DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS

201

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS
201

KRISTJAN VÄLK

Gene expression profiling and genome-wide association studies of non-small cell lung cancer

TARTU UNIVERSITY PRESS

Institute of Molecular and Cell Biology, University of Tartu, Estonia

Dissertation is accepted for the commencement of the degree of Doctor of Philosophy (in molecular diagnostics) on 19.07.2011 by the Council of the Institute of Molecular and Cell Biology, University of Tartu.

Supervisor: Prof. Andres Metspalu, MD, PhD
Department of Biotechnology, Institute of Molecular and Cell Biology, and Estonian Genome Center, University of Tartu, Estonia

Opponent: Dr. Jörg Hoheisel, PhD.
Division of Functonal Genome Analysis, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

Commencement: Room No 217, 23 Riia Str., Tartu, on August 25th 2011, at 12.00

The publication of this dissertation is granted by the University of Tartu.

European Union
European Social Fund

Investing in your future

CAPACITIES

FP7-REGPOT-2009-1
Grant no. 245536

ARCHIMEDES

DoRa

The publication of the current thesis was supported by Targeted Financing from the Estonian Ministry of Education and Research (SF0180142s08), the Estonian Science Foundation grant ETF7076 and the Graduate School of Biomedicine and Biotechnology, the European Science Foundation Frontiers of Functional Genomics Exchange grant 2182 and DoRa program activity 6, the EU FP7 grant ECOGENE (#205419, EBC), the EU FP7 grant OPENGENE (#245536, EGC/University of Tartu), and by the EU via the European Regional Development Fund grant to the Centre of Excellence in Genomics, the Estonian Biocentre, and the University of Tartu.

ISSN 1024–6479
ISBN 978–9949–19–832–0 (trükis)
ISBN 978–9949–19–833–7 (PDF)

Tartu Ülikooli Kirjastus
www.tyk.ee
Tellimus nr. 486

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

The current dissertation is based on the following publications:

I. Landi, M. T., Chatterjee, N., Yu, K., Goldin, L. R., Goldstein, A. M., Rotunno, M., Mirabello, L., Jacobs, K., Wheeler, W., Yeager, M., Bergen, A. W., Li, Q., Consonni, D., Pesatori, A. C., Wacholder, S., Thun, M., Diver, R., Oken, M., Virtamo, J., Albanes, D., Wang, Z., Burdette, L., Doheny, K. F., Pugh, E. W., Laurie, C., Brennan, P., Hung, R., Gaborieau, V., McKay, J. D., Lathrop, M., McLaughlin, J., Wang, Y., Tsao, M. S., Spitz, M. R., Krokan, H., Vatten, L., Skorpen, F., Arnesen, E., Benhamou, S., Bouchard, C., Metspalu, A., Vooder, T., Nelis, M., **Välk, K.**, Field, J. K., Chen, C., Goodman, G., Sulem, P., Thorleifsson, G., Rafnar, T., Eisen, T., Sauter, W., Rosenberger, A., Bickeboller, H., Risch, A., Chang-Claude, J., Wichmann, H. E., Stefansson, K., Houlston, R., Amos, C. I., Fraumeni, J. F., Jr., Savage, S. A., Bertazzi, P. A., Tucker, M. A., Chanock, S., and Caporaso, N. E. (2009). **A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma.** Am J Hum Genet 85(5), 679–91.

II. **Välk, K.***, Vooder, T.*, Kolde, R., Reintam, M.A., Petzold, C., Vilo, J., and Metspalu, A. (2010). **Gene Expression Profiles of Non-Small Cell Lung Cancer: Survival Prediction and New Biomarkers.** Oncology 2010;79:283–292.

III. Vooder, T.*, **Välk, K.***, Kolde, R., Roosipuu, R., Vilo, J., and Metspalu, A. (2010). **Gene Expression-Based Approaches in Differentiation of Metastases and Second Primary Tumour.** Case Rep Oncol 3(2), 255–261.

* These authors contributed equally to this work.

The articles are reprinted with the permission of the copyright owners.
My contributions to the articles are as follows:

I. Sample preparation, single nucleotide polymorphism genotyping analysis, and writing of the manuscript.

II. Design of the study, conducted laboratory experiments, and primary quality control of the array data; participated in the analysis of data, interpretation of the results, and preparation of the manuscript. Shared first authorship.

III. Designed and conducted most of the laboratory experiments and primary quality control of the data. Participated in the analysis of the data, interpretation of the results, and preparation of the manuscript. Shared first authorship.

LIST OF ABBREVIATIONS

AC	Adenocarcinoma
ASC	Adenosquamous Carcinoma
BAC	Bronchioalveolar Carcinoma
BP	Biological Process
DNA	Deoxyribonucleic Acid
GO	Gene Ontology
GWAS	Genome-Wide Association Study
HPV	Human Papilloma Virus
LCC	Large-Cell Carcinoma
LD	Linkage Disequilibrium
NSCLC	Non-Small Cell Lung Cancer
OR	Odds Ratio
PCA	Principal Component Analysis
RT-PCR	Reverse Transcriptase Polymerase Chain Reaction
SCC	Squamous Cell Cancer
SCLC	Small Cell Lung Cancer
SNP	Single Nucleotide Polymorphism
tagSNP	Tagging Single Nucleotide Polymorphism
TNM	Tumour, Node, Metastasis staging system
WG	Whole Genome

INTRODUCTION

Cancer is the second leading cause of death after cardiovascular disease, and lung cancer is one of the most commonly diagnosed cancers worldwide. Currently there are more than 1.61 million new cases of lung cancer diagnosed each year (Ferlay et al., 2010). The most prevalent cause of death due to cancer (18.2%) is also caused by lung cancer, that accounts for 1.38 million deaths worldwide each year (Jemal et al., 2011). In Europe alone, it is predicted that lung cancer will be responsible for the deaths of 182 000 men and 76 000 women in 2011 (Malvezzi et al., 2011). Moreover, lung cancer is a widespread disease in both developed and developing countries, thereby constituting a major public health problem and economic burden.

The incidence of lung cancer, globally, started to rise during the 1950s, and was consistent with the increase in cigarette consumption that started at the beginning of the century (Proctor, 2004). According to the World Health Organization (WHO), the early export of tobacco from America to the rest of the world by Christopher Columbus now results in more than 5 million annual premature deaths, and is the main cause of lung cancer (Ezzati and Lopez, 2003). Moreover, approximately 90% of lung cancer patients have a history of smoking at least one pack a day for 20 years, although not all smokers develop lung cancer. For example, there are lung cancer patients who have never been exposed to tobacco (Hecht, 1999), and this indicates that other factors are involved in the pathogenesis of this disease. Correspondingly, previous studies have identified that exposure to asbestos, heavy metals, silica, diesel exhaust, painting work, or cooking fumes, can represent significant risk factors in the development of lung cancer (Mollberg et al., 2011; Tse et al., 2011). In recent years, human papillomaviruses (HPVs), 16 and 18, which are mostly known as risk factors for cervical cancer, have also been associated with lung cancer, especially in Asian populations (Cheng et al., 2001; Klein, Amin Kotb, and Petersen, 2009; Koshiol et al., 2011). There have also been, several familial aggregation, candidate gene, and genome-wide association studies (GWAS) that have identified several heritable components of lung cancer (Amos et al., 2008; Bailey-Wilson et al., 2004; Hung et al., 2008; Lorenzo Bermejo and Hemminki, 2005; Schwartz et al., 2007).

Currently, the diagnosis and prognosis of lung cancer relies on anatomical Tumour size, Node involvement, and presence of Metastasis (TNM) staging system (Rami-Porta, Chansky, and Goldstraw, 2009; Tanoue and Detterbeck, 2009), as well as histological classification (Brambilla et al., 2001). Histologically, lung cancer can be divided into small-cell lung cancer (SCLC), which is not radically treatable, and non-small cell lung cancer (NSCLC). The latter is further categorized into squamous cell cancer (SCC), large-cell carcinoma (LCC), and adenocarcinoma (AC) (which also includes a bronchioalveolar carcinoma (BAC) subtype) (details in Chapter 1.5.2) (Lacroix, Commo, and Soria, 2008). Currently, the most prevalent histological forms of lung cancer detected are SCC (30%) and AC (30%), while SCLC, LCC, and BAC equally account for the remaining 40% (Garber et al., 2001). Despite improvements in lung cancer histology and staging, the prognosis and response to therapy by patients within the same subgroup has been observed to vary substantially (Cox et al., 2001; Hou et al., 2010). In addition, a considerable number of histological samples exhibit mixed features, which makes it difficult to obtain a clear classification (Bhattacharjee et al., 2001).

In recent years, several candidate gene approaches and hypothesis-free GWAS have successfully identified genetic markers associated with an increased risk for lung cancer. In addition, expression profiling has elucidated several new potential biomarkers from blood, saliva, and biopsy samples. However, despite the development of novel targeted treatments, the survival and recurrence-free periods for patients with lung cancer have not improved substantially, and standard treatment options have remained largely unchanged.

In the current thesis, the subsequent literature review outlines the principles, challenges, and current state of GWAS and whole genome (WG) gene expression profiling of NSCLC. The experimental part of the current study was approved by the Ethics Review Committee on Human Research of the University of Tartu, and is based on samples collected from lung cancer patients between November 2002 and December 2006 from the Estonian population. These samples were subjected to GWAS and WG gene expression profiling, and these results are presented and discussed.

I. REVIEW OF LITERATURE

1.1 Principles of genome-wide association studies (GWAS)

In 2001, the Human Genome Project (HGP) reported an initial draft containing more than 1.4 million single nucleotide polymorphisms (SNPs) (Lander et al., 2001). Subsequently, phase I and phase II Hap Map Projects validated these SNPs with linkage disequilibrium (LD) studies which were completed in 2005 and 2007, respectively. Today, there are more than 4 million validated SNPs available for GWAS, however, only 0.5–1 million tagging SNPs (tagSNPs) are necessary to describe the genome (Li, Li, and Guan, 2008). Moreover, recent studies have established that missing genotypes can be covered by genome wide imputation (Anderson et al., 2008; Zhao et al., 2008), a method which predicts non-identified SNPs based on LD. As a result, several cancer-associated LD blocks have been identified (Varghese and Easton, 2010), of which locus 8q24 is one of the most cited (Figure 1).

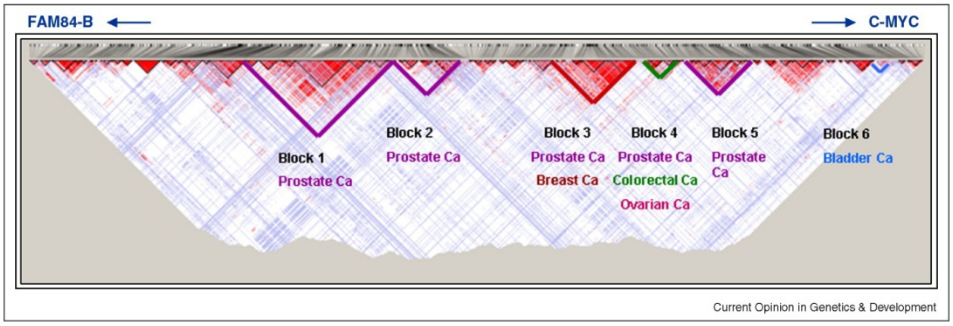


Figure 1. LD plot of the 8q24 region adapted from Varghese and Easton (2010). Six blocks are represented which refer to regions associated with increased cancer risk. For example, SNP rs6983267 in Block 4 is associated with prostate cancer, colorectal cancer, and ovarian cancer. Separate susceptibility variants for breast and prostate cancer are associated with Block 3, and SNP rs9642880 in Block 6 is associated with an increased risk for bladder cancer.

In 2005, the first GWAS that used commercially available genotyping microarrays included 100 000 SNPs (Klein et al., 2005). This study of 96 age-related macular degeneration (AMD) cases and 50 controls identified the association of complement factor H (CFH) with age-related macular degeneration. Moreover, the effect size of the risk allele was found to be 4.6 for the heterozygous state, and 7.4 for the homozygous state. Today, more than 500 GWAS, and associations of more than 2000 SNPs, have been published, and these numbers are growing rapidly. However, with odds ratios (OR) of < 1.5 although the study and control groups are much larger than in initial studies (Hindorff et al., 2009). It is also important to note that the vast majority of GWAS have identified chromosomal regions implicated with various traits and more extensive studies are needed to locate and describe the actual genetic markers responsible for a particular phenotype. For issues involving missing heritability, several strategies are available, which include next-generation whole genome sequencing that can capture rare and non-coding variants, analyses of structural variations and re-arrangements, as well as investigations of gene-gene and gene-environment interactions (Manolio et al., 2009).

Typically, GWAS use a case-control design in which cases are determined by the trait of interest. Ideally, the controls should be as similar to the cases being analysed as possible, except for the trait being studied. Recently, the use of reference cohort subjects has become popular (2007), and data for controls used in other studies can be obtained from the genotype-phenotype association database (http://www.ncbi.nlm.nih.gov/gap), the Illumina iControl database, and/or the Genetic Association Information Network and Wellcome Trust Case Control Consortium (http://www.wtccc.org.uk/). However, when using reference data obtained from different populations, the allele frequency may not vary according to the trait, but rather due to differences in ancestry. This population stratification can be controlled for by programs like PLINK (Purcell et al., 2007) and EIGENSOFT (Price et al., 2006). Another important consideration is that the majority of SNPs included in currently available arrays are selected with the goal of covering the genome equally, which is why only a small proportion of them affect protein structure or gene expression. As a result, chromosomal regions, rather than actual genetic variants responsible for a trait, are identified. Therefore, most findings from GWAS require further study, such as fine mapping or sequencing, in order to elucidate the true functional variant of interest (Frazer et al., 2009).

The goal for data analyses of GWAS is to test every individual SNP in order to select those that exhibit a statistically significant association with a particular trait. As a result, hundreds of thousands of tests are generated, and a correction for multiple testing is applied to reduce the number of false positives obtained. For the Bonferroni correction, the p-value is divided by the number of tests conducted (Cardon and Bell, 2001). A less conservative approach is to use a false discovery rate, which provides the proportion of false positive associations present among the detected significant associations (Benjamini et al., 2001; Tusher, Tibshirani, and Chu, 2001).As a result of the potential confounding results due to multiple testing, population stratification, or technical errors in WGAS, positive findings need to be replicated in independent studies. Alternatively, meta-analysis can be performed, with the data from different studies combined to identify novel significant findings.

1.2 GWAS of NSCLC

GWAS have led to the identification of more than 100 common, low-penetrance loci for cancer. At these loci, common genetic variants have been associated with a moderate increased risk for cancer, typically < 1.3-fold (Chung et al., 2010; Varghese and Easton, 2010). However, the majority of these findings have not been replicated, nor translated to currently known pathways of carcinogenesis (Galvan, Ioannidis, and Dragani, 2010). In a series of familial aggregation studies of lung cancer, adjustments for family smoking patterns were conducted (Matakidou, Eisen, and Houlston, 2005; Mayne, Buenconsejo, and Janerich, 1999; Tokuhata and Lilienfeld, 1963), and three types of variants identified. Most of the findings are speculated to be associated with lung cancer in three different ways (Rafnar et al., 2011):

- variants that affect the risk of lung cancer regardless of smoking status;
- variants that increase the vulnerability of smokers to the harmful effects of smoking;
- variants that affect smoking behaviour.

However, most of these studies were conducted in European-derived populations, which limits the worldwide applicability of these findings, as well as the range of marker alleles possibly associated with cancer.

Most inherited cancer syndromes are associated with rare and highly penetrant monogenic mutations. For example, among the sporadic cancers, thyroid cancer is associated with a 53% heritable genetic component, while the heritable component of lung cancer is estimated to be 8% (Czene, Lichtenstein, and Hemminki, 2002). For lung cancer, several GWAS have been conducted, most of which have used a meta-analysis approach. In these studies, more than a thousand cases and controls were analysed, and a genome-wide significance level was achieved ($p < 5 \times 10^{-7}$) (2007). The results of these studies that have been replicated are listed in Table 1.

Table 1. Results of WGAS performed to evaluate lung cancer risk.

LC Gene	Locus	SNP	OR	Reference
CLPTM1L **TERT**	5p15.33	rs401681	1.3	(Broderick et al., 2009; Landi et al., 2009; McKay et al., 2008; Wang et al., 2008; Yoon et al., 2010)
BAT3 **MSH5**	6p21.33	rs3117582	1.24	(Broderick et al., 2009; Wang et al., 2008)
CHRNA3 **CHRNA5**	15q24-25.1	rs8034191	1.3	(Amos et al., 2008; Broderick et al., 2009; Hung et al., 2008; Thorgeirsson et al., 2008)
TP53BP1	15q15.2	rs748404	1.15	(Broderick et al., 2009; Rafnar et al., 2011)
C3orf21	3q29	rs2131877	1.33	(Yoon et al., 2010)

1.2.1 Locus 6p21 and lung cancer

In 2008 and 2009, Wang et al. and Broderick et al., respectively, published GWAS that identified a lung cancer risk locus at 6p21.33. The initial study had consisted of 1952 cases and 1438 controls that were pooled with data from two other GWAS (which included 5095 cases and 5200 controls). In addition, replication experiments with 2484 cases and 3036 controls were conducted, and rs3117582 was identified as a lung cancer-associated risk locus ($p = 4.97 \times 10^{-10}$; OR = 1.2 for the AC genotype and OR = 1.8 for the CC genotype). This locus is located in the first intron of *BAT3* that exhibits a strong LD with *MSH5*. Moreover, BAT3 is associated with the p53-mediated response to DNA damage (Sasaki et al., 2007), while MSH5 is associated with DNA mismatch repair (Wang et al., 2006). However, in 2009, when the association between smoking behaviour and the 6p locus was investigated, no correlation was found.

1.2.2 Locus 15q15 and lung cancer

15q15.2 is another lung cancer risk locus that has recently been identified in two independent multistep studies (Broderick et al., 2009; Rafnar et al., 2011). In the first meta-analysis study, conducted by Peter Broderick and co-workers the SNP, rs748404 ($p = 1.08 \times 10^{-6}$, OR = 1.15), was identified, which is located between two transglutaminase genes, *TGM5* and *TGM7*. In the second study, by Rafnar and co-workers, the best association involving the 15q locus was mapped to the same SNP. This association reached a genome-wide significance level ($p = 1.1 \times 10^{-9}$), and the OR for the T allele was 1.15. However, no association between rs748404 and patient gender, age at diagnosis, or smoking quantity was detected. Based on these results, the authors hypothesized that multiple lung cancer risk loci are present in the region.

1.2.3 Locus 15q25 and lung cancer

Perhaps the most intriguing lung cancer risk locus found in GWAS has been 15q25, which includes three nicotine acetylcholine receptor subunit genes (*CHRNA3*, *CHRNA4*, and *CHRNA5*) (Amos et al., 2008; Hung et al., 2008; Thorgeirsson et al., 2008). In three studies, the OR for 15q25 was found to be approximately 1.3. Moreover, the association of smoking habit with the 15q25 locus has been confirmed in multiple studies (Saccone et al., 2007; Thorgeirsson et al., 2008). For example, Thorgeirsson and co-workers demonstrated a significant association ($p = 5 \times 10^{-16}$) between the T allele of SNP rs1051730 and the number of cigarettes smoked per day and a nicotine dependence scale. However, recent WGAS of lung cancer by Amos et al. and Hung et al. reached the opposite conclusion – namely, that the association of the 15q25 locus is with lung cancer, and not with smoking status. Due to inconsistencies in the data regarding this locus, larger samples of non-smoking lung cancer patients are needed.

1.2.4 Locus 5p15 and lung cancer

The most cited lung cancer locus mapped by WGAS is 5p15.33 (Broderick et al., 2009; Landi et al., 2009; McKay et al., 2008; Wang et al., 2008; Yoon et al., 2010). In this locus SNP rs2736100 is located within the coding region of the cleft lip and palate transmembrane protein 1-like protein (CLPTM1L), also known as cisplatin resistance-related protein 9, while SNP rs402710 is located within the telomerase reverse transcriptase (TERT) gene. Both of these genes have putative roles in cancer predisposition.

One of the first GWAS that reported an association between 5p15.33 and lung cancer was conducted by James D. McKay and co-workers (2008). The SNP contained in this locus, rs402710 ($p = 4 \times 10^{-6}$), was tested for possible associations with histology, patient gender, smoking exposure, and age of onset of lung cancer as co-factors. However, no associations were detected for any of these factors. In the same volume of Nature, another study that also reported an association between 5p15.33 and lung cancer risk was published (Wang et al., 2008). In this study, SNP rs401681 reached a statistically significant level ($p = 7.9 \times 10^{-9}$), and ORs for the GA and AA genotypes were 0.86 and 0.77, respectively. In a Korean population, GWAS also found 5p15 to be a susceptibility locus for lung cancer (Yoon et al., 2010), however, genome-wide significance was not reached. In addition to these findings, Broderick and co-authors (2009) have estimated the individual risk of other major lung cancer-associated loci. According to these estimates, both the 5p15.33 and 6p21.33 loci account for 1% of excess familial risk, while the 15q25.1 locus accounts for 5%.

In conclusion, lung cancer GWAS have mapped several important genomic loci. However, it appears that the limits of GWAS have been reached using the technology currently available. Therefore, deep sequencing of entire genomes, large meta-analyses, and the study of patients of non-European descent, need to be included in future studies in order to elucidate novel and rarer variants.

1.3 Genome-wide gene expression profiling

1.3.1 Gene expression analysis methods

Several methods have been developed for the study of gene expression, and these can be divided according to their throughput. For example, northern blotting (Alwine, Kemp, and Stark, 1977) and reverse transcriptase polymerase chain reaction (RT-PCR) assays are considered low throughput methods, based on the relatively small number of samples and/or probes that can be analysed simultaneously. However, RT-PCR does have an advantage over methods associated with a higher throughput due to its precision and cost of the assay when only a few genes are analysed, and when absolute quantification is needed (Livak and Schmittgen, 2001). High-throughput gene expression methods include serial analysis of gene expression (Horan, 2009), cap analysis of gene expression (Kodzius et al., 2006), massively parallel signature sequencing

(Reinartz et al., 2002), microarray-based technologies (DeRisi et al., 1996; Schena et al., 1995), ribonucleic acid sequencing (RNA-Seq) (Ozsolak and Milos, 2011), and whole exome and transcriptome sequencing. Moreover, the application of a whole genome approach facilitates hypothesis-free study designs, and the discovery of molecular patterns and uncharacterized transcripts, rather than confirmation of single gene expression profiles. However, the disadvantages of most high-throughput methods are that they are not as precise as RT-PCR, and next-generation sequencing is still relatively expensive for routine use.

1.3.2 Experimental steps and critical aspects of gene expression analysis using microarrays

Overall, the principle steps of most gene expression experiments are similar (Figure 2). However, to obtain high-quality, meaningful, and statistically correct results from gene expression microarray experiments, several considerations need to be addressed prior to performing experiments.

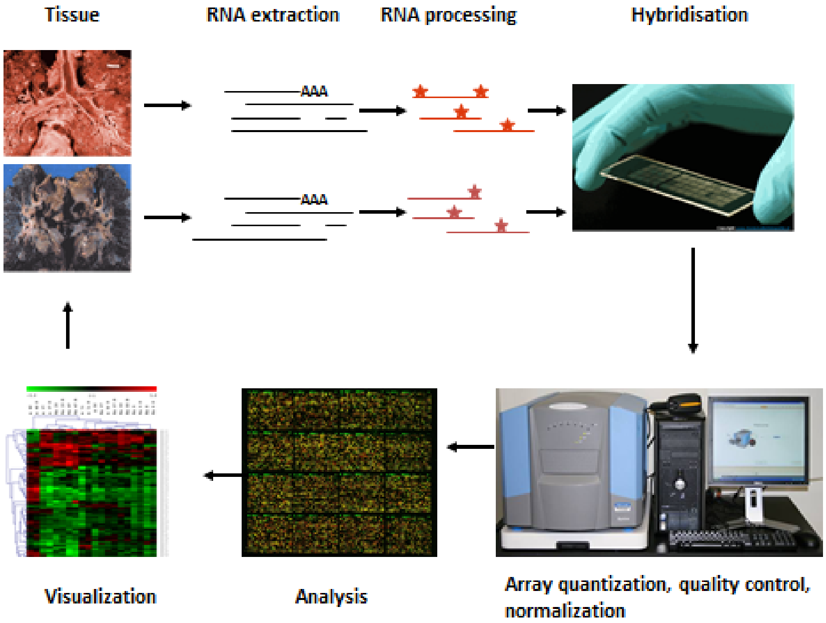


Figure 2. Overview of the steps associated with microarray-based gene expression profiling.

One of the first challenges in design and realization of gene expression experiments is the selection and availability of biological material. In lung cancer studies, the sample collection options include needle biopsy, tissue samples excision during surgery, blood sample or sputum sample collection. While the first three methods are most commonly used, they do represent invasive methods. In contrast, the collection of sputum samples is not invasive and can therefore be easily applied in routine use, yet cancerous cells from the periphery of lung may not always be available.

It is also important to consider the selection of control material, specifically whether controls will be collected from the same individual providing the cancer sample, or from another group. In the former case, environmental factors such as smoking exposure can be eliminated, thus facilitating a more precise analysis of direct gene expression changes that occur between two conditions.

Thirdly, collection and storage methods need to be determined. Once RNA and other biomolecule samples are extracted, their quantity and quality is assessed. For RNA, the subsequent steps involve amplification and labelling. Currently, Affymetrix, Illumina, Roche, and Solid are leading suppliers of whole genome analysis platforms. Despite differences in the designs of each application available, signal intensities detected by each platform are transformed into numerical values representing the gene expression. Furthermore, primary quality control for each experiment, as well as array performance, can be evaluated using software compatible with the array platform used.

1.3.3 Data analysis of WG gene expression experiments

Analysis of WG gene expression profiling consists of both a statistical phase and a descriptive phase. At the start of a statistical analysis, quality assessment and data normalisation are performed to check and smooth raw data values within and between the hybridisations, or calls, depending on the platform used. If the normalised signal intensities still have differing distributions, the outlier is reported and the sample is usually excluded from further analyses. To detect and avoid artefacts in gene expression array data, several methods are available. These include dye swap experiments for two colour array platforms, biological and technical replicates, and different normalisation algorithms (Do and Choi, 2006; Dudoit and Speed, 2000; Schmid et al., 2010). Although the aim of WG gene expression profiling is the identification of larger patterns in the data, the statistical importance of every gene in a cluster can also be assessed. Therefore, correction for multiple testing, e.g. Bonferroni correction, is essential for the identification of significant results and the exclusion of false positives.

In the second phase of gene expression analysis, grouping of genes and samples based on expression similarities and dissimilarities is performed to discover new relationships present in the samples. However, due to the matrix size associated with WG gene expression studies, visualisation methods are needed to identify patterns. These visualisation methods can include clustering (Figure 3), principal component analysis (PCA), as well as correlation analysis.

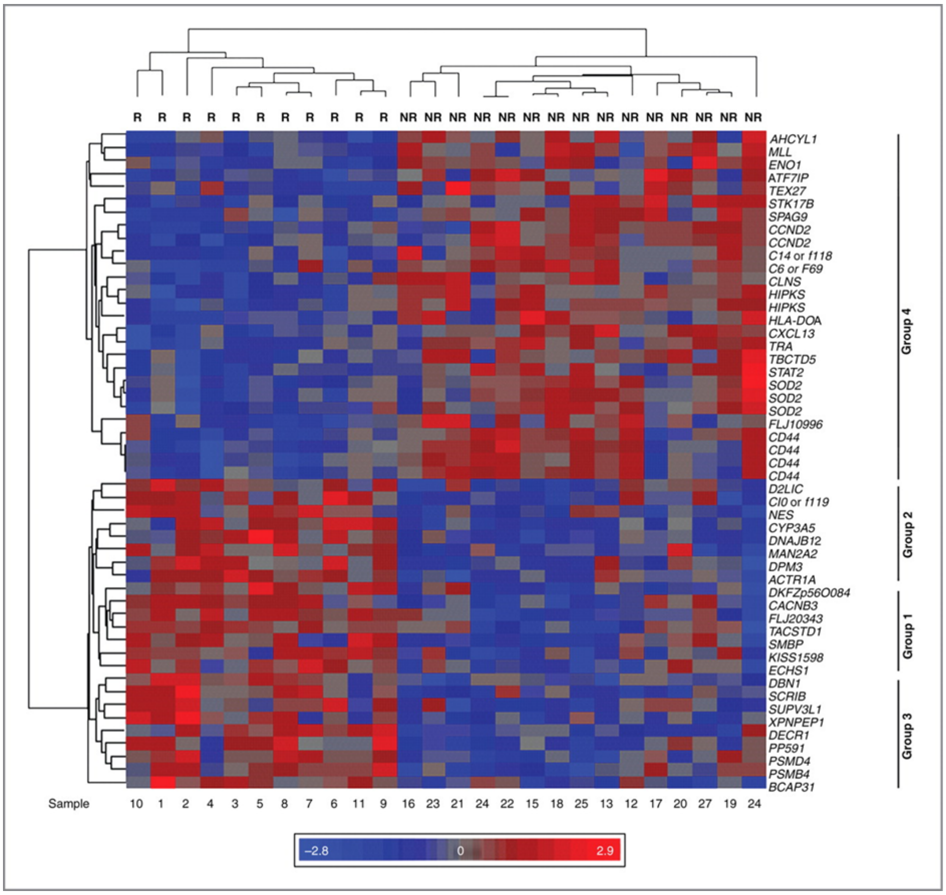


Figure 3. Example of a gene expression profile and hierarchical clustering of NSCLC data separated according to recurrent (R) and non-recurrent (NR) groups on the basis of discriminatory genes adopted from Mitra et al. (2011). Gene symbols are indicated along the right side of the figure. Red and blue are used to represent increased and decreased levels of gene expression, respectively, relative to the mean level of gene expression indicated in grey.

1.3.4 Visualisation of WG gene expression data

Clustering methods organize complex expression data sets into subgroups, or clusters, of genes that share similar expression patterns. As a result, patterns of co-regulation and possible common biological functions are identified (Eisen et al., 1998). The clustering algorithms used can either be hierarchical or partitional.

Hierarchical clustering can be agglomerative (Ramoni, Sebastiani, and Kohane, 2002) or divisive (Herrero, Valencia, and Dopazo, 2001). In the former case, the clustering process is initiated with the lowest hierarchical levels first, and upon resolution, clustering is applied to increasingly higher levels. In contrast, divisive clustering involves resolution of the highest hierarchical levels before lower hierarchical levels are resolved. The results of hierarchical clustering are displayed as dendrograms (to illustrate the hierarchy between genes and samples), (Figure 4 a) and/or heat-maps (where gene expression values are transformed into colour intensities). Additionally, distinctions between supervised and un-supervised clustering are used. Typically, supervised clustering is applied when there is additional data to consider in addition to gene expression values.

Currently, K-means clustering is the most widely used partitional clustering method (Brazma and Vilo, 2000). The advantage of this method is that the scientist can assign the number of clusters. In every run, the algorithm computes the centroid, or average, of all data points in a particular cluster. Data are then reassigned so that the centroid is surrounded by similar samples. Computation is complete when samples remain in the same cluster.

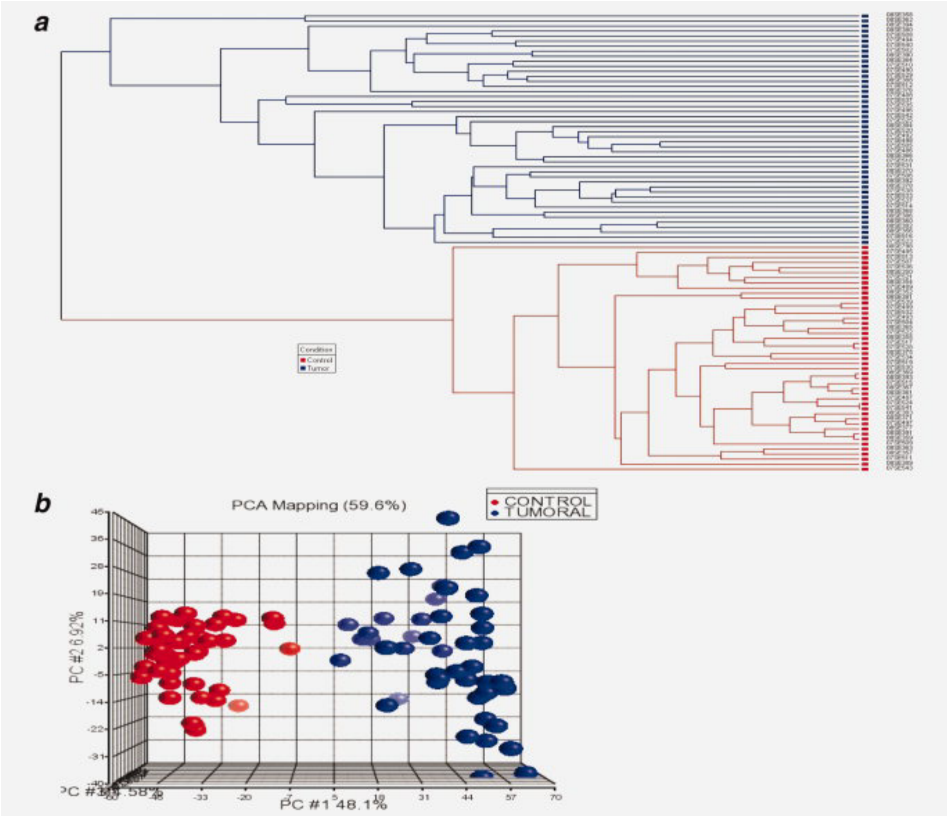


Figure 4. Adopted from Sanchez-Palencia (2010). Unsupervised hierarchical clustering (a) and (b) PCA of 91 samples using 10,263 differentially expressed sequences. Samples are colour-coded for their experimental condition: red bar: non-tumour tissue sample; blue bar: tumour tissue sample.

By performing PCA (Figure 4 b), dimensionality is reduced by recognising the most variable components present in the data (Raychaudhuri, Stuart, and Altman, 2000). As a result of this transformation, a two- or three-dimensional scatter plot is obtained, with each axis representing the direction in which the data vary the most. Typically, the first two to four principal components are highlighted, as the subsequent ones are not informative.

Another method that can be used to visualize WG gene expression data is correlation analysis. Similar to a gene expression matrix, each sample and gene is treated as a vector of many dimensions, or a point in a higher dimensional space. Correspondingly, the closer two points are the smaller the distance between them, and vice versa. The most prominent algorithms used for gene expression data correlation analysis are the Pearson correlation and the Spearman rank correlation (Hack, 2004).

1.4 Clinically important aspects of lung cancer

An accurate diagnosis of lung cancer is a prerequisite for optimal therapy. Currently, a diagnosis of lung cancer is based on radiological imaging (e.g., computed tomography, magnetic resonance tomography, positron emission tomography) and evaluation of histomorphology (e.g., histology, dedifferentiation status, other molecular features) (Vollmer et al., 2010). The goal of radiological imaging is to determine the exact location and extent of the cancer present, to evaluate the structure and vasculature of the cancer and surrounding tissues, and to explore the presence of metastasis. Based on the radiographic findings obtained, TNM staging is determined, which addresses the size of the primary cancer, the involvement of lymph nodes, and the presence or absence of metastasis, respectively (Haberkorn and Schoenberg, 2001). Currently, TNM staging is the best method available for obtaining a prognosis for lung cancer, although patients with the same stage often experience different outcomes. Correspondingly, TNM staging is constantly being re-evaluated, and the increased need for molecular markers to be incorporated into the staging system has been discussed (Tanoue and Detterbeck, 2009).

Histological evaluations can be performed prior to surgery using appropriate methods such as cryo-biopsy or needle aspiration, although the final, definitive histology of a cancer and its margins is assigned after surgery (Rivera, Detterbeck, and Mehta, 2003). However, histological evaluations can be difficult to assess. For example, it can be hard to distinguish between SCC and NSCLC, or to determine the correct subtype of NSCLC (Franklin, 2000). Correspondingly, this can have a significant impact on decisions regarding patient treatment. As a result, the importance of identifying new molecular markers, including new gene expression profiles, has become increasingly apparent, and an important area of investigation in order to apply these gene signatures to everyday clinical practice.

Once lung cancer is diagnosed, the most important issue is to select for the effective, and tolerable, therapy for the patient. Secondly, prognosis and prediction need to be addressed. Prognosis describes the outcome of the disease regardless of the treatment, and prediction refers to the efficacy, or toxicity, estimated for a given treatment (Ferte, Andre, and Soria, 2010). However, it is known that staging, as well as histomorphology of lung cancer are not informative enough to provide robust accuracy in the evaluation and prognosis of each lung cancer case. For example, while radiology can estimate the size and extent of the cancer present, and histomorphology can provide information regarding the morphology of the cancer, it is information from molecular methods that are based on DNA and RNA profiling that provides the most detailed characteristics of a cancerous tissue. Therefore, the development of novel approaches to evaluate genomic and transcriptome data are needed to evaluate disease pathogenesis.

1.5 Genome-wide gene expression profiling of NSCLC

During the past decade, a substantial number of gene expression profiling studies of lung cancer have been published. However, due to the hypothesis-free nature of WG profiling, the diversity and number of findings associated with these studies, as well as the corresponding interpretations, are difficult to succinctly summarise. Therefore, for the purpose of this literature review, only the clinically important aspects of these studies that relate to diagnosis (that relies on staging and histology), prognosis, and prediction of lung cancer will be presented.

1.5.1 Gene expression profiles and TNM staging of NSCLC

TNM staging is currently the gold standard for predicting lung cancer prognosis, and is used to select patient treatment (Tanoue, 2008). Therefore, TNM stage-associated gene expression profiles have been sought in most lung cancer studies. Unfortunately, the number of positive results obtained has been limited, suggesting that patients with the same stage of disease do not share the same molecular characteristics (Raponi et al., 2006; Tomida et al., 2004). This is further supported by the observation that staging is a poor prognostic factor (Raponi et al., 2006). However, in some studies, gene expression profiling has been evaluated in relation to nodal status and distant metastasis. Since lung cancer metastasis in local lymph nodes is a critical parameter for evaluating patient prognosis and treatment, the accurate evaluation and prediction of micro-metastases not detectable by histological and radiological methods could significantly improve lung cancer management.

Currently, Takefumi Kikuchi is one of the leading scientists in the field of gene expression profiling studies of lung cancer metastasis and nodal status. In 2003, he published a study of lung cancer lymph node profiles and sensitivity to anti-cancer drugs (Kikuchi et al., 2003). For this work, Kikuchi and co-workers used laser capture microdissection technology to harvest cancerous tissues. Then, using gene expression profiling, ACs were distinguished from SCCs. Moreover, for the ACs, two clusters were detected that represented metastasis-negative, and metastasis-positive, lymph node-associated genes. As mentioned above, lung cancer can give rise to distant metastasis, with the most common sites involving the brain (60%), breast (20%), skin (10%), and colon (5%) (Nathoo et al., 2005; Subramanian et al., 2002). In 2006, Kikuchi and colleagues used gene expression profiling and microdissection to characterize molecular patterns of primary lung AC and non-matched lung AC brain metastases (Kikuchi et al., 2006). Based on the results obtained, metastatic tumour cells were found to vary considerably from primary tumours.

In another study of NSCLC using gene expression profiling, Minoru Takada and co-workers collected larger amounts of primary cancer sections (up to 5 sections of 5 × 5 × 5 mm in size) from each patient studied for RNA extraction that enables to evaluate the broader molecular profiles of cancer (Takada et al., 2004). Analysis of these samples revealed expression profiles associated with lymph-node positive, and lymph node-negative, ACs and SCCs. In addition, profiles associated with various tumour sizes were observed. However, these results were not validated using an independent sample set.

In 2009, Yasumitsu Moriya and co-workers analysed 41 AC samples (Moriya et al., 2009) using gene expression profiling. A total of 15 predictor genes for lymph node metastasis were identified, and these findings were also evaluated using independent samples. Using these predictor genes, the accuracy in evaluating lymph node status was found to be 71%. Although the number of samples analysed in this study was relatively small, the authors concluded that a combined analysis of pathology and molecular classification has the potential to provide additional information, a better diagnosis, as well as improved treatment, for patients with NSCLC.

1.5.2 Gene expression profiles, prognosis, and histology of NSCLC

A second important clinical factor in the diagnosis of lung cancer is histological classification, where lung cancer can be categorized as small cell lung cancer (SCLC) or NSCLC (Figure 5). In 80% of lung cancer cases, NSCLC is the diagnosis. Moreover, NSCLC can be further classified as AC, SCC, or giant cell carcinoma (Mitsuuchi and Testa, 2002; Rosell et al., 2004). In the former case, a bronchioalveolar subtype of AC also exists, that is associated with a more favourable prognosis. In general, adenocarcinoma arises from bronchioles and alveolus, and is localized in the periphery of the lung. In contrast, SCC can be associated with smoking, it is usually localized to the central airways, and originates from large or medium-sized bronchial epithelium (Nacht et al., 2001). Lastly, the origin of giant cell carcinoma is currently under debate, but it exhibits neuroendocrine features and is associated with a poor prognosis (Franklin, 2000).

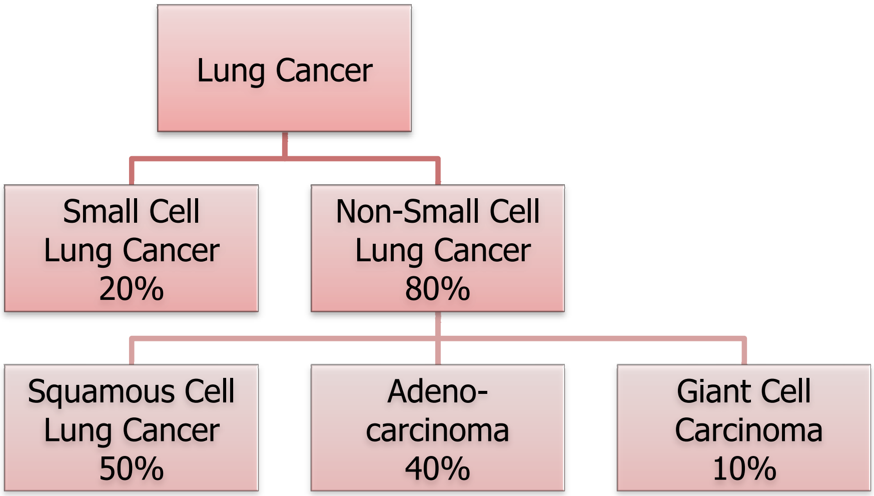

Lung Cancer
Small Cell Lung Cancer 20%
Non-Small Cell Lung Cancer 80%
Squamous Cell Lung Cancer 50%
Adeno-carcinoma 40%
Giant Cell Carcinoma 10%

Figure 5. Main histological types of lung cancer.

Histological evaluations of lung cancer is usually straight-forward for experienced pathologists. However, there are cancer samples that have mixed, or confusing, histological features. For example, in a study published by Sorensen and colleagues, three lung pathologists independently examining the same set of samples consistently agreed only on a histology of lung AC in less than 50% of cases (Sorensen et al., 1993). It is hypothesized that the use of global gene expression profiling to identify molecular subtypes not detectable by histology could eventually help classify histologically mixed samples by identifying the tissue of origin and distinguishing cancer subtypes to obtain a more precise prognosis. Correspondingly, a study of NSCLC histology and molecular profiles was conducted by Dracheva et al. (2007) to identify distinct sets of genes associated with cancerous versus control lung tissue samples (Dracheva et al., 2007). To improve the validity of their study, four additional, independent datasets were evaluated. As a result, 20 genes were identified that provided robust discrimination of gene expression associated with cancerous tissue versus non-cancerous tissue. The development of this gene assay for the clinic is anticipated since only~300 cells are needed for the RNA extraction used in this protocol.

In addition, there have been several publications that have identified distinct molecular profiles for AC versus SCC, as well as different molecular subtypes within a particular histological type (Table 2).

Table 2. List of publications containing gene expression data of lung cancers used to identify molecular subtypes. AC-adenocarcinoma, SCC-squamous cell lung cancer, SCLC- small cell lung cancer, LCC-large cell lung cancer, NSCLC- non-small cell lung cancer.

Histology analysed	No. samples	Molecular profiles	Survival prediction	Article
AC/SCC/Norm/SCLC/LCC	67	7	+	(Garber et al., 2001)
AC/SCC/Norm/SCLC/ Carcinoid/Metastasis	220	9	+	(Bhattacharjee et al., 2001)
AC	86	3	+	(Beer et al., 2002)
NSCLC	39	2	+	(Wigle et al., 2002)
AC/SCC	50	6	+	(Tomida et al., 2004)
AC/SCC	58	2	–	(Kuner et al., 2009)
AC/SCC/LCC/BAC/Carcinoid	178	5	+	(Hou et al., 2010)
AC/SCC/ASC (study on rats)	19	3	–	(Bastide et al., 2010)

In addition to the studies listed in Table 2, gene expression profiles that describe various prognostic gene sets have also been reported. For example, two sets of gene signatures including 35 genes versus 6 genes have been associated with the prediction of NSCLC recurrence (Guo et al., 2008; Lee et al., 2008), 64 genes have been used to evaluate stage I NSCLC survival (involving a good versus bad prognosis) (Lu et al., 2006), 50 genes versus 12 genes have been associated with a prognostic signature for SCC (Raponi et al., 2006; Zhu et al., 2010b), and 10 genes have been shown to provide a prognosis for stage I AC (Bianchi et al., 2007).

In combination, these results demonstrate that molecular classification of tumours is a promising approach for identifying sub-classifications of histological groups of lung cancer. However, these results have not been reproducible and consistent (Ransohoff, 2007). For example, using molecular classification of ACs, Bhattacharjee et al. identified four subgroups, while Garber et al. identified three subgroups (Bhattacharjee et al., 2001; Garber et al., 2001). Moreover, different sets of genes are associated with each study. In a recent review article, 16 sets of gene expression-based prognostic signatures for lung cancer were compared (Subramanian and Simon, 2010). The authors concluded that serious flaws existed in the design and analysis used in each of these studies, and the use of these signatures in clinical practice would not be reasonable. Therefore, it has become apparent that in order to achieve meaningful results associated with reduced noise, WG expression studies need to include a similar number of cases and controls as is used in WGAS, and new methods such as next-generation transcriptome sequencing need to be incorporated.

1.5.3 Gene expression profiles and predicted treatment responses of NSCLC

Despite the apparent ability of surgeons to completely resect cancerous lung tissue, 33% of stage IA patients, and 77% of stage IIIA patients, die within 5 years of initial diagnosis (Xie and Minna, 2010). This has mainly been attributed to metastatic disease present at the time of surgical resection. Furthermore, when adjuvant chemotherapy is administered following resection, patient survival has been shown to improve, yet is often accompanied by serious adverse effects (Douillard, 2010; Douillard et al., 2006; Winton et al., 2005). Therefore, a predictive biomarker, or combination of biomarkers, that could identify patient groups according to predicted treatment responses, would have a significant impact on clinical decisions concerning treatment selection. Currently, the most promising predictive gene expression marker for lung cancer is excision repair cross-complementation group 1 (ERCC1) (Olaussen et al., 2006), whose gene product has been shown to play a role in the repair of cisplatin-generated DNA adducts. Correspondingly, ERCC1-negative NSCLCs have been shown to be more responsive to cisplatin-based chemotherapy compared with ERCC1-positive NSCLCs (Figure 6).

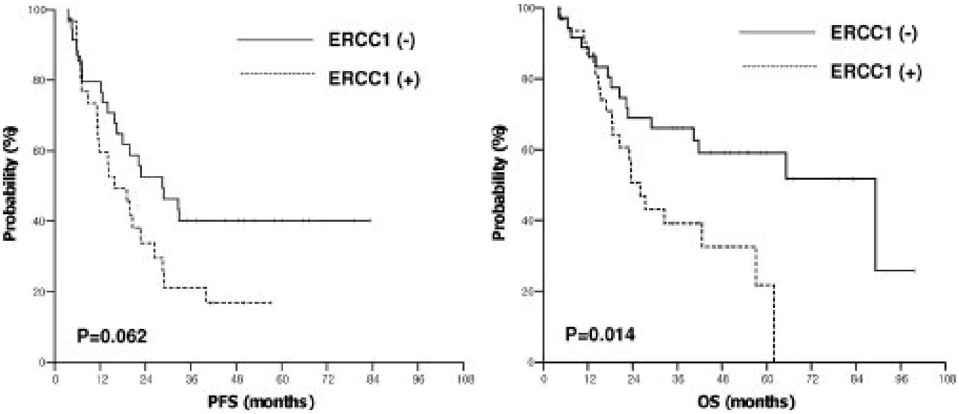


Figure 6. Progression-free survival (PFS) and overall survival (OS) curves of patients receiving platinum-based treatments according to ERCC1 expression. This figure is adopted from Hwang et al. (2008).

Recently, Zhu and co-workers have described a predictive 15-gene expression signature (Zhu et al., 2010a) for evaluating patient prognosis. In this study, patients whose tumours were predicted to have a poor prognosis, yet received adjuvant chemotherapy, exhibited an improved response compared with patients whose tumours showed a poor prognosis signature and did not receive treatment. In contrast, patients with tumours associated with a good prognosis signature that received adjuvant chemotherapy did significantly worse than patients with a good prognosis signature that did not receive treatment. Therefore, this gene expression signature appears to identify patients with a good prognosis who have an increased risk of performing worse following treatment, versus patients with a poor prognosis who would benefit from additional treatment.

In addition, for genome-wide profiling, a large number of patients with accompanying clinical data, as well as good quality biological samples, are needed. Since treatment decisions and patient outcome are dependent on evaluations of these signatures, an independent training cohort and validation of the results obtained are also critical.

2. AIMS OF THE CURRENT STUDY

The aims of the present studies included:

I. To use genome-wide association study approach to identify new SNPs and LD blocks associated with predisposition to NSCLC.

II. To use genome-wide gene expression profiling to discover novel lung cancer-associated genes and molecular patterns associated with different clinical features of NSCLC.

III. To evaluate the degree of variance in gene expression profiles between three NSCLC samples and two lung tissue control samples obtained from the same patient before and after the administration of radio-chemotherapy.

3. RESULTS AND DISCUSSION

3.1 Lung cancer patients, clinical data, and biological samples used

A total of 146 patients diagnosed with NSCLC underwent surgery between November 28, 2002 and December 31, 2006 at the Clinic of Cardiovascular and Thoracic Surgery of Tartu University Hospital, Estonia. All patients gave their written informed consent to participate in the study, to allow their biological samples to be genetically analysed, and for their clinical data to be reviewed. The Ethics Review Committee on Human Research of the University of Tartu approved the current study. Both tumour specimens and control samples were collected during the surgeries performed, and the departmental pathologist promptly examined all specimens. Tumour histology and stage for each collected sample were estimated according to WHO guidelines (Travis, Travis, and Devesa, 1995) and TNM staging criteria according to International Union Against Cancer (UICC) classifications (Mountain, 2000). Furthermore, the same pathologist determined all histological classifications. Control samples were also obtained from each cancer patient at a site distant from the tumour, and were approved as control samples by the pathologist.

3.2 GWAS meta-analysis of lung cancer associated loci. Ref. I

Although tobacco smoking is a major risk factor for lung cancer, an individual's genetic background also plays an important role. For example, many WGAS have identified pathways and genes associated with lung cancer that can contribute to smoking addictions and smoking-related carcinogen-induced damage repair. Correspondingly, it is hypothesized that the study of subjects with a susceptible genetic background to smoking and lung cancer will represent an important step in improving personalised genetics and medicine. One approach to elucidate statistically significant novel loci, and to confirm previous results of WGAS of lung cancer, is the analysis of large, pooled datasets consisting of tens of thousands of samples and controls. Correspondingly, a meta-analysis of 14 lung cancer studies involving individuals of European descent (described in Reference I) was performed. In this study, 13 300 primary lung cancer cases and 19 666 controls were included, as well as 109 Estonian samples and 874 controls (Ref. I, Table 1). Genome-wide analysis of the data confirmed all previous findings of lung cancer association studies that identified significant roles for the 5p15, 15q25, and 6p21 chromosomal regions, yet no novel statistically significant loci were identified (Ref. I, Suppl. Table 5). However, new candidate genes and SNPs of lung cancer susceptibility emerged (Ref. I, Suppl. Table 5–7), and these will need to be further investigated in fine mapping studies. Lung cancer association studies will also need to be performed for samples of non-European descent for comparison with previously published findings.

3.3 Gene expression profiles of NSCLC: survival prediction and new biomarkers. Ref. II

Despite the well-defined histology associated with subtypes of NSCLC, a given stage is often associated with survival rates and treatment outcomes that vary considerably from patient to patient (Brambilla et al., 2001). In addition, a broad spectrum of lung cancer morphologies have been observed, with many tumours being atypical, or characterized by a lack of morphologic features necessary for an improved differential diagnosis. Therefore, lung cancer diagnoses based solely on morphological features are usually insufficient (D'Amico, 2008). As a result, there is an increased demand for the discovery and identification of new informative biomarkers that could be applied independently, or in combination, with histological and morphological evaluations for diagnosis and prognosis of lung malignancies.

In the study presented in Reference II, an Illumina BeadChip platform and corresponding Human-6 Expression Whole-Genome arrays containing more than 48 000 transcript probes were employed to elucidate molecular profiles and novel biomarkers associated with NSCLC. After the exclusion of samples due to pre-operative chemotherapy, RNA degradation, and final diagnosis, 81 samples were available for analysis (Ref. II, Table 1). Histology confirmed that 13 cases involved BACs, 8 cases involved ACs, and 60 cases involved SCCs (Table 3).

Table 3. Detailed clinical and pathological characteristics of patients enrolled in the gene expression study of Ref. II (N = 81) following medical exclusion and RNA integrity number (RIN) cut-off.

Clinicopathological characteristics	No. of Patients	%
Histology		
Adenocarcinoma	8	9.90%
Bronchioloalveolar carcinoma	13	16.00%
Squamous cell carcinoma	60	74.10%
Lymph node		
Positive	13	16.00%
Negative	68	84.00%
Differentiation		
Well/moderate	76	94.00%
Poor/undifferentiated	5	6.00%
Stage		
Ia	13	16.10%
Ib	46	56.80%

Clinicopathological characteristics	**No. of Patients**	**%**
IIa	1	1.20%
IIb	3	3.70%
IIIa	7	8.60%
IIIb	6	7.40%
IV	5	6.20%
T1	15	18.50%
T2	56	69.10%
T3	5	6.20%
T4	5	6.20%
Tumour size (mm)		
< 30	36	44.40%
> 30	45	55.60%
Surgical procedure		
Wedge resection	6	7.40%
Lobectomy	54	66.70%
Bilobectomy	3	3.70%
Pneumonectomy	18	22.20%
Gender		
Female	9	11.10%
Male	72	88.90%
Age, years		
Range	38-81	
Mean	65.8	
Median	68	
< 39	1	1.20%
40–49	5	6.20%
50–59	13	16.00%
60–69	27	33.30%
> 70	35	43.20%
Smoking status		
Non-smoker	2	2.50%
Smoker	79	97.50%
Family history of cancer	9	11.10%
Occupational exposure	9	11.10%
None	72	88.90%
Possible	9	11.10%

3.3.1 Differentially expressed NSCLC genes

A total of 997 statistically significant, differentially expressed transcripts were identified from a comparison of paired NSCLC samples and control samples obtained from each individual. Of these, 326 involved up-regulated genes and 671 involved down-regulated genes. Moreover, a large number of previously described NSCLC-associated genes were identified (Ref. II, Suppl. Table). Novel, up-regulated genes included *SPAG5*, *POLQ*, *KIF23*, and *RAD54L*, which are associated with mitotic spindle formation, DNA repair, chromosome segregation, and dsDNA brake repair, respectively. The down-regulated genes included *SGCG*, *NLRC4*, *SFTPA1B*, *MMRN1*, and *SFTPD*, which have roles in extracellular matrix formation, apoptosis, blood vessel leakage, and inflammation, respectively (Table 4).

Table 4. Novel up- and down-regulated genes associated with NSCLC.

Adjusted p-value	Mean fold change	Gene symbol	Gene name and source
Up-regulated in cancer tissues			
5,41E-16	2,2	C6ORF129	Chromosome 6 open reading frame 129
9,50E-14	3,6	SPAG5	Sperm-associated antigen 5, map126, deepest
1,11E-11	3,3	POLQ	DNA polymerase theta, [source: uniprot/swissprot; acc:o75417]
5,42E-11	2,1	C6ORF125	Uncharacterized protein c6orf125
3,13E-10	2,9	KIF23	Kinesin-like protein kif23
1,01E-09	2,2	RAD54L	DNA repair and recombination protein rad54-like
1,24E-09	2,2	C12ORF48	upf0419 protein c12orf48
2,04E-09	2,0	C16ORF33	u11/u12 snrnp 25 kDa protein (minus-99 protein)
4,38E-09	2,0	RAB26	Ras-related protein rab-26
4,94E-09	2,0	ARHGEF19	Rho guanine nucleotide exchange factor 19
Down-regulated in cancer tissues			
3,14E-20	2,4	SGCG	Gamma-sarcoglycan
2,22E-17	3,5	NLRC4	Caspase recruitment domain-containing protein 12
1,75E-16	2,1	VAPA	Vesicle-associated membrane protein-associated protein
3,11E-15	9,5	SFTPA1B	Pulmonary surfactant-associated protein a1 precursor
2,65E-12	2,0	MMRN1	Multimerin-1 precursor (endothelial cell multimerin 1)
3,52E-08	10,9	SFTPD	Pulmonary surfactant-associated protein d precursor
6,44E-07	2,0	SELPLG	P-selectin glycoprotein ligand 1 precursor
7,99E-07	2,1	PCDH17	Protocadherin-17 precursor (protocadherin-68)

A hierarchical cluster analysis was also performed (Figure 7) to represent the distribution of the 997 differentially expressed transcripts identified. Two distinct groups of NSCLC genes were found (e.g., Group 1 and Group 2), while the control samples were associated with a single cluster. Different histological types were also observed to be randomly positioned. Based on this analysis, no clear association between the NSCLC gene expression profiles obtained and NSCLC stage, smoking cessation, or patient gender were observed.

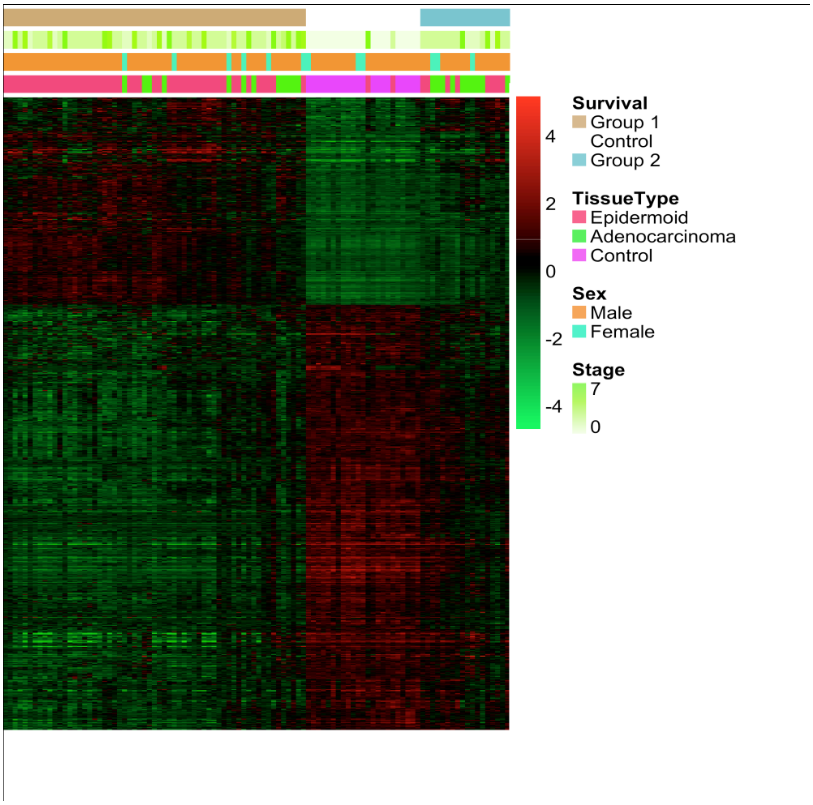


Figure 7. Gene expression heat-map of NSCLC and control samples classified by TNM stage. Control sample = stage 0; Ia = stage 1; Ib = stage 2; IIa = stage 3; IIb = stage 4; IIIa = stage 5; IIIb = stage 6, and IV = stage 7.

Since survival prediction is one of the key parameters associated with molecular diagnostics, Kaplan-Meier survival curves were generated based on the gene expression profiles associated with Group 1 and Group 2 (identified above), as well as histology. Although the time between the initial surgical resection and survival analysis was limited to less than 7 years, and the results of the analysis did not achieve statistical significance (p = 0.0691 for expression-based groups and p = 0.0198 for histology-based groups), enhanced predictive p-values were detected for a group that was selected based on gene expression profiles (Figure 8).

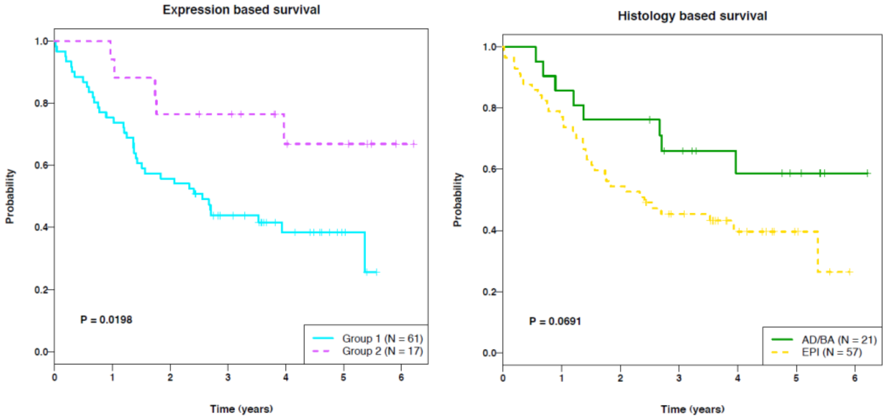


Figure 8. Kaplan-Meier survival curves for NSCLC patients. Patients were grouped according to the gene expression profiles of statistically significant up- and down-regulated genes (n = 997). It was observed that survival curves based on gene expression profiles of NSCLC patients yielded improved survival predictions than groupings based on histology. AD/BA = AC and bronchioalveolar cancer; EPI = epidermoid cancer (SCC).

3.3.2 Tumour RNA degradation and prognosis of NSCLC

To investigate the hypothesis that RNA degradation in NSCLC specimens is associated with disease prognosis, a survival analysis was performed for patients with lung AC. Based on these data, it was observed that patient survival associated with cancer samples containing intact RNA was found to be significantly improved compared with patients that had RNA degradation detected in their cancer samples (p = 0.0474) (Figure 9).

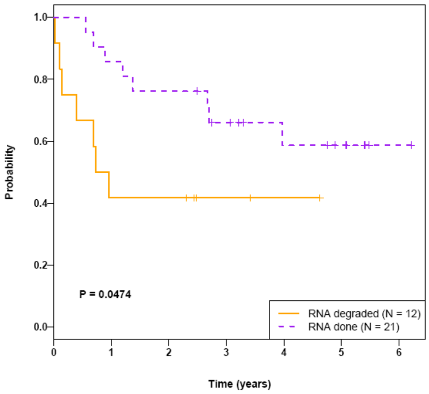


Figure 9. Kaplan-Meier survival curves for patients with lung AC grouped according to RNA integrity. Twelve 12 AC samples were associated with low RNA integrity (RIN < 7), while 21 samples exhibited high levels of RNA integrity. Lung AC patients whose tumour specimens contained intact RNA had a statistically significantly higher survival prediction (p = 0.0474).

3.4 Gene expression-based approaches for the differentiation of metastases versus a second primary tumour site. Ref. III

Cancer treatment schemes differ substantially for metastasis events versus primary tumours, for differentiated versus de-differentiated tumours, and for AC versus SCC. Moreover, an accurate diagnosis for patients with multiple cancers at different sites, whether within the same organ or not, can be extremely difficult. In the latter case, the combination of histological evaluations and gene expression profiling has the potential to improve diagnosis. In addition, since gene expression profiling can detect thousands of genes simultaneously, treatment decisions can be made according to a specific individual's genetic background and disease nature.

In the study described in Reference III, an Illumina whole-genome HumanHT-12 v3 Expression BeadChip was used to explore the gene expression profiles of three consecutive NSCLC samples, and three paired control samples, obtained from the same patient (Figure 10). Based on the histological patterns and clinical performance of this patient (which were better than predicted for a patient with metastatic cancer), the presence of a second primary disease, rather than metastasis, was considered. Therefore, gene ontology (GO) and PCA were performed to elucidate and compare the biological patterns and clinical make-up of the samples collected.

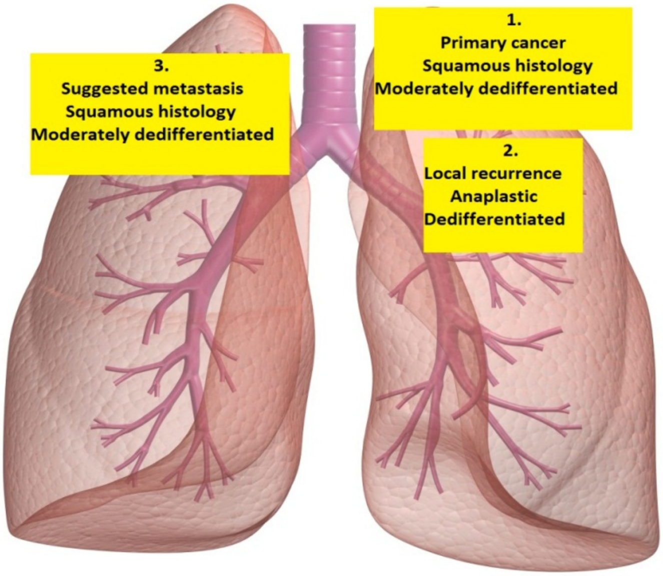
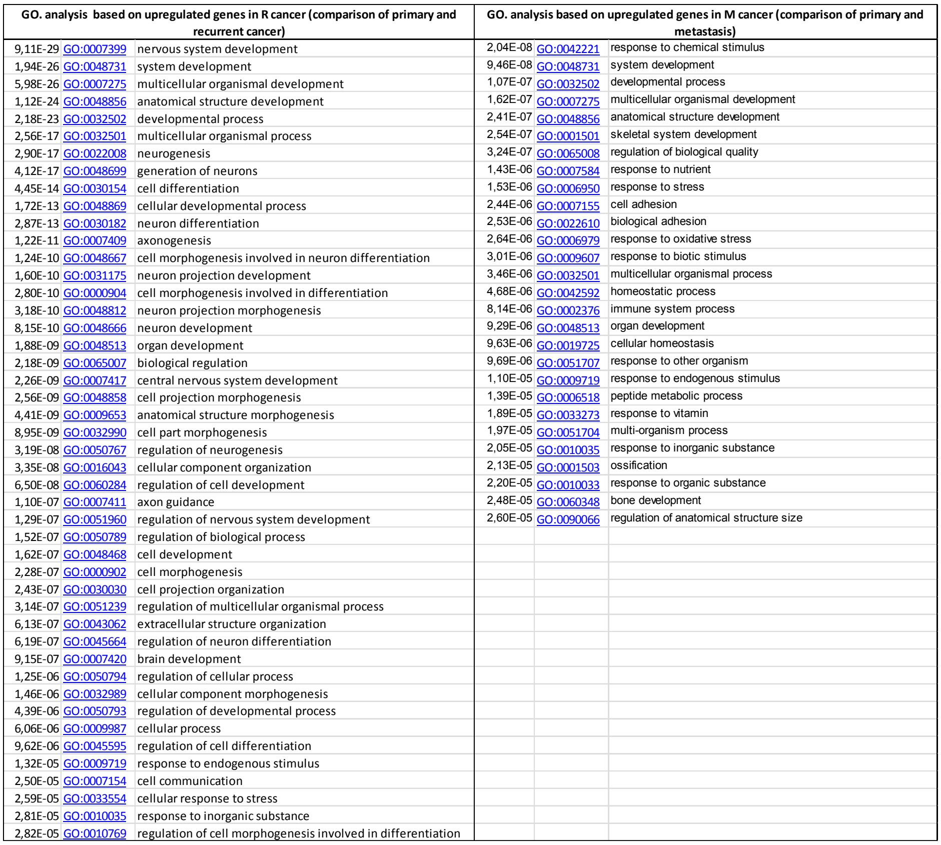


Figure 10. Clinical aspects of a patient that presented with NSCLC for evaluation. Three boxes represent the location and order of the cancers diagnosed.

GO analysis (Table 5) of the primary cancer and a potential metastasis identified gene expression changes associated with system and organ development, adhesion, oxidative stress, homeostasis, as well as ossification of the metastasis sample. However, well-characterized biological hallmarks of metastasis-associated processes such as dedifferentiation, extensive metabolism, DNA synthesis, and inflammation were not noted. In addition, the genes that were found to be down-regulated in the proposed metastasis sample involved deactivation of genes with a role in cellular localization, cytoskeleton and organelle organization, glucose catabolism, and locomotion. Therefore, this profile was more consistent with an active primary cancer than a metastasis. Correspondingly, GO analyses also did not support the presence of a metastasis.

Table 5. GO analysis of a primary cancer versus a recurrent or a metastatic cancer.



GO. analysis based on upregulated genes in R cancer (comparison of primary and recurrent cancer)			GO. analysis based on upregulated genes in M cancer (comparison of primary and metastasis)		
9,11E-29	GO:0007399	nervous system development	2,04E-08	GO:0042221	response to chemical stimulus
1,94E-26	GO:0048731	system development	9,46E-08	GO:0048731	system development
5,98E-26	GO:0007275	multicellular organismal development	1,07E-07	GO:0032502	developmental process
1,12E-24	GO:0048856	anatomical structure development	1,62E-07	GO:0007275	multicellular organismal development
2,18E-23	GO:0032502	developmental process	2,41E-07	GO:0048856	anatomical structure development
2,56E-17	GO:0032501	multicellular organismal process	2,54E-07	GO:0001501	skeletal system development
2,90E-17	GO:0022008	neurogenesis	3,24E-07	GO:0065008	regulation of biological quality
4,12E-17	GO:0048699	generation of neurons	1,43E-06	GO:0007584	response to nutrient
4,45E-14	GO:0030154	cell differentiation	1,53E-06	GO:0006950	response to stress
1,72E-13	GO:0048869	cellular developmental process	2,44E-06	GO:0007155	cell adhesion
2,87E-13	GO:0030182	neuron differentiation	2,53E-06	GO:0022610	biological adhesion
1,22E-11	GO:0007409	axonogenesis	2,64E-06	GO:0006979	response to oxidative stress
1,24E-10	GO:0048667	cell morphogenesis involved in neuron differentiation	3,01E-06	GO:0009607	response to biotic stimulus
1,60E-10	GO:0031175	neuron projection development	3,46E-06	GO:0032501	multicellular organismal process
2,80E-10	GO:0000904	cell morphogenesis involved in differentiation	4,68E-06	GO:0042592	homeostatic process
3,18E-10	GO:0048812	neuron projection morphogenesis	8,14E-06	GO:0002376	immune system process
8,15E-10	GO:0048666	neuron development	9,29E-06	GO:0048513	organ development
1,88E-09	GO:0048513	organ development	9,63E-06	GO:0019725	cellular homeostasis
2,18E-09	GO:0065007	biological regulation	9,69E-06	GO:0051707	response to other organism
2,26E-09	GO:0007417	central nervous system development	1,10E-05	GO:0009719	response to endogenous stimulus
2,56E-09	GO:0048858	cell projection morphogenesis	1,39E-05	GO:0006518	peptide metabolic process
4,41E-09	GO:0009653	anatomical structure morphogenesis	1,89E-05	GO:0033273	response to vitamin
8,95E-09	GO:0032990	cell part morphogenesis	1,97E-05	GO:0051704	multi-organism process
3,19E-08	GO:0050767	regulation of neurogenesis	2,05E-05	GO:0010035	response to inorganic substance
3,35E-08	GO:0016043	cellular component organization	2,13E-05	GO:0001503	ossification
6,50E-08	GO:0060284	regulation of cell development	2,20E-05	GO:0010033	response to organic substance
1,10E-07	GO:0007411	axon guidance	2,48E-05	GO:0060348	bone development
1,29E-07	GO:0051960	regulation of nervous system development	2,60E-05	GO:0090066	regulation of anatomical structure size
1,52E-07	GO:0050789	regulation of biological process			
1,62E-07	GO:0048468	cell development			
2,28E-07	GO:0000902	cell morphogenesis			
2,43E-07	GO:0030030	cell projection organization			
3,14E-07	GO:0051239	regulation of multicellular organismal process			
6,13E-07	GO:0043062	extracellular structure organization			
6,19E-07	GO:0045664	regulation of neuron differentiation			
9,15E-07	GO:0007420	brain development			
1,25E-06	GO:0050794	regulation of cellular process			
1,46E-06	GO:0032989	cellular component morphogenesis			
4,39E-06	GO:0050793	regulation of developmental process			
6,06E-06	GO:0009987	cellular process			
9,62E-06	GO:0045595	regulation of cell differentiation			
1,32E-05	GO:0009719	response to endogenous stimulus			
2,50E-05	GO:0007154	cell communication			
2,59E-05	GO:0033554	cellular response to stress			
2,81E-05	GO:0010035	response to inorganic substance			
2,82E-05	GO:0010769	regulation of cell morphogenesis involved in differentiation			

PCA (Figure 11) revealed similar gene expression changes in control samples excised during a first and third operation, despite two chemotherapy treatments that were administered between these two collections. The recurrent dedifferentiated cancer sample that was removed during the second operation showed the largest difference (distance) from the controls. Moreover, in PCA chart the potential metastasis sample was located much closer to the control sample than the primary cancer.

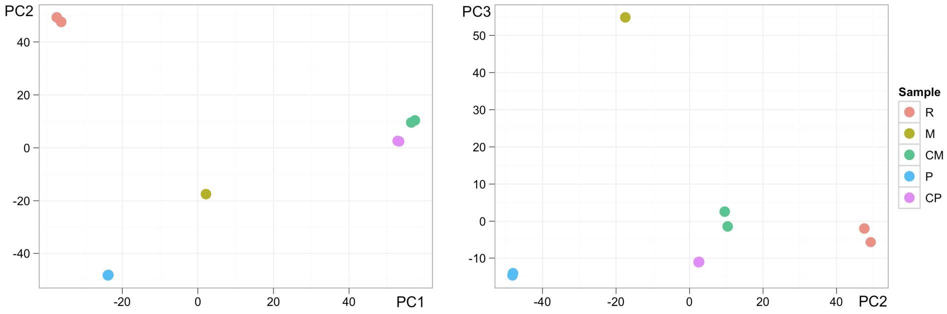


Figure 11. PCA of recurrent (R), metastasis (M), control of metastasis (CM), primary (P) and primary control (CP) samples. Replicate array data were available for all of these except the M sample.

Thus, according to GO analyses, the potential metastasis only differed minimally from the primary cancer, and no activation of processes characteristic of metastatic cancers, such as matrix remodelling, metastasis, dedifferentiation, mitosis, etc., were detected. These data were consistent with the interpretation of the PCA, thereby supporting the need to re-evaluate the proposed metastasis as a second primary cancer. In addition, although there were no histological signs of chemotherapy administration in the proposed metastasis sample, the GO analysis revealed that processes associated with a chemical stimulus were up-regulated. Therefore, it is hypothesized that the proposed metastasis was present, but not yet detectable, at the time of treatment for the recurrent cancer.

CONCLUSIONS

The following conclusions are made based on the studies conducted in the current thesis:

I. In WGAS of lung cancer described in the current thesis, 13 300 primary lung cancer cases and 19 666 controls were analysed. Genome-wide analysis of these data confirmed the findings of previous lung cancer association studies that mapped to the 5p15, 15q25, and 6p21 chromosomal regions, and no additional statistically significant loci were identified. Based on these results, one can hypothesize that, with the technology applied, the limits of WGAS have been reached. Therefore, the discovery of new associations will require deep sequencing of whole genomes, larger meta-analyses, and an analysis of patients of non-European descent.

II. Diagnosis, prognosis, and prediction are key aspects for the accurate treatment of any disease. WG expression profiling of lung cancer has been successfully applied to address all of these aspects, although overlap of the results obtained has been relatively modest. This may be due to the small number of studies that have been performed, and substantial differences that exist between individuals with lung cancer. However, one approach that has not been employed in the analysis of lung cancer gene expression profiling is a large meta-analysis of thousands of samples and controls. In expression profiling experiments of whole genomes, which are described in the current thesis, novel, potential lung cancer biomarkers were identified, as well as molecular profiles that predicted the outcome of patients with greater accuracy than histology. Moreover, an analysis of AC samples identified statistically significant correlations between RNA degradation and patient survival.

III. In the third publication of this thesis, PCA and GO analyses were applied to WG expression profiling data obtained from a patient presenting with NSCLC in order to investigate the presence of a metastatic tumour versus a second primary tumour. Moreover, due to the availability of different samples, we were able to demonstrate that gene expression profiling represents a valuable method for the diagnosis of complicated tumour samples.

SUMMARY IN ESTONIAN

Mitteväikerakulise kopsuvähi kogugenoomi geeniekspressiooni- ja assotsiatsiooniuuringud

Maailmas avastatakse igal aastal 1.6 miljonit uut kopsuvähi juhtu, Eestis ligikaudu 700, millest valdav enamus on diagnoosimise hetkeks juba kaugele arenenud ning radikaalset ravi ei rakendata. Kuigi kopsuvähi peamisteks riskifaktoriteks on suitsetamine ja õhusaaste, kokkupuude asbesti ja raskmetallidega, mängib haiguse tekkes ja arengus olulist rolli ka indiviidi geneetiline taust.

Nii nagu enamus haiguste puhul, on ka kopsuvähi ravistrateegia valiku ja prognoosi aluseks täpne diagnoos. Kopsuvähi määratlemise aluseks on TNM klassifikatsioon ja histoloogiline analüüs. Kuigi TNM klassifikatsioon on aastate jookul oluliselt täpsustunud, on siiski gruppide sisene ravivastus ja elulemus väga varieeruv. Samuti esineb märkimisväärsel hulgal vähkkasvajaid, mille histoloogiline määratlemine on võimatu, kas siis dediferentseerumise või segatüübilise kasvaja tõttu. Sellest tulenevalt on asutud otsima uusi prognostilisi ja diagnostilisi molekulaarseid markereid, mis võimaldaksid täpsemalt klassifitseerida kopsuvähki ja juhtida raviarsti sobilikuma ravi valikul.

Käesolevas doktoritöös on kasutatud kogu genoomi hõlmavat assotsiatsiooni- ja geeniekspressiooni analüüsi eesmärkidega identifitseerida kopsuvähi riskialleele ning vähkkasvajate spetsiifilisi ekspressiooniprofiile.

Geneetilisest taustast tingitud kopsuvähi ja suitsetamisest kergesti sõltuvesse jäävate isikute tuvastamine on oluline aspekt nii personaalgeneetikas kui ka – meditsiinis. Enamus tuvastatud kopsuvähi riskialleele asub kromosoomipiirkondades, kus paiknevad geenid, mis on seotud kas tubakasuitsetamise sõltuvusega või tubakasuitsust tingitud kartsinogeense toime neutraliseerimisega. Käesolevas töös käsitletud assostsiatsiooniuuringu metaanalüüsi käigus kasutati 13 300 kopsuvähi ja 19 666 kontrollindiviidi genoomi andmeid, mille tulemusel leidsid kinnitust kõik eelnevalt identifitseeritud kopsuvähi riskilookused. Samuti tuvastati uusi potentsiaalseid kopsuvähiga seotud kromosoomilookuseid, mis aga kahjuks ei osutunud statistiliselt olulisteks. Sellest tulenevalt võib järeldada, et käsoleval hetkel rakendatavate metoodikatega on kopsuvähi valdkonnas jõutud piirini, kus uusi riskiallele on raske tuvastada. Kogu genoomi sekveneerimine, veelgi suuremad metaanalüüsid ning mitteeuroopa päritolu patsientide kasutamine on tõenäoliselt moodusteks, mis võimaldavad kopsuvähi assotsiatsiooniuuringutes uusi positiivseid tulemusi.

Kopsuvähi kogugenoomi ekspressiooniprofiilide analüüsid on andnud paljutõotavaid tulemusi nii vähkkasvajate diferentseerimise, prognoosi, diagnoosi kui ka ravivastuse osas. Kahjuks on enamus tulemusi omavahel võrreldamatud või mittekattuvad, mille põhjusteks võib olla iga individuaalse kopsuvähi ekspressioonimustri suur varieeruvus või siiski veel väike uuringute üldarv. Käesoleva töö raames teostatud kogugenoomi geeniekspressiooni analüüsil identifitseeriti uued potentsiaalsed kopsuvähi biomarkerid ning molekulaarsed profiilid, mis ennustasid patsientide elulemust paremini kui histoloogiline klassifitseerimine.

Lisaks avastati, et patsiendid, kellede adenokartsinoomi proovides oli RNA lagunenud, omasid ka statistiliselt olulist kehvemat prognoosi.

Käesoleva doktoritöö kolmandas artiklis on uuritud ühe patsiendi kahe kontrollkoe ja kolme kopsuvähi kasvaja kogugenoomi geeniekspressiooni profiile selgitamaks metastaasi või teise primaarkasvaja hüpoteesi. Lisaks geeniekspressiooni profiilide rakendamise testimisest personaalses meditsiinis oli töö ajendatud ka asjaolust, et nii prognoos kui ka ravistrateegia on metastaseerunud ja mittemetastaseerunud kopsuvähi korral väga erinev. Geeniontoloogia, peakomponent- ning korrelatsioonianalüüs viitasid üheselt uue primaarkasvaja tekkele, mida kinnitab ka tänaseni säilinud patsiendi suhteliselt hea kliiniline pilt.

REFERENCES

(2–05). A haplotype map of the human genome. *Nature* **437**(7063)**,** 1299–320.

(2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**(7145)**,** 661–78.

Alwine, J. C., Kemp, D. J., and Stark, G. R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci U S A* **74**(12)**,** 5350–4.

Amos, C. I., Wu, X., Broderick, P., Gorlov, I. P., Gu, J., Eisen, T., Dong, Q., Zhang, Q., Gu, X., Vijayakrishnan, J., Sullivan, K., Matakidou, A., Wang, Y., Mills, G., Doheny, K., Tsai, Y. Y., Chen, W. V., Shete, S., Spitz, M. R., and Houlston, R. S. (2008). Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet* **40**(5)**,** 616–22.

Anderson, C. A., Pettersson, F. H., Barrett, J. C., Zhuang, J. J., Ragoussis, J., Cardon, L. R., and Morris, A. P. (2008). Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *Am J Hum Genet* **83**(1)**,** 112–9.

Bailey-Wilson, J. E., Amos, C. I., Pinney, S. M., Petersen, G. M., de Andrade, M., Wiest, J. S., Fain, P., Schwartz, A. G., You, M., Franklin, W., Klein, C., Gazdar, A., Rothschild, H., Mandal, D., Coons, T., Slusser, J., Lee, J., Gaba, C., Kupert, E., Perez, A., Zhou, X., Zeng, D., Liu, Q., Zhang, Q., Seminara, D., Minna, J., and Anderson, M. W. (2004). A major lung cancer susceptibility locus maps to chromosome 6q23–25. *Am J Hum Genet* **75**(3)**,** 460–74.

Bastide, K., Ugolin, N., Levalois, C., Bernaudin, J. F., and Chevillard, S. (2010). Are adenosquamous lung carcinomas a simple mix of adenocarcinomas and squamous cell carcinomas, or more complex at the molecular level? *Lung Cancer* **68**(1)**,** 1–9.

Beer, D. G., Kardia, S. L., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M., Iannettoni, M. D., Orringer, M. B., and Hanash, S. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* **8**(8)**,** 816–24.

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behav Brain Res* **125**(1–2)**,** 279–84.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J., and Meyerson, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A* **98**(24)**,** 13790–5.

Bianchi, F., Nuciforo, P., Vecchi, M., Bernard, L., Tizzoni, L., Marchetti, A., Buttitta, F., Felicioni, L., Nicassio, F., and Di Fiore, P. P. (2007). Survival prediction of stage I lung adenocarcinomas by expression of 10 genes. *J Clin Invest* **117**(11)**,** 3436–44.

Brambilla, E., Travis, W. D., Colby, T. V., Corrin, B., and Shimosato, Y. (2001). The new World Health Organization classification of lung tumours. *Eur Respir J* **18**(6)**,** 1059–68.

Brazma, A., and Vilo, J. (2000). Gene expression data analysis. *FEBS Lett* **480**(1)**,** 17–24.

Broderick, P., Wang, Y., Vijayakrishnan, J., Matakidou, A., Spitz, M. R., Eisen, T., Amos, C. I., and Houlston, R. S. (2009). Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study. *Cancer Res* **69**(16)**,** 6633–41.

Cardon, L. R., and Bell, J. I. (2001). Association study designs for complex diseases. *Nat Rev Genet* **2**(2)**,** 91–9.

Cheng, Y. W., Chiou, H. L., Sheu, G. T., Hsieh, L. L., Chen, J. T., Chen, C. Y., Su, J. M., and Lee, H. (2001). The association of human papillomavirus 16/18 infection with lung cancer among nonsmoking Taiwanese women. *Cancer Res* **61**(7)**,** 2799–803.

Chung, C. C., Magalhaes, W. C., Gonzalez-Bosquet, J., and Chanock, S. J. (2010). Genome-wide association studies in cancer--current and future directions. *Carcinogenesis* **31**(1)**,** 111–20.

Cox, G., Jones, J. L., Andi, A., Waller, D. A., and O'Byrne, K. J. (2001). A biological staging model for operable non-small cell lung cancer. *Thorax* **56**(7)**,** 561–6.

Crick, F. (1970). Central dogma of molecular biology. *Nature* **227**(5258)**,** 561–3.

Czene, K., Lichtenstein, P., and Hemminki, K. (2002). Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int J Cancer* **99**(2)**,** 260–6.

D'Amico, T. A. (2008). Molecular biologic staging of lung cancer. *Ann Thorac Surg* **85**(2)**,** S737–42.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* **14**(4)**,** 457–60.

Do, J. H., and Choi, D. K. (2006). Normalization of microarray data: single-labeled and dual-labeled arrays. *Mol Cells* **22**(3)**,** 254–61.

Douillard, J. Y. (2010). Adjuvant chemotherapy for non-small-cell lung cancer: it does not always fade with time. *J Clin Oncol* **28**(1)**,** 3–5.

Douillard, J. Y., Rosell, R., De Lena, M., Carpagnano, F., Ramlau, R., Gonzales-Larriba, J. L., Grodzki, T., Pereira, J. R., Le Groumellec, A., Lorusso, V., Clary, C., Torres, A. J., Dahabreh, J., Souquet, P. J., Astudillo, J., Fournel, P., Artal-Cortes, A., Jassem, J., Koubkova, L., His, P., Riggi, M., and Hurteloup, P. (2006). Adjuvant vinorelbine plus cisplatin versus observation in patients with completely resected stage IB-IIIA non-small-cell lung cancer (Adjuvant Navelbine International Trialist Association [ANITA]): a randomised controlled trial. *Lancet Oncol* **7**(9)**,** 719–27.

Dracheva, T., Philip, R., Xiao, W., Gee, A. G., McCarthy, J., Yang, P., Wang, Y., Dong, G., Yang, H., and Jen, J. (2007). Distinguishing lung tumours from normal lung based on a small set of genes. *Lung Cancer* **55**(2)**,** 157–64.

Dudoit, S., and Speed, T. P. (2000). A score test for the linkage analysis of qualitative and quantitative traits based on identity by descent data from sib-pairs. *Biostatistics* **1**(1)**,** 1–26.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**(25)**,** 14863–8.

Ezzati, M., and Lopez, A. D. (2003). Estimates of global mortality attributable to smoking in 2000. *Lancet* **362**(9387)**,** 847–52.

Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C., and Parkin, D. M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer*.

Ferte, C., Andre, F., and Soria, J. C. (2010). Molecular circuits of solid tumors: prognostic and predictive tools for bedside use. *Nat Rev Clin Oncol* **7**(7)**,** 367–80.

Franklin, W. A. (2000). Diagnosis of lung cancer: pathology of invasive and preinvasive neoplasia. *Chest* **117**(4 Suppl 1)**,** 80S–89S.

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Sun, W., Wang, H., Wang, Y., Xiong, X., Xu, L., Waye, M. M., Tsui, S. K., Xue, H., Wong, J. T., Galver, L. M., Fan, J. B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J. F., Phillips, M. S., Roumy, S., Sallee, C., Verner, A., Hudson, T. J., Kwok, P. Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L. C., Mak, W., Song, Y. Q., Tam, P. K., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., de Bakker, P. I., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda, E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Yakub, I., Birren, B. W., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archeveque, P., Bellemare, G., Saeki, K., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R., and Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**(7164)**,** 851–61.

Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nat Rev Genet* **10**(4)**,** 241–51.

Galvan, A., Ioannidis, J. P., and Dragani, T. A. (2010). Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. *Trends Genet* **26**(3)**,** 132–41.

Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G. D., Perou, C. M., Whyte, R. I., Altman,

R. B., Brown, P. O., Botstein, D., and Petersen, I. (2001). Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci U S A* **98**(24)**,** 13784–9.

Guo, N. L., Wan, Y. W., Tosun, K., Lin, H., Msiska, Z., Flynn, D. C., Remick, S. C., Vallyathan, V., Dowlati, A., Shi, X., Castranova, V., Beer, D. G., and Qian, Y. (2008). Confirmation of gene expression-based prediction of survival in non-small cell lung cancer. *Clin Cancer Res* **14**(24)**,** 8213–20.

Haberkorn, U., and Schoenberg, S. O. (2001). Imaging of lung cancer with CT, MRT and PET. *Lung Cancer* **34 Suppl 3,** S13–23.

Hack, C. J. (2004). Integrated transcriptome and proteome data: the challenges ahead. *Brief Funct Genomic Proteomic* **3**(3)**,** 212–9.

Hecht, S. S. (1999). Tobacco smoke carcinogens and lung cancer. *J Natl Cancer Inst* **91**(14)**,** 1194–210.

Herrero, J., Valencia, A., and Dopazo, J. (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics* **17**(2)**,** 126–36.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**(23)**,** 9362–7.

Horan, M. P. (2009). Application of serial analysis of gene expression to the study of human genetic disease. *Hum Genet* **126**(5)**,** 605–14.

Hou, J., Aerts, J., den Hamer, B., van Ijcken, W., den Bakker, M., Riegman, P., van der Leest, C., van der Spek, P., Foekens, J. A., Hoogsteden, H. C., Grosveld, F., and Philipsen, S. (2010). Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS One* **5**(4)**,** e10312.

Hung, R. J., McKay, J. D., Gaborieau, V., Boffetta, P., Hashibe, M., Zaridze, D., Mukeria, A., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., Fabianova, E., Mates, D., Bencko, V., Foretova, L., Janout, V., Chen, C., Goodman, G., Field, J. K., Liloglou, T., Xinarianos, G., Cassidy, A., McLaughlin, J., Liu, G., Narod, S., Krokan, H. E., Skorpen, F., Elvestad, M. B., Hveem, K., Vatten, L., Linseisen, J., Clavel-Chapelon, F., Vineis, P., Bueno-de-Mesquita, H. B., Lund, E., Martinez, C., Bingham, S., Rasmuson, T., Hainaut, P., Riboli, E., Ahrens, W., Benhamou, S., Lagiou, P., Trichopoulos, D., Holcatova, I., Merletti, F., Kjaerheim, K., Agudo, A., Macfarlane, G., Talamini, R., Simonato, L., Lowry, R., Conway, D. I., Znaor, A., Healy, C., Zelenika, D., Boland, A., Delepine, M., Foglio, M., Lechner, D., Matsuda, F., Blanche, H., Gut, I., Heath, S., Lathrop, M., and Brennan, P. (2008). A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452**(7187)**,** 633–7.

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA Cancer J Clin* **61**(2)**,** 69–90.

Kikuchi, T., Daigo, Y., Ishikawa, N., Katagiri, T., Tsunoda, T., Yoshida, S., and Nakamura, Y. (2006). Expression profiles of metastatic brain tumor from lung adenocarcinomas on cDNA microarray. *Int J Oncol* **28**(4)**,** 799–805.

Kikuchi, T., Daigo, Y., Katagiri, T., Tsunoda, T., Okada, K., Kakiuchi, S., Zembutsu, H., Furukawa, Y., Kawamura, M., Kobayashi, K., Imai, K., and Nakamura, Y. (2003). Expression profiles of non-small cell lung cancers on cDNA microarrays: identification of genes for prediction of lymph-node metastasis and sensitivity to anti-cancer drugs. *Oncogene* **22**(14)**,** 2192–205.

Klein, F., Amin Kotb, W. F., and Petersen, I. (2009). Incidence of human papilloma virus in lung cancer. *Lung Cancer* **65**(1)**,** 13–8.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., and Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* **308**(5720)**,** 385–9.

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y., and Carninci, P. (2006). CAGE: cap analysis of gene expression. *Nat Methods* **3**(3)**,** 211–22.

Koshiol, J., Rotunno, M., Gillison, M. L., Van Doorn, L. J., Chaturvedi, A. K., Tarantini, L., Song, H., Quint, W. G., Struijk, L., Goldstein, A. M., Hildesheim, A., Taylor, P. R., Wacholder, S., Bertazzi, P. A., Landi, M. T., and Caporaso, N. E. (2011). Assessment of human papillomavirus in lung tumor tissue. *J Natl Cancer Inst* **103**(6)**,** 501–7.

Kuner, R., Muley, T., Meister, M., Ruschhaupt, M., Buness, A., Xu, E. C., Schnabel, P., Warth, A., Poustka, A., Sultmann, H., and Hoffmann, H. (2009). Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung Cancer* **63**(1)**,** 32–8.

Lacroix, L., Commo, F., and Soria, J. C. (2008). Gene expression profiling of non-small-cell lung cancer. *Expert Rev Mol Diagn* **8**(2)**,** 167–78.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blocker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., and Chen, Y. J. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**(6822)**,** 860–921.

Landi, M. T., Chatterjee, N., Yu, K., Goldin, L. R., Goldstein, A. M., Rotunno, M., Mirabello, L., Jacobs, K., Wheeler, W., Yeager, M., Bergen, A. W., Li, Q., Consonni, D., Pesatori, A. C., Wacholder, S., Thun, M., Diver, R., Oken, M., Virtamo, J., Albanes, D., Wang, Z., Burdette, L., Doheny, K. F., Pugh, E. W., Laurie, C., Brennan, P., Hung, R., Gaborieau, V., McKay, J. D., Lathrop, M., McLaughlin, J., Wang, Y., Tsao, M. S., Spitz, M. R., Krokan, H., Vatten, L., Skorpen, F., Arnesen, E., Benhamou, S., Bouchard, C., Metsapalu, A., Vooder, T., Nelis, M., Valk, K., Field, J. K., Chen, C., Goodman, G., Sulem, P., Thorleifsson, G., Rafnar, T., Eisen, T., Sauter, W., Rosenberger, A., Bickeboller, H., Risch, A., Chang-Claude, J., Wichmann, H. E., Stefansson, K., Houlston, R., Amos, C. I., Fraumeni, J. F., Jr., Savage, S. A., Bertazzi, P. A., Tucker, M. A., Chanock, S., and Caporaso, N. E. (2009). A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet* **85**(5)**,** 679–91.

Lee, E. S., Son, D. S., Kim, S. H., Lee, J., Jo, J., Han, J., Kim, H., Lee, H. J., Choi, H. Y., Jung, Y., Park, M., Lim, Y. S., Kim, K., Shim, Y., Kim, B. C., Lee, K., Huh, N., Ko, C., Park, K., Lee, J. W., Choi, Y. S., and Kim, J. (2008). Prediction of recurrence-free survival in postoperative non-small cell lung cancer patients by using an integrated model of clinical information and gene expression. *Clin Cancer Res* **14**(22)**,** 7397–404.

Li, M., Li, C., and Guan, W. (2008). Evaluation of coverage variation of SNP chips for genome-wide association studies. *Eur J Hum Genet* **16**(5)**,** 635–43.

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**(4)**,** 402–8.

Lorenzo Bermejo, J., and Hemminki, K. (2005). Familial lung cancer and aggregation of smoking habits: a simulation of the effect of shared environmental factors on the familial risk of cancer. *Cancer Epidemiol Biomarkers Prev* **14**(7)**,** 1738–40.

Lu, Y., Lemon, W., Liu, P. Y., Yi, Y., Morrison, C., Yang, P., Sun, Z., Szoke, J., Gerald, W. L., Watson, M., Govindan, R., and You, M. (2006). A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. *PLoS Med* **3**(12)**,** e467.

Malvezzi, M., Arfe, A., Bertuccio, P., Levi, F., La Vecchia, C., and Negri, E. (2011). European cancer mortality predictions for the year 2011. *Ann Oncol.*

Matakidou, A., Eisen, T., and Houlston, R. S. (2005). Systematic review of the relationship between family history and lung cancer risk. *Br J Cancer* **93**(7)**,** 825–33.

Mayne, S. T., Buenconsejo, J., and Janerich, D. T. (1999). Familial cancer history and lung cancer risk in United States nonsmoking men and women. *Cancer Epidemiol Biomarkers Prev* **8**(12)**,** 1065–9.

McKay, J. D., Hung, R. J., Gaborieau, V., Boffetta, P., Chabrier, A., Byrnes, G., Zaridze, D., Mukeria, A., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., Fabianova, E., Mates, D., Bencko, V., Foretova, L., Janout, V., McLaughlin, J., Shepherd, F., Montpetit, A., Narod, S., Krokan, H. E., Skorpen, F., Elvestad, M. B., Vatten, L., Njolstad, I., Axelsson, T., Chen, C., Goodman, G., Barnett, M., Loomis, M. M., Lubinski, J., Matyjasik, J., Lener, M., Oszutowska, D., Field, J., Liloglou, T., Xinarianos, G., Cassidy, A., Vineis, P., Clavel-Chapelon, F., Palli, D., Tumino, R., Krogh, V., Panico, S., Gonzalez, C. A., Ramon Quiros, J., Martinez, C., Navarro, C., Ardanaz, E., Larranaga, N., Kham, K. T., Key, T., Bueno-de-Mesquita, H. B., Peeters, P. H., Trichopoulou, A., Linseisen, J., Boeing, H., Hallmans, G., Overvad, K., Tjonneland, A., Kumle, M., Riboli, E., Zelenika, D., Boland, A., Delepine, M., Foglio, M., Lechner, D., Matsuda, F., Blanche, H., Gut, I., Heath, S., Lathrop, M., and Brennan, P. (2008). Lung cancer susceptibility locus at 5p15.33. *Nat Genet* **40**(12)**,** 1404–6.

Mitsuuchi, Y., and Testa, J. R. (2002). Cytogenetics and molecular genetics of lung cancer. *Am J Med Genet* **115**(3)**,** 183–8.

Mollberg, N., Surati, M., Demchuk, C., Fathi, R., Salama, A. K., Husain, A. N., Hensing, T., and Salgia, R. (2011). Mind-mapping for lung cancer: Towards a personalized therapeutics approach. *Adv Ther* **28**(3)**,** 173–94.

Moriya, Y., Iyoda, A., Kasai, Y., Sugimoto, T., Hashida, J., Nimura, Y., Kato, M., Takiguchi, M., Fujisawa, T., Seki, N., and Yoshino, I. (2009). Prediction of lymph node metastasis by gene expression profiling in patients with primary resected lung cancer. *Lung Cancer* **64**(1)**,** 86–91.

Mountain, C. F. (2000). The international system for staging lung cancer. *Semin Surg Oncol* **18**(2)**,** 106–15.

Nacht, M., Dracheva, T., Gao, Y., Fujii, T., Chen, Y., Player, A., Akmaev, V., Cook, B., Dufault, M., Zhang, M., Zhang, W., Guo, M., Curran, J., Han, S., Sidransky, D., Buetow, K., Madden, S. L., and Jen, J. (2001). Molecular characteristics of non-small cell lung cancer. *Proc Natl Acad Sci U S A* **98**(26)**,** 15203–8.

Nathoo, N., Chahlavi, A., Barnett, G. H., and Toms, S. A. (2005). Pathobiology of brain metastases. *J Clin Pathol* **58**(3)**,** 237–42.

Olaussen, K. A., Dunant, A., Fouret, P., Brambilla, E., Andre, F., Haddad, V., Taranchon, E., Filipits, M., Pirker, R., Popper, H. H., Stahel, R., Sabatier, L., Pignon, J. P., Tursz, T., Le Chevalier, T., and Soria, J. C. (2006). DNA repair by ERCC1 in non-small-cell lung cancer and cisplatin-based adjuvant chemotherapy. *N Engl J Med* **355**(10)**,** 983–91.

Ozsolak, F., and Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12**(2)**,** 87–98.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**(8)**,** 904–9.

Proctor, R. N. (2004). The global smoking epidemic: a history and status report. *Clin Lung Cancer* **5**(6)**,** 371–6.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**(3)**,** 559–75.

Rafnar, T., Sulem, P., Besenbacher, S., Gudbjartsson, D. F., Zanon, C., Gudmundsson, J., Stacey, S. N., Kostic, J. P., Thorgeirsson, T. E., Thorleifsson, G., Bjarnason, H., Skuladottir, H., Gudbjartsson, T., Isaksson, H. J., Isla, D., Murillo, L., Garcia-Prats, M. D., Panadero, A., Aben, K. K., Vermeulen, S. H., van der Heijden, H. F., Feser, W. J., Miller, Y. E., Bunn, P. A., Kong, A., Wolf, H. J., Franklin, W. A., Mayordomo, J. I., Kiemeney, L. A., Jonsson, S., Thorsteinsdottir, U., and Stefansson, K. (2011). Genome-wide significant association between a sequence variant at 15q15.2 and lung cancer risk. *Cancer Res* **71**(4)**,** 1356–61.

Rami-Porta, R., Chansky, K., and Goldstraw, P. (2009). Updated lung cancer staging system. *Future Oncol* **5**(10)**,** 1545–53.

Ramoni, M. F., Sebastiani, P., and Kohane, I. S. (2002). Cluster analysis of gene expression dynamics. *Proc Natl Acad Sci U S A* **99**(14)**,** 9121–6.

Ransohoff, D. F. (2007). How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. *J Clin Epidemiol* **60**(12)**,** 1205–19.

Raponi, M., Zhang, Y., Yu, J., Chen, G., Lee, G., Taylor, J. M., Macdonald, J., Thomas, D., Moskaluk, C., Wang, Y., and Beer, D. G. (2006). Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res* **66**(15)**,** 7466–72.

Raychaudhuri, S., Stuart, J. M., and Altman, R. B. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput***,** 455–66.

Reinartz, J., Bruyns, E., Lin, J. Z., Burcham, T., Brenner, S., Bowen, B., Kramer, M., and Woychik, R. (2002). Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Brief Funct Genomic Proteomic* **1**(1)**,** 95–104.

Rivera, M. P., Detterbeck, F., and Mehta, A. C. (2003). Diagnosis of lung cancer: the guidelines. *Chest* **123**(1 Suppl)**,** 129S–136S.

Rosell, R., Felip, E., Garcia-Campelo, R., and Balana, C. (2004). The biology of non-small-cell lung cancer: identifying new targets for rational therapy. *Lung Cancer* **46**(2)**,** 135–48.

Saccone, S. F., Hinrichs, A. L., Saccone, N. L., Chase, G. A., Konvicka, K., Madden, P. A., Breslau, N., Johnson, E. O., Hatsukami, D., Pomerleau, O., Swan, G. E., Goate, A. M., Rutter, J., Bertelsen, S., Fox, L., Fugman, D., Martin, N. G., Montgomery, G. W., Wang, J. C., Ballinger, D. G., Rice, J. P., and Bierut, L. J. (2007). Cholinergic nicotinic receptor genes implicated in a nicotine dependence association study targeting 348 candidate genes with 3713 SNPs. *Hum Mol Genet* **16**(1)**,** 36–49.

Sasaki, T., Gan, E. C., Wakeham, A., Kornbluth, S., Mak, T. W., and Okada, H. (2007). HLA-B-associated transcript 3 (Bat3)/Scythe is essential for p300-mediated acetylation of p53. *Genes Dev* **21**(7)**,** 848–61.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**(5235)**,** 467–70.

Schmid, R., Baum, P., Ittrich, C., Fundel-Clemens, K., Huber, W., Brors, B., Eils, R., Weith, A., Mennerich, D., and Quast, K. (2010). Comparison of normalization methods for Illumina BeadChip HumanHT–12 v3. *BMC Genomics* **11,** 349.

Schwartz, A. G., Prysak, G. M., Bock, C. H., and Cote, M. L. (2007). The molecular epidemiology of lung cancer. *Carcinogenesis* **28**(3)**,** 507–18.

Sorensen, J. B., Hirsch, F. R., Gazdar, A., and Olsen, J. E. (1993). Interobserver variability in histopathologic subtyping and grading of pulmonary adenocarcinoma. *Cancer* **71**(10)**,** 2971–6.

Subramanian, A., Harris, A., Piggott, K., Shieff, C., and Bradford, R. (2002). Metastasis to and from the central nervous system--the 'relatively protected site'. *Lancet Oncol* **3**(8)**,** 498–507.

Subramanian, J., and Simon, R. (2010). Gene expression-based prognostic signatures in lung cancer: ready for clinical use? *J Natl Cancer Inst* **102**(7)**,** 464–74.

Zhao, Z., Timofeev, N., Hartley, S. W., Chui, D. H., Fucharoen, S., Perls, T. T., Steinberg, M. H., Baldwin, C. T., and Sebastiani, P. (2008). Imputation of missing genotypes: an empirical evaluation of IMPUTE. *BMC Genet* **9,** 85.

Zhu, C. Q., Ding, K., Strumpf, D., Weir, B. A., Meyerson, M., Pennell, N., Thomas, R. K., Naoki, K., Ladd-Acosta, C., Liu, N., Pintilie, M., Der, S., Seymour, L., Jurisica, I., Shepherd, F. A., and Tsao, M. S. (2010a). Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. *J Clin Oncol* **28**(29)**,** 4417–24.

Zhu, C. Q., Strumpf, D., Li, C. Y., Li, Q., Liu, N., Der, S., Shepherd, F. A., Tsao, M. S., and Jurisica, I. (2010b). Prognostic gene expression signature for squamous cell carcinoma of lung. *Clin Cancer Res* **16**(20)**,** 5038–47.

Takada, M., Tada, M., Tamoto, E., Kawakami, A., Murakawa, K., Shindoh, G., Teramoto, K., Matsunaga, A., Komuro, K., Kanai, M., Fujiwara, Y., Shirata, K., Nishimura, N., Miyamoto, M., Okushiba, S., Kondo, S., Hamada, J., Katoh, H., Yoshiki, T., and Moriuchi, T. (2004). Prediction of lymph node metastasis by analysis of gene expression profiles in non-small cell lung cancer. *J Surg Res* **122**(1)**,** 61–9.

Tanoue, L. T. (2008). Staging of non-small cell lung cancer. *Semin Respir Crit Care Med* **29**(3)**,** 248–60.

Tanoue, L. T., and Detterbeck, F. C. (2009). New TNM classification for non-small-cell lung cancer. *Expert Rev Anticancer Ther* **9**(4)**,** 413–23.

Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., Manolescu, A., Thorleifsson, G., Stefansson, H., Ingason, A., Stacey, S. N., Bergthorsson, J. T., Thorlacius, S., Gudmundsson, J., Jonsson, T., Jakobsdottir, M., Saemundsdottir, J., Olafsdottir, O., Gudmundsson, L. J., Bjornsdottir, G., Kristjansson, K., Skuladottir, H., Isaksson, H. J., Gudbjartsson, T., Jones, G. T., Mueller, T., Gottsater, A., Flex, A., Aben, K. K., de Vegt, F., Mulders, P. F., Isla, D., Vidal, M. J., Asin, L., Saez, B., Murillo, L., Blondal, T., Kolbeinsson, H., Stefansson, J. G., Hansdottir, I., Runarsdottir, V., Pola, R., Lindblad, B., van Rij, A. M., Dieplinger, B., Haltmayer, M., Mayordomo, J. I., Kiemeney, L. A., Matthiasson, S. E., Oskarsson, H., Tyrfingsson, T., Gudbjartsson, D. F., Gulcher, J. R., Jonsson, S., Thorsteinsdottir, U., Kong, A., and Stefansson, K. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452**(7187)**,** 638–42.

Tokuhata, G. K., and Lilienfeld, A. M. (1963). Familial aggregation of lung cancer in humans. *J Natl Cancer Inst* **30,** 289–312.

Tomida, S., Koshikawa, K., Yatabe, Y., Harano, T., Ogura, N., Mitsudomi, T., Some, M., Yanagisawa, K., Takahashi, T., and Osada, H. (2004). Gene expression-based, individualized outcome prediction for surgically treated lung cancer patients. *Oncogene* **23**(31)**,** 5360–70.

Travis, W. D., Travis, L. B., and Devesa, S. S. (1995). Lung cancer. *Cancer* **75**(1 Suppl)**,** 191–202.

Tse, L. A., Yu, I. S., Au, J. S., Qiu, H., and Wang, X. R. (2011). Silica dust, diesel exhaust, and painting work are the significant occupational risk factors for lung cancer in nonsmoking Chinese men. *Br J Cancer* **104**(1)**,** 208–13.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**(9)**,** 5116–21.

Wang, Y., Broderick, P., Webb, E., Wu, X., Vijayakrishnan, J., Matakidou, A., Qureshi, M., Dong, Q., Gu, X., Chen, W. V., Spitz, M. R., Eisen, T., Amos, C. I., and Houlston, R. S. (2008). Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet* **40**(12)**,** 1407–9.

Wang, Y. C., Hsu, H. S., Chen, T. P., and Chen, J. T. (2006). Molecular diagnostic markers for lung cancer in sputum and plasma. *Ann N Y Acad Sci* **1075,** 179–84.

Varghese, J. S., and Easton, D. F. (2010). Genome-wide association studies in common cancers--what have we learnt? *Curr Opin Genet Dev* **20**(3)**,** 201–9.

Wigle, D. A., Jurisica, I., Radulovich, N., Pintilie, M., Rossant, J., Liu, N., Lu, C., Woodgett, J., Seiden, I., Johnston, M., Keshavjee, S., Darling, G., Winton, T., Breitkreutz, B. J., Jorgenson, P., Tyers, M., Shepherd, F. A., and Tsao, M. S. (2002). Molecular profiling of non-small cell lung cancer and correlation with disease-free survival. *Cancer Res* **62**(11)**,** 3005–8.

Winton, T., Livingston, R., Johnson, D., Rigas, J., Johnston, M., Butts, C., Cormier, Y., Goss, G., Inculet, R., Vallieres, E., Fry, W., Bethune, D., Ayoub, J., Ding, K., Seymour, L., Graham, B., Tsao, M. S., Gandara, D., Kesler, K., Demmy, T., and Shepherd, F. (2005). Vinorelbine plus cisplatin vs. observation in resected non-small-cell lung cancer. *N Engl J Med* **352**(25)**,** 2589–97.

Vollmer, E., Schultz, H., Stellmacher, F., Kahler, D., Abdullah, M., Galle, J., Lang, D. S., and Goldmann, T. (2010). Tumors in the lung – morphologic features and the challenge of integrating biomarker signatures into diagnostics. *Rom J Morphol Embryol* **51**(4)**,** 607–14.

Xie, Y., and Minna, J. D. (2010). Non-small-cell lung cancer mRNA expression signature predicting response to adjuvant chemotherapy. *J Clin Oncol* **28**(29)**,** 4404–7.

Yoon, K. A., Park, J. H., Han, J., Park, S., Lee, G. K., Han, J. Y., Zo, J. I., Kim, J., Lee, J. E., Takahashi, A., Kubo, M., Nakamura, Y., and Lee, J. S. (2010). A genome-wide association study reveals susceptibility variants for non-small cell lung cancer in the Korean population. *Hum Mol Genet* **19**(24)**,** 4948–54.

ACKNOWLEDGEMENTS

First, I would like to express my gratitude to my supervisor, Prof. Andres Metspalu, for giving me the opportunity to do research on a topic I was interested in, and for his encouragement to go on and beyond. Secondly, I would like to stress my gratefulness to Dr. Tõnu Vooder who collected the samples that were the foundation for the studies of this thesis. Thank you Tõnu also for the fruitful discussions on the deep nature of cancer, medicine, and life. I wish to thank Raivo Kolde for bioinformatical support and Dr. Lili Milani for her constructive suggestions on how to improve the manuscript of this thesis. I would also like to thank all of the co-authors of the papers on which this thesis is based, and all of the people on our lung cancer team and in the Department of Biotechnology. I also want to acknowledge the Competence Centre on Reproductive Medicine and Biology that gave me time to compile this thesis.

My special thanks will go to Triinu Temberg. Thank you, Triinu for filling my time of studies with pleasant friendship and kind motivation. Last, but not least, I would like to express my deep gratitude to my parents who gave me the possibility to educate myself.

This study was financially supported by Targeted Financing from the Estonian Ministry of Education and Research (SF0180142s08), an Estonian Science Foundation grant ETF6465, an EU FP7 grant ECOGENE (#205419, EBC), and by the EU via the European Regional Development Fund grant to the Centre of Excellence in Genomics, Estonian Biocentre, and University of Tartu.

PUBLICATIONS

CURRICULUM VITAE

Kristjan Välk

Date and Place of birth:	30.08.1980, Tartu, Estonia
Address:	Department of Biotechnology, Institute of Molecular and Cell Biology, University of Tartu, Riia Street 23, 51010, Tartu, Estonia
Phone:	+372 5341 2722
E-mail:	kristjan.valk@ut.ee

Education

1988–1999	Tartu Tamme Gymnasium (Class of Environmental Technology)
1999–2003	*B.Sc.* in Gene Technology, Chair of Biotechnology, IMCB, University of Tartu
2003–2005	*M.Sc.* in Gene Technology, Chair of Biotechnology, IMCB, University of Tartu
2005–2011	*Ph.D.* Student in Gene Technology, Chair of Biotechnology, IMCB, University of Tartu

Professional Employment

2003–2009	Estonian Biocentre, University of Tartu, Scientist
2007–2009	Asper Biotech Ltd., Head of the Division of Research and Development
2009–present	Competence Centre on Reproductive Medicine and Biology Ltd. (CCRMB), Scientist

Scientific Work

In recent years I have been actively involved in the preparation of positively assessed grant applications for the generation of a human hepatitis C transgenic mouse model (KPA Scientific) and a colorectal cancer genetic testing portfolio (Asper Biotech). My current work at the CCRMB involves phage display experiments designed to discover new biomarkers for various infertility-associated diseases and conditions. My scientific interests include personalized medicine, pharmacogenomics, translational genetics, and biomarker discovery. During my doctoral studies, I have mainly focused on the application of WGAS and gene expression profiling to NSCLC.

List of Publications

I. Landi, M. T., Chatterjee, N., Yu, K., Goldin, L. R., Goldstein, A. M., Rotunno, M., Mirabello, L., Jacobs, K., Wheeler, W., Yeager, M., Bergen, A. W., Li, Q., Consonni, D., Pesatori, A. C., Wacholder, S., Thun, M., Diver, R., Oken, M., Virtamo, J., Albanes, D., Wang, Z., Burdette, L., Doheny, K. F., Pugh, E. W., Laurie, C., Brennan, P., Hung, R., Gaborieau, V., McKay, J. D., Lathrop, M., McLaughlin, J., Wang, Y., Tsao, M. S., Spitz, M. R., Krokan, H., Vatten, L., Skorpen, F., Arnesen, E., Benhamou, S., Bouchard, C., Metspalu, A., Vooder, T., Nelis, M., **Välk, K.**, Field, J. K., Chen, C., Goodman, G., Sulem, P., Thorleifsson, G., Rafnar, T., Eisen, T., Sauter, W., Rosenberger, A., Bickeboller, H., Risch, A., Chang-Claude, J., Wichmann, H. E., Stefansson, K., Houlston, R., Amos, C. I., Fraumeni, J. F., Jr., Savage, S. A., Bertazzi, P. A., Tucker, M. A., Chanock, S., and Caporaso, N. E. (2009). **A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma.** Am J Hum Genet 85(5), 679–91.

II. Vooder, T.*, **Välk, K.***, Kolde, R., Roosipuu, R., Vilo, J., and Metspalu, A. (2010). **Gene Expression-Based Approaches in Differentiation of Metastases and Second Primary Tumour.** Case Rep Oncol 3(2), 255–261.

III. **Välk, K.** *, Vooder, T. *, Kolde, R., Reintam, M.A., Petzold, C., Vilo, J., and Metspalu, A. (2010). **Gene Expression Profiles of Non-Small Cell Lung Cancer: Survival Prediction and New Biomarkers.** . Oncology 2010;79:283–292.

IV. McKay, J. D., Truong, T., Gaborieau, V., Chabrier, A., Chuang, S. C., Byrnes, G., Zaridze, D., Shangina, O., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., Fabianova, E., Bucur, A., Bencko, V., Holcatova, I., Janout, V., Foretova, L., Lagiou, P., Trichopoulos, D., Benhamou, S., Bouchardy, C., Ahrens, W., Merletti, F., Richiardi, L., Talamini, R., Barzan, L., Kjaerheim, K., Macfarlane, G. J., Macfarlane, T. V., Simonato, L., Canova, C., Agudo, A., Castellsague, X., Lowry, R., Conway, D. I., McKinney, P. A., Healy, C. M., Toner, M. E., Znaor, A., Curado, M. P., Koifman, S., Menezes, A., Wunsch-Filho, V., Neto, J. E., Garrote, L. F., Boccia, S., Cadoni, G., Arzani, D., Olshan, A. F., Weissler, M. C., Funkhouser, W. K., Luo, J., Lubinski, J., Trubicka, J., Lener, M., Oszutowska, D., Schwartz, S. M., Chen, C., Fish, S., Doody, D. R., Muscat, J. E., Lazarus, P., Gallagher, C. J., Chang, S. C., Zhang, Z. F., Wei, Q., Sturgis, E. M., Wang, L. E., Franceschi, S., Herrero, R., Kelsey, K. T., McClean, M. D., Marsit, C. J., Nelson, H. H., Romkes, M., Buch, S., Nukui, T., Zhong, S., Lacko, M., Manni, J. J., Peters, W. H., Hung, R. J., McLaughlin, J., Vatten, L., Njolstad, I., Goodman, G. E., Field, J. K., Liloglou, T., Vineis, P., Clavel-Chapelon, F., Palli, D., Tumino, R., Krogh, V., Panico, S., Gonzalez, C. A., Quiros, J. R., Martinez, C., Navarro, C., Ardanaz, E., Larranaga, N., Khaw, K. T., Key, T., Bueno-de-Mesquita, H. B., Peeters, P. H., Trichopoulou, A., Linseisen, J., Boeing, H., Hallmans, G., Overvad, K., Tjonneland, A., Kumle, M., Riboli, E., **Välk, K**., Voodern, T., Metspalu, A., Zelenika, D., Boland, A., Delepine, M., Foglio, M., Lechner, D., Blanche, H., Gut, I. G., Galan, P., Heath, S., Hashibe, M., Hayes, R. B., Boffetta, P., Lathrop, M., and Brennan, P. (2011). **A Genome-Wide Association Study of Upper Aerodigestive Tract Cancers Conducted within the INHANCE Consortium**. *PLoS Genet* 7(3)**,** e1001333.

V. Urgard, E., Vooder, T., Võsa, U., **Välk, K.**, Liu, M., Luo, C., Hoti, F., Roosipuu, R., Annilo, T., Laine, J., Frenz, M.C., Zhang, L., Metspalu, A. (2011). **Metagenes associated with survival in NSCLC**. Cancer Informatics 2011:10 175–183.

VI. Võsa, U., Vooder, T., Kolde, R., Fischer, K., **Välk, K.**, Tõnisson, N., Roosipuu, R., Vilo, J., Metspalu, A., Annilo, T., (2011). **Identification of MiR-374a as a Prognostic Marker for Survival in Patients with Early-Stage Nonsmall Cell Lung Cancer**. Genes Chromosomes and Cancer (accepted).

* These authors contributed equally to this work.

Referee

Priit Palta B.Sc. thesis “Statistical methods for DNA copy-number detection”, 2007, University of Tartu, Institute of Molecular and Cell Biology, Department of Bioinformatics.

Courses

I. 23.08.04–25.08.04 University of Tartu, Centre of Molecular and Clinical Medicine „DNA Microarray Applications in Biomedical Studies”

II. 24.01.05–04.02.05 Visiting Scientist at the University of Helsinki Biomedicum Biochip Centre

III. 16.04.05–24.04.05 Wellcome Trust advanced course „Microarrays and Transcriptome” (Hinxton)

IV. 22.01.07–08.02.07 Federation of European Laboratory Animal Science Associations Course (FELASA, C category)

V. 19.03.07–22.03.07 CSC Turku Biotechnology Centre “DNA Microarray Data Analysis using R/Bioconductor”

VI. 18.12.07–20.12.07 Sean McCarthy course on "How to write a competitive proposal for FP7" and "How to negotiate and administer FP7 grant agreements".

Oral Presentations at International Conferences

I. Understanding the Genome: Microarray applications from the Biology to the Clinics, 11–13 November 2005, Genoa, Italy, "Preliminary data of Non-Small Cell Lung Cancer gene expression pilot study".

II. 1st International Lung Cancer Conference, 09–12 July 2008 Liverpool, United Kingdom "NSCLC gene expression study in Estonia".

Conferences Attended with Poster Presentation

I. European Human Genetic Conference 16.06.07–19.06.07 Nice, France.

II. 12th World Conference on Lung Cancer 02.09.07–06.09.07 Seoul, Korea.

III. 1st European Lung Cancer Conference 23.04.08–26.04.07 Geneva, Switzerland.

IV. 1st International Lung Cancer Conference 09.07.08–12.07.08 Liverpool, United Kingdom.

V. 14th World Conference on Lung Cancer 03.07.11–07.07.11 Amsterdam, Netherlands.

ELULOOKIRJELDUS

Kristjan Välk

Sünniaeg ja koht:	30.08.1980 Tartu
Kodakondsus:	Eesti
Address:	Biotehnoloogia õppetool, Tartu Ülikooli Molekulaar- ja Rakubioloogia Instituut, Tartu Ülikool, Riia 23, 51010, Tartu, Eesti
Telefon:	5341 2722
E-mail:	kristjan.valk@ut.ee

Haridus

1988–1999 Tartu Tamme Gümnaasium (keskkonnatehnoloogia klass)

1999–2003 B.Sc. geenitehnoloogia erialal, Biotehnoloogia õppetool, Tartu Ülikooli Molekulaar- ja Rakubioloogia Instituut

2003–2005 M.Sc. geenitehnoloogia erialal, Biotehnoloogia õppetool, Tartu Ülikooli Molekulaar- ja Rakubioloogia Instituut

2005–2011 Ph.D. tudeng geenitehnoloogia erialal, Biotehnoloogia õppetool, Tartu Ülikooli Molekulaar- ja Rakubioloogia Instituut

Keelteoskus: eesti keel, vene keel, inglise keel

Töökogemus

2003–2009	Teadur, Eesti Biokeskus, Tartu Ülikool
2007–2009	Teadus- ja arendusosakonna juht, AS Asper Biotech
2009–	Teadur, Reproduktiivmeditsiini TAK AS

Teaduslik ja arendustegevus

Peamised uurimisvaldkonnad: vähkkasvaja molekulaarne iseloomustamine, personaalne meditsiin, farmakogenoomika.

Publikatsioonide loetelu

I Landi, M. T., Chatterjee, N., Yu, K., Goldin, L. R., Goldstein, A. M., Rotunno, M., Mirabello, L., Jacobs, K., Wheeler, W., Yeager, M., Bergen, A. W., Li, Q., Consonni, D., Pesatori, A. C., Wacholder, S., Thun, M., Diver, R., Oken, M., Virtamo, J., Albanes, D., Wang, Z., Burdette, L., Doheny, K. F., Pugh, E. W., Laurie, C., Brennan, P., Hung, R., Gaborieau, V., McKay, J. D., Lathrop, M., McLaughlin, J., Wang, Y., Tsao, M. S., Spitz, M. R., Krokan, H., Vatten, L., Skorpen, F., Arnesen, E., Benhamou, S., Bouchard, C., Metspalu, A., Vooder, T., Nelis, M., **Välk, K.**, Field, J. K., Chen, C., Goodman, G., Sulem, P., Thorleifsson, G., Rafnar, T., Eisen, T., Sauter, W., Rosenberger, A., Bickeboller, H., Risch, A., Chang-Claude, J., Wichmann, H. E., Stefansson, K., Houlston, R., Amos, C. I., Fraumeni, J. F., Jr., Savage, S. A., Bertazzi, P. A., Tucker, M. A., Chanock, S., and Caporaso, N. E. (2009). A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. Am J Hum Genet 85(5), 679–91.

II Vooder, T.*, **Välk, K.***, Kolde, R., Roosipuu, R., Vilo, J., and Metspalu, A. (2010). Gene Expression-Based Approaches in Differentiation of Metastases and Second Primary Tumour. Case Rep Oncol 3(2), 255–261.

III **Välk, K.** *, Vooder, T. *, Kolde, R., Reintam, M.A., Petzold, C., Vilo, J., and Metspalu, A. (2010). Gene Expression Profiles of Non-Small Cell Lung Cancer: Survival Prediction and New Biomarkers. . Oncology 2010;79:283–292.

IV McKay, J. D., Truong, T., Gaborieau, V., Chabrier, A., Chuang, S. C., Byrnes, G., Zaridze, D., Shangina, O., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., Fabianova, E., Bucur, A., Bencko, V., Holcatova, I., Janout, V., Foretova, L., Lagiou, P., Trichopoulos, D., Benhamou, S., Bouchardy, C., Ahrens, W., Merletti, F., Richiardi, L., Talamini, R., Barzan, L., Kjaerheim, K., Macfarlane, G. J., Macfarlane, T. V., Simonato, L., Canova, C., Agudo, A., Castellsague, X., Lowry, R., Conway, D. I., McKinney, P. A., Healy, C. M., Toner, M. E., Znaor, A., Curado, M. P., Koifman, S., Menezes, A., Wunsch-Filho, V., Neto, J. E., Garrote, L. F., Boccia, S., Cadoni, G., Arzani, D., Olshan, A. F., Weissler, M. C., Funkhouser, W. K., Luo, J., Lubinski, J., Trubicka, J., Lener, M., Oszutowska, D., Schwartz, S. M., Chen, C., Fish, S., Doody, D. R., Muscat, J. E., Lazarus, P., Gallagher, C. J., Chang, S. C., Zhang, Z. F., Wei, Q., Sturgis, E. M., Wang, L. E., Franceschi, S., Herrero, R., Kelsey, K. T., McClean, M. D., Marsit, C. J., Nelson, H. H., Romkes, M., Buch, S., Nukui, T., Zhong, S., Lacko, M., Manni, J. J., Peters, W. H., Hung, R. J., McLaughlin, J., Vatten, L., Njolstad, I., Goodman, G. E., Field, J. K., Liloglou, T., Vineis, P., Clavel-Chapelon, F., Palli, D., Tumino, R., Krogh, V., Panico, S., Gonzalez, C. A., Quiros, J. R., Martinez, C., Navarro, C., Ardanaz, E., Larranaga, N., Khaw, K. T., Key, T., Bueno-de-Mesquita, H. B., Peeters, P. H., Trichopoulou, A., Linseisen, J., Boeing, H., Hallmans, G., Overvad, K., Tjonneland, A., Kumle, M., Riboli, E., **Välk, K.**, Voodern, T., Metspalu, A., Zelenika, D., Boland, A., Delepine, M., Foglio, M., Lechner, D., Blanche, H., Gut, I. G., Galan, P., Heath, S., Hashibe, M., Hayes, R. B., Boffetta, P., Lathrop, M., and Brennan, P. (2011). A Genome-Wide Association Study of Upper Aerodigestive Tract Cancers Conducted within the INHANCE Consortium. PLoS Genet 7(3), e1001333.

V Urgard, E., Vooder, T., Võsa, U., **Välk, K.**, Liu, M., Luo, C., Hoti, F., Roosipuu, R., Annilo, T., Laine, J., Frenz, M.C., Zhang, L., Metspalu, A. (2011). Metagenes associated with survival in NSCLC. Cancer Informatics 2011:10 175–183.

VI Võsa, U., Vooder, T., Kolde, R., Fischer, K., **Välk, K**., Tõnisson, N., Roosipuu, R., Vilo, J., Metspalu, A., Annilo, T., (2011). Identification of MiR-374a as a Prognostic Marker for Survival in Patients with Early-Stage Nonsmall Cell Lung Cancer. Genes Chromosomes and Cancer accepted.

Muu teaduslik organisatsiooniline ja erialane tegevus (konverentside ettekanded, osalemine erialastes seltsides, seadusloome jms.)

Suulised konverentsiettekanded

I. Understanding the Genome: Microarray applications from the Biology to the Clinics, 11–13 November 2005, Genoa, Italy, "Mitteväikerakulise kopsuvähi geeniekspressiooni pilootuuringu esmased tulemused".

II. 1st International Lung Cancer Conference, 09–12 July 2008 Liverpool, United Kingdom, "Mitteväikerakulise kopsuvähi geeniekspressiooni uuring Eestis".

Posterettekandega osaletud konverentsid

I. Euroopa Inimesegeneetika Konverents 16.06.07–19.06.07 Nizza, Prantsusmaa.

II. XII Maailma Kopsuvähi Konverents 02.09.07–06.09.07 Söul, Korea.

III. I Euroopa Kopsuvähi Konverents 23.04.08–26.04.07 Genf, Šveits.

IV. I Rahvusvaheline Kopsuvähi Konverents 09.07.08–12.07.08 Liverpool, UK.

V. XIV Maailma Kopsuvähi Konverents 03.07.11–07.07.11 Amsterdam, Holland.

Erialane enesetäiendus

1. 23.08.04–25.08.04 TÜ Molekulaarse ja Kliinilise Meditsiini keskuse kursus „DNA mikrokiipide kasutus biomeditsiinilistes uuringutes"
2. 24.01.05–04.02.05 Täiendkoolitus Helsingi Ülikooli Biomeedikumi Biokiibi Keskuses
3. 16.04.05–24.04.05 Wellcome Trust edasijõudnute kursus „Mikrokiibid ja transkriptoom" (Hinxton, Inglismaa)
4. 22.01.07–08.02.07 katseloomateaduse kursus (FELASA, C kategooria)
5. 19.03.07–22.03.2007 CSC Turu Biotehnoloogia Keskus "DNA mikrokiipide analüüs R keskkonnas"
6. 18.12.07–20.12.07 Sean McCarthy kursus "Kuidas kirjutada konkurentsivõimelist FP7 granditaotlust" ja "Kuidas pidada FP7 granditaotluse läbirääkimisi ja administreerida projekti".

DISSERTATIONES BIOLOGICAE UNIVERSITATIS TARTUENSIS

1. **Toivo Maimets**. Studies of human oncoprotein p53. Tartu, 1991, 96 p.
2. **Enn K. Seppet**. Thyroid state control over energy metabolism, ion transport and contractile functions in rat heart. Tartu, 1991, 135 p.
3. **Kristjan Zobel**. Epifüütsete makrosamblike väärtus õhu saastuse indikaatoritena Hamar-Dobani boreaalsetes mägimetsades. Tartu, 1992, 131 lk.
4. **Andres Mäe**. Conjugal mobilization of catabolic plasmids by transposable elements in helper plasmids. Tartu, 1992, 91 p.
5. **Maia Kivisaar**. Studies on phenol degradation genes of *Pseudomonas* sp. strain EST 1001. Tartu, 1992, 61 p.
6. **Allan Nurk**. Nucleotide sequences of phenol degradative genes from *Pseudomonas sp.* strain EST 1001 and their transcriptional activation in *Pseudomonas putida.* Tartu, 1992, 72 p.
7. **Ülo Tamm**. The genus *Populus* L. in Estonia: variation of the species biology and introduction. Tartu, 1993, 91 p.
8. **Jaanus Remme**. Studies on the peptidyltransferase centre of the *E.coli* ribosome. Tartu, 1993, 68 p.
9. **Ülo Langel**. Galanin and galanin antagonists. Tartu, 1993, 97 p.
10. **Arvo Käärd**. The development of an automatic online dynamic fluorescense-based pH-dependent fiber optic penicillin flowthrought biosensor for the control of the benzylpenicillin hydrolysis. Tartu, 1993, 117 p.
11. **Lilian Järvekülg**. Antigenic analysis and development of sensitive immunoassay for potato viruses. Tartu, 1993, 147 p.
12. **Jaak Palumets**. Analysis of phytomass partition in Norway spruce. Tartu, 1993, 47 p.
13. **Arne Sellin**. Variation in hydraulic architecture of *Picea abies* (L.) Karst. trees grown under different enviromental conditions. Tartu, 1994, 119 p.

13. **Mati Reeben**. Regulation of light neurofilament gene expression. Tartu, 1994, 108 p.
14. **Urmas Tartes**. Respiration rhytms in insects. Tartu, 1995, 109 p.
15. **Ülo Puurand.** The complete nucleotide sequence and infections *in vitro* transcripts from cloned cDNA of a potato A potyvirus. Tartu, 1995, 96 p.
16. **Peeter Hõrak**. Pathways of selection in avian reproduction: a functional framework and its application in the population study of the great tit (*Parus major*). Tartu, 1995, 118 p.
17. **Erkki Truve**. Studies on specific and broad spectrum virus resistance in transgenic plants. Tartu, 1996, 158 p.
18. **Illar Pata**. Cloning and characterization of human and mouse ribosomal protein S6-encoding genes. Tartu, 1996, 60 p.
19. **Ülo Niinemets**. Importance of structural features of leaves and canopy in determining species shade-tolerance in temperature deciduous woody taxa. Tartu, 1996, 150 p.

20. **Ants Kurg**. Bovine leukemia virus: molecular studies on the packaging region and DNA diagnostics in cattle. Tartu, 1996, 104 p.
21. **Ene Ustav**. E2 as the modulator of the BPV1 DNA replication. Tartu, 1996, 100 p.
22. **Aksel Soosaar**. Role of helix-loop-helix and nuclear hormone receptor transcription factors in neurogenesis. Tartu, 1996, 109 p.
23. **Maido Remm**. Human papillomavirus type 18: replication, transformation and gene expression. Tartu, 1997, 117 p.
24. **Tiiu Kull**. Population dynamics in *Cypripedium calceolus* L. Tartu, 1997, 124 p.
25. **Kalle Olli**. Evolutionary life-strategies of autotrophic planktonic microorganisms in the Baltic Sea. Tartu, 1997, 180 p.
26. **Meelis Pärtel**. Species diversity and community dynamics in calcareous grassland communities in Western Estonia. Tartu, 1997, 124 p.
27. **Malle Leht**. The Genus *Potentilla* L. in Estonia, Latvia and Lithuania: distribution, morphology and taxonomy. Tartu, 1997, 186 p.
28. **Tanel Tenson**. Ribosomes, peptides and antibiotic resistance. Tartu, 1997, 80 p.
29. **Arvo Tuvikene**. Assessment of inland water pollution using biomarker responses in fish *in vivo* and *in vitro.* Tartu, 1997, 160 p.
30. **Urmas Saarma**. Tuning ribosomal elongation cycle by mutagenesis of 23S rRNA. Tartu, 1997, 134 p.
31. **Henn Ojaveer**. Composition and dynamics of fish stocks in the gulf of Riga ecosystem. Tartu, 1997, 138 p.
32. **Lembi Lõugas**. Post-glacial development of vertebrate fauna in Estonian water bodies. Tartu, 1997, 138 p.
33. **Margus Pooga**. Cell penetrating peptide, transportan, and its predecessors, galanin-based chimeric peptides. Tartu, 1998, 110 p.
34. **Andres Saag**. Evolutionary relationships in some cetrarioid genera (Lichenized Ascomycota). Tartu, 1998, 196 p.
35. **Aivar Liiv**. Ribosomal large subunit assembly *in vivo*. Tartu, 1998, 158 p.
36. **Tatjana Oja**. Isoenzyme diversity and phylogenetic affinities among the eurasian annual bromes (*Bromus* L., Poaceae). Tartu, 1998, 92 p.
37. **Mari Moora**. The influence of arbuscular mycorrhizal (AM) symbiosis on the competition and coexistence of calcareous crassland plant species. Tartu, 1998, 78 p.
38. **Olavi Kurina**. Fungus gnats in Estonia (*Diptera: Bolitophilidae, Keroplatidae, Macroceridae, Ditomyiidae, Diadocidiidae, Mycetophilidae*). Tartu, 1998, 200 p.
39. **Andrus Tasa**. Biological leaching of shales: black shale and oil shale. Tartu, 1998, 98 p.
40. **Arnold Kristjuhan.** Studies on transcriptional activator properties of tumor suppressor protein p53. Tartu, 1998, 86 p.

41. **Sulev Ingerpuu.** Characterization of some human myeloid cell surface and nuclear differentiation antigens. Tartu, 1998, 163 p.
42. **Veljo Kisand.** Responses of planktonic bacteria to the abiotic and biotic factors in the shallow lake Võrtsjärv. Tartu, 1998, 118 p.
43. **Kadri Põldmaa.** Studies in the systematics of hypomyces and allied genera (Hypocreales, Ascomycota). Tartu, 1998, 178 p.
44. **Markus Vetemaa.** Reproduction parameters of fish as indicators in environmental monitoring. Tartu, 1998, 117 p.
45. **Heli Talvik.** Prepatent periods and species composition of different *Oesophagostomum* spp. populations in Estonia and Denmark. Tartu, 1998, 104 p.
46. **Katrin Heinsoo.** Cuticular and stomatal antechamber conductance to water vapour diffusion in *Picea abies* (L.) karst. Tartu, 1999, 133 p.
47. **Tarmo Annilo.** Studies on mammalian ribosomal protein S7. Tartu, 1998, 77 p.
48. **Indrek Ots.** Health state indicies of reproducing great tits (*Parus major*): sources of variation and connections with life-history traits. Tartu, 1999, 117 p.
49. **Juan Jose Cantero.** Plant community diversity and habitat relationships in central Argentina grasslands. Tartu, 1999, 161 p.
50. **Rein Kalamees.** Seed bank, seed rain and community regeneration in Estonian calcareous grasslands. Tartu, 1999, 107 p.
51. **Sulev Kõks.** Cholecystokinin (CCK) — induced anxiety in rats: influence of environmental stimuli and involvement of endopioid mechanisms and erotonin. Tartu, 1999, 123 p.
52. **Ebe Sild.** Impact of increasing concentrations of O_3 and CO_2 on wheat, clover and pasture. Tartu, 1999, 123 p.
53. **Ljudmilla Timofejeva.** Electron microscopical analysis of the synaptonemal complex formation in cereals. Tartu, 1999, 99 p.
54. **Andres Valkna.** Interactions of galanin receptor with ligands and G-proteins: studies with synthetic peptides. Tartu, 1999, 103 p.
55. **Taavi Virro.** Life cycles of planktonic rotifers in lake Peipsi. Tartu, 1999, 101 p.
56. **Ana Rebane.** Mammalian ribosomal protein S3a genes and intron-encoded small nucleolar RNAs U73 and U82. Tartu, 1999, 85 p.
57. **Tiina Tamm.** Cocksfoot mottle virus: the genome organisation and translational strategies. Tartu, 2000, 101 p.
58. **Reet Kurg.** Structure-function relationship of the bovine papilloma virus E2 protein. Tartu, 2000, 89 p.
59. **Toomas Kivisild.** The origins of Southern and Western Eurasian populations: an mtDNA study. Tartu, 2000, 121 p.
60. **Niilo Kaldalu.** Studies of the TOL plasmid transcription factor XylS. Tartu 2000. 88 p.

61. **Dina Lepik.** Modulation of viral DNA replication by tumor suppressor protein p53. Tartu 2000. 106 p.
62. **Kai Vellak.** Influence of different factors on the diversity of the bryophyte vegetation in forest and wooded meadow communities. Tartu 2000. 122 p.
63. **Jonne Kotta.** Impact of eutrophication and biological invasionas on the structure and functions of benthic macrofauna. Tartu 2000. 160 p.
64. **Georg Martin.** Phytobenthic communities of the Gulf of Riga and the inner sea the West-Estonian archipelago. Tartu, 2000. 139 p.
65. **Silvia Sepp.** Morphological and genetical variation of *Alchemilla L.* in Estonia. Tartu, 2000. 124 p.
66. **Jaan Liira.** On the determinants of structure and diversity in herbaceous plant communities. Tartu, 2000. 96 p.
67. **Priit Zingel.** The role of planktonic ciliates in lake ecosystems. Tartu 2001. 111 p.
68. **Tiit Teder.** Direct and indirect effects in Host-parasitoid interactions: ecological and evolutionary consequences. Tartu 2001. 122 p.
69. **Hannes Kollist.** Leaf apoplastic ascorbate as ozone scavenger and its transport across the plasma membrane. Tartu 2001. 80 p.
70. **Reet Marits.** Role of two-component regulator system PehR-PehS and extracellular protease PrtW in virulence of *Erwinia Carotovora* subsp. *Carotovora*. Tartu 2001. 112 p.
71. **Vallo Tilgar.** Effect of calcium supplementation on reproductive performance of the pied flycatcher *Ficedula hypoleuca* and the great tit *Parus major,* breeding in Nothern temperate forests. Tartu, 2002. 126 p.
72. **Rita Hõrak.** Regulation of transposition of transposon Tn*4652* in *Pseudomonas putida*. Tartu, 2002. 108 p.
73. **Liina Eek-Piirsoo.** The effect of fertilization, mowing and additional illumination on the structure of a species-rich grassland community. Tartu, 2002. 74 p.
74. **Krõõt Aasamaa.** Shoot hydraulic conductance and stomatal conductance of six temperate deciduous tree species. Tartu, 2002. 110 p.
75. **Nele Ingerpuu.** Bryophyte diversity and vascular plants. Tartu, 2002. 112 p.
76. **Neeme Tõnisson.** Mutation detection by primer extension on oligonucleotide microarrays. Tartu, 2002. 124 p.
77. **Margus Pensa.** Variation in needle retention of Scots pine in relation to leaf morphology, nitrogen conservation and tree age. Tartu, 2003. 110 p.
78. **Asko Lõhmus.** Habitat preferences and quality for birds of prey: from principles to applications. Tartu, 2003. 168 p.
79. **Viljar Jaks.** p53 — a switch in cellular circuit. Tartu, 2003. 160 p.
80. **Jaana Männik.** Characterization and genetic studies of four ATP-binding cassette (ABC) transporters. Tartu, 2003. 140 p.
81. **Marek Sammul.** Competition and coexistence of clonal plants in relation to productivity. Tartu, 2003. 159 p

82. **Ivar Ilves.** Virus-cell interactions in the replication cycle of bovine papillomavirus type 1. Tartu, 2003. 89 p.
83. **Andres Männik.** Design and characterization of a novel vector system based on the stable replicator of bovine papillomavirus type 1. Tartu, 2003. 109 p.
84. **Ivika Ostonen.** Fine root structure, dynamics and proportion in net primary production of Norway spruce forest ecosystem in relation to site conditions. Tartu, 2003. 158 p.
85. **Gudrun Veldre.** Somatic status of 12–15-year-old Tartu schoolchildren. Tartu, 2003. 199 p.
86. **Ülo Väli.** The greater spotted eagle *Aquila clanga* and the lesser spotted eagle *A. pomarina*: taxonomy, phylogeography and ecology. Tartu, 2004. 159 p.
87. **Aare Abroi.** The determinants for the native activities of the bovine papillomavirus type 1 E2 protein are separable. Tartu, 2004. 135 p.
88. **Tiina Kahre.** Cystic fibrosis in Estonia. Tartu, 2004. 116 p.
89. **Helen Orav-Kotta.** Habitat choice and feeding activity of benthic suspension feeders and mesograzers in the northern Baltic Sea. Tartu, 2004. 117 p.
90. **Maarja Öpik.** Diversity of arbuscular mycorrhizal fungi in the roots of perennial plants and their effect on plant performance. Tartu, 2004. 175 p.
91. **Kadri Tali.** Species structure of *Neotinea ustulata*. Tartu, 2004. 109 p.
92. **Kristiina Tambets.** Towards the understanding of post-glacial spread of human mitochondrial DNA haplogroups in Europe and beyond: a phylogeographic approach. Tartu, 2004. 163 p.
93. **Arvi Jõers.** Regulation of p53-dependent transcription. Tartu, 2004. 103 p.
94. **Lilian Kadaja.** Studies on modulation of the activity of tumor suppressor protein p53. Tartu, 2004. 103 p.
95. **Jaak Truu.** Oil shale industry wastewater: impact on river microbial community and possibilities for bioremediation. Tartu, 2004. 128 p.
96. **Maire Peters.** Natural horizontal transfer of the *pheBA* operon. Tartu, 2004. 105 p.
97. **Ülo Maiväli.** Studies on the structure-function relationship of the bacterial ribosome. Tartu, 2004. 130 p.
98. **Merit Otsus.** Plant community regeneration and species diversity in dry calcareous grasslands. Tartu, 2004. 103 p.
99. **Mikk Heidemaa.** Systematic studies on sawflies of the genera *Dolerus, Empria,* and *Caliroa* (Hymenoptera: Tenthredinidae). Tartu, 2004. 167 p.
100. **Ilmar Tõnno.** The impact of nitrogen and phosphorus concentration and N/P ratio on cyanobacterial dominance and N_2 fixation in some Estonian lakes. Tartu, 2004. 111 p.
101. **Lauri Saks.** Immune function, parasites, and carotenoid-based ornaments in greenfinches. Tartu, 2004. 144 p.

102. **Siiri Rootsi.** Human Y-chromosomal variation in European populations. Tartu, 2004. 142 p.

103. **Eve Vedler.** Structure of the 2,4-dichloro-phenoxyacetic acid-degradative plasmid pEST4011. Tartu, 2005. 106 p.

104. **Andres Tover.** Regulation of transcription of the phenol degradation *pheBA* operon in *Pseudomonas putida*. Tartu, 2005. 126 p.

105. **Helen Udras.** Hexose kinases and glucose transport in the yeast *Hansenula polymorpha*. Tartu, 2005. 100 p.

106. **Ave Suija.** Lichens and lichenicolous fungi in Estonia: diversity, distribution patterns, taxonomy. Tartu, 2005. 162 p.

107. **Piret Lõhmus.** Forest lichens and their substrata in Estonia. Tartu, 2005. 162 p.

108. **Inga Lips.** Abiotic factors controlling the cyanobacterial bloom occurrence in the Gulf of Finland. Tartu, 2005. 156 p.

109. **Kaasik, Krista.** Circadian clock genes in mammalian clockwork, metabolism and behaviour. Tartu, 2005. 121 p.

110. **Juhan Javoiš.** The effects of experience on host acceptance in ovipositing moths. Tartu, 2005. 112 p.

111. **Tiina Sedman.** Characterization of the yeast *Saccharomyces cerevisiae* mitochondrial DNA helicase Hmi1. Tartu, 2005. 103 p.

112. **Ruth Aguraiuja.** Hawaiian endemic fern lineage *Diellia* (Aspleniaceae): distribution, population structure and ecology. Tartu, 2005. 112 p.

113. **Riho Teras.** Regulation of transcription from the fusion promoters generated by transposition of Tn*4652* into the upstream region of *pheBA* operon in *Pseudomonas putida*. Tartu, 2005. 106 p.

114. **Mait Metspalu.** Through the course of prehistory in india: tracing the mtDNA trail. Tartu, 2005. 138 p.

115. **Elin Lõhmussaar.** The comparative patterns of linkage disequilibrium in European populations and its implication for genetic association studies. Tartu, 2006. 124 p.

116. **Priit Kupper.** Hydraulic and environmental limitations to leaf water relations in trees with respect to canopy position. Tartu, 2006. 126 p.

117. **Heili Ilves.** Stress-induced transposition of Tn*4652* in *Pseudomonas Putida.* Tartu, 2006. 120 p.

118. **Silja Kuusk.** Biochemical properties of Hmi1p, a DNA helicase from *Saccharomyces cerevisiae* mitochondria. Tartu, 2006. 126 p.

119. **Kersti Püssa.** Forest edges on medium resolution landsat thematic mapper satellite images. Tartu, 2006. 90 p.

120. **Lea Tummeleht.** Physiological condition and immune function in great tits (*Parus major* l.): Sources of variation and trade-offs in relation to growth. Tartu, 2006. 94 p.

121. **Toomas Esperk.** Larval instar as a key element of insect growth schedules. Tartu, 2006. 186 p.

122. **Harri Valdmann.** Lynx (*Lynx lynx*) and wolf (*Canis lupus*) in the Baltic region: Diets, helminth parasites and genetic variation. Tartu, 2006. 102 p.
123. **Priit Jõers.** Studies of the mitochondrial helicase Hmi1p in *Candida albicans* and *Saccharomyces cerevisia*. Tartu, 2006. 113 p.
124. **Kersti Lilleväli.** Gata3 and Gata2 in inner ear development. Tartu, 2007. 123 p.
125. **Kai Rünk.** Comparative ecology of three fern species: *Dryopteris carthusiana* (Vill.) H.P. Fuchs, *D. expansa* (C. Presl) Fraser-Jenkins & Jermy and *D. dilatata* (Hoffm.) A. Gray (Dryopteridaceae). Tartu, 2007. 143 p.
126. **Aveliina Helm.** Formation and persistence of dry grassland diversity: role of human history and landscape structure. Tartu, 2007. 89 p.
127. **Leho Tedersoo.** Ectomycorrhizal fungi: diversity and community structure in Estonia, Seychelles and Australia. Tartu, 2007. 233 p.
128. **Marko Mägi.** The habitat-related variation of reproductive performance of great tits in a deciduous-coniferous forest mosaic: looking for causes and consequences. Tartu, 2007. 135 p.
129. **Valeria Lulla.** Replication strategies and applications of Semliki Forest virus. Tartu, 2007. 109 p.
130. **Ülle Reier**. Estonian threatened vascular plant species: causes of rarity and conservation. Tartu, 2007. 79 p.
131. **Inga Jüriado**. Diversity of lichen species in Estonia: influence of regional and local factors. Tartu, 2007. 171 p.
132. **Tatjana Krama.** Mobbing behaviour in birds: costs and reciprocity based cooperation. Tartu, 2007. 112 p.
133. **Signe Saumaa.** The role of DNA mismatch repair and oxidative DNA damage defense systems in avoidance of stationary phase mutations in *Pseudomonas putida.* Tartu, 2007. 172 p.
134. **Reedik Mägi**. The linkage disequilibrium and the selection of genetic markers for association studies in european populations. Tartu, 2007. 96 p.
135. **Priit Kilgas.** Blood parameters as indicators of physiological condition and skeletal development in great tits (*Parus major*): natural variation and application in the reproductive ecology of birds. Tartu, 2007. 129 p.
136. **Anu Albert**. The role of water salinity in structuring eastern Baltic coastal fish communities. Tartu, 2007. 95 p.
137. **Kärt Padari.** Protein transduction mechanisms of transportans. Tartu, 2008. 128 p.
138. **Siiri-Lii Sandre.** Selective forces on larval colouration in a moth. Tartu, 2008. 125 p.
139. **Ülle Jõgar.** Conservation and restoration of semi-natural floodplain meadows and their rare plant species. Tartu, 2008. 99 p.
140. **Lauri Laanisto.** Macroecological approach in vegetation science: generality of ecological relationships at the global scale. Tartu, 2008. 133 p.
141. **Reidar Andreson**. Methods and software for predicting PCR failure rate in large genomes. Tartu, 2008. 105 p.

142. **Birgot Paavel.** Bio-optical properties of turbid lakes. Tartu, 2008. 175 p.
143. **Kaire Torn.** Distribution and ecology of charophytes in the Baltic Sea. Tartu, 2008, 98 p.
144. **Vladimir Vimberg.** Peptide mediated macrolide resistance. Tartu, 2008, 190 p.
145. **Daima Örd.** Studies on the stress-inducible pseudokinase TRB3, a novel inhibitor of transcription factor ATF4. Tartu, 2008, 108 p.
146. **Lauri Saag.** Taxonomic and ecologic problems in the genus *Lepraria* (*Stereocaulaceae*, lichenised *Ascomycota*). Tartu, 2008, 175 p.
147. **Ulvi Karu.** Antioxidant protection, carotenoids and coccidians in greenfinches – assessment of the costs of immune activation and mechanisms of parasite resistance in a passerine with carotenoid-based ornaments. Tartu, 2008, 124 p.
148. **Jaanus Remm.** Tree-cavities in forests: density, characteristics and occupancy by animals. Tartu, 2008, 128 p.
149. **Epp Moks.** Tapeworm parasites *Echinococcus multilocularis* and *E. granulosus* in Estonia: phylogenetic relationships and occurrence in wild carnivores and ungulates. Tartu, 2008, 82 p.
150. **Eve Eensalu.** Acclimation of stomatal structure and function in tree canopy: effect of light and CO_2 concentration. Tartu, 2008, 108 p.
151. **Janne Pullat**. Design, functionlization and application of an *in situ* synthesized oligonucleotide microarray. Tartu, 2008, 108 p.
152. **Marta Putrinš.** Responses of *Pseudomonas putida* to phenol-induced metabolic and stress signals. Tartu, 2008, 142 p.
153. **Marina Semtšenko.** Plant root behaviour: responses to neighbours and physical obstructions. Tartu, 2008, 106 p.
154. **Marge Starast.** Influence of cultivation techniques on productivity and fruit quality of some *Vaccinium* and *Rubus* taxa. Tartu, 2008, 154 p.
155. **Age Tats.** Sequence motifs influencing the efficiency of translation. Tartu, 2009, 104 p.
156. **Radi Tegova.** The role of specialized DNA polymerases in mutagenesis in *Pseudomonas putida.* Tartu, 2009, 124 p.
157. **Tsipe Aavik.** Plant species richness, composition and functional trait pattern in agricultural landscapes – the role of land use intensity and landscape structure. Tartu, 2008, 112 p.
158. **Kaja Kiiver.** Semliki forest virus based vectors and cell lines for studying the replication and interactions of alphaviruses and hepaciviruses. Tartu, 2009, 104 p.
159. **Meelis Kadaja.** Papillomavirus Replication Machinery Induces Genomic Instability in its Host Cell. Tartu, 2009, 126 p.
160. **Pille Hallast.** Human and chimpanzee Luteinizing hormone/Chorionic Gonadotropin beta (*LHB/CGB*) gene clusters: diversity and divergence of young duplicated genes. Tartu, 2009, 168 p.

161. **Ain Vellak.** Spatial and temporal aspects of plant species conservation. Tartu, 2009, 86 p.

162. **Triinu Remmel.** Body size evolution in insects with different colouration strategies: the role of predation risk. Tartu, 2009, 168 p.

163. **Jaana Salujõe.** Zooplankton as the indicator of ecological quality and fish predation in lake ecosystems. Tartu, 2009, 129 p.

164. **Ele Vahtmäe.** Mapping benthic habitat with remote sensing in optically complex coastal environments. Tartu, 2009, 109 p.

165. **Liisa Metsamaa.** Model-based assessment to improve the use of remote sensing in recognition and quantitative mapping of cyanobacteria. Tartu, 2009, 114 p.

166. **Pille Säälik.** The role of endocytosis in the protein transduction by cell-penetrating peptides. Tartu, 2009, 155 p.

167. **Lauri Peil.** Ribosome assembly factors in *Escherichia coli.* Tartu, 2009, 147 p.

168. **Lea Hallik.** Generality and specificity in light harvesting, carbon gain capacity and shade tolerance among plant functional groups. Tartu, 2009, 99 p.

169. **Mariliis Tark.** Mutagenic potential of DNA damage repair and tolerance mechanisms under starvation stress. Tartu, 2009, 191 p.

170. **Riinu Rannap.** Impacts of habitat loss and restoration on amphibian populations. Tartu, 2009, 117 p.

171. **Maarja Adojaan.** Molecular variation of HIV-1 and the use of this knowledge in vaccine development. Tartu, 2009, 95 p.

172. **Signe Altmäe.** Genomics and transcriptomics of human induced ovarian folliculogenesis. Tartu, 2010, 179 p.

173. **Triin Suvi.** Mycorrhizal fungi of native and introduced trees in the Seychelles Islands. Tartu, 2010, 107 p.

174. **Velda Lauringson.** Role of suspension feeding in a brackish-water coastal sea. Tartu, 2010, 123 p.

175. **Eero Talts.** Photosynthetic cyclic electron transport – measurement and variably proton-coupled mechanism. Tartu, 2010, 121 p.

176. **Mari Nelis.** Genetic structure of the Estonian population and genetic distance from other populations of European descent. Tartu, 2010, 97 p.

177. **Kaarel Krjutškov.** Arrayed Primer Extension-2 as a multiplex PCR-based method for nucleic acid variation analysis: method and applications. Tartu, 2010, 129 p.

178. **Egle Köster.** Morphological and genetical variation within species complexes: *Anthyllis vulneraria* s. l. and *Alchemilla vulgaris* (coll.). Tartu, 2010, 101 p.

179. **Erki Õunap.** Systematic studies on the subfamily Sterrhinae (Lepidoptera: Geometridae). Tartu, 2010, 111 p.

180. **Merike Jõesaar.** Diversity of key catabolic genes at degradation of phenol and *p*-cresol in pseudomonads. Tartu, 2010, 125 p.

181. **Kristjan Herkül.** Effects of physical disturbance and habitat-modifying species on sediment properties and benthic communities in the northern Baltic Sea. Tartu, 2010, 123 p.
182. **Arto Pulk.** Studies on bacterial ribosomes by chemical modification approaches. Tartu, 2010, 161 p.
183. **Maria Põllupüü.** Ecological relations of cladocerans in a brackish-water ecosystem. Tartu, 2010, 126 p.
184. **Toomas Silla.** Study of the segregation mechanism of the Bovine Papillomavirus Type 1. Tartu, 2010, 188 p.
185. **Gyaneshwer Chaubey.** The demographic history of India: A perspective based on genetic evidence. Tartu, 2010, 184 p.
186. **Katrin Kepp.** Genes involved in cardiovascular traits: detection of genetic variation in Estonian and Czech populations. Tartu, 2010, 164 p.
187. **Virve Sõber.** The role of biotic interactions in plant reproductive performance. Tartu, 2010, 92 p.
188. **Kersti Kangro.** The response of phytoplankton community to the changes in nutrient loading. Tartu, 2010, 144 p.
189. **Joachim M. Gerhold.** Replication and Recombination of mitochondrial DNA in Yeast. Tartu, 2010, 120 p.
190. **Helen Tammert.** Ecological role of physiological and phylogenetic diversity in aquatic bacterial communities. Tartu, 2010, 140 p.
191. **Elle Rajandu.** Factors determining plant and lichen species diversity and composition in Estonian *Calamagrostis* and *Hepatica* site type forests. Tartu, 2010, 123 p.
192. **Paula Ann Kivistik.** ColR-ColS signalling system and transposition of Tn*4652* in the adaptation of *Pseudomonas putida.* Tartu, 2010, 118 p.
193. **Siim Sõber.** Blood pressure genetics: from candidate genes to genome-wide association studies. Tartu, 2011, 120 p.
194. **Kalle Kipper.** Studies on the role of helix 69 of 23S rRNA in the factor-dependent stages of translation initiation, elongation, and termination. Tartu, 2011, 178 p.
195. **Triinu Siibak.** Effect of antibiotics on ribosome assembly is indirect. Tartu, 2011, 134 p.
196. **Tambet Tõnissoo.** Identification and molecular analysis of the role of guanine nucleotide exchange factor RIC-8 in mouse development and neural function. Tartu, 2011, 110 p.
197. **Helin Räägel.** Multiple faces of cell-penetrating peptides – their intracellular trafficking, stability and endosomal escape during protein transduction. Tartu, 2011, 161 p.
198. **Andres Jaanus.** Phytoplankton in Estonian coastal waters – variability, trends and response to environmental pressures. Tartu, 2011, 157 p.
199. **Tiit Nikopensius.** Genetic predisposition to nonsyndromic orofacial clefts. Tartu, 2011, 152 p.
200. **Signe Värv.** Studies on the mechanisms of RNA polymerase II-dependent transcription elongation. Tartu, 2011, 108 p.