidate genes in the Cuban population: Secondly, the study of the correlation between the severity of the disease, race and age of debut. Thirdly to evaluate the response against different treatments to reduce blood pressure in order to prepare and evaluate the predictability of certain SNPs for the response of each treatment.

## *Genetic dissection of common diseases: Now it can be done!*

Ariel Darvasi

The Life Sciences Institute, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

The identification of genes affecting common diseases such as diabetes, asthma, cancer, schizophrenia, etc., as well as other complex traits, is considered to be one of the major challenges of contemporary genetics. In the past decade several attempts were made to identify such genes, with relatively little success. This state is now changing due to advances in knowledge of the human genome (a consequence of the Human Genome Project), new high throughput molecular technologies, and advanced computational and statistical approaches. We have comprehensively studied these elements in order to establish an effective general strategy for the genetic dissection of complex traits. In particular, we have studied linkage disequilibrium patterns across the genome and across human populations. We studied a well-characterized homogeneous population of Ashkenazi Jews in comparison with Caucasians and African Americans. We suggest that studying homogeneous populations provides significant advantages for gene discovery. To overcome the genotyping bottleneck, we examined several genotyping technologies for allele frequency estimation in DNA pools. Instead of genotyping each individual, pooling the DNA of many individuals together enables the comparison of cases and controls allele frequencies, with only a few reactions. We have combined these approaches and established unprecedented statistically significant gene-disease associations in various diseases. We expect that biomedical impact of these field will revolutionized medicine by completely dissecting the biochemical pathways involved and thus device novel therapies and preventive medicine.

## *Kir Genotyping*

Derek Middleton

Anthony Nolan Research Institute, Royal Free Hospital, Pond Street, Hampstead, NW3 2QG, London, UK

Natural Killer (NK) cells are an important component of the innate immune system. Two NK receptor families use HLA as their ligand – killer inhibitory receptor (KIR) and killer lectin like receptor (KLR) (CD94/NKG2). Both families have activating and inhibiting receptors.

The KIR receptors are very polymorphic both in number of genes expressed in an individual and the alleles present in each gene. Initially we described frequencies of each of the KIR genes. We have now developed SSOP techniques to detect the alleles of 2DL3, 2DL4, 2DS4, 3DL1/S1 and 3DL2. During the course of this study we have found many novel events – multiple copies of some of the genes in a number of individuals, deletion events leading to non-expression – in addition to finding many new alleles. A nomenclature committee similar to that for HLA has been formed to assign names to new alleles. Of the alleles already named there are several that we have not detected despite looking in different populations.

Two main haplotypes (A and B) have previously been described for the KIR receptor genes. Our work has shown that these are a useful guide but that there is much plasticity. Interestingly at the allele level there is not the linkage disequilibruim as seen in HLA (e.g. HLA-A*0101 with –B*0801). Our family haplotypes show many different haplotypes whereby one allele from one gene can be inherited with any of the alleles from another gene. A website was initially developed to capture the frequency data of HLA alleles in different populations. Data from studies of KIR genes and their alleles in different populations has been added to this website.

## *HLA class I and class II polymorphism in the Cuban population and their association with rheumatoid arthritis and celiac disease*

Marcelo Nazábal, Annia Ferrer, Hanlet Camacho, Alberto Cintado, Osmel Campanioni, Marta Dueñas., Lester Leal, Adelaida Villarreal, Tamara Díaz, Jesús Benítez, Racmar Casalvilla, Luis Sorell, Maria C. Domínguez and Lidia I. Novoa

Center for Genetic Engineering and Biotechnology, P.O.Box 6162, Cubanacán, Playa, La Habana 10600, Cuba

Human leukocyte antigens (HLA) allele determination is becoming an increasingly important aspect in the field of transplantation as well as in the area of HLA association with a number of diseases. Through Cuba's history, this country, situated at a crossroads of the route from Europe to America, has been a host for various populations of different ethnicities. The aim of our study is to determine whether allele polymorphisms in the Cuban population present a distinguishing feature. Although data on HLA phenotypic polymorphisms in Cuba have been reported in the literature, our study is the first to examine frequencies of HLA polymorphisms in the country at the molecular level and determining class I and class II alleles. Allele frequencies of the Cuban population were analyzed and compared with those of other populations and the previous reports of our population. HLA class I and class II genotyping was performed using PCR followed by SSO with commercial kit from Dynal. The subjects were 211 unrelated healthy blood donors of both sexes and of different regions in Cuba, mainly from Havana city.

We also investigated the HLA-DRB, and DQB polymorphism and haplotypes in RA subjects of the same origin. Using the same technique as the population study. We studied 46 subjects affected with the disease. We compared the allele and haplotype frequencies of the affected subjects with the healthy donors group as controls.

The celiac disease (CD) is a multifactor disease resulting from a life time abnormal immune response to gluten accompanied by autoimmune characteristics, which can in sensitive individuals evoke small bowel mucosa morphologic changes. The genetically sensitive individual to CD has not been defined yet, it is obvious, however, that this illness is closely linked to the HLA class II genes. The objective of our study was to detect associations of HLA class II alleles and haplotypes DRB1/DQA1/DQB1 in Cuban CD patients. A group of Cuban CD patients diagnosed according to ESPGHAN criteria was genotyped HLA for alleles of DRB1/DQA1/DQB1 loci. The DQA1/DQB1 typing was performed by SSP PCR and we also used the PCR SSO from Dynal to perform the DRB1\DQB1 typing in the same samples. The results obtained with both techniques were compared. This is the first report of this kind of study made in Cuban population. We compared our results with the allele and haplotypes frequencies reported in others populations.

The results have shown great diversity of HLA haplotypes in Cuban population which can be the result of admixture with neighborhood immigrating populations during the history.

## *Human Genome Annotation at the Sanger Institute*

Jane Lovel

Sanger Institute, United Kingdom

The role of the Human and Vertebrate Analysis and Annotation (HAVANA) group at the Sanger Institute is to provide high quality manually curated annotation of genomic sequence. The Sanger institute has sequenced approx third of the human genome, and published full analysis and annotation of chromosomes 20, 22, 6, 13. Work is under way to publish annotation and global analysis of chromosome 9,10 and X before the end of the year. Our analysis is performed on a clone by clone basis using a combination of similarity searches against DNA and protein databases together with a number of ab-initio gene predictions, prior to manual annotation by the HAVANA group based on supporting evidence. Given the importance of alternative splicing in providing functional diversification of the human gene set, one of the objectives of the HAVANA Group is to identify new alternative splice variants. Also annotated are processed and unprocessed pseudogenes which are difficult to detect automatically. When the annotation of a chromosome has been completed the data is submitted to EMBL and also available to browse in the Vertebrate Genome Annotation (VEGA) database. This database is designed to be a central repository for manual annotation of different vertebrate finished genome sequence. At present human chromosomes 6, 13, 14, 20 and 22 are available in VEGA. The sequence data may be browsed separately or alongside the predictions produced by the ensembl database. Since the data is manually curated it can be reviewed and updated frequently to give an up to date view of the genome.

## *The role of Genetics and Genomics in Health Care*

Klaus Lindpaintner

Roche Genetics, F. Hoffmann-La Roche, Ltd, Basel, Switzerland. klaus.lindpaintner@roche.com

The more fundamental understanding of molecular cell biology that advanced molecular and cell biological research is providing us with will, in due time, translate into a more differentiated understanding of pathology, and thus shift the emphasis towards molecular based diagnoses which will interrogate causative of contributory parameters. This will result in a rising importance of molecular diagnostics as a key tool to diagnose pathological entities. Furthermore, an enhanced and more causative understanding of disease will allow the targeting of novel, contributory mechanisms in our search for new medicines. The application of these medicines to clinical practice may then often depend on first establishing the molecular diagnosis for which the drug is applicable. The development of the Her-2-neu antibody, trastuzimab (Herceptin®) may serve as a paradigmatic example of this approach.

Whilst we look forward to additional, incremental progress being made along these lines, it is prudent no to exaggerate the rate at which this will be happening. We are still ignorant of the biological function of the vast majority of all genes, let alone of the clinical impact of any one of the millions of SNPs characterized so far. In addition, we must not overestimate the overall minor and secondary role which genetic predispositions, as compared to lifestyle and environment play in the majority of all common complex diseases. To generate a more realistic expectation than is commonly portrayed in various media, a renewed effort at public education about genetics, long neglected by geneticists, is urgently needed, to avoid unwarranted hopes as well as fears, and to combat "genetic exceptionalism", the sentiment that genetics represents a fundamentally different aspect of biology and medicine.

## *Experimental Gene Resources at the Sanger Institute*

Graeme Bethel

The Sanger Centre, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK

In order to completely annotate the human genome, the structure of every gene requires identification and confirmation to a high degree of confidence. Experimental verification of genes provides essential evidence for intronexon structure and has been significantly aided by the cDNA sequencing projects of the MGC, HUGE, NEDO and DKFZ. However, genes expressed at low levels may be missed using this random cDNA sequencing approach and furthermore the cloned cDNAs, available from these sequencing projects, require significant downstream manipulation before they can be used for protein expression and other functional studies. In conjunction with the HAVANA group, we are working on a number of complementary projects including verification of predicted genes and alternative variants, confirmation and full length ORF cloning of novel genes and the development of resources to aid the annotation of transcription start sites. The ultimate aim of this work is to provide a publicly available resource of cDNA clones, representing each gene in the human genome, to be used by research groups for further functional characterisation.