SMP: DNA

Project: Sequencing/Finishing Large Regions of the Chimpanzee X Chromosome and Medically Relevant Loci of the Chimp Genome

Marie-Laure Yaspo – MPI for Molecular Genetics, Berlin - yaspo@molgen.mpg.de
Matthias Platzer – Leibniz Institute for Age Research - Fritz-Lipmann-Inst., Jena - mplatzer@fli-leibniz.de
Helmut Blöcker – Gesellschaft für Biotechnologische Forschung, Braunschweig - bloecker@gbf.de

Introduction

The chimpanzee (*Pan troglodytes*) is our closest living relative. It is generally accepted that comparative analysis of the human and chimpanzee genomes might help to identify human disease genes as well as genes of evolutionary interest. A finished chimpanzee sequence provides a unique angle from which to look at the human genome and to draw conclusions about our recent evolution. The longer-term hope is to identify those sequence changes that could account for unique human features, such as highly developed cognitive functions, the use of complex language, bipedalism or for the differences in disease susceptibility between the two species. Our knowledge about the human genome will be greatly advanced by the availability of a second hominid genome sequence.

Over the last two years, major national and international efforts have been carried out to decipher the chimpanzee genome. The first finished sequence of an entire chimpanzee chromosome, namely chr.22, was published in 2004 (Watanabe H et. al 2004) by a consortium coordinated by the RIKEN institute (Yokohama, Japan), and including the German Genome Sequence Analysis Consortium (GGSAC: MPIMG Berlin, FLI formerly IMB Jena, GBF Braunschweig), the MPI for Molecular Anthropology (Leipzig), the Genome Research Center (Daejon, Korea), the Chinese National Human Genome Center (Shangai), the National Health Research Institute (Taipei, Taiwan) and the NIG (Tokyo) In this study, the comparison of chimp chr.22 with its orthologous human chromosome (chr.21) revealed key features of the recent genome evolution (see below).

In parallel, a low coverage sequence of the whole chimp genome was produced in the USA. Since November 2003, alignments of the chimpanzee draft assembly to the human assembly (v. July 2003) were available (http://genome.ucsc.edu/cgi-bin/hgGateway) and a first comparison of the chimpanzee sequence draft with the finished human genome has been recently published (The Chimpanzee Sequencing and Analysis Consortium 2005).

Currently, an international consortium for finishing the chimpanzee genome is being set up and a clone repository is organised at the NCBI (http://www.ncbi.nlm.nih.gov/ genome/clone/). The German Genome Sequence Analysis Consortium (GGSAC: MPIMG Berlin, IMB Jena, GBF Braunschweig) aims at the generation of ~ 40 Mb of high-quality chimp sequence (1.5% of the genome), primarily from the X chromosome and selected medically-relevant regions. Compared to the autosomes, the sequence coverage of the X chromosome is significantly lower by a factor two and the gap frequency is higher due to the under-representation of chromosome X in the whole genome shotgun effort which was focussed on a male individual. Thus, gap closure and finishing efforts is particularly relevant for this chromosome. . There was a long-standing German contribution and ongoing interest in mapping (Heidelberg, Munich) and sequencing (Jena, Berlin) of the human chromosome X (Ross M et al 2005) as well as in studying X inactivation (Sudbrak R et al. 2001). This knowledge and the available clone resource, trace data and sequence assemblies will be instrumental for finishing the orthologous regions in the chimpanzee:

Chromosome X is of special interest with respect to the large number of X-linked hereditary diseases (838 entries in OMIM by 01/2004), in particular many forms of syndromic and non-syndromic mental retardation (for review see Chelly J and Mandel JL 2001). About 20 to 25% of all known loci involved in these disease forms are located on this chromosome and direct analysis of the genes and their promoter regions should be of great value for the scientific community. In addition, we propose to finish the sequence of selected autosomal regions of special interest for other SMPs or KGs within the NGFN2 clinical network.

It was commonly admitted that the chimpanzee sequence was about 98.8% identical to the human sequence (Chen et al. 2001, Am.J.Hum.Genet. 68: 444; Ebersberger et al. 2002, Am.J.Hum.Genet. 70: 1490; Sakate et al. 2003, Genome Res. 13: 1022), in contrast to dramatic biological differences seen between the two species. The outcome of the comparison of human chr.21 with chimp chr.22 showed that a high number of insertions and deletions occurred in both genomes, indicating a divergence of about 5% when we take these events into accounts (Watanabe et al. 2004). However, in most regions we found that 1.44 % of the chromsome consists of single-base substitutions. Our study showed that ca. 40 genes out of 232 analysed featured structural changes in their coding regions (Watanabe et.al. 2004). The 98.8% identity between man and chimp is thus valid for most genes but not all. Such striking differences have to be expected also in genes encoded by other chromosomes.

Recent studies, comparing a set of exons representing nearly 8.000 chimpanzee genes with their human and mouse orthologs revealed a set of gene classes with different patterns of substitution on the human lineage (Clark et al. 2003). In addition, gene classes with accelerated evolution could be identified, showing that sequence comparisons between man and chimpanzee is instrumental to shed light on key molecular events that occurred during recent evolution.

In addition, a body of evidences suggest that at least part of the evolutionary changes must have occurred at the level of gene regulation and extends the focus of interest also to analysing regulatory elements such as gene promoter regions and transcription factors.

Project Status

In 2002, the RIKEN institute (in coop. with the MPIMG) constructed and analysed a human-chimpanzee comparative clone map using 77.500 end-sequences from BAC clones of libraries PTB and RPCI-43 (Fujiyama A et al. 2002). In the first two month of the project (July 2005), the GGSAC started to identify large-insert clones by sequence comparison of the human sequence of chromosome X to the end sequences of clones from these two libraries. In addition, the end-sequences from the CHORI-251 library were used to complement the clone set (ftp://genome.wustl.edu/pub/seqmgr/chimp/BAC_ends/). A standardized procedure was established to retrieve all available trace data from the panTro1 chimpanzee working draft (November 2003) that will be used to achieve a mean coverage of 8-10, indispensable for the required accuracy of >99.99% of the finished sequence. The GGSAC works also in cooperation with RIKEN and the Genome Research Center (Daejon, Korea), for finishing and analysing regions of chr.X, were RIKEN will provide additional ready-mapped clones to the consortium.

The MPIMG will concentrate on chromosomal bands Xq27.3 (5 Mb) and Xq28 (10 Mb), one of the most gene-rich region of the entire genome. In addition, we will work on selected regions associated with syndromic and non-syndromic forms of mental retardation. So far, we have completed library

preparation for 23 clones and the shotgun sequencing has been started (status from htgs1 to finished). We committed 28 additional minimum tiling path clones at the NCBI clone registry (http://www.ncbi.nlm.nih. gov/genome/clone/). We have also identified additional 29 minimum tiling path clones by electronic mapping. At the FLI, the chimpanzee regions orthologous to human Xp11.4, Xp11.3 and Xp11.23 (chrX:37.4-49.5 Mb) will be finished. These regions contain loci associated with syndromic and non-syndromic forms of mental retardation (e.g. TSPAN7, ATP6AP2, MAO-A / B, NDP), several breakpoints associated to translocations in synovial sarcoma (SSX), the GAGE and PAGE gene clusters associated with melanoma and other cancer forms and CACNA1F, a gene in which mutations cause incomplete X-linked congenital stationary night blindness. The FLI has identified 175 clones covering 87% of the corresponding ortholog regions of the ape ("electronical mapping"), that are committed at the NCBI clone registry. Among them, 78 clones represent a minimum tiling path that will be sequenced with a coverage ("sequence depth") of about 4-5 fold. For the 34 clones building the minimum tiling path of the Xp11.4 orthologous region (4.7 Mb) library preparation has been completed and the shotgun sequencing has started. In addition, selected comparative human-chimpanzee analyses were performed in order to investigate breakpoints corresponding to pericentric inversions (Kehrer-Sawatzki et al. 2002 and 2005).

Fig 1: *The common chimpanzee (Pan troglodytes) – our closest living evolutionary relative.*

GBF will finish about 10 Mb of the chimpanzee working draft region in Xp21.1 (chrX: 27-37 Mb), orthologous to the human regions carrying the MAGE gene cluster and the Duchenne/Becker Muscular Dystrophy (DMD) locus. Further sequence comparison of the human sequence (chrX: 27-37 Mb) to the end sequences of two different chimpanzee BAC/PAC libraries (PTB, RPCI43) yielded additional 36 seed clones.

The GGSAC will annotate the chimp sequences according to standard procedures, and will use whenever possible the curated annotation available in VEGA (Sanger Center), such as the one available for the human Xp11.4 region (Wen et al in press).

Outlook

The mapping of the clones were primarily done using so – called electronic mapping using the three different chimpanzee PAC/BAC libraries allowing the construction of a first minimum tiling path. Sequencing of this path together with the available working draft will produce a first sequence assembly which will be completed by using additional clones. Gaps will be closed during the sequencing and finishing process using PCR screening (or hybridization on library filters) in cooperation with the RIKEN institute.

Our final goal is to produce a finished chimpanzee consensus sequence with an accuracy corresponding to the one achieved in the Human Genome Project (<1 error per 10.000 bp) and a minimum of gaps.

The available draft chimpanzee sequence is a very valuable resource, but is still imperfect and incomplete. Owing to the sequence similarity between man and chimp, narrowing down important evolutionary changes including pseudogene formation, gene family expansion, segmental duplications, but also subtle changes in gene and promoter sequences will require high-quality finished sequence. The hardest question to answer is “What makes us human?” Genomic comparison allows the search for functionally important differences between species, but specific biological insights will be needed to screen the large list of candidates to separate adaptive changes from neutral background. To tackle the question by many different approaches the EU applied within the 6th Framework Programme (FP6) the NEST PATHFINDER research programme “What is means to be human”. The know-how and experience about comparative sequence analysis of human and chimpanzee sequences accumulated within NGFN1 and 2 lead to the successful proposal “APES”. Together with the NGFN SMP DNA sequencing, these projects will bring together several fields of expertise and analysis levels including patterns of gene expression and promoter studies, allowing the exploitation of sequence data for understanding gene function.

Lit.: ***1.*** *Watanabe H et al. DNA sequence and comparative analysis of chimpanzee chromosome 22. Nature. 2004 May 27;429(6990):382-8.* ***2.*** *The Chimpanzee Sequencing and Analysis Consortium Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 2005 437:69-87* ***3.*** *Hattori M et al. The DNA sequence of human chromosome 21. Nature (2000) May 18;405(6784):311-9.* ***4.*** *Chelly J and Mandel JL Monogenic causes of X-linked mental retardation. Nature Reviews Genetics 2001 Sep;2(9):669-80.* ***5.*** *Ross MT et al. The DNA sequence of the human X chromosome. Nature (2005) Mar 17;434(7031):325-37.* ***6.*** *Sudbrak R et al. X chromosome-specific cDNA arrays: identification of genes that escape from X-inactivation and other applications. Hum Mol Genet (2001) Jan 1;10(1):77-83.* ***7.*** *Chen F-C and Li W-H Genomic Divergences between Humans and Other Hominoids and the Effective Population Size of the Common Ancestor of Humans and Chimpanzees. Am J Hum Genet (2001) 68:444-456* ***8.*** *Ebersberger I et al. Genomewide comparison of DNA sequences between humans and chimpanzees. Am J Hum Genet (2002) Jun;70(6):1490-7.* ***9.*** *Sakate R et al. Analysis of 5'-end sequences of chimpanzee cDNAs. Genome Res. 2003 May;13(5):1022-6.* ***10.*** *Clark AG et al. Inferring Non-neutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios. Science (2003) Dec 12;302(5652):1960-3.* ***11.*** *Fujiyama A et al. Construction and analysis of a human-chimpanzee comparative clone map (2002) Jan 4;295(5552):131-4.* ***12.*** *Kehrer-Sawatzki, H. et al. Molecular characterization of the pericentric inversion that causes differences between chimpanzee chromosome 19 and human chromosome 17. (2002)Am J Hum Genet 71,:375-388.* ***13.*** *Kehrer-Sawatzki, H., Szamalek, J. M., Tänzer, S., Platzer, M. & Hameister, H. Molecular characterization of the pericentric inversion of chimpanzee chromosome 11 homologous to human chromosome 9. (2005) Genomics 85: 542-550.* ***14.*** *Wen G. et al. Validation of mRNA/EST based gene predictions in human Xp11.4 revealed differences to the organisation of the orthologous mouse locus. Mammalian Genome in press.* ***15****. Kehrer-Sawatzki, H. et al. Breakpoint analysis of the pericentric inversion distinguishing human chromosome 4 from the homologous chromosome in the chimpanzee (Pan troglodytes). (2005) Hum Mutat 25: 45-55.*