# Disease gene-oriented genomic sequence analysis of medically important regions of the human genome and homologous regions of the mouse genome.

**Helmut Blöcker** [1], **Hans Lehrach** [2], **Matthias Platzer** [3]
[1] GBF – Gesellschaft für Biotechnologische Forschung, Abt. Genomanalyse, Braunschweig
[2] Max-Planck-Institut für Molekulare Genetik, Berlin [3] Institut für Molekulare Biotechnologie (IMB), Abt. Genomanalyse, Jena

Since the DNA sequence of the genome of each organism contains the blueprint for all biological processes, knowledge of the entire genome sequence of key organisms will be an essential complement to the information generated on transcript, on proteins, and on mutant phenotypes. Especially sequence comparison between genomes can identify functionally important elements as regions of sequence conservation, and therefore play an essential role in identifying new biological processes. Ultimately, only the genomic sequence will allow the identification and analysis of most regulatory elements, essential to understand many biological processes. Genomic sequence analysis makes a major contribution to gene discovery in general since EST/cDNA sequencing and chip-based RNA expression profiling programmes may not be appropriate to unravel genes expressed primarily at specific time points of the development or at a very low level. The genomic sequence also constitutes an essential link between clinical genetics, allowing the identification of chromosomal regions involved in specific diseases, or carrying specific mutations in model organisms, and the molecular identification of the specific gene or genes responsible for the particular disease or phenotype.

The German Genomic Sequence Analysis Consortium (GGSAC) was founded in 1997 by three academic groups from the Institute for Molecular Biotechnology (IMB, Jena), the Max-Planck-Institute for Molecular Genetics (MPIMG, Berlin) and the German Research Centre for Biotechnology (GBF, Braunschweig). The GGSAC is characterised by its visible contributions to international large scale sequence analysis (human chr 21: Nature 2000, draft of the human genome: Nature 2001) as well as functional analysis projects. In the framework of the DHGP we have established sequence contigs of medically important regions on human chromosomes 3, 8, 9, 17, and X. Our genomics-based approach has discovered among others disease genes for NBS (Cell, 1998), TRPS1 (Nat. Genet., 2000), HHD (Hum. Mol. Genet., 2000), HMSNL (Am. J. Hum. Genet., 2001), BSND (Nat. Genet., 2001) and PCD (Nat. Genet., 2002). For selected loci comparative sequence analyses of orthologous regions on mouse chromosomes are in progress.
After publishing a draft sequence of the human genome in spring 2001 (IHGSC, 2001), which the German consortium (IMB, MPIMG, GBF) contributed to by sequencing regions of chromosomes 3, 8, 9, 17, 21 and X, the focus of the International Human Genome Sequencing Consortium (IHGSC) has shifted to generating contiguous and high qualitiy sequence data, the so called "finished" sequence, which will be done by 2003. In the framework of the second phase of the DHGP the targets of the genomic sequence analysis consortium comprise the finishing of selected disease-related regions of human chromosomes 3, 8, 9, 17. Moreover, the project aims at the sequence analysis of the homologous regions in the mouse genome to enable functional studies like deletion mutagenesis in medically important mouse chromosome regions, knock-out/knock-in experiments and to identify novel elements involved in gene regulation.

IMB (Jena) targets are the finishing and analysis of selected disease-related regions of human chromosome 8: p22-p21 (defensin, DEF; keratolytic winter erythema, KWE), p21-p11 (fragile sites involved in breast cancer, BRCA) and q23-q24 (spastic paraplegia, SPG8).

A major target of the GBF (Braunschweig) is finishing of the region around the interferon gene cluster on human chromosome 9 (p arm) and its syntenic region on mouse chromosome 4. Moreover, we are functionally analysing the interferon gene clusters for conservation of higher order DNA structures, their nuclear localisation and function by a combination of techniques (SIDD, in vitro S/MAR-binding, EMSA and halo-FISH). Furthermore we plan to analyse a relevant portion of the IgH locus on mouse chromosome 12. The ultimate aim of this activity is to sequence and to annotate the immunoglobuline heavy chain (IgH) locus of the 129 mouse strain and to compare this region of this strain to the corresponding sequence of the C57BL/6 mouse strain, which is being sequenced by the mouse sequencing consortium. By comparing the IgH locus of the two strains we hope to be able to understand processes involved in the evolution of immunoglobuline variable regions.

The MPIMG (Berlin) targets are disease-related regions of human chromosome 3q (Malignant Hyperthermia Susceptibility MHS4; adolescent Nephronophthisis, NPHP3, Myotonic Dystrophy DM2, deafness DFNA18,

evolutionary breakpoints between human and orang-utan), and 17q (Smith-Magenis Syndrome) and other medically important regions (e.g. chromosome 1, Bartter Syndrome). Moreover, the project aims for the sequence analysis of regions in the mouse genome associated with phenotypes of the embryonic signal transduction in the notochord (chromosome 2, Danforth short tail phenotype, SD and chromosome 6, truncate phenotype, TC).

For the entire consortium, we proposed to deliver about 15 Mb of finished sequence and perform functional analysis in selected region or genes of interest, including the characterisation of the genomic organisation of genes, expression analysis, SNP and mutation detection, and genotype-phenotype correlation. The sequences represent the German contribution to the ongoing international sequencing projects of man and mouse.

With the complete finished human sequence and a working draft version of the mouse genome available by the end of 2003, disease gene-oriented projects need, in addition to this information, tools and facilities to step from a list of candidate disease genes to the disease causing mutations. Targeted sequencing of candidate genes in a cohort of patients is a proven quick and cost-effective method for identifying disease-causing variations. In the future we would like to study genetic loci related to deafness, cystic kidney diseases and primary ciliary dyskinesia by high-throughput targeted sequencing of patient and control cohorts.

International Human Genome Sequencing Consortium. **Initial sequencing and analysis of the human genome.** Nature 2001, 409: 860-921

The International Human Genome Mapping Consortium. **A physical map of the human genome.** Nature 2001, 409: 934-94

International Human Chromosome 21 Mapping And Sequencing Consortium. **The DNA sequence of human chromosome 21.** Nature 2000, 405: 311-319

For a complete list of publications see one of the following URLs:
**http://genome.imb-jena.de/publ/ggsacdhgp.html**
**http://genome.gbf.de/ggsacdhgp/pub.html**
**http://seq.molgen.mpg.de/hgs/publication.html**

In the core structure of the NGFN, rat, chimpanzee and rhesus monkey are regarded as model organisms with relevance for disease-oriented research. Currently, we contribute to the analysis of the chimp chromosome 22 (corresponds to human chromosome 21). We would like to extend our DHGP-funded comparative analysis of the human/mouse defensin gene cluster to non-human primates (Rhesus monkey), which are the most important model organism in immunological research to study viral infectious diseases. Finally, we will characterise in rat the MHC region and selected hot spot regions of chromosomal recombination, fragility, translocations related to cancer and immune response.

Moreover, the DHGP-based know-how and equipment may provide additional resources for disease gene oriented sequencing of patient and control cohorts in the DHGP-NGFN transition period in 2004.


Fig 1: **Nature cover (February 2001) on the occasion of the publication of the first "working draft" of the human genome.** Permission kindly granted by Nature and Macmillan Magazine Ltd.


Fig 2: **Nature cover (May 2000) on the occasion of the publication of the sequence analysis of the human chromosome 21.** Permission kindly granted by Nature and Macmillan Magazine Ltd.