Genetic differences in lung adenocarcinoma cells from patients of African and European ancestry

Karabo Diseko

Supervisor: Prof. Nicola Mulder

Co-supervisor: Dr Musalula Sinkala

Submitted for the degree of Master of Science in Medicine

specialising in Bioinformatics

Division of Computational Biology

Department of Integrative Biomedical Sciences

Faculty of Health Sciences

University of Cape Town

12 November 2023

Declaration

I, Karabo Mompati Diseko, hereby declare that the work on which this dissertation is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: Signed by candidate .....

Date: 12 November 2023

Acknowledgements

Thank you:

To my parents whose genuine love and support is unconditional,

To my brothers, nephews, and family who deeply inspire me,

To my supervisors, who effectively guided me through this process,

To my CBIO lab mates, who were my friendly companions along this journey.

Your presence and contributions were foundational to the completion of this thesis, and for that I am grateful.

Thank you!

Abstract

In the past two decades, advancements in cancer genetics research have significantly enhanced our molecular comprehension of human cancers. This progress has led to the development of improved clinical tools for the precise diagnosis, prognosis prediction, and tailored treatment of cancers. However, the predominant focus of this research has been on individuals of European ancestry, inadvertently marginalizing the diverse genetic landscapes represented by other ethnic populations. Given minor differences in the genetic makeup across diverse ethnicities, specific cancer genetic variants prevalent in certain ethnic groups may remain overlooked within the current research.

Some studies have indeed illuminated nuanced distinctions in the genetic architecture of cancers among patients of varying ethnic backgrounds. Disparities in cancer incidence and outcome between patients of different ethnicities have also been identified. These distinctions stem from a combination of environmental and biological factors, collectively shaping the intricate interplay of cancer genetics and its clinical manifestations. This study endeavours to elucidate clinically significant disparities in lung adenocarcinoma (LUAD) genetics across distinct ethnicities, particularly focusing on African ancestry (AA) and European ancestry (EA) populations. A meticulous comparison of genetic traits within LUAD cells derived from these ethnic groups is conducted to pinpoint genetic variances that hold potential biological relevance.

Leveraging data from The Cancer Genome Atlas' lung adenocarcinoma (TCGA-LUAD) study, samples were stratified based on self-reported racial classifications into African ancestry (AA) and European ancestry (EA) groups. Propensity score matching (PSM) was meticulously employed to mitigate disparities in crucial clinical attributes, ensuring a balanced basis for subsequent genetic comparisons. A total of 147 EA and 49 AA samples were extracted following PSM, forming the basis for comprehensive comparisons of gene expression, copy number alterations, and mutation frequencies between the two ethnic cohorts.

Key genetic disparities between the two groups were discerned, including 371 significantly differentially expressed (SDE) genes, a higher incidence of copy number alterations in the AA group compared to the EA group, and 101 genes exhibiting varying mutation frequencies between the two groups. An analysis of the biological functions impacted by these genetic variances revealed involvement in critical processes such as cellular response to xenobiotics, hormone metabolism and regulation, mitochondrial energy production, and epithelial-mesenchymal transition. We posit that clinically relevant biological distinctions in LUAD tumours between AA and EA patients stem from differential expression and mutations in genes encoding pivotal proteins such as UDP glucuronosyltransferases and cytochrome P450s, among others. Variations in the sequence and

expression of these genes can significantly influence drug response and hallmark cancer cell characteristics, including energy production and epithelial-mesenchymal transition.

Despite the limitation of a relatively small sample size, this study illuminates genetic disparities that underpin clinically significant differences in tumour biology between LUAD patients of African and European ancestry.

List of abbreviations

AA	African Ancestry
CYP	Cytochrome P450
DNA	Deoxyribonucleic acid
EA	European Ancestry
EMT	Epithelial to Mesenchymal Transition
GO-BP	Gene Ontology Biological Process
GSEA	Gene Set Enrichment Analysis
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
miRNA	Micro Ribonucleic Acid
mRNA	Messenger Ribonucleic Acid
NSCLC	Non-Small Cell Lung Cancer
ORA	Over-Representation Analysis
PSM	Propensity Score Matching
RecCNA	Recurrently Copy Number Altered
RNA	Ribonucleic acid
SCNA	Somatic Copy Number Alteration
SDE	Significantly Differentially Expressed
TCGA	The Cancer Genome Atlas
TCGA-LUAD	The Cancer Genome Atlas' Lung Adenocarcinoma Study

Table of contents

1. Literature Review

1.1 Lung Cancer

Cancer is the second leading cause of death worldwide (after cardiovascular disease) ("WHO cancer fact sheet," 2022). Therefore, scientific research into the causes, diagnosis, and progression of cancer is a major focus. Cancer results from the transformation of normal cells into malignant cells, through gene alterations that affect essential cellular processes like cell growth and division (Ponder, 2001). Genetic changes linked to cancer arise from the interplay between genes and their surrounding environment. Carcinogens, which can trigger cancer development, are environmental factors that can be physical (like ultraviolet and ionizing rays), chemical (such as asbestos, alcohol, or tobacco smoke), or biological (like infections from viruses, bacteria, or parasites) ("NIH About Cancers: Understanding Cancers," 2022; "WHO cancer fact sheet," 2022).

Within the realm of cancers, lung cancer ranks second in prevalence and records the highest death rate, with only 19% of patients surviving beyond five years. There are two primary classifications of lung cancer: non-small cell lung cancer (NSCLC), accounting for 85% of diagnoses, and small-cell lung cancer (SCLC), representing the remaining 15% (Fig 1.1). NSCLC can be further categorised into three histological variants: large-cell lung cancer, squamous-cell lung carcinoma, and lung adenocarcinoma (LUAD). (Herbst et al., 2008; Schabath and Cote, 2019).

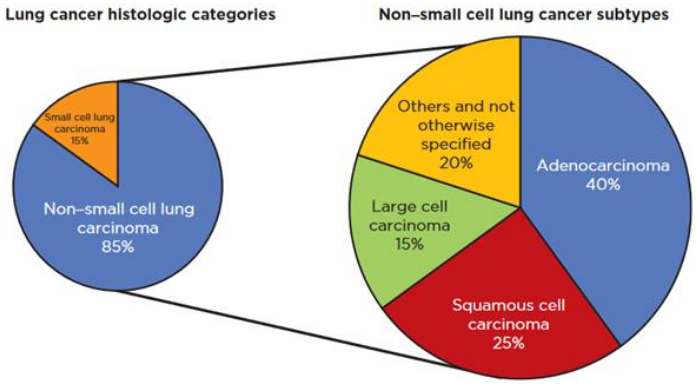


Figure 1.1. Distribution of NSCLC histologic types. Adapted from Schabath et al.

Some primary risk factors associated with the development of lung cancer include exposure to tobacco smoke (including second-hand), radon gas, and asbestos, as well as a history of respiratory ailments such as asthma, pneumonia and tuberculosis (Schabath and Cote, 2019). Over time, interactions between lung cells and their environment, particularly the above-cited risk factors, can induce genetic modifications, turning benign cells into malignant ones. These malignant cells divide

in a dysregulated manner and eventually form a tumour (Ponder, 2001). By pinpointing genetic alterations linked to lung cancer incidence, prognosis, and treatment response, researchers have enhanced the clinical screening and treatment approaches for lung cancer (Šutić et al., 2021).

1.2 Lung cancer genetics

The improvement of tools for detecting alterations in gene sequences, gene expression and protein expression in cells has allowed for more specific characterisation of cancers. This improved molecular characterisation of cancers can be considered alongside histology to better predict cancer susceptibility, prognosis, and sensitivity to various treatment regimens. (Shames and Wistuba, 2014). In this introductory section, we focus specifically on the genetic alterations linked to NSCLC as the primary subject of this study, LUAD, falls under the NSCLC category. Here, we describe the tools employed to identify cancer-associated genetic changes, highlight certain pertinent gene alterations observed, and discuss the clinical relevance of several of these genetic changes.

1.2.1 Tools for genetic characterisation of cells

Genetic changes can happen in three primary ways: through the alteration of the DNA sequence itself (genomic), changes in the regulation of gene expression (epigenetic), or changes in mRNA expression or structure (transcriptomic). Each of these genetic changes can influence the protein expression in cells, potentially transforming a normal cell into a malignant one (Climente-González et al., 2017; Li et al., 2015; Ponder, 2001). In our study, we focus on genomic and transcriptomic modifications. Below, we provide a brief overview of the main tools employed by researchers to gather genomic and transcriptomic data from cancer cells.

Genomic alterations, encompassing the deletion, addition, substitution, or translocation of single or multiple DNA bases, are identified using DNA sequencing (DNA-seq) techniques. These techniques include whole-genome sequencing, which examines the entire DNA sequence, whole-exome sequencing that focuses on the coding regions, and single nucleotide polymorphism-based platforms (McCarroll et al., 2008; Shendure and Ji, 2008). DNA-seq is a high throughput method capable of sequencing the complete exome or genome from a sample, while single nucleotide polymorphism-based platforms detect single nucleotide variations at specified DNA regions (McCarroll et al., 2008; Shendure and Ji, 2008).

Transcriptomic (RNA) alterations – which involve changes in the rate of mRNA formation (gene expression) or changes in the sections of a gene used to form mRNA (alternative splicing) - are usually measured through high-throughput sequencing, microarray-based tools, or quantitative reverse transcription polymerase chain reactions (qRT-PCR) (Gunaratne et al., 2012; Persson et al., 2005; Provenzano and Mocellin, 2007; Slonim and Yanai, 2009; Wang et al., 2009). High-throughput

sequencing technologies, like RNA-seq and MicroRNA-seq, are used to determine the nucleotide sequence and abundance of all protein-coding and/or non-protein-coding RNA in the cell (Gunaratne et al., 2012; Wang et al., 2009). Microarray-based tools use probes complementary to an RNA sequence of interest to quantify its abundance (Slonim and Yanai, 2009). QRT-PCR uses reverse transcription to convert RNA into complementary DNA (cDNA) , then infers the abundance of a particular RNA sequence through the amplification and detection of its cDNA sequence (Persson et al., 2005; Provenzano and Mocellin, 2007).

1.2.2 Genetic changes associated with non-small cell lung cancer (NSCLC)

Two categories of genes, tumour suppressors and oncogenic driver genes, play a role in the onset and progression of cancer. Tumour suppressor genes help contain cell growth and proliferation to within normal levels. Cells can become cancerous when these genes are rendered inactive due to loss-of-function mutations and alterations. Conversely, oncogenes promote cellular growth and proliferation in normal cells, but gain-of-function mutations or alterations leading to their overexpression can predispose cells to become cancerous. Tumour suppressor genes whose mutations have been associated with NSCLC include *LKB1*, *RASSF1A*, *FHIT*, *CDKN2A*, and *TP53* (El-Telbany and Ma, 2012; Herbst et al., 2008; Shames and Wistuba, 2014). Conversely, oncogenic driver genes whose mutations are frequently associated with NSCLC include *EGFR*, *KRAS*, *BRAF*, *PIK3CA*, *ALK*, *RET* and *VEGF* (El-Telbany and Ma, 2012; Herbst et al., 2008; Sinkala, 2023).

Over the past two decades, Improvements in the quality and cost of high-throughput transcriptome sequencing have made it possible for researchers to associate certain transcriptome changes with NSCLC. Chen et al. (2014) conducted a meta-analysis looking into gene expression changes associated with LUAD cells, by comparing 783 LUAD samples to 127 normal samples, and found 11 genes that were significantly (FDR<0.00001) overexpressed in the LUAD cells. They highlighted *PTK7* as particularly highly expressed and specific to LUAD. The alteration in the normal splicing of several genes, including *cyclin-D1*, *KLF6*, *VEGF-A*, *CD44*, *FHIT*, *BCL-X*, and *MET*, has been shown to contribute to the onset and progression of NSCLC tumours (de Fraipont et al., 2019).

Research into DNA sequence and RNA expression alterations continues, probing the mechanisms that transform normal cells into malignant ones. As more data emerges, our understanding of lung cancer's molecular nuances deepens, reinforcing specific lung cancer-associated genetic shifts, unveiling new ones, and diminishing the importance of others. This cyclical progression in cancer genetic research reflects the ever-evolving grasp of cancer genetics and its clinical implications.

1.2.3 Clinical relevance of NSCLC genetic research

Various factors in the detection, treatment, and monitoring of NSCLC contribute to its poor survival rate. For example, the diagnosis of NSCLC at a late-stage relative to other cancers and the resistance to treatment that NSCLCs often develop. Foundational in dealing with these issues is improving the molecular understanding of lung cancer. Research has shown that testing for lung cancer-associated genetic variation in circulating genetic material and resected lung tissue can aid in early diagnosis, prognosis forecasting, and formulating treatment plans for patients, thereby mitigating the disease's typically high mortality rate.

1.2.3.1 Diagnosis of cancer

Diagnosis involves the detection and characterisation of a cancer. The early diagnosis of cancer is important and can significantly improve clinical outcomes. Most lung cancer patients don't display noticeable symptoms (apart from coughing) during its initial stages (Restrepo et al., 2023). This makes early diagnosis of lung cancer challenging. Currently, the initial detection of lung cancer in patients is primarily done through imaging techniques such as a chest X-ray, computed tomography(CT) and/or positron emission tomography-computer tomography (PET-CT) (Šutić et al., 2021). Following the initial detection of cancer, the subsequent phase in diagnosis involves characterising the specific nature of the cancer.

Characterisation involves determining the malignancy, histologic subtype, size, and cancer stage of cancer. This is done by imaging, assessing morphology by histology and/or cytology, protein biomarker identification, and genetic biomarker identification (Šutić et al., 2021). Morphology and protein expression differences are the most popular means of differentiating between lung cancer histologic subtypes. In instances where tissue samples are scarce or the cancer displays poor morphological differentiation, merely evaluating the morphology and protein expression of the cancer tissue might fall short of providing an accurate histologic subtype classification. This is where the non-invasive and non-tissue-requiring method of using liquid biopsies can be useful in characterising cancers. Liquid biopsies involve isolating and assessing any tumour-derived material circulating in the blood or other bodily fluids. Liquid blood biopsies can contain circulating tumour cells, exomes (with RNA from tumours), tumour associate proteins, and cell-free tumour-derived genetic material such as circulating tumour DNA, micro RNAs and single-strand RNAs (Restrepo et al., 2023) (Šutić et al., 2021). Among others, circulating protein biomarkers like *CYFRA 21-1*, *CEA*, and *CA125* are used individually or in combination for the detection of lung cancer via liquid biopsies (Chen et al., 2015; Okamura et al., 2013; Rastel et al., 1994). The potential to use circulating tumour-derived microRNAs for the detection and characterisation of lung cancer in the future is promising, owing to their stability and research showing the capacity of miRNA signatures to differentiate

between cancer patients, non-cancer individuals, and different lung cancer subtypes (Gao et al., 2016; Gilad et al., 2012; Šutić et al., 2021; Wang et al., 2015; Weber et al., 2010). Although it's a promising tool, using microRNA and circulating microRNAs as a lung cancer diagnostic and characterisation tool is not ready for clinical use. Research efforts toward standardising the type of sample (serum vs. plasma) and sample preparation protocols, as well as larger scale validation of potential miRNA biomarkers material is needed for the potential of this non-invasive diagnostic biomarker to be realised (Restrepo et al., 2023).

1.2.3.2 Cancer therapies

Various anti-cancer treatments have been developed that target and neutralise the oncogenic gene alterations responsible for tumour growth. These therapies have demonstrated efficacy in extending the survival duration of cancer patients (Barbar et al., 2022; Šutić et al., 2021). Such targeted therapies include the *EGFR* inhibitors erlotinib and gefitinib, as well as the *ALK* inhibitor crizotinib. Other oncogenic gene alterations against which inhibitors have been developed include ROS1 gene fusions, *RET* fusions, *MET* copy number amplification, and *P1K3CA* mutations. Since *EGFR* mutations, as well as *ALK* and *RET* fusions, are amongst the more common genetic causes of LUAD, many clinical guidelines recommend multigene testing for alterations in those genes and others for all LUADs in order to identify the best therapeutic measures (Šutić et al., 2021). Similar genetic biomarker test recommendations exist for the other lung cancer types. The genetically complex nature of NSCLC tumours often makes them resistant to the targeted therapies meant to counteract the effects of a particular cancer-related gene mutation to kill that cancer cell. Employing combined therapies, which utilise multiple targeted gene treatments in conjunction with other treatments like immunotherapy or chemotherapy, provides a more potent NSCLC treatment approach less susceptible to resistance. The precision of targeted therapies that are tailored to an individual's tumour enhances their effectiveness. Over time, more research to illuminate the genetic landscape of cancer will improve cancer treatment by uncovering additional molecular targets that can be used to kill cancer cells (Shames and Wistuba, 2014; Barbar et al., 2022)

1.2.3.3 Prognosis

Prognostic biomarkers are those which give insight into the likely clinical outcome for a patient. A patient's clinical outcome is often defined as the overall survival, progression-free survival, or disease-free survival rate. Prognostic biomarkers are useful because they help anticipate different outcomes for different groups of patients (Califf, 2018; Šutić et al., 2021).While prognostic biomarkers can provide insights into potential patient outcomes, they do not gauge a patient's response to treatment in the same way predictive biomarkers can. The routinely used biomarkers for prognosis assessment include the T, N and M stage of cancer, and the performance status of a

patient. Molecular prognostic biomarkers include proteins, genes, mRNA, and miRNA. For example, Der et al. (2014) identified a 15-gene mRNA expression signature in early NSCLC tumour tissue that was able to separate patients into high and low-risk subgroups based on their 5-year overall survival. Roepman et al. (2008) developed a 72-gene mRNA expression signature that could differentiate between patient tumours with high or low risk of disease recurrence after tumour resection; Kratz et al. (2012) developed a 14-gene mRNA expression signature that was able to identify early stage non-squamous NSCLC patients at high risk of mortality after surgical resection. The prognostic power of these gene signatures is promising, but further validation is essential before they can be translated into clinical tools. Some studies have singled out interesting genes as potential prognostic markers, including p53 mutational status, which are associated with poor survival and increased resistance to cancer therapy in NSCLCs (Gu et al., 2016; Xu et al., 2020). The overexpression of VEGF, *TUBB3* and *Ki-67* have independently been associated with poor NSCLC prognosis (Reiman et al., 2012; Wei et al., 2018; Zhan et al., 2009). Similar to the gene expression signatures, these genetic prognostic biomarkers have yet to find clinical application, as substantial research is still required for their validation.

1.3 Ethnic disparities in cancer genetic traits, cancer research, and cancer outcomes

1.3.1 Ethnic differences in cancer genetic traits

While the majority of alleles are common across population groups from diverse geographical ancestries, subtle allele frequency variations across multiple genetic loci often distinguish these groups (Rosenberg et al., 2002). It is therefore possible that the slight differences in the genomic background and cancer-causing mutations of different populations may contribute to disparities in cancer susceptibility, development, and outcome. Various studies have investigated the genomic landscape of cancer (at a DNA, RNA, and epigenetic level) and how this might vary based on ethnicity/genetic ancestry. Yuan et al. (2018) assessed the influence of genetic ancestry on genomic alterations across ten cancer types with the use of genetic data from African ancestry (AA) and European ancestry (EA) patients. The variations they noted between the groups across the cancers included: distinct alteration frequencies in three focal somatic copy number changes; disparities in alteration rates of five recurrently mutated genes (RMGs), a notably higher frequency of *TP53* mutations in AAs, and a reduced mutation rate in the *PI3K* pathway for AAs. Ethnicity-based differences in the frequency of chromosomal instability and alterations of recurrently mutated genes were also observed at a cancer-specific level. Another pan-cancer study identified increased microRNA expression levels in cancers from AA vs. EA patients (Lara et al., 2020). At a cancer-specific level, investigations into genetic variations between AA and EA patients have been conducted for several of the more common cancers, such as breast, lung, and prostate cancer (Araujo and

Carbone, 2017; Bollig-Fischer et al., 2015; Keenan et al., 2015). These studies support the growing body of evidence that there may be some genetic differences between cancers from patients of AA and EA. In the following section, we delve deeper into intriguing studies on ethnicity-associated variations in NSCLC genetics.

1.3.1.1 NSCLC genomics: insights into ethnic differences

Studies indicating a reduced incidence of documented oncogenic drivers in AA compared to EA NSCLC patients (Steuer et al., 2016; Costa et al., 2021) have spurred further investigation into the consistency of these differences and their potential influence on lung cancer outcome disparities across ethnicities. Lusk *et al.* (2019) used a defined Sequenom Mass Array system, tissue staining (for detection of *ALK* fusion) and real-time PCR (for copy number analysis of *FGFR1*) to test for the presence of known driver mutations in NSCLC tissue from 193 AA patients. 34.2% (66/193) of the AA's tumours in this study harboured known pathogenic mutations. Another study that assessed the presence of classic driver mutations in 260 AAs NSCLC patients found these mutations in only 23.5% of the tumours (Araujo et al., 2015a). In both of these studies, the observed frequency of known driver mutations detected in the AA group was lower than reported in EA NSCLC patients. Bollig-Fischer *et al.* (2015) studied 335 EA and 137 AA Americans afflicted with NSCLC and found that known driver mutations were carried in 32% of African Americans and 41% of European Americans. Despite these differences in overall driver mutation burden between the two groups, the frequency of *EGFR*, *KRAS* and *PIK3CA* driver mutations was similar in AA and EA patients (Araujo et al., 2015a) (Bollig-Fischer et al., 2015).

Conversely, findings from certain studies indicate that there might not be a marked discrepancy in the occurrence of mutations in oncogenic driver genes between EA and AA tumour specimens. Araujo, Timmers, *et al.* (2015) tested for NSCLC-associated driver mutations in AAs using a custom panel of 81 NSCLC-related genes. The driver mutation frequency among this NSCLC AA group was compared with that of EAs in the TCGA database, revealing no significant differences. Additionally, Campbell *et al.* (2017) analysed 504 cancer genes in 245 AA and 264 EA tumours, and found no significant difference in the overall frequency of gene mutations and copy number changes between the two groups.

Utilising gene panels to compare the frequency of known driver mutations associated with NSCLC in AA vs. EA-derived tumours has sometimes unveiled a lower prevalence of driver mutations in the AA group, while other times no significant difference was reported. Although much of the cancer-related somatic driver mutation landscape is shared between AA and EA tumours, certain studies have uncovered differences in potentially important oncogenes and tumour suppressors. For

example, whole exome sequencing has revealed a significantly greater frequency of *STK11* mutations in AA NSCLC tumours (Arauz et al., 2020). *SKT11* inactivation through mutations has been shown to be a major driver of immune escape and resistance to anti-PD-L1 therapies in LUADs (Skoulidis et al., 2018). The higher prevalence of *STK11* mutations in AA compared to EA-derived tumours could influence disparities in cancer development and response to treatment. Furthermore, utilising a specific gene array to study known NSCLC-related driver mutations in AA patients, Lusk *et al.* (2019) only found mutations in 66 out of 193 cases. After array analysis, whole exome sequencing on cases without known driver mutations revealed that 50% had nonsynonymous mutations in driver genes, predicted to be likely damaging and undetectable using a defined array. Furthermore, by whole exome sequencing, the study found 88 genes significantly mutated among the AA NSCLC patients, 85 of which have not been previously identified as cancer driver genes. These findings indicate the presence of novel oncogenic mutations in driver genes among AA NSCLC patients, as well as novel potential oncogenic driver genes, that were previously unidentified in research focused mainly on EA cohorts.

Several factors might explain the observed disparities in cancer genomics, for example: the smaller sample size of AA participants in many studies, and potential nuances in the oncogenic driver landscape of AA patients that aren't adequately captured in current NSCLC gene lists and arrays predominantly based on EA patients. Comprehensive whole-genome sequencing on a larger cohort of AA NSCLC patients could unveil previously unidentified non-synonymous mutations in known oncogenic driver genes and even reveal new driver genes. Over time, such endeavours will enrich the representation of AA derived data in cancer genetics databases and enhance our comprehension of cancer genetics across various ethnic backgrounds.

1.3.1.2 NSCLC transcriptomics: insights into ethnic differences

In a seminal study, Deveaux *et al.* (2021) used an Affymetrix Human Clarion D Array to assess transcriptome-wide differential expression and differential splicing of genes in lung squamous cell carcinoma (LUSC) tissue from 21 AA and 20 EA patients. The study identified 4829 differentially spliced genes (DSGs), 267 differentially expressed genes (DEGs), and 208 genes that were differentially expressed and differentially spliced between the AA and EA groups. These ethnicity related DSGs and DEGs were validated using The Cancer Genome Atlas (TCGA) database. The study discovered and validated that *CRADD*, *LYRM1* and *OAS2* splice variants were much more prevalent in AA vs. EA patients. In addition, some DSGs previously identified as cancer-related were found to be differentially spliced between the groups. Among others, these included *MET*, *PTEN*, and *BCL2*. In total, 11% of the validated DEGs and 18% of the DSGs have been previously implicated in cancer. The researchers found that 355 of the race-related splicing events, and 18 of the race-related

differentially expressed genes were potentially associated with overall survival in lung squamous cell carcinoma patients.

In another study, Mitchell *et al.* (2017) investigated mRNA and miRNA expression differences in 22 AA vs. 19 EA NSCLC patient tumour tissues. Differential expression of mRNA and miRNA was independently assessed between tumour cells and adjacent normal cells for both AA and EA samples. While there was considerable overlap in DEGs between the two groups, 637/3500 and 1844/4797 coding genes were distinctly and significantly differentially expressed (SDE) in AA and EA samples alone, respectively. DEGs from tumours in the AA population were enriched in stem cell and invasion pathways, while those from the EA population were enriched in proliferation, cell cycle, and mitosis pathways. The study also validated the differential expression of a few genes previously shown to be differentially expressed between AA and EA tumours in other cancers. Connectivity Map software (Lamb et al., 2006) was used to predict drug response in AA and EA patients by assessing their gene expression profiles. An inverse relationship to drug response between AA and EA patients was predicted for 53 drugs. For these 53 drugs, AA individuals were predicted to be resistant, while EA individuals were predicted to be sensitive. With respect to the non-coding region of the transcriptome, 7 miRNAs were differentially expressed in the AA group only, while 10 miRNAs were differentially expressed in the EA group only. The researchers hypothesize that the observed race-specific mRNA expression may be driven, in part, by race-specific miRNA expression.

From surveying the available literature, these (Deveaux et al., 2021; Mitchell et al., 2017) were the only studies which primarily focused on comparing NSCLC global gene expression (alongside alternative splicing and miRNA expression) between AA and EA patients. These studies point toward ethnicity-based differences in transcriptomic traits like the mRNA expression, alternative splicing, and miRNA expression of NSCLC tumours. Studies looking into these types of transcriptomic variations in NSCLCs between ethnic groups, with greater sample sizes and statistical power, will provide valuable insight into population-level variation in cancer genetics, which could prove clinically relevant.

1.3.2 Disparities in representation in research

Cancer genetic research has proved to be a valuable tool in illuminating our molecular understanding of cancers and our ability to better treat cancer in a patient-specific way. An assessment of the cancer genetic research databases and studies reveals a bias toward European ancestry (EA) samples and an underrepresentation of African Ancestry (AA) samples and other ancestral (ethnic) groups. Large genomic databases like TCGA, the Genome-Wide Association Study Catalogue (GWAC), and the database of Genotypes and Phenotypes (DGP) are used to identify

relevant cancer-related gene mutations. A look at the research studies that comprise the GWAC and DGP databases reveals an underrepresentation of genomic cancer studies involving individuals of AA and other non-EA ethnic groups (Landry et al., 2018). A similar pattern can be seen in the TCGA database, where Spratt et al. (2016) investigated the genomic data of 10 common cancers and found that 77%, 12% and 3% of the samples were from European, African, and Asian ancestry populations, respectively. Within the assessed cancer studies from the TCGA database, the absolute size of African ancestry and Asian ancestry populations is often inadequate for use in detecting even relatively common somatic mutations specific to those underrepresented groups (Spratt et al., 2016). The inadequate representation of the diversity that exists globally in cancer genetic research means that our understanding of cancer genetics is largely biased toward cancer genetic information from samples of European Ancestry. Improving the representation of African ancestry, Asian ancestry, and other groups in cancer genetics studies will ensure that more individuals can benefit from advancements in cancer diagnosis, prognosis, and treatment.

1.3.3 Disparities in cancer outcomes

For many cancer types, studies have identified interesting differences in incidence, aggressiveness, and mortality of cancers between different ethnic groups (Zavala et al., 2021). These result from differences in a combination of factors such as access to healthcare, socioeconomic status, socioenvironmental conditions, behavioural factors, and biological factors.

Individuals of EA have been shown to have the highest breast cancer incidence (136.3/100000), followed by AA(128.3/100000), then Asian and pacific islander ancestry individuals(102.9/100000) and Hispanic individuals (98.5/100000) (“Cancer of the Breast (Female) - Cancer Stat Facts,” n.d.). Regarding the aggressiveness of cancer, women of AA are twice as likely to be diagnosed with aggressive triple-negative breast cancer than women of EA (Newman and Kaljee, 2017). In the case of prostate cancer, a study in the USA revealed that AA men have a 73% higher incidence and a two times higher mortality rate than EA men (Siegel et al., 2022). For lung cancer, amongst males, those of AA have the highest incidence (68.3/100,000), followed by males of EA and other ancestral groups (61.5-31.8/100,000). Among females, the highest incidence is among those of EA (52.7/100,000), followed by those of AA and other ancestral groups (44-22.3/100,000)(“Cancer of the Lung and Bronchus - Cancer Stat Facts,” n.d.). After an analysis of almost 5,000 lung cancer cases in the USA, Stram et al. (2019) found that among patients with lower smoking intensities (10 cigarettes per day), Native Hawaiians and African Americans were at higher risk of smoking-related lung cancer than White Americans, Japanese Americans and Latinos. Research has also found ethnicity-based disparities in cancer incidence and outcome for colorectal, pancreatic, gastric, leukaemia, and liver cancer, among many others (Zavala et al., 2021).

More broadly, A pan-cancer study (Lara et al., 2020) looking at slightly more than 2 million patients from the Surveillance Epidemiology and End Results Database (SEER) as well as the National Cancer Database (NCDB) found that for 21 cancers, African American patients had an increased risk of death when compared to European American patients. This difference in risk of death between the ethnic groups persisted after controlling for environmental factors such as socio-economic factors, access to health care, and insurance status. The result supports the growing body of evidence that alongside differences in environmental factors, biological differences in cancers between ethnicities may significantly contribute to the sometimes-observed disparate aggressiveness and outcomes of cancer between different ethnic groups.

The interplay between environmental and biological factors and how these may contribute to cancer health disparities is a complex multi-trait question. Work looking into both environmental and biological factors contributing to these differences will improve our overall, as well as our population-specific, understanding of lung cancer.

1.4 Importance of representative genetic research/ consequences of non-rep cancer genetics research

Advancements in cancer genetic research have improved our molecular understanding of cancer. These advancements provide clinically relevant insight to improve diagnosis, prognosis, treatment development and treatment recommendation. However, due to social, logistical, and historical reasons, most of this cancer genetic research has been done on patients of European ancestry. The differences in genetic architecture between groups of different ancestry, because of natural selection acting on independent alterations to their genome (resulting from processes like genetic drift and bottleneck effects) over generations, means that the association or the strength of an association of a genetic variant to a phenotype may not always be transferable between populations (Sirugo et al., 2019). Therefore, polygenic risk scores, diagnostic gene signatures, and genetic predictors of drug response developed from studies in one ethnic group may not be as applicable to other ethnic groups (Wojcik et al., 2019). This limits the transferability of cancer genetic insights between different ethnic groups.

For the knowledge and clinical benefits of cancer genetic research to benefit the diverse groups of people globally, more genetic research involving patients of African ancestry, Asian ancestry and other underrepresented groups is needed. Understanding the genetic differences and commonalities between these different ancestral groups increases the power to identify novel disease-associated variants on a global and population-specific level and develop more representative and applicable clinical tools for personalised cancer treatment. Such representative cancer genetic research may

also help understand and mitigate the disparity in cancer incidence and outcomes between different ancestral groups.

1.5 Study aims and objectives

Comparative research outlining similarities and differences in cancer associated genetics between different ethnic groups is an important first step in developing a more representative molecular understanding of cancer. Toward this end, this study aimed to identify what clinically relevant genetic differences exist between LUAD cells derived from patients of African and European ancestry.

More specifically, the objectives of the study were to identify whether any differences in gene expression, copy number alteration and mutations exist between LUAD cell samples derived from patients of African and European ancestry. We hypothesized that there will be differences observed in the expression of genes between the two groups. We also hypothesized that there will be a greater frequency of copy number alterations, and mutations among the AA group than the EA group, as AA individuals have been observed to have a greater diversity of single nucleotide polymorphisms and genetic variation in their genomes that EA individuals (Hughes et al., 2008). After comparing gene expression, copy number alterations, and mutations between the two groups we identified what molecular processes the observed genetic differences were associated with, and what clinical relevance they may have. All genetic and clinical data used in this study was derived from the TCGA database.

2. Data Source and Selection

2.1 Introduction

Observational studies that look to associate genetic variations with certain phenotypic traits have to account for differences in environmental and clinical characteristics between the treatment and control groups, in order to be confident in the identified associations. Methods such as regression analysis and propensity score matching (PSM) are popular methods for achieving such balance (Amoah et al., 2020).

This study aims to identify whether any interesting genetic differences exist between LUAD tumours from African ancestry (AA) and European ancestry (EA) patients. It is therefore imperative to minimize clinical differences between the two sample groups, in order to enhance the reliability of our findings.

In this chapter, we explain where and how we obtained the samples for this study, and how PSM was used to balance out the distribution of clinical variables between the two groups.

2.2 The Cancer Genome Atlas (TCGA)

TCGA is a public project that aims to facilitate the discovery and validation of cancer-associated genetic traits (Tomczak et al., 2015). Toward this end, TCGA brings together several cooperating centres that work to enrol eligible cancer patients, collect biological specimens (blood and tissue) and clinical information, process samples and validate quality; and obtain molecular analytes for genomic characterization and high throughput sequencing. The generated data is then available to the research community and Genome Data Analysis Centres. Genome Data Analysis Centres provide tools for the processing, analysis, and visualisation of the data. For the research community, the data is also publicly accessible through free access databases like the National Cancer Institute's Genomic Data Commons portal and the FireBrowse data portal ("FireBrowse," n.d.; "GDC," n.d.; Heath et al., 2021). This allows the research community to easily access the data and implement novel analyses that can contribute to cancer genetic research.

The TCGA has data from over 10,000 cancer patients spanning 33 cancer types. Alongside the clinical information collected for each patient, DNA, RNA, and proteins are isolated for genomic, transcriptomic, epigenetic and proteomic analysis using various 'omics' platforms (Tomczak et al., 2015).

2.2.1 Downloading and initial filtering of data

Data from the TCGA-LUAD study contained clinical and molecular information for 566 LUAD patients (Collisson et al., 2014). Subsets of the TCGA-LUAD study samples and their clinical and molecular information were used for all the analyses performed in this study. Approval for this study was obtained from the Human Research Ethics Committee of the University of Cape Town (HREC reference number: 523/2022).

In this chapter, clinical data for samples of the TCGA-LUAD study (version 2016_01_28) was downloaded from the Broad Institute of Harvard and MIT's FireBrowse data portal (http://firebrowse.org). The data was then handled using R Studio (version 4.2.1).

Samples of the downloaded TCGA-LUAD data were filtered in 4 steps. Firstly, only samples for which patients had self-reported their race to be “black” or “white” were kept, leaving 393 white samples and 53 black samples. This study refers to self-reported white and black patient samples as European ancestry (EA) and African ancestry (AA), respectively. Secondly, to ensure that only samples for which gene expression can be compared are used, only samples that had read count data from gene expression analysis were kept, leaving 389 EA and 52 AA samples. Thirdly, only samples with clinical information on age, gender, tobacco smoking history, T pathologic stage, N pathologic stage and M pathologic stage were kept, leaving 371 EA and 49 AA samples. Finally, only the samples obtained after PSM (which is outlined in the subsequent section) were retained.

2.3 Propensity score matching (PSM)

2.3.1 Defining and understanding PSM

In any experiment, the goal is to ascertain if a specific treatment causes a particular outcome by examining if there's a significant difference in results between a treatment and a control group. To confidently attribute the observed difference in outcome to the treatment, it's essential to minimise the possibility that other differences between the treatment and control groups are influencing the results. In a randomised control trial, the assignment of participants to the treatment and control group is random. This randomization balances the distribution of clinical traits between the two groups, minimizing confounding factors in the experiment. With observational studies, the assignment of a sample to a treatment or control group is not random but is instead based on a certain feature (clinical characteristic) of the sample. This non-random treatment assignment increases the likelihood that other associated features will differ between the two groups and, therefore, have a confounding effect on the observed outcome. With observational studies, minimising the confounding effect of certain features on the outcome is necessary.

A popular method for minimizing the effect of confounding variables in observational studies is PSM. Rosenbaum and Rubin (1983) define the propensity score as the probability of treatment assignment, conditional on the selected baseline characteristics. The closer the propensity scores of the two samples are, the more similar their baseline characteristics. PSM involves forming matched sets of treatment and control samples with similar propensity scores. By ensuring the propensity scores of samples in the treatment group are similar to those in the control group, PSM minimizes the confounding effect of the considered clinical characteristics between the two groups (Austin, 2011). During PSM, each treatment group sample is matched to a specified number of samples (one or more) from the control group with a similar propensity score. Greedy and optimal matching are the two main types of PSM techniques. With greedy matching, a treated sample is selected randomly and matched with the specified number of control samples with a propensity score closest to it. This is done until all the treated samples have been matched with control samples or until no more control samples are available for matching. This matching technique is called greedy because at each step, once a sample of the treatment group is matched with samples of the control group, the matched control group samples cannot be matched to any subsequent treatment group sample even if they have closer propensity scores. Optimal matching, on the other hand, performs matching between the treatment and control group to minimize the sum of differences in the propensity score between the matched treatment and control samples. There is not much difference in the ability of these two methods to minimize the difference in clinical characteristics between a treatment and control group by matching (Gu and Rosenbaum, 1993).

To minimise the cofounding effect of clinical variable in this study, PSM was selected over other propensity score-based techniques and traditional logistic regression analysis. Firstly, PSM was selected because propensity score-based methods have been shown to produce similar and sometimes better correction of confounding variables than traditional logistic regression analysis(Amoah et al., 2020). Secondly, PSM in particular was selected over other propensity-score based methods due to its proven effectiveness in balancing the confounding of multiple clinical traits between groups and the relative simplicity of applying a PSM based approach (Reeve et al., 2008). Despite the relative simplicity of applying PSM, a downside to using this technique is the reduction in the sample size. When dealing with small sample sizes, propensity score-based methods such as propensity weighting or covariate adjustment using the propensity score are viable methods to consider, as they do not lead to sample size reduction(ROSENBAUM and RUBIN, 1983; Weeks et al., 2015). Despite the reduction in sample size associated with PSM, it was chosen for this study due to its proven effectiveness in balancing clinical characteristics in observational studies, and the relative ease implementation.

2.3.2 Selecting and executing a PSM technique

To begin with, in order to minimise the difference in some important clinical variables between the African ancestry (AA) and European ancestry (EA) samples, the extent to which some clinical variables differ between the two groups was determined. Six clinical variables were considered, namely age, gender, and tobacco smoking history of patients, as well as the T pathologic stage, N pathologic stage and M pathologic stage of their tumours. The T, N and M pathologic stages are categorical variables describing the anatomy of a tumour (American Joint Committee on Cancer, 2002). Finally, the different PSM techniques were compared so that the one that provided the best balancing of the considered clinical variables between the two groups could be selected.

To begin with, tests for any statistically supported difference between the AA and EA groups of the unmatched cohort in any of the 6 considered clinical variables were performed. The resulting p-values supported the null hypothesis that within the unmatched cohort, there is no difference in sex, pathologic T stage, pathologic N stage and pathologic M stage (p-value >0.05) between the AA and EA groups. However, for age and tobacco smoking history, the null hypothesis was rejected (p-value< 0.05), indicating a statistically supported difference in these clinical variables between the two groups (Table 2.1).

To neutralise the observed difference in age and tobacco smoking history between the two groups, six PSM techniques were evaluated to determine which would be most effective. More specifically, 1:2, 1:3 and 1:4 optimal PSM were performed, each matching one AA patient with two, three or four EA patients respectively. These matching techniques were each performed in two ways, either with age and tobacco smoking history being used to calculate propensity scores, or with all the six clinical variables (age, sex, tobacco smoking history, pathologic T stage, pathologic N stage, and pathologic M stage) being used to calculate propensity scores. When matching AA samples to EA samples, PSM considered age to be a discrete variable, and all the other five clinical characteristics (sex, tobacco smoking history, pathologic T stage, pathologic N stage, and pathologic M stage) to be categorical variables as they appear in Table 2.1.

Of these 6 PSM techniques, 1:3 optimal PSM with 6 clinical variables provided the best balancing of clinical variables between the EA and AA groups. This PSM technique produced AA and EA groups with less difference in age and tobacco smoking history (based on the p-value) than all the other matched cohorts, as well as the unmatched cohort (Supplementary tables 1 and 2). To further assess the effectiveness of different propensity matching techniques in balancing clinical variables between the two groups, the mean p-values from statistical tests examining differences in clinical variables between the AA and EA groups for each technique were computed. The bar chart in Figure

2.1 shows that the 1:3 PSM technique with 6 clinical variables had the highest average p-value when comparing all the techniques. Based on this, it was inferred that this matching technique best balances clinical variables between the two groups.

In summary, of the six clinical characteristics we evaluated, only two —age and tobacco smoking history—met the criteria for confounding between the two groups, evidenced by a p-value < 0.05 after testing for differences in the clinical trait between the two groups. The best PSM method for minimising the confounding effect of age and tobacco smoking history between the two groups was determined by comparing the statistical differences in each of the six considered clinical variables between the two groups after PSM was performed. Among the six PSM methods compared in this study, the one deemed most effective resulted in an increase in the p-values of age and tobacco smoking history to greater than 0.05 and provided the highest p-value mean of the six considered clinical variables after PSM (Figure 2.1, Supplementary Table 1, Supplementary Table 2). Based on these criteria for minimising the likelihood of confounding between the two groups, 1:3 propensity score matching using all the six considered clinical variables was selected as the best matching technique. We, therefore, decided to use 1:3 propensity-matched samples (matched using 6 clinical variables) for all downstream analyses that followed.

Table 2.1. Clinical traits of samples before and after 1:3 PSM. The displayed p-values are from statistical tests performed to determine whether the null hypothesis that there is no difference in the distribution of a particular variable between the two groups is supported. For age, the p-value was determined using the Wilcoxon Rank Sum test. For all the other variables the p-values were determined using the Chi-squared test.

	Unmatched Cohort			*1:3 Matched cohort*		
	AFR Ancestry	**EUR Ancestry**	**P-value**	**AFR Ancestry**	**EUR Ancestry**	**P-value**
Patients	49 (11.7%)	371 (88.3)		49(25%)	147(75%)	
Age (mean)	60	65.8	0.0003	60	60.2	0.93
Sex			0.98			1
Male	21(42.9%)	164(44.2%)		21(42.9%)	62(42.2%)	
Female	28(57.1%)	207(55.8%)		28(57.1%)	85(57.8%)	
Pathologic T stage			0.613			0.95
T1	22(44.9%)	135(36.4%)		22(44.9%)	60(40.8%)	
T2	20(40.8%)	191(51.5%)		20(40.8%)	66(44.9%)	
T3	6(12.2%)	33(8.9%)		6(12.2%)	17(11.6%)	
T4	1(2%)	10(2.7%)		1(2%)	4(2.7%)	
TX	0	2(0.5%)		0	0	
Tobacco smoking history			0.016			0.72
1	3(6.1%)	60(16.17%)		3(6.1%)	14(9.5%)	
2	17(34.7%)	83(22.4%)		17(34.7%)	54(36.7%)	
3	10(20.4%)	100(27%)		10(20.4%)	25(17%)	
4	17(34.7%)	126(34%)		17(34.7%)	52(35.4%)	
5	2(4.1%)	2(0.5%)		2(4.1%)	2(1.4%)	
Pathologic N stage			0.63			0.98
N0	29(59.2%)	253(68.2%)		29(59.2%)	84(57.1%)	
N1	12(24.5%)	59(15.9%)		12(24.5%)	35(23.8%)	
N2	7(14.3%)	50(13.5%)		7(14.3%)	25(17%)	
N3	0	1(0.3%)		0	0	
NX	1(2%)	8(2.2%)		1(2%)	3(2%)	
Pathologic M stage			0.178			0.71
M0	25(51%)	241(65%)		25(51%)	82(55.8)	
M1	2(4.1%)	16(4.3%)		2(4.1%)	9(6.1%)	
MX	21(42.9%)	112(30.2%)		21(42.9%)	55(37.4%)	
Unknown	1(2%)	2(0.5%)		1(2%)	1(0.68%)	

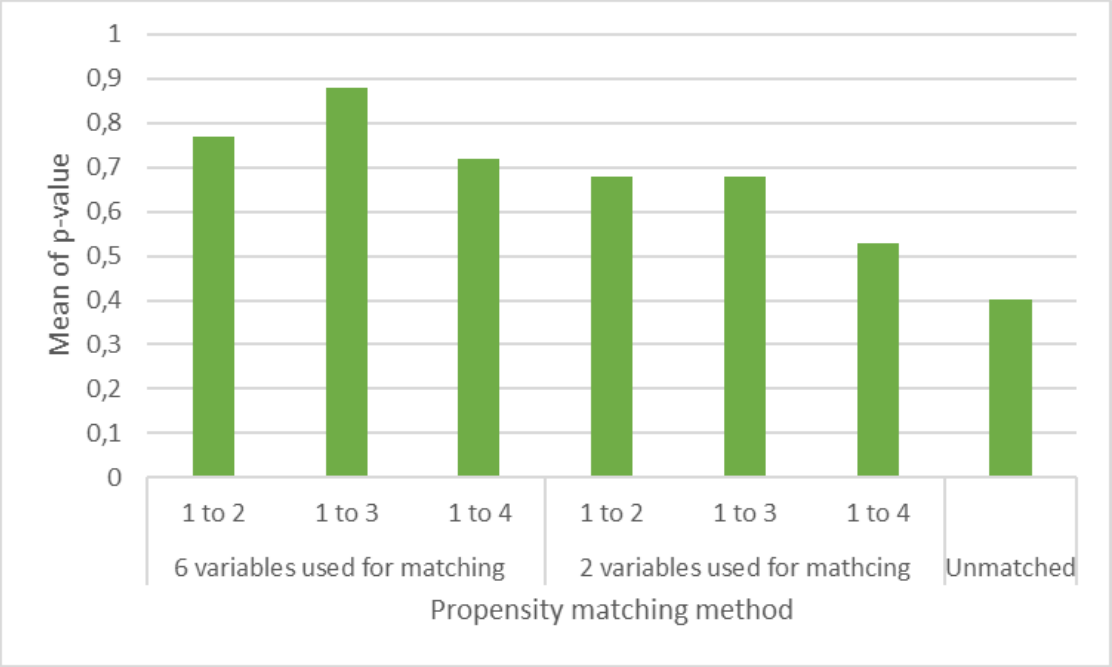


Figure 2.1. Mean of p-values from statistical tests comparing the difference in clinical variables between EA and AA groups for each PSM technique.

2.4 Discussion

In this section, we detail the process of downloading and filtering the samples used for all subsequent analyses in this study. The main component of this filtering of samples was PSM which aimed to minimize clinical differences between the self-reported AA and EA groups. A comparison of different PSM techniques revealed that 1:3 optimal PSM, which considered 6 clinical variables (age, gender, tobacco smoking history, and the tumours T Stage, N stage, and M stage), provided the best balancing of clinical variables between the two groups.

Following the application of the 1:3 PSM, 147 EA and 49 AA samples were selected, with no statistically significant differences observed in the six evaluated clinical characteristics. Although it is impossible to account for all the clinical and environmental factors that may differ between the two groups, we felt that ensuring a balanced distribution of the selected 6 variables would help minimize important biological and tumour differences between them. This is because patient age and tobacco smoking history are primary contributors to lung cancer incidence and mortality , and the T N and M stages of cancer define the stage and aggressiveness of a cancer (American Joint Committee on Cancer, 2002). These characteristics are therefore useful to consider when balancing differences in patient characteristics and tumour biology between samples. By limiting the difference in important clinical characteristics between the two groups, we maximised the likelihood that any differences in

genetic traits that may be observed in downstream analyses, were indeed a result of differences in ethnicity between the two groups.

Some limitations exist in selecting the 49 AA and 147 EA matched samples we used for downstream analysis. Firstly, the prevalence of genetic admixture within populations means that self-reported race sometimes doesn't align well with one's genetic ancestry. When trying to find associations between genetic ancestry and a particular phenotype, using ancestry informative genetic markers is a more reliable way of separating individuals by genetic ancestry than self-reported race (Mersha and Abebe, 2015). We categorised samples based on self-reported race. Employing ancestry informative genetic markers might offer a more accurate method for this categorisation and enable a more precise correlation of genetic differences with genetic ancestry. Secondly, although we managed to minimise the difference in important clinical variables between the two groups, the absence of patient environmental conditions and exposures prevented us from being able to minimise certain important environmental differences between the two groups that might contribute to differences in tumour genetics (i.e., socio-economic status, access to healthcare etc.).

3. Differential Expression and Enrichment analysis

3.1 Introduction

Gene expression analysis can provide useful information on the up and down-regulation of important molecular pathways and functions in cancer cells. Assessments of gene expression differences between samples of AA and EA have been used to better illuminate ethnicity-based differences in the molecular nature of multiple cancers such as breast, kidney, brain, prostate, and lung cancer (Deveaux et al., 2021; Krishnan et al., 2016; Mitchell et al., 2017; Ping et al., 2020; Wu et al., 2019; Yuan et al., 2020).

Research looking at differences in lung cancer genetics between tumours of AA and EA has predominantly looked at DNA-based differences and not as much at gene expression-based (RNA abundance or transcriptome) differences. Only 2 studies have primarily focused on identifying differences in gene expression between the two groups. Deveaux et al. (2021) pinpointed genes that exhibited differential expression and splicing patterns in lung squamous cell carcinoma tumours from AA and EA patients. Of the identified genes, they highlighted 853 out of 4829 differentially spliced genes (DSGs) and 29 out of 267 differentially expressed genes (DEGs) as likely contributors to cancer progression. Mitchell et al. (2017) independently examined differential gene expression between NSCLC tumour cells and adjacent normal cells from samples of AA and EA. Among the DEGs associated with tumour development, in samples of AA, stem cell and invasion pathways were enriched, whereas in samples of EA proliferation, cell cycle and mitosis pathways were enriched. In addition, drug response differences between AA and EA patients were predicted based on the differences in the gene expression profiles of their tumours.

With the advancement of cancer genetics research, the utility of transcriptome profiles in characterising the molecular nature of patient tumours continues to increase. The paucity of NSCLC transcriptomic profiles from patients of AA limits the extent to which this group can benefit from the advances in the clinical handling of NSCLC stemming from NSCLC transcriptomic research. Developing an understanding of the subtle differences in NSCLC gene expression between samples of AA and EA is useful in facilitating “diversity-aware” genetic research (Sirugo et al., 2019; Wojcik et al., 2019). In this chapter, we explored the variations in gene expression between LUAD samples from AA and EA individuals. Through enrichment analysis, we aimed to decipher the potential molecular and clinical ramifications of the identified differential gene expression.

3.2 Methods

3.2.1 Downloading raw count RNA-Seq data

Raw RNA-Seq count data from the TCGA-LUAD study, was downloaded from the Broad Institute of Harvard and MIT's FireBrowse data portal (http://firebrowse.org/). More specifically, the "illuminahiseq_rnaseqv2-RSEM_genes" file was downloaded. From this file, raw RNA-Seq count data from the 1:3 propensity score matched samples (49 AA and 147 EA) were isolated and used downstream for differential gene expression analysis.

3.2.2 Differential gene expression analysis

DESeq2 (version 1.36.0) (Love et al., 2014) was used in R Studio to perform differential gene expression analysis on raw RNA-Seq count data from the 49 AA and 147 EA-matched samples. DESeq2 uses negative binomial distribution to identify differential gene expression between groups.

Before differential gene expression analysis, a DESeq object was created using the DESeqDataSetFromMatrix function. A count matrix and sample table were created as input to this function. The count matrix contained RSEM raw gene counts for each sample, rounded to the nearest integer. The count matrix had genes along its rows and sample barcodes across its columns. The sample table consisted of two columns containing the sample barcode and self-reported race of each sample. Using the count matrix, sample table and specification of ethnicity as the variable that would define our experiment (EA vs. AA), the DESeqDataSetFromMatrix function was used to create a DESeq object for differential gene expression analysis. After creating the DESeq object, the EA samples were designated as the reference group for the differential gene expression analysis. Additionally, any genes with a cumulative count of fewer than 10 across all samples were excluded from the DESeq object.

Using the DESeq function, and the DESeq object as its input, DESeq2 was utilised to perform differential gene expression analysis between the AA and EA groups. From the differential gene expression analysis results, genes with an adjusted p-value < 0.05 and -1> log2-transformed fold change >1 were considered SDE. Among the list of SDE genes, NSCLC associated genes were identified with reference to the Kyoto Encyclopaedia of Genes and Genomes (KEGG) Pathways NSCLC gene list (hsa05223).

3.2.3 Enrichment analysis

Enrichment analysis was performed to identify biological processes and molecular pathways associated with the differential gene expression results. Two types of enrichment analysis were performed: over-representation analysis (ORA) and gene set enrichment analysis (GSEA).

3.2.3.1 Over-representation analysis (ORA)

ORA is a method employed to discern which established biological functions and processes, represented as gene sets, are disproportionately represented within a given list of genes obtained from experimental data (Boyle et al., 2004). The ClusterProfiler package (version 4.4.4) (Wu et al., 2021) was used to identify which gene sets from the KEGG pathway and the Gene Ontology Biological Process (GO-BP) databases were overrepresented (enriched) in our list of SDE genes. Cluster profiler utilises a hypergeometric test (also called a Fisher's exact test) to determine p-values of the over-represented gene sets that result from ORA. In our analysis, gene sets that displayed an adjusted p-value < 0.05 were deemed significantly enriched, indicating their over-representation.

3.2.3.2 Gene set enrichment analysis (GSEA)

GSEA (Subramanian et al., 2005) is another approach for identifying biological processes and molecular pathways that are enriched among differentially expressed genes. Unlike ORA, which seeks enrichment in a supplied list of SDE genes, GSEA analyses the entirety of the differential gene expression results to detect enrichment. This is achieved by ordering the differential gene expression analysis output by fold change or test statistic in descending order to create an ordered gene list (*L)*. Given gene set *S* (a group of genes associated with a particular function or pathway), GSEA aims to determine whether the members of *S* are randomly distributed throughout *L*, or are primarily found at the top or bottom of *L*. If a large proportion of the gene members of *S* are found at the top of *L*, then the gene set *S* will be considered enriched and upregulated. On the other hand, if a large proportion of the gene members of S are found towards the bottom of *L*, then the gene set *S* will be considered enriched and downregulated. This method allows GSEA to identify situations where multiple genes in a gene set are differentially expressed in a small but coordinated way that may be biologically meaningful.

The ClusterProfiler package (version 4.4.4) was used to perform GSEA in RStudio. With the differential gene expression results as an input, the gseGO and gseKEGG functions were used to investigate the enrichment of gene sets from the GO-BP terms and KEGG pathway database respectively. An adjusted p-value threshold of <0.05 was applied to the GSEA results.

3.3 Results

3.3.1 Differential gene expression

DESeq2 was used to perform differential gene expression analysis on LUAD cells from 49 AA and 147 EA samples, to identify differentially expressed genes between the AA and EA groups. Genes with an adjusted p-value < 0.05 and a -1 > log2-transformed fold change > 1 were considered SDE.

A total of 371 genes were identified to be SDE between the two groups (Supplementary table 3). 171 genes were significantly upregulated, and 200 genes were significantly downregulated in the AA group relative to the EA group, as seen in the volcano plot below (Fig. 3.1). The top 10 significantly downregulated and upregulated genes are listed in table 3.1.

Among the 371 SDE genes, *CDKN2A* was the only one associated with NSCLC according to the KEGG pathway NSCLC gene list. *CDKN2A* encodes a cyclin dependent kinase inhibitor (also known as p16) that acts as a tumour suppressor protein. *CDKN2A* expression was upregulated in the AA vs. EA group, with a log2-transformed fold change of 1.29 (2.45 absolute fold change) between the two groups.

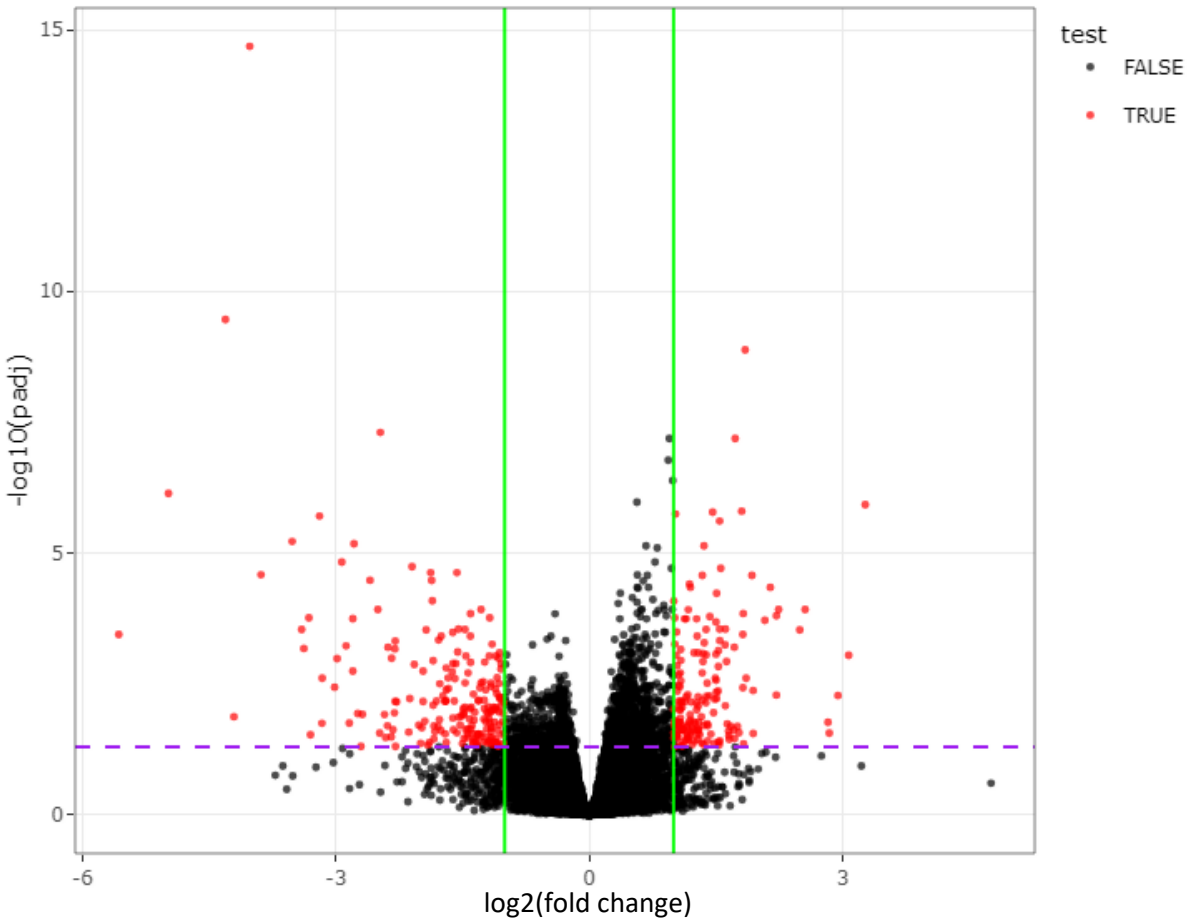


Figure 3.1. Volcano Plot showing the result of differential gene expression analysis between the AA and EA groups. The 371 SDE genes (adjusted p-value < 0.05 and a -1> log2-transformed fold change >1) are represented by red dots. In the figure, adjusted p-value is log10-transformed. The green lines represent the fold change threshold for significant differential expression, and the purple dotted line represents the p-value threshold for significant differential expression.

Table 3.1. The top 10 upregulated and downregulated SDE genes between the AA and EA groups. Differential gene expression analysis was performed with the EA group as a reference, so upregulated or downregulated genes are upregulated or downregulated in the AA group compared to the EA group.

Top upregulated genes		Top downregulated genes	
Gene name	**Log2-transformed fold change**	**Gene name**	**Log2-transformed fold change**
LBP	3.3	*LCN15*	-5.5
FMR1NB	3.1	*DLK1*	-5.0
RHOXF2B	2.9	*CALB1*	-4.3
TUBA3C, TUBA3D	2.8	*CPN1*	-4.2
TFF2	2.8	*CDH17*	-4
NLRP11	2.6	*NEUROD1*	-3.9
MUCL1	2.5	*LINC00162*	-3.5
CLPSL2	2.2	*CALCB*	-3.4
UTS2R	2.2	*F2*	-3.4
C6orf68	2.2	*TM4SF20*	-3.3

3.3.2 Enrichment analysis

3.3.2.1 Over-representation analysis (ORA)

ORA was performed to evaluate the biological processes and molecular pathways, in the form of gene sets, which were enriched (overrepresented) in the list of genes that were SDE between the AA and EA tumours.

ORA analysis against the KEGG pathway database revealed that 9 KEGG pathway gene sets were enriched in our list of SDE genes (supplementary table 4). These enriched KEGG pathways and the number of SDE genes that were a part of them are listed in Figure 3.2a. Five of these KEGG pathways are involved in the metabolism of a various compounds. Of these five, four (drug metabolism – cytochrome p450; metabolism of xenobiotics by cytochrome p450; drug metabolism – other enzymes; and retinol metabolism) have to do with the metabolism of xenobiotics such as drugs and vitamins, while the other is involved in the metabolism of the locally produced porphyrin macromolecule (an important precursor for the formation of hemoproteins like haemoglobin and cytochrome P450) (Phillips, 2019). Notably, the steroid biosynthesis KEGG pathway was also enriched. Most of the SDE genes, which form a part of the top five most enriched KEGG pathways, are shared between at least two of these enriched KEGG pathways (Fig. 3.2b). Indicating that there is a lot of overlap in the genes that made up the different enriched KEGG pathways. The top five most enriched KEGG pathways contain a mixture of upregulated and downregulated genes.

ORA of GO-BP terms uncovered 48 GO-BP terms that were enriched in our list of SDE genes (supplementary table 5). The ten most significantly enriched terms and the number of SDE genes that were a part of them, are listed in Figure 3.3a. The top ten enriched biological processes fell into

four distinct categories: firstly, three related to the response to xenobiotics; secondly, three pertaining to hormone metabolism and regulation; thirdly, two associated with cell-to-cell adhesion; and fourthly, two linked to nervous system activities. When looking closer at the SDE genes that make up the five most enriched GO-BP terms, we see that the three biological processes involved in the response to xenobiotics share many genes between them (Fig. 3.3b). In figure 3.3b, we also see that a combination of up and down-regulated genes make up the top five enriched GO-BP terms.

ORA on the list of 371 genes that were SDE between our AA and EA groups, revealed some pathways and biological processes through which these genes may result in differences in tumour biology between the two groups. Gene sets that relate to the response to xenobiotics and those relating to hormone metabolism and regulation were enriched in both KEGG Pathway and GO-BP ORA, making these enriched gene sets interesting candidates for further investigation.

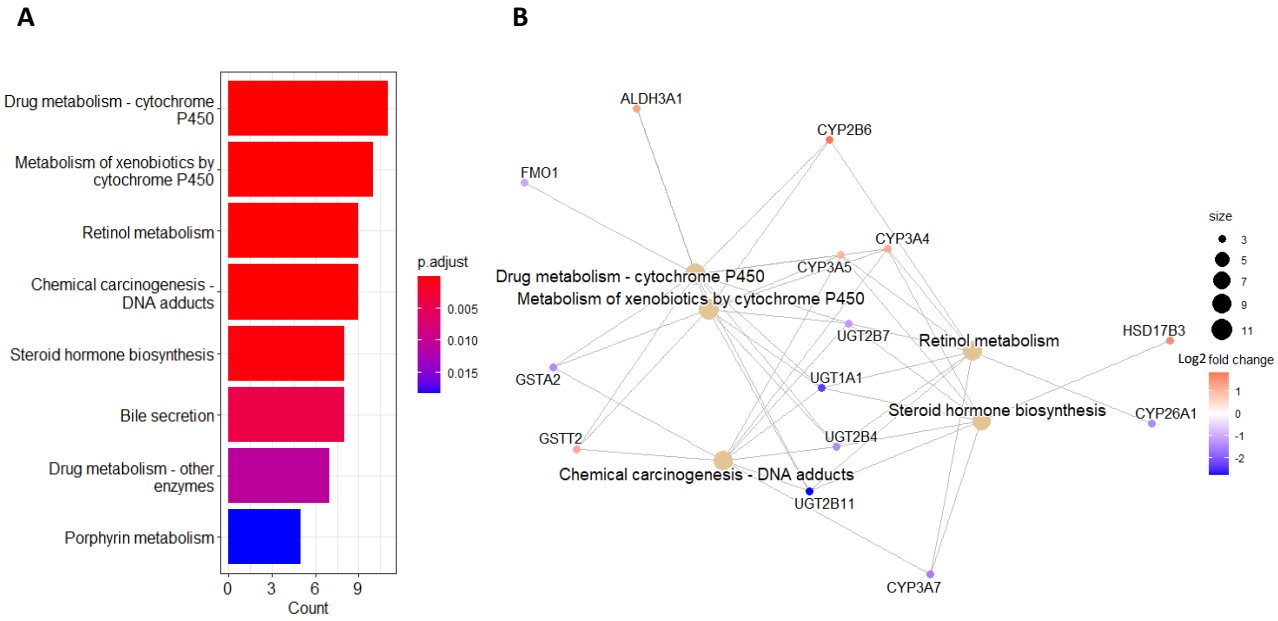


Figure 3.2. **Result from ORA using the KEGG Pathway database**. (a) All 9 enriched KEGG Pathways. The count value on the x axis is the number of SDE genes that form part of a particular gene set. (b) A cnet plot showing the top 5 enriched KEGG pathways, the SDE genes that are a part of them, and how some of these genes are shared by 2 or more gene sets.

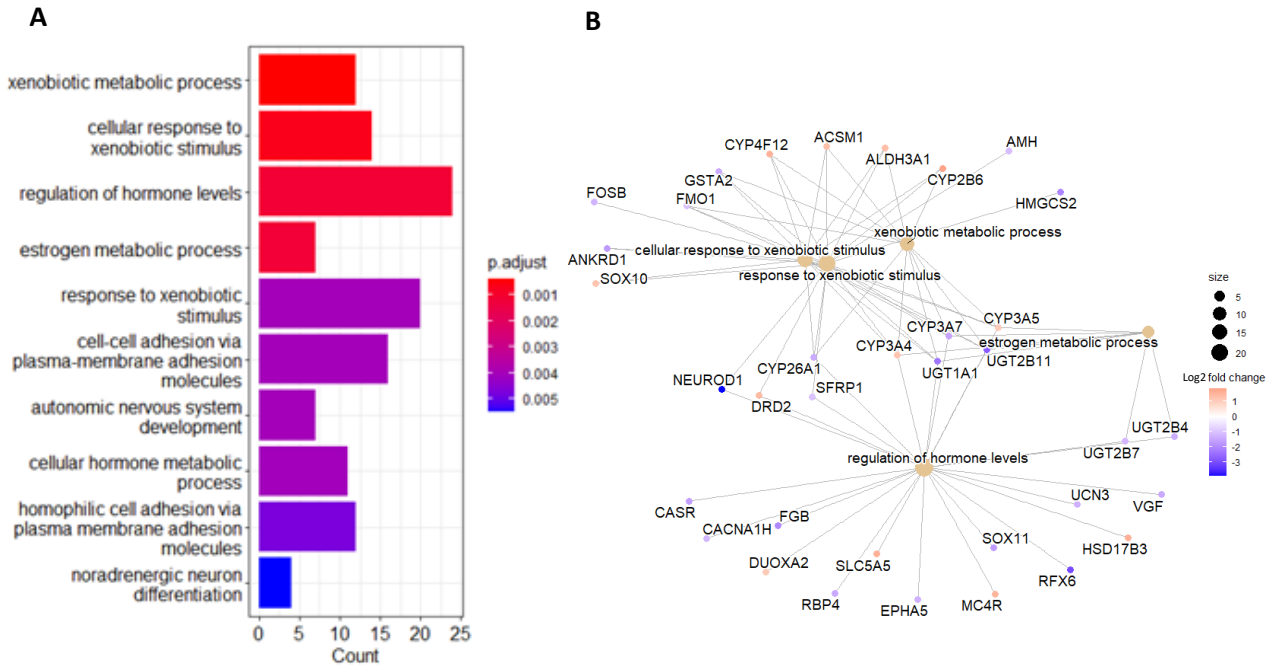


Figure 3.3. Result from ORA using the GO-BP terms. (a) The top 10 most enriched processes (out of 48 enriched processes). The count value on the x axis is the number of SDE genes that form part of a particular gene set. (b) A cnet plot showing the top 5 enriched GO-BP's, the SDE genes that are a part of them, and how some of these genes are shared by 2 or more gene sets

3.3.2.2 Gene Set Enrichment Analysis (GSEA)

GSEA was performed to identify biological processes and molecular pathways that were enriched in the differential gene expression results we obtained. Unlike ORA, GSEA can detect instances where many genes in a gene set exhibit subtle yet coordinated expression shifts, potentially signifying biologically relevant changes. This allows for the prediction of biological alterations stemming from less pronounced gene expression variations.

GSEA against the KEGG Pathways database identified 84 enriched pathways (supplementary figure 6). Figure 3.4 shows the top enriched pathways, as well as the number (Fig. 3.4a), differential expression (Fig. 3.4b), and sharing of the differentially expressed genes that make them up (Fig. 3.4c). Among the 10 most significantly enriched pathways (Fig. 3.4a and 3.4b), three were associated with neurodegenerative conditions (Parkinson's, Huntington's disease, and lateral sclerosis). These gene sets were upregulated in AA vs. EA tumours. The clustering of these gene sets and other neurodegenerative condition gene sets like Alzheimer's and Prion disease, in Figure 3.4c's enrichment map, indicated an overlap in the genes that make up these pathways. Among this cluster of neurodegenerative condition pathways is an oxidative phosphorylation pathway, which was shown to share genes with most of the neurodegenerative condition pathways in the enrichment

map. The oxidative phosphorylation pathway ranked among the top 10 for significant enrichment (Fig. 3.4a and Fig. 3.4b). Similar to the pathways related to neurodegenerative conditions, it displayed upregulation in AA compared to EA samples (Fig. 3.4b). Among the top 10 enriched KEGG pathways, two notable downregulated pathways were the focal adhesion and actin cytoskeleton pathways. Both play pivotal roles in transmitting signals from a cell's extracellular matrix to its intracellular components, facilitating a suitable cellular response (Janiszewska et al., 2020). In addition, the P13K-Akt signalling pathway, MAPK signalling pathway, and proteoglycans in cancer pathway, were also downregulated and among the most significantly enriched pathways.

Performing GSEA against the GO-BP terms revealed 976 enriched biological processes (supplementary figure 7) . Our analysis was narrowed to the most significantly enriched processes, with the highest probability of being biologically relevant based on their adjusted p-value. Figure 3.5 displays what the top enriched processes were, as well as the number (Fig. 3.5a), the direction of differential expression (Fig. 3.5b), and the sharing of the differentially expressed genes that make up these processes (Fig. 3.5c). Of the 10 most significantly enriched processes, five were involved in cellular respiration, and three were involved in cell-to-cell adhesion. The cellular respiration related processes were upregulated in the group of AA relative to the EA group and shared many genes between them, while the cell-to-cell adhesion related processes were downregulated in the group of AA relative to the group of EA (Fig. 3.5).

Overall, through GSEA using the GO-BP terms and KEGG pathway database, we identified numerous gene sets with subtle yet coordinated differential expression between the AA and EA groups. Notably, gene sets associated with cellular respiration and neurodegenerative conditions were upregulated, while those related to cell-to-cell adhesion and cell signalling were downregulated in the AA group relative to the EA group.

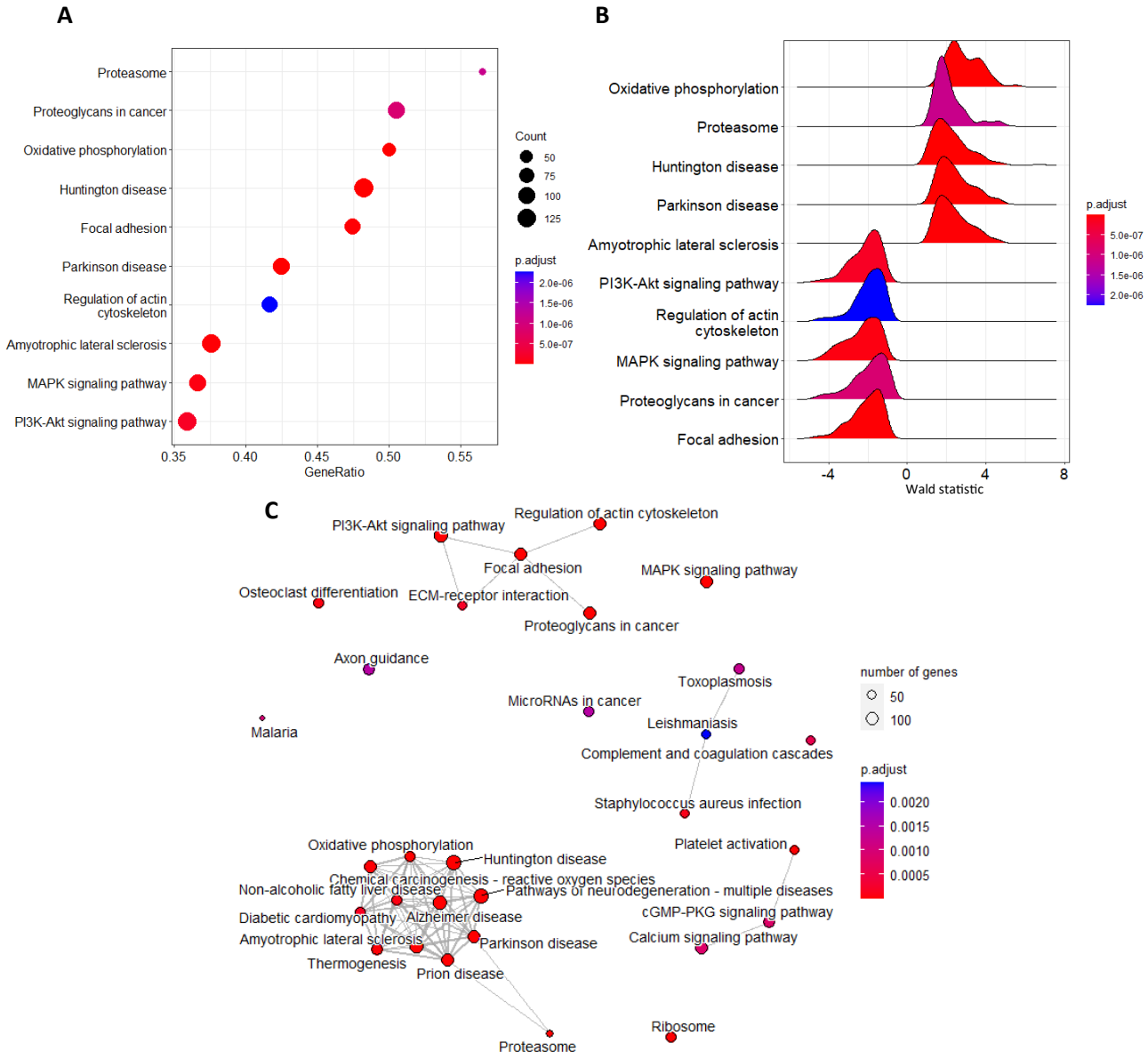


Figure 3.4. Results of GSEA, which was performed on the output of differential gene expression analysis between EA and AA groups, using the KEGG pathway database. (a) Dot plot showing the top 10 most enriched KEGG pathways. The gene ratio is the ratio of the differentially expressed genes which make up the enriched pathway, to the total number of genes that make up the pathway. (b) Ridge Plot showing the top 10 most enriched pathways, and the distribution of the Wald statistic (stat value) of the differentially expressed genes that make them up. (c) Enrichment map showing the top enriched pathways. Lines connecting two pathways indicate that differentially expressed genes are shared between them.

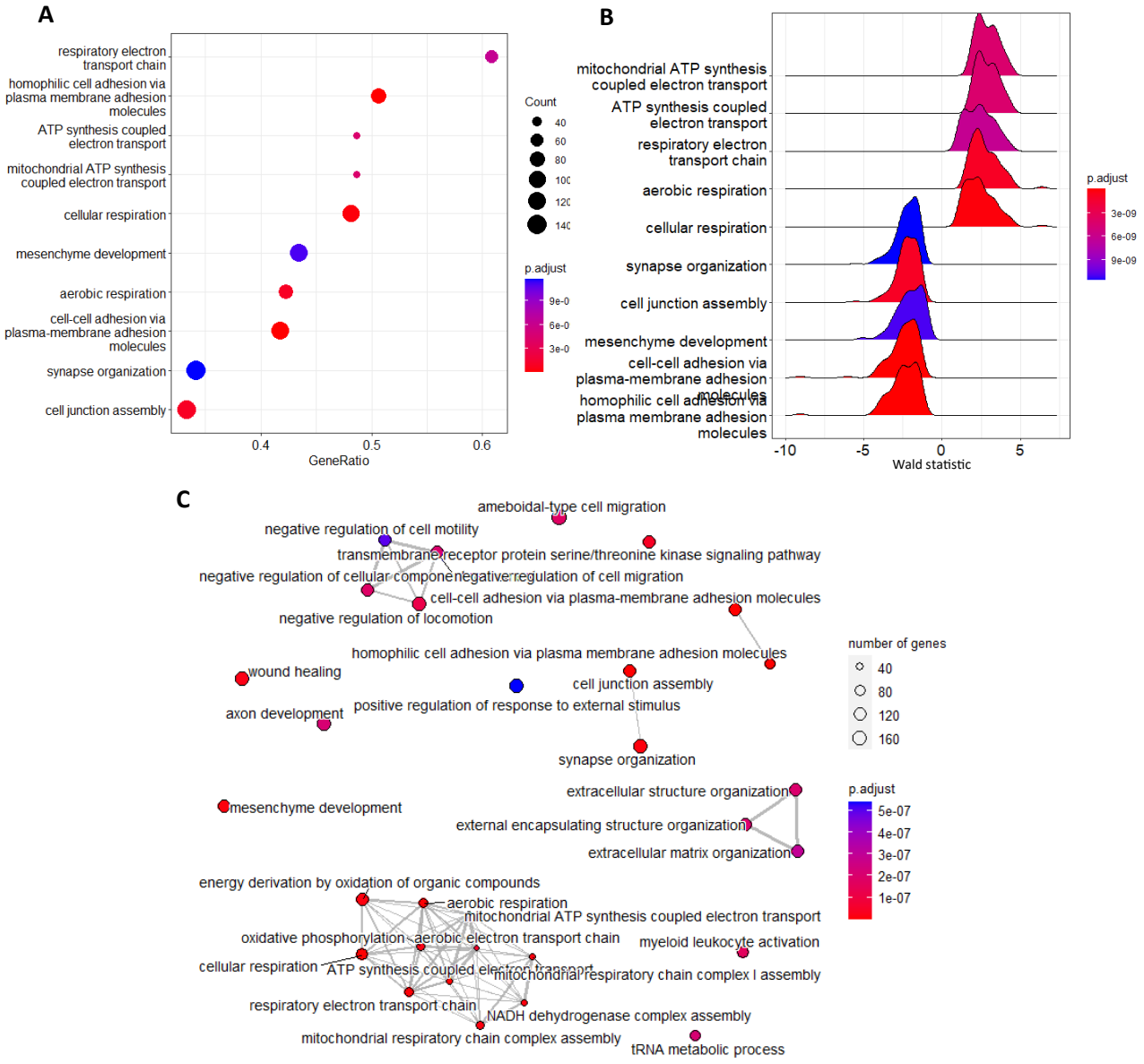


Figure 3.5. Results of GSEA, which was performed on the output of differential gene expression analysis between EA and AA groups, against the GO-BP terms. (a) Dot plot showing the top 10 most enriched biological processes. The gene ratio is the ratio of differentially expressed genes which make up the enriched biological process, to the total number of genes which make up the biological process. (b) Ridge Plot showing the top 10 most enriched biological processes, and the distribution of the Wald statistic (stat value) of the differentially expressed genes that make them up. (c) Enrichment map showing the top enriched biological processes. Lines connecting two biological processes indicate that differentially expressed genes are shared between them.

3.4 Discussion

We looked at differential gene expression between the EA and AA LUAD tumour samples, then performed enrichment analysis on the results of differential gene expression analysis to identify molecular pathways and biological processes through which differences in gene expression between the groups may affect tumour biology.

3.4.1 Differential gene expression analysis

Among the 371 genes our analysis identified as SDE between the AA and EA groups, 171 are upregulated, and 200 are downregulated in AA patients relative to EA patients. A standout gene among these SDE genes is *CDKN2A* because it is the only SDE gene that is associated with NSCLC, according to the KEGG pathway database.

CDKN2A's expression is upregulated in AA samples relative to EA samples. The *CDKN2A* gene codes for the p16 protein, which is a tumour suppressor gene whose deletion is implicated in the development of a variety of cancers such as leukaemia, melanoma, oesophageal and lung cancer (Shi et al., 2022). P16 is a cyclin-dependent kinase inhibitor that acts as a tumour suppressor by inhibiting cell progression through the S phase of the cell cycle. Molecularly, p16 achieves this by inhibiting the activity of cyclin-dependent kinase 4 (CDK4), which in turn blocks off the downstream phosphorylation and activation of retinoblastoma (RB) proteins, which, when phosphorylated would otherwise facilitate the transition of cells from the G1 to the S phase of the cell cycle (Romagosa et al., 2011). Although the association of p16 deletion with the transition from pre-malignant to malignant cells has implicated p16 as an essential tumour suppressor gene, on the other hand, p16 overexpression has also been observed in malignant cells (Romagosa et al., 2011). The overexpression of p16 associated with a malignant cell is counterintuitive because p16 is a tumour suppressor. It is thought that this overexpression of p16 in malignant cells results from the dysregulation of the p16 – Rb pathway, when Rb loss (by mutation or deletion) results in positive feedback that increases the production of p16 (Romagosa et al., 2011). So, in this case, p16 overexpression is a failed attempt to prevent uncontrolled proliferation caused by Rb pathway failure. In the context of this study, Rb does not display reduced expression or increased deletion in the AA samples vs. the EA samples, so it is not obvious that p16 overexpression in AA vs. EA tumours is directly tied to Rb loss.

Concerning its effect on cancer prognosis, p16 protein overexpression has been identified as an unfavourable prognostic indicator in colon adenocarcinomas, breast cancer, astrocytomas and gastrointestinal cancer (Arifin et al., 2006; Lam et al., 2008; Milde-Langosch et al., 2001; Romagosa

et al., 2011; Steigen et al., 2008). Studies examining ethnicity-based differences in cancer gene expression have shown CDKN2A overexpression in AA vs. EA breast cancer and endometrial cancer tumours (Grunda et al., 2012; Javadian et al., 2021; Martin et al., 2009), consistent with our results.

Being the only cancer-associated gene we detected as SDE between the TCGA-LUAD AA and EA samples, the overexpression of *CDKN2A* (p16) was of particular interest. In the context of cancer and lung cancer, p16 deletions are often associated with cancer. Although p16 overexpression is less common in cancers, it has been associated with poor cancer prognosis on several occasions. The CDKN2A overexpression that we have identified in AA vs EA NSCLC tumours may therefore represent a clinically relevant difference in tumour biology between the two groups.

3.4.2 Enrichment analysis

3.4.2.1 Over-representation analysis (ORA)

In the results of ORA on the list of SDE genes, we saw the enrichment of two KEGG pathways and three GO-BP's that have to do with the cellular metabolism of drugs and xenobiotics. Among the genes comprising these enriched drug and xenobiotics metabolism-related gene sets, were some Cytochrome P450 (CYP) genes. CYPs are important enzymes in cancer formation and treatment because they mediate the activation of multiple procarcinogens and are involved in activating and inactivating anticancer drugs (Rodriguez-Antona and Ingelman-Sundberg, 2006). Therefore, altered CYP expression in tumour cells could alter drug efficacy. Of the CYP genes found in the many enriched xenobiotics metabolism-related gene sets, two are of particular interest because of previous research findings about them. These two genes are *CYP2B6* and *CYP3A4*. From our results, their expression was significantly upregulated in AA samples relative to EA samples. *CYP2B6* is involved in the metabolism of some procarcinogens and various therapeutic drugs. *CYP3A4* is involved in the metabolism and breakdown of various anti-cancer drugs into inactive forms. Among others, some important anti-cancer drugs that *CYP3A4* can deactivate are the chemotherapies docetaxel and irinotecan and the tyrosine kinase inhibitor gefitinib. Because these drugs are used in the treatment of NSCLC, the overexpression of *CYP3A4* in AAs compared to EAs might predict lower efficacy of these drugs against the AA patient-derived NSCLC tumours (Rodriguez-Antona and Ingelman-Sundberg, 2006).

In the results of ORA, within the enriched gene sets involved in the response to xenobiotics as well as hormone regulation and metabolism, we saw several *UDP glucuronosyltransferase* (*UGT*) family genes that were highly downregulated in the AA group relative to the EA group. *UGTs* are enzymes that catalyse the conjugation of glucuronic acid to various target molecules, including steroid hormones, bile acids, bilirubin, carcinogens, and therapeutic drugs. This conjugation acts on these

compounds by abolishing their biological activity and enhancing their solubility, which ultimately facilitates their excretion from the body (Allain et al., 2020). Reports investigating metabolic alterations in the transcriptome and metabolome of multiple tumour types found pentose and glucoronate interconversion pathways, to which all *UGT* genes belong, to be amongst the most perturbed (Rosario et al., 2018; Wikoff et al., 2015). *UGT*, therefore, plays an important role in mediating the bioavailability and bioactivity of a variety of endogenous compounds, drugs, and xenobiotics within a tumour microenvironment. *UGTs* are both overexpressed and under-expressed in tumours.

Some SDE *UGTs* that form part of the overrepresented response to xenobiotics and hormone regulation gene sets are *UGT1A*, *UGT1A8* and *UGT2B11*. These UGTs are downregulated in AA samples compared to EA samples. With regards to the onset of lung cancers, deletions or low-activity *UGT1A* mutations have been linked with an increased risk of lung cancer, due to their decreased activity in the metabolism of smoking-derived carcinogens (Liu et al., 2022; Wassenaar et al., 2015). Regarding cancer drug resistance, increased pre-treatment expression of *UGT1A* has been associated with a diminished response to the epidermal growth factor (*EGFR*) inhibitor erlotinib in NSCLC and head and neck cancer patients (López-Ayllón et al., 2015; Thomas et al., 2013). Additionally, because of *UGT1A's* role in converting the toxic anti-cancer irinotecan drug into a less toxic form, small cell lung cancer patients with low-activity *UGT1A* mutant alleles experience increased irinotecan-related toxicity (Liu et al., 2022). Some UGTs downregulated in the AA vs. EA tumours are involved in hormone metabolism. *UGT1A8* is central to the metabolism and excretion of estrogen, which has pro-growth effects on lung tumour cells (Liu et al., 2022). UGT2B11 is involved in the glucuronidation and excretion of androgens, steroids, and toxic compounds (Liu et al., 2022). Increased expression of *UGT2* proteins (which are downregulated in AAs vs. EAs in this study) has been associated with increased risk of metastasis and worse survival in other cancers like chronic lymphocytic leukaemia (CLL), prostate, bladder, and breast cancer (Allain et al., 2020). The important role of UGTs in the metabolism of various drugs, hormones, and other compounds is well established. It is thus foreseeable that the downregulation of *UGTs* that we see in AA patients vs. EA patients may alter their NSCLC tumour microenvironment in clinically relevant ways. (Allain et al., 2020) (Liu et al., 2022).

The presence and activity of endogenous hormones and xenobiotics, such as drugs, affect the microenvironment of a tumour. Differences in the expression of proteins that metabolise, activate, or deactivate such hormones and drugs allow cancer cells to respond uniquely to these substances, and affect the abundance of these substances in their microenvironment. As has been outlined, the enrichment of KEGG pathways and GO-BPs associated with the response to xenobiotics, as well as

the metabolism and regulation of hormone levels among the genes that are SDE between the AA and EA tumours, shows that these groups of LUADs may respond differently to anti-cancer drugs and endogenous hormones. In particular, this difference is likely to result from the differential expression of *cytochrome p450* and *UDP glucuronosyltransferase* family genes. Differential expression of *cytochrome P450s* and *UDP glucuronosyltransferases* has been associated with differences in prognosis and treatment response of cancers (Liu et al., 2022; López-Ayllón et al., 2015; Rodriguez-Antona and Ingelman-Sundberg, 2006; Thomas et al., 2013; Wassenaar et al., 2015).

3.4.2.2 Gene set enrichment analysis (GSEA)

To look more generally at the effects of differential gene expression between the two groups, we performed GSEA using gene sets from the KEGG pathway database and GO-BP annotations.

Amongst the top enriched and upregulated gene sets, outputted after performing GSEA, are gene sets that relate to the process of cellular respiration, particularly mitochondrial energy production. To appreciate the significance of this, it's essential to delve into the Warburg Effect, a phenomenon long regarded as a pivotal shift in cellular metabolism during the transition of a normal to a cancerous state (Warburg, 1956). One of the hallmarks of cancer development is the reprogramming of energy metabolism. The Warburg effect refers to cells displaying an increased rate of ATP (energy) production from glucose via glycolysis (aerobic glycolysis), at the expense of ATP production via mitochondrial respiration, which normal cells would generally rely on more. According to the Warburg effect, to enhance their rate of proliferation, cancer cells tend toward producing ATP through aerobic glycolysis, even though it produces 16x less ATP than mitochondrial ATP synthesis, because aerobic glycolysis metabolises glucose at a much faster rate than mitochondrial respiration and by so doing produces a variety of by-products that are useful for cell proliferation. The Warburg effect suggests that the rapid glucose metabolism through aerobic glycolysis is favoured by cancer cells over the slower glucose metabolism via mitochondrial respiration, because the by-products of aerobic glycolysis support cancer cell growth(Liberti and Locasale, 2016). Some evidence partially supports the Warburg effect in some cancers (Xintaropoulou et al., 2018) (Liberti and Locasale, 2016). However, there is also evidence that in some cases, mitochondrial ATP production is not compromised and is instead central to cancer cell proliferation. For instance, a study in a mouse model showed that the disruption in the mitochondrial function of LUAD cells reduced tumorigenesis (Weinberg et al., 2010). Another study showed that mitochondrial DNA-depleted LUAD cells would take up wild-type mitochondrial DNA from neighbouring cells to sustain their mitochondrial activity and cell replication (Tan et al., 2015). These examples show the central role that mitochondrial activity can have in the activity of cancer cells. In our case, the upregulation of

multiple mitochondrial respiration-related gene sets among our samples of AA vs. our samples of EA suggests that ATP production is less glycolytic and more mitochondrial in the AA tumours. This is further supported by the observed downregulation of the PI3K-AKT signalling KEGG pathway among AA vs. EA samples, as PI3K-AKT upregulation is primary for increasing glucose uptake by cells to support increased rates of aerobic glycolysis in glycolytic cancer cells (Rascio et al., 2021). The upregulation and enrichment of the oxidative phosphorylation KEGG pathway also supports the proposed mitochondrial ATP production bias in the AA tumours. The repurposing of oxidative phosphorylation-inhibiting drugs to target cancer cells with increased mitochondrial respiration has been suggested as a potentially viable anti-cancer therapy (Bedi et al., 2022). There are, however, concerns about the specificity of such drugs, and the resistance to the drug that can come from cancer cells becoming more glycolytic to avoid detection. The upregulation of multiple mitochondrial respiration-associated gene sets, as well as the downregulation of the PI3K-Akt signalling pathway in the AA tumours vs. the EA tumours, suggests that the AA LUAD tumours rely more on mitochondrial respiration for energy production than the EA tumours do. This difference in energy metabolism might correlate with differences in cancer progression and might call for varied treatment to target the primary energy production pathways and tools that sustain cancer cell proliferation.

The GSEA results also show the enrichment and upregulation of pathways associated with neurodegenerative conditions like Huntington's disease, Parkinson's disease, Alzheimer's, and others. Many genes that make up these pathways are shared with the enriched oxidative phosphorylation and mitochondrial energy production-related pathways. Mitochondrial dysfunction is a central property of neurodegenerative conditions, leading to neurons not receiving appropriate amounts of energy to function effectively and becoming damaged (Johri and Beal, 2012; Mattson, 2000). In our findings, the mitochondrial genes within the enriched neurodegenerative condition-related gene sets demonstrate upregulation, suggesting heightened mitochondrial activity. The neurodegenerative condition-related gene sets are enriched largely because of the upregulation of many mitochondrial respiration-related genes. We therefore postulate that the enrichment of these neurodegenerative condition-related gene sets doesn't relate to increased incidence or likelihood of the neurodegenerative conditions among the AA group, as that would instead be associated with the downregulation of the genes in these gene sets. The enrichment and upregulation of the neurodegenerative condition gene sets is likely an artefact of the many mitochondrial energy production-related gene sets that are upregulated and enriched. The overlap in genes associated with lung cancer and Alzheimer's and their inverse gene expression in these conditions has been observed in many other studies and is not unusual (Li et al., 2022). Studies showing a reduced incidence of some cancers among patients with neurodegenerative conditions further suggests that

there is an inverse alteration of similar genes in neurodegenerative conditions vs. cancers (Bennett and Leurgans, 2010; Inzelberg and Jankovic, 2007; Sorensen et al., 1999). Considering this, it is also possible that there is an upregulation of certain pro lung cancer genes among AA tumours, whose downregulation is associated with Alzheimer's. The upregulation of these genes can lead to the enrichment of the Alzheimer's gene set but does not directly have to do with Alzheimer's.

Our GSEA results also show the enrichment and downregulation of gene sets associated with cell-to-cell adhesion in the AA group vs. the EA group. This is interesting because the loss of cell-to-cell adhesion and anchorage-independent growth of cells are both hallmarks of cancer and result from the dysregulation of cell adhesion molecules (Janiszewska et al., 2020). Cell adhesion molecules help connect cells to maintain tissue structure and are involved in extracellular matrix-to-cell signal transduction. Dysregulation of cell adhesion molecules can, therefore, disrupt tissue structure, cell anchorage, and signal transduction into cells and, in this way, facilitate malignancy and metastases. One of the major consequences of cell adhesion molecule dysregulation is the epithelial-mesenchymal transition (EMT), which makes cancer cells more motile and invasive (Kalluri and Weinberg, 2009). It might be tempting to link the downregulation of the gene sets "cell-cell adhesion via plasma membrane adhesion molecules" and "homophilic cell adhesion via plasma membrane adhesion molecules" in AA vs. EA tumour cells to reduced cell-to-cell adhesion and communication. Such a change could potentially correlate with heightened malignancy and an increased likelihood of metastasis. Our interpretation of the result cannot be this simple as the combination of upregulation and downregulation of different cell adhesion molecules is needed to disrupt the epithelial nature of cells and make them lean toward a more malignant and metastatic mesenchymal state. To more conclusively link our findings with diminished cell-cell adhesion in the AA group, it would be necessary to observe the downregulation of genes known to be associated with this function, such as *CDH1*, which encodes E-cadherin. Among the SDE genes in the enriched cell-to-cell adhesion gene sets is *CDH17*, which encodes N-cadherin. Research has associated the downregulation of E-cadherin, and the upregulation of N-cadherin with the epithelial-mesenchymal transition, a process where cells shift from a localised epithelial state to a more dispersed mesenchymal state. While the precise impact on cell-to-cell adhesion remains elusive, the notable downregulation of N-cadherin in AA samples compared to EA samples suggests potential differences in cell adhesion dynamics, where AA tumours may be less mesenchymal than EA tumours. This notion is further supported by the enrichment and downregulation of the mesenchymal development GO-BP term in AA vs. EA samples. Expression changes associated with decreased cell-to-cell adhesion and mesenchymal development vary according to tissue type. In this context, it's challenging to assert with confidence that the observed changes indicate a decrease in mesenchyme development in AA tumours

compared to EA tumours. The downregulation and enrichment of the mesenchyme development GO-BP and the downregulation of the N-cadherin encoding gene do, however, suggest this. The dysregulation of cell adhesion molecules is primary in the development of malignant and metastatic cancer cells. Our GSEA results show that interesting differences in this important aspect of tumour development and metastasis may exist between the two groups.

4. Comparison of somatic copy number alterations (SCNA)

4.1 Introduction

Copy number alterations are increases (amplifications) or decreases (deletions) in the number of copies of a stretch of DNA. Somatic copy number alterations (SCNAs) are among the DNA sequence changes associated with the transformation of normal cells into cancerous cells.

The frequency of SCNAs in lung cells has been observed to rise as they transform from healthy cells to malignant ones, and further as the cancer progresses from early to advanced stages (Huang et al., 2011). The association between SCNAs and the nature of NSCLC is further reiterated by studies that have pinpointed specific SCNAs capable of differentiating between normal lung cancer cells, squamous cell carcinoma cells and adenocarcinoma cells (Li et al., 2014; Qiu et al., 2017). Common oncogenic gene SCNAs associated with NSCLC include the amplification of *MET*, *FGFR*, *EGFR*, *HER2*, and *TITF-1*, as well as the deletion of PTEN (Herbst et al., 2008; Shames and Wistuba, 2014). Some of these SCNAs have shown potential for informing NSCLC treatment selection. For example, *MET* amplification and *PTEN* deletion have been associated with resistance to EGFR tyrosine kinase inhibitors, while on the other hand, *HER2* amplifications have been associated with sensitivity to *EGFR* tyrosine kinase inhibitors (Bean et al., 2007; Cappuzzo et al., 2005; Engelman et al., 2007; Sos et al., 2009).

Research examining ethnicity-based differences in SCNAs has yielded varied conclusions. After comparing 245 AA samples to 264 EA samples from NSCLC tumours, Campbell et al. (2017) found no differences in the frequency of SCNAs per sample between the two ethnic groups. Sinha et al. (2020) evaluated CNA genomic instability (characterised by the proportion of the genome with a non-diploid copy number). In their findings, among Squamous cell carcinoma samples, those of AA exhibited significantly more copy number-associated genomic instability than their EA counterparts. However, this ethnicity-based difference was not observed among LUAD patients. Various studies have identified significant differences in the chromosomal positioning of SCNA regions between AA and EA NSCLC samples (Shi et al., 2022) (Sinha et al., 2020). Regarding cancer-related genes, a comparison of 126 AA and 96 EA NSCLC samples observed that *KRAS* amplifications and *PTEN* deletions occur in both EA and AA samples but occur with a higher frequency in samples of AA (Sinha et al., 2020). In the same study, deletions of *CDKN2A* were observed in both groups, but were more prevalent in AA samples. Furthermore, the frequency of cancer-associated *MET* amplification has been found to be consistent across ethnic groups (Steuer et al., 2016).

Copy number alterations can be clinically useful in categorising cancers into their subtype and informing treatment selection for NSCLCs. Although research is limited, studies have presented varying conclusions on whether differences in the frequency and type of SCNAs exist between different ethnic groups. In this chapter, we explore potential disparities in the SCNA profiles between AA and EA LUAD specimens sourced from the TCGA-LUAD study.

4.2 Methods

4.2.1 Downloading somatic copy number alteration data

SCNA data for samples of the TCGA-LUAD study were downloaded from the Broad Institute of Harvard and MIT's FireBrowse data portal (http://firebrowse.org).

To generate SCNA data for normal and tumour cells of samples, the Broad Institute used the Affymetrix SNP Array 6.0. The SNP Array has probes complementary to sequences across the human genome. To determine which discrete chromosomal regions were copy number altered, hybridization of DNA to the array's probes was first quantified. Based on these hybridization levels, gene segments representing amplified or deleted stretches of the genome were determined. Finally, gene segments were assigned segment means, indicating their amplification or deletion magnitude.

The SCNA data, containing segment means of normal and tumour cells from the TCGA-LUAD study, was downloaded from the Broad Institute's FireBrowse data portal and subsequently filtered and processed in R Studio (version 4.2.1). To begin with, the SCNA data linked to the tumour cells from our 1:3 propensity-matched AA and EA samples were isolated. Among the propensity-matched samples, SCNA data was available for 146 out of 147 EA samples and all 49 AA samples, forming the basis for subsequent SCNA analyses. A binary interpretation of SCNA segments was implemented and classified SCNA segments with a segment mean of >0.3 as amplified and those with segment means of < -0.3 as deleted. For every sample, segments that did not meet the amplification and deletion thresholds were removed. The resulting data contained CNA segments classified as amplifications and deletions for each of the 146 EA and 49 AA samples, which served as the starting point for subsequent SCNA comparisons.

4.2.2 Comparison of somatic copy number alteration burden

To identify whether there is a difference in the median number of SCNA segments per sample between the EA and AA groups, the number of SCNA segments per sample were first compiled, based on our binary classification of amplifications and deletions. Box plots were generated using the in-house RStudio boxplot function to visually compare SCNA segments per sample between the AA and EA groups. Subsequently, a Wilcoxon rank-sum test (Mann-Whitney U test) assessed if there

was a statistically significant difference (p-value < 0,05) in the median number of SCNA segments per sample between the two groups.

4.2.3 Identification and comparison of recurrently copy number altered regions

The processed SCNA data was used to independently identify and plot recurrent copy number alterations for samples of AA and EA. A modified version of Silva et al's (2016) TCGA Workflow for Genomic Analysis was used to do this. The steps that were followed are outlined in more detail below.

The GAIA (version 2.39.0) (Morganella et al., 2011) package was utilised to determine recurrently copy numbered altered (recCNA) regions in the AA and EA groups independently. GAIA was executed on the University of Cape Town's High-Performance Cluster using the R programming environment (version 4.1.1). Prior to running GAIA, a copy number variant matrix and a markers object were created. The copy number variant matrix detailed the position and type of copy number alterations for each sample, and the markers object outlined all the probes used for copy number alteration measurement. The markers object was edited to exclude all probes that were considered part of common somatic copy number variants in non-cancer cells according to the Broad Institute (available at ftp://ftp.broadinstitut.org/pub/GISTIC2.0/hg19_support/). Using this edited markers object excluded SCNAs previously identified as prevalent in non-cancer samples from the analysis.

GAIA was then run to identify recCNA regions, with the copy number matrix and edited markers objects as its inputs. RecCNA regions were defined by a false discovery rate (q-value) < 0.15 after 10 iterations of running GAIA. After running GAIA independently for the AA and EA groups, a list of recCNA regions for both the AA and the EA groups was obtained.

For each group, AA and EA, the resulting file of recCNA regions outputted by GAIA was loaded into R studio where it was used to plot the type and position of recCNA regions. A custom function similar to that used by Silva et al. (2016) was used to plot recurrent amplifications and deletions onto a copy number alteration plot.

4.2.4 Comparison of genes in recurrently copy number altered regions

Genes located within recCNA regions (q value < 0.15) in the tumour samples of AA and EA were pinpointed. Additionally, we ascertained whether these recCNA occurrences were shared between the AA and EA tumours.

To begin with, BiomaRt (version 2.52.0) was utilised in R studio to download a list of all genes in the human grch37 (hg) genome (Durinck et al., 2009). Each gene's gene symbol, chromosomal number, nucleotide start position, and nucleotide end position was obtained. The GenomicRanges (1.48.0)

package was then used to store all that information in a GenomicRanges object (Lawrence et al., 2013). The genomic coordinates of the recCNA regions among AA and EA samples identified by GAIA, were also stored in their own GenomicRanges objects. The findoverlaps function, from the GenomicRanges package, was then used to find overlaps between the genomic coordinates of the recCNA regions and the grch37 genes. By doing this, the recurrently copy number altered segments in the groups of AA and EA were annotated with the names of the genes that fell within these segments. Venn diagrams were made to show the number of amplified and deleted genes that were recurrently copy number altered in both groups or in the AA or EA group alone. The KEGG pathway NSCLC gene list was referred to when identifying NSCLC-associated recCNA genes.

4.3 Results

4.3.1 Comparison of somatic copy number alteration burden

Our analysis aimed to understand the differences in the genomic landscape between AA and EA LUAD samples. Box plots visually comparing the median number of SCNAs per sample between the two groups showed that the AA samples had a greater SCNA burden than the EA samples (Fig 4.1). A Wilcoxon Rank sum test was performed to determine whether this difference could be statistically supported. A p-value of 0.013 from the Wilcoxon Rank Sum test provided statistical support for the observed difference in SCNA burden between the two groups.

4.3.2 Comparison of recurrently copy number altered regions

RecCNA regions for the groups of AA and EA were determined and plotted independently using the GAIA package. Any SCNA regions identified as occurring in normal non-cancer cells by the Broad Institute were excluded from this process.

The outputted recCNA plots (Fig. 4.2) showed differences in the type and position of recCNA regions between the AA and EA samples. Between the two groups, chromosomes 1, 5, 8, 12, and 17 had recurrent copy number alterations in the opposite direction, while chromosomes 4 and 18 were only recCNA in the EA group and not the AA group.

4.3.3 Comparison of genes in recurrently copy number altered regions

Venn diagrams showing the recurrently amplified and deleted genes shared between the AA and EA groups or only present in one of them were constructed. The samples of EA had many more genes in their recCNA regions than the samples of AA (Fig. 4.3). This is because there were just under three times as many EA samples as AA samples in our analysis, and this allowed GAIA to identify more recCNA segments among EA samples (2347) than AA samples (769). The majority of recurrently amplified and deleted genes were shared between both groups. Among the shared and unique

recCNA genes, several were associated with NSCLC (Fig. 4.3). The samples of EA had many more recurrently amplified and deleted genes that were NSCLC-associated than the AA samples. This was in line with the greater number of recCNA regions identified among the EA group compared to the AA group.

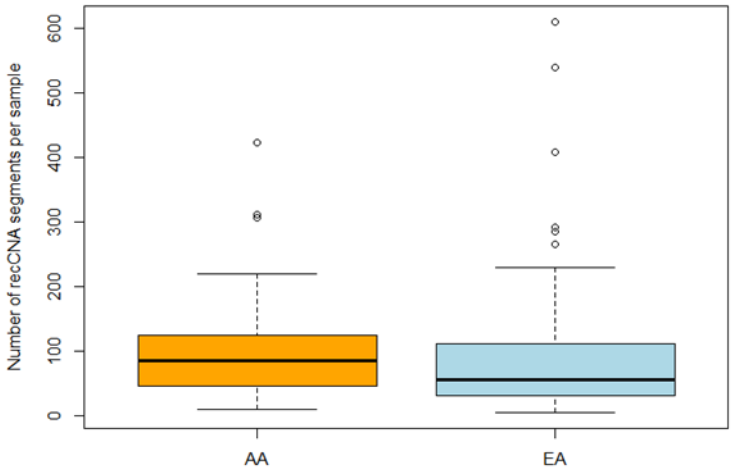


Figure 4.1. Box plots showing the somatic copy number alterations per sample for the AA and EA groups. The greater median number of SCNA per sample in the group of AA vs. EA group was statistically supported by a Wilcoxon rank-sum test which produced a p-value of 0.013.

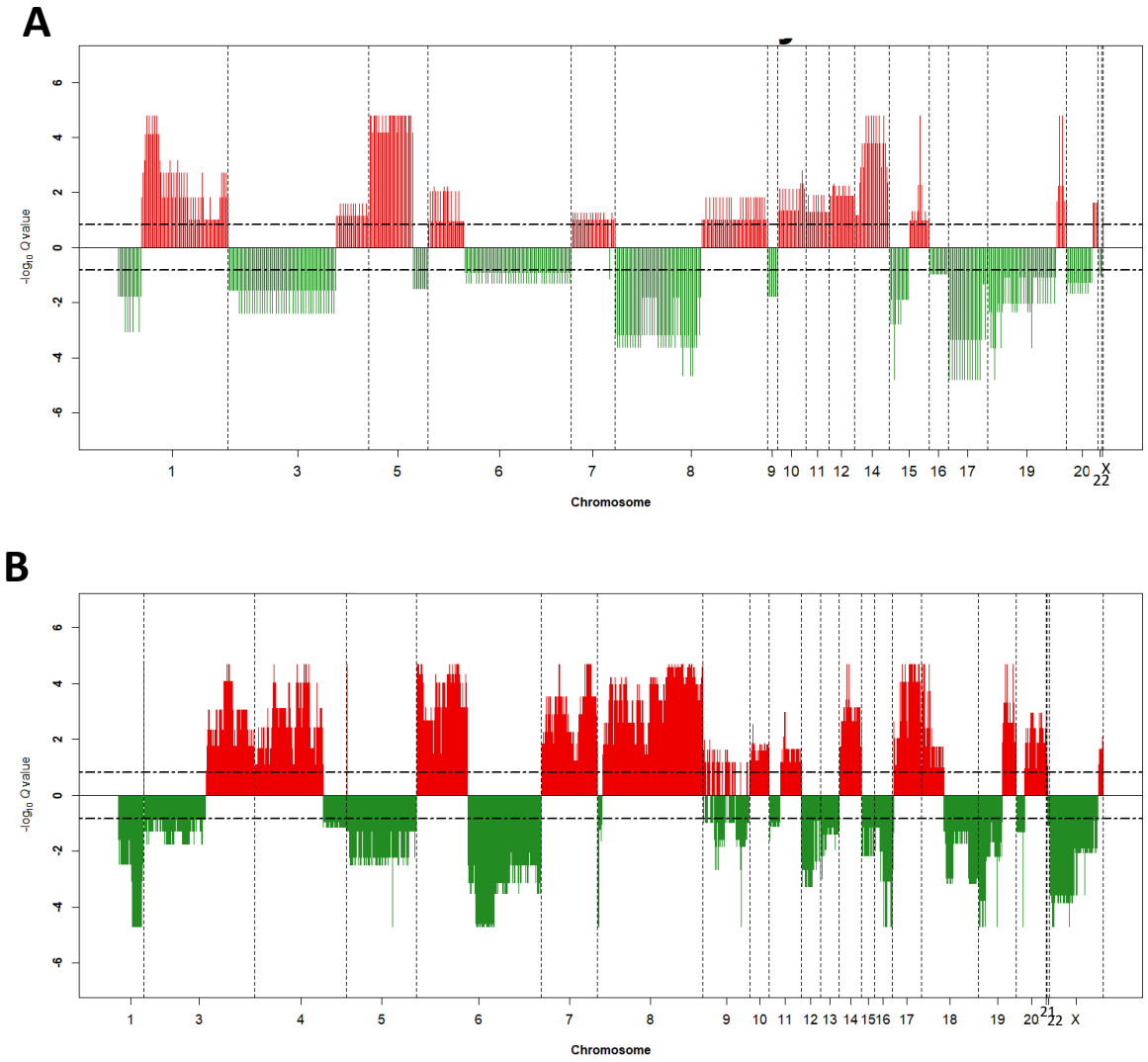
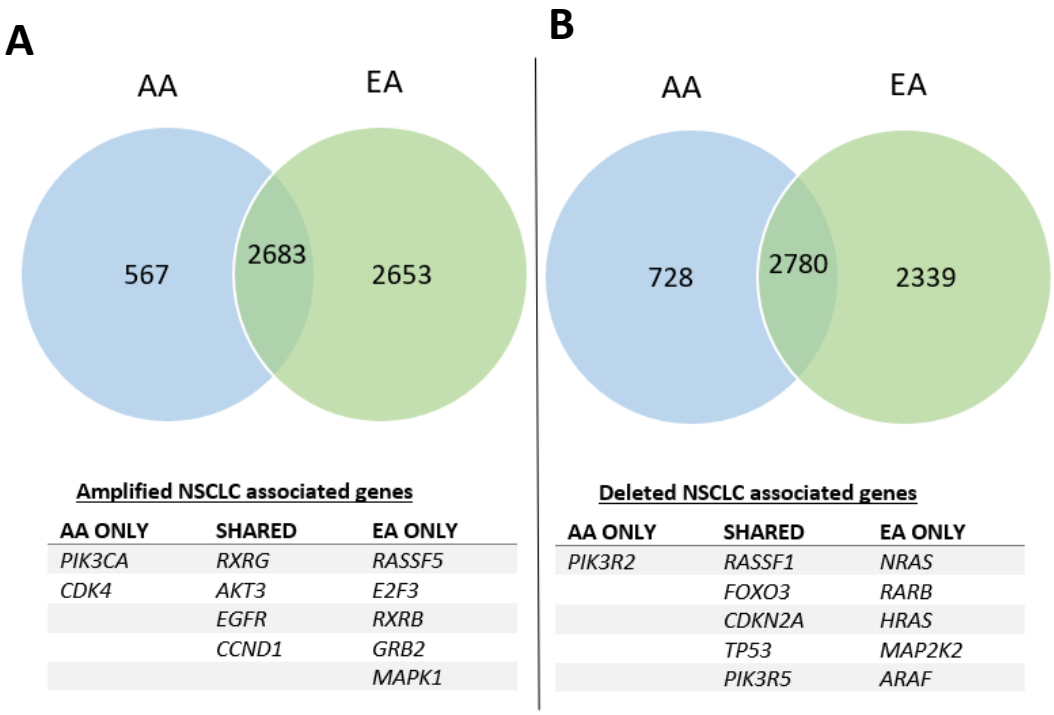


Figure 4.2. Recurrently copy number altered regions in each ethnic group. Recurrent copy number alteration (recCNA) plots for AA samples (a) and EA samples (b). RecCNAs were determined and plotted using GAIA, with SCNAs known to be present in normal cells filtered out. The horizontal dotted line represents the q-value cutoff of <0.15 (log10-transformed in the figure) that was used as a threshold when determining recCNA regions.



Amplified NSCLC associated genes

AA ONLY	SHARED	EA ONLY
PIK3CA	*RXRG*	*RASSF5*
CDK4	*AKT3*	*E2F3*
	EGFR	*RXRB*
	CCND1	*GRB2*
		MAPK1

Deleted NSCLC associated genes

AA ONLY	SHARED	EA ONLY
PIK3R2	*RASSF1*	*NRAS*
	FOXO3	*RARB*
	CDKN2A	*HRAS*
	TP53	*MAP2K2*
	PIK3R5	*ARAF*

Figure 4.3. Genes in recurrently copy number altered (recCNA) regions. (a) Venn diagram showing the number of genes that make up recurrently copy number amplified regions in AA and EA samples, as well as a table showing the NSCLC associated genes in these regions. (b) Venn diagram showing the number of genes that make up recurrently copy number deleted regions in AA and EA samples, as well as a table showing the NSCLC associated genes in these regions.

4.4 Discussion

In order to gain further insight into what genetic differences exist between our TCGA-LUAD derived AA and EA samples, we compared their copy number altered regions and genes. We found that the samples of AA had a greater SCNA burden than samples of EA, and we found differences in the position, type, and genes of recCNA regions.

In an analysis of 300 NSCLC tumours, Huang et al. observed that the frequency of SCNA in a tumour escalated with the onset and progression of the disease. Moreover, they identified functional and positional clustering of these alterations within the tumours. Their findings, suggest that an increase in certain SCNA functionally facilitates cancer progression. In this study we found that AA LUAD samples had a greater median number of SCNA per sample than EA LUAD samples. This difference is unlikely to be because of a difference in the cancer stage of samples in the two groups, as balancing of T, N and M stages was successfully achieved by PSM. The increased number of SCNAs per sample in AAs vs. EAs may point to the presence of cancer cells that are more genetically altered and require more robust treatment approaches among AA samples.

Comparing the NSCLC-associated genes within the recCNA regions between the AA and EA groups revealed certain NSCLC-associated genes that were copy number altered in one group but not in the other. We further investigated these genes to identify any that also showed notable differential gene expression between the two groups. Such NSCLC-associated genes, being both differentially copy number altered and expressed, are most likely to hold biological significance in the context of NSCLC disparities between the groups. We identified two NSCLC-associated genes, *MAPK1* and *HRAS*, that satisfied these criteria. None of these were SDE between the two groups, but the adjusted p-values from their differential gene expression results were less than 0.05, and their direction of differential gene expression corresponded with the difference in the CNA of these genes between the groups.

MAPK1 was recurrently amplified exclusively in the EA group, and when comparing gene expression in the AA group to the EA group, it exhibited a log2-transformed fold change of -0.33 (absolute fold change of 0.8) with an adjusted p-value of 0.0027. Although this gene expression difference is small, the amplification in EAs alone and reduced expression in AAs indicate that this gene's products are likely more prevalent in the EA group than in the AA group. *MAPK1* amplification has been associated with resistance to anti-cancer EGFR tyrosine kinase inhibitor (TKI) treatments (Blakely and Bivona, 2012). Therefore, more EA tumour patients may benefit from the combination of *EGFR*-TKI and trametinib anti-cancer therapies to maximise treatment effectiveness (Minari et al., 2016).

The *HRAS* gene is recurrently deleted exclusively in the EA samples, and its gene expression in the AA group relative to the EA group presents a log2-transformed fold change of 0.45 (absolute fold change = 1.37) with an adjusted p-value of 0.003. The exclusive recurrent deletion of this gene in EA samples and its increased expression in AA samples tells us that the products of *HRAS* are likely to be more prevalent in AA samples than in EA samples. HRAS is part of a large group of RAS proteins, including *KRAS* and *NRAS*, which, when dysregulated by mutations exist in a constitutively active state that promotes malignancy and tumour growth (Keeton et al., 2017). Recent work by Tang et al. (2023) showed how *HRAS* and *NRAS* suppress *KRAS*-driven lung cancer growth *in vivo*. Interestingly, *NRAS* and *HRAS* are only recurrently deleted among the EA samples. Together with the diminished expression of HRAS in EA compared to AA tumours, this suggests that EA tumours might exhibit diminished *KRAS* suppression, potentially leading to a higher prevalence of *KRAS*-driven tumours.

From our comparison of SCNA between AA and EA LUAD samples, we see a greater burden of SCNA in AA vs. EA samples, as well as differences in the recCNA and gene expression of two NSCLC genes, *MAPK1* and *HRAS*. A main limitation of our SCNA analysis stems from the limited number of AA samples in comparison to the EA samples. This difference in sample size correlated with the number of recurrent copy number alterations that GAIA could identify in AA (769) vs. EA (2347) samples. In

future, to allow for a better comparison of recCNAs between the groups, working with a sample size of AA samples similar to that of EA samples, or incorporating a normalization technique that accounts for the difference in sample size when identifying recCNAs will be useful.

5. Comparison of Mutations

5.1 Introduction

Among the genetic alterations that make cells cancerous, DNA mutations are well studied and have provided a lot of insight into the molecular nature of cancers. As is mentioned in chapter 1.2.2, gene mutations frequently associated with the onset of NSCLC involve tumour suppressors like *LKB1*, *RASSF1A*, *FHIT*, p16 and p53, as well as oncogenic drivers such as *EGFR*, *KRAS*, *BRAF*, *PIK3CA*, *ALK*, *RET* and *VEGF* (El-Telbany and Ma, 2012; Herbst et al., 2008; Shames and Wistuba, 2014; Sinkala, 2023).

Studies examining differences in the presence of NSCLC-related mutations between samples of African ancestry (AA) and European ancestry (EA) have produced varied results. Some studies have identified a lower amount of cancer-related mutations in AA samples relative to EA samples (Araujo et al., 2015a; Costa et al., 2021; Lusk et al., 2019; Steuer et al., 2016), while other studies could not identify any such difference between the groups (Araujo et al., 2015b; Campbell et al., 2017). Cross-ethnic consistency in the frequency of mutations in important NSCLC-related genes such as *EGFR*, *KRAS*, and *PIK3CA* have been reported (Araujo et al., 2015a; Campbell et al., 2017)

Existing research suggests that while the NSCLC genetic mutation landscape is largely similar between tumours of AA and EA, differences in the frequency of cancer driver mutations between the two groups are observed. Considering the underrepresentation of AA patient samples in cancer genetic research, whether ethnicity-based molecular differences in NSCLC contribute to discrepancies in NSCLC incidence and outcomes between different ethnic groups remains a relevant question, and will help inform the clinical utilisation of mutation screening to inform cancer treatment. In this chapter, we compare the mutational burden and frequency of gene mutations in TCGA-LUAD AA and EA samples to see if any relevant differences can be uncovered.

5.2 Methods

5.2.1 Downloading and processing mutation related data for analysis

In order to compare the occurrence of mutations between the two groups, a mutation annotation file (MAF) containing somatic mutations of the TCGA-LUAD samples was downloaded from the NCI Genomic Data Commons (GDC) data portal using the TCGAbiolinks (version 2.24.3) package in RStudio (Colaprico et al., 2016). More specifically, the GDCquery function was used to retrieve a list of MAFs from the TCGA-LUAD study that were open access, had simple somatic mutation information, and were classified as legacy data by TCGA. After assessing the available files,

GDCdownload was used to download the automated somatic MAF file, which contained mutation data for 569 samples from the TCGA-LUAD study. The selected MAF file prioritized sensitivity over specificity unlike the other MAF files, which went through more stringent manual and automatic curation of the identified mutations. The automated somatic MAF file was used because the less stringent filtering of its mutation data provided more opportunities to detect potentially interesting differences in mutations between the AA and EA groups in the analysis that followed. The Maftools package (version 2.12.0) was then used to subset the selected MAF file and create two separate MAF files, one containing the mutation information of the propensity-matched samples of EA and one containing the mutation information of the propensity-matched samples of AA (Mayakonda et al., 2018). The subsequent mutational landscape analyses utilised the resulting MAF files, which encompassed 147 EA and 49 AA samples.

5.2.2 Comparison of mutational burden

To assess the mutational burden of the LUAD cells from the two groups, the median number of mutations (deletions, insertions, and substitutions) per sample for each group was determined then compared. The number of mutations per sample for the AA and EA groups was obtained from their MAF files. Box plots showing mutations per sample for each group were made using RStudio's in-house boxplot function. Subsequently, a Wilcoxon rank-sum test (Mann-Whitney U test) assessed if there was a statistically significant difference (p-value < 0.05) in the median number of mutations per sample between the two groups.

5.2.3 Comparison of mutational frequency

The Maftools package was used to compare the frequency of gene mutations between the groups. Maftools' mafcompare function was utilised to determine which genes had a different mutation frequency between the two groups. Mafcompare uses Fisher's exact statistical test to do this. A p-value threshold of 0.05 was used to identify genes with differing mutation frequencies between the two groups.

5.3 Results

5.3.1 Comparison of mutation burden

To discern potential differences in mutational burden between LUAD samples of AA and EA, we analysed the number of mutations per sample for each group. The box plots in Figure 5.1 show that the median number of mutations per sample was greater in the AA group than in the EA group. The result of the Wilcoxon rank sum test provided a p-value of 0.125, indicating no statistically significant difference in the number of mutations per sample between the two groups.

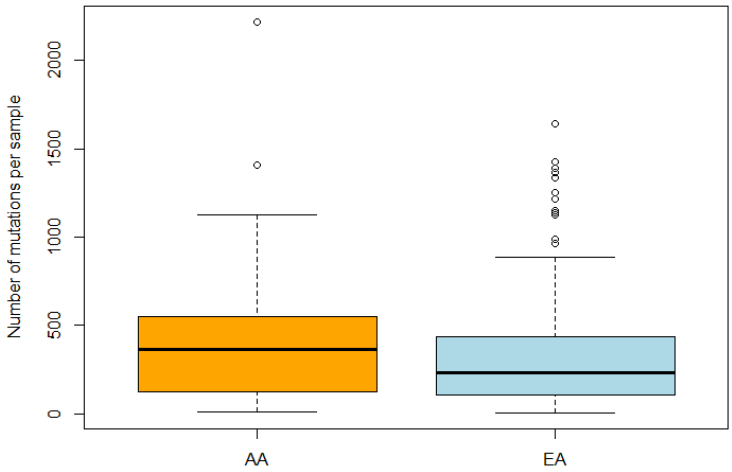


Figure 5.1 Box plots showing the number of mutations per sample for the AA and EA groups. A Wilcoxon rank sum test indicated that there was no statistically significant difference in the median number of mutations per samples between the two groups.

5.3.2 Comparison of mutational frequency

To further explore the potential mutation differences between the two groups, we analysed the mutation frequency of each gene and pinpointed genes that exhibited significantly varied mutation rates between the two groups. There was a large overlap in the mutation frequencies of genes between the two groups. Among the top 10 mutated genes in each of the two groups, 8 were shared and had similar frequencies of mutations in the two groups (Fig. 5.2).

Use of the Fishers exact test revealed that a total of 101 genes were mutated at different frequencies between the two groups (supplementary table 7), with a p-value of < 0.05. Of these, 98 had a higher mutation frequency in the AA group than the EA group, and only 3 had a higher mutation frequency in the EA group than the AA group (Table 5.3). Amongst the differentially mutated genes, no genes formed part of the KEGG pathway NSCLC gene list.

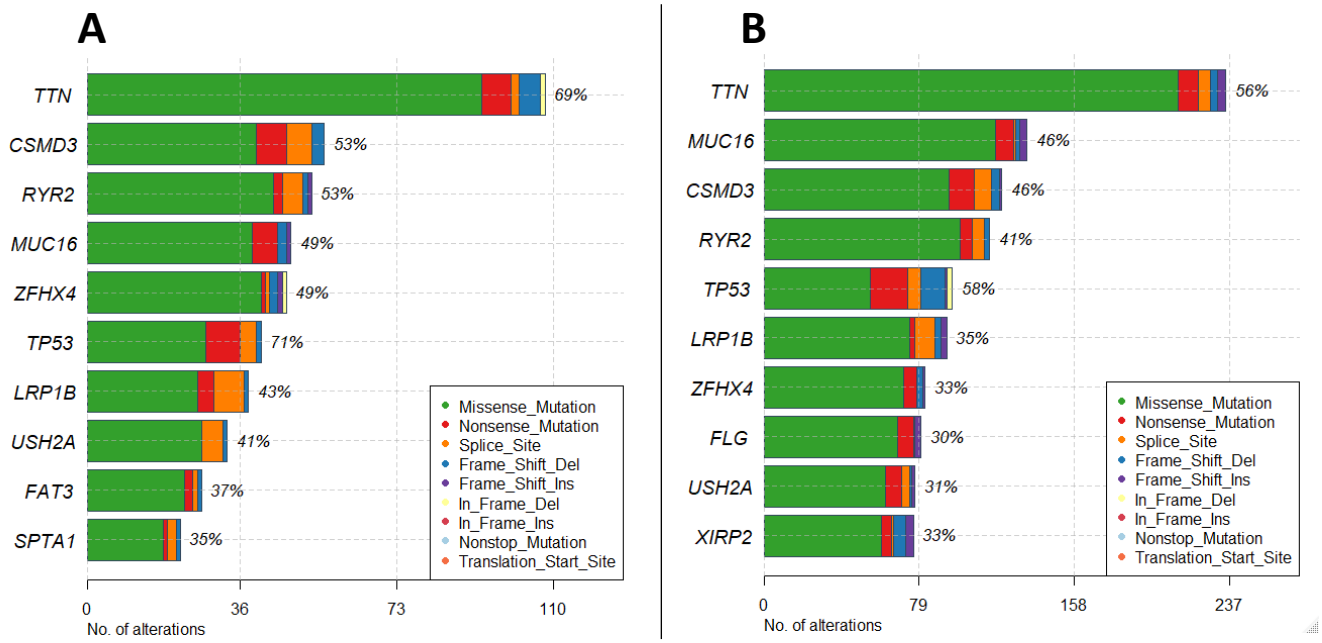


Figure 5.2. Top 10 most frequently mutated genes in AA group (a) and EA group (b). Percentages indicate the proportion of samples in each group with a mutation in a particular gene, and the x axis indicates the number of alterations to that gene among the group.

Table 5.1 Top differentially mutated genes: The tables show the top differentially mutated genes that are more mutated in the AA group than the EA group (a) and those that are more mutated in the EA group than the AA group (b). Genes with a p-value of < 0.05 were considered differentially mutated after performing Fishers' exact test.

A

GENE SYMBOL	AA	EA	PVAL	OR
NME8	9 (18%)	3(2%)	0,0002	10,6
NLRP6	5 (10%)	0	0,0008	Inf
C9ORF3	7(14%)	2(1%)	0,001	11,9
KIAA0226	7(14%)	2(1%)	0,001	11,9
MCM10	7(14%)	2(1%)	0,001	11,9
OPLAH	7(14%)	2(1%)	0,001	11,9
FCGR3A	6(12%)	1(1%)	0,001	20
FLRT3	6(12%)	1(1%)	0,001	20
PCDHGB3	7(14%)	3(2%)	0,003	7,9
OR6F1	9(18%)	6(4%)	0,003	5,2
CHST1	5(10%)	1(1%)	0,004	16,3
OXCT1	5(10%)	1(1%)	0,004	16,3
SLC17A3	5(10%)	1(1%)	0,004	16,3
SLC3A1	5(10%)	1(1%)	0,004	16,3
TTLL5	5(10%)	1(1%)	0,004	16,3

B

GENE SYMBOL	AA	EA	PVAL	OR
AGMO	0	14(10%)	0,023	0
CREBBP	0	12(8%)	0,04	0
KIAA1210	1(2%)	17(12%)	0,05	0,16

5.4 Discussion

In this chapter, we set out to identify whether there are differences in the number of mutations per sample and in the mutation frequency of any gene between cancer cells from the TCGA-LUAD AA and EA samples. The observed higher median number of mutations per sample in AA tumours compared to EA tumours was not statistically significant. Although some studies have reported a difference in mutational burden between AA and EA lung cancer tumours (Araujo et al., 2015a; Costa et al., 2021; Lusk et al., 2019; Steuer et al., 2016), our results align with other studies that have identified no such difference (Araujo et al., 2015b; Campbell et al., 2017). Differences in mutational burden between the two groups that have been observed in the past may in part be due to the fact that among the AA population there is generally a greater diversity of single nucleotide polymorphisms, and more genetic variation than among the EA population group (Hughes et al., 2008).

While the mutation frequency for most genes was consistent between the two groups, we identified 101 genes with differing mutation frequencies between them. We found that none of these genes with differing mutation frequencies between the two groups were part of the KEGG NSCLC gene list, indicating that the frequency of mutation of cancer-related genes is similar between the two groups.

To identify which of the differentially mutated genes may contribute to biological differences between the AA And EA groups, we isolated genes that were both differentially mutated and SDE (from chapter 3) between the two groups. Four genes matched this criteria: *UGT2B11*, *PCDHB6*, *KIAA1210* and *NTNG2*.

NTNG2 had a greater mutation rate among the AA group than the EA group (odds ratio = 8.1, p-value = 0.01), and its gene expression was significantly upregulated (log2-transformed fold change = 1.2, adjusted p-value = 0,0056) in the AA group relative to the EA group. *NTNG2* encodes a netrin protein and its differential expression has been observed in a number of cancers, including lung cancer (Hao et al., 2020). In a pan cancer study looking into the activity of Netrins in cancer, Hao et al. (2020) postulated that *NTNG1* and *NTNG2* are potential diagnostic biomarkers markers and therapeutic targets because of their association with multiple cancers and involvement in drug response pathways. The researchers found that *NTNG2* is strongly involved in cell cycle arrest, the inhibition of estrogen, and the activation of the epithelial-mesenchymal transition. Genes implicated in these pathways also emerged as probable contributors to the biological disparities between the two groups in our differential gene expression and enrichment analyses. This highlights the potential significance of the differential mutation and expression of *NTNG2* between the groups.

UGT2B11 had a greater mutation frequency in AA vs. EA tumours (odds ratio = 6.1, p-value = 0.008), and its gene expression was significantly downregulated (log2-transformed fold change = -2.7, adjusted p-value = 0.012) in the AA relative to the EA group. Interestingly this gene, as well as other *UDP glucuronosyltransferase* (*UGT*) family members, formed part of the xenobiotics and hormone regulation-related gene sets that were enriched among our SDE genes. Considering the vital role of *UGT* proteins in modifying steroid hormones and drugs in order to abolish their biological activity and facilitate their excretion from the body, we postulate that the differential expression of this group of genes would be central in creating differences in the tumour microenvironment and response to drugs of the AA vs. EA tumours. The differential mutation rate and differential expression of *UGT2B11* between the two groups makes this gene a likely contributor to clinically relevant biological differences that may exist between these two groups.

The *PCDHB6* gene is mutated significantly more frequently in the AA group than in the EA group (odds ratio = 4.7, p-value = 0.01), and its expression is also significantly lower in the AA group (log2-transformed fold change = -1, adjusted p-value = 0.0016). The *PCDHB6* gene forms part of the protocadherin beta gene cluster which encodes cadherin like cell adhesion proteins that likely play an important role in cell-cell neural connections (Frank and Kemler, 2002). A pre-print describing a study by Huang et al. (2022) associated increased expression of *PCDHB6* with poor prognosis in gastric cancer, as well as an increased level of cancer fibroblast invasion in LUADs and a number of other cancers. In chapter 3.4 we postulated that the decreased expression of certain genes in the cell-cell adhesion gene sets in AAs vs EAs, as well as the decreased expression of the mesenchyme development gene set in AAs vs EAs may relate to the AA tumours being less mesenchymal than EA tumours. Since cancer fibroblast invasion is linked to the epithelial-mesenchymal transition of cancer cells, the association that Huang et al. (2022) found between increased *PCDHB6* expression and cancer fibroblast invasion in LUADs, as well as the decreased gene expression of *PCDHB6* in AAs vs. EAs that we observe, support our theory that the AA tumours in this study are less mesenchymal than the EA tumours. The greater mutation rate of this gene in AAs may contribute to the difference in expression of this gene between the two groups.

KIAA1210 is mutated more frequently in EAs (odds ratio = 0.16, p-value = 0.05) and exhibits significantly lower expression in the AA group compared to the EA group (log2-transformed fold change = -1.2, adjusted p-value = 0.038). This gene is primarily expressed in the testis, where it is an important cell junction protein involved in cell-cell adhesion (Iwamori et al., 2017). *KIAA1210* hasn't been implicated in cancer genetics research. In this case, its involvement in cell-cell adhesion, and differential expression and mutation frequency between the AA and EA groups means it may play a

role in the more epithelial and less mesenchymal nature of AA tumour cells than EA that we postulate in this study.

Overall, despite the largely similar mutational landscape between the AA and EA groups, several genes that displayed both differential mutation and significant differential expression between the groups emerged as potential drivers of biological differences in tumours from these populations.

6. Conclusion

Cancer genetics research has facilitated an improved molecular understanding of cancer which has allowed for better treatment recommendation and novel treatment development to better deal with cancer. The majority of this research has been done on cohorts that are predominantly of European ancestry. Subtle differences in the genetic traits of populations of different ethnic groups, means that cancer associated variants specific to other ethnic groups (i.e., African and Asian ancestry) may be overlooked in research that is biased toward any particular ethnic group. Understanding what types of differences, if any, exist in cancer genetics between different ethnic groups has recently become a research topic of interest that aims to identify cross ethnicity, and ethnicity specific genetic variants to improve clinical interventions against cancer.

In this study, we aimed to identify if any genetic differences exist between lung adenocarcinoma (LUAD) from patients of African ancestry (AA) and European ancestry (EA), and if these differences are potentially biologically and clinically meaningful. While there was substantial genetic similarity in LUAD cells between the two groups, we identified genes and gene sets that exhibited differential expression, distinct copy number alterations, or varied mutation frequencies between the two populations. The main biological processes through which we predict these genes may result in differences in tumour biology between these two groups are the metabolism of xenobiotics (i.e. drugs) and hormones, epithelial to mesenchymal transition (EMT), and mitochondrial energy production.

The upregulation of mitochondrial energy production related genes in AA samples potentially points toward a fundamental difference in the glucose metabolism and energy production of the AA tumours vs. the EA tumours. This result suggests that the EA tumours utilise the traditional aerobic glycolysis method of energy production in cancer cells more than the AA samples, which in this case rely more on mitochondrial energy production. The change in glucose metabolism is a hallmark of cancer cell development, so such a difference in the biology of the cancers would not be a trivial one.

We observed differences in gene expression and mutation patterns within genes associated with the cellular metabolism of xenobiotics and the regulation of hormone levels. The metabolism of drugs targeting cancer cells can influence the cell's sensitivity or resistance to these drugs. Population level differences in the activity of drug metabolising genes may therefore influence the efficacy of drugs in different populations. Hormone levels in a tumour’s microenvironment affect the tumours

proliferation, growth, and sensitivity to treatments. Differences in hormone levels that a tumour is exposed to and the way it responds to these hormones alter the pathways it relies on for growth and proliferation, and therefore affect the types of treatment strategies suitable to kill the tumour.

The downregulated expression and differential mutation rate of certain genes suggested a difference in EMT between the two groups. EMT is a hallmark of cancer development, and is primary for the development of metastasis in cancers. Genetic differences between the two groups in this study suggest that the AA tumours are more epithelial and less mesenchymal than the EA tumours. Given that we ensured there was no difference in the metastasis of cancers between the two groups, this result suggests that AA tumours may rely on a slightly different mechanism to establish metastasis compared to EA tumours. Whatever the case, the difference in genetic traits relating to this essential process in cancer development is an interesting one.

In future, certain adjustments to the study design may help further validate the results and provide more robust findings. Although genetic differences can be useful markers of differences in biology between tumours, further support for the biological significance of these genetic changes can be gained though the identification of concordant protein expression differences between the groups of interest. Furthermore, a larger sample size, particularly for the AA group, would provide more statistical power to the analysis, and be able to detect genetic variations that are less frequent between the two groups. Since clinically relevant population level differences in disease associated genetic traits often come from rare variants, increasing sample size to improve detection of rare variants will prove useful. Additionally, the use of ancestry informative genetic markers instead of self-reported race to separate samples into AA and EA will help better separate samples by genetic ancestry, and the controlling for more exposure/environmental differences between the two groups will reduce confounding in the analysis. The ability to implement the mentioned improvements in a study is often limited when conducting a secondary analysis of previously collected data. To maximize the robustness of gene-disease associations derived from secondary analyses, primary data collectors should explore feasible ways to enhance information collection, thereby improving the utility of the data. In turn, researchers conducting secondary analyses should take into account the aforementioned considerations when selecting a sample source. They may also consider the option of combining data from multiple sources to bolster the strength of their analysis.

The study's findings regarding the genetic differences between LUAD patients of AA and EA offer valuable insights into the complex interplay of cancer genetics and ethnicity. While the potential implications for early detection and tailored therapies are intriguing, it is important not to jump toward proposing ethnicity-specific cancer therapies based solely on genetic differences as cancer

stems from a complex combination of the biological and environmental factors that an individual is subject to. Instead, the observed differences in cancer genetics should prompt a broader discussion on how to incorporate ethnic diversity into cancer research and treatment paradigms. Backed by findings such as those communicated in this study, these discussions should develop a level of awareness and sensitivity to ethnicity-based genetic differences in cancers to encourage more inclusive research, and a consideration of ethnicity based biological differences when recommending treatment regimens. Considering the nascent nature of research associating the difference in cancer genetics between ethnic groups to differences in the efficacy of certain cancer treatment regimens, the use of ethnicity-based cancer genetic differences to inform differential treatment regimens should not be prescriptive, but should rather be a point of consideration that clinicians are aware of and can draw on when deciding on the best treatments for a patient. With time, the information gleaned from diverse populations should inform a more comprehensive understanding of cancer genetics that can inform more personalised and effective means of detecting and treating cancer based on both the biological characteristics (including ethnicity) and environmental exposures of a patient. A focus on engagement with relevant research and clinical stakeholders is necessary to facilitate a diversity aware approach to cancer genetics research and cancer treatment to allow for the continuous improvement in the efficacy of personalised cancer treatment regimens for all.

Overall, we identified genetic differences between AA and EA LUADs from the TCGA-LUAD study. The involvement of these genes in critical cellular functions, tumour processes, and drug response pathways underscores the potential clinical relevance of their differences between the two groups. These findings highlight the need for the representation of adequate numbers of different ancestral populations in cancer genetic studies to better understand the overlap and differences in cancer-genetics associations between different groups. It is through such studies that the validation or disapproval of the differences we identified will be possible. Furthermore, these studies will contribute to both our general understanding of lung cancer genetics and our understanding of how these genetic factors may vary across different populations.

Supplementary tables

Supplementary table 1. Clinical information of samples before and after PSM, which considered patient age and tobacco smoking history.

		Unmatched Cohort		*1:2 propensity matched*		*1:3 propensity matched*		*1:4 propensity matched*	
	AFR Ancestry	**EUR Ancestry**	**P-value**	**EUR ancestry**	**P-value**	**EUR**	**P-value**	**EUR**	**P-value**
Patients	49	371 (88.3)	0.0003	98(66.7%)	0.89	147(75%)	0.9	196(80%)	0.47
Age (mean)	60	65.8		59.8		60.5		61.3	
Sex			0.98		0.86		0.87		0.77
Male	21(42.9%)	164(44.2%)		45(45.9%)		67(45.6%)		91(46.4%)	
Female	28(57.1%)	207(55.8%)		53(54.1%)		80(54.4%)		105(53.6%)	
Pathologic T stage			0.613		0.55		0.88		0.62
T1	22(44.9%)	135(36.4%)		38(38.8%)		61(41.5%)		75(38.3%)	
T2	20(40.8%)	191(51.5%)		51(52%)		69(46.9%)		99(50.5%)	
T3	6(12.2%)	33(8.9%)		7(7.1%)		15(10.2%)		17(8.7%)	
T4	1(2%)	10(2.7%)		2(2%)		2(1.4%)		5(2.6%)	
TX	0	2(0.5%)		0		0		0	
Tobacco smoking history			0.016		0.8		0.74		0.56
1	3(6.1%)	60(16.17%)		8(8.2%)		12(8.2%)		18(9.2%)	
2	17(34.7%)	83(22.4%)		41(41.8%)		58(39.5%)		64(32.7%)	
3	10(20.4%)	100(27%)		15(15.3%)		25(17%)		36(18.4%)	
4	17(34.7)	126(34%)		32(32.7%)		50(34%)		76(38.8%)	
5	2(4.1%)	2(0.5%)		2(2%)		2(1.4%)		2(0.01%)	
Pathologic N stage			0.63		0.71		0.61		0.65
N0	29(59.2%)	253(68.2%)		67(68.4%)		102(69.4%)		136(69.4%)	
N1	12(24.5%)	59(15.9%)		20(20.4%)		26(17.7%)		34(17.3%)	
N2	7(14.3%)	50(13.5%)		10(10.2%)		17(11.6%)		23(11.7%)	
N3	0	1(0.3%)		0		1(0.7%)		1(0.5%)	
NX	1(2%)	8(2.2%)		1(1%)		1(0.7%)		2(1%)	
Pathologic M stage			0.178		0.24		0.09		0.12
M0	25(51%)	241(65%)		62(63.3%)		101(68.7%)		131(66.8%)	
M1	2(4.1%)	16(4.3%)		8(8.2%)		8(5.4%)		10(5.1%)	
MX	21(42.9%)	112(30.2%)		27(27.6%)		37(25.2%)		54(27.6%)	
Unknown	1(2%)	2(0.5%)		1(1%)		1(0.7%)		1(0.5%)	
Mean p-value			0.403		0.68		0.68		0.53

Supplementary Table 2 Clinical information of samples before and after PSM, which considered patient age, sex, pathologic T stage, pathologic N stage, pathologic M stage and tobacco smoking history.

		Unmatched Cohort		*Propensity-matched 1:2*		*Propensity- matched 1:3*		*Propensity Matched 1:4*	
	AFR Ancestry	**EUR Ancestry**	**P-value**	**EUR ancestry**	**P-value**	**EUR**	**P-value**	**EUR**	**P-value**
Patients	49 (11.7%)	371 (88.3)		98(66.7%)		147(75%)		196(80%)	
Age (mean)	60	65.8	0.0003	58.8	0.49	60.2	0.93	61.7	0.33
Sex			0.98		1		1		0.82
Male	21(42.9%)	164(44.2%)		42(42.9%)		62(42.2%)		90(45.9%)	
Female	28(57.1%)	207(55.8%)		56(57.1%)		85(57.8%)		106(54.1%)	
Pathologic T stage		0.613			0.89		0.95		0.93
T1	22(44.9%)	135(36.4%)		38(38.8%)		60(40.8%)		81(41.3%)	
T2	20(40.8%)	191(51.5%)		45(45.9%)		66(44.9%)		90(45.9%)	
T3	6(12.2%)	33(8.9%)		12(12.2%)		17(11.6%)		22(11.2%)	
T4	1(2%)	10(2.7%)		3(3.1%)		4(2.7%)		3(1.5%)	
TX	0	2(0.5%)		0		0		0	
Tobacco smoking history			0.016		0.51		0.72		0.55
1	3(6.1%)	60(16.17%)		10(10.2%)		14(9.5%)		19(9.7%)	
2	17(34.7%)	83(22.4%)		36(36.7%)		54(36.7%)		61(31.1%)	
3	10(20.4%)	100(27%)		11(11.2%)		25(17%)		42(21.4%)	
4	17(34.7)	126(34%)		39(39.8%)		52(35.4%)		72(36.7%)	
5	2(4.1%)	2(0.5%)		2(2%)		2(1.4%)		2 (1%)	
Pathologic N stage			0.63		0.9		0.98		0.98
N0	29(59.2%)	253(68.2%)		54(55.1%)		84(57.1%)		116(59.2%)	
N1	12(24.5%)	59(15.9%)		23(23.5%)		35(23.8%)		45(30%)	
N2	7(14.3%)	50(13.5%)		19(19.4%)		25(17%)		32(16.3%)	
N3	0	1(0.3%)				0		0	
NX	1(2%)	8(2.2%)		2(2%)		3(2%)		3(2%)	
Pathologic M stage			0.178		0.83		0.71		0.73
M0	25(51%)	241(65%)		57(58.2%)		82(55.8)		105(53.6%)	
M1	2(4.1%)	16(4.3%)		4(4.1%)		9(6.1%)		10(5.1%)	
MX	21(42.9%)	112(30.2%)		36(36.7%)		55(37.4%)		80(40.8%)	
Unknown	1(2%)	2(0.5%)		1(1%)		1(0.68%)		1(0.5%)	
Mean p-value			0.403		0.77		0.88		0.72

Supplementary table 3. List of SDE genes with their log2-transformed fold change and adjusted p-values. Results show differential expression in AA samples relative to EA samples.

Entrez gene id	Gene symbol	Log2-transformed FoldChange	Adjusted p-value
1015	*CDH17*	-4	0,000000000000002
793	*CALB1*	-4,3	0,0000000003
1415	*CRYBB2*	1,8	0,000000001
5320	*PLA2G2A*	-2,5	0,00000005
462	*SERPINC1*	1,7	0,000000065
8788	*DLK1*	-5	0,00000073
3929	*LBP*	3,3	0,0000012
23145	*SSPOP*	1,8	0,0000016
441273	*SPDYE1*	1,5	0,0000016
192683	*SCAMP5*	1	0,0000018
1088	*CEACAM8*	-3,2	0,000002
23615	*PYY2*	1,5	0,0000024
378825	*LINC00162*	-3,5	0,000006
3642	*INSM1*	-2,8	0,0000066
131831	*FAM194A*	1,4	0,0000072
4922	*NTS*	-2,9	0,000015
152573	*SHISA3*	-2,1	0,000018
5913	*RAPSN*	1,6	0,000019
7499	*XG*	-1,9	0,000024
119	*ADD2*	-1,6	0,000024
4760	*NEUROD1*	-3,9	0,000026
126147	*NTN5*	1,3	0,000026
285596	*FAM153A*	1,9	0,000026
312	*ANXA13*	-2,6	0,000033
122402	*TDRD9*	-1,9	0,000033
148170	*CDC42EP5*	1,2	0,000039
191585	*PLAC4*	2,1	0,000045
2953	*GSTT2*	1,2	0,000045
25837	*RAB26*	1,5	0,000058
7490	*WT1*	-1,9	0,000081
440307	*TTLL13*	1	0,000082
389383	*CLPSL2*	2,2	0,00012
130013	*ACMSD*	-2,5	0,00012
10326	*SIRPB1*	-1,3	0,00012
204801	*NLRP11*	2,6	0,00012
3299	*HSF4*	1,2	0,00012
1555	*CYP2B6*	1,8	0,00014
56670	*SUCNR1*	-1,4	0,00014
2837	*UTS2R*	2,2	0,00016
339674	*LINC00634*	1,4	0,00016
126248	*WDR88*	1	0,00017
162632	*null*	-1,2	0,00017
79853	*TM4SF20*	-3,3	0,00017
9421	*HAND1*	-2,8	0,00018
100134938	*UPK3BL*	1,1	0,00018
728882	*FAM182A*	1,3	0,00018
728875	*RP11-640M9.1*	1,1	0,00018
4588	*MUC6*	2,1	0,00019
284359	*IZUMO1*	1,5	0,00021
415	*ARSE*	-1,5	0,00028
3293	*HSD17B3*	1,6	0,00028
66002	*CYP4F12*	1,5	0,00028
797	*CALCB*	-3,4	0,00029
152404	*IGSF11*	1,4	0,00029
7857	*SCG2*	-1,5	0,00029

Entrez gene id	Gene symbol	Log2-transformed FoldChange	Adjusted p-value
401105	*FLJ42393*	-1,9	0,00029
118430	*MUCL1*	2,5	0,00029
221481	*ARMC12*	1	0,00032
2984	*GUCY2C*	-1,6	0,00033
389812	*LCN15*	-5,6	0,00036
202134	*FAM153B*	1,8	0,00036
6328	*SCN3A*	-1,8	0,00038
116285	*ACSM1*	1,3	0,00038
7448	*VTN*	-1,4	0,00039
3598	*IL13RA2*	-1,8	0,00046
163688	*CALML6*	1,5	0,00047
386617	*KCTD8*	-2,3	0,00048
92291	*CAPN13*	1,4	0,00052
6553	*SLC9A5*	1	0,00055
155185	*AMZ1*	-1,1	0,00055
100127888	*RP11-93B14.5*	1,6	0,00057
6555	*SLC10A2*	-2,9	0,00059
8345	*HIST1H2BH*	1,7	0,00063
57467	*HHATL*	-2,4	0,00063
2147	*F2*	-3,4	0,00066
116534	*MRGPRE*	-2,3	0,00067
126129	*CPT1C*	1,1	0,00069
124773	*C17orf64*	1,5	0,00073
6326	*SCN2A*	-1,6	0,00077
8928	*FOXH1*	1,3	0,0008
126006	*PCP2*	1,2	0,0008
644172	*AC139677.11*	-1,1	0,0008
374286	*CDRT1*	1,3	0,00083
10517	*FBXW10*	1,5	0,00085
54798	*DCHS2*	1,4	0,00088
158521	*FMR1NB*	3,1	0,0009
84848	*MIR503HG*	1	0,0009
8912	*CACNA1H*	-1,2	0,00092
8685	*MARCO*	-1,5	0,00092
349196	*LINC00965*	-1,1	0,00093
84889	*SLC7A3*	-2,3	0,001
56062	*KLHL4*	-1,1	0,001
54474	*KRT20*	-3	0,001
100130958	*SYCE1L*	1,1	0,0011
3960	*LGALS4*	-1,8	0,0011
1638	*DCT*	1,6	0,0012
51233	*C22orf43*	1,3	0,0012
54898	*ELOVL2*	-1,2	0,0012
345557	*PLCXD3*	-1,4	0,0012
5179	*PENK*	-1,6	0,0013
2326	*FMO1*	-1	0,0013
816	*CAMK2B*	-1,6	0,0013
3158	*HMGCS2*	-2,1	0,0013
57007	*ACKR3*	1	0,0014
100144604	*LINC00930*	1,5	0,0015
460	*ASTN1*	-1,7	0,0016
56130	*PCDHB6*	-1	0,0016
55600	*ITLN1*	-2	0,0018
9947	*MAGEC1*	-2,8	0,0018
9543	*IGDCC3*	-1,6	0,0018
1294	*COL7A1*	1,3	0,0019
414060	*TBC1D3*	1,1	0,0019
5950	*RBP4*	-1,4	0,0019

Entrez gene id	Gene symbol	Log2-transformed FoldChange	Adjusted p-value
9048	*ARTN*	1,1	0,0021
290	*ANPEP*	-1,1	0,0024
1826	*DSCAM*	-1,7	0,0024
340508	*RP11-520B13.4*	1,9	0,0024
360226	*PRSS41*	-3,2	0,0024
285696	*AC091878.1*	1,5	0,0024
199675	*C19orf59*	-1,6	0,0025
4160	*MC4R*	1,5	0,0027
51237	*MZB1*	1,1	0,0027
57451	*TENM2*	-1,3	0,0029
3990	*LIPC*	-1,2	0,003
1404	*HAPLN1*	-1,3	0,003
139599	*MAGEE2*	-1,5	0,003
147111	*NOTUM*	1,4	0,0031
728215	*FAM155A*	-1,1	0,0031
2167	*FABP4*	-1,8	0,0031
387644	*LINC00202-1*	1,1	0,0034
152330	*CNTN4*	-1,2	0,0035
3008	*HIST1H1E*	1,1	0,0035
222546	*RFX6*	-3	0,0036
5593	*PRKG2*	-1,3	0,0037
479	*ATP12A*	1,8	0,0037
6664	*SOX11*	-1,7	0,0038
6769	*STAC*	-1,1	0,0039
29953	*TRHDE*	-1,7	0,0041
56969	*RPL23AP32*	-1,2	0,0041
653499	*LGALS7*	1,9	0,0042
85009	*MGC16025*	1	0,0043
10255	*HCG9*	1,5	0,0043
154197	*PNLDC1*	1,5	0,0045
2201	*FBN2*	-1,1	0,0048
4317	*MMP8*	-1,4	0,0049
51066	*SSUH2*	1,2	0,0051
352999	*C6orf58*	2,2	0,0051
727940	*RHOXF2B*	2,9	0,0052
51617	*NSG2*	-1,7	0,0053
6043	*EIF4A2*	1,4	0,0053
400224	*PLEKHD1*	1,4	0,0054
84628	*NTNG2*	1,2	0,0056
5801	*PTPRR*	-1,1	0,0056
1029	*CDKN2A*	1,3	0,0056
96626	*LIMS3*	-1	0,0059
57526	*PCDH19*	-1,1	0,006
9464	*HAND2*	-2,1	0,006
169166	*SNX31*	1,5	0,006
128864	*C20orf144*	1,5	0,0062
84623	*KIRREL3*	-1,2	0,0064
4504	*MT3*	1,3	0,0064
3481	*IGF2*	-1,3	0,0064
27063	*ANKRD1*	-1,7	0,0066
200010	*SLC5A9*	-1,3	0,0066
341208	*HEPHL1*	-1,6	0,0067
116931	*MED12L*	1,1	0,0067
348980	*HCN1*	-1,7	0,0067
125115	*KRT40*	-1,8	0,0067
405753	*DUOXA2*	1,1	0,0067
257629	*ANKS4B*	-2,3	0,0068
732	*C8B*	-1,7	0,0068

Entrez gene id	Gene symbol	Log2-transformed FoldChange	Adjusted p-value
3078	*CFHR1*	-2	0,0068
8190	*MIA*	1,1	0,0068
677847	*SNORA81*	1,2	0,0069
9271	*PIWIL1*	-2,3	0,0069
114770	*PGLYRP2*	1,1	0,007
4360	*MRC1*	-1	0,0071
1813	*DRD2*	1,4	0,0073
125972	*CALR3*	1,3	0,0073
643236	*TMEM72*	-1,8	0,008
389058	*SP5*	1,1	0,0081
6003	*RGS13*	-1	0,0081
6528	*SLC5A5*	1,6	0,0082
8424	*BBOX1*	1,3	0,0083
8707	*B3GALT2*	-1,2	0,0086
1821	*DRP2*	-1,1	0,0087
1592	*CYP26A1*	-1,4	0,0088
563	*AZGP1*	-1,5	0,0088
158830	*CXorf65*	1,1	0,0088
781	*CACNA2D1*	-1,1	0,0089
84658	*EMR3*	-1,2	0,0093
7425	*VGF*	-1,4	0,0094
117156	*SCGB3A2*	-1,5	0,0094
58512	*DLGAP3*	1	0,0096
57453	*DSCAML1*	-1,1	0,0097
167681	*PRSS35*	-1,1	0,0097
222171	*PRR15*	1,1	0,0098
401562	*LCNL1*	1,1	0,0098
644974	*ALG1L2*	1,3	0,0099
10882	*C1QL1*	-1,2	0,01
108	*ADCY2*	-1,1	0,011
57864	*SLC46A2*	-1,2	0,011
646	*BNC1*	-1,4	0,011
1000	*CDH2*	-1,2	0,011
29118	*DDX25*	1,5	0,011
64102	*TNMD*	-2,3	0,011
51352	*WT1-AS*	-1,5	0,011
25758	*KIAA1549L*	-1	0,011
10720	*UGT2B11*	-2,7	0,012
284111	*SLC13A5*	-1,2	0,012
3237	*HOXD11*	-2,7	0,012
360205	*PRAC2*	-2,4	0,012
1991	*ELANE*	-1,5	0,012
6663	*SOX10*	1,2	0,012
652995	*UCA1*	-1,5	0,013
1369	*CPN1*	-4,2	0,013
64090	*GAL3ST2*	1,3	0,013
1551	*CYP3A7*	-1,5	0,014
270	*AMPD1*	1,1	0,014
846	*CASR*	-1,6	0,014
57094	*CPA6*	-1,4	0,015
93978	*CLEC6A*	-1,2	0,015
2354	*FOSB*	-1,2	0,015
84570	*COL25A1*	-1,8	0,015
285852	*TREML4*	1,2	0,015
54457	*TAF7L*	-1,4	0,015
1299	*COL9A3*	1	0,016
10917	*BTNL3*	-1,9	0,016
148823	*GCSAML*	-1	0,016

Entrez gene id	Gene symbol	Log2-transformed FoldChange	Adjusted p-value
1776	*DNASE1L3*	-1,2	0,016
388951	*TSPYL6*	1,2	0,017
54456	*MOV10L1*	1	0,017
56163	*RNF17*	1,3	0,017
7032	*TFF2*	2,8	0,017
3222	*HOXC5*	-1,1	0,017
5276	*SERPINI2*	1,2	0,017
56147	*PCDHA1*	-1,2	0,017
131034	*CPNE4*	1,1	0,017
414062	*CCL3L1*	1,1	0,017
112609	*MRAP2*	1,1	0,018
5083	*PAX9*	1	0,018
64850	*ETNPPL*	-2,8	0,018
2566	*GABRG2*	-3,2	0,018
4892	*NRAP*	1,3	0,018
645528	*LINC00264*	1,1	0,018
81551	*STMN4*	1,6	0,018
4319	*MMP10*	1,4	0,018
26577	*PCOLCE2*	-1,1	0,019
730130	*TMEM229A*	-2	0,019
1261	*CNGA3*	-1,5	0,019
218	*ALDH3A1*	1,4	0,02
7363	*UGT2B4*	-1,4	0,02
339977	*LRRC66*	1	0,02
729522	*AACSP1*	1,4	0,02
389197	*C4orf50*	1,2	0,02
6278	*S100A7*	-2,4	0,02
165257	*C1QL2*	-1,8	0,02
341350	*OVCH1*	-1,8	0,02
425054	*VCX3B*	1,8	0,02
2835	*GPR12*	1,7	0,02
285180	*RUFY4*	1	0,021
5169	*ENPP3*	1,2	0,021
10223	*GPA33*	-1,2	0,021
2244	*FGB*	-2	0,022
6540	*SLC6A13*	-1,2	0,022
317772	*HIST2H2AB*	1,3	0,022
2571	*GAD1*	1,1	0,022
401565	*FAM166A*	1,2	0,022
7364	*UGT2B7*	-1,2	0,023
6474	*SHOX2*	1,1	0,023
2069	*EREG*	-1,4	0,023
3381	*IBSP*	-1,4	0,023
63970	*TP53AIP1*	1,1	0,023
5288	*PIK3C2G*	-1,6	0,024
54658	*UGT1A1*	-2,3	0,024
643763	*NKAIN3-IT1*	-1,5	0,024
7125	*TNNC2*	1,2	0,024
64220	*STRA6*	1,1	0,024
677814	*SNORA31*	1,7	0,024
29986	*SLC39A2*	1,1	0,024
84691	*FAM71F1*	1,3	0,024
7015	*TERT*	1,1	0,024
53904	*MYO3A*	1,3	0,024
404203	*SPINK6*	-1,8	0,025
575	*BAI1*	-1,1	0,025
547	*KIF1A*	1,7	0,025
2562	*GABRB3*	-1,2	0,025

Entrez gene id	Gene symbol	Log2-transformed FoldChange	Adjusted p-value
286749	*STON1-GTF2A1L*	-1,8	0,026
29991	*OBP2A*	-2,1	0,026
642636	*RAD21L1*	1,8	0,027
1576	*CYP3A4*	1,2	0,027
1178	*CLC*	-1,6	0,027
1401	*CRP*	1	0,027
7278	*TUBA3C*	2,8	0,027
170626	*XAGE3*	-2,5	0,028
6549	*SLC9A2*	1,2	0,028
56918	*C2orf83*	1,9	0,028
10642	*IGF2BP1*	-1,7	0,028
131405	*TRIM71*	-1,4	0,028
91948	*LINC00923*	1,3	0,028
2939	*GSTA2*	-1,4	0,028
84676	*TRIM63*	-1,3	0,028
147744	*TMEM190*	1,4	0,029
84530	*SRRM4*	-1,3	0,029
5473	*PPBP*	-1,3	0,029
6422	*SFRP1*	-1	0,029
554202	*MIR31HG*	1,3	0,029
2044	*EPHA5*	-1,3	0,029
29974	*A1CF*	-3,3	0,03
7069	*THRSP*	1	0,03
1577	*CYP3A5*	1	0,03
56311	*ANKRD7*	1	0,03
84634	*KISS1R*	1,2	0,03
143379	*C10orf82*	1,7	0,031
11166	*SOX21*	1,2	0,032
375567	*VWC2*	-1,2	0,032
388015	*RTL1*	-2,3	0,032
5063	*PAK3*	-1,3	0,032
145957	*NRG4*	1,1	0,032
181	*AGRP*	-1,2	0,033
5308	*PITX2*	-1,4	0,033
91464	*ISX*	-2,4	0,033
90226	*UCN2*	1	0,034
79190	*IRX6*	-1,1	0,034
150280	*HORMAD2*	1,2	0,035
51378	*ANGPT4*	-1,5	0,035
57554	*LRRC7*	1,7	0,035
429	*ASCL1*	-1,8	0,035
347376	*H3P44*	-1,4	0,036
140832	*WFDC10A*	1,7	0,036
91646	*TDRD12*	1,3	0,036
30813	*VSX1*	1,1	0,037
157739	*TDH*	1,2	0,037
186	*AGTR2*	-1,3	0,038
57481	*KIAA1210*	-1,2	0,038
8358	*HIST1H3B*	1,2	0,039
100192379	*PP12613*	1,4	0,039
7180	*CRISP2*	1,5	0,039
10117	*ENAM*	-1,1	0,04
29106	*SCG3*	-1,2	0,04
84944	*MAEL*	-1,4	0,04
441476	*C9orf173*	1,1	0,04
100303453	*TSNAX-DISC1*	1,5	0,041
192666	*KRT24*	-1,9	0,041
81033	*KCNH6*	-1,1	0,041

Entrez gene id	Gene symbol	Log2-transformed FoldChange	Adjusted p-value
54207	*KCNK10*	-1,3	0,042
400120	*SERTM1*	-1,7	0,042
79785	*RERGL*	-1,2	0,042
6861	*SYT5*	1,3	0,042
268	*AMH*	-1,1	0,042
1244	*ABCC2*	-1,2	0,042
2262	*GPC5*	-1,1	0,043
204962	*SLC44A5*	-1	0,043
3543	*IGLL1*	1,5	0,043
7652	*ZNF99*	-2	0,043
341032	*C11orf53*	-1,5	0,043
84665	*MYPN*	-1,4	0,044
89869	*PLCZ1*	1,8	0,044
340811	*AKR1CL1*	1,4	0,045
123346	*HIGD2B*	1,2	0,045
8715	*NOL4*	-1,1	0,046
374308	*PTCHD3*	1,5	0,046
11211	*FZD10*	1,1	0,046
347731	*LRRTM3*	-1,9	0,046
114131	*UCN3*	-1,3	0,047
57582	*KCNT1*	-1,1	0,048
196335	*OR56B4*	1,1	0,048
79782	*LRRC31*	-1,1	0,049
5672	*PSG4*	-2,7	0,049
56896	*DPYSL5*	-2,3	0,049
348840	*ANKRD18DP*	1	0,05

Supplementary table 4. Result of ORA against the KEGG Pathway database. Displayed are the KEGG gene-sets that are overrepresented among the list of SDE genes.

ID	Description	GeneRatio	p.adjust	qvalue
hsa00982	Drug metabolism - cytochrome P450	11/137	0,000005	0,0000048
hsa00980	Metabolism of xenobiotics by cytochrome P450	10/137	0,000061	0,000058
hsa00830	Retinol metabolism	9/137	0,000099	0,000094
hsa05204	Chemical carcinogenesis - DNA adducts	9/137	0,000099	0,000094
hsa00140	Steroid hormone biosynthesis	8/137	0,00033	0,00031
hsa04976	Bile secretion	8/137	0,0039	0,0037
hsa00983	Drug metabolism - other enzymes	7/137	0,011	0,01
hsa00860	Porphyrin metabolism	5/137	0,018	0,017
hsa00053	Ascorbate and aldarate metabolism	4/137	0,034	0,032

Supplementary table 5. Result of ORA against the GO-BP database. Displayed are the GO-BP gene-sets that are overrepresented among the list of SDE genes.

ID	Description	GeneRatio	p.adjust	qvalue
GO:0006805	xenobiotic metabolic process	12/294	0,00039	0,00036
GO:0071466	cellular response to xenobiotic stimulus	14/294	0,00065	0,00059
GO:0010817	regulation of hormone levels	24/294	0,0011	0,001
GO:0008210	estrogen metabolic process	7/294	0,0011	0,001
GO:0009410	response to xenobiotic stimulus	20/294	0,004	0,0037
GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	16/294	0,004	0,0037
GO:0048483	autonomic nervous system development	7/294	0,004	0,0037
GO:0034754	cellular hormone metabolic process	11/294	0,0041	0,0037
GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	12/294	0,0047	0,0044
GO:0003357	noradrenergic neuron differentiation	4/294	0,0055	0,005
GO:0034587	piRNA metabolic process	5/294	0,0068	0,0062
GO:0042445	hormone metabolic process	13/294	0,018	0,017
GO:0051591	response to cAMP	8/294	0,018	0,017
GO:0050808	synapse organization	18/294	0,024	0,023
GO:0070593	dendrite self-avoidance	4/294	0,024	0,023
GO:0007416	synapse assembly	11/294	0,025	0,023
GO:0010038	response to metal ion	16/294	0,025	0,023
GO:0006814	sodium ion transport	13/294	0,027	0,024
GO:0007189	adenylate cyclase-activating G protein-coupled receptor signaling pathway	10/294	0,027	0,024
GO:0042573	retinoic acid metabolic process	5/294	0,027	0,024
GO:0071692	protein localization to extracellular region	16/294	0,027	0,024
GO:0043046	DNA methylation involved in gamete generation	4/294	0,03	0,028
GO:0052695	cellular glucuronidation	4/294	0,03	0,028
GO:0035270	endocrine system development	9/294	0,03	0,028
GO:0006063	uronic acid metabolic process	4/294	0,038	0,035
GO:0019585	glucuronate metabolic process	4/294	0,038	0,035
GO:0030277	maintenance of gastrointestinal epithelium	4/294	0,038	0,035
GO:0048485	sympathetic nervous system development	4/294	0,038	0,035
GO:2000177	regulation of neural precursor cell proliferation	7/294	0,04	0,036
GO:0003211	cardiac ventricle formation	3/294	0,041	0,037
GO:0034776	response to histamine	3/294	0,041	0,037
GO:0061351	neural precursor cell proliferation	9/294	0,041	0,037
GO:0009306	protein secretion	15/294	0,041	0,037
GO:0035592	establishment of protein localization to extracellular region	15/294	0,041	0,037
GO:0007188	adenylate cyclase-modulating G protein-coupled receptor signaling pathway	12/294	0,042	0,038
GO:0060563	neuroepithelial cell differentiation	4/294	0,042	0,038
GO:0030073	insulin secretion	10/294	0,044	0,041
GO:0042178	xenobiotic catabolic process	4/294	0,045	0,042
GO:0015833	peptide transport	12/294	0,045	0,042
GO:0008202	steroid metabolic process	14/294	0,045	0,042
GO:0002065	columnar/cuboidal epithelial cell differentiation	7/294	0,045	0,042

GO:0048593	camera-type eye morphogenesis	8/294	0,049	0,045
GO:0030072	peptide hormone secretion	11/294	0,049	0,045
GO:0048592	eye morphogenesis	9/294	0,049	0,045
GO:0043010	camera-type eye development	14/294	0,049	0,045
GO:0046660	female sex differentiation	8/294	0,049	0,045
GO:0046683	response to organophosphorus	8/294	0,049	0,045
GO:0030325	adrenal gland development	4/294	0,05	0,046

Supplementary table 6. Top 50 (out of 84) enriched gene sets from GSEA against the KEGG Pathway database.

ID	Description	setSize	enrichmentScore	NES	p.adjust	qvalue
hsa00190	Oxidative phosphorylation	120	0,58	2,6	0,000000011	0,000000008
hsa05016	Huntington disease	290	0,45	2,2	0,000000011	0,000000008
hsa05012	Parkinson disease	252	0,43	2,1	0,000000011	0,000000008
hsa04510	Focal adhesion	198	-0,42	-2,2	0,000000025	0,000000017
hsa05014	Amyotrophic lateral sclerosis	346	0,38	1,9	0,00000003	0,00000002
hsa04010	MAPK signaling pathway	292	-0,36	-1,9	0,000000075	0,000000051
hsa04151	PI3K-Akt signaling pathway	340	-0,33	-1,8	0,00000019	0,00000013
hsa05205	Proteoglycans in cancer	200	-0,39	-2	0,00000089	0,00000061
hsa03050	Proteasome	46	0,64	2,4	0,0000012	0,00000078
hsa04810	Regulation of actin cytoskeleton	223	-0,37	-1,9	0,0000023	0,0000015
hsa05010	Alzheimer disease	366	0,35	1,8	0,0000033	0,0000022
hsa05020	Prion disease	258	0,38	1,9	0,0000035	0,0000024
hsa05022	Pathways of neurodegeneration - multiple diseases	456	0,34	1,7	0,0000038	0,0000026
hsa04611	Platelet activation	122	-0,43	-2,1	0,0000077	0,0000052
hsa03010	Ribosome	131	0,45	2	0,000011	0,0000074
hsa05208	Chemical carcinogenesis - reactive oxygen species	208	0,39	1,9	0,000018	0,000012
hsa04714	Thermogenesis	218	0,38	1,8	0,000018	0,000012
hsa04380	Osteoclast differentiation	127	-0,4	-1,9	0,000062	0,000042
hsa04932	Non-alcoholic fatty liver disease	150	0,41	1,9	0,00007	0,000048
hsa05150	Staphylococcus aureus infection	83	-0,45	-2	0,00011	0,000077
hsa05415	Diabetic cardiomyopathy	187	0,38	1,8	0,00016	0,00011
hsa04512	ECM-receptor interaction	87	-0,44	-2	0,0002	0,00014
hsa04610	Complement and coagulation cascades	83	-0,43	-1,9	0,00058	0,0004
hsa04022	cGMP-PKG signaling pathway	166	-0,35	-1,8	0,00084	0,00057
hsa04020	Calcium signaling pathway	237	-0,3	-1,6	0,00087	0,00059
hsa05144	Malaria	49	-0,5	-2	0,00099	0,00067
hsa05145	Toxoplasmosis	110	-0,39	-1,8	0,0013	0,00086
hsa04360	Axon guidance	181	-0,32	-1,7	0,0015	0,001
hsa05206	MicroRNAs in cancer	160	-0,33	-1,7	0,0015	0,001
hsa05140	Leishmaniasis	74	-0,41	-1,8	0,0024	0,0016
hsa04014	Ras signaling pathway	231	-0,3	-1,6	0,0024	0,0016
hsa05165	Human papillomavirus infection	320	-0,27	-1,5	0,0025	0,0017
hsa00040	Pentose and glucuronate interconversions	33	-0,54	-2	0,003	0,002

hsa04350	TGF-beta signaling pathway	93	-0,4	-1,8	0,003	0,002
hsa04727	GABAergic synapse	87	-0,4	-1,8	0,003	0,002
hsa04713	Circadian entrainment	97	-0,37	-1,7	0,0036	0,0024
hsa03040	Spliceosome	129	0,37	1,7	0,0039	0,0027
hsa04724	Glutamatergic synapse	113	-0,35	-1,7	0,0039	0,0027
hsa04520	Adherens junction	71	-0,4	-1,8	0,0049	0,0033
hsa05146	Amoebiasis	100	-0,36	-1,7	0,0049	0,0033
hsa04921	Oxytocin signaling pathway	152	-0,33	-1,6	0,0049	0,0033
hsa04145	Phagosome	148	-0,32	-1,6	0,0049	0,0033
hsa05152	Tuberculosis	168	-0,31	-1,6	0,005	0,0034
hsa04080	Neuroactive ligand-receptor interaction	336	-0,26	-1,4	0,0054	0,0037
hsa04933	AGE-RAGE signaling pathway in diabetic complications	100	-0,36	-1,7	0,0055	0,0037
hsa04514	Cell adhesion molecules	148	-0,32	-1,6	0,0056	0,0038
hsa03410	Base excision repair	33	0,53	1,9	0,0056	0,0038
hsa04976	Bile secretion	85	-0,39	-1,8	0,0062	0,0042
hsa04614	Renin-angiotensin system	23	-0,58	-2	0,007	0,0047
hsa04974	Protein digestion and absorption	101	-0,36	-1,7	0,0075	0,0051

Supplementary table 7. Top 50 (out of 976) enriched gene sets from GSEA against the GO-BP database.

ID	Description	setSize	Enrichment Score	NES	p.adjust	qvalue
GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	275	-0,47	-2,5	0,000000000000002	0,000000000000002
GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	168	-0,54	-2,7	0,00000000000002	0,00000000000002
GO:0045333	cellular respiration	212	0,48	2,3	0,0000000002	0,0000000002
GO:0009060	aerobic respiration	168	0,51	2,3	0,0000000009	0,0000000007
GO:0034329	cell junction assembly	411	-0,36	-2	0,0000000009	0,0000000007
GO:0042773	ATP synthesis coupled electron transport	76	0,62	2,5	0,000000004	0,000000003
GO:0042775	mitochondrial ATP synthesis coupled electron transport	76	0,62	2,5	0,000000004	0,000000003
GO:0022904	respiratory electron transport chain	97	0,58	2,5	0,000000006	0,000000005
GO:0060485	mesenchyme development	283	-0,39	-2,1	0,000000011	0,000000009
GO:0050808	synapse organization	415	-0,35	-1,9	0,000000012	0,000000009
GO:0015980	energy derivation by oxidation of organic compounds	297	0,41	2	0,000000012	0,000000009
GO:0010257	NADH dehydrogenase complex assembly	53	0,68	2,6	0,000000013	0,00000001
GO:0032981	mitochondrial respiratory chain complex I assembly	53	0,68	2,6	0,000000013	0,00000001
GO:0033108	mitochondrial respiratory chain complex assembly	87	0,59	2,4	0,000000013	0,00000001
GO:0006119	oxidative phosphorylation	121	0,54	2,4	0,000000019	0,000000014
GO:0042060	wound healing	414	-0,35	-1,9	0,000000021	0,000000016
GO:0007178	transmembrane receptor protein serine/threonine kinase signaling pathway	346	-0,36	-1,9	0,000000043	0,000000033
GO:0019646	aerobic electron transport chain	68	0,62	2,4	0,000000052	0,000000039
GO:0040013	negative regulation of locomotion	313	-0,36	-1,9	0,00000012	0,000000093
GO:0002274	myeloid leukocyte activation	225	-0,4	-2,1	0,00000016	0,00000013
GO:0001667	ameboidal-type cell migration	411	-0,34	-1,9	0,00000018	0,00000014
GO:0051271	negative regulation of cellular component movement	286	-0,36	-1,9	0,00000019	0,00000014

GO:0030336	negative regulation of cell migration	263	-0,37	-2	0,0000002	0,00000016
GO:0043062	extracellular structure organization	300	-0,36	-1,9	0,0000002	0,00000016
GO:0045229	external encapsulating structure organization	302	-0,36	-1,9	0,00000021	0,00000016
GO:0006399	tRNA metabolic process	185	0,45	2,1	0,00000022	0,00000017
GO:0061564	axon development	474	-0,32	-1,8	0,00000022	0,00000017
GO:0030198	extracellular matrix organization	299	-0,36	-1,9	0,0000003	0,00000023
GO:2000146	negative regulation of cell motility	278	-0,37	-1,9	0,0000005	0,00000038
GO:0032103	positive regulation of response to external stimulus	419	-0,32	-1,8	0,00000054	0,00000041
GO:0007409	axonogenesis	429	-0,32	-1,8	0,00000071	0,00000054
GO:0046034	ATP metabolic process	252	0,4	2	0,00000082	0,00000063
GO:0090287	regulation of cellular response to growth factor stimulus	284	-0,35	-1,9	0,00000082	0,00000063
GO:0008015	blood circulation	487	-0,3	-1,7	0,00000082	0,00000063
GO:0001503	ossification	402	-0,32	-1,8	0,00000084	0,00000064
GO:0048762	mesenchymal cell differentiation	232	-0,38	-2	0,0000009	0,00000069
GO:0050900	leukocyte migration	372	-0,33	-1,8	0,00000098	0,00000075
GO:0001655	urogenital system development	350	-0,33	-1,8	0,00000099	0,00000075
GO:0071559	response to transforming growth factor beta	240	-0,37	-2	0,000001	0,0000008
GO:0045765	regulation of angiogenesis	276	-0,36	-1,9	0,0000011	0,0000008
GO:0030900	forebrain development	366	-0,33	-1,8	0,0000011	0,00000081
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	42	0,68	2,5	0,0000012	0,0000009
GO:1901342	regulation of vasculature development	280	-0,36	-1,9	0,0000012	0,0000009
GO:0022900	electron transport chain	151	0,46	2,1	0,0000012	0,00000095
GO:0035265	organ growth	155	-0,44	-2,1	0,0000014	0,0000011
GO:0072001	renal system development	310	-0,34	-1,8	0,0000019	0,0000014
GO:0007416	synapse assembly	178	-0,41	-2	0,0000025	0,0000019
GO:0003279	cardiac septum development	105	-0,47	-2,2	0,0000029	0,0000022
GO:1901888	regulation of cell junction assembly	197	-0,39	-2	0,0000031	0,0000024
GO:0001822	kidney development	301	-0,34	-1,8	0,0000036	0,0000027

Supplementary table 7. List of genes mutated with differing frequencies between the two groups. Odds ratio indicates the odds that the gene is mutated in the AA group.

Gene symbol	AA	EA	P-val	Odds ratio
NME8	9	3	0.00024	11
NLRP6	5	0	0.00083	Inf
C9orf3	7	2	0.001	12
KIAA0226	7	2	0.001	12
MCM10	7	2	0.001	12
OPLAH	7	2	0.001	12
FCGR3A	6	1	0.0011	20
FLRT3	6	1	0.0011	20
PCDHGB3	7	3	0.0027	7.9
OR6F1	9	6	0.0029	5.2

CHST1	5	1	0.004	16
OXCT1	5	1	0.004	16
SLC17A3	5	1	0.004	16
SLC3A1	5	1	0.004	16
TTLL5	5	1	0.004	16
BCL11B	8	5	0.0042	5.5
CNTN1	9	7	0.0053	4.5
OR4A15	9	7	0.0053	4.5
SEMA3A	7	4	0.006	5.9
SYT10	7	4	0.006	5.9
CNTN5	8	6	0.0079	4.5
MUC2	8	6	0.0079	4.5
ABCA13	15	19	0.0081	3
SRGAP1	6	3	0.0084	6.6
UGT2B11	6	3	0.0084	6.6
FAT3	18	26	0.0094	2.7
C6orf222	5	2	0.011	8.1
DARC	5	2	0.011	8.1
KIAA2018	5	2	0.011	8.1
NTNG2	5	2	0.011	8.1
PTPRK	5	2	0.011	8.1
SMYD1	5	2	0.011	8.1
AMPH	7	5	0.012	4.7
PCDHB6	7	5	0.012	4.7
TLR4	10	10	0.012	3.5
DSG3	8	7	0.014	3.9
GRM6	8	7	0.014	3.9
OR2L13	8	7	0.014	3.9
SCN1A	8	7	0.014	3.9
CDH23	9	8	0.015	3.9
RYR1	14	19	0.015	2.7
DBC1	6	4	0.017	4.9
KRTAP4-11	6	4	0.017	4.9
MGAT5B	6	4	0.017	4.9
CNGB3	7	6	0.02	3.9
DACH2	7	6	0.02	3.9
HFM1	7	6	0.02	3.9
AGMO	0	14	0.023	0
AGBL1	5	3	0.025	5.4
CUX1	5	3	0.025	5.4
DNAJC10	5	3	0.025	5.4
FLRT2	5	3	0.025	5.4
HS3ST4	5	3	0.025	5.4
IRX2	5	3	0.025	5.4
KIAA1009	5	3	0.025	5.4
MMP2	5	3	0.025	5.4
OR11L1	5	3	0.025	5.4
OR2G3	5	3	0.025	5.4

OR4P4	5	3	0.025	5.4
OR5M11	5	3	0.025	5.4
RET	5	3	0.025	5.4
SLC8A2	5	3	0.025	5.4
TRPC3	5	3	0.025	5.4
WDR96	5	3	0.025	5.4
WWP1	5	3	0.025	5.4
ASH1L	8	8	0.03	3.4
EPHA3	8	8	0.03	3.4
NDST4	8	8	0.03	3.4
SELP	8	8	0.03	3.4
ZFHX3	8	8	0.03	3.4
CEP170	6	5	0.03	3.9
FCRLA	6	5	0.03	3.9
GABRA3	6	5	0.03	3.9
GABRB1	6	5	0.03	3.9
OR14K1	6	5	0.03	3.9
RB1CC1	6	5	0.03	3.9
SLC32A1	6	5	0.03	3.9
CTNNA2	12	16	0.031	2.6
CTNND2	10	12	0.033	2.9
ASPM	12	17	0.036	2.5
CCDC108	8	9	0.039	3
CREBBP	0	12	0.04	0
ATR	5	4	0.045	4
CFHR5	5	4	0.045	4
DNAJC6	5	4	0.045	4
FHL5	5	4	0.045	4
MAATS1	5	4	0.045	4
OR56A4	5	4	0.045	4
PRSS55	5	4	0.045	4
REG3G	5	4	0.045	4
SYCP2	5	4	0.045	4
ZNF33A	5	4	0.045	4
KIAA1210	1	17	0.048	0.16
CR1	7	7	0.048	3.3
KIAA1211	7	7	0.048	3.3
OR2W3	7	7	0.048	3.3
OR4S2	7	7	0.048	3.3
OR8U1	7	7	0.048	3.3
PCDHB2	7	7	0.048	3.3
TAF1	7	7	0.048	3.3
PTPRT	11	15	0.049	2.5

References

Allain, E.P., Rouleau, M., Lévesque, E., Guillemette, C., 2020. Emerging roles for UDP-glucuronosyltransferases in drug resistance and cancer progression. Br J Cancer 122, 1277–1287. https://doi.org/10.1038/s41416-019-0722-0

American Joint Committee on Cancer, 2002. AJCC Cancer Staging Manual. Springer, New York, NY. https://doi.org/10.1007/978-1-4757-3656-4

Amoah, J., Stuart, E.A., Cosgrove, S.E., Harris, A.D., Han, J.H., Lautenbach, E., Tamma, P.D., 2020. Comparing Propensity Score Methods Versus Traditional Regression Analysis for the Evaluation of Observational Data: A Case Study Evaluating the Treatment of Gram-Negative Bloodstream Infections. Clin Infect Dis 71, e497–e505. https://doi.org/10.1093/cid/ciaa169

Araujo, L.H., Carbone, D.P., 2017. Non-small cell lung cancer genomics around the globe: focus on ethnicity. J. Thorac. Dis. 9, E392–E394. https://doi.org/10.21037/jtd.2017.03.143

Araujo, L.H., Lammers, P.E., Matthews-Smith, V., Eisenberg, R., Gonzalez, A., Schwartz, A.G., Timmers, C., Shilo, K., Zhao, W., Natarajan, T.G., Zhang, J., Yilmaz, A.S., Liu, T., Coombes, K., Carbone, D.P., 2015a. Somatic Mutation Spectrum of Non–Small-Cell Lung Cancer in African Americans: A Pooled Analysis. Journal of Thoracic Oncology 10, 1430–1436. https://doi.org/10.1097/JTO.0000000000000650

Araujo, L.H., Timmers, C., Bell, E.H., Shilo, K., Lammers, P.E., Zhao, W., Natarajan, T.G., Miller, C.J., Zhang, J., Yilmaz, A.S., Liu, T., Coombes, K., Amann, J., Carbone, D.P., 2015b. Genomic Characterization of Non–Small-Cell Lung Cancer in African Americans by Targeted Massively Parallel Sequencing. JCO 33, 1966–1973. https://doi.org/10.1200/JCO.2014.59.2444

Arauz, R.F., Byun, J.S., Tandon, M., Sinha, S., Kuhn, S., Taylor, S., Zingone, A., Mitchell, K.A., Pine, S.R., Gardner, K., Perez-Stable, E.J., Napoles, A.M., Ryan, B.M., 2020. Whole-Exome Profiling of NSCLC Among African Americans. Journal of Thoracic Oncology 15, 1880–1892. https://doi.org/10.1016/j.jtho.2020.08.029

Arifin, M.T., Hama, S., Kajiwara, Y., Sugiyama, K., Saito, T., Matsuura, S., Yamasaki, F., Arita, K., Kurisu, K., 2006. Cytoplasmic, but not nuclear, p16 expression may signal poor prognosis in high-grade astrocytomas. J Neurooncol 77, 273–277. https://doi.org/10.1007/s11060-005-9037-5

Austin, P.C., 2011. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research 46, 399–424. https://doi.org/10.1080/00273171.2011.568786

Barbar, J., Armach, M., Hodroj, M.H., Assi, S., El Nakib, C., Chamseddine, N., Assi, H.I., 2022. Emerging genetic biomarkers in lung adenocarcinoma. SAGE Open Medicine 10, 20503121221132352. https://doi.org/10.1177/20503121221132352

Bean, J., Brennan, C., Shih, J.-Y., Riely, G., Viale, A., Wang, L., Chitale, D., Motoi, N., Szoke, J., Broderick, S., Balak, M., Chang, W.-C., Yu, C.-J., Gazdar, A., Pass, H., Rusch, V., Gerald, W., Huang, S.-F., Yang, P.-C., Miller, V., Ladanyi, M., Yang, C.-H., Pao, W., 2007. MET amplification occurs with or without T790M mutations in EGFR mutant lung tumors with acquired resistance to gefitinib or erlotinib. Proc Natl Acad Sci U S A 104, 20932–20937. https://doi.org/10.1073/pnas.0710370104

Bedi, M., Ray, M., Ghosh, A., 2022. Active mitochondrial respiration in cancer: a target for the drug. Mol Cell Biochem 477, 345–361. https://doi.org/10.1007/s11010-021-04281-4

Bennett, D.A., Leurgans, S., 2010. Is there a link between cancer and Alzheimer disease? Neurology 74, 100–101. https://doi.org/10.1212/WNL.0b013e3181cbb89a

Blakely, C.M., Bivona, T.G., 2012. Resiliency of Lung Cancers to EGFR Inhibitor Treatment Unveiled, Offering Opportunities to Divide and Conquer EGFR Inhibitor Resistance. Cancer Discovery 2, 872–875. https://doi.org/10.1158/2159-8290.CD-12-0387

Bollig-Fischer, A., Chen, W., Gadgeel, S.M., Wenzlaff, A.S., Cote, M.L., Schwartz, A.G., Bepler, G., 2015. Racial Diversity of Actionable Mutations in Non–Small Cell Lung Cancer. Journal of Thoracic Oncology 10, 250–255. https://doi.org/10.1097/JTO.0000000000000420

Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., Sherlock, G., 2004. GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20, 3710–3715. https://doi.org/10.1093/bioinformatics/bth456

Califf, R.M., 2018. Biomarker definitions and their applications. Exp Biol Med (Maywood) 243, 213–221. https://doi.org/10.1177/1535370217750088

Campbell, J.D., Lathan, C., Sholl, L., Ducar, M., Vega, M., Sunkavalli, A., Lin, L., Hanna, M., Schubert, L., Thorner, A., Faris, N., Williams, D.R., Osarogiagbon, R.U., van Hummelen, P., Meyerson, M., MacConaill, L., 2017. Comparison of Prevalence and Types of Mutations in Lung Cancers Among Black and White Populations. JAMA Oncol 3, 801. https://doi.org/10.1001/jamaoncol.2016.6108

Cancer of the Breast (Female) - Cancer Stat Facts [WWW Document], n.d. . SEER. URL https://seer.cancer.gov/statfacts/html/breast.html (accessed 8.12.23).

Cancer of the Lung and Bronchus - Cancer Stat Facts [WWW Document], n.d. . SEER. URL https://seer.cancer.gov/statfacts/html/lungb.html (accessed 8.12.23).

Cappuzzo, F., Varella-Garcia, M., Shigematsu, H., Domenichini, I., Bartolini, S., Ceresoli, G.L., Rossi, E., Ludovini, V., Gregorc, V., Toschi, L., Franklin, W.A., Crino, L., Gazdar, A.F., Bunn, P.A., Hirsch, F.R., 2005. Increased HER2 Gene Copy Number Is Associated With Response to Gefitinib Therapy in Epidermal Growth Factor Receptor–Positive Non–Small-Cell Lung Cancer Patients. JCO 23, 5007–5018. https://doi.org/10.1200/JCO.2005.09.111

Chen, F., Wang, X.-Y., Han, X.-H., Wang, H., Qi, J., 2015. Diagnostic value of Cyfra21-1, SCC and CEA for differentiation of early-stage NSCLC from benign lung disease. Int J Clin Exp Med 8, 11295–11300.

Chen, R., Khatri, P., Mazur, P.K., Polin, M., Zheng, Y., Vaka, D., Hoang, C.D., Shrager, J., Xu, Y., Vicent, S., Butte, A.J., Sweet-Cordero, E.A., 2014. A Meta-analysis of Lung Cancer Gene Expression Identifies PTK7 as a Survival Gene in Lung Adenocarcinoma. Cancer Research 74, 2892–2902. https://doi.org/10.1158/0008-5472.CAN-13-2775

Climente-González, H., Porta-Pardo, E., Godzik, A., Eyras, E., 2017. The Functional Impact of Alternative Splicing in Cancer. Cell Reports 20, 2215–2226. https://doi.org/10.1016/j.celrep.2017.08.012

Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I., Ceccarelli, M., Bontempi, G., Noushmehr, H., 2016. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res 44, e71. https://doi.org/10.1093/nar/gkv1507

Collisson, E.A., Campbell, J.D., Brooks, A.N., Berger, A.H., Lee, W., Chmielecki, J., Beer, D.G., Cope, L., Creighton, C.J., Danilova, L., Ding, L., Getz, G., Hammerman, P.S., Neil Hayes, D., Hernandez, B., Herman, J.G., Heymach, J.V., Jurisica, I., Kucherlapati, R., Kwiatkowski, D., Ladanyi, M., Robertson, G., Schultz, N., Shen, R., Sinha, R., Sougnez, C., Tsao, M.-S., Travis, W.D., Weinstein, J.N., Wigle, D.A., Wilkerson, M.D., Chu, A., Cherniack, A.D., Hadjipanayis, A., Rosenberg, M., Weisenberger, D.J., Laird, P.W., Radenbaugh, A., Ma, S., Stuart, J.M., Averett Byers, L., Baylin, S.B., Govindan, R., Meyerson, M., Rosenberg, M., Gabriel, S.B., Cibulskis, K., Sougnez, C., Kim, J., Stewart, C., Lichtenstein, L., Lander, E.S., Lawrence, M.S., Getz, G., Kandoth, C., Fulton, R., Fulton, L.L., McLellan, M.D., Wilson, R.K., Ye, K., Fronick, C.C., Maher, C.A., Miller, C.A., Wendl, M.C., Cabanski, C., Ding, L., Mardis, E., Govindan, R., Creighton, C.J., Wheeler, D., Balasundaram, M., Butterfield, Y.S.N., Carlsen, R., Chu, A., Chuah, E., Dhalla, N., Guin, R., Hirst, C., Lee, D., Li, H.I., Mayo, M., Moore, R.A., Mungall, A.J., Schein, J.E., Sipahimalani, P., Tam, A., Varhol, R., Gordon Robertson, A., Wye, N., Thiessen, N., Holt, R.A., Jones, S.J.M., Marra, M.A., Campbell, J.D., Brooks, A.N., Chmielecki, J., Imielinski, M.,

Onofrio, R.C., Hodis, E., Zack, T., Sougnez, C., Helman, E., Sekhar Pedamallu, C., Mesirov, J., Cherniack, A.D., Saksena, G., Schumacher, S.E., Carter, S.L., Hernandez, B., Garraway, L., Beroukhim, R., Gabriel, S.B., Getz, G., Meyerson, M., Hadjipanayis, A., Lee, S., Mahadeshwar, H.S., Pantazi, A., Protopopov, A., Ren, X., Seth, S., Song, X., Tang, J., Yang, L., Zhang, J., Chen, P.-C., Parfenov, M., Wei Xu, A., Santoso, N., Chin, L., Park, P.J., Kucherlapati, R., Hoadley, K.A., Todd Auman, J., Meng, S., Shi, Y., Buda, E., Waring, S., Veluvolu, U., Tan, D., Mieczkowski, P.A., Jones, C.D., Simons, J.V., Soloway, M.G., Bodenheimer, T., Jefferys, S.R., Roach, J., Hoyle, A.P., Wu, J., Balu, S., Singh, D., Prins, J.F., Marron, J.S., Parker, J.S., Neil Hayes, D., Perou, C.M., Liu, J., Cope, L., Danilova, L., Weisenberger, D.J., Maglinte, D.T., Lai, P.H., Bootwalla, M.S., Van Den Berg, D.J., Triche Jr, T., Baylin, S.B., Laird, P.W., Rosenberg, M., Chin, L., Zhang, J., Cho, J., DiCara, D., Heiman, D., Lin, P., Mallard, W., Voet, D., Zhang, H., Zou, L., Noble, M.S., Lawrence, M.S., Saksena, G., Gehlenborg, N., Thorvaldsdottir, H., Mesirov, J., Nazaire, M.-D., Robinson, J., Getz, G., Lee, W., Arman Aksoy, B., Ciriello, G., Taylor, B.S., Dresdner, G., Gao, J., Gross, B., Seshan, V.E., Ladanyi, M., Reva, B., Sinha, R., Onur Sumer, S., Weinhold, N., Schultz, N., Shen, R., Sander, C., Ng, S., Ma, S., Zhu, J., Radenbaugh, A., Stuart, J.M., Benz, C.C., Yau, C., Haussler, D., Spellman, P.T., Wilkerson, M.D., Parker, J.S., Hoadley, K.A., Kimes, P.K., Neil Hayes, D., Perou, C.M., Broom, B.M., Wang, J., Lu, Y., Kwok Shing Ng, P., Diao, L., Averett Byers, L., Liu, W., Heymach, J.V., Amos, C.I., Weinstein, J.N., Akbani, R., Mills, G.B., Curley, E., Paulauskis, J., Lau, K., Morris, S., Shelton, T., Mallery, D., Gardner, J., Penny, R., Saller, C., Tarvin, K., Richards, W.G., Cerfolio, R., Bryant, A., Raymond, D.P., Pennell, N.A., Farver, C., Czerwinski, C., Huelsenbeck-Dill, L., Iacocca, M., Petrelli, N., Rabeno, B., Brown, J., Bauer, T., Dolzhanskiy, O., Potapova, O., Rotin, D., Voronina, O., Nemirovich-Danchenko, E., Fedosenko, K.V., Gal, A., Behera, M., Ramalingam, S.S., Sica, G., Flieder, D., Boyd, J., Weaver, J., Kohl, B., Huy Quoc Thinh, D., Sandusky, G., Juhl, H., The Cancer Genome Atlas Research Network, Disease analysis working group, Genome sequencing centres: The Eli & Edythe L. Broad Institute, Washington University in St. Louis, Baylor College of Medicine, Genome characterization centres: Canada's Michael Smith Genome Sciences Centre, B.C.C.A., The Eli & Edythe L. Broad Institute, Harvard Medical School/Brigham & Women's Hospital/MD Anderson Cancer Center, University of North Carolina, C.H., University of Kentucky, The USC/JHU Epigenome Characterization Center, Genome data analysis centres: The Eli & Edythe L. Broad Institute, Memorial Sloan-Kettering Cancer Center, University of California, S.C.I., Oregon Health & Sciences University, The University of Texas MD Anderson Cancer Center, Biospecimen core resource: International Genomics Consortium, Tissue source sites: Analytical Biological Service, Inc., Brigham & Women's Hospital, University of Alabama at Birmingham, Cleveland Clinic, Christiana Care, Cureline, Emory University, Fox Chase Cancer Center, ILSbio, Indiana University, Indivumed, John Flynn Hospital, 2014. Comprehensive molecular profiling of lung adenocarcinoma. Nature 511, 543–550. https://doi.org/10.1038/nature13385

Costa, P.A., Saul, E.E., Paul, Y., Iyer, S., 2021. Prevalence of Targetable Mutations in Black Patients With Lung Cancer: A Systematic Review and Meta-Analysis 17, 9.

de Fraipont, F., Gazzeri, S., Cho, W.C., Eymin, B., 2019. Circular RNAs and RNA Splice Variants as Biomarkers for Prognosis and Therapeutic Response in the Liquid Biopsies of Lung Cancer Patients. Frontiers in Genetics 10.

Der, S.D., Sykes, J., Pintilie, M., Zhu, C.-Q., Strumpf, D., Liu, N., Jurisica, I., Shepherd, F.A., Tsao, M.-S., 2014. Validation of a Histology-Independent Prognostic Gene Signature for Early-Stage, Non–Small-Cell Lung Cancer Including Stage IA Patients. Journal of Thoracic Oncology 9, 59–64. https://doi.org/10.1097/JTO.0000000000000042

Deveaux, A.E., Allen, T.A., Al Abo, M., Qin, X., Zhang, D., Patierno, B.M., Gu, L., Gray, J.E., Pecot, C.V., Dressman, H.K., McCall, S.J., Kittles, R.A., Hyslop, T., Owzar, K., Crawford, J., Patierno, S.R., Clarke, J.M., Freedman, J.A., 2021. RNA splicing and aggregate gene expression differences in

lung squamous cell carcinoma between patients of West African and European ancestry. Lung Cancer 153, 90–98. https://doi.org/10.1016/j.lungcan.2021.01.015
Durinck, S., Spellman, P.T., Birney, E., Huber, W., 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc 4, 1184–1191. https://doi.org/10.1038/nprot.2009.97
El-Telbany, A., Ma, P.C., 2012. Cancer Genes in Lung Cancer: Racial Disparities: Are There Any? Genes & Cancer 3, 467–480. https://doi.org/10.1177/1947601912465177
Engelman, J.A., Zejnullahu, K., Mitsudomi, T., Song, Y., Hyland, C., Park, J.O., Lindeman, N., Gale, C.-M., Zhao, X., Christensen, J., Kosaka, T., Holmes, A.J., Rogers, A.M., Cappuzzo, F., Mok, T., Lee, C., Johnson, B.E., Cantley, L.C., Jänne, P.A., 2007. MET Amplification Leads to Gefitinib Resistance in Lung Cancer by Activating ERBB3 Signaling. Science 316, 1039–1043. https://doi.org/10.1126/science.1141478
FireBrowse [WWW Document], n.d. URL http://firebrowse.org/ (accessed 4.21.23).
Frank, M., Kemler, R., 2002. Protocadherins. Curr Opin Cell Biol 14, 557–562. https://doi.org/10.1016/s0955-0674(02)00365-4
Gao, X., Wang, Y., Zhao, H., Wei, F., Zhang, X., Su, Y., Wang, C., Li, H., Ren, X., 2016. Plasma miR-324-3p and miR-1285 as diagnostic and prognostic biomarkers for early stage lung squamous cell carcinoma. Oncotarget 7, 59664–59675. https://doi.org/10.18632/oncotarget.11198
GDC [WWW Document], n.d. URL https://portal.gdc.cancer.gov/ (accessed 11.4.23).
Gilad, S., Lithwick-Yanai, G., Barshack, I., Benjamin, S., Krivitsky, I., Edmonston, T.B., Bibbo, M., Thurm, C., Horowitz, L., Huang, Y., Feinmesser, M., Hou, J.S., St Cyr, B., Burnstein, I., Gibori, H., Dromi, N., Sanden, M., Kushnir, M., Aharonov, R., 2012. Classification of the four main types of lung cancer using a microRNA-based diagnostic assay. J Mol Diagn 14, 510–517. https://doi.org/10.1016/j.jmoldx.2012.03.004
Grunda, J.M., Steg, A.D., He, Q., Steciuk, M.R., Byan-Parker, S., Johnson, M.R., Grizzle, W.E., 2012. Differential expression of breast cancer-associated genes between stage- and age-matched tumor specimens from African- and Caucasian-American Women diagnosed with breast cancer. BMC Res Notes 5, 248. https://doi.org/10.1186/1756-0500-5-248
Gu, J., Zhou, Y., Huang, L., Ou, W., Wu, J., Li, S., Xu, J., Feng, J., Liu, B., 2016. TP53 mutation is associated with a poor clinical outcome for non-small cell lung cancer: Evidence from a meta-analysis. Molecular and Clinical Oncology 5, 705–713. https://doi.org/10.3892/mco.2016.1057
Gu, X.S., Rosenbaum, P.R., 1993. Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms. Journal of Computational and Graphical Statistics 2, 405–420. https://doi.org/10.1080/10618600.1993.10474623
Gunaratne, P.H., Coarfa, C., Soibam, B., Tandon, A., 2012. miRNA Data Analysis: Next-Gen Sequencing, in: Fan, J.-B. (Ed.), Next-Generation MicroRNA Expression Profiling Technology: Methods and Protocols, Methods in Molecular Biology. Humana Press, Totowa, NJ, pp. 273–288. https://doi.org/10.1007/978-1-61779-427-8_19
Hao, W., Yu, M., Lin, J., Liu, B., Xing, H., Yang, J., Sun, D., Chen, F., Jiang, M., Tang, C., Zhang, X., Zhao, Y., Zhu, Y., 2020. The pan-cancer landscape of netrin family reveals potential oncogenic biomarkers. Sci Rep 10, 5224. https://doi.org/10.1038/s41598-020-62117-5
Heath, A.P., Ferretti, V., Agrawal, S., An, M., Angelakos, J.C., Arya, R., Bajari, R., Baqar, B., Barnowski, J.H.B., Burt, J., Catton, A., Chan, B.F., Chu, F., Cullion, K., Davidsen, T., Do, P.-M., Dompierre, C., Ferguson, M.L., Fitzsimons, M.S., Ford, M., Fukuma, M., Gaheen, S., Ganji, G.L., Garcia, T.I., George, S.S., Gerhard, D.S., Gerthoffert, F., Gomez, F., Han, K., Hernandez, K.M., Issac, B., Jackson, R., Jensen, M.A., Joshi, S., Kadam, A., Khurana, A., Kim, K.M.J., Kraft, V.E., Li, S., Lichtenberg, T.M., Lodato, J., Lolla, L., Martinov, P., Mazzone, J.A., Miller, D.P., Miller, I., Miller, J.S., Miyauchi, K., Murphy, M.W., Nullet, T., Ogwara, R.O., Ortuño, F.M., Pedrosa, J., Pham, P.L., Popov, M.Y., Porter, J.J., Powell, R., Rademacher, K., Reid, C.P., Rich, S., Rogel, B., Sahni, H., Savage, J.H., Schmitt, K.A., Simmons, T.J., Sislow, J., Spring, J., Stein, L., Sullivan, S.,

Tang, Y., Thiagarajan, M., Troyer, H.D., Wang, C., Wang, Z., West, B.L., Wilmer, A., Wilson, S., Wu, K., Wysocki, W.P., Xiang, L., Yamada, J.T., Yang, L., Yu, C., Yung, C.K., Zenklusen, J.C., Zhang, J., Zhang, Z., Zhao, Y., Zubair, A., Staudt, L.M., Grossman, R.L., 2021. The NCI Genomic Data Commons. Nat Genet 53, 257–262. https://doi.org/10.1038/s41588-021-00791-5

Herbst, R.S., Heymach, J.V., Lippman, S.M., 2008. Lung Cancer. N Engl J Med 359, 1367–1380. https://doi.org/10.1056/NEJMra0802714

Huang, L., Liang, W., Wei, J., Xu, Z., Sha, Y., Deng, Y., Ou, M., 2022. Bioinformatics study of PCDHB6 as a prognostic marker for gastric cancer (preprint). In Review. https://doi.org/10.21203/rs.3.rs-2019985/v1

Huang, Y.-T., Lin, X., Chirieac, L.R., McGovern, R., Wain, J.C., Heist, R.S., Skaug, V., Zienolddiny, S., Haugen, A., Su, L., Christiani, D.C., 2011. Impact on Disease Development, Genomic Location and Biological Function of Copy Number Alterations in NonSmall Cell Lung Cancer. PLOS ONE 6, e22961. https://doi.org/10.1371/journal.pone.0022961

Hughes, A.L., Welch, R., Puri, V., Matthews, C., Haque, K., Chanock, S.J., Yeager, M., 2008. Genome-wide SNP typing reveals signatures of population history. Genomics 92, 1–8. https://doi.org/10.1016/j.ygeno.2008.03.005

Inzelberg, R., Jankovic, J., 2007. Are Parkinson disease patients protected from some but not all cancers? Neurology 69, 1542–1550. https://doi.org/10.1212/01.wnl.0000277638.63767.b8

Iwamori, T., Iwamori, N., Matsumoto, M., Ono, E., Matzuk, M.M., 2017. Identification of KIAA1210 as a novel X-chromosome-linked protein that localizes to the acrosome and associates with the ectoplasmic specialization in testes. Biol Reprod 96, 469–477. https://doi.org/10.1095/biolreprod.116.145458

Janiszewska, M., Primi, M.C., Izard, T., 2020. Cell adhesion in cancer: Beyond the migration of single cells. J Biol Chem 295, 2495–2505. https://doi.org/10.1074/jbc.REV119.007759

Javadian, P., Washington, C., Mukasa, S., Benbrook, D.M., 2021. Histopathologic, Genetic and Molecular Characterization of Endometrial Cancer Racial Disparity. Cancers 13, 1900. https://doi.org/10.3390/cancers13081900

Johri, A., Beal, M.F., 2012. Mitochondrial Dysfunction in Neurodegenerative Diseases. J Pharmacol Exp Ther 342, 619–630. https://doi.org/10.1124/jpet.112.192138

Kalluri, R., Weinberg, R.A., 2009. The basics of epithelial-mesenchymal transition. Journal of Clinical Investigation 119, 1420–1429.

Keenan, T., Moy, B., Mroz, E.A., Ross, K., Niemierko, A., Rocco, J.W., Isakoff, S., Ellisen, L.W., Bardia, A., 2015. Comparison of the Genomic Landscape Between Primary Breast Cancer in African American Versus White Women and the Association of Racial Differences With Tumor Recurrence. JCO 33, 3621–3627. https://doi.org/10.1200/JCO.2015.62.2126

Keeton, A.B., Salter, E.A., Piazza, G.A., 2017. The RAS–Effector Interaction as a Drug Target. Cancer Research 77, 221–226. https://doi.org/10.1158/0008-5472.CAN-16-0938

Kratz, J.R., He, J., Eeden, S.K.V.D., Zhu, Z.-H., Gao, W., Pham, P.T., Mulvihill, M.S., Ziaei, F., Zhang, H., Su, B., Zhi, X., Quesenberry, C.P., Habel, L.A., Deng, Q., Wang, Z., Zhou, J., Li, H., Huang, M.-C., Yeh, C.-C., Segal, M.R., Ray, M.R., Jones, K.D., Raz, D.J., Xu, Z., Jahan, T.M., Berryman, D., He, B., Mann, M.J., Jablons, D.M., 2012. A practical molecular assay to predict survival in resected non-squamous, non-small-cell lung cancer: development and international validation studies. The Lancet 379, 823–832. https://doi.org/10.1016/S0140-6736(11)61941-7

Krishnan, B., Rose, T.L., Kardos, J., Milowsky, M.I., Kim, W.Y., 2016. Intrinsic Genomic Differences Between African American and White Patients With Clear Cell Renal Cell Carcinoma. JAMA Oncol 2, 664. https://doi.org/10.1001/jamaoncol.2016.0005

Lam, A.K.-Y., Ong, K., Giv, M.J., Ho, Y.-H., 2008. p16 expression in colorectal adenocarcinoma: marker of aggressiveness and morphological types. Pathology 40, 580–585. https://doi.org/10.1080/00313020802320713

Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K.N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S.A., Haggarty, S.J., Clemons, P.A., Wei, R., Carr, S.A., Lander, E.S., Golub, T.R., 2006. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. Science 313, 1929–1935. https://doi.org/10.1126/science.1132939

Landry, L.G., Ali, N., Williams, D.R., Rehm, H.L., Bonham, V.L., 2018. Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice. Health Aff (Millwood) 37, 780–785. https://doi.org/10.1377/hlthaff.2017.1595

Lara, O.D., Wang, Y., Asare, A., Xu, T., Chiu, H.-S., Liu, Y., Hu, W., Sumazin, P., Uppal, S., Zhang, L., Rauh-Hain, J.A., Sood, A.K., 2020. Pan-cancer clinical and molecular analysis of racial disparities. Cancer 126, 800–807. https://doi.org/10.1002/cncr.32598

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., Carey, V.J., 2013. Software for Computing and Annotating Genomic Ranges. PLoS Comput Biol 9, e1003118. https://doi.org/10.1371/journal.pcbi.1003118

Li, B.-Q., You, J., Huang, T., Cai, Y.-D., 2014. Classification of Non-Small Cell Lung Cancer Based on Copy Number Alterations. PLOS ONE 9, e88300. https://doi.org/10.1371/journal.pone.0088300

Li, J., Gao, X., Tang, M., Wang, C., Liu, W., Tian, S., 2022. Autoencoder Networks Decipher the Association between Lung Cancer and Alzheimer's Disease. Computational Intelligence and Neuroscience 2022, e2009545. https://doi.org/10.1155/2022/2009545

Li, M., Fu, S., Xiao, H., 2015. Genome-wide analysis of microRNA and mRNA expression signatures in cancer. Acta Pharmacol Sin 36, 1200–1211. https://doi.org/10.1038/aps.2015.67

Liberti, M.V., Locasale, J.W., 2016. The Warburg Effect: How Does it Benefit Cancer Cells? Trends in Biochemical Sciences 41, 211–218. https://doi.org/10.1016/j.tibs.2015.12.001

Liu, W., Li, J., Zhao, R., Lu, Y., Huang, P., 2022. The Uridine diphosphate (UDP)-glycosyltransferases (UGTs) superfamily: the role in tumor cell metabolism. Front Oncol 12, 1088458. https://doi.org/10.3389/fonc.2022.1088458

López-Ayllón, B.D., de Castro-Carpeño, J., Rodriguez, C., Pernía, O., de Cáceres, I.I., Belda-Iniesta, C., Perona, R., Sastre, L., 2015. Biomarkers of erlotinib response in non-small cell lung cancer tumors that do not harbor the more common epidermal growth factor receptor mutations. Int J Clin Exp Pathol 8, 2888–2898.

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15, 550. https://doi.org/10.1186/s13059-014-0550-8

Lusk, C.M., Watza, D., Dyson, G., Craig, D., Ratliff, V., Wenzlaff, A.S., Lonardo, F., Bollig-Fischer, A., Bepler, G., Purrington, K., Gadgeel, S., Schwartz, A.G., 2019. Profiling the Mutational Landscape in Known Driver Genes and Novel Genes in African American Non–Small Cell Lung Cancer Patients. Clin Cancer Res 25, 4300–4308. https://doi.org/10.1158/1078-0432.CCR-18-2439

Martin, D.N., Boersma, B.J., Yi, M., Reimers, M., Howe, T.M., Yfantis, H.G., Tsai, Y.C., Williams, E.H., Lee, D.H., Stephens, R.M., Weissman, A.M., Ambs, S., 2009. Differences in the Tumor Microenvironment between African-American and European-American Breast Cancer Patients. PLoS ONE 4, e4531. https://doi.org/10.1371/journal.pone.0004531

Mattson, M.P., 2000. Apoptosis in neurodegenerative disorders. Nat Rev Mol Cell Biol 1, 120–130. https://doi.org/10.1038/35040009

Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., Koeffler, H.P., 2018. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res 28, 1747–1756. https://doi.org/10.1101/gr.239244.118

McCarroll, S.A., Kuruvilla, F.G., Korn, J.M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M.H., de Bakker, P.I.W., Maller, J.B., Kirby, A., Elliott, A.L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P.J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K.W., Rava, R.,

Daly, M.J., Gabriel, S.B., Altshuler, D., 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40, 1166–1174. https://doi.org/10.1038/ng.238
Mersha, T.B., Abebe, T., 2015. Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. Human Genomics 9, 1. https://doi.org/10.1186/s40246-014-0023-x
Milde-Langosch, K., Bamberger, A.-M., Rieck, G., Kelp, B., Löning, T., 2001. Overexpression of the p16 Cell Cycle Inhibitor in Breast Cancer is Associated with a More Malignant Phenotype. Breast Cancer Res Treat 67, 61–70. https://doi.org/10.1023/A:1010623308275
Minari, R., Bordi, P., Tiseo, M., 2016. Third-generation epidermal growth factor receptor-tyrosine kinase inhibitors in T790M-positive non-small cell lung cancer: review on emerged mechanisms of resistance. Transl Lung Cancer Res 5, 695–708. https://doi.org/10.21037/tlcr.2016.12.02
Mitchell, K.A., Zingone, A., Toulabi, L., Boeckelman, J., Ryan, B.M., 2017. Comparative Transcriptome Profiling Reveals Coding and Noncoding RNA Differences in NSCLC from African Americans and European Americans. Clin Cancer Res 23, 7412–7425. https://doi.org/10.1158/1078-0432.CCR-17-0527
Morganella, S., Pagnotta, S.M., Ceccarelli, M., 2011. Finding recurrent copy number alterations preserving within-sample homogeneity. Bioinformatics 27, 2949–2956. https://doi.org/10.1093/bioinformatics/btr488
Newman, L.A., Kaljee, L.M., 2017. Health Disparities and Triple-Negative Breast Cancer in African American Women: A Review. JAMA Surgery 152, 485–493. https://doi.org/10.1001/jamasurg.2017.0005
NIH About Cancers: Understanding Cancers, 2022. URL https://www.cancer.gov/about-cancer/understanding/what-is-cancer#:~:text=Cancer%20is%20caused%20by%20certain,tightly%20packed%20DNA%20called%20chromosomes.&text=Cancer%20is%20a%20genetic%20disease,how%20they%20grow%20and%20divide.
Okamura, K., Takayama, K., Izumi, M., Harada, T., Furuyama, K., Nakanishi, Y., 2013. Diagnostic value of CEA and CYFRA 21-1 tumor markers in primary lung cancer. Lung Cancer 80, 45–49. https://doi.org/10.1016/j.lungcan.2013.01.002
Persson, K., Hamby, K., Ugozzoli, L.A., 2005. Four-color multiplex reverse transcription polymerase chain reaction—Overcoming its limitations. Analytical Biochemistry 344, 33–42. https://doi.org/10.1016/j.ab.2005.06.026
Phillips, J.D., 2019. Heme biosynthesis and the porphyrias. Mol Genet Metab 128, 164–177. https://doi.org/10.1016/j.ymgme.2019.04.008
Ping, J., Guo, X., Ye, F., Long, J., Lipworth, L., Cai, Q., Blot, W., Shu, X.-O., Zheng, W., 2020. Differences in gene-expression profiles in breast cancer between African and European-ancestry women. Carcinogenesis 41, 887–893. https://doi.org/10.1093/carcin/bgaa035
Ponder, B.A.J., 2001. Cancer genetics. Nature 411, 336–341. https://doi.org/10.1038/35077207
Provenzano, M., Mocellin, S., 2007. Complementary Techniques, in: Mocellin, S. (Ed.), Microarray Technology and Cancer Gene Profiling, Advances in Experimental Medicine and Biology. Springer, New York, NY, pp. 66–73. https://doi.org/10.1007/978-0-387-39978-2_7
Qiu, Z.-W., Bi, J.-H., Gazdar, A.F., Song, K., 2017. Genome-wide copy number variation pattern analysis and a classification signature for non-small cell lung cancer. Genes, Chromosomes and Cancer 56, 559–569. https://doi.org/10.1002/gcc.22460
Rascio, F., Spadaccino, F., Rocchetti, M.T., Castellano, G., Stallone, G., Netti, G.S., Ranieri, E., 2021. The Pathogenic Role of PI3K/AKT Pathway in Cancer Onset and Drug Resistance: An Updated Review. Cancers (Basel) 13, 3949. https://doi.org/10.3390/cancers13163949
Rastel, D., Ramaioli, A., Cornillie, F., Thirion, B., 1994. CYFRA 21-1, a sensitive and specific new tumour marker for squamous cell lung cancer. Report of the first European multicentre

evaluation. European Journal of Cancer 30, 601–606. https://doi.org/10.1016/0959-8049(94)90528-2
Reeve, B.B., Smith, A.W., Arora, N.K., Hays, R.D., 2008. Reducing Bias in Cancer Research: Application of Propensity Score Matching. Health Care Financ Rev 29, 69–80.
Reiman, T., Lai, R., Veillard, A.S., Paris, E., Soria, J.C., Rosell, R., Taron, M., Graziano, S., Kratzke, R., Seymour, L., Shepherd, F.A., Pignon, J.P., Sève, P., 2012. Cross-validation study of class III beta-tubulin as a predictive marker for benefit from adjuvant chemotherapy in resected non-small-cell lung cancer: analysis of four randomized trials. Annals of Oncology 23, 86–93. https://doi.org/10.1093/annonc/mdr033
Restrepo, J.C., Dueñas, D., Corredor, Z., Liscano, Y., 2023. Advances in Genomic Data and Biomarkers: Revolutionizing NSCLC Diagnosis and Treatment. Cancers 15, 3474. https://doi.org/10.3390/cancers15133474
Rodriguez-Antona, C., Ingelman-Sundberg, M., 2006. Cytochrome P450 pharmacogenetics and cancer. Oncogene 25, 1679–1691. https://doi.org/10.1038/sj.onc.1209377
Roepman, P., Jassem, J., Smit, E.F., Muley, T., Niklinski, J., van de Velde, T., Witteveen, A.T., Rzyman, W., Floore, A., Burgers, S., Giaccone, G., Meister, M., Dienemann, H., Skrzypski, M., Kozlowski, M., Mooi, W.J., van Zandwijk, N., 2008. An Immune Response Enriched 72-Gene Prognostic Profile for Early-Stage Non–Small-Cell Lung Cancer. Clinical Cancer Research 15, 284–290. https://doi.org/10.1158/1078-0432.CCR-08-1258
Romagosa, C., Simonetti, S., López-Vicente, L., Mazo, A., Lleonart, M.E., Castellvi, J., Ramon y Cajal, S., 2011. p16Ink4a overexpression in cancer: a tumor suppressor gene associated with senescence and high-grade tumors. Oncogene 30, 2087–2097. https://doi.org/10.1038/onc.2010.614
Rosario, S.R., Long, M.D., Affronti, H.C., Rowsam, A.M., Eng, K.H., Smiraglia, D.J., 2018. Pan-cancer analysis of transcriptional metabolic dysregulation using The Cancer Genome Atlas. Nat Commun 9, 5330. https://doi.org/10.1038/s41467-018-07232-8
ROSENBAUM, P.R., RUBIN, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55. https://doi.org/10.1093/biomet/70.1.41
Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., Feldman, M.W., 2002. Genetic Structure of Human Populations. Science 298, 2381–2385. https://doi.org/10.1126/science.1078311
Schabath, M.B., Cote, M.L., 2019. Cancer Progress and Priorities: Lung Cancer. Cancer Epidemiology, Biomarkers & Prevention 28, 1563–1579. https://doi.org/10.1158/1055-9965.EPI-19-0221
Shames, D.S., Wistuba, I.I., 2014. The evolving genomic classification of lung cancer. J Pathol 232, 121–133. https://doi.org/10.1002/path.4275
Shendure, J., Ji, H., 2008. Next-generation DNA sequencing. Nat Biotechnol 26, 1135–1145. https://doi.org/10.1038/nbt1486
Shi, H., Seegobin, K., Heng, F., Zhou, K., Chen, R., Qin, H., Manochakian, R., Zhao, Y., Lou, Y., 2022. Genomic landscape of lung adenocarcinomas in different races. Frontiers in Oncology 12.
Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A., 2022. Cancer statistics, 2022. CA A Cancer J Clinicians 72, 7–33. https://doi.org/10.3322/caac.21708
Silva, T.C., Colaprico, A., Olsen, C., D'Angelo, F., Bontempi, G., Ceccarelli, M., Noushmehr, H., 2016. *TCGA Workflow*: Analyze cancer genomics and epigenomics data using Bioconductor packages. https://doi.org/10.12688/f1000research.8923.2
Sinha, S., Mitchell, K.A., Zingone, A., Bowman, E., Sinha, N., Schäffer, A.A., Lee, J.S., Ruppin, E., Ryan, B.M., 2020. Higher prevalence of homologous recombination deficiency in tumors from African Americans versus European Americans. Nat Cancer 1, 112–121. https://doi.org/10.1038/s43018-019-0009-7
Sinkala, M., 2023. Mutational landscape of cancer-driver genes across human cancers. Sci Rep 13, 12742. https://doi.org/10.1038/s41598-023-39608-2

Sirugo, G., Williams, S.M., Tishkoff, S.A., 2019. The Missing Diversity in Human Genetic Studies. Cell 177, 26–31. https://doi.org/10.1016/j.cell.2019.02.048

Skoulidis, F., Goldberg, M.E., Greenawalt, D.M., Hellmann, M.D., Awad, M.M., Gainor, J.F., Schrock, A.B., Hartmaier, R.J., Trabucco, S.E., Gay, L., Ali, S.M., Elvin, J.A., Singal, G., Ross, J.S., Fabrizio, D., Szabo, P.M., Chang, H., Sasson, A., Srinivasan, S., Kirov, S., Szustakowski, J., Vitazka, P., Edwards, R., Bufill, J.A., Sharma, N., Ou, S.-H.I., Peled, N., Spigel, D.R., Rizvi, H., Aguilar, E.J., Carter, B.W., Erasmus, J., Halpenny, D.F., Plodkowski, A.J., Long, N.M., Nishino, M., Denning, W.L., Galan-Cobo, A., Hamdi, H., Hirz, T., Tong, P., Wang, J., Rodriguez-Canales, J., Villalobos, P.A., Parra, E.R., Kalhor, N., Sholl, L.M., Sauter, J.L., Jungbluth, A.A., Mino-Kenudson, M., Azimi, R., Elamin, Y.Y., Zhang, J., Leonardi, G.C., Jiang, F., Wong, K.-K., Lee, J.J., Papadimitrakopoulou, V.A., Wistuba, I.I., Miller, V.A., Frampton, G.M., Wolchok, J.D., Shaw, A.T., Jänne, P.A., Stephens, P.J., Rudin, C.M., Geese, W.J., Albacker, L.A., Heymach, J.V., 2018. STK11/LKB1 Mutations and PD-1 Inhibitor Resistance in KRAS-Mutant Lung Adenocarcinoma. Cancer Discovery 8, 822–835. https://doi.org/10.1158/2159-8290.CD-18-0099

Slonim, D.K., Yanai, I., 2009. Getting Started in Gene Expression Microarray Analysis. PLoS Comput Biol 5, e1000543. https://doi.org/10.1371/journal.pcbi.1000543

Sorensen, S.A., Fenger, K., Olsen, J.H., 1999. Significantly lower incidence of cancer among patients with Huntington disease: An apoptotic effect of an expanded polyglutamine tract? Cancer 86, 1342–1346. https://doi.org/10.1002/(SICI)1097-0142(19991001)86:7<1342::AID-CNCR33>3.0.CO;2-3

Sos, M.L., Koker, M., Weir, B.A., Heynck, S., Rabinovsky, R., Zander, T., Seeger, J.M., Weiss, J., Fischer, F., Frommolt, P., Michel, K., Peifer, M., Mermel, C., Girard, L., Peyton, M., Gazdar, A.F., Minna, J.D., Garraway, L.A., Kashkar, H., Pao, W., Meyerson, M., Thomas, R.K., 2009. PTEN Loss Contributes to Erlotinib Resistance in EGFR-Mutant Lung Cancer by Activation of Akt and EGFR. Cancer Research 69, 3256–3261. https://doi.org/10.1158/0008-5472.CAN-08-4055

Spratt, D.E., Chan, T., Waldron, L., Speers, C., Feng, F.Y., Ogunwobi, O.O., Osborne, J.R., 2016. Racial/Ethnic Disparities in Genomic Sequencing. JAMA Oncol 2, 1070–1074. https://doi.org/10.1001/jamaoncol.2016.1854

Steigen, S.E., Bjerkehagen, B., Haugland, H.K., Nordrum, I.S., Løberg, E.M., Isaksen, V., Eide, T.J., Nielsen, T.O., 2008. Diagnostic and prognostic markers for gastrointestinal stromal tumors in Norway. Modern Pathology 21, 46–53. https://doi.org/10.1038/modpathol.3800976

Steuer, C.E., Behera, M., Berry, L., Kim, S., Rossi, M., Sica, G., Owonikoko, T.K., Johnson, B.E., Kris, M.G., Bunn, P.A., Khuri, F.R., Garon, E.B., Ramalingam, S.S., 2016. Role of race in oncogenic driver prevalence and outcomes in lung adenocarcinoma: Results from the Lung Cancer Mutation Consortium: Race and Genomics in Lung Cancer. Cancer 122, 766–772. https://doi.org/10.1002/cncr.29812

Stram, D.O., Park, S.L., Haiman, C.A., Murphy, S.E., Patel, Y., Hecht, S.S., Le Marchand, L., 2019. Racial/Ethnic Differences in Lung Cancer Incidence in the Multiethnic Cohort Study: An Update. JNCI: Journal of the National Cancer Institute 111, 811–819. https://doi.org/10.1093/jnci/djy206

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U.S.A. 102, 15545–15550. https://doi.org/10.1073/pnas.0506580102

Šutić, M., Vukić, A., Baranašić, J., Försti, A., Džubur, F., Samaržija, M., Jakopović, M., Brčić, L., Knežević, J., 2021. Diagnostic, Predictive, and Prognostic Biomarkers in Non-Small Cell Lung Cancer (NSCLC) Management. Journal of Personalized Medicine 11, 1102. https://doi.org/10.3390/jpm11111102

Tan, A.S., Baty, J.W., Dong, L.-F., Bezawork-Geleta, A., Endaya, B., Goodwin, J., Bajzikova, M., Kovarova, J., Peterka, M., Yan, B., Pesdar, E.A., Sobol, M., Filimonenko, A., Stuart, S., Vondrusova, M., Kluckova, K., Sachaphibulkij, K., Rohlena, J., Hozak, P., Truksa, J., Eccles, D., Haupt, L.M., Griffiths, L.R., Neuzil, J., Berridge, M.V., 2015. Mitochondrial Genome Acquisition Restores Respiratory Function and Tumorigenic Potential of Cancer Cells without Mitochondrial DNA. Cell Metabolism 21, 81–94. https://doi.org/10.1016/j.cmet.2014.12.003

Tang, R., Shuldiner, E.G., Kelly, M., Murray, C.W., Hebert, J.D., Andrejka, L., Tsai, M.K., Hughes, N.W., Parker, M.I., Cai, H., Li, Y.-C., Wahl, G.M., Dunbrack, R.L., Jackson, P.K., Petrov, D.A., Winslow, M.M., 2023. Multiplexed screens identify RAS paralogues HRAS and NRAS as suppressors of KRAS-driven lung cancer growth. Nat Cell Biol 25, 159–169. https://doi.org/10.1038/s41556-022-01049-w

Thomas, F., Delmar, P., Vergez, S., Rochaix, P., Hennebelle, I., McLoughlin, P., Benlyazid, A., Sarini, J., Delord, J., 2013. Gene expression profiling on pre- and post-erlotinib tumors from patients with head and neck squamous cell carcinoma. Head & Neck 35, 809–818. https://doi.org/10.1002/hed.23036

Tomczak, K., Czerwińska, P., Wiznerowicz, M., 2015. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn) 19, A68–A77. https://doi.org/10.5114/wo.2014.47136

Wang, C., Ding, M., Xia, M., Chen, S., Van Le, A., Soto-Gil, R., Shen, Y., Wang, N., Wang, J., Gu, W., Wang, X., Zhang, Y., Zen, K., Chen, X., Zhang, C., Zhang, C.-Y., 2015. A Five-miRNA Panel Identified From a Multicentric Case-control Study Serves as a Novel Diagnostic Tool for Ethnically Diverse Non-small-cell Lung Cancer Patients. EBioMedicine 2, 1377–1385. https://doi.org/10.1016/j.ebiom.2015.07.034

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10, 57–63. https://doi.org/10.1038/nrg2484

Warburg, O., 1956. On Respiratory Impairment in Cancer Cells. Science 124, 269–270. https://doi.org/10.1126/science.124.3215.269

Wassenaar, C.A., Conti, D.V., Das, S., Chen, P., Cook, E.H., Ratain, M.J., Benowitz, N.L., Tyndale, R.F., 2015. UGT1A and UGT2B genetic variation alters nicotine and nitrosamine glucuronidation in European and African American smokers. Cancer Epidemiol Biomarkers Prev 24, 94–104. https://doi.org/10.1158/1055-9965.EPI-14-0804

Weber, J.A., Baxter, D.H., Zhang, S., Huang, D.Y., How Huang, K., Jen Lee, M., Galas, D.J., Wang, K., 2010. The MicroRNA Spectrum in 12 Body Fluids. Clinical Chemistry 56, 1733–1741. https://doi.org/10.1373/clinchem.2010.147405

Weeks, W.B., Tosteson, T.D., Whedon, J.M., Leininger, B., Lurie, J.D., Swenson, R., Goertz, C.M., O'Malley, A.J., 2015. Comparing Propensity Score Methods for Creating Comparable Cohorts of Chiropractic Users and Nonusers in Older, Multiply Comorbid Medicare Patients With Chronic Low Back Pain. Journal of Manipulative and Physiological Therapeutics, Special Issue: Adverse Events 38, 620–628. https://doi.org/10.1016/j.jmpt.2015.10.005

Wei, D., Chen, W., Meng, R., Zhao, N., Zhang, X., Liao, D., Chen, G., 2018. Augmented expression of Ki-67 is correlated with clinicopathological characteristics and prognosis for lung cancer patients: an up-dated systematic review and meta-analysis with 108 studies and 14,732 patients. Respiratory Research 19, 150. https://doi.org/10.1186/s12931-018-0843-7

Weinberg, F., Hamanaka, R., Wheaton, W.W., Weinberg, S., Joseph, J., Lopez, M., Kalyanaraman, B., Mutlu, G.M., Budinger, G.R.S., Chandel, N.S., 2010. Mitochondrial metabolism and ROS generation are essential for Kras-mediated tumorigenicity. Proceedings of the National Academy of Sciences 107, 8788–8793. https://doi.org/10.1073/pnas.1003428107

WHO cancer fact sheet, 2022. URL https://www.who.int/news-room/fact-sheets/detail/cancer

Wikoff, W.R., Grapov, D., Fahrmann, J.F., DeFelice, B., Rom, W.N., Pass, H.I., Kim, K., Nguyen, U., Taylor, S.L., Gandara, D.R., Kelly, K., Fiehn, O., Miyamoto, S., 2015. Metabolomic Markers of

Altered Nucleotide Metabolism in Early Stage Adenocarcinoma. Cancer Prevention Research 8, 410–418. https://doi.org/10.1158/1940-6207.CAPR-14-0329

Wojcik, G.L., Graff, M., Nishimura, K.K., Tao, R., Haessler, J., Gignoux, C.R., Highland, H.M., Patel, Y.M., Sorokin, E.P., Avery, C.L., Belbin, G.M., Bien, S.A., Cheng, I., Cullina, S., Hodonsky, C.J., Hu, Y., Huckins, L.M., Jeff, J., Justice, A.E., Kocarnik, J.M., Lim, U., Lin, B.M., Lu, Y., Nelson, S.C., Park, S.-S.L., Poisner, H., Preuss, M.H., Richard, M.A., Schurmann, C., Setiawan, V.W., Sockell, A., Vahi, K., Verbanck, M., Vishnu, A., Walker, R.W., Young, K.L., Zubair, N., Acuña-Alonso, V., Ambite, J.L., Barnes, K.C., Boerwinkle, E., Bottinger, E.P., Bustamante, C.D., Caberto, C., Canizales-Quinteros, S., Conomos, M.P., Deelman, E., Do, R., Doheny, K., Fernández-Rhodes, L., Fornage, M., Hailu, B., Heiss, G., Henn, B.M., Hindorff, L.A., Jackson, R.D., Laurie, C.A., Laurie, C.C., Li, Y., Lin, D.-Y., Moreno-Estrada, A., Nadkarni, G., Norman, P.J., Pooler, L.C., Reiner, A.P., Romm, J., Sabatti, C., Sandoval, K., Sheng, X., Stahl, E.A., Stram, D.O., Thornton, T.A., Wassel, C.L., Wilkens, L.R., Winkler, C.A., Yoneyama, S., Buyske, S., Haiman, C.A., Kooperberg, C., Le Marchand, L., Loos, R.J.F., Matise, T.C., North, K.E., Peters, U., Kenny, E.E., Carlson, C.S., 2019. Genetic analyses of diverse populations improves discovery for complex traits. Nature 570, 514–518. https://doi.org/10.1038/s41586-019-1310-4

Wu, M., Miska, J., Xiao, T., Zhang, P., Kane, J.R., Balyasnikova, I.V., Chandler, J.P., Horbinski, C.M., Lesniak, M.S., 2019. Race influences survival in glioblastoma patients with KPS ≥ 80 and associates with genetic markers of retinoic acid metabolism. J Neurooncol 142, 375–384. https://doi.org/10.1007/s11060-019-03110-5

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., Yu, G., 2021. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation 2, 100141. https://doi.org/10.1016/j.xinn.2021.100141

Xintaropoulou, C., Ward, C., Wise, A., Queckborner, S., Turnbull, A., Michie, C.O., Williams, A.R.W., Rye, T., Gourley, C., Langdon, S.P., 2018. Expression of glycolytic enzymes in ovarian cancers and evaluation of the glycolytic pathway as a strategy for ovarian cancer treatment. BMC Cancer 18, 636. https://doi.org/10.1186/s12885-018-4521-4

Xu, F., Lin, H., He, P., He, L., Chen, J., Lin, L., Chen, Y., 2020. A TP53-associated gene signature for prediction of prognosis and therapeutic responses in lung squamous cell carcinoma. OncoImmunology 9, 1731943. https://doi.org/10.1080/2162402X.2020.1731943

Yuan, J., Hu, Z., Mahal, B.A., Zhao, S.D., Kensler, K.H., Pi, J., Hu, X., Zhang, Youyou, Wang, Y., Jiang, J., Li, C., Zhong, X., Montone, K.T., Guan, G., Tanyi, J.L., Fan, Y., Xu, X., Morgan, M.A., Long, M., Zhang, Yuzhen, Zhang, R., Sood, A.K., Rebbeck, T.R., Dang, C.V., Zhang, L., 2018. Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers. Cancer Cell 34, 549-560.e9. https://doi.org/10.1016/j.ccell.2018.08.019

Yuan, J., Kensler, K.H., Hu, Z., Zhang, Y., Zhang, T., Jiang, J., Xu, M., Pan, Y., Long, M., Montone, K.T., Tanyi, J.L., Fan, Y., Zhang, R., Hu, X., Rebbeck, T.R., Zhang, L., 2020. Integrative comparison of the genomic and transcriptomic landscape between prostate cancer patients of predominantly African or European genetic ancestry. PLoS Genet 16, e1008641. https://doi.org/10.1371/journal.pgen.1008641

Zavala, V.A., Bracci, P.M., Carethers, J.M., Carvajal-Carmona, L., Coggins, N.B., Cruz-Correa, M.R., Davis, M., de Smith, A.J., Dutil, J., Figueiredo, J.C., Fox, R., Graves, K.D., Gomez, S.L., Llera, A., Neuhausen, S.L., Newman, L., Nguyen, T., Palmer, J.R., Palmer, N.R., Pérez-Stable, E.J., Piawah, S., Rodriquez, E.J., Sanabria-Salas, M.C., Schmit, S.L., Serrano-Gomez, S.J., Stern, M.C., Weitzel, J., Yang, J.J., Zabaleta, J., Ziv, E., Fejerman, L., 2021. Cancer health disparities in racial/ethnic minorities in the United States. Br J Cancer 124, 315–332. https://doi.org/10.1038/s41416-020-01038-6

Zhan, P., Wang, J., Lv, X., Wang, Q., Qiu, L., Lin, X., Yu, L., Song, Y., 2009. Prognostic Value of Vascular Endothelial Growth Factor Expression in Patients with Lung Cancer: A Systematic Review

with Meta-Analysis. Journal of Thoracic Oncology 4, 1094–1103.
https://doi.org/10.1097/JTO.0b013e3181a97e31