EXPLORING THE CONTRIBUTION OF GENETIC AND ENVIRONMENTAL FACTORS TO CANCER RISK AND DEVELOPMENT

Dissertation

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)

eingereicht an der
Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

von

M.Sc. Maria Stella de Biase

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr. Julia von Blumenthal

Dekan der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin
Prof. Dr. Dr. Christian Ulrichs

Gutachter
1. Prof. Dr. Markus Landthaler
2. Prof. Dr. Roland Schwarz
3. Prof. Dr. Dieter Beule

Tag der mündlichen Prüfung
10.01.2023

Declaration of authorship

I hereby declare that I have written this thesis independently and that I have used no other sources or aids than those reported. I have not applied for a doctoral degree elsewhere and do not have a corresponding doctoral degree. I have not submitted the doctoral thesis, or parts of it, to another academic institution and the thesis has not been accepted or rejected. I declare that I have acknowledged the Doctoral Degree Regulations which underlie the procedure of the Faculty of Life Sciences of Humboldt-Universität zu Berlin, as amended on 5th March 2015. Furthermore, I declare that no collaboration with commercial doctoral degree supervisors took place, and that the principles of Humboldt-Universität zu Berlin for ensuring good academic practice were abided by.

Hiermit erkläre ich, die Dissertation selbstständig und nur unter Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt zu haben. Ich habe mich anderwärts nicht um einen Doktorgrad beworben und besitze keinen entsprechenden Doktorgrad. Ich erkläre, dass ich die Dissertation oder Teile davon nicht bereits bei einer anderen wissenschaftlichen Einrichtung eingereicht habe und dass sie dort weder angenommen noch abgelehnt wurde. Ich erkläre die Kenntnisnahme der dem Verfahren zugrunde liegenden Promotionsordnung der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin vom 5. März 2015. Weiterhin erkläre ich, dass keine Zusammenarbeit mit gewerblichen Promotionsberaterinnen/Promotionsberatern stattgefunden hat und dass die Grundsätze der Humboldt-Universität zu Berlin zur Sicherung guter wissenschaftlicher Praxis eingehalten wurden.

February 2023

Acknowledgements

Many people contributed to making my PhD, and these six years in Berlin, a wonderful experience and I would like to say thanks to them here.
First, I would like to thank the Helmoltz Foundation and the MDC-NYU Exchange Program for funding my PhD.
To my supervisor, Roland Schwarz: thank you for the trust you gave me in making me one of the founding members of the Schwarzlab, for the support and encouragement in tough times, the heated scientific discussions, the constant push to get out of my shell, have confidence in my ideas, and 'just be a postdoc'.
To my NYU co-supervisor, Mark Siegal: for your kindness, hospitality, and for always being available to help and support, even at an ocean of distance.
To Gaetano Gargiulo, for being part of my thesis advisory committee and enriching project discussions with insightful comments and advice.
To Markus Landthaler and Dieter Beule, for offering their time to read and review this thesis.
To Bruce Ponder and Robert Rintoul: it was an honour to collaborate with you. Thank you for the always fruitful and exciting back and forth of ideas that kept our project moving forward.
To Eugene Plavskin, the best collaborator I could have wished for: thank you for sharing your knowledge with me, for your patience and your willingness to answer my questions over and over again.
To Florian Massip, always and forever my favourite postdoc (even if you are not a postdoc anymore): you have been a pillar during my PhD, always there to support me, scientifically and personally. I don't know what I would have done without you.
To the present and past members of the Schwarzlab: Marina Petkovic, Martin Burkert, Matt Huska, Cindy Wei, Victoria Dombrowe, Maja Stöber, Tom Kaufmann, Adam Streck. If my PhD was such a joyous experience, it was thanks to you.
To Julia Markowski: in these years we shared our doubts, our crazy thoughts, our anger and frustration, our important milestones, and so so many happy moments. What an incredible friend you are I cannot express in one sentence; thank you for simply being you.
To my mom and dad, who have always encouraged me to follow my passions, always supported me in my decisions and always made sure I knew, wherever I am, they

are my safe haven of unconditional love. Mamma, papá, grazie per avermi sempre appoggiata in tutte le mie decisioni, per essere sempre il mio punto fisso, il mio porto sicuro, la certezza di un amore senza limiti e senza condizioni.
To Nico, best brother in the universe: thank you for always being there for me, I am so lucky to have you.
Finally, to the most wonderful gift this Berlin adventure has given me: my wife Lorena Sofía López Zepeda. Thank you for always believing in me, for understanding me so well, for constantly inspiring me to be better. We shared this journey from the very first day, and I could not have asked for a better partner. I would not choose to live all the new adventures that life will bring with anyone but you.

Abstract

Cancer is a complex disease, initiated and sustained by the interplay of germline and somatic mutations and environmental factors. Understanding the phenotypic consequences of somatic mutations, how their spectrum and effects are influenced by germline variation and the role of environmental factors in increasing cancer risk is key to improving cancer prevention, early detection, and treatment. In this thesis, I explore both the genetic and environmental influences on cellular phenotypes and cancer risk.

First, I take advantage of a simple yeast model to investigate the effects of spontaneously accumulating mutations on cell growth. I describe a mutation accumulation experiment conducted in yeast strains with a defective *MutS*β complex of the mismatch repair system (MMR). *MutS*β-deficient lines accumulate mutations mainly in simple sequence repeats (SSRs), inherently unstable regions involved in normal eukaryotic transcriptional regulation and several human pathologies, including cancer. I show that abrogating *MutS*β function leads to an increased SSR mutation rate, with a bias towards deletions. I also report a drastic increase in mutation rate in SSR loci longer that 8-bp, which suggests *MutS*β is primarily responsible for mismatch repair at longer repeat loci. Finally, I show that SSR mutations have mostly small deleterious effects on cell growth, and propose a role for the combined effects of many mildly deleterious passenger mutations in the favourable prognosis observed for microsatellite unstable cancers.

Next, I investigate the effects of cigarette smoke on the transcriptome of an accessible airway tissue, nasal epithelium, in a cohort of healthy volunteers and patients with suspected or diagnosed lung cancer. I find that smoke injury response is strikingly different in healthy individuals and clinic patients, with genes and biological functions affected by smoking showing a slower reversal to healthy baseline level in clinic patients. In particular, I find persistent smoking-associated immune alterations to be a hallmark of the clinic patients. Finally, I show that a classifier including nasal expression of smoke-injury-associated genes to predict lung cancer performs better than a model based exclusively on clinical information, providing evidence for the potential of nasal epithelial gene expression to improve population-level risk stratification with the use of a non-invasive test.

Zusammenfassung

Krebs ist eine komplexe Krankheit, die durch das Zusammenspiel von Keimbahn- und somatischen Mutationen sowie Umweltfaktoren ausgelöst und aufrechterhalten wird. Das Verständnis der phänotypischen Folgen somatischer Mutationen, der Beeinflussung ihres Spektrums und ihrer Auswirkungen durch Keimbahnvarianten sowie der Rolle von Umweltfaktoren bei der Steigerung des Krebsrisikos ist der Schlüssel zur Verbesserung der Prävention, Früherkennung und Behandlung von Krebserkrankungen. In dieser Arbeit untersuche ich sowohl die genetischen als auch die umweltbedingten Einflüsse auf zelluläre Phänotypen und das Krebsrisiko.

Zu Beginn nutze ich ein einfaches Hefemodell, um die Auswirkungen von spontan akkumulierenden Mutationen auf das Zellwachstum zu untersuchen. Ich beschreibe ein Mutationsakkumulationsexperiment, das in Hefestämmen mit einem defekten *MutSβ*-Komplex des Mismatch-Reparatursystems (MMR) durchgeführt wurde. *MutSβ*-defiziente Linien häufen Mutationen hauptsächlich in einfachen Sequenzwiederholungen (SSRs) an, welche inhärent instabilen Regionen darstellen, die an der normalen eukaryotischen Transkriptionsregulation und mehreren menschlichen Pathologien, einschließlich Krebs, beteiligt sind. Ich zeige, dass die Aufhebung der *MutSβ*-Funktion zu einer erhöhten SSR-Mutationsrate führt, mit einer Tendenz zu Deletionen. Zudem zeige ich einen drastischen Anstieg der Mutationsrate in SSR-Loci, die länger als 8 bp sind, was darauf hindeutet, dass *MutSβ* hauptsächlich für die Reparatur von Fehlpaarungen an längeren Repeat-Loci verantwortlich ist. Abschließend zeige ich, dass SSR-Mutationen meist geringe, negative Auswirkungen auf das Zellwachstum haben, und schlage vor, dass die kombinierten, leicht negativen Effekte vieler Passagiermutationen eine Rolle bei der günstigen Prognose spielen, die für Krebsarten mit Mikrosatelliteninstabilität beobachtet wird.

Als nächstes untersuche ich die Auswirkungen von Zigarettenrauch auf das Transkriptom eines leicht zugänglichen Atemwegsgewebes, des Nasenepithels, in einer Kohorte von gesunden Freiwilligen und Patienten mit Verdacht auf oder diagnostiziertem Lungenkrebs. Ich stelle fest, dass die Reaktion auf durch Rauch hervorgerufene Verletzungen bei gesunden Personen und Klinikpatienten auffallend unterschiedlich ist, wobei Gene und biologische Funktionen, die durch das Rauchen beeinträchtigt werden, bei Klinikpatienten langsamer auf ein gesundes Ausgangsniveau zurückkehren. Ich

stelle zudem fest, dass anhaltende, rauchassoziierte Immunveränderungen ein klares Kennzeichen der Klinikpatienten sind. Schließlich zeige ich, dass ein Klassifikator, der die nasale Expression von mit Rauchverletzungen assoziierten Genen zur Vorhersage von Lungenkrebs einbezieht, bessere Ergebnisse erzielt als ein Modell, das ausschließlich auf klinischen Informationen basiert. Damit belege ich das Potenzial der Einbeziehung der Genexpression des Nasenepithels zur Verbesserung der Risikostratifizierung auf Bevölkerungsebene mit Hilfe eines nicht-invasiven Tests.

Table of contents

List of figures

List of tables

List of abbreviations

AUC Area Under the Curve

CA Cessation-associated

COPD Chronic obstructive pulmonary disease

CRC Colorectal cancer

CS Current smoker

ctDNA Circulating tumour DNA

CV Cross-validation

ECM Extracellular matrix

EMT Epithelial-mesenchymal transition

FS Former smoker

GATK Genome Analysis ToolKit

GFP Green fluorescent protein

GO Gene Ontology

HPV Human papilloma virus

HV Healthy volunteer

ICGC International Cancer Genome Consortium

IGV Integrative Genome Browser

INDEL Insertion/deletion

IR Irreversible

LDCT Low-dose computer tomography

LOH Loss of heterozygosity

LS Lynch syndrome

LUAD Lung adenocarcinoma

MA Mutation accumulation

MMR Mismatch repair

MSI Microsatellite instability

NGS Next-generation sequencing

NSCLC Non-small cell lung cancer

NV Never smoker

PCAWG Pan-Cancer Analysis of Whole Genomes

PCA Principal component analysis

PCR Polymerase chain reaction

PR Precision-recall

PSA Prostate-specific antigen

ROC Receiver operating characteristic

ROS Reactive oxygen species

RR Rapidly reversible

SCC Squamous cell carcinoma

SCLC Small cell lung cancer

SC Synthetic complete medium

SNM Single-nucleotide mutation

SR Slowly reversible

SSR Simple sequence repeat

TCGA The Cancer Genome Atlas

TF Transcription factor

TRED Triplet repeat expansion disorder

US Unaffected by smoking

WT Wild type

YPD Yeast extract Peptone Dextrose

Chapter 1

Introduction

1.1 Genes and environment in cancer

The definition of cancer as a "disease of the genome" is today very well established. Malignant transformation starts with the accumulation of somatic mutations in the genome of a normal cell. While some mutations are inconsequential, others lead to changes which confer the cell a selective advantage. The accumulation of these mutations eventually leads to uncontrolled expansion of the cell of origin, which forms a "clone" that grows, giving rise to the tumour. In 2000, Hanahan and Weinberg described six fundamental characteristics, or hallmarks, of malignant cells (Hanahan and Weinberg, 2000). These hallmarks are sustained growth, evasion of growth-regulating signals, replicative immortality, resistance to cell death, induction of angiogenesis, and the ability to invade and metastasise other tissues. In the following years, two additional characteristics were included: metabolic reprogramming and immune evasion (Hanahan and Weinberg, 2011). These are characteristics of the cancer cells themselves. However, a tumour does not exist as an isolated entity. It is part of a complex environment that includes a variety of surrounding cell types and intercellular components such as stromal, endothelial and immune cells and the extracellular matrix (ECM). The back and forth interaction between malignant cells and non-malignant components of this "tumour microenvironment" sustains tumour growth and promotes its development and invasiveness (Balkwill et al., 2012). Therefore, factors that modify and shape the tumour microenvironment also contribute to cancer development. An important example is tumour-associated inflammation. While the immune system plays a role in the initial anti-tumoral response, attempting to eradicate malignant cells from healthy tissue, it is now clear that inflammation at the tumour site promotes tumorigenesis by producing proliferation-sustaining, pro-angiogenic and ECM-degrading factors that enhance growth and invasiveness, together with cytokines and chemokines

that sustain the inflammatory state making the tumour site a never-healing wound (Grivennikov et al., 2010).

Mutations that confer cells a fitness advantage, which often coincides with the acquisition of the above-mentioned hallmark capabilities, are defined as 'driver' mutations, as opposed to 'passenger' mutations, which have no effect on cellular phenotype. Driver mutations frequently affect two types of genes: proto-oncogenes and tumour suppressors (Kopnin, 2000). Proto-oncogenes are genes whose normal functions promote cell proliferation and growth. Common examples of proto-oncogenes encode for growth factor and angiogenic factor receptors such as EGFR and VHL, intracellular signalling molecules such as RAS and RAF, and anti-apoptotic factors such as BCL-2. Cancer-promoting mutations in these genes are typically gain-of-function mutations, which lead to an over-activation of the gene product and thus enhanced growth and proliferation, and resistance to apoptosis and anti-proliferative signals. Mutations in proto-oncogenes are also generally dominant, with only one mutated copy of the gene being sufficient to produce the pro-tumorigenic effect (Alberts et al., 2002). Tumour suppressors, on the other hand, have anti-proliferative and pro-apoptotic functions, or are involved in DNA repair and maintenance of genome stability. The most typical examples of tumour suppressor genes are *TP53* and *RB1*, which are involved in growth, apoptosis and DNA repair regulation. Mutations observed in tumour suppressor genes are loss-of-function mutations, and generally both copies of the gene must be inactivated to observe an effect on the phenotype (Alberts et al., 2002; Kopnin, 2000).

A diverse range of mutations is observed in cancer genomes and can potentially affect driver genes. These include single nucleotide substitutions, short insertions and deletions, copy-number alterations involving large regions of chromosomes, chromosome arms or entire chromosomes, and a variety of complex structural variants. These mutations can also indirectly affect the function of driver genes. Mutations occurring in non-coding regions of the genome can potentially modify the behaviour of regulatory regions such as enhancers and insulators, reshape chromatin organisation and affect the expression of non-coding regulatory RNAs (Elliott and Larsson, 2021). Changes in genome sequence are not the only way to acquire hallmarks of cancer. Functional alteration of oncogenes and tumour suppressors can also be caused by epigenetic alterations (Nishiyama and Nakanishi, 2021). For instance, promoter hypermethylation has been shown to be a mechanism for silencing of tumour suppressor genes in several cancer types (Esteller et al., 2000; Herman et al., 1994; Merlo et al., 1995; Zhang et al., 2008). Patterns of global hypomethylation have also been observed in cancer genomes. This

widespread loss of methylation has been linked to increased chromosomal instability; it can also lead to enhanced transcription of the regions involved, and the potential overexpression of oncogenes located in these regions (Eden et al., 2003; Nishiyama and Nakanishi, 2021).

Although the accumulation of somatic alterations is ultimately what leads to tumour development, germline variation can favour malignant transformation. For example, germline variants that affect the DNA repair system can increase the overall somatic mutational burden of the genome (Curtin, 2012; de Boer and Hoeijmakers, 2000; Roy et al., 2011). In other cases, a mutation in a tumour suppressor is inherited, making the cell already closer to a malignant state. That is the case for some hereditary cancers, most famously retinoblastoma. While studying this cancer, Alfred Knudson first formulated his "two hit" hypothesis, according to which tumour suppressor genes are recessive in nature and, in order to observe a phenotypic change in the cell, both copies of the gene need to be inactivated, through either mutation or epigenetic mechanisms (Knudson, 1971). Germline variation can influence tumorigenesis by acting not only on the cell-of-origin itself, but on components of the tumour microenvironment, such as ECM components, stromal cells, immune cells and blood vessels. Polymorphisms in genes involved in ECM structure, such as metalloproteinases and adhesion molecules, have been found to be implicated in tumorigenesis, to influence tumour characteristics, survival time and response to therapy for several cancer types (Han et al., 2011; Kida et al., 2014; Ricketts et al., 2009). Associations between polymorphisms and cancer risk were also described for genes that encode proteins produced by stromal cells, such as SDF1 and TGFβ1, and vascular factors such as VEGF (Eng et al., 2012; Krishna et al., 2020; Teng et al., 2009; Verboom et al., 2017). Variants affecting immune system functions also weigh on cancer risk and development. These variants can influence both the immune anti-tumoral response and the cancer-associated inflammation within the tumour microenvironment (Duell et al., 2006; Eaton et al., 2018; Frank et al., 2010; Korobeinikova et al., 2020; Kwon et al., 2011; Sayaman et al., 2021; Shahamatdar et al., 2020).

Mutations accumulate randomly in normal cells due to endogenous mutagenic processes such as oxidative stress, erroneous DNA repair and ineffective DNA polymerase proofreading (Barnes et al., 2018). However, they can also be induced by environmental carcinogens, such as cigarette smoke and UV radiation. Both endogenous and exogenous processes increase the frequency of certain alterations occurring, creating very specific mutational patterns or "signatures" (Alexandrov et al., 2013, 2020). Diet and obesity are also linked to increased risk of some cancers, as are infections with

certain viruses and bacteria (De Pergola and Silvestris, 2013; Krump and You, 2018). Even if some of these factors, such as obesity, do not directly induce the occurrence of mutations, most of them lead to chronic inflammation, creating a pro-tumorigenic environment (Bosch et al., 2002; Coussens and Werb, 2002; Howe et al., 2013; Lakatos and Lakatos, 2008; Lee et al., 2012). Interaction between genetic and environmental factors can also play a role in cancer initiation and development, as germline variation can modify the response to environmental carcinogens. For example, polymorphisms in *NQO1*, a gene encoding for an enzyme involved in cellular detoxification, have been shown to affect the sensitivity to tobacco carcinogens, thus influencing the risk of smoking-associated lung cancer (Yamamoto et al., 2017). Similarly, polymorphisms in the vitamin D receptor gene *VDR* were associated with an increased risk of UV-induced skin cancer (Denzer et al., 2011).

Understanding the consequences of germline and somatic variation, the effect of environmental factors, and the interaction between the two, on cellular phenotype is fundamental to devise preventive measures, early detection and therapeutic strategies for human cancers.

1.2 Cancer genomics and its potential clinical applications

The notion that cancer is a disease of the genome, together with the availability of the complete sequence of the human genome at the beginning of the millennium, and the subsequent advent of next-generation sequencing (NGS), has led to a shift in the way we study and treat cancer. It became possible to characterise cancer cell genomes, transcriptomes, proteomes, and epigenomes, and to compare them with those of normal cells, thus identifying the key features underlying malignancy and disease progression. Single-cell sequencing provided an additional layer to tumour characterisation, allowing one to study tumour heterogeneity, the evolutionary processes underlying cancer development, the spatial organisation of the tumour mass and tumour microenvironment. In 2006 and 2008, the massive efforts of The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) were started, followed and integrated by the Pan-Cancer Analysis of Whole Genomes Consortium (PCAWG) (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020; Weinstein et al., 2013). These efforts led to the collection of genomic, epigenomic, transcriptomic, and proteomic data from more than 2500 samples across

39 cancer types. In 2019 ARGO, a new phase of ICGC, was launched (ICGC-ARGO, 2019). The ARGO project will analyse samples from 100,000 cancer patients using standardised analysis pipelines, in order to gain knowledge that will ultimately improve personalised treatment, prevention and early detection strategies.The availability of this vast amount of data and the technological tools to produce it led to the discovery of a wide spectrum of actionable alterations and to the birth and development of precision oncology. During the last two decades, several cancer driver genes have been identified and the effects of their alteration on cellular phenotype characterised. This led to the development of therapeutic agents specifically targeting these alterations, which complemented or replaced traditional cytotoxic chemotherapy. Two key examples of targeted drugs are trastuzumab and imatinib, developed, respectively, to treat HER2-amplified breast cancer and BCR-ABL-fusion-positive leukaemia (Druker et al., 2001; Slamon et al., 2001).

Analysing the genome and transcriptome of healthy and cancerous tissues allows us to understand the mechanisms of disease initiation and progression, and thus devise so-called biomarkers. A biomarker is a feature, or a short list of features, which carries information about the status of the disease, and can therefore be of clinical use in several settings, such as cancer screening and risk assessment, diagnosis, prognosis, monitoring of disease progression and prediction of response to therapy. A biomarker can be based on the presence/absence of germline variants and somatic mutations, on gene expression, protein activity or metabolic profiles (Henry and Hayes, 2012). The ideal biomarker, specially for risk assessment and early detection, is non-invasive or minimally invasive, allowing for easy and frequent testing of non-symptomatic subjects; these might be individuals at high risk of developing certain cancer types (e.g. cigarette smokers) or part of the general population. For example, in recent years, a lot of research has focused on biomarkers based on circulating tumour DNA (ctDNA), genetic material from malignant cells that, if detected in the subject's blood stream, can inform diagnosis and clinical decisions (Pessoa et al., 2020).

Despite all the advances and discoveries described, and their indubitable potential to improve cancer patient care, translating genomic-based biomarkers from a research setting to actual clinical use is not an easy task, and few biomarkers are translated into actual clinical practice. This can be due to the biomarker failing to meet the required criteria of high sensitivity, specificity and cost effectiveness (Diamandis, 2012). Moreover, while early detection of cancer is generally desirable and conducive to timely treatment and better survival, there is a risk of overdiagnosis and overtreatment. A prominent example is the controversial use of prostate-specific antigen (PSA) for the

screening of prostate cancer in men. The improvement in early detection of prostate cancer using PSA screening and the benefits derived from it were negatively balanced by the increase in unnecessary medical procedures, both diagnostic and curative, for a typically indolent and slow-growing malignancy (Loeb et al., 2014). Another point to consider is that often, at the point in disease development where a biomarker is capable of detection, the cancer is already at an advanced stage, effectively limiting the benefits of the biomarker assessment (Diamandis, 2012). Furthermore, biomarkers are usually built by leveraging features of the tumour mass itself. However, biomarkers for screening and assessing the risk of cancer occurrence and recurrence would greatly benefit from taking into consideration other tumour-associated features, such as the characteristics of cancer stem cells, the tumour microenvironment, and tissue alterations preceding and leading to the actual malignancy (Brooks, 2012).

In this context, the study of the processes occurring before, and leading to, cancer insurgence is another promising application of cancer 'omics. Malignant transformation is usually a long and slow process, and alterations linked to tumorigenesis can occur years before a tumour is diagnosed (Gerstung et al., 2020). Often, in particular for cancer types associated with environmental factors, such as colorectal, oesophageal and lung cancer, pre-malignant lesions are present in the involved tissue before the appearance of a frank malignancy. Pre-malignant lesions are areas of tissue with an abnormal appearance, that frequently harbour similar driver mutations, pathway alterations and hallmark characteristics as the cancer itself, but lack the characteristic uncontrolled expansion and invasiveness of malignant tissue (Ryan and Faupel-Badger, 2016). Thus, an individual presenting such lesions is at higher risk of developing cancer. Examples of pre-malignant conditions are Barrett's oesophagus, oral leukoplakia, colorectal ulcerative colitis, cervical dysplasia, and lung squamous metaplasia (Burd, 2003; Hnatyszyn et al., 2019; Kaz et al., 2015; Mustafa et al., 2021; Wistuba et al., 1997). However, not all pre-malignant lesions will progress to cancer: it is possible for the lesions to regress to a healthy morphology or persist in their pre-malignant state without ever developing further. One of the factors in determining the final outcome of a pre-malignant lesion is the microenvironment in which it resides. Current knowledge on the topic points to a switch of the involved cellular components from an anti-tumorigenic to a pro-tumorigenic behaviour as the lesion progresses towards malignancy (Jones et al., 2021). A fundamental role is played by the immune system, both its innate and adaptive components. While early lesions are characterised by strong immune surveillance, as they progress, the microenvironment switches to an immunosuppressive state that is fully established and observable in

the tumour microenvironment. Being able to identify pre-malignant lesions based on their morphology is already a valuable tool in clinical practice. One example is the "Pap test", currently used in routine screening for pre-cancerous lesions caused by the human papilloma virus (HPV) in the uterine cervix (Kitchen and Cox, 2021). However, molecular typing of pre-malignant lesions has the potential to add further clinical benefits, by providing insight into the possible outcome of the lesions, and indication for possible chemopreventive strategies. Colorectal cancer (CRC), for example, develops from precursor lesions exhibiting clear genetic, transcriptional and epigenetic alterations compared to normal tissue. Among these alterations is the overexpression of nitric oxide synthase (iNOS) and lipoxygenase (5-LOX), pro-inflammatory enzymes. Inhibitors of these enzymes have shown promising chemopreventive effects in preclinical models of CRC (Gao et al., 2019; Gounaris et al., 2015). Thanks to advances in NGS technology, alterations present in precursor CRC lesions can be detected in stool samples, allowing the development of non-invasive screening tools (Imperiale et al., 2014). Another example of chemopreventive agent targeting molecular alterations in pre-malignant lesions is lapatinib, a tyrosine kinase inhibitor used in the treatment of HER2-positive breast cancer, which was shown to be also effective in preventing the progression of precursor lesions in mammary tissue (Decensi et al., 2011; Ma et al., 2017). Smoking-induced lung cancer also develops from precursor dysplastic lesions. Increased activity of the phosphatidylinositol 3-kinase (PI3K) pathway was observed in normal-appearing bronchial tissue of subjects with dysplastic lesions. Patients treated with PI3K-inhibiting myo-inositol showed significant regression of dysplastic lesions, indicating the chemopreventive potential if this molecule (Gustafson et al., 2010).

While the characteristics of cancer tissue have been extensively studied, the genomic an transcriptomic landscape of pre-malignant tissue, and its interaction with the surrounding microenvironment, are still largely uncharacterized. Identifying the sequence and effects of events leading from healthy tissue to pre-malignancy, and ultimately to frank malignancy, will open the possibility to detect cancer, or determine the risk of developing cancer, before it presents, and thus be able to implement effective prevention or early intervention strategies.

1.3 The utility of yeast as a model organism in cancer research

It is hard to unravel genetic and environmental contributions to complex diseases like cancer in complex organisms like humans. Cell cycle, transcription, protein synthesis, mitochondrial function, autophagy and many other basic cellular processes are conserved between species, even evolutionarily distant ones. Therefore, studying these processes in simpler model organisms is easier, but still translates well to human biology. For this, model organisms can be a valuable resource in cancer research (Dolinski and Botstein, 2007).

Throughout the years, research involving model organisms has led to important discoveries which advanced our understanding of cancer, ranging from basic knowledge of cellular processes to the identification and characterisation of oncogenes and tumour suppressors to drug discovery and testing. Although more complex organisms such as fruit fly and mouse allow to study genes, pathways and responses to administered substances in a systemic context, simple model organisms have played a major role in biomedical research. A prime example of the utility of simple model organisms for cancer research is baker's yeast, *Saccharomyces cerevisiae*. This organism has been extensively used as a model for eukaryotic cell processes and in particular for the study of cancer-related cellular events (Hartwell, 2002). The use of yeast as a model organism has several advantages. Its unicellular nature and short division time make it easy and cost-effective to grow and maintain. It is easy for the experimenter to control its environmental conditions and even modify its genetic background. Moreover, its basic cellular functions are remarkably similar to those of human cells, with thousands of genes having corresponding human orthologs (Kachroo et al., 2015; Sonnhammer and Östlund, 2015). In fact, the genes, phases and checkpoints of the cell division cycle were first discovered and described in *S. cerevisiae* by Leland Hartwell, Paul Nurse and Tim Hunt, and were later found to be very similar to those of human cells (Hartwell, 2002; Hunt; Nurse, 2002). Hartwell's studies were also the first to provide insight into the role of DNA repair defects in genome fidelity and cancer susceptibility (Hartwell, 2002). The extensive study of yeast uncovered other aspects that closely match human cancer-related characteristics. For example, a hallmark of cancer is the reprogramming of energy metabolism. Hypoxia, mitochondrial dysfunctions and other stress-inducing stimuli are often associated with metabolic reprogramming in tumour cells (Diaz-Ruiz et al., 2009). A lot of processes occurring in yeast cells during adaptation to external

environmental conditions can be compared to the metabolic switch occurring in human cells during malignant transformation. In yeast, under stress conditions, the 'retrograde response' (RTG) is activated. This is a communication pathway between mitochondria and the nucleus, an its activation promotes metabolic adaptation and pro-survival signalling. This system was first characterised in *S. cerevisiae* (Liu and Butow, 2006), and it was later found to be comparable to the more specialised NFkB pathway in mammalian cells (Srinivasan et al., 2010). Yeast switches from oxidative metabolism to fermentation when glucose supply is high. Ras proteins are involved in this switch, by activating a response to nutrient availability: when glucose is present, they activate a pathway that leads to cell growth, differentiation and survival (Rolland et al., 2002). This metabolic switch is comparable to the Warburg effect in tumour cells, which consists in increased glycolysis with conversion of glucose primarily to L-lactate, and is mediated by the direct human homologs of Ras proteins (Liberti and Locasale, 2016). Signalling leading to apoptosis is also similar in yeast and mammalian cells, with many orthologous genes involved in this process (Carmona-Gutierrez et al., 2010). These and many other similarities exist between the cellular machinery in yeast and mammals, which make yeast a perfect model for studying some of the aspects of tumorigenesis in a simple, easily controllable context.

A very common and effective way to study the effects of mutations in cancer genes is the use of functional assays in yeast lines. Since many human genes have a homologous counterpart in yeast, it is possible to "humanise" yeast strains by introducing human cancer-associated mutations into these genes. A functional assay is based on the comparison of the phenotype produced by the mutated gene with the phenotype of a wild-type strain, and it can have several possible readouts (Cervelli et al., 2020). For instance, the effect of mutations in DNA-repair genes can be investigated with DNA damage sensitivity assays. In these assays, the effect of a mutation on the phenotype is tested by exposing the strains carrying the mutation to DNA-damaging agents and subsequently measuring their growth compared to that of wild-type strains (Kim et al., 2018; Lee et al., 2012). The function of DNA-repair genes can also be studied using forward and reverse mutation assays. In both assays, defects in genes leading to an increase in mutation rate are detected by using reporter genes. In forward mutation assays, the reporter could be a gene conferring sensitivity to an antibiotic (e.g. *CAN1* or *URA3*). Mutations in the reporter gene that occur in the DNA-repair-defective strains abolish its function and confer resistance to the antibiotic. The mutation rate can then be estimated by counting resistant colonies. In reverse mutation assays, the reporter gene (e.g. lacZ) is initially not functional. An increase in mutations in

the DNA-repair-defective strains will restore gene function by chance, producing an observable and quantifiable phenotype (e.g production of β-galactosidase and blue appearance of colonies) (Gammie et al., 2007; Shimodaira et al., 1998).

Using endogenous genes, as just described, is the preferred approach, as it allows one to study the effects of mutations occurring in their native genomic context, where genes are under the control of their natural promoters, and all context-dependent gene regulation mechanisms are preserved (Cervelli et al., 2020). However, when a direct homolog to a human gene of interest is missing, it is possible to produce humanised yeast strains by introducing plasmids carrying the mutated human gene under the control of a yeast promoter (Hamza et al., 2015; Laurent et al., 2016). This is the case for the tumour suppressors *BRCA1* and *BRCA2*. Several assays were designed to study the effects of mutations in these genes, using different readouts such as transcriptional activation, colony size and protein localisation (Carvalho et al., 2007; Coyne et al., 2004; Monteiro et al., 2020).

Assays based on reporter genes, such as the ones described above, are very useful for providing a simplified model in which to explore the function of cancer genes. They do, however, have limitations. For example, they do not allow to observe the full spectrum of mutations arising in DNA-repair-defective strains, nor to gain insight into mutation rate biases due to genomic context. A powerful approach that can overcome some limitations of reporter assays is the use of mutation accumulation (MA) experiments (Halligan and Keightley, 2009; Mukai, 1964; Ohnishi, 1977). In these experiments, several lines are derived from a single ancestor strain and propagated in parallel for many generations, during which they spontaneously accumulate mutations. Importantly, all lines are subjected to frequent single-cell bottlenecks, obtained by picking a single colony (derived from a single cell) and re-streaking it on a fresh plate. This step drastically reduces the effects of natural selection, allowing all non-lethal mutations to accumulate in the MA lines, even if their effect reduces fitness. Although MA experiments allow the effects of mutations arising across the entire genome to be assessed, in multiple parallel lines at the same time, initially they still relied on phenotypic readouts for the identification of mutational events (Katju and Bergthorsson, 2019). This means that only indirect estimation of mutation rates was possible. Moreover, identification of the full spectrum of mutations was still difficult, as neutral or nearly neutral mutations do not produce an observable phenotypic change. The advent of next-generation sequencing opened the possibility of sequencing the entire genome of MA lines and identifying all mutations that occur in it by comparing it with the genome of the ancestor strain. This allowed observation of the full spectrum of

mutations and direct estimation of mutation rates (Serero et al., 2014; Zhu et al., 2014); it also made it possible to estimate rates for different mutation types (e.g. substitutions, insertions/deletions, copy-number changes), genomic regions (e.g. exonic, intronic, intergenic) and cellular compartments (e.g. nuclear, mitochondrial) (Lynch et al., 2008, 2016; Zhu et al., 2014). Sequencing the transcriptome of MA lines can also provide insight into the transcriptional consequences of mutations (Konrad et al., 2018).

Another aspect of cancer biology that MA experiments can help investigate is the study of passenger mutations and their contribution to cancer progression. As discussed in Section 1.1, mutations in driver genes lead to the formation and expansion of clones during tumour progression. However, driver events are rare and the vast majority of mutations accumulated in a tumour are passengers (ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). Being non-recurrent alterations, and not directly leading to clonal expansion, passengers have been assumed to be phenotypically neutral and play a minor role in tumour progression. However, while the effect of individual passengers on the phenotype might be negligible, the combined effects of many mildly advantageous or deleterious mutations could significantly affect tumour progression (Castro-Giner et al., 2015; Kumar et al., 2020; McFarland et al., 2013). In particular, the accumulation of deleterious passengers with effects not strong enough to be eradicated by negative selection can lead to phenomena that slow down cancer progression, such as proteotoxic stress (Brancolini and Iuliano, 2020) and neoantigen-induced immune response (Jiang et al., 2019). Events like long dormancy, slow progression and spontaneous regression might be explained by the presence of these deleterious passengers mitigating the pro-proliferative and anti-apoptotic effects of driver mutations. Understanding the phenotypic effects of passenger mutations might thus have important clinical applications, potentially leading to the development of cancer treatments that use deleterious passengers to the patient's advantage, by enhancing the anti-cancer activity of passengers. With MA experiments it is possible to observe deleterious mutations with mild effects, and to determine their combined effects on phenotype.

The type of basic research on simple model organisms described in this section can have, and has had, a major impact on clinical cancer management. A prominent example is the history of Lynch syndrome (LS) (Boland and Lynch, 2013). LS was first qualitatively described as a hereditary predisposition to developing colorectal cancer in the early 1900s. No information about the causes or mechanisms of the disease was found until the early 1990s, when LS tumours were associated with the presence of instability in the length of simple repetitive sequences (microsatellites).

This phenotype was recognised by basic scientists studying genetics in bacteria and yeast (Heinen, 2016). Previous studies in these model organisms had described the DNA mismatch repair system (MMR), and observed that mutations in MMR genes led to increased microsatellite instability. Research efforts were then focused on finding and studying the human homolog of the MMR pathway, and eventually led to improvements in the diagnosis and treatment of LS-associated and MMR-defective sporadic cancers. The main human MMR genes were characterised, as were the molecular phenotypes associated with mutations in these genes, such as loss of protein stability. The diagnosis of LS, previously only based on patient's age and family history, could then be aided by molecular tests, such as immunohistochemistry to detect the presence of the main MMR proteins in a patient's sample (Li et al., 2020). Molecular diagnosis also made it possible to identify tumours associated with sporadic alterations in MMR genes (Heinen, 2016; Li et al., 2020). Basic research on simple models also had an impact on the choice of treatment for microsatellite-unstable tumours, as it showed that MMR-defective cells do not respond to certain chemotherapeutic agents (Aebi et al., 1996; Brown et al., 1997).

1.4 Thesis outline

In this thesis I will present two projects exploring the genetic and environmental contributions to cancer-associated phenotypes.

Chapter 2 focuses on the genetic component of cancer by studying the effect of mutations on cellular growth rate in the model organism *S. cerevisiae*. Specifically, the focus is on simple sequence repeats (SSRs), regions particularly prone to mutate, with a known involvement in normal and pathological eukaryotic biology, including cancer. By using data from a mutation accumulation experiment conducted in DNA-repair-deficient yeast strains, we estimated the rate and spectrum of contractions/expansions occurring in SSRs and measured their effect on cell fitness.

Chapter 3 focuses on a cancer type with a known environmental component: lung cancer. Here, I explore the potential use of nasal epithelium as a non-invasive alternative to deeper airway tissues for improving lung cancer risk stratification. We investigated cigarette smoke-induced transcriptional alterations in the airways of healthy volunteers and patients with suspected or diagnosed lung cancer, and used these alterations to predict lung cancer risk and to gain insight into the mechanisms leading to increased risk.

Finally, in **Concluding remarks** I give a brief summary and outlook of the findings described in the thesis.

Chapter 2

The effect of SSR mutations on growth phenotype in *S. cerevisiae*

Contributions

The work presented in this chapter is part of a collaborative project within the MDC-NYU PhD Exchange Program. It was conducted in collaboration with Eugene Plavskin (NYU Center for Genomics and Systems Biology, New York, United States), under the supervision of Roland Schwarz (MDC, Berlin, Germany) and Mark Siegal (NYU Center for Genomics and Systems Biology, New York, United States). Sequencing of the MA lines with no previously identified mutations and the spores derived from their cross, and growth rate assay on these lines were performed by me and Eugene Plavskin. Additional experiments and modelling of mutation rates were performed by Eugene Plavskin. Generation of the SSR reference list was performed by me and Eugene Plavskin. Mutation calling using FreeBayes, including filtration of the raw variant list to obtain final callsets, was performed by Eugene Plavskin. Mutation calling using Muver, MSIsensor and GATK, including the preceding processing of raw sequencing data (quality control, pre-processing, alignment) and subsequent analysis of the results were performed by me. In the text, the first person plural is used when work was performed by a collaborator, or jointly with a collaborator.

Part of this work is reported in Plavskin, de Biase et al. (2022), available as a preprint on bioRxiv.

2.1 Introduction

2.1.1 Studying the effects of spontaneous mutations in yeast

Cancer is caused by the accumulation of spontaneous or induced mutations in the genome. Predisposing germline mutations in genes controlling cell growth and DNA repair and environmental mutagens can increase the rate of mutations. For this reason, studying the way mutations spontaneously accumulate, and how their rate and spectrum are influenced by genetic and environmental factors, is key for cancer research.

Mutation accumulation (MA) experiments (Section 1.3) conducted in simple model organisms such as yeast, in combination with NGS, are a powerful tool for tightly controlling environmental conditions and genetic background while collecting large amounts of data from several parallel lines. Since mutations normally appear at a very slow rate, growing many strains in parallel makes it possible to observe even rare mutations. Moreover, it is possible to assess the effects of the accumulated mutations on phenotype by performing functional assays on the final MA lines. Due to the reduced effect of selection during the MA process, deleterious mutations of moderate and small effects can be observed, allowing investigation of the cumulative impact of passenger-like mutations on the cellular phenotype.

In 2014, Zhu et al. (2014) sequenced the whole genome of 145 diploid MA lines that were propagated for $\sim$2000 generations. They identified a wide spectrum of mutations in the MA lines, including substitutions, small insertions and deletions, and aneuploidies. The large number of lines and generations, and the consequent high number of mutations identified, allowed the authors to estimate diploid mutation rates for each class of mutations. Despite the power of combining MA experiments with NGS, some limitations remain. For instance, certain classes of mutations were excluded from the analysis by Zhu et al. (2014) due to the challenges associated with their sequencing and genotyping. This is the case for mutations occurring in repeat regions, such as simple sequence repeats (SSRs).

Another limiting factor for MA experiments is the availability of appropriate phenotypic measurements to assess the effects of the accumulated mutations. Great accuracy and sensitivity are necessary to measure very small mutational effects, thus there is a high chance of missing such effects. To overcome this issue, the Siegal lab has developed a high-throughput assay based on live imaging of micro-colonies, which allows for very precise estimation of growth rates (Levy et al., 2012; Sartori et al.,

2021). The Siegal lab expanded the work of Zhu et al. (2014) by measuring the growth rate changes of 70 of the sequenced MA strains compared to their ancestor. They found that a few MA strains showed changes in growth rate while apparently harbouring no mutations. They also performed modelling of the mutational effects of single-nucleotide mutations (SNMs)(unpublished). The results of their modelling suggested that an additional class of mutations, not considered in their analysis, contributed significantly to the observed phenotypic effects.

The aim of my collaboration with the Siegal lab was to identify additional mutations potentially causing the unexplained growth rate changes observed in the MA lines from Zhu et al. (2014). We focused particularly on mutations occurring in SSRs, given the functional significance of these regions in normal and pathological eukaryotic biology (described in the following section), and as repeat regions were excluded from the analysis in Zhu et al. (2014).

2.1.2 Simple sequence repeats and their role in human disease

Simple sequence repeats, also known as microsatellites, are tandem repeats of 1-6bp motifs present in the genome of all organisms, from the simplest prokaryotes to higher order eukaryotes (Field and Wills, 1996; Tautz and Renz, 1984). SSRs can be "perfect", when they present as an uninterrupted streak of identical motif repeats, "imperfect", if one or multiple non-motif nucleotides interrupt the motif repeats, and "composite", when they consist of adjacent repeats of two distinct motifs (Vieira et al., 2016). A key characteristic of SSRs is their hyper-variability: they are highly polymorphic in populations, with several alleles presenting different repeat copy numbers, which leads to high heterozygosity. Because of this characteristic, for decades they have been used as polymorphic markers for applications such as genome mapping and DNA fingerprinting (Ellegren, 2004).

SSRs are found both in coding and non-coding regions of the genome. However, a depletion of SSRs, with the exception of tri- and hexanucleotide motifs, is observed in coding regions for several eukaryotic species (Li et al., 2004; Metzgar et al., 2000). This is an indication of evolutionary pressure aimed at preventing deleterious frameshift mutations. Trinucleotide repeats are particularly common in coding regions. Excessive expansion of trinucleotide repeats in genes can lead to the production of a faulty protein product, a phenomenon associated with several diseases in humans, referred to as triplet repeat expansion disorders (TREDs). Common examples of TREDs

are Huntington's disease and Fragile X syndrome, both associated with neurological disorders (Budworth and McMurray, 2013).

SSRs were initially considered to be evolutionarily neutral markers. However, later on, several functional roles for these repeats were discovered, leading to the theory that SSRs provide a continued source of quantitative genetic variation, thus acting as drivers of adaptive evolution (Kashi et al., 1997; Kashi and King, 2006). Trinucleotide repeats are more frequent than expected within coding regions, in particular in genes involved in transcriptional regulation (Kozlowski et al., 2010; Young et al., 2000). SSRs in the promoter, intronic, and 5'-UTR regions were found to be actively involved in regulation of gene expression. In some cases, repeat copy number influences the strength of expression. An example is the human gene *PAX6*, for which a nine-fold increase in expression is observed when the number of repeats of an SSR within its promoter exceeds a certain threshold (Okladnova et al., 1998). In other cases, a specific number of motif repeats is required for transcription to happen. This is the case for the *E. coli lacZ* gene: expression will occur only when 12-13 copies of a GAA repeat are present within its promoter, while a lower or higher number of repeats will result in lack of expression (Liu et al., 2000). The length of an SSR locus can also influence the interaction with DNA-binding proteins, enhancing or inhibiting expression (Lue et al., 1989; Winter and Varshavsky, 1989). Other SSR loci within coding regions are implicated in translational regulation (Timchenko et al., 1999) and chromatin organisation (Gao et al., 2013).

Due to their repetitive nature, SSRs are inherently unstable, leading to their observed hyper-variability within populations. The main mutational mechanism for SSRs is 'replication slippage' (Eisen, 1999; Strand et al., 1993). During DNA replication, the lagging strand may shortly separate from the leading strand; when the 2 strands join again, misalignment may occur, creating a loop of one or more repeats on the leading or lagging strand, resulting in a contraction or expansion of the repeat number, respectively. Some of these errors are not corrected by DNA mismatch repair mechanisms or DNA polymerase proofreading, and are thus retained in the replicated sequence. Replication slippage errors are corrected primarily by the mismatch repair pathway (MMR), through recognition, excision and synthesis of the erroneous nucleotides (Liu et al., 2017). The MMR system is highly conserved in all organisms, from bacteria to humans (Jiricny, 2013). In eukaryotes, mismatches are recognised by one of two MutS heterodimers: *MutS*α (*msh2*/*msh6*) and *MutS*β (*msh2*/*msh3*). *MutS*α primarily repairs single-base mismatches and very small (1-3 nt) insertions and deletions, while *MutS*β repairs larger (4-13 nt) insertions and deletions (Jiricny, 2013). The *MutS*

heterodimers then recruit a MutL heterodimer (encoded by the *MLH* genes) and form a complex at the mismatch site, which will perform the repair. Polymerase proofreading is also involved in DNA repair at SSR loci, but it seems to have a more minor role (Eisen, 1999; Strand et al., 1993). In addition to replication slippage, recombination, in particular gene conversion (non-reciprocal transfer of information), is also a source of SSR instability (Jakupciak and Wells, 1999).

When MMR is deficient, errors generated during DNA replication fail to be corrected, leading to genomic instability (Loeb, 2001; Strand et al., 1993). SSR length alterations increase in frequency, leading to a mutator phenotype called microsatellite instability (MSI). Given their role in maintaining genome integrity, MMR genes are considered tumour suppressor genes, and the MSI phenotype is a characteristic of several human cancer types. The first and most known example is colorectal cancer (CRC) (Boland and Goel, 2010). Fifteen percent of CRC present with MSI. While a small fraction of these cases is linked with hereditary mutations in MMR genes, the rest are sporadic cases in which the MMR genes are inactivated, usually through promoter methylation (Lynch et al., 1993; Toyota et al., 1999). Many other cancer types are associated with MSI, including melanoma, gastric, ovarian and lung cancer (Bonneville et al., 2017). The presence of the MSI phenotype has prognostic value: patients with microsatellite-unstable tumours tend to have longer survival compared to those with microsatellite-stable tumours (Choi et al., 2014; Guastadisegni et al., 2010). This apparent paradox could be explained by the fact that MSI tumours, having a higher mutational burden, produce a larger repertoire of immunogenic neoantigens, leading to better response to immunotherapy (Lee et al., 2016; McGrail et al., 2020). Therefore, studying SSR mutations and their phenotypic impact, and understanding the function of the different components of the MMR complex, is of great interest for translational cancer research.

A large number of studies in yeast have described the function of the MMR complex in a wild-type background, and the effect of inactivation of different MMR genes on SSR mutation rate. Initially, these studies were conducted using reporter gene assays (Section 1.3), often combined with strains containing MMR gene mutations. Reporter gene assays led to several observations, such as a bias toward deletions when certain MMR genes are defective and changes in SSR mutation rate due to repeat unit size (Sia et al., 1997; Strand et al., 1993, 1995). MA experiments (Section 1.3) and whole-genome sequencing of the MA strains produced more accurate mutation rate estimations and the ability to observe full SSR mutational spectra in the desired genetic background (Katju and Bergthorsson, 2019). MA experiments have been

used to explore the mutation rate in wild-type and MMR-deficient strains (Haye and Gammie, 2015; Ma et al., 2012; Serero et al., 2014; Zanders et al., 2010; Zhu et al., 2014), leading to the estimation that the MMR system is very efficient, repairing >98% of replication errors (Lujan and Kunkel, 2021).

2.2 Results

2.2.1 Growth rate effects of unidentified mutations

As mentioned in Section 2.1.1, the objective of this project was to identify mutations that could potentially explain the changes in growth rate observed in the apparently mutation-free MA lines from Zhu et al. (2014), with a particular focus on mutations occurring in SSR regions.

As a first step, we assessed two MA lines from Zhu et al. (2014) showing unexplained growth rate changes for the presence of potential unidentified mutations. We reasoned that, if any mutations were present in the two lines, a progeny derived from their cross would carry a random re-assortment of the mutations present in the parents, including potential SSR mutations, and thus exhibit a range of growth rate changes. The two diploid MA lines were therefore sporulated to produce haploids and crossed to produce a diploid, which was again sporulated. Twenty haploid spores were selected. A growth rate assay performed on the cross spores showed a significant difference in their growth rate compared to a group of strains directly derived from the ancestor of the MA experiment (thus not harbouring any mutations, **Figure 2.1**). We also observed a greater variability in the growth rate of the cross spores compared to the ancestor-derived strains, likely determined by unidentified mutations present in the parental strains and randomly segregating in the progeny.

To identify the mutations associated with the observed growth rate changes, we performed whole-genome sequencing on the haploid parental strains, the 20 spores, and a haploid ancestor-derived strain. I then performed mutation calling using Muver, a pipeline specifically designed for mutation calling on data from MA experiments (Burkholder et al., 2018). I identified 4 previously unobserved substitutions in the parental strains of the spores (**Figure 2.2**). I was likely able to identify these mutations due to the higher coverage of our sequencing experiment ($\sim$30x) compared to Zhu et al. (2014) ($\sim$10x). As expected, I observed that these mutations randomly segregated in the cross progeny. Surprisingly, I also found 14 *de novo* mutations across the

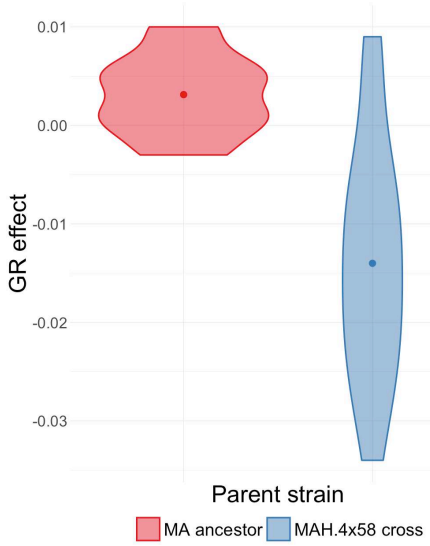


Fig. 2.1 **Growth rate of the cross spores of 2 MA strains from Zhu et al. (2014) with no identified mutations.** Distribution of growth rates measured for 16 cross spores and 15 control strains derived from the ancestor of the MA experiment. Growth rate is relative to one strain chosen from the control group.

20 spores, each private to a single spore, and one mutation that was shared by 3 spores, but observed in neither of the parents **Figure 2.2**). Although these mutations could have arisen during propagation of the diploid strain that underwent sporulation, this is rather unlikely given the small number of generations for which this strain was propagated before sporulation and its reported single-nucleotide mutation rate ($\sim$ 1 mutation/diploid genome/250 generations, Zhu et al. (2014)). Another possible explanation for the 14 private mutations could be that they occurred during sporulation. One possibility was that some of the mutations identified in the parental strains affected the function of genes involved in DNA-repair or maintenance of genome integrity during meiosis. To check for this possibility, I looked at the genomic location of the 4 parental mutations. One of them occurred in the body of the *NMD2* gene, involved in the nonsense-mediated decay pathway and telomere maintenance. Defects in this gene have been shown to increase chromosomal instability (Strome et al., 2008). However, there was no difference in the number of private mutations in spores harbouring or not harbouring the *NMD2* mutation. Therefore, it is more likely that the higher-than-expected number of *de novo* mutations is intrinsic to the meiotic process itself, a hypothesis corroborated by a previous report showing that the meiotic mutation rate in yeast is higher than the mitotic rate (Rattray et al., 2015). Although this hypothesis might explain part of the mutations observed in the cross spores, the meiotic mutation rate reported in Rattray et al. (2015) is still too low to match our observed rate. It is

possible that, in combination with the increased meiotic mutation rate, some mutations were already present in the two haploid colonies used as parental strains, which were propagated for an unknown number of generations before mating and sporulation.

The newly identified mutations in the two MA lines originally sequenced by Zhu et al. (2014) could explain the observed changes in their growth rate. These mutations segregating in the cross progeny, together with the identified *de novo* mutations, also explained some of the variability observed in the cross strains. However, it was still possible that other mutation types, such as SSR mutations, contributed to the effects on growth rate, although the high rate of *de novo* mutations that arises when performing strain crossing complicates the study of these effects.

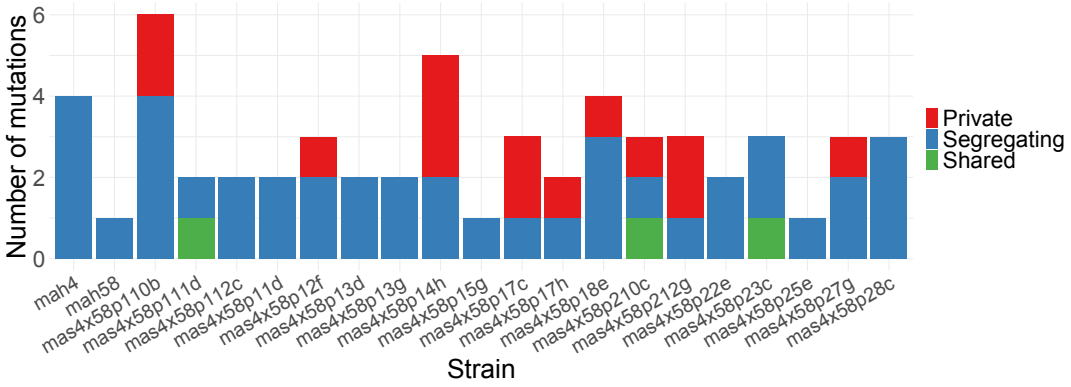


Fig. 2.2 **Number and type of mutations identified across the 20 cross spores.** Mah4 and mah58 are the 2 haploid parents that were crossed to produced the 20 *mas* spores. Mutations are classified as *Segregating*, if they are present in either mah4 or mah58 and passed on to the progeny, *Private*, if they occur in a single spore and are absent from both parental strains; one mutation is labelled as *Shared*, as it is absent from both parental strains, but harboured by multiple cross spores.

2.2.2 Studying SSR mutations in *msh3*-deficient strains

To study SSR mutations in a more isolated context, we decided to take advantage of yeast strains with a defective MMR system, in particular strains missing the *msh3* gene, encoding for one of the components of the *MutSβ* heterodimer. The *MutSβ* complex is primarily responsible for INDEL mutations caused by polymerase slippage, and *msh3* mutants have been previously shown, using reporter assays, to have an increased SSR mutation rate (∼10 times the SNM rate), while the SNM rate appeared not significantly affected (Harrington and Kolodner, 2007; Haye and Gammie, 2015; Sia et al., 2001). We reasoned that propagating *msh3*-mutant strains for ∼200 generations would allow

SSR mutations to accumulate but would keep the probability of SNMs low, thus facilitating the evaluation of the effects of these mutations on growth phenotype.

Moreover, most MA studies investigating the function of the MMR system focused on strains with defects in either both eukaryotic *MutS* complexes (by using *msh2*-deficient strains), or only in *MutS*α (by using *msh6*-deficient strains). By using *msh3* mutants, we would be able to investigate the role of the *MutS*β complex in SSR mutation repair, to estimate the SSR mutation rate in *msh3*-deficient strains and to model its dependency on several characteristics of the SSR locus.

Therefore, we carried out a mutation accumulation experiment on 39 haploid *S. cerevisiae* lines derived from two genetic backgrounds. Five lines were derived from a wild-type (WT) ancestor, carrying no alterations in the MMR complex genes, while the remaining 34 were derived from an ancestor carrying a deletion of the *msh3* gene (*msh3*Δ). All lines were propagated for a total of $\sim$ 200 generations. Bottlenecks were performed every $\sim$ 20 generations by picking a randomly chosen, single-cell-derived colony, and re-streaking it on a fresh plate (**Figure 2.3**). Whole-genome, 150bp paired-end sequencing was then performed on the final MA lines.

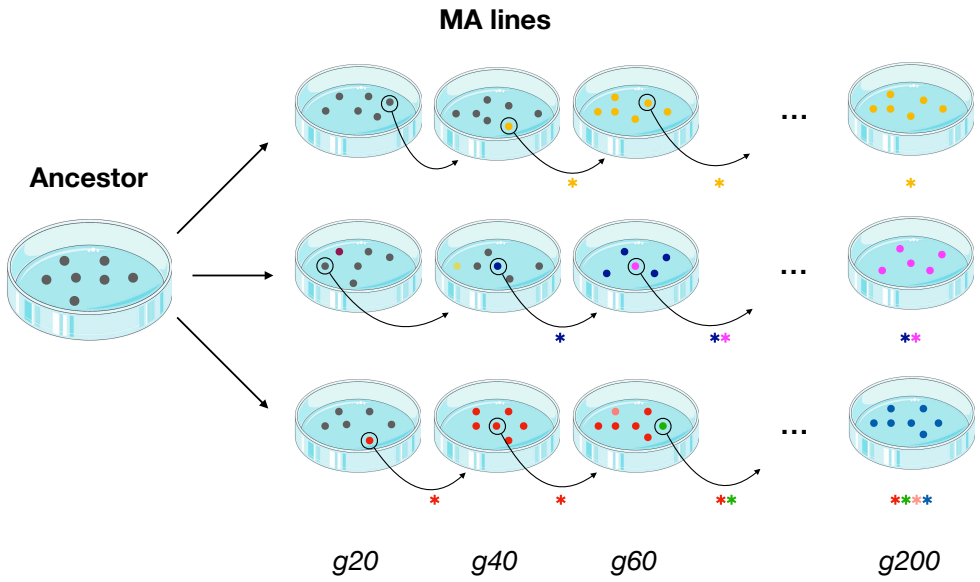


Fig. 2.3 **Schematic of the mutation accumulation experiment.** Multiple lines were derived from a single cell of an ancestor strain (either WT or *msh3*Δ) and propagated in parallel for 200 generations, with single-cell bottlenecks applied by re-plating a single colony every $\sim$ 20 generations. Every dot on the plates represents a single-cell colony; change in colour indicates a colony acquiring a new mutation; asterisks represent mutations carried on to the new plate after a bottleneck.

2.2.3 SSR loci in the ancestor strain genome

To be able to correctly call mutations inside and outside simple sequence repeats, we first created a reference list of SSR loci, by searching for all 1-4 bp motifs repeated in tandem at least 3 times in the genome of the ancestor strain of the mutation accumulation experiment (Section 2.4). Penta- and hexanucleotide repeats are rare in *S. cerevisiae*'s genome (Karaoglu et al., 2005), and previous studies similar to the one described in this chapter also focused on SSRs with motifs shorter than 5bp (Lang et al., 2013; Lujan et al., 2015; Lynch et al., 2008). We identified 270,796 SSRs in the nuclear genome of the ancestor, with the majority of loci being mononucleotide repeats (**Figure 2.4a**). Accurate genotyping of SSR loci requires the sequencing reads to span the entire length of the locus. The mean length of all repeats was 5.1bp, with just 2.6% of loci longer than 10bp. Only 9 loci were longer than 100bp, making genotyping feasible for almost all SSRs in the ancestor's genome using 150bp reads (**Figure 2.4b**).

Most of the identified SSR loci (68%) were within gene regions. This number is in line with the fraction of *S. cerevisiae*'s genome in gene regions (∼73%, (Alexander et al., 2010)) However, the proportion of SSRs in genes was dependent on motif length and on total length of the repeat (**Figure 2.4c**). In particular, trinucleotide repeats were more frequent than non-trinucleotide repeats inside genes (χ^2 test $P <$ .001). Furthermore, SSRs with repeat copy number of 10 or higher were less frequent than expected within genes, but this trend was only observable for non-trinucleotide repeats (χ^2 test $P < .001$). This bias against non-trinucleotide repeats in gene regions is consistent with the presence of selection against potentially damaging frameshift mutations, and was observed before by Metzgar et al. (2000).

2.2.4 Single-nucleotide mutations and short INDELs outside SSR loci

I next performed mutation calling on all WT and *msh3*Δ MA strains and their ancestor strains to identify mutations outside repetitive regions. Using Muver, I identified 28 substitutions in non-SSR regions (Section 2.4). Two of these mutations were in one WT strain, the remaining 26 occurred across 17 *msh3*Δ strains. We also called mutations using FreeBayes, a classical variant caller (Garrison and Marth, 2012), using the genome of the strain from which both MA ancestors were derived as a reference. We called mutations in the MA strains by selecting loci at which the genotype of the MA strain was different from the genotype of its ancestor. We observed a very good

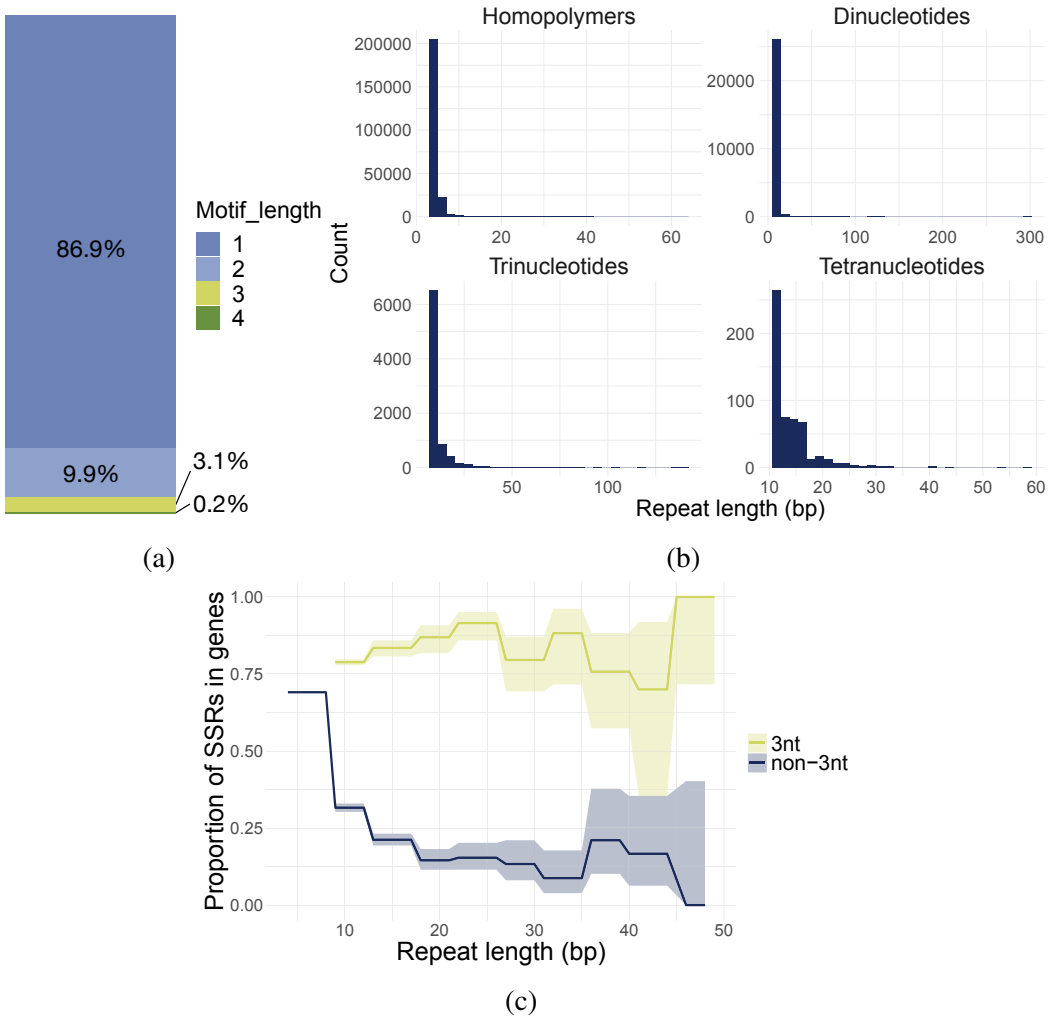


Fig. 2.4 **SSRs in the ancestor strain genome. (a)** Proportion of SSRs with different motif sizes; **(b)** Distribution of total repeat length for SSRs with different motif sizes; **(c)** Proportion of trinucleotide and non-trinucleotide SSRs in coding regions of the genome, as a function of total repeat length.

overlap between Muver's and FreeBayes' results. After filtering out low-confidence mutations and mutations in repetitive regions, including SSRs, FreeBayes identified 27 SNMs (Section 2.4). Of these, 2 were the same mutations identified by Muver in one of the WT strains, and 24 were the same identified by Muver in the *msh3*Δ strains. FreeBayes identified one additional substitution in an *msh3*Δ strain. It also identified and 5 INDELs in non-SSR regions; however, upon closer inspection, these mutations were in fact falling within SSR loci not present in our SSR reference list, likely due to them having a very small number of motif repeats or a motif length greater than 4bp. All mutations identified by both tools, the additional mutation identified by FreeBayes, and the 2 mutations identified by Muver but not by FreeBayes were tested via Sanger

sequencing. The 2 mutations identified by Muver were revealed to be erroneous calls. We thus considered the final non-SSR mutation callset to be the 26 mutations identified by Muver and the one additional mutation identified by FreeBayes.

To test whether the single-nucleotide mutation rate in *msh3*Δ strains significantly differed from that of WT strains, we ran a logistic regression model predicting mutation status for each position in the genome, in each of the MA strains, based on *msh3* status of the strain (deleted or WT). We then compared this model with an intercept-only model. There was no significant difference between the two models, and thus between the WT and *msh3*Δ SNM rates.

The overall substitution rate across all strains was 3.5×10^{-10} mutations/bp/generation (95% CI: 2.3-4.9 $\times 10^{-10}$) or $\sim$1 mutation/haploid genome/250 generations.

These results are consistent with those of the reporter assay studies of Sia et al. (2001) and Harrington and Kolodner (2007) showing that the substitution rate is not affected in *msh3*-deficient strains. They also confirm that the *MutS*β complex plays a minor role in the repair of single-nucleotide mutations.

2.2.5 Calling mutations within SSR loci

As mentioned in the introduction to this chapter, calling mutations in SSR regions is not a trivial task. The repetitive nature of these regions increases the chances of errors in the alignment of the sequencing reads to the reference genome. Moreover, due to the main mutational mechanism in SSR regions being DNA polymerase slippage, errors are also likely to occur during DNA sequencing itself, at the *in vitro* amplification step, producing reads differing in length from the original sequence, a phenomenon known as "PCR stutter" (Ellegren, 2004). For this reason, we decided to test different tools to call mutations in SSRs, to eventually choose the most suitable one for our purposes.

I first called mutations at all SSR loci in the genome of the MA ancestor strain by using Genome Analysis Toolkit (GATK) germline short variant discovery pipeline (Poplin et al., 2018). Although this pipeline is not specifically designed to call mutations in tandem repeats, its "joint genotyping" mode allows potential mutations to be jointly called across all input samples, which increases sensitivity over low coverage regions and improves filtering accuracy (Poplin et al., 2018). As for the non-SSR mutation calling, I used the genome of the strain from which the WT and *msh3*Δ MA ancestors were derived, and called a mutation when the genotype of the MA strain differed from that of its ancestor. After filtering out variants with poor genotype quality

(Section 2.4), GATK identified 682 SSR mutations, 95 across the WT strains and 587 across the *msh3*Δ strains. Of these 682, 125 mutations occurred in a single MA strain. The remaining 557 mutations occurred across 71 loci, and in all cases the same locus was mutated in multiple MA strains. Given that all MA lines were propagated independently, the occurrence of mutations in multiple lines at the same locus is highly unlikely. This suggests that shared mutations identified by GATK are the product of unreliable calls, and should not be considered in the final callset. To check for the accuracy of the 125 private mutations identified, I visually inspected a random sample of 50 mutations using IGV (Integrative Genome Browser). Only 4/50 inspected calls appeared to be supported by the data. These observations indicate that the SSR mutation calling results produced by GATK are unreliable and inaccurate, and thus not suitable to our aims.

Next, I decided to perform SSR mutation calling using MSIsensor, a tool designed to detect microsatellite instability caused by replication slippage in tumour samples compared to matched normal samples (Niu et al., 2014). I ran MSIsensor using on all possible pairs of ancestor-MA strain, specifying the MA strain as the "tumour" sample and the ancestor as the "normal" sample. MSIsensor identified no SSR length changes (INDEL mutations) in the WT strains, and 58 mutations at 56 loci in the *msh3*Δ strains. Most mutations were private to one strain, with only 2 mutations shared by two strains. Upon visual inspection in IGV, all but 2 mutations appeared to be supported by the data. The 2 mutations shared by 2 strains were also present upon IGV inspection. This might be a case of shared ancestry: a mutation might have arisen in one of the few initial cell divisions of the ancestor, and colonies harbouring the same mutation might have been picked to generate 2 of the MA lines, resulting in them sharing the mutation. While MSIsensor's SSR mutation calls appear to be reliable, this tool has some limitations. Since its intended purpose is to identify microsatellite unstable tumour samples, and not strictly to produce variant calls, some useful information is missing from the output, such as the number of repeat units inserted or deleted at the SSR locus in the MA strain. For the same reason, the only mutations identified by MSIsensor are expansions and contractions of SSRs, excluding possible single-base substitutions within SSR loci, which are also of interest for this project. Moreover, MSIsensor only considers perfect SSR loci, where the repeated stretch of motif units is uninterrupted, thus missing information about possible mutations occurring in imperfect repeats.

For these reasons, we decided to also look at SSR mutations called by FreeBayes. FreeBayes is able to call INDEL mutations, including those occurring within SSRs, although it is not specifically tailored to call mutations in tandem repeats. However,

we reasoned that FreeBayes would provide the information missing from MSIsensor's output, and that a comparison with MSIsensor's results would increase our confidence in the calls produced by FreeBayes. We also set up a custom filtering strategy to retain high-confidence SSR mutation calls from FreeBayes's results. The extent of PCR stutter errors directly correlates with repeat number and inversely correlates with motif length (Shinde et al., 2003). This means that longer loci, which are more likely to be mutated *in vivo*, will also have a larger stutter noise, leading to lower-confidence mutation calls. Moreover, the fact that accurate SSR genotyping requires reads to span the entire repeat contributes to the lower confidence calls for longer loci, since these loci are more likely to have fewer spanning reads. Thus, using the same filtration cutoff for all SSR loci would lead to bias against naturally more mutable loci. In our custom filtration strategy, we grouped SSR loci based on shared properties which have been shown to affect the SSR mutation rate, such as length of the locus and AT proportion, and applied appropriate confidence score cutoffs to each group (see Section 2.4 for a detailed description of the filtration strategy).

After filtering, we found 35 SSR mutations, distributed across 18 *msh3*Δ strains (0-5 mutations/strain). One of the mutations was shared by 2 strains. As explained above, this might be a case of shared ancestry. All identified mutations were confirmed by visual inspection in IGV. No mutations passed filtration in WT strains.

We then compared the final mutation callset obtained by using FreeBayes and our custom filtration strategy with the mutations identified by Muver and MSIsensor within SSR regions. Of the 35 mutations called by FreeBayes, 29 were identified by Muver and 26 by MSIsensor. There were thus 6 and 9 SSR mutations identified by FreeBayes but not MSIsensor and Muver, respectively. These differences might be due to the different methods and filtering steps of the different tools. Indeed, all of the 9 mutations not called by MSIsensor were in loci excluded by MSIsensor's callset for technical reasons, such as the presence of an imperfect repeat at the locus (which, as mentioned above, is not supported by the tool) and 2 of the mutations not identified by Muver were in regions filtered out because of insufficient coverage according to Muver's algorithm. Finally, there were 62 and 31 mutations identified by Muver and MSIsensor, respectively, (and confirmed visually with IGV) but not FreeBayes. This might be a result of the more stringent filtering applied to FreeBayes' results. Although very stringent and possibly leading to discarding several real mutations, our custom filtration strategy is based on FreeBayes' confidence scores for calls at all SSR loci, including both the reference and alternative allele calls. Thus, it provides us not only with the number of high-confidence mutation calls at a certain threshold, but also

with the total number of SSR loci for which there is enough confidence to assign a genotype, irregardless of whether the genotype is reference or alternative. So, while not all SSR mutations occurring during MA will be included in the final callset, we will have high confidence in the ratio of detected mutations over total calls, leading to consistent, stable estimates of mutation rate that are not very sensitive to the choice of threshold (Section 2.4). Since our goal was to correctly estimate the SSR mutation rate, we decided to continue our analysis using the more stringent set of calls produced by FreeBayes.

2.2.6 Rate and spectrum of SSR mutations in *msh3*Δ strains

We then sought to estimate the SSR mutation rate in *msh3*Δ strains, and identify the properties of SSR loci that contribute most to the mutation rate. To this aim, we built a logistic regression model predicting mutation probability at each SSR locus based on AT proportion and motif length. Previous studies of *msh2* mutants have shown that replication errors in short SSRs are usually corrected by polymerase proofreading rather than the MMR system (Lujan et al., 2015). Thus, we added an additional binary variable to our model, indicating whether an SSR was longer than 7bp (Section 2.4). We did not include the 5 INDELs identified in unannotated SSR loci in this model (Section 2.2.4). Since we did not have a complete list of loci with similar properties (e.g. with motif size > 4) we would have not been able to apply the same filtering criteria as for the other SSR mutations, which could have led to incorrect mutation rate estimate.

The only property with a significant effect on the odds of mutation was whether the SSR was longer than 7bp: we observed a 32-fold increase in the odds of observing a mutation in loci longer than 7bp (**Table 2.1**). Using this model to predict the mutation rate for each SSR locus in our reference list gave us an estimate for the mutation rate in *msh3*Δ strains of $\sim$1 mutation/genome/120 generations. Because no mutations were identified in WT strains, not enough data was available to estimate the mutation rate in these strains. We thus calculated a lower bound (95% CI bound) for the *msh3* status effect on mutation rate, which is a 1.8-fold increase in *msh3*Δ compared to WT strains (**Table 2.1**, Section 2.4). This indicates that defect in *MutS*β function significantly increase the mutation rate at SSR loci.

Next, we looked at the spectrum of SSR mutations accumulated in *msh3*-deficient strains. **Figure 2.5** shows the distribution of motif-number insertions and deletions found in homopolymer, dinucleotide and trinucleotide repeats (no mutations were

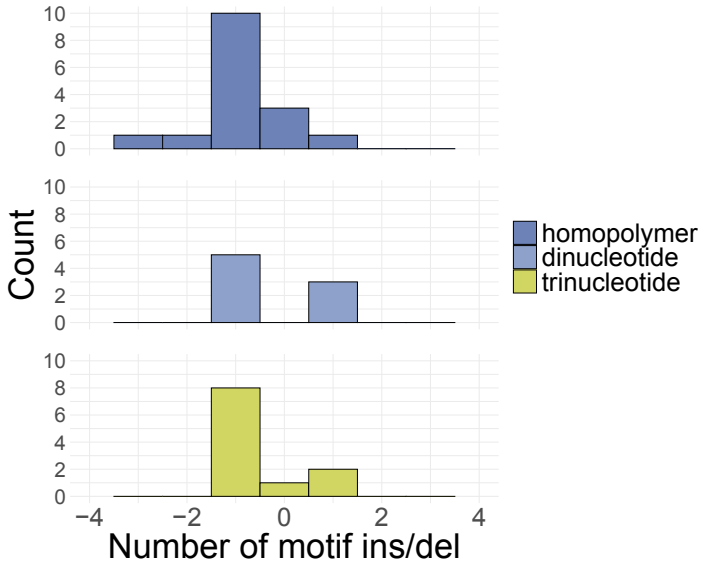
	p-value	fold-change in odds of mutation
msh3Δ	0.003	17×10^6 (1.8-Inf)
SSR >7 bp	<0.001	32 (14-94)
+1 bp motif length	0.5	0.9 (0.6-1.3)
proportion A/T	0.27	0.5 (0.1-1.9)

Table 2.1 **Effects of SSR properties and msh3 status on the odds of mutation.** Fold changes in the odds of mutation, with associated p-values and 95% confidence intervals, for msh3 status and various properties of SSR loci included in our model. Because no SSR mutations were found in WT strains, we calculated a lower bound for the effect of msh3 status; the msh3Δ coefficient reported in the table is thus is arbitrarily large, and has no upper bound.

identified in tetranucleotide repeats). A bias towards deletions can be observed for all motif lengths, with the most frequent event being deletion of a single motif copy. The increase in deletion rate compared to insertion rate, however, was only significant for homopolymers ($P = .002$) and borderline for trinucleotides ($P = .055$). We also observed 4 instances of substitution mutations occurring within SSR loci, but the estimated difference in substitution rate in SSR and non-SSR loci was not significant.

I then looked at the genomic locations where SSR mutations occurred in the *msh3*Δ strains. Given that during the MA experiment only weak selection should occur, due to stringent bottlenecking, we expected to find no bias against mutations falling within coding regions. Since ∼73% of *S. cerevisiae*'s genome is genic, we should expect a similar proportion of the occurring mutations to fall into gene regions. However, Zhu et al. (2014) observed a bias against INDELs in gene regions, suggesting that some INDELs are deleterious enough to still be eradicated under reduced selection. I also observed that SSR mutations are less likely than expected to occur within gene regions, with only 44% (15/34) falling within genes (χ^2 test $P < .001$). I also tested whether the frequency of genic mutations was influenced by motif length. Of the 14 SSR mutations falling within genes, nine were in trinucleotide repeats, 5 in homopolymers and 1 in a dinucleotide repeat. The bias was only present for non-trinucleotide mutations (χ^2 test $P < .001$), while trinucleotide mutations had the same frequency inside and outside gene bodies (χ^2 test $P = 0.91$). However, as mentioned in Section 2.2.3, the proportion of non-trinucleotide SSRs in genes depends on the total length of the repeat, and on average only ∼25% of non-trinucleotide SSRs longer than 10bp fall within genes

(**Figure 2.4c**). This proportion is close to the proportion of mutations we observe in non-trinucleotide SSRs and in genic regions (6/21). Since loci longer than 7bp are the most likely to mutate (as shown above), this dependency of the SSR genic proportion on repeat length likely explains the observed bias against INDELs in genic regions.



(a)

Fig. 2.5 **Spectrum of SSR mutations in *msh3*Δ strains.** Mutations with 0 motifs inserted or deleted represent substitutions within SSR loci.

2.2.7 Growth rate changes associated with SSR mutations

We then sought to investigate the phenotypic effect of SSR mutations in the MA lines. We selected 14 strains for which no SNMs were identified. During the MA experiment, samples of the strains were collected and frozen every 20 generations. This allowed us to select 4 additional strains at generation 100, when they had not yet acquired those SNMs. We then performed a growth rate assay on these 18 lines to test for changes in their growth rate compared to their ancestor (Section 2.4). In the absence of any known SNMs, potential changes in growth rate in these lines must be caused by other mutation types, very likely SSR mutations.

Out of the 18 lines, the majority (13 lines) showed a significant change in growth rate compared to the ancestor (**Figure 2.6**). Given the small number of mutations per strain, this suggests that most SSR mutations are non-neutral. The majority of the observed changes were decreases in growth rate, with a mean effect of -0.01, which indicates that SSR mutations have mostly small deleterious effects.

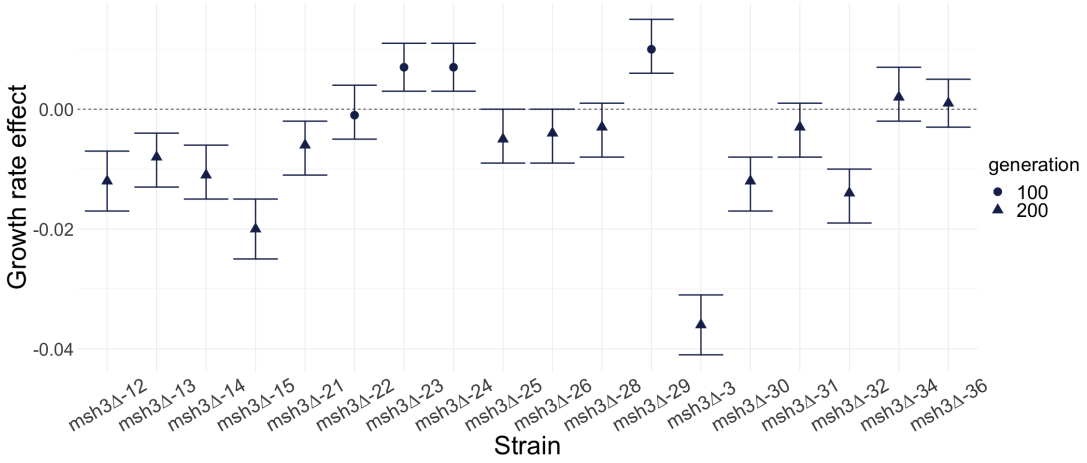


Fig. 2.6 **Phenotypic effects of SSR mutations.** Growth rate changes of 18 MA lines with no SNMs compared to their ancestor (dotted line); four lines (round points) were derived from frozen colonies isolated at generation 100 of the MA experiment, at which point they had not accumulated any SNMs.

2.3 Discussion

In this chapter, I presented work aimed at identifying the full spectrum of spontaneously occurring mutations in a yeast model, and the effects of these mutations on growth phenotype. In particular, the focus of this chapter was on simple sequence repeat (SSR) mutations, a class of mutations that is usually excluded by similar studies due to the challenges associated with sequencing and genotyping of repetitive regions.

In order to study spontaneously occurring SSR mutations in relative isolation, we performed a mutation accumulation experiment in *msh3*-deficient *S. cerevisiae* lines. This allowed the strains to accumulate mutations in SSR regions faster than WT strains, as *msh3* is a component of the *MutS*β heterodimer, involved in eukaryotic mismatch repair system. It also allowed us to explore the role of *msh3* in the repair of mutations occurring in SSR regions. Before this, only one study investigated the effects of defects in the *msh3* gene on SSR mutation rate using a mutation accumulation experiment combined with whole-genome sequencing (Haye and Gammie, 2015); this study, however, was performed on a single *msh3*-deficient MA line.

We identified single-nucleotide mutations outside SSR regions, and compared mutation rate in *msh3*-deficient and wild-type lines. We found no significant difference in mutation rate, confirming the results of previous reporter assay studies, reporting that the *MutS*β complex, unlike *MutS*α, is primarily involved in the repair of short INDELs, but does not play a major role in the repair of substitutions (Harrington and

Kolodner, 2007). Our estimated substitution rate across all MA lines was 3.5×10^{-10} mutations/bp/generation ($\sim$ 1 mutation/haploid genome/250 generations), which is consistent with a previously reported estimate of the haploid mutation rate in yeast (Sharp et al., 2018).

We then focused on the identification of SSR mutations arising in the MA lines. We estimated the SSR mutation rate in *msh3*-deficient strains to be $\sim$1 mutation/halploid genome/120 generations. We observed a bias towards deletions in the identified mutations, which was significant for homopolymers and had borderline significance for trinucleotide repeats. This bias was observed for *msh3*-deficient strains in a reporter assay study (Strand et al., 1995) as well as in a study using oligonucleotide transformation to introduce repeat INDELs at one locus (Romanova and Crouse, 2013). *Msh2*-deficient strains have also been shown to exhibit a similar bias towards deletions, but only in homopolymer repeats (Lang et al., 2013).

We also observed a very strong effect of repeat length on mutation rate, with loci longer than 8bp having a 32-fold increase in mutation rate. A similar pattern was observed by Lujan et al. (2015) in *msh2*-deficient strains. With an MA experiment, the authors showed that loci longer than 10bp are much more affected by MMR defects (*msh2*-only or *msh3+msh6* deletions) than shorter loci. Additionally, they demonstrated that mutation rate in shorter loci increases when an additional defect in polymerase proofreading function is introduced. Their results suggest that INDELs at shorter loci are repaired by polymerase proofreading, while longer loci require the MMR system to be repaired. Our results support this model, and show that even the abrogation of the sole *MutS*β complex reduces repair efficiency at longer loci.

This is one of the only studies assessing the phenotypic effects of SSR mutations in MA lines. We measured growth rate in strains free of single-nucleotide mutations and compared them to their ancestor, and showed that most SSR mutations are likely non-neutral, and have small deleterious effects. This result has interesting implications when read in the context of cancer biology. As mentioned in the introduction to this chapter (Section 2.1.2), microsatellite-unstable cancers are usually less aggressive and have better prognosis than microsatellite-stable cancers, a phenomenon that is hypothesised to be linked to increased neoantigen load and consequent enhanced anti-tumour immune response. Another factor contributing to the less aggressive phenotype of MSI cancers could be the combined effect of the several mutations accumulated in a tumour with MMR deficiency. While a few of these mutations might affect the function of oncogenes and tumour suppressors, the majority will likely be passengers. Our results indicate that they would likely be mildly deleterious passengers. As discussed

in McFarland et al. (2013), the joint effects of these passengers might counter-balance the tumour-promoting effects of driver mutations, and slow down tumour progression.

One factor to consider when interpreting our results is the ploidy of our MA strains. Our estimated SSR mutation rate, and the observed effects on growth phenotype, refer to haploid strains. Sharp et al. (2018) showed that substantial differences exist in the rate, spectrum and effects of spontaneously accumulated mutations between haploid and diploid *S. cerevisiae* strains. Haploid strains were more prone to SNMs, while diploid strains were more prone to structural changes, and mutational effects appeared to be more deleterious in diploids compared to haploids. It is therefore possible that diploid *S. cerevisiae* strains would exhibit different spectrum and effects of SSR mutations compared to those described in this chapter.

Finally, the focus of the work by Zhu et al. (2014), the subsequent analysis conducted by the Siegal Lab, and this chapter was on changes in the DNA sequence of the MA strains. However, previous studies have shown that epigenetic changes also spontaneously accumulate during MA experiments (van der Graaf et al., 2015). It thus cannot be excluded that the observed changes in growth rate described in this chapter and in previous MA studies are at least partially due to epigenetic changes.

2.4 Methods

Derivation and analysis of MAH.58x4 cross spores

To construct the cross to assess the effects of unidentified mutations on the MA lines from Zhu et al. (2014), haploid strains MAH.58 and MAH.4 were mated. Diploids derived from this cross were grown and sporulated, and the deriving tetrads were then dissected. One SNM segregating in this cross was previously identified in the MAH.4's diploid parent (ChrIV.831520C>T). This SNM, as well as the mating type locus, were genotyped within each tetrad, and only complete tetrads segregating all three alleles for both loci were used. A single a spore per tetrad, without the SNM on ChrIV, was selected, for a total of 20 spores. Growth rate assay was performed on 16 of these spores and 15 control strains derived from the ancestor strain. Sequencing was performed on all 20 spores (see below).

After quality control with FastQC (FastQC, 2010), mutation calling was performed using Muver (see below).

Mutation accumulation experiment

The parent strain of the MA experiments is a haploid line derived from a single spore of the MA ancestor, described in Hall et al. (2008); Joseph and Hall (2004), with genotype *ade2*, *lys2-801*, *his3-ΔD200*, *leu2–3.112*, *ura3–52*, *ho*. A single colony derived from this ancestor founded strain s.EP049, which is the WT ancestral strain used in this study.

To construct the *msh3*Δ ancestor, s.EP049 was transformed with a linear construct containing homology upstream and downstream of the *msh3* gene flanking a Hygromycin/ 5-fluorodeoxyuridine positive/negative selection cassette (Alexander et al., 2014) that was flanked by two 50-bp internal homology sites; spontaneous recombination between these sites results in the excision of the selection cassette, leaving behind a single copy of the internal homology and a Cyc1 terminator sequence. *msh3*Δ transformants were selected on hygromycin, genotyped, and re-selected on 50 µg/mL 5-fluorodeoxyuridine to select for strains with the selection cassette removed. The resulting *msh3*Δ::Cyc1T strain, founded by a single colony, was designated s.EP060.3.

To perform mutation accumulation, single YPD(Yeast Extract Peptone Dextrose)-grown colonies of s.EP049 and s.EP060.3 were re-streaked on YPD; each resulting colony founded a single MA line, with 5 WT lines and 36 *msh3*Δ lines. Respiring *ade2* mutant colonies have a pink tint after two days of growth on YPD, developing a distinct red colour after an additional two days. As in Hall et al. (2008); Joseph and Hall (2004), we used this colour difference to ensure selection of respiring non-petites during mutation accumulation. Petites are lines with mutations affecting aerobic respiration and thus unable to grow on non-fermentable carbon sources. To facilitate this visual discrimination, a single WT petite line was passed through mutation accumulation alongside the others as a reference. Each transfer was performed in duplicate: to avoid unconscious bias in the subculture procedure, the two pink colonies that were closest to a pre-marked spot on the plate were chosen at every passage. Colonies of each line were transferred every two days, and both the two colonies that were re-streaked were frozen in 50% YPD, 15% glycerol. The two colonies from each line used at each transfer were designated as ‘primary’ and ‘secondary’, and only the ‘primary’ re-streak of each line was used in the following transfer, except in cases when it turned out to be a petite, in which case a colony from the ‘secondary’ streak was used. A single line was petite in both the primary and secondary transfer near the last generation, and was removed from further analysis. One additional MA line was excluded due to potential contamination. After each transfer, a mixture of many yeast colonies from the transfer plate was also frozen in 50% YPD, 15% glycerol. Transfers were interrupted for 3

months after the first 6 transfers (∼120 generations) due to the closing of the university as a result of the COVID pandemic. After re-opening, mutation accumulation was re-started from frozen whole colonies. A total of 10 transfers (∼200 generations of mutation accumulation) were performed.

Sequencing

Cultures derived from single frozen colonies from the final MA transfer, as well as from ancestral MA strains, were grown in SC media and DNA extraction was performed as in Schwartz and Sherlock (2016). Whole-genome sequencing, with 150bp paired-end reads, was performed on 34 *msh3*Δ and 5 WT MA lines.

For both the MA lines and the 20 MAH.58x4 cross spores, Nextera library preparation was performed as in Baym et al. (2015), but with 14 PCR cycles instead of 13. Bead cleanup was modified to optimise selection of 500-600 bp fragments: libraries were initially incubated with 0.53x volume AmpPure beads. Supernatant was saved, beads were washed with water, and then supernatant was incubated with original beads + 0.1x original volume AmpPure beads, followed by washing beads with 75% Ethanol and elution of DNA in 10 mM Tris pH 8, 1 mM EDTA, 0.05% Tween-20.

Quality control on the raw sequencing data was performed with FastQC.

Identification of SSR loci in ancestor genome

To identify SSR loci, Tandem Repeat Finder (TRF) (Benson, 1999) v4.09 was run with suggested parameters, except minimum alignment score, which was set to 3. TRF fails to identify a large number of short SSRs; therefore, a genome-wide string search for homopolymers with 4-10 repeats, and di- and tri-nucleotides with 3-10 repeats was also performed. The results of this search were joined with TRF results. Repeats were filtered as suggested in Willems et al. (2017). Finally, any overlapping loci with non-identical motifs were split in such a way as to maximise the combined alignment score of the two motifs.

SNM and SSR mutation calling with FreeBayes

An ancestral reference genome was built by incorporating the mutations identified in the MA ancestor strain in Zhu et al. (2014) in the S. cerevisiae S288C reference genome. Sequencing reads were aligned to this reference using bwa-mem (v0.7.17) (Li, 2013) and duplicate reads were removed using GATK (Van der Auwera and O'Connor, 2020).

FreeBayes (v1.3.4) (Garrison and Marth, 2012) was initially run with the default parameters (except the parameter *min_mapping_quality*, which was set to 1). All calls

with a QUAL value greater than 1 were used in downstream analysis. The output of FreeBayes was a list of loci (including SSR loci) that were called as mutated in at least one MA strain compared to the reference genome.

In order to get likelihood values for non-mutant SSR loci, loci called as mutant were removed from the full list of SSRs; the list was then converted to a VCF, with the alternative allele at each locus listed as a missing value. FreeBayes was then re-run to obtain calls at these loci: the non-mutant SSR list was provided as input for the *–variant-input* parameter and specified *–only-use-input-alleles*, *–min-alternate-count* 0, *–min-alternate-fraction* 0, and *–min-coverage* 0. In this mode (and in the absence of a provided alternative allele), FreeBayes evaluated the likelihood of each unmutated SSR locus as compared to a version of the locus one motif repeat shorter than the original. The resulting calls and likelihood values were joined with the list of calls from the initial round of FreeBayes analysis.

Any call identified within 100 base pairs of a telomere, centromere, or LTR transposon, as well as calls falling inside the rDNA-containing regions of chromosome XII, and in the mitochondrial genome, were removed. Only calls sequenced with a read depth of at least 10x were retained. A small number of mutations falling in non-SSR repetitive regions and having low call confidence (differences <20 between log-likelihood of top calls) were also removed. Some non-SSR mutations passing filtration were found in multiple MA strains, or in very close loci across multiple strains (within 50bp distance). These events are highly unlikely to occur, given that all lines are propagated independently during the MA experiment. These calls were further examined by grouping SNMs found within 50bp and counting the number of strains with mutations in these regions, allowing for the identification of recurrently mutated regions. Unlike mutations found in a single strain, all these loci had a mix of reads supporting two different alleles, with the proportion of reads supporting the not-called alleles being >25% of the total mapped reads. This observation suggests that these calls are unreliable; they were thus removed from the final callset.

For SSR mutations, we applied a custom filtration strategy (**Figure 2.7a**). One of FreeBayes' measures of confidence in its calls is reported as the difference in genotype likelihood between the two most likely alleles (ΔGL). We thus set up filtering thresholds based on the distribution of ΔGL values for each group of SSR calls with similar properties. We grouped SSR loci based on three properties: motif length, repeat copy number, and AT-proportion, which have all been shown to impact SSR mutation rate (Lang et al., 2013; Lujan et al., 2014). To avoid SSR groups with small numbers of calls, for each unique motif copy number we considered a window

of -2.5 - +2.5 repeats around the actual copy number. Moreover, we removed SSR groups including less than 25 loci. For each ΔGL distribution, we calculated quantile thresholds corresponding to the removal of 0 to 95% of all group calls. For each given threshold, we merged the calls passing the threshold in each group and estimated the mutation rate, producing estimates at different filtration stringency levels. While at very stringent ΔGL thresholds the low number of mutations passing filtration led to unstable estimates of mutation rate, for thresholds corresponding to the removal of 35-60% of all calls the calculated mutation rate was largely stable (**Figure 2.7b**); this shows that our filtration strategy is not very sensitive to the choice of threshold above a certain ΔGL value. We thus selected the ΔGL value corresponding to the lowest stringency in the stable range, and removed the 35% of calls with the lowest ΔGL values in each SSR group.

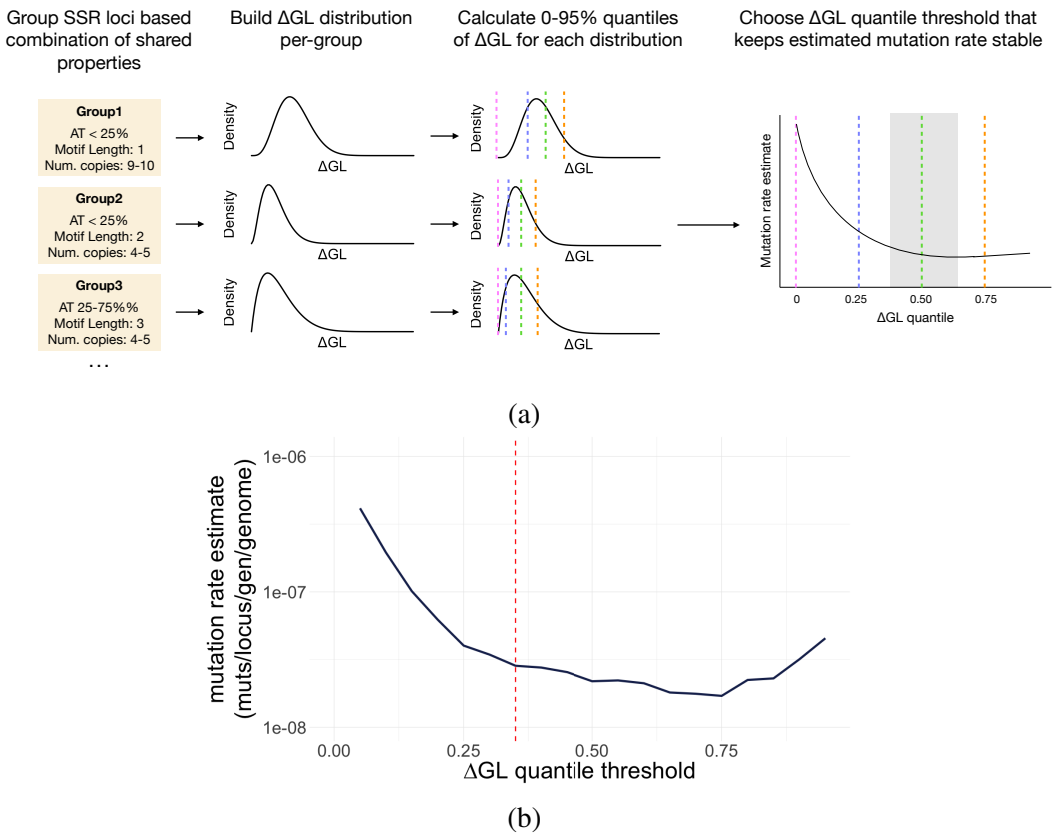


Fig. 2.7 **Filtration strategy for SSR mutation calls.(a)** Overview of the filtration strategy; **(b)** Estimated mutation rate at all possible ΔGL quantile thresholds; the red dashed line represents the chosen threshold, corresponding to the removal of the lowest 35% of calls in each SSR group.

Mutation calling with Muver

In addition to FreeBayes, mutation calling was performed for all MA strains using Muver. Muver's pipeline includes an alignment step, performed with Bowtie2 (version 2.3.5), and a variant calling step, during which SNMs and small INDELs are identified using GATK (version 3.8) (Langmead and Salzberg, 2012). Muver allows the user to indicate the ancestor strain of the experiment and, in a final step, calls mutations occurring in the MA strains compared to the ancestor. Muver was run on the WT and *msh3*Δ strains separately, specifying s.EP049 and s.EP060.3 as the ancestor strain, respectively. Muver's results were filtered to exclude mutations called in low mappability regions of the genome, including centromeric and telomeric regions and LTRs. Mutations occurring on mitochondrial DNA were also filtered out. Among the WT strains, only C3 had 3 mutations compared to the ancestor, while Muver identified 117 mutations across the 34 *msh3*Δ strains. 91 of the called mutations fall into loci that we classified in our analysis as SSRs, leaving 26 non-SSR loci mutated in the *msh3*Δ strains.

SSR mutation calling with GATK and MSIsensor

To call mutations in SSR regions with GATK (v4.2.1), sequencing reads were first aligned to the ancestral reference genome using bwa-mem2 (v2.2.1) (Vasimuddin et al., 2019). The resulting BAM filed were further processed using picard (v2.25.7) (Pic, 2019): they were sorted sorted with *SortSam*, duplicates were marked with *MarkDuplicates*, read groups were added with *AddOrReplaceReadGroups* and an index was built with samtools (v1.9) *index*. *HaplotypeCaller* was ran in GVCF mode (joint genotyping), providing our SSR reference list as the intervals in which to perform variant calling, followed by *GenomicsDBImport* and the final *GenotypeGVCFs* step. The ploidy of all samples was specified as 1 (haploid strains). The resulting VCF file was filtered using *VariantFiltration* to remove low confidence INDELs according to the GATK best practices (QD < 2.0, FS > 200.0, ReadPosRankSum < -20.0, SOR > 10.0).

In addition, mutations in SSR regions were identified with MSIsensor (Niu et al., 2014), a tool designed to identify microsatellite instability in paired tumor-normal sequence data. First, the *scan* command was used to identify all SSR regions in the reference genome, with the following parameters: -l 4, -m 50, -r 3, -s 4. MSIsensor only considers perfect repeats. The *msi* command was run on all possible ancestor-MA strain pairs for both the WT and *msh3*Δ strains, specifying the following parameters: -c 15, -l 4, -p 4, -m 120, -q 3, -s 3, -w 120, -f 0.1.

The resulting files for both tools were filtered to remove loci within LTRs, telomeres, centromeres, and rDNA repeats on Chrom XII, as well as mutations occurring on mitochondrial DNA, using bedtools (v2.29.2) (Quinlan and Hall, 2010).

Modelling of SSR mutation rate

All modelling was performed using a binomial generalised linear model in R.

We modelled the per-SSR-locus odds of mutation as a per-base pair odds of mutation multiplied by the length of the SSR. Per-base-pair odds of mutation are based on three locus properties: AT proportion, motif length and a binary variable indicating whether the total repeat length is smaller than 8bp (Equation 2.1). Since no mutations were found in the WT strains, this modelling was performed on data from the *msh3*Δ strains only.

$$\log\left[\frac{P(\text{mutation})}{1-P(\text{mutation})}\right] = \alpha + \beta_1(\text{AT\_proportion}) + \beta_2(\text{motif\_length}) + \beta_3(\text{short\_ssr}) + \text{offset}(\log(\text{repeat\_length})) \tag{2.1}$$

The only property with a significant effect on SSR mutation odds (locus $< 8bp$) was then used to build a model on the full dataset, including *msh3* status of the strains, in order to estimate the effect of the *msh3* deletion on mutation rate. Since not enough data was available to estimate the effect of *msh3* status directly, we calculated a lower bound on the fold-difference between the WT and *msh3*Δ mutations rates, by choosing an offset for the *msh3* status beta coefficient such that the difference in log likelihood from the model excluding *msh3* status was 2.5 (corresponding to the 95% confidence interval bound) (Equation 2.2).

$$\log\left[\frac{P(\text{mutation})}{1-P(\text{mutation})}\right] = \alpha + \text{offset}(\log(\text{repeat\_length})) + \beta_1(\text{short\_ssr}) + \text{offset}(\text{msh3\_beta\_fixed*msh3\_status}) \tag{2.2}$$

Growth rate assay

Strains were randomised into 96-well U-bottom plates and stored frozen at -80°C in 20 µL of 50% YPD + 15% Glycerol. 'Petite-only' control strains were included in each experimental plate. Three days before each growth rate assay, a plate each of

MA and reference strains was thawed and 180 µL SC media supplemented with an extra 50 mg/L Adenine (SC+Ade) was added to each well (adding Adenine decreases selection pressure for Ade+ phenotypes, including [PSI+] cells). After 1 day of growth in a shaking incubator at 30°C, each strain was diluted 1:10 in SC+Ade in a new plate. The experiment was performed following an additional 2 days of growth from the resulting saturated cells of each line. The microscope growth rate assay was performed largely as described in Sartori et al. (2021). On the day of the growth rate assay, MA line and reference strains were mixed in a 2:1 ratio and diluted $\sim 1 \times 10^{-4}$-fold with vigorous mixing. Cells were imaged hourly for 10 hours in brightfield, followed by a single GFP exposure, as described previously (Levy et al., 2012). Image analysis was performed using the PIE software (Plavskin et al., 2021).

We calculated the growth rate change in the MA lines compared to their ancestor strain as:

$$GR = \frac{\mu_{MA}}{\mu_{anc}} - 1 \tag{2.3}$$

where μ_{MA} and μ_{anc} are the mean growth rates of the non-petite colonies in the MA lines and ancestor strain, respectively.

Chapter 3

The smoking-induced field of injury and its implications for lung cancer risk

Contributions

The work presented in this chapter is the result of a collaboration with Cancer Research UK - Cambridge Institute, the University of Cambridge and the Royal Papworth Hospital NHS Foundation Trust. The work was supervised by Roland Schwarz (University Hospital Cologne, Germany), Bruce Ponder (CRUK-Cambridge Institute) and Robert Rintoul (Royal Papworth Hospital, University of Cambridge) and conducted by me and Florian Massip (Institut Curie, Mines ParisTech, France). Many people contributed to the collection a pre-processing of clinical and molecular data: Rory Stark and the Bioinformatics Core at CRUK-Cambridge contributed to experimental and study design and pre-processed the raw sequencing data; Amanda Stone oversaw all sample and data collection; Amy Gladwell processed patient data and provided clinical classification; Kerstin Meyer and Florian Markowetz helped design and implement the study.

In the text, the first person plural is used when work was performed by a collaborator, or jointly with a collaborator.

Part of this work is reported in de Biase, Massip et al. (2021), available as a preprint on bioRxiv.

3.1 Introduction

3.1.1 Lung cancer and its link to cigarette smoke

To date, lung cancer is the leading cause of cancer-related death worldwide (Sung et al., 2021). Lung cancer is classified into two major groups based on histology: small cell lung cancer (SCLC), observed in 15% of cases and originating from neuroendocrine cells, and non-small cell lung cancer (NSCLC), the most frequently observed in the population (85% of cases), originating from epithelial cells. NSCLC is additionally classified into three subtypes: lung adenocarcinoma (LUAD), usually originating from alveolar type II (AT2) epithelial cells, squamous cell carcinoma (SCC), originating from basal cells, and large cell carcinoma (LCC), originating from various epithelial cell types.

Lung cancer is a perfect example of neoplasia with a strong environmental component. Cigarette smoke has long been established as the main risk factor for all lung cancer subtypes. The hundreds of carcinogenic chemicals contained in cigarette smoke cause the accumulation of mutations in lung tissue. The occurrence of mutations in oncogenes and tumour suppressors such as *TP53* and *KRAS* is the first step towards the development of a malignancy. In addition, carcinogenic molecules and reactive oxygen species (ROS) contained in cigarette smoke lead to persistent inflammation, which in turn leads to the development of lung pathologies such as chronic obstructive pulmonary disease (COPD) and emphysema (Walser et al., 2008). These diseases perpetuate the inflammatory environment and therefore further increase the risk of lung cancer insurgence. Both tobacco carcinogens themselves and the inflammatory mediators produced by the host in response to them are linked to epithelial-mesenchymal transition (EMT), which is involved in early events in carcinogenesis and in determining invasiveness and metastatic potential (Krysan et al., 2008; Yoshino et al., 2007). The frequent loss of *TP53* and *KRAS* in the early stages of lung cancer development also contributes to the creation and maintenance of pulmonary inflammation: p53 is a suppressor of NF-κB, one of the most prominent mediators of inflammation, while loss of *KRAS* has been shown to induce the production of COX-2 and downstream inflammatory mediators, which in turn results in increased EMT and immunosuppression (Walser et al., 2008).

Another important characteristic of the link between smoking and lung cancer is that smoking cessation, while reducing the risk of developing cancer, never reverts it back

to baseline. Over 40% of lung cancer cases occur in former smokers more than 15 years after smoking cessation (Siegel et al., 2020; Tindle et al., 2018), suggesting that the damage to the pulmonary environment caused by smoking could, at least in part, be irreversible.

The clear link between cigarette smoke and lung cancer risk creates a defined high-risk population that benefits from regular screening. In fact, great effort was put into the implementation of such screening plans, and monitoring with low-dose CT scans (LDCT) revealed to be very effective, determining a 26% reduction in mortality (de Koning et al., 2020; National Lung Screening Trial Research Team et al., 2011). However, there are several drawbacks to the use of frequent LDCT screening, including the high costs and the risks associated with cumulative radiation exposure. In addition, diagnostic procedures following the observation of a suspicious lesion on a CT scan, most commonly fiber-optic bronchoscopy, are invasive and often do not yield a definitive diagnostic response, especially for peripheral lesions (Rivera et al., 2013). This leads to further invasive diagnostic procedures, even in the event of a benign lesion. All these considerations, together with the fact that evident symptoms of lung cancer present at late stages of the disease, contribute to the low 5-year survival rate which, today, remains $\sim$ 19% overall and $\sim$ 5% for advanced stages (Siegel et al., 2020).

3.1.2 The airway field of injury

The concept of field of injury has its origin in Danely Slaughter's seminal studies during the 1950s, where he described the peculiar characteristics of the insurgence and spread of oral squamous cancers (Slaughter et al., 1953). He observed that oral cancers tend to have a lateral rather than vertical spread, and that benign epithelium outside the margins of a cancer lesion also showed abnormalities such as epithelial hyperplasia and hyperkeratinization. He also observed that pre-invasive lesions were present at multiple foci within the benign tissue. His observations of what he defined as 'field cancerization' prompted the hypothesis that a pre-conditioned, injured, epithelium represents a favourable environment for pre-malignancies to arise and eventually develop into cancer. This hypothesis explains the tendency of squamous cancers to have multifocal growth, as well as the high frequency of local re-occurrence. Field cancerization refers to abnormalities in benign tissue in the proximity of a neoplasm. However, in later studies following Slaughter's first observations, more widespread alterations were observed in tissues exposed to damaging agents, even in the absence of a frank malignancy. These observations expanded the concept of field cancerization

to the more general concept of "field of injury", defined as an array of molecular alterations observable throughout the tissue exposed to the damaging agent, reflecting the host's response to injury (Steiling et al., 2008).

As oral cancer, lung cancer presents a clear causal link to an external damaging agent, cigarette smoke. The first report suggesting an airway field of injury produced by the exposure to cigarette smoke is a publication from 1961, in which the authors described extensive alterations in the bronchial epithelium of smokers and the presence of multiple independent foci of pre-malignancy, even when death occurred in the absence of lung cancer (Auerbach et al., 1961). Since then, the airway field of injury and its link to lung cancer have been thoroughly described. Different studies have shown that bronchial epithelium of cancer-free current and former smokers exhibits a wide array of alterations, including point mutations, allelic losses, microsatellite instability, changes in promoter methylation and altered telomerase activity (Franklin et al., 1997; Miyazu et al., 2005; Powell et al., 1999; Wistuba et al., 1997). Often these alterations involve important oncogenes and tumour suppressors; an example is loss of heterozygosity (LOH) of chromosome 13q (*RB* gene) and 17p (*TP53* gene). Moreover, identical alterations can be found in lung tumour samples and histologically normal airway tissue from the same patients (Nelson et al., 1996; Tang et al., 2005).

The plethora of genetic and epigenetic alterations, along with the acute and chronic inflammation caused by cigarette smoke, produce extensive changes in the airway transcriptome. Spira et al. (2004) described the gene expression alterations occurring in bronchial epithelial tissue of smokers without lung cancer. Genes up-regulated in smokers were involved mainly in secretion, oxidative stress response and xenobiotic metabolism, while down-regulated genes were involved in regulation of inflammation. Moreover, Beane et al. (2007) performed a study on the reversibility of the bronchial injury after smoking cessation, by observing the behaviour of affected genes in former smokers. The authors found that genes that rapidly returned to normal expression levels after smoking cessation were mostly involved in the detoxification of tobacco smoke components, for example the cytochrome p450 genes *CYP1A1* and *CYP1B1*, the NADPH oxidoreductases genes *AKR1B10* and *AKR1C1*, and the aldehyde dehydrogenase gene *ALDH3A1*. On the other hand, several genes were found to be slowly reversible or irreversible in former smokers, including ones encoding for adhesion molecules, such as *CEACAM5*, metalloproteinases, such as *MMP10*, and immune-related molecules, such as *CX3CL1*. Additional studies showed that smoking-induced expression changes in the healthy-appearing bronchus of patients with suspected lung

cancer can serve as a lung cancer biomarker, complementing bronchoscopy results and potentially avoiding further invasive procedures (Spira et al., 2007).

Corroborating the presence of a field of injury in the airway of smokers, similar transcriptomic changes to those observed in the bronchus were found in other tissues exposed to smoke, such as nasal and buccal epithelium (Sridhar et al., 2008; Zhang et al., 2010). This observation opened the way to studies on the potential use of easily accessible tissues of the airway as non-invasive proxies to aid lung cancer diagnosis. Perez-Rogers et al. (2017) investigated the bronchial and nasal epithelium of current and former smokers with lesions suspicious for lung cancer and found concordant cancer-associated gene expression changes in the two tissues. The authors also showed that the use of nasal expression of these genes in a clinico-genomic classifier improved lung cancer prediction compared to the use of clinical information only. This first "proof-of-principle" study demonstrated that accessible airway field of injury alterations can be predictive of the presence of cancer.

Despite the clear link between cigarette smoke, airway field of injury and lung cancer, only $\sim 15\%$ of smokers develop lung cancer (Shields, 1999), suggesting that the patient's genetic background and the way they respond to smoke-induced injury could play an important role in determining the risk of cancer insurgence. In this chapter, I will describe a study investigating the airway field of injury in bronchial and nasal tissue of healthy volunteers and clinic-referred patients with suspected lung cancer, with a diverse smoking history. I will present a thorough characterisation of the smoke injury response in the nasal epithelium of the healthy population and compare it to the response in clinic-referred patients. I will then describe how nasal expression of these genes affected by smoking, in particular those exhibiting a different behaviour in healthy and clinic subjects, can potentially be used to improve lung cancer risk stratification.

3.1.3 The CRUKPAP dataset

The following study was conducted on samples collected as part of the CRUKPAP study. The CRUKPAP dataset includes 487 subjects, among which 114 healthy volunteers (HV group) recruited from the Cambridge Bioresource (cam) and 373 patients referred to the out-patient clinic at the Royal Papworth Hospital (Cambridge, UK) or Peterborough City Hospital on suspicion of lung cancer (clinic group). Within the clinic group, 72 subjects showed benign conditions (clinic benign) and 301 were diagnosed with lung cancer (clinic cancer). Nasal epithelial samples were collected by

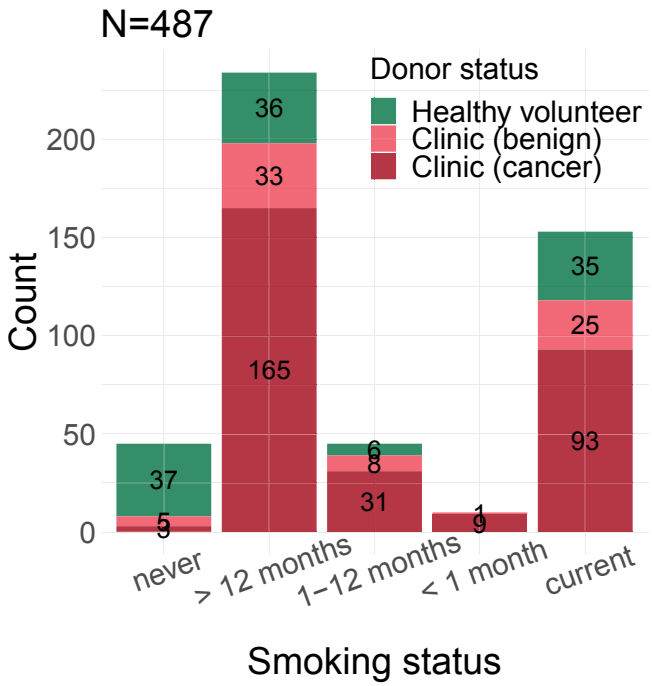


Fig. 3.1 **Overview of study subjects.** Number of individuals with different smoking status recruited for the study, stratified by donor's population of origin: healthy volunteer group and clinic-referred patients with a diagnosis of benign condition or lung cancer.

mini-curette from all 114 HV donors and 299 Clinic patients, and bronchial brushings from 236 clinic patients (Section 3.5). Both nasal and bronchial samples were obtained from 162 of the clinic patients. Smoking history was obtained for all subjects: these included never smokers (NV), current smokers (CS) and former smokers (FS) with time-since-quit ranging from less than 1 month to more than 30 years. Within this study, former smokers were stratified into 3 categories based on their time-since-quit: former smokers who quit less than one month, 1 to 12 months or more than 1 year prior to sample collection (**Figure 3.1**). Smoking intensity was also recorded in terms of pack-years, and stratified into 4 categories: none, less than 10 years, 1-30 years and more than 30 years. In addition to smoking status, sex, age, lung cancer subtype and stage and presence of chronic obstructive pulmonary disease (COPD) were recorded (**Table 3.1**). While most clinic cancer subjects were diagnosed with non-small cell lung cancer (NSCLC), 56 subjects presented a metastatic mass in their lung from a different cancer type, or were diagnosed with small-cell lung cancer (SCLC): these subjects (annotated as having "ineligible" cancer status) were included in all analyses investigating smoke injury response, but discarded for lung cancer risk prediction. Clinic benign patients were followed up for a minimum of 1 year to confirm

the absence of cancer. Airway samples underwent RNA sequencing using standard protocols (Section 3.5). Total gene expression was quantified as variance-stabilised counts and corrected for batch effects in all downstream analyses (Section 3.5).

3.2 Results

3.2.1 Transcriptome exploration of nasal and bronchial tissue from subjects with different smoking and disease status

First, I explored the contribution of the clinical and environmental variables known for our study subjects to total gene expression in nasal and bronchial tissue. To do so, I performed a variance components analysis across all genes and all samples (Section 3.5). I specifically tested for the contribution of tissue of origin, cancer status, smoking status, cumulative smoke exposure (measured in pack-years), sex and age. The proportion of total variance explained across all samples was 22,2%. The large percentage of unexplained variance is likely due to subject-specific differences. The variable contributing most to gene expression was tissue of origin (70% of the total explained variance, **Figure 3.2a**), followed by the donor's population of origin (healthy volunteer or clinic-referred patient, 15% of the total explained variance) and smoking status (14% of the total explained variance). Cancer status only accounted for 5% of the total explained variance, following sex and age. Similar results were obtained when performing the analysis on bronchial and nasal samples separately (**Figures 3.2b and 3.2c**). In particular, in nasal samples the proportion of total variance explained was 4%, the main contributors being again healthy volunteer status and smoking status.

To further explore the nasal epithelial transcriptome of healthy and clinic donors, I performed a differential expression analysis comparing clinic current and former smokers to healthy volunteer current and former smokers, correcting for age, sex, smoking status and pack-years. There were 5359 genes differentially expressed between clinic patients and healthy volunteers ($P < .05$, Section 3.5). I performed Gene Ontology (GO) term enrichment analysis on the list of differentially expressed genes and found that genes up-regulated in clinic patients enriched in GO terms related to cilium assembly and organisation and chromatin modification, while down-regulated genes enriched in oxidative phosphorylation and several immune-related terms, such as *Neutrophil activation*, *Antigen processing and presentation* and *Response to interferon gamma* (**Figure 3.3** and Supplementary table 1). I then performed the same comparison

	Healthy volunteers	Clinic group	
		Without Cancer	With Cancer
Sex			
Male	60	52	193
Female	54	20	108
Age			
(24.9, 41.5]	9	2	1
(41.5, 58]	38	18	44
(58, 74.5]	64	39	176
(74.5, 91.1]	3	13	80
Smoking status			
never	37	5	3
> 12 months	36	33	165
1-12 months	6	8	31
< 1 month	0	1	9
current	35	25	93
Pack-years			
None	37	5	3
0-10	19	14	20
11-30	35	20	72
> 30	22	32	206
Unknown	1	1	0
Tissue			
Nasal	114	13	125
Bronchial	0	16	58
Both	0	43	119
Cancer status and subtype			
No cancer	114	72	0
Adenocarcinoma	0	0	126
Squamous cell carcinoma	0	0	99
Not specified	0	0	20
Ineligible	0	0	56
Cancer Stage			
None	114	72	0
Stage 1	0	0	50
Stage 2	0	0	38
Stage 3	0	0	79
Stage 4	0	0	62
Mix or Unknown	0	0	16
Ineligible			56
COPD			
None	93	25	103
Mild	9	7	47
Moderate	4	18	66
Severe	2	6	32
Past history	0	5	17
Unknown	6	11	36

Table 3.1 **Clinical and demographic characteristics of the study subjects.**

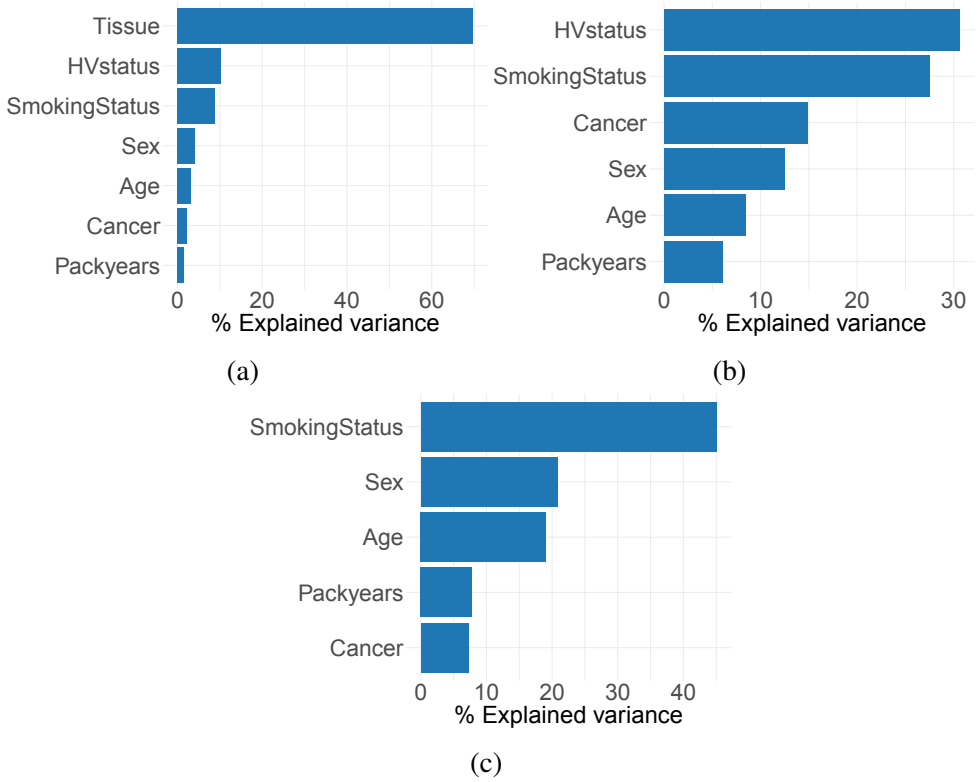


Fig. 3.2 **Variance components analysis.** Contribution of different clinical variables to the total explained variance in gene expression calculated using a random model on all samples **(a)**, nasal samples **(b)** and bronchial samples **(c)**.

separately in current smokers and former smokers who had quit for more than 1 year, to identify possible differences in transcriptome alterations in the two populations. For current smokers, GO enrichment in the genes with increased and reduced expression in clinic donors was similar to the previous comparison. For former smokers, no enrichment was found for genes related to ciliary function within the up-regulated set, but genes with reduced expression were enriched for immune pathways such as *Inflammatory response*, *Neutrophil activation* and *Response to interferon gamma*. These results suggest the presence of an immunosuppressed state in subjects from the clinic group, and that this state can be detected at a distal airway site both during active smoking and after smoking cessation.

Next, I performed a differential expression analysis comparing clinic donors with and without cancer. Only 28 genes were significantly altered ($P < .05$) in the bronchus, and no genes were significantly differentially expressed in the nose. Among the 28 differentially expressed genes in the bronchus, 3 were up-regulated in patients with cancer: *MMP13*, encoding for a metalloproteinase known to increase lung cancer invasion and metastasis (Merchant et al., 2017), *EDA2R*, encoding for a member of

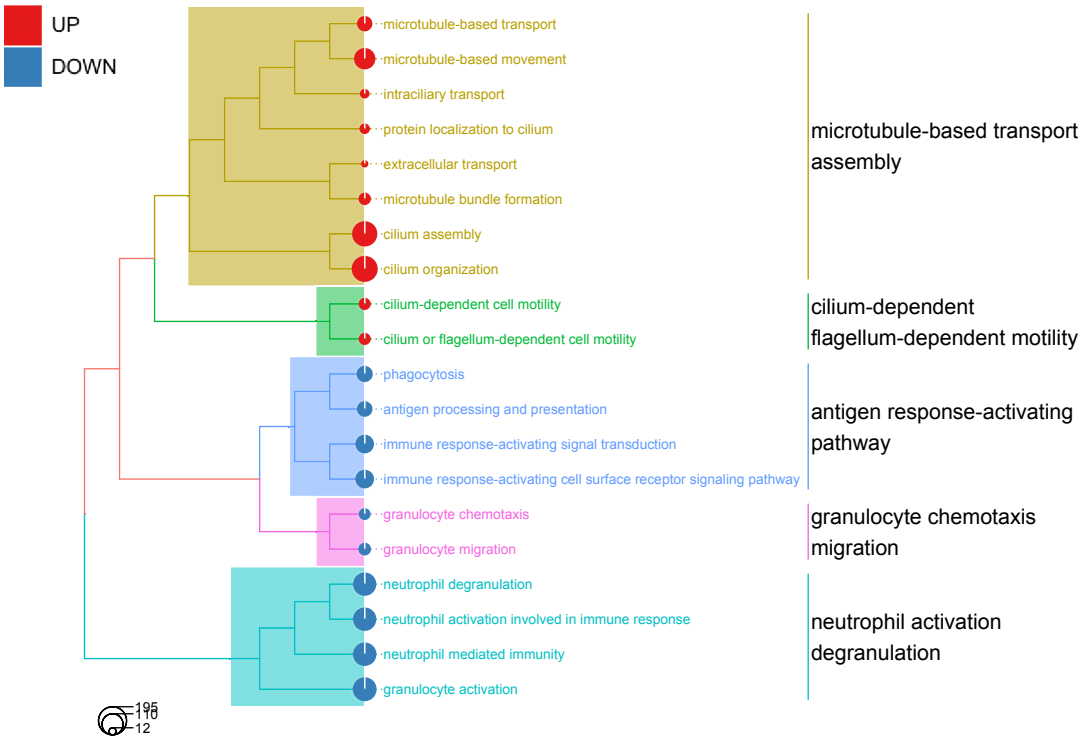


Fig. 3.3 **GO enrichment of differentially expressed genes in clinic compared to healthy subjects.** Tree plot showing the top enriched GO terms among up- (red) and down-regulated (blue) genes in clinic patients compared to healthy volunteers. Dot size indicates number of genes within a term.

the tumour necrosis factor (TNF) receptor superfamily, members of which modulate immune response in the tumour microenvironment and might serve as biomarkers for immunotherapy in lung cancer (Zhang et al., 2020), and *CTSL*, encoding for a lysosomal cysteine protease involved in EMT (Sullivan et al., 2009). The 25 genes down-regulated in cancer patients were enriched in immune-related GO terms, in particular neutrophil-mediated immunity, consistent with the findings in the comparison between clinic patients and healthy volunteers in nasal tissue (Supplementary table 2).

Taken together, these results show that the strongest gene expression changes observable in nasal epithelium are between healthy volunteers and clinic patients, while differences between donors with and without cancer appear to be too subtle to be detected at a distal site such as nasal epithelium with a gene-level differential expression analysis.

Given the extensive differences in the transcriptome of healthy and clinic subjects with a similar smoking history, we argued that these individuals could also differ in their response to smoke injury, and that these differences might reflect on their risk

of developing lung disease and, in particular, lung cancer. In the next sections, I characterise and compare the smoke injury response in the two groups of subjects at the level of nasal epithelium.

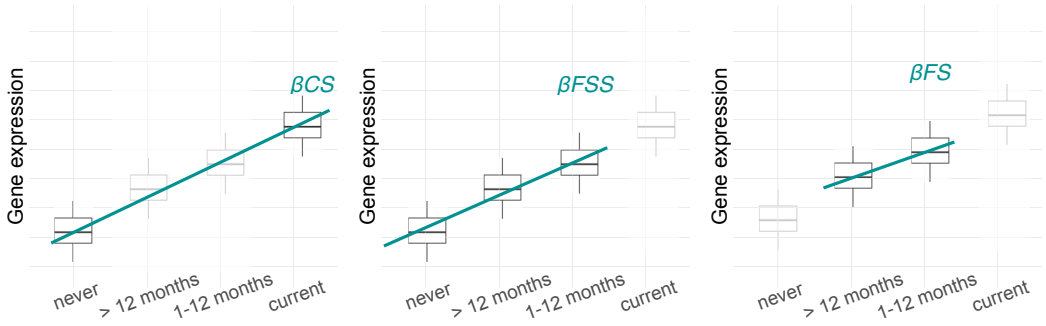
3.2.2 Smoke injury response and reversibility of damage in healthy current and former smokers

In order to investigate the smoke injury response in the nasal epithelium of healthy subjects, as well as its long-term response after smoking cessation, I employed multivariate linear regression. I modelled gene expression changes over smoking status for each of 18,072 protein-coding genes individually, using gene expression in healthy never smokers as a baseline. I encoded smoking status into three binary variables (**Figure 3.4a**): CS (current smoker status, 0/1), FSS (former smoker status, 0/1) and FS (former-smoker's time since quit, 0 for current and never smokers, 1 for ex smokers 1-12 months, 2 for ex smokers > 12 months). Additionally, I included age, sex and experimental batch as confounding variables (Equation (3.1)).

$$\text{gxp} = \alpha + \beta_1(\text{CS}) + \beta_2(\text{FSS}) + \beta_3(\text{FS}) + \beta_4(\text{sex}) + \beta_5(\text{age}) + \beta_6(\text{batch}) + \varepsilon \quad (3.1)$$

To classify genes based on their behaviour over smoking status, I employed a Bayesian approach to model selection. I tested for the inclusion of each of the three variables into the model and inferred posterior probabilities for all eight possible models to retrieve the most likely reversibility dynamic of gene expression changes for each gene. Each combination, or group of combinations, of variables was associated to a gene class. Each gene was assigned to the class with the highest posterior probability (**Figure 3.4b**). Genes with no discernible behaviour over smoking status were classified as *unaffected by smoking* (US). Genes for which a difference in expression was observed in current compared to never smokers were assigned to one of three reversibility classes: *rapidly reversible* (RR), if no difference could be observed between former and never smokers; *slowly reversible* (SR), if a slope across the current and former smoker categories showed a trend of returning to never-smoker expression levels; *irreversible* (IR) if a difference was observed in former compared to never smokers but no difference is observed between current and former smokers and across the former smoker categories. In addition, genes were classified as *cessation-associated* (CA) if no difference was

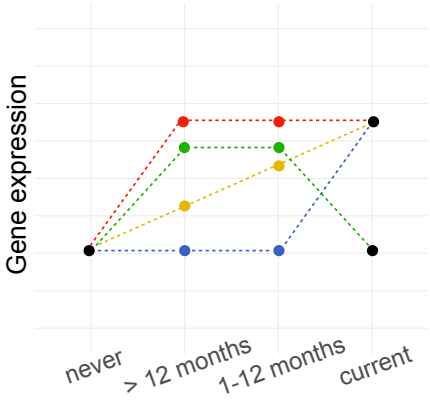
present between current and never smokers, but elevated or reduced expression was observed in former smokers (**Figure 3.4c**).



(a)

FSS	FS	CS	PP	
0	0	1	a	→ RR
0	1	1	b	→ SR
1	1	1	c	
1	0	1	d	→ IR
1	0	0	e	→ CA
1	1	0	f	
0	0	0	g	→ US
0	1	0	h	

(b)



(c)

Fig. 3.4 **Smoke injury reversibility analysis. (a)** The slope coefficients associated to the three smoking status variables included in the Bayesian model (CS: current smoker status, FSS: former smoker status, FS: former smoker's time since quit); **(b)** Description of the model selection procedure used to assign each gene to a reversibility class: the table shows all possible combinations of inclusion/exclusion of the three smoking status variables; **(c)** In blue, yellow and red, schematic of a gene with altered expression in current compared to never smokers, and the three possible trajectories after smoking cessation, corresponding to the RR, SR and IR reversibility classes; in green, schematic of a gene with no expression difference in current compared to never smokers, but altered expression in former smokers, corresponding to the CA class. US: not affected; RR: rapidly reversible SR: slowly reversible; IR:irreversible; CA: cessation-associated; PP: posterior probability.

In total, 5755 genes were found to be affected by smoke in healthy volunteers, of which 513 genes showed a strong difference between never smokers and current smokers, or former and never smokers in the case of cessation-associated genes (Section 3.5, Supplementary table 3). The majority (485/513) of the affected genes were classified as rapidly reversible. This result is in line with a previous study by Beane et al. (2007), in which a reversibility classification was performed for genes affected by smoke in

bronchial tissue, and shows that there is a similarity in the reversibility of damage in bronchial and nasal tissue. Rapidly reversible genes showing up-regulation in current smokers were involved in cellular detoxification, response to oxidative stress (e.g. *CYP1A1*, *CYP1B1*, *AHRR*, *NQO1*, *GPX2*, *ALDH3A1*) and keratinization (e.g. *KRT6A*, *KRT13*, *KRT17*, *SPRR1A*, *SPRR1B*, *CSTA*), while down-regulated genes were involved in cilium organisation (e.g *FOXJ1*, *DNAH6*, *IFT81*, *CEP290*, *UBXN10*), extracellular matrix organisation (e.g. *FN1*, *COL3A1*, *COL5A1*, *COL9A2*) and interferon signalling (e.g. *IFI6*, *IFIT1*, *IFI44*, *RSAD2*). Genes involved in inflammatory response were found in both the up-regulated (*IL36A*, *IL36G*, *S100A8*, *S100A9*, *CLU*) and down-regulated (*SAA1*, *SAA2*, *IL33*) genes. Of the remaining 28 genes, 6 were classified as slowly reversible (*CCK*, *STATH*, *CXCL13*, *SRCRB4D*, *CLU*, *PLCB2*), 2 (*SULF1*, *FRMD3*) as irreversible and 20 as cessation-associated. Among the slowly reversible genes, of particular interest is *CXCL13*, which has been shown to play an important role in carcinogenesis induced by polycyclic aromatic hydrocarbon (Wang et al., 2015). Both the irreversible genes, *SULF1* and *FRMD3*, showed down-regulation in current smokers in our study. *SULF1* is down-regulated in many cancers and has been shown to impede angiogenesis and carcinogenesis both in vitro and in vivo (Lai et al., 2008). *FRMD3* is a tumour suppressor frequently silenced in non-small cell lung cancers and has been shown to reduce clonogenicity (Haase et al., 2007).

3.2.3 Deviations from healthy smoke injury response in clinic subjects

By conducting this analysis in healthy volunteer subjects, I was able to describe the expected injury response of a population of healthy current and former smokers. Despite being exposed to the same damaging agent, not all smokers develop lung cancer, suggesting that the response in higher-risk subjects might deviate from this expected response. Therefore, I repeated the analysis just described to model the smoke injury response in patients from the clinic group, aiming at finding differences from the healthy smoke injury response. As in the previous section, I used healthy never smokers as the 'baseline' group.

There were 4112 genes with smoking-dependent expression changes in the clinic group, of which 584 showed a large effect size (Section 3.5, Supplementary table 3).

To verify that the model correctly classified the reversibility of the genes affected by smoking, I performed a principal component analysis (PCA) of nasal samples using

genes in the different reversibility classes independently. I performed this PCA in both healthy and clinic subjects, using the respective reversibility classifications. Since only 2 genes were classified as irreversible in healthy subjects, for that group I performed PCA for slowly reversible and irreversible genes jointly. As expected, the PCA showed a clear separation of current smokers from all other subjects for rapidly reversible genes; slowly reversible genes placed patients on a trajectory from never smokers to current smokers; irreversible genes separated never smokers from former and current smokers (**Figures 3.5a and 3.5b**). Additionally, PCA on the bronchial samples of clinic subjects, using the same set of genes, showed a similar pattern, confirming that nasal and bronchial epithelium have similar responses to short and long term smoke-induced injury (**Figure 3.5c**).

Overall, we observed a shift towards slower reversibility for the smoke-affected genes in the clinic group compared to the healthy volunteers: 190 genes were found to be rapidly reversible, 107 slowly reversible, 102 irreversible and 185 cessation associated. The smoke-injury genes identified in the clinic group also showed a significant overlap with those found in the healthy volunteer group, with 233 shared genes (χ^2 test $P <$ .001). Within these 233 genes, 227 were rapidly reversible in the healthy volunteer group; of those 227, only 112 remained rapidly reversible in the clinic group. The remaining 115 genes were classified differently in the clinic compared to the healthy group: 22 genes became slowly reversible, 1 gene irreversible and 92 genes cessation-associated. For example, *CYP1B1*, a well-known detoxification gene, and *BMP7*, a gene previously shown to have a role in immunoregulation (Cortez et al., 2020), appear to be rapidly reversible in healthy volunteers but slowly reversible in the clinic group (**Figure 3.6b**). *WNT5A* and *SUSD2* genes show a similar behaviour. *WNT5A* was up-regulated in current smokers; its over-expression has been shown to induce epithelial-mesenchymal transition and invasiveness in NSCLC (Wang et al., 2017). *SUSD2*, down-regulated in current smokers, was identified as a tumour suppressor in NSCLC (Cai et al., 2015; Cheng et al., 2016). The 92 genes switching their classification from rapidly reversible in healthy volunteers to cessation-associated in clinic subjects showed a strong enrichment for cilia structure and function (**Figure 3.6c**, Supplementary table 4).

Notably, 351 genes with a smoking-dependent expression change in the clinic group had no smoking-dependent change in healthy volunteers (**Figure 3.6a**). These genes were strongly enriched in extracellular matrix organisation and immune-related genes (including response to interferon gamma, neutrophil activation, chemotaxis and inflammation (**Figure 3.6c**). For example, the expression of *GBP6*, an interferon-

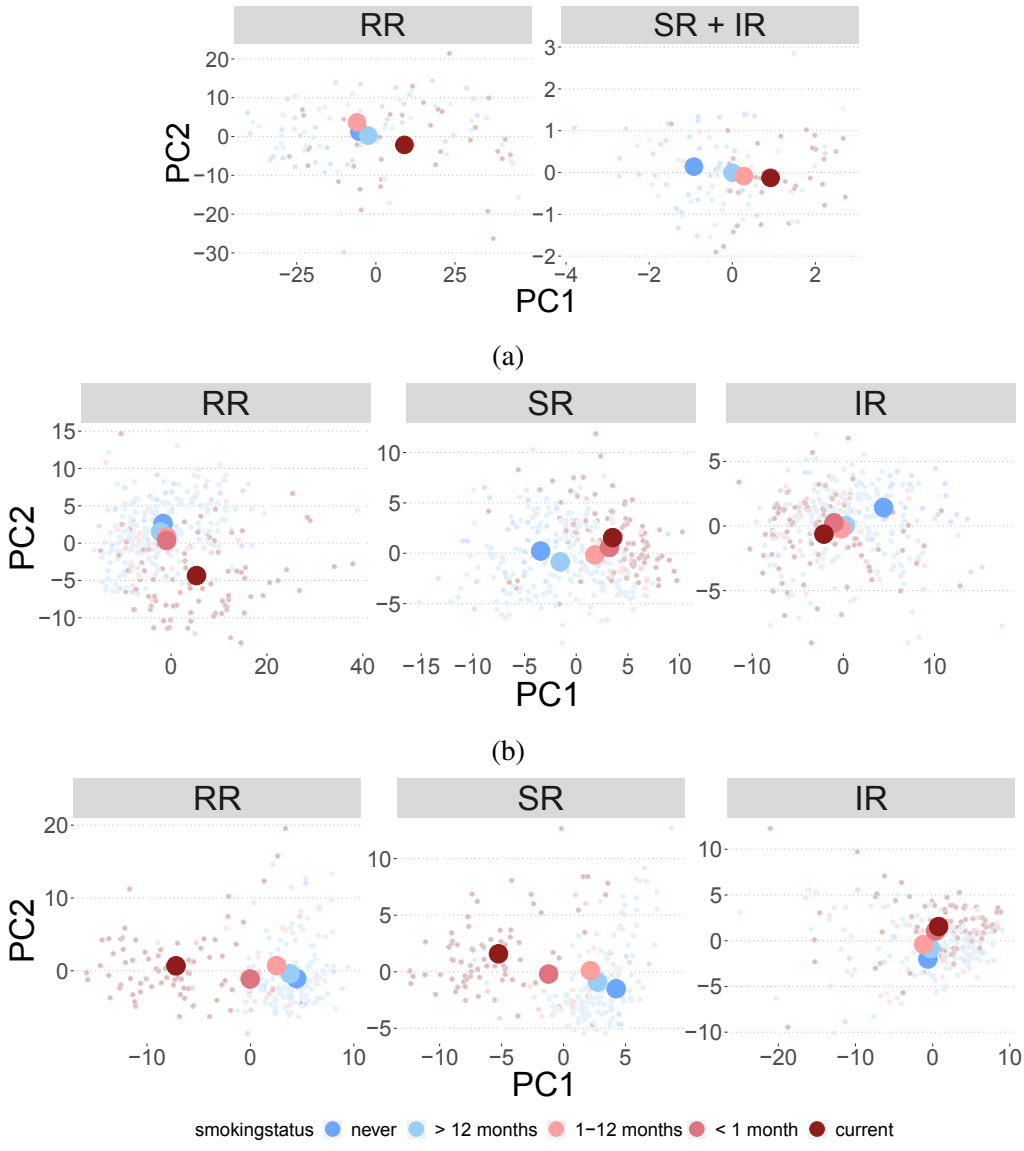


Fig. 3.5 **Principal component analysis on the genes belonging to different reversibility classes.** RR: Rapidly reversible genes, SR: Slowly reversible genes; IR:irreversible genes. Each small dot is a patient and colours indicate the smoking status of the patient. Large dots represent the mean of all patients for each smoking class. **(a)**: Nasal samples from healthy volunteers, using the reversibility classes from the Bayesian model on the healthy volunteer group. **(b)**: Nasal samples from clinic subjects (cancer + benign), using the reversibility classes from the Bayesian model on the clinic group. **(c)**: Bronchial samples from clinic subjects (cancer + benign), using the reversibility classes from the Bayesian model on the clinic group.

induced gene, was constant over smoking status in healthy volunteers, while reduced in smokers in the clinic group and showed a slowly reversible post-cessation dynamic (**Figure 3.6b**). Down-regulation of *GBP6* was associated with reduced overall survival in squamous cell carcinoma of the head and neck (Wu et al., 2020). These results align with the differences between the clinic and healthy volunteer groups observed by differential expression analysis and again highlight the presence of immune alterations in clinic subjects.

Overall, a strikingly different response to smoke can be observed in clinic patients compared to healthy subjects, with altered genes in clinic patients showing slower reversibility post-cessation. Moreover, a large number of alterations appear to be specific to the clinic group.

From here on, I define the union of the genes that show smoking-dependent expression changes in the healthy and clinic groups as *smoke-injury genes* (N=864).

3.2.4 Reversibility of pathways affected by smoking

In the previous sections I described how I identified a set of genes involved in the smoke injury response in the healthy volunteer and clinic groups, along with the major disrupted cellular pathways and functions. Next, I set out to assess the overall behaviour and post-cessation reversibility of these pathways, in an effort to identify regulatory mechanisms underlying the different response to smoke in the two donor groups. Therefore, I performed a pathway analysis by aggregating the expression of genes belonging to pathways of interest and looking at their behaviour over smoking status. I calculated geneset metascores by averaging the expression of genes belonging to each of 8 GO terms: *Keratinization*, *Oxidative stress response*, *Extracellular matrix organization*, *Cilium organization* and the immune-related *Response to interferon gamma*, *Neutrophil-mediated immunity*, *Antigen processing and presentation* and *Inflammatory response*. The metascore trends over smoking status mirrored the gene-level observations described in Section 3.2.2 and Section 3.2.3. *Keratinization* showed similar, rapidly reversible, dynamics in the two donor groups, while *Oxidative stress response* showed a slower reversibility in the clinic compared to the healthy group. *Cilium organization* appeared rapidly reversible in healthy volunteers while displaying a cessation-associated trajectory in clinic subjects, with seemingly increased expression in former smokers compared to current and never smokers. All immune-related genesets had reduced expression and were uniquely disrupted in clinic patients (**Figure 3.7a**). In particular, the metascore of *Response to interferon gamma* and

Antigen processing and presentation in clinic patients did not revert back to healthy never-smoker level even longer than 10 years after smoking cessation (**Figure 3.7b**).

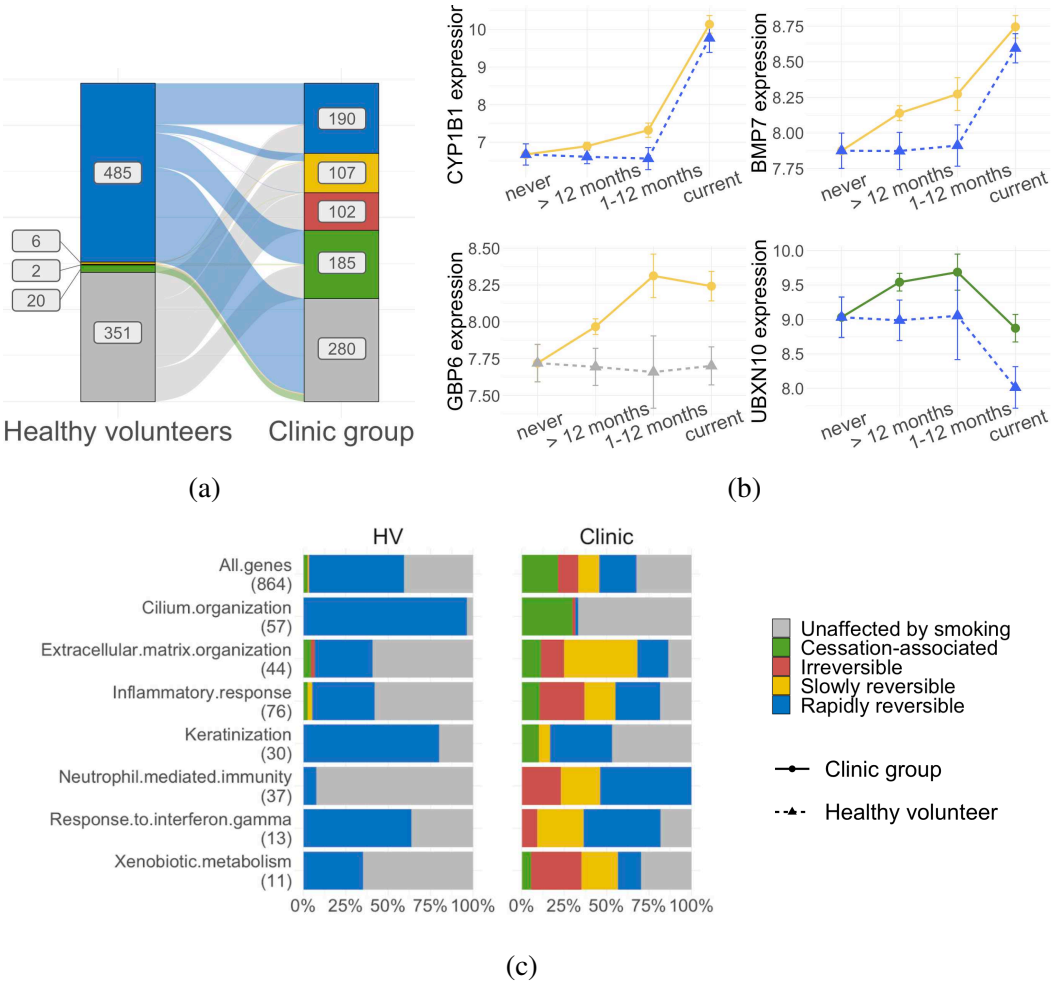
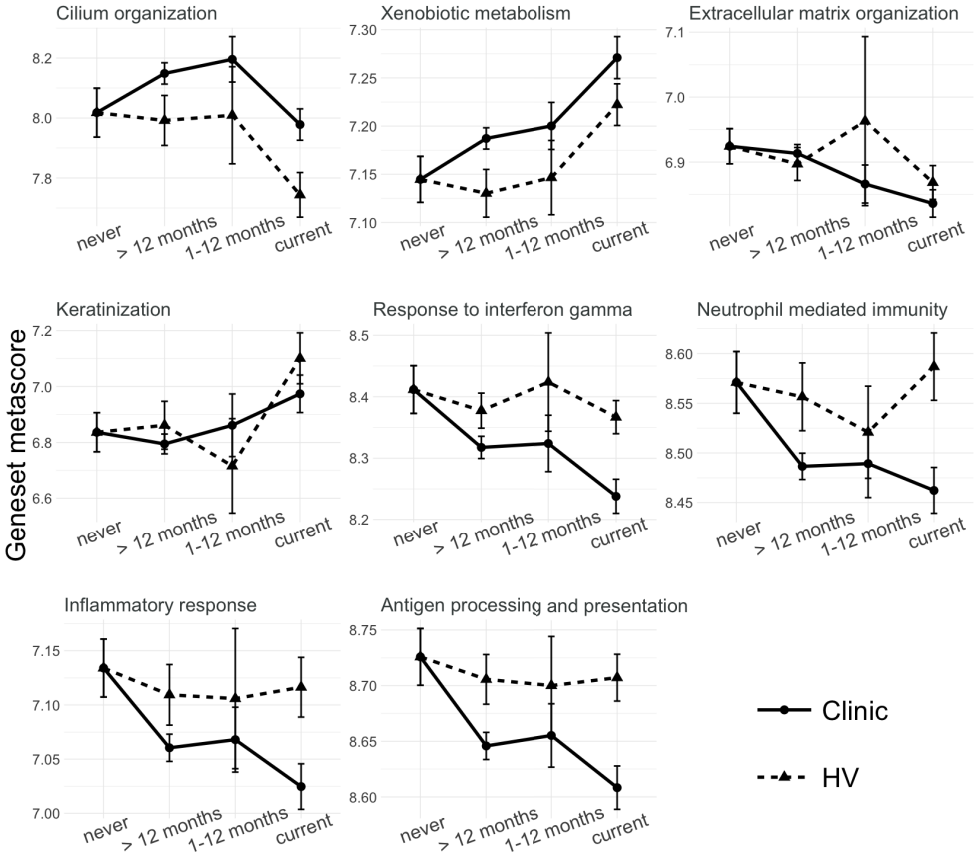
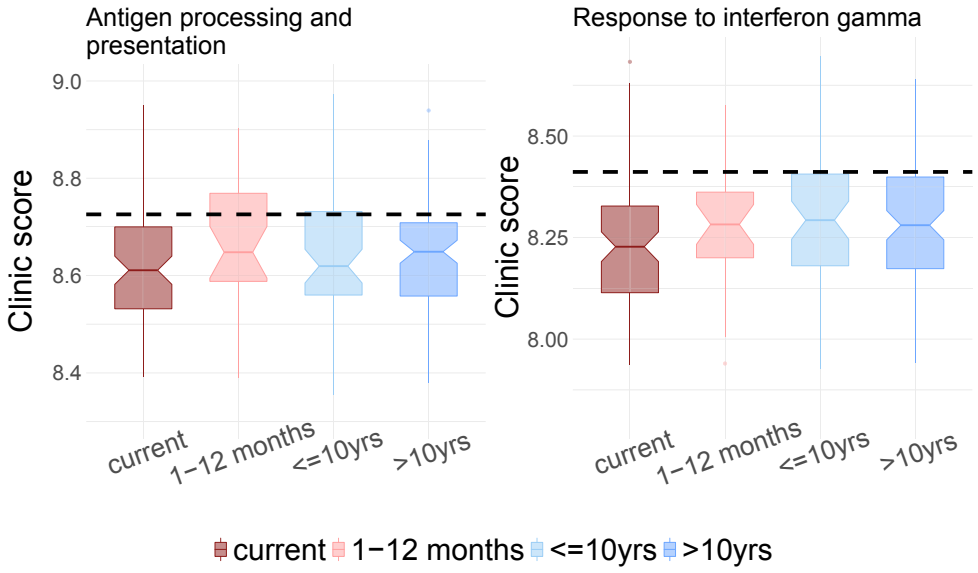


Fig. 3.6 **Smoke injury dynamics in nasal epithelium.** **(a)** Plot showing a change of reversibility dynamics for the same genes in the two donor groups; **(b)** Normalised gene expression over smoking status for 4 smoke injury genes with different post-cessation dynamics in the clinic and healthy groups, with line-type and shape representing donor status and colours representing the reversibility classes assigned to genes; **(c)** Proportion of RR, SR and IR genes within different GO categories involved in smoke injury response. The top row shows the proportions for all smoke-injury genes, and the numbers in parentheses give the number of smoke-injury genes found in each GO category. In **(a)** and **(c)** genes classified as unaffected by smoking in both donor groups were removed.


Cilium organization
Xenobiotic metabolism
Extracellular matrix organization
Keratinization
Response to interferon gamma
Neutrophil mediated immunity
Inflammatory response
Antigen processing and presentation
Geneset metascore
never
> 12 months
1-12 months
current
Clinic
HV

(a)


Antigen processing and presentation
Response to interferon gamma
Clinic score
current
1−12 months
<=10yrs
>10yrs

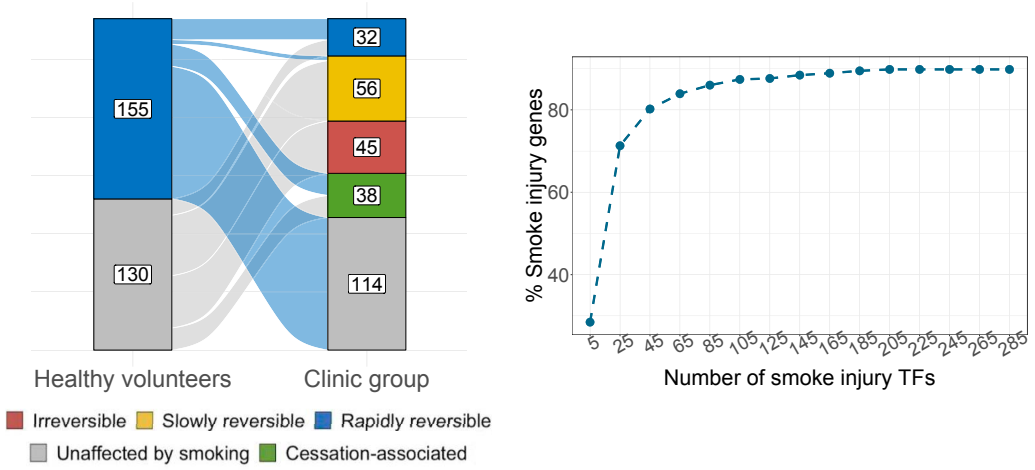
(b)

Fig. 3.7 *(previous page)* **Reversibility of pathways affected by smoking. (a)** Geneset metascore over smoking status for 8 GO terms describing pathways involved in smoke injury response; the plots show a comparison of pathway metascore dynamics in healthy and clinic subjects. **(b)** Pathway metascore for *Response to interferon gamma* and *Antigen processing and presentation* in clinic patients, with >1 year former smokers divided in subjects who quit more or less than 10 years before sample collection; black dashed line represents the average metascore value in healthy never smokers.

3.2.5 Core transcriptional regulators of smoke injury response

Next, I sought a different approach to understand the processes involved in smoke injury response and their post-cessation reversibility by identifying the transcriptional regulators orchestrating the observed expression changes in current and former smokers. To this end, I built a transcription factor (TF)-target interaction network specific to nasal epithelium, and inferred the activity of each TF in the network from nasal gene expression data (Section 3.5). I then used the Bayesian model selection approach described in Section 3.2.2 to categorise each TF into reversibility classes (US, RR, SR, IR, CA), separately in healthy volunteers and clinic subjects. As before, I used healthy never smokers as the baseline for comparison. I found 155 TFs with smoking-dependent activity changes in the healthy volunteer group (at the effect size thresholds defined in Section 3.5). All of them were classified as rapidly reversible. Similarly to the results at the gene level, in the clinic group I observed a shift towards slower reversibility: of 171 TFs with smoking-dependent activity changes, 32 were classified as rapidly reversible, 56 as slowly reversible, 45 as irreversible and 38 as cessation-associated (**Figure 3.8a**).

Since the TF-target interaction network was built from gene expression data from nasal samples, I expected the smoke injury TFs to summarise the expression alterations observed at the gene level. To confirm this, I performed a hypergeometric test for over-representation of all 864 smoke injury genes identified in the healthy volunteer and clinic groups within the targets of each of the 285 smoke injury TFs. Significant enrichment was observed for 130/285 TFs ($P < .05$). Notably, $\sim$70% of the smoke injury genes (616/864) were contained within the targets of 25 smoke injury TFs (**Figure 3.8b**). I defined these TFs as *master regulators* of the smoke injury response, as they are likely the drivers of the transcriptional changes caused by cigarette smoke in nasal epithelium. When visualised in an interaction network where nodes are TFs and edges represent the overlap of their target genes, the master regulator TFs



(a) (b)

Fig. 3.8 **Smoke injury dynamics at TF-level. (a)** Plot showing a change of reversibility dynamics for the same TFs in the two donor groups (TFs classified as unaffected by smoke in both donor groups were excluded); **(b)** Accumulation curve showing the percentage of smoke injury genes (union of genes found in healthy and clinic group, n=864) covered by an increasing number of smoke injury TFs (union of TFs found in healthy and clinic group, n=285). The 'elbow' of the curve is reached at 25 TFs, which cover ∼70% of smoke injury genes.

form smaller groups, each regulating one of the main biological functions identified as disrupted by smoking: ciliary function, keratinization, oxidative stress response, extracellular matrix organisation and immune response and inflammation (**Figure 3.9a**, Supplementary table 5).

I then looked at the activity dynamics over smoking status for the 25 master regulator TFs (4 examples are shown in **Figure 3.9b**). Five TFs showed the same dynamic in healthy and clinic subjects, being rapidly reversible in both groups (2/5 passed the effect size threshold, Section 3.5). Four of these TFs are involved in the regulation of keratinization. Keratinization is a mark of squamous metaplasia, a pre-malignant alteration of the airway epithelium induced by exposure to cigarette smoke (Leube and Rustad, 1991; Peters et al., 1993). Rapid reversibility of the activity of these TFs suggests that squamous metaplasia induced by smoke in the airway epithelium, at least in nasal epithelium, is quickly resolved once the damaging agent is removed. The remaining TFs showed different dynamics in the healthy and clinic groups. Confirming gene-level observations, 11 regulators of ciliary function were rapidly reversible in healthy subjects while cessation-associated in clinic subjects (7/11 passed the effect size threshold, Section 3.5). Two regulators of xenobiotic detoxification and oxidative stress response, *PIR* and *LHX6*, switched from rapidly to slowly reversible in the

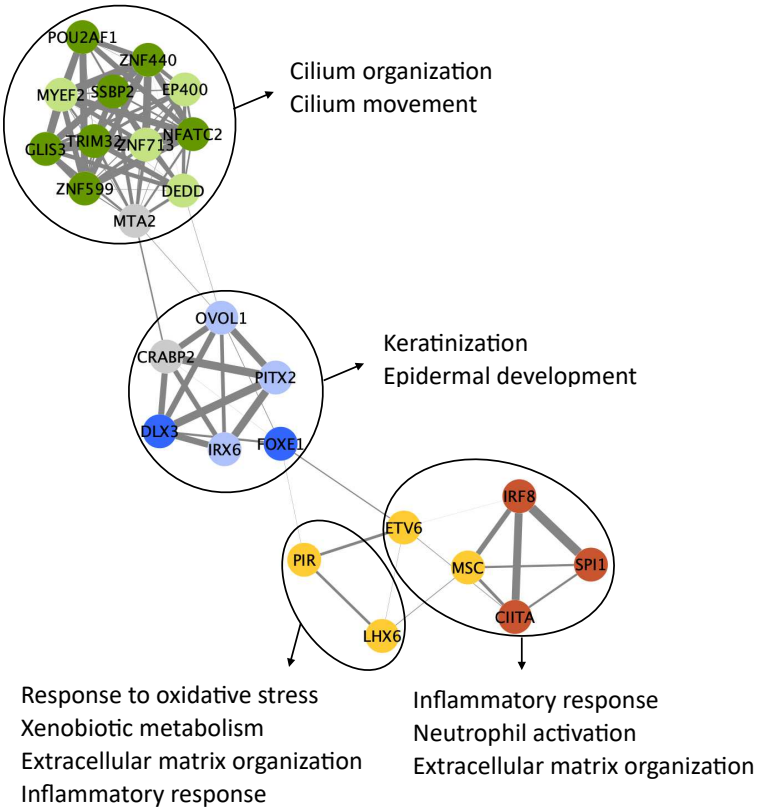
clinic group, suggesting that alterations of these pathways persist longer in clinic patients than in healthy subjects. Again, as observed at the gene level, a group of TFs showed no smoking-dependent change in healthy subjects, while being classified as slowly reversible or irreversible in the clinic group. All these TFs had lower activity in clinic current and ex smokers compared to healthy never smokers, and all of them are regulators of immune functions, including interferon-mediated signalling.

Together with the observations at the gene level and pathway level, these results indicate smoking-associated immune alterations as a specific response of the clinic group and a potential contributor to lung cancer risk in both current and former smokers.

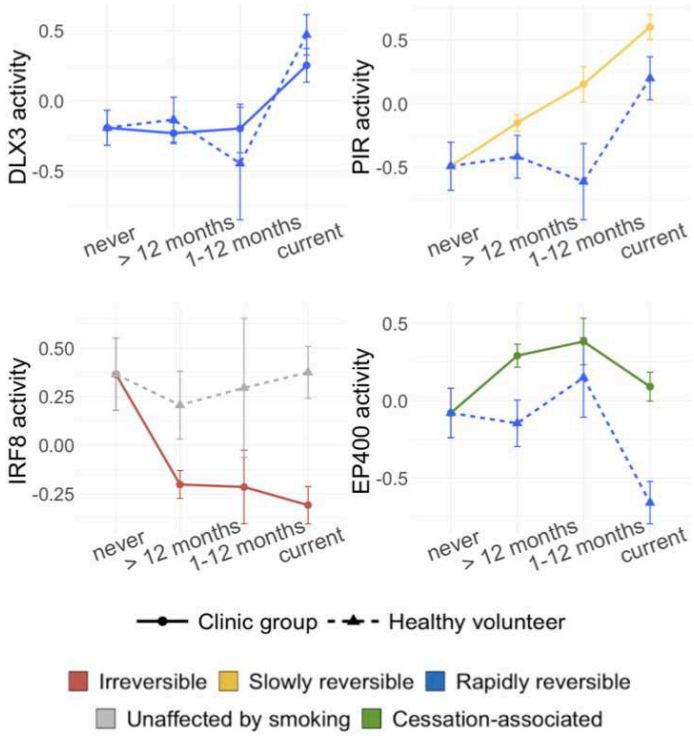
3.2.6 Using slowly reversible genes as a biomarker of past smoke exposure

Most studies investigating the effects of cigarette smoke on health rely on self-report for the annotation of patients' smoking history (Connor Gorber et al., 2009). Several studies have shown that self-reported information is often inaccurate, which can lead to misinterpretation of clinical outcomes and contradictory results across studies. Patients tend to misrepresent or minimise their cigarette use, specially if they belong to certain populations for which smoking is particularly frowned upon, such as pregnant women and individuals affected by lung diseases such as COPD and asthma (Aurrekoetxea et al., 2013; Stelmach et al., 2015). Self-report of smoking habits can be misleading even for truthful patients, as it does not consider other sources of exposure such as second-hand smoke. For these reasons, self-reported smoking status is often confirmed by a biochemical test in which the levels of cotinine, a metabolite of nicotine, are measured in the patient's blood or urine (Connor Gorber et al., 2009). However, cotinine can only be detected for a few days after smoking cessation, making it a valid biomarker for recent smoke exposure only. Moreover, being a metabolite of nicotine, cotinine can also be found in the fluids of non-smokers on nicotine-replacement therapy and of e-cigarette users.

In previous sections I reported a subset of genes affected by smoking whose expression slowly reverts back to healthy never smoker levels after smoking cessation. When these slowly reversible genes are used to perform a PCA, they order patients according to their smoking status, from the most to the least recently exposed to smoke (**Figure 3.5**). Therefore, these genes could be used as a biomarker of past smoke exposure, potentially


POU2AF1
ZNF440
MYEF2
SSBP2
EP400
NFATC2
GLIS3
TRIM33
ZNF713
ZNF599
DEDD
MTA2
Cilium organization
Cilium movement
OVOL1
CRABP2
PITX2
DLX3
IRX6
FOXE1
Keratinization
Epidermal development
IRF8
ETV6
PIR
MSC
SPI1
CIITA
LHX6
Response to oxidative stress
Xenobiotic metabolism
Extracellular matrix organization
Inflammatory response
Inflammatory response
Neutrophil activation
Extracellular matrix organization

(a)


DLX3 activity
PIR activity
IRF8 activity
EP400 activity
0.5
0.0
-0.5
0.50
0.25
0.00
-0.25
never
> 12 months
1-12 months
current
Clinic group
Healthy volunteer
Irreversible
Slowly reversible
Rapidly reversible
Unaffected by smoking
Cessation-associated

(b)

Fig. 3.9 *(previous page)* **Master regulators of smoke injury response. (a)** Network representation of the 25 master regulator TFs summarising the smoke injury response in nasal epithelium. Edge thickness indicates the overlap of TF target genes (Jaccard coefficient). Nodes are coloured based on their reversibility class in the clinic group; a lighter colour shade is used for TFs that are classified as RR, SR, IR or CA but do not pass the effect size thresholds; **(b)** activity over smoking status for 4 smoke injury MRs with different post-cessation dynamics in the clinic and healthy groups, with line-type and shape representing donor status and colours representing the reversibility classes assigned to genes.

able to detect current or very recent exposure, as well as exposure that occurred months or even years in the past. Such biomarker could be used to confirm self-reports by collecting a nasal sample in a non-invasive way.

To confirm this potential of our slowly reversible genes, I performed a pseudotime analysis (Section 3.5). In this analysis, patients are ordered along a pseudotemporal trajectory defined by their transcriptional state, usually reflecting the progression through a biological process. As input for the pseudotime analysis, I provided the nasal expression of the slowly reversible genes identified in Section 3.2.3. The resulting pseudotemporal ordering of the patients correlates with the patients' smoking status, with never smokers exhibiting the lowest pseudotime values and current smokers and former smokers who quit for less than 1 month the highest (**Figure 3.10**).

Smoking history of patients can thus be accurately described by the expression of our slowly reversible genes, and summarised via pseudotemporal ordering into a single value representing a "smoke exposure score".

3.2.7 Using nasal expression of smoke injury genes for disease risk prediction

In the previous sections, I described how I characterised the gene expression alterations induced by smoke injury at the level of nasal epithelium, and postulated that the observed differences in injury response in healthy and clinic subjects might explain their difference in lung cancer risk. In this section and the following, I assess the potential of using the expression of nasal smoke injury genes with a different behaviour in healthy and clinic subjects (749 genes, here referred to as "risk genes") to predict lung cancer risk. If nasal epithelium of current and former smokers contained valuable information on their disease status, it would be possible to devise a non-invasive lung

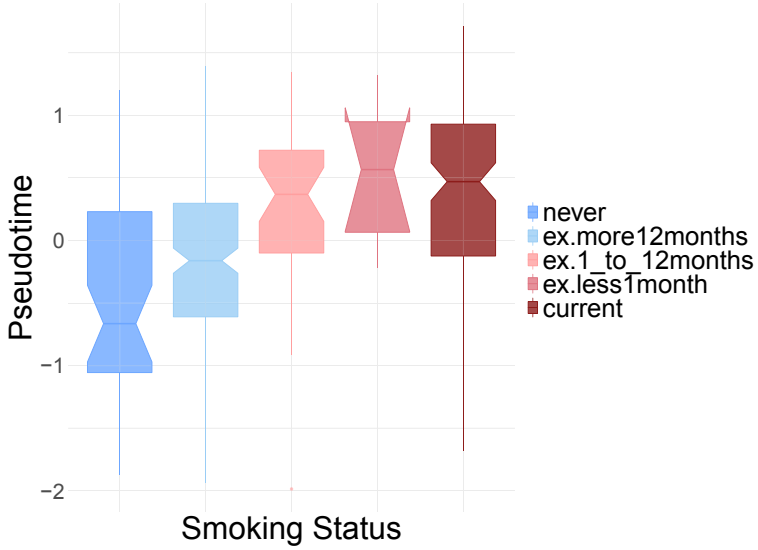


Fig. 3.10 **Distribution of pseudotime values stratified by smoking status of the subjects.** Pseudotime was inferred using genes classified as slowly reversible in Section 3.2.3.

cancer biomarker. Such biomarker could be used to assess the general population of current and former smokers, to establish a systematic pre-screening to prevent low-risk individuals from undergoing unnecessary follow-up. In the clinic context, where patients already show evidence of lung disease, the biomarker would help clinicians in the identification of patients with the highest need for further investigation.

Thus, we trained two independent classifiers using the 749 risk genes: a "population classifier" that predicts the donor group from which the samples were taken (clinic subject or healthy volunteer) and a "clinic classifier" that predicts the cancer status of each patient. For both classifiers, we used lasso-penalised multivariate logistic regression controlling for sex, age, smoking status and pack-years (Section 3.5), and derived a per-subject score from each classifier.

As expected, the population score clearly separated healthy volunteers from clinic subjects (**Figure 3.11a**), while the clinic score (**Figure 3.11b**) additionally distinguished cancer patients from patients with benign conditions within the clinic group. Interestingly, this classifier placed clinic benign subjects between healthy volunteers and clinic cancer subjects. Both classifiers were still able to separate the donor groups after regressing out all clinical covariates, showing that nasal gene expression data improves classification compared to the use of clinical covariates alone (**Figures 3.11c and 3.11d**). The two scores were highly correlated (0.88 Pearson correlation, p-value <1e-16), as expected, since the same set of genes is used as predictors (**Figure 3.11e**).

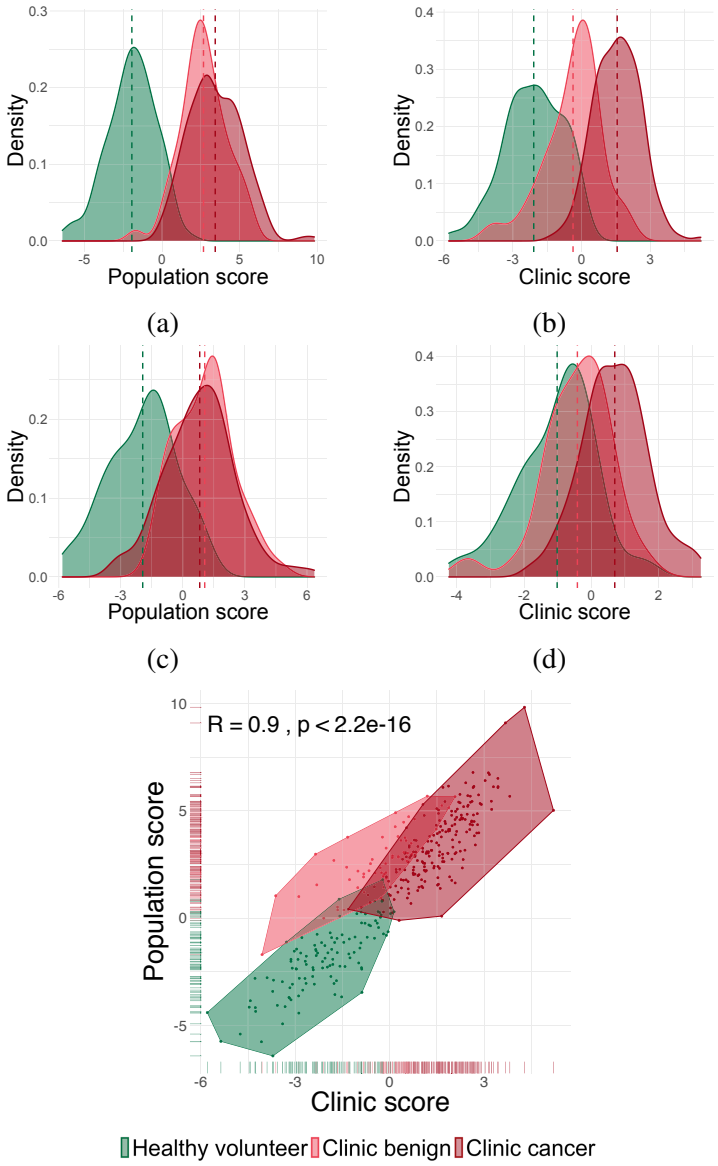
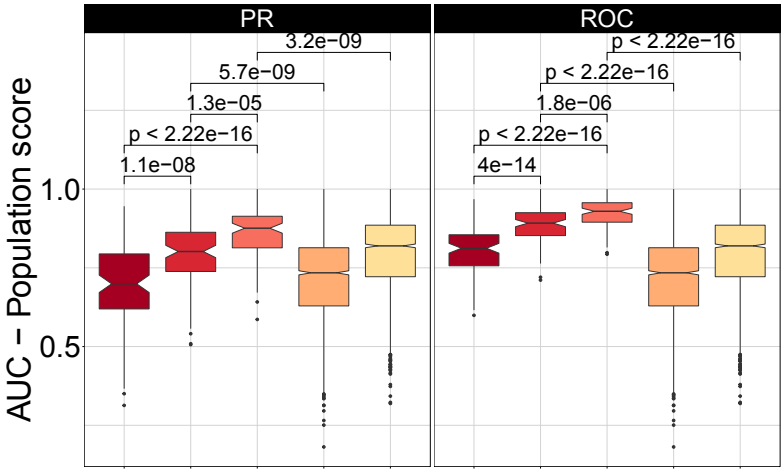


Fig. 3.11 **Population and clinic risk scores.** Risk score distributions for the population **(a)** and clinic **(b)** classifiers, predicted using clinical variables the expression of smoke injury 'risk genes'; **(c)** and **(d)** show the same distributions after regressing out the clinical covariates; **(e)** correlation between the population and clinic risk scores.

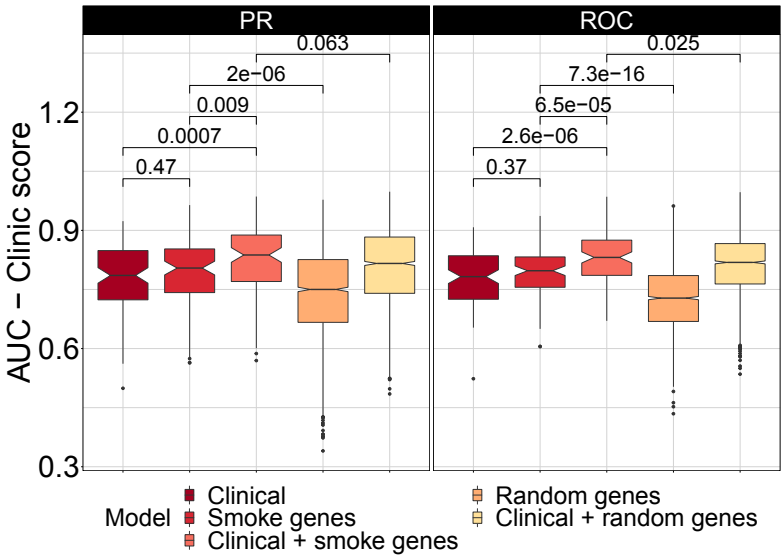
For both scores, we performed 10-fold cross-validation and calculated the area under the curve (AUC) for precision-recall (PR) and receiver operating characteristic curve (ROC). Both scores have mean AUC values greater than 0.8 over cross-validation (**Figure 3.12**). Moreover, we found that models that incorporate the expression of the risk genes performed significantly better than models built on clinical covariates alone (**Figure 3.12**). To confirm the contribution of the 749 risk genes to the predictive power,

we also compared our model with an equivalent one built using the expression of 749 randomly selected genes. Both the clinic and the population classifiers significantly outperformed the models based on random genes (**Figure 3.12**).

Furthermore, I separately looked at the distributions of risk scores in current and former smokers. The population risk score showed a difference in the score of healthy and clinic subjects for both smoking groups, observable even after regressing out clinical covariates (**Figures 3.13a and 3.13b**). Although the clinic score also separates benign from cancer patients in both smoking groups, the added value of gene expression information appears less important, particularly in former smokers (**Figures 3.13c and 3.13d**).



(a)



(b)

Fig. 3.12 *(previous page)* **Performance of different classification models based on clinical covariates and nasal gene expression.** Distribution of AUC values over 100 rounds of cross-validation for the population **(a)** and clinic **(b)** risk classifiers. Performance is shown for models including clinical variables only (Clinical), expression of the 749 risk genes only (Smoke genes) and a combination of clinical variables and expression of the 749 risk genes (Clinical + smoke genes); Additionally, performance is shown for a model trained on the expression of a set of 749 randomly selected genes (Random genes) and a model trained on a combination of clinical variables and a set of 749 randomly selected genes (Clinical + random genes). For both classifiers, both the area under the precision-recall (PR) and receiver-operating characteristics curves are reported.

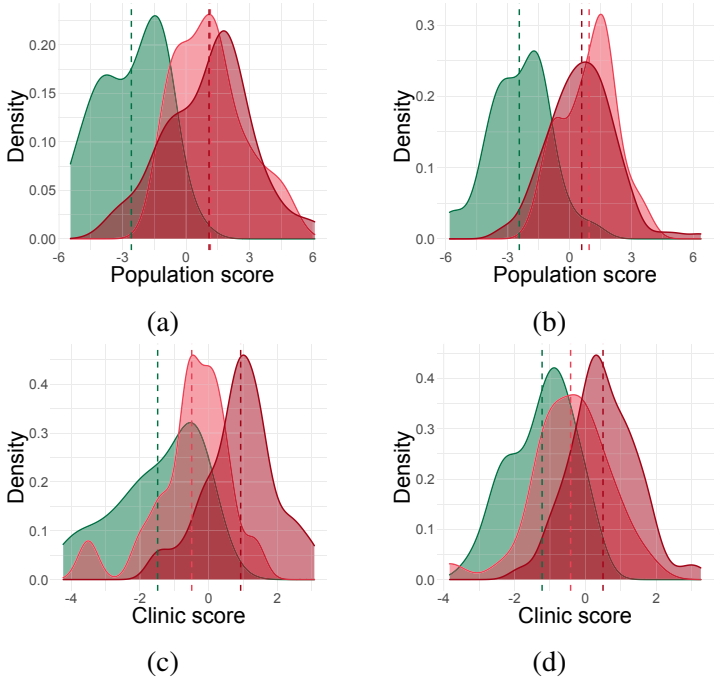
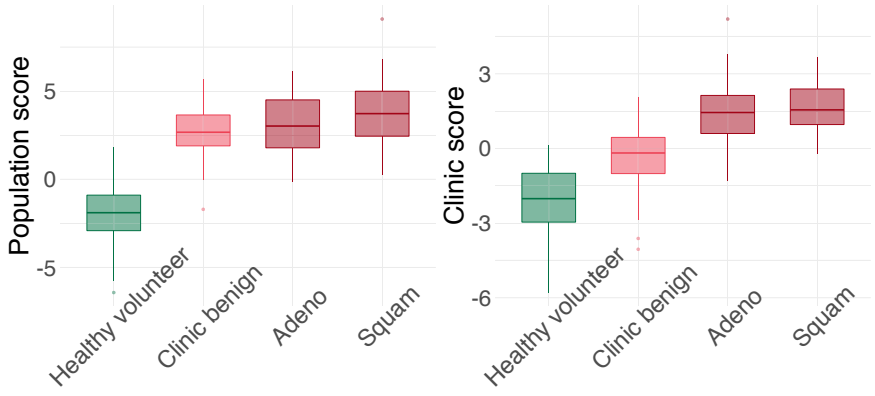


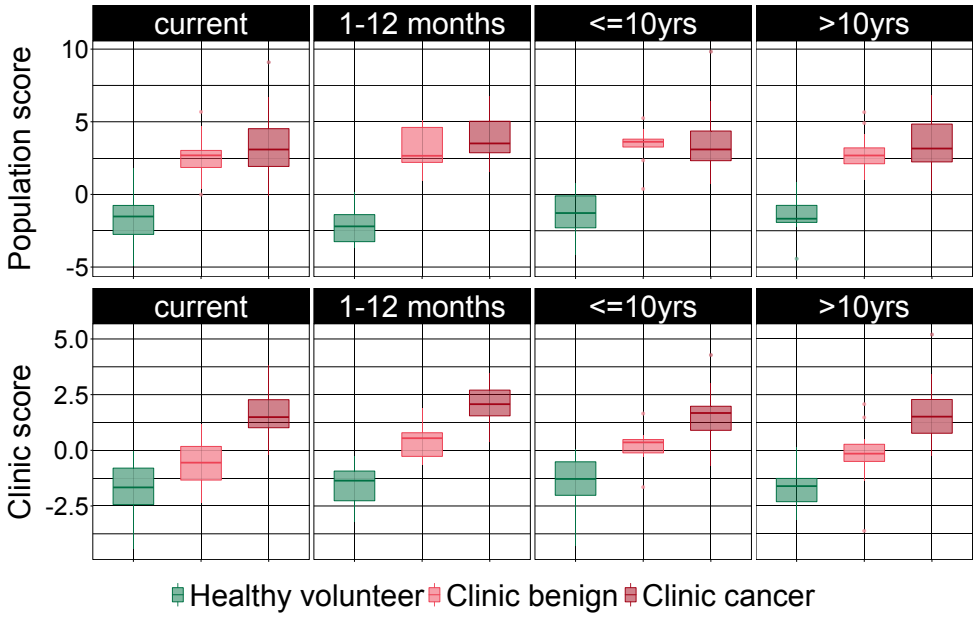
Fig. 3.13 **Risk score distributions in current and former smokers.** Population and clinic risk score distributions for current smokers **(a-c)** and former smokers with time-since-quit longer than 1 year **(b-d)** after regressing out all clinical covariates.

Additionally, our clinic classifier separates clinic benign from clinic cancer patients regardless of their cancer type (squamous carcinoma or adenocarcinoma, **Figure 3.14a**), and both classifiers detect increased risk in subjects who quit smoking even for more than 10 years (**Figure 3.14b**).

Finally, the work presented in this chapter is based on the notion that cigarette smoke produces an extensive field of injury, which is comparable in lower airway tissues, such as bronchial epithelium, as well as more accessible upper airway tissues, such as nasal epithelium. Thus, I tested our clinic model on bronchial samples collected from



(a)



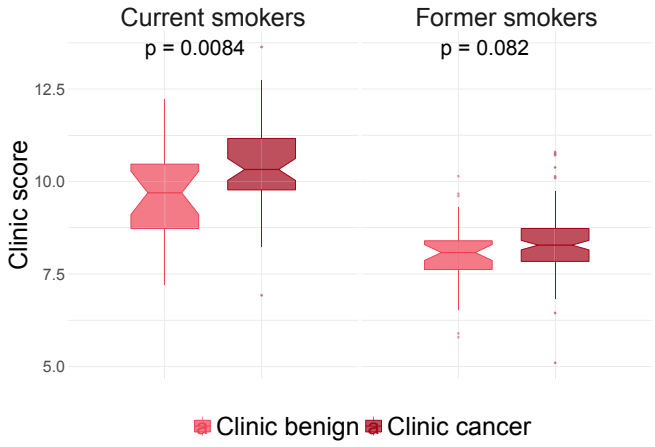
(b)

Fig. 3.14 **Distributions of population and clinic risk scores in different NSCLC subtypes and in long term ex smokers.** Population and clinic risk scores distinguish cancer patients independently of NSCLC subtype **(a)** and smoking status **(b)**, including former smokers with time-since-quit longer than 10 years.

clinic patients. The model separated patients with cancer from patients with benign conditions in both current and former smokers (**Figure 3.15a**). Even a model that included only gene expression information distinguished between the two groups, although the difference was not significant for former smokers (P = .082, **Figure 3.15b**). Showing that our classifier works in a deeper airway tissue confirms that gene expression changes similar to those occurring in the lower airway, closer to the tumour site, can be observed at the level of the nasal epithelium.

Overall, our results demonstrate that classifiers based on smoking-induced transcriptional alterations in nasal epithelium have the potential to improve risk stratification of current and former smokers, both in the context of a population screening and in a clinical setting.

(a)



(b)

Fig. 3.15 **Clinic risk score applied to bronchial samples from clinic patients.** Distribution of patients' risk scores with **(a)** and without **(b)** including clinical covariates in the model.

3.2.8 Immune alterations drive lung cancer risk classification

To better understand the mechanisms leading to increased risk, I identified genes contributing most to the population and clinic scores. In particular, I looked at genes that were selected by the lasso procedure in more than 80% of CV rounds **Figure 3.16**. Among the risk genes selected most frequently in the population model were the interferon-regulated *IFI44* and *SAA2*, encoding for a protein found at higher levels in

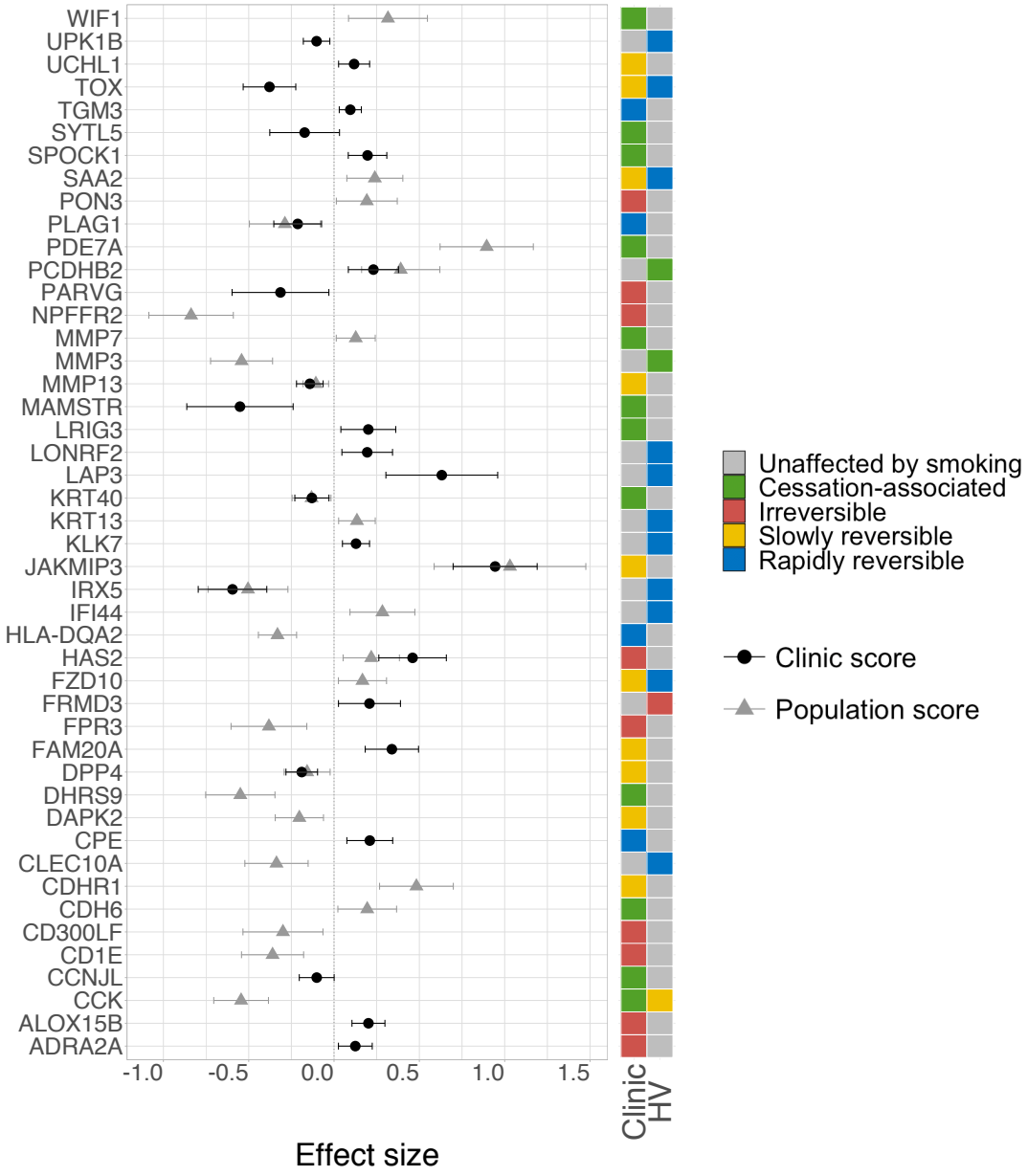
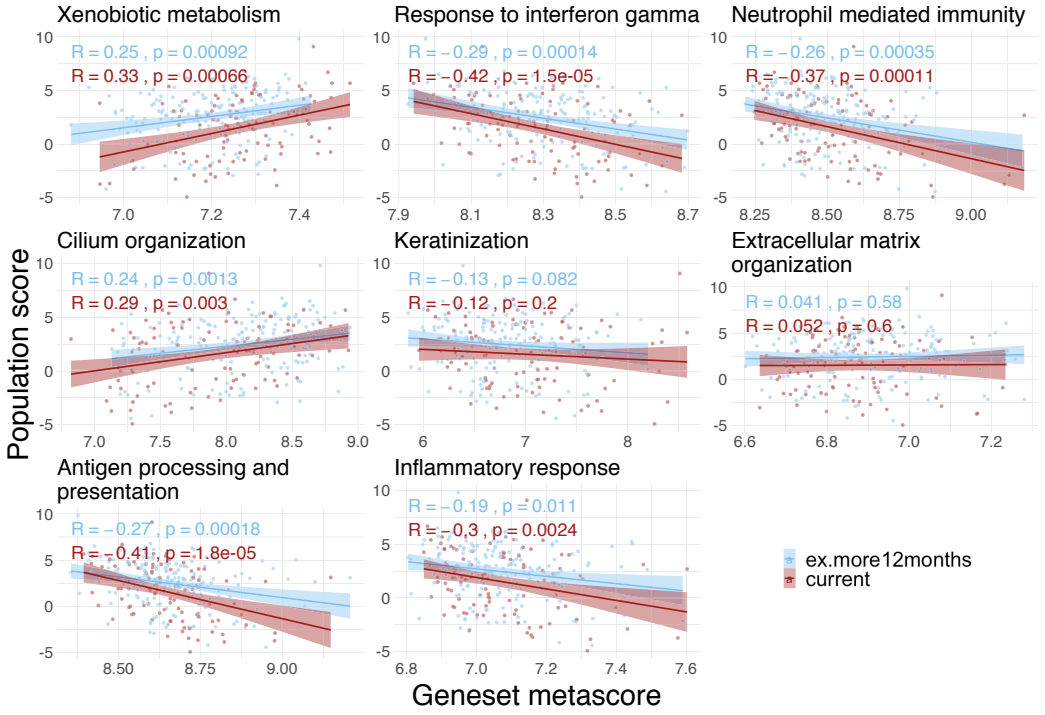


Fig. 3.16 **Top contributing genes to the population and clinic risk classifiers.** The weight of the genes selected in more than 80% of cross validations in the population and clinic classifiers; the presented value is the mean over all cross validation and the error bars represent standard deviation; the annotation track on the right shows the reversibility classes of the genes in the healthy and clinic groups.

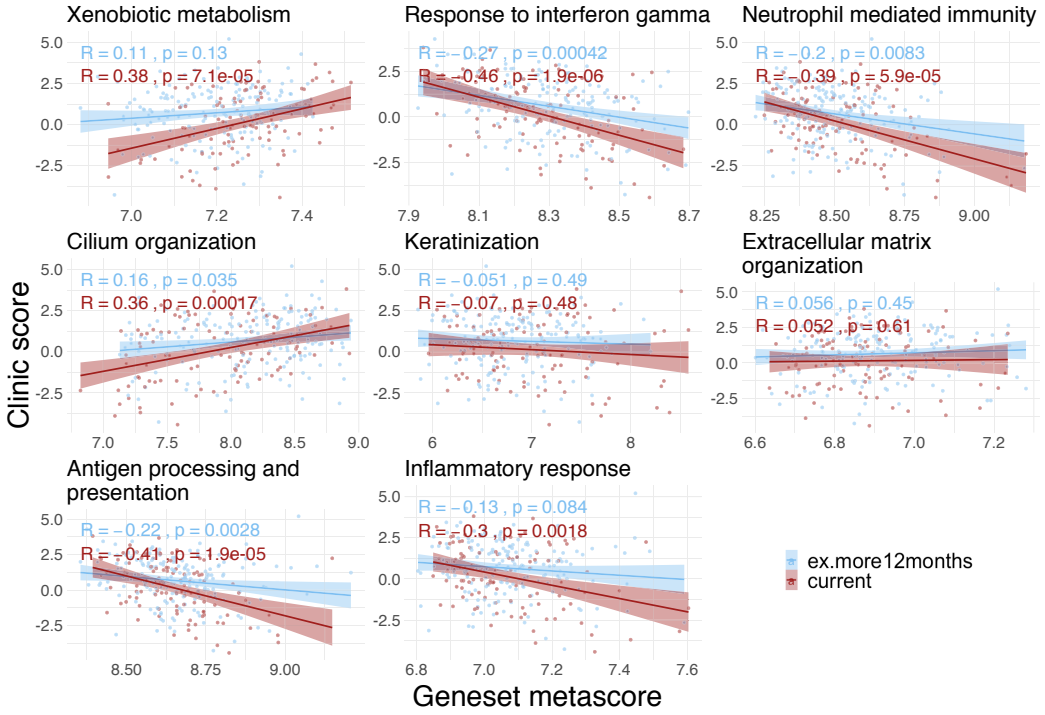
the blood and tissues of lung cancer patients compared to healthy subjects (Critchley-Thorne et al., 2009; Sung et al., 2011). Risk genes selected most frequently for the clinic score included *TGM3*, a gene involved in EMT in several cancer types (Feng et al., 2020; Hu et al., 2020; Uemura et al., 2009; Wu et al., 2013) and *PDE7A*, also shown to be involved in EMT (Kolosionek et al., 2009). Some of the risk genes selected most frequently for both population and clinic score were *MMP13*, encoding a metalloproteinase known for its involvement in NSCLC (Merchant et al., 2017), and *HLA-DQA2*, a member of the major histocompatibility complex (MHC); several studies have shown the involvement of the MHC in tumour development (Seliger et al., 2017).

Since all risk genes used as predictors for our classifiers are involved in the response to smoke injury, it is difficult to gain insight into the mechanisms of risk by only looking at genes frequently selected by the lasso procedure. In order to identify, among the cellular processes disrupted by smoking, the ones that contribute most to increased risk, I used the geneset metascores calculated for the 8 smoking-associated GO terms mentioned in Section 2.2.4. Then I calculated the correlation between the per-subject geneset metascores and the population and clinic risk scores. I calculated these correlations for current and former smokers (> 12 months) separately, to be able to identify differences in geneset contribution to risk in the two groups that might reflect differences between acute smoke injury response and the long-term consequences of past smoke exposure (**Figures 3.17a and 3.17b**). In current smokers, while *Keratinization* and *Extracellular matrix organization* did not significantly correlate with either risk score, the remaining four genesets tested showed moderate but significant correlation with both risk scores, pointing to alterations of the xenobiotic detoxification pathways, ciliary function and immune response as the main contributors to patient-specific differences in risk. In former smokers, the population risk score correlated with the same 4 GO terms, indicating that detoxification pathways, ciliary function and immune response are the main contributors to risk. In contrast, only pathways related to immune alterations, (*Response to interferon gamma* and *Neutrophil-mediated immunity*), correlated with the clinic risk score in former smokers, while no correlation was observed with Xenobiotic metabolism, and only a very weak correlation with *Cilium organization*.

These results indicate that immune alterations are significant contributors to the risk of cancer in both current and former smokers in the clinic group.



(a)



(b)

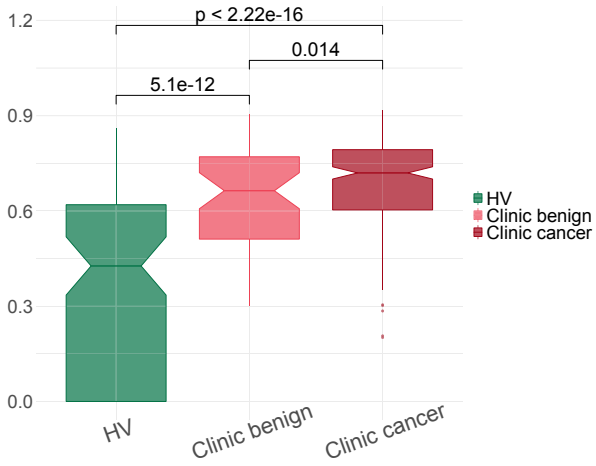
Fig. 3.17 **Correlation of geneset metascore and risk scores.** Correlation between the population **(a)** and clinic **(b)** risk score and geneset metascore for 8 genesets representing biological functions altered by smoking; correlation is shown separately for current (red) and former (blue) smokers; shaded areas around the fitted line indicate 95% confidence interval.

3.3 Comparison with existing lung cancer classifiers based on nasal gene expression

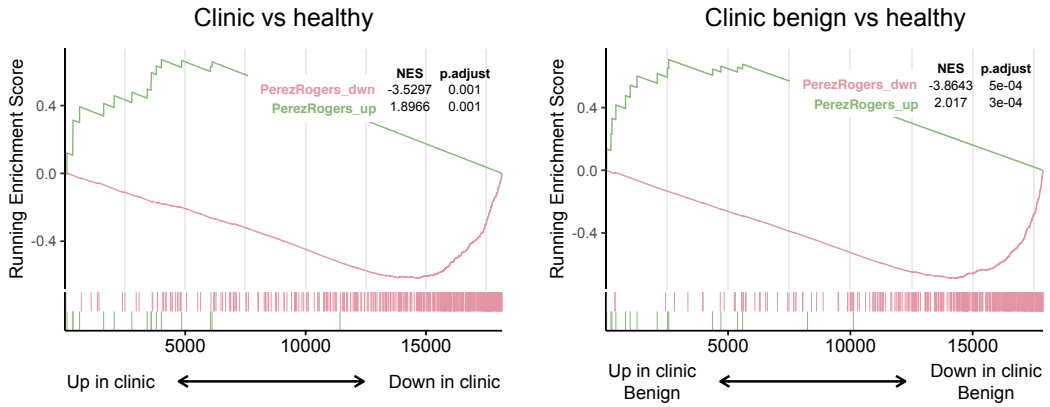
In 2017, Perez-Rogers et al. (2017) described a lung cancer classifier based on gene expression changes in the nasal epithelium of cancer patients compared to patients with benign conditions. The classifier was trained on microarray expression profiling data from the AEGIS study, where nasal epithelial brushings were collected from current and former smokers undergoing diagnostic procedures for pulmonary lesions suspicious of lung cancer. The authors found $\sim$500 genes with differential expression between cancer patients and patients with benign disease, and built their classifier using 30 of these genes, in addition to clinical information (age, smoking status, time since quit, size of lesion).

When applying the classifier from Perez-Rogers et al. (2017) to our cohort, I observed a clear separation of healthy volunteer subjects from clinic patients, regardless of cancer status, while only a weak separation of clinic cancer from clinic benign patients (**Figure 3.18a**). I hypothesised that the reason lay in the different composition of the "benign" group in the two cohorts. The AEGIS benign group might be closer in composition to our healthy volunteer subjects than to our clinic benign patients. I tested this hypothesis by using gene-set enrichment analysis (GSEA). I ranked all genes based on their fold-change in clinic compared to healthy subjects, and tested for enrichment of Perez-Rogers' cancer signature gene at the top and bottom of the ranked list. I found a significant enrichment of Perez-Rogers' up-regulated and down-regulated cancer genes at the top and bottom of the ranked gene list, respectively. Moreover, I performed a similar GSEA, ranking genes by their fold-change in clinic patients, benign only, compared to healthy volunteers. I observed an even stronger enrichment of Perez-Rogers' cancer signature genes at the top and bottom of the ranked list (**Figure 3.18b**). These results suggest that Perez-Rogers' cancer signature reflects differences between healthy current and former smokers and people who developed smoking-associated respiratory symptoms and pathologies, including lung cancer.

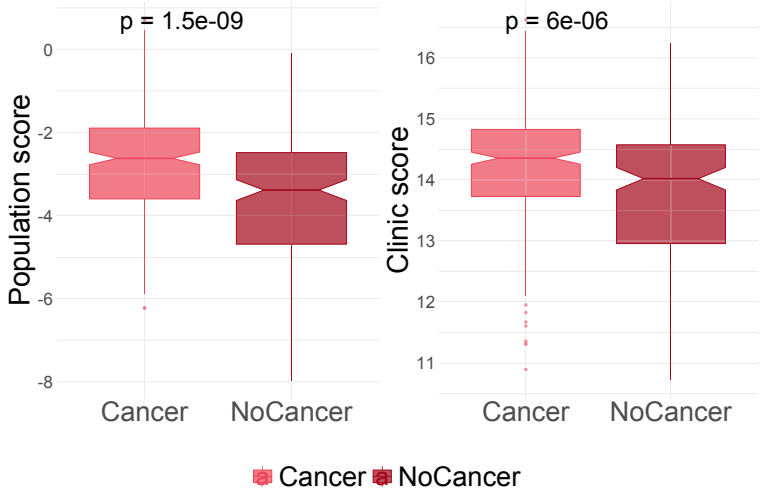
I then tested our population and clinic classifiers on the AEGIS cohort used in Perez-Rogers et al. (2017). Even though the AEGIS cohort was microarray-based, and the samples derived from a different clinical context, both classifiers distinguished patients with and without cancer (**Figure 3.18c**). The strongest separation was observed with the population classifier ($P = 1.5 \times 10^{-9}$ compared to $P = 6 \times 10^{-6}$ for the clinic classifier), as expected from the observations described above in this section.


p < 2.22e-16
0.014
5.1e-12
1.2
0.9
0.6
0.3
0.0
HV
Clinic benign
Clinic cancer
HV
Clinic benign
Clinic cancer

(a)


Clinic vs healthy
Running Enrichment Score
NES p.adjust
PerezRogers_dwn -3.5297 0.001
PerezRogers_up 1.8966 0.001
0.4
0.0
-0.4
5000
10000
15000
Up in clinic
Down in clinic
Clinic benign vs healthy
Running Enrichment Score
NES p.adjust
PerezRogers_dwn -3.8643 5e-04
PerezRogers_up 2.017 3e-04
0.4
0.0
-0.4
5000
10000
15000
Up in clinic Benign
Down in clinic Benign

(b)


p = 1.5e-09
Population score
0
-2
-4
-6
-8
Cancer
NoCancer
p = 6e-06
Clinic score
16
15
14
13
12
11
Cancer
NoCancer
Cancer
NoCancer

(c)

Fig. 3.18 *(previous page)* **Comparison with nasal lung cancer classifier from Perez-Rogers et al. (2017). (a)** Perez-Roger's classifier applied to the nasal samples in the our cohort. **(b)** GSEA plot showing the concordance between Perez-Roger's nasal cancer signature (535 genes) and the genes differentially expressed between the CRUKPAP clinic patients and healthy volunteers (5359 genes); the bottom track represents a list of all protein-coding genes ranked according to their fold-change in clinic compared to healthy subjects; vertical bars represent the position in the ranked list of genes up- (green) and down-regulated (pink) in cancer patients in Perez-Roger's signature; pink and green solid lines represent the running enrichment score for Perez-Roger's signature genes moving down the ranked list; in the top right corners are reported the final enrichment scores and the associated p-values. **(c)** Our population and clinic classifiers applied to the AEGIS cohort.

3.4 Discussion

In this chapter, I explored the field of injury created in the airway by an environmental damaging agent, cigarette smoke, and its involvement in determining lung cancer risk. I focused on the field of injury at the level of nasal epithelium, a tissue of the upper airway, accessible through non-invasive procedures, and thus potentially useful as a diagnostic proxy for lower, less accessible, regions of the airway. The cohort analysed included subjects with a wide range of smoking histories, and it was the first to date to include healthy subjects, with no indication of lung disease, and clinic-referred patients, with symptoms suggesting the presence of lung cancer or other lung diseases. The particular composition of this cohort allowed us not only to study the effects of smoking in nasal epithelium, and the reversibility of the smoking-associated injury after smoking cessation, but to compare these dynamics in subjects exposed to the same insult, but with different outcomes, namely healthy and clinic-referred subjects, to try to discover differences in the two groups that could explain differences in the risk of developing lung cancer.

First, I compared the nasal transcriptional landscape of healthy and clinic subjects to identify overall differences in gene expression between the two groups. I found that one of the major features distinguishing healthy and clinic subjects is a decrease in the expression of genes involved in immune response in clinic subjects, which suggests the presence of immunosuppression in current and former smokers who develop lung disease.

I then thoroughly characterised the smoke injury response, at the level of nasal epithelium, in healthy volunteers. The genes identified as affected by smoking, and

the biological pathways they were enriched in, were largely consistent with those from other studies of 'smoke injury genes' (Beane et al., 2007). I also showed that most of the genes whose expression was affected in current smokers reverted back to never-smoker level rapidly after smoking cessation.

I compared this "healthy" response to smoke injury with the response observable in clinic patients with symptoms of lung disease. As for the analysis in healthy volunteers, here we also used healthy never smoker gene expression as the baseline value. While it might be argued that a group of never smokers from the clinic group would be a better control group to identify smoking-associated expression changes in the clinic group, we chose to use healthy never smokers for two reasons. Firstly, only a few never smokers in our cohort belong to the clinic group, the number being too small to grant enough power for the analysis. Secondly, and more importantly, even a larger number of clinic-referred never smokers would not constitute an appropriate control group for this analysis: although they were never actively smoking, these subjects show respiratory symptoms typically associated with smoke exposure (including COPD and cancer), and thus their samples are not representative of healthy airway unaffected by smoke. Thus, in the absence of longitudinal data, never smokers from the healthy group best serve as a baseline for this analysis.

A strikingly different response to smoke was observed in clinic patients compared to healthy volunteers, with a larger number of affected genes whose expression appeared slowly reversible or irreversible post-cessation in the clinic group. Notably, most of the slowly reversible and irreversible genes in the clinic group were unaffected by smoking in healthy volunteers. This suggests that these genes might be affected by smoking exclusively in higher-risk subjects, or exhibit an overall expression difference between individuals in the healthy and clinic groups (maybe determined by germline variation), or a combination or interaction of the two. Most of these genes showed reduced expression in current smokers, and pathway analysis showed enrichment for immune-related genes including response to interferon gamma, neutrophil activation, chemotaxis and inflammation. Analysis at the level of transcription factor activity within a nasal epithelium gene regulatory network showed a similar pattern, suggesting again a role for immune-depression in determining risk of lung disease.

Also interesting to note is the large number of cilia-related genes classified as cessation-associated in the clinic group. The unusual trajectory of these genes is due to their conflicting behaviour over smoking status and between healthy and clinic subjects. Consistent with cigarette smoke damaging airway cilia (Prasetyo et al., 2021), cilia-related genes were down-regulated in current smokers in both donor groups. However,

cilia-related genes showed increased expression in the clinic group compared to healthy volunteers, both in current and former smokers. This increased expression of cilia genes in the clinic group might be due to the decreased expression of interferon-gamma-related genes in the same group, as it has been shown that interferon gamma suppresses ciliogenesis and ciliary movement (Chen et al., 2020).

Finally, by using genes exhibiting a different response to smoke in healthy and clinic subjects, we devised two lung cancer risk classifiers, with potential application in different clinical contexts. The population classifier, in particular, is the first to address lung cancer risk stratification in the healthy current and former smoker population, and it has an average cross-validated AUC (ROC) of 0.92, meaning that it identifies 95% of high-risk individuals with a false positive rate of $\sim$40%. We also show that our population classifier is effective in both current and former smokers. In particular, in line with evidence indicating persistent cancer risk long after smoking cessation, clinic patients have an elevated risk score more than 10 years after smoking cessation (Peto, 2011). The classifier is also equally efficient at identifying individuals with early or late stage disease and squamous or adenocarcinoma. This classifier, if validated in an independent group of healthy volunteers with a history of smoking, could help improve population-level risk stratification with the use of a non-invasive test.

We compared our classifiers with the one described in Perez-Rogers et al. (2017). Interestingly, our population classifier performed better than our clinic classifier in separating cancer and benign patients within the AEGIS cohort. At the same time, Perez-Rogers' classifier clearly separated our healthy volunteer group from our clinic-referred patients, but the separation was weaker for benign and cancer patients within the clinic group. These results might reflect the different composition of the cohorts compared. The AEGIS cohort is entirely composed of subjects presenting with pulmonary lesions and referred to clinical investigation for diagnosis. These patients would thus place within the clinic group, if compared to our CRUKPAP cohort. However, we could speculate that country-dependent differences in screening frequency and diagnostic protocols could have had an impact on the composition of these benign groups, with the AEGIS benign patients being closer to a healthy current/former smoker population than the CRUKPAP benign patients.

Throughout this study, alteration of certain immune-related functions appeared to be a key feature of the clinic-referred group of patients, which clearly distinguishes them from the healthy current and former smokers. Genes involved in these immune functions were also identified as the main contributors to our risk scores in both current and former smokers. The two pathways recurrently emerging throughout our analysis

were *Neutrophil mediated immunity*, *Response to interferon gamma* and *Antigen processing and presentation*. All these pathways are known for their involvement in lung cancer development.

Neutrophils play a complex role in the tumour immune microenvironment. There are two different types of tumour-associated neutrophils, named N1 and N2, that exhibit opposing pro-tumorigenic and anti-tumorigenic effects (Gonzalez et al., 2018; Mackey et al., 2019; Rosales, 2018). N1 neutrophils are pro-inflammatory cells, exerting their anti-tumorigenic function by enhancing cytotoxycity and increasing the secretion of immuno-activating cytokines. N2 neutrophils were shown to promote proliferation, extracellular matrix remodelling and angiogenesis and to inhibit anti-tumoral immune response. Furthermore, these two populations of neutrophils play their roles at different stages of tumour development. N1 cells are more abundant at early stages and, as the tumour progresses, they are displaced by N2 cells, a phenomenon known as neutrophil polarisation. In our study, we observed decreased expression of neutrophil-related genesets in the clinic group compared to healthy volunteers. Since the tissue examined is distant from the cancer site, and the decrease is observable also in patients with benign conditions, we can speculate that the anti-tumour neutrophil population is less active in clinic-referred patients, making them at higher risk to contract lung cancer.

IFN-γ is a molecule with an important role in anti-tumour immune response. It activates cellular immunity in an inflammatory environment, has anti-proliferative, pro-apoptotic properties and it has been shown to inhibit angiogenesis (Jorgovanovic et al., 2020). The lower expression of IFN-γ we observe in clinic patients might play a role in increasing their risk of developing lung cancer. A decrease in expression of genes involved in IFN-γ signalling, as well as in antigen presentation, was also observed by Pennycuick et al. (2020) in persistent and progressive bronchial premalignant lesions. The authors concluded that an immunosuppressive microenvironment, which has been documented before in the presence of lung cancer (Altorki et al., 2019), is already present during premalignancy, and might promote the progression to invasive disease. We observe these alterations at an even earlier step of carcinogenesis, in healthy-appearing, upper airway tissue affected by the smoking-associated field of injury. Moreover, further work conducted in our group linked known lung cancer GWAS variants to changes in the expression of 41 genes in our cohort (de Biase, Massip et al., 2021). These genes overlapped with the list of smoke injury genes described in this chapter; in particular, they were enriched for genes involved in response to IFN-γ and antigen presentation. These findings provide a first evidence of a causal link between germline variants and individual response to smoke, with

patient-specific genetic background increasing the risk of lung cancer insurgence by creating an immunosuppressive environment.

3.5 Methods

Cohort and sample collection

487 donors were recruited into the CRUKPAP cohort at Royal Papworth Hospital, Cambridge (UK), including 114 healthy volunteers (HV) and 337 patients being investigated for suspicion of lung cancer. From these donors 413 nasal epithelial curettages were collected using Arlington Scientific ASI Rhino-pro nasal curettes. Briefly, the nostril is opened with a nasal speculum to identify the inferior turbinate. Under direct vision the tip of the nasal curette is gently scraped over the turbinate to obtain a 'peel or curl' of epithelial tissue. The curl of tissue is then removed by flicking the curette while the tip is submerged in RNAlater™ collection medium and presence of the curl floating in the medium is confirmed by visual inspection. This procedure is repeated twice for each nostril per donor. RNA integrity (RIN) was checked for all samples and we found >80% of samples to have RIN 6 or better.

Bronchial brushings were collected using 2.0mm brush diameter cytology brushes (Olympus Medical, UK) from 236 patients undergoing flexible bronchoscopy as part of investigations for suspected lung cancer.

For 162 donors, both nasal and bronchial samples were available. All samples underwent short-read total RNA sequencing using Illumina TruSeq library generation for the Illumina HiSeq 2500 platform. Blood samples were taken from 467 donors and germline genotyped using the Illumina Infinium Oncoarray platform at 450K tagging germline variants. Total gene expression levels (TPM and variance stabilised) were determined for 18,072 protein coding genes for all samples using DeSeq2 v1.26.0 Love et al. (2014). Research ethics approvals for sample collection from participants in this study were given by East of England Cambridge Central REC 13/EE/0012 and the National Research Ethics Service Committee South East Coast – Surrey 13/LO/0889.

RNA extraction and sequencing

Tissue samples from bronchial brushings and nasal curettes were stored in 500μl RNALater overnight at 4 °C, and then at -80 °C for longer-term storage. RNA was extracted using Qiagen MiRNeasy columns according to the manufacturer's protocols. Briefly, bronchial brushes were rinsed in PBS, brushes transferred into 700μl Qiazol

and cells lysed by vortexing twice for 30 seconds. For nasal samples the RNALater containing nasal tissue (500μl) was diluted with 2ml of PBS and spun at 10,000 rpm for 10 min. The cell pellet was lysed by re-suspension in 700μl Qiazol. For both types of samples, the Qiazol lysate was applied to a QiaShredder tube (#217004) and spun at 13,000 rpm for 2 mins. The homogenate was kept at room temperature for 5 mins, followed by chloroform extraction using PhaseLock tubes. Nucleic acids in the aqueous phase were precipitated using 1.5 volumes of 100% ethanol and DNA was digested using DNAse I. Finally, RNA was isolated from the mixture using RNAeasy mini spin columns. RNA was quantified using a Qbit measurement and quality assessed using an Agilent Bioanalyzer. For samples with a RIN greater than 7, a total of 500ng of RNA was used for Illumina TruSeq Library generation. Sequencing was carried out on HiSeq 2500 Illumina sequencers. Sequencing was carried out in two separate multiplexed experiments.

RNA sequencing data processing

Alignment was carried out with TopHat2 (Kim et al., 2013), using as reference the human genome version GRCh37. Read counts were computed for all protein-coding genes with subread featureCounts v1.6.0 (Liao et al., 2014). The data was produced in two experimental batches, producing a strong batch effect that can be observed in the raw data. Moreover, a group of samples from one batch has lower total counts compared to the other samples. Raw counts were normalised using DESeq2's variance-stabilising transformation, which had the advantage of partially correcting the previously mentioned batch effects. Genes with across-samples log variance smaller than -4 were discarded from further analysis. Total gene expression levels (variance stabilised) were determined for 18,072 protein-coding genes for all samples.

Variance components analysis and differential expression analysis

Variance components analysis was performed using R package variancePartition v1.16.1 (Hoffman and Schadt, 2016). The experimental batch effect was regressed out of the vst-normalised expression before extracting variance components.
All differential expression analyses were performed with DESeq2 v1.26.0. Age, experimental batch, sex and pack-years were included as confounding variables. Genes with multiple-testing-adjusted (Benjamini-Hochberg) p-values < 0.05 were considered differentially expressed. For differential expression between clinic cancer and clinic benign in bronchial samples, 8 genes had artificially high (>20) absolute fold-change, due to their very low average expression across samples. These genes were removed from the list of differentially expressed genes.

Modelling time-dependent dynamics of smoke injury in nasal tissue

To identify genes affected by smoke and characterise their post-cessation expression dynamics, we applied Bayesian linear regression and model selection using R package BAS v1.5.3 (Clyde, 2018). To identify genes for which smoking has the strongest effect, we applied a threshold on the beta coefficient and retained only genes with a beta CS greater than 0.4 for rapidly reversible, slowly reversible and irreversible genes, and a beta FSS greater than 0.25 for cessation-associated genes.

To model the dynamics of transcription factor activity over smoking status, a context-specific protein-protein interaction network for nasal epithelium was built using ARACNe-AP (Lachmann et al., 2016) on the vst-normalised expression data and a list of 1988 human transcriptional regulators, compiled using information available in public databases (Ravasi et al., 2010). ARACNe-AP was able to infer nasal context-specific interactions across 1548 regulators. The activity of each of these regulators in each nasal sample was inferred using VIPER v1.20.0 (Alvarez et al., 2016). A Bayesian regression and model selection approach was used, similarly to what described above, to model transcription factor activity on smoking status and assign each transcription factor to a reversibility class among unaffected by smoking, rapidly reversible, slowly reversible, irreversible or cessation-associated. The same thresholds for effect sizes applied to the gene expression results were applied here.

To test for enrichment of the smoke injury genes within the regulons of smoke injury TFs, we performed a hypergeometric test; we corrected p-values for multiple testing using the Benjamini-Hochberg method. Network representations of TF-TF and TF-targets interactions were produced with Cytoscape v3.8.1 (Shannon et al., 2003).

For the network representation of the 25 smoke injury master regulators, the Jaccard coefficient was calculated for each pair of TFs as the intersection of the targets of the pair, divided by their union. Based on the Jaccard coefficient, the TFs appear to cluster in smaller groups on the network. We manually defined the groups, and performed functional enrichment analysis (with Gene Ontology terms) on the union of each group's target genes to identify the biological functions regulated by each group.

Pseudotime analysis was performed with R package phenopath v1.10.0 (Campbell and Yau, 2018), using a constant value for the covariate vector.

Derivation of population and clinic risk scores

L1-penalised multivariate logistic regression was performed with R package glmnet 3.0-2 (Simon et al., 2011) using only the nasal gene expression data. Patient status

was encoded with a binary variable (cancer: 1, no cancer: 0 for the clinic classifier; clinic patient: 1, healthy volunteer: 0 for the population classifier), and patients with *Ineligible* status were excluded from the analysis. In the gene expression classifiers, the status of each patient was predicted based on the expression of the 749 risk genes and 4 clinical covariates, namely sex, age, smoking status and pack-years, all of which were encoded as numerical variables (smoking status encoding: Never smoker: 0, former smoker >1year: 1, former smoker 1-12months: 2, former smoker <1m: 3, current smoker: 4). For the clinical classifier we also used a lasso regression, using only sex, age, smoking status and pack-years as predictors. The lasso shrinkage parameter (λ) was chosen to minimise the mean cross-validated error ("lambda-min" option in the cv.glmnet function).

Area under the receiver operating characteristic curve and precision recall curves were computed using R package PRROC (Grau et al., 2015), after 10 rounds of 10-fold cross validation experiments. To compare performances of the risk genes to performances on random genes, we randomly drew 20 sets of 749 genes among the 18,072 protein-coding genes retained for all analyses, and cross validations experiments were conducted on the same test and training set as the one used with the risk genes.

The clinic classifier was applied to bronchial samples by choosing genes selected by the lasso procedure in >80% of CV rounds, averaging the β coefficients of these genes across CV, and using these values to calculate the per-patient risk score.

The population and clinic classifiers were applied to the AEGIS cohort as described for bronchial samples. Age, pack-years, sex and smoking status information reported for AEGIS were encoded to match as closely as possible our encoding of clinical variables.

The clinicogenomic classifier described in Perez-Rogers et al. (2017) was applied to our data as described in the paper, including all clinical variables except "mass size", which is not available for the CRUKPAP cohort. Since the AEGIS data is microarray-based, quantile normalisation was performed on the CRUKPAP data. The R function *normalize.quantiles.use.target* from the *preprocessCore* package (Bolstad, 2019) was used on the CRUKPAP data with the AEGIS expression matrix as the target distribution.

Gene ontology analysis and pathway analysis

All Gene Ontology (GO) enrichment analyses were performed using clusterProfiler v3.14.3 (Yu et al., 2012). GO terms with adjusted (Benjamini-Hochberg) p-values < 0.05 were considered enriched. Pathway metascores were calculated by averaging

vst-normalised gene expression of genes belonging to the selected genesets, after regressing out experimental batch effect.

Concluding remarks

In this thesis, I explored the effects of genetic and environmental factors on cellular phenotypes associated with cancer risk and development.

In Chapter 2, I showed that mutations occurring in simple sequence repeats have a significant effect on cell fitness, which is observable in yeast strains as a mildly negative impact on growth rate. Other mutation accumulation experiments in model organisms show that only few mutations that spontaneously accumulate in a genome have strong advantageous effects on cell fitness (Eyre-Walker and Keightley, 2007), mirroring the tumour landscape, where a multitude of mutations accumulate before and after malignant transformation, but only a few of them act as drivers. These studies also assign mainly deleterious effects to the majority of accumulated mutations Keightley and Lynch (2003).

The majority of mutations observed in cancer genomes are passengers, alterations which are often considered neutral and inconsequential for tumour development. However, passengers have the potential to significantly participate in shaping the tumour landscape, when a large number of their small individual effects are combined (McFarland et al., 2013). The cumulative deleterious effect of passenger mutations might help explain the paradox of MSI tumours showing a better prognosis than non-MSI tumours, despite having an increased genome instability. The main hypothesis proposed in current literature is that this outcome is associated with the large number of neoantigens produced by the high mutational load, and the consequent increase in anti-tumour immune response (Lee et al., 2016; McGrail et al., 2020). However, McFarland et al. (2017) have shown that a high number of copy-number passenger alterations reduces fitness in cancer cell lines and slows down cancer progression in mouse models, even in the absence of enhanced immunity.

Our results thus provide an alternative, or complementary hypothesis, for the favourable outcome associated with MSI tumours. We showed that most mutations in SSR regions have an effect on growth rate, although very small; MSI tumours, however, accumulate a large number of these mutations. We suggest that the load of small deleterious mutational effects likely present in MSI tumours acts as a counter-balance to the large effects of few drivers, effectively slowing down tumour growth and spread.

Nevertheless, more investigation is needed to confirm this hypothesis. Mutation rates

and phenotypic effects of mutations are dependent on many factors, including species, ploidy and environmental conditions (Liu and Zhang, 2019; Martin and Lenormand, 2006; Sharp et al., 2018). Our results were obtained in yeast, in very controlled experimental conditions, with free availability of nutrients and resources and under reduced selective pressure. On the other hand, tumour cells reside in a complex microenvironment, in constant competition for resources and under strong selective pressure. Therefore, the results obtained in this simplified model should be verified by directly measuring the effect of passenger mutations in SSRs in more accurate models of cancer such as tumour cell lines harbouring MMR mutations, mouse models and patient-derived cell-lines from MSI tumours.

In Chapter 3 I showed that valuable information regarding the response to smoke injury and the associated risk of lung cancer can be inferred from gene expression at the level of an accessible airway tissue: nasal epithelium. Knowledge about the field of injury caused by cigarette smoke across airway tissues has been proposed as an aid to cancer diagnosis, for example in patients undergoing bronchoscopy for the evaluation of suspicious lesions. In this context, a molecular biomarker based on bronchial gene expression has shown potential to improve the diagnostic sensitivity of bronchoscopy, a procedure that is often inconclusive and leads to additional invasive and costly procedures (Spira et al., 2007). Similarly, the potential of nasal gene expression to improve classification of cancer patients was explored in a clinic-referred population with suspicious lesions (Perez-Rogers et al., 2017).

The particular composition of our study cohort allowed us to explore another application for the information that can be derived from the field of injury: risk stratification in the general population of smokers and ex smokers. Our results show that there is a range of responses to smoke injury also within a healthy population, with some individuals having responses closer to the clinic-referred patients. Currently, high-risk individuals in the general population are defined by age and history of heavy smoking. Our results suggest that pre-screening with a nasal gene expression biomarker might help improve the selection of asymptomatic individuals at risk, who would benefit from frequent check-ups and LCDT screening. Such an application would require extensive validation of the results presented in this thesis, ideally with an independent cohort of healthy volunteers drawn from the general population, with follow-up information regarding their cancer status.

Even though we found that significant information could be gained from analysis of protein-coding genes, many other aspects of the airway field of injury remain to be explored. Non-coding RNAs also have been shown to play a role in lung cancer, during

tumour progression and in premalignancy (Jiang et al., 2021; Mascaux et al., 2009; Perdomo et al., 2013; Wu et al., 2019). Cigarette smoke also causes epigenetic alterations, which could be used as well as biomarkers for risk stratification (Zong et al., 2019). Another topic for future investigation is the link between the transcriptional alterations observed in the airway, in particular the individual response to smoke injury, to germline variation, and the possible interactions between genotype and response to injury.

References

Cambridge bioresource. https://www.cambridgebioresource.group.cam.ac.uk/. Accessed: 2022-1-18.

Picard toolkit. https://broadinstitute.github.io/picard/, 2019.

S Aebi, B Kurdi-Haidar, R Gordon, B Cenni, H Zheng, D Fink, R D Christen, C R Boland, M Koi, R Fishel, and S B Howell. Loss of DNA mismatch repair in acquired resistance to cisplatin. *Cancer Res.*, 56(13):3087–3090, July 1996.

Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Finding the Cancer-Critical Genes*. Garland Science, 2002.

Roger P Alexander, Gang Fang, Joel Rozowsky, Michael Snyder, and Mark B Gerstein. Annotating non-coding regions of the genome. *Nat. Rev. Genet.*, 11(8):559–571, August 2010.

William G Alexander, Drew T Doering, and Chris Todd Hittinger. High-efficiency genome editing and allele replacement in prototrophic and wild strains of saccharomyces. *Genetics*, 198(3):859–866, November 2014.

Ludmil B Alexandrov, Serena Nik-Zainal, David C Wedge, Peter J Campbell, and Michael R Stratton. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.*, 3(1):246–259, January 2013.

Ludmil B Alexandrov, Jaegil Kim, Nicholas J Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, Kyle R Covington, Dmitry A Gordenin, Erik N Bergstrom, S M Ashiqul Islam, Nuria Lopez-Bigas, Leszek J Klimczak, John R McPherson, Sandro Morganella, Radhakrishnan Sabarinathan, David A Wheeler, Ville Mustonen, PCAWG Mutational Signatures Working Group, Gad Getz, Steven G Rozen, Michael R Stratton, and PCAWG Consortium. The repertoire of mutational signatures in human cancer. *Nature*, 578(7793):94–101, February 2020.

Nasser K Altorki, Geoffrey J Markowitz, Dingcheng Gao, Jeffrey L Port, Ashish Saxena, Brendon Stiles, Timothy McGraw, and Vivek Mittal. The lung microenvironment: an important regulator of tumour growth and metastasis. *Nat. Rev. Cancer*, 19(1):9–31, January 2019.

Mariano J Alvarez, Yao Shen, Federico M Giorgi, Alexander Lachmann, B Belinda Ding, B Hilda Ye, and Andrea Califano. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.*, 48(8):838–847, August 2016.

O Auerbach, A P Stout, E C Hammond, and L Garfinkel. Changes in bronchial epithelium in relation to cigarette smoking and in relation to lung cancer. *N. Engl. J. Med.*, 265:253–267, August 1961.

Juan J Aurrekoetxea, Mario Murcia, Marisa Rebagliato, María José López, Ane Miren Castilla, Loreto Santa-Marina, Mónica Guxens, Ana Fernández-Somoano, Mercedes Espada, Aitana Lertxundi, Adonina Tardón, and Ferran Ballester. Determinants of self-reported smoking and misclassification during pregnancy, and analysis of optimal cut-off points for urinary cotinine: a cross-sectional study. *BMJ Open*, 3(1), January 2013.

Frances R Balkwill, Melania Capasso, and Thorsten Hagemann. The tumor microenvironment at a glance. *J. Cell Sci.*, 125(Pt 23):5591–5596, December 2012.

Jessica L Barnes, Maria Zubair, Kaarthik John, Miriam C Poirier, and Francis L Martin. Carcinogens and DNA damage. *Biochem. Soc. Trans.*, 46(5):1213–1224, October 2018.

Michael Baym, Sergey Kryazhimskiy, Tami D Lieberman, Hattie Chung, Michael M Desai, and Roy Kishony. Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One*, 10(5):e0128036, May 2015.

Jennifer Beane, Paola Sebastiani, Gang Liu, Jerome S Brody, Marc E Lenburg, and Avrum Spira. Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biol.*, 8(9):R201, 2007.

G Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, 27(2):573–580, January 1999.

C Richard Boland and Ajay Goel. Microsatellite instability in colorectal cancer. *Gastroenterology*, 138(6):2073–2087.e3, June 2010.

C Richard Boland and Henry T Lynch. The history of lynch syndrome. *Fam. Cancer*, 12(2):145–157, June 2013.

Ben Bolstad. *preprocessCore: A collection of pre-processing functions*, 2019. URL https://github.com/bmbolstad/preprocessCore. R package version 1.48.0.

Russell Bonneville, Melanie A Krook, Esko A Kautto, Jharna Miya, Michele R Wing, Hui-Zi Chen, Julie W Reeser, Lianbo Yu, and Sameek Roychowdhury. Landscape of microsatellite instability across 39 cancer types, 2017.

F X Bosch, A Lorincz, N Muñoz, C J L M Meijer, and K V Shah. The causal relation between human papillomavirus and cervical cancer. *J. Clin. Pathol.*, 55(4):244–265, April 2002.

Claudio Brancolini and Luca Iuliano. Proteotoxic stress and cell death in cancer cells. *Cancers*, 12(9), August 2020.

James D Brooks. Translational genomics: the challenge of developing cancer biomarkers. *Genome Res.*, 22(2):183–187, February 2012.

R Brown, G L Hirst, W M Gallagher, A J McIlwrath, G P Margison, A G van der Zee, and D A Anthoney. hMLH1 expression and cellular responses of ovarian tumour cells to treatment with cytotoxic anticancer agents. *Oncogene*, 15(1):45–52, July 1997.

Helen Budworth and Cynthia T McMurray. A brief history of triplet repeat diseases. *Methods Mol. Biol.*, 1010:3–17, 2013.

Eileen M Burd. Human papillomavirus and cervical cancer. *Clin. Microbiol. Rev.*, 16 (1):1–17, January 2003.

Adam B Burkholder, Scott A Lujan, Christopher A Lavender, Sara A Grimm, Thomas A Kunkel, and David C Fargo. Muver, a computational framework for accurately calling accumulated mutations. *BMC Genomics*, 19(1):345, May 2018.

Cuixia Cai, Rong Shi, Yuan Gao, Jun Zeng, Min Wei, Handuo Wang, Wenling Zheng, and Wenli Ma. Reduced expression of sushi domain containing 2 is associated with progression of non-small cell lung cancer. *Oncol. Lett.*, 10(6):3619–3624, December 2015.

Kieran R Campbell and Christopher Yau. Uncovering pseudotemporal trajectories with covariates from single cell and bulk expression data. *Nat. Commun.*, 9(1):2442, June 2018.

D Carmona-Gutierrez, T Eisenberg, S Büttner, C Meisinger, G Kroemer, and F Madeo. Apoptosis in yeast: triggers, pathways, subroutines. *Cell Death Differ.*, 17(5): 763–773, May 2010.

Marcelo A Carvalho, Sylvia M Marsillac, Rachel Karchin, Siranoush Manoukian, Scott Grist, Ramona F Swaby, Turan P Urmenyi, Edson Rondinelli, Rosane Silva, Luis Gayol, Lisa Baumbach, Rebecca Sutphen, Jennifer L Pickard-Brzosowicz, Katherine L Nathanson, Andrej Sali, David Goldgar, Fergus J Couch, Paolo Radice, and Alvaro N A Monteiro. Determination of cancer risk associated with germ line BRCA1 missense variants by functional analysis. *Cancer Res.*, 67(4):1494–1501, February 2007.

Francesc Castro-Giner, Peter Ratcliffe, and Ian Tomlinson. The mini-driver model of polygenic cancer evolution. *Nat. Rev. Cancer*, 15(11):680–685, November 2015.

Tiziana Cervelli, Samuele Lodovichi, Francesca Bellè, and Alvaro Galli. Yeast-based assays for the functional characterization of cancer-associated variants of human DNA repair genes. *Microb. Cell Fact.*, 7(7):162–174, May 2020.

Qianmin Chen, Kai Sen Tan, Jing Liu, Hsiao Hui Ong, Suizi Zhou, Hongming Huang, Hailing Chen, Yew Kwang Ong, Mark Thong, Vincent T Chow, Qianhui Qiu, and De-Yun Wang. Host antiviral response suppresses ciliogenesis and motile ciliary functions in the nasal epithelium. *Front Cell Dev Biol*, 8:581340, December 2020.

Yingying Cheng, Xiaolin Wang, Pingzhang Wang, Ting Li, Fengzhan Hu, Qiang Liu, Fan Yang, Jun Wang, Tao Xu, and Wenling Han. SUSD2 is frequently downregulated and functions as a tumor suppressor in RCC and lung cancer. *Tumour Biol.*, 37(7): 9919–9930, July 2016.

Yoon Young Choi, Jung Min Bae, Ji Yeong An, In Gyu Kwon, In Cho, Hyun Beak Shin, Tanaka Eiji, Mohammad Aburahmah, Hyung-Il Kim, Jae-Ho Cheong, Woo Jin Hyung, and Sung Hoon Noh. Is microsatellite instability a prognostic marker in gastric cancer? a systematic review with meta-analysis. *J. Surg. Oncol.*, 110(2): 129–135, August 2014.

Merlise Clyde. *BAS: Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling*, 2018. R package version 1.5.3.

Sarah Connor Gorber, Sean Schofield-Hurwitz, Jill Hardt, Geneviève Levasseur, and Mark Tremblay. The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status. *Nicotine Tob. Res.*, 11(1):12–24, January 2009.

Maria Angelica Cortez, Fatemeh Masrorpour, Cristina Ivan, Jie Zhang, Ahmed I Younes, Yue Lu, Marcos R Estecio, Hampartsoum B Barsoumian, Hari Menon, Mauricio da Silva Caetano, Rishab Ramapriyan, Jonathan E Schoenhals, Xiaohong Wang, Ferdinandos Skoulidis, Mark D Wasley, George Calin, Patrick Hwu, and James W Welsh. Bone morphogenetic protein 7 promotes resistance to immunotherapy. *Nat. Commun.*, 11(1):4840, September 2020.

Lisa M Coussens and Zena Werb. Inflammation and cancer. *Nature*, 420(6917): 860–867, 2002.

Robert S Coyne, Heather B McDonald, Keith Edgemon, and Lawrence C Brody. Functional characterization of BRCA1 sequence variants using a yeast small colony phenotype assay. *Cancer Biol. Ther.*, 3(5):453–457, May 2004.

Rebecca J Critchley-Thorne, Diana L Simons, Ning Yan, Andrea K Miyahira, Frederick M Dirbas, Denise L Johnson, Susan M Swetter, Robert W Carlson, George A Fisher, Albert Koong, Susan Holmes, and Peter P Lee. Impaired interferon signaling is a common immune defect in human cancer. *Proc. Natl. Acad. Sci. U. S. A.*, 106 (22):9010–9015, June 2009.

Nicola J Curtin. DNA repair dysregulation from cancer driver to therapeutic target. *Nat. Rev. Cancer*, 12(12):801–817, December 2012.

de Biase, Massip, Tzu-Ting Wei, Federico M Giorgi, Rory Stark, Amanda Stone, Amy Gladwell, Martin O'Reilly, Ines de Santiago, Kerstin Meyer, Florian Markowetz, Bruce A J Ponder, Robert C Rintoul, and Roland F Schwarz. Smoking-dependent expression alterations in nasal epithelium reveal immune impairment linked to germline variation and lung cancer risk. November 2021.

J de Boer and J H Hoeijmakers. Nucleotide excision repair and human syndromes. *Carcinogenesis*, 21(3):453–460, March 2000.

Harry J de Koning, Carlijn M van der Aalst, Pim A de Jong, Ernst T Scholten, Kristiaan Nackaerts, Marjolein A Heuvelmans, Jan-Willem J Lammers, Carla Weenink, Uraujh Yousaf-Khan, Nanda Horeweg, Susan van 't Westeinde, Mathias Prokop, Willem P Mali, Firdaus A A Mohamed Hoesein, Peter M A van Ooijen, Joachim G J V Aerts, Michael A den Bakker, Erik Thunnissen, Johny Verschakelen, Rozemarijn Vliegenthart, Joan E Walter, Kevin Ten Haaf, Harry J M Groen, and Matthijs Oudkerk. Reduced Lung-Cancer mortality with volume CT screening in a randomized trial. *N. Engl. J. Med.*, 382(6):503–513, February 2020.

Giovanni De Pergola and Franco Silvestris. Obesity as a major risk factor for cancer. *J. Obes.*, 2013:291546, August 2013.

Andrea Decensi, Matteo Puntoni, Giancarlo Pruneri, Aliana Guerrieri-Gonzaga, Matteo Lazzeroni, Davide Serrano, Debora Macis, Harriet Johansson, Oriana Pala, Alberto Luini, Paolo Veronesi, Viviana Galimberti, Maria Cristina Dotti, Giuseppe Viale, and Bernardo Bonanni. Lapatinib activity in premalignant lesions and HER-2-positive cancer of the breast in a randomized, placebo-controlled presurgical trial. *Cancer Prev. Res.*, 4(8):1181–1189, August 2011.

Nicole Denzer, Thomas Vogt, and Jörg Reichrath. Vitamin D receptor (VDR) polymorphisms and skin cancer, 2011.

Eleftherios P Diamandis. The failure of protein cancer biomarkers to reach the clinic: why, and what can be done to address the problem? *BMC Med.*, 10:87, August 2012.

R Diaz-Ruiz, S Uribe-Carvajal, A Devin, and M Rigoulet. Tumor cell energy metabolism and its common features with yeast metabolism. *Biochim. Biophys. Acta*, 1796(2):252–265, December 2009.

Kara Dolinski and David Botstein. Orthology and functional conservation in eukaryotes. *Annu. Rev. Genet.*, 41:465–507, 2007.

B J Druker, M Talpaz, D J Resta, B Peng, E Buchdunger, J M Ford, N B Lydon, H Kantarjian, R Capdeville, S Ohno-Jones, and C L Sawyers. Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. *N. Engl. J. Med.*, 344(14):1031–1037, April 2001.

Eric J Duell, Daniel P Casella, Robert D Burk, Karl T Kelsey, and Elizabeth A Holly. Inflammation, genetic polymorphisms in proinflammatory genes TNF-A, RANTES, and CCR5, and risk of pancreatic adenocarcinoma, 2006.

Keith D Eaton, Perrin E Romine, Gary E Goodman, Mark D Thornquist, Matt J Barnett, and Effie W Petersdorf. Inflammatory gene polymorphisms in lung cancer susceptibility. *J. Thorac. Oncol.*, 13(5):649–659, May 2018.

Amir Eden, François Gaudet, Alpana Waghmare, and Rudolf Jaenisch. Chromosomal instability and tumors promoted by DNA hypomethylation. *Science*, 300(5618):455, April 2003.

J A Eisen. Mechanistic basis for microsatellite instability. *Microsatellites : Evolution and Applications*, pages 34–48, 1999.

Hans Ellegren. Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.*, 5(6):435–445, June 2004.

Kerryn Elliott and Erik Larsson. Non-coding driver mutations in human cancer. *Nat. Rev. Cancer*, 21(8):500–509, August 2021.

Lawson Eng, Abul Kalam Azad, Steven Habbous, Vincent Pang, Wei Xu, Anke Maitland-van der Zee, Sevtap Savas, Helen J Mackay, Eitan Amir, and Geoffrey Liu. Vascular endothelial growth factor pathway polymorphisms as prognostic and pharmacogenetic factors in cancer: A systematic review and meta-analysis, 2012.

M Esteller, J M Silva, G Dominguez, F Bonilla, X Matias-Guiu, E Lerma, E Bussaglia, J Prat, I C Harkes, E A Repasky, E Gabrielson, M Schutte, S B Baylin, and J G Herman. Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *J. Natl. Cancer Inst.*, 92(7):564–569, April 2000.

Adam Eyre-Walker and Peter D Keightley. The distribution of fitness effects of new mutations. *Nat. Rev. Genet.*, 8(8):610–618, August 2007.

FastQC. FastQC, a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, 2010.

Yifei Feng, Dongjian Ji, Yuanjian Huang, Bing Ji, Yue Zhang, Jie Li, Wen Peng, Chuan Zhang, Dongsheng Zhang, Yueming Sun, and Ziwei Xu. TGM3 functions as a tumor suppressor by repressing epithelial-to-mesenchymal transition and the PI3K/AKT signaling pathway in colorectal cancer. *Oncol. Rep.*, 43(3):864–876, March 2020.

D Field and C Wills. Long, polymorphic microsatellites in simple organisms. *Proc. Biol. Sci.*, 263(1367):209–215, February 1996.

Bernd Frank, Michael Hoffmeister, Norman Klopp, Thomas Illig, Jenny Chang-Claude, and Hermann Brenner. Polymorphisms in inflammatory pathway genes and their association with colorectal cancer risk, 2010.

W A Franklin, A F Gazdar, J Haney, I I Wistuba, F G La Rosa, T Kennedy, D M Ritchey, and Y E Miller. Widely dispersed p53 mutation in respiratory epithelium. a novel mechanism for field carcinogenesis. *J. Clin. Invest.*, 100(8):2133–2137, October 1997.

Alison E Gammie, Naz Erdeniz, Julia Beaver, Barbara Devlin, Afshan Nanji, and Mark D Rose. Functional characterization of pathogenic human MSH2 missense mutations in saccharomyces cerevisiae. *Genetics*, 177(2):707–721, October 2007.

Caihua Gao, Xiaodong Ren, Annaliese S Mason, Jiana Li, Wei Wang, Meili Xiao, and Donghui Fu. Revisiting an important component of plant genomes: microsatellites. *Funct. Plant Biol.*, 40(7):645–661, July 2013.

Yanfeng Gao, Shuang Zhou, Yi Xu, Sen Sheng, Steven Y Qian, and Xiongwei Huo. Nitric oxide synthase inhibitors 1400W and L-NIO inhibit angiogenesis pathway of colorectal cancer. *Nitric Oxide*, 83:33–39, February 2019.

E Garrison and G Marth. Haplotype-based variant detection from short-read sequencing. arxiv 1207.3907. *Preprint posted online July*, 12, 2012.

Moritz Gerstung, Clemency Jolly, Ignaty Leshchiner, Stefan C Dentro, Santiago Gonzalez, Daniel Rosebrock, Thomas J Mitchell, Yulia Rubanova, Pavana Anur, Kaixian Yu, Maxime Tarabichi, Amit Deshwar, Jeff Wintersinger, Kortine Kleinheinz, Ignacio Vázquez-García, Kerstin Haase, Lara Jerman, Subhajit Sengupta, Geoff Macintyre, Salem Malikic, Nilgun Donmez, Dimitri G Livitz, Marek Cmero, Jonas Demeulemeester, Steven Schumacher, Yu Fan, Xiaotong Yao, Juhee Lee, Matthias Schlesner, Paul C Boutros, David D Bowtell, Hongtu Zhu, Gad Getz, Marcin Imielinski, Rameen Beroukhim, S Cenk Sahinalp, Yuan Ji, Martin Peifer, Florian Markowetz, Ville Mustonen, Ke Yuan, Wenyi Wang, Quaid D Morris,

PCAWG Evolution & Heterogeneity Working Group, Paul T Spellman, David C Wedge, Peter Van Loo, and PCAWG Consortium. The evolutionary history of 2,658 cancers. *Nature*, 578(7793):122–128, February 2020.

Hugo Gonzalez, Catharina Hagerling, and Zena Werb. Roles of the immune system in cancer: from tumor initiation to metastatic progression. *Genes Dev.*, 32(19-20): 1267–1284, October 2018.

Elias Gounaris, Michael J Heiferman, Jeffrey R Heiferman, Manisha Shrivastav, Dominic Vitello, Nichole R Blatner, Lawrence M Knab, Joseph D Phillips, Eric C Cheon, Paul J Grippo, Khashayarsha Khazaie, Hidayatullah G Munshi, and David J Bentrem. Zileuton, 5-lipoxygenase inhibitor, acts as a chemopreventive agent in intestinal polyposis, by modulating polyp and systemic inflammation. *PLoS One*, 10(3):e0121402, March 2015.

Jan Grau, Ivo Grosse, and Jens Keilwagen. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*, 31(15):2595–2597, August 2015.

Sergei I Grivennikov, Florian R Greten, and Michael Karin. Immunity, inflammation, and cancer. *Cell*, 140(6):883–899, March 2010.

Cecilia Guastadisegni, Mauro Colafranceschi, Laura Ottini, and Eugenia Dogliotti. Microsatellite instability as a marker of prognosis and response to therapy: a meta-analysis of colorectal cancer survival data. *Eur. J. Cancer*, 46(15):2788–2798, October 2010.

Adam M Gustafson, Raffaella Soldi, Christina Anderlind, Mary Beth Scholand, Jun Qian, Xiaohui Zhang, Kendal Cooper, Darren Walker, Annette McWilliams, Gang Liu, Eva Szabo, Jerome Brody, Pierre P Massion, Marc E Lenburg, Stephen Lam, Andrea H Bild, and Avrum Spira. Airway PI3K pathway activation is an early and reversible event in lung cancer development. *Sci. Transl. Med.*, 2(26):26ra25, April 2010.

D Haase, M Meister, T Muley, J Hess, S Teurich, P Schnabel, B Hartenstein, and P Angel. FRMD3, a novel putative tumour suppressor in NSCLC. *Oncogene*, 26 (30):4464–4468, June 2007.

David W Hall, Rod Mahmoudizad, Andrew W Hurd, and Sarah B Joseph. Spontaneous mutations in diploid saccharomyces cerevisiae: another thousand cell generations. *Genet. Res.*, 90(3):229–241, June 2008.

Daniel L Halligan and Peter D Keightley. Spontaneous mutation accumulation studies in evolutionary genetics. *Annu. Rev. Ecol. Evol. Syst.*, 40(1):151–172, December 2009.

Akil Hamza, Erik Tammpere, Megan Kofoed, Christelle Keong, Jennifer Chiang, Guri Giaever, Corey Nislow, and Philip Hieter. Complementation of yeast genes with human genes as an experimental platform for functional testing of human genetic variants. *Genetics*, 201(3):1263–1274, November 2015.

Chan H Han, Yu-Jing Huang, Karen H Lu, Zhensheng Liu, Gordon B Mills, Qingyi Wei, and Li-E Wang. Polymorphisms in the SULF1 gene are associated with early age of onset and survival of ovarian cancer. *J. Exp. Clin. Cancer Res.*, 30:5, January 2011.

D Hanahan and R A Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, January 2000.

Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, March 2011.

Jill M Harrington and Richard D Kolodner. Saccharomyces cerevisiae Msh2-Msh3 acts in repair of base-base mispairs. *Mol. Cell. Biol.*, 27(18):6546–6554, September 2007.

Leland H Hartwell. Yeast and cancer. *Biosci. Rep.*, 22(3-4):373–394, 2002.

Joanna E Haye and Alison E Gammie. The eukaryotic mismatch recognition complexes track with the replisome during DNA synthesis. *PLoS Genet.*, 11(12):e1005719, December 2015.

Christopher D Heinen. Mismatch repair defects and lynch syndrome: The role of the basic scientist in the battle against cancer. *DNA Repair*, 38:127–134, February 2016.

N Lynn Henry and Daniel F Hayes. Cancer biomarkers. *Mol. Oncol.*, 6(2):140–146, April 2012.

J G Herman, F Latif, Y Weng, M I Lerman, B Zbar, S Liu, D Samid, D S Duan, J R Gnarra, and W M Linehan. Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. *Proc. Natl. Acad. Sci. U. S. A.*, 91(21):9700–9704, October 1994.

Andrzej Hnatyszyn, Szymon Hryhorowicz, Marta Kaczmarek-Ryś, Emilia Lis, Ryszard Słomski, Rodney J Scott, and Andrzej Pławski. Colorectal carcinoma in the course of inflammatory bowel diseases. *Hered. Cancer Clin. Pract.*, 17:18, July 2019.

Gabriel E Hoffman and Eric E Schadt. variancepartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics*, 17(1):483, November 2016.

Louise R Howe, Kotha Subbaramaiah, Clifford A Hudis, and Andrew J Dannenberg. Molecular pathways: adipose inflammation as a mediator of obesity-associated cancer. *Clin. Cancer Res.*, 19(22):6074–6083, November 2013.

Jin-Wu Hu, Zhang-Fu Yang, Jia Li, Bo Hu, Chu-Bin Luo, Kai Zhu, Zhi Dai, Jia-Bin Cai, Hao Zhan, Zhi-Qiang Hu, Jie Hu, Ya Cao, Shuang-Jian Qiu, Jian Zhou, Jia Fan, and Xiao-Wu Huang. TGM3 promotes epithelial–mesenchymal transition and hepatocellular carcinogenesis and predicts poor prognosis for patients after curative resection. *Dig. Liver Dis.*, 52(6):668–676, June 2020.

Hunt. Protein synthesis, proteolysis, and cell cycle transitions: Nobel lecture. *See http://www. nobelprize. org/nobel_prizes/medicine*.

ICGC-ARGO. ICGC ARGO. http://platform.icgc-argo.org, 2019. Accessed: 2022-4-15.

ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82–93, February 2020.

Thomas F Imperiale, David F Ransohoff, Steven H Itzkowitz, Theodore R Levin, Philip Lavin, Graham P Lidgard, David A Ahlquist, and Barry M Berger. Multitarget stool DNA testing for colorectal-cancer screening. *N. Engl. J. Med.*, 370(14):1287–1297, April 2014.

J P Jakupciak and R D Wells. Genetic instabilities in (CTG· CAG) repeats occur by recombination. *J. Biol. Chem.*, 1999.

Jun Jiang, Yuan Lu, Fang Zhang, Jie Huang, Xin-Ling Ren, and Rui Zhang. The emerging roles of long noncoding RNAs as hallmarks of lung cancer. *Front. Oncol.*, 11:761582, October 2021.

Tao Jiang, Tao Shi, Henghui Zhang, Jie Hu, Yuanlin Song, Jia Wei, Shengxiang Ren, and Caicun Zhou. Tumor neoantigens: from basic research to clinical applications. *J. Hematol. Oncol.*, 12(1):93, September 2019.

Josef Jiricny. Postreplicative mismatch repair. *Cold Spring Harb. Perspect. Biol.*, 5(4): a012633, April 2013.

James O Jones, William M Moody, and Jacqueline D Shields. Microenvironmental modulation of the developing tumour: an immune-stromal dialogue. *Mol. Oncol.*, 15(10):2600–2633, October 2021.

Dragica Jorgovanovic, Mengjia Song, Liping Wang, and Yi Zhang. Roles of IFN-γ in tumor progression and regression: a review. *Biomark Res*, 8:49, September 2020.

Sarah B Joseph and David W Hall. Spontaneous mutations in diploid saccharomyces cerevisiae: more beneficial than expected. *Genetics*, 168(4):1817–1825, December 2004.

Aashiq H Kachroo, Jon M Laurent, Christopher M Yellman, Austin G Meyer, Claus O Wilke, and Edward M Marcotte. Evolution. systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science*, 348(6237):921–925, May 2015.

Haydar Karaoglu, Crystal Man Ying Lee, and Wieland Meyer. Survey of simple sequence repeats in completed fungal genomes. *Mol. Biol. Evol.*, 22(3):639–649, March 2005.

Y Kashi, D King, and M Soller. Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet.*, 13(2):74–78, February 1997.

Yechezkel Kashi and David G King. Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.*, 22(5):253–259, May 2006.

Vaishali Katju and Ulfar Bergthorsson. Old trade, new tricks: Insights into the spontaneous mutation process from the partnering of classical mutation accumulation experiments with High-Throughput genomic approaches. *Genome Biol. Evol.*, 11 (1):136–165, January 2019.

Andrew M Kaz, William M Grady, Matthew D Stachler, and Adam J Bass. Genetic and epigenetic alterations in barrett's esophagus and esophageal adenocarcinoma, 2015.

Peter D Keightley and Michael Lynch. Toward a realistic model of mutations affecting fitness. *Evolution*, 57(3):683–5; discussion 686–9, March 2003.

Hiroyuki Kida, Yuki Takano, Ken Yamamoto, Masaki Mori, Katsuhiko Yanaga, Jun-Ichi Tanaka, Shin-Ei Kudo, and Koshi Mimori. A single nucleotide polymorphism in fibronectin 1 determines tumor shape in colorectal cancer. *Oncol. Rep.*, 32(2): 548–552, August 2014.

Changshin Kim, Jinmo Yang, Su-Hyun Jeong, Hayoung Kim, Geun-Hee Park, Hwa Beom Shin, Myungja Ro, Kyoung-Yeon Kim, Youngjoon Park, Keun Pil Kim, and Kyubum Kwack. Yeast-based assays for characterization of the functional effects of single nucleotide polymorphisms in human DNA repair genes. *PLoS One*, 13(3):e0193823, March 2018.

Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, 14(4):R36, April 2013.

Felisha L Kitchen and Christina M Cox. Papanicolaou smear. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), October 2021.

A G Knudson, Jr. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci. U. S. A.*, 68(4):820–823, April 1971.

Ewa Kolosionek, Rajkumar Savai, Hossein Ardeschir Ghofrani, Norbert Weissmann, Andreas Guenther, Friedrich Grimminger, Werner Seeger, Gamal Andre Banat, Ralph Theo Schermuly, and Soni Savai Pullamsetti. Expression and activity of phosphodiesterase isoforms during epithelial mesenchymal transition: the role of phosphodiesterase 4. *Mol. Biol. Cell*, 20(22):4751–4765, November 2009.

Anke Konrad, Stephane Flibotte, Jon Taylor, Robert H Waterston, Donald G Moerman, Ulfar Bergthorsson, and Vaishali Katju. Mutational and transcriptional landscape of spontaneous gene duplications and deletions in caenorhabditis elegans. *Proc. Natl. Acad. Sci. U. S. A.*, 115(28):7386–7391, July 2018.

B P Kopnin. Targets of oncogenes and tumor suppressors: key for understanding basic mechanisms of carcinogenesis. *Biochemistry*, 65(1):2–27, January 2000.

Erika Korobeinikova, Rasa Ugenskiene, Ruta Insodaite, Viktoras Rudzianskas, Evelina Jaselske, Lina Poskiene, and Elona Juozaityte. Association of angiogenesis and inflammation-related gene functional polymorphisms with early-stage breast cancer prognosis. *Oncol. Lett.*, 19(6):3687–3700, June 2020.

Piotr Kozlowski, Mateusz de Mezer, and Wlodzimierz J Krzyzosiak. Trinucleotide repeats in human genome and exome. *Nucleic Acids Res.*, 38(12):4027–4039, July 2010.

B Madhu Krishna, Samir Jana, Aditya K Panda, David Horne, Sanjay Awasthi, Ravi Salgia, and Sharad S Singhal. Association of TGF-β1 polymorphisms with breast cancer risk: A Meta-Analysis of Case–Control studies. *Cancers*, 12(2):471, February 2020.

Nathan A Krump and Jianxin You. Molecular mechanisms of viral oncogenesis in humans. *Nat. Rev. Microbiol.*, 16(11):684–698, November 2018.

Kostyantyn Krysan, Jay M Lee, Mariam Dohadwala, Brian K Gardner, Karen L Reckamp, Edward Garon, Maie St John, Sherven Sharma, and Steven M Dubinett. Inflammation, epithelial to mesenchymal transition, and epidermal growth factor receptor tyrosine kinase inhibitor resistance. *J. Thorac. Oncol.*, 3(2):107–110, February 2008.

Sushant Kumar, Jonathan Warrell, Shantao Li, Patrick D McGillivray, William Meyerson, Leonidas Salichos, Arif Harmanci, Alexander Martinez-Fundichely, Calvin W Y Chan, Morten Muhlig Nielsen, Lucas Lochovsky, Yan Zhang, Xiaotong Li, Shaoke Lou, Jakob Skou Pedersen, Carl Herrmann, Gad Getz, Ekta Khurana, and Mark B Gerstein. Passenger mutations in more than 2,500 cancer genomes: Overall molecular functional impact and consequences. *Cell*, 180(5):915–927.e16, March 2020.

Erika M Kwon, Claudia A Salinas, Suzanne Kolb, Rong Fu, Ziding Feng, Janet L Stanford, and Elaine A Ostrander. Genetic polymorphisms in inflammation pathway genes and prostate cancer risk. *Cancer Epidemiol. Biomarkers Prev.*, 20(5):923–933, May 2011.

Alexander Lachmann, Federico M Giorgi, Gonzalo Lopez, and Andrea Califano. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics*, 32(14):2233–2235, July 2016.

Jin-Ping Lai, Dalbir S Sandhu, Abdirashid M Shire, and Lewis R Roberts. The tumor suppressor function of human sulfatase 1 (SULF1) in carcinogenesis. *J. Gastrointest. Cancer*, 39(1-4):149–158, 2008.

Peter-Laszlo Lakatos and Laszlo Lakatos. Risk for colorectal cancer in ulcerative colitis: changes, causes and management strategies. *World J. Gastroenterol.*, 14(25): 3937–3947, July 2008.

Gregory I Lang, Lance Parsons, and Alison E Gammie. Mutation rates, spectra, and Genome-Wide distribution of spontaneous mutations in mismatch repair deficient yeast, 2013.

Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9(4):357–359, March 2012.

Jon M Laurent, Jonathan H Young, Aashiq H Kachroo, and Edward M Marcotte. Efforts to make and apply humanized yeast. *Brief. Funct. Genomics*, 15(2):155–163, March 2016.

J Lee, V Taneja, and R Vassallo. Cigarette smoking and inflammation: cellular and molecular mechanisms. *J. Dent. Res.*, 91(2):142–149, February 2012.

Valerie Lee, Adrian Murphy, Dung T Le, and Luis A Diaz, Jr. Mismatch repair deficiency and response to immune checkpoint blockade. *Oncologist*, 21(10):1200–1211, October 2016.

R E Leube and T J Rustad. Squamous cell metaplasia in the human lung: molecular characteristics of epithelial stratification. *Virchows Arch. B Cell Pathol. Incl. Mol. Pathol.*, 61(4):227–253, 1991.

Sasha F Levy, Naomi Ziv, and Mark L Siegal. Bet hedging in yeast by heterogeneous, age-correlated expression of a stress protectant. *PLoS Biol.*, 10(5):e1001325, May 2012.

Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. March 2013.

Kai Li, Haiqing Luo, Lianfang Huang, Hui Luo, and Xiao Zhu. Microsatellite instability: a review of what the oncologist should know. *Cancer Cell Int.*, 20: 16, January 2020.

You-Chun Li, Abraham B Korol, Tzion Fahima, and Eviatar Nevo. Microsatellites within genes: structure, function, and evolution. *Mol. Biol. Evol.*, 21(6):991–1007, June 2004.

Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7): 923–930, April 2014.

Maria V Liberti and Jason W Locasale. The warburg effect: How does it benefit cancer cells? *Trends Biochem. Sci.*, 41(3):211–218, March 2016.

Dekang Liu, Guido Keijzers, and Lene Juel Rasmussen. DNA mismatch repair and its many roles in eukaryotic cells. *Mutat. Res. - Rev. Mut. Res.*, 773:174–187, July 2017.

Haoxuan Liu and Jianzhi Zhang. Yeast spontaneous mutation rate and spectrum vary with environment. *Curr. Biol.*, 29(10):1584–1591.e3, May 2019.

L Liu, K Dybvig, V S Panangala, and others. GAA trinucleotide repeat region regulates M9/pMGA gene expression in mycoplasma gallisepticum. *Infection*, 2000.

Zhengchang Liu and Ronald A Butow. Mitochondrial retrograde signaling. *Annu. Rev. Genet.*, 40:159–185, 2006.

L A Loeb. A mutator phenotype in cancer. *Cancer Res.*, 61(8):3230–3239, April 2001.

Stacy Loeb, Marc A Bjurlin, Joseph Nicholson, Teuvo L Tammela, David F Penson, H Ballentine Carter, Peter Carroll, and Ruth Etzioni. Overdiagnosis and overtreatment of prostate cancer. *Eur. Urol.*, 65(6):1046–1055, June 2014.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, 15(12):550, 2014.

N F Lue, A R Buchman, and R D Kornberg. Activation of yeast RNA polymerase II transcription by a thymidine-rich upstream element in vitro. *Proc. Natl. Acad. Sci. U. S. A.*, 86(2):486–490, January 1989.

Scott A Lujan, Anders R Clausen, Alan B Clark, Heather K MacAlpine, David M MacAlpine, Ewa P Malc, Piotr A Mieczkowski, Adam B Burkholder, David C Fargo, Dmitry A Gordenin, and Thomas A Kunkel. Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. *Genome Res.*, 24(11):1751–1764, November 2014.

Scott A Lujan, Alan B Clark, and Thomas A Kunkel. Differences in genome-wide repeat sequence instability conferred by proofreading and mismatch repair defects. *Nucleic Acids Res.*, 43(8):4067–4074, April 2015.

Scott Alexander Lujan and Thomas A Kunkel. Stability across the whole nuclear genome in the presence and absence of DNA mismatch repair. *Cells*, 10(5), May 2021.

H T Lynch, T C Smyrk, P Watson, S J Lanspa, J F Lynch, P M Lynch, R J Cavalieri, and C R Boland. Genetics, natural history, tumor spectrum, and pathology of hereditary nonpolyposis colorectal cancer: an updated review. *Gastroenterology*, 104(5):1535–1549, May 1993.

Michael Lynch, Way Sung, Krystalynne Morris, Nicole Coffey, Christian R Landry, Erik B Dopman, W Joseph Dickinson, Kazufusa Okamoto, Shilpa Kulkarni, Daniel L Hartl, and W Kelley Thomas. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci. U. S. A.*, 105(27):9272–9277, July 2008.

Michael Lynch, Matthew S Ackerman, Jean-Francois Gout, Hongan Long, Way Sung, W Kelley Thomas, and Patricia L Foster. Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.*, 17(11):704–714, October 2016.

Xin Ma, Maria V Rogacheva, K T Nishant, Sarah Zanders, Carlos D Bustamante, and Eric Alani. Mutation hot spots in yeast caused by long-range clustering of homopolymeric sequences. *Cell Rep.*, 1(1):36–42, January 2012.

Zhikun Ma, Amanda B Parris, Zhengzheng Xiao, Erin W Howard, Stanley D Kosanke, Xiaoshan Feng, and Xiaohe Yang. Short-term early exposure to lapatinib confers lifelong protection from mammary tumor development in MMTV-erbB-2 transgenic mice. *J. Exp. Clin. Cancer Res.*, 36(1):6, January 2017.

John B G Mackey, Seth B Coffelt, and Leo M Carlin. Neutrophil maturity in cancer. *Front. Immunol.*, 10:1912, August 2019.

Guillaume Martin and Thomas Lenormand. The fitness effect of mutations across environments: a survey in light of fitness landscape models. *Evolution*, 60(12): 2413–2427, December 2006.

C Mascaux, J F Laes, G Anthoine, A Haller, V Ninane, A Burny, and J P Sculier. Evolution of microRNA expression during human bronchial squamous carcinogenesis. *Eur. Respir. J.*, 33(2):352–359, February 2009.

Christopher D McFarland, Kirill S Korolev, Gregory V Kryukov, Shamil R Sunyaev, and Leonid A Mirny. Impact of deleterious passenger mutations on cancer progression. *Proc. Natl. Acad. Sci. U. S. A.*, 110(8):2910–2915, February 2013.

Christopher D McFarland, Julia A Yaglom, Jonathan W Wojtkowiak, Jacob G Scott, David L Morse, Michael Y Sherman, and Leonid A Mirny. The damaging effect of passenger mutations on cancer progression. *Cancer Res.*, 77(18):4763–4772, September 2017.

Daniel J McGrail, Jeannine Garnett, Jun Yin, Hui Dai, David J H Shih, Truong Nguyen Anh Lam, Yang Li, Chaoyang Sun, Yongsheng Li, Rosemarie Schmandt, Ji Yuan Wu, Limei Hu, Yulong Liang, Guang Peng, Eric Jonasch, David Menter, Melinda S Yates, Scott Kopetz, Karen H Lu, Russell Broaddus, Gordon B Mills, Nidhi Sahni, and Shiaw-Yih Lin. Proteome instability is a therapeutic vulnerability in mismatch Repair-Deficient cancer. *Cancer Cell*, 37(3):371–386.e12, March 2020.

Neha Merchant, Ganji Purnachandra Nagaraju, Balney Rajitha, Saipriya Lammata, Kishore Kumar Jella, Zachary S Buchwald, Sajani S Lakka, and Arif N Ali. Matrix metalloproteinases: their functional role in lung cancer. *Carcinogenesis*, 38(8): 766–780, August 2017.

A Merlo, J G Herman, L Mao, D J Lee, E Gabrielson, P C Burger, S B Baylin, and D Sidransky. 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. *Nat. Med.*, 1(7): 686–692, July 1995.

D Metzgar, J Bytof, and C Wills. Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Res.*, 10(1):72–80, January 2000.

Yuka Matsuoka Miyazu, Teruomi Miyazawa, Keiko Hiyama, Noriaki Kurimoto, Yasuo Iwamoto, Hiroo Matsuura, Koji Kanoh, Nobuoki Kohno, Masahiko Nishiyama, and Eiso Hiyama. Telomerase expression in noncancerous bronchial epithelia is a possible marker of early development of lung cancer. *Cancer Res.*, 65(21): 9623–9627, November 2005.

Alvaro N Monteiro, Peter Bouwman, Arne N Kousholt, Diana M Eccles, Gael A Millot, Jean-Yves Masson, Marjanka K Schmidt, Shyam K Sharan, Ralph Scully, Lisa Wiesmüller, Fergus Couch, and Maaike P G Vreeswijk. Variants of uncertain clinical significance in hereditary breast and ovarian cancer genes: best practices in functional analysis for clinical annotation. *J. Med. Genet.*, 57(8):509–518, August 2020.

T Mukai. THE GENETIC STRUCTURE OF NATURAL POPULATIONS OF DROSOPHILA MELANOGASTER. i. SPONTANEOUS MUTATION RATE OF POLYGENES CONTROLLING VIABILITY. *Genetics*, 50:1–19, July 1964.

El Mustafa, Sat Parmar, and Prav Praveen. Premalignant lesions and conditions of the oral cavity. In Krishnamurthy Bonanthaya, Elavenil Panneerselvam, Suvy Manuel, Vinay V Kumar, and Anshul Rai, editors, *Oral and Maxillofacial Surgery for the Clinician*, pages 1845–1852. Springer Singapore, Singapore, 2021.

National Lung Screening Trial Research Team, Denise R Aberle, Amanda M Adams, Christine D Berg, William C Black, Jonathan D Clapp, Richard M Fagerstrom,

Ilana F Gareen, Constantine Gatsonis, Pamela M Marcus, and Jorean D Sicks. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.*, 365(5):395–409, August 2011.

M A Nelson, J Wymer, and N Clements, Jr. Detection of k-ras gene mutations in non-neoplastic lung tissue and lung cancers. *Cancer Lett.*, 103(1):115–121, May 1996.

Atsuya Nishiyama and Makoto Nakanishi. Navigating the DNA methylation landscape of cancer. *Trends Genet.*, 37(11):1012–1027, November 2021.

Beifang Niu, Kai Ye, Qunyuan Zhang, Charles Lu, Mingchao Xie, Michael D McLellan, Michael C Wendl, and Li Ding. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics*, 30(7):1015–1016, April 2014.

Paul Nurse. Cyclin dependent kinases and cell cycle control (nobel lecture). *Chembiochem*, 3(7):596–603, July 2002.

O Ohnishi. Spontaneous and ethyl methanesulfonate-induced mutations controlling viability in drosophila melanogaster. i. recessive lethal mutations. *Genetics*, 87(3): 519–527, November 1977.

Olga Okladnova, Yana V Syagailo, Michael Tranitz, Gerald Stöber, Peter Riederer, Rainald Mössner, and Klaus-Peter Lesch. A Promoter-Associated polymorphic repeat ModulatesPAX-6Expression in human brain. *Biochem. Biophys. Res. Commun.*, 248(2):402–405, July 1998.

Adam Pennycuick, Vitor H Teixeira, Khalid AbdulJabbar, Shan E Ahmed Raza, Tom Lund, Ayse U Akarca, Rachel Rosenthal, Lukas Kalinke, Deepak P Chandrasekharan, Christodoulos P Pipinikas, Henry Lee-Six, Robert E Hynds, Kate H C Gowers, Jake Y Henry, Fraser R Millar, Yeman B Hagos, Celine Denais, Mary Falzon, David A Moore, Sophia Antoniou, Pascal F Durrenberger, Andrew J Furness, Bernadette Carroll, Claire Marceaux, Marie-Liesse Asselin-Labat, William Larson, Courtney Betts, Lisa M Coussens, Ricky M Thakrar, Jeremy George, Charles Swanton, Christina Thirlwell, Peter J Campbell, Teresa Marafioti, Yinyin Yuan, Sergio A Quezada, Nicholas McGranahan, and Sam M Janes. Immune surveillance in clinical regression of preinvasive squamous cell lung cancer. *Cancer Discov.*, 10 (10):1489–1499, October 2020.

Catalina Perdomo, Joshua D Campbell, Joseph Gerrein, Carmen S Tellez, Carly B Garrison, Tonya C Walser, Eduard Drizik, Huiqing Si, Adam C Gower, Jessica Vick, Christina Anderlind, George R Jackson, Courtney Mankus, Frank Schembri, Carl O'Hara, Brigitte N Gomperts, Steven M Dubinett, Patrick Hayden, Steven A Belinsky, Marc E Lenburg, and Avrum Spira. MicroRNA 4423 is a primate-specific regulator of airway epithelial cell differentiation and lung carcinogenesis. *Proc. Natl. Acad. Sci. U. S. A.*, 110(47):18946–18951, November 2013.

Joseph F Perez-Rogers, Joseph Gerrein, Christina Anderlind, Gang Liu, Sherry Zhang, Yuriy Alekseyev, Kate Porta Smith, Duncan Whitney, W Evan Johnson, David A Elashoff, Steven M Dubinett, Jerome Brody, Avrum Spira, Marc E Lenburg, and for the AEGIS Study Team. Shared gene expression alterations in nasal and bronchial epithelium for lung cancer detection. *J. Natl. Cancer Inst.*, 109(7), February 2017.

Luciana Santos Pessoa, Manoela Heringer, and Valéria Pereira Ferrer. ctDNA as a cancer biomarker: A broad overview. *Crit. Rev. Oncol. Hematol.*, 155:103109, November 2020.

E J Peters, R Morice, S E Benner, S Lippman, J Lukeman, J S Lee, J Y Ro, and W K Hong. Squamous metaplasia of the bronchial mucosa and its relationship to smoking. *Chest*, 103(5):1429–1432, May 1993.

J Peto. That lung cancer incidence falls in ex-smokers: misconceptions 2. *Br. J. Cancer*, 104(3):389, February 2011.

Yevgeniy Plavskin, Shuang Li, Hyun Jung, Federica M O Sartori, Cassandra Buzby, Heiko Müller, Naomi Ziv, Sasha F Levy, and Mark L Siegal. High-throughput microcolony growth analysis from suboptimal low-magnification micrographs. June 2021.

Plavskin, de Biase, Roland F Schwarz, and Mark L Siegal. The rate of spontaneous mutations in yeast deficient for MutSβ function. August 2022.

Ryan Poplin, Valentin Ruano-Rubio, Mark A DePristo, Tim J Fennell, Mauricio O Carneiro, Geraldine A Van der Auwera, David E Kling, Laura D Gauthier, Ami Levy-Moonshine, David Roazen, Khalid Shakir, Joel Thibault, Sheila Chandran, Chris Whelan, Monkol Lek, Stacey Gabriel, Mark J Daly, Ben Neale, Daniel G MacArthur, and Eric Banks. Scaling accurate genetic variant discovery to tens of thousands of samples. July 2018.

C A Powell, S Klares, G O'Connor, and J S Brody. Loss of heterozygosity in epithelial cells obtained by bronchial brushing: clinical utility in lung cancer. *Clin. Cancer Res.*, 5(8):2025–2034, August 1999.

Awal Prasetyo, Udadi Sadhana, and Jethro Budiman. Nasal mucociliary clearance in smokers: A systematic review. *Int Arch Otorhinolaryngol*, 25(1):e160–e169, January 2021.

Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010.

Alison Rattray, Gustavo Santoyo, Brenda Shafer, and Jeffrey N Strathern. Elevated mutation rate during meiosis in saccharomyces cerevisiae. *PLoS Genet.*, 11(1): e1004910, January 2015.

Timothy Ravasi, Harukazu Suzuki, Carlo Vittorio Cannistraci, Shintaro Katayama, Vladimir B Bajic, Kai Tan, Altuna Akalin, Sebastian Schmeier, Mutsumi Kanamori-Katayama, Nicolas Bertin, Piero Carninci, Carsten O Daub, Alistair R R Forrest, Julian Gough, Sean Grimmond, Jung-Hoon Han, Takehiro Hashimoto, Winston Hide, Oliver Hofmann, Atanas Kamburov, Mandeep Kaur, Hideya Kawaji, Atsutaka Kubosaki, Timo Lassmann, Erik van Nimwegen, Cameron Ross MacPherson, Chihiro Ogawa, Aleksandar Radovanovic, Ariel Schwartz, Rohan D Teasdale, Jesper Tegnér, Boris Lenhard, Sarah A Teichmann, Takahiro Arakawa, Noriko Ninomiya, Kayoko Murakami, Michihira Tagami, Shiro Fukuda, Kengo Imamura, Chikatoshi Kai, Ryoko Ishihara, Yayoi Kitazume, Jun Kawai, David A Hume, Trey Ideker, and Yoshihide Hayashizaki. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–752, March 2010.

Christopher Ricketts, Maurice P Zeegers, Jan Lubinski, and Eamonn R Maher. Analysis of germline variants in CDH1, IGFBP3, MMP1, MMP3, STK15 and VEGF in familial and sporadic renal cell carcinoma. *PLoS One*, 4(6):e6037, June 2009.

M Patricia Rivera, Atul C Mehta, and Momen M Wahidi. Establishing the diagnosis of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American college of chest physicians evidence-based clinical practice guidelines. *Chest*, 143(5 Suppl): e142S–e165S, May 2013.

F Rolland, J Winderickx, and J Thevelein. Glucose-sensing and -signalling mechanisms in yeast, 2002.

Nina V Romanova and Gray F Crouse. Different roles of eukaryotic MutS and MutL complexes in repair of small insertion and deletion loops in yeast. *PLoS Genet.*, 9 (10):e1003920, October 2013.

Carlos Rosales. Neutrophil: A cell with many roles in inflammation or several cell types? *Front. Physiol.*, 9:113, February 2018.

Rohini Roy, Jarin Chun, and Simon N Powell. BRCA1 and BRCA2: different roles in a common pathway of genome protection. *Nat. Rev. Cancer*, 12(1):68–78, December 2011.

Bríd M Ryan and Jessica M Faupel-Badger. The hallmarks of premalignant conditions: a molecular basis for cancer prevention. *Semin. Oncol.*, 43(1):22–35, February 2016.

Federica M O Sartori, Cassandra Buzby, Yevgeniy Plavskin, and Mark L Siegal. High-Throughput live imaging of microcolonies to measure heterogeneity in growth and gene expression. *J. Vis. Exp.*, (170), April 2021.

Rosalyn W Sayaman, Mohamad Saad, Vésteinn Thorsson, Donglei Hu, Wouter Hendrickx, Jessica Roelands, Eduard Porta-Pardo, Younes Mokrab, Farshad Farshidfar, Tomas Kirchhoff, Randy F Sweis, Oliver F Bathe, Carolina Heimann, Michael J Campbell, Cynthia Stretch, Scott Huntsman, Rebecca E Graff, Najeeb Syed, Laszlo Radvanyi, Simon Shelley, Denise Wolf, Francesco M Marincola, Michele Ceccarelli, Jérôme Galon, Elad Ziv, and Davide Bedognetti. Germline genetic contribution to the immune landscape of cancer. *Immunity*, 54(2):367–386.e8, February 2021.

Katja Schwartz and Gavin Sherlock. Preparation of yeast DNA sequencing libraries. *Cold Spring Harb. Protoc.*, 2016(10), October 2016.

Barbara Seliger, Matthias Kloor, and Soldano Ferrone. HLA class II antigen-processing pathway in tumors: Molecular defects and clinical relevance. *Oncoimmunology*, 6 (2):e1171447, February 2017.

Alexandre Serero, Claire Jubin, Sophie Loeillet, Patricia Legoix-Né, and Alain G Nicolas. Mutational landscape of yeast mutator strains, 2014.

Sahar Shahamatdar, Meng Xiao He, Matthew A Reyna, Alexander Gusev, Saud H AlDubayan, Eliezer M Van Allen, and Sohini Ramachandran. Germline features associated with immune infiltration in solid tumors. *Cell Rep.*, 30(9):2900–2908.e4, March 2020.

Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, November 2003.

Nathaniel P Sharp, Linnea Sandell, Christopher G James, and Sarah P Otto. The genome-wide rate and spectrum of spontaneous mutations differ between haploid and diploid yeast. *Proc. Natl. Acad. Sci. U. S. A.*, 115(22):E5046–E5055, May 2018.

P G Shields. Molecular epidemiology of lung cancer. *Ann. Oncol.*, 10 Suppl 5:S7–11, 1999.

H Shimodaira, N Filosi, H Shibata, T Suzuki, P Radice, R Kanamaru, S H Friend, R D Kolodner, and C Ishioka. Functional analysis of human MLH1 mutations in saccharomyces cerevisiae. *Nat. Genet.*, 19(4):384–389, August 1998.

D Shinde, Y Lai, F Sun, and N Arnheim. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. *Nucleic Acids Res.*, 2003.

E A Sia, R J Kokoska, M Dominska, P Greenwell, and T D Petes. Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. *Mol. Cell. Biol.*, 17(5):2851–2858, May 1997.

E A Sia, M Dominska, L Stefanovic, and T D Petes. Isolation and characterization of point mutations in mismatch repair genes that destabilize microsatellites in yeast. *Mol. Cell. Biol.*, 21(23):8157–8167, December 2001.

Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2020. *CA Cancer J. Clin.*, 70(1):7–30, January 2020.

Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *J. Stat. Softw.*, 39(5):1–13, March 2011.

D J Slamon, B Leyland-Jones, S Shak, H Fuchs, V Paton, A Bajamonde, T Fleming, W Eiermann, J Wolter, M Pegram, J Baselga, and L Norton. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.*, 344(11):783–792, March 2001.

D P Slaughter, H W Southwick, and W Smejkal. Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. *Cancer*, 6(5): 963–968, September 1953.

Erik L L Sonnhammer and Gabriel Östlund. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res.*, 43(Database issue):D234–9, January 2015.

Avrum Spira, Jennifer Beane, Vishal Shah, Gang Liu, Frank Schembri, Xuemei Yang, John Palma, and Jerome S Brody. Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc. Natl. Acad. Sci. U. S. A.*, 101(27):10143–10148, July 2004.

Avrum Spira, Jennifer E Beane, Vishal Shah, Katrina Steiling, Gang Liu, Frank Schembri, Sean Gilman, Yves-Martine Dumas, Paul Calner, Paola Sebastiani, Sriram Sridhar, John Beamis, Carla Lamb, Timothy Anderson, Norman Gerry, Joseph Keane, Marc E Lenburg, and Jerome S Brody. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat. Med.*, 13(3): 361–366, March 2007.

Sriram Sridhar, Frank Schembri, Julie Zeskind, Vishal Shah, Adam M Gustafson, Katrina Steiling, Gang Liu, Yves-Martine Dumas, Xiaohui Zhang, Jerome S Brody, Marc E Lenburg, and Avrum Spira. Smoking-induced gene expression changes in the bronchial airway are reflected in nasal and buccal epithelium. *BMC Genomics*, 9:259, May 2008.

Visish Srinivasan, Andres Kriete, Ahmet Sacan, and S Michal Jazwinski. Comparing the yeast retrograde response and NF-κB stress responses: implications for aging. *Aging Cell*, 9(6):933–941, December 2010.

Katrina Steiling, John Ryan, Jerome S Brody, and Avrum Spira. The field of tissue injury in the lung and airway. *Cancer Prev. Res.*, 1(6):396–403, November 2008.

Rafael Stelmach, Frederico Leon Arrabal Fernandes, Regina Maria Carvalho-Pinto, Rodrigo Abensur Athanazio, Samia Zahi Rached, Gustavo Faibischew Prado, and Alberto Cukier. Comparison between objective measures of smoking and self-reported smoking status in patients with asthma or COPD: are our patients telling us the truth? *J. Bras. Pneumol.*, 41(2):124–132, March 2015.

M Strand, T A Prolla, R M Liskay, and T D Petes. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature*, 365 (6443):274–276, September 1993.

M Strand, M C Earley, G F Crouse, and T D Petes. Mutations in the MSH3 gene preferentially lead to deletions within tracts of simple repetitive DNA in saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U. S. A.*, 92(22):10418–10421, October 1995.

Erin D Strome, Xiaowei Wu, Marek Kimmel, and Sharon E Plon. Heterozygous screen in saccharomyces cerevisiae identifies dosage-sensitive genes that affect chromosome stability. *Genetics*, 178(3):1193–1207, March 2008.

Shane Sullivan, Miriam Tosetto, David Kevans, Alan Coss, Laimun Wang, Diarmuid O'Donoghue, John Hyland, Kieran Sheahan, Hugh Mulcahy, and Jacintha O'Sullivan. Localization of nuclear cathepsin L and its association with disease progression and poor outcome in colorectal cancer. *Int. J. Cancer*, 125(1):54–61, July 2009.

Hye-Jin Sung, Jung-Mo Ahn, Yeon-Hee Yoon, Tai-Youn Rhim, Choon-Sik Park, Jae-Yong Park, Soo-Youn Lee, Jong-Won Kim, and Je-Yoel Cho. Identification and validation of SAA as a potential lung cancer biomarker and its involvement in metastatic pathogenesis of lung cancer. *J. Proteome Res.*, 10(3):1383–1395, March 2011.

Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.*, 71(3):209–249, May 2021.

Ximing Tang, Hisayuki Shigematsu, B Nebiyou Bekele, Jack A Roth, John D Minna, Waun Ki Hong, Adi F Gazdar, and Ignacio I Wistuba. EGFR tyrosine kinase domain mutations are detected in histologically normal respiratory epithelium in lung cancer patients. *Cancer Res.*, 65(17):7568–7572, September 2005.

D Tautz and M Renz. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Res.*, 12(10):4127–4138, May 1984.

Ying-Hock Teng, Te-Hsiung Liu, Hsien-Chun Tseng, Tsung-Te Chung, Chia-Ming Yeh, Yu-Chiung Li, Yu-Hsiang Ou, Long-Yau Lin, Hsiu-Ting Tsai, and Shun-Fa Yang. Contribution of genetic polymorphisms of stromal cell-derived factor-1 and its receptor, CXCR4, to the susceptibility and clinicopathologic development of oral cancer, 2009.

N A Timchenko, A L Lu, X Welm, and L T Timchenko. CUG repeat binding protein (CUGBP1) interacts with the 5' region of C/EBP mRNA and regulates translation of C/EBP isoforms, 1999.

Hilary A Tindle, Meredith Stevenson Duncan, Robert A Greevy, Ramachandran S Vasan, Suman Kundu, Pierre P Massion, and Matthew S Freiberg. Lifetime smoking history and risk of lung cancer: Results from the framingham heart study. *J. Natl. Cancer Inst.*, 110(11):1201–1207, November 2018.

M Toyota, N Ahuja, M Ohe-Toyota, J G Herman, S B Baylin, and J P Issa. CpG island methylator phenotype in colorectal cancer. *Proc. Natl. Acad. Sci. U. S. A.*, 96(15): 8681–8686, July 1999.

Norihisa Uemura, Yukihiro Nakanishi, Hoichi Kato, Shigeru Saito, Masato Nagino, Setsuo Hirohashi, and Tadashi Kondo. Transglutaminase 3 as a prognostic biomarker in esophageal cancer revealed by proteomics, 2009.

Geraldine A Van der Auwera and Brian D O'Connor. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. "O'Reilly Media, Inc.", April 2020.

Adriaan van der Graaf, René Wardenaar, Drexel A Neumann, Aaron Taudt, Ruth G Shaw, Ritsert C Jansen, Robert J Schmitz, Maria Colomé-Tatché, and Frank Johannes. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. *Proceedings of the National Academy of Sciences*, 112(21):6676–6681, 2015.

Md Vasimuddin, Sanchit Misra, Heng Li, and Srinivas Aluru. Efficient Architecture-Aware acceleration of BWA-MEM for multicore systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pages 314–324, May 2019.

Michiel C Verboom, Jacqueline S L Kloth, Jesse J Swen, Tahar van der Straaten, Judith V M Bovée, Stefan Sleijfer, Anna K L Reyners, Ron H J Mathijssen, Henk-Jan Guchelaar, Neeltje Steeghs, and Hans Gelderblom. Genetic polymorphisms in angiogenesis-related genes are associated with worse progression-free survival of patients with advanced gastrointestinal stromal tumours treated with imatinib, 2017.

Maria Lucia Carneiro Vieira, Luciane Santini, Augusto Lima Diniz, and Carla de Freitas Munhoz. Microsatellite markers: what they mean and why they are so useful. *Genet. Mol. Biol.*, 39(3):312–328, August 2016.

Tonya Walser, Xiaoyan Cui, Jane Yanagawa, Jay M Lee, Eileen Heinrich, Gina Lee, Sherven Sharma, and Steven M Dubinett. Smoking and lung cancer: the role of inflammation. *Proc. Am. Thorac. Soc.*, 5(8):811–815, December 2008.

Biao Wang, Zhen Tang, Huiyuan Gong, Li Zhu, and Xuegang Liu. Wnt5a promotes epithelial-to-mesenchymal transition and metastasis in non-small-cell lung cancer. *Biosci. Rep.*, 37(6), December 2017.

Gui-Zhen Wang, Xin Cheng, Bo Zhou, Zhe-Sheng Wen, Yun-Chao Huang, Hao-Bin Chen, Gao-Feng Li, Zhi-Liang Huang, Yong-Chun Zhou, Lin Feng, Ming-Ming Wei, Li-Wei Qu, Yi Cao, and Guang-Biao Zhou. The chemokine CXCL13 in lung cancers associated with environmental polycyclic aromatic hydrocarbons pollution. *Elife*, 4, November 2015.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas Pan-Cancer analysis project. *Nat. Genet.*, 45(10):1113–1120, September 2013.

Thomas Willems, Dina Zielinski, Jie Yuan, Assaf Gordon, Melissa Gymrek, and Yaniv Erlich. Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods*, 14(6):590–592, June 2017.

E Winter and A Varshavsky. A DNA binding protein that recognizes oligo(dA).oligo(dT) tracts. *EMBO J.*, 8(6):1867–1877, June 1989.

I I Wistuba, S Lam, C Behrens, A K Virmani, K M Fong, J LeRiche, J M Samet, S Srivastava, J D Minna, and A F Gazdar. Molecular damage in the bronchial epithelium of current and former smokers. *J. Natl. Cancer Inst.*, 89(18):1366–1373, September 1997.

Kuan-Li Wu, Ying-Ming Tsai, Chi-Tun Lien, Po-Lin Kuo, Hung, and Jen-Yu. The roles of MicroRNA in lung cancer. *Int. J. Mol. Sci.*, 20(7), March 2019.

Xiangbing Wu, Wei Cao, Xu Wang, Jianjun Zhang, Zhongjing Lv, Xing Qin, Yadi Wu, and Wantao Chen. TGM3, a candidate tumor suppressor gene, contributes to human head and neck cancer. *Mol. Cancer*, 12(1):151, December 2013.

Zeng-Hong Wu, Fucheng Cai, and Yi Zhong. Comprehensive analysis of the expression and prognosis for GBPs in head and neck squamous cell carcinoma. *Sci. Rep.*, 10 (1):6085, April 2020.

Yuzo Yamamoto, Chikako Kiyohara, Saiko Suetsugu-Ogata, Naoki Hamada, and Yoichi Nakanishi. Biological interaction of cigarette smoking on the association between genetic polymorphisms involved in inflammation and the risk of lung cancer: A case-control study in japan. *Oncol. Lett.*, 13(5):3873–3881, May 2017.

Ichiro Yoshino, Takuro Kometani, Fumihiro Shoji, Atsushi Osoegawa, Taro Ohba, Hidenori Kouso, Tomoyoshi Takenaka, Tomofumi Yohena, and Yoshihiko Maehara. Induction of epithelial-mesenchymal transition-related genes by benzo[a]pyrene in lung cancer cells. *Cancer*, 110(2):369–374, July 2007.

E T Young, J S Sloan, and K Van Riper. Trinucleotide repeats are clustered in regulatory genes in saccharomyces cerevisiae. *Genetics*, 154(3):1053–1068, March 2000.

Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofiler: an R package for comparing biological themes among gene clusters. *OMICS*, 16(5): 284–287, May 2012.

Sarah Zanders, Xin Ma, Arindam Roychoudhury, Ryan D Hernandez, Ann Demogines, Brandon Barker, Zhenglong Gu, Carlos D Bustamante, and Eric Alani. Detection of heterozygous mutations in the genome of mismatch repair defective diploid yeast using a bayesian approach. *Genetics*, 186(2):493–503, October 2010.

Chaoqi Zhang, Guochao Zhang, Nan Sun, Zhen Zhang, Zhihui Zhang, Yuejun Luo, Yun Che, Qi Xue, and Jie He. Comprehensive molecular analyses of a TNF family-based signature with regard to prognosis, immune features, and biomarkers for immunotherapy in lung adenocarcinoma. *EBioMedicine*, 59:102959, September 2020.

Wei Zhang, Sabine C Glöckner, Mingzhou Guo, Emi Ota Machida, David H Wang, Hariharan Easwaran, Leander Van Neste, James G Herman, Kornel E Schuebel, D Neil Watkins, Nita Ahuja, and Stephen B Baylin. Epigenetic inactivation of the canonical wnt antagonist SRY-box containing gene 17 in colorectal cancer. *Cancer Res.*, 68(8):2764–2772, April 2008.

Xiaoling Zhang, Paola Sebastiani, Gang Liu, Frank Schembri, Xiaohui Zhang, Yves Martine Dumas, Erika M Langer, Yuriy Alekseyev, George T O'Connor, Daniel R Brooks, Marc E Lenburg, and Avrum Spira. Similarities and differences between smoking-related gene expression in nasal and bronchial epithelium. *Physiol. Genomics*, 41(1):1–8, March 2010.

Yuan O Zhu, Mark L Siegal, David W Hall, and Dmitri A Petrov. Precise estimates of mutation rate and spectrum in yeast. *Proc. Natl. Acad. Sci. U. S. A.*, 111(22): E2310–8, June 2014.

Dandan Zong, Xiangming Liu, Jinhua Li, Ruoyun Ouyang, and Ping Chen. The role of cigarette smoke-induced epigenetic alterations in inflammation. *Epigenetics Chromatin*, 12(1):65, November 2019.

List of publications

Yevgeniy Plavskin*, **Maria Stella de Biase***, Roland F. Schwarz, Mark L. Siegal, *The rate of spontaneous mutations in yeast deficient for MutSβ function.* bioRxiv (2022), doi:10.1101/2022.08.25.505291

Maria Stella de Biase*, Florian Massip*, Tzu-Ting Wei, Federico M. Giorgi, Rory Stark, Amanda Stone, Amy Gladwell, Martin O'Reilly, Ines de Santiago, Kerstin Meyer, Florian Markowetz, Bruce A.J. Ponder, Robert C. Rintoul, Roland F. Schwarz, *Smoking-dependent expression alterations in nasal epithelium reveal immune impairment linked to germline variation and lung cancer risk.* bioRxiv (2021), doi:10.1101/2021.11.24.21266740.

Hananeh Aliee*, Florian Massip*, Cancan Qi*, **Maria Stella de Biase***, Jos van Nijnatten*, Elin T.G. Kersten*, Nazanin Z. Kermani*, Basil Khuder*, Judith M. Vonk, Roel C.H. Vermeulen, U-BIOPRED study group, Cambridge Lung Cancer Early Detection Programme, INER-Ciencias Mexican Lung Program, Margaret Neighbors, Gaik W. Tew, Michele Grimbaldeston, Nick H.T. Ten Hacken, Sile Hu, Yike Guo, Xiaoyu Zhang, Kai Sun, Pieter S. Hiemstra, Bruce A.J. Ponder, Mika J. Makela, Kristiina Malmstrom, Robert C. Rintoul, Paul A. Reyfman, Fabian J. Theis, Corry-Anke Brandsma, Ian Adcock, Wim Timens, Cheng J. Xu, Maarten van den Berge, Roland F. Schwarz, Gerard H. Koppelman, Martijn C. Nawijn, Alen Faiz, *Determinants of expression of SARS-CoV-2 entry-related genes in upper and lower airways.* Allergy. 77, 690–694 (2022), https://doi.org/10.1111/all.15152

* indicates co-first authorship

Appendix A

List of supplementary tables

All supplementary tables are available as part of the digital supplementary material.

Supplementary table 1: GO terms enriched in the list of genes differentially expressed in clinic-referred patients compared to healthy volunteers.

Supplementary table 2: GO terms enriched in the list of genes differentially expressed in bronchial samples from clinic cancer compared to clinic benign patients.

Supplementary table 3: List of genes classified as affected by smoking in healthy volunteers, clinic-referred patients, or both groups (Sections 3.2.2 and 3.2.3). The *Change_HV* and *change_Clinic* columns report the direction of the expression change in current smokers relative to healthy never smokers; for CA genes, the columns report whether there is an increase or decrease of expression in ex smokers. The *class_HV* and *class_Clinic* columns report the reversibility class to which the gene has been assigned in healthy and clinic subjects.

Supplementary table 4: GO terms enriched in genes changing classification from RR in healthy volunteers to CA in clinic-referred patients.

Supplementary table 5: GO terms enriched in the targets of the 25 smoke injury master regulator TFs. Enrichment analysis was performed on the list of target genes of the 4 TF groups appearing in the network representation in Figure 3.9a.