SMP: Cell

Project: THE GERMAN cDNA CONSORTIUM: MODELING OF GENE STRUCTURES

Ingo Schupp – Deutsches Krebsforschungszentrum (DKFZ), Heidelberg – i.schupp@dkfz.de

Introduction

The sequencing of the human genome (1, 2) and those of other model organisms (3, 4) is nearly finished, and over 99% of the human genes are contained in the current genome assembly. However, the annotation of gene structures remains an open challenge towards the completion of a comprehensive catalog of all human genes (5). The analysis of the transcriptome in several large-scale projects for human (6-8) and other model organisms, e.g. the mouse (9), has proven an essential means for the identification of genes, as most gene prediction software are based on available cDNA sequence information. An international consortium makes use of these cDNA transcripts to annotate human gene structures manually in an high throughput manner (10). Yet, only about half of the expected 24,000 (protein coding) genes have been more or less comprehensively described at the structural level, and a similar fraction is available to the scientific community as "full-length" cDNAs. A strong bias exists as to the under-representation of long and rarely expressed genes and transcripts, which were not covered in the random sampling approach of the large-scale cDNA projects. The structure of many genes is therefore either still completely unknown or not correctly annotated. Further challenges arise from biological noise, i.e. the presence of incomplete or aberantly spliced transcripts in living cells (11, 12), and from experimental noise, i.e. the cDNA generation and cloning processes that generate clones being 5' or 3' truncated, containing frameshift errors, or which are internally deleted.

This project aims at the identification of thus far incompletely annotated genes, the modeling of their structures and the definition of the encoded protein coding regions, as an essential process at the very beginning of the molecular and cellular functional gene analysis pipeline of SMP-Cell. The major focus is on genes with disease relevance that has been established within SMP-Cell or in collaborative projects with other SMPs and KGs.

Project Status

A number of automated software solutions has been implemented to annotate genomes and genes. However, the quality of gene predictions is tightly associated with the availability and quality of cDNA sequence information as most gene prediction is based on cDNA sequences. We perform a manual annotation and modeling of genes and gene structures, and pass the identified structures on to the directed cloning of ORFs, and the subsequent experimental validation of structures, functions and disease relations.

Special emphasis is also on the identification and the molecular and cellular functional gene analysis of splice variants (Fig. 1), as the tissue distribution, functionality, and disease relation of many genes has already been shown to be associated with variant gene products (12-14).

We manually annotate gene structures (Fig 2) using: i. sequence information from cDNA and EST sequencing of mRNAs that are available from diverse species, ii. comparative genome analysis to identify conserved features within ORFs and putative proteins, and iii. computational gene predictions, mainly visualized by the UCSC genome browser (http://genome.ucsc.edu). We further use the HUSAR software package (http://husar.dkfz-heidelberg.de) for analysis, mainly its BLAST and ORF prediction tools. The combined data is used to create gene models, to generate virtual templates, and to finally predict functional ORFs for subsequent cloning and sequence validation. To this end we select promising cDNAs or 5'-EST clones and/or a suitable RNA sources for RT-PCR that may serve as templates for the amplification of ORFs. In addition, primers for ORF amplification and for sequence verification

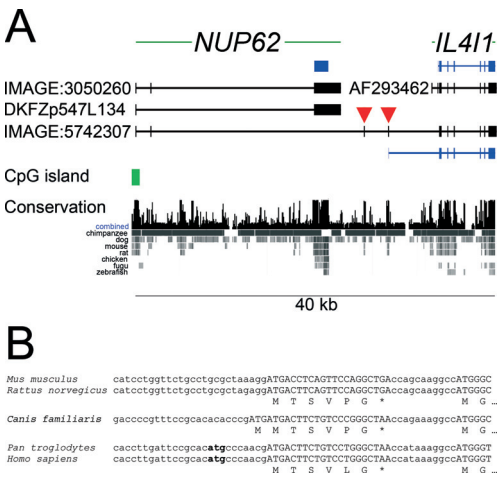

Fig 1: *(A) Transcript "fusion" of the NUP62 and IL4I1 genes caused by the pre-mRNA processing machinery (12). The splicing of representative sequence (IMAGE:5742307) is cell-type specific and induces the affected gene (IL4I1) to likely take up a new functionality in expressing cells. (B) 5' UTR of the IL4I1_2 gene transcript in the indicated organisms. All eutherian species analyzed thus far contain a short upstream ORF in the UTR. While human and chimpanzee have an additional upstream ATG start codon (in bold) that would extend the potential protein encoded, this ATG is not conserved in mouse, rat and dog. It remains to be determined whether or not that ATG is utilized in hominids and if the upstream ORF has a biological function.*

of ORFs after cloning into entry vectors are designed, using our Pride Software (http://pride.molgen.mpg.de/pride.html) that we have implemented in the Staden software (http://staden.sourceforge.net). All information computed by the different algorithms and the manual annotation generated in this project are stored in a dedicated annotation database that we have implemented. This data feeds into the cloning and sequence validation pipeline of SMP-Cell.

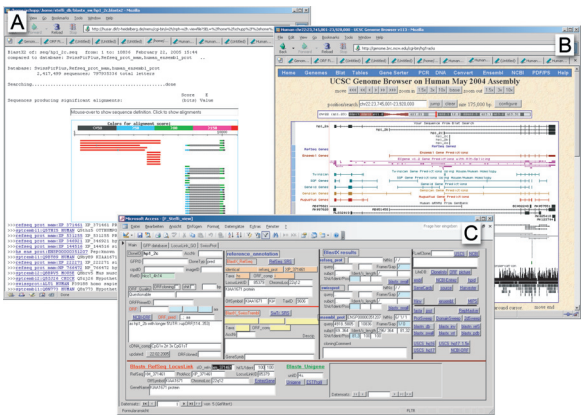

Fig 2: *Tools for the annotation and modeling of gene structures: (A) Blast output, (B) UCSC Genome Browser, (C) in house annotation database and interface.*

NGFN Nationales Genomforschungsnetz

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung

In a pilot study we generated 13 gene models, representing nine gene loci with ORFs between 3 and 12 kb in length. No clone resources were available that would support these models and that could be employed as templates for amplification. Instead the ORFs needed to be amplified by RT-PCR and then cloned into the Gateway system. Eight of these gene structures (representing six gene loci) were successfully verified by RT-PCR, for six of these the ORFs could be cloned and sequence validated. Fig. 3 shows the example of a novel gene locus, encoding a typical long (longest transcript hp1_2a > 5 kb) and apparently lowly expressed gene (only little and noisy cDNA data available). We predicted three different gene models for this gene, all of which could be verified by RT-PCR, cloning, and sequence validation. The products of the three different variants of that gene have since been entered into the cellular functional gene analysis pipeline of SMP-Cell.

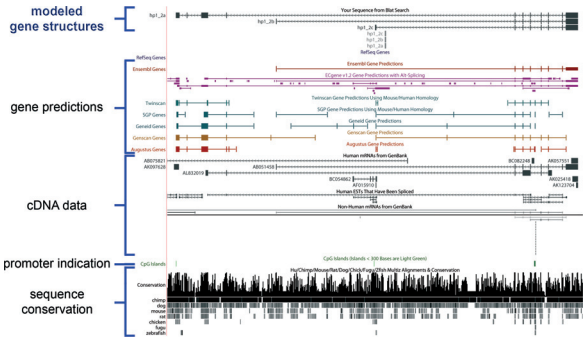

Fig 3: *Screen shot of the UCSC Genome browser displaying a novel gene <hp1_2>, which we predicted to be expressed in three variant transcripts and was consequently characterized by three gene models (hp1_2a-c). The gene models show different transcription start sites, resulting in different N-terminal ends of the encoded proteins. All three models could be verified by RT-PCR, and the cloning and sequence validation of Gateway entry clones.*

A total of 400 gene models have been generated thus far in this project. The information on their sequences, potential templates and primers for amplification and sequence validation have been parsed to the cloning and sequencing partners of the German cDNA Consortium, where the cloning and validation process is in progress. The same information is also fed into the LIFEdb database system (http://www.dkfz-heidelberg.de/LIFEdb) to permit the prioritization of targets for cellular functional gene analysis once the sequence validated ORF-clones become available. Further 900 “genes” with disease association have been identified by partners of SMP-Cell, SMP-RNA and collaborating KGs, these are currently in the process of model generation and will be entered into the validation pipeline.

Outlook

The completion of the human gene catalog and the availability of full-length ORF resources for experiments necessitates the identification of gene structures, and includes the huge number of gene and splice variants. This project provides our partners within the German cDNA Consortium, the SMP-Cell, and beyond with models of genes and ORF sequences that are likely to exist and that can be verified experimentally in molecular functional gene analysis. Once verified, the resource is available for subsequent characterization in the cellular functional gene analysis pipeline of SMP-Cell. Modeling is the first step in the processes leading from gene identification to their comprehensive functional characterization and elucidation of disease relation, and is thus key for the successful work of the SMP-Cell and its collaboration partners from other SMPs and from KGs. Further, the validated gene models of SMP-Cell significantly add to the gene catalog that shall constitute a comprehensive collection of gene and transcript sequences, and the physical clone resources that await exploitation in functional genomics and proteomics experiments.

Lit.: ***1.*** *Lander ES et al. Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium. Nature. 2001 Feb 15;409(6822):860-921.* ***2.*** *Venter JC et al. The Sequence of the Human Genome. Science. 2001 Feb 16;291(5507):1304-51.* ***3.*** *Waterston RH et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002 Dec 5;420(6915):520-62.* ***4.*** *Watanabe H et al. DNA sequence and comparative analysis of chimpanzee chromosome 22. Nature. 2004 May 27;429(6990):382-8.* ***5.*** *Ashurst JL et al. Gene annotation: prediction and testing. Annu Rev Genomics Hum Genet. 2003 4:69-88.* ***6.*** *Wiemann S et al. Toward a Catalog of Human Genes and Proteins: Sequencing and Analysis of 500 Novel Complete Protein Coding Human cDNAs. Genome Res. 2001 Mar;11(3):422-35.* ***7.*** *Strausberg RL et al. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. Proc Natl Acad Sci USA. 2002 Dec 24;99(26):16899-903.* ***8.*** *Ota T et al. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nature Genet. 2004 Jan;36(1):40-5.* ***9.*** *Okazaki Y et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature. 2002 Dec 5;420(6915):563-73.* ***10.*** *Imanishi T et al. Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones. PLoS Biol. 2004 Apr 20;2(6):856-75.* ***11.*** *Modrek B et al. A genomic view of alternative splicing. Nature Genet. 2002 Jan;30(1):13-9.* ***12.*** *Wiemann S et al. Alternative pre-mRNA processing regulates cell-type specific expression of the IL4I1 and NUP62 genes. BMC Biol. 2005 Jul 19;3(1):16.* ***13.*** *Neubrand VE et al. Gamma-BAR, a novel AP-1 interacting protein involved in post-Golgi trafficking. EMBO J. 2005 24:1122-33.* ***14.*** *Kalnina Z et al. Alterations of pre-mRNA splicing in cancer. Genes Chromosomes Cancer. 2005 Apr;42(4):342-57.*