Chimia 57 (2003) 19–21
© Schweizerische Chemische Gesellschaft
ISSN 0009–4293

Focal Point: Biotechnology

October 16, 2002

Technologies for Genome Analysis: Applications in Biomedical Research

Organized by: Prof. Ulrich Certa[a]*, F. Hoffmann-La Roche Ltd., Basel
Dr. Beat Wipf[b]*, F. Hoffmann-La Roche Ltd., Basel
Chairperson: Prof. Ulrich Certa, F. Hoffmann-La Roche Ltd., Basel

With the deciphering of whole genomes, a tremendous amount of data has been made available for research. In order to analyze all this information within a reasonable time frame and at reasonable cost, new technologies in DNA analysis as well as enzyme assays have been developed in the last few years in a symbiosis of biological, automation and miniaturization technologies. Examples are chip technologies for DNA analysis and microplate enzyme assays for high-throughput screening. Applications in biomedical research and in microbial diagnostics were presented.

Keywords: Biotechnology · DNA microarrays · Functional genomics · Protein expression · Single nucleotide polymorphisms · Transcript imaging

The Relationship between the NMYC Proto-Oncogene and the Neuroblastoma Transcriptome

Philip Day
Royal University Manchester, UK
E-Mail: philip.day@bli.unizh.ch

Neuroblastoma, the most common solid extracranial neoplasm in children, is remarkable for its clinical heterogeneity. Complex patterns of genetic abnormalities interact to determine the clinical phenotype. The molecular biology of neuroblastoma is characterized by somatically acquired genetic events that lead to gene overexpression (oncogenes), gene inactivation (tumor suppressor genes), or alterations in gene expression. Amplification of the NMYC proto-oncogene occurs in 20% to 25% of neuroblastomas and is a reliable marker of aggressive clinical behavior. In more than 50% of primary tumors the chromosome locus 17q23 is amplified and results in enhanced transcription of the oncogene.

The determination of the copy number of known genes is accomplished by quantitative PCR techniques. Modulation of gene activity is determined by quantitation of the mRNA levels in a cell, as the direct determination of protein concentrations is not yet feasible, especially for less abundant proteins. RNA concentrations are determined indirectly *via* isolation of RNA, reverse transcription and fluorolabelling of the resulting cDNA, hybridization with the probes on a DNA – chip and detection/quantitation of the label of each spot on the chip.

A potential drawback can be the translational regulation of gene activity resulting in variations of the relation of mRNA – and protein concentrations or the regulation of protein activity itself. In addition, this technique needs careful normalization of the level of a single RNA compared to total RNA.

A set of maintenance genes was found with a constant RNA level in normal and tumor cells, suitable for normalization of RNA levels of potential tumor genes. Using this technology with a commercially available medium density gene chip with 4608 sequence identities (Yale chip), 120 gene candidates for up-regulation and 140 for down-regulation were found in various tumor cells.

Further exploitation will hopefully lead to the identification of additional genes which are modulated in neuroblastoma and thus may result in a better understanding of this tumor and eventually in more specific treatment strategies for children with neuroblastoma.

SNP Discovery and SNP Screening Technologies in Pharma Research and Development

Dorothee Förnzler
Roche Center for Medical Genomics (RCMG), CH–4070 Basel
E-Mail: dorothe.foernzler@roche.com

Single nucleotide polymorphisms (SNPs) are single base exchanges in the human DNA that are stable inherited and occur with an average of 1 SNP every 1kb in our genomes. With the availability of a draft of the human genome a lot of such small ge-

Correspondence: [a]Prof. U. Certa
F. Hoffmann-La Roche Ltd.
Roche Center for Medical Genomics (RCMG)
CH–4070 Basel
Tel.: +41 61 688 53 40
Fax: +41 61 688 14 48
E-Mail: ulrich.certa@roche.com
[b]Dr. B. Wipf
F. Hoffmann-La Roche Ltd.
CH–4070 Basel
Tel.: +41 61 688 61 63
E-Mail: beat.wipf@roche com

netic variants were discovered, up to now more than 5 Mio.! The availability of a dense SNP map and the development of high-throughput technologies during the last years have opened the door for large-scale genotyping studies in individuals with well-characterized phenotypes to identify new genes involved in complex common diseases (*e.g.* type 2 diabetes, schizophrenia, Alzheimer, *etc.*) as well as the cause for inter-individual differences in drug response (pharmacogenetics).

The large number of SNPs available from public databases appear to render experimental SNP discovery redundant. However, 30–50% of public SNPs represent sequencing errors. On the other hand, the coverage of SNPs within genes is not very high – and gene-associated SNPs are often the most important if located in coding or regulatory regions, since these are the most likely ones for a medical phenotype. Out of 1081 SNPs in 78 genes discovered in our lab, only 28% could be detected in dbSNP, the most comprehensive of the public SNP databases. The necessity of automated, high-throughput SNP discovery (96 or 384-well format) is unquestioned, the gold standard is still DNA sequencing using capillary electrophoresis. For the analysis of SNPs we use Polyphred software. dsDNA sequencing in standard populations of different ethnic background does not only deliver SNPs, but also important validation data such as frequency, location and population specificity – parameters which are extremely helpful to select genetic variants for genotyping studies.

Nowadays, more than 20 different genotyping technologies are available and new ones seem to emerge very quickly. However, there is no single technology that fits all needs and applications. Choice of a technology will depend on the specific application. Furthermore, it is helpful to work with at least two different technologies for assay development and validation reasons. For our pharmacogenetic applications we rely on allele-specific PCR and primer extension followed by a mass spectrometric detection, both in 384-well format assays to ensure an appropriate throughput. Allele-specific PCR is a fast, cheap, single-tube assay, based on discrimination of 3'end primer mismatches by a specific Taq polymerase. However, match-mismatch discrimination is not good enough for all SNPs to unambiguously identify 2 different SNP alleles and the choice of primer sequences is very restricted. Primer extension with mass spectrometry detection is more robust and can be practically applied for any SNP, but needs more elaborate sample preparation. Another advantage is the determination of the correct mass of an extended oligonucleotide which avoids false positives.

The availability of affordable high-throughput SNP genotyping technologies means that the bottleneck for association studies moves to data storage, handling and automated genotyping analysis and to the accessibility of well-characterized and large enough study populations. To master the tremendous amounts of samples (often several 1000 for one project) and data (several 1000 genotypes per day) a well-functioning LIMS system is the prerequisite for successful high-throughput SNP discovery and screening projects.

We use SNP discovery and genotyping predominantly to support our clinical studies with pharmacogenetic information. We try to identify genetic factors underlying inter-individual differences in drug response and adverse effects. Further, SNP studies help us in drug target validation, in the stratification of the drug development process and in a better understanding of molecular mechanisms involved in drug response and/or disease. For this purpose we have established our own Roche Sample Repository, in which we store blood samples from participants in Roche's clinical trials. This allows us to extract DNA and genotype these samples whenever we suspect genetic factors to be involved in different responses to a certain drug.

Applications of DNA Microarrays in Functional Genomics: An Overview

Ulrich Certa
Roche Center for Medical Genomics (RCMG), CH–4070 Basel
E-Mail: ulrich.certa@roche.com

Functional Genomics is defined as the conversion of linear, genetic information into the building plans of cells and complex organisms. This parallel approach requires the application of sophisticated technologies able to handle vast amounts of information encoded by the genome of a living creature. DNA microarrays combine methods of molecular biology with semi-conductor physics to enable the transcriptional analysis of entire genomes on a single silicon microchip with several thousands of immobilized DNA probes. These are either DNA fragments of 300–500 base pairs in length ('spotted arrays') or else oligonucleotides synthesized directly on the silicon surface of the chip either chemically or by photolithography. Today, a single DNA array has the capacity to analyze the transcription of about 30000 genes, which corresponds basically to an entire eukaryotic genome. Microarrays can also be used for re-sequencing DNA or point mutation analysis. Complex computer programs were developed to analyze and display the chip results for further biological analysis.

The technology was applied in pharmaceutical research to analyze the expression patterns of interferon inducible genes in sensitive and resistant human melanoma cell lines, which led to the identification of distinct expression modes in response to interferon-α. In addition, several genes were identified that are not inducible in resistant lines and these are possibly related to the clinical failure of interferon treatment in non-responding patients. In another example it was shown that the inhibition of tumor growth by interferons in a mouse model is associated with differential expression of a number of genes which represent potential novel oncogenes (down-regulation) or else tumor suppressor genes (TSGs; induction). The technology was also applied to identify surrogate interferon response transcript markers (TMs) by selecting genes which are differentially expressed only in sensitive or resistant melanoma lines without cytokine stimulation.

Finally, oligonucleotide chips were utilized to study the molecular consequences of ischemic stroke in brain regions of rats after an experimental insult by artery occlusion. A wave of gene expression is triggered which leads ultimately to the activation of genes involved in tissue repair and neuroregeneration. This suggests the possibility to induce the protective responses by drugs targeting the signal transduction molecules. Induction of gene expression can be confirmed directly by *in situ* hybridization, which adds spatial information to the chip expression data.

A limitation and the major bottle-neck of these high-throughput approaches are the follow-up experiments using conventional analytical methods such as RT-PCR, Northern blot analysis or immunodetection of gene products. The most obvious solution is the consistent automation of experiments and data integration by sophisticated bioinformatics.

From Gene to Screen

Jeremy Beauchamp
Pharmaresearch,
F. Hoffmann-La Roche Ltd.,
CH–4070 Basel
E-Mail: jeremy.beauchamp@roche.com

High-throughput screening of chemical compound libraries for an effect on a target protein is a highly effective way of obtaining lead compounds for a pharmaceutical product. Once such a protein target has been identified, it is necessary to first develop an assay for its activity of interest and for this assay, protein will invariably be needed. Although this protein may be sufficiently abundant in a natural source, it will often be necessary to produce it artificially and currently, molecular biological manipulation of living cells gives the best means to do this.

With access to the information from the human genome project and PCR, we are able to easily produce virtually unlimited quantities of DNA corresponding to the protein of interest. Once this is done, the DNA is introduced into a specific site of an expression plasmid – a circle of DNA capable of autonomous replication in the cell. Expression plasmids have several features in common: some ‘housekeeping’ sequences to help maintain the plasmid in the cell; a promoter to switch on the production of protein in the presence of a stimulus (*e.g.* addition of a chemical) and an antibiotic resistance gene to ensure that there is an advantage to the cell to keep the plasmid. This expression plasmid or vector is the template for the production of the protein in a cell system of our choice.

Cells useful for protein expression are broadly divided into the eukaryotic and prokaryotic kingdoms. While prokaryotic expression is high in yield, cheap, easy and quick, it can give protein that is misfolded or lacking in essential post-translational modifications such as glycosylation or proteolytic cleavage. If large amounts of protein are required, this is, however, likely to be the best option. Conversely, higher eukaryotic cell expression of human proteins is often low-yield, difficult and expensive, but usually gives the expected protein. If the assay is a cellular assay – for example measuring the cellular response to a receptor activation – then higher eukaryotic expression is necessary. There are also the mid-way solutions of expression in insect or yeast cells and the possibility of cell-free expression.

Once an expression system is chosen, the expression plasmid is ‘transformed’ into the cells. Transformation can be by a number of mechanisms, of which the most useful are electroporation (using an electric shock to open holes in a cell wall and allow DNA to enter the cell), chemical methods (where DNA is concentrated at the cell surface and is taken into the cell by endocytosis-like mechanisms) and viral (in which the expression plasmid is also a viral genome designed to deliver the gene of interest to the cell cytoplasm). Once the DNA is incorporated, the cells are allowed to proliferate before the promoter in the expression plasmid is switched on. Then after a delay, the cells are harvested and either used for a cell-based assay or broken open and the protein of interest purified.

Protein purification techniques are well beyond the scope of this overview, but in the last decade molecular biology has greatly simplified this process by the ability to introduce tags and make other alterations to the protein of interest. Important examples include:

- Adding six, ten or more histidine residues in a row designed to bind to an immobilised metal affinity chromatography column.
- An *in vivo* biotinylation signal appended to the C-terminus to bind to immobilised streptavidin.
- A short peptide sequence added to the N-terminus that will be recognized by a specific monoclonal antibody and can be used for Western blotting or purification.
- Truncation of the protein to a length that limited proteolysis or structural analysis indicates will give a more stable product.
- Fusion of a protein that is targeted to a specific cellular compartment to enhance expression or folding.
- Fusion of the protein to green fluorescent protein so that its location in the cell can be identified or highly-expressing cells can be isolated.

Whether the assay used requires a highly purified enzyme or a defined cell line displaying a few receptors on its surface, the molecular biology techniques described will be a vital part of the toolkit that is used to develop the compound screen that ultimately leads to a pharmaceutical product.

Analysis of Pathogenic Bacteria by Gene Chip Technology

Joachim Frey
Institute for Veterinary Bacteriology,
University of Bern
E-Mail: joachim.frey@vbi.unibe.ch

Severe infections by pathogenic bacteria have been shifted out of our perception by the overwhelming success of antibiotic therapy. But with recent outbreaks of new or re-emerging bacteria such as *Escherichia coli O157H7*, *Vibrio cholerae* or *Legionella pneumophila* the infectious diseases have regained attention.

Pathogenicity can be a matter of the genus as well as on the species level: Whereas *Lactococcus lactis* is non-pathogenic and used as a food bacterium, *Bacillus anthracis* is well known as a highly pathogenic organism. For the ubiquitous intestinal bacterium *Escherichia coli* there are well-known laboratory strains (K-12 derivatives) devoid of any virulence factors, normal intestinal flora of healthy individuals, generally also without virulence factors, as well as highly pathogenic serotype O157H7 of which only 10–100 ingested bacteria can lead to a severe and potentially fatal infection.

Fast and accurate identification of bacteria can be of crucial importance in therapy but also in epidemiology. Classical identification methods include biochemical and serological tests, both only indirect measures of pathogenicity. The direct detection of genes of virulence factors and toxins by PCR is far too laborious for fast and routine analysis.

Immobilization of fragments (400–800 bp) of genes of *E. coli* on membranes (macroarrays) or glass (microarrays) has been developed. Suitable genes were chosen for the identification of *E. coli* and for each of the known pathogen or virulence genes of all the variants of pathogenic strains. Hybridization of DNA from isolated strains on these arrays allows the fast differentiation between laboratory strains (no pathogen gene) and the various enterotoxigenic, enterohaemolytic and uropathogenic *E. coli* strains having different sets of virulence factor or toxin genes.

This technology has already proved its value in tracking the origin of infection in clinical cases or establishing suitability of industrial production strains by showing absence of any virulence gene. Further development of DNA-chip technology will be a generalized chip for clinical applications integrating specific virulence genes and phylogenetic, taxonomic marker genes for the fast and detailed identification of most all pathogens.