International Review of Psychiatry (2004), 16(4), 294–300

BrunnerRoutledge Taylor & Francis healthsciences

Genomics and the Human Genome Project: implications for psychiatry

JOHN R. KELSOE

Departments of Psychiatry, University of California, San Diego and San Diego VA Healthcare System, La Jolla, California, USA

Summary

In the past decade the Human Genome Project has made extraordinary strides in understanding of fundamental human genetics. The complete human genetic sequence has been determined, and the chromosomal location of almost all human genes identified. Presently, a large international consortium, the HapMap Project, is working to identify a large portion of genetic variation in different human populations and the structure and relationship of these variants to each other. The Human Genome Project has approached human genetics on a scale not previously seen in biology. This has been made possible by dramatic advances in high throughput technology and bio-informatics. Tools such as gene chips and micro-arrays have spawned an entirely new strategy to examine the function and expression of genes in a massively parallel fashion. Together these tools have dramatically advanced our knowledge about the human genome. They promise powerful new approaches to complex genetic traits such as psychiatric illness. The goals and progress of the Human Genome Project and the technology involved are reviewed. The implications of this science for psychiatric genetics are discussed.

Goals of the Human Genome Project

The Human Genome Project has been one the most exciting and rapidly growing areas of knowledge in medicine and biology. It promises revolutions in all aspects of medicine including psychiatry. Much of this issue is devoted to the impact of these powerful tools. As these articles detail, psychiatric clinicians will undoubtedly see their day-to-day practice of medicine change dramatically. In this paper, the goals, science and tools of the Human Genome Project will be reviewed. Application of the tools derived from this powerful engine of biology will also be discussed.

In 1953, Watson and Crick famously first described the double helical structure of deoxyribonucleic acid (DNA). This scientific event is widely used to mark the beginning of molecular genetics and the study of genomes. However, for many decades, the tools for exploring this unknown territory were slow and crude and progress correspondingly slow. Furthermore, the extent of that unknown territory remained a mystery. Just as early seafaring explorers had no idea as to the size of the world and feared falling off the end, early molecular biologists had only very indirect guesses as to the size and structure of the genome. The very invention of that word similarly marks a new phase in biology. A genome is the complete set of genetic material and information in an organism. Genomics has been variously defined as the study of the structure and function of whole genomes. Until 15–20 years ago, such a thing would be a futuristic fantasy. But the marvel of genomics is that fantasy becomes reality faster than most would expect.

The Human Genome Project was begun as an international effort to characterize the human genome. The primary goals were to determine the complete sequence of all human DNA and to identify and map all human genes. But first, a little refresher on genetics and DNA will be presented. DNA is a long molecule comprised of a chain of smaller molecules called bases. There are four bases to select from: adenosine, cytosine, guanidine, and thymidine. The order of these bases can define the structure of a protein. Three bases are required to define one of twenty possible amino acids. In this way, a series of triplets of bases can define the amino acids that make up a protein. Such sequences are termed coding sequence. A gene is a segment of DNA comprised of coding sequences that code for the amino acids in a protein that is the product of that gene. Such a protein may be an enzyme or a structural cellular element. A gene is expressed when it is transcribed into messenger ribonucleic acid (mRNA), and the mRNA is subsequently translated into protein. In bacteria, all the coding sequence is contiguous, however, in eukaryotes, the coding sequence of genes is interrupted. Small regions of coding sequence, termed exons, are interspersed among non-coding sequences called introns. After the DNA is transcribed into ribonucleic acid (RNA), the introns are spliced out and the exons spliced together to form a functional mRNA. However,

Correspondence to: John R. Kelsoe, Department of Psychiatry, 0603, UCSD, La Jolla, CA 02093, USA. Tel: + 1 858 534 5927; Fax: + 1 858 534 5527; E-mail: jkelsoe@ucsd.edu

ISSN 0954–0261 print/ISSN 1369–1627 online/04/040294–07 © Institute of Psychiatry
DOI: 10.1080/09540260400014385

coding sequence comprises only a small portion of the total DNA. Other regions, may have no function or unknown function, or may serve to regulate the expression of genes, i.e., when the gene is transcribed into mRNA. Though these non-coding regions have been termed 'junk' DNA because they presumably have no function, it is becoming increasingly apparent that much of this DNA may have important regulatory functions that have been previously unappreciated.

DNA is arranged into 23 pairs of long molecules that in metaphase assume the rod like shape of chromosomes and are visible by light microscopy. On each chromosome, the genes are arrayed in a consistent order and the sequence of base pairs relatively consistent within species. Therefore, these primary goals of the Human Genome Project in many ways served to define the size of the genome. It had been estimated indirectly for many years that there were approximately three billion base pairs and 100,000 genes. But it was unclear if this was correct. Furthermore, the exact sequence and identity of these bases was unknown, as was the identity of the genes. Such information would be of great value in the challenge of mapping disease genes as outlined elsewhere in this issue. Such a problem would change from one of unknown scope to one of large but finite scope. More importantly, it would efficiently direct researchers to the genes of interest within chromosomal regions of interest.

Sequencing the human genome

So how was a project of such enormity approached and accomplished? Two basic strategies were taken towards the sequencing of the genome. One, termed 'top-down', was pursued by a large consortium of academic centers. The other, 'bottom-up' or 'whole genome shotgun' was pursued by a private company—Celera. In the top-down approach, chromosomes were assigned to different centers, and each chromosome then broken up into smaller pieces, which were ordered into a physical map. These smaller pieces were approximately 100,000 base pairs in length and cloned into vectors called bacterial artificial chromosomes (BAC). Once the order of the BACs was known, then each BAC was sequenced and the whole sequence finally assembled. In the whole genome shotgun approach, rather than dividing up the project into medium size pieces, the entire genome was fragmented into many very small pieces, each about 500 bases in size. These were then all sequenced in a very high throughput parallel fashion. Then computers were used to arrange these millions of snippets of sequence together based on overlapping regions. The challenge and risk associated with this approach was that it was not clear that enough overlap would be present and computers would be fast enough to make such a massive reassembly possible. In the end, both approaches were successful (Venter *et al*., 2001; Lander *et al*., 2001). In fact, their complementary strengths and weaknesses made the final combined sequence more accurate.

The completion of the sequencing led to a number of surprising results. While it had long been indirectly estimated that there were approximately 100,000 human genes, in fact, there turned out to be only about 35,000 genes. These genes underwent extensive splicing in which different exons were included in the final mRNA to produce functionally different proteins. Though alternate splicing was a well-known phenomenon, it is more extensively employed in order to achieve protein diversity in man. Other intriguing results included numerous footprints of viruses and other rogue DNA elements that had inserted themselves into human DNA through the millennia. Approximately, half of these genes have been shown to be expressed in the brain and are of potential relevance to psychiatric illness. The product was a spectacular accomplishment and a road map for gene mapping.

Mapping human variation

Facilitating the mapping of disease genes was a driving force behind the Human Genome Project. Human diseases and all human traits result from genetic variation in human populations. Therefore, in order to further advance disease mapping, the next goal of the Human Genome Project was to characterize this variation in different populations. Knowledge of such variation is essential to mapping in two ways. First, a given base pair substitution, might actually be the mutation that changes the proteins function and directly contributes to disease susceptibility. Alternatively, and much more likely, it may merely be nearby the functional variant and serve as a marker. The genetic strategies of linkage and linkage disequilibrium are described elsewhere in this issue. Both approaches rely on genetic markers to map illnesses. The success of disease gene mapping over the last two decades has resulted directly from the development of methods to study such markers and determine their genotype.

The vast majority of variation in the genome consists of one base being substituted for another. These have recently been termed single nucleotide polymorphisms (SNPs). Much of the current focus of the Human Genome Project has been the identification of a large number of SNP markers for mapping purposes. Large scale sequencing capacity has been shifted from sequencing the human genome in general to sequencing many humans from different racial and ethnic backgrounds in order to identify a large number of SNPs. To date over nine million SNPs have been identified; over four million of these have been validated (www.ncbi.nlm.nih.gov/SNP).

This provides a very high-density map of markers for disease mapping and will ultimately identify most of the functional variants in the human genome.

Mapping by linkage disequilibrium relies on the non-random co-occurrence of SNP variants. If two SNPs are close enough to each other in the genome, in the order of several thousand bases, then they will be propagated together through successive generations in the population. With many generations, eventually they may occur together no more often than expected by chance at which point they are said to have reached linkage equilibrium. The further apart they are in the genome the faster in terms of generations that this will be achieved. Linkage disequilibrium can be used to map disease genes by testing for association of a specific marker allele with a disease. If the marker is close enough to the disease susceptibility mutation, then it will be observed more often among ill individuals. The advantage of this approach is that the marker must be very close for this to be detected making a high resolution of mapping possible.

However, linkage disequilibrium has been shown to not be uniformly distributed across the genome. In much of the genome, it occurs in blocks, termed haplotype blocks, where SNP variants occur together in a limited number of haplotypes (Gabriel *et al.*, 2002). In other regions, linkage disequilibrium is much less preserved. Figure 1 illustrates the haplotype block structure across a region of chromosome 22. Haplotype blocks are of great value in mapping studies because a larger region can be tested using a smaller number of markers. It is also important in the analysis of such data to understand the haplotype block structure in a region in order to more efficiently analyze association data for the presence of susceptibility genes. For these reasons, a large international effort, the HapMap project, has begun to genotype millions of SNPs in several different populations (Anonymous, 2003). These data will then be used to determine the extent and structure of linkage disequilibrium in different populations. The exact origin of this haplotype structure is currently unclear. It may arise from recombination hotspots in the genome, where crossovers between chromosomes are disproportionately likely to occur and disrupt linkage disequilibrium. It may also be the result of the history of populations as they have migrated and gone through various population bottlenecks. This work may help elucidate some of these mechanisms. It may also provide intriguing information about the ancient history of different human populations as recorded in patterns of DNA variation.

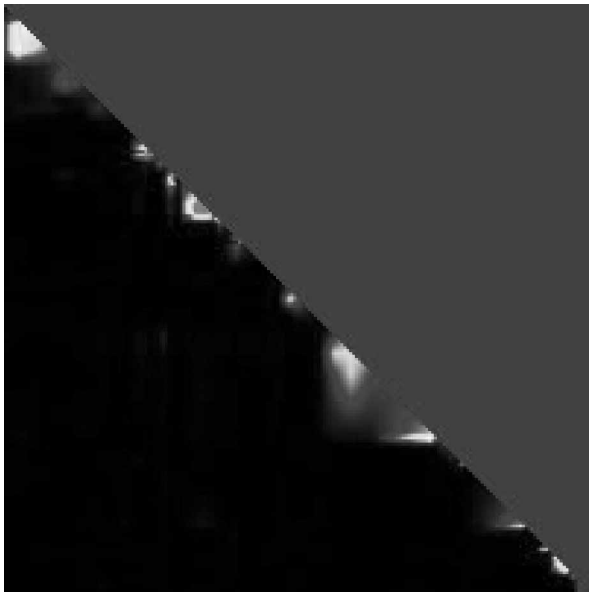

Figure 1. *Haplotype block structure of a region of chromosome 22. The figure illustrates the haplotype block structure of an approximately 1 Mb region on chromosome 22. Each point in the figure represents the strength of linkage disequilibrium between SNP markers whose positions on each axis intersect at that point. Hence, the other half of the diagonal would be identical. Strength of linkage disequilibrium is depict from black to red using a color scale. Regions of high linkage disequilibrium fall into blocks, interspersed by non-block regions of low linkage disequilibrium.*

Application to mapping disease genes

Data and tools from the Human Genome Project accelerates gene mapping in several major ways. The first is fine mapping of genes within linkage peaks. Linkage peaks in psychiatric disorder have tended to be quite large, often extending over 20–30 million base pairs. This is consistent with other complex genetic traits and with theoretical predictions (Terwilliger *et al.*, 1997). Often such peaks can contain hundreds of genes. Prior to the Human Genome Project, investigators had to first physically map the region of interest by cloning it into BACs or other similar vectors and sequence large regions of it. The SNP markers would have to be identified for association studies. Other methods such as exon trapping and cDNA selection were employed to identify possible genes in the region. The Human Genome Project has obviated most of these steps. The exact contents of the region of interest in terms of genes and sequence are now available. Possible candidate genes within the region and their intron/exon structure are now known. The tissues where they are expressed are known. Most regions are now covered by a very high density of SNP markers suitable for association and linkage disequilibrium studies. Soon the haplotype block structure for most regions will be available, as well as knowledge of the most informative SNPs. Together, these tools dramatically accelerate the process of fine mapping of disease genes. Once a likely gene is identified, the next stage in developing proof of the gene's involvement is identifying a likely functional mutation. This requires extensive sequencing of the gene in ill individuals. The complete sequence of the gene

enables investigators to readily amplify and sequence the regions in question and to have sequence for comparison.

However, the most exciting contribution of the Human Genome Project to gene mapping is the prospect of completely novel approaches to mapping. Most notable among these is whole genome association. Risch and Merikangas (1996) proposed that with a large enough number of markers, it would be possible to survey the entire genome using association. Association has the advantage of having a much higher sensitivity to small gene effects. Linkage is very powerful for single gene disorders, but it is becoming increasingly apparent that psychiatric disorders, as with other complex traits, are likely more polygenic in nature with many genes each contributing a small amount. Association also has the advantage of operating at the gene level of resolution. If dozens or hundreds of genes are involved in illness, then a linkage peak may contain several susceptibility genes and be unable to resolve a single locus. The limitation to this strategy to date has been the large number of markers necessary. If linkage disequilibrium extends over 10–50 kilobases on average, then 50,000–500,000 markers may be necessary in order to cover the entire genome. Recently, the number of available SNP markers has reached a density in this range. The HapMap project also will soon deliver necessary information about the best SNPs to choose. Moreover, new developments in technology now make it possible to genotype such a large number of markers.

Technology

The Human Genome Project has advanced hand in hand with the development of high throughput technologies. Completion of the sequence was aided greatly by the development of capillary-based electrophoresis methods for sequencing and high throughput robotics. However, some of the most dramatic advances in technology have been the development of micro-arrays. Methods have been developed to place thousands of DNA probes on a slide or a chip to be used for either genotyping SNPs or measuring RNA. In one approach, thousands of small fragments of DNA or oligonucleotides are synthesized in a massively parallel fashion onto microchips (McGall & Christians, 2002). This approach borrows photolithographic methods from the semiconductor industry in order to achieve a high level of miniaturization. Such oligo probes can then be used to measure RNA levels as described below or to genotype SNP markers, termed a SNP chip. The RNA is labeled with a fluorescent tag and hybridized to the chip. After non-specifically bound RNA is washed off, the chip is read with a laser scanning reader to read the levels of thousands of genes simultaneously.

An alternative technology for micro-array based RNA quantification involves the robotic spotting of thousands of cDNA, or DNA copies of mRNA's, onto a glass slide. Typically, two RNA samples are compared and each labeled with a different color fluorescent probe. The slide is washed to remove non-specifically bound RNA and read using a laser scanning reader. The ratio of RNase is determined as a ratio of the signals from the two fluors.

Several strategies have also been developed for high throughput chip based genotyping. Some rely on the SNP decreasing the hybridization efficiency under specific conditions such that the labeled DNA does not bind (Matsuzaki *et al.*, 2004). Using this approach it is now possible to genotype approximately 100,000 SNPs individually. Figure 2 illustrates a SNP chip that can genotype 10,000 SNPs simultaneously. Other strategies pool many DNA samples from ill individuals and estimate the allele frequencies based on the differential signal for the SNP allelic variants. Over 1.7 million SNPs can be read in this fashion using some current technologies (Patil *et al.*, 2001). Another widely employed SNP genotyping methods employ fiber optic bundles. Beads with specific oligo tags on them are bound to a fiber optic bundle (Oliphant *et al.*, 2002). Each bead is sized so that only one will bind to the end of each of 64,000 fibers in the bundle and can be individually read out by fluorescent methods. The SNPs are detected by a ligation based

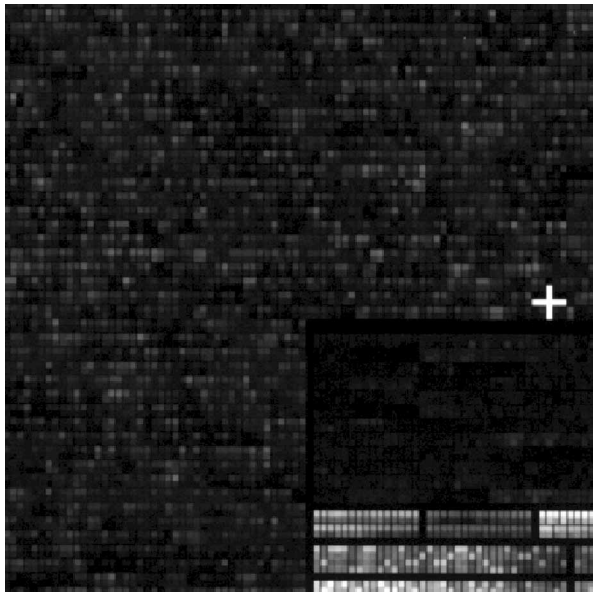

Figure 2. *SNP chip close-up. A portion of the Affymetrix 10K human SNP chip is illustrated with substantial magnification. Each small square represents a distinct oligonucleotide that specifically detects a single SNP based on hybridization intensity. The fluorescently labeled test DNA sample will hybridize to the oligo probe on the chip only if the SNP base is an exact match. In order to maximize the ability to detect SNPs, each SNP is interrogated with several oligo probes. The rectangle of markings on the lower right are calibration signals to line up and register the chip for accurate reading of features and to calibrate it for signal intensity.*

chemistry and then hybridized to specific tags on the bundle. These exciting new methods provide the high throughput necessary to measure genes and SNPs at the genomic level.

Functional genomics and expression profiling

The invention of the micro-array RNA quantification methods described above has led to the development of an entirely new approach to gene function at the genomic level. It is now possible to simultaneously measure the levels of mRNA for all genes in the genomes of several species including man and mouse. Levels of RNA of course do not access all aspects of gene function. Gene function is regulated not only in how much RNA is made but also in the splicing of that RNA. Many levels of regulation also exist in the translation of the mRNA into protein, post-translation modification of the protein and regulation of the proteins function itself through phosphorylation and many other mechanisms. Nevertheless, transcription is a very important and initial stage, which can now be approached with an unprecedented level of comprehensiveness.

Some examples may help illustrate the usefulness and application of this method. An animal model can be used to determine which genes in the genome are involved in pathways functionally relevant to psychiatric illness. We have previously employed acute amphetamine administration as a model of mania (Niculescu *et al.*, 2000). Amphetamine recapitulates many of the symptoms of mania including: increased energy, decreased need for sleep, irritability, grandiosity, euphoria, risk taking, racing thoughts and increased speech. It has also been proposed that chronic treatment models the evolution of mania into psychotic mania. We treated rats with methamphetamine and sacrificed them 24 hours later. Twenty four hours was chosen as the time point at which they would already have begun to show a sensitized response to amphetamine and a time point where adaptive as opposed to immediate early gene responses might better be observed. The RNA was prepared from prefrontal cortex and amygdala, fluorescently labeled and hybridized to an oligonucleotide-based chip. Out of 8,000 genes examined on the chip, several genes had large changes in expression. One of these, G-protein receptor kinase 3 (GRK3), mapped to a region on chromosome 22 that had already showed evidence of linkage in a set of 20 extended bipolar families (Kelsoe *et al.*, 2001). This enabled us to identify this gene out of the hundreds in this region as being functionally relevant to bipolar disorder. Based on this information, we sequenced the gene in bipolar patients from families with evidence of linkage and found several SNP variants in the promoter of the gene or that region involved in its transcriptional regulation. One of these SNPs then showed association in two independent samples of families (Barrett *et al.*, 2003). This evidence strongly suggests the role of the GRK3 gene in bipolar disorder, and it may be one of the first susceptibility genes identified. G-protein receptor kinase 3 is a particularly intriguing candidate as it is involved in the desensitization of a variety of G protein coupled receptors as part of a homeostatic response to agonist stimulation. Our hypothesis is that a defect in this gene leads to a failure of homeostatic regulation and a super-sensitivity to dopamine and other neurotransmitters. We have termed this approach of combining functional genomics studies with linkage and positional cloning, as Convergent Functional Genomics. It holds promise to facilitate the identification of other disease genes.

Another major application of expression profiling to psychiatric illness is the study of post mortem brain. The expression of all genes in the genome in relevant brain regions can be compared between patients with psychiatric illness and matched control subjects. The collection of such samples is quite difficult and sample sizes for these studies tend to be small. But there is no substitute for examination of human brain despite numerous confounds such as time after death, uncertainty of diagnosis and drug treatment. Mirnics *et al.*, (2001) examined brain regions of patients with schizophrenia as compared to controls. They identified the gene RGS4 as being differentially expressed in schizophrenia. Using a similar strategy as that described above for bipolar disorder, they compared the chromosomal map position of the RGS4 gene to linkage results for schizophrenia and found that it mapped near a locus on chromosome 1. Subsequent studies of SNP markers in the gene have shown evidence of association suggesting that it may be a susceptibility gene for schizophrenia (Chowdari *et al.*, 2002).

More broadly, micro-array expression profiling studies can be used to identify genes and networks of genes that are relevant to the pathophysiology of disease or drug action. Comparison with positional cloning results as above may help identify susceptibility genes, but the broader goal of both positional and functional studies is to better understand functional pathways. The real power of the functional genomics approach is to identify how genes may interact with each other in functionally relevant pathways. Completely novel pathways may thereby be discovered. These may point to possible susceptibility genes, they may also point to novel points of intervention for the development of therapeutics with completely new mechanisms of action.

Bio-informatics

The common element among all of genomics is the vast amount of data generated. The management of such data becomes a formidable challenge in and of itself. Completely new statistical approaches must be

developed in order to understand such large amounts of data. This challenge has in large part spawned the new and rapidly expanding field of bio-informatics.

One of the first major challenges facing the Human Genome Project was managing the large amounts of data generated by the sequencing effort. For the 'top-down' approach, the location and sequence of each BAC or other cloning vector had to be tracked. Their relationship to each other had to be determined by seeking sequence overlap and linking them to each other in a 'tiling' pattern. The computational effort for the 'bottom-up' or 'whole genome shotgun' approach was even larger. One of the largest assemblies of computers and greatest computational capacity had to be created in order to assemble the millions of small fragment sequences into entire chromosomes. Once the sequence was assembled, new analytic and computational methods had to be developed in order to better understand it. Thousands of sequences of cDNAs representing expressed genes and coding sequence had to be mapped to the genomic sequence. This process and others required for the understanding of the genome has been termed 'annotation' of the genome. This led to a map of all the genes and all their exons on all the chromosomes, which has been invaluable in disease gene mapping. Methods have also been developed to predict the location of an exon based on statistical sequence commonalities even in the absence of an actual cDNA sequence.

A significant challenge now facing such annotation efforts is an understanding of the regulatory regions in the genome. Such sequences are highly variable and this variability has made it quite difficult to predict their presence based strictly on sequence. It also remains unclear as described above exactly how much of the non-coding genome may serve such functions. This is a large, difficult and important challenge. It may be particularly relevant to complex genetic traits, as mentioned above, and it has been suggested that regulatory mutations may be more common in such traits.

Similar challenges are faced in the management of data regarding the millions of SNPs identified to date. Simply naming, mapping and tracking them in an organized fashion is formidable. Large-scale studies of haplotype block structure have required the development of entirely new statistical approaches. How should a haplotype block be defined? How are its borders defined? Methods to map genes using large numbers of such markers are just now being developed. It is unclear how best to employ such haplotype block information in analyses of disease association. The coming prospect of whole genome amplification promises data management and analysis challenges that are orders of magnitude greater. How does one track and manage data on 500,000 SNPs in several thousand individuals? How is this data best analyzed? The most serious problem of all may be dealing with the multiple comparisons issues. If 500,000 tests are conducted then under simple assumptions, 25,000 of them will be false positives at a $p < 0.05$ level. Appropriate correction for such a large number of tests will require extremely low p values. Very large sample sizes may be necessary in order to achieve this. In other words, in order to achieve the hoped for benefits of whole genome association analyses, the overall scale of the project will be much larger than those conducted to date.

Analysis of micro-array data also presents numerous new challenges of scale. A typical experiment may generate 25,000 data points for each of 5–20 samples examined. Each chip may vary in its overall level of hybridization; therefore, data must be normalized to a common standard so that it can be meaningfully compared. Many such approaches exist, but the optimal one is not clear at present. A simple comparison of 25,000 genes presents a formidable multiple comparisons problem. How strong must a result be in order to be meaningful in the context of so many tests? A variety of simulation methods have been developed in order to empirically determine this. One of the greatest problems facing such experiments is simply naming things. Genes, cDNA, genomic sequences all have evolved numerous names, numbers and identifiers. Matching these correctly with each other is key, and challenging. In order to better understand the volumes of micro-array data being generated, methods have been developed to record experimental conditions and results in a standardized fashion so they can be curated in databases for meta-analyses.

Lastly, as described above, the important results from such experiments are not about individual genes, but rather about biological pathways and networks of genes. Little is currently known about the function of most genes. Given that many or most genes may serve multiple even diverse functions, this is a large gap in knowledge. Array experiments promise the discovery of new pathways, but novel statistical methods must be developed in order to make sense of the emerging data in the context of existing biological data and in order to create new functional pathways. These in turn will be used in order to interpret and expand on pathways already known.

The future of genomics

One future direction already well underway is the sequencing of other species. In addition to the human genome those of mouse, rat, dog, chimpanzee, *Drosophila*, *C. elegans* are either already complete or underway. These data will be invaluable for at least two major reasons. First they will facilitate genetic mapping experiments in these valuable model organisms. Secondly, comparison of man to

other species will aid in understanding the functional role of genomic regions. One of the most powerful ways to identify likely functionally relevant regions is the observation of their evolutionary conservation.

The steady advance of technology also promises to make experiments only now dreamed about, soon a reality. In particular this is true for high throughput genotyping and its impact on the feasibility of whole genome association analysis. These experiments using current technology and at current prices are in the tens of millions of dollars. As the cost of genotyping drops below $0.01/genotype, then such very large experiments will begin to be more tractable. It is not hard to imagine in the not too distant future that obtaining the complete DNA sequence of a patient may be part of a routine workup.

Ultimately, the problem of complex genetics must have a large though finite scope. There are a very large though finite number of SNPs in the world's population. If all SNPs were known, then all possible functional SNPs would number among them. The problem boils down to mapping these SNPs to the traits of interest in a large enough number of people. Though this seems science fiction in the size of the project, so did many other things in the not too distant past that have now been accomplished. Amazing things likely lie ahead.

Web sites of interest

The human genome can be accessed via several web sites:

- www.ensembl.org
- www.hapmap.org
- www.ncbi.nlm.nih.gov
- www.genome.ucsc.edu

References

ANONYMOUS. (2003). The International HapMap Project. *Nature*, *426*, 789–796.

BARRETT, T.B., HAUGER, R.L., KENNEDY, J.L., *et al.* (2003). Evidence that a single nucleotide polymorphism in the promoter of the G protein receptor kinase 3 gene is associated with bipolar disorder. *Molecular Psychiatry*, *8*, 546–557.

CHOWDARI, K.V., MIRNICS, K., SEMWAL, P., *et al.* (2002). Association and linkage analyses of RGS4 polymorphisms in schizophrenia. *Human Molecular Genetics*, *11*, 1373–1380.

GABRIEL, S.B., SCHAFFNER, S.F., NGUYEN, H., *et al.* (2002). The structure of haplotype blocks in the human genome. *Science*, *296*, 2225–2229.

KELSOE, J.R., SPENCE, M.A., LOETSCHER, E., *et al.* (2001). A genome survey indicates a possible susceptibility locus for bipolar disorder on chromosome 22. *Proceedings of the National Academy of Science, USA*, *98*, 585–590.

LANDER, E.S., LINTON, L.M., BIRREN, B., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921.

MATSUZAKI, H., LOI, H., DONG, S., *et al.* (2004). Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome Research*, *14*, 414–425.

MCGALL, G.H. & CHRISTIANS, F.C. (2002). High-density genechip oligonucleotide probe arrays. *Advances in Biochemical Engineering & Biotechnology*, *77*, 21–42.

MIRNICS, K., MIDDLETON, F.A., STANWOOD, G.D., LEWIS, D.A. & LEVITT, P. (2001). Disease-specific changes in regulator of G-protein signaling 4 (RGS4) expression in schizophrenia. *Molecular Psychiatry*, *6*, 293–301.

NICULESCU, A.B., III, SEGAL, D.S., KUCZENSKI, R., BARRETT, T., HAUGER, R.L. & KELSOE, J.R. (2000). Identifying a series of candidate genes for mania and psychosis: a convergent functional genomics approach. *Physiology & Genomics*, *4*, 83–91.

OLIPHANT, A., BARKER, D.L., STUELPNAGEL, J.R., & CHEE, M.S. (2002). BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques*, *Suppl 56–1*.

PATIL, N., BERNO, A.J., HINDS, D.A., *et al.* (1996). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, *294*, 1719–1723.

RISCH, N. & MERIKANGAS, K. (1996). The future of genetic studies of complex human diseases. *Science*, *273*, 1516–1517.

TERWILLIGER, J.D., SHANNON, W.D., LATHROP, G.M., *et al.* (1997). True and false positive peaks in genome-wide scans: applications of length-biased sampling to linkage mapping. *American Journal of Human Genetics*, *61*, 430–438.

VENTER, J.C., ADAMS, M.D., MYERS, E.W., *et al.* (2001). The sequence of the human genome. *Science*, *291*, 1304–1351.