Genetic variants associated with chemo-resistance/recurrence in head and neck squamous cell carcinoma (HNSCC)

A thesis submitted

by

Cathal King

to

The Discipline of Bioinformatics,
School of Mathematics, Statistics & Applied Mathematics
National University *of* Ireland, Galway

in partial fulfillment of the requirements for the degree of

M.Sc. in Computational Genomics

August 10th 2018

Thesis Supervisor(s): Dr. Pilib O Broin

Declaration

I, Cathal King, declare that this thesis titled, 'Genetic variants associated with chemo-resistance/recurrence in head and neck squamous cell carcinoma (HNSCC)' submitted to the Discipline of Bioinformatics, School of Mathematics, Statistics and Applied Mathematics National University of Ireland, Galway in partial fulfilment of the requirements for the degree of M.Sc in Computational Genomics is entirely my own work.

I have acknowledged all main sources used to complete this project. Should it be requested, I freely give permission to the library at NUIG that this thesis may be loaned or copied

Signed: _____

Cathal King

August 2018

Abstract

This project involved the analysis of Whole Exome Data from a pilot study of three HNSCC patients, with matched normal, tumour, and recurrence biopsies. Comparison of tumour-normal, recurrence-normal, and recurrence-tumour data provided information on the specific variants that may have allowed these tumours to evade treatment/recur following chemo-radiation.

Little is known about specific variants that confer HNSCC tumours with chemo-resistance. To identify these mutational signatures, the Genome Analysis Toolkit (GATK) best practice workflow along with the Picard suite of preprocessing tools was followed to generate analysis ready reads. A collection of databases and publicly available resources of human genetic variation was used to annotate, analyse and visualise genetic variants classify significance.

154 variants were shown to have been exposed to chemo-radiation and conferred the tumour with resistance to treatment. This gave insights into the mechanisms at play such cell cycle control and mitotic timing events and provides potential for therapeutically actionable events for future research and treatment of HNSCC. Mutations were found on important genes such as *CTAGE5*, mucin family genes, *OR2L3*, *CDC27*, *NBPF10* and *NOTCH2A* and on important pathways such as the olfactory receptor pathway and the Wnt signalling pathway.

Contents

List of Figures

List of Tables

1 Introduction

1.0.1 Head and Neck Squamous Cell Carcinoma

Cancer arises throughout the accumulation of genetic and epigenetic changes in coding and non-coding genes. These genes play important roles in many signalling pathways and changes in them lead to cancinogenesis. These changes can be broadly defined as mutations; gains or losses at the nucleotide base pair level and SNP's, indels, copy number abberations and chromosome gain or loss are common carcinogenic mutations. In the past, the importance of mutations with respect to cancer development seemed elusive. What can be said now is that mutations, especially in tumour suppressor genes, are significant and contribute toward cancer evolution[1].

In the head and neck region, the most common type of cancer found is one that originates from squamous cells[2]. As a type of epithelial malignancy, a squamous cell carcinoma develops in the mucosal linings of the upper aerodigestive tract, the tumours often grow within preneoplastic fields (lesions) of genetically altered cells and are unexpectedly heterogeneous in nature [2],[3]. Head and Neck cancer can also originate from salivary glands [1]. However, they are uncommon and further categorised based on the type of cell they originate from. Head and Neck Squamous Cell Carcinoma (HNSCC), however, is the most pervasive and can be further classified by anatomical site. Tumours are classified by location and typically occur in the oral cavity, oropharynx, nasal cavity and paranasal sinuses, nasopharynx, larynx, or hypopharynx [1] and can be divided into 14 more subsites.

Most HNSCC patients present with tumours de novo, however in some cases precancerous lesions often develop in the mucosal linings that are often visible

to the naked eye. These precursors to carconogenesis contain tumour associated mutations that should be monitored [4]. Tumours are more likely to develop in a lesion such as leukoplakia rather than normal tissue. These lesions present an opportunity to define the timing of genetic events and thus characterise the molecular landscape of HNSCC. The genomic predictive model of HNSCC, published in 2017, contains details of mutations with accompanying genomic regions.

1.0.2 Etiology and Epidemiology

Excessive tobacco use and alcohol consumption are considered to be the typical risk factors of HNSCC. Certain high-risk human papillomaviruses (HPVs) are involved in a more substantial subgroup of these tumours[5]. HNSCCs are divided into HPV-negative (HPV-ve) and HPV-positive (HPV+ve) diseases [2] where the tumours are different at a molecular level and the clinical outcome is different and HPV+ve HNSCC tumours are seen to have a more favourable prognosis. HPV+ve cases are seen to be biologically distinct and originate from different genetic alterations than those associated with alcohol and tobacco related cases [6]. However, the quantitative assessment of HPV related HNSCC can be challenging given the multifactorial etiology largely driven by tobacco and alcohol usage [5]. Other risk factors include preserved or salted foods, infection with the Epstein-Barr virus and poor oral hygiene [1]. Asian descent, or Southern Chinese descent more specifically, is also considered to be a risk factor for nasopharyngeal carcinoma. This is due to inherited genetic susceptibility coupled with gene-environment interactions[7].

Worldwide, HNSCC is the sixth leading cancer by incidence and affects approximately 600,000 patients annually. Cases are more common in the developed world and only 40-50% of patients will survive for 5 years[3]. HNSCC incidence rates are seen to be higher in areas of the globe that have high levels of tobacco and alcohol consumption [8]. Certain areas have witnessed a decline in oral cavity cancer correlated to a decline in tobacco usage. However, countries such as the USA, the UK, Denmark, Sweden, Norway and more have seen increased rates of oropharyngeal and oral cavity cancer despite declining smoking rates [9].

That further associates HPV with increased HNSCC risk and is why geographical and socioeconomic factors are considered for each patient, especially in HPV+ve cases.

1.0.3 Detection & Treatment

Symptoms of HNSCC are agressive, varied and usually related to the location of the tumour but mainly include pains in the neck or throat, swelling and paralyzed face muscles. A late-stage diagnosis is common with HNSCC. In fact, more than half of HNSCC's are diagnosed at a late stage [10]. This is problematic as prognosis, survival rate and quality of life after treatment is directly related to its stage of diagnosis. Late diagnoses are largely due to late presentation or early diagnosis of a late stage tumour. In other cancers, a late stage diagnosis can be attributed to low socio-economic status or limited access to health care and this may well be the case for HNSCC's also.

An operative needle biopsy under anaesthesia is the traditional method of diagnosis. More recently, an ultrasound guided needle biopsy is becoming a more accurate method of obtaining a tissue diagnosis due to the additional imaging details that support accurate disease staging [11]. Tumours are staged according to Tumour Node Metastasis (TNM) staging; a notation system that gives details of the cancer such as original tumour size and metastatic tumour characteristics.

TNM stage and anatomical location are important factors used to direct treatment decisions. Comorbidity, age and institutional factors also play an important part in the treatent decision making process. Newer tumour protocols which came into effect in January 2018 also highlighted the importance of tumour depth of invasion when determining staging and treatment options, especially in the case of oral cavity tumours (1)[2]. Early stage tumours are treated with surgery or radiotherapy. For advanced tumours, surgery combined with postoperative chemoradiation or upfront chemoradiation is the most common protocol. As HNSCC is an epithelial malignancy, induction chemotherapy is seen to have no benefit[12]. Other clinical treatment options include transoral robotic resection where a surgical robot is used to remove tumours from difficult to access areas such as the

throat and application of the epidermal growth factor receptor (EGFR)-specific antibody cetuximab in combination with radiotherapy [13]. The classical clinical tumour characteristics are the mainstays when it comes to staging and clinical diagnosis overall. This is lacking in terms of the wider array of treatment options that are becoming available for this diverse disease, especially inadequate in an era of precision medicine where patients require personalised treatment.

1.0.4 Genetic factors of HNSCC

Drivers and Passengers in HNSCC

Mutations, genes and pathways can be classified as driver or passenger. Driver mutations are ones that contribute directly to the development of cancer. For example, a typical driver gene in HNSCC is *TP53*, which mutates early in cancer and typically mutates in one or both alleles in all cancer cells [2]. Passenger mutations play a secondary role in cancer, usually occur during the growth of the cancer and far exceed the number of driver mutations present [14]. Driver and passenger genes are ones that contain such mutations and contribute to carcinogenic pathways [15].

Identifying driver genes or mutations is a complex procedure and techniques vary. The relevance of the vast array of somatic mutations that appeared needed to be considered. Many mutations were appearing in olfactory receptor genes and the driver passenger concept aimed to quantify these mutations and account for biases. Driver mutations are typically classified based on their frequency of occurrence in a cohort of samples or according to their predicted functional impact on protein sequence or structure [16]. Driver gene sets meet the criteria of covering a large number of samples with high coverage and the mutations that they contain usually exhibit mutual exclusivity. This means a single mutation would be enough to disturb one pathway [17], [18],[19]. Driver gene sets then follow on from this approach. A technique was developed that combines the coverage and exclusivity of a gene set and maximizes it using a Markov Chain Monte Carlo (MCMC) method to extract driver gene sets. The assumptions made are that most patients will have a mutation in some gene in the cancer pathway and

that a single driver mutation in that pathway is enough to disrupt the pathway [20]. This MCMC approach is used to sample from a set of genes according to a distribution that results in a higher probability given two sets of genes that have high coverage and exclusivity. More recent methods to classify driver and passengers include incorporating the concept of Variant Allele Frequency (VAF), which is the frequency of a variant at a particular locus in the population[2]. A tumour cell is expected to have a VAF of approximately 50% or 100%. However, given the fact that tumours contain stroma and driver mutations may occur in sub-populations of cancer cells, a tumour cell line will have a VAF of close to 5% or 10% [2].

The frequency of cancer mutations tend to be associated with tumour type and mutagen exposure[21]. A study found that somatic mutations appear in genome regions that replicate late during S-phase which is thought to be related to gene expression related DNA repair and/or nucleotide shortage during replication[21]. Correction algorithms were used that take into account the number of biases presented. When tested on sequencing data from a cancer data set, the number of driver genes presented shrank from 450 to 11. This was due to the fact that many of the original 450 genes contained olfactory genes and genes that were highly enriched that encoded large proteins [21]. Therefore these algorithms available are expected to be used with caution. However, they do provide opportunities to analyse large cancer cohorts and account for specific biases.

Identifying cancer drivers is important in precision oncology. The most comprehensive characterisation of cancer driver genes and mutations to date was released this year. A landscape of cancer driver genes were found including 299 genes and details of a cancer driver gene discovery workflow was included [22]. The fact that not all mutations in a cancer driver gene have equal impact [23] and that subsequent consequences depend on the mutations position within the protein and amino acid [24] was the foundation of the approach to driver mutation discovery. 26 different computational tools from a variety of institutions were used. The first phase, driver gene set discovery, used 8 computational tools based on mutational frequency, features, clustering and externally defined regions. The second phase, gene and in-silico mutational validation, used 16 computational tools including algorithms to identify clinically actionable events.

In the case of HNSCC, between 50 and 100 genes are found to be mutated regularly and are candidate cancer driver genes[2]. Many of these genes are mutated at low frequencies and are only considered candidate driver genes of HNSCC. A collection of molecular evidence such as a gene being linked to a pathway or its presence in a cancer database such as the Catalogue of Somatic Mutations in Cancer (COSMIC) as well as some correction algorithms was the approach used in the last large HNSCC molecular catalogue[3]. Also, this study did not consider the genetic subgroup of HNSCC (HPV+ve, HPV-ve etc.) which has shown to be an important factor when evaluating this disease[25]. In fact, genes which are frequently mutated in HNSCC were shown to be mostly unaffected in HPV+ve cases[26]. These include *TP53* and *CDKN2A*.

Mutational Signatures in HNSCC

Given the metastatic nature and high recurrence rate of most types of HNSCC's, molecular signatures and biomarkers continue to be highly sought after to predict and prevent advancement of the disease. Most are specific to the type and etiology of the cancer. In 2015, The Cancer Genome Atlas consortium published an in depth molecular catalogue on HNSCC [26]. In that, HPV associated tumours were shown to be dominated by helical domain mutations of the oncogene *PIKC3A*, novel alterations involving loss of the protein coding gene *TRAF3*, and amplification of the cell cycle gene *E2F1* [1]. Smoking related tumours show near universal loss of function *TP53* mutations and *CDKN2A* inactivation with frequent copy number alterations. Subgroups of HNSCC contained loss of function alterations in Wnt signalling pathway genes *AJUBA* and *FAT1* and also activation of nuclear factor gene *NFE2L2*, mostly in laryngeal tumours. Most mutations, as expected, were associated with tumour supressor genes in cell cycle control, cellular growth and survival.

The most significantly mutated genes in HNSCC are depicted in figure 1.1. HNSCC genomes show high instability with a mean of 141 Copy Number Aberrations (CNA's) (amplifications or deletions) and 62 structural aberrations (chromosomal fusions) per tumour from high coverage whole genome sequencing (WGS)[1]. 39 regions of recurrent copy number loss and 23 regions of copy number gain were also observed. Structural alterations such as fusion events in known oncogenes

such as *ALK*, *ROS* and *RET* were not observed in HNSCC. However, known *FGFR3-TACC3* fusion events were present in some HPV+vs tumours (paper 3). Whole exome sequencing (WES) revealed mutated genes in regions of CNA's and listed in the COSMIC database (figure 1.1). Transversions at CpG sites occurred more in HPV-ve tumours and an excess of TpC mutations were present in HPV+ve tumours. A frequently mutated gene, the nuclear receptor binding SET domain protein 1 (*NSD1*), was identified in 33 HNSCC's.

The interactions that HNSCC tumours have with the host immune system is the most recent area of focus. In general, HNSCC tumours are grouped into HPV+ve and HPV-ve tumours where HPV-ve tumours are ones that are characterised based on tobacco and alcohol related mutations in genes such as *TP53* and *CDKN2A*. HPV+ve tumours occur due to the integration of the viral genome from HPV into the host genome which results in expression of the E6 and E7 viral oncoproteins [25]. The immune microenvironment and HNSCC development is observed by changes in immune cell populations and checkpoints among other indicators. This year, more biomarkers have emerged that help predict benefit from therapy, mainly in HPV+ve HNSCC. These include PD-L1 expression, PD-L2 expression and the interferon gamma gene signature [25]. Ignoring HPV+ve tumours, there are little reliable biomarkers available to date. Similarities have been found with tobacco and alcohol related tumours (HPV-ve) to lung squamous cell carcinoma [27]. Interestingly, the lung is the region where most metastastic HNSCC tumours occur. Overall however, the range of mutations found in HNSCC is complex and varied.

1.0.5 Recurrence and metastasis in HNSCC

In terms of recurrent tumours, biomarkers are again grouped based on HPV status. For the case of HPV+ve tumours, immune therapies such as nivolumab and durvalumab target the programmed death 1 receptor (PD1) or its ligand PD-L1, which is involved in programmed cell death, and has shown positive outcomes in a subset of recurrent HNSCC's. These results are not as reliable for HPV-ve cases, as is the trend, and recurrent tumours still have the reputation of the most difficult ones to treat.

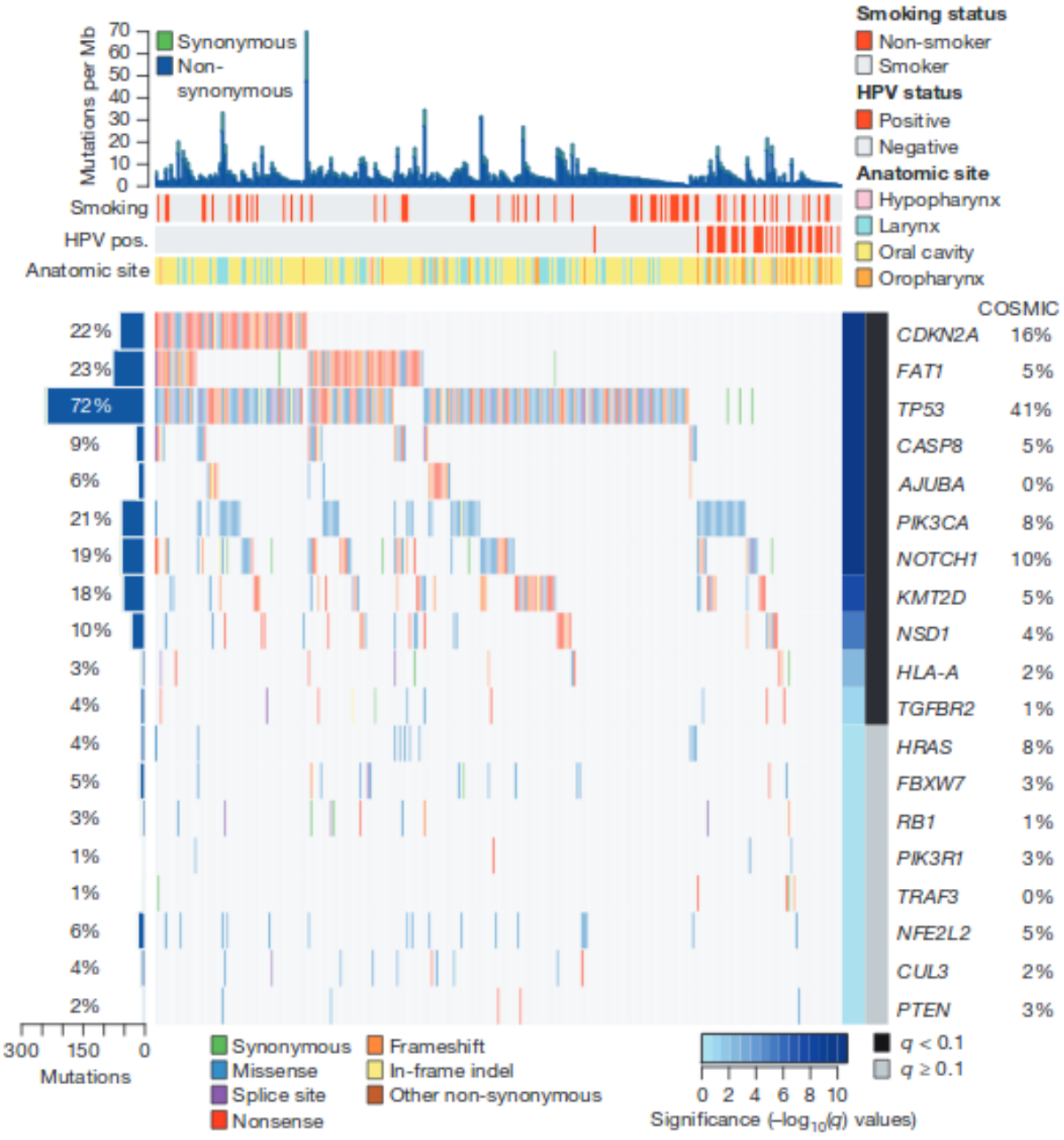

Figure 1.1: Significantly mutated Genes (rows) in HNSCC are ordered by q value. Columns (samples) are arranged to represent mutual exclusivity among mutations. On the left is the mutational percentage in TCGA. The right shows mutational percentage in COSMIC. The top shows overall number of mutations per megabase with the colour coding representing mutation type.[1]

HNSCC has a high likelihood of recurring post treatemt. The disease will recur in one half of patients with recurrence at the locoregion most common i.e. at the same site, followed by a high likelihood of distant metastasis. Recurrent tumours are notoriously difficult to treat, harbouring mutational signatures that evaded the treatment of choice. The likelihood of preventing recurrence increases with punctual tumour diagnosis. Consistent HNSCC genomic profiles build a predictive model, bring about earlier diagnoses and assist in individualised clinical management. Aside from the psychological impact, a recurrent or matastatic HNSCC tumour has poor prognosis with a median survival time of usually 6 to 12 months[25]. Salvage therapy for these patients depends on a number of factors including the presence of a distant metastasis, if the tumour is locally recurrent and the response to the initial treatment. Indeed, biomarkers in recurrent HNSCC tumours do little to influence the next step of treatment at present, however research is active in this area.

A tumour biopsy may not be a single cell type but a heterogenous mix of cells. This poses difficulties when detecting variants. A sample could be a group of cells that are in different stages of differentiation or even clonal populations that have been selected for because they confer the tumour with some type of evolutionary advantage that may provide resistance to therapeutics for example. Sub clones might not contain variants and cross contamination of tumoural DNA into normal germline samples are some problems that may arise when exposing variants. Intratumoural heterogenity refers to the different morphological and phenotypic profiles presented when taking biopsies. It is also a problem during the treatment of cancer and may explain why cells within a tumour show different therapeutic responses to chemotherapy and radiation therapy. Genetic changes in tumours that occurred when treated may explain this[28]. Others hypothesise the existence of cell populations with stem cell properties to be the reason[29]. The relevance of this hererogenity amoungst cells in the context of HNSCC is shown in a mathematical algorithm developed that correlates greater tumoural heterogenity with poorer outcomes in HNSCC patients [30].

Field cancerization is a term used to refer to the likelihood of a tumour recurring post treatment and the likelihood that multiple tumours will develop in

the head and neck mucosa[3]. A study linked the frequent dysplastic (abnormal) changes around oral cancer tumours to the occurance of local recurrences and multiple primary tumours [31]. Field cancerisation can be thought of as the molecular signatures observed in tumours. Loss of heterozygosity at chromosomes 3p, 9p and 17p tend to occur in dysplasia indicating early carcinogenesis and alterations at chromosomes 11q, 4q and of chromosome 8 were usually evident in existing carcinomas [32]. These genetic signatures coupled with TP53 mutations showed that in at least 35% of HNSCC's analysed, the cancer was surrounded by mucosal epithelium that had genetic changes [33]. This epithelium had a normal appearance but were histologically dysplastic. Interestingly, these recurrent cancerous fields are often found in the surgical margins when the tumour is excised meaning that these genetic changes can remain in the patient [33]. Other studies highlighted the importance of these fields in terms of local recurrences and the second primary tumours so often seen in HNSCC patients[34].

The existence of one or more mucosal areas consisting of epithelial cells that have carcinogenic genetic alterations[3] is an important aspect of field cancerisation and HNSCC. A precursor field is monoclonal in origin and does not show invasive growth or metastatic behaviour. A field may have histological aberrations but not necessarily. A field may be the source of local recurrences or second primary after surgical removal of the initial carcinoma [3]. Extra mutations are required to change a field into a new carcinoma. Importantly, the field and the primary tumour share genetic alterations and have a common clonal origin [3]. Mutations at chromosome 9p, decreased cytokeratin 4 expression, decreased comulin expression and p53 immunopositivity are genetic markers that may predict the risk of a field developing into a cancer [35]. Furthermore, the existence and number of mutations, typically chromosome 9p loss, chromosome 3p loss and chromosome 17p loss are all associated with the risk of progression in HNSCC and all cancer patients [36]. What precedes these fields is uncertain, however, small, p-53 positive focal patches in tumour-adjacent mucosal epithelium were observed [37]. TP53 mutations were found but were not identical to the mutations in the tumours themselves suggesting that these field patches are not clonally related to the tumours. These field patches are considered to be 'clonal units' meaning a family of cells derived from a common progenitor cell or adult

stem cell that comprises the squamous epithelium [38] and are detectable by mutations of p53. These 'clonal units' represent the first carcinogenic changes in the mucosa and with the field/patch tumour metastasis, provide a model for HNSCC development.

HNSCC behaves typically in terms of other cancers in terms of metastasis. Most metastatic tumours tend to develop near lymph nodes [39]. The presence and number of lymph node metastases are important prognositc markers of distant disease and survival[3]. In HNSCC, the *CSMD1* gene, located on chromosome 8, has been linked with invasion and metastases [40]. The *CSMD1* locus was shown to be significant and associated with outcome in patients with squamous cell carcinoma of the supraglottic larynx [41]. However, as of late no cancer gene has been definitively linked to invasion and metastasis in HNSCC. However the act of metastasis does seem to be a biology driven process in that a certain expression profile in the primary tumour can predict the presence or absence of lymph node metastasis, suggesting that certain genes might drive metastasis. These profiles contained a large number of genes associated with the process of epithelial to mesenchymal transition (EMT) which is a process where cells change from an epithelial phenotype to a mesenchymal phentype. EMT is frequently seen in cancer cells and more relevantly linked to invasion and metastasis. The tyrosine kinase NTRK2 and its ligand seem to have a crucial role in this process in HNSCC. Also, the TGFβ pathway has been shown to be a huge factor of the EMT transition[42].

1.0.6 Pathogenesis & oncogenic pathways in HNSCC

A signalling pathway is a regulatory system that contains core processes such as genome maintenance, evasion of growth supressors and angiogenesis. Cancerous pathways develop when pathways are activated or inactivated by genetic or epigenetic mutations. Knowledge on cancerous pathways developed from exploration of the functions of cancerous genes. Gene families, such as the protein kinases, are recurrent players in most cancer gene sets. Also, cancer genes tend to cluster on signalling pathways such as the MAPK/ERK pathway [43]. Other common cancer pathways include the TP53 regulatory system, the PI3K/AKT pathway [44]

and the cell cycle regulatory network centred around RB1; a common tumour supressor gene. They all interact with each other and other pathways are connected to them [45]. Common cancerous genes cluster on some signalling pathways. For example, upstream mutations are found in cell-membrane bound receptor tyrosine kinases such as *EGFR*, *ERBB2*, *FGFR2* and more. A mutational analysis on gliomas (a type of tumour) found that almost all cases have a mutation at one of the genes on the critical signalling pathways [46]. Pathways can also be classified into driver and passenger whereby driver pathways are groups of genes that contain driver mutations [47].

1.0.7 Cell cycle control

The most common mutations in HNSCC are losses of chromosomes 3p, 9p and mutations of TP53 [32]. CDKN2A, a tumour supressor gene, is located on chromosome arm 9p21 and encodes the p16 protein that binds and disturbs the cyclin D-CDK4 and cyclin D-CDK6 complexes. Loss of CDKN2A combined with the frequently observed amplification of cyclin D1 on 11q13 drives cells through the G1-S checkpoint of cell cycle and is a factor in unscheduled DNA replication [2]. This unwanted DNA replication typically leads to DNA damage and p53 activation [48]. The p53 protein is a key tumour supressor with a lot of functions including the induction of p21, another CDK inhibitor that arrests the cell cycle (38)[49]. p53 is also an inducor of apoptosis [49]. The TP53 gene is frequently mutated and inactivated in HNSCC, mostly by missense mutations and allelic loss [33]. 60-80% of HNSCC's contain somatic mutations in the TP53 gene and in 84% of HPV-ve HNSCC's [3]. Apart from the proteins disturbing the cell cycle itself, the genes of several growth factor receptors have been identified as candidate driver genes which might influence HNSCC onset, including EGFR and MET. However, due to their low mutational frequency, they have not been identified as candidate driver genes. EGFR, however, has seen to be frequently amplified and is supported as a driver gene in HNSCC [26]. EGFR also has pleitropic functions and links to intracellular pathways such as MAPK signalling, PI3K signalling and nuclear signalling [50]. A target of MAPK signalling is cyclin

D1 which also suggests that growth factor receptors are functionally linked to cell cycle control [26]. As cell cycle control is deregulated at the G1-S transition in HNSCC, cell cycle regulation will depend on the S phase to ensure proper cell division and may present cell cycle checkpoints that may represent targetable functional mechanisms [51].

1.0.8 The Wnt signalling pathway in HNSCC

The *FAT1* gene, which is a cadherin related gene and is located on chromosome 4q35.2 has shown a plethora of inactivating mutations[26]. This gene has shown to be mutated in 23% of HNSCC cases and lost or deleted in 8% of cases [52]. One of its purposes is to encode a large membrane protein which is part of the larger cadherin superfamily and these proteins contain a multitude of E-cadherin domains [53]. E-cadherin is the main player in the cadherin group which are calcium dependent transmembrane adhesion molecules which can form homotypic and heterotypic adhesion structures (43)[53]. *FAT1* has been associated with cell-cell contacts and regulation of actin dynamics[54] and of late, *FAT1* has been shown to have a role in Wnt signalling [55]. Essentially, recurrent somatic mutations of FAT1 in carcinogenesis is associated with abnormal Wnt activation.

The LIM domain containing protein *AJUBA* is a tumour supressor protein that is linked to HNSCC and is identified in the Wnt pathway [26]. There are three *AJUBA* family members; *AJUBA* itself, LIM domain containing protein 1 (LIMD1) and Wilms tumour protein 1-interacting protein (WTIP). They are all part of the cytosolic LIM domain protein family that are involved in a multitude of cellular functions[56], including cell division, cell-matrix adhesion, and cell-cell adhesion [57]. These proteins are also associated with the Hippo signalling pathway [58]. In keratinocytes (an epidermal cell that produces keratin), *AJUBA* also interacts with cadherin dependent cell-cell adhesive complexes through α-catenin that moves to the nucleus near to β-catenin (49) [59] and determines cell fate during early development [60]. Inactivation of *AJUBA* might cause increased β-catenin levels [2]. *FAT1* and *AJUBA* alterations are mutually exclusive, meaning that both events cannot occur and suggesting that the proteins encoded by these

two genes function in the same pathway [55] [61]. However, *AJUBA* has multiple functions and its precise role in HNSCC is not fully understood[2].

Another tumour supressor gene associated with the Wnt pathway that may play a role in HNSCC is *NOTCH1*. The Notch family has four main receptors namely *NOTCH1* - *NOTCH4* [2]. The Notch receptors bind to membrane bound ligands on other cells, are cleaved, and the Notch intracellular domain translocates to the nucleus and acts as a transcription fator[62][63]. In HNSCC, *NOTCH1* is classified as a tumour supressor gene [64]. More recently, the function of *NOTCH1* is more uncertain [65] as a more thorough evaluation of the mutations is needed. A link was found between Notch and β-catenin whereby membrane bound Notch can bind to active β-catenin and regulate its expression by degradation [66]. There has also been links demonstrated between the NOTCH1 and Wnt signalling pathways [67]. Following on from that, inactivating mutations in HNSCC were found in both *FAT1* and *NOTCH1*, both of which are associated with the Wnt pathway as previously stated. This suggests that carcinogenesis may develop through inactivation of FAT1 coupled with inactivation of NOTCH1 which increases β-catenin levels.

1.0.9 Somatic mutation rate

The fundamentals of tumour and carcinogenic development stem from mistakes during the DNA replication and repair processes; these mistakes presenting in the form of mutations. Traditionally, it was assumed that mutations occurred at the same rate across exons and introns. However, somatic mutations were shown to occur less frequently in exons than expected due to a higher mismatch-repair (MMR) activity in these genomic regions. A study published in 2017 [68] found that mutation rate varies across the genome due to replication timing, chromatin compaction and level of gene expression. That prior assumption has an impact not only on evolutionary biology but on DNA repair mechanisms and cancer genomics for example in driver gene detection, especially true given the fact that the study used somatic mutations found in colorectal and uterine cancerous tumours [68]. It was demonstrated that even in the absence of purifying selection,

exons acquired fewer mutations than expected across seven tumour types. Conditional probabilities were used to calculate the expected mutation rate and found a larger decrease in mutation rate in MMR-proficient tumours and no decrease in MMR-deficient tumours[68].

1.0.10 My Project

This project will involve the analysis of WES data from a pilot study of three HNSCC patients, with matched normal, tumour, and recurrence biopsies. Comparison of tumour-normal, recurrence-normal, and recurrence-tumour data, will provide information on the specific variants that may have allowed these tumours to evade treatment / recur following chemoradiation. This will provide important insight into the mechanisms at play, and potentially, help to identify alternative treatment targets.

To establish the somatic variants presents, the pipeline largely followed that of the Broad Institutes GATK best practices workflow. The initial stage of the pipeline will involve quality control (QC) analysis, trimming of sequence reads where necessary, mapping reads to the reference human genome and pre-processing of reads with the Broad's Picard suite of tools to generate analysis reads reads. These reads are then prepared for the accurate DREAM winning challenge somatic mutation caller Mutect2. The samples are labelled HN51, HN60 and HN72; one for each patient. Raw sequence data is presented in the form of paired end reads; referring to the two ends of the same DNA molecule. Sample HN72 was split across multiple flow cells/lanes in the sequencer and is referred to as HN72AC and HN72AH in places until it was combined into a single BAM file with the MarkDuplicates pre-processing tool step, explained below.

The overall aim of the project is to add to the wealth of knowledge available in Cancer Genomics. Accurate somatic variant calling is at the forefront of genomics as we see clinical settings worldwide develop workflows for prompt analysis, diagnosis and treatment of patients as it can provide a solution to many of the traditional hurdles presented by these diseases.

2 Methods

The pipeline used for this analysis is based on a combination of different tools and packages commonly used in variant calling work-flows. This analysis was primarily performed using parameters that were tried with different approaches before deciding on the final analysis. The tools used were typically downloaded using the wget command and analysis was primarily performed through the Sun Grid Engine (SGE) queing system and RStudio. As the entire pipeline is exhaustive, some code is included in this chapter. Nonetheless, the complete pipeline with accompanying parameters can be viewed at the GitHub repository https://github.com/cathalgking/MSc_THESIS

2.0.1 Sequencing Data

The sequencing data in this project is generated from 100bp PE libraries on an Illumina HiSeq machine. The analysis is on whole-exome-sequencing data (WES) from three HNSCC patients, with matched normal, tumour, and recurrence biopsies. The files are in fastq format which is usually the starting format for most NGS sequencing analyses due to its simplicity and readability (ref). The format is text based with four lines containing the sequence identifier (read group), beginning with an @ symbol, the raw nucleotide sequence, a '+' sign and the corresponding Phred quality score for each nucleotide which is the measure of the probability that a nucleotide is incorrectly called. Each fastq file is ~ 30G. Samples such as HN72 were split across multiple lanes/flowcells and were combined in the MarkDuplicates stage of the analysis. The mkdir command was used to group the samples into three sets; Normal, Tumour and Recurrence the following pipeline was repeated for each sample.

2.0.2 Multi QC and Trimming Reads

A summary of the sequencing data is made prior to analysis using the FASTQC package. This produces a report of the sequencing data to achieve an overall understanding of each samples quality and to determine the pre-processing steps required. MULTIQC combines the individual FASTQC outputs and makes one inclusive HTML report.

Trimming/adapter removal from sequence reads is necessary to ensure a high read quality and Trimmomatic was the tool used to do this. Trimmomatic was downloaded using wget to the local directory and contained data on the adapter sequences made by Illumina. These adapter sequences are identified by the MULTIQC report and are removed to prevent interference with read mapping in further downstream analysis. A Phred score of 33 was set, meaning any piece of read below 33 will get trimmed. Other parameters such as the sliding window was set to 4:15. This means trimming when the average quality per base drops below 15 in a 4 base window. Other parameters were set such as 'minlen', which drops reads below a specified length of 36 base pairs.

Two input fastq files with the samples to be trimmed and the path to them were provided. They refer to P1(forward reads) and P2 (reverse reads). Then, four output fastq files are generated (forward paired, forward unpaired, reverse paired and reverse unpaired). Paired refers to the reads that were not trimmed or kept. Unpaired refers to the reads that were discarded as they fell below a certain threshold. Details of trimmed reads can be found in the generated trimlog file, for example 'tumourHN60.trimlog', which contains details such as the location of the first surviving base, the location of the last surviving base and the amount trimmed from the end.

2.0.3 Aligning Reads to the Human Genome

There are many different alignment programs that have been developed that accurately map sequences to a reference genome. The most widely accepted gold standard in terms of speed and accuracy is the Burrows-Wheeler Aligner (BWA). BWA was used to align the reads to the reference genome (hg19.fa). BWA has three algorithms with the one most applicable to this project is BWA-MEM

(maximal exact matches). It tolerates errors given longer alignment and will work well given 2% error for a 100bp alignment. BWA uses a computational method known as 'indexing' to speed up its mapping algorithm, created by the Burrows-Wheeler transform. Using this transform, the entire genome is compressed into an index of 2.2GB.

BWA-MEM follows the well understood seed-and-extend paradigm. It initially seeds an alignment with supermaximal exact matches (SMEMs) which essentially finds at each query position the longest exact match covering the position. However, occasionally the true alignment may not contain any SMEMs. Group of seeds that are co-linear and close to each other are a chain. The seeds are chained while seeding and then short chains are filtered that are usually contained in a long chain and are much worse than the long chain (shorter in bp length than the long chain). Chain filtering aims to reduce unsuccessful seed extension at a later step. Each chain may not always correspond to a final hit. Seeds are then ranked by the length of the chain it belongs to and then by the seed length. In a ranked list, a seed is dropped if it is already contained in an alignment found before or extend the seed with a banded affine-gap-penalty dynamic programming (DP) if it potentially leads to a new alignment.

Different parameters are specified depending on how rigorous and specific the alignment is required to be. The output format for the alignments is specified as a BAM (Binary Alignment Map) file, the binary version of a SAM (Sequence Alignment Map) file. A BAM file can be converted to a SAM file easily. A SAM file is a readable format consisting of a header, which provides information regarding the project and the genome and an alignment section which is split into eleven fields providing information regarding the template name, the location, strand and quality of each sequence (ref-thesis bookmarks). Figure 2.1 shows the code used to map one of the fastq files to the reference sequence. The input is two fastq files, one forward and one reverse sequenced. The output is a combined BAM file.

```
#mappingHN51
bwa mem -M -t 8 -R '@RG\tID:BD1LYPACXX.3\tSM:HN52\tLB:I12' \
hg19.fa HN51_S1_tumor.BD1LYPACXX.lane_2_P1_I11_For_PAIRED.fastq \
HN51_S1_tumor.BD1LYPACXX.lane_2_P2_I11_Rev_PAIRED.fastq | samtools view -Sb - > HN51.tumour.bam
```

Figure 2.1: The command used to map sample HN51 to the human genome using BWA-MEM. -M marks split hits as secondary and is essential for Picard compatibility downstream. -t specifies the number of threads or cores the CPU should use. -R is the read group of the sample; explained below. The output is piped to samtools which generates the aligned BAM file.

2.0.4 Methods for Somatic Variant Calling

An overview of pipeline used to generate filtered, analysis ready variants can be seen in figure 2.2.

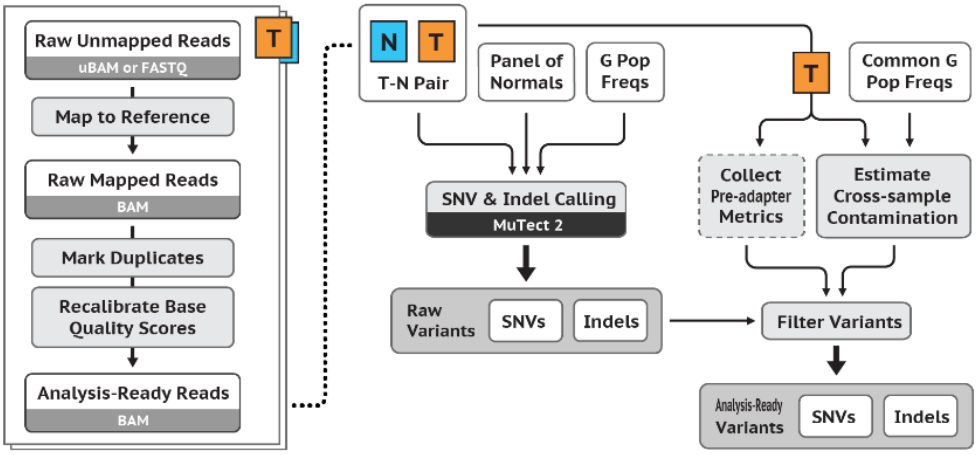

Figure 2.2: Flowchat showing the workflow from the Broad institute used and the main steps in generating filtered somatic variant calls. To the left are the main pre-processing steps that generates analysis ready reads. The Picard suite of tools was utilizes to pre-process the data. The reads were mapped to the reference using the BWA-MEM algorithm and base quality scores were corrected using GATK's BQSR and ApplyBQSR. Mutect2 was the variant caller used and reads were filtered to take tumour contamination into account and improve accuracy using various GATK tools.

Pre-processing with Picard

Picard is a set of command line tools developed by the Broad Institute for the variant calling process. It is used to pre-process data in order to generate analysis

ready reads. The picard.jar file was obtained with wget. As the picard.jar file is a java archive (or jar file), the 'java -jar' command is used to run it. Java is therefore required and loaded with the 'module load' command beforehand.

The main Picard tools used to pre-process files were:

- CreateSequenceDictionary
- AddOrReplaceReadGroups
- CollectAlignmentSummaryMetrics
- SortSam
- MarkDuplicates
- CollectHsMetrics
- BuildBamIndex

CreateSequenceDictionary was used to make the sequence dictionary (hg19.dict) for a reference sequence with the '.dict' extension. It makes it based on the reference sequence provided in FASTA format. The output file contains a header but no SAM records and is required for other Picard tools.

The sequencer will assign identification tags to each set of sequence reads. They exist in the title of each read in the FASTQ file. Picard and other downstream analysis tools require the read groups in a certain format which is often different than the way the sequencer machine assigns them.
AddOrReplaceReadGroups is a tool used to format and include necessary information in a way that is compatible with the suite of tools provided by the Broad Institute. The command 'samtools view -H sample.bam | grep "@RG"' was useful in checking the current state of the read group.

SortSam sorted the reads in the input BAM file by co-ordinate. It can also be sorted by query name or some other property of the SAM record. The SortOrder is found in the SAM file header tag @HD in the field labelled SO. The

input is the mapped BAM file and the output is also a BAM file, that is sorted.

MarkDuplicates identifies and tags duplicate reads in the sorted BAM/SAM file. Duplicate reads are those that originated from a single fragment of DNA. Duplicates can arise during library prep during PCR for example or from a single amplification cluster that are incorrectly detected as multiple clusters by the optical sensor of the sequencing machine. The MarkDuplicates tool compares sequences in the 5 prime positions of both reads and read pairs in the SAM/BAM file. A barcode tag is optional which facilitates duplicate marking using molecular barcodes. When the reads are collected, the tool differentiates the primary and the duplicate read using an algorithm that ranks reads based on their base quality scores.

CollectHsMetrics is used to generate statistics on the quality of the hybrid selection; the lab technique that was used to capture exon-specific sequences for this study. It requires a BAM/SAM file and bait and target files in Picard interval list format. The protocol used was Nimblegen SeqCap EZ Exome v3. The design files were downloaded here https://bit.ly/2LMtnG7. As they were in BED format they were converted to the required interval list format using the Picard tool **BedToIntervalList**.

The output metrics largely fall into three groups. Basic sequencing metrics generated as a baseline against which to evaluate other metrics such as genome size, number of reads etc. Metrics that evaluate the performance of the wet lab assay that generated the data such as number of bases mapping on/off/near bait regions and hs penalty metrics. Metrics that assess target coverage as a proxy for how well the data is likely to perform in downstream analysis such as mean target coverage and the percentage of bases excluded by various filters.

BuildBamIndex generates a '.bai' index file for the current BAM file. This facilitates fast look up of data in the BAM file. The input BAM must have been sorted in co-ordinate order in the SortSam step.

GATK

The Genome Analysis Toolkit (GATK) suite of tools was utilized to generate analysis ready reads and call somatic variants. GATK 4.0.2.1 was the version used. The GATK Resource bundle is a collection of standard files used for working with human sequencing data. It contains files such as the current best set of known indels (Mills_and_1000G_gold_standard.indels.hg19.sites.vcf) and each bundle is specific to the reference genome used. Other files such as the reference sequence (hg19.fa) and the common mutation sites VCF files were obtained from the bundle also. The bundle is available at ftp://ftp.broadinstitute.org/bundle/ The key tools used in the GATK suite were:

- BaseQualityScoreRecalibration
- ApplyBQSR
- CreateSomaticPanelOfNormals
- Mutect2

Base Quality Score Recalibration

BQSR stands for Base Quality Score Recalibration. It is an important data pre-processing step that detects systematic errors made by the sequencer when it estimates the quality score of each base call. Sequencers produce quality scores that are subject to various sources of systematic, non-random technical error. This leads to over or under estimated base quality scores in the data. This estimated quality score is a Phred score and represents an error probability. A Phred score of Q20 represents a 99.9% accuracy. Errors can be due to the physics or chemistry of how the sequencing reaction works and some may be due to manufacturing flaws in the equipment. The criteria for assignment of base quality scores from the sequencer remains concealed by the manufacturing companies.

BQSR applies machine learning to model these errors and adjust error scores accordingly. For example, when two A nucleotides get called, the next base has a 1% higher rate of error. Any base that follows an AA base call will have its Phred score reduced by 1%. This process is done by analysing the covariation among

several features of a base such as reported quality score, the position within a read and the preceding and current nucleotide called. A base may have its Phred score increased for one reason and decreased for another. BQSR does not correct the base call itself but it can tell the variant caller how much it can trust the base that was called. False confidence is often the case on base quality score calls, with the sequencer issuing higher than expected Phred scores. Also, base mismatches tend to occur at the end of the reads more often than at the beginning.

BQSR has two main steps. The first program builds a model of covariation based on the data and a set of known variants. Then it adjusts the base quality scores in the data based on the model. Files with known variants are input to the algorithm. They are dbsnp and Mills indel files downloaded from the GATK resource bundle. The known variants are used to mask out bases at sites of expected variants so that real variants are not counted as errors. All other sites that have mismatches are counted as errors. This process involves two GATK tools; **BaseRecalibrator** and **ApplyBQSR**. BaseRecalibrator takes the known sites of common mutations and outputs a recalibration table which has recalibrated score statistics. ApplyBQSR then uses this table to recalibrate the quality scores in the input BAM file and writes out a new BAM file with recalibrated QUAL field values. This was then repeated for all tumour and recurrence files.

Mutect2

A tumour biopsy may not be a single cell type but a heterogenous mix of cells. Samples might contain sub clones that do not contain variants or a group of cells that are in different stages of differentiation. This introduces the problem of allelic fraction; the fraction of DNA molecules harbouring an alteration. Mutect2 is an algorithm that applies a Bayesian classifier to detect somatic mutations with very low allele fractions requiring only a few supporting reads, followed by carefully tuned filters that ensure high specificity. Mutect2 was released in 2017 and advances upon the original MuTect algorithm. It combines the proven somatic modeling algorithm MuTect with the haplotype-centric logic of the GATK's leading germline variant caller, HaplotypeCaller.

Allelic fraction has been reported as low as 0.05 for highly impure tumours.

Mutect2 considers these cancer subclones and their evolution in exome and genomic data. A panel of normals (PoN) is first built by calling Mutect2 in 'tumor-only' mode on each normal sample. This means 'tumor-only' mode is called three times; for HN51, HN60 and HN72. The output for each is a VCF file. The three VCF files are then combined to make the PoN (threesamplepon.vcf.gz) and its index file with the tool **CreateSomaticPanelOfNormals**. This PoN is a list of the common recurrent technical artifacts and is used to improve the results of the variant calling analysis downstream. It is made from normal samples i.e. the ones that are derived from healthy tissue that should not contain any somatic mutations.

The command 'samtools index sample.bam' and the tool **IndexFeatureFile** was used to index BAM files that were input to Mutect2.

```
java -jar gatk-package-4.0.2.1-local.jar Mutect2 \
-R hg19.fa \
-I HN60.dedup_reads.bqsr.tumour.bam \
-tumor HN60_tumour \
-I HN60.dedup_reads.bqsr.normal.bam \
-normal HN60 \
-pon threesamplepon.vcf.gz \
--germline-resource af-only-gnomad.raw.sites.b37.vcf.gz \
-L SeqCap_EZ_Exome_v3_hg19_primary_targets.bed \
-O HN60_tumour_somatic.vcf.gz
```

Figure 2.3: The command line Mutect2 call for patient HN60 primary tumour. Required and optional arguments are shown. -R is the reference human genome, -I is the input BAM, -tumor is the tumour sample name, -normal is the normal sample name, -pon the panel of normals VCF file, –germline resource is the site specific population germline resource, -L specifies certain regions to analyse, -O is the output VCF file.

Once the PoN was made, Mutect2 was called on the tumour and recurrence data set. Mutect2 then takes that PoN (threesamplepon.vcf.gz) along with two inputs (the tumour BAM and its matched normal BAM), the reference human genome file and a germline resource which contains common population specific mutations. This outputs a VCF file containing unfiltered variants. The code used for a Mutect2 call on the HN60 patient that generated the unfiltered variant set can be seen in figure 2.3. The germline resource was downloaded from gnomad,

the genome aggregation database, also developed by the Broad institute.

Mutect2 operates on each genomic locus independently (exonic regions were specified) and consists of four key steps: (i) removal of low quality sequence data; (ii) variant detection in the tumour sample using a Bayesian classifier; (iii) filtering to remove false positives resulting from correlated sequencing artifacts to remove false positives resulting from correlated sequencing artifacts that are not captured by the error model; (iv) classifying the variants as somatic or germline based on another Bayesian classifier.

Calculating Tumour Contamination

At this point, unfiltered Mutect2 callsets with annotations have been generated in VCF files. Tumours can be contaminated with non tumourous DNA and steps need to be taken to filter for more confident calls. Filtering tools are required to identify which variants can be trusted as real somatic variants. The GATK tools used to filter variant calls according to the best practices work-flow were:

- GetPileupSummaries
- CalculateContamination
- FilterMutectCalls

GetPileupSummaries tabulates pileup metrics for inferring contamination. It makes a six column tabulated file that contains information on each variant. It summarizes counts of reads that support reference, alternate and other alleles for given sites. A germline variant resource (small_exac_common_3_b37.vcf.gz) is used to limit analyses to those sites that commonly have variants and ignores the filter status of the variant calls in this germline resource. As Genome Reference Consortium Human Build 37 (GRCh37) had more than one release, the germline resource had to be lifted over to the reference build. This was done with the tool **LiftoverVcf**.

CalculateContamination uses the results from the GetPileupSummaries table and calculates the fraction of reads coming from cross-sample contamination. This tool estimates contamination based on the signal from reference reads

at homozygous alternate sites. It relaxes the assumptions that its predecessor tool, ContEst, used which used a probabilistic model that assumes a diploid genotype with no copy number variation and independent contaminating reads i.e. drawn randomly and independently from a different human. CalculateContamination uses a simpler estimate of contamination and works in the presence of copy number variations with an arbitrary number of contaminating samples. The idea is to count reference reads at homozygous alternate sites and subtract the number of reference reads expected from sequencing error to obtain the number of reference reads contaminating these homozygous alternate sites. Then the allele frequencies are used to account for the fact that some contaminating reads have the alternate allele. The resulting contamination table contains contamination and error values and is used with FilterMutectCalls.

FilterMutectCalls determines whether a variant call is a confident somatic call. The tool uses various hard filters to eliminate the vast majority of sequencing artifacts. These filters reject sites if some annotation is out of an allowable range. Some are optional and are relevant to the study being performed. Two of them are unadjustable; the panel of normals filter which removes all alleles at blacklisted sites generated in the panel of normals previously and the Short Tandem Repeat (STR) filter which removes variants that are the deletion of a single repeat unit of an STR when this repeat unit contains more than one base.

2.0.5 Statistical Analysis of Variants

Analysis was then done on the resultant VCF files that were output from the somatic calling best practice workflow. Variant calls were input into Ensemble's Variant Effect Predictor (VEP). The three samples, HN51, HN60 and HN72, from both the tumour and tumour recurrence data sets were input individually. Then the tumour data variant calls were combined using bcftools. The same was done with the tumour recurrence variant calls. Then the variants common to both groups i.e. variants that would have survived chemotherapy was obtained by using the 'isec' tool in the bcftools suite. This was then input into VEP and the variants were filtered according to Impact; which was set to HIGH.

Maftools is an R package used to visualise, summarize and analyse MAF

files. MAF stands for mutation annotation format and the maftools package is commonly used to analyse somatic variant calls in cancer genomics. Once the VCF files were converted to MAF format, the analysis was completed on RStudio and the R Script can be found in the GitHub Repo.

3 Results

3.0.1 QC Analysis before and after trimming/adapter removal

MultiQC reports were generated before and after trimming of reads. They provided opportunities to assess each sample and identify any issues that might effect downstream analysis. The summary is reported via traffic light colours where green indicates good, amber warning and red fail. MultiQC was performed on three groups of samples: normal, tumour and recurrence. Prior to trimming, it was clear that the HN72 samples in all three groups were of a lower quality as can be seen in the general statistics and sequence quality histogram section of the reports in figure 3.1. Other factors such as the reported Phred score and the per sequence content aided the trimming process that followed. Theoretically, trimming/removing adaptors from sequence reads based on a set of predefined parameters should improve the quality of the reads. This is because the quality of the sequence read reduces towards the end naturally. Reads contaminated with adaptors will not map but may affect some mappers speeds and/or quality. All MultiQC HTML reports can be seen at the provided GitHub repo.

For the normal samples, the main differences seen post trimming was lower/equal GC content in all samples, less duplicate reads in all samples and less number of sequences, all as expected. Samples HN72 were of lower quality in all groups pre trimming. One sample, HN72AC failed prior to trimming and passed after trimming as can be seen from the red line on first plot in figure 3.1. All samples were above median Phred score for quality. No samples were found with adapter contamination after trimming. Per base N content was higher in all samples before trimming, 7 samples passed and 1 sample passed with warnings

before trimming while all samples passed with no warnings after trimming. No differences were seen in terms of Per Sequence Quality Scores, Overrepresented sequences, sequence Duplication levels and Per Base Sequence Content before and after trimming.

For the tumour biopsy samples, all reads had a higher Phred score after trimming. The HN72 samples were again problematic with a failed result before trimming. Per base sequence content QC did not change before and after trimming with the same 2 samples failing, 4 passing with warnings and 2 samples passing. The HN60 forward and reverse files failed before and after trimming for sequence duplication levels. The HN60 forward read passed with warnings for the overrepresented sequences test. Overall, the HN72 samples showed the poorest results and the HN60 files showed poorer than their equivalent normal for this set of tumour biopsy samples.

For the recurrence tumour biopsy samples, the HN72AC sample failed by dropping below the Phred score quality before trimming but not after. Trimming made no difference to the Per Base Sequence Content test as expected with HN51 and HN60 failing. Per Base N Content failed HN72AC before trimming but passed it after trimming. HN51 and HN60 failed the Sequence Duplication Levels test before trimming but not after. Several metrics gave a failed result for the three groups of samples. However, trimming/adapter removal is an important step in somatic variant calling and improved the reads adequately for the analysis to proceed.

3.0.2 Alignment to the Human genome

The Picard tool **CollectAlignmentSummaryMetrics** was used to generate alignment statistics for each sample. It takes a BAM or SAM file and outputs a text file with high level metrics about the alignment of reads as well as the proportion of the reads that passed machine signal-to-noise threshold quality filters. These filters are specific to Illumina data. Table 3.1 shows the number of PF Reads, that is the number of reads that passed Illumina's platform filter (PF) for quality. To the right of the table is the percentage of those reads that were successfully aligned to the reference sequence. As can be seen, 99% of

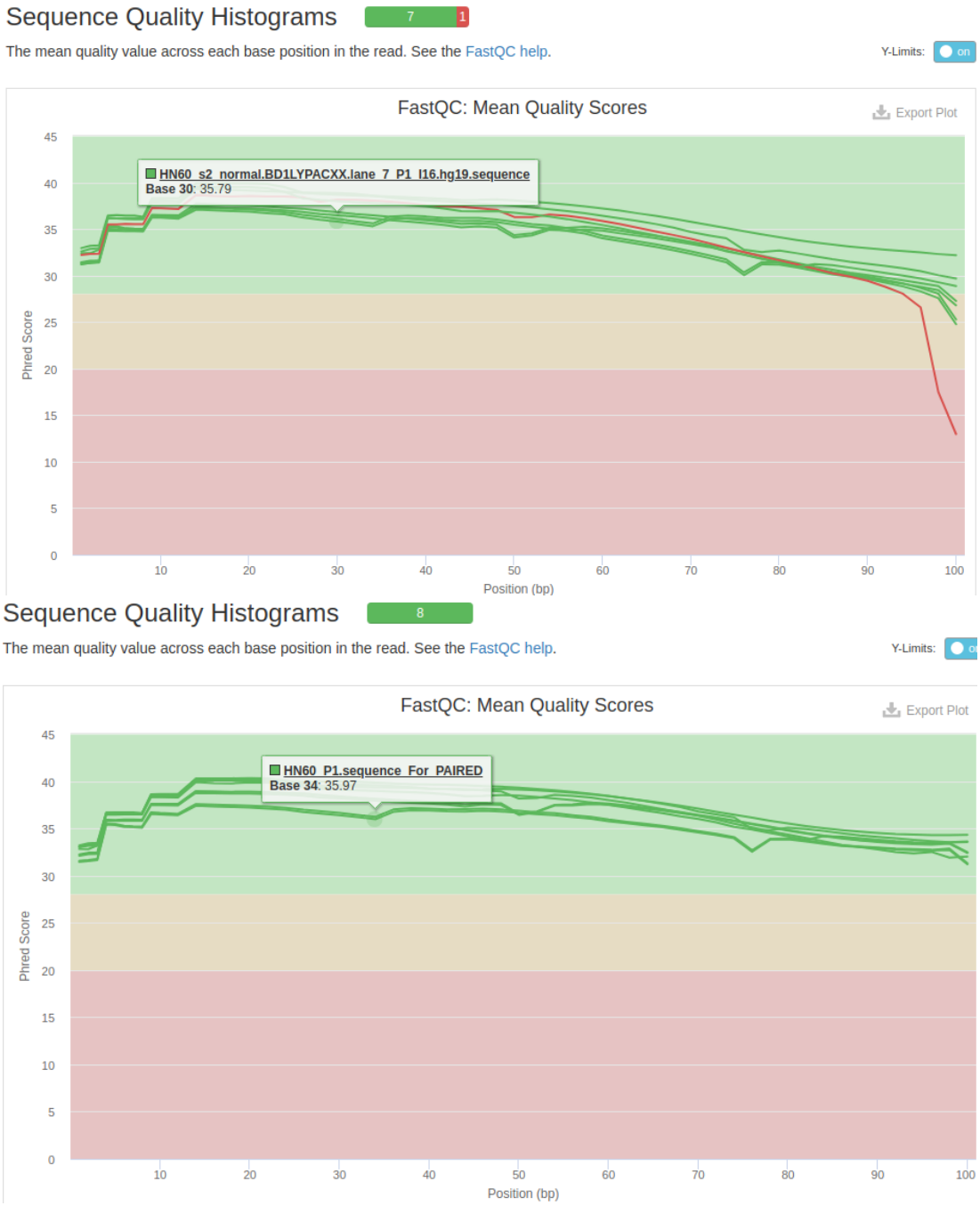

Figure 3.1: Top: Sequence Quality Histogram pre trimming. Bottom: Sequence Quality Histogram post trimming. The y-axis shows the Phred scores while the x-axis shows the position of the read in base pairs. The red line is showing a failed sample. After the sample was trimmed the sample passed.

Sample	PF Reads	% of PF Reads Aligned
HN51 Tumour	251944834	0.999823
HN60 Tumour	249117458	0.999828
HN72AC Tumour	227447536	0.994754
HN72AH Tumour	8517624	0.931442
HN51 Recurrence	237673672	0.999839
HN60 Recurrence	216782354	0.999831
HN72AC Recurrence	153610920	0.992234
HN72AH Recurrence	8524554	0.911665

Table 3.1: Table showing number reads that passed Illumina's platform filter. Table also shows the percentage of those reads that were aligned.

reads were aligned to the human genome for most samples. HN72AH Tumour and Recurrence were the other portion of the HN72AC samples. They were sequenced on a different flow cell/lane and contained less reads, showed a lower Phred quality score in the MULTIQC reports and successfully mapped less reads (91%) than all other samples.

3.0.3 Picard metrics

The tool **CollectHsMetrics** was used to gather information on the quality of the hybrid-selection (HS); the lab technique used to capture exon-specific data. Table 3.2 shows some HS metrics. The number of on target bases is shown on the left for each sample. On the right is the percentage of bases that were mapped away from any baited exonic region. As can be seen, a small percentage of bases were off target overall. This means that most or 98% of bases were actually from exonic regions. Again, the HN72 samples were flagged with lower quality in earlier analysis and the HS metrics could explain why. Less bases were mapped to exonic regions for those samples. The percentage of bases off bait was 78% and 79%, for samples HN72 Tumour and HN72 Recurrence respectively. All other samples, however, showed all metrics that were confident of accurate downstream analysis. Full HS metrics text files can be seen in the GitHub repo.

Sample	On Target Bases	% of bases Off Bait
HN51 Tumour	9195406720	0.117766
HN60 Tumour	9122609166	0.113001
HN72 Tumour	2128559802	0.778129
HN51 Recurrence	8583651820	0.125968
HN60 Recurrence	8012452417	0.127963
HN72 Recurrence	1350747256	0.791361

Table 3.2: Table showing number of bases on target i.e. aligned to the exon regions of the human genome and the percentage of bases that are off bait i.e. not on target.

Sample	No. of Variant Calls
HN51 Tumour	10741
HN60 Tumour	11035
HN72 Tumour	28677
HN51 Recurrence	10678
HN60 Recurrence	10627
HN72 Recurrence	2830

Table 3.3: Number of variant calls produced in each VCF file in primary tumour and tumour recurrence data sets.

3.0.4 Somatic Variants called with Mutect2

Mutect2 provides its own filters along with the parameters used such as the PoN. Then the three steps outlined in the pipeline; GetPileupSummaries, Calculate-Contamination and FilterMutectCalls were used to further filter calls to take tumour contamination into account. The end of the best practices pipeline produced a number of variants that can be seen in Table 3.3. The number of variants produced was less in each recurrence set. For example with HN51, 10741 variants were called in the primary tumour set and 10678 variants were called in the tumour recurrence data set.

3.0.5 Tumour VS Normal

bcftools was used to generate a VCF file with the variants that are common to all normal germline samples. The same was done for the primary tumour cohort. Then, bcftools found the variants that are common to the tumour and normal set (commonTumourandNormal.vcf). These common variants were put into VEP and filtered for 'HIGH' impact. The complete VEP output can be viewed at https://bit.ly/2vLQNRH and on GitHub.

620 variants were processed and when the 'HIGH' impact filter was applied, 12 remained. These included the genes *KANSL3*, *ZDHHC11*, *MGAM2* and *MUC6* and the variant types comprised of stop lost, frameshift variants and Nonsense mediated decay (NMD) Transcript variants. A summary of these variants including genomic location and exon number can be seen in figure 3.2.

Location	Allele	Consequence	Impact	Symbol	Gene	Feature	Biotype	Exon
2:96610810-96610811	-	frameshift_variant, NMD_transcript_variant	HIGH	KANSL3	ENSG00000114982	ENST00000420155	nonsense_mediated_decay	11/21
2:96610810-96610811	-	frameshift_variant	HIGH	KANSL3	ENSG00000114982	ENST00000431828	protein_coding	11/21
5:796047-796049	-	frameshift_variant	HIGH	ZDHHC11	ENSG00000188818	ENST00000424784	protein_coding	7/11
7:142143879-142143879	T	stop_lost	HIGH	MGAM2	ENSG00000257743	ENST00000477922	protein_coding	13/48
7:142143879-142143879	T	stop_lost	HIGH	MGAM2	ENSG00000257743	ENST00000550469	protein_coding	13/13
11:1017040-1017042	AGT	frameshift_variant	HIGH	MUC6	ENSG00000184956	ENST00000421673	protein_coding	31/33
11:1017040-1017042	-	frameshift_variant	HIGH	MUC6	ENSG00000184956	ENST00000421673	protein_coding	31/33

Figure 3.2: Chart showing variants that are common to the primary tumour and normal cohort. Variants shown are filtered for 'HIGH' impact, as shown on chart.

3.0.6 Recurrence VS Normal

The same process was completed for the tumour recurrence and normal cohorts. The bcftools script used to generate these VCF files is located on GitHub. The common variants to the recurrent tumour and normal samples is called normalandRecurrencevariants.vcf. Again, VEP was used to process 174 variants which can be found at https://bit.ly/2vqikJj. 12 variants were present as HIGH impact.

Figure 3.3 shows two pie charts depicting consequences of common variants for tumour VS normal (top) and tumour VS recurrence (bottom). The tumour VS normal group had a higher percentage of intron variants, non-coding transcript variants, NMD transcript variants and upstream gene variants. Both the tumour

VS normal and tumour recurrence VS normal common variant lists were then compared to the tumour VS recurrence variants.

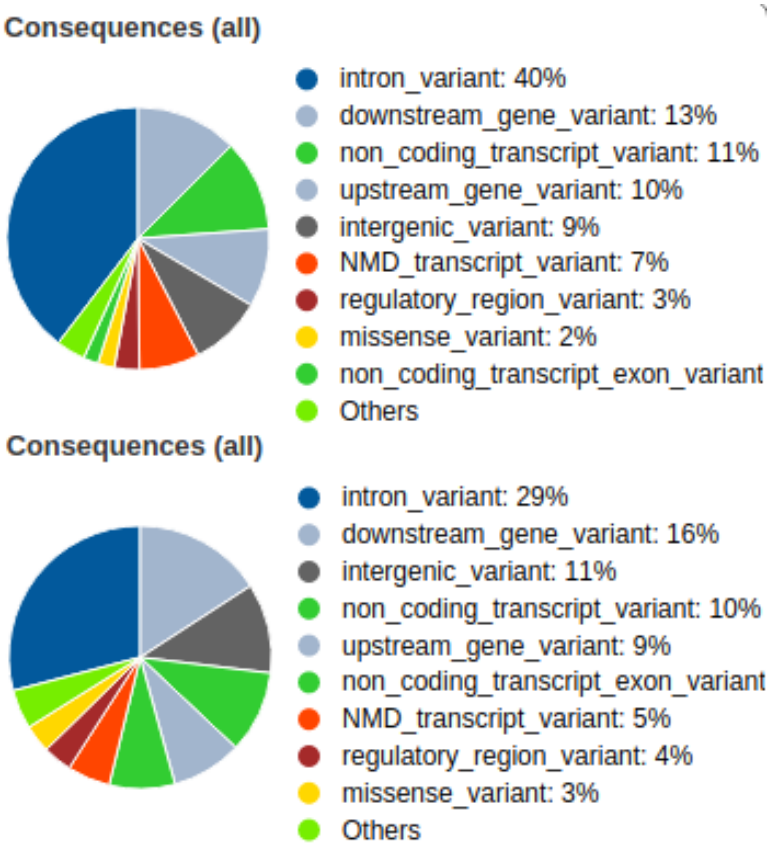

Figure 3.3: comparing tumour and normal and tumour and recurrence

3.0.7 Tumour VS Recurrence and the Genetic Variants associated with chemo-resistance/recurrence in HNSCC.

bcftools was used to generate a VCF file that contained variants that were common to all samples in the primary tumour group and variants that were common to all samples in the tumour recurrence group. That way, primary tumour and recurrent tumour variants could be compared. They were then put through VEP for analysis. The primary tumour group contained 1388 variants while the recurrence tumour group had only 465 variants. The primary tumour variants

contained 375 variants found on overlapping genes while the tumour recurrence equivalent was 961. 1302 variants were found on overlapping transcripts in the primary tumour set and the tumour recurrence set had 3436. Coding consequences for both groups of variants can be compared in figure 3.4. Both contain the same portion of missense, synonymous and stop-gained variants with similar portions of other variant types being shared. VCF files containing common tumour variants (commontumourvariants.vcf) and common recurrence tumour variants (commonRecurrencevariants.vcf) can be found on GitHub.

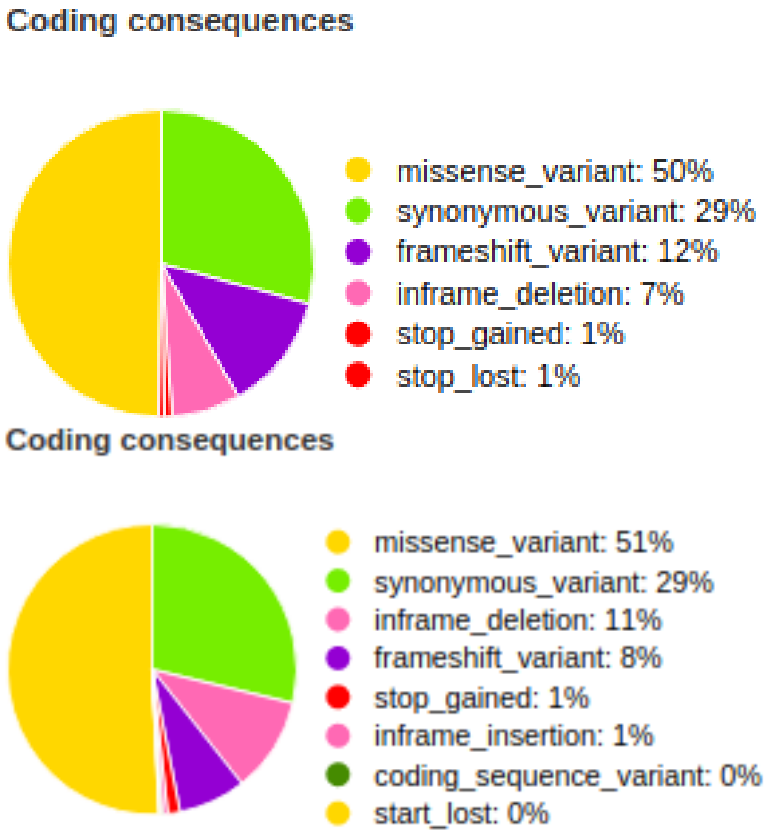

Figure 3.4: Top: coding consequences for the primary tumour variants. Bottom: coding consequences for the recurrence tumour variants.

Then, bcftools was again used to generate a VCF file that contained variants that were common to both the primary tumour and tumour recurrence data set. This VCF file (tumourANDrecurrencevariants.vcf) can be found on GitHub and contain somatic variants that existed in the primary tumour biopsy

samples as well as the tumour recurrence biopsy samples. The fact that the variants existed in both cohorts must mean that they were exposed to chemotherapy and were unaffected. The VEP output can be found on GitHub and here https://bit.ly/2OfXL9p.

Initial inspection yields 154 variants that were processed and 0 variants were filtered out. 142 variants are on overlapping genes and 426 are on overlapping transcripts. These variants are ones that answers this projects question i.e. the variants that are associated with chemo-resistance/reccurence in head and neck squamous cell carcinoma. As expected, no variants were found on the X or Y chromosomes. Figure 3.5 shows a summary of the coding consequences of these 154 variants. 50% of the group contain missense mutations, while the remainder is comprised of synonymous variants (37%), frameshift variants (12%) and inframe deletions (1%).

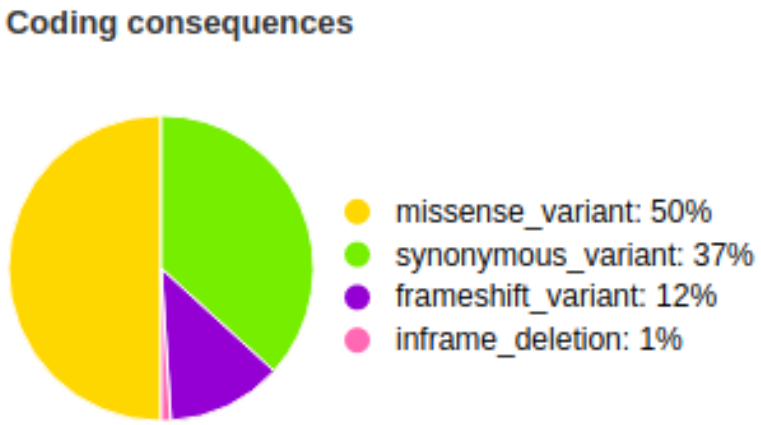

Figure 3.5: Pie chart showing coding consequences of the 154 variants that are present in tumour and recurrence cohorts.

Of the coding portion of the set, 50% of variants were missense mutations i.e. point mutations that code for a different amino acid. 37% are synonymous variants (a substitution of one base for another in an exon region). 12% are frameshift variants (an indel that is not divisible by 3 that will alter the amino acid) and only 1% of them are inframe deletions (deletions in a multiple of 3 meaning only some amino acids are changed and the protein may still be able to function.) When the results are filtered for 'HIGH' impact, 12 variants remain. Full results of these high impact cohort can be seen in the file

(VEP_IMPACT_is_HIGH_VARIANTS.csv) and a synopsis of such file can be seen in figure 3.6. The highest impact variant is located on the *OR13C5* gene. The next seven variants are all located on the *CTAGE5*. The next highest impact variant is located on *RP11-407N17.3*, a CTAGE family member. The final three variants of the high impact 12 are also located on the *RP11-407N17.3* and *CTAGE5* genes. To make sure, the tumour VS normal and recurrence VS normal common variants were compared to this set of 12 variants. None of them were found in both the tumour VS normal and tumour VS recurrence set or the recurrence VS normal and tumour VS recurrence set. This means that the 12 high impact variants found must in fact be harmful as they do not contain any normal germline samples.

Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Biotype	Exon
.	9:107361178-107361182	-	frameshift_variant	HIGH	OR13C5	ENSG00000255800	Transcript	ENST00000374779	protein_coding	1/1
.	14:39746242-39746242	T	frameshift_variant	HIGH	CTAGE5	ENSG00000150527	Transcript	ENST00000280083	protein_coding	2/24
.	14:39746242-39746242	T	frameshift_variant	HIGH	CTAGE5	ENSG00000150527	Transcript	ENST00000341502	protein_coding	2/24
.	14:39746242-39746242	T	frameshift_variant	HIGH	CTAGE5	ENSG00000150527	Transcript	ENST00000341749	protein_coding	2/24
.	14:39746242-39746242	T	frameshift_variant	HIGH	CTAGE5	ENSG00000150527	Transcript	ENST00000348007	protein_coding	2/23
.	14:39746242-39746242	T	frameshift_variant	HIGH	CTAGE5	ENSG00000150527	Transcript	ENST00000396158	protein_coding	2/24
.	14:39746242-39746242	T	frameshift_variant	HIGH	CTAGE5	ENSG00000150527	Transcript	ENST00000396165	protein_coding	2/24
.	14:39746242-39746242	T	frameshift_variant	HIGH	CTAGE5	ENSG00000150527	Transcript	ENST00000553352	protein_coding	1/23
.	14:39746242-39746242	T	frameshift_variant	HIGH	RP11-407N17.3	ENSG00000258941	Transcript	ENST00000553728	protein_coding	6/28
.	14:39746242-39746242	T	frameshift_variant	HIGH	CTAGE5	ENSG00000150527	Transcript	ENST00000554392	protein_coding	2/7
.	14:39746242-39746242	T	frameshift_variant	HIGH	CTAGE5	ENSG00000150527	Transcript	ENST00000555716	protein_coding	2/6
.	14:39746242-39746242	T	frameshift_variant	HIGH	RP11-407N17.3	ENSG00000258941	Transcript	ENST00000603904	protein_coding	2/24

Figure 3.6: VEP output for variants that are filtered for 'HIGH' impact. Gene names and genomic location can be seen also.

3.0.8 Analysis and visualisation of results with Maftools

The maftools package is designed to take larger cohorts; in fact it is a requirement to provide more than 2 samples. All samples that could contain somatic variants (tumour and tumour recurrence twice filtered VCF files) were merged using the cat command on the terminal. They were then converted to MAF format and

analysed. Summary statistics of MAF were generated with the mafSummary command and can be seen in figure 3.7. Missense mutations dominate the cohort. In frame deletions and frame shift deletions were the next most commonly found variant class. SNP's were the most commonly found type, as expected. This summary call also generated a SNV class plot which can be seen in figure 3.8. C>T transitions were the most popular SNV class found. This was followed up by T>C transitions.

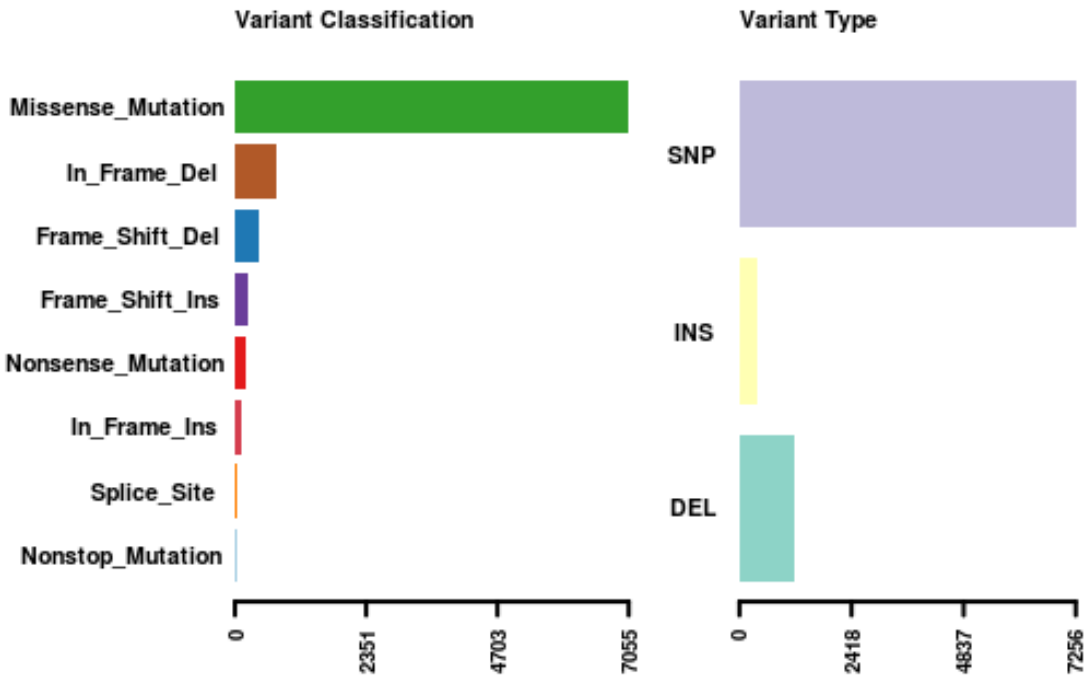

Figure 3.7: A summary of the variant classification and mutation type produced by maftools for all somatic variants found.

Amino acid changes

Understanding the effects mutations have on protein structure is an important part of evaluating the effects of variants. The function call lollipopPlot was used to show mutation spots on protein structure for a specified gene. A ggplot was returned for the high impact *OR13C5* and is shown in figure 3.9. Mutations such as F32L and L69M are labelled on each point along the protein structure. The number of mutations is seen on the y-axis. Below the x-axis gives the legend

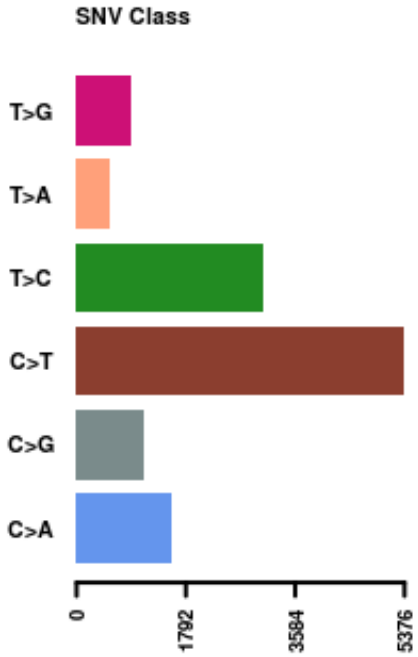

Figure 3.8: Plot showing most common SNV classes for all variants.

corresponding to the mutational type. Blue corresponds to a Frame shift deletion and Red corresponds to a missense mutation.

Hyper mutated regions on the cancer genome

Cancer genomes, especially tumourous ones, can be shown to have loci with localized hyper mutations (ref). These mutated regions can be visualised by plotting intervariant distance on a linear genomic scale. A function call, rainfallPlot, was used to highlight regions where potential changes in inter-event distances are located and the plot can be seen in figure 3.10. Highlighted regions include chromosomes 1-3, 6-8 and 13-19. No variants occurred on the sex chromosomes as expected.

Detecting cancer driver genes

The function call oncodrive identifies cancer driver genes in a given MAF file. Most variants in cancer genes are enriched at a few specific loci. This algorithm uses these positions to identify cancer driver genes. The plot produced can be

lollipop.png lollipop.png

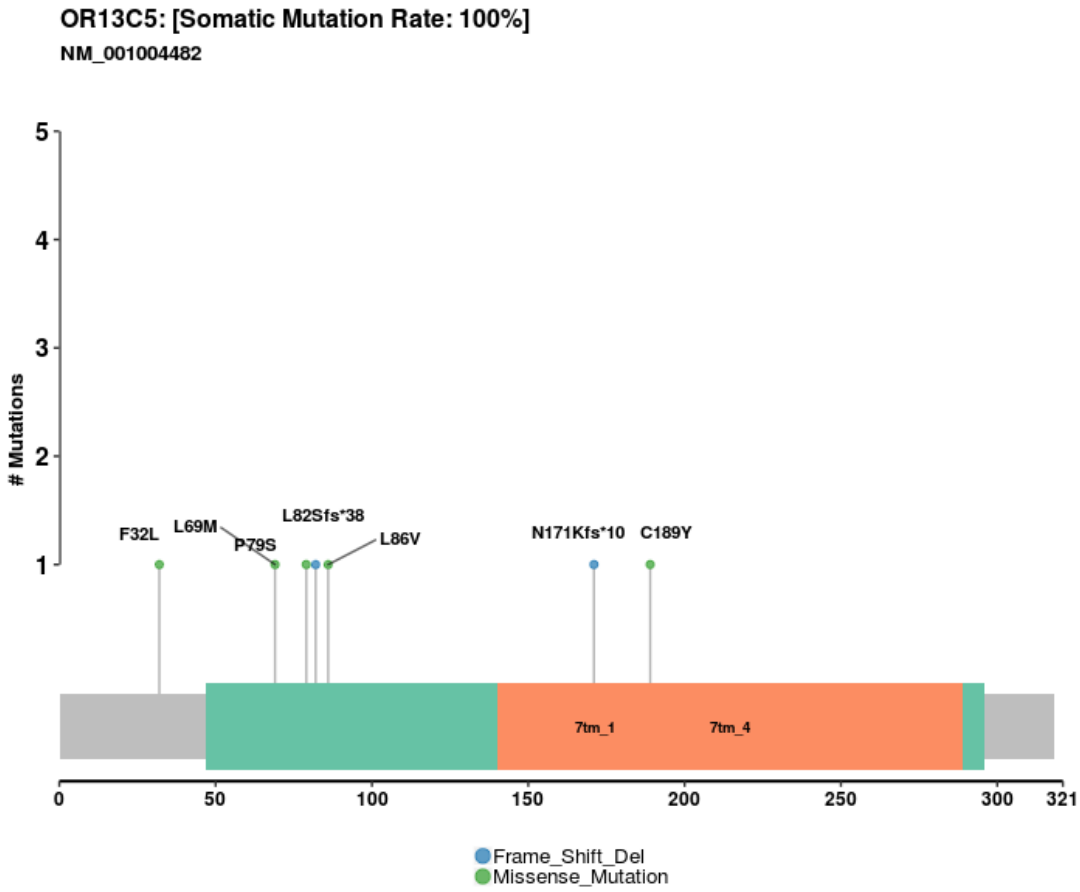

Figure 3.9: Amino acid changes for the *OR13C5* gene.

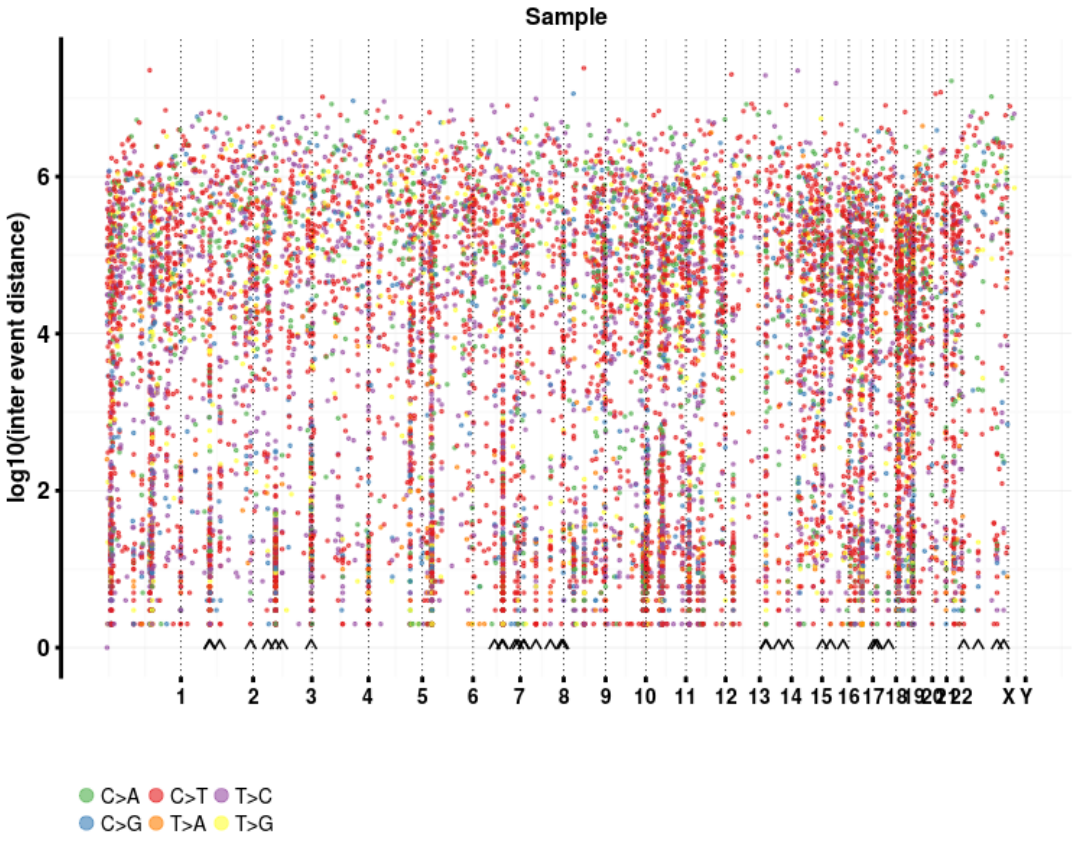

Figure 3.10: Plot showing hyper mutated regions in the HNSCC cancer genome from this study. As detectChangePoints was set to TRUE, the plot highlights changes with arrows along the x axis. The colour of the points corresponds to the SNV class and the legend is located under the x-axis.

seen in figure 3.11. *ZNF880* and *GBP4* were two driver genes estimated by this method. The size of the points on the plot is proportional to the number of clusters found in the genes. The two driver gene estimates contain more mutations per cluster than other genes.

To visualise the effect these genes may have on the protein structure, the function call lollipopPlot was again used on these flagged driver genes; *ZNF880* and *GBP4*. The plots can be seen in figure 3.12. In this plot, frame shift deletions are plotted in blue, frame shift insertions plotted in purple and missense mutations plotted in green. Cluster scores were estimated from non-synonymous variants

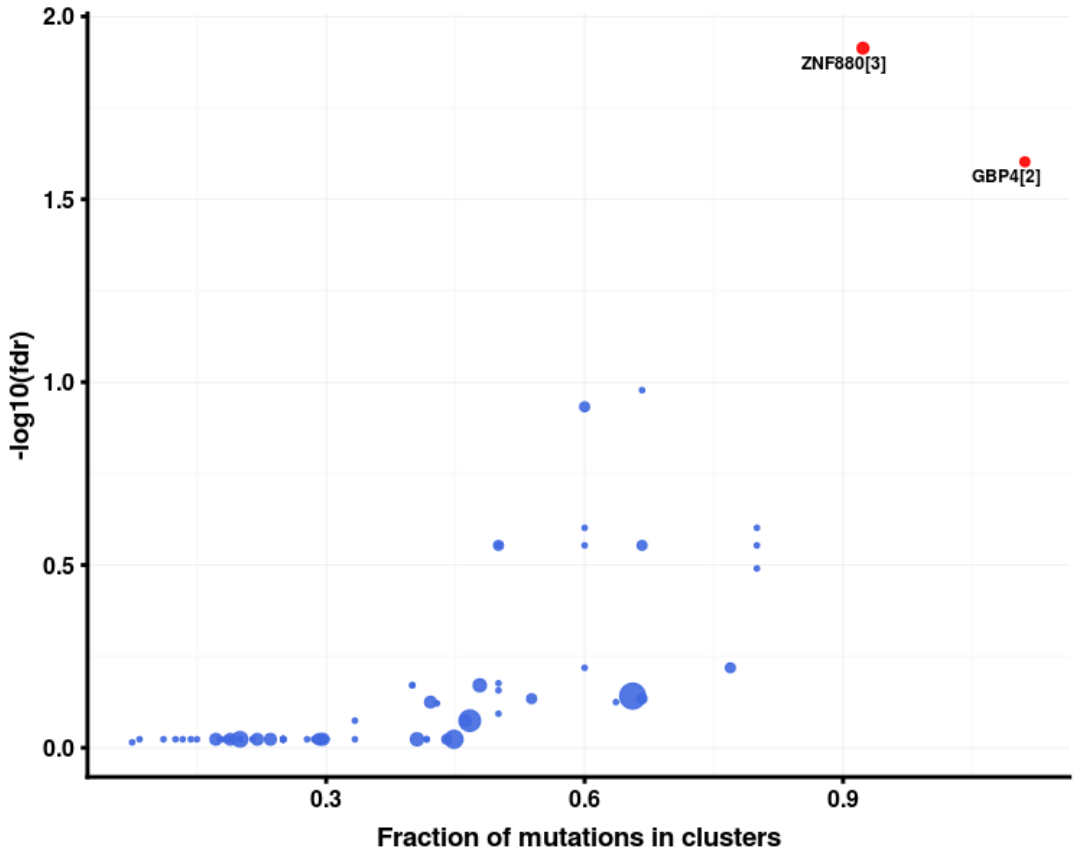

Figure 3.11: Scatter plot produced showing mutational clusters. Driver genes are estimated as *ZNF880* and *GBP4*. The size of the points are proportional to the number of clusters found in the genes.

lollipop.png lollipop.png

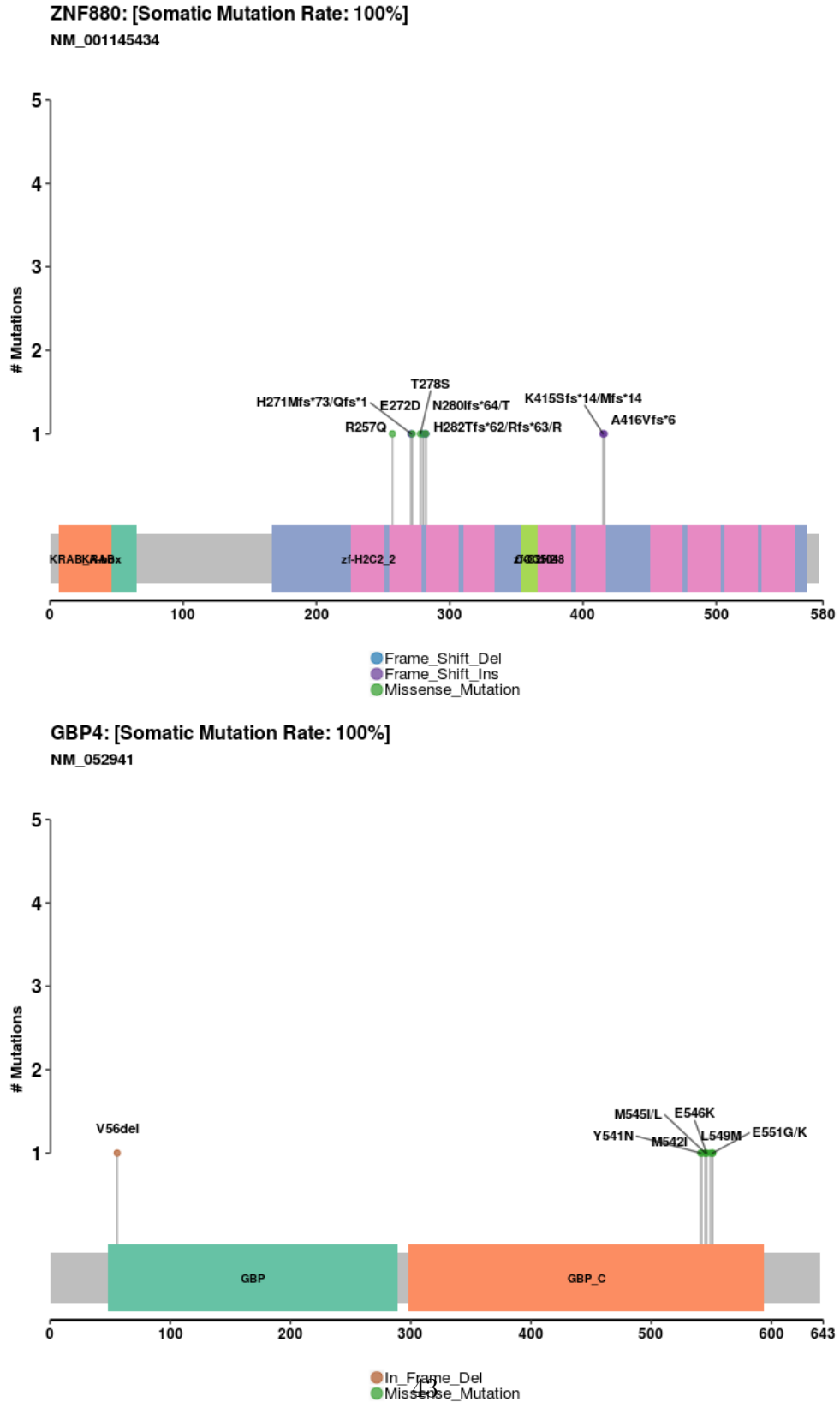

Figure 3.12: Lollipop plot shown for the genes *ZNF880* (top) and *GBP4* (bottom).

Analysing tumour heterogeneity and inferring contamination

Tumours are generally heterogeneous in nature (ref) i.e. they will consist of multiple clones. This can be estimated by clustering VAF scores. VAF can be though of as the number of reads that contains the variant allele. The function call inferHeterogenity clusters variants using VAF information which infers clonality and generates the plot in figure 3.13. Overall, most VAF scores are less than 0.5. The majority cluster between 0 and 0.25. Not all clusters are in this range. Cluster 1 to 4 are under 0.25.

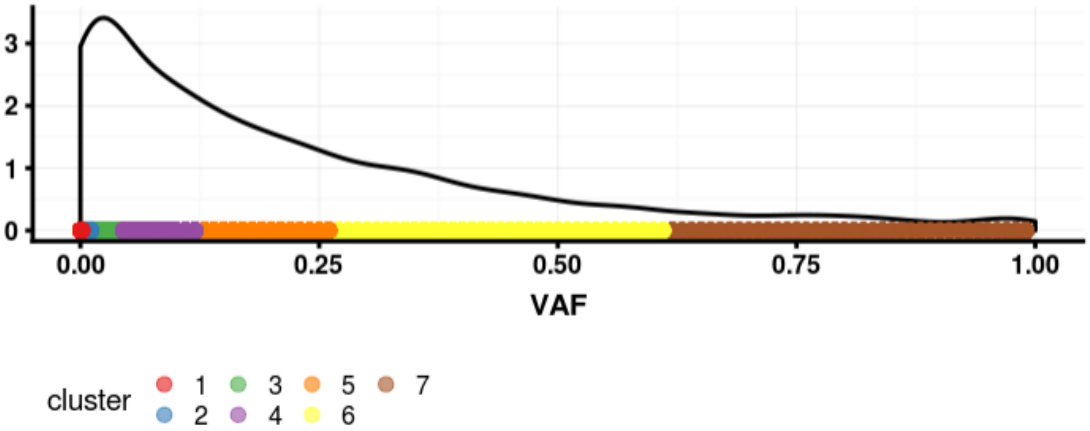

Figure 3.13: Plot showing VAF scores and clustering of variants. Most VAF scores recorded are in the 0 - 0.25 range where there are at least 5 clusters.

Compare mutational load to TCGA cohorts

It is informative to see how mutation load compares against TCGA cohorts. The function tcgaCompare was used which draws distribution of variants compiled from over 10,000 WES studies across 33 TCGA landmark cohorts. The plot generated can be seen in figure 3.14. This study is entitled 'Project' along the x-axis. The mutational load seen is much higher than other cancer genomic studies. This could be due to the small sample size and the fact that the projects cohort contains primary and relapse tumours which are known to have heavy mutational loads. Also, the other cohorts represent patients that are at different stages of cancer and mostly contain primary tumours.

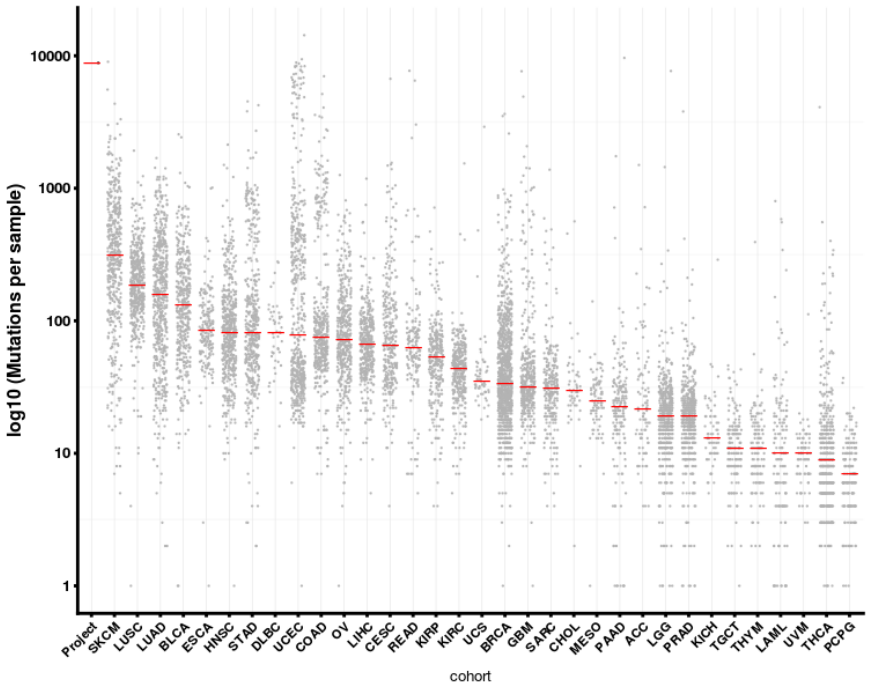

Figure 3.14: The mutational load found in this project, seen on the left of the x-axis compared to over 10,000 WES samples from TCGA. The mutational load seen in this project is much higher than other cohorts.

4 Discussion

Somatic variant callers are usually challenged by the balance between detecting true low-allelic somatic variants and maintaining accuracy in the mutation calling procedure in order to keep the false positive rate low. Mutect2 provides a compromise that has proven to be accurate. Of course, the nature of the cancer samples can influence the results. Overall, the samples in this project presented with high quality. The mean coverage (80x) of the samples is adequate for Mutect2 to handle as it is known to have shown superior sensitivity and accuracy at 50x coverage.

The first step was to trim the sequence reads. This ensured that they remained above a Phred quality score of 33. Other metrics examined by the fastqc package such as Per Sequence GC Content and Per base N Content all passed, at least after the trimming of reads. However, before trimming, the HN72 primary and relapse tumour samples initially failed the quality score reported by MULTIQC. The same samples also showed poor HS metrics with only 22% of the bases called being actually from exonic regions. This metric could not be corrected for and the analysis proceeded with sequence alignment. BWA-MEM, the gold standard mapper, proved its worth with a typical 99% score for the amount of reads successfully mapped as reported by the resourceful Picard tool; CollectAlignmentSummaryMetrics. All other quality checks investigated including MarkDuplicates metrics, BQSR and bcftools statistics were as expected and provided welcomed confidence moving through the workflow. Thus it can be confirmed that the workflow was followed according to the industry and best practices standard and the analysis ready reads and variants are as accurate as possible, given the cancer cohort.

The molecular genetics of cancer is rapidly evolving and it is important to recognize that levels of evidence in prognosis, disease and therapy is continually changing. Maintaining confidence in the consequences of somatic variants is done by modifying impact status regularly. To understand the cohort and the significance of the mutations found in this project, a collection of databases and publicly available resources of human genetic variation was used including dbSNP, Omim, COSMIC, ClinVar, VEP, PolyPhen2 and more. A four tiered system to categorize somatic cancer variants was followed, convened by the Association for Molecular Pathology with liaison representation from the American College of Medical Genetics and Genomics[69]. The four tiers were (i) variants with strong clinical significance, (ii) variants with potential clinical significance, (iii) variants of unknown clinical significance and (iv) variants deemed benign or likely benign. This model proved useful as guidance when sifting through the large amounts of variants and previous clinical links made to a variant and/or gene.

A suitable comparison of all variants called is required in order to accurately conclude the genetic variants that are associated with chemo-resistance/recurrence in HNSCC. Firstly, the variants common to all normal germline samples were compared to the same from the tumour variants. This led to a list of variants that were found in a tumour and a normal sample. The variants common all came from protein coding genes. Of them, included the *MUC6* gene. A member of the mucin protein family, *MUC6* has been shown to be strongly correlated with other carcinomas such as gastric adenocarcinoma [70] and duodenal adenocarcinoma[71]. In fact, variants were observed frequently on the mucin family genes in the normal germline samples.

The recurrence vs normal cohort produced a smaller list of variants. This was expected as germline variants are not considered to hold the same level of chemo-resistance as somatic variants do. 12 variants were high impact. *MUC6* was again frequently mutated as well as the *NBPF10* gene; a gene correlated with lung adenocarcinoma and lung squamous cell carcinoma [72] which is closely related to relapse HNSCC. The impact status was 'moderate' and 'low' for most variants in these two cohorts. For all missense variants, the largest abberation found, were shown to have at least a 'moderate' impact status and were found mainly on the genes *CDC27*, *OR2L3* and various *MUC* family members across

pathways including the Innate Immune System super pathway and the KEGG pathway. The Tumour VS Normal and Recurrence VS Normal analysis provided extra assurance when assessing the Tumour VS Recurrence variants.

154 variants were found that were listed in the primary tumour and recurrence tumour groups. These variants were exposed to chemoradiation and were unaffected. The list comprised of mostly missense variants which is likewise to other somatic variant calling studies on cancer[73]. 12 high impact variants were found and were all frameshift variants. The *CTAGE5* gene was one that featured in the high impact list more than once. Interestingly, it is one that has been associated with squamous cell carcinoma of the lung[74], a cancer that is closely related to HNSCC and is a common metastatic cancer for HNSCC patients. Mutations on this gene have been shown to be enhancers of other cancers such as Cutaneous T-Cell Lymphoma and colon carcinoma[75]. Cutaneous T-Cell Lymphoma symptoms include enlarged lymph nodes and as outlined in the introduction, HNSCC tumours tend to develop near lymph nodes and the presence and number of them are important prognostic markers of distant disease and survival [3]. Most variants found on *CTAGE5* are frameshift variants which cause disruptions of the translational reading frame as the insertion or deletion is not a multiple of three. Notably, other high impact variants lie on the *RP11-407N17.3* gene, a *CTAGE* family member and linked to Cutaneous T-Cell Lymphoma. Given the high likelihood of HNSCC relapsing post treatment and the existing link with *CTAGE5* and lung squamous cell carcinoma, a potential treatment target would involve this gene family, especially in patients that are late onset.

The *OR13C5* gene was another mutated gene on the chemo-resistant variant list. The *OR13C5* gene is an olfactory receptor gene that interacts with odorant molecules in the nose, to initiate a neuronal response that triggers the perception of a smell. It is already strongly correlated with respiratory tract diseases and the SNP rs199665292 on *OR13C5* is the strongest variant for Respiratory syncytial virus (RSV) in infants [76]. Noteably, *MUC4* and *MUC6* were two other mutated genes associated with respiratory syncytial virus in the same study. Some studies suggest olfactory receptors are drivers of cancer[77]. What can be said is that olfactory receptor genes are upstream effectors of the MAPK signalling cascade and are potential novel targets for pain in HNSCC patients[78].

NBPF10, an alias of the *NOTCH2A* gene, was found to harbour missense mutations. These mutations are also apparently chemo-resistant and can be seen when the filter is set to moderate. As explained in chapter 1, the NOTCH receptors bind to membrane bound ligands on other cells and act as a transcription factors and in HNSCC is classified as a tumour supressor gene[64]. Links were found between *NOTCH* genes and the Wnt signalling pathway. Inactivating mutations were also found in *NOTCH1*. Carcinogenesis with these patients may have developed through inactivation of a *NOTCH* gene. However the link is not clear and a larger cohort may have provided more evidence. The *OR2L3* gene, another olfactory receptor gene contained missense variants. When no filters were applied, genes such as*MUC*, *MUC4* and *MUC3A* were frequently found with mutations. The mucin family of genes are glycosylated proteins that play an important role in forming protective mucous barriers on epithelial surfaces. At the most basic level, HNSCC is an epithelial malignancy and these abberations are somewhat expected. Also, the *FAM41C* gene contained mutations that were previously linked to recurrence in thyroid cancer [79] and found in this study.

The *CDC27* gene was found to be one of the most mutated genes in the list of the chemo-resistant cohort. Its impact status is classed as a modifier and it was also found in the tumour VS normal and recurrence VS normal lists. *CDC27* is directly linked to esophageal squamous cell carcinoma [80] and metastasis in colorectal cancer [81]. More relevantly, *CDC27* was shown to interact with mitotic checkpoint proteins including Mad2, p55CDC and BUBR1, which means it might be involved in controlling the timing of mitosis and is a key regulator of the exit from mitosis [82]. *CDK* family members such as *CDK2AP2P2* appear mutated regularly and are involved in cell cycle control. Efficient cell proliferation, especially during S-phase, was shown to be a contributing factor to HNSCC developement as explained in chapter 1.

Maftools provided an excellent analysis package to summarise the cohort of somatic variants called. Figure 3.8 shows the most common SNV classes. This is not surprising as C>T transitions have been known to be common mutational types in cancer [83]. A genomic analyses of oesophageal squamous-cell carcinoma identified the most common mutations to be C>T and C>G transitions at TpCpW sites.

Detecting cancer driver genes is a process that varies considerably and was reviewed in Chapter 1. Nonetheless, the function oncodrive was used to detect potential oncogenic drivers in the cohort. In figure 3.11, a plot was generated which showed mutational clusters. The genes *ZNF880* and *GBP4* were flagged as driver genes. Traditional methods of driver gene detection identify genes that are more frequently mutated than the background rate. Maftools, however, detects genes that are significantly clustered in specific regions of the amino acid sequence. Figure 3.12 shows the mutational clusters on the amino acid of the detected driver genes. Clusters of mutations are evident for example with the T278S and E272D mutations on the ZNF880 amino acid sequence. A cluster also seems apparent on the *GBP4* gene with mutations E546K and Y541N. However, the two flagged driver genes do not appear often on the final VEP reports. They would be expected to present frequent mutations in some variant cohorts. The reason for this could be the sample size. It is clear that there are in fact mutations due to these two genes from figure 3.12. A cohort of more patients or a confirmatory analysis may provide a better insight into the driver genes in HNSCC.

The hyper mutated regions are presented by the rainfall plot in figure 3.10. They depicted the most mutated regions to be at chromosome 1 to 3, 6 to 8 and 13 to 19. This confirms the most common variants found were from those regions. For example, *MUC4* is located on chromosome 3 and *CDC27* is located on chromosome 17. It is also clear that the most common mutation type is the C>T as can be seen from the red dots and confirms the earlier figure 3.8 (SNV class).

Most tumour samples are thought to be somewhat contaminated with matched normals. Of course, normal samples can be impure also. Figure 3.11 shows the VAF scores clustered for the cohort. In normal cells, VAF scores cluster around 50%, tumour cells vary considerable with them commonly clustering around 0% ans 100% [84]. Therefore it would seem that the there is some evidence for tumour contamination in my samples. Again, this may have come from the HN72 samples and it is expected to have contamination in a biopsy of this nature and the level of contamination appears low considering. However, clustering VAF's treat each subpopulation independently and do not consider that these frequencies are correlated by partitions of the same cellular population. Therefore, these

approaches are also limited. Also, it is possible that there is no clear variant caused founder clone in the tumour and that other events such as epigenetic changes could have played a part.

The excess mutational load observed in figure 3.14, when comapred to TCGA appears high. This could have been due to the fact that my samples were in a small cohort and contained late onset and recurrent HNSCC mutations. Significant pathways in HNSCC are the Wnt and signalling pathway and the cadherin related cell cycle mitotic super pathway. Future HNSCC research could consider the Olfactory signalling pathway and the Innate Immune System super pathway.

5 Conclusions

Significant pathways such as the unfolded protein response pathway associated with the *DDX11* gene, the signalling by gpcr pathway and the olfactory signalling pathway with genes such as *OR1C35* and more were pinpointed in this study. Some are not strongly studied wrt HNSCC and offer up potential for further research. Mutations were found on important genes such as *CTAGE5*, mucin family genes, *OR2L3*, *CDC27*, *NBPF10* and *NOTCH2A*. They provided confirmatory evidence of existing link with not only HNSCC but other closely related squamous cell carcinomas and were found to be related to existing pathways such as the Wnt and KEGG pathways.

These 154 variants were shown to have been exposed to chemo-radiation and conferred the tumour with resistance to therapeutics. This gave insights into the mechanisms at play such cell cycle control and mitotic timing events and provides potential for therapeutically actionable events for future research and treatment of HNSCC. A more evolved gene panel for HNSCC, targeted panels for re-sequencing and clearer evidence for causality with genes/variants are things that can be advanced moving forward. It was somewhat overwhelming when sifting through the vast amounts of databases when attempting to classify a variant based on impact, feature type, clinical significance etc.

The Broad institute provide an easy to use, accurate somatic variant caller along with a host of other tools to process data. For example, it took Mutect2 just 2.2 hours (1,132 minutes) to produce a VCF file containing somatic variant calls for sample HN51. This cements its good reputation and leaves it as a good option for clinical pipelines.

Limitations in the study included the low quality of the HN72 samples. This was seen with the poor HS metrics scores and Phred scores. Ultimately, this

ruled out one third of the data and effected accuracy. Nonetheless, Mutect2 allocated strict filtering scores and removed variants accordingly. Given the fact that the tumour samples came coupled with a matched normal meant that variants would be true variants. A larger cohort would be more useful in terms of identifying driver genes, comparing a cohort against TCGA and other useful utilities in maftools. With more time, the effect of microsatellite instability on mutational load across the genome may be looked at, which can determine eligibility for immunotherapy. Also, a confirmatory analysis with deep sequencing data would be used from the Ion Torrent platform. Tumour contamination is a concern in somatic calling pipelines and a tumour is expected to have a certain level of contamination. The advancement of single cell sequencing is recommended and should reduce tumour heterogenity and improve the accuracy in somatic variant calling pipelines.

Bibliography

[1] J. Y. Jang, N. Choi, Y. H. Ko, M. K. Chung, Y. I. Son, C. H. Baek, K. H. Baek, and H. S. Jeong. Treatment outcomes in metastatic and localized high-grade salivary gland cancer: high chance of cure with surgery and post-operative radiation in T1-2 N0 high-grade salivary gland cancer. *BMC Cancer*, 18(1):672, Jun 2018.

[2] C. R. Leemans, P. J. F. Snijders, and R. H. Brakenhoff. The molecular landscape of head and neck cancer. *Nat. Rev. Cancer*, 18(5):269–282, May 2018.

[3] C. R. Leemans, B. J. Braakhuis, and R. H. Brakenhoff. The molecular biology of head and neck cancer. *Nat. Rev. Cancer*, 11(1):9–22, Jan 2011.

[4] E. Brouns, J. Baart, K. h. Karagozoglu, I. Aartman, E. Bloemena, and I. van der Waal. Malignant transformation of oral leukoplakia in a well-defined cohort of 144 patients. *Oral Dis*, 20(3):19–24, Apr 2014.

[5] Xavier CastellsaguÃ©, Laia Alemany, Miquel Quer, Gordana Halec, Beatriz QuirÃ³s, $SaraTous, OmarClavero, LlciaAl^2s, ThorstenBiegner, TomaszSzafarowski, M$ $Cabezas, IsabelAlvarado - Cabrero, Chang - SukKang, Jin - KyoungOh, MarcialGarcia - Rojo, ErminaIljazovic, Oluseyi F.Ajayi, FloraDuarte, AshrafunNessa, LeopoldoTinoco$ $Padilla, Edyta C.Pirog, HalinaViarheichyk, HeslerMorales, Val©rieCostes, AnaF©li$ $Revilla, VclavMandys, Manuel E.Gonzlez, JulioVelasco, Ignacio G.Bravo, WimQuint, M$ $oz, Silvia deSanjos©, F. Xavier Bosch, theICOInternationalHPV inHead, andNeckCa$ $Comprehensiveassessmentofbiomarkersin3680patients.JNCI : JournaloftheNational$ $djv403, 2016.$

[6] C. H. Chung and M. L. Gillison. Human papillomavirus in head and neck cancer: its role in pathogenesis and clinical implications. *Clin. Cancer Res.*, 15(22):6758–6762, Nov 2009.

[7] M. C. Yu and J. M. Yuan. Epidemiology of nasopharyngeal carcinoma. *Semin. Cancer Biol.*, 12(6):421–429, Dec 2002.

[8] M. Hashibe, P. Brennan, S. C. Chuang, S. Boccia, X. Castellsague, C. Chen, M. P. Curado, L. Dal Maso, A. W. Daudt, E. Fabianova, L. Fernandez, V. Wunsch-Filho, S. Franceschi, R. B. Hayes, R. Herrero, K. Kelsey, S. Koifman, C. La Vecchia, P. Lazarus, F. Levi, J. J. Lence, D. Mates, E. Matos, A. Menezes, M. D. McClean, J. Muscat, J. Eluf-Neto, A. F. Olshan, M. Purdue, P. Rudnai, S. M. Schwartz, E. Smith, E. M. Sturgis, N. Szeszenia-Dabrowska, R. Talamini, Q. Wei, D. M. Winn, O. Shangina, A. Pilarska, Z. F. Zhang, G. Ferro, J. Berthiller, and P. Boffetta. Interaction between tobacco and alcohol use and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium. *Cancer Epidemiol. Biomarkers Prev.*, 18(2):541–550, Feb 2009.

[9] A. K. Chaturvedi, W. F. Anderson, J. Lortet-Tieulent, M. P. Curado, J. Ferlay, S. Franceschi, P. S. Rosenberg, F. Bray, and M. L. Gillison. Worldwide trends in incidence rates for oral cavity and oropharyngeal cancers. *J. Clin. Oncol.*, 31(36):4550–4559, Dec 2013.

[10] J. Adrien, C. Bertolus, L. Gambotti, A. Mallet, and B. Baujat. Why are head and neck squamous cell carcinoma diagnosed so late? Influence of health care disparities and socio-economic factors. *Oral Oncol.*, 50(2):90–97, Feb 2014.

[11] Aaron Smith, Anthony Grady, Francisco Vieira, and Merry Sebelik. Ultrasound-guided needle biopsy for diagnosis of advanced-stage malignancies of the upper aerodigestive tract. *OTO Open*, 1(1):2473974X17690132, 2017.

[12] W. Budach, E. Bolke, K. Kammers, P. A. Gerber, K. Orth, S. Gripp, and C. Matuschek. Induction chemotherapy followed by concurrent radio-

chemotherapy versus concurrent radio-chemotherapy alone as treatment of locally advanced squamous cell carcinoma of the head and neck (HNSCC): A meta-analysis of randomized trials. *Radiother Oncol*, 118(2):238–243, Feb 2016.

[13] J. A. Bonner, P. M. Harari, J. Giralt, N. Azarnia, D. M. Shin, R. B. Cohen, C. U. Jones, R. Sur, D. Raben, J. Jassem, R. Ove, M. S. Kies, J. Baselga, H. Youssoufian, N. Amellal, E. K. Rowinsky, and K. K. Ang. Radiotherapy plus cetuximab for squamous-cell carcinoma of the head and neck. *N. Engl. J. Med.*, 354(6):567–578, Feb 2006.

[14] J. Zhang and S. Zhang. Discovery of cancer common and specific driver gene sets. *Nucleic Acids Res.*, 45(10):e86, Jun 2017.

[15] Y. Chen, J. McGee, X. Chen, T. N. Doman, X. Gong, Y. Zhang, N. Hamm, X. Ma, R. E. Higgs, S. V. Bhagwat, S. Buchanan, S. B. Peng, K. A. Staschke, V. Yadav, Y. Yue, and H. Kouros-Mehr. Identification of druggable cancer driver genes amplified across TCGA datasets. *PLoS ONE*, 9(5):e98293, 2014.

[16] B. J. Raphael, J. R. Dobson, L. Oesper, and F. Vandin. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome Med*, 6(1):5, 2014.

[17] B. Vogelstein and K. W. Kinzler. Cancer genes and the pathways they control. *Nat. Med.*, 10(8):789–799, Aug 2004.

[18] C. H. Yeang, F. McCormick, and A. Levine. Combinatorial patterns of somatic gene mutations in cancer. *FASEB J.*, 22(8):2605–2622, Aug 2008.

[19] J. Zhang and S. Zhang. The Discovery of Mutated Driver Pathways in Cancer: Models and Algorithms. *IEEE/ACM Trans Comput Biol Bioinform*, 15(3):988–998, 2018.

[20] F. Vandin, E. Upfal, and B. J. Raphael. De novo discovery of mutated driver pathways in cancer. *Genome Res.*, 22(2):375–385, Feb 2012.

[21] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts,

A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortes, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D. A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. A. McCarroll, J. Mora, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. M. Roberts, J. A. Biegel, K. Stegmaier, A. J. Bass, L. A. Garraway, M. Meyerson, T. R. Golub, D. A. Gordenin, S. Sunyaev, E. S. Lander, and G. Getz. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, Jul 2013.

[22] M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, M. C. Wendl, J. Kim, B. Reardon, P. K. Ng, K. J. Jeong, S. Cao, Z. Wang, J. Gao, Q. Gao, F. Wang, E. M. Liu, L. Mularoni, C. Rubio-Perez, N. Nagarajan, I. Cortes-Ciriano, D. C. Zhou, W. W. Liang, J. M. Hess, V. D. Yellapantula, D. Tamborero, A. Gonzalez-Perez, C. Suphavilai, J. Y. Ko, E. Khurana, P. J. Park, E. M. Van Allen, H. Liang, M. S. Lawrence, A. Godzik, N. Lopez-Bigas, J. Stuart, D. Wheeler, G. Getz, K. Chen, A. J. Lazar, G. B. Mills, R. Karchin, L. Ding, S. J. Caesar-Johnson, J. A. Demchok, I. Felau, M. Kasapi, M. L. Ferguson, C. M. Hutter, H. J. Sofia, R. Tarnuzzer, Z. Wang, L. Yang, J. C. Zenklusen, J. J. Zhang, S. Chudamani, J. Liu, L. Lolla, R. Naresh, T. Pihl, Q. Sun, Y. Wan, Y. Wu, J. Cho, T. DeFreitas, S. Frazer, N. Gehlenborg, G. Getz, D. I. Heiman, J. Kim, M. S. Lawrence, P. Lin, S. Meier, M. S. Noble, G. Saksena, D. Voet, H. Zhang, B. Bernard, N. Chambwe, V. Dhankani, T. Knijnenburg, R. Kramer, K. Leinonen, Y. Liu, M. Miller, S. Reynolds, I. Shmulevich, V. Thorsson, W. Zhang, R. Akbani, B. M. Broom, A. M. Hegde, Z. Ju, R. S. Kanchi, A. Korkut, J. Li, H. Liang, S. Ling, W. Liu, Y. Lu, G. B. Mills, K. S. Ng, A. Rao, M. Ryan, J. Wang, J. N. Weinstein, J. Zhang, A. Abeshouse, J. Armenia, D. Chakravarty, W. K. Chatila, I. de Bruijn, J. Gao, B. E. Gross, Z. J. Heins, R. Kundra, K. La, M. Ladanyi, A. Luna, M. G. Nissan, A. Ochoa, S. M. Phillips, E. Reznik, F. Sanchez-Vega, C. Sander, N. Schultz,

R. Sheridan, S. O. Sumer, Y. Sun, B. S. Taylor, J. Wang, H. Zhang, P. Anur, M. Peto, P. Spellman, C. Benz, J. M. Stuart, C. K. Wong, C. Yau, D. N. Hayes, J. S. Parker, M. D. Wilkerson, A. Ally, M. Balasundaram, R. Bowlby, D. Brooks, R. Carlsen, E. Chuah, N. Dhalla, R. Holt, S. J. M. Jones, K. Kasaian, D. Lee, Y. Ma, M. A. Marra, M. Mayo, R. A. Moore, A. J. Mungall, K. Mungall, A. G. Robertson, S. Sadeghi, J. E. Schein, P. Sipahimalani, A. Tam, N. Thiessen, K. Tse, T. Wong, A. C. Berger, R. Beroukhim, A. D. Cherniack, C. Cibulskis, S. B. Gabriel, G. F. Gao, G. Ha, M. Meyerson, S. E. Schumacher, J. Shih, M. H. Kucherlapati, R. S. Kucherlapati, S. Baylin, L. Cope, L. Danilova, M. S. Bootwalla, P. H. Lai, D. T. Maglinte, D. J. Van Den Berg, D. J. Weisenberger, J. T. Auman, S. Balu, T. Bodenheimer, C. Fan, K. A. Hoadley, A. P. Hoyle, S. R. Jefferys, C. D. Jones, S. Meng, P. A. Mieczkowski, L. E. Mose, A. H. Perou, C. M. Perou, J. Roach, Y. Shi, J. V. Simons, T. Skelly, M. G. Soloway, D. Tan, U. Veluvolu, H. Fan, T. Hinoue, P. W. Laird, H. Shen, W. Zhou, M. Bellair, K. Chang, K. Covington, C. J. Creighton, H. Dinh, H. Doddapaneni, L. A. Donehower, J. Drummond, R. A. Gibbs, R. Glenn, W. Hale, Y. Han, J. Hu, V. Korchina, S. Lee, L. Lewis, W. Li, X. Liu, M. Morgan, D. Morton, D. Muzny, J. Santibanez, M. Sheth, E. Shinbrot, L. Wang, M. Wang, D. A. Wheeler, L. Xi, F. Zhao, J. Hess, E. L. Appelbaum, M. Bailey, M. G. Cordes, L. Ding, C. C. Fronick, L. A. Fulton, R. S. Fulton, C. Kandoth, E. R. Mardis, M. D. McLellan, C. A. Miller, H. K. Schmidt, R. K. Wilson, D. Crain, E. Curley, J. Gardner, K. Lau, D. Mallery, S. Morris, J. Paulauskis, R. Penny, C. Shelton, T. Shelton, M. Sherman, E. Thompson, P. Yena, J. Bowen, J. M. Gastier-Foster, M. Gerken, K. M. Leraas, T. M. Lichtenberg, N. C. Ramirez, L. Wise, E. Zmuda, N. Corcoran, T. Costello, C. Hovens, A. L. Carvalho, A. C. de Carvalho, J. H. Fregnani, A. Longatto-Filho, R. M. Reis, C. Scapulatempo-Neto, H. C. S. Silveira, D. O. Vidal, A. Burnette, J. Eschbacher, B. Hermes, A. Noss, R. Singh, M. L. Anderson, P. D. Castro, M. Ittmann, D. Huntsman, B. Kohl, X. Le, R. Thorp, C. Andry, E. R. Duffy, V. Lyadov, O. Paklina, G. Setdikova, A. Shabunin, M. Tavobilov, C. McPherson, R. Warnick, R. Berkowitz, D. Cramer, C. Feltmate, N. Horowitz, A. Kibel, M. Muto, C. P. Raut, A. Malykh, J. S. Barnholtz-Sloan, W. Barrett, K. Devine, J. Fulop, Q. T.

Ostrom, K. Shimmel, Y. Wolinsky, A. E. Sloan, A. De Rose, F. Giuliante, M. Goodman, B. Y. Karlan, C. H. Hagedorn, J. Eckman, J. Harr, J. Myers, K. Tucker, L. A. Zach, B. Deyarmin, H. Hu, L. Kvecher, C. Larson, R. J. Mural, S. Somiari, A. Vicha, T. Zelinka, J. Bennett, M. Iacocca, B. Rabeno, P. Swanson, M. Latour, L. Lacombe, B. Tetu, A. Bergeron, M. McGraw, S. M. Staugaitis, J. Chabot, H. Hibshoosh, A. Sepulveda, T. Su, T. Wang, O. Potapova, O. Voronina, L. Desjardins, O. Mariani, S. Roman-Roman, X. Sastre, M. H. Stern, F. Cheng, S. Signoretti, A. Berchuck, D. Bigner, E. Lipp, J. Marks, S. McCall, R. McLendon, A. Secord, A. Sharp, M. Behera, D. J. Brat, A. Chen, K. Delman, S. Force, F. Khuri, K. Magliocca, S. Maithel, J. J. Olson, T. Owonikoko, A. Pickens, S. Ramalingam, D. M. Shin, G. Sica, E. G. Van Meir, H. Zhang, W. Eijckenboom, A. Gillis, E. Korpershoek, L. Looijenga, W. Oosterhuis, H. Stoop, K. E. van Kessel, E. C. Zwarthoff, C. Calatozzolo, L. Cuppini, S. Cuzzubbo, F. DiMeco, G. Finocchiaro, L. Mattei, A. Perin, B. Pollo, C. Chen, J. Houck, P. Lohavanichbutr, A. Hartmann, C. Stoehr, R. Stoehr, H. Taubert, S. Wach, B. Wullich, W. Kycler, D. Murawa, M. Wiznerowicz, K. Chung, W. J. Edenfield, J. Martin, E. Baudin, G. Bubley, R. Bueno, A. De Rienzo, W. G. Richards, S. Kalkanis, T. Mikkelsen, H. Noushmehr, L. Scarpace, N. Girard, M. Aymerich, E. Campo, E. Gine, A. L. Guillermo, N. Van Bang, P. T. Hanh, B. D. Phu, Y. Tang, H. Colman, K. Evason, P. R. Dottino, J. A. Martignetti, H. Gabra, H. Juhl, T. Akeredolu, S. Stepa, D. Hoon, K. Ahn, K. J. Kang, F. Beuschlein, A. Breggia, M. Birrer, D. Bell, M. Borad, A. H. Bryce, E. Castle, V. Chandan, J. Cheville, J. A. Copland, M. Farnell, T. Flotte, N. Giama, T. Ho, M. Kendrick, J. P. Kocher, K. Kopp, C. Moser, D. Nagorney, D. O'Brien, B. P. O'Neill, T. Patel, G. Petersen, F. Que, M. Rivera, L. Roberts, R. Smallridge, T. Smyrk, M. Stanton, R. H. Thompson, M. Torbenson, J. D. Yang, L. Zhang, F. Brimo, J. A. Ajani, A. M. A. Gonzalez, C. Behrens, J. Bondaruk, R. Broaddus, B. Czerniak, B. Esmaeli, J. Fujimoto, J. Gershenwald, C. Guo, A. J. Lazar, C. Logothetis, F. Meric-Bernstam, C. Moran, L. Ramondetta, D. Rice, A. Sood, P. Tamboli, T. Thompson, P. Troncoso, A. Tsao, I. Wistuba, C. Carter, L. Haydu, P. Hersey, V. Jakrot, H. Kakavand, R. Kefford, K. Lee, G. Long, G. Mann, M. Quinn, R. Saw, R. Scolyer, K. Shan-

non, A. Spillane, J. Stretch, M. Synott, J. Thompson, J. Wilmott, H. Al-Ahmadie, T. A. Chan, R. Ghossein, A. Gopalan, D. A. Levine, V. Reuter, S. Singer, B. Singh, N. V. Tien, T. Broudy, C. Mirsaidi, P. Nair, P. Drwiega, J. Miller, J. Smith, H. Zaren, J. W. Park, N. P. Hung, E. Kebebew, W. M. Linehan, A. R. Metwalli, K. Pacak, P. A. Pinto, M. Schiffman, L. S. Schmidt, C. D. Vocke, N. Wentzensen, R. Worrell, H. Yang, M. Moncrieff, C. Goparaju, J. Melamed, H. Pass, N. Botnariuc, I. Caraman, M. Cernat, I. Chemencedji, A. Clipca, S. Doruc, G. Gorincioi, S. Mura, M. Pirtac, I. Stancul, D. Tcaciuc, M. Albert, I. Alexopoulou, A. Arnaout, J. Bartlett, J. Engel, S. Gilbert, J. Parfitt, H. Sekhon, G. Thomas, D. M. Rassl, R. C. Rintoul, C. Bifulco, R. Tamakawa, W. Urba, N. Hayward, H. Timmers, A. Antenucci, F. Facciolo, G. Grazi, M. Marino, R. Merola, R. de Krijger, A. P. Gimenez-Roqueplo, A. Piche, S. Chevalier, G. McKercher, K. Birsoy, G. Barnett, C. Brewer, C. Farver, T. Naska, N. A. Pennell, D. Raymond, C. Schilero, K. Smolenski, F. Williams, C. Morrison, J. A. Borgia, M. J. Liptay, M. Pool, C. W. Seder, K. Junker, L. Omberg, M. Dinkin, G. Manikhas, D. Alvaro, M. C. Bragazzi, V. Cardinale, G. Carpino, E. Gaudio, D. Chesla, S. Cottingham, M. Dubina, F. Moiseenko, R. Dhanasekaran, K. F. Becker, K. P. Janssen, J. Slotta-Huspenina, M. H. Abdel-Rahman, D. Aziz, S. Bell, C. M. Cebulla, A. Davis, R. Duell, J. B. Elder, J. Hilty, B. Kumar, J. Lang, N. L. Lehman, R. Mandt, P. Nguyen, R. Pilarski, K. Rai, L. Schoenfield, K. Senecal, P. Wakely, P. Hansen, R. Lechan, J. Powers, A. Tischler, W. E. Grizzle, K. C. Sexton, A. Kastl, J. Henderson, S. Porten, J. Waldmann, M. Fassnacht, S. L. Asa, D. Schadendorf, M. Couce, M. Graefen, H. Huland, G. Sauter, T. Schlomm, R. Simon, P. Tennstedt, O. Olabode, M. Nelson, O. Bathe, P. R. Carroll, J. M. Chan, P. Disaia, P. Glenn, R. K. Kelley, C. N. Landen, J. Phillips, M. Prados, J. Simko, K. Smith-McCune, S. VandenBerg, K. Roggin, A. Fehrenbach, A. Kendler, S. Sifri, R. Steele, A. Jimeno, F. Carey, I. Forgie, M. Mannelli, M. Carney, B. Hernandez, B. Campos, C. Herold-Mende, C. Jungk, A. Unterberg, A. von Deimling, A. Bossler, J. Galbraith, L. Jacobus, M. Knudson, T. Knutson, D. Ma, M. Milhem, R. Sigmund, A. K. Godwin, R. Madan, H. G. Rosenthal, C. Adebamowo, S. N. Adebamowo, A. Boussioutas, D. Beer, T. Gior-

dano, A. M. Mes-Masson, F. Saad, T. Bocklage, L. Landrum, R. Mannel, K. Moore, K. Moxley, R. Postier, J. Walker, R. Zuna, M. Feldman, F. Valdivieso, R. Dhir, J. Luketich, E. M. M. Pinero, M. Quintero-Aguilo, C. G. Carlotti, J. S. Dos Santos, R. Kemp, A. Sankarankuty, D. Tirapelli, J. Catto, K. Agnew, E. Swisher, J. Creaney, B. Robinson, C. S. Shelley, E. M. Godwin, S. Kendall, C. Shipman, C. Bradford, T. Carey, A. Haddad, J. Moyer, L. Peterson, M. Prince, L. Rozek, G. Wolf, R. Bowman, K. M. Fong, I. Yang, R. Korst, W. K. Rathmell, J. L. Fantacone-Campbell, J. A. Hooke, A. J. Kovatich, C. D. Shriver, J. DiPersio, B. Drake, R. Govindan, S. Heath, T. Ley, B. Van Tine, P. Westervelt, M. A. Rubin, J. I. Lee, N. D. Aredes, and A. Mariamidze. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, 173(2):371–385, Apr 2018.

[23] A. Torkamani and N. J. Schork. Prediction of cancer driver mutations in protein kinases. *Cancer Res.*, 68(6):1675–1682, Mar 2008.

[24] H. Carter, S. Chen, L. Isik, S. Tyekucheva, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, and R. Karchin. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.*, 69(16):6660–6667, Aug 2009.

[25] B. Solomon, R. J. Young, and D. Rischin. Head and neck squamous cell carcinoma: Genomics and emerging biomarkers for immunomodulatory cancer treatments. *Semin. Cancer Biol.*, Jan 2018.

[26] M. S. Lawrence, C. Sougnez, L. Lichtenstein, K. Cibulskis, E. Lander, S. B. Gabriel, G. Getz, A. Ally, M. Balasundaram, I. Birol, R. Bowlby, D. Brooks, Y. S. Butterfield, R. Carlsen, D. Cheng, A. Chu, N. Dhalla, R. Guin, R. A. Holt, S. J. Jones, D. Lee, H. I. Li, M. A. Marra, M. Mayo, R. A. Moore, A. J. Mungall, A. G. Robertson, J. E. Schein, P. Sipahimalani, A. Tam, N. Thiessen, T. Wong, A. Protopopov, N. Santoso, S. Lee, M. Parfenov, J. Zhang, H. S. Mahadeshwar, J. Tang, X. Ren, S. Seth, P. Haseley, D. Zeng, L. Yang, A. W. Xu, X. Song, A. Pantazi, C. A. Bristow, A. Hadjipanayis, J. Seidman, L. Chin, P. J. Park, R. Kucherlapati, R. Akbani, T. Casasent, W. Liu, Y. Lu, G. Mills, T. Motter, J. Weinstein, L. Diao,

J. Wang, Y. Hong Fan, J. Liu, K. Wang, J. T. Auman, S. Balu, T. Bodenheimer, E. Buda, D. N. Hayes, K. A. Hoadley, A. P. Hoyle, S. R. Jefferys, C. D. Jones, P. K. Kimes, Y. Liu, J. S. Marron, S. Meng, P. A. Mieczkowski, L. E. Mose, J. S. Parker, C. M. Perou, J. F. Prins, J. Roach, Y. Shi, J. V. Simons, D. Singh, M. G. Soloway, D. Tan, U. Veluvolu, V. Walter, S. Waring, M. D. Wilkerson, J. Wu, N. Zhao, A. D. Cherniack, P. S. Hammerman, A. D. Tward, C. Sekhar Pedamallu, G. Saksena, J. Jung, A. I. Ojesina, S. L. Carter, T. I. Zack, S. E. Schumacher, R. Beroukhim, S. S. Freeman, M. Meyerson, J. Cho, L. Chin, G. Getz, M. S. Noble, D. DiCara, H. Zhang, D. I. Heiman, N. Gehlenborg, D. Voet, P. Lin, S. Frazer, P. Stojanov, Y. Liu, L. Zou, J. Kim, C. Sougnez, S. B. Gabriel, M. S. Lawrence, D. Muzny, H. Doddapaneni, C. Kovar, J. Reid, D. Morton, Y. Han, W. Hale, H. Chao, K. Chang, J. A. Drummond, R. A. Gibbs, N. Kakkar, D. Wheeler, L. Xi, G. Ciriello, M. Ladanyi, W. Lee, R. Ramirez, C. Sander, R. Shen, R. Sinha, N. Weinhold, B. S. Taylor, B. A. Aksoy, G. Dresdner, J. Gao, B. Gross, A. Jacobsen, B. Reva, N. Schultz, S. O. Sumer, Y. Sun, T. A. Chan, L. G. Morris, J. Stuart, S. Benz, S. Ng, C. Benz, C. Yau, S. B. Baylin, L. Cope, L. Danilova, J. G. Herman, M. Bootwalla, D. T. Maglinte, P. W. Laird, T. Triche, D. J. Weisenberger, D. J. Van Den Berg, N. Agrawal, J. Bishop, P. C. Boutros, J. P. Bruce, L. Averett Byers, J. Califano, T. E. Carey, Z. Chen, H. Cheng, S. I. Chiosea, E. Cohen, B. Diergaarde, A. M. Egloff, A. K. El-Naggar, R. L. Ferris, M. J. Frederick, J. R. Grandis, Y. Guo, R. I. Haddad, P. S. Hammerman, T. Harris, D. N. Hayes, A. B. Hui, J. J. Lee, S. M. Lippman, F. F. Liu, J. B. McHugh, J. Myers, P. Kwok Shing Ng, B. Perez-Ordonez, C. R. Pickering, M. Prystowsky, M. Romkes, A. D. Saleh, M. A. Sartor, R. Seethala, T. Y. Seiwert, H. Si, A. D. Tward, C. Van Waes, D. M. Waggott, M. Wiznerowicz, W. G. Yarbrough, J. Zhang, Z. Zuo, K. Burnett, D. Crain, J. Gardner, K. Lau, D. Mallery, S. Morris, J. Paulauskis, R. Penny, C. Shelton, T. Shelton, M. Sherman, P. Yena, A. D. Black, J. Bowen, J. Frick, J. M. Gastier-Foster, H. A. Harper, K. Leraas, T. M. Lichtenberg, N. C. Ramirez, L. Wise, E. Zmuda, J. Baboud, M. A. Jensen, A. B. Kahn, T. D. Pihl, D. A. Pot, D. Srinivasan, J. S. Walton, Y. Wan, R. A. Burton, T. Davidsen, J. A. Demchok, G. Eley, M. L. Ferguson, K. R. Mills Shaw, B. A. Ozenberger,

M. Sheth, H. J. Sofia, R. Tarnuzzer, Z. Wang, L. Yang, J. C. Zenklusen, C. Saller, K. Tarvin, C. Chen, R. Bollag, P. Weinberger, W. Golusi?ski, P. Golusi?ski, M. Ibbs, K. Korski, A. Mackiewicz, W. Suchorska, B. Szybiak, M. Wiznerowicz, K. Burnett, E. Curley, J. Gardner, D. Mallery, R. Penny, T. Shelton, P. Yena, C. Beard, C. Mitchell, G. Sandusky, N. Agrawal, J. Ahn, J. Bishop, J. Califano, Z. Khan, J. P. Bruce, A. B. Hui, J. Irish, F. F. Liu, B. Perez-Ordonez, J. Waldron, P. C. Boutros, D. M. Waggott, J. Myers, W. N. William, S. M. Lippman, S. Egea, C. Gomez-Fernandez, L. Herbert, C. R. Bradford, T. E. Carey, D. B. Chepeha, A. S. Haddad, T. R. Jones, C. M. Komarck, M. Malakh, J. B. McHugh, J. S. Moyer, A. Nguyen, L. A. Peterson, M. E. Prince, L. S. Rozek, M. A. Sartor, E. G. Taylor, H. M. Walline, G. T. Wolf, L. Boice, B. S. Chera, W. K. Funkhouser, M. L. Gulley, T. G. Hackman, D. N. Hayes, M. C. Hayward, M. Huang, W. K. Rathmell, A. H. Salazar, W. W. Shockley, C. G. Shores, L. Thorne, M. C. Weissler, S. Wrenn, A. M. Zanation, S. I. Chiosea, B. Diergaarde, A. M. Egloff, R. L. Ferris, M. Romkes, R. Seethala, B. T. Brown, Y. Guo, M. Pham, and W. G. Yarbrough. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517(7536):576–582, Jan 2015.

[27] P. S. Hammerman, M. S. Lawrence, D. Voet, R. Jing, K. Cibulskis, A. Sivachenko, P. Stojanov, A. McKenna, E. S. Lander, S. Gabriel, G. Getz, C. Sougnez, M. Imielinski, E. Helman, B. Hernandez, N. H. Pho, M. Meyerson, A. Chu, H. J. Chun, A. J. Mungall, E. Pleasance, A. Robertson, P. Sipahimalani, D. Stoll, M. Balasundaram, I. Birol, Y. S. Butterfield, E. Chuah, R. J. Coope, R. Corbett, N. Dhalla, R. Guin, A. He, C. Hirst, M. Hirst, R. A. Holt, D. Lee, H. I. Li, M. Mayo, R. A. Moore, K. Mungall, K. M. Nip, A. Olshen, J. E. Schein, J. R. Slobodan, A. Tam, N. Thiessen, R. Varhol, T. Zeng, Y. Zhao, S. J. Jones, M. A. Marra, G. Saksena, A. D. Cherniack, S. E. Schumacher, B. Tabak, S. L. Carter, N. H. Pho, H. Nguyen, R. C. Onofrio, A. Crenshaw, K. Ardlie, R. Beroukhim, W. Winckler, P. S. Hammerman, G. Getz, M. Meyerson, A. Protopopov, J. Zhang, A. Hadjipanayis, S. Lee, R. Xi, L. Yang, X. Ren, H. Zhang, S. Shukla, P. C. Chen, P. Haseley, E. Lee, L. Chin, P. J. Park, R. Kucherlapati, N. D. Socci, Y. Liang, N. Schultz, L. Borsu, A. E. Lash, A. Viale, C. Sander, M. Ladanyi,

T. Auman, K. A. Hoadley, M. D. Wilkerson, Y. Shi, C. Liquori, S. Meng, L. Li, Y. J. Turman, M. D. Topal, D. Tan, S. Waring, E. Buda, J. Walsh, C. D. Jones, P. A. Mieczkowski, D. Singh, J. Wu, A. Gulabani, P. Dolina, T. Bodenheimer, A. P. Hoyle, J. V. Simons, M. G. Soloway, L. E. Mose, S. R. Jefferys, S. Balu, B. D. O'Connor, J. F. Prins, J. Liu, D. Y. Chiang, D. Hayes, C. M. Perou, L. Cope, L. Danilova, D. J. Weisenberger, D. T. Maglinte, F. Pan, D. J. Van Den Berg, T. Triche, J. G. Herman, S. B. Baylin, P. W. Laird, G. Getz, M. Noble, D. Voet, G. Saksena, N. Gehlenborg, D. DiCara, J. Zhang, H. Zhang, C. J. Wu, S. Y. Liu, M. S. Lawrence, L. Zou, A. Sivachenko, P. Lin, P. Stojanov, R. Jing, J. Cho, M. D. Nazaire, J. Robinson, H. Thorvaldsdottir, J. Mesirov, P. J. Park, L. Chin, N. Schultz, R. Sinha, G. Ciriello, E. Cerami, B. Gross, A. Jacobsen, J. Gao, B. Aksoy, N. Weinhold, R. Ramirez, B. S. Taylor, Y. Antipin, B. Reva, R. Shen, Q. Mo, V. Seshan, P. K. Paik, M. Ladanyi, C. Sander, R. Akbani, N. Zhang, B. M. Broom, T. Casasent, A. Unruh, C. Wakefield, R. Cason, K. A. Baggerly, J. N. Weinstein, D. Haussler, C. C. Benz, J. M. Stuart, J. Zhu, C. Szeto, G. K. Scott, C. Yau, S. Ng, T. Goldstein, P. Waltman, A. Sokolov, K. Ellrott, E. A. Collisson, D. Zerbino, C. Wilks, S. Ma, B. Craft, M. D. Wilkerson, J. Auman, K. A. Hoadley, Y. Du, C. Cabanski, V. Walter, D. Singh, J. Wu, A. Gulabani, T. Bodenheimer, A. P. Hoyle, J. V. Simons, M. G. Soloway, L. E. Mose, S. R. Jefferys, S. Balu, J. Marron, Y. Liu, K. Wang, J. Liu, J. F. Prins, D. Hayes, C. M. Perou, C. J. Creighton, Y. Zhang, W. D. Travis, N. Rekhtman, J. Yi, M. C. Aubry, R. Cheney, S. Dacic, D. Flieder, W. Funkhouser, P. Illei, J. Myers, M. S. Tsao, R. Penny, D. Mallery, T. Shelton, M. Hatfield, S. Morris, P. Yena, C. Shelton, M. Sherman, J. Paulauskis, M. Meyerson, S. B. Baylin, R. Govindan, R. Akbani, I. Azodo, D. Beer, R. Bose, L. A. Byers, D. Carbone, L. W. Chang, D. Chiang, A. Chu, E. Chun, E. Collisson, L. Cope, C. J. Creighton, L. Danilova, L. Ding, G. Getz, P. S. Hammerman, D. Hayes, B. Hernandez, J. G. Herman, J. Heymach, C. Ida, M. Imielinski, B. Johnson, I. Jurisica, J. Kaufman, F. Kosari, R. Kucherlapati, D. Kwiatkowski, M. Ladanyi, M. S. Lawrence, C. A. Maher, A. Mungall, S. Ng, W. Pao, M. Peifer, R. Penny, G. Robertson, V. Rusch, C. Sander, N. Schultz, R. Shen, J. Siegfried, R. Sinha, A. Sivachenko, C. Sougnez,

D. Stoll, J. Stuart, R. K. Thomas, S. Tomaszek, M. S. Tsao, W. D. Travis, C. Vaske, J. N. Weinstein, D. Weisenberger, D. Wheeler, D. A. Wigle, M. D. Wilkerson, C. Wilks, P. Yang, J. J. Zhang, M. A. Jensen, R. Sfeir, A. B. Kahn, A. L. Chu, P. Kothiyal, Z. Wang, E. E. Snyder, J. Pontius, T. D. Pihl, B. Ayala, M. Backus, J. Walton, J. Baboud, D. L. Berton, M. C. Nicholls, D. Srinivasan, R. Raman, S. Girshik, P. A. Kigonya, S. Alonso, R. N. Sanbhadti, S. P. Barletta, J. M. Greene, D. A. Pot, M. S. Tsao, B. Bandarchi-Chamkhaleh, J. Boyd, J. Weaver, D. A. Wigle, I. A. Azodo, S. C. Tomaszek, M. C. Aubry, C. M. Ida, P. Yang, F. Kosari, M. V. Brock, K. Rodgers, M. Rutledge, T. Brown, B. Lee, J. Shin, D. Trusty, R. Dhir, J. M. Siegfried, O. Potapova, K. V. Fedosenko, E. Nemirovich-Danchenko, V. Rusch, M. Zakowski, M. V. Iacocca, J. Brown, B. Rabeno, C. Czerwinski, N. Petrelli, Z. Fan, N. Todaro, J. Eckman, J. Myers, W. Rathmell, L. B. Thorne, M. Huang, L. Boice, A. Hill, R. Penny, D. Mallery, E. Curley, C. Shelton, P. Yena, C. Morrison, C. Gaudioso, J. M. Bartlett, S. Kodeeswaran, B. Zanke, H. Sekhon, K. David, H. Juhl, X. Van Le, B. Kohl, R. Thorp, V. T. Nguyen, V. B. Nguyen, H. Sussman, B. D. Phu, R. Hajek, P. H. Nguyen, K. Z. Khan, T. Muley, K. R. Shaw, M. Sheth, L. Yang, K. Buetow, T. Davidsen, J. A. Demchok, G. Eley, M. Ferguson, L. A. Dillon, C. Schaefer, M. S. Guyer, B. A. Ozenberger, J. D. Palchik, J. Peterson, H. J. Sofia, and E. Thomson. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, Sep 2012.

[28] N. Andor, T. A. Graham, M. Jansen, L. C. Xia, C. A. Aktipis, C. Petritsch, H. P. Ji, and C. C. Maley. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.*, 22(1):105–113, Jan 2016.

[29] M. E. Prince, R. Sivanandan, A. Kaczorowski, G. T. Wolf, M. J. Kaplan, P. Dalerba, I. L. Weissman, M. F. Clarke, and L. E. Ailles. Identification of a subpopulation of cells with cancer stem cell properties in head and neck squamous cell carcinoma. *Proc. Natl. Acad. Sci. U.S.A.*, 104(3):973–978, Jan 2007.

[30] E. A. Mroz and J. W. Rocco. MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous

cell carcinoma. *Oral Oncol.*, 49(3):211–215, Mar 2013.

[31] D. P. SLAUGHTER, H. W. SOUTHWICK, and W. SMEJKAL. Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. *Cancer*, 6(5):963–968, Sep 1953.

[32] J. Califano, P. van der Riet, W. Westra, H. Nawroz, G. Clayman, S. Piantadosi, R. Corio, D. Lee, B. Greenberg, W. Koch, and D. Sidransky. Genetic progression model for head and neck cancer: implications for field cancerization. *Cancer Res.*, 56(11):2488–2492, Jun 1996.

[33] M. P. Tabor, R. H. Brakenhoff, V. M. van Houten, J. A. Kummer, M. H. Snel, P. J. Snijders, G. B. Snow, C. R. Leemans, and B. J. Braakhuis. Persistence of genetically altered fields in head and neck cancer patients: biological and clinical implications. *Clin. Cancer Res.*, 7(6):1523–1532, Jun 2001.

[34] M. P. Tabor, R. H. Brakenhoff, H. J. Ruijter-Schippers, J. A. Kummer, C. R. Leemans, and B. J. Braakhuis. Genetically altered fields as origin of locally recurrent head and neck cancer: a retrospective study. *Clin. Cancer Res.*, 10(11):3607–3613, Jun 2004.

[35] A. P. Graveland, P. J. Golusinski, M. Buijze, R. Douma, N. Sons, D. J. Kuik, E. Bloemena, C. R. Leemans, R. H. Brakenhoff, and B. J. Braakhuis. Loss of heterozygosity at 9p and p53 immunopositivity in surgical margins predict local relapse in head and neck squamous cell carcinoma. *Int. J. Cancer*, 128(8):1852–1859, Apr 2011.

[36] M. Partridge, S. Pateromichelakis, E. Phillips, G. G. Emilion, R. P. A'Hern, and J. D. Langdon. A case-control study confirms that microsatellite assay can identify patients at risk of developing oral squamous cell carcinoma within a field of cancerization. *Cancer Res.*, 60(14):3893–3898, Jul 2000.

[37] V. M. van Houten, M. P. Tabor, M. W. van den Brekel, J. A. Kummer, F. Denkers, J. Dijkstra, R. Leemans, I. van der Waal, G. B. Snow, and R. H. Brakenhoff. Mutated p53 as a molecular marker for the diagnosis of head and neck cancer. *J. Pathol.*, 198(4):476–486, Dec 2002.

[38] A. S. Jonason, S. Kunala, G. J. Price, R. J. Restifo, H. M. Spinelli, J. A. Persing, D. J. Leffell, R. E. Tarone, and D. E. Brash. Frequent clones of p53-mutated keratinocytes in normal human skin. *Proc. Natl. Acad. Sci. U.S.A.*, 93(24):14025–14029, Nov 1996.

[39] K. Pantel and R. H. Brakenhoff. Dissecting the metastatic cascade. *Nat. Rev. Cancer*, 4(6):448–456, Jun 2004.

[40] P. C. Sun, R. Uppaluri, A. P. Schmidt, M. E. Pashia, E. C. Quant, J. B. Sunwoo, S. M. Gollin, and S. B. Scholnick. Transcript map of the 8p23 putative tumor suppressor region. *Genomics*, 75(1-3):17–25, Jul 2001.

[41] S. B. Scholnick, B. H. Haughey, J. B. Sunwoo, S. K. el Mofty, J. D. Baty, J. F. Piccirillo, and M. R. Zequeira. Chromosome 8 allelic loss and the outcome of patients with squamous cell carcinoma of the supraglottic larynx. *J. Natl. Cancer Inst.*, 88(22):1676–1682, Nov 1996.

[42] H. Ikushima and K. Miyazono. TGFbeta signalling: a complex web in cancer progression. *Nat. Rev. Cancer*, 10(6):415–424, Jun 2010.

[43] G. L. Johnson and R. Lapadat. Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases. *Science*, 298(5600):1911–1912, Dec 2002.

[44] M. Martini, M. C. De Santis, L. Braccini, F. Gulluni, and E. Hirsch. PI3K/AKT signaling pathway and cancer: an updated review. *Ann. Med.*, 46(6):372–383, Sep 2014.

[45] M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. *Nature*, 458(7239):719–724, Apr 2009.

[46] R. McLendon, A. Friedman, D. Bigner, E. G. Van Meir, D. J. Brat, G. M. Mastrogianakis, J. J. Olson, T. Mikkelsen, N. Lehman, K. Aldape, W. K. Yung, O. Bogler, J. N. Weinstein, S. VandenBerg, M. Berger, M. Prados, D. Muzny, M. Morgan, S. Scherer, A. Sabo, L. Nazareth, L. Lewis, O. Hall, Y. Zhu, Y. Ren, O. Alvi, J. Yao, A. Hawes, S. Jhangiani, G. Fowler, A. San Lucas, C. Kovar, A. Cree, H. Dinh, J. Santibanez, V. Joshi,

M. L. Gonzalez-Garay, C. A. Miller, A. Milosavljevic, L. Donehower, D. A. Wheeler, R. A. Gibbs, K. Cibulskis, C. Sougnez, T. Fennell, S. Mahan, J. Wilkinson, L. Ziaugra, R. Onofrio, T. Bloom, R. Nicol, K. Ardlie, J. Baldwin, S. Gabriel, E. S. Lander, L. Ding, R. S. Fulton, M. D. McLellan, J. Wallis, D. E. Larson, X. Shi, R. Abbott, L. Fulton, K. Chen, D. C. Koboldt, M. C. Wendl, R. Meyer, Y. Tang, L. Lin, J. R. Osborne, B. H. Dunford-Shore, T. L. Miner, K. Delehaunty, C. Markovic, G. Swift, W. Courtney, C. Pohl, S. Abbott, A. Hawkins, S. Leong, C. Haipek, H. Schmidt, M. Wiechert, T. Vickery, S. Scott, D. J. Dooling, A. Chinwalla, G. M. Weinstock, E. R. Mardis, R. K. Wilson, G. Getz, W. Winckler, R. G. Verhaak, M. S. Lawrence, M. O'Kelly, J. Robinson, G. Alexe, R. Beroukhim, S. Carter, D. Chiang, J. Gould, S. Gupta, J. Korn, C. Mermel, J. Mesirov, S. Monti, H. Nguyen, M. Parkin, M. Reich, N. Stransky, B. A. Weir, L. Garraway, T. Golub, M. Meyerson, L. Chin, A. Protopopov, J. Zhang, I. Perna, S. Aronson, N. Sathiamoorthy, G. Ren, J. Yao, W. R. Wiedemeyer, H. Kim, S. W. Kong, Y. Xiao, I. S. Kohane, J. Seidman, P. J. Park, R. Kucherlapati, P. W. Laird, L. Cope, J. G. Herman, D. J. Weisenberger, F. Pan, D. Van den Berg, L. Van Neste, J. M. Yi, K. E. Schuebel, S. B. Baylin, D. M. Absher, J. Z. Li, A. Southwick, S. Brady, A. Aggarwal, T. Chung, G. Sherlock, J. D. Brooks, R. M. Myers, P. T. Spellman, E. Purdom, L. R. Jakkula, A. V. Lapuk, H. Marr, S. Dorton, Y. G. Choi, J. Han, A. Ray, V. Wang, S. Durinck, M. Robinson, N. J. Wang, K. Vranizan, V. Peng, E. Van Name, G. V. Fontenay, J. Ngai, J. G. Conboy, B. Parvin, H. S. Feiler, T. P. Speed, J. W. Gray, C. Brennan, N. D. Socci, A. Olshen, B. S. Taylor, A. Lash, N. Schultz, B. Reva, Y. Antipin, A. Stukalov, B. Gross, E. Cerami, W. Q. Wang, L. X. Qin, V. E. Seshan, L. Villafania, M. Cavatore, L. Borsu, A. Viale, W. Gerald, C. Sander, M. Ladanyi, C. M. Perou, D. N. Hayes, M. D. Topal, K. A. Hoadley, Y. Qi, S. Balu, Y. Shi, J. Wu, R. Penny, M. Bittner, T. Shelton, E. Lenkiewicz, S. Morris, D. Beasley, S. Sanders, A. Kahn, R. Sfeir, J. Chen, D. Nassau, L. Feng, E. Hickey, A. Barker, D. S. Gerhard, J. Vockley, C. Compton, J. Vaught, P. Fielding, M. L. Ferguson, C. Schaefer, J. Zhang, S. Madhavan, K. H. Buetow, F. Collins, P. Good, M. Guyer, B. Ozenberger, J. Peterson, and E. Thomson. Comprehensive ge-

nomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, Oct 2008.

[47] F. Vandin, E. Upfal, and B. J. Raphael. Finding driver pathways in cancer: models and algorithms. *Algorithms Mol Biol*, 7(1):23, Sep 2012.

[48] L. Toledo, K. J. Neelsen, and J. Lukas. Replication Catastrophe: When a Checkpoint Fails because of Exhaustion. *Mol. Cell*, 66(6):735–749, Jun 2017.

[49] A. J. Levine and M. Oren. The first 30 years of p53: growing ever more complex. *Nat. Rev. Cancer*, 9(10):749–758, 10 2009.

[50] S. Y. Lin, K. Makino, W. Xia, A. Matin, Y. Wen, K. Y. Kwong, L. Bourguignon, and M. C. Hung. Nuclear localization of EGF receptor and its potential new role as a transcription factor. *Nat. Cell Biol.*, 3(9):802–808, Sep 2001.

[51] R. Moser, C. Xu, M. Kao, J. Annis, L. A. Lerma, C. M. Schaupp, K. E. Gurley, I. S. Jang, A. Biktasova, W. G. Yarbrough, A. A. Margolin, C. Grandori, C. J. Kemp, and E. Mendez. Functional kinomics identifies candidate therapeutic targets in head and neck cancer. *Clin. Cancer Res.*, 20(16):4274–4288, Aug 2014.

[52] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*, 6(269):pl1, Apr 2013.

[53] F. van Roy and G. Berx. The cell-cell adhesion molecule E-cadherin. *Cell. Mol. Life Sci.*, 65(23):3756–3788, Nov 2008.

[54] T. Tanoue and M. Takeichi. Mammalian Fat1 cadherin regulates actin dynamics and cell-cell contact. *J. Cell Biol.*, 165(4):517–528, May 2004.

[55] L. G. Morris, A. M. Kaufman, Y. Gong, D. Ramaswami, L. A. Walsh, ?. Turcan, S. Eng, K. Kannan, Y. Zou, L. Peng, V. E. Banuchi, P. Paty, Z. Zeng,

E. Vakiani, D. Solit, B. Singh, I. Ganly, L. Liau, T. C. Cloughesy, P. S. Mischel, I. K. Mellinghoff, and T. A. Chan. Recurrent somatic mutation of FAT1 in multiple human cancers leads to aberrant Wnt activation. *Nat. Genet.*, 45(3):253–261, Mar 2013.

[56] G. V. Schimizzi and G. D. Longmore. Ajuba proteins. *Curr. Biol.*, 25(11):R445–446, Jun 2015.

[57] T. Hirota, N. Kunitoku, T. Sasayama, T. Marumoto, D. Zhang, M. Nitta, K. Hatakeyama, and H. Saya. Aurora-A and an interacting activator, the LIM protein Ajuba, are required for mitotic commitment in human cells. *Cell*, 114(5):585–598, Sep 2003.

[58] G. Sun and K. D. Irvine. Ajuba family proteins link JNK to Hippo signaling. *Sci Signal*, 6(292):ra81, Sep 2013.

[59] H. Marie, S. J. Pratt, M. Betson, H. Epple, J. T. Kittler, L. Meek, S. J. Moss, S. Troyanovsky, D. Attwell, G. D. Longmore, and V. M. Braga. The LIM protein Ajuba is recruited to cadherin-dependent cell junctions through an association with alpha-catenin. *J. Biol. Chem.*, 278(2):1220–1228, Jan 2003.

[60] J. Kanungo, S. J. Pratt, H. Marie, and G. D. Longmore. Ajuba, a cytosolic LIM protein, shuttles into the nucleus and affects embryonal cell proliferation and fate decisions. *Mol. Biol. Cell*, 11(10):3299–3313, Oct 2000.

[61] K. Haraguchi, M. Ohsugi, Y. Abe, K. Semba, T. Akiyama, and T. Yamamoto. Ajuba negatively regulates the Wnt signaling pathway by promoting GSK-3beta-mediated phosphorylation of beta-catenin. *Oncogene*, 27(3):274–284, Jan 2008.

[62] R. Kopan and M. X. Ilagan. The canonical Notch signaling pathway: unfolding the activation mechanism. *Cell*, 137(2):216–233, Apr 2009.

[63] P. Ntziachristos, J. S. Lim, J. Sage, and I. Aifantis. From fly wings to targeted cancer therapies: a centennial for notch signaling. *Cancer Cell*, 25(3):318–334, Mar 2014.

[64] N. Stransky, A. M. Egloff, A. D. Tward, A. D. Kostic, K. Cibulskis, A. Sivachenko, G. V. Kryukov, M. S. Lawrence, C. Sougnez, A. McKenna, E. Shefler, A. H. Ramos, P. Stojanov, S. L. Carter, D. Voet, M. L. Cortes, D. Auclair, M. F. Berger, G. Saksena, C. Guiducci, R. C. Onofrio, M. Parkin, M. Romkes, J. L. Weissfeld, R. R. Seethala, L. Wang, C. Rangel-Escareno, J. C. Fernandez-Lopez, A. Hidalgo-Miranda, J. Melendez-Zajgla, W. Winckler, K. Ardlie, S. B. Gabriel, M. Meyerson, E. S. Lander, G. Getz, T. R. Golub, L. A. Garraway, and J. R. Grandis. The mutational landscape of head and neck squamous cell carcinoma. *Science*, 333(6046):1157–1160, Aug 2011.

[65] W. Sun, D. A. Gaykalova, M. F. Ochs, E. Mambo, D. Arnaoutakis, Y. Liu, M. Loyo, N. Agrawal, J. Howard, R. Li, S. Ahn, E. Fertig, D. Sidransky, J. Houghton, K. Buddavarapu, T. Sanford, A. Choudhary, W. Darden, A. Adai, G. Latham, J. Bishop, R. Sharma, W. H. Westra, P. Hennessey, C. H. Chung, and J. A. Califano. Activation of the NOTCH pathway in head and neck cancer. *Cancer Res.*, 74(4):1091–1104, Feb 2014.

[66] C. Kwon, P. Cheng, I. N. King, P. Andersen, L. Shenje, V. Nigam, and D. Srivastava. Notch post-translationally regulates ÃÂ² − $cateninproteininstemandprogenitorcells.\ Nat.CellBiol.$, 13(10) : 1244 − −1251, Aug2011.

[67] T. Borggrefe, M. Lauth, A. Zwijsen, D. Huylebroeck, F. Oswald, and B. D. Giaimo. The Notch intracellular domain integrates signals from Wnt, Hedgehog, TGFÃÂ²/$BMPandhypoxiapathways.\ Biochim.Biophys.Acta$, 1863(2) : 303 − −313, Feb2016.

[68] J. Frigola, R. Sabarinathan, L. Mularoni, F. Muinos, A. Gonzalez-Perez, and N. Lopez-Bigas. Reduced mutation rate in exons due to differential mismatch repair. *Nat. Genet.*, 49(12):1684–1692, Dec 2017.

[69] M. M. Li, M. Datto, E. J. Duncavage, S. Kulkarni, N. I. Lindeman, S. Roy, A. M. Tsimberidou, C. L. Vnencak-Jones, D. J. Wolff, A. Younes, and M. N.

Nikiforova. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J Mol Diagn*, 19(1):4–23, 01 2017.

[70] K. Kai, M. Satake, and O. Tokunaga. Gastric adenocarcinoma of fundic gland type with signet-ring cell carcinoma component: A case report and review of the literature. *World J. Gastroenterol.*, 24(26):2915–2920, Jul 2018.

[71] C. Minatsuki, N. Yamamichi, K. I. Inada, Y. Takahashi, K. Sakurai, T. Shimamoto, Y. Tsuji, K. Shiogama, S. Kodashima, Y. Sakaguchi, K. Niimi, S. Ono, T. Niwa, K. Ohata, N. Matsuhashi, M. Ichinose, M. Fujishiro, Y. Tsutsumi, and K. Koike. Expression of Gastric Markers Is Associated with Malignant Potential of Nonampullary Duodenal Adenocarcinoma. *Dig. Dis. Sci.*, Jun 2018.

[72] Y. Zhang, H. Wang, J. Wang, L. Bao, L. Wang, J. Huo, and X. Wang. Global analysis of chromosome 1 genes among patients with lung adenocarcinoma, squamous carcinoma, large-cell carcinoma, small-cell carcinoma, or non-cancer. *Cancer Metastasis Rev.*, 34(2):249–264, Jun 2015.

[73] K. L. Kanchi, K. J. Johnson, C. Lu, M. D. McLellan, M. D. Leiserson, M. C. Wendl, Q. Zhang, D. C. Koboldt, M. Xie, C. Kandoth, J. F. McMichael, M. A. Wyczalkowski, D. E. Larson, H. K. Schmidt, C. A. Miller, R. S. Fulton, P. T. Spellman, E. R. Mardis, T. E. Druley, T. A. Graubert, P. J. Goodfellow, B. J. Raphael, R. K. Wilson, and L. Ding. Integrated analysis of germline and somatic variants in ovarian cancer. *Nat Commun*, 5:3156, 2014.

[74] D. Zhang, L. Qu, B. Zhou, G. Wang, and G. Zhou. Genomic variations in the counterpart normal controls of lung squamous cell carcinomas. *Front Med*, 12(3):280–288, Jun 2018.

[75] D. Usener, D. Schadendorf, J. Koch, S. Dubel, and S. Eichmuller. cTAGE: a cutaneous T cell lymphoma associated antigen family with tumor-specific splicing. *J. Invest. Dermatol.*, 121(1):198–206, Jul 2003.

[76] A. Salas, J. Pardo-Seco, M. Cebey-Lopez, A. Gomez-Carballa, P. Obando-Pacheco, I. Rivero-Calle, M. J. Curras-Tuala, J. Amigo, J. Gomez-Rial, F. Martinon-Torres, A. Justicia-Grande, B. Morillo, L. Redondo-Collazo, C. Rodriguez-Tenreiro, R. Barral-Arca, S. Pischedda, J. Pena-Guitian, C. Curros Novo, M. Puente-Puig, R. Leis-Trabazo, N. Martinon-Torres, J. M. Martinon-Sanchez, M. F. Fraga-Rodriguez, J. R. Antunez, E. Bernaola-Iturbe, L. Moreno-Galarraga, J. Alvarez, T. Gonzalez-Lopez, D. Suarez-Vazquez, A. Vazquez Vazquez, S. Rey-Garcia, F. Gimenez-Sanchez, M. S. Forte, C. Calvo-Rey, M. L. Garcia-Garcia, I. Oulego-Erroz, D. Naranjo Vivas, S. Lapena, P. Alonso-Quintela, J. Martinez-Saenz de Jubera, E. Garrido-Garcia, C. Calvo Monge, E. Onate-Vergara, J. de la Cruz Moreno, M. D. C. Martinez-Padilla, M. Baca-Cots, D. Moreno-Perez, S. Beatriz-Reyes, and M. C. Leon-Leon. Whole Exome Sequencing reveals new candidate genes in host genomic susceptibility to Respiratory Syncytial Virus Disease. *Sci Rep*, 7(1):15888, Nov 2017.

[77] M. Ranzani, V. Iyer, X. Ibarra-Soria, M. Del Castillo Velasco-Herrera, M. Garnett, D. Logan, and D. J. Adams. Revisiting olfactory receptors as putative drivers of cancer. *Wellcome Open Res*, 2:9, Feb 2017.

[78] C. C. Reyes-Gibby, J. Wang, M. R. Silvas, R. K. Yu, E. Y. Hanna, and S. Shete. Genome-wide association study suggests common variants within RP11-634B7.4 gene influencing severe pre-treatment pain in head and neck cancer patients. *Sci Rep*, 6:34206, Sep 2016.

[79] B. Ma, T. Liao, D. Wen, C. Dong, L. Zhou, S. Yang, Y. Wang, and Q. Ji. Long intergenic non-coding RNA 271 is predictive of a poorer prognosis of papillary thyroid cancer. *Sci Rep*, 6:36973, 11 2016.

[80] J. Guo, J. Huang, Y. Zhou, Y. Zhou, L. Yu, H. Li, L. Hou, L. Zhu, D. Ge, Y. Zeng, B. Guleng, and Q. Li. Germline and somatic variations influence the somatic mutational signatures of esophageal squamous cell carcinomas in a Chinese population. *BMC Genomics*, 19(1):538, Jul 2018.

[81] L. Qiu, X. Tan, J. Lin, R. Y. Liu, S. Chen, R. Geng, J. Wu, and W. Huang. CDC27 Induces Metastasis and Invasion in Colorectal Cancer via the Promo-

tion of Epithelial-To-Mesenchymal Transition. *J Cancer*, 8(13):2626–2635, 2017.

[82] A. B. Georgi, P. T. Stukenberg, and M. W. Kirschner. Timing of events in mitosis. *Curr. Biol.*, 12(2):105–114, Jan 2002.

[83] T. M. Runger. C–>T transition mutations are not solely UVB-signature mutations, because they are also generated by UVA. *J. Invest. Dermatol.*, 128(9):2138–2140, Sep 2008.

[84] I. Hajirasouliha, A. Mahmoody, and B. J. Raphael. A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, 30(12):78–86, Jun 2014.