Impact of the genome project on the identification of disease genes

Cognitive defects and neurological diseases represent a major issue for human health, especially in aging populations. An estimated 15% of people >65 years are affected by mild-to-severe conditions of genetic origin affecting the central nervous system. Etiological factors of common neurological and psychiatric disorders remain elusive, apart from a few genes associated with rare disorders, such as one form of Alzheimer's disease *(APP)*, a form of amyotrophic lateral sclerosis *(SOD1)*, expanded polyglutamine track in Huntington's disease, and several types of ataxia or ion channel–associated conditions. With the human DNA sequence unveiled as a huge book of 3 gigabases,[1,2] how can we exploit the genome readout to identify disease-associated alleles and what is the projected impact for clinical genetics? The "book of life" is complete to >90% of euchromatic gene-rich regions, opening unprecedented possibilities for the characterization of all genes. The emerging human catalogue is thought to contain about 30 000 genes. Until now, factors underlying inherited conditions were mostly identified by positional cloning without prior knowledge of their biochemical function, and the catalogue unlocks the door to fast in silico searching *(Figure 1, Part I)*.

Complex molecular processes govern organogenesis and fitness builds upon the correct orchestration of gene actions throughout life. Most clinical phenotypes result from alterations of genetic instructions perturbing this tightly regulated system, while being strongly influenced by individual genetic makeup. The profound transition seen with the sequence information is the ability to foster novel concepts in our way of addressing biology as a global entity. Comprehensive studies of genome landscape and common polymorphisms will help identify causal and susceptibility factors at a much greater pace *(Figure 1, Parts II and III)*.

Although 60% of human genes have no characterized function yet, the sequence provides a body of information for the design of global strategies in functional genomics, for instance, using molecular evolution to underpin function by inference. Comparative genomics is one of the most powerful approaches to deciphering the molecular basis of disease pathogenesis *(Figure 2)*.

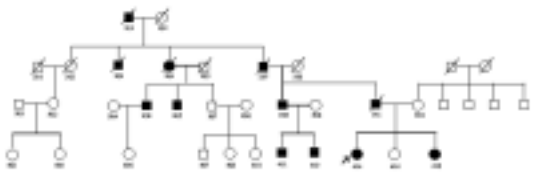
Another essential approach to extracting biological meaning from the genetic message is illustrated by global transcriptome analysis *(Figure 3)*. Grasping how global gene expression is processed into phenotype will be essential to any progress in molecular medicine. Hunting for disease-associated alleles by surveying dynamic biological systems at all relevant developmental stages and in all relevant tissues brings novel perspectives that will allow the correlation of molecular phenotype with clinical phenotype.

Perspectives

Dissecting the complex genetic architecture of common diseases represents a massive endeavor that will profoundly influence the next decades of research in molecular medicine. The strategic approaches described here will become incredibly informative when integrated with proteosome studies clinical records, neuroimaging data, and physiology. Genome research and bioinformatics are the cement bridging all these disciplines together toward the establishment of disease profiles, from molecules to phenotypes, for assessing disease susceptibility, developing accurate diagnosis, and novel personalized treatments.

Poster by: Marie-Laure Yaspo, PhD
Max-Planck-Institut für Molekulare Genetik
Ihnestrasse 73, D-14195 Berlin-Dahlem, Germany
(e-mail: yaspo@molgen.mpg.de)

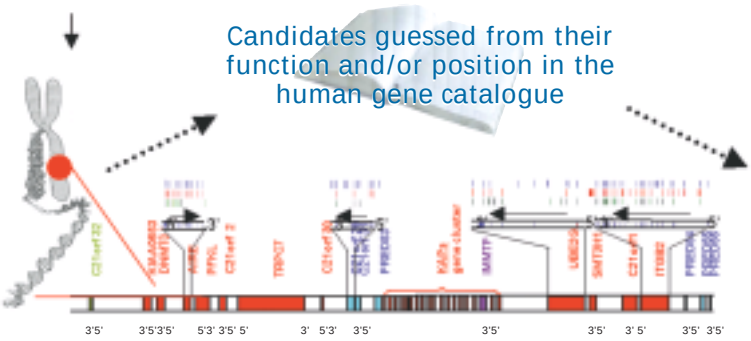
I. Linkage analysis or genome-wide scanning followed by positional/in silico cloning strategy



- Monogenic disorders, when a single gene is necessary and sufficient to cause a disease
  - Rare genetic disorders with a mendelian pattern of inheritance
  - Genetic instabilities (eg, triplet expansion in neurodegenerative conditions, such as Huntington disease)
- Disease loci mapped by chromosomal rearrangements (eg, translocation breakpoints in cancer)
- Modifying factors associated with multifactorial diseases (in rare cases when mapping was possible)

Chromosomal location

Candidates guessed from their function and/or position in the human gene catalogue



Gene predictions using computer programs

May 2001: 1500 phenotypes were mapped and 1100 disease genes identified in the Online Mendelian Inheritance in Man database (OMIM)[3]

- 343 listed phenotypes match the keywords "mental retardation" (168), "psychiatric" (21), or "neurologic" (154).[4]
- Sequence information expedites positional cloning strategies, shifting from prior gene-finding methods (eg, exon trapping, cDNA selection) to "in silico cloning," by scanning critical regions for known genes or novel predicted genes. Matching cDNA sequences and clones are often available from distribution centers,[5] bypassing tedious cloning steps for the users. With the gene structure information at hand, mutations can be searched effectively, also for those genetic lesions falling outside of coding regions (splice, regulation sites, etc).

II. Assessing the impact of common sequence variants (SNPs) on phenotypic traits

Single nucleotide polymorphisms (SNPs) are distributed across the genome. Dense and precise SNP maps will pave the way to indexing global haplotype variations, essential bases for population genetics, and epidemiological studies

1.42 million SNPs in the genome;
only 60 000 SNPs in exons[6]

Disease-modifying factors
Linkage disequilibrium

III. Identification of susceptibility factors in complex disorders

Most common diseases, including behavior and psychiatric disorders, show a complex inheritance pattern resulting of subtle interactions between several genetic loci and environmental factors. Linkage is often not possible since the penetrance effect of individual genes is not large enough. Allelic association can detect genes that account for as little as 1% of the variance in a given trait; polymorphisms with phenotypic associations allow the identification of individual variations in the genome and the assesment of the contribution of different genes in different families. The human sequence, the availability of SNP maps, and novel high-throughput genotyping systems will boost the power of association studies using large cohorts, allowing the analysis of multiple variations for locating quantitative trait loci (QTL) components. Deciphering genetic repertoires is a key step to understanding how variations in our genetic program influence disease development, progression, and behavior.

The challenge for the next decade will be to establish a genetic risk or predisposition profiling for complex diseases. However, this approach has intrinsic limitations. Progress on the genetics of complex diseases will largely come from the analysis of animal models recapitulating human pathogenic traits (see next page).

Figure 1. The human genome catalogue unlocks the door to fast in silico searching and the design of novel high-throughput genotyping strategies.

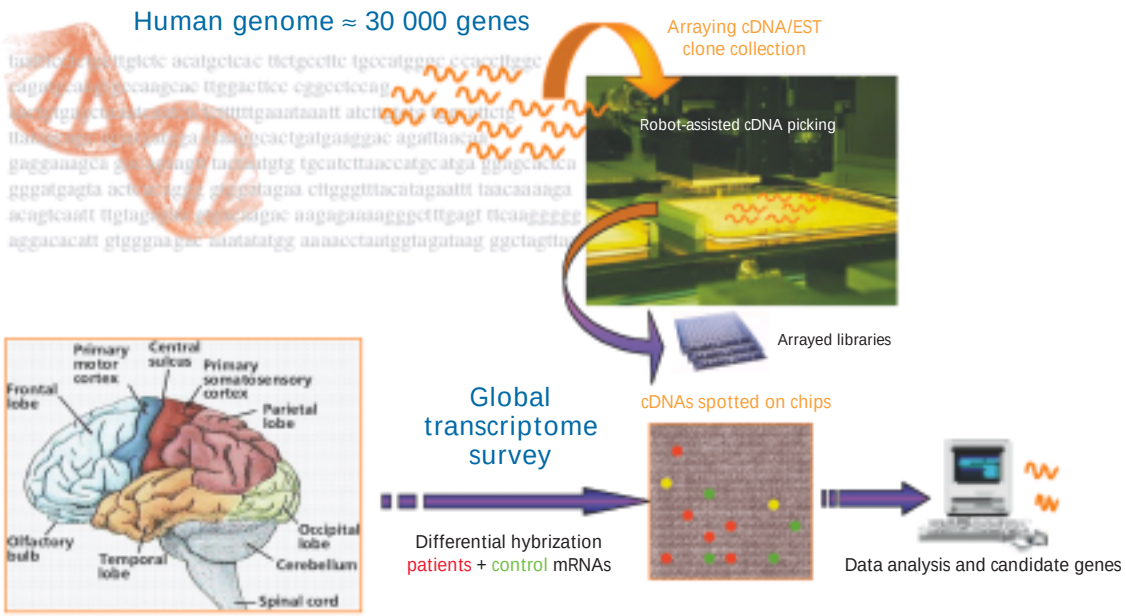
Animal models—spontaneous or engineered mutants (eg, knockout)—are essential tools for dissecting components of complex diseases. Mouse genetics allows the mapping of candidate genes (quantitative trait loci), and back testing of orthologous genes in patient cohorts.

The large mutant collections obtained in *Drosophila* represent a breakthrough for deciphering gene function, in pathways involved in development and neurogenesis.[7]

Human diseases

The complete DNA sequences of about 60 prokaryotic and eukaryotic organisms are now available. Comparing genomic sequences of different species helps to identify genes, since protein-coding exons are usually more conserved than the rest of the genome. Comparative mapping is used for establishing correspondence of disease loci.[8] Mapping of genes in one species often allows the identification and testing of the corresponding genes in humans as candidates for involvement in a disease. Besides, the networks of processes disturbed in a disease can only be understood by using powerful genetic tools available in model organisms like mouse or even in the evolutionarily distant *Drosophila melanogaster* or *Caenorhabditis elegans*. Since a number of functional networks have been remarkably conserved during evolution, the transfer of data from model organisms to humans represents an efficient route to elucidating the molecular basis of genetic diseases. For instance, striking parallels between human, mouse, and *Drosophila* are observed for genes involved in the development of the central nervous system (eg, *sonic, Gli*). Further information can be found in specific genome databases; the main addresses are listed on the Internet.[9]

Figure 2. Genome sequences boost the power of model organisms and comparative genomics for identifying disease genes and understanding their function.



Transcript profiling and molecular phenotyping: chip-based strategies allow the monitoring of variations of gene expression in affected tissues (eg, brain regions) by analyzing several thousands genes at a time; anatomical maps of gene expression can be established. Genes that are either down- or upregulated in a given condition will identify a pool of potential candidates. Relevant cDNAs are isolated from their original array position, giving immediate access to cloned material. This approach provides essential entry points for dissecting the molecular basis of disease pathogenesis and deciphering affected pathways. Global transcript profiles can help stratifying patient samples according to their molecular phenotypes and defining disease-specific subtypes, eg, to classify cancer subtypes.

Figure 3. Global analysis of the transcriptome by complex hybridization on assays: identifying and spotting all of the ≈ 16 000 to 20 000 genes that could be expressed in the human brain.

REFERENCES
1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature.* 2001;409:860-921.
2. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science.* 2001;291:1304-1351.
3. National Center for Biotechnology Information. Online Mendelian Inheritance in Man. http://www.ncbi.nlm.nih.gov/Omim. Accessed July 2, 2001.
4. Antonarakis SE, McKusick VA. OMIM passes the 1000-disease-gene mark. *Nat Genet.* 2000;25:11.
5. Deutsches Ressourcenzentrum für Genomgorschung GmbH. Resource Center/Primary Database. http://www.rzpd.de. Accessed July 2, 2001.
6. The International SNP Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature.* 2001;409:928-933.
7. Berkeley *Drosophila* Genome Project. http://www.fruitfly.org/. Accessed July 2, 2001.
8. Mouse Genome Informatics. The Jackson Laboratory. http://www.informatics.jax.org. Accessed July 2, 2001.
9. Japan Science and Technology Corporation. The Genome Database. Human Genome Project Resources. http://gdb.jst.go.jp/gdb/hgpResources.html. Accessed July 5, 2001.