View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Helsingin yliopiston digitaalinen arkisto

Microarrays in Lung Cancer Research:
From Comparative Analyses to Verified Findings

Salla Ruosaari

Health and Work Ability
Finnish Institute of Occupational Health
Helsinki, Finland

Department of Information and Computer Science
Helsinki University of Technology
Espoo, Finland

Department of Biological and Environmental Sciences
Faculty of Biosciences
University of Helsinki, Finland

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Biosciences of the University of Helsinki, for public criticism in Lecture Hall 1041 of Biocenter 2, Viikinkaari 5, Helsinki, on June 12th, 2008, at 12 o'clock noon.

Helsinki 2008

SUPERVISED BY

Docent Sisko Anttila
Health and Work Ability
Finnish Institute of Occupational Health
Helsinki, Finland

Jaakko Hollmén, D.Sc.(Tech.)
Department of Information and Computer Science
Helsinki University of Technology
Espoo, Finland

REVIEWED BY

Docent Outi Monni
Institute of Biomedicine and Biomedicum Biochip Center
University of Helsinki
Helsinki, Finland

Docent Sampsa Hautaniemi
Computational Systems Biology Laboratory
Institute of Biomedicine and Genome-Scale Biology Research Program
University of Helsinki
Helsinki, Finland

OPPONENT

Professor Jorma Isola
Institute of Medical Technology
University of Tampere
Tampere, Finland

ISBN 978-952-92-3994-8 (paperback)
ISBN 978-952-10-4739-8 (PDF)
http://ethesis.helsinki.fi
Helsinki 2008
Yliopistopaino

CONTENTS

ABSTRACT **5**

LIST OF ORIGINAL PUBLICATIONS **7**

ABBREVIATIONS **8**

REVIEW OF LITERATURE **10**

1 CANCER GENETICS 10
1.1 Oncogenes and tumor suppressor genes *10*
1.2 Genetic instability *12*
1.3 Genomic alterations *13*
1.4 Fragile sites *14*
1.5 Signaling pathways *15*
2 LUNG CANCER 16
2.1 Histological classification of LC *16*
2.2 Risk factors for LC *17*
2.3 Genetic alterations in LC *17*
Changes in LC related to smoking *18*
Changes in LC related to asbestos *19*
3 MICROARRAYS IN CANCER RESEARCH 20
3.1 Microarray data processing *21*
3.2 Pathway analysis of microarray data *22*
Gene Ontology *23*
3.3 Combining array CGH and gene expression *24*
3.4 Considerations related to microarray analyses *25*
3.5 Microarrays in lung cancer *26*

AIMS OF THE STUDY **28**

MATERIALS **29**

1 STUDY SUBJECTS 29
1.1 Asbestos related studies (II-V) *29*
1.2 Studies related to disseminating LCs (VI) *29*
2 MICROARRAY ANALYSIS METHODS 31
2.1 Three-color image analysis (I) *31*
2.2 Array CGH (II, IV & VI) *32*
2.3 Gene expression microarrays (III, V & VI) *33*
2.4 Differential region finding method (II, III & VI) *34*
2.5 ROC curves (III & VI) *35*
2.6 GO analysis (V) *35*
3 VERIFICATION METHODS 36
3.1 Microsatellite analysis of 19p (III) *36*
3.2 FISH (IV & VI) *37*
3.3 Evaluation of UBA1 and UBA7 levels by Western blotting (V) *37*
3.4 Studies on human bronchial epithelial BEAS 2B cells (IV) *38*

RESULTS **39**

1 THREE-COLOR IMAGE ANALYSIS (I) 39
2 DIFFERENTIAL REGION FINDING (II, III & VI) 39
3 ASBESTOS RELATED CHANGES 40
3.1 Molecular changes of asbestos related tumors (II & III) *40*
3.2 Region 19p in asbestos related carcinogenesis (III & IV) *41*
19p aberrations revealed by array CGH (IV) *41*
Microsatellite analysis (III) *42*
FISH (IV) *43*
Micronuclei induced by crocidolite asbestos (IV) *44*
3.3 Asbestos related deregulated pathways (V) *45*
4 MOLECULAR CHANGES OF TUMORS WITH DTCS IN BM (VI) 46

DISCUSSION **48**

1 MICROARRAY ANALYSIS 48
2 ALTERATIONS INDUCED BY ASBESTOS EXPOSURE 49
2.1 Aberrations characteristic of asbestos related cancers (II - IV) *49*
2.2 Asbestos related deregulated pathways (V) *52*
3 ALTERATIONS UNDERLYING DISSEMINATION INTO BM (VI) 54

SUMMARY AND CONCLUSIONS 56

ACKNOWLEDGEMENTS 58

REFERENCES 59

ABSTRACT

Lung cancers (LCs) represent a heterogeneous collection of tumors that are characterized by a large number of abnormalities of both chromosome number and structure. The genetic alterations displayed by a given tumor are the result of a combination of changes that are directly or indirectly caused by inducing factors, such as tobacco carcinogens, and those that rise up secondarily as a consequence of defects in genes that maintain genomic stability. Although a number of genes have been identified to be recurrently aberrated in LCs, numerous genes contributing to lung carcinogenesis are yet to be revealed.

In this thesis, opposite categories of lung tumors were studied using comparative microarray analyses with the aim of identifying aberrations that are of significance in distinct types of LCs. Microarrays were used as they provide information about virtually all of the genes of the subject of the study in a single assay and enable genome wide studies without *a priori* information of the affected genes. The studies involved LCs in two separate themes: those induced by asbestos exposure and those that disseminate into bone marrow (BM) in the early stages of tumorigenesis. The category specific alterations were identified by using a differential region finding (DRF) method designed for comparative studies of DNA and gene expression level high-throughput data that was developed in this thesis.

The studies carried out for this thesis showed that comparative analyses of opposite tumor categories are practical in the identification of molecular changes that are characteristic of sub-categories of LCs. Distinct chromosome regions were found to be more frequently aberrated in the LCs of asbestos exposed than nonexposed patients. Aberrations of distinct regions also differentiated lung adenocarcinomas (ACs) from patients with (BM-positive) and without (BM-negative) evidence of disseminated tumor cells (DTCs) in their BM. In both study settings, further verifications and characterizations of the findings were performed on one of the putative sites of preferential aberrations.

In the asbestos related studies, microsatellite and fluorescence *in situ* hybridization (FISH) analyses verified that aberrations of 19p, caused by losses, were significantly more frequent in tumors of the asbestos exposed than of the nonexposed patients. We showed also that 19p aberrations can be induced *in vitro* by means of a crocidolite asbestos treatment. Furthermore, a Gene Ontology (GO) analysis revealed a number of differentially regulated biological processes and molecular functions between the tumor groups with differences in protein ubiquitination and ion transport especially highlighted.

In the subsequent dissemination related studies, FISH analyses of the 4q region were performed on both primary LCs and brain metastases of LCs. The connection between the loss of 4q and the presence of DTCs in BM in ACs could be verified but 4q loss was also demonstrated to be a common feature of BM-positive tumors across different histological types of LCs. Losses of 4q were also frequently observed in brain metastases of LCs indicating that the aberration could be a universal feature of spreading LCs.

The results of this thesis imply that LCs that have been influenced by asbestos exposure and those that disseminate to BM have distinct molecular changes. However, the research initiated in thesis needs to be continued in order to uncover the target genes of the preferential aberrations related to asbestos and dissemination.

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following original publications, which are referred to in the text by the Roman numerals I-VI as indicated below.

I Gupta R*, **Ruosaari S***, Kulathinal S, Hollmén J, and Auvinen P. (2007) Microarray image segmentation using additional dye - An experimental study. *Mol Cell Probes*, **21 (5-6)**, 321-8.

II Nymark P*, Wikman H*, **Ruosaari S**, Hollmén J, Vanhala E, Karjalainen A, Anttila S, and Knuutila S. (2006) Identification of specific gene copy number changes in asbestos-related lung cancer. *Cancer Res*, **66 (11)**, 5737-43.

III Wikman H*, **Ruosaari S*,** Nymark P, Sarhadi VK, Saharinen J, Vanhala E, Karjalainen A, Hollmén J, Knuutila S, and Anttila S. (2007) Gene expression and copy number profiling suggests the importance of allelic imbalance in 19p in asbestos-associated lung cancer. *Oncogene*, **26 (32)**, 4730-7.

IV **Ruosaari S**, Nymark P, Aavikko M, Kettunen E, Knuutila S, Hollmén J, Norppa H, and Anttila S. Aberrations of chromosome 19 in asbestos-associated lung cancer and in asbestos-induced micronuclei of bronchial epithelial cells *in vitro*. (2008) *Carcinogenesis*, Mar 13, [Epub ahead of print].

V **Ruosaari S***, Hienonen-Kempas T*, Puustinen A, Sarhadi VK, Hollmén J, Knuutila S, Saharinen J, Wikman H, and Anttila S. Pathways affected by asbestos exposure in normal and tumour tissue of lung cancer patients. Manuscript.

VI Kraemling M, **Ruosaari S**, Eijk PP, Kaifi JT, Hollmén J, Yekebas EF, Izbicki JR, Brakenhoff RH, Streichert T, Riethdorf S, Glatzel M, Ystra B, Pantel K, and Wikman H. Genomic profiles associated with early micrometastatis in lung cancer: Relevance of 4q deletion. Manuscript.

* The authors contributed equally to the study

ABBREVIATIONS

8-OHdG	8-hydroxy-guanine
ABL	c-abl oncogene 1, receptor tyrosine kinase
AC	adenocarcinoma
AI	allelic imbalance
BAC	bacterial artificial chromosome
BAT-26	"Big A-Tract", a marker of 26 adenine residues
BCR	breakpoint cluster region
BAX	BCL2-associated X protein
BER	base-excision repair
BM	bone marrow
bp	base pair
BRAF	v-raf murine sarcoma viral oncogene homolog B1
C^-	without centromere signal
C^+	with centromere signal
CCND1	cyclin D1
CDK4	cyclin-dependent kinase 4
CDKN2A	cyclin-dependent kinase inhibitor 2A
CGH	comparative genomic hybridization
CIN	chromosomal instability
Cy3	Cyanine3
Cy5	Cyanine5
Cyt-B	cytochalasin-B
DAG	directed acyclic graph
DRF	differential region finding
DTC	disseminated tumor cell
EGFR	epidermal growth factor receptor
FHIT	fragile histidine triad gene
FISH	fluorescence *in situ* hybridization
FRA3B	fragile site, aphidicolin type, common, fra(3)(p14.2)
FRA6E	fragile site, aphidicolin type, common, fra(6)(q26)
FRA7H	fragile site, aphidicolin type, common, fra(7)(q32.3)
FRA16D	fragile site, aphidicolin type, common, fra(16)(q23.2)
FRAXB	site, aphidicolin type, common, fra(X)(p22.31) B
GO	Gene Ontology
HSR	Homogeneously staining region
iGA	Iterative Group Analysis
KRAS	v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
LC	lung cancer
LCLC	large cell lung carcinoma
LKB1/*STK11*	serine/threonine kinase 11
LOH	loss of heterozygosity
LOWESS	locally weighted linear regression
LN	lymph node
M-FISH	multicolor fluorescence *in situ* hybridization
MAPK	mitogen activated protein kinase
MLH1	mutL homolog 1
MMC	mitomycin C
MMR	mismatch repair
MN	micronucleus

MSH2	mutS homolog 2
MSI	microsatellite instability
MYC	v-myc myelocytomatosis viral oncogene homolog (avian)
NER	nucleotide-excision repair
NF-κB	nuclear factor kappaB
NSCLC	non-small cell lung carcinoma
PAH	polycyclic aromatic hydrocarbon
PCR	polymarase chain reaction
PI3K	phosphatidylinositol 3-kinase
RAS	rat sarcoma viral oncogene homolog
RASSF1A	Ras association (RalGDS/AF-6) domain family member 1
RB1	retinoblastoma 1
RMA	robust multi-array average
RNS	reactive nitrogen species
ROC	Receiver Operating Characteristic
ROS	reactive oxygen species
SCC	squamous cell carcinoma
SCLC	small cell lung carcinoma
SNP	single nucleotide polymorphism
TGFBR2	transforming growth factor, beta receptor II (70/80kDa)
TP53	tumor protein p53
TSG	tumor suppressor gene
UBA1	ubiquitin-like modifier activating enzyme 1
UBA7	ubiquitin-like modifier activating enzyme 7
UPD	uniparental disomy

REVIEW OF LITERATURE

1 CANCER GENETICS

Cancer is a genetic disease of somatic cells that develops through multiple successive advantageous changes (Nowell 1976). Although the changes occurring in different tumors may vary, distinct alterations in cell physiology are needed for a tumor to develop. These are self-sufficiency in growth signals, insensitivity to anti-growth signals, evasion of apoptosis, limitless replicative potential, sustained angiogenesis, tissue invasion and metastasis (reviewed in Hanahan & Weinberg 2000).

Cancer cells originate from a single ancestral cell that gains a selective growth advantage over its neighboring normal cells (Nowell 1976). Only rare populations of cells within a tumor have been shown to possess the potential to initiate and sustain cancer growth (Lapidot *et al*. 1994; Al-Hajj *et al*. 2003; O'Brien *et al*. 2007) indicating that hierarchy of the cells persists also in tumors. No single cause can be usually pinpointed for somatic cancers, but instead, environmental factors and lifestyle choices predispose individuals to the disease (Weinberg 2007). A heritable predisposition is also known for virtually every form of cancer (reviewed in Knudson 2002).

From a histological perspective, tumors can be grouped into four major categories on the basis of their tissue of origin. The epithelium derived carcinomas are the most common group accounting for 80% of all incidences of cancer. The non-epithelial cancers include connective tissue derived sarcomas, leukemias and lymphomas derived from blood forming tissue and neuroectodermal tumors of the central and peripheral nervous systems (Weinberg 2007). The emphasis of this thesis is on carcinomas, particularly lung carcinomas, and features relevant to this category of cancers are mainly discussed.

1.1 Oncogenes and tumor suppressor genes

Molecular changes in two groups of genes with opposite functions, namely proto-oncogenes and tumor suppressor genes (TSGs) play a pivotal role in the development of cancer (Vogelstein 1988; Kim *et al.* 1994). Proto-oncogenes stimulate cell growth and inhibit apoptosis, whereas TSGs inhibit growth and maintain the integrity of the cell. Deregulation of these genes may alter the cell's capacity for controlled growth and differentiation (Anderson *et al.* 1992) and thereby lead to tumorigenic transformation of the cell.

Proto-oncogenes were first identified in mammalian cells through homology with retroviral oncogenes (Stehelin *et al*. 1976). They function for instance as growth factors, transcription factors and cell cycle regulators, and can transform into cancerous oncogenes when activated. The oncogenic change can occur through either regulatory or structural mechanisms, which include amplifications, point mutations, chromosomal translocations and loss of methylation, and result in an increase in protein concentration or in the creation of an altered protein (Roth 1995). One mutated allele is generally sufficient to active the proto-oncogene.

TSGs can be classified into caretakers and gatekeepers based on gene functions (Kinzler & Vogelstein 1997) and, contrary to oncogenes, TSGs are inactivated in cancer cells. Gatekeepers directly inhibit growth or promote cell death and are the most classical group of TSGs. Caretakers, on their part, are involved in DNA repair and replication, and thereby in the maintenance of the genomic integrity of the cell. The class includes mismatch repair (MMR), nucleotide-excision repair (NER) and base-excision repair (BER) genes responsible for correcting mistakes of normal DNA replication or those induced by mutagens. Inactivation of a caretaker gene promotes tumor initiation indirectly through an increased mutation rate of other genes including those involved in tumorigenesis (Kinzler & Vogelstein 1997; reviewed in Friedberg 2003). Furthermore, a third group of TSGs called landscapers has been proposed to exist, with contributions to neoplastic transformation through an abnormal microenvironment (Kinzler & Vogelstein 1998). TSG inactivation mechanisms include intragenic mutations, losses of wild type chromosomes and somatic recombinations (Weinberg 2007). Gene inactivation may also occur through epigenetic changes such as hyper-methylation of CpG islands, histone or chromatin modifications. These changes hinder the binding of transcription factors and therefore influence gene expression. Contrary to oncogenes, inactivation of both the maternal and the paternal alleles of a TSG are generally required to produce a phenotypic effect as initially proposed by Knudson (1971).

At the time of writing, altogether 367 human genes have been implicated in cancer via mutation, which accounts for more than 1% of genes in the human genome (reviewed in Futreal 2004; Cancer Gene Census at http://www.sanger.ac.uk/genetics/CGP/Census/). In terms of tumor development, the cancer genes seem to possess substitutive roles as no single gene is known to be activated or deleted from all cancers. However, differences in mutation rates between cancer genes are substantial. For instance, *TP53* is mutated in most human cancers, e.g. in lung cancers, *TP53* mutations are detected in up to 60% of cases

(Hollstein *et al.* 1991; Greenblatt *et al.* 1994), whereas the translocation formed oncogene *BCR-ABL* (Rowley 1973), for example, is characteristic of chronic myelogenous leukemias.

1.2 Genetic instability

Normal cells replicate their DNA with exceptional accuracy but carcinomas are genetically remarkably unstable (Loeb *et al.* 1974; Loeb 1991; Stoler *et al.* 1999). This was illustrated for instance in a recent analysis of 13,023 genes in 11 breast and 11 colorectal cancers, where an average tumor was shown to harbor about 90 mutated genes (Sjöblom *et al.* 2006). The increased mutation rate is thought to be related to malfunctions of the genes that are involved in the maintenance of genomic stability, i.e. have functions such as DNA repair, replication, chromosomal segregation, and cell cycle regulation (reviewed in Lengauer *et al*. 1998; Loeb & Loeb 1999; Bielas & Loeb 2005). The breakdowns of the maintenance genes necessitate a "mutator phenotype" that facilitates formation of new aberrations (Loeb 1991). It is, however, still debated whether genetic instability is the driving force of cancer progression or a result of it (Loeb 1991; Tomlinson *et al.* 1996; Sieber *et al.* 2002).

Genetic instability was proposed to have a role in tumorigenesis already in the 70's (Nowell 1976). The fundamental differences between the aberration profiles detected in cancers have further indicated that there are two types of genetic instability, microsatellite instability (MSI) and chromosomal instability (CIN) (Lengauer *et al* 1997). MSI and CIN define two distinct pathways for cancer development caused by a failure of the DNA MMR (Levinson & Gutman 1987; Parsons *et al.* 1993) and a deficiency of mitotic segregation (Lengauer *et al.* 1997; Cahill *et al*. 1998), respectively. One tumor may harbor both MSI and CIN but the rare co-existence of the instabilities indicates that tumorigenesis can be driven by just one of them (Lengauer *et al.* 1997).

MSI was first reported in colorectal cancers (Thibodeau *et al.* 1993), but has since been recognized in various other cancers including that of the lung (Hansen *et al.* 2003; Ninomiya *et al.* 2006). Cancers with MSI accumulate single nucleotide mutations and length variations in microsatellite sequences (Boland *et al*. 1998), which are polymorphic tandem repeats of one to six bases that occur ubiquitously throughout the genome (Tautz & Renz 1984). Microsatellites situated at critical coding regions expose the corresponding genes to frameshift mutations induced by mismatch (Malkhosyan *et al.* 1996) as presented for instance for *TGFBR2* (Wang *et al.* 1995), *BAX* (Ionov *et al.* 2000), and the MMR genes

themselves. The MSI phenotype is predominantly caused by an inactivation of the MMR gene *MLH1* or *MSH2* (Dietmaier *et al.* 1997), which in lung tumors occurs by promoter methylation (Wang *et al.* 2003; Hsu *et al.* 2005).

In contrast to the rather small DNA loci affected in the MSI types of tumors, CIN driven tumors display strikingly abnormal karyotypes. This instability is characterized by both chromosome and chromatid rearrangements, and losses and gains of whole chromosomes or chromosome segments (Lengauer *et al.* 1997; reviewed in Lengauer *et al.* 1998). Consequently, CIN driven tumors may also display altered ploidy, loss of heterozygosity (LOH), and uniparental disomy (UPD). The underlying causes of CIN, which are likely to be more heterogeneous than that of MSI, are thought to be related to mitotic defects that give rise to chromosome segregating errors (reviewed in Lengauer *et al.* 1998). Oncogens and TSGs that disrupt the normal function and numerical integrity of centrosomes have been suggested to be involved owing to the central role of centrosomes in chromosome segregation (Fukasawa 2007).

1.3 Genomic alterations

A copy number aberration may involve any size of chromosome material from micro level changes that are not visible using a regular light microscope to changes of whole chromosomes. A segment of a chromosome that is amplified many times may be fused in long arrays within a chromosomal segment forming a homogeneously staining region (HSR) or be cleaved out of the cell and form an autonomously replicating structure called a double minute (Weinberg 2007). Alternatively, a chromosome or chromosome fragment may be excluded from the nucleus if it lacks the ability to travel to the spindle poles during cell division. The lagging fragment may form its own micronucleus (MN), which resembles the main nuclei morphologically apart from the size (Fenech 2000). Also, dicentric chromosomes generated via translocations can end up in an MN through the formation of nucleoplasmic bridges (Fenech 2000).

In a number of cancers, gene copy number is known to have a significant impact on gene expression (Hyman *et al.* 2002; Pollack *et al.* 2002; Wolf *et al.* 2004). Particularly, highly amplified regions have been shown to induce gene expression changes in 20-62% of the corresponding genes (Hyman *et al.* 2002; Pollack *et al.* 2002; Wolf *et al.* 2004).On the other hand, the effects on gene expression of low level changes in copy number were shown to be less proportionate but this observation could also relate to the difficulty of

detecting slight increases in gene expression. A change in gene dosage is not always proportional to the alterations observed in gene or protein expression. Particularly, transcriptional regulation is influenced by chromatin structure and accessibility to DNA binding sites that control transcription, presence of transcriptional activators and repressors, and post-transcriptional modifications (Berg *et al.* 2001). Hence, apart from homozygous deletions that cause a lack of expression, gene or protein expression cannot be deduced directly from the DNA copy number.

1.4 Fragile sites

Fragile sites are specific chromosomal loci that preferentially exhibit breakage following a partial inhibition of DNA synthesis (Glover *et al.* 1984; Durkin and Glover 2007). The sites are generally categorized into common and rare types based on their population frequency and pattern of inheritance. At present, the genome database (www.gdb.org) contains 88 common fragile sites present in all individuals and 28 rare fragile sites, which appear in <5% of individuals and segregate in Mendelian manner (Kremer *et al.* 1991; reviewed in Sutherland *et al.* 1998). The breakage of the fragile sites is thought to be related to the presence of sequences that are difficult to replicate, owing, for instance, to secondary structure formation (Gacy *et al.* 1995; Hewett *et al.* 1998). Replication of these loci has been suggested to be especially complicated in the presence of replication inhibitors or with a deficiency in DNA repair and cell cycle checkpoint pathways (Casper *et al.* 2002; reviewed in Arlt *et al.* 2006).

The fragile sites have been implicated as being loci of frequent rearrangements also in cancers (Yunis & Soreng 1984; reviewed in Durkin and Glover 2007). Furthermore, LOH has been reported to occur preferentially at these sites in early lesions (Bartkova *et al.* 2005; Gorgoulis *et al.* 2005). Their significance in cancer is further highlighted by the fact that several fragile sites span genes that function as TSGs and stress responders (reviewed in Smith *et al*. 2006). For example, the most unstable fragile site, FRA3B at 3p14, lies within the tumor suppressor *FHIT* (Ohta *et al.* 1996), which has been shown to be inactivated in 60% of human cancers (reviewed in Pekarsky *et al.* 2002). Other highly breakable fragile sites include FRA16D at 16q23, FRA6E at 6q26, FRA7H at 7q23.3, and FRAXB at Xp23.3 (Glover *et al.* 1984). In addition to TSG inactivation, fragile sites may be of relevance in oncogene activation (Hellman *et al.* 2002).

1.5 Signaling pathways

In normal cells, virtually all aspects of cell behavior are regulated by the signals the cell receives from its surroundings. The extra-cellular signals are transferred to intracellular effectors through sequential activation of mediator proteins leading to responses also governing focal processes such as cell proliferation, differentiation, and apoptosis (Weinberg 2007). Aberrations of key genes cause perturbations in these systems enabling the cell to escape normal growth regulation (Anderson *et al.* 1992).

The signals are transmitted in modules of multilayered signaling cascades (reviewed in Pawson & Saxton 1999). The pathways exhibit cross-talk and redundancy (Weinberg 2007), building up an intricately connected signaling network. There are fewer signaling pathways than proteins, and one protein may be involved in several of them. The structure is extremely robust to alterations in general (Albert *et al.* 2000) but easily wounded when a highly connected component such as the TP53 is affected (Vogelstein *et al.* 2000). TP53 functions primarily as a transcription factor regulating processes such as apoptosis and DNA damage response (reviewed in Sengupta & Harris 2005) and is, as previously mentioned, one of the most commonly mutated genes in human cancers.

A signaling pathway can be deregulated in several different ways in tumors (Downward 2006). Interestingly, different cancer types accumulate mutations preferentially in specific genes of a given pathway, which reflects a functional redundancy of certain mutations (Ichimura *et al.* 2000; Eberhard *et al.* 2005; Suzuki *et al.* 2006). Mutually exclusive mutations have been recognized for instance in the RAS signaling cascades that link EGFR activation to cell cycle progression and survival (reviewed in Downward 2003). RAS family genes lie downstream of EGFR and interact with several families of effector proteins such as RAF and PI3K. *KRAS* mutations are found preferentially in tumors of the pancreas, whereas melanomas and ovarian cancers show a tendency for *BRAF* and *PI3K* mutations respectively (reviewed in Downward 2003). Lung adenocarcinomas (ACs) develop mutations both in *KRAS* and *EGFR*, but they are preferentially mutated in smokers and non-smokers respectively (Westra *et al.* 1993; Kosaka *et al.* 2004; Shigematsu *et al.* 2005). Similarly, the RB pathway can be inactivated through mutations of *RB1* itself, as in the majority of small cell lung cancers (SCLCs) but deregulation can be equally achieved, for example, through the inactivation of *CDKN2A* as shown in 30-70% of non small cell lung cancers (NSCLCs), or through the activation of *CCND1* or *CDK4* (Harbour *et al.* 1988; Shapiro *et al.* 1995; González-Quevedo *et al.* 2002; reviewed in Wikman & Kettunen 2006). The differences in preferential genetic alterations between tumors of various organs

and between lung ACs of smokers and non-smokers indicate that the microenvironment also plays a role in the advantageousness of a given alteration.

The complexity of the signaling systems enables malignant transformation to be driven by a combination of different sets of genes. Accordingly, the accumulation of mutations in a distinct group of genes as proposed by the earlier models, e.g. for the development of colorectal cancer (Fearon & Vogelstein 1990), have proven limited and alternative pathways for carcinogenesis have been implicated (Smith *et al.* 2002; Leslie *et al.* 2003) instead. Knowledge of the pathways and their affected components are essential in designing a tailored treatment for a given tumor patient and is also likely to facilitate diagnostics because different combinations of genetic alterations have an impact on the tumor phenotype (Conlin *et al.* 2005).

2 LUNG CANCER

Lung cancer (LC) is the leading category of cancer causing death in men and comes in third place in women. Overall, it accounts for 20% of deaths caused by cancer in Europe (Boyle & Ferlay 2005). The overall incidence of LC in the Finnish population has been declining among men since the early 1980's, but among women the incidence rate has remained stable since the beginning of the 2000's (Finnish Cancer Registry, www.cancerregistry.fi). The overall five year survival rate for all LC patients is less than 15% (Finnish Cancer Registry), but it is 60% for patients who show no evidence of lymph node (LN) metastases at the time of surgery. However, owing to lack of early detection methods and the asymptomatic early stage of cancer, 65% of patients are diagnosed with an advanced disease (Naruke 2003).

2.1 Histological classification of LC

LCs derive from the epithelial tissue of the central and peripheral lung, but represent a rather heterogeneous collection of tumors. They are commonly divided into two broad categories, SCLC and NSCLC. NSCLCs are further classified histologically into subcategories of which the largest ones are squamous cell carcinoma (SCC), AC, and large cell lung carcinoma (LCLC) (Travis *et al.* 1999). ACs and LCLCs usually arise in the periphery of the lung, whereas SCCs and SCLCs are more centralized. The clinical features, treatment and prognosis depend on the histological tumor type and the stage of the disease.

2.2 Risk factors for LC

Tobacco smoking is the greatest risk factor associated with all histological types of LCs (Lubin & Blot 1984). 90% of lung cancer deaths in men and about 80% in women are attributable to smoking (Thun *et al.* 2002). The lifetime probabilities of dying of lung cancer by age 85 are estimated to be 15% for male smokers and 8% for females (Thun *et al.* 2002). For lifelong non-smokers, the probability is about 1%. However, environmental tobacco smoke also increases the risk of LC (Hackshaw *et al.* 1997). Cigarette smoke contains more than 4000 chemical compounds of which over 50, including polycyclic aromatic hydrocarbons (PAHs) and tobacco specific *N*-nitrosamines, are known human carcinogens (Hoffman & Hoffman 1997). Carcinogen metabolic activation leads to covalent modifications in DNA, i.e. DNA adducts, which can lead to mutations if not repaired. Also reactive oxygen and nitrogen species (ROS/RNS) are produced, which are involved in causing various types of DNA and chromosomal damage including 8-hydroxy-guanine (8-OHdG) adducts, DNA cross-links and single and double strand breaks in DNA (Janssen *et al.* 1993; Jaurand 1997; reviewed in Marnett 2000). Genetic factors such as variations in the DNA repair capacity and in the ability to metabolize carcinogens, may affect an individual's susceptibility to cancer (reviewed in Schwartz *et al.* 2007).

In addition to tobacco smoke, occupational and environmental exposure to other carcinogenic factors such as asbestos, nickel chromates, radon and uranium contribute to the risk of lung cancer (Saccomanno *et al.* 1976; Mossman *et al.* 1996). Concurrent exposure to several factors, as in the case of tobacco smoking and asbestos, may act synergistically to induce cancer (Selikoff *et al.* 1968; Vainio & Boffetta 1994; reviewed in Lee 2001).

2.3 Genetic alterations in LC

The molecular mechanisms of lung carcinogenesis are complex and involve a large number of genetic changes. The tumors may exhibit multiple abnormalities of both chromosome number and structure (reviewed in Testa *et al.* 1997), which can affect any chromosome and are mainly unbalanced. Other common features include ploidy alterations in about 50% of LCs (Mitelman *et al.* 2007), LOH, promoter methylation (Zöchbauer-Müller *et al.* 2001; Liu *et al.* 2007), isochromosomes and double minutes (Nielsen *et al.* 1993; reviewed in Testa *et al.* 1997). Altogether, the findings indicate high genomic instability, specifically CIN, which also shows as an increase in aberrations with tumor development. Additionally,

although MSI of multiple loci has been reported to occur only in 1-10% of all LCs, MSI of individual markers appear more frequent (Sekido *et al.* 1998).

A recurrent trend of aberration has been revealed for most chromosome arms in lung cancers. Irrespective of LC category, losses are frequently detected at 3p, 5q, 6q, 8p, 9p, 13q, 17p, and 18q, whereas 1q, 5p, 7p, 8q, and 20q are commonly gained (Garnis *et al.* 2006; Balsara & Testa 2002; Coe *et al.* 2006; Weir *et al.* 2007). However, the functional roles of the genes amplified and deleted in LC are still mostly unknown. Known target genes include TSGs *FHIT* at 3p14 (Sozzi *et al.* 1996), *CDKN2A* at 9p23 (Hayashi *et al.* 1994), *TP53* at 17p13 (Takahashi *et al.* 1989), and *LKB1* at 19p13 in ACs (Sanchez-Cespedes *et al.* 2002) in addition to oncogenes *MYC* at 8q24 (Little *et al.* 1983), *EGFR* at 7p12 (Merlino *et al.* 1985), and *KRAS* at 12p12 (Pulciani *et al.* 1985). It is, however, possible that the regions harbor several genes that contribute to tumorigenesis. Several aberration hot spots have been identified for instance at the 3p region, the most common site of aberration in LCs (Hibi *et al.* 1992; Balsara & Testa 2002), namely 3p25-26, 3p21.3, and 3p12-14 (Whang-Peng *et al.* 1982; Kok *et al.* 1997; Zabarovsky *et al.* 2002).

Although different histological types of LCs follow the same general overall pattern of losses and gains, distinct chromosomal regions have been shown to be preferentially aberrated in certain types (Petersen *et al.* 1997; Luk *et al.* 2001; Balsara & Testa 2002; Coe *et al.* 2006; Garnis *et al.* 2006). For example, LOH of 3p has been reported in 69% of SCCs and in 35% of ACs (Mitsudomi *et al.* 1996), whereas the region has been found to be lost in 90% of SCLCs (Balsara & Testa 2002). Similar findings have been presented for multiple other sites, e.g. 1p, 6q, 8p, and 19p (Balsara & Testa 2002; Coe *et al.* 2006; Garnis *et al.* 2006). Differences between the histological types of LCs are also evident at the level of single genes and proteins, which is illustrated here by three examples. *TP53* is mutated in about 90% of SCLCs but only in about 50% of SCCs and 40% of ACs (Tammemagi *et al.* 1999; Le Calvez *et al.* 2005). Additionally, mutations of the EGFR/RAS pathway genes *KRAS* and *EGFR* are observed in NSCLCs, especially ACs, but infrequently in other LC types (Kosaka *et al.* 2004; Shigematsu *et al.* 2005). The Fhit protein has been found to be lost in about 90% of SCCs, 60% of ACs, and 40% of SCLCs (Sozzi *et al.* 1998).

Changes in LC related to smoking

Differential changes have been recognized in tumors with and without a history of smoking. Exposure to tobacco smoke may induce specific changes early in carcinogenesis

as LOH of 3p has been detected already in the normal epithelium of smokers without lung cancer (Mao *et al.* 1997; Wistuba *et al.* 1997). In LCs, 80% of tumors from current smokers, but only 22% of tumors from patients who have never smoked show LOH at 3p (Sozzi *et al.* 1997). Furthermore, *TP53* mutations of smokers differ from those in non-smokers in terms of frequency, hot spots and tendency for distinct transversions (Gao *et al.* 1997; reviewed in Bennett *et al.* 1999; Gealy 1999; Husgafvel-Pursiainen *et al.* 2000; Le Calvez *et al.* 2005). *KRAS* mutations, on the other hand, are almost exclusively detected in the ACs of smokers (Westra *et al.* 1993; Gealy *et al.* 1999; Ahrendt *et al.* 2001), whereas *EGFR* mutations occur particularly in the ACs of non-smokers (Kosaka *et al.* 2004; Pao *et al.* 2004; Shigematsu *et al.* 2005). Tumors of smokers and non-smokers also display differences in the methylation patterns of TSGs including *CDKN2A* and *RASSF1A* (Zöchbauer-Müller *et al.* 2001; Toyooka *et al.* 2006).

Changes in LC related to asbestos

Another significant risk factor for LC is asbestos. The asbestos fibers enter the body by inhaling or swallowing and are capable of penetrating into the lung, where they are thought to interact with the lung epithelial cells directly and indirectly (Mossman *et al.* 1997). The length, texture, and chemical properties of the fibers contribute to their toxicity. The exact carcinogenic mechanisms are not fully known, but they are thought to include the formation of ROS and RNS, alterations in mitochondrial function, and physical disturbance to the cell cycle (Mossman *et al.* 1997; Shukla *et al.* 2003; reviewed in Upadhyay & Kamp 2003). Asbestos fibers cause DNA breakage and chromosomal damage (Adachi *et al.* 1994; Kamp *et al.* 1995; Ollikainen *et al.* 1999; Levresse *et al.* 2000). They also induce changes in several signaling pathways (Nymark *et al.* 2007) such as the mitogen activated protein kinase (MAPK) and nuclear factor kappaB (NF-κB) cascades (Mossman *et al.* 1997, 2006; Janssen *et al.* 1995). Particularly, MAPK lies downstream of EGFR, which receptor has been shown to be activated by oxidative stress induced by asbestos (Zanella *et al.* 1996; Mossman *et al.* 1997). Only limited information is, however, available of the putative preferential genetic alterations that may be involved in asbestos related carcinogenesis. Although extensive LOH of 3p21 (Marsit *et al.* 2004) and mutations of *TP53*, *KRAS*, and *FHIT* have been suggested to be related with asbestos exposure (Wang *et al.* 1995; Nelson *et al.* 1998, 1999; Husgafvel-Pursiainen *et al.* 1999; Pylkkänen *et al.* 2002), the effects induced by asbestos have been difficult to decipher. The fact that most asbestos exposed

LC patients have a definite history of smoking hinders the uncovering of asbestos specific effects.

3 MICROARRAYS IN CANCER RESEARCH

Recent advances in genomics, most importantly the sequencing of human and other organisms' genomes, have provided revolutionary information that enables the characterization of diseases at the molecular level. This data is elaborately exploited by DNA microarrays (Schena *et al.* 1995), i.e. miniature measurement devices that contain probes for practically all genes of the study subject and enable genome wide studies of a given subject in a single assay. Soon after the introduction of the first arrays for gene expression analysis (Schena *et al.* 1995), several different modifications of the high-throughput platform have been introduced, including array comparative genomic hybridization (array CGH) for DNA copy number (Solinas-Toldo *et al.* 1997; Pinkel *et al.* 1998), single nucleotide polymorphism (SNP) array for genotyping (Wang *et al.* 1998), and the methylation array for epigenetic (Hatada *et al.* 2002) assays. Special platforms have also been designed for protein and cell studies.

DNA microarrays contain nucleotide probes at predefined positions, which, upon hybridization, bind preferentially with their complementary single stranded nucleotide sequences. With known probe sequences, genome wide screening studies can be carried out without *a priori* knowledge of the affected genes in the condition of interest. The probes are typically either oligonucleotide (Carvalho *et al.* 2004), cDNA (Pollack *et al.* 1999) or bacterial artificial chromosome (BAC) based sequences (Solinas-Toldo *et al.* 1997). Oligonucleotides provide the best resolution due to the short length of oligos and their independency of coding sequences, whereas longer BACs provide more intense signals (Pinkel & Albertson 2005). Current arrays may even contain more than two million probes (NimbleGen HD2 Arrays, http://www.nimblegen.com).

Depending on the array platform, all samples are either hybridized to separate arrays (one-color array) or each sample and the corresponding reference are hybridized together to a single array (two-color array). Two-color platforms utilize competitive hybridization of the sample and a differently labeled reference (Schena *et al.* 1995), with Cyanine3 (Cy3) and Cyanine5 (Cy5) being the most commonly used dyes. On the other hand, biotin labeled samples are used in the Affymetrix one-color GeneChip® system (Affymetrix 2004). Owing to differences in platform design, the sensitivity, specificity and reproducibility

obtainable by various array types varies. Thus considerable divergence of the results across platforms has been previously reported (Tan *et al.* 2003; Marshall 2004). The issue has been recently addressed by a collaborative MicroArray Quality Control (MAQC) project (MAQC Consortium 2006). On the basis of data from seven different microarray platforms and about 60 hybridizations per platform, intra- and inter-platform comparability of the results could be demonstrated (MAQC Consortium 2006).

3.1 Microarray data processing

In the experimental setup, DNA or reverse transcribed mRNA of the sample of interest or reference is labeled and allowed to hybridize with the probes on the array. After hybridization, the microarray is scanned with a laser that excites the fluorescently labeled sample (and reference). A digital image is formed where the intensities reflect the expression or copy number of the genes. Image analysis techniques are applied to locate, segment, and quantify the spot intensities. At the time of segmentation and intensity quantification, the quality of the measurements can be assessed using the spatial features of the spot, e.g. its shape or size, or by comparing the spot foreground and background intensity distributions (Ruosaari & Hollmén 2002; Wang *et al.* 2001; Hautaniemi *et al.* 2003; Li *et al.* 2005). Steps in the further analysis cannot compensate for possible sources of error such as scratches on the slide and therefore image quality needs to be assessed at this stage.

Prior to further analysis, data preprocessing or normalization is usually performed (reviewed in Quackenbush 2002). Normalization methods generally calculate a scaling factor or function to correct for nonbiological effects in the data. No standard procedures have been defined, and therefore the methods are currently chosen to suit each individual study. For two-color microarrays, the method can involve simply adjusting the median and variance of the logarithmic ratios of the sample and reference across the dataset to obtain similarly scaled data distributions. Locally weighted scatterplot smoothing (LOWESS) (Cleveland 1979) has become a commonly used option in cases where there is, for example, a need to adjust for the dye bias. To enhance the comparability of the measurements between the arrays of one-color Affymetrix slides, robust multi-array average (RMA) (Irizarry *et al.* 2003), and its derivations such as GC-RMA (Wu *et al.* 2004), have become popular. Recently, the standardization of microarray experiments has been addressed and the Minimum Information About a Microarray Experiment (MIAME) has been established. It describes the information needed to enable the results of an experiment to be interpreted

unambiguously and to allow the reproduction of the experiments (Brazma *et al.* 2001, 2003; Barrett *et al.* 2005).

Microarray data are commonly presented as a logarithmic ratio of the measurements between the sample and the reference. Various statistical approaches are available and can be used to identify changes associated with a specific outcome such as survival or to reveal global patterns without regard to a given outcome. Comparative studies may be advantageous when distinct study populations such as different histological types of tumors of a given organ or tumors of patients exposed and not exposed to a carcinogen, are investigated. On the other hand, subclasses of a seemingly uniform group of specimens may also be sought on the basis of their aberration profiles.

The earliest and simplest methods use thresholds to detect changes in gene expression or copy number, or apply statistical hypothesis testing such as detecting differences in the means between the conditions of interest with a t-test. Another approach for analyzing the array data is the Receiver Operating Characteristic (ROC) curve (Swets 1988), which provides a non-parametric approach for analyzing the diagnostic value of the probes in a two-group classification setting. The curve displays the relationship between the proportion of true and false positive classifications, thus the area under the curve yields an estimate of a correct diagnosis when the probe is used to classify the groups. Past studies have, however, indicated that gene by gene analyses produce discordant results (Cahan *et al.* 2005) in apparently similar studies (e.g. Bhattacharjee *et al.* 2001; Garber *et al.* 2001; Beer *et al.* 2002; Takeuchi *et al.* 2006). Thus, more sophisticated methods have been developed, that integrate different types of data including measurements of DNA, mRNA and protein levels from multiple studies (Hanash 2004).

3.2 Pathway analysis of microarray data

One commonly applied data integration approach is to group genes into biologically meaningful categories and test the categories for deregulation or enrichment of differentially expressed genes. Categories with biological relevance could represent, for instance, signaling pathways discussed earlier in this thesis or genes that are involved in the same cellular processes. Investigations of groups of genes are justified as a change in the expression of one gene or protein is not an independent event but also influences the expression of other genes or proteins through the signaling cascades. Indeed, the

deregulation of a signaling cascade cannot be evaluated based on studies of isolated genes and therefore holistic studies of the pathways should be performed (Vogelstein *et al.* 2000).

Whilst high-throughput protein expression measurement techniques are still evolving, pathway analyses are frequently performed on gene expression data. Yet, pathway analyses have also been performed on array CGH data (Kaur *et al.* 2006). Analyses of distinct groups of genes instead of single genes are practical as moderate changes in a number of components of one pathway may be enough to indicate differential regulation of the whole pathway. On the contrary, single gene analyses, e.g. fold change or t-test, require significant differences in the expression of a gene between the conditions of interest for the change to be detected. Furthermore, it has been shown that the consistency of analyses across independent data of a similar kind are remarkably improved by analyzing a defined set of genes that share a biological function, chromosomal location or regulation instead of single genes (Subramanian *et al.* 2005). However, to ensure that a specific finding is significant in the condition of interest, further verification of the results by studies on independent populations is always needed (Hayes *et al.* 2006).

Gene Ontology

The Gene Ontology (GO) project is a collaborative effort that aims at consistent descriptions of gene products in different databases (Ashburner *et al.* 2000; Gene Ontology Consortium 2006). The project has developed controlled vocabularies for describing the biological processes, molecular functions and cellular locations of genes and gene products in a species independent manner. Information about the deregulated pathways can be obtained by testing GO categories for the enrichment of differentially expressed genes. A common choice is to apply the hypergeometric test to assess the representation of genes with a given GO annotation relative to the whole data (Breitling *et al.* 2004; Subramanian *et al.* 2005). The enrichment calculations can be performed on a predefined list of putative, differentially expressed genes or on rank lists of differentially expressed genes which include the whole data (Breitling *et al.* 2004).

The GO terms can be ordered as a directed acyclic graph (DAG), where detailed terms branch off from the more general terms. The detailed “child” terms are specific descriptions of the more general “parent” terms indicating that a gene with a detailed GO description can also be described by the more general terms of the branch. Figure 1 illustrates the connections of the GO terms for biological processes related to protein ubiquitination, a

process of relevance to this thesis. The figure shows that ubiquitination processes may be described by multiple levels of specificity, which is also a feature of other GO categories.

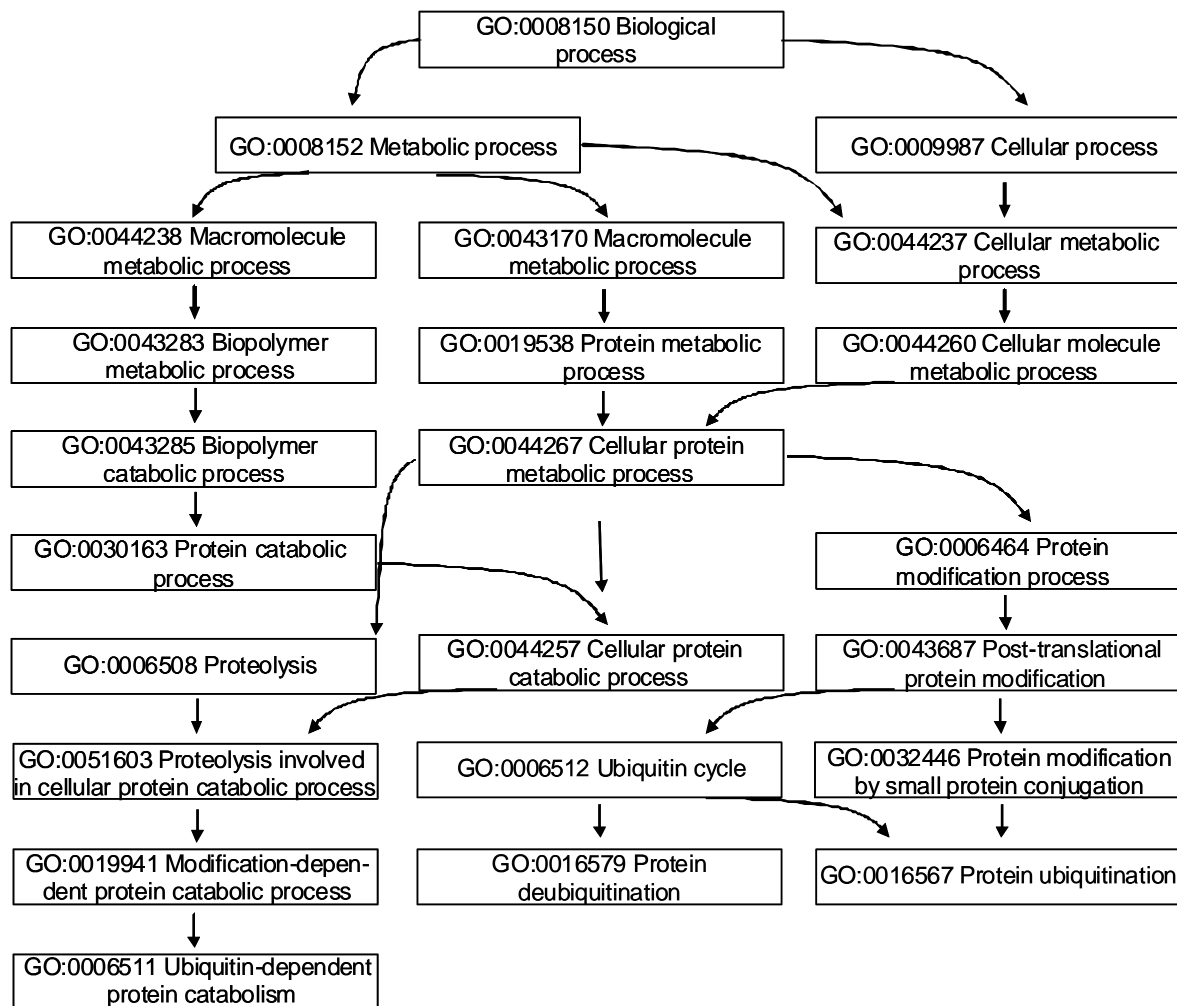

Figure 1. Gene Ontology illustration of biological processes related to protein ubiquitination. The GO DAG branches off from the most general term (GO:0008150) and provides several levels of specificity to describe the ubiquitination process. Most GO terms have also child terms not shown in this figure due to limited space.

3.3 Combining array CGH and gene expression

Although changes in gene expression can occur through various mechanisms as presented in Section 1.1, recurring DNA level aberrations are thought to target specific genes that contribute to tumorigenesis. The microarray technology provides a powerful means for detecting genes whose expression is affected by the copy number change and which may thus induce changes at the cellular level also. Indeed, various approaches have been presented for the integration of high-throughput array CGH and gene expression data (Pollack *et al.* 2002; Hautaniemi *et al.* 2004; Myers *et al.* 2004; Furge *et al.* 2005; Berger *et*

al. 2006). The approaches generally fall into three categories that share the following properties. In the first category, the two data types are analyzed separately and the results are combined to identify regions displaying both DNA and expression level changes (Pollack *et al.* 2002; Myers *et al.* 2004; Furge *et al.* 2005). In the second category, the regions with copy number changes are determined, enabling the detection of genes whose expression levels can be explained by the copy number change (Hautaniemi *et al.* 2004). In the third category, the array CGH and expression data are combined as such using transformation algorithms, after which the data projections that correspond to similar patterns of variations are sought (Berger *et al.* 2006).

3.4 Considerations related to microarray analyses

Microarray data analysis typically requires the testing of a large number of hypotheses. For instance, when differentially expressed genes are sought, a statistical test is commonly performed for all the probes on the array. The testing procedure requires the formulation of the null and the alternative hypotheses, which are commonly phrased so that the null represents no change and the alternative is a real effect, and a test statistic that is able to provide an answer to the problem under consideration is used. The test yields a p-value, which is the probability of obtaining a value of the test statistic that is at least as extreme as the one observed under the circumstances that the null hypothesis were true.

Based on the outcome, the null hypothesis is either rejected in favor of the alternative hypothesis or it is accepted at the level of significance of the test. Two types of errors are associated with the testing: the null hypothesis may be erroneously rejected (Type I) or the null hypothesis may not be rejected even if it is false (Type II). The level of significance is specified by setting an acceptable maximum probability of making a Type I error, which depends on the study setting and, owing to an inverse relation between the error types, also affects Type II errors.

When multiple hypotheses are tested, the probability that some Type I errors are committed increases with the number of hypotheses (Shaffer 1995). Hence, when a cutoff of 0.05 is used to detect over and underexpressed genes from the high-throughput microarray data, there is a 5% chance of a Type I error. Thus hundreds of the so called significant findings are expected to have come up by chance alone. Several multiple test correction procedures such as Bonferroni or Bonferroni-Holm corrections (Holm 1979; Benjamini & Hochberg 1995) have been developed. In the context of gene expression analysis, approaches that

adjust the p-values by multiplying them with the number of performed tests may, however, provide too stringent a control. However, the significance of the results can also be assessed using permutation testing (Good 2000). An empirical probability distribution can be constructed for any test statistic by randomizing the observed data points, calculating the outcome, and repeating the process an adequate number of times until a stable distribution is obtained.

Another consideration regarding gene expression microarrays relates to the use of a limited number of samples. As a result, the identified molecular signatures, i.e. the subset of genes found to be most differentially expressed in patients with different outcomes, are unstable and depend on the selection of the patients in the training set (Michiels *et al.* 2005). Also flaws in the statistical design of the experiments (Dupuy & Simon 2007), variability due to the development stage of the tumor (Beer *et al.* 2002), differences in patient cohorts and inadequate validation of the findings (Michiels *et al.* 2005) can cause variability in the expression signatures of seemingly similar study populations. Studies on lung ACs by Bhattacharjee *et al.* (2001), Garber *et al.* (2001), Beer *et al.* (2002), and Takeuchi *et al.* (2006) illustrate this variability by showing a lack of consistency in survival gene expression signatures. However, it is possible that several signatures consisting of different genes from distinct pathways produce equal classification capability.

Also array CGH has its limitations. Although the aberration breakpoints are powerfully revealed by array CGH, other factors such as balanced translocations, UPD, ploidy alterations and micro-level copy number transitions beyond the resolution of the platform cannot be detected using this system.

3.5 Microarrays in lung cancer

As described in Section 2.3, previous studies using chromosomal CGH and LOH analyses have indicated that lung tumors harbor specific gains and losses. Array CGH analyses have confirmed the trend of recurring gains and losses of the chromosome arms (Tonon *et al.* 2005; Coe *et al.* 2006; Garnis *et al.* 2006; Weir *et al.* 2007). In addition to confirming previous findings, the high-resolution techniques have revealed novel regions of frequent gains and losses overlooked by the gross analysis methods. For instance, Weir *et al.* (2007) revealed 31 recurrent focal aberrations of which only six were associated with a known mutation in LCs. Similarly, Tonon *et al.* (2005) identified 93 focal copy number alterations. Also, comparative microarray studies of LCs that represent different histological types have

uncovered genetic alterations prevalent in different LC subtypes (Coe *et al.* 2006; Garnis *et al.* 2006). On the basis of the recurrent amplification and deletions identified by array CGH, a few candidate tumor genes have been identified (Tonon *et al.* 2005; Weir *et al.* 2007). However, the majority of the genes targeted by the aberrations were recently suggested as being undiscovered (Weir *et al.* 2007).

On the gene expression side, various approaches have been taken to identify genes that are differentially expressed in LCs and likely to be of relevance in lung carcinogenesis. For instance, gene expression signatures that distinguish lung tumors from normal lung (Dracheva *et al*. 2007), separate different histological types of tumors (Coe *et al.* 2006) or that partition patients into prognostic groups (Bhattacharjee *et al.* 2001; Garber *et al.* 2001; Beer *et al.* 2002; Takeuchi *et al*. 2006) have been presented. Depending on the study, the presented signatures may include as few as three genes (Lau *et al*. 2006) but may also contain more than 100 genes (Coe *et al*. 2006) that show differences in expression between the studied subgroups.

Translation of the findings into clinical practice has, however, been problematic owing to the low concordance of the results. One illustration of this was given by Lau *et al*. (2006), who compared the various prognostic gene expression signatures that had been presented previously for the classification of NSCLC patients. The study showed that the prognostic gene lists presented in different reports had minimal overlap with only one common gene that had been identified in four separate studies. Recent results are, however, more promising and show that microarrays are of value in LC research. Reproducible LC subtypes could be revealed in a study involving multiple independent patient cohorts with the subtypes showing correlation to clinically relevant covariates such as stage specific survival (Hayes *et al.* 2006).

AIMS OF THE STUDY

The identification of molecular changes related to LCs in two separate themes have been studied for this thesis. The themes are changes induced by asbestos exposure and those that disseminate into bone marrow (BM) in the early stages of tumorigenesis. Additionally, methods were developed for microarray data analysis.

The specific aims of this thesis were:

- To study whether an additional DNA binding dye can aid in the image analysis of gene expression microarrays (I)
- To develop and apply microarray methods in the comparative analyses of distinct LC subgroups, i.e. those associated with asbestos exposure and those that disseminate into BM in the early stages of tumorigenesis (II-VI)
- To validate the key findings of the screening using independent methods and thereby prove the efficacy of the comparative study approach (III-VI)

MATERIALS

1 STUDY SUBJECTS

Tumors of two distinct groups of LCs were investigated: lung tumors of patients with 1) a history of asbestos exposure (Publications II-V) and 2) evidence of disseminated tumor cells (DTCs) in BM (Publication VI). Lung tumors of patients not exposed to asbestos and without DTCs in BM were, respectively, used as a reference in the respective analyses. The study protocols had been approved by ethics review boards and the permission to use diagnostics samples had been granted.

1.1 Asbestos related studies (II-V)

Tumors of asbestos exposed and non-exposed LC patients, who were matched for smoking history, histological tumor type and stage of disease, were screened for differential aberrations using microarray methods (Table 1). The main findings were verified on larger study populations. Representatives of different histological lung tumor types were included as the goal was to investigate asbestos exposure associated aberrations in LC in general.

Three exposure groups were studied. These were heavily asbestos exposed (referred to as asbestos exposed in text), moderately exposed, and non-exposed. The samples were assigned to adequate groups based on their pulmonary fiber counts per gram of dry lung tissue being above 5 million, between 1 and 5 million, and below 0.5 million for the three groups respectively (Karjalainen *et al.* 1994).

1.2 Studies related to disseminating LCs (VI)

Lung tumors of patients, with and without evidence of DTCs in their BM, were screened for differential aberrations using microarray methods and the main findings were verified on a larger study population (Table 2). In the microarray analyses, primary AC specimens were studied, whereas samples of different histological LC types were studied in the subsequent verifications. Brain metastases of LCs were also analyzed in the subsequent studies. The BM status of the patients was determined by means of immunocytochemical staining (Pantel *et al.* 1994).

Table 1. Characteristics of lung tumors used in Publications II-IV.

		CGH		Affymetrix		FISH			Microsatellite		
		Asbestos-exposed n = 11	Non-exposed n = 9	Asbestos-exposed n = 14	Non-exposed n = 14	Asbestos-exposed n = 28	Non-exposed n = 25	Moderately asbestos-exposed n = 5	Asbestos-exposed n = 25	Non-exposed n = 29	Moderately asbestos-exposed n = 8
Age	median	63	67	63.5	65	65	64	64	64	65	69.5
	(range)	(57 – 67)	(41 – 72)	(57 – 67)	(41 – 72)	(53 – 85)	(40 – 81)	(58 – 78)	(53 – 79)	(36 – 81)	(47 – 79)
Asbestos fiber count [I]	median	10.8	0.0	11.7	0.0	10.8	0.0	2.1	12.8	0.0	2.3
	(range)	(5.9 – 90)	(0.0 – 0.5)	(5.9 – 145)	(0.0 – 0.5)	(5.1 – 184)	(0.0 – 0.5)	(1.3 – 4.6)	(5.9 – 8000)	(0.0 – 0.5)	(1.2 – 4.3)
Histology	AC	4	4	5	6	11	10	2	9	12	4
	SCC	4	2	4	4	7	9	2	5	11	3
	LCLC	3	2	3	2	5	2	--	6	2	1
	SCLC	--	--	1	1	3	3	1	1	2	--
	Others [II]	--	1	1	1	2	1	--	4	2	--

[I] Pulmonary asbestos fiber count in million per gram of dried lung. [II] Other types include adenosquamous, giant cell and pleomporhic carcinomas.

Table 2. Characteristics of lung tumors used in Publication VI.

		CGH		Affymetrix		FISH		
		BM-positive n = 14	BM-negative n = 16	BM-positive n = 7	BM-negative n = 9	BM-positive n = 22	BM-negative n = 21	Brain metastases n = 36
Age	median	65	59	67	63	58	65	
	(range)	(51 – 75)	(49 – 78)	(55 – 75)	(49 – 76)	(48 – 81)	(37 – 78)	
Histology [I]	AC	14	16	7	9	8	14	14
	SCC	--	--	--	--	14	6	10
	LCLC	--	--	--	--	--	1	0
	SCLC	--	--	--	--	--	0	7

[I] Histology information is missing for five brain metastases

2 MICROARRAY ANALYSIS METHODS

2.1 Three-color image analysis (I)

In this study, an additional dye, SYBR green RNA II, was used to improve the assessment of spot quality in gene expression microarrays. SYBR green RNA II is a fluorescent dye with high detection sensitivity and binds non-specifically to all DNA molecules including ssDNA. If a microarray slide is stained with SYBR, all spots that have nucleotide probes become labeled with SYBR in contrast to the sample labeling dyes that label spots depending on the amount of particular mRNA in the sample.

As this study contributes to image analysis, the experimental procedures are not described. The microarray experiment was performed using cDNA arrays consisting of 16,000 human cDNAs that have been spotted in duplicates (Human 16K slides, Turku Centre for Biotechnology, University of Turku, Finland). After the images corresponding to the sample and reference had been obtained, the microarray was stained with SYBR green RNA II and rescanned to obtain the SYBR image.

In our approach, microarray image segmentation was performed on the SYBR image and the results were applied to the corresponding Cy3 and Cy5 images. The spots were analyzed separately and each spot was assumed to be located within a predetermined size of pixel block around its center coordinates, which had been obtained from pre-analysis using QuantArray (Packard Biosciences Technologies). The block size was adjusted to include both spot and local background pixels, and to allow for some imprecision in the QuantArray coordinates.

The following approach was applied to segment the spots and to assess the quality of the spots: 1) Pixels with extreme values, here regarded as those with intensities more than three times higher than the average of row maxima, were removed. 2) An intensity histogram was formed and smoothed with a Gaussian filter of length 15 and standard deviation 0.4. 3) Pixels with intensities within the range of the higher intensity were assigned as peak, as putative foreground pixels. 4) The largest connected component of foreground pixels was assigned as the spot foreground. A visual representation of the spot segmentation is shown in Figure 2.

To identify faulty spots, the shape of the intensity histogram and spot features were assessed. Spots with one-peaked intensity histograms, with more than one unconnected

region of foreground pixels of size larger than 50 pixels were labeled faulty. Additionally, spots larger than 1300 or smaller than 350 pixels or with background borders of less than 180 pixels were labeled faulty. The spots that passed the quality inspection were quantified for foreground and background intensities in the Cy3 and Cy5 images by using the segmentation results of the SYBR image. For evaluation of the SYBR based faulty spot detection, the results were compared with those obtained using GenePix Pro 6.0 (Axon Instruments Inc.) with the irregular gridding feature.

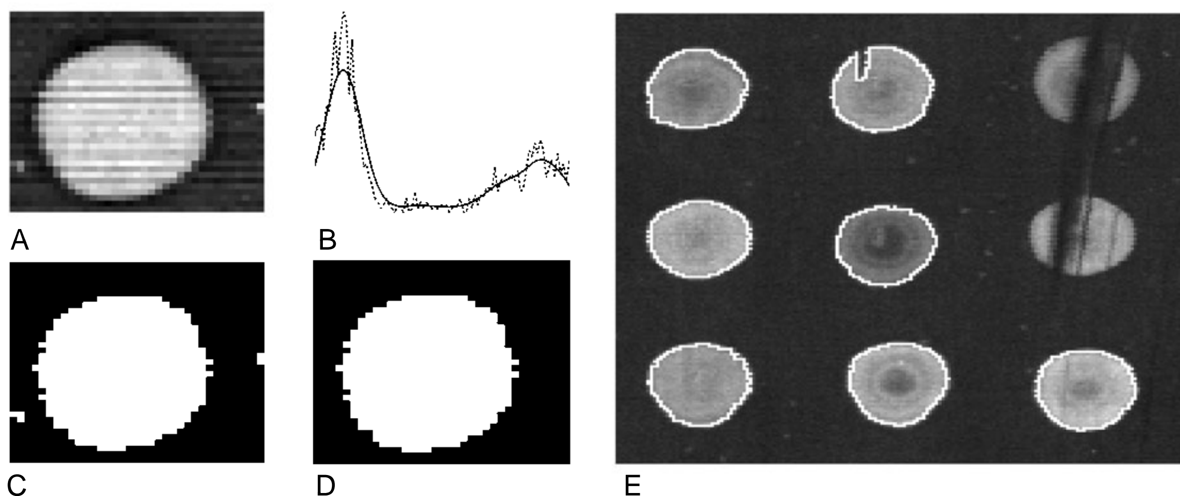

Figure 2. (A) Example of the spot on gene expression array. (B) An intensity histogram was formed from the remaining pixels (dashed line) and the histogram was smoothened (solid line). (C) If the histogram was two peaked, pixels with intensities in the range of the higher intensity peak were assigned as putative foreground pixels (white), while the pixels with lower intensities were assigned as background (black). (D) The largest connected component of foreground pixels was detected and assigned as the spot foreground. Spots with several large connected components or abnormal size were discarded. (E) The borders of the segmented spots are shown (white). Spots at top right did not pass the quality inspection.

2.2 Array CGH (II, IV & VI)

In Publication II, array CGH was performed using the Human 1.0 cDNA system (Agilent Technologies, Palo Alto, CA, USA) on 20 LC samples (Table 1). In Publication IV, 5 asbestos exposed and 5 non-exposed LCs were selected for high resolution copy number profiling using oligonucleotide arrays (Human Genome CGH Microarray Kit 44B, Agilent). Measurements detected as unreliable by Feature Extraction (Agilent Technologies), i.e. the signal was saturated or the spot was not uniform, were removed from further analyses. Additionally, owing to a need for further removal of morphologically ununiform spots from the cDNA arrays, our own image analysis method (Ruosaari &

Hollmén 2002) that discards spots based on the same morphological features as the three-color analysis presented in Section 2.1 of the Results, was applied. After preprocessing, the data from each tumor sample was scaled against a common reference consisting of a pool of normal samples, log2 transformed and LOWESS normalized (Cleveland 1979; Yang *et al*. 2001).

In Publication VI, array CGH was performed using an oligonucleotide based array platform (van den Ijssel *et al.* 2005) on 30 lung ACs (Table 2). Measurements with a quality flag <1 or the Confidence value <0.1, given by BlueFuse (BlueGnome, Cambridge, USA), and morphologically ununiform spots defined by our own image analysis (Ruosaari & Hollmén 2002) were removed from further analyses. The data from each patient were scaled with a common reference consisting of a pool of normal samples to obtain an equal median and variance for the middle 80 percentiles of autosomes and the log2 transformed.

2.3 Gene expression microarrays (III, V & VI)

In Publication III, the gene expression array experiment was performed using Affymetrix HG-U133A GeneChips (Affymetrix, Santa Clara, CA, USA) on 28 LC samples (Table 1) and corresponding normal samples, which were used as a reference. Probe sets with present calls by the Affymetrix MAS 5.0 software (p-value < 0.04) in at least one third of the exposed or non-exposed samples were included in subsequent analyses. In Publication V, the data was preprocessed using the GC-RMA (Wu *et al.* 2004) and log2 transformed.

In Publication VI, gene expression array experiments were performed using Affymetrix HG-U133 Plus 2.0 GeneChips (Affymetrix) on 16 lung ACs (Table 2). The data was preprocessed using the GC-RMA and log2 transformed. The median signal of three normal patients was used as a reference.

Owing to problems at the probe level in the early annotation files provided by Affymetrix (Dai *et al*. 2005), the probe sequences were re-annotated in Publication III by mapping the probe sequences against human genome build 34. In Publications VI and V, the annotation information was extracted from updated Affymetrix annotation files, where the probe positions were given with regard to human genome build 35.

2.4 Differential region finding method (II, III & VI)

A DRF method was developed in Publication II to identify chromosomal regions that are preferentially aberrated in either the asbestos exposed or non-exposed group of patients from array CGH data. Since the DRF was applied successfully in Publication II, it was also used in Publications III and VI. In these publications, the DRF was also applied to gene expression data. As an input, the DRF takes genome wide data from two conditions, e.g. microarray data from tumors of the asbestos exposed and non-exposed or BM-positive or BM-negative patients. The method detects regions that show preferential copy number or gene expression level alterations between the conditions and gives the regions as an output. The phrase preferential alteration covers changes that are specific to either of the compared conditions or occur more frequently in either of the conditions.

The DRF compares data from two conditions at group level in overlapping segments of 0.5 to 1 Mbp, which are tested sequentially for differences between the conditions. In practice, a sliding window is used, which moves from the p-arm of each chromosome towards the q-arm. The window is stopped at the locus of each probe and the probes within a 0.5 Mbp and 1 Mbp window defined by the probe locus are detected. Both window sizes are used owing to differences in probe distribution along the chromosome and to enable detection of focal aberrations at probe rich sites. However, a window needs to contain at least five probes to be included in the testing. The tips of the chromosomes are analyzed by the same procedure but probes will obviously be detected on the centromeric side of the window.

Next, the performance of the segment in differentiating the given conditions is assessed. Each probe is used in a classifier, where the minimum number of incorrectly classified patients is determined. The average number of patients correctly classified by the probes in the segment is calculated and used as a test statistic. The regions with differences between the compared groups are identified by hypothesis testing. In the two-tailed testing, the null and the alternative hypotheses were set as “the two groups do not display differences in the region" and “the two groups display differences in the region” respectively. The regions likely to be associated with the condition of interest are identified using a permutation test with 10,000 permutations of the group labels (Good 2000). The empirical p-values are calculated by comparing the test statistic to the permutation distribution. Multiple hypothesis correction is not performed because the method is only used to identify regions that, according to the p-values, are most likely to have relevance in the condition studied. The relevance of the findings is verified by using independent laboratory techniques.

2.5 ROC curves (III & VI)

Differentially expressed genes were identified using ROC curves (Swets 1988) by measuring the diagnostic value of the probes of the tumors of the asbestos exposed and non-exposed, and the tumors of the BM-positive and BM-negative patients. The ROC curve was constructed by sweeping the threshold used to distinguish the conditions to obtain the fraction of true positives as a function of false positives. Here the area under the curve corresponding to the probability of correctly diagnosing a patient by the gene expression measures was used as the test statistic. For evaluation of the significance of the findings, empirical p-values were calculated by comparing the test statistic to the permutation distribution using 10,000 permutations of the group labels.

In Publication VI, the p-values obtained from the ROC analysis were also used to determine whether differentially expressed genes ($p < 0.05$) were enriched at specific loci. The number of differentially expressed genes found at a locus was determined and compared to 10,000 random loci containing the same amount of genes as the test region.

2.6 GO analysis (V)

The gene expression data was combined with GO vocabulary to detect molecular functions, biological processes and cellular locations that are differentially regulated between specimens from asbestos exposed and non-exposed lung cancer patients. Data from both the normal and tumor samples were assessed for both up- and down-regulated pathways (GO terms). Custom made analysis methods and Iterative Group Analysis (iGA, Breitling *et al.* 2004) were used.

First, the genes were rank ordered with respect to differences in gene expression between the two groups of patients. Both p-values from the t-test and the fold-change between the medians of the two groups were used in performing the ranking, thus two rank lists were obtained. iGA was then applied to test the categories defined by GO for enrichment of over or under expressed genes, separately for the t-test and fold-change based rank lists. iGA calculates hypergeometric p-values iteratively by considering all possible options for the number of differentially expressed genes for each category and determines the optimum p-value. No defining of thresholds for differential expression is thus needed. The significance of the findings was evaluated by means of hypothesis testing using 50,000 permutations of the group labels.

In the second phase, related functions, processes and locations showing deregulation on different levels of specificity on the GO DAG were sought. Branches that contained at least three GO terms with permuted $p < 0.05$ obtained using both t-test and fold change based ranking were detected. Only branches with significant GO terms of less than 100 genes were considered in order to identify small categories from which targets for further analyses can be selected. Multiple hypothesis correction was not performed because the analysis is used to identify categories that, according to the p-values, are most likely to be differentially regulated between the asbestos exposed and non-exposed patients. For verification of the findings, further analysis using independent laboratory techniques is needed.

3 VERIFICATION METHODS

3.1 Microsatellite analysis of 19p (III)

The allelic balance of the p arm of chromosome 19 (chr19:539869-22271313 bp) was assessed using 5-19 microsatellite markers on 62 tumor and respective normal samples. The study specimens included 28 macrodissected tumor samples also used in the gene expression analyses and microdissected tumor samples from 34 additional cases (Table 1). Microdissection was performed using the automated Veritas™ system (Arcturus Bioscience, Inc., Mountain View, CA, USA) on tissue sections stained with fresh filtered 1% toluidine blue (Sigma-Aldrich, Deisenhofen, Germany) and 0.2% methylene blue (Merck, Darmstadt, Germany) solution.

The microsatellites were first amplified using FAM or HEX end labeled primer pairs, which were designed to produce fragments of 80 - 300 bp in length. PCR was performed with an initial 10 min 95ºC denaturation step followed by 35 cycles at 95ºC for 40 s, 40 s at the optimized annealing temperature and 1 min at 72ºC. The fragment separation was performed with a 310 or 3100 Avant Genetic Analyzer (Applied Biosystems). GeneMapper Analysis Software version 3.5 (Applied Biosystems) was used to study the alleles. The determination of allelic imbalance (AI) was performed for heterozygous markers by calculating the ratio of the peak heights of the tumor and normal alleles with ratios of 1.5 or higher being scored as AI. Tumors of patients who showed AI in at least one fourth of the informative microsatellite markers were determined carriers of AI. The markers displaying emergence of novel alleles in the tumor tissue were scored as MSI. The mononucleotide repeat BAT-26 was used to test its correlation with the specimens that harbor MSI.

3.2 FISH (IV & VI)

In Publication IV, a BAC targeting the telomeric region of chromosomal arm 19p (RP11-333F10) was used for quantification of 19p copy numbers. The DNA was labeled directly with fluorescent dUTPs using a Vysis nick translation kit (Vysis, Downers Grove, IL). Tumor ploidy was assessed by a combination of 2-5 corresponding centromeric probes for chromosomes 2, 3, 9, 10, and 15. FISH analyses were performed both on fresh frozen sections and on specimens embedded in paraffin mounted on tissue microarrays obtained from a total of 58 primary tumors. A specimen with a signal to ploidy ratio ≥ 1.3 or with a signal to ploidy difference ≥ 0.9 was considered to carry a gain, whereas a specimen with signal to ploidy ratio of ≤ 0.75 or with a signal to ploidy difference ≤ -0.9 was considered to harbor a loss. The thresholds were chosen to enable detection of one allele copy number changes in cells with polyploid genomes.

In Publication VI, a BAC probe targeting 4q21 (RP11-570L13) was used for detecting copy number changes in 4q. The DNA was labeled by random priming with fluorescent d-UTPs using the BioPrime Labeling System (Invitrogen). Centromeric probes for chromosome 15 and 17 (Vysis) were used to assess tumor ploidy. FISH analyses were performed both on fresh frozen sections from 43 primary tumors and on a tissue microarray containing 36 paraffin embedded brain metastases. Tumors containing a signal to ploidy ratio ≥ 1.5 were considered to carry a gain whereas a ratio of < 0.75 was scored as a loss. The thresholds were chosen to enable detection of one allele copy number changes, but in contrast with Publication IV, the tumors were assumed to exhibit less polyploidy owing to an earlier average stage of the LCs, allowing the use of higher ratios.

3.3 Evaluation of UBA1 and UBA7 levels by Western blotting (V)

Western blotting was performed to study the involvement of UBA1 and UBA7 in exposure related deregulation of protein ubiquitination. Cytoplasmic and nuclear protein fractions were extracted from six asbestos exposed and six non-exposed cases with both groups containing three ACs and three SCCs. Equal amounts of protein were loaded into Tris-HCl gels and blotting was done onto an Immobilon-P PVDF membrane (Millipore, Billerica, MA, USA). UBA1 and UBA7 detections were performed using rabbit polyclonal antibodies to UBA1 (ab16849) and UBA7 (ab12199) (Abcam, Cambridge, UK). Rabbit polyclonal antibody to GAPDH (ab9485, Abcam) was utilized as an internal control. Biotinylated anti-rabbit IgG (Vector Laboratories Inc., Burlingame, CA, USA) was used

together with streptavidin-horseradish peroxidase conjugate (Amersham Biosciences, NJ, USA). ECL reagent (Amersham Biosciences) was used to detect the signals. ImageMaster (Amersham Biosciences, NJ, USA) was used to quantify the intensities of the protein bands.

3.4 Studies on human bronchial epithelial BEAS 2B cells (IV)

Human bronchial epithelial BEAS 2B cells were treated with crocidolite asbestos *in vitro* to assess whether 19p was a preferential site of asbestos induced chromosomal breakage. Prior to the studies, BEAS 2B cells were karyotyped by multicolor FISH (M-FISH) on metaphase spreads to confirm that the cell line had two copies of chromosome 19 and that significant changes in chromosome number and structure were not present. For M-FISH, 24XCyte-MetaSystems 24-colour kit (MetaSystems GmbH, Altlussheim, Germany) was applied and used as recommended by the manufacturer.

For the exposure, 20,000 cells were plated onto Lab-Tek 2-well chamber slides (Nalge Nunc International, Naperville, IL) and maintained in the growth medium. At 48 h, one third of the slides were exposed to 2.0 μg/cm^2 crocidolite fibers, one third to 250 ng/ml mitomycin C (MMC) (positive control), and the remaining third received no exposure (negative control). All cultures were incubated for 48h in the presence of 9 μg/ml cytochalasin-B (Cyt-B), which blocks cytokinesis after nuclear division (Fenech & Morley 1985).

FISH was applied to analyze the contents of MN in BEAS 2B cells treated with crocidolite fibers and in control cultures similarly to the previous description (Lindberg *et al.* 2007). Two probes were used, the BAC probe RP11-333F10 that identifies the telomeric region of chromosomal arm 19p, used also in studies of tumor specimens (Publication IV), and a Human Chromosome Pan-Centromeric paint (Cambio), which labels all human centromeres. To evaluate whether 19p was a preferential site of asbestos induced chromosomal breakage in crocidolite induced MN in binucleate BEAS 2B cells, MN were classified as follows: MN with only a 19p signal were classified as having a 19p fragment. MN with a 19p and a centromere signal were classified as having a whole chromosome 19. MN without signal were classified as having a fragment other than 19p. MN with only a centromere signal were classified as having a whole chromosome other than chromosome 19.

RESULTS

1 THREE-COLOR IMAGE ANALYSIS (I)

An image analysis approach based on the use of an additional dye that is not dependent on the amount of any particular mRNA in the sample was used to improve the assessment of spot quality in gene expression microarrays. Owing to the correspondence of the images, spot segmentation could be performed on the additional, SYBR green II image and the results could be directly applied to the Cy3 and Cy5 images. The SYBR image allowed easy detection of the spot foreground pixels and therefore also facilitated estimation of gene expression in the sample and reference. Additionally, the information provided by the use of SYBR was taken advantage of in the detection of faulty and missing spots.

The performance of the SYBR based identification of faulty spots was compared with that of GenePix Pro 6.0. The comparison showed that although the two methods produced similar estimates for gene expression, the use of an additional dye improved the performance of identification of spots that had not been printed on the slide for one reason or another. While the SYBR based method discarded these spots, GenePix was observed to give high sample to reference ratios to a number of such spots, indicating a possible source of error in the downstream analyses.

2 DIFFERENTIAL REGION FINDING (II, III & VI)

A DRF method was introduced for detection of aberrations that occur preferentially in a given subgroup of tumors with respect to another subgroup. The method analyzes microarray data at group level, i.e. the aberrations are not identified case by case, but instead regions harboring more than an expected number of probes that are able to separate the compared groups are revealed. The detected aberrations may be specific to either of the groups or occur more frequently in either group in comparison to the other.

The DRF was successfully used to define regions with copy number differences between asbestos exposed and non-exposed patients in Publication II. In Publication III, the method was extended for gene expression analyses and was used to define regions with differences in gene expression between the asbestos exposed and non-exposed patients. In Publication VI, the method was used to locate regions of difference in terms of both copy number and gene expression between the BM-positive and BM-negative patients. The verification of the

key results using independent laboratory techniques (in the sections to follow) proved the efficacy of the approach.

3 ASBESTOS RELATED CHANGES

3.1 Molecular changes of asbestos related tumors (II & III)

Array CGH and gene expression microarray analysis was performed to detect aberrations that occur preferentially in lung tumors of either asbestos exposed or non-exposed patients. Exposure related DNA copy number and gene expression changes were detected in18 and 35 regions ($p < 0.01$, two-tailed test) respectively, using the DRF method. To assess whether the asbestos related copy number aberrations could be related to the breakage of fragile sites, an enrichment calculation was performed. We found that 11 of the 125 fragile sites coincided with the asbestos related DNA level aberrations and calculated whether the overlap was higher than would be expected by chance. A one-tailed test suggested an over-representation of the fragile sites within the asbestos related regions ($p = 0.08$).

Table 3. Chromosome regions with DNA copy number and gene expression alterations compared between the tumors of asbestos exposed and non-exposed patients. The positions of the regions are described according to the NCBI Build 34.

Region	Position (Mbp)	Size (Mbp)	Asbestos exposed	Non-exposed
2p21	45.7 – 47.4	1.7	Gain	Loss
3p21.31	48.5 – 49.2	0.6	Loss	No aberration
5q35.2	175.8 – 176.4	0.5	Loss	No aberration
16p13.3	0.3 – 0.8	0.5	No aberration	Gain
19p13.3-p13.11	0.6 – 18.8	18.2	Loss	Gain
22q12.3	34.8 – 34.9	0.1	No aberration	Gain

The results from the separate DNA and gene expression level analyses were compared to detect regions that displayed both expression and DNA level changes related to the exposure status. Altogether, six common regions discriminated the LCs of asbestos exposed and non-exposed patients (Table 3). Owing to the noise level of the data, judging whether an aberration had occurred in an individual sample was not feasible for all the specimens. However, by using the DRF method, relative information about the discriminating aberrations was obtained. On the basis of these results, regions 3p21.31, 5q35.2, 16p13.3, 19p13.3-p13.11, and 22q12.3 were found to show lower DNA copy numbers and contain

genes that were less expressed in the asbestos exposed than in the non-exposed group of tumors. The region 2p21 showed opposite behavior.

3.2 Region 19p in asbestos related carcinogenesis (III & IV)

The 19p region was chosen as the target of downstream studies. The objective of these analyses was to both verify the DRF findings and to characterize the aberration site further. However, as all the six regions appeared to be interesting candidates for further studies on asbestos related cancers, the selection process was arbitrary. The subsequent analyses involved use of high-resolution array CGH for detailed analysis of 19p aberrations, microsatellite analysis for initial verification of the array findings using an independent laboratory technique, and FISH for determining whether the changes detected using microsatellite analysis were caused by losses or gains. Additionally, FISH characterization of asbestos induced micronuclei in human bronchial epithelial BEAS 2B cells was performed to examine whether asbestos is capable of inducing 19p aberrations *in vitro*.

19p aberrations revealed by array CGH (IV)

To identify putative micro level aberrations in the 19p region that could remain unidentified using the cDNA platform that was utilized in the initial screening, high-resolution copy number profiling was performed. The other objective of the study was to localize the aberration break points within chromosome arm 19p. Altogether five asbestos exposed and five non-exposed LC specimens were analyzed using the Agilent 44B microarrays containing 44,000 clones.

The aberration break points were localized to the centromeric region of chromosome 19, but copy number changes smaller than 1 Mbp were not observed. Owing to the presence of only gross level changes, also detectable by the cDNA arrays used in Publication II, the results obtained using the cDNA and oligo platforms were combined in order to get data from additional samples with regard to the 19p aberrations in the LCs. In the combined data set, losses involving 19p were found to occur more commonly in tumors of the asbestos exposed than in those of the non-exposed patients (42%; 5/12 vs. 14%; 2/14). Albeit suggestive, the difference was not statistically significant.

Microsatellite analysis (III)

Microsatellite analysis was used as an initial independent laboratory technique for verification of the connection between asbestos exposure and 19p aberrations. Altogether 62 patients' tumor and normal tissues were analyzed for AI. The analysis involved use of 19 microsatellite markers spanning a 21.7 Mbp region on 19p13.3-p13.11.

The AI degree for individual markers ranged between 50 and 90% in asbestos exposed, 40 and 100% in moderately exposed and 20 and 50% in non-exposed patients' tumor samples (only informative markers are taken into account). In most cases, AI seemed to extend throughout the whole investigated region and no apparent aberration hot spots could be revealed.

As the whole chromosome arm was found to be involved in the aberration, patients were categorized as carriers of AI in 19p if at least 25% of the informative microsatellite markers were AI-positive and otherwise normal. Using this criterion, AI ranging through 19p was present in 80% (20/25) of the heavily exposed, in 75% (6/8) of the moderately exposed, and in 45% (13/29) of the non-exposed patients' tumor samples ($p < 0.01$, Fisher's exact test for the difference in AI between all exposed and non-exposed patients).

Table 4. Prevalence of allelic imbalance in the 19p region in different histological tumor types

Histological tumor type	ASBESTOS EXPOSED patients with AI in 19p	NON-EXPOSED patients with AI in 19p
All	79% (26/33)	45% (13/29)
AC	70% (9/13)	67% (8/12)
SCC	75% (6/8)	36% (4/11)
LCLC	86% (6/7)	0% (0/2)
Other types [I]	100% (5/5)	25% (1/4)

[I] The other types contained SCLC (1 exposed, 2 non-exposed), adenosquamous carcinoma (1,1), giant cell carcinoma (1,0), and pleomoprhic carcinoma (2,1).

Differences were also observed in the prevalence of AI between the histological tumor types included in the study. Among the tumors of asbestos exposed patients, AI of 19p was detected in tumors of all histological types, whereas in the tumors of non-exposed patients, AI seemed to be restricted to certain histological types, notably ACs. Of the non-AC types, 85% (17/20) of the exposed and 29% (5/17) of the non-exposed cases showed AI ($p < 0.01$, Fisher's exact test). The results are summarized in Table 4.

In addition to AI, microsatellite analysis also gives information about MSI occurring in the microsatellites analyzed. 10% (3/29) of the non-exposed specimens displayed MSI throughout the 19p region (>50% of the markers MSI-positive), whereas one non-exposed and three moderately exposed cases showed MSI in one or two markers. MSI was also detected in the BAT-26 marker in two of the three tumors displaying MSI throughout 19p and in two of the four tumors displaying MSI in single markers.

FISH (IV)

FISH analysis was carried out to find which types of aberrations, i.e. gains or losses, the tumors of the asbestos exposed and non-exposed patients carried in the 19p region. This information could not be obtained using microsatellite analysis, as the approach utilized in this thesis provided only semi-quantitative data of the aberrations and was therefore incapable of distinguishing allelic gains from losses.

The FISH analysis was performed on 58 tumor specimens using a BAC probe targeting the telomeric region of 19p. As our previous array CGH and microsatellite analyses had shown that the 19p aberrations typically involve the whole chromosome arm, the 19p aberrations were defined using only one probe. Additionally, two to five centromeric enumeration probes were used to assess the ploidy of each specimen. At least 75 cells were scored for each tumor sample whose average count was used as the estimate of the copy number of the locus. Specimens with copy number losses and gains were detected by scaling the 19p copy numbers against the average of the centromeric counts, which served as an estimate of the ploidy. Results from the moderately exposed samples were combined with the highly exposed samples owing to the limited number of moderately exposed cases included in the study (five).

In the asbestos exposed patients' tumors, losses were detected in 39% (13/33) of the cases, compared with 12% (3/25) of the non-exposed patients' tumors ($p = 0.04$, Fisher's exact test). Gains were detected in 18% (6/33) of the exposed and 40% (10/25) of the non-exposed cases but the difference was not statistically significant. Table 5 shows the results separately for each histological tumor type.

Table 5. Prevalence of copy number losses and gains in the 19p region in different histological tumor types

Histological tumor type	ASBESTOS EXPOSED 19p loss	ASBESTOS EXPOSED 19p gain	NON-EXPOSED 19p loss	NON-EXPOSED 19p gain
All	39% (13/33)	18% (6/33)	12% (3/25)	40% (10/25)
AC	30% (4/13)	23% (3/13)	10% (1/10)	50% (5/10)
SCC	33% (3/9)	11% (1/9)	0% (0/9)	56% (5/9)
LCLC	80% (4/5)	0% (0/5)	0% (0/2)	0% (0/2)
Other types [I]	33% (2/6)	33% (2/6)	50%(2/4)	0%(0/4)

[I] The other types contained SCLC (4 exposed, 3 non-exposed), adenosquamous carcinoma (1,1), and pleomoprhic carcinoma (1,0).

Micronuclei induced by crocidolite asbestos (IV)

After showing that aberrations of 19p, especially losses of the chromosome arm, were more prevalent in the tumors of the asbestos exposed patients than in those of the non-exposed ones, human bronchial epithelial BEAS 2B cells were exposed to crocidolite asbestos to assess whether asbestos is capable of inducing 19p aberrations *in vitro*. Prior to the treatment, M-FISH analysis was performed to ascertain that the BEAS 2B cell line harbored two copies of chromosome 19. In the 12 analyzed metaphases, chromosome 19 was present in two copies, and no rearrangements affecting the chromosome were observed.

The positive control substance MMC increased the frequency of binucleate cells harboring MN with no centromere signals (C^{-}), as expected, and there was a 4.8-fold difference to the untreated control ($p < 0.001$, Fisher's exact test). Also, crocidolite induced a 1.7-fold increase in C^{-} MN ($p < 0.01$). Binucleate cells harboring MN with centromeres (C^{+}) were rare and their frequency remained fairly similar between the different treatments. The results indicated that crocidolite and MMC induce MN that harbor chromosomal fragments.

Among C^{-} MN of binucleate cells, fragments of 19p were detected in 3.4% (6/176) in the crocidolite treatment, 1.4% (2/139) in the MMC treatment, and 0.6% (1/159) in the untreated control cultures (Table 6). The proportion of MN that contained a 19p fragment was 5.4 times higher in crocidolite treated cells than in the untreated control ($p = 0.079$; Fisher's exact test). When the total number of cells scored to obtain the numbers of MN shown in Table 6 were considered, the frequency of cells harboring MN with a 19p fragment was found to be about 10 times higher in the crocidolite treated cultures (about

6000 cells scored) than in the untreated controls (about 10,000 cells scored; $p = 0.01$, Fisher's exact test).

Table 6. Percentage of micronuclei (MN) in BEAS 2B cells found to harbor 19p signals by FISH. MN with (C^+) and without (C^-) centromere signals were analyzed separately

Treatment	**No. MN scored**	**MN with 19p among**	
		C^- MN	**C^+ MN**
Untreated control	177	5.6% (1/18)	0.6% (1/159)
Crocidolite (2.0 µg/cm^2)	206	3.3% (1/30)	3.4% (6/176)
Mitomycin C (250 ng/ml)	169	0% (0/30)	1.4% (2/139)

3.3 Asbestos related deregulated pathways (V)

We performed a pathway analysis to obtain a wider, mechanistic view of the asbestos related alterations at the gene expression level than provided by the DRF analysis, which centered on regions that also displayed aberrations on the DNA level. Biological processes, molecular functions and cellular localizations defined by Gene Ontology were analyzed for enrichment of differentially expressed genes between the asbestos exposed and non-exposed patients. Enrichment analyses were performed on gene expression data from both normal and tumor tissue samples.

Hundreds of differentially regulated categories were identified both in the normal and tumor tissue between the two groups when GO terms from all levels of specificity were considered. The findings included changes in the NF-κB pathway, DNA repair and mitochondrial functions that have already been previously implicated in asbestos related carcinogenesis. However, as selection of targets for subsequent analyses from hundreds of putatively significant findings is not practical, the results were further processed.

The GO terms were ordered as a directed acyclic graph based on the relations of the terms. Branches that contained at least three GO terms with permuted p-values <0.05 that were obtained using both t-test and fold change based ranking, in both the tumor and normal tissues, were then sought. When only the most specific GO terms were considered, 24 up-regulated and 8 down-regulated cellular processes and molecular functions between the asbestos exposed and non-exposed patient samples were revealed. Six out of the 24 up-regulated processes and molecular functions were ion transport related, whereas three out of the eight down-regulated terms were related to protein ubiquitination.

Two proteins, UBA1 and UBA7, involved in the down-regulated ubiquitination processes were chosen for subsequent analyses. Six asbestos exposed and six non-exposed lung tumor and respective normal tissue samples were analyzed for differences in the protein levels by Western analysis. No exposure related differences in UBA1 or UBA7 levels in either the normal or tumor tissue samples could be detected. In the SCC group, expression of both UBA1 and UBA7 were significantly lower in the tumor tissue than in the normal tissue, $p = 0.02$ and $p = 0.01$ (two-sided t-test) respectively.

4 MOLECULAR CHANGES OF TUMORS WITH DTCS IN BM (VI)

Array CGH and gene expression microarray data were analyzed to identify molecular genetic alterations characteristic of lung ACs with or without DTCs in BM. Similarly to the asbestos related studies, the DNA and gene expression level data were analyzed separately using the DRF method. The results of the analyses were compared to discover discriminating regions that might contain genes with a role in metastasizing.

Table 7 shows the altered regions in the BM-positive and BM-negative cases. Three regions, i.e. 4q12-32, 10p12-p11, and 10q21-q22, appeared to be lost in the BM-positive groups of tumors but gained in BM-negative tumors. The gains at 17q21 and 20q11-q13 were more frequent in the BM-positive than in the BM-negative group of tumors. In terms of prevalence of differentially expressed genes (i.e. ROC permuted p-value < 0.05), the 4q and 10p regions were found to be the hot spots. 20% of the genes located at these two regions were found to be differentially expressed between the two groups, whereas on average, 5% of the genes were differentially expressed.

Table 7. Chromosome regions with DNA copy number and gene expression alterations between the tumors of BM-positive and BM-negative patients. The positions of the regions are described according to the NCBI Build 35.

Region	Position (Mbp)	Size (Mbp)	BM-positive	BM-negative
4q12-32	53.3 – 160.4	107.1	Loss	Gain
10p12-p11	21.1 – 33.5	12.4	Loss	Gain
10q21-q22	69.9 – 81.5	11.7	Loss	Gain
17q21	35.4 – 46.4	11.0	Gain	No aberration
20q11-q13	30.6 – 56.4	25.8	Gain	No aberration

The 4q region was chosen as the target of further verifications owing to being a hot spot of differentially expressed genes and because the region coincided with previously reported metastatic aberration signatures (Petersen *et al.* 2000; Goeze *et al.* 2002). Both primary LCs and brain metastases of LCs were characterized for 4q21 copy number changes using FISH. 4q was frequently lost in the BM-positive group of tumors consisting of both ACs and SCCs, while gains were more common than losses in the BM-negative groups of tumors (Table 8). The loss of 4q was significantly associated with the presence of DTCs in the BM (11/22; 50% vs. 2/21; 10%, $p < 0.001$, Fisher's exact test). 4q was also lost in 39% (13/36) of the brain metastases of LCs comprised of ACs, SCCs, and SCLCs. The results are summarized in Table 8.

Table 8. Prevalence of copy number losses and gains in the 4q region in different histological tumor types.

Histological tumor type	BM-positive n=22 4q loss	BM-positive n=22 4q gain	BM-negative n=21 4q loss	BM-negative n=21 4q gain	Brain metastases n=36 4q loss	Brain metastases n=36 4q gain
All	50% (11/22)	5% (1/22)	10% (2/21)	14% (3/21)	39% (13/36)	3% (2/36)
AC	38% (3/8)	0% (0/8)	7% (1/14)	21% (3/14)	36% (5/14)	0% (0/14)
SCC	57% (8/14)	7% (1/14)	17% (1/6)	0% (0/6)	50% (5/10)	10% (1/10)
LCLC	-	-	0% (0/1)	0% (0/1)	-	-
SCLC	-	-	-	-	43% (3/7)	0% (0/7)
Not known[1]	-	-	-	-	0% (0/5)	20% (1/5)

[1] The information of tumor histology was missing for five brain metastases

DISCUSSION

1 MICROARRAY ANALYSIS

Microarrays provide an attractive means of conducting genome wide studies of multiple samples in parallel. Information can be obtained from virtually all genes of the study subject making the technology especially suitable for studies where the affected genes are not known *a priori*. False positive findings can be controlled at different levels, which include detection of faulty spots by means of image analysis as well as statistical significance tests and multiple testing corrections and validations of the findings using independent methods and study specimens. Especially when microarrays of lower quality are used, analysis of spatial features of the spots should be performed to detect and discard faulty spots that may otherwise show as noise in the subsequent analyses. As Publication I showed, spot analysis of gene expression microarrays may be performed with the aid of an additional dye, which is not dependent on the amount of mRNA in the sample. On the other hand, image analysis of two-color DNA copy number arrays is likely to succeed without the use of a third dye, as samples with normal genomes are typically used as the reference.

In this thesis, a differential region finding method was developed for comparative studies of DNA and mRNA level high-throughput data. The DRF method was designed for comparative study settings and functions by comparing data from two conditions at group level, while other region finding methods typically analyze each sample separately (Pollack *et al*. 2002; Myers *et al*. 2004; Furge *et al*. 2005). The method does not require defining of thresholds to call aberrations because the putative regions of difference are detected by means of hypothesis testing. As the output of the method is regions, analysis of array CGH and gene expression data from different platforms is also enabled, provided that the array probes have been mapped against the same genome build. Furthermore, the region level analysis was considered more suitable for our purposes than gene by gene analysis such as that presented by Berger *et al.* (2006). To enable gene by gene analyses, the corresponding probes between different platforms would have had to be determined in the studies of this thesis. Owing to a lack of one to one mapping between probes from different array platforms, the region level approach was considered more feasible. We are aware of the fact that the DRF method may fail to detect micro level changes that discriminate between the groups under comparison when segments of 0.5-1.0 Mbp in size are used as in this study. However, use of smaller segments was observed to result in an increase in positive findings

that, based on visual inspection, appeared to be false positive results brought about by noise in the data.

This thesis applied the comparative study setting to identify DNA and mRNA level changes that could be involved in 1) development of asbestos related LCs and 2) dissemination of tumor cells into BM. The aim was to detect regions where changes in DNA copy number were associated with gene expression alterations to identify genes with a putative role in lung carcinogenesis.

In the following sections, the findings made in this thesis with regard to asbestos related and dissemination related changes are further discussed.

2 ALTERATIONS INDUCED BY ASBESTOS EXPOSURE

Comparative microarray analyses of lung tumors from asbestos exposed and non-exposed patients were carried out to investigate whether distinct changes are characteristic of the asbestos related tumors. Presence of such changes could indicate that some of the driver genes may be different in the carcinogenic process if influenced by asbestos, which knowledge is required for better understanding of the disease.

2.1 Aberrations characteristic of asbestos related cancers (II - IV)

Array CGH and gene expression microarrays were performed on groups of primary asbestos related and unrelated tumors matched for patient and tumor characteristics. To detect chromosome regions with both gene copy number and gene expression changes, the two data types were analyzed using the DRF method and the results were combined. Six regions were revealed that contained asbestos related copy number and gene expression changes. To verify the results obtained using the DRF and to further characterize the findings, one of the discriminating regions was chosen for subsequent analyses. The characterizations specified that especially the loss of 19p is related to asbestos exposure.

Despite only a limited number of samples having been analyzed using microarrays, meaning that the identified putative asbestos related regions could represent random aberrations, several factors in addition to the successful verification studies of the 19p region support the importance of the findings in asbestos related carcinogenesis. Firstly, in accordance with the chromosomal deletions and DNA breakage induced by asbestos fibers (Adachi *et al*. 1994; Marczynski *et al.* 1994; Levresse *et al*. 2000), only one of the six

regions showed gains in the tumors of asbestos exposed patients. Secondly, the regions were identified as changed at both the mRNA and DNA levels, which reduces the possibility of the findings having been brought about by noise in the data. Thirdly, a change at the level of gene expression is required for a cellular response (Pinkel & Albertson 2005), indicating that the regions harbor candidate genes for further analysis.

Owing to the genomic instability characteristic of LCs, copy number changes can be detected throughout the entire tumor genomes. Not surprisingly, the putative asbestos related regions 2p, 3p, 5q, 16p, 19p, and 22q have also been described in LCs without a history of asbestos exposure (Luk *et al.* 2001; Petersen *et al.* 1997; Balsara & Testa 2002; Garnis *et al*. 2006; Coe *et al.* 2006; Weir *et al.* 2007). These studies indicate that the 3p region is among the most typically aberrated regions in LCs and may be aberrated in up to 90% of the tumors. The five other putative asbestos related regions displayed copy number changes in about 10-50% of LCs, but the aberration frequencies of distinct regions depend on the tumor types included in the analysis (Luk *et al.* 2001; Petersen *et al.* 1997; Balsara & Testa 2002; Garnis *et al*. 2006; Coe *et al.* 2006; Weir *et al.* 2007). Accordingly, the asbestos related aberrations revealed in this study are suggested to occur at higher frequencies among the asbestos exposed than the non-exposed group of tumors, rather than being exclusively induced by asbestos exposure.

As described in Section 2.3 of the Review of Literature, differential genetic changes have previously been identified between LCs from smokers and non-smokers. The findings suggest that different exposures can give rise to their mutation signatures, which are apparent in the tumor genome. Indeed, *in vitro* studies have shown that asbestos fibers cause breaks in chromosomes 1 and 9 (Dopp *et al.* 1997; Dopp & Schiffmann 1998; Lohani *et al.* 2002). Furthermore, monosomy of chromosome 19 has been suggested to have relevance in the tumorigenic transformation of human bronchial epithelial BEP2D cells (Suzuki *et al.* 2001) induced by asbestos. Additionally, one of the only studies where a chromosome region has been characterized for differences between the asbestos exposed and non-exposed patients, revealed a connection between extensive LOH of 3p21 and asbestos exposure (Marsit *et al.* 2004). The previous findings support our results as we also detected asbestos related mRNA and DNA level changes at regions 3p21 and 19p. Furthermore, DNA level changes were detected in chromosomes 1 and 9. Nevertheless, our study was the first genome wide investigation of asbestos related LCs and was thus expected to reveal novel candidates for asbestos related target regions.

The subsequent microsatellite analysis revealed that AI of 19p was related to asbestos exposure in histological tumor types other than ACs. Interestingly, AI of 19p has previously been shown to be associated with smoking in lung ACs (Sanchez-Cespedes *et al.* 2001). As the tumor samples included in our analyses originated from patients with a definite smoking history, our results agree with those of Sanchez-Cespedes *et al.* (2001). It therefore appears that smoking drives aberrations of 19p in ACs, but exposure to asbestos does so in non-AC histological types. Further evidence for 19p as a preferential site of asbestos induced chromosomal breakage was obtained from our *in vitro* characterizations of crocidolite induced MN. An increase in the frequency of MN harboring the 19p fragment was detected already after a 48h treatment, in comparison with the untreated control, which could imply that loss of 19p occurs early in asbestos related carcinogenesis.

Although no genes located in the 19p region have been associated with asbestos exposure, the gene targeted by 19p deletions in LCs in general is thought to be the TSG *LKB1* (Sanchez-Cespedes *et al.* 2002). Previously, mutations of *LKB1* have been reported to occur preferentially in lung tumors of smoking males (Matsumoto *et al.* 2007). This suggests that loss of 19p and inactivation of *LKB1* could belong to an alternative route of carcinogenesis similar to what has been observed between *KRAS* and *EGFR* in lung cancers of smokers and non-smokers (Ahrendt *et al.* 2001; Pao *et al.* 2004; Shigematsu *et al.* 2005; reviewed in Sun *et al.* 2007). Although we did not perform studies on asbestos related target genes, one of the microsatellite markers that separated the tumors of asbestos exposed and non-exposed patients the best resided proximal to *LKB1*. Therefore, *LKB1* may be a target gene in asbestos related carcinogenesis also, since the 19p region was found to be preferentially lost in tumors of asbestos exposed patients. If the loss of 19p and the inactivation of *LKB1* belong to a specific route of carcinogenesis, tumors influenced by asbestos could be driven through this route. However, it is possible that the target of asbestos related 19p aberrations is another yet unknown TSG because dozens of genes located at this region were found be down-regulated in tumors of asbestos exposed patients.

Microsatellite analysis also revealed that altogether 11% (7/62) of the tumors showed MSI in the 19p region, which frequency agrees with previous findings (Sekido *et al.* 1998). Tumors of both non-exposed and moderately asbestos exposed patients were found to be affected, suggesting that asbestos exposure is unlikely to be related to the occurrence of MSI. 30% of the tumors displaying MSI were found to harbor triploid tumors by FISH, indicating simultaneous alterations in both CIN and MSI pathways. Although cell line investigations have indicated that either instability appears to be sufficient to drive

tumorigenesis (Lengauer *et al.* 1997), both of the instabilities may be developed as the disease progresses.

In all, the studies of this thesis suggest that non-random targeting of distinct loci such as 19p occurs in asbestos related lung cancers. Although the thesis is not capable of providing an explanation for why asbestos exposure drives specific aberrations, the regions detected in this study could be related to breakage of fragile sites owing to their proximity to fragile sites. Indeed, although fragile sites have been implicated as being loci of frequent rearrangements also in cancers (Yunis & Soreng 1984; reviewed in Durkin and Glover 2007), specific conditions are known to induce their breakage (Glover *et al.* 1984). The most fragile site, FRA3B located in 3p14 has, for instance, been previously suggested as a preferential target of tobacco smoke (Sozzi *et al.* 1997), suggesting that other substances could also induce similar effects. Further studies should be carried out to further characterize whether a connection between breakage of the fragile sites and asbestos exists.

2.2 Asbestos related deregulated pathways (V)

The Gene Ontology enrichment analysis provides an attractive means for detecting changes that have occurred in a biologically related group of genes between two conditions of interest. Moderate changes in expression may be sufficient to show deregulation of a category as opposed to single gene approaches that require significant changes. Furthermore, analyses of groups of genes instead of single genes are justified as alterations in the expression of genes do not occur as independent events. Also, in contrast to the copy number driven expression changes sought for in Publication III, the aim of this substudy was to identify gene expression changes induced by any mechanism.

GO terms were analyzed for enrichment of differentially expressed genes between the asbestos exposed and the non-exposed lung cancer patients. To enable the selection of targets for subsequent analyses, the search was narrowed down to those molecular processes and cellular functions that were deregulated on at least three levels of specificity in the GO graph. Owing to the child terms being specific descriptions of the parent terms, related parent and child terms are expected to show deregulation when truly altered. We focused into the categories that were deregulated in both normal and tumor tissue, because such findings may imply the early involvement of these mediators in asbestos related carcinogenesis.

On the basis of the enrichment scores of related GO terms, down-regulation of protein ubiquitination and up-regulation of ion channels were considered to be the main findings. Protein ubiquitination is best known in tagging proteins for degradation (Hershko & Ciechanover 1998), but ubiquitin modifications have recently been recognized as more versatile regulators of protein activity in processes related e.g. to cell cycle, transcription, and DNA repair (reviewed in Haglund & Dikic 2005). Deregulated ubiquitination has been linked to a variety of diseases including cancers (reviewed in Weissman 2001) including the asbestos exposure linked mesothelioma (Borczuk *et al.* 2006; Wali *et al.* 2007). Also our recent studies of time dependent gene expression changes of lung derived cell lines after exposure to asbestos fibers showed down-regulation of ubiquitination, 24h or 48h after the exposure to asbestos fibers (Nymark *et al.* 2007). Concordant findings from several asbestos related studies suggest the importance of protein ubiquitination in asbestos mediated carcinogenesis albeit the link between protein ubiquitination and asbestos is yet to be characterized.

UBA1 and UBA7 were tested for involvement in the asbestos related deregulation of protein ubiquitination. These enzymes were chosen owing to their roles at the early stages of ubiquitin and ubiquitin like tagging processes. Additionally, both of these proteins have been annotated to the ubiquitin cycle, one of the most specific protein ubiquitination related GO terms that were found down-regulated in this study. Although Western analysis did not reveal differences in UBA1 or UBA7 expression between the asbestos exposed and non-exposed patients, the role of protein ubiquitination in asbestos related carcinogenesis cannot be ruled out, as hundreds of other enzymes are also involved in the process. Future analyses of a similar kind could thus benefit from more holistic approaches.

As the GO analysis was performed on the same gene expression as the asbestos related regions discussed in Section 2.1 were detected from, the findings are dependent. To unravel the connection, we examined whether the genes involved in the most specific protein ubiquitination related GO terms were enriched in the six genomic loci found to be related to asbestos exposure, but no enrichment was detected. The analysis was, however, crude and involved all the hundreds of genes annotated to the most detailed ubiquitination related GO terms. The analysis could be refined, for instance, by assessing only those genes involved in the ubiquitination process that share a common transcriptional regulator. Promoter sequence analysis of the differentially expressed genes annotated to the down-regulated ubiquitination related GO terms could be worth pursuing as it may pinpoint a subgroup of genes that are affected.

Despite the fact that the asbestos related deregulation of ubiquitination could not be verified, the GO analysis indicated that the asbestos related and unrelated tumors display differences at the level of pathways as well. The finding therefore supports the presence of genetic differences between tumors of patients exposed and not exposed to asbestos, as implicated in Publications II-IV.

3 ALTERATIONS UNDERLYING DISSEMINATION INTO BM (VI)

After the successful detection of asbestos related aberrations, a similar comparative approach combining array CGH and gene expression microarrays was used to identify genetic alterations that may be related to the early dissemination of tumor cells into BM in LCs. Lung ACs of BM-positive and BM-negative patients were compared, facilitating the detection of differential alterations displayed by the primary tumors that could provide a tumor cell with the capability required in the metastatic cascade. The soundness of the division is supported by clinical evidence, which suggests that BM-positive tumors are more likely to relapse after surgery than BM-negative ones (Cote *et al.* 1995; Pantel *et al.* 1996; Kubuschok *et al.* 1999).

The combined gene expression and DNA copy number microarray screening performed in this thesis revealed five regions, where changes at both the mRNA and DNA levels associated with the BM status. Although comparative studies have been performed previously between LCs with and without distant metastases (Petersen *et al.* 2000; Goeze *et al.* 2002), our study was the first one where changes related to early dissemination were studied. However, early dissemination and emergence of distant metastases are likely to be connected phenomena owing to an association between disease relapse and the presence of DTCs in BM at the time or surgery (Cote *et al.* 1995; Pantel *et al.* 1996; Kubuschok *et al.* 1999). Therefore, division of the metastatic and non-metastatic tumors on the basis of the BM status is likely to provide a finer separation than the presence of macro-metastases, but either comparison should reveal predominant metastasizing related changes, provided that they are displayed by the primary tumors. Indeed, in accordance with our results, comparative studies of macro-metastatic and non-metastatic LCs have also identified losses of 4q and 10q among the changes differentiating the tumor groups (Petersen *et al.* 2000; Goeze *et al.* 2002). This gives confidence in our findings.

Further investigations were carried out on the 4q region owing to an overlap with the previous metastatic aberration signatures (Petersen *et al.* 2000; Goeze *et al.* 2002) and

because the expression of a notable fraction of the genes at the 4q locus (20%) was found to be influenced by the copy number loss. FISH characterizations validated the array finding of the connection between the loss of 4q and the presence of DTCs in BM in lung ACs, but additionally showed that the loss occurred in BM-positive SCCs. Furthermore, losses of 4q were detected in brain metastases of LCs as commonly as in the primary tumors, i.e. in about 40% of the cases, and likewise across different histological tumor types. The aberration thus appears to be commonly carried by spreading LCs and could possibly even be a more universal characteristic of advanced tumors, as loss of 4q has been shown to be associated with progression in mesothelioma and esophageal adenocarcinoma also (Shivapurkar *et al*. 1999; Sterian *et al*. 2006).

The recurrence of 4q loss in the metastatic tumor aberration signature suggests that a gene or genes in the 4q region could confer a capability required in the metastatic cascade. However, to pass through the metastatic cascade, multiple capabilities are needed (reviewed in Fidler 2003). Thus, for instance the 4q aberration alone is not likely to provide all the required abilities, but may, together with other alterations, produce metastasis competent cells. This could explain why 7% of the BM-negative tumors displayed loss of 4q, i.e. they lack some of the competencies needed in the dissemination process. Also, BM-positive and BM-negative breast tumors have been shown to display distinct gene expression alterations (Woelfle *et al*. 2003; Naume *et al*. 2007), indicating that the tendency for dissemination is likely to be governed by the genetic alterations of the primary tumors.

In all, the results indicate that loss of 4q, and thus the genetic alterations displayed by the primary tumor, have relevance in the dissemination of LCs. Yet, further knowledge about the genes targeted by the aberrations is required for improved understanding of the biology behind metastasizing.

SUMMARY AND CONCLUSIONS

This thesis aimed to develop and apply microarray data analysis methods suitable for comparative studies of two opposite categories of tumors, and to further characterize the main findings using independent laboratory techniques. Two distinct categories of LCs were studied, those related to asbestos exposure and those with evidence of DTCs in BM.

The DRF method introduced in this thesis was proven efficient in generating hypotheses about both DNA and mRNA level differences between specimens of two opposite groups, and is likely to suit future comparative study settings. Yet, further verifications of the findings are necessary to exclude false positive findings and to confirm the relevance of the results in the condition being studied.

On the basis of microarray analyses, DNA copy number and gene expression changes in chromosome regions 2p21, 3p21.31, 5q35.2, 16p13.3, 19p13.3-p13.11, and 22q12.3 are suggested to have relevance in LCs of patients with a history of asbestos exposure. The non-random targeting of distinct chromosomal loci indicates that some of the driver genes may be different in the lung carcinogenesis influenced by asbestos. In subsequent verification studies, asbestos exposed patients' tumors were found to display AI in 19p caused by losses, significantly more frequently than the non-exposed patients' tumors. The aberration could occur early in carcinogenesis as BEAS 2B cells treated with crocidolite asbestos preferentiality showed this chromosomal breakage. In addition to the preferentially aberrated regions, tumors of asbestos exposed and non-exposed patients were found to display differential regulation of multiple biological processes and molecular functions, including protein ubiquitination and ion channels. Altogether, the studies of this thesis suggest differences in asbestos related and unrelated carcinogenic processes. In the future, it will be interesting to see whether alternative routes are involved in asbestos related carcinogenesis, similarly to what has been observed between smokers and non-smokers (Ahrendt *et al*. 2001; Pao *et al.* 2004; Shigematsu *et al.* 2005; reviewed in Sun *et al.* 2007).

Comparative studies of patients with and without evidence of DTCs in BM showed that regions 4q12-q32, 10p12-p11, 10q21-q22, 17q21, and 20q11-q13 may be of relevance in the metastatic cascade. In the further characterization of 4q, loss of the region was found a typical feature of BM-positive primary LCs and brain metastases across different histological types. The results suggest that loss of 4q has relevance in the dissemination of LCs and that changes displayed by the primary tumors may assist identification of aggressive tumors.

The results obtained in this thesis demonstrate that distinct molecular changes are characteristic of LCs influenced by asbestos exposure and of those that disseminate to BM. However, although knowledge of chromosomal regions associated with a condition may facilitate prognosis or assist in the recognition of distinct types of tumors, for example, the regions as such do not provide accurate information for treating the disease. Further action should thus be taken to propel research to reveal the genes targeted by the aberrations, as they may uncover the biology behind the development of lung tumors.

ACKNOWLEDGEMENTS

The work presented in this thesis was carried out at Finnish Institute of Occupational Health (FIOH) in Helsinki, Finland and at Department of Information and Computer Science at Helsinki University of Technology (HUT) in Espoo, Finland. I want to thank the heads of these organizations and their personnel for providing excellent research facilities. Especially team leaders Kirsti Husgafvel-Pursiainen at FIOH and Heikki Mannila at HUT are thanked for making the work possible. I also gratefully acknowledge the financial support received from Graduate School in Computational Biology, Bioinformatics, and Biometrics (ComBi), Finnish Cultural Foundation and Academy of Finland.

I want to thank my supervisors, Sisko Anttila and Jaakko Hollmén, for their support during the course of the thesis and for giving me the freedom to work my way. Jaakko is also thanked for organizing financial support despite of my transition to Meilahti.

I am also very grateful to Harriet Wikman for her valuable ideas on various aspects related to the studies of this thesis. It has been a great pleasure to work with Haiju and for being able to share thoughts of all sorts with someone on the same wavelength.

The official reviewers, Outi Monni and Sampsa Hautaniemi, are thanked for critical review of the thesis and for their valuable comments.

I thank all the coauthors of the original publications included in this thesis for pleasant collaboration. Huge additional thanks to Hannu Norppa for help with the micronuclei studies and to Sakari Knuutila for providing facilities for the microarray experiments. Also Michaela Krämling and Penny Nymark deserve additional thanks for good collaboration. Technical assistance received from Päivi Tuominen, Tuula Stjernvall, Helinä Urhonen, Hilkka Järventaus, Satu Suhonen, and Tiina Wirtanen is gladly acknowledged.

My fellow research mates are acknowledged for fruitful discussions and for contributing to my thinking of research in an incredible way. Present and former colleagues from Meilahti region and Otaniemi are thanked for support of all kind. I especially want to thank Anu Usvasalo for inspiring discussions and Sippy Kaur for exchange of thoughts.

My dear friends outside the scientific world are thanked for fun times spent together. More or less extreme doings are delightedly memorized and coming ones likewise awaited J . I also want to thank my family for their endless support. Above all, I owe my deepest gratitude for Terho for putting up with me during the past busy months.

REFERENCES

Adachi S, Yoshida S, Kawamura K, Takahashi M, Uchida H, Odagiri Y, and Takemoto K. (1994) Inductions of oxidative DNA damage and mesothelioma by crocidolite, with special reference to the presence of iron inside and outside of asbestos fiber. *Carcinogenesis*, **15**, 753-8.

Affymetrix. (2004) GeneChip® Expression Analysis Technical Manual. Technical Report 701021 Rev. 5.

Ahrendt SA, Decker PA, Alawi EA, Zhu Yr YR, Sanchez-Cespedes M, Yang SC, Haasler GB, Kajdacsy-Balla A, Demeure MJ, and Sidransky D. (2001) Cigarette smoking is strongly associated with mutation of the K-*ras* gene in patients with primary adenocarcinoma of the lung. *Cancer*, **92**, 1525-30.

Albert R, Jeong H, and Barabasi AL. (2000) Error and attack tolerance of complex networks. *Nature*, **406**, 378-82.

Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ, and Clarke MF. (2003) Prospective identification of tumorigenic breast cancer cells. *Proc Natl Acad Sci USA*, **100**, 3983-8.

Anderson MW, Reynolds SH, Ming Y, and Maronpot RM. (1992) Role of proto-oncogene activation in carcinogenesis. *Environ Health Perspect*, **98,** 13-24

Arlt MF, Durkin SG, Ragland RL, and Glover TW. (2006) Common fragile sites as targets for chromosome rearrangements. *DNA Repair*, **5**, 1126-35.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, and Sherlock G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25-29.

Balsara BR and Testa JR. (2002) Chromosomal imbalances in human lung cancer. *Oncogene*, **21**, 6877-83.

Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, and Edgar R. (2005) NCBI GEO: mining millions of expression profiles-database and tools. *Nucleic Acids Res*, **33**¸ D562-6.

Bartkova J, Horejsí Z, Koed K, Krämer A, Tort F, Zieger K, Guldberg P, Sehested M, Nesland JM, Lukas C, Ørntoft T, Lukas J, and Bartek J. (2005) DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. *Nature*, **434**, 864-70.

Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, and Meyerson M. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA*, **98**, 13790-5.

Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JMG, Iannettoni MD, Orringer MB, and Hanash S. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med*, **8**, 816-24.

Benjamini Y and Hochberg Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Statist Soc B*, **57**, 289-300.

Bennett WP, Hussain SP, Vahakangas KH, Khan MA, Shields PG, and Harris CC. (1999) Molecular epidemiology of human cancer risk: gene-environment interactions and p53 mutation spectrum in human lung cancer. *J Pathol*, **187**, 8-18.

Berg JM, Tymoczko JL, and Stryer L. (2002) *Biochemistry*, 5th Ed. WH Freeman and Company, New York, NY, USA.

Berger JA, Hautaniemi S, Mitra SK, and Astola J. (2006) Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. *IEEE/ACM Trans Comput Biol Bioinform*, **3**, 2-16.

Bielas JH and Loeb LA. (2005) Mutator phenotype in cancer: timing and perspectives. *Environ Mol Mutagen*, **45**, 206-13.

Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN, and Srivastava S. (1998) A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res*, **58**, 5248-57.

Borczuk AC, Cappellini GC, Kim HK, Hesdorffer M, Taub RN, and Powell CA. (2007) Molecular profiling of malignant peritoneal mesothelioma identifies the ubiquitin-proteasome pathway as a therapeutic target in poor prognosis tumors. *Oncogene*, **26,** 610-7.

Boyle P and Ferlay J. (2005) Cancer incidence and mortality in Europe, 2004. *Ann Oncol*, **16**, 481-8.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, and Vingron M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, **29**, 365-71.

Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, and Sansone SA. (2003) ArrayExpress-a public repository for microarray gene expression data at the EBI. *Nucleic Acid Res*, **31**, 68-71.

Breitling R, Amtmann A, and Herzyk P. (2004) Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics*, **5,** 34.

Cahan P, Ahmad AM, Burke H, Fu S, Lai Y, Florea L, Dharker N, Kobrinski T, Kale P, and McCaffrey TL. (2005) List of lists-annotated (LOLA): A database for annotation comparison of published microarray gene lists. *Gene*, **360**, 78–82.

Cahill DP, Lengauer C, Yu J, Riggins GJ, Willson JK, Markowitz SD, Kinzler KW, and Vogelstein B. (1998) Mutations of mitotic checkpoint genes in human cancers. *Nature*, **392**, 300-3.

Carvalho B, Ouwerkerk E, Meijer GA, and Ylstra B. (2004) High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *J Clin Pathol*, **57**, 644-6.

Casper AM, Nghiem P, Arlt MF, and Glover TW. (2002) ATR regulates fragile site stability. *Cell*, **111**, 779-89.

Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, Cheng CL, Wang CH, Terng HJ, Kao SF, Chan WK, Li HN, Liu CC, Singh S, Chen WJ, Chen JJ, Yang PC. (2007) A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med*, **356**, 11-20.

Cistulli CA, Sorger T, Marsella JM, Vaslet CA, and Kane AB. (1996) Spontaneous p53 mutation in murine mesothelial cells: increased sensitivity to DNA damage induced by asbestos and ionizing radiation. *Toxicol Appl Pharmacol*, **141**, 264-71.

Cleveland W. (1979) Robust locally weighted regression and smoothing scatterplots. *J Am Stat Association*, **74,** 829-836.

Coe BP, Lockwood WW, Girard L, Chari R, MacAulay C, Lam S, Gazdar AF, Minna JD, and Lam WL. (2006) Differential disruption of cell cycle pathways in small cell and non-small cell lung cancer. *Br J Cancer*, **94**, 1927-35.

Conlin A, Smith G, Carey FA, Wolf CR, and Steele RJ. (2005) The prognostic significance of K-ras, p53, and APC mutations in colorectal carcinoma. *Gut*, **54**, 1283-6.

Cote RJ, Beattie EJ, Chaiwun B, Shi SR, Harvey J, Chen SC, Sherrod AE, Groshen S and Taylor CR. (1995) Detection of occult bone marrow micrometastases in patients with operable lung carcinoma. *Ann Surg*, **222**, 415-423.

Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, and Meng F. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*, **33**, e175.

Dietmaier W, Wallinger S, Bocker T, Kullman F, Fishel R, and Rüschoff J. (1997) Diagnostic microsatellite instability: definition and correlation with mismatch repair protein expression. *Cancer Res*, **57**, 4749-56.

Dopp E, Schuler M, Schiffmann D, and Eastmond DA. (1997) Induction of micronuclei, hyperdiploidy and chromosomal breakage affecting the centric/pericentric regions of chromosomes 1 and 9 in human amniotic fluid cells after treatment with asbestos and ceramic fibers. *Mutat Res*, **377**, 77-87.

Dopp E and Schiffmann D. (1998) Analysis of chromosomal alterations induced by asbestos and ceramic fibers. *Toxicol Lett*, **96-97**, 155-62.

Downward J. (2006) Cancer biology: signatures guide drug choice. *Nature*, **439**, 274-5.

Downward J. (2003) Targeting RAS signalling pathways in cancer therapy. *Nat Rev Cancer*, **3**, 11-22.

Dracheva T, Philip R, Xiao W, Gee AG, 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**, 157-64.

Dupuy A and Simon RM. (2007) Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst*, **99**, 147-57.

Durkin SG and Glover TW. (2007) Chromosome fragile sites. *Annu Rev Genet*, **41**, 169-92.

Eberhard DA, Johnson BE, Amler LC, Goddard AD, Heldens SL, Herbst RS, Ince WL, Jänne PA, Januario T, Johnson DH, Klein P, Miller VA, Ostland MA, Ramies DA, Sebisanovic D, Stinson JA, Zhang YR, Seshagiri S, and Hillan KJ. (2005) Mutations in the epidermal growth factor receptor and in KRAS are predictive and prognostic indicators in patients with non-small-cell lung cancer treated with chemotherapy alone and in combination with erlotinib. *J Clin Oncol*, **23**, 5900-9.

Fearon ER and Vogelstein B. (1990) A genetic model for colorectal tumorigenesis. *Cell*, **61**, 759-67.

Fenech M and Morley AA. (1985) Measurement of micronuclei in lymphocytes. *Mutat Res*, **147**, 29-36.

Fenech M. (2000) The in vitro micronucleus technique. *Mutat Res*, **455**, 81-95.

Fidler IJ and Kripke ML. (1997) Metastasis results from preexisting variant cells within a malignant tumor. *Science*, **197**, 893-5.

Fidler IJ. (2003) The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. *Nat Rev Cancer*, **3**, 453-8.

Friedberg EC. (2003) DNA damage and repair. *Nature*, **421**, 436-40.

Fukasawa K. (2007) Oncogenes and tumour suppressors take on centrosomes. *Nat Rev Cancer*, **7**, 911-24.

Furge KA, Dykema KJ, HO C, and Chen X. (2005) Comparison of array-based comparative genomic hybridization with gene expression-based regional expression biases to identify genetic abnormalities in hepatocellular carcinoma. *BMC Genomics*, **6**, 67.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, and Stratton MR. (2004) A census of human cancer genes. *Nat Rev Cancer*, **4**, 177-83.

Gacy AM, Goellner G, Juranic N, Macura S, and McMurray CT. (1995) Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell*, **81**, 533-40.

Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI, Altman RB, Brown PO, Botstein D, and Petersen I. (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc Natl Acad Sci USA*, **98**, 13784-9.

Garnis C, Lockwood WW, Vucic E, Ge Y, Girard L, Minna JD, Gazdar AF, Lam S, MauAulay C, and Lam WL. (2006) High resolution analysis of non-small cell lung cancer cell lines by whole genome tiling path array CGH. *Intl J Cancer*, **118**, 1556-64.

Gao HG, Chen JK, Stewart J, Song B, Rayappa C, Whong WZ, and Ong T. (1997) Distribution of p53 and K-ras mutations in human lung cancer tissues. *Carcinogenesis*, **18**, 473-8.

Gazdar AF, Shigematsu H, Herz J, and Minna JD. (2004) Mutations and addiction to EGFR: the Achilles 'heal' of the lung cancers? *Trends Mol Med*, **10**, 481-6.

Gealy R, Zhang L, Siegfried JM, Luketich JD, and Keohavong P. (1999) Comparison of mutations in the p53 and K-ras genes in lung carcinomas from smoking and nonsmoking women. *Cancer Epidemiol Biomarkers Prev*, **8**, 297-302.

Gene Ontology Consortium. (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res*, **34,** D322-326.

Glover TW, Berger C, Coyle J, and Echo B. (1984) DNA polymerase a inhibition by aphidicolin induces gaps and breaks at common fragile sites in human chromosomes. *Hum Genet*, **67**, 136-42

Goeze A, Schluns K, Wolf G, Thäsler Z, Petersen S, and Petersen I. (2002) Chromosomal imbalances of primary and metastatic lung adenocarcinomas. *J Pathol*, **196**, 8-16.

González-Quevedo R, Iniesta P, Morán A, de Juan C, Sánchez-Pernaute A, Fernández C, Torres A, Díaz-Rubio E, Balibrea JL, and Benito M. (2002) Cooperative role of telomerase activity and p16 expression in the prognosis of non-small-cell lung cancer. *J Clin Oncol*, **20**, 254-62.

Good P. (2000) Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, 2nd ed. Springer-Verlag, New York, USA.

Gorgoulis VG, Vassiliou LV, Karakaidos P, Zacharatos P, Kotsinas A, Liloglou T, Venere M, Ditullio RA Jr, Kastrinakis NG, Levy B, Kletsas D, Yoneta A, Herlyn M, Kittas C, and Halazonetis TD. (2005) Activation of the DNA damage checkpoint and genomic instability in human precancerous lesions. *Nature*, **434**, 907-13.

Greenblatt MS, Bennett WP, Hollstein M, and Harris CC. (1994) Mutations in the *p53* tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. *Cancer Res,* **54**, 4877-78.

Hackshaw AK, Law MR, and Wald NJ. (1997) The accumulated evidence on lung cancer and environmental tobacco smoke. *BMJ*, **315**¸ 980-8.

Haglund K and Dikic I. (2005) Ubiquitylation and cell signaling. *EMBO J*, **24**, 3353-9.

Hanahan D and Weinberg RA. (2000) The hallmarks of cancer. *Cell*, **100**, 57-70.

Hanash S. (2004) Integrated global profiling of cancer. *Nat Rev Cancer*, **4**, 638-44.

Hansen LT, Thykjaer T, Ørntoft TF, Rasmussen LJ, Keller P, Spang-Thomsen M, Edmonston TB, Schmutte C, Fishel R, and Petersen LN. (2003) The role of micmatch repair in small-cell lung cancer. *Eur J Cancer*, **39**, 1456-67.

Harbour JW, Lai SL, Whang-Peng J, Gazdar AF, Minna JD, and Kaye FJ. (1988) Abnormalities in structure and expression of the human retinoblastoma gene in SCLC. *Science*, **241**, 353-7.

Hatada I, Kato A, Morita S, Obata Y, Nagaoka K, Sakurada A, Sato M, Horii A, Tsujimoto A, and Matsubara K. (2002) A microarray-based method for detecting methylated loci. *J Hum Genet*, **47**, 448-51.

Hautaniemi S, Edgren H, Vesanen P, Wolf M, Järvinen A-K, Yli-Harja O, Astola J, Kallioniemi O, and Monni O. (2003) A novel strategy for microarray quality control using Bayesian networks. *Bioinformatics*, **19**, 2031-8.

Hautaniemi S, Ringrén M, Kauraniemi P, Autio R, Edgren R, Yli-Harja O, Astola J, Kallioniemi A, and Kallioniemi O-P. (2004) A strategy for identifying putative causes of gene expression variation in human cancers. *J Franklin Inst*, **341**, 77-88.

Hayashi N, Sugimoto Y, Tsuchiya E, Ogawa M, and Nakamura Y. (1994) Somatic mutations of the MTS (multiple tumor suppressor) 1/CDK4l (cyclin-dependent kinase-4 inhibitor) gene in human primary non-small cell lung carcinomas. *Biochem Biophys Res Commun*, **202**, 1426-30.

Hayes DN, Monti S, Parmigiani G, Gilks CB, Naoki K, Bhattacharjee A, Socinski MA, Perou C, and Meyerson M. (2006) Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. *J Clin Oncol*, **24**, 5079-90.

Hellman A, Zlotorynski E, Scherer SW, Cheung J, Vincent JB, Smith DI, Trakhtenbrot L, and Kerem B. (2002) A role for common fragile site induction in amplification of human oncogenes. *Cancer Cell*, **1**, 89-97.

Hershko A and Ciechanover A. (1998) The ubiquitin system. *Annu Rev Biochem*, **67**, 425-79.

Hewett DR, Handt O, Hobson L, Mangelsdorf M, Eyre HJ, Baker E, Sutherland GR, Schuffenhauer S, Mao JI, and Richards RI. (1998) FRA10B structure reveals common elements in repeat expansion and chromosomal fragile site genesis. *Mol Cell*, **1**, 773-81.

Hibi K, Takahashi T, Yamakawa K, Ueda R, Sekido Y, Ariyoshi Y, Suyama M, Takagi H, Nakamura Y, and Takahashi T. (1992) Three distinct regions involved in 3p deletion in human lung cancer. *Oncogene*, **7**, 445-9.

Hoffman D and Hoffman I. (1997) The changing cigarette, 1950-1995. *J Toxicol Environ Health*, **50**, 307-64.

Holm S. (1979) A simple sequentially rejective multiple test procedure. *Scand J Statist*, **6**, 65-70.

Hollstein M, Sidransky D, Vogelstein B, and Harris CC. (1991) p53 mutations in human cancers. *Science*, **253**, 49-53.

Hsu HS, Wen CK, Tang YA, Lin RK, Li WY, Hsu WH, and Wang YC. (2005) Promoter hypermethylation is the predominant mechanism in *hMLH1* and *hMSH2* deregulation and is a poor prognostic factor in nonsmoking lung cancer. *Clin Cancer Res*, **11**, 5410-6.

Husgafvel-Pursiainen K, Karjalainen A, Kannio A, Anttila S, Partanen T, Ojajärvi A, and Vainio H. (1999) Lung cancer and past occupational exposure to asbestos. Role of *p53* and *K-ras* mutations. *Am J Respir Cell Mol Biol*, **20**, 667-74.

Husgafvel-Pursiainen K, Boffetta P, Kannio A, Nyberg F, Pershagen G, Mukeria A, Constantinescu V, Fortes C, Benhamou S. (2000) *p53* mutations and exposure to environmental tobacco smoke in a multicenter study on lung cancer. *Cancer Res*, **60**, 2906-11.

Hyman E, Kauraniemi P, Hautaniemi S, Wolf M, Mousses S, Rozenblum E, Rignér M, Sauter G, Monni O, Elkhloun A, Kallioniemi OP, and Kallioniemi A. (2002) Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res*, **62**, 6240–5.

Ichimura K, Bolin MB, Goike HM, Schmidt EE, Moshref A, and Collins VP. (2000) Deregulation of the p14ARF/MDM2/p53 pathway is a prerequisite for human astrocytic gliomas with G1-S transition control gene abnormalities. *Cancer Res*, **60**, 417-24.

Ionov Y, Yamamoto H, Krajewski S, Reed JC, and Perucho M. (2000) Mutational inactivation of the proapoptotic gene *BAX* confers selective advantage during tumor clonal evolution. *Proc Natl Acad Sci USA*, **97**, 10872-7.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, and Speed TP. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucl Acids Res*, **31**, e15.

Janssen YM, Van Houten B, Borm PJ, and Mossman BT. (1993) Cell and tissue responses to oxidative damage. *Lab Invest*, **69**, 261-74.

Janssen YM, Barchowsky A, Treadwell M, Driscoll KE, and Mossman BT. (1995) Asbestos induces nuclear factor κB (NF-κB) DNA-binding activity and NF-κB-dependent gene expression in tracheal epithelial cells. *Proc Natl Acad Sci USA*, **29**, 8458-62.

Jaurand MC. (1997) Mechanisms of fiber-induced genotoxicity. *Environ Health Perspect*, **105,** 1073-84.

Kamp DW, Israbian VA, Preusen SE, Zhang CX, and Weitzman SA. (1995) Asbestos causes DNA strand breaks in cultured pulmonary epithelial cells: role of iron-catalyzed free radicals. *Am J Physiol*, **268**, 471-80.

Karjalainen A, Anttila S, Vanhala E, and Vainio H. (1994) Asbestos exposure and the risk of lung cancer in a general urban population. *Scand J Work Environ Health*, **20**, 243-50.

Kaur S, Larramendy ML, Gentile M, Svarvar C, Koivisto-Korander R, Vauhkonen H, Scheinin I, Leminen A, Bützow R, Böhling T, and Knuutila S. (2006) New insights into the cellular pathways affected in primary uterine leiomyosarcoma**.** *Cancer Genomics Proteomics* **3** 347-354.

Kim NW, Piatyszek MA, Prowse KR, Harley CB, West MD, Ho PLC, Coviello GM, Wright WE, Weinrich SL, and Shay JW. (1994) Specific association of human telomerase activity with immortal cells and cancer. *Science*, **266**, 2011-4

Kinzler KW and Vogelstein B. (1997) Cancer-susceptibility genes. Gatekeepers and caretakers. *Nature*, **386** , 761-3.

Kinzler KW and Vogelstein B. (1998) Landscaping the cancer terrain. *Science*, **280**, 1036-7.

Knudson AG Jr. (1971) Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci USA*, **68**, 820-3.

Knudson, A.G. (2002) Cancer genetics. *Am J Med Genet*, **111**, 96-102.

Kok K, Naylon SL, Buys CH. (1997) Deletions of the short arm of chromosome 3 in solid tumors and the search for suppressor genes. *Adv Cancer Res*, **71**, 27-92.

Kosaka T, Yatabe Y, Endoh H, Kuwano H, Takahashi T, and Mitsudomi T. (2004) Mutations of the epidermal growth factor receptor gene in lung cancer: biological and clinical implications. *Cancer Res*, 64, 8919-23.

Kremer EJ, Pritchard M, Lynch M, Yu S, Holman K, Baker E, Warren ST, Schlessinger D, Sutherland GR, and Richards RI. (1991) Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. *Science*, **252**, 1711-4.

Kubuschok B, Passlick B, Izbicki JR, Thetter O, and Pantel K. (1999) Disseminated tumor cells in lymph nodes as a determinant for survival in surgically resected non-small-cell lung cancer. *J Clin Oncol*, 17, 19-24.

Lapidot T, Sirard C, Vormoor J, Murdoch B, Hoang T, Caceres-Cortes J, Minden M, Paterson B, Caligiuri MA, and Dick JE. (1994) A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature*, **367**, 645-8.

Lau SK, Boutros PC, Pintilie M, Blackhall FH, Zhu CQ, Strumpf D, Johnston MR, Darling G, Keshavjee S, Waddell TK, Liu N, Lau D, Penn LZ, Shepherd FA, Jurisica I, Der SD, and Tsao MS. (2007) Three-gene prognostic classifier for early-stage non small-cell lung cancer. *J Clin Oncol*, **25**, 5562-9.

Le Calvez F, Mukeria A, Hunt JD, Kelm O, Hung RJ, Tanière P, Brennan P, Boffetta P, Zaridze DG, and Hainaut P. (2005) TP53 and KRAS mutation load and types in lung cancers in relation to tobacco smoke: distinct patterns in never, former, and current smokers. *Cancer Res*, **65**, 5076-83.

Lee PN. (2001) Relation between exposure to asbestos and smoking jointly and the risk of lung cancer. *Occup Environ Med*, **58**, 145-53.

Lengauer C, Kinzler KW, and Vogelstein B. (1997) Genetic instability in colorectal cancers. *Nature,* **386**, 623-7.

Lengauer C, Kinzler KW, and Vogelstein B. (1998) Genetic instabilities in human cancer. *Nature*, **396**, 643-9.

Leslie A, Pratt NR, Gillespie K, Sales M, Kernohan NM, Smith G, Wolf CR, Carey FA, and Steele RJ. (2003) Mutations of APC, K-ras, and p53 are associated with specific chromosomal aberrations in colorectal adenocarcinomas. *Cancer Res*, **63**, 4656-61.

Levinson G and Gutman GA. (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol*, **4**, 203-21.

Levresse V, Renier A, Levy F, Broaddus VC, and Jaurand M. (2000) DNA breakage in asbestos-treated normal and transformed (TSV40) rat pleural mesothelial cells. *Mutagenesis*, **15**, 239-44.

Li Q, Fraley C, Bumgarner RE, Yeung KY, and Raftery AE. (2005) Donuts, scratches and blanks: robust model-based segmentation of microarray images. *Bioinformatics*, **21**¸2875-82.

Lindberg HK, Wang X, Järventaus H, Falck GC-M, Norppa H and Fenech M. (2007) Origin of nuclear buds and micronuclei in normal and folate-deprived human lymphocytes. *Mutat Res*, **617** 33-45.

Little CD, Nau MM, Carney DN, Gazdar AF, and Minna JD. (1983) Amplification and expression of the c-myc oncogene in human lung cancer cell lines. *Nature*, **306**, 194-6.

Liu Y, Gao W, Siegfried JM, Weissfeld JL, Luketich JD, and Keohavong P. (2007) Promoter methylation of RASSF1A and DAPK and mutations of K-ras, p53, and EGFR in lung tumors from smokers and never-smokers. *BMC Cancer*, **7**, 74.

Loeb LA, Springgate CF, and Battula N. (1974) Errors in DNA replication as a basis of malignant changes. *Cancer Res*, **34**, 2311-21.

Loeb LA. (1991) Mutator phenotype may be required for multistage carcinogenesis. *Cancer Res*, **51**, 3075-9.

Loeb KR and Loeb LA. (1999) Genetic instability and the mutator phenotype. Studies in ulcerative colitis. *Am J Pathol*, **154**, 1621-6.

Lohani M, Dopp E, Becker EE, Seth K, Schiffmann D, and Rahman Q. (2002) Smoking enhances asbestos-induced genotoxicity, relative involvement of chromosome 1: a study using multicolor FISH with tandem labeling. *Toxicol Lett,* **136**, 55-63.

Lubin JH and Blot WJ. (1984) Assessment of lung cancer risk factors by histologic category. *J Natl Cancer Inst*, **73**, 383-9.

Luk C, Tsao M-S, Bayani J, Shepherd F, and Squire JA. (2001) Molecular cytogenetic analysis of non-small cell lung carcinoma by spectral karyotyping and comparative genomic hybridization. *Cancer Genet Cytogenet*, **125**, 87-99.

MAQC Consortium. (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol*, **24**, 1151-61.

Malkhosyan S, Rampino N, Yamamoto H, and Perucho M. (1996) Frameshift mutator mutations. *Nature*, **382**, 499-500.

Mao L, Lee JS, Kurie JM, Fan YH, Lippman SM, Lee JJ, Ro JY, Broxson A, Yu R, Morice RC, Kemp BL, Khuri FR, Walsh GL, Hittelman WN, Hong WK. (1997) Clonal genetic alterations in the lungs of current and former smokers. *J Natl Cancer Inst*, **89**, 857-62.

Marczynski B, Czuppon AB, Marek W, Reichel G, and Baur X. (1994) Increased incidence of DNA double-strand breaks and anti-ds DNA antibodies in blood of workers occupationally exposed to asbestos. *Hum Exp Toxicol*, **13,** 3-9.

Marnett LJ. (2000) Oxyradicals and DNA damage. *Carcinogenesis*, **23**, 361-70.

Marshall E. (2004) Getting the noise out of gene arrays. *Science*, 3**06**, 630-1.

Marsit CJ, Hasegawa M, Hirao T, Kim DH, Aldape K, Hinds PW, Wiencke JK, Nelson HH, and Kelsey KT. (2004) Loss of heterozygosity of chromosome 3p21 is associated with mutant TP53 and better patient survival in non-small-cell lung cancer. *Cancer Res*, **64**, 8702-7.

Matsumoto S, Iwakawa R, Takahashi K, Kohno T, Nakanishi Y, Matsuno Y, Suzuki K, Nakamoto M, Shimizu E, Minna JD, and Yokota J. (2007) Prevalence and specificity of LKB1 alterations in lung cancers. *Oncogene*, **26**, 5911-8.

Merlino GT, Xu YH, Richert N, Clark AJ, Ishii S, Banks-Schlegel S, and Pastan I. (1985) Elevated epidermal growth factor receptor gene copy number and expression in a squamous carcinoma cell line. *J Clin Invest*, **75**, 1077-9.

Michiels S, Koscielny S, and Hill C. (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**, 488–92.

Mitelman F, Johansson B, and Mertens F. (eds.) (2007) Mitelman Database of Chromosome Aberrations in Cancer, http://cgap.nci.nih.gov/Chromosomes/Mitelman.

Mitsudomi T, Oyama T, Nishida K, Ogami A, Osaki T, Sugio K, Yasumoto K, Sugimachi K, and Gazdar AF. (1996) Loss of heterozygosity at 3p in non-small cell lung cancer and its prognostic implication. *Clin Cancer Res*, **2**, 1185-9.

Mossman BT, Kamp DW, and Weitzman SA. (1996) Mechanisms of carcinogenesis and clinical features of asbestos-associated cancers. *Cancer Invest*, **14**, 466-80.

Mossman BT, Faux S, Janssen Y, Jimenez LA, Timblin C, Zanella C, Goldberg J, Walsh E, Barchowsky A, and Driscoll K. (1997) Cell signaling pathways elicited by asbestos. *Environ Health Perspect*, **105**, 1121-5.

Mossman BT, Lounsbury KM, and Reddy SP. (2006) Oxidants and signaling by mitogen-activated protein kinases in lung epithelium. *Am. J. Respir. Cell. Mol. Biol.,* **34**, 666-669.

Myers CL, Dunham MJ, Kung SY, and Troyanskaya OG. (2004) Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics*, **20**, 3533-43.

Myllykangas S, Himberg J, Böhling T, Nagy B, Hollmén J, and Knuutila S. (2006) DNA copy number amplification profiling of human neoplasms. *Oncogene*, **25**, 7324-32.

Naruke T. (2003) Surgery in locally advanced non-small cell lung cancer. *Lung Cancer*, **42**, S11-5.

Naume B, Zhao X, Synnestvedt M, Borgen E, Russnes HG, Lingjarde OC, Stromberg M, Wiedswang G, Kvalheim G, Kåresen R, Nesland JM, Borresen-Dale A-L, and Sorlie T. (2007) Presence of bone marrow micrometastasis is associated with different recurrence risk within molecular subtypes of breast cancer. *Molecular Oncology*, **1**, 160-71.

Nelson HH, Wiencke JK, Gunn L, Wain JC, Christiani DC, and Kelsey KT. (1998) Chromosome 3p14 alterations in lung cancer: evidence that FHIT exon deletion is a target of tobacco carcinogens and asbestos. *Cancer Res*, **58**, 1804-7.

Nelson HH, Christiani DC, Wiencke JK, Mark EJ, Wain JC, and Kelsey KT. (1999) k-*ras* mutation and occupational asbestos exposure in lung adenocarcinoma: asbestos-related cancer without asbestosis. *Cancer Res*, **59**, 4570-3.

Nielsen JL, Walsh JT, Degen DR, Drabek SM, McGill JR, and von Hoff DD. (1993) Evidence of gene amplification in the form of double minute chromosomes is frequently observed in lung cancer. *Cancer Genet Cytogenet*, **65**, 120-4.

Ninomiya H, Nomura K, Satoh Y, Okumura S, Nakagawa K, Fujiwara M, Tsuchiya E, and Ishikawa Y. (2006) Genetic instability in lung cancer: concurrent analysis of chromosomal, mini- and microsatellite instability and loss of heterozygosity. *Br J Cancer*, **94**, 1485-91.

Nowell PC. (1976) The clonal evolution of tumor cell populations. *Science*, **194**, 23-8.

Nymark P, Lindholm P, Korpela MV, Lahti L, Ruosaari S, Kaski S, Hollmén J, Anttila S, Kinnula V, and Knuutila S. (2007) Gene expression profiles in asbestos-exposed epithelial and mesothelial lung cell lines. *BMC Genomics*, **8**, 62.

O'Brien CA, Pollett A, Gallinger S, and Dick JE. (2007) A human colon cancer cell capable of initiating tumour growth in immunodeficient mice. *Nature*, **445**, 106-10.

Ohta M, Inoue H, Cotticelli MG, Kastury K, Baffa R, Palazzo J, Siprashvili Z, Mori M, McCue P, Druck T, Croce CM, and Huebner K. (1996) The FHIT gene, spanning the chromosome 3p14.2 fragile site and renal carcinoma-associated t(3;8) breakpoint, is abnormal in digestive tract cancers. *Cell*, **84**, 587-97.

Ollikainen T, Linnainmaa K, and Kinnula VL. (1999) DNA single strand breaks induced by asbestos fibers in human pleural mesothelial cells in vitro. *Environ Mol Mutagen*, **33**, 153-60.

Pantel K, Schlimok G, Angstwurm M, Weckermann D, Schmaus W, Gath H, Passlick B, Izbicki JR, and Riethmüller G. (1994) Methodological analysis of immunocytochemical screening for disseminated epithelial tumor cells in bone marrow. *J Hematother*, **3**,165-73.

Pantel K, Izbicki J, Passlick B, Angstwurm M, Häussinger K, Thetter O, and Riethmüller G. (1996) Frequency and prognostic significance of isolated tumour cells in bone marrow of patients with non-small-cell lung cancer without overt metastases. *Lancet*, **347**, 649-53.

Pao W, Miller W, Zakowski M, Doherty J, Politi K, Sarkaria I, Singh B, Heelan R, Rusch V, Fulton L, Mardis E, Kupfer D, Wilson R, Kris M, and Varmus H. (2004) EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proc Natl Acad Sci USA*, **101**, 13306-11.

Parsons R, Li GM, Longley MJ, Fang WH, Papadopoulos N, Jen J, de la Chapelle A, Kinzler KW, Vogelstein B, and Modrich P. (1993) Hypermutability and mismatch repair deficiency in RER+ tumor cells. *Cell*, **75**, 1227-36.

Pawson T and Saxton TM. (1999) Signaling networks - do all roads lead to the same genes? *Cell*, **97**, 675-8.

Pei J, Balsara BR, Li W, Litwin S, Gabrielson E, Feder M, Jen J, and Testa JR. (2001) Genomic imbalances in human lung adenocarcinomas and squamous cell carcinomas. *Genes Chromosomes Cancer*, **31**, 282-7.

Pekarsky Y, Zanesi N, Palamarchuk A, Huebner K, and Croce CM. (2002) FHIT: from gene discovery to cancer treatment and prevention. *Lancet Oncol*, **3**, 748-54.

Petersen I, Bujard M, Petersen S, Wolf G, Goeze A, Schwendel A, Langreck H, Gellert K, Reichel M, Just K, du Manor S, Cremer T, Dietel M, and Ried T. (1997) Patterns of chromosomal imbalances in adenocarcinoma and squamous cell carcinoma of the lung. *Cancer Res,* **57**, 2331-5.

Petersen S, Aninat-Meyer M, Schlüns K, Gellert K, Dietel M, and Petersen I. (2000) Chromosomal alterations in the clonal evolution to the metastatic stage of squamous cell carcinomas of the lung. *Br J Cancer*, **82**, 65-73.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, and Albertson DG. (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*, **20**, 207-11.

Pinkel D and Albertson DG. (2005) Array comparative hybridization and its applications in cancer. *Nat Genet*, **37**, S11-17.

Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, and Brown PO. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet*, **23**, 41-6.

Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, and Brown PO. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci USA*, **99**, 12963–68.

Pulciani S, Santos E, Long LK, Sorrentino V, and Barbacid M. (1985) *ras* gene amplification and malignant transformation. *Mol Cell Biol*, **5**, 2836-41.

Pylkkänen L, Wolff H, Stjernvall T, Tuominen P, Sioris R, Karjalainen A, Anttila S and Husgafvel-Pursiainen K. (2002) Reduced Fhit protein expression and loss of heterozygosity at FHIT gene in tumors from smoking and asbestos-exposed lung cancer patients. *Int J Oncol*, **20**, 285-90.

Quackenbush J. (2002) Microarray data normalization and transformation. *Nat Genet*, **32** Suppl, 496-501.

Rodriguez E, Sreekantaiah C, and Chaganti RS. (1994) Genetic changes in epithelial solid neoplasia. *Cancer Res*, **54**, 3398-406.

Roth. (1995) Molecular events in lung cancer. *Lung Cancer*, **12**, 3-15.

Rowley JD. (1973) A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature*, **243**, 290-3.

Ruosaari S and Hollmén J. (2002) Image analysis for detecting faulty spots from microarray images. In Steffen Lange, Ken Satoh, and Carl H. Smith, editors, *Proceedings of the 5th International Conference on Discovery Science (DS 2002)*, volume 2534 of *Lecture Notes in Computer Science* 2534, 259 - 266. Springer-Verlag, Berlin / Heidelberg, Germany.

Saccomanno G, Archer VE, Saunders RP, Auerbach O, and Klein MG. (1976) Early indices of cancer risk among uranium miners with reference to modifying factors. *Ann N Y Acad Sci,* **271**, 377-83.

Sanchez-Cespedes M, Parrella P, Esteller M, Nomoto S, Trink B, Engles JM, Westra WH, Herman JG, and Sidransky D. (2002) Inactivation of LKB1/STK11 is a common event in adenocarcinomas of the lung. *Cancer Res*, **62**, 3659-62.

Sandberg AA, Turc-Carel C, and Gemmill RM. (1988) Chromosomes in solid tumors and beyond. *Cancer Res*, **48**, 1049-59.

Schena M, Shalon D, Davis RW, and Brown PO. (1995) Quantitative monitoring of gene expression patterns with complimentary DNA microarray. *Science*, **270**, 467-70.

Schwartz AG, Prysak GM, Bock CH, and Cote ML. (2007) The molecular epidemiology of lung cancer. *Carcinogenesis*, **28**, 507-18.

Sekido Y, Fong KM, and Minna JD. (1998) Progress in understanding the molecular pathogenesis of human lung cancer. *Biochim Biophys Acta*, **1378**, F21-59.

Selikoff IJ, Hammond EC, and Churg J. (1968) Asbestos exposure, smoking, and neoplasia. *JAMA*, **204,** 106-12.

Sengupta S and Harris CC. (2005) p53: traffic cop at the crossroads of DNA repair and recombination. *Nature Rev Mol Cell Biol*, **6**, 44-55.

Shaffer JP. (1995) Multiple hypothesis testing. *Annu Rev Psychol*, **46**, 561-84.

Shapiro GI, Edwards CD, Kobzik L, Godleski J, Richards W, Sugarbaker DJ, Rollins BJ. (1995) Reciprocal Rb inactivation and p16INK4 expression in primary lung cancers and cell lines. *Cancer Res*, **55**, 505-9.

Shigematsu H, Lin L, Takahashi T, Nomura M, Suzuki M, Wistuba II, Fong KM, Lee H, Toyooka S, Shimizu N, Fujisawa T, Feng Z, Roth JA, Herz J, Minna JD, and Gazdar AF. (2005) Clinical and biological features associated with epidermal growth factor receptor gene mutations in lung cancers. *J Natl Cancer Inst*, **97**, 339-46.

Shivapurkar N, Virmani AK, Wistuba, II, Milchgrub S, Mackay B, Minna JD, and Gazdar AF. (1999) Deletions of chromosome 4 at multiple sites are frequent in malignant mesothelioma and small cell lung carcinoma. *Clin Cancer Res*, **5**, 17-23.

Shukla A, Gulumian M, Hei TK, Kamp D, Rahman Q, and Mossman BT. (2003) Multiple roles of oxidants in the pathogenesis of asbestos-induced diseases. *Free Radic Biol Med*, **34**, 1117–29

Sieber OM, Heinimann K, Gorman P, Lamlum H, Crabtree M, Simpson CA, Davies D, Neale K, Hodgson SV, Roylance RR, Phillips RK, Bodmer WF, Tomlinson IP. (2002) Analysis of chromosomal instability in human colorectal adenomas with two mutational hits at APC. *Proc Natl Acad Sci USA*, **99**, 16910-5.

Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JK, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, and Velculescu VE. (2006) The consensus coding sequences of human breast and colorectal cancers. *Science*, **314**, 268-74.

Smith G, Carey FA, Beattie J, Wilkie MJ, Lightfoot TJ, Coxhead J, Garner RC, Steele RJ, and Wolf CR. (2002) Mutations in APC, Kirsten-ras, and p53--alternative genetic pathways to colorectal cancer. *Proc Natl Acad Sci USA*, **99**, 9433-8.

Smith DI, Zhu Y, McAvoy S, and Kuhn R. (2006) Common fragile sites, extremely large genes, neural development and cancer. *Cancer Lett*, **232**, 48-57.

Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Döhner H, Cremer T, and Lichter P. (1997) Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer*, **20**, 399-407.

Sozzi G, Veronese ML, Negrini M, Baffa R, Cotticelli MG, Inoue H, Tornielli S, Pilotti S, De Gregorio L, Pastorino U, Pierotti MA, Ohta M, Huebner K, and Croce CM. (1996) The FHIT gene 3p14.2 is abnormal in lung cancer. *Cell*, **85**, 17-26.

Sozzi G, Sard L, De Gregorio L, Marchetti A, Musso K, Buttitta F, Tornielli S, Pellegrini S, Veronese ML, Manenti G, Incarbone M, Chella A, Angeletti CA, Pastorino U, Huebner K, Bevilaqua G, Pilotti S, Croce CM, and Pierotti MA. (1997) Association between cigarette smoking and FHIT gene alterations in lung cancer. *Cancer Res*, **57**, 2121-3.

Sozzi G, Pastorino U, Moiraghi L, Tagliabue E, Pezzella F, Ghirelli C, Tornielli S, Sard L, Huebner K, Pierotti MA, Croce CM, and Pilotti S. (1998) Loss of FHIT function in lung cancer and preinvasive bronchial lesions. *Cancer Res*, **58**, 5032-7.

Stark M and Hayward N. (2007) Genome-wide loss of heterozygosity and copy number analysis in melanoma using high-density single-nucleotide polymorphism arrays. *Cancer Res*, **67**, 2632-42.

Stehelin D, Varmus HE, Bishop JM, and Vogt PK. (1976) DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*, **260**, 170-3.

Sterian A, Kan T, Berki AT, Mori Y, Olaru A, Schulmann K, Sato F, Wang S, Paun B, Cai K, Hamilton JP, Abraham JM, and Meltzer SJ. (2006) Mutational and LOH analyses of the chromosome 4q region in esophageal adenocarcinoma. *Oncology*, **70**, 168-72.

Stoler DL, Chen N, Basik M, Kahlenberg MS, Rodriguez-Bigas MA, Petrelli NJ, and Anderson GR. (1999) The onset and extent of genomic instability in sporadic colorectal tumor progression. *Proc Natl Acad Sci USA*, **96**, 15121-6.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, and Mesirov JP. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, **102**, 15545-50.

Sun S, Schiller JH, and Gazdar AF. (2007) Lung cancer in never smokers -- a different disease. *Nat Rev Cancer*, **7**, 778-90.

Sutherland GR, Baker E, and Richards RI. (1998) Fragile sites still breaking. *Trends Genet*, **14**, 501-6.

Suzuki M, Piao CQ, Zhao YL, and Hei TK. (2001) Karyotype analysis of tumorigenic human bronchial epithelial cells transformed by chrysotile asbestos using chemically induced premature chromosome condensation technique. *Int J Mol Med*, **8**, 43-7.

Suzuki M, Shigematsu H, Iizasa T, Hiroshima K, Nakatani Y, Minna JD, Gazdar AF, Fujisawa T. (2006) Exclusive mutation in epidermal growth factor receptor gene, HER-2, and KRAS, and synchronous methylation of nonsmall cell lung cancer. *Cancer*, **106**, 2200-7.

Swets J. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285-93.

Takahashi T, Nau MM, Chiba I, Birrer MJ, Rosenberg RK, Vinocour M, Levitt M, Pass H, Gazdar AF, and Minna JD. (1989) p53: a frequent target for genetic abnormalities in lung cancer. *Science*, **246***:* 491-4.

Takeuchi T, Tomida S, Yatabe Y, Kosaka T, Osada H, Yanagisawa K, Mitsudomi T, and Takahashi T. (2006) Expression profile-defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors. *J Clin Oncol*, **24**, 1679-88.

Tammemagi MC, McLaughlin JR, and Bull SB. (1999) Meta-analyses of p53 tumor suppressor gene alterations and clinicopathological features in resected lung cancers. *Cancer Epidemiol Biomarkers Prev*, **8**, 625-34.

Tan PK, Downey TJ, Spitznagel EL Jr, Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, and Cam MC. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res*, **31**, 5676-84.

Tautz D and Renz M. (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Res*, **12**, 4127-38

Testa JR, Liu Z, Feder M, Bell DW, Balsara B, Cheng JQ, and Taguchi T. (1997) Advances in the analysis of chromosome alterations in human lung carcinomas. *Cancer Genet Cytogenet*, **97**, 20-32.

Thibodeau SN, Bren G, and Schaid D. (1993) Microsatellite instability in cancer of the proximal colon. *Science*, **260**, 816-9.

Thun MJ, Henley SJ, and Calle EE. (2002) Tobacco use and cancer: an epidemiologic perspective for geneticists. *Oncogene*, **21**, 7307-25.

Tomlinson IP, Novelli MR, and Bodmer WF. (1996) The mutation rate and cancer. *Proc Natl Acad Sci USA*, **93**, 14800-3.

Tonon G, Wong KK, Maulik G, Brennan C, Feng B, Zhang Y, Khatry DB, Protopopov A, You MJ, Aguirre AJ, Martin ES, Yang Z, Ji H, Chin L, Depinho RA. (2005) High-resolution genomic profiles of human lung cancer. *Proc Natl Acad Sci USA*, **102**, 9625-30.

Toyooka S, Tokumo M, Shigematsu H, Matsuo K, Asano H, Tomii K, Ichihara S, Suzuki M, Aoe M, Date H, Gazdar AF, Shimizu N. (2006) Mutational and epigenetic evidence for independent pathways for lung adenocarcinomas arising in smokers and never smokers. *Cancer Res*, **66**, 1371-5.

Travis WD, Colby TV, Corrin B, Shimosato Y, and Brambilla E. (1999) Histological typing of lung and pleural tumours. *World health organization international histological classification of tumours*, 3rd ed. Springer-Verlag, Berlin, Germany.

Upadhyay D and Kamp DW. (2003) Asbestos-induced pulmonary toxicity: role of DNA damage and apoptosis. *Exp Biol Med*, **228**, 650-9.

Vainio H and Boffetta P. (1994) Mechanisms of the combined effect of asbestos and smoking in the etiology of lung cancer. *Scand J Work Environ Health*, **20**, 235-42.

van den Ijssel P, Tijssen M, Chin SF, Eijk P, Carvalho B, Hopmans E, Holstege H, Bangarusamy DK, Jonkers J, Meijer GA, Caldas C, and Ylstra B. (2005) Human and mouse oligonucleotide-based array CGH. *Nucleic Acids Res*, **33**, e192.

Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, Nakamura Y, White R, Smits AM, and Bos JL. (1988) Genetic alterations during colorectal-tumor development. *N Engl J Med,* **319**, 525-32

Vogelstein B, Lane D, and Levine AJ. (2000) Surfing the p53 network. *Nature*, **408**, 307-10.

Wali A, Wang Y, Kanakapalli D, Liu Z, Sharma S, Lonardo F, Rishi AK, Majumdar AP, Ping Dou Q, Carbone M, Ruckdeschel JC, and Pass HI. (2007) Role of Ubiquitin-proteasome proteolytic pathway in malignant pleural mesothelioma. (Abstract). In: *American Association for Cancer Research Annual Meeting Proceedings*, Abstract nr 4406.

Wang X, Christiani DC, Wiencke JK, Fischbein M, Xu X, Cheng TJ, Mark E, Wain JC, and Kelsey KT. (1995) Mutations in the p53 gene in lung cancer are associated with cigarette smoking and asbestos exposure. *Cancer Epidemiol Biomarkers Prev*, **4**, 543-8.

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, and Lander ES. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**, 1077-82.

Wang X; Ghosh S, and Guo SW. (2001) Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res*, **29**, E75-5.

Wang YC, Lu YP, Tseng RC, Lin RK, Chang JW, Chen JT, Shih CM, and Chen CY. (2003) Inactivation of *hMLH1* and *hMSH2* by promoter methylation in primary non-small cell lung tumors and matched sputum samples. *J Clin Invest*, **111**, 887-95.

Weinberg RA. (2007) *The Biology of Cancer.* Garland Science, Taylor & Francis Group, LLC, New York, NY, USA.

Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhim R, Lin WM, Province MA, Kraja A, Johnson LA, Shah K, Sato M, Thomas RK, Barletta JA, Borecki IB, Broderick S, Chang AC, Chiang DY,

Chirieac LR, Cho J, Fujii Y, Gazdar AF, Giordano T, Greulich H, Hanna M, Johnson BE, Kris MG, Lash A, Lin L, Lindeman N, Mardis ER, McPherson JD, Minna JD, Morgan MB, Nadel M, Orringer MB, Osborne JR, Ozenberger B, Ramos AH, Robinson J, Roth JA, Rusch V, Sasaki H, Shepherd F, Sougnez C, Spitz MR, Tsao MS, Twomey D, Verhaak RG, Weinstock GM, Wheeler DA, Winckler W, Yoshizawa A, Yu S, Zakowski MF, Zhang Q, Beer DG, Wistuba II, Watson MA, Garraway LA, Ladanyi M, Travis WD, Pao W, Rubin MA, Gabriel SB, Gibbs RA, Varmus HE, Wilson RK, Lander ES, and Meyerson M. (2007) Characterizing the cancer genome in lung adenocarcinoma. *Nature*, **450**, 893-8.

Weissman AM. (2001) Themes and variations on ubiquitylation. *Nat Rev Mol Cell Biol*, **2**, 169-78.

Westra WH, Slebos RJ, Offerhaus GJ, Goodman SG, Evers SG, Kensler TW, Askin FB, Rodenhuis S, and Hruban RH. (1993) *K-ras* oncogene activation in lung adenocarcinomas from former smokers. Evidence that *K-ras* mutations are an early an irreversible event in the development of adenocarcinoma of the lung. *Cancer*, **72**, 432-8.

Whang-Peng J, Bunn PA Jr, Kao-Shan CS, Lee EC, Carney DN, Gazdar A, and Minna JD. (1982) A nonrandom chromosomal abnormality, del 3p(14-23), in human small cell lung cancer (SCLC). *Cancer Genet Cytogenet*, **6**, 119-34.

Wikman H and Kettunen E. (2006) Regulation of the G1/S phase of the cell cycle and alterations in the RB pathway in human lung cancer. *Expert Rev Anticancer Ther*, **6**, 515-30.

Wistuba II, Lam S, Behrens C, Virmani AK, Fong KM, LeRiche J, Samet JM, Srivastava S, Minna JD, and Gazdar AF. (1997) Molecular damage in the bronchial epithelium of current and former smokers. *J Natl Cancer Inst*, **89**, 1366-73.

Woelfle U, Cloos J, Sauter G, Riethdorf L, Janicke F, van Diest P, Brakenhoff R, and Pantel K. (2003) Molecular signature associated with bone marrow micrometastasis in human breast cancer. *Cancer Res*, **63**, 5679-84.

Wolf M, Mousses S, Hautaniemi S, Karhu R, Huusko P, Allinen M, Elkahloun A, Monni O, Chen Y, Kallioniemi A, and Kallioniemi OP. (2004) High-resolution analysis of gene copy number alterations in human prostate cancer using CGH on cDNA microarrays: impact of copy number on gene expression. *Neoplasia*, **6**, 240-7.

Wu Z, Irizarry RA, Gentleman R, Murillo FM, and Spencer F. (2004) A model based background adjustment for oligonucleotide expression arrays. *J Amer Stat Assoc*, **99**, 909-17.

Yang YH, Dudoit S, Luu P and Speed TP. (2001) Normalization for cDNA data. In Bittner ML, Chen Y, Dorsel AN, Dougherty ER (eds), *Microarrays: Optical Technologies and Informatics*, volume 4266 of *Proceedings of SPIE*, 141-52. SPIE, Society for Optical Engineering, San Jose, CA.

Yunis JJ and Soreng AL. (1984) Constitutive fragile sites and cancer. *Science*, **226**, 1199-204.

Zanella CL, Posada J, Tritton TR, and Mossman BT. (1996) Asbestos Causes Stimulation of the Extracellular Signal-regulated Kinase 1 Mitogen-activated Protein Kinase Cascade after Phosphorylation of the Epidermal Growth Factor Receptor. *Cancer Res*, **56**, 5334-8.

Zabarovsky ER, Lerman MI, and Minna JD. (2002) Tumor suppressor genes on chromosome 3p involved in the pathogenesis of lung and other cancers. *Oncogene*, **21**, 6915-35.

Zöchbauer-Müller S, Fong KM, Virmani AK, Geradts J, Gazdar AF, and Minna JD. (2001) Aberrant promoter methylation of multiple genes in non-small cell lung cancers. *Cancer Res*, **61**, 249-55.