

This newsletter from the Norwegian Microarray Consortium (NMC) is distributed in paper and electronic form through the Norwegian scientific community. If you are not on our mailing list and wish to receive further issues, please go to microarray.no and sign up for an electronic subscription.

NMC Services

In this issue of the NMC newsletter, we will focus on our lab and bioinformatics services and on new challenges related to high-throughput sequencing.

NMC Microarray Lab Service

by Rita Holdhus and Helle Lybæk, Norwegian Microarray Consortium, Center for Medical Genetics and Molecular Medicine, University of Bergen.

NMC offers user-friendly profiling service (RNA-in Data-out) for expression studies and SNP service (DNA- in Data- out) for genomic studies.

After an initial consultation, the user provides good quality RNA/DNA from the samples to be examined, and NMC performs the experimental procedures and low-level data analysis. The steps in this procedure include quality control of RNA/DNA, labelling, hybridization, scanning and image analysis. NMC offers a variety of different arrays, labelling procedures and hybridizations dependent on the customer's needs.

The NMC can provide profiling and SNP services on different platforms such as: Agilent Technologies, Affymetrix GeneChip, Illumina BeadArrays and Nimblegen. Details can be found at our web page: www.microarray.no. NMC has recently completed the Illumina CSpProTM certification for Gene Expression and SNP genotyping, gaining entry to an elite group of Illumina genomics service providers globally. More information can be found at page 4. Lab services are offered at all NMC nodes (Bergen, Oslo and Trondheim).

NMC offers Real-Time verification of your Microarray gene expression data. Using the Applied Biosystems Low Density Array, up to 380 genes can be rapidly examined on the same array using the well-established ABI TaqMan[®] gene expression system. Several different Gene Signature Arrays are available for human, mouse and rat, and MicroRNA is available for human. In addition one can design custom arrays with your favourite genes (note that custom arrays require a minimum order of 10 arrays).

The Bergen node of the Norwegian Microarray Consortium is pleased to announce a new service for custom-based SNP genotyping on the Illumina GoldenGate technology, which enables examination of up to 96-, 384-, 768- or 1536 SNP loci simultaneously. Illumina has re-

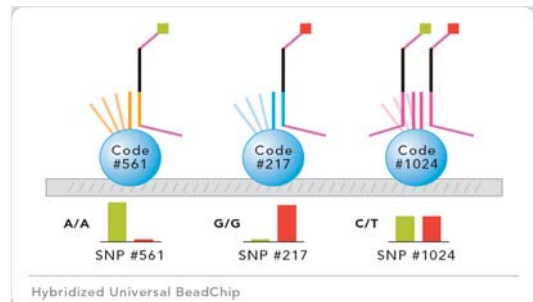


Figure 1 : Hybridisation of three different allele-specific GoldenGate SNP assay products to Illumina Universal BeadChips

cently launched an improved assay that allows multiplexing of several DNA samples. The service is therefore a cost-efficient alternative for medium scale SNP projects, e.g., to replicate findings in genome-wide association studies (GWAS) or for screening more limited numbers of candidate genes. After selection of the SNPs to be included in the study, the individual genotyping assays are designed by the customer, using the Illumina Assay Design Tool (for more details, please see information on www.illumina.com). All assays are approved by Illumina before they are produced on their custom-made chips. The users then provide NMC with genomic DNA from the relevant samples to be examined (500ng DNA per sample). We will perform all GoldenGate-related lab procedures, such as DNA activation, allele-specific oligo hybridisation, extension and ligation, product amplification, and finally hybridisation onto Illumina Universal BeadChips in a 16- or 32 sample format, with subsequent analysis of the fluorescence signals and genotype calling. NMC may also perform a primary QC of the genotype data, which can be provided as Excel spreadsheets or files compatible with the GenomeStudio analysis software from Illumina.

If requested, the NMC UiB node can advise the initiation of the custom design of the SNP genotyping assays and we may also assist in the bioinformatic analysis of the genotyping data. For more details, please contact the UiB node: Core Facility Manager Rita.Holdhus@uib.no.

Bioinformatics Service

by Endre Anderssen, Norwegian Microarray Consortium, Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology

The greatest effort in a microarray study is often not the hybridisations themselves but the process of making sense of the flood of data that the study generates. Often, the data analysis process is presented as a stepwise procedure going from raw data through quality control and normalisation and on to hypothesis testing, resulting in a list of significantly differentially expressed genes. The biologist can then take this list and perhaps do experiments to verify the findings and publish. Although this approach sometimes works, hypothesis testing alone will often just produce a confusing collection of gene lists that contain too many genes to manage. Also, perhaps the most important contribution from a bioinformaticist is to the experimental design.

At NMC we try to have a meeting with every customer before any experiments are done, at this stage we discuss the experimental design including the number of replicates needed and what analysis strategy will best suit the biological question. Microarray experiments typically fall into one of the following types:

- **Group comparisons:** Find differentially expressed genes between different treatments or clearly defined groups of individuals.
- **Time series:** Find genes that respond to a treatment. Typically used to get information about gene regulatory systems.
- **Phenotype relationships:** Find genes that relate to a phenotype such as survival time or metabolite levels.

Once an experimental strategy has been chosen analysis of already published datasets studying similar systems that can give important information about the level of biological noise, the magnitude of biological differences and which biological processes are involved. Published data can also be helpful when evaluating which time points to look at or if there are any special quality issues with certain tissues etc. All this information will be used to suggest an experimental design including how to randomise or block experimental work to avoid technical bias in the results, and suggesting a number of replicates that will give a reasonable compromise between reliability and cost for the analysis strategy agreed upon. After the experimental work has been completed the data is analysed by explorative methods and methods for hypothesis testing. The explorative analysis focuses on forming an overall understanding of the whole dataset.

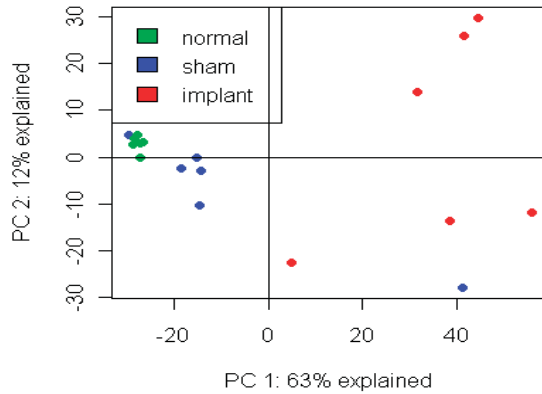


Figure 2. Score plot from a principal component analysis shows an approximation of the relationships between samples. Different colours are used to represent the different groups in the experiment and samples that do not fit in can be detected.

Methods such as principal component analysis (PCA) or clustering can be used to model and visualise the whole dataset. The aim in this phase is to get a birds-eye view of the dataset and understand how different and how homogenous the various groups of samples are, and which major biological processes or pathways are causing the differences. This process lets the researcher know if the experiment has gone as expected biologically, and any unexpected findings can be evaluated. Sometimes they might be unexpected side effects of the way experiments are carried out, at other times they may be interesting starting points for finding new knowledge. After the overview of the data has been obtained a more focused analysis can be done. Most customers ask for a list of differentially expressed genes, but a number of methods are also available that can find important functional gene categories, including GO terms and pathways and detect important regulators.

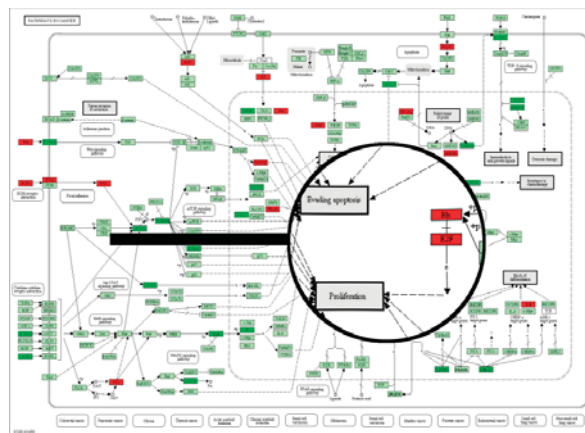


Figure 3. Pathway analysis. By highlighting differentially expressed genes in known pathways changes that affect regulatory or metabolic paths can be detected.

Massively Parallel Sequencing of Arrayed DNA Molecules

by Leonardo A. Meza-Zepeda and Ola Myklebost, Norwegian Microarray Consortium, Oslo University Hospital Rikshospitalet, and Institute for Molecular Biosciences, University of Oslo.

During the last years, high-throughput sequencing has been developed into a massively parallel format, allowing the new generation of sequencing machines to determine billions of basepairs of sequence, or a complete genome, in a few days. Although the technologies produce shorter reads than traditional sequencing, a huge increase in output per experiment is obtained by processing millions of separate sequencing reactions in a microarray format. This technology is revolutionising research with its single base resolution and quantitative reads, and it is currently being implemented in an increasing number of applications.

Massively parallel sequencing generates Gbp of sequence that can be used for de novo or resequencing of whole or large segments of the genome. The single base resolution and the large number of sequence reads facilitates also the identification and discovery of genetic variation, mutations and genomic rearrangements (i.e. DNA copy number changes, SNPs, indels and balanced translocations). Recently, novel sequence capture protocols have been developed to enrich for specific genes, exons, chromosome segments or all the coding segments of the genome. These new sequencing technologies are not only changing genome sequencing approaches but are also rapidly being adopted to study other aspects of genome dynamics. Within the field of transcriptomics, massively parallel sequencing can be used to accurately quantify gene expression and for transcript discovery. The single base sequence resolution of these technologies allows the detection of mutations and SNPs within expressed genes (and thus in many cases allele-specific expression), as well as the identification and characterisation of fusion genes and alternative splice variants. Sequencing provides an excellent linear dynamic range for quantitative analysis, as well as unprecedented sensitivity for detection and measurement of low-abundant transcripts. Another area where sequencing is having a dramatic impact is epigenomics. This technology has allowed for the first time to sequence the methylome at a single base resolution, and is widely being used in combination with chromatin immunoprecipitation to map chromatin state genome wide (ChIP-Seq).

The huge amounts of data that are generated by massively parallel sequencing experiments impose major challenges related to storage, assem-

bly and analysis. It is therefore critical to establish procedures of efficient data storage and pipelines, as well as server-based analysis software.

Although massively parallel sequencing gives higher precision and better sensitivity than traditional hybridisation-based microarray techniques, it is still quite expensive, and will for a long time be **complementary to microarrays**. Microarrays are a very cost effective tool for studying large numbers of samples that would be economically impossible to pursue by sequencing. In addition, the new information generated by the sequencing projects will translate into new improved microarray content.

At present, three robust technologies share most of the market of "second-generation" sequencers; the Roche GS FLX platform (formerly 454 Life Sciences), the Illumina Genome Analyzer (formerly Solexa), and the SOLiD technology from Applied Biosystems. In addition, new emerging technologies are being developed but have not reached the market as robust technologies (i.e. Helicos and Pacific Biosciences using single molecule sequencing, or OxfordNanopore Technologies' using a nanopore-based system).

Sequencing starts by generating genomic libraries of adapter ligated DNA fragments from a variety of sources (i.e. fragments of genomic DNA, cDNA, chromatin immunoprecipitates). Individual DNA molecules are then amplified by emulsion PCR (Roche and Applied Biosystems) or bridge amplification (Illumina) before being sequenced. The **Roche GS FLX** sequencer was the first commercially available sequencer and works in the principle of pyrosequencing. This technology produces longer sequence reads, 250-500 bp, and is capable of sequencing from about 0.5 million to over a million reads per run, thereby generating 100-500 Mbp of sequence. The **Illumina Genome Analyzer** is based on the sequencing-by-synthesis principle. Today's instrument produces 5-18 Gbp of sequence in a single run, by reading up to 2x75 bp in 96-120 million DNA clusters. The **Applied Biosystems SOLiD** sequencer is one of the newest in the market. The technology is based on sequencing by oligonucleotide ligation and detection, and it is able to generate up to 20 Gbp of sequence by analysing 400 million tags per run.

For information about sequencing at the NMC, see page 4.

Sequencing at the NMC

Since the year 2000, the Norwegian Microarray Consortium has provided state-of-the-art technology and delivered high-throughput genomic services to the Norwegian scientific community. According to our FUGE-2 road map, we are planning to introduce massively parallel sequencing as a complementary technology to the services provided today. The NMC has applied for funding for a gradually increasing but extensive set of equipment to support users of our current applications, and the Oslo facility is collaborating with the Cancer Biomedicine Centre of Excellence on establishing the Illumina technology at Radiumhospitalet early this fall. The endeavour into this complex technology is aided by our experienced international advisors, and the data storage and analysis is planned in collaboration with the FUGE Bioinformatics platform.

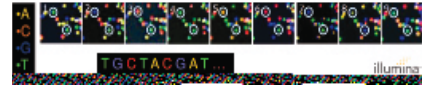
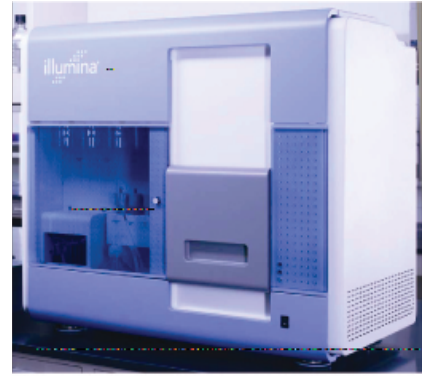


Figure 4. Shows the Illumina Genome Analyzer

NMC Achieves Illumina CPro Status

With this NMC announces that we have completed Illumina CPro™ certification for Gene Expression and SNP genotyping (Infinium), gaining entry to an elite group of Illumina genomics service providers globally. Illumina Inc., a San Diego-based company, provides leading-edge genetic analysis tools to genomics centers worldwide.

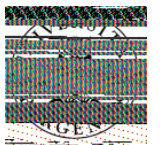
Illumina CPro is the collaborative service provider partnership dedicated to ensure the delivery of the highest-quality data available for genetic analysis applications. Illumina CPro participants undergo a rigorous two-phase certification process that includes minimum data generation, data certification, and an on-site audit of the facility and processes. Illumina CPro certification provides a competitive advantage for service providers and ensures that customers who use Illumina CPro services receive the industry-leading data quality and service they have come to expect from Illumina.

Quote from the provider: "Illumina CPro recognizes organizations that provide customers with industry-leading data quality and service in genetic analysis", said Karen Possemato, Illumina's Director of Corporate Marketing. "NMC is now a certified service provider in Norway able to deliver gene expression, SNP and data analysis services using Illumina technology. Now NMC is a global CPro partner, we are excited to work with them to make it easier for researchers in Norway and the Nordic countries to access the power of Illumina's genetic analysis technologies".



www.microarray.no

Microarray.no



NTNU



Microarray Services

Expression profiling service

- Illumina
- Affymetrix
- Agilent
- Nimblegen
- Data analysis

Custom printing service

Genomic profiling or SNP service

- Illumina
- Affymetrix
- Agilent
- Nimblegen
- Data analysis



Norwegian Microarray Consortium

Bergen +47-55975326
 Oslo +47-22781769
 Trondheim +47-72576514
 NMC News Editor (Vidar Beisvåg)

www.microarray.no

bergen@microarray.no
 oslo@microarray.no
 trondheim@microarray.no
 vidarbe@microarray.no