Genome-wide identification of functional polymorphisms modulating individual risk and prognosis of lung cancer

Thesis

The Open University

FONDAZIONE IRCCS
ISTITUTO NAZIONALE
DEI TUMORI

GENOME-WIDE IDENTIFICATION OF FUNCTIONAL POLYMORPHISMS MODULATING INDIVIDUAL RISK AND PROGNOSIS OF LUNG CANCER

Frullanti Elisa

Thesis Presented To

The Open University of London

for the Degree of Doctor of Philosophy

Discipline: Life Sciences

February 7th, 2011

Director of study: **Dr. Tommaso A. Dragani**

External Supervisor: **Dr. Aage Haugen**

Affiliated Research Centre:

Istituto Nazionale Tumori, Milan (Italy)

DATE of SUBMISSION: 7 FEBRUARY 2011
DATE of AWARD: 4 AUGUST 2011

PAGE/PAGES EXCLUDED UNDER INSTRUCTION FROM THE UNIVERSITY

LIST OF CONTENTS

LIST OF CONTENTS ..... **p. II**

ABSTRACT ..... **p. 1**

STUDENT'S DECLARATION ..... **p. 3**

CHAPTER 1. INTRODUCTION ..... **p. 4**

1.1 EPIDEMIOLOGY ..... p. 4

1.2 CANCER GENETICS AND TUMOUR PROGRESSION ..... p. 10

1.3 POLYGENIC MODEL OF INHERITED PREDISPOSITION TO CANCER ..... p. 30

1.4 ENVIRONMENTAL CONTRIBUTION TO CANCER ..... p. 32

1.5 LUNG CANCER ..... p. 37

1.6 SINGLE NUCLEOTIDE POLYMORPHISMS ..... p. 45

1.7 OBSERVATIONAL STUDIES ..... p. 53

1.8 GENOME-WIDE ANALYSIS ..... p. 61

AIM OF THE PROJECT ..... **p. 70**

CHAPTER 2. MATERIALS AND METHODS ..... **p. 71**

2.1 PATIENTS AND SAMPLES CHARACTERISTICS ..... p. 71

2.1.1 Genomic DNA extraction and quantification ..... p. 75

2.1.2 Total RNA extraction and quantification ..... p. 75

2.1.3 Preparation of DNA and RNA pools ..... p. 76

2.2 POPULATION-BASED ASSOCIATION STUDY FOR LUNG CANCER RISK ..... p. 79

2.2.1 Genome-wide SNPs analysis ..... p. 79

2.2.2 Independent confirmation on DNA pools ..... p. 82

2.2.3 Individual genotyping ..... p. 85

2.2.4 Statistical analysis ..... p. 86

2.3 FAMILY-BASED ASSOCIATION STUDY FOR LUNG CANCER RISK ..... p. 87

2.3.1 Genome-wide SNPs analysis ..... p. 87

2.3.2 Individual genotyping ..... p. 88

2.3.3 Statistical analysis ..... p. 88

2.4 CASE-ONLY ASSOCIATION STUDY FOR LUNG CANCER PROGNOSIS ..... p. 89

2.4.1 Genome-wide SNPs analysis ..... p. 89

2.4.2 Individual genotyping ..... p. 89

2.4.3 Gene expression profile with microarray analysis ..... p. 89

2.4.4 Quantitative Real-Time PCR (qRT-PCR) ..... p. 90

2.4.5 Immunohistochemical Analysis ..... p. 93

2.4.6. Statistical analysis ..... p. 93

CHAPTER 3. RESULTS ..... p. 97

3.1 RESULTS OF POPULATION-BASED ASSOCIATION STUDY FOR LUNG CANCER RISK ..... p. 97

3.1.1 Multiple unlinked SNPs are associated with a decreased lung ADCA risk ..... p. 97

3.1.2 The confirmed SNPs point to a polygenic model with additive and interchangeable effects ..... p. 102

3.2 RESULTS OF FAMILY-BASED ASSOCIATION STUDY FOR LUNG CANCER RISK ..... p. 103

3.2.1 Multiple unlinked SNPs are associated with lung ADCA risk ..... p. 103

3.2.2 Four SNPs were confirmed in population series ..... p. 106

3.2.3 The polygenic model explains lung cancer risk in discordant sib-ships ..... p. 107

3.3 RESULTS OF GENOME-WIDE SNPs ANALYSIS IN CASE-ONLY ASSOCIATION STUDY FOR LUNG CANCER PROGNOSIS ..... p. 109

3.3.1 Multiple unlinked SNPs are associated with lung ADCA prognosis ..... p. 109

3.3.2 Differences in lung ADCA outcome are associated with patients' genetic profile ..... p. 116

3.4 RESULTS OF GENOME-WIDE MICROARRAY ANALYSIS IN PATIENT-BASED ASSOCIATION STUDY ..... p. 119

3.4.1 A gene expression profile of normal lung is associated with clinical stage ..... p. 119

3.4.2 Differential expression profiles of cytokine and cytokine-related genes according to clinical stage ..... p. 123

3.4.3 Differential expression between normal and tumour tissue ..... p. 127

3.4.4 Integration of GWAS and gene expression profiling ..... p. 129

CHAPTER 4. DISCUSSION ..... p. 137

4.1 POPULATION-BASED AND FAMILY-BASED ASSOCIATION STUDIES FOR LUNG CANCER RISK ..... p. 138

4.2 CASE-ONLY ASSOCIATION STUDY FOR LUNG CANCER PROGNOSIS ..... p. 149

CONCLUSIONS ..... p. 164

REFERENCES ..... p. 166

LIST OF FIGURES ..... p. 198

LIST OF TABLES ..... p. 202

LIST OF WEB SITES ..... p. 204

ABBREVIATIONS ..... p. 205

ACKNOWLEDGEMENTS ..... p. 208

APPENDIX ..... p. 210

ABSTRACT

Lung cancer is a leading cause of cancer death in Western countries. Although most cases are due to tobacco smoking, complex genetics may modulate this disease, as suggested by epidemiological studies and findings obtained in mouse models.

Systematic population-based association studies testing several thousands of genetic markers dispersed genome-wide have recently become a powerful and widely-used approach to identify genetic factors affecting common diseases.

In this thesis, the role of genetic polymorphisms and risk of cancer were investigated through a case-control association study in Italian lung adenocarcinoma (ADCA) patients and unrelated controls from general population and through a case-control association family-based study in lung cancer patients and unaffected sibs as controls. I confirmed the relevance of a polygenic model characterized by additive and interchangeable effects of rare alleles in the modulation of individual risk of lung ADCA identifying multiple inherited susceptibility alleles linked to lung cancer.

Additionally, I studied the role of genetic polymorphisms modulating individual lung cancer prognosis through a case-only association study in lung ADCA patients with clinical stage I versus higher clinical stage. In particular, I identify two genes (FCN3 and TMEM100) down-regulated up to 1.8-fold in normal lung of stage>I as compared to stage I patients. These results suggest that clinical stage may be genetically determined as

reflected in germ-line variations as well as in the transcriptional profile of normal lung tissue.

Although clinical application of these results awaits replication in independent and large populations, I found that genetic variants may be involved in the modulation of not only individual risk of lung cancer but also clinical staging. The newly identified individual genetic profiles associated with risk and/or prognosis of lung cancer may thus represent new diagnostic tools and suggest molecular targets for the development of new therapies against lung cancer.

STUDENT'S DECLARATION

The present thesis is the result of collaborations between different groups and researchers of Dr. Dragani's laboratory: in particular, I personally carried out the genetics analysis, whereas my colleagues carried out the statistical analyses.

Parts of the research work included in this thesis were published or have been already submitted for publication in the following papers or manuscripts:

- Galvan A, Falvella FS, Spinola M, **Frullanti E**, Leoni V, Noci S, Alonso MR, Zolin A, Spada E, Milani S, Pastorino U, Incarbone M, Santambrogio L, Gonzalez Neira A, Dragani TA. Polygenic model with common variants may predict lung adenocarcinoma risk in humans. *Int J Cancer*. 2008 Nov 15;123(10):2327-30.

- Galvan A, Falvella FS, **Frullanti E**, Spinola M, Incarbone M, Nosotti M, Santambrogio L, Conti B, Pastorino U, Neira AG, Dragani TA. Genome-wide association study in discordant sibships identifies multiple inherited susceptibility alleles linked to lung cancer. *Carcinogenesis*. 2010 Mar;31(3):462-5.

- **Frullanti E**, Galvan A, Falvella FS, Colombo F, Manenti G, Colombo F, Vannelli A, Incarbone M, Alloisio M, Nosotti M, Santambrogio L, Neira AG, Pastorino U, Dragani TA. Multiple genetic loci modulate lung adenocarcinoma clinical staging. *Clin Cancer Res*. 2011 Apr 15;17(8):2410-6.

- **Frullanti E**, Falvella FS, Colombo F, Galvan A, De Cecco L, Noci S, Incarbone M, Alloisio M, Tosi D, Nosotti M, Santambrogio, Pastorino U, Dragani TA. Lung adenocarcinoma clinical stage is associated with gene expression pattern in adjacent normal lung tissue. Submitted for publication in *Int J Cancer* (June 26, 2011).

The three published papers and the submitted manuscript, listed here, were attached at the end of this thesis, as appendix.

1. INTRODUCTION

1.1 EPIDEMIOLOGY

The word "*epidemiology*" derives from the Greek language "*epi demos logos*" in which "*epi*" means "on", "among" or "upon"; "*demos*" = "people" and "*logos*" = "study of" and so is defined "the study of what is upon the people", suggesting that it applies only to human populations (1). It is now widely recognised as the tool used to measure the public health impact of disease and to study the distribution and determinants of a disease, injury, and other health outcomes in human populations. It is the basic science of preventive medicine involving studying groups of people in order to identify causal or/and risk factors of a disease or trait (2). These factors may be a characteristic of individuals (e.g., their genetic background) or the exposure to external agents.

According to Buck et al. (1988) the first published use of the word "epidemiology" was the Spanish "epidemiologia" in a study of bubonic plague in Spain in 1598 (3). Originally used as the term for the study of epidemic disease, epidemiology is a continually evolving discipline evidenced by the changing definitions. It concerns itself with populations rather than individuals (4). The subject of epidemiology has developed and can be applied to the control of health conditions, disease distribution and threats to public health thanks to the methodological development of techniques such as statistics and clinical epidemiology (5, 6).

The observation that patterns of disease observed in a community are often caused by interaction of several factors in a multiple causation or multi-factorial aetiology of disease (4) gives the opportunity of epidemiological approaches to describe the natural history of specific diseases in populations analysing the aetiological determinants of disease (2). The field of epidemiology has then expanded considerably in scope to cover the description and causation of not only epidemic disease, but of disease in general, and even many common non-disease health-related conditions, such as high blood pressure and obesity.

The history of epidemiology has been documented in 1978 by Lilienfeld (7), which investigated the origins of epidemiology, and described it as the coming together of clinical and statistical sciences. The establishment of epidemiology as a science was also shown to have progressed in tandem with public health developments (8). In 1850 the science of epidemiology was officially "born" by the clinicians when they founded the London Epidemiological Society. These clinicians included John Snow, the "father of epidemiology", who firstly investigated the cholera epidemic in Soho London in 1854 by plotting cases on a map of the area (7). During the first half of the 20th century the field of epidemiology and public health were consolidated on monitoring and tackling major diseases, and began to move towards cancer research. The first case-control study was carried out in 1926 in London and Glasgow by the investigation of aetiology of breast cancer (9). Also during the second half of the 20th century, with improved sanitation, vaccination, and antibiotics, attention turned to chronic diseases such as cancer and coronary heart disease. Doll

and Hill's carried out case-control and cohort studies establishing the aetiological link between smoking and lung cancer (10-12). Towards the end of the 20th century and in the third millennium, the field has further developed and has been applied to pursuit genetic, lifestyle and biomedical phenomena (13); and exploring the wider, social determinants of health and disease (14).

The history of the topic of cancer epidemiological research has been intricately detailed by dos Santos Silva (15), where she noted that the concept of cancer incidence, as a formal topic for scientific study, is relatively new. Until the nineteenth and twentieth centuries, cancer was a relatively rare since it mainly occurs in older people (7). At the beginning of the 19th century life expectancy was around 35 years in Europe. Thus many of those who may have got cancer later in life had died at earlier ages due to infectious diseases, malnutrition, or accidents (7). By the 20th century, however, pathogenesis of cancer was studied, and epidemiologists sought to describe the distribution of the disease in populations and to analyse potential causes (15). The principal purposes of cancer epidemiology are to describe the burden of the disease in various human population groups, generate and test hypotheses on its cause, and testing effectiveness of treatments and interventions (13). With the sequencing of human genome, the huge challenge to understand the complex interactions among genes, environment, and behaviours, in the causation of cancer became central in cancer epidemiology.

Typically, epidemiology is divided into three main branches: descriptive, analytic and experimental epidemiology.

- *Descriptive epidemiology* describes the distribution and the frequency of sanitary events (deaths, disease, etc.) in terms of person, place, and time (2). These three pillars correspond to the questions "who?", "where?", and "when?" and are used to describe and explain health events. Person characteristics include socio-demographic data such as age, ethnicity, education, income, occupational status, and marital status as well as behaviours such as diet, substance abuse, use of health care services, etc. They are used to describe whether a particular risk factor or outcome is more prevalent in one population than another. Place characteristics include geographic location, population density, different features of the geography, and location of worksites, schools, and health facilities. Finally, time characteristics include cyclical changes, long term secular trends, and even daily or hourly occurrences during an epidemic. Sometimes descriptive epidemiology investigated also the questions "what?", and "how many?" (13). Since epidemiology jointly considers person, place, and time, it advances the idea that health and disease as result from the interaction between individuals and their environment. The aims of descriptive epidemiology are to describe the extent and spectrum of disease; describe the natural history of disease; identify disease aetiological factors through generating hypotheses for further study; predict disease trends; identify health needs of a community; and evaluate public health intervention programmes (4). A series of methods have been developed for study design, statistical analyses, data collection, classification, synthesis, tabulation and presentation, followed by inference, and interpretation (13). Descriptive epidemiology has an important surveillance role, particularly in

terms of cancer surveillance. Interpretation of findings from descriptive epidemiology needs to be done with caution and all potential sources of bias, confounding, and artefacts in the data need to be explored. To these ends, it is important that the methods of data collection, collation, and processing are understood (15). Descriptive epidemiology should not be considered an end in itself but should be regarded as a means of monitoring the burden of disease in the population, in addition to generating hypotheses or highlighting areas for further study and investigation. These areas could subsequently be explored using methods of analytical epidemiology.

- *Analytical epidemiology* takes hypotheses generated by descriptive means and tests them through an analytical approach. The main aim is to determine causal factors in the form of aetiological risk factors for a disease, through investigation of exposure and disease outcome at the individual level (2). Analytical studies aims to determine whether particular exposures (variables) such as environmental or behavioural factors (including physical, chemical, or biological agents) are associated to a disease outcome (13). Such an association does not necessarily indicate causation, as chance, confounding and bias need to be considered as possible sources of the relationship (16). Thus mathematical tools and appropriate statistical analytical methods were developed for quantifying and minimizing the uncertainty in the relationship between exposure and outcome. Epidemiologists further test possible bias by teasing out spurious or indirect causes described as confounders by increasing certainty through repetition of observations in different populations; by increasing the number

of subjects under observation which reduces the effects of random variation and uncertainty and through developing a better understanding of the underlying biological mechanisms. The criteria for causation in public health and epidemiology were set down originally by Hill (17). Hills' nine criteria for associations to be considered as causes were: strength, consistency, temporality, specificity, biologic gradient (dose-response ratio), plausibility (biological explanation), coherence (with previous research), experiment (e.g., further indication from removing exposure), analogy (with previous results in other settings) (17). These criteria have been adopted by epidemiologists as a pragmatic approach to assess associations and causation. Finally, it is important to consider and minimize bias that includes any systematic error in an epidemiologic study due to an incorrect estimate of the association (18). Bias comes in many forms and is a particular challenge in case-control studies (see paragraph 1.7) where selection bias can be related to the controls' selection, the comparability between cases and controls, and the statistical efficiency (18).

- *Experimental epidemiology* aims to evaluate sanitary interventions either with preventive objective (e.g., vaccinations, sanitary educational) or therapeutic objective (e.g., testing of new drugs, new surgical techniques) using intervention studies that explore the associations between interventions and outcomes such as clinical trials (2).

1.2 CANCER GENETICS AND TUMOUR PROGRESSION

The origins of the term 'cancer' are in the writings of the early Ancient Greek physician and philosopher Hippocrates (460-377 BC) who

used the Greek word for crab, "karkinoma", to describe the radiating antennae-like growths of the blood vessels extending "out of control in all directions" from some breast tumours (19).

Cancer is a common and devastating disease that represents one of the major public health problems in industrialized countries (20). The disease accounted worldwide to about 7.9 million deaths (around 13% of all deaths) in 2007, with an estimation of 12 million deaths in 2030 (World Health Organization - WHO: http://www.who.int/cancer/en/). The most prevalent form of cancer is lung carcinoma, representing about 29% and 26% of all cancer in men and women, respectively (Fig. 1). Among women, breast cancer is the second most prevalent cause of cancer-related deaths, while among men, the second most prevalent form of cancer is prostate cancer (www. cancer.org; (20); Fig. 1).

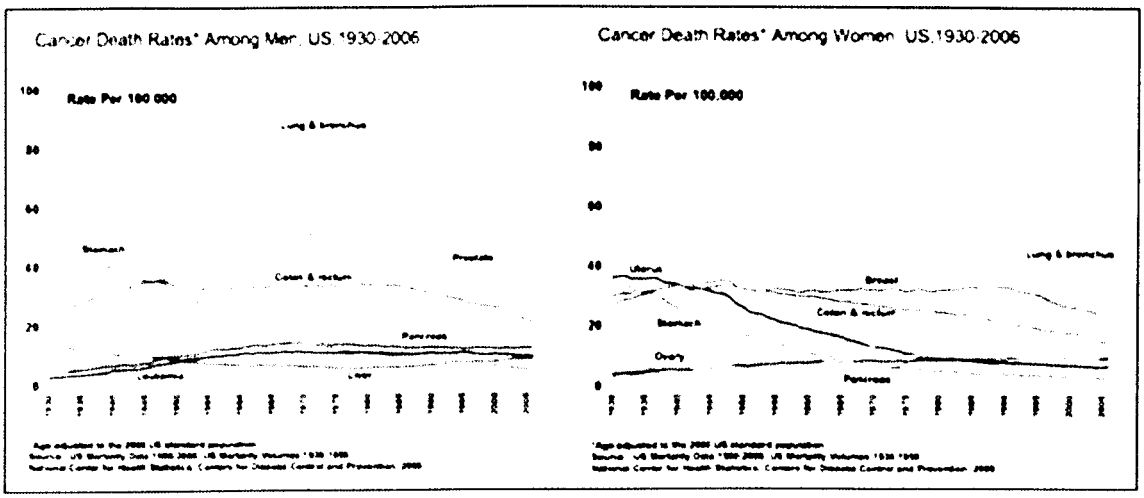

Fig. 1 Cancer related death rates in the United States, from 1930 until 2006 (20).

Cancer is **characterized** by uncontrolled growth and spread of transformed malignant cells which are capable of invasion and destruction of the adjacent tissue, and metastasizing far from the primitive origin through blood or lymphatic vessels. Since the middle of 1900, cancer has

been understood as genetic disease resulting from the dynamic accumulation of several changes that can affect structure and expression of key genes (review in (21)). The list of potentially factors, as we will see, leading to genetic alterations and associated with cancer includes genetics (e.g., family history), behaviour (e.g., tobacco habit) and environment (e.g., radiation).

Tumours grow from a single cell as the result of one or more mutations which confer a selective growth advantage on its progeny through a **clonal evolution process** (22). The transition of a normal cell towards a neoplastic and malignant phenotype is a multistep process influenced by several factors that can occur spontaneously inside the cell or can be induced by external agents (carcinogens) and that can alter either the probability of transformation or the effects of the transforming events. (23). *Internal* spontaneous changes can occur through various genetic and epigenetic mechanisms, such as point mutation, gene amplification, translocation, deletion, chromosomal loss, somatic recombination, gene conversion, or DNA methylation (24). Internal influences include defects in cell-cycle control and DNA repair mechanisms, defects in regulation of epigenetic events, variations in metabolism of exogenous carcinogens and in production or destruction of endogenous mutagens (23). *External* factors are instead represented by environmental exposures to exogenous carcinogens, interaction with surrounding cells and microenvironment, mechanisms of immune system cellular defence against tumour cells, and levels of circulating hormones or growth factors (23). All these causal events may act together or in sequence to initiate or promote

carcinogenesis. Two or more events are necessary before a cell becomes malignant. It has been estimated, for example, that between four and seven rate-limiting genetic events are required for common epithelial cancers development (25). The pattern of alterations that transform a cell is

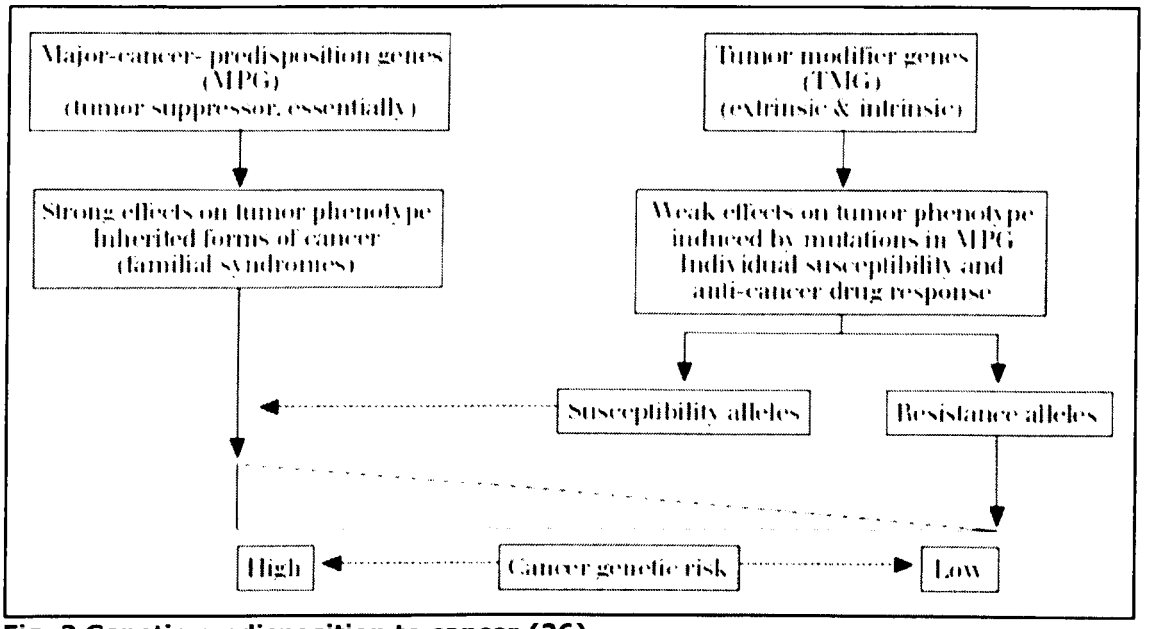

Fig. 2 Genetic predisposition to cancer (26).

not random and is peculiar not only to each type of cancer but can also differ between cancers of the same type. Tumour development seems to be analogous to a Darwinian evolution process in which each genetic change, that confers a growth advantage, is maintained and leads to a progressive switch of normal cells into cancer cells (22). The entire process of transformation can take years to decades in humans (23).

There are two main categories of genes that influence the appearance of cancer (26): major cancer genes with frequent somatic mutations exerting a strong and evident effect on tumour development (*Major Cancer Predisposition Genes* or MPG) and cancer modifier genes (*Tumour Modifier Genes* or TMG) that are characterized by naturally occurring germ line

variants playing a less perceptible role on tumour phenotypes (Fig. 2).

Somatic changes in MPG genes have strong effects on tumour phenotype. These genes can be further divided in oncogenes and tumour suppressor genes.

- *Oncogenes* are characterized by *gain-of-function* events. They become the constitutionally and inappropriately activated counterparts of normal cellular genes, named "proto-oncogenes", that control normal cellular growth and differentiation (review in (21)). In this way, oncogenes encode proteins that strongly promote cell proliferation, increasing the chance that a normal cell will develop into a tumour cell. Mechanisms of oncogene activation range from single-point mutations to chromosomal abnormalities, such as translocation and amplification (24). Oncogenes and their products are highly unregulated in many cancer cells.

- *Tumour suppressor genes* (anti-oncogenes) are involved in tumourigenesis by *loss-of-function* events. They normally functions to limit cell proliferation, so in this case the loss of function takes away the control and facilitates cancer development, usually in combination with other genetic changes (28). Functional activity of tumour suppressor genes can be lost through several mechanisms such as introduction of inactivating mutations, loss of chromosomal material, epigenetic silencing and haploinsufficiency (29). Kinzler and Vogelstein (30) proposed a new subdivision of this vast gene family in two different categories: *gatekeepers* (such as p53, RB and APC) that directly control cellular proliferation by inhibiting growth or promoting cell death; and *caretakers* (such as Mismatch Repair Genes and Nucleotide Excision Repair Genes) that maintain genome

integrity during DNA replication, repair or recombination, in telomeres maintenance, or in chromatin assembly thus controlling cell proliferation and cell apoptosis indirectly. Inactivation of a gatekeeper gene is a limiting step in the initiation of cancer, whereas inactivation of caretakers indirectly promotes tumourigenesis through genome instability that results in an elevated mutation rate of all genes, including gatekeeper genes or oncogenes (30).

Germ line mutations in TMG genes have instead a weak effect and fine-tuning in tumour phenotype modulation, influencing the expression or activity of other genes through allele-specific effects. These genes are capable of either affecting the probability that cancer will develop (conferring susceptibility or resistance), influencing the severity of tumour phenotypes as well as the differential response to environmental compounds or drug treatments (26). They are involved in a variety of functions, such as control of the cellular properties, exposure to carcinogen, diet or lifestyle factors, systemic molecules (e.g., hormones and growth factors), local events affecting cancer cells (e.g., chronic inflammation), or the vulnerability to viruses and bacteria recognized as risk factors for cancer (26). The modifier genes often have at least two alleles, one of which that has no effects on tumour phenotype and one that exacerbates or suppresses disease. The first evidence of the existence of cancer modifier genes was obtained in mouse models characterized for their genetic susceptibility or resistance to spontaneous and induced tumourigenesis. In laboratory animals, modifier effects are usually attributed to genetic background and can be inherited as Mendelian or polygenic traits. At

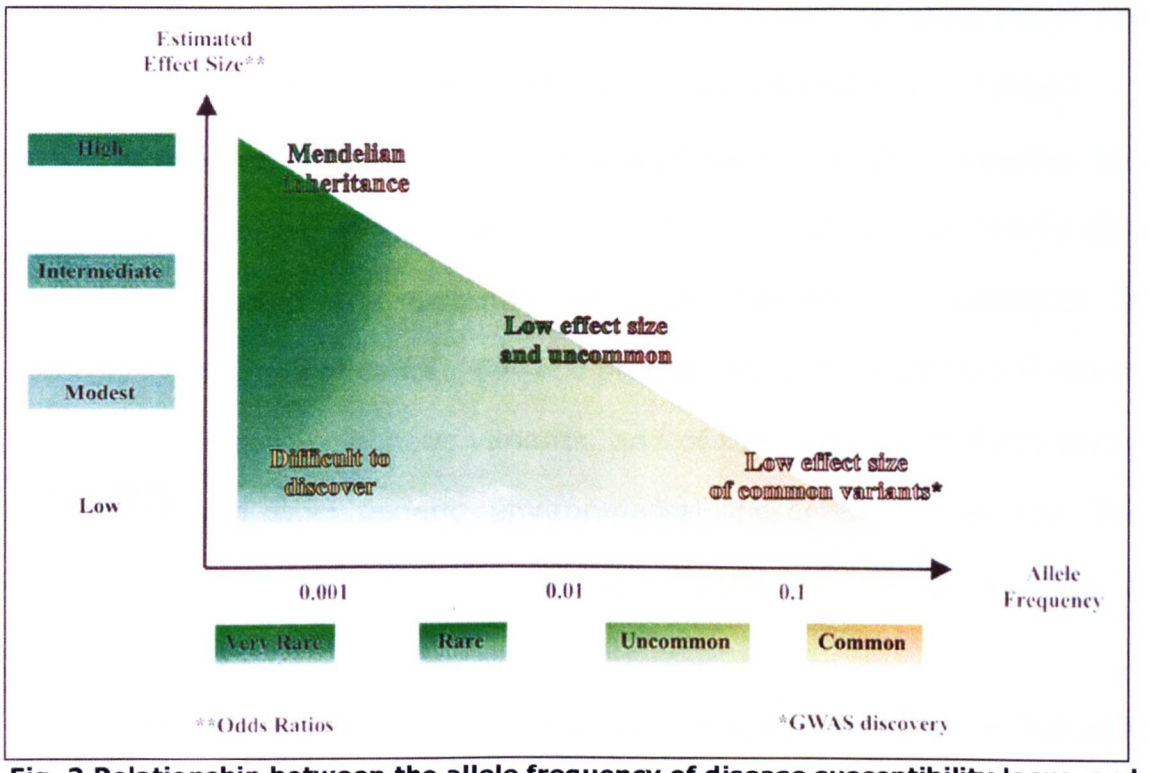

Fig. 3 Relationship between the allele frequency of disease susceptibility locus and their estimated effect size (35).

present, more than 100 mouse loci have been detected that can affect different types of tumourigenesis and at different tumour phenotypes (number, size, stage, latency period, survival time) (31). First indirect evidence for the existence of cancer modifier genes also in humans derived from epidemiological studies reporting the increase in relative risk values for first-degree relatives of cancer patients consistent with a polygenic model of inherited predisposition to cancer (32, 33) and from Crabtree's study (34) reporting segregation in families that could not be explained by germ line mutations. Functional allele variants of major cancer predisposing genes associated with increased or decreased tumour development risk are present in general population and the modulation of tumour phenotype is often due to different combinations of multiple allelic variant of predisposition or resistance (see paragraph 1.3). Identification of cancer

modifier genes could be very important in order to understand the biochemical mechanisms of inherited resistance/susceptibility to cancer.

The interest in **genetic predisposition** to common cancers has constantly increased in the last decades. At the start of the 1990's first findings supported the potential role of hereditary components in determining the risk of cancer. Genetic predisposition is based on the role of one or more genes or genetic variants, and of the interplay of these genes or variants with other genetic, environmental and/or lifestyle factors. The degree of involvement of genetic factors depends on penetrance (Fig. 3 (35)).

- *High penetrance* is due to rare variants (with frequency <1%) with a high effect on risk (e.g., BRCA1, Odds Ratio (OR)~5) (35).

- *Low penetrance* is due to common variants (with frequency >1%) that individually confer a low effect on risk (OR~1.3-1.8). Low-penetrance genetic factors characterize the bulk of inherited cancer risk according to the polygenic model (see paragraph 1.3) and to the hypothesis of "common disease, common variant", which suggests that genetic influences on common traits are at least partially due to a limited number of allelic variants with a frequency more than 1-5% in a population (35).

Based on familial clustering, three main categories of cancer genetic predisposition can be distinguished: inherited cancer syndromes, familiar cancer and predisposition without evident family clustering (23).

- *Inherited cancer syndromes* account less than 5% of all cancer cases consist of rare cancers or combination of cancers with strong familiar history. Examples are retinoblastoma, familial adenomatous polyposis,

Wilms' tumour syndrome, and Li- Fraumeni syndrome. The genetic changes are confined to a particular tissue and take place over several cell generations with specific phenotypic abnormalities. An inherited cancer predisposing genetic mutation is present in somatic cell and germ line cells, and therefore can be passed onto a proportion of the offspring through a well-defined pattern of inheritance as the effect of a single highly penetrant

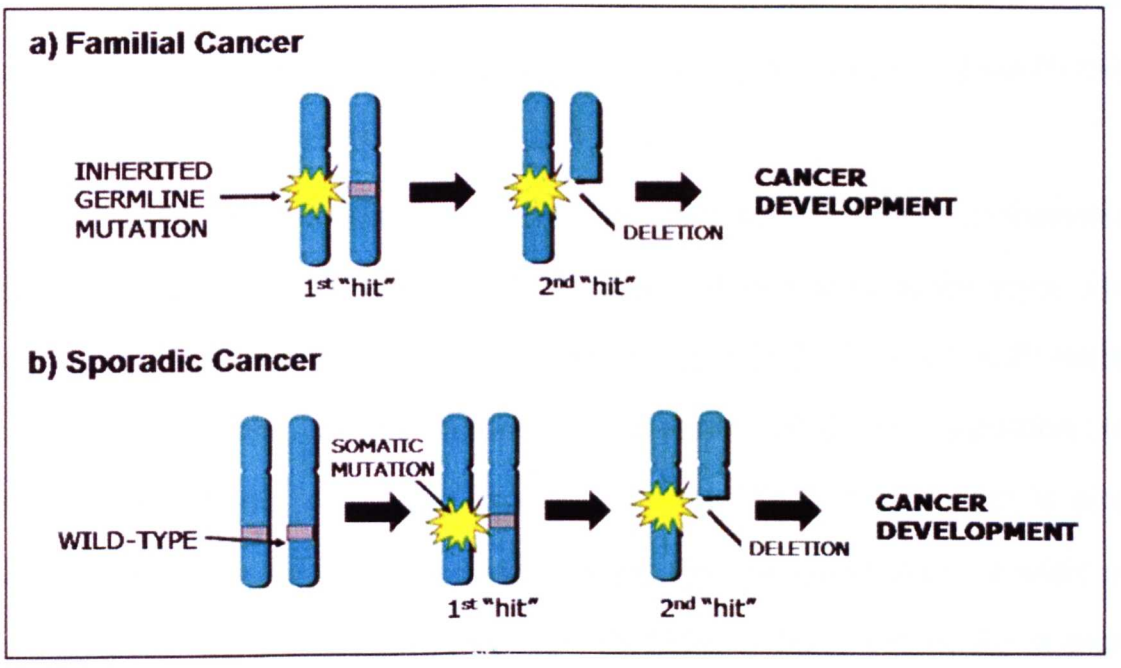

Fig. 4 Knudson's two-hit hypothesis for tumourigenesis.

autosomal dominant allele according to the Mendelian dominant inheritance (36). Knudson explained the genetic mechanism underlying predisposition by highly-penetrant variants studying retinoblastoma (37). Knudson proposed the "two-hits model" (Fig. 4) where the first "hit" affecting the gene responsible for the development of familial retinoblastoma is inherited through the germ line and the second "hit" occurs somatically in the other allele of the same gene (38). Studies have been shown that this second somatic event may arise by a variety of molecular mechanisms, for example

new intragenic mutations, gene deletions, chromosomal loss or somatic recombination (37, 38). The model extends also to sporadic forms of cancer that are explained to initiate only after two somatic "hits" arise independently in the two alleles. Knudson's hypothesis was confirmed when the *RB* gene was cloned on human chromosome 13 and both copies were found to be mutated in the tumours (39). It was understandable that people who inherit an inactivated copy of a tumour suppressor gene had a higher risk of developing the associated form of cancer than people born with two normal copies, as postulated in "two-hit model".

Indeed, it was shown that in the tumours of these predisposed patients, the remaining wild-type copy of the tumour suppressor gene was lost, a process referred to as *loss of heterozygosity* (LOH) (38). LOH leads to either deletion of the tumour suppressor locus or "reduction to homozygosity" (40, 41). Later studies confirmed that this concept is also suitable for other tumour suppressor genes. Genetic variants responsible of these cancer syndromes are very rare (1:1000 or less) but confer a high risk to develop cancer and the age at onset of hereditary cases is, on average, earlier since the inheritance of predisposing genetic mutations through the germ line can accelerate the process of carcinogenesis (42).

- *Familial cancers* are characterized by evident clustering in families of common cancers. The main clinical features of familial cancers are two or more close relatives affected, early age at onset, cancer of specific type occurring together, multiple or bilateral cancers in one individual. The pattern in families is generally consistent with dominant inheritance (23). However, large epidemiological population-based studies on breast cancer

indicate that only 15-20% of the observed familial risk depends on mutations in strong predisposing genes, such as BRCA1 and BRCA2 (43). The remaining 80-85% of the familial risk is attributable to other genetic determinants conferring a low relative risk and to environmental origin.

- *Inherited predisposition without family clustering*, also named as sporadic, is a common trait of the great majority of cancer cases (about 95%). However, the adjective "sporadic" does not mean that there is no hereditary genetic determinant of predisposition, but only that there is no family history. Predisposition to non-hereditary sporadic forms of cancer can be described by **polygenic model** in which the combination of multiple genetic predisposing factors and environmental risk factors has a main role in the pathogenesis of the disease (44) (see paragraph 1.3). In the last years, the scientific community focuses its attention on the hypothesis of "common disease, common variant" in which susceptibility to common diseases, as cancer, is the result of a joint "work" of several common genetic variants each with a low effect (low penetrant variants), rather than a result of rare genetic variants with high effect (high penetrant variants, under the hypothesis of "common disease, rare variant") (see paragraph 1.2).

In the last few years the genetics of cancer predisposition has experienced great progress. The greatest discoveries in the genetics of common inherited cancers relate to breast, ovarian and colorectal cancer. Of particular note are mutations in two genes for breast and ovarian cancer (BRCA1 and BRCA2), in the APC gene for familial adenomatous polyposis

and in several mismatch repair genes for hereditary non-polyposis colon cancer (HNPCC).

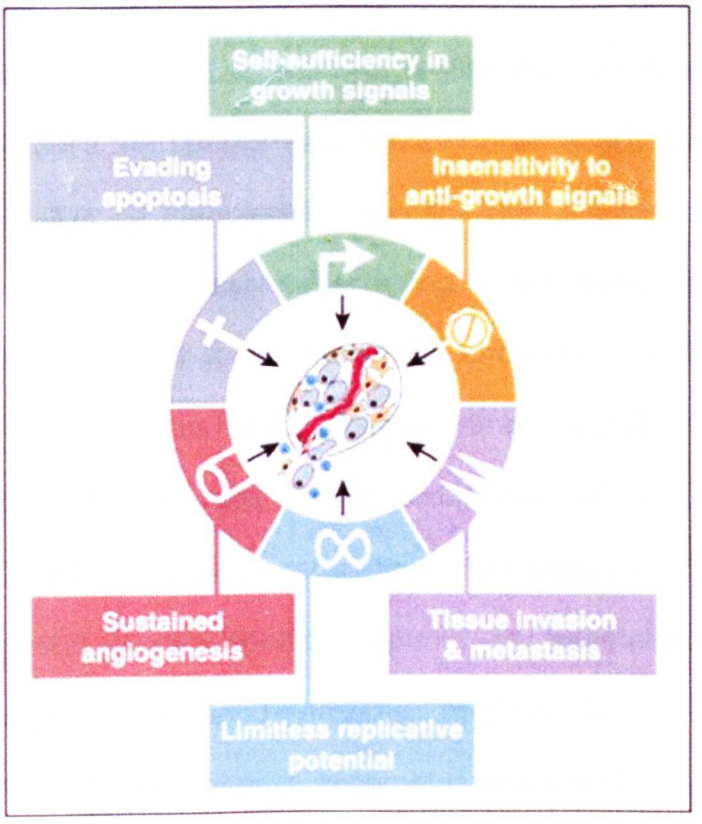

Fig. 5 Acquired Capabilities of Cancer during progression (45).

The malignant cell can be distinguished from its normal healthy counterpart because of **abnormal properties**, shared by almost all cancer cells and that determine the transition towards a more aggressive behaviour. These are known as the "hallmarks of cancer" (45-47) (Fig. 5):

- *Genome instability*: cancer cells escape the mechanisms aimed at the maintenance of genome integrity to acquire an increased mutability and raise the possibility of further mutations (48). Genome instability is attributed to loss of function of genes involved in sensing and repairing DNA damages, in assuring correct chromosomal segregation during mitosis and

in cell cycle checkpoints (49).

- *Limitless replicative potential*: normal cells have a finite replicative potential, named as Hayflick limit (50). After a certain number of divisions, cells stop proliferating and become non-proliferative going into senescence, presumably because the telomeres reach a critical length due to the inability of DNA polymerase to completely replicate the 3'-ends of chromosomes during S phase. The progressive shortening of telomeres during successive cycles of replication leads to chromosomal anomalies, such as end-to-end chromosomal fusions, karyotypic disarray and ultimately to cell death. Tumour cells acquire the capability to proliferate without limit since they maintain telomeres through up-regulation of expression of the telomerase enzyme that adds hexanucleotide repeats at telomerase ends or the Alternative Lengthening of Telomeres method (ALT) that maintains telomeres with inter-chromosomal recombination events. As a consequence tumour cells undergo an immortalization process (51).

- *Loss of differentiation*: metabolic functions necessary for specialized activities often disappear or decrease in tumour cells that seems to evade from anti-proliferative and pro-differentiation signals as they acquire self-sufficiency in growth signals and insensitive to anti-growth signals.

- *Evasion of apoptosis*: the capability of tumour cells to expand is determined by the imbalance of both cell proliferation and cell death. Apoptosis (programmed cell death) represents a major barrier to cancer growth and defects in this mechanism play important roles in a wide variety of tumour types. Resistance to apoptosis can be acquired through several

strategies including increase of growth factor secretion, loss of tumour suppressor genes and oncogene activation. One of the most affected pathways involves the p53 tumour suppressor gene that represents a key sensor in detecting DNA damages and inducing the cascade of apoptotic effectors (52).

- *Sustained angiogenesis*: cancer progression, in the case of solid tumours, is invariably dependent on the formation of new blood vessels from the pre-existing vessels (angiogenesis) because oxygen and nutrients, supplied by the vasculature, are crucial for cell function and survival. It has been recognized that the tumour vasculature often exhibits distinct morphological and biochemical properties as compared to the normal vasculature, including the increased expression of various cell surface proteins (integrins and adhesion molecules), growth factor receptors, and matrix metalloproteinases (45).

- *Tissue invasion and metastasis*: the acquisition of the capacity to escape the primary tumour, invade surrounding tissues and colonize distant new sites is the fundamental definition of malignancy (46). Invasion and metastasis are complex and not completely understood processes that comprise the activation of extracellular proteases and the involvement of numerous cell-cell adhesion and cell-extracellular matrix molecules (CAMs, integrins and cadherins) and cell-microenvironment interactions. All these processes help tumour cells to acquire the capability to detach from the primary tumour, penetrate the basement membrane and the vascular lumen, and survive outside of their normal microenvironment through different adaptation mechanisms (53). Metastasis causes 90% of human

cancer deaths (45).

- *Acquisition of drug resistance*: treatment of malignancies with chemotherapy can be limited by drug resistance of cancer cells. Important mechanisms of drug resistance include apoptosis regulation, cellular stress response, and cell survival signals. Tumours can be either intrinsically resistant to many of cytotoxic agents used in cancer therapy or acquire this property during late stages of development so that therapeutic agents are no longer effective (54);

- *Escape from the host immune system*: the capacity of tumour cells to evade the host immune surveillance involves multiple pathways and mechanisms: reduction of MHC class I expression, loss of costimulatory factors, suppression of the immune response and tolerance development in the host versus tumour antigens (55).

These capabilities are shared by most types of human tumours. The paths, however, which cells take on their way to becoming malignant, are highly variable. Mutations in certain oncogenes and tumour suppressor genes can occur early in some tumour progression pathways and late in others, so that the acquisition of specific biological capabilities may appear at different times during progression (56). Neoplastic cells can remain in a quiescent state or evolve towards more aggressive clinical behaviour and malignant characteristics.

Tumour progression is a dynamic multistep and complex process, which starts with the transformation of a benign into a malignant cell and potentially leads to surrounding tissue destruction and invasion, metastasis and finally death (47). Each step is characterized by the acquisition of new

properties on the level of either single tumour cells or whole tumour tissue (Fig. 6) (NIH, modified from (28)).

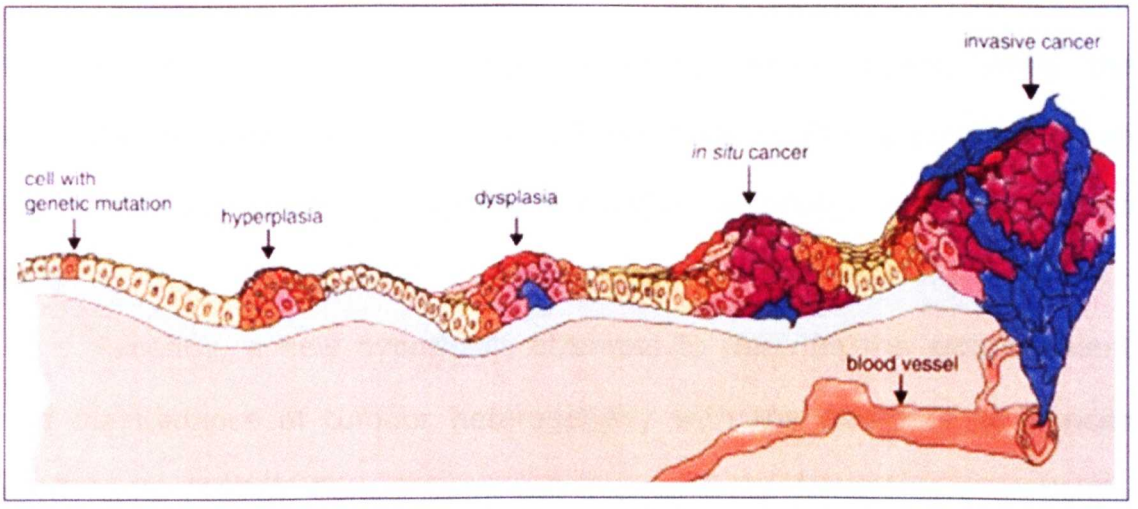

Fig. 6 Stages of tumor progression (NIH, modified from (28)).

Cell migration is the result of a continuous cycle of repetitive steps. First, the cell becomes polarized and it elongates. Cell protrusions containing filamentous actin and structural and signalling proteins are formed, which initiate the recognition of and interaction with the extracellular matrix (ECM). Then, the leading edge or the whole cell contracts, and moves forward. Now the cell has to survive in the blood stream and finally extravasates again to colonize the secondary organ and grow out (57). The acquisition of features above described are the same for all forms of cancer cell, but the molecular mechanism may vary from one invasion pattern to another and the entire process may be quite variable, thus determining the differences in aggressiveness and malignancy among tumours. Indeed, there are tumours that acquire the properties of advanced malignancy before reaching macroscopic size; tumours that may persist for years in a large well-differentiated quiescent state before shifting to a more

malignant state and tumours characterized by a strong heterogeneity with part of the cell population showing a later degree of tumour progression than other components. Changes in the cell-cycle control and immortalization are more significant during early stages, while the alteration on migration and adhesion resulting in the acquisition of an invasive and metastatic phenotype is typically associated with later events (47).

Recently, a new hypothesis attempts to describe the establishment and maintenance of tumour heterogeneity with the existence of **cancer stem cells** (CSCs). Cancer stem cell population is defined as a particular rare subset of undifferentiated tumour cells with stem cell-like properties that are thought to be responsible of tumour initiation, progression, maintenance, spreading, resistance to therapy, recurrence, and metastasis. They are also called cancer initiating cells (CICs). This type of cells is characterized, as normal stem cells, by self-renewal capacity and the ability to differentiate leading to the production of all cell types of a tumour, and thus generating tumour heterogeneity (58). Cancer stem cells are thought to arise from normal stem or non-stem progenitor cells of an organ and to persist in a tumour as a small side population of cells that sustains tumour growth. Although some similarities are evident in cancer stem cell theory and clonal evolution model (22), several differences are evident. CSCs explain tumour heterogeneity with different mechanisms, either by a program of aberrant differentiation or by a competition among neighbours. Under this hypothesis, normal stem and progenitor cells are considered the most likely targets of transformation. The cancer stem cell hypothesis states

that only the "cancer stem cells," contribute to tumour progression, while the clonal evolution model supposes that any tumour cell has the potential to become more aggressive, since all may further mutate. The two theories also explain therapeutic resistance differently: either cancer stem cells are inherently drug resistant or therapy selects for resistant clones (59). However, it has been hypothesized that an integration of the two models should be more successful in oncological research in the next future (60).

Finally, it is important also to focus on the fundamental role of **microenvironment** in affecting the efficiency of tumour formation, growth, invasiveness and metastatic potential. A typical example of the microenvironment leading to cancer is chronic inflammatory status in response to tissue injury (e.g., irradiation) or infection. In fact, many cancers arise from sites of infection such as stomach cancer caused by *Helicobacter pylori* infection in stomach and liver cancer after chronic inflammation caused by *hepatitis C* infection of the liver (61) (see paragraph 1.4). The hypothesis is that inflammatory cells act as powerful tumour promoters facilitating genomic instability and DNA damages through their generation of reactive oxygen and nitrogen species to fight infection (61). In other experiments, injection of non-transformed mammary epithelial cells into irradiated mammary stromal fat pads resulted in increased tumour growth when compared to those injected into contralateral, non irradiated mammary fat pads (62). The authors concluded that irradiation induces no reversible changes in stromal cells altering the microenvironment and leading to tumour promotion. In some cases, the trigger for neoplastic progression is speculated to come from

signals within the stromal microenvironment (62). Summarizing the described data suggest that microenvironment is crucial not only in tumour invasion and metastasis, but also in the earlier steps of tumourigenesis.

Cancer is a heterogeneous disease difficult to treat. For that reason it is necessary to understand the metabolic pathways that are altered in cancer cells in order to adapt therapies with targeting of multiple and/or specific pathways. The primary objectives of **cancer treatment** are cure, prolongation of life and improvement of the quality of life. Treatment of cancer usually includes surgery, radiotherapy, hormonal treatment, immunotherapy and chemotherapy, often used in combination. Despite recent progress in its treatments, so far few types are curable. Thus, cancer is under intense research because of the high prevalence and severe consequences leading to death. Most research aims to apply the knowledge about cancer in order to allow early diagnosis and understanding the mechanism of tumour development.

Two complementary **analytical methods** are used to detect the specific genetic regions and genes that are involved in the disease process: linkage analyses and association studies.

- *Linkage analyses* identify chromosomal regions that co-segregate with the disease in many affected families or over many generations of an extended pedigree. The hypothesis is that the disease locus will lie in the region of the genome that is shared by all affected members of a family or pedigree. Generally, the number of observed crossovers is small unless to have numerous families, or very large multi-generation pedigrees, with the resulting gene being mapped to a large interval (63). This approach is

useful for Mendelian diseases but is particularly unhelpful for complex diseases where the involvement of many genes and the possible influence of environmental factors in the pathogenesis mean that large multi-generation pedigrees or wide detailed family histories are harder to recover. An analysis by Risch and Merikangas (64) suggested that, in a linkage study, the number of pedigrees required to map the genes of minor effect that probably underlie susceptibility to common diseases would be prohibitively large.

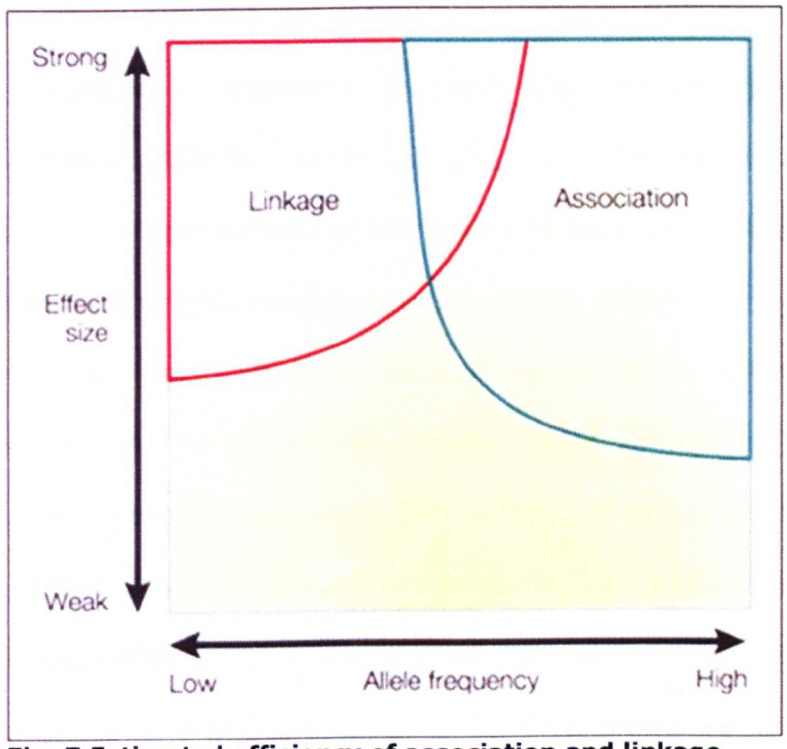

Fig. 7 Estimated efficiency of association and linkage analysis in relationship to the allele frequency of disease susceptibility locus (63).

- *Association studies* (see paragraph 1.7) perform a "genetic dissection of complex traits" without involving familial inheritance patterns but comparing frequency of genetic variants in diseased individuals (cases) and healthy subjects (controls) (65). Generally, tests of association are

more powerful than linkage studies when the disease alleles are common (Fig. 7 (63)).

1.3 POLYGENIC MODEL OF INHERITED PREDISPOSITION TO CANCER

Polygenic traits or diseases controlled by a single major gene or biochemical pathway are called Mendelian or single-gene traits. Complicating factors, such as incomplete penetrance and variable age of onset, are often present in single-gene traits, but they show basic Mendelian segregation patterns. By contrast, the polygenic model of inherited predisposition to cancer is based on the assumption that the combination of multiple genetic predisposing factors and environmental risk factors has a main role in the pathogenesis of the disease (44).

In the past, there was no consensus regarding the genetic model that can account for this increased risk, in particular whether it is caused by aggregation of multiple rare alleles in a subset of genes that have strong effects ("common disease, rare variant model"), by combination of common alleles with weak effects ("common disease, common variant model"), or by some mixture of these hypotheses (43, 66). More recently, complex polygenic diseases and traits result principally from genetic common variants in the population rather than being due to specific and relatively rare mutations, under the hypothesis of "common disease, common variant". The combined effects of many genetic variants, each with an individual modest effect, determine the major portion of susceptibility to cancer. The polygenic model predicts in fact a very high risk for individuals

inheriting the appropriate combination of susceptibility alleles associated with a specific disease, but only a marginal increase in the relative risk of the same disease in the progeny carrying a half of the genetic background of the affected parent. The phenotype is genetically controlled, but it does not run in families, as actually observed for common cancers (44). Moreover, this model is in good agreement with epidemiological studies reporting that the risk of cancer for people with affected first-degree relatives is about 2- to 4-fold higher as compared to those without a family history (44).

Polygenic inheritance models of predisposition to diseases is difficult to demonstrate in humans since genetic heterogeneity, epistasis and gene-environment interactions may mask the role of genetic factors but it has been successfully and extensively studied and demonstrated in animal models (67).

The list of complex diseases controlled by the polygenic model in humans embraces a large fraction of the common causes of morbidity and death and includes atherosclerosis, hypertension, psychiatric disorders, Alzheimer disease, type I and type II diabetes, asthma, rheumatoid arthritis, and cancer (68-73). The difference between Mendelian and complex traits is not in the fact that the involved genes in complex diseases disobey the rules of Mendelian inheritance, but that the pattern of inheritance is not simple (74).

The main challenge for medical genetics in the last decade was to systematically search and identify genes or genetic determinants responsible for the hereditary contributions to complex polygenic traits.

Indeed, the problems in defining specific gene variants that contribute to the propensity for common complex disorders are multiple and difficult. In 1994 and 1995, Lander and Schork (67) and also Weissman (75) in their reviews about the genetics of complex diseases gave a partial list of difficulties, the main of these are the following:

- *incomplete penetrance and phenocopy*: some individuals can inherit a predisposing allele without manifesting the disease (incomplete penetrance), whereas others can manifest the disease without the predisposing allele but as a result of environmental or random causes (phenocopy). The genotype at a given locus may therefore affect the probability of disease but not fully determine the outcome (67).

- *heterogeneity of causation*: the same genes may not be contributing to the disease process in all the affected individuals, thus non overlapping combinations of gene variants may contribute to the increased propensity for the same disease in different individuals (e.g., in breast cancer) (75).

1.4 ENVIRONMENTAL CONTRIBUTION TO CANCER

Differential rates of cancer incidence between different populations and the observation that immigrants tend to acquire the same cancer risk of their new country led epidemiologists to conclude that an important cause of cancer is environmental and that changes in lifestyle and environment could be helpful for prevention (76). Lung cancer also occurs in non-smokers and only about 10% of smokers develop lung cancer. Additional genetic, environmental, hormonal factors and chance (mutation is to some

extent a stochastic process) determine the ultimate development of cancer in mutation carriers (77). Humans are in fact daily exposed to a wide range of potential natural or synthetic toxicants that are carcinogens and so can increase the incidence of cancer. Many environmental causes of cancer are now ascertained, the best characterized being smoking as a risk factor for lung cancer (see paragraph 1.5), alcohol consumption for liver cancer, and intense exposure to sunlight for skin cancer. Generally, the known environmental causes indicated as major etiologic factors in the development of sporadic tumours include exogenous chemicals, diet, workplace, radiation, oxidative agents, chronic inflammation and infections (78-81).

The most relevant **environmental and lifestyle factors** that play a role in tumour development are briefly summarized below (Fig. 8) (81):

- *Diet and nutrition*: it is a general consensus that about 35% of cancers may be preventable by changing our diet. However, no single dietary factor have shown a strong and consistent effect to establish it unequivocally as an important carcinogen or anti-carcinogen, except for drinking alcohol and consumption of foods contaminated with aflatoxin (82). There is a general consensus that some types of cancer are commoner in people who are overweight such as cancers of the oesophagus, colorectal, endometrial, breast, and kidney. A high intake of red meat and fats has been related to increased risk of several cancers such as stomach, and colorectal cancers, whereas alcohol consumption is associated with cancers of the oral cavity, pharynx, larynx, oesophagus, and liver. On the contrary, adequate consumption of fresh fruits and vegetables is regarded as a

protective cancer factor since they contain important antioxidants (76, 83);

- *Oxidative agents*: oxidant by-products of normal metabolism, such as reactive oxygen species (ROS), cause extensive damage to DNA, proteins and lipids. The damages are mostly repaired by enzymes and occur naturally several times per cell and day. Unrepaired damage or modification of DNA bases may cause genetic mutation in semi-conservative replication processes of DNA. Oxidative stress is an important mutagenic or carcinogenic lesion *in vivo* and is associated with as many as half of all

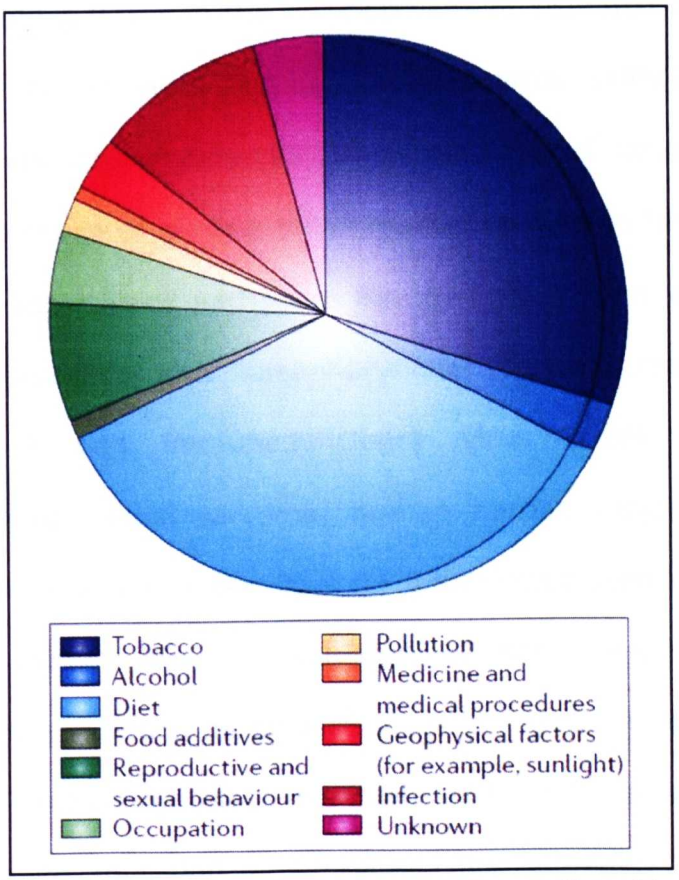

Fig. 8 Proportion of cancer mortality attributable to environmental and lifestyle factors (81).

human cancers (84). Due to higher cellular metabolic rates and deficiency in redox systems, cancer cells may contain increased levels of ROS that continue to generate elevated levels of oxidative DNA lesions which may

disrupt normal cellular replication and lead to double-strand breaks and further chromosome abnormalities. Oxidative endogenous damage is also estimated to be a major contributor to aging and to degenerative diseases of aging, since antioxidant defences remove most, but not all, of these lesions that accumulate in macromolecules with time (78);

- *Chronic infection and inflammation:* different pathogens have been clearly correlated to specific cancer risks. For instance, chronic gastric infection caused by *Helicobacter pylori* (*H. Pylori*) causes gastric ulcers and is a contributing factor in the development of stomach cancer; a subgroup of sexually transmitted human papillomavirus (HPV) is detectable in virtually all cervical cancers (85); hepatitis B and C viruses (HBV and HCV) are a major cause of chronic inflammation leading to hepatocellular cancer (78). Other ascertained pathogens include Epstein-Barr virus (HHV-EBV) for B-cell malignancies and nasopharyngeal cancer, malaria for Burkitt's lymphoma, human immunodeficiency virus (HIV) for non-Hodgkin's lymphoma and Kaposi sarcoma, human herpes simplex virus (HSV) for Kaposi's sarcoma, schistosomiasis for inflammation associated with bladder and colon cancer (79). In addition, it has been demonstrated that inflammatory conditions predispose to cancer since stimulate cytokines and chemokines that contribute to development of malignant disease influencing survival, growth, mutation, proliferation, differentiation, interaction with the extracellular matrix, and movement of cells (61);

- *Environmental chemical carcinogens*: the environment contains many potentially carcinogenic compounds including polycyclic aromatic hydrocarbons (PAHs), heterocyclic amines, and aromatic amines that

represent important classes of carcinogens (85). Like most other xenobiotic substances, these chemical compounds are not carcinogenic *per se* but undergo metabolic activation producing reactive intermediate metabolites that can bind covalently DNA and form DNA adducts leading to genetic mutations. Activation is performed by the phase I enzymes and consists mainly of an oxidation reaction catalyzed by the enzymes of the cytochrome P450 (CYP) or by microsomal epoxide hydroxylase (mEH). If DNA adducts escape cellular repair mechanisms and persist, they may lead to miscoding, resulting in permanent mutations (86). Detoxification is performed by the phase II enzymes, such as the glutathione-S-transferases (GSTs) and the N-acetyl transferases (NATs), which favour the elimination of reactive intermediates by conjugating them with endogenous molecules. Some of these enzymes could play a dual role in detoxification and activation;

- *Radiation*: a small fraction of all neoplasia seems to be correlated to DNA damages of exposure to radiations (e.g., ultraviolet for skin cancer and ionizing radiations for many forms of cancer) (79).

Often ten or more years pass between exposure to environmental factors and detectable cancer. The interaction of susceptibility factors and exposure to carcinogenic environmental agents may lead to the initiation of cancer development. Inherited genetic variants can affect genes that are involved in metabolism of xenobiotics and DNA repair. Additional genes that contribute to carcinogenic process belong to the DNA modification and cell proliferation control groups.

1.5 LUNG CANCER

Lung cancer is an important public health problem and the most common form of cancer in the world accounting for approximately 1.5 million new cases in 2006, 12% of total cancer diagnoses (87), and is a major cause of cancer deaths in the Western countries (20) (Fig. 1). It is characterized by late diagnosis and poor prognosis and therapeutic strategies have shown only a limited effect. Indeed, most cases are diagnosed at late stages often related to metastases. The overall five-year survival rates are only 5-15% (88) and it has not significantly improved in the last 20 years. However, long-term survival of patients who undergo resection of lung tumours at early stages are higher than 80% (89).

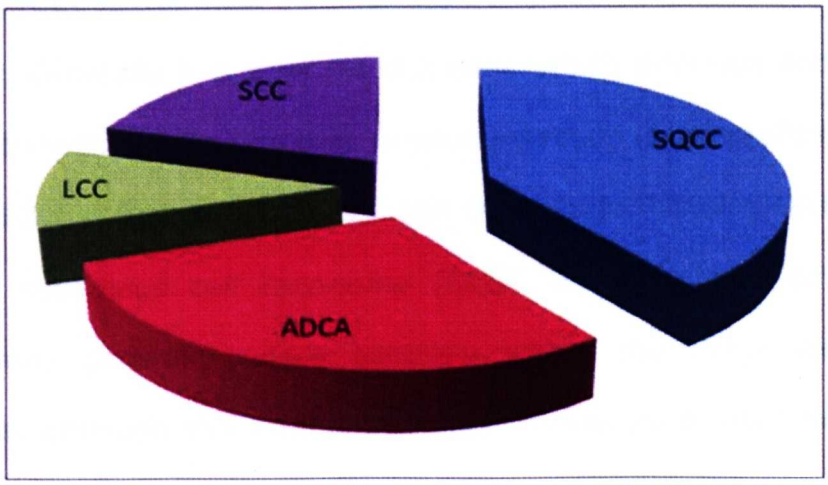

Fig. 9 Incidence of lung histologic subtypes (Modified from (90) and (91)). SCC indicates SCLC.

Lung cancer is generally classified in two major histological types: non-small cell lung cancer (NSCLC) and small cell lung cancer (SCLC) (Fig. 9, modified from (90) and (91)).

SCLC or microcytoma (also named 'oat cell' carcinoma), is so defined because of the characteristic shape of its cells. The incidence of this

histology is about 20% of all newly diagnosed lung cancers. SCLC is most frequent in males and in heavy smokers. SCLC mostly arises centrally in a large bronchus and tends to be more aggressive than the NSCLC. It is characterised by rapid growth and is more likely to spread to other organs and it is usually disseminated at time of presentation (92). Just because of its aggressivity, surgical resection is generally not indicated and systemic therapy is required, especially chemotherapy and radiation therapy. SCLC responds very well to chemotherapy, but nevertheless the disease is recurrent after a period which varies from person to person (92). In almost all cases this type of tumour has a severe prognosis with a 3-year survival of less than 10%.

NSCLC is the most common type and accounts for 75-80% of all lung cancers. Generally it is a localized tumour which develops and spreads out more slowly than SCLC, so that surgical resection is the preferred treatment (90). NSCLC is further subdivided into three major histological subtypes:

- *squamous cell carcinoma* (SQCC or epidermoid or spinocellular carcinoma) generally arises centrally within the lungs inside a large bronchus although the tumour may sometimes be located peripheral and involves the squamous epithelium of lung. SQCC most widespread in men and accounts for approximately 30-40% of all lung tumours (90);

- *adenocarcinoma* (ADCA or AD or AC) tends to occur in more peripheral locations arising from smaller airways but it can be found centrally in a main bronchus and involves glandular tissue forming recognizable glandular patterns. ADCA accounts about 30-40% of all lung cancers and it is the most frequent among never smokers, women, and

young people (93). Today ADCA is the most frequent form of lung cancer in the world, and the frequency of adenocarcinoma is constantly increasing, probably due to a change in cigarette production and composition (94).

- *large cell carcinoma* (LCC) is composed of round large-sized poorly-differentiated cells and lacks the diagnostic features of the other subtypes. It accounts for about 15% of all lung cancers.

The histological distinction between NSCLC and SCLC is also important for therapeutic choices, since there are substantial differences between the two groups in both treatment and prognosis (95).

Lung cancer cells originate from airway epithelia of bronchi, bronchioles or alveoli. Differences in site of origin of lung cancer reflect histological differences. Indeed, the NSCLC histological types all have the phenotypic features of the differentiated cell types in normal or injured bronchial epithelium, whereas SCLC cells have neuroendocrine markers common to endocrine cells that are found in normal bronchial mucosa. Thus, one possibility is that each of the four major histological types arises from alterations in its pre-existing normal counterpart. An alternative hypothesis is that the four types of lung cancer arise from a common stem cell and are related through a common differentiation pathway (96).

Cancer staging describes the anatomical extent or spread of a cancer at the time of diagnosis and attempts to group together patients with similar prognosis. Proper staging is essential to determine the type of therapy and to assess the prognosis. The staging system for lung cancer is based on the TNM (tumour size, lymph nodes, metastasis) classification system, according to UICC (International union against cancer) criteria, and

takes into account the degree of spread of the primary tumour (T), the extent of regional lymph node involvement (N), and the presence or absence of distant metastases (M). Information on each parameter is attributed separately and later combined together to assign an overall stage of I, II, III or IV. Tumours of stage I have a maximum primary tumour size of 5 cm with the exclusion of local or distant metastasis (T1-2 N0 M0). For Stage II cases, the primary tumour has a minimum dimension of 5 cm or extends to the breast wall or skin (T3-4 N0 M0). Stage III includes primary tumours of any size with local metastases affecting lymph nodes (T1-4 N1-2 M0). The highest stage tumours (Stage IV) present distant metastases in liver, skeleton, brain or adrenal glands (T1-4 N0-3 M1) (97).

There are limitations in the use of the TNM classification. For example, recent reports observed that N1 patients actually behave as a heterogeneous subgroup with different lymph node involvement. Indeed, patients with microscopic N1 and single-node N1 diseases show the same survival of patients with pathologic N0 disease, whereas patients with multiple-node N1 disease are similar to N2 patients (98). Such heterogeneity of N1 lung cancer could lead to an underestimation of the effects of genetic variants affecting nodal status, as the "good prognosis" variant could be over-represented in N1 patients. These problems make it necessary to revise regularly the staging system since the development of more accurate diagnostic methodologies can lead to the identification of discrepancies among patients belonging to the same group and allow performing a more homogenous classification. Indeed, TNM system currently in use for the classification of NSCLC was first proposed several

decades ago and has been modified and refined over the years until the last revision was approved in 1997 (99).

Genetic changes acquired by lung cancers are complex and heterogeneous. There are molecular lesions that are common to different lung tumour types and others that are relatively specific to only one of them. For instance, alterations of c-MYC, E2F1 and RB genes are more frequent in SCLC, whereas alterations in EGFR, K-RAS and p16/Ink4 genes are mainly detected in NSCLC, while mutations of p53 can be detected in both histological types (100). Besides smoking, a small number of genetic polymorphisms have been associated with modest increases in lung cancer risk, thus excluding existence of highly-penetrant, strongly-predisposing genetic variants for this type of cancer (101). The main deregulated signalling pathways in lung cancer cells include positive and negative signallers of cell growth and proliferation, apoptosis, senescence, angiogenesis, invasion, metastasis, genomic instability, DNA repair pathways, autocrine and paracrine growth factor circuits (102).

Epidemiological research has convincingly established that **tobacco smoking** is the main cause of lung cancer (10, 11, 103, 104); today we know that about 85% of lung cancer cases arises in current or former cigarette smokers (105). Overall risk of lung cancer for smokers depends on several factors. A lifetime smoker has a 20- to 30-fold increased risk of developing lung cancer compared to a lifetime non-smokers. Risk increases with both the duration of smoking and the number of cigarettes smoked per day, although the former is predicted to have a much stronger effect (106). Smoking cessation results in decreased risk after a lag period of about 7

years (107). However, the decreased risk never reaches baseline levels and risk of lung cancer among former smokers remains elevated as compared to never smokers.

Tobacco smoke contains an array of biologically active components: carbon monoxide, benzene, nicotine, polycyclic aromatic hydrocarbons (PAHs), aromatic amines, N-nitrosamines, aldehydes, oxidative radicals, butadiene, and heavy metals. Tobacco-specific nitrosamines and PAHs are the major risk factors (108). Most tobacco carcinogens require metabolic activation to exert their carcinogenic effects forming mutagenic DNA adducts. Although the predominant cause of lung cancer is well-ascertained (i.e., tobacco smoking), there are other factors known to increase the risk of lung cancer. Exposure to xenobiotics may also increase the risk of cancer. Occupational agents (e.g., asbestos), some metals (e.g., nickel, arsenic, cadmium, chromium), chemical elements (e.g., beryllium), ionizing radiation (e.g., radon), and outdoor and indoor air pollution play an important role in the causality of lung cancer (109-112). Some of these agents act in concert with smoking to synergistically increase the risk.

Although 80-90% cases of lung cancer develops among smokers, only about 10-15% of heavy smokers develop lung cancer and lung cancer are also observed among non-smokers (113, 114) suggesting that **genetic factors** have effects on lung cancer susceptibility.

The first evidence for a genetic control of lung cancer susceptibility and progression comes from mouse inbred strains that provide an essential tool for the dissection of the determinants underlying the complex genetic nature of lung cancer. Several susceptibility and resistance loci have been

mapped in several crosses between different strains. The major locus affecting lung cancer susceptibility, the *Pulmonary Adenoma Susceptibility 1* (*Pas1*), was identified in the distal region of mouse chromosome 6 and it is linked to both tumour multiplicity and volume; therefore, it can affect both lung cancer risk and lung cancer growth (115). The *Las1* gene and the *Kras2* gene were then indicated as primary candidates for the *Pas1* locus, the former affecting lung tumour multiplicity and the latter determining lung tumour progression (116). More recently it has been hypothesized that *Pas1* constitutes a genetic cluster composed of six candidate tumour modifier genes (*Bcat1*, *Lrmp1*, *Las1*, *Ghiso*, *Kras2* and *Lmna-rs1*) and it has been demonstrated that polymorphisms in these genes might confer susceptibility or resistance to lung tumourigenesis (117). Population-based association studies were carried out using genetic markers in the human homologous region on chromosome 12 and demonstrated this locus to be most likely involved in the genetic control of human lung carcinogenesis (118, 119). Mapping in other genetic crosses identified *Pulmonary adenoma resistance* (*Par*) loci that inhibit genetic predisposition to lung cancer provided by the *Pas1* susceptibility allele: *Par1* on chromosome 11 (120), *Par2* on chromosome 18, and *Par4* on chromosome 6 (121). In addition, a locus specifically associated with lung tumour growth was mapped on the central region of mouse chromosome 4 and named *Pulmonary adenoma progression 1* (*Papg1*) (121). Besides the major susceptibility and resistance genes, other minor loci have been mapped confirming that lung cancer in mice is a complex trait controlled by multiple genes with additive and/or counteracting effects (122).

In humans, the inherited genetic susceptibility to lung cancer was first suggested more than 40 years ago following epidemiological evidence for familial aggregation of lung cancer (123). Family-based studies indicate that relatives of lung cancer patients are 2-5 times more likely to develop lung cancer than relatives of control participants (124). This is in agreement with models of polygenic inheritance supporting the role of multiple predisposing genes that modulate the development and growth of neoplastic lesions and the response to environmental carcinogens. Subsequent linkage analyses of high-risk families identified a locus in chromosomal region 6q23-25 as potential lung cancer susceptibility (125, 126). Some tumour suppressor genes (e.g., p53), genes linked to the metabolism of tobacco carcinogens, and DNA repair genes are associated with an increase in lung cancer risk (101), however most associations have not been robustly replicated (127, 128). In the last years, under the hypothesis of "common disease, common variant", several genome-wide association studies (GWASs) identified three main lung cancer susceptibility loci at 15q25 (129-131), 6p21 (132, 133), and 5p15.33 (132, 134), providing further powerful evidence of a genetic contribution to lung cancer, even if with some discrepancies due to ethnicity, smoking habits, and tumour histology (135). More recently, genetic variants at 13q31.3 have been interestingly reported to be associated with susceptibility to lung cancer in never-smokers and to modify the expression of the glypican 5 (GPC5) gene (involved in cell division and cell growth regulation) (136), providing further evidence that the genetic factors for risk in smokers and never-smokers may be different. Together these data strongly indicate that

lung cancer is a complex multi-factorial disease characterized by the interplay of environmental and genetic contribution.

Also in **lung cancer progression**, as for lung cancer susceptibility, it is possible that genetic polymorphisms or other inherited genetic factors, which occur in genes controlling basic cellular process, together with environmental, psychological, social, and biological factors, might have a role influencing neoplastic development and leading to differences in patients' prognosis and in their survival rates. In addition, it has been demonstrated that genetic factors alter treatment response, affecting disease prognosis and outcome (137, 138). In the last years, several inherited genetic variants, are being assessed as predictors of different cancer outcome phenotypes such as MYCL1 for cell growth (139); FGFR4 for tissue invasion (140, 141); VEGF for tumour angiogenesis (142); KRAS (143) and p53 (144) for tumour prognosis. At the moment, except for our work, no GWASs for the identification of lung cancer prognostic germ line variations have been published.

In the last ten years, a lot of progress has been made in the treatment of lung cancer such as adjuvant chemotherapy, targeted therapy, and individualized therapy. However, lung cancer is still today the leading cause of death due to cancer remaining a main medical, scientific, and social problem (90) (Fig. 1).

1.6 SINGLE NUCLEOTIDE POLYMORPHISMS

In 2001 the first two reference versions of the human DNA were published (145, 146), but both these sequences did not report genetic

variants that differ among individuals. Subsequent studies, that completed human genome sequencing, focused on identification of human genetic variants. The HapMap project (147, 148, http://hapmap.ncbi.nlm.nih.gov/) aimed to localize and validate variants throughout the genome.

The most common sequence variations in the human genome are substitutions of a single base called **SNPs** or single nucleotide polymorphisms (Fig. 10) that occur with a frequency of more than 1% in at least one population (149). The different sequence alternatives in a SNP are named "alleles". SNPs could be bi-, tri-, or tetra-allelic polymorphisms. However, in humans, tri-allelic and tetra-allelic SNPs are rare almost to the point of non-existence (reviewed in (149)). Observed data indicate a clear bias towards transitions (i.e., purine-purine or pyrimidine-pyrimidine changes) instead of transversions (purine-pyrimidine or pyrimidine-purine exchanges) (150, 151). One probable explanation is the high spontaneous rate of deamination of 5-methyl cytosine to thymidine in the CpG dinucleotides, leading to the generation of high levels of C/T SNPs, seen as G/A SNPs on the reverse strand (152, 153).

The major conceptual change arose from two critical events early in the 1980s: Kan and Dozy (154) demonstrated how DNA polymorphisms could be identified in non-coding DNA and Botstein et al. (155) proposed that these DNA polymorphisms could be used as the basis for defining molecular markers. Before 1978, all known human polymorphisms were within gene products; however, evolutionary selection on genes does not lead to high polymorphism. The possibility that molecular genetic methods could be used to detect polymorphism within any arbitrary segment of DNA

(154) was of great value since without the constraints of evolutionary selection, polymorphism rates could be much higher in intervening regions than within genes.

In 1980s, restriction enzymes were used to identify single base-pair changes in genomic DNA fragment by the ability of a segment of DNA to be cut, or not, by a specific restriction enzyme that recognises between 4-6 specific DNA base pairs (155). These nucleotide variants were called "restriction fragment length polymorphisms" (RFLPs). The discovery of the Polymerase Chain Reaction (PCR) methodology then made possible the rapid development of highly informative markers for genetic mapping like single nucleotide polymorphisms. Since SNPs are stable (with a low rate of recurrent mutation), frequent, and easy to automatically genotyped, they are the markers of choice for a variety of genetic studies including those on susceptibility to polygenic diseases and poor drug reactions in order to understand disease causation and facilitate a more accurate drug prescribing or development of new drugs (156, 157).

It has been estimated that the human genome contains at least 11 million SNPs, with about 7 million of these occurring with a *minor allele frequency* (MAF) of over 5% (158). The distribution of allele frequencies can vary greatly among different population (159). Depending on their localization, SNPs can be defined as anonymous variants with no effect on gene products or functional substitutions affecting either the amino acid sequence of the protein product or the expression of the gene. Many polymorphisms lie outside genes and are silent, with no effect on gene products (160). It has been estimated that only between 60,000 and

240,000 common SNPs could have a biological effect, as non-synonymous coding variants, regulators of gene expression, or affecting RNA splicing, mRNA stability or mRNA translation (161). Most likely there are substantial differences in SNP densities across the genome with the great majority of variations localized in non-coding regions and having no functional consequences on the activity and expression of proteins (145, 146, 162).

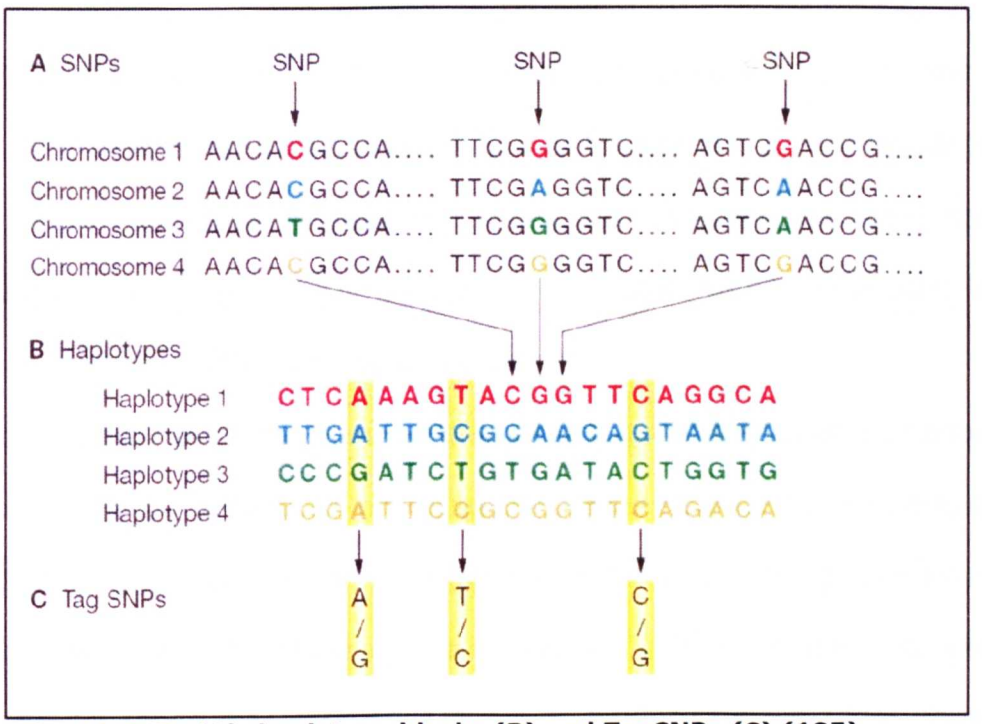

Fig. 10 SNPs (A), haplotype blocks (B) and TagSNPs (C) (185).

It has been observed that SNPs located in the same genomic interval are not inherited independently but often tend to be associated with each other in a set of SNPs called haplotype block (Fig. 10). This correlation structure is named as **linkage disequilibrium** (LD), and refers to the fact that particular alleles at nearby sites can co-occur on the same haplotype more often than expected by chance (63, 163, 164). When a particular allele of one SNP is found together on the same chromosome with a specific

allele of a second SNP, these alleles are said to be in disequilibrium. The extent of LD in populations is expected to decrease with both time and recombination distance between markers. Nevertheless stochastic factors predominate in the behaviour of LD over short distances. Consequently, although a trend towards decreasing disequilibrium with increasing distance between markers has generally been observed in empirical data, closely "linked" markers are not always in LD (165-167). By contrast, in other instances, LD has been reported between quite distant markers (163, 168-170). This variability is due to the fact that the factors governing LD among any specific collection of loci are numerous, complex and only partially understood. A range of demographic, molecular and evolutionary forces have a significant effect on the LD patterns:

- *Genetic drift*: frequencies of genotypes and haplotypes change in a population every generation owing to the random sampling of gametes that occurs during the production of a finite number of offspring, particularly in small populations. In general, the increased drift of small, not growing populations tends to increase LD since haplotypes are lost from the population. But the applicability of this phenomenon to gene mapping has not been well characterized (63).

- *Natural selection:* natural selection can have a hitchhiking effect in which an entire haplotype that flanks a favoured variant can be rapidly swept to high frequency or even fixation (171). Selection against deleterious variants can also inflate LD, as the deleterious haplotypes are lost in the population. The second way in which selection can affect LD is through epistatic selection for combination of alleles at two or more loci on

the same chromosome (172). This form of selection has been shown to lead to the association of particular alleles at different loci in Drosophila, not yet in humans (173).

- *Population structure:* various aspects of population structure are believed to influence LD. Population subdivision is likely to have been an important factor in establishing the patterns of LD in humans, but most of our limited information comes from studies of model organisms (63).

- *Admixture or migration:* admixture is the introduction of mates from one previously distinct population into another. Admixture and migration (gene flow), between populations can create LD. Initially, LD is proportional to the allele frequency differences between the populations, and is unrelated to the distance between markers. In subsequent generations, the "spurious" LD between unlinked markers quickly dissipates, while LD between nearby markers is more slowly broken down by recombination. In theory, this would allow the mapping of disease genes in hybrid populations without using many genetic markers (174-177). In practice, the diseases and circumstances for which this mapping approach will be feasible might turn out to be quite rare and exceptional.

- *Variable recombination rates:* recombination rates are known to vary by more than an order of magnitude across the genome. It is even possible that recombination is largely confined to highly localised hot spots, with little recombination elsewhere. According to this view, LD will be strong across the non-recombining regions and break down at hot spots. There are indications that this reflects the situation for some regions (178), but the generality of the hot-spot phenomenon, the strength of recombination in

and outside hot spots, and the length distributions of these regions remain to be determined.

- *Variable mutation rates:* some SNPs, such as those at CpG dinucleotides, might have high mutation rates and therefore show little or no LD with nearby markers, even in the absence of historical recombination.

- *Gene conversion:* a short stretch of one copy of a chromosome is transferred to the other copy during meiosis process. The effect is equivalent to two very closely spaced recombination events, and can break down LD in a manner similar to recombination or recurrent mutation. It has been recently shown that rates of gene conversion in humans are high and are important in LD between very tightly linked markers (167, 179, 180).

It has been observed that LD varies across the genome. A "haplotype-block" model has been proposed that suggests the genome might be structured into discrete regions of high LD, with a mean size of 5-20 kb in length, separated by regions of recombination hotspots and breakdown of LD (160). Furthermore, LD can vary considerably also among different populations, reflecting the effects of population size, structure and migration history. Some results showed LD between single nucleotide polymorphisms to be usually limited to short distances (3-5 kb) (166, 181), although in certain populations it may extend to longer distances, up to 1 Mb (119, 182, 183).

LD is commonly measured by one of two estimators, D' or r^2, that represent the proportion of variation in one SNP explained by another SNP, or the proportion of observations in which two specific pairs of their alleles occur together. D' or r^2 can range from zero (no association between the

two SNPs) to one (perfect correlation among SNPs) but their interpretation is slightly different (63). D'=1, known as a complete linkage disequilibrium, means that two markers have not been separated by recombination during the history of sample and occurs only when some haplotypes have frequency equals to zero. r^2 summarizes both recombinational and mutational history representing the statistical correlation between two sites. In general, $r^2=1$ is used to measure the statistical association between pairs of markers and reflects the proportion of information provided by one locus about the other and takes into account differences in allele frequencies at the two locus (63).

It has been determined that the majority of 7 million SNPs with a MAF more than of 5% could be reduced about to 550,000 haplotype blocks for European and Asian population and to 1,100,000 haplotype blocks for African population (184). For each LD block, it is determined a **tagSNP** (Fig. 10) (185), a representative SNP, and its genotyping is sufficient to know over 80% of SNPs with a MAF >5% in the same LD block.

In June 2002 Gabriel et al. (186) reported in Science, the construction of a haplotype map (HapMap) of the human genome with the use of common SNP markers and up to now more than 5 million human SNPs were validated with genotyping assay by the International HapMap Project's SNP Consortium (184). Information and data about each SNP and LD blocks are publicly available online (http://hapmap.ncbi.nlm.nih.gov/). The challenge is to determine which genetic variant is responsible for the inherited components of certain phenotype.

1.7 OBSERVATIONAL STUDIES

There are two primary non-experimental, observational study designs which are the mainstay of analytical epidemiology: the cohort study and the case-control association study (13).

Cohort studies or prospective studies are longitudinal population-based studies, in which a group of individuals is identified based on exposure to a suspected risk factor for a disease (*exposure* → *disease*) (187). This group is selected before disease onset and then followed forward in time, together with a group of unexposed individuals, to ascertain the occurrence of the disease of interest, and their individual prior exposure information can be related to the subsequent disease development. With this basic design, there are also a number of different variations based on whether the design is prospective from the present time into the future, or defines a cohort and their experiences from historic records. In addition comparison groups can be identified from within the same cohort (internal group), i.e., those not exposed. When the whole cohort has similar exposure experience, an external comparison group is needed. This is particularly used in occupational cohort studies where a cohort from one company or industry, may be compared to those from another company outside the cohort (187). Since data are collected from a population that is free of disease, it is possible to follow the cohorts of exposed and unexposed individuals from exposure to outcome, and to calculate the incidence of the outcome in both the exposed and unexposed groups (188). Thus, in cohort study the measure of association is the ratio of these two risks, named relative risk (RR) (see paragraph 1.9). Since the

exposure is always assessed prior to disease development, this type of study also allows the advantage of examine rare exposure events, multiple disease outcomes and incidence and relative risk in exposed and unexposed, in that way avoiding problems due to selection bias of control population (15). The main limitations are that cohort studies are expensive and time-consuming, particularly in prospective designs; the need to consider changes in exposure status during the time of follow-up that require repeated measurements; and bias from loss to follow-up, and from outcome information being influenced by knowledge of exposures (information bias). Finally, cohort studies have limited utility in conducting a detailed investigation of risk factors related for outcomes which are rare or have long induction periods. In such circumstances, where a cohort study is not feasible, the best option is a case-control study (15). Cohort studies allow for calculating either *cumulative incidence* (i.e., the number of events per number of exposed individuals per time) or *incidence rate* (i.e., the number of events over a certain time of exposure) (see paragraph 1.9).

Case-control association studies (or retrospective) require two different and quite large groups of individuals, selected on the bases of whether they do (cases) or not (controls) develop a particular disease or trait (*disease* → *exposure*) (189). Under the hypothesis that affected individuals carrier genetic variants associated with disease, the aim of association studies is identifying genetic determinants that make different patients from healthy subjects and so that are more or less frequent in patients (189). The analysis consists of a comparison of allele frequencies between individuals with a disease or trait of interest and disease- or trait-

free comparison group, in search of a statistical difference that can be reflected in an estimated effect size (usually quite small, see paragraph 1.8). If most affected individuals in a population share the same mutant allele at a causative locus in respect to control groups that allele results "associated" with disease. A significant association with risk or prognosis of a disease may indicate that a marker plays a role in pathogenesis or aetiology of the disease or, if this is not a functional marker, it could be in LD with the functional one. In this case, it is possible to perform a fine mapping of the genetic interval around the disease locus in order to find the functional SNP (63). Carriers of a particular disease associated variant will not necessary develop the disease, but they have an increased/decreased risk since the genetic variant confers susceptibility or resistance to given disease or phenotype.

The important aspects of case-control studies are: defining the study hypothesis; definition and selection of cases; definition and selection of controls; measurement of exposures or presence of a genetic variant; analyses; interpretation and reporting. The major strengths of a case-control study include its direct application to humans, its ability to study diseases with a very long latency period, and its "informativeness" and efficiency, such that one study can simultaneously evaluate multiple hypotheses and interactions (15). Another advantage of case-control studies is that they allow the evaluation of casual significance, even with relatively low risk factor exposure or disease prevalence. Rare diseases with a wide-range of potential risk factors are also particularly suitable for case-control design. The case-control study is more commonly used than the

cohort study because it is considered relatively quicker and less expensive to accumulate cases of an outcome of interest and subsequently gather controls who are similar enough to the cases to allow for a comparison of differential exposure (162). On the other hand, the main limitation of case-control studies is their susceptibility to bias. Since case-control studies are mounted when the disease is manifested, disease incidence cannot be calculated and the relative risk cannot be used as a measure of association. Instead, the measure of association between exposure and outcome used as an approximation of the relative risk is the odds ratio (see paragraph 1.9). Two different designs of case-control association studies can be carried out according to the selection of controls to be representative of the study population or for their comparability with cases:

- *Population-based association studies* look for differences in frequency of genetic variants between affected individuals and unrelated healthy controls testing for the co-occurrence of a marker and disease at the population level. Population controls are also considered more suitable than hospital controls, as they avoid the bias arising from the factors which lead people to use health services, although cost and effectiveness in terms of participation are recognised issues (190). Exposure variables are ascertained through questionnaires, interviews and examination of health records. Interpretation of results of case-control studies should be done taking into account the potential biases in the form of selection bias in the choice of cases and controls, information bias in the collection of data, and confounding factors (16). Ideally, a case-control association study should match cases and controls by ethnicity, age at disease onset, gender, and

smoking status in the case of lung cancer, since the risk of tobacco smoking seems to be higher than any risk factor (191). A high quality case-control study can provide informative results, if cases and controls can be selected independently of the exposure and controls are selected at random from the same defined study population as the cases came from, thus the results would be unbiased and equivalent to a cohort study.

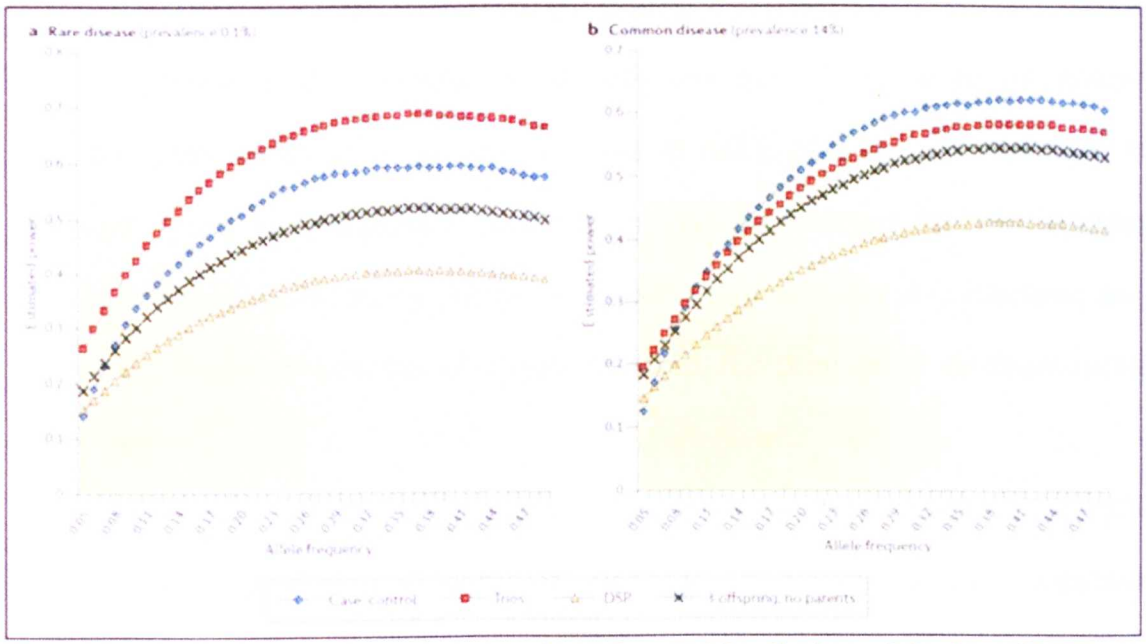

Fig. 11 Estimated power in case–control studies and family-based designs (192).

- *Family-based association studies* generally test associations using genotype information from affected individuals and both of their parents ("trio design"), estimating the frequency with which an allele is transmitted to the affected offspring. When parents are missing, an alternative family-based design is looking for genetic differences between affected individuals and their unaffected sibs as control. The discordant sibling pairs (DSPs) design is less powerful than trio in case of rare disease; but is more efficient when the prevalence of disease is high (192) (Fig. 11). Since cases and controls derive from the same pedigree, family-based studies are not biased

by population admixture and stratification, and the observed DNA differences in genetic polymorphisms are putatively responsible for the disease status. Thus, it may represent an alternative design to population-based association studies in studying sporadic cancer. However, the poor feasibility of the recruitment of healthy sibs or parents for cancer patients, due to late age at cancer onset for most of sporadic cancer cases, make difficult to carry out such type of study (192).

In those studies aiming to identify the genetic variants correlated with the progression of a disease (instead of risk), a *case-only approach* is required since no prognostic parameters can be defined for the control group. In this specific study design, frequencies of polymorphic markers are compared between subsets of cases selected for their poor or favourable prognosis.

Association studies have been widely used in the attempt to identify genetic loci contributing to complex diseases. However, so far, negative results have been more frequent than positive outcomes and the main criticism of this approach relies on the lack of replication of significant findings in independent studies (193). The absence of reproducibility is generally ascribed to inadequate statistical power, biological and phenotypic complexity, population-specific linkage disequilibrium patterns, population stratification, and other biases that can lead to spurious associations (194).

Inadequate statistical power of single studies in detecting weak effects of common variants (195) may be partially resolved using meta-analyses that summarize data of previous independent studies (in order to increase the sample size under analysis). Indeed, if the power is low, there

is a low chance to detect a difference between groups (or an association) if one exists (196). The reliability of results from meta-analyses depends on the validity of the primary studies included and on rigorous methodology. Meta-analyses suffer of several limitations, such as potential heterogeneity of the studies in the diagnostic criteria, patient selection, laboratory methods (197). However, these combined analyses present the advantage of an overall assessment of the potential role of a given polymorphism in a specific disease increasing the power of single studies. Indeed, most of the confounding variables present in individual studies, such as population stratification and population-specific LD, are expected to balance and reduce their effects (198, 199).

The problem of *population stratification* (presence in population of distinct groups with limited inbreeding) arises when cases and controls are unknowingly sampled from different populations or variations in allele frequency between groups are present (200). For instance, if a disease is unique to (or more frequent in) one population and controls have a different ethnic origin, an *association* study will most likely produce a positive result at many loci throughout the genome reflecting the "genetic distance" between the two populations rather than a real correlation between variants and the investigated phenotype. The disparity in frequencies among populations is a well known event and arises from genetic and social features unique to each population. However, the amount of bias attributed to stratification is likely to be small and not substantial in case-control studies with unrelated controls (201). Stratification can be controlled using either family-based controls or testing a set of unlinked genetic markers in

the study population (202). Thus, if frequency differences are observed for randomly selected and anonymous markers, one could infer that populations have genetic differences consistent with stratification.

Biases affecting association studies fell into three broad categories: recall (information) bias, selection (including response) bias, and analytical bias (including confounding effects) (13): *recall* or *information bias*, where the case subjects have a differential ability to remember details about their past life history and this affects the accuracy of information; *selection bias* relates to the way cases and controls are selected or not. If they are not representative of the population from which the cases come, the results are likely to be distorted. *Analytical bias* issues include the potential problems of lack of precision and validity of results can be improved by increasing sample size (utilising a pre-study power calculation) and by getting better study design or efficiency (including matching control group e.g., by age) (13).

Finally, *confounding* is the most important consideration in the analysis and interpretation of case-control studies. Basically, confounding variable is related independently to the risk factor or exposure and the outcome variable under investigation, and can create an apparent association or mask real one (203).

Although association studies suffer from several limitations and need corrections in order to gain more power and reliability, positive findings have been published and support the use of this approach (204).

As with Consolidated Standards of Reporting Trials (CONSORT) guidelines, which are widely adopted and improved the quality of clinical

trial reporting internationally (205), a group of epidemiologists in Europe have begun to develop similar guidelines with similar aims for reporting observational studies. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines (206), set out standards for reporting of observational studies, including case-control studies for all the issues covered above (http://www.strobe-statement.org/).

1.8 GENOME-WIDE ANALYSIS

Genome-wide association study is the study of genetic common variation across the entire genome that is designed to associate genetic variations with phenotypic traits (such as blood pressure or weight) or with the presence or absence of disease or condition (207). The National Institute of Health defined GWAS as *"Study of common genetic variation across the entire human genome designed to identify genetic associations with observable traits"* (208).

Interest in GWASs started in 1996, when Risch N and Merikangas K, reviewing the statistical framework of association studies, evidenced that association studies have greater statistical power than linkage analysis to detect genetic variants with small or moderate effect on a disease or trait testing a large number of variants across the genome (64).

GWASs represented the most widely used approach to study relationship between genetic variations and phenotypic diversity (209) based upon the "common disease, common variant" hypothesis (see paragraph 1.2). Important disease-causing variants, that are rarer than 1-5% in population, are not detected with GWA approach.

As a traditional association study, a typical GWA study consists of four principal phases: selection of individuals with the trait of interest and a reference group for comparison; DNA isolation, genotyping and data review; statistical analysis for association between SNPs and trait of interest and replication of identified associations or their functional characterization (Fig. 12) (210).

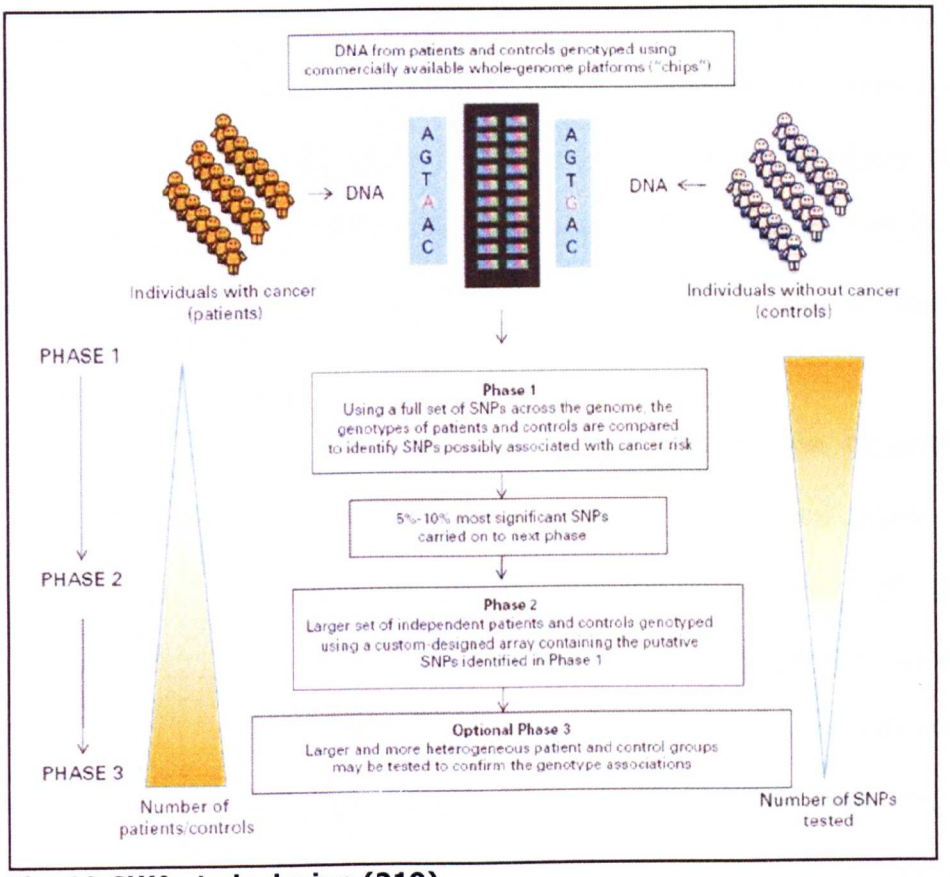

Fig. 12 GWA study design (210).

Over the past 3 years these studies have identified significant statistical association between more than one hundred of loci across the genome and common complex traits. Over the last years, the number of complex traits investigated through GWA studies is widely increased with a total of 658 genomic regions identified for 135 phenotypes up to December

2009, of which 96 are cancer hits with frequency of more than 10% (http://hugenavigator.net).

Most of cancer associated loci are tissue-specific, but some are also common in different types of cancer, such as the 8q24 region in prostate, colorectal and breast cancer, suggesting possible unsuspected relationships and common pathways among different diseases not previously implicated, such as the autophagy pathway in inflammatory bowel disease (211). Other loci seem associated only with certain histological subtype. For example, the locus 5p15.33 was found significantly associated in ADCA histotype but not in SQCC (212, 213). Up to now, three regions were identified with by GWA studies associated with lung cancer (see paragraph 1.5): 15q25, 5p15 and 6p21 (reviewed in (134)).

Of course, whole genome information may offer the potential for discovery of new regions associated with disease (individual SNPs, gene-gene interactions, high-risk haplotype) establishing utility of genetic markers for risk and outcome prediction. In addition, it increased understanding of basic biological processes and molecular pathways of disease causation with the future promise of personalized medicine, differential pharmacological intervention (pharmacogenetics) and new drug targets. Once genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease.

The success of GWA study is due to the development and upgrading of the high throughput **SNP genotyping platforms** commercially available, that allow genotyping of hundreds of thousands of tagSNPs with their

relative comprehensive annotations and that have constantly increased the number of variants that can be typed at once. The main companies producing SNP platforms are Affymetrix (Santa Clara, CA, USA) and Illumina (San Diego, CA, USA) that developed a last array containing more than one million of tagSNPs and offer a coverage of 67-89% of SNPs with a MAF >5% in European and Asian populations and of 50% in African population (185). Even with a very low cost for each SNP, the total cost for genotyping 1 million of SNPs in a large sample size is prohibitive and **DNA pooling approach** can be used as initial screening in order to reduce costs as compared to the analysis of individual samples at the same power of study and with a robust estimation of allele frequency (214, 215). In this way, an equal amount of DNA samples from cases and controls were pooled and genotyped together to determine imbalance among allele frequencies of the two groups.

Despite its success, GWA studies have several limitations, that are typical of a traditional association study (see paragraph 1.7), but more evident in this high throughput approach. Indeed, SNP associations identified in one population are frequently not transferable to other populations because both allele frequencies and LD blocks are different from population to population (216). However, most of associations found with GWA analyses have problems on reproducibility in different series from the discovery, even though within the same population. This could be due to the bulk of the **genetic heterogeneity** or existence of phenocopies across individuals from the same population that have not been accounted for in GWA studies and whereby multiple variants in the same or different loci can

contribute to the same phenotype. In this case, failure in replication does not mean that the initial findings was spurious, but that replication series was substantially different from the discovery and thus invalidates test for association giving different results. In addition, **population stratification** or selection for subgroups reflecting population history with different characteristics alters association analyses, providing associations even for unlinked loci (false-positive or spurious associations). Moreover, there is great difficulty moving beyond statistical associations to identifying the **functional and biological explanation** of link between a genomic locus and a given complex trait. It is important also to call attention to the fact that GWA approach identifies significant statistical associations for a tagSNP of entire LD block and does not give information about the exact associated polymorphic or structural variant in the region and moving from tagSNPs to disease causal variants is often difficult. Moreover, most associated SNPs are not localized within a gene or regulatory regions; markers are often located in introns or in intergenic region. For example, variants on 8q24 that were found associated with multiple solid tumours risk is 300 kb from the nearest gene (MYC) (217, 218). Another debated topic any of GWA studies tend to minimize false-positive associations paying attention only on the highest statistically significant associated SNPs and carrying over from initial screening into the replication step only these. Indeed, the most robust findings are often not in the "top" associations and this approach can cause false-negative results. Another hotly debated topic of GWA approach is that variants found to be associated with a given disease have a limited impact on its susceptibility. The mean contribution to the overall risk

variation estimated for most of these associations is modest (OR in order of 1.1-1.3 for heterozygous genotype and 1.5-1.6 for homozygous genotype, http://hugenavigator.net). Each variant by itself has in general a small effect; however, the combination of several low-risk alleles could have a strong effect and identify individuals with a substantially increased risk (219).

On the other hand, in the last years, only a small part of inherited risk has been explained by GWA studies (220). Estimates of residual missing heritability suggest that numerous other variants, including rare, structural or other common variants, remain to be found for most of complex disease.

To perform an excellent GWA study we need to develop a **robust study design** to obtain a high power to detect genes of modest risk minimizing the potential of false association signals due to testing large numbers of markers and to high genetic heterogeneity interindividual. The key components are:

- *Sufficient sample sizes*: since the relatively modest effect sizes of common genetic variants in modulation of complex disease susceptibility/resistance, very large samples sizes need to detect them. Since statistical power of a study is function of MAF, sample size and supposed genetic effect, a GWA study from general population (with MAF > 5% and OR in order of 1.2-1.5) requires more than 10,000 individuals for group (221). Ioannidis JPA et al. (222) estimated that a median sample size of 15,000 participants is needed to have a power of study of 90%.

- *Rigorous phenotypes* (cases and matched controls): since misclassification of case and control participants can widely reduce study power, the two groups should be careful defined, selected and matched for confounding factors (gender, age, smoking, ethnicity, etc.) in order to focus on differences really associated with the given trait and to minimize phenotypic heterogeneity and population stratification (35). For lung cancer, where the contribution of tobacco smoking is important in definition of risk, it could be useful to use smokers as controls.

- *Accurate high throughput genotyping technologies* (comprehensive maps, rigorous assessment of genome-wide signatures, rapid algorithms for data analysis): quality control measures for genotyping step require a SNP call rate > 95%, concordance in replicas > 99.5%, MAF > 5%, Hardy-Weinberg equilibrium and Mendelian inheritance in trio studies. Statistical analysis of dense genotyping data can be performed with publicly available tools as SNPLims (http://www.itb.cnr.it/snplims), Genotype Library and Utilities (GLU) or PLINK (223), that allow archiving, management and basic analysis of datasets.

- *Replication*: an important step in evaluating the reliability of results is replication of initial associations study in independent series in the same population changing recruitment centre, genotyping platform or method and then extending results in other population. In the last years much interest has been focused with contrasting results on the advantages/disadvantages of splitting the initial available series of samples in two series: a "testing series" for GWA study and a "validation series" to perform the replication step for the most promising SNPs (194, 224, 225). An innovative approach

to the replication-based analysis, proposed by Skol AD et al. (226), could be to jointly analyze the results from both series. The joint analysis provides greater power than replication-based analysis of only the replication series, although a more stringent significance level is required. Replication represents the major challenge of GWA studies in the last years since the extensive lack of reproducibility.

- *Functional studies*: it is important investigate the functional and biological value of statistical associations found with a genome-wide approach to confirm their role and increase understanding of their mechanisms and their possible interactions with other genes or environmental factors. Indeed, GWA studies find significant statistically association for a tagSNP of an entire LD block and without giving suggestions about the exact associated polymorphic or structural variant in the region. Moreover, most associations are often located in introns or in intergenic region, rather than within a gene or regulatory regions such as the 5′ or 3′ untranslated regions (UTRs), the promoter, or the splicing donor/acceptor sites (218, 227).

Software provided with the SNP platforms is sufficient for management of row data while management and statistical analysis of data post-genotyping could be done with publicly free tools (e.g., PLINK), that allow tests for allelic, genotypic, dominant, recessive or additive model associations with permutations, multiple testing corrections, and test for LD and Hardy-Weinberg equilibrium (HWE) analyses.

Data from GWA studies are available consulting the Database of Genotype and Phenotype (dbGaP) of the National Center for Biotechnologies

Information (NCBI) (228) (http://view.ncbi.nlm.nih.gov/dbgap) and updated information on published GWA findings are released online to the scientific community through the National Genome Research Institute site (http://www.genome.gov/GWAstudies/).

AIM OF THE PROJECT

In this project I examined the genetic profile related to lung cancer risk and prognosis. More specifically, the overall aim of this project was the identification of genetic profiles predictive of individual risk of lung cancer or associated to patients' prognosis through genome-wide analysis of DNA and RNA pools from different groups of Italian lung cancer cases and controls, followed by individual genotyping of candidate SNPs and by individual assessment of the transcript levels of candidate genes.

According to these aims, my project is divided in two fundamental tasks. The first task investigates human genetic variants that may play a role in lung cancer risk through GWA in a case-control association studies in Italian lung ADCA patients and unrelated controls from general population and a case-control association family-based studies in Italian lung patients and unaffected sibs as controls. The second task investigates a genetic profile that can explain the differences in cancer prognosis through GWA in a case-only association studies in Italian lung ADCA patients with clinical stage I versus higher clinical stage and a whole-genome expression profile in normal lung tissue of these patients.

2. MATERIALS AND METHODS

2.1 PATIENTS AND SAMPLES CHARACTERISTICS

The entire project involved pathologically and clinically documented Italian lung ADCA patients who underwent surgical resection at three Institutes in Milan (Italy): Istituto Nazionale Tumori, Istituto Clinico Humanitas and Ospedale Maggiore Policlinico. Control subjects from general population were enrolled among healthy blood donors or subjects participating in a computed tomography screening for lung cancer prevention (272) at the same Istituto Nazionale Tumori (Milan, Italy) and matched to the group of cases for their district of birth, age at diagnosis, gender, and smoking status. Characteristics of lung ADCA patients and control subjects used in the population-based case-control association study are summarized in Table 1.

Table 1. Characteristics of lung adenocarcinoma patients and control subjects population-based association study.

Subject characteristics		Controls	Cases
No. of subjects		522	482
Median age (range) [a]		59 (31-77)	63 (34-77)
Gender			
	Male	389	361
	Female	127	121
Smoker status			
	Never	25	68
	Ever	485	398
Clinical stage			
	I	NA	252
	II	NA	85
	III	NA	93
	IV	NA	43

[a] Age in years. NA, not applicable.

For the family-based case-control association study, series consisted of 80 Italian lung cancer patients and their healthy sibs as control. This population was recruited, on a voluntary basis, with the help of Marta Nurizzo Association (Brugherio, Italy, http://www.martalive.org/foreign.htm) according to recruitment criteria consisting in the non-smoking status and young age (<60 years) of lung cancer cases (Table 2).

Table 2. Characteristics of discordant sibs series in the family-based association case-control series.

Subject characteristics	Controls	Cases
No. of subjects	80	80
Median age (range) *	51 (31 - 73)	52 (31 - 80)
Gender		
Male	29	19
Female	51	61
Smoker status		
Never	54	75
Ever	26	5
Histological type §		
ADCA	NA	47
NSCLC	NA	30
SCLC	NA	3
Clinical stage	NA	Unknown

* Age in years. § ADCA, adenocarcinoma; NSCLC, non-small cell lung carcinoma; SCLC, small-cell lung carcinoma; NA, not applicable.

Lung cancer patients used in case-only association study consisted of pathologically documented 1174 Italian lung cancer patients distributed in a first series composed of 600 lung ADCA patients (discovery series) and in two additional independent validation series composed of 317 lung ADCA and 257 lung SQCC patients (Table 3).

Gene expression profile analysis in normal tissue was performed in a series of RNAs from 120 lung ADCA patients derived from the discovery seires of previous GWA, divided in two groups according to their clinical stage (I or >I) (Table 4). We selected only smokers to avoid bias in gene expression associated to the smoking habit (273).

Table 3. Characteristics of lung cancer patients in case-only association study.

Subject characteristics	All patients (N=1174)		
	Discovery ADCA series	Validation ADCA series	Validation SQCC series
No. of subjects	600	317	257
Age at diagnosis (years)			
Median	63	65	67.5
Range	20 - 81	34 - 84	44 - 84
Gender			
Male	442	233	233
Female	156	84	21
Smoker status			
Never	98	50	7
Ever	494	262	241
Histological type			
ADCA	600	317	0
SQCC	0	0	257
Clinical stage			
1	300	160	109
>1	300	141	130
Follow-up at 60 months			
No. patients alive	316	183	157
Median duration (months)	59.1	60	55.8
Range	4.4 - 60	1.7 - 60	1.9 - 60

For 27 out of 120 cases, we had also available the matched lung ADCA tissue for analysis of gene expression in matched couples of lung ADCA tissue and adjacent normal lung tissue (Table 5).

Files were recorded to get personal and clinical data. Study protocols were approved by the institute ethics committee and written informed

consent was obtained from each subject for the use of their biological samples for research purposes.

Table 4. Characteristics of lung ADCA patients used in gene expression profile analysis of normal tissue.	
Parameter	Values
No. of patients	120
Median age (range) [a]	65 (36 - 81)
Gender	
Male	99
Female	21
Smoking status	
Never	0
Ever	120
Clinical stage	
I	60
II	15
III	35
IV	10
[a]Age in years.	

Table 5. Characteristics of 27 out of 120 cases lung ADCA patients used for paired analysis of the gene expression of lung ADCA tissue and adjacent normal lung tissue.	
Parameter	Values
No. of patients	27
Median age (range) [a]	63 (44 - 76)
Gender	
Male	23
Female	4
Smoking status	
Never	0
Ever	27
Clinical stage	
I	13
II	4
III	6
IV	3
[a]Age in years.	

2.1.1 Genomic DNA extraction and quantification

Genomic DNA was extracted from peripheral blood sample or from a small piece of non-tumour lung parenchyma excised during surgery using the DNeasy Blood & Tissue kit (QIAGEN, Valencia, CA, USA), according to the manufacturer's instructions. Quality of genomic DNAs was checked on 1% agarose gel stained with ethidium bromide (EtBr) and DNAs were quantified using Picogreen dsDNA Quantitation Kit (Invitrogen, Carlsbad, CA, USA) in fluorimetry. The method allows the estimation of DNA concentration by comparison of the fluorescent signal obtain from each sample with that collected using a dilution of a DNA standard. Signal can be measured with a fluorescent microplate reader using excitation wavelength 484 nm, emission wavelength 538 nm, according to the protocol. The purified DNA was stored at -20°C.

2.1.2 Total RNA extraction and quantification

A small section of lung tumour tissue and normal lung parenchyma distant from the macroscopic lung cancer tissue was removed at surgery and stored frozen or in RNAlater solution (Ambion, Austin, TX, USA). Total RNA was extracted from normal lung or ADCA tissue with the RNeasy Midi kit (Qiagen) and quantified by Nanodrop Spettrophotometer ND-1000 (Thermo Scientific, Wilmington, DE, USA). The integrity of the total RNA obtained was evaluated with spectrophotometric analysis using the RNA 6000 Nano Assay Kit (Agilent Technologies, Palo Alto, CA, USA). The purified RNA was stored at -80°C.

2.1.3 Preparation of DNA and RNA pools

A genome-wide DNA pooling strategy was used in our case-control association studies as initial screening in order to minimize interindividual sample variability and to reduce costs and time as compared to the analysis of individual samples at the same power of study and with a robust estimation of allele frequency (214). Then, we confirmed putative associations by individual genotyping.

DNAs from patients and control series of population-based association study (Table 1) were respectively pooled to form 4 different pools (two from cases and two from controls), each constituted by 200 individuals contributing 30 ng of DNA to the pool. Cases and controls in pools were matched for gender, age and smoking habits to minimize phenotypic heterogeneity and population stratification (Table 6) (191). We have so applied a joint analysis of two experiments, each one including a pool of either lung ADCA cases or matched healthy controls.

DNA of 80 discordant sibs of family-based study (Table 2) was used to generate two pools (cases or controls) containing 30 ng of each DNA sample.

Patients of the discovery series of case-only association study, composed of 600 lung ADCA patients (Table 3), were divided into two groups according to their clinical stage (I or >I). As expected, Kaplan-Meier survival curves (Fig. 13) and Cox regression analysis of survival indicated poorer survival among patients with higher clinical stage compared with patients with stage I (P=6.47x10^{-6} and P=2.32x10^{-05} respectively). For each

sample 15 ng of DNA was used to create a DNA pool of 300 stage I patients and a DNA pool of 300 patients at higher clinical stages. Since the accuracy of analyses using a DNA pooling strategy depends heavily on the estimates of DNA concentration, we performed serial dilutions of each DNA sample (284). DNAs were first diluted to 15 ng/ul and their concentrations re-estimated by using Picogreen dsDNA Quantitation Kit (Invitrogen) in fluorimetry. Samples were then diluted to 5 ng/ul, re-quantified and finally 15 ng of each DNA were combined. Pools were quantified to check their correct concentration.

Table 6. Characteristics of Italian lung cancer patients and controls used for DNA pools.

	First experiment		Second experiment	
Subject characteristics	Cases	Controls	Cases	Controls
No. of subjects	200	200	200	200
Median age (years)	61.0	59.0	62.0	61.0
Gender				
Male	158	158	140	140
Female	42	42	60	60
Smoker status				
Never	29	29	0	0
Ever	171	171	200	200
Histology [a]				
ADCA	200	NA	200	NA
Lymph node status [b]				
N0	128	NA	101	NA
N1	69	NA	75	NA
Clinical stage				NA
I	107	NA	102	NA
II	39	NA	29	NA
III	36	NA	42	NA
IV	15	NA	25	NA
Follow-up (months) [c]	89.1 (n=61)	NA	87.3 (n=64)	NA

[a] ADCA, adenocarcinoma. [b] N0, absence of nodal metastasis, N1, presence of nodal metastasis. [c] Median for patients alive at the end of follow-up. NA, not applicable.

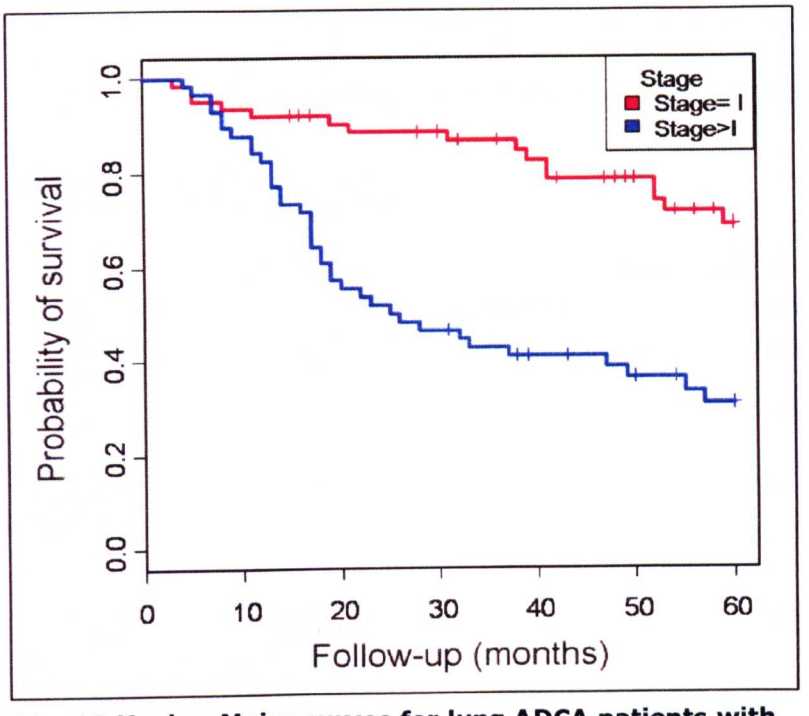

Fig. 13 Kaplan-Meier curves for lung ADCA patients with stage I (red line, number of patients = 60) or with higher clinical stage (blue line, number of patients = 60). Log-rank test showed a significant difference between the two curves (P=6.47 x 10^{-6}).

Whole-gene expression analysis for initial screening of transcriptome was performed on RNA pools obtained using equal amounts of each RNA sample. In the first experiment (A), the 120 RNA samples from normal lung (Table 4) were combined in 24 small pools: 12 pools constituted by patients with stage I and 12 pools by patients with higher clinical stage (5 samples per pool). These pools were analyzed on Sentrix Bead Chip HumanHT-12 (Illumina). In the second experiment (B), the same 120 samples were combined to form only two pool (60 samples per pool) representing stage I and stage > I patients, respectively. These pools were analyzed in quadruplicate on the Sentrix Bead Chip HumanRef_8_v2 (Illumina).

We used different pooling approach for the three genome-wide scanning and for the whole-gene expression analysis (Table 7). This was

due to the characteristics of each population and to the fact that each subsequent study reflects the experience acquired in the previous study.

Table 7. Summary of pooling approaches used in the thesis.

Study	Outcome	N° Experiment	Pool	samples/ pool	Replicas
Population-based case-control association study (GWAS)	Risk	*I*	2	200	2
		II	2	200	2
Family-based case-control association study (GWAS)	Risk	*I*	1	80	4
Case-only association study (GWAS)	Prognosis	*I*	1	300	12
Case-only association study (whole transcriptome analysis)	Prognosis	*I*	24	5	0
		II	2	60	4

2.2 POPULATION-BASED ASSOCIATION STUDY FOR LUNG CANCER RISK

2.2.1 Genome-wide SNPs analysis

In order to map genetic variation across human populations to identify variants associated with lung cancer risk or staging, we performed genome-wide association study using Illumina platform in collaboration with the CNIO Genotyping Unit in Madrid, where the Illumina platform is already available.

Genome-wide genotyping for initial screening was carried out in DNA pools (see paragraph 2.1.3) and 800 ng of DNA per pool was hybridized using the Infinium II Assay 300K on the Sentrix BeadChip platform

(Illumina, San Diego, CA, USA), that allows for the analysis of more than 318,000 tagSNPs chosen from the International HapMap Project (274). The Infinium II Whole-Genome Genotyping Assay used a single bead type and dual colour channel approach. The DNA samples were isothermally amplified in an overnight step using random primers and then fragmented by a controlled enzymatic process that does not require gel electrophoresis (Fig. 14, (275)).

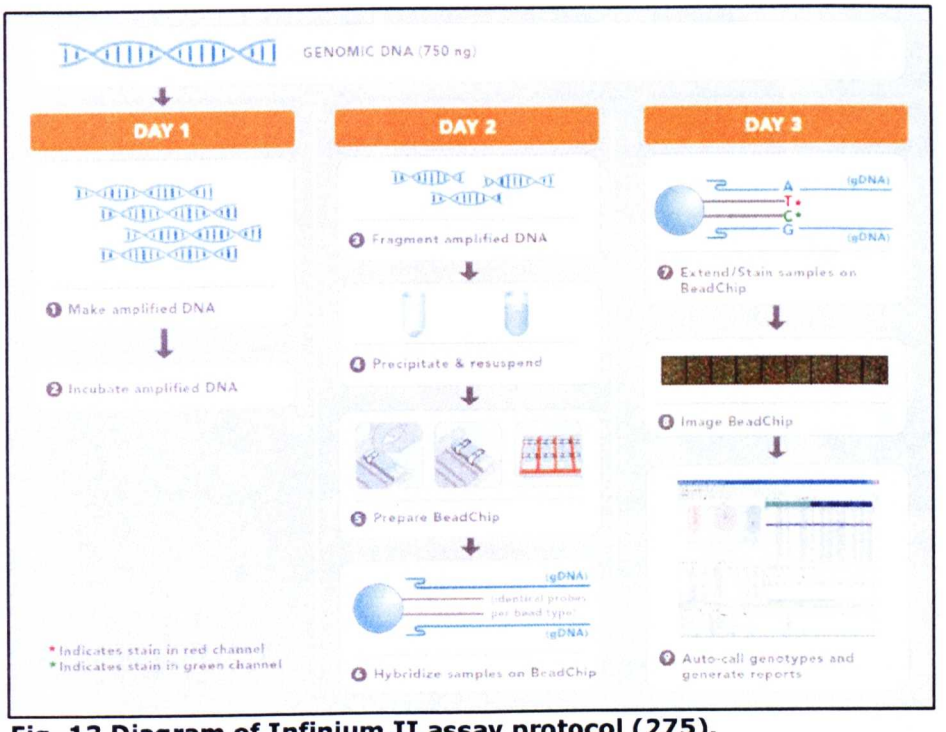

Fig. 13 Diagram of Infinium II assay protocol (275).

Briefly, after alcohol precipitation and resuspension, the amplified and fragmented DNA of 300-600 bp are hot denatured and the BeadChip is prepared for hybridization in the capillary flow-through chamber. Samples are applied to BeadChips and incubated overnight to permit the annealing of these to locus-specific 50-mers covalently linked to one of over 500,000 beadtypes. One bead type corresponds to each allele per SNP locus. After locus-specific hybridization capture, each SNP locus is "scored" by an

enzymatic single-base extension assay using labelled nucleotides that confers allelic specificity. These labelled products are subsequently visualized by fluorescent staining with a sandwich-based immunohistochemistry (IHC) that increases the overall sensitivity of the assay (Fig. 15).

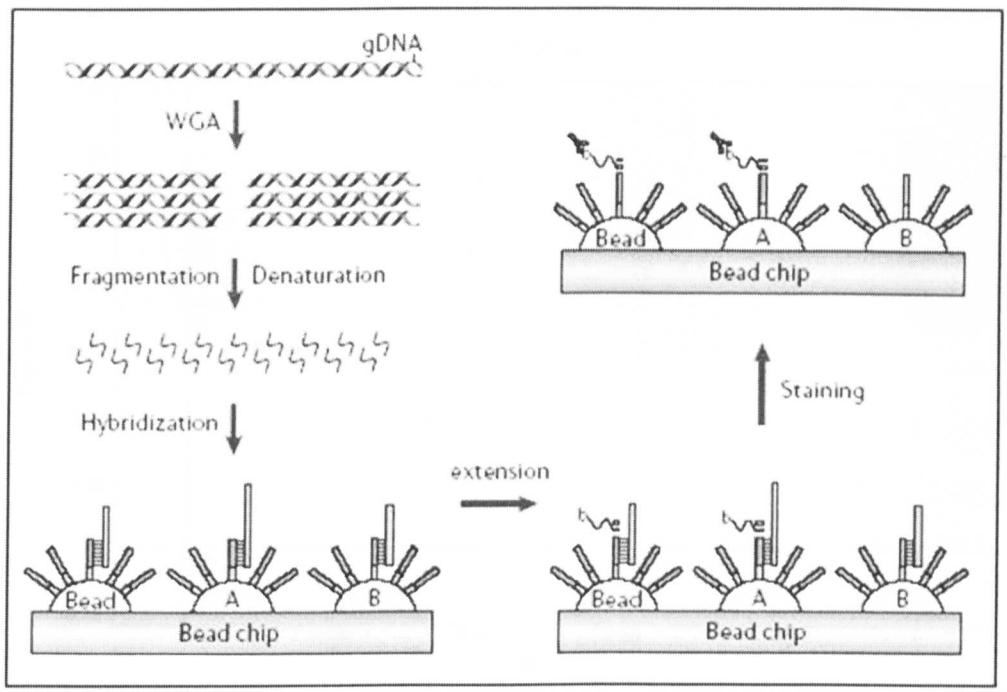

Fig. 14 Whole-Genome Genotyping steps (385).

The intensities of the beads' fluorescence are detected by the Illumina BeadArray Reader, and are in turn analyzed using Illumina's software for automated genotype clustering calling (275). The software represents samples in cluster of homozygous (red and blue points) and of heterozygous (violet points) according to the fluorescent signals in a diagram with normalized intensity in y and "theta" value in x axes. In case of pool genotyping, the diagram has intermediated values (grey points) (Fig. 16).

Data were obtained in the form of intensity files, which were used to determine the allele frequencies of each SNP and to reconstruct the number of chromosomes carrying each of the two possible alleles. For each DNA pool, SNP array analysis was carried out in duplicate to verify genotype reproducibility and estimate technical variability.

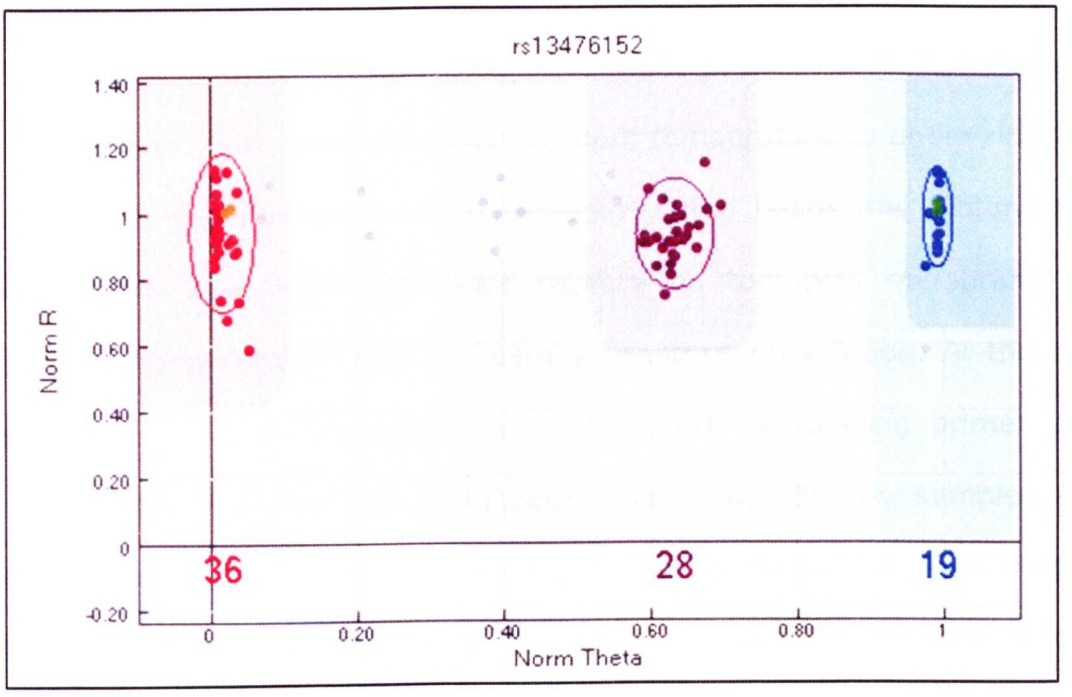

Fig. 15 Report of a SNP genotyping (275).

2.2.2 Independent confirmation on DNA pools

Independent confirmation of allele frequency estimation was carried out in the same DNA pools by PCR amplification of SNP-containing fragments, followed by pyrosequencing analysis on a PSQ96MA system (Biotage AB, Uppsala, Sweden), according to the manufacturer's instructions using specific primers reported in Supplementary Table 1 (at the end of this chapter).

Pyrosequencing technology is sequencing by synthesis, a simple to use technique for accurate and quantitative analysis of DNA sequences performed on PSQ96MA system. PCR assays were performed with primers (with one of them modified by addition of a biotinylated group at 5'-end) that amplified a short region contained the SNP. Then 20 µl of PCR products were mixed with 4 µl of streptavidin-coated beads (Biotage) and 36 µl of binding buffer and the volume is adjusted to 100 µl with water. Samples are then vortex at 1100 rpm for 10 min at room temperature to optimizing the formation of complex between streptavidin-coated beads and biotinylated PCR product. The complexes were capture on tool pins membrane by vacuum filtration and purified by using a denaturation solution. At the end the complexes were released in a solution with sequencing primer and samples were analysed with Pyrosequencing assay. Briefly, samples are denaturated at 80° C and sequencing primer is hybridized to a single-stranded PCR amplicon that serves as a template, and incubated with four enzymes (DNA polymerase, ATP sulfurylase, luciferase and apyrase) as well as the substrates adenosine 5' phosphosulfate (APS) and luciferin (Fig. 17, (276)). The first deoxribonucleotide triphosphate (dNTP) is added to the reaction. DNA polymerase catalyzes the incorporation of the dNTP into the DNA strand, if it is complementary to the base in the template strand. Each incorporation event is accompanied by release of pyrophosphate (PPi) in a quantity equimolar to the amount of incorporated nucleotide. ATP (adenosine triphosphate) sulfurylase converts PPi to ATP in the presence of adenosine 5' phosphosulfate and drives the luciferase-mediated conversion of luciferin to oxyluciferin that generates visible light in amounts that are

proportional to the amount of ATP (277). The light produced in the luciferase-catalyzed reaction is detected by a charge coupled device (CCD) chip and seen as a peak in the raw data output (Pyrogram). The height of each peak (light signal) is proportional to the number of nucleotides incorporated. Apyrase, a nucleotide-degrading enzyme, continuously degrades unincorporated nucleotides and ATP. When degradation is complete, another nucleotide is added. Addition of dNTPs is performed

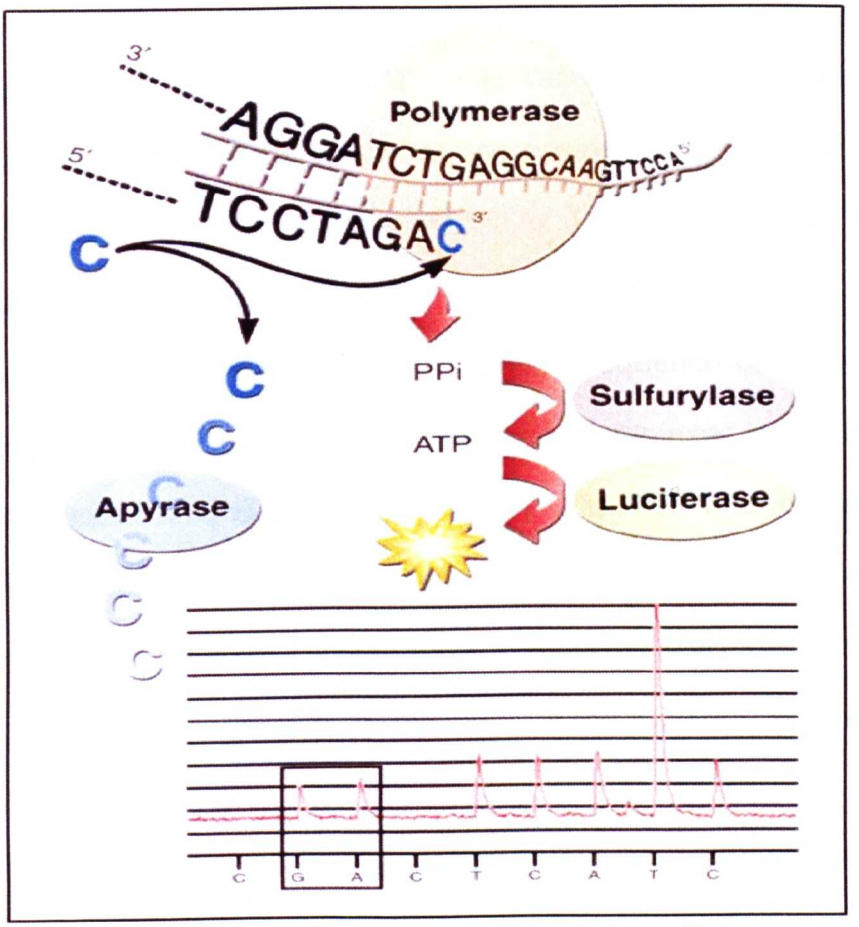

Fig. 17 Pyrosequencing: reactions and principles (276).

sequentially. As the process continues, the complementary DNA strand is built up and the nucleotide sequence is determined from the signal peaks in the Pyrogram trace (Fig. 17).

Primers designed to confirm allele frequencies of 47 putative associated SNPs in the same DNA pools for confirmation were listed in Supplementary Table 1.

2.2.3 Individual genotyping

Validation of statistically significant associated SNPs in individual samples was performed using MassARRAY Sequenom assay (Sequenom, San Diego, CA, USA).

Genotyping of the selected SNPs was carried out following published protocols applying the multiplex genotyping assay iPLEX™ for use with the MassARRAY platform (278). Briefly, multiplex PCR assays were designed using Sequenom SpectroDESIGNER software by entering sequence containing the SNP site and 100 bp of flanking sequence on either side of the SNP (Fig. 18). The SNPs were grouped into multiplexes according to the mass of the extension product over the SNP site. PCR was carried out in 384-well reaction plates in a volume of 5 µl using 2.5 ng of genomic DNA. All reactions are ended after a single base extension (SBE) into the SNP site and SBE products are separated by their mass differences allowing to genotype (Fig. 18, (278)).

The extension products were spotted onto a 384-well spectroCHIP bioarray before analysis by MALDI-TOF mass spectrometry (Sequenom). To guarantee quality of genotyping, all samples plus a series of duplicates were genotyped in the same batch.

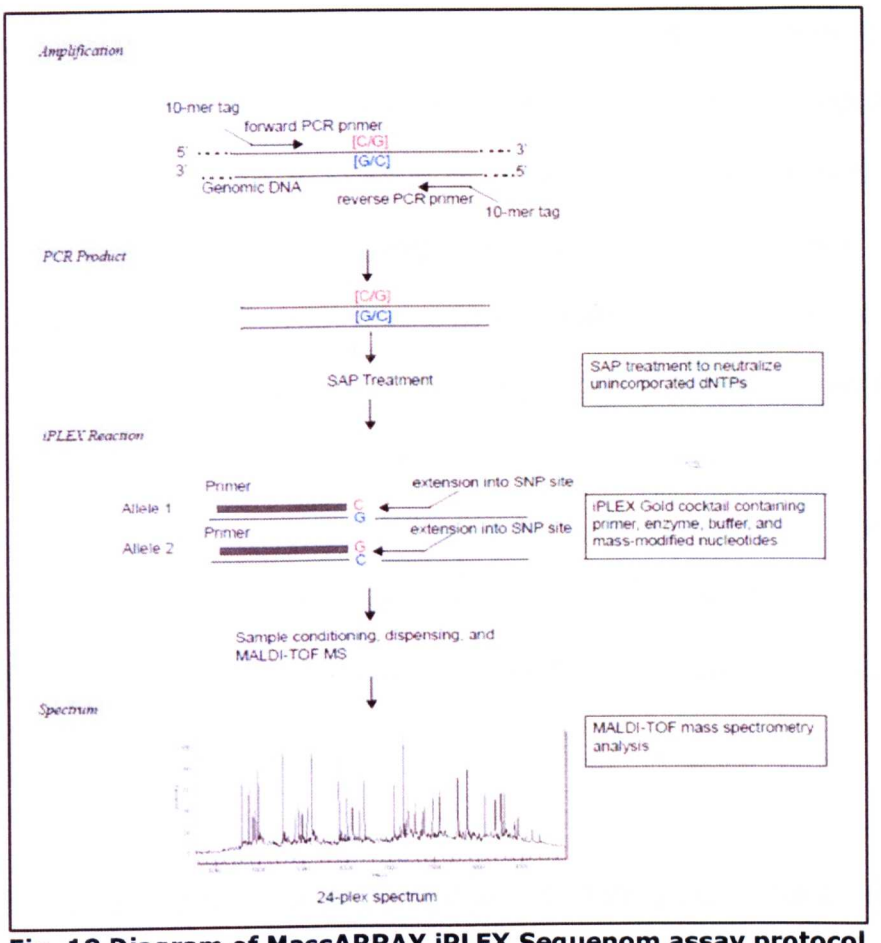

Fig. 18 Diagram of MassARRAY iPLEX Sequenom assay protocol (278).

2.2.4 Statistical analysis

The consistency of genotype frequencies at each SNP locus with respect to the Hardy–Weinberg equilibrium was tested (279). The correlation of the allelic frequencies within and among experiments, or between allele frequency data obtained by SNP array and pyrosequencing analyses, was tested by the Pearson's coefficient. Differences in allelic frequencies between case and control groups were analyzed by the Fisher's exact test or by chi-square analysis when the normal approximation was appropriate. Technical component was estimated as mean of the variance

between replicates within experiment, whereas the component due to sampling was obtained by comparison between mean allelic frequencies of the two experiments. The estimation of number of chromosomes in cases and controls was carried out using the 2x2 contingency table analysis. Association between each SNP and disease status, computing odds ratios and 95% confidence intervals, was assessed using the logistic regression or the extended Mantel-Haenszel chi-square for linear trend. LD of the SNPs with the surrounding region was assessed using HapMap data (CEU population). The Kaplan-Meier product-limit method and the log-rank test (280, 281) were used to evaluate the effect of the genotypes on overall survival of lung cancer patients.

2.3 FAMILY-BASED ASSOCIATION STUDY FOR LUNG CANCER RISK

2.3.1 Genome-wide SNPs analysis

Genome-wide genotyping was carried out in DNA pools prepared from cases and controls of the family-based series (see paragraph 2.1.3). 200 ng of DNA per sample was hybridized using the Infinium II Assay Human610-Quad BeadChip on the Sentrix BeadChip platform (Illumina, see paragraph 2.2.2), which allows analysis of 620,901 genetic markers chosen from the International HapMap release 23. For each DNA pool, SNP array analysis was carried out in quadruplicate.

2.3.2 Individual genotyping

Individual samples were genotyped using MassARRAY (Sequenom) (see paragraph 2.2.3).

2.3.3 Statistical analysis

Differences between lung cancer cases and their sib controls in allelic frequencies assessed in SNP array hybridization were analyzed using random variance t-statistics (282) and BRB ArrayTools developed by Dr. Richard Simon and Amy Peng Lam (http://linus.nci.nih.gov/BRBArrayTools.html). Differences in allele frequencies, estimated from SNP-array analysis of DNA pools, between cases and controls were tested by Fisher's exact test or by chi-square analysis when the normal approximation was appropriate. The correlation of the allelic frequencies between SNP array and individual genotypes was expressed as a Pearson's coefficient. Association analyses were carried out using PLINK software (223), which included analysis of HWE, family-based TDT (258) and population-based association analyses between disease status and genotype/allelotype. A generalized linear model with binomial errors was used to test the relationship between genetic susceptibility score and proportion of lung cancer cases; the mean values of genetic susceptibility scores were also analyzed using the Kruskal–Wallis test. The age was down-coded to binary dummy variables (age in decades), which were used as covariates in logistic analyses. Linkage disequilibrium between SNP markers was evaluated using JLIN program, version 1.6.0 (http://www.genepi.org.au/jlin.html) (283).

2.4 CASE-ONLY ASSOCIATION STUDY FOR LUNG CANCER PROGNOSIS

2.4.1 Genome-wide SNPs analysis

DNA pools obtained from 600 lung ADCA patients according to their clinical stage (see paragraph 2.1.3) were analyzed using the Human610-Quad BeadChip array (Illumina), as for family-based series (see paragraph 2.3.2). Twelve SNP-array hybridizations were performed for each DNA pool as described in paragraph 2.2.2.

2.4.2 Individual genotyping

Selected SNPs were genotyped in individual samples using MassARRAY Sequrnom assay (Sequenom) as described in paragraph 2.2.4.

2.4.3 Gene expression profile with microarray analysis

Microarray gene expression analysis was carried out in RNA pool from 120 lung ADCA patients (see paragraph 2.1.3). Each RNA pool was reverse-transcribed, labelled with biotin and amplified overnight (14 h) using the Illumina Total Prep RNA Amplification kit (Ambion) according to the manufacturer's protocol. A mixture of 1.5 μg of the biotinylated cRNA samples were hybridized according to manufacturer's protocol to Sentrix Bead Chip HumanHT-12 (experiment A) or to Sentrix Bead Chip HumanRef_8_v2 (experiment B) (Illumina). The arrays contain more than 48,000 or 22,000 bead types representing 47,231 or 18,196 unique sequences, respectively, derived from human genes in the NCBI Reference

Sequence (RefSeq) Release 38 or 22, respectively. Array chips were scanned with an Illumina BeadArray Reader. Intensity values of each hybridization were quality-checked and the data set was normalized using the cubic spline algorithm in the BeadStudio Version 3 software. A detection *P*-value <0.05 was set as a cut-off to filter the reliable genes, yielding a matrix containing 12,244 genes and 13,035 detectable transcripts, respectively. Data were analyzed using BRB ArrayTools (see paragraph 2.4.8). Microarray results were validated by quantitative real-time PCR (qRT-PCR) as described in paragraph 2.4.6.

2.4.4 Quantitative Real-Time PCR

From each sample, 1 µg of RNA was used to synthesize cDNA by reverse-transcription using Transcriptor First Strand cDNA Synthesis Kit (Roche, Basel, Switzerland) with a 1:1 mix of oligo(dT) and random hexamer primers, according to the manufacturer's instructions.

Real-time PCR analysis was performed using customized TaqMan® Low Density Arrays on the 7900HT System (Applied Biosystems, Foster City, CA, USA). TaqMan Gene Expression Assays used, spotted onto a 384-well card, are listed in Table 8 and in Table 9. Eight cDNA samples were analyzed per card. Each sample was measured in duplicate in a single RT-PCR run. 2.5 ng of cDNA template, mixed with TaqMan® Universal PCR Master Mix (Applied Biosystems), in a total volume of 100 µl, was loaded *per* sample loading port. Thermal cycling and fluorescence detection was performed on the microfluidic card sample block in the Applied Biosystems ABI Prism 7900HT Sequence Detection System (SDS) with ABI Prism

7900HT SDS Software 2.2 (Applied Biosystems). The thermal cycling conditions were 2 min at 50 °C and 10 min at 94.5 °C, followed by 40 cycles of 30 s at 97 °C and 1 min at 59.7 °C. Relative expression levels were calculated using the comparative Ct method calibrating the samples relative to a cDNA pool from normal lung tissue (calibrator). The raw gene expression values were normalized according to the expression of hypoxanthine phosphoribosyltransferase 1 (HPRT1, Hs99999909_m1) gene as endogenous reference (housekeeping) (Table 8 and Table 9).

The amount of a target gene in a sample, normalized to an endogenous reference and relative to a calibrator, is given by (285): $2^{-\Delta\Delta C_t}$ where C_t, or threshold cycle, is "the fractional cycle number at which the amount of amplified target reaches a fixed threshold" and $\Delta\Delta C_t = \Delta C_{t(\text{target gene in sample})} - \Delta C_{t(\text{target gene in calibrator})}$. The ΔC_t value is calculated as $C_{t(\text{target gene})} - C_{t(\text{housekeeping})}$ for each samples and for calibrator sample.

Table 8. Genes present on the TaqMan® LowDensity Array for Microarray validation.

Gene symbol	Gene name	TaqMan® Gene Expression Assay No.
Gcom1	GRINL1A complex locus	Hs00291311_m1
MSX1	Msh homeobox 1	Hs00427183_m1
TMEM100	Transmembrane protein 100	Hs00388033_m1
SMAD6	SMAD family member 6	Hs00178579_m1
IDH1	Isocitrate dehydrogenase 1 (NADP+), soluble	Hs00271858_m1
VIPR1	Vasoactive intestinal peptide receptor 1	Hs00270351_m1
SLC14A1	Solute carrier family 14 (urea transporter), member 1 (Kidd blood group)	Hs00210608_m1
BCL3	B-cell CLL/lymphoma 3	Hs00180403_m1
PLEKHO2	Pleckstrin homology domain containing, family O member 2	Hs00368811_m1
SFTPA2B	Surfactant protein A2B	Hs00359837_m1
SBNO2	Strawberry notch homolog 2 (Drosophila)	Hs00209130_m1
RRP12	Ribosomal RNA processing 12 homolog (S. cerevisiae)	Hs00958380_m1

DACT1	Dapper, antagonist of beta-catenin, homolog 1 (Xenopus laevis)	Hs00420410_m1
ITLN1	Intelectin 1 (galactofuranose binding)	Hs00914745_m1
C20orf114	Chromosome 20 open reading frame 114	Hs01113243_m1
SELE	Selectin E	Hs00950401_m1
FCN3	ficolin (collagen/fibrinogen domain containing) 3 (Hakata antigen)	Hs00892390_m1
COL1A1	Collagen, type I, alpha 1	Hs01076777_m1
DEFA3/DEFA1	defensin, alpha 3, neutrophil-specific/ defensin, alpha 1	Hs00414018_m1
TXNIP	thioredoxin interacting protein	Hs00197750_m1
LZTS1	leucine zipper, putative tumor suppressor 1	Hs00232762_m1
INHBB	Inhibin, beta B	Hs00173582_m1
HPRT1	Hypoxanthine phosphoribosyltransferase 1	Hs99999909_m1

Table 9. Genes present on the TaqMan® LowDensity Array for cytokine-cytokine receptor pathway validation.

Gene symbol	Gene name	TaqMan® Gene Expression Assay No.
CXCL2	chemokine (C-X-C motif) ligand 2	Hs00601975_m1
CCL2	chemokine (C-C motif) ligand 2	Hs00234140_m1
CXCL14	chemokine (C-X-C motif) ligand 14	Hs00171135_m1
TNFRSF12A	tumor necrosis factor receptor superfamily, member 12A	Hs00171993_m1
CCL3	chemokine (C-C motif) ligand 3	Hs00234142_m1
CSF3R	colony stimulating factor 3 receptor (granulocyte)	Hs01114427_m1
TNFSF10	tumor necrosis factor (ligand) superfamily, member 10	Hs00234356_m1
TGFB3	transforming growth factor, beta 3	Hs01086000_m1
CSF3	colony stimulating factor 3 (granulocyte)	Hs99999083_m1
IL7R	interleukin 7 receptor	Hs00233682_m1
IL1R1	interleukin 1 receptor, type I	Hs00991010_m1
IL8	interleukin 8	Hs00174103_m1
CCL21	chemokine (C-C motif) ligand 21	Hs99999110_m1
CX3CR1	chemokine (C-X3-C motif) receptor 1	Hs00365842_m1
IL6	interleukin 6	Hs00174131_m1
IL1RL1	interleukin 1 receptor-like 1	Hs01073300_m1
CXCR7	chemokine (C-X-C motif) receptor 7	Hs00604567_m1
ICAM1	intercellular adhesion molecule 1	Hs00164932_m1
ICAM4	intercellular adhesion molecule 4	Hs00169941_m1
CXCL1	chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha)	Hs00236937_m1
CCL4L1	chemokine (C-C motif) ligand 4-like 1	Hs00237011_m1
CXCL13	chemokine (C-X-C motif) ligand 13	Hs00757930_m1
HPRT1	Hypoxanthine phosphoribosyltransferase 1	Hs99999909_m1

2.4.5 Immunohistochemical Analysis

Immunohistochemical staining was performed on paraffin-embedded tissue sections of lung ADCA and surrounding normal lung tissue retrieved from the archives of our Department of Pathology. Antibodies used were anti-SLC14A1 (AV48116; diluted 1:1500) and anti-SMAD6 (AV100717; diluted 1:250) from Sigma-Aldrich™ (Sigma-Aldrich™, St. Louis, MO, USA), and anti-FCN3 (sc-55202; diluted 1:50) from Santa Cruz Biotecnology, Inc (Santa Cruz, CA, USA). Immunoreactive signals were detected with Chem-Mate DAB (Dako, Glostrup, Denmark).

2.4.6 Statistic analysis

Differences between stage I and stage>I lung ADCA cases in allelic frequencies assessed in SNP-array hybridization were analyzed using random variance t-statistics (282) and BRB Array Tools (http://linus.nci.nih.gov/BRB-ArrayTools.html). Differences in chromosome counts between the two groups were tested by Fisher's exact test or by chi-square analysis when the normal approximation was appropriate. The correlation between SNP-array and individual genotype allelic frequencies was expressed as a Spearman's coefficient. Association between clinical stage (I or >I) and confounding variables was analyzed using ANOVA (analysis of variance) or logistic analysis, whereas association between SNPs and clinical stage was analyzed using PLINK software (223), which included analysis of HWE, LD between SNPs, and population-based association between prognosis factors and genotype/allelotype. Age at cancer diagnosis was down-coded to binary dummy variables (age in

decades), which were used as covariates in logistic regression analyses. The average genetic risk score of clinical stage >1 for individuals was calculated using the "score" procedure of PLINK, i.e., the sum, across the 22 statistically significantly (P <0.01) associated SNPs in the joint analysis, of the number of minor alleles (0,1 or 2) at any SNP multiplied by the log of the odds ratio for that SNP. The reliability of the model was assessed by bootstrap re-sampling with replacement (286). Overall survival was assessed using Cox regression analysis and the "survival" package in R, with follow-up cut at 60 months to reduce bias due to mortality caused by non-cancer-related factors. All statistical tests were 2-sided.

Analyses of gene expression data were performed using BRB Array Tools version 3.8.1 (http://linus.nci.nih.gov/BRB-ArrayTools.html). Pathway analyses were carried out using the DAVID (Database for Annotation, Visualization and Integrated Discovery) Functional Annotation Tool (DAVID Bioinformatics Resources 6.7, NIAID/NIH, http://david.abcc.ncifcrf.gov/ (287)) and the Ingenuity Pathway Analysis tool (IPA, Ingenuity System, https://analysis.ingenuity.com). Differences in mRNA levels determined in qRT-PCR were assessed by ANOVA on relative quantification (RQ) data. Correlation between microarray and qRT-PCR results was assessed using Pearson's correlation coefficient, r. Kruskal-Wallis test was carried out using R packages.

SUPPLEMENTARY TABLES

Supplementary Table 1. Primers used for PCR amplification and genotyping of 47 SNPs

SNP	Gene	Forward PCR primer[1] (seq 5'->3')	Reverse PCR primer[1] (seq 5'->3')	Pyrosequencing primer (seq 5'->3')
rs10518668		B-cacttgccttaatcagatggtca	tctcccctccttaaataaatgatg	aaagcaattcaatttctct
rs11119493	HHAT	gtgctctgtattcaaaagccattt	B-aaacatcccaaaatatgggtgaga	atttgatatttggttaaatt
rs12556578		tgtggtggcaaatattctttgtt	B-gctggtgtcacttgctgaataat	tccctttttctatttgg
rs12680976		gggatgcagctagagcaatactta	B-aactgtgagtcctccaactttgt	taatacaaataaaatcagaa
rs1385049		cctgaatcaaacttgcgtaatg	B-atgtatcagttcttagccactgga	aattccctatgaaaatatta
rs1433184		gagcaaatgtgggatgattcaaag	B-acaccatcagcctgttgttttta	tgcatttttgtttacaaatt
rs1584586		tgtctggacccaagcttttc	B-gaaacggttcatcactaccagtc	tcttagttcagttttggtaa
rs17199134		B-tctttttgactcctcatttcttta	gataaacttctctccttttttgtaa	aatatgattcaaagttaata
rs2038256		atatttttgggtaggcaacagact	B-accctctagctttgctgtttgt	ggagttgaggaatttcag
rs3797832		B-ttttcaccaggcttaagacattg	tcacaataaatgggaaggatga	aattaaaagaaaacgaatt
rs3804479		B-agccaaacaaaggtacagtatgc	tcataaaagcaagcgagatttcca	agcgagatttccaaaa
rs4897493		ctcttttgctttcacacacaagtt	B-gatcacaggatacaaagcacacat	atggcatttagaagaaaa
rs1877116		B-tctccttctgccaaattctgct	gggctgttgacattgagctact	gttatctcaagaatacctg
rs2588767		B-gcctaatattagtttggcactgaa	gtattgggaagggattcttcaa	cagagtcttaaaaaaacgc
rs3130517		tttttgaagactagccatgacact	B-gcctctctctgtgggctattaa	tccttttaagacttgatgtt
rs2418422	C9orf27	B-gaggaagaggaaatcaggtagaaa	ttctccactatcctcttgcatcta	ctcttgcatctagtgctt
rs6488007		B-tacggagtttcactttgggatgat	tgctagtccacctgaattctctta	acttaaccatatttttgaga
rs16918924		B-ggccatctgcttcaccaa	tagagtgggcagcctgaaga	tgaagaaaattgtaaatgtt
rs8027776	SEC11A	B-cactgcgtccagcctaaaaac	cgttctgtcccacctctaatg	atcatagagtatgtattctt
rs132470		tttcccaaaagaccctaaatagct	B-ccatcctggtgtcataaaacatct	cactgtgtgttagcagaa
rs4823406	PHF21B	caagcatgcagggactagaat	B-accccacacacccacatt	gaaattcagtgaagaaaaac
rs6654096	GPM6B	B-caaagctgccattgcctttta	tggaaagattggggttgaagg	tcattattgccctgag
rs5945306	ZNF275	aggtctggttgccagtttaggtg	B-aaccacatgcctcgctcttt	ctaccaaaacttgaagg
rs2172706		B-gatccagagcttgctgaagtga	ataccatgtgagggaagagtaacc	tcctactctcttctcttcc
rs1470037		B-atgtcagattttcccctacaacaa	cactagaagggcaataggcaaga	ggcaataggcaagatg
rs1584586		tgtctggacccaagcttttc	B-gaaacggttcatcactaccagtc	tcttagttcagttttggtaa
rs1428053		gcagaaaaaagagcaaaataaacc	B-tggatctgtagtgatagcccattt	aatttgtttctttagaatgc
rs1033822		catgcttattcattcaggaacatc	B-agctagccagtattgtgacattga	aaaatacactctgtatgagc
rs1520	KIF6	B-gcttccataaactacccaaggata	ggctcccttgcattttataaga	attcatgaaatataattcac
rs210798	MYB	B-ctctgtccatggatttgactacaa	tggcaccagctgtttgtaaa	tttaataaaaagcacagga
rs4731775		gtcccccaaatggtttttattgt	B-tttctttctcttgcctgactgct	caatattgtgaaagcat
rs3812278	CNOT4	gcctcagttaaccaaatctaagga	B-tgtcatgttggagttggagtttt	ttgggataaagtcatttaat
rs7011544		taaaacatagtggggacctccaa	B-tgtcattgcacatatcacactcag	gggacctccaagaatct
rs302917		B-aaaggattggaggagaaatattca	atggtgctgtttttgaaaacttg	tggaaataaatttgacaga
rs7907321	CTNNA3	ggggaattgttgtatttgaaagaa	B-caatcatgcttcaaattgcttaga	tgacgtgtatattttgttat
rs2515373	CNTN5	B-tcttagggggcacgtggt	cacagagggtggaatgttga	gttgaaattctcctacctc
rs7670329		tctagatgagcaccaataatgcc	B-gaatgttgctcatttgaaacagtt	gcaccaataatgcctt
rs8062660		B-ccaaactcatgggtgtctgaa	agagggataggaaatgtgtcagg	gagtaccagctgcca
rs2869832		B-tctgttgcctttcccttctg	tggcccaggtttctcctgtt	gagaaaatcacgactatcaa
rs2139875	LOC388458	B-gaacttggtaacactgccaacaat	cctcagcatcaggcaatatacat	tttaatttccagattttcac
rs8113515		gttttaccacattattggcagttg	B-cacaaatggctgtctcatctgta	aaggcaatttccttt
rs967785	GK	B-ccgtagagatcctgtgtgaagtag	gcattttgcagggagtgg	gggagtggtttagcaa
rs1199508	DGAT2L6	tagcaataaacaagctgattcaaa	B-gccacatatggacatcaaattttc	ccatctgattttgagactta
rs5949639		B-tccctctgagctttcaagatcact	tggggcattaaatgaattaaggt	agacaaattactgttattgg

rs1933800	gaatgttcccaacacaaagaaat	B-atgatattatgattcgtggagtgc	acaccataatagaaccca
rs5943261	acagctttttgcattgtgattacc	B-tctctctcaatcccctatcattct	ttttgcattgtgattacc
rs6918015	acacaggaagcagtggctagatga	B-aggcttcaagtgcttctgaatcaa	agaggcaaaggccac

[1]For each SNP one PCR primer is modified by addition of a biotinylated group at the 5' extremity (B).

3. RESULTS

3.1 RESULTS OF POPULATION-BASED ASSOCIATION STUDY FOR LUNG CANCER RISK

3.1.1 Multiple unlinked SNPs are associated with a decreased lung ADCA risk

SNP-array hybridization design was carried out using a DNA pooling strategy in a joint analysis of 2 experiments, each conducted in duplicate in pools of either 200 lung ADCA cases or 200 matched healthy controls extracted from our series (Table 6). Correlation of the allelic frequencies between replicates within each experiment was 0.991 in both experiments, whereas correlation between experiments was 0.984.

Statistical analyses of allelic frequencies of the two independent experiments pointed to 235 SNPs statistically correlated ($P < 0.01$) within and between experiments and significant allelic imbalance between case and control DNA pools. From these SNPs putative associated with lung cancer risk, we selected a subset of 47 SNPs to limit costs of the experiment (Table 10), giving priority to SNPs with the highest statistical association ($P < 1 \times 10^{-7}$ in at least 1 replica), reduced variation of allele frequency between the two experiments (coefficient of variation, i.e., standard deviation/mean allele frequency, ≤5%), frequency of the rare allele ≥0.20, and close vicinity (<1 Mb) to another SNP of the 235 SNP list.

Table 10. 47 SNPs putative associated with lung cancer risk in GWA analysis of the two experiments.

SNP name [a]	Chromosome	Position (Mb) [b]	Gene	*P* value [c]	
				First experiment	Second experiment
rs10518668	1	83,1		$1,31 \times 10^{-13}$	$3,33 \times 10^{-16}$
rs11119493	1	210,6	HHAT	$1,59 \times 10^{-12}$	$6,16 \times 10^{-12}$
rs2172706	1	154,7	KCNN3 *	$4,03 \times 10^{-6}$	$8,80 \times 10^{-3}$
rs1470037	2	81,9		$6,59 \times 10^{-3}$	$2,74 \times 10^{-6}$
rs1584586	3	150,2	TSC22D2 *	$3,54 \times 10^{-11}$	$9,00 \times 10^{-3}$
rs17199134	4	172,3		$7,77 \times 10^{-16}$	$4,08 \times 10^{-9}$
rs1877116	4	66,0		$1,46 \times 10^{-3}$	$6,59 \times 10^{-3}$
rs2588767	4	66,9		$1,49 \times 10^{-3}$	$2,62 \times 10^{-6}$
rs7670329	4	180,3		$1,94 \times 10^{-6}$	$6,78 \times 10^{-3}$
rs3797832	5	108,4	FER	$1,52 \times 10^{-10}$	$6,16 \times 10^{-11}$
rs1428053	5	116,7		$9,00 \times 10^{-3}$	$3,78 \times 10^{-3}$
rs3804479	6	6,6	LY86	$1,11 \times 10^{-16}$	$9,00 \times 10^{-3}$
rs4897493	6	131,4	EPB41L2 *	$6,63 \times 10^{-10}$	$3,36 \times 10^{-10}$
rs3130517	6	31,3		$1,81 \times 10^{-6}$	$7,15 \times 10^{-6}$
rs1033822	6	18,8		$8,00 \times 10^{-8}$	$2,75 \times 10^{-4}$
rs1520	6	39,5	KIF6	$4,10 \times 10^{-7}$	$5,00 \times 10^{-8}$
rs210798	6	135,5	MYB	$9,00 \times 10^{-3}$	$1,00 \times 10^{-8}$
rs6918015	6	151,5		$1,34 \times 10^{-4}$	$5,50 \times 10^{-5}$
rs4731775	7	130,9	MKLN1	$1,82 \times 10^{-3}$	$5,62 \times 10^{-4}$
rs3812278	7	135,1	CNOT4	$1,14 \times 10^{-6}$	$6,90 \times 10^{-4}$
rs12680976	8	140,0		$2,54 \times 10^{-9}$	$4,70 \times 10^{-14}$
rs1385049	8	51,8	SNTG1 *	$1,59 \times 10^{-12}$	$9,00 \times 10^{-3}$
rs1433184	8	108,5	ANGPT1	$2,54 \times 10^{-11}$	$2,41 \times 10^{-13}$
rs7011544	8	111,1		$9,27 \times 10^{-3}$	$4,06 \times 10^{-4}$
rs2418422	9	118,7	C9orf27	$4,10 \times 10^{-3}$	$2,35 \times 10^{-3}$
rs302917	9	135,6	GTF3C4, DDX31 *	$2,00 \times 10^{-8}$	$4,46 \times 10^{-5}$
rs7907321	10	68,0	CTNNA3	$4,77 \times 10^{-3}$	$2,00 \times 10^{-8}$
rs2515373	11	99,5	CNTN5	$1,00 \times 10^{-8}$	$7,14 \times 10^{-5}$
rs6488007	12	31,9		$1,21 \times 10^{-3}$	$2,16 \times 10^{-3}$
rs16918924	12	31,9		$4,15 \times 10^{-3}$	$1,09 \times 10^{-3}$
rs2038256	14	29,2	C14orf23, FOXG1B *	$2,35 \times 10^{-11}$	$2,83 \times 10^{-12}$
rs8027776	15	85,3	SEC11A	$9,05 \times 10^{-5}$	$8,99 \times 10^{-6}$
rs8062660	16	59,2		$3,73 \times 10^{-3}$	$1,18 \times 10^{-5}$
rs2869832	17	63,3		$2,04 \times 10^{-5}$	$4,92 \times 10^{-3}$
rs2139875	18	4,2		$9,00 \times 10^{-3}$	$2,00 \times 10^{-8}$
rs8113515	19	43,8	PSG9 *	$6,98 \times 10^{-6}$	$1,20 \times 10^{-3}$
rs132470	22	45,2	ARHGAP8, PRR5	$3,71 \times 10^{-3}$	$3,62 \times 10^{-4}$
rs4823406	22	45,3	PHF21B	$7,30 \times 10^{-4}$	$9,87 \times 10^{-3}$
rs12556578	X	62,7		$7,08 \times 10^{-10}$	$1,39 \times 10^{-8}$
rs6654096	X	14,0	GPM6B	$1,41 \times 10^{-6}$	$4,14 \times 10^{-4}$
rs5945306	X	152,7		$4,07 \times 10^{-4}$	$3,19 \times 10^{-4}$
rs967785	X	30,7	GK	$3,83 \times 10^{-4}$	$4,17 \times 10^{-3}$
rs1199508	X	69,4	DGAT2L6	$1,00 \times 10^{-7}$	$8,77 \times 10^{-4}$
rs5949639	X	95,0		$9,00 \times 10^{-3}$	$3,12 \times 10^{-3}$
rs1933800	X	97,7		$1,90 \times 10^{-5}$	$6,00 \times 10^{-8}$

rs5943261	X	108,4	$6,15 \times 10^{-3}$	$2,66 \times 10^{-4}$

[a] SNPs sorted by chromosome and position; * gene in LD with relative SNP (HapMap3 Genome Browser release #2); [b] Position in megabases according to Ensembl release 59; [c] *P* values obtained by the Fisher's exact test or by chi-square analysis when the normal approximation was appropriate.

The independent confirmation of allele frequency obtained from 300K SNP assay was carried out using pyrosequencing analysis. Primers designed (Supplementary Table 1) to analyze the 47 SNPs in the same DNA pools led to the selection of 16 SNPs, based on strength of association (-log *P* > 1.5, Table 11) and concordance between the two experiments, and analyzed in the individual samples by MassARRAY Sequenom Assay (Sequenom).

Table 11. 16 SNP validated in the pyrosequencing analysis on DNA pools.

SNP [a]	Chromosome	Mb [b]	Gene	*P* Fisher	-log*P*
rs2172706	1	154,7	KCNN3 *	3.30×10^{-3}	2,48
rs1470037	2	81,9		$8,85 \times 10^{-5}$	4,05
rs1877116	4	66,0		1.20×10^{-3}	2,92
rs3130517 §	6	31,3		$2,45 \times 10^{-4}$	3,61
rs4897493	6	131,4	EPB41L2 *	1.09×10^{-2}	1,96
rs6918015	6	151,5		1.39×10^{-2}	1,86
rs4731775 †	7	130,9	MKLN1	$6,15 \times 10^{-4}$	3,21
rs2515373	11	99,5	CNTN5	$8,11 \times 10^{-5}$	4,09
rs16918924	12	31,9		$2,69 \times 10^{-4}$	3,57
rs6488007	12	31,9		1.76×10^{-2}	1,75
rs8062660	16	59,2		1.80×10^{-2}	1,74
rs2139875 °	18	4,2		1.31×10^{-2}	1,88
rs8113515 °	19	43,8	PSG9 *	$4,31 \times 10^{-4}$	3,37
rs5945306	X	152,7		$2,54 \times 10^{-6}$	5,60
rs967785	X	30,7	GK	2.62×10^{-2}	1,58
rs5943261	X	108,4		3.16×10^{-2}	1,50

[a] SNPs sorted by chromosome and position; § failed pre-extend (Sequenom); ° failed in homogenous mass extension (hME) design (Sequenom); † failed genotyping; * gene mapping in the LD region with relative SNP (HapMap3 Genome Browser release #2); [b] Position in megabases according to Ensembl release 59.

Among these 16 SNPs, 1 SNP failed in the pre-extend analysis, 2 SNP failed in homogenous mass extension (hME) design, and 1 SNP failed in the genotyping. Frequency of the rare alleles of the 12 remaining SNPs in controls ranged from 0.07 to 0.36. None of these SNPs showed significant deviations from the Hardy–Weinberg equilibrium. Paired analysis for

possible LD between the 12 SNPs detected a significant LD between rs6488007 and rs16918924, mapping at a 13.6 kb distance on chromosomes 12, with a preferential segregation of the rare alleles of the two SNPs in the same individuals (P<0.0001), in both controls (D' = 0.88) and cases (D' = 0.79). Although these SNPs were linked, their genotypes matched only partially in the series of controls and cases, indicating that each of the two SNPs contained a distinct set of genetic information. Thus, they were maintained in the study and analyzed separately. The SNPs selection steps are summarized in Fig. 19.

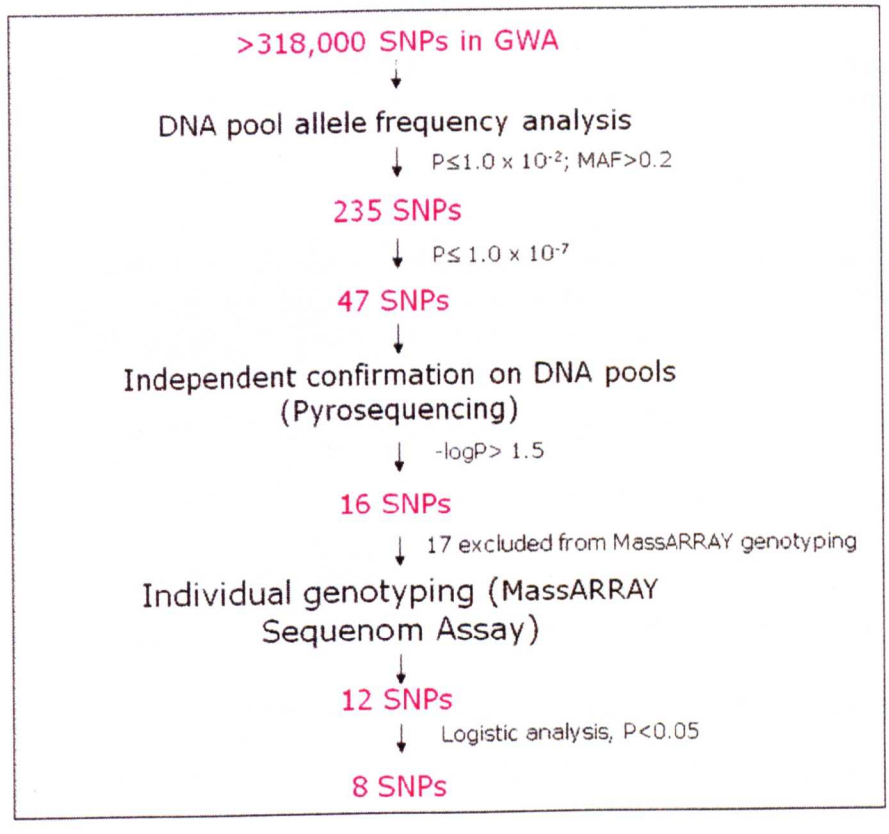

Fig. 19 Schematic representation of SNPs selection in population-based case-control association study.

Significant allelic associations with lung ADCA risk were observed for 8 SNPs (P≤0.05); 7 autosomal and one on chromosomes X (Table 12). The SNP on chromosomes X (rs5945306) showed significant allelic association in both sexes. None of the 8 SNPs showed statistical associations with the confounding variables of gender, smoking habit and age, except for SNPs mapping on chromosomes X, for which logistic analyses were adjusted by sex.

Table 12. List of 8 SNPs showing significant association with lung ADCA risk after MassARRAY Sequenom assay on individual samples.

SNP [a]	Chromosome	Position (Mb) [b]	Gene
rs2172706	1	154,7	KCNN3 *
rs1470037	2	81,9	
rs1877116	4	66,0	
rs4897493	6	131,4	EPB41L2 *
rs2515373	11	99,5	CNTN5
rs6488007	12	31,9	
rs16918924	12	31,9	
rs5945306	X	152,7	

[a] SNPs sorted by chromosome and position; [b] Position in megabases according to Ensembl release 59; * gene mapping in the LD region with relative SNP (HapMap3 Genome Browser release #2).

We assessed the lung cancer risk by genotype or allele status testing a genetic model based on dominant or codominant effects of the rare allele on this risk. A significant association between the rare allele carrier status and decreased risk of lung ADCA was found for the 8 SNPs (OR ~0.6-0.8, $P < 0.05$; Fig. 20), except for the SNP rs2515373 that shows a borderline association (OR = 0.73; 95% CI: 0.52–1.02; Fig. 20). No significant association was observed between rare allele carrier status and survival rate of lung ADCA patients at any SNPs.

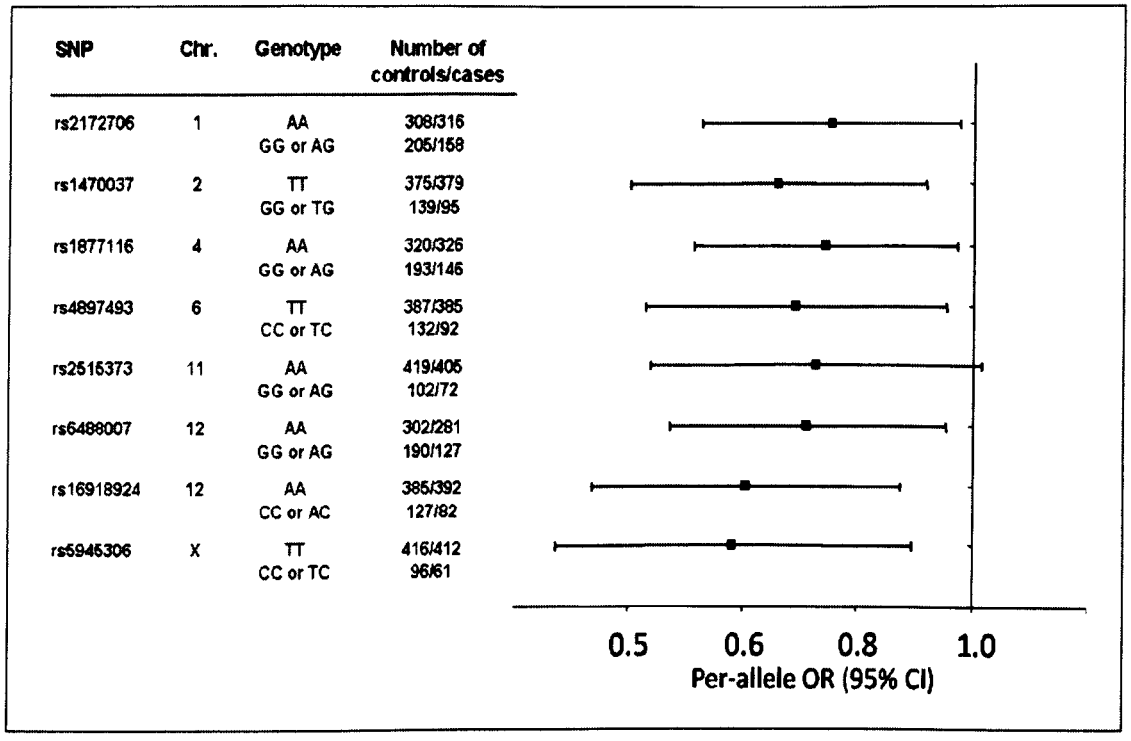

Fig. 20 Plot of the risk of lung cancer associated with the rare allele carrier status at each of 8 SNPs identified by genome-wide scan, in a series constituted by Italian cases and controls. Shaded squares denote odd ratios (ORs). Horizontal lines represent 95% CIs. The vertical line indicates the null effect (OR=1.0).

3.1.2 The confirmed SNPs point to a polygenic model with additive and interchangeable effects

We tested the hypothesis of a polygenic model under the assumption of additive and independent effects of rare alleles of the 8 SNPs that showed associations with lung ADCA risk. Gender-adjusted allele dosage–response analysis (i.e., number of alleles versus risk of lung ADCA) evidenced that risk of lung cancer significantly decreased according to the number of rare alleles carried (Fig. 21, $P = 5.3 \times 10^{-9}$). In particular, carriers of 2 to 6 rare alleles showed a statistically significant decreased risk of lung cancer, with a decreasing trend up to an OR=0.29; (95% CI 0.13–0.67) for individuals carrying 5-6 rare alleles (Fig. 21). Overall, carriers of 2 or more rare alleles (n=344/229, controls/cases) versus carrier of 0 or 1 rare allele

(n=137/177, controls/cases) showed about 2-fold lower risk of lung ADCA (OR=0.52, 95% CI 0.39 – 0.68, $P = 2.8x10^{-6}$).

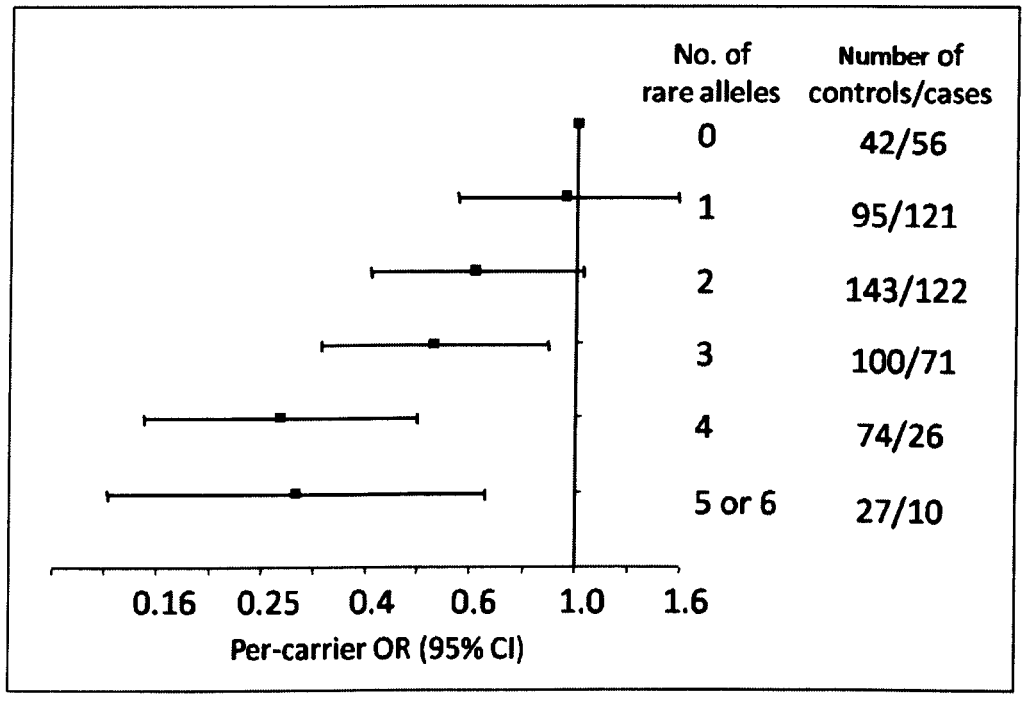

Fig. 21 Plot of the risk of lung cancer associated with the carrier status of each rare allele of the 8 SNPs reported in Fig. 21.

3.2 RESULTS OF FAMILY-BASED ASSOCIATION STUDY FOR LUNG CANCER RISK

3.2.1 Multiple unlinked SNPs are associated with lung ADCA risk

Genome-wide analysis of allelic frequencies of each SNP from case and sib-control DNA pools, deleting SNPs whose minor allele frequency in both cases and controls was <0.1, revealed 659 SNPs with parametric *P*-values $\leq 1.0 \times 10^{-7}$ (equivalent at a false discovery rate $P = 0.0008$). For these 659 SNPs, we estimated the number of chromosomes in cases and controls and obtained 82 SNPs putatively associated with disease at $P \leq 0.001$. All of them were assayed for mass spectrometry analysis on

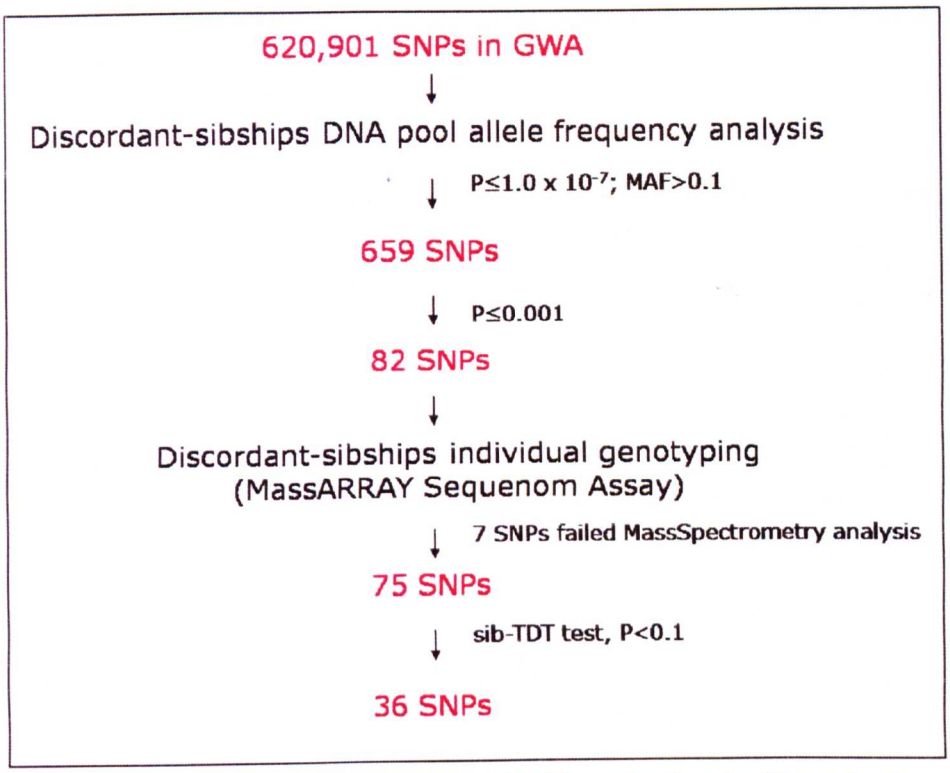

Fig. 22 Schematic representation of SNPs selection in discordant sib-pairs study.

MassARRAY Sequenom assay (Sequenom): three SNPs failed PCR or MassEXTEND primer design and, therefore, 79 SNPs were genotyped by MassARRAY in 80 cases and their respective healthy sib controls. After MassARRAY genotyping, two SNPs failed genotyping, one SNP was monomorphic and one SNP showed highly significant deviation from the Hardy–Weinberg equilibrium (most likely because of bias in genotype calls due to preferential allele amplification or to technical problems in their assays), and were therefore removed from the study, reducing the number of markers to 75 SNPs. The SNPs selection steps are summarized in Fig. 22.

Correlation analysis of the minor allele frequencies estimated in cases and controls either in DNA pools by SNP array analysis or in individual samples by MassARRAY for the 75 SNPs associated with lung cancer

demonstrated the reliability of the pooling approach ($r = 0.78$, $P < 2.2 \times 10^{-16}$, Fig 23).

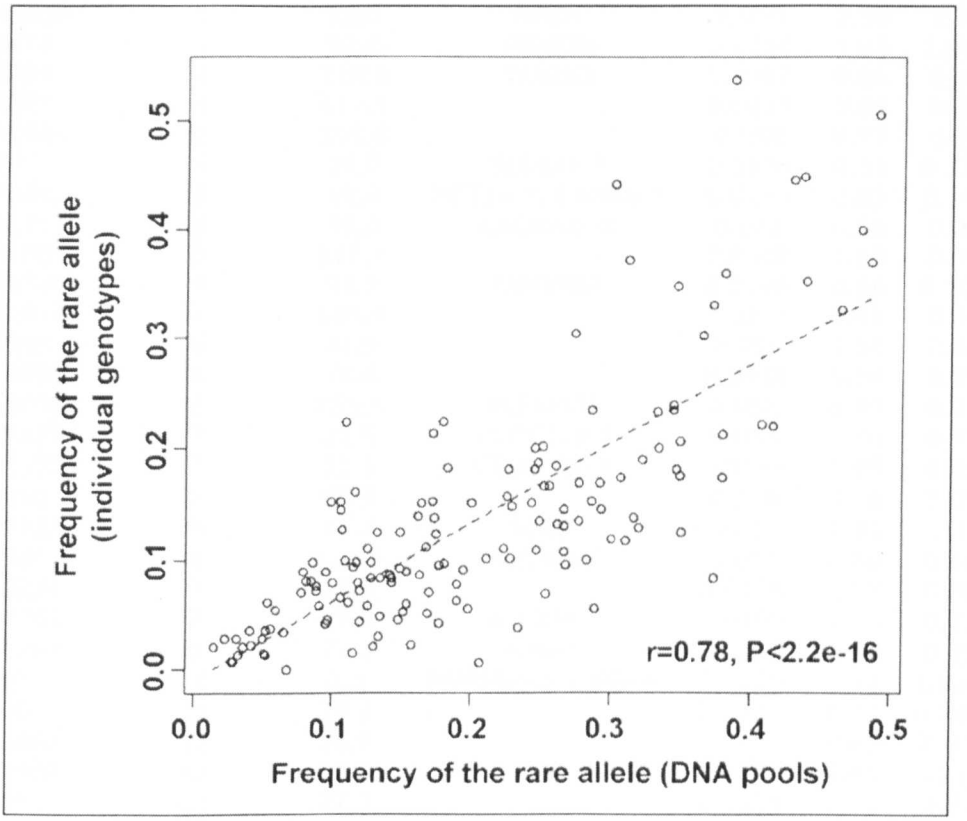

Fig. 23 Correlation between SNP frequencies measured by SNP array analysis of DNA pools and frequencies measured by genotyping of individual samples. Plotted data represent frequencies of the rare allele of 75 SNPs putatively associated with lung cancer risk.

Single-point analysis using the sib-TDT test (258) indicated that 36 of the 75 genotyped SNPs were significantly associated with disease status (Table 13). The strongest associations were observed for SNPs rs11833102 mapping in the carboxypeptidase M (CPM) gene on chromosomes 12, rs17120323 in the sarcoglycan zeta (SGCZ) gene on chromosome 8, rs12445758 within cadherin 13, H-cadherin (heart) (CDH13) gene on chromosome 16, and rs325702 mapping in the cyclic nucleotide-gated channel alpha 4 (CNGA4) gene on chromosome 11.

Table 13. 36 SNPs showing statistically significant association with lung cancer risk in the discordant sibs series.

SNP [a]	Chromosome	Position (Mb) [§]	Gene	*P*-value [b]	OR [c]	95% CI
rs639739	1	4,4		0.0124	0.5	0.24 - 1.06
rs12748434	1	52,3	NRD1	0.058	2.30	0.78 - 6.81
rs1261411	1	57,0	PPAP2B	0.0114	1.65	1.01 - 2.69 [d]
rs2765529	1	119,6	WARS2	0.0482	0.66	0.39 - 1.14
rs6676647	1	119,8		0.0233	0.64	0.37 - 1.09
rs10931664	2	195,8		0.058	0.57	0.24 - 1.36
rs721377	3	14,4	SLC6A6 *	0.0126	0.33	0.12 - 0.95 [d]
rs9813644	3	49,9	MST1R *, CAMKV *	0.0254	3.63	0.74 - 17.80
rs1918071	3	55,0	CACNA2D3	0.072	0.55	0.26 - 1.14
rs1456196	3	117,7		0.0348	1.68	0.78 - 3.60
rs12648320	4	92,3	FAM190A	0.0196	0.54	0.30 - 0.97 [d]
rs28475332	4	188,4		0.0201	0.41	0.16 - 1.02
rs7713580	5	41,9		0.096	1.56	0.68 - 3.60
rs16889292	6	78,4		0.0339	0.14	0.02 - 1.12
rs12663498	6	151,0	PLEKHG1	0.055	0.49	0.27 - 0.91
rs17160175 [†]	7	31,5	CCDC129 *	0.033	1.86	0.92 - 3.76
rs11773530	7	31,5	CCDC129 *	0.0164	1.95	0.97 - 3.93
rs4330610	7	85,3		0.059	0.36	0.09 - 1.39
rs17120323	8	14,7	SGCZ	0.0011 [e]	1.94	1.11 - 3.38 [d]
rs3019885	8	118,0	SLC30A8	0.052	0.69	0.42 - 1.13
rs12342234	9	13,3		0.0126	1.83	0.88 - 3.81
rs16937762	9	19,8	SLC24A2	0.0196	0.35	0.11 - 1.11
rs12001157	9	72,1	APBA1	0.090	1.63	0.78 - 3.42
rs325702	11	6,3	FAM160A2, CNGA4	0.0045	2.41	1.06 - 5.49 [d]
rs820900	11	38,2		0.0114	0.31	0.10 - 0.99 [d]
rs10842402	12	24,9		0.061	0.63	0.33 - 1.19
rs11833102	12	69,3	CPM	0.0006 [e]	2.44	1.21 - 4.94 [d]
rs9544359	13	77,3		0.0254	1.91	0.93 - 3.94
rs1958226	14	82,2		0.052	0.55	0.24 - 1.26
rs11074274	15	95,0	MCTP2	0.0348	2.54	0.87 - 7.43
rs12445758	16	83,3	CDH13	0.0016 [e]	1.93	1.16 - 3.22 [d]
rs790097	17	71,6	SDK2	0.096	0.48	0.16 - 1.45
rs4426464	19	1,8	ONECUT3	0.0067	2.57	1.17 - 5.65 [d]
rs755032	20	24,0		0.0067	2.04	0.91 - 4.57
rs2516542	22	21,4	TOP3B	0.0076	3.1	1.18 - 8.11 [d]
rs4823153	22	44,3		0.0046	0.51	0.27 - 0.97

[a] SNPs sorted by chromosome and position; [b] DFAM procedure in PLINK toolset, nominal *P*-values. SNPs sorted by chromosome and position. [c] Based on allelic test for association. [d] $P<0.05$, logistic regression procedure in PLINK toolset, based on allelic test for association, i.e., rare allele versus common allele. [e] $P<0.05$ by 20,000 permutations of the whole series (75 SNPs). [†] SNP rs17160175 excluded from the polygenic model due to its high linkage disequilibrium with rs11773530 ($D'= 1.0$, $r^2 = 0.97$). [§] Position in megabases according to Ensembl release 59; * gene mapping in the LD region with relative SNP (HapMap3 Genome Browser release #2).

3.2.2 Four SNPs were confirmed in population series

The 36 statistically associated SNPs with lung cancer in the discordant sibs analysis (Table 13) were replicated in a population-based lung ADCA

case-control series (Table 1). None of them showed significant deviation from the Hardy-Weinberg equilibrium, except for rs11074274 (P = 0.005, in cases only). Unadjusted logistic analysis indicated that 4 SNPs (rs12748434, rs1261411, rs4330610, and rs3019885) were statistically significant associated (P<0.05, Table 14). When adjusted for sex, age and smoking habit only the last three SNPs were significantly associated in the population-based series (P < 0.05, Table 14).

Table 14. SNPs showing statistically significant association with lung adenocarcinoma risk in the population-based series.

SNP [a]	Chromosome	Position (Mb) [§]	Gene	OR [b]	95% CI	P-value [c]
rs12748434	1	52,3	NRD1	0.70	0.51 - 0.96	0.026
rs1261411 [†]	1	57,0	PPAP2B	1.23	1.03 - 1.48	0.024
rs4330610 [†]	7	85,3		0.57	0.35 - 0.94	0.025
rs3019885 [†]	8	118,0	SLC30A8	1.21	1.01 - 1.45	0.036

[a] SNPs sorted by chromosome and position; [b] Based on allelic test for association. [c] Logistic regression procedure in PLINK toolset; nominal P-values; [†] Statistically associated SNPs in adjusted analysis for sex, age and smoking habit; [§] Position in megabases according to Ensembl release 59.

Comparison of the cases with early tumour onset (age up to 60 years; n=198) versus the whole controls confirmed the association of the SNP rs3019885 on chromosome 8 (P = 0.029) and detected the association of the SNP rs16937762 on chromosome 9 (P = 0.024).

3.2.3 The polygenic model explains lung cancer risk in discordant sibships

In the family-based series we tested the previous proposed polygenic model for the interpretation of individual risk of lung cancer.

Analysis included a total of 35 SNPs (Table 13) and 151 subjects. SNP rs17160175 was excluded for its high linkage disequilibrium with rs11773530 (D'= 1.0, r^2 = 0.97). Five controls and four cases were removed from the dataset because >80% genotypes were missing.

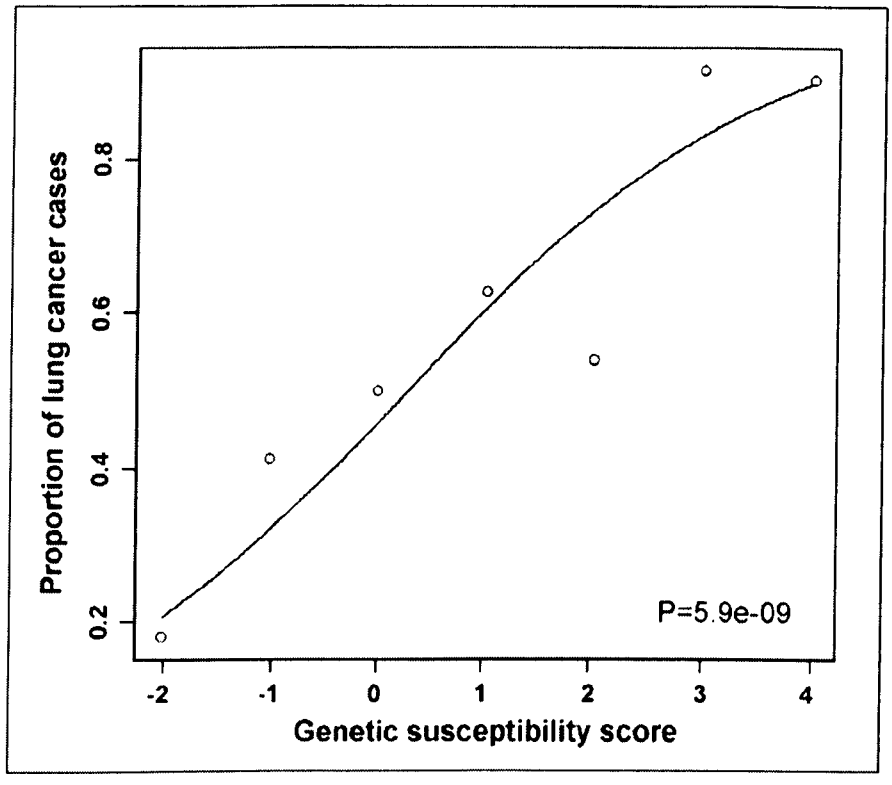

Fig. 24 A polygenic inheritance model with additive and interchangeable effects of rare alleles at lung cancer modifier loci explains the individual risk of lung cancer in the family-based series. Scatterplot shows the proportions of subjects that are cases as a function of genetic susceptibility score and the fitted line.

To test the model, a score of +1 or -1 was attributed to the rare allele of each SNP based on its association with increased or decreased lung cancer risk, respectively. For each subject, the sum of the scores for all 35 SNPs was obtained as a general estimator of the individual genetic risk. The average estimator was -1.6 ± 0.3 (mean ± standard error) in controls and 2.0 ± 0.3 in cases, respectively (P = 2.0 x 10^{-11}, Kruskal–Wallis test). The

proportion of lung cancer cases increased at higher genetic susceptibility scores (Fig. 24; $P = 5.9 \times 10^{-9}$).

Then, we tested the same model in the replication study that has been carried out in the population-based series, by calculating for each individual the genetic susceptibility score as we have done in the family-based series. Applying the same polygenic model in the population-based series, by using the three confirmed SNPs (Table 14), we found that the average estimator was 1.25 ± 0.03 in controls and 1.40 ± 0.03 in cases, respectively ($P = 0.0019$, Kruskal–Wallis test). Analysis limited only to non-smoker cases (n=66) versus all controls (n=503) or versus non-smoker controls (n=25) gave similar results.

3.3 RESULTS OF GENOME-WIDE SNPs ANALYSIS IN CASE-ONLY ASSOCIATION STUDY FOR LUNG CANCER PROGNOSIS

3.3.1 Multiple unlinked SNPs are associated with lung ADCA prognosis

Analysis to test for possible associations between clinical stage and confounding variables, such as gender, age at diagnosis and smoking habit, revealed a relatively weak statistical association between clinical stage and smoking status, with a borderline significant decrease of ever-smokers in clinical stage >I (OR=0.61, 95% CI 0.39 - 0.95, P=0.030, logistic analysis). No statistically significant associations were observed between clinical stage and either age at diagnosis or gender.

Genome-wide SNP array analysis conducted in 12 replicas of DNA pools from lung cancer cases at clinical stage I or at higher clinical stages (Table 3), respectively, allowed the screening of 620,901 SNPs. Analysis of allelic frequencies of each SNP from DNA pools, deleting SNPs whose minor allele frequency was >0.10 in the pools, revealed 10,571 SNPs at parametric *P*-values $\leq 1.0 \times 10^{-5}$. For these 10,571 SNPs, using a 2x2 contingency table analysis to reconstructed the number of chromosomes in the two groups we identified 80 most statistically associated SNPs with clinical stage at $P \leq 1.0 \times 10^{-4}$. The SNPs selection steps are summarized in Fig. 25.

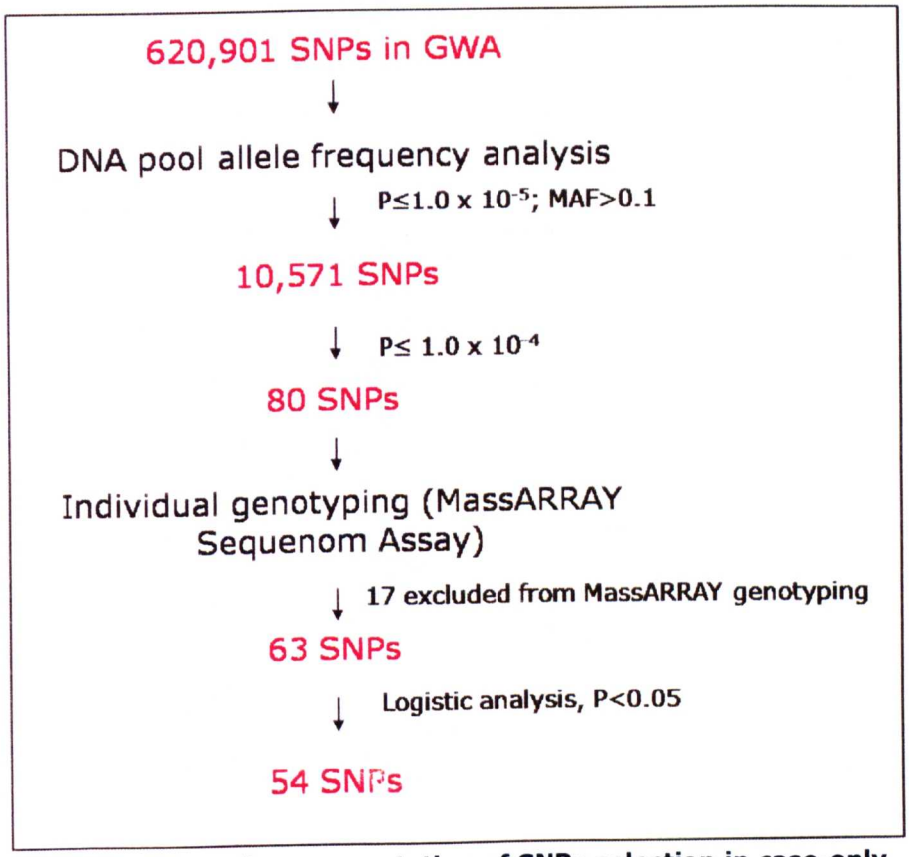

Fig. 25 Schematic representation of SNPs selection in case-only GWAS.

Table 15. 63 SNPs putative associated with lung cancer staging.

SNP [a]	Chromosome	Position (Mb) [b]	Gene	*P* value [c]	OR [d]	95% IC [d]	*P* value [d]
rs951774	2	102,9	IL1RL1*, IL1RL2*	1.48×10^{-5}	0.67	0.48 to 0.95	2.6×10^{-2}
rs10187901	2	140,6		7.26×10^{-5}	0.60	0.39 to 0.93	2.8×10^{-2}
rs13390491	2	179,6	TTN	9.33×10^{-5}	1.65	1.20 to 2.27	1.8×10^{-3}
rs10498217	2	228,0	COL4A4	1.23×10^{-5}	0.56	0.37 to 0.84	6.9×10^{-3}
rs16843438	2	242,1	TMEM16G	8.48×10^{-6}	0.56	0.39 to 0.80	1.6×10^{-3}
rs2574711	3	11,7	VGLL4	5.28×10^{-5}	0.55	0.36 to 0.85	3.5×10^{-3}
rs7694589	4	30,3		4.45×10^{-5}	0.68	0.52 to 0.90	9.4×10^{-3}
rs11722134	4	73,6		2.79×10^{-5}	1.68	1.05 to 2.69	2.5×10^{-2}
rs1994854	4	77,8		7.59×10^{-7}	0.55	0.40 to 0.77	2.1×10^{-4}
rs423997	4	86,3		2.88×10^{-6}	1.34	1.05 to 1.72	1.2×10^{-2}
rs4505911	5	68,2		5.01×10^{-5}	0.45	0.26 to 0.78	4.1×10^{-3}
rs10900886	5	105,2		2.40×10^{-5}	0.54	0.38 to 0.78	1.0×10^{-4}
rs13189604	5	107,2	FBXL17	3.54×10^{-6}			7.3×10^{-1}
rs3823111	6	53,5	KLHL31	4.95×10^{-6}	2.33	1.26 to 4.33	5.0×10^{-3}
rs806435	6	88,8	SPACA1	6.59×10^{-6}	1.96	1.34 to 2.86	4.8×10^{-4}
rs458523	6	95,1		1.72×10^{-5}	1.54	1.20 to 1.97	5.8×10^{-4}
rs565968	6	125,4	IBRDC1	1.24×10^{-6}	0.78	0.62 to 0.98	4.3×10^{-2}
rs10278557	7	15,7	MEOX2	3.42×10^{-5}	0.50	0.37 to 0.67	5.0×10^{-6}
rs13438238	7	54,3		4.14×10^{-10}			3.3×10^{-1}
rs2877213	7	54,9		4.02×10^{-5}			6.7×10^{-1}
rs845559	7	55,2	EGFR	8.98×10^{-6}			4.9×10^{-1}
rs17819684	7	82,7	PCLO	9.86×10^{-5}	1.45	1.12 to 1.87	2.4×10^{-3}
rs2299297	7	104,7	MLL5	6.56×10^{-5}	1.64	1.24 to 2.16	4.2×10^{-4}
rs2648	7	128,8	TSPAN33	7.57×10^{-6}	2.29	1.43 to 3.68	8.0×10^{-4}
rs17125699	8	17,7		3.87×10^{-5}	0.52	0.28 to 0.98	4.4×10^{-2}
rs10738132	8	98,2		4.54×10^{-7}			3.6×10^{-1}
rs972519	9	4,5	SLC1A1	7.14×10^{-6}	0.60	0.41 to 0.89	7.3×10^{-3}
rs824249	9	28,8		1.78×10^{-5}	0.55	0.37 to 0.82	3.1×10^{-3}
rs10491726	9	114,3	LTB4DH, ZNF483	3.87×10^{-5}			1.2×10^{-1}
rs10987191	9	129,0		2.28×10^{-7}	0.47	0.29 to 0.77	2.7×10^{-3}
rs11259181	10	14,6	FAM107B	6.87×10^{-5}	0.60	0.38 to 0.95	2.2×10^{-2}
rs2797902	10	31,3	ZNF468 *	5.05×10^{-5}			1.4×10^{-1}
rs10832757	11	17,3	NUCB2	3.33×10^{-5}	0.54	0.40 to 0.72	4.6×10^{-5}
rs7107350	11	21,2	NELL1	5.52×10^{-6}	2.01	1.34 to 3.00	7.2×10^{-4}
rs3808996	11	125,0	SLC37A2	6.72×10^{-5}	0.60	0.40 to 0.89	1.3×10^{-2}
rs3825305	12	63,0	PPM1H	6.13×10^{-5}	0.50	0.31 to 0.79	3.3×10^{-3}
rs9596742	13	53,6		1.01×10^{-5}	0.43	0.28 to 0.67	1.3×10^{-4}
rs2391875	13	111,5		3.03×10^{-5}	0.53	0.37 to 0.75	2.0×10^{-4}
rs8020076	14	28,5		6.34×10^{-6}	1.66	1.28 to 2.14	8.2×10^{-5}
rs718998	14	37,4	SLC25A21	9.70×10^{-5}	1.45	1.07 to 1.96	2.0×10^{-2}
rs1255641	14	64,0	PPP2R5E	6.62×10^{-5}	1.75	1.24 to 2.47	1.1×10^{-3}
rs10520058	15	38,6	SPRED1	7.82×10^{-7}	0.35	0.20 to 0.64	5.4×10^{-4}
rs2937940	15	86,4		7.86×10^{-5}	1.88	1.40 to 2.54	1.9×10^{-5}
rs9927531	16	26,5		7.19×10^{-5}	1.76	1.23 to 2.51	1.6×10^{-3}
rs1183259	16	60,4		5.28×10^{-5}	0.51	0.33 to 0.77	1.8×10^{-3}
rs4788587	16	72,0	PKD1L3	2.22×10^{-5}	0.66	0.48 to 0.91	1.2×10^{-2}
rs10514440	16	78,7	WWOX	1.48×10^{-5}	3.38	1.67 to 6.82	5.4×10^{-4}
rs1860444	17	48,9		3.57×10^{-6}	3.10	1.71 to 5.63	4.3×10^{-4}
rs16950191	17	49,7	CA10	4.69×10^{-5}	1.38	1.03 to 1.84	3.2×10^{-2}
rs12610723	19	3,8	MATK	1.13×10^{-5}	2.41	1.36 to 4.26	2.8×10^{-3}
rs2287700	19	14,6	PKN1	8.46×10^{-5}	1.99	1.21 to 3.26	7.9×10^{-3}
rs4805442	19	30,1		6.30×10^{-5}	0.48	0.32 to 0.72	4.3×10^{-4}
rs6030680	20	41,8	PTPRT	2.28×10^{-6}	1.61	1.21 to 2.13	1.1×10^{-3}
rs4553110	X	6,5		7.85×10^{-5}	0.62	0.44 to 0.88	5.2×10^{-3}
rs12687904	X	6,8		4.26×10^{-5}	0.51	0.28 to 0.90	1.4×10^{-2}
rs4830793	X	12,7	FRMPD4	9.56×10^{-5}	2.19	1.25 to 3.81	4.0×10^{-3}

rs7887846	X	22,7		5.58×10^{-5}	0.62	0.43 to 0.89	1.3×10^{-2}
rs5972356	X	31,3	DMD	3.89×10^{-7}	0.56	0.35 to 0.87	7.4×10^{-3}
rs5927730	X	31,4	DMD	4.19×10^{-5}	0.66	0.47 to 0.92	1.8×10^{-2}
rs5906595	X	48,1		6.65×10^{-5}			5.8×10^{-2}
rs5969041	X	86,2		7.37×10^{-5}			7.2×10^{-2}
rs404481	X	102,5	TCEAL8	8.31×10^{-5}	0.66	0.48 to 0.89	7.8×10^{-3}
rs2207031	X	128,0		3.30×10^{-6}	2.24	1.45 to 3.47	1.9×10^{-4}

[a] SNPs sorted by chromosome and position; * gene in LD with relative SNP (HapMap3 Genome Browser release #2); [b] Position in megabases according to Ensembl release 59; [c] *P* values obtained by chi-square analysis from GWA; [d] OR, odds-ratio obtained by logistic regression procedure of PLINK toolset, based on allelic test for association, i.e., rare allele versus common allele, adjusted by age at tumour diagnosis in decades, and smoking status; CI, confidence interval; [§] SNPs showing statistically significant ($P<0.05$) association with clinical stage in individual genotyping.

To validate the SNP-array findings in DNA pools, the 80 SNPs were selected for genotyping in the individual cases by MassARRAY. Of the 80 SNPs, 2 mitochondrial SNPs, 1 SNP on chromosome Y, and 9 redundant SNPs in tight LD with closely SNPs (<58 kb distance) in the same locus in the HapMap Caucasian (CEU) population were excluded. One SNP failed PCR or MassEXTEND primer design and 4 additional SNPs failed genotyping, reducing the number of markers to 63 SNPs (Table 15).

A good correlation of the minor allele frequencies obtained by MassARRAY genotyping in single individuals or by SNP-array analysis in DNA pools was observed ($r=0.79$, $P<2.2 \times 10^{-16}$), demonstrating the reliability of the DNA pooling approach (Fig. 26). None of the selected SNPs showed significant deviation from the HWE, except for rs565968 ($P=0.00076$). No statistically significant LD was observed between any SNP pairs ($r^2<0.1$).

Association analysis using a logistic model adjusted for smoking status indicated that 54 of 63 SNPs were significantly associated with clinical stage status (Table 15, $P<0.05$). The strongest association was observed for SNP rs10278557 ($P=5.0 \times 10^{-6}$), which maps in the mesenchyme homeobox 2 (MEOX2) gene on chromosome 7.

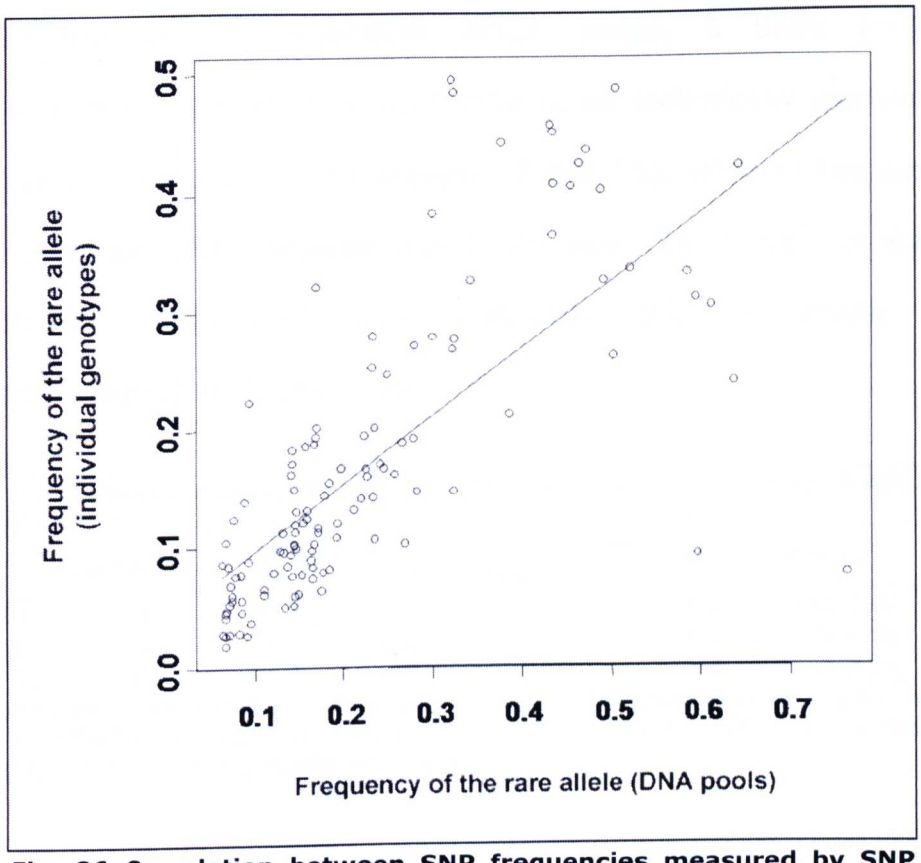

Fig. 26 Correlation between SNP frequencies measured by SNP array analysis of DNA pools and frequencies measured by genotyping of individual samples. Plotted data represent frequencies of the rare allele of 63 SNPs putatively associated with lung cancer risk.

The 63 SNPs were then tested in two independent series of 317 lung ADCA and 257 lung SQCC patients (Table 3). The ADCA series of patients showed similar phenotypic characteristics as compared to the discovery series, whereas the SQCC series had an older age at diagnosis and a higher frequency of males and of ever-smokers as compared to the ADCA series. A statistically significant association was found between clinical stage and age at tumour diagnosis in the SQCC series (P=0.005, Kruskal-Wallis test), but not in the ADCA series; neither gender nor smoking status was associated with clinical stage in either of the two series. No statistically significant deviation (P<0.01) from the HWE was observed in the two series for any of the 63 SNPs.

In the small independent ADCA series, 3 SNPs (rs7694589, rs3823111, rs16950191) were confirmed to be statistically associated with clinical stage ($P<0.05$, logistic analysis, Table 16), while in the lung SQCC series, 3 other SNPs showed significant association with clinical stage (rs10491726, rs10520058, and rs4805442; $P<0.05$, logistic analysis adjusted for age at diagnosis, Table 17).

Table 16. SNPs showing statistically significant association with clinical stage in ADCA independent series.

SNP [a]	Chromosome	Position (Mb) [b]	Gene	Rare/Common allele	OR [c]	95% CI [c]	*P* value [c]
rs7694589	4	30,3		T/C	2.1	1.38 to 3.06	4.1×10^{-4}
rs3823111	6	53,5	KLHL31	T/C	2.7	1.13 to 6.38	2.6×10^{-2}
rs16950191	17	49,7	CA10	A/C	1.6	1.06 to 2.29	2.4×10^{-2}

[a] SNPs sorted by chromosome and position; [b] Position in megabases according to Ensembl release 59; [c] Unadjusted logistic regression procedure in PLINK toolset; *P* values obtained by allelic test for association; CI, confidence interval.

Table 17. SNPs showing statistically significant association with clinical stage in SQCC independent series.

SNP [a]	Chromosome	Position (Mb) [b]	Gene	Rare/ Common allele	OR [c]	95% CI [c]	*P* value [c]
rs10491726	9	114,3	LTB4DH, ZNF483	T/C	0.5325	0,29-0,98	0.04266
rs10520058	15	38,6	SPRED1	A/C	0.3582	0,14-0,91	0.03014
rs4805442	19	30,1		A/G	0.475	0,27-0,83	0.009019

[a] SNPs sorted by chromosome and position; [b] Position in megabases according to Ensembl release 59; [c] Logistic regression procedure in PLINK toolset, adjusted for age at diagnosis; *P* values obtained by allelic test for association; CI, confidence interval.

To test for possible heterogeneity between the ADCA series and to increase the statistical power of association analyses (288), we carried out a joint analysis of the GWA and ADCA replication series bringing the total sample size of 917 lung ADCA patients (Table 3). Logistic analysis in the whole series adjusted for age at diagnosis and smoking status revealed 22 SNPs showing statistically significant association with clinical stage at statistical threshold of $P<0.01$ (Table 18). The strongest association

remained for the SNP rs10278557 (P=1.1 x 10^{-5}) mapping in the MEOX2 gene.

Table 18. 22 SNPs associated with lung ADCA clinical stage in the joint analysis of the GWA and replication ADCA series and used to build up the polygenic model with additive effects of SNP rare alleles on risk of clinical stage >1.

SNP [a]	Chromosome	Position (Mb) [b]	Gene	Rare Allele	OR [c]	95% CI [c]	P-value [c]
rs951774	2	102,9	IL1RL1*, IL1RL2*	A	0.7	0.5 to 0.9	6.9 x 10^{-3}
rs13390491	2	179,6	TTN	T	1.4	1.1 to 1.8	6.3 x 10^{-3}
rs10498217	2	228.0	COL4A4	T	0.7	0.5 to 0.9	9.4 x 10^{-3}
rs1994854	4	77,8		A	0.7	0.5 to 0.9	2.8 x 10^{-3}
rs4505911	5	68,2		A	0.5	0.3 to 0.8	4.0 x 10^{-3}
rs10900886	5	105,2		A	0.7	0.5 to 0.9	7.7 x 10^{-3}
rs3823111	6	53,5	KLHL31	T	2.6	1.6 to 4.3	2.1 x 10^{-4}
rs806435	6	88,8	SPACA1	T	1.8	1.3 to 2.4	2.5 x 10^{-4}
rs10278557	7	15,7	MEOX2	A	0.6	0.4 to 0.7	1.1 x 10^{-5}
rs2299297	7	104,7	MLL5	T	1.6	1.3 to 2.0	7.5 x 10^{-5}
rs824249	9	28,8		T	0.6	0.5 to 0.9	9.5 x 10^{-3}
rs10987191	9	129		A	0.5	0.4 to 0.8	2.5 x 10^{-3}
rs10832757	11	17,3	NUCB2	A	0.6	0.5 to 0.8	5.8 x 10^{-5}
rs9596742	13	53,6		A	0.6	0.5 to 0.9	7.7 x 10^{-3}
rs2391875	13	111,5		A	0.6	0.4 to 0.8	8.8 x 10^{-5}
rs8020076	14	28,5		C	1.3	1.1 to 1.6	9.8 x 10^{-3}
rs10520058	15	38,6	SPRED1	A	0.5	0.3 to 0.8	5.8 x 10^{-3}
rs9927531	16	26,5		A	1.7	1.3 to 2.3	2.5 x 10^{-4}
rs10514440	16	78,7	WWOX	T	2.4	1.4 to 4.1	1.1 x 10^{-3}
rs16950191	17	49,7	CA10	A	1.5	1.2 to 1.8	1.6 x 10^{-3}
rs7887846	X	22,7		A	0.7	0.5 to 0.9	7.8 x 10^{-3}
rs2207031	X	128		A	1.8	1.2 to 2.5	1.4 x 10^{-3}

[a] SNPs sorted by chromosome and position; [b] Position in megabases according to Ensembl release 59; [c] Logistic regression procedure in PLINK toolset, based on allelic test for association with clinical stage, adjusted for age at cancer diagnosis and smoking status. SNPs selected based on P <0.01 threshold for association.

3.3.2 Differences in lung ADCA outcome are associated with patients' genetic profile

Using our polygenic model (289), we evaluate additive effects of these 22 SNP in modulating individual clinical stage. 81 of 917 patients with more than 30% missing genotypes were removed from the dataset. For each patient, the allele-based odds ratio (Table 18) was attributed to the carrier status of an allele of each SNP associated with clinical stage status, based on its association with the probability of carrying a stage >1 lung ADCA. To test the model, a score of +1 was attributed to the carrier status of a risk allele of each SNP based on its association with increased probability of developing lung cancer with higher clinical stage. For each

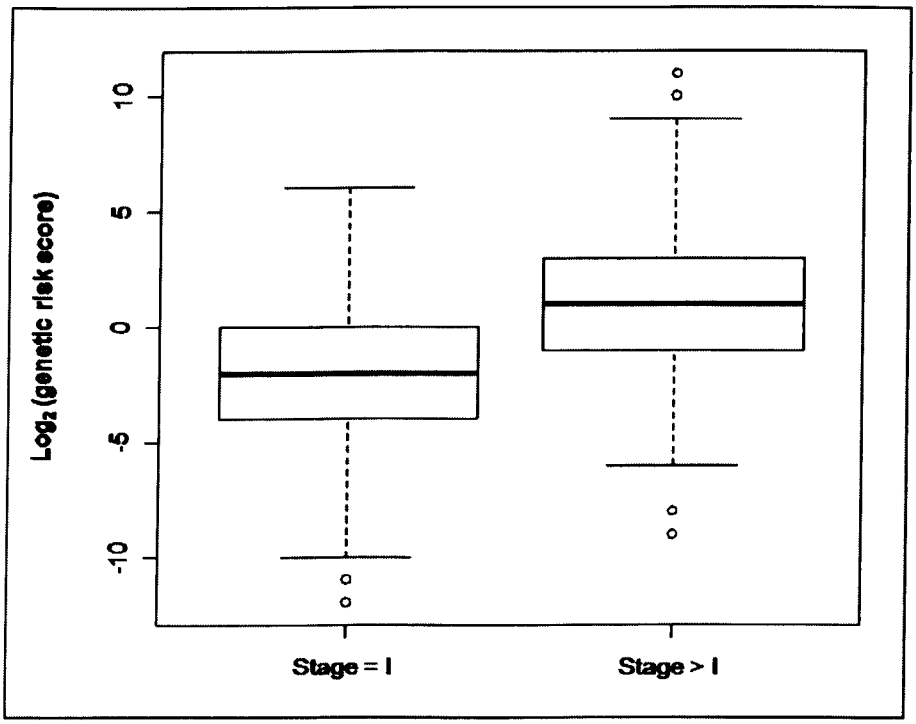

Fig. 27 Genetic risk score in patients with clinical stage I and in patients with higher clinical stage. The horizontal line within the box represents the median value of the genetic estimator of outcome (in base 2 logarithmic units); the upper and lower boundaries of each box represent 75th and 25th percentile, respectively; upper and lower bars indicate the relative highest and lowest values, respectively ($P<2.2 \times 10^{-16}$).

patient, a genetic risk score was calculated as the sum of the scores for all 22 SNPs to obtain a general estimator of individual outcome. The average genetic estimator was $-7.9 \times 10^{-3} \pm 5.4 \times 10^{-4}$ (mean ± standard error) for patients with clinical stage I (n=418) and $3.2 \times 10^{-3} \pm 5.3 \times 10^{-4}$ for patients with higher clinical stage (n=403) ($P<2.2 \times 10^{-16}$, ANOVA analysis, Fig. 27). The 22 SNPs explained 20.7% of the phenotypic variance in clinical staging. Although with a lower size effect as compared to the first series and to the whole series, the genetic estimator was statistically associated to clinical stage in the second ADCA series alone (P = 0.0006, ANOVA analysis), suggesting the predictive value of the 22-SNPs genetic profile on clinical staging of lung cancer patients.

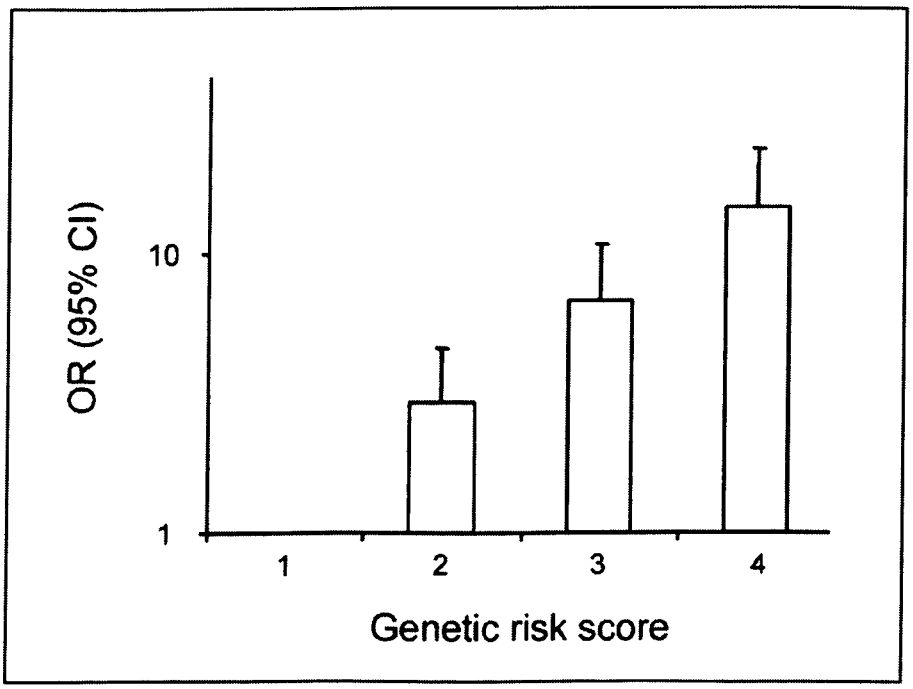

Fig. 28 Genetic risk of developing a more aggressive lung ADCA (clinical stage >1) in patients grouped according to the quartiles of genetic risk score, with the lowest quartile as the reference group. Bars denote ORs. Vertical lines represent 95% CIs.

To verify the robustness of the model in our series, since we did not have sufficient available lung ADCA samples for a larger second replication

step, we carried out an empirical replication using bootstrap samples (B=2000 resamplings), as proposed in a recent paper (290). We found that the difference in the genetic estimator between stage I and stage >I patients was = -11.1×10^{-3}, 95% confidence interval (CI) = -12.7×10^{-3} to -9.7×10^{-3}, $P_{diff} = 0.0005$.

Subjects were divided in 4 groups based on the quartiles of the genetic risk score. Application of the generalized linear model to the quartile groups, with the lowest quartile as the reference, revealed a significant association between the genetic estimator and increased probability to develop a more aggressive lung ADCA cancer (OR= 2.9, 95% CI 1.9 - 4.6, $P=2.7 \times 10^{-6}$ for the second quartile, OR= 6.8, 95% CI 4.4 - 10.7, $P<2 \times$

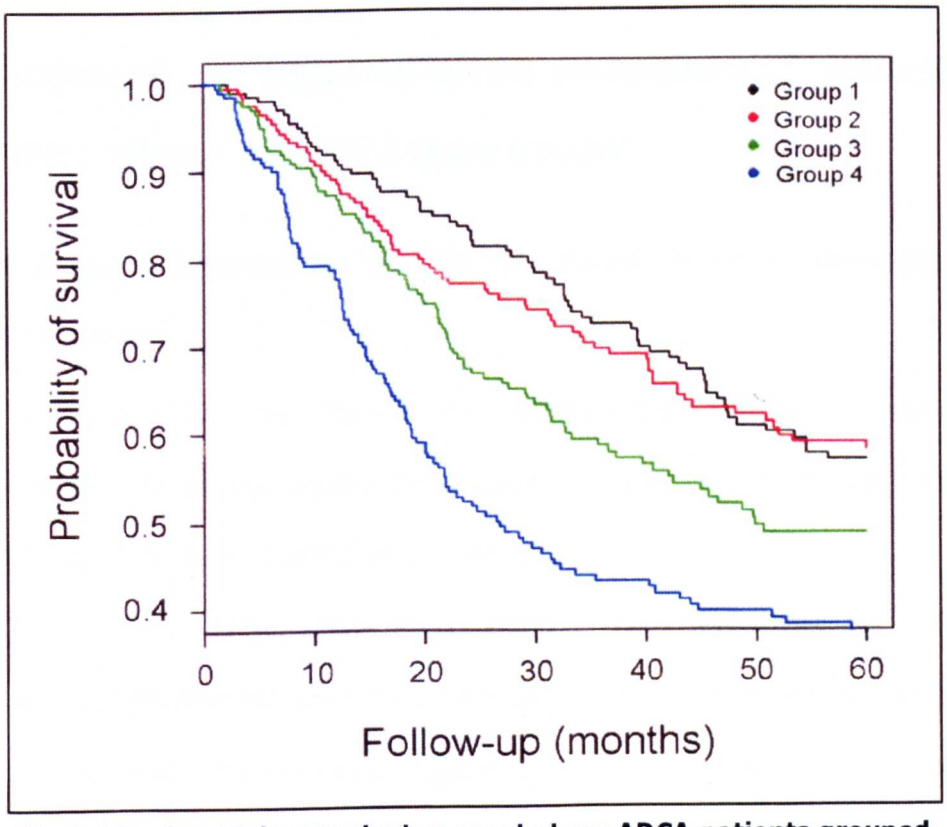

Fig. 29 Kaplan-Meier survival curves in lung ADCA patients grouped as in Fig 3.9. Follow-up is shown truncated at 60 months ($P = 8.0 \times 10^{-8}$, logrank test).

10^{-16} for the third quartile; and OR= 14.5, 95% CI 9.1 – 23.6, $P < 2 \times 10^{-16}$ for the forth quartile group, Fig. 28).

Finally, Kaplan-Meier curves showed statistically significant association of the genetic risk score, in quartiles, and overall survival ($P = 8.0 \times 10^{-8}$, log-rank test; Fig. 29).

Use of multivariate Cox proportional hazard models for survival (adjusted for age and smoking habit) to evaluate the association between the genetic risk score and overall survival showed that the risk of death for the quartiles 3 to 4 (HR= 1.5, 95% CI 1.1 – 2.0, $P = 0.016$; HR= 2.3, 95% CI 1.7 – 3.0, $P = 8.7 \times 10^{-8}$, respectively) was statistically significant higher from that of the lowest quartile.

3.4 RESULTS OF GENOME-WIDE MICROARRAY ANALYSIS IN PATIENT-BASED ASSOCIATION STUDY

3.4.1 A gene expression profile of normal lung is associated with clinical stage

Preliminarily, we found no statistically significant associations between gender or age with clinical stage, indicating that either variables do not modulate clinical staging in our series.

To identify stage-associated genes, we performed a microarray analysis of 120 normal lung tissues from lung ADCA patients, differentially grouped in two microarrays experiments. In normal lung, statistically significant differences in expression levels between clinical stage I and >I patients were detected for 55 (Fig. 30 A) in experiment A and for 361 (the

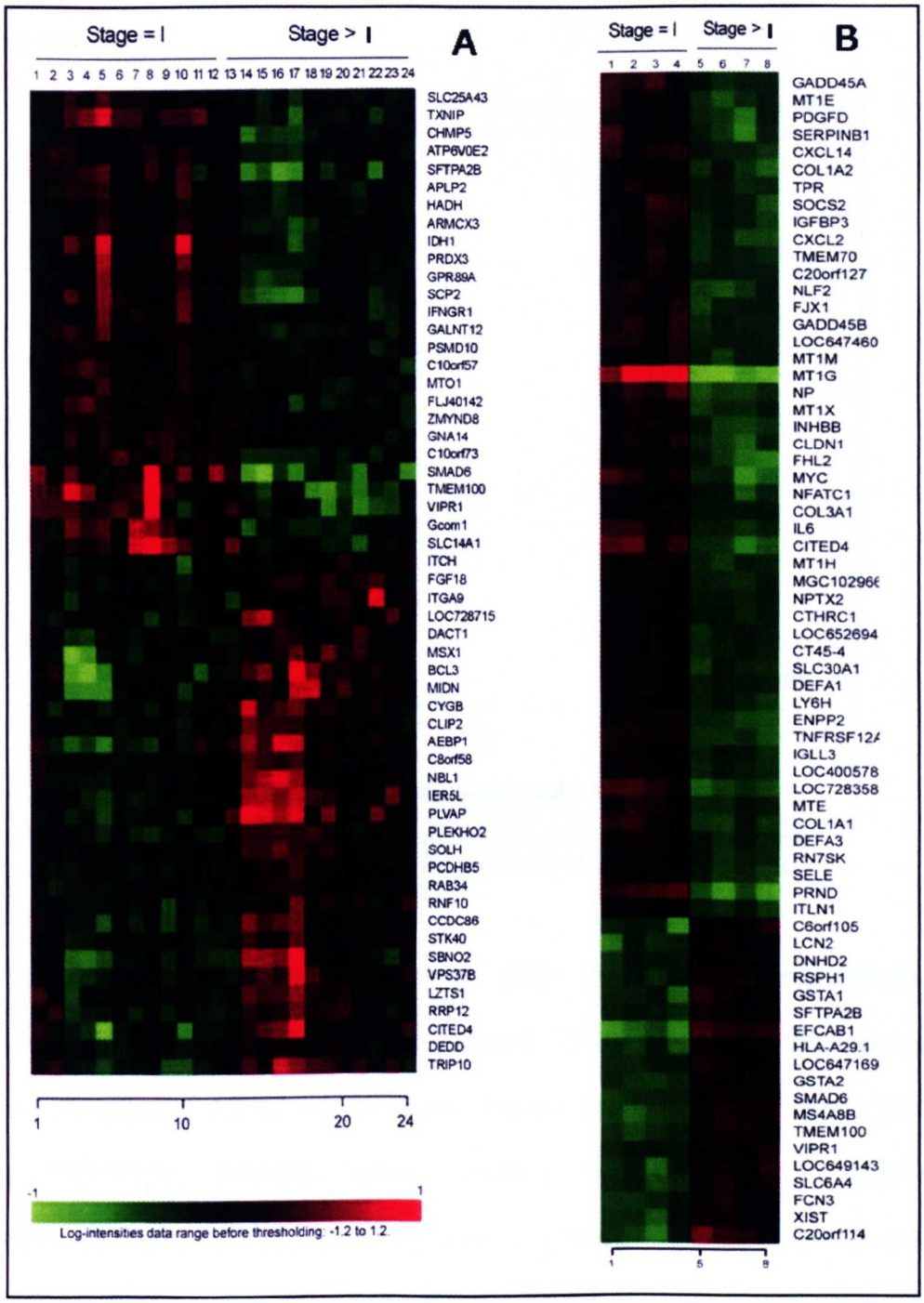

Fig. 30 Heat map of 55 transcripts (at threshold nominal level of $P<1 \times 10^{-3}$) and of top 68 out of 361 transcripts (at threshold nominal level of $P<1 \times 10^{-5}$) whose expression levels showed statistically significant differences in normal lung of stage I as compared to stage >I patients in the first experiment (A) and in the second experiment (B). Gene names are given on the right. Expression levels of the listed genes are indicated by the color bar (green, low; red, high).

top 68 genes with $P < 1 \times 10^{-5}$ are listed in Fig. 30 B) transcripts in experiment B ($P<0.001$, see also Supplementary Table 2 and 3 at the end

of this chapter).

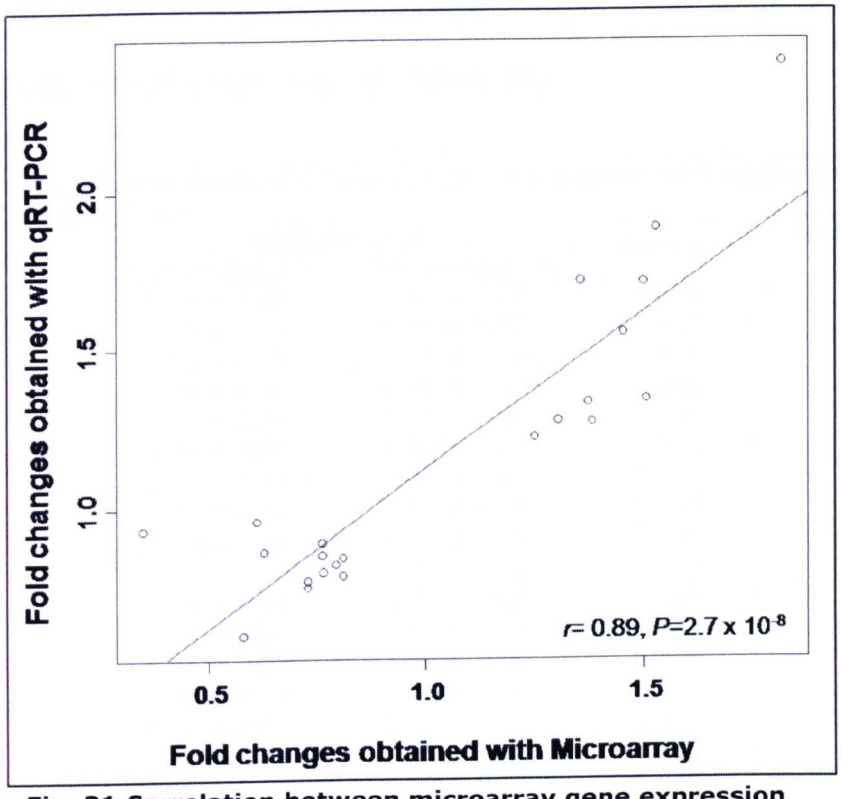

Fig. 31 Correlation between microarray gene expression data obtained on RNA pools and qRT-PCR on individual RNA samples for 22 genes.

To validate the results obtained with microarray experiments, we performed qRT-PCR on the 120 individual RNAs samples. We selected a total of 22 genes, 13 of which were shared by both gene lists (GCOM1, MSX1, TMEM100, SMAD6, IDH1, VIPR1, SLC14A1, BCL3, PLEKHO2, SFTPA2B, SBNO2, RRP12, DACT1), and 2 (TXNIP, LZTS1) and 7 (ITLN1, C20orf114, SELE, FCN3, COL1A1, DEFA3, INHBB) genes that showed the most significant statistical associations in experiment A and B, respectively (Table 8). The correlation between microarray assay and qRT-PCR in detected levels was excellent (r=0.89, $P=2.7\times10^{-8}$, Fig. 31), indicating that microarray analysis detected real variations and that expression data were reproducible.

Among the 22 assayed genes, 11 genes showed a statistically significant difference in expression between stage I and stage > I patients ($P<0.05$, Kruskal-Wallis test, Fig. 32, Table 19).

Table 19. Gene expression results of 22 assayed genes in patients with stage I and higher clinical stage using qRT-PCR.

Gene symbol	RQ mean ± SE		*P*-value		Fold change [a]
	Clinical stage I	Clinical stage >I			
BCL3	0.70 ± 0.06	0.91 ± 0.07	0.052		0.77
C20orf114	1.68 ± 0.43	0.95 ± 0.23	0.014	*	1.77
COL1A1	0.50 ± 0.07	0.75 ± 0.19	0.698		0.67
DACT1	0.83 ± 0.07	0.94 ± 0.07	0.096		0.88
DEFA3	0.51 ± 0.11	1.25 ± 0.51	0.964		0.41
FCN3	1.49 ± 0.18	0.81 ± 0.11	0.005	*	1.84
GCOM1	1.19 ± 0.14	0.85 ± 0.07	0.116		1.40
IDH1	1.13 ± 0.09	0.88 ± 0.04	0.042	*	1.28
INHBB	0.6 ±0 0.08	0.85 ± 0.14	0.198		0.71
ITLN1	0.65 ± 0.21	2.39 ± 1.12	0.610		0.27
LZTS1	0.66 ± 0.09	0.74 ± 0.06	0.045	*	0.89
MSX1	0.63 ± 0.07	0.75 ± 0.06	0.038	*	0.84
PLEKHO2	0.93 ± 0.07	1.00 ± 0.04	0.209		0.93
RRP12	0.80 ± 0.04	1.00 ± 0.06	0.051		0.80
SBNO2	1.02 ± 0.07	1.03 ± 0.05	0.470		0.99
SELE	0.82 ± 0.11	1.29 ± 0.2	0.040	*	0.64
SFTPA2B	1.45 ± 0.26	1.09 ± 0.15	0.280		1.33
SLC14A1	1.46 ± 0.13	1.03 ± 0.12	0.016	*	1.42
SMAD6	1.34 ± 0.16	0.89 ± 0.11	0.044	*	1.51
TMEM100	1.57 ± 0.23	0.86 ± 0.11	0.005	*	1.83
TXNIP	1.18 ± 0.11	0.84 ± 0.06	0.013	*	1.40
VIPR1	1.20 ± 0.17	0.71 ± 0.09	0.040	*	1.69

RQ, relative quantification. * Genes showing statistically significant variation ($P<0.05$) of expression between patients with stage I and higher clinical stage patients using Kruskal-Wallis test. [a] Clinical stage I vs. clinical stage >I.

Among the statistically significant associated genes, FCN3 (ficolin 3) and TMEM100 (transmembrane protein 100) showed the stronger differences between stage I and stage>I patients (fold change>1.8, Table 19) and the best statistical associations with clinical stage ($P=0.005$, Table 19, Fig. 32).

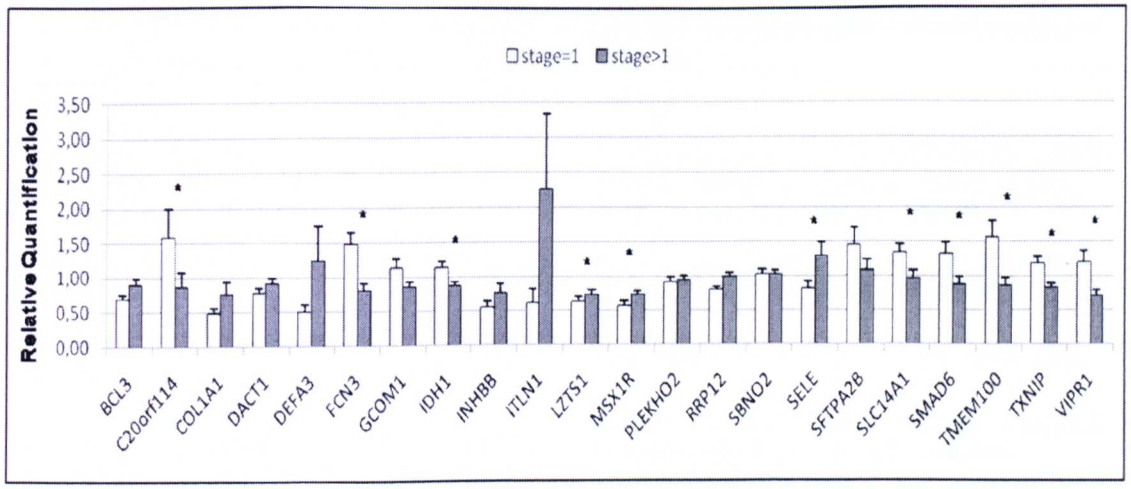

Fig. 32 mRNA expression levels (mean ± S.E.) of genes in normal lung tissue of lung ADCA patients assessed by qRT-PCR by clinical stage. Asterisks indicate statistically significant differences ($P < 0.05$) as compared to the reference group (open bars).

3.4.2 Differential expression profiles of cytokine and cytokine-related genes according to clinical stage

The gene lists obtained from the two experiments in relationship to lung cancer stage were uploaded into the IPA tool for gene network search and into the DAVID Functional Annotation Tool for pathway analysis. The first list of 55 genes revealed no statistically significant results, whereas the first network identified by IPA tool uploading the second list of 361 genes was "Antigen Presentation, Cell-mediated Immune Response, Humoral Immune Response" (Fig. 33). In addition, in overall analysis of biochemical pathways (KEGG database in DAVID tool) the best statistically associated pathway identified in the list of 361 genes was "Cytokine-cytokine receptor interaction" (hsa04060, P=0.00057; Table 20).

Table 20. KEGG pathway analysis by DAVID of the 361 statistically differentially expressed genes between stage I and higher clinical stage patients

KEGG pathway	Gene count	*P* *	Genes
hsa04060:Cytokine-cytokine receptor interaction	15	5.7×10^{-4}	CXCL2, CCL2, CXCL14, TNFRSF12A, CCL3, CSF3R, TNFSF10, INHBB, TGFB3, CSF3, IL7R, IL1R1, IL8, CCL21, CX3CR1
hsa04514:Cell adhesion molecules (CAMs)	8	0.015	ICAM1, HLA-DRB5, CLDN1, SELE, HLA-A29.1, HLA-E, HLA-DMA, VCAM1,
hsa04940:Type I diabetes mellitus	4	0.044	HLA-DRB5, HLA-A29.1, HLA-E, HLA-DMA,
hsa04510:Focal adhesion	9	0.049	LAMB1, COL3A1, THBS2, PDGFD, COL1A2, COL6A2, BCAR1, COL1A1, COL6A3,

* Calculated by the DAVID functional annotation tool, using a modifier Fisher exact test.

Based on this information, we used qRT-PCR on customized TaqMan® Low Density Arrays (assays listed in Table 9) to analyze the 120 individual RNAs samples for expression of 22 cytokine-related genes highlighted by the KEGG biochemical pathway analysis (Table 20) and by IPA tool analysis. Analysis of qRT-PCR in normal lung tissue found that the expressions of 6 genes were statistically different between stage I and stage >I patients ($P<0.05$, Kruskal-Wallis test, Table 21 and Fig. 34), with TNFSF10/TRAIL (tumour necrosis factor ligand superfamily, member 10) showing the best statistical association ($P=0.007$, Table 21, Fig. 34) and IL6 (interleukin 6) showing the higher modulation (~1.5-fold up-regulation in stage >I patients, Table 21).

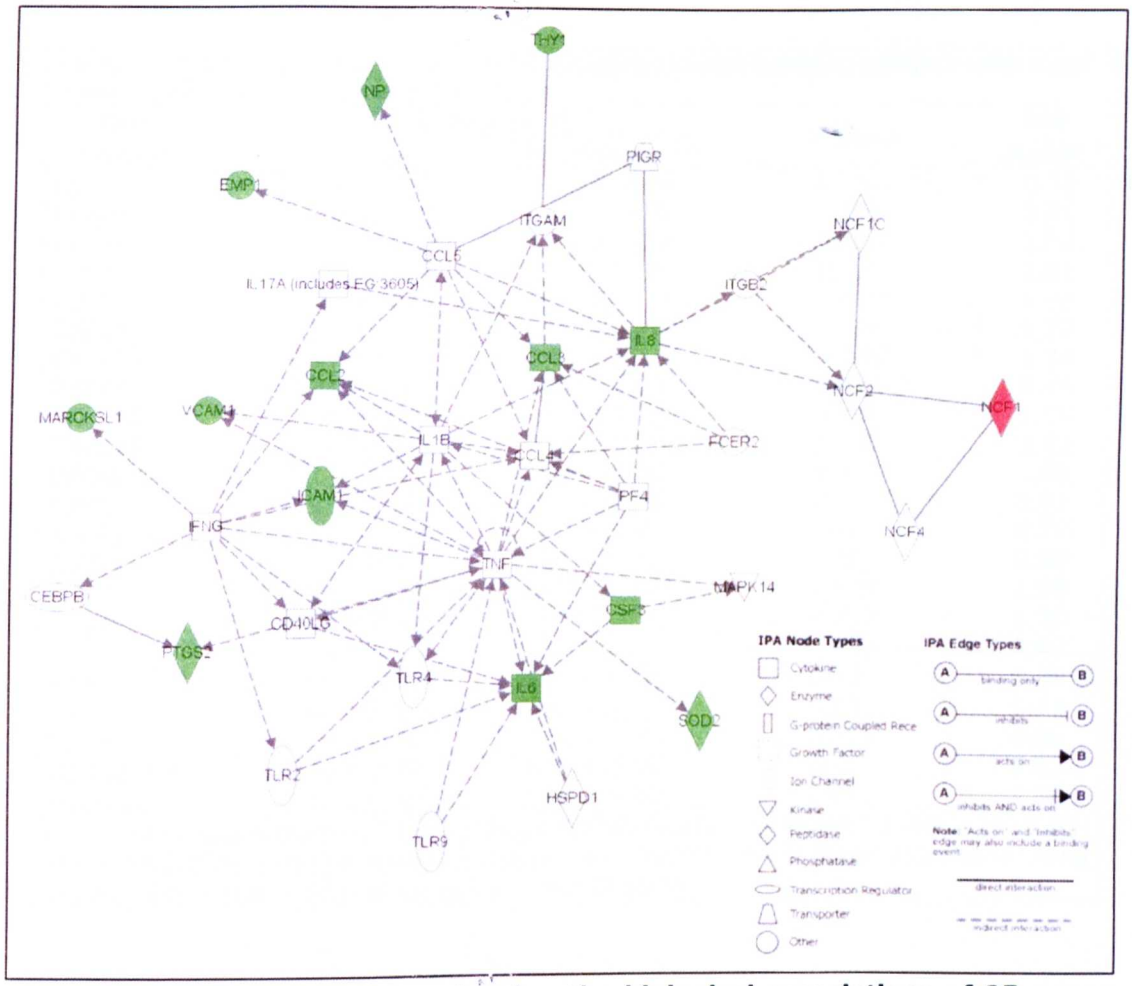

Fig. 33 IPA network diagram showing the biological associations of 35 genes associated with "Antigen Presentation, Cell-mediated Immune Response, Humoral Immune Response". Genes that showed up-regulation or down-regulation in our samples are in red or in green, respectively. The significance of the nodes are displayed using various shapes that represent the functional classes of the gene products as shown in the key.

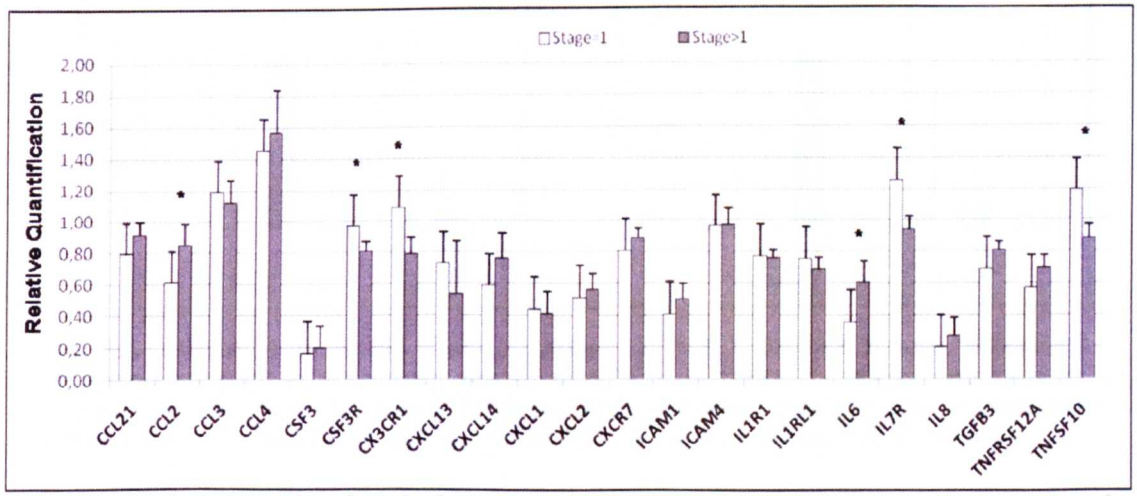

Fig. 34 mRNA expression levels (mean ± S.E.) of cytokine-related genes in normal lung tissue of lung ADCA patients assessed by qRT-PCR by clinical stage. Data are given as in Fig. 34.

Table 21. Gene expression results of 22 cytokine-related genes in patients with stage I and higher clinical stage using qRT-PCR.

Gene symbol	RQ mean ± SE		*P*-value		Fold-change [a]
	Clinical stage I	Clinical stage >I			
CCL2	0.89 ± 0.11	1.23 ± 0.14	0.020	*	0.72
CCL21	0.93 ± 0.08	1.09 ± 0.08	0.079		0.85
CCL3	1.53 ± 0.17	1.50 ± 0.14	0.834		1.02
CCL4L1	2.15 ± 0.28	2.12 ± 0.27	0.629		1.01
CSF3	0.51 ± 0.11	0.64 ± 0.13	0.515		0.80
CSF3R	1.12 ± 0.07	0.92 ± 0.06	0.024	*	1.22
CX3CR1	1.39 ± 0.12	1.04 ± 0.10	0.049	*	1.34
CXCL1	0.73 ± 0.11	0.78 ± 0.14	0.946		0.94
CXCL13	2.09 ± 0.48	1.50 ± 0.34	0.326		1.39
CXCL14	0.90 ± 0.13	1.24 ± 0.16	0.147		0.73
CXCL2	0.73 ± 0.08	0.83 ± 0.10	0.400		0.88
CXCR7	0.91 ± 0.06	1.00 ± 0.06	0.123		0.91
ICAM1	0.62 ± 0.08	0.79 ± 0.09	0.229		0.78
ICAM4	1.21 ± 0.09	1.24 ± 0.10	0.832		0.98
IL1R1	0.85 ± 0.04	0.85 ± 0.05	0.828		1.00
IL1RL1	1.09 ± 0.13	0.84 ± 0.07	0.737		1.30
IL6	0.70 ± 0.10	1.04 ± 0.13	0.032	*	0.67
IL7R	1.54 ± 0.13	1.11 ± 0.08	0.029	*	1.39
IL8	0.48 ± 0.11	0.68 ± 0.12	0.152		0.71
TGFB3	0.79 ± 0.05	0.92 ± 0.06	0.065		0.86
TNFRSF12A	0.75 ± 0.07	0.90 ± 0.08	0.128		0.83
TNFSF10	1.41 ± 0.10	1.04 ± 0.08	0.007	*	1.36

RQ, relative quantification. * Genes showing statistically significant variation ($P<0.05$) of expression between patients with stage I and higher clinical stage patients in using Kruskal-Wallis test. [a] Clinical stage I vs. clinical stage >I.

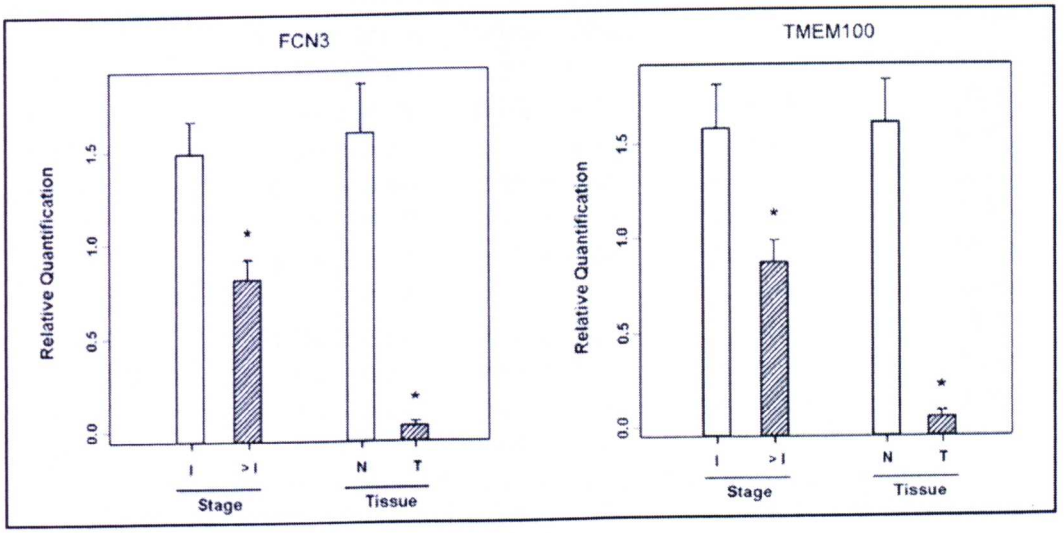

Fig. 35 mRNA expression levels (mean ± S.E.) of FCN3 and TMEM100 assessed by qRT-PCR in lung tissue of ADCA patients by stage and type of tissue (N, normal; T, tumor). Asterisks indicate statistically significant differences ($P < 0.01$) as compared to the reference group (open bars).

3.4.3 Differential expression between normal and tumour tissue

To determine whether the mRNA levels of the 11 modulated genes between stage I and stage > I (and validated by qRT-PCR) are modulated in tumour tissue, we compared gene expression levels in 27 matched pairs of lung ADCA and adjacent normal lung tissue by qRT-PCR. Most of the assayed genes, i.e., 9 out of 11, showed statistically significant deregulation in ADCA tissue as compared to normal tissue ($P<0.05$, Kruskal-Wallis test, Table 22). All deregulated genes, except for IDH1, showed down-regulation in ADCA tissue as compared to normal lung tissue (Table 22). FCN3, SELE and TMEM100 showed ≥40-fold lower mRNA levels in lung ADCA than in normal lung tissue ($P<0.0001$, Table 22, Fig. 35).

Table 22. Gene expression results of 22 assayed genes in lung ADCA tissue and adjacent normal lung tissue using qRT-PCR.

Gene symbol	RQ mean ± SE		*P*		Fold-change [a]
	Normal tissue	Tumour tissue			
C20orf114	1.74 ± 0.69	1.27 ± 0.43	0.849		1.37
FCN3	1.59 ± 0.25	0.02 ± 0.01	2.8×10^{-10}	*	79.5
IDH1	1.29 ± 0.14	1.73 ± 0.25	0.416		0.75
LZTS1	0.97 ± 0.19	0.35 ± 0.04	9.8×10^{-06}	*	2.77
MSX1	0.78 ± 0.17	0.12 ± 0.02	4.0×10^{-09}	*	6.50
SELE	1.33 ± 0.37	0.02 ± 0.00	8.1×10^{-09}	*	66.5
SLC14A1	1.37 ± 0.17	0.12 ± 0.02	1.3×10^{-09}	*	11.4
SMAD6	1.70 ± 0.27	0.15 ± 0.02	1.3×10^{-09}	*	11.3
TMEM100	1.60 ± 0.23	0.04 ± 0.02	5.2×10^{-10}	*	40.0
TXNIP	1.16 ± 0.16	0.27 ± 0.05	2.6×10^{-08}	*	4.30
VIPR1	1.54 ± 0.24	0.05 ± 0.01	2.8×10^{-10}	*	30.8

RQ, relative quantification. * Genes showing statistically significant variation ($P<0.05$) of expression between lung ADCA and adjacent normal lung tissue using Kruskal-Wallis test. [a] Normal tissue vs. tumour tissue.

To select promising candidate genes for further analysis, we decided to perform validation at the protein level for the 9 genes modulated between normal tissue and ADCA tissue using immunohistochemistry. Only

commercial antibodies tested for IHC were selected. Immunohistochemical staining was performed for FCN3, SLC14A1 and SMAD6 on paraffin-embedded tissue sections of lung ADCA and surrounding normal lung tissue to determine whether mRNA over-expression was reflected by an increase of their corresponding proteins in normal and tumour tissue. We confirmed

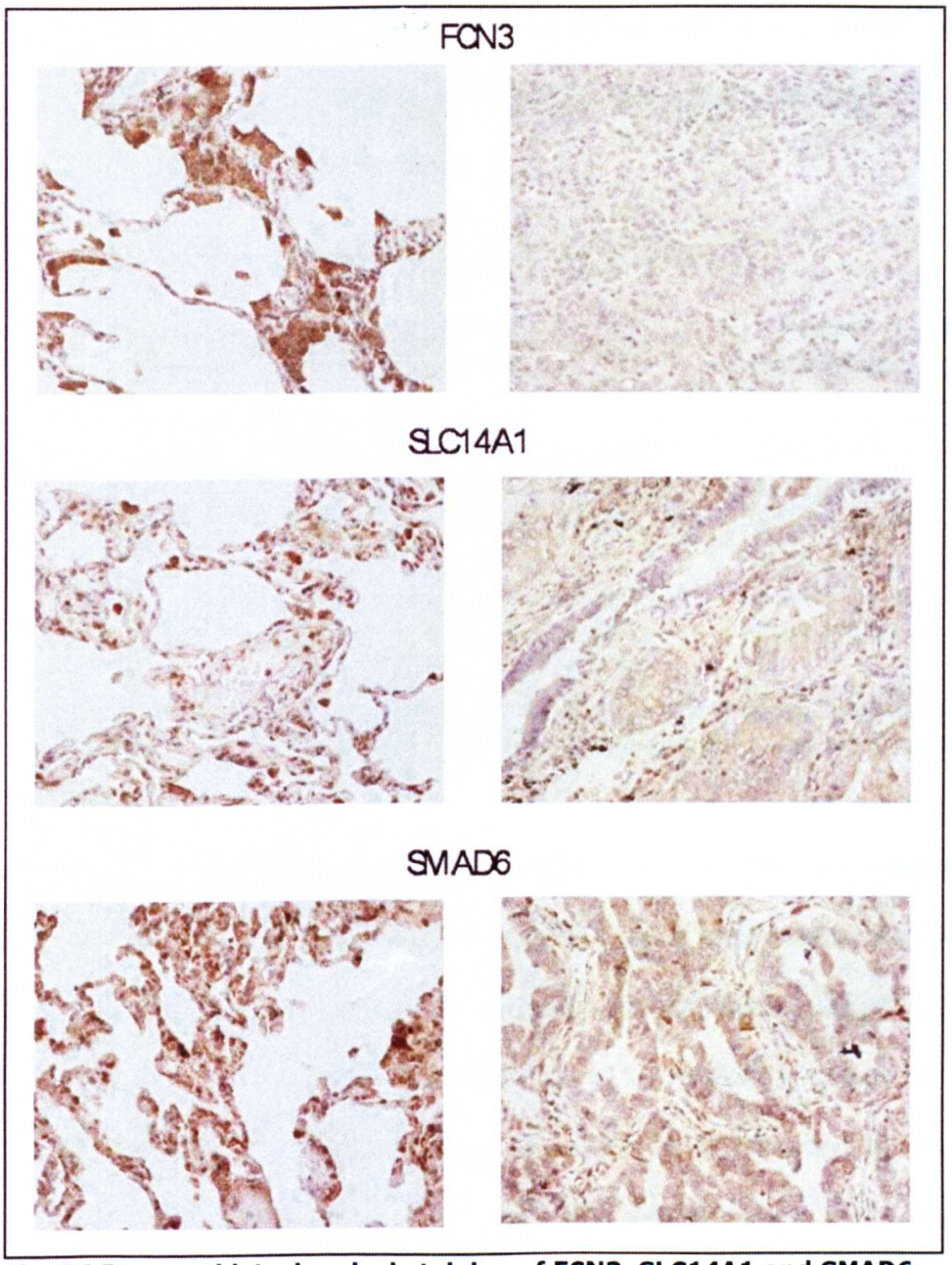

Fig. 36 Immunohistochemical staining of FCN3, SLC14A1 and SMAD6 proteins. No or few proteins were detected in tumour tissues (right panels), whereas a clear staining patterns were observed in normal tissues for each proteins (left panel).

differential expression of proteins FCN3, SLC14A1 and SMAD6 between normal and tumour tissue (Fig. 36).

3.4.4 Integration of GWAS and gene expression profiling

By integrating data from GWA with gene expression signatures from case-only study, we found 6 SNPs, among those with $P \leq 1 \times 10^{-7}$ in GWA, that map within genes slightly differentially expressed $P<0.001$, Table 23) in the second microarray experiment.

Table 23. Integration of GWA data with microarray results.

SNP [a]	P-value GWA	Chromosome	Position (Mb)	Gene*	P-value Microarray	Ratio stage=I vs. stage>I
rs4340697	1×10^{-7}	3	64.60	ADAMTS9	1.45×10^{-4}	0.81
rs11140860	$< 1 \times 10^{-7}$	9	72.52	C9orf135	1.72×10^{-5}	1.30
rs16927500	$< 1 \times 10^{-7}$	11	35.49	PAMR1	5.30×10^{-4}	0.83
rs7305739	$< 1 \times 10^{-7}$	12	13.36	EMP1	3.68×10^{-4}	0.80
rs2839531	$< 1 \times 10^{-7}$	21	43.89	RSPH1	1.80×10^{-6}	1.36
rs1799969	$< 1 \times 10^{-7}$	19	10.39	ICAM1, ICAM4*	2.99×10^{-4}	0.79,0.81

[a] SNPs sorted by chromosome and position; [b] Position in megabases according to Ensembl release 59; * gene in LD with relative SNP (HapMap3 Genome Browser release #2)

Focusing on regions where we identified the 54 most associated SNPs with clinical stage (Table 15), we found detectable expression of 18 of 30 known genes (Table 24) in normal lung tissue of lung ADCA patients, but we found no statistically significant differences in mRNA expression levels between the clinical stage I and stage >I patients at any of the 18 genes (Table 24), suggesting that the SNP candidacy may rest on non-synonymous variations that are in linkage disequilibrium with the identified

SNPs, or on splicing alternative variants rather than alteration of transcript regulation.

Then, we undertook a more direct approach to link GWAS and microarray and to assess whether GWAS and microarray analyses have identified similar sets of genes performing DAVID functional annotations pathway analyses using the 555 unique genes out of 854 genes identified with GWAS among the top 1,334 SNPs (with $P \leq 1 \times 10^{-7}$), and the top 361 genes identified in gene expression data.

Table 24. Integration of microarray results with GWA data.

Gene*	Fold change microarray	SNP	P-value GWA
IL1RL1*, IL1RL2*	1.26	rs951774	4.00×10^{-7}
VGLL4	0.97	rs2574711	$< 1 \times 10^{-7}$
MLL5	1.00	rs2299297	$< 1 \times 10^{-7}$
TSPAN33	1.04	rs2648	$< 1 \times 10^{-7}$
SLC1A1	0.88	rs972519	$< 1 \times 10^{-7}$
LTB4DH, ZNF483	1.07, 1.02	rs10491726	6.00×10^{-7}
FAM107B	1.03	rs11259181	$< 1 \times 10^{-7}$
NUCB2	0.84	rs10832757	$< 1 \times 10^{-7}$
NELL1	0.99	rs7107350	$< 1 \times 10^{-7}$
PPM1H	0.95	rs3825305	$< 1 \times 10^{-7}$
PPP2R5E	1.03	rs1255641	$< 1 \times 10^{-7}$
SPRED1	1.04	rs10520058	$< 1 \times 10^{-7}$
WWOX	1.08	rs10514440	$< 1 \times 10^{-7}$
MATK	1.05	rs12610723	$< 1 \times 10^{-7}$
PKN1	1.03	rs2287700	$< 1 \times 10^{-7}$
PTPRT	0.96	rs6030680	1×10^{-7}
DMD	1.11	rs5972356, rs5927730	$< 1 \times 10^{-7}$, 7×10^{-7}
TCEAL8	1.01	rs404481	1×10^{-7}

*gene in LD with relative SNP (HapMap3 Genome Browser release #2)

Results of functional annotation clustering sorted by statistically significance ($P < 0.05$) can be found in Fig. 37. Interestingly, in both list, most of the top functional clusters derived from GWAS and microarray data

are directly or indirectly related to cell adhesion: cell adhesion molecules and focal adhesion in GWAS results and cell adhesion molecules, extracellular matrix-receptor interaction, and focal adhesion in gene expression data.

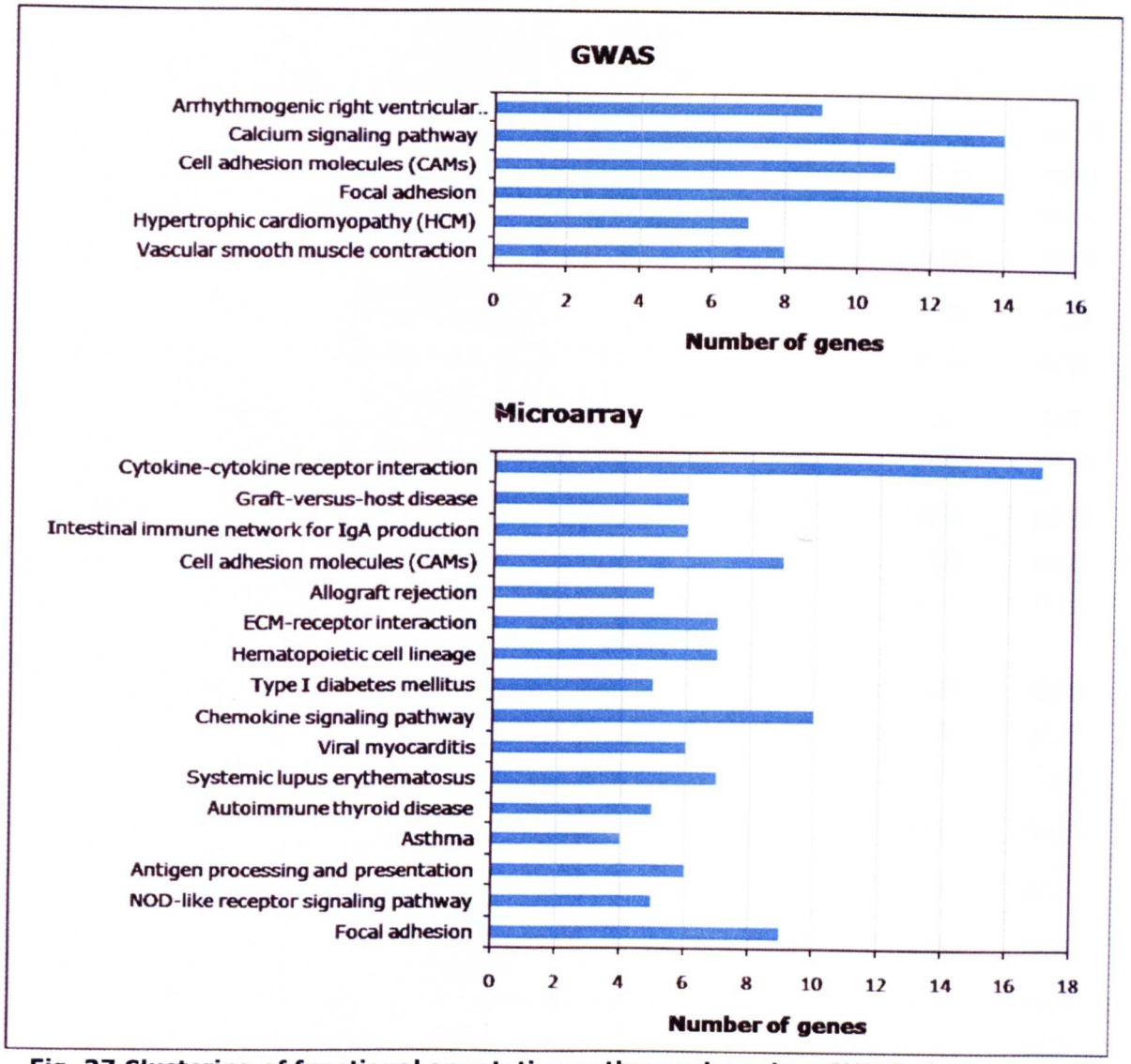

Fig. 37 Clustering of functional annotation pathways based on GWAS- (upper panel) and microarray-dericed genes (lower panel) (with $P<0.05$).

Supplementary Table 2. List of genes differentially expressed between clinical stage I and >I patients in the first experiment.

Gene symbol [a]	P-value [b]	stage=I / stage>I ratio	Gene name	Chromosome	Gene Start (Mb) [c]	Gene End (Mb) [c]
NBL1	3.32×10^{-4}	0.76	Neuroblastoma, suppression of tumorigenicity 1	1	19.97	19.98
STK40	6.31×10^{-4}	0.83	Serine/threonine kinase 40	1	36.81	36.85
CITED4	2.27×10^{-4}	0.69	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 4	1	41.33	41.33
SCP2	7.00×10^{-4}	1.32	Sterol carrier protein 2	1	53.39	53.52
TXNIP	2.33×10^{-5}	1.37	Thioredoxin interacting protein	1	145.44	145.44
GPR89A	8.86×10^{-4}	1.22	G protein-coupled receptor 89A	1	145.76	145.83
DEDD	5.02×10^{-4}	0.86	Death effector domain containing	1	161.09	161.10
IDH1	1.59×10^{-4}	1.33	Isocitrate dehydrogenase 1 (NADP+), soluble	2	209.10	209.13
ITGA9	5.91×10^{-4}	0.80	Integrin, alpha 9	3	37.49	37.87
VIPR1	1.79×10^{-4}	1.45	Vasoactive intestinal peptide receptor 1	3	42.53	42.58
MSX1	8.35×10^{-5}	0.71	Msh homeobox 1	4	4.86	4.87
HADH	2.40×10^{-4}	1.19	Hydroxyacyl-Coenzyme A dehydrogenase	4	108.91	108.96
PCDHB5	5.07×10^{-4}	0.85	Protocadherin beta 5	5	140.51	140.52
FGF18	7.86×10^{-4}	0.83	Fibroblast growth factor 18	5	170.85	170.88
MTO1	7.62×10^{-4}	1.17	Mitochondrial translation optimization 1 homolog (S. cerevisiae)	6	74.17	74.22
IFNGR1	7.86×10^{-4}	1.20	Interferon gamma receptor 1	6	137.52	137.54
AEBP1	2.86×10^{-4}	0.70	AE binding protein 1	7	44.14	44.15
CLIP2	4.97×10^{-4}	0.81	CAP-GLY domain containing linker protein 2	7	73.70	73.82
ATP6V0E2	3.40×10^{-4}	1.20	ATPase, H+ transporting V0 subunit e2	7	149.57	149.58
LZTS1	3.38×10^{-5}	0.76	Leucine zipper, putative tumor suppressor 1	8	20.10	20.16
C8orf58	1.53×10^{-4}	0.82	Chromosome 8 open reading frame 58	8	22.46	22.46
CHMP5	9.70×10^{-4}	1.20	Chromatin modifying protein 5	9	33.26	33.28
GNA14	7.01×10^{-4}	1.18	Guanine nucleotide binding protein (G protein)	9	80.04	80.26
GALNT12	8.77×10^{-4}	1.21	UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 12 (GalNAc-T12)	9	101.57	101.61
IER5L	3.09×10^{-4}	0.73	Immediate early response 5-like	9	131.94	131.94
C10orf73	5.35×10^{-4}	1.27	PREDICTED: chromosome 10 open reading frame 73	10	50.34	50.34
C10orf57	4.96×10^{-4}	1.19	Chromosome 10 open reading frame 57	10	82.17	82.19
RRP12	5.31×10^{-4}	0.78	Ribosomal RNA processing 12	10	99.12	99.16

			homolog (S. cerevisiae)			
PRDX3	2.27×10^{-4}	1.25	Peroxiredoxin 3	10	120.93	120.94
SFTPA2B	4.80×10^{-4}	1.39	Surfactant, pulmonary-associated protein A2B	10	81.32	81.32
CCDC86	2.22×10^{-4}	0.76	Coiled-coil domain containing 86	11	60.61	60.62
APLP2	6.74×10^{-4}	1.21	Amyloid beta (A4) precursor-like protein 2	11	129.94	130.01
LOC728715	9.81×10^{-4}	0.79	PREDICTED: similar to hCG38149	12	31.26	31.36
FLJ40142	6.54×10^{-4}	1.21	FLJ40142 protein	12	110.48	110.51
RNF10	5.94×10^{-4}	0.81	Ring finger protein 10	12	120.97	121.02
VPS37B	2.61×10^{-4}	0.76	Vacuolar protein sorting 37 homolog B (S. cerevisiae)	12	123.35	123.38
DACT1	5.32×10^{-4}	0.81	Dapper, antagonist of beta-catenin, homolog 1 (Xenopus laevis)	14	59.10	59.12
GCOM1	6.27×10^{-5}	1.40	GRINL1A combined protein	15	57.88	58.07
PLEKHO2	3.80×10^{-4}	0.80	Pleckstrin homology domain containing, family O member 2	15	65.13	65.21
SMAD6	1.34×10^{-4}	1.59	SMAD family member 6	15	66.99	67.07
SOLH	3.05×10^{-4}	0.83	Small optic lobes homolog (Drosophila)	16	0.58	0.60
RAB34	5.24×10^{-4}	0.86	RAB34, member RAS oncogene family	17	27.04	27.05
TMEM100	1.08×10^{-4}	1.55	Transmembrane protein 100	17	53.79	53.80
CYGB	9.47×10^{-4}	0.78	Cytoglobin	17	74.52	74.53
SLC14A1	2.04×10^{-4}	1.43	Solute carrier family 14 (urea transporter), member 1 (Kidd blood group)	18	43.30	43.33
SBNO2	4.99×10^{-4}	0.70	Strawberry notch homolog 2 (Drosophila)	19	1.11	1.17
MIDN	4.30×10^{-4}	0.71	Midnolin	19	1.25	1.26
TRIP10	6.12×10^{-5}	0.72	Thyroid hormone receptor interactor 10	19	6.74	6.75
PLVAP	5.73×10^{-4}	0.71	Plasmalemma vesicle associated protein	19	17.46	17.49
BCL3	2.98×10^{-4}	0.64	B-cell CLL/lymphoma 3	19	45.25	45.26
ITCH	6.80×10^{-4}	0.84	Itchy homolog E3 ubiquitin protein ligase (mouse)	20	32.95	33.10
ZMYND8	7.12×10^{-4}	1.16	Zinc finger, MYND-type containing 8	20	45.84	45.99
ARMCX3	3.76×10^{-4}	1.22	Armadillo repeat containing, X-linked 3	X	100.88	100.88
PSMD10	8.65×10^{-4}	1.16	Proteasome (prosome, macropain) 26S subunit, non-ATPase, 10	X	107.33	107.33
SLC25A43	6.42×10^{-4}	1.18	Solute carrier family 25, member 43	X	118.53	118.59

[a] Gene sorted by chromosome and position; [b] p-value obtained by Class Comparison Analysis using BRB ArrayTools version 3.8.1; c Position in megabases according to Ensemble Release 60.

Supplementary Table 3. List of genes differentially expressed between clinical stage I and >I patients in the second experiment.

Gene symbol [a]	P-value [b]	stage=I / stage>I ratio	Gene name	Chromosome	Gene Start (Mb) [c]	Gene End (Mb) [c]
MTE	$< 1 \times 10^{-7}$	0.66	Metallothionein E	1	14.03	14.15
FCN3	$< 1 \times 10^{-7}$	1.53	Ficolin (collagen/fibrinogen domain containing) 3 (Hakata antigen)	1	27.7	27.7
CITED4	1.4×10^{-6}	0.72	Cbp/p300-interacting transactivator, with Glu/Asp-rich carboxy-terminal domain, 4	1	41.33	41.33
GADD45A	9.00×10^{-6}	0.79	Growth arrest and DNA-damage-inducible, alpha	1	68.15	68.15
ITLN1	$< 1 \times 10^{-7}$	0.35	Intelectin 1 (galactofuranose binding)	1	160.85	160.85
SELE	$< 1 \times 10^{-7}$	0.58	Selectin E (endothelial adhesion molecule 1)	1	169.69	169.7
TPR	3.9×10^{-6}	0.77	Translocated promoter region (to activated MET oncogene)	1	186.28	186.34
SLC30A1	5×10^{-7}	0.7	Solute carrier family 30 (zinc transporter), member 1	1	211.74	211.75
LOC652694	6.1×10^{-6}	0.71	PREDICTED: similar to Ig kappa chain V-I region HK102 precursor	2	0.004	0.005
FHL2	3.8×10^{-6}	0.72	Four and a half LIM domains 2	2	105.97	106.05
INHBB	3.1×10^{-6}	0.73	inhibin, beta B (activin AB beta polypeptide) (INHBB).	2	121.1	121.11
COL3A1	3×10^{-7}	0.72	Collagen, type III, alpha 1 (Ehlers-Danlos syndrome type IV, autosomal dominant)	2	189.84	189.88
VIPR1	1×10^{-7}	1.45	Vasoactive intestinal peptide receptor 1	3	42.53	42.58
DNHD2	8.9×10^{-6}	1.34	PREDICTED: dynein heavy chain domain 2	3	55,31	59,53
CLDN1	4×10^{-6}	0.73	Claudin 1	3	190.02	190.04
CXCL2	3.5×10^{-6}	0.77	Chemokine (C-X-C motif) ligand 2	4	74.96	74.97
CXCL14	7.2×10^{-6}	0.77	Chemokine (C-X-C motif) ligand 14	5	134.91	134.91
LOC649143	1.1×10^{-6}	1.46	PREDICTED: similar to HLA class II histocompatibility antigen, DRB1-9 beta chain precursor (MHC class I antigen DRB1*9) (DR-9) (DR9)	6		
SERPINB1	7.6×10^{-6}	0.78	Serpin peptidase inhibitor, clade B (ovalbumin), member 1	6	2.83	2.84
C6orf105	6.9×10^{-6}	1.3	Chromosome 6 open reading frame 105	6	11.71	11.81
HLA-A29.1	5×10^{-7}	1.4	Major histocompatibility complex class I HLA-A29.1	6	29.9	29.9
GSTA2	1×10^{-7}	1.41	Glutathione S-transferase A2	6	52.61	52.63
GSTA1	6.4×10^{-6}	1.36	Glutathione S-transferase A1	6	52.66	52.67

RN7SK	$< 1 \times 10^{-7}$	0.6	RNA, 7SK small nuclear (RN7SK) on chromosome 6	6	52.86	52.86
LOC647169	5×10^{-7}	1.4	PREDICTED: similar to Chain A, Glutathione Transferase A1-1 Complexed With An Ethacrynic Acid Glutathione Conjugate (Mutant R15k)	6	52.63	52.64
IL6	6×10^{-7}	0.72	Interleukin 6 (interferon, beta 2)	7	22.77	22.77
IGFBP3	3.9×10^{-6}	0.77	Insulin-like growth factor binding protein 3	7	45.95	45.96
COL1A2	6.4×10^{-6}	0.77	Collagen, type I, alpha 2	7	94.02	94.06
NPTX2	5×10^{-7}	0.71	Neuronal pentraxin II	7	98.25	98.26
DEFA1	2×10^{-7}	0.7	Defensin, alpha 1	8	6.84	6.84
LOC728358	$< 1 \times 10^{-7}$	0.67	Defensin, alpha 1 (DEFA1B)	8	6.85	6.86
DEFA3	$< 1 \times 10^{-7}$	0.61	Defensin, alpha 3, neutrophil-specific	8	6.87	6.88
EFCAB1	6×10^{-7}	1.39	EF-hand calcium binding domain 1	8	49.62	49.65
TMEM70	5.2×10^{-6}	0.76	Transmembrane protein 70	8	74.88	74.9
CTHRC1	1×10^{-7}	0.71	Collagen triple helix repeat containing 1	8	104.38	104.4
ENPP2	2×10^{-7}	0.69	Ectonucleotide pyrophosphatase/phosphodiesterase 2 (autotaxin)	8	120.57	120.69
MYC	1.4×10^{-6}	0.72	V-myc myelocytomatosis viral oncogene homolog (avian)	8	128.75	128.75
LY6H	8×10^{-7}	0.7	Lymphocyte antigen 6 complex, locus H	8	144.24	144.24
LCN2	4.8×10^{-6}	1.31	Lipocalin 2	9	130.91	130.92
SFTPA2B	5.8×10^{-6}	1.38	Surfactant, pulmonary-associated protein A2B	10	81,31	81,32
FJX1	3.1×10^{-6}	0.75	Four jointed box 1 (Drosophila)	11	35.64	35.64
MS4A8B	5.9×10^{-6}	1.43	Membrane-spanning 4-domains, subfamily A, member 8B	11	60.47	60.48
PDGFD	6.4×10^{-6}	0.78	Platelet derived growth factor D	11	103.78	104.04
SOCS2	4.1×10^{-6}	0.77	Suppressor of cytokine signaling 2	12	93.96	93.97
NP	3.4×10^{-6}	0.73	Nucleoside phosphorylase	14	20.94	20.95
NLF2	2.3×10^{-6}	0.75	PREDICTED: nuclear localized factor 2	15	62.46	62.46
SMAD6	$< 1 \times 10^{-7}$	1.43	SMAD family member 6	15	66.99	67.07
TNFRSF12A	6×10^{-7}	0.68	Tumor necrosis factor receptor superfamily, member 12A	16	3.07	3.07
MT1E	6.5×10^{-6}	0.78	Metallothionein 1E	16	56.66	56.66
MT1M	1×10^{-6}	0.74	Metallothionein 1M	16	56.67	56.67
MT1G	4×10^{-7}	0.74	Metallothionein 1G	16	56.7	56.7
MT1H	3.2×10^{-6}	0.72	Metallothionein 1H	16	56.7	56.71
MT1X	1.7×10^{-6}	0.73	Metallothionein 1X	16	56.71	56.72

LOC400578	< 1 x 10^{-7}	0.67	PREDICTED: similar to Keratin, type I cytoskeletal 14 (Cytokeratin-14) (CK-14) (Keratin-14) (K14)	17	16.73	16.74
MGC102966	6 X 10^{-7}	0.72	PREDICTED: similar to Keratin, type I cytoskeletal 16 (Cytokeratin-16) (CK-16) (Keratin-16) (K16)	17	20.4	20.41
SLC6A4	3 X 10^{-7}	1.47	Solute carrier family 6 (neurotransmitter transporter, serotonin), member 4	17	28.52	28.56
COL1A1	< 1 x 10^{-7}	0.63	Collagen, type I, alpha 1	17	48.26	48.28
TMEM100	1 X 10^{-7}	1.45	Transmembrane protein 100	17	53.8	53.81
NFATC1	9.1 X 10^{-6}	0.72	Nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 1	18	77.16	77.29
GADD45B	6.8 X 10^{-6}	0.74	Growth arrest and DNA-damage-inducible, beta	19	2.48	2.48
PRND	< 1 x 10^{-7}	0.54	Prion protein 2 (dublet)	20	4.7	4.71
C20orf114	< 1 x 10^{-7}	1.82	Chromosome 20 open reading frame 114	20	31.86	31.9
C20orf127	3.8 X 10^{-6}	0.75	Chromosome 20 open reading frame 127	20	33.81	33.81
RSPH1	1.8 X 10^{-6}	1.36	Radial spoke head 1 homolog (Chlamydomonas)	21	43.89	43.92
IGLL3	2.2 X 10^{-6}	0.68	Immunoglobulin lambda-like polypeptide 3	22	23.92	23.98
XIST	8 X 10^{-7}	1.59	X (inactive)-specific transcript (non-protein coding)	X	73.04	73.07
CT45-4	9 X 10^{-7}	0.7	Cancer/testis antigen CT45-4	X	134.93	134.95
LOC647460	5.2 X 10^{-6}	0.74	PREDICTED: similar to Ig kappa chain V-I region HK101 precursor			

[a] Gene sorted by chromosome and position; [b] p-value obtained by Class Comparison Analysis using BRB ArrayTools version 3.8.1; c Position in megabases according to Ensemble Release 60.

4. DISCUSSION

It is now believed that both cancer initiation risk and later neoplastic events (tumour growth, invasion, metastatic spread, response to therapeutic interventions, and survival) may be strongly influenced by factors predetermined by individual's genetic background. Recent progress in decoding the human genome has provided information about thousands of potentially important gene polymorphisms affecting both normal physiological mechanisms and cancer pathogenesis. These variants act through their products involved in various regulatory systems and metabolic chains at different levels of biological organization. It seems likely that combinations of these common polymorphic gene variants frequently found in populations may exert regulation of basic processes such as proliferation, differentiation, and apoptosis, and may influence different stages of carcinogenesis, as supported by several reports of significant associations between germ line variations and risk or prognosis of different cancer types.

At the beginning of my project, population-based association studies were widely-used approach for the identification of common genetic factors affecting common diseases, such as cancer (291), but only few genome-wide studies were carried out. In 2006, this research group has reported a study on >80,000 SNPs that led to the identification of a functional association between the region containing the PDCD5 (programmed cell death 5) gene and lung cancer risk in two independent Caucasian populations (292). In 2007, the same research group reported an

association between SNPs on Kruppel-like factor 6 (KLF6) gene and reduced risk of lung cancer suggesting its candidacy in modulated lung cancer susceptibility (293) using one of the first Affimetrix platforms (100K SNP array). These preliminary studies opened new prospects to carry out additional genome-wide scans in order to investigate the hypothesis of a polygenic inheritance of susceptibility to lung cancer in humans (294).

At the moment, as at the beginning of my project, no GWASs for the identification of lung cancer prognostic germ line variations have been published. However, some studies have suggested the involvement of genetic elements influencing of neoplastic development and leading to differences in patients' prognosis, treatment response and survival rates (141, 142, 292, 295).

4.1 POPULATION-BASED AND FAMILY-BASED ASSOCIATION STUDIES FOR LUNG CANCER RISK

To address the initial aim of identifying genetic determinants of lung ADCA risk we have carried out a genome-wide association study (GWAS) in Italian lung ADCA patients (population-based study) and healthy unrelated controls and a GWAS in Italian lung patients and their unaffected sibs as controls (family-based study).

In the population-based GWAS, we decided to focus specifically on lung ADCA, instead of lung cancer patients in general, to avoid problems deriving from histotype admixture and because this histotype is most likely the one where inherited components play the stronger role as suggested by

its high frequency among young and non-smoker lung cancer patients (94, 296, 297). We used a joint analysis of two independent populations rather than a replication-based analysis to increase power to detect genetic association (226). However, although the correlation between the measurements of allelic frequencies was high, concordance between two different experiments in the identification of SNPs associated to the risk of lung ADCA was not high, leading to 235 SNPs only with *P* values < 0.01 in both experiments. This result might be due to either technical variance representing almost one third of the observed variance, or to the wide genetic differences that make not plausible the comparison of groups of individuals although taken from the same population. Indeed, the compared groups differed much more than expected because of the effect of sampling.

We identified 12 SNPs putatively associated with lung cancer risk. Genotyping of these SNPs in individual samples led to statistical confirmation of 8 of 12 (67%) SNPs. This result provided evidence that the screening system was sufficiently accurate to determine real differences in allele frequency between cases and controls.

The 8 SNPs associated with lung ADCA risk identified several chromosomal regions putatively associated with lung cancer risk (Table 12). Most of the 8 SNPs do not have an apparent functional activity but they most likely represent genetic markers in significant linkage disequilibrium (LD) with the genomic regions containing the functional variations. Among these, SNP rs2515373 on Chromosome 11 maps within intron 3 of the contactin 5 (CNTN5) gene, which encodes a glycosylphosphatidylinositol-anchored neuronal membrane protein that functions as a cell-adhesion

molecule. A role for this protein in tumour invasion and metastasis is possible, since another gene of this same family (contacting 1) encodes a product that modulates invasion and metastasis of lung ADCA cells (298). The other SNPs show significant LD with genomic regions containing genes that may carry functional variations. For instance, SNP rs2172706 maps on Chromosome 1 at a distance of 10 kb from the 30-end of the KCNN3 gene (potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3), and at a distance of 70 kb from the 50 of the ADAR gene (adenosine deaminase RNA-specific). A wide LD region (140 kb) around rs2172706 is observed, preferentially including the ADAR gene and partially including the 30 region of KCNN3. SNP rs4897493 on Chromosome 6 is in LD with the EPB41L2 gene, a member of the protein 4.1 superfamily involved in linking cell surface glycoproteins to the actin cytoskeleton and acting in tumour suppression (299). At present, it is unknown whether the effects of single SNPs on lung cancer risk are mediated by encoded proteins or by non-coding RNAs in LD with the relative SNPs.

These preliminary findings suggest the involvement of multiple common alleles in the inherited modulation of lung ADCA risk in the general population. Indeed, the rare allele carrier status at each of the 8 confirmed SNPs was associated with a significant modulation of lung cancer risk (Fig. 20), suggesting that multiple, and unlinked genetic loci may control individual susceptibility to lung cancer in humans. These findings would not exclude that rare germ line mutations could provide a high risk of lung cancer in carriers; however, such putative mutations would have a low

impact on the risk of the disease at the general population level and would not be detectable by GWA studies.

The results of this first GWAS are consistent with a polygenic inheritance model characterized by dominant or co-dominant effects of rare alleles at 8 unlinked markers. Indeed, we found a significant trend of decreased lung ADCA risk by the carrier status of two or more rare alleles, with effects particularly strong for carriers of four or more rare alleles (OR < 0.3; Fig. 21). Thus, the combination of multiple genetic variants may have a strong effect on lung ADCA risk. Dosage effects and interchangeability of rare alleles in the same individual in modulation of lung ADCA risk suggest that candidate genes act on independent biochemical pathways, as the known functions of genes in LD with the associated SNPs would predict.

It should also be considered that the 8 SNPs associated with lung ADCA risk in our study may show different LD in different ethnic groups, and consequently, may be relevant only in certain populations, because ethnic-related loci are plausible under the assumption of the polygenic model.

Our results are in agreement with findings obtained in the well characterized model constituted by mouse inbred strains, where the polygenic nature of control of strain susceptibility to carcinogen-induced lung tumourigenesis has been clearly demonstrated (31, 300). In addition, we have reported that genetic variants causing an inhibition of genetic susceptibility to lung tumourigenesis are common in inbred strains (301), consistent with the present findings of the reduction of lung cancer risk by the rare alleles in humans. Therefore, the same type (polygenic) of genetic

control of susceptibility to lung tumourigenesis may be operative in both mammalian species, although the genetic loci involved may differ.

Our results suggest that a polygenic control of susceptibility to lung cancer may also operate in humans, leading to possible strong effects on cancer risk of the combination of multiple genetic variants at the individual level and, consequently, representing an important determinant of lung cancer risk in the general population.

Since most of the lung cancer cases that we have analyzed consist of smokers, the identified loci may affect susceptibility to smoking-induced lung cancer.

Of course, this first analysis would not provide exhaustive coverage of the genetic components affecting lung cancer risk, but it would represent a demonstration of the plausibility of the polygenic model of lung cancer risk in the general population and a first example of how genome-wide screens could represent a useful approach to dissect the genetic determinants underlying the susceptibility to common complex diseases. Studies in large population series are needed to confirm our results that could represent a first step toward the definition of a genetic profile for the estimation of individual genetic risk of lung cancer. The future possibility of an estimation of the individual risk of lung cancer could be helpful for the control of lung cancer incidence at population level, since high risk individuals may be more motivated to stop smoking and to undergo early diagnostic procedures (302).

In the second GWAS we planned to use a sibling-based study design to detected loci statistically associated with lung cancer risk. Unlike the

previous GWAS, given the difficult to recruit healthy sibs due to old age at diagnosis for most of the cases, we included all lung tumour histotypes. Despite the poor feasibility of the recruitment of healthy sibs controls to carry out family-based genome-wide association studies in lung cancer, , the possible benefits resulting from the appropriate matching of cases and controls may justify the effort. Even a small size of population (Table 2), this type of study offers complete robustness to potential population heterogeneity eliminating problems that are related to case-control studies with controls from the general population. In particular, an important advantage of the discordant sibs design is the possibility to exclude the potential for bias due to population stratification, which is common in population-based studies (258). Indeed, cases and controls derive from the same pedigree whose DNA differences may lie in genetic polymorphisms putatively responsible for the disease status. The effect (lung cancer risk estimation) detected by the discordant sib pair design (1:1 case:control ratio) and sib transmission disequilibrium test is due to the combined presence of linkage and association (258). In addition, even the limited number of sib-pairs, all cases are non-smokers and younger lung cancer patients. Thus genetic factors may most likely have played a role in lung cancer development in these cases, as they did not smoke and suffered from lung cancer at young age.

In this genome-wide association study using DNA pools, we identified 36 SNPs that showed significant linkage/association in the family-based series (Table 13). Individual genotyping confirmed the robustness of our pooling approach (Fig. 23), demonstrating that this method produces

reliable results and is time- and cost-effective. Of the 36 genetic markers, 13 mapped within genes. The most significantly associated SNPs ($P \leq 0.0045$), i.e., rs11833102, rs17120323, rs12445758 and rs325702, mapped in carboxypeptidase M (CPM), sarcoglycan zeta (SGCZ), cadherin 13, H-cadherin (heart) (CDH13) and cyclic nucleotide-gated channel alpha 4 (CNGA4) genes, respectively. Overexpression of CPM was recently reported to correlate negatively with disease survival in human lung ADCA patients (303), and aberrant methylation of the CDH13 gene was observed in lung ADCA (304). Thus, our findings point to the relevance of genetic components in the modulation of individual lung cancer risk in non-smokers.

Interestingly, we found that one of the associated SNPs (rs12663498, $P = 0.055$) maps to 6q25.1, the same locus previously linked to lung cancer risk in pedigrees with multiple lung cancer members (126). The SNP maps within the pleckstrin homology domain containing family G (with RhoGef domain) member 1 (PLEKHG1) gene, which lies 2 Mb from RGS17, the major candidate gene for the familial lung cancer susceptibility locus (127).

The application of the previously proposed polygenic model to the 35 SNPs associated in the discordant sib-based series showed a highly statistically significant association between the genetic susceptibility score and the proportion of lung cancer cases (Fig. 24).

Our single-point analysis confirmed in the population-based series only 3 of 36 SNPs that were statistically associated in the family-based series (Tables 14). This result could be expected if we consider the differences between these two series, i.e., the family-based series is constituted by young non-smoker lung cancer patients whereas the

population-based series is constituted by mostly smokers with a higher median age at tumour diagnosis. In addition, since the results of the family-based series are not biased by population structure and most of the detected SNPs presumably represent real associations, the scarce effects of the same SNPs in the population-based series rest on either the existence of significant population admixture, masking real associations, or the existence of phenocopies and a high degree of genetic heterogeneity in the general population. In the latter case, a model of "private" genetic epidemiology (305) may account for the genetic effects detected in lung cancer families. Interesting, in mouse models, our group recently detected a high degree of genetic heterogeneity affecting genetic susceptibility to skin tumourigenesis and to inflammatory response, i.e., the same phenotype being linked to different loci in different mouse lines (306), thus supporting the role of the "private" genetic epidemiology in an experimental model.

Another aspect that we should take into consideration is the role of genetic heterogeneity in the predisposition to cancer. Indeed, independent loci may modulate the risk of sporadic and of familial cancer, as the model of breast cancer susceptibility demonstrated (307, 308). Also, we should consider the great impact of the major environmental risk factor, i.e., smoking habit, and the difficulty in separating the genetic and environmental contributions to lung cancer risk. Indeed, a study in monozygotic and dizygotic twins showed that the possible sharing of the same environmental risk factors may play a major role in lung cancer risk (309).

In 2008, three separate GWASs on several thousand of samples were published and all three studies found a region on chromosome 15q25 associated with lung cancer risk (130-132). Two of the GWASs identified polymorphisms directly associated with lung cancer (130, 131), whereas the third study identified an association between the same genetic region and nicotine dependence and concluded that the association with lung cancer goes indirectly through smoking (132).

Even if not included in the 47 SNPs with $P < 1 \times 10^{-7}$ and thus not considered for further analysis, SNPs of that region (rs12916375) was included in our initial top list of 235 SNPs ($P < 0.01$) in population-based study, indicating that also in our series the chromosome 15q25 region may be involved in lung ADCA risk in general population. To further test the candidacy of these region in our population and having already available DNA pools from our population-based case-control study, we analyzed two coding polymorphisms reportedly associated with lung cancer risk: rs1051730, a synonymous change within the CHRNA3 gene (130-132), and the rs16969968, a D398N polymorphism of the CHRNA5 gene (131). Because the rs1051730 showed a slightly weaker statistical association with lung cancer risk as compared with the rs16969968 and because of an almost complete linkage disequilibrium between these two single-nucleotide polymorphisms in the European population, only the CHRNA5 polymorphism was analyzed in individuals of the whole series (Table 1). The frequency of the A (398Asn) allele differed significantly between controls and cases (0.41 and 0.48, respectively; $P = 0.0001$) with the homozygosity status of the A allele significantly associated with lung ADCA risk (OR=1.9, 95% CI 1.3-

2.7; P = 0.0003) as well as the heterozygosity status (OR=1.4, 95% CI 1.0-1.9; P = 0.024) when compared with GG (Asp398Asp) homozygous subjects. Comparison of subjects carrying the A (398Asn) allele as dominant effect versus GG (Asp398Asp) homozygous subjects also showed a significant association of the A allele with lung ADCA risk (OR=1.5, 95% CI 1.2-2.0; P = 0.002). No significant associations of the CHRNA5 D398N polymorphism with patients' clinical stage or overall survival were detected. (310, 311). Our findings in non-smokers discordant sib pairs did not confirm the previously reported population-based association of lung cancer risk with the chromosome 15q25 nicotinic receptor locus. These results are consistent with a recent meta-analysis in >1000 never-smoker cases and >1800 controls (312) and by a recent pooled analysis (136), showing that this locus is not associated with lung cancer risk in never-smokers. Interestingly, a large GWAS in never-smokers found statistically significant association between lung cancer and a locus at 13q31.3 (137). All these findings suggest that the genetic factors for risk in smokers and never-smokers may be different and that lung cancer risk in non-smokers may have an inherited susceptibility component that may take the place of the strong role played by the smoking habit in smokers (313), as reviewing in (314).

Subsequent GWASs identified lung cancer susceptibility loci also at 6p21 (133, 134), and 5p15.33 (133, 135), providing further powerful evidence of a genetic contribution to lung cancer, even if several discrepancies due to population characteristic such as ethnicity and smoking (136). In fact, in Asian population the association has not been confirmed

for the variants in 15q25 reported in the Caucasian studies, due to their rare allele frequencies, and no variants in 6p21 were replicated in Caucasians (136). In our two GWA studies, we did not find any associations among the top SNPs between polymorphisms in 5p15.33 and 6p21.33 and lung cancer risk. When we genotyped rs4016181 (in 5p15.33, CLPTM1L gene) and rs3117582 (in 6p21.33, BAT3-MSH5 gene) in individual samples from our population-based case-control study we found a borderline association with lung ADCA risk only for the SNP in 5p15.33 locus P=0.02).

Since analysis of candidate genes located in these regions by individual studies has had only limited success in identifying specific variants that are conclusively associated with lung cancer risk, the International Lung Cancer Consortium (ILCCO) recently conducted a genotyping study in a total of 8,431 lung cancer cases and 11,072 controls of European and Asian ethnic groups (128). This study suggests that only the SNP rs560191 (TP53BP1) of ten variants tested is associated to lung cancer risk and refuses all other associations focusing on the importance of consortia and of great case-control studies in replicating or refuting published genetic associations. Notwithstanding the identification of these loci (15q25, 5p15 and 6p21) associated with a modulation of lung cancer risk in particular population, a model explaining the complex genetics of lung cancer predisposition different to our proposed polygenic model is still waiting to be defined.

4.2 CASE-ONLY ASSOCIATION STUDY FOR LUNG CANCER PROGNOSIS

Tumour progression is defined as the dynamic stepwise process through which neoplastic cells evolve towards more malignant characteristics and more aggressive clinical behaviour (315). This process is a critical point in clinical cancer management since most cancer deaths still result from metastasis and the spread of cancer to other parts of the body begins early in the growth of the primary tumour (316). In the last years, variations in cancer aggressiveness and malignancy have been mainly associated with the accumulation of multiple somatic alterations and epigenetic changes in the neoplastic cells (317). Indeed, most studies aimed to identify factors that affect cancer patient's outcome/survival are focused on genetic alterations or transcriptional changes in cancer tissues. However, such studies have ignored the fact that cancer is a mass of heterogeneous cells whose growth is dependent upon reciprocal interactions between genetically transformed cells and the microenvironment in which they live. Indeed, genetic studies carried out in experimental mouse models support the biological plausibility of a genetic modulation of cancer progression (295), suggesting that germ line variations may also play a role in the control of lung cancer patients' outcome. Although in the last years several GWA studies have focused on genetic risk for lung cancer, none has examined the possible genetic modulation of lung cancer staging, that is the most powerful prognostic factor in cancer patients (318).

To address the hypothesis stating that genetic constitution might also contribute to tumour development, we planned to investigate the role of

genetic control in lung cancer progression through case-only association studies in a group of Italian patients affected by lung ADCA. We collected an initial relatively large case series containing patients of the same lung cancer histotype, i.e., ADCA, and of the same ethnicity, which follow-up and epidemiological data were available. Clinical stage according to the tumour-node-metastasis (TNM) system (100, 319) is the most powerful prognostic factor in cancer patients and the main determinant of lung cancer patient's survival. Thus, in order to look for possible coding SNPs that could explain the different effects on lung tumour prognosis, we divided our series according to their clinical staging, comparing stage I and higher clinical stage patients, and we investigated SNPs with different allele frequencies between the two groups.

We identified 63 SNPs putatively associated with clinical stage. Genotyping in individual samples led to statistical confirmation of 54 of 63 (85.7%) SNPs, demonstrating the robustness of our pooling approach (Fig. 26). The most significantly associated SNPs ($P \leq 5.0 \times 10^{-6}$), i.e., rs10278557, maps to chromosome 7 in the intronic region of the mesenchyme homeobox 2 (MEOX2) gene, also known as growth arrest-specific homeobox (GAX) gene, which encodes a member of a subfamily of non-clustered, diverged, antennapedia-like homeobox-containing genes. The encoded MEOX2 protein is a key regulator of vascular-cell function. MEOX2 has been proposed as a candidate tumour suppressor gene in Wilms tumour, and showed differential expression and aberrant methylation in lung cancer (320, 321). To test the reproducibility of our results we chose two smaller lung cancer populations with different lung histological type

(317 ADCA and 257 SQCC, Table 3). Even though ADCA and SQCC belong to the same main lung cancer histological group of non-small cell lung cancer (NSCLC), differences in etiologic, clinical and molecular characteristics have been widely reported (322). Indeed, one of the loci recently associated with lung cancer risk, the 5p15.33 locus, was found significantly associated in ADCA subtype but not in squamous cell carcinoma (212, 213). Also the results of our confirmation analyses identify different associations between ADCA and SQCC series, reflecting the differences across histology. In fact, though the loss of statistical power in confirmation series due to the smaller number of subjects compared to discovery series, logistic regression analysis of the same 54 SNPs pointed to 3 SNPs significant associated with clinical stage in ADCA series that were not confirmed by analysis in SQCC series. Our findings suggested that several loci are involved in the modulation of lung tumour progression in general population and that the involvement is strongly histology-specific.

Joint analysis of the GWA and replication series to increase the statistical power of the study and to obtain an overall unbiased estimate (288) identified 22 SNPs that, at nominal statistical P-value <0.01, showed statistical association with clinical stage (Table 19). Analysis of additive effects of risk associated to the minor alleles of these 22 SNPs using a polygenic model (289, 323) in 917 lung cancer patients (Table 3) revealed a statistically significant association between the general estimator score and increased risk of higher clinical stage (Fig. 27 and 28) and with risk of death (Fig. 29), suggesting the complex genetic control of lung ADCA patients' clinical prognosis. The predictive value of the genetic estimator calculated

on the 22 SNPs genetic profile was statistically associated to clinical stage also in the second smaller ADCA series only (P=0.0006).

Empirical replication using bootstrap samples from the original data, rather than replication in independent samples, has been proposed in association studies since bootstrap samples likely share the same population structure of original data, whereas an independent series may be characterized by a different population structure and, thus, lead to false-negative results on analysis (290). Our empirical replication using bootstrap samples confirmed the statistically significant difference between stage I and stage >I patients in their genetic estimator based on 22 SNPs.

Of the 22 candidate SNPs, ten mapped within genes. Among these, the most significantly associated SNP in the joint analysis (rs10278557, P = 1.1 x 10^{-5}, Table 19) maps on chromosome 7 in the intronic region of the mesenchyme homeobox 2 (MEOX2) gene, described above. Other important genes are the myeloid/lymphoid or mixed-lineage leukemia 5 (trithorax homolog, Drosophila) (MLL5), the sprouty-related, EVH1 domain containing 1 (SPRED1, rs10520058, Table 4), and WW domain containing oxidoreductase (WWOX, rs10514440, Table 4) candidacies are also of interest.

Indeed, MLL5 gene belongs to the evolutionarily conserved trithorax family of human genes that activate and regulate diverse genes, including homeobox (HOX) genes that are important in oncogenesis and tumour suppression (324, 325). MLL5 is located on chromosome 7q22, which frequently is deleted in myeloid leukaemia, and recent studies demonstrate that MLL5 is a key regulator of normal haematopoiesis (326).

SPRED1 gene negatively regulates the Ras-ERK signalling pathway, cell motility, and metastasis, and its germ line loss-of-function mutations cause a neurofibromatosis 1-like syndrome (327, 328).

WWOX gene acts as a tumour suppressor in different tumour types and plays a regulatory role in a wide variety of cellular functions such as protein degradation, transcription, and RNA splicing (see review in (329)). More recently, WWOX was found to be often altered or silenced by promoter methylation in NSCLC (330).

At present, it is unknown whether the observed associations between SNPs and lung cancer clinical stage underlie effects of non-synonymous or regulatory variants in linkage disequilibrium with these SNPs. Replication in large cohorts of patients and on different types of cancer would provide strong information whether a SNP may have a role on cancer prognosis and whether this effect is specific only for a subset of tumour types.

Together, these results indicate for the first time that clinical staging of lung ADCA can be under genetic control, with each individual patient displaying a own tendency toward a low or high clinical stage, modulated by individual genetic variations. Indeed, it presented the first effort to identify whole genomic alterations that determine different outcome in lung cancer patients and would allow to draw a SNP profile associated with lung cancer clinical stage and overall survival, representing a first step towards the possible clinical use of such a profile for the personalized follow-up of patients at genetic risk of poor clinical outcome. The significant association of the 22 SNPs with lung cancer clinical stage and survival opens the possibility that the functional products of the genes linked to these SNPs

use novel biochemical pathways associated with lung cancer patients' outcome, and that the identification of these pathways might provide gene targets for therapies to counter lung cancer progression.

It seems that much of the genetic architecture of complex traits remains unexplained. A new strategy should be developed for estimating the degree of false positive finding. In order to analyze the role of genetic heterogeneity, SNPs panels assembled in the last few years that permit to identify ethnic and sub-ethnic group, as well as individuals in paternity testing could be useful (331). The use of these panels has been proposed in controlling for admixture in association studies (332). Genotyping such a panel of SNPs in our series would allow identification of genetically-related subgroups of individuals. In turn, adjusting by genetic clusters may allow highlighting genetic differences between cases and controls that would be masked by genetic heterogeneity. Thus, further clarification of the role of genetic mechanisms in lung cancer patients' outcome may hold the promise of improved therapy and disease outcome.

It is also known that SNPs in regulatory elements can affect gene expression levels. Therefore, we planned to analyze whole-genome expression profiles in normal lung tissue from patients with different clinical stage, in order to identify transcripts whose expression levels are associated with lung ADCA prognosis.

The identification of candidate genes by the transcriptional profile analyses allows tracing possible biochemical pathways that are associated with lung cancer prognosis. This could overcome the genetic heterogeneity of this disease, reducing its complex genetic architecture to fewer

pathways. Over the last years there has been an increase in the use of microarray technology in cancer research for transcriptional analysis of primary tumours. Indeed, most of the studies use microarray analysis of tumor tissues compared with normal tissues for profiling of molecular characteristics in order to identify possible classifiers for prognosis (333,334) or to predict for aggressive forms of different stage of cancer (335). In lung cancer, the microarray analysis has identified gene expression profiles related to disease recurrence, prognosis and survival in ADCA (336) and in SQCC patients (337). The number of publications relating to the use of microarrays for analysis of normal tissue is much more limited. There are some studies that used normal tissue to generate gene signatures that discriminated cell populations in sensitive and resistant to radiotherapy or to identify genes and pathways involved in tissue response to radiation injury (reviewed in 338). Recently, some studies analyzed gene expression profiles comparing normal breast tissues from cancer patients with normal breast tissues from non-cancer patients and indicated that gene alteration associated with tumor development is already detected in normal tissue, leading to higher risk for development of a malignant disease in the breast (339,340).

Under this hypothesis that the dynamic microenvironment in which cancer grows may influence its aggressiveness and that individual genetic constitution may affect the expression profile of normal lung and also explain differences in the cancer outcome, we performed a genome-wide transcriptional analysis in normal tissue comparing gene expression profiles of lung ADCA patients with different stages. Although we cannot exclude

possible presence of micro-metastases in tumour adjacent normal tissue, macroscopic analysis in our samples indicated no apparent contamination of cancer cells. Lung gene expression patterns could be altered by genetic heterogeneity of human population and by environmental factors, first of all exposure to cigarette smoke. We attempted to minimize all these confounding factors by studying a relatively large number of well-characterized Italian ADCA patients and performing gene expression analysis only among smokers. Thus, we analyzed expression level of 120 normal tissues from smoker ADCA patients comparing stage I with higher (Table 4), in order to examine relationships between gene expression profiles in normal tissue and staging.

In order to verify microarray reproducibility and estimate technical variability we used a joint analysis of two independent microarray experiments in RNA pools and we identified a set of 11 stage-associated genes able to distinguish patients with stage I from patients with higher stage (Table 21). This gene set included genes that are biologically plausible contributing to pathogenesis of disease. Indeed, of the two genes whose transcript levels in normal lung tissue showed the higher modulation between stage I and stage>I patients and the best statistical association with patients' clinical stage, FCN3 (ficolin 3) encodes a collagen-like defence molecule that is involved in the maintenance of tissue homeostasis and of the innate immune system and acting as recognition molecules in the complement system (335). FCN3 might play a protective role against the development of autoimmunity (336) and FCN3 deficiency is associated with immunodeficiency and with susceptibility to fever, neutropenia, and

infection (337, 338). Interestingly, susceptibility to infection has been reported to increase the risk of cancer, including lung cancer (339).

The other best statistically associated and most modulated gene, TMEM100, encodes a transmembrane protein of unknown function(s); in the developing mouse embryo, TMEM100 is expressed predominantly in endothelial cells and thus might be involved in angiogenesis (340).

Among the other genes, C20orf114 gene, known also as LPLUNC1, encodes for a protein that is expressed in the upper respiratory tract and oral cavity, and that may function in host defence (341).

IDH1 (isocitrate dehydrogenase 1) gene encodes for a NADP(+)-dependent isocitrate dehydrogenase, that has a significant role in cytoplasmic NADPH production and in peroxisomal NADPH regeneration and whose coding mutation at the arginine in 132, that results in loss of the enzyme's catalytic activity, was associated with malignant gliomas (342) and thyroid cancers (343).

LZTS1 (leucine zipper, putative tumour suppressor 1) encodes a tumour suppressor protein ubiquitously expressed in normal tissues and its expression is often much lower in tumour tissues (344) confirming our results in ADCA lung tissues. It may have a role in cell-cycle control by interacting with the Cdk1/cyclinB1 complex and preventing the uncontrolled cell proliferation. Loss of heterozygosity (LOH) in the LZTS1's locus is a common characteristic of many types of cancer as ovarian carcinoma (345), oral squamous cell carcinomas (346) and bladder cancer (347).

MSX1 (msh homeobox 1) encodes a small member of the muscle segment homeobox gene family that functions as a transcriptional repressor

during embryogenesis interacting with components of the core transcription complex (348).

SELE gene encodes for E-selectin, a protein involved in cell adhesion and responsible for the accumulation of blood leukocytes at sites of inflammation by mediating the adhesion of cells to the vascular lining. High serum E-selectin levels had prognostic significance and could be a potential prognosis factor in NSCLC patients (349).

SLC14A1 (solute carrier family 14, member 1) gene encodes for a membrane transporter.

SMAD6 (SMAD family member 6) gene encodes for a signal transducer, whose expression effects the progression of oesophageal squamous cell carcinoma (350) and high expression levels are associated to prognosis and improved survival in oral squamous cell carcinoma patients (351).

TXNIP (thioredoxin interacting protein), also known as vitamin D3 up-regulated protein 1 (VDUP1), is a known tumour suppressor gene, that is involved in redox stress responses (352), regulation of cellular proliferation (353), and in the differentiation of myeloid and macrophage lineages (354). Its expression is frequently lost in cancer tissue including breast, gastrointestinal, renal, and liver tumours (355-358). Our findings of the down-regulation of TXNIP expression in lung cancer are in agreement with similar observations reported in small series of NSCLC (359).

VIPR1 (vasoactive intestinal peptide receptor 1) gene encodes for a small neuropeptide involved in ion flux in lung and intestinal epithelia that was proposed as tumour suppressor since it was found down-regulated in

lung adenocarcinoma (360). Most of these genes seem involved in the control of the response of innate or acquired cellular immunities; therefore, immunity response as detected in normal lung by transcriptional analysis may be associated with clinical stage in lung ADCA patients.

Biochemical pathway analysis of the whole transcriptional profile indicated the involvement of cytokines and cytokine receptors. Overall, the biochemical pathways of genes in normal lung tissue that were associated with clinical staging in lung ADCA patients are involved in the control of inflammation and infection (Table 20). Of the detectable genes, most were found to be up-regulated in normal tissues from patients with higher clinical stage, indicating the crucial role of these inflammatory mediators in tumour growth and progression. It is known that lung tissue samples subjected to gene profiling may contain an abundance of migratory inflammatory cells and blood vessels so that analysis of whole lung tissue represents an amalgam of expression by all of these cell types. However it is interesting that inflammatory responses were more evidence in normal tissue surrounding the tumour at advanced stages. These findings are consistent with reports showing that non-malignant lung stromal areas in advanced-stage non-squamous cell lung carcinoma contain high levels of neutrophil infiltration and vascular endothelial cells recruited by chemokines/cytokines (361-364). Their expression was differentially regulated in the tumour and lymph node sites during the progression of tumour growth (365). Further analysis of genes involved in this pathway identified 6 additional genes whose expression in normal lung was statistically associated with clinical staging (Table 4). Among these genes, IL6 (interleukin 6) showed the

stronger modulation (Table 23). This gene encodes a cytokine implicated in a wide variety of inflammation-associated disease states. Moreover, high serum IL6 level was found to correlate with tumour invasiveness, size, and grade and with clinical stage and survival in patients with gastric (366), colorectal (367), and breast cancer (368).

TNFSF10 (tumour necrosis factor (ligand) superfamily, member 10), also known as tumour necrosis factor-related apoptosis inducing ligand (TRAIL), is a member of the TNF superfamily of cytokines that induces apoptosis in about 50% of investigated tumour cell lines and play an important role in tumour surveillance (reviewed in (369)). TNFSF10/TRAIL is a key regulator of inflammatory response (370) and its expression has been implicated in asthma (371), and a specific haplotype of this gene is associated with risk of asthma (372)

CCL2 (chemokine (C-C motif) ligand 2), also known as monocyte chemoattractant protein-1 (MCP-1), has been previously demonstrated to increase tumour growth and bone metastasis through its chemotactic activity for monocytes/macrophages and basophiles to tumour sites (373). Monitoring of CCL2 concentration in serum may enable prediction of clinical course of interstitial lung disease (374). CCL2 is also involved in the advanced stage of atherosclerotic cerebro-vascular disease (375) and is associated with poor prognosis in associated small vessel vasculitis (376).

In our study, we found that the expression level of these molecules is just different in the normal tissue from lung cancer patients with different clinical stage and our observation that a signature is associated with clinical stage across heterogeneous population of patients is encouraging and it

could be an important marker of prognosis following further clinical validation.

We further investigated the expression of the 11 modulated genes according to clinical stage in tumour tissue, and, interestingly, all of these genes except for IDH1 were down-regulated in lung ADCA tissue as compared to normal lung (Table 22); such down-regulation in tumour tissue paralleled the decreased expression levels of the same genes, except SELE, in normal lung of stage >I as compared to stage I patients (Table 19). These findings were also confirmed at level of proteins (Fig. 36).

Our findings suggest that clinical staging of lung ADCA patients may be genetically modulated, at least partially, and that a transcriptional profile signature associated with clinical staging is detectable in normal lung tissue of lung ADCA patients. Such a signature may underlie individual genetic predisposition to low or high clinical stage. Characterization of the identified candidate genes whose expression is associated with clinical stage might shed light on the genetic mechanisms underlying individual predisposition to tumour aggressiveness and might define new genetic targets for drugs aimed at countering cancer progression.

Our GWAS and microarray analyses both allowed identification of candidate genes and pathways associated with lung cancer clinical stage. Each of these two different approaches have several advantages and weakness, thus by combining data from the two analyses we could identify a small fraction of genes putatively involved in lung cancer outcome.

To prioritize the discovery of candidate loci associated with lung cancer prognosis we carried out an integration of data from GWA with gene

expression signatures from case-only study. We found 6 SNPs, among those with $P \leq 1 \times 10^{-7}$ in GWA, that map within genes slightly differentially expressed ($P<0.001$, Table 23) in the second microarray experiment. Among these, ADAMTS9 gene encodes a member of the disintegrin and metalloproteinase with thrombospondin motifs protein family. Members of the ADAMTS family have been implicated in the cleavage of proteoglycans, the control of organ shape during development and angiogenesis. In particular ADAMTS9 contributes to the inhibition of angiogenesis in the tumour microenvironment (377). Recently, ADAMTS9 has been characterized as a novel tumour suppressor gene in esophageal squamous cell carcinoma and has been shown to be epigenetically silenced in association with lymph node metastases in nasopharyngeal carcinoma (378). This gene is localized to chromosome 3p14.3-p14.2, an area known to be lost in hereditary renal tumours (379). Interestingly, the SNP rs1799969 on chromosome 19 maps within ICAM1 gene and in LD with the near ICAM4 gene, two genes belonging to the intercellular adhesion molecule protein family. Both are down-regulated in normal tissue from higher clinical stage patients. These genes are candidates for additional studies that could clarify their role and function related to tumour progression.

When we focused on regions where we identified the 54 most associated SNPs with clinical stage (Table 15), we found on the Sentrix Bead Chip HumanRef_8_v2 (Illumina) detectable expression of 18 of 30 known genes (Table 24) in normal lung tissue of our lung ADCA patients, but we found no statistically significant differences in mRNA expression

levels between the clinical stage I and stage >I patients at any of the 18 genes (Table 24), suggesting that the SNP candidacy may rest on non-synonymous variations that are in linkage disequilibrium with the identified SNPs, or on splicing alternative variants rather than alteration of transcript regulation.

In order to test whether these two approaches identified common pathways, DAVID functional annotation tool was used for pathways analyses of GWAS and microarray data. We found that the top GWAS and differentially expressed genes were enriched in cell adhesion molecules focusing in different aspects such as focal adhesion, extracellular matrix and cell adhesion itself. The involvement of cell adhesion system in cancer progression is now well ascertained (380). In fact, integrins play an important role in different aspects of tumourigenesis such as cell proliferation, cell motility, and apoptosis (381), and cadherins were found involved in tumour cell proliferation through cyclins and cyclin-dependent kinases (382). In addition, modulation of cell adhesion was found to be involved in angiogenesis (383) and to play an important role in epithelial-to-mesenchymal transition that is thought to be a key step in malignant transformation (384). This result suggest that functional annotation analyses using candidate genes identified by GWAS and by gene expression profiling can help to refine the identification of candidate genes or pathways associated with a certain phenotype.

CONCLUSIONS

The results showed in the present thesis indicated that genetic constitution plays an important role in lung cancer susceptibility and progression. We suggested and confirmed the relevance of a polygenic model characterized by additive and interchangeable effects of rare alleles in the modulation of individual risk of lung ADCA identifying multiple inherited susceptibility alleles linked to lung cancer. Furthermore, we detected 22 genetic variants that together explained a large individual variation in clinical stage and that were also associated with overall survival, demonstrating that the individual genetic constitution may affect clinical stage of lung cancer patients.

In the second part of this thesis I addressed the critical question of whether a gene expression profile of normal lung tissue can be associated with clinical stage in lung adenocarcinoma (ADCA) patients. The results of such analysis pointed to 11 differentially expressed genes, with FCN3 and TMEM100 showing the best statistical association with clinical stage and the higher modulation. The same FCN3 and TMEM100 genes were also >40-fold down-regulated in lung ADCA tissue as compared to normal tissue. Moreover, analysis of biochemical pathways pointed to a transcriptional signature involving cytokines and cytokine receptors. In addition, combining GWAS and microarray data, we identified cell adhesion as a common biological function and this new approach can help to refine the identification of candidate genes and/or functions involved in tumour development.

These findings provided evidence that clinical stage may be at least partially genetically determined as reflected in the transcriptional profile of normal lung tissue and in the germ line polymorphisms.

The elucidation of the molecular events controlling cancer prognosis and susceptibility could have a great impact on methods for a better prediction of lung cancer outcome and diagnosis and on adequate therapeutic choices. In particular, the identification of the genetic variations and of genes differentially expressed in inherited constitution is essential knowledge concerning tumour initiation and progression in lung ADCA cancers. This should help ultimately to identify new potential target areas for the cancer therapy, design of new efficient drugs to cure cancer with personalized chemotherapeutic and preventive strategies, based on individual genetic constitution.

The identification and subsequent functional characterization of the genetic factors modulating individual risk of lung cancer and/or associated to patients' prognosis represents an important step toward a better understanding of the biological and molecular basis of lung cancer development and progression.

REFERENCES

1. Mawson AR. On not taking the world as you find it-epidemiology in its place. J Clin Epidemiol 2002;55(1):1-4.

2. Last JM, Spasoff RA, Harris SS. A dictionary of epidemiology. 4th ed. Oxford ; New York: Oxford University Press; 2001.

3. Buck C. The challenge of epidemiology : Issues and selected readings. Pan American Sanitary Bureau., editor. Pan American Health Organization; 1988.

4. Detels R. Epidemiology: The foundation of public health. In: Detels R, Holland WW, McEwen J, Omenn GS, editors. Oxford textbook of public health. Vol.2 The methods of public health. 3rd ed. Oxford: Oxford University Press; 1997. p. 501-6.

5. Greenwood M. The application of mathematics to epidemiology. Nature 1916;97(2429):243.

6. Susser M. Epidemiology in the united states after world war II: The evolution of technique. Epidemiol Rev 1985;7:147-77.

7. Lilienfeld DE. "The greening of epidemiology": Sanitary physicians and the london epidemiological society (1830-1870). Bull Hist Med 1978;52(4):503-28.

8. Winkelstein W,Jr. Interface of epidemiology and history: A commentary on past, present, and future. Epidemiol Rev 2000;22(1):2-6.

9. Paneth N, Susser E, Susser M. Origins and early development of the case-control study: Part 2, the case-control study from lane-claypon to 1950. Soz Praventivmed 2002;47(6):359-65.

10. Doll R, Hill AB. Smoking and carcinoma of the lung; preliminary report. Br Med J 1950;2(4682):739-48.

11. Doll R, Hill AB. A study of the aetiology of carcinoma of the lung. Br Med J 1952;2(4797):1271-86.

12. Doll R, Hill AB. Mortality in relation to smoking: Ten years' observations of british doctors. Br Med J 1964;1(5395):1399-410.

13. Rothman KJ. Modern epidemiology. 2nd ed. Greenland S, editor. Philadelphia: Lippincott-Raven; 1998.

14. Berkman LF, Kawachi I. Social epidemiology. Berkman LF and Kawachi I, editors. Oxford: Oxford University Press; 2000.

15. Santos Silva Id, International Agency for Research on Cancer. Cancer epidemiology : Principles and methods. 2nd ed. Lyon: I.A.R.C; 1999.

16. Lagiou P, Adami HO, Trichopoulos D. Causality in cancer epidemiology. Eur J Epidemiol 2005;20(7):565-74.

17. HILL AB. The environment and disease: Association or causation? Proc R Soc Med 1965;58:295-300.

18. Sackett DL. Bias in analytic research. J Chronic Dis 1979;32(1-2):51-63.

19. Weinberg RA. One renegade cell : The quest for the origins of cancer. London: Phoenix; 1999.

20. Jemal A, Siegel R, Xu J, Ward E. Cancer statistics, 2010. CA Cancer J Clin 2010;60(5):277-300.

21. Weinberg RA. The genetic bases of cancer. Arch Surg 1990;125:257-60.

22. Nowell PC. The clonal evolution of tumor cell populations. Science 1976;194:23-8.

23. Ponder BA. Cancer genetics. Nature 2001;411:336-41.

24. Todd R, Wong DT. Oncogenes. Anticancer Res 1999;19:4729-46.

25. Renan MJ. How many mutations are required for tumorigenesis? implications from human cancer data. Mol Carcinog 1993;7(3):139-46.

26. Fernández-Piqueras J, Santos Hernández J. Tumor modifier genes. Rev Oncol 2002;4(7):349-57.

28. Weinberg RA. How cancer arises. Sci Am 1996;275(3):62-70.

29. Balmain A, Gray J, Ponder B. The genetics and genomics of cancer. Nat Genet 2003;33 Suppl:238-44.

30. Kinzler KW, Vogelstein B. Cancer-susceptibility genes. gatekeepers and caretakers. Nature 1997;386:761, 763.

31. Dragani TA. 10 years of mouse cancer modifier loci: Human relevance. Cancer Res 2003;63:3011-8.

32. Pharoah PD, Day NE, Duffy S, Easton DF, Ponder BA. Family history and the risk of breast cancer: A systematic review and meta-analysis. Int J Cancer 1997;71:800-9.

33. Johns LE, Houlston RS. A systematic review and meta-analysis of familial colorectal cancer risk. Am J Gastroenterol 2001;96:2992-3003.

34. Crabtree MD, Tomlinson IP, Hodgson SV, Neale K, Phillips RK, Houlston RS. Explaining variation in familial adenomatous polyposis: Relationship between genotype and phenotype and evidence for modifier genes. Gut 2002;51:420-3.

35. Chung CC, Magalhaes WC, Gonzalez-Bosquet J, Chanock SJ. Genome-wide association studies in cancer--current and future directions. Carcinogenesis 2010;31(1):111-20.

36. Garber JE, Offit K. Hereditary cancer predisposition syndromes. J Clin Oncol 2005;23(2):276-92.

37. Knudson AG,Jr. Mutation and cancer: Statistical study of retinoblastoma. Proc Natl Acad Sci U S A 1971 Apr 1971;68:820-3.

38. Knudson AG. Hereditary cancer: Two hits revisited. J Cancer Res Clin Oncol 1996;122(3):135-40.

39. Weinberg RA. Tumor suppressor genes. Science 1991;254:1138-46.

40. Tomlinson IP, Roylance R, Houlston RS. Two hits revisited again. J Med Genet 2001;38(2):81-5.

41. Cavenee WK, Hansen MF, Scrable HJ, James CD. Loss of genetic information in cancer. Ciba Found Symp 1989;142:79,88; discussion 88-92.

42. Russo A, Zanna I, Tubiolo C, et al. Hereditary common cancers: Molecular and clinical genetics. Anticancer Res 2000;20(6C):4841-51.

43. Easton DF. How many more breast cancer predisposition genes are there? Breast Cancer Res 1999 1999;1:14-7.

44. Dragani TA, Canzian F, Pierotti MA. A polygenic model of inherited predisposition to cancer. FASEB J 1996;10:865-70.

45. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100:57-70.

46. Classon M, Settleman J. Emerging concepts in tumor progression and therapy. Semin Cancer Biol 2000;10:393-7.

47. Nowell PC. Tumor progression: A brief historical perspective. Semin Cancer Biol 2002;12:261-6.

48. Loeb LA. A mutator phenotype in cancer. Cancer Res 2001;61:3230-9.

49. Lengauer C, Kinzler KW, Vogelstein B. Genetic instabilities in human cancers. Nature 1998;396:643-9.

50. Hayflick L. The cell biology of aging. J Invest Dermatol 1979;73(1):8-14.

51. Bryan TM, Reddel RR. Telomere dynamics and telomerase activity in in vitro immortalised human cells. Eur J Cancer 1997;33(5):767-73.

52. Harris CC. P53 tumor suppressor gene: From the basic research laboratory to the clinic--an abridged historical perspective. Carcinogenesis 1996;17(6):1187-98.

53. Yokota J. Tumor progression and metastasis. Carcinogenesis 2000;21:497-503.

54. Moscow JA, Cowan KH. Multidrug resistance. J Natl Cancer Inst 1988;80(1):14-20.

55. Burnet FM. The concept of immunological surveillance. Prog Exp Tumor Res 1970;13:1-27.

56. Loktionov A. Common gene polymorphisms, cancer progression and prognosis. Cancer Lett 2004 May 10 2004;208:1-33.

57. Lauffenburger DA, Horwitz AF. Cell migration: A physically integrated molecular process. Cell 1996;84(3):359-69.

58. Clarke MF, Dick JE, Dirks PB, et al. Cancer stem cells--perspectives on current status and future directions: AACR workshop on cancer stem cells. Cancer Res 2006;66(19):9339-44.

59. Campbell LL, Polyak K. Breast tumor heterogeneity: Cancer stem cells or clonal evolution? Cell Cycle 2007;6(19):2332-8.

60. Vermeulen L, Sprick MR, Kemper K, Stassi G, Medema JP. Cancer stem cells--old concepts, new insights. Cell Death Differ 2008;15(6):947-58.

61. Coussens LM, Werb Z. Inflammation and cancer. Nature 2002;420(6917):860-7.

62. Barcellos-Hoff MH, Ravani SA. Irradiated mammary gland stroma promotes the expression of tumorigenic potential by unirradiated epithelial cells. Cancer Res 2000;60(5):1254-60.

63. Ardlie KG, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. Nat Rev Genet 2002;3:299-309.

64. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273:1516-7.

65. Pharoah PD, Dunning AM, Ponder BA, Easton DF. Association studies for finding cancer-susceptibility genetic variants. Nat Rev Cancer 2004;4(11):850-60.

66. Wright AF, Hastie ND. Complex genetic diseases: Controversy over the croesus code. Genome Biol 2001;2:COMMENT2007.

67. Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994;265:2037-48.

68. Cordell HJ, Todd JA. Multifactorial inheritance in type 1 diabetes. Trends Genet 1995;11:499-504.

69. Lifton RP. Genetic determinants of human hypertension. Proc Natl Acad Sci USA 1995;92:8545-51.

70. Pericak-Vance M, Haines JL. Genetic susceptibility to alzheimer disease. Trends Genet 1995;11:504-8.

71. Weyand CM, Goronzy JJ. Inherited and noninherited risk factors in rheumatoid arthritis. Current Opinion Rheum 1995;7:206-13.

72. Guillausseau PJ, Tielmans D, Virally-Monod M, Assayag M. Diabetes: From phenotypes to genotypes. Diabetes Metab 1997;23:14-21.

73. Poser CM. Notes on the pathogenesis of multiple sclerosis. Clin Neurosci 1994;2(3-4):258-65.

74. Thomson G, Esposito MS. The genetics of complex diseases. Trends Cell Biol 1999 Dec 1999;9:M17-20.

75. Weissman SM. Genetic bases for common polygenic diseases. Proc Natl Acad Sci USA 1995;92:8543-4.

76. Doll R, Peto R. The causes of cancer: Quantitative estimates of avoidable risks of cancer in the united states today. J Natl Cancer Inst 1981;66:1191-308.

77. Emery J, Lucassen A, Murphy M. Common hereditary cancers and implications for primary care. Lancet 2001;358(9275):56-63.

78. Ames BN, Gold LS, Willett WC. The causes and prevention of cancer. Proc Natl Acad Sci USA 1995;92:5258-65.

79. Peto J. Cancer epidemiology in the last century and the next decade. Nature 2001;411:390-5.

80. Rushton L. How much does the environment contribute to cancer? Occup Environ Med 2003;60.

81. Colditz GA, Sellers TA, Trapido E. Epidemiology - identifying the causes and preventability of cancer? Nat Rev Cancer 2006;6(1):75-83.

82. McCullough ML, Giovannucci EL. Diet and cancer prevention. Oncogene 2004;23(38):6349-64.

83. Key TJ, Allen NE, Spencer EA, Travis RC. The effect of diet on risk of cancer. Lancet 2002;360(9336):861-8.

84. Giese B. Long-distance charge transport in DNA: The hopping mechanism. Acc Chem Res 2000;33(9):631-6.

85. Chen YC, Hunter DJ. Molecular epidemiology of cancer. CA Cancer J Clin 2005;55(1):45,54; quiz 57.

86. Wogan GN, Hecht SS, Felton JS, Conney AH, Loeb LA. Environmental and chemical carcinogenesis. Semin Cancer Biol 2004;14(6):473-86.

87. Statistics for 2006. cancer facts & figures 2006.[homepage on the Internet].

88. Janssen-Heijnen ML, Coebergh JW. The changing epidemiology of lung cancer in europe. Lung Cancer 2003;41:245-58.

89. Patz EF,Jr., Rossi S, Harpole DH,Jr., Herndon JE, Goodman PC. Correlation of tumor size and survival in patients with stage IA non-small cell lung cancer. Chest 2000 Jun 2000;117:1568-71.

90. Hammerschmidt S, Wirtz H. Lung cancer: Current diagnosis and treatment. Dtsch Arztebl Int 2009;106(49):809,18; quiz 819-20.

91. Wahbah M, Boroumand N, Castro C, El-Zeky F, Eltorky M. Changing trends in the distribution of the histologic types of lung cancer: A review of 4,439 cases. Ann Diagn Pathol 2007;11(2):89-96.

92. Hinson JA,Jr, Perry MC. Small cell lung cancer. CA Cancer J Clin 1993;43(4):216-25.

93. Charloux A, Quoix E, Wolkove N, Small D, Pauli G, Kreisman H. The increasing incidence of lung adenocarcinoma: Reality or artefact? A review of the epidemiology of lung adenocarcinoma. Int J Epidem 1997;26:14-23.

94. Hoffmann D, Djordjevic MV, Hoffmann I. The changing cigarette. Prev Med 1997;26(4):427-34.

95. Maggiore C, Mule A, Fadda G, et al. Histological classification of lung cancer. Rays 2004;29(4):353-5.

96. Carney DN, De Leij L. Lung cancer biology. Semin Oncol 1988;15:199-214.

97. Sihoe AD, Yim AP. Lung cancer staging. J Surg Res 2004;117:92-106.

98. Osaki T, Nagashima A, Yoshimatsu T, Tashima Y, Yasumoto K. Survival and characteristics of lymph node involvement in patients with N1 non-small cell lung cancer. Lung Cancer 2004;43:151-7.

99. Mountain CF. Revisions in the international system for staging lung cancer. Chest 1997;111(6):1710-7.

100. Kitamura H, Yazawa T, Okudela K, Shimoyamada H, Sato H. Molecular and genetic pathogenesis of lung cancer: Differences between small-cell and non-small-cell carcinomas. The Open Pathology Journal 2008;2:106-114.

101. Kiyohara C, Otsu A, Shirakawa T, Fukuda S, Hopkin JM. Genetic polymorphisms and lung cancer susceptibility: A review. Lung Cancer 2002;37(3):241-56.

102. Sekido Y, Fong KM, Minna JD. Molecular genetics of lung cancer. Annu Rev Med 2003;54:73-87.

103. Wynder EL, Graham EA. Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma; a study of 684 proved cases. J Am Med Assoc 1950;143(4):329-36.

104. Levin ML, Goldstein H, Gerhardt PR. Cancer and tobacco smoking; a preliminary report. J Am Med Assoc 1950;143(4):336-8.

105. Khuder SA, Mutgi AB. Effect of smoking cessation on major histologic types of lung cancer. Chest 2001;120(5):1577-83.

106. Doll R, Peto R. Cigarette smoking and bronchial carcinoma: Dose and time relationships among regular smokers and lifelong non-smokers. J Epidemiol Community Health 1978;32:303-13.

107. Peto R, Chen ZM, Boreham J. Tobacco--the growing epidemic. Nat Med 1999;5:15-7.

108. Hecht SS. Tobacco smoke carcinogens and lung cancer. J Natl Cancer Inst 1999;91:1194-210.

109. Alberg AJ, Samet JM. Epidemiology of lung cancer. Chest 2003;123:21S-49S.

110. International Agency for Research on Cancer. Cancer : Causes, occurrence and control. Tomatis L, editor. International Agency for Research on Cancer; 1990.

111. Samet JM. The epidemiology of lung cancer. Chest 1993;103(1 Suppl):20S-9S.

112. Kabat GC. Recent developments in the epidemiology of lung cancer. Semin Surg Oncol 1993;9(2):73-9.

113. Mattson ME, Pollack ES, Cullen JW. What are the odds that smoking will kill you? Am J Public Health 1987;77(4):425-31.

114. Peto R, Darby S, Deo H, Silcocks P, Whitley E, Doll R. Smoking, smoking cessation, and lung cancer in the UK since 1950: Combination of national statistics with two case-control studies. Br Med J 2000;321:323-9.

115. Gariboldi M, Manenti G, Canzian F, et al. A major susceptibility locus to murine lung carcinogenesis maps on chromosome 6. Nature Genet 1993;3:132-6.

116. Zhang Z, Futamura M, Vikis HG, et al. Positional cloning of the major quantitative trait locus underlying lung tumor susceptibility in mice. Proc Natl Acad Sci U S A 2003;100:12642-7.

117. Manenti G, Galbiati F, Giannì Barrera R, Pettinicchio A, Acevedo A, Dragani TA. Haplotype sharing suggests that a genomic segment containing six genes accounts for the pulmonary adenoma susceptibility 1 (Pas1) locus activity in mice. Oncogene 2004;23:4495-504.

118. Manenti G, De Gregorio L, Pilotti S, et al. Association of chromosome 12p genetic polymorphisms with lung adenocarcinoma risk and prognosis. Carcinogenesis 1997;18(10):1917-20.

119. Dragani TA, Hirohashi S, Juji T, et al. Population-based mapping of pulmonary adenoma susceptibility 1 (PAS1) locus. Cancer Res 2000;60:5017-20.

120. Manenti G, Gariboldi M, Elango R, et al. Genetic mapping of a pulmonary adenoma resistance locus (Par1) in mouse. Nature Genet 1996;12:455-7.

121. Manenti G, Gariboldi M, Fiorino A, Zanesi N, Pierotti MA, Dragani TA. Genetic mapping of lung cancer modifier loci specifically affecting tumor initiation and progression. Cancer Res 1997;57:4164-6.

122. Tripodis N, Hart AA, Fijneman RJ, Demant P. Complexity of lung cancer modifiers: Mapping of thirty genes and twenty-five interactions in half of the mouse genome. J Natl Cancer Inst 2001;93:1484-91.

123. Tokuhata GK, Lillienfeld AM. Familial aggregation of lung cancer in humans. J Natl Cancer Inst 1963;30:289-312.

124. Goldgar DE, Easton DF, Cannon-Albright LA, Skolnick MH. Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. J Natl Cancer Inst 1994;86:1600-8.

125. Bailey-Wilson JE, Amos CI, Pinney SM, et al. A major lung cancer susceptibility locus maps to chromosome 6q23-25. Am J Hum Genet 2004;75:460-74.

126. You M, Wang D, Liu P, et al. Fine mapping of chromosome 6q23-25 region in familial lung cancer families reveals RGS17 as a likely candidate gene. Clin Cancer Res 2009;15(8):2666-74.

127. Truong T, Sauter W, McKay JD, et al. International lung cancer consortium: Coordinated association study of 10 potential lung cancer susceptibility variants. Carcinogenesis 2010;31(4):625-33.

128. Neddermann P, Gallinari P, Lettieri P, et al. Cloning and expression of human G/T mismatch-specific thymine DNA glycosylase. J Biol Chem 1996;in press.

129. Amos CI, Wu X, Broderick P, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. Nat Genet 2008;40:616-22.

130. Hung RJ, McKay JD, Gaborieau V, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 2008;452:633-7.

131. Thorgeirsson TE, Geller F, Sulem P, et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature 2008;452:638-42.

132. Wang Y, Broderick P, Webb E, et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. Nat Genet 2008;40(12):1407-9.

133. Wang Y, Broderick P, Matakidou A, Eisen T, Houlston RS. Role of 5p15.33 (TERT-CLPTM1L), 6p21.33 and 15q25.1 (CHRNA5-CHRNA3) variation and lung cancer risk in never-smokers. Carcinogenesis 2010;31(2):234-8.

134. McKay JD, Hung RJ, Gaborieau V, et al. Lung cancer susceptibility locus at 5p15.33. Nat Genet 2008;40(12):1404-6.

135. Truong T, Hung RJ, Amos CI, et al. Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: A pooled analysis from the international lung cancer consortium. J Natl Cancer Inst 2010;102(13):959-71.

136. Li Y, Sheu CC, Ye Y, et al. Genetic variants and risk of lung cancer in never smokers: A genome-wide association study. Lancet Oncol 2010;11(4):321-30.

137. Savas S, Liu G. Genetic variations as cancer prognostic markers: Review and update. Hum Mutat 2009;30(10):1369-77.

138. Savas S, Liu G. Studying genetic variations in cancer prognosis (and risk): A primer for clinicians. Oncologist 2009;14(7):657-66.

139. Spinola M, Conti B, Ravagnani F, et al. A new polymorphism (Ser362Thr) of the L-myc gene is not associated with lung adenocarcinoma risk and prognosis. Eur J Cancer Prev 2004;13:87-9.

140. Spinola M, Leoni V, Pignatiello C, et al. Functional FGFR4 Gly388Arg polymorphism predicts prognosis in lung adenocarcinoma patients. J Clin Oncol 2005;23:7307-11.

141. Spinola M, Leoni VP, Tanuma J, et al. FGFR4 Gly388Arg polymorphism and prognosis of breast and colorectal cancer. Oncol Rep 2005;14:415-9.

142. Heist RS, Zhai R, Liu G, et al. VEGF polymorphisms and survival in early-stage non-small-cell lung cancer. J Clin Oncol 2008;26(6):856-62.

143. Manenti G, Nomoto T, De Gregorio L, et al. Predisposition to lung tumorigenesis. Toxicol Lett 2000;112-113:257-63.

144. Campling BG, el-Deiry WS. Clinical implications of p53 mutations in lung cancer. Methods Mol Med 2003;75:53-77.

145. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001 Feb 15 2001;409:860-921.

146. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. Science 2001;291:1304-51.

147. The International HC. The international HapMap project. Nature 2003;426:789-96.

148. International HapMap Consortium. A haplotype map of the human genome. Nature 2005;437(7063):1299-320.

149. Brookes AJ. The essence of SNPs. Gene 1999 1999;234:177-86.

150. Collins DW, Jukes TH. Rates of transition and transversion in coding sequences since the human-rodent divergence. Genomics 1994 1994;20:386-96.

151. Picoult-Newberg L, Ideker TE, Pohl MG, et al. Mining SNPs from EST databases. Genome Res 1999;9:167-74.

152. Cooper DN, Krawczak M. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum Genet 1989;83:181-8.

153. Wang DG. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 1998;280:1077-82.

154. Kan YW, Dozy AM. Polymorphism of DNA sequence adjacent to human beta-globin structural gene: Relationship to sickle mutation. Proc Natl Acad Sci U S A 1978;75(11):5631-5.

155. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 1980;32(3):314-31.

156. Roses AD. Pharmacogenetics and pharmacogenomics in the discovery and development of medicines. Novartis Found Symp 2000;229:63,6; discussion 66-70.

157. Roses AD. Pharmacogenetics and the practice of medicine. Nature 2000;405(6788):857-65.

158. Kruglyak L, Nickerson DA. Variation is the spice of life. Nat Genet 2001;27(3):234-6.

159. Hinds DA, Stuve LL, Nilsen GB, et al. Whole-genome patterns of common DNA variation in three human populations. Science 2005;307(5712):1072-9.

160. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. Nat Genet 2001;29:229-32.

161. Risch NJ. Searching for genetic determinants in the new millennium. Nature 2000;405(6788):847-56.

162. Sachidanandam R, Weissman D, Schmidt SC, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001 2001;409:928-33.

163. Reich DE, Cargill M, Bolk S, et al. Linkage disequilibrium in the human genome. Nature 2001;411(6834):199-204.

164. Wall JD, Pritchard JK. Haplotype blocks and linkage disequilibrium in the human genome. Nat Rev Genet 2003 Aug 2003;4:587-97.

165. Clark AG, Weiss KM, Nickerson DA, et al. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 1998;63(2):595-612.

166. Dunning AM, Durocher F, Healey CS, et al. The extent of linkage disequilibrium in four populations with distinct demographic histories. Am J Hum Genet 2000;67:1544-54.

167. Ardlie K, Liu-Cordero SN, Eberle MA, et al. Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. Am J Hum Genet 2001;69:582-9.

168. Collins A, Lonjou C, Morton NE. Genetic epidemiology of single-nucleotide polymorphisms. Proc Natl Acad Sci USA 1999;96:15173-7.

169. Abecasis GR, Noguchi E, Heinzmann A, et al. Extent and distribution of linkage disequilibrium in three genomic regions. Am J Hum Genet 2001;68:191-7.

170. Stephens JC, Schneider JA, Tanguay DA, et al. Haplotype variation and linkage disequilibrium in 313 human genes. Science 2001;293(5529):489-93.

171. Parsch J, Meiklejohn CD, Hartl DL. Patterns of DNA sequence variation suggest the recent action of positive selection in the janus-ocnus region of drosophila simulans. Genetics 2001;159(2):647-57.

172. Cannon GB. The effects of natural selection on linkage disequilibrium and relative fitness in experimental populations of drosophila melanogaster. Genetics 1963;48:1201-16.

173. Lewontin RC. The interaction of selection and linkage. I. general considerations; heterotic models. Genetics 1964;49(1):49-67.

174. Chakraborty R, Weiss KM. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. Proc Natl Acad Sci U S A 1988;85(23):9119-23.

175. Stephens JC, Briscoe D, O'Brien SJ. Mapping by admixture linkage disequilibrium in human populations: Limits and guidelines. Am J Hum Genet 1994;55(4):809-24.

176. Pfaff CL, Parra EJ, Bonilla C, et al. Population structure in admixed populations: Effect of admixture dynamics on the pattern of linkage disequilibrium. Am J Hum Genet 2001;68(1):198-207.

177. Wilson JF, Goldstein DB. Consistent long-range linkage disequilibrium generated by admixture in a bantu-semitic hybrid population. Am J Hum Genet 2000;67(4):926-35.

178. Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 2001 Oct 2001;29:217-22.

179. Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am J Hum Genet 2001;69:831-43.

180. Quintana PJ, Neuwirth EA, Grosovsky AJ. Interchromosomal gene conversion at an endogenous human cell locus. Genetics 2001;158(2):757-67.

181. Kruglyak L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genet 1999;22:139-44.

182. Gordon D, Simonic I, Ott J. Significant evidence for linkage disequilibrium over a 5-cM region among afrikaners. Genomics 2000;66:87-92.

183. Tenesa A, Wright AF, Knott SA, et al. Extent of linkage disequilibrium in a sardinian sub-isolate: Sampling and methodological considerations. Hum Mol Genet 2004;13:25-33.

184. International HapMap Consortium, Frazer KA, Ballinger DG, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449(7164):851-61.

185. International HapMap Consortium. The International HapMap Project. Nature 2003;426(6968):789-96.

186. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. The structure of haplotype blocks in the human genome. Science 2002;296(5576):2225-9.

187. Breslow NE. Statistical methods in cancer research. vol. 2. the design and analysis of cohort studies. Day NE, editor. Lyon: International Agency for Research on Cancer; 1987.

188. Hennekens CH. Epidemiology in medicine. Buring JE and Mayrent SL, editors. Little, Brown; 1987.

189. Breslow NE. Statistical methods in cancer research. vol. 1 the analysis of case-control studies. Day NE, editor. Lyon: International Agency for Research on Cancer; 1980.

190. Wacholder S, Silverman DT, McLaughlin JK, Mandel JS. Selection of controls in case-control studies. II. types of controls. Am J Epidemiol 1992;135(9):1029-41.

191. Daly AK, Day CP. Candidate gene case-control association studies: Advantages and potential pitfalls. Br J Clin Pharmacol 2001;52(5):489-99.

192. Laird NM, Lange C. Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet 2006;7(5):385-94.

193. Weiss KM, Terwilliger JD. How many diseases does it take to map a gene with SNPs? Nat Genet 2000 Oct 2000;26:151-7.

194. NCI-NHGRI Working Group on Replication in Association Studies, Chanock SJ, Manolio T, et al. Replicating genotype-phenotype associations. Nature 2007;447(7145):655-60.

195. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 2003;33:177-82.

196. Whitley E, Ball J. Statistics review 4: Sample size calculations. Crit Care 2002;6(4):335-41.

197. Taioli E, Bonassi S. Methodological issues in pooled analysis of biomarker studies. Mutat Res 2002;512:85-92.

198. d'Errico A, Malats N, Vineis P, Boffetta P. Review of studies of selected metabolic polymorphisms and cancer. IARC Sci Publ 1999:323-93.

199. Houlston RS. CYP1A1 polymorphisms and lung cancer risk: A meta-analysis. Pharmacogenetics 2000;10:105-14.

200. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet 2003 Feb 15 2003;361:598-604.

201. Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: Quantification of bias. J Natl Cancer Inst 2000;92:1151-8.

202. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 1999;65:220-8.

203. Strom BL. Pharmacoepidemiology. 3rd ed. Strom BL, editor. Chichester: Wiley; 2000.

204. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease. Nat Genet 2003;33 Suppl:228-37.

205. Moher D, Schulz KF, Altman DG, CONSORT Group (Consolidated Standards of Reporting Trials). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. J Am Podiatr Med Assoc 2001;91(8):437-42.

206. von Elm E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. Lancet 2007;370(9596):1453-7.

207. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 2005;6(2):95-108.

208. National Institutes of Health. Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS). Federal Regist 2007;72(166):49290-7. Available from: http://www.grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html.

209. McCarthy MI, Abecasis GR, Cardon LR, et al. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. Nat Rev Genet 2008;9:356-69.

210. Stadler ZK, Gallagher DJ, Thom P, Offit K. Genome-wide association studies of cancer: Principles and potential utility. Oncology (Williston Park) 2010;24(7):629-37.

211. Rioux JD, Xavier RJ, Taylor KD, et al. Genome-wide association study identifies new susceptibility loci for crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 2007;39(5):596-604.

212. Broderick P, Wang Y, Vijayakrishnan J, et al. Deciphering the impact of common genetic variation on lung cancer risk: A genome-wide association study. Cancer Res 2009;69(16):6633-41.

213. Landi MT, Chatterjee N, Yu K, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. Am J Hum Genet 2009;85(5):679-91.

214. Norton N, Williams NM, Williams HJ, et al. Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. Hum Genet 2002;110(5):471-8.

215. Bansal A, van den BD, Kammerer S, et al. Association testing by DNA pooling: An effective initial screen. Proc Natl Acad Sci U S A 2002;99:16871-4.

216. Myles S, Davison D, Barrett J, Stoneking M, Timpson N. Worldwide population differentiation at disease-associated SNPs. BMC Med Genomics 2008;1:22.

217. Amundadottir LT, Sulem P, Gudmundsson J, et al. A common variant associated with prostate cancer in european and african populations. Nat Genet 2006;38(6):652-8.

218. Freedman ML, Haiman CA, Patterson N, et al. Admixture mapping identifies 8q24 as a prostate cancer risk locus in african-american men. Proc Natl Acad Sci U S A 2006;103(38):14068-73.

219. Scott LJ, Mohlke KL, Bonnycastle LL, et al. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. Science 2007;316(5829):1341-5.

220. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet 2008;9(4):255-66.

221. Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: Theoretical and practical concerns. Nat Rev Genet 2005;6(2):109-18.

222. Ioannidis JP, Trikalinos TA, Khoury MJ. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. Am J Epidemiol 2006;164(7):609-14.

223. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-75.

224. Xu J, Wiesch DG, Meyers DA. Genetics of complex human diseases: Genome screening, association studies and fine mapping. Clin Exp Allergy 1998;28 Suppl 5:1,5; discussion 26-8.

225. Kallberg H, Alfredsson L, Feychting M, Ahlbom A. Don't split your data. Eur J Epidemiol 2010;25(5):283-4.

226. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 2006;38:209-13.

227. Cooper DN. Functional intronic polymorphisms: Buried treasure awaiting discovery within our genes. Hum Genomics 2010;4(5):284-8.

228. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 2007;39(10):1181-6.

229. Rosenberg D, Handler A. Descriptive epidemiology and statistical estimation. In: Handler A, Monahan C, editors. Analytic Methods in Maternal and Child Health. Vienna: Margaret M. Maloney; 1998.

230. Whitley E, Ball J. Statistics review 2: Samples and populations. Crit Care 2002;6(2):143-8.

231. Clayton D. Statistical models in epidemiology. H, editor. Oxford University Press; 1993.

232. Parkin DM, Bray FI. International patterns of cancer incidence and mortality. In: Schottenfeld D, Fraumeni JF, editors. Cancer epidemiology and prevention. 3rd ed. Oxford: Oxford University Press; 2006. p. 1392.

233. Boyle P, Parkin DM. Statistical methods for registries. In: Jensen OM, Parkin DM, MacLennan R, Muir CS, Skeet RG, editors. Cancer registration : principles and methods. Lyon: IARC; 1991.

234. International Agency for Research on Cancer., International Association of Cancer Registries. Cancer incidence in five continents. Waterhouse JAH and Davis W,Ph.D., editors. Lyon; London: IARC; Distributed by H.M.S.O.; 1976.

235. Common Services Agency for the Scottish Health Service. Scottish Cancer Intelligence Unit., National Health Service in Scotland. Information & Statistics Division. Cancer registration statistics scotland 1986-1995. Harris V, Sandridge AL, Black RJ, Brewster DH, Gould A, editors. Edinburgh: National Health Service in Scotland, Information & Statistics Division; 1998.

236. Pharoah PD. Genetic susceptibility, predicting risk and preventing cancer. Recent Results Cancer Res 2003;163:7,18; discussion 264-6.

237. Bewick V, Cheek L, Ball J. Statistics review 11: Assessing risk. Crit Care 2004;8(4):287-91.

238. Spruance SL, Reid JE, Grace M, Samore M. Hazard ratio in clinical trials. Antimicrob Agents Chemother 2004;48(8):2787-92.

239. Kirkwood BR. Essential medical statistics. 2nd ed. Sterne JAC and Kirkwood,Betty R..Essentials of medical statistics, editors. Malden, Mass. ; Oxford: Blackwell Science; 2003.

240. Whitley E, Ball J. Statistics review 3: Hypothesis testing and P values. Crit Care 2002;6(3):222-5.

241. Wilcox RR. Fundamentals of modern statistical methods : Substantially improving power and accuracy. New York ; London: Springer; 2001.

242. Whitley E, Ball J. Statistics review 6: Nonparametric methods. Crit Care 2002;6(6):509-13.

243. Larson MG. Analysis of variance. Circulation 2008;117(1):115-21.

244. Bewick V, Cheek L, Ball J. Statistics review 9: One-way analysis of variance. Crit Care 2004;8(2):130-6.

245. Zar JH. Biostatistical analysis. 4th ed. Upper Saddler River, N.J.; London: Prentice Hall; Prentice Hall International (UK); 1999.

246. Krustal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc 1952;47:583-621.

247. Bewick V, Cheek L, Ball J. Statistics review 8: Qualitative data - tests of association. Crit Care 2004;8(1):46-53.

248. Bewick V, Cheek L, Ball J. Statistics review 7: Correlation and regression. Crit Care 2003;7(6):451-9.

249. Bewick V, Cheek L, Ball J. Statistics review 14: Logistic regression. Crit Care 2005;9(1):112-8.

250. Bland JM, Altman DG. Multiple significance tests: The bonferroni method. BMJ 1995;310(6973):170.

251. Duggal P, Gillanders EM, Holmes TN, Bailey-Wilson JE. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. BMC Genomics 2008;9:516.

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

253. Ioannidis JP. Non-replication and inconsistency in the genome-wide association setting. Hum Hered 2007;64(4):203-13.

254. Dudbridge F. A note on permutation tests in multistage association scans. Am J Hum Genet 2006;78(6):1094,5; author reply 1096.

255. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 2007;39(7):906-13.

256. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 1993;52:506-16.

257. Scott WK, Pericak-Vance MA, Haines JL. Genetic analysis of complex diseases. Science 1997;275(5304):1327; author reply 1329-30.

258. Spielman RS, Ewens WJ. A sibship test for linkage in the presence of association: The sib transmission/disequilibrium test. Am J Hum Genet 1998;62:450-8.

259. Lander E, Kruglyak L. Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. Nature Genet 1995;11:241-7.

260. Bewick V, Cheek L, Ball J. Statistics review 12: Survival analysis. Crit Care 2004;8(5):389-94.

261. Lee ET. Statistical methods for survival data analysis. 3rd ed. Wang JW, editor. New Jersey ; Great Britain: Wiley-Interscience; 2003.

262. Ahlbom A, Lichtenstein P, Malmstrom H, Feychting M, Hemminki K, Pedersen NL. Cancer in twins: Genetic and nongenetic familial risk factors. J Natl Cancer Inst 1997;89:287-93.

263. Sellers TA, Potter JD, Bailey-Wilson JE, Rich SS, Rothschild H, Eltson RC. Lung cancer detection and prevention: Evidence for an interaction between smoking and genetic predisposition. Cancer Res 1992;52:2694s-7s.

264. Sellers TA, Bailey-Wilson JE, Elston RC, Wilson AE, Rothschild H. Evidence for mendelian inheritance in the pathogenesis of lung cancer. J Natl Cancer Inst 1990;82:1272-9.

265. Yang P, Schwartz AG, McAllister AE, Swanson GM, Aston CE. Lung cancer risk in families of nonsmoking probands: Heterogeneity by age at diagnosis. Genet Epidemiol 1999;17:253-73.

266. Gauderman WJ, Morrison JL. Evidence for age-specific genetic relative risks in lung cancer. Am J Epidemiol 2000;151:41-9.

267. Zhang Z, Wang Y, Herzog CR, et al. A strong candidate gene for the Papg1 locus on mouse chromosome 4 affecting lung tumor progression. Oncogene 2002;21:5960-6.

268. Sankila R, Aaltonen LA, Jarvinen HJ, Mecklin JP. Better survival rates in patients with MLH1-associated hereditary colorectal cancer. Gastroenterology 1996 Mar 1996;110:682-7.

269. Aarnio M, Mustonen H, Mecklin JP, Jarvinen HJ. Prognosis of colorectal cancer varies in different high-risk conditions. Ann Med 1998;30:75-80.

270. Meaburn E, Butcher LM, Schalkwyk LC, Plomin R. Genotyping pooled DNA using 100K SNP microarrays: A step towards genomewide association scans. Nucleic Acids Res 2006;34:e27.

271. Kirov G, Nikolov I, Georgieva L, Moskvina V, Owen MJ, O'donovan MC. Pooled DNA genotyping on affymetrix SNP genotyping arrays. BMC Genomics 2006;7:27.

272. Pastorino U. Does screening for stage I lung cancer improve survival in a high-risk population? Nat Clin Pract Oncol 2007;4:218-9.

273. Spira A, Beane J, Shah V, et al. Effects of cigarette smoke on the human airway epithelial cell transcriptome. Proc Natl Acad Sci U S A 2004;101:10143-8.

274. Steemers FJ, Gunderson KL. Whole genome genotyping technologies on the BeadArray platform. Biotechnol J 2007;2:41-9.

275. Illumina I. Illumina SNP genotyping. infinium II assay workflow. 2006 Dec 07, 2006. Report No.: Pub. No. 370-2006-027.

276. Lavebratt C, Sengul S. Single nucleotide polymorphism (SNP) allele frequency estimation in DNA pools using pyrosequencing™ Nature Protocols 2007;1:2573-82.

277. Nyren P. Apyrase immobilized on paramagnetic beads used to improve detection limits in bioluminometric ATP monitoring. J Biolumin Chemilumin 1994;9(1):29-34.

278. Oeth P, Beaulieu M, Park C, et al. iPLEX™ assay: Increased plexing efficiency and flexibility for MassARRAY system through single base primer extension with mass-modified terminators. 2005 April 28, 2005. Report No.: Doc. No. 8876-006.

279. Weir BS. Genetic data analysis 2: Methods for discrete population genetic data. Sunderlands, MA: Sinauer Associates, Inc.; 1996.

280. Kaplan EL, Meier P. Non-parametric estimation from incomplete observations. J Am Stat Assoc 1958;3:457-81.

281. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. J R Stat Soc Ser A Stat Soc 1972;135:185-206.

282. Wright GW, Simon RM. A random variance model for detection of differential gene expression in small microarray experiments. Bioinformatics 2003;19:2448-55.

283. Carter KW, McCaskie PA, Palmer LJ. JLIN: A java based linkage disequilibrium plotter. BMC Bioinformatics 2006;7:60.

284. Sham P, Bader JS, Craig I, O'Donovan M, Owen M. DNA pooling: A tool for large-scale association studies. Nat Rev Genet 2002;3(11):862-71.

285. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-delta delta C(T)) method. Methods 2001;25(4):402-8.

286. Efron B. Bootstrap methods: Another look at the jackknife. Ann Statist 1979;7(1):1-26.

287. Dennis G,Jr., Sherman BT, Hosack DA, et al. DAVID: Database for annotation, visualization, and integrated discovery. Genome Biol 2003;4:3.

288. Bowden J, Dudbridge F. Unbiased estimation of odds ratios: Combining genomewide association scans with replication studies. Genet Epidemiol 2009;33(5):406-18.

289. Galvan A, Falvella FS, Spinola M, et al. A polygenic model with common variants may predict lung adenocarcinoma risk in humans. Int J Cancer 2008;123:2327-30.

290. Cho S, Kim K, Kim YJ, et al. Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. Ann Hum Genet 2010;74(5):416-28.

291. Hunter DJ, Riboli E, Haiman CA, et al. A candidate gene approach to searching for low-penetrance breast and prostate cancer genes. Nat Rev Cancer 2005;5:977-85.

292. Spinola M, Meyer P, Kammerer S, et al. Association of the PDCD5 locus with lung cancer risk and prognosis in smokers. J Clin Oncol 2006;24:1672-8.

293. Spinola M, Leoni VP, Galvan A, et al. Genome-wide single nucleotide polymorphism analysis of lung cancer risk detects the KLF6 gene. Cancer Lett 2007;251:311-6.

294. Dragani TA, Manenti G, Pierotti MA. Polygenic inheritance of predisposition to lung cancer. Ann Ist Sup Sanita' 1996;32(1):145-50.

295. Crawford NP, Hunter KW. New perspectives on hereditary influences in metastatic progression. Trends Genet 2006;22(10):555-61.

296. Wingo PA, Ries LA, Giovino GA, et al. Annual report to the nation on the status of cancer, 1973-1996, with a special section on lung cancer and tobacco smoking. J Natl Cancer Inst 1999;91:675-90.

297. Travis WD, Travis LB, Devesa SS. Lung cancer. Cancer 1995 Jan 1 1995;75:191-202.

298. Su JL, Yang CY, Shih JY, et al. Knockdown of contactin-1 expression suppresses invasion and metastasis of lung adenocarcinoma. Cancer Res 2006;66:2553-61.

299. Sun CX, Robb VA, Gutmann DH. Protein 4.1 tumor suppressors: Getting a FERM grip on growth regulation. J Cell Sci 2002;115:3991-4000.

300. Manenti G, Dragani TA. Pas1 haplotype-dependent genetic predisposition to lung tumorigenesis in rodents: A meta-analysis. Carcinogenesis 2005;26:875-82.

301. Manenti G, Acevedo A, Galbiati F, Giannì Barrera R, Noci S, Dragani TA. Cancer modifier alleles inhibiting lung tumorigenesis are common in mouse inbred strains. Int J Cancer 2002;99:555-9.

302. Pastorino U, Bellomi M, Landoni C, et al. Early lung-cancer detection with spiral CT and positron emission tomography in heavy smokers: 2-year results. Lancet 2003;362:593-7.

303. Tsakiris I, Soos G, Nemes Z, et al. The presence of carboxypeptidase-M in tumour cells signifies epidermal growth factor receptor expression in

lung adenocarcinomas: The coexistence predicts a poor prognosis regardless of EGFR levels. J Cancer Res Clin Oncol 2008;134:439-51.

304. Kubo T, Yamamoto H, Ichimura K, et al. DNA methylation in small lung adenocarcinoma with bronchioloalveolar carcinoma components. Lung Cancer 2009;65(3):328-32.

305. Ioannidis JP. Commentary: Grading the credibility of molecular evidence for complex diseases. Int J Epidemiol 2006;35:572-7.

306. Galvan A, Vorraro F, Cabrera WH, et al. Genetic heterogeneity of inflammatory response and skin tumorigenesis in phenotypically selected mouse lines. Cancer Lett 2010.

307. Hunter DJ, Kraft P, Jacobs KB, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 2007;39(7):870-4.

308. Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA. Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet 2002;31:33-6.

309. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from sweden, denmark, and finland. N Engl J Med 2000;343:78-85.

310. Falvella FS, Galvan A, Frullanti E, et al. Transcription deregulation at the 15q25 locus in association with lung adenocarcinoma risk. Clin Cancer Res 2009;15:1837-42.

311. Falvella FS, Galvan A, Frullanti E, Dragani TA. Re: Variants weakly correlated with CHRNA5 D398N polymorphism should be considered in transcriptional deregulation at the 15q25 locus associated with lung cancer risk. Clin.Cancer Res. 2009; 15:5599.

312. Galvan A, Dragani TA. Nicotine dependence may link the 15q25 locus to lung cancer risk. Carcinogenesis 2009.

313. Matakidou A, Eisen T, Houlston RS. Systematic review of the relationship between family history and lung cancer risk. Br J Cancer 2005;93:825-33.

314. Sun S, Schiller JH, Gazdar AF. Lung cancer in never smokers--a different disease. Nat Rev Cancer 2007;7(10):778-90.

315. Foulds L. Tumor progression. Cancer Res 1957;17:355-6.

316. Bernards R, Weinberg RA. A progression puzzle. Nature 2002;418(6900):823.

317. Sidransky D. Emerging molecular markers of cancer. Nat Rev Cancer 2002;2:210-9.

318. Brundage MD, Davies D, Mackillop WJ. Prognostic factors in non-small cell lung cancer: A decade of progress. Chest 2002;122(3):1037-57.

319. Edge SB, Compton CC. The american joint committee on cancer: The 7th edition of the AJCC cancer staging manual and the future of TNM. Ann Surg Oncol 2010;17(6):1471-4.

320. Ohshima J, Haruta M, Arai Y, et al. Two candidate tumor suppressor genes, MEOX2 and SOSTDC1, identified in a 7p21 homozygous deletion region in a wilms tumor. Genes Chromosomes Cancer 2009;48(12):1037-50.

321. Cortese R, Hartmann O, Berlin K, Eckhardt F. Correlative gene expression and DNA methylation profiling in lung development nominate new biomarkers in lung cancer. Int J Biochem Cell Biol 2008;40(8):1494-508.

322. Gabrielson E. Worldwide trends in lung cancer pathology. Respirology 2006;11(5):533-8.

323. Galvan A, Falvella FS, Frullanti E, et al. Genome-wide association study in discordant sibships identifies multiple inherited susceptibility alleles linked to lung cancer. Carcinogenesis 2010;31(3):462-5.

324. Ansari KI, Mandal SS. Mixed lineage leukemia: Roles in gene expression, hormone signaling and mRNA processing. FEBS J 2010;277(8):1790-804.

325. Shah N, Sukumar S. The hox genes and their roles in oncogenesis. Nat Rev Cancer 2010;10(5):361-71.

326. Heuser M, Yap DB, Leung M, et al. Loss of MLL5 results in pleiotropic hematopoietic defects, reduced neutrophil immune function, and extreme sensitivity to DNA demethylation. Blood 2009;113(7):1432-43.

327. Bundschu K, Walter U, Schuh K. Getting a first clue about SPRED functions. Bioessays 2007;29(9):897-907.

328. Brems H, Chmara M, Sahbatou M, et al. Germline loss-of-function mutations in SPRED1 cause a neurofibromatosis 1-like phenotype. Nat Genet 2007;39(9):1120-6.

329. Del Mare S, Salah Z, Aqeilan RI. WWOX: Its genomics, partners, and functions. J Cell Biochem 2009;108(4):737-45.

330. Baykara O, Demirkaya A, Kaynak K, Tanju S, Toker A, Buyru N. WWOX gene may contribute to progression of non-small-cell lung cancer (NSCLC). Tumour Biol 2010;31(4):315-20.

331. Pakstis AJ, Speed WC, Fang R, et al. SNPs for a universal individual identification panel. Hum Genet 2010;127(3):315-24.

332. Nassir R, Kosoy R, Tian C, et al. An ancestry informative marker set for determining continental origin: Validation and extension using human genome diversity panels. BMC Genet 2009;10:39.

333. Stokes A, Joutsa J, Ala-Aho R, Pitchers M, Pennington CJ, et al. Expression profiles and clinical correlations of degradome components in the tumor microenvironment of head and neck squamous cell carcinoma. Clin Cancer Res. 2010;16(7):2022-35.

334. Whitfield ML, George LK, Grant GD, Perou CM. Common markers of proliferation. Nat Rev Cancer. 2006;6(2):99-106.

335. Wurmbach E, Chen YB, Khitrov G, Zhang W, Roayaie S, Schwartz M, Fiel I, Thung S, et al. Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma. Hepatology. 2007;45(4):938-47.

336. Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8:816-24.

337. Raponi M, Zhang Y, Yu J, et al. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer Res 2006;66(15):7466-72.

338. Kruse JJ, Stewart FA. Gene expression arrays as a tool to unravel mechanisms of normal tissue radiation injury and prediction of response. World J Gastroenterol. 2007;13(19):2669-74.

339. Riis ML, Lüders T, Nesbakken AJ, Vollan HS, Kristensen V, Bukholm IR. Expression of BMI-1 and Mel-18 in breast tissue--a diagnostic marker in patients with breast cancer. BMC Cancer. 2010;10:686.

340. Schummer M, Green A, Beatty JD, Karlan BY, Karlan S, Gross J, et al. Comparison of breast cancer to healthy control tissue discovers novel markers with potential for prognosis and early detection. PLoS One. 2010;5(2):e9122.335. Garred P, Honore C, Ma YJ, et al. The genetics of ficolins. J Innate Immun 2009;2(1):3-16.

336. Honore C, Hummelshoj T, Hansen BE, Madsen HO, Eggleton P, Garred P. The innate immune component ficolin 3 (hakata antigen) mediates the clearance of late apoptotic cells. Arthritis Rheum 2007;56(5):1598-607.

337. Munthe-Fog L, Hummelshoj T, Honore C, Madsen HO, Permin H, Garred P. Immunodeficiency associated with FCN3 mutation and ficolin-3 deficiency. N Engl J Med 2009;360(25):2637-44.

338. Schlapbach LJ, Aebi C, Hansen AG, Hirt A, Jensenius JC, Ammann RA. H-ficolin serum concentration and susceptibility to fever and neutropenia in paediatric cancer patients. Clin Exp Immunol 2009;157(1):83-9.

339. Wu CY, Hu HY, Pu CY, et al. Pulmonary tuberculosis increases the risk of lung cancer: A population-based cohort study. Cancer 2010.

340. Moon EH, Kim MJ, Ko KS, et al. Generation of mice with a conditional and reporter allele for Tmem100. Genesis 2010;48(11):673-8.

341. Bingle CD, Wilson K, Lunn H, et al. Human LPLUNC1 is a secreted product of goblet cells and minor glands of the respiratory and upper aerodigestive tracts. Histochem Cell Biol 2010;133(5):505-15.

342. Parsons DW, Jones S, Zhang X, et al. An integrated genomic analysis of human glioblastoma multiforme. Science 2008;321(5897):1807-12.

343. Murugan AK, Bojdani E, Xing M. Identification and functional characterization of isocitrate dehydrogenase 1 (IDH1) mutations in thyroid cancer. Biochem Biophys Res Commun 2010;393(3):555-9.

344. Ishii H, Baffa R, Numata SI, et al. The FEZ1 gene at chromosome 8p22 encodes a leucine-zipper protein, and its expression is altered in multiple human tumors. Proc Natl Acad Sci U S A 1999;96(7):3928-33.

345. Califano D, Pignata S, Pisano C, et al. FEZ1/LZTS1 protein expression in ovarian cancer. J Cell Physiol 2010;222(2):382-6.

346. Ono K, Uzawa K, Nakatsuru M, et al. Down-regulation of FEZ1/LZTS1 gene with frequent loss of heterozygosity in oral squamous cell carcinomas. Int J Oncol 2003;23(2):297-302.

347. Knowles MA, Aveyard JS, Taylor CF, Harnden P, Bass S. Mutation analysis of the 8p candidate tumour suppressor genes DBC2 (RHOBTB2) and LZTS1 in bladder cancer. Cancer Lett 2005;225(1):121-30.

348. Robert B, Sassoon D, Jacq B, Gehring W, Buckingham M. Hox-7, a mouse homeobox gene with a novel pattern of expression during embryogenesis. EMBO J 1989;8(1):91-100.

349. Guney N, Soydinc HO, Derin D, et al. Serum levels of intercellular adhesion molecule ICAM-1 and E-selectin in advanced stage non-small cell lung cancer. Med Oncol 2008;25(2):194-200.

350. Osawa H, Nakajima M, Kato H, Fukuchi M, Kuwano H. Prognostic value of the expression of Smad6 and Smad7, as inhibitory smads of the TGF-beta superfamily, in esophageal squamous cell carcinoma. Anticancer Res 2004;24(6):3703-9.

351. Mangone FR, Walder F, Maistro S, et al. Smad2 and Smad6 as predictors of overall survival in oral squamous cell carcinoma patients. Mol Cancer 2010;9:106.

352. Nishiyama A, Matsui M, Iwata S, et al. Identification of thioredoxin-binding protein-2/vitamin D(3) up-regulated protein 1 as a negative regulator of thioredoxin function and expression. J Biol Chem 1999;274(31):21645-50.

353. Kim SY, Suh HW, Chung JW, Yoon SR, Choi I. Diverse functions of VDUP1 in cell proliferation, differentiation, and diseases. Cell Mol Immunol 2007;4(5):345-51.

354. Han SH, Jeon JH, Ju HR, et al. VDUP1 upregulated by TGF-beta1 and 1,25-dihydorxyvitamin D3 inhibits tumor cell growth by blocking cell-cycle progression. Oncogene 2003;22(26):4035-46.

355. Yang X, Young LH, Voigt JM. Expression of a vitamin D-regulated gene (VDUP-1) in untreated- and MNU-treated rat mammary tissue. Breast Cancer Res Treat 1998;48(1):33-44.

356. Ikarashi M, Takahashi Y, Ishii Y, Nagata T, Asai S, Ishikawa K. Vitamin D3 up-regulated protein 1 (VDUP1) expression in gastrointestinal cancer and its relation to stage of disease. Anticancer Res 2002;22(6C):4045-8.

357. Kwon HJ, Won YS, Suh HW, et al. Vitamin D3 upregulated protein 1 suppresses TNF-alpha-induced NF-kappaB activation in hepatocarcinogenesis. J Immunol 2010;185(7):3980-9.

358. Cadenas C, Franckenstein D, Schmidt M, et al. Role of thioredoxin reductase 1 and thioredoxin interacting protein in prognosis of breast cancer. Breast Cancer Res 2010;12(3):R44.

359. Kopantzev EP, Monastyrskaya GS, Vinogradova TV, et al. Differences in gene expression levels between early and later stages of human lung development are opposite to those between normal lung tissue and non-small lung cell carcinoma. Lung Cancer 2008;62(1):23-34.

360. Mlakar V, Strazisar M, Sok M, Glavac D. Oligonucleotide DNA microarray profiling of lung adenocarcinoma revealed significant downregulation and deletions of vasoactive intestinal peptide receptor 1. Cancer Invest 2010;28(5):487-94.

361. Bellocq A, Antoine M, Flahault A, et al. Neutrophil alveolitis in bronchioloalveolar carcinoma: Induction by tumor-derived interleukin-8 and relation to clinical outcome. Am J Pathol 1998;152(1):83-92.

362. Wislez M, Rabbe N, Marchal J, et al. Hepatocyte growth factor production by neutrophils infiltrating bronchioloalveolar subtype pulmonary adenocarcinoma: Role in tumor progression and death. Cancer Res 2003;63(6):1405-12.

363. Esendagli G, Bruderek K, Goldmann T, et al. Malignant and non-malignant lung tissue areas are differentially populated by natural killer cells and regulatory T cells in non-small cell lung cancer. Lung Cancer 2008;59(1):32-40.

364. Al-Shibli K, Al-Saad S, Donnem T, Persson M, Bremnes RM, Busund LT. The prognostic value of intraepithelial and stromal innate immune system cells in non-small cell lung carcinoma. Histopathology 2009;55(3):301-12.

365. Jung MY, Kim SH, Cho D, Kim TS. Analysis of the expression profiles of cytokines and cytokine-related genes during the progression of breast cancer growth in mice. Oncol Rep 2009;22(5):1141-7.

366. Ikeguchi M, Hatada T, Yamamoto M, et al. Serum interleukin-6 and -10 levels in patients with gastric cancer. Gastric Cancer 2009;12(2):95-100.

367. Knupfer H, Preiss R. Serum interleukin-6 levels in colorectal cancer patients--a summary of published results. Int J Colorectal Dis 2010;25(2):135-40.

368. Sullivan NJ, Sasser AK, Axel AE, et al. Interleukin-6 induces an epithelial-mesenchymal transition phenotype in human breast cancer cells. Oncogene 2009;28(33):2940-7.

369. Gerspach J, Wajant H, Pfizenmaier K. Death ligands designed to kill: Development and application of targeted cancer therapeutics based on proapoptotic TNF family ligands. Results Probl Cell Differ 2009;49:241-73.

370. Collison A, Foster PS, Mattes J. Emerging role of tumour necrosis factor-related apoptosis-inducing ligand (TRAIL) as a key regulator of inflammatory responses. Clin Exp Pharmacol Physiol 2009;36(11):1049-53.

371. Weckmann M, Collison A, Simpson JL, et al. Critical link between TRAIL and CCL20 for the activation of TH2 cells and the expression of allergic airway disease. Nat Med 2007;13(11):1308-15.

372. Weckmann M, Kopp MV, Heinzmann A, Mattes J. Haplotypes covering the TNFSF10 gene are associated with bronchial asthma. Pediatr Allergy Immunol 2010.

373. Mizutani K, Sud S, McGregor NA, et al. The chemokine CCL2 increases prostate tumor growth and bone metastasis through macrophage and osteoclast recruitment. Neoplasia 2009;11(11):1235-42.

374. Suga M, Iyonaga K, Ichiyasu H, Saita N, Yamasaki H, Ando M. Clinical significance of MCP-1 levels in BALF and serum in patients with interstitial lung diseases. Eur Respir J 1999;14(2):376-82.

375. Davi G, Tuttolomondo A, Santilli F, et al. CD40 ligand and MCP-1 as predictors of cardiovascular events in diabetic patients with stroke. J Atheroscler Thromb 2009;16(6):707-13.

376. Ohlsson S, Bakoush O, Tencer J, Torffvit O, Segelmark M. Monocyte chemoattractant protein 1 is a prognostic marker in ANCA-associated small vessel vasculitis. Mediators Inflamm 2009;2009:584916.

377. Lo PH, Lung HL, Cheung AK, et al. Extracellular protease ADAMTS9 suppresses esophageal and nasopharyngeal carcinoma tumor formation by inhibiting angiogenesis. Cancer Res 2010;70(13):5567-76.

378. Lo PH, Leung AC, Kwok CY, et al. Identification of a tumor suppressive critical region mapping to 3p14.2 in esophageal squamous cell carcinoma and studies of a candidate tumor suppressor gene, ADAMTS9. Oncogene 2007;26(1):148-57.

379. Yamato T, Orikasa K, Fukushige S, Orikasa S, Horii A. Isolation and characterization of the novel gene, TU3A, in a commonly deleted region on 3p14.3-->p14.2 in renal cell carcinoma. Cytogenet Cell Genet 1999;87(3-4):291-5.

380. Zigler M, Dobroff AS, Bar-Eli M. Cell adhesion: Implication in tumor progression. Minerva Med 2010;101(3):149-62.

381. Moschos SJ, Drogowski LM, Reppert SL, Kirkwood JM. Integrins and cancer. Oncology (Williston Park) 2007;21(9 Suppl 3):13-20.

382. Mason MD, Davies G, Jiang WG. Cell adhesion molecules and adhesion abnormalities in prostate cancer. Crit Rev Oncol Hematol 2002;41(1):11-28.

383. Ramjaun AR, Hodivala-Dilke K. The role of cell adhesion pathways in angiogenesis. Int J Biochem Cell Biol 2009;41(3):521-30.

384. Ke XS, Qu Y, Goldfinger N, et al. Epithelial to mesenchymal transition of a primary prostate cell line with switches of cell adhesion modules but without malignant transformation. PLoS One 2008;3(10):e3368.

385. Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS. A genome-wide scalable SNP genotyping assay using microarray technology. Nat Genet 2005;37:549-54.

LIST OF FIGURES

Fig. 1 Cancer related death rates in the United States, from 1930 until 2006 (20).....p. 11
Fig. 2 Genetic predisposition to cancer (26).....p. 13
Fig. 3 Relationship between the allele frequency of disease susceptibility locus and their estimated effect size (35).....p. 16
Fig. 4 Knudson's two-hit hypothesis for tumourigenesis.....p. 18
Fig. 5 Acquired Capabilities of Cancer during progression (45).....p. 21
Fig. 6 Stages of tumour progression (NIH, modified by (28)).....p. 25
Fig. 7 Estimated efficiency of association and linkage analysis in relationship to the allele frequency of disease susceptibility locus (63) .....p. 29
Fig. 8 Proportion of cancer mortality attributable to environmental and lifestyle factors (76).....p. 34
Fig. 9 Incidence of lung histological subtypes (Modified from (91) and (92)). SCC indicates SCLC.....p. 37
Fig. 10 SNPs (A), haplotype blocks (B) and TagSNPs (C) (186).....p. 48
Fig. 11 Estimated power in case–control studies and family-based designs (192).....p. 57
Fig. 12 GWA study design (210) .....p. 62
Fig. 13 Kaplan-Meier curves for lung ADCA patients with stage I (red line, number of patients = 60) or with higher clinical stage (blue line, number of patients = 60). Log-rank test showed a significant difference between the two curves ($P = 6.47 \times 10^{-6}$).....p. 78
Fig. 14 Diagram of Infinium II assay protocol (275).....p. 80
Fig. 15 Whole-Genome Genotyping steps (385).....p. 81
Fig. 16 Report of a SNP genotyping (275).....p. 82
Fig. 17 Pyrosequencing: reactions and principles (276).....p. 84
Fig. 18 Diagram of MassARRAY iPLEX Sequenom assay protocol (278)..p. 86
Fig. 19 Schematic representation of SNPs selection in population-based case-control association study.....p. 100
Fig. 20 Plot of the risk of lung cancer associated with the rare allele carrier status at each of 8 SNPs identified by genome-wide scan, in a series

constituted by Italian cases and controls. Shaded squares denote odd ratios (ORs). Horizontal lines represent 95% CIs. The vertical line indicates the null effect (OR=1.0).....p. 102

Fig. 21 Plot of the risk of lung cancer associated with the carrier status of each rare allele of the 8 SNPs reported in Fig. 21.....p. 103

Fig. 22 Schematic representation of SNPs selection in discordant sib-pairs study.....p. 104

Fig. 23 Correlation between SNP frequencies measured by SNP array analysis of DNA pools and frequencies measured by genotyping of individual samples. Plotted data represent frequencies of the rare allele of 75 SNPs putatively associated with lung cancer risk...p. 105

Fig. 24 A polygenic inheritance model with additive and interchangeable effects of rare alleles at lung cancer modifier loci explains the individual risk of lung cancer in the family-based series. Scatter plot shows the proportions of subjects that are cases as a function of genetic susceptibility score and the fitted line.....p. 108

Fig. 25 Schematic representation of SNPs selection in case-only GWAS.....p. 110

Fig. 26 Correlation between SNP frequencies measured by SNP array analysis of DNA pools and frequencies measured by genotyping of individual samples. Plotted data represent frequencies of the rare allele of 63 SNPs.....p. 113

Fig. 27 Genetic risk score in patients with clinical stage I and in patients with higher clinical stage. The horizontal line within the box represents the median value of the genetic estimator of outcome (in base 2 logarithmic units); the upper and lower boundaries of each box represent 75th and 25th percentile, respectively; upper and lower bars indicate the relative highest and lowest values, respectively ($P<2.2 \times 10^{-16}$).....p. 116

Fig. 28 Genetic risk of developing a more aggressive lung ADCA (clinical stage >1) in patients grouped according to the quartiles of genetic risk score, with the lowest quartile as the reference group. Bars denote ORs. Vertical lines represent 95% CIs.....p. 117

Fig 29 Kaplan-Meier survival curves in lung ADCA patients grouped as in Fig 3.9. Follow-up is shown truncated at 60 months (P = 8.0 x 10-8, log-rank test).....p. 118

Fig. 30 Heat map of 55 transcripts (at threshold nominal level of P<1 x 10-3) and of top 68 out of 361 transcripts (at threshold nominal level of P<1 x 10-5) whose expression levels showed statistically significant differences in normal lung of stage I as compared to stage >I patients in the first experiment (A) and in the second experiment (B). Gene names are given on the right. Expression levels of the listed genes are indicated by the colour bar (green, low; red, high).....p. 120

Fig. 31 Correlation between microarray gene expression data obtained on RNA pools and qRT-PCR on individual RNA samples for 22 genes.....p. 121

Fig. 32 mRNA expression levels (mean ± S.E.) of genes in normal lung tissue of lung ADCA patients assessed by qRT-PCR by clinical stage. Asterisks indicate statistically significant differences (P < 0.05) as compared to the reference group (open bars).....p. 123

Fig. 33 IPA network diagram showing the biological associations of 35 genes associated with "Antigen Presentation, Cell-mediated Immune Response, Humoral Immune Response". Genes that showed up-regulation or down-regulation in our samples are in red or in green, respectively. The significance of the nodes are displayed using various shapes that represent the functional classes of the gene products as shown in the key.....p. 125

Fig. 34 mRNA expression levels (mean ± S.E.) of cytokine-related genes in normal lung tissue of lung ADCA patients assessed by qRT-PCR by clinical stage. Data are given as in Fig. 34.....p. 125

Fig. 35 mRNA expression levels (mean ± S.E.) of FCN3 and TMEM100 assessed by qRT-PCR in lung tissue of ADCA patients by stage and type of tissue (N, normal; T, tumour). Asterisks indicate significant differences (P < 0.01) as compared to the reference group (open bars).....p. 126

Fig. 36 Immunohistochemical staining of FCN3, SLC14A1 and SMAD6 proteins. No or few proteins were detected in tumour tissues (right panels), whereas a clear staining patterns were observed in normal tissues for each proteins (left panel).....p. 128

Fig. 37 Clustering of functional annotation pathways based on GWAS- (upper panel) and microarray-derived genes (lower panel) (with p-value < 0.05).....p.131

LIST OF TABLES

Table 1. Characteristics of lung adenocarcinoma patients and control subjects population-based association study.....p. 71

Table 2. Characteristics of discordant sibs series in the family-based association case-control series.....p. 72

Table 3. Characteristics of lung cancer patients in case-only association study.....p. 73

Table 4. Characteristics of lung ADCA patients used in gene expression profile analysis of normal tissue.....p. 74

Table 5. Characteristics of 27 out of 120 cases lung ADCA patients used for paired analysis of the gene expression of lung ADCA tissue and adjacent normal lung tissue.....p. 74

Table 6. Characteristics of Italian lung cancer patients and controls used for DNA pools.....p. 77

Table 7. Summary of pooling approaches used in the thesis.....p. 79

Table 8. Genes present on the TaqMan® Low Density Array for Microarray validation.....p. 91

Table 9. Genes present on the TaqMan® Low Density Array for cytokine-cytokine receptor pathway validation.....p. 92

Table 10. 47 SNPs putative associated with lung cancer risk in GWA analysis of the two experiments.....p. 98

Table 11. 16 SNP validated in the pyrosequencing analysis on DNA pools.....p. 99

Table 12. List of 8 SNPs showing significant association with lung ADCA risk after MassARRAY Sequenom assay on individual samples.....p. 101

Table 13. 36 SNPs showing statistically significant association with lung cancer risk in the discordant sibs series.....p. 106

Table 14. SNPs showing statistically significant association with lung adenocarcinoma risk in the population-based series.....p. 107

Table 15. 63 SNPs putative associated with lung cancer staging.....p. 111

Table 16. SNPs showing statistically significant association with clinical stage in ADCA independent series.....p. 114

Table 17. SNPs showing statistically significant association with clinical stage in SQCC independent series.....p. 114

Table 18. 22 SNPs associated with lung ADCA clinical stage in the joint analysis of the GWA and replication ADCA series and used to build up the polygenic model with additive effects of SNP rare alleles on risk of clinical stage >1.....p. 115

Table 19. Gene expression results of 22 assayed genes in patients with stage I and higher clinical stage using qRT-PCR .....p. 122

Table 20. KEGG pathway analysis by DAVID of the 361 statistically differentially expressed genes between stage I and higher clinical stage patients.....p. 124

Table 21. Gene expression results of 22 cytokine-related genes in patients with stage I and higher clinical stage using qRT-PCR. .....p. 126

Table 22. Gene expression results of 22 assayed genes in lung ADCA tissue and adjacent normal lung tissue using qRT-PCR.....p. 127

Table 23. Integration of GWA data with microarray results.....p. 129

Table 24. Integration of microarray results with GWA data.....p. 130

Supplementary Table 1. Primers used for PCR amplification and genotyping of 47 SNPs.....p. 95

Supplementary Table 2. List of genes differentially expressed between clinical stage I and >I patients in the first experiment.....p. 132

Supplementary Table 3. List of genes differentially expressed between clinical stage I and >I patients in the second experiment.....p. 134

LIST OF WEB SITES

- http://www.who.int/cancer/en/ and www. cancer.org for information about cancer epidemiology and statistics.
- http://hapmap.ncbi.nlm.nih.gov/ for information on SNPs allele frequencies and linkage disequilibrium data.
- http://www.strobe-statement.org/ for The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines.
- http://www.itb.cnr.it/snplims for statistical analysis of dense genotyping data using SNPLims tool.
- http://hugenavigator.net for updated information on GWA studies.
- http://view.ncbi.nlm.nih.gov/dbgap for information on SNPs.
- http://www.genome.gov/GWAstudies/ for updated information on published GWA findings.
- http://www.martalive.org/foreign.htm for information on Marta Nurizzo Association.
- http://linus.nci.nih.gov/BRB-ArrayTools.html for microarray data analyses.
- http://www.genepi.org.au/jlin.html for linkage disequilibrium analyses using JLIN program, version 1.6.0.
- http://david.abcc.ncifcrf.gov/ for pathways and gene ontology analyses of gene expression data.
- https://analysis.ingenuity.com for pathways and networks analyses of gene expression data.

ABBREVIATIONS

ADCA	Adenocarcinoma
ALT	Alternative Lengthening of Telomeres
ANOVA	analysis of variance
APS	adenosine 5' phosphosulfate
AR	Attributable Risk
ATP	Adenosine 5'-Triphosphate
bp	base pairs
CCD	charge coupled device
CEU	HapMap Caucasian population
CI	confidence interval
CIC	Cancer initiating cell
CONSORT	Consolidated Standards of Reporting Trials
CSC	Cancer stem cell
CYP	cytochrome P450
DAVID	Database for Annotation, Visualization and Integrated Discovery
dbGaP	Database of Genotype and Phenotype
dNTP	2'-deoxynucleoside-5'triphosphate
DSPs	discordant sibling pairs
ESP	European standard population
EtBr	Ethidium Bromide
FBATs	family-based association tests
FDR	False Discovery Rate
λ	hazard rate
GLU	Genotype Library and Utilities
GST	glutathione-S-transferase
GWA	Genome-wide association
GWAS	Genome-wide association study
H. Pylori	Helicobacter pylori

HBV	hepatitis B virus
HCV	hepatitis C virus
HIV	human immunodeficiency virus
HHV-EBV	Epstein-Barr virus
hME	homogenous mass extension
HNPCC	hereditary non-polyposis colon cancer
HPV	human papillomavirus
HR	Hazard Ratio
HSV	herpes simplex virus
HWE	Hardy–Weinberg equilibrium
Kb	kilobase
IPA	Ingenuity Pathway Analysis
LCC	large cell carcinoma
LD	Linkage Disequilibrium
ln	natural logarithm
LOH	loss of heterozygosity
MAF	minor allele frequency
Mb	Megabases
mEH	microsomal epoxide hydroxylase
miRNA	microRNA
MPG	Major Cancer Predisposition Genes
NAT	N-acetyl transferase
NCBI	National Center for Biotechnologies Information
NIH	National Institute of Health
NSCLC	non-small cell lung cancer
OR	Odds Ratio
P	*P*-value
PCR	Polymerase chain reaction
PPi	pyrophosphate
PSQ	Pyrosequencing
q	quantile
qRT-PCR	quantitative Real-Time PCR
r	correlation coefficient

RefSeq	Reference Sequence
RFLP	restriction fragment length polymorphism
ROS	reactive oxygen species
RQ	relative quantification
RR	Relative Risk
SCLC	small cell lung cancer
SD	Standard Deviation
SE	Standard Error
Seq	sequence
SQCC	squamous cell carcinoma
SNP	single nucleotide polymorphism
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
TDT	transmission disequilibrium test
TMG	Tumour Modifier Genes
TNM	tumour, nodes, metastasis
UTRs	untranslated regions
WSP	World standard population
χ^2	Chi-square

ACKNOWLEDGEMENTS

The present work was carried out under the supervision of Dr. Tommaso A. Dragani and Prof. Aage Haugen. Through the times I have spent in the Laboratory of Molecular Bases of Genetic Risk and Polygenic model at the Department of Predictive and Preventive Medicine (Fondazione IRCCS Istituto Nazionale Tumori in Milan) I have made lots of special friends. I thank all of them to share the working experiences and scientific knowledge with me and thank all the support they gave to me.

First, I would like to thank my director of study Dr. Tommaso A. Dragani for giving me the opportunity to perform my PhD thesis work in his Laboratory and research group. I would like to express my gratitude to for his continuous, patient guidance, scientific vision, encouragement, help, and great degree of freedom that he offered me.

I would like to thank my supervisor Professor Aage Haugen for his advices during my PhD time, valuable criticism to manuscripts, as well as for general expert help and valuable discussions during my study and who gave their valuable time to read, examine and evaluate this thesis.

I thank our department services, in particular to Dr. Loris De Cecco and Dr. Lucia Gioiosa for technical assistance in gene expression microarrays and immunohistochemistry, respectively; Harvard-Partners Center for Genetics and Genomics Genotyping Facility (Cambridge, MA) for custom genotyping by MassARRAY; and Dr. Anna Gonzalez Neira for her support in genome-wide analysis.

I really thank all patients and healthy donors for their participation in this study, as well as their respective doctors (Dr. Ugo Pastorino, Dr. Luigi Santambrogio e Dr. Matteo Incarbone) for contributing clinical information.

I would like to express my thanks to all my friend-colleagues for sharing their practical knowledge and experience throughout my studies. In particular, I thank Dr. Antonella Galvan and Dr. Stefania Falvella, for their great help, effective cooperation and helpful scientific suggestions. A very special thank goes to Dr. Francesca Colombo who shared with me hopes and experiences, fears and joys. It has been lovely to work with you and thank you for sharing part of your life with me.

I deeply express my sincere thanks to my family for their truly unconditional love and unimaginable constant support and encouragement during my PhD studies. I am very lucky to have a such loving and supportive family, and I thank each of them for always believing me.

Last but not least, I would like to thank my dearest husband for his constant support, patience, help, motivation and being always the most important person for me. I have never been able to express my appreciation enough for his great supports. Thank you!

This study could never have been achieved without all these people.

This research was supported in part by grants from Associazione and Fondazione Italiana Ricerca Cancro (AIRC and FIRC) and by AIRC Triennal Fellowship 2010-2012 "ANTONIETTA ANDREOLI" awarded to me.

ELISA FRULLANTI