



**Karolinska  
Institutet**

Department of Medical Epidemiology and Biostatistics

## Methods for deep examination of DNA

AKADEMISK AVHANDLING

som för avläggande av teknologie doktorsexamen vid Karolinska Institutet offentligens försvaras i hörsal Atrium, Nobels väg 12B

**Torsdagen den 30 maj, 2013, kl 09:00**

av

**Mårten Neiman**

*Tekn.Lic.*

*Huvudhandledare*

Professor Henrik Grönberg  
Karolinska Institutet  
Institutionen för medicinsk epidemiologi och biostatistik

*Bihandledare*

Dr. Johan Lindberg  
Karolinska Institutet  
Institutionen för medicinsk epidemiologi och biostatistik

Dr. Daniel Klevebring  
Karolinska Institutet  
Institutionen för medicinsk epidemiologi och biostatistik

*Fakultetsopponent*

Professor Olli Kallioniemi  
University of Helsinki  
Institute for Molecular Medicine Finland

*Betygsnämnd*

Docent Ann Nordgren  
Karolinska Institutet  
Institutionen för molekylär medicin och kirurgi

Professor Per-Åke Nygren  
Kungliga Tekniska Högskolan  
Molekylär bioteknologi

Docent Lukas Käll  
Kungliga Tekniska Högskolan  
Genteknologi

**Stockholm 2013**

Mårten Neiman (2013): **Methods for deep examination of DNA**  
ISBN 978-91-7549-173-8

## Abstract

The development of sequencing technology has had a rapid pace during the last years and today, the sequencing instruments harbors enormous capacity. This thesis is about the development of methods to make the most out of this capacity and to use it for various applications.

In **paper I**, a dual tagging system for sequencing large sample sets was developed. To proof the concept, 4,700 dogs were subjected to amplicon sequencing of the 2nd exon of the gene DLA-DRB1 using a 454 genome sequencer. By using a combination of two tags, 4,992 samples can be analyzed within the same run using only 148 tagging sequences. In our experiment, 95% of the generated PCR-products achieved a sufficient read depth ( $\geq 20\times$ ) for variant calling.

**Paper II** solved a problem coupled to amplicon sequencing experiments, namely contamination of sequence data from unwanted by-product formation. By hybridizing an oligo nucleotide coupled to a fluorescent dye to the wanted PCR products, fluorescence activated cell sorting (FACS) could be used to enrich the target molecules after emulsion PCR. The resulted in a nearly three-fold increase of quality reads from a sorted library compared to a non-sorted.

Since the cost of sequence data has decreased during the last years, the budget item of library preparation have become more abundant in sequencing experiments. In **paper III**, library preparation using cheap bulk enzymes was explored regarding various factors affecting process efficiency. Multiplex sequence capture was also evaluated for processing up to eight samples at the cost of one single reaction.

Using circulating tumor DNA for measuring systemic tumor load have been proposed and investigated by several studies. In **paper IV**, the utility of exome sequencing for this application was assessed. The findings suggests that the ability of detecting low-frequent alleles is insufficient for clinical use where detection levels of  $< 0.5\%$  is required.

**Keywords:** massive parallel sequencing, cancer genomes, prostate cancer, biomarkers, library preparation, amplicon sequencing, DNA barcodes, circulating DNA, enrichment, targeted sequencing, exome sequencing.